# Cross-linguistic analysis of discourse variation across registers

*Kerstin Kunz, Heidelberg University and Ekaterina Lapshinova-Koltunski, Saarland University*

Abstract

The present study deals with variation in discourse relations in different registers of English and German. Our previous analyses have been concerned with the systemic contrasts between English and German, cf. Kunz & Steiner (2013 a/b), Kunz & Lapshinova (to appear) and have addressed some cross-linguistic differences with regard to textual realizations of selected subtypes of cohesion. In our current work, our focus is on the empirical analysis of cross-linguistic variation between registers. In order to obtain a more comprehensive picture, we investigate three main types of cohesion in combination: co-reference, substitution and conjunction and their subtypes, cf. Halliday & Hasan (1976). We extract instantiations of cohesive devices from an English-German corpus of spoken and written registers. The data is analyzed with statistical procedures which show that subcorpora can be grouped along particular combinations of cohesive devices.

## *1. Research objective*

Our aim in this study is to interpret corpus data about three types of cohesion—co-reference, substitution and conjunction—as well as their subtypes. Recent years have seen an increase in the number of works employing corpus-based methods for the study of cohesion. However, multilingual works are mostly concerned with individual cohesive devices in particular registers, e.g. Neumann et al for repetition, Zinsmeister et al. (2012) for abstract anaphors, Bührig & House (2004) for particular cohesive conjunctions or adverbs, and Taboada & Gómez-González (2012) for particular coherence relations. Most studies that analyse particular types of cohesion work with one language only. For instance, corpus-based works concerned with conjunction are Stede (2008), Dipper & Stede (2006), Bestgen et al. (2006) and works on co-reference are Eckert & Strube (2000), Gundel et al. (2004). Works on substitution in the sense of Halliday & Hasan (1976) are rare (see Kunz & Steiner 2013b), also because it seems to be a less frequent cohesive phenomenon.

The main objective of our previous studies was to identify German-English contrasts in the realization of particular types and subtypes of

Kunz, Kerstin and Ekaterina Lapshinova-Koltunski. 2015. "Cross-linguistic analysis of discourse variation across registers." *Nordic Journal of English Studies* 14(1):258-288.

cohesion both from a systemic and textual perspective, c.f. Kunz and Steiner (2013a) for co-reference, Kunz and Steiner (2013b) for substitution and Lapshinova & Kunz (2014) and Kunz & Lapshinova (to appear), for conjunctive relations. These types will now be examined in combination in order to obtain a more comprehensive expression of textual features. The focus of this study is on the analysis of variation in the registers that were collected in the GECCo corpus - a bilingual corpus of English and German that comprises 12 different registers of English and German. Although our current interest lies in the cohesive devices which serve as explicit indicators of textual relations across grammatical domains, we also consider the ties (and chains in case of co-reference) established by these devices. There are several corpora which have already been annotated on the level of discourse. cf. Doddington et al. (2004) and Pradhan et al. (2011), the OntoNotes Corpus for English, Arabic and Chinese, (Weischedel et al. 2013) and the TüBa-D/Z corpus (Hinrichs et al. 2005) or the Prague dependency Treebank, and the Penn Discourse Treebank for English, cf. Prasad et al. (2008). These corpora cannot be employed for our research on English and German as (1) they do not contain comparable registers across languages; (2) most of them do not contain annotations of different types of cohesion; and (3) they do not provide enough register variation to permit analyzing register as a variable. To our knowledge, our corpus is the only existing resource that allows for an investigation of different cohesive phenomena cross-linguistically and across different registers at the same time.

For this study, only those ten registers from the GECCo corpus were selected for analysis where the annotation obtained from semi-automatic annotation procedures has already undergone in-depth correction phases by human annotators. The data obtained has been evaluated by a combination of different statistical evaluation procedures. This methodology permits comparison of distinctive main types and subtypes of cohesive devices and clustering of registers according to particular types. It thus facilitates a sound interpretation that goes beyond the level of grammar towards more abstract conceptual ranks.

Most importantly, it permits us to address the research objectives pursued in the frame of the current study. The main question we are interested in is whether contrasts are more pronounced between different registers independent of language or whether more differences are identified in one and the same register between English and German.

Targeting this question, several others arise. For instance, we intend to identify which registers show most similarities/ differences, within one language and across languages. Next, we want to identify those registers that are most pronounced in the realization of particular features. And finally, we are also interested in the features that contribute to the observed differences/ commonalities.

Before we deal with these questions, we will provide a brief definition of cohesion which is followed by a conceptual clarification of the three cohesive types under investigation. We will describe in short our corpus resource and the procedures employed for annotating co-reference, substitution and conjunction. The centerpiece of our study will be the evaluation of the obtained data via various statistical methods such as descriptive analysis and correspondence analysis and their interpretation in terms of the questions highlighted above.

## *2. Cohesion and cohesive subtypes*

Let us now move on to a short discussion of the concept of cohesion and the subtypes under investigation. Note that details on systemic differences between English and German can be found in Kunz & Steiner (2013a) for co-reference, Kunz & Steiner (2013b) for substitution and Kunz & Lapshinova (to appear) for conjunction.

Language producers employ particular lexicogrammatical items (**cohesive devices**) which indicate a linguistic relation to other textual elements across grammatical domains. The explicit linguistic ties or chains which are created by language producer on the text surface (**cohesive relations**) help recipients in their cognitive interpretation as to how different thematic concepts are connected. Cohesion can be signaled in texts by grammatical items such as personal and demonstrative pronouns and modifiers, substitute forms, elliptical constructions and conjunctions or by lexical devices such as verbs, nouns and adjectives. These cohesive devices trigger different semantic relationships, whose borderlines however are often blurred.

As mentioned above, the focus of this study is on three main types of cohesion (see further Halliday & Hasan 1976, Halliday & Matthiessen 2013): co-reference, substitution and conjunction and on their subtypes. Therefore, we will now examine their peculiarities in more detail.

Let us start by looking at the features shared by all three types under investigation. What co-reference, substitution and conjunction have in common is that there are explicit linguistic devices signalling particular conceptual relations to linguistic elements in other clauses, sentences or paragraphs of the same text/ discourse, and that the interpretation of the cohesive devices is dependent on the elements they tie up with (see Halliday & Hasan 1976). All three types are regarded in the literature as being grammar-driven (see e.g. Louwerse and Graesser 2005, Brinker 2005, Schubert 2008) since the devices that trigger the cohesive relations belong to a closed class of functional items, in contrast to devices of lexical cohesion, which comprise open classes of nouns, verbs, adjectives and adverbs. Grammar-driven items are quite often semantically weak (see Halliday & Hasan 1976) and it is this semantic reduction which initiates a search for other linguistic elements in the text on the basis of which the intended meaning can be fully interpreted. For an illustration, consider (1) to (3) below, where cohesive devices are marked in bold and antecedents/ elements connected by a cohesive device are in square brackets.

