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

Action number: CA 18209

Grantee name: Thierry Declerck

Details of the STSM

Title: : Interlinking Latin Language Data in the Linguistic Linked Open Data cloud

Start and end date: 25/09/2022 to 03/10/2022

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.

The main topics to be covered by the stay of Mr Thierry Declerck at the Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE), directed by Prof. Marco C. Passarotti, were the extension of the formal representation of language data delivered by the ERC Project LiLa (LiLa: Linking Latin Building a Knowledge Base of Linguistic Resources for Latin, https://lila-erc.eu) in the context of the OntoLex-Lemon framework, which has been developed in the context of a W3C Community Group (https://www.w3.org/2016/05/ontolex/). The goals of the STSM, formulated in the first months of 2020 (but the STSM had to be postponed due to the pandemics), have been precised and the focus was on the use of the morphology (https://www.w3.org/community/ontolex/wiki/Morphology) and the FrAC (https://acoli-repo.github.io/ontolexfrac/) modules, both in an advanced development stadium, for LiLa data, with the aim to support the cross-linking of lexical and corpus data in LiLa.

We had first an introductory session, with Prof. Marco C. Passarotti and Thierry Declerck, resulting in the delimitation of the topics for the discussions to be held during the STSM. We had then a general session, with Prof. Marco C. Passarotti, Francesco Mambrini (lexicon and FrAC), Matteo Pellegrini (morphology and FrAC) and Thierry Declerck. This was followed then by bi-lateral sessions between Matteo Pellegrini and Thierry Declerck and between Francesco Mambrini and Thierry Declerck. A final session took place with Prof. Marco C. Passarotti, Matteo Pellegrini and Thierry Declerck. Informal sessions took also place with Federica Iurescia and Flavio M. Cecchini (who are working in the field of Universal Dependencies for LiLa corpus data).

[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.

Discussions on the use of the morphology module were dealing mainly with the issue of underspecification of morphological information contained in the lexical knowledge base of LiLa. This is particularly relevant when it comes to the issue of relating Latin corpora data to the lexical knowledge base of LiLa. This relation is to be represented via the FrAC module. The issue is that different corpora (of Latin language) have different types of annotation associated with them. Textual content of only some corpora has been lemmatized. Some corpora contain from zero morphological annotation to more complex morpho-syntactic annotation strategies.

For example, if the form “rosa” (rose) occur in a corpus, but only lemmatized (to “rosa”) and with no morphological information, linking it to a fully specified (RDF encoded) lexicon of LiLa is not trivial, if the lexicon is describing three ontolex:Form variants: one for the nominative-singular, one for the vocative-singular and one for the ablative-singular declensions of “rosa”. We discussed therefore ways to describe a morphology pattern that would allow to represent “rosa” as being either nominative-singular or vocative-singular or ablative-singular.

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.

An encoding (with a suggested class for the morphology module -- morph:UnderspecifiedMorph) of the above mentioned topics could be:

```
underspecified_morph:Form_rosa_nom_voc_abl
  rdf:type morph:UnderspecifiedMorph ;
  lexinfo:case lexinfo:nominativeCase ;
  lexinfo:case lexinfo:vocativeCase ;
  lexinfo:case lexinfo:ablativeCase ;
  lexinfo:gender lexinfo:feminine ;
  lexinfo:number lexinfo:singular ;
  ontolex:writtenRep "\"rosa\""@la ;
.
```

This type of encoding led us to discuss the use of FrAC for encoding, among others, the frequency of occurrences of instances of the so-called “frac:Observable” class. As a fully specified fine-grained morphosyntactic information associated with a lexical entry is very seldomly (if at all) to be found as annotation in a corpus, we need to take a subset of the information included in the LiLa lexical knowledge base, and to adapt it to the type of annotation of language data that one can find in a specific corpus. In this, the RDF code displayed above can function as an instance of frac:Observable.

If the wordform “rosa” is found in the corpus without morpho-syntactic annotation, the underspecified linguistic object (sketched above) can be associated with it, as no decision is taken with respect of the case (nominative or vocative or ablative) of the word/string in the corpus (if not correspondingly annotated). Therefore, a frequency value can be associated with the underspecified lexical object, regardless of the amount of morpho-syntactic information included in the corpus.

On the other hand, if a wordform “rosa” is found in a corpus annotated for fine-grained morpho-syntactic information, and we know it is, say, a nominative, than the idea would be to generate a new object consisting of the pairing of the (underspecified) form with the morphosyntactic properties that it expresses (in this case, nominative and singular), and having this new object as a frac:Observable – since this class is open to user-defined items. Furthermore, the frequency information can also be projected to the corresponding underspecified form (i.e., this would also count as an attestation of the underspecified “rosa”).

This would allow us to be able to express information about frequency corresponding to different levels of granularity of the annotation, without having to make the lexical representation more complex, as the second type of frac:Observable could be generated on the basis of the data provided by the available corpora, rather than being stored in the lexicon for all wordforms.

As a main result, it appears that we can usefully link aspects of the morph and the FrAC modules, by representing underspecified morphology of lexical objects as instances of the class frac:Observable.

We plan to extend our joint work to a large-scale implementation of our discussed approaches to underspecification, especially for generating FrAC representations about frequency and attestation to another language having a rich morphology, like German.

We will discuss about how information on vowel length might make it possible to narrow down the range of the possible morpho-syntactic features expressed by the forms.

We will present some of our findings in the telcos of the W3C Ontology-Lexica Working Group.

We will write a collaborative paper to be submitted to a relevant conference or workshop.