COREFERENCE CHAINS IN CZECH, ENGLISH AND RUSSIAN: PRELIMINARY FINDINGS[1]

Anna Nedoluzhko (nedoluzko@ufal.mff.cuni.cz)[1],
Svetlana Toldova (stoldova@hse.ru)[2],
Michal Novák (mnovak@ufal.mff.cuni.cz)[1]

[1]Charles University in Prague
[2]National Research University "Higher School of Economics"

This paper is a pilot comparative study on coreference chaining in three languages, namely, Czech, English and Russian. We have analyzed 16 parallel English-Czech newspaper texts and 16 texts in Russian (similar to the English-Czech ones in length and topics). Our motivation was to find out what the linguistic structure of coreference chains in different languages is and what types of distinctions we should take into account for advancing the development of systems for coreference resolution. Taking into account theoretical approaches to the phenomenon of coreference we based our research on the following assumption: the recognition of coreference links for different structural types of noun phrases is regulated by different language mechanisms. The other starting point was that different languages allow pronominal chaining of different length and that coreference chains properties differ for the languages with different strategies for zero anaphora and different systems for definiteness marking. This work reports our first findings within the task of the structural NP types' distribution comparison in three languages under analysis.

Keywords: coreference, coreference resolution, zero anaphora, NP-structural types distribution, cross-language comparison

> "Statistical models of anaphora resolution so far have only scratched the surface of the phenomenon, and the contributions to the linguistic understanding of the phenomenon have been few". (Poesio et al., 2011)

1. Introduction

Coreferential—anaphoric in particular—relations are common to most languages. However, the means of expressing these relations in languages can be different. The use of anaphoric expressions is influenced by different factors. These factors are not limited to such relatively language-independent ones, as context, pragmatic situation and semantics of referring expressions, but also by language-dependent

[1] The reported study was partially supported by RFBR, research project No. 15-07-09306

factors, such as pro-drop character of a language, different kinds of syntactic constructions common for a language in question, and so on. The present research is a pilot study aiming at comparison of the coreference chains structure in three languages—English, Czech and Russian. The goal of our research is twofold. From the linguistic point of view, hypotheses on coreferential expressions and coreferential chains will be formulated. From a computational perspective, this work will help us find specific features related to anaphoric expressions in text that can be further used as background knowledge for the development of a multilingual tool for coreference and anaphora resolution. In this paper, we will try to find out whether there are any language specific parameters of coreference chaining in different languages, what they are in particular, and what language phenomena they could be accounted for.

By a corefential chain we mean all the mentions of one and the same entity (or in some case mental entity or notion) through the whole text, irrespective of the features of a particular noun phrase (NP):

(1) (English) *Police say a husband fatally shot his wife and another man before* [$\varnothing_{PRO}$] *killing himself in a central Pennsylvania motel room.* […] *The York County Coroner's Office says 35-year-old Donnell Graham shot his wife.*

(2) (Russian) … *banker Lamberto Albuccani zastrelil svoju zhenu,…, a potom* [$\varnothing_{PRO}$] *zastrelilsya sam.*
[*lit.* The banker Lamberto Ambuccani shot his wife and then shot himself]

In (1), all the underlined NPs refer to the same entity (Donnell Graham), the first NP is an indefinite description denoting an entity's social role (*husband*), then we have a possessive anaphoric pronoun *his*, a reflexive pronoun *himself*, a zero inexpressible pronoun $\varnothing_{PRO}$[2] as a subject of a clause with a gerund *killing*, and a full NP including a proper name in the second sentence (*35-year-old Donnell Graham*). As for Russian Example (2), we have the proper name within a full NP (*banker Lamberto Albuccani*), a possessive reflexive svoju 'his', a zero pronoun as an agent of a finite verb *zastrelilsya*.

2. Motivation

Two basic issues, determining different parameters of coreferential expressions and chains, served us as motivation for the beginning of our research.

