Literary Studies Meet Corpus Linguistics: Estonian Pilot Project of Private Letters in KORP[*]

Marin Laak[1], Kaarel Veskis[1],
Olga Gerassimenko[2], Neeme Kahusk[2], and Kadri Vider[2]

[1] Estonian Literary Museum
Vanemuise 42, 51003 Tartu, Estonia
{marin.laak,kaarel.veskis}@kirmus.ee
[2] Institute of Computer Science, University of Tartu,
J. Liivi 2, 50409 Tartu, Estonia
{olga.gerassimenko,neeme.kahusk,kadri.vider}@ut.ee

Abstract. Digitalisation of cultural heritage in Estonia has been in progress during recent years, and we will see expansive mass digitalisation of printed books and handwritten documents in the very near future.

This situation and potential actualises questions of the usage of the literary heritage. In our paper, we consider benefits for digital literary research that arise from representing a literary text collection as an annotated language resource. We discuss the pilot project of creating a text corpus based on private letters between two Estonian avantgarde writers in the beginning of the 20th century.

The advantages and possibilities of corpus query system KORP that we have chosen for representing and searching literary heritage DH corpora as a language resource are described. Challenges that the application of Natural Language Processing and Text and Data Mining imposes on the preparation and representation of texts are discussed along with benefits for the research.

Keywords: literary studies · private letters · digital heritage collections · corpus linguistics · natural language processing

1   Introduction

Application of Natural Language Processing (NLP) and Text and Data Mining (TDM) methods and tools is gradually increasing and becoming more relevant in DH research in Estonia. Similarly to the Nordic and Baltic countries, Estonia can expect an explosive growth in digital heritage and text resources: preparations for massive digitisation of cultural heritage started as a national programme in 2019. The creation of these new digital resources will be the priority for all Estonian leading memory institutions and the scope of this huge project includes different types of cultural heritage (printed books, archival documents, photo and film heritage, ethnographic and fine art objects) [6,10] Digitised resources will be made accessible on the internet as open data, but it is still an open question how and for what purpose can such digital resources be used [3].

What are the new challenges and the new knowledge offered by using the tools of NLP in analysing the digital literary heritage? Our interdisciplinary project "Literary Studies Meet Corpus Linguistics" concentrates on using cultural heritage, especially literary history sources in research with the help of the NLP&TDM methods. The primary question is how to bridge the gap between the research possibilities offered by the contemporary NLP&TDM and the ever increasing amount of digitised texts and other digital data, produced by memory institutions. This has proved to be a complicated task and international practice has shown that literary scholars are slower to embrace new practices than linguists for whom corpus-based research is already a professional standard. In general, literary scholars are used to working with texts using traditional methods, analysing them as undivided poetical and semantic entities [10]. Digitising of texts often preserves them as undivided and unannotated entities with occasionally added metadata - as a result, the digitised materials are human-readable rather than machine-readable.

NLP&TDM methods require treating literary works or texts as data, which can be analysed and processed with computer programmes. To process data for statistics and for finding trends, a literary scholar needs either to do a close reading of literary texts as huge data amounts, or to develop new professional skills in data analysis methods. This imposes rethinking the approach to empirical object in literary studies in general and posing new and different research questions. The possibility to compare text strategies, rhetorical and stylistic patterns in literary, religious and political text corpora might give us new insights into intertwining ideology, rhetoric and identity presentations. One of the important trends of automated textual analysis in DH is Sentiment Analysis which, for instance, allows to measure emotion in parliamentary debates [7].

2   Private letters as a literary source

The empirical foundation of our pilot project of private letters in KORP is based on the archival ego-documents. This is a rare collection of handwritten private letters: the correspondence between two Estonian writers Johannes Semper

(1892–1970) and Johannes Barbarus (1890–1946). The authors were well-known avant-garde writers in Estonian literature, best friends and school classmates. They shared the same intellectual attitudes, views, and values. For example, they were Francophiles and during several decades they both reviewed and translated French literature into Estonian.