(1) *Do not turn on [the computer]. Turning* ***it*** *on before you're finished assembling the system could …*

(2) *[We have to be honest about the challenges facing Europe].* ***And*** *[we have to listen to what Europe's voters are telling us].*

(3) *'I don't have a [car], 'he said. 'If I borrowed* ***one****, would you ...?'*

In (1) the referential device employed is the neuter personal pronoun *it*. *It* refers to an entity whose semantic class (e.g. computer vs. printer) can only be identified by looking at the antecedent *the computer*. The cohesive item *and* in (2) belongs to the closed class of coordinators and only signals that there is an additive relation between two entities without but does not provide any information as to which kind of entities are involved. In (3) the substitutional form *one* is an indefinite pronoun that indicates similarity with or selection of a referent out of a particular class but again, does not give a clue as to the intended class or referents.

Although many devices of co-reference, substitution and conjunction conform to the above descriptions, there also items which are semantically richer and therefore less grammatical. This is particularly the case with comparative co-reference and also various conjunctive

adverbials, which are on the borderline between grammatical and lexical cohesion.

Co-reference, substitution and conjunction however differ in their lexico-grammatical realizations and, most importantly, in the type of semantic relation they preferentially express. These differences are discussed in the following.

### *2.1 Co-reference*

The conceptual relation created by **co-reference** is one of identity. Co-referential devices signal that they point to a referent which has already been mentioned by a referring expression in another (mostly) preceding textual part (antecedent), hence to a referent that is textually old or given (cf. Prince 1981, Gundel et al. 1993, Ariel 1990). The referential devices can therefore be considered as search instructions to other textual elements on the basis of which the intended referent is cognitively identified (cf. Schwarz 2000:43). Following the classifications by Halliday & Hasan (1976), we distinguish three subtypes of co-reference:

- **personal**: relations triggered by personal pronouns (e.g. *it*/ *es*, *they*/ *sie*), possessive pronouns and modifiers (e.g. *his*/ *sein(e,r,s)*)
- **demonstrative**: triggered by definite article, demonstrative pronouns (e.g. *this, that/ dies, das*) and modifiers (e.g. *this/ diese (r,s)*), local and temporal adverbs (e.g. *here, then/ hier, da*) and pronominal adverbs (e.g. *herewith/ hiermit)*
- **comparative**: triggered adjectives and adverbs of comparison (such as *similar/ ähnlich* or *such*/ *solche*[1]

It has to be noted that the category of comparative reference is semantically distinct from the other two main types as it does not create identity of reference but rather evokes a relation of similarity and comparison between referents, events or propositions of the same type (see e.g. Halliday & Matthiessen 2004:560, Schubert 2008:35). For instance in (4), *another* in combination with *explanation* - device of

[1] See Kunz & Steiner (2013) for a more detailed discussion of semantic and functional differences.

lexical cohesion - ties with the preceding sentence but does not create identity.

(4) *It is said that the French soldiers saw the Welsh women from a distance in their tall hats, thought they were soldiers and surrendered! There may be* ***another*** *explanation but …*

Devices of comparative reference and also of demonstrative reference are combined with devices of lexical cohesion in case the referential device is a modifier (e.g. *these explanations*) or a comparative adjective (e.g. *another explanation*).

*2.2 Substitution*

The main difference between co-reference and **substitution** concerns the semantic type of relation: in contrast to co-reference the tie between the cohesive device and its antecedent does not trigger identity between instantiated referents but similarity between referents belonging to the same class (cf. Kunz & Steiner 2013b, de Beaugrande & Dressler 1981). It therefore exhibits some similarities with comparative reference in the semantic meaning relations established. Substitution can additionally be differentiated from co-reference because of its syntactic constraints, in which it resembles cohesive ellipsis. The formal options available in English (and also in German) for establishing substitution are very limited. We analyze three main subtypes of substitution:

- **nominal:** e.g. signaled by the same and one(s)/ Das Gleiche/ dasselbe, eine(r,s) in German
- **verbal:** by *do (so)/ tun* and *machen* in German
- **clausal:** mainly with the form *so* in English, more variation in German

*2.3 Conjunction*

The semantic relation of **conjunction** differs from co-reference and substitution in that conjunctive devices do not refer themselves and therefore do not have an antecedent. They indicate relations between two other textual elements and explicitate logico-semantic relations between referents, which are semantically rather complex such as states,

processes and events (cf. Pasch et al. 2003, Blühdorn 2008). Note that the term 'conjunction' used in this study deviates in its conceptualization from most grammars. We depart from the meaning relation established, in the sense of Halliday and Hasan (1976) and Halliday and (Matthiessen 2013: 593ff), who include all forms that signal a cohesive relation between linguistic elements. These forms are termed "conjunctive device" in our work, while "conjunction" refers to the relation as such.
The relations that are explicitated by conjunctive devices can be mainly grouped into the following categories (cf. Halliday & Hasan 1976):

- **additive**: relation of addition, for two events that are true/ not true at the same time (conjunctive devices indicating such a relation are e.g. *and, in addition, und, außerdem*)
- **adversative**: relation of contrast/ alternative, for two events which are not true at the same time (*yet, although, by contrast, doch, obwohl, im Gegensatz dazu*)
- **causal**: relation of causality/ dependence between (because, therefore, that's why, weil, deshalb, aus diesem Grund)
- **temporal**: temporal relation between events (after, afterwards, at the same time, nachdem, danach, gleichzeitig)
- **modal**: This latter category subsumes devices that are not included in most grammars. The meaning rather is an interpersonal or pragmatic one (see e.g. Martin 1992: 178ff) in which conjunctive devices connect events by an evaluation of the speaker (*well, sure, klar, sicher*). In the literature, they often fall under the category of 'discourse markers' and are called 'continuatives' by Halliday & Hasan (1976) and Halliday & Matthiessen (2013).

Apart from these semantic peculiarities conjunctive devices exhibit distinctive lexicogrammatical features. There is more variation with respect to different forms available as well as the number of elements contained in the conjunctive device. This particularly the case with conjunctive adverbials, which may consist of one adverb only, e.g. *therefore*, or may be multiword constructions, e.g. *for this reason* or *that's why*, which may contain lexical as well as referential items. For this reason, there is also more variation in terms of the semantic explicitness of the devices. However, the set of devices linking main

clauses is relatively small. In our analysis we distinguish the two main structural types of conjunctive devices mentioned above:

- **coordinators**: link textual elements in a paratactic construction (e.g. *and, but, neither … nor,und, aber, weder …noch etc*)
- **adverbials**: link clause complexes (sentences) or even elements on a higher textual level (e.g. *therefore*, *by contrast*, *deshalb*, *im Gegensatz dazu*, etc.)

Subordinators, which link main and subordinating clauses, are not included in the present study since they are generally not regarded in the literature as establishing relations of cohesion.

Conjunctive devices also exhibit more restrictions in their syntactic function and position: coordinators do not serve as fully-fledged syntactic constituents, and conjunctive adverbials only take on the function of a syntactic adverbial. The most common position of conjunctive devices is between the first and the second textual element they link, although there is more variation for conjunctive adverbials; see Kunz & Lapshinova (to appear).

If we look at instantiations of co-reference, substitution and conjunction, we sometimes notice that the borderlines between the three cohesive types are blurred to the extent that one and the same cohesive form may serve as referential, substitutional or conjunctive device, dependent on the context in which it is realized.

