Multiword Expressions as Discourse Markers in Hebrew and Lithuanian

Giedrė Valūnaitė-Oleškevičienė, Chaya Liebeskind

Abstract

Multiword expressions are of key importance in language generation and processing and could also operate as discourse markers. We combined the alignment model of the phrase-based statistical machine translation and manual treatment of the data in order to examine English multiword discourse markers and their equivalents in Lithuanian and Hebrew, by researching their changes in translation. We focused on the two most frequent: 'I think' and 'you know' aiming to research if they demonstrate their functional stability as discourse markers in translation and what changes they undergo in Lithuanian and Hebrew translation.

Keywords: *multilingual corpus; multiword expression; discourse relation; discourse marker; translation.*

Research on multiword expressions has identified that language is not produced just word by word but it usually involves generating certain chunks using a lot of formulaic constructions (Barlow 2011). Native speakers have a multitude of memorized sequences to perform various functions within language, for example, organizing discourse (Nattinger and DeCarrico 1992), or processing language by the speaker and the hearer (Siyanova-Chanturia, Conklin, and Van Heuven 2011). Formulaic language includes idioms and proverbs, various clichés and collocations, lexical bundles, and phrasal verbs. Biber et al. (2004) observed that lexical bundles constitute a high percentage of the produced language and the authors identified that one function of lexical bundles is to organize discourse by providing an example of such bundles, for example, *I think*, which relates to the research on discourse markers. Phrases such as *you know* and *I think* have also been classified as discourse markers that perform certain discourse organising functions. However, Maschler & Schiffrin (2015) observe that there is no a priori theoretical classification of discourse markers and the analysis of function in the data is necessary. Research on discourse markers as tools of discourse management prove that they carry several functions, including signposting, signalling, 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 multiword expressions as discourse markers in a corpus of spoken Slovene, identifying structurally fixed discourse marking multiword expressions.

The purpose of the current research is to examine multiword expressions used as discourse markers in TED talk English transcripts focusing on 'I think' and 'you know' and compare them with their counterparts in Lithuanian and Hebrew by following Maschler & Schiffrin (2015) observation on the necessity of closer investigation on their function as discourse markers. 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 analyse their translations in Lithuanian and Hebrew to determine if they function as discourse markers and are also multiword expressions or one word translations, or if they acquire any other linguistic forms. An additional benefit 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.

Theoretical background

The literature overview briefly takes into account the research languages, studies related to multiword 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

First, it is necessary to briefly discuss the cultural heritage of the languages of the research, which, in a way, guided the choice of languages for our study. According to Bieliauskienė (2012), Jewish and Lithuanian cultures coexisted on the same territory from the first half of the 14th century. The author stressed that from 19th century onwards, in the Republic of Lithuania, Vilnius was called Lithuania's Jerusalem, attracting knowledgeable people in the field of education and inspiring a flourishing high culture, for example, in theatre, art, and literature. In fact, both languages, Lithuanian and Hebrew, formed the cultural heritage of the region. In this study, we research the Lithuanian and Hebrew corpus in parallel with pivotal English.

Lithuanian is an old surviving Baltic language, retaining forms related to Sanskrit and Latin and preserving the most phonological and morphological aspects of the Proto-Indo-European language. Thus, it has gained importance in Indo-European language studies and has been researched by many scientists so far, including Ferdinand de Saussure, who considered Lithuanian "the Galapagos of linguistic evolution" (Joseph 2009). Lithuanian is rich in declensions and cases inside the declensions and the oldest layer of the Lithuanian language vocabulary is related to the Indo-European language, which is dated to be approximately over 5000 years old.

Hebrew is a very old, northwest Semitic language belonging to the group of Canaanite languages; the first examples of Paleo-Hebrew date back to the 10th century. It is a successful example of a revived dead language. It survived in the medieval period as the language of religious scriptures, being revived, in the 19th century, into a spoken and literary language (Joslyn-Siemiatkoski 2007). Hebrew is an important language for researchers specializing in Middle East civilizations and Christian theology studies.

Multiword expressions as discourse markers

The research areas of natural language processing (NLP), linguistics, and translation are closely related to discourse research, focusing on discourse relations between clauses or sentences. NLP research focuses more and in depth on multiple language-related areas, such as semantic phenomena, dialogue exchange structure, and discourse textual structure (Webber and Joshi 2012). 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 word by word; it is also formulaic. Speakers possess multiple learnt 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 finalized 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 & Li (2013), formulaic language covers approximately 60% of written texts in their researched corpus of 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. Based on the evidence of the formulaic nature of language for communication, research has turned to investigating multiword expressions used as discourse markers (Dobrovoljc 2017), identifying structurally fixed discourse marking multiword expressions.

