CLARIN Annual Conference Proceedings

2023

Edited by

Krister Lindén, Jyrki Niemi, and Thalassia Kontino

16 – 18 October 2023
Leuven, Belgium

Please cite as:
CLARIN Annual Conference Proceedings, 2023. ISSN 2773-2177 (online).
Eds. Krister Lindén, Jyrki Niemi, and Thalassia Kontino.
Leuven, Belgium, 2023.

Programme Committee

Chair:

- Krister Lindén, University of Helsinki (FI)

Members:

- Starkaður Barkarson, Árni Magnússon Institute for Icelandic Studies (IS)
- Lars Borin, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Tomaž Erjavec, Jožef Stefan Institute (SI)
- Cristina Grisot, University of Zurich (CH)
- Eva Hajičová, Charles University Prague (CZ)
- Monica Monachini, Institute of Computational Linguistics "A. Zampolli" (IT)
- Karlheinz Mörth, Austrian Academy of Sciences (AT)
- Costanza Navarretta, University of Copenhagen (DK)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- Gijsbert Rutten, Leiden University (NL)
- Kiril Simov, IICT, Bulgarian Academy of Sciences (BG)
- Inguna Skadiņa, University of Latvia (LV)
- Koenraad De Smedt, University of Bergen (NO)
- Marko Tadič, University of Zagreb (HR)
- Jurgita Vaičenonienė, Vytautas Magnus University (LT)
- Vincent Vandeghinste, Instituut voor de Nederlandse Taal (Dutch Language Institute), the Netherlands & KU Leuven (BE)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Joshua Wilbur, Univerity of Tartu (EE)
- Andreas Witt, University of Mannheim (DE)
- Friedel Wolff, South African Centre for Digital Language Resources, North-West University (ZA)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Starkaður Barkarson, IS
- Lars Borin, SE
- António Branco, PT
- Tomaž Erjavec, SI
- Cristina Grisot, CH
- Eva Hajičová, CZ
- Krister Lindén, FI
- Monica Monachini, IT
- Karlheinz Mörth, AT
- Costanza Navarretta, DK
- Maciej Piasecki, PL
- Stelios Piperidis, GR
- Gijsbert Rutten, NL
- Kiril Simov, BG
- Inguna Skadiņa, LV
- Koenraad De Smedt, NO
- Marko Tadić, HR
- Jurgita Vaičenonienė, LT
- Vincent Vandeghinste, BE
- Tamás Váradi, HU
- Joshua Wilbur, EE
- Andreas Witt, DE
- Friedel Wolff, ZA
- Martin Wynne, UK

Subreviewers:

- Ilze Auzina, LV
- Federico Boschetti, IT
- Riccardo Del Gratta, IT
- Amelie Dorn, AT
- Maria Gavriilidou, GR
- Luís Gomes, PT
- Marissa Griesel, ZA
- Kinga Jelencsik-Mátyus, HU
- Mateja Jemec Tomazin, SI
- Fahad Khan, IT
- Penny Labropoulou, GR
- László János Laki, HU
- Kristine Levane-Petrova, LV
- Noémi Ligeti-Nagy, HU
- Amália Mendes, PT
- Hannes Pirker, AT
- Valeria Quochi, IT
- João Rodrigues, PT
- Rodrigo Santos, PT
- João Silva, PT
- Juan Steyn, ZA
- Benito Trollip, ZA
- Menno van Zaanen, ZA

CLARIN 2023 submissions, review process and acceptance

- Call for abstracts: 23 January 2023 first call published on CLARIN website, disseminated, and submission system open
- Submission deadline: 28 April 2023
- In total 52 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 9 June 2023
- Notifications to authors: 30 June 2023
- 37 accepted submissions

More details on the paper selection procedure and the conference can be found at https://www.clarin.eu/event/2023/clarin-annual-conference-2023.

DBBErt: Part-of-Speech Tagging of Pre-Modern Greek Text

Colin Swaelens and **Els Lefever**
Language Technology & Translation Team
Ghent University, Belgium
firstname.lastname@ugent.be

Ilse De Vos
Department of Linguistics
Ghent University, Belgium
i.devos@ugent.be

Abstract

This contribution presents DBBErt, a machine-learning approach to linguistic annotation for pre-Modern Greek, which provides a part-of-speech and fine-grained morphological analysis of Greek tokens. To this end, transformer-based language models were built on both pre-Modern and Modern Greek text and further fine-tuned on annotated treebanks. The experimental results look very promising on a gold standard of Byzantine book epigrams, with an F-score of 83% for coarse-grained part-of-speech-tagging and of 69% for fine-grained morphological analysis. The resulting pipeline and models will be added to the CLARIN infrastructure to stimulate further research in NLP for Ancient and Medieval Greek.

1 Introduction

The Database of Byzantine Book Epigrams or DBBE (Ricceri et al., 2023) contains over 12,000 epigrams. They are stored both as *occurrences* – the epigrams exactly as they occur in the manuscripts – and as *types* – their orthographically normalised counterparts. The relationship between occurrences and types is not one-to-one. For instance, Example 1 (DBBE Type 2148, translated by the authors) represents 70 two-verse occurrences of the ὥσπερ ξένοι epigram which was used widely by scribes to mark their joy of reaching the end of the manuscript and thus of their copying task.

