Across Languages and Genres: Creating a Universal Annotation Scheme for Textual Relations

Ekaterina Lapshinova-Koltunski
Saarland University
Universitat Campus A2.2
66123 Saarbrucken
e.lapshinova
@mx.uni-saarland.de

Anna Nedoluzhko
Charles University in Prague
Malostranske nam. 25,
CZ-11800 Prague 1
nedoluzko
@ufal.mff.cuni.cz

Kerstin Anna Kunz
University of Heidelberg
Ploeck 57a
DE-69117 Heidelberg
kerstin.kunz
@iued.uni-heidelberg.de

Abstract

The present paper describes an attempt to create an interoperable scheme using existing annotations of textual phenomena across languages and genres including non-canonical ones. Such a kind of analysis requires annotated multilingual resources which are costly. Therefore, we make use of annotations already available in the resources for English, German and Czech. As the annotations in these corpora are based on different conceptual and methodological backgrounds, we need an interoperable scheme that covers existing categories and at the same time allows a comparison of the resources. In this paper, we describe how this interoperable scheme was created and which problematic cases we had to consider. The resulting scheme is supposed to be applied in the future to explore contrasts between the three languages under analysis, for which we expect the greatest differences in the degree of variation between non-canonical and canonical language.

1 Aims and Motivation

The aim of the present study is to create a scheme which will allow us to use existing annotations of textual phenomena, and which will be applicable to multiple languages and genres, including non-canonical ones. The annotations were created within two separate projects: German-English Contrasts in Cohesion (GECCo, Lapshinova and Kunz (2014)) whose focus was on English and German on the one hand, and the Prague Dependency Treebank (PDT 3.0, Bejček et al. (2013)) with the analysis of Czech, on the other hand.

The resulting scheme will serve our overarching goal to unify the two approaches in a joint analysis of contrasts between English, German and Czech on the level of discourse. We assume that the greatest differences between these languages lie in the degree of variation between non-canonical and canonical language (here we especially mean spoken language). Previous findings on lexico-grammatical and also cohesive phenomena have evidenced that there is more variation between written and spoken dimensions in German than in English, even though they are closely related, cf. Mair (2006) or Kunz et al. (forthcoming). Studies with respect to spoken and written Czech (see, e.g., Cvrček et al. (2010)) suggest that the differences between written and spoken language are even more pronounced in Czech than in German, at least with respect to lexico-grammar, hence we expect that this also holds for the level of text/ discourse.

We therefore suggest that if we draw a line of differences between spoken and written English, German and Czech, we would observe a continuum in the degree of variation between these languages, as seen in Figure 1. The graph also reflects the above assumption that the differences are less pronounced between English and German than if we compare English and German with Czech. The reasons for this lie in the linguistic heritage of these languages (English and German have a common West-Germanic origin while Czech is a Slavic language) and in sociolinguistic factors that influenced their evolution (for example, Czech purism at the beginning of the 20th century, described, e.g., in Havránek and Weingart (1932)). To our knowledge, there is no

Proceedings of LAW IX - The 9th Linguistic Annotation Workshop, pages 168–177,
Denver, Colorado, June 5, 2015. ©2015 Association for Computational Linguistics

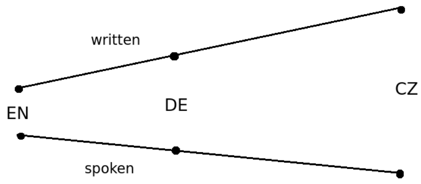

Figure 1: Differences between spoken and written English, German and Czech

research testing these assumptions. We believe that a cross-language analysis based on the interoperable scheme proposed in this work will help to fill this gap.

