Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5304–5313
Marseille, 20-25 June 2022

Named Entity Recognition in Estonian 19th Century Parish Court Records

Siim Orasmaa, Kadri Muischnek, Kristjan Poska, Anna Edela
Institute of Computer Science, University of Tartu, Narva mnt 18, 51009 Tartu, Estonia,
Institute of Estonian and General Linguistics, University of Tartu, Jakobi 2, 51005, Tartu, Estonia,
{siim.orasmaa, kadri.muischnek}@ut.ee ,
{kristjanposka, anna.edela}@gmail.com

Abstract

This paper presents a new historical language resource, a corpus of Estonian Parish Court records from the years 1821-1920, annotated for named entities (NE), and reports on named entity recognition (NER) experiments using this corpus. The hand-written records have been transcribed manually via a crowdsourcing project, so the transcripts are of high quality, but the variation of language and spelling is high in these documents due to dialectal variation and the fact that there was a considerable change in Estonian spelling conventions during the time of their writing. The typology of NEs for manual annotation includes 7 categories, but the inter-annotator agreement is as good as 95.0 (mean F1-score). We experimented with fine-tuning BERT-like transfer learning approaches for NER, and found modern Estonian BERT models highly applicable, despite the difficulty of the historical material. Our best model, finetuned Est-RoBERTa, achieved microaverage F1 score of 93.6, which is comparable to state-of-the-art NER performance on the contemporary Estonian.

Keywords: historical language processing, named entity recognition, digital humanities, corpus annotation

1. Introduction

In this paper we present a new resource – a corpus of 19th century Parish Court records annotated for named entities (NE) in Estonian – and report on experiments on named entity recognition (NER) using this dataset.

The corpus contains a subpart of the Estonian Parish Court records from the 1820s to the beginning of the 20th century, the majority of texts originate from the period 1860-1890. The original documents are handwritten and the digitization was done manually by volunteers. Although the digitized documents are not completely error-free, manual digitization gives better results than handwritten text recognition or even OCR. The texts of the Parish Court records are very heterogeneous in terms of spelling conventions, capitalization and dialectal variation, which can hamper both manual NE annotation and NER.

This corpus is a valuable source of information for historians, linguists interested in language history and the historical development of written language, and also for the public at large. Of course it is also a valuable resource for historical language processing.

Annotating it with NEs enables better search queries for the users, also various data analysis and visualization procedures. By developing a specific NER tool for these texts we make it possible to annotate the NEs in all Parish Court records that are being constantly digitized.

While NER for contemporary texts is an established task in information retrieval and information extraction, there has also been an increasing amount of work done for NER in historical documents during recent years. Ehrmann et al. (2021) published a thorough survey on the subject which perhaps exemplifies the rising need for research carried out on it.

The task of named entity recognition usually includes the classification of these entities into types or categories. The most simple set of categories includes *Person*, *Location* and *Organization*, but there are also more complex, multi-layered and/or domain-specific typologies.

The Parish Court records make up a text type in its own right: they document the arguments and agreements between people who are identified as representatives of certain communities that in turn are linked to or originate from certain locations. The arguments and agreements are often about ownership of certain objects. In order to capture information about these entities, we have developed a specifically tailored set of NE categories for our corpus, containing seven categories: *Person*, *Location*, *Organization*, *Location-Organization*, *Artefact*, *Other* and *Unknown*.

Finally, we present experimental NER results on the dataset, which involve finetuning BERT transfer learning models and comparing their performance to a traditional machine learning baseline. Our results show that despite the difficulty of the historical material, transfer learning can achieve NER performance comparable to the state-of-the-art on the modern language.

Thus our contribution can be summarized as follows: (1) We have created a new language resource - a corpus of Estonian Parish Court records, annotated for named entities. This resource has value for historians, linguists, digital humanists and researchers working with historical language processing. (2) We have developed a NE taxonomy specifically tailored for this text-type and show that it is possible to annotate texts manually into these NE categories with good inter-annotator

agreement. (3) We have shown that BERT models fine-tuned for NER on this historical material achieve performance levels comparable to the state-of-the-art on contemporary language.

2. Related work

In this Section we give a brief overview of NE typologies and annotation schemes for historical texts, NE-annotated historical corpora, and developing NER systems for historical texts. We also describe state-of-the-art in Estonian NER, as our experiments build upon this work.

Resources for historical NER. By resources for NER we mean NE typologies and NE-annotated corpora.

Quite a few NE typologies exist for NE annotation and NER on present-day documents, but in their thorough overview of NE-related resources and approaches to NER in historical documents, Ehrmann et al. (2021) admit that, to their knowledge, very few **typologies and guidelines designed for historical texts** were publicly released by the time of writing their overview. They list the Quaero, SoNAR and Impresso guidelines. All these three annotation schemes for historical corpora include the three "classical" NE categories of *Person*, *Location*, *Organization* and they all have also a category for entities created by humans.