(5) *And we also we have the Dee river on one side of the peninsula and the Mersey on the other. - **So** the peninsula is between the two rivers. - Yes, yes. Right. (causal conjunction)*

(6) *It's a financially driven issue that local authorities who are responsible for providing care will **do so** of course in the least cost way possible. (clausal substitution)*

(7) ***So** pflegen Sie Ihr Gerät … : Es genügt, wenn Sie das Gerät nur feucht abwischen. (demonstrative reference or clausal substitution)*
***So** ('in this way') you take care of your instrument… It suffices when you wipe the instrument damp. (literal translation)*

As examples (5) to (7) illustrate, this is especially the case with the form *so* in English and German, and also with pronominal adverbs in German, which may either serve as devices of co-reference or as devices of conjunction.

## *3. Data and methods*

In the following section, we will first provide information about the GECCo corpus and about the procedures to extract the different types of cohesive devices before we will shortly describe the statistical methods applied for data evaluation.

### *3.1 Corpus resources*

As mentioned in section 2 above, our research is based on data extracted from the GECCo corpus (cf. Kunz and Lapshinova-Koltunski, 2011 and Lapshinova et al., 2012), a German - English corpus of different written and spoken registers. The corpus contains ca. 1.3m tokens. For this particular study, we analyze four subcorpora only: German written originals (GO), English written originals (EO), English spoken originals (EO-SPOKEN) and German spoken originals (GO-SPOKEN). The two written subcorpora consist of texts from eight registers: popular-scientific texts (POPSCI), tourism leaflets (TOU), prepared speeches (SPEECH), political essays (ESSAYS), fictional texts (FICTION), corporate communication (SHARE), instruction manuals (INSTR) and corporate websites (WEB). The two spoken subcorpora contain academic speeches (ACADEMIC) and interviews (INTERVIEW). The corpus also contains two further subcorpora: English and German translations of EO and GO. However, in this study, we analyze cohesive phenomena in subcorpora of originals only, which provide a good database for our English-German contrastive analysis. The subcorpora, from which we extract frequency information on the occurrence of cohesive phenomena, are presented in Table 1.

The corpus is annotated on several levels, and annotations include information on tokens, lemmas, morpho-syntactic features (e.g. case, number, etc.), parts-of-speech, phrase chunks and their grammatical functions, as well as and sentence boundaries. The annotation of the written part was partly imported from CroCo, whereas for the spoken

part, we use Stanford POS Tagger (Toutanova et al., 2003) and the Stanford Parser (Klein and Manning, 2003). The corpus is encoded in the CWB format (CWB, 2010) and can be queried with Corpus Query Processor (CQP) (Evert, 2005). These annotation levels provide us with additional information on cohesive phenomena and cohesive relations, i.e. for co-reference: morpho-syntactic preferences of referring expressions, such as their positions in a clause, etc.

Table 1. Variables for corpus-based analysis

| EO | GO |
|---|---|
| EO_ACADEMIC | GO_ACADEMIC |
| EO_FICTION | GO_FICTION |
| EO_ESSAY | GO_ESSAY |
| EO_INSTR | GO_INSTR |
| EO_INTERVIEW | GO_INTERVIEW |
| EO_POPSCI | GO_POPSCI |
| EO_SHARE | GO_SHARE |
| EO_SPEECH | GO_SPEECH |
| EO_TOU | GO_TOU |
| EO_WEB | GO_WEB |

Moreover, the corpus is annotated with information on cohesive devices. For this, semi-automatic procedures were applied, which include a rule-based tagging of cohesion candidates and their manual post-correction by humans. A description of the procedures is given in Lapshinova and Kunz (2014). For this study, we deploy the annotation of cohesive devices establishing co-reference, substitution and conjunction, whose subtypes and functions as annotated in the corpus are given in Table 2. These subtypes serve as categories for our corpus-based analysis described in section 5 below.

Table 2. Categories of phenomena under analysis

| device | co-reference | conjunction | substitution |
|---|---|---|---|
| category | personal head,<br>personal modifier<br>demonstrative head<br>demonstrative modifier<br>demonstrative local<br>demonstrative temporal<br>pronominal adverbs<br>definite articles<br>comparative general<br>comparative particular | additive connectives<br>additive adverbials<br>adversative connectives<br>adversative adverbials<br>causal connectives<br>causal adverbials<br>temporal adverbials<br>modal adverbials | nominal<br>verbal<br>clausal |

*3.2 Data extraction*

The instantiations of the categories presented in Table 2, can be easily extracted from the corpus, as their information is annotated and can be queried with CQP. In Table 3, we provide examples of queries used for the data extraction.

Table 3. Query examples used to extract the categories under analysis

| | Query | Explanation |
|---|---|---|
| 1 | [_.reference_func="poss.*"] | personal reference with a modifying function (pers_mod) |
| 2 | [_.reference_func="temporal"] | temporal demonstrative reference (dem_temporal) |
| 3 | [_.conj_func="additive" &_.conj_type="connect"] | additive coordinating conjunctions (additive_connect) |
| 4 | [_.conj_func="additive" &_.conj_type="adverbial" | additive adverbials (additive_adverbial) |
| 5 | [_.substitution_type="verbal"] | all cases of verbal substitution |
| 6 | [_.substitution_type ="clausal"] | all cases of clausal substitution |

For instance, with the help of query 1, we can identify how many referential devices function as personal modifiers (possessive determiners), whereas query 2 is used to identify all cases of referential devices with a temporal function (temporal adverbs). Queries 3 and 4 are built to differentiate between coordinating conjunctions expressing additive relations and conjunctive adverbials expressing additive relations. We apply queries like those in 5 and 6 to extract different types of substitution and ellipsis.

### *3.3 Statistical methods*

Descriptive data analysis is employed to compare frequency distributions of main cohesive types and of cohesive subtypes. This allows us to obtain a first insight into differences and commonalities between registers within one language and across languages with respect to preferences of particular types of cohesion.

To validate our data statistically, we use an unsupervised technique – correspondence analysis (CA), cf. Baayen (2008). This statistical procedure allows us, on the one hand, to see which registers have more commonalities and which significantly differ from each other. On the other hand, CA permits identifying the features (in our case cohesive categories) which contribute to these differences or commonalities. This also allows us to distinguish features as indicators of register and language variation. Moreover, we are able to trace the interplay of categories of the cohesive devices under analysis.

We use the CA package (cf. Nenadic and Greenacre, 2007) to perform correspondence analysis in the R environment. An input for CA is frequencies of the categories under analysis across registers. The output of the correspondence analysis is plotted into a two dimensional graph with arrows representing the observed frequencies of cohesive devices and triangles representing the subcorpora. The triangle position to the arrows and their length allow us to interpret their correlation. The length of the arrows indicates how pronounced a cohesive device is, see Jenset and McGillivray (2012) for details. The position of the triangles in relation to the arrows indicates the relative importance of a cohesive device for a subcorpus. The arrows pointing in the direction of an axis indicate a high contribution to the respective dimension.