However, this kind of study requires corpora that are annotated for textual phenomena. As the creation of such corpora is a time-consuming task, we decide to take advantage of existing resources, i.e. corpora, which already contain annotations of these phenomena. However, while capturing the same phenomena, the annotations in the corpora at hand were created in the frame of two different projects (GECCo and PDT, see Section 2). Moreover, both existing annotation schemes only account for the systemic peculiarities and realizational options of the languages analysed and hence are not general enough to permit a comparison across Germanic and Slavic languages. For this reason, we need to unify the categories in these schemes to create an interoperable one which can be applicable to multiple languages and text registers, including spoken ones. The scheme will allow us to profit from the existing annotated resources and at the same time will enable the contrastive analysis of the languages involved. We believe that the resulting scheme will find application not only in our research, but also in further linguistic studies and in cross-language NLP applications. It is beyond the scope of this paper to include the contrastive language analysis, which will follow from the unified scheme in our future work.

2 Theoretical Background

In this section, we describe the frameworks for the analysis of English, German and Czech. They were used in the development of the resources at hand (which are described later in Section 3) and will serve as a basis for our interoperable scheme.

2.1 Frameworks for the analysis of English and German

The analysis of textual phenomena in GECCo is based on the definition of cohesion. The concept was established by Halliday and Hasan (1976) for English, in the frame of Systemic Functional Linguistics. It concerns textual relations between linguistic expressions across grammatical domains. Additionally, the categories under analysis are based on the conceptualisations of de Beaugrande and Dressler (1981), who consider cohesion as an explicit linguistic signal on the text surface to establish coherence or textuality. Cohesion always involves a linguistic trigger (cohesive device) that links up to other linguistic expressions in the same text. The main categories used in the analysis include coreference to create relations of identity, comparative reference, substitution and ellipsis to create relations of comparison between referents belonging to the same type, conjunction for logico-semantic relations between propositions, and lexical cohesion for similarity between different types of referents. The adaptation of these categories and their subcategories to the bilingual comparison of English and German have been described in Kunz et al. (forthcoming). For coreference, ellipsis and lexical cohesion, not only cohesive devices were considered, but also the linguistic expressions they tie up with as well as the cohesive relations. The relations may contain more than just two linguistic expressions and form cohesive chains that stretch over longer passages of text.

2.2 Framework for the analysis of Czech

In the framework for the analysis of Czech, the following textual phenomena are included: ellipses, information structure, grammatical and textual coreference, bridging relations (associative anaphora) and discourse relations. Their definition is based on Functional Generative Description as described in Sgall et al. (1986). The approach uses syntactic as well as semantic criteria for text analysis and considers three layers of text representation: morphological, analytical and tectogrammatical (deep syntactic). At the tectogrammatical layer, the meaning of the sentence is represented as a dependency tree structure, in which nodes represent autosemantic words and are labelled with a large set of attributes. This layer of representation is especially important for elliptical constructions, as they are captured here in reconstructions (Mikulová, 2014). Besides that, the tectogrammatical layer also covers information on structural attributes (in terms of contextually bound or contextually non-bound nodes). The approach to textual phenomena exceeding the sentence boundary is two-fold for the Czech framework. On the one hand, the conception of discourse relations is based on the Penn-style discourse lexically-grounded approach, as described in Prasad et al. (2008). According to this approach, only those relations that are signaled by explicit markers (connectives) are considered as discourse relations. However, in contrast to the Penn-style approach, the set of connectives is an open list, see Poláková et al. (2013), and the treatment of coreference and bridging relations includes both explicit and implicit categories. Language expressions that refer to the same discourse entity are considered to be coreferent. As for bridging relations, their definition has been taken from Clark (1975).

3 Data and Experiment

As already mentioned in Section 1, we aim to take advantage of the existing corpora annotated for textual phenomena to avoid the time-consuming creation of such resources. The existing German and English data are annotated with the GECCo framework described in 2.1, whereas the data for Czech are annotated in the PDT style described in section 2.2 above. The current section provides a brief description of these resources at hand.

