From Interoperable Annotations towards Interoperable Resources: A Multilingual Approach to the Analysis of Discourse

Ekaterina Lapshinova-Koltunski*, Kerstin Anna Kunz, Anna Nedoluzhko*****

Saarland University*, University of Heidelberg**, Charles University in Prague, Faculty of Mathematics and Physics***

e.lapshinova@mx.uni-saarland.de, kerstin.kunz@iued.uni-heidelberg.de, nedoluzko@ufal.mff.cuni.cz

Abstract

In the present paper, we analyse variation of discourse phenomena in two typologically different languages, i.e. in German and Czech. The novelty of our approach lies in the nature of the resources we are using. Advantage is taken of existing resources, which are, however, annotated on the basis of two different frameworks. We use an interoperable scheme unifying discourse phenomena in both frameworks into more abstract categories and considering only those phenomena that have a direct match in German and Czech. The discourse properties we focus on are relations of identity, semantic similarity, ellipsis and discourse relations. Our study shows that the application of interoperable schemes allows an exploitation of discourse-related phenomena analysed in different projects and on the basis of different frameworks. As corpus compilation and annotation is a time-consuming task, positive results of this experiment open up new paths for contrastive linguistics, translation studies and NLP, including machine translation.

Keywords: interoperability, linguistic annotation, multilingual resources, discourse, coreference, German-Czech contrasts

1. Introduction

This paper aims at a cross-lingual analysis of discourse phenomena in the two typologically different languages – German and Czech. The discourse properties in focus are relations of identity and non-identity (semantic similarity) of discourse entities, ellipsis and discourse relations (types of conjunctions). Information on differences between the two languages in terms of discourse structuring devices is beneficial to contrastive linguistics, translation studies and multilingual natural language processing.

The novelty of our analysis lies in the nature of the resources used. Quantitative contrastive analyses on the level of discourse require annotated corpora involving time-consuming compilation and annotation, especially in a multilingual setting. Therefore, we take advantage of the existing resources reflecting systemic peculiarities and realisational options of the languages under analysis. We use Czech and German data annotated on the basis of two different frameworks: Functional Generative Description, see (Sgall et al., 1986), for Czech, and textual cohesion, see (Halliday and Hasan, 1976), for German. In our previous work, see (Lapshinova et al., 2015), we have shown that annotations of the involved resources are comparable if abstract categories are used and only the phenomena with a direct match in the frameworks for German and Czech are taken into consideration. We have also shown that although being not general enough to permit a comparison across Germanic and Slavic languages, the existing annotated resources capture the same phenomena, and the creation of an interoperable scheme is possible if more abstract categories are taken into consideration. Hence, we make use of this scheme to perform a comparison between German and Czech.

Our analysis is a first step towards unifying separate analyses of discourse relations in Germanic and Slavic languages. At the same time, it demonstrates that the application of 'theoretically' different resources is possible in one contrastive analysis. This is especially valuable for NLP, which uses annotated resources to train language models for various tools. Training of language models with more complex linguistic annotation often requires manually annotated corpora, which is time consuming and costly, especially if more than one language is involved. Therefore, development of interoperable schemes that enable usage of the existing annotated resources is very important.

2. Related Work

Generally speaking, Slavic languages have a richer, more fusional morphology than Germanic languages. Even though German has conserved more of the inflectional morphology of Proto-Indo-European than other Germanic languages such as English, it has a more isolating character than Czech. The morphological reduction in German partially results in a less flexible constituent word order as compared to Czech, although more positional options are possible than, e.g., in English. We expect these contrasts to have an effect on the creation of discourse properties (see interpretations below).

