Michal Novák
Coreference from Cross-lingual Perspective
Doctoral thesis

Reviewed by Alexandr Rosen

The thesis investigates coreference by linguistically informed computational methods of coreference resolution (CR), based on parallel bilingual treebanks (generated by available tools even when a manually checked version of treebank was available). There were two goals: (i) to add the coreferential dimension to a multilingual contrastive project, answering the question of whether two specific cross-lingual methods can quantify similarities and differences between the languages; and (ii) to explore bilingually informed CR as a way towards coreference annotation of parallel corpora. The author succeeded in both. As for (i), both methods proved to be useful for the comparison of English and Czech, yielding interesting results, such as the fact that English is more informative than Czech for CR. As for (ii), the experiments showed that bilingually informed CR outperforms monolingual approaches (using the same scenario otherwise). There were also several spin-off results: improvement of the Czech monolingual coreference resolver, a newly designed supervised learning method targeting selected coreferential expressions outperforming traditional approaches, and a dataset of manually annotated Czech-English correspondences between coreferential expressions.

The thesis is logically structured, starting from impressive introductory parts on theoretical foundations, related work, data resources, tools and evaluation measures, continued by quantitative and qualitative analysis of English and Czech anaphora in the gold parallel treebank data and a description of cross-lingual alignment of co-referring expressions, before finally focusing on the two methods of using parallel texts to improve CR. The methodology seems sound. The experiments are described and evaluated in a clear, concise and fair way. With just a few exceptions the text is clear enough. Typos and other mishaps are quite rare (see the few cases I spotted below).

There are at least three main merits of the thesis: (i) The combination of several advanced machine learning methods, concerned with real texts and open to evaluation, with a strong role of linguistic analysis, reflected in the use of highly abstract representation with empty elements. (ii) Interesting linguistic observations, including some truly revealing quantitative results, due to their factoring by types of anaphors. These include especially the higher success of English-informed CR of Czech as compared to Czech-informed CR of English, as exemplified by example 7.1 on page 102, where the neuter gender of English *it* provides unambiguous reference to the antecedent, unlike the ambiguous gender of its Czech equivalent *ho*, which has two potential antecedents within the same sentence. The author correctly says that *English pronoun's gender thus serves rather as an animacy feature, which cannot be reconstructed solely from the Czech pronoun.* Page 117: *With respect to coreference, English is more informative for Czech than vice versa.* (iii) Impressive results of the cross-lingual CR.

In the following I focus on specific issues. Any corrections or comments should be taken in the context of the overwhelmingly respectable piece of work.

Theory

The thesis is concerned with linguistic notions and actually verges on a translation study. To explain some phenomena found in the parallel corpus, some more insights from translatology in addition to the Explicitation Hypothesis quoted on page 112 could be useful.

The distinction between grammatical and textual coreference is described on pages 14–15, but is not reflected in the models of coreference or in the analyses of the results. An option of treating references by relative or reflexive pronouns by a specialized model within a cascade of models is suggested on page 86, but the option is not implemented. The reviewer found no other

specific mention corresponding to a theoretically founded binding theory or syntactically determined rules of coreferences. Is it because the distinction is not useful for practical CR?

The Czech and English relative possessive pronominal forms (*jehož, jejíž, …*; *whose*) are not listed in §2.4 under the possessive or relative type, and they are not mentioned elsewhere either.

It is not quite clear why first and second person anaphora are excluded (page 20, last paragraph).

Page 64, paragraph 2, about 20% less possessives in Czech: this seems to be a trivial fact – no decent translator would render *raise your hand* as *zvedni svou ruku*. The preference of dative personal pronouns to possessives in Czech is also part of translators' wisdom.

Incorrect or imprecise claims:

- The author seems to assume that only infinitives participate in grammatical control (bullet 3 on page 14), while (2.4) *John cannot stop laughing* is standardly treated as a case of raising, indistinguished from control by FGD. Similarly in Czech – page 47, paragraph 5. Moreover, there is a wrong assumption that a verb governing the controlled infinitive is always finite.
- Page 85, paragraph 1: *…anaphoric pronouns tend to agree with their antecedent in morphological gender and number…* – also in person
- Ibid.: *…reflexive pronouns point to a sentence subject…* – they may point to the subject of an embedded clause, including subjects of non-finite clauses *she told him to shave himself*
- Page 102, last paragraph: *The analysis also shows that English syntax, which is more strict and thus easier to reconstruct, often helps in determining the correct antecedent.* The claim is based on example 7.2, where the only difference of the relevant structure is in the Czech use of genitive instead of the English use of a PP. Perhaps it would be more acceptable to use replace *syntax* by *word order*?

Data

Page 6, paragraph 4: *Czech and English are actually a good choice of language also from the linguistic point of view. The way how they realize coreference relations on the surface could not differ more.* This is a strong but unsubstantiated claim. From a more distant perspective, English and Czech are in the same language family. Comparison with any non-Indo-European language might falsify the claim and show that Czech and English are actually not that far, not even according to the maps showing the way they express coreference.

