doi:10.5128/ERYa17.14

QUOTE EXTRACTION FROM ESTONIAN MEDIA: ANALYSIS AND TOOLS

**Dage Särg, Karmen Kink,
Karl-Oskar Masing**

Abstract. This paper describes the identification, adaptation and creation of tools that are needed for creating a quote extractor for Estonian media texts that would be able to properly extract both direct and indirect quotes and attribute them to the correct person identified by full name and profession. This includes named entity recognition and resolution as well as grammar-based extraction of direct and indirect quotes. To get a further understanding of indirect speech in Estonian media, we also performed a corpus linguistic analysis of the quotes extracted with our tools from one week of Estonian news.*

Keywords: quote extraction, indirect speech, named entity recognition, information extraction, corpus linguistics, computational linguistics, Estonian

1. Introduction

The aim of this paper is to give an overview of the experiments conducted for creating an automatic quote extractor for the Estonian language. Quote extraction is a relevant topic both in linguistics and other fields. From the linguistic point of view, the main question is how language users use the language to express that the message is someone else's and not their own. While there do exist grammatical descriptions of how a language (e.g. Estonian) enables the user to convey indirect speech, automatic quote extraction allows us to study actual language use both quantitatively and qualitatively.

In addition to linguistics, indirect speech is also an important research object in journalism and scientific writing, where automatic extraction of quotes and citations allows researchers to process significantly larger amounts of data compared to manual extraction. Finally, extracted quotes can be used for analysis in almost any field where it is important to know what opinions different participants hold,

* This study was supported by STACC OÜ and by Eesti Keeletehnoloogia Riiklik Programm (National Programme for Estonian Language Technology) under the project EKTB14.

be it political science, public health, education, or any other field that gets enough attention, whether in journalism or in focused research papers.

This paper focuses on the work done for developing an automatic quote extractor for Estonian news texts. The goal of the extractor would be to extract quotes together with the name of the citee – the person being cited – as well as the person's profession, if possible. This goal is derived from the interests of our partner Baltic Mediamonitoring Group (BMMG), for whom it is important to find out when and how different persons and organizations are cited.

The rest of the paper is organized as follows. Section 2 gives an overview of related works on indirect speech and quote extraction. Section 3 describes the data and tools that were used for this study. As we want to extract some extra information besides quotes, Section 4 focuses on named entity extraction and normalization and Section 5 on profession extraction from the texts. Section 6 describes the extraction and evaluation of direct and Section 7 of indirect speech quotes. This is followed by an analysis of quotes extracted from one week of Estonian news in Section 8, and the paper then ends with a conclusion.

2. Related research

While extracting quotes has been a relatively well-studied topic for bigger languages such as English (Pareti et al. 2013, Pareti 2016, Papay, Padó 2019, 2020) or German (Schricker et al. 2019, Jannidis et al. 2018, Tu et al. 2019), to our knowledge, there have been no previous attempts for Estonian. Therefore, in the related works overview, we focus on various approaches used for other languages that have inspired us the most.

One of those is the paper by Andrew Salway et al. (2017) that describes a methodology for Norwegian quote extraction and attribution. Salway et al. had a goal of extracting quotes of a limited set of politicians from Norwegian bokmål newspapers to provide social science researchers a new means of research. They aimed to extract both direct and indirect quotes. For this, they composed a list of speech verbs and then used the subjects and complements of the speech verbs to identify the speaker and the indirect speech. For speakers, they used a look-up table with variants of the politicians' names and their genders. Salway et al. achieved a very high precision of 95.9% and a recall of 57.0%.

For English, there exist at least two annotated corpora for quote extraction, the Penn Attribution Relations Corpus (PARC, Silvia Pareti 2016) of news texts and RiQuA (Papary, Padó 2020) of literary texts. These open the possibilities of training machine learning models without the researchers having to label the data themselves. For example, Chris Newell et al. (2018) have used the PARC corpus to train and evaluate the state-of-the-art Convolutional Neural Network (CNN) classifier for verb cues. For quote content and source, they use CRF classifiers. They find that while classifiers are reliable for easy and straightforward sentences, more complex grammatical structures lead to errors. The overall precision of their quote extractor is 62.1% and recall 52.2%. Arun Chaganty and Grace Muzny (2020) have also tried neural networks for English quote attribution and found that they did not outperform or even perform as well as more traditional approaches.

While we are not aware of any automatic quote extraction systems for Estonian, there do exist linguistic works describing indirect speech. Aino Admann has described the phenomenon in Estonian (Admann 1975, 1976) and EKS (2017: 689–694) also provides a contemporary overview of Estonian direct and indirect speech. In addition, Denys Teptiuk has thoroughly studied new quotative indexes in Estonian as well as in other Finno-Ugric languages (Teptiuk 2019a, 2019b).

3. Data and tools

For our experiments, we are using the data from Estonian newspapers received in different ways. To develop and evaluate our tools, we extracted data from the corpus Estonian Web 2019 (Kallas, Koppel 2020). An exception was the final evaluation of indirect quote extraction, for which we used manually collected news articles from the web (see Section 7.2).