Ehrmann et al. (2021) list 17 **NE-annotated historical corpora**. The texts originate mostly from 19th-20th centuries, but there are also two corpora from 3rd-5th centuries and from 1st century BC – 2nd century. The languages of these corpora include English, French, German, Italian, Czech, Portuguese, Finnish, Dutch, also Coptic and Latin. The texts come from the domains of newspapers, literary texts, but also specialized domains like medical literature or travelogues.

The annotated NE categories depend on the domain of the corpus, but the most common annotated NE categories are *Person*, *Location* and *Organization*, sometimes only *Person* and *Location*, e.g. in the corpus of Finnish newspapers (Ruokolainen and Kettunen, 2018).

As an example of more complex NE taxonomy for general domain, one could mention the corpus of Czech newspapers from the year 1872 (Hubková et al., 2020) or the NewsEye (Hamdi et al., 2021) dataset.

The NEs in the Czech newspaper corpus were annotated manually, using five NE categories: *Person, Institution*, *Geographical name*, *Time expression*, *Artefact name/Object* and *Ambiguous name*.

The NewsEye dataset is multilingual, consisting of newspaper articles in French, German, Finnish and Swedish from the middle of the 19th century up to the middle of the 20th century. The annotated NE categories include *Person*, *Location*, *Organization* and *Human Product*.

So the annotation schemes of those two corpora include a category corresponding to our category *Artefact*. However, to our knowledge no historical corpus is annotated for the NE category *Location-Organization* used in our corpus. A similar category - *Geo-Political Entity* (GPE) - has been used while annotating Norwegian treebank NorNe (Jørgensen et al., 2020). GPE stands for complex entities that can refer both to a location and an organization or simply a group of people associated with that location. Names of states and cities are typical examples of GPEs.

Automatic NER on historical texts. In their overview, Ehrmann et al. (2021) describe a variety of rule-based, traditional machine learning and deep learning systems proposed for NER in historical documents, and they observe that the rule-based and traditional machine learning system performance F1 scores range from 60% to 70% on average, and the best neural systems exceed 80%. But systems' performances largely depend on characteristics of historical documents, e.g. digitisation errors, language dynamics in the collection, and document domain.

Many previous NER efforts have focused on historical newspapers, which have been made machine-readable via massive digitisation.

The HIPE-2020 shared task addressed named entity recognition in ca. 200 years worth of OCRed historical newspapers in French, English and German. Median F1 scores of the participating systems ranged from 46% to 67% across languages, with the highest reported F1 score of 84% (Ehrmann et al., 2020). In general, neural systems prevailed the campaign, and most of the best systems were using or incorporating BERT (Bidirectional Encoder Representations from Transformers) models (Devlin et al., 2018).

While NER results on historical OCR-ed documents can be moderate or even low, there are also some projects reporting results close to the modern state-of-the-art on historical material.

Aguilar et al. (2016) trained a Conditional Random Field (CRF) based NER model for recognising person and location names in manually transcribed Latin medieval charters (from 10th to 13th century), and reported F1-score performance levels around 90%. It is likely that formal nature of these documents and the quality of manual transcription supported the high-quality NER in that collection.

Swaileh et al. (2020) addressed NER task on OCRed French and German financial yearbooks from the beginning of and mid of 20th century. Their best model was a hybrid BiLSTM-CRF (Huang et al., 2015) model, which employed contextual character-level embeddings (Akbik et al., 2018) (pretrained on modern language), and achieved average F1-scores ranging from 86% to 96%. The regularity of the language used in financial documents and good quality OCR results likely contributed to the high performance of their model.

Estonian NER. Recent research on modern Estonian NER has also seen a rise of deep neural net-

works, and transfer learning based approaches. Kittask et al. (2020) show that multilingual BERT models finetuned for NER outperform the traditional CRF approach (Tkachenko et al., 2013) on the benchmark corpus of modern newspapers. Tanvir et al. (2020) report that monolingual Estonian BERT models (EstBERT and WikiBERT-et) outperform multilingual ones and F1 scores rise over 90% with the monolingual models. Ulčar et al. (2021) compare non-contextual (fastText) and contextual embedding methods (ELMO and BERT-like models) on Estonian NER, and report BERT-like models outperforming other apporaches. In particular, they report the new highest result for Estonian NER in news domain – F1 score of 93.6% – obtained by fine-tuning the monolingual Est-RoBERTa model for the task. These results also motivate us to focus on BERT-like models in our experiments.

3. Annotation project and data

This Section gives an overview of our corpus of Parish Court records, describes the creation process of the corpus and outlines its specific linguistic and spelling-related features. Then, in Section 3.2 we present the annotation project – the NE categories to be annotated and the annotation procedure. In order to find out the quality of the annotation, we have also measured inter-annotator agreement.

