Issues in Informing Science + Information Technology

An Official Publication of the Informing Science Institute InformingScience.org

IISIT.org

Volume 20, 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.

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."

Accepting Editor: Eli Cohen │ Received: January 25, 2023 │ Revised: May 24, 2023 │ Accepted: June 4, 2023

Cite as: 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

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

INTRODUCTION

Foreign language acquisition via the use of corpora has been extensively researched over the last few decades (Aijmer, 2009; Bunt & Prasad, 2016; Gessler et al., 2021; Sinclair, 2004; Zeldes et al., 2021). It is emphasized that corpora are crucial resources for both indirect and direct usage by foreign language learners (McEnery & Xiao, 2011; Römer, 2008). Bilingual parallel corpora are beneficial to foreign language instruction because they allow the comparison of language patterns in foreign and native languages, fostering an understanding of their similarities and differences and facilitating the analysis and evaluation of translations.

Helping students become aware of discourse modes – how to regulate discourse, and the pragmatic properties of discourse components – is one of the most crucial aspects of teaching a foreign language to more advanced learners. This implies that they should be able to comprehend and use the appropriate language tools to represent knowledge logically and articulate the pragmatic aspects of communication. These types of gadgets are termed "discourse markers" in scientific writing. Although their primary function is to ensure that communication makes sense, they may also be utilized to convey how one feels about the message and the recipient (Crible & Degand, 2019).

Research on discourse markers in translation has been acknowledged as a way that may provide certain insights into the different shades of meanings and functions of discourse markers. Noël (2003) stressed the value of translation corpora in discourse marker research getting into certain linguistic dimensions. It is discussed by Biber et al. (2004) that the produced language includes a high percentage of lexical bundles and that lexical bundles among many functions demonstrate a function of organizing discourse this way, functioning as discourse markers. For example, such bundles as *I think,* etc. relate to the research on discourse markers because such phrases as *you know*, *I think*, etc. perform certain discourse organizing functions and have also been classified as discourse markers in discourse research. Research on discourse markers as tools of discourse management proves that they carry several functions, including signposting, signaling, and rephrasing. Furthermore, there are ongoing attempts to investigate the importance of discourse layers in language production, communication, second language learning, and translation. Additionally, Dobrovoljc (2017) has recently attempted to research multi-word expressions as discourse markers in a corpus of spoken Slovene, identifying structurally fixed discourse marking multi-word expressions.

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, focusing on the cases when English multi-word expressions used as discourse markers in social media texts remain multi-word expression in Lithuanian and Hebrew translation, and searching for reasons for the changes of discourse markers in translation. The research was carried out focusing mainly on the structure of the phrases, 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 they did were the objectives of the research, as parallel data add value to corpus-driven research and teach learners how to manage discourse. An additional benefit of the study was extending the available resources to 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.

THEORETICAL BACKGROUND

The related research provides an overview of an account of the research languages, multi-word expressions and their use as discourse markers, the importance of discourse markers for discourse management, and certain insights into discourse marker translation.

CULTURAL HERITAGE AND LANGUAGES OF THE RESEARCH

It is important to briefly address the cultural legacy of the research languages, which in some ways drove the selection of languages for our study. From the first part of the 14th century, Jewish and Lithuanian civilizations coexisted in the same region, according to Bieliauskienė (2012). The author emphasized that beginning in the nineteenth century, Vilnius was known as Lithuania's Jerusalem, drawing competent individuals in the area of education and stimulating a thriving high culture, such as theater, art, and literature. In reality, both languages, Lithuanian and Hebrew, contributed to the region's cultural history.

In this work, we investigate the Lithuanian and Hebrew parallel corpora alongside pivotal English. Lithuanian is an ancient Baltic language that retains forms linked to Sanskrit and Latin while maintaining the majority of Proto-Indo-European phonological and morphological features. As a result, it has achieved prominence in Indo-European language studies and has been studied by several scholars, including Ferdinand de Saussure, who referred to Lithuanian as "the Galapagos of linguistic evolution" (Joseph, 2009). Lithuanian is rich in declensions and cases inside declensions, and the earliest layer of the Lithuanian language lexicon is tied to the Indo-European language, which is over 5,000 years old.

Hebrew is a Northwest Semitic language that belongs to the Canaanite branch of the Afroasiatic language family. It was the native language of the Israelites and was regularly used by their descendants, the Jews and the Samaritans, until after 200 CE, when it became extinct. However, Hebrew was mainly preserved as a liturgical language in Judaism and Samaritanism. Since the 19th century, when Hebrew stopped to be a dead language, it functions as a large-scale example of a successful linguistic revival (Joslyn-Siemiatkoski, 2007). Hebrew is also an essential language for scholars who study Middle Eastern civilizations and Christian theology.

The Hebrew alphabet, also known as the Ktav Ashuri, Jewish script, square script, and block script by various scholars, is an abjad script used to write the Hebrew language. It is a descendant of the Imperial Aramaic alphabet, which thrived during the Achaemenid Empire and derived from the Phoenician alphabet. The Hebrew alphabet has 22 letters, while the English alphabet has 26. It is lacking case. Five letters take on distinct forms at the end of a word. Hebrew is written from right to left.

MULTI-WORD EXPRESSIONS AS DISCOURSE MARKERS

The effectiveness of spoken communication is strongly dependent on the proper use of discourse markers. Consequently, their acquisition is of the utmost significance in foreign language learning, particularly at more advanced levels. Acquisition of discourse markers enables learners of a second language to better comprehend and interpret the speech of native speakers as well as generate more fluent and coherent speech themselves. However, the multifunctionality and culture-bound character of discourse markers make it difficult for foreign language learners to acquire them.

NLP recognizes that language is not just placing words in the right order but getting the meaning and deeper textual relations as well as organizing ideas into a logical textual flow. According to researchers (Barlow, 2011; Sinclair, 1991), language is not just generated according to compositional principles; it is also formulaic. Speakers possess multiple learned formulaic sequences which, according to Siyanova-Chanturia et al. (2011), are important in organizing discourse and help the language producer and recipient to manage language processing. However, formulaic language is not easy to manage and categorize for NLP research, as it may seem at first sight, since the sequences that could be considered formulaic vary in length, meaning, fixedness, etc., and the consensual definition of formulaic language has not fully crystallized. It could be considered as an umbrella term embracing idioms, proverbs, clichés, phrasal verbs, collocations, and lexical bundles (Wray, 2012). According to Wei and Li (2013), formulaic language covers approximately 60% of written texts in their researched corpus of the English academic language. According to Biber et al. (1994, 1999), lexical bundles are groups of words that show a statistical tendency to co-occur and could be considered as extended collocations, for example, *I think*. Biber et al. (2004) identify that lexical bundles have functional purposes, such as organizing discourse, expressing stance, and referential meaning.

