Me, myself and I: A corpus-based, contrastive study of English and German computer-mediated communication from a Systemic Functional perspective

GENEHMIGTE DISSERTATION

zur Erlangung eines Grades des Doktors der Philosophie
im Fachbereich Gesellschafts- und Geschichtswissenschaften
an der Technischen Universität Darmstadt

Referentinnen:
Prof. Dr. Andrea Rapp
Prof. Dr. Elke Teich

vorgelegt von
Anke Schulz M.A.
aus Verden (Aller)

Tag der Einreichung: 29.08.2014
Tag der mündlichen Prüfung: 01.06.2015

D17

Darmstadt
2015

Abstract

Use of the Internet has opened countless possibilities to access information and to connect with other people. In earlier days, contact was limited to people in the immediate surroundings. New media, like paper, radio or telephones, have opened new channels for communication, and so has the Internet. We no longer need to move our physical bodies in order to see and speak to people who live elsewhere. Physical borders are not relevant for Internet communication. What impact does this have on the language people use? Can we still find differences in the use of two closely related languages, English and German, even though Internet communication may have blurred boundaries?

As the language model against which to compare English and German the author chose Systemic Functional Grammar (SFG). The main assumption in SFG is that any option in a language system serves a certain function for the language user. SFG speaks of three broad functions in human communication, called *metafunctions*: the experiential, the interpersonal and the textual metafunction. With the experiential metafunction, we describe the world around us and inside us; this is realized by the system of transitivity, i.e. process types and participant roles. With the interpersonal metafunction, we establish a relationship between us and our listeners or readers. This is realized by the two systems of modality and negation. Finally, the textual metafunction serves to produce cohesion and is represented by the theme-rheme structure.

The aim of this contrastive study is to show the similarities and differences of language use in a bilingual corpus of computer-mediated communication (CMC). The *Englische und deutsche Newsgroup Texte – Annotiertes Korpus* (EDNA) holds 2 x 10,000 words of newsgroup texts in which people write about either eating disorders or relationship problems. The entire EDNA corpus is manually annotated; the annotation was carried out with the help of the UAM corpus tool. The manual annotation covers all four systems representing the three metafunctions. The analysis is twofold: the first part is a qualitative analysis of transitivity, modality, negation and theme-rheme structure, including a test for statistical significance. The second part is an analysis of the lexical

items which are most frequently used to express the systems described in the first part.

The results suggest that the German writers use significantly more modality and negation than the English writers. Relational processes (processes of being and having) are the most frequent ones in both sub-corpora. Following these, German writers prefer action processes (processes of doing) to mental processes (processes of thinking, feeling and perceiving), whereas English writers use more mental than action processes. The first and main participant roles, usually serving as the subject, are almost exclusively realized by pronouns, most commonly *I* / *ich,* and thus say little about the content of the text. In the newsgroup texts by German writers, there are more marked topical themes, i.e. constituents other than subjects stand in the first position of a declarative clause. In the English texts, these marked topical themes are mainly temporal circumstances, while in the German texts, writers refer to themselves with words like *mir, mich, für mich*.

The present study is a comprehensive contrastive analysis of a new register, CMC, in English and in German. It does not limit itself to selected grammatical or lexical features but gives an extensive description and comparison of the language systems and language use in a corpus of CMC by using SFG as linguistic model. There are differences in the language systems, and differences in the frequencies of using the available options. These, however, are outnumbered by the similarities.

Zusammenfassung

Das Internet hat unzählige Möglichkeiten eröffnet, z.B. den Zugang zu Informationen und die Kontaktaufnahme mit anderen Menschen. In früheren Zeiten war eine Kontaktaufnahme beschränkt auf Menschen in der unmittelbaren Umgebung. Neue Medien wie das Papier, Radio oder Telefon haben neue Kanäle für die Kommunikation geschaffen, ebenso das Internet. Niemand muss sich selbst mehr bewegen, um Menschen zu sehen und zu sprechen, die woanders leben. Geographische Grenzen spielen in der Internetkommunikation kaum eine Rolle. Welchen Einfluss hat das auf den Sprachgebrauch der Menschen? Lassen sich noch Unterschiede feststellen zwischen zwei eng verwandten Sprachen, Englisch und Deutsch, obwohl Grenzen durch die Internetkommunikation verwischt werden?

Das Modell von Sprache, das dem Vergleich zugrunde liegt, ist die systemisch-funktionale Grammatik (SFG). Die grundsätzliche Annahme in der SFG ist, dass jede Option in einem Sprachsystem eine bestimmte Funktion für die Benutzerinnen erfüllt. Die SFG kennt drei übergeordnete Funktionen, genannt *Metafunktionen*: die inhaltliche, die zwischenmenschliche und die textbildende Metafunktion. Durch die inhaltliche Metafunktion beschreiben wir die Welt um uns herum und in uns, dies wird durch das System der Transitivität, d.h. Prozesstypen und Teilnehmerrollen, dargestellt. Mit den Systemen der zwischenmenschlichen Metafunktion, Modalität und Negation, stellen wir eine Beziehung her zwischen uns und unseren Gesprächspartnerinnen oder Leserinnen. Die textbildende Metafunktion schließlich dient dazu, aus einzelnen Worten einen zusammenhängenden Text zu schaffen, hier ist die Thema-Rhema-Struktur von Bedeutung.

Es ist das Ziel dieser kontrastiven Studie, Gemeinsamkeiten und Unterschiede im Sprachgebrauch zu zeigen in einem bilingualen Korpus computervermittelter Kommunikation (*computer-mediated communication*, CMC). Das *Englische und deutsche Newsgroup-Texte – Annotiertes Korpus* (EDNA) beinhaltet 2 x 10.000 Wörter aus Texten aus Diskussionsforen. In diesen Foren schreiben Personen entweder über ihre Essstörungen oder über ihre Beziehungsprobleme. Das gesamte EDNA Korpus ist manuell annotiert mit Hilfe des UAM Cor-

pus Tools. Die manuelle Annotation umfasst alle vier Systeme, die die drei Metafunktionen repräsentieren. Die Auswertung besteht aus zwei Schritten: Der erste Teil ist eine quantitative Auswertung der Transitivität, der Modalität, der Negation und der Thema-Rhema-Struktur. Sie beinhaltet einen Test zur statistischen Signifikanz der Ergebnisse. Im zweiten Teil werden die Wörter untersucht, die am häufigsten innerhalb der vier Systeme gebraucht werden.

Die Ergebnisse zeigen, dass die deutschen Autorinnen signifikant mehr Modalität und Negation ausdrücken im Vergleich zu den englischen. Konstante Prozesse (*relational processes*, Sein und Haben) sind in beiden Sub-Korpora die am häufigsten verwendeten Prozesse. Darauf folgen in den deutschen Texten die handelnden Prozesse (*action processes*) vor den mentalen (*mental processes*, Prozesse des Fühlens, Denkens und Wahrnehmens). In den englischen Texten jedoch werden mehr mentale als handelnde Prozesse verwendet. Die erste und wichtigste Teilnehmerrolle, die in der Regel das Subjekt des Satzes ist, wird fast ausschließlich durch Personalpronomen, insbesondere *I / ich*, realisiert und sagt somit wenig über den Inhalt der Texte. In den deutschen Texte aus den Newsgroups gibt es signifikant mehr markierte topikalische Themen, d.h. ein Objekt, Komplement oder Adjunkt steht an erster Stelle im Deklarativsatz, nicht das Subjekt. In den englischen Texten sind dies Zeitangaben, während die deutschen Schreiberinnen auf sich selbst Bezug nehmen mit Worten wie *mir, mich, für mich*.

Die vorliegende Arbeit ist eine umfassende kontrastive Analyse eines neuen Registers, computer-vermittelter Kommunikation, in Englisch und in Deutsch. Die Arbeit beschränkt sich nicht auf einzelne grammatikalische oder lexikalische Merkmale. Vielmehr liefert sie ausführliche Beschreibungen und Vergleiche der Sprachsysteme und des Sprachgebrauchs auf Grundlage von Daten in einem Korpus, aufbauend auf die systemisch-funktionale Grammatik-Theorie. Es werden Unterschiede in den Sprachsystemen gezeigt, ebenso wie im Sprachgebrauch. Diese Unterschiede werden jedoch zahlenmäßig von den Gemeinsamkeiten übertroffen.

Acknowledgements

First and foremost, I am most grateful for Prof. Dr. Elke Teich's intellectual guidance, moral support, and continuous encouragement at every step of the way. She is the best role model, both for academic and social abilities. Second, I am obliged to Prof. Dr. Andrea Rapp, who did not hesitate to step in as a reviewer of this PhD thesis and who provided me with the incentive to finally finish. Thank you also to Prof. Dr. Dieter Schott and Prof. Dr. Nina Janich for volunteering to join the examination board.

I am indebted to Robin Fawcett and Uwe Helm Petersen for letting me use their yet unpublished manuscripts, and to Michael O'Donnell and Geoff Thompson. They all generously shared their knowledge with me, inspired me and provided practical support. A big thank-you goes to Paul Rayson, for letting me use the free online CLAWS PoS-tagger on much longer texts than had been intended.

I also express my gratitude towards the people I was lucky enough to meet along the way; Sabine Bartsch, for countless good ideas and discussions; Mônica Holtz, for innumerable occasions where she offered practical help with all sorts of things; Richard Eckart de Castilho, for help with PoS-tagging the German texts with the STTS tag set and many other computer-related challenges; Tatsiana Markovic, who did a great job annotating the EDNA corpus; Andrew Harrison and Patrick Bal at TU Darmstadt for their friendship; Susanne Dyka, Marcus Callies and John Bateman at the University of Bremen, who kept supporting me even when my PhD thesis was long overdue, and Sandra James for being such a wonderful proofreader.

Finally, I would not be where I am today without my wonderful friends and family. I owe all of them a debt of gratitude: Dietmar Pöppke, for never ceasing to love and to support me; Sabine Korell and Heike Krichbaum, for taking me out to dance; Maria Fröhlich, for always believing I could do this, and last but not least Elke, Hans, Anneliese, Herbert, Thorsten, Christa, Hannes and Thea. A special thank you to Marie Könemann.

Wissenschaftlicher Werdegang

Anke Schulz

geboren in Verden / Aller

1989	Allgemeine Hochschulreife, Berufsbildende Schulen Verden
1993-1996	Diplomstudium Übersetzen und Dolmetschen, Fachhochschule Köln, ohne Abschluss
1996-2003	Magisterstudium Anglistik und Romanistik, Schwerpunkt Sprachwissenschaft, Universität Bremen
2006-2010	Wissenschaftliche Mitarbeiterin und Doktorandin von Prof. Dr. Elke Teich, Englische Sprachwissenschaft, Institut für Sprach- und Literaturwissenschaft, Fachbereich 2, Gesellschafts- und Geschichtswissenschaften, Technische Universität Darmstadt
seit 2011	Universitätslektorin für Englische Sprachwissenschaft mit halber Stelle, Fachbereich 10, Sprach- und Literaturwissenschaften, Universität Bremen

Contents

List of Abbreviations

CAMA	Computer-assisted manual annotation
CL	Contrastive linguistics
CMC	Computer-mediated communication
EDNA	Englische und deutsche Newsgroup-Texte – annotiertes Korpus
EN	English part of EDNA
GN	German part of EDNA
PoS	Part-of-speech
PR	Participant role
PrEx	Process extension
SFG	Systemic Functional Grammar
SFL	Systemic Functional Linguistics
UAM CT	Universidad de Autónoma de Madrid corpus tool

1 Introduction

> "The research literature is characterized
>
> by a great deal of theoretical speculation
>
> but relatively few empirical studies."
>
> David Crystal 2011: 13

Use of the Internet has opened countless possibilities to access information and to connect with other people. In earlier days, contact was limited to people in the immediate surroundings. New media, like paper, radio or telephones, has opened new channels for communication, and so has the Internet. Crystal (2011, 238) calls Internet communication a "development of millennial significance. […] A new medium of linguistic communication does not arrive very often in the history of the race." We no longer need to move our physical bodies in order to see and speak to people who live elsewhere. Physical borders are not relevant for Internet communication. What impact does this have on the language people use? Can we still find differences in the use of two closely related languages, English and German, even if Internet communication may have blurred boundaries? This study has a descriptive background. "But it is an appropriate background to have, for the one thing Internet language needs, more than anything else, is good descriptions" (Crystal 2011, x).

1.1 Objectives of the study

In this study, I want to explore the differences and similarities in the use of the English and the German language in a rather new register of language, i.e. Internet language as it is used in newsgroups. If we consider that English and German are closely related languages, and that the topics discussed in these newsgroups are the same, what differences (or similarities) can we find? Can these differences be explained by differences in the language systems, or by differences in the use of the languages? The overall aim is to contribute to a comprehensive contrastive linguistic description of English and German based on corpus data.

The use of corpora in contrastive linguistic studies has only just begun, e.g. Johansson and Hofland (1994, 141), Granger et al. (2003), Johansson (2007). In the past, linguistic studies were carried out with rather small databases as a source of authentic language material, simply because larger amounts of accessible language data were not available before the 1960s (Matthews 2005, 78). Since then, the development of personal computers has led to the advent of corpus studies, i.e. "any systematic collection of speech or writing in a language or variety of a language" (Matthews 2005, 78). This new approach to language description, corpus linguistics,

> aims to base accounts of languages on corpora derived from systematic recordings of real conversations and real discourse of other kinds, as opposed to examples obtained by introspection, by the judgement of the grammarians, or by haphazard observation. (Matthews 2005, 78)

Contrastive linguistics benefits from systematic recordings of real discourse to provide sound evidence. So far, however, there are hardly any corpora of Internet language, let alone in languages other than English. For my research, I use a DIY corpus. It is a corpus of *computer-mediated communication* (CMC), i.e. of communication using computer technology. The name of my self-made corpus is EDNA (*Englische und Deutsche Newsgroup Texte – Annotiertes Korpus*). EDNA is a bilingual comparable corpus with 2 x 10,000 words of texts on relationship problems or eating disorders. These texts have been collected from Internet newsgroups in 2004-6. Thus, the first of three goals (apart from the overall aim) of my research is to base a contrastive linguistic description of the English and German language on corpus data.

In the past, contrastive studies of languages often had to be limited to individual aspects of the language systems involved, e.g. tense and aspect (Dahl 2000), modal verbs (Salkie 2008), or cohesive substitution (Kunz and Steiner 2013). Differences in the language system, i.e. the options that are available to speakers, however, say little about the use of the options. Furthermore, they are but pieces of a puzzle that are waiting to be put together. The present study uses a comprehensive model of language to describe the systems and the use of the options within the systems to overcome those limitations: the *Systemic Functional Grammar* theory.

Systemic Functional Grammar (SFG) is a model of language that is based on the assumption that every option from a linguistic system which is available to a speaker or writer does serve a certain purpose. SFG defines three so-called metafunctions, i.e. three general functions of any human language (Halliday 1994): first, to communicate a speaker's ideas and experiences (experiential metafunction), second, to relate to the other person(-s) involved in the discourse (interpersonal metafunction), and third, to structure discourse in a way that makes it coherent and cohesive (textual metafunction). With this broad approach, SFG provides a very rich theoretical background for the description of a language. A contrastive study of two languages based on SFG will in turn become more comprehensive than descriptions which focus on individual systems alone. In this respect, the present study is very innovative. The English and the German language have been the subjects of many contrastive studies over a number of years. Few were done from a Systemic Functional Grammar viewpoint, e.g. Steiner (1987), Steiner (2001), Teich (1999), Teich (2003). Therefore, the second goal is to contribute to the field of contrastive linguistics by approaching the data with a specific theoretical framework, namely Systemic Functional Grammar.

Both the work with corpora and the model of SFG have developed over the past fifty years. Corpora in the sense that we use today did not begin to emerge until the 1960s. Quirk started working on the ground-breaking *Survey of English Usage* (SEU) in 1960, Francis and Kucera began their work on the *Brown Corpus* of written American English in 1961, Svartvik used the SEU to build the *London-Lund Corpus* of spoken texts beginning in 1975. All these projects took several years to completion (Baker, Hardie, and McEnery 2006) (McEnery and Hardie 2012).

Michael Halliday started to build up SFG theory beginning in the 1960s (Halliday 1961). The first publication of his *An Introduction to Functional Grammar* dates back to 1985 (Halliday 1985) and is now available in its fourth edition (Halliday and Matthiessen 2014). Systemic Functional Grammar theory and the methodology of using corpora for linguistic studies, however, have not been joined to any satisfying degree yet. One possible explanation for this lack of merging the two successfully is this: Whereas in corpus linguistics a lot of effort is put into the development of automatic annotation and queries, the identification of SFG features, e.g. the semantic roles of participants in a process, cannot be done automatically – at least not until computers have learnt to distinguish not only strings of words, but also different meanings that words can have. Following from this, the third goal of the present study is to apply SFG theory to the annotation of a corpus and to find out whether the results can contribute to a better description of the two languages, English and German.

EDNA (*Englische und Deutsche Newsgroup Texte – Annotiertes Korpus*) is the first corpus which is deeply annotated for SFG features. There is now a corpus of 10,000 words of English texts with annotation of SFG features (named EN), and also a German annotated corpus of the same size (named GN). This German part of the EDNA corpus is a novelty in two ways: first, it is the first corpus which is annotated for SFG features. Second, the German part of EDNA is pioneering work because a Systemic Functional Grammar of the German language remains underdeveloped at the time of writing (2014). Guidelines for the annotation had to be written with very little reference material. Thus, the German part of EDNA does not only serve the contrastive study at hand alone, but also contributes to a better description of the German language in SFG terms.

If the aim is a more all-embracing description of language, we must expose ourselves to large amounts of ‘real’ data with an open mind to the outcome. The study cannot be limited by using only individual example sentences. Thus, corpus linguistics is the methodology of choice here. A corpus gives access to whatever is there to be found, perhaps even some aspects that had not been imagined at the beginning of the study. With SFG as the theoretical background, the study covers a wide range of systems and options. It contributes empirical evidence to the new field of Internet linguistics in two languages, English and German.

1.2 The hypotheses

The description of the language systems of English and German according to SFG and the use of language in the EDNA corpus proceeds in the following way: The first step is a description of the systems that realize the metafunction. For the interpersonal metafunction, we look at modality and polarity. For the textual metafunction, there is a study of the theme-rheme structure. Finally, the experiential metafunction is described through the system of transitivity, i.e. process types and participant roles. The second step is a quantitative investigation of the systems. The third step is an investigation of the lexical items that realize the individual SFG features. Herein, the focus is on the most frequent modal auxiliaries and adverbs as well as the most frequent negation markers, the lexical items which most frequently realize the different themes and the words or phrases which most frequently realize the different process types and participant roles. For each of the four systems, there is a hypothesis that can be tested.

My hypothesis with regard to the interpersonal metafunction is that both English and German newsgroup text writers use modality in the same way and to the same extent. The relationship between writers and readers of the newsgroup texts is the same in both languages; therefore I expect that writers position themselves in a similar way towards their readers through the use of modal auxiliaries and modal adjuncts.

My second hypothesis is about the use of negative polarity, i.e. syntactic and morphological negation. The English and the German language have the same options for expressing the negative, for example syntactic negation markers on clause rank like *not* / *nicht* and on phrase rank like *no* / *keine*, as well as morphological negation markers like the prefixes *un-* / *im-* / *dis-* / *non-* in English and *un-* / *miss-* / *des-* in German. The topics which are discussed in the newsgroup texts collected in EDNA are the same. Therefore I expect that syntactic and morphological negation markers are used in the same way and to the same extent in the two EDNA subcorpora.

The third hypothesis concerns the textual metafunction, represented in the texts through the structure of theme and rheme. In this system, I do expect a difference in the English and German texts, hence a difference in the use of marked and unmarked topical themes, i.e. the position of subjects, objects, complements and adjuncts in a clause. In the English language, the word order of subject, verb and objects is much more fixed than in the German language:

> There are several ways of encoding grammatical relations, two of which are exemplified by the languages under comparison. German uses case marking for that purpose. The nominative case identifies the subject, while objects are commonly encoded in the accusative (direct object) or the dative (indirect object […]). […] In English, by contrast, grammatical relations are identified by linear order. The subject is that constituent which precedes the finite verb, whereas objects follow the main verb. (König and Gast 2007, 103)

Thus, considering the greater freedom within the German language to place the subject before or after the verb, we may expect that the German newsgroup writers make use of this freedom and put other constituents before the verbs for emphasis.

Finally, the fourth hypothesis relates to the system of transitivity which represents the experiential metafunction. I assume that the world which is represented through language is more or less similar for English and German writers. The English writers come from the UK or the USA. The German writers come from Germany or Austria or Switzerland. All of these countries are fully developed ‘Western’ countries, and industrialized nations. Consequently, we may assume that their cultures do not differ to any a great extent. Both English and German writers in EDNA talk about the same problems, eating disorders and relationship problems, and I expect that they do so in similar ways due to their similar cultural background. Therefore, my hypothesis is that both parts of EDNA contain roughly the same kinds of process types to the same extent. I expect that the English and German writers use mostly relational processes to describe the state of their eating disorder or relationship problem. The second most frequent type is probably mental processes, as I assume that writers describe how they feel about their problems. I do not assume that action processes are more frequent than relational or mental processes, as these texts are not narrating stories but describing problems.

As for participant roles in these processes, I anticipate that both English and German writers use personal pronouns to talk about the people involved in these problematic situations, e.g. *I, she* and *he*, as well as words referring to their problems, for example *gaining weight, kilogram, food*, or *partner, girlfriend, marriage, divorce*. Both parts of EDNA should contain roughly the same lexemes in the semantic roles (participant roles) in their descriptions of problems with food and eating or with their respective partners.

These are my four basic hypotheses. By using a deeply annotated corpus, I will be able to verify or falsify my hypotheses by the end of the study. In addition, however, I hope to find other details which were not predictable, in order to give a more comprehensive description of the use of the English and German language in this kind of discourse, newsgroup texts.

1.3 Organization of the thesis

Following this introduction, chapter 2 gives a brief overview of the theoretical framework and the state-of-the-art of the four fields relevant to this project: Systemic Functional Grammar, corpus linguistics, contrastive linguistics and computer-mediated communication. Chapter 3 describes the methodology for this research project, including a description of the EDNA corpus. Chapter 4 provides some basic corpus statistics and an analysis of the lexical density and grammatical intricacy of the EDNA corpus (Halliday 1989a). Chapter 5 is the first chapter that presents results from the study, namely the results for modality and negative polarity, i.e. for the interpersonal metafunction. Chapter 6 then continues with the presentation of the results from the study of theme-rheme structure, representing the textual metafunction. Chapter 7 is about process types and participant roles, thus the experiential metafunction. Finally, chapter 8 discusses the overall results of the research project. Was it possible to provide a comprehensive description of the English and the German language as they are used in computer-mediated communication by studying an annotated corpus? Is the SFG framework useful for corpus annotation? The next chapter thus begins with the theoretical framework which has been found to be most useful for this study, Systemic Functional Grammar (Halliday 1994).

2 Theoretical background and state-of-the-art

In this chapter, the theory and the main fields of linguistics which are important for the present study are briefly introduced. We begin with Systemic Functional Grammar theory (SFG), followed by corpus linguistics, contrastive linguistics and computer-mediated communication (CMC). Each successive subchapter begins with a definition of the relevant terminology, followed by a summary of the state-of-the-art in the respective field.

2.1 Systemic Functional Grammar

2.1.1 Defining the terms in SFG

Systemic Functional Grammar (SFG) or Systemic Functional Linguistics (SFL) has developed from Michael Halliday's work, beginning in the 1960s (Halliday 1961). Halliday's language model grew from the works of J.R. Firth, Bronislaw Malinowski, Louis Hjelmslev, and the Prague School of Linguistics (Neumann 2003, 46). This new theory of language, more descriptive than prescriptive, is based on the assumption that people use language for a purpose, to achieve a goal. From its origins in the UK and Australia, it was embraced quickly and globally. More and more research was conducted using SFG as a framework, by a growing number of scientists. By the mid-1980s, this theory of language description had also been used for languages other than English (Halliday and Matthiessen 2014, xiii). The basic reference book, Halliday's *Introduction to Functional Grammar* (IFG), first published in 1985, with a second edition in 1994, the third in 2004 with Christian Matthiessen, is now in its fourth edition (Halliday and Matthiessen 2014).

Systemic Functional was chosen as the theoretical framework because it is most suitable for a systematic and extensive description of language in use. Halliday says about the IFG that "[t]he aim has been to construct a grammar for purposes of text analysis: one that would make it possible to say sensible and useful things about any text, spoken or written, in modern English" (Halliday 1994, xv). Halliday goes on to explain why the theory is called 'functional'. To start with, the grammar is functional in three respects (Halliday 1994, xiii-xiv):

> (1) It is functional in the sense that it is designed to account for how the language is **used**. […] Language has evolved to satisfy human needs; and the way it is organized is functional with respect to these needs – it is not arbitrary.
>
> (2) […], the fundamental components of **meaning** in language are functional components. All languages are organized around two main kinds of meaning, the 'ideational' or reflective, and the 'interpersonal' or active. Human beings use language to reflect on their environment, and to interact with others in their environment.
>
> (3) Thirdly, each **element** in a language is explained by reference to its function in the total linguistic system.

To take up Halliday's second point, every time a speaker or writer uses language it serves two functions simultaneously. These general functions of language are called the *ideational* and *interpersonal* metafunction. The third metafunction in SFG framework is the *textual* metafunction.

With the ideational metafunction people describe their experience of the world around and within them. This is expressed by process types, i.e. types of verbs such as *material*, *relational* or *mental*, and the semantic roles (or participant roles) involved in these processes, e.g. *actor* and *goal*, *senser* and *phenomenon*, *identifier* and *identified*. Note that these are the terms of the so-called Sydney Grammar, and those used in the IFG (Halliday and Matthiessen 2004). In my work with the experiential metafunction here, however, I will use the terminology and framework provided by what is called the Cardiff Grammar, based on work by Fawcett (Fawcett forthcoming) and his colleagues. Fawcett employs the terms *action*, *relational* and *mental* process, the participant roles in these main processes are *agent* and *affected*, *carrier* and *attribute*, and *emoter* and *phenomenon*.

With the second metafunction people express their relationship to the other participants in a discourse. Such relationships are formed by social hierarchies, closeness and frequency of contact. The interpersonal metafunction is realized through sentence type (mood), modality and polarity.

Finally, coherent and cohesive discourse is created through the textual metafunction; this is mainly represented by the theme-rheme structure. Details regarding the three metafunctions will be explained in the respective chapters.

The Systemic Functional Grammar is 'systemic' in the way it describes the options that speakers can choose from whenever they say or write something. For example, speakers can state whether something is or is not the case, as in *dancing is allowed* / *dancing is not allowed*. In SFG, the options are displayed as system networks, thus a very simple system network for polarity would look as in figure 2.1 below.

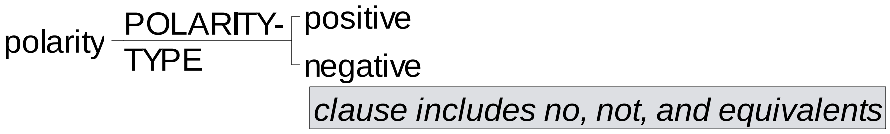

Figure 2.1 Simple system network for polarity

All systems which are relevant during the course of this study will be explained in later chapters. For a more detailed introduction to SFG, the interested reader is referred to the wealth of introduction books to SFG, especially Thompson (2014), Fontaine (2013) and Coffin, Donohue and North (2009), which are written in English. For introductions to the topic written in German, see Steiner (1983) and Neumann (2003).

2.1.2 Previous work on SFG of the English language

The main reference work itself, Halliday's *Introduction to Functional Grammar* (IFG), has undergone some revisions and improvements. Matthiessen (Halliday and Matthiessen 2014, xiv) states that by the time he and Halliday were working on IFG 3 (2004), the "ecological niche in which IFG operates [had] thus changed considerably". We cannot call it a niche anymore, for a start. While preparing IFG 4, Matthiessen continued working with corpora, among these the Australian Corpus of English (ACE), the Corpus of Contemporary American English (COCA), and the International Corpus of English (ICE) (Halliday and Matthiessen 2014, xv). Corpora in the IFG, however, serve only as a source for authentic examples. The IFG does not describe corpus work in the sense of frequency statistics or concordance analyses, for instance.

There is a multitude of studies based on Halliday's SFG, both developing the theory further for the description of English and other languages, and applying the theory to practical tasks like language teaching, translator training, and text analyses of all kinds. Yan and Webster's (2014) *Developing Systemic Functional Linguistics* "sums up SFL's fifty years of refinement as a theory" (from the publisher's description). SFL is strong in the English-speaking countries, and I will not go into more detail here. In German-speaking countries, however, the situation is different.

2.1.3 Previous work on SFG of the German language

In Neumann (2003, 56) we find her paraphrasing Halliday's statement that even though IFG started out as an introduction to the functional grammar of English, it can be read as a functional grammar in general, with "English as the language of illustration" (Halliday 1994, xxxiii). Neumann says that Systemic Functional Linguistics started from the description of the English language. Nevertheless, she says, this description can be the starting point for the description of any other language, since the basic functions of language are comparable. The grammatical structures that are used to realize these functions, however, may well be different (Neumann 2003, 56).

Systemic Functional Grammars of languages other than English do exist; Caffarel, Martin and Matthiessen (2004) provide an extensive collection, including languages such as, for example, French, Japanese, Chinese, and Pitjantjatjara, a language spoken by Australian Aborigines. A full SFG of French followed in 2006 (Caffarel 2006), and the SFG of Spanish was published in 2010 (Lavid, Jorge, and Zamorano-Mansilla 2010).

Unfortunately, at the time of writing (2014), there is still no comprehensive SFG of the German language. The best attempt at such a grammar is Petersen's manuscript, which awaits publication (Petersen forthcoming). Petersen, a Danish scholar, is among the first to write about the functional grammar of German (Petersen 2004). It is his manuscript that serves as the main reference for SFG of German in this study.

Credits for the very first publications concerned with an SFG of German, however, must go to the scholars at Saarland University, with Steiner, Eckert, Weck and Winter (1988) using SFG for work within the EUROTRA-D project, and Teich (1991) on how an SFG of German can be used in text generation. Issues in natural language generation have triggered research using SFG in Germany, for example Teich (1999), who also deals with the German language. Steiner and Teich were the scholars contributing to Caffarel, Martin and Matthiessen's (2004) collection in *Language Typology*, with their chapter named "Metafunctional profile of the grammar of German". Their work serves as theoretical background for later chapters in this study, alongside Petersen's. In recent years, the systemic functional framework is most actively being used by a small group of teachers in Switzerland, e.g. Alan Hess (2014) and Eckart Störmer (2009). As a theory, it is most influential at Saarland University, see the references in chapter 2.3.1.

Following the description of individual languages, contrasting the commonalities and differences of two (or more) languages using SFG is the logical next step. Neumann (2003) presents a corpus analysis of English and German travel guides, with SFG as the underlying theory. Her book is written in German and therefore one of the very few sources for scholars who are interested in SFG but speak no or little English. The lack of literature written in German about SFG in general, and about the SFG of the German language in particular may be the major reason why this theory is not more widespread in German-speaking countries. Teich (2003) provides a framework based on SFG to study how and to what extent translations of English and German texts differ from originals written in these two languages. Finally, Neumann (2014) makes extensive use of quantitative analyses of linguistic features to describe register variation across texts written in English and German, as well as translations. The present study is different in that it uses a manually annotated corpus, which had not been previously available.

> Stated in other terms, a grammar is an attempt to crack the code. Each language has its own semantic code, although languages that share a common culture tend to have codes that are closely related. […] The main problem for linguists is to give an objective account of the code. (Halliday 1994, xxx)

Thus, SFG categories provides the basis for an objective account of the code of English and German as it is used in the EDNA corpus. We shall see just how closely related the codes of these two languages are, and how similar the cultures they represent.

2.2 Working with a corpus

2.2.1 Defining 'corpus'

First of all, what is a corpus? Baker, Hardie and McEnery (2006, 48) explain that "[i]n linguistics a corpus is a collection of texts (a 'body' of language) stored in an electronic database". Working with a collection of existing texts, instead of a set of invented examples, serves one basic function. To quote Baker (2010, 95), "[m]ost research questions in corpus linguistics are based around one overarching question: 'how do people *really* use language?'" In a corpus, large amounts of 'real' language data can be collected and investigated. This data can be either written or spoken language. Biber (1990) says that corpus studies have their beginnings in the early 20th century, with Boas and Sapir collecting spoken material of Native American languages as database for their linguistic studies. In the 21st century, a corpus is usually understood to be machine-readable, i.e. stored on a computer and processed with computer software programmes.

There are two ways to approach the data in a corpus (Baker 2010, 95). First, a scientist can explore the language data without any preconception of existing language theories; this is called a corpus-driven study.

Second, a scientist can start investigating the corpus with a theory of language as framework. Hypotheses derived from the theory can then be verified or falsified; this is called a corpus-based study. As with so many things, the distinction is not bipolar, but rather a matter of locating one's study on a continuum between corpus-driven and corpus-based. The present study is clearly corpus-based, as there is a theory of language, i.e. Systemic Functional Grammar, which serves as the backdrop to describe the corpus.

Another distinction that linguists make when talking about corpus linguistics is whether it is a methodology or a theory. In my work, I use corpus studies as a methodology to answer a set of research questions, and thus agree with McEnery and Hardie (2012, 6), who state the following:

> Corpus-based studies typically use corpus data in order to explore a theory or hypothesis, typically one established in the current literature, in order to validate it, refute it or refine it. The definition of corpus linguistics as a *method* underpins this approach to the use of corpus data in linguistics.

2.2.2 Different types of corpora

There are various types of corpora. One of the main ways to distinguish corpora is mentioned by Baker (2010, 99): General corpora are built in order to be representative of a language (variety), e.g. the British National Corpus (BNC). They consist of several million words. Specialized corpora are smaller and often built of a certain text type, collected from a certain time period, or from a specified language variety. Specialized corpora serve to answer particular research questions, whereas general corpora are large enough to answer a variety of different questions. Usually, a general corpus is used as a reference corpus for results from a specialized corpus. Researchers can show "what forms of language (e.g. lexis, grammar, topics) are over- or underrepresented in the smaller corpus" (Baker 2010, 99). Following this distinction, the EDNA corpus is a **specialized** corpus.

Baker, Hardie and McEnery (2006, 49) mention further types of corpora apart from general (here called *reference* corpus) and specialized: "Types of corpora include specialised, reference, multilingual, parallel, learner, diachronic and monitor." In addition to being a specialized corpus, EDNA is a **bilingual** corpus, as it contains texts in two languages (Baker, Hardie, and McEnery 2006, 119). Multilingual corpora must contain texts in at least three languages (McEnery and Hardie 2012, 19).

Baker, Hardie and McEnery (2006, 126-7) differentiate further between parallel and comparable corpora. A parallel corpus "consists of the same documents in a number of languages, that is, a set of texts and their translations". The corpus which is used for the present study, *Englische und Deutsche Newsgroup Texte – Annotiertes Korpus* (EDNA) is, however, not a parallel corpus, as it holds no translations, but a **comparable** corpus. A comparable corpus consists of the same kinds of text in different languages, but the texts are not translations of each other. In a comparable corpus like EDNA, the same sampling frame is used for all data: "the *same proportions* of the texts of the *same genres* in the *same domains* in a range of *different languages* in the *same sampling period*" (McEnery and Hardie 2012, 20). Thus, the term 'comparable corpus' is used in the way it is used in contrastive linguistics, because it is a bilingual corpus of original texts. In contrast, a comparable corpus in translation studies is monolingual and contains original and translated texts in the same language (Granger 2003, 20). Details of the sampling method are described in the chapter on methodology.

Another difference between corpora is whether they are dynamic or static. Dynamic corpora are growing continually; new data is added annually (or in other regular intervals). These are also called monitor corpora. Static corpora (diachronic corpora), on the other hand, do not grow once they have been built (Baker, Hardie, and McEnery 2006, 64). EDNA is clearly a **static** corpus. The EDNA corpus can best be described as an opportunistic corpus, even if the term emerged much later than the EDNA corpus. **Opportunistic** corpora "represent nothing more nor less than the data that it was possible to gather for a specific task" (McEnery and Hardie 2012, 11). I can only agree when they say that "it should be noted, and accepted, that the corpora that we use and construct must sometimes be determined by pragmatic considerations" (McEnery and Hardie 2012, 13).

Finally, corpora can be distinguished by the kind of language that is collected in them. The texts in the corpus may be "spoken, written or computer-mediated texts (such as emails, text messages or website) or a mixture of all three" (Baker 2010, 99). All three major types of text can come from many different registers of course. EDNA, however, consists of computer-mediated texts, in particular texts written in newsgroups. More details are given in the methodology chapter.

For an overview of the historical development of corpus linguistics and existing corpora, readers are referred to McEnery and Hardie 2012.

If we locate the EDNA corpus on the continua between the respective opposing features, the matrix would look like in figure 2.2. Naturally, the positioning of the features on the right or left hand side does not indicate an evaluation that one of them is superior to the other.

corpus based	X	corpus driven
corpus as methodology	X	corpus as theory
specialized corpus	X	general corpus
static	X	dynamic
bilingual	X	monolingual
comparable	X	parallel
opportunistic	X	balanced

Figure 2.2 Locating the EDNA corpus

2.2.3 Some issues in corpus linguistics

Balance: Baker, Hardie and McEnery (2006, 18) define a balanced corpus as a corpus which consists of texts from as many different registers, written and spoken, as possible. Thus, a reference corpus would aim at being balanced. Since the EDNA corpus is a specialized corpus, it cannot be not balanced according to this definition.

Representativeness: A corpus should aim to be a "representative sample of a particular language variety" (Baker, Hardie, and McEnery 2006, 139). Representativeness, as Biber (1992, 174) puts it, "refers to the extent to which a sample includes the full range of variability in a population". The population in the corpus, and its boundaries, must be clearly defined before starting to collect samples (Biber 1992, 174). What texts do we want to include in our corpus? What exactly is 'the entire population'? Defining a population and choosing a sampling technique, like corpus annotation, "are an act of interpretation on the part of the corpus builder" (McEnery and Wilson 2001, 78). Once a population and a sampling frame have been defined, however, probabilistic random sampling techniques can be used to collect data. Biber (1992, 175) says that "[i]n a simple random sampling, all texts in the population have an equal chance of being selected".

One challenge here is that in order to collect (random) samples, the entirety of the population to be sampled must be finite. Whatever the size of the population, it must not grow; otherwise the change would skew any attempt at collecting a representative sample of the population. Thus it is virtually impossible to achieve perfect representativeness.

Having said that, I cannot claim that the EDNA corpus is representative, first, because the population of newsgroup texts is growing daily, and rapidly, and second, because texts that were included in the corpus had to fulfil certain criteria with regard to text length and topic. Thus, not every text had the same chance of being selected. The selection criteria are explained in the methodology chapter. There, readers will also find the definition of the population and sampling frame. At this point, I can only take solace from what McEnery and Hardie (2012, 10) state:

> Balance, representativeness and comparability are ideals which corpus builders strive for but rarely, if ever, attain. In truth, the measures of balance and representativeness are matters of degree. […] Similarly, while some corpora designed to be comparable to each other can clearly make a claim for balance and representativeness, others may only do so to a degree. (McEnery and Hardie 2012, 10)

Total accountability (Leech 1992, 112 in McEnery and Hardie 2012, 14) means that it is not permissible to choose from the corpus only those examples that suit one's hypothesis (confirmation bias);

> The principle of total accountability is, simply, that we must not select a favourable subset of the data in this way. When approaching the corpus with a hypothesis, one way of satisfying falsifiability is to use the entire corpus – and all relevant evidence emerging from analysis of the corpus – to test the hypothesis. (McEnery and Hardie 2012, 15)

The thorough way in which the entire EDNA corpus was annotated throughout should provide us with the best possible total accountability.

Replicability: McEnery and Hardie (2012, 16) explain that

> [a] result is considered replicable if a reapplication of the methods that led to it consistently produces the same result. This process of checking and rechecking may be done with the same dataset or it may be done with new datasets.

Results from the EDNA corpus should be replicable by other researchers using the same approach, but bear in mind that manual annotation is never free of inter- and intra-annotator disagreement to a certain degree.

Qualitative versus quantitative analysis: According to McEnery and Wilson's (2001, 76) definition below, my study is a quantitative corpus analysis:

> The difference between qualitative and quantitative corpus analysis, as the terms themselves imply, is that in qualitative research no attempt is made to assign frequencies to the linguistic features which are identified in the data. Whereas in quantitative research we classify features, count them and even construct more complex statistical models in an attempt to explain what is observed, in qualitative research the data are used only as a basis for identifying and describing aspects of usage in the language and to provide 'real-life' examples of particular phenomena. (McEnery and Wilson 2001, 76)

Apart from being a quantitative study, however, EDNA provides us with a wealth of "real-life" examples of particular phenomena. Perhaps this study can best be described as a multi-method approach, where quantitative and qualitative approaches are combined, which does make sense:

> It will be appreciated from this brief discussion that both qualitative and quantitative analyses have something to contribute to corpus study. Qualitative analysis can provide greater richness and precision, whereas quantitative analysis can provide statistically reliable and generalizable results. (McEnery and Wilson 2001, 76)

In this subchapter, I have located my corpus with regard to main concepts in corpus linguistics. In the following, one of the first studies combining SFG and corpus linguistics is briefly summarized.

2.2.4 First steps in combining SFG and corpus linguistics

The combination of SFG and corpus linguistics is still in its early days. Thompson and Hunston (2006) gather contributions by linguists from either an SFL or a corpus studies background that shed some light on questions connecting both fields, from different angles, but these chapters are of a more theoretical nature. One of the few 'hands-on' studies using both SFG as a theoretical framework and a 'corpus' as data is the work by Andrew Goatly (Goatly 2004). In his article, Goatly describes how he used J.K. Rowling's book *Harry Potter and the Philosopher's Stone,* first published in 1997, as data for investigating how ideology is perpetuated and reinforced in children's literature. He investigates concordance lines and word frequencies of the most frequent lexical verbs that realize the main processes in *Harry Potter and the Philosopher's Stone*, as well as participants and circumstances. One of his results is concerned with how human beings make use of their natural environment:

> The point seems to be that animals are represented as doing something significant only if they are magical. So the existence of magical animals does nothing to undermine the pattern that, unless exploitable like owls, or behaving like humans, ordinary animals are not worth attention. (Goatly 2004, 131)

In addition, he shows that girls and women are far more likely to cry, scream and shriek than boys and men are. The boys in *Harry Potter and the Philosopher's Stone*, on the other hand, are more likely to break the rules than the girls. Intense rivalry and hatred between groups is encouraged, and, as Goatly (2004,140) states, that "[t]his produces a mind-set not far removed from that of the actors in the Palestine-Israeli conflict, with tendencies to collective punishment and fascism on both sides". Other topics in Goatly's work are self-control, how the food in the book is predominantly English food, and that students are expected not to be late. Goatly concludes by saying that the world which is reflected in the concordance lines seems to be attractive to young readers, even if not to him, and that "the stance here, apart from tokenism, is fundamentally sexist, and certainly speciesist" (Goatly 2004, 149). Using the categories of SFG for lexical verbs, participants and circumstances and investigating word frequencies and concordance lines has proven to be helpful in revealing this invisible, underlying ideology in J.K. Rowling's book.

2.3 Contrastive linguistics

2.3.1 Contrastive linguistics: Definition and challenges

Contrastive linguistics is defined as "any investigation in which the structures of two languages are compared" (Matthews 2005, 74). The comparison not only describes differences, but also commonalities between two or more languages: "Contrastive Linguistics is the systematic comparison of two or more languages, with the aim of describing their similarities and differences" (Johansson 2003, 32).

In the 1960s and 1970s, the focus in contrastive linguistics was on language typologies, translations and language teaching. Linguists tried to predict learner's errors by describing the contrasts between the native language (L1) and the target language (L2). One of the major works was James' (1980) *Contrastive Analysis*. Hawkins' (1986) book was a major contribution to the contrastive study of English and German. The prediction of learner's errors, however, was less successful than had been expected. Some of the predicted errors did not actually occur in learner's use of L2, whereas other errors occurred which had not been predicted. "[T]he lack of predictive power (of learners' mistakes) led to widespread disillusionment" (Schmied 2008). Interest in contrastive linguistics waned, until Stig Johansson and his colleagues built the English-Norwegian Parallel Corpus in the early 1990s (Johansson and Hofland 1994) and initiated a new era in contrastive linguistics. According to Aijmer and Altenberg (2013), this first parallel corpus "placed contrastive analysis on a sound empirical footing". Through the use of corpora and corpus-linguistic methods, linguists were able to compare language use on a much greater scale than had been previously possible.

The beginning of the 21st century has seen a growing interest in corpus-based contrastive studies, fuelled by the advances in corpus linguistics; see for example Granger, Lerot and Petch-Tyson (eds.) (2003), Johansson (2007), Xiao (ed.)

(2010), Marzo, Heylen and De Sutter (eds.) (2012), Aijmer and Altenberg (eds.) (2013). Marzo, Heylen and De Sutter (2012, 1) attest that

> [t]he field of contrastive linguistics has already witnessed a clear shift from […] corpus-illustrated work, using hand-picked corpus examples, to corpus-based analyses, characterised by a systematic analysis of corpus instances and empirical verification of theoretically grounded hypotheses.