*4. Data analysis*

We now present the results from the statistical tests discussed above and interpret them in the light of our research questions, which, for the sake of convenience are repeated below:

a) Main question: Are contrasts more pronounced between different registers independent of language or are more differences identified in one and the same register between English and German?
b) Which register(s) are more similar to each other and which registers are more different?
c) Which register(s) is (/are) most pronounced in the realization of particular features, across English and German?
d) Which features contribute to the observed differences/ commonalities?

*4.1 Frequency distribution*

We begin with the results obtained from the descriptive methods. Figure 1 below shows the frequency distributions of cohesive types and subtypes extracted from the subcorpora mentioned in section 4.1. The distributions are grouped according to the registers in which they occur. Subtypes of co-reference are marked in shades of blue, substitution types are marked in green, and conjunction types are marked in red.

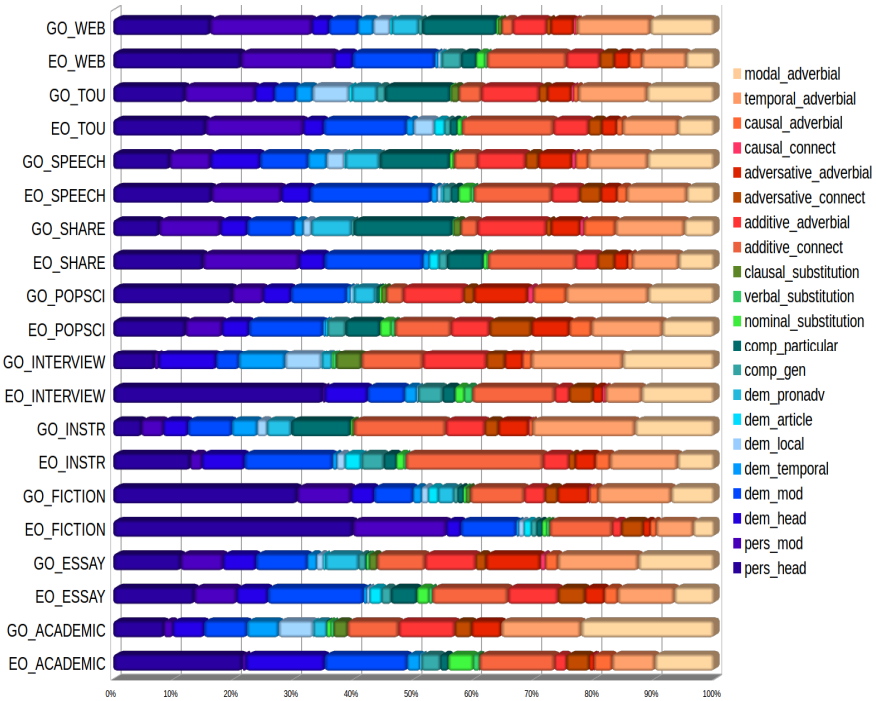

Figure 1. Frequency distribution of cohesive devices per register

Figure 1 first of all reveals that substitution plays a minor role in all registers of both languages, and that high frequencies can be found for co-reference and conjunction. In addition German seems to show a preference for relations of conjunction, whereas English seems to favour relations of co-reference. Another general observation is that there are considerable differences between the two languages in the distribution of subtypes. For instance, more co-reference relations seem to be realized in English by personal pronouns (personal head) and demonstrative determiners (demonstrative modifier), more substitution relations by nominal substitutes and more conjunctive relations by additive connects. By contrast more co-reference relations are expressed in the German subcorpora via demonstrative pronouns (demonstrative heads) and pronominal adverbs and comparative particular, more relations of substitution by clausal substitutes and more conjunctive relations via modal, additive and adversative adverbials. These language peculiarities

also translate into differences within one register between languages. However, we also see that registers can be identified across languages on the basis of particular cohesive subtypes. In particular, EO_FICTION and GO_FICTION stand out in terms of a heterogeneous distribution of cohesive subtypes, both are characterized by very high frequencies of personal reference and low frequencies of subtypes of comparative reference. The markedness of the register of fictional texts in terms of cohesion in both languages seems to be in line with observations by Neumann (2013:230ff) and Biber (1995: 151ff) in terms of lexicogrammatical features. Taken together, these features seem to reflect communication between non-experts and narrative style. In contrast to FICTION, EO_POPCSI and also GO_POPSCI seem to contain more even distributions of several cohesive subtypes. This goes along with higher frequencies for less reduced cohesive devices, such as demonstrative modifiers and different logico-semantic types of conjunctive adverbials, and may point to a more informational and/ or expository type of production (see Neumann 2013:231ff and Biber 1995:141ff).

### *4.2 Correspondence analysis*

Using the descriptive methods above we obtain information on general differences in frequency distributions between languages and registers. Yet, they do not give any information on the correlation between registers and features. Moreover, it is difficult to trace the groupings of languages and registers, as well as the discriminatory features, i.e. cohesive devices responsible for these groupings. For this, a multivariate statistic method is needed. We therefore use correspondence analysis in order to identify correlations between particular cohesive subtypes and subcorpora.

#### *4.2.1 Analysis across registers per language*

We start with the analysis of the correlation of cohesive subtypes in EO and GO separately. Figure 2 demonstrates a two-dimensional graph for English originals.

The *x*-axis reveals a distinction between English written and spoken subcorpora, although EO_POPSCI and EO_INSTR are also situated on

the left side of the borderline. However, the summary information for CA reveals their low representation in the graph (mass of 40-60), which means that they are not represented well in this subdivision. The features which contribute to this division include demonstrative reference with a local function, as well as personal reference with a modifier function as being specific for written registers of English. By contrast, demonstrative reference, with head and temporal function, nominal and verbal substitution, general comparative reference and modal conjunctions are specific for the English spoken registers EO_INTERVIEW and EO_ACADEMIC. From the grouping of these features we can conclude that spoken registers can be distinguished from written registers in English by a preference for focusing on complex referents (events). These are marked as particularly relevant or important via demonstratives heads - *that* and *these* as in (8). We also notice a tendency towards expressing relations of comparison and similarity via substitution, e.g. *do* in (9) and comparative reference, e.g. *different* in (8). In addition cohesion is often employed as a means for marking interpersonal relations, via modal conjunctions as *I mean* in (9).

(8) *if you're not convinced by **that** let me give you a second, way of packing **these**, if I don't look a - if I look at a **different** angle you see a hexagon. [EO_ACADEMIC]*

(9) *Yes, I like this little figure, yes, I definitely **do**. Some people describe, **I mean**, to me is like a little puppet version of myself. OK, OK. And you **do**, you get quite attached. [EO_INTERVIEW]*

Hence, semantically reduced forms are combined with cohesive types that mark an 'involved' style (Biber 1995). The latter, however, prevails in EO_INTERVIEW, as can also be seen in figure 1.

