CONTRASTING COREFERENCE IN CZECH AND GERMAN: FROM DIFFERENT FRAMEWORKS TO JOINT RESULTS

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz), Charles University in Prague, Prague, Czech Republic
Lapshinova-Koltunski E. (e.lapshinova@mx.uni-saarland.de), Saarland University, Saarbrücken, Germany

Abstract

In this paper, we analyse coreference patterns in Czech and German. We specifically focus on different types of coreference chains and their properties, for instance, their length, number and functional subtypes of the elements inside these chains. We use two datasets annotated within different annotation frameworks, showing that this approach is possible if an interoperable analysis scheme is applied.

Key words: coreference, anaphora, antecedent, abstract anaphora, Czech, German

КОРЕФЕРЕНТНОСТЬ В ЧЕШСКОМ И НЕМЕЦКОМ ЯЗЫКАХ: ОТ РАЗЛИЧЫХ ПОДХОДОВ К ОБЩИМ РЕЗУЛЬТАТАМ

Недолужко А. (nedoluzko@ufal.mff.cuni.cz), Карлов университет в Праге, Прага, Чехия
Лапшинова-Колтунски Е. (e.lapshinova@mx.uni-saarland.de), Саарбрюкенский университет, Саарбрюкен, Германия

Ключевые слова: кореферентность, анафора, антецедент, абстрактная анафора, чешский язык, немецкий язык

1. Aims and Motivation

The main aim of this study is to analyse cross-lingual variation in the features of coreference chains, which are supposed to reflect variation on coherence. Such features as length and number of coreference chains provide us with the information on how certain contents are distributed in a discourse, how strong the relation between various elements in coreference chains is.

Our previous analyses have shown that there are cross-lingual differences in the features of coreference chains in Czech, English and Russian, see Nedoluzhko et al. (2015). This was a pilot study where we performed analyses on several texts revealing some regularities concerning the structure of coreference chains. This revealed a series of new research questions that should be thereupon addressed to. However, our research was based on the texts of different nature: translations from English into Czech and texts originally authored in English and Russian. This mixture of partly comparable and parallel texts appeared to be quite problematic when interpreting the results for English-Russian-Czech contrasts, since the phenomena originating from translation process see studies of translationese, e.g. Baker (1995), have an impact on the outcome. For this reason, we address coreference relations in a set of comparable (non-translated) texts in Czech and German in this study.

More generally speaking, we aim to analyse language contrasts in the context of textual coherence. Our intention here is to see if there are any differences between German and Czech with respect to coherence-related phenomena. These phenomena are measured by different properties of coreference chains, e.g. their length, syntactic function of their members and others, see Section 3 below. Our notion of coreference includes reference not only to entities expressed with nominal phrases (NPs) or pronouns, but also to larger texts and discontinuous strings or clauses and sentences, like the one demonstrated in example (1). The inclusion of these different coreference types is justified by various reasons. On the one hand, we aim at a comprehensive analysis of all types of discourse relations expressed through coreference. On the other hand, both entity and abstract coreference are expressed with similar language means (pronouns and nominal phrases). And finally, the annotation of both coreference types are available in the resources at hand, which enables a multilingual analysis of these phenomena.

(1) *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]*.

The differences in the properties of coreference chains are expected, because Czech as a Slavic language has a richer, more fusional morphology than German (a Germanic language). Even though German has conserved more of the inflectional morphology of Proto-Indo-European than other Germanic languages, 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 some positional options are possible. We expect these contrasts to have an effect on the creation of referring expressions. There is a vast number of

theoretical studies comparing Germanic and Slavic languages on a general level and in anaphoric relations, whereas quantitative comparisons are rare.

The remainder of this paper is organised as follows: Section 2 presents the related work on coreference, in Section 3 we present the methods and the data used in our analyses described in Section 4. Finally, in Section 5, we pose a number of arising questions and provide an outlook for our future work.