3.1 GECCo - German and English corpora

The GECCo corpus annotated for textual phenomena with the framework described in 2.1 represents a continuum of different text types (registers in the sense of Systemic Functional linguistics) from written to spoken discourse. More precisely, it includes English and German texts of ten registers, eight of which represent written discourse and include fictional texts, political essays, instruction manuals, popular-scientific texts, letters to shareholders, prepared political speeches, tourism leaflets and texts from corporate websites. This part contains not only original texts, but also their translations in both directions. The registers of spoken discourse include recorded and transcribed interviews and academic speeches described in Lapshinova-Koltunski et al. (2012), as well as transcriptions of television talkshows, texts from internet forums, medical consultations and sermon texts. The total number of words contained in the corpus comprises ca. 1,6 Mio (including translations). The corpus is annotated on several levels, which include morphological, syntactical, structural and textual information (i.e. information on cohesion as described above). The information on the latter was annotated with the help of semi-automatic procedures described by Lapshinova-Koltunski and Kunz (2014). These result from an integration of the systemic peculiarities of English and German and at the same time account for textual variation in terms of canonical written and non-canonical spoken language. The rich annotation allows capturing information about the structural and syntactic features of cohesive devices (and also antecedents) and about how they are mapped onto information structure. Moreover, it yields information about chain features, e.g. number of elements in chains, distance between chain elements and number of different chains.

3.2 Prague Dependency Treebanks

There is a number of corpora annotated according to the Prague annotation scenario described in section 2.2 above. These include PDT 3.0 – Prague Dependency Treebank (Bejček et al., 2013), PCEDT 2.0 – Prague English Dependency Treebank (Hajič et al., 2012) and PDTSL – Prague Dependency Treebank of Spoken Language (Hajič et al., 2009). All these corpora consist of original texts (Czech and English respectively) extracted from newspaper articles (PDT), Wall Street Journal (PCEDT) and transcribed and reconstructed spontaneous dialogue speech in Czech and English. PCEDT 2.0 also contains translations from English into Czech. The total number of words in written corpora comprises ca. 3,2 Mio (including translations) and spoken corpora for English and Czech total ca. 770 thousand tokens. The written corpora are annotated with morphological, analytical and tectogrammatical information, whereas each sentence is represented as a dependency tree structure. The tectogrammatical layer of PDT 3.0 also contains annotation of information structure attributes and the following discourse phenomena: extended (nominal) textual coreference, bridging relations, discourse connectives and the discourse units linked by them, and semantic relations between these units, see Poláková et al. (2013) for details.

3.3 Experiment settings

The creation of an interoperable scheme requires a comparison of the underlying annotations. We therefore annotate the same data set on the basis of both conceptions, and identify those categories that cover the same phenomena. For this, we have selected texts in English (both originals) belonging to two different genres – journalism and fiction and annotated them in accordance with the guidelines of the Prague and GECCo conceptions. Journalistic texts represent written discourse, whereas the fictional texts we selected are closer to spoken language and other non-canonical genres, e.g., internet blogs or tweets. They are partially narrative and partially dialogic, and hence contain turns, but also reformulations, elaboration and other spoken language features. We believe that this data constellation ensures a good base for our future analysis (aimed at comparison of spoken vs. written dimensions). We decide for texts in English, as English data is available in both underlying resources, hence allowing us to unify the annotated categories afterwards. The journalistic sample contains texts exported from PCEDT 2.0 (see section 3.2), with a size of around 100 sentences. A sample of fictional texts of the same size was exported from the GECCo corpus described in 3.1. For the sake of convenience, we used different annotation tools for the two different frameworks – TrEd (Pajas and Štěpánek, 2008) for the framework described in 2.2, as it allows visualisation of trees, and MMAX2 (Müller and Strube, 2006) for the framework described in 2.1, as this enables visualisation of longer cohesive chains. The annotations were carried out manually by four trained annotators. Then, the parallelly created annotations were compared and analysed qualitatively and quantitatively. The results of this analysis are presented in section 4 below.

