Quantitative register analysis across languages

STELLA NEUMANN
Universität des Saarlandes

1. Quantifying register analysis

The aim of the present paper is to show to what degree quantitative processing of a bilingual corpus of varied registers can serve as a basis of learning more about register variation both within a given language and across the languages English and German. Register here is understood in the systemic sense as functional variation of language use in a given context of situation. The study builds on the view of register mediating between the options of the language system and its realisation in textual instances. In this view, register pre-selects, or, in Matthiessen's terms (1993), blocks certain options of the system. Registers can thus be described as subsystems of the language system or, when viewed from below, as types of instances (cf. Matthiessen, 2001).

The study is based on the assumption that the description of types requires generalisations to be made from a representative number of exemplars of the assumed register, i.e. the examination of a representative corpus. Studies based on a limited amount of textual data, i.e. qualitative studies, can analyse their data in an interpretative fashion, using categories that can only be assigned on the basis of human interpretation. Quantification, however, requires a different approach. The processing of large amounts of textual data calls for the use of computational tools to automate the analysis to the greatest extent possible. Those tools capable of delivering reliable results cannot provide the level of interpretation that can be achieved by a human analyst. The quantitative approach thus intrinsical-

ly splits the process into two activities: Firstly, data analysis, which often does not follow theory-internal categories and, secondly, the human interpretation process. Here, the main challenge consists in formulating queries geared to the complexities of the linguistic annotation and in relating the findings to more abstract and theory-driven concepts. This approach is pursued by the CroCo Project (cf. Hansen-Schirra et al., 2007) and also by the present study which forms part of the CroCo Project[1].

The remainder of the paper is organized as follows: Section 2 discusses the state of the art. Here, we will first discuss Douglas Biber's work on cross-linguistic register variation and clarify why founding register analyses on a theoretical framework is desirable in terms of both theory and methodology. We will then discuss register linguistics in the systemic functional framework as a means to address the issues outlined above, before moving on to explain that this framework requires some specifications for the quantitative study of registers. Section 3 is dedicated to the research methodology and its implications for corpus design and exploitation. Section 4, the bulk of the paper, exemplifies the specifications for two registers in English and German. We will pick out exemplary features for each of the three register parameters field, tenor and mode of discourse to discuss the available quantifications and interpret the findings. The paper will be rounded off by some conclusions and an outlook on future work (section 5).

2. State of the art

2.1 Cross-linguistic register variation

The quantitative study of register variation is not a novel object of investigation. Particularly Douglas Biber's work (1988, 1995) in this field has had a major impact – and has also attracted criticism (cf. Lee, 2009). In what follows we will introduce this seminal work and explain some shortcomings that an alternative approach on the basis of systemic functional register analysis could overcome.

In 1988, Douglas Biber published a corpus-based study on variation across speech and writing. This study proposes a genuinely inductive approach to the investigation of the continuum of spoken and written discourse in English. Biber selected 67 mainly lexical and grammatical features mentioned in the relevant literature as indicators for this continuum. He combined the LOB corpus (Johansson et al., 1978), the London-Lund corpus (Greenbaum and Svartvik, 1990) and – in Biber (1995) – a collection of personal and professional letters and annotated the whole corpus with linguistic information on the 67 features.

Biber proceeded to process the frequencies of the features in a factor analysis to determine co-occurrences of the linguistic features. The resulting groupings of co-occurring features are interpreted in the form of factors or dimensions. Factor analysis is one of the multivariate statistical techniques serving to reduce large data sets to a smaller number of factors that are assumed to reflect patterns of relationships in the data. It thus works largely inductively since, apart from the theoretical assumptions implicated in the selection of the linguistic features analysed in the data sets, factor analysis does not build on any further derivation from abstract concepts.

The whole research, from the linguistic features to their interpretation in terms of dimensions, is taken up in Biber (1995) under the heading of register variation. It needs to be stressed here that Biber does not alter the research layout in this more recent study: In fact, he does not report any new analysis of new features but rather reports the results of the 1988 study of spoken and written language changing the perspective of the interpretation to a more general comparison of registers in English. He also retains the same seven dimensions from the 1988 study, i.e. Involved versus Informational Production, Narrative versus Non-Narrative Concerns, Explicit versus Situation-Dependent Reference, Overt Expression of Persuasion, Abstract versus Non-Abstract Information, On-line Informational Elaboration, Academic Hedging. Biber explains that the variation between different registers is reflected in their position along these seven dimensions.

The main focus of the 1995 study is on the comparison of four different languages: English, Nukulaelae Tuvaluan, Korean and Somali. After the discussion of the dimensions identified independently for each language, Biber compares the dimensions across the four languages. Due to his inductive method the dimensions are not directly comparable across the four languages. Biber therefore introduces so-called communicative functions without further clarifying this term. With respect to the oral/literate dimension[2] he names four types marking interactiveness, production circumstances, stance and finally functions particular to each language. Two additional functions refer to narration and argumentation/persuasion. Biber then compares the monolingual dimensions, their relevant features and characteristic registers in the four languages along these functional types. In the next step, he compares what he calls 'equivalent registers' (1995: 237) across the languages on the basis of their respective position on the dimension. For instance, face-to-face conversations are compared along the six functional types. A final set of intralingual interpretations is concerned with the internal variation of the registers and text types which reflect similarities of linguistic features.

Biber thus offers a far-reaching discussion of different aspects of intra- and interlingual register variation. His approach is innovative in that it does not only include quantitative, bottom-up analyses and detailed interpretations but also introduces statistical techniques into linguistic interpretation. His work, however, also raises a number of questions: the linguistic features Biber uses for the general analysis of English registers are taken from the 1988 study on the spoken-written continuum. This results in a bias of the whole study towards this continuum (see also Lee, 2009), even though many of the features under investigation are not restricted to the spoken-written continuum or, to put it in systemic terms, to medium of discourse. A more serious reservation is that certain important features serving as indicators for other aspects of register variation are not included.

From the point of view of statistical techniques a question with Biber's – very convincing – findings remains, namely which level of measurement, i.e. the nature of information contained within numbers assigned to objects, is appropriate for linguistic frequencies. It determines the statistical technique apt for the purpose of investigating the respective data (cf. Butler, 1985: 12). Nominal (i.e. discrete variables) is the lowest level, referring to names that are given to objects. Parts-of-speech are, for instance, categorical labels for words. Interval level, by

contrast, allows insightful comparisons of differences between arbitrary pairs of measurements. Vermunt and Magidson (2005) remark that factor analysis is frequently used with variables that do not achieve the interval level and may consequently bias the parameter estimates[3]. However, Baroni and Evert (2009) explain that it is necessary to compute relations between the frequency of the respective linguistic feature and the number of linguistic units it occurs in. This way, comparability between the differing texts in a corpus is ensured and, as a side effect, the data reach interval level. In this view, Biber's use of factor analysis appears admissible.

Biber's inductive approach is more problematic from the contrastive perspective – not to mention the general issue of using features that have been deduced in other theory-based studies, which again highlights the fact that there is no such thing as an exclusively inductive study. Biber has to introduce communicative functions which are in turn determined inductively on the basis of the findings supplied by his contrastive comparisons. As a result of the functions being derived from these findings, their scope is limited to generalisations that merely apply to the features originally investigated. A combined deductive and inductive methodology – particularly one that is functional[4] – would safeguard a systematic and comprehensive contrastive comparison, while still taking into account much-needed insights from the study of empirical evidence.

A theoretical framework well-suited to complement and motivate bottom-up analyses is the systemic functional one, certain aspects of which are related to Biber's work, for instance, the functional view of language and the role of contextual information. As will be shown below, the concept of register in systemic functional terms requires some adaptations for quantitative studies, for which Biber's approach can serve as a valuable example.

2.2 Register linguistics in the systemic functional framework

The concept of register stems from the notion that the context of situation determines language use. For the description of a register three parameters were introduced (Halliday et al., 1964; Halliday, 1978; Halliday and Hasan, 1989): *field of discourse* specifying the topic of the linguistic exchange in the given situation, *tenor of discourse* characterising the relationship between the participants in the situation and finally *mode of discourse* describing the way in which the exchange is transmitted. Just as situations tend to recur and thus form types, registers represent recurring ways of using language in a given situation. The language system can even be grouped into typical co-occurrences and non-occurrences according to the respective situation. Registers can thus be described as sub-systems of the language system or, when viewed from below, as types of instantiated texts reflecting a similar situation (cf. Matthiessen, 2001). The concept of types (of situations or of instantiated texts) implies a certain frequency of recurrence and repetition of features or patterns. In a methodological sense this means that, strictly speaking, a description of these types requires a quantification of their characteristic features. Otherwise, we can only describe a given specimen of the assumed type, which does not permit any statements on the type itself.