To perform a final analysis of a set of extracted quotes, we used data provided by our partner BMMG. To keep the computational resources and time needed to perform our experiments under control, we used data of one week: we extracted a dataset of 14,395 news articles from the first week of May 2020. The nature of this data is very similar to the news data contained in the Estonian Web 19 (Kallas, Koppel 2020) corpus as it includes news texts from the same main sources. Therefore, we assume that the results of the analysis would not be significantly different if performed on any other corpus of the same timeframe, but we opted for the data provided by our partner instead of collecting it on our own to save time on data scraping and cleaning issues.

To perform the experiments, we used the EstNLTK (Orasmaa et al. 2016) toolkit of Python libraries for processing Estonian. For this work, as the processing speed was not of utmost importance, we used the latest version 1.6 (1.6.6b during the implementation of our experiments) of the toolkit (Laur et al. 2020), which is not optimized yet and therefore slower but more accurate than the older and more widely used version 1.4.

4. Named entity recognition and resolution

For named entity recognition, we used the CRF-based tool developed by Tkachenko et al. (2013) that is included in the EstNLTK toolkit. It extracts three types of named entities – persons (PER), organizations (ORG), and locations (LOC). It has been trained on 572 news articles from the two biggest Estonian online newspapers, Delfi and Postimees. Its overall F1-score is 87%, but it works significantly better for detecting person names and locations compared to the detection of organization names. As we are primarily interested in PER labels, its precision for those is 90.2% and recall is 91.6%. While the citee can also be an organization, for ORG labels, the tool's precision is 80.0% and recall as low as 74.7%.

In addition to recognizing personal names from a text, for citation extraction, it is also important to be able to join the possible variants of one person's name into one entity. For a morphologically rich language like Estonian, this starts from proper lemmatization. Instead of the default lemmatization offered by the EstNLTK toolkit, we are using a corpus-based approach described by Kaalep et al. (2012) which considers the other lemmas appearing in the corpus in addition to the one we are looking at at the moment. Kaalep et al. (2012) have found that this approach manages to solve 52% of proper name lemma ambiguities with precision of 97%, therefore, using it can only increase the overall performance of the system.

In addition to lemmatization, person names can also be abbreviated by using only a part of the name. In Estonian media texts, it is possible to use either a first name or a surname alone, depending on the context. There are many Estonian names that can serve both as a first name and as a surname, therefore, it is not possible to say with confidence which one it is if just one name is mentioned. However, if the same name is mentioned in the same text as part of a multi-token name, we can derive if it is a first or last name from that.

With this approach, we managed to correctly disambiguate 61.0% of one-word names and there were no false positives in our testset. Out of the one-word person names that could not be disambiguated, 47.8% were actually not persons but locations, organizations, or not named entities at all. This means that this approach could also be used for post-corrections of person names.

5. Profession extraction

In addition to names, professions are commonly used in the media texts to refer to the speaker, especially when it comes to more prominent professions like president, ministers, etc. Knowing the professions of the people who have been quoted in a text also helps to determine the topic of the text and to decide about its relevance for further analysis. To extract professions from the text, we developed a combined approach of rule-based matching and CRF that is described in the following subsections.

5.1. Creating a professions lexicon

The first step for profession extraction was to create a comprehensive lexicon using the existing resources. On the one hand, we used data that we were able to scrape from the web pages of different authorities, e.g. Estonian Unemployment Insurance Fund, Estonian Qualifications Authority, etc. On the other hand, we used dedicated language resources, namely Estonian Wordnet (Orav et al. 2018) and Estonian word2vec models[1].

As there is no dedicated synset for all the professions in Wordnet, we looked through the direct hyponyms of the ‘person’ (*inimene.n.01*) synset and made a list of relevant synsets, e.g. ‘professional’ (*professionaal.n.01*), ‘intellectual’ (*intellektuaal.n.01*), ‘farmer’ (*põllumees.n.01*), etc. We extracted all the lemmas of the hyponyms of these synsets and extended the list recursively by adding hyponyms of existing elements until nothing could be added. This way, we achieved a list of almost 3000 different professions.

[1] https://github.com/eleriaedmaa/embeddings (26.10.2021).

However, as Wordnet definitely does not contain all the possible professions, we decided to use a word2vec model (Mikolov *et al.* 2013) to expand the list with words appearing in similar contexts. For this, we used the professions extracted from Wordnet as positive examples, and extracted the 5000 most similar words from a 100-dimensional CBOW model[2] trained on the Estonian Web 2013 corpus (Muischnek 2016).

The extracted words contained a lot of noise and therefore needed additional filtering. As the Estonian language has a very productive system of compound word formation, we decided to filter out the words whose last token was already appearing in our Wordnet list - e.g. we already had *dirigent* 'conductor' in our list, and therefore *abidirigent* 'assistant conductor' was also added from the word2vec list. We repeated this process twice as after the second iteration, it no longer gave relevant words.

Finally, after some semi-automatic cleaning of the lists, we joined the scraped professions and the professions from Wordnet and word2vec, and had a list of 5659 professions. To our surprise, the overlap between the scraped list and the list extracted from language resources was only about 500 elements.

5.2. Retraining the CRF-based NER tool for professions

While our composed list was quite comprehensive, it was clear that it cannot contain all the possible professions – especially because new ones emerge from time to time, be it because of new industries (e.g. *vlogija* 'vlogger') or just more specific names for already known professions (*insener* 'engineer' → *andmeinsener* 'data engineer'). Therefore, we decided to retrain the CRF-based NER tool described in Section 4 to extract professions.

