SHORT TERM SCIENTIFIC MISSION (STSM)
SCIENTIFIC REPORT

This report is submitted for approval by the STSM applicant to the STSM coordinator

Action number: CA18209
STSM title: Creating a multilingual corpus for formulaic language (multiword expressions) research
STSM start and end date: 22/02/2020 to 05/03/2020
Grantee name: Giedre Valunaite Oleskeviciene

PURPOSE OF THE STSM:

(max.200 words)
The purpose of the STSM was extending the available resources and providing linguistic processing for several languages by creating a multilingual parallel corpus (including English Lithuanian and Hebrew) based on social media texts and working on multiword expressions in social media texts. We also, had an objective of making the created corpus a part of LLOD by sharing it and interlinking via CLARIN open language resources in a semantically interoperable manner.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

(max.500 words)
First the parallel texts in English, Lithuanian and Hebrew were extracted from TED talks transcripts and then the sentences were aligned to make parallel corpus for further research. The corpus contains 87230 aligned sentences and it is going to be shared publicly on CLARIN. Then further, we focused on multiword expressions and narrowed our research focusing on multi word expressions which are used as discourse markers. The research of discourse markers is relevant to multiple research areas such as linguistics, translation, and Natural Language Processing (NLP). Discourse marker and relations enable the researchers in linguistics to focuse on pragmatics and to analyze how texts are organized beyond the sentence level, and how textual coherence is ensured. NLP researchers and practitioners focus on the structures that bind together multiple sentences and ensure the connection of ideas. Discourse markers ensure textual cohesion and according to Fraser (2009) relate separate discourse messages, for example, such phrases as *you know, I mean, of course,* etc. which are characteristic of spoken language (Furkó & Abuczki (2014), Huang (2011)). Thus, 3314 aligned sentences containing the earlier mentioned multi word expressions were extracted and then manually annotated spotting the cases when the expressions are used as discourse markers, for example in case (1) the multi word expression *you know* is used to introduce a new discourse message, while in case (2) they are content words fully integrated into the sentence.

(1) You know, but really it's the kind of same old crap we've had for the last 30 years.

(2) I'll let you know when you can look again.

After that, the variations of the translations of discourse markers into Lithuanian and Hebrew were extracted for comparative study spotting out the variations in translation.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

It was identified that English multi word expressions used as discourse markers demonstrate variability in Lithuanian and Hebrew translations: they are either translated into multi word expressions or in one inflected word in the target languages, or are omitted at all. For example, in Lithuanian multiword expression discourse marker *you know* splits into a number of multi word expressions and also one word translations. Multi word expressions could be classified into cases representing pronoun-verb phrase *jūs žinote, jūs suprantate, jūs įsivaizduojate, jūs esate girdėję* or particle-verb phrase: (*na/juk/ir*) *žinote, suprantate*, or connective-verb phrase (*kaip, kad*) *žinote, matote* where connective could be used in a pre- or post- position to the verb.
One word translations mainly include verbs, for example, *žinote, suprantate, įsivaizduojate*, and etc., which due to Lithuanian being a highly inflected language (Zinkevičius, Daudaravičius & Rimkutė, 2005) fully represent the verb-pronoun cases. It should be noted that Lithuanian translations of pronoun-verb multi word expressions and one word – verb cases could be considered as almost word for word translations. So, it could be said that more interesting cases which represent translator choices of particle-verb or connective-verb multi word expressions which due to the use of particles and conjunctions also carry out certain rhetorical discourse meaning.

FUTURE COLLABORATIONS (if applicable)

The collaboration with the host institution will continue. We are expecting to filter, classify and analyze the translations of the multiword expressions used as discourse markers. There is also a tentative plan to invite other researchers from our COST action who could be interested in joining and working on a case study of multiword expressions in social media texts and continue with a view on modelling the corpus and the annotations with a LD approach.
We are aiming at a paper as the results of research within this STSM and it will be sent either to the prominent journals (e.g. Journal of Pragmatics, etc.) or conferences/workshops in the field. We are also planning to present our work at one of the following COST events