ISSN 1392-0340
E-ISSN 2029-0551

Pedagogika / Pedagogy
2018, t. 131, Nr. 3, p. 99–109 / Vol. 131, No. 3, pp. 99–109, 2018

LIETUVOS EDUKOLOGIJOS UNIVERSITETAS

TED-MDB Multilingual Discourse-annotated Corpus for Language Learning and Teaching

Giedrė Valūnaitė-Oleškevičienė[1], Andrius Puksas[2], Dalia Gulbinskienė[3]

[1] Mykolas Romeris University, Lithuania, gentrygiedre@gmail.com
[2] Mykolas Romeris University, Lithuania, andrius_puksas@mruni.eu
[3] General Jonas Žemaitis Military Academy of Lithuania, Lithuania, dalia.gulbinskiene@lka.lt

Abstract. The newest research has proved the importance of discourse for second language learning and teaching at more advanced levels. TED-MDB multilingual discourse-annotated corpus, created within the framework of TextLink, COST action IS1312, appears to be a handy tool for illuminating qualitative differences between the first and the second language discourse marker use. The current research focuses on the pragmatic use of English discourse marker *and* with its Lithuanian counterpart *ir* paying attention to the cases when the connectives *and/ir* perform different functions, e. g., discourse structuring, etc. It was established by the research that the use and translation of the connectives vary. They may be translated by the discourse markers provided by dictionaries, or might be transferred into a different discourse marker or are simply omitted. The research leads to a conclusion that language learners and teachers may use the corpus resources for raising linguistic awareness about the pragmatic use of discourse markers.

Keywords: *Discourse marker; corpus based research; pragmatic; teaching and learning; translation.*

Introduction

Globalization has been shaping our world and increasing the need for mastering languages to ensure international communication. The international relations induce the need for mastering languages at more advanced levels to ensure communication among people coming from different cultural settings. The importance of discourse

ISSN 1392-0340
E-ISSN 2029-0551

awareness becomes essential while teaching and learning languages at more advanced levels as successful discourse management is closely related to coherence, cohesion and textual rhetorical features. The development of corpora and corpus based research which is focused on researching various linguistic patterns including pragmatics and textual features. Discourse annotated corpora illuminate qualitative differences between the first language and the second language discourse marker use, especially in the complex cases and could be used as supplemental teaching/learning material for raising discourse management awareness of more advanced learners. The object of the research is English discourse marker *and* with its Lithuanian counterpart *ir*, the research of which my illuminate the importance of linguistic discourse awareness while teaching and learning languages at more advanced levels.

The present research is aimed at revealing the pragmatic use of English discourse connective *and* with its Lithuanian counterpart *ir* by comparing the discourse marker *and* with its Lithuanian counterparts and analyzing the translations of *and* into Lithuanian. To achieve this aim the following objectives have been set: 1) to compare discourse marker *and* with its Lithuanian counterparts by applying Crible's (2017) taxonomy of domains and functions of discourse markers; to analyze the translations of *and* into Lithuanian by examining English transcripts of TED talks and their Lithuanian counterparts. Since the research demonstrates the complexity of the connective pragmatics and peculiarities of translation and reveals the importance of raising pragmatic textual awareness in language teaching and learning it naturally leads to supporting the idea of direct corpus use in language teaching and learning.

Theoretical background

According to the Oxford English Dictionary, the first recorded written example of the word corpus, understood as "the body of written or spoken material upon which a linguistic analysis is based", dates back to 1956. There are many ways to define a corpus, but most scholars agree that a corpus is a collection of machine-readable, authentic texts, chosen to characterize or represent a state or variety of a language. Corpus linguistics can be described as the study of language based on text corpora.

The development of large language databases known as corpora revealed the potential of language research using the corpus techniques focused on researching patterns of lexis, grammar, semantics, pragmatics, and textual features. Many corpora are coded according to the parts of speech or analyzed for grammatical structure, or examined focusing on the pragmatic features.