Few corpus-based contrastive studies, however, have been conducted using the language pair of English and German. The team connected to Saarland University is certainly the most productive in this area of studies, see for example Čulo, Hansen-Schirra, Neumann and Maksymski (2011), Neumann (2014), Kunz and Steiner (2013).

Bußmann (2002) summarizes the main challenges that contrastive linguistics has to address: The choice of an adequate grammar model for the description of the languages or language areas, the decision about what can serve as a *tertium comparationis*, and finally the selection of criteria to judge the formal, pragmatic and communicative equivalence of sentences or utterances.

Tekin (2012) identifies three prerequisites for successful language comparison: First, in order to compare two languages, both must be described individually before comparing the descriptions. There are two ways to do this (Tekin 2012, 133): first, to describe and compare both languages in one step ("beschreibend-vergleichend"), and second, to first describe both languages and then compare them in a second step ("beschreibend und vergleichend"). He prefers to describe and compare in one step, because this method averts overly lengthy descriptions of the individual languages, which, as he says, is not the work for contrastive linguists. In my study, however, I have to choose the second method, to first describe and then compare, simply because there is not yet a full description of the German language in SFG terms. The non-availability of corpora which can be used for contrastive corpus studies is a complication also mentioned by Granger (2003, 22). She expresses her hopes that this is only a temporary problem, and the present study contributes one corpus to the ever growing number of corpora to be of good use in contrastive linguistics.

Second, Tekin (2012, 115) agrees with Bußmann on the need for an underlying theory for the description of languages. Having referred to SFG in the previous paragraph, this language theory will provide me with the required grammar model. Based on this theory, the descriptions of the English and the German language will become comparable. Tekin (2012, 132) cites Hjelmslev (1974), who said that it does not matter which theory you adhere to, as long as it is "without contradictions, comprehensive and as simple as possible". In my view, SFG fulfils these requirements for a grammar model and can thus serve well in the present contrastive study.

Third, Tekin also emphasizes the need for a tertium comparationis (t.c.) in contrastive studies of languages; this need is also expressed by Schmied (2008). Tekin's main concern about finding a valid t.c., a common ground for comparing languages, is the fact that regardless of which level of language is chosen as t.c., other languages may use other levels to express the same meaning. In my work, however, I presume that English and German are so closely related languages, with so many typological commonalities, that most of the time meaning is expressed by the same forms. This view is shared by König and Gast (2007) in their book *Understanding English-German Contrasts*. König and Gast (2007, 6) use both semantic criteria like temporal relations or co-reference, and formal criteria like case, function words, and word order. I am confident that the present study will show that the categories which are provided by SFG apply equally well to both English and German and will be a sufficient t.c. for a thorough contrastive study. Furthermore, we can address this concern about what is equivalent, as expressed by Johansson:

> One of the most serious problems of contrastive studies is the problem of equivalence. How do we know what to compare? What is expressed in one language by, for example, modal auxiliaries could be expressed in other languages in quite different ways. In this case a comparison of modal auxiliaries does not take us very far. (Johansson 2003, 34)

While this is generally true, carrying out a comparison of elements which have the same form, and annotating them throughout the corpus, as in the present study, we will become aware of other forms which may have similar functions.

This, in turn, will help us to compare better in the future. The use of a corpus provides empirical evidence "to achieve a higher degree of descriptive adequacy" (Granger 2003, 19). Furthermore, corpora will make native-speaker competence less important (Schmied 2008, 1143). Schmied (2008, 1155) concludes by saying that "many more detailed analyses need to be carried out before we can come to a more comprehensive understanding of the qualitative and quantitative contrasts between languages". The present study will contribute to a more comprehensive description and comparison of the language pair English-German.

2.3.2 First steps in combining CL and corpus linguistics

One of the first few corpus-based contrastive studies of the English and the German language is the one by Salkie (2008). In his article, Salkie reports on his investigation of modal auxiliaries in English and German. He develops the concept of a 'typological cluster' which originates in prototype theory (Salkie 2008, 77). He explains that a typological cluster is a set of criteria which can be used to study any language feature, and which can serve as tertium comparationis. Thus, the typological cluster can be used to test how many of the criteria are met by a language feature, thereby evaluating how unmarked or marked that feature is. A typological cluster is language-independent and therefore extremely useful for contrastive studies.

Salkie's focus lies on modal auxiliaries in English and German, a feature which is also of interest to me in the present study. In his view, the fact that a modal auxiliary belongs to a certain word class, and has certain morpho-syntactic criteria, tells us nothing about how that modal auxiliary is used. Thus, we must identify the centrality of each use of a modal auxiliary. One form may be used differently in each example sentence. Not all of these uses necessarily express modality. The unmarkedness of a modal auxiliary is not a question of frequency or marginality or stability over time, but of how many of the criteria apply. Salkie (2008, 89) provides us with the following four criteria to judge how central a modal auxiliary is:

A. They express possibility or necessity.
B. They are epistemic or deontic.
C. They are subjective, involving

I. commitment by the speaker,
II. primary pragmatic processes,
III. a sharp distinction between the modal expression and the propositional content.

D. They are located at one of the extremes of a modal scale.

He ascertains that in the English language, *must* and *may* are central because they meet all four criteria most of the time, *should* is in between, whereas most uses of *can* do not meet the criteria. In the German language, Salkie says, *müssen* is central, *können* is in between, and the majority of uses of *sollen* do not meet the criteria. Salkie (2008, 95) concludes by saying that

> [t]he typological cluster approach invites us, as a minimum, to re-examine what we mean by "modality". It offers a set of explicit criteria which can be revised, removed or augmented as research progresses. It can help to resolve disagreements between scholars who take a narrow view of modality and those who want a broader approach.

Salkie's study is a rather theoretical discussion, and even though he uses his own INTERSECT corpus (Salkie 2006), he only uses it to choose examples, thus, his work is corpus-illustrated, but not corpus-based (Salkie 2008, 92). In fact, the INTERSECT corpus is a parallel corpus with original texts and translations. For reasons which remain unclear, he opts for examples in English, without the translation into German, to illustrate the use of modal auxiliaries in English. For the study of modal auxiliaries in the German language, however, he makes use of the German original sentences and their translations into English. A contrastive study does not require translation corpora, and it is not clear why he uses the translation in one direction but not the other, without paying attention to the differences between original texts and translations.

2.4 Computer-mediated communication (CMC)

2.4.1 Computer-mediated communication – defining the term

The term *computer-mediated communication* (CMC) stands for any kind of communication that uses an electronic device for transmission, thus one which is neither oral communication nor written communication on paper. Originally only used for communication via devices such as computers, it is now also used to describe communication using mobile phones, smartphones and tablet computers. The term emerged in the 1990s and is still in use (Barton and Lee 2013), although it is being criticized for being too broad (Crystal 2011, 1). Crystal uses the term *Internet linguistics*, which is also the title of his book. In addition to *Internet linguistics*, scientists write about *Cyberspeak*, *Netspeak*, *electronically mediated communication* and *digitally mediated communication*. Fraas and Meier and Pentzold (2012) published their book, which is written in German, under the title of *Online-Kommunikation*; they found the term *computervermittelte Kommunikation* too unwieldy. I will use the term CMC throughout my work, as this was the term which was in broad usage when I began working on my project. Since academic research on language in the Internet is still in its infancy (Crystal 2011, 3), we may assume that other new terms will be created, and only time will show which of these terms survive.

In CMC studies, researchers distinguish between synchronous and asynchronous CMC. Asynchronous means that "users do not have to be online at the same time to communicate; the addressee of a message may both read and respond to it at a later time" (Beißwenger and Storrer 2008, 293). The newsgroups which provided me with texts for the EDNA corpus are asynchronous CMC.

2.4.2 State-of-the-art in CMC research

The first article reporting a study of CMC was "Computer mediated communication as a force in language change" by Naomi Baron, published in 1984 (Baron 1984). But Herring (2001, 613), as cited in Frehner (2008, 18), states that it was not until 1991 that linguists became interested in computer-mediated discourse (CMD), inspired by an article named "Interactive written discourse as an emergent register" by Kathleen Ferrara, Hans Brunner and Greg Whittemore (1991). The flames were fanned from that year on.

Barton and Lee (2013) identify three key directions in the study of Internet language: structural features (written / spoken language), social variation, and language ideologies, including a metalanguage to talk about Internet language. There have been many different theoretical foundations on which researchers have based their CMC studies, e.g. sociolinguistics, conversation analysis, with Herring's (2004) framework for analysing CMD, and pragmatics, with the handbook by Herring, Stein and Virtanen (2013).

One study which is related to my work is "Conversation Analysis and Community of Practice as Approaches to Studying Online Community" by Wyke Stommel (2008). Stommel uses a conversation analysis approach, based on Herring (2004), to show how in a forum on eating disorders writers create a common ground that helps them to feel like part of the group. The article is written in English, and all her examples from the eating disorder forums are in English. It turns out, however, that her corpus consists of texts written in German, which have been translated for the article. This makes one wonder what might have been 'lost in translation'.

Most studies in CMC research, however, have investigated the English language. It is only in more recent years that other languages which are used on the Internet have become the focus of linguistic studies, e.g. Danet and Herring (2007) and French, German and Japanese data in the 2008 issue of the journal *Language@Internet* (including Stommel 2008), later followed by studies on Jordanian chats and Dutch emails in 2012 (http://www.languageatinternet.org/articles/2012, 29.05.2014). This may not be surprising if we consider that almost all editors of the journal are based at European or US-American universities, with the exception of only one editor who is based in Hong Kong. Contrastive studies of two languages, to my knowledge, have not been carried out to any larger extent. Thus, my project can fill a gap and contribute to the description of CMC as well as to the discussion about methodologies used.

2.4.3 The challenges in CMC research

Even if Internet language has become a rich resource for linguists, research in CMC is not without challenges and obstacles. Crystal (2011, 10) names a few:

- The sheer amount of data, the ever-growing, enormous size of the corpus
- The diversity of languages on the Internet (English, Chinese, Spanish, German, …)

- The stylistic range: web sites, email, chat rooms, discussion forums, virtual worlds, blogs, instant messaging, mobile texting, wikis, tweets, social networking platforms, online dictionaries and encyclopaedias, multimedia sharing sites
- The speed of change, making it difficult to define the start and end of a variety
- In-/ accessibility, legal and commercial constraints, protection of privacy, ethical considerations
- Anonymity: Age, gender, class, ethnicity, non-/ native speaker status are hidden from others, but crucial for (socio-) linguistic studies. Crystal (2011, 13) elaborates: "But in a medium where a large number of participants hide their identity, or where we cannot trust the self-disclosed information about themselves which they place online, it is difficult to know how to interpret observed usage."

In addition to challenges provided by the data itself, Androutsopoulos and Beißwenger (2008), and Herring (2013) and Fraas et al. (2012), diagnose that reflection on methodological issues in CMC research is largely missing. These issues, they say, involve data collection, i.e. the size and representativeness of samples, data processing techniques, ethical issues, and the required amount of contextual information needed.

> Much research in the area has been based on small, ad-hoc data sets; there is a lack of standard guidelines for CMD corpus design and a lack of publicly-available CMD corpora. […] What is largely lacking, however, is critical reflection on the problems and challenges that arise when these research traditions are applied to the new settings and environments of CMD. (Androutsopoulos and Beißwenger 2008, 1)

2.4.4 CMC – written or spoken language?

Barton and Lee (2013) inform us that in the early years of research on CMC, the focus was to compare CMC to existing modes of communication, and to investigate whether CMC should be seen as speech or writing or a hybrid of both (Herring 1996, Baron 2003 in Barton and Lee 2013, 4). The earliest corpus-based study comparing spoken, written and computer-mediated communication was probably Yates (1996). Many more linguists have contributed to the discussion whether CMC is more of a spoken or a written variety of language, and some of these early studies were criticized for pretending that all CMC was rather similar (Barton and Lee 2013, 5). Crystal's point of view is the following:

> Whatever facts were established about, say, the differences between spoken and written vocabulary and grammar, these now have to be revisited, because the way we use language on the Internet is different in salient respects from the way we use it in traditional speech and writing. (Crystal 2011, 14)

Thus, I agree with him when he says that Internet language is a whole new variety and may call for a third category, in addition to spoken language and written language in the classical meaning:

> On the whole, Internet language is better seen as writing which has been pulled some way in the direction of speech rather than as speech which has been written down. However, expressing the question in terms of the traditional dichotomy is misleading. Internet language is identical to neither speech nor writing, but selectively and adaptively displays properties of both. It is more than an aggregate of spoken and written features. It does things that neither of the other mediums does. (Crystal 2011, 21)

The major changes that the Internet brought to the ways we communicate may justify adding a third category, which combines features of both spoken and written language:

> The language of the Internet cannot be identified with either spoken language or written language, even though it shares some features with both. The electronic medium constrains and facilitates human strategies of communication in unprecedented ways. Among the constraints are limited message size, message lag, and lack of simultaneous feedback. Among the facilitations are hypertext links, emoticons, and the opportunities provided by multiple conversations and multiply authored texts. (Crystal 2011, 32)

To conclude, the present study contributes to the discussion of whether CMC is spoken or written language (e.g. in chapter 4), but that is not its main aim.

2.4.5 CMC research and corpus studies

One empirical study on CMC is the work by Frehner (2008). She investigates a corpus of private email, text messages and MMS from Germany, Switzerland and the UK, written by students between 17 and 27 years of age and provided on a voluntary basis. Her aim is to characterize these new text types and to show that even though they are written, they gradually shift towards spontaneous spoken language. Frehner uses Koch and Oesterreicher's (1994) popular model, combining qualitative and quantitative methods. In the end, she concludes that

> [t]he phenomenon of Netspeak [a term coined by David Crystal, m.A.] has been variously analysed and many linguists have concluded that it is a hybrid register which makes use of both spoken as well as written language. (Frehner 2008, 26)

In addition, Frehner wants to know in what ways these new media types have changed the English language, and whether they lead to the deterioration of the English language. Why she needs the German texts and emails, and those from Switzerland, is not entirely clear. They serve as a reference, and to back up the English data on those occasions where it is insufficient (MMS). Furthermore, she compares SMS to telegrams, with the result that the two have little in common. She also does a study of Anglicisms in the German and Swiss texts and mails. Her conclusion is that Anglicisms are especially common in text messages, particularly within salutation formulas. In my opinion, Frehner to some extent attempts to compensate for a lack of resources by grasping at any available data – her methodology is good, though, and her use of a corpus for her empirical study is meritorious.

2.4.6 Available CMC corpora

Beißwenger and Storrer (2008, 294) state that CMC research is a relatively new field in linguistics and that CMC is generally not included in large balanced corpora, neither in English ones like the British National Corpus (Consortium 2007) nor in any German representative corpus, e.g. the COSMAS II (IDS 2014).

> Thus, at the present time, the assortment of large accessible corpora that were exclusively designed for analysing CMC phenomena is rather unsatisfactory. Therefore, for empirical studies, corpora often have to be individually acquired from the Internet or obtained from users of CMC facilities. (Beißwenger and Storrer 2008, 295)

As early as 1995, Feldweg, Kibiger and Thielen (Feldweg, Kibiger, and Thielen) created the *Korpus deutschsprachiger Newsgroups,* with raw data for general use. Another work is Pankow (2003), who uses a contrastive German-Swedish IRC corpus. There are, however, some German CMC corpora, e.g. the Düsseldorf CMC Corpus (Zitzen 2004) and the Dortmund Chat Corpus (www.chatkorpus.uni-dortmund.de). A website provided by Beißwenger and Storrer (2008, 293) lists available CMC corpora at www.cmc-corpora.de. Despite these attempts, the general lack of available corpora of CMC – especially a parallel one containing English and German texts – made the building of the EDNA corpus inevitable. That has taken a good deal of time, and Crystal (2011, 11) warns that "[l]inguistic studies of the Internet always run the risk of being out of date as soon as they are written", let alone published. My data is ten years old by the time of writing (2014), but apparently, there are still not that many corpus-based studies of CMC.

Other scholars see the same risk of CMC studies quickly being out of date, which poses a dilemma. On the one hand, the material in the EDNA corpus is outdated. On the other hand, however, no one has done a similar study with a similar corpus and theoretical approach, and in this respect, my work is still highly relevant.

3 Methodology

The data basis for my contrastive research is provided by a corpus of newsgroup texts from the Internet, i.e. computer-mediated communication (CMC), annotated for SFG features. The corpus will be introduced in 3.1, and I will briefly talk about ethical consideration in 3.2. This will be followed by a description of the population in the corpus using Halliday's (1989b) parameters of register, of the criteria for selecting texts, as well as of the size of the samples and the corpus. The procedure of choice in this work is computer-assisted manual annotation (CAMA) with the use of the UAM corpus tool (O'Donnell 2008). Guidelines to ensure inter- and intra-annotator agreement have been developed, these will be described in 3.5. Subchapter 3.6 outlines the method for quantitative analysis. The linguistic features under analysis will be described in detail in the respective chapters later on.

3.1 The BTC and the EDNA Corpus

BTC stands for *Bremen Translation Corpus*. The BTC was designed and compiled at the University of Bremen, Germany, by Kerstin Fischer, Anatol Stefanowitsch and Anke Schulz in 2003-4. The compilation of newsgroup texts for the BTC was agreed on because, in 2004, they were a new register with interesting features of both spoken and written language and promised to be a fertile ground for answering a range of research questions we were interested in at that time. Furthermore, newsgroup texts were easy to collect and easy to process due to the fact that they already came in a digital format, thereby providing first class machine-readable data. We decided to keep it as easy as possible; Baker (2010, 109) also confirms that "corpora can be time-consuming, expensive and difficult to build".

The corpus consists of a comparable and a parallel part. In the comparable part of the corpus, there are about 10,000 words of English texts and another 10,000 words of German texts, taken from the same register. This small, comparable part of the BTC is called the EDNA corpus (*Englische and Deutsche Newsgroup-Texte - Annotiertes Korpus*). The EDNA corpus with its original texts in English and German is the basis for the much larger parallel corpus, the BTC. The BTC is parallel in two ways: first, it contains the 30 individual original texts in English and German and their translations into the other language. Second, there is not only one translation: Each of these individual texts, about 250 words long, has been translated by five different non-professional translators who were native speakers of the target language, German or English respectively. The translated part of the corpus amounts to approximately 80,000 words, thus giving the entire corpus a size of 100,000 words. For my research I use the smaller corpus of original texts, the EDNA corpus.

Biber (1990) has shown that 1,000-word text samples are sufficient for most linguistic analyses. He used 1,000-word samples to investigate ten linguistic features, from very common ones like nouns and prepositions to rare ones, e.g. WH- relative clauses and conditional subordination. "Overall, the results [...] indicate a high level of stability for these linguistic feature counts across 1,000-word sub-samples of texts. This stability holds generally across linguistic features and across text categories" (Biber 1990, 261). This was the starting point for collecting the samples in EDNA. Contributions to newsgroups, however, are usually quite short and it proved difficult to find 1,000-word samples. Therefore, what I did was to collect 250-word samples, so that texts would be equally long and have a beginning and an end. The English and the German subcorpora comprise approximately 10,000 words each. Based on Biber's (1990) work, I assume that the linguistic features I plan to study will be represented reliably in 10 times 1,000 words. Furthermore, "a corpus has to be of a manageable size for manual analysis" (McEnery and Wilson 2001, 79).

"Finally, there may be more pragmatic reasons for building a corpus of a particular size – depending on what texts are available, how much money or time we have to devote to a project or whether we can obtain permission from copyright holders [...]" (Baker 2010, 96).

3.2 Ethical considerations and copyright

When building the Bremen Translation Corpus, it was assumed that written contributions to newsgroups are published on the Internet and therefore authors must expect to be read by a potentially infinite number of people. Downloading texts, however, seems to be another matter. In opposition to earlier claims that most of the material published on the Internet is not copyright protected (McEnery and Xiao and Tono 2006, 78), more recently voices are saying that "copyright laws apply to documents available on the web exactly as they do to print documents" (McEnery and Hardie 2012, 58). To avoid legal action, the BTC and EDNA are only used for non-profit-making academic research and will not be made available to the general public, in agreement with McEnery and Hardie's (2012, 59) suggestion: "The third approach is to collect data without any regard to seek permission, and not to distribute it, but instead to make it available to other researchers through a tool that does not allow copyright to be breached". There is no tool that allows the use of EDNA without displaying the full texts, thus the only solution seems to be to not distribute the corpus. McEnery and Hardie (2012, 61) suggest, and I support that view, that

> no test case has been taken to court that we know of. There are several reasons for this: […], corpus linguistics is fairly obscure in the grand scheme of things and most text producers probably don't even know if their text ends up in a corpus that is searchable online; and, ultimately, corpus linguistics are unlikely to have enough money to be worth suing.

Beißwenger and Storrer (2008, 300) assume that it is unrealistic and not practicable to receive written consent prior to collecting the data, and that it may change the data if the writer is aware of being observed. Receiving written consent after collection is just as impracticable, since writers are often registered under pseudonyms. Still, even data from writers using a pseudonym should not be used without taking measures to protect the anonymity of the writer. Such measures include anonymization of the participants' real names, nicknames or pseudonyms, omission of details about the location of the discussion groups, or only processing data with statistical software.

To guarantee anonymity of the writers as much as possible, all names have been changed in the corpus. Names in the EDNA corpus only indicate the same gender as the original name. Writers only gave first names or nicknames. All these of course may be false identities. All email-addresses which were given in the texts have been anonymized. Details about the age, where made explicit, have not been changed. It is the custom in some of the eating disorder platforms that any mention of a number (age, weight, height and others) is prohibited, in order to discourage individuals from entering into a weight-loss competition. As a result, in some of the texts in the corpus the [*] that replaced a forbidden number has been replaced by the number 99, so that part-of-speech taggers would correctly assign the number-tag. All emoticons have been deleted, as these were of little interest to us and would only mess up the PoS-tagging.

3.3 Definition of the population of samples in EDNA

The population of samples in the EDNA corpus of newsgroup texts can be described using Halliday and Hasan's (1989b) parameters of register description. Halliday and Hasan relate any piece of discourse, written or spoken, to the situation in which it is used to communicate. They describe three "abstract components of the context of situation" (Halliday and Hasan 1989b, 29). These three components are called field, tenor and mode. The field describes what is going on in a situation, the tenor indicates the relation between the people involved in the situation, and finally the mode shows how the discourse is realized. In the following paragraphs the field, tenor and mode of the newsgroup texts collected in the EDNA corpus are outlined.

Field: The experiential domain in half of the texts is that of eating disorders, the second half of the texts deals with relationship problems. These two domains were chosen because one of the goals of building the corpus was to investigate how negative experience is expressed. These two domains lend themselves to such an investigation because the authors reveal their problems and experiences. The social activity from which the discourse emerges can probably best be described as 'self-therapy by public display of experiences, thereby inviting feedback and support from others'.

Tenor: Writers in these newsgroups have rare to regular contact, use informal language, are in an equal, non-hierarchical power relationship, there is minimal social distance. Contributions are made to address 'fellow sufferers', high affective involvement is displayed: Writers divulge secrets and display weaknesses, possibly encouraged by the fact that the writer does not have to sit face to face with the addressee, one way of 'saving face'.

Mode: The medium is written, although digitally, informal, written-as-if-spoken. The spacial distance is defined by the absence of addressee (or addressor, respectively), there is no immediate feedback, no visual or aural contact. The language in this register is an indispensable constituent, there is only language, but no action to accompany it, hence no facial expressions or body language or movement to be read.

3.4 Criteria for selecting samples

The samples from newsgroups, sometimes referred to as discussion groups, were collected at random from various different newsgroups in 2003-4. Some criteria were established for the selection of samples:

The topic; text samples had to be about either eating disorders or relationship problems.

The language; English or German. In the selection process, it was not possible to control whether the writers were native speakers of English or German, but only texts that seemed to display native speaker competence were included. Neither was it possible to identify at that stage what variety of English or German writers used or which part of the world they came from. In the thirty English texts, two writers stated their location as U.S.A. (Iowa and Minnesota); one lived in New South Wales, Australia. In the German texts, only one person referred to his location, the Ruhrgebiet, Germany. In this kind of discourse, the location of a person is not considered to be a relevant factor.

The length; around 250 words for a text that is complete in itself, i.e. that has a greeting at the beginning, then goes on to talk about the problem at hand, then ends with a good-bye. The contributions had to be the beginning of a new discussion. A person introduces herself and her problem. Those contributions in which people react to a previous one were not included. In a few texts, people referred to something that was written at an earlier stage, made a reference to someone else's contribution, but only briefly, not giving advice at length.

The gender; female or male. The writers either stated these in their posts, or it was revealed by the personal pronouns or words like *wife* or *husband* with which the writers referred to their partners in the relationship problem texts. It was assumed, without any evidence, that the writers were in a heterosexual relationship. In the eating disorder group, there are two texts were the gender is unclear, but assumed to be female. (Keep in mind that any of these identities may well be false ones.) It was our goal to have a balance of female and male writers on both subjects.

The age of the writers; there was no way of knowing the age of the writers unless it was stated in the contribution, which was done in 15 out of 30 English texts and in 9 out of 30 German texts. The ages that were explicitly given range from 20 to 47 in the English texts, and from 19 to 43 in the German texts.

The social and educational background; no information about the educational or social background of the writers was available. It seems that these factors are considered irrelevant in the discussion of eating disorders or relationship problems. People do not necessarily reveal all their offline identity features like age, sex, location when writing online. This is not necessarily a sign that fraud is intended, but for creativity and playfulness, trying on a different identity, or safety issues (Barton and Lee 2013, 69). Fraas et al. (2012) cite Thibaut and Kelly (1959), who coined the term 'stranger-on-the-train phenomenon'. This term describes the phenomenon that we find it easier to give more – and more intimate – information about ourselves to strangers than we would give to friends and acquaintances. As soon as we assume that we will not meet a person again, and that she is not part of our social circles, we believe that our self-disclosure will not have any consequences for ourselves. We need not fear a loss of face. This is exactly what happens in the newsgroups on eating disorders and relationship problems.

3.5 Annotation of the corpus

Collecting the texts for EDNA was only the first step. Plain texts in a corpus are perfectly sufficient for many corpus queries, e.g. concordances or word lists. However, a corpus can be enriched with additional information; this is known as corpus annotation. There are many different ways to add information to a corpus, the easiest is probably by adding part-of-speech tags, and even that is not without challenges. Let us start by considering the advantages and disadvantages of corpus annotation. McEnery, Xiao and Tono (2006, 29) make the point that corpus annotation must be considered an interpretative act; annotation is the result of a human's understanding of a text. They name four advantages of corpus annotation (McEnery, Xiao, and Tono 2006, 30):

1. Information can be extracted more easily.
2. Annotation is reusable information.
3. The annotation is multifunctional and can be used for answering different research questions.
4. The annotations are a record of the interpretation and are available for scrutiny, criticizm, and, most importantly, reproducibility.

The authors also name four criticizms that have been brought forward, but do not fail to invalidate those (McEnery, Xiao, and Tono 2006, 31):

1. The annotations clutter the raw texts. Most corpus tools are able to display raw text only, though.
2. An annotation imposes an interpretation of the data on the next user. However, any following user can agree or disagree with an annotation, the advantage of making the interpretation recoverable and visible compensates for the risk of patronizing corpus users.
3. Third, an annotated corpus cannot so easily be expanded and updated. Most corpora, however, are static and once the data is collected and annotated, there is no need for expansion.
4. The final criticism for annotating corpora is concerned with the accuracy and consistency of annotation. Neither automatic nor computer-assisted nor manual annotation is 100% accurate. But then, no linguistic analysis has ever been free of errors. Using machine-readable data and making the annotation machine-readable results in a better chance of finding and correcting any errors in the analysis.

McEnery, Xiao and Tono (2006, 33) give a final word of warning about manual annotation that I can only agree with: “As manual annotation is expensive and time-consuming, it is typically only feasible for small corpora.”

The annotation of the data in the EDNA corpus was done by way of computer-assisted manual annotation (CAMA). The UAM corpus tool, developed by Michael O’Donnell from the Universidad Autonoma de Madrid (O'Donnell 2008), was chosen because it had proven its suitability in the pilot study. The main advantage of the UAM CT is the possibility for researchers to build their own annotation schemes depending on their needs. This was necessary because there was no software that could automatically identify SFG features when the annotation of EDNA was carried out (2006-9).

Following the development of suitable annotation schemes, the entire 20,000 words in EDNA were annotated manually. A team of two annotators (Anke Schulz and Tatsiana Markovic) worked on the task. Note that each text has only been annotated by one of the annotators. The annotated corpus was then revised once to improve consistency. There are three annotated layers; the first is for the theme-rheme annotation (Halliday 1994), the second for the annotation of modality and negation (Halliday 1994), the third is annotated for participant roles and process types according to the Cardiff Grammar (Fawcett forthcoming). The coding schemes which were constructed for the annotation with UAM CT will be shown and explained in the respective chapters on the three SFG metafunctions.

3.5.1 Consistency of manual corpus annotation

The quality of the corpus annotation is an important element in any empirical corpus study. The key term is consistency. “*Consistency* here means that the same linguistic phenomena are annotated in the same way, […]” (Zinsmeister et al. 2008, 764). Zinsmeister et al. (2008, 764) name four techniques that can be used to make corpus annotation more consistent:

I. Annotation guidelines

II. Semi-automatic annotation

III. Manual or automatic consistency checking

IV. Multiple annotation by different annotators

To make sure that the manual annotation of the EDNA corpus would be sufficiently consistent, annotation guidelines were written and the annotations were manually checked for consistency. In an annotation guideline, the phenomena under investigation must be clearly defined and easily identifiable. This, in turn, will promote a high inter- and intra-annotator agreement. The inter-annotator agreement, also known as inter-coder reliability, refers to "the degree to which the different annotators agree on a single annotation for a specific sentence or paragraph" (Zinsmeister et al. 2008, 766). The same is true for the intra-annotator agreement, i.e. not two people agreeing on one annotation, but one annotator agreeing with her-/himself over a longer period of time. High levels of inter- and intra-annotator agreement are a sign of good quality of the annotation and repeatability of the study:

> If the analysis is manual, and if the annotations were undertaken by a linguist or linguists working to an agreed set of guidelines for applying the annotation, then we can be much more confident in the consistency of the analysis, […] (McEnery and Hardie 2012, 32).

Note that calculations for inter-annotator agreement, usually done by measuring the *kappa statistic* (Cohen 1960 in Zinsmeister et al. 2008, 766), were not carried out, as the aim of working with two annotators was to speed up the annotation process, not to test the validity of the guidelines. The quality of the annotation was improved by a second round of checking by the author, but it was not measured.

Writing the guidelines for the annotation of the English texts was easy compared with writing the guidelines for the annotation of the German texts. The guidelines for English basically summarize Halliday's (1994) *An Introduction to Functional Grammar* and Fawcett's (forthcoming) *The Functional Semantics Handbook: Analyzing English at the level of meaning.* There are now three guidelines for English, one for the annotation of theme-rheme structure (Halliday 1994), one for modality and negation (Halliday 1994) and another one for process types (Fawcett forthcoming).

Writing the three respective guidelines for German was more challenging because at the time of writing the guidelines in 2007 no Systemic Functional Grammar of the German language had been published. A draft written by Uwe Helm Petersen (Petersen forthcoming) on the Systemic Functional Grammar of German was used. Apart from that, the guidelines for German had to be developed from scratch using the guidelines for English as a starting point and adjusting the annotation schemes to adequately represent the German language system. The guidelines for manual annotation of SFG features in the German language describe how to identify the constituents and label them appropriately, just as the English guidelines do. The guidelines for English and German can be found in the appendix.

With the help of the annotation guidelines and a manual consistency check, we tried to make the results reproducible. They are, however, still not free from errors and mistakes, but, to use McEnery and Hardie's (2012, 32) words:

> In fact, it is inevitable that, from time to time, manual annotations will be inconsistent to some degree. Quite apart from considerations of human error, this is due to a property of all linguistic analyses, namely that an analysis typically represents one choice among a variety of plausible analyses.

3.6 Quantitative analysis with the chi-squared test

Following the manual annotation with the help of the UAM corpus tool, EDNA was ready for quantitative and qualitative analysis of all systems representing the experiential, interpersonal and textual metafunctions. In addition to the raw frequencies, relative frequencies were calculated, these are expressed as percentages. We also want to know, however, whether the results indicate differences in the two subcorpora, or whether any deviation in number is simple due to chance. The chi-squared test (χ^2 test) was chosen as a statistical test to answer that question. It is a simple test that takes observed and expected frequencies into account: "The greater the difference between the observed values and the expected values, the less likely it is that any difference is due to chance" (Baker, Hardie, and McEnery 2006, 31).

Hence, the tables in the subsequent chapters on quantitative analysis give the raw frequencies (F) as well as the relative numbers as per cent. They indicate which of the two subcorpora in EDNA makes predominant use of the feature, indicated by the word 'overuse', for lack of a better term. One of the columns displays the chi-squared value, and the last column in all tables with a chi-squared value indicates whether the result is not significant (-), significant at the 0.05 level (+), the 0.01 level (++) or the 0.001 level (+++) (McEnery and Wilson 2001, 84-85), (Oakes 1998, 24-25).

The different levels of χ^2 indicate how sure we can be that the result actually describes a meaningful deviation from the expected frequencies:

0.05 = 99.5% certainty that the result is not due to chance or error (+).

0.01 = 99.9% certainty that the result is not due to chance or error (++).

0.001 = 99.99% certainty that the result is not due to chance or error (+++).

This chapter described and explained the data that is used for the study as well as the population of samples in the EDNA corpus, how the samples were collected and annotated, and how the results were processed. The next chapter begins the description of computer-mediated communication, i.e. the newsgroup texts in EDNA, with some basic numbers.

4 Basic corpus statistics

In this chapter we look at a few basic numbers which help to describe the EDNA corpus. A description of the EDNA corpus is done by way of the type-token-ratio, words per sentence and per clause, and by lexical density and grammatical intricacy (Halliday 1989a).

4.1 Type-token ratio

The type-token ratio is a basic parameter to describe corpora, and is included here for the sake of a comprehensive description of EDNA. The English and the German parts of EDNA were lemmatised using the TreeTagger software (Schmid 2013), and the number of tokens (all words) and types (different words) were counted. This allows the calculation of the type-token ratio (McEnery and Wilson 2001, 82) by dividing the number of types by the number of tokens and multiplying the result by 100 to receive a percentage:

(Type / Token) * 100 = TTR.

The calculation does not include cardinal and ordinal numbers or punctuation. The result is shown in table 4.1 below. Note that TTR is sensitive to corpus size. The larger the corpus, the more words keep repeating themselves, especially functional words (Baker, Hardie, and McEnery 2006, 162). EDNA is a small corpus, and both subcorpora have a comparable size, thus we can assume that the TTR are comparable.

Corpus	**Types**	**Tokens**	**TTR**
EN	1,797	10,360	17.34
GN	2,057	10,425	19.73

Table 4.1 Type-token ratio in EDNA

The German newsgroup texts display a slightly higher TTR, which means that there is greater lexical variation in the German texts than in the English newsgroup texts. We cannot, however, tell whether the difference is statistically significant, since the TTR is already a relative number.

4.2 Words, clauses and sentences

Apart from the type-token ratio, it might be interesting to know the total number of words, clauses, and sentences; these are shown in table 4.2. Clauses are defined as a stretch of text consisting of a finite verbal group and the groups and words that depend on it. Sentences may consist of one clause (clause simplex) or more than one clause (clause complex). Sentences were counted by counting sentence final punctuation (.!?).

Feature	**EN**	**GN**
Words	10,360	10,425
Sentences	666	779
Clauses (Rhemes)	1,578	1,499

Table 4.2 Sample size of the corpus

The German part of EDNA has a greater number of sentences, but fewer clauses than the English part. The German writers seem to prefer writing simple rather than complex sentences. This assumption is verified by a calculation of the word-per-sentence, word-per-clause and clause-per-sentence ratios, as shown in table 4.3 below.

Feature	**EN**	**GN**
Words per sentence	15.56	13.38
Clause per sentence	2.37	1.92
Words per clause	6.57	6.95

Table 4.3 Some ratios

The word-per-sentence ratio is in fact smaller in the German part of EDNA; this is due to the smaller number of clauses per sentence. If we consider, however, that a sentence is often no more than a clause simplex, i.e. one clause is one sentence, the German clauses in the newsgroup texts are actually slightly longer than the English ones. The English newsgroup texts are composed of clause complexes more often than the German ones.

4.3 Grammatical intricacy

What can the number of words per clause or the number of clauses per sentence tell us about newsgroup texts? Are newsgroup texts more like spoken or written language? Halliday (1989a, 87) says the following about the relationship between lexical density and grammatical intricacy:

> The complexity of the written language is static and dense. That of the spoken language is dynamic and intricate. Grammatical intricacy takes the place of lexical density. The highly information-packed, lexically dense passages of writing often tend to be extremely simple in their grammatical structure, as far as the organization of the sentence (clause complex) is concerned. [...] The complexity of the written language is its density of substance, solid like that of a diamond formed under pressure. By contrast, the complexity of spoken language is its intricacy of movement, liquid like that of a rapidly running river.

Remember that a word and a lexical item are not necessarily the same stretch of language. A word here is understood as a string of letters, preceded and followed by a blank space. A lexical item is a clause constituent of variable length, it may consist of more than one word, e.g. a phrasal verb like *put down* or an adverb like *at all*. In the calculation of grammatical intricacy and lexical density, words rather than lexical items were counted in order to simplify the procedure.

As shown in table 4.3 , grammatical intricacy in EDNA is low, with only 2.37 clauses per sentence in the English texts, and 1.92 clauses per sentence in the German texts. The sentence structure is rather simple, not intricate, not "liquid like that of a rapidly running river" (Halliday 1989a, 87). What about the lexical density, though? Are newsgroup texts lexically dense like other written discourse?

4.4 Lexical density

If the newsgroup texts lack grammatical intricacy, are they lexically dense like other written discourse? Lexical density is the relation between lexical and functional words per clause. The group of lexical words includes all adverbs, adjectives, nouns and verbs. All other words go into the functional word group, including cardinal and ordinal numbers as well as interjections. The PoS-tagged EDNA was thus divided into lexical and functional words. The total number for each word class as well as the percentages can be seen in table 4.4 below. Differences in the total number of the whole corpus as compared to total numbers in the TTR calculation are due to the fact that in the calculations of lexical density, the cardinal and ordinal numbers were included, whereas they were excluded in the TTR calculations. The deviation in numbers is smaller than 2% in both corpora, which is acceptable for the present study.

Word class	**EN**		**GN**	
Lexical	5,344	51%	5,514	54%
Functional	5,146	49%	4,707	46%
Total	10,490	100%	10,221	100%

Table 4.4 Lexical density in EDNA

The lexical density of the newsgroup texts can be compared to the numbers given in Halliday (1989a). Halliday (1989a, 80) states the following, referring to lexical words per clause:

> [A] typical average lexical density for spoken English is between 1.5 and 2, whereas the figure for written English settles down somewhere between 3 and 6, depending on the level of formality in the writing.

In EN, the lexical density per clause is 3.39, in GN it is 3.68. The texts in EDNA are thus more closely related to written discourse, but located more towards less formal language in the continuum.

Another source, the *Longman Grammar of Spoken and Written English* (LGSWE) (Biber et al. 1999, 62), does not give percentages, except for two small text samples of less than 100 words, but simply states that "[c]onversation has by far the lowest lexical density. News has the highest lexical density". It is difficult to say whether the percentages of lexical and functional words in EDNA point towards high or low lexical density. There is, however, a statement in the LGSWE that is more useful: "The proportion of the lexical word classes varies with register: In conversation, nouns and verbs are about equally frequent. In news reportage and academic prose, there are three to four nouns per lexical verb" (Biber et al. 1999, 62).

Table 4.5 shows that the ratio of noun-to-verb in the English texts is roughly 2:3. Clearly, the newsgroup texts are more similar to conversation in the LGSWE approach.

Word class	**EN**	
Nouns	1,453	14%
Lexical verbs	2,329	22%
Other word class	6,708	64%
Total corpus size	10,490	100%

Table 4.5 The noun-to-verb ratio in the English newsgroup texts

The numbers in table 4.5 above, however, are distorted. The newsgroup texts have been PoS-tagged with the BNC C5 tag set. The number in table 4.5 includes all words tagged as lexical verbs, excluding only the modal auxiliaries. This is problematic, because some of the verbs, namely forms of *be*, *have* and *do*, can be used either as lexical verbs, or as primary auxiliaries, and the tags do not reveal how they were used. Here is an example:

(1) I'm 30, he's 52. He's been married twice.

I_PNP 'm_VBB 30_CRD ,_, he_PNP 's_VBZ 52_CRD ._.

He_PNP 's_VHZ been_VBN married_VVN twice_AV0

The first two clauses have only one lexical verb (a form of *to be*), the third clause has a verbal group comprising three verbs; the forms of *to have* and *to be* function as primary auxiliary, *married* is the lexical verb. There is no indication of how the authors of the LGSWE dealt with the problem.

Another statement in the LGSWE is the following: "A high ratio of nouns to verbs corresponds to longer clauses and more complex phrases embedded in clauses" (Biber et al. 1999, 66). In the English newsgroup texts, the noun-to-verb ratio is low, thereby suggesting that the corpus consists of comparatively short clauses, with less complex phrases embedded in clauses. This again suggests the proximity of the newsgroup texts to conversation.

Let us look at the noun-to-verb ratio in the German newsgroup texts in table 4.6. The tag set used for German, i.e. the STTS tag set, does distinguish between primary auxiliaries, lexical verbs and modal verbs (not included in the calculation here). Therefore, the numbers for the German noun-to-verb ratio were easier to calculate.

Word class	GN	
Nouns	1,352	13%
Lexical verbs	1,223	12%
Auxiliary verbs	622	6%
Other word class	7,024	69%
Total size of corpus	10,221	100%

Table 4.6 The noun-to-verb ratio in the German newsgroup texts

The noun-to-verb ratio for the German newsgroup texts is almost exactly 1:1. If we assume that registers in German behave similar to their English counterparts, the newsgroup texts are more like spoken than written discourse. The low noun-to-verb ratio implies that there are rather short clauses without complex embedded phrases. Comparing the English and German corpus, they have 14% and 13% nouns respectively, the English texts have slightly more verbs (22%) while the German texts have only 18% verbs (auxiliaries plus lexical verbs, similar to the English counting). We can conclude that in the German texts, there are slightly more nouns per clause than there are in the English texts.

Feature	Spoken discourse	Written discourse
Grammatical intricacy (Halliday 1989a)		X
Noun-to-verb ratio (Biber et al. 1999)	X	
Lexical items per clause (Halliday 1989a)		X

Figure 4.1 Categorisation of newsgroup texts

Figure 4.1 lists the three features investigated above and suggests that we cannot conclude once and for all whether the newsgroup texts are more like conversation / spoken discourse or more like written discourse. The newsgroup texts seem to be a register that incorporates features from both.

5 Modality and negation – the interpersonal metafunction

In this chapter, the focus is on the first of the three metafunctions, the interpersonal metafunction, represented by the systems of modality and negation.

Note that there is in fact a third grammatical system, that of mood, which represents the interpersonal metafunction. Halliday (1994, 68) explains:

> […], the clause is also organized as an interactive event involving the speaker, or writer, and audience. […] In the act of speaking, the speaker adopts for himself a particular speech role, and in so doing assigns to the listener a complementary role which he wishes him to adopt in turn.

There are three basic speech roles that a speaker can adopt; giving information with a declarative sentence, asking for information with a question, or asking for a service by issuing a command. Even though Halliday (1994, 68) calls the mood system the principal grammatical system in the interpersonal metafunction, it is not included in the present study because after conducting a pilot study, it was found to show little divergence between the English and German texts in EDNA. In order to save time, the system of mood is therefore not investigated in the course of this research project.

Hence, the first part of the chapter is concerned with modality and the second with negation. To begin, it is necessary to explain the terms used when talking about modality. Depraetere and Reed (2006, 269) give a good definition:

> The term 'modality' is a cover term for a range of semantic notions such as ability, possibility, hypotheticality, obligation, and imperative meaning. […] modal meaning crucially involves the notions of necessity and possibility […], or rather, involves a speaker's judgement that a proposition is possibly or necessarily true or that the actualization of a situation is necessary or possible.

According to Halliday and Matthiessen, modality adds subjectivity to a discourse; speakers give an evaluation of a situation, thus making a statement more personal.

> Modality is a rich resource for speakers to intrude their own views into the discourse: their assessments of what is likely or typical, their judgement of the rights and wrongs of the situation and of where other people stand in this regard. (Halliday and Matthiessen 1999, 526)

The following subchapter explains the different types of modality and the different ways in which modality is realized in the English and German language.

5.1 Modality types – theoretical background

Two main types of modality are generally distinguished: epistemic and root modality. Quirk et al. (1985, 219) call these extrinsic (human judgement of what is or is not likely to happen) and intrinsic (some kind of human control over events). Halliday (1994, 88) calls them modalization (probability and usuality) and modulation (obligation and inclination). I find the terms modalization and modulation confusing, therefore I will use the terms of traditional grammar, epistemic and root modality.