To retrain the model, we relabelled the existing NER corpus with professions from our professions list. As can be anticipated, some professions were very frequent while most of them did not appear in the corpus even once. The automatic relabelling was also not perfectly accurate. We found two main sources of errors – ambiguous words (e.g. *juht* – either 'driver' or 'case') and proper names (e.g. *President* – head of state or a coffee brand, *Kalamees* meaning 'fisherman' or being a surname). However, due to the lack of resources, we did not perform any manual post-corrections on the relabelled corpus.

To understand if training the CRF model on our corpus gave any improvement compared to lemma-based tagging of professions, we performed an evaluation on a set of 600 paragraphs from different types of news texts from spring 2020. The results can be seen in Table 1, which shows that both the precision and recall of the list-based approach are considerably higher than the ones of the retrained CRF. However, the combined approach where we consider all words that have been marked by either the list-based tagger or the CRF as professions works the best. While its precision is not as good as for the list-based approach only, the huge increase in recall also increases the F1-score by 8.2%.

[2] http://193.40.33.66/pretrained/cbow_100_5_10_20.zip (26.10.2021).

Table 1. Results of profession extraction

	List-based	CRF	Combined
Precision	91.1%	81.5%	86.8%
Recall	68.1%	51.6%	85.4%
F1-score	77.9%	63.2%	86.1%

This means that while the CRF model was not able to learn to detect all the professions, which was anticipated, as most of them did not appear in the training corpus, it did learn to detect new professions in addition to the ones in the list. This is illustrated by the fact that it detects professions that were marginally or not at all represented in the news before but became very important during the COVID-19 emergency situation when our test data was extracted. For example, professions like *viroloogiaprofessor* 'professor of virology', *kommunikatsiooniekspert* 'communication expert', and *hotellitöötaja* 'hotel employee' were not present in our professions list but were detected by the retrained CRF model.

When it comes to the professions that were not detected, many of them were rather untraditional, like *raskerokkar* 'hard rocker' or *šamaan* 'shaman', and it could be argued whether they even are professions or not. The only frequently occurring profession that was not detected was *poliitik* 'politician', other missed professions were mostly compound words that occur rarely, e.g. *juhtivprokurör* 'lead prosecutor' or *tegevprodutsent* 'executive producer'.

Among the false positives, the list-based approach brought only the ambiguous words mentioned previously, while the CRF model identified some proper names as professions, which can be solved by comparing the current CRF annotations with the ones done before adding professions and removing the words that overlap with proper names. In addition, it incorrectly learned some names of industries (*ehitussektor* 'construction sector') as professions, and, following the example of *linnapea* 'mayor' but literally 'head of town' some words that resemble professions but are not, like *maapea* – from the expression *maapeal* 'on the earth' but literally 'head of land'; *tainapea* – 'fool' but literally 'head of dough'. While those issues could be solved by labelling a bigger corpus, we found the accuracy to be good enough for our purposes for now.

6. Extraction of direct quotes

6.1. Methodology

The most intuitive task in quote extraction is to extract direct quotes that are placed inside quotation marks. Of course, as has also been reported by Salway et al. (2017) for Norwegian, direct quote extraction is not as straightforward as only looking at quotation marks, as in addition to quotes, there can also be titles, foreign words, slang words, and other items denoted by quotation marks in the texts.

Fortunately, in Estonian, if the reporting clause precedes a direct quote, there is a colon after the reporting clause (1). When the reporting clause comes after the quote, there is either a comma, an exclamation mark, or a question mark before the ending quotation marks (2).

(1) Kohtunik küsis üle: “Te rääkisite põlemisest?”
judge asked over you spoke.PST.PL3 burning.ELAT
‘The judge reconfirmed: “Did you speak about burning?”’

(2) “Praegu töö jätkub,“ lisas peaminister.
now work continues added prime-minister
‘“Right now, the work goes on,” added the Prime Minister.’

Therefore, thanks to the punctuation, we were able to extract direct quotes using a simple regular expression based approach on an already sentence-segmented text: the content of the quotation marks was tagged as a direct quote if preceded or followed by a suitable punctuation mark. The rest of the first/last sentence of a detected quote was tagged as a reporting clause.

6.2. Evaluation

We evaluated the results on a 55,000-word test corpus that contained 118 texts with 393 direct quotes from 4 of the biggest Estonian news websites: postimees.ee, delfi.ee, ohtuleht.ee, and err.ee, extracted from the Estonian Web 2019 corpus (Kallas, Koppel 2020). For precision, we manually checked all the text segments labelled as direct quotes by our system. For recall, we checked the text segments placed inside quotation marks that were not labelled as quotes by our system, as by definition, direct quotes should be included in quotation marks. As was expected, the precision and recall for direct quotes are very high (see Table 2), and the only direct quotes not detected were the ones with some technical issues. The ones that were detected partially (e.g. due to the use of quotation marks inside the quote or the rare cases where the reporting clause is placed in the middle of a quote) were still counted as true positives, they made up ~5% of the true positives.

Table 2. Results of quote extraction

	Direct	Indirect
Precision	97.5%	93.5%
Recall	92.7%	77.4%
F1-score	95.0%	84.7%

7. Extraction of indirect quotes

7.1. Methodology

We decided to use a grammar-based approach to extract indirect quotes despite the fact that indirect speech can be reported in a variety of ways. This decision was motivated by several factors. The most important, of course, is the fact that we did not have any labelled training data for Estonian. Therefore, to train machine learning models, we would need to start by labelling data. On the other hand, if we manage to build a grammar-based system with reasonable precision, we can use its output as training material for future machine learning systems that would not need to follow specific rules but would look at textual similarities.

