Cross-Lingual Coreference: The Case of Bulgarian and English

Zara Kancheva
LMaKP
IICT-BAS
Sofia, Bulgaria
zara@bultreebank.org

Abstract

The paper presents several common approaches towards cross- and multi-lingual coreference resolution in a search of the most effective practices to be applied within the work on Bulgarian-English manual coreference annotation of a short story. The work aims at outlining the typology of the differences in the annotated parallel texts. The results of the research prove to be comparable with the tendencies observed in similar works on other Slavic languages and show surprising differences between the types of markables and their frequency in Bulgarian and English.

1 Introduction

Coreference tends to be a common subject of research nowadays due to its various NLP applications like text summarization, question answering, information extraction, machine translation, named entity recognition, etc. For the accomplishment of these applications many coreference annotated corpora have been built and a number of annotation schemes have been created.

Many recent investigations focus on the coreference resolution in parallel corpora or translated texts with multiple languages (major and less wide-spread) and thus face a number of challenges like choosing between automatic and manual annotation, between different genres and size of the data, guidelines, tools and methods for analysis.

In the current research, the original English[1] text and the translated Bulgarian version of "The Adventure of the Speckled Band" by Sir A.C. Doyle are taken as a starting point for finding cross-lingual coreference similarities and differences. For this task, OntoNotes guidelines have been adapted to accommodate for the specifics of the two languages. Both texts have been manually annotated within WebAnno system. The manual approach to the coreference annotation would contribute to the future work on the automatic processing by improving the evaluation process as a gold annotation. The investigation will be used for facilitating the creation of a coreference resolver for Bulgarian.

The paper is structured as follows: the next section presents relevant related work; section 3 presents the dataset and the annotation process; section 4 illustrates the typology of the differences observed; section 5 shows directions for future work, and concludes the paper.

2 Related Work

One of the main methods for treating coreference in parallel texts is the projection. Formerly used for various purposes as POS tags projection (Yarowsky et al., 2001), dependency structures projection (Hwa et al., 2005) or semantic roles projection (Pado and Lapata, 2005), this approach proves effective also for projecting coreference chains.

The work of (Postolache et al., 2006) is based on that method and applied to coreference for the first time, using a parallel corpus, containing three parts of the English original and Romanian translation of the novel "1984". The researches focus only on noun phrases and do automatic word alignment with a Romanian-English aligner; they extract the corresponding referential expressions and transfer the English coreference chains to Romanian.

(Grishina and Stede, 2015) apply knowledge-lean projection of coreference chains across three languages – English, German and Russian. In this research the specifics of the genre are also considered and thus argumentative newspaper articles,

[1] In the text we refer to English as a source language and to Bulgarian as a target one.

Proceedings of the Student Research Workshop associated with RANLP-2019, pages 32–38,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/issn.2603-2821.2019_006

narratives and medicine instruction leaflets are annotated and later aligned with the commonly used for this type of investigation tool GIZA++ (Och and Ney, 2003).

In following work, (Grishina and Stede, 2017) expand their approach with a new method. They present an annotation projection from multiple sources again with a trilingual parallel corpus of English, German and Russian. In both their articles, the authors use an annotation scheme similar to the guidelines of OntoNotes as it is the case in the work presented - an adapted version that holds also for Bulgarian is used.

Another corpus-based approach is employed by (Novak, 2018) with about 100 times bigger data compared to the previously mentioned works from the Prague Czech-English Dependency Treebank 2.0 (Nedoluzhko et al., 2016). Here, the word alignment is done again with GIZA++ and the analysis of the mention types is inspired by (Grishina and Stede, 2017) and even expanded with a new category, anaphoric zeros, which is essential for a pro-drop language like Czech.

In their next project, (Nedoluzhko et al., 2018) further investigate the cross-lingual coreference with the PAWS parallel treebank with texts in four languages - English, Czech, Russian and Polish - by annotating and analysing not only noun, but also verb phrases.

A different approach is presented by (Lapshinova-Koltunski et al., 2019) who use an English-German parallel corpus annotated manually with coreference information (ParCorFull) in order to discover, analyse and introduce a typology of differences in the coreference chains (referred to as ‘incongruences’).

Another line of research has its focus on the type of pronouns of the referential entities (Novak and Nedoluzhko, 2015). The authors thoroughly investigate the nature of the correspondences between the Polish and English chains by manually annotated alignments of coreferential expressions. Since the aim of the current research is to offer a preliminary outline of some specifics of coreference annotation in parallel English-Bulgarian texts, the model of analysis used in (Lapshinova-Koltunski et al., 2019) and the one of (Novak and Nedoluzhko, 2015) is applied in combination. The result is a typology of differences between Bulgarian and English coreference chains.