5.1.1 Epistemic modality

"Epistemic modality reflects the speaker's judgement of the likelihood that the proposition underlying the utterance is true" (Depraetere and Reed 2006, 274).

Halliday's (1994) probability and usuality (modalization) are included in epistemic modality, see examples 2 and 3. Furthermore, prediction and hypothetical clauses are annotated as epistemic modality, see example 4.

(2) Probability: *I sure as hell don't want to be fat too.*

(3) Usuality: *I weigh and measure myself several times a day.*

(4) Hypothetical: *But if she is cheating, I will leave, no second chances.*

5.1.2 Root modality

Root modality implies that some sort of authority (humans, social norms, general circumstances) controls the situation, and is basically the same as modulation in Halliday's (1994) terms. Root modality in my annotation scheme has three subcategories: Obligation and permission, including both deontic and non-deontic modality, inclination (volition), and ability. Depraetere and Reed (2006, 274) give the following definition of deontic and non-deontic modality:

> Deontic modality also implies an authority, or 'deontic source' – which may be a person, a set of rules, or something as vague as a social norm – responsible for imposing the necessity (obligation) or granting the possibility (permission). […] Non-deontic root possibility […] and non-deontic root necessity […] concern possibility and necessity that arise, not via a particular authority but due to circumstances in general.

Example 5 shows a clause from EDNA which expresses obligation.

(5) *I had to leave my six young kids.*

The second type of root modality, inclination, means that a person is willing and intends to do something; the person herself is the source of the impetus, see example 6.

(6) *Other times I just need to be alone.*

Finally, the third type of root modality, ability, expresses that there are enabling or disabling factors internal to the speaker, such as physical and mental abilities or skills. Ability is included in root modality because it implies that the speaker has control over the situation. See example 7 below for illustration of ability.

(7) *Can you believe I was actually scared of eating it?*

5.2 Types of realization of modality – theoretical background

In English and in German, there are several ways to express epistemic or root modality: Modal auxiliaries, modal adjuncts and grammatical metaphor. In addition to those, German also uses modal particles and the subjunctive verb form to express modality. Let us look at the different types of realizing modality. A more detailed description of the realization types can be found in the annotation guidelines for modality in the appendix.

5.2.1 Modal auxiliaries in English

In English, nine central modal auxiliary verbs are used to express modality: *can, could, may, might, must, shall, should, will, would* (Biber et al. 1999, 483). In addition to these, there are marginal auxiliary verbs, e.g. *need to, ought to, dare to, used to* and a number of fixed idiomatic phrases with functions similar to those of modals: *(had) better, have to, (have) got to, be supposed to, be going to* (Biber et al. 1999, 484).

In the annotation of the BTC, the marginal auxiliaries and the fixed idiomatic phrases were taken into account in addition to the nine central modal auxiliaries dealt with in Halliday (1994).

5.2.2 Modal auxiliaries in German

In German, there are six central modal auxiliaries, i.e. *dürfen, können, mögen, müssen, sollen, wollen.* Furthermore, there are three marginal modal auxiliaries, i.e. *brauchen, haben* and *sein* + *zu*-Infinitiv (Götze and Hess-Lüttich 1999, 64). Fixed idiomatic expressions were not taken into consideration in the annotation of the BTC because I found no comprehensive list to refer to.

In analogy to the English verb *will*, which has been annotated as modal verb, not as future tense marker, the German verb *werden* + infinitive has been annotated as modal verb, not future tense marker. In this view I agree with Vater (1975, 94), who says that the construction *werden* + infinitive does not differentiate tense and does not behave differently from other modal verbs. He concludes that *werden* + infinitive is not a tense marker, but a modal auxiliary.

Both the indicative and the subjunctive forms of the modal auxiliaries have been annotated simply as modal auxiliary. In my annotations only lexical verbs were annotated for subjunctive (see below), because otherwise I would have had to assign two features from the same set of options to one word, which would have made the annotation scheme more complicated without having a positive impact on the results.

5.2.3 Modal adjuncts in English

Modal adjuncts are typically realized by an adverbial group or prepositional phrase. Halliday (1994, 82) gives an overview of modal adjuncts of polarity, modality, temporality, and mood:

Adjuncts of polarity and modality:

(a) Polarity:	*not, yes, no, so*
(b) Probability:	*probably, possibly, certainly, perhaps, maybe*
(c) Usuality:	*usually, sometimes, always, never, seldom, rarely*
(d) Readiness:	*willingly, readily, gladly, certainly, easily*
(e) Obligation:	*definitely, absolutely, possibly, by all means*

Adjuncts of temporality:

(f) Time:	*yet, still, already, once, soon, just*
(g) Typicality:	*occasionally, generally, mainly, for the most part*

Adjuncts of mood:

(h) Obviousness:	*of course, surely, obviously, clearly*
(i) Intensity:	*just, simply, merely, only, even, really, in fact*
(j) Degree:	*quite, almost, nearly, hardly, absolutely, totally*

This list served as the basis for the annotation guidelines written for the current research project. A few more lexical items which seemed to be synonymous to words from the previous list have also been annotated as modal adjuncts, e.g. *every day* (= regularly) or *a few times a year* (= sometimes), they express usuality. The words which express polarity are covered in the annotation of negation, not the one for modality. Not all adverbial groups, however, express modality. Some express a sort of appraisal (e.g. *understandably, fortunately*), called comment adjuncts in Halliday (1994, 83). In the annotation of modality in the EDNA corpus, appraisal is excluded from the annotation. The lexical items which function as modal adjuncts in EDNA are investigated in section 6.4.2.

5.2.4 Modal adjuncts in German

As in English, adverbial groups and prepositional phrases can express modality as well as appraisal in German. Only those adjuncts that express modality in the German newsgroup texts in EDNA were annotated, for example *wieder, ständig, eigentlich, immer, im Prinzip, Tag für Tag, vermutlich, klar, wirklich, tatsächlich*. In the annotation guidelines, we established a list of lexical items which seemed to express the same meaning as the modal adjuncts described by Halliday (1994, 82).

5.2.5 Grammatical metaphor

Matthews (2005, 224) defines metaphor as a "[f]igure of speech in which a word or expression normally used of one kind of object, action, etc. is extended to another". We speak of grammatical metaphor when we find one grammatical construction in a clause which is used to express some meaning that is typically expressed by other grammatical means. In our case here, we express modality not by a modal adjunct or modal auxiliary but by a superordinate clause containing a mental cognitive process, e.g. *I personally believe, I know, I think, I guess*. Consider the following examples 8 and 9:

(8) *I think a lot of it is water weight and muscle loss.*

(9) *I feel like I'm faking it staying with her.*

With the superordinate clause of *I think* in example 8 the person expresses probability, which is usually done with a modal adjunct, see example 10.

(10) *A lot of it is probably water weight and muscle loss.*

The same kind of paraphrasing can easily be done with example 9, where the first clause can be replaced by a modal adverb, see example 11.

(11) *I'm certainly faking it staying with her.*

The same kind of grammatical metaphor is possible in German, where a superordinate clause adds modality to the subordinate clause, as for example in 12 and 14. Both 12 and 14 can be paraphrased with a clause that contains a modal adjunct instead of the superordinate clause to express modality, see examples 13 and 15.

(12) *Ich hab die Vermutung, dass mein Körper so jede Woche aufs Neue bei Null anfängt.*

(13) *Mein Körper fängt wahrscheinlich so jede Woche aufs Neue bei Null an.*

(14) *Ich weiß, dass er keine neue Freundin hat.*

(15) *Er hat ganz sicher keine neue Freundin.*

The three types of realization of epistemic or root modality are comparable in English and German. Figure 5.1 below displays the complete annotation scheme for modality in English, which includes both the type of modality and the type of realization. In the end, each clause containing modality is annotated for the two features, see examples 16, 17 and 18.

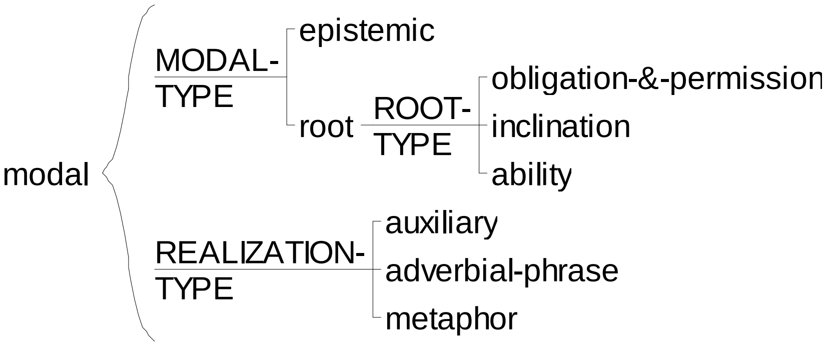

Figure 5.1 The annotation scheme for modality in English

(16) *Root modality, obligation, modal auxiliary: Two weeks ago I <u>had to</u> leave.*

(17) *Epistemic modality, adverbial phrase: I <u>usually</u> have a salad for lunch at school.*

(18) *Epistemic modality, grammatical metaphor: <u>I knew</u> I didn't want the night to end so early.*

5.2.6 Modal particles in German

In addition to the three realization types mentioned above, we find little words called *modal particles* in German. According to the Duden (Fabricius-Hansen et al. 2006, 597), modal particles (*Abtönungspartikel*) are particularly common in spoken language and they are not fillers without a function. Modal particles express stance, evaluation, expectation, assumptions and surprise. The multi-functionality of modal particles is also mentioned in Götze and Hess-Lüttich (1999, 328), who give the following examples (A – C):

A Surprise: Das ist *aber* eine Menge Geld.

B Evaluation: Das ist *ja* eine Katastrophe!

C Modality, likelihood: Das wird *schon* gut gehen.

The Duden (Fabricius-Hansen et al. 2006, 598) lists the most commonly used modal particles: *ja, denn, wohl, doch, aber, nur, halt, eben, mal, schon, auch, bloß, eigentlich, etwa, nicht, vielleicht, ruhig*.

In the annotation of EDNA, we did our best to annotate only those modal particles that expressed modality. There are, however, no tests to separate one function of those particles from another, meaning that the annotation may be less than objective.

5.2.7 Subjunctive verb forms in German

Matthews (2005, 360) defines the subjunctive as "[m]ood, especially in European languages, whose central role is to mark a clause as expressing something other than a statement of what is certain".

The subjunctive (*Konjunktiv*) is used to a greater extent in German than in English and had to be taken into account for the annotation of modality in my project. Fabricius-Hansen et al. (2006, 506) say in the *Duden Grammatik* that the verb moods indicative, subjunctive and imperative belong to the functional dimension of modality, the same as modal auxiliaries, clause types, modal adverbs and particles. They elaborate that the indicative mood is the unmarked mood, which is used as long as there is no reason to use a different mood. The subjunctive I (*Konjunktiv I*) is used in reported speech, to make clear that the speaker herself is not the source of the statement, thereby adding an element of uncertainty. Subjunctive II (*Konjunktiv II*) expresses that the statement is not true. (Fabricius-Hansen et al. 2006)

Figure 5.2 below shows the extended annotation scheme for modality in the German newsgroup corpus, with modal particles and subjunctive mood added to the realization types, thus, all five realization types are capable of expressing both types of modality, epistemic and root. This was the assumption at the beginning of the annotation process. The results of my study will show whether auxiliaries, adverbial phrases, metaphors, modal particles and the subjunctive verb form actually do express both types. Examples 19 to 23 demonstrate each realization type with a clause from EDNA.

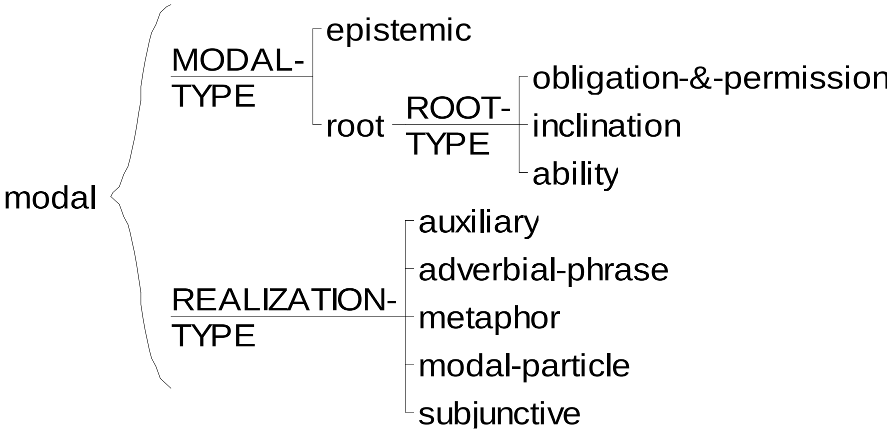

Figure 5.2 The annotation scheme for modality in German

(19) Root modality, ability, auxiliary: *Ich kann das Essen auch einfach nicht genießen.*

(20) Epistemic modality, adverbial phrase: *Vielleicht hat ja einer ähnliche Erfahrungen gemacht.*

(21) Epistemic modality, grammatical metaphor: *…, und ich glaub, ich hab mich wieder erkannt.*

(22) Epistemic modality, modal particle: *Ich kenne ihn doch schon mein ganzes Leben.*

(23) Epistemic modality, subjunctive: *Neulich hätten wir uns beinahe geküsst.*

Following from the fact that English has nine central modal auxiliaries, plus four marginal ones and some fixed idiomatic phrases, whereas German has only six central and three marginal modal auxiliaries, there must be differences in the use of the existing repertoire of the two languages to cover the same range of meanings. Then again, German has two more ways to realize modality. Do the German writers express modality more often, or are only the ways to express modality different? The results from the modality annotation of the EDNA corpus and the statistical tests for significance are presented next.

5.3 Quantitative analysis of modality

We start with the null hypothesis, which represents the basic assumption that there are no differences at all between the English and the German newsgroup texts in regard to the use of modality. More precisely, the null hypothesis here is that in both corpora we find the same amount of clauses containing a modal marker. The total amount for the calculation of clauses containing modality is the number of processes.

Please note that the number of processes (EN 1543, GN 1479) is slightly smaller than the number of rhemes (EN 1578, GN 1499). When I annotated processes, I annotated every process on the level of the main clause, and those that were on the first level of subordination, but not those that were subordinate in a subordinate clause. An example from EN would be *We both agreed that the sex (we had) was like making love with someone (you cared about).* The processes *to agree* and *to be* were annotated, but the processes *to have* and *to care* were not. When I annotated rhemes, however, all finite clauses were annotated, regardless of whether they were main or subordinate clauses. Thus, in the example above, there are two processes but four rhemes. That is the explanation for the difference in the total amount of processes and rhemes. Minor clauses were excluded altogether.

The tests of statistical significance for the experiential metafunction will definitely have to be based on the number of process types. For the sake of consistency, the tests of statistical significance for the interpersonal metafunction will also be based on the number of processes. The tests of statistical significance for the textual metafunction do not need to be based on the number of processes or rhemes because there, the rhemes are a feature inside the scheme, not outside of it. The total number of modal markers is 364 in EN and 687 in GN. The number of clauses without an instance of modality, however, cannot be calculated by simply subtracting the number of modal markers from the total number of clauses (processes). One clause may contain more than one instance of modality, therefore the instances per clause had to be studied; examples are shown in 24 to 29.

One modal marker per clause:

(24) EN: *At first, I would be devastated*

(25) GN: *Ich weiß wirklich nicht mehr weiter*

Two modal markers per clause:

(26) EN: *Rarely would she apologize*

(27) GN: *Ich will doch gesund werden*

Three modal markers per clause:

(28) EN: *I really really need to know how to purge better*

(29) GN: *Sicherlich würde ich ja in ein paar Monaten wieder zu Hause sein*

Thus, table 5.1 shows the results of the significance test for clauses with at least one modal marker per clause, but possibly two or three, and the total number of clauses without any modal marker. The null hypothesis suggests there is no significant difference between EN and GN.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	χ^2	**S**
+ Modality	322	575	21	39	GN	82.34	+ + +
- Modality	1,220	904	79	61	EN	34.77	+ + +
Column total	1,542	1,479	100	100			

Table 5.1 Raw and relative numbers, χ^2 and significance for clauses with /without modal marker

It becomes clear that the German newsgroup text writers use significantly more modal auxiliaries, adverbs, grammatical metaphors (and modal particles and subjunctive verb forms) than the English writers, with the threshold of 10.83 for $p < 0.001$ (df=1). In GN, almost 40% of all clauses carry at least one modal marker, whereas in EN, there are only half as many clauses (21%) with a modal marker. Now how do these modal markers distribute over the number of clauses with more than one modal marker? The results can be seen in Table 5.2, where the column total is the total number of instances of modality in the newsgroup texts and the null hypothesis the same as above, i.e., no significant difference between the two subcorpora.

Feature	EN F	GN F	EN%	GN%	Overuse	χ^2	S
First modal marker	322	575	89	83	-	0.63	-
Second modal marker	38	101	10	15	-	3.27	-
Third modal marker	4	11	1	2	-	0.42	-
Column total	364	687	100	100			

Table 5.2 Raw and relative numbers, χ^2 and significance for instances of modal markers per clause

Although the raw and relative numbers of clauses with one or more than one modal marker (see table 6.1) differ to a substantial degree in EN and GN, the distribution of more than one modal marker per clause does not. There are not significantly more clauses with two or even three modal markers in the German newsgroup texts, contrary to what we might have expected. On the basis of these results, the null hypothesis cannot be rejected; the difference in the frequency of modal markers is considerable, but most of the time there is only one modal marker per clause in both EN and GN.

5.3.1 Types of modality

We continue by taking a closer look at the types of modality, with the results displayed in table 5.3. The null hypothesis is that epistemic and root modality is expressed to the same degree in both corpora.

Feature	EN F	GN F	EN%	GN%	Overuse	χ^2	S
Epistemic modality	255	485	70	70	-	0.01	-
Root modality	109	202	30	30	-	0.02	-
Column total	364	687	100	100			

Table 5.3 Raw and relative numbers, χ^2 and significance for main types of modality

Although the total number of instances of modality is much higher in German, the distribution of epistemic and root modality in the two corpora is the same. The difference is not significant (3.84 threshold for $p < 0.05$, df=1). The null hypothesis cannot be rejected; in both corpora epistemic and root modality is expressed to the same extent, with epistemic modality more than twice as frequent as root modality. Both the English and the German authors tend to strengthen or weaken their statements, rather than expressing obligation, permission, ability or inclination.

5.3.2 Types of root modality

Although root modality is expressed in only a third of all clauses with a modal marker, we can further investigate how the numbers are distributed for the three types of root modality, i.e. obligation and permission, inclination and ability. The null hypothesis is that all three types occur equally frequently in both corpora.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Obligation	51	86	47	43	EN	0.29	-
Inclination	19	74	17	36	GN	8.73	+
Ability	39	42	36	21	EN	6.11	+
Column total	109	202	100	100			

Table 5.4 Raw and relative numbers, χ^2 and significance for types of root modality

The results in table 5.4 give reason to reject the null hypothesis. The expression of obligation and permission does not differ to a significant degree. The German authors, however, express inclination significantly more often than their English counterparts (5.99 threshold for $p < 0.05$, df=2), whereas the English writers express ability more often than the German writers do (5.99 threshold for $p < 0.05$, df=2). The German authors seem to have a tendency towards saying that they are willing to do something, without saying that they can, whereas the English authors tend to say something can be done, without claiming that they want to do it. The next section will show how the instances of modality are realized.

5.3.3 Realization types

Finally, tables 5.5 and 5.6 show the results for the different ways of realizing modality. The English language does not provide writers with a set of modal particles, and the subjunctive form is used to a much lesser extent in English than in German, but are the other three types used equally frequently in the two corpora? The null hypothesis is that they are. In the calculation of the chi-squared value, only modal auxiliaries, modal adjuncts and grammatical metaphor are considered because the χ^2 value cannot be calculated with zero frequencies (Gries 2008, 171).

Feature	EN F	GN F	EN%	GN%
Modal auxiliary	198	256	55	37
Modal adjunct	99	172	27	25
Grammatical metaphor	67	55	18	8
Modal particle	-	178	-	26
Subjunctive	-	26	-	4
Column total	364	687	100	100

Table 5.5 Raw and relative numbers for modality realization types in EN and GN

Feature	EN F	GN F	EN%	GN%	Overuse	χ^2	S
Modal auxiliary	198	256	55	53	EN	0.08	-
Modal adjunct	99	172	27	36	GN	4.59	-
Grammatical metaphor	67	55	18	11	EN	7.10	+
Column total	364	483	100	100			

Table 5.6 Raw and relative numbers, χ^2 and significance for modality realization types in EN and GN

Our null hypothesis can once again be rejected; there is one statistically significant difference between the two corpora. The χ^2 value, however, is only of limited value due to the fact that the English language does not have modal particles and the subjunctive has not been annotated because it is hardly ever used in English newsgroup texts. With this in mind, we see that in the English corpus, modality is mostly expressed with modal auxiliaries (55%), followed by modal adjuncts (27%) and grammatical metaphor (18%). In the German newsgroup texts, modal auxiliaries are the most frequent realization type (37%), but far less frequent than in the English texts. Modal adjuncts (25%) and modal particles (26%) together account for more than half of all expressions of modality. The grammatical metaphor (8%) is used less often in the German than in the English texts, and the subjunctive form (4%) of the verb is not very frequent at all in GN. The only statistically significant difference lies in the use of grammatical metaphors to express modality, this feature is more frequent in the English newsgroup texts, with the threshold for $p < 0.05$ at 5.99 (df=2). We can reject the null hypothesis on these grounds, although the differences in both the raw and relative numbers are small.

5.3.4 Summary

The tests for statistical significance with regard to modality have revealed a few interesting results, which are summarized in figure 5.3.

System	**Divergence**	**Divergent feature**	**Divergent corpus**
Modal / non-modal	Significant	Modal markers	GN
More than one modal marker per clause	Not significant	-	-
Epistemic / root modality	Not significant	-	-
Type of root modality	Significant Significant	Inclination Ability	GN EN
Realization types	Significant	Grammatical metaphor	EN

Figure 5.3 Summary of tests of statistical significance of modality

Following the quantitative study of the system of modality, we investigate the lexical items which are used to express modality.

5.4 Analysis of lexical items used to express modality

5.4.1 Modal auxiliaries

After looking at the quantitative results of the study of the newsgroup corpora, we focus on how modality is expressed differently in the English and German newsgroup texts. We begin by looking at the modal auxiliaries, and investigate which modal auxiliary is most frequently used to express which type of modality. Following that, we study the most frequent modal adverbs in EN and GN, the most frequent modal particles in GN, the subjunctive verb forms in GN, and finally the grammatical metaphors in both EN and GN.

The first table in this set is table 5.7. It shows the most frequent modal auxiliaries that express epistemic modality in the EDNA corpus. The first lexical item that occurs only once in these frequency lists is used as the cut-off point (note that the lexical items are ordered alphabetically if they occur equally frequently). Therefore, the lists for EN and GN may not be of the same length.

N	Word EN	F	%	Word GN	F	%
1	WOULD	40	44.94	WERDEN	37	68.52
2	WILL	21	23.60	KOENNEN	11	20.37
3	MAY	6	6.74	MUESSEN	2	3.70
4	COULD	5	5.62	SOLLEN	2	3.70
5	GOINGTO	5	5.62	DUERFEN	1	1.85
6	SHOULD	4	4.49	MOEGEN	1	1.85
7	MIGHT	2	2.25			
8	USEDTO	2	2.25			
9	WOULDLIKETO	2	2.25			
10	CAN	1	1.12			
11	HAVETO	1	1.12			
	Total	89		Total	54	

Table 5.7 Frequency of modal auxiliaries used to express epistemic modality

We see that in both languages, a strong certainty that the statement is true (or rather, will become true) is expressed, with *would* and *will* in EN and *werden* in GN. In the English texts, *would* and *will* are followed in frequency by auxiliaries which express a weaker conviction of the certainty or likelihood that the statement is true, i.e. *may, could, should, might, used to, would like to*. *Going to* is a stronger auxiliary and falls into the same category as *will* and *would*, i.e. it expresses a strong conviction that a statement will become true. These three modal auxiliaries are also used as future tense markers. In my work, they are considered modal markers, following the assumption that there is no such thing as a definite future. Any statement about what will happen in the future is a hypothesis, and writers/speakers only express a high degree of certainty that the hypothesis will be verified at some point in time, see the discussion in chapters 5.2.1 and 5.2.2.

In the German texts, we find a smaller range of auxiliaries to express epistemic modality than in the English texts. Apart from *müssen*, which occurs twice and is a strong modal marker, there are only *können, sollen, dürfen* and *mögen;* all of these express a weak conviction that what is said is true.

Table 5.8 below gives the frequency of modal auxiliaries used to express obligation and permission in the newsgroup texts.

N	Word EN	F	%	Word GN	F	%
1	CAN	16	31.37	KOENNEN	30	34.88
2	HAVETO	16	31.37	SOLLEN	30	34.88
3	NEEDTO	11	21.57	MUESSEN	21	24.42
4	COULD	5	9.80	HABENZU	4	4.65
5	MUST	2	3.92	DUERFEN	1	1.16
6	SHOULD	1	1.96			
	Total	51		Total	86	

Table 5.8 Frequency of modal auxiliaries used to express root modality; obligation

Although the most frequent auxiliaries, *can* and *können*, are also used to express ability, they are multifunctional and apparently they are frequently used to express obligation. Furthermore, obligation is expressed with the auxiliaries *have to, need to, must* and *should* in the English texts and *sollen, müssen, haben zu* + infinitive in the German texts.

Table 5.9 gives the results for the most frequent modal auxiliaries used to express inclination, i.e. speakers express that they are willing to do something. In the analysis in section 5.3.2 it became clear that the German writers state inclination significantly more often than the English writers.

N	Word EN	F	%	Word GN	F	%
1	NEEDTO	5	26.32	WOLLEN	53	71.62
2	WOULD	4	21.05	MOEGEN	11	14.86
3	GOINGTO	3	15.79	WUERDEGERN	6	8.11
4	HAVETO	3	15.79	KOENNEN	3	4.05
5	WILL	2	10.53	MUESSEN	1	1.35
6	CAN	1	5.26			
7	SHOULD	1	5.26			
	Total	10		Total	74	

Table 5.9 Frequency of modal auxiliaries used to express root modality; inclination

Similar to the expression of epistemic modality, obligation and permission, the English writers also used a greater variety of modal auxiliaries to express inclination. The most frequent modal auxiliary in GN is *wollen*, followed by *mögen*, e.g. *ich möchte einfach nicht noch einmal von zuhause weg,* and *würde gern*.

When we consider the German auxiliary *wollen*, we might come to think that a very likely equivalent to the German *wollen* + main verb is the English lexical verb *want*, i.e. *want to* + lexical verb. A look at the annotated corpus reveals that in the construction *want to* + lexical verb, *want* has not been annotated as modal auxiliary, but as the main verb in a mental process, e.g. *I want to get over my ex completely*. *Want to* + lexical verb occurs 29 times, *wanted to* + lexical verb 4 times, and *wanting to* + lexical verb 2 times, thus 35 times altogether. *Want to* + lexical verb is a borderline case. It has been annotated as lexical verb in a mental process, but it could just as well be a modal auxiliary expressing inclination. Had it been annotated as modal auxiliary expressing inclination in EN, then there probably would not have been a significant difference in frequency in EN and GN with regard to this aspect. *Want* also appears in combination with a nominal group, e.g. *I don't want this direct rejection*, here it is more like a main verb, and more equivalent to the German construction of *wollen* + nominal group, e.g. *doch will ich diese Hilfe?* In both constructions, *want* / *wollen* + nominal group have been annotated as lexical / main verb.

One conclusion, however, is that the German writers 'want to do' a lot more than the English writers. English writers, on the other hand, express ability significantly more often than their German counterparts, as has been revealed in chapter 5.3.2. In table 5.10 below the modal auxiliaries used to express ability are shown.

N	**Word EN**	**F**	**%**	**Word GN**	**F**	**%**
1	CAN	37	94.87	KOENNEN	41	97.62
2	COULD	2	5.13	MUESSEN	1	2.38
	Total	39		Total	42	

Table 5.10 Frequency of modal auxiliaries used to express root modality; ability

The most obvious, and most frequent auxiliary in both corpora for expressing ability is the auxiliary *can / können*, there is no striking difference in regard to how ability is expressed.

I would like to briefly refer to the annotation guidelines which were used to annotate the EDNA corpus, and evaluate the list of modal auxiliaries in the guidelines. The list contains 18 central and marginal auxiliaries in the English language and ten in the German language. Out of the nine central modal auxiliaries in the English language taken from Biber et al. (1999), eight occurred in EN. There was no instance of the ninth central modal auxiliary, *shall*. Of the marginal auxiliaries and fixed idiomatic expressions (Biber et al. 1999) there were *need to, used to, have to, had better* and *be going to* in EDNA. There was no instance of *ought to, dare to, have got to* and *be supposed to*. One other construction has been added to the list of modal auxiliaries, *would like to* + lexical verb, e.g. *I would like to tell her all my feelings*. Out of the ten central and marginal modal auxiliaries in the German language (Fabricius-Hansen et al. 2006), eight were found in GN: *dürfen, können, mögen, müssen, sollen, wollen, haben* + *zu, werden*. The other two, *brauchen* and *sein* + *zu,* were not found. One construction has been added to the list of modal auxiliaries, *würde gern* + lexical verb, e.g. *ich würde gern zu ihm fahren*.

Interestingly, all instances of root modality were realized by a modal auxiliary in both the English and the German newsgroup texts. There is not a single instance in either corpus where obligation, permission, inclination or ability was expressed by a modal adjunct or grammatical metaphor.

5.4.2 Modal adjuncts

We now proceed to focus on the modal adjuncts which were used to express modality in the EDNA corpus. Modal adjuncts in both languages express epistemic modality only. The most frequently used modal adjuncts are shown in table 5.11 below.

N	Word	F	%	Word	F	%
1	REALLY	24	24.24	WIEDER	26	15.12
2	ALWAYS	9	9.09	IMMER	23	13.37
3	SOMETIMES	9	9.09	VIELLEICHT	15	8.72
4	ACTUALLY	6	6.06	WIRKLICH	14	8.14
5	USUALLY	6	6.06	OFT	12	6.98
6	MAYBE	5	5.05	IRGENDWIE	9	5.23
7	ALLTHETIME	3	3.03	EIGENTLICH	6	3.49
8	ALMOST	3	3.03	IMMERWIEDER	6	3.49
9	CERTAINLY	2	2.02	MANCHMAL	6	3.49
10	EVERYDAY	2	2.02	REGELMAESSIG	4	2.33
11	LITERALLY	2	2.02	SICHER	4	2.33
12	MANYTIMES	2	2.02	MEISTENS	3	1.74
13	MOSTOFTHETIME	2	2.02	NATUERLICH	3	1.74
14	OFTEN	2	2.02	SICHERLICH	3	1.74
15	PROBABLY	2	2.02	STAENDIG	3	1.74
16	RARELY	2	2.02	ABUNDZU	2	1.16
17	TOTALLY	2	2.02	IMPRINZIP	2	1.16
18	AFEWTIMESAYEAR	1	1.01	JEDENTAG	2	1.16
				OEFTERS	2	1.16
				ANDAUERND	1	0.58
		84	84.84		146	84.87
	Other	15	15.16	Other	26	15.13
	Total	99	100	Total	172	100

Table 5.11 The most frequently used modal adjuncts in the EDNA corpus

In the English newsgroup texts, it seems as if the authors use modal adjuncts mostly to strengthen their statements, with adjuncts like *really, always, actually, usually, all the time, certainly, every day, literally, many times, most of the time, often, totally*, adding up to 62 out of 99 modal adjuncts. The other modal adjuncts express less conviction by the writer and function to articulate vagueness, e.g. *sometimes, maybe, almost, probably, rarely*, with 21 instances altogether. There are 16 modal adjuncts which occur only once and which were not included in the table (apart from *a few times a year* as the cut-off point). It has not been assessed whether their function is weakening or strengthening. But even if all of the remaining 16 modal adjuncts were used to weaken a statement, the strengthening adjuncts would still outnumber them.

Just like the writers in the English newsgroup texts, the German writers mostly use modal adjuncts to strengthen a statement, with lexical items such as *wieder, immer, wirklich, oft, immer wieder, regelmäßig, sicher, meistens, natürlich, sicherlich, ständig, im Prinzip, jeden Tag, öfters*, all instances together make up 108 out of the 172 modal adjuncts found in GN. To a much lesser extent, the writers weaken their statements with a modal adjunct like *vielleicht, irgendwie, eigentlich, manchmal, ab und zu*, these add up to only 38 instances in 172. There are 27 other modal adjuncts that occur only once. Surprisingly, modal adjuncts that occur only once make up 15% of all modal adjuncts in both corpora. A special case in the German newsgroup texts may be the modal adjunct *wieder*. It was annotated as modal adjunct due to its close relationships to phrases like *immer, immer noch, immer wieder*, which express that something always happens, or something happens repeatedly, thereby expressing a strong likelihood that it will happen again, and giving the statement a strong probability. *Wieder*, however, expresses the return to a state that a person has been in before, see examples 30 to 34 below.

(30) Ich könnte schon wieder total viel Sport machen

(31) Sicherlich würde ich ja in ein paar Monaten wieder zu Hause sein

(32) Jetzt bin ich wieder total verwirrt

(33) Wir kommen heute wieder gut klar

(34) Gestern Abend ist sie wieder fort gegangen

It is different from the other modal adjuncts that express the probability or usuality of an event or action in that it expresses that the event or action has happened in the past, then stopped, and now begins again. It does not give an evaluation of the likelihood of a statement. Perhaps we cannot even count *wieder* as a modal adjunct, but this is a dilemma I was unaware of at the beginning of the annotation process, and which will have to be resolved in the next annotation of a German corpus.

5.4.3 Modal particles in GN

Modal particles are exclusive to the German newsgroup text corpus and are not used in the English language. Table 5.12 displays the most frequently used modal particles in GN.

N	Word	F	%
1	AUCH	53	29.78
2	EINFACH	32	17.98
3	JA	17	9.55
4	DOCH	15	8.43
5	MAL	9	5.06
6	ABER	7	3.93
7	ZWAR	7	3.93
8	NUR	5	2.81
9	WIEDER	5	2.81
10	HALT	4	2.25
11	EIGENTLICH	3	1.69
12	NAEMLICH	3	1.69
13	WOHL	3	1.69
14	DENN	2	1.12
15	SCHON	2	1.12
16	ALSO	1	0.56
		168	94.4
	Other	10	5.6
	Total	178	100

Table 5.12 Most frequent modal particles in the German newsgroup texts

The German modal particle *auch* is the most frequently used one in the German part of EDNA, followed by *einfach, ja* and *doch*. There are more modal particles (178) than modal adjuncts (172). It would be interesting to compare these results to a reference corpus, to see whether modal particles dominate in other registers as well. Due to time constraints, this has to be postponed to future studies. At present, this result and the frequencies of the individual modal particles do not lend themselves to an obvious conclusion.

Within the group of modal particles, there are five instances of *wieder*, shown in the examples 35 to 39.

(35) *Alle Wunden sind wieder aufgebrochen*

(36) *Irgendwann schaffte ich es, mein Leben wieder in den Griff zu bekommen.*

(37) *Leider hatten wir irgendwann letztes Jahr wieder Kontakt.*

(38) *Und es ist wieder dasselbe passiert.*

(39) *[…], warum ich hier wieder aufschlage.*

In these examples, *wieder* is used in the same way that was discussed in the previous section about modal adjuncts, it signals the return to a state or event. There is no difference in function, we must therefore consider either these five instances as annotation mistakes, or the 26 instances where *wieder* was annotated as modal adjunct. This inconsistency in the annotation is due to the less-than-clear status of *wieder*, and has to be considered in future annotation projects.

5.4.4 Subjunctive verb forms in GN

The subjunctive verb forms in the German texts that express modality are not as numerous as one might have expected; there are only 26 instances in GN, see table 5.13, plus another 25 instances of the auxiliary *werden* in subjunctive verb form.

N	Word	F	%
1	HAETTE	9	34.62
2	WAERE	8	30.77
3	SEI	3	11.54
4	WUENSCHTE	2	7.69
5	BRAEUCHTE	1	3.85
6	BRAUCHE	1	3.85
7	HAETTEN	1	3.85
8	SEIEN	1	3.85
	Total	26	100

Table 5.13 The frequencies of subjunctive verb forms in GN

Not all 26 instances of subjunctive verb forms, however, are used to express modality, i.e. hypothesis or potentiality. Out of the 26 instances, 8 are used in reported speech. We find five instances of subjunctive I (Konjunktiv I), i.e. *sei, seien, brauche*. Of these, one *sei* is used in a clause expressing potentiality, see example 40.

(40) *…, sei es auch gewesen, was es will*

The other four subjunctive I verb forms appear in reported speech, clarifying that the writer is only passing on information given by somebody else, not information she herself knows to be true, see examples 41 and 42.

(41) *..., und meinte, ich sei eifersüchtig*

(42) *Sie sagte, dass dies halt eine (Art) von Beziehung sei und sie ihre Freiheit brauche*

As with the subjunctive I verb forms in the German language, the subjunctive II verb forms also have two functions. The first use is in reported speech (or reported thought), and we find four examples for this function in our corpus, see 43 to 45 below.

(43) *Dann sollte sie mir aber auch signalisieren, ich hätte die „ Macht "*

(44) *..., aber sie meinte, es wäre für mich sehr langweilig*

(45) *... meinte er dann, dass ich wohl was falsches da hinein interpretiert hätte. Er hätte mir alles gesagt.*

The second function is the expression of hypothesis or potentiality, i.e. to express epistemic modality. There are 17 instances in GN; some examples are given in 46 to 48 below.

(46) *…, und das wäre dann alles kaputt*

(47) *Neulich hätten wir uns beinahe geküsst*

(48) *Das wäre schon aus finanzieller Sicht besser*

In conclusion, 8 out of 26 subjunctive verb forms are used in reported speech, not to express epistemic modality. It is striking that although there are 26 instances, these are realized by only four different verbs, i.e. by *haben, sein, wünschen* and *brauchen*.

A fifth verb that frequently occurs in the subjunctive II verb form in GN is *werden*. The subjunctive form of the auxiliary verb *werden*, i.e. *würden*, has been annotated as *auxiliary* in the German part of the EDNA corpus for practical reasons, not as both *auxiliary* and *subjunctive*. Table 5.14 demonstrates the frequency of *würden* in its inflected verb forms and shows that *würden* is more frequent than any of the other subjunctive verb forms discussed above.

N	**Word**	**F**	**%**
1	WUERDE	22	88
2	WUERDET	2	8
3	WUERDEN	1	4
	Total	25	100

Table 5.14 The modal auxiliary *werden* in subjunctive verb form in GN

The subjunctive II verb forms of *werden* serve two functions as well. The first is to express hypothesis or potentiality; there are 17 instances of this in GN, see examples 49 to 51.

(49) *Sicherlich würde ich ja in ein paar Monaten wieder zu Hause sein*

(50) *Ich würde so gerne mit ihm darüber sprechen*

(51) *Was würdet ihr mir raten?*

The second function is the use in reported speech or thought, and although not as frequent as the first function, there are eight instances of this in GN, see examples 52 to 54 below.

(52) *Die meisten denken, ich würde einfach nur kotzen, um abzunehmen*

(53) *..., und ich bin mir nicht sicher, ob ich es nicht noch mal tun würde*

(54) *Ich habe ihr gesagt, dass dieses Modell von alle zwei Wochen für mich auch nicht in Frage kommen würde*

The subjunctive I and II are not used frequently in the corpus to express epistemic modality, and of all the instances of subjunctive verb forms, about 1/3 is used in reported speech rather than in hypothetical clauses. Is it possible that the subjunctive is not used in spoken language as much as in written language? Fabricius-Hansen et al. (2006, 529) state that in written language, the subjunctive is predominantly used in reported speech, which is not the case in the EDNA corpus. The results in the EDNA corpus do, however, support a claim in Götze and Hess-Lüttich (1999, 132) that in spontaneous spoken language, a form of *würde* is predominant in reported speech: "In Textsorten spontan gesprochener Sprache gilt: bei der indirekten Rede herrscht die *würde*-Form vor."

5.4.5 Grammatical metaphor

In what follows, the focus is a qualitative study of the grammatical metaphors in the EDNA corpus which are used to express epistemic modality. This is done by way of a superordinate clause which adds a degree of likelihood to the subordinate clause, instead of a modal auxiliary or adjunct which could serve the same purpose. Only those superordinate clauses which have the 1st person singular pronoun as subject have been annotated as grammatical metaphor, and a few expressions which serve the same function, e.g. *it seems, it appears* in EN and *mag sein* in GN. Table 5.15 below shows the results.

N	Word EN	F	%	Word GN	F	%
1	IKNOW	13	19.40	ICHWEISS	21	38.18
2	ITHINK	13	19.40	ICHBINMIRSICHER	4	7.27
3	IMSURE	7	10.45	ICHFRAGEMICH	4	7.27
4	IBELIEVE	5	7.46	ICHGLAUBE	4	7.27
5	IFEEL	5	7.46	ICHHABEDASGEFUEHL	4	7.27
6	IFEELLIKE	5	7.46	ICHDENKEMIR	2	3.64
7	IKNEW	5	7.46	ICHNEHMEAN	2	3.64
8	IGUESS	4	5.97	ICHBINSICHER	1	1.82
9	ITHOUGHT	3	4.48	ICHBINUEBERZEUGT	1	1.82
10	ITSEEMS	3	4.48	ICHDACHTEMIR	1	1.82
11	BELIEVEME	1	1.49	ICHHABEANGSTDASS	1	1.82
12	IREALIZE	1	1.49	ICHHABEBEDENKEN	1	1.82
13	ISUPPOSE	1	1.49	ICHHABEDIEVERMUTUNG	1	1.82
14	ITAPPEARS	1	1.49	ICHHABEKEINEAHNUNG	1	1.82
15				ICHHATTEMIREINGEBILDET	1	1.82
16				ICHHOFFE	1	1.82
17				ICHSAGSEUCH	1	1.82
18				ICHVERMUTE	1	1.82
19				ICHWUSSTE	1	1.82
20				MAGSEIN	1	1.82
21				MIRISTKLAR	1	1.82
	Total	67	100	Total	55	100

Table 5.15 The frequency of superordinate clauses used as grammatical metaphor to express epistemic modality in the EDNA corpus

We see that although in the English newsgroup texts there is a higher total number of grammatical metaphors, the variety of clauses is smaller than in the German texts, where writers use more different main clauses to express a degree of probability. Just as with the modal auxiliaries and adjuncts, these grammatical metaphors convey a stronger or weaker likelihood that the following statement is true.

5.5 Negation – theoretical background

Together with modality, polarity is the second aspect that realizes the interpersonal metafunction in Systemic Functional Grammar. In this subchapter, I introduce the theoretical background of syntactic negation, morphological negation, and negation by conjunctions, located in the textual metafunction.

5.5.1 Syntactic negation in English

Serving as our starting point once again, Halliday (1994) distinguishes negative polarity, i.e. clauses including syntactic negation, and positive polarity, i.e. clauses without a negation marker:

> Polarity is the choice between positive and negative, as in *is / isn´t, do / don´t*. Typically, in English, polarity is expressed in the Finite element. […] However, the possibilities are not limited to a choice between yes and no. There are intermediate degrees: various kinds of indeterminacy that fall in between, like ´sometimes´ or ´maybe´. These intermediate degrees, between the positive and negative poles, are known collectively as MODALITY. (Halliday 1994, 88)

According to Halliday (1994), the interpersonal metafunction is realized along a continuum, from the positive pole on one end to the negative pole on the other, as demonstrated on the following example from EDNA: *we never plan our meetings.*

We plan our meetings.
*We **always** plan our meetings.*
*We **sometimes** plan our meetings.*
*We **rarely** plan our meetings.*
*We **never** plan our meetings.*
*We **do not** plan our meetings.*

After discussing the grey area between *yes* and *no* in the previous subchapter on modality, here we will be concerned with the negative pole of this cline. The example above illustrates that a positive clause is the unmarked option, and by far the most frequently used, whereas a negative clause is less frequent and a marker of polarity is essential (Matthiessen 1995, 477). Both Halliday (1994, 88) and Matthiessen (1995) position the polarity marker inside the verbal group; Matthiessen calls this the pure negative, comprising *not* and *n't*. Fawcett (2008, 127) calls *not* the strong negative and *n't* the unmarked negative.

Matthiessen (1995) adds another two categories for polarity markers, which he calls combined negative, i.e. negative occurrence and negative specificity. Negative occurrence includes the modal adjuncts of degree (*hardly, scarcely, barely*, …), of usuality (*seldom, rarely, never*, …) and of time (*no longer, no more*). Except for the last two examples, *no longer* and *no more*, which include the negation marker *no*, the modal adjuncts are included in the annotation of modality, but not negation, in this work. There needs to be a clear line to separate polarity from modality in an annotation project.

The second type of combined negative, negative specificity, lists determiners and pronouns, which Matthiessen (1995) calls non-specific determination, e.g. *few, little, no, none, nobody, nothing*. Here again, the line is fuzzy and cannot be used in an annotation project. In the annotation of EDNA, only those pronouns that mean "zero" are annotated for negative polarity.

