ISSN 1392-0340 (Print)
ISSN 2029-0551 (Online)
https://doi.org/10.15823/p.2022.145.7

Pedagogika / Pedagogy
2022, t. 145, Nr. 1, p. 117–134 / Vol. 145, No. 1, pp. 117–134, 2022

VYTAUTO DIDŽIOJO
UNIVERSITETO
ŠVIETIMO
AKADEMIJA

Lithuanian Discourse Markers in Parallel Corpus for Teaching Translation Awareness

Giedrė Valūnaitė Oleškevičienė[1], Vitalija Karaciejūtė[2], Dalia Gulbinskienė[3], Nagaletchimee Annamalai[4]

[1] Mykolas Romeris University, Ateities g. 20, LT-08303, Vilnius, Lithuania, gvalunaite@mruni.eu
[2] Mykolas Romeris University, Ateities g. 20, LT-08303 Vilnius, Lithuania, vitalija.karaciejute@mruni.eu
[3] Vilnius Gediminas Technical University, Saulėtekio al. 11, LT-10223 Vilnius, Lithuania, dalia.gulbinskiene@vilniustech.lt
[4] Malaysia University of Science, 11800 USM Penang, Malaysia, naga@usm.my

Annotation. The contrastive analysis of discourse markers in Lithuanian and English provides data on the functions of discourse markers to language teachers and translators. The study reveals that there is a tentative tendency for translators to rarely choose to run an explicit discourse marker and leave it only in the implied translation. It is essential that philology students are taught in detail about discourse markers, and TED-MBD is an excellent medium for teaching text coherence.

Keywords: *discourse markers, discourse relations, parallel corpus, annotation, PDTB sense hierarchy, translation.*

Introduction

The development, research, and application of discourse annotated corpora is a comparatively new research area, which also could be used in teaching translation awareness (Kubler & Foucou, 2003; Boulton & Tyne, 2014). Effective discourse management in any language is characterized by clear connections between sentences and a cohesive, coherent language structure. However, in different languages, the connections and structure of discourse are ensured by different linguistic means. Various dictionaries and grammar textbooks introduce the peculiarities of words and sentences, and the connections of discourse layer still lack being discussed. It should also be noted that discourse research raises awareness of pragmatic categories, not just typically relying on grammatical lists of conjunctions to describe certain functions of text concatenation and coherence (Crible & Degand, 2019). Working in this field of knowledge, teams of scholars from different countries create lexicons of their language discourse markers (Roze et al. (2010) developed a lexicon of French discourse markers). Lithuanian researchers are also beginning to delve into the comparative research of discourse connections and discourse markers in Lithuanian and other languages on the basis of textual data (Šolienė, 2018). The aim of the article is to present the multilingual discourse-annotated corpus (TED-MDB), developed in collaboration with the international scientific community, and reveal its value to discourse awareness in translation studies. Each text coherence is a very important but often overlooked area in the study of philology and translation, as senior students and prospective philologists must be more widely taught the peculiarities of text coherence and discourse markers. Also, it is expected that researchers will get more widely interested in discourse marker research in the Lithuanian language, as has already been done in other languages, which would be a significant contribution to the development of modern resources in the Lithuanian language and also a vital aid for translation studies and translators.

Discourse relations in text

Theoretical background

Discourse markers, or connecting discourse elements, form a functional category of lexical elements that are used to denote relationships between units of text or discourse that provide text coherence, such as explanation, contrast, and so on. (Mann &Thomson, 1988; Sanders, 2000; Hunston, 2002). Although most languages have sets of such elements, the number of connecting elements, their use, and the relationships of discourse expressed vary widely. In addition, a well – known feature of discourse markers is that they are often multifunctional and can convey multiple discourse relationships. For example, the English discourse marker since can convey not only causal but also temporal meaning. However, in Lithuanian these two meanings require different translations. In some cases, the same link is conveyed by different discourse markers. The literature suggests that some languages tend to express discourse relationships with implicit unexpressed discourse markers, leading to a more complex perception of coherence in the language structure, while others prefer to use explicit discourse markers that mark discourse relationships between structures as clearly visible, resulting in simpler coherence. For example, Baker (2011), discussing the difference between languages, considers that some languages prefer to present information in smaller parts of discourse using explicit discourse markers that clearly signal discourse connections. In other languages, large groups of discourse are preferred, using less pronounced discourse markers, leaving more implicit discourse relations. Thus, the question arises as to how translators deal with discourse relationships when there are many expressed discourse markers in the source text, or, conversely, how they convey discourse relations when there are a limited number of discourse markers in the source text.

Given that discourse markers relate to the logic and interpretation of the text, the process of matching discourse markers, given the specifics of the target language and the type of text in the target language, is a complex process. Translators can choose certain strategies. For smooth and clear translation, they may try to insert additional discourse markers, even if they are not used in the original text, or they may choose to translate the original text discourse markers literally, even though the resulting translation into the target language may seem foreign to that language. In practice, translators tend to choose to either use one of the strategies mentioned or to seek balance and use some of all the methods mentioned (Baker, 2011). In addition, the translation process involves procedures for translating words or syntactic structures from one language to another in a variety of ways. Such a transfer involves transformations that help to understand the main meaning in the target language. Grammatical transformation involves alteration, transfer, omission, and addition. An amendment means the use of a different word or expression to convey an identical meaning in the target language, and a transfer involves a change in the structure of a sentence translated into the target language. The omission hardly needs an explanation; suffice it to note that this article deals with the abandonment of the expressed discourse marker as a connecting element in the target text. Finally, addition means the use of additional words to better reveal the context.