Another important issue in NLP is discourse management, which is related to discourse relations, connecting ideas between sentences and bigger parts of the text. Based on the evidence of the formulaic nature of language for communication, research has turned to investigating multi-word expressions used as discourse markers (Dobrovoljc, 2017), identifying structurally fixed discourse marking multi-word expressions. Discourse relations may remain implicit or be expressed explicitly through discourse markers, which help textual coherence and discourse management, and are used for making coherent speech appropriately segmented to enable textual understanding. In the current study, we rely on the definition formulated by Crible and Degand (2019) that discourse markers form an open-class category of pragmatic expressions generally defined by two main features: syntactic optionality or weak clause association, and procedural meaning based on Fraser (1996), Schiffrin (1987) and Schourup (1999). Discourse markers are connecting discourse elements, which form a category of lexical elements used to identify relations between discourse segments and provide text coherence, such as explanation, contrast, and so on (Crible & Degand, 2019; Sanders & Noordman, 2000). Discourse markers perform important functions, such as signposting, signaling, and rephrasing, by facilitating discourse organization. They are mainly drawn from syntactic classes of conjunctions, adverbials, and prepositional phrases (Fraser, 2009), as well as expressions such as *you know*, *you see*, and *I mean* (Hasselgren, 2002; Schiffrin, 2001). Hasselgren (2002) advocated that discourse markers signal better fluency, which contributes to interaction and even makes the speaker sound more 'native-like'. Recently, discourse relations and discourse maker research has gained certain impetus with corpora annotation for exploring discourse structure in texts; for example, the Penn Discourse Tree Bank (PDTB) (Webber et al., 2016). Furthermore, there was a rise in annotated multilingual corpora for researching different means of expressing discourse relations and managing discourse (Gessler et al., 2021; Oleškevičienė et al., 2018; Stede et al., 2016; Zeldes et al., 2021; Zeyrek et al., 2020; Zufferey & Degand, 2017).

Language, especially spoken, is characterized by discourse marker use; however, some of them (e.g., *you know, I think, well*) are sometimes referred to in a critical manner, indicating a lack of fluency (O'Donnell & Todd, 2013). Still, discourse markers are abundantly used and, according to Crystal (1988), they enhance communication if used appropriately and should not be considered unnecessary or undesirable. As Biber (2006) observed, discourse markers, such as *you know*, or *well*, are very rare in written language. However, they are quite common in spoken discourse and should not be treated as just fancy words since they serve the function of organizing discourse by signaling, rephrasing, marking, or relating ideas. Svartvik (1980) observed that, if a foreign language learner makes a mistake (e.g., *he goed*), it can be easily identified and redeemed by the native speaker; however, if a learner uses words such as *you know*, or *well* inappropriately, the native speaker cannot identify any error and the speech might sound impolite or even dogmatic. The same idea is also supported by Hasselgren (2002), who observed that discourse markers enhance interaction. Furthermore, it has also been researched using learner corpora to demonstrate the importance of discourse-level knowledge, especially at more advanced levels of language learning (Cobb & Boulton, 2015; Granger, 2015).

Translation Issues of Discourse Markers

Discourse markers are used in both written texts and spoken discourse to connect ideas and guide the reader or the listener through expression by ensuring that the ideas are grasped correctly. 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). Discourse markers have been researched by applying various theoretical approaches, such as Rhetorical Structure Theory (Mann & Thompson, 1988), Segmented Discourse Representation Theory (Asher & Lascarides, 2003), and Penn Discourse Treebank (PDTB) (Prasad et al., 2008), first focusing on the monolingual approach, which resulted in multilingual studies focusing on translation (Das & Taboada, 2019; Degand & Pander Maat, 2003; Dixon, 2009; Pit, 2007; Zufferey & Cartoni, 2012). As Zufferey and Cartoni (2012) observed, multilingual studies are more complicated as languages differ in the use of discourse markers and their expression. The authors also added that often discourse markers are poly-semic, which means that a single expression of a discourse marker may perform in expressing various discourse relations. They provided an example of the English *since*, which could express temporal or causal discourse relations depending on the surrounding contexts.

Recently, much research has gained interest in using parallel translated corpora. For example, Dupont and Zufferey (2017) focused on the investigation of translation corpora to study if the effect of register, translation direction, or translator's expertise could influence the shifts of meaning and omissions of English and French markers of concession. Hoek et al. (2017) investigated a parallel corpus of English parliamentary debates translated into Dutch, German, French, and Spanish, searching for what types of discourse markers might have a higher tendency to be more frequently omitted in the translation. Baker (2018), in her extensive studies on translation, observed that discourse markers could be used to signal different relations and these relations could be expressed by a variety of means. The author provided the example that, in English, the expression of causality could be realized through content verbs, such as *cause* or *lead*, or more simply, through a discourse marker signaling the causality relation. Further, different languages demonstrate different tendencies – some languages prefer using simpler structures connected by a variety of discourse markers, while other languages favor complex structures, sparsely using explicit discourse markers.

Therefore, translation poses a challenge in adapting various preferences of the source and target languages. Translators face various choices of inserting discourse markers to make the flow of the ideas smoother in the target text; however, they risk making the translation sound foreign or transposing the grammatical syntactic structure, ending up using different means of expressing discourse markers or simply omitting them. The phenomenon of the implication of discourse markers across languages, including the Lithuanian language, is analyzed by Crible et al. (2019). It appears that it is not always possible to use the word-for-word technique and natural changes in translation are sometimes inevitable. The insights in semantics provided by Noël (2003) stress the importance of cross-linguistic and translation studies of discourse markers as such an approach may give light on contextual dimensions of the researched discourse markers. Evers-Vermeul et al. (2011) identify that translation correspondence of discourse markers may provide information on the pragmatic content because usually certain translator choices are guided by certain meanings that guide the translator while looking for the equivalents or making the corresponding choices in the target context.

RESEARCH METHODOLOGY

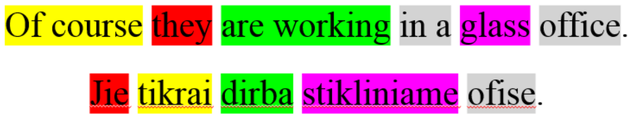
To achieve the research aim of examining multi-word expressions used as discourse markers in TED Talk English transcripts and comparing them with their counterparts in Lithuanian and Hebrew, there was a need to achieve the double objectives of creating the parallel corpus for the research data and carrying out the research on multi-word expressions used as discourse markers in the studied languages. A recently proposed methodology for annotating and researching parallel corpora was adapted for the current study (Montechiari et al., 2022; Silvano et al., 2022). Initially, the list of multi-word and one-word expressions that could potentially be used as discourse markers were generated relying on theoretical insights by Schiffrin (1987), supported by further analysis of phrases such as *you know*, *I mean*, *of course* as characteristic of spoken language (Furkó & Abuczki, 2014; Huang, 2011), and the classification provided by Fraser (2009). Fraser's extensive classification was taken as a basis, and Huang's (2011) theoretical analysis of discourse marker characteristics for spoken discourse, for example, *you know, you see, I mean, I think*, was also included.

First, a parallel corpus meeting the research aim needed to be created. We decided to use TED Talk transcripts, as they are publicly available and provide appropriate material for parallel data. In order to create a substantial parallel corpus containing data in English, Lithuanian, and Hebrew, the talks were extracted automatically using a special code, which ensured that English sentences with the candidate discourse markers from the theoretically based list were extracted and matched with their Lithuanian and Hebrew counterparts. The process of creating the parallel corpus could be viewed as an innovative achievement as it allows parallelizing the data of any researched languages. While building the corpus, the parallel texts in English, Lithuanian, and Hebrew were extracted from TED talk transcripts. Then, the sentences were aligned to make a parallel corpus for further research. The corpus contains 87,142 aligned sentences (published in LINDAT/CLARIN-LT repository, http://hdl.handle.net/20.500.11821/34).

