# Multiword expressions as discourse markers in Hebrew and Lithuanian

**Giedre Valunaite Oleskeviciene**
Institute of Humanities,
Mykolas Romeris university
Ateities 20 LT-08303,
Vilnius, Lietuva
gvalunaite@mruni.eu

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

## Abstract

Multiword expressions are of key importance in language generation and processing. Certain multiword expressions also could operate as discourse markers. In this research, 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. After establishing a full list of multiword discourse markers in our generated parallel corpus, we focused on the two most frequent ones functioning as stance attitudinal discourse markers: *I think* and *you know* aiming to research if they demonstrate their functional stability as stance attitudinal discourse markers in translation and what changes they undergo in Lithuanian and Hebrew translation. Our research proves that the examined multiword discourse markers preserve their function as stance attitudinal discourse markers and tend to remain multiword discourse markers in the Hebrew translation but turn into one-word discourse markers in Lithuanian due to the translation tendency relying on inflections.

## 1 Introduction

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 et al., 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 (DMs). Phrases such as *you know* and *I think* have also been classified as DMs that perform certain discourse organising functions. However, Maschler and Schiffrin (2015) observe that there is no a priori theoretical classification of DMs and the analysis of function in the data is necessary. Research on DMs 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 DMs in a corpus of spoken Slovene, identifying structurally fixed discourse marking multiword expressions.

The underlying assumption is that DMs *I think* and *you know* are indicators of stance in discourse used to express and understand points of view and beliefs. The purpose of the current research is to examine multiword expressions used as DMs in TED talk English transcripts focusing on stance attitudinal DMs *I think* and *you know* and compare them with their counterparts in Lithuanian and Hebrew by following Maschler and Schiffrin (2015) observation on the necessity of closer investigation on their function as stance DMs. To achieve the aim of the research, the set objectives were to create a parallel research corpus to identify multiword expressions used as stance attitudinal DMs and to analyse their translations in Lithuanian and Hebrew to determine if they function as stance DMs 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); the created corpus is shared and interlinked via CLARIN open language resources. What is more, the current research could be extended to other languages. The future research envisions applying machine learning and using the model for discourse marker identification in other languages to research how stance signalling is treated.

## 2 Theoretical background

The literature overview briefly takes into account the research languages, studies related to multiword expressions and their use as DMs, the importance of DMs for discourse management, and certain insights into DM translation.

### 2.1 Cultural heritage and research languages

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 Semitic language and 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.

### 2.2 Multiword expressions as DMs

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 and 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 DMs (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. DMs 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 DM signalled fluency contributes to interaction and even makes the speaker sound more 'native-like'. Recently, discourse relations and DM 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).

Language, especially spoken, is characterised by DM 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, DMs 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, DMs, 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 DMs 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).

### 2.3 Translation issues of DMs

DMs 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. DMs have been researched by applying various theoretical approaches, such as Rhetorical Structure Theory (Mann and Thompson, 1988), Segmented Discourse Representation Theory (Asher et al., 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 and Cartoni (2012) observed, multilingual studies are more complicated as languages differ in the use of DMs and their expression. The authors also added that often DMs are poly-semic, which means that a single expression of a DM 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 on English parliamentary debates translated into Dutch, German, French, and Spanish, searching what types of DMs might have a higher tendency to be more frequently omitted in translation. Baker (2018), in her extensive studies on translation, observed that DMs 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 DM signalling the causality relation. Further, different languages demonstrate different tendencies – some languages prefer using simpler structures connected by a variety of DMs, while other languages favour complex structures, sparsely using explicit DMs. The author analysed the example of an evident difference between English and Arabic, identifying that, while English prefers signalling discourse relation through DMs, Arabic prefers grouping the information into bigger grammatical chunks and using fewer DMs. The finding is supported by Al-Saif and 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 DMs 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 DMs 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 עשר 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.

### 2.4 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 et al., 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 and Zufferey (2017) identified that such talks contain features of both spoken and written language, as they are semi-prepared speeches by nature. Additionally, Lefer and 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 et al., 2019; Chu et al., 2017; Hoang et al., 2018; Khayrallah et al., 2018; Tan et al., 2018; Xiong et al., 2019; Zhang et al., 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.

## 3 Research methodology

The detailed description of the research procedures is provided in the research methodology section. 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 research aim comprised examining multiword expressions used as DMs 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 DMs 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 DMs 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 (2011) theoretical analysis of DM characteristics for spoken discourse, for example, *you know*, *you see*, *I mean*, *I think*, was also included.

### 3.1 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 DMs from the theoretically based list were extracted and matched with their Lithuanian and Hebrew counterparts. The process of creating the parallel corpus 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`).