As for the discourse relationship annotation process, in cases where the discourse marker is missing as a connecting element in the target language (the discourse marker is implicit only), the annotator must select the discourse marker as a connecting element and thus can either mark the same discourse meaning as the original or to choose a different meaning of a discourse relation. Furko (2020) claims that using explicit discourse markers in the target text can lead to annotating additional discourse relationships in translated texts while using a transformation in translation can lead to annotating different discourse relations or it can mean that annotations may not be available because discourse relations may be lost due to grammatical structure change.

The analysis of discourse markers in translation provides some theoretical and practical insights, as well as advantages in comparative language research. For example, a recent study by Hoek et al. (2017) investigate the types of discourse markers that are most often omitted during translation. The authors hypothesize that cognitively simple discourse relations could be denoted by expressed discourse markers, and the latter could be omitted more often than those denoting more complex discourse relations.

The classical method of annotating discourse markers consists of independent annotation of several annotators by assigning a value from a list of discourse relationships to a particular discourse marker. Typically, such annotations are performed by more than one annotator, and in the evaluation phase, the reliability of the annotation is assessed by measuring the overlap of multiple annotations.

Discourse relations can be annotated based on several known discourse models, such as Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) and Segmented Discourse Representation Theory (SDRT) (Asher et al., 2003). However, these two models aim to provide a comprehensive theoretical picture of the discourse relationship, and the Penn Discourse Treebank (PDTB), developed by English scholars, allows for a greater consideration of the meaning of connecting elements. This system is based on a lexical approach to discourse relations or an approach based on the lexical meaning of discourse markers (even implied discourse relations are expressed by a possible conjunction (discourse marker)) and does not assume a global discourse structure, hence the theoretically neutral approach.

One of the most important sources with an annotation of discourse markers is the Penn Discourse Treebank (PDTB). The PDTB provides annotations at the discourse level in the Wall Street Journal Corpus (WSJ). Discourse annotation consists of manually annotated discourse relation values – about 100 types of discourse markers in the text and implicit discourse relations that link discourse arguments. The entire WSJ Corpus, which includes 1,000,000 annotated tags, contains 18,459 annotated discourse markers in the text and 16,053 annotated implicit discourse relations. The values that can be signaled by discourse markers are organized into a hierarchical structure of values consisting of three levels of detail, with four top-level values (temporal, contingency, comparison, and expansion), followed by 16 subtypes at the second level and 23 detailed secondary values at the third level.

According to Bello et al. (2019), discourse markers in the PDTB annotation scheme include several categories of discourse markers. First, explicit discourse markers that belong to well-defined syntactic classes and implicit discourse markers that can be inserted between paragraphs or sentences or compound sentences within internal sentence pairs and that are not directly related to defined syntactic classes and defined sets of discourse markers are discussed. In the case of implied discourse markers, the annotator should attempt to construct a discourse relation between adjacent sentences or parts of the discourse, and the annotation consists of the insertion of a connecting discourse marker that best conveys the implied discourse relation. The connecting discourse markers inserted in this way are called implicit discourse markers. Webber et al. (2008) also discuss multiple discourse relations, where there may be cases where the annotator may see multiple discourse relations and suggest the inclusion of multiple implicit discourse markers. Adjacent pairs of sentences or larger discourse elements between which the annotator cannot see the implied discourse marker are further subdivided as follows: (a) AltLex (so-called alternative lexicalization) that the implied relation of discourse is already expressed in another lexical expression or form, which may be called alternative lexicalization; (b) EntRel (so-called entity relationships), where no inference can be drawn about the existence of a particular discourse relations, but the second sentence or larger element of the discourse is intended only to provide some further description of the first element; and c) NoRel (so-called no relation), when there is no discourse relation between adjacent sentences and it is not even possible to identify the integrity connection, then a lack of connection is concluded.

Since there are no generally accepted abstract semantic categories for the classification of so-called arguments (sentences or parts of discourse), such as expressed discourse markers or, for example, categorization of discourse into agent, recipient, subject, frame, etc., discourse markers combine elements or arguments simply denoted Arg2 and Arg1. Arg2 is an argument that is in a sentence that is syntactically related to the discourse marker, and Arg1 is just another argument. Arg1 and Arg2 are defined as marked text material that is relevant and minimally necessary to explain the relation of discourse. No other supplementary text is marked.

The annotation of the expressed discourse markers and their arguments consists of the selection of the relevant parts of the text in the texts to be worked with and their assignment to Arg1 and Arg2 and the assignment of the discourse relation value to the corresponding discourse marker. Annotation of implicit discourse markers begins by first selecting the first part of Arg2 text for the implied discourse marker, then selecting the text segment Arg1 and finally identifying the value of the discourse relation expressing the Arg1 and Arg2 relationship by providing a word or phrase to express that relationship. In the case of AltLex, instead of presenting a word or phrase to express a discourse relation, a section of text in Arg2 that expresses the discourse relation is selected and marked. In the case of EntRel and NoRel, annotation involves first selecting the text part of Arg2 and then selecting and marking adjacent sentences or parts of the text as Arg1.

Thus, in summary, discourse relationships are denoted by expressed discourse markers, implicit discourse markers, and Altlex, the so-called alternative lexicalization. In the case of EntRel and NoRel, no discourse relations are identified. The values or labels of discourse markers are selected from a grouping of hierarchical discourse relationship values at three levels of hierarchical classification, where discourse markers are divided into classes, types, and subtypes according to the expressed discourse relation, and values from all three levels of the hierarchy are selected during annotation.