In addition, there is no good proof yet that machine learning would be more successful in this task. While neural models have taken over in several fields, e.g. syntactic parsing and text classification, the breakthrough in quote extraction is yet to be achieved even for languages that do have training data. For example, Newell *et al.* (2018) found that their CNN classifier detected only the most simple constructions but more rare ones went undetected. Using a grammar-based approach is very useful in this case as it allows us to find a variety of cases, especially since unlike labelled training data, we do have thorough theoretical grammar descriptions for quotes in Estonian that were mentioned in Section 2.

Based on the theoretical sources, as well as our own linguistic intuition, we addressed three types of indirect quotes:

1) indicated by a reporting verb together with a relevant conjunction (3);
2) indicated by a noun together with a relevant conjunction (4);
3) indicated by specific reporting constructions (5).

For type 1, the defining features are a reporting verb and a conjunction. To get a list of possible reporting verbs, we first used the same methodology as for composing the list of professions described in Section 5.1 with the use of Wordnet and word2vec for extending a small, manually composed list.

(3) Minister ütles, et olukord on lootusrikas.
Minister said that situation is hopeful
'The minister said that the situation is promising.'

(4) Ta saatis teate, et ta ei tule.
he/she sent message that he/she not come
'He/she sent a message that he/she wouldn't come.'

(5) Ministri sõnul on olukord lootusrikas.
Minister words.ADE is situation hopeful
'According to the minister, the situation is promising.'

In addition, following the example of Salway et al. (2017), we used the direct quotes extracted in a relatively simple way that was described in the previous section as input for extracting the more complicated indirect quotes. The idea is that, mostly, the same verbs are used for both direct and indirect speech. We extracted the verb lemmas from reporting clauses of direct quotes, and made a frequency list. As can be anticipated, from the top of the list we found the verbs that we already had received from Wordnet and word2vec.

However, we looked through the end of the list together with usage examples from our corpus and made some additions. In total, we obtained a 240-word list of verb cues to use to extract reporting clauses of indirect speech.

Type 2 indirect quotes are similar to type 1 but the defining features are a suitable noun together with a conjunction. For type 3, we relied on literature as well as our own observations to get a finite set of possible other reporting constructions. In total, we considered 8 nouns and 10 types of other constructions indicated either by an adverb or a noun in a specific case.

7.2. Evaluation

We evaluated the results of indirect quote extraction on a corpus of 69 Estonian news texts from different online news agencies from August 2020. The set was hand-picked by our colleagues, whom we asked to find news that contained at least one indirect quote. We provided them with a few examples to illustrate what we were looking for but let them decide on their own what an indirect quote was. This approach was used instead of picking random articles to ensure that we could evaluate the recall in a reasonable way on the not very sizable corpus that we were going to label manually.

The corpus was in total 29,300 words (punctuation included) and the individual news texts ranged from 37 words up to 2,941 words, with an average of 430 words. It was manually labelled by one linguistically trained and experienced annotator who labelled 495 indirect quotes. Our system detected 417 indirect quotes, 27 of which were incorrect. This gives us a precision of 93.5%. 105 quotes labelled by the annotator were undetected by the system, resulting in a recall of 78.8% and an F1 score of 85.5%.

The false positives contained ambiguous words and expressions that can be used for reporting speech but also in other meanings. In (6), *leidma* ‘find’ could be used to report someone’s speech, but here it is used in the sense of agreeing on a meeting time. This could be solved by using syntactic information about the sentence in addition to morphological and lexical information. A more complicated case is (7). When someone’s feelings are reported, without semantic knowledge, we are not aware if the person communicated those feelings verbally or they are just assumed by the reporter.

(6) Teen ettepaneku leida aeg, et ta saaks
make.1SG proposal.GEN find time that he/she could
kohtumisel osaleda.
meeting.ADE participate
‘I propose to find a time so that he/she could participate at the meeting.’

(7) Kui lehm midagi näeb, mis tekitab temas tunde,
when cow something sees which makes she.IN feeling
et tuleks piima anda, siis too hormoon vallandubki.
that should milk.PART give then that hormone releases
‘When a cow sees something that makes her feel that she should give milk, then this hormone is released.’

Most of the false negatives – the indirect quotes that our system was not able to detect – were cases where the distance between the verb-cue and the conjunction was too large, like in (8). This issue could be solved by clause segmentation, as the distance maximum was set in place to reduce the amount of false positives where the conjunction is not in the clause that directly follows the verb or noun cue.

In addition, there are some types of constructions that our approach did not consider. (9) is a sentence with a verb that can be used for quoting someone without a conjunction-led subordinate clause but including the quote as a direct object or an adverbial. We also did not include phrasal verbs in our approach where the verb on its own cannot be used to quote someone. Those issues could be solved by expanding the construction types that our system is able to detect.

(8) Neljapäeval teatasid Eesti suured
Thursday.ADE announced.3PL Estonian large.PL
ehitusfirmad Nordecon ja Merko
construction-companies.PL Nordecon and Merko
Tallinna Börsile, et ehituses
Tallinn.GEN Stock-Exchange.ALL that construction.IN
läheb paremini.
goes better
'On Thursday, big Estonian building companies Nordecon and Merko informed Tallinn Stock Exchange that the construction industry is doing better.'