The decision to link multiple occurrences to a single type was both pragmatic and conceptual. Creating fewer types not only freed up time to trace new occurrences, it was also a straightforward way to group similar occurrences. Soon however, this all-or-nothing system ran against its limitations: What exactly does "similar" mean? How "similar" do occurrences need to be for them to be put under the same type? The ὥσπερ ξένοι epigram for example circulated in many different versions, some counting three or four verses. To deal with this variety, increasingly more types were created, each of them covering different subsets of occurrences.

To (re)connect these subsets, a complementary system was introduced to link individual verses regardless of the type their occurrence belongs to. As for the ὥσπερ ξένοι epigram, no less than 202 instances of its first verse are to be found in DBBE. Although a huge step forward, this system still treats similarity as a dichotomy whereas it clearly is a continuum. Also, it does not allow to adequately visualise variation within more complex lists of "similar" verses nor to take into account different parameters, both textual and other. In order to add linguistic information enabling more advanced similarity detection and visualisation, we developed the first annotation tool for non-normalised Byzantine Greek.

(1) Ὥσπερ ξένοι χαίρουσιν ἰδεῖν πατρίδα,
οὕτως καὶ οἱ γράφοντες βιβλίου τέλος.
Just as strangers rejoice upon seeing their homeland,
so do writers upon completion of a book.

2 Related research

The last two decades witnessed a number of initiatives for compiling pre-Modern Greek corpora and making them accessible in open-source format, which stimulated NLP research on Ancient and Medieval Greek texts. An important corpus initiative is the Open Greek and Latin Project,[1] which consists of the Perseus Digital Library (G. R. Crane, 2022), a collection of more than 13,5M tokens of mostly classical Greek prose and poetry, and the First1K Project, containing 25,5M tokens of classical and post-classical Greek prose and poetry.[2]

In addition, various treebanks were developed. The Ancient Greek Dependency Treebank (AGDT) (Bamman & Crane, 2011) stores 560,000 tokens from both classical prose and poetry, all of which were tagged manually. The PROIEL (Haug & Jøhndal, 2008) treebank has a more specific content, as it stores 277,000 tokens of the New Testament in Greek and four other languages. The Gorman treebank (Gorman, 2020) on the other hand contains around 550,000 tokens of exclusively classical Greek prose. Finally, the Pedalion Trees (Keersmaekers et al., 2019) count around 320,000 tokens of annotated texts complementary to the AGDT, among which Trismegistos (Depauw & Gheldof, 2014), a database of papyri displaying the original text with all its idiosyncrasies including *errors*, just like the DBBE occurrences. All of the above mentioned treebanks are annotated in accordance with the Universal Dependencies principles and guidelines (Nivre et al., 2017).

A widely known and used tool for automatic linguistic annotation of Ancient Greek is Morpheus (G. Crane, 1991), a rule- and dictionary-based system that performs part-of-speech tagging and morphological analysis. It has, however, two important shortcomings: (1) it does not disambiguate ambiguous forms, but instead returns all possible analyses, and (2) it cannot analyse out-of-vocabulary words. In order to cope with this lack of flexibility, researchers have started to develop machine-learning systems for Greek part-of-speech tagging. Celano et al. (2016) did a comparative study, which showed that MateTagger (Bohnet & Nivre, 2012) outperformed Hunpos tagger (Halácsy et al., 2007), RFTagger (Schmid & Laws, 2008), the OpenNLP part-of-speech tagger[3] and NLTK Unigram tagger (Bird, 2006) on Ancient, normalised Greek data. Keersmaekers (2019), however, obtained different results when applying various taggers on papyrological data, with RFTagger clearly outperforming the other taggers on this specific data type. More recent approaches rely on neural networks, such as RNN tagger Schmid (2019), and the transformer-based part-of-speech tagger developed by Singh et al. (2021), which showed very promising results on an evaluation set containing normalised DBBE types.

3 Part-of-Speech Tagger

To develop a part-of-speech tagger for Ancient and Byzantine Greek, we compared three different transformer-based language models with embedding representations: BERT (Devlin et al., 2018), ELECTRA (Clark et al., 2020), and RoBERTa (Liu et al., 2019). These were then fine-tuned on the task of both coarse-grained part-of-speech tagging and fine-grained morphological analysis. To train these models, two data sets were compiled: one consisting of all Ancient and Byzantine Greek text corpora described in Section 2, and that same set complemented with the Modern Greek Wikipedia data. This allowed us to ascertain whether or not Modern Greek contributes to the modelling of Byzantine Greek.

