Abstract Coreference in a Multilingual Perspective: a View on Czech and German

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

Ekaterina Lapshinova-Koltunski
Saarland University
A2.2 University Campus
D-66123, Saarbrücken, Germany
e.lapshinova@mx.uni-saarland.de

Abstract

This paper aims at a cross-lingual analysis of coreference to abstract entities in Czech and German, two languages that are typologically not very close, since they belong to two different language groups – Slavic and Germanic. We will specifically focus on coreference chains to abstract entities, i.e. verbal phrases, clauses, sentences or even longer text passages. To our knowledge, this type of relation is underinvestigated in the current state-of-the-art literature.

1 Introduction

The main aim of this study is to enhance knowledge on abstract coreference in a multilingual perspective. One of the examples of abstract coreference in German is given in (1). Here, the anaphoric pronoun *dies [this]* referes to the whole preceding sentence and not to nominal phrases (NPs) or pronouns, which are analysed in most studies on coreference.

(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].*

Although there exists a number of analyses of such cases (see Section 2), the majority of studies are monolingual or they do not include Germanic and Slavic languages. Information on differences between Czech and German in terms of abstract coreference will be beneficial to contrastive linguistics, translation studies and multilingual natural language processing.

The paper is organised as follows: related work and the definition of the phenomena under analysis are presented in Section 2, data and research questions are detailed in Section 3, followed by the analysis in Section 4. The discussion of the outcome and future work are provided in Section 5.

2 Related Work

There are a number of works on coreference relations other than identity, i.e. concerning references to abstract entities or extended reference. Most of them concentrate on the analysis of abstract anaphora. For instance, Botley (2006) distinguishes three main types of abstract anaphora: "label" anaphora, which encapsulates stretches of text (following Francis (1994)); "situation" anaphora and "text deixis". Following Fraurud (1992), "situation" anaphora is classified into eventuality and factuality. Hedberg et al. (2007), Navarretta & Olsen (2008), and Dipper & Zinsmeister (2009) present a similar distinction concerning "situation" anaphora subtypes. Dipper & Zinsmeister (2009) provide a survey of corpus-based studies on this topic, structur-

Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016), co-located with NAACL 2016, pages 47–52,
San Diego, California, June 16, 2016.

ing them according to the form of anaphoric expressions (demonstratives, personal pronouns, etc.) and antecedents (verbal phrases, clauses, arbitary sequences or larger sequences). Most of these studies take into account only some particular forms of anaphors and antecedents, e.g. Hedberg et al. (2007) and Müller (2008) concentrate exclusively on *it*, *this* and *that*. The analyses of Müller (2008), Kučová & Hajičová (2004) and Pradhan et al. (2007) are limited to coreference to verbal phrases. Viera et al. (2005), Hedberg et al. (2007) and Poesio & Artstein (2008) concentrate on arbitary sequences. Byron (2003), Poesio & Artstein (2008) as well as Navarretta & Olsen (2008) include clauses into their analysis. The only work known to us that provides a description of coreference to various forms of abstract antecedents is (Taulé et al., 2008).

Dipper & Zinsmeister (2009) also describe the languages involved in the studies on abstract coreference. It is obvious that English predominates over other languages. Multilingual approaches are presented in (Vieira et al., 2005), (Navarretta and Olsen, 2008) and (Taulé et al., 2008) only and do not involve the language pair analysed in the present paper.

We analyse properties of abstract anaphora and antecedents from a multilingual perspective comparing Czech and German. Coreference to abstract entities such as events, states, situations, facts and propositions are referred to as *abstract coreference*. These include coreference to (i) verbal phrases as in example (2-a), where *these purposes* refers to *answering of these questions*, (ii) finite clauses and sentences as in example (1) above, in which the German demonstrative pronoun *dies [this]* refers to the whole preceding sentence, and (iii) larger text passages and discontinuous strings as in example (2-b). Here, the modified NP *these goals* refers to the three preceding sentences.