During the years covered by the correspondence, in 1911–1939, Barbarus published the total of 17 collections of poetry, and Semper published four collections of poetry and four books of short stories and novels. The letters were written mainly at different locations in Estonia, and also during travels: they both travelled extensively in Europe (and other parts of the world) and in their letters they described to each other the sights they saw as well as the literary events at home. They also wrote about their work, everyday life, health and prescriptions, hobbies (hunting, swimming, skating, etc.), and guests.

The temporal and contextual frames of this correspondence are the period in Europe between two World Wars and, according to the French philosopher Pierre Bourdieu, the creation of the Estonian cultural field in the Estonian Republic in years 1918–1940. As such, the correspondence offers semantically rich and multidimensional content for literary scholars.

Generally, we have been inspired by the question how this kind of sources (private letters, correspondences, manuscripts, etc.) in different kinds of machine-readable formats (pdf, docx, txt etc.) can be used for creating new knowledge in the interdisciplinary field of literary studies, including biographies and life writing studies [8].

The private correspondence of two Estonian writers during 28 years, as empirical material, is extremely rich in the themes and possibilities to pose the traditional as well as the new research questions: the factors of subjectivity and emotionality, verbal and poetic creativeness of the both authors, etc.

The original letters are held at different archives: the letters of Barbarus to Semper belong to the Estonian Cultural History Archives of the Estonian Literary Museum and the letters of Semper to Barbarus belong to the Estonian National Archives due to the unique literary and historical value of this correspondence. Their correspondence consists of 670 letters with, all in all, more than 1,100 pages and more than 310 000 tokens or about 250 000 words. The number of tokens and words is approximate, as we will discuss later.

3   KORP as a tool for corpus linguistics

KORP is a corpus query analysis system that allows to find concordances and build various statistics from differently annotated corpora using text metadata (author, year of publishing, text type etc) and linguistic annotation (splitting into sentences and words, punctuation, morphology, syntax and semantics). Technically, KORP is a frontend tool that uses much of IMS Open Corpus Workbench [2] as backend. It was created and maintained at Swedish Language Bank Språkbanken [1] and is being developed and customised by language technology

networks in several countries: Sweden[3] , Finland[4], Norway[5], Estonia[6], Denmark[7] and Island[8].

Estonian KORP is hosted by the Centre of Estonian Language Resources[9]. The Estonian data available in KORP currently consists of more than 850 million tokens. As Estonian is a highly inflective language with a rich morphology and free word order, almost all texts are automatically annotated and disambiguated on the morphological level to enable form- and baseform-based search.

As a pilot project for the needs of literary studies in addition to the corpora designed for linguistic research, we have created Correspondence Corpus of private letters (Semper and Barbarus, 1911–1940). The original handwritten manuscripts where already typed in. They were to transform to machine-readable text files by manually adding metadata and automatically analysing and disambiguating morphological categories.

KORP as a corpus query system was chosen because of its open source value, search flexibility and easiness, graphical overview of search results in subcorpora, easy switching between concordances and broader context (as well as between frequency statistics and concordances), broad possibilities to group absolute frequencies of tokens or occurrences and automatically calculated relative frequencies (occurrences per million tokens). Search results and statistics from KORP are exportable into Comma Separated Values (CSV) and JSon format.

Text fragments cited in KORP are limited to a sentence or a passage, so that KORP does not infringe copyrights. Metadata cited in the query results allow exact pinpointing of the source of the sentence, and it is possible to link whole texts hosted elsewhere to the search metadata for close reading of a broader context when needed.

KORP helps to analyse and objectively verify or refute different research arguments, for example, such as presented by Estonian literary scholars in the 1980s: (1) The correspondence of Semper and Barbarus is subjective and emotional, the letters reveal the character and state of mind of the authors at the moment of writing them. (2) The letters demonstrate the authors awareness of topical problems in Estonia and in Europe. (3) The subject range of the letters includes everyday life, health, hobbies, visits and visitors, literary work, books and reading, and the literary, economic and political life in Estonia and Europe.