(9) Harris süüdistas Bidenit kahtlases koostöös.
Harris accused Biden.PART suspicious.IN collaboration.IN
'Harris accused Biden of suspicious collaboration.'

Overall, we find that while the precision and recall for indirect quotes are much lower than the direct quotes, they are good enough to perform an analysis of indirect speech on a slightly larger corpus that is presented in the next section.

8. Overview of indirect speech in Estonian news

In this section, we give an overview of indirect speech in Estonian news from both the corpus linguistic point of view as well as from the practical needs of quote extraction. We are using the 14,395-article corpus of news texts received from our partner BMMG that was described in Section 3.

From the corpus, our system extracted a total of 61,467 quotes. 64% of the extracted quotes were indirect quotes and 36% were direct. While based on the evaluations presented in two previous sections we know that we are not able to automatically find about 7% of direct and 23% of indirect quotes, these numbers illustrate why we still have to pay a lot of attention to indirect quotes, despite them being more complex.

8.1. Lengths of quotes and reporting clauses

Out of the extracted direct quotes, 88% were the type where the quote is followed by the reporting clause while only 12% contained the reporting clause first followed by the quote. For indirect quotes, the distribution was the other way round: the first type made up 18% of the quotes and the second one 82%. This probably means that direct speech is mostly used when the emphasis is on the message as this is put first, while indirect speech is used to set the focus on the source of the message.

The word counts of the reporting clauses and quotes for all the types are presented in Table 3. From all the word counts here and later, punctuation has been excluded as otherwise commas, colons and quotation marks would influence the counts too much. As can be seen from Table 3, for both direct and indirect quotes, the reporting clause of the more frequent type is shorter than the other.

Table 3. The average length of a reporting clause and a quote (word count). 'Quote-first' signifies the cases where the quote comes before the reporting clause (as shown in (2)), 'Quote-last' signifies the opposite (1)

Quote	Quote-first		Quote-last	
	Reporting clause	Quote	Reporting clause	Quote
Direct	4.3	23.2	9.0	23.8
Indirect	7.6	13.4	6.5	12.8

For quotes, there is not that much difference in length depending on where the reporting clause is positioned. However, there is a large difference between the lengths of direct and indirect quotes. This is due to the extraction methodology as well as how information is reported. While for direct quotes, we included all the sentences inside quotation marks, for indirect quotes, we considered only one sentence at a time. This was a conscious decision and resulted from the fact that while direct speech is limited by quotation marks, there is no such marker for indirect speech. Therefore, if indirect speech is reported via several sentences, each of those sentences should signify again that it is still indirect speech. Thus, if the indirect speech is reported in several sentences, we extract this as separate quotes.

The distribution of the length of reporting clauses over all the types can be seen from Figure 1. It shows that, as can be expected, frequency is negatively correlated with length. 2-word reporting clauses make up almost a third (32.2%) of all the reporting clauses. They are most prominent among direct quote-first quotes where they make up more than half (55.9%) of reporting clauses for this type. This comes down to the most common reporting structure containing a reporting verb and a one-word subject.

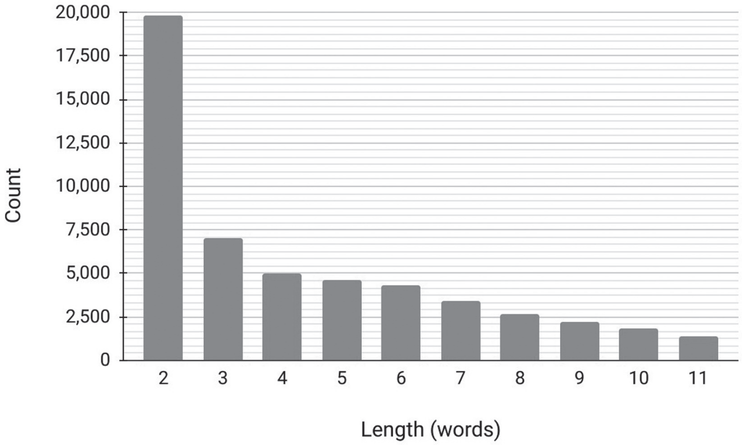

Figure 1. Distribution of length of reporting clauses

8.2. Subjects in reporting clauses

It is clear that for proper quote attribution, a high-accuracy co-reference resolution is needed. While 57.2% of 2-word reporting clauses do contain a proper noun for a subject, the fact that we are looking at 2-word clauses means that most likely, the proper noun is expressed by only one token. We did address this issue in Section 4, showing that 61.0% of those cases in our data can be solved by looking at the other named entities in the same text.

For further analysis, we only considered reporting clauses of 7 words or less because otherwise it is probably not one clause but a compound sentence. As can be seen from Figure 1, these make up about 75% of all the reporting clauses. 60.2% of those have a proper noun for a subject, but for the rest there is either a common noun, a pronoun, or no subject at all.

11.8% of reporting clauses have a common noun for the subject. These common nouns are mostly professions like *juhataja* 'manager', *president* 'president', *minister* 'minister', *prokurör* 'prosecutor', etc, but also just markers of gender: *mees* 'man' and *naine* 'woman'. In addition, there are a few words denoting organizations that cannot be attributed to a single person like *valitsus* 'government', *ettevõte* 'company', *politsei* 'police', *ajaleht* 'newspaper'.