For the supervised task of part-of-speech tagging and morphological analysis, we compiled a training set based on the treebanks described in Section 2 and completed it with a small set of 2,000 manually annotated tokens from DBBE occurrences. To train the part-of-speech tagger, we made use of the FLAIR framework (Akbik et al., 2019), where the contextual token embeddings from the language models are stacked with randomly initialised character embeddings. These are processed by a bi-directional long short-term memory encoder (hidden size of 256) and a

[1] https://opengreekandlatin.org
[2] https://opengreekandlatin.github.io/First1KGreek/
[3] https://opennlp.apache.org

conditional random field decoder. For evaluation, a gold standard containing 10,000 tokens of non-normalised Byzantine Greek epigrams out of the DBBE corpus was compiled, manually annotated and validated through an inter-annotator agreement study.

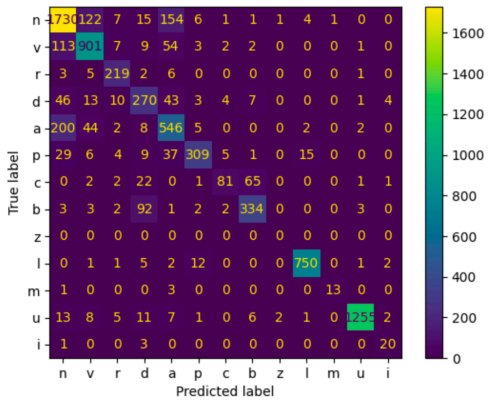

Figure 1: Confusion matrix of the coarse-grained part-of-speech labels

The BERT model trained on Classical, Medieval and Modern Greek performs best on this task with an F-score of 83% on the coarse-grained part-of-speech tagging and 69% on the morphological analysis. This model is hence called DBBErt.[4] Figure 1 shows the confusion matrix of the coarse-grained part-of-speech tagging, which reveals some expected trends (e.g., lot of confusion of the label noun (n) with the label adjective (a)).

4 Conclusion and Future Research

This paper introduced DBBErt, a transformer-based Part-of-Speech tagger for Ancient and Byzantine Greek. The evaluation of the tool on a novel gold standard containing occurrences of Byzantine Epigrams showed very promising results. In future research, we will keep improving DBBErt, since we believe that automatic linguistic annotation of non-normalised text will be very valuable for NLP research on historic languages. With respect to the DBBE, the enriching of the epigrams with linguistic information will supply important additional information for the next step in our research: measuring similarity between lexical variants of words in order to detect relationships between verses in various occurrences.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59.

Bamman, D., & Crane, G. (2011). The ancient greek and latin dependency treebanks. In C. Sporleder, A. van den Bosch, & K. Zervanou (Eds.), *Language technology for cultural heritage* (pp. 79–98). Springer Berlin Heidelberg.

Bird, S. (2006). Nltk: The natural language toolkit. *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 69–72.

Bohnet, B., & Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1455–1465.

[4]The model is made available at https://huggingface.co/colinswaelens

Celano, G. G. A., Crane, G., & Majidi, S. (2016). Part of speech tagging for ancient greek. *Open Linguistics*, *2*(1).

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Crane, G. (1991). Generating and Parsing Classical Greek. *Literary and Linguistic Computing*, *6*(4), 243–245.

Crane, G. R. (2022). Perseus digital library [Last accessed 14 October 2022].

Depauw, M., & Gheldof, T. (2014). Trismegistos: An interdisciplinary platform for ancient world texts and related information. *Theory and Practice of Digital Libraries – TPDL 2013 Selected Workshops*, 40–52.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gorman, V. B. (2020). Dependency treebanks of ancient greek prose. *Journal of Open Humanities Data*, *6*(1).

Halácsy, P., Kornai, A., & Oravecz, C. (2007). Hunpos: An open source trigram tagger. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 209–212.

Haug, D. T., & Jøhndal, M. (2008). Creating a parallel treebank of the old indo-european bible translations. *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, 27–34.

Keersmaekers, A. (2019). Creating a richly annotated corpus of papyrological Greek: The possibilities of natural language processing approaches to a highly inflected historical language. *Digital Scholarship in the Humanities*, *35*(1), 67–82.

Keersmaekers, A., Mercelis, W., Swaelens, C., & Van Hal, T. (2019). Creating, enriching and valorizing treebanks of Ancient Greek. *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, 109–117.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nivre, J., Zeman, D., Ginter, F., & Tyers, F. (2017). Universal Dependencies. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*.

Ricceri, R., Bentein, K., Bernard, F., Bronselaer, A., De Paermentier, E., De Potter, P., De Tré, G., De Vos, I., Deforche, M., Demoen, K., Lefever, E., Rouckhout, A.-S., & Swaelens, C. (2023). *The database of byzantine book epigrams project: Principles, challenges, opportunities* [working paper or preprint].

Schmid, H. (2019). Deep learning-based morphological taggers and lemmatizers for annotating historical texts. *Proceedings of the 3rd international conference on digital access to textual cultural heritage*, 133–137.

Schmid, H., & Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, 777–784.

Singh, P., Rutten, G., & Lefever, E. (2021). A pilot study for bert language modelling and morphological analysis for ancient and medieval greek. *The 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, co-located with EMNLP 2021*, 128–137.