2. Related work

In theoretical linguistics, the analysis of coreferential chains most closely relates to referent activation theories (see e.g. Givon, 1983; Ariel, 2001; Kibrik, 2011; Kibrik, 1997, etc.). These studies suggest the model of referential choice (the choice of a particular NP type) based on the degree of referent salience. Some studies analyse the predictability of upcoming referents in relation to the choice of coreferring expressions and its status in the information structure of an utterance (see the algorithm, determining the degree of salience in Hajičová et al., 2006; Lambrecht, 1994; Strube and Hahn, 1999, etc.).

In corpus-based approaches, there is a large amount of annotated data for coreference, anaphoric relations, event anaphora (or discourse deixis, reference to events), bridging relations (associative anaphora) and so on. However, as far as we know, there is a very little number of studies that analyse the structure of coreferential chains as a whole. Coreference chains have been annotated in a number of frameworks (including Zikánová et al. 2015 and Lapshinova-Koltunski and Kunz, 2014). However, they mostly address this phenomenon in one language: an extensive description, especially in a multilingual perspective, is missing.

Most of the analyses on coreference chains concentrate on entity anaphora only and do not cover coreference to events, states and situation. This type of coreference is often analysed within separate studies on non-entity anaphora[1]. For instance, Botley (2006) distinguishes three main types of abstract anaphora: "label" anaphora, which encapsulates stretches of text (following Francis (1994)); "situation" anaphora and (iii) "text deixis". Following Fraurud (1992), "situation" anaphora is divided into eventuality and factuality. Hedberg et al. (2007), Navarretta & Olsen (2008), and Dipper & Zinsmeister (2009) present a similar distinction concerning "situation" anaphora subtypes. In the latter work, the authors describe annotation of these subtypes to the abstract anaphors and their antecedents. In the Prague Dependency Treebank (henceforth PDT, textual phenomena annotation described in Zikánová et al. 2015) and the GECCo corpus (Lapshinova-Koltunski and Kunz, 2014), which we use for our analysis, references to events were annotated within the phenomenon and in the same way as textual coreference in case the antecedent does not exceed one sentence. References to larger antecedents were marked with the special label *segm* in PDT, but the antecedent itself was not annotated. The characteristics of non-nominal coreference in these approaches were recently addressed to in Nedoluzhko et al. (2016).

Properties of antecedents of abstract anaphora and extended references have been analysed in few studies only. Most of the works are concerned with the marking span, since they lie within annotation frameworks, e.g. Müller (2007), Pradhan et al. (2007), (Byron, 2003) or Dipper and Zinsmeister (2009).

[1] There is quite a large terminological variation concerning this phenomenon in the literature. References to non-nominal entities can be also referred to as *abstract anaphora* (Zinsmeister et al. 2012), *discourse anaphora* (Dipper et al. 2009), *event anaphora* (Caselli - Prodanof, 2010), *situational deixis* (e.g. Linke et al., 2004), *discourse deixis* (Recasens et al. 2007), etc.

We analyse properties of coreference chains in a multilingual perspective comparing Czech and German. A coreference chain consists of an *antecedent* (the first mention in the chain) and *anaphors* (co-referring expressions). Coreference to abstract entities such as events, states, situations, facts and propositions are referred to as *abstract anaphora*.

3. Methods and resources

For our analysis, several texts of written discourse (essays) with comparable topics on economic, political and social issues have been selected.

For the German data, 8 texts were excerpted from GECCo (Lapshinova-Koltunski and Kunz, 2014), comprising 12243 tokens and 645 sentences in total. The GECCo corpus contains texts of various 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, including 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). The textual information is represented in form of cohesive devices, such as coreference, conjunction, substitution, ellipsis and lexical cohesion. The annotated structures contain information on 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. The annotation of textual coreference contains not only relations of identity between entities but also abstract and situation anaphora. Therefore, we may corefer to nominal phrases (NPs) along with coreference to clauses, clause complexes and larger textual chunks, as illustrated in example (1) above.

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 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. Since texts are shorter in PDT than in GECCo, 15 texts were excerpted to arrive at a similar number of tokens and sentences (11399 and 628 respectively).

Both German and Czech texts under analysis include all levels of annotations (i.e. morphological, syntactical, POS, textual phenomena, etc.) along the corresponding frameworks.