There is a vast number of theoretical studies comparing Germanic and Slavic languages on a rather general level, such as Štícha (2003) and Engel (1999). Apart from these general comparative studies, a special focus on anaphoric relations between Czech and German was done by Komárek (1994). Yet, quantitative comparisons of Germanic and Slavic languages are very rare. The only works, known to us, include the comparison of English and Czech by Novák and Nedoluzhko (2015) and the comparison of English, Czech and Russian by Nedoluzhko et al. (2015). There are almost no corpus-based approaches to the comparison of the language-pair German and Czech, especially if the properties of discourse are concerned. A number of corpus-based analyses exists for different Germanic languages, e.g. the one for particular cohesive conjunctions or adverbs in prepared speeches by Bührig and House (2004) or that for abstract anaphors in parliament debates by Zinsmeister et al. (2012). Other corpus-based studies compare Romance languages, e.g. (Taboada and Gómez-González,

2012) for particular coherence relations. In addition, there are various studies related to human and machine translation, see for instance, (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012), or (Webber et al., 2013) and and (Webber et al., 2015).

Although the analysis of human and machine translation is beyond the scope of this work, we do not exclude the application of our findings for these research areas.

3. Methodology

3.1. Data

For our analysis, several texts of written discourse (essays) with comparable topics on economic, political and social issues were selected. For the German data, nine texts were excerpted from the corpus GECCo, comprising 14930 tokens and 736 sentences in total, see Table 1. The whole corpus represents a continuum of different text types including written discourse, described in (Hansen-Schirra et al., 2012) and spoken discourse described in (Lapshinova-Koltunski et al., 2012). The corpus is annotated on several levels, which include morphological, syntactical, structural and textual information. 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. Textual information is represented in form of cohesive devices, such as coreference, conjunction, substitution, ellipsis and lexical cohesion. The annotated structures contain information about morpho-syntactic features of devices (including antecedents) and allow yielding information on the chain features, i.e. number of elements in chains, distance between chain elements, etc. Annotation of textual coreference contains not only relations of identity between entities but also abstract and situation anaphora. Therefore, we may coreference to nominal phrases (NPs) along with coreference to clauses, clause complexes and larger textual chunks, as illustrated in example (1-a) for German and (1-b) for Czech.

(1) a. GO: *Gleichzeitig brauchen wir mindestens eine Verdoppelung des Wohlstands. Wenn wir die Armutsgegenden der Erde anschauen, weiß jeder sofort, dass dies das Mindeste an moralischer Herausforderung ist. [At the same time, we need to double the current level of prosperity. One look at the poor regions throughout the world is enough to make anyone realize that this is the most urgent moral challenge we face].*

b. CZ: *Cizinci podstatně přispěli k německému hospodářskému a kulturnímu vývoji, proč jejich počet naopak ve statistikách nezdůrazňovat a tím veřejně uznat jejich zásluhy o německou hospodářskou a politickou demokracii? [Foreigners have contributed significantly to the German economic and cultural development, so why not to emphasize their number in statistics, and to acknowledge their merit of the German economic and political democracy by this?*

The Czech texts were taken from the Prague Dependency Treebank (PDT 3.0, (Bejček et al., 2013)). They 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 inter-sentential relations: pronominal, zero and nominal coreference, abstract anaphora, bridging relations and discourse relations (including connectives, discourse units linked by them, and semantic relations between these units), see Zikánová et al. (2015) for details.

Since texts are shorter in PDT than in GECCo, 17 texts were excerpted to arrive at a similar number of tokens and sentences (11769 and 763 respectively), see Table 2. Both German and Czech texts under analysis include all levels of annotations (i.e. morphological, syntactical, POS, textual phenomena, etc.) along the corresponding frameworks.

textID	topics	sent	tok
GO1	Germany and social market economy	121	2035
GO2	Optimistic remarks on globalisation	47	971
GO3	Politics and globalisation	103	1871
GO4	Globalisation and new challenges	27	478
GO5	The biggest currency changeover	85	1460
GO6	Globalisation and market economy	80	1782
GO7	Global market and technical progress	108	1851
GO8	Economic and technological changes	73	1795
GO9	Doctors and medical system	92	2687
GO	TOTAL: all texts	736	14930

Table 1: German dataset

textID	topics	sent	tok
CZ1-5	Germany, politics and history	170	687
CZ6	Housing	83	1644
CZ7-8	Technological changes	73	1795
CZ9-12	Politics	121	1854
CZ13-14	Economics	149	2568
CZ15-16	Unemployment	112	2252
CZ17	Television	55	969
CZ	TOTAL: all texts	763	11769

Table 2: Czech dataset

Although these two data sets were annotated within two different frameworks, the data is comparable, see our discussion (Lapshinova et al., 2015).