The first issue concerns the structural types of coreferring expressions. Although the quality of coreference resolution is usually evaluated as a whole, some researchers claim that syntactic and semantic structure of coreferring expressions might influence crucially the quality of coreference relations extraction. [Poesio et al. (2011)] suggests that recognition of different types of referring expressions (e.g., syntactic

[2] For zero NPs and their status see Section 2.

anaphora, named entities, metaphors, etc.) should be evaluated separately. We consider this approach very reasonable. Indeed, we should take into account that the referential choice for different NP classes (e.g., reflexive pronouns vs. anaphoric ones) is regulated by different language mechanisms that are studied within quite different linguistic paradigms. For example, syntactic anaphora (e.g., bound anaphora such as reflexives, some types of zero pronouns) is in the domain of formal syntactic investigation (see, e.g., in Chomsky (1981), Reuland (2011)). The discourse anaphora (i.e. 3rd person pronouns or demonstratives) is in the focus of referent activation theories (see, for example, Givon (1983), Ariel (2001), Kibrik (1997)). The recognition of metaphoric and metonymic expressions within coreference chains includes the semantic similarity detection (e.g., NP 'treasure' used to refer to a golden ring). To sum up, there are different language mechanisms, responsible for different NP types referential choice. Thus, the data on different NP types distribution within coreference chains for a particular language might be useful both for coreference chaining theoretical issues and for the multilingual coreference resolution task.

The second issue concerns language-specific zero pronominal elements in particular. For example, in our case, the three languages under discussion differ in the Subject ellipsis[3] options. In English, the syntactic subject should be always expressed explicitly. In other languages, like Czech and Russian, the subject can be omitted, or, in other words, free zero-pronoun (∅)[4] is used in some contexts (see Examples (3a-c)):

(3) *a.* (English) *Peter came home. *(∅) Watched TV and went to sleep.*
b. (Czech) *Petr se vrátil domů. ∅ Podíval se televizi a šel spát.*
c. (Russian) *Petya prishel domoj.* (On/∅) *Posmotrel televisor i poshel spat.*

We should also take into account another case of non-overt Subjects, coreferential to a NP in a previous context, that is the distribution of unexpressed Null Subjects (PRO) regulated by syntactic rules (e.g., PRO in infinitival constructions). To sum up, the syntactic properties of a language can influence the coreference chaining distribution in this language.

Thus, the purpose of our pilot study is just to compare the distribution of NP structural types (including free zero pronouns) in coreference chains for three languages.

3. Theoretical Background

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. They mostly address this phenomenon in one language. Some studies analyze the predictability

[3] Languages like Czech are called pro-drop languages. In pro-drop languages a pronoun (primarily in subject position) could be omitted in contexts where they could be pragmatically inferred.

[4] We do not draw a distinction between zero-pronoun and ellipsis here

of upcoming referents in relation to the choice of coreferring expressions and its status in information structure of an utterance (see the algorithm, determining the degree of salience in Hajičová et al., 2006; Lambrecht, 1994; Strube—Hahn, 1999, etc.). The referential choice regulation in subject position for Russian and its comparison to other world languages, first of all, to Germanic ones, is provided in Kibrik (2013). The deeper diachronic analyses of subject reference in Russian can be found in Sidorova (2013) and Kibrik (2013). A contrastive research of coreference and anaphoric reference of demonstratives in French and Portuguese is presented in Salmon-Alt et al. (2005).

As for corpus approaches, there is a large amount of large-scale 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, analyzing the difference between anaphoric expressions, based on large-scale annotated parallel corpora. The comparison of pronominal and zero coreferential expressions in Czech and English has been recently provided in Novák—Nedoluzhko (forthcoming in 2015). However, this work focuses on mappings between certain classes of coreferential expressions, and it does not take into account the structure of coreferential chains as a whole. Conversely, coreference chains have been included in the statistical analysis of cohesive devices in Kunz et al. (2015) for German-English corpus, containing written and spoken texts (GECCo), where the number of chains and chain lengths have been computed, but the collected numbers have not been analyzed yet.