4 Analyses

4.1 Overall comparison

Both GECCo and PDT frameworks include annotations of ellipses, coreference relations and discourse connectives. The category of lexical cohesion in the German-English framework (see section 2.1) can be partially mapped to bridging relations in the Czech framework (see 2.2), although lexical cohesion is much more lexically grounded than bridging. Substitution is the only phenomenon which is asymmetric in the frameworks. It is not covered by the definition of textual relations in the framework for Czech, as this device is common for English and (less so) for German but not relevant at all for Czech. We provide a mapping of the phenomena available in both frameworks in Table 1.

GECCo	PDT
coreference	coreference
lexical cohesion	bridging
ellipsis	ellipsis (in dependency trees)
connectives, relations	connectives, arguments, relations
substitution	-

Table 1: Mapping of the phenomena

We count the occurrences of these categories in the experimental dataset and compare absolute numbers for both frameworks, see Figure 2. The numbers in Figure 2 reveal the preferences for certain types of relations in the two approaches involved. At the same time, we are able to observe the similarities between the types.

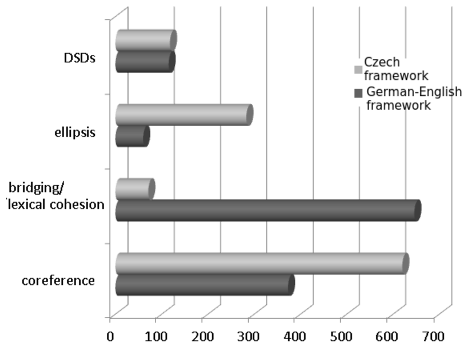

Figure 2: Overall annotation statistics

What is most evident from the figure is that the number of **discourse relations** expressed by connectives[1] annotated in both approaches is very similar. This is mainly due to the fact that the typology of discourse relations of the main categories is similar in both approaches. Neglecting the terminology, there are four main relations in both approaches: temporal, causal, adversative and additive. In GECCo, also modal DSDs are distinguished (such as *well, sure, of course, surely*, etc.). They are especially frequent in spoken genres. However, they only provide a rather vague link to the two arguments, as they primarily carry an emotional meaning. For this particular reason, this type of textual devices is not included in the PDT framework, where a DSD always requires a clear linkage of two arguments, and in which the scope of discourse arguments is taken into account. If modal DSDs were substracted, the number of connectives for the German-English framework would slightly change. However, it does not change the comparison considerably. The other difference observed in the approach to discourse relations is that, in the Penn-style, the four main categories are further differentiated into more detailed relations, whereas in the German-English framework, only the general categories are considered.

The numbers for the other textual phenomena reveal more differences. For example, the frequencies of **ellipses** and coreference relations annotated within the PDT framework prevail over those of the other types. This is justified by the representation of the phenomena according to the framework: Apart from textual ellipses (*Did she open the door? No, she did not [open the door]*), it also contains various grammatical types of elliptical constructions, e.g. structural ellipses (ellipses of governing verbs and nouns), different kinds of anaphoric zeros (*Their reaction was 0 to do nothing and 0 ride it out*), including arguments with control constructions (*Peter want to [Peter] sleep*), general arguments (*Jane sells at Bata [what] [to whom]*), etc. These are reconstructed on the deep syntactic level. The GECCo approach is based on signals to textual cohesion, and therefore, ellipses are annotated only in the case of textual relations across grammatical domains. Besides, anaphoric zeros are not reconstructed in syntactic structures.

For our contrastive analysis, we will consider cases of textual ellipsis only, which are expected to contribute especially to the differences between spoken and written language. We expect textual ellipsis to be more common in spoken genres, as our previous analyses for English and German have already evidenced, cf. Kunz et al. (forthcoming). Example (1) demonstrates a case of textual ellipsis considered in both approaches.