3.2. Scheme for Analysis

In (Lapshinova et al., 2015), an attempt was made to unify the Czech and the German-English frameworks for the an-

	featID	framework for Czech	framework for German
IDENTITY	id1	coreference with pronouns	coreference with heads (no extended reference)
	id2	pronouns with arrows to segments and events	reference to verb phrases and longer segments
	id3	NP coreference	coreference with modifiers or def.articles
	id4	coreference with the word *same*	general comp.reference
	id5	coreference with local and temporal adverbs	coreference with local and temporal adverbs
NON-IDENTITY	nonid1	relations of MERONYMY	relations of MERONYMY
	nonid2	bridging CONTRAST	particular comparative reference and antonyms
DISCOURSE RELATIONS	temp	temporal	temporal
	cont	contingency	causal
	comp	comparison (contrast)	adversative
	expan	expansion	additive
ELLIPSIS	ellipsis	textual ellipsis	cohesive ellipsis

Table 3: Categories of the interoperable scheme

notation of discourse properties. The creation of an interoperable scheme requires a comparison of the underlying annotations. So, we annotated a small corpus of comparable English texts according to the two separate frameworks. This dataset served us as basis for identifying overlapping annotation categories and creating an interoperable scheme. In the present study, we use this scheme to test whether this can be applied for contrastive analyses of Czech and German, which can be extended to more general comparisons of Germanic and Slavic languages in the future. The whole scheme is illustrated in Table 3. The main categories are labelled as IDENTITY, NON-IDENTITY, ELLIPSIS and DISCOURSE RELATIONS.
The category of IDENTITY, or coreferential relations, are further specified into five groups according to the form of anaphoric expressions:

- Pronominal coreference (**id1**) with pronouns referring to nominal antecedents, e.g. *Ludwig Erhard – er [he]* in example (2).

(2) a. GO: *Als Superstar der sozialen Marktwirtschaft gilt aus gutem Grund Ludwig Erhard. Er hatte.. in den 50er Jahren... die produktiven Kräfte der Unternehmen entfesselt und daraus ein Wirtschaftswunder gezaubert... [Ludwig Erhard is regarded as the superstar of the social market economy, and for good reasons. ...in the nineteen-fifties..., he had unleashed the productive forces of business and in this way conjured up an economic miracle...]*

b. CZ: *Ta přijala strategii Bílého domu v domnění, že je to nejjistější cesta k vítězství. [She endorsed the White House strategy, believing it to be the surest way to victory.*

- Abstract coreference (**id2**) with pronominal anaphors linking up to complex antecedents such as clauses, sentences and longer stretches of text, see example (1) above.

- Nominal coreference, where anaphors are realised in text by nouns with (in German, see *Gewerkschaften – den Gewrkschaften* in example (3-a)) or without a modifier, as in Czech, see Prahu – Prahy in example (3-b) (**id3**).

(3) a. GO: *Staatstragender können Gewerkschaften kaum sein. Auch wenn... Ludwig Erhard von den Gewerkschaften nicht viel hielt... [Greater loyalty to the state can hardly be expected of a trade union. Despite the fact that... Ludwig Erhard did not think much of the trade unions...]*

b. CZ: *Zaím se posunuje stále více za Prahu... Po dálnici bychom se měli svézt z Prahy až do Českých Budějovic... [So far, people are moving away from Prague... [Highways should take us from Prague all the way to České Budějovice...*

- coreference with anaphors including the word *same* (**id4**), see (4), and

(4) *And then we do this process again. It's really exactly the same process every time.*

- coreference with local and temporal adverbs as anaphors (**id5**), e.g. *Lissabon – there.*

The NON-IDENTITY category includes the relations of MERONYMY (**nonid1**) and CONTRAST (**nonid2**) as these categories correspond in both frameworks. Meronymy relations are generally taken part-whole relations between lexical items, such as *Germany – the Ger-*

mans in (5-a), *studio apartments – kitchens* in (5-b) and so on.

