LDK 2023

Multi-word Expressions as Discourse Markers in Multilingual TED-ELH Parallel Corpus

Giedrė Valūnaitė Oleškevičienė
Institute of Humanities
Mykolas Romeris university
Ateities 20, LT-08303
Vilnius, Lietuva
gvalunaite@mruni.eu

Chaya Liebeskind
Department of Computer Science
Jerusalem College of Technology
21 Havaad Haleumi st., 9116001
Jerusalem, Israel
liebchaya@gmail.com

Abstract

In this paper, we present the outcome of the research inspired by the Nexus Linguarum network. As a theoretical basis, we discuss the multi-word word expressions as a part of the formulaic language used as discourse markers for organizing discourse. We also identify that parallel research in multiple languages may provide inter-lingual insights. We created a parallel multilingual corpus TED-ELH for our research and applied a parallel corpus alignment algorithm to extract multi-word discourse markers and their translations in Lithuanian and Hebrew. The analysis of the translations of multi-word discourse markers allowed us to identify that they demonstrate certain variability and either remain multi-word expressions or turn into one-word translations due to the linguistic characteristics of the target languages.

1 Introduction

One of natural language processing (NLP) research trends focuses on textual coherence including the relatedness of dialogical speech and also discourse relations between sentences and bigger pieces of text. Discourse relations both explicit and implicit facilitate a better understanding of the underlying relations among ideas in spoken or written texts. While implicit discourse relations could be inferred relying on the surrounding context, explicit discourse relations are realized through explicit discourse markers that belong to a number of linguistic classes including multi-word expressions. Currently, the researchers are working on both monolingual and multilingual resources. Monolingual studies and the development of the resources of discourse makers (Prasad et al., 2014; Webber et al., 2016) gave rise to multilingual studies creating multilingual corpora and comparing the use of discourse markers in various languages (Stede et al., 2016; Zufferey, 2016; Oleskeviciene et al., 2018; Zeyrek et al., 2019).

The purpose of the current study is extending the available resources working towards low-resource languages and providing linguistic processing for several languages by creating a multilingual parallel corpus (including English Lithuanian and Hebrew) based on social media texts and working on multi-word expressions in social media texts by exploring how multi-word expressions are used as discourse markers and if they remain multi-word expressions in the languages of the TED-ELH Parallel Corpus.

2 Related research

The rise of corpus linguistics and NLP brought the understanding that formulaic language plays an important role and that language users have memorized sequences which enable language generation process (Biber et al., 1999). In fact formulaic language is used as an umbrella term which covers collocations, idioms, lexical bundles or multiword expressions and etc. Lexical bundles or multiword expressions often perform discourse organizing functions (Biber et al., 2004) so in such cases they operate as discourse markers. As discourse markers signal discourse relations and organization researchers expect that obtaining parallel findings in different languages may serve as substantial evidence of discourse marker discourse organizing role (Zufferey, 2016). This generated research focusing on cross-linguistic mapping of discourse markers (Nedoluzhko and Lapshinova-Koltunski, 2018; Meyer and Poláková, 2013). The insights in semantic provided by Noel (Noël, 2003) stress the importance of cross-linguistic and translation studies of discourse markers as such approach may give light on contextual dimensions of the researched discourse markers. Evers-Vermeul et al. (Evers-Vermeul et al., 2011) identify that translation correspondence of discourse markers may provide the information on the pragmatic content because usually certain translator choices are guided by certain

meanings which guide the translator while looking for the equivalents or making the corresponding choices in the target context.

The research on coherence relations also stimulated research on multi-word expressions used as discourse markers (Dobrovoljc, 2017). Initially, only secondary status was given to multi-word expressions serving as discourse markers and performing pragmatic functions in corpus linguistics research. However, Wray (Wray, 2013) pointed out that multi-word discourse markers require empirical research and reconsideration. Corpus-driven research on formulaic language led to understanding that certain multi-word expressions perform discourse signaling and organizing function (Csomay, 2013; Schnur, 2014).

3 Methodology

First, the parallel texts in English, Lithuanian, and Hebrew were extracted from TED talks by using the transcripts, and then the sentences were aligned to make a parallel corpus for further research. The corpus contains 87230 aligned sentences (published in LINDAT/CLARIN-LT repository). Then further, we focused on multi-word expressions and narrowed our research focusing on multi-word expressions which are used as discourse markers to ensure textual cohesion and according to Fraser (Fraser, 2009) relate separate discourse messages, for example, such phrases as *you know, I mean, of course,* etc. which are characteristic of spoken language (Furkó and Abuczki, 2014; Huang, 2011). Thus, 3314 aligned sentences containing the earlier mentioned multi-word expressions were extracted and then manually annotated spotting the cases when the expressions are used as discourse markers, for example in case (1) the multi-word expression you know is used to introduce a new discourse message, while in case (2) they are content words fully integrated into the sentence.

