Fuzzy Boundaries in the Different Functions and Translations of the Discourse Marker *and* (Annotated in TED Talks)

Giedrė Valūnaitė Oleškevičienė, Mykolas Romeris University, Lithuania, gvalunaite@mruni.eu

Sigita Rackevičienė, Mykolas Romeris University, Lithuania, sigita.rackeviciene@mruni.eu

Agnes Abuczki, Károli Gáspár University, Hungary, abuczki.agnes@gmail.com

Nijolė Burkšaitienė, Mykolas Romeris University, Lithuania, n.burksaitiene@mruni.eu

Ludivine Crible, Université Catholique de Louvain, Belgium, ludivine.crible@uclouvain.be

Peter Furko, Károli Gáspár University, Hungary, furko.peter@gmail.com

Viktorija Mažeikienė, Mykolas Romeris University, Lithuania, vmazeikiene@mruni.eu

Liudmila Mockienė, Mykolas Romeris University, Lithuania, liudmila@mruni.eu

Jolita Šliogerienė, Vilnius Gediminas Technical University, Lithuania, j.sliogeriene@gmail.com

Šárka Zikánová, Charles University, Czech Republic, zikanova@ufal.mff.cuni.cz

Abstract

The present research focuses on the multiple functions performed by the discourse marker *and* in annotated spoken-like texts of TED Talks in English and Lithuanian. The annotation of TED Talks in Lithuanian has started only recently, which results in the limitation regarding the quantity of annotated texts. The research findings show that *and* and its Lithuanian counterparts perform multiple fuzzy functions, including the function of addition, discourse management and structuring discourse.

It was also established that the most frequent variants of translation of the discourse marker *and* are those provided by bilingual English–Lithuanian dictionaries and that translators choose paraphrases to convey the pragmatics of the spoken-like texts.

The research has been conducted within the framework of TextLink COST Action IS1312.

Key words: annotated discourse, TED Talks, translation, English, Lithuanian, discourse marker *and*

Introduction

The development of corpora and corpus-based research is focused on various linguistic phenomena, including pragmatics and textual features. Discourse annotated corpora reveal qualitative differences of the use of discourse markers in different languages. The difficulty of conducting cross-linguistic comparisons of discourse markers is determined by their polysemy and the ways of expressing coherence relations used in different languages. The former, i.e. polysemy of discourse markers means that a single lexical item can be used to convey several coherence relations, whereas the latter means that when coherence relations are not expressed by any discourse marker, the relations are implicit and have to be reconstructed by inference (Zufferey and Degand, 2013). The use of discourse markers is challenging in translations from a source language to a target language as translators who have to adapt them to a new language and culture, in which textual strategies involving their use are often different from those of the source text (Baker, 1993; Mason, 1998; Halverson, 2004, cited in Zufferey and Degand, 2013). Hence, discourse marker analysis is relevant in the field of discourse analysis, which becomes pivotal in case of cross-cultural communication and translation.

The object of the present research is the English discourse marker *and* and its Lithuanian counterpart *ir* and it is aimed at revealing pragmatic use of these discourse markers. To achieve this aim, the following two objectives have been set: to compare discourse marker *and* with its counterparts in Lithuanian by applying Crible and Degand's (2017) functional taxonomy of domains and functions of discourse markers and to analyse the translations of *and* into Lithuanian by examining English and Lithuanian transcripts of TED Talks. The research demonstrates the complexity of the connective pragmatics and peculiarities of translation.

Theoretical background

The development of large language databases known as corpora revealed the potential of language research using corpus techniques focused on researching various patterns of lexis, grammar, semantics, pragmatics, and textual features. Many corpora are coded according to word classes, analysed for grammatical structure or examined with a focus on pragmatic features. Corpora development has enriched our knowledge of lexis, grammar, semantics, pragmatics, and textual features (Sinclair, 1991; Stubbs, 2004). Corpus linguistics is based on the premise that language varies according to the context related to space and time, which sustain the infinite potential for establishing new facts about it. As dictionaries and grammars do not have the capacity to fully describe language, corpus-based approach is beneficial as it provides real data of real language used in certain contexts (Aston 2001).