The research questions we address in this analysis include:

1) Are there any language contrasts with respect to textual coherence between German and Czech, if we consider coreference chain features as indicators of coherence phenomena?
2) Are there any differences between German and Czech if abstract coreference is concerned? Since abstract coreference reflects the scopus, we will see if the scopus in textual coherence is bigger in German than in Czech or vice versa.

To answer these questions, we define a number of features (operationalisations) that include properties of chains, antecedents and anaphors under analysis:

1. number of coreference chains (total number of chains in German and Czech texts),
2. length of coreference chains (calculated as an average length of coreference chains per text),
3. longest chain (chains with the greatest number of elements in a text),

4. number of coreference pairs (= number of anaphorically referring expressions),
5. types of anaphors (e.g. pronouns, nominal modifiers, temporal and local adverbs and so on),
6. number of anaphora referring to antecedents other than NP/pronoun,
7. types of anaphora referring to antecedents other than NP/pronoun.

Features 1-4 will help us answer the first research question, and will give some hints on how the topics are construed in the analysed texts. We assume that if we have shorter coreference chains, we have more various topics. Longer chains, and their smaller amount would indicate an opposite phenomenon. We will also pay attention to the means of expressing coreference in both languages, as we assume that systemic differences between the two languages would have an influence on the devices available in the texts. Features 6 and 7 deal with nominal groups referring to non-nominal antecedents and will provide us with the information required for the research question 2. We will compare the frequency of referring to clauses, sentences and larger textual segments in German and Czech and describe the types of referential devices in both languages.

4. Analyses

4.1. General analysis of coreference chains

In the first step, we analyse chain properties in German and Czech texts under analysis. In Table 1, we provide numeric data on features 1 to 6 for both languages.

	German	**Czech**
1. number of coreference chains (nr of antecedents)	225	550
2. length of coreference chains (mean value per text)	2.45	3.2
3. longest chain	11	27
4. shortest chain	2	2
5. number of coreference relations (nr of anaphors)	327 (996 if repetitions included)	1231 (1027 if zero anaphora excluded)

Table 1: Chain properties in German and Czech

As can be seen from the table, Czech texts in our data contain significantly more coreference chains (feature 1) and relations (feature 5) than the German ones. We believe that the reason is the definition of coreference elements underlying annotations available: annotations in the German texts do not include lexical repetitions, whereas Czech texts do. These cohesive means belong to lexical cohesion according to the theory of Halliday and Hasan (1976) which is underlying the framework used in the annotations. Indeed, upon looking at lexical cohesion in the German texts, we have found 669 annotated repetitions, which partly compensates the smaller number of coreferential chains as compared to Czech. Generally, the correlations of lexical cohesion in German and Czech annotation schemes applied to texts under analysis were addressed to in Lapshinova-Koltunski et al. (2015). The comparison of the approaches provided in the study has shown that the category of lexical cohesion, although also present

in the Prague Dependency Treebank in the form of bridging relations, is much more detailed in the scheme applied for the annotation of GECCo, and some types annotated as coreference in PDT are considered as lexical cohesion in GECCo.

We achieve comparable numbers (996 vs. 1027), if repetitions from the category of lexical cohesion annotated in the German data is added, and zero anaphors annotated in the Czech data are excluded. The latter is a coreference means which does not exist in the German language.

The mean value of coreference chains length in Czech is significantly higher than in German. However, if we look at the character of the chain elements, we observe that the longer the chain is, the higher is the probability that it includes named entities, mostly hyperthematic, and these are the cases that are not annotated as coreference in the German annotation scheme (but lexical cohesion). For example, the longest chain of 27 links consists of expressions referring to Slovakia. In the 29-sentences long text, the chain consists of 14 repetitions of the words *Slovensko (Slovakia)* and *Slovenská republika* (*Slovak republic*), and 15 referential adjectives *slovenský (Slovak, Slovakian)* derived from *Slovensko (Slovakia)* and used in nominal groups such as *slovensky parliament* (*the parliament of Slovakia*), *slovenská historie* (*the history of Slovakia*), *slovenští politici* (*the politicians of Slovakia*) and so on. In this text, the chain does not contain a single personal pronoun or ellipsis, however it should rather be considered as an exception.