However, we also note a separation between the combination of EO_FICTION and EO_INTERVIEW and the rest of the subcorpora with respect to the *y*-axis. The most prominent cohesive devices in these two registers are devices of personal reference, which serve as nominal heads while the other eight English registers are characterized by the frequent use of various conjunctive subtypes. This feature distinction could mark the separation between dialogic and non-dialogic registers contained in the GECCo corpus, as fictional texts and interviews contain dialogues. However the two registers can be distinguished by clausal substitution as a typical feature of EO_FICTION and causal conjunctions as a

distinctive feature of EO_INTERVIEW, which may potentially reflect the boundary between argumentative and non-argumentative style (see Biber 1995).

(10) ***So*** *it's basically assessors coming in and looking at the school as a whole? - Yes, coming in to observe. - Yes, yes, that's it. - Is that quite stressful? Very stressful, yes. It's we had it before Christmas actually last year - and it was the most, because it was my first Ofsted, it was the most stressful thing I think I've encountered at school, so. [EO_INTERVIEW]*

Example (10) illustrates quite nicely that causal conjunctions and personal pronouns are employed in combination in EO_INTERVIEW. Quite often, very short and semantically weak forms are used, such as the neuter personal pronoun *it*, establishing a rather vague connection to previous utterances. Furthermore, conjunctions such as *so* also carry an interpersonal meaning, in that they mark the beginning or the end of a speaker turn or the willingness of the speaker to continue with his/ her speech. Hence, the involved style seems to occur in combination with argumentation here.

From these observations we can conclude that some clusters of cohesive subtypes group registers in English with respect to different modes of production (written vs. spoken) while other features of cohesion reflect other aspects of typical contexts of situation, such as speaker interaction.

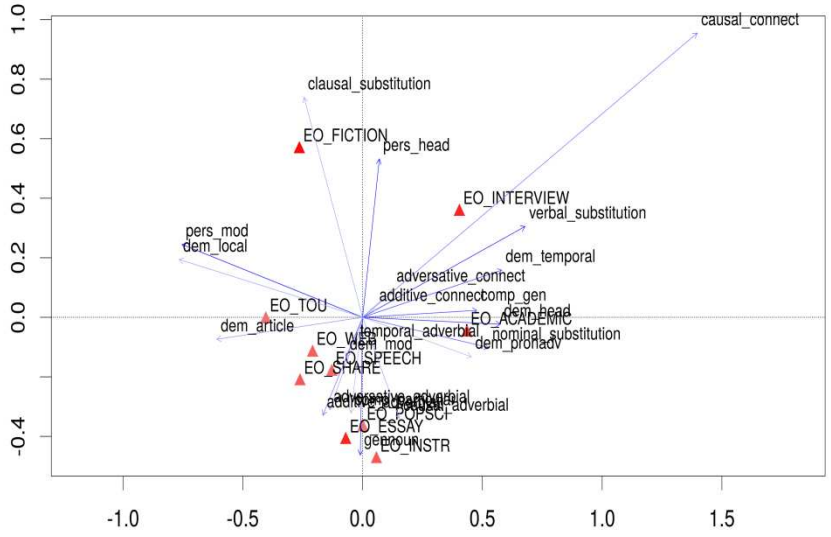


Figure 2. Cohesive subtypes in EO

The two-dimensional graph for GO registers is shown in Figure 3. Here, we also see a clear distinction between written and spoken subcorpora along the *x*-axis. Yet, partially differing features seem to contribute to this distinction in German, as compared to English: German written registers are mainly characterized by co-reference relations that are established by personal pronouns and modifiers. As in English, we count more distinctive features for the German spoken registers: verbal and clausal substitution, local and temporal reference and modal and temporal adverbials. Although the distinctive features partially differ from those for English, the realizations in the corpus reveal that the cohesive devices serve to express similar meaning relations.

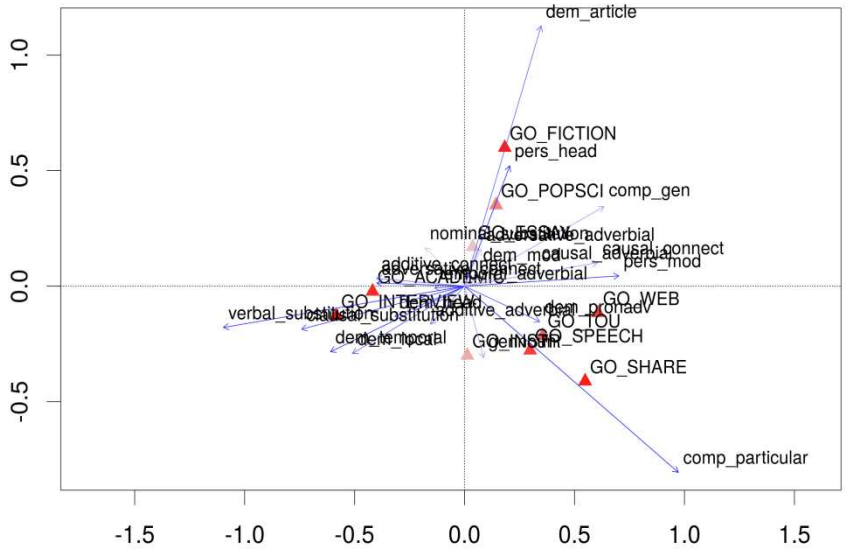


Figure 3. Cohesive subtypes in GO

As in English, we can attest a tendency in the spoken registers for highlighting important referents/ events by demonstratives, for expressing relations of comparison (e.g. by *besser* ('better'), as in (12)) and for marking speaker turns and speaker continuity, e.g. *also* ('well') and *und* ('and') (examples 11 and 12). Yet, these devices differ from those of English in the specific meaning relations they express. In contrast to English, German spoken registers additionally show a tendency towards expressing more cohesive relations of location and time. However, this does not necessarily point to a more content-oriented style of production (see House 1997). For an illustration see the textual passages below:

(11) *Und gibt es **da** noch sprachliche Probleme bei denen oder sind die sprachlich völlig integriert? - **Also da** sind die Probleme massiv und **da** sieht man eigentlich auch, wo des Problem bei uns in Deutschland liegt, ja, dass die einfach mit zu geringen Sprachkenntnissen [GO_INTERVIEW] …*

(12) *And are **there** also linguistic problems amongst them or are they fully integrated linguistically? **Well**, **there** the problems are*

*huge and **there** you can also see, what the problem really is in Germany …*[2]

*(13) **Und da** nehm ich an, ich könnte auch sagen eine endliche Menge. Aber irgendwie ist es **besser**, wenn ich sage eine unendliche Menge [GO_ACADEMIC]*

*(14) **And there** I expect, I could also say a finite number. But somehow it is **better** if I say an infinite number …*

In examples (11) and (12) we find the locative adverb *da* ('there'), which is used with a very high frequency in the spoken registers of German. If we compare instantiations of *da*, we observe that it is employed with a multitude of cohesive meanings, sometimes it tends to express time (in the sense of *then*) rather than location, and sometimes its meaning is rather metaphorical and mainly textual (as in 12). These observations suggest a combination of our empirical findings with insight from quantitative analysis in the future. Statistical evidence for differences in the spoken registers between the two languages are examined with the combined analysis below.

