Proceedings of the Informing Science + Information Technology Education Conference

An Official Publication of the Informing Science Institute
InformingScience.org

InformingScience.org/Publications

Online July 5 – 6, 2023

CORPUS PROCESSING OF MULTI-WORD DISCOURSE MARKERS FOR ADVANCED LEARNERS

Chaya Liebeskind*	Jerusalem College of Technology, Jerusalem, Israel	liebchaya@gmail.com
Giedrė Valūnaitė-Oleškevičienė	Mykolas Romeris University, Vilnius, Lithuania	gvalunaite@mruni.eu

* Corresponding author

ABSTRACT

Aim/Purpose	The most crucial aspects of teaching a foreign language to more advanced learners are building an awareness of discourse modes, how to regulate discourse, and the pragmatic properties of discourse components. However, in different languages, the connections and structure of discourse are ensured by different linguistic means which makes matters complicated for the learner.
Background	By uncovering regularities in a foreign language and comparing them with patterns in one's own tongue, the corpus research method offers the student unique opportunities to acquire linguistic knowledge about discourse markers. This paper reports on an investigation of the functions of multi-word discourse markers.
Methodology	In our 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 translations by researching their changes in translation. After establishing the full list of multi-word discourse markers in our generated parallel corpus, we research how the multi-word discourse markers are treated in translation.
Contribution	Creating a parallel research corpus to identify multi-word expressions used as discourse markers, analyzing how they are translated into Lithuanian and Hebrew, and attempting to determine why the translators made the choices add value to corpus-driven research and how to manage discourse.

The full paper has been published as the following and is being presented at this conference:

Liebeskind, C., & Valūnaitė-Oleškevičienė, G. (2023). Corpus processing of multi-word discourse markers for advanced learners. *Issues in Informing Science and Information Technology, 20,* 149-169. https://doi.org/10.28945/5144

Abstract published in *Proceedings of InSITE 2023: Informing Science and Information Technology Education Conference*, July 5-6 [online], Article 16. Informing Science Institute. https://doi.org/10.28945/5125

Findings	Our research proves that there is a possible context-based influence guiding the translation to choose a particle or other lexical item integration in Lithuanian or Hebrew translated discourse markers to express the rhetorical domain which could be related to the so-called phenomenon of "over-specification."
Recommendations for Practitioners	The comparative examination of discourse markers provides language instructors and translators with more specific information about the roles of discourse markers.
Recommendations for Researchers	Understanding the multifunctionality of discourse markers provides new avenues for discourse marker application in translation research.
Impact on Society	The current study may be a useful method to strengthen students' language awareness and analytic skills and is particularly important for students specializing in English philology or translation. Beyond the empirical research, an extensive parallel data resource has been created to be openly used.
Future Research	It should be noted that the observed phenomenon of "over-specification" could be analyzed further in future research.
Keywords	translation, corpus, multi-word expression, discourse, discourse marker

AUTHORS

Dr. Chaya Liebeskind is a lecturer and researcher in the Department of Computer Science at the Jerusalem College of Technology. Her research interests span both Natural Language Processing and Data Mining. Especially, her scientific interests include Semantic Similarity, Language Technology for Cultural Heritage, Morphologically Rich Languages (MRL), Multi-Word Expressions (MWEs), Information Retrieval (IR), and Text Classification (TC). Much of her recent work has been on analyzing discourse markers. To this end, the researcher's released corpus comprises parallel alignments of TED Talk scripts in multiple languages. The researcher devoted great attention to the analysis of MWE's use as discourse markers. The researcher has published a variety of previous articles and several additional articles are under review or in preparation. The researcher is a member of several international research actions funded by the EU.

Dr. Giedrė Valūnaitė Oleškevičienė is a professor at the Institute of Humanities, Mykolas Romeris University. Her scientific interests in the domain of humanities include discourse analysis, discourse annotated corpora, professional English, and legal English, and in the domain of social sciences, and educational science her scientific interests include social research methodology, modern education, philosophical issues, creativity development in modern education system, etc. The researcher is actively engaged in second language teaching and learning research, linguistics, and translation research. The researcher coordinated international research projects funded by the EU, published scientific articles, and participated as a presenter in scientific conferences.