(5) a. GO: *...praktisch wird es dazu nicht kommen – dafür ist in Deutschland die Bereitschaft zur Solidarität, der Glaube an das "für alle" zu groß. Eine andere Gefahr ist da weit realer: daß die Deutschen... [In practice... it will not come to that – the readiness to practice solidarity, and people's belief in the "for all" is too pronounced in Germany. Another danger is much more real, however: that the Germans...]*

b. CZ: *Jednotlivá studia v apartmánech jsou vybavena kuchyní, takže je možná individuální příprava stravy. [Studio apartments are equipped with kitchens, so everyone can prepare their own food.]*

CONTRAST covers (again, generally taken) antonymy between nominal groups (such as *Halbierung – Verdoppelung [halving – doubling]* in example (6-a)) and relations termed as comparative reference, e.g. *cars – a smaller car*.

(6) a. GO: *Dazu gehören zum Beispiel die Halbierung der Energie- und Rohstoffintensität bis 2020 gegenüber 1990 (bzw. 1994) und die Verdoppelung des Anteils erneuerbarer Energien am Energieverbrauch bis 2010. [For example, halving the amount of power and raw material consumption by 2020 compared to 1990 (or 1994) levels and doubling the percentage of renewable energy used as part of total energy consumption by 2010.]*

b. CZ: *Saldo běžného účtu platební bilance podle odhadu dosáhlo vloni cca 600 USD... I když letos a příští rok je nutné počítat se zpomalením růstu vývozu, prognózujeme, že saldo pžesto zůstane kladné. [The balance of the current account deficit is estimated to reach $600 last year ... Although this and the next yeas we expect the slowdown in export growth, we forecast that the deficit will still remain positive.]*

Similarly, we include four subclasses of DISCOURSE RELATIONS, i.e. logico-semantic relations that are signalled by a discourse marker or a conjunction:

- temporal relations (temp), e.g. *als [when]* in (7-a) for German or *potom [then]* in (7-b) for Czech.

  (7) a. GO: *Als in Osteuropa der Kommunismus stürzte, hätten viele, die dabei mittaten, gerne etwas von ihm gerettet. [When communism collapsed in Eastern Europe, many of the people involved would gladly have kept individual aspects of it].*

  b. CZ: *Posluchač musí přistoupit na pozici, že vše je dovoleno. Potom se pobaví a také pochopí, že drama znázorňuje ztrátu reálné komunikace. [The listener has to accept the fact that everything is permitted. Then he can enjoy himself and also understand that the drama symbolizes the loss of a real-life communication.]*

- relations of contingency or cause (cont), e.g. *deshalb [this is why]* in (8-a) for German or *proto [therefore]* in (8-b) for Czech.

  (8) a. GO: *Aber nur in den wenigsten ist diese Organisation ein dynamisches Element der Volkswirtschaft. Deshalb irritiert ausländische Beobachter auch oft... [This is why foreign observers are often confused...]*

  b. CZ: *Zatímco většina fotbalových reprezentací vstupuje do kvalifikace pro ME 1996 nyní v září, boj o účast v Anglii vypukl již dříve. (...) Před opravdovým rozjezdem kvalifikace proto přinášíme přehled, jak často spolu celky v jednotlivých skupinách už v soutěžích ME a MS v minulosti hrály. [While most national football teams enter the qualification for the 1996 European Championship now, in September, the fight for a place at the competition in England started earlier.Before the real start of the qualification, we therefore provide an overview of how often the teams in each group had played each other at European and World Championships in the past.]*

- relations of contrast (comp), e.g. *aber* in (9-a) for German, and *však [however]* in (9-b) for Czech.

  (9) a. GO: *Arbeiten wie die Polen, aber leben wie die Japaner... [Work like the Poles, but live like the Japanese...]*

  b. CZ: *Poslední statistické sčítání dopravy proběhlo v roce 1990. Za poslední tři roky se však na českých silnicích zvýšil provoz. [The latest statistical traffic census took place in 1990. Over the past three years, however, traffic on Czech roads has increased.*

- relations of expansion or addition (expan), such as *ebenso* in example (10-a) in German or a [and] in (10-b) in Czech.