3.1. Characteristics of the Corpus

The corpus annotated for named entities is a subpart of the Parish Court records from 1821 to 1920 digitized via a crowdsourcing project[1] at the Estonian National Archives. The corpus contains ca 320 000 words in 1500 documents.

The Parish Courts, also called Community Courts at that time, tried peasants for their minor offenses and solved their civil disputes, claims, and family matters. They also registered agreements and wills. (Traat, 1980)

So the court records are a rich historical resource shedding light on the everyday lives of the peasantry and the common sources of quarrels and disputes. Annotating named entities in these texts helps the researchers to find documents containing information about the same subjects or similar events. Annotations are also prerequisite for linking named entities across different documents, and for linking names to other historical knowledge sources, such as parish registers or church records.

It is worth noting that the volunteers who manually digitize the records can choose freely the manuscripts they want to digitize from a large pool of Parish Court records. So it might be the case that noisier documents are not chosen for digitization.

The volunteers are supposed to maintain the original spelling and capitalization in digitizations. They are also encouraged to annotate names of persons and locations in the documents, but this annotation is quite inconsistent, meaning that some volunteers annotate these named entities and some do not and some annotate only the first mentions of named entities (cf Inter-annotator agreement section below).

For every court record, metadata includes the time of writing the document and the name of the parish, which enables us to group the texts according to the dialects spoken in those parishes.

For NE-annotated corpus the digitized texts were chosen randomly, but maintaining the proportion of distribution of court records between the parishes.

Linguistic and spelling-related variation The level of linguistic and spelling-related variation is high in these texts. This variation is influenced by three factors.

First, Estonian language at that time was divided into local dialects and the lexicon and especially the inflectional properties of the words differ from dialect to dialect. Most prominent differences are between North and South dialect groups, and during 17th-18th centuries there were two separate written languages – that of North Estonian and that of South Estonian. Throughout the 19th century the usage of written South Estonian receded and by the end of the century written North Estonian had gained the status of nation-wide official language (Raag, 2008), pp. 28-53.

During the period of writing the Parish Court records the two written language variants existed side-by-side also in South Estonia (Raag, 2008), pp 57-59. What complicates the situation is, that in some regions of South Estonia, the written South Estonian was used also in schools, whereas in other parts of South Estonia North Estonian written language was used as the schooling language (Raag, 2008), p 41.

Both in North and South Estonia, there existed also dialectal differences between parishes that are apparent in the texts of different Parish Court records (Pilvik et al., 2019).

A second factor contributing to the variation is spelling reform. In the beginning of the 19th century, both Written South Estonian and Written North Estonian followed the German spelling conventions, which could not adequately represent Estonian pronunciation. In 1843 Eduard Ahrens (Ahrens, 1843) proposed a new spelling similar to that of Finnish, that belongs to the same, Finnic group of the Finno-Ugric language family. The so-called new spelling became dominant in published texts during 1870-1880. (Raag, 2008), pp 57-60, but in the Parish Court records the two spelling conventions are both present, sometimes even in the same text, which means that the writer was not sure what the correct spelling was or did not care much about it.

And the third factor contributing to the linguistic and spelling variation of the texts is the dialectal background and the education of the clerks. There is lit-

[1] https://www.ra.ee/vallakohtud/ (2022-01-04)

tle information about them and it is apparent from the texts that their writing experience was quite different – some court records are really fluent written language, whereas others are rather clumsy.

The capitalization conventions also vary from text to text. The general rule at that time was that proper nouns should be capitalized and usually they are, but often capitalization is used randomly: most of the proper nouns are capitalized, but less so in the earlier documents; common nouns are usually not capitalized, but again especially in the earlier documents capitalization is used randomly as exemplified in (1)

(1) *Tulli Sare Metsa ülle Waatja Waltman*
Came Sare Wood over Seer Waltman
ette ja Kaibas et olla, Kuddina
forward and Complained that were Kuddina
Metsa jau sees Kaks mändi ärra Warrastat
Wood part in Two pines away Stolen

'Sare wood overseer Waltman came forward and complained that two pines were stolen in the part of the wood called Kudina' (document 21335)

3.2. Corpus Annotation

For manual NE-annotation, 1500 Parish Court records were chosen randomly, maintaining the the proportion of distribution of court records between the parishes.

Annotation guidelines Seven categories of named entities were annotated in the Parish Court records: *Person* (tag: PER), *Location* (LOC), *Organization* (ORG), *Location-organization* (LOC-ORG), *Artefact*, *Other* and *Unknown*.

A **person** (PER) can be referred to using the full version of the name, which most often consists of one forename and a surname. The further mentions of the same person in the same document can use only the forename or only the surname. In certain parishes also the initial of the forename, followed by a full stop can be used instead of the first name, e.g. *A. Kalew, J.E. Treiblat*.