4   Challenges for annotation in manuscript corpus

Which distinctive features of older correspondences need to be considered in preparing them to be used as a corpus? Several steps are necessary to convert the

---

[3] https://spraakbanken.gu.se/korp/

[4] http://korp.csc.fi

[5] http://gtweb.uit.no/korp/

[6] http://korp.keeleressursid.ee

[7] http://alf.hum.ku.dk/korp/

[8] http://malheildir.arnastofnun.is/

[9] https://www.keeleressursid.ee

digitised text collection into a morphologically analysed corpus, and to provide it with essential metadata about the sources.

1. The first step in digitising different kinds of manuscripts, incl. private letters, is retyping the handwritten originals in order to transform them into machine-readable format. Digitised typed texts can be subjected to the character recognition (OCR).
2. In order to enable statistics and to find different linguistic phenomena in the texts, it is important to tokenise at least sentences and words, and to preserve meaningful units (chapters, articles, verses, letters).
3. The process of annotation was carried out semi-automatically. At the first stage, sentence boundaries were determined, then words were tokenised. Tokenised words were lemmatised and tagged by part-of-speech and other grammatical features.

Different levels of annotation make it possible to search for different linguistic or other phenomena. In a tokenised text we can search by word forms (or their parts), but not by lemmas. In a lemmatised text we can search also by lemmas, but not by morphological characteristics (e.g., the case, the number). When the text has been morphologically analysed and disambiguated, we can search also by morphological characteristics and parts of speech. The automatic morphological analysis and automatic disambiguation that enable searching for different morphological forms of a word based on the lemma, seriously improve the usability and quality of the corpus. The accuracy of automatic morphological disambiguation in the Estonian standard language corpora reaches up to 93–98% [9]. The accuracy for these texts is probably much lower.

Meaningful units, e.g., temporal expressions or named entities, found in the text can also be used in searches. If all the meanings of words were tagged, it would be possible to search for a specific meaning of a word and discard all other meanings.

5   Annotation types specific for private letters

The private letters of writers, the correspondence as whole was not addressed for a wider audience. Creators of the corpus met serious difficulties because often, both correspondents were very familiar with the subjects mentioned in the letters, so that sometimes they only hinted at the events or persons and used plenty of abbreviations known only to themselves. Authors of letters reflect both the historical period and the personal context and idiolects. Being avant-garde poets, Semper & Barbarus both experimented with the language a lot.

In the following sections we are going to present the distinctive features of the texts that would be of great interest for both linguists and literary scholars. See Fig. 1 for a passage from a letter by Barbarus to Semper with annotations and English translation.

5.1   Abbreviations

In this corpus, there are a lot of nonstandard abbreviations that were comprehensible to the authors of the letters. The abbreviations pose a challenge to the morphological analysis and disambiguation of abbreviations itself, but also of the neighbouring words, the disambiguated analysis of which often relies on the context (for instance, "is. Linde" should be analysed as "isand (i.e master) Linde" and not as a sentence ending and a starting of a new sentence although there is no common abbreviation "is." in standard Estonian).

5.2   Date and other temporal expressions

The date format of the manuscript letters varies greatly: "11. jaanuar 1927", "19/XII.37", "2. 2. 1934", "Jõulu 3. pühal 1934" (on the 3rd day of Christmas 1934). The date in the format chosen by the letter authors needs to be saved as a part of the text, but in addition to that, all the date formats need to be normalised so that they could be present in the letter metadata in a standardised format (i.e., YYYY-MM-DD) whenever possible. When it is not possible, we can use the category "undefined" for the analysis transparency. Temporal expressions in the text of letters can be automatically recognised with the help of EstNLTK library for Python [5] and a software tool developed by Siim Orasmaa [4].

5.3   Named Entity Recognition

Special interest for the literary scholars lies in the proper names used in the letters: their usage combined with the metadata (time period, author) might lead to recognising the important patterns and tendencies.

Named Entity Recognition (NER) annotations can be produced with the EstNLTK toolkit and include the types of entities: person (PER), organisation (ORG), and location (LOC).