Discourse markers have been analysed by a number of researchers using different perspectives, therefore, their definitions vary. To illustrate, Schiffrin (2006) defines discourse markers in two ways, i.e., in her operational definition they are referred to as "<…> sequentially dependent elements that bracket units of talk, i.e., non-obligatory utterance-initial items that function in relation to ongoing talk and text" (p. 321). In her theoretical definition, the researcher specified the conditions that allow a word to be used as a discourse marker, i.e., such a word is syntactically detachable, is used in the initial position within an utterance, has a range of prosodic contours, and operates on different levels of discourse. Schiffrin (2006) also stresses that discourse markers have primary domains within which they function as well as that they can connect utterances either within a single domain or across different domains, which helps to create coherence (p. 322).

Crible (2014), on the other hand, defines discourse markers as "grammatically heterogeneous, multifunctional type of pragmatic markers" that signal "a discourse relation between the host unit and its context <…>, expliciting the structural sequencing of discourse segments, expressing the speaker's meta-comment on his phrasing, or contributing to interpersonal collaboration" (Crible, 2014, pp. 3–4). In her work, the author distinguishes two main subcategories of discourse markers, mainly *relational discourse markers* (RDMs) and *non-relational discourse markers* (NRDMs) (Crible, 2014, p. 10). Besides, the researcher describes a group of discourse markers that "belong somewhere between RDM-NRDM extremes and are thus difficult to situate" and ascribes them to a separate category of discourse markers which perform both relational and non-relational functions (Crible, 2014, p. 16).

Thus, *relational discourse markers* signal "a two-place relation" on the content level (e.g. a cause between two events), on the textual level (e.g. a thematic shift) or on the meta-discursive level (a reformulation of a previous statement) (p. 15). *Non-relational discourse markers*, on the other hand, comprise miscellaneous lexical items performing various functions, e.g. interactive verbal expressions (e.g. *you know*), interjectional punctuators (e.g. *well*), and other meta-discursive elements (e.g. *actually*). In other words, relational discourse devices connect two explicit textual units while non-relational markers signal relations between assumptions (Crible, 2014, p. 15). Discourse markers of the "in-between" category perform both types of functions, whereas the hyperonym "discourse marker" highlights the similarity of the functions of all the three hyponyms (Crible, 2014, p. 16).

For the present research, translation spotting is important. It is a technique of disambiguation of ambiguous discourse markers in one language using parallel data from another language. It is also defined as "an annotation method that makes use of translation of specific lexical items in order to disambiguate them" (Cartoni et al, 2013, p. 68). The theoretical rationale behind this method is that differences in translation can disclose semantic features of a source language (Noël, 2003; Cartoni et al, 2013) and help to identify semantic features of the discourse markers denoting coherence relations since translation relies on the decisions made by the translators, who are experts in their own language (Behrens and Fabricius-Hansen, 2003).

The term 'translation spotting' was originally coined by Veronis and Langlais (2000) to refer to the automatic extraction of translation equivalents in a parallel corpus (Veronis and Langlais, 2000, cited by Cartoni et al, 2013, p. 69). Danlos and Roze (2011), on the other hand, recommend conducting translation spotting manually as there exist several possible translation variants, ranging from various paraphrases and syntactic constructions to no translation or omission. In their research, Cartoni et al (2013) also performed translation spotting manually, which provided reliable results that disclosed both a number of advantages over the classical sense annotation and a number of limitations. The advantages included a low level of annotator disagreements and absence of labels of senses that are set a priori, relying on decisions made by the translator who is an expert in his/her language and whose translation choices are made based on the knowledge of the whole text. The most important limitations were related to the issue of disambiguation. That is, this annotation method provided a direct disambiguation for an item only when the language of translation was less ambiguous than the source language, and that only one translation variant was possible for each meaning of the source language. To resolve this limitation, the authors proposed an additional step of analysis, i.e. to conduct interchangeability tests. It was concluded that the tests allowed distinguishing equivalent translations reflecting the same meaning in the source language from translations that were not equivalent (or interchangeable) and reflected different meanings of the discourse maker in the source language in a more reliable way than traditional sense annotation. The authors also suggested that if extended to a larger number of languages and discourse markers in more diverse genres, the technique would allow making more empirically grounded generalisations regarding discourse relations in the world's languages (Cartoni et al, 2013, p. 83).

Research methodology

In the present study, the methodological decisions regarding the choice of the corpus and the annotation method were determined by the aim of the research. That is, to annotate the English spoken discourse marker *and*, to compare its meanings with their counterparts in Lithuanian as well as to analyse the translations of *and* into Lithuanian, the multilingual corpus of TED Talks was chosen. This choice was made on the premise that parallel corpora are considered to be ideal for optimal comparability between languages as they provide more flexible and accurate ways to compare discourse markers (Zufferey and Degand, 2013).