In some parishes and during certain periods of time, the full version of a person name consists of more than two parts, containing also the patronymic name. There are different patterns how the patronymic name is integrated into the full version of a person name, e.g. *Peeter Kristjani p Peterson* 'Peeter Kristjan's s Peterson', *Mihkel Tõnise poeg Reinberg* 'Mihkel Tõnis's son Reinberg' and, in parishes with Russian population, *Iwan Fedorow Poljakow* 'Iwan Fedor's son Poljakow'. However, if the person name with patronymic name follows the pattern 'Forname Surname Father's son/daughter', then these are annotated as two separate named entities: 'Forname Surname' and 'Father'.

Place-names are divided into two categories: **Locations** (LOC) and the place-names that can also refer to the people or organization connected with this place - **Location-Organization** (LOC-ORG).

By Location-Organization we mean a place-name that can be used for referring to a certain location, and also for referring to a group of people connected with this location. So our motivation for introducing the category LOC-ORG is to distinguish place-names that can be used for identifying people. LOC-ORG is similar to the category Geo-Political Entity (GPE), described in Section 2.

A person could be identified, in addition to his name, by his farmstead, village or parish or all those combined, e.g. (2)

(2) *Mihkel Rauba Sare küllän Rauba tallo*
Mihkel Rauba Sare village-in Rauba farmstead
perremees oma Tolama Wallast Rein
owner has Tolama Parish-from Rein
Otsingiga sedda möda lepno ...
Otsing-with that way agreed ...

'Mihkel Rauba, owner of Rauba farmstead in Sare village has made an agreement with Rein Otsing from Tolama parish ...' (document 1170)

While annotating place-related named entities LOC and LOC-ORG, the determinative part of the name (farmstead, village, parish, etc.), if present, is included in the annotated entity. So, in the previous example sentence 2, the LOC-ORG entities are *Sare küllän* 'in Sare village', *Rauba tallo* 'Rauba farmstead' and *Tolama Wallast* 'from Tolama parish'. Note the inconsistent capitalization of the determinative part of the name.

In the annotation guidelines we designated a set of location types that should be annotated as LOC-ORG, mainly names of farmsteads, villages, parishes and manors, plus other minor types of locations. However, recognizing LOC-ORG named entities was not so easy as it might seem because a lot of synonyms for the determinative word meaning farmstead were used in those texts, mainly due to dialectal variation and farmstead was by far the most frequent LOC-ORG type named entity present in those texts.

Estonian peasants got their surnames in the beginning of the 19th century, before that people were usually identified by the name of the farmsteads they lived at. When a person was officially given a surname, quite often it was the same name of the farmstead.

So, the same person could be called *Tamme Jaan*, meaning that his first name was *Jaan* and he came from *Tamme* ('Oak's') farmstead. When choosing himself a surname, he could probably choose *Tamm* 'Oak' or *Tamme* 'Oak's' as his last name and was then referred to in documents as *Jaan Tamm* or *Jaan Tamme*.

In our corpus of parish court records we differentiate between these two usages. In the first case, *Tamme Jaan* is annotated as two different named entities: *Tamme*, the farmstead name as LOC-ORG and *Jaan* as PER. However, *Jaan Tamm* is annotated as one named entity PER.

The category **Location** (LOC) includes mainly names of hills, rivers, islands and other landscape objects, but also names of countries and towns as these entities never act as organizations in those texts, they are mentioned only as places.
Organizations (ORG) in the Parish Court records include mostly names of courts, both parish courts and higher courts.
Names of **Artefacts** refer to entities made by humans. In our corpus, the names of ships, books and newspapers were frequent in this category.
Category **Other** includes names of events, mostly fairs, and names of laws or collections of laws.
Category **Unknown** was used for annotating tokens and phrases that definitely were named entities, but their category could not be established with certainty. Sometimes a place was mentioned in a court record, but without any clue for interpreting it as belonging to the category LOC-ORG or, alternatively, to the category LOC. And a few documents, mainly from the earliest period, contained capitalized tokens that could be interpreted as proper nouns, but the name type and even the meaning of the larger phrase or sentence remained somewhat unclear.
After developing the initial guidelines, a linguist familiar with older versions of Estonian annotated NEs in the texts using the brat rapid annotation tool (Stenetorp et al., 2012). Uncertain cases were discussed with another linguist and a historian. The guidelines were constantly improved and refined during the annotation process and the previous annotations were adjusted, if necessary.
Final statistics of the annotated corpus are depicted in Table 1.

NE category	Number of entities	Proportion
PER	23 126	84.0%
LOC-ORG	2 733	9.9%
LOC	1 008	3.7%
ORG	419	1.5%
MISC	254	0.9%
# total	27 540	100%

Table 1: Distribution of named entity categories in the corpus. Category MISC aggregates entities of the 3 least frequent categories: Other, Unknown and Artefact.