Concerning the y-axis, we observe a similar tendency as in English: fictional texts are opposed to the other registers. GO_POPSCI, GO_INSTR and GO_ESSAY, which are located on the same side of the separation line as GO_FICTION are poorly represented in this dimension. This means that German FICTION is a very distinctive register in terms of cohesive categories and, in contrast to English, does not cluster with EO_INTERVIEW. Most prominent features, which contribute to this distinction, are cohesive personal heads and demonstrative articles. Hence, this separation is no reflection of dialogicity but may rather have to do with the narrative passages contained in the German fictional texts. The passages contain rather long co-reference chains which denote the same protagonist again and again, via semantically reduced pronouns. The protagonist is involved in events or concerned with various objects which are realized in smaller co-

[2] Note that the English translations for passages from the German spoken registers are literal in order to reflect German peculiarities of cohesion. The translations for the passages for German written registers are taken from the GECCo corpus.

reference chains that are triggered by combinations of definite articles and lexical cohesion.

(15) *Man konnte [den Schatten] vor uns zusehen, wie **er** sich näherte, bis **er** unter [der Motorhaube] verschwand, kurz darauf **die** Motorhaube erklomm, über die Windschutzscheibe kroch, auf unsere Gesichter, und schließlich den Wagen verschluckte, rücksichtslos, wie **er** alles verschluckte, was vor **ihm** lag. [GO_FICTION]*

(16) *The shadow in front of us could be seen approaching until **it** vanished below the hood, climbed the hood a moment later, crawled across the windshield onto our faces, and finally swallowed the car as ruthlessly as **it** swallowed all that lay before **it**, the shadow of that wide roof, of the building that straddled the road and blocked our view. [ETRANS_FICTION]*

In the passage presented in (13), the protagonist is a personified object – *der Schatten* ('the shadow'), which is taken up by masculine personal pronouns in singular (*er*, *ihm*). It performs several actions in which objects such as *die Motorhaube* ('the hood') are involved.

*4.2.2 Analysis across registers and languages*

In the next step, we compare all subcorpora under analysis, which is represented in the two-dimensional graph in Figure 4.

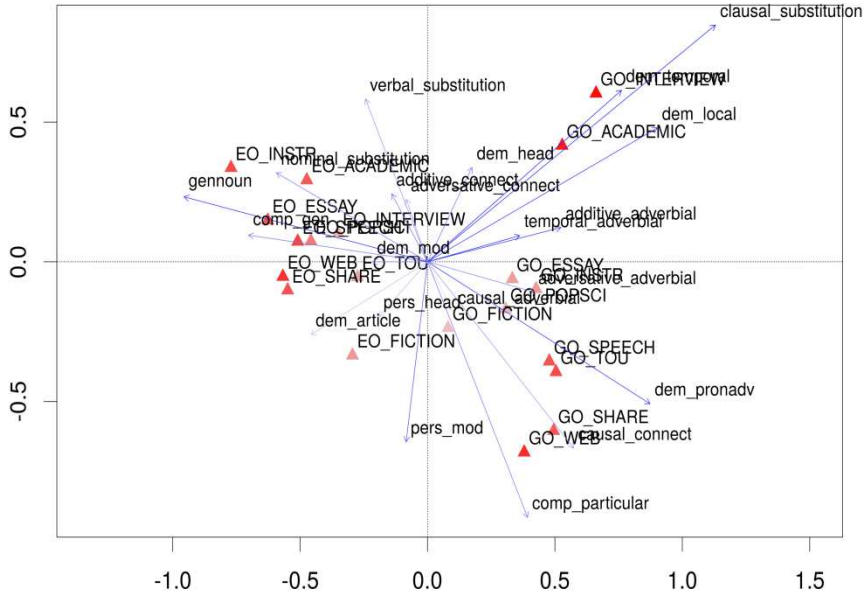

Figure 4. Cohesive devices in EO and GO

A clear separation between the two languages is observed along the *x*-axis. Characteristic cohesive features for German registers include demonstrative reference with local and temporal functions and pronominal adverbs, conjunctive relations expressed with adverbials as well as clausal substitution. In English, personal heads, definite articles and general comparative reference contribute to distinction. This corroborates our earlier observations based on the descriptive analyses described in section 5.1 above.

Hence, from a semantic perspective German prefers to mark recurring referents as particularly relevant or important by focus lifters such as demonstrative adverbs, e.g. *da* in (14), demonstrative pronouns, as *der* in (14), which does not have an equivalent in English and pronominal adverbs, as in (15). The high frequency of the latter is a peculiarity of the German language system.

(17) *und plötzlich verwandelt sich einer von den Protagonisten in ein Nashorn. Und plötzlich verwandeln sich immer mehr in ein Nashorn, können danach nicht mehr sprechen. Sie haben also* ***da*** *auch das Problem der Sprache und es gibt einen einzigen,* ***der***

*bleibt übrig. **Der** verwandelt sich nicht eh und **der** denkt darüber nach,... [GO_ACADEMIC]*

*(18) And suddenly, one of the protagonists turns into a hippopotamus. And suddenly, more of them are transformed into a hippopotamus, not able to speak afterwards. So **there** they have the problem with language and there is only one person, **he** remains. **He** does not transform and he thinks about …*

*(19) Andererseits müssen wir den Dialog gerade mit der islamischen Welt verstärken und intensivieren. **Dabei** geht es darum - … - die sämtlichen Weltkulturen gemeinsamen Werte sichtbar zu machen. **Dazu** gehört auch das unzweideutige Eintreten für die Menschenrechte … [GO_SPEECH]*

*(20) On the other hand, we have to strenghten dialogue with the Islamic world. **This** is a matter of visualizing the values shared by all cultures of the world. **This** also includes defending human rights unambiguously. [ETRANS_SPEECH]*

Fewer constructions are available in English, and they are furthermore instantiated less often. Hence, English texts seem to be less marked and more neutral in terms of taking up referents in the textual world. Instead personal pronouns or modifiers seem to be employed more often to establish identity between animate referents and definite articles (in combination with lexical cohesion) seem to be used more often for inanimate referents.

*(21) When we left my mother said, "Of what? What does **he** have to be careful about? **They** put the tray out so people can look at the things, don't they? So what does **he** have to be careful about?"[EO_FICTION]*

*(22) The superb river poised in such elegance and folded to roar down its fight full throat of rapid, they will try crossing here because it is so narrow, they are certain they could throw a stone over **it** if they still had their usual strength. [EO_FICTION]*

In (16) personal pronouns are employed where the demonstrative pronouns *der* and *die* could be used in German. In (17), the meaning of the neuter pronoun *it*, which refers to *the superb river*, could be realized with the pronominal adverb *darüber* in German.