The choice of the functional approach to be used for this investigation was predetermined by the specific nature of discourse markers, which covers some specific features, e.g. even though most languages possess discourse markers, they have a high degree of contextual variation (Crible and Degand, 2017). The general approach proposed by Crible (2017) describes discourse markers as functioning in four "domains", including the ideational domain which is 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 revised functional taxonomy (Crible and Degand 2017), 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 predetermining the choice of this particular annotation scheme used for the present research.

Concerning the approach used to analyse translation, theoretical insights provided by Noël (2003), Behrens & Fabricius-Hansen (2003) and Danlos and Roze (2011) were important. Noël (2003) stated that translation spotting can reveal differences in translation that could be used to disclose semantic features of the source language or translation could be used to elicit some semantic features of content words in the source language. On the other hand, Behrens & Fabricius-Hansen (2003) observed 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. Finally, translation spotting reveals the existing discrepancies between the languages, especially in the case discourse markers, when there are no one-to-one translation equivalents and where exist a number of possible translations (Danlos and Roze, 2011).

Research findings

The research was conducted in three stages. First, sentences in English were extracted from TED Multilingual Discourse Bank (TED-MDB) of TED talks, then English sentences and their counterparts in Lithuanian were annotated by applying Crible's revised functional taxonomy of domains and functions of discourse markers (Crible and Degand 2017). Finally, the extracted cases of *and* were analysed. All the annotated values of the discourse marker *and* are presented in Figure 1).

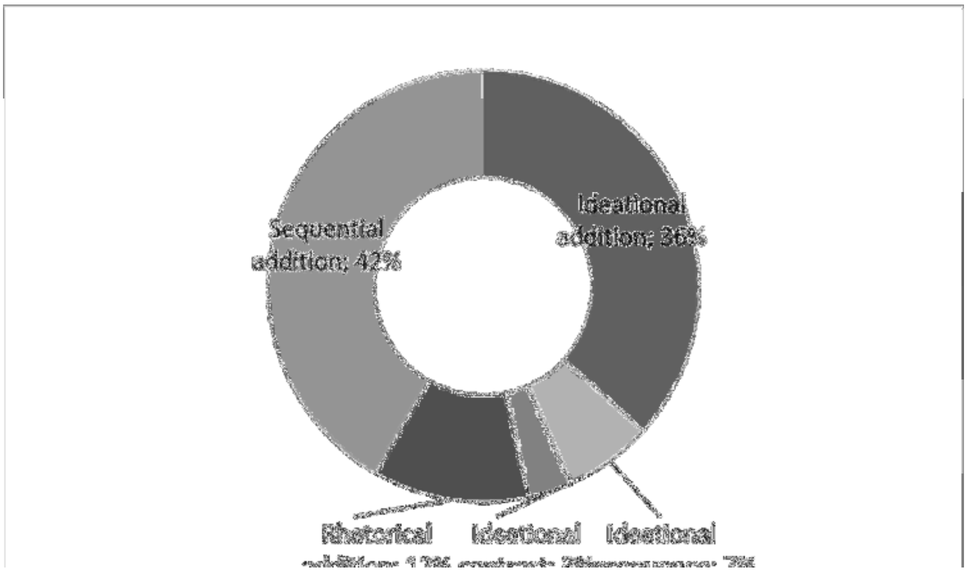

Figure 1. Annotated values of the discourse marker *and*

The research results show that the discourse marker *and* and its Lithuanian counterpart *ir* are used approximately equally both in the ideational domain (36%), representing factual information, and in the sequential domain (42%), representing cases of structuring local and global units of discourse. Besides, 12% of occurrences are related to the rhetorical domain representing the speaker's subjectivity. The research results also reveal that *and* and its Lithuanian counterpart *ir* are used in the function of addition: 36% of the occurrences in the annotated sample express ideational addition in English and 20% in Lithuanian. Besides, in the Lithuanian sample, there are some cases of omissions and cases where *ir* is not functioning as a discourse marker. It should be stressed that the ideational domain is related to real-world events, thus ideational addition expresses an additive meaning based on real world facts, for example:

(1) Now, because they're mathematicians, they have been collecting data on everybody who uses their site for almost a decade, [and] they've been trying to search for patterns in the way that we talk about ourselves and the way that we interact with each other on an online dating website.