In a later publication, Halliday and Matthiessen (1999, 64) are clearer about the categories of modality and polarity:

> [T]he process element is either polar (positive/negative) or modal (some intermediate degree between positive and negative); it may embody phase, or aspect; and it will refer to past, present or future time. Polarity and modality derive from the interpersonal perspective on the process.

König and Gast say nothing about the line between polarity and modality or about the frequency of negation markers, they just point out the difference in the freedom of the word order: whereas "*not* is placed after the first auxiliary verb in English"(König and Gast 2007, 182), in German "the negative particle *nicht* can be moved around relatively freely in the Middle Field, depending on its scope" (König and Gast 2007, 183). Figure 5.4 displays the system network for the options of building a verbal group, of which polarity is one and modality another, as in the present work.

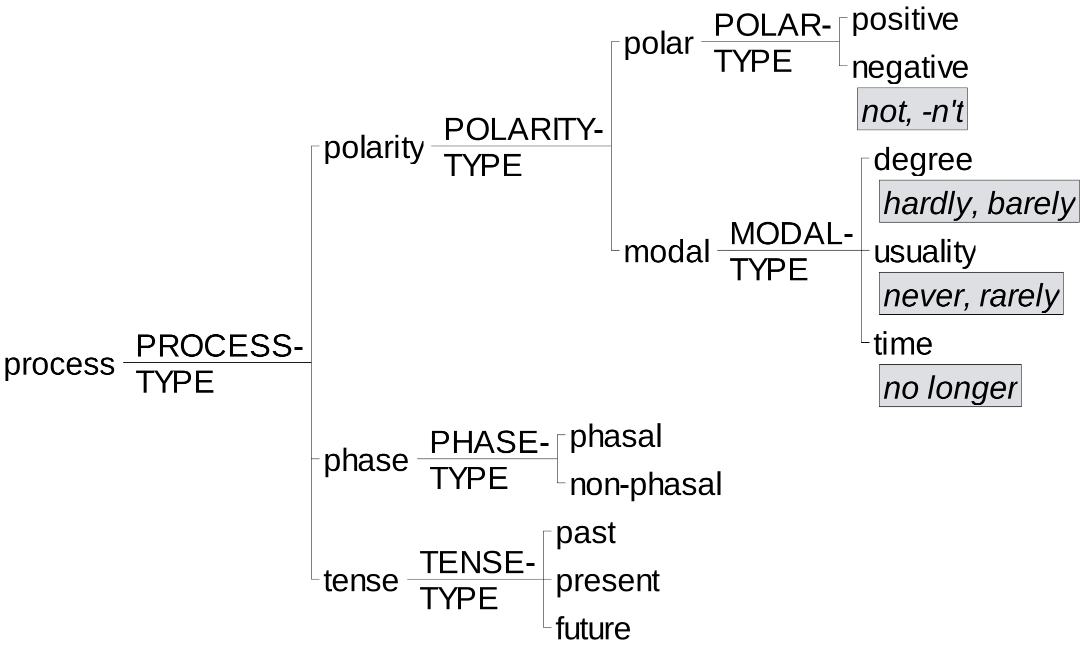

Figure 5.4 System network for verbal group, from Halliday and Matthiessen (1999, 65).

There are two main types of syntactic negation. The first is the *not*-negation, where the negation marker *not* / *n't* is inside the verbal group and negates the whole clause, see examples 55 and 56.

(55) *Funny how starving yourself doesn't seem silly.*

(56) *I did not restore any weight.*

With the syntactic negation marker *not* in the verbal group, we occasionally come across a phenomenon called 'negative raising' or 'negative transfer'. Negative transfer takes place when speakers / writers do not negate the clause where the negation marker logically belongs, but a superordinate clause, as in the following examples 57 and 58. It is not the thinking that is being negated, i.e. the person is not saying that she is not thinking. It is only the negation marker which travelled from the subordinate to the superordinate clause. The annotation scheme for the BTC includes both direct and transferred negation of the verbal group.

(57) *I don´t think I´m getting enough veggies, though.*

= I think I´m not getting enough veggies, though.

(58) *However, I don't think that is the case.*

= However, I think that is not the case.

The second main type of syntactic negation is the *no*-negation, where the negation marker is part of a nominal group, i.e. either a noun with the article *no*, or a pronoun: *no, nobody, no one, none, nothing, nowhere,* see examples 59 and 60 below.

(59) *There was basically no physical relationship.*

(60) *No one noticed what was happening to me.*

This type negates only a phrase inside the clause, but not the clause itself, i.e. the negation is shifted down in rank. Halliday (1985) includes a system network for deicticity, required for nominal groups, see figure 5.5. One of the options is 'total amount', which can be negative, e.g. there is *nothing* or *no problem*.

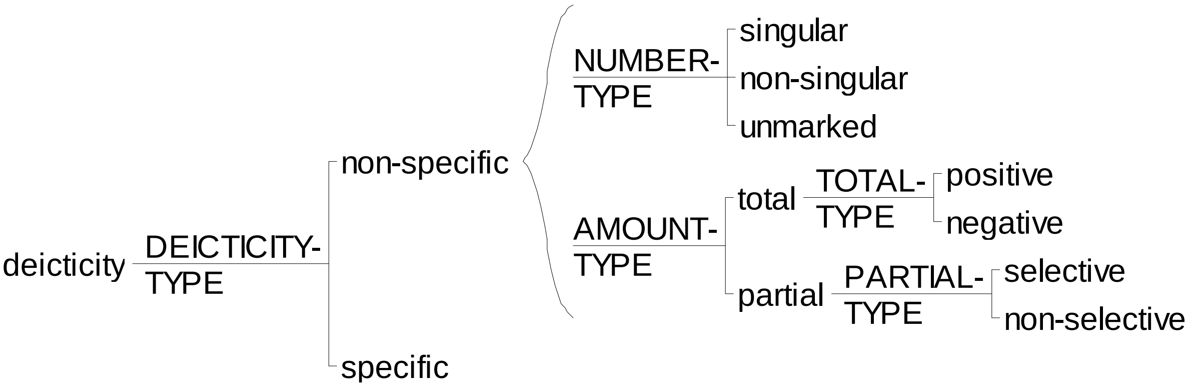

Figure 5.5 The system network for deicticity, relevant for nominal groups, from Halliday (Halliday 1985, 160-161)

Biber et al. (1999) give a percentage of the varying ratio of *not*-negation to *no*-negation, depending on the register, see table 5.16. The ratio for the newsgroup texts is close to the register of conversation.

	Conversation	Fiction	News	Academic	Newsgroup
Not-negation	90%	75%	65%	75%	84%
No-negation	10%	25%	35%	25%	16%

Table 5.16 Ratio of no- and not-negation, from (Biber et al. 1999, 170)

Quirk et al. (1985) distinguish three types of negation: clause negation (*not*-negation), local negation (*no*-negation) and predication negation. Predication negation is a rare form that occurs with denials or permissions where one person grants another permission to not do something, e.g. *they may not go swimming* (= they are allowed to not go). Since predication negation is rare and does depend on intonation for correct interpretation, I did not include this type in the annotation of the written texts in the EDNA corpus.

5.5.2 Method for finding all syntactic negation markers in EN

Syntactic negation markers have a huge advantage to them, they are easy to identify in a clause because of the finite number of different forms in which they can occur. Therefore, in order to find all the negation markers in EDNA, both the English and the German part of EDNA were PoS-tagged. The English original texts were tagged with the BNC C5 tag set, using the CLAWS tagger (Rayson 2010). Table 5.6 shows the part-of-speech tags relevant for finding syntactic negation markers in English texts.

BNC C5 tag	**Word class**	**Group class**	**Example**
XX0	Negative particle (adverb)	Verbal group	*not, n't*
PNI	Indefinite pronoun	Nominal group	*no one, nothing*
AT0	Article	Nominal group	*no point*
AV0	Adverb	Adverbial group	*no longer, never*
CJS	Conjunction	-	*nor, whether or not*
ITJ	Interjection	-	*no!*

Figure 5.6 The part-of-speech tags for negation markers in the BNC C5 tag set

Thus, after tagging a text, a clause would resemble the ones in examples 61 and 62, with a negation marker in the verbal group:

(61) *He_PNP just_AV0 does_VDZ n't_XX0 love_VVI me_PNP ._.*

(62) *My_DPS wife_NN1 feels_VVZ that_CJT a_AT0 married_AJ0 man_NN1 should_VM0 not_XX0 have_VHI female_AJ0 friends_NN2 ,_,*

The automatic PoS-tagger finds negation markers not only in verbal groups, of course, but also in nominal groups, as articles and indefinite pronouns, see examples 63 and 64:

(63) *[…] and_CJC has_VHZ had_VHN no_AT0 problem_NN1 with_PRP this_DT0 until_PRP recently_AV0 ._.*

(64) *[…] when_CJS I_PNP try_VVB NOTHING_PNI HAPPENS_VVZ !_!*

The PoS-tagging reveals that syntactic negation markers can also be found in other places in the clause, i.e. in adverbial groups, which means that they are a kind of *not*-negation rather than *no*-negation. Examples 65 to 67 show the proximity to syntactic negation of the verbal group, and also to epistemic modality:

(65) *Well_AV0 ,_, he_PNP 's_VBZ never_AV0 said_VVN to_PRP me_PNP that_CJT he_PNP loves_VVZ me_PNP ,_,*

(66) *I_PNP did_VDD have_VHI a_AT0 bout_NN1 of_PRF anorexia_NN1 as_PRP a_AT0 teenager_NN1 ,_, but_CJC that_DT0 is_VBZ no_AV021 longer_AV022 the_AT0 case_NN1 ._.*

(67) *I_PNP 'm_VBB by_AV031 no_AV032 means_AV033 overweight_AJ0 ,_,*

PoS-tagging is an adequate method to quickly find all clauses with any sort of syntactic negation. Apart from syntactic negation, however, there are other language options to express negation in a language.

5.5.3 Morphological negation in English

Apart from the synthetic negation discussed above, where a negation marker is added to the clause or phrase, there is another way to negate a phrase, or, more strictly speaking, a word, and that is the morphological negation. Quirk et al. (1985, 1540) mention five negative prefixes, namely *a-, dis-, in-, non-* and *un-*. The prefix *a-* adds the meaning of 'lack of' to a word, the other four add the meaning of 'not'. Morphological negation is also mentioned in the LGSWE (Biber et al. 1999, 531): "[…] adjectives can be derived from other adjectives, especially by the negative prefixes *un-, in-* and *non-* (e.g. *unhappy, insensitive, nonstandard*)."

The negative prefixes *un-, in-, il-, im-* and *dis-* can all be found in the EN corpus, but there is no instance of a word with the prefix *non-* or *a-.* There are, however, some words with a negative suffix, i.e. *–less.* These negating affixes can be found in all lexical word classes, see examples 68 to 76 below.

Predicative adjectives:

(68) Am I being impatient to expect him to […]

(69) I think my spouse is being unfaithful.

(70) […] or is it hopeless?

Adverbs:

(71) It is unfortunately a bit unsettling at times.

Verbs:

(72) […] which I disagreed with.

(73) It would make those feelings disappear.

Attributive Adjectives:

(74) *The eating is followed by an even more extreme anxiety and a ruthless, usually violent purge.*

(75) *I am now at an extremely uncomfortable 160.*

Nouns:

(76) *There is a stigma attached to men with eating disorders.*

The noun *eating disorder* was the only instance of a noun with the prefix *dis-* in EDNA. It has not been annotated as a morphologically negated noun on the grounds that there is no opposite, positive term like **eating order*. *Eating disorder* is a proper noun in its own right, rather than the negation of another noun. In EDNA, we did not annotate those words that do have a negative affix, but where leaving the affix away does not render a positive word, e.g. *disappointing, *appointing,* or *disgusting, * gusting*.

Prepositions:

(77) *The longest I went without eating between binges would range from two days to a week.*

(78) *So I ate and ate and without thinking about it, I threw up my food.*

(79) *She has friends and wants to do other things without me from time to time.*

Of course, prepositions do not belong to the class of lexical words, but for the sake of convenience they were added to the annotation scheme together with the other lexical words that can be negated morphologically.

5.5.4 Negation markers in the textual metafunction in EN

With the help of the PoS-tagging, another two locations for negation markers were found. The first is the negation *whether or not*, which is a conjunction and needs to be related to the textual metafunction. There was only one instance in the EDNA corpus, though, see example 80:

(80) *(…] our main distancing difference being whether or not to have an open relationship.*

The second additional location, or word class, for a negation marker is as a continuative, or interjection, see examples 81 and 82. These two are the only instances in the English newsgroup texts.

(81) *I don´t know who to tell, my parents? HELL* *NO!*

(82) *[…] and he said that* *no, he didn't love me.*

Figure 5.7 displays the annotation scheme for negation, which is identical for English and German.

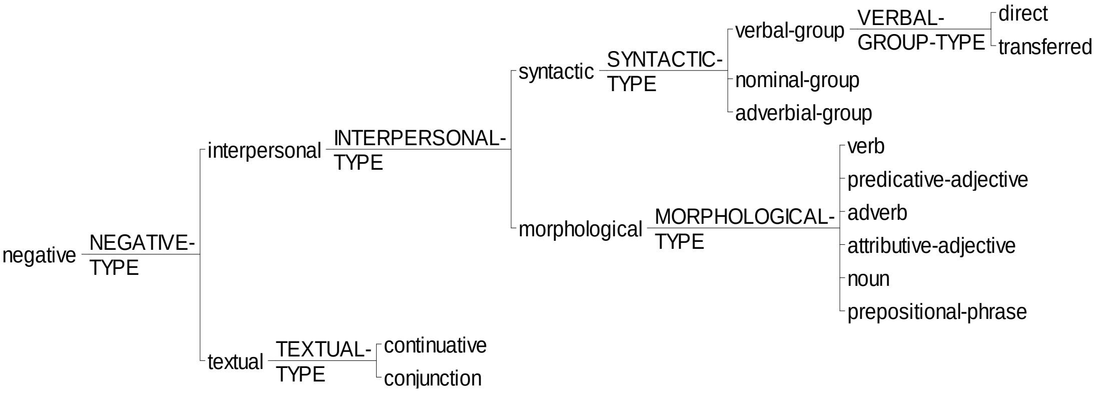

Figure 5.7 The annotation scheme for negation in English and German

5.5.5 Syntactic negation in German

The annotation scheme for negation in English can be transferred without changes to the annotation of the German part of the EDNA corpus. The German equivalent for a syntactic negation marker in the verbal group is the negative polarity marker *nicht* (Götze and Hess-Lüttich 1999, 920). See some examples for a direct syntactic negation of the verbal group in examples 83 and 84.

(83) *Das ganze klappt nicht so, wie ich will.*

(84) *Ich will meinen Schatz nicht einengen.*

The German language, like the English, does allow the negation marker *nicht* to be transferred from the subordinate to the superordinate clause, see example 85 for a transferred syntactic negation.

(85) *Ich glaube nicht, dass das stimmt.*

The second main type of syntactic negation, where the negation marker is down-shifted to the nominal group, is realized in German with a form of *kein*, see the following examples:

(86) *[…] obwohl es dazu rein esstechnisch keinen Grund gibt.*

(87) *Wir hatten so gut wie keinen Kontakt mehr.*

Furthermore, indefinite pronouns like *kein, keiner, keins, nichts, niemand, nirgends* can express negation inside nominal groups, as shown in examples 88 and 89.

(88) *Auf jeden Fall glaubt mir keiner.*

(89) *Ich esse fast gar nichts.*

As a third type of syntactic negation there is the negative polarity marker *nicht* inside adverbial groups, see examples 90 and 91.

(90) *Ich will sie mein ganzes Leben lang nicht mehr sehen.*

(91) *Ich bin noch nicht über dich hinweg.*

Surprisingly, out of the 30 instances of syntactic negation inside an adverbial group, 28 were the adverbial group *nicht mehr*, and only two were *noch nicht*. There is a striking difference between the frequencies of the adverbial group *nicht mehr* and the English equivalent *not anymore*, which occurs only 10 times, plus one instance of *no longer*. My first suspicion was that in the English news-groups, the writers would use the influential process *to stop* to express 'not anymore', but this could not be verified. *To stop* occurs only 6 times, plus one instance of *to give up*. There are just as many, i.e. six, influential processes of *aufhören* in the German corpus. Is it true that the German writers just like the concept of *nicht mehr*, and is there truly no alternative way to express that in English? These questions may be the focus of future studies.

5.5.6 Method for finding all syntactic negation markers in GN

The German texts have also been tagged for part-of-speech to make it easier to find all syntactic negation markers in the 10,000 word corpus. The German texts were tagged using the Stuttgart-Tübingen Tag Set (STTS). The BNC C5 tag set and the STTS are comparable, although the tags are not completely identical. In both languages, the tags revealed all 'locations' of syntactic negation markers, which are, as shown above and below, not limited to verbal and nominal groups; see examples 92 to 94 from the German newsgroup texts.

Verbal group

(92) *Dieses_PDAT Gefühl_NN ,_$, geliebt_VVPP zu_PTKZU werden_VAINF ,_$, bekomme_VVFIN ich_PPER nicht_PTKNEG ._$.*

Nominal group

(93) *Ich_PPER bekomme_VVFIN nur_ADV noch_ADV flüchtige_ADJA Küsse_NN ,_$, keine_PIAT zärtlichen_ADJA Küsse_NN mehr_ADV ,_$, keine_PIAT Umarmungen_NN oder_KON Kuscheln_NN ,_$,*

Adverbial group

(94) *Ich_PPER will_VMFIN sie_PPER mein_PPOSAT ganzes_ADJA Leben_NN lang_APPO nicht_PTKNEG mehr_ADV sehen_VVINF ._$.*

5.5.7 Morphological negation in German

Similar to the English newsgroup texts, the German newsgroup text writers occasionally use an affix to negate a word. The negative prefixes in German include *a-*, *un-*, *miss-*, *des-* and *dis-*, of which only *un-* is found in EDNA. Additionally, there is one instance of the suffix *–los*. The preposition *ohne* has been annotated as morphological negation in a prepositional phrase, analogous to the English texts, to ensure comparability. Words with a negative affix which were not the opposite of the same word without the affix have not been annotated, e.g. *Diskussion, *Kussion*. In the German EDNA subcorpus, negative affixes are only found on predicative and attributive adjectives and on adverbs, but not on verbs or nouns, see examples 95 to 99.

Predicative adjective

(95) *Er weiß, dass das Geschehene für mich untragbar ist.*

(96) *Ich bin ziemlich ratlos.*

Adverb

(97) *[…], aber ich zeige mich ungern in der Öffentlichkeit.*

Attributive adjective

(98) *Hinzu kam, dass ich seit genau diesem Tag eine neue ungewohnte, stärkere und engere Brille habe.*

Preposition

(99) *Ich habe das aber ohne Therapie einigermaßen wieder in den Griff bekommen.*

5.5.8 Negation markers in the textual metafunction in GN

The two types of negation mentioned above, syntactic and morphological, realize aspects of the interpersonal metafunction. With the help of PoS-tags, however, it is easy to see that *nein, ohne* and some equivalents also realize the textual metafunction, as continuatives (interjections) and conjunctions, see examples 100 to 103.

(100) Continuative / Interjection: *Nicht nur irgendein One-Night-Stand, nein, ein Verhältnis in etwa 180 km Entfernung.*

(101) Continuative / Interjection: *Er sagte nein, das will er nicht.*

(102) Conjunction: *[…], ohne sich den ganzen Tag Gedanken darum zu*

machen.

(103) Conjunction: *Ich bin weder bei mir, noch bei meiner Freundin zuhause.*

After finding these three types of negation, we turn to the quantitative and qualitative analysis of negation in the EDNA corpus in the coming sections, but before that, multiple negation must be mentioned.

5.5.9 Multiple negation in EDNA

This term traditionally describes clauses with two syntactic negation markers neutralizing each other, e.g. *I have never not given him anything*. In fact, this example is the only instance of a classical multiple negation in the corpus. Other combinations, however, are also possible. Consider the invented example *it does not seem to be hopeless though.* Here, a syntactical negation combines with a morphological negation. A look into the two corpora, however, reveals that the phenomenon of combining two negations in one clause is a very rare one, and therefore does not allow us to draw any conclusions about it due to lack of data. See the following examples 104 to 108, which display all the clauses from the EDNA corpus that do include more than one instance of a negation.

EN, syntactic negation marker verbal group and adverbial group: one instance

(104) *I have never not given him anything.*

EN, syntactic negation marker verbal group and nominal group: one instance

(105) *After a month of eating nothing but salads and fruits... my tummy can't seem to handle meat anymore.*

EN, syntactic negation marker verbal group and morphological negation prepositional phrase: two instances

(106) *If I am going to eat, I am going to eat until I cannot shove one more crumb into my body without bursting.*

(107) *I don't want to visit my dad without being able to say I've lost another 30 pounds since the last time I've seen him.*

EN, one instance of where a clause carries a double negation, signalling uncertainty

(108) *I'm not quite sure or not if it's a major problem or not.*

With only five clauses that contain more than one negation marker, I assumed that the total number of positive clauses can be calculated by subtracting the number of negation markers from the total number of clauses. In other words, no clause carries two negation markers, as opposed to modal markers, where we find more than one in some clauses.

In GN, there are only two instances of a clause with two negation markers; both clauses carry two syntactic negation markers in the nominal group. The first clause (example 109) is a logical mistake made by the writer, and the second (example 110) is a simple coordination of two negated nominal groups. In both examples, the two negation markers do not 'eliminate' each other.

(109) *Ich finde einfach kein Mittelmaß zwischen hungern und nichts essen.*

(110) *Das gemeinsame Lebensmodel sah keine Heirat und keine Kinder vor.*

As with EN, the total number of positive clauses was calculated by subtracting the number of negated clauses from the total number.

5.6 Quantitative analysis of negation

5.6.1 Positive and negative clauses

The study of multiple negation in the EDNA corpus has shown that there are virtually no clauses with more than one instance of negation, therefore it is assumed that the number of positive clauses (in the sense of lacking a negation marker) can be calculated from subtracting the number of clauses with a negation marker from the total number of clauses. Polarity is what Halliday and

James (1993) call a skew system, stating that "grammatical systems fell largely into two types: those where the options were equally probable – there being no 'unmarked term', in the quantitative sense; and those where the options were skew, one term being unmarked" (Halliday and James 1993, 35). Their hypothesis is that in a skew system, the ratio of unmarked and marked option is 0.9 : 0.1, thus, 10% of all finite clauses would have negative polarity, the marked option. In their study, Halliday and James (1993, 60) find 12.4% negative clauses (negation in the verbal group with *no*) in the COBUILD corpus of written texts.

The total number of clauses in the study at hand is based on the number of process types, not rhemes (see explanation in 5.3). Table 5.17 displays the raw and relative numbers and the result of the chi-squared test. The null hypothesis is that there is no significant variation in the frequency of negation across the two corpora.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Negative clause	238	300	15	20	GN	9.97	+ +
Positive clause	1,304	1,179	85	80	EN	2.16	-
Column total	1,542	1,479	100	100			

Table 5.17 Raw and relative numbers, χ^2 and significance for positive / negative clauses

20% of all clauses in the German newsgroup texts carry a negation marker, whereas in the English newsgroup texts, only 15% of all clauses have a negation marker. The result is significant at the level of $p < 0.01$, with the threshold at 6.64 (df=1). The null hypothesis can be rejected, as there is a difference in the frequency of negation markers which is not due to chance with 99.9% certainty.

It may seem illogical in a system with mutually exclusive features, i.e. a clause can be positive or negative but not both, to have significant variation in the frequency of negative clauses but not positive clauses. The reason lies in the total amount of each feature. The total number of negative clauses is small, and even a small divergence from the expected frequency has an impact. The total number of positive clauses is much higher, therefore, even if the divergence is as large (or small) as the one in the number of negative clauses, it results in only a small divergence from the expected frequency for the positive clauses and is therefore not significant.

The LGSWE (Biber et al. 1999), unfortunately, does not say anything about the ratio of positive and negative clauses. Syntactic *not*-negation, however, is only a fraction of the full picture of negation. Halliday and James (1993, 60) concede "that a significant number of clauses containing negative words, such as *never, nobody, hardly*, should be interpreted as negative clauses". Consequently, the scheme for the annotation of negative polarity in the EDNA corpus also includes other types of negation in addition to negation in the verbal group.

The next two features in the annotation scheme for negation are the negation related to the interpersonal metafunction (syntactic and morphological negation) and the negation related to the textual metafunction (interjections and conjunctions). The null hypothesis states that there is no difference in frequency between the English and German corpus, see table 5.18 for the calculation of statistical significance.

Feature	**EN F**	**GN F**	**EN%**	**GN%**
Interpersonal negation	235	296	99	99
Textual negation	3	4	1	1
Column total	238	300	100	100

Table 5.18 Raw and relative numbers for negation related to interpersonal or textual metafunction

In the two corpora, the numbers for negation related to the interpersonal and textual metafunction are in fact too small to allow a calculation of statistical significance. No test of statistical significance was carried out for the difference in frequency between continuative and conjunction either due to numbers being as small as they are. The total numbers for syntactic and morphological negation in EDNA are slightly higher, thus the statistical significance can be calculated, see table 5.19. The null hypothesis states that there is no variation between the English and the German texts at a significant level.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Syntactic negation	209	279	89	94	GN	0.40	-
Morphological negation	26	17	11	6	EN	4.58	+
Column total	235	296	100	100			

Table 5.19 Raw and relative numbers, χ^2 and significance for syntactic and morphological negation

The χ^2 value reveals that in the English corpus, there are significantly more instances of morphological negation, compared to the German corpus, with a threshold of 3.84 for $p < 0.05$ (df=1). The null hypothesis can be rejected. The English writers use morphological negation more often than the German writers.

5.6.2 Syntactic negation markers

Since the syntactic negation marker is the most common means to negate a statement, where can it be found most often, in the verbal, the nominal or the adverbial group? The null hypothesis will be that there is no difference in the frequency of use between the English and German part of the EDNA corpus, see table 5.20 below for the results.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Verbal group	176	169	84	61	EN	9.44	++
Nominal group	15	71	7	25	GN	22.63	+ + +
Adverbial group	18	39	9	14	GN	2.95	-
Column total	209	279	100	100			

Table 5.20 Raw and relative numbers, χ^2 and significance for position of syntactic negation marker

The results in table 5.20 show that it is more common in the English texts to negate the verbal group than it is in the German texts, at a highly significant level, with the threshold at 9.21 for $p < 0.01$ (df=2). In the German texts, there are significantly more syntactic negation markers in a nominal group, with the threshold for $p < 0.001$ at 13.82 (df=2), i.e. the negation is down-shifted to a lower rank, 'hidden' more deeply in the clause structure. For the negation inside adverbial groups, the observed frequencies did not differ to a significant degree from the expected frequencies.

Biber et al. (1999) base their corpus findings of *not* / *n't* versus other negative forms on occurrences per million words. It is not clear what the authors of the LGSWE mean by 'other negative forms'. The following table thus shows only the occurrences of *not* / *n't* in the four registers described in the LGSWE, compared to the English part of EDNA.

Register	**Occurrences per million words**
LGSWE Conversation	19,500
LGSWE Fiction	9,500
LGSWE News	4,500
LGSWE Academic	3,500
EDNA Newsgroup texts	17,000

Table 5.21 Occurrences of *not*-negation per register in the LGSWE (Biber et al. 1999, 159) and EDNA corpus

It becomes clear that the English newsgroup texts in EDNA are more similar to spoken conversation than to one of the other registers, to judge by the number of the syntactic negation marker *not* alone. For comparison, the German newsgroup texts have 16,200 instances of *nicht* per million words.

5.6.3 Direct and transferred negation

In the annotation scheme for negation, the verbal group is an entry condition for two more options, i.e. direct and transferred negation. Table 5.22 gives the calculations for this system. The null hypothesis once more predicts no significant difference in the frequencies of direct or transferred negation between the two EDNA subcorpora.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Direct negation	164	153	93	90	EN	0.07	-
Transferred negation	12	16	7	10	GN	0.75	-
Column total	176	169	100	100			

Table 5.22 Raw and relative numbers, χ^2 and significance for direct / transferred negation of the verbal group

The phenomenon of transferring a syntactic negation from the verbal group of the subordinate clause to the superordinate clause is not a very frequent one.

The English writers did transfer 7% of the negation markers in the verbal group, whereas a transfer to the superordinate clause happens with 10% of all syntactic negation markers in the German texts. The χ^2 value, however, shows that the difference between the English and German newsgroup texts is not statistically significant.

5.6.4 Morphological negation marker

The last system in the annotation of negation in EDNA is the position of the morphological negation marker. The null hypothesis is that there is no difference between the two corpora at any significant level. See table 5.23 for results.

Feature	**EN F**	**GN F**	**EN%**	**GN%**
Verb	1	0	4	0
Predicative adjective	12	11	46	65
Adverb	1	1	4	6
Attributive adjective	2	1	8	6
Noun	0	0	0	0
Prepositional phrase	10	4	38	24
Column total	26	17	100	100

Table 5.23 Raw and relative numbers, χ^2 and significance for position of morphological negation marker

The total numbers for morphological negation markers are rather small. The table shows that in the German texts, there are more predicative adjectives with a negative affix than in the English texts, and that in the English texts, the preposition *without* is used more often than the German equivalent *ohne*. These two, predicative adjectives and prepositional phrases, are more frequent than the other options in both subcorpora. The χ^2 value, however, cannot be calculated because 8 out of the 12 cells of expected frequencies have a value < 5, which distorts the calculation (Gries 2008, 157). Consequently, the null hypothesis can neither be verified nor rejected.

5.6.5 Summary

Figure 5.8 below displays the most prominent results from the tests for statistical significance for the system of negation.

System	**Divergence**	**Divergent feature**	**Divergent corpus**
Negative / positive clause	Significant	Negative clause	GN + +
Interpersonal / textual negation	-	-	-
Syntactic / morphological negation	Significant	Morphological negation	EN +
Position of syntactic negation	Significant Significant	Verbal group Nominal group	EN + + GN + + +
Direct / transferred negation	-	-	-
Position of morphological negation	-	-	-

Figure 5.8 Summary of tests of statistical significance of the system of negation

The German corpus has more clauses with a negation marker, compared to the English corpus. The syntactic negation marker is more often down-shifted to a nominal group in the German texts, thereby making the negation less obvious, less negotiable. The English newsgroup writers prefer negation in the verbal group, and they tended to use more morphological negation markers than the German writers.

5.7 Analysis of lexical items used to express negative polarity

5.7.1 Negation markers in adverbial and nominal groups

In the following sections, the lexical items that realize syntactic negation markers inside adverbial and nominal groups will be investigated in more detail. This is followed by a study of transferred negation and finally of morphological negation markers. The negation of the verbal group is always realized by the syntactic negation marker *not*, or **n't* in EN and *nicht* in GN, therefore we need not study that any further.

Let us begin by looking at the adverbial groups containing a negation marker. Table 5.24 shows all of those adverbial groups in order of frequency.

N	Word EN	F	%	Word GN	F	%
1	NEVER	16	88.89	NICHTMEHR	27	69.23
2	BYNOMEANS	1	5.56	NIE	10	25.64
3	NOLONGER	1	5.56	NIEMALS	1	2.56
				NOCHNICHT	1	2.56
	Total	18	100	Total	39	100

Table 5.24 Syntactic negation marker inside adverbial groups in EN and GN

The table shows us that in the English newsgroup texts, *never* is the most frequent adverbial group expressing negation. The German equivalents, *nie* and *niemals*, are less frequent, and also not the most frequently used adverbial group expressing negation in the German newsgroup texts. The German writers use *nicht mehr* more often, thereby expressing that a state of affairs has ended. This concept is not expressed by the English writers, at least not with an adverbial group; there is only one instance of *no longer*.

Table 5.25 shows the frequency of different syntactic negation markers inside nominal groups. Note that in the list of nominal groups in the German newsgroup texts, *keine, keine mehr* and *niemand* stand for all instances of the word which have been stripped of case marking (e.g. *keine* stands for *kein, keine, keiner, keinen, keinem*).

N	Word EN	F	%	Word GN	F	%
1	NO	6	40.00	KEINE	44	61.97
2	NOTHING	5	33.33	KEINEMEHR	9	12.68
3	NOONE	2	13.33	NICHTS	8	11.27
4	NOTALLFATPEOPLE	1	6.67	NIEMAND	6	8.45
5	NOTEVERYONE	1	6.67	NICHTSMEHR	2	2.82
				NIX	2	2.82
	Total	15	100	Total	71	100

Table 5.25 Syntactic negation marker inside nominal groups in EN and GN

In EN, there are far less syntactic negation markers forming part of the nominal group. The most frequent one is *no*, followed by nouns like for example *problem, relationship, chance*. In GN, the most frequent negation markers in a nominal group are the forms of *keine*, followed by nouns, e.g. *Mittelmaß, Ende, Nahrung, Beziehung, Lösung*. What is striking, though, is the use of *keine mehr* (9 times) and *nichts mehr* (2 times). As with the adverbial group *nicht mehr* (27 times), the German writers describe the ending of a state of affairs, i.e. something that was once there is no longer. The English equivalent *not anymore* occurs 11 times in connection with the syntactic negation of the verbal group, but not in adverbial or nominal groups. *No longer* occurs once, and there are no occurrences of *no more*.

Another interesting aspect is the appearance of the syntactic negation marker *kein* inside the nominal group *auf keinen Fall*. *Auf keinen Fall* by itself has been annotated as adverbial phrase expressing epistemic modality. As can be seen in the examples 113 and 114 below, *auf keinen Fall* expresses epistemic modality and negation at the same time, its positive counterpart, *auf jeden Fall*, also expresses epistemic modality and can appear in clauses where there is also a syntactic negation, see examples 111 and 112.

(111) *Und ich hab irgendwie keine Kraft mehr oder auf jeden Fall weniger*

(112) *Auf jeden Fall glaubt mir keiner*

(113) *Verlieren will ich Mars aber auf keinen Fall*

(114) *[…], dass er nach einer Trennung auf keinem Fall eine Freundschaft will bzw. haben kann mit mir*

It would be interesting to investigate what triggers the use of *auf keinen Fall* and *auf jeden Fall nicht*, but it cannot be done within the present study.

5.7.2 Transferred syntactic negation

The term grammatical metaphor has been explained in section 5.4.5. It refers to clause complexes where the superordinate clause realizes a mental process which functions as a grammatical metaphor of modality, adding an evaluation of likelihood to the subordinated clause, see examples 115 and 116. In these examples, the negation marker is part of the subordinate clause.

(115) *Lately I feel like I can't be productive if I am full.*

(116) *[...] und deshalb denke ich mir, dass jetzt nicht der richtige Zeitpunkt ist.*

Transferred syntactic negation describes the process of shifting the syntactic negation of the verbal group from the subordinate to the superordinate clause, see examples 117 and 118. Thus, instead of saying *I may or may not be qualified as having an eating disorder*, using *may* as modal marker and the syntactic negation marker in the subordinate clause, the writer negates the superordinate clause, the grammatical metaphor. Example 118 shows a German clause with transferred syntactic negation. Instead of saying *Ich bin vielleicht essgestört*, where epistemic modality is expressed with *vielleicht*, the writer negates the superordinate clause. This has the same effect; she expresses uncertainty.

(117) *I'm not sure if I am qualified as having an eating disorder.*

(118) *Ich weiß nicht, ob ich essgestört bin oder nicht.*

Table 5.26 displays the most frequent superordinate clauses that contain transferred negation.

N	**Word EN**	F	%	**Word GN**	F	%
1	IMNOTSURE	5	42.00	ICHWEIßNICHT	13	81.00
2	IDONTKNOW	3	25.00	ICHBINNICHTSICHER	2	13
3	IDONTTHINK	2	17.00	ICHHABEKEINEAHNUNG	1	6.00
4	ICANTBELIEVE	1	8.00			
5	ITDOESNTAPPEAR	1	8.00			
	Total	12	100	Total	16	100

Table 5.26 Superordinate clauses containing transferred negation

In EN, there are 67 clauses with a mental process that have been annotated as grammatical metaphor of modality. Out of these 67 clauses, 12 are negated, i.e. 18%. In GN, there are 55 clauses with a mental process that have been annotated as grammatical metaphor. Out of these, 16 are negated, i.e. 29%. Transferred negation seems to be slightly more popular among the German newsgroup writers. The variation of lexical verbs in the negated superordinate clauses, however, is greater in the English texts, where we find *be sure, know, think, be-*

lieve and *it appears*. In the German texts, there is only the most frequent lexical verb, *wissen*, and synonyms of *wissen*, i.e. *sicher sein* and *Ahnung haben*.

5.7.3 Morphological negation markers

Morphological negation is most frequently found with predicative adjectives. In EN, there are three negative morphemes, i.e. *–less*, *im-* and *un-*, see examples below. In GN, there are only two negative morphemes, i.e. *-los* and *un-*, see examples below.

-less: *needless* (2x), *hopeless*
Im-: *impossible* (2x), *impatient*
Un-: *unsure, uncaring, unfaithful, unfair, unsettling, uncomfortable*

-los: *lustloser, nutzlos, ratlos, hilflos*
Un-: *unsicher* (2x), *unbeweglicher, unglaublich, untragbar, unwichtig, unglücklich*

The numbers of predicative adjectives in EN (12) and GN (11), however, are too small to lend themselves to any meaningful interpretation of the findings.

The preposition *without* is used 10 times as a negation marker in EN. *Without* occurs at the beginning of a non-finite clause, or as part of a prepositional phrase, see examples 119 to 122 below.

Without introducing a non-finite clause

(119) […] and without thinking about it, I threw up my food

(120) I don't want to visit my dad without being able to say […]

Without as part of a prepositional group

(121) I knew that without clothes I still had bulges of fat

(122) Without a regular eating routine I was having up to 3 seizures a day

In the German newsgroup texts, the preposition *ohne* (as the equivalent of *without*) is used only 4 times as part of a prepositional group, thereby expressing the absence of something, see examples 123 to 126.

(123) Ich habe das aber ohne Therapie einigermaßen wieder in den Griff be-

kommen

(124) *[…], oder ich stopfe mich ohne Ende voll*

(125) *Dann machte er ohne ersichtlichen Grund von einem Tag auf den nächsten Schluss*

(126) *Und dies 13 Jahre alleine, d.h. ohne Vater*

The English newsgroup text writers use the preposition *without* more frequently than the German writers use the German equivalent *ohne*, but apart from that, the numbers are small. Morphological negation does not seem to be a favorite means to express negation, neither in EN nor in GN.

5.8 The combination of modality and negation

5.8.1 Modal and syntactic negation markers

From the analyses in chapters 5.3 and 5.6 on modality and negation we know that in the German newsgroup texts there are significantly more negation markers (significance level 0.01) and also significantly more modal markers (significance level 0.001). In the following section, the interplay of modal markers and syntactic negation markers will be investigated. The investigation involves only the syntactic negation markers because morphological negation markers and textual negation (conjunctions and continuatives) are too small in number. Furthermore, the query of the EDNA corpus is only for clauses which contain a first modal marker, i.e. we look at clauses with at least one modal marker, but do not specify whether there is also a second or third modal marker in a clause. With around 15%, the numbers for clauses with a second or third modal marker, however, are small. 85% of all clauses have only one modal marker. Table 5.27 displays, as a summary, the raw numbers for clauses with negation marker only, modal marker only, both markers, or neither. This table is the basis for the following calculations.

Feature	EN F	GN F
Negation marker only	129	141
Modal marker only	238	437
Both neg and modal marker	80	138
No neg or modal marker	1,095	763
Total (processes)	1,542	1,479

Table 5.27 Raw numbers for clauses with/without syntactic negation marker and modal marker

Let us start with the significance test for the frequency of clauses containing at least a first modal marker (type of modality and type of realization unspecified) and syntactic negation (position unspecified), see examples 127 to 129 in EN and 130 to 132 in GN:

(127) I still can't let go

(128) […] which is in truth not all that often

(129) We never really argued

(130) Das ist doch auch keine Lösung

(131) […], denn mir ging es ja nicht schlecht

(132) Ich weiß wirklich nicht mehr weiter

The null hypothesis states that there will be no significant variation in numbers, table 5.28 shows the results.

Feature	EN F	GN F	EN%	GN%	Overuse	χ^2	S
Modality + negation	80	138	25	24	-	0.11	-
Modality - negation	238	437	75	76	-	0.04	-
Column total	318	578	100	100			

Table 5.28 Raw and relative numbers, χ^2 and significance for clauses with modal marker (first modal only) and syntactic negation marker (any position)

In both EN and GN, a quarter of all clauses with a modal marker additionally include a syntactic negation marker, either in the verbal, the nominal or an adverbial group. There is no statistically significant deviation in numbers and the null hypothesis can therefore not be rejected.

In the second table (5.29) of this section, the results of the test for the query of non-modal clauses plus a syntactic negation (position unspecified) are shown, see example 133 from EN and 134 from GN:

(133) *I never had any physical pain*

(134) *Er ist nicht mein Traummann*

We use the default null hypothesis that there is no statistically significant variation in numbers when comparing the two subcorpora.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Non-modal + negation	129	141	8	10	-	1.16	-
Non-modal - negation	1,413	1,338	92	90	-	0.11	-
Column total	1,542	1,479	100	100			

Table 5.29 Raw and relative numbers, χ^2 and significance for clauses without a modal marker and syntactic negation

Again, there is no statistically significant variation in the frequency of clauses with only a syntactic negation marker, but without a modal marker. In both corpora, there are 8% / 10% of non-modal clauses with a syntactic negation marker in the clause, and 92% / 90% of all clauses have neither a modal marker nor a syntactic negation marker. The null hypothesis cannot be rejected.

The frequency of clauses with/without modal marker and/or syntactic negation marker, however, can also be looked at from another angle. Above in table 5.28, the column totals were for clauses with or without a modal marker (EN 318, GN 578) plus (or not) a syntactic negation marker. In table 5.29, however, the column totals are the number of clauses with or without a syntactic negation (EN 209, GN 279) plus (or not) a modal marker. Table 5.30 displays the results for the query of clauses with a syntactic negation (position unspecified) and at least a first modal marker (type of modality and type of realization unspecified). No statistically significant difference is predicted by the null hypothesis.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Negation + modality	80	138	38	49	-	3.35	-
Negation – modality	129	141	62	51	-	2.70	-
Column total	209	279	100	100			

Table 5.30 Raw and relative numbers, χ^2 and significance for clauses with syntactic negation marker (any position) and modal marker (first modal only)

In our third possible combination of features, there is no statistically significant difference between the two newsgroup text corpora. What is interesting to note is that of all clauses with a modal marker, only 25% are negated (see table 5.28 above), but of all clauses with a syntactic negation, as many as almost 40% in EN and as many as almost 50% in GN additionally carry a modal marker. These results suggest that the writers have a tendency to either strengthen or weaken negated statements, to indicate that the negated statement is not 100% unquestionably true (epistemic modality is expressed by 70% of all modal markers). It would be interesting to know whether the modal markers were strengthening ones, like *really, must* or *will* for English, or whether they were weakening the negated statement, like the modal markers *might* or *maybe* in English. Unfortunately, the modal markers in the EDNA corpus were not annotated for their strength, so this aspect cannot be examined in this study.

Allow me to repeat here that the German newsgroup texts have a significantly larger number of clauses with a negation marker (20%) than the English newsgroup texts (15%). The higher total number of negative clauses (in the sense of including a syntactic negation marker) in GN probably explains why GN has a higher number of clauses with both a syntactic negation marker and a modal marker (49%) compared to EN (38%).

The last study of statistical significance involves the frequency of positive clauses (in the sense of lacking a syntactic negation marker) and how many of these contain at least one modal marker. Examples for clauses with only a modal marker, but no syntactic negation marker, are *I rarely leave the house, my god you'll hate what you see, something needs to change* in the EN corpus and *für mich ist Essen meistens ein Qual und so überflüssig, ich soll doch ein Vorbild für meine Kinder sein,* […] *waren es ja bisher auch immer* for the GN corpus. The null hypothesis suggests that there is no statistically significant difference in the frequency of such clauses in both corpora; see table 5.31 below for raw and relative numbers and the χ^2 value calculation.

Feature	EN F	GN F	EN%	GN%	Overuse	χ^2	S
No negation + modality	238	437	18	36	GN	81.83	+ + +
No negation - modality	1,095	763	82	64	EN	29.71	+ + +
Column total	1,333	1,200	100	100			

Table 5.31 Raw and relative numbers, χ^2 and significance for clauses without syntactic negation marker but with modal marker (first modal only)

It probably comes as no surprise to see that modal markers occur most frequently in positive clauses, since positive clauses (85% in EN and 80% in GN) are far more frequent than negative ones, i.e. clauses with a negation marker. The null hypothesis can safely be rejected, there is a statistically significant deviation in the number of clauses with a modal marker. This is not surprising, either; although GN has fewer positive clauses in total compared to EN, GN has a much higher number of modal markers, and thus the modal markers appear significantly more often in positive clauses in GN, with 10.83 being the threshold for a significance of 0.001 (df=1). Due to the fact that in EN there are significantly fewer modal markers, a significant amount of positive clauses carry no modal marker, compared to GN (again, 10.83 being the threshold for a significance of 0.001 (df=1)).