A pronoun is used as a subject in 13.3% of the studied reporting clauses. This is more common in shorter clauses: more than a quarter of 2-word reporting clauses have a pronoun as a subject. This means that pronoun resolution would be needed to successfully attribute the quotes. Unfortunately, while there have been some studies on pronoun resolution for Estonian (e.g. Puolakainen 2015, Freienthal 2020), there is still no usable tool for that. On the positive side, as it is mostly the pronoun *tema* 'he/she' that is used in reporting clauses, we would only need to solve this subtask of pronoun resolution.

14.7% of reporting clauses do not have a subject. These can be divided into two groups. The first group are clauses where the source of the message is expressed in the locative case, mostly denoting a written piece of information like *pressiteade* 'press release', *avaldus* 'statement', *kiri* 'letter'. The second group are clauses where the verb ending shows that it is the first person and the subject has been omitted – those sentences come mostly from opinion pieces and can be both in singular or plural. While the singular means the author of the text, the plural can mean anything from the author plus one to all of mankind.

8.3. Verbs in reporting clauses

In our extracted reporting clauses, 198 different verbs were used at least five times. The most common verbs used are, as can be anticipated, the neutral ones: *ütlema*, 'say', *rääkima* 'tell', *lisama* 'add', *sõnama* 'utter', *selgitama* 'explain', *märkima* 'note' etc. While we limited the set of verbs that we considered possible for indirect reporting, the verbs for direct reporting clauses were not limited as their use is much more productive. This appears also in the news corpus. From the end of the frequency list, we find emotionally loaded verbs like *kurjustama* 'scold', *kiruma* 'curse', *ironiseerima* 'ironize', *nukrutsema* 'grieve', *lajatama* 'burst out', *tusatsema* 'pout', *vinguma* 'whine'. In addition, as direct speech does not require the verb of the reporting clause to be related to talking as strictly as indirect speech, there also appear verbs that are related to the actions or emotions of the speaker and not the speaking itself, like *jooksma* 'run' in (10) or *särama* 'shine' in (11).

(10) Korraga jooksis väravatest stjuardess:
suddenly ran gates.PL.ELAT stewardess
"Kas teie soovisite sõita Moskvasse?"
did you wish.PST.PL2 ride Moscow.ILL
'Suddenly, a stewardess ran through the gates: "Did you wish to fly to Moscow?"'

(11) "Kirjutate, et see valmistaks teile rõõmu,"
write.PL2 that this make.COND you.PL.ALL joy.PART
särab Ljudmila.
shines Ljudmila
'"You write that it would make you happy", beams Ljudmila.'

86.1% of the extracted reporting clauses contain a verb. The rest are either direct clauses stating only the speaker's name with a colon or indirect clauses of the third type that could more accurately be labelled as 'reporting constructions' instead of 'reporting clauses'.

When there is a verb in a reporting clause, in 61.7% of cases it is in the past tense. As there is no grammatical future in Estonian, the rest are in the present tense. Most often the verb is in singular, but in 6.0% of the clauses it is in plural and in 4.4% in impersonal form. Impersonal is generally used to quote written pieces of information, and the source of the message could be extracted by looking at the locative phrase.

9. Conclusion and future work

In this work, we made the first attempts in creating the tools and resources that are needed for building a reliable quote extractor for Estonian media texts that would extract both direct and indirect quotes together with the name of the quoted person and their profession.

First, we used the standard CRF-based Estonian NER-tagger from the EstNLTK toolkit and added a disambiguator that finds the person's full name from the text if the reporting clause only contains a partial name (a first or a last name). This way, we were able to solve 61.0% of the partial names. This could be improved by adding a corpus-based disambiguator that, if the ambiguity cannot be solved based on the text, uses corpus-based probabilities to find the full name.

Second, we created a lexicon of Estonian profession words, and using that lexicon, relabelled the standard NER training corpus with professions. By evaluating the profession extraction results, we concluded that it is best to use a combined approach of retrained NER together with the professions list. Profession extraction, as well as named entity recognition in general, could be improved by labelling more training data or using another machine learning approach instead of CRF. While our attempts in NER and profession extraction were concentrated only on Estonian, the approach, in principle, could also be used for other languages.

Third, we explored the possibility to extract both direct and indirect quotes. For direct quotes, we used a simple regular expression based approach. As direct quotes in Estonian are clearly marked by punctuation, this approach gave us a high F1 score of 95% that was mostly influenced by technical or spelling issues.

From the reporting clauses of direct quotes, we got input to extract indirect quotes based on verb, noun, and adverb cues. We used a grammar-based approach to detect indirect quotes and received a good precision of 93.5% and a recall of 77.4%. While the recall is far from perfect, exploring the false negatives from the manually labelled test set, we found several types of constructions that could be added to the grammar to improve the results. In addition, the approach could be used to create a training set for machine learning approaches in a semi-automatic manner.

Finally, as the precision of both direct and indirect quote extraction was fairly good, we performed an analysis of reporting constructions on one week of Estonian news texts, exploring the lexical, morphological, syntactic and semantic traits of reporting clauses.

In the future, in addition to the improvements of the individual tools that were mentioned in the previous subsections, this study would hopefully lead to a full pipeline of Estonian quote extraction and attribution in the EstNLTK toolkit.

