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

Action number: CA18209

Grantee name: Aleksandra Tomaszewska

Details of the STSM

Title: **Towards 'Universal Discourse' – the Project of a Multilingual Discourse Description**

Start and end date: 12.10.2022 to 24.10.2022

Description of the work carried out during the STSM

Description of the activities carried out during the STSM. Any deviations from the initial working plan shall also be described in this section.

(max. 500 words)

The training component (Phase 1) included activities designed to familiarize the participant with relevant literature, and infrastructure and resources available at the host institution. Linguistic Linked Open Data (LLOD) technologies, introduced in Chiarcos, Hellmann, and Nordhoff (2011: 245–275), prior studies and implementations, and challenges that scholars currently face were addressed. Examples of materials discussed include Gromann et al. (2022, under review). This stage focused on a field literature review in which relevant books, articles, journals, and resources (corpora and databases), as well as annotation

[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.

tools used by the host, were presented. This broad overview of projects on discourse markers, e.g., **ISO-Based Annotated Multilingual Corpus For Discourse Markers** (Silvano et al. 2022) and discourse relations, e.g., **TED Multilingual Discourse Bank** (Zeyrek et al. 2019; Zeyrek et al. 2022) was followed by an introduction to LLOD technologies and applications, particularly in data publishing and its integration, e.g. in **recent studies presented at LLODREAM**.

The emphasis in Phase 2 was on skill-building and cooperation. The theoretical insights along with the materials and tools were used to design potential follow-up actions. Cooperation and expertise exchange on completed projects in both institutions, and future plans for annotation and research, were discussed. The projects conducted within the Linguistic Engineering (LE) Group (http://zil.ipipan.waw.pl/) at the Institute of Computer Science of the Polish Academy of Sciences (https://www.ipipan.waw.pl/en) and its resources were presented (http://clip.ipipan.waw.pl/):

(1) General corpora of Polish: (A) **The National Corpus of Polish** (http://nkjp.pl/poliqarp/) and (B) **a corpus of materials from the last decade** (http://korpus-dekady.ipipan.waw.pl/).
(2) The PCC (**Polish Coreference Corpus**)(Ogrodniczuk et al. 2015), is available in Open Access (http://cothec.nlp.ipipan.waw.pl/index.xhtml#/) together with Coreference Tools (http://zil.ipipan.waw.pl/PolishCoreferenceTools).
(3) The PDC (**Polish Discourse Corpus**), based on the PCC and described by Celina Heliasz and Maciej Ogrodniczuk (2019) is available at: http://zil.ipipan.waw.pl/PolishDiscourseCorpus). The annotation was completed using Discann (http://zil.ipipan.waw.pl/Discann).

Experience from PDC is of particular relevance for the cooperation as it will focus on annotating discourse relations in Polish and Portuguese (and potentially in other languages).

Another phase included creating a **conceptual framework for annotating discourse relations** in multilingual data that both sides could contribute to. The utilization of various annotation resources was explored, and possible materials were selected. The work continued with establishing the joint research work allowing for the description of discourse relations across languages. It was established that:

(1) The study will **expand on prior achievements in discourse description** and its constraints for Polish and Portuguese.

(2) First, the aim will be to **test the ISO standard for annotation for both languages** using the following frameworks: Language resource management — Semantic annotation framework (SemAF) — Part 5: Discourse structure (SemAF-DS), https://www.iso.org/standard/57083.html, and Part 8: Semantic relations in discourse, core annotation schema (DR-core), https://www.iso.org/standard/60780.html.

(3) Research material will be **multilingual corpus use**, e.g., Europarl (Koehn 2005), which includes versions in 21 European languages and can be the basis for multilingual study.

The work will be conducted **using the same annotation tool.** Inforex web-based software (http://www.inforex-work.clarin-pl.eu, Marcińczuk and Oleksy 2019) used for ISO annotation of the Polish corpus was proposed for testing also for Portuguese.

References:

(1) Chiarcos, Christian, Hellmann, Sebastian, and Nordhoff, Sebastian (2011). Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group. TAL (*Traitement Automatique des Langues*), 52(3), 245-275.

(2) Europarl: A Parallel Corpus for Statistical Machine Translation, Philipp Koehn, MT Summit 2005, pdf, https://www.statmt.org/europarl/.

(3) Gromann, Dagmar et al. (2022, under review). Multilinguality and LLOD: A Survey Across Linguistic Description Levels https://www.semantic-web-journal.net/system/files/swj3259.pdf.