Another important issue in NLP is discourse management, which is related to discourse relations, connecting ideas between sentences and bigger parts of the text. 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. Discourse markers perform important functions, such as signposting, signalling, 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* (Schiffrin 2001; Hasselgren 2002; Maschler and Schiffrin 2015). Hasselgren (2002) advocated that better discourse marker signal fluency 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 (Stede et al. 2016; Zufferey and Degand 2017; Oleskeviciene et al. 2018; Zeyrek et al. 2019).

Language, especially spoken, is characterised by discourse marker use; however, some of them (e.g., *you know*, *I think*, *well*) are sometimes referred to in a critical manner, as indicating a lack of fluency (O'Donnell and 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 signalling, 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 misses words such as *you know*, or *well*, 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 (Granger 2015; Cobb and Boulton 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. Discourse markers have been researched by applying various theoretical approaches, such as Rhetorical Structure Theory (Mann and Thompson 1988), Segmented Discourse Representation Theory (Asher, Asher, and Lascarides 2003), and PDTB (Prasad et al. 2008), first focusing on the monolingual approach, which resulted in multilingual studies focusing on translation (Degand and Pander Maat 2003; Pit 2007; Dixon 2009; Zufferey and Cartoni 2012). As Zufferey & 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 & 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 on English parliamentary debates translated into Dutch, German, French, and Spanish, searching what types of discourse connectives might have a higher tendency to be more frequently omitted in 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 signalling 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 favour complex structures, sparsely using explicit discourse markers. The author analysed the example of an evident difference between English and Arabic, identifying that, while English prefers signalling discourse relation through discourse markers, Arabic prefers grouping the information into bigger grammatical chunks and using fewer discourse markers. The finding is supported by (Al-Saif & Markert (2010), who observed that, in Arabic, many discourse relations are expressed via prepositions with nominalizations. 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. It appears that it is not always possible to use the word for word technique and natural changes in translation are sometimes inevitable. According to Baker (2018), grammatical changes in translation involve certain techniques, such as substitution, transposition, omission, and supplementation. Substitution is the change of the grammatical category of the source unit in translation. For example, active voice is more common in Lithuanian; therefore, English passive voice units could be changed into active units:

(1) He was told the news. – jam pranešė naujienas

Similarly, in the following example, the verb in the source language is changed into a noun in Hebrew translation.

(2) We should have broken ten minutes before. – היינו צריכים לצאת להפסקה לפני 10 דקות

Transposition represents a change of position in the order of elements of the source textual unit or changing the part of speech in translation, which implies the change in the order of the elements in the translated text. In Lithuanian translation, we observe a change in the order of the elements in the sentence.

(3) After he had left – Jam išėjus.

In the case of Hebrew translation, the change of the order of the elements could be observed in the following example.

(4) Classical music – מוזיקה קלאסית

Omission occurs when some elements of the original text could be considered excessive or redundant in translation. In the Lithuanian translation example, the whole phrase *I thought* is omitted.

(5) I thought you said you were alright. – Bet tu sakei, kad viskas gerai.

In the following example in Hebrew, the translation of *are* is omitted.

(6) We still are – אנחנו עדיין

Supplementation involves changes when new elements, which are non-existent in the source text, appear in the translated text in order to ensure structural adequacy of the latter. Such modifications are usually considered structurally or contextually motivated. For example, due to the elliptical nature of the English language, the Lithuanian translation should use supplementation to make the translation understandable.

(7) Soap star – muilo operos žvaigždė (although the word opera is omitted in English due to ellipsis, it should be added in Lithuanian translation to make it contextually coherent).

The same technique should be applied in the Hebrew translation.

(8) Soap star – כוכב אופרת סבון

As shown above, translation is not a mere process of transposing words from one language into another but requires certain motivated changes. Thus, translation involves grammatical transformations, as a result of the process of looking for approximate correspondences in the translated texts.

Research data resources

It should be stressed that parallel data resources are not extensive, and researchers still need to work on creating parallel corpora for their research, especially if they would like to cover the variety of languages and areas. One of the most prized parallel multilingual resources is Europarl (Koehn 2005). It comprises the translations of the European Parliament proceedings (at most 50 million words) in most European languages; however, it covers just one specific domain of parliamentary proceedings.