Furthermore, German seems to prefer more explicit devices to mark conjunctive relations than English since adverbials are less reduced semantically than coordinating conjunctions. This is illustrated in (18) and (19). Addition is more often expressed in German by a conjunctive adverbial, while English more often employs the simple coordinator *and*:

(23) *Aufgrund der veränderten Sicherheitslage konnte die Mannschaftsstärke um 40 Prozent reduziert werden.* ***Außerdem*** *wurden, …, knapp 11000 ehemalige Soldaten der Nationalen Volksarmee der DDR in die nun gesamtdeutsche Bundeswehr integriert. [GO_ESSAY]*

(24) *The new security situation made it possible to reduce personnel by 40 %.* ***Furthermore****, almost 11,000 former soldiers from the GDR's National People's Army (NVA), excluding higher ranking officers, were integrated into the new all-German Bundeswehr. [ETRANS_ESSAY]*

(25) *Mr. Bush has time and again demonstrated his commitment to open trade.* ***And*** *he is determined to extend the benefits of open markets to the world's poorer nations. [EO_ESSAY]*

The y-axis represents the separation between spoken and written registers in both languages: we have the constellation of EO_ACADEMIC, EO_INTERVIEW, GO_ACADEMIC and GO_INTERVIEW on the one side, which are opposite to the other registers in both languages. Further registers, e.g. EO_INSTR or EO_SPEECH can be seen on the graph but are poorly represented in this dimension. Demonstrative heads and conjunctive devices expressing additive and adversative relations contribute to the commonality between English and German spoken registers, whereas comparative particular as illustrated in (20) for German and (21) for English distinguishes the written registers from the spoken registers.

(26) *Wenn sich in der Folge trotzdem ein Sinn zeigte, dann auf eine viel* ***kompliziertere*** *und* ***fragwürdigere*** *Weise. [GO_POPSCI]*

(27) *If subsequently a meaning nevertheless appeared, then in a much* ***more complicated and dubious*** *way. [ETRANS_POPSCI]*

(28) *Today interest rates which were 4 per cent above those of the euro area are now 1.75 per cent* ***higher****. [EO_ESSAY]*

In addition, we note that devices of substitution play a role for the distinction of spoken and written registers both in English and German although their general distributions as observed in 5.1 above are rather low in relation to devices of co-reference and conjunction. Yet, spoken English is more characterized by the use of verbal (and to a lesser degree nominal substitution) as shown in (22), whereas spoken German clearly favors clausal substitution, as illustrated with cataphoric *sowas* in (23).

*(29) Yes, I like this little figure, yes, I definitely* ***do****. Some people describe, I mean, to me is like a little puppet version of myself. OK, OK. And you* ***do****, you get quite attached. [EO_INTERVIEW]*

*(30) Ehm was machen Sie damit? - Ja zum Beispiel* ***sowas****: Wir nehmen einen ganz einfachen Prozess, das Proz - dieser Prozess ist ein Streichholz, …[GO_INTEVRIEW]*

*(31) What do you do with it? – Well, for instance,* ***something like this****: We take a simple process, the proc (truncated word) – the process is a quick match …*

The commonalities observed across languages in the spoken registers point to preferences for realizing particular semantic meaning relations, as already discussed above. They have a tendency towards using cohesive items with an interpersonal or rhetorical function. In addition higher frequencies for cohesive devices of substitution and general comparative reference seem to point to a lower degree of explicitness of cohesive devices, which may generally create more vagueness in terms of the relations expressed by cohesion in spoken as compared to written registers. Mode of production therefore seems to be the variable along which registers can be differentiated cross-linguistically.

However, the difference between the languages is greater than that between spoken and written modes. We comprehend it from the eigenvalues of the two dimensions represented in the graph: the first dimension contributes ca. 38%, whereas the second contributes around 26%.

Figure 4 additionally shows that registers in English cluster more densely than German registers and also that the distance between spoken and written registers is more pronounced in German than in English. These observations therefore seem to support earlier assumptions about contrastive tendencies in lexicogrammar by Mair (2006) and Leech et al.

(2009), and reflect a higher degree of variation between registers in general and written and spoken registers, in German than English.

## *5. Summary and conclusions*

The study presented above has been concerned with the corpus-based analysis of cohesion in different registers of English and German. Our aim has been to identify contrasts and commonalities in the registers under investigation with respect to types and subtypes of co-reference, substitution and conjunction.

For this purpose, we have interpreted corpus data which was evaluated by different statistical methods. Descriptive methods have been employed to observe types and subtypes of cohesive devices in registers within and across languages. The results of our correspondence analyses show that we can observe additional variables, which include combinations of subcorpora under analysis: mode (spoken vs. written), language (German vs. English), registers (FICTION vs. others); and if we observe languages separately, other dimensions come into the fore, e.g. monologic vs. dialogic texts in English and narration vs. other speaker goals in German.

Concerning our main research question formulated at the outset of this study, we can generally state that contrasts are more pronounced between the two languages English and German than between registers. The main differences are attested in terms of the preferred meaning relations: a preference for explicitly realizing logico-semantic relations by cohesive devices of conjunction and tendency towards realizing more relations of identity by cohesive devices of referents. In addition, different subtypes are preferred in the two languages for realizing similar meaning relations.

Yet – and this observation targets our second question - mode of production plays an essential role for the grouping of particular registers in the two languages separately and also across languages. The spoken registers of ACADEMIC and INTERVIEW stand out in their preference for highlighting referents and events, comparing them and evaluating them by means of cohesive relations. Their lexicogrammatical realizations via cohesive devices, however, are partially language-specific.

Furthermore, the register of FICTION seems to be marked by distinct cohesive features which reflect peculiar contextual configurations in both languages. Again there are language-specific preferences which reflect distinctions on a more semantic or pragmatic level: dialogicity in English, and narration in German. These observations suggest a need for further analysis, in which we should take into account the representation of particular registers in the dimensions. For example, we should exclude POPSCI and INSTR when analyzing registers within a language, and study them separately.

Our analyses have also shown that the two languages differ as to the degree of variation between individual registers. We can find more variation in the realization of cohesive devices in German than English. In order to obtain further empirical evidence of this tendency, we will have to combine insights about cohesive devices with studies of the elements they tie up with, and the cohesive relation as such. This will yield empirical evidence in terms of distance between elements in cohesive chains, chain size and chain length. In addition, our quantitative analyses have to be accompanied by qualitative analyses. They will allow us to investigate the structural environments of the cohesive relations as well as the specific functions and meaning relations expressed.

*References*

Ariel, M. (1990). *Accessing Noun-phrase Antecedents*. London/New York: Routledge.

Baayen, H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Bestgen, Y., Degand, L. & W. Spooren. (2006). Towards Automatic Determination of the Semantics of Connectives in Large Newspaper Corpora. In: *Discourse Processes*. 41, 175-193.

Biber, D. (1995). *Dimensions of Register Variation. A cross-linguistic comparison*. Cambridge. CUP.

Blühdorn, H. (2008). *Syntax und Semantik der Konnektoren. Ein Überblick*. Mannheim: Institut für Deutsche Sprache, Manuskript .

Brinker, K. (2005). *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden*. Erich Schmidt, Berlin. 6 edition.

Bührig, K. & J. House. (2004). Connectivity in Translation: Transitions from Orality to Literacy. In: House, J. and J. Rehbein (eds), *Multilingual Communication*. Amsterdam: Benjamins. 87-114.

CWB. (2010). The IMS Open Corpus Workbench. http://www.cwb.sourceforge.net.

Dearborn, F, Chicago Nenadic, O. & M. Greenacre. (2007). Correspondence Analysis in R, with two- and three-dimensional Graphics: The CA Package. In: *Journal of Statistical Software*, 20(3):1-13.