Related to its focus on contrastive linguistic issues is the choice of the data. From the viewpoint of translatology or corpus linguistics, the corpus at the heart of the thesis (PCEDT) fails in several respects. Translation per se may be problematic as linguistic evidence, but when accepted, texts are carefully selected, usually in balanced proportions of text types and controlled shares of originals and translations for individual languages, often examined separately for each direction of translation. PCEDT includes texts of a single type, with its Czech side translated according to rules to suit a specific purpose of statistical machine translation, with a result that the Czech part of PCEDT is biased towards English. These facts are noticeable from some examples (see below) and may distort some of the reported results. Statistics by genre, based on the much larger, varied and 'properly' translated CzEng, are presented in an attachment, but this is the only role of the CzEng corpus in the thesis.

Some examples include misleading annotations or translation:

- Example 1.1 on page 6, represented as tectogrammatical tree on page 13 and 41: zero actor of *using* in the English version refers to *formula*, resulting in the unlikely reading *a formula uses the Coke*, whereas in Czech it is the company which uses the formula for its beverage. Most likely, in English it is the company using its new Coke as an application for the formula.
- Example 2.13 on page 13: *takov**é** auta* 'such cars' should be translated as *takov**á** auta*.
- Figure 7.1 on page 85: *…the magazine in January would begin publishing without advertising.* The governor of the node representing *without advertising* is the node for *that would begin*. The reviewer believes the governor should rather be the node representing the word *publishing*.
- In 7.3 on page 103, it is very hard to identify the antecedent of the Czech possessive *jeho* with the bare *GM*, as suggested by the author. This is probably due to the repetition of the semantically more likely but gender-wise incompatible candidate noun *společnost* 'company' much closer to the pronoun and the presence of several other compatible candidates along the way. It seems to be a case at least of a suboptimal translation.

Sometimes a better example could be picked:

- Example 5.11 on page 66 is supposed to illustrate the fact that Czech personal pronouns need not or cannot be expressed in English: *he would write the letters as ordered* is translated as *dopisy napíše tak, jak **mu** bylo nařízeno*. However, the Czech pronoun *mu* is also optional, perhaps the translator decided it should be present for stylistic reasons. Anyway, a better example could be used.

Clarification needed

- Page 5: cross-lingual projection vs. bilingually informed resolution – the non-expert reader may be lost about the difference here, a brief explanation or at least a pointer to where the methods are described would help
- Page 12, bullet 2: why is gender a semantically indispensable feature of nouns? Is it because semantic nouns include pronouns?
- Page 20, line 1: as a curtesy to the reader, the notion of quasi-control could be briefly explained, especially if it is part of the research.
- Page 21, paragraph 3: *coordination root* – may not be understandable to someone unfamiliar with PDT or FGD.
- Page 41, line 6: *Coreference annotation for PCEDT 2.0 Coref has proceeded in two stages. In the first stage resulting in PCEDT 2.0, …* But PCEDT 2.0 existed before this task started.
- Page 46, bullet 6: *Valency frames specify how a verb is connected with its arguments and modifiers in a particular sense. It can then be used to check if the tectogrammatical structure complies with the estimated valency frame.* – Does *it* in the second sentence really refer to *sense*? Has this been actually done?
- Page 114, paragraph: shouldn't the modifiers such as *společnost* be part of the named entity?

Formal, typographical

Mishaps or blemishes very rare:

- page 7, bullet 1, line 3: word salad
- page 11, beginning of §2.2: unintended editing note
- page 15, last paragraph: reference to Table 2.1 should include page number
- page 43, paragraph 2: missing period after *Section 4.3.1*

- page 46, bullet 4: missing space after *respectively).*
- page 102, paragraph 3: *Tables 7.6 and 7.6* – should be 7.6 and 7.7?
- page 114, paragraph 2: *ven if they are correctly aligned*
- word examples in the text rendered as "*kde /where/*", maybe *kde* 'where' would be more preferable
- short space after *e.g.* etc. ("etc.\ "in latex)
- en-dash instead of hyphen in intervals such as 2-3 (> 2–3)

English

A good command of specialist language in most sections, including terminology, except for a single case:

- page 11, line ^6: *automatic machine approaches*

However, the language is often non-idiomatic, some use of articles does not sound well; a native proofreader would help:

- page 11, line 6: *…the measure accepts if a mention is linked…*
- page 121, line 5: *…which concerns with…*

Errors in grammar or typos are rare:

- page 6, line 9: *are>is*
- page 11, line ^7: *if>when*
- page 21, last sentence: *Last but definitely no least*
- page 22, paragraph 4: *why this work mostly limit to these categories*
- page 74, last sentence: *An many Czech zeros…*
- page 83, line ^11: *unpropable*
- page 109, line 4: *This would inevitably caused…*

Conclusion

Coreference is a timely and important topic, theoretically interesting and practically useful. Its cross-lingual setting, employing advanced machine learning methods and sophisticated linguistic expertise, make the thesis a respectable piece of work with a significant creative contribution. Its results deserve attention due to their novel and revealing character and solid methodology. The linguistic and formal level is adequate. The author himself has already suggested some realistic and desirable potential of this work. Acknowledging his proven experience in the field the reader feels like wishing him good luck. The thesis clearly fulfills the requirements for a dissertation and the author has clearly demonstrated his abilities for independent scientific work.

26 August 2018

Alexandr Rosen