TED talks subtitles to their videos seem to be a growing resource of parallel linguistic material, covering a multitude of languages. In addition, being an open and a developing

resource, TED talks attract attention of researchers and their subtitles cover a wide variety of knowledge fields (Cettolo, Girardi, and Federico 2012), which makes the data of the talks widely applicable. However, researchers should keep in mind that the talks are translated by volunteers although with administratively managed quality checks, and the translation is mostly unidirectional from source English subtitles to other target languages. Furthermore, Dupont & Zufferey (2017) identified that such talks contain features of both spoken and written language, as they are semi-prepared speeches by nature. Additionally, (Lefer & Grabar (2015) observed that subtitle translation bears certain specificity in itself. Even by taking into account the features of TED talks discussed by researchers, TED talks are extensively useful as they are an open resource and could provide large amounts of parallel data for research. Besides, parallel corpora are employed as a pool of data for statistical machine translation systems and TED talks is one of the most frequent data resources referred to explore multilingual Neural MT (NMT) (Aharoni, Johnson, and Firat 2019; Chu, Dabre, and Kurohashi 2017; Hoang et al. 2018; Khayrallah et al. 2018; Tan et al. 2018; Xiong et al. 2019; Zhang, Meng, and Liu 2019). NMT, as currently the newest technique of MT, stems from the model of the functioning of the human brain neural networks, which place information into different layers for processing it before generating the outcome. With the technological advancements, NMT gained impetus, as it used to be, resource and computation wise, too costly to outdo phrase-based MT, which operates on the basis of translating entire sequences of words. Now, the neural approach of NMT started challenging the long-lasting prevalence of phrase-based MT techniques. However, in the current research, phrase-based MT was applied relying on two main reasons: NMT techniques do not allow extensive processing of phrases and NMT procedures are not as explicit as phrase-based MT processes. The current study does not involve the full set of phrase-based MT systematic procedures, as it is used just for a phrase table construction, which is a single step of the phrase-based MT paradigm. The detailed description of the research procedures is provided in the research methodology section.

Research methodology

The research aim comprised examining multiword expressions used as discourse markers in TED talk English transcripts and comparing them with their counterparts in Lithuanian and Hebrew. Thus, there was a need to achieve the double objectives of creating the parallel corpus for the research data and carrying out the research on multiword expressions used as discourse markers in the studied languages. Unlike working on one language and using statistical methods we used parallel corpus knowledge alignment algorithm. Initially, the list of multiword and one word expressions that could potentially be used as discourse markers was generated relying on theoretical insights by Schiffrin (1987) 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.

Parallel Corpus creation

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 87.230 aligned sentences (published in LINDAT/CLARIN-LT repository http://hdl.handle.net/20.500.11821/34).

Multiword discourse marker extraction

Another stage of the research focuses on multiword expressions that are used as discourse markers to ensure textual cohesion and, according to Fraser (2009), to relate separate discourse messages. For example, phrases such as *you know*, *I mean*, *of course*, are characteristic of spoken language (Maschler and Schiffrin 2015; Furkó and Abuczki 2014; Huang 2011). Thus, 3.314 aligned sentences containing the earlier mentioned multiword 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 much challenge; however, turning to multiword expressions, they 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 multiword 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, this is really an infinite thing.
(2) You know exactly what you want to do from one moment to the other.

After that, the variations 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.

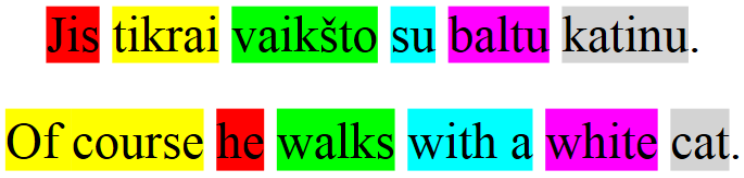

Figure 1. Lithuanian – English phrase alignment

Figure 1 visualizes Lithuanian – English corresponding phrases marked in respective colours.

Figure 2. English – Hebrew phrase alignment

Figure 2 shows English – Hebrew respective phrase alignment, with a note for the reader that Hebrew text should be read from right to left.

The model applies the segmentation of the input into sequences of words, which are called phrases, and then each phrase is translated into English 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 translations were manually filtered to reject the wrong translation variants and 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 analyse their use.

While analysing the data, we noticed that there was a small amount of data left which did not fit the variations of possible translations. The first supposition was that it might represent the cases of omissions; however, we decided to analyse it closely to verify. We checked manually the extracted non-attached data and established that most of the analysed cases involved omission with some minor grammatical transformation cases, incorrect translations, and some phrases not included in the possible translations by the machine.

Research findings

Multiword discourse markers in the corpus

The most frequent multiword expressions used in the study corpus have been extracted and are presented in the table below.