Another example is a 26-link long coreference chain referring to the Slovak national uprising, an anti-fascist event in the year 1944. Here, the situation is more heterogeneous. Although unique, this is a quite typical event, which is referred to with a full nominal group (*Slovenské národní povstání* [*Slovak national uprising*], 2x), its abbreviation *SNP* (14x), a common noun *povstání* [*uprising*] (6x) and its synonym *puč* [*putsh*] (1x), as well as with an anaphoric zero (3x).

By contrast, longer chains in the German data mostly contain personal and demonstrative pronouns or definite nominal phrases, e.g. *die soziale Marktwirtschaft [the social market economy]*. For instance, the longest chain contains 11 elements, 10 of which is the repetition of this definite phrase (*die soziale Marktwirtschaft*) and only one personal pronoun *sie* (3rd person singular). The relation is built up with the help of the definite article *die* serving as a demonstrative modifier in this coreferring expression. Another example is a chain with 10 elements, which contains a mixture of personal pronouns and definite nominal phrases: *Gewerkschaften [trade unions]* (antecedent) - *die Gewerkschaften [the trade unions*] (3x) - *sie [they]* (6x).

Let us now look at the numbers of chains of different length in German and Czech data under analysis. The most frequent are naturally the shortest two-element chains, making 69% out of all coreference chains in German (156 cases) and 62% in Czech (343 cases). Figure 1 summarises the numbers of coreference chains of the length higher than 2 elements (from 3 to 27) in our data. Naturally, the shorter the chain is, the more frequent it is. In Czech, coreference chains of the length of 3 elements occur 78 times; in German this number is 20. Long chains (longer than 5 for German and longer than 13 for the Czech data) are rather seldom. In the annotated texts, their number varies from 0 to one.

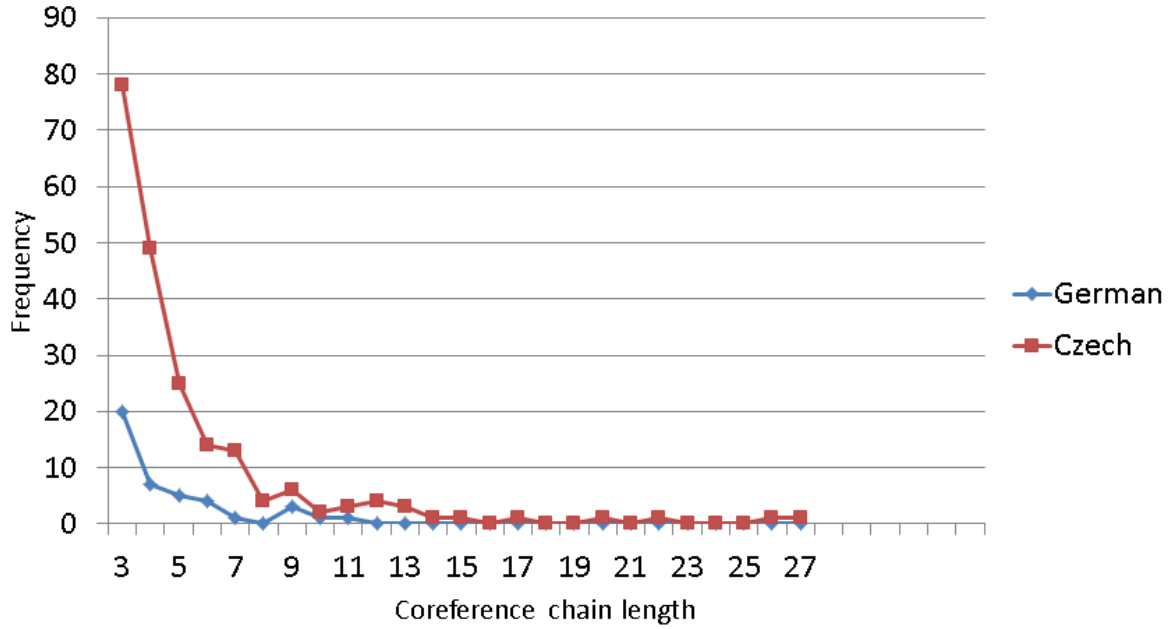

Figure 1: Frequencies of coreference chains of various length in German and Czech