4. Data

4.1. Description of the Corpora

Prague Czech-English Treebank (PCEDT) is a manually parsed Czech-English parallel corpus of 1.2 million words in almost 50,000 sentences for each language. The English part consists of the Wall Street Journal (WSJ) section of the Penn Treebank (Linguistic Data Consortium, 1999). The Czech part was translated from the English source sentence by sentence. PCEDT 2.0 is annotated on three layers; the most abstract (tectogrammatical) layer includes the annotation of coreferential links. For the detailed overview of the underlying linguistic theory, see Hajič et al. (2012).

For the Russian part of our investigation, we took the data from the Russian Coreference Corpus (RuCor). The corpus is the Gold Standard corpus for coreference resolution evaluation for Russian (Toldova et al., 2014). It consists of two parts, both manually annotated for coreference: the learning set and the evaluation set, 185 texts (200 000 tokens) in total. The corpus contains automatic morphological annotation. The set of tools, developed by S. Sharoff for Russian, was used, which includes a tokenizer, a TreeTagger-based (Schmid, 1994) part-of-speech tagger, and a lemmatizer, based on CSTLemma (Jongejan—Dalianis, 2009). Some of the tagger mistakes, e.g. in anaphoric pronouns POS detection, were corrected manually.

4.2. Coreference annotation in PCEDT and RuCor

In PCEDT, coreference links are annotated (mostly manually) for both the Czech and English parts separately. The coreference annotation captures the so-called grammatical and textual coreference. The grammatical coreference typically occurs within a single sentence, with the antecedent being able to be derived on the basis of grammar rules of a given language. It includes relative pronouns, verbs of control, reflexive pronouns, reciprocity and verbal complements. On the other hand, the textual coreference is not expressed by grammatical means alone, but also via context. Annotation of textual coreference in PCEDT captures coreferential relations between personal and possessive pronouns, anaphoric zeros, noun phrases with nominal head, numerals and pronominal adverbs with demonstrative meaning (there, then, etc.). Also the cases of event anaphora (anaphoric reference to clauses) have been annotated. A detailed description of the types of grammatical and textual coreference annotated in PCEDT can be found in Nedoluzhko et al. (2014).

In RuCor, the grammatical and textual coreference was annotated manually. The following coreferential relations have been taken into consideration: reflexive and relative pronouns (included in the so-called syntactic anaphora), 3rd person pronouns and 3rd person possessive pronouns, anaphoric zero, and noun phrases of different types. There is no syntactic zero (PRO) or event anaphora annotation. More details on coreference annotation in RuCor can be found in Toldova et al. (2014).

4.3. Dataset used for the experiment

For the aim of our analysis, we have chosen comparable texts for each corpus. Taking into account the length of the texts and their genre specification, we have extracted 16 parallel English-Czech texts from PCEDT and 16 Russian texts from RuCor. The thematics, genre, type and size of texts were determinants of the excerption. The average length of the texts is 31.8 sentences for PCEDT and 31.4 for Russian, the shortest text consisting of 14 sentences and the longest one—of 64 sentences. As for genres, these are all journalistic texts; the topics are distributed as shown in Table 1[5].

Table 1. The thematic structure of texts in RuCor and PCEDT (in sentences)

	PCEDT	**RuCor**
economics texts	161	166
political news	230	231
other news	112	105
TOTAL	503	499

[5] The choice of texts has been based on the annotated data we have at our disposal. We are fully aware that the fact that Czech texts are translated from English and the Russian texts are comparable but not parallel may considerably influence the numbers.

5. Quantitative characteristics of coreference chains

We started by gathering general statistics on the number and length of coreferential chains[6]. Although these data were just the preliminary stage of our study the results have turned out to be worth of a special analysis. They are presented in Table 2.