The named entities in our corpus are distributed unevenly between the categories. Person names (PER) are by far the most frequent category in these documents, making up 84% of all annotated named entities. 9.9% of the annotated named entities belong to the category LOC-ORG and 3.7% to the category LOC. Only 1.5% of the annotated named entities are names of organizations (ORG), and the remaining three categories - Artefacts, Other and Unknown - constitute less than 1% of all annotated names.

Inter-annotator agreements. A portion of manually annotated corpus – 250 documents – was re-annotated by another linguist in order to measure inter-annotator agreements. Agreements were calculated as mean pairwise F1-scores (Hripcsak and Rothschild, 2005), using the *bratiaa* tool[2] (Kolditz et al., 2019). We used strict ("instance-based") measure: two annotation instances were considered as matching only iff their start and end locations in document were exactly matching.
Two annotators following our guidelines obtained overall mean F1-score of 95.0. Inspecting results category-wise, we noted highest agreements on PER (F1=98.3) and ORG (F1=87.1) categories. The agreement on the category Other was also relatively high (F1=83.6), but most of the annotated entities in that category were abbreviations referring to a peasant law (such as *T.S.R = Tallorahwa seädusse ramato* (*Peasant Law Codex*)). The agreement on the LOC-ORG category (F1=81.6) was notably higher than the agreement on the LOC category (F1=74.2). Agreements were lowest on Unknown (F1=54.5) and Artefact (F1=20.0) categories, which is expected, as these categories cover problematic and rare entities.
We also compared annotations of our experts with the crowd-sourced annotations of the Estonian National Archives. The crowd-sourcing project uses only 2 named entity categories – Person and Location – so we reduced our annotations to these categories. We merged LOC-ORG and LOC categories into one and excluded annotations of ORG, Other, Unknown and Artefact categories from the evaluation. Overall, mean F1-score agreement between annotators following our guidelines and crowd-sourcing annotators was 0.68. This is much lower than agreement among linguist annotators, indicating difficulties on establishing consistent annotations via crowd-sourcing, and also outlining the need for automation of the annotation process.

4. Experiments

Our experimental work here builds upon the setup of machine learning experiments reported in Kristjan Poska's thesis (2021). While Poska experimented with traditional machine learning on the dataset, in this work, we focus on the deep learning, more specifically, on BERT-like transfer learning approaches[3], which have recently shown the best performance on Estonian named entity recognition (Tanvir et al., 2020; Ulčar et al., 2021).

4.1. Data preprocessing

For conducting NER experiments, we converted documents from the brat annotation format to word level annotations, following the IOB2 representation: words part of a named entity got label prefixes B (beginning

[2] https://github.com/kldtz/bratiaa (2021-12-20)

[3] Source code of our experiments: https://github.com/soras/vk_ner_lrec_2022 (2022-04-19)

of a named entity) and I (inside a named entity), and words not part of any name entity were labelled as O (out).

For tokenising documents into sentences and words, we used EstNLTK's (Laur et al., 2020) segmentation tools. We made adaptions to the default segmentation rules in order to align name annotations with word tokens. For instance, we added rules for splitting name tokens that were mistakenly joined with non-name tokens, such as *'talumeesNikolai'* → *'talumees Nikolai'* 'farmer Nikolai'. We also fixed EstNLTK's automatic compounding of names with initials into single token (e.g. person names *'M. Nipman'* and *'J. Pader'*), because the heuristic frequently produced wrong tokenization due to roman numerals. For instance, the text snippet *'I. Jaan Rand'* was mistakenly split into tokens *'I. Jaan'* and *'Rand'*, although *'I.'* was actually a roman numeral (as the name appeared in an enumeration of names), so the expected tokenization is *'I'*, *'.'*, *'Jaan'* and *'Rand'*.

It must be noted that our tokenization adaption focused only on most frequent issues related to the task at hand, and a complete tokenization adaptation to the historical language was out of the scope of this work.

Overall, we used 5 named entity categories. As categories Other, Unknown and Artefact were relatively infrequent (when accumulated, they make up less than 1% of all entities), we replaced these categories with a single category Misc(ellaneous) for the experiments.

4.2. Experimental setup

We keep the test set same is in Poska's (2021) experiments, and split the remaining data into 90% training and 10% development set[4]. Document boundaries were preserved while splitting the data. Table 2 gives overview of the statistics resulting from the data split.

	train	dev	test	*total*
# documents	1 125	125	250	1 500
# sentences	16 040	2 336	3 170	21 546
# words	240 614	28 891	50 900	320 405
# named entities	20 944	2 357	4 239	27 540

Table 2: Corpus statistics for training, development and test sets. Word and sentence counts are based on (adapted) EstNLTK's text segmentation.

