Pre-Modern Data: Applying Language Modeling and Named Entity Recognition on Criminal Records in the City of Bern

Hodel, Tobias
tobias.hodel@unibe.ch
University of Bern, Switzerland

Prada Ziegler, Ismail
ismail.prada@unibe.ch
University of Bern, Switzerland; University of Basel, Switzerland

Schneider, Christa
christa.schneider@unibe.ch
University of Bern, Switzerland; Paris Lodron Universität Salzburg, Austria

The end of the Middle Ages and the beginning of Early Modern times see an up-swing in the production and use of administrative documents in most parts of Europe. One typical example of frequently produced documents is trial records. In the Early Modern time, some criminals were brought for trial to the Tower of Bern (part of today's Switzerland), and a multitude of protocols for these trials was collected in the so-called tower books.

Even though these protocols are today available to the public, systematic research from a historical, linguistic, or jurisprudential perspective is still missing. Reasons for the absence of research may be due to the size of the corpus (approx. 300'000 pages, State Archive of Bern: B IX 423 - B XI 587) and certainly also because of the type of handwriting used in the documents - a form of German Kurrent - that cannot be described as very regular due to the circumstances of the writing process (often as part of an ongoing interrogation, in a tower, sometimes at night or in winter). Furthermore, the documents are not legible for modern readers.

From a scholarly point of view, the documents are of high interest from different perspectives, which makes the lack of research even more regrettable. There is only very little research on early forms of the written language (pre-standard) in Switzerland (Haas 2000: 109-138). Since these records also contain the usage of words originating in Alemannic dialects, we understand the collection of crucial to be able to understand parts of the language history in early modern Switzerland. Besides the linguistic issues, the collection also allows to address historical questions, covering specific cases such as witchcraft, persecution of homosexuals, etc., but also broader questions, like the geographic origin of the delinquents and their social background and circumstances or reasons for such prosecutions and its consequences to name but a few (Schwerhoff 2011).

In a first step, we applied Handwritten Text Recognition, utilizing large models that were built on several thousands of pages of similarly written documents (available on Transkribus: https://readcoop.eu/model/german-kurrent-16th-18th/, for the application of general models, see also (Hodel et al. 2021)). Parts of the automatically transcribed texts were corrected and tagged for named entities. Only through the combination of other (openly available, (Binz-Wohlhauser / Dorthe 2022)) material and in-house produced data from similar sources (minutes from Engelberg), a language model could be built that supported the training of a NER tagger (language model and tagger were built using the FLAIR framework (Akbik et al. 2019)).

Table 1. Results on the NER task, based on different types of data augmentation.

	Precision				Recall				F1			
Model	Micro-Avg	PER	LOC	ORG	Micro-Avg	PER	LOC	ORG	Micro-Avg	PER	LOC	ORG
Only sentences	80.07%	79.69%	90.41%	24.59%	83.50%	87.29%	83.83%	39.47%	81.75%	83.31%	86.99%	30.30%
Augmentation method 1	81.21%	82.46%	88.51%	28.81%	83.99%	88.51%	83.02%	44.74%	82.57%	85.38%	85.67%	35.05%
Augmentation method 2	81.29%	82.11%	89.97%	25.42%	82.89%	87.53%	82.21%	39.47%	82.08%	84.73%	85.92%	30.93%
Augmentation method 3	80.02%	79.96%	88.95%	31.75%	83.74%	87.78%	82.48%	52.63%	81.84%	83.68%	85.59%	39.60%
Augmentation method 4	80.51%	82.48%	87.10%	30.65%	81.78%	86.31%	80.05%	50.00%	81.14%	84.35%	83.43%	38.00%

The success of the tagger has not only been measured on a test set derived from the material, but also on uncorrected automatic transcription, recognized by the HTR. On corrected text we get state-of-the art results compared to similar datasets (Chastang / Aguilar / Tannier 2021; Torres Aguilar / Stutzmann 2021) with a F1-score of 0.8257 on the three categories "PERSON", "PLACE", "ORGANISATION". Since different forms of embeddings (fasttext, BERT, and from scratch language models) were tested, using different parameters (hidden size of embeddings, combination of input for language models etc.), we can report about strategies to build domain as well as time specific taggers.