5.4   Detect other languages

Semper & Barbarus were both promoters of French literature and culture in Estonia. They both worked as translators and travelled a lot. They scattered their letters with words, phrases, sentences and longer citations in many other languages than Estonian that were familiar to them (Latin, French, Russian, German). This is a challenge for the automatic language recognition and annotation tools which are currently oriented only to the analysis of standard Estonian. These tools do not help much in case of excerpts from other languages, but if such code-switching pieces were precisely recognised, lemmatised and analysed, that information would be useful both to linguists and literary scholars.

" *Bifur* "i kiri saabus *Sinu kirjaga ühel päeval* . Palutakse *kohe* luule üle artikkel ära saata ja kedagi paluda, kes teisi küsimusi (*sur la vie en général*) käsitaks. Neil olla tõlkija, nii siis võivat eesti keeleski kirjutada. Et mul artikkel juba valmis oli, siis saatsin ta *täna* minema. Kui Sul lusti midagi saata, siis läkita *kohe* , — ehk novelli tõlge (maksavad 50 *fr.* leheküljest), ehk siis mahutavad neljandamasse *nr* -isse; ehk viskad proosa & teaatri ülegi artikli. Aadress : " *Bifur* " , *Éditions du Carrefour* , *199 , boul . St.-Germain , Paris* (VIe) (*M–eur le rédacteur en chef Ribemont Dessaignes*). Küsisin kirjas, kas nende tõlkija luuletisi tõlkida võiks, siis võiksime valiku teha, ehk Suitsi ilmuva antoloogia neile saata. " *Bifur* " harrastab küll rohkem proosat & informatsioonilaadilisi ülevaateid, nii siis vaevalt nad luule liimile lähevad, aga eks ole ju veel teisi žurnaale pääle " *Bifur* "i, kus avaldada saaks, kui aga tõlkija leiduks. Mis teeb see *E.K.L.* propagandakomitee? Kas peab viimaks nende liikmete seas propagandeerima hakkama? *Visnapuu* kirjutab " *V.-M* aas" *Igori* pötserduse puhul, et vaja propagandeerida, aga, kui ma ei eksi, oli ta ise selles komitees?

Legend:

green — Named Entity: Person, Organisation, Title

yellow — Named Entity: Location, Address

blue — Abbreviations

orange — other languages than Estonian

magenta — temporal expression

"The letter from Bifur arrived on the same day as yours. They asked to send immediately the article about poetry, and to find somebody who could treat other questions (sur la vie en général). They are supposed to have a translator, thus it could be written in Estonian. Since I had already finished my article, I sent it to them today. If you think youd like to send something, do it at once, — maybe a translation of some short story (they pay 50 francs a page), and perhaps they can fit it into the fourth issue; or maybe you will even rush an article about prose and theatre. Address: Bifur, Éditions du Carrefour, 199, boul. St.-Germain, Paris (VIe) (M—eur le rédacteur en chef Ribemont Dessaignes). I asked them in my letter whether their translator can translate poems, in this case we could select something, or we can send them the anthology by Suits, which will come out soon. Bifur is more into prose and information-like reviews, so it's unlikely that they can be cajoled into accepting poetry, but there are other magazines besides Bifur where we could publish, if only we found a translator. What is the E.W.U.'s [Estonian Writers' Union] propaganda committee up to now? Should we by any chance start propagandising among their members? Visnapuu writes in the V.-Maa about Igor's handiwork that it is necessary to propagandise, but if I'm not mistaken, wasn't he on the committee himself?"

Fig. 1. Excerpt from letter (from Barbarus to Semper, Nov 8., 1929) with different annotation types. See Table 1 for metadata associated to the letter the passage comes from.

Table 1. Text attributes and their values for example presented at Fig. 1

Text Attribute	Value
author	Barbarus
recipient	Semper
catalogue no.	333
original date	8. november 1929
category	[empty]
date	1929-11-08
year	1929
location	Pärnu
notes	[empty]

6   Using private letters as text corpus

The creation of the text corpus of the Semper & Barbarus correspondence had two wider objectives.

From the perspective of corpus linguistics, to study the effect of the distinctive characteristics of private letters in creating a text corpus: what can and what cannot be achieved? What difficulties may arise in the course of such work? How can a digitised collection be converted into a language resource to study the linguistic phenomena in their literary and cultural contexts?

