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

Action number: CA19102

Grantee name: Hiwa Aasdpour

Details of the STSM

Title: **Designing a shared task on multimodal argument mining in low-resource and endangered language of Western Asia**

Start and end date: 06/10/2022 to 20/10/2022

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)

In this STSM the main goal which was establishing collaboration and networking between the grantee and the host is established. We discussed various topics, for example, designing a shared database where we can put our data with the aim of making it available for a shared task on multi-modal machine learning. My language expertise is on low-resourced Iranian languages and the host's expertise is on Arabic and low-resourced languages such as Domari and other Romani languages. We found out a common interest on working with free speech data which is essential in designing systematic corpus of free speech which can be useful for research and marketing purposes. For example, various linguists and researchers can benefit from the data for different projects. Different IT experts in the field of computational linguistics can use the data for NLP and machine learning purposes. Beside setting up a systematic database we also discussed various topics related to publicizing the data (see below the planned tasks). Such data can be useful for multilateral queries on topics related to language contact phenomena. For example, the languages which the host and me are focusing on are in contact with several other languages for centuries. So it requires a careful consideration how to treat the annotation of these languages in the database. To do that, we discussed the collaboration between different scholars in our networking and we aimed at setting up a proposal for a prospective project.

[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 aim of this project is to conceive a research on processing complex structures in the aforementioned languages that will supply the database with all the relevant data. This enormous task can only be accomplished collaboratively.

During the visit, we established the basis for a networking initiation between Inalco and Goethe University Frankfurt. We agreed on a dataset to be created / adapted and devise a financial plan on how to cover the costs of this effort, where decided that it will involve human annotation (e.g., though crowdsourcing). We also discussed and proposed evaluation methodologies for the shared task and for a baseline approach for the shared task. To do that, we planned a draft timeframe for the task (training data release, test data release, results submission period, evaluation period, etc.). In the meetings, we also described the shared task, to be publicized on a website and through mailing lists inviting participants in the weeks following the visit. We outlined a plan of research for our collaborative work in the area. This mutual linguistic and computational data exchange will provide a solid base for multimodal data analysis and enhance the efficiency of Asadpour's project by reducing working steps. The synchronization of the remaining working steps will be achieved by stipulating a common framework and by refining the methodology for the visualization of the results. One of the main outcomes for database designing is the creation and development of a Comparative Corpus of IST Languages[2]. This will form a unique corpus for those languages, freely accessible and usable for a wide range of digital humanities-related research (WG1, 2, 3, 4, 5, 6, 7, 8).

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)

In this STSM mission, I explained my current research to the host, obtained valuable feedback and found common research interests. During the STSM, we had a discussion of the main workload concerning the analysis of data of Kurdish, Arabic, and Domari. We discussed the topic of commitment to the maintenance of the technical infrastructure (corpora and digital software). For example, during this time, I obtained an agreement from Max-Planck Institute for Psycholinguistics for technical support and archiving the data. MPI gives the platform to securely archive the data and to make it available to the public by giving various types of permission such as full type permission, middle or less type permission to the researchers and people. Furthermore, MPI's platform gives the opportunity for annotating the data to be used for various purposes. Having archived the data in MPI's archive platform facilitates issues related to the copyright etc. by considering all potential standards and considerations. This is highly significant especially in case of the local varieties such as Kurdish and Domari as they are low-resource languages in geo-politically sensitive regions where fieldwork is very hard for foreigners and western researchers. Even the same applies for local varieties of Iraqi and Syrian Arabic because there is no rich corpus and database for these varieties.

Aside from archiving and open access to the data, we also discussed how we can have a better integration of language and computer modelling for the future results in a more robust way. To do that, we thought about collaboration and networking with other departments especially computer engineering. Moreover, we discussed mutual support in the analysis of the

[2] IST is an abbreviation for Indo-European, Semitic, Turkic languages

aforementioned languages, e.g., their structures such as passive and relative clause structures. To achieve our goals, we planned to organize a workshop meeting in 2023 or possibly 2024 with several intended colleagues in the field of empirical, historical, and computational linguistics. The date will be finalized based on the availability of the intended participants. In this workshop, we aim to gain an opportunity to discuss and question the findings and details of the work. The aim is the preparation of a joint project in future on 'Multiscale Visioning of Contact Linguistics in morphosyntactic change' and the intention is to shed light on the computational analysis of contact situation in Indo-European and Semitic languages with a focus on Western Asia region. Additionally, we planned to publish a volume on a topic such as passive and relative clause structures in collaboration with the networking members. We aim to invite the intended collaborators for this purpose.