5.8.2 Type of modality and syntactic negation

We now turn to the interplay of type of modality, i.e. epistemic or root modality, and syntactic negation markers. Later on we will look in more detail at the types of root modality, i.e. obligation and permission, inclination and ability, in combination with a syntactic negation marker. As in the previous section, morphological and textual negation will be disregarded due to the small numbers of these features.

We begin with the types of modality. Remember from chapter 5.3 that 70% of all modal markers in both corpora express epistemic modality and the other 30% of modal markers express root modality; the variation in numbers between EN and GN is not statistically significant. In GN, however, there is a larger number of syntactic negation markers, which is statistically significant at a significance level of 0.01. Does the larger number of syntactic negation markers lead to a significant difference in the frequency of clauses combining both one type of modality and a syntactic negation marker? The null hypothesis states that this is not the case, in both corpora, epistemic and root modality will occur in a clause with syntactic negation markers to the same extent. Note that the numbers of modal markers expressing epistemic or root modality have not been checked for whether they are the first, second or third modal marker in a clause, therefore, the column total is of instances of modal markers, not number of clauses carrying one. However, less than 11% of all instances are second or third modal markers in EN and less than 17% are second or third modal marker in GN. This lead me to the conclusion that this feature is of minor importance and that a detailed study would cost more in terms of time spent on it than it would gain us in terms of knowledge. Tables 5.32 and 5.33 display the results.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Epistemic + negation	41	103	16	21	-	2.29	-
Epistemic - negation	214	382	84	79	-	0.55	-
Column total	255	485	100	100			

Table 5.32 Raw and relative numbers, χ^2 and significance for instances expressing epistemic modality plus a syntactic negation marker (any position) in same clause

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Root + negation	45	63	41	31	-	2.08	-
Root - negation	64	139	59	69	-	1.11	-
Column total	109	202	100	100			

Table 5.33 Raw and relative numbers, χ^2 and significance for instances expressing root modality plus a syntactic negation marker (any position) in same clause

It is somewhat unexpected to see that in both EN and GN, root modality combines with a syntactic negation marker more often than epistemic modality does. Furthermore, a higher percentage of modal markers expressing epistemic modality combine with a syntactic negation in the German newsgroup texts, whereas in the English newsgroup texts, modal markers expressing root modality combine with a syntactic negation marker more often. The deviation in frequencies is not statistically significant, though.

In the next paragraph, we will look at the combination of modal markers expressing different types of root modality and any type of syntactic negation markers. In the case of root modality, the modal marker is always a modal auxiliary. For this reason, there are no clauses with two or even three modal markers expressing root modality. The null hypothesis predicts no divergence in the frequencies across the three types of root modality; tables 5.34, 5.35 and 5.36 show the results.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Obligation + negation	14	16	28	19	EN	1.14	-
Obligation - negation	37	70	72	81	GN	0.32	-
Column total	51	86	100	100			

Table 5.34 Raw and relative numbers, χ^2 and significance for instances expressing obligation and permission plus a syntactic negation marker (any position)

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Inclination + negation	7	25	37	34	EN	0.04	-
Inclination - negation	12	49	63	66	GN	0.02	-
Column total	19	74	100	100			

Table 5.35 Raw and relative numbers, χ^2 and significance for instances expressing inclination plus a syntactic negation marker (any position)

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Ability + negation	25	22	64	52	EN	0.48	-
Ability - negation	14	20	36	48	GN	0.66	-
Column total	39	42	100	100			

Table 5.36 Raw and relative numbers, χ^2 and significance for instances expressing root modality plus a syntactic negation marker (any position)

The calculation of the χ^2 values reveals no statistically significant variation in frequency of combining root modal markers and syntactic negation in EN and GN; the null hypothesis cannot be rejected. But this is not the only conclusion we can draw from the results. From table 5.4 we learned that in both corpora, writers express obligation more or less to the same extent (EN 47% and GN

43% of all root modal markers), but the German authors express inclination significantly more often (significance level 0.05) than their English counterparts (EN 17%, GN 36% of all root modal markers). The English writers, on the other hand, express ability more often than the German writers do (significance level 0.05, EN 36%, GN 21% of all root modal markers). But although in GN there are significantly more clauses with a syntactic negation marker compared to EN, the English newsgroup text writers negate each single type of root modality more frequently than their German counterparts. The English writers do express ability much more frequently than the German writers, but 64% of the clauses are in fact negated. The English writers do not say they can do something, but instead say that they cannot do something.

5.8.3 Type of realization of modality and syntactic negation

In the next set of calculations of raw and relative numbers and χ^2 values for the frequencies of occurrence, shown in tables 5.37 to 5.41, we look at the different ways of realizing modality, i.e. modal auxiliaries, modal adjuncts and grammatical metaphor in EN and GN, and also modal particles and subjunctive verb forms in GN and how they combine in a clause with any kind of syntactic negation markers, i.e. syntactic negation inside the verbal, nominal or adverbial group. The study of modality realized by means of grammatical metaphor in combination with a syntactic negation in the superordinate clause has been carried out already in table 5.6.3 on transferred negation above. It is included here once more for the sake of a comprehensive overview. The null hypothesis is that there are no meaningful differences in the frequencies of occurrence in the two EDNA corpora. The chi-squared value cannot be calculated for modal particles or subjunctive verb forms plus syntactic negation because in EN, there are no instances of those.

Feature	EN F	GN F	EN%	GN%	Overuse	χ^2	S
Auxiliary + negation	60	72	30	28	-	0.18	-
Auxiliary - negation	138	184	70	72	-	0.07	-
Column total	198	256	100	100			

Table 5.37 Raw and relative numbers, χ^2 and significance for instances of modal auxiliaries plus a syntactic negation marker (any position) in one clause

Feature	EN F	GN F	EN%	GN%	Overuse	χ^2	S
Adjunct + negation	14	22	14	13	-	0.09	-
Adjunct - negation	85	150	86	87	-	0.01	-
Column total	99	172	100	100			

Table 5.38 Raw and relative numbers, χ^2 and significance for instances of modal adjuncts plus a syntactic negation marker (any position) in one clause

Feature	EN F	GN F	EN%	GN%	Overuse	χ^2	S
Metaphor + negation	12	16	18	29	GN	1.65	-
Metaphor - negation	55	39	82	71	EN	0.49	-
Column total	67	55	100	100			

Table 5.39 Raw and relative numbers, χ^2 and significance for instances of grammatical metaphors of modality plus a syntactic negation marker (any position) in the superordinate clause

Feature	EN F	GN F	EN%	GN%
Particle + negation	0	55	0	31
Particle - negation	0	123	0	69
Column total	0	178	0	100

Table 5.40 Raw and relative numbers for instances of modal particles plus a syntactic negation marker (any position) in one clause in GN

Feature	EN F	GN F	EN%	GN%
Subjunctive + negation	0	1	0	4
Subjunctive - negation	0	25	0	96
Column total	0	26	0	100

Table 5.41 Raw and relative numbers for instances of subjunctive verb forms plus a syntactic negation marker (any position) in one clause in GN

On the basis of the calculations in the tables above, the null hypothesis cannot be rejected; in both the English and the German newsgroup texts, the various types of realizing modality tend to occur to the same extent in clauses which also carry a syntactic negation marker. One interesting finding is that in both corpora, modal auxiliaries and grammatical metaphors occur together with a syntactic negation marker around 30% of the time (only 18% of negated grammatical metaphors in EN though). In GN, the modal particles coincide with syntactic negation markers in 31% of all instances also (where two or three instances may occasionally be found in one clause, but not frequently). In contrast to the other types, however, modal adjuncts, i.e. adverbial phrases expressing modality, combine with a syntactic negation marker only half as frequently; only 14% in EN and 13% in GN can be found in clauses with syntactic negation markers. This is consistent in both corpora. Subjunctive verb forms in the German newsgroup texts seem to repel syntactic negation even stronger, only one of the 26 clauses (4%) with a subjunctive verb form is syntactically negated, but the total number is probably too small to allow generalization.

5.8.4 Position of syntactic negation markers and modal markers

The next step in the present study is to look at the combination of modal markers and syntactic negation markers from the other angle: how many syntactic negation markers inside verbal groups, inside nominal groups, inside adverbial groups combine with a modal marker in one clause? This time, the type of modality expressed by the modal marker and the type of realization is unspecified. The null hypothesis for all three possible positions suggests that the frequencies show no statistically significant differences between the English and German newsgroup text corpora; see tables 5.42 to 5.44 for the results of the calculations. The first table (5.42) gives the numbers for clauses that contain both a syntactic negation in a verbal group and a modal marker, see example 135 from EN and 136 from GN:

(135) *You're really not alone with those feelings*

(136) *Es ist vielleicht nicht gut für dich*

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Verbal group + modal	77	120	44	71	GN	11.22	+ + +
Verbal group - modal	99	49	56	29	EN	14.93	+ + +
Column total	176	169	100	100			

Table 5.42 Raw and relative numbers, χ^2 and significance for syntactic negation marker in verbal group and modal marker (type unspecified) in one clause

In table 5.20 in chapter 5.6.2 it was shown that syntactic negation in the verbal group is by far the most frequent position of syntactic negation markers in both corpora, and that there are more in the English newsgroup texts than in the German ones (significance level 0.01). When we look at clauses with both a syntactic negation marker and a modal marker, however, the results are somewhat surprising. Although in EN, there is more *not*-negation, these combine with a modal marker in only 44% of all clauses, whereas in GN, as many as 71% of all clauses with *not*-negation additionally carry a modal marker. The result is significant from the statistical point of view, with 10.83 as the threshold for $p < 0.001$ (df=1). Do these findings suggest that the English writers do not hesitate to put forward a negated statement, but that the German writers feel the need to modify almost three out of four negated statements? Which type of modal marker is added to a negated clause; one indicating epistemic modality, or one indicating root modality? The following table (5.43) gives the number of clauses combining *not*-negation with either epistemic or root modality, the null hypothesis says that there is no deviation in either of the corpora.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Verbal group + epistemic	35	75	45	63	GN	2.44	-
Verbal group + root	42	45	55	38	EN	3.09	-
Column total	77	120	100	100			

Table 5.43 Raw and relative numbers, χ^2 and significance for syntactic negation marker in verbal group and modal marker expressing epistemic or root modality

The calculations reveal no difference in the frequencies of any statistical significance. Nevertheless, in the English newsgroup texts, root modality is more frequent in combination with a syntactic negation marker of the verbal group than epistemic modality (also shown in tables 5.32 and 5.33), whereas in the German newsgroup texts this is the other way around, *not*-negation combines more frequently with epistemic modality. Could this phenomenon be interpreted in two ways? Is it true that the English writers feel that root modality should be negated, i.e. obligation, permission, inclination and ability, whereas the German writers feel that negations should be weakened (or strengthened) using epistemic modality? Such an interpretation must be dealt with cautiously, of course, because the total numbers are rather small. The assumption would need to be verified with data from a much larger corpus.

Now back to the second possible position for syntactic negation markers, the nominal group, and to how these are joined by modal markers in a clause. Examples for such clauses from the English newsgroup texts include […] *so no one can see my stomach; I will leave, no second chances*, and in the German texts, we find clauses like *ich bin doch keine 16 mehr, also ich wollte keine Freundschaft, und meistens habe ich keinen Hunger*. See the results for this analysis in table 5.44.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Nominal group + modal	2	27	13	38	GN	2.24	-
Nominal group - modal	13	44	87	62	EN	1.14	-
Column total	15	71	100	100			

Table 5.44 Raw and relative numbers, χ^2 and significance for syntactic negation marker in nominal group and modal marker (type unspecified) in one clause

None of the two corpora deviates in frequencies from the other to any statistically significant degree, although, as we know from table 5.20, GN has significantly more *no*-negation (i.e. syntactic negation inside a nominal group with 25% of all syntactic negation markers) than EN (7%). This certainly explains why there are more clauses with both *no*-negation and a modal marker in GN (38%), compared to EN (13%). In the end, the null hypothesis here was verified.

In the next table (5.45), the focus is on syntactic negation inside adverbial groups in combination with any modal marker (not specified). We find clauses like the following ones in EN: […] *that I would never fall in love again, I've never really been a member of a pro ED site before*. In GN, examples include *Ich will sie mein ganzes Leben lang nicht mehr sehen, […] weil ich das ganze einfach nicht mehr ertragen und aushalten kann; aber ob das Liebe ist, bin ich mir nicht mehr sicher*.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	**χ^2**	**S**
Adverbial group + modal	7	19	39	49	GN	0.26	-
Adverbial group - modal	11	20	61	51	EN	0.22	-
Column total	18	39	100	100			

Table 5.45 Raw and relative numbers, χ^2 and significance for syntactic negation marker in adverbial group and modal marker (type unspecified) in one clause

Again, the null hypothesis was verified, there is no significant difference in frequencies, and total numbers are small. This analysis does not readily offer any interpretation.

Summarizing, table 5.46 shows that in both EN and GN, syntactic negation in the verbal group is the most frequent to coincide with a modal marker in one clause, probably to no surprise as it is also the most frequent type of position for syntactic negation markers. Furthermore, syntactic negation markers in adverbial groups combine more often with modal markers than those in nominal groups do, although the latter ones are more frequent.

Feature	**EN %**	**GN %**
Verbal group + modal	44%	71%
Adverbial group + modal	39%	49%
Nominal group + modal	13%	38%

Table 5.46 Ranking of position of syntactic negation and modal marker in one clause

Coming to the end of this section, I want to point out a few more possible analyses which have not been carried out in the framework of this study because the total numbers of features would have been too small to allow a correct calculation of chi-squared values. First of all, the interplay between morphological negation markers and modality and between textual negation (conjunctions and continuative) and modality has been neglected. I did not go into detail with second and third modal markers in a clause, or which realization type occurs with which other realization type of modality. How do second and third modal markers interplay with syntactic negation in the same clause? Furthermore, one could look at the different positions of syntactic negation markers and how these combine with epistemic or root modality, or with the different realization types of modality. With a more substantial corpus the answers to these questions may bring interesting new insights.

6 Theme-rheme structure – the textual metafunction

6.1 Theme and rheme – theoretical background

The textual metafunction serves the purpose of creating a cohesive and coherent text out of individual words and phrases. Cohesion is achieved by arranging the clause constituents into theme and rheme components. Theme-rheme structure is, of course, not the only cohesive device of languages. However, it is a major aspect. The theme has special status in a clause and is put first, the rheme follows the theme. Halliday (1994, 37) explains that

> the theme is indicated by position in the clause. In speaking or writing English we signal that an item has thematic status by putting it first. […] The Theme is the element which serves as the point of departure of the message; it is that with which the clause is concerned.

The theme, thus, is the starting point, often containing information that has been previously mentioned or can easily be gathered from the context of the situation. The rheme follows the theme and gives information about the theme, often new information. The German *Duden Grammatik* (Fabricius-Hansen et al. 2006, 1130) agrees with Halliday's description, calling theme-rheme structure *Funktionale Satzperspektive (FSP, functional sentence perspective*).

Apart from the theme being the starting point of a message, it is also special in that it is hard to challenge what is said in the theme, it is hard to argue about it. When we question the statement of a clause, we question the information in the rheme, not that in the theme. Compare the following examples, where it is hard to express doubts about the 'fact' given in the theme that the show is brand new:

*a) <u>This brand new show</u> promises to get up close and personal to some of the stars. (*From *What's On Warwickshire*, July 2013, p. 35)

b) <u>This brand new show</u> promises to get up close and personal to some of the stars, doesn't it?

c) *This brand new show promises to get up close and personal to some of the stars. No, it does not.*

This special status makes the theme valuable for the current study. In Systemic Functional Grammar, there are three types of theme: topical, interpersonal and textual (Halliday 1994,52). The topical theme is the one that is obligatory in a clause. It contains exactly one constituent that also plays a role in the experiential metafunction, i.e. a participant, process or circumstance. These are realized by nominal phrases that function as subject or object, by verbal phrases that function as predicator or by adverbial phrases or prepositional phrases that function as complement or adjunct (Halliday 1994, 52). *This brand new show* in the example above would be a topical theme. Topical themes can be unmarked or marked or unmarked structural (Halliday 1994, 44), see section 7.1.1.

Interpersonal themes establish a relationship with the reader, they are not obligatory and do not appear in every clause. Vocatives and modal adjuncts, if they precede the topical theme, are interpersonal themes (Halliday 1994, 53). Vocatives are used to address the reader or listener, while modal adjuncts add personal opinion to what is being said. In *Helen, this brand new show promises …, Helen* would be a vocative and thus an interpersonal theme. In *Of course this brand new show promises…, of course* is a modal adjunct and thereby also an interpersonal theme. Interpersonal themes must stand before the topical theme, otherwise, they are just part of the rheme and not taken into account. Interpersonal themes are explained in more detail in chapter 6.1.3.

Last but not least, there are textual themes. These help to 'glue' clauses and sentences together. Not every clause has a textual theme, they are not obligatory. Structural conjunctives (coordinating and subordinating conjunctions), conjunctive adjuncts and continuatives are textual themes, but only if they precede the topical theme (Halliday 1994, 53). Examples for structural conjunctives would be *and, or, but, since, because*. Examples of conjunctive adjuncts include lexical items such as *in other words, anyway, also, moreover, for this reason*. Finally, continuatives are what we usually call discourse markers in linguistics, e.g. *well, now, oh, yeah*. The textual themes will be investigated in more detail in chapter 6.1.2.

The following section studies the three different kinds of topical theme and discusses differences between the English and German language with regard to unmarked and marked topical themes.

6.1.1 Topical theme

6.1.1.1 Topical theme in English

Topical themes, thus the theme that expresses the content of the clause, what the clause is about, can be unmarked or marked. The unmarked option is a subject in a clause that stands before the finite verb, with the finite verb being the border between theme and rheme (Halliday 1994, 43). For example, in *Pigs / can fly*, *pigs* is the unmarked topical theme because it is the subject, and the rest of the clause, *can fly*, is the rheme.

In the English language system, a marked topical theme is "[a] Theme that is something other than the Subject" (Halliday 1994, 44), i.e. an object, complement or adjunct standing before the subject. Special focus is given to the marked topical theme, e.g. in *From this day on, / pigs can fly*; *from this day on* is not the subject of the clause but an adjunct, we call it a marked topical theme. In this case, the rheme begins with the subject; *pigs can fly* is the rheme in this example.

Apart from unmarked and marked topical themes, there is a third option for topical themes, called **unmarked structural topical themes.** These unmarked structural topical themes are realized by relative pronouns, which have a connecting function, but also function as subject, adjunct or complement of the clause (Halliday 1994, 50).

Type	**Examples**
Definite	*which, who, that, whose, when, where, (why, how)*
Indefinite	*whatever, whichever, whoever, whosever, whenever, wherever, however*

Figure 6.1 List of relative pronouns (Halliday 1994, 50)

Consider the example *The woman who lives downstairs has five cats*. The main clause is *The woman / has five cats*. The embedded clause *who / lives downstairs* has *who* as the unmarked structural topical theme; *who* connects the two clauses and is at the same time the subject. There is not much choice about where to put the unmarked structural topical theme, it has to stand at the beginning of a clause. Any given clause can have only one of the topical themes, either unmarked, marked or unmarked structural.

The annotation scheme for English in figure 6.2 reflects the options given in Halliday (1994, 37-67) for theme-rheme structure. Figure 6.3 is an example of an annotated English clause from the EN corpus. The interpersonal and textual theme types are explained in the following sections, but before that, we need to look at topical themes in the German language.

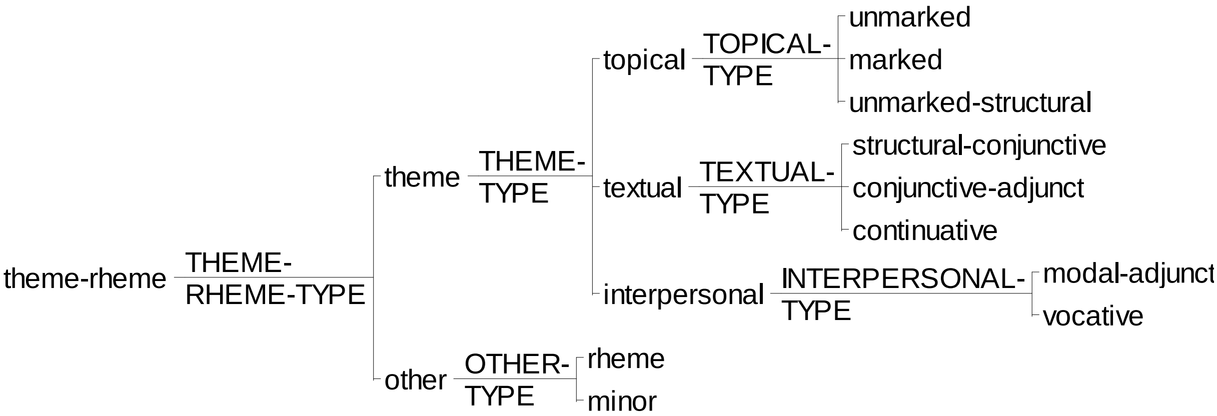

Figure 6.2 Annotation scheme for theme in English

Textual theme	**Topical theme**	**Rheme**
	She (unmarked)	*is 31*
and		*has had health issues*
	that (unmarked structural)	*may make having kids impossible*
if	*she* (unmarked)	*doesn't act now.*

Figure 6.3 Example of theme-annotation in EN

6.1.1.2 Topical theme in German

The annotation scheme for theme in English cannot be directly applied to the annotation of the German corpus, due to differences in the word order typology of English and German. The basic word order of English in declarative clauses is subject-verb-object (SVO) (Biber et al. 1999, 899), and the word order is described as 'fixed', i.e. "the placement of the core elements of the clause is strictly regulated" (Biber et al. 1999, 898). German word order, on the other hand, is less strictly regulated. The German standard reference for grammar, the *Duden Grammatik* (Fabricius-Hansen et al. 2006, 1134), speaks of the German word order (*Grundreihenfolge der deutschen Wortstellung*): In an independent declarative clause, the basic word order is subject > finite verb part > adverbial > objects > non-finite verb part. In front of the finite verb, at the beginning, we find the thematic constituent, usually the subject.

The finite verb is fixed to the second position in an independent declarative clause in German. The finite verb in second position plus any non-finite parts of the verbal group that may stand in the last position of the clause together form the *Satzklammer*, e.g. *Ich bin schon immer etwas schwierig gewesen.* In German grammar, we speak of the *Satzklammer*, the *Vor-, Mittel-* and *Nachfeld*, see the example in figure 6.4:

Vorfeld	**Satzklammer-Anfang**	**Mittelfeld**	**Satzklammer-Ende**	**Nachfeld**
Ich	*bin*	*schon immer etwas schwierig*	*gewesen*	*weil ich …*

Figure 6.4 Basic word order in German

The *Vorfeld* is the position in a German clause where we would typically find the unmarked topical theme, i.e. the subject. Götze and Hess-Lüttich (1999, 481) give the following example:

(a) *Müller hat eine Million im Lotto gewonnen.*

The *Vorfeld*, however, does not have to be filled by the subject. Götze and Hess-Lüttich (1999, 481) explain that if a constituent other than the subject moves to the *Vorfeld*, we speak of inversion. The finite verb in second position functions like an axis around which the subject and the other constituent revolve. They give this example:

(b) *Eine Million hat Müller im Lotto gewonnen.*

or: (c) *Im Lotto hat Müller eine Million gewonnen.*

Inversion serves the speaker to put the emphasis on the constituent in the *Vorfeld*, which may not always be the subject (Götze and Hess-Lüttich 1999, 481). Thus, if the subject stands before the finite verb, the subject is the unmarked topical theme. If a speaker or writer uses inversion, i.e. if she puts a constituent other than the subject in *Vorfeld* position, that constituent would be a marked topical theme. But "[f]unctional elements other than the subject can easily be conflated with Theme, often without creating a particularly marked word order" (Steiner and Teich 2004, 143).

They continue by saying that

> […] it is not the Subject in German which realizes the Mood element together with Finite, but rather the position of the Finite itself. The Subject in German does not have a high functional load for expressing Mood, and is therefore not tied to a preverbal position in declaratives. For the same reason, German has no obligatory ideational element within the Theme. (Steiner and Teich 2004, 180)

In the view of Steiner and Teich (2004, 121), the topical theme in German is unmarked if the most inherent / least oblique participant role stands before the finite verb. The most inherent / least oblique participant (1st participant role) is either

a) the subject which is required by intransitive, transitive and ditransitive verbs, e.g *sie schläft, er küsst sie, sie gibt ihm eine Ohrfeige.*

b) the indirect object (nominal group in dative case) which is required by a small group of intransitive verbs, e.g. *mir ist kalt/schwindelig/übel*.

Steiner and Teich (2004) suggest to speaking of an unmarked theme if the *Vorfeld* is filled with either '1st participant role' (i.e. the most inherent), circumstance, textual or interpersonal elements. This view is shared by Neumann (2003). Traditional grammars suggest that we should draw the line between unmarked and marked topical themes between subject in *Vorfeld* position, since this is the most frequent choice, and anything else in *Vorfeld* in declarative clauses. But some clauses, as Steiner and Teich´s examples show, do not require a subject but an object in *Vorfeld* position, e.g. *Mir ist kalt* (Steiner and Teich 2004, 155).

In order to gain more clarity about where the line between unmarked and marked topical theme should be drawn in German, the annotation scheme of theme for German was extended to add information on which constituent occupies the *Vorfeld* position, see figure 6.5 below. The additional information enables us to study the frequency of the different kinds of constituents in *Vorfeld* position, and to draw the line between unmarked and marked topical theme where it seems appropriate. More information on how to annotate constituents in the German theme-rheme structure can be found in the appendix.

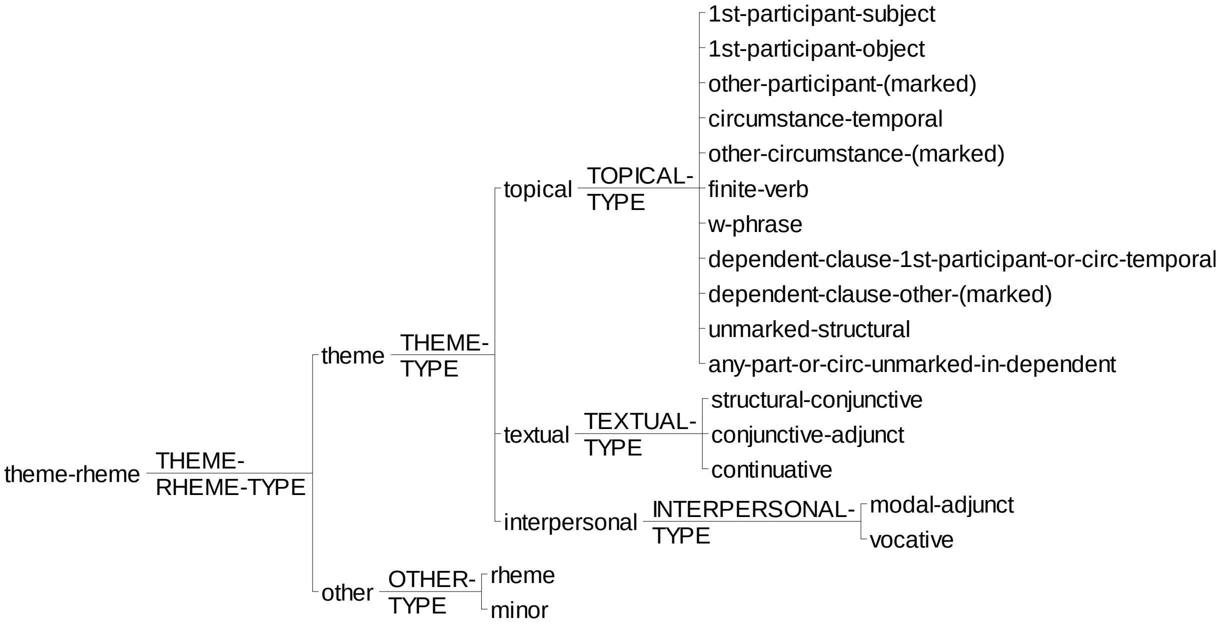

Figure 6.5 Annotation scheme for theme in German

The main types of constituents which can be found frequently in the first position in a clause in GN are demonstrated with examples 137 to 143 below.

(137) 1st participant subject in the Vorfeld: *<u>Meine Kindheit</u> war super.* *('My childhood was great.')*

(138) 1st participant object in the Vorfeld, required by a small number

of intransitive verbs: *Mir ist immer so schwindelig. ('Me is always so dizzy.')*

(139) Temporal circumstance in the Vorfeld: *Letzte Woche habe ich ihm davon erzählt ('Last week have I him of it told.')*

(140) Dependent finite clause expressing 1st participant or temporal circumstance in the Vorfeld: *Als ich noch geraucht habe wog ich 75 kg. ('When I still smoked have weighed I 75 kg.')*

(141) Other participant, i.e. objects of transitive and ditransitive verbs in the Vorfeld: *Einen von beiden werde ich verlieren. ('One of the two will I loose').*

(142) Other circumstance in the Vorfeld: *In meinem Kopf kreisen so viele Gedanken. ('In my head circle so many thoughts.')*

(143) Dependent finite clause expressing other participant or circumstance in the Vorfeld: *Wo wir gehen, wachsen Blümchen. ('Where we go grow flowers.')*

Table 6.1 gives the raw and relative numbers for different constituents in the *Vorfeld* position in the German newsgroup corpus.

Constituent in Vorfeld	**GN F**	**GN%**
1st participant subject	819	67
1st participant object	2	0
Circumstance temporal	71	6
Finite verb in interrogative	56	5
W-phrase in interrogative	41	3
Dependent clause 1st part. or circ. temporal	9	1
Any unmarked in dependent clause	10	1
Other participant	52	4
Other circumstance	34	3
Dependent clause other part. or circ.	14	1
Unmarked structural theme	116	9
Total	1,224	100

Table 6.1 Raw and relative numbers for different constituents in *Vorfeld* position

Clearly the subject in theme position is the most frequent choice, and therefore the most unmarked type of topical theme with 67% of all clauses, exceeding even Petersen's (forthcoming) estimate of 60% of subjects in *Vorfeld* position. There are only two clauses in the entire GN corpus that have an object as the most inherent participant; these must be counted as unmarked topical themes since certain verbs require this construction. The temporal circumstances are most frequently found in *Vorfeld* position, twice as often (6%) as all other circumstances combined (3%). They are thus counted as unmarked topical themes. There are also a few dependent finite clauses that function as subject or temporal circumstance (1%), plus a few unclear cases (1%). All these constituents together with the finite verbs and w-phrase in interrogatives make up 83% of all topical themes. I will conclude for the present study that they form the unmarked topical themes in German.

The marked topical themes in German include participants which are objects of transitive and ditransitive verbs as well as all circumstances except temporal ones and finally dependent finite clauses that function as object or other circumstance. Unmarked structural topical themes, as in English, are relative pronouns and relative pro-adverbials and relative adverbials which at the same time function as subject of a clause and as connecting word. Table 6.6 gives a list of these lexical items that can be unmarked structural topical themes in German.

Type	**Examples**
Relative pronoun	*der/die/das, wer/was, welche/welches/welcher*
Relative pro-adverbial	*wo, wohin, wann, wie, warum, wieso, woher, weswegen, weshalb,*
Relative adverbial	*wogegen, worauf, wonach, wodurch*

Figure 6.6 List of relatives functioning as unmarked structural topical theme in German (Fabricius-Hansen et al. 2006, 310, 584)

If constituents in the German annotation of topical themes are grouped together in the way described here, the raw and relative numbers are comparable to the numbers from the English theme annotation, see table 6.2 below.

Topical theme	**EN F**	**GN F**	**EN%**	**GN%**
unmarked	1,251	1,008	88	83
marked	76	100	5	8
unmarked structural	102	116	7	9
Total	1,429	1,224	100	100

Table 6.2 Comparison of raw and relative numbers for types of topical theme

6.1.2 Textual theme

Textual themes are the connecting words and phrases that 'glue' clauses together. They have no function inside the clause, and they are optional. Not every clause has a textual theme. Textual themes stand at the beginning of clauses, they precede the topical theme. Thus, in the EDNA annotations, if there is a textual theme in a clause complex, it counts as part of the clause that follows. In the *Introduction to Functional Grammar* (IFG), Halliday (1994, 53) distinguishes three types of textual themes: structural conjunctives, conjunctive adjuncts and continuatives. Structural conjunctives are a word class of their own and unite two parts of the same class into one. In our case, they unite two clauses into one clause complex. Table 6.3 shows the conjunctions as they are given in Halliday (1994, 50) for the English language, and table 5.4 shows the equivalent set for the German language.

Type	**Examples**
Co-ordinator	*and, or, nor, neither, but, yet, so, then*
Subordinator	*when, while, before, after, until, because, if, although, unless, since, that, whether, (in order) to,*
	even if, in case, supposing (that), assuming (that), seeing (that), given that, provided (that), in spite of the fact that, in the event that, so that

Table 6.3 List of structural conjunctives in English (Halliday 1994, 50)

Type	**Examples**
Co-ordinator	*und, oder, aber, sowie, sowohl … als auch, weder … noch, wenn auch,*
Subordinator	*wenn, bis, dass, wo, ob, außer, während, als, wie, nachdem, seit, bevor, um, anstatt, ohne, weil, indem, als ob,*

Table 6.4 List of structural conjunctives in German (Fabricius-Hansen et al. 2006, 628-640)

The second type of textual themes, the conjunctive adjuncts, "set up a semantic relationship with what precedes" (Halliday 1994, 50). Halliday also gives a list of examples of conjunctive adjuncts, see figure 6.7 below. The different types of conjunctive adjuncts have not been distinguished in the process of annotating EDNA. Table 6.8 shows the list of conjunctive adjuncts used in the process of annotating the German texts in EDNA; it is a simple translation of the English terms in the previous list.

Type of conjunctive adjuncts	**Meaning**	**Examples**
Appositive	′ i. e.′, ‘e. g.'	*that is, in other words, for instance*
Corrective	'rather'	*or rather, at least, to be precise*
Dismissive	'in any case'	*in any case, anyway, leaving that aside*
Summative	'in short'	*briefly, to sum up, in conclusion*
Verifactive	‘actually'	*actually, in fact, as a matter of fact*
Additive	'and'	*also, moreover, in addition, besides*
Adversative	'but'	*on the other hand, however, conversely*
Variative	'instead'	*instead, alternatively*
Temporal	'then'	*meanwhile, before that, later on, next, soon, finally*
Comparative	'likewise'	*likewise, in the same way*
Causal	'so'	*therefore, for this reason, as a result, with this in mindeden, . n process. ve adjuncts have not been distinguished in the annotation process. e of the fact that, in the ev*
Conditional	'(If ...) then'	*in that case, under these circumstances, otherwise*
Concessive	'yet'	*nevertheless, despite that*
Respective	'as to that’	*in this respect, as far as that’s concerned*

Figure 6.7 List of conjunctive adjuncts in English (Halliday 1994, 49)

Type of conjunctive adjuncts	**Meaning**	**Examples**
Appositive	´i.e.´, ‘e.g.´	*mit anderen Worten, z.B., bzw.*
Corrective	´rather´	*oder eher, mindestens, um genau zu sein*
Dismissive	´in any case´	*auf jeden Fall, sowieso, davon abgesehen*
Summative	´in short´	*kurzum, zum Abschluss, abschließend*
Verifactive	´actually´	*eigentlich, genau genommen, übrigens*
Additive	´and´	*auch, überdies, zudem, außerdem*
Adversative	´but´	*einerseits, andererseits, jedoch, umgekehrt, zwar*
Variative	´instead´	*stattdessen, ersatzweise*
Temporal	´then´	*inzwischen, vorher, dann, endlich, als Nächstes, danach, zum Schluss, anfangs*
Comparative	´likewise´	*gleichfalls, ebenso*
Causal	´so´	*deshalb, aus diesem Grund, als Folge, folglich*
Conditional	´if-then´	*in diesem Fall, unter diesen Umständen, ansonsten*
Concessive	´yet´	*nichtsdestoweniger, trotz, trotzdem*
Respective	'as to that´	*in dieser Hinsicht, was X betrifft,*

Figure 6.8 List of conjunctive adjuncts in German, translated from Halliday (1994, 49)

The third type of textual themes are the continuatives. Halliday (1994, 53) gives *yes, no, well, oh, now* as examples. Continuatives are also called discourse markers. Discourse markers can be interjections or words that are not syntactically independent (Bußmann 2002, 173). German continuatives / discourse markers would be lexical items like *ja, nein, also, ok, übrigens, ach, oh*.

6.1.3 Interpersonal theme

Interpersonal themes build a connection to the listener or reader. There are three types (Halliday 1994, 53): vocatives, modal adjuncts and mood-marking elements. Interpersonal themes are optional, not every clause has an interpersonal theme. Vocatives and modal adjuncts only count as interpersonal theme if they precede the topical theme. For example, in the clause *Maybe we will join you for lunch*, *maybe* is an interpersonal theme, it stands before the topical theme *we*. But in a clause like *We will certainly be hungry by then, certainly* stands after the topical theme and thus is part of the rheme, no special emphasis is given to *certainly*. Examples of modal adjuncts in English that may be found in interpersonal themes include *probably, maybe, sometimes, always, seldom, really*. A list can be found in Halliday (1994, 49). These are the same lexical items that have been annotated as expressing modality, see chapter 6. If they stand before the topical theme, these words are annotated as modal markers (interpersonal metafunction), and also as interpersonal theme (textual metafunction). German examples for modal adjuncts as interpersonal themes include *vielleicht, manchmal, auf jeden Fall, allerdings*. In the German part of EDNA, modal adjuncts only count as interpersonal theme if they stand in *Vorfeld* position, i.e. before the finite verb in a declarative clause. They are also annotated as modal markers (interpersonal metafunction).

Vocatives can also be interpersonal themes (Halliday 1994, 53). They serve to address the listener and are more common in spoken language. For example, *Guys, why don't you just shut up?* has a vocative, *guys*. The example *Helen, will you join us for dinner tonight* also has a vocative, *Helen*. Such names are textual themes if they stand in front of the topical theme. The same applies for the German language.

Mood-marking elements are interpersonal themes too, according to Halliday (1994, 53): "A mood-marking theme is a finite verbal operator, if preceding the topical theme; or a WH-interrogative (or imperative *let's*) when not preceded by another experiential element (i.e. when functioning simultaneously as topical theme)." This double function as topical and interpersonal theme is what caused the mood-marking elements to be excluded from the annotation of interpersonal theme in this study for the sake of simplicity. They are only annotated as topical theme.

6.1.4 Other clause elements

In a nutshell, the rheme of a clause is everything that is not a textual, interpersonal or topical theme. The cut-off point is behind the topical theme (Halliday 1994, 37). That means that in a declarative sentence, the rheme begins with the finite verb, e.g. <u>Helen</u> / *is in Birmingham today*, the rheme is *is in Birmingham today*, it follows the topical theme *Helen*. In an interrogative sentence, the finite verb or WH-word is the topical theme, and thus, everything that follows is the rheme: <u>Where</u> / *is Helen today?* In an imperative sentence, the non-finite verb or *let's* is the topical theme and everything that follows is the rheme, e.g. <u>*Close*</u> / *the window, please!* <u>*Let's*</u> / *go for lunch!* This principle can be applied to the German language as well: whatever follows the topical theme is the rheme.

The last element in the discussion of theme-rheme structure is the minor clause. According to Halliday (1994, 63), minor clauses have no finite verb (they are not imperatives, either), and therefore no mood structure. They often function as greetings or exclamations; they can be formulas like *Good night! Thank you. Hi sweetheart!* Minor clauses, lacking mood, cannot be analysed in terms of theme-rheme structure, but are annotated as minor clausse in the EDNA corpus. Of course, the German language also has minor clauses, the criteria are the same as for the minor clauses of the English language.

6.2 Quantitative analysis of theme-rheme structure

6.2.1 Types of theme

In what follows, we go through the system (represented in the annotation scheme) of theme, moving from left to right, testing in each step the null hypothesis which says that there is no meaningful deviation between the English and the German part of the EDNA corpus of newsgroup texts. The first of our tables (6.5) displays the numbers and results of the chi-squared test for theme and other, i.e. rhemes and minor clauses.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	χ^2	**S**
Theme	2,155	1,953	56	54	EN	1.09	-
Other	1,677	1,635	44	46	GN	1.36	-
Column total	3,832	3,588	100	100			

Table 6.5 Raw and relative numbers, χ^2 and significance for *theme* and *other*

It is difficult to tell whether the English and German newsgroup texts differ to a substantial degree by looking at the raw numbers alone. However, if we look at the χ^2 values, it becomes clear that the differences are not significant. The distribution of theme and other across EN and GN is fairly even, there is hardly any deviance from the expected frequencies. With a threshold of 3.84 for $p < 0.05$ (df=1), the null hypothesis cannot be rejected.

Next, table 6.6 below shows the calculation and results of the chi-squared test with regard to the different major types of theme.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	χ^2	**S**
Topical theme	1,429	1,224	66	63	EN	2,10	-
Textual theme	695	690	32	35	GN	2,88	-
Interpersonal theme	31	39	1	2	GN	1,87	-
Column total	2,155	1,953	100	100			

Table 6.6 Distribution of major types of theme

Although there are small differences in the relative numbers of EN and GN, the χ^2 values are smaller than the threshold for $p < 0.05$ (df= 2), 5.99. Therefore, with a degree of certainty of 99.5%, we cannot reject the null hypothesis and must assume that the three types of theme are distributed equally in the English and German newsgroup texts.

Taking up the features and numbers from table 6.2 once more, we can now test whether the distribution of unmarked, marked and unmarked structural topical themes diverges to a statistically significant degree between the English and German part of EDNA. The results are displayed in table 6.7.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	χ^2	**S**
Unmarked theme	1,251	1,008	88	83	EN	2,09	-
Marked theme	76	100	5	8	GN	8,08	+
Unmarked structural	102	116	7	9	GN	4,39	-
Column total	1,429	1,224	100	100			

Table 6.7 Distribution of types of topical theme across EN and GN

The difference in the frequency of unmarked themes and unmarked structural themes is not significant, but the German subcorpus has more marked themes, at the level of significance of $p < 0.05$ (df =2), where the threshold is 5.99. This must be due to greater word order freedom in German. The null hypothesis can be rejected for the types of topical themes in EDNA.

Following the investigation of the types of topical theme, we now focus on the three types of textual theme. Table 6.8 below gives the results.

Feature	EN F	GN F	EN%	GN%	Overuse	χ^2	S
Structural conjunction	641	570	92	83	EN	3,67	-
Conjunctive adjunct	31	91	4	13	GN	29,94	+ + +
Continuative	23	29	3	4	GN	0,74	-
Column total	695	690	100	100			

Table 6.8 Distribution of types of textual theme across EN and GN

The structural conjunctions and the continuatives do not show any significant difference ($p < 0.05$, df = 2, the threshold is 5.99). The conjunctive adjuncts, however, are used more frequently in GN to a highly significant degree ($p < 0.001$, df =2, the threshold is 13.82). Thus, we can reject the null hypothesis. Examples for conjunctive adjuncts in German would include *nun, sonst, also, zumindest, danach, im Gegenteil*; English examples would include *however, by the way, just, also, even though, especially*. The German authors use about three times the amount of the English authors.

The calculation of statistical significance of the third theme system, interpersonal theme, is not possible because the expected frequencies are below 5 in two of the four cells. For such small numbers the χ^2 test cannot be used (Gries 2008, 157). Table 6.9 displays the raw and relative numbers for the interpersonal themes in EDNA.

Feature	EN F	GN F	EN%	GN%
Modal adjunct	28	39	90	100
Vocative	3	0	10	0
Column total	31	39	100	100

Table 6.9 Observed and relative frequencies of interpersonal theme in EN and GN

The three vocatives used in EN are *Girls; Ladies, Ana friends; Ana Love* (where *Ana* stands for *Anorexia)*, and modal adjuncts in EN include *especially, please, maybe, hopefully*. In GN, modal adjuncts include *manchmal, allerdings, in Wahrheit, vermutlich*, and there is not a single vocative in 10,000 words. Although these newsgroup texts are rather informal discourse, interpersonal themes are not used excessively.

Below, table 6.10 shows the results for the other clause elements within the theme-rheme structure, i.e. rheme or minor clause.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	χ^2	**S**
Rheme	1,578	1,499	94	92	EN	0,52	-
Minor	99	136	6	8	GN	6,80	+ +
Column total	1,677	1,635	100	100			

Table 6.10 Distribution of types of other across EN and GN

We can reject the null hypothesis that rheme and minor clauses occur to an equal degree in both corpora. In GN, there are significantly more minor clauses (6.64 is the threshold for $p < 0.01$, df = 1). The difference in raw numbers, however, is only 37. The minor clauses in the EN subcorpus are mostly salutations, formulas and the names of the authors, e.g. *hi guys, take care everyone, Liz*. In the German texts, apart from the salutations, formulas and names of authors that also appear here, there are quite a few clauses where the verb is missing, as in the following examples:

(144) Oft auch von mir aus.

(145) Also gemeinsam, und doch getrennt.

(146) Treffen nur an öffentlichen Orten und nur zum Lernen.

(147) Zumindest noch einmal.

(148) Und dann die vielen Telefonate.