1. You know, I'm not even ashamed of that.

2. You know the little plastic drawers you can get at Target.

After that, the variations of the translations of discourse markers into Lithuanian and Hebrew were extracted for comparative study spotting out the variations in translation.

4 Research findings

At the initial stage of the research the manual annotation revealed the distribution of multi-word expressions used as discourse markers and content words (see Figure 1). The research revealed that some multi-word expressions are used as discourse markers more often while other multi-word expressions have a tendency to remain content words in the research corpus. The most frequent multiword expressions used as discourse markers appear to be *I think* and *you know*. It is visible in Figure 1 that such multi-word expressions as *that is* or *you see* are seldom used as discourse markers in the researched corpus, instead they are mostly content words.

Also it was identified that English multi-word expressions used as discourse markers demonstrate variability in Lithuanian and Hebrew translations: they are either translated into multi-word expressions or in one inflected word in the target languages or are omitted at all. For example, in Lithuanian multi-word expression discourse marker *you know* splits into a number of multi-word expressions and also one-word translations. Multi-word expressions could be classified into cases representing pronoun-verb phrase jūs žinote (you know), jūs suprantate (you understand), jūs įsivaizduojate (you imagine), jūs esate girdėję (you have heard) or particle-verb phrase: (na (well), juk (after all), ir (and)) žinote (you know), suprantate (you understand), or connective-verb phrase (kaip (how), kad (that)) žinote (you know), matote (you see) where connective could be used in a pre- or post- position to the verb.

One-word translations mainly include verbs, for example, žinote (you know), suprantate (you understand), įsivaizduojate (you imagine), and etc., which due to Lithuanian being a highly inflected language (Zinkevičius et al., 2005) fully represent the verb-pronoun cases. It should be noted that Lithuanian translations of pronoun-verb multiword expressions and one-word verb cases could be considered as almost word for word translations. It could be said that more interesting cases which represent translator choices of particle-verb or connective-verb multi-word expressions which due to the use of particles and conjunctions also carry out certain rhetorical discourse meaning which needs to be researched further.

In Hebrew multi-word discourse marker translations demonstrate the tendency to remain multi-

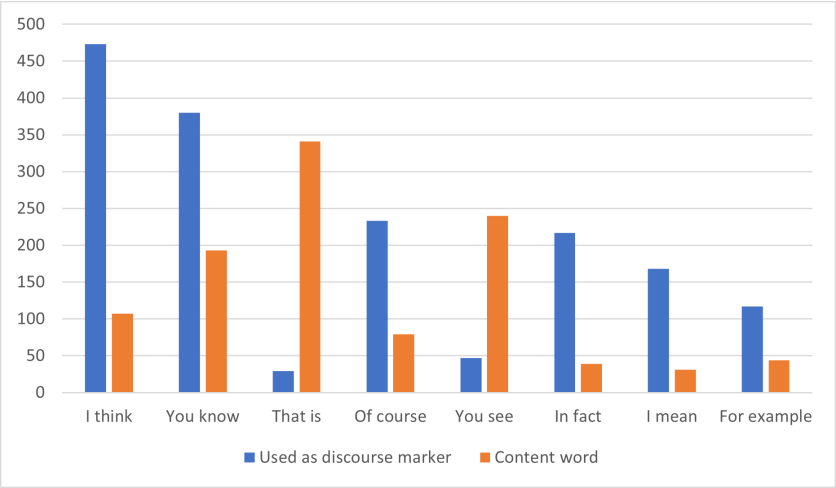

Figure 1: Multiword expressions used as discourse markers and content words

word discourse markers with a little number of one word translations. The distinctive pattern in Hebrew is the prevalence of male gender in discourse marker translations, for example, the translations of the discourse marker *you know* are mostly expressed using male gender in plural **אתם יודעים** and in singular **אתה יודע** which reveals that the translators demonstrate preference for male gender in their translation choices. Similarly to Lithuanian there are cases in Hebrew translation when a connective is added to the multi-word expression for example, **ואנו יודעים** (and we know) which also relate to the rhetorical discourse nature so further research is required to investigate the cases of additional particles and connectives used in the translation.