Table 1. Most frequent multiword expressions in the corpus

Multiword expression	Frequency
I think	580
You know	573
That is	370
Of course	312
You see	287
In fact	256
I mean	199

For example	161

It could be seen in Table 1 that the two most frequent multiword expressions in the corpus are *I think* and *you know*.

As mentioned earlier, multiword expressions needed to be manually annotated, spotting the cases when the expressions were used as discourse markers. The manual annotation revealed that some multiword expressions are used as discourse markers more frequently while others more often used as content words fully integrated into sentences.

Table 2. Most frequent multiword expressions used as discourse markers

Multiword expression	Used as discourse marker	Content word
I think	473	107
You know	380	193
That is	29	341
Of course	233	79
You see	47	240
In fact	217	39
I mean	168	31
For example	117	44

It is visible in Table 2 that multiword expressions *That is* and *You see* although identified as discourse markers by the theoretical literature, in this study, they demonstrate a weak tendency to be used as discourse markers and are mainly used as content words in the current corpus, while multiword expressions *I think* and *you know* demonstrate a high tendency of being used as discourse markers and the stability of remaining discourse markers in Lithuanian and Hebrew translation.

The translations of discourse marker "I think"

Further, following our research aim, we present a detailed analysis of the translations of the two most frequent multiword expressions used as discourse markers – *I think* and *you know*. The alignment approach allowed extracting direct output of the translations together with the figures of the translation frequency. First, we explore the translations of the most frequent multiword discourse marker, *I think*.

Table 3. Translations of discourse marker *I think*

Lithuanian			
Discourse marker	Translation variants		Number of cases used
I think	Mano manymu	In my opinion	17
	Man atrodo	It seems to me	7

	Man rodos	It seems to me (different derivation)	6
	Mano nuomone	In my opinion (different derivation)	20
	Mano galva	In my head	2
	Aš galvoju	I think	8
	Aš susimąstau	I reflect	1
	Aš tikiu	I believe	1
	Manau	I think (different derivation)	350
	Tikiu	I believe (different derivation)	3
	Atrodo	It seems	4
	Galvoju	I consider	8
	Manyčiau	I would think	1
	Prisimenu	I remember	1
	Omission	48	48
	Grammatical transformation	3	3
Hebrew			
Discourse marker	Translation variants		Number of cases used
I think	אני חושב	I think (male)	215
	אני חושב ש	I think that	4
	אני חושבת	I think (female)	51
	ואני חושב	And I think	70
	אני מאמינה	I believe (female)	1
	אני משוכנע	I am convinced (male)	1
	אני משערת	I assume (female)	1
	אני סבור	I think (male)	17
	אני סבורה	I think (female)	4
	כך אני סבור	So I think (male)	1
	שאני סבור	As I think	1
	דעתי	In my opinion	55
	כמדומני	It seems to me	2
	לטעמי	to one's taste	1
	נדמה	It seems	2

	אבל נראה לי	But it seems to me	2
	נראה לי	It seems to me	13
	Omissions		23
	Grammatical transformation		1
	Missing derivations		6
	Missing phrases		2

The most frequent multiword expression in the researched corpus, *I think*, has a number of translation variants in both researched languages, Hebrew and Lithuanian. The most frequent one in Lithuanian is a one-word expression – an inflected verb, *manau*, which, due to Lithuanian being a highly inflected language (Zinkevičius, Daudaravičius, and Rimkutė 2005), fully represents the verb-pronoun cases. Other one-verb variants and multiword expressions do not demonstrate high. A separate case is represented by omission, which comprises 48 situations, showing that such a technique is also chosen by the translators.

Referring to Hebrew, the most frequent translation is **אני חושב**, which refers to a male derivative, while the female derivate, **אני חושבת**, comprises only 51 cases. The prevalence of male derivatives could be explained by the nature of the Hebrew language, which has the feature that male derivatives are used while addressing purely male and mixed audiences (Tobin 2001). However, Hebrew translation variant choices differ from the Lithuanian ones, as they mostly remain multiword expressions in translation. Another interesting observation in Hebrew is that a number of 70 cases include the additionally integrated connective *and* into the derivative **ואני חושב**. It reveals that sometimes translators prefer inserting additional information into the translation, which could be related not to the direct semantic meaning of addition of *and* but more to the pragmatic inferences drawn by the translators form the surrounding contexts, which relates to the observations of (Blakemore & Carston (1999), and Moeschler (1989). Hebrew demonstrates less omission cases than Lithuanian for the discourse marker *I think* as the number of omissions in Hebrew is 23, almost half of the Lithuanian omission number**.**

The translations of discourse marker 'you know'

Another commonly used multiword discourse marker, *you know*, demonstrates far more variable translations.