3 Annotation

As (Lapshinova-Koltunski et al., 2019) defines, no matter what pairs of languages are exemplified in the parallel texts for coreference annotation, there will always be some language-typology and translation-process-driven differences in the coreference chains. Besides the type of language and the type of translation (machine- or human-translated), the genre of the text has considerable impact as well. In the present research a piece of fictional literature is used, similarly to the approach of (Postolache et al., 2006).

Lots of parallel corpora used for cross-lingual coreference resolution consist of news articles (Nedoluzhko et al., 2018), (Novak, 2018) and some of them contain more than one type of texts (Grishina and Stede, 2015), (Grishina and Stede, 2017) .

The preliminary hypotheses concerning the types of annotation differences are:

- a missing coreference chain in the source text;
- a missing coreference chain in the target text;
- an identical coreference chain in both texts, but with different types of referential expressions;
- an identical coreference chain in both texts, but with different number of referential expressions;
- mismatching annotators decisions;
- annotation errors.

Some of the most obvious differences between the original and the translated text are as follows: a) the size of the text: “The Adventure of the Speckled Band” in English contains 608 sentences while the Bulgarian version - 647, and b) the total number of referential entities: 2133 in the first text, and only 1089 in the second. The source text has 329 coreference chains while the target text – only 190.

The texts were manually annotated with coreferences by two annotators working at first independently from each other and later - together with the web-based annotation tool WebAnno 2.3.1. (Yimam et al., 2014). The consequent analysis was done with the XML-based software system

CLaRK[2] (Simov et al., 2004). The additional processing with CLaRK was necessary because it works well with large texts and has no issues with languages and different encodings, which is not the case with WebAnno.

The annotation was done in accordance with the OntoNotes guidelines; at the same time some necessary modifications were made. Noun, adjective, adverb and pronoun antecedents and anaphora were annotated. The extension with event coreference and bridging anaphora is considered as one of the directions for future work. The modifications of the annotation scheme affect:

- subordinate clauses - in the OntoNotes guidelines these cases are taken for markables only if they contain the relative pronouns *which* and *who*, but in the current annotation, constructions with *when*, *where* and *that* are treated in the same way as the previous two;

- constructions of the type *only* + *plural noun* are considered generic;

- in constructions of the type *each of* + *noun/pronoun* only the noun or pronoun from the phrase are marked as referential expressions.

4 Typology of Differences

Annotation differences could be analysed and classified from various points of view. One possible approach is that of (Lapshinova-Koltunski et al., 2019), inspired by the work of (Klaudy and Karoly, 2005) on explicitation and implicitation in translation. (Lapshinova-Koltunski et al., 2019) present a typology of incongruences, outlining four types:

1. explicitation - it takes place when the translation contains more specific or new (not present in the source text) linguistic units; phrases are extended and sentences are split into two sentences;

2. implicitation - the translation is shorter than the source;

3. different interpretations - this is the case when annotators interpret the parallel texts in a different way;

4. annotation error - this concerns errors done during the manual annotation of the texts.

As considered by (Klaudy and Karoly, 2005), explicitations and implicitations may be:

- *obligatory* - their presence is motivated by the characteristics of the language and they serve to make the translation more comprehensible;

- *optional* - (Klaudy and Karoly, 2005) point out that in this case translators decide whether to apply explicitation or implicitation based on differences in language use, discourse structure, and background information.

The classification of (Novak and Nedoluzhko, 2015) distinguishes between three types:

1. central pronouns - this class includes personal, possessive, reflexive and reflexive possessive pronouns; the study shows that more than the half of all personal English pronouns turn out to be Czech anaphoric zeros.

2. relative pronouns - here pronominal adverbs are also added;

3. anaphoric zeros.

In our study the latter approach is applied, but also a deeper analysis of the nature of the annotation differences and examples is presented. It was stated earlier that the source text has almost 50 percent more coreference chains and referential entities than the target text. Most likely this substantial difference in quantity is due to the typical for Bulgarian zero anaphora. A lot of research has been devoted to that phenomenon, and it still seems to be the most sophisticated variety of anaphora, as noted by (Mitkov, 2002). For that reason, ellipsis is considered a separate class in the typology of annotation differences.

Zero Anaphora

This type of difference in the cross-lingual coreference annotation is very common. Probably every translator's basic aim is to give the translated text the most natural form possible, so the annotated Bulgarian version of "The Adventure of the Speckled Band" has lots of ellipses, especially zero pronominal anaphora:

[2] http://bultreebank.org/en/clark/

(1) Както виждам , пристигнали сте
As see-I , arrive-Part were-you
с утринния влак .
with morning-the train .
'You have come in by train this morning, I see.'

The frequent omission of personal pronouns in the text illustrated in (1) results in shorter coreference chains (with less referential entities) in the translated story compared to the original. In the next example, two phenomena can be observed: a pronominal zero anaphora and an implicitation.