References

Admann, Aino 1975. Otsene kõne ja saatelause [‘Direct speech and quotative clause’]. – Emakeele Seltsi aastaraamat, 9–20 (1973–1974), 63–71.

Admann, Aino 1976. Otsese ja kaudse kõne segavormidest eesti kirjakeeles [‘On the mixed forms of direct and indirect speech in the Estonian literary language’]. – Emakeele Seltsi aastaraamat, 21, 71–79.

Chaganty, Arun; Muzny, Grace 2015. Quote Attribution for Literary Text with Neural Networks https://cs224d.stanford.edu/reports/ChagantyArun.pdf (24.9.2020).

EKS 2017 = Erelt, Mati; Metslang, Helle (Eds.). Eesti keele süntaks [‘The Syntax of Estonian’]. Eesti keele varamu 3. Tartu: Tartu Ülikooli kirjastus.

Freienthal, Linda 2020. Pronominaalsete viitesuhete automaatne lahendamine eesti keeles närvivõrkude abil [‘Pronominal coreference resolution in Estonian with neural networks’]. Master’s thesis. Tartu: University of Tartu. http://hdl.handle.net/10062/67637

Jannidis, Fotis; Zehe, Albin; Konle, Leonard; Hotho, Andreas; Krug, Markus 2018. Analysing direct speech in German novels. – Digital Humanities im deutschsprachigen Raum. Kritik der digitalen Vernunft. Konferenzabstracts. Köln: Universität zu Köln, 114–118.

Kaalep, Heiki-Jaan; Kirt, Riin; Muischnek, Kadri 2012. A trivial method for choosing the right lemma. – Arvi Tavast, Kadri Muischnek, Mare Koit (Eds.), Human Language Technologies – The Baltic Perspective. Frontiers in Artificial Intelligence and Applications 247. IOS Press, 82–89. https://doi.org/10.3233/978-1-61499-133-5-82

Kallas, Jelena; Koppel, Kristina 2020. Eesti keele ühendkorpus 2019. Center of Estonian Language Resources. https://doi.org/10.15155/3-00-0000-0000-0000-08489L

Laur, Sven; Orasmaa, Siim; Särg, Dage; Tammo, Paul 2020. EstNLTK 1.6: Remastered Estonian NLP Pipeline. – Proceedings of The 12th Language Resources and Evaluation Conference (LREC’20). European Language Resources Association (ELRA), 7152–7160.

Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey 2013. Efficient estimation of word representations in vector space. – Proceedings of Workshop at ICLR 2013. https://arxiv.org/pdf/1301.3781.pdf (25.9.2020).

Muischnek, Kadri 2016. Eesti veeb 2013 (etTenTen) korpus, morfoloogiliselt ühestatud. Center of Estonian Language Resources. https://doi.org/10.15155/1-00-0000-0000-0000-0012EL

Newell, Chris; Cowlishaw, Tim; Man, David 2018. Quote extraction and analysis for news. – DSJM, August 2018, London, UK. https://research.signal-ai.com/assets/RnD_at_the_BBC__and_quotes.pdf (24.9.2020)

Orasmaa, Siim; Petmanson, Timo; Tkachenko, Alexander; Laur, Sven; Kaalep, Heiki-Jaan 2016. EstNLTK - NLP Toolkit for Estonian. – Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), 2460–2466.

Orav, Heili; Vare, Kadri; Zupping, Sirli 2018. Estonian Wordnet: Current state and future prospects. – Proceedings of the 9th Global WordNet Conference. Global Wordnet Association. compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/GWC2018_paper_37.pdf (25.9.2020).

Papay, Sean; Padó, Sebastian 2019. Quotation detection and classification with a corpus-agnostic model. – Proceedings of Recent Advances in Natural Language Processing. Varna, Bulgaria, September 2–4. INCOMA Ltd, 888–894. https://doi.org/10.26615/978-954-452-056-4_103

Papay, Sean; Padó, Sebastian 2020. RiQuA: A corpus of rich quotation annotation for English literary text. – Proceedings of the 12th Conference on Language Resources and Evaluation (LREC'20). European Language Resources Association (ELRA), 835–841. https://aclanthology.org/2020.lrec-1.104

Pareti, Silvia; O'Keefe, Tom; Konstas, Ioannis; Curran, James R.; Koprinska, Irena 2013. Automatically detecting and attributing indirect quotation. – Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 989–999.

Pareti, Silvia 2016. PARC 3.0: A corpus of attribution relations. – Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), 3914–3920.

Puolakainen, Tiina 2015. Anaphora resolution experiment with CG rules. NODALIDA 2015. https://www.hf.uio.no/iln/om/organisasjon/tekstlab/aktuelt/arrangementer/2015/nodalida15_submission_99.pdf (25.9.2020).

Salway, Andrew; Meurer, Paul; Hofland, Knut; Reigem, Øystein 2017. Quote extraction and attribution from Norwegian newspapers. – Proceedings of the 21st Nordic Conference on Computational Linguistics. Association for Computational Linguistics, 293–297.

Schricker, Luise; Stede, Manfred; Trilcke, Peer 2019. Extraction and classification of speech, thought, and writing in German narrative texts. – Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019). German Society for Computational Linguistics & Language Technology, 183–192.

Teptiuk, Denys 2019a. New quotatives in Finnish and Estonian. – Finnisch-ugrische Mitteilungen, 42, 207–249.

