Applied Deep Learning: Project introduction

Benedikt Baumgartner (01549291)

October 2021

1 Description of the project and data set

Tocharian A and B were languages spoken in Northwestern China, present-day Xinjian. Although one might assume that these languages are Sino-Tibetan or Turkic due to their geographical location, these languages are Indo-European and form a separate branch from the others. They were identified as indo-european only at the start of the 20th century by the german linguists Emil Sieg and Wilhelm Siegling. Due to the geopgraphical spread of the language, directly on the Silk Road, the speakers have not remained isolated from the cultural developments of the surrounding area. Especially the rise of Buddhism had a big influence of the literary production of Tocharian speakers, as we can see in the preserved documents of buddhist monasteries. But we can also find commercial documents and medical texts, indicating a highly developed literacy culture. As these languages were living non-standardized languages the documents feature variants, which we have to explain and can help us reconstruct the indo-european urlanguage. Since some variants stem from different interlanguage dialects, spoken by different persons, we would like to identify the writers of each text document.

To help the Tocharian linguist community, my Project Idea would be to use *Deep Learning* to identify different authors in the Tocharian Corpus. When I talked to Professor Hannes A. Fellner from the Indo-European Department of the University of Vienna, where I study as well, he told me that the human eye is only tuned for this task up to a certain degree and that for some corpus languages *Deep Learning* were already successfully used for this task. Before I can start the *Deep Learning*-approach I must get the data before I can preprocess it. Unfortunately I can't be more specific as I don't know yet, how the data will look like. But you can already take a look at Tocharian A and B manuscripts at `https://www.univie.ac.at/tocharian/?manuscripts`

2 Time Table

Task	Time estimate (in hours)
Getting access to the Tocharian data base	3
Familiarizing myself with the data base	2
Getting the data into machine readable form	20
Reading and understanding research papers	5
implementing and fine-tuning model	30

References

[1] Hannes A. Fellner. "The Expeditions to Tocharistan". In: Jan. 2007, pp. 13–36. ISBN: 978-3825352998.

[2] Malihe Javidi and Mahdi Jampour. "A deep learning framework for text-independent writer identification". In: *Engineering Applications of Artificial Intelligence* 95 (2020), p. 103912. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2020.103912. URL: https://www.sciencedirect.com/science/article/pii/S0952197620302463.

[3] Derek S Prijatelj et al. "Handwriting Recognition with Novelty". In: *arXiv preprint arXiv:2105.06582* (2021).

[4] Mohammad Abuzar Shaikh et al. "Attention based Writer Independent Handwriting Verification". In: *arXiv preprint arXiv:2009.04532* (2020).

[5] S. TRACY et al. "Identifying hands on ancient Athenian inscriptions: First steps towards a digital approach". In: *Archaeometry* 49 (Nov. 2007), pp. 749–764. DOI: 10.1111/j.1475-4754.2007.00333.x.