Contemporizing 20-th Century Estonian

Heiki-Jaan Kaalep
University of Tartu
Tartu, Estonia
heiki-jaan.kaalep@ut.ee

Abstract

The paper describes a contemporization effort of a 1.9 million word corpus of Estonian parliament minutes from 100 years ago. The paper describes the corpus of Asutaw Kogu (the Constitutional Assembly) and the main differences of language that require one to contemporize it for modern researchers. The effort is implemented as a work flow that combines a freely available speller lexicon, hand-crafted transformation rules and various corpus-based word lists into finite state transducers. Evaluation on a 53,000 token subset of the corpus showed that 0.02% of text tokens ended up with an incorrect contemporary form, corresponding to 0.05% of the corpus vocabulary. However, if we count only the tokens that actually need changing in the contemporization process, we see that 0.12% end up being incorrect, corresponding to 0.15% of the corpus vocabulary. An additional experiment with generative AI showed that using it as a contemporization tool results in a content-preserving, but more formal version of the original minutes.

Keywords: Estonian, finite-state transducer, historical corpus

1. Introduction

Digitization and text recognition (OCR or typing in) creates possibilities for a researcher to "1) perform a full-text search in their digitized corpora, or 2) to perform distant reading: using Natural Language Processing (NLP) tools, automatically process documents in order to make analyses over large corpora. One major obstacle to this research design is language change. The older the historical texts are, the more they diachronically deviate from the current standard language." (Ehrmanntraut, 2024)

The current paper is about contemporizing the language used by the founders of the Republic of Estonia a hundred years ago – language used in the Asutaw Kogu 'the Constituent Assembly' (AK).

A contemporary researcher who would like to find passages in the transcripts that might be relevant for the research question, is confronted with the question: what linguistic expression(s) should one use for querying? Formulating a query requires an hypothesis about how speakers might have used language to express the notions that the researcher is interested in.

Choosing between the linguistic options requires some intuition about the language – its orthographic norms, vocabulary, inflectional and derivational morphology. It is unrealistic to expect a contemporary researcher to have good intuition about a past language. So, it would be nice to have some oracle providing guidance, e.g. providing modern orthographic variants and lemmas of historical words. The oracle could be seen to provide modern paraphrases of the original text.

By way of terminology one should note that it would be inaccurate to use the term "normalize" for describing the changes necessary for making near-past Estonian similar to contemporary one, because it was already a standard normative language, albeit one that is different from the one used today.

2. Related Work

Approaches to historical text normalization can be divided into type- or lexicon-based and token- or text-based ones. A lexicon-based approach tries to utilize the observation that orthographies tend to have more or less regular sound-letter mapping conventions. It has been common for this approach to write rules that correspond to historical changes and variations in these mappings, notably using weighted finite state transducers (FST) (Beesley and Karttunen, 2003), e.g. (Porta et al., 2013) for historical Spanish, (Etxeberria et al., 2016) for historical Basque, (Koskenniemi and Kuutti, 2017) for historical Finnish. A text-based approach views the task as a special type of machine translation. This has evolved to using LLMs, e.g. (Bawden et al., 2022) for early French, (Bracke, 2025) for historical German.

Makarov and Clematide (2020) experimented with historical German, English, Spanish, Icelandic, Portuguese, Hungarian, Slovene and Swedish, representing periods between 14th and 19th centuries. They found that the upper bound on accuracy for a non-contextual oracle that selected the most frequent normalization for each historical word was 97.0 on average, indicating that lexical normalization is a very reasonable strategy.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 400–406.

3. Corpus of Asutaw Kogu

The Constituent Assembly was the first parliament of Estonia. It acted as the national representative body and the legislative power of Estonia from April 23rd 1919 until December 20th 1920. Its task was to lay the foundations for the Estonian statehood, to adopt the Constitution and the Land Act. The elections for the Constituent Assembly were held from 5th to 7th April 1919.

The Constituent Assembly held five sessions in 170 sittings, in which 88 laws were passed. During their term of office, the members of the Constituent Assembly also sent greetings to foreign states, discussed the problems connected with research and culture, followed the events in foreign policy, and expressed their opinion about them.[1]

In 2009, students at the Graduate School of Humanities and Sociology, University of Tokyo, under the guidance of Prof. Kazuto Matsumura, created a corpus of Asutaw Kogu by typing in the published minutes of the 170 sittings (Kogu, 1919-1920) (Asutawa Kogu protokollid I-V), a total of 1.93 million words. The corpus is