Table 2. Number and length of coreference chains in the analyzed texts

Chain length	English	Czech	Russian
number of 2-elements chains	254 (60.5%)	304 (62.7%)	139 (52.9%)
number of chains of length 3–4	108 (25.7%)	109 (22.5%)	64 (24.3%)
number of chains of length 5–8	39 (9.3%)	47 (9.7%)	33 (12.5%)
number of chains longer than 8	19 (4.5%)	25 (5.1%)	28 (10.6%)
TOTAL number of chains[7]	**420 (100.0%)**	**485 (100.0%)**	**263 (100.0%)**

As we can see from the table, the distribution of chains with different length for Czech, English and Russian is quite similar, at least comparable. More than 50% of chains consist of two elements in all three languages, about 25% of chains consist of three or four elements, and the rest is distributed between longer coreference chains of 5 and more elements. There is a slightly stronger tendency for chains longer than 5 items in Russian. The long chains are typical for the referents that are main topics of a text. Thus, the difference could be due to the texts thematic structure (there are two texts in Russian that have 4 and 6 chains longer than 8 items).

It is noteworthy that the proportion of two-member chains in Russian is 10 percentage points less than in English and Czech. Besides, there is a substantial difference in the absolute numbers of chains. The total number of coreference chains in English and Czech is almost twice as much as the number of coreference chains in the Russian texts.

Ex facte, this difference may be due to the fact that English and Czech texts come from the parallel dataset. More than that, an attempt was made to preserve the original structure of the English texts in their Czech translations. The same approach to coreference annotation has been applied to the texts in both languages. However, even the number of chains for Czech and English differ in 15% though these texts are from the parallel datasets. Thus, the difference in chains quantity for the languages under discussion could not be accounted for by the difference in themes, referents mentioned etc. It needs some other explanation.

The preliminary analysis of sentence structure and zero pronouns distribution in three languages has drawn us to the following hypothesizes: the observed disproportion in chain numbers might be attributed to (a) more extensive use of non-finite constructions

---

[6] For all analyzed languages, coreference chain is considered to consist of all coreferential expressions including pronouns and full NPs in the gives discourse.

[7] Coreference chains statistics presented in Table 3 for English and Czech texts takes into account only those types of coreferential relations that are annotated also for Russian. For this reason, for example, syntactic (grammatical) coreference of arguments in control constructions and event anaphora (reference to sentences and clauses) have been excluded.

in English introducing elided nodes which are involved in coreference annotation (see Section 6), and even more extensive use of non-finite constructions with an inexpressible PRO in Russian, (b) and to more extensive use of free zero pronoun in Czech.

6. Qualitative characteristics of coreference chains

As claimed in typological literature (e.g. Givon, 1983), languages differ in their repertoire of the anaphoric and coreference maintenance devices, especially, pronominal and syntactic devices. For the pro-drop languages, the zero pronoun is one of the basic means of reference maintenance, while for the other languages, zero is only possible in a limited number of syntactic positions (e.g., with infinitives). Another possible difference is the frequency of anaphoric pronouns in a chain. One more point of variation is the difference in the distribution of pronominal and nominal NPs across languages. Besides, the structure of nominal NPs could vary depending on whether a language has obligatory definiteness marking, and depending on what means a language uses to compensate for the absence of grammatical definiteness.

Taking in consideration all these possibilities, we compared the structure of markables in the three languages under analysis. Table 3 shows the distribution of structural types of noun phrases in English, Czech and Russian. In the table, all markables are considered, including initial antecedents and all anaphoric mentions.