Table 4. Translations of discourse marker *you know*

Lithuanian			
Discourse marker	Translation variants		Number of cases used
You know	Na jūs žinot	Just you know	2

	Jūs žinot/e	You know	7
	Kaip žinote	As you know	8
	Jūs suprantat	You understand	2
	Ar ne	Isn't it	3
	Ar žinot	Do you know	2
	Norėtumėte žinoti	You would like to know	1
	Na suprantate	you just understand	2
	Kaip matote	As you see	1
	Bet žinote	But you know	7
	Žinote	You know (different derivation)	116
	Žinot	You know (different derivation)	16
	Na	Particle (just)	71
	išties	right	2
	Žinai	You know (different derivation)	26
	Žinoma	It's known	1
	Matote	You see	3
	Greičiausiai	Probably	1
	juk	Particle (yeah)	5
	žinokite	Just know	1
	suprantama	It's understandable	1
	suprantat	You understand (different derivation)	2
	omission	31	31
	Grammatical transformation	8	8
	Missing derivation	1	1
Hebrew			
Discourse marker	Translation variants		Number of cases used
You know	אתם יודעים	You know (plural, male)	191
	אתן יודעות	You know (plural, female)	2
	אתה יודע	You know (singular, male)	26
	את יודעת	You know (singular, female)	17

	אתם מבינים	You understand (plural, male)	2
	אתם מכירים	You know (plural, male)	1
	כידוע	As you know	1
	Omissions		113
	Grammatical transformation		21
	Missing derivations		5
	Missing phrases		0
	Typo		1

A closer investigation into the translations of discourse marker *you know* reveals that the most common ones in Lithuanian are also one-word verbs žinote/ žinai/ žinot, which represent verb-pronoun cases. Another quite frequent translator choice is the single particle *na*. Although not numerous, very interesting cases of multiword expressions with particles could be found, such as *na jūs žinote* or *na suprantate*, or a single particle *juk*. Even a single particle is used as discourse marker, which is characteristic of the Lithuanian language. There are also cases of multiword expressions involving a connective and inflected verb phrases, for example, *kaip žinote*, *bet žinote*. The translator's choice to additionally use particles or connectives is obviously related not to the translation of semantic meaning but more to the pragmatic meaning inferred by them from the surrounding context. It connotes with the deep observation made by (Nau & Ostrowski (2010) that Lithuanian particles contain the component of subjectivity and inter-subjectivity, and their meaning is mostly coloured by the surrounding context.

In Hebrew, the translation variants for the discourse marker *you know* are not as variable. The most frequent ones, again, are the variants referring to the male gender, including both plural (191) **אתם יודעים** and singular (26) **אתה יודע**, which by far exceeds the number of female derivatives in plural (2) **אתן יודעות** and singular (17) **את יודעת**. In Hebrew, this discourse marker is much prone to omission, as the number of omissions amounts to 113 cases, which are a bit less than the number of the translated cases. Again, multiword expressions remain multiword expressions with just one case of one-word choice in translation.

The translation choices for the multiword expression serving as a discourse marker *you know* are more versatile than those of *I think* and certain cases of grammatical transformation could be observed in the case of the former in Table 5.

Table 5. Grammatical changes in translation of the multiword discourse marker *you know*

Lithuanian			
Discourse marker	Translation variants with grammatical change		Number of cases used
You know	t.y.	That is	1

	Kaip sakiau	As I said	1
	taigi	so	2
	įsivaizduokit	You imagine	1
	laikoma	It is thought	1
	Iš tiesų	really	1
	gerai	okay	1
Hebrew			
Discourse marker	Translation variants with grammatical change		Number of cases used
You know	טוב נו,	colloquial in Hebrew okay, well,	1
	ואז כמובן,	Then of course	1
	לדוגמא,	For example	2
	וכמובן	And of course	1
	הרי	Indeed, therefore	3
	אם יודעים	If you know (plural, male)	1
	כאילו	As if	2
	ואנו יודעים	And we know (plural, male)	1
	נחשו מה,	guess what	1
	נוטים להיות	Tend to be	1
	למעשה	In fact	1
	איך לומר	How to say	1
	נו	well	1
	ברור	clearly	1
	תראו	look	1
	לידיעתכם	For your information	1
	ככה	This way	1

In Lithuanian, eight cases of grammatical changes were found and, even amongst those, one-word discourse connectives prevail. The multiword discourse marker *you know* is translated also into a conjunction, *taigi* (so), and adverbs *gerai* (okay) and *iš tiesų* (really). However, such translator choices are absolutely rare, considering the size of the dataset.