These seem to be examples of sentence fragments, i.e., "a sentence reduced by ellipsis to an incomplete form" (Matthews 2005, 338). Such sentence fragments seem to suggest a proximity of the newsgroup texts to spoken more than to written discourse.

6.2.2 Summary

After investigating all the options in the theme-rheme structure step by step, we can summarize the results as in figure 6.9.

System	Divergence	Divergent feature	Divergent corpus
Theme / other	Not significant		
Type of theme	Not significant		
Topical theme	Significant	Marked theme	GN +
Textual theme	Significant	Conjunctive adjunct	GN + + +
Interpersonal theme	-	-	-
Other	Significant	Minor clause	GN ++

Figure 6.9 Summary of tests of statistical significance of theme / other

The English and German newsgroup texts in EDNA are strikingly similar with respect to the numbers of the major types of themes: topical, textual and interpersonal ones. The German writers use more marked topical themes, which is easy in German due to the greater word order freedom. The German writers also use more conjunctive adjuncts, thereby making their texts more cohesive than the English ones. And finally, German writers use more minor clauses, many of which are sentence fragments. In the following subchapter, the focus is on the lexical items that are used as topical, textual, and interpersonal themes.

6.3 Analysis of lexical items used as theme

After the quantitative approach in the previous chapter, we will now focus on the analysis of the lexical items in the theme-rheme structure in the English and German parts of the EDNA corpus.

6.3.1 Unmarked topical theme in EN

We will begin with the unmarked topical theme in the EDNA corpus. Table 6.11 below presents the most frequently used unmarked topical themes in EDNA, and will help us to see whether the newsgroup text writers use many pronouns, like in conversation, or rather nominal elements, like in written texts.

N	Word EN	F	%
1	I	641	51.24
2	IT	82	6.55
3	SHE	82	6.55
4	HE	70	5.6
5	WE	63	5.04
6	THEY	19	1.52
7	THERE	17	1.36
8	THIS	17	1.36
9	YOU	16	1.28
10	WHAT	11	0.88
		936	74.8
	Other	315	25.2
	Total	1,251	100

Table 6.11 Heads of nominal groups filling the unmarked topical theme position in EN

The heads of the nominal groups in subject position, i.e. the unmarked topical themes in the English EDNA corpus, verify Halliday's (1994, 44) assumption that the unmarked topical theme in spoken language is most often the pronoun *I* or *you*, or, the other way around, that the EDNA texts are more spoken than written discourse. *I* is the pronoun in 51% of all unmarked topical themes with 641 instances. Following second are *it* and *she* with only 82 instances each. Within the ten most frequent heads of nominal groups, we find seven pronouns. *This* and *that* also function as pronouns for anaphoric reference here, not as determiners or relative pronouns. Furthermore, there is the existential *there*, e.g. *there is this guy*, and the WH-pronoun / question word *what*, e.g. *what would you do in my situation*. *What* is used more frequently than any other question word. Writers ask for *what*, not for *why* or *how* or *when*, for example.

6.3.2 Unmarked topical theme in GN

In this section, the focus is on the unmarked topical themes in the German part of EDNA. The extraction of the heads of nominal groups has been more complicated than it was with the English newsgroup texts, because the annotation of the German texts is more detailed. As has been described earlier, this procedure was chosen to enable us to decide what would be unmarked and marked topical themes in the German texts in the first place. In this study, the following constituents of a clause are considered unmarked topical themes in German:

First position in a declarative clause:

- Subjects
- Objects required by an intransitive verb needing an object but no subject
- Temporal circumstances standing before the subject
- A dependent clauses functioning as subject in an independent declarative clause
- A nominal or prepositional group at the beginning of a dependent clause.

First position in an interrogative clause:

- The finite verb
- The WH-pronoun (e.g. *was, wann, warum*)

First position in an imperative clause:

- The non-finite verb

Table 6.12 below displays the results for the frequency of heads of nominal groups in unmarked topical theme position in the German newsgroup texts in the EDNA corpus.

N	Word GN	F	%
1	ICH	400	39.72
2	ER	79	7.85
3	ES	58	5.76
4	SIE	46	4.57
5	DAS	28	2.78
6	WIR	24	2.38
7	MAN	11	1.09
8	WAS	11	1.09
9	WARUM	9	0.89
10	ALLES	6	0.6
		672	66.7
	Other	336	33.3
	Total	1,008	100

Table 6.12 Heads of nominal groups filling the unmarked topical theme position in GN

In the German newsgroup texts, the pronoun *ich* is the most frequent head in unmarked topical themes with 400 out of 1008 unmarked topical themes. With 40%, however, it is far less frequent than *I* in the English texts with 51%. Similar to the English newsgroup texts, the pronouns *er*, *es* and *sie* are the next most frequent unmarked topical themes, although far less frequent than *ich*, with 79, 58 and 46 occurrences. Of the ten most frequent heads of nominal groups, eight are pronouns, including *das* and *man*. In contrast to the English newsgroup texts, where we find only *what* as a WH-pronoun, in the German texts we have *was* and *warum*, which introduce questions, among the more frequent unmarked topical themes. Finally, there is the word *alles*, which seems to be a pronoun used for endophoric or exophoric reference, e.g. *Alles tut weh. Alles dreht sich nur um's Essen und um's Zunehmen*, thus, another pronoun. Looking at the ten most frequently used unmarked topical themes alone would not tell any reader what these texts are about.

When we look at the total number of unmarked topical themes in the EDNA corpus, we realize that the German authors use a greater variety of heads of groups as unmarked topical themes than the English authors. In the English newsgroup texts, the 10 most frequent heads together make up for 75% of all unmarked topical themes. In the German newsgroup texts, the 10 most frequent heads of nominal groups together form only 66% of all unmarked topical themes.

6.3.3 Marked topical theme

Even though less frequent than unmarked topical themes, it can be rewarding to study the marked topical themes in texts, and this is what we do now, beginning with the English texts. In the English language, marked topical themes are realized by a nominal, prepositional or adverbial group which do not have the grammatical function of subject in the clause. Table 6.13 below shows the groups used as marked topical themes in the English newsgroup texts. The table shows the entire group, not only the head. Marked topical themes are rare and too much information would have been lost if the groups had been reduced to their heads only.

N	Word EN	F	%
1	NOW	8	10.53
2	AMONTHAGO	2	2.63
3	LATELY	2	2.63
4	THEHARDER	2	2.63
5	TODAY	2	2.63
6	XWEEKSAGO	2	2.63
7	XYEARSAGO	2	2.63
8	AFTERAXYEARRELATIONSHIP	1	1.32
9	AFTERLOOKINGAT[…]	1	1.32
10	AFTERMANYTESTS	1	1.32
		23	30.2
	Other	53	69.8
	Total	76	100

Table 6.13 Groups realizing the marked topical theme in EN

In the English newsgroup texts, marked topical themes are not very numerous. Out of the 1429 topical themes in the EDNA corpus, 1251 are unmarked topical themes (88%) and only 76 are marked topical themes (5%), plus 102 unmarked structural themes (7%). Therefore it is not surprising to find that the most frequent marked topical theme, *now*, has only 8 instances, the other groups appear only twice or once. What is striking, however, is that the large majority of groups realize temporal circumstances, locating the process in time, e.g. *a* *month ago, lately, today, x years ago, x weeks ago, at the time, before*. We can attest a similarity to the German newsgroup texts here. As a result of the annotation of the German texts, temporal circumstances are considered unmarked topical themes in German since they appear twice as often (6% of all topical themes) as all other circumstances taken together (3% of all topical themes), see table 6.1. The annotation of the English texts show that temporal circumstances are also the type of circumstance most frequently used before the subject in the English newsgroup texts, thereby making it less marked than any other circumstance in that position. Had temporal circumstance been considered unmarked topical themes in the English EDNA corpus, hardly any marked topical themes would have been left. One of the groups must probably be considered an annotation mistake: *the harder*, where the two instances come from the construction *and the harder I tried not to fall for him, the harder I did*, where the marked topical theme is probably more of a textual theme.

Let us turn to the marked topical theme in GN. A marked topical theme in the German newsgroup texts can be any nominal, adverbial or prepositional group that stands before the finite verb, and that does not function as subject. The finite verb has to stand in the second position in an independent declarative clause in the German language system, but the position before it, the *Vorfeld*, can be filled by either a topical, interpersonal or textual theme, thereby moving the subject to a position after the finite verb. Table 6.14 displays the groups which are marked topical themes in the German newsgroup texts. As with the English marked topical themes, the groups have not been reduced to their heads in order not to lose information.

N	**Word GN**	**F**	**%**
1	MIR	12	12
2	DA	8	8
3	IRGENDWIE	6	6
4	FUERMICH	4	4
5	DAS	3	3
6	MICH	3	3
7	ABGENOMMEN	2	2
8	ES	2	2
9	ALSZWEITENSCHRITT	1	1
10	ANEINZUGMEINERSEITS	1	1
		42	42
	Other	58	58
	Total	100	100

Table 6.14 Groups realizing the marked topical theme in GN

There are significantly more marked topical themes in the German EDNA corpus than there are in the English part. Out of 1224 topical themes in total in the German texts, 1008 are unmarked topical themes (83%), 100 are marked topical themes (8%), and 116 are unmarked structural themes (9%). We must not forget that temporal circumstances in theme position add to the number of marked topical themes in the English corpus, but not in the German one. If temporal circumstance had been added to marked topical themes in German, the total number would have been even higher. We can conclude, however, that there are more marked topical themes in the German newsgroup texts than in the English ones, this is certainly due to the word order typology of German which allows a greater variety of groups to be put in *Vorfeld* position. A closer look at the marked topical themes reveals that most of them have the function of putting the focus on the writer. The writer starts the clause with *mir, für mich, mich*.

When it comes to marked topical themes, as opposed to unmarked topical themes, English has a greater diversity of marked topical themes. The most frequent marked topical themes together make up only 30% of all, whereas most marked topical themes occur only once (70%). In the German texts, the lexical items that are most frequent as marked topical themes make up as much as 42%, groups which occur only once make up 58% of all marked topical themes. In GN, we can see a clear trend of putting the focus on the writer with marked topical themes, something that we cannot say about the English marked topical themes. They more often than not express a location in time.

6.3.4 Unmarked structural theme

Finally, in addition to the unmarked and marked topical themes, there are the unmarked structural themes. Their function is both textual and topical at the same time; they introduce a subordinate clause with a relative pronoun or an equivalent pronominal group. Table 6.15 below shows the results.

N	Word EN	F	%	Word GN	F	%
1	WHAT	28	27.45	WAS	28	24.14
2	THAT	19	18.63	DIE	21	18.1
3	WHO	12	11.76	DER	9	7.76
4	HOW	9	8.82	WIE	9	7.76
5	WHICH	9	8.82	DAS	6	5.17
6	WHERE	5	4.9	ANDENEN	4	3.45
7	WHENEVER	4	3.92	WO	4	3.45
8	HOWMUCH	3	2.94	DEN	3	2.59
9	WHY	3	2.94	MITDEM	3	2.59
10	DURINGWHICH	1	0.98	MITDER	2	1.72
		93	91.18		89	76.72
	Other	9	8.82	Other	27	23.28
	Total	102	100	Total	116	100

Table 6.15 Unmarked structural themes in EDNA

Obviously, there are similarities between English and German newsgroup texts. *What* (*was*) is the most frequent relative pronoun, followed by *that* and *who* with 31 instances in the English newsgroup texts and the German equivalent, *die, der, das,* with 36 instances in the German texts. The German newsgroup texts have a greater variety of different pronominal groups as unmarked structural themes compared to the English texts. This is probably due to the existence of gender specific relative pronouns in the German language system.

6.3.5 Textual theme

Apart from topical themes, there are textual and interpersonal themes in EDNA. In the following, we look at textual themes first and then interpersonal themes. The main type of textual theme is the structural conjunctive (coordinating and subordinating conjunctions). Table 6.16 shows the ten most frequent structural conjunctives in EN and GN.

N	Word EN	F	%	Word GN	F	%
1	AND	221	32.76	UND	193	33.80
2	BUT	76	11.86	DASS	102	17.86
3	THAT	76	11.86	ABER	59	10.33
4	IF	31	4.84	DENN	21	3.68
5	WHEN	28	4.37	WENN	19	3.33
6	BECAUSE	27	4.21	DA	17	2.98
7	AS	16	2.50	WEIL	16	2.80
8	SO	13	2.03	OB	15	2.63
9	OR	13	2.03	ODER	14	2.45
10	UNTIL	9	1.40	WIE	10	1.75
		510	79.56		466	81.75
	Other	131	20.43	Other	104	18.25

Table 6.16 Structural conjunctives in EDNA

In both the English and German newsgroup texts, 47% of the ten most frequent structural conjunctives are coordinating conjunctions (*and, but, or* in EN; *und, aber, oder* in GN). *And* as well as *und* make up for one third of structural conjunctives in the tables. *But* and *aber* account for about 12% / 10%, and the coordination with *or* and *oder* is far less frequent, with only about 2% in both subcorpora. These results are in agreement with what Biber et al. (1999, 81) say: "*And* is by far the most common coordinator in all the registers […]. *But* is most frequent in conversation and fiction, and least frequent in academic prose. *Or* is far more common in academic prose than in the other registers." EDNA has roughly 21,000 times *and* (EN) and 18,500 times *und* (GN) per million words. *But* occurs about 7,000 times (EN), *aber* about 5,600 times (GN) per million words. The least frequent coordinator is *or* with roughly 1,200 (EN) and *oder* with about 1,300 occurrences (GN) per million words. It is hard to tell from the bar chart on page 81 in Biber et al. (1999) exactly how many times *and*, *but* and *or* occur per million words, but it seems that the results from EDNA correspond best to the columns for the registers of conversation and news, and that the results from GN do not differ to a great extent from the results in EN. Unfortunately, Fabricius-Hansen et al. (2006) and Götze and Hess-Lüttich (1999) include no statements about frequencies of coordinating or subordinating conjunctions.

Subordination of clauses is slightly more frequent in EDNA than coordination, with 53% of the most frequently used structural conjunctives introducing subordinate clauses. For subordinating conjunctions, there are no frequencies given, neither in Biber et al. (1999) nor Fabricius-Hansen et al. (2006) or Götze and Hess-Lüttich (1999).

The next table (6.17) shows the five most frequent conjunctive adjuncts which were used in the English newsgroup texts. Another 11 lexical items appeared only once. The same table also shows the five most frequent conjunctive adjuncts in GN. These are significantly more frequent in GN, compared to EN. Apart from the five most frequent ones, there are another 17 conjunctive adjuncts which appear twice, and 37 which appear only once. GN has a much greater variety of different conjunctive adjuncts.

N	**Word EN**	**F**	**%**	**Word GN**	**F**	**%**
1	HOWEVER	7	22.58	DANN	15	16.48
2	ANYWAY	6	19.35	NUN	11	12.09
3	ESPECIALLY	3	9.68	NUR	10	10.99
4	ALSO	2	6.45	SEITDEM	3	3.30
5	BY THE WAY	2	6.45	EIGENTLICH	3	3.30
		20	64.52		42	46.15
	Other	11	35.48	Other	49	53.85

Table 6.17 Conjunctive adjuncts in EDNA

As the final step in this chapter, we look at the continuatives. Table 6.18 below shows all the continuatives which were used by the writers in EDNA.

N	Word EN	F	%	Word GN	F	%
1	WELL	9	39.13	ALSO	8	27.59
2	NOW	3	13.04	SO	4	13.79
3	YEAH	3	13.04	NAJA	4	13.79
4	SO	2	8.70	NUNJA	2	6.90
5	HEY	1	4.35	ALSOLEUTE	1	3.45
6	WELL SORRY	1	4.35			
7	OH	1	4.35			
8	GEE	1	4.35			
9	YES	1	4.35			
10	OKAY	1	4.35			
		23	100		19	65.52
	Other	0	0	Other	10	34.48
	Total	23	100	Total	29	100

Table 6.18 Continuatives in EDNA

We see also the five most frequent continuatives from the German newsgroup texts, there are another 10 which are used only once. Thus, once again, the German writers use a greater variety of different lexical items than the English writers.

6.3.6 Interpersonal theme

Table 6.19 shows the five most frequently used modal adjuncts which function as interpersonal theme in EDNA. Another 14 lexical items (50%) in EN appear only once. In GN, there are another 26 modal adjuncts (66%) which are used as interpersonal theme twice or once, clearly a greater variety than the English writers in EDNA use.

N	Word EN	F	%	Word GN	F	%
1	SOMETIMES	6	21.42	VIELLEICHT	5	12.82
2	MAYBE	3	10.71	ALLERDINGS	2	5.13
3	PLEASE	2	7.14	BITTE	2	5.13
4	USUALLY	2	7.14	IMPRINZIP	2	5.13
5	CERTAINLY	1	3.57	MANCHMAL	2	5.13
		14	50		13	33.33
	Other	14	50	Other	26	66.67
	Total	28	100	Total	39	100

Table 6.19 Modal adjuncts as interpersonal theme in EDNA

To sum up the previous chapter, we can conclude that in both the English and German newsgroup texts, the unmarked topical themes are mostly personal pronouns, and of these, mostly *I* / *ich* was used. In the English newsgroup texts, the marked topical themes were mainly adjuncts that expressed a temporal circumstance, whereas in the German texts, marked topical themes were often used to put the focus on the writer (*mir, mich, für mich*). The annotation guidelines differed in this respect, though, turning temporal circumstances in the *Vorfeld* position in GN into unmarked topical themes. With regard to textual themes, it has become clear that coordinating conjunctions are less frequent (47%) than subordinating conjunctions (53%). The results for the coordinators *and, but, or* (EN) agreed with Biber et al.'s (1999) statements, and did not differ much from the results for *und, aber, oder* (GN). For the other two types of textual themes, conjunctive adjuncts and continuatives, the German part of EDNA contained a greater variety of different lexical items expressing these functions. Also for interpersonal themes, in particular modal adjuncts in theme position, the German newsgroup texts had a greater variety of different modal adjuncts, compared to the English newsgroup texts.

7 Participant roles and process types – the experiential metafunction

7.1 How to identify process types in 'Cardiff Grammar'

The concept of 'transitivity' in SFG terms, as opposed to traditional grammar terms, has been introduced in the chapter on theoretical background. The two basic ideas are as follows. First, that "language enables human beings to build a mental picture of reality, to make sense of what goes on around them and inside them" (Halliday 1994, 106), and second, that the clause "embodies a general principle for modelling experience – namely, the principle that reality is made up of processes" (Halliday 1994, 106).

Thus, the system of transitivity arranges human experiences into clauses with processes and participants. Over time, two different systems for describing transitivity have developed in Systemic Functional Grammar. The first and more widely known is the so-called Sydney Grammar, named so because its mastermind, Michael A.K. Halliday, worked at the University of Sydney, Australia. The second approach has been named Cardiff Grammar, because the main figures, Robin Fawcett and Gordon Tucker, worked at Cardiff University in Wales, UK.

In both the Sydney Grammar and the Cardiff Grammar approach to transitivity, there are three major process types: action / material, mental and relational processes. Halliday (1994, 106) describes a material process as a process of the external world, where "things happen, and people, or other actors, do things, or make them happen". Mental processes, on the other hand, reflect the internal world, the processes of the consciousness. In their minds, people (or other conscious beings) record, react to and reflect experiences of the outer world (Halliday 1994, 107). The third main process type is that of relational process. These relational processes are processes of *being* and *having*, used "to relate one fragment of experience to another: this is the same as that, this is a kind of the other" (Halliday 1994, 107). Examples 149 to 154 illustrate the three major process types with clauses from EN and GN.

Material / action process

(149) *I measure myself several times a day*

(150) *[…], wie sie sich ziemlich lange geküsst haben.*

Mental process

(151) *I am terrified of being fat*

(152) *Ich ärgere mich tierisch über mich selbst.*

Relational process

(153) *I'm a self-centered jerk*

(154) *Ich habe irgendwie keine Kraft mehr.*

Apart from the three major process types, there are a small number of other, less frequent process types. The transitivity system in Sydney Grammar includes seven process types, i.e. *material, behavioural, mental, verbal, relational, existential,* and *meteorological* processes. The Cardiff Grammar transitivity system includes six process types, i.e. *action, mental, relational, influential, event-relating* and *environmental* processes. Both systems describe the whole range of possible human experiences, only in a few cases one process falls into a different process type category. Figure 7.1 shows how the two systems relate to each other.

Sydney Grammar		**Cardiff Grammar**	
Main Process	**Subtype**	**Main Process**	**Subtype**
		Influential	
Material		Action	
Behavioural		Action	
Mental		Mental	
	Mental: affection		Mental: emotion
	Mental: perception		Mental: two-role perception
			Mental: three-role perception
	Mental: cognition		Mental: two-role cognition
			Mental: three-role cognition
Verbal			Mental: three-role cognition (communication processes)
Relational		Relational	
	Relational: intensive		Relational: attributive
	Relational: circumstantial		Relational: locational
			Relational: directional
	Relational: possessive		Relational: possessive
			Relational: matching
Existential			Relational: locational
		Event-relating	
Meteorological		Environmental	

Figure 7.1 Mapping Sydney Grammar with Cardiff Grammar process types

We see that the Cardiff Grammar categorization is more fine-grained and includes two process types which are not seen as a separate process types in Sydney Grammar: *influential* and *event-relating* processes. An influential process in Cardiff Grammar is a process that describes the beginning or ending of an event, often another, subordinate process, see examples 155 and 156.

(155) *A few months ago, the passion started to die.*

(156) *Ich höre momentan zu rauchen auf.*

An event-relating process is rarely found in discourse, examples 157 and 158.

(157) *The problem is that there is this guy that I want so much.*

(158) *Eine Beziehung wegzuschmeißen ist doch zu einfach.*

The least frequent process type, the environmental (meteorological) process like *it's raining,* does not occur in the EDNA corpus.

The Cardiff Grammar *action* process includes both physical and social action, i.e. both *material* and *behavioural* processes of Sydney Grammar. One of the process types in Sydney Grammar, the *verbal* process, is only a subtype of another process type, i.e. a *mental* process, in Cardiff Grammar. A *mental* process can be generalized as 'someone knows / perceives / feels something'. The *verbal* process of Sydney Grammar is understood as 'someone causes someone else to know something' in Cardiff Grammar, i.e. it is a *mental three-role cognition* process. Examples 159 to 170 from EDNA show the subtypes of mental processes. Names like 'two-role' or 'three-role' indicate the number of participant roles required for this type of process: One or two conscious beings plus the phenomenon that is being heard, seen, said or felt (note that the phenomenon itself can be a process, see examples 164, 165 and 166).

Mental, emotion

(159) *I didn't really enjoy it.*

(160) *Ich liebe diese Tage.*

Mental, two-role perception

(161) *He looks at me.*

(162) *In so einer Verfassung hatte ich ihn noch nie erlebt*

Mental, three-role perception

(163) *He greets me with 'Hi Honey'.*

(164) *Dennoch hat er mir immer gezeigt, [dass er seine Entscheidung nicht wirklich bereut hat].*

Mental, two-role cognition

(165) *She was aware [we needed to get back early too].*

(166) *Ich weiß, [dass es falsch ist]*

Mental, three-role cognition

(167) *Mike reminds me much of my brother.*

(168) *[…], dass ich mir eingestanden habe, schwer krank zu sein.*

Mental, three-role cognition (communication)

(169) *She informed me [that I gained 55 lbs].*

(170) *Ich habe sie gefragt, [ob sie mich noch liebt].*

The third major process type is the relational process, which has five subtypes in the Cardiff Grammar, each illustrated below (examples 171 to 180) with an example from EDNA.

Relational, attributive

(171) *My boyfriend and I are both around 40.*

(172) *Das ist doch voll schädlich.*

Relational, locational

(173) *I am at home with the food most of the time.*

(174) *Das Familienbild steht noch immer auf seinem Platz.*

Relational, directional

(175) *I go to Men's group every Friday.*

(176) *Ich falle nicht mehr jeden Tag in dieses schwarze Loch.*

Relational, possessive

(177) *A married man should not have female friends.*

(178) *[…], da ich selber auch einiges an Hausrat besitze*

Relational, matching

(179) *He separated from his wife a year ago.*

(180) *Ich bin mit meiner Freundin nun seit 8 Monaten zusammen.*

Sydney Grammar's *existential* process is simply considered a *relational, locational* process in Cardiff Grammar, *there* is a dummy participant role, the location in time or space sometimes omitted, e.g. *there are family members all over the place trying to feed me more*.

The reason why the Cardiff Grammar approach to transitivity has been given preference over the Sydney Grammar approach is the existence of a set of tests to identify process types in Cardiff Grammar, which at the present time does not exist to the same extent for Sydney Grammar. The guidelines for identifying process types (Fawcett forthcoming) involve tests to identify the participant roles (PR) in each clause. The process type can then be derived from a constellation of participant roles which is specific for the process type in question. Guidelines that enable the annotator to achieve a high degree of intra- and also inter-annotator agreement are essential to a corpus-driven investigation like the present one. They ensure consistent annotation over large amounts of text and repeatability of the annotation process. Examples 181 and 182 below give two examples of how to identify a process type, based on Fawcett (forthcoming). For further information the reader is referred to the full guidelines in the appendix. After reading through the clause, we follow a three-step procedure (remember that *need to* is a modal auxiliary).

(181) I need to lose weight for health reasons, not just for looks.

Step one: How many PR does the process require? Use the re-expression test: "In this process of *losing*, we expect someone *to lose* something". Two PR: someone, i.e. *I*, and something, i.e. *weight*.

Step two: What kind of PR do *I* and *weight* represent? Use the re-expression tests:

- Carrier: The thing about X is that … "The thing about me is that I lose weight". X = *I* is PR Carrier.
- Possessed: X is what Y had/had on/lacked (as a result). "Weight is what I lacked as a result". X= *weight* is PR Possessed.

Step three: What kind of process does the constellation of PR Carrier + Process + PR Possessed represent? We look this up in the list provided with the guidelines. PR Carrier + Process + PR Possessed = relational possessive process.

Here is another example to demonstrate the three-step procedure:

(182) I'm not seeing him at all for the next couple of weeks.

Step one: How many PR does the process require? Use the re-expression test: "In this process of *seeing*, we expect someone *to see* someone". Two PR: someone, i.e. *I*, and someone, i.e. *him*.

Step two: What kind of PR do *I* and *him* represent? Use the re-expression tests:

- Agent: What X did was to … "What I did was to see him". X = *I* is PR Agent.
- Affected: What happened to X was that ... "What happened to him was that I saw him". X= *him* is PR Affected.

Step three: What kind of process does the constellation of PR Agent + Process + PR Affected represent? Look it up in the list. PR Agent + Process + PR Affected = action process.

Due to the fact that two corpora had to be annotated for process types, and one of them was in German, the guidelines had to be adapted for German. The guidelines for German are basically a translation of the English guidelines (see both in the appendix). Apart from differences in word order typology and resulting difficulties with discontinuous verbal groups, the re-expression tests could be applied to the German texts well. The ease with which the re-expression test could be transferred from English to German seems to confirm the assumption that the two languages, which share the same roots and conceptualize similar realities, would express human reality in similar ways. See examples 183 and 184 below for the procedure of identifying the process type via the constellation of PR (named TR, Teilnehmerrolle) in German.

(183) Nun bin ich aus zeitlichen Gründen ewig nicht mehr zu den Treffen gegangen.

Step one: How many TR does the process require? Use the re-expression test: "In diesem Prozess des *Gehens* erwarten wir, dass jemand dorthin *geht*". Two TR: someone, i.e. *ich*, and somewhere, i.e. *zu den Treffen*.

Step two: What kind of TR do *ich* and *zu den Treffen* represent? Use the re-expression tests:

- Trägerin: Man kann über X sagen, dass... „Man kann über mich sagen, dass ich zu den Treffen gehe." X = *ich* ist die TR Trägerin.
- Endpunkt: Y (Trägerin) bewegte sich / ging in Richtung X. "Ich bewegte mich zu den Treffen." X = *zu den Treffen* ist die TR Endpunkt.

Step three: What kind of process does the constellation of TR Trägerin and TR Endpunkt represent? Look this up in the list provided with the guidelines. TR Trägerin + process + TR Endpunkt = relational process, directional.

Here is another example to demonstrate the three-step procedure:

(184) *[…] und davor habe ich so eine Angst [davor refers to ,putting on weight'.]*

Step one: How many TR does the process require? Use the re-expression test: "In diesem Prozess des *Angst-habens* erwarten wir, dass jemand vor etwas *Angst hat*". Two TR: someone, i.e. *ich*, and being afraid of something, i.e. *davor*.

Step two: What kind of TR do *ich* and *davor* represent? Use the re-expression tests:

- Empfindende: X hatte ein gutes / schlechtes Gefühl bei dem Gedanken an Y. „Ich hatte ein schlechtes Gefühl bei dem Gedanken daran [,an das Zunehmen']". X = *ich* ist die TR Empfindende.
- Phänomen: Y hatte ein schlechtes Gefühl bei dem Gedanken an X. „Ich hatte ein schlechtes Gefühl bei dem Gedanken daran [,an das Zunehmen']". X = daran / *davor* ist die TR Phänomen.

Step three: What kind of process does the constellation of TR Empfindende and TR Phänomen represent? Look it up in the list. TR Empfindende + Prozess + TR Phänomen = mentaler Prozess des Empfindens (emotional, desiderativ).

Figure 7.2 and 7.3 display the annotation schemes for process type used for the annotation of the English and German newsgroup text corpora. Although all terms have been translated to German for the annotation of the German corpus, in the course of this work only the English terms will be used for ease of understanding.

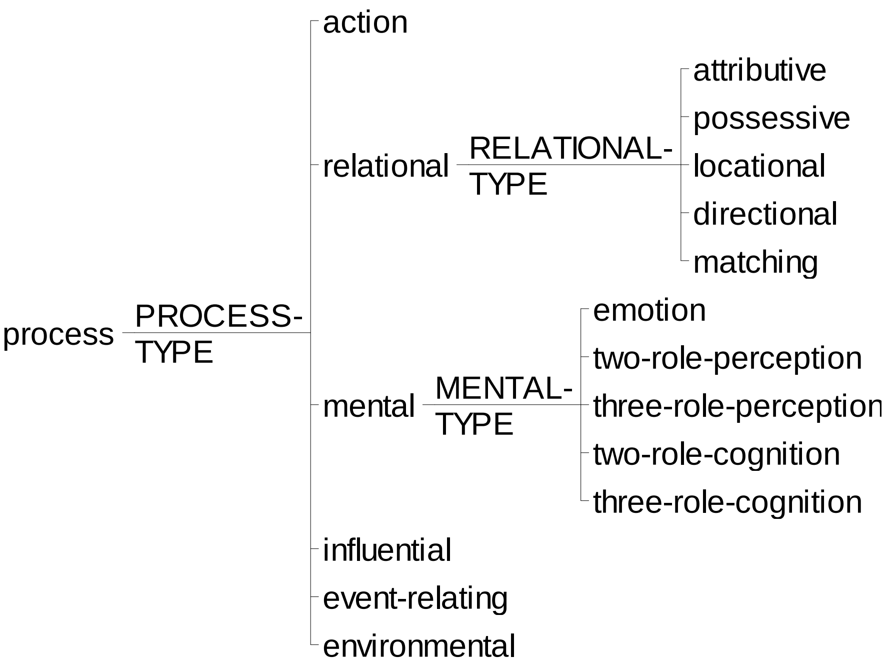

Figure 7.2 The annotation scheme for process types in English

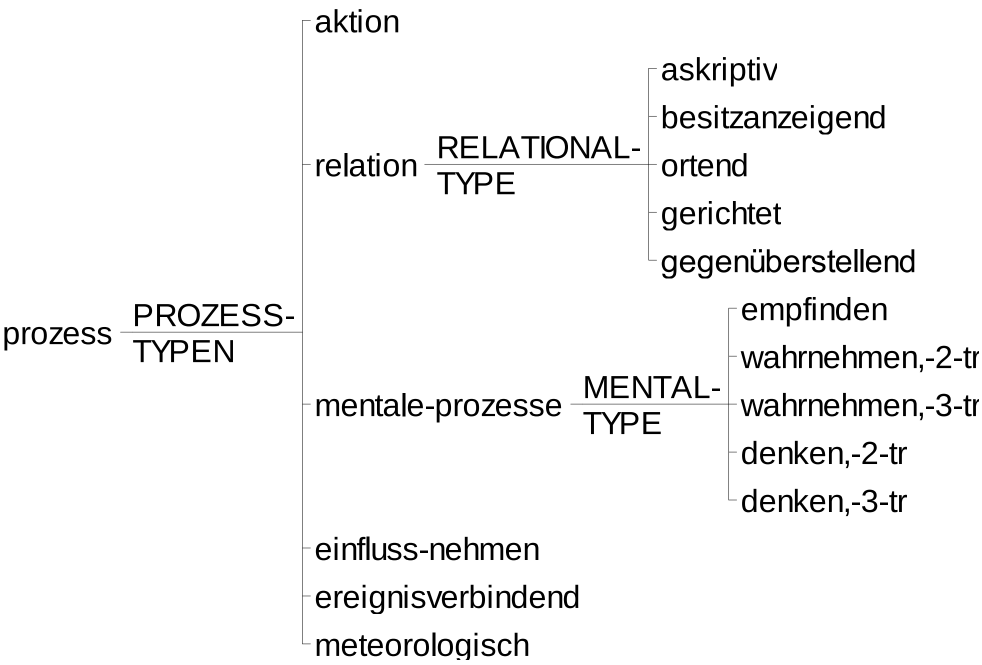

Figure 7.3 The annotation scheme for process types in German

Thus, all participant roles and process types in the EDNA corpus in two languages were annotated with the procedure described above, using the UAM corpus tool (O'Donnell 2008). The results are shown and discussed in the following subchapter.

7.2 Quantitative analysis of process types in EDNA

This subchapter deals with the calculation of statistical significance for the frequency of the different process types in the two EDNA corpora. We start with the main process types, and the null hypothesis is once again that there is no significant variation in the use of different process types. Table 7.1 shows the results of the analysis and calculation. The total number in this set of calculations is the number of processes, not the number of rhemes (see explanation in 5.3).

Feature	EN F	GN F	EN%	GN%	Overuse	χ^2	S
Action process	383	456	25	31	GN	9.82	+
Relational process	572	579	37	39	GN	0.86	-
Mental process	473	382	31	26	EN	6.22	-
Influential process	87	48	6	3	EN	9.68	+
Event-relating process	28	14	2	1	EN	4.09	-
Environmental process	0	0	0	0	-	-	-
Column total	1,543	1,479	100	100			

Table 7.1 Raw and relative numbers, χ^2 and significance for main process types in EN and GN

A look at the relative numbers tells us that in the German newsgroup text corpus, there are 6% more action processes, but 5% fewer mental processes, compared to the English newsgroup text corpus, and that the English corpus has more influential processes. With a total number of 0, the environmental processes have been excluded from the calculation of statistical significance (Gries 2008, 157). The χ^2 value, however, reveals that the differences are rather small, and apart from the frequency of action processes and influential processes, not significant, with the threshold for $p < 0.05$ at 9.49 (df = 4). We might say that the German writers use more action processes, which are processes of the outside world, whereas the English writers, by using more mental processes, reflect more on what goes on in their minds. The frequency differences, however, are only small, again supporting the assumption that English and German writers construct the reality that surrounds them in similar ways, and reflect on it in similar ways.

We shall now investigate the subtypes of relational and mental processes. Table 7.2 shows the results of the calculations for the subtypes of relational processes. The null hypothesis predicts no variation in frequency to a significant degree.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	χ^2	**S**
Relational, attributive	370	319	65	55	EN	4.42	-
Relational, possessive	108	118	19	20	GN	0.33	-
Relational, locational	54	95	9	16	GN	10.79	+
Relational, directional	38	29	7	5	EN	1.32	-
Relational, matching	2	18	0	3	GN	12.61	+
Column total	572	579	100	100			

Table 7.2 Raw and relative numbers, χ^2 and significance for subtypes of relational process type in EN and GN

The null hypothesis can be rejected, there is statistically significant deviation. The number of relational, locational processes is significantly greater in GN, where the threshold for $p < 0.05$ is 9.49 (df = 4). The German writers seem to have a tendency to situate processes in place and time more often than the English writers, to make clear where and when something happened. Also, the difference in frequency of occurrence of relational, matching process types is significant (threshold for $p < 0.05$ is 9.49 (df = 4)), but the raw numbers are rather small.

In table 7.3 below, the focus is on the subtypes of mental processes, with the null hypothesis being that there is no significant variation in frequency of occurrence.

Feature	EN F	GN F	EN%	GN%	Overuse	χ^2	S
Mental, emotion	175	131	37	34	EN	0.43	-
Mental, 2-role percept.	22	20	5	5	-	0.15	-
Mental, 3-role percept.	2	2	0	1	GN	0.05	-
Mental, 2-role cognit.	183	133	39	35	EN	0.86	-
Mental, 3-role cognit.	91	96	19	25	GN	3.35	-
Column total	473	382	100	100			

Table 7.3 Raw and relative numbers, χ^2 and significance for subtypes of mental process type in EN and GN

It is surprising how little variation there is in the use of the different mental process types in the two languages. No mental process subtype is used more often, compared to the other newsgroup text corpus, to any significant degree. What little variation there is in the relative numbers is purely due to random variation. Both the English and German writers seem to construct their inner reality in the same way, no group of writers puts more emphasis, on, for example, mental emotional processes, or mental, three-role cognition (communication) processes. Clearly we cannot reject the null hypothesis this time.

7.3 Constituents of the verbal group in English and German

7.3.1 Constituents of the verbal group – theoretical background

For the sake of completeness, during the annotation of transitivity in the newsgroup corpora we also annotated the constituents that the verbal group was constructed of. Examples 185 and 186 illustrate the verbal group constituents: the auxiliary verb, the main (or lexical) verb, and the process extension (PrEx). In example 185 the verbal group consists of a modal auxiliary; *need to*, and a main verb; *lose*. In example 186 there is a primary auxiliary; *am* in its contracted from *'m*, and a main verb; *seeing*. Both modal and primary auxiliaries have been annotated simply as auxiliary in the English and German corpora, because the difference between the two types of auxiliaries is not relevant for the analysis of transitivity.

(185) *I need to lose weight for health reasons, not just for looks.*

(186) *I'm not seeing him at all for the next couple of weeks.*

Example 187 below from the German corpus shows a primary auxiliary; *bin*, and a main verb; *gegangen*. Furthermore, we see how in German the verbal group is discontinuous, with the auxiliary in second position in the clause and the main verb at the end (if there is no auxiliary, the main verb stands in second position and is then finite). Example 188 shows what in Cardiff Grammar is called a process extension: there is a main verb; *habe*, and the process extension; *so eine Angst*. The process is not the *having*, where the process would require the PR of a Carrier and a Possessed, but *Angst haben*, 'to be afraid', which is a mental process and requires the PR of Emoter and Phenomenon. Therefore, *Angst* is not the PR Possessed, but part of the process; we call this a process extension.

(187) *Nun bin ich aus zeitlichen Gründen ewig nicht mehr zu den Treffen gegangen.*

(188) *[…] und davor habe ich so eine Angst.*

A process extension can be distinguished from a participant role by asking whether the constituent in question would exist even if the process did not take place. Consider the example *I may not be eating as much rubbish*. The *rubbish* is a participant role; the *rubbish* would exist even if *I* would not eat it. In the clause *I fell instantly in love*, however, the *love* would not exist without *I* feeling it (or falling into it). Therefore the *love* is a process extension and part of the process, not a participant role.

Figure 7.4 shows the annotation scheme for the constituents of the English verbal group, and figure 7.5 shows the German equivalent, which is no more than a translation of the English scheme, since verbal groups are constructed of the same constituents in both languages.

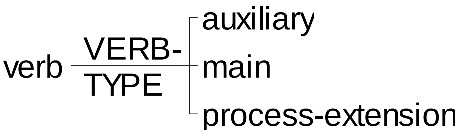

Figure 7.4 The annotation scheme for the verbal group in English

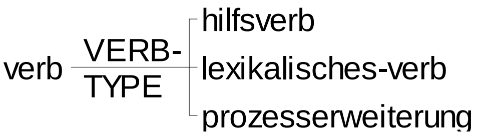

Figure 7.5 The annotation scheme for the verbal group in German

7.3.2 Quantitative analysis of the constituents of the verbal group

In table 7.4 below we see the raw and relative numbers and the tests of statistical significance in the two newsgroup text corpora. The null hypothesis states that there is no difference in frequency of use of the different constituents of the verbal group.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	χ^2	**S**
Auxiliary	520	436	26	20	EN	12.36	+ +
Lexical verb	1,498	1,431	73	67	EN	6.94	+
Process extension	21	280	1	13	GN	209.84	+ + +
Column total	2,039	2,147	100	100			

Table 7.4 Raw and relative numbers, χ^2 and significance for constituents of the verbal group in EN and GN, including process extensions

It seems that there are significantly more auxiliary verbs in the English corpus, with the threshold for $p < 0.01$ at 9.21 (df = 2). These may be either modal or primary auxiliary verbs. There also seem to be significantly more lexical verbs in the English corpus, the threshold for $p < 0.05$ is at 5.99 (df = 2). The reason for the higher percentage of both auxiliary and lexical verbs in EN, however, can be found in the high number of process extensions in the German newsgroup corpus. Due to the high number of process extensions in GN, the total number and percentage of the other two options are necessarily smaller. GN has a significantly larger amount of process extensions, with $p < 0.001$ at 13.82 (df= 2). Table 7.5 below excludes the process extensions from the calculations and reveals that the difference between the two corpora concerning the frequency of auxiliary and lexical verbs is not significant.

Feature	EN F	GN F	EN%	GN%	Overuse	χ^2	S
Auxiliary	520	436	26	23	EN	2.30	-
Lexical verb	1,498	1,431	74	77	GN	0.75	-
Column total	2,018	1,867	100	100			

Table 7.5 Raw and relative numbers, χ^2 and significance for constituents of the verbal group in EN and GN, excluding process extensions

The English and German newsgroup texts contain roughly the same number of auxiliary and lexical verbs (and thus processes), but GN has significantly more processes made up of a lexical verb and a process extension.

7.3.3 Qualitative analysis of the constituents of the verbal group

The results call for a closer look at the process extensions. In both languages, the particles or prepositions that accompany a phrasal or prepositional verb make up a part of the process extensions. A second group are the nominal groups that are strongly connected to the verb, forming a process together. In German traditional grammar these constructions are called 'Funktionsverbgefüge' (Götze and Hess-Lüttich 1999, 424). A third large group of words which are annotated as process extension are the reflexive pronouns, i.e. *mir, mich, dir, dich, sich, uns, euch* in GN. Figures 7.6 and 7.7 below show the different types of process extensions and some examples.

Type	Examples
Phrasal / prepositional verb	*Let go, substitute X for Y, get out, keep down, be like, put back on, bring up, invite X in, go out, vent out, greet X with, equate X with*
Nominal group	*Have in mind, fall in love, give one more try, make a choice for, be in love with*

Figure 7.6 Examples of process extensions in the English newsgroup corpus

Type	**Examples**
Phrasal / prepositional verb	*kreisen um, mit X zusammen sein, gegen X verstoßen*
Nominal group	*in den Griff bekommen, unter Kontrolle haben, unter Druck setzen, in den Arm nehmen*
Reflexive pronoun	*in sich hinein fressen, sich streiten, zu sich nehmen, sich zurückziehen, sich wünschen, sich X durchlesen, sich X einfallen lassen, sich X fühlen, sich umstellen*

Figure 7.7 Examples of process extensions in the German newsgroup corpus

All these reflexive pronouns which are annotated as process extensions in GN, but not in EN, raise the question of why reflexive pronouns are not annotated as process extensions in EN. Are there less reflexive pronouns in EN? Are they annotated as participant roles (PR) instead of process extensions (PrEx)? A string-based search in the UAM corpus tool reveals the following numbers, see table 7.6.

Reflexive pronoun EN	**Total number**	**of which PrEx**	**Reflexive pronoun GN**	**Total number**	**of which PrEx**
myself	15	-	*mir /mich*	157 / 134	14 / 32
yourself	1	-	*dir / dich*	- / 4	- / -
him-/herself	1/-	-	*sich*	58	39
ourselves	-	-	*uns*	22	5
yourselves	-	-	*euch*	11	1
themselves	-	-			
Total number	17	-		386	91

Table 7.6 Reflexive pronouns in the English and German newsgroup corpus

In fact, we find considerably more reflexive pronouns in the German corpus, a small number of which serve as process extension, while the rest serve as participant roles. Examples 189 and 190 demonstrate the difference. The lexical verb *lieben* in example 189 is a mental process and requires two PR, an Emoter (*man*), and a Phenomenon (*sich*). The lexical verb *anlehnen* in example 190, however, is an action process and requires only one PR, an actor (*sie*), thus, *sich* is a reflexive pronoun and a process extension.