  (10) a. GO: *Tendenziell ist der Anteil der deutschen Pharmabranche an den globalen Forschungsausgaben der Branche, ebenso wie der Anteil an der Zahl neuer Wirkstoffe, aber rückläufig. [Even so, its share of global expenditure on pharmaceutical research, as well as its share of new active-substance discoveries, is declining.]*

b. CZ: *Vládní plán je podle Jana Švarce ambiciózní a počítá v této oblasti s investicemi 85 miliard korun do roku 2005. [According to Jan Švarc, the government plan is ambitious and it envisages in this area the investment of 85 billion crowns in 2005.]*

Note that all kinds of structural types are analysed, such as connectives of main clauses, subordinators and also adverbials.

NOMINAL ELLIPSIS includes only nominal constructions as this type is available in both frameworks. We demonstrate an example of a nominal elliptical construction in example (11-a).

(11) a. GO: *All das ist eine kleine Revolution. Die grössere [] ist diese: [But there is also a bigger [revolution], and it is this:]*

b. CZ: *Klienti pojišťoven, které ukončí svou činnost, se automaticky vrátí k Všeobecné []. [Clients of insurance companies which shut down will automatically return to the General [one].*

4. Analyses and Results

We now analyse the categories in both languages with respect to their overall distribution, the degree of explicitness, as well as the type of textual categories that are preferred. Moreover, we examine variation in the degree of dependence of these textual phenomena on lexico-grammatical constraints or pragmatic peculiarities.

First, we compare the distributional characteristics of the German and Czech data. We produce box plots for analysing variance and significant differences between both data sets. Box plots are median-oriented graphics that represent a convenient way of depicting groups of numerical data through their quartiles which are the three points that divide the data set into four equal groups, each group comprising a quarter of the data. Box plots have lines extending vertically from the boxes (*whiskers*), indicating variability outside the upper and lower quartiles. *Notched* boxplots reveal if the differences between the variables under analysis are significant: if two boxes' notches overlap, then there is no 'strong evidence' that their medians differ, see Chambers et al. (1983). The boxplots in Figure 1 demonstrate that German (GO) and Czech (CZ) texts do not differ significantly in their overall degree of cohesiveness if all four categories are taken together.

The differences get pronounced if we compare the distributions for each category, see the barplot based on the normalised (per 10000) overall frequencies per relation in Figure 2. So, we observe more variation for identity and discourse relations, while the frequency distribution for ellipsis and non-identity is similarly low.

Taking a closer look into the subcategories in Table 4, illustrating overall frequencies per category (normalised per total number of words in texts), we find that the higher frequencies of IDENTITY relations in Czech exclusively stem from id3, as numbers are higher in German in all other identity types. Qualitative analyses show that more coreference relations are underspecified in Czech than German in terms of explicit accessibility markers, since the definite article does not exist in Czech and accessibility of referents is indicated by information structure more often than in German.