The data in Figure 1 show the tendency of the analysed German texts to shorter chains: while two-element chains are more frequent in German than in Czech texts, we can observe a drastic decrease in frequency for German three-element chains. However, this record primarily reflects the fact that longer coreference chains in the data at hand contain, most probably, repetitions. In the German texts, these are partly captured in the annotation of coreference (see our example on *die soziale Marktwirtschaft* or *Gewerkschaften* above). However, if coreference chains contain repetitions of named entities or other types of relations (e.g. type-entity as *Ludwig Erhard* and *Wirtschaftsminister [minister of economics]* in example (2)), this relations are not captured within coreference in German framework which is based on the concept of cohesion, hence an explicit trigger of a relation.

(2) *Als Superstar der sozialen Marktwirtschaft gilt aus gutem Grund Ludwig Erhard. Er hatte in der Anfangszeit der Bundesrepublik, in den 50er Jahren, als Wirtschaftsminister die produktiven Kräfte der Unternehmen entfesselt und daraus ein Wirtschaftswunder gezaubert... Erhards Philosophie war nicht einfach ein singulärer Geistesblitz [Ludwig Erhard is regarded as the superstar of the social market economy, and for good reasons. As minister of economics in the nineteen-fifties, the early days of the Federal Republic, he had unleashed the productive forces of business and in this way conjured up an economic miracle...For Erhard's philosophy was not just a singular flash of inspiration]*.

Nevertheless, this type of relation is annotated within lexical cohesion chains. The number of the elements in German coreference chains containing also members of lexical chains equals 165, which is approximately a half of the total number of coreference relations (see Table 1 above). This means that part of the chains in the German data at hand might be extended to longer chains, if lexical cohesion is considered. However, qualitative analysis of the chains in the data shows that this does not substantially increase the length of the German coreference chains. For instance, in the text with the chain consisting of *Gewerkschaften - die Gewerkschaften - sie*, mentioned above, later on, there is *Gewerkschaften* which is used in a general meaning (*Gewerkschaften gibt es in vielen Ländern [There are trade unions in many*

countries]) and is a part of the same lexical chain as the other mentionings of Gewerkschaften, but is not coreferent and cannot be considered as an extension for the German coreference chain here.

Another reason for the observed differences in coreference chains in the Czech and German data is the absence of grammatical definiteness in Czech. In languages with a definite article, anaphoric expressions mostly (but not in all cases) contain a formal definite marker (article, definite description, demonstrative modifier, as *die* in *die soziale Marktwirtschaft*) which allows to (even automatically) extract most candidates for anaphoric expressions from the corpus. In most approaches to the coreference annotation for such languages, bare nouns and generics are mostly dismissed, see e.g. Ontonotes (BBN Technologies 2006). However, in our German data, the only cases of bare nouns in coreference chains are repetitions of named entities, e.g. *Ludwig Erhard/Erhard* in example (2) above. Generics, if coreferent, would be marked with a demonstrative modifier (definite article or demonstrative pronoun), and be therefore included into the annotated chains. 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, annotating 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.

Overall, we could see that the properties of coreferential chains in our Czech and German data are comparable (with some exceptions). We therefore believe that variation in the properties of coreference chains is rather genre- or register-dependent. Similarities or differences observed in the number of chains, and the length of chains follow from thematic similarities or differences of the selected texts. We suppose that there are no typological differences in the chain properties, at least in the data at hand. However, we need further analyses of comparable texts belonging to other genres to verify this assumption, which goes beyond the scope of this work.

Yet, we do observe interesting phenomena of contrast in our data: German and Czech seem to show different preferences for the type of the device expressing the same coreferential relations. Therefore, in the next analysis part, we will concentrate on these differences.