We used the HuggingFace Transformers library (Wolf et al., 2020) for fine-tuning transfer learning models. Hyperparameters were selected based on fine-tuning a model for 3 epochs on the training set and evaluating on the development set, using a grid search over learning rate values (5e-5, 3e-5, 1e-5) and batch size values (8, 16, 32). After that, a model with the best-performing hyperparameters was fine-tuned until F1-score no longer improved on the development set (with the limit of 10 epochs at maximum), and then evaluated on the test set.

For evaluation, we used the $nervaluate$ package[5], which implements the entity-level evaluation metrics used in the SemEval-2013 Task 9 (Segura Bedmar et al., 2013). We report results from the strict evaluation, which requires exact match of named entity string boundaries and entity types.

4.3. Methods

4.3.1. CRF (baseline)

We used the named entity recognition model proposed by Tkachenko et al. (2013) as a baseline method[6]. The model uses CRFs as the learning algorithm and employs features based on word's surface form, morphological analysis, appearance in a large name gazetteer and word's other occurrences in the document. These features were originally developed for analysing modern Estonian news, and their detailed description can be found in Tkachenko et al. (2013). In order to keep the settings comparable with BERT models, we trained the model from the scratch on the training part of the corpus.

4.3.2. EstBERT

EstBERT[7] is a language-specific BERT model for Estonian, which finetuning has been shown to achieve state-of-the-art results for multiple NLP tasks, including named entity recognition (Tanvir et al., 2020). EstBERT was pre-trained on a 1.1 billion word Estonian National Corpus 2017, which consisted of Estonian Web Corpora (2013 and 2017), Estonian Wikipedia 2017 and Estonian Reference Corpus (1990-2008) (Kallas and Koppel, 2018).

Our best performing EstBERT NER model was fine-tuned for 9 epochs, using the batch size 8 and learning rate 5e-05.

4.3.3. WikiBERT-et

WikiBERT-et[8] is a BERT model pre-trained exclusively on the Estonian Wikipedia (Pyysalo et al., 2020), which consisted of 38 million words at the time of the pre-training. Despite the small size of the pre-traning

[4]More specifically, Poska (2021) used a crossvalidation on 1250 documents, and kept 250 documents as a hold out set for final evaluation; we use the same hold out set, but split the 1250 documents into training and development sets.

[5]`https://github.com/MantisAI/nervaluate` (2021-12-20)

[6]Although Poska (2021) proposed an adaptation to Tkachenko et al. (2013)'s model, and showed it outperforming the original model, our experiments on the new data split did not confirm the superiority of the adapted model. So, we chose the original model as our baseline.

[7]`https://huggingface.co/tartuNLP/EstBERT` (2022-01-03)

[8]`https://huggingface.co/TurkuNLP/wikibert-base-et-cased` (2022-01-03)

dataset, finetuning the model for named entity recognition has shown very competitive results, and even outperforming EstBERT, notably on recognizing the Organisation category (Tanvir et al., 2020).
Our best performing WikiBERT-et NER model was fine-tuned for 10 epochs, using the batch size 8 and learning rate 5e-05.

4.3.4. Est-RoBERTa

Est-RoBERTa[9] is a large monolingual BERT-like model, which has been pre-trained on a 2.51 billion token corpus, containing mainly Estonian news articles. According to an evaluation conducted by Ulčar et al. (2021), Est-RoBERTa fine-tuned for NER significantly outperforms EstBERT on the task.
Our best performing Est-RoBERTa NER model was fine-tuned for 8 epochs, using the batch size 16 and learning rate 5e-05.

4.4. Results

Overall results. Table 3 reports micro-averaged NER results on the test set. The highest performing model was Est-RoBERTa (F1-score of 93.6%), and WikiBERT-et model obtained the second best result (91.63%). While EstBERT outperformed the CRF baseline with higher recall (91.2% vs 88.2%), the CRF baseline obtained higher precision (91.6% vs 89.7%).

model	precision	recall	F1
CRF (baseline)	91.57	88.18	89.84
EstBERT	89.74	91.15	90.44
WikiBERT-et	91.29	91.98	91.63
Est-RoBERTa	92.97	94.24	93.60

Table 3: Evaluation results on the test set. Microaveraged over all test documents and across name categories.

These transfer learning results are relatively high and on par with the NER results obtained on modern Estonian (Ulčar et al., 2021; Tanvir et al., 2020)[10]. The high results may come as surprising, considering that all the BERT models we experimented with have been pre-trained on the modern Estonian, not on the historical language. While do not know the exact reasons of these results, few hypotheses can be put forward. First, the quality of manually transcribed texts is relatively good (unaffected by OCR errors), and that likely enables high performance. From this perspective, our experimental settings are similar to the settings of Nissim et al. (2004) and Aguilar et al. (2016), where high NER performances were obtained on manually transcribed historical documents. Second, while our corpus is characterised by high linguistic and spelling variability, the overall structure of a Parish Court record is relatively regular, reflecting the order of court procedures prescribed by law (Pilvik et al., 2019), and this may also have contributed to high scores.