The applications of corpora for language teaching have been discussed, for example, by Leech (1997), Römer (2008) and McEnery and Xiao (2011), who differentiate between indirect and direct uses. Corpora are being indirectly used, for example, for the design of

ISSN 1392-0340
E-ISSN 2029-0551

teaching syllabi with an emphasis on communicative competence, or when representing the frequency of occurrence of language items in grammar and usage handbooks. Other indirect applications are found in Language for Specific Purposes (LSP) corpora, learner corpora and translation corpora, each with different implications for the language classroom. On the other hand, learner and translation corpora are two of the most widely employed upshots of corpus linguistics for language pedagogy, and offer a variety of practical uses, such as learner dictionaries, syllabus design or the creation of teaching materials based on error analysis. As regards the direct applications of corpora, scholars have often reported positive results when students are faced with hands-on tools such as online corpora or when they are able to retrieve and discuss concordance lines on a relevant topic. It seems that this kind of data-driven learning furthers an autonomous and interactive kind of learning between students and language data, while teachers are able to move from the role of information provider to that of facilitator.

Corpora development has enriched the knowledge concerning lexis, grammar, semantics, pragmatics, and textual features. (Sinclair, 1991; Stubbs, 2004). Corpus linguistics is based on the theory that language varies according to the context related to space and time, which sustain the infinite potential for establishing new facts about language. If the same theoretical insight is applied to language teaching and learning practices, then the use of corpora in teaching and learning languages becomes very significant. Dictionaries and grammars do not have the capacity to fully describe the language. So while applying corpora for teaching and learning languages both teachers and learners can identify certain regularities and irregularities of the language relaying on the corpora data. Also, according to Aston (2001) another benefit is that corpus-based approach provides real data of real language used in certain contexts. The author also stresses the importance of the frequency information which might be helpful while making teaching/learning choices. Scott and Tribble (2006) observe that at more advanced levels it is also important to acquire certain knowledge of genres and registers. The idea is well-supported by the learner corpus research (Granger, 2015) which reveals that most patterns used by relatively advanced language learners exemplify stylistic discrepancies rather that grammatical problems. The problematic areas for advanced language learners include coherence, cohesion and textual rhetorical features. Cohesive devices, discourse markers attract researcher attention as the tools for ensuring textual and discourse management looking for answers what and how could be taught at more advanced levels concerning the matters of textual features. Recent research suggestions have lead to the idea of direct corpus use by language learners sand teachers. The studies by Cobb and Boulton (2015) have shown that the application of such an innovative idea in teaching and learning has proved to be effective and efficient. The authors have revealed that learners better acquired linking adverbs by using corpus concordances rather than using bilingual dictionaries or grammars. The development of discourse-annotated corpora could present a substantial

supplement in the surplus of teaching/learning materials, especially for more advanced learners in dealing with textual cohesion and coherence.

TED-MDB multilingual discourse-annotated corpus created by Deniz Zeyrek and Amalia Mendes within the framework of TextLink, COST action IS1312 appears to be a handy tool for illuminating qualitative differences between the first language and the second language discourse marker use, especially in the complex cases of discourse markers (Zeyrek, 2017). The corpus provides invaluable resources both for language learning/teaching and research, and is still a developing corpus extended by other TextLink action languages. It was used in the present research as a source for research data.

Research methodology

The research consisted of two stages: 1) comparing discourse marker *and* with its Lithuanian counterparts by applying Crible's taxonomy of domains and functions of discourse markers. 2) analyzing the translations of *and* into Lithuanian found in TED talks translations. The general approach proposed by Crible (2017) describes discourse markers as functioning in four "domains": ideational – related to real-world events, rhetorical – related to expressing the speaker's subjectivity and metadiscursive effects, sequential – concerning the structuring of local and global units of discourse and interpersonal – related to managing the speaker-hearer relationship. According to Crible (2017) the four domains correspond to overall discourse intentions or entities, which depend on what the speaker is targeting: content (ideational), illocutionary value (rhetorical), discourse structure (sequential) or inter-subjective inferences (interpersonal). While applying Crible's (2017) revised taxonomy, annotators can choose to start at domain-level or function-level, to annotate both levels simultaneously or independently, and could even decide to stop at one level if a particular discourse marker value is under-specified for the other level. This feature makes the approach more flexible than inter-dependent or hierarchical taxonomies, which was the main reason guiding the choice of the annotation scheme for the research.