4.2 Anaphora types in German and Czech

As can be seen from Figure 2 (frequencies are given normalised against the total number of anaphors per thousand tokens), the main differences between chain members expressing referential relations in the Czech and German data at hand are observed in their structural and functional subtypes, particularly concerning full nominal phrases. In the Czech texts, preferences are given to bare nouns, whereas German operates with nouns modified with a definite article or a demonstrative pronoun. This difference originates from the language system, as Czech does not possess the grammatical category of definiteness, as it was already explained in Section 4.1 above.

Systemic differences also explain the prevalence of Czech in the category of personal heads - this happens due to a great number of zero anaphors in this language, which are also counted in this category. Interestingly, both languages employ a similar number of local and temporal devices in coreference chains, as well as demonstrative heads. However, we observe a difference in the type of antecedent of these anaphors. 72% of all demonstrative heads in German refer to abstract entities, whereas in Czech only 39% do so,[2] see Nedoluzhko & Lapshinova (2016). They are compensated by modified nominal phrases (with a

[2] The reason for such a low number of demonstrative heads in this position is that many Czech demonstrative heads refer to deverbal nouns (*correction*, *murder*, etc.) or syntactic constructions with other formal heads.

demonstrative modifier). The proportion of such anaphors referring to abstract entities is ca. 37% out of modified NPs.

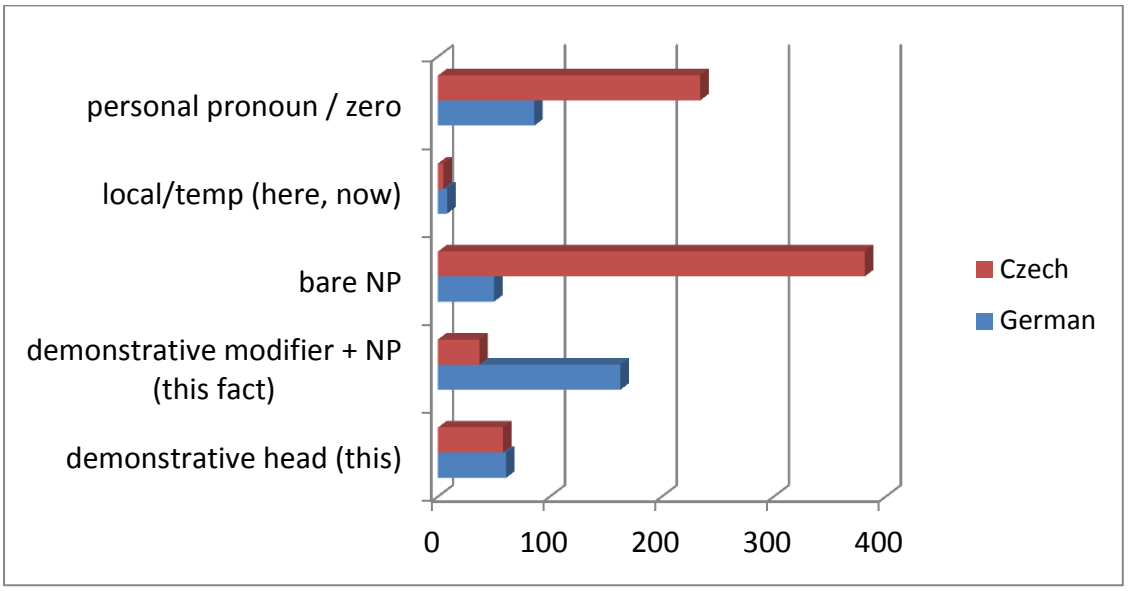

Figure 2: Distribution of anaphors in German and Czech

This is also confirmed by data on various types of anaphoric means that are used in both languages to refer to non-nominal antecedents that we summarise in Table 3.

