Named Entity Recognition in 19th Century Parish Court Records

Siim Orasmaa, Kadri Muischnek, Kristjan Poska, Anna Edela

University of Tartu
LREC 2022

19th Century Parish Court Records

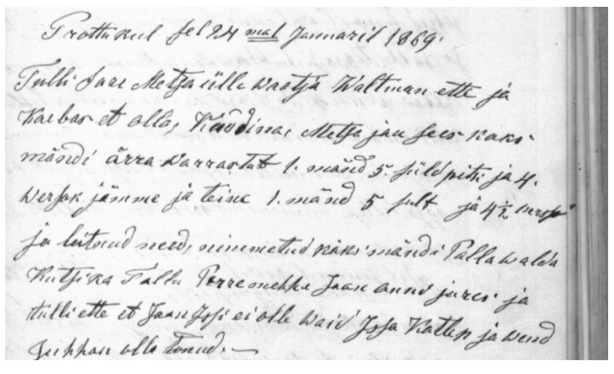
A historical resource documenting court cases of Estonian peasantry:
- which minor offences they were tried for
- how did they solve their civil disputes and family matters

Digitized minute books at the National Archives of Estonia

A crowdsourcing project:
- https://www.ra.ee/vallakohtud (in Estonian)
- manual transcription of digitized minute books

Interesting source of knowledge for historians:
- everyday lives of the peasantry
- social networks

Interesting for linguists:
- change of ortography during the 19th century:
  - ► transition from German style to Finnish style spelling
  - ► volatile writing conventions, inconsistent capitalization
- language variation and dialects
  - ► language standard not fully established:
  - ► big differences between northern and southern dialects



PROTOKOLL (minutes)

Leidandmed (reference code)	EAA.3089.1.3	
Kaader (frame number)	57	
Daatum (date)	24.01.1869	
Protokolli number		
Protokolli teema (topic)	Varavastased kuriteod	(crimes against property)
Märkused (comments)		

Tulli Sare Metsa ülle Waatja Waltman ette ja Kaibas et olla, Küddina Metsa jau sees Kaks mändi ärra Warrastat. 1. mänd 5. süld ja 4. wersok jämme ja teine 1. mänd 5 sült ja 4 1/2 wersok.Ja leitud need ‚nimmetud kaks mändi Palla Walda Kutsika Tallu Perremehhe Jaan

Anni jures ja tulli ette et Jaan Issi ei olle Waid Issa Kotlep ja wend Juhhan olla tonud. Noutis Metsa ülle Waatja Waltman nende kahhe Männi peält 6. Rubla hõbbe. Moistis Koggokonna Kohhos nende Kahe männi peält 3. Rubla hõbbe puu Trahwi ja Gotlep Anni omma pojaga Kumpakile 50. Kop Trahwi laega.

Figure 1,2: Example document
Source: https://www.ra.ee/vallakohtud/index.php/record/view?id=21335

Named entity annotation project

1,500 documents, randomly chosen, period: 1821 to 1920

Named entity types:
- **PER** – person names
- **LOC_ORG** – a placename referring to a certain location, and also to a group of people (community) connected with the location
  - ► a farmstead, a village, a parish
  - ► ≈ GPE (geopolitical entity)
- **LOC** – rest of the place names
- **ORG** – mostly court names
- **MISC** – artefacts (human-made entities), other (events), unknown

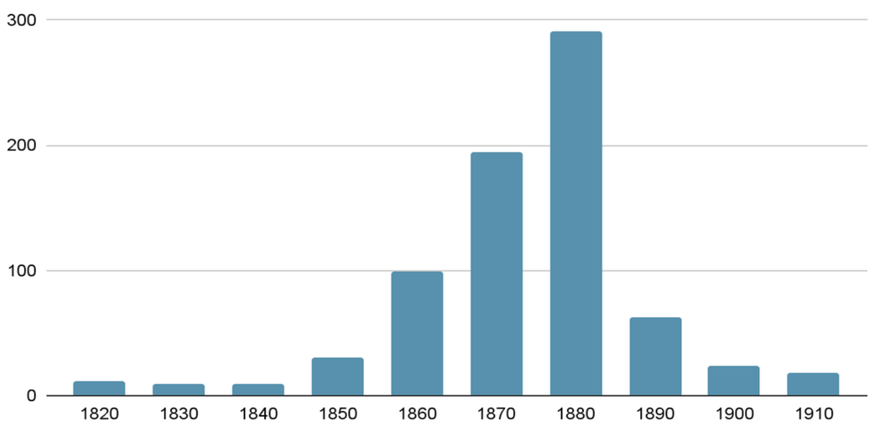


Figure 3: Temporal distribution of minute books (our corpus)

Manual annotation
- brat annotation tool (Stenetorp et al., 2012)
- 27,540 named entities:
  - ► **84% PER**; ► 9.9% LOC_ORG; ► 3.7% LOC;
  - ► 1.5% ORG ; ► <1% MISC