Concerning the approach to analysis of translation, according to Noël (2003) the theoretical background of translation spotting is that differences in translation could be used to reveal semantic features of the source language or translation could be used to elicit some semantic features of content words in the source language. However, Behrens, and Fabricius-Hansen (2003) observe that using translated data can also help to identify the semantic features of the discourse markers denoting coherence relations since the translation relies on the decisions made by the translators, who are experts in their own languages, and they make translation choices according to the entire context of the whole text and their professional knowledge in the target language. Danlos and Roze (2011) suggest that it is difficult to spot automatically and translation spotting is

ISSN 1392-0340
E-ISSN 2029-0551

preferably carried out manually because there exit a number of possible translations of the connectives, ranging from various paraphrases and syntactic constructions to no translation or omission. Translation spotting gives an interesting view of the existing discrepancies between the languages, especially in the case of connectives or discourse markers, when there are no one-to-one translation equivalences.

Research findings

Sentences in the English language were extracted from TED Multilingual Discourse Bank (TED-MDB) of TED talks. The sentences and their Lithuanian counterparts were annotated by applying Crible's (2017) taxonomy of domains and functions of discourse markers and the extracted cases of *and* were analyzed. The diagram presents the research results revealing the prevailing annotated values of the discourse marker *and*.

As it could be seen from the research results that discourse marker *and* with its Lithuanian counterpart *ir* are used approximately equally both in the ideational domain 36.7%, representing factual information, and sequential domain 41.7%, representing the structuring of local and global units of discourse. Also a certain number of occurrences 11.7% are related to rhetorical domain representing the speaker's subjectivity. The research results also reveal that *and* with its Lithuanian counterpart *ir* are closely associated with the function of addition as 36.7% of the occurrences in the annotated sample express ideational addition in the English language and 20% in the Lithuanian language with minor cases of omissions and the Lithuanian counterpart *ir* not functioning as a discourse marker. It is very important to observe that ideational domain is related to real-world events, thus ideational addition expresses additive meaning based on real world facts, for example:

So let's imagine then, that you start dating when you're 15, [and] ideally, you'd like to be married by the time that you're 35.

Įsivaizduokite, kad pradedate susitikinėti kai jums 15, [ir] idealiu atveju norėtumėt susituokti kai jums 35.

The prevalence of sequential domain in the occurrences 41.7% reveal that *and* with its Lithuanian counterpart *ir* are often used for discourse structuring purposes with the purpose to join the bigger units of discourse. That is why

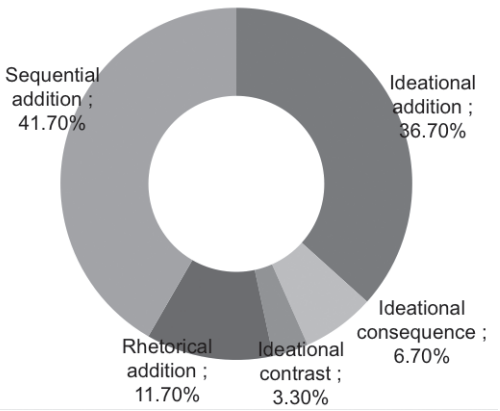

Fig. 1. Annotated values of the discourse marker *and*

ISSN 1392-0340
E-ISSN 2029-0551

it is important to reveal such functions of certain discourse markers, especially while learning and teaching languages at more advanced levels.

We give ourselves a little bit of time to play the field, get a feel for the marketplace or whatever when we're young. [And] then we only start looking seriously at potential marriage candidates once we hit our mid-to-late 20s.

Leidžiam sau šiek tiek išsilakstyti, kol esam jauni, leidžiam suprasti, kas yra rinkoje, ar panašiai. [Ir] tuomet vėlesniame dvidešimtmetyje pradedame į vedybų kandidatus žiūrėti rimtai.

A certain number, 11.7% of occurrences in the sample are associated with rhetorical addition. Rhetorical domain means that discourse markers are used to express the speaker's subjectivity and other meta-discursive effects. It shows that rhetorical addition is related to speaker's subjective perception and gives the effect of subjective discourse management, for example:

There'd be a huge spread in her scores. [And] [actually] it's this spread that counts.
Jos balai būtų visiškai pasiskirstę. [Ir] [išties], svarbus būtent tas pasiskirstymas.