The grammatical transformation cases are more numerous, comprising of 21 occurrences, and much more versatile in Hebrew. The most interesting cases include: **טוב נו,** (okay), which is a usual colloquial saying in Hebrew, **נחשו מה,** (guess what), and two conjunctions used successively, **כאילו** (as if). There are also some cases when a conjunction is just added as in the following example**, ואז כמובן** (then of course), which could be done by the translator simply to stress the discourse management role of the discourse marker used or possibly attaches a rhetorical function to the integrated conjunction. Even among the limited cases of grammatical transformation, multiword expressions as discourse markers prevail in

Hebrew. What is similar to Lithuanian is that there are also adverbs used in the Hebrew translation: **הרי** (indeed), **נו** (well), **ברור** (clearly). Reflecting why different discourse markers demonstrate different translation choices could be based on the nature of the target language into which the texts are translated; for example, Lithuanian is rich in particles and, as the analysis has demonstrated, translators choose to additionally integrate particles into discourse markers to add supplementary discourse expressions**.**
In Hebrew, the male gender prevails in translation, and translators automatically give preference to male derivatives as in English; the gender is not expressed and the choice of the gender of the derivative is completely the translator's choice. Another observation regarding Hebrew is that multiword discourse markers remain multiword because of the translator choice to relay more on word for word translation, while in Lithuanian there is a tendency to omit the pronoun by using just an inflected verb, and this way, multiword discourse markers turn into one-word discourse markers.

Conclusions

The study results showed that English multiword expressions 'I think' and 'you know', identified as discourse markers according to Maschler and Schiffrin (2015) function-based approach, remain discourse markers in Lithuanian and Hebrew translation but they demonstrate variability in Lithuanian and Hebrew translations: they are either translated into multiword expressions or one inflected word, or they are completely omitted. In Hebrew, the translation of multiword discourse markers prevail, and there is a clear tendency for translators to give preference to male over female derivatives, which is due to the nature of the Hebrew language (Tobin 2001). However, it should be stressed that, in Lithuanian, there is a clear tendency observed for one-word discourse markers in translation. One-word translations mainly include verbs, for example, *žinote; suprantate, įsivaizduojate*, which, due to Lithuanian being a highly inflected language (Zinkevičius, Daudaravičius, and Rimkutė 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 almost word-for-word translations.
More interesting cases include translator choices of particle-verb or connective-verb multiword expressions, which, due to the use of additionally integrated particles and conjunctions, also carry out certain additional discourse meaning. For example, in Lithuanian, the multiword expression discourse marker *you know* splits into a number of multiword expressions and one-word translations. Multiword expressions could be classified into cases representing pronoun-verb phrases – *jūs žinote, jūs suprantate, jūs įsivaizduojate, jūs esate girdėję* – (which do not have additional colouring), particle-verb phrases – (*na/juk/ir*) *žinote, suprantate* – or connective-verb phrases – (*kaip, kad*) *žinote, matote* – in which connectives could be used in a pre- or post-position relative to the verb (which carry additional discourse meaning due to the integrated particle or connective). In addition, in Hebrew translations, the connective *and* is integrated into the derivate in quite a significant number of occurrences, and there are cases of integration of other connectives. The integration of particles for Lithuanian and connectives for both languages evidently carries the pragmatic meaning that could have been inferred from the surrounding contexts by the translators (Nau and Ostrowski 2010; Blakemore and Carston 1999; Moeschler 1989). Concerning discourse layer, based on the results of the current study revealing the cases

where translators chose to insert particles in Lithuanian and connectives in Hebrew, both of which carrying a certain additional discourse meaning in the translation, it seems that translator choices might be also guided by the inner discourse managing system of the target language.

Referring to omissions, they are moderate in number except for surprisingly high occurrences of *you know* omissions in the Hebrew translation, which could be explained by the fact that such a discourse connective is not naturally used in Hebrew. Consequently, translators choose either omission or grammatical transformation, which is also a bit higher in number in this case.

Future research

The translator's choice to insert particles and connectives needs closer investigation and might be studied in future research. Furthermore, keeping in mind that each language is a unique system with unique features, research could be carried out without English as a pivotal language, which means furthering the current research and using linguistically linked open data (LLOD) and thus accessing related linguistic data directly and comparing the languages. This has already been done for related languages; for example, Snyder et al (2010) analysed Ugaritic (an ancient Semitic language spoken in the second millennium BCE) through resources originally developed for Hebrew. However, linked data provide a sound basis and potential for interoperable resources relating across various languages and enable research across languages and areas**.**

References:

Aharoni, Roee, Melvin Johnson, and Orhan Firat. 2019. "Massively Multilingual Neural Machine Translation." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3874–84.