As for the multiple expressed discourse markers adjacent in one place, they are all annotated separately. When there are several expressed discourse markers in the same place (e.g., two or more discourse adverbs or a conjunction and a discourse adverb, etc. (e.g., yes, for example; but then; and more; earlier, for example, etc.)), then each discourse marker is denoted separately according to its two arguments. It should be noted, however, that this ignores the real possibility that discourse markers may be dependent, that one discourse marker may be dependent on another, and have different arguments, but the PDTB annotation scheme does not distinguish between dependent and independent discourse markers and their arguments in a sentence. In the case of implicit discourse markers, even if the annotator wants to insert a multiple discourse marker, such a case is annotated with one discourse relationship value from the hierarchy of discourse values. In summary, research shows that the PDTB annotation scheme provides good insights into discourse relations in text and discourse markers identifying these discourse relations.

PDTB annotators are allowed to freely select values from all levels, including the ability to annotate with two value characters (from any level of the hierarchy) to account for ambiguous cases. Thus, in principle, combinations of 129 values are possible. A similar methodology has been implemented to annotate the discourse relations of many other languages, such as Hindi, Czech, Arabic, and Italian (Webber & Joshi, 2012). In addition, Zufferey and Degand (2017) conducted a multilingual annotation experiment with five Indo-European languages belonging to Germanic and Romance language families: English, French, German, Dutch, Italian. In all these studies, it was observed that the cases of discrepancies between the different annotators are similar and the number is not large. These results suggest that the PDTB methodology and results can be replicated and applied to other languages.

Annotation of Lithuanian language discourse relations in TED-MBD corpus

Research methodology

The research first deals with the possibilities of expressing discourse relations by using discourse markers as their linguistic realization in different languages, discussing possible choices of translators, taking into account discourse relations in translation and the use of different linguistic means. The article presents the parallel multilingual corpus TED-MBD (Multilingual discourse-annotated corpus), which is annotated at the discourse level, in accordance with the objectives and principles of PDTB (Penn Discourse Treebank) discourse annotation. The article discusses in detail the annotation system of PDTB discourse markers, the reader is introduced to the hierarchy of senses of discourse relations, the principles of annotation and insights into the application of the PDTB scheme. It also widely describes the Lithuanian part of the corpus and its annotation principles in accordance with the PDTB discourse annotation rules; the first results related to the expression of discourse relations and the use of discourse markers are discussed. The article also presents the first research insights on comparing Lithuanian and English discourse annotated texts in order to understand translation tendencies at the discourse level.

It should be noted that research on discourse relations and discourse markers is an active field of research based on corpus linguistics and computational linguistics tools.

However, corpora in different languages are usually created for different text types or genres, and the data is annotated using different annotation schemes. Thus, although existing corpora can be used for comparative studies at a general level, it is hardly possible to make a detailed comparative analysis of such a variety of data. The TED Multilingual Discourse Textbook (TED-MDB) is a parallel corpus annotated at the discourse level, following the objectives and principles of PDTB discourse annotation (Zeyrek et al., 2018). TED-MDB is developed based on Webber et al. (2016) PDTB discourse relations hierarchy and already includes 7 languages: Turkish, English, Polish, German, Russian, Portuguese, and Lithuanian. Thus, this corpus makes it possible to compare discourse annotated translated texts with the English discourse annotated source text in order to understand translation trends, as well as to analyze different languages of the text. Thus, the TED-MDB analysis offers a better perspective of comparative studies because the same type of texts are used for annotation. Text-type similarity is a major advantage in analyzing language discourse relations and discourse markers; therefore, research on TED-MDB annotated texts should lead to a better understanding of many discourse-related phenomena.

According to the principles of the TED-MDB project, Lithuanian texts were annotated with the main types of discourse relations (expressed (discourse marker in the text, expressed lexically), implicit (not expressed in a text lexically), alternative lexicalization, entity relation, no relation) and their highest level values (temporal, contingency, comparison, and expansion), as well as second- and third-level values in PDTB style.

Research Findings and Discussion

Discourse markers expressed in Lithuanian (lexically include lexical units from four grammatical classes: subordinating conjunctions – e.g., when, until, because, etc.; coordination conjunctions – and, however, etc; sentential relatives – so that, at the time when, etc.; and adverbs – in fact, in the end, etc. The main task of annotation is to find out whether annotated words and phrases act as discourse markers. Like the PDTB, five types of discourse relationships are identified and annotated: expressed discourse relations, implicit discourse relations, alternative lexicalizations, entity relations, and no relations. When denoting discourse arguments, both in the case of expressed discourse markers and alternative lexicalizations, the rule is that the Arg2 label is assigned to an argument that is in a sentence syntactically related to the discourse marker; the next argument is denoted Arg1. As in the PDTB scheme, in the TED-MDB corpus adverbs called "discourse markers" are not annotated because they indicate the organizational structure of the discourse rather than the discourse relationships that link the two arguments semantically, e.g. and it's English equivalent now (see Examples 1 and 2):

1. *Dabar, jie artėja prie 100 procentų tvaraus investavimo, sistemingai integravę ASV visose fondo veiklose.*

2. *Now, they are moving toward 100 percent sustainable investment by systematically integrated ESG across the entire fund.*

According to PDTB annotation guidelines in the case of implicit discourse relation the annotator has to insert a discourse marker which best expresses the implied discourse relation (see example) (Arg1 is given in *italics*, and Arg2 is given in **bold**):

3. *Aplikosauga apima energijos vartojimą, prieigą prie vandens, atliekų tvarkymą ir taršą ir ekonomišką išteklių naudojimą.* [Implicit = Ir] **Socialinė pusė – žmogiškasis kapitalas, įdarbinimo klausimai ir gebėjimas imtis inovacijų, taip pat tiekimo grandinės valdymas ir darbuotojų teisės bei žmogaus teisės** (Implicit) (Expansion: Conjunction)).