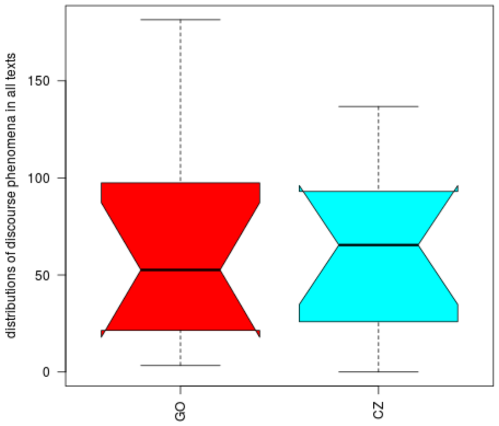

Figure 1: Discourse phenomena in German and Czech

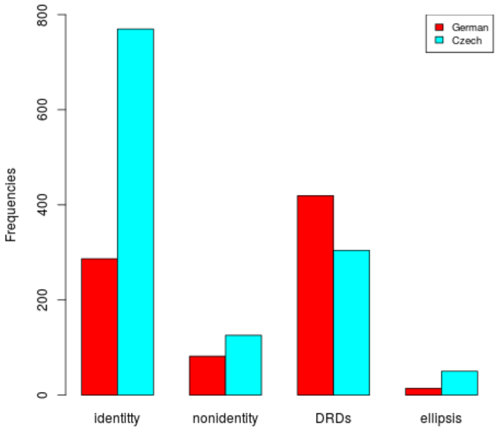

Figure 2: Discourse phenomena in German and Czech

On the one hand, the difference in the amount of identity relations is due to the discrepancies in the annotation framework: repetitions of named entities are annotated in the German framework within lexical cohesion. The other repetitions, if coreferent, are included into the annotation of identity relations, since they will be almost always modified either with a demonstrative pronoun or a definite article. Qualitative analysis of the chains in the data shows that, for instance, in the text containing the chain consisting of Gewerkschaften – die Gewerkschaften, part of which was illustrated in example (3) above, later on, there is another mention of Gewerkschaften without an article or a demonstrative, which is used in a general meaning (*Gewerkschaften gibt es in vielen Ländern [There are*

featID	German	Czech
id1	88.41	97.71
id2	38.18	64.58
id3	144.68	597.33
id4	3.35	0.00
id5	12.06	10.20
nonid1	52.91	88.37
nonid2	28.80	37.39
temp	106.50	14.44
cont	52.24	66.28
comp	79.04	86.67
expan	181.51	136.80
ellipsis	14.07	50.13

Table 4: Frequencies of discourse categories

trade unions in many countries]). This would be a part of the same lexical chain as the other mentions of Gewerkschaften, but is not coreferent and, cannot be considered as an extension for the German coreference chain here. In languages with the definite article, anaphoric expressions mostly contain a formal definite marker which allows to (even automatically) extract most anaphors from the corpus. Czech, as a Slavic language without definite article, does not dispose a formal means with the help of which anaphoric expressions can be easily found and annotated. Thus, the annotation is completed on the base of semantic and referential criteria: everything that refers to the same discourse entity, according to the annotator, is marked as coreferential.

By contrast, the frequencies for DISCOURSE RELATIONS are higher in German than in Czech. As a similar tendency was observed in comparison to English (see e.g. Kunz et al. (in press)), German seems to be exceptional in signaling logico-semantic relations by an explicit discourse marker, especially in terms of temporal relations or relations of expansion, as it is seen in Figure 3.

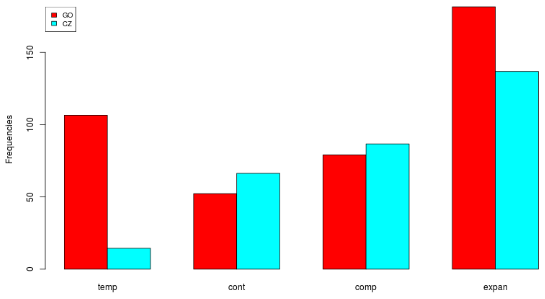

Figure 3: Discourse relations of in German and Czech

As for NON-IDENTITY, we foresee much higher frequencies when integrating further relations in the future. Finally, the higher number of NOMINAL ELLIPSIS in Czech than German points to a higher preference for expressing comparison by fragments. This tendency towards implicitness may, however, stem from the greater syntactic flexibility of Czech relative to German.

5. Conclusion and Discussion

We have performed a cross-lingual analysis of discourse phenomena, using resources annotated along two different frameworks. Our preliminary results show that interoperable schemes like the one used here permit a multilingual analysis of discourse-annotated corpora originating from different approaches. On the one hand, we are able to validate the interoperable scheme in an application. On the other hand, the successful application of the scheme indicates possible interoperability in existing resources. In this way, our methodology saves time and effort as no compilation of additional resources is required. This is especially valuable for multilingual NLP which usually requires multilingual data sets annotated according to the same scheme to build appropriate language models. Creation of such data sets is costly and time-consuming, and our approach can be a good solution in this case. Furthermore, the results yield first insights into differences between German and Czech in terms of the annotated phenomena. At the same time, we are aware of the limitations the dataset at hand provides: although the texts are from the same text genres and have similar topics, the variation observed may be author- or source dependent, since the size of the dataset is small. Some of the differences could also be explained by the differences in the conceptualisations in the schemes. Nevertheless, our work is an important first step towards better comparing and harmonising available resources that are already enriched with annotations. Our future plans include an expansion of the analyses in terms of corpus size, languages and factors influencing variation, e.g. authors, topics. A deeper analysis of textual examples in German and Czech will help us to improve and to refine the analysed categories. Moreover, we plan to include spoken data into our analyses, and compare the distribution of discourse relations across spoken and written dimensions in both languages. Additionally, we intend to test this scheme in further applications, e.g. for machine translation or other NLP areas.

