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

Action number: CA18209

Grantee name: Flavia Sciolette

Details of the STSM

Title: Post data resurgo: The annotated texts as a source for lexical resources and their linking.

Start and end date: 01/10/2023 to 30/10/2023

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.

The aims of this STSM were primarily: i) the study of existing and ongoing models for the creation of lexical resources compliant as Linked Data, particularly with regard to the concept of "attestation" as “a special form of citation that provide evidence for the existence of a certain lexical phenomena”[2]; ii) the application of these models to a sample of a multilingual terminological resource on historical varieties.

The first activity was a survey of existing resources on historical linguistic varieties. Although the number of such projects and dedicated publications is on the rise, it is important to acknowledge that when stricter parameters are established (e.g. freely available, compliant with the current standards), the framework appears to narrow significantly. For the survey, I considered the “Virtual Language Observatory[3]” of CLARIN, existing clouds for resources (e.g. LOD cloud), with the addition of repositories of GitHub, where it needed. The survey will be available as a catalogue on my GitHub.

These resources still appear to be poorly represented as Linked Data and, in general, more accessible for consultation purposes rather than for analysis or exploitation[4] in linguistic tasks (with some notable exceptions, such as for the Latin language). I believe that this survey activity has been crucial in understanding the need to establish guidelines and best practices, even in a philological/humanistic context, which is typically less considered.

[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.
[2] https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#attestation
[3] https://vlo.clarin.eu/;jsessionid=6D9061E3D4FCCABAEBC5C799DDCEC8C0?0
[4] https://lod-cloud.net/

Subsequently, two case studies were selected: i) the scientific terminology in two medieval texts; ii) the terminology in the Italian translation of the Babylonian Talmud. The latter was available from the work of Giovannetti et al. 2020 (Giovannetti, E., Bellandi, A., Dattilo, D., Del Grosso, A.M., Marchi, S., Pecchioli, A., and Piccini, S. 2020. The Terminology of the Babylonian Talmud: Extraction, Representation and Use in the Context of Computational Linguistics. Materia Giudaica, XXV), while in the first case, an extraction phase from texts annotated in TEI-XML language was required. The extraction was performed using a Python script created specifically for this purpose, which included all the scientific terms identified by human annotators. It is essential to note that this initial sample had to serve as a gold standard. A sample of approximately 100 terms in Castilian and medieval Italian was obtained.

For the subsequent phase of formalizing the entries, the OntoLex-Lemon[5] model was chosen. In this regard, discussions with Professor Christian Chiarcos were fundamental to set the work, as well as the opportunity to participate in meetings of the Lexica group on the ongoing frequency-attestation-corpora (FrAC) module. In particular, two main aspects were considered: i) the formalization of metrics in the context of the module, such as TF-IDF (important for topic modeling tasks, especially in multi-topic texts like the Talmud); ii) the formalization of attestation for forms in texts with extensive spelling variation, as is the case in manuscript witnesses.

Following these discussions, it was decided to formalize entries, senses, forms (from the core module), multiword expressions (from the decomp module), and TF-IDF (from the FrAC module) for Talmudic terminology. For medieval terminology, the core module, including reference and lexical concept, and attestation with the FrAC module[6] were chosen. Attestations are based on annotated text portions linked to the lexical entry through the Maia tool, an integrated environment for creating annotated corpora and lexical resources, under development by the KLAB of CNR-ILC. Regarding the export formats of these materials, I evaluated CoNLL-U+[7], CoNLL-RDF[8], and the NIF[9] annotation standard.

I presented the results of the mid-term project and some of my previous activities during the ILKA seminars at Augsburg (https://www.uni-augsburg.de/de/fakultaet/philhist/studium/vortragsreihen/interdisziplinares-linguistisches-kolloquium-augsburg), where I was invited by Professor Chiarcos. I also had the opportunity to participate in the kick-off meeting of the Computational Linguistics laboratory ([https://coli-augsburg.github.io/)]) and the Digital Humanities group, held on October 4th and 11th, respectively.

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

The first result achieved during the STSM consists of the creation of two terminological resources: one with approximately 6,000 formalised terms previously extracted from the Italian translation of the Talmud and the second with about 100 terms in Castilian and Old Italian. The first resource will be employed in the context of Talmud project, the latter will be made available through CLARIN and on GitHub. I also believe that this experience has been particularly beneficial for the reflection on lexical linking methods, especially for the export formats for this type of annotation, which will undoubtedly enrich the development of the Maia tool currently under development at our institute.

Currently, a review of existing projects on historical varieties is also in progress, which could be made available as an additional technical report. I believe that promoting and disseminating best practices in these projects can align with the goals of the COST Nexus Linguarum action.

We are also awaiting the publication of calls for papers for the LREC-COLING workshops for the submission of a contribution, which will later be expanded into a journal article.

[5] https://www.w3.org/2016/05/ontolex/
[6] https://github.com/ontolex/frequency-attestation-corpus-information
[7] https://universaldependencies.org/format.html#extensions
[8] https://github.com/acoli-repo/conll-rdf
[9] https://bpmlod.github.io/report/nif-corpus/index.html

The next steps of work include focusing on multiword expressions, given the importance of formulae and expressions in technical texts.