(1) Kadangi jie matematikai, jie beveik dešimtmetį rinko duomenis apie žmones, kurie naudojosi jų portalu, [ir] jie bandė surasti dėsningumus mūsų bendravime, kai kalbame apie save ir kaip bendraujame tarpusavyje būtent pažinčių portaluose.

The prevalence of sequential domain, which was established in 42% of all occurrences, reveals that *and* and its Lithuanian counterpart *ir* are often used for the purpose of structuring discourse, i.e. for joining smaller discourse units into bigger ones. For example:

(2) So these equations, they predict how the wife or husband is going to respond in their next turn of the conversation, how positive or negative they're going to be. [And] these equations, they depend on the mood of the person when they're on their own, the mood of the person when they're with their partner, but most importantly, they depend on how much the husband and wife influence one another.

(2) Šios formulės nuspėja, kaip vyras arba žmona reaguos, kai ateis eilė jiems šnekėti – kaip pozityviai arba negatyviai jie bendraus. [Ir] šios formulės priklauso nuo žmogaus nuotaikos, kai yra tiesiog su savimi, ir žmogaus nuotaikos, kai jie su savo partneriu, bet svarbiausia, jos priklauso nuo kaip stipriai vyras arba žmona daro vienas kitam įtaką.

The findings also illustrate that 12% of occurrences in the sample denote addition used in the rhetorical domain, which means that discourse markers in these cases are used to express the speaker's subjectivity and other meta-discursive effects, i.e. that rhetorical addition refers to the speaker's subjective perception and produces the effect of subjective discourse management. For example:

(3) There'd be a huge spread in her scores. [And] [actually] it's this spread that counts.

(3) 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. Even though sometimes it is difficult to isolate a discourse marker from the whole context of the argument, example (3) provided above reveals that the adverb phrase *actually* provides a clear association to the subjective perception.

Analysis of the translation of the discourse marker *and* revealed some possible translations. All translation values of the discourse marker *and* are presented in Figure 2.

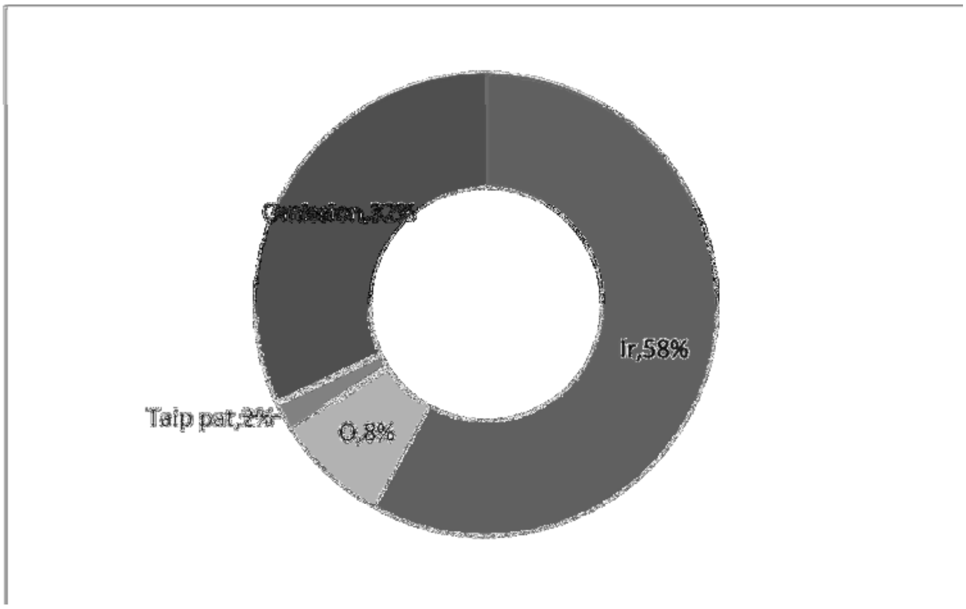

Figure 2. Translation values of the discourse marker *and*

It has been established that the most frequent variants of translation of the discourse marker *and* into Lithuanian is *ir* which is the variant provided by the bilingual English – Lithuanian dictionaries. Also, the translation variant *o* provided by the bilingual dictionaries is present among the identified values. Interestingly, the discourse marker *o* in Lithuanian has the meaning of contrast. The examples of the dictionary-based translations are provided below:

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

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

(5) 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?

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

Example (5) provided above demonstrates how the contrastive meaning of the English discourse marker *but* used at the beginning of the first sentence influences the translation of the following discourse marker *and* which is rendered as a contrastive *o* in Lithuanian.