The global definition of the three parameters remains essentially unchanged with some work having been carried out in this area aimed to provide more detailed elaboration (Matthiessen, 1993 gives a comprehensive overview of the development of register theory). Halliday (1978) refers to possible subdimensions specifying the three variables. Again, from a methodological point of view, these subdimensions are necessary concretions of the highly abstract and broad register parameters. The subdimensions constitute latent variables, variables that are still too abstract to be directly observable in instances. These subdimensions are briefly introduced in Halliday and Hasan (1989), further discussed with a strong focus on tenor of discourse in Martin (1992) and taken up and sometimes modified in several descriptions of individual registers (cf. for instance several contributions in Ghadessy, 1988 and 1993). Butt (2004) approaches the specifications of the register parameters from a semantic point of view. He develops a 'semantic network', i.e. a system network representing semantic choices that contains subdimensions as well as more fine-grained semantic elaborations. Although Butt's (2004) semantic networks include specifications of the subdimensions, even in the most fine-grained branches of the networks it remains unclear how we can relate the criteria to observable data in reality, i.e. map the networks on instances (of situations, registers, texts).

Subdimensions represent a crucial step in deriving concrete indicators in terms of features observable in texts. While an example-based study may interpret these latent variables directly on the basis of human interpretation (even if this may leave the relationship between indicators and variables somewhat implicit), a quantitative study requires operationalisations in the form of observable indicators, i.e. an additional step of derivation.

For instance, the dimension of social hierarchy, sometimes also called social role relationship, has been named as one concretion of tenor. The options of this relationship have been described as either equal or unequal. However, we cannot observe this directly in linguistic data. Another level of abstraction is required before arriving at the level of concrete, measurable indicators which may then be interpreted as pointing to either an equal or an unequal social relationship. One of the criteria used for the subclassification of the abstract construct is level of expertise (cf. Steiner, 2004), which is still not an observable variable. We have to define lexico-grammatical realisations, for instance technical vocabulary. The explication of all intermediate steps results in a transparent derivation of concrete, observable (and potentially also measurable) linguistic data capable of serving as evidence for the abstract construct.

Martin (1992), for instance, names – particularly for tenor – subcategories and relates them to concrete features on the different strata. Steiner (2004) gives an overview of a wide range of subdimensions naming possible operationalisations for most of the subdimensions. Both Martin and Steiner, however, confine themselves to listing the features without stating which realisation points to which interpretation. For instance, naming the mood system as an operationalisation of the distinction between spoken and written medium (under mode of discourse) does not tell us which mood option points to which end of the spoken-written continuum.

The steps required to derive observable indicators resemble the inter-stratal realisations Matthiessen (1993) explicates. Conceptualising the theoretical framework of register analysis, he describes the way in which

> language in context is interpreted as a system of systems ordered in symbolic abstraction. (…) Each system has its own internal organization (…) but it is related to other systems in a realizational chain: it realizes a higher system (unless it is the highest system) and it is realized by a lower one (unless it is the lowest system). This chain of inter-stratal realizations bridges the gap between semiotic in high-level cultural meanings and the material, either in speaking or in writing, through a series of intermediate strata. (Matthiessen, 1993: 226)

In the same way that the different strata bridge the gap between the high-level meanings and low-level material, the different methodological steps in deduction reduce the distance between the abstract construct and the concrete linguistic evidence.

We suggest that there is a difference between giving an example for an intellectual application of a given category of register analysis and supplying an operationalisation for an observable indicator of the same concept. While an example only serves as an illustration of the interpretation and can be difficult to repeat by a different analyst, an operationalisation should allow replication in the sense of applying the same operationalisation to a different instantiation and – provided the linguistic constellation is comparable – yielding the same results.

3. Research design

3.1 Deriving indicators

This study is thus guided by the idea that the indicators used to describe the registers under investigation should be replicable, i.e. applicable to other registers, possibly also in other languages. As mentioned above, qualitative studies offer the advantage of permitting in-depth analyses of a given feature. In quantitative studies, there is a risk of using indicators that are easily countable but do not allow sensible and reliable conclusions on the abstract concept in question. Type-token ratio is a case in point. This measure is frequently used in quantitative studies as a means to characterize differences between registers. However, it is subject to many, sometimes offsetting factors such as diverging text lengths, frequency of function words etc. – not to mention contrastive differences in writing conventions, a major obstacle in cross-linguistic studies based on shallow features. The present paper explores to what degree automatic analyses and queries enable (and advance) register analysis.

Any quantitative empirical enterprise should strive to satisfy three quality factors of empirical work:

- reliability, i.e. repetitions of the study should yield the same results. In manual analyses this means that different analysts should interpret a given feature in the same way. In automatic analyses, the computer tool should always produce the same results on the same data;

- intersubjective verifiability (replacing the elusive concept of objectivity): the categories should yield the same results irrespectively of the person carrying out the analysis. The more human interpretation is required in a given study, the more important this concept becomes, and
- validity, a concept that has two equally important aspects: is the choice of method appropriate to the phenomenon under investigation? Do the chosen indicators actually measure the concept under investigation (and not a confounding factor)?

These factors are taken into account in this study by making the features as transparent as possible, thus permitting replications. Reliability is further supported by the use of automatic tools that may not be faultless but repeat both correct and incorrect results in a traceable manner.

3.2 The CroCo Corpus

The research presented here was carried out in the framework of the CroCo project concerned with the investigation of typical properties of translations as compared to originals in the language pair English-German. The investigation of translations requires controlling the context of situation, i.e. the register, in which the translations are produced in order to exclude confounding variables influencing the make-up of the translations. The corpus compiled for these purposes, the CroCo Corpus, meets this requirement by including texts from eight different registers that are translated on a regular basis: political essays (ESSAY), fictional texts (FICTION), instructional texts (INSTR), popular scientific writings (POPSCI), shareholder communication (SHARE), prepared speeches (SPEECH), tourism leaflets (TOU) and websites (WEB). In addition to being translated in both directions, these registers foreground different registerial variations (cf. Neumann and Hansen-Schirra 2005). The corpus is thus divided into four subcorpora: English originals (EO), German translations (GTrans), German originals (GO) and English translations (ETrans). In line with Biber's calculations (1990, 1993), each register in each of the subcorpora contains at least 10 texts of 3,125 words at the most. The complete corpus thus contains approximately 1 million words, i.e. 31,250 words per register in each of the four subcorpora. The core corpus is complemented by two small reference corpora in both languages, each containing 2,000 word samples from 17 different registers. The reference corpora are register-neutral in that they represent a wide registerial spread. Although 17 different registers cannot reflect the complete register variation in German or English, they can at least serve as a basis of comparison for determining specific characteristics of a register under investigation. For the purposes of the CroCo project, the reference corpora described in Neumann (2003) were amended by two additional registers (court decisions and travel guides) and doubled in size. The whole corpus is annotated with various types of linguistic information (see section 3.3).

There is one paradox in compiling a corpus for the study of registers that continues to spark discussions[5]: The selection of the texts in the CroCo Corpus is

based on the assumption that they belong to the same register, yet only an in-depth analysis of their defining features can determine whether they actually do. CroCo addresses this aspect by including a cursory register characterisation in the metadata (see section 3.3). Additionally, statistical processing of the corpus findings can help identify outliers.

Since the present paper focuses on originals from the registers of SHARE and FICTION, these two parts of the corpus are explained in some more detail:

- SHARE contains letters from the CEOS of various companies to their shareholders. These texts inform the shareholders on the performance of the company in the last accounting period. Beyond this, they also have a persuasive character because their goal is to convince the shareholder of the successful or, under difficult circumstances, necessary management. Although the managers as authors are the experts, they address the reader in a respectful if not formal way, since the management is accountable to the shareholder as the addressee. The texts are in written mode. The German subcorpus contains 11 texts; the English subcorpus contains 13 texts.
- The FICTION register contains contemporary literary texts of which translations into the respective other language exist. An additional criterion for selecting the texts was literary quality with the assumption that sophisticated literature stretches the options of lexico-grammar to a greater degree and thus contains more linguistic variation than light fiction. The field of these texts is diverse, their audience as well. Again, all texts are in written mode. Each of the two subcorpora contains 10 samples from longer fictional texts.

