▶ Processing Multi-Word Discourse Markers in Translation: English to Hebrew and Lithuanian

Giedrė Valūnaitė-Oleškevičienė,
Mykolas Romeris University, Ateities g. 20, LT-08303, Vilnius, Lietuva, gvalunaite@mruni.eu,
Chaya Liebeskind,
Jerusalem College of Technology, 21 Havaad Haleumi str., 9116001, Jerusalem, Israel, liebchaya@gmail.com

Purpose: It has been proved that multi-word expressions are of key importance in language generation and processing. They also could perform a function of discourse organization, and certain multi-word expressions operate as discourse markers. The purpose of the current research is to examine multi-word expressions used as discourse markers in TED talk English transcripts and compare them with their counterparts in their Lithuanian and Hebrew translations, identifying if English multi-word expressions used as discourse markers in social media texts remain multi-word expressions in Lithuanian and Hebrew translation and searching for reasons for the changes of discourse markers in translation. We follow the research question of how English multiword discourse markers are processed in Hebrew and Lithuanian translation.

Approach: To achieve the aim of the research, the set objectives were to create a parallel research corpus to identify multiword expressions used as discourse markers and to analyze their translations in Lithuanian and Hebrew to determine if they are also multiword expressions or one-word translations, or if they acquire any other linguistic forms, and look for the possible reasons for the translator choices. In the research, we combine the alignment model of the phrase-based statistical machine translation and manual treatment of the data in order to examine English multi-word discourse markers and their equivalents in Lithuanian and Hebrew translation by researching their changes in translation. After establishing the full list of multi-word discourse markers in our generated parallel corpus, we research how multi-word discourse markers are treated in translation. We apply the method of Corpus research and phrase-based statistical machine translation/research corpus available at LINDAT/CLARIN-LT repository http://hdl.handle.net/20.500.11821/34

Findings: Our research proves that the examined multi-word discourse markers have different translation tendencies due to the different grammars of the researched languages. There is a trend to remain multi-word in Hebrew translation, but due to the translation choices relying on inflections, they are one-word discourse markers in Lithuanian. There is also possible context-based influence guiding the translator to choose a particle or other lexical item integration in Lithuanian or Hebrew translated discourse markers to express the rhetorical domain, but the observed phenomenon of "over-specification" requires further research. Beyond the empirical research, an extensive parallel

data resource has been created to be openly used.

Value: The valuable outcome of the study was extending the available resources and providing linguistic processing for several languages by creating a multilingual parallel corpus (including English, Lithuanian, and Hebrew) based on social media texts; the created corpus is shared and interlinked via CLARIN open language resources.

Keywords: *Translation, corpus, multi-word expression, discourse relation, discourse marker.*

Research type: Research paper