Al-Saif, Amal, and Katja Markert. 2010. "The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic." In *LREC*, 2046–53.

Asher, Nicholas, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Baker, Mona. 2018. *In Other Words: A Coursebook on Translation*. Routledge.

Barlow, Michael. 2011. "Corpus Linguistics and Theoretical Linguistics." *International Journal of Corpus Linguistics* 16 (1): 3–44.

Biber, Douglas. 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, Douglas, Susan Conrad, and Viviana Cortes. 2004. “If You Look At…: Lexical Bundles in University Teaching and Textbooks.” *Applied Linguistics* 25 (3): 371–405.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1994. “Corpus-Based Approaches to Issues in Applied Linguistics.” *Applied Linguistics* 15 (2): 169–89.

Biber, Douglas, Stig Johansson, Geoffrey Leech, S. Conrad, Eclwarcl Finegan, and Randolph Quirk. 1999. “Longman.” *Grammar of Spoken and Written English*.

Bieliauskienė, Roza. 2012. “Vilnius–Jidiš Kalbos Jeruzalė.” *Krantai*, no. 4: 56–61.

Blakemore, Diane, and Robyn Carston. 1999. “The Interpretation of And-Conjunctions.” *Iten, C. &*.

Cettolo, Mauro, Christian Girardi, and Marcello Federico. 2012. “Wit3: Web Inventory of Transcribed and Translated Talks.” In *Conference of European Association for Machine Translation*, 261–68.

Chu, Chenhui, Raj Dabre, and Sadao Kurohashi. 2017. “An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation.” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 385–91.

Cobb, Tom, and Alex Boulton. 2015. *Classroom Applications of Corpus Analysis*.

Crystal, David. 1988. “Another Look at, Well, You Know….” *English Today* 4 (1): 47–49.

Degand, Liesbeth, and Henk Pander Maat. 2003. “A Contrastive Study of Dutch and French Causal Connectives on the Speaker Involvement Scale.” *LOT Occasional Series* 1: 175–99.

Dixon, Robert MW. 2009. “The Semantics of Clause Linking in Typological Perspective.” *The Semantics of Clause Linking: A Cross-Linguistic Typology*, 1–55.

Dobrovoljc, Kaja. 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): 551–82.

Dupont, Maïté, and Sandrine Zufferey. 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): 270–97.

Fraser, Bruce. 2009. “An Account of Discourse Markers.” *International Review of Pragmatics* 1 (2): 293–320.

Furkó, Péter, and Ágnes Abuczki. 2014. “English Discourse Markers in Mediatised Political Interviews.”

Granger, Sylviane. 2015. “Contrastive Interlanguage Analysis: A Reappraisal.” *International Journal of Learner Corpus Research* 1 (1): 7–24.

Hasselgren, Angela. 2002. “Learner Corpora and Language Testing: Smallwords as Markers of Learner Fluency.” *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 143–74.

Hoang, Vu Cong Duy, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. “Iterative Back-Translation for Neural Machine Translation.” In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, 18–24.

Hoek, Jet, Sandrine Zufferey, Jacqueline Evers-Vermeul, and Ted JM Sanders. 2017. “Cognitive Complexity and the Linguistic Marking of Coherence Relations: A Parallel Corpus Study.” *Journal of Pragmatics* 121: 113–31.

Huang, Lan Fen. 2011. “Discourse Markers in Spoken English: A Corpus Study of Native Speakers and Chinese Non-Native Speakers.” PhD Thesis, University of Birmingham.

Joseph, John E. 2009. “Why Lithuanian Accentuation Mattered to Saussure.” *Language & History* 52 (2): 182–98.

Joslyn-Siemiatkoski, Daniel. 2007. “The Cambridge History of Judaism: The Late Roman-Rabbinic Period.” *Theological Studies* 68 (4): 924.

Khayrallah, Huda, Brian Thompson, Kevin Duh, and Philipp Koehn. 2018. “Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation.” In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, 36–44

.

Koehn, Philipp. 2005. “Europarl: A Parallel Corpus for Statistical Machine Translation.” In *MT Summit*, 5:79–86. Citeseer.

Koehn, Philipp, Franz J. Och, and Daniel Marcu. 2003. “Statistical Phrase-Based Translation.” University of Southern California Marina Del Rey Information Science Inst.

Lefer, Marie-Aude, and Natalia Grabar. 2015. “Super-Creative and over-Bureaucratic: A Cross-Genre Corpus-Based Study on the Use and Translation of Evaluative Prefixation in TED Talks and EU Parliamentary Debates.” *Across Languages and Cultures* 16 (2): 187–208.