Alternative lexicalization (AltLex) involves implicit discourse markers between adjacent sentences where redundancy occurs if an attempt is made to insert an expressed discourse marker. The reason for this redundancy is that the discourse connection is already expressed in some form of alternative lexicalization, rather than the usual expression of the discourse marker (see Example 4):

4. *Daugybė žmonių su amputuotomis galūnėmis mano šalyje nesinaudos savo protezais.* [Priežastis buvo ta], **jų protezų jungtys jiems kėlė skausmą, nes netiko jų forma** (AltLex) (Contingency: Cause: Reason).

Entity relations (EntRel) are annotated between adjacent sentences when the subject or object in one argument is further described in the following argument (see Example 5):

5. *Tad pasakysiu kai ką, kas gali jus nustebinti: galios balansas, galintis išties paveikti tvarumą, yra institucinių investuotojų rankos.* **Tai tokie didieji investuotojai kaip pensijų fondai, kiti fondai ir labdaros fondai** (EntRel).

No relation (NoRel) is identified if the annotator cannot see any discourse relation between the adjacent sentences (see Example 6):

6. *Gavau šią nuotrauką prieš kelias savaites ir joje matote, kas vyksta San Fransisko gatvėse, ir manau, kad tai galima suprasti žiūrint į šiuos žemėlapiu.* **Pažvelkime į Rio de Žaneirą** (NoRel).

TED-MDB adds a new top-level category to the PDTB discourse relation hierarchy, the so-called hypophore. This category is designed to capture rhetorical question-and-answer pairs, where a rhetorical question is asked and the speaker himself answers it. TED-MDB annotates the hypophore as an AltLex case expressed in question. If possible and necessary, another additional question-answer pair discourse may be added. According to the TED-MDB annotation instructions, in Lithuanian, we annotate the question as Arg2, the answer – as Arg1. The question is marked Arg2 because the word expressing AltLex is part of the question. The question word (or a special word, used in Yes / No questions, which can also be used as an expressed discourse marker in Lithuanian (see Example 7) is denoted as AltLex because it expresses the discourse relationship between the question and the answer (see Example 8):

7. *Nieko nepadarysi*, [ar] **tu bandysi kažką keisti**, [ar] **tu nebandysi nieko keisti** (Explicit) (Expansion: Disjunction).

Further examples illustrate how hypofora is annotated in Lithuanian (see Examples 8 and 9):

8. [Ar] **investuotojai, ypač instituciniai investuotojai, į tai įsitraukia**? *Atsakymas yra kai kurie – „taip"* (Explicit) (Altlex: Ar; (Expansion: Level-of-detail: Arg1-as-detail; Hypophora)).

9. [Kodėl] **jie taip padar**ė? – *Atsakymas į šį klausimą gali būti, kad vandens verslas žada didesnį augimą nei elektros įrankiai* (Explicit) (Altlex: Hypofora (Contingency: Cause: Reason; Hypophora).

Examples of the use of annotated corpus for research and teaching

Such corpus can be used for translation research and teaching discourse awareness in translation. At the beginning, it is possible to review and compare the whole set of annotated texts in English and Lithuanian and present the frequencies of annotated discourse relationship types and PDTB top-level discourse relationship values in figures (see Figures 1 and 2).

Figure 1
Frequency of Annotated Discourse Relation Types in English and Lithuanian

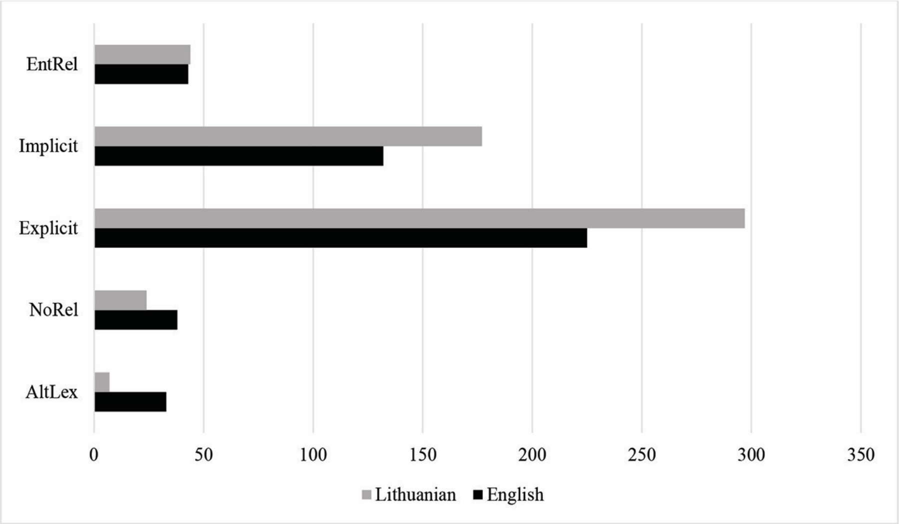
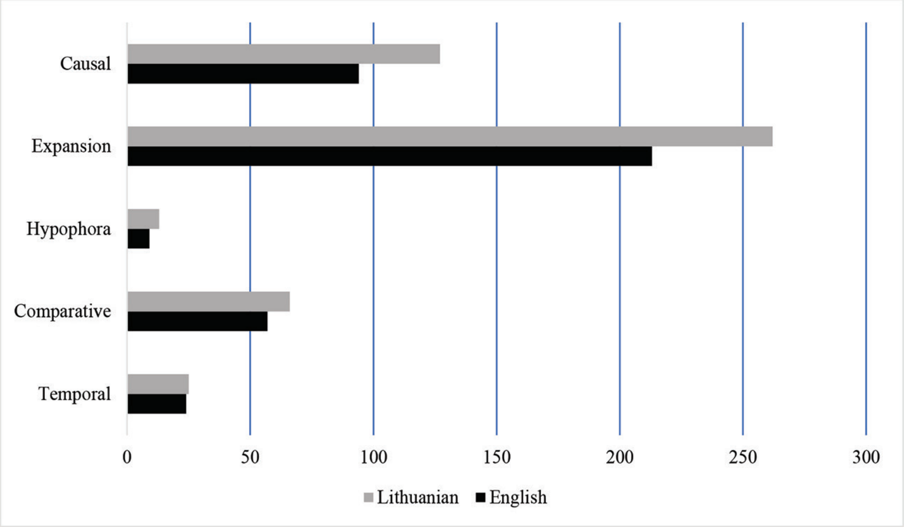


