Corpus Research on Multiword Discourse Markers for Raising Translation Awareness

Giedrė Valūnaitė Oleškevičienė [a[0000-0001-5688-2469]] and Chaya Liebeskind [b[0000-0003-0476-3796]],
[a]Mykolas Romeris university, Ateities 20, LT-08303, Vilnius, Lietuva; [b]Jerusalem College of Technology, 21 Havaad Haleumi str., 9116001, Jerusalem, Israel

Key words: multilingual corpus; multiword expression; discourse relation; discourse marker; translation.

1. INTRODUCTION

The development, research and application of discourse annotated corpora is a comparatively new research area which includes the study of discourse markers and requires the competences of scientists not only in creating and annotating texts, but also in exploring the application possibilities of texts annotated with discourse markers [1], [2]. Effective discourse management in any language is characterized by clear connections between sentences and a cohesive, coherent language structure. However, in different languages, the connections and structure of discourse are ensured by different linguistic means. Various dictionaries and grammar textbooks introduce the peculiarities of words and sentences, and the connections of discourse layer still lack being discussed. It should also be noted that discourse research raises awareness of pragmatic categories, not just typically relying on grammatical lists of conjunctions to describe certain functions of text cohesiveness and coherence [3]. Discourse markers are tools of discourse management and their functions include signposting, signalling, rephrasing, etc. Their importance affects language production, communication, second language learning, and translation. Dobrovoljc has recently researched multiword expressions as identifying structurally fixed discourse marking multiword expressions in a corpus of spoken Slovene [4]. According to Mona Baker (2011), during translation, the realities of a situation, the realities of the context, as well as language-specific aspects need to be considered [5]. Thus, the question that needs to be answered empirically is: In translation, what are the shifts of multi-word discourse markers in their lexical form? Establishing what lexical forms multiword discourse markers acquire in translation helps to produce a sound basis for future research investigating the possible reasons for the particular lexical forms in translation.

The current research examined multiword expressions used as discourse markers in English social media texts. We used transcripts of TED talks and compared them with their counterparts in Lithuanian and Hebrew. Our research the objectives were: to create a parallel corpus to identify multiword expressions used as discourse markers and to analyse their translations in Lithuanian and Hebrew. Our focus was to investigate if multiword expressions remain multiword or become one-word expressions in translation to Lithuanian and Hebrew so that to raise translator awareness in translation studies.

2. RESEARCH METHODOLOGY

The research process included three stages – the parallel corpus creation, establishing the candidates of multiword expressions potentially used as discourse markers, the extraction and the analysis of a sub-corpus with the established multiword expressions as discourse marker candidates. We decided to use TED Talk transcripts because they are publicly available and provide appropriate material for parallel data. To create a substantial parallel corpus containing data in English, Lithuanian, and Hebrew, talk transcripts were extracted automatically using a language-independent method that permits parallelizing data for any researched language. The talk transcripts were automatically extracted by using a special code which ensured that English sentences with the candidate discourse markers from the theoretically-based list were extracted and matched their Lithuanian and Hebrew counterparts. The process of the compiling of the parallel corpus could be considered innovative because it allows parallelizing data from any researched language. After the corpus creation, the variations of the translations of discourse markers into Lithuanian and Hebrew were extracted automatically for comparative study, identifying the variations in translation.

3. RESULTS AND CONCLUSIONS

The article discusses discourse research in relation to raising text coherence awareness in translation, and also to introduce the developed corpora resources. Therefore, the study first deals with the possibilities of expressing discourse relations by using multiword discourse markers as their linguistic realization in different languages, discussing possible choices of translators, taking into account the use of different linguistic means in translation. The article also presents the first research insights on comparing English, Lithuanian and Hebrew multiword discourse markers in order to understand translation tendencies at the discourse level.

English multiword expressions used as discourse markers demonstrate variability in Lithuanian and Hebrew translations: they either remain multiword expressions in the target languages or are translated as one inflected word, or omitted. In Hebrew translations, due to the nature of the Hebrew language, multiword discourse markers prevail and there is a clear tendency for translators to give preference to male derivatives [6]. However, in Lithuanian, there is a clear tendency for one-word discourse markers in translation. Lithuanian translations of pronoun-verb multi word expressions into one-word verb cases may be considered as almost word for word translations due to Lithuanian being a highly inflected (or null-subject) language [7] which fully represent the verb-pronoun cases. However, there are still cases where the subject is preserved in the Lithuanian translation and the discourse marker remains a multiword expression. Reflecting on why different discourse markers demonstrate different translation choices might be based on the nature of the target language into which the texts are translated, for example Lithuanian is rich in particles, and as the analysis has demonstrated, translators choose to integrate particles into discourse markers to mark the supplementary discourse expression. In addition, in English the gender is not expressed, thus when translating from English to Hebrew, the choice of the gender of the derivative is totally a translator’s choice. However, since in Hebrew male gender prevails, translators automatically give preference to male derivatives. Another observation for Hebrew is that multiword discourse markers remain multiword because of the translator choice to rely more on word for word translation; while in Lithuanian there is a tendency to omit the pronoun by using an inflected verb which is how multiword discourse markers become one-word discourse markers.

Concerning discourse layer, based on the results of the current study revealing cases where translators chose to insert particles in Lithuanian and connectives in Hebrew which bear an additional discourse meaning in the translation, that translator’s choices might be also guided by the internal discourse managing system of the target language.

REFERENCES

1. Webber, B., Prasad, R., Lee, A., & Joshi, A. (2016). A discourse-annotated corpus of conjoined VPs. *Proceedings of the 10th Linguistic Annotation Workshop Held in Conjunction with ACL 2016 (LAW-X 2016)*, 22–31. http://dx.doi.org/10.18653/v1/W16-1704
2. Zufferey, S., & Degand, L. (2017). Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*, *13*(2), 399–422. http://dx.doi.org/10.1515/cllt-2013-0022
3. Crible, L., & Degand, L. (2019). Domains and functions: A two-dimensional account of discourse markers. *Discours. Revue de Linguistique, Psycholinguistique et Informatique. A Journal of Linguistics, Psycholinguistics and Computational Linguistics*, *24*. http://dx.doi.org/10.4000/discours.9997
4. Dobrovoljc, K. (2017). Multi-word discourse markers and their corpus-driven identification: The case of MWDM extraction from the reference corpus of spoken Slovene. *International Journal of Corpus Linguistics*, *22*(4), 551–582. http://dx.doi.org/10.1075/ijcl.16127.dob
5. Baker, M. (2018). *In other words: A coursebook on translation*. Routledge.
6. Tobin, Y. (2001). Gender switch in modern Hebrew. *Gender across Languages: The Linguistic Representation of Women and Men, 1*, 177–198.
7. Zinkevičius, V., Daudaravičius, V., & Rimkutė, E. (2005). The Morphologically annotated Lithuanian Corpus. *Proceedings of The Second Baltic Conference on Human Language Technologies*, 365–370.

ISBN: 978-84-09-63206-0