Table 3. The distribution of NP-structural types for English, Czech and Russian

NP type \ language		English		Czech		Russian	
central pronouns	Subj	121	8.6%	2	0.1%	39	3.8%
	non-subj	129	9.2%	125	7.8%	95	9.3%
Relative		96	6.9%	136	8.5%	42	4.1%
anaphoric zero		28	2.0%	304	19.1%	13	1.3%
bare noun		75	5.4%	208	13.1%	164	16.0%
NP with determiner		315	22.5%	63	4.0%	20	1.9%
NP with other modif		119	8.5%	313	19.7%	172	16.8%
NP including NE		104	7.4%	141	8.9%	80	7.8%
NE		331	23.7%	216	13.6%	379	37.0%
other		81	5.8%	83	5.2%	21	2.0%
TOTAL		1,399	100%	1,591	100%	1,025	100%

6.1. Table description and observations

Central explicit pronouns (central pronouns) include 3rd person pronouns and 3rd person possessive and reflexive pronouns that are explicitly expressed in the sentence. For subject position this group is very scarce in Czech (only two instances), thus

corroborating the pro-drop character of Czech. In Russian texts, the subject anaphora is not as rare as in Czech. However, it is substantially more frequent in English. Only 30% of Russian anaphoric central pronouns have been used in the subject position, while in English subject and non-subject anaphoric pronouns are distributed nearly equally.

Relative pronouns (relative) as anaphoric expressions in coreference chains. The number of relative pronouns in Czech coreference chains is larger than that of relative pronouns in English and in Russian. One of the reasons for this discrepancy is that English non-finite clauses are often translated to Czech as relative clauses (see Example 4)[8].

(4) *a.* (English)*: Mr. Bush had been holding out for a bill ∅ boosting the wage floor to $ 4.25 an hour.*
b. (Czech)*: Bush trval na zákoně, který by zvyšoval dolní hranici mzdy na 4.25 dolaru za hodinu.*

As will be shown in Section 7, Russian is closer to English in this respect, both of the languages preferring non-finite clauses. There are only 40 cases of clauses with the relative pronoun "kotoryj" in our texts, while there are nearly 100 participial clauses.

Anaphoric zeros (anaphoric zero) are most frequent in Czech, compensating for the lack of explicitly expressed anaphoric pronouns in the subject position, and again supporting the idea of the pro-drop character of Czech. In this respect, the results for Russian are especially interesting. Though Russian is also considered to be a pro-drop language and zero anaphora is possible in Russian (see Example 1c and 3c), there are very few examples of textual anaphoric ellipsis in Russian, not more than one or two per text. Moreover, we have no zero in the subject position? of the main clause in our text collection.

(5) *a.* (English)*: The recent explosion of country funds mirrors the "closed-end fund mania" of the 1920s* [...] *They fell into oblivion after the 1929 crash.*
b. (Czech)*: Současná exploze národních fondů je stejná jako "mánie uzavřených fondů" ve 20. letech* [...] *Po krachu v roce 1929 ∅ upadly v zapomnění.*
c. (Russian)*: Strany Nato prevoshodili Jugoslaviju …, odnako cherez 2.5 mesyaca voiny ∅ byli na predele vozmoznostej* (='NATO countries overpowered Yugoslavia …, however, after 2.5 months (they) *stretched too thin')*

The number of anaphoric zeros in English is positive even though English is not a pro-drop language. Moreover, this number seems to be relatively high. However, all 28 English anaphoric zeros in our texts are arguments of nonfinite clauses, where the syntactic subject cannot be expressed explicitly (PRO). Thus, the reason for such a high number of zeros in English is rather technical, reflecting a slightly different conception of PCEDT in understanding the distinction between PRO and ∅.

The group of **nouns** is the set of non-pronominal non-zero markables. For the analysis of coreference chains and anaphoric expressions, the following subsets

[8] On a larger corpus and with statistically more representative results, this fact was addressed in Novák—Nedoluzhko (2015).

of nominal expressions seem to be relevant: bare nouns, NPs with a determiner, NPs with other modifiers, named entities (NEs), and noun phrases that include a named entity as a dependent element (marked as NP including NE in Table 3).