(1) *He'd never even bothered to read it. But Truman had [].*

The difference here lies in the representation of the missing element. In the GECCo approach, this case is annotated as verbal ellipsis. The missing parts of the verbal phrase could either be *bothered to read it* or *read it*. In the PDT approach, the whole verbal phrase is reconstructed in the dependency tree, see Figure 3, connected to the antecedents of verbs by the arrows of grammatical and textual coreference. Note that this type of ellipsis, where only the operator is kept (termed as lexical ellipsis by Hallidday & Hasan (1976)), is available in English, but neither in German nor Czech.

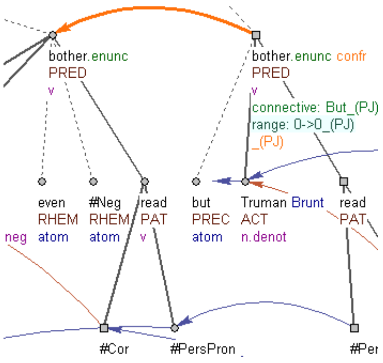

Figure 3: Ellipsis in the dependency tree representation (PDT-style)

The differences in the annotations of **coreference** are due to the diverging definitions of coreferring expressions. In GECCo, only the mentions with an explicit marker, the cohesive device (e.g. definite articles, pronouns, demonstratives, etc.), are taken into account. This implies, for instance, that relations between named entities or between nominal construc-

[1] hereinafter referred to as discourse-structuring devices (DSDs).

tions in plural which are not introduced by a determiner are excluded from the annotation of coreference. They are, however, annotated as devices of lexical cohesion (see below). Moreover, as a cohesive relation to the antecedent is indicated by a cohesive device, only this explicit marker is annotated but not the other elements of the anaphoric nominal phrase. Hence, if an anaphoric expression consists of a definite article and a nominal head, the former is annotated as corerential device and the noun as lexical cohesion (see *the* and *manuscript* in example (2)). In the PDT approach, both implicit and explicit relations of coreference are annotated, including indefinite NPs. In addition, the whole anaphoric expression is annotated as one coreferential element, as illustrated in example (2).

(2) *Twenty years I have been working on [this book]," and he leaned over to rap [[the] [manuscript]] with a thick proprietary finger," and you can sit home in Peterskill and read it when it's published.*

Lexical repetitions (which belong to the level of lexical cohesion in GECCo) are also annotated as coreferent if they refer to the same discourse entity.

We assume that the differences in the annotation of coreference are also related to the contrasts that we observe for **bridging/lexical cohesion**, see Figure 2. Although there is a partial intersection of sets of the relations, the different conceptions are clearly seen in the annotations: in lexical cohesion, lexico-semantic properties of mentions in text are important. The semantic relations (e.g., meronymy, hyponymy, synonymy, etc.) assigned to the mentions are based on the context-free sense relations into which lexical words or patterns can enter, whereas their contextual meaning and referential properties are neglected. By contrast, bridging relations are based on the information instantiated in the text, which means that only those conceptual relations are considered which hold between entities mentioned in the same discourse. Nevertheless, we noticed that relations not marked as lexical cohesion are compensated by the annotation of coreference relations in GECCo, and taken together, they are comparable to the relations of bridging and coreference in the PDT framework. For example, repetitions, which are a subcategory of lexical cohesion, are marked as coreference relations in the PDT framework (see above).

Summing up, there are numerous similarities and overlaps in the categories of textual phenomena in both approaches, despite of the differences discussed earlier. This leads us to conclude that textual phenomena are reflected in both approaches in a very similar way although they are annotated with diverging terminology that stems from different theoretical backgrounds. The following section (4.2) illustrates in more detail some of the cases which are especially interesting for a cross-lingual analysis of spoken and written language.

4.2 Case studies

Coreference and bridging / lexical cohesion The interplay between coreference and bridging or lexical cohesion is especially interesting if we compare spoken and written genres, as we expect certain preferences due to contextual settings (short-time memory, presence of all speech participants in the communication situation, etc.). In Table 2, we demonstrate the statistics (numbers are counted for one journalistic text consisting of 43 sentences) for coreference chains identified with both annotation schemes.