(189) *PR Phenomenon: Ist das nicht normal, wenn man sich liebt?*

(190) *PrEx: […] und sie sich an meine Schulter anlehnen will.*

We find very few reflexive pronouns in the English corpus, and the few that are actually there are part of a PR, not a PrEx. Consider examples 191 and 192:

(191) *PR Affected: I cannot control myself.*

(192) *Part of PR Carrier: Starving yourself doesn't seem silly.*

There is to be a typological difference in the use of reflexive pronouns between English and German (König and Gast 2007, 141). For future SFL annotations of German corpora it might be helpful to point this difference out to the annotators. Reflexive pronouns account for 32% out of the 280 PrEx in the German corpus. The other 68% of PrEx are either particles or prepositions which accompany phrasal or prepositional verbs or 'Funktionsverbgefüge', a verb plus nominal group.

7.4 Quantitative analysis of major participant roles

Following the test of statistical significance of the frequency of process types in EDNA, we can also study the frequency of the major participant roles (PR) involved in these processes. This calculation may not by entirely meaningful, because the frequency of process types controls the frequency of participant roles. However, for the sake of comprehensiveness, it shall be included. The first of the three main process types is the action process, the related PR are investigated in table 7.7. The second are the relational processes, see table 7.8, followed by the PR involved in the third main process type to be investigated, mental processes, in table 7.9. Only those PR that appear at least 100 times in at least one of the two corpora are taken into consideration here. The row named 'other' includes all other PR in every process in EDNA and thus the number is the same for all three sets of calculations. For all three calculations, the null hypothesis is that there is no significant variation in the use of PR.

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	χ^2	**S**
Agent	460	470	17	18	GN	1.16	-
Affected	155	144	6	5	EN	0.05	-
Other	2,168	2,036	78	77	-	0.20	-
Column total	2,783	2,650	100	100			

Table 7.7 Raw and relative numbers, χ^2 and significance for the main PR in action processes

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	χ^2	**S**
Carrier	481	466	17	18	GN	0.07	-
Attribute	366	315	13	12	EN	1.73	-
Possessed	84	103	3	4	GN	2.97	-
Other	1,852	1,766	67	67	-	0.00	-
Column total	2,783	2,650	100	100			

Table 7.8 Raw and relative numbers, χ^2 and significance for the main PR in relational processes

Feature	**EN F**	**GN F**	**EN%**	**GN%**	**Overuse**	χ^2	**S**
Emoter	148	125	5	5	-	0.98	-
Cognizant	151	126	5	5	-	1.20	-
Phenomenon	541	469	19	18	-	2.21	-
Other	1,943	1,930	71	72	-	1.73	-
Column total	2,783	2,650	100	100			

Table 7.9 Raw and relative numbers, χ^2 and significance for the main PR in mental processes

The null hypothesis cannot be rejected for any of the three sets, as there is no significant divergence in the frequency of any of the PR investigated here. On the contrary, the numbers are strikingly similar, comparing the English and German newsgroup texts. It would be difficult for any PR to occur significantly more frequently, since all PR are restrained by the related process type. What these tests do show us, however, is the ranking of the most frequent participant roles in the newsgroup corpus. Note that the PR Phenomenon is used mainly in mental processes, but also in other processes, i.e. influential and event-relating processes, which are not frequent in EDNA.

1. PR Phenomenon EN 19%, GN 18%
2. PR Agent EN 17%, GN 18%
3. PR Carrier EN 17%, GN 18%
4. PR Attribute EN 13%, GN 12%
5. PR Affected EN 6%, GN 5%
6. PR Emoter EN 5%, GN 5%
7. PR Cognizant EN 5%, GN 5%

7.4.1 Summary

Figure 7.8 below summarizes the test of statistical significance for the system of transitivity.

System	**Divergence**	**Divergent feature**	**Divergent corpus**
Main process types	Significant	Action processes Influential processes	GN + EN +
Relational process subtypes	Significant	Relational, locational process Relational, matching process	GN + GN +
Mental process subtypes	Not significant	-	-
Verbal group constituents	Significant	Process extensions	GN + + +
Frequency of PR	Not significant	-	-

Figure 7.8 Summary of tests of statistical significance of the system of transitivity

Following the quantitative analysis of process types and participant roles in the EDNA corpus, we will study the lexical items that function as process or PR in the EDNA corpus in the next subchapter.

7.5 Analysis of lexical items used as process types

The following subchapters are dedicated to the analysis, first, of the lexical verbs that realize the five different process types, excluding environmental processes which did not occur in EDNA, and second, of the nominal groups that realize the most prominent participant roles (PR) in the five process types: PR Agent, PR Affected, PR Carrier, PR Attribute, PR Possessed, PR Emoter and PR Cognizant. The most frequent PR, PR Phenomenon, has been excluded from the study of lexical realization because a PR Phenomenon usually is an entire subordinate (projected) clause. Therefore, no single head of a group can be identified, and no clause occurs more than once.

The first table (7.10) in this subchapter displays the ten most frequently used lexical verbs that realize action processes in EN and GN.

N	Word EN	F	%	Word GN	F	%
1	DO	24	6.33	ESSEN	23	5.04
2	EAT	15	3.96	HELFEN	13	2.85
3	LEAVE	12	3.17	TUN	12	2.63
4	SEE	9	2.37	ZUNEHMEN	10	2.19
5	HELP	8	2.11	ABNEHMEN	9	1.97
6	MEET	7	1.85	MACHEN	8	1.75
7	PURGE	7	1.85	ANFANGEN	6	1.32
8	START	7	1.85	AUFHÖREN	5	1.10
9	HAPPEN	6	1.85	GEHEN	5	1.10
10	HAVE	6	1.85	REDENMIT	5	1.10
		101	26.37		96	21.06
	Others	282	73.63	Others	360	78.94
	Total	383	100	Total	456	100

Table 7.10 Lexical verbs realizing action processes

In the English newsgroup texts, the unspecific lexical verb *do* is the most frequent one to express an action process, followed by *eat, leave, see* (~ meet someone) and *help*. In the German newsgroup texts, we find as the most frequent lexical verbs in action processes *essen, helfen, tun, abnehmen* and *zunehmen*. The lexical verbs point to the topic of the discourses: eating, losing and gaining weight, seeing and leaving someone, and helping, or rather asking for help. But even the most frequent lexical verb accounts for no more than 6% in EN and 5% in GN of all lexical verbs in action processes. The 10 most frequent lexical verbs in EN make up only 26% of all lexical verbs in action processes, in GN, the 10 most frequent lexical verbs account for only 21% of all the lexical verbs in action processes. These results indicate that a great variety of lexical verbs is used to realize action processes.

The most frequent lexical verbs in relational processes, shown in table 7.11 below for EN and GN, paint a different picture of relational processes, compared to the action processes.

N	Word EN	F	%	Word GN	F	%
1	BE	330	58.00	SEIN	239	43.22
2	HAVE	58	10.19	HABEN	67	12.12
3	FEEL	22	3.87	GEHEN	24	4.34
4	GET	13	2.28	WERDEN	22	3.98
5	GO	13	2.28	GEBEN	19	3.44
6	LOSE	13	2.28	SICHFÜHLEN	13	2.35
7	LIVE	7	1.23	ZUSAMMENSEIN	8	1.45
8	BECOME	6	1.05	FINDEN	7	1.27
9	MAKE	6	1.07	BEKOMMEN	6	1.08
10	GOT	5	0.88	ZUSAMMENSEINMIT	5	0.90
		473	82.69		410	70.81
	Others	99	17.31	Others	169	29.19
	Total	572	100	Total	579	100

Table 7.11 Lexical verbs realizing relational processes

Not surprisingly, *be* and *have* are the most frequent lexical verbs in EN realizing relational processes, with *be* alone accounting for 58% of all lexical verbs in relational processes. Together with *have,* these two lexical verbs are found in 68% of all relational processes. In GN, *sein*, the German equivalent of *be*, is also the most frequent lexical verb in a relational process, followed by *haben;* these two lexical verbs together account for 55% of all lexical verbs in relational processes. All ten most frequent lexical verbs in EN make up as much as 83% of all lexical verbs in relational processes of the different types. In GN, the variety of lexical verbs in relational processes is slightly greater, with the 10 most frequent ones covering 70% of all relational processes. Thus, it is easier to predict which lexical verbs realize a relational process than it is to predict the lexical verbs in action processes, where the variety is much greater. Most of the lexical verbs in relational processes in EN and GN express the same meaning; they convey the basic idea of what a relational process is: *be – sein, have – haben, feel X / feel like X– sich X fühlen, get / got – werden / bekommen, become – werden, go – gehen, lose – verlieren, find – finden, give – geben*.

Next, the most frequent lexical verbs found in mental processes in the newsgroup texts are shown in table 7.12.

N	Word EN	F	%	Word GN	F	%
1	KNOW	45	9.51	WISSEN	52	13.61
2	WANTTO+Inf	38	8.03	SAGEN	42	10.99
3	THINK	34	7.19	LIEBEN	19	4.97
4	SAY	25	5.29	DENKEN	13	3.40
5	WANT+NG	22	4.65	BRAUCHEN	12	3.14
6	TELL	21	4.44	WOLLEN	12	3.14
7	LOVE	20	4.23	MEINEN	9	2.36
8	FEEL	15	3.17	GLAUBEN	8	2.09
9	ASK	13	2.75	KENNEN	6	1.57
10	FEELLIKE	13	2.75	MERKEN	6	1.57
		246	52.00		179	46.86
	Others	227	48.00	Others	203	53.14
	Total	473	100	Total	382	100

Table 7.12 Lexical verbs realizing mental processes

The most frequent lexical verbs in mental processes are a little more frequent than those in action processes, with *know, want to* + infinitive and *think* in EN each accounting for about 10% of all lexical verbs, and *wissen* and *sagen* also accounting for about 10% of all lexical verbs in GN. With regard to the variety of different lexical verbs realizing a certain process type, mental processes are in between action and relation processes. The top 10 of most frequent lexical verbs in EN accounts for 52% in EN and 47% in GN, thus, other lexical verbs account for roughly half of the lexical verbs.

Readers will have noticed that the lexical verb *feel* appears in both the table of relational and mental processes. We need to look at some examples to find an explanation why this is so; see examples 193 to 195 and 196 to 198 for *feel* and *feel like* in relational processes, and examples 199 to 201 and 202 to 204 for *feel* and *feel like* in mental processes.

Relational *feel*:

(193) *I feel awful*

(194) *[…] and then feel guilty*

(195) *I feel young, sexy and amazing*

Relational *feel like*:

(196) *I feel like an old maid*

(197) *I feel like such a fool*

(198) *I feel like a junkie hurting for a fix*

Mental *feel*:

(199) *[…] and I really feel that the sparkle is not there anymore*

(200) *[…] because I feel we are doomed to fail*

(201) *I really feel that I am there in so many ways*

Mental *feel like*:

(202) *I feel like I can't be productive if I am full*

(203) *I feel like if I leave her I'm a self-centered jerk*

(204) *I feel like I'm faking it staying with her*

We see that in the relational processes, *feel* is followed by an adjective, e.g. *guilty, young, amazing,* thus the person feeling *guilty* or *young* ascribes the attribute to herself, like saying 'I am disgusting' or 'I am young', but with a degree of subjectivity added: 'It is only I who thinks I am disgusting, not everybody'. In this sense, *feel* + adjective is a relational process with a pinch of modality thrown in. The phrasal verb *feel like* in relational processes is found preceding a nominal phrase, e.g. *an old maid, such a fool*, thus the writer ascribes the attribute to herself, adding a bit of modality: 'I am an old maid, I think' or 'I am such a fool, but that is only my opinion'.

The lexical verb *feel* behaves differently in a mental process, here it is followed by a PR Phenomenon, i.e. by a subordinate finite clause, e.g. *that the sparkle is not there anymore, we are doomed to fail*. *Feel* in these mental two-role cognition clauses is a synonym for *think*. *Feel like* in the mental two-role cognition processes behaves similar to *feel*: It is followed by a PR Phenomenon, i.e. a subordinate finite clause, therefore we can consider *feel like* to be a synonym of *think* as well when followed by a subordinate finite clause.

The next table (7.13) in this subchapter displays the most frequently used lexical verbs in influential processes used in the EDNA corpus.

N	**Word EN**	**F**	**%**	**Word GN**	**F**	**%**
1	START	17	18.68	VERSUCHEN	11	22.92
2	MAKE	12	13.19	ANFANGEN	6	12.50
3	TRY	10	10.99	AUFHÖREN	5	10.42
4	STOP	6	6.59	SCHAFFEN	4	8.33
5	PLAN	5	5.49	SICHZWINGEN	2	4.17
6	DO	4	4.40	TUN	2	4.17
7	QUIT	3	3.30	VERBIETEN	2	4.17
8	BEGIN	2	2.20	AUFGEBEN	1	2.08
9	END	2	2.20			
10	ENDUP	2	2.20			
		66	72.54		34	70.84
	Others	21	27.46	Others	14	29.16
	Total	87	100	Total	48	100

Table 7.13 Lexical verbs realizing influential processes

We find the lexical verb *start* to be the most frequently used one to realize influential processes in EN, with 19% of all lexical verbs in influential processes, followed by *make* and *try* with about 10% each in EN. In GN, we have *versuchen* as the most frequent lexical verb in influential processes with 23%, and *anfangen, aufhören* and *schaffen* with about 10% each. The English part of EDNA contains significantly more influential processes. The German part has few of those, and only the top four occur more often than just once or twice.

The last types of processes, the event-relating processes, are few in numbers, but for the sake of comprehensiveness table 7.14 displays the lexical verbs used to realize this process type.

N	Word EN	F	%	Word GN	F	%
1	BE	20	71.43	SEIN	6	42.86
2	BELIKE	2	7.14	ANGEHEN	1	7.14
3	ENTAIL	2	7.14	MITETWASGEHEN	1	7.14
4	FOLLOW	1	3.57	REICHEN	1	7.14
5	INVOLVE	1	3.57	SICHABWECHSELNMIT	1	7.14
6	MAKE	1	3.57	SICHDREHENUM	1	7.14
7	RESULTIN	1	3.57	UMETWASKREISEN	1	7.14
				VONETWASKOMMEN	1	7.14
				VORSICHHABEN	1	7.14
		28	100		14	100
	Others	0	0	Others	0	0
	Total	28	100	Total	14	100

Table 7.14 Lexical verbs realizing event-relating processes

The lexical verbs *be* in EN and *sein* in GN are the most frequently used ones in event-relating processes, which demonstrates the semantic proximity with relational processes. Examples of event-relating processes are given in 205 to 208 for EN and 209 to 212 for GN. The difference to relational processes is that the second PR, which usually follows the verbal group, is a PR Phenomenon, not a PR Attribute or Possessed or one of the other possible PR in a relational process. In event-relating processes, the second PR has the characteristics of an event, rather than an object.

(205) One of the main stresses is [the cookie marathon]

(206) The reason I say this is [because I had a hunch and read some of her email]

(207) The sex we had was like [making love with someone you cared about]

(208) This entails [spending time learning to read each other's communication]

(209) *[Eine Beziehung wegzuschmeißen] ist doch zu einfach*

(210) *Das Ende vom Lied war, [dass ich bulimiekrank wurde]*

(211) *[…] die sich abwechseln [mit Tagen, in denen ich alles unter Kontrolle habe]*

(212) *Alles dreht sich nur [ums Essen und ums Zunehmen]*

To conclude this chapter, figure 7.9 gives a graphical display of the results of the analysis of the most frequently used lexical verbs in the five different process types in the English and German newsgroup texts. The most frequent lexical items are those top ten lexical items shown in tables 7.10 to 7.19 above.

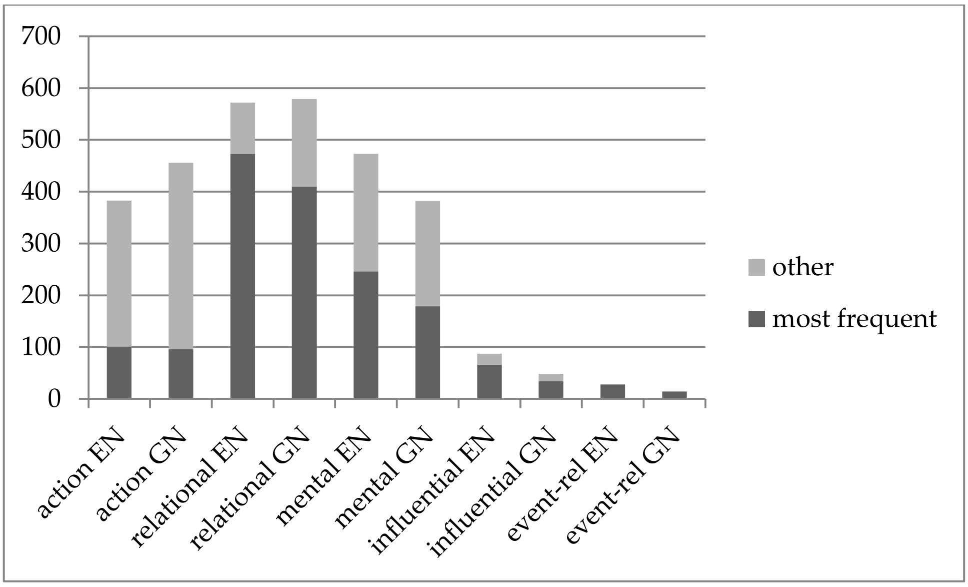

Figure 7.9 Most frequently used lexical verbs in the EDNA corpus (raw numbers)

The figure summarizes what was said in the previous sections: In both EN and GN, the most frequent processes are relational processes. In EN, these are followed by mental processes and action processes, but in GN, action processes are more frequent than mental processes. The action processes are realized by the greatest variety of different lexical verbs, followed by mental processes to a smaller extent. The German newsgroup texts have a greater number of different lexical verbs in any of the main process types compared to the English newsgroup texts. This fact can be seen even better in figure 7.10, which shows the percentages of most frequently used and other lexical verbs.

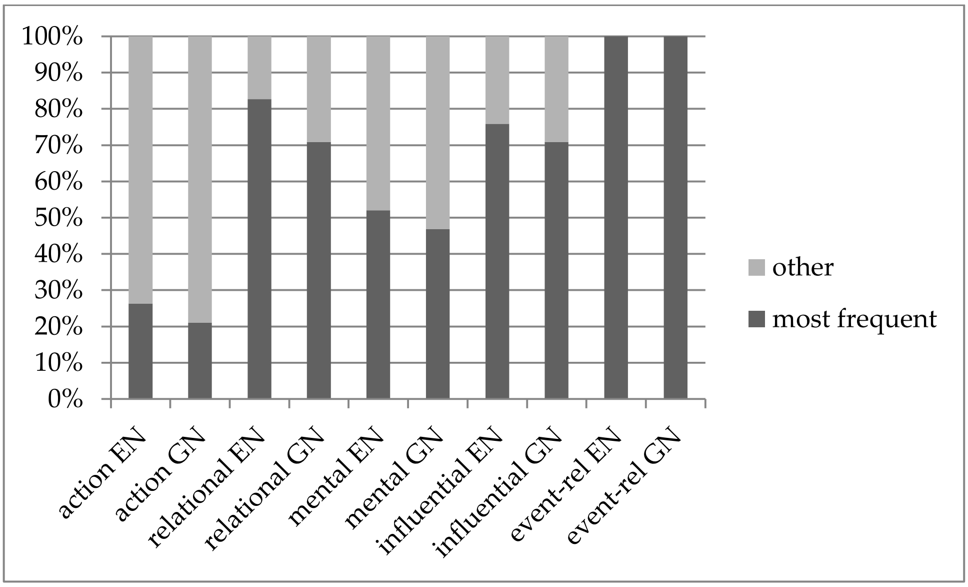

Figure 7.10 Most frequently used and other lexical verbs in the EDNA corpus (100%)

In this subchapter, we focused on the analysis of the lexical verbs which realize the different process types in EDNA. We will now look at the nominal groups which realize the main participant roles in these processes.

7.6 Analysis of lexical items used as participant roles

In what follows, we concentrate on the nominal groups realizing the most prominent participant roles (PR) in the three main process types: PR Agent and PR Affected in action processes; PR Carrier, PR Attribute and PR Possessed in relational processes (attributive and possessive ones), and finally PR Emoter and PR Cognizant in mental processes (of emotion and two/three-role cognition).

7.6.1 Participant roles in action processes

The most frequent participant roles (PR) that are involved in action processes, i.e. PR Agent and PR Affected, are the first to be investigated. Table 7.20 displays the heads of nominal groups functioning as PR Agent in (mostly) action processes in the English newsgroup texts collected in EDNA. Note that there are more PR Agents (460) than there are action processes (383) in EN. The reason for this is that PR Agents can also be involved in relational, influential and, most importantly, mental three-role cognition (communication) processes. Thus, 77 (16% of all) PR Agents come from processes other than action processes in the English texts, but are included here in the list. 16% should not distort the big picture too much. The same is true for GN, where there are 470 PR Agents, but only 456 action processes. Thus, 14 (3% of all) PR Agents appear in processes other than action in the German texts.

N	Word EN	F	%	Word GN	F	%
1	I	232	50.43	ICH	242	51.49
2	WE	41	8.91	ER	49	10.43
3	SHE	40	8.7	SIE	28	5.96
4	HE	38	8.26	WIR	21	4.47
5	IT	16	3.48	IHR	13	2.77
6	YOU	12	2.61	DAS	10	2.13
7	THEY	10	2.17	ES	10	2.13
8	WHAT	6	1.3	DIE	9	1.91
9	FRIENDS	5	1.09	MAN	8	1.7
10	THAT	5	1.09	ALLES	5	1.06
		405	88.04		395	84.04
	Other	55	11.96	Other	75	15.96
	Total	460	100	Total	470	100

Table 7.15 Heads of nominal groups functioning as PR Agent

The most frequent heads of nominal groups in PR Agent come as no surprise after our study of unmarked and marked topical themes in chapter 6.3. In both corpora, personal pronouns are most frequently used as PR Agent in action processes. In EN, the pronoun *I* accounts for 50% of all PR Agents, followed by *we, she, he* with each accounting for around 8%. Similarly, in GN, the pronoun *ich* accounts for 51% of all PR Agents, followed by *er* making up 10% and *sie, wir* accounting for about 5% each. Thus, the writers talk mostly about themselves doing something (to/with someone). Out of all action processes in EN, 44% of the action processes have no second participant role, for example in clauses like *I exercise regularly*. 40% of the action processes have a PR Agent plus a PR Affected, and the remaining 16% have a PR Agent and one of the other possible PR; Carrier, Created, Range or Manner. In GN, 50% of the action processes involve only the PR Agent. This shows that the German writers use more intransitive lexical verbs. The PR Affected is found in 32% of the action processes, and the remaining 18% of action processes involve one of the other possible PR.

The next table (7.16) displays the most frequent heads of nominal groups in PR Affected as they are used in EN and GN. The PR Affected is used in action and influential processes.

N	**Word EN**	**F**	**%**	**Word GN**	**F**	**%**
1	ME	25	16.03	MICH	25	17.36
2	HIM	10	6.41	MIR	23	15.97
3	I	10	6.41	SIE	8	5.56
4	IT	8	5.13	ICH	7	4.86
5	HER	5	3.21	UNS	7	4.86
6	MYSELF	5	3.21	IHN	5	3.47
7	SALAD	5	3.21	SICH	5	3.47
8	EACHOTHER	4	2.56	DAS	3	2.08
9	FRIEND	4	2.56	IHM	3	2.08
10	FOOD	3	1.92	NICHTS	3	2.08
		79	50.97		89	61.81
	Other	76	49.03	Other	55	38.19
	Total	155	100	Total	144	100

Table 7.16 Heads of nominal groups functioning as PR Affected

We see that in EN *me, I, myself* account for 25% of all PR Affected, in GN, the pronouns *mich, mir, ich* account for 38% of all PR Affected. Apart from 3rd person pronouns *him* and *her*, some nominal group heads indicate the topic of the discourse in the English newsgroup texts: *salad, friend, food*. In the German newsgroup texts, there are only pronouns as PR Affected. This seems to make the German texts vaguer than the English texts. Most PR Affected in EN and GN must have been mentioned before, or are easily retrievable from the context; otherwise the writers could not use only pronouns to refer to someone / something. By looking at the PR Affected alone, readers would not know who is being affected by the action.

7.6.2 Participant roles in relational processes

In this section, we turn to the second main process type, the relational process, and the three most frequent PR in this process type, the PR Carrier, PR Attribute and PR Possessed. Note that there are only 480 PR Carrier in EN, even though there are 572 relational processes. Apart from PR Carrier, 16% of the relational processes have a PR Affected-Carrier or PR Agent-Carrier as the main participant. In GN, there are 466 PR Carrier in 572 relational processes, thus, 18% PR Affected-Carrier or PR Agent-Carrier in the relational processes. A PR Carrier can also occur in an event-relating process, but these are rare in EDNA. See table 7.17 for the results of the nominal heads which realize the PR Carrier.

N	**Word EN**	**F**	%	**Word GN**	**F**	%
1	I	187	38.96	ICH	168	36.05
2	IT	51	10.63	ES	34	7.3
3	WE	25	5.21	DAS	27	5.79
4	HE	22	4.58	ER	26	5.58
5	SHE	22	4.58	MIR	19	4.08
6	THIS	12	2.5	SIE	12	2.58
7	THAT	10	2.08	WIR	12	2.58
8	YOU	9	1.88	ALLES	10	2.15
9	WHO	8	1.67	DIE	7	1.5
10	PROBLEM	6	1.25	DIES	7	1.5
		352	73.33		322	69.10
	Other	128	26.67	Other	144	30.90
	Total	480	100	Total	466	100

Table 7.17 Heads of nominal groups functioning as PR Carrier

The results from this investigation show us that the PR Carrier in the English and German newsgroup texts are almost exclusively pronouns, thus, they create exophoric or endophoric reference. Relational processes, which either attribute a quality, describe a possession or locate the process in time or space, are used once the context is made clear in the surrounding discourse, or is clear from the situation. The PR Carrier tells the reader nothing about the topic of the discourse, eating disorders or relationship problems. And again, as with the action processes, the 1st person singular pronoun is used most often, i.e. writers themselves are the carriers of attributes or possessions or locate themselves in space or time.

In the following, let us have a look at the PR Attribute, which is typically associated with a PR Carrier in relational, attributive processes, i.e., processes of being something. Please see table 7.18 for the most frequent heads of nominal or adjectival groups in EDNA.

N	Word EN	F	%	Word GN	F	%
1	XYEARSOLD	19	5.19	XJAHREALT	12	3.81
2	THIS	9	2.46	GUT	7	2.22
3	CONFUSED	7	1.91	NORMAL	6	1.9
4	NEW	7	1.91	SCHLECHT	5	1.59
5	XPOUNDS	7	1.91	BESSER	4	1.27
6	MARRIED	6	1.64	MENSCH	4	1.27
7	FULL	5	1.37	WEG	4	1.27
8	GREAT	5	1.37	XKILOSCHWER	4	1.27
9	NICE	5	1.37	ANDERS	3	0.95
10	TOGETHER	5	1.37	DICK	3	0.95
		75	20.50		52	16.51
	Others	291	79.50	Other	263	83.49
	Total	366	100	Total	315	100

Table 7.18 Heads of nominal groups functioning as PR Attribute

The most noticeable conclusion we can draw from the table is probably the fact that there is a great number of different attributes used in relational processes; even the most frequent adjectival group *X years old* accounts for no more than 5% of all PR Attribute in EN and *X Jahre alt* for only 4% in GN. Apart from the 10 most frequent lexical items used as PR Attribute, there are another 80% (EN) / 84% (GN) of other adjectives or nouns realizing a PR Attribute.

We saw that the PR Carrier is most often a pronoun of some sort. The PR Attribute has more semantic weight. Writers finally make clear what they are talking about, and they talk about a lot of different things. Interestingly, the English writers use only *nice* and *great* to openly assign an evaluation to a PR

Carrier, whereas the German writers use the whole range: *gut, normal, schlecht, besser*.

The next most frequent PR after PR Carrier and PR Attribute in a relational process is the PR Possessed. This occurs in relational, possessive processes where writers state that they have something, or that they do not have something, see table 7.19.

N	Word EN	F	%	Word GN	F	%
1	CHILDREN	5	5.95	PROBLEME	6	5.83
2	PROBLEM	5	5.95	PROBLEM	5	4.85
3	EATINGDISORDERS	3	3.57	XKILO	5	4.85
4	FRIENDS	3	3.57	KRAFT	4	3.88
5	TIME	3	3.57	ZEIT	4	3.88
6	ADVICE	2	2.38	DAS	3	2.91
7	ANYTHING	2	2.38	FREUND	3	2.91
8	FEELINGS	2	2.38	MAGERSUCHT	3	2.91
9	HEART	2	2.38	ES	2	1.94
10	SOMETHING	2	2.38	FREUNDIN	2	1.94
		34	34.52		37	35.92
	Others	55	65.48	Other	66	64.08
	Total	84	100	Total	103	100

Table 7.19 Heads of nominal groups functioning as PR Possessed

In the English newsgroup texts, writers have *children, a problem, eating disorders*, and the majority of the most frequent nominal groups are nouns. We only find *anything* and *something* in the top 10 to refer to information given in the context. In the German newsgroup texts, writers also have a *Problem, Probleme, X Kilo*, and they do not have *Kraft* or *Zeit*. There are two pronouns among the more frequent groups realizing a PR Possessed, i.e. *das, es*. Both groups of writers speak about time, usually saying they do not have time. The PR Possessed indicates clearly the topic of the discourse in the newsgroups. The lexical items used to indicate a possession (or the lack thereof) vary greatly, there are many different nominal groups in PR Possessed. This variety is what PR Possessed and PR Attribute have in common. We may conclude that in relational processes, the second PR, i.e. the one following the finite verb, is much more telling than the first PR, i.e. the PR Carrier. This comes as no surprise, knowing that new information is usually put in the rheme of a clause. It is in the rheme that writers say how they are (PR Attribute), e.g. *confused, married,* and what they have (PR Possessed), e.g. *children, problems, eating disorders.*

7.6.3 Participant roles in mental processes

The third set of participant roles to look at are the ones involved in mental processes; the PR Emoter in mental, emotional processes and the PR Cognizant in mental two- or three-role cognition processes. Hardly any mental, perception processes occur in the EDNA corpus, and therefore hardly any PR Perceiver. In consequence, the PR Perceiver has been excluded from this study. The PR Phenomenon is the third most frequent PR in the EDNA corpus after PR Agent and PR Carrier; it appears in more than one process, i.e. in action, mental, influential and event-relating processes. A PR Phenomenon more often than not is an entire subordinate (projected) clause. Therefore, no single head of a group can be identified, and no clause occurs more than once. For these reasons, the PR Phenomenon is excluded from an investigation of most frequently occurring head of groups realizing a PR. But let us turn now to the PR which we can actually study. Table 7.20 displays the most frequent heads of nominal groups used as PR Emoter in the English and German newsgroup texts.

N	Word EN	F	%	Word GN	F	%
1	I	105	70.95	ICH	77	61.6
2	HE	12	8.11	ER	12	9.6
3	SHE	9	6.08	SIE	10	8.0
4	YOU	9	6.08	MIR	8	6.4
5	ME	5	3.38	MAN	5	4.0
6	HUSBAND	2	1.35	MICH	4	3.2
7	WE	2	1.35	DER	3	2.4
8	ANYONE	1	0.68	IHM	1	0.8
9	EVERYONE	1	0.68	IHN	1	0.8
10	PEOPLE	1	0.68	IHRERSEITS	1	0.8
11	WIFE	1	0.68	SCHWESTER	1	0.8
				WER	1	0.8
				WIR	1	0.8
		148	100		125	100
	Total	148	100	Total	125	100

Table 7.20 Heads of nominal groups functioning as PR Emoter

The PR Emoter is different from the other participant roles. There are only 11 different nominal groups in EN and 13 different ones in GN which constitute the entirety of this PR. All of these 11 / 13 nominal groups describe human beings with a consciousness capable of feeling something (but then, the discourse is not about animals or plants, but about human problems). Most of the PR Emoter are realized by personal pronouns: In EN we find *I* and *me* (74% of all), *he, she, you, we* as well as *anyone* and *everyone*. Three nouns refer to human beings, too: *husband, wife, people*. In GN, *ich, mir, mich* account for 71% of all PR Emoter, followed by personal pronouns *er, sie, ihm, ihn, ihrerseits, wir* and other types of pronoun; *wer, der, man*. Only one noun in the German texts functions as PR Emoter, the noun *Schwester*.

The second main PR in mental processes is that of PR Cognizant. This one appears in mental two-role cognition processes, i.e. someone knows something, and mental three-role cognition processes (communication processes), i.e. someone tells someone something (someone knows something as a result). Table 7.21 below presents the heads of nominal groups in PR Cognizant.

N	**Word EN**	F	%	**Word GN**	F	%
1	I	115	76.16	ICH	97	76.98
2	SHE	12	7.95	ER	7	5.56
3	YOU	7	4.64	SIE	5	3.97
4	WE	4	2.65	IHR	4	3.17
5	PEOPLE	2	1.32	MIR	3	2.38
6	THEY	2	1.32	WIR	3	2.38
7	WIFE	2	1.32	BEOBACHTER	1	0.79
8	ANYONE	1	0.66	DIEMEISTEN	1	0.79
9	EVERYONE	1	0.66	ELTERN	1	0.79
10	EX	1	0.66	IRGENDWER	1	0.79
11	HE	1	0.66	JEMAND	1	0.79
12	HUSBAND	1	0.66	MAN	1	0.79
13	NEUROLOGIST	1	0.66	VIELE	1	0.79
14	NOONE	1	0.66			
		151	100		126	100
	Total	151	100	Total	126	100

Table 7.21 Heads of nominal groups functioning as PR Cognizant

These 14 (EN) / 13 (GN) different heads of nominal groups are not only the most frequent, but the only ones functioning as PR Cognizant in EDNA. They all refer to human beings bestowed with a consciousness. In EN, there are personal pronouns, mainly *I* (76%) followed by *she, you, we, they, he,* and indefinite pronouns, namely *anyone, everyone, no one*. (Note by the way that there are not many instances of *he* in EN as PR Cognizant – does this suggest that women do not put their male partners in the position of PR Cognizant, of someone who knows something?) We find a few nouns in EN as PR Cognizant, i.e. *people, wife, ex, husband* and *neurologist*. In the German newsgroup texts, writers also place themselves in the position of the PR Cognizant most of the time, with *ich* and *mir* accounting for 79%, followed by *er, sie, ihr, mir, wir*, and other pronouns including *die meisten, irgendwer, jemand, man, viele* and two nouns: *Beobachter, Eltern*. Similar to the PR Carrier, the two main PR in mental processes are mostly realized by pronouns, thus the person feeling or knowing something is clear from the context or situation. What is being felt or know, i.e. what is put into the PR Phenomenon, stands in the rheme. Once more, the rheme contains the new information, the PR Phenomenon.

The reader will have noticed a recurring nominal group in the study of heads of groups in participant roles. In the English newsgroup texts, the pronoun *you* functions as PR Emoter, PR Carrier and PR Agent quite often. The frequent use of *you* suggests an interaction with readers, e.g. in sentences like [*I was wondering*] *if any of you have conquered binge eating, I understand how you feel*, [*my email is xcom*] *if you want to have a private discussion*.

The German newsgroup text writers do not use *du* or *ihr* to the same extent to connect with the reader, there is only the pronoun *ihr* used frequently in the PR Agent, but not any of the other participant roles. Do the German writers not want to establish a connection with their readers to the same degree as the English authors? Or do they use different means to do so? Some of the uses of *you* in the English texts do not address the reader but are used in a sentence which in German would be expressed by using the indefinite pronoun *man*, see example 213 and a possible translation of it into German.

(213) *I felt that strange sensation when you want to purge*

~ Ich hatte ein komische Gefühl, wie wenn man dringend auf Toilette muss.

This phenomenon calls for future investigations. We may assume for the moment that participant roles do not only function to express meaning in the experiential metafunction, to describe experiences, but also in the ideational metafunction, by connecting to the reader, or by not connecting with her / him.

One of the conclusions to be drawn from the analysis of participant roles in the EDNA corpus is that the first PR, the one with the grammatical function of subject, is not particularly revealing. We find mostly personal (or other) pronouns. The second PR, thus the ones which function as object or complement in a clause, tell us more clearly what the discourse is actually about. This result is in agreement with what was said earlier about the theme-rheme structure of clauses: The rheme, i.e. the objects and complements in an unmarked declarative clause, offers the new information. The theme holds given information, e.g. personal pronouns referring to the writer or someone who was mentioned before.

8 Conclusion

8.1 Summary

In this thesis, the aim was a comprehensive description of language as it is used in a new type of medium, Internet language, in particular the language of newsgroup texts. It was a contrastive study, describing and comparing the English and German language systems and language use. For this purpose, a corpus of newsgroup texts in these two languages was built, manually annotated and studied. As the underlying theory Systemic Functional Grammar was chosen because it is more comprehensive than other models. In a first step, the features of the systems under investigation were described quantitatively, followed in a second step by an investigation of the lexical items that most frequently realized the features. In this way, differences and similarities were found.

In the following subchapters, we return to the four main hypotheses from the introduction and examine to what extent they were confirmed or rejected. In fact, all hypotheses were partly confirmed, and partly rejected. In the end, the hypotheses turned out to be too much of a simplification. Nevertheless, the English and German newsgroup text writers, whose texts are collected in the EDNA corpus (*Englische und Deutsche Newsgroup Texte – Annotiertes Korpus*), gave a very detailed picture. Their use of language reveals that even though the cultures that writers came from may be quite similar, and the language systems comparable, the writers do talk about the same topics in different ways. Let us focus in on the details, and then step back to see the bigger picture.

8.2 Summary of results for the interpersonal metafunction: modality

The first hypothesis was concerned with the interpersonal metafunction, which describes how speakers or writers relate to their audience (Halliday 1994). One of the two systems which realize such interpersonal relationships is that of modality. Modal markers serve to weaken or strengthen a statement. By using modal auxiliaries, modal adverbs, modal particles and subjunctive verb forms, writers express probability and likelihood (epistemic modality) or obligation, permission, ability and inclination to do something (deontic or root modality). The hypothesis claimed that both English and German newsgroup text writers use modality in the same way and to the same extent. The relationship between the writers and the readers in these newsgroup texts is the same in both languages, it is non-hierarchical. Therefore, I assumed that there are more similarities than differences.

8.2.1 Similar use of modality

To start with, the analysis has shown that modal markers in both subcorpora of EDNA express root modality 30% of the time and epistemic modality 70% of the time. Modal auxiliaries are the most frequent modal markers (EN 55%, GN 37% of all modal markers). Modal adjuncts account for a quarter of all modal markers in the EDNA corpus. In both EN and GN, the modal auxiliaries *will, would* (EN) and *werden* (GN) respectively make up 70% of all modal auxiliaries expressing epistemic modality. These auxiliaries express a high certainty and strengthen the statements. In addition to the modal auxiliaries, two thirds of the modal adjuncts have the function of strengthening the writers' accounts. Some examples are *really, always, actually, usually* (EN) and *wieder, immer, wirklich, oft, immer wieder* (GN).

The annotation of EDNA with the UAM corpus tool has revealed that for some reason, root modality is expressed only by modal auxiliaries, whereas epistemic modality is expressed by all types of modal markers (modal auxiliaries and adjuncts, grammatical metaphor, and, in GN, modal particles and the subjunctive verb forms).

In both subcorpora, obligation and permission account for almost half of all instances of root modality (47% EN / 43% GN). A qualitative analysis has shown that it is rather obligation than permission which is expressed with modal auxiliaries like *have to, need to* and *can* in EN and *sollen, müssen, können* in GN.

8.2.2 Differing use of modality

In addition to the similarities in the use of modality in the English and German newsgroup texts in EDNA, there are also a few differences. The first and most striking difference is the fact that the German writers use about twice as many modal markers as the English writers. In GN, 39% of all clauses have at least one modal marker, sometimes two or even three. In EN, only 21% of all clauses have at least one modal marker. This is a highly significant divergence, from the statistical point of view. German writers seem to feel the need to strengthen or play down what they say. They seem to find it harder to simply say *'it is so'*.

Furthermore, the means to express epistemic or root modality differ in the English and German language systems. In the English language, writers can make use of modal auxiliaries, modal adjuncts and grammatical metaphor. In addition to these three options, German writers can also use modal particles and the subjunctive verb form to express certainty or likelihood (epistemic modality). It comes as no surprise, then, that modal auxiliaries are the most frequently used modal marker in both subcorpora (EN 55%, GN 37%). Following the modal auxiliaries, modal adjuncts are the second most frequent type of modal marker (EN 27%, GN 25%). The German writers use many modal particles (GN 26%). Modal particles and modal adjuncts together make up 51% of all modal markers in GN and are thus more frequent than modal auxiliaries in GN. Their share is almost as large as the share of modal auxiliaries in EN (55%).

The qualitative analysis has shown that the English writers in EDNA use significantly more grammatical metaphors (18% of all modal markers in EN, 8% in GN). The variety of superordinate clauses which express epistemic modality, however, is greater in the German texts. Some examples of such superordinate clauses as grammatical metaphors of modality include *I know, I think, I am sure* (EN) and *ich weiß, ich bin mir sicher, ich frage mich, ich glaube* (GN). Altogether we find 21 different superordinate clauses in GN compared to 14 different ones in EN. By contrast, the English newsgroup writers in EDNA use a greater variety of modal auxiliaries and semi-modals to express modality.

Possibly the most amusing result is that when the writers express root modality, the English writers express significantly more ability than the German writers (EN 36%, GN 21% of all modal markers expressing root modality). The German writers, on the other hand, express significantly more inclination than the English writers (EN 17%, GN 36% of all modal markers expressing root modality). The English writers feel they can do something, but may not want to, whereas the German writers claim that they want to do something, but seem to be unable to.

In the end, we cannot simply confirm the hypothesis that the English and German writers who contributed to EDNA use modality in the same way and to the same extent. Even though there are similarities, there are also many differences in the use of modal markers.

8.3 Summary of results for the interpersonal metafunction: negation

The second system that realizes interpersonal relationships is that of polarity. Speakers and writers can either make a positive statement (positive polarity), e.g. *I love her*, or negate a statement (negative polarity), e.g. *I do not love her*. Positive and negative polarity are the two complementary ends of a continuum, with modality covering everything in between, e.g. *I certainly love her*. The hypothesis was that both English and German newsgroup text writers express positive and negative polarity in the same way and to the same extent, because they talk about the same topics to the same audience.

8.3.1 Similar use of negation

In fact, there are few things that the English and German newsgroup texts in EDNA have in common with regard to negation. First of all, syntactic negation markers like *not, no, never* (EN 89%) and *nicht, kein, nicht mehr* (GN 94%) account for the majority of all negation markers in EDNA. Only 1% of all negation markers in EN and GN were classified as representing the textual metafunction, i.e. conjunctions like *whether or not* (EN) and *ohne, weder noch* (GN). Another aspect that is negligible in EDNA is multiple negation. EN has only 5 and GN only 2 clauses with two negation markers out of about 1.500 clauses.

9% of all syntactic negation markers in EN and 14% in GN are realized by adverbial groups, e.g. *never, by no means, no longer* (EN) and *nicht mehr, nie, nie-*

mals, noch nicht (GN), hence no statistically significant difference. What is striking, though, is that in EN, writers tend to state that something has never been, with *never* being the most frequently used adverb (16/18). In GN, however, writers indicate that something has come to an end by using *nicht mehr* as the most frequent adverb (27/39). When we add the instances of negation markers in the nominal groups in GN *keine mehr* (9 times) and *nichts mehr* (2 times), this focus on the end of a state of affairs is even clearer. The English equivalent in EDNA, *not anymore*, occurs only 11 times, *no more* and *no longer* only once.

The total numbers for morphological negation are rather small in EDNA. In both subcorpora, predicative adjectives are the most likely word class to carry a negating affix, for example *hopeless, impossible, unfaithful* (EN, 11 instances) and *nutzlos, unsicher, untragbar* (GN, 12 instances). Another word class with a negative affix in both EN and GN are the prepositions, e.g. *without* (EN, 10 instances) and *ohne* (GN, 4 instances). The total number for morphological negation, however, is really small. Even though both languages allow negating affixes on verbs, adverbs and nouns, writers make no use of these options, and only very little use of adjectives and prepositions with negating affixes.

8.3.2 Differing use of negation

In addition to the similarities in the use of syntactic and morphological negation described above, the study of EDNA revealed some differences. In EN, 15% of all clauses carry a negation marker, whereas in GN, as many as 20% of all clauses have a negation marker; this is a statistically significant difference. German writers in EDNA make considerably more negative statements.

Another difference is the use of syntactic negation markers (*no, not,* and *nicht, keine*) and morphological negation markers (affixes like *un-, -less,* or *-los*). Morphological negation accounts for 11% of all negation markers in EN, whereas it accounts for only 6% in GN. This is a statistically significant result. However, the total numbers for words with a negative affix are rather small, and do not allow for a generalization.

A third difference between the English and German texts in EDNA is the position of the syntactic negation. In EN, as many as 84% of all syntactic negation markers are on clause level, within the verbal group, with the negation marker *not*. In GN, negation with *nicht* on clause level, within the verbal group, accounts for only 61% of syntactic negation. Statistically, this difference is highly significant. The German writers in EDNA use more syntactic negation on phrase level, within the nominal group, i.e. *keine, nichts, niemand* (25% of all instances of syntactic negation). The English writers use negation on phrase level, within the nominal group with *no, no one, nothing* only 7% of the time. Again, this divergence is statistically highly significant. The German writers negate considerably more of their clauses. But it seems as if they try to downshift the negation to a deeper rank inside the clause. Is this done deliberately? Is this a way to make the negation less obvious, less of a face-threatening act? This would be an interesting question for future studies.