It should be noted that rhetorical subjectivity is also related to the whole argument. The example reveals that the phrase *actually* provides a clear association to subjective perception of the discourse marker *and*. Sometimes it is not so easy to isolate discourse marker from the whole context of the argument.

It is also important to investigate the translation of the connectives. In the case of the discourse marker *and* some possible translations have been spotted out. The diagram presents the translation values of the discourse marker *and*.

As it is shown in the diagram the most frequent variants of translation in the sample is *ir* which is the variant provided by the bilingual English–Lithuanian dictionaries. Also, the variant *o* provided by the bilingual dictionaries is present among the identified values. However, the Lithuanian *o* expresses the meaning of contrast. The examples of the dictionary based translations are provided below.

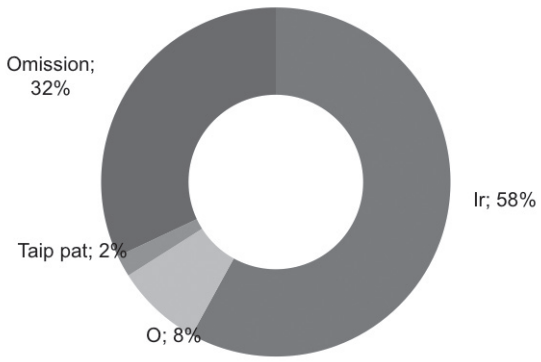

Fig. 2. Translation values of the discourse marker *and*

Okay, so let's imagine then that you picked your perfect partner [and] you're settling into a lifelong relationship with them.

ISSN 1392-0340
E-ISSN 2029-0551

Įsivaizduokime, kad išsirinkote savo idealų partnerį [ir] pradėjote santykius iki gyvenimo galo.

But the question arises of how do you then convert that success into longer-term happiness [and] in particular, how do you decide when is the right time to settle down?

Bet iškyla klausimas, kaip jums tą sėkmę paversti į ilgalaikę laimę, [o] ypač, kaip nuspręsti, kada tinkamas laikas susitupėti?

The example provided above demonstrates the case how the contrastive meaning of the connective *but* used at the beginning of the first sentence also influences the translation of the following discourse marker *and* which is rendered as a contrastive *o* in Lithuanian.

The adverb *taip pat* is also observed translation value in the sample. In the example the translator renders the ideational value of the discourse marker *and* in the source language using the adverb *taip pat,* which expresses another variant of addition and helps to avoid repetition since the discourse structuring marker *now* is already rendered into *ir* at the beginning of the sentence.

[Now] the rules are that once you cash in and get married, you can't look ahead to see what you could have had [and] equally, you can't go back and change your mind.

[Ir] yra taisyklė, kad kai susituokiat, jūs negalite pažiūrėti, ką galėjote turėti. [Taip pat] jūs negalite grįžti ir pakeisti savo sprendimo.

The cases of omission seem to be twofold. There are cases when there is more than one discourse markers used to introduce an argument, which is especially characteristic of spoken-like speech where discourse markers are used abundantly. In such cases just one discourse marker is rendered into the target language which could be the translator's choice because of the requirements of synchronizing the subtitles and making them concise. The example demonstrates the translator's choice to render the temporal discourse marker *while* into a concessive *nors* and omit the sequential addition *and* rendering the concessive meaning, which again is a successful choice having in mind the requirements of the synchronization.

The study found that even in companies with diversity policies and inclusion programs, employees struggle to be themselves at work because they believe conformity is critical to their long-term career advancement. [And] [while] I was surprised that so many people just like me waste so much energy trying to hide themselves, I was scared when I discovered that my silence has life-or-death consequences and long-term social repercussions.

Tyrimas parodė, kad kompanijose, kuriose pripažįstama įvairovė ir skatinama priimti skirtumus, darbuotojai patiria sunkumų stengdamiesi būti savi. [Nors] mane stebino tai, kad tiek daug žmonių kaip aš taip stengėsi slėpti tiesą apie save, aš išsigandau sužinojusi, kad mano tylėjimas gali lemti gyvenimą ar mirtį ir turėti ilgalaikių socialinių pasekmių.