(2) a. *Polling is essential for answering both of these questions. ...the technique most frequently employed for these purposes is the "cross-sectional" survey.*

b. *Germany is seeking to achieve a 40% reduction of greenhouse gases in Germany by 2020, assuming the EU commits to a reduction of 30%. The German renewable energy act sets a new target of 20% electricity from renewables by 2020. Germany's Sustainable Strategy intends to halve overall energy consumption by 2050. The scale of effort needed to meet these goals demonstrates the degree of commitment of both our nations.*

Anaphoric expressions referring to abstract entities in our approach include mostly pronouns, nouns, nominal groups and pronominal adverbs.

3 Data and Research Questions

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 the corpus CroCo (Hansen-Schirra et al., 2012), comprising 12243 tokens and 645 sentences in total. The corpus is annotated on several levels, which include morphological, syntactic, structural and textual information. The information on the latter was annotated with the help of semi-automatic procedures described by Lapshinova-Koltunski & Kunz (2014). Textual information is represented in the form of cohesive devices, such as coreference, connectives, 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 have coreference to nominal phrases (NPs) along with coreference to clauses, clause complexes or sentences as the one 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, textual coreference of different types, bridging relations and PDTB-style discourse rela-

tions (discourse connectives, the discourse units linked by them, and semantic relations between these units), see (Poláková et al., 2013) for details. Since texts are shorter in the Czech data than in the German data, 15 texts were excerpted to arrive at a similar number of tokens and sentences (11399 and 628 respectively).

Although these two data sets were annotated within two different frameworks, the data on the abstract entities are comparable, since they contain information on the structural types of antecedents (if they are clauses, sentences or longer segments), as well as the structural and functional types of referring devices, i.e. demonstratives or other linguistic means. The comparability of the data was proved and discussed in Lapshinova et al. (2015).

4 Analyses

The total number of abstract coreference, i.e. the cases where anaphoric devices point on antecedents other than nominal groups or pronouns, is similar in the analysed Czech and German texts: 63 and 68 respectively. However, the scopus of the segments they refer to demonstrates variation, as seen from Table 1.

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 with longer segments (ca. 48%). On the one hand, these difference may 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.

For the structural types of antecedents, we observe a general tendency of demonstratives to refer to abstract entities in both languages. 72% of all demonstrative heads in our German data refer to abstract entities, whereas 39% do so in the Czech data. They are compensated by modified nominal phrases (with a demonstrative modifier) whose proportion estimates ca. 37% out of all modified NPs in the Czech texts.

Now, we will have a look at various types of anaphoric means that are used in both languages to refer to non-nominal antecedents, see Table 2.

In Czech, most of the explicitly expressed references to clauses (except for one) are realized by 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 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.]*

In some cases (ca. 16%), reference to abstract entities in Czech is expressed by a non-modified NP. In coreference chains in German, these are mostly named entities which never refer to abstract entities.

German shows a clear preference for demonstrative heads (like *dies* in example (1) in Section 1 above) to refer to abstract entities (ca. 65%). Another device within this category is a pronominal adverb, e.g. *dazu, dabei* which represents a com-

	German		Czech	
	abs.	**in %**	**abs.**	**in %**
to clauses	5	7.35	13	20.63
to sentences	45	66.18	20	31.75
to bigger segments	18	26.47	30	47.62
total	68	100.00	63	100.00

Table 1: Number of anaphors referring to the antecedents other than NP and pronoun and their subtypes

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

Table 2: Distribution of anaphora types referring to abstract entities in German and Czech

position of a preposition and the definite article, and is very common in German. Most of them in our data (over 55%) refer to sentences or even larger segments, although NPs can also be their antecedents, see example (5).

(5) *Diese "Euro-Münzhaushaltsmischung" kostet 20 DM. Dafür bekommt man 20 Münzen zwischen 1 Cent und 2 Euro. [This household set of euro coins will cost 20 marks. For this, you get 20 coins between 1 cent and 2 euros in value].*

Anaphoric expressions in the form of nominal phrases modified by a demonstrative pronoun or a definite article, mostly contain a general noun, e.g. *Weise [way]* in example (6).

(6) *Die neue deutsche Truppe wurde vollständig in die militärischen Strukturen der Nato integriert. Auf diese Weise konnte das Ziel erreicht werden. [The new German units were fully integrated into NATO military structures. In this way it was possible to achieve the goal of...]*

This structure can refer both to longer segments as in example (6), and to clauses as in example (7), where we have an infinitive clause as an antecedent of *dieses Feld [this area]*.

