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

Action number: CA18209

Grantee name: Paola Marongiu

Details of the STSM

Title: Lexical semantic change detection in Latin: a use-case on medical latin

Start and end date: 8/05/2023 to 29/05/2023

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.

(max. 500 words)

Grantee enters max 500 word summary here.

During the STSM I performed the following activities (details about each activity can be found in the Supporting Document "Extensive report"):

I conducted a preliminary literature review of the existing applications of word embedding for lexical semantic change research in Latin. In particular Rodda et al. (2019), Sprugnoli et al. (2019), Ribary and McGillivray (2020), Sprugnoli et al. (2020).

I have checked the language resources to perform the study, as detailed below.

1) I identified the LatinISE corpus (McGillivray and Kilgarriff, 2013) as a corpus of reference to perform the analysis. Then I identified the medical texts in LatinISE, as per table below where '1' in the last column refers to medical texts and '0' to non-medical texts:

[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.

	id	title	creator	date	type	file	medical text
221	IT-LAT0382	De medicina	Celsus, Cornelius Aulus	30	prose	lat_0030_IT-LAT0382.txt	1
562	IT-LAT0895	Medicina ex oleribus et pomis	Gargilius Martialis, Quintus	260	prose	lat_0260_IT-LAT0895.txt	1
634	IT-LAT0987	De observatione ciborum	Anthimus	550	prose	lat_0550_IT-LAT0987.txt	1
837	ITMQDQ-247	medicamina faciei	Ovidius Naso, Publius	30	poetry	lat_+0030_MQDQ-247.txt	1
1221	ITMQDQ-440	liber medicinalis	unknown	350	poetry	lat_0350.0_MQDQ-440.txt	1

Table. 1 Medical texts in LatinISE corpus (McGillivray and Kilgarriff, 2013)

2) I identified the list of 25 words of interest for the analysis. The list includes *patella* 'small bowl' then 'knee-cap'; *lenticula* 'lentil' then 'freckle'; *pupilla* 'young girl, ward' then 'pupil of the eye'. The words were taken from Langslow (2000), an extensive and (as far as I know) unique work on Latin medical lexicon. For each word I recorded its meaning in non specialised language and its meaning in medical contexts (see section 'List of words' in supporting document "Extensive report").
3) I prepared a Gold Standard (GS) for the evaluation of the embeddings (Supporting document "Gold Standard"). Based on previous work on legal Latin (Ribary ad McGillivray 2020), I created a GS for medical Latin parallel to the benchmark created by Sprugnoli et al. (2019). In order to define the synonyms for each target word I used the Latin dictionaries of synonyms used in Sprugnoli et al. (2020). Medical words are rare in the Latin vocabulary, so sometimes they are not recorded in the dictionaries of synonyms. In these cases, I used the Thesaurus Linguae Latinae (ThLL 1900–), a monolingual Latin dictionary.
4) I adapted an existing code from McGillivray and Nowak (2022) to the use-case on medical Latin. I split the corpus into the sub-corpora for medical vs. non medical texts as indicated in point 1. Then I performed some preliminary analyses on the corpus:
    i) distribution of texts per year
    ii) distribution of texts by genre (medical vs. non medical).
5) I trained word embeddings on the entire corpus using fastText with the method cbow. I performed some tests to find the best combination of parameters for the type of data used for this use-case, considering the reduced size of the sub-corpus for medical texts (only 5 texts). See next section for the results.
6) I trained the algorithm on the two sub-corpora (medical and non-medical) and I obtained the 10 closest neighbours for each word in the wordlist.
7) I performed a qualitative evaluation of the closest neighbours, using the Gold Standard created in step 3.

References:

Langslow, D. R. (2000). *Medical Latin in the Roman Empire*. Oxford: Oxford University Press.

McGillivray, B. & Kilgarriff, A. (2013). Tools for historical corpus research, and a corpus of Latin, in: P. Bennett, M. Durrell, S. Scheible, R. J. Whitt (Eds.), *New methods in Historical Corpus Linguistics*. Narr,

Tübingen, 2013.

McGillivray, B. & Nowak, K. (2022). Tracing the semantic change of socio-political terms from Classical to early Medieval Latin with computational methods. In *Latin vulgaire – latin tardif XIV. 14th International Colloquium on Late and Vulgar Latin. September 5-9, 2022, Ghent University.* Book of Abstracts. Ghent University. https://www.lvlt14.ugent.be/wp-content/uploads/2022/09/LVLT14-Book-of-abstracts.pdf (last accessed date: 31/01/2023).