To sum up, we can say that even though there are similarities between the texts in EN and GN with regard to negation, there are just as many differences. We cannot say that the writers in EDNA use positive and negative polarity in the same way. The hypothesis was proven wrong.

8.4 Summary of results for the combination of modality and negation

In chapter 5, which focuses on modality and negation, we also investigated how modal markers and syntactic negation markers combine in clauses. Hence, we want to summarize the outcome here as well.

8.4.1 Similar use of the combination of modality and negation

On the most general level, no statistically significant difference between the English and the German part of EDNA with respect to modality and negation could be detected. The vast majority of clauses are positive clauses, in the sense of not containing a negation marker (EN 85%, GN 80%). In addition, the large majority of clauses do not contain a modal marker (EN 79%, GN 61%). In the remaining minority of clauses, how do the modal markers and negation markers combine?

To begin, there are a significantly higher number of clauses in GN with a modal marker (EN 21%, GN 39% of all clauses). Also, in GN, there are a significantly higher number of clauses with a negation marker (EN 15%, GN 20% of all clauses). But in both GN and EN, exactly 25% of all clauses that contain at least one modal marker additionally contain a negation marker. Remarkably, the percentage of clauses with both modal and negation markers is the same in both languages.

We can look at these numbers from a second angle. Of all clauses with a syntactic negation, as many as 40% in EN and 50% in GN additionally carry a modal marker. Thus, writers seem to feel the need to modify statements with negative polarity. By adding a modal marker to a clause with negative polarity, writers weaken or strengthen their statements, and the German writers do that even more than the English writers. Remember that 70% of all modal markers express epistemic modality. This means that writers express likelihood and certainty, or uncertainty, with regard to what is said in the negated clause.

The study at hand has also shown that some modal markers attract negation markers more than others do. On average, 25% of all clauses with a modal marker are negated. Modal auxiliaries have a stronger attraction for negative polarity; of all clauses with a modal auxiliary, 30% in EN and 28% in GN are also negated. In the German newsgroup texts, 31% of all clauses with a modal particle have a negation added to them. Modal adjuncts are less likely to be found in a negated clause. This combination occurs only half as often as the other options (EN 13%, GN 14% of clauses with modal adjunct and a negation marker). Subjunctive verb forms in the German newsgroup texts seem to repel syntactic negation, only one of the 26 clauses with a subjunctive verb form is syntactically negated (i.e. 4%). We can conclude, however, that the writers combine negation markers and the types of modal markers in one clause in a similar way.

Another similarity between the two subcorpora lies in the way that epistemic and root modality markers combine with negation markers. The results show no statistically significant differences. Root modal markers attract negation markers more strongly than epistemic modal markers do. In EN, 16% and in GN, 21% of all clauses with epistemic modal markers have a negation as well, whereas in EN, 41% and in GN, 31% of all clauses with root modal markers attract a negation marker.

Focusing in on the three types of root modality an astonishing detail became apparent. For all three types of root modality, EN has higher numbers of root modal marker plus negation marker in one clause (EN 41%, GN 31%). The least likely type of root modality to occur in a clause together with a negation is obligation or permission (EN 28%, GN 19% of all clauses with root modality; obligation and permission). When writers state what must be done, they are rather straight-forward with a positive clause. Next, clauses where a modal marker expresses inclination have a 30% chance of being negated (EN 37%, GN 34% of all clauses with root modality; inclination). In one third of all clauses where writers claim they want to do something, they in fact do not want to. The third type of root modality, where writers express ability, is negated more often than not. Remember that English writers express ability significantly more often than the German writers in EDNA (EN 36%, GN 21% of all modal markers expressing root modality). In fact, 64% of all clauses in EN indicating ability also carry a negation marker; in GN 52%. Thus, writers do not claim they can do something, instead they admit that they cannot do it. English writers admit inability even more often than the German writers. Do bear in mind, though, that total numbers for such clauses are rather small, and that the differences between EN and GN are not statistically significant. Still, these are rather unforeseen results.

8.4.2 Differing use of the combination of modality and negation

Clearly there are more similarities than differences when we investigate how modality and negation combine in one clause in the English and German newsgroup texts. There is one striking difference, though. The English subcorpus has a significantly higher number of syntactic negation markers in the verbal group, i.e. negation with *not* (84% of all types of syntactic negation), compared to the German subcorpus (61%). Of all clauses containing *not* in EN, 44% also include a modal marker. In GN, the number is much higher; 71% of all clauses containing *nicht* also include a modal marker. This result is statistically highly significant. The German writers seem incapable of just saying no without feeling the need to modify their negated statement.

8.5 Summary of results for the textual metafunction

The hypothesis concerning the theme-rheme structure is different from the previous ones in that it did predict differences in the use of marked and unmarked topical themes. The hypothesis reflects the difference in the English and German language systems. German has a more flexible word order, compared to the fixed word order in declarative clauses of subject – verb – object of the English language. It was therefore reasonable to expect that the German newsgroup text writers would make use of this freedom by putting other clause constituents into the topical theme position, thereby placing emphasis on what they find most important. As with my first two hypotheses, the results partly confirmed the third hypothesis, and partly rejected it.

8.5.1 Similar use of themes and rhemes

First of all, the three types of theme, topical, textual and interpersonal, are distributed equally in the English and German texts in EDNA. The overall numbers of topical themes in EN and GN do not show statistically significant differences. In EN we find 66%, in GN 63% topical themes, i.e. unmarked and marked as well as unmarked structural topical themes. The unmarked topical themes are the most frequent type of topical theme (EN 88%, GN 83%), i.e. the subject stands before the finite verb in declarative clauses (see chapter 6 for a definition of unmarked topical themes in interrogative and imperative clauses). Halliday (1994, 44) says that in spoken language, the personal pronouns *I* and *you* are most frequently the unmarked topical theme. This could be partly confirmed with the data from EDNA. In EN, *I* accounts for 51% of all unmarked topical themes, followed by *it* (7%), *she* (7%), *he* (6%), and *we* (5%). In GN, even though *ich* is also the most frequent unmarked topical theme with 40%, the amount is considerably smaller than in EN. The next most frequent unmarked topical themes in GN are *er* (8%), *es* (6%) and *sie* (5%). There is no *you* in EN, though, and there is no *du* or *ihr* in GN, either. Obviously, the newsgroup texts are different from spoken language insofar as the writers do not address the readers nor refer to them in their texts. The writers just talk about themselves, mainly. These texts clearly are not dialogues, as can be seen from the use of pronouns.

Apart from unmarked and marked topical themes, there is a third type of topical theme, namely unmarked structural themes. These are the relative pronouns that introduce subordinate clauses, for example *what, that* and *who* in EN or *die, der, das* in GN. The study of the use of relative pronouns showed a similar use in both subcorpora of EDNA.

Furthermore, in EN 41% of clauses are introduced with a conjunction, compared to 38% in GN. Conjunctions, i.e. structural conjunctives, are the most frequent type of textual theme (EN 92% and GN 83% of all textual themes). The difference is not statistically significant. In both subcorpora of EDNA, coordinating conjunctions account for 47% of all structural conjunctives. *And* in EN and *und* in GN account for one third of all structural conjunctives. The use of *and* / *und* to connect clauses is the easiest way to build clause complexes; this is preferred in the EDNA corpus. Clause coordination, however, is less frequent in EDNA than subordination with 53% in both EN and GN. In EN, *that, if, when* (among others) are used to introduce subordinate clauses, in GN, we find *dass, denn, wenn* as the more frequent subordinating conjunctions.

The second type of textual themes, conjunctive adjuncts, was discussed in chapter 6.3.5. The third type of textual theme is the continuative. They are equally rare in EN (3% of all textual themes) and GN (4%). In EN, there are *well* (9 times), *now* (3), and *yeah* (3), and in GN, there are *also* (8 times), *so* (4) and *naja* (4).

There are so few interpersonal themes in both EN (1% of all themes) and GN (2%) that it is hardly worth mentioning these, and the numbers do not differ to any statistically significant degree. We turn to the differences in the theme-rheme structure in the two corpora in EDNA in the next section.

8.5.2 Differing use of themes and rhemes

First, the second major type of topical theme, the marked topical theme, is more frequent in GN (8%) than in EN (5%) to a statistically significant degree. A look at the marked topical themes in the English newsgroup texts reveals that most of them express a temporal circumstance, like *now, a month ago, lately, today, at the time, before*. In EN, these temporal circumstances were annotated as marked topical theme, whereas in GN, the equivalents were annotated as unmarked topical theme. This was the result of a pilot study which had to be conducted due to the lack of adequate descriptions of the textual metafunction of the German language. In this pilot study, it became clear that temporal circumstances were used twice as often as other circumstances to stand before the finite verb in declarative clauses. I concluded that this would make the temporal circumstances unmarked topical themes, unlike in English, where they are marked topical themes if they precede the subject. In the end, it seems that temporal circumstances are the least marked topical themes in EN, too, as they are more frequent than other circumstances in the position before the subject. Had I decided to annotate them as unmarked topical themes in EN, like I did in GN, hardly any marked topical themes in EN would have remained.

The ten most frequent marked topical themes in GN together account for 42% of the marked topical themes; there is a great variety of phrases in the position before the finite verb. Many of these phrases put the emphasis on the writer, e.g. *mir* (12%), *da* (8%), *irgendwie* (6%), *für mich* (4%), *mich* (3%). Obviously, the writers in the German newsgroup texts use marked topical themes to put focus on themselves. This compensates for the lower number of *ich* as unmarked topical themes with 40% in GN, compared to *I* with 51% in EN of all unmarked topical themes.

A second difference between the English and the German newsgroup texts in EDNA is the number of conjunctive adjuncts which are used as textual themes. The German writers use about three times more conjunctive adjuncts than the English writers (EN 4%, GN 13% of all textual themes); this difference is statistically significant. Examples from EN include *however* (7 times), *anyway* (6), and *especially* (3). Examples from GN include *dann* (15 times), *nun* (11), and *nur* (10). The German texts appear more coherent through these conjunctive adjuncts.

Finally, the third difference is the number of minor clauses. In GN, 8% of the constituents which are not a theme are minor clauses, the remaining 92% are rhemes. In EN, we find only 6% of minor clauses. This difference is statistically significant. In the English newsgroup texts, minor clauses are mainly salutations like *hi guys, take care everyone, thanks*. In the German texts, apart from salutations and formulas, there are a large number of clauses without a finite verb, like for example *oft auch von mir aus; also gemeinsam, und doch getrennt;*

zumindest noch einmal. The German writers make more use of verb ellipsis than the English writers. This may suggest that the German newsgroup texts are more like spoken language than the English texts.

To sum up, we can say that the hypothesis was more confirmed than rejected; there are differences in the English and German newsgroup texts. In particular, there are significantly more marked topical themes in GN, as expected, and also more conjunctive adjuncts. There are, however, also many similarities, especially the heavy use of the personal pronouns *I* and *ich* as unmarked topical themes.

8.6 Summary of results for the experiential metafunction

Finally, this subchapter summarizes the results of the study of the transitivity system. The general hypothesis was that in the English and German newsgroup texts, writers would use the same proportions of process types; mostly relational processes, followed by mental and action processes. I expected writers to use more or less the same lexical items to talk about their experiences of the world around them and inside them. This hypothesis was based on the assumption that English and German are closely related languages, and that the English and German writers who contributed to the newsgroups live in cultures which are fairly similar, and write about similar topics. This should be reflected in a similar use of PR and process types.

8.6.1 Similar use of PR and process types

Even though the share of the major process types is different in EN and GN, as will be summarized in the next section, the frequency of the major participant roles (PR) in the two subcorpora is strikingly similar. The ranking is shown below. Note that the PR Phenomenon most often consists of a subordinate clause and is used in more than one process type. All other PR appear (with few exceptions) in only one process type.

1. PR Phenomenon EN 19%, GN 18%
2. PR Agent EN 17%, GN 18%
3. PR Carrier EN 17%, GN 18%
4. PR Attribute EN 13%, GN 12%
5. PR Affected EN 6%, GN 5%
6. PR Emoter EN 5%, GN 5%
7. PR Cognizant EN 5%, GN 5%

Similarities between the English and German newsgroup texts can also be observed when it comes to the lexical verbs that realize the different types of processes. Action processes are realized by a great variety of different lexical verbs. The 10 most frequent lexical verbs account for only 26% in EN and 21% in GN of all lexical verbs realizing action processes. In EN, the top ten of lexical verbs in action processes include *do, eat, leave, see, help, meet, purge, start, happen, have*. In GN, the top ten include *essen, helfen, tun, zunehmen, abnehmen, machen, anfangen, aufhören, gehen, reden mit*.

The lexical verbs that realize relational processes form a group which is similar in EN and GN, but different from the group of lexical verbs in action processes. The variety is much smaller here. The ten most frequent lexical verbs account for 83% of all relational processes in EN, they include *be, have, feel, get, go, lose, live, become, make, got*. In GN, the ten most frequent lexical verbs make up 70% of all relational processes; these verbs include *sein, haben, gehen, werden, geben, sich X fühlen, zusammen sein, finden, bekommen, zusammen sein mit*. In EN, *be* and *have* together take the largest share with 58%. In GN, *sein* and *haben* make up 55% of all lexical verbs in relational processes.

With regard to the variety of different lexical verbs realizing a process type, mental processes are in between action and relation processes. The top ten of most frequent lexical verbs in mental processes in EN account for 52%, these verbs are *know, want to* + infinitive, *think, say, want* + nominal group, *tell, love, feel, ask, feel like*. The top ten lexical verbs in mental processes in GN account for 47%, these verbs include *wissen, sagen, lieben, denken, brauchen, wollen, meinen, glauben, kennen, merken*.

The German newsgroup texts have a greater number of different lexical verbs in any of the main process types compared to the English newsgroup texts. In a future study, it may be worth investigating whether this can be explained with the general phenomenon of over- or under-specification in one of the two languages (König and Gast 2007, 222).

We now proceed to consider the participant roles (PR) that are most frequently used in the EDNA corpus. In this respect, the English and German subcorpora show little deviation. We can distinguish between the first major PR, which is usually the subject of a clause, and the second PR, which is usually realized by an object or complement, see examples 214 and 215.

(214) She_S / PR Carrier had two children_O / PR Possessed with this man

(215) Ich_S / PR Emoter fühle mich dann wohler und mir_O / PR Emoter geht es_S / ES richtig gut_C / PR Phen

In chapter 7 it became clear that most major PR in the newsgroup texts are realized by personal pronouns. Table 8.1 below shows the percentage of 1st and 3rd person singular and 1st person plural pronouns used as PR Agent, PR Carrier, PR Emoter and PR Cognizant in the EDNA corpus. Note that theses pronouns include not only personal pronouns in nominative case, but also in accusative and dative case. Even though most major first PR are realized by subjects, there are exceptions. Thus, in EN, the personal pronouns used as major PR incluce *I, she, he, it* and *we*. In GN, we find *ich, sie, er, es* and *wir*, and also *mir, mich, ihr* and *ihm*.

	PR Agent	PR Carrier	PR Emoter	PR Cognizant
1st person singular EN %	50	39	74	76
3rd person singular EN %	20	20	14	8
1st person plural EN %	9	5	1	3
1st person singular GN %	51	40	71	79
3rd person singular GN %	23	24	22	10
1st person plural GN %	4	3	1	2

Table 8.1 Percentage of personal pronouns realizing different PR in EDNA

The table above shows that in the newsgroup texts where people write about problems with their eating disorders or with their relationships, they talk about *me, myself and I,* as the famous song title goes. The writers themselves do, they are, and they think or feel. This may not be very surprising, but in fact, it could not be foreseen. This is the first quantitative study based on a manually annotated corpus of process types and participant roles and this result, even if not surprising, is very valuable.

In the following, the results for the second PR in the clauses collected in EDNA are summarized. The main ones are PR Affected in action processes, and PR Attribute and PR Possessed in relational processes. These results cannot be summarized as easily as those for the first main PR because they show a much greater variety. For the PR Affected in EN, we have *me, myself* and *I* accounting for 26%, and *it, her* and *him* together make up 15%. Thus, 1st person and 3rd person singular pronouns are not as frequently used for PR Affected as they are used for PR Agent or PR Carrier. The last four of the ten most frequent lexical items as PR Affected are *salad, each other, friend* and *food*. In GN, 1st person singular pronouns *ich, mich* and *mir* account for 38% and 3rd person singular pronouns *sie, ihn, ihm* account for 11%. Furthermore, the ten most frequent lexical items used as PR Affected include *uns, sich, das* and *nichts*. The PR Affected does not tell us what the discourse is about, except that people are being referred to. In this respect, PR Affected are similar to PR Agent, PR Carrier, PR Emoter and PR Cognizant.

A closer look at the lexical items that express PR Attribute and PR Possessed in relational processes is more revealing. Both PR Attribute and PR Possessed are realized by a great variety of different lexical items. Among the more frequent ones, we find *x years old, x pounds, married, confused, new, full, great* and *nice* as PR Attribute in EN. Writers describe how they are. In GN, the more frequent PR Attribute include *x Jahre alt, dick, x Kilo schwer, gut, normal, schlecht, besser, anders*. A similar picture is painted with the lexical items used as PR Possessed. In EN, the more frequent ones include *children, problem, eating disorders, friends* and *time*. In GN, we have the lexical items *Probleme, X Kilo, Kraft, Zeit, Freund, Freundin* and *Magersucht*. Writers describe what they have, or do not have. The fact that the second main PR (objects and complements) are realized by a much greater variety of lexical items than the first main PR (subjects) is connected to the theme-rheme structure of the clauses. In general, the first main PR (subject) stands in theme position before the finite verb in a declarative clause. Themes tend to pick up the topic of preceding clauses, therefore, we find many personal pronouns. The second main PR (objects and complements) stands in the rheme position after the finite verb in declarative clauses. In the rheme, new information is given about the theme, therefore, we find a great variety of different lexical items.

8.6.2 Differing use of PR and process types

One of the dissimilarities, which could not be foreseen, is the use of major process types. The hypothesis assumed relational processes to be the most frequently used process type in EDNA; writers would use these to state what problems they have. I expected mental processes to be the second most frequent process type, which writers use to talk about what they feel and think. Finally, I expected action processes to be the third most frequently used process type; writers use action processes to describe what they are doing and what is happening.

Relational processes are indeed the most frequent process type in EDNA, and writers in EN and GN use relational processes roughly to the same extent (EN 37%, GN 39%). When we look at the subtypes of relational processes, however, we see a difference. Relational locational processes are significantly more frequent in GN (16%) than in EN (9%). The German writers in EDNA state much more often where things are, or happen, in place or in time.

The second most frequently used process type differs in EN and GN. The German writers use 6% more action processes (EN 25%, GN 31%) than the English writers. Therefore action processes are the second most frequent type in GN, but only the third most frequent in EN. The English writers in EDNA use 5% more mental processes (EN 31%, GN 26%). Mental processes are thus the second most frequently used process type in EN, but only the third most frequent in GN. This is a statistically significant deviation. The results suggest that the German texts in EDNA are more about what is happening in the world, whereas the English texts are more about thoughts and feelings.

Furthermore, EN and GN are different when it comes to processes with process type extensions, i.e. processes which consist not only of a lexical verb, but need a preposition or noun or reflexive pronoun to be complete, e.g. *to take a shower, to cut off* in EN, *in den Arm nehmen, sich wünschen* in GN. The German writers in EDNA use significantly more process type extensions (PrEx), and reflexive pronouns in particular. In GN, there are 386 reflexive pronouns, in EN, only 17. Of the 386 reflexive pronouns, 91 are PrEx, and the other 295 were annotated as participant role. Reflexive pronouns account for 32% of the process extensions in GN (91 out of 280 process extensions). It is not surprising to see so many more reflexive pronouns being used in GN. König and Gast (2007, 141-159) discuss the overlap of domains of use of the reflexive pronouns in the English and the German languages. They state that there is some overlap, but that both languages also use the reflexive pronoun for different purposes, in addition to the overlapping functions. They explain this with the historical development of the languages. It is impressive to see this difference surviving in the modern age of computer-mediated communication.

All in all, with regard to the experiential metafunction that expresses how people see the world around them and inside them, we can conclude that the English and German texts collected in EDNA show many similarities but also some differences. German writers use slightly more action processes, whereas English writers use more mental processes. The main PRs (Ag, Ca, Em, Cog) in a clause tell us little about the topic, they are mostly personal pronouns. The second PRs (Af, At) are realized by a much greater variety of different words and phrases and thus tell us more about the topic of the discourse.

8.7 The bigger picture – summary of the summary

The present study has revealed many interesting details, some of which were unexpected. Since the annotation of a corpus of CMC with SFG features is pioneering work, many details of the description have been shown for the first time. This subchapter condenses the details to show the bigger picture.

To begin with, the German writers who contributed to the EDNA corpus use significantly more modality than the English writers. The German writers also use significantly more negative polarity; many of the syntactic negation markers are shifted from clause level to phrase level, possibly in order to hide them. In the German texts, 50% of all the clauses that carry a negation marker also include a modal marker, in the English texts, the share is lower (40%). Writers in both subcorpora seem to feel the need to indicate that 'no' is not one end of a dichotomy, but rather somewhere between 'yes' and 'no'. This is even more obvious for the syntactic negation marker *nicht* in the German texts. 71% of all clauses with *nicht* additionally carry a modal marker, and 70% of all modal markers indicate epistemic modality. The German newsgroup text writers in EDNA strongly avoid just saying 'no'. We may conclude that the English writers in EDNA are a bit more straight-forward in making a statement clear.

The German writers make use of the greater word order freedom that the German language system offers by putting constituents to the front of a clause for emphasis, as marked topical themes. Many of these marked topical themes include *mir, mich, für mich* and thus compensate for the lower number of *ich* as unmarked topical theme (40%), compared to *I* in the English texts (51%). In EDNA, writers just like to talk about themselves.

Writers in EDNA mostly say how they are or what they have; relational processes are the most frequent ones (EN 37%, GN 39%). This makes the texts more static than dynamic. It may indicate that writers do think about their problems, but do not feel like they *could do*, or indeed *do*, much about those problems.

8.8 Evaluation of methodology

The present study has contributed a thorough description of computer-mediated language in English and in German to this new field of linguistic research. The study demonstrated the usefulness of a corpus for a comprehensive description and contrastive linguistic analysis. The study has been successful in this respect. Of course, the results presented here are only valid for the texts in the EDNA corpus of computer-mediated communication.

The SFG framework has proven to be very useful for corpus annotation since it provides us with clear networks of options. The major challenge was the amount of time that was necessary for the manual analysis of 2 x 10,000 words in the EDNA corpus. In this study, the UAM corpus tool (O'Donnell 2008) was used for the computer-assisted manual annotation. Manual annotation, however, is not feasible for annotating corpora of a very large size. In the future, in order to be able to use corpus data based on SFG theory, we must find ways for annotating or retrieving SFG features automatically. One step in this direction is the *Parsimonious Vole* parser which is being developed by Costetchi (2013).

I hope that this study can fuel insights and provide training data for English and German. Do not hesitate to contact the author if you require access to the EDNA corpus, or the annotation guidelines.

8.9 Future work

One question that could not be answered in the course of this thesis is whether data from a corpus provides evidence that differs from statements about grammatical features in grammar books. More time and effort has to be put into the comparison of corpus data, e.g. the data collected in EDNA, and grammar books.

Furthermore, the results from this study need to be compared to reference corpora in order to be able to say whether they are restricted to newsgroup texts, or whether they present common phenomena. The reason why I have not done that, apart from time constraints, is the absence of corpora which are annotated for SFG criteria. Some linguistic features, like for example modal auxiliaries and modal adjuncts, could have been retrieved from other corpora with a simple concordance search. Even then, however, we would not know which type of modality was expressed by them. Other features, e.g. the semantic roles of phrases in a clause or the identification of unmarked and marked topical themes, are much more difficult if not impossible to retrieve automatically. This is one of the major obstacles that SFG faces in the future.

During the work on the current project, each of the three metafunctions has pointed to further questions and material for more detailed studies. The aim of this study was a comprehensive, overall description of CMC. There were, however, many tempting paths to the left and right to get side-tracked. Apart from synchronic studies of the metafunctions in one register or another, a diachronic study of computer-mediated communication could show whether, or in what way, the English and German languages change due to technical developments like the Internet.

In the end, this project has brought new insights in the fields of contrastive linguistics, computer-mediated communication and Systemic Functional Linguistics for the English and German languages. Guidelines for manual annotation of SFG features in English and German were written and tested. A valuable corpus has emerged for future studies, the EDNA corpus (*Englische und Deutsche Newsgroup Texte – Annotiertes Korpus*). Without doubt corpus linguistics will develop and prosper as a methodology in linguistic studies. Only the future will tell to what extent Systemic Functional Grammar as a language theory will be able to profit from corpus studies, and vice versa.

References

Aijmer, Karin, and Bengt Altenberg, eds. 2013. *Advances in Corpus-based Contrastive Linguistics: Studies in Honour of Stig Johansson*. Amsterdam; Philadelphia: John Benjamins.

Androutsopoulos, Jannis, and Michael Beißwenger. 2008. "Introduction: Data and Methods in Computer-Mediated Discourse Analysis." *Language@Internet* 5, article 9.

Baker, Paul. 2010. "Corpus Methods in Linguistics." In *Research Methods in Linguistics*, edited by Lia Litosseliti, 93-113. London; New York: Continuum.

Baker, Paul, Andrew Hardie, and Tony McEnery. 2006. *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baron, Naomi. 1984. "Computer Mediated Communication as a Force in Language Change." *Visible Language* 18 (2):118-141.

Barton, David, and Carmen Lee. 2013. *Language Online: Investigating Digital Texts and Practices*. London; New York: Routledge.

Beißwenger, Michael, and Angelika Storrer. 2008. "Corpora of Computer-Mediated Communication." In *Corpus Linguistics. An International Handbook. Handbücher zur Sprach- und Kommunikationswissenschaft 29.1*, edited by Anke Lüdeling and Merja Kyto, 292-309. Berlin; New York: Walter de Gruyter.

Biber, Douglas. 1992. "Representativeness in Corpus Design." In *Corpus Linguistics: Readings in a Widening Discipline*, edited by Geoffrey Sampson and Diana McCarthy, 174-197. London; New York: Continuum.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow, England; New York: Longman.

Bußmann, Hadumod. 2002. *Lexikon der Sprachwissenschaft*. Stuttgart: Alfred Kroener.

Caffarel, Alice. 2006. *A Systemic Functional Grammar of French: From Grammar to Discourse*. London: Continuum.

Caffarel, Alice, J. R. Martin, and Christian M. I. M. Matthiessen, eds. 2004. *Language Typology: A Functional Perspective*. Amsterdam; Philadelphia: John Benjamins.

Coffin, Caroline, Jim Donohue, and Sarah North. 2009. *Exploring English Grammar: From Formal to Functional*. London; New York: Routledge.

Consortium, BNC. 2007. The British National Corpus, Version 3 (BNC XML Edition). Oxford University Computing Services.

Costetchi, Eugeniu. 2013. "Semantic Role Labelling as SFL Transitivity Analysis." ESSLLI, Duesseldorf, Germany.

Crystal, David. 2011. *Internet Linguistics: A Student Guide*. London; New York: Routledge.

Culo, Oliver, Silvia Hansen-Schirra, Stella Neumann, and Karin Maksymski. 2011. "Querying the CroCo Corpus for Translation Shifts. Beyond Corpus Construction: Exploitation and Maintenance of Parallel Corpora." *Translation: Corpora, Computation, Cognition* 1 (1):75-104.

Dahl, Oesten, ed. 2000. *Tense and Aspect in the Languages of Europe*. Berlin: Mouton de Gruyter.

Danet, Brenda, and Susan C. Herring, eds. 2007. *The Multilingual Internet: Language, Culture and Communication Online*. Oxford: Oxford University Press.

Depraetere, Ilse, and Susan Reed. 2006. "Mood and Modality in English." In *The Handbook of English Linguistics*, edited by Bas Aarts and April McMahon, 269-290. Malden: Blackwell.

Duden. 2006. *Band 4: Die Grammatik*. 7th ed. Mannheim, Leipzig, Wien, Zürich: Duden Verlag.

Fabricius-Hansen, Cathrine, Irmhild Barz, Damaris Nübling, Thomas A. Fritz, Reinhard Fiehler, Peter Gallmann, Jörg Peters, and Peter Eisenberg. 2006. *Duden Band 4: Die Grammatik*. Edited by Dudenredaktion. Mannheim: Bibliographisches Institut & F.A. Brockhaus AG.

Fawcett, Robin P. 2008. *Invitation to Systemic Functional Linguistics through the Cardiff Grammar : an Extension and Simplification of Halliday's Systemic Functional Grammar*. 3rd ed. London: Equinox.

Fawcett, Robin P. forthcoming. *The Functional Semantics Handbook: Analyzing English at the Level of Meaning*. London: Equinox.

Feldweg, Helmut, Ralf Kibiger, and Christine Thielen. 1995. "Zum Sprachgebrauch in deutschen Newsgruppen." *Osnabrücker Beiträge zur Sprachtheorie* 50:143-154.

Ferrara, Kathleen, Hans Brunner, and Greg Whittemore. 1991. "Interactive Written Discourse as an Emergent Register." *Written Communication* 8 (1):8-33.

Fontaine, Lise. 2013. *Analysing English Grammar: A Systemic Functional Introduction*. Cambridge: Cambridge University Press.

Fraas, Claudia, Stefan Meier, and Christian Pentzold. 2012. *Online-Kommunikation: Grundlagen, Praxisfelder und Methoden*. München: Oldenbourg Verlag.

Frehner, Carmen. 2008. *Email - SMS - MMS: The Linguistic Creativity of Asynchronous Discourse in the New Media Age*. Bern; Berlin: Peter Lang.

Goatly, Andrew. 2004. "Corpus Linguistics, Systemic Functional Grammar and Literary Meaning: A Critical Analysis of Harry Potter and the Philosopher's Stone." *Journal of English Language, Literature in English and Cultural Studies* 46:115-154.

Götze, Lutz, and Ernest W. B. Hess-Lüttich. 1999. *Grammatik der deutschen Sprache : Sprachsystem und Sprachgebrauch*. Gütersloh: Bertelsmann.

Granger, Sylviane. 2003. "The Corpus Approach: A Common Way Forward for Contrastive Linguistics and Translation Studies?" In *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, edited by Sylviane Granger, Jacques Lerot and Stephanie Petch-Tyson, 17-29. Amsterdam; New York: Rodopi.

Granger, Sylviane, Jacques Lerot, and Stephanie Petch-Tyson, eds. 2003. *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam; New York: Rodopi.

Gries, Stefan Th. 2008. *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoeck & Ruprecht.

Halliday, M. A. K. 1961. "Categories of the theory of grammar." *Word* 17 (3):241-292.

Halliday, M. A. K. 1985. *An Introduction to Functional Grammar*. London: Edward Arnold.

Halliday, M. A. K. 1989a. *Spoken and Written Language*. 2nd ed, *Language education*. Oxford: Oxford University Press.

Halliday, M. A. K. 1994. *An Introduction to Functional Grammar*. 2nd ed. London: Edward Arnold.

Halliday, M. A. K., and Ruqaiya Hasan. 1989b. *Language, Context, and Text : Aspects of Language in a Social-semiotic Perspective*. 2nd ed, *Language Education*. Oxford: Oxford University Press.

Halliday, M. A. K., and Christian M. I. M. Matthiessen. 1999. *Construing Experience through Meaning : a Language-based Approach to Cognition*, *Open Linguistics Series*. London; New York: Cassell.

Halliday, M. A. K., and Christian M. I. M. Matthiessen. 2004. *An Introduction to Functional Grammar*. 3rd ed. London: Arnold.

Halliday, M. A. K., and Christian M. I. M. Matthiessen. 2014. *Halliday's Introduction to Functional Grammar*. 4th ed. London; New York: Routledge.

Halliday, M.A.K., and Z.L. James. 1993. "A Quantitative Study of Polarity and Primary Tense in the English Finite Clause." In *Techniques of Description*, edited by John Sinclair, 32-66. London; New York: Routledge.

Hawkins, John. 1986. *A Comparative Typology of English and German: Unifying the Contrasts*. London; Sydney: Croom Helm.

Herring, Susan C. 2001. "Computer-mediated Discourse." In *The Handbook of Discourse Analysis*, edited by Deborah Schiffrin, Deborah Tannen and Heidi E. Hamilton, 612-634. Oxford: Blackwell.

Herring, Susan C. 2004. "Computer-mediated Discourse Analysis: An Approach to Researching Online Behavior." In *Designing for Virtual Communities in the Service of Learning*, edited by Sasha Barab, Rob Kling and James H. Gray, 338-376. Cambridge; New York: Cambridge University Press.

Herring, Susan C. 2013. "Discourse in Web 2.0: Familiar, Reconfigured, and Emergent." In *Discourse 2.0: Language and New Media*, edited by Deborah Tannen and Anna Marie Trester, 1-25. Washington, DC: Georgetown University Press.

Herring, Susan C. , Dieter Stein, and Tuija Virtanen, eds. 2013. *Pragmatics of Computer-Mediated Communication*. Berlin: De Gruyter Mouton.

Hess, Alan. 2014. "Netzwerk Deutsche SFL: Foren und Wikis zur Diskussion der deutschen SFL." Accessed 09 June 2014. http://manxman.ch/moodle2/course/view.php?id=18

Hjelmslev, Louis. 1974. *Prolegomena zu einer Sprachtheorie*. München: Hueber. Original edition, 1943.

IDS. 2014. COSMAS II. Mannheim: IDS Mannheim.

James, Carl. 1980. *Contrastive Analysis*. London: Longman.

Johansson, Stig. 2003. "Contrastive Linguistics and Corpora " In *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, edited by Sylviane Granger, Jacques Lerot and Stephanie Petch-Tyson, 31-44. Amsterdam; New York: Rodopi.

Johansson, Stig. 2007. *Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam; Philadelphia: John Benjamins.

Johansson, Stig, and Knut Hofland. 1994. "Towards an English-Norwegian Parallel Corpus." In *Creating and Using English Language Corpora* edited by Udo Fries, Gunnel Tottie and Peter Schneider, 25-37. Lund: Lund University Press.

Koch, Peter, and Wulf Oesterreicher. 1994. "Schriftlichkeit und Sprache." In *Schrift und Schriftlichkeit. Handbücher für Sprach- und Kommunikationswissenschaft 1*, edited by Hartmut Günther and Otto Ludwig, 588-604. Berlin; New York: Walter de Gruyter.

König, Ekkehard, and Volker Gast. 2007. *Understanding English-German Contrasts*. Berlin: Erich Schmidt Verlag.

Kunz, Kerstin, and Erich Steiner. 2013. "Cohesive Substitution in English and German: A Contrastive and Corpus-based Perspective." In *Advances in Corpus-based Contrastive Linguistics: Studies in Honour of Stig Johansson*, edited by Karin Aijmer and Bengt Altenberg, 201-231. Amsterdam; Philadelphia: John Benjamins.

Lavid, Julia, Arús Jorge, and Juan Rafael Zamorano-Mansilla. 2010. *Systemic Functional Grammar of Spanish: A Contrastive Study with English*. London: Continuum.

Marzo, Stefania, Kris Heylen, and Gert De Sutter, eds. 2012. *Corpus Studies in Contrastive Linguistics*. Amsterdam; Philadelphia: John Benjamins.

Matthews, P. H. 2005. *Oxford Concise Dictionary of Linguistics*. Oxford: Oxford University Press.

Matthiessen, Christian M. I. M. 1995. *Lexicogrammatical Cartography: English Systems*. Tokyo: International Language Science Publishers.

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, Tony, and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction*. 2nd ed. Edinburgh: Edinburgh University Press.

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London; New York: Routledge.

Neumann, Stella. 2003. *Textsorten und Übersetzen. Eine Korpusanalyse englischer und deutscher Reiseführer*. Frankfurt: Peter Lang.

Neumann, Stella. 2014. *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. Berlin: De Gruyter.

O'Donnell, Michael. 2008. "The UAM CorpusTool: Software for Corpus Annotation and Exploration." In *Applied Linguistics Now: Understanding Language and Mind*, edited by Carmen M. Bretones Callejas, 1433-1447. Almería: Universidad de Almería.

Oakes, Michael P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press. Reprint, 2003.

Pankow, Christiane. 2003. "Zur Darstellung nonverbalen Verhaltens in deutschen und schwedischen IRC-Chats. Eine Korpusuntersuchung." *Linguistik Online* 15.

Petersen, Uwe Helm. 2004. "Überlegungen zu einer Systemisch Funktionalen Grammatik des Deutschen." In *Thema mit Variationen. Dokumentation des VI. Nordischen Germanistentreffens in Jyväskylä vom 4. -9. Juni 2002*, edited by A. Jäntti and J. Nurminen. Frankfurt am Main: Peter Lang.

Petersen, Uwe Helm. forthcoming. *Bedeutung, Funktion und Form*. Odense: University of Southern Denmark.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Harlow: Longman.

Rayson, Paul. 2010. "CLAWS tagger." UCREL, Lancaster University Accessed 09 June 2014. http://ucrel.lancs.ac.uk/claws/trial.html.

Salkie, Raphael. 2006. INTERSECT: Multilingual Corpora and Contrastive Linguistics. A Project at the University of Brighton.

Salkie, Raphael. 2008. "Modals and Typology: English and German in Contrast." In *Current Trends in Contrastive Linguistics: Functional and Cognitive Perspectives*, edited by Maria L. A. Gómez-Gonzáles, J. L. Mackenzie and E. M. González Alvarez, 77-98. Amsterdam; Philadelphia: John Benjamins.

Schmid, Helmut. 2013. "TreeTagger - a language independent part-of-speech tagger." Accessed 22 May 2013. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

Schmied, Josef. 2008. "Contrastive Corpus Studies." In *Corpus Linguistics: An International Handbook*, edited by Anke Lüdeling and Merja Kytö, 1140-1159. Berlin; New York: Walter De Gruyter.

Steiner, Erich. 1983. *Die Entwicklung des Britischen Kontextualismus*. Heidelberg: Julius Groos.

Steiner, Erich. 1987. "Zur Zuweisung satzsemantischer Rollen im maschinellen Übersetzungssystem EUROTRA-D." In *Computerlinguistik und Philologische Datenverarbeitung*, edited by Ursula Klenk. Hildesheim: Olms.

Steiner, Erich. 2001. "Translations English-German: Investigating the Relative Importance of Systemic Contrasts and of the Text-type 'Translation'. Papers from the 2000 Symposium on Information Structure Across Languages." *SPRIK Reports University of Oslo* 7.

Steiner, Erich, U. Eckert, B. Weck, and J. Winter. 1988. "The Development of the EUROTRA-D System of Semantic Relations." In *From Syntax to Semantics: Insight from Machine Translation*, edited by Erich Steiner, Paul Schmidt and Cornelia Zelinsky-Wibbelt. London: Pinter.

Steiner, Erich, and Elke Teich. 2004. "Metafunctional Profile of the Grammar of German." In *Language Typology: A Functional Perspective*, edited by Alice Caffarel, J. R. Martin and Christian M. I. M. Matthiessen. Amsterdam; Philadelphia: John Benjamins.

Stommel, Wyke. 2008. "Conversation Analysis and Community of Practice as Approaches to Studying Online Community." *Language@Internet*.

Störmer, Eckart. 2009. "Bedeuten lernen - Die Funktionale Grammatik von Michael Halliday." http://manxman.ch/moodle2/course/view.php?id=20.

Teich, Elke. 1991. "A Systemic Grammar of German for Text Generation." In *Occasional Papers in Systemic Linguistics*, edited by Dirk Noël. Nottingham: University of Nottingham.

Teich, Elke. 1999. *Systemic Functional Grammar in Natural Language Generation: Linguistic Description and Computational Representation*. London; New York: Cassell.

Teich, Elke. 2003. *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Berlin; New York: Mouton de Gruyter.

Tekin, Özlem. 2012. *Grundlagen der Kontrastiven Linguistik in Theorie und Praxis*. Tübingen: Stauffenburg.

Thibaut, John W., and Harold H. Kelly. 1959. *The Social Psychology of Groups*. New York: Wiley.

Thompson, Geoff. 2014. *Introducing Functional Grammar*. 3rd ed. London; New York: Routledge.

Thompson, Geoff, and Susan Hunston, eds. 2006. *System and Corpus: Exploring Connections*. London; Oakville: Equinox.

Vater, Heinz. 1975. "Werden als Modalverb." In *Aspekte der Modalität*, edited by Joseph Calbert and Heinz Vater, 71-148. Tübingen: Gunter Narr.

Xiao, Richard. 2010. *Using Corpora in Contrastive and Translation Studies*. Newcastle: Cambridge Scholars Publishing.

Yan, Fang, and Jonathan Webster, eds. 2014. *Developing Systemic Functional Linguistics: Theory and Application*. London; Oakville: Equinox.

Yates, Simeon J. 1996. "Oral and Written Linguistic Aspects of Computer Conferencing." In *Computer-mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, edited by Susan C. Herring, 29-46. Amsterdam; Philadelphia: John Benjamins.

Zinsmeister, Heike, Erhard Hinrichs, Sandra Kübler, and Andreas Witt. 2008. "Linguistically Annotated Corpora: Quality Assurance, Reusability and Sustainability." In *Corpus Linguistics: An International Handbook*, edited by Anke Lüdeling and Merja Kytö, 759-776. Berlin; New York: Walter de Gruyter.

Zitzen, Michaela. 2004. "Topic Shift Markers in Asynchronous and Synchronous Computer-Mediated Communication (CMC)." Doktorarbeit, Universität Düsseldorf.

Appendix

The appendix on the CD that is enclosed includes the following:

1) *Me, myself and I*: the complete dissertation

2) Annotation guidelines Theme English
3) Annotation guidelines Theme German
4) Annotation guidelines Modality and Negation English
5) Annotation guidelines Modality and Negation German
6) Annotation guidelines Transitivity English
7) Annotation guidelines Transitivity German

The EDNA corpus, in the form of two UAM Corpus Tool projects.

8) Annotated EN clean gold UAM 1_33
9) Annotated GN clean gold UAM 1_33

In order to open the projects, it is essential to have the UAM Corpus Tool installed on the computer beforehand. It is available from Mick O'Donnell's website http://www.wagsoft.com/CorpusTool/index.html.

Please choose version 2.8.14, the older version. You can open the EDNA project by choosing 'open project' and selecting the file that ends in *.ctpr* from the EDNA folder (the round blue icon).

You will find that the first letter in each segment is missing. This error is due to the conversion from version 1.33 to 2.8.14.

There is a short manual for the UAM CT 2.8.x on the next page.

A short manual for the UAM CT Version 2.8.x

- This manual was written by Anke Schulz for use in class at the University of Bremen.

First of all, you need a text to analyse. Find a text, *copy-and paste* or type to MS Word or other text processing software. Save your text as *text only* (.txt), save as *other encoding*, save as *UTF-8*. Then close text file.

Next, download the UAM corpus tool for free from Mick O'Donnell's website http://www.wagsoft.com/CorpusTool/ (27.08.2014). There is also a manual available from that website that explains how to work with the tool. Install UAM ct.

Third, open the UAM ct by *Start New Project*. Go through the steps the assistant suggests. Next time you want to work on the same project you use *Open Last Project*.

Now that your software is displayed on your screen, the first thing you want to do is add your text to the software. Press *Extend Corpus*. For *Add Single Text File*, you need to find the place where you saved your text. Follow the assistant through. Do not forget to press *Incorporate All*, otherwise your text is there, but you cannot work on it.

The next thing to do is *Add Layer*, to add an annotation scheme. The assistant will ask you to *Provide A Name*, type in *transitivity* or *theme* or any other name. Next, choose *Annotate Segments* and *Plain Text Segments*. Now, either choose *Create New Scheme* or *Copy Existing Scheme* if you have one. For copying previous schemes, you need to find the project folder of the previous project, inside of which there is a folder called *Schemes*. Open that folder. Click on the scheme you want, then open (or double-click on scheme).

If you want to look at, or change, your annotation scheme, press *Edit* under *Layers in this project*.

Now you are ready to annotate your text. Click on the blue (or purple) button with the name for the layer. Once you have fully annotated the text, the button will turn white to indicate there are no unannotated segments left. This methodology is called c*omputer-assisted manual annotation*.

Once you have finished your annotation, you want to look at the results. You can go to *Statistics,* choose *Describe A Dataset* and instead of *General Text Statistics*, choose *Feature Coding* and *Local*. Then press *Show*. Or you can go to *Search,* click on *theme* or *transitivity*, and a list will appear. Choose from the list what you want to have displayed and click on it. Press *Show*.

Do not forget to put a safety copy of your UAM project on a memory stick or second computer or hard disk.

If you want to send your project to someone by email, the entire (the main) folder must be zipped beforehand, and if someone sends you a zipped UAM project folder, you must unzip / unpack it, otherwise it will not work properly.

Doing the annotation properly is only the first step, although it requires a lot of work. The more interesting second step is looking at the results and doing an interpretation. What does that tell you? The second step shows how much you thought about the topic, and how much time you gave yourself to do so.