ISSN 1392-0340
E-ISSN 2029-0551

Other cases of omission seem to occur when the translator rendered the source language meanings into different grammatical structures which in their own right required different translator choices in rendering the discourse markers. In the example, it could be observed that that the translator successfully chose to change the whole argument to render the meaning of the source argument in such a way omitting discourse marker *and*.

So it is fitting and scary that I have returned to this city 16 years later [and] I have chosen this stage to finally stop hiding.

Dabar pats laikas ir šiek tiek baisu, kad po 16 metų grįžusi į šį miestą, pasirinkau šią sceną, kad nustočiau slapstytis.

There also were interesting cases observed that omission of a discourse marker in the translated Lithuanian texts was used more often in the case of sequential addition. The phenomenon might be explained by the requirements of synchronizing the subtitles and keeping to the goal of creating subtitles that are easily read, well-rounded bits of text. It means that the translator chooses to omit the discourse structuring and obeying the synchronization requirements and relying on the implicit contextual meaning. Such a feature is also important to discuss while learning and teaching translation peculiarities.

How does her father feel? I don't know, because I was never honest with them about who I am. And that shakes me to the core.

Ką jos tėvas galvoja? To nežinau, nes niekada su jais nekalbėjau apie tai, kas aš esu. Tai mane nepaprastai gąsdina.

The above example demonstrates how sometimes it is difficult for a translator to make a decision by observing synchronysing rules. In the Lithuanian translation the discourse structuring maker *ir* is omitted relying on the contextual meaning; however, it gives the Lithuanian reader an impression of a slight chunky feeling.

The final observation of the research findings points out that most translator choices are really successful in conveying both the semantic and pragmatic values of the discourse markers and also gives an interesting view of the existing linguistic spaces between the languages. All the discussed features become important in teaching language at more advanced levels, especially working with students majoring in translation studies. Raising learner awareness about such a phenomena as discourse structuring, coherence, cohesion and textual rhetorical features, demonstrating how certain discourse markers are used for joining bigger units of discourse allows preparing more advanced language users.

ISSN 1392-0340
E-ISSN 2029-0551

Conclusion

As the research reveals the discourse marker *and* with its Lithuanian counterpart are closely associated with additive meaning as 36,7% of the occurrences in the annotated sample express ideational addition. However, the prevalence of sequential domain in the occurrences 41,7% reveal that *and* with its Lithuanian counterpart *ir* are often used for discourse structuring purposes with the purpose to join the bigger units of discourse. Also, a certain number of 11,7% of occurrences in the sample are associated with rhetorical addition which is related to the expression of the speaker's subjectivity Such results demonstrate how important it is to reveal such functions of certain discourse markers, especially while learning and teaching languages at more advanced levels. The most frequent variant of translation in the sample is *ir* and *o* which are the variants provided by the bilingual English – Lithuanian dictionaries. Totally the dictionary provided choices make up 64% of the values in the sample. Also, omission technique is used abundantly observing the requirements of synchronizing the subtitles and keeping to the goal of creating subtitles that are easily read, concise bits of text. Such features also become important in teaching language at more advanced levels, especially working with students majoring in translation studies.

Acknowledgements

This research is funded by the European Social Fund under the No 09.3.3-LMT-K-712 "Development of Competences of Scientists, other Researchers and Students through Practical Research Activities" measure.

Also, for training in annotation and generating ideas for future research we acknowledge the support of TextLink COST action IS1312.

References

Abdulhay, H. (2015). Panorama of Constructivism in Language Education. *International Journal of Pedagogy, Innovation and New Technologies*, *2*(2), 73–77. DOI: 10.5604/23920092.1187859

Aston, G. (2001). Learning with corpora: An overview. In G. Aston (ed.), *Learning With Corpora* (pp. 7–45). Houston: Athelstan.