	German		**Czech**	
	abs.	**in%**	**abs.**	**in%**
demonstrative head (*dies, dazu/ ten [this]*)	44	64.71	28 (22+6)[3]	44.44
demonstrative modifier + NP (*diese Frage/ tato otázka [this question]*)	16	23.53	17 (3+14)	26.98
bare NP	0	0.00	10 (4+6)	15.87
temporal/local (*hier, da, nun/ tam, tady [here, there, now]*)	3	4.41	4 (0+4)	6.35
personal pronoun *(er, sie [he she]*), etc. / zero	3	4.41	2 (2+0)	3.17
gen/part	2	2.94	2 (2+0)	3.17
total	68	100.00	63 (33+30)	100.00

Table 3: Distribution of anaphora types referring to abstract entities in German and Czech

[3] The first number in brackets represents the number of references to clauses/sentences; the second number shows references to larger textual segments.

In the German data, the most occurring cases of abstract anaphors (ca. 66%) refer to segments of one sentence, whereas in the Czech texts, there are more cases of coreferences to longer segments (ca. 48%). On the one hand, these differences have a technical origin. By marking references to longer segments in the data for Czech, annotators did not have to mark the antecedent, which could result in a greater number of abstract anaphors in Czech in general. On the other hand, this could also mean that the authors of texts in Czech summarise larger textual passages more often than those of the German texts. This needs to be tested qualitatively by a closer look at the data at hand.

In Czech, most of the explicitly expressed references to clauses (except one) are realized with a demonstrative pronoun *ten [it/this]*. This is quite expectable, because these are mostly references to clauses within the same sentence, so the antecedent is close to the anaphor and should be neither repeated nor emphasized by other demonstratives, cf. example (3), where *ten [it/this]* refers to the immediately preceding antecedent *proč jejich počet naopak ve statistikách nezdůrazňovat* [*why not to emphasize their number in statistics*]. The remaining sentence is the case of nominalisation (*pokles [decline]*) in example (4), used without a demonstrative pronoun, also because the antecedent clause immediately precedes the anaphoric noun.

(3) *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?*]

(4) *Dnes se tento počet snížil na asi půl milionu, jenže důvodem <poklesu> je především skutečnost, že ten, kdo není zaměstnán déle než rok, již podporu nedostane. [Today, that number dropped to about half a million, but the reason for the decline is the fact that anyone who is not employed for more than a year, gets no support anymore.]*

5. Discussion and future work

In this paper, we analysed cross-lingually variation in coreference chains by considering two languages that are not very close typologically, and by using data sets annotated within two different frameworks. We had to deal with a number of technical issues that result in the differences in frequency data for the features under analysis. However, we could find a solution, achieving comparable results in the end.

Our findings show that the differences of typological character (absence of definiteness or pro-drops) also have influence on the properties of coreference chains in the two languages. However, the main differences in the length and number of chains (reflecting the topic structuring) are not of typological character. It may be rather genre- or even domain-dependent. Our future work will include analysis of further genres and domains for the evidence for this assumption. Moreover, we plan to include more texts to have a better data representation. We assume that if the analysis of a greater number of texts (on different topics and from different genres) has a similar outcome (the same tendencies in the distribution), the observed differences are not due to the properties of specific texts, but rather language- or scheme-specific. Although the identification of scheme-specific properties is not within the scope of the present analysis, it is our overarching goal, since this information is important

for defining interoperability of the existing annotated resources. Creating interoperable annotation schemes is one of the goals of the TextLink COST Action.[4]

We believe that the knowledge of the observed differences, e.g. in the preferences for certain functional or structural types expressing coreference, is important for various areas of linguistics, including contrastive studies, translatology and NLP, i.e. machine translation. For instance, when translating from Czech to German, demonstrative heads should be used for summarisation of sentences or longer text segments instead of full nominal phrases. It would be interesting to have a look at translations from Czech to German to see if we would also see changes in preferences for abstract anaphora in translated German, as was shown by Zinsmeister et al (2012) for translations from English into German. The authors show that although demonstrative heads are more common for the originally authored texts in German, translated German reveals a higher number of personal heads expressed with es, the direct translation of the English it, which is used in English for coreference to abstract entities. Both translation scholars and machine translation developers should be aware of such discrepancies to avoid producing texts which sound less natural for the target language.