Category-wise results. As can be observed from Table 4, Est-RoBERTa also achieves the highest scores on all named entity categories. All models perform very well (F1 scores over 92%) on the most frequent category PER, and Est-RoBERTa achieves the top F1 score of 96% on the category.
Models achieve the second best performance (F1 scores ranging from 83% to 95%) on the ORG category, despite its low frequency. This result can be explained by regularity of organisation names: most names in this category were court names, which tended to appear in certain contexts in the document (frequently at the beginning of the document).
The performance on LOC-ORG category ranged from 71% to 81%. Difficulties that hampered models from achieving higher scores on the category may reflect those of human annotators, as the inter-annotator agreement for LOC-ORG also peaked at F1 score of 81%.
The category LOC was the most difficult one to learn, with F1 scores ranging from 58% to 66%. We hypothesise that this difficulty stems from rareness and high variability of names in that category, as the category covered names of local landscape objects (such as rivers and hills) as well as names of larger geographical entities, such as towns and countries.
Models' F1-scores on the least frequent category MISC ranged from 61% to 74%. A large portion of names in that category were re-occurring mentions of the peasant law codex, which likely contributed to achieving higher F1 scores than on the (more frequent) LOC category.

4.5. Error analysis

We conducted a preliminary analysis of errors in the output of our best model, Est-RoBERTa, dividing the errors in the test set into automatically detectable categories (cf table 5) and then examined them manually, trying to establish recurrent patterns.
The most common error type is the erroneous determination of the named entity boundaries. The common examples of this kind of error are the person names containing a patronymic name and an abbreviation standing for 'son' or 'daughter'. For example a person name *Dawid Peetri p. Lawasson* is divided into two separate person names *Dawid Peetri p.* and *Lawasson*. It might be the full stop after the abbreviation that confuses the model as the same error appears also in cases of abbreviated forenames. For example a person name *Hindr. Laari* (abbreviation *Hindr.* stands for *Hindrek*) is divided into two person names – *Hindr.* and *Laari* by the model.
The model has sometimes also included an extra word

[9] https://huggingface.co/EMBEDDIA/est-roberta (2022-01-03)

[10] Note also that the baseline CRF obtains a very high score: for comparison, the highest average F1 score reported by Tkachenko et al. (2013) on the domain of contemporary news was 87%; although their settings were different due to the usage of crossvalidation.

model	PER			LOC-ORG			LOC			ORG			MISC		
	p	r	f1	p	r	f1	p	r	f1	p	r	f1	p	r	f1
CRF (baseline)	93.72	92.15	92.93	81.69	70.78	75.84	64.96	52.41	58.02	87.88	79.45	83.45	69.44	55.56	61.73
EstBERT	93.50	94.86	94.17	72.68	71.03	71.85	55.90	62.07	58.82	85.00	93.15	88.89	63.04	64.44	63.74
WikiBERT-et	94.42	95.47	94.94	76.21	75.06	75.63	59.18	60.00	59.59	89.33	91.78	90.54	76.92	66.67	71.43
Est-RoBERTa	95.70	96.93	96.31	80.30	82.12	81.20	65.75	66.21	65.98	92.31	98.63	95.36	76.19	71.11	73.56

Table 4: Category-wise precisions (p), recalls (r) and F1 scores (f1) on the test set.

Error type	Number of errors	%
Wrong boundaries	115	39%
Redundant entity	66	22%
Wrong label	50	17%
Missing entity	46	15%
Wrong label and wrong boundaries	20	7%
All	297	100%

Table 5: Types of Est-RoBERTa errors in the test set

into a named entity, e.g. in the phrase *Rein õuewärawas küsinud*, lit. 'Rein at the gate asked', meaning 'Rein asked at the gate' the model has annotated the token *õueväravas* 'at the gate' as part of the person name entity.

There are also several cases where the model has placed the boundary of a named entity in the middle of a word. This is the result of Est-RoBERTa's tokenization – the model is tokenizing words into subwords and decides for every subword whether it is in the beginning, in or out of a named entity. In principle, such errors could be fixed by a post-processing step that forces alignments between named entity annotations and word tokens[11]. However, because our tokenization adaptation did not completely solve the tokenization of the historical language, we did not force such alignment. It remains a future work to investigate whether a better tokenization adaptation along with forced tokenization constraints on named entities helps to improve model's results.

Other, less frequent types of errors are redundant named entities and categorization errors. A typical example of a redundant named entity is a capitalized token in the middle of a sentence, annotated by the model as a named entity. Most frequent categorization errors are a name of a person or a location categorized as a name of a location-organization.