Another stage of the research focuses on multi-word expressions that are used as discourse markers to ensure textual cohesion and, according to Fraser (2009), to relate separate discourse messages. The author provides extensive analysis and examples of multi-word discourse markers relating to separate discourse messages. For research purposes, 3,314 aligned sentences containing the earlier mentioned multi-word expressions were extracted and manually annotated, spotting the cases in which the expressions were used as discourse markers. One-word discourse marker identification did not represent many challenges; however, turning to multi-word expressions certainly caused challenges. For example, to identify if the expression *you know* is used as a connective, the context in which it occurs should be examined by identifying if the expression serves as a discourse marker. As such, two situations arise: (1) the multi-word expression *you know* is used to introduce a new discourse message, or (2) they are content words fully integrated into the sentence.

1. You know, but really, it's the kind of same old crap we've had for the last 30 years.
2. I'll let you know when you can look again.

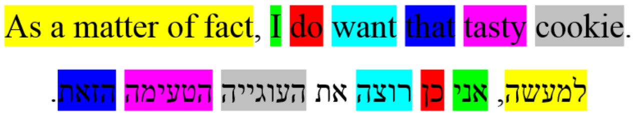
After that, the variants of the translations of discourse markers into Lithuanian and Hebrew were extracted automatically for a comparative study, determining the variations in translation. We ran an NLP word-alignment algorithm to extract a phrase table of all the possible translations of the researched discourse markers, using our parallel corpus (in our case, source = English, target = Lithuanian/Hebrew). The extraction of the translation variations was dependent on the phrase-based statistical machine translation model introduced by Koehn et al. (2003). The model could be visually represented in the research languages by the figures below. Figure 1 visualizes Lithuanian–English corresponding phrases marked in respective colors. Figure 2 shows English–Hebrew respective phrase alignment, with a note for the reader that Hebrew text should be read from right to left.



Of course they are working in a glass office.

Jie tikrai dirba stikliniame ofise.

Figure 1. English – Lithuanian phrase alignment



As a matter of fact, I do want that tasty cookie.

למעשה, אני כן רוצה את העוגייה הטעימה הזאת.

Figure 2. English – Hebrew phrase alignment

The model applies the segmentation of the input into sequences of words, which are called phrases, and then each phrase is translated into target language phrases that could later be reordered in the output. Such a model ensures the correspondence between the units of phrases.

After being extracted, all the possible phrase variants were manually filtered to eliminate any inconsistencies and to prepare the data for the machine analysis stage. This helped us extract sentences with translations of the researched discourse markers from the target language corpus and analyze their use.

While analyzing the data, we noticed that there was a small amount of data related to unexpected combinations of discourse markers with integrated particles both in Lithuanian and Hebrew translations. We collected all the cases and annotated them using the cross-domain function taxonomy of discourse markers by Crible (2017) (Table 1). The inter-annotator agreement was reliable with Cohen's kappa = 0.78. Then we analyzed the annotated data searching if certain domains and functions might be related to the translator's choices to integrate additional particles or other lexical items into the translation of discourse markers.

Table 1. Cross-domain functions taxonomy (Crible & Degand, 2019)

IDEATIONAL	RHETORICAL	SEQUENTIAL	INTERPERSONAL
Addition / alternative / cause / closing / concession / condition / consequence / contrast / enumeration / opening / punctuation / resuming / temporal / topic-shift / specification			

Table 1 schematically represents the functional taxonomy (Crible, 2017) used for the annotation in the current study. The taxonomy describes discourse markers as functioning in four domains which include: the ideational domain related to real-world events; the rhetorical domain related to the speaker's expressed subjectivity and meta-discursive effects; the sequential domain concerning the structuring of local and global units of discourse; and the interpersonal domain related to managing the speaker-hearer relationship. These four domains correspond to the overall discourse intentions or entities, which depend on what the speaker is targeting: content (the ideational domain); illocutionary value (the rhetorical domain); discourse structure (the sequential domain); or inter-subjective inferences (the interpersonal domain) (Crible, 2017). The functions can perform in any domain which is why the taxonomy is identified as a cross-domain one.

Research Findings

Multi-Word Discourse Markers in the Corpus

While analyzing the extracted multi-word expressions, first we examined how frequent they are in the research corpus. The list of frequencies of multi-word expressions used in the study corpus has been extracted and is presented in Figure 3.

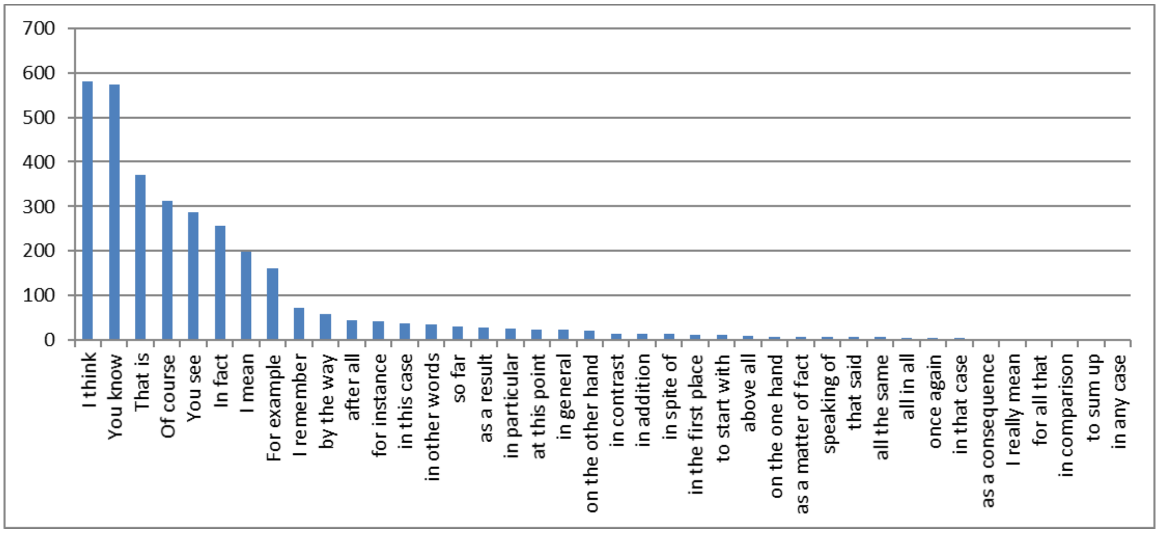


Figure 3. Multi-word expressions and their frequency in the corpus

It can be seen in Figure 3 that the two most frequent multi-word expressions in the corpus are *I think* and *you know* followed by the other multi-word expressions.

As mentioned earlier, multi-word expressions needed to be manually annotated, spotting the cases when the expressions were used as discourse markers. The manual annotation revealed that some multi-word expressions are used as discourse markers more frequently, while others tend to be used as content words fully integrated into sentences. Figure 4 presents the multi-word expressions which have a tendency to be used as discourse markers rather than content words and their distribution in their use as discourse markers.