Generally speaking, it is possible to apply NLP methodologies based on deep learning to pre-modern documents (for French, as an example, see (Gabay et al. 2022; Cafiero et al. 2021)). Similar to larger datasets containing modern languages, the hidden size of the neural network improves the results, be it only slightly. But the need to have large datasets for training of specific language models available remains. Only little attention has been brought so far to pre-modern documents, which in consequence led to a lack of available language models.

As the first results, from a linguistic as well as a historical perspective, we can assume that there was an "office language" in use that was intermingled with interferences of the historic Bernese Swiss German dialect. Furthermore, person and place names in the documents hint at lower classes that mainly were on trial, but with a sizable spatial background (especially from the North, Southern Germany and the Alsace), demonstrating the movements of esp. lower class people.

At this stage in particular, NLP for pre-modern documents still needs to be described as in its infancy and with a rather specialized data set, not building on any previous language models or taggers, we needed to rely on open research data that allowed for the enlargement of available training, validation, and test material.
The result of this paper is not only the experiences gained, but also another openly available dataset (partly tagged, currently in publication) and a published tagger (Prada 2022). By making aware of our results that clearly demonstrate that larger datasets result

in more precise models, we hope to further push other projects and institutions to publish their available data of pre-modern documents or even proper language models. Only through collaboration and communication of available data can improvements be made in the long run. At the same time, we strive for more communication about the re-usage of data sets to push from that perspective for the explainability of language models.

Bibliography

Akbik, Alan / **Bergmann, Tanja** / **Blythe, Duncan** / **Rasul, Kashif** / **Schweter, Stefan** / **Vollgraf, Roland** (2019). «FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP». In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4010.

Binz-Wohlhauser, Rita / **Dorthe, Lionel** (2022). *Freiburger Hexenprozesse 15.–18. Jahrhundert. Die Rechtsquellen des Kantons Freiburg. Erster Teil: Stadtrechte. Zweite Reihe: Das Recht der Stadt Freiburg.* Bd. 1. Teilband. Sammlung Schweizerischer Rechtsquellen. Basel: Schwabe. SSRQ FR I/2/8.

Cafiero, Florian / **Clérice, Thibault** / **Fièvre, Paul** / **Gabay, Simon** / **Camps, Jean-Baptiste** (2021). «Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre». *Journal of Data Mining & Digital Humanities* 2021 (Februar). https://doi.org/10.46298/jdmdh.6485.

Chastang, Pierre / **Torres Aguilar, Sergio** / **Tannier, Xavier** (2021). «A Named Entity Recognition Model for Medieval Latin Charters». *Digital Humanities Quarterly* 15 (4). http://www.digitalhumanities.org/dhq/vol/15/4/000574/000574.html.

Gabay, Simon / **Ortiz Suarez, Pedro** / **Bartz, Alexandre** / **Chagué, Alix** / **Bawden, Rachel** / **Gambette, Philippe** / **Sagot, Benoît** (2022). «From FreEM to D'AlemBERT». In *13th Language Resources and Evaluation Conference - LREC 2022*, European Language Resources Association, Jun 2022, Marseille, France. pp.3367-3374. https://hal.inria.fr/hal-03596653 .

Haas, Walter (2000). «Kurze Geschichte der deutschen Schriftsprache in der Schweiz». In: Bickel, Hans and Schläpfer, Robert (ed.): Die viersprachige Schweiz. Aarau, Frankfurt am Main, Salzburg, Sauerländer: 109-138.

Hodel, Tobias / **Schoch, David** / **Schneider, Christa** / **Purcell, Jake** (2021). «General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example». *Journal of Open Humanities Data*, Journal of Open Humanities Data, 7. https://doi.org/10.5334/johd.46.

Prada, Ismail (2022). «dh-unibe/turmbuecher-ner-v1». Huggingface. https://huggingface.co/dh-unibe/turmbuecher-ner-v1.

Schwerhoff, Gerd (2011). *Historische Kriminalitätsforschung*. 1. Aufl. Frankfurt am Main: Campus Verlag.

Torres Aguilar, Sergio / **Stutzmann, Dominique** (2021). «Named Entity Recognition for French medieval charters». In *Workshop on Natural Language Processing for Digital Humanities*. Workshop on Natural Language Processing for Digital Humanities Proceedings of the Workshop. Helsinki, Finland. https://hal.archives-ouvertes.fr/hal-03503055.