	GECCo-style	**PDT-style**
coref.chains	23	46
aver.chain length1	3,48	4,20
aver.chain length2	6,25	7,05

Table 2: Annotation statistics for coreference chains

We compare the total number of chains and the average chain length[2] which are higher in the PDT framework than in the GECCo approach for German and English. This coincides with the results that we observed in Section 2 above, as the total number of coreference elements is much lower in the GECCo framework.

If we go into detail and analyse the subtypes of anaphora, we find some fine-grained differences in the annotation. For example, event anaphora are annotated in both frameworks. However, the largest

[2] **aver.chain length1** is used for all chains, whereas **aver.chain length2** indicates statistics for chains containing more than two elements.

scope of the antecedent of this anaphora type is limited to the extension of a sentence in the tree-based approach while cohesion-based annotations also include larger textual antecedents.

The above mentioned (see Section 4.1) overlap between coreference and bridging can be illustrated by the example in (3). The relation in (3-a) is covered by a combination of comparative reference and lexical cohesion in the GECCo framework, and by contrastive bridging in the PDT framework. At the same time, comparative reference also includes such cases as (3-b) and (3-c), combined with lexical cohesion in (3-b) and coreference and lexical cohesion in (3-c). Both are cases of bridging anaphora and common textual coreference in the PDT framework.

(3) a. *a presentation – a better presentation, an example – other examples*
b. *some case – such/similar cases.*
c. *one hand – the same hand*

Another illustration of this overlap can be seen in (4), where *she, her children, her war-damaged husband* and *their* are marked as a bridging relation (type subset - set) in one approach, whereas *she, her, her* and *their* are annotated as coreference in the other, *their* with a split antecedent.

(4) *Although [she] was kind and playful to [her] children, she was dreadful to [her war-damaged husband]; she openly brought her lover into [their] home.*

The relation between *The World War II* and *that* in (5) shows how coreference signaled by a demonstrative pronoun in the GECCo approach may coincide with the bridging relation in the PDT approach. In the latter, an explicit anaphor is marked as signalling a bridging and not a coreference relation since it is not entirely clear whether the event (*The World War II* in (5)) is identical with *that time*.

(5) *[The World War II] remained one of the most tragic events in the history. But at [[that] time] nobody thought about it.*

A minor difference between the approaches can be found within the field of event anaphora annotation. In the PDT approach, an antecedent can be explicitly annotated only when it is not longer than one sentence. In the GECCo approach, the scope of the antecedent is annotated independently of the size of the antecedent.

Discourse relations As already mentioned above, the greatest similarities between the two approaches were observed in terms of the total number of identified discourse relations in both schemes. The differences are discovered here on the level of types of relations involved. For example, the connective *and* in (6) is assigned a reason-result relation in the PDT framework, while the GECCo framework considers it as an additive conjunction.

(6) *William Gates and Paul Allen in 1975 developed an early language-housekeeper system for PCs, [and] Gates became an industry billionaire six years after IBM adapted one of these versions in 1981.*

In Table 3, we demonstrate the number of relations identified per approach and per text genre, as we suppose that the detected differences can be genre-sensitive.

	GECCo-style		**PDT-style**	
	journ.	**fict.**	**journ.**	**fict.**
temporal	6	11	5	5
contin./caus.	9	6	19	4
comp./adver.	16	10	15	17
expan./addit.	22	24	19	22
modal	7	4	-	-

Table 3: Annotation statistics for discourse connectives

For instance, both frameworks identify approximately the same number of temporal relations in the journalistic texts. Yet, deviating numbers for this relation are obtained for the fictional texts. The same tendency is observed for relations of contrast (adversative). In case of contingency or causal relations, the situation is different: the number of relations coincide here for fiction rather than journalism.