6. Acknowledgement

We acknowledge the 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).

References

BBN Technologies. Coreference Guidelines for English OntoNotes – Version 6.0. Linguistic Data Consortium. BBN Pronoun Coreference and Entity Type Corpus. 2006.

Baker, M. (1995), Corpora in translation studies: An overview and some suggestions for future research. Target 7(2). 223–243.

Botley, S. (2006), Indirect anaphora: Testing the limits of corpus-based linguistics, International Journal of Corpus Linguistics, 11(1), pp. 73–112.

Caselli, T., Prodanof I. (2010), Annotating Event Anaphora: A Case Study, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta.

Dipper S., Zinsmeister H. (2009), Annotating discourse anaphora. Proceedings of LAW III, pp. 166–169.

Francis G. (1994), Labelling discourse: an aspect of nominal group lexical cohesion. Coulthard M. (ed.), Advances in Written Text Analysis, London: Routledge, pp. 83–101.

Fraurud K. (1992), Situation reference: What does 'it' refer to? GAP Working Paper No 24, Fachbereich Informatik, Universit¨at Hamburg.

[4] http://textlink.ii.metu.edu.tr/

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

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.

Hedberg N., Jeanette K. G., and Zacharski, R. (2007), Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles, Proceedings of DAARC-2007, pp. 31–36.

Kučová L., Hajičová, E. (2004), Coreferential Relations in the Prague Dependency Treebank. Proceedings of DAARC-2004, pp. 97–102.

Lapshinova-Koltunski E., Nedoluzhko A., Kunz K. A. (2015), Across Languages and Genres: Creating a Universal Annotation Scheme for Textual Relations. Proceedings of the The 9th Linguistic Annotation Workshop (LAW IX 2015) Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 168-177.

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

Lapshinova-Koltunski, E., Kunz, K., and Amoia, M. (2012). Compiling a multilingual spoken corpus. In Mello H., Pettorino M., T. R. (ed.), Proceedings of the VIIth GSCP International Conference: Speech and corpora, Firenze, Firenze University Press, pp. 79–84.

Müller Ch. (2008), Fully Automatic Resolution of ’it’, ’this’, and ’that’ in Unrestricted Multi-Party Dialog. Ph.D. thesis, University of Tübingen.

Navarretta C., Olsen S. (2009), The annotation of pronominal abstract anaphora in Danish texts and dialogues. DAD report 1. Centre for Language Technology, University of Copenhagen.

Nedoluzhko A. (2011), Extended textual coreference and bridging anaphora [Rozšířená textová koreference a asociační anafora], Prague, Czech Republic, 268 pp., Dec 2011

Nedoluzhko A., Toldova S., Novák M. (2015), Coreference chains in Czech, English and Russian: Preliminary findings. Computational Linguistics and Intellectual Technologies, Vol. 14, No. 21, RGGU, Moscow, Russia, ISSN 2221-7932, pp. 474-486.

Nedoluzhko, A., Lapshinova-Koltunski, E. (2016), Abstract Coreference in a Multilingual Perspective: a View on Czech and German. In Proceedings of the CORBON Workshop at NAACL2016, San Diego, California.

Pradhan S., Ramshaw L., Weischedel R., MacBride J., Micciulla L. (2007), Unrestricted coreference: Identifying entities and events in OntoNotes. In Proceedings of the IEEE-ICSC.

Recasens M., Martí, A. (2010), AnCoraCO: Coreferentially annotated corpora for Spanish and Catalan. Language Resources & Evaluation.

Recasens, M., Martí, M.-A., Taulé, A.M. (2007), Text as Scene: Discourse Deixis and Bridging Relations. Procesamiento del Lenguaje Natural, 39.

Zinsmeister, H., Dipper, S., 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).

Zikánová Š., Hajičová E., Hladká B., Jínová P., Mírovský J., Nedoluzhko A., Poláková L., Rysová K., Rysová M., Václ J. (2015), Discourse and Coherence. From the Sentence Structure to Relations in Text. ÚFAL, Prague, Czech Republic, 274 pp.