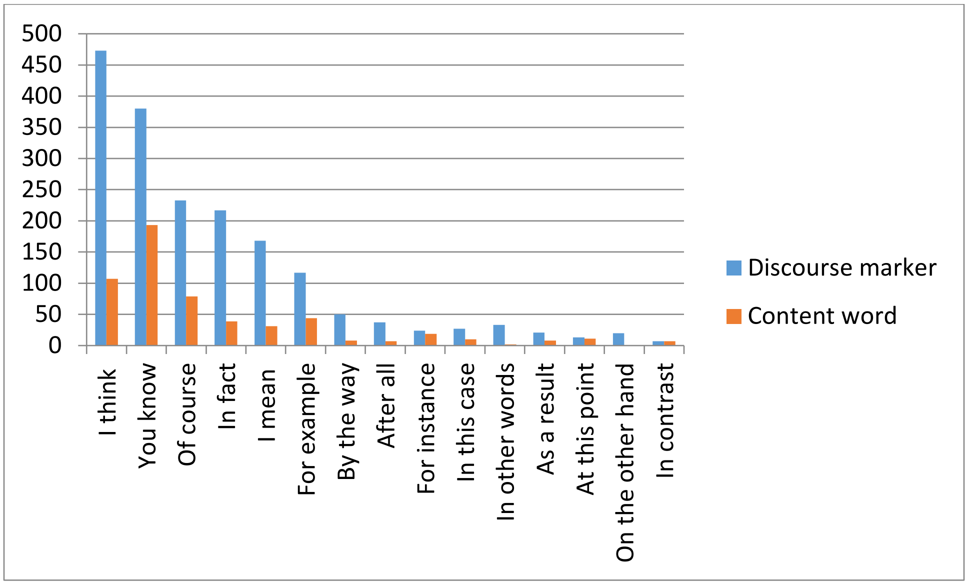


Figure 4. Multi-word expressions used as discourse markers

The high numbers of the discourse marker *you know* could be explained by the fact that in the English language, we may say "well, y'know…" which is rooted in the way English speakers talk which is less likely to follow the same ways in other languages. Another reason could be that language can serve an attention function and the interlocutors may aim to keep the communicator engaged in the interaction by not directly requiring a communicative response but just following the function of keeping the other person's attention.

The other group of multi-word expressions includes the multi-word expressions used as discourse markers less frequently and more often used as content words fully integrated into sentences. Figure 5 shows that although the multi-word expressions *that is, you see, so far, in particular, in general, in the first place* and *to start with* are identified as discourse markers by the theoretical literature, in the current corpus of this study they demonstrate a weak tendency to be used as discourse markers and are primarily used as content words.

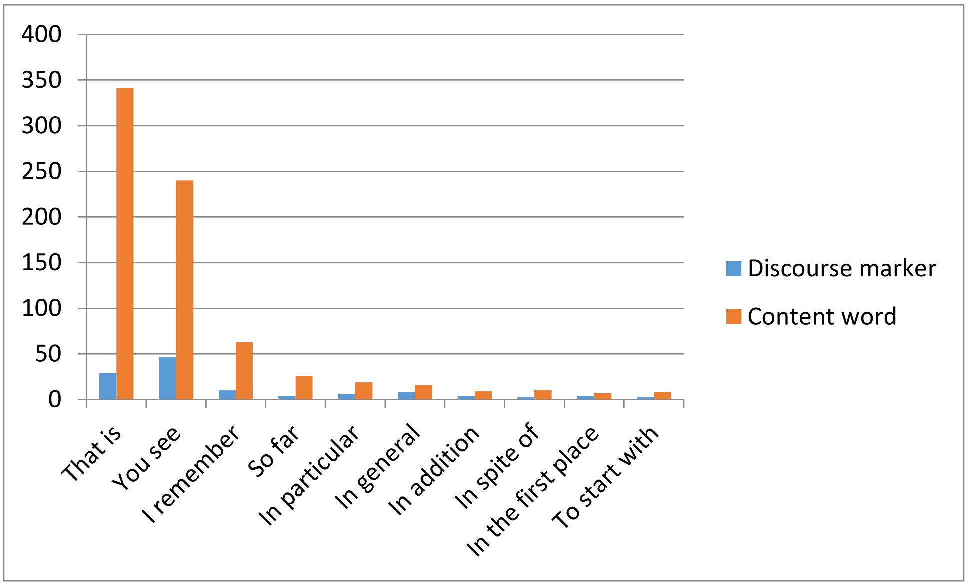


Figure 5. Multi-word expressions used as content words

Discourse Marker Combinations in Lithuanian and Hebrew Translation

While analyzing the translation of discourse markers, we noticed that there are cases when discourse markers are not translated simply using the equivalents provided by the dictionaries. There are cases when additional particles or additional discourse markers are added to the dictionary variants or particles and different discourse markers are used. The observed phenomenon of the "over-specification" could be analyzed in further research deeper; however, in the current study, we selected all such cases and analyzed the functions expressed by the discourse markers as our presupposition was that such translation could be guided by some rhetorical functions expressed by the original discourse markers.

There are not numerous but very interesting cases of multi-word expressions including particles *na* ("well/just") or *juk* ("just") in Lithuanian translation. Even a single particle is used as a discourse marker which is characteristic of the Lithuanian language to employ particles as discourse markers. There are also cases of multi-word expressions involving connective and inflected verb phrases, for example, *kaip žinote* ("as you know"), *bet žinote* ("but you know"), etc. We looked closer and analyzed such cases with the assumption that the translator's choices could be guided by the function of the rhetorical domain. Table 2 shows the cases of translation with additional elements of discourse markers in Lithuanian.

Table 2. Discourse markers with additional elements in Lithuanian translation

You know	Kaip žinote	As you know	8
You know	Bet žinote	But you know	7
You know	Na jūs žinot	Just you know	2
You know	Na suprantate	you just understand	2
You know	Kaip matote	As you see	1
You know	Na	Particle (well/ just)	71
You know	juk	Particle (just)	5
For instance	kaip pavyzdžiui	As an example	1
I mean	juk	Particle (just)	7
In fact	Ir iš tiesų	And truly	1
In fact	Ir išties	And truly	2
After all	juk	Particle (just)	5
On the other hand	Kai tuo tarpu	When after all	1

Analyzing the cases of the translation *kaip žinote* ("as you know"), we found one case of word-for-word translation as in English the combination of two discourse markers (as you know) is used.

1. How does this all come about? Well, *as you know*, the brain is made up of neurons.

   Kaip visa tai vyksta? Ką gi, *kaip žinote* ("as you know"), smegenys yra sudarytos iš neuronų

The rest of the cases fall under the function of rhetorical specification like in the example below:

2. They have the vocabulary for sense of purpose, ikigai, like the Okinawans. *You know* they are the two most dangerous years of your life.

   Jie turi žodžių tikslingumui nusakyti, kaip "ikigai" Okinavoje. *Kaip žinote* ("as you know"), tai yra du patys pavojingiausi metai tavo gyvenime.

The example demonstrates that the rhetorical domain might have guided the translator to insert an additional connective *kaip* ("as") to convey the specification function of the rhetorical domain.

A different situation is with the translation *bet žinote* ("but you know"), as almost all the cases involve either a combination of the lexical items *but you know / you know but* or *and you know* in English which expresses ideational contrast. Example (3) shows the translation of the two successive discourse markers.

3. That's fantastic. I love Big Placebo. *But, you know*, it's really a serious thing because this stuff is crap and we spend billions of dollars on it.

   Fantastiška. Man patinka didysis placebas. *Bet, žinote*, ("but you know") tai rimtas dalykas, nes tie preparatai yra mėšlas, o mes jiems išleidžiame milijardus dolerių.

Examples (3) and (4) contain the only difference in the order of the discourse markers so we observe the combination of English discourse markers*you know*, *but* translated into Lithuanian *bet žinote* ("but you know").

4. And OK, so you could sex it up and like go to a much more likable Mac,*you know*, *but* really it's the kind of same old crap we've had for the last,' you know, 30 years.

Na, gerai, galima jam suteikti seksualumo, kaip šiuo metu atrodantis Mac darbastalis, *bet žinote* ("but you know") iš tikrųjų tai tas pats šlamštas, kurį mes regime jau 30 metų.

