Creating a corpus of communal court minute books: a challenge for digital humanities

Gerth Jaanimäe, Liina Lindström, Kadri Muischnek, Siim Orasmaa, Maarja-Liisa Pilvik (University of Tartu),
Kersti Lust (The National Archives of Estonia)

This paper presents the work of a digital humanities project concerned with the digitization of Estonian communal court minute books. The local communal courts in Estonia came into being through the peasant laws of the early 19th century and were the first instance class-specific courts, that tried peasants. Rather than being merely judicial institutions, the communal courts were at first institutions for the self-government of peasants, since they also dealt with police and administrative matters. After the municipal reform of 1866, however, the communal courts were emancipated from the noble tutelage and the court became a strictly judicial institution, that tried peasants for their minor offences and solved their civil disputes, claims and family matters. The communal courts in their earlier form ceased to exist in 1918, when Estonia became independent from the Russian rule.

The National Archives of Estonia holds almost 400 archives of communal courts from the pre-independence period. They have been preserved very unevenly and not all of them include minute books. The minute books themselves are also written in an inconsistent manner, the earlier minute books are often written in German and the writing is strongly dependent on the skills and will of the parish clerk. However, the materials from the period starting with the year 1866, when the creation of the minute books became more systematic, are a massive and rich source shedding light on the everyday lives of the peasantry. Still, at the moment, the users of the minute books meet serious difficulties in finding relevant information since there are no indexes and one has to go through all the materials manually. The minute books are also a fascinating resource for linguists, both dialectologists and computational linguists: the books contain regional varieties tied to specific genre and early time period (making it possible to detect linguistic expressions, which are rare in atlases, for example, and also in dialect corpus, which represents language from about 100 years later) while also being a written resource, reflecting the writing traditions of the old spelling system. This is also what makes these texts complex and challenging for automatic analysis methods, which are otherwise quite well-established in contemporary corpus linguistics.

In our talk we present a project dealing with the digitization and analysis of the minute books from the period between 1866 and 1890. The texts were first digitized in the 2000s and preserved in a server in html-format, which is good for viewing, but not so good for automatic processing. After the server crashed, the texts were rescued via web archives and the structure of the minute books was used to convert the documents automatically into a more functional format using xml-markup and separating the body text with tags referring to information about the titles, dates, indexes, participants, content and topical keywords, which indicate the purview of the communal courts in that period.

We discuss the workflow of creating a digital resource in a standardized and maximally functional format as well as challenges, such as automatic text processing for cleaning and annotating the corpus in order to distinguish the relevant layers of information. Tools developed for Estonian morphological analysis are trained on contemporary written standard Estonian. Communal court minute books, however, include language variants, which are a mixture of dialectal language, inconsistent spelling and the old spelling system. In the presentation, we introduce the results of our first attempts to apply the automatic text analysis tools to the materials of communal court minute books, the problems that we've run into, and provide solutions for overcoming these problems. Similar experiments of testing tools developed for contemporary language on historical texts have been conducted on other languages as well, e.g. Icelandic (Lofsson 2013; Rögnvaldsson and Helgadóttir 2011), German (Scheible et al. 2011; Bollmann 2013), Swedish (Petterson 2016), and Spanish (Sánchez-Marco et al. 2011). To achieve better accuracy rate for tagging, the two main proposed solutions have been text normalization (Scheible et al. 2011; Bollmann 2013; Petterson 2016) and tool adaption (Rögnvaldsson and Helgadóttir 2011; Sánchez-Marco et al. 2011).

As the National Archives have a considerable amount of communal court minute books, which are thus far only in a scanned form, the digitized minute books collection is planned to expand using crowdsourcing oportunities. The final aim of the project is to create a multifunctional source, which could be of interest for researchers of different fields within the humanities. In addition to comprising important lingustic information, it enables systematic studying of family matters, taxes, granary loans, old age support, tenancy, damages, contractual relationships, credit relations, inventories, living conditions, and minor criminal offences. Furthermore, the database can be linked to individual level demographic databases as well as genealogical databases of various sorts. Thus, both professional and hobby researchers have a great potential to benefit from the project. For example, identification of the kin relationships of otherwise 'anonymous' people as well as of their household and class affiliation allows analyzing the events recorded in the minute books in terms of class, kin and networks.