Inter-annotator agreement
- Mean pairwise F1-scores (Hripcsak and Rothschild, 2005)
- 2 annotators, 250 documents, mean F1-score: 95%
- Agreement with crowdsourcing annotators: 68%

Examples of named entities

PER: *Jaan Tamm, J.E. Treiblat, Peeter Kristjani p Peterson*
LOC_ORG: *Kalama tallo, Altra talu, Sepa päriskoht, Hintsu talukoha, Kiiwita kuha, Nigula koha, Willemi perre, Tire pere, Adraku krundi, Annuka asse* - all those are names of farmsteads, note the multiple synonyms and ortographic variants for ‘farmstead’
LOC: *Peipsi järwe* ‘Peipsi lake’, *Ihhasallo rannast* ‘Ihasalu shore’
ORG: *Sauga Koggokonna kohhus* ‘Sauga Community court’
MISC: *laewa “Eduard”* ‘ship “Eduard”’

NER experiments: preprocessing and setup

Convert brat-annotated documents to IOB2 representation:
- Tokenisation by EstNLTK library (Laur et al., 2020);
  - ► Tokenisation fixes, e.g. separate name and non-name tokens:
    ’talumeesNikolai’ → *’talumees Nikolai’* ’farmer Nikolai’
- Data split:

	training	development	test
# documents	1,125	125	250
# words	240,614	28,891	50,900
# named entities	20,944	2,357	4,239

Methods

- **Baseline (traditional machine learning):** A CRF-based model with manually designed feature extraction (Tkachenko et al. 2013);
- **EstBERT:** BERT model pretrained on a 1.1 billion word Estonian National Corpus 2017 (Tanvir et al., 2020);
- **WikiBERT-et:** BERT model pre-trained on 38 million words of the Estonian Wikipedia (Pyysalo et al., 2020);
- **Est-RoBERTa:** BERT-like model pre-trained on a 2.51 billion token corpus (mostly Estonian news).

Transfer learning settings

- HuggingFace Transformers library (Wolf et al., 2020) for fine-tuning BERT models;
- Hyperparameter selection: grid search over learning rates (5e-5, 3e-5, 1e-5), batch sizes (8, 16, 32) and 3 epochs;
- Best model fine-tuning until F1-score no longer improved on development set (10 epochs at maximum).

Results

Model results on the test set →

model	*precision*	*recall*	*F1-score*
CRF (baseline)	91.57	88.18	89.84
EstBERT	89.74	91.15	90.44
WikiBERT-et	91.29	91.98	91.63
EstRoBERTa	92.97	92.24	93.60

Category-wise F1-scores of the best model (finetuned EstRoBERTa) →

PER	*LOC_ORG*	*LOC*	*ORG*	*MISC*
96.31	81.2	65.98	95.36	73.56

Discussion of results

- NER performances close to the state of art in modern language. Possible reasons:
  - ► Noise-free texts due to manual transcription;
  - ► Overall regular textual structure of a Parish Court record;
- Very good performance on person names;
- Low performance on location names:
  - ► rareness and high variability of location names.

Acknowledgements

- This research has been supported by the Centre of Excellence in Estonian Studies (CEES, European Regional Development Fund) and by the national programme ”Estonian language and cultural memory” project EKKD29.
- The BERT models were fine-tuned on the UT Rocket cluster of the High-performance Computing Center at the University of Tartu (University of Tartu, 2018).

References

- Hripcsak, G. and Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. Journal of the American medical informatics association, 12(3): 296–298.
- Laur, S., Orasmaa, S., Särg, D., and Tammo, P. (2020). EstNLTK 1.6: Remastered Estonian NLP pipeline. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 7152–7160.
- Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for NLP-assisted text annotation. In Proceedings of the Demonstrations Session at EACL 2012, Avignon, France, April. Association for Computational Linguistics.
- Pilvik, M.-L., Muischnek, K., Jaanimae, G., Lindström, L., Lust, K., Orasmaa, S., and Türna, T. (2019). Moistus sai kuulotedu: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine [creating a digital resource from the 19th century parish court records]. Eesti Rakenduslingvistika Ühingu aastaraamat, 15:139–158.
- Pyysalo, S., Kanerva, J., Virtanen, A., and Ginter, F. (2020). Wikibert models: deep transfer learning for many languages. arXiv preprint arXiv:2006.01538.
- Tanvir, H., Kittask, C., Eiche, S., and Sirts, K. (2020). EstBERT: A Pretrained Language-Specific BERT for Estonian. arXiv preprint arXiv:2011.04784.
- Tkachenko, A., Petmanson, T., and Laur, S. (2013). Named entity recognition in Estonian. In Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pages 78–83.
- University of Tartu. (2018). UT Rocket. doi: 10.23673/PH6N-0144, https://share.neic.no/marketplace-public-offering/c8107e145e0d41f7a016b72825072287/ .
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference onEmpirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October. Association for Computational Linguistics.