(2) Не издържам повече , ще полудея .
No stand-I no longer , shall go-I mad .
'Sir, I can stand this strain no longer; I shall go mad if it continues.'

Because of the dropped personal pronoun, the omission of the title *sir* and the phrase *this strain*, there are no referential entities in the first clause and the second clause has a short coreference chain (*the strain, it*) with no analogue chain in the target text.

An example for a zero noun anaphora with *cases* as an antecedent is found in the following sentence:

(3) Сред всичките тези случаи един от
Among all-the these cases one of
най-интересните безспорно е с
most-interesting-the undoubtedly is with
известния род [...] .
famous-the family [...] .
'Of all these varied cases, however, I cannot recall any which presented more singular features [...] .'

All the types of zero anaphora defined in (Mitkov, 2002) - pronominal, noun, verb, verb phrase anaphora - are present in the target text, however the ones including verbs are not in the focus of this survey.

Explicitation and Implicitation

Numerous cases of explicitation and implicitation were observed in the Bulgarian translation of A.C. Doyle's story. Most of them seem to be optional. This could be explained with the translator's decision, not necessarily with the specifics of the language, as the following example illustrates:

(4) Настъпи дълго мълчание . Холмс
(Followed a-long silence . Holmes
седеше вторачен в огъня .
was-sitting staring in fire-the.)
'There was a long silence, during which Holmes leaned his chin upon his hands and stared upon the crackling fire.'

The "details" that the translator skipped (*his chin, his hands*) would actually be parts of the coreferential chain if present in the Bulgarian sentence.

Other cases of explicitation might be observed in examples like the next one where an English sentence with a subordinate clause is divided into two shorter sentences with the subordinate clause transformed to main clause in Bulgarian:

(5) Жената , в черни дрехи и с
Woman-the , in black clothes and with
плътен воал , седеше до
thick veil , was-sitting-she by
прозореца . Когато ни видя ,
window-the . When us saw-she ,
веднага стана .
immediately rose-she .
'A lady dressed in black and heavily veiled, who had been sitting by the window, rose as we entered.'

The Bulgarian version has a new markable, *the veil*, which does not have an analogue in the English one.

The translation could rather easily lower the number of markables with the means of explicitation:

(6) А когато една млада жена се
And when one young woman se.Refl
втурне толкова рано сутринта през
rushes so early morning-the through
столицата да буди спящите , със
capital-the to wake up sleeping-the , with
сигурност има да съобщи нещо
certainty has to announce something
много важно .
very important .
'Now, when young ladies wander about the metropolis at this hour of the morning and knock sleepy people up out of their beds, I presume that it is something very pressing which they have to communicate.'

In this example the markables *young ladies/they, sleepy people/their* do not have any correspondences in the target text.

Most Frequent Markables

Next sentences hint to one possible explanation about why (and how) the target text ends up with lower number of coreference chains and different markables than the source text:

Pronoun type	Bulgarian	Mentions	English	Mentions
Personal	аз	18	I	224
Personal	ти	8	you	99
Personal	той	41	he	108
Personal	тя	22	she	56
Personal	ние	5	we	70

Table 1: Most frequent pronouns in the coreference chains.

(7) Майка ни умря скоро след нашето
Mother our died-she shortly after our
завръщане в Англия . Загина преди
return to England . Perish-she before
осем години при железопътна
eight years in railway
катастрофа недалеч от Кру .
accident not far from Crewe .

‘Shortly after our return to England my mother died — she was killed eight years ago in a railway accident near Crewe.’

Here the combination of explicitation and zero anaphora lead to the presence of new markables and chains:

- *our (mother)* - (refers to *Helen and Julia*) will not have correspondence with the English *my (mother)* - referring only to *Helen*;
- *our (return)* - (refers to *Helen, Julia, their mother* and *their father*) will not have analogue in the Bulgarian text;
- *the mother* is literally mentioned once in the target text, because of the zero pronoun anaphora, and in the source text there are two expressions referring to her - *my mother, she*.

The following example illustrates a case of implicitation, in particular - a simplification of a phrase:

(8) Тя спусна плътния си черен воал и
She dropped thick-the her black veil and
излезе .
left-she .

‘She dropped her black thick veil over her face and glided for the room .’

It cannot be concluded that the implicitation in this case is obligatory - if the translation was literal it would not make the sentence incomprehensible or lead to unnecessary repetitions. The translator’s approach leads to the lack of two markables in the target text - *her face and* the room.

The opposite process is also frequently observed:

(9) Усмивката се разля още
Smile-the se-Refl. spread more
по-широко върху лицето на Холмс .
wider on face-the of Holmes .

‘His smile broadened .’

With the optional explicitation the Bulgarian sentence has three additional markables - *smile*, *face* and *Holmes* unlike the English sentence with two.