5 Conclusions

In conclusion, the analysis of multi-word expressions used as discourse markers identifies that there is a certain distribution of multi-word expressions used as discourse markers in the researched corpus. The analyzed multi-word expressions fall into two groups: the multi-word expression with the tendency of being used as discourse markers in the researched corpus and the multi-word expressions with the tendency of being used as content words in the researched corpus.

The initial research also reveals that in Ted talks translated transcripts English multi-word discourse markers may be translated into one-word expression probably due to the rich in inflections target languages of the research. The analysis of the translations of the multi-word expressions used as discourse markers in Lithuanian and Hebrew reveals that there is a tendency in Lithuanian to turn them into one word discourse markers due to translator preferences to use inflected verb forms. While in Hebrew the tendency is to keep the multi-word form of discourse markers just mainly choosing the male gender both in singular and plural forms of the discourse marker translations which could be socio-culturally guided translator choice.

There are also cases of additional particles and connectives used in the translation of multi-word expressions both in Lithuanian and Hebrew. Such translator choices could be guided by the contextual pragmatic features; however, further research is needed to investigate the cases further. The mentioned cases are interesting for the research as they require insights and specific annotation to investigate which contextual pragmatic factors guided the translator choices.

The corpus building method and the extraction method of the multi-word expressions used as discourse markers tested on social media texts such as TED talks scripts can be applied to other languages. Also, it relates to expanding resources by working towards low-resource languages as the parallel corpus embracing English, Lithuanian, and Hebrew was build and it could be used as a resource for multiple scientific research.

Acknowledgements

This study is based upon work from COST Action NexusLinguarum - European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology, https://www.cost.eu/).

References

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3):371–405. Publisher: Oxford University Press.

Douglas Biber, Stig Johansson, Geoffrey Leech, S. Conrad, Eclwarcl Finegan, and Randolph Quirk. 1999. Longman. *Grammar of spoken and written english*.

Eniko Csomay. 2013. Lexical bundles in discourse structure: A corpus-based study of classroom discourse. *Applied linguistics*, 34(3):369–388.

Kaja Dobrovoljc. 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.

Jacqueline Evers-Vermeul, Liesbeth Degand, Benjamin Fagard, and Liesbeth Mortier. 2011. Historical and comparative perspectives on subjectification: A corpus-based analysis of dutch and french causal connectives. *Linguistics*, 49(2):445–478.

Bruce Fraser. 2009. An account of discourse markers. *International review of Pragmatics*, 1(2):293–320. Publisher: Brill.

Péter Furkó and Ágnes Abuczki. 2014. English discourse markers in mediatised political interviews. *Brno Studies in English*, 40(1).

Lan Fen Huang. 2011. *Discourse markers in spoken English: A corpus study of native speakers and Chinese non-native speakers*. PhD Thesis, University of Birmingham.

Thomas Meyer and Lucie Poláková. 2013. Machine translation with many manually labeled discourse connectives. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 43–50.

A Nedoluzhko and E Lapshinova-Koltunski. 2018. Pronominal adverbs in german and their equivalents in english, czech and russian: Evidence from the parallel corpus. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue"(Moscow*, pages 522–532.

Dirk Noël. 2003. Translations as evidence for semantics: an illustration. *Linguistics*, 41(4):757–785.

Giedre Valunaite Oleskeviciene, Deniz Zeyrek, Viktorija Mazeikiene, and Murathan Kurfalı. 2018. Observations on the annotation of discourse relational devices in TED talk transcripts in Lithuanian. In *Proceedings of the workshop on annotation in digital humanities co-located with ESSLLI*, volume 2155, pages 53–58.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950. Publisher: MIT Press.

Erin Schnur. 2014. Phraseological signaling of discourse organization in academic lectures: A comparison of lexical bundles in authentic lectures and eap listening materials. *Yearbook of Phraseology*, 5(1):95–122.

Manfred Stede, Stergos Afantenos, Andreas Peldzsus, Nicholas Asher, and Jérémy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1051–1058.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A discourse-annotated corpus of conjoined vps. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31.

Alison Wray. 2013. Formulaic language. *Language Teaching*, 46(3):316–334.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (tedmdb): A parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–27.

Vytautas Zinkevičius, Vidas Daudaravičius, and Erika Rimkutė. 2005. The Morphologically annotated Lithuanian Corpus. In *Proceedings of The Second Baltic Conference on Human Language Technologies*, pages 365–370.

Sandrine Zufferey. 2016. Discourse connectives across languages: factors influencing their explicit or implicit translation. *Languages in Contrast*, 16(2):264–279.