5. Conclusions

We have presented a new language resource, the corpus of 19th century Estonian Parish Court records, annotated for named entities. The texts have been digitized manually, but they are very heterogeneous in terms of spelling conventions, capitalization and dialectal variation.

Although our taxonomy of named entities extends beyond the most used NE types of *Person*, *Location*, *Organization*, we have shown that it is still possible to maintain high inter-annotator agreement.

We have used this annotated corpus to finetune and evaluate several BERT-like transfer learning models. Although all the models were pre-trained on modern Estonian texts, they performed surprisingly well on historical Parish Court records: the best-performing model, EstRoBERTa, achieved 92.97% precision, 94.24% recall, the F1 score was 93.60. The fact that manual digitization has produced noise-free texts is one of the possible explanations for the high quality of NER. Also the overall regular textual structure of a Parish Court record could contribute to this.

6. 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).

7. Bibliographical References

Aguilar, S. T., Tannier, X., and Chastang, P. (2016). Named entity recognition applied on a data base of medieval latin charters. the case of chartae burgundiae. In *3rd International Workshop on Computational History (HistoInformatics 2016)*.

Ahrens, E. (1843). *Grammatik der Ehstnischen Sprache Revalschen Dialektes.* Laakmann, Tallinn.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

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

Ehrmann, M., Romanello, M., Flückiger, A., and Clematide, S. (2020). Extended overview of clef hipe 2020: named entity processing on historical

[11] We would like to thank an anonymous reviewer for noting this point.

newspapers. In *CEUR Workshop Proceedings*, number 2696. CEUR-WS.

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., and Doucet, A. (2021). Named entity recognition and classification on historical documents: A survey. *CoRR*, abs/2109.11406.

Hamdi, A., Boroş, E., Pontes, E. L., Nguyen, T. T. H., Hackl, G., Moreno, J. G., and Doucet, A. (2021). A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In *Proceedings of the 44rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Hubková, H., Kral, P., and Pettersson, E. (2020). Czech historical named entity corpus v 1.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4458–4465, Marseille, France, May. European Language Resources Association.

Jørgensen, F., Aasmoe, T., Ruud Husevåg, A.-S., Øvrelid, L., and Velldal, E. (2020). NorNE: Annotating named entities for Norwegian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4547–4556, Marseille, France, May. European Language Resources Association.

Kittask, C., Milintsevich, K., and Sirts, K. (2020). Evaluating Multilingual BERT for Estonian. In *Baltic HLT*, pages 19–26.

Kolditz, T., Lohr, C., Hellrich, J., Modersohn, L., Betz, B., Kiehntopf, M., and Hahn, U. (2019). Annotating german clinical documents for de-identification. In *MedInfo 2019 – Proceedings of the 17th World Congress on Medical and Health Informatics*, pages 203–207.

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.

Nissim, M., Matheson, C., and Reid, J. (2004). Recognising geographical entities in scottish historical documents. In *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*, volume 35.

Pilvik, M.-L., Muischnek, K., Jaanimäe, G., Lindström, L., Lust, K., Orasmaa, S., and Türna, T. (2019). Möistus 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.

Poska, K. (2021). Nimeolemite tuvastamine 19. sajandi vallakohtu protokollides (Named Entity Recognition in 19th Century Parish Court Protocols). Bachelor's thesis, University of Tartu.

Pyysalo, S., Kanerva, J., Virtanen, A., and Ginter, F. (2020). Wikibert models: deep transfer learning for many languages. *arXiv preprint arXiv:2006.01538*.

Raag, R. (2008). *Talurahva keelest riigikeeleks [From the language of peasants to state language]*. Atlex, Tartu.

Ruokolainen, T. and Kettunen, K. (2018). À la recherche du nom perdu–searching for named entities with stanford ner in a finnish historical newspaper and journal collection. In *13th IAPR International Workshop on Document Analysis Systems*.

Segura Bedmar, I., Martínez, P., and Herrero Zazo, M. (2013). Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.

Stenetorp, P., Pyysalo, S., Topić, 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.

Swaileh, W., Paquet, T., Adam, S., and Camacho, A. R. (2020). A named entity extraction system for historical financial data. In *International Workshop on Document Analysis Systems*, pages 324–340. Springer.

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.

Traat, A. (1980). *Vallakohus eestis 18. sajandi keskpaigast kuni 1966. aasta reformini [Parish courts in Estonia from mid-18th century to the 1866 resorm]*. Eesti Raamat, Tallinn.

Ulčar, M., Žagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., and Robnik-Šikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. *arXiv preprint arXiv:2107.10614*.

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 on*

Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October. Association for Computational Linguistics.

8. Language Resource References

Kallas, J. and Koppel, K. (2018). Eesti keele ühendkorpus 2017 (*Estonian National Corpus 2017*). https://doi.org/10.15155/3-00-0000-0000-0000-071E7L.