Not surprisingly, the number of **bare nouns** in Russian and Czech is much larger than in English (75,208 and 164 in English, Czech and Russian, respectively). As a language with the grammatical category of definiteness, English does not use bare nouns very often. The group of coreferential English bare nouns in our text selection consists mostly of plural nouns, nouns of time (Tuesday, yesterday, etc.) that could be also considered as named entities. On the contrary, NPs with determiners prevail in English, because many elements of coreference chains in English are used with the definite article. For Czech and Russian, noun phrases with demonstratives, corresponding to "this" and "that" have been counted. The structure of NPs with other modifications need deeper investigation. These NPs can include evaluative adjectives which do not contribute to the NP definite/indefinite interpretation, or geographic and other adjectives that serve for the referent identification.

The number of **named entity** roots (**NEs**) for Czech is substantially less than in English and Russian. Even though English and Russian behave similarly in many aspects, the reason for the discrepancy between these two languages and Czech probably lies somewhere else. The high frequency of named entities in Russian may stem from the text specificity and the difference in annotation scheme. Another possible reason is the tendency in Russian to repeat full named entities in cases where Czech uses anaphoric devices. On the other hand, the difference between English and Czech is affected by the differences in sets of categories used in automatic annotation of NEs. This was provided by the tools NameTag (Straková et al., 2014) and Stanford NER (Finkel et al., 2005) for Czech and English, respectively.

The category **other** for Czech and English includes mostly coordinative and appositive structures, such as coreferential expressions and clauses (sentences, verbal phrases) as antecedents. These are also the years and **other** numerals in substantive function, local and temporal adverbs like there, then and so on.

7. Discussion

7.1. Pro-drop properties

As it has been mentioned above, one of the important differences among the three languages is the difference in zero NPs distribution. First of all, all three languages differ a lot in their pro-drop properties (see Example 1). This property is crucial for Czech: the Table 3 shows that 19% of anaphoric NPs in Czech are zeros. This affects the difference in explicit pronoun distribution for Czech and Russian. Though Russian is also a pro-drop language, the proportion of zeros is very little. Moreover, there is no ∅ in the main clause in subject position in Russian in our data. Thus, we can assume that Russian is a pro-drop language to a lesser extent than Czech. There are very few cases of zero anaphora even in subordinate clauses in Russian. Our hypothesis is that the difference in clausal structure for these two languages could play the role.

7.2. Zeros and clause structure

Another important dissimilarity in chain distribution is that the number of chains in Russian differs from those in English and Czech. In our calculations, we do not take into account the inexpressible zeros (PRO). This type of anaphora was absent in annotation scheme for RusCor. However, we can try to compensate for the lack of information on PRO distribution by taking into account the distribution of finite/non-finite forms in the languages. As far as syntactic anaphora is concerned, the distribution of syntactically regulating pronouns (PRO and some others) depends on the sentence complexity. The non-finite subordinate clauses in Russian, such as infinitival constructions or participial constructions, presuppose an inexpressible PRO in the subject position (for the PRO distribution in Russian non-finite constructions see Testelets 2001).

Thus, as it is mentioned in 5 one of our hypothesises is that the difference in coreference chaining is strongly influenced by the clause structure of a sentence.

To check this hypothesis, we have counted the number of finite and nonfinite clauses in the three languages. The total number of sentences in all three collections was approximately the same (see Table 1). The results are given in Table 4.

Table 4. Finite and nonfinite clauses in texts

	Czech	**English**	**Russian**
number of finite clauses	1,166	1,005	663
number of nonfinite clauses	97	200	379

As seen in the table, there are 379 non-finite verb forms in Russian texts (among which there are 106 participial clauses, 59 short form participles, 17 converbs and 197 infinitive clauses), thus, we expect approximately 350 PROs.

It is interesting to observe that the Czech sentence in Example 6b can be hardly reformulated using an infinite clause (it is possible, but it will be stylistically marked, see Example 6c), while in Russian, either finite subordinate clause with relative pronoun (Example 6d), infinitive (Example 6e), or infinite participial clause (Example 6f) can be used. This fact supports the hypothesis that the relatively small number of coreference chains in Russian is caused by the frequent use of nonfinite clauses in this language, the arguments of which are not annotated for coreference in Russian coreference corpus.