Example (5) is very similar to example (3) with the only difference being that the discourse marker *and* is translated into *bet* ("but") which helps to express ideational contrast.

5. My volunteering got me to the front of the line. *And, you know*, I'm not even ashamed of that.

   Dėl savanoriavimo atsidūriau eilės priešakyje. *Bet žinote* ("but you know"), to nesigėdiju.

The translation of *na jūs žinot* (well you know) represents rhetorical domain, examples (6) and (7). Example (6) shows rhetorical specification and in example (7) we see rhetorical addition.

6. So one of the things you can do to our icons, just like paper, is crease them and fold them, just like paper. Remember, *you know*, something for later.

   Vienas iš dalykų kuriuos galima daryti su mūsų piktogramomis tai jas lankstyti ir perlenkti, lyg popierių. Priminimas, *na jūs žinot* ("well you know"), kažkas vėlesniam laikui.

7. She came to this country in the late 60s, and she was, *you know*, she found herself pregnant, as women did in the late 60s.

   Ji atvyko į šią šalį 60-ųjų antroje pusėje, ir ji, *na suprantate* ("well you know"), ji tapo nėščia, kaip ir visos moterys 60-aisiais.

The Lithuanian particle *na* ("well / just") appears either as a translation of *and* in a combination of discourse markers with *and* shown in example (8).

8. Also just like paper, around our workspace well pin things up to the wall to remember them later,' and I can do the same thing here, *and you know*, you'll see post-it notes and things like that around people's offices.'

   Taip pat kaip ir popieriukai aplink mūsų darbo vietą kuriuos galite tiesiog prisegti prie sienos ir prisiminti vėliau, ta patį aš galiu ir čia, *na žinot* ("well you know"), jūs matysite priminimus taip kaip įprasta matyti žmonių biuruose.

Or again *na* ("well /just") is added to stress the rhetorical specification as in example (9)

9. I realized I was seeing the same archetypal events depicted again and again and again. *You know*: weddings, births, funerals, the first car, the first kiss,

   suvokiau, kad matau tuos pačius archetipinius įvykius, vaizduojamus vėl, ir vėl, ir vėl. *Na žinote* ("well you know"): vestuves, gimimą, laidotuves, pirmąjį automobilį, pirmąjį bučinį,

In a similar way, the translation for instance gains additional particle *kaip pavyzdžiui* because of the rhetorical specification.

10. And in fact, interaction design is what I've been trying to insert in the collection of the Museum of Modern Art, *for instance*, you see here the way a machine plays chess with itself, works by Martin Wattenberg.

    Ir išties interaktyvusis dizainas yra tai, ką bandžiau įtraukti į MoMA kolekciją, *kaip pavyzdžiui* ("as for example"), jūs matote čia mašina žaidžia šachmatais su pačia savimi, Martin Wtterberg darbas.

The Lithuanian particle *juk* ("just") in the below-analyzed examples is used to mark the function of rhetorical cause. It could be used as a translation of the discourse markers *you know* and *I mean* as in examples (11) and (14) or could be added to the direct translation as in examples (12) and (13).

11. That's ridiculous, *you know*, this is New York.

Tai kvaila. *Juk* ("just") čia Niujorkas.

12. and my fingers were literally the size of sausages because -- *you know*, were made partially of water

    jie tiesiogine prasme buvo dešrelių dydžio, nes – *juk žinote* ("just you know"), kad mūsų kūną iš dalies sudaro vanduo

13. What does it mean to have a disability? I mean, people -- Pamela Anderson has more prosthetics in her body than I do.

    Ką reiškia būti neįgaliam? *Sakau, juk* ("I say just") Pamelos Anderson kūne daugiau dirbtinų elementų nei manajame.

14. Okay, step back a minute. *I mean*, it's really not news for me to tell you that innovation emerges out of groups.

    Gerai, stabtelėkime minutei. *Juk* ("just") tai tikrai nebėra naujiena tai, kad inovacijos kyla iš grupių.

The translation of *in fact* with the integrated Lithuanian connective *ir* ("and") also marks the function of rhetorical specification.

15. We don't have to choose between' inspired employees and sizable profits, we can have both. *In fact*, inspired employees quite often help make sizable profits.

    Mes galime turėti abu. *Ir išties* ("and truly"), įkvėpti darbuotojai gan dažnai padeda sukurti žymius pelnus.

The other cases with the discourse marker *in fact* could be considered as word for word translations as in the original text we observe a combination of the two discourses markers *and/in fact* as in example (16)

16. And it's interesting how so much of what we're talking about tonight is not simply design but interaction design. *And in fact*, interaction design is what I've been trying to insert into the collection of the Museum of Modern Art

    Įdomu, kad tiek daug, apie ką kalbame šįvakar nėra tik dizainas, bet interaktyvusis dizainas. *Ir išties* ("and truly") interaktyvusis dizainas yra tai, ką bandžiau įtraukti į MoMA kolekciją

The translation of *after all* acquires either additional particle *juk* ("just") in translation (example (18)) or is simply translated into *juk* (just) (example (17)) as it helps to express rhetorical cause.

17. God thought that it would be best to create the world only with the divine attribute of justice. Because, *after all*, God is just.

    Jis pamanė, kad reiktų jį kurti tik iš dieviškojo teisingumo atributo. Nes *juk* ("just") Dievas teisingas.

18. if we want to have a citizen-centric Internet in the future, we need a broader and more sustained Internet freedom movement. *After all*, companies didn't stop polluting groundwater

    jeigu ateityje mes norime turėti pilietiškai orientuotą internetą, mums reikia platesnio ir ilgalaikio internetinės laisvės judėjimo. *Galiausiai juk* ("finally just") kompanijos nenustojo teršti

Speaking about the translation of *on the other hand* into the Lithuanian *kai tuo tarpu* ("when after all"), the use of the additional *when* could not be explained by rhetorical domain as in this case the discourse markers express ideational contrast (example (19)). The translator's choice to include an additional discourse marker could be explained by choosing not a word-for-word translation but the equivalent for the English after all and thus probably feeling a need to insert an additional discourse marker in order to stress the function of contrast.

19. This is simply because my mother came from a poor background, and she had no choice. My father, *on the other hand*, was rich.

    Kilusi iš neturtingos šeimos, kito pasirinkimo ji neturėjo, *kai tuo tarpu* ("when after all") mano tėvas buvo turtingas,

The translator's choice to additionally use particles is obviously related not to the translation of semantic meaning but more to the pragmatic meaning inferred by the translator from the surrounding context. It connotes the deep observation by Nau and Ostrowski (2010) that Lithuanian particles contain the component of subjectivity and inter-subjectivity and mostly their meaning is colored by the surrounding context.

In the Hebrew translation, the additional particles are not used but in a similar way additional lexical items and question words are sometimes inserted by the translators. Thus, there are cases of multi-word expressions involving an additional connective such as *or*, *as*, *and*, *but*, etc. So again, we need to analyze such cases searching if certain pragmatic meanings could have guided the translator's choices. Table 3 shows the cases of translation with additional elements of discourse markers in Hebrew.

