# 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: Researching discourse markers expressing opinion with machine learning techniques in a multilingual corpus**
**STSM start and end date: 16/08/2020 to 30/08/2020**
**Grantee name: Giedre Valunaite Oleskeviciene**

**PURPOSE OF THE STSM:**

(max.200 words)
The STSM was carried within the framework of work group WG4, task 4.2 "Use Cases in Humanities and Social Sciences", use case UC4.2.2 "Use Case on Social Sciences" focusing on researching social attitudes through attitudinal discourse markers expressing opinions. The purpose of the STSM was providing linguistic processing for several languages analyzing multilingual corpus data in English, Bulgarian and Lithuanian in preparation of the data for the research of automatic detection of discourse markers expressing opinion by using machine learning. We also, had an objective of enriching the existing multilingual corpus - TED-EHL parallel corpus published in LINDAT/CLARIN-LT repository as an open language resource.

**DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS**

(max.500 words)
First, we enriched TED-EHL parallel corpus based social media texts with 4 languages so that the multilingual corpus contains alignments of Lithuanian, Bulgarian, Hebrew, Portuguese, Macedonian, and German languages with English as pivot language with a size of 1.3 million sentences. Then the part of the enriched multilingual corpus comprising 2428 English-Bulgarian-Lithuanian aligned sentences containing the multiword expressions (MWE) as discourse markers or content expressions was manually annotated (1 or 0) in preparation for the machine learning experiments. Example (1) below classifies the multiword expression *you know* as a discourse marker (annotated 1) used to introduce a new discourse message, whereas example (2) represents content words (annotated 0) fully integrated into the sentence.

(1) That's ridiculous. *You know*, this is New York, this chair will be empty, nobody has time to sit in front of you.
(2) *You know* some people who say "Well"

The manual annotation of the data was carried out in order to refine the data for the successive elaboration aiming to reach automatic detection of discourse markers expressing opinion. Finally, the discourse markers expressing opinion and their semantics will be compared cross-

linguistically looking for regularities and striking features aiming at drafting a publication within the framework of the COST action of NexusLinguarum. Also, following the manual annotation the results of the experiment of detection of discourse markers expressing opinion are going to be presented in our COST organized workshop in Skopje September 2021.

**DESCRIPTION OF THE MAIN RESULTS OBTAINED**

The manual annotation was carried out in order to provide the data for machine training and a dataset of size 1450 KB and 1870 entries was prepared. Table 1 shows an example of annotated corpus prepared for machine learning experiment.
Table 1: Extract of annotated corpus entries

| MWE | Sentence chunk | Context | DM presence |
|---|---|---|---|
| I remember | I remember so many huge, hollowed out, haunted eyes | I remember so many patients, their names still vivid on my tongue, their faces still so clear. I remember so many huge, hollowed out, haunted eyes staring up at me as I performed this ritual. And then the next day, | 0 |
| I think | I think he surely must have known by then | It was an offering, an invitation. I did not decline. I percussed. I palpated. I listened to the chest. I think he surely must have known by then that it was vital for me just as it was necessary for him. | 1 |

We also started working on semantics of discourse markers and we are adopting ISO 24618-8 annotation scheme to semantically annotate discourse relations as identifiers of speaker attitudes in English so that later we could use Chiarcos (2014) methodology to represent them as LLOD. (Chiarcos, C. (2014, May). Towards interoperable discourse annotation. Discourse features in the Ontologies of Linguistic Annotation. In *LREC* (pp. 4569-4577).
The example of manually annotated data according to ISO 2461-8 scheme is provided in table 2 below.
Table 2. Extract of annotated corpus according to ISO 246-8 scheme

| Sentence chunk | Context | Presence of DM | Argument1 | Discourse marker | Argument2 | DM label | Arg1 label | Arg2 label |
|---|---|---|---|---|---|---|---|---|
| — for those of you doing the numbers, you know that's 132 Americans | gave 10,000 dollars or more to federal candidates, and in this election cycle, my favorite statistic is .000042 percent — for those of | 1 | my favorite statistic is .000042 percent — for those of you doing the numbers | You know | that's 132 Americans | expansion | narrative | expander |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | you doing the numbers, you know that's 132 Americans — gave 60 percent of the Super PAC money spent in the cycle we have just seen ending. | | | | | | | |
| turned, you know, a part of the evil empire | Two were particularly inspiring to me. One was Ray Anderson, who turned -- (Applause) -- turned, you know, a part of the evil empire into a zero-footprint, or almost zero-footprint business. Why? Because it was the right thing to do. | 0 | | | | | | |

**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. We are aiming at a paper presenting the results of research within this STSM and it will be sent either to the prominent journals or conferences/workshops in the field. We are also planning to present our work at the following COST events