Discourse research: a case study on attitudinal multiword discourse markers

Prof. Dr. Giedrė Valūnaitė Oleškevičienė

Mykolas Romeris University

Institute of Humanities

Research conducted in the context of the COST Action CA18209 “Nexus Linguarum” (“European network for Web-centered linguistic data science”) together with colleagues Chaya Liebeskind, Jerusalem College of Technology, Department of Computer Science. Further working with a group: Dimitar Trajanov, Ss. Cyril and Methodius University; Purificação Silvano University of Porto; Christian Chiarcos University of Frankfurt; Mariana Damova, Mozaika Ltd.

Personal scientific background

- scientific interests in the domain of humanities include discourse analysis, discourse annotated copora, lexicology, professional English and legal English,
- and in the domain of social sciencies, educational science the scientific interests include social research methodology, modern education, philosophical issues, creativity development in modern education system, etc
- Publications are available: https://www.researchgate.net/profile/Giedre_Valunaite_Oleskeviciene/research

Linking on discourse level

Information flow across multiple languages is a need –

How can linguistic and discourse research handle it? How can linguistic data (LD) in combination with natural language processing (NLP) techniques and multilingual language resources (LRs) (bilingual dictionaries, multilingual corpora, terminologies, etc.) allow it?

Challenges:

Linking in width and in depth of the available linguistic discourse level resources - on the “depth” level the linguistic discourse resources vary and there is a challenge to preserve the information by still keeping the linked data comparable.

- resources are unevenly developed in different languages and some languages may not have the material developed available for linking.

Discourse research background

Discourse markers (DMs) are used in both written texts and spoken discourse to connect ideas and guide the reader or the listener through expression by ensuring that the ideas are grasped correctly.

Multilingual studies are more complicated as languages differ in the use of DMs and their expression (Zufferey

and Cartoni 2012).

Recently, research on parallel translated corpora has gained interest (Dupont and Zufferey , 2017) (investigated the effect of register, translation direction, the shifts of meaning of DMs.

Multiword DMs case study validation

Language is generated in chunks by using formulaic constructions so called lexical bundles or multiword expressions

Some multiword expressions are used to organize discourse as DMs

Discourse layer is important in language production, communication, second language learning, and translation.

It is related to discourse relations, connecting ideas between sentences and bigger parts of the text.

DMs are connecting discourse elements, which form a category of lexical elements used to identify relations between discourse segments and provide text coherence.

DMs in language?

Observation:

If a learner misses words such as *you know*, or *well*, the native speaker cannot identify any error but the speech might sound impolite or even dogmatic.

Research on learner corpora to demonstrate the importance of discourse level knowledge, especially at more advanced levels of language learning (Granger, 2015; Cobb and Boulton, 2015)..

Contradictory discourse on DMs (*I think* and *You know)*

Language, especially spoken, is characterized by DM use; however, some of them (e.g., you know, I think, well) are sometimes referred to in

a critical manner, as indicating a lack of fluency.

- (O'Donnell and Todd, 2013)

Still, DMs are abundantly used and they enhance communication if used appropriately.

- (Hasselgren, 2002)

Assumption

DMs *I think* and *you know* are attitudinal indicators in discourse used to express and understand points of view and beliefs.

Research question

If stance attitudinal discourse markers: *I think* and *you know* demonstrate their functional stability as stance attitudinal discourse markers in translation and what changes they undergo in Lithuanian and Hebrew translation.

Research objectives

To create a parallel research corpus for the research

To identify multiword expressions used as stance attitudinal DMs.

To to analyze DM translations in

Lithuanian and Hebrew and to determine if they function as stance attitudinal DMs.

To disclose what linguistic forms they acquire in Lithuanian and Hebrew translation.