(7) *Es ist eine der wichtigsten Aufgaben des Staates, die Erhaltung des freien Wettbewerbs sicherzustellen. Versagt der Staat auf diesem Felde, dann ist es bald um die soziale Marktwirtschaft geschehen. [Protecting free competition is one of the state' s most important tasks. If the state fails in this area, the social market economy will soon be lost].*

In all cases observed in our data, devices referring to abstract entities occur immediately in the following segment (either a clause or a sentence). No cases of longer distances were discovered in the data at hand.

5 Discussion and future work

In this paper, we present preliminary results of cross-lingual analysis of variation in abstract coreference. We analysed a small portion of texts in two languages that are not very close typologically using data sets annotated within two different frameworks. Our findings show that the differences of typological character (absence of definiteness or pro-drops) have also influence on the preferences for certain functional or structural types expressing coreference. We believe that the knowledge on the difference observed here is important for various areas of linguistics, including contrastive studies, translatology and multilingual NLP, especially machine translation. For instance, when translating from Czech

into 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 (e.g. using a discriminative translation model of *it* designed in Novak et al. (2013)) to see if we would also see changes in the preferences for abstract anaphora in translated German, as it was shown by Zinsmeister et al. (2012) for the 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 differences to avoid production of texts which sound less natural for the target language.

In our future work, we will also consider if the observed phenomena are genre- or domain-dependent. Coreference to abstract entities seems to be specific for the data at hand: abstract anaphora refer to the most central concepts in the analysed discourse. However, we need to have a look at further genres and domains, as well as at larger number of texts, for the evidence for this assumption.

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 work on this project was partially supported by the grant "Multilingual Corpus Annotation as a Support for Language Technologies" (LH14011, Ministry of Education, Youth and Sports). The project GECCo has been supported through a grant from the Deutsche Forschungsgemeinschaft (German Research Society).

References

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.

Simon Botley. 2006. Proceedings of the 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation. *International Journal of Corpus Linguistics*, 11(1):73–112.

Donna Byron. 2003. Annotation of pronouns and their antecedents: A comparison of two domains. Technical report, University of Rochester.

Stefanie Dipper and Heike Zinsmeister. 2009. Proceedings of the third linguistic annotation workshop (law iii).

Gill Francis. 1994. Labelling discourse: an aspect of nominal group lexical cohesion. In Malcolm Coulthard, editor, *Advances in Written Text Analysis*, pages 83–101, London:Routledge.

Kari Fraurud. 1992. Situation reference: What does it refer to? In *GAP Working Paper*, Fachbereich Informatik, Universitat Hamburg.

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

Nancy Hedberg, Jeanette K Gundel, and Ron Zacharski. 2007. Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles. *Proceedings of DAARC*, pages 31–36.

Lucie Kučová and Eva Hajičová. 2004. Coreferential Relations in the Prague Dependency Treebank. In *Proceedings of the 5th International Conference on Discourse Anaphora and Anaphor Resolution Colloquium*, pages 97–102, San Miguel. Edies Colibri.

Ekaterina Lapshinova, Anna Nedoluzhko, and Kerstin Kunz. 2015. cross languages and genres: Creating a universal annotation scheme for textual relations. In Ines Rehbein and Heike Zinsmeister, editors, *Proceedings of the Workshop on Linguistic Annotations, NAACL-2015*, Denver, USA.

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.

M.C. Müller. 2008. *Fully Automatic Resolution of It, this and that in Unrestricted Multi-party Dialog*. Ph.D. thesis, University of Tübingen.

Costanza Navarretta and Sussi Olsen. 2008. Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of LREC-08*.

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. 2013. Translation of "it" in a deep

syntax framework. In Bonnie L. Webber, editor, *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Workshop on Discourse in Machine Translation*, pages 51–59, Sofija, Bulgaria. Bălgarska akademija na naukite, Omnipress, Inc.

Massimo Poesio and Ron Artstein. 2008. Anaphoric Annotation in the ARRAU Corpus. In *International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May.

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.

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

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *LREC*.

Renata Vieira, Susanne Salmon-Alt, Caroline Gasperin, Emmanuel Schang, and Gabriel Othero. 2005. Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. *Anaphora Processing: linguistic, cognitive and computational modeling*, pages 385–403.

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