Figure 2

Frequency of Annotated Top-Level Discourse Relations in English and Lithuanian



The small frequency of annotated AltLex in Lithuanian shown in figure 1 could indicate a certain tendency reflecting the choices of translators when translating discourse markers. Translators do not seem to be inclined to use alternative lexicalization and have shown a tendency to convey discourse markers in the versions provided by dictionaries. This resonates with Baker's (2011) observation that translators may choose to combine discourse markers with the nature of the target language (the language being translated).

Another interesting observation is that there are more pronounced discourse markers in Lithuanian than in the English version. This could be explained by the translators' efforts to convey the implicit discourse relations in English with explicit discourse markers in Lithuanian (see Example 10):

10. *Rezultatas geras, tiesa*? [Bet] **Mums reikia geresnio** (Explicit) (Palyginimas: nuolaida: Arg2_kaip_paneigimas) (Comparison: Concession: Arg2_as_denier).

The result is good, right? [(Implicit) = But] **We need a better one** (Implicit) (Comparison: Concession: Arg2_as_denier).

Therefore, it could be hypothesized that perhaps the Lithuanian language may be more prone to the use of expressed discourse markers, and some discourse markers of the original language may also have a similar tendency to become expressed in translation. However, after parallelizing the discourse relations in both languages, it became clear that there are simply more annotated discourse relations in the Lithuanian language, probably due to the peculiarities of the language. And after analyzing the data, it was noticed that the so-called expression of discourse markers in translation is a rather rare phenomenon, occurring in only 5% of all cases when an unexpressed (implicit) discourse marker in English becomes an expressed (explicit) discourse marker in Lithuanian translation. We also find 8% of EntRel cases where English implicit discourse markers in the Lithuanian translation are denoted by the annotator as EntRel instead of "Implicit". Again, the absence of a discourse marker causes some annotation instability when the implied discourse relation in the original language in the translation is annotated with an EntRel connection. There are also 2.9% of NoRel cases: when there is no connection and the annotator could simply not perceive the existence of any discourse connection.

On the other hand, we have noticed cases when the explicit English discourse marker is conveyed by an implicit discourse marker in Lithuanian, and sometimes this leads to the loss of discourse meaning present in the original English text but not in the Lithuanian translation (see Example 11):

11. *Atsižvelgimas tik į rasinius skirtumus nepadeda bandant vystyti visuomenės įvairovę ir tolerenciją*. [Implicit = Taigi] **Bandydami įvairovę naudoti sudėtingesnių problemų sprendimui, turime imti kitaip interpretuoti įvairovę ir sieti ją su tolerancija** (Numanomas) (Implicit) (Priežastinis: priežastis: rezultatas (Contingency: Cause: Result)).

Simply taking race into account doesn't really help to develop the diversity of the society. So if **we would like to try using diversity as a possible way to solve some of the complicated problems of our society, we need to start interpreting diversity in a completely new way by relating it to tolerance** (So (Explicit) (Contingency: Cause: Result) if (Explicit) (Contingency: Condition: Arg2_as_condition).

Thus, example 11 shows that the translator chose not to convey the English explicit discourse markers *so* and *if*, and although the "result" discourse relation remains implicit, a loss of meaning of the "condition" discourse relation can be observed as the "condition" discourse relation is completely lost.

Therefore, we decided to analyze the parallel discourse relations in both languages. The following discusses the cases of explicit discourse relations denoted by expressed discourse markers in English, which have been translated as implicit discourse relations with implied discourse markers. This means that the discourse marker is omitted in the translation, but the annotator nonetheless determines the type of implicit discourse relation and suggests a possible discourse marker in that context. We reviewed our data with some hypothesis in mind that perhaps there is a tendency in translation to use more implicit discourse relations and discourse markers, turning explicit discourse relations in the original into discourse markers denoting them into implicit in translation. It is also possible that some discourse relations will show a greater tendency to become implicit in translation. However, the results of the analysis of parallelized explicit discourse relations marked with expressed discourse markers show a weak tendency of implication of the analyzed discourse relations in Lithuanian translation. It was found that in 80% of cases the English explicit discourse markers were translated into the explicit discourse markers into Lithuanian, and the remaining 20% were converted into implicit discourse relations. An interesting observation is that out of all implicit cases not expressed by discourse markers, half (50%) of the cases have resulted from the English discourse marker and (and); however, when annotating an implied discourse relations in translation, the annotator selects the appropriate implied discourse marker and retains the same meaning (Implicit) (Expansion: Conjunction) as implicit (see Example 12).