Table 3. Discourse markers with additional elements in Hebrew translation

You know	אתם יודעים מה	You know what	1
You know	אתם מבינים למה	You know (understand) why	1
You know	כמו שהבנתם	As you know (understand)	1
You see	כפי שאתם רואים	As you see	1
In other words	או במילים אחרות	Or in other words	1
I think	שלדעתי/שאני חושב	As I think	2
I think	ואני חושב	And I think	1
I think	כך אני סבור	I think so	1
I think	אבל נראה לי	But I think	1
I mean	אני מתכוון לכך	I mean that	1
By the way	ודרך אגב	And by the way	1
Of course	וכמובן	And of course	2
Of course	אבל כמובן	But of course	2
In fact	ולמעשה	And in fact	3
In fact	שבעצם	As in fact	1

The Hebrew translation of *you know* seems to be also guided by the rhetorical domain, which is related to rhetorical specification, example (20), where in the Hebrew translation there is an additional question word *you know what.*

20. We took all that down, and we found beautiful wooden floors, whitewashed beams and it had the look -- while we were renovating this place, somebody said, "*You know*, it really kind of looks like the hull of a ship."

פירקנו הכל, וגילינו ריצפת עץ נהדרת, קורות עץ מסוידות, וזה נראה -- כששיפצנו את המקום, מישהו אמר,
“*אתם יודעים מה*? (“you know what”) זה נראה כמו גוף של אוניה”

The following example (21) shows that the translation of rhetorical cause includes a different question word related to the reason *you know why*.

21. I was in the park, and I was dressed in my biblical clothing, so sandals and sort of a white robe, *you know*, because again, the outer affects the inner.

זה קרה, בזמן שהייתי בפארק, לבוש בבגדים התנ”כיים שלי. סנדלים וגלימה לבנה. *אתם מבינים למה*,

(“you know why”) בגלל מה שאמרתי מקודם, החיצוניות משפיעה על הפנימיות.

Both examples demonstrate that the translator chose to insert an additional question word to convey the function of the rhetorical domain.

An additional example (22) involves ideational cause which in the Hebrew translation acquires an additional connective *as you know* that could be explained by the translator's choice for a more formal option as the topic is related to science, although generally, TED talks contain rather informal expressions.

22. This one was very unexpected because, *you know*, I grew up with the scientific worldview

השיעור הזה היה מאוד לא צפוי, בגלל, *כמו שהבנתם*, (“as you know”) אני גדלתי עם ראיית עולם מאוד
מדעית

The following few examples (23), (24), and (25) represent the translation of rhetorical specification for the multi-word discourse markers *you see*, *in other words* and *I mean* featuring again the use of additional lexical items.

23. As a matter of fact, this is something that about half of you, more than half of you will not be familiar with. It's a beard trimmer, *you see*?'

למעשה, זה משהו שבערך מחצית מכם, יותר ממחצית מכם, לא יכירו. *כפי שאתם רואים*,

(“as you see”) זוהי מכונת תספורת לזקן.

It should be noted that example (23) represents an interesting case when in Hebrew due to the translator's choice of a more formal variant *as you see* the discourse marker function in Hebrew translation turns into ideational specification.

24. And then there is self-similarity across the scales -- *in other words*, from one skin of the onion to another one.

יש דימיון עצמי במעבר בין קני מידה שונים, *או במילים אחרות*, (“or in other words”) בין קליפת בצל אחת
לאחרת.

25. I'm not sure which way it's going to go.' *I mean*, there's enormous pressures'

אני לא בטוח לאיזה כיוון זה ילך. *אני מתכוון לכך* (“I mean that”) שיש לחצים עצומים.

The multi-word discourse marker *I think* also demonstrates the translator's choices to insert an additional connective related to the rhetorical discourse meaning such as rhetorical cause in example (26), rhetorical specification in example (25), and rhetorical addition in example (28).

26. I'm very proud because *I think* I'm the only person in America'

אני מאוד גאה *מכיוון שאני חושב* (“as I think”) שאני הבן אדם היחיד באמריקה

Again in example (26) due to a more formal choice in Hebrew translation we observe the change into ideational cause.

Rhetorical specification:

27. And the problem with it is, *I think* we are setting ourselves up for a kind of disaster.

והבעיה עם זה היא, *שלדעתי* (“as I think”) אנו מכינים לעצמנו אסון מן הסוג

Rhetorical addition:

28. I want to urge everybody here to apply your passion, your knowledge, and your skills to areas like cymatics. *I think* collectively we can build a global community.

ואני רוצה להאיץ בכולם כאן ליישם את התשוקה שלכם, הידע והמיומנות שלכם לתחומים כמו סיימטיקה. *ואני חושב* (“and I think”) שבאופן קולקטיבי אפשר לבנות כפר גלובלי.

Another interesting case is related to the rhetorical consequence expressed by the two discourse markers: *so* in the initial position and *I think* in the middle position of the sentence. In a way the Hebrew translation *I think so* could be considered word-for-word translation just the translator choosing to use both discourse markers together. In Hebrew it is more formal.

29. It can still really make your day. So, this possibility of a new type of global recognition, *I think*, is driving huge amounts of effort.

עדיין זה עושה לך את היום. לכן, אפשרות זו של סוג חדש של הכרה עולמית, *כך אני סבור*, (“I think so”) גורמת להשקעת מאמצים גדולים מאוד.

There is also a case of ideational contrast which is not expressed by the direct connective so the translator chose to insert the additional connective *but* example (30).

30. Now maybe cows have a really rich internal mental life and are so smart that they choose not to let us realize it, but we eat them. *I think* most people will agree that chimps are capable of much more complex, elaborate, and flexible behaviors than cows are.

ייתכן שלפרות יש עושר גדול של חיי מחשבה, והן כה חכמות שהן בוחרות לא לאפשר לנו לגלות את חכמתן, ומעדיפות שנאכל אותן. *אבל נראה לי* (“but I think”) שרוב האנשים יסכימו שלשימפנזים יש יכולת להתנהג ברמת מורכבות, הבנת פרטים וגמישות רבה יותר מלפרות.

A similar situation could be observed in the Hebrew translation of the multi-word discourse marker *by the way*. It goes together with the discourse marker also and helps to stress ideational addition which leads to the translator’s choice for *and by the way* which could be considered as word-for-word translation.

31. Because the inability to experience regret is actually one of the diagnostic characteristics of sociopaths. It’s also, *by the way*, a characteristic of certain kinds of brain damage.

בגלל שחוסר היכולת להרגיש חרטה היא אחד המאפיינים שעל ידם מאבחנים סוציאופתיה. *ודרך אגב*,

(“and by the way”) זה גם אחד המאפיינים של סוגים מסוימים של נזק מוחי.

The use of additional lexical items with the multi-word discourse marker *of course* in Hebrew translation is also related to the rhetorical discourse meaning. It comprises the case of rhetorical addition translated into *and of course* in Hebrew, example (32) and cases of rhetorical contrast translated into both *and of course* and *but of course* into Hebrew, examples (33), (34), and (35).

32. There are few things more glamorous than the horizon --except, possibly, multiple horizons. *Of course*, here you don’t feel the cold, or the heat --’

יש מעט דברים יותר זוהרים מאופק פרט אולי למספר של אופקים *וכמובן*, (“and of course”) כאן אתה לא מרגיש את הקור, או את החום --

33. I remember thinking; my friend could have explained that entire experiment with a dance. *Of course*, there never seem to be any dancers around when you need them.

אני זוכר שחשבתי שחברי יכול היה להסביר את הניסוי כולו באמצעות מחול. *וכמובן*,

(“and of course”) לעולם אין בנמצא רקדנים בדיוק כשצריך אותם.

34. And I remember reading, after the Lara Croft movies, how Angelina Jolie would go home completely black and blue. *Of course*, they covered that with make-up.