De Beaugrande, R.-A. & W.U. Dressler. (1981). *Einführung in die Textlinguistik.* Tübingen: Niemeyer.

Dipper, S. & M. Stede. (2006). Disambiguating Potential Connectives. In *Proceedings of KONVENS-06*. Konstanz, Germany. 167-173.

Doddington, G, Mitchell, A., Przybocki, M, Ramshaw, L., Strassel, S & R. Weischedel. (2004). The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2004). 837-840.

Eckert, M. & M. Strube. (2000). Dialogue Acts, Synchronising Units and Anaphora Resolution. In: *Journal of Semantics*. 17(1):51–89.

Evert, S. (2005). *The CQP Query Language Tutorial*. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, April. CWB version 2.2.b90.

Gundel, J., N. Hedberg & R. Zacharski. (1993). Cognitive Status and the Form of Referring Expressions in Discourse. In: *Language*. 69/2: 274-307

Gundel, J.K., Hedberg, N. & R. Zacharski. (2004). Demonstrative Pronouns in Natural Discourse. In: *Proceedings of the 5th Discourse Anaphora and Anaphora Resolution Colloquium*. 81 – 86.

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

Halliday M. A. K. & Ch. Matthiessen. (2013). *Halliday's Introduction to Functional Grammar*. London: Routledge.

Hansen-Schirra, S., Neumann, S. & E. Steiner. (2012). *Cross-linguistic Corpora for the Study of Translations. Insights from the language pair English - German.* Series Text, Translation, Computational Processing. Berlin, New York: Mouton de Gruyter.

Hinrichs, E.W., Kübler, S. & K. Naumann. (2005). Representation for Morphological, Syntactic, Semantic, and Referential Annotations. In: *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*. 13-20.

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

Jenset, G. B. & B. McGillivray. (2012). Multivariate Analyses of Affix Productivity in Translated English. In: M. P. Oakes and M. Ji, (eds.), *Quantitative Methods in Corpus-Based Translation Studies*. John Benjamins. 301-324.

Klein, D. & C. D. Manning. (2003). Accurate Unlexicalized Parsing. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL 2003, Stroudsburg, PA, USA. Association for Computational Linguistics. 423-430.

Kunz, K. & E. Steiner. (2013a). Towards a Comparison of Cohesive Reference in English and German: System and Text. In: Taboada, M., Doval Suárez, S. and E. González Álvarez (eds). *Contrastive Discourse Analysis. Functional and Corpus Perspectives.* London: Equinox, 208-239

Kunz, K. & E. Steiner. (2013b). Cohesive Substitution in English and German a Contrastive and Corpus-based Perspective. In: Aijmer K. and B. Altenberg (eds), *Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson.* Amsterdam: John Benjamins, 201-232

Kunz, K. & E. Lapshinova-Koltunski. (2014). Cohesive Conjunctions in English and German: Systemic Contrasts and Textual Differences. In: Davidse, K, Gentens, C., Kimps, C & and L. Vandelanotte, (eds.). *Recent Advances in Corpus Linguistics: Developing and Exploiting Corpora*. Rodopi, Amsterdam, 229-262.

Kunz, K. & E. Lapshinova-Koltunski. (2011). Tools to Analyse German-English Contrasts in Cohesion. In: Hedeland, H., Schmidt, T. and K. Worner (eds.). *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language technology* (GSCL) 2011. 243-246.

Lapshinova-Koltunski, E. & K. Kunz. (2014). Conjunctions across Languages, Registers and Modes: Semiautomatic Extraction and Annotation. In: Diaz Negrillo, A. & F. J. Daz Prez, (eds.).

*Specialisation and Variation Language Corpora.* Peter Lang. Papers from the CILC2012. 77-104.

Lapshinova-Koltunski, E., K. Kunz & M. Amoia. (2012). Compiling a Multilingual Spoken Corpus. In: Mello, H., and M. Pettorino (eds). *Proceedings of the VIIth GSCP-2012 International Conference: Speech and Corpora*. Firenze: Firenze University Press. 79-84.

Leech, G., M. Hundt, C. Mair, & N. Smith. (2009). *Change in Contemporary English. A Grammatical Study*. Cambridge: CUP.

Louwerse, M.M. & A.C. Graesser. (2005). Coherence in Discourse. In: P Strazny, editor, *Encyclopedia of Linguistics*, 216–218.

Mair, C. (2006). *Twentieth-Century English. History, Variation and Standardization*. Cambridge: CUP.

Martin, J.R. (1992). *English Text.* Amsterdam: Benjamins.

Nenadic, O. & M. Greenacre. (2007). Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package. In: *Journal of Statistical Software*. 20(3).

Neumann, S. (2013). *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Berlin/ Boston: de Gruyter

Pasch, R., Braußе, U., Breindl, E. and U.H. Waßner. (2003). *Handbuch der deutschen Konnektoren: Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln)*. Berlin: Walter de Gruyter.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R & N. Xue. (2011). Shared Task: Modeling Unrestricted Coreference in OntoNotes. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (CoNLL 2011), 1-27.

Prasad, R., Dinesh, N, Lee, A., Miltsakaki, E. Robaldo, L., Joshi, A & B. Webber. (2008). The Penn Discourse Treebank 2.0. *In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*. Marrakech, Morocco.

Prince, E.F. (1981). Towards a Taxonomy of Given-New Information. In: Cole, P. (ed.), *Radical Pragmatics.* New York: Academic Press, 223-255.

Schubert, C. (2008). *Englische Textlinguistik. Eine Einführung*. Berlin: Erich Schmid Verlag.

Schwarz, M. (2000). *Indirekte Anaphern in Texten: Studien zur domänengebundenen Referenz und Kohärenz im Deutschen*. Tübingen: Niemeyer.

Stede, M. (2008). Connective-Based Local Coherence Analysis: A Lexicon for Recognizing Causal Relationships. In: Bos, J. and R. Delmonte (eds.). *Semantics in Text Processing (STEP-2008). Research in Computational Semantics Series*. London: College Publications. 221-237.

Taboada, M. & MLA Gómez-González. (2012). Discourse Markers and Coherence Relations: Comparison across Markers, Languages and Modalities. In: *Linguistics and the Human Sciences* 6 (1-3): 17-41.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer. (2003). Feature-rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: *NAACL 2003: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Morristown, NJ, USA. Association for Computational Linguistics. 173-180.

Venables, W. N. and D. M. Smith. (2010). *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics*. Electronic edition.

Weischedel, R., Palmer, M., Marcus, M.; Hovy, E., Pradhan, S., Ramshaw, L., N., Xue, Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R. and A. Houston. (2013). OntoNotes Release 5.0 LDC2013T19. Web Download. Philadelphia: Linguistic Data Consortium.

Zinsmeister, H., Dipper S. & M. Seiss. (2012). Abstract Pronominal Anaphors and Label Nouns in German and English: Selected Case Studies and Quantitative Investigations. In: *Translation: Corpora, Computation, Cognition.* Special Issue on the Crossroads between Contrastive Linguistics, Translation Studies, and Machine Translation. 47-80.