12. *These photographs are in color* [and] **they portray a community swirling across the country, fiercely alive and creatively free, seeing sides of America that no one else gets to see** (Explicit) (Expansion: Conjunction)

Šios nuotraukos spalvotos, [Implicit = Ir] **jos parodo per visą šalį klajojančią bendruomenę** (Implicit) (Expansion: Conjunction).

Another interesting observation is that when there are several discourse markers in one place in the original text, one explicit discourse marker is left in the translation and the other becomes implicit (see Example 13).

13. *I had a deep feeling of restlessness or an essential fear that my life might get into a course of routine and boredom* [And] [so] **many of my early childhood and youth memories are related to my dreams of walking across borders, wandering in nature, and meeting all kinds of unconventional people living their lives on the road** (Explicit) (Expansion: Conjunction).

Negalėjau nusėdėti vietoje, labai bijojau, kad rutina ir nuobodulys praris mano gyvenimą. [No implicit discourse] [Todėl] **dauguma mano vaikystės prisiminimų yra susiję su vaizdiniais apie kitų šali**ų sienų kirtimą, klajones gamtoje, susitikimus su nesuvaržytais žmonėmis, gyvenančiais kelyje (Implicit, no annotation).

These examples are consistent with the hypothesis that some discourse relations and the discourse markers that denote them tend to become implicit in translation. However, further research on the implication of certain discourse markers is needed to finally substantiate this.

It is also interesting to review the changes in discourse relations in translation and to observe certain trends. Analyzing the shift of discourse relations, we first pay attention to the same type of discourse relations in English and Lithuanian, and only then we examine different discourse relations in a language pair.

a) We first review the explicit discourse markers, which are also expressed in the translation.It should be noted that when translating explicit discourse markers into the same expressed discourse markers, there is little change in meaning in discourse relations - 18%, which occurs at the first level of discourse relation, because the translator chooses a different discourse marker that gives a change in value for discourse relation (see Example 14).

14. *My dreams became reality* [through] [per] **my profession which is a documentary photographer** (Explicit) (Expansion: Manner: Arg2_as_manner).

Mano svajonės tapo realybe, [kai] **aš įgijau dokumentikos fotografės profesiją** (Explicit) (Temporal: Synchronous).

Thus, we see that in Example 14, due to the discourse marker in Lithuanian chosen by the translator, the meaning changes from the meaning of manner discourse relation in English to the meaning of time in Lithuanian. Thus, after analyzing the data, we see that changes in the meaning of discourse relations are due to two reasons: the translator's choice of discourse marker determines the change in the meaning of first-level discourse relations, also differences in the choice of discourse relation meanings occur in the work of annotators, especially at lower levels of the hierarchy of discourse relation meanings. Cases of a shift in the meanings of second-level discourse relations include such changes as: Extension: Merger shift to Extension: Detailing; Causality: Cause + Belief: Result + Confidence shift to Causality: Cause: Result; Extension: Merge Shift to Comparison: Discount or Comparison: Contrast; Causality: Cause + Belief: Result + Belief Shift to Causality: Cause + Language Act: Result + Language Act.

15. *And then the Flamengo football team is also represented here.* [So] **you have that same kind of spread of sports and civics and the arts and music, but it's represented in a very different way, and I think that maybe fits with our understanding of Rio as being a very multicultural, musically diverse city** (Explicit) (Contingency: Cause + Belief: Result + Belief).

Flamingo futbolo komanda taip pat čia. [Taigi] **turime tą patį pasiskirstymą tarp sporto, pilietinių teisių, menų ir muzikos, bet tai pavaizduota visai kitaip, ir manau, kad tai gerai atitinka mūsų Rio de Žaneiro suvokimą, kad tai labai daugiakultūrinis ir muzikos prasme įvairus miestas** (Explicit) (Contingency: Cause + Speech Act: Result + Speech Act).

Example 15 shows that in the English text, the discourse relation associated with the discourse marker *so* is annotated as Contingency: Cause + Belief: Result + Belief; however, in Lithuanian this discourse relation together with the discourse marker is thus named not as a belief, but rather as a speech act, and is annotated accordingly. Therefore, we can observe that at lower levels of the hierarchy of discourse relation values, the differences in annotation in Lithuanian are more pronounced when annotating the explicit discourse markers.

b) Another case that also needs to be analyzed is the implicit discourse relations and discourse markers, which remain implicit in the translation as well.

For implicit discourse markers, changes in the meaning of discourse relations are observed in only 12% of cases, which are shifts in the meaning of the hierarchy of discourse relations on the first level or shifts in the meaning on the second level of the hierarchy of discourse relations. Shifts in the meaning on the first level of the hierarchy of discourse relations are determined by how the annotator chooses to insert a different discourse marker that allows different interpretations of the relations of the combined discourse arguments. Particularly, noticeable first-level change includes: Extension: Conjunction shift to Causality: Cause and vice versa in both Lithuanian and English. The shifts of discourse relations in the Lithuanian translation at the level of the second level of hierarchy of discourse relations also occur due to the choice of annotator, because there is no discourse marker as a connecting element that would provide the annotator with information on interpretation of discourse relations between sentences.