Teptiuk, Denys 2019b. Quotative Indexes in Finno-Ugric (Komi, Udmurt, Hungarian, Finnish and Estonian). Dissertationes philologiae uralicae Universitatis Tartuensis 21. Tartu: University of Tartu Press. http://hdl.handle.net/10062/66597

Tkachenko, Alexander; Petmanson, Timo; Laur, Sven 2013. Named entity recognition in Estonian. – Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria, 8–9 August 2013. Association for Computational Linguistics, 78–83.

Tu, Ngoc Duyen Tanja; Krug, Markus; Brunner, Annelen 2019. Automatic recognition of direct speech without quotation marks. A rule-based approach. – Digital Humanities: multimedial & multimodal. Konferenzabstracts. Universitäten zu Mainz und Frankfurt, 87–89.

TSITAATIDE ERALDAMINE EESTIKEELSETEST MEEDIATEKSTIDEST: ANALÜÜS JA TÖÖVAHENDID

Dage Särg, Karmen Kink,
Karl-Oskar Masing
Tartu Ülikool

Artikkel annab ülevaate eesti keele tsitaadituvastaja loomise esimesest etapist. Tsitaadituvastaja eesmärk on eraldada nii otseses kui kaudses kõnes väljendatud tsitaate koos tsiteeritud isiku täisnime ning võimalusel ka ametiga. Artiklis selgitasime, milliseid komponente tsitaadituvastaja jaoks oleks vaja ning vastavalt sellele testisime ja kohandasime olemasolevaid ning lõime veel puuduvaid töövahendeid. Samuti identifitseerisime tsitaadituvastaja arenduseks vajalikud parandused ja lisatööriistad ning analüüsisime uudistes otsese ja kaudse kõne edastamiseks kasutatavaid saatelauseid.

Isikunimede leidmiseks kasutasime EstNLTK teegi standardset CRF-põhist nimeüksuste märgendajat. Lisasime sellele ühestaja, mis leiab tekstist tsiteeritud isiku täisnime, juhul kui saatelauses on kasutatud ainult ees- või perekonnanime. Sel moel suutsime ära lahendada 61,0% ühesõnalistest isikunimedest.

Elukutsete leidmiseks lõime 5659 sõna suuruse eestikeelse ametite leksikoni ning märgendasime selle põhjal nimeüksuste tuvastaja treeningkorpuses ka elukutsed. Seejärel treenisime nimeüksuste tuvastaja ümber tuvastama ka elukutseid. Tulemusi hinnates leidsime, et kõige parem on kasutada leksikonipõhist lähenemist koos ümbertreenitud CRF-märgendajaga, mis andis elukutsete tuvastamise F1-skooriks 86,1%.

Otsekõne eraldamiseks kasutasime regulaaravaldistepõhist lähenemist. Kuna otsekõne on jutumärkidega selgelt markeeritud, saime sel moel 95,0%-se F1-skoori. Otsekõne saatelausetest saime sisendit kaudse kõne eraldamiseks: lõime verbide, nimisõnade ning määruste leksikoni, mis viitavad, et lause edastab vahendatud mõtteid. Grammatikapõhise lähenemisega saime F1-skooriks 84,7%, seejuures oli täpsus 93,5%. Uurides tuvastamata jäänud kaudse kõne lauseid, leidsime veel mitut tüüpi konstruktsioone, mida saagise parandamiseks grammatikapõhises lähenemises käsitleda võiks.

Lõpetuseks analüüsisime vastloodud töövahendite abil ühe nädala Eesti meediatekstidest eraldatud tsitaate ja nende saatelauseid, käsitledes nii leksikaalseid, morfoloogilisi, süntaktilisi kui ka semantilisi jooni.

Tulevikus on plaanis peale eraldiseisvate töövahendite parandamise luua ka vabalt kasutatav terviklahendus eestikeelsete tsitaatide ja tsiteeritute tuvastamiseks.

Võtmesõnad: tsitaatide tuvastamine, vahendatud kõne, nimeüksuste tuvastamine, info eraldamine, korpuslingvistika, arvutilingvistika, eesti keel

Dage Särg is a PhD student in computational linguistics at the University of Tartu. She is also working as an NLP data scientist in the machine learning and data science company STACC OÜ. Her research interests include automatic information extraction, dependency parsing, and processing of spontaneous unedited written texts. She has participated in various research projects, including the development of the EstNLTK toolkit, as well as designed and taught a course on natural language processing at the University of Tartu.
Narva mnt 18/20, 51009 Tartu, Estonia
dage.sarg@ut.ee

Karmen Kink, MSc, is an NLP data scientist at STACC OÜ. Her research interests include semantic text representations, transfer learning methods, named entity recognition and linking, and automatic summarization.
Narva mnt 18/20, 51009 Tartu, Estonia
karmen.kink@stacc.ee

Karl-Oskar Masing is an MSc student in computer science at the University of Tartu. He is the team lead of the NLP group at STACC OÜ, with prior experience in data science, data engineering, and scientific computations. His research interests include automatic information extraction, text algorithms, text retrieval, and business processes. He has participated in various research projects, contributed to the development of the EstNLTK toolkit, and taught various courses at the University of Tartu.
Narva mnt 18/20, 51009 Tartu, Estonia
karl-oskar.masing@stacc.ee