One might argue that the character of literary texts as creating an imaginary world and situating the (linguistic) action within this world prohibits grouping – and analysing – these texts as a register in its own right. However, like any other register, literary texts also reflect a given context of situation including authors/writers and readers. As Halliday and Hasan (1989: 40) argue, 'we are never selecting with complete freedom from all the resources of our linguistic system'. This should also apply to literary texts which despite their originality reflect the situation and can be said to be related to other literary texts in terms of intertextuality.

3.3 SFL-interpretation of non-SFL annotation

The CroCo Corpus contains several annotation and alignment layers on word, chunk, clause and sentence level, in addition to metadata on the basis of the TEI standard[6] for each text in the corpus. The metadata include a specification of the publication and translation as well as a first, very shallow register analysis based on inspection (Klinger et al., 2006). The annotation on word level comprises part-of-speech tagging and morphology. On the chunk level – our neutral term for groups and phrases – the annotation is carried out manually with the help of a tool for creating linguistic annotation. The annotation comprises the formal classification in terms of phrase structure as well as the assignment of gramma-

tical functions on the highest level within the sentence [7]. The various annotations are stored in individual files making it possible to retrieve either each type of linguistic information separately or in combination with other files (cf. Vela et al., 2007 for a detailed description of the corpus enrichment).

The annotation is thus clearly not compliant with systemic functional categories. A main assumption of the current research is the interpretation of theory-driven abstract concepts on the basis of indicators that are not specifically part of the same theory (see also Hansen-Schirra et al., 2007), an assumption that is further corroborated by the present study. The resource used for this purpose is the CroCo Corpus with all its annotation and – where applicable – its alignment. The guiding principle of its compilation is to make the resource available to a broad range of research questions. This means that linguistic information was added to the texts in the corpus in such a way as to minimise any bias towards a particular linguistic theory. A theory-driven view on the corpus data comes into play in the form of deriving indicators by way of queries into the corpus, and their interpretation in terms of the theory-guided research question. For the present study this means that theory-internal concepts like the thematic structure of texts are queried on the basis of the available information, i.e. grammatical functions in combination with word order (see section 4.3.3).

4. Some findings for two registers in English and German

In what follows, we will discuss operationalisations for all subdimensions of the three register parameters and pick out one subdimension under each parameter to exemplify the results yielded from the corpus exploitation.

4.1 Field of discourse

Field is the variable covering the description of what the register, i.e. a given situational context, is about, which experiential meanings are conveyed in the given situation, with what goal. Halliday and Hasan (1989: 56) call it the nature of the social activity and the goals to be achieved by this activity. It is related to the ideational metafunction of language and is thus expected, in Martin's (1992) words, to put the systems of transitivity, circumstantiation and agency at risk. Steiner (2004) names the subdimensions of *experiential domain*, *goal orientation* and *social activity* with some operationalisations for each dimension. *Social activity* is a rather problematic category that is sometimes used instead of *experiential domain* (e.g. by Halliday and Hasan, 1989), thus highlighting the social and action-oriented character of field. Sometimes it is also used to refer to a rather abstract idea of activity relevant within society like production, exchange, communication, reproduction and consumption (see Steiner, 2004). There seems to be a rather wide gap between the textual instances of a given register and very general types of activity in this sense. It remains unclear how these types of activity can be related to observable indicators in a methodologically sound way. While it

would be interesting to investigate the types of activity represented by various registers, this aspect of register analysis is left aside here due to methodological concerns.

4.1.1 Experiential domain

This category's task is to elicit the subject matter of the register. It is related to what Halliday and Hasan (1989) call the nature of the social activity. However, they do not expand on how to elicit this activity, neither does Martin (1992), who proposes a taxonomy of fields. While this is a very interesting approach, he concedes that he can only provide a very sketchy coverage. We argue that a taxonomy must inevitably be too abstract to be applied to the analysis of concrete texts, particularly because the field covered by a text attributed to a given register will be rather specific. So describing the subject matter of a register from below seems preferable even if this description is more particular and thus less generalisable. Steiner (2004) proposes the following operationalisations: lexical fields, terminology, lexical chains, transitivity, headings, paragraphing, expressions of time, sometimes perspective and *Aktionsart*. Of these, the most insightful operationalisations seem to be the first four. In the present study, we will therefore focus on these.

Lexical field and terminology

Terminology is interpreted here as the specific vocabulary used in a given (professional) domain. Broadly speaking 'lexical field' refers to vocabulary belonging to the same semantic area. Obvious candidates for specifying this are repetitions (also in compounds) and synonyms of a given frequently occurring word. In a broader interpretation all sense relations as described by Halliday and Hasan (1976) in relation to lexical cohesion should cover the relevant semantic area. Since a manual analysis of sense relations is not feasible as part of a quantitative study, we will concentrate on interpreting repetitions[8]. Repetitions of lexical items are retrieved by counting the frequency of lemmas. In CroCo, lemmatisation is part of the morphology annotation. The script used for this query only retrieves those items annotated with open-class part-of-speech tags. It thus yields the most frequent lexical words in the respective corpus. Table 1 displays the 10 most frequent lexical words in the two registers SHARE and FICTION in both languages[9].

EO__FICTION		EO__SHARE		GO__FICTION		GO__SHARE	
say	145	year	216	sagen	109	jahr	179
go	100	company	208	kommen	78	neu	142
man	96	business	191	vater	76	unternehmen	120
see	89	service	168	sehen	73	gut	98
come	82	customer	126	gehen	68	euro	90
day	80	percent	113	jahr	59	%	87
know	79	new	106	machen	59	konzern	79
time	64	product	106	kind	58	weit	78
look	61	market	105	geben	55	aktionär	74
take	60	baker	100	kopf	51	hoch	69

Tab. 1 The 10 most frequent lexical words

In general terms, there seems to be a tendency towards lower frequencies in the two German registers than in the English registers. This may reflect a tendency of the German language towards more lexical variation and a tendency of the English language towards more lexical repetition (as reflected also by the relative importance of general nouns in English). The comparison of lexical fields within each register across the two languages reveals registerial similarities:

- The texts in the context of corporate communication display a rather frequent occurrence of terminology from the field of economy. We find lexical items like *company, business, service, customer, product* and *market* in the English register and related items like *unternehmen* (*company/ business*), *konzern* (*corporate group*) and *aktionär* (*shareholder*) in the German register as well as less specific vocabulary that can still be attributed to the subject matter of reporting to the shareholder like *year* and *percent* (*jahr, %* and *euro* in the German register). Among the lexical items we also find adjectives with an evaluative character like *new* (*neu*), and the German *gut* (*good*), *weit* (*far*) and *hoch* (*high*).
- The fictional texts do not reveal clear lexical fields pointing to greater lexical diversity. However we still find interesting patterns obviously common to the various texts: In both languages there are many verbs, the most frequent word being *say* (*sagen*), which may point to a verbal process[10] (see Transitivity below). Other verbs seem to be material processes like *go (gehen*), *come* (*kommen*), *take* and, in German, *machen* and *geben*. Finally, there are lexical verbs, particularly in the English register, representing mental processes like *see (sehen*), *know* and *look*. Additionally we find general nouns like *man, day* and *time* in the English register and superordinates like *vater* (*father*), *jahr* (*year*), *kind* (*child*) and *kopf* (*head*) in the Ger-

man register. A possible explanation for the prevalence of verbs could be a tendency to more procedural descriptions in FICTION. The overall number of verbs (including those verb classes that do not appear in Table 1 because they are counted as function words) shows that in both languages FICTION contains clearly more verbs (6,102.57 in English and 5,399.45 in German vs. 4,859.24 and 4,059.73 in SHARE; normalised frequencies). We should, however, bear in mind that the greater lexical spread in FICTION[11] overly emphasises those lexical items common to the different texts contained in the corpus, items that are rather general in meaning.

Transitivity/lexical verbs

Transitivity is concerned with the goings-on a clause expresses (cf. Halliday and Matthiessen 2004). We are thus mainly interested in the processes a clause represents. From the point of view of experiential meaning, we are also interested in the participants and circumstances realised in the clause – and more generally in the text. However, since this analysis is very time-consuming, and we aim at a broad coverage of all indicators, we shall focus on the processes. Process types can be indicative of different kinds of linguistically realised interactions. For instance, a text which, in relation to a basis of comparison, contains significantly more relational processes will probably deal with descriptions of entities rather than with material or verbal action.