Behrens, B., & Fabricius-Hansen, C. (2003). Translation equivalents as empirical data for semantic/pragmatic theory. In K. Jaszczolt & J. Turner (eds.), *Meaning through Language Contrast* (pp. 463–477). Amsterdam: Benjamins.

Cartoni, B., Zufferey, S., & Meyer, T. (2013). Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue* & *Discourse*, *4*(2), 65–86.

Cobb, Th., & Boulton, A. (2015). Classroom Applicatios of Corpus Analysis. In D. Biber-Reppen (ed.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 478–497). Cambridge, Cambridge University press.

Crible, L. (2017). *Discourse markers, (dis)fluency and the non-linear structure of speech: a contrastive usage-based study in English and French*. Louvain-la-Neuve: Université catholique de Louvain.

Danlos, L., & Roze, C. (2011). Traduction (automatique) des connecteurs de discours. In *Proceedings of TALN 2011*, Montpellier, France.

Granger, S. (2015). Contrastive interlanguage analysis: a reappraisal. *International Journal of learner Corpus Research*, *1*, 7–24.

Houterman, M. (2015). Learning a second language in theory and practice. A teacher's perspective on the importance of spoken language in a multilingual classroom, *International Journal of Pedagogy, Innovation and New Technologies*, *2*(1), 50–61. DOI: 10.5604/23920092.1159140

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

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Teaching*. Amsterdam/Philadelphia: John Benjamins.

Sinclair, J. M. (1991). *Corpus, Concordance Collocation*. Oxford: Oxford University Press.

Stubbs, M. (2004). Language corpora. In A. Davies & C. Elder (eds.), *The Handbook of Appliedlinguistics* (pp. 106–32). Oxford: Blackwell.

Zeyrek, D. (April 3rd, 2017). TED Multilingual Discourse Bank (TED-MDB): A parallel annotated in the PDTB style. In *11th Linguistic Annotation Workshop (LAW)*. European Chapter of the Association of Computational Linguistics., Valencia. Retrieved from https://www.researchgate.net/publication/296694747_Introduction_The_use_of_corpora_for_language_teaching_and_learning [accessed Jul 02 2018].

ISSN 1392-0340
E-ISSN 2029-0551

TED-MDB daugiakalbio diskurso ryšiais anotuoto tekstyno naudojimas anglų kalbos mokymui(si)

Giedrė Valūnaitė-Oleškevičienė[1], Andrius Puksas[2], Dalia Gulbinskienė[3]

[1] Mykolo Romerio universitetas, gentrygiedre@gmail.com
[2] Mykolo Romerio universitetas, andrius_puksas@mruni.eu
[3] Generolo Jono Žemaičio Lietuvos karo akademija, dalia.gulbinskiene@lka.lt

Santrauka

Naujausiais tyrimais įrodyta, kad svarbu kalbėti apie antrosios kalbos mokymą(si) pažengusio vartotojo lygmeniu. TED-MDB daugiakalbio diskurso ryšiais anotuotas tekstynas naudojamas pagal TextLink, COST veiksmus IS1312 ir yra patogi priemonė, nurodanti kokybinius pirmosios ir antrosios kalbos diskurso žymeklių vartojimo skirtumus. Dabartiniai tyrimai skirti pragmatiniam anglų kalbos diskurso žymeklių vartojimui nustatyti, kartu atkreipiant dėmesį į atvejus, kai jungtukai and / ir atlieka skirtingas funkcijas, pvz., diskurso struktūrizavimo funkciją ir kt. Šiuo tyrimu buvo nustatyta, kad jungtukų vartojimas ir vertimas skiriasi. Jie gali būti išversti žodynuose pateikiamais diskursų žymenimis, pakeisti kitu diskursų žymekliu arba yra tiesiog praleisti. Tyrimas leidžia daryti išvadą, kad besimokantieji anglų kalbą ir mokytojai turėtų naudoti tekstyno išteklius, skatinančius kalbinį supratimą apie pragmatinį diskurso žymenų vartojimą.

Esminiai žodžiai: *diskurso žymeklis, tekstyno tyrimai, pragmatiškas, mokymas ir mokymasis, vertimas.*

Gauta 2018 03 14 / Received 14 03 2018
Priimta 2018 09 03 / Accepted 03 09 2018