Research structural representation

Research context. Literature overview.	• Studies related to multiword expressions and their use as DMs; • The importance of DMs for discourse management; • Insights into DM translation.
Research methodology.	• Creating parallel corpus; parallel corpus knowledge alignment algorithm. • Automatic extraction of translations of DMs into Lithuanian and Hebrew; phrase-based MT for for a phrase table construction Koehn et al. (2003); • Annotation if multiword expressions *I think* and *you know* function as DMs or content words.
Research findings.	• Multiword expressions *I think* and *You know* used as DMs more frequently than other multiword expressions; • Their translation variants and stance attitudinal function.
Discussion.	• Some translations contain additional particles or connectives which could be related more to the pragmatic inferences drawn by the translators form the surrounding contexts, which relates to the observations by Blakemore and Carston (1999).
Conclusions.	• Multiword expressions *I think* and *you know* remain stance attitudinal DMs in Lithuanian and Hebrew translation but they demonstrate lexical variability in Lithuanian and Hebrew translations.

Translation challenge

Translator choices of inserting DMs to make the flow of the ideas smoother in the target text, however, they risk making the translation sound foreign or

Transposing the grammatical syntactic structure, ending up using different means of expressing DMs or simply omitting them.

Research methodology

Phrase-based MT was applied relying on two main reasons:

NMT techniques do not allow extensive processing of phrases and

NMT procedures are not as explicit as phrase-based MT processes.

The current study does not involve the full set of phrase-based MT systematic procedures, as it is used just for a phrase table construction, which is a single step of the phrase-based MT paradigm.

Research findings

Tendency to function as DMs	• *I think* and *you know* demonstrate a high tendency of being used as DMs and • the stability of remaining stance attitudinal DMs in Lithuanian and Hebrew translation.
Lexical form in translation	• either translated into multiword expressions or • one inflected word, or • they are completely omitted.
Pragmatic inferences	• additional particles or connectives which could be related more to the contextual pragmatic inferences.

Conclusions

I think and *you know* demonstrate a tendency to remain stance DMs in translation;

In Hebrew translation there is a tendency to use multiword DM translations, and there is a clear tendency for translators to give preference to male over female derivatives, which is due to the nature of the Hebrew language (Tobin, 2001);

In Lithuanian, there is a clear tendency observed for one-word DMs in translation;

Particle or connective integration into multiword expressions for both languages might carry the pragmatic meaning inferred from the context.

Future Research

- semantically annotate discourse relations as carriers of speaker attitudes in English.
- apply machine learning experiments aiming to reach automatic detection of stance discourse markers for further detecting opinions.
- adopt Chiarcos (2014) methodology to represent them as LLOD

References

Blakemore, D., & Carston, R. (1999). The pragmatics of and-conjunctions: The non-narrative cases. *CILISC1 (L'économie dans les structures, les computations et l'utilisation du langage*, 12-15.

Christian Chiarcos (2014). Towards interoperable discourse annotation. Discourse features in the Ontologies of Linguistic Annotation *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* http://www.lrec-conf.org/proceedings/lrec2014/pdf/893_Paper.pdf

Cobb, T., & Boulton, A. (2015). *Classroom applications of corpus analysis*.

Dupont, M., & Zufferey, S. (2017). Methodological issues in the use of directional parallel corpora: A case study of English and French concessive connectives. *International Journal of Corpus Linguistics*, *22*(2), 270–297.

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

Hasselgren, A. (2002). Learner corpora and language testing: Small words as markers of learner fluency. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 143–174.

Koehn, P., Och, F. J., & Marcu, D. (2003). *Statistical phrase-based translation*. University of Southern California Marina Del Rey Information Science Inst.

O'Donnell, W. R., & Todd, L. (2013). *Variety in contemporary English*. Routledge.

Tobin, Y. (2001). Gender switch in modern Hebrew. *Gender across Languages: The Linguistic Representation of Women and Men*, *1*, 177–198.

Zufferey, S., & Cartoni, B. (2012). English and French causal connectives in contrast. *Languages in Contrast*, *12*(2), 232–250.

Thank you ☺