### 3.2 Multiword DM extraction

Another stage of the research focuses on multiword expressions that are used as DMs 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 DM. One-word DM 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 DM, the context in which it occurs should be examined by identifying if the expression serves as a DM. 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 DMs 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 DMs, 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

Jis tikrai vaikšto su baltu katinu.

Of course he walks with a white cat.

Figure 2: English – Hebrew phrase alignment

In my opinion, you will not regret your quick decision.

לדעתי, לא תתחרט על החלטתך המהירה.

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

## 4 Research findings

### 4.1 Multiword DM distribution

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

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 DMs. The manual annotation revealed that some multiword expressions were used as DMs more frequently while others were more often used as content words fully integrated into sentences.

| 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 |

Table 1: Multiword expressions in the corpus

| Multiword expression | 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 |

Table 2: Multiword expressions used as DMs

It is visible in Table 2 that multiword expressions *That is* and *You see* although identified as DMs by the theoretical literature, in this study, they demonstrate a weak tendency to be used as DMs 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 DMs and the stability of remaining DMs in Lithuanian and Hebrew translation.

### 4.2 DM 'I think' translations

Further, following our research aim, we present a detailed analysis of the translations of the two most frequent multiword expressions used DMs – *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 DM, *I think*.

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 et al., 2005), fully represents the verb-pronoun cases. Other one-verb variants and multiword expressions do not demonstrate high numbers. 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 assumption could be that the choice of gender in first person pronouns depends on the gender of the speaker. 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 and Carston (1999), and Moeschler (1989). Hebrew demonstrates less omission cases than Lithuanian for the DM *I think*. The number of omissions in Hebrew is 23, while the Lithuanian omission number is approximately double in the parallelized corpus sentences.

### 4.3 DM 'you know' translations

Another commonly used multiword DM, *you know*, demonstrates far more variable translations. A closer investigation into the translations of DM *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 supran-tate*, or a single particle *juk*. Even a single particle is used as a DM, 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 and Ostrowski (2010b) 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 DM 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. 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). In Hebrew, this DM 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 DM *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 Lithuanian, eight cases of grammatical changes were found and, even amongst those, one-word DMs prevail. The multiword DM *you know* is translated also into a connective, *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 connectives used successively, כאילו (as if). There are also some cases when a connective 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 DM used or possibly attaches a rhetorical function to the integrated connective. Even among the limited cases of grammatical transformation, multiword expressions as DMs 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 DMs 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 DMs 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 in English and the choice of the gender of the derivative is completely the translator's choice. Another observation regarding Hebrew is that multiword DMs 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 DMs turn into one-word DMs.

## 5 Conclusions and Future research

The study results showed that English multiword expressions *I think* and *you know*, identified as DMs according to Maschler and Schiffrin (2015) function-based approach, remain stance attitudinal DMs 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 translation there is a tendency to use multiword discourse marker translations to express stance, 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, in Lithuanian, there is a clear tendency observed for one-word DMs in translation. One-word translations mainly include verbs, which, due to Lithuanian being a highly inflected language (Zinkevičius et al., 2005), fully represent the verb-pronoun cases. It should be noted that Lithuanian translations of pronoun-verb multiword expressions and one-word verb cases could be considered almost word-for-word translations. Concerning translation modelling the research reveals stance signalling in discourse preserved as an important element in translation.

More interesting cases include translator choices of particle or connective integration into multiword expressions. The integration of particles for Lithuanian and connectives for both languages might carry the pragmatic meaning that could have been inferred from the surrounding contexts by the translators (Nau and Ostrowski, 2010a; Blakemore and Carston, 1999; Moeschler, 1989), or translator choices might be also guided by the inner discourse managing system of the target language. 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.

## Acknowledgments

The research described in this paper was conducted in the context of the COST Action CA18209 "Nexus Linguarum" ("European network for Web-centred linguistic data science").

## References

Roee Aharoni, 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)*, pages 3874–3884.

Amal Al-Saif and Katja Markert. 2010. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *LREC*, pages 2046–2053.

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

Mona Baker. 2018. *In other words: A coursebook on translation*. Routledge.

Michael Barlow. 2011. Corpus linguistics and theoretical linguistics. *International Journal of Corpus Linguistics*, 16(1):3–44. Publisher: John Benjamins.

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

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

Douglas Biber, Susan Conrad, and Randi Reppen. 1994. Corpus-based approaches to issues in applied linguistics. *Applied linguistics*, 15(2):169–189. Publisher: Oxford University Press.

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

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

Diane Blakemore and Robyn Carston. 1999. The interpretation of and-conjunctions. *Iten, C. &*.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Conference of european association for machine translation*, pages 261–268.

