Dealing with Inconsistencies in Parallel Annotations

Ekaterina Lapshinova-Koltunski*, Sharid Loáiciga,**
Christian Hardmeier*, **Marie-Pauline Krielke***
Saarland University*, CLASP-University of Gothenburg**, Uppsala University***
e.lapshinova@mx.uni-saarland.de, sharid.loaiciga@gu.se, christian.hardmeier@lingfil.uu.se,
mariepauline.krielke@uni-saarland.de

We describe the problem of inconsistencies in annotations that appear when working with parallel corpora. In a broader scope, we are interested in automated methods of discovering any kind of inconsistencies that exist in the annotation of parallel data. In a more focused context, we deal with the automatic detection of inconsistencies in the manual annotation of coreference chains in the parallel corpus ParCorFull (Lapshinova-Koltunski et al. 2018). This corpus contains English-to-German translations that were annotated for full coreference chains. However, the annotation process itself was not performed in parallel - the English source and the German target texts were annotated independently. As a consequence, the texts in both languages were in some cases interpreted differently, resulting in variation of the category tags assigned to the annotated structures. For instance, the source text in example (1) contains one chain with the antecedent *moments of our life* and *them* referring to this antecedent. The annotation of the target text reveals some variation. First, the scope of the antecedent in both languages is marked differently (*moments of our life* in English vs. *die meisten Augenblicke unseres Lebens*-'most moments of our life' in German). Besides, the pronominal adverb *davon* in the target refers not to the antecedent *die meisten Augenblicke unseres Lebens* (like its equivalent *of them* in the source), but to another antecedent (*die psychologische Gegenwart*-'the psychological present'), building in this way a second chain.

1. EN: *I mean, most of [the moments of our life] - and I calculated - you know, the psychological present is said to be about three seconds long. Which means that, you know, in a life there, are about 600 million of [them]. In a month, there are about 600,000. Most of [them] don't leave a trace.*
   DE: *Ich meine, [die meisten Augenblicke unseres Lebens] - und ich habe das ausgerechnet - Sie wissen, <die psychologische Gegenwart> gilt als ungefähr drei Sekunden lang. Was bedeutet, wissen Sie, dass es in einem Leben ungefähr 600 Millionen <davon> gibt. In einem Monat gibt es ungefähr <600.000>. Die meisten von [ihnen] hinterlassen keine Spur.*

On the one hand, such inconsistencies may emerge when each side of the parallel data is annotated independently as explained and illustrated in (1). On the other hand, they can also be motivated by the following factors: a) Different interpretation of coreference chains by the translator, b) Language typological transformation processes, c) Translation errors. To our knowledge, none of the existing projects dealing with parallel annotation of discourse-related structures (Dipper and Zinsmeister, 2010; Zikánová et al., 2015; Grishina and Stede, 2017) address this issue. An automatic solution for detecting such inconsistencies applicable to all kinds of parallel annotations would be useful not only for corpus linguistics, but also for natural language processing. In this paper, we propose such a method. Applying automatic alignment tools on the annotated structures on the source and target sides, we detect possible inconsistencies in the annotations. Automatic identification of inconsistencies in annotated parallel data could be helpful for detecting translation and linguistic phenomena such as translation errors, translationese, language contrasts,

which are of great interest for translation studies, translator training and contrastive linguists. In this way, this can be seen as a bottom-up way to discover such phenomena. In our presentation, we will provide details on our methodology, and quantitative and qualitative analysis of the inconsistencies in the annotation of the English-German coreference chains in ParCorFull.

Keywords:
annotation inconsistencies, parallel annotation, coreference annotation, parallel data, English-German, automatic alignment of inconsistencies.

Thematic areas: 1 or 2

References

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), pages 41-50. Valencia, Spain: Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 423-428, Miyazaki, Japan. European Language Resources Association (ELRA).

Stefanie Dipper and Heike Zinsmeister. 2010. Towards a standard for annotating abstract anaphora. In *Proceedings of the LREC Workshop on Language Resource and Language Technology Standards state of the art, emerging needs, and future developments*, LREC10-W4, pages 54–59, Valletta, Malta. European Language Resources Association (ELRA).

Zikánová Šárka, Hajičová Eva, Hladká Barbora, Jínová Pavlína, Mírovský Jiří, Nedoluzhko Anna, Poláková Lucie, Rysová Kateřina, Rysová Magdaléna, Václ Jan. 2015. *Discourse and Coherence. From the Sentence Structure to Relations in Text.* Copyright © ÚFAL, Praha, Czechia, ISBN 978-80-904571-8-8, 274 pages.