ואני זוכרת שקראתי, אחרי סרטי לארה קרופט, איך אנג’לינה ג’ולי הלכה הביתה עם חבלות שחורות
וכחולות. *אבל כמובן* (“but of course”) שהם כיסו זאת עם איפור,

35. Namely, those who are looking at Spam think potato chips are going to be quite tasty; those who are looking at Godiva chocolate think they won’t be nearly so tasty. *Of course*, what happens when they eat the potato chips?

כלומר, אלו שמסתכלים בקופסת הלוף חושבים שהתפוצ’יפס יהיה מאוד טעים; אלו שמסתכלים בשוקולד לא
חושבים שהתפוצ’יפס יהיה עד כדי כך טעים. *אבל כמובן*, (translated *but of course*) מה קורה כשהם אוכלים
את התפוצ’יפס?

The translation of the discourse marker *in fact* is mostly related to the cases of rhetorical specification where in the Hebrew translation we find additional lexical items – *and in fact* and *as in fact*, examples (36) and (37).

36. We don’t have to choose between’ inspired employees and sizable profits, we can have both. *In fact*, inspired employees quite often help make sizable profits.

אין סיבה שנצטרך לבחור בין עובדים שמוצאים סיפוק בעבודה שלהם לבין רווח כלכלי משמעותי בעסק
שלנו. אנחנו יכולים לקבל את שני הדברים. *ולמעשה*, (“and in fact”) לעתים די קרובות, עובדים שממוצאים
סיפוק בעבודה שלהם עוזרים לחברה בהגדלת הרווחים.

37. And then God looked to the future and realized that, *in fact*, if the world were just filled with compassion, there would be anarchy and chaos.

ואז אלוקים התבונן אל העתיד והבין, *שבעצם*,(“as in fact”) אם העולם יהיה מלא רק בחמלה, יהיו אנרכיה
וכאוס.

There is also a case of rhetorical contrast translated into Hebrew with the additional connective *and in fact*, example (38)

38. Increases in material well-being don’t seem to affect how happy people are.’ *In fact*, you can find that the lack of basic resources, material resources, contributes to unhappiness, but the increase in material resources does not increase happiness.

תוספות ברווחה חומרית אינן משפיעות על כמה אנשים שמחים. *ולמעשה*, (“and in fact”) אתם יכולים למצוא
שהיעדר של משאבים בסיסים משאבים חומריים, תורמים לחוסר אושר. אך הגידול במשאבים החומריים לא
מגביר את האושר.

It could be observed that similar to the Lithuanian translation, the Hebrew translation also includes additional lexical items used together with the multi-word discourse markers to convey the rhetorical discourse meaning. The translator’s choice to additionally use particles or other additional lexical items could be related more to the translation of the pragmatic meaning guided by the surrounding context rather than to the semantics of the discourse markers. As Nau and Ostrowski (2010) point out that Lithuanian particles express certain aspects of subjectivity and inter-subjectivity, the meaning of which is mostly influenced by the surrounding context. The current analysis demonstrates the importance of comparative investigation of discourse markers for raising students’ cultural language awareness that the context and the characteristics of the languages may lead the translator choices to reflect the source message in the target translated text.

Conclusions

The comparative examination of discourse markers offers language instructors and translators more specific information about the roles of discourse markers. Understanding the multifunctionality of discourse markers provides a greater understanding of their application and translation. This form of case study may be a useful way to strengthen students' language awareness and analytic skills and is particularly pertinent for students specializing in English philology or translation.

The current study allows us to observe that there is a tendency of integrating additional particles or other additional lexical items in Lithuanian or Hebrew translated discourse markers (the so-called phenomenon of "over-specification") which has been observed in the translation studies in relation to the differences in the nature of languages. In the current study, it may be observed that the translator choices could be guided by the relatedness of discourse marker functions to the rhetorical domain. Lithuanian is rich in particles and, as the analysis has demonstrated, translators choose to additionally integrate particles into discourse markers to stress the expression of the rhetorical domain. The additional particles and other lexical items in Lithuanian translation help to emphasize the rhetorical domain of discourse marker functions. In a similar mode, additional lexical items are used in Hebrew translations to convey the rhetorical domain. However, Hebrew translations also demonstrate the translator's choices of more formal discourse marker expressions which could be related to the scientific content of certain TED talks. Generally, it seems that, in both languages, translators decipher the rhetorical domain and feel that it should be reflected by additional particles or other lexical items added to the multi-word discourse markers in the translation. It should be also stressed that the observed phenomenon of "over-specification" could be analyzed further in future research.

The current study focuses more on the structure of phrases, although admittedly discourse marker usage is heavily influenced by time periods, available resources in the environment, traditions, and other cultural linguistic factors. However, it is difficult to draw overarching culture-related conclusions because there is a need for more extensive future research on the phenomenon in translations of rendering discourse markers even though they may not have been used in the target language All in all, the current study allows increasing students' awareness of language differences guiding translator choices to convey not only the semantics but also the discourse meaning.

References

Aijmer, K. (Ed.). (2009). *Corpora and language teaching* (Vol. 33). John Benjamins. https://doi.org/10.1075/scl.33

Asher, N. M., & Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Baker, M. (2018). *In other words: A coursebook on translation*. Routledge. https://doi.org/10.4324/9781315619187

Barlow, M. (2011). Corpus linguistics and theoretical linguistics. *International Journal of Corpus Linguistics*, *16*(1), 3–44. https://doi.org/10.1075/ijcl.16.1.02bar

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, *5*(2), 97–116. https://doi.org/10.1016/j.jeap.2006.05.001

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Conrad, S., & Reppen, R. (1994). Corpus-based approaches to issues in applied linguistics. *Applied Linguistics*, *15*(2), 169–189. https://doi.org/10.1093/applin/15.2.169

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Grammar of spoken and written English*. Longman.

Bieliauskienė, R. (2012). Vilnius–jidiš kalbos Jeruzalė. *Krantai*, *4*, 56–61.

Bunt, H., & Prasad, R. (2016, May). ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. *Proceedings of the 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, *Portoroz, Slovenia,* 45–54.

Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber, & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 478-497). Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.027

Crible, L. (2017). Discourse markers and (dis)fluency in English and French: Variation and combination in the DisFrEn corpus. *International Journal of Corpus Linguistics*, *22*(2), 242–269. https://doi.org/10.1075/ijcl.22.2.04cri

Crible, L., Abuczki, Á., Burkšaitienė, N., Furkó, P., Nedoluzhko, A., Rackevičienė, S., Oleškevičienė, G. V., & Zikánová, Š. (2019). Functions and translations of discourse markers in TED Talks: A parallel corpus study of underspecification in five languages. *Journal of Pragmatics*, *142*, 139–155. https://doi.org/10.1016/j.pragma.2019.01.012

Crible, L., & Degand, L. (2019). Domains and functions: A two-dimensional account of discourse markers. *Discours. Revue de Linguistique, Psycholinguistique et Informatique, 24*. https://doi.org/10.4000/discours.9997

Crystal, D. (1988). Another look at, well, you know … *English Today*, *4*(1), 47–49. https://doi.org/10.1017/S0266078400003321

Das, D., & Taboada, M. (2019). Multiple signals of coherence relations. Discours. *Revue de Linguistique, Psycholinguistique et Informatique, 24*. https://doi.org/10.4000/discours.10032

Degand, L., & Pander Maat, H. (2003). A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale. *LOT Occasional Series*, *1*, 175–199.