From the perspective of literary studies, to test the suitability of the methods of corpus linguistics in solving the problems that literary scholars face.

One of the corpus usage examples is verification of the index of proper names mentioned in the correspondence. The index was created manually more than 30 years ago and it lists the foreign writers mentioned in the correspondence. From the visualisation of index statistics (Fig 2) we can only see that Gide is the most discussed French writer in the whole correspondence.

Query of KORP allows us to see the concordances of the Gide mentions, their absolute and relative frequency in the corpus, but not only. KORP allows us to organise the statistics by all the categories used in the corpus, including metadata categories. To know who of the authors and when mentioned Andre Gide, we only have to add those metadata categories ("sender" and "date") to the statistics criteria; other metadata and linguistic categories can also be used for more sophisticated statistics that is connected to the concordances and broader context text blocks.

From the statistics we see that it was mainly Semper who talked about Gide and the time interval is 1926 to 1930. It aligns well with the literary history: in 1928 Semper graduated as Master of Arts (*magister artium*) from the University of Tartu (with a masters diploma in literary studies, "The structure of the literary style of André Gide") and continued his work at the university as an aesthetics and stylistics lecturer.
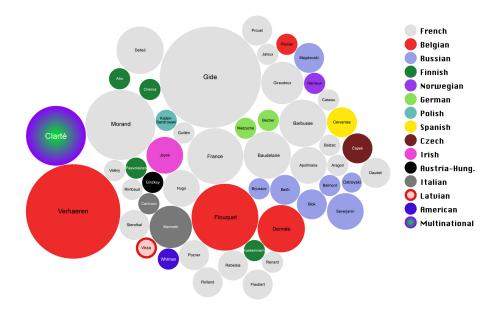
Fig. 2. Foreign writers and countries mentioned in the correspondence based on the proper name index. The size of the circle depends on the number of occurences in the corpus. The graph was generated using open source Javascript library D3 from a predefined list of authors and number of letters to Barbarus. The colours were later fine-tuned in Adobe Illustrator. We would like to thank Kristjan-Julius Laak for providing us with this infographics.

Table 2. Statistics of mentioning Gide in KORP, organised by word form, sender and date.

word	sender	date	occurrences (absolute frequency)
Gide'i	Semper	–	6.4 (2)
Gide'i	Semper	1926-12-13	3.2 (1)
Gide'i	Semper	1928-05-26	3.2 (1)
Gide'il	Barbarus	1928-05-26	3.2 (1)
Gide'i	Semper	1929-04-03	3.2 (1)
Gide	Semper	1929-05-06	3.2 (1)
Gide'i-raamatule	Semper	1929-04-03	3.2 (1)
Gide'i	Semper	1929-10-29	3.2 (1)
Gide'	Semper	1929-11-10	3.2 (1)
Gide'i	Barbarus	1929-12-08	3.2 (1)
Gide	Semper	1930-01-10	3.2 (1)
Gide'i	Semper	1930-10-13	3.2 (1)
Gide'i	Barbarus	1930-10-19	3.2 (1)
Gide'i	Semper	1931-07-03	3.2 (1)
		Total:	51.5 (16)

7   Current workflow and future work

At present stage we have applied the standard procedures for converting text into KORP-compatible corpus to the Correspondence corpus as well with some minor additions. The standard procedure is a pipeline consisting of tokeniser, sentence splitter, morphological analyser (including POS-tagger and lemmatiser) and morphological disambiguator. Sentence splitter was modified according to non-standard abbreviations found from the Correspondence texts, but morphological analyser was used with standard options: guesser and forced proper-name finder.

Sentence splitter and tokeniser are tools that determine the quality of the corpus. They can be adjusted by properties of corpus and annotation aims, e.g. we can annotate a date as one token, or as a series of several tokens. This would give different results of number of tokens and number of sentences.

Guesser makes the analyser find answers for words not in analyser's lexicon, and proper-name finder gives precedence to proper names, if analysis is ambiguous and word starts with a capital letter. The corpus is not properly evaluated, but quick overview revealed at least two problems in POS tagging and lemmatisation. First, the POS tagging seem to be too optimistic about proper names, and second, foreign words and phrases are analysed as Estonian.