The qualitative investigator will manually assign values for process types (and possibly not exclude participant roles and circumstances), allowing a detailed view of experiential aspects of the texts under investigation. A study focussing on the quantification of register characteristics in a larger corpus cannot achieve a complete analysis of process types – interesting though it may be. The present study approximates the interpretation of process types by interpreting the most frequent lexical verbs using the same query as that for all lexical words, this time, however, only retrieving those parts of speech tagged as lexical verbs. More in-depth studies could further this analysis by loading concordance lines of e.g. the five most frequent lexical verbs into an annotation tool like the UAM CorpusTool[12] and then analysing process types for these verbs. The present study, however, is limited to a first interpretation of the lexical meaning of the most frequent lexical verbs. As mentioned before, the query does not retrieve relational uses of *be* or *have* or their German equivalents *sein* and *haben* since these receive part-of-speech tags that are sorted out as function words. In order to at least get some insight into the frequency of relational processes in the four subcorpora under investigation, we searched for the 3rd person singular form for *be* and *sein* (*is* and *ist*) in the corpora and counted the relational uses[13].The query thus leaves out other tenses as well as other relational processes.

This interpretation is manageable on a large scale and can shed some light on the goings-on, i.e. the typical process types, pertaining to one register – and possibly provide information on distinguishing differences between two registers or a comparable register in two languages. However, only a fully-fledged transitivity analysis can reveal general patterns with respect to the distribution of process types.

Table 2 displays the five most frequent lexical verbs in the two registers under investigation as well as the results from concordancing the 3rd person singular of *be* and *sein*. The lower frequencies of all verbs in the SHARE subcorpora might suggest that there are more different (and consequently less frequent) verbs in this register. In fact, these reduced frequencies rather reflect a generally reduced frequency of verbs in SHARE as compared to FICTION.

EO__FICTION		EO__SHARE		GO__FICTION		GO__SHARE	
say	123.0	"is"	178.1	"ist"	171.8	"ist"	173.3
"is"	104.0	make	76.7	sagen	108.4	erreichen	38.7
see	87.5	continue	62.6	kommen	73.6	führen	35.7
go	85.0	increase	53.8	sehen	70.5	machen	35.7
come	71.8	provide	50.3	gehen	64.4	liegen	34.7
know	71.0	help	48.5	machen	56.2	bleiben	33.6

Table 2. 5 most frequent lexical verbs plus relational *is/ist* (normalised frequencies)

Both the English and German fictional texts use more different verbs that can be assigned to a greater range of process types, the most frequent lexical verb in English literary texts being *say*, followed by the relational use of the 3rd person singular form of *be*. The remaining verbs in the English texts point to mental (*see* and *know*) and material processes (*go* and *come*). German fictional texts most frequently use the relational process represented by the 3rd person singular of *sein*. The second most frequent lexical verb is *sagen* (*say*). In the German register, there seems to be slightly more emphasis on material procedures with three verbs pointing to this process type (*kommen*, *gehen* and *machen*). *Sehen*, the German equivalent for *see*, takes the third place in the list, representing a mental process. All of these verbs in both languages are rather general in meaning. As mentioned in the context of lexical fields, the lexical verbs in the register FICTION account for more verbs than in the SHARE register in both languages.

As to the register SHARE, in both languages the relational *is/ist* are the most frequent verbs. While they are approximately comparable in frequency with the German FICTION register, their frequency stands out starkly in the economics texts as compared to the other most frequent lexical verbs. From their lexical meaning all of these remaining lexical verbs in both languages point to material processes. A detailed analysis will probably result in a more differentiated picture with some of the realisations being used as relational processes as in (1).

(1) Unsere einmalige Fähigkeit liegt in der Vernetzung der drei Bereiche. (GO__SHARE)

The English verb *continue* often expresses aspect (see example 2) and is particularly typical for corporate communication where the management signals to the shareholder that it will maintain its success or carry on with a consolidation process etc.

(2) To become the most respected global financial services company, we must continue to advance our strategic goals-to expand our international franchise, to continue to grow our consumer business, and to ensure that our corporate and investment banking business is best in class. (EO__SHARE)

While this analysis cannot capture the overall frequencies of the different process types, we can identify similarities in terms of verbs expressing certain goings-on within each of the two registers in English and German as well as differences between the two registers irrespective of the languages.

Lexical chains

Closely related to lexical fields are lexical chains. While fields are mainly semantically defined, chains are identified by the patterns lexical items build within a text. The most frequent vocabulary gives us an important indication of the experiential domain covered by the register. It is straightforward to query vocabulary by running frequency word lists over each individual text with the help of a concordance tool like WordSmith (Scott, 2004). Beyond the mere frequency of the respective lemma in a word list, lexical chains, i.e. sequences of related words (Morris and Hirst, 1991), provide information on whether a frequent lemma forms a topical thread throughout the text or whether it is repeated only locally. In the former case the chain underpins the lemma's relevance for determining the referential meaning of the whole text. In the latter case the lemma forming a chain represents only a local strand of referential meaning.

For a comprehensive analysis of lexical chains in each text, the corpus has to be annotated with sense relations since semantically related items like synonyms, hyponyms etc. must be interpreted as contributing to a lexical chain (e.g. with the help of a WordNet and GermaNet annotation, which is currently not available for the CroCo Corpus). Teich and Fankhauser (2004) describe how chains of lexically related words can be processed and analysed automatically on the basis of WordNet. The present account concentrates on chains created by repetitions of the same lemma. In Halliday and Hasan's terms (1989: 84), this represents a similarity chain made up of items that refer to members of non-identical but related classes.

> The items in a similarity chain belong to the same general field of meaning, referring to (related/similar) actions, events, and objects and their attributes. Halliday and Hasan (1989: 85)

Of course, 'how much of such a [semantic] grouping will appear in the shape of similarity chains in a particular text' (Halliday and Hasan, 1989) is open to variation.

Our query looks for each lemma and the ID of the sentence in which the given lemma appears. Since the sentence IDs correspond to a sentence's linear position in a text, the IDs of two consecutive occurrences represent the distance between the two links in the chain. Morris and Hirst (1991) also mention the span from the first to the last occurrence of the lemma within the text. This span is an additional cue to the relevance of the chain for the overall referential meaning of the text. We interpret the repetition of the lemma as a link in a continuous chain if

the distance between the occurrences is less than four sentences. If the distance is longer, the new occurrence of the lemma is interpreted as a return to an existing chain (Morris and Hirs, 1991: 32). Chain lengths thus result from the addition of occurrences in sentences less than four sentences away.

	EO		GO	
	FICTION	SHARE	FICTION	SHARE
av. no. of sentences	181.40	114.54	206.30	157.64
av. frequency	21.90	37.08	20.50	31.64
av. span	157.90	106.69	165.80	141.64
av. chain length	1.90	3.8	1.86	3.29
av. distance between occurences	7.49	3.15	8.94	2.73

Table 3. Lexical chain statistics for most frequent lexical word per text

Table 3 displays statistics for the most frequent lexical word in each text of the SHARE and the FICTION sub-corpora. First, the average number of sentences per text is of interest since it gives us the base line for the span of lexical chains as well as the distance between occurrences. Besides this information, this figure is of little relevance to us since the texts in the FICTION subcorpora are samples. We can see that the average frequency of the most frequent lexical word per text is higher in the SHARE texts in both languages than in the FICTION corpora.

While occurring less frequently, the words under investigation in the fictional texts span longer passages of text than in the SHARE texts. Consequentially, the distance between the occurrences is clearly higher in the FICTION texts as compared to the SHARE texts. Inspection of the figures in Table 2 suggests that there may be a correlation between text lengths in terms of sentences and span of a lexical item. Nevertheless, frequent repetition of a lexical item within a short span is also plausible and would, as mentioned above, point to a local strand of meaning. The resulting chain lengths supply information on the importance of the given lexical item. A longer average length of the lexical chain signals more importance[14]. It seems plausible that full lexical repetition – irrespective of its being part of a lexical chain – can be a feature of a given register, whereas another register may rely more on pronominalisation. As will be seen below in section 4.3.3 fictional texts contain more pronouns than SHARE texts. We can therefore tentatively interpret the lexical chains in the economics texts as pointing towards more lexicalised reference. Example (3) shows how the most frequent lexical word in this text, the company's brand name, is repeated in two consecutive sentences.

