Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: CA18209

Grantee name: Maria da Purificação Moura Silvano

Details of the STSM

Title: Towards a multilingual lexicon of discourse markers

Start and end date: 10/09/2020 to 17/09/2022

Description of the work carried out during the STSM

The two main objectives of the STSM were to: (i) work on the construction of a multilingual semantic vocabulary of discourse markers as LLOD, and (ii) prepare a roadmap for future research.

Regarding the first objective, I and Mariana Damova, at Mozaika, in line with the work we have been pursuing within working group 4.2.2., developed a vocabulary of discourse markers for English, Portuguese and Bulgarian. To this end, we used the multilingual parallel corpus with data from nine languages, Bulgarian, Lithuanian, German, European Portuguese, Hebrew, Romanian, Polish, and Macedonian, with English as a pivot language, that was previously built by our working group using the publicly available TED Talk transcripts to study discourse markers. In order to represent the meaning of the discourse markers, we applied an annotation scheme that comprises discourse relations from ISO 24617-8 with a plug-in to ISO 24617-2 for communicative functions. This annotation scheme was also developed by our working group and already published at LREC 2022 (Silvano, Purificação; Damova, Mariana; Oleškevičienė, Giedrė Valūnaitė; Liebeskind, Chaya; Chiarcos, Christian; Trajanov, Dimitar; Truică, Ciprian-Octavian; Apostol, Elena-Simona & Baczkowska, Anna (2022). "ISO-Based Annotated Multilingual Corpus For Discourse Markers". In Proceedings of the 13th Edition Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-24 June, European Language Resources Association (ELRA), ACL Anthology).

The first step was agreeing on some general annotation guidelines and some annotation procedures. The next step was to assess the different semantic and pragmatic values conveyed by the discourse markers. During the process of identifying the different values of discourse markers across the three languages we discussed some examples to assert the degree of inter-annotator agreement. The final step was a comparative quantitative and qualitative analysis of the results obtained from the analysis.

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

The annotated datasets are evidence of the operability of the annotation scheme, and will serve as exemplary cases for the other seven languages datasets (Hebrew, Lithuanian, Macedonian, German, Romanian, Polish and Italian) that constitute our corpus. These steps will allow us to formalize the ISO-based annotation scheme in an Web Ontology Language (OWL) ontology for publishing and integrating data, and to convert annotations to RDF, link with ontology and perform conjoint queries.

Regarding the second objective, we looked at projects calls that may be of interest having in mind the research we have been doing within *NexusLinguarum* and intend to continue pursuing, debated possible topics, thought of potential stakeholders, and established a timeline to outline a project proposal.

Description of the STSM main achievements and planned follow-up activities

The STSM achieved its planned goals and expected outcomes. Thus, due to this STMS, I in collaboration with Mariana Damova, at Mozaika, were able to: (i) test the reliability of a comprehensive interoperable Discourse Markers taxonomy able to represent not only the semantic meaning of discourse markers but also their pragmatic meaning in a sample of the multilingual parallel corpus created by working work 4.2.2.; (ii) create a parallel vocabulary of discourse markers, in three languages, English, Portuguese and Bulgarian, being two of them low-resourced languages; (iii) develop a taxonomy prototype to be applied to other datasets of different languages Lithuanian, German, Hebrew, Romanian, Polish, Macedonian, Italian, some of which are under-resourced languages, as well; (iv) perform a quantitative and qualitative study of the semantic and pragmatic role of discourse markers across three languages; and (v) devise a plan to formulate a proposal of future research that can follow-up the investigation that we have been conducting in working group 4.2.2. led by Mariana Damova.

During this STMS, due to Mariana Damova's expertise, I got the opportunity to become better acquainted with semantic representations of discourse markers, and with processing methods to extract semantic and pragmatic information related to discourse markers. Furthermore, the work that we did will enable an OWL ontology if ISO discourse relations and communicative functions and LLOD modelling annotations.

The results of the tasks accomplished during the STSM will be presented at the *International Scientific Interdisciplinary Conference on "LLOD approaches for language data research and management (LLODREAM2022)*. We also intend to publish the outcomes in a special issue of the journal *Discourse Processes*.

Overall, the assessment of the STSM is very satisfactory, because we advanced on the work of the use case for linguistic data science, while simultaneously contributing to the achievement of the scientific objectives of *NexusLinguarum* Action, and we also strengthened the collaboration between me and Mariana Damova, having plans to develop together a project that can continue the investigation made possible by this Cost Action.