We compared the number of proper names with manually disambiguated corpus, and found the general number to be rather normal. Although being far more than in fiction, it fell between relative frequency of proper nouns in informational texts and journal texts (newspapers). The over-estimation of proper nouns showed up at the introductory parts of the letters, where capitalised word "Armas" stands before name of the recipient. The English equivalent would be "Dear" at the very beginning of letter. Well, "Armas" is a first name used in Estonia, but it is a extremely rare one, and it was really never used as such in this correspondence. We adjusted the workflow accordingly by switching off forced proper-name finder, and with some decline of the number of proper nouns, got rid of analysis of "Armas" as name. How much did it affect the recall of proper names, needs some further examination.

As it was mentioned before, both of the authors used foreign language in their letters as well. Just now it is difficult to tell, how much, because the Estonian morphological analyser was too eager to find suitable analyses. It would be possible to apply some language detecting tool before morphological analysis, but it is possible that foreign phrases are not so numerous and this strategy would generate too much noise.

8   For conclusion

The number of words and sentences in this corpus is rather big to annotate it entirely manually, but at least by some extent it would be necessary. Without manually annotated data it would be difficult, if not impossible, to evaluate automatic annotation that is trained on nowadays texts. We are going to annotate

at least some parts of the texts manually to find better methods to improve automatic analysis.

Such archival materials as a private correspondences of writers have a great cultural value, but promise an incredible linguistic importance as well. The elaborated search possibilities of KORP may be used to verify objectively, via linguistic analysis, the conclusions which have so far been drawn by using traditional methods of literary research, mainly using the well known method of slow close reading of the texts.

Application of corpus linguistic methods in literary studies based on archival sources requires meticulous preparation of the material and transforming it into a text corpus. The text corpus of the correspondence of the writers, translators and friends Johannes Semper and Johannes Barbarus allows to go further with principally new type of research questions, about the networks of the writers, or the ideas, that have influenced the original works of both authors, leading more abstract semantic research. This could be the next phrase of the work with KORP.

References

1. Borin, L., Forsberg, M., Roxendal, J.: Korp the corpus infrastructure of språkbanken. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (may 2012)
2. Hardie, A.: CQPweb combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics **17**(3), 380–409 (2012). https://doi.org/doi:10.1075/ijcl.17.3.04har
3. Laak, M., Viires, P.: Digital culture as part of estonian cultural space in 20042014: current state and forecasts (2016), https://www.kogu.ee/vana/wp-content/uploads/2015/06/EIA-ENG_OK-1.pdf, last accessed 2018-10-31
4. Orasmaa, S.: Towards an integration of syntactic and temporal annotations in estonian. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (may 2014)
5. Orasmaa, S., Petmanson, T., Tkachenko, A., Laur, S., Kaalep, H.J.: Estnltk — nlp toolkit for estonian. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)
6. Rand, E.: A third of Estonias cultural heritage to be available digitally in five years. Press release (2018), https://www.kul.ee/en/news/third-estonias-cultural-heritage-be-available-digitally-five-years
7. Rheault, L., Beelen, K., Cochrane, C., Hirst, G.: Measuring emotion in parliamentary debates with automated textual analysis. PLOS ONE **11**(12), e0168843 (dec 2016). https://doi.org/10.1371/journal.pone.0168843

8. Schreibman, S., Siemens, R., Unsworth, J. (eds.): A New Companion to Digital Humanities. John Wiley & Sons, Ltd (dec 2015). https://doi.org/10.1002/9781118680605

9. Veskis, K., Liba, E.: Automatic tagger evaluation. NLP course assignment report (2008), https://entu.keeleressursid.ee/public-document/entity-7052, last accessed 2018-10-31

10. Viires, P., Laak, M.: Digital humanities meet literary studies: Challenges facing estonian scholarship. In: Mkel, E., Tolonen, M., Tuominen, J. (eds.) DHN Helsinki 2018. Book of Abstracts (2018), https://www.helsinki.fi/sites/default/files/atoms/files/dhn2018-book-of-abstracts.pdf, last accessed 2018-10-31