(3) This environment of dynamic change poses many challenges to **Citigroup**, but also creates great opportunities to serve our customers, to provide exciting careers for our employees, and to do well for our shareholders. **Citigroup** has a long and storied history in its constituent parts, including pioneering in international banking

> at Schroders and Citibank, leading the way in many areas of trading, asset management, and investment banking at Salomon Brothers, bringing modern banking to entire nations through Banamex and Bank Handlowy, and offering thoughtful advice in wealth management to clients through Smith Barney and **Citigroup**'s Private Bank. (EO__SHARE)

A more precise picture of lexical chains will result from including sense relations beyond mere repetitions of a given lemma. This will also allow interpretation in terms of cohesion.

The indicators for experiential domain discussed here are apt to narrow down the subject matter of a given register – even in a vague register like FICTION. This became particularly evident from the intralingual comparison of the two registers under investigation. Even the investigation of lexical verbs which is rather shallow from the systemic point of view yielded valuable results.

4.1.2 Goal orientation

While the question of what goal the (linguistic) action is directed at is included in all accounts of register theory, this category is described in various ways by different authors. Like Halliday and Hasan (1989), Martin (1992) is more concerned with tenor of discourse and does not offer a detailed description. Hasan (1999: 234ff) explicates goal as an inherent aspect of human social action and thus as an important component of a text's relevant context. She points out that the concept is 'riddled with problems' mentioning among others the potential invisibility of goals. In line with this view, Butt (2004) offers different parameters for the description of goal(s) in his semantic network. He characterises goal orientation as 'an attempt to elucidate the (outward) indices of the inner controls and organization of meaning in a context' (Butt, 2004: 34). However, it remains unclear how the options in Butt's network can be operationalised in terms of observable evidence in the texts avoiding introspective interpretation by the investigator that may not be intersubjectively verifiable. It may well be the case that an interview with authors will afford insight into those goals not explicitly marked in the text but more often than not we may be limited to analysing the product of interaction, i.e. the text.

Steiner (2004) lists types of goals related to what is often called 'text types' (e.g. Werlich, 1976). While they seem to represent a rather static classification, they provide a sound and comprehensive basis for operationalisations of invisible, implicit goals. It should be possible, for instance, to identify features of argumentation in text. Possible operationalisations could be the appropriate thematic structure of the text (cf. Lavid, 1994 for an extensive discussion of thematic progression in different text types). Other observable indicators include modality, mood, voice and pronominalisation. The same derivation of indicators is applicable to the other goal types. Each goal type should thus exhibit a characteristic constellation of lexico-grammatical features.

4.2 Tenor of discourse

Tenor of discourse comprises those criteria capturing the nature of the relationship between the interactants. The relationship is analysed with respect to agentive as well as social roles borne by the interactants, their social distance as well as appraisal present in the interaction.

The interpretation of these subdimensions is subject to the following restrictions. Mostly, we will expect that one of the interactants has a more active role and that there is a (group of) interactant(s) re-acting to what the active interactant produces. The writer[15] is the active part while the reader is the re-active part. Interactants may swap roles in the course of the interaction, particularly in certain spatio-temporal constellations. Here, the analyst can interpret both the writer's and the reader's part in the interaction.

In written texts, however, the roles are fixed with the author of the text taking the agentive role and the reader being restricted to reaction[16].This has consequences both for the interactants as well as for the analyst of written texts. As Hasan (1999: 230) explains, the reader is unable to influence the process of text production. In the case of, say, published texts written without any knowledge of the concrete readers, the writer can only imagine a prototypical reader, or as Hasan (1999: 229) puts it, 'the intended addressee of this text has an imaginary being'. And finally the analyst can only look at the active participant's output, not at the anonymous reader's reaction to it.[17] From the interpretation of the findings about the writer the analyst may extrapolate on the audience. Or, in Hasan's words (1999: 238):

> where the addressee is virtual, all aspects of the interactant relation – their respective status, their social distance, the specific attributes of the addressee – are *logically* entirely *created* by the language of the text, none having a basis in reality for obvious reasons.

This means that for the analysis of texts written for anonymous readers, as is the case with the texts under investigation here, statements on the tenor of discourse will be focussed on the writer. The conclusions we draw for the reader are only indirect inferences obtained from interpreting what we assume to be the writer's projections of his/her reader(s).

4.2.1 Agentive roles

Referring to *agent* roles, Halliday and Hasan (1989: 56) elaborate that '[t]his social activity [i.e. buying food-stuffs] is institutionalised. And so the nature of the activity predicates the set of roles relevant to the unfolding of the activity'. They continue to identify vendor and customer for the text in question. The roles here seem to be related to participant roles in lexico-grammar. In Hasan (1999: 247), the focus shifts slightly to a more abstract view that helps shed additional light on the interactants' roles in communication. Referring to a mother and child interaction, Hasan explains that 'it is difficult to identify one single relation of the agentive kind which would apply constantly to the entire dialogue'. Particular-

ly in a context where both speaker/writer and addressee/reader are present in some form, the question of who is the agentive participant in the interaction becomes relevant. Agentivity, i.e. the active control of the interaction, can vary independently of the interactants' social role or their social distance, thus requiring a separate analysis. Indicators could be the proportion of turns per interactant and distribution of different mood and modality options among the interactants. In written texts where the reader remains virtual in the sense discussed above, the analysis of agentive roles is pointless since the interaction is sustained entirely by the writer.

4.2.2 Social hierarchy

Social hierarchy, sometimes also referred to as social role relationship, is concerned with the degree of control (or power) one interactant has over the other (Halliday and Hasan, 1989: 57). This may, as Halliday and Hasan (ibid) write, be 'almost by virtue of their agent role relationship', but other agentivity constellations are conceivable as well. Therefore, it is useful to analyse the two subdimensions of agentive role and social hierarchy separately. Social hierarchy allows insight into whether writer and reader hold equal social roles or whether they are rather in a hierarchical relationship. These roles should be reflected in the linguistic choices the interactants make. Social roles depend on a person's level of authority, expertise and education. Other aspects contributing to an individual's position in the social hierarchy are religion, gender, sexual orientation etc. The analysis of level of expertise, for instance, should show whether there is a difference in expert knowledge between the interactants[18]. It should be stressed, though, that an analysis of the product of interaction in the form of written texts can only give insight into the author's level of expertise as well as his/her expectation of the reader's expertise as shown by the presence or absence of explanations of technical terms. Whether the social relationship between the interactants is hierarchical or not is thus difficult to extract from a monologic product of communication.

Observable indicators for level of expertise are features of language for specialised purposes (LSP) like LSP terminology and LSP grammar. These give information about the technicality of the text which in turn requires a certain expertise at least on the part of the writer. While terminology can be detected with the help of a concordance tool, e.g. with a key word analysis in WordSmith (Scott, 2004), LSP grammar can be queried on the basis of the CroCo annotation using phrase chunking as well as sentence and clause segmentation. Grammatical structures typical for LSP texts have been described as packing more information into noun groups and at the same time reducing the complexity of clause structure (cf. Halliday and Martin, 1993; Ventola, 1996). Steiner refers to this phenomenon as informational density, assuming that informationally dense texts have a high proportion of '"Intermediate phrase types" (groups, phrases, rather than words or clauses) per clause' (2005: 22).

Our query retrieves the number of phrases per clause and sentence, the number of words per sentence, clause and phrase etc. In LSP registers we would ex-

pect less grammatical density in terms of fewer clauses and more chunks per sentence, reflecting their tendency to package information into nominal phrases rather than spread it over clauses. Consistently with this expectation, we would also expect a higher number of words per chunk. Table 4 displays average figures for these proportions.

	EO		GO	
	FICTION	SHARE	FICTION	SHARE
chunks per sentence (av.)	5.35	4.87	4.68	5.39
chunks per clause (av.)	1.95	1.99	2.35	3.19
clauses per sentence (av.)	2.74	2.45	1.99	1.69
words per sentence (av.)	20.36	24.05	16.07	20.31
word per clause (av.)	7.43	9.81	8.06	12.02
word per chunk (av.)	3.81	4.94	3.43	3.77

Table 4. Proportions of grammatical density

The German registers confirm this expectation: on average, fictional texts contain slightly more clauses per sentence. On the other hand, the economic texts contain more chunks per sentence and clause as well as more words per sentence, clause and chunk. The comparison in English is somewhat less clear-cut. While SHARE texts do have more words per sentence/clause/chunk, FICTION texts contain more chunks per sentence, with chunks per clause being level. As predicted, the fictional texts contain more clauses per sentence. Examples (4) and (5) below reflect the differences between the two registers. In (4), the 44 words are spread over 5 sentences, 11 clauses and 27 chunks many of which consist only of single words. Example (5) is one single sentence with 41 words, 3 clauses and 9 chunks. The chunks contain quite complex nominal groups. The first one is both pre- and postmodified. Here, the premodification represents a participial construction typical for German LSP texts which would be realised as a postmodifying relative clause in more general language use. The last group in the sentence is discontinuous with the predicator inserted between the head and the postmodification.