5 Resulting Scheme and Discussion

Summarising all the cases analysed in the data that were annotated with both frameworks, we create an intersection scheme, covering all overlapping categories. This scheme is illustrated in Table 4. The main categories here are labelled as

IDENTITY, NON-IDENTITY, ELLIPSIS and DISCOURSE RELATIONS. These general categories also include subclasses on a more fine-grained level, e.g. METONYMY or CONTRAST, which can be derived from the existing annotation. For the time being, we exclude the categories without correspondence, i.e. which exist in one approach but not in the other.

As can be seen from the table, the annotation schemes based on both frameworks can be merged even though there are differences in the terminology used for specific features, in the level of granularity and in the method of annotation.

However, without the categories we had to exclude because there was no correspondence between the two approaches, we cannot cover all the cases of textual phenomena. For instance, modal discourse markers, which are especially important for spoken genres cannot be captured by our interoperable scheme for the time being.

One of the main reasons for the incompatibility of the excluded categories lies in the nature of the phenomenon itself: the GECCo approach takes a linguistic signal into account, while the PDT framework includes a more abstract level of coherence. This is especially reflected in the relations of IDENTITY which are not marked by a referring item, e.g. definite article, pronoun, etc. In turn, the GECCo framework captures more semantic relations, e.g. hyponomy, synonymy, etc. that are purely based on sense relations and not on relations between instantiated referents, thus allowing a more fine-grained view on the thematic progression in a text, see Figure 4.

As already stated above, the conceptual dissimilarities discovered in this study seem to result, at least partially, from the systemic differences between Germanic and Slavic languages with respect to the language devices available for expressing textual phenomena. For instance, English uses a very closed class of explicit markers for establishing a relation of comparison, labeled as substitution (*the shirt – the red one*). German is more heterogeneous with respect to the linguistic items available, while Czech has no corresponding structures and makes use of ellipsis instead. We expect that these differences will be even more apparent when integrating the analysis of non-canonical spoken varieties into our trilingual study.

Our future work will include the application of the resulting scheme to our contrastive analysis of naturally occurring texts of English, German and Czech. We are particularly interested in comparing the textual phenomena realized in texts with plain written style with those occurring in non-canonical texts that are produced spontaneously, with a high degree of interaction between varying numbers of speech participants, such as talkshows or private conversation. Moreover, we intend to investigate language production in between spoken and written, such as forums, blogs or interviews. We expect that the most significant differences between languages and genres are tied to varying contextual configurations of mode, e.g. number of speech participants, private vs. public conversation, time laps between production and reception). They may be reflected in textual phenomena with respect to their overall number, the degree of explicitness, as well as the type of textual categories that are preferred. Moreover, we intend to examine variation in the degree of dependence of these textual phenomena on lexicogrammatical constraints or pragmatic peculiarities. The scheme developed in this paper is a first step towards unifying different frameworks that result from separate analyses of Germanic languages and a Slavic language. It therefore reflects a level of generalisation that is applicable to trilingual analysis, which will, however, be broken into more delicate subcategories to permit an identification of fine-grained contrasts.

6 Acknowledgement

This work was made possible by a grant on Short Term Scientific Missions received within the Textlink Action (ISCH COST Action IS1312)[3]. We also acknowledge support from the Grant Agency of the Czech Republic (grant P406/12/0658). This work has been using language resources developed and stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013). The project GECCo has been supported through a grant from the Deutsche Forschungsgemeinschaft (German Research Society).