Ribary, Marton, and Barbara McGillivray. (2020). A corpus approach to Roman law based on Justinian's digest. *Informatics*. Vol. 7. No. 4. MDPI.

Rodda, Martina, Philomen Probert, and Barbara McGillivray. (2019). "Vector space models of Ancient Greek word meaning, and a case study on Homer." *Traitement Automatique Des Langues* 60.3

Sprugnoli, Rachele, Marco Passarotti, and Giovanni Moretti. (2019). Vir is to Moderatus as Mulier is to Intemperans-Lemma Embeddings for Latin. *CLiC-it*.

Sprugnoli, Rachele, Giovanni Moretti, and Marco Passarotti. (2020). Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas. *IJCoL. Italian Journal of Computational Linguistics* 6.6-1: 29-45.

ThLL = Thesaurusbüro München Internationale Thesaurus-Kommission (Ed.) (1900–). *Thesaurus Linguae Latinae*. Berlin: De Gruyter.

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)

Grantee enters max 500 word summary here.

The main achievements of this STSM can contribute to the activities of Task 4 of the Use Case on Humanities within Working Group 4 of the Action (Lexical Semantic Change Detection). The code and the GS can be found at https://github.com/paoma370/Semantic-change-medical-Latin

1) I provided an overview of previous applications of word embedding to historical languages i.e., Latin and Ancient Greek, which could complement the state of the art for UC 4.2.1.
2) I prepared a GS specific for medical Latin lexicon, which can be used for the evaluation of word embedding algorithms on this specific use-case. The methodology can be replicated to create other GS on different use-cases or other historical languages e.g. Ancient Greek. The Gold Standard contains 25 lexical items which have been described in Langslow (2000) as words that have specialised their meaning in the medical domain through various types of semantic change (e.g. metaphor *patella* from 'small bowl' to 'knee-cap'). Some words have been excluded from the sample because the process of semantic specialisation is sometimes too subtle to be captured by word embeddings. This is the case e.g. for *album* and *nigrum*, which indicate respectively the colours 'white' and 'black', but then specialise in medical context to indicate the sclera (white) and the iris/pupil (black) in the human eye. The structure of the Gold Standard follows Ribary and McGillivray (2020): the lemma of the Latin medical word; a direct synonym of the lemma; two words that are not semantically related to the target word and are randomly selected from the benchmark in Sprugnoli et al. 2020. The GS can also be used to perform quantitative analyses on this use-case e.g. computing cosine similarity between the embeddings trained on the two subcorpora.
3) The tests on the LatinISE corpus allowed me to find the best combination of parameters to obtain the most accurate results possible. The tests were run both on the entire corpus and on the two subcorpora (medical vs. non-medical). The best combination of parameters is minimal frequency 5 (min_count = 5)–the target words are rare and the corpus is small– with subwords turned off (max_n = 0, min_n = 0) to avoid having orthographically similar words among the results.

4) I performed a qualitative analysis on the list of 25 words (for the specific results on each word see Supporting document "Extensive report"). The synonyms listed in the GS almost never appear among the closest neighbours in the medical subcorpus (only one case out of 25). The reason is that the medical subcorpus is extremely small, and medical words are rare. In some cases, however, the neighbours suggest a change of meaning, e.g. *malum* 'a bad thing' has among the closest neighbours in the medical subcorpus *sanesco* 'to recover', which points towards the semantic shift of *malum* towards the notions of 'disease'.

Outputs: I have submitted an abstract with Barbara McGillivray to Digital Classicist summer seminar series titled "Lexical semantic change detection in Latin: a use-case on medical Latin". The abstract was accepted for a presentation and the program is available at https://www.digitalclassicist.org/wip/wip2023.html

Future outcomes: using the Gold Standard I will be able to expand this work with a quantitative evaluation of the algorithm, calculating cosine similarity for the embeddings trained on the two sub corpora.

Plans for future follow-up collaborations: during the STSM Dr. Khan visited the King's College and Dr. Khan, Dr. McGillivray and were able to start discussing a future collaboration on a proposal for modelling lexical semantic change in ontologies, developing work that has already been carried out in the context of Nexus Use Case 4.2.1. The work will build on a paper already drafted by dr. Khan on ontologies and lexical semantic change, which would however benefit from concrete examples of lexical semantic change. The cases presented in this case study, together with others examples from languages other than Latin, could complement dr. Khan's work in the future. This would contribute to the COST Action activities, in particular to Task 5 of UC4.2.1 (Ontological Constructs for Concept Evolution).