16. *There are other ways of thinking about maps of the cities and how they could be made* [Implicit = and] **Today, I would like to demonstrate new types of maps and mapping** (Implicit) (Expansion: Conjunction).

Galvodami apie miestų žemėlapius, dažniausiai turime mintyse gatves ir pastatus, žmonių gyvenamųjų vietų plėtrą, ir kaip miestas kūrėsi, arba galime galvoti apie įdomius urbanistinius sprendimus ir vizijas, bet yra ir kiti būdai apmąstyti miestų žemėlapius ir kaip jie galėjo būti sukurti. [Implicit = taigi] Šiandien aš norėčiau jums parodyti naujo tipo žemėlapį ir jo sudarymą (Implicit) (Contingency: Cause + Speech Act: Result + Speech Act)).

Example 16 shows that in the English text the annotator chose the implicit discourse marker as a connecting element *and*, thus, annotating the implicit discourse relation as Extension: Conjunction, while in the Lithuanian text the annotator chose the implicit discourse marker Thus, annotating it as: Contingency: Cause + Speech Act: Result + Speech Act.

Analyzing the cases where the explicit discourse relations in the translation are conveyed as implicit discourse relations, we observe similar tendencies mentioned in the analysis of the cases of annotation of implicit discourse relations.

c) When the explicit discourse relation in English is conveyed as implicit in Lithuanian, in 11% of cases the meaning of the discourse connection changes. This percentage is similar to cases where implicit discourse relations remain the same in both languages. Changes affect the first and second levels of the discourse relation hierarchy. Indeed, the same top-level change found in discourse relation contexts is related to Extension: Conjunction shifting to Contingency, while the lower levels of the discourse relation hierarchy differ according to the discourse marker selected by the annotator and the detail of the annotation level. Although the explicit discourse marker is in most cases the conjunction *and*, in Lithuanian the corresponding implicit discourse marker proposed by the annotator shows more lexical diversity.

d) Finally, we review the implicit discourse markers in English, which are conveyed as explicit discourse markers in the Lithuanian translation.

As mentioned earlier, there are quite a few such cases and the translator selects the expressed discourse marker quite well, so the meanings of the discourse relations chosen by the annotators are the same in both languages.

At the end, we must mention the grammatical transformation – the transfer in translation, which is quite rare in the texts we are considering. It can be noticed that transformations are rarely used in the translations into Lithuanian. In most cases, translated texts follow the structure of the original English version, except in a few cases where verbs are translated into noun forms and therefore simply become homogeneous parts of the sentence and do not require annotation, as they no longer express any discourse events (see Example 17).

16. ... maim (and) kill (Explicit) Expansion: Conjunction.

... mutilation (noun) and massacre (noun) on freight trains (No annotation).

In the Lithuanian translation, this part of the sentence does not express any discourse relation and is not annotated.

In this article, we have presented only a few examples that reveal the possibilities of translation research and raising translator discourse awareness in teaching while using a parallel discourse annotated corpus. Also, as mentioned at the beginning, such corpus allows to study the peculiarities of the expression of Lithuanian discourse markers and is recommended to be used in translation studies. Case studies of this type may serve as an effective method to develop students' linguistic awareness as well as analytic abilities and are especially relevant for students majoring in English philology or translation.

Conclusions

The development and research of discourse annotated corpora is a relatively new field, therefore Lithuanian researchers seek to supplement the existing corpora resources and look for ways to study discourse relations by linking and comparing them with their counterparts in other languages because in different languages discourse relations are realized by different linguistic means. The article discusses discourse research and its relation to translation studies as raising text coherence awareness in translation is of a key importance, and also to introduce the developed corpora resources.

TED-MDB parallel multilingual discourse annotated corpus allows to study Lithuanian discourse markers, as well as to compare them with discourse markers of other corpus languages. The analyzed examples of Lithuanian and English annotations presented in the article show that Lithuanian discourse markers sometimes become explicit instead of English implicit ones, which could be explained by the translators' efforts to translate the implicit discourse relations into the relations with the explicit discourse markers. In addition, it can be observed that the rendering of explicit discourse markers in to the implicit ones can indirectly impair the rendering of the meaning of discourse relations in the text. However, further research and insights into such translator choices would be needed as a follow-up. It should be borne in mind that certain stylistic provisions may be the choice of individual translators, for example, some translators may wish to use more explicit discourse markers, others less so further research is needed to establish the prevailing tendencies.

Another observation relates to the change in the meaning of the discourse relation, which depends on two reasons: the choices of translators and the choices of annotators. When a translator chooses a different discourse marker or a discourse marker with multiple meanings, then the annotator, relying on the discourse marker, naturally marks the different value of the discourse relation. When the translator selects an equivalent discourse marker as a connecting element, the annotator typically marks the same value of the discourse relation. Implicit discourse markers cause noticeable changes in discourse relation values in the annotation, including some shifts in the meaning on the first level of the hierarchy of discourse relations in Lithuanian and some cases of shifts in the meaning on the second level of the hierarchy of discourse relations in Lithuanian.

In the future, it is possible to delve into the comparative research of Lithuanian texts and other languages of TED-MDB multilingual corpus, it is expected to reveal and refine more translation trends. Finally, it is vital to teaching philology students and senior students about discourse markers, and TED-MBD is an excellent medium to study text coherence.