The Lithuanian adverb *taip pat* was also spotted as a translation value in the sample. In the example presented below, the translator renders the ideational value of the discourse marker *and* in the source language using the adverb *taip pat* in the target language, which is 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.

(6) [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.

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

Omission cases seem to be twofold. There are cases in which more than one discourse marker is used to introduce an argument, which is particularly characteristic of spoken-like speech where discourse markers are used abundantly. In such cases, only one discourse marker is rendered into the target language, which could be the translator's choice predetermined by the requirements of synchronising subtitles and making them concise. Example (7) presented below demonstrates the translator's choice to render the temporal discourse marker *while* into a concessive *nors* and to omit the sequential addition *and*, which is a successful choice having in mind the requirements of the synchronisation.

(7) 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.

(7) 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.

Other cases of omission seem to occur when the translator rendered the meanings of the source language by making a shift in grammatical structures that in their own right required different translator choices in rendering the discourse markers. In example (8), it can be observed that the translator successfully chose to change the whole argument to render the meaning of the source argument and, in doing so, omitted the discourse marker *and*.

(8) 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.

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

There were some interesting cases in which omission of the discourse marker in the translated Lithuanian texts was used more often and these cases denoted sequential addition. The phenomenon might be explained by the requirements of synchronising the subtitles and pursuing the goal of creating subtitles that are easily read, well-rounded bits of text. This means that the translator chose to omit discourse structuring in order to follow synchronisation requirements and relied on the contextual meaning. Example (9) demonstrates how difficult it is for a translator to make a decision by observing synchronisation rules. In the Lithuanian translation the discourse structuring maker *ir* is omitted relying on the contextual meaning; however, it sounds strange to a Lithuanian reader.

(9) 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.

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

The final observation regarding the research findings is that most translator choices are really successful in conveying both the semantic and pragmatic values of the discourse markers and also produces an interesting view of the existing linguistic spaces between the languages as all the discussed features become important in translation research.

Conclusion

The findings of the present research reveal that the discourse marker *and* and its Lithuanian counterparts have the additive meaning as 36% of the occurrences in the annotated sample express ideational addition. On the other hand, the prevalence of the sequential domain established in 42% of occurrences illustrates that *and* and its Lithuanian counterpart *ir* are often used for discourse structuring purposes with the purpose to create bigger discourse units. The present research also shows that 12% 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 multiple functions of certain discourse markers, especially while raising translator awareness in choosing certain translation options.

The most frequent variant of translation (66% of the values in the sample) of the discourse marker *and* in the sample is *ir* and *o,* both of which are the variants provided by the bilingual English–Lithuanian dictionaries. Besides, the omission technique is used abundantly, which may be predetermined by the requirements of synchronising the subtitles and pursuing the goal of creating subtitles that are easily read, concise bits of text. Such features also are relevant, especially for translation studies.

Acknowledgments

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

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 Jaszczolt K, Turner Jen (editors), *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 Applications of Corpus Analysis, in D. Biber-Reppen (ed.), The Cambridge Handbook of English Corpus Linguistics. Cambridge, Cambridge University press, 2015,478-497.

Crible, L. (2014). *Identifying and describing discourse markers in spoken corpora*. Université Catholique de Louvain, Louvain-la-Neuve.

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.

Crible, L., & Degand, L. (2017). Reliability vs. Granularity in discourse annotation: What is the trade-off? *Corpus Linguistics and Linguistic Theory* 14(2): 1–29.

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.

Hansen, M.-B. M. (2006). *A dynamic polysemy approach to the lexical semantics of discourse markers (with exemplary analysis of French toujurs).* In K. Fischer (Ed.): *Approaches to discourse particles* (pp. 21–41). Amsterdam: Elsevier, How to Compress Subtitles? Available at http://translations.ted.org/wiki/How_to_Compress_Subtitles [accessed 3 December, 2017].

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.

Schiffrin, D. (2006). Discourse maker research and theory: revisiting and. In K. Fisher (Ed.), *Approaches to Discourse Particles* (pp. 315-339). Amsterdam: Elsevier.

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 applied linguistics* (pp. 106–32). Oxford: Blackwell.

Zufferey, S., & Degand, L. (2013). Annotating the meaning of discourse connectives in multilingual corpora. In: *Corpus Linguistics and Linguistic Theory* (pp. 1–24). Available at https://dial.uclouvain.be/pr/boreal/object/boreal%3A137126/datastream/PDF_01/view [accessed 24 April, 2017].