[3] http://textlinkcost.wix.com/textlink

	Czech framework	**German-English framework**
IDENTITY	coreference with pronouns	coreference with pers. and demo. heads (except extended reference)
	pronouns with arrows to segments and events	extended reference
	NP coreference	coreference with pers./ dem. modifiers or def.art.+hyperonymy/ repetition/ synonymy
	coreference of NEs	repetitions of named entities
	coreference with the word same	comp.reference with the word same
	coreference with demonstrative local and temporal adverbs (tam, tehdy)	coreference with demonstrative local and temporal adverbs
NON-IDENTITY	contextual relations of MERONYMY between lexical items	contextual relations of MERONYMY between lexical items
	bridging CONTRAST with comparative adjective	comparative reference excluding cases with the word same
	bridging CONTRAST without comparative adjective	antonyms in lex.coh
DISCOURSE RELATIONS	temporal	temporal
	contingency	causal
	comparison (contrast)	adversative
	expansion	additive
ELLIPSIS	textual ellipsis (nominal, verbal, clausal)	cohesive ellipsis (nominal, verbal, clausal)

Table 4: Categories for the language- and genre-insensitive scheme

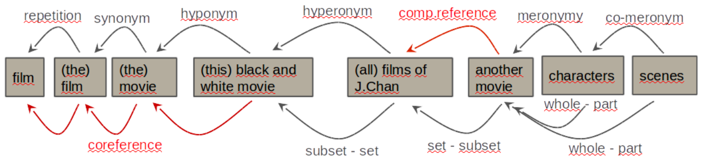

Figure 4: Coreferential and lexical relations in both approaches

References

Robert-Alain de Beaugrande and Wolfgang Ulrich Dressler. 1981. *Einführung in die Textlinguistik*. Niemeyer, Tübingen.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. Prague dependency treebank 3.0.

Herbert H. Clark. 1975. Bridging. In *Theoretical Issues in Natural Language Processing*, pages 169–174.

Václav Cvrček, Vileém Kodýtek, Marie Kopivová, Dominika Kováříková, Petr Sgall, Michal Šulc, Jan Táborský, Jan Volín, and Martina Waclawičová. 2010. *Mluvnice současné češtiny/Grammar of Contemporary Czech/*. Karolinum, Prague.

Jan Hajič, Petr Pajas, David Mareček, Marie Mikulová, Zdeňka Urešová, and Petr Podveský. 2009. Prague dependency treebank of spoken language (PDTSL) 0.5.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association.

M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London, New York.

Bohuslav Havránek and Miloš Weingart. 1932. *Spisovná čeština a jazyková kultura / Standard Czech and language culture*. Melantrich, Prague.

Kerstin Kunz, Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunski, Katrin Menzel, and Erich

Steiner. forthcoming. Gecco – an empirically-based comparison of english-german cohesion. In G. De Sutter, I. Delaere, and M.-A. Lefer, editors, *New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies*. Mouton de Gruyter. TILSM series.

Ekaterina Lapshinova-Koltunski and Kerstin Kunz. 2014. Annotating cohesion for multillingual analysis. In *Proceedings of the 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, Reykjavik, Iceland, May. LREC.

Ekaterina Lapshinova-Koltunski, Kerstin Kunz, and Marilisa Amoia. 2012. Compiling a multilingual spoken corpus. In Tommaso Raso Heliana Mello, Massimo Pettorino, editor, *Proceedings of the VIIth GSCP International Conference: Speech and corpora*, pages 79–84, Firenze. Firenze University Press.

Christian Mair. 2006. *Twentieth-Century English: History, Variation and Standardization*. Cambridge University Press, Cambridge.

Marie Mikulová. 2014. Semantic representation of ellipsis in the prague dependency treebanks. In *Proceedings of the Twenty-Sixth Conference on Computational Linguistics and Speech Processing ROCLING XXVI (2014)*, pages 125–138, Taipei, Taiwan. Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Petr Pajas and Jan Štěpánek. 2008. Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 673–680, Manchester, UK. Coling-2008 Organizing Committee.

Lucie Poláková, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. 2013. Introducing the prague discourse treebank 1.0. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, Japan. Asian Federation of Natural Language Processing, Asian Federation of Natural Language Processing.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2961–2968, Marrakech, Morocco. European Language Resources Association.

Petr Sgall, Eva Hajicova, and Jarmila Panevovǎ. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Reidel, Dordrecht.