Preliminary analyses

In order to enable queries with different degrees of specificity in the corpus, the texts also need to be linguistically analyzed. For historical texts written in languages with small number of inflectional forms per lemma, handling spelling variation and canonicalization (Piotrowski 2012: 73-78) is sufficient for word and string based searches, frequency lists etc. However, for languages with rich morphological system that is not enough and the text needs to be lemmatized, which in case of rich inflectional morphology, is done by performing morphological analysis and disambiguation. For both named entity recognition (NER), which enables network analysis and links the events described in the materials to geospatial information, and morphological annotation, which makes it possible to perform queries based on lemmas or grammatical information, we have applied the EstNLTK library (Orasmaa et al. 2016) in Python, which is developed for processing contemporary written standard Estonian. NER's performance was satisfactory, i.e. it found recognized names well, even though it systematically overrecognized organization names. The most complicated issue so far has been the morphological analysis of word forms. Apart from being simply an older version of

Estonian, the language represented by Estonian communal court minute books contains a lot of twofold variation: spelling variation and dialectal variation. Estonian dialects are divided into two main groups: Northern and Southern; Standard Estonian is based on the Northern dialects. Although the parish clerk not necessarily was a speaker of the local dialect, this seems quite often to have been the case. One may therefore assume that automatic morphological analysis and lemmatization of the texts from the Northern parishes should give better results than that of the texts from the Southern parishes.

In order to roughly estimate the quality of the outcome of morphological analysis of communal court minute books, we processed the texts using EstNLTK morphological analyzer without the guesser. Running the analyser without the guesser means that the out-of-vocabulary words, apart from compounds and regular derivations, are tagged as unknown words. The percentage of unknown words per parish is presented in Table 1.

Table 1. Results of automatic morphological analysis for communal court minute books

Parish	Dialectal group	% of words with morphological analysis	of them unambiguous, %	% of unknown words	of them with capital initial letter, %
Laiuse	Northern	93	49	7	74
Mihkli	Northern	90	50	10	56
Navesti	Northern	89	48	11	38
Uue-Suislepa	Southern	86	51	14	38
Alatskivi	Northern	84	54	16	46
Kiuma	Southern	83	53	17	35
Maasi	Northern	83	46	17	34
Vastse-Nõo	Southern	81	53	19	27
Laeva	Southern	80	53	19	43
Aru	Southern	79	53	21	34
Tarvastu	Southern	75	56	25	23
Pangodi	Southern	69	54	31	27
Kahkva	Southern	67	57	33	26

Kärevere	Southern	61	56	39	29
Mäksa	Southern	61	59	39	21
Joosu	Southern	61	48	39	24
Haaslava	Southern	60	61	40	24
Kokora	Southern	60	51	40	29
Valguta	Southern	59	61	41	16
Luke	Southern	53	64	47	27
Suure-Konguta	Southern	49	57	51	27
Väike-Rõngu	Southern	48	53	52	18

The percentage of recognized words varies from 93% for the texts of Laiuse belonging to the Northern dialect group to 48% for the texts from Väike-Rõngu belonging to the Southern dialect group. This perhaps means that due to wide inter-parish variation, we will have to use different means for lemmatizing and/or normalizing texts from different parishes or parish groups.

For comparison, we also processed fiction and newspaper texts in present-day Standard Estonian. The results are presented in Table 2. The comparison of these results with those of the parish court texts shows that the analyzer recognizes minimally *ca.* 4% more word forms when analysing present-day Estonian texts and out of the out-of-vocabulary words *ca.* 12% more word forms begin with a capital letter, i.e. are probably proper nouns.

Table 2. Results of automatic morphological analysis for present-day Standard Estonian texts

text class	% of words with morphological analysis	of them unambiguous, %	% of unknown words	Of them with capital initial letter, %
fiction	98	54	2	86
newspaper	97	60	3	87

References:

Bollmann, Marcel (2013). POS Tagging for Historical Texts with Sparse Training Data. - Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pp 11–18.

Loftsson, Hrafn (2013). Tagging the Past: Experiments Using the Saga Corpus. - Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA-2013).

Orasmaa, Siim, Timo Petmanson, Alexander Tkachenko, Sven Laur, Heiki-Jaan Kaalep (2016). ESTNLTK - NLP Toolkit for Estonian. - Proceedings of LREC 2016, pp 2460–2466.

Petterson, Eva (2016). Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction. Doctoral Thesis, Uppsala University.

Piotrowski, Michael (2012). Natural Language Processing for Historical Texts. Morgan & Claypool Publishers.

Rögnvaldsson, Eirikur and Sigrún Helgadóttir (2011). Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. - C. Sporleder, A. van den Bosch, and K. Zervanou, editors, Language Technology for Cultural Heritage, Theory and Applications of Natural Language Processing, pp 63–76.

Sánchez-Marco, Cristina, Gemma Boleda and Lluís Padró (2011). Extending the tool, or how to annotate historical language varieties. - Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pp 1–9.

Scheible, Silke, Richard J. Whitt, Martin Durrell and Paul Bennett (2011). Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. - Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pp 19–23.