(6) *a.* (English) *He left a message PRO accusing Mr. Darman of selling out.*
b. (Czech) ***Ø*** *Zanechal mu zprávu, ve které* ***Ø*** *viní Darmana ze zaprodanosti.*
?c. (Czech) ***Ø*** *Zanechal mu zprávu,* **PRO** *obviňující Darmana ze zaprodanosti.*
d. (Russian) ***On*** *ostavil soobschenije, v* ***kotorom*** *obvinjajet Darmana v prodazhnosti.*
e. (Russian) ***On*** *ostavil soobschenije, chtoby* **PRO** *obvinit' Darmana v prodazhnosti.*
f. (Russian) ***On*** *ostavil soobschenije,* **PRO** *obvinjajuscheje Darmana v prodazhnosti.*

Some differences in the properties of coreferential chains are also caused by the differences in annotation styles, which should not be neglected. We have not addressed these in a sufficient detail, which should be one of the aims of future research.

8. Conclusions

Our pilot study has shown that the inter-language comparison of coreference chains distribution reveals a systematic difference in coreference for languages with different syntactic properties. One of very important syntactic features that should be taken into account in modelling multi-lingual anaphora resolution is the language pro-drop properties. Another influential factor is the sentence structure, especially the distribution of finite vs. non-finite verb forms. The question of contrastive analysis of anaphoric chains is very interesting. Having been touched upon here they deserve more detailed qualitative and quantitative analysis. Our future work will be to analyze this topic in more detail by addressing separately more extensive parallel and non-parallel texts in the languages.

Acknowledgements

We acknowledge support from the Grant Agency of the Czech Republic (grant P406/12/0658), GAUK 3389/2015, EU (grant FP7-ICT-2013-10-610516—QTLeap) and SVV project number 260 224. On the Russian side, the study was supported by the Russian Foundation for Basic research (grant No. 15-07-09306). This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013). We thank the Lomonosov Moscow University students Ru-eval team for Russian data annotation, and Dmitrij Gorshkov for software support (the RuCor creation and Russian data management).

References

1. *Ariel, M.* (2001), Accessibility theory: An overview, in Sanders, T., Schliperoord, J. & W. Spooren (eds.) Text representation: Linguistic and psycholinguistic aspects, Amsterdam, Philadelphia: John Benjamins Publishing, pp. 29–87.
2. *Chomsky N.* (1981), Lectures on government and binding, Dordrecht, 1981.
3. *Finkel J. R., Grenager T., Manning Ch.* (2005), Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling, Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363–370.
4. *Givón T.* (1983), Topic continuity in discourse: introduction, in Topic continuity in discourse: Quantified cross-language studies, Amsterdam.
5. *Hajič J., Hajičová E., Panevová J., Sgall P., Bojar O.j, Cinková S., Fučíková E., Mikulová M., Pajas P., Popelka J., Semecký J., Šindlerová J., Štěpánek J., Toman J., Urešová Z., Žabokrtský Z.* (2012), Announcing Prague Czech-English Dependency Treebank 2.0, Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Copyright © European Language Resources Association, İstanbul, Turkey, ISBN 978-2-9517408-7-7, pp. 3153–3160.