Dixon, R. M. W. (2009). The semantics of clause linking in typological perspective. In A. Aikhenvald, & R. M. W. Dixon (Eds.), *The semantics of clause linking: A cross-linguistic typology* (pp. 1–55). Oxford University Press.

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. https://doi.org/10.1075/ijcl.16127.dob

Dupont, M., & Zufferey, S. (2017). Methodological issues in the use of directional parallel corpora: A case study of English and French concessive connectives. *International Journal of Corpus Linguistics*, *22*(2), 270–297. https://doi.org/10.1075/ijcl.22.2.05dup

Evers-Vermeul, J., Degand, L., Fagard, B., & Mortier, L. (2011). Historical and comparative perspectives on subjectification: A corpus-based analysis of Dutch and French causal connectives. *Linguistics*, *49*. https://doi.org/10.1515/ling.2011.014

Fraser, B. (1996). Pragmatic markers. *Pragmatics*, *6*(2), 167–190. https://doi.org/10.1075/prag.6.2.03fra

Fraser, B. (2009). An account of discourse markers. *International Review of Pragmatics*, *1*(2), 293–320. https://doi.org/10.1163/187730909X12538045489818

Furkó, P., & Abuczki, Á. (2014). English discourse markers in mediatised political interviews. *Brno Studies in English*, *40*(1), 45-64. https://doi.org/10.5817/BSE2014-1-3

Gessler, L., Behzad, S., Liu, Y. J., Peng, S., Zhu, Y., & Zeldes, A. (2021). DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking* (pp. 51–62). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.disrpt-1.6

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7–24. https://doi.org/10.1075/ijlcr.1.1.01gra

Hasselgren, A. (2002). Learner corpora and language testing: Smallwords as markers of learner fluency. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer-learner corpora, second language acquisition and foreign language teaching* (pp. 143–174). John Benjamins. https://doi.org/10.1075/lllt.6.11has

Hoek, J., Zufferey, S., Evers-Vermeul, J., & Sanders, T. J. (2017). Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics*, *121*, 113–131. https://doi.org/10.1016/j.pragma.2017.10.010

Huang, L. F. (2011). *Discourse markers in spoken English: A corpus study of native speakers and Chinese non-native speakers* [Doctoral dissertation, University of Birmingham].

Joseph, J. E. (2009). Why Lithuanian accentuation mattered to Saussure. *Language & History*, *52*(2), 182–198. https://doi.org/10.1179/175975309X452067

Joslyn-Siemiatkoski, D. (2007). Book review: The Cambridge history of Judaism: The late Roman-Rabbinic period. *Theological Studies*, *68*(4), 924-925. https://doi.org/10.1177/004056390706800413

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 48-54. https://doi.org/10.21236/ADA461156

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*(3), 243–281. https://doi.org/10.1515/text.1.1988.8.3.243

McEnery, T., & Xiao, R. (2011). What corpora can offer in language teaching and learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 364–380). Routledge. https://doi.org/10.4324/9780203836507

Montechiari, E. A., Stankov, S., Mishev, K., & Damova, M. (2022). Machine learning methods for discourse marker detection in Italian. *Proceedings of the International Scientific Interdisciplinary Conference,* 74-80.

Nau, N., & Ostrowski, N. (2010). Background and perspectives for the study of particles and connectives in Baltic languages. In N. Nau, & N. Ostrowski (Eds.), *Particles and connectives in Baltic* (pp. 1–37). Academia Salensis.

Noël, D. (2003). Translations as evidence for semantics: An illustration. *Linguistics*, *41*, 757–785. https://doi.org/10.1515/ling.2003.024

O'Donnell, W. R., & Todd, L. (2013). *Variety in contemporary English*. Routledge. https://doi.org/10.4324/9780203135433

Oleškevičienė, G. V., Zeyrek, D., Mažeikienė, V., & Kurfalı, M. (2018). Observations on the annotation of discourse relational devices in TED talk transcripts in Lithuanian. *Proceedings of the Workshop on Annotation in Digital Humanities Co-Located with ESSLLI*, 3–58.

Pit, M. (2007). Cross-linguistic analyses of backward causal connectives in Dutch, German and French. *Languages in Contrast*, *7*(1), 53–82. https://doi.org/10.1075/lic.7.1.04pit

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., & Webber, B. L. (2008). The Penn Discourse TreeBank 2.0. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. European Language Resources Association.

Römer, U. (2008). Corpora and language teaching. In A. Lüdeling, & M. Kytö (Eds.). *Corpus linguistics. An international handbook* (Volume 1, pp. 112–130). Mouton de Gruyter.

Sanders, T. J., & Noordman, L. G. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, *29*(1), 37–60. https://doi.org/10.1207/S15326950dp2901_3

Schiffrin, D. (1987). *Discourse markers*. Cambridge University Press. https://doi.org/10.1017/CBO9780511611841

Schiffrin, D. (2001). Discourse markers: Language, meaning, and context. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 54–75). Blackwell Publishers. https://doi.org/10.1002/9780470753460.ch4

Schourup, L. (1999). Discourse markers. *Lingua*, *107*(3–4), 227–265. https://doi.org/10.1016/S0024-3841(96)90026-1

Silvano, M. da P., Damova, M., Valunaité, G. O., Liebeskind, C., Chiarcos, C., Trajanov, D., Ciprian-Octavian, T., Apostol, E.-S., & Baczkowska, A. (2022, June). ISO-based annotated multilingual parallel corpus for discourse markers. *Proceedings of the 13th Language Resources and Evaluation Conference, Marseille, France,* 2739–2749.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J. (Ed.). (2004). *How to use corpora in language teaching*. John Benjamins. https://doi.org/10.1075/scl.12

Siyanova-Chanturia, A., Conklin, K., & Van Heuven, W. J. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 776. https://doi.org/10.1037/a0022531

Stede, M., Afantenos, S., Peldzsus, A., Asher, N., & Perret, J. (2016). Parallel discourse annotations on a corpus of short texts. *Proceedings of the 10th International Conference on Language Resources and Evaluation* (pp. 1051–1058). European Language Resources Association.

Svartvik, J. (1980). *Well* in conversation. In S. Greenbaum, G. Leech, & J. Svartvik (Eds.) *Studies in English Linguistics* (pp. 167–177). Longman.

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2016). A discourse-annotated corpus of conjoined VPs. *Proceedings of the 10th Linguistic Annotation Workshop* (pp. 22–31). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-1704

Wei, N., & Li, J. (2013). A new computing method for extracting contiguous phraseological sequences from academic text corpora. *International Journal of Corpus Linguistics*, *18*(4), 506–535. https://doi.org/10.1075/ijcl.18.4.03wei

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, *32*(1), 231–254. https://doi.org/10.1017/S026719051200013X

Zeldes, A., Liu, Y. J., Iruskieta, M., Muller, P., Braud, C., & Badene, S. (2021). The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking*, 1–12.

Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., & Ogrodniczuk, M. (2020). TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, *54*, 587–613. https://doi.org/10.1007/s10579-019-09445-9

Zufferey, S. (2016). Discourse connectives across languages: Factors influencing their explicit or implicit translation. *Languages in Contrast*, *16*(2), 264–279. https://doi.org/10.1075/lic.16.2.05zuf

Zufferey, S., & Cartoni, B. (2012). English and French causal connectives in contrast. *Languages in Contrast*, *12*(2), 232–250. https://doi.org/10.1075/lic.12.2.06zuf

Zufferey, S., & Degand, L. (2017). Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*, *13*(2), 399–422. https://doi.org/10.1515/cllt-2013-0022

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.