Mann, William C., and Sandra A. Thompson. 1988. “Rhetorical Structure Theory: Toward a Functional Theory of Text Organization.” *Text* 8 (3): 243–81.

Maschler, Yael, and Deborah Schiffrin. 2015. “Discourse Markers: Language, Meaning, and Context.” *The Handbook of Discourse Analysis* 2: 189–221.

Moeschler, Jacques. 1989. “Pragmatic Connectives, Argumentative Coherence and Relevance.” *Argumentation* 3 (3): 321–39.

Nattinger, James R., and Jeanette S. DeCarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford University Press.

Nau, Nicole, and Norbert Ostrowski. 2010. “Background and Perspectives for the Study of Particles and Connectives in Baltic Languages.” *Particles and Connectives in Baltic*, 1–37.

O’Donnell, William R., and Loreto Todd. 2013. *Variety in Contemporary English*. Routledge.

Oleskeviciene, Giedre Valunaite, 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*, 2155:53–58.

Pit, Mirna. 2007. “Cross-Linguistic Analyses of Backward Causal Connectives in Dutch, German and French.” *Languages in Contrast* 7 (1): 53–82.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. “The Penn Discourse TreeBank 2.0.” In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Schiffrin, Deborah. 1987. *Discourse Markers*. 5. Cambridge University Press.

———. 2001. “Discourse Markers: Language, Meaning, and Context.” *The Handbook of Discourse Analysis* 1: 54–75.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Siyanova-Chanturia, Anna, Kathy Conklin, and Walter JB Van Heuven. 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.

Snyder, Benjamin, Regina Barzilay, and Kevin Knight. 2010. “A Statistical Model for Lost Language Decipherment.” In . Association for Computational Linguistics.

Stede, Manfred, 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)*, 1051–58.

Svartvik, Jan. 1980. “Well in Conversation.” *Studies in English Linguistics for Randolph Quirk* 5: 167–77.

Tan, Xu, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2018. “Multilingual Neural Machine Translation with Knowledge Distillation.” In *International Conference on Learning Representations*.

Tobin, Yishai. 2001. “Gender Switch in Modern Hebrew.” Gender across Languages: The Linguistic Representation of Women and Men 1: 177–98.

Webber, Bonnie, and Aravind Joshi. 2012. “Discourse Structure and Computation: Past, Present and Future.” In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, 42–54.

Webber, Bonnie, 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)*, 22–31.

Wei, Naixing, and Jingjie Li. 2013. “A New Computing Method for Extracting Contiguous Phraseological Sequences from Academic Text Corpora.” *International Journal of Corpus Linguistics* 18 (4): 506–35.

Wray, Alison. 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): 231–54.

Xiong, Hao, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. “Modeling Coherence for Discourse Neural Machine Translation.” In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:7338–45.

Zeyrek, Deniz, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. “TED Multilingual Discourse Bank (TED-MDB): A Parallel Corpus Annotated in the PDTB Style.” *Language Resources and Evaluation*, 1–27.

Zhang, Yuqi, Kui Meng, and Gongshen Liu. 2019. “Paragraph-Level Hierarchical Neural Machine Translation.” In *International Conference on Neural Information Processing*, 328–39. Springer.

Zinkevičius, Vytautas, Vidas Daudaravičius, and Erika Rimkutė. 2005. “The Morphologically Annotated Lithuanian Corpus.” In *Proceedings of The Second Baltic Conference on Human Language Technologies*, 365–70.

Zufferey, Sandrine, and Bruno Cartoni. 2012. “English and French Causal Connectives in Contrast.” *Languages in Contrast* 12 (2): 232–50.

Zufferey, Sandrine, and Liesbeth Degand. 2017. “Annotating the Meaning of Discourse Connectives in Multilingual Corpora.” *Corpus Linguistics and Linguistic Theory* 13 (2): 399–422.

Giedrė Valūnaitė-Oleškevičienė
Mykolas Romeris University
Ateities 20
LT-08303, Vilnius, Lietuva
e-mail: gvalunaite@mruni.eu

Chaya Liebeskind
Jerusalem College of Technology
21 Havaad Haleumi str.
9116001, Jerusalem, Israel
e-mail: liebchaya@gmail.com

In SKASE Journal of Translation and Interpretation [online]. 2022, vol. 15, no. 2 [cit. 2022 12-12]. Available online at http://www.skase.sk/Volumes/JTI23/pdf_doc/03.pdf. ISSN 1336-7811.