6. *Hajičová E., Hladká B., Kučová L.* (2006), An Annotated Corpus as a Test Bed for Discourse Structure Analysis, Proceedings of the Workshop on Constraints in Discourse, Copyright © National University of Ireland, Maynooth, Ireland, pp. 82–89.
7. *Jongejan B., Dalianis H.* (2009), Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.—Singapore: Association for Computational Linguistics, 2009.—P. 145–153.
8. *Kibrik, A. A.* (2011), Reference in discourse, Oxford, Oxford University Press.
9. *Kibrik, A. A.* (2013), Peculiarities and origins of the Russian referential system, in Dik Bakker and Martin Haspelmath (eds.) Languages Across Boundaries: Studies in Memory of Anna Siewierska, Berlin, Mouton de Gruyter.
10. *Kibrik, A. A.* (1997), Modelling of multifactor processes: referential choice in Russian discourse [Modelirovaniye multifaktornogo protsessa: vybor referentsial'nogo sredstva v russkom diskurse], MSU reporter [Vestnik MGU], Vol. 4.
11. *Kunz, K., Degaetano-Ortlieb, S., Lapshinova-Koltunski, E., Menzel, K. and Steiner, E.* (to appear 2015), GECCo—an empirically-based comparison of English-German cohesion, in De Sutter, G. and Delaere, I. and Lefer, M.-A. (eds.). New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies. TILSM series. Mouton de Gruyter.
12. *Lambrecht, K.* (1994), Information structure and sentence form. Topic, focus and the mental representation of discourse referents, Cambridge, Cambridge University Press.
13. *Linguistic Data Consortium* (1999), Penn Treebank 3. LDC99T42.
14. *Mikulová M., Bémová A., Hajič J., Hajičová E., Havelka J., Kolářová V., Lopatková M., Pajas P., Panevová J., Razímová M., Sgall P., Štěpánek J., Urešová Z., Veselá K., Žabokrtský Z., Kučová L.* (2005), Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technical report no. 2005/TR-2005-28, Copyright © ÚFAL MFF UK, Prague, ISSN 1214-5521, 1185 pp.
15. *Nedoluzhko A., Mírovský J., Fučíková E., Pergler J.* (2014), Annotation of coreference in Prague Czech-English Dependency Treebank. Technical report no. 2014/TR-2014-57, Copyright © ÚFAL MFF UK, ISSN 1214-5521, 41 pp.
16. *Novák M., Nedoluzhko A.* (to appear in 2015), Comparison of coreferential expressions in Czech and English, in Discours, Vol. 16.
17. *Poesio M., Ponzetto, S. Versley, Y.* (2011), Computational models of anaphora resolution: A survey Linguistic Issues in Language Technology, available at http://cswww.essex.ac.uk/poesio/papers.html.
18. *Reuland, Eric J.* (2011), Anaphora and language design, Cambridge, MA: MIT Press.
19. *Salmon-Alt S., Vieira R.* (2002), Nominal Expressions in Multilingual Corpora: Definites and Demonstratives, in Language resources and evaluation conference LREC 2002, Las Palmas, Spain.
20. *Salmon-Alt S., Vieira R., Gasperin C.* (2005), Coreferent and anaphoric demonstrative NPs, in António Branco, Tony McEnery, Ruslan Mitkov (eds.) Anaphora Processing: Linguistic, Cognitive and Computational Modelling, Jonh Benjamons, Lisbon, Portugal.

21. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of International Conference on New Methods in Language Processing. Manchester. 1994. Vol. 12, Issue 4. P. 44–49.
22. *Sidorova E. V.* (2013), Evolution of subject reference in Russian [Evolutsiya sub'yektnoy referentsii v russkom yazyke], Master thesis.
23. *Straková J., Straka M. and Hajič J.* (2014), Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition, Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 13–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
24. *Strube, M. and U. Hahn* (1999), Functional Centering: Grounding Referential Coherence in Information Structure. Computational Linguistics 25/3, pp. 309–344.
25. *Testelets, Ya.* (2001), Introduction to general syntax [Vvedenije v obščˇij sintaksis], Moscow: RGGU. 2001.
26. *Toldova, S., Grishina Ju., Ladygina A., Vasilyeva M., Nedoluzhko A., Rojtberg A., Azerkovich I., Kurzukov M., Ivanova A.,* (2014):RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian. In: Computational Linguistics and Intellectual Technologies,ISSN 2221-7932, 13 (20), pp. 681–694