Chenhui Chu, 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)*, pages 385–391.

Tom Cobb and Alex Boulton. 2015. *Classroom applications of corpus analysis*.

David Crystal. 1988. Another look at, well, you know…. *English Today*, 4(1):47–49. Publisher: Cambridge University Press.

Liesbeth Degand and Henk Pander Maat. 2003. A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale. *LOT Occasional Series*, 1:175–199. Publisher: LOT, Netherlands Graduate School of Linguistics.

Robert MW Dixon. 2009. The semantics of clause linking in typological perspective. *The semantics of clause linking: A cross-linguistic typology*, pages 1–55. Publisher: Oxford University Press Oxford.

Kaja Dobrovoljc. 2017. Multi-word discourse markers and their corpus-driven identification: The case of MWDM extraction from the reference corpus of spoken Slovene. *International journal of corpus linguistics*, 22(4):551–582. Publisher: John Benjamins.

Maïté Dupont 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):270–297. Publisher: John Benjamins.

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

Péter Furkó and Ágnes Abuczki. 2014. English discourse markers in mediatised political interviews.

Sylviane Granger. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1):7–24. Publisher: John Benjamins.

Angela Hasselgren. 2002. Learner corpora and language testing: Smallwords as markers of learner fluency. *Computer learner corpora, second language acquisition and foreign language teaching*, pages 143–174. Publisher: John Benjamins Amsterdam, The Netherlands.

Vu Cong Duy Hoang, 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*, pages 18–24.

Jet Hoek, 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*, 121:113–131. Publisher: Elsevier.

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

John E. Joseph. 2009. Why Lithuanian accentuation mattered to Saussure. *Language & History*, 52(2):182–198. Publisher: Taylor & Francis.

Daniel Joslyn-Siemiatkoski. 2007. The Cambridge history of Judaism: The late Roman-rabbinic period. *Theological Studies*, 68(4):924. Publisher: Sage Publications Ltd.

Huda Khayrallah, 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*, pages 36–44.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. Technical report, University of Southern California Marina Del Rey Information Science Inst.

Marie-Aude Lefer 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. Publisher: Akadémiai Kiadó.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281. Publisher: Berlin.

Yael Maschler and Deborah Schiffrin. 2015. Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*, 2:189–221. Publisher: Wiley Online Library.

Jacques Moeschler. 1989. Pragmatic connectives, argumentative coherence and relevance. *Argumentation*, 3(3):321–339. Publisher: Springer.

James R. Nattinger and Jeanette S. DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford University Press.

Nicole Nau and Norbert Ostrowski. 2010a. Background and perspectives for the study of particles and connectives in Baltic languages. *Particles and connectives in Baltic*, pages 1–37. Publisher: Vilnius: Vilniaus universitetas, Academia Salensis.

Nicole Nau and Norbert Ostrowski. 2010b. Particles and connectives in Baltic.

William R. O'Donnell and Loreto Todd. 2013. *Variety in contemporary English*. Routledge.

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

Mirna Pit. 2007. Cross-linguistic analyses of backward causal connectives in Dutch, German and French. *Languages in Contrast*, 7(1):53–82. Publisher: John Benjamins.

Rashmi Prasad, 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).

Deborah Schiffrin. 2001. Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*, 1:54–75. Publisher: Wiley Online Library.

John Sinclair. 1991. *Corpus, concordance, collocation*. Oxford University Press.

Anna Siyanova-Chanturia, 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. Publisher: American Psychological Association.

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

Jan Svartvik. 1980. Well in conversation. *Studies in English Linguistics for Randolph Quirk*, 5:167–177. Publisher: London: Longman.

Xu Tan, 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*.

Yishai Tobin. 2001. Gender switch in modern Hebrew. *Gender across languages: The linguistic representation of women and men*, 1:177–198.

Bonnie Webber 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*, pages 42–54.

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

Naixing Wei and Jingjie Li. 2013. A new computing method for extracting contiguous phraseological sequences from academic text corpora. *International Journal of Corpus Linguistics*, 18(4):506–535. Publisher: John Benjamins.

Alison Wray. 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. Publisher: Cambridge University Press.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345.

Yuqi Zhang, Kui Meng, and Gongshen Liu. 2019. Paragraph-Level Hierarchical Neural Machine Translation. In *International Conference on Neural Information Processing*, pages 328–339. Springer.

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

Sandrine Zufferey and Bruno Cartoni. 2012. English and French causal connectives in contrast. *Languages in contrast*, 12(2):232–250. Publisher: John Benjamins.

Sandrine Zufferey and Liesbeth Degand. 2017. Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*, 13(2):399–422. Publisher: De Gruyter Mouton.