The observations with respect to the grammatical category of the annotated referential expressions in the Bulgarian text show that some of the most frequent markables are proper nouns - mainly names of the characters in the story, but also names of locations. The total number of proper nouns for main characters (which form the longest coreference chains) is 135, of which, predictably, 68 are referring to Sherlock Holmes. In the English version the results are similar - 122 proper nouns for character’s names and 62 of them referring to Holmes.

Other frequent markable from the class of personal pronouns is *he* (той) with 41 uses in Bulgarian and 108 in English, followed by the plural personal pronoun *you* (ви) with 123 mentions in the source text and 26 in the target one. Another English possessive pronoun, *his*, has the remarkably high frequency of 95 mentions. In the translation, it is expressed by the Bulgarian reflexive possessive particle си (15 mentions) or with the short form of the non-reflexive possessive form му (28 mentions).

As previously stated, subordinate English sentences are usually transformed and simplified in the Bulgarian translation. The analysis of the pronoun markables proves this observation once again - there are 90 mentions of the relative pronoun *which*, and the other pronouns of this class are also very common. However, the Bulgarian corresponding forms are actually rare.

It can be concluded that zero pronoun anaphora are the main reason for the pronounced difference in terms of pronoun mentions frequency in the two languages. The results of the analysis of the central pronouns are very similar to the conclusions of (Novak and Nedoluzhko, 2015) based on the comparison between Czech and English coreference chains. The personal pronoun *I* has the highest rate of mentions in English while its Bulgarian analogue is rarely mentioned; the explanation for this phenomenon could be illustrated with examples of this kind:

(10)	Познах	гласа	на	сестра	си
	Recognized-I	voice-the	of	sister	si.Refl
	[...] .				

‘I knew that it was my sister’s voice .’

5 Conclusions

The results from the observations made in the current study serve to support the creation of bigger quantities of coreference parallel corpora with Bulgarian as member of the language pair.

The existence of such corpora will allow for training a coreference resolver for Bulgarian and consequent experiments on the cross-lingual coreference resolution with Bulgarian.

This preliminary work might serve as a first draft for coreference annotation guidelines for Bulgarian, for the semi-automatic annotation of basic coreference chains, and with the creation of a larger bilingual corpus – for the fully automatic processing, as these are the directions of our future work. The next stage of the research is planned to include investigation of event coreference and bridging anaphora.

The work performed in this study is intended to serve as a base for a Ph.D. dissertation that would provide a thorough insight on the subject.

Acknowledgments

This work was partially supported by the Bulgarian Ministry of Education and Science under the National Research Programme “Young scientists and postdoctoral students” approved by DCM # 577 / 17.08.2018. Also the research reported here was partially funded by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-164/28.08.2018.

References

Yulia Grishina and Manfred Stede. 2015. Knowledge-lean projection of cireference chains across languages. In *Proceedings of the Eight Workshop on Building and Using Comparable Corpora*. Association for Coputational Linguistics.

Yulia Grishina and Manfred Stede. 2017. Multi-source annotation projection of coreference chains: Assessing strategies and testing opportunities. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, Valencia, Spain. Association for Computational Linguistics.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*.

Kinga Klaudy and Krisztina Karoly. 2005. Implicitation in translation: Empirical evidence for operational asymmetry in translation. *Across Languages and Cultures*.

Ekaterina Lapshinova-Koltunski, Sharid Loáiciga, Christian Hardmeier, and Pauline Krielke. 2019. Cross-lingual incongruences in the annotation of coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*. Association for Computational Linguistics.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman, Great Britain.

Anna Nedoluzhko, Michal Novak, Silvie Cinkova, Marie Mikulova, and Jiri Mirovsky. 2016. Coreference in prague czech-english dependency treebank. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Anna Nedoluzhko, Michal Novak, and Maciej Ogrodniczuk. 2018. Paws: A multi-lingual parallel treebank with anaphoric relations. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*. Association for Computational Linguistics.

Michal Novak. 2018. A fine-grained large-scale analysis of coreference projection. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*. Association for Computational Linguistics.

Michal Novak and Anna Nedoluzhko. 2015. Correspondences between czech and english coreferential expressions. *Discours. Revue de linguistique, psycholinguistique et informatique*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sebastian Pado and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Transferring coreference chains through word alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. European Language Resources Association.

Kiril Simov, Alexander Simov, Hristo Ganev, Krasimira Ivanova, and Ilko Grigorov. 2004. The CLaRK System: XML-based Corpora Development System for Rapid Prototyping. *Proceedings of LREC 2004*, pages 235–238.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Seid Muhie Yimam, Richard Eckart de Castilho, Iryna Gurevych, and Chris Biemann. 2014. Automatic annotation suggestions and custom annotation layers in webanno. In *Proceedings of ACL-2014*. Association for Computational Linguistics.