References

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

Baker, M. (2011). *In other words: A coursebook on translation* (2nd ed.). London & New York: Routletge.

Bello, I., Bernales, C., Calvi, M., & Landone, E. (2019). *Cognitive insights into discourse markers and second language acquisition.* Peter Lang Ltd., International Academic Publishers.

Crible, L., & Degand, L. (2019). Domains and functions: A two-dimensional account for discourse markers, *Discours, 24,* 3–35. https://journals.openedition.org/discours/9997

Boulton, A., &Tyne, H. (2014). *Corpus-based study of language and teacher education*. New York: Routledge.

Furko, P. B. (2020). *Discourse markers and beyond: Descriptive and critical perspectives on discourse- pragmatic devices across genres and languages.* London: Palgrave Macmillan.

Hoek, J., Zufferey, S., Evers-Vermeul, J., & Sanders, T. (2017). Cognitive complexity and the linguistic marking of coherence relations: a parallel corpus study, *Journal of Pragmatics*, *121,* 113–131.

Hunston, S., (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Kubler, N., & P. Y. Foucou. (2003). Teaching English verbs with bilingual corpora: Examples in the computer science area. *Contrastive Linguistics and Translation Studies*, 1–15. Amsterdam: Rodopi.

Roze, Ch., Danlos, L., & Muller, Ph. (2010). LEXCONN: a french lexicon of discourse connectives, *Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010), Moissac, France.* https://journals.openedition.org/discours/8645

Sanders, T., J., M., Noordman, L., & G., M. (2000). The role of coherence relations and their linguistic markers in text processing, *Discourse Processes*, *29*, 37–60.

Šolienė, *A.* (2018). Diskurso žymikliai: ar viskas išverčiama? *Gimtoji kalba*, *11,* 7–10.

Webber, B., & Joshi, A. (2012). Discourse structure and computation: past, present and future, *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, Association for Computational Linguistics*, 42–54.

Webber, B., Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., & Joshi, A. (2008). The penn discourse TreeBank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation. European Language Resources Association, Marrakech*, 2961–2968.

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2016). A discourse-annotated corpus of conjoined VPs. *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016). Association for Computational Linguistics*, 22–31.

Zeyrek, D., Mendes, A., & Kurfalı, M. (2018). *Multilingual extension of PDTB-style annotation: The case of TED Multilingual Discourse Bank, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association* (ELRA).

Zufferey, S., & Degand, L. (2017). Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory, 13*(2), 399–422.

Lietuvių kalbos diskurso žymikliai daugiakalbiame tekstyne vertimo mokymui

Giedrė Valūnaitė Oleškevičienė[1], Vitalija Karaciejūtė[2], Dalia Gulbinskienė[3], Nagaletchimee Annamalai[4]

[1] Mykolo Romerio universitetas, Ateities g. 20, 08303, Vilnius, Lietuva, gvalunaite@mruni.eu

[2] Mykolo Romerio universitetas, Ateities g. 20, LT-08303, Vilnius, Lietuva, vitalija.karaciejute@mruni.eu

[3] Vilniaus Gedimino technikos universitetas, Saulėtekio al. 11, LT-10223, Vilnius, Lietuva, dalia.gulbinskiene@vilniustech.lt

[4] Malaizijos mokslų universitetas, Penangas, Malaizija, naga@usm.my

Santrauka

Diskurso ryšiais anotuotų tekstynų kūrimas ir tyrimai yra pakankamai nauja sritis, todėl Lietuvos mokslininkai siekia papildyti esamus tekstynų resursus ir ieško būdų, kaip būtų galima tyrinėti diskurso ryšius siejant ir lyginant juos su kitomis kalbomis. Straipsnyje pristatomo tyrimo tikslas yra aptarti diskurso tyrimus pabrėžiant diskurso suvokimo svarbą vertimo studijose, taip pat pristatyti tekstynų išteklius ir paskatinti diskurso jungtukų tyrimus lietuvių kalboje, remiantis užsienio mokslininkų patirtimi. Pirmiausia aptariamos diskurso ryšių ir juos išreiškiančių diskurso žymiklių raiškos galimybės skirtingose kalbose, atskleidžiant galimus vertėjų pasirinkimus, atsižvegiant į diskurso ryšius vertime ir skirtingų kalbinių priemonių vartojimą. Straipsnyje pristatomas lygiagretusis daugiakalbis tekstynas TED-MBD (angl. *Multilingual discourse-annotated corpus*), kuris yra anotuotas diskurso lygmeniu, laikantis PDTB (angl. *Penn Discourse Treebank*) diskurso anotavimo tikslų ir principų. Straipsnyje aptariama PDTB diskurso žymiklių anotavimo sistema, supažindinama su diskurso ryšių reikšmių hierarchija, anotavimo principais ir PDTB schemos taikymo įžvalgomis. Lietuvių kalbos diskurso žymeklių anotacija yra naudinga ne tik vertėjams, bet ir užsienio kalbų mokytojams, nes leidžia geriau susipažinti su teksto rišlumu bei pragmatiniu diskurso aspektu.

Esminiai žodžiai: *diskurso žymikliai, diskurso ryšiai, lygiagretusis tekstynas, anotavimas, PDTB diskurso ryšių hierarchija, vertimas.*

Gauta 2021 08 09 / Received 09 08 2021
Priimta 2022 03 15 / Accepted 15 03 2022