(4) Heliasz, Celina and Maciej Ogrodniczuk. Eksplicytność a implicytność w świetle analizy korpusowej (meta)tekstu. Linguistica Copernicana, 16:75–100, 2019, https://apcz.umk.pl/czasopisma/index.php/LinCop/article/download/LinCop.2019.004/25483.

(5) LLODREAM2022 (2022). LLOD approaches for language data research and management. Conference Program, https://llodapproaches2022.mruni.eu/?page_id=323.

(6) Marcińczuk, M. & Oleksy, M. (2019). Inforex — a Collaborative Systemfor Text Corpora Annotation and Analysis Goes Open. In Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, pages 711—719. Varna, Bulgaria. INCOMA Ltd.

(7) Ogrodniczuk, Maciej et. al. (2015). Coreference in Polish: Annotation, Resolution and Evaluation. Walter De Gruyter.

(8) Silvano, Purificação et al. (2022). ISO-based Annotated Multilingual Parallel Corpus for Discourse Markers, Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2739–2749, Marseille, 20-25 June 2022, http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.293.pdf.

(9) Zeyrek, Deniz, Amália Mendes, Giedrė Valūnaitė Oleškevičienė, and Sibel Özer. "An Exploratory Analysis of TED Talks in English and Lithuanian, Portuguese and Turkish Translations", Contrastive Pragmatics 3, 3 (2022): 452-479, doi: https://doi.org/10.1163/26660393-bja10052

(10) Zeyrek, Deniz, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. Language Resources and Evaluation 54(2): 587–613.

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

Description and assessment of whether the STSM achieved its planned goals and expected outcomes, including specific contribution to Action objective and deliverables, or publications resulting from the STSM. Agreed plans for future follow-up collaborations shall also be described in this section.

(max. 500 words)

The primary purpose was to create **a basic framework for a project on the description of discourse relations in multilingual data.** A substantial portion of the mission was spent working with current linguistic resources and tools, and being familiarized with the goals, annotation layers, and methodologies applied in the host institution that may be useful in the planned project. Conducted in response to the lack of resources including information for many languages as well as a multi-layer representation beyond the morphosyntactic description, this STSM helped identify possible ways of designing a new study and solutions to the existing challenges.

Another objective was investigating the potential of LLOD technologies for publishing and integrating data. Recent studies conducted as part of NexusLinguarum, as well as current advancements in the field, were highlighted. As the host's work prioritizes multilingual corpora and features one of the first implementations of LLOD technology for discourse annotation, this

STSM was an opportunity to get acquainted with the general principles of the strategies and challenges in research of this type.

Importantly, the STSM's helped identify areas of cooperation, including:

(1) The need of **unifying formalisms** for both languages

(2) **Testing ISO standards** across languages

(3) **Testing the same annotation software**

(4) Sharing and discussing project results, limitations, and resources

(5) **Presenting** results (jointly)

(6) **Designing future research** including more languages.

Candidates from existing open-source multilingual corpora (Europarl) were chosen to serve as a pilot study material for annotating in accordance with the SemAF standard. The ISO categories were compared, and it was found that the requirements for both languages and software needed to be revised. The goals of the pilot study were discussed: **testing the sufficiency of the ISO standard for research in equivalent (but not parallel) data for both languages** to determine whether its application is the right solution for both Polish and Portuguese and would **allow for future standardization of the description of discourse relations across languages.**

The project was divided in six phases. The first four phases include joint annotation work and discussions on its result as well as working on joint resources:

Phase 1: Deciding on the scope and limitations of research material and establishing a detailed research procedure
Phase 2: Testing annotation for both languages on the research material
Phase 3: Comparing results
Phase 4: Deciding on adjustments to the standards applied

Joint dissemination and future research areas will be conducted:

Phase 5: Presenting cooperation effects and proposing adjustments to the standard at an international conference (e.g., ISO Workshop on Interoperable Semantic Annotation (https://sigsem.uvt.nl/isa18/))

Phase 6: Designing a multilingual project

If the test is successful, another phase will be **incorporating more languages** and **establishing partnerships with more institutions**, e.g. from the NexusLinguarum working groups. The final plan includes testing the ontology for discourse relations, https://purl.org/olia/discourse/discourse.Nexus.owl. In line with activities of WG 4.2.2., LLOD transformation and linking of the annotations will be conducted and presented at LDK (http://2023.ldk-conf.org). Finally, data gathered, a joint presentation and deliverables created while working on the study will be **published in Open Access** and will feature links to other resources.