Acknowledgements

This work was made possible within the Textlink Action (ISCH COST Action IS1312)[1]. We also acknowledge support from the Grant Agency of the Czech Republic (grant 16-05394S). 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 LM2015071). The project GECCo has been supported through a grant from the Deutsche Forschungsgemeinschaft (German Research Society). The German corpus can be queried via the CQPWeb interface provided through CLARIN-D (`hdl: 11858/00-246C-0000-0023-8CF7-A`).

6. Bibliographical References

Bührig, K. and House, J. (2004). Connectivity in translation: Transitions from orality to literacy. In J. House

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

et al., editors, *Multilingual Communication*, pages 87–114. Benjamins, Amsterdam.

Chambers, J., Cleveland, W., Kleiner, B., and Tukey, P. (1983). *Graphical Methods for Data Analysis*. Wadsworth.

Engel, U. (1999). *Deutsch-polnische kontrastive Grammatik*. Groos, Heidelberg.

Guillou, L. (2012). Improving pronoun translation for statistical machine translation. In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, pages 1–10.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman, London, New York.

Hardmeier, C. and Federico, M. (2010). Modelling pronominal anaphora in statistical machine translation. In *IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010.*, pages 283–289.

Komárek, M. (1994). On relative pronouns in czech and german. In Světla Čmejrková et al., editors, *The Syntax of Sentence and Text*. Jonh Benjamins.

Kunz, K., Degaetano-Ortlieb, S., Lapshinova-Koltunski, E., Menzel, K., and Steiner, E. (in press). Gecco – an empirically-based comparison of english-german cohesion. In G. De Sutter, et al., editors, *New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies*. Mouton de Gruyter. TILSM series.

Lapshinova, E., Nedoluzhko, A., and Kunz, K. (2015). cross languages and genres: Creating a universal annotation scheme for textual relations. In Ines Rehbein et al., editors, *Proceedings of the Workshop on Linguistic Annotations, NAACL-2015*, Denver, USA.

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

Le Nagard, R. and Koehn, P. (2010). Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 252–261, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nedoluzhko, A., Toldova, S., and Novák, M. (2015). Coreference chains in czech, english and russian: Preliminary findings. *Computational Linguistics and Intellectual Technologies*, 14(21):474–486.

Novák, M. and Nedoluzhko, A. (2015). Comparison of coreferential expressions in czech and english. *Discours: Revue de linguistique, psycholinguistique et informatique.*, 16:1–24.

Sgall, P., Hajicova, E., and Panevovă, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Reidel, Dordrecht.

Štícha, F. (2003). *Äesko-nÄmeckÃ¡ srovnÃ¡vacÃ gramatika*. Argo, Prague.

Taboada, M. and Gómez-González, M. (2012). Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences*, 6 (1-3):17–41.

Bonnie Webber, et al., editors. (2013). *Proceedings of the Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, August.

Bonnie Webber, et al., editors. (2015). *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, September.

Zikánová, Š., Hajičová, E., Hladká, B., Jínová, P., Mírovský, J., Nedoluzhko, A., Poláková, L., Rysová, K., Rysová, M., and Václ, J. (2015). *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Studies in Computational and Theoretical Linguistics. ÚFAL, Prague, Czech Republic.

Zinsmeister, H., Dipper, S., and Seiss, M. (2012). Abstract pronominal anaphors and label nouns in german and english: selected case studies and quantitative investigations. *Translation: Computation, Corpora, Cognition*, 2(1).

7. Language Resource References

Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, Š. (2013). Prague dependency treebank 3.0.

Hansen-Schirra, S., Neumann, S., and Steiner, E. (2012). *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. de Gruyter, Berlin, New York.

Lapshinova-Koltunski, E., Kunz, K., and Amoia, M. (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.