(4)[I] [had to] [be] [quiet] [all the time]. [Twice] [they] [took] [me] [and] [I] [stayed] [the whole three weeks]. [It] [stopped], [though]. [My mother and father] [talked] [about quitting] [but] [they] [did][n't]. [They] [got] [my grandmother] [to move in and watch over me]. (EO__FICTION)

(5)[Mit der quer durch das gesamte Unternehmen gehenden Bildung kleiner, marktnaher und mit hohem Verantwortungsumfang ausgestatteter Einheiten] [wurden] [nicht nur] [die Flexibilität der Gesamtorganisation] [deutlich] [erhöht], [sondern auch] [unternehmerischer Geist] [freigesetzt], [der sich nicht zuletzt in einem völlig neuen Kostenbewußtsein niederschlägt]. (GO__SHARE)

The forms of grammatical density we can observe in the SHARE texts characterize the writers' level of authority expressed in their command of specific language beyond terminology. The writers do not adapt to general language use but demand a certain degree of familiarity with this kind of language use on the part of the reader.

Steiner (1998: 243) adds other realizations like modality, mood, forms of address, formality of text, level of education in text. This latter aspect again requires additional operationalisation in the form of e.g. elaborate vocabulary, complex/intricate grammatical structures, lexical density etc.

4.2.3 Social distance

The subdimension social distance encodes the interactants' relationship based on their mutual interactive history. Butt (2004: 16) characterises this category as classifying the extent of the relationship between the participants in terms of density as well as formality of context. He also takes into account whether the participants have or can be expected to have shared and distinct codes. Martin (1992) uses the term 'contact' for this subdimension. It represents, as Halliday and Hasan (1989: 57) write, a continuum whose end-points they call maximal and minimal. House (1997: 41f) specifies this continuum on the basis of Joos' (1961) categorisation of levels of formality, namely frozen, formal, consultative, casual and intimate. In the framework of the CroCo project this categorisation is slightly modified to cover those styles not marked with respect to formality, mainly because the relationship between writer and reader is not realised explicitly by any linguistic means. The categorization thus contains (in order of increasing distance): intimate, colloquial, casual, consultative and formal[19]. For each of these options observable indicators have to be determined. Martin (1992) mentions tone, accent, ellipsis, vocation and terminology as indicators. Steiner (2004) lists tagging, forms of address, modality, accents, dialects and sociolects. As for goal orientation (see section 4.1.2) we should be able to describe profiles for each of the options of social distance reflecting certain values of the observable indicators.

4.2.4 Appraisal

Appraisal in SFL theory is concerned with those features of interaction which contain evaluative meaning, be it emotional (affect system), ethical (judgement) or aesthetic (appreciation). Concentrating on affect, Martin (1992) names tone, attitude, comment, intensification, repetition, mental affection, manner degree and attitudinal lexis as indicators. In appraisal theory[20] the interpersonal force the writer attaches to an utterance (graduation) and those meanings which vary the terms of the writer's engagement with their utterances (engagement) are included as well. Steiner (1998) lists lexical selections, grammatical choices and rhetorical devices such as repetitions, parallelisms etc. as operationalisations. As to lexical selections, the interesting and challenging part is to tease out not only

explicitly evaluative lexis but also implicitly conveyed evaluative meaning in a systematic way.

Recent years have seen various studies of this highly important aspect of social interaction (cf. for instance Thompson and Hunston, 2003; Martin and White, 2005). However, it remains unclear how to identify indicators that are adequately operationalised to be processed on a large amount of data. Moreover, this subdimension particularly highlights issues in contrastive comparison. Little work has been done in comparing the options of evaluation in English and German (one exception being Bublitz, 1978). The study of the language of evaluation in German (under the heading of 'Sprechereinstellung') is mainly concerned with modality and does not operate in the systemic framework.

Since it is not within the scope of the present paper to provide a more detailed comparison, we will only discuss some possible indicators. Possible and comparable indicators are mental processes included in the present analysis under the heading of lexical verbs – evaluative lexis and evaluative patterns (cf. Bednarek, 2007). This latter indicator works on the basis of pattern grammar and aims at extracting word order patterns characteristic of evaluative contexts. It is presumably due to language typological differences between English and German that this feature works better for English than German with its more flexible word order, which allows more variation in terms of positional patterns. It is possible to replicate Bednarek's (2007) evaluative patterns using the CroCo annotation, also for the German language. However, these structures are rare in German suggesting that either evaluative meaning does not play the same role in German texts as in English ones, or that – more plausibly – patterns do not adequately reflect evaluative meaning in German.

4.3 Mode of discourse

The final of the three register parameters is concerned with the organisation of language to reflect the social action between writer and reader. This parameter is based on the assumption that the means of message transmission has an impact on the text's language. We are therefore interested in how much language contributes to accomplishing the intended social action (language role), how the text is transmitted (channel) and finally, whether it is produced in written or spoken mode (medium), since this influences the organisation of the text.

4.3.1 Language role

We can distinguish different situations where we rely to a greater or lesser degree on language to achieve our goal. Extreme examples would be a case where a nod may be sufficient to convey the intended meaning or a comic strip where the story may be realised almost completely in pictures with language only coming into play at some focal point like the punch line. At the other end of the continuum, social action may be realised entirely linguistically without any kind of material action supporting the verbal action. Hasan (1999: 281f) argues quite

stringently that this subdimension belongs to field of discourse. However, her main point is that 'the so-called rhetorical modes such as explaining, defining, generalising, reporting, recounting, narrating, chronicling etc. are best viewed as constitutive verbal actions' and thus as specifying the nature of the social activity which is accordingly analysed under field of discourse. In our account these modes are treated as *goals* of verbal action and analysed as the subdimension 'goal orientation' (see section 4.1.2). From the point of view of written text – the view taken up by this study – language role as analysed under mode of discourse covers a different aspect of verbal action, one that belongs more clearly to the study of the textual make-up of the register. We are concerned with the interaction between verbal, i.e. linguistic, parts of a text and other semiotic modes like photos, figures, graphs etc. The presence or absence of these modes should have an impact on the language used in the register. If, for instance in instruction manuals (the INSTR register in the CroCo project), the meaning is to a greater extent realised in the form of graphical presentations of the device described, the verbal parts of the text may not explicitly describe the device but rather use endophoric reference in the form of deictics, pronouns etc. to complement the figures (cf. Bartsch's (2007) discussion of cohesion between the different modalities in texts from the field of mechanical engineering).

Usually, two options are named for language role, namely ancillary and constitutive. It seems plausible to view the options as two ends of a cline, but since this is a matter of interpreting the findings, it is not of great importance here. Observable indicators for the role of language are ellipsis, mood, theme and reference.

4.3.2 Channel

This subdimension is concerned with the physical conditions of the communication. It is relevant to the study of registers because different channels offer and constrain choices in meanings and their realisation in different ways (cf. Steiner 2004). The phonic channel, i.e. transmission via sound waves, requires different linguistic expressions from a graphic setting. For instance, material action is of no consequence if the interaction is transmitted via paper. In most cases, electronic environments share the characteristics of the graphic channel, being only graphic texts provided in electronic form. There are, however, interactions like chat room conversations with their real time transmission of written turns that are probably unique to the electronic channel, thus making a third option, 'electronic', necessary.

While we can name criteria for assigning texts to one of the three options, in most cases the assignment should be possible on mere inspection of the texts without needing to go into a detailed analysis of the text. The texts analysed in this paper are all transmitted via the graphic channel.

4.3.3 Medium

As Halliday and Hasan (1989: 58) convincingly point out, the subdimension channel does not cover the spoken-written distinction, which has to be investigated separately under the heading 'medium' because transmission and production are different – though related – aspects. This becomes obvious when we look at cross-classifications: a text may be transmitted through the phonic channel but still present more characteristics of the written medium. Although transmitted by sound waves, the text may thus be produced *as if* written. SPEECH, the register in the CroCo Corpus representing prepared speeches, is a case in point. The speeches in this register are written to be spoken. They are originally transmitted in the phonic channel but are previously prepared in writing and thus bear characteristics of the written medium, reflected for instance in a rather high lexical density (cf. Vela et al., 2007). Other indicators for medium besides lexical density are thematic structure, reference and certain types of clause complexity.

Thematic structure can be analysed by retrieving the occurrences of different grammatical functions in sentence-initial position. Our expectations with respect to this category go along the following lines: The word order characteristics of the two languages suggest that the initial position should show more variation in the German registers. Apart from this potential contrastive difference, we should also be able to detect registerial differences. Possibly, the fictional texts stretch the grammatical options and constraints more than economic texts with their focus on conveying factual information. The frequency of the lexical item *year/jahr* (see section 4.1.1) suggests that there should also be variation in the area of adverbials: the SHARE texts report on the last financial period and could therefore contain a certain amount of temporal adverbials in theme position, setting the scene for the reported aspect of the company's performance.

The expectations discussed so far only address general aspects of textual variation. This is broadly related to mode of discourse. Adding the intermediate subdimension of medium specifies the general registerial consideration in terms of the spoken-written distinction. In this narrower sense we would expect the FICTION texts – particularly in the direct speech passages – to contain more finites, conjunctions and certain types of adverbials in theme position pointing to a more interpersonal orientation that can be assumed to occur in spoken discourse.

Reference is of particular interest concerning the subdimension medium, since spoken registers are said to rely more on pronominal reference than written registers. Biber (1988: 225f) gives an overview of studies discussing the role of various pronouns in spoken and written registers. Particularly the distinction between exophoric and endophoric reference (Halliday and Hasan, 1976) should shed light on the spoken-written distinction. We can also assume that – beyond this – syn-semantic, i.e. pronominal, versus full lexical reference is an indicator for this distinction with pronominal reference pointing to spoken, situation-dependent registers and lexical reference being more likely to feature in factual, written registers (cf. Hansen-Schirra et al., 2007). The overall number of pronouns should therefore be higher in spoken registers.

The two registers under investigation here are both transmitted through the graphic channel and can thus be expected to exhibit more characteristics of written language. We can, however, assume that fictional texts make rather extensive use of direct speech, which is not to be expected in the SHARE texts. Although it presumably belongs to the written medium, FICTION should therefore contain noticeably more pronouns than SHARE.

The present study concentrates on personal pronouns (thus excluding demonstratives and indefinite pronouns). The query is based on part-of-speech tagging and includes tags for possessive personal pronouns. The English fictional texts contain 2,477.34 personal pronouns and the German 2,297.46. SHARE contains 941.48 in English and 1,009.98 in German (all values expressed as normalised frequencies[21]). The comparison of the overall frequencies of personal pronouns reveals a distinct difference between the two registers.

	EO		GO	
	FICTION	SHARE	FICTION	SHARE
1st person sing	25.59	6.87	31.29	3.92
3rd person sing masc	18.18	1.98	25.49	2.35
3rd person sing fem	13.90	0.10	10.15	2.55
3rd person sing neutr	15.50	16.13	13.70	10.59
1st person plur	5.49	61.19	6.87	68.43
3rd person plur	13.70	10.61	6.03	4.61
2nd person	7.65	3.12	6.47	7.55

Table 5. Distribution of personal pronouns in per cent

Table 5 displays the distribution of the different types of personal pronouns in the two registers under investigation. From this table we can gather interesting differences in the distribution of forms of the personal pronouns in the two registers, with contrastive differences present but rather gradual than categorical. Among these differences are the following: First person singular pronouns seem to be typical of FICTION texts (cf. example 6), while first person plural pronouns are very frequent in SHARE. This is probably owing to the fact that the management of the respective company reports on behalf of the company (cf. example 7).

(6) If **I** were to attempt this description - but no, **I** cannot - yet **I** must, for **I** am your chronicler, bound to recount to you, what? (EO__FICTION)

(7) But in the areas that **we** can control and influence, **we** believe **we** are well positioned for greater future success. (EO__SHARE)

4.4 An outlook on register variation

This paper only discussed a selection of indicators for the different subdimensions for two registers in the languages English and German. It allows the following preliminary assumptions about register variation within a given language and across two languages.

Virtually all indicators discussed in this paper exhibit variation in the intralingual comparison of the two registers. The two registers FICTION and SHARE are thus anchored in quite different situational contexts. More insight on register variation in the language pair English-German is to be expected from the inclusion of other registers that may be more closely related.

The two registers compared here show some, but not highly marked contrastive differences with the differences seeming to be gradual rather than categorical. An analysis of all subdimensions is required as well as a validation against a basis of comparison (see Neumann, 2008).

5. Conclusions and future work

The features analysed in the present study belong to some kind of intermediate level of abstraction. They are not as specific and interpretative as those features typically used in manual, example-based analyses, but not as shallow as features often used in quantitative studies on the basis of raw, i.e. un-annotated texts.

The present study highlights the usefulness of a sound theoretical foundation in register studies. While very similar to Douglas Biber's (1988, 1995) work on register variation, it extends the scope of register description beyond features relevant for the spoken-written distinction. Biber's use of data reduction techniques is replaced by more human interpretation relating the quantitative empirical findings to the abstract concepts expressed by the subdimensions of the register parameters. The methodology is apt to identify differences and commonalities between registers and allows cross-linguistic comparisons because the more abstract functional concepts in the theoretical framework of register linguistics are valid interlingually.

The theoretical framework yields comprehensive findings in that it permits the combination of lexico-grammatical analyses with statements on higher level units like texts and registers.

Space did not allow the discussion of all subdimensions and all linguistic indicators relevant to register analysis. This will have to be shown in future work. Future work will also have to encompass a wider variation of registers as well as translations that may prove to represent distinct registers differing from originals in both languages under investigation. Finally, statistical procedures have to be applied to the corpus in order to treat outliers as well as to test the significance of the quantitative findings. It remains to be seen whether data reduction procedures more appropriate than the ones used by Biber can be applied to the data.

Acknowledgements

The present research was mainly undertaken in the framework of the project "CroCo – Linguistic Properties of Translations" sponsored by the German Research Foundation as project no. STE 840/5-2, particularly during a research visit at Macquarie University, Sydney, Australia. It greatly profited from continuous discussions with people at Macquarie University as well as Erich Steiner and Silvia Hansen-Schirra. I am indebted to Mihaela Vela for her help with posing the queries. I would like to thank Mary Mondt for her thorough and thoughtful proofreading as well as two anonymous reviewers for their valuable comments.

NOTES

1 http://fr46.uni-saarland.de/croco

2 "The term *oral* refers to stereotypically spoken discourse – that is, conversation – while the term *literate* refers to stereotypically written discourse – that is, informational exposition." (Biber 1995:238)

3 Butler (1985:12f) points out that linguists commonly assume a higher level of measurement: "The reason for this is that the parametric tests suitable for interval data are more powerful than non-parametric tests. As often happens in the application of statistical methods to real problems, practical considerations frequently outweigh the concerns of the theoretical purist."

4 See Neumann (2003) for a discussion of the advantages of founding contrastive comparisons on the systemic functional approach

5 Cf. postings on the sysfling mailing list (http://listserv.uam.es/archives/sysfling-l.html) in July 2007.

6 http://www.tei-c.org/

7 At the time of writing this paper the annotation was not yet available for the reference corpora.

8 In the near future, the CroCo Corpus will be annotated with sense relations according to WordNet and GermaNet. This annotation will allow retrieving all sense relations of a frequent lexical item in a register.

9 We keep small case for all German tokens as assigned by the automatic lemmatiser.

10 Since the present study does not analyse whole clauses in their textual surroundings, we can only conjecture that the verbs discussed here represent a given process type on the basis of their lexical meaning. We do not claim that the process types mentioned hereafter do actually obtain for the lexical verbs discussed.

11 Lexical spread is measured here as the number of different lexical types occurring in a register. In normalised frequencies, FICTION contains 3,810.38 (English) and 4,231.47 (German) different lexical types, while SHARE only contains 2,773.11 (English) and 3,805.70 (German).

12 http://www.wagsoft.com/CorpusTool/index.html

13 Since this analysis takes into account the whole clause, albeit only for this form of the verbs *be* and *sein*, it is admissible to determine the process type.

14 In terms of cohesion the subject matter of the text is probably maintained by other cohesive devices, particularly reference.

15 Since all texts in the corpus are published in writing, we use the terms 'writer' and 'reader' instead of 'speaker' and 'addressee'.

16 He or she may react to the writer, but this would be considered a different text (cf. Hasan 1999 for a discussion of related texts).

17 An additional problem for the analyst may arise from the fact that there may be little or no information on the writer aside from his/her output, the text under investigation.

18 This part of the analysis is often subsumed under the heading "agentive roles". However, expert knowledge does not necessarily give insight into who is the agent in the context of situation, but rather into who bears which role in society and is therefore analysed as part of the subdimension "social hierarchy".

19 There seems to have been a shift towards less formal interaction since Joos came up with his categorisation in the early 1960s. This became obvious from inspection of the texts in the corpus and resulted in a shift towards more informal categories.

20 http://grammatics.com/appraisal/AppraisalOutline/UnFramed/AppraisalOutline.htm, last visited: 7.10.2007

21 The normalised corpus size is 31,250 tokens per register.

REFERENCES

Baroni, M., Evert, S. (2009), "Statistical methods for corpus exploitation", in Lüdeling A. , Kytö M. (eds.) *Corpus Linguistics. An International Handbook.* Berlin: Mouton de Gruyter, 777-802.

Bartsch, S. (2007), "Intermodal Relations in Multimodal Text". Paper Presented at *Corpus Linguistics 2007*, Birmingham, UK, 27 – 30 July 2007.

Bednarek, M. (2007) "Local grammar and register variation: explorations in broadsheet and tabloid newspaper discourse". *ELR Journal* 1 (1). Accessible at http://ejournals.org.uk/ELR/article/2007/1.

Biber, D. (1988), *Variation across Speech and Writing.* Cambridge: Cambridge Univ. Press.

Biber, D. (1990), "Methodological issues regarding corpus-based analyses of linguistic variation", in *Literary and Linguistic Computing* 5/3, 257-269.

Biber, D. (1993), "Representativeness in corpus design", in *Literary and Linguistic Computing* 8/4, 243-257.

Biber, D. (1995), *Dimensions of Register Variation.* Cambridge: Cambridge Univ. Press.

Bublitz, W. (1978), *Ausdrucksweisen der Sprechereinstellung im Deutschen und Englischen.* Tübingen: Niemeyer.

Butler, C. (1985), *Statistics in Linguistics.* Oxford: Blackwell. Web edition accessible at http://www.uwe.ac.uk/hlss/llas/statistics-in-linguistics/bkindex.shtml.

Butt, D. (2004), *Parameters of Context: On Establishing the Similarities and Differences between Social Processes.* Manuscript. North Ryde: Centre for Language in Social Life, Macquarie University.

Ghadessy, M. (ed.), (1988), *Registers of Written English.* London, New York: Pinter.

Ghadessy, M. (ed.) (1993), *Register Analysis. Theory and Practise.* London, New York: Pinter.

Greenbaum, S., Svartvik J. (1990), "The London-Lund Corpus of Spoken English", in Svartvik J. (ed.), *The London-Lund Corpus of Spoken English: Description and Research.* Lund: Lund University Press.

Halliday, M.A.K. (1978), *Language as Social Semiotic. The Social Interpretation of Language and Meaning.* London: Arnold.

Halliday, M.A.K., McIntosh A., Strevens P. (1964), *The Linguistic Sciences and Language Teaching.* London: Longman.

Halliday, M.A.K., Hasan R. (1976), *Cohesion in English.* London: Longman.

Halliday, M.A.K., Hasan R. (1989), *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective.* Oxford: Oxford Univ. Press.

Halliday, M.A.K., Martin, J.R. (1993), *Writing Science: Literacy and Discursive Power.* London, Washington, D.C.: Falmer Press.

Halliday, M.A.K., Matthiessen, C.M.I.M. (2004), *An Introduction to Functional Grammar.* London: Arnold. (earlier versions by Halliday 1985, 1994).

Hansen-Schirra, S., Neumann S., Steiner E. (2007), "Cohesive explicitness and explicitation in an English-German translation corpus", in *Languages in Contrast* 7:2 (2007), 241-265.

Hasan, R. (1999), "Speaking with reference to Context". Ghadessy, M.(ed.), *Text and Context in Functional Linguistics.* Amsterdam, Philadelphia: Benjamins. 219-328.

House, J. (1997), *Translation Quality Assessment. A Model Revisited.* Tübingen: Narr.

Johansson, S., Leech G. N., Goodluck H. (1978), *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers.* Oslo: Department of English, University of Oslo. Accessible at http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM

Klinger, A., Vela M., Hansen-Schirra S. (2006), *Kodierung von Metainformation.* Deliverable no. 2 of the CroCo Project. Accessible

at http://fr46.uni-saarland.de/croco/corpus_meta.pdf.

Lavid, J. (1994), *Thematic Development in Texts*. Technical Report. Saarbrücken: Universität des Saarlandes. (ESPRIT Basic Research Action: Dandelion, EP6665; Deliverable R1.2.1). Accessible at http://www.uni-saarland.de/fak4/fr46/steiner/dandelion/r121-1.ps.

Lee, D. (2009), *Modelling Variation in Spoken and Written English. The Multi-Dimensional Approach Revisited*. London: Routledge.

Martin, J. R. (1992), *English Text. System and Structure*. Amsterdam, Philadelphia: Benjamins.

Martin, J. R., White P. R.R. (2005), *The Language of Evaluation. Appraisal in English*. Basingstoke: Palgrave Macmillan.

Matthiessen, C. M.I.M. (1993), "Register in the round: diversity in a unified theory of register analysis". Ghadessy, M. (ed.), in *Register Analysis. Theory and Practice*. London: Pinter. 221-292.

Matthiessen, C.M.I.M. (2001), "The environments of translation". Steiner E., Yallop C. (eds.), *Exploring Translation and Multilingual Text Production: Beyond Content*. Berlin, New York: Mouton de Gruyter. 41-124.

Morris, J., Hirst, G. (1991), "Lexical cohesion computed by thesaural relations as an indicator of the structure of text". *Computational Linguistics* 17:1, 21-48.

Neumann, S. (2003), *Textsorten und Übersetzen. Eine Korpusanalyse englischer und deutscher Reiseführer*. Frankfurt/M.: Peter Lang.

Neumann, S. (2008), *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Unpublished habilitation thesis. Saarbrücken: Universität des Saarlandes.

Neumann, S., Hansen-Schirra S. (2005), "The CroCo Project: Cross-linguistic corpora for the investigation of explicitation in translations". *Proceedings from the Corpus Linguistics Conference Series*, Vol. 1, no. 1, ISSN 1747-9398.

Scott, M.(2004), *WordSmith Tools*. Oxford: Oxford Univ. Press.

Steiner, E. (1998), "How much variation can a text tolerate before it becomes a different text? An exercise in making meaningful choices", in Schulze, R. (ed.), *Making Meaningful Choices in English*. Tübingen: Narr. 235-257.

Steiner, E. (2004), *Translated Texts. Properties, Variants, Evaluations*. Frankfurt/M.: Peter Lang.

Steiner, E. (2005), "Explicitation, its lexicogrammatical realization, and its determining (independent) variables – towards an empirical and corpus-based methodology", in *SPRIKreport* No. 36, December 2005. Accessible at http://www.hf.uio.no/forskningsprosjekter/sprik/docs/pdf/Report_36_ESteiner.pdf.

Teich, E., Fankhauser P. (2004), "Exploring lexical patterns in text: lexical cohesion analysis with WordNet", in *Interdisciplinary Studies on Information Structure* 2 (2005):129-145.

Thompson, G., Hunston S.(eds.) (2003), *Evaluation in Text. Authorial Stance and the Construction of Discourse*. Oxford: Oxford Univ. Press.

Vela, M., Neumann S., Hansen-Schirra S., (2007), "Querying multi-layer annotation and alignment in translation corpora". *Proceedings of the Corpus Linguistics Conference CL 2007*. Birmingham, UK, 27-30 July 2007. Article #97.

Ventola, E. (1996), "Packing and unpacking of information in academic texts". Ventola, E., Mauranen, A. (eds.), *Academic Writing. Intercultural and Textual Issues*. Amsterdam, Philadelphia: Benjamins, 153-194.

Vermunt, J. K., Magidson, J. (2005), "Factor analysis with categorical indicators: a comparison between traditional and latent class approaches", in van der Ark, L.A., Croon, M.A., Sijtsma, K. (eds.), *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences*. Erlbaum, 41-62.

Werlich, E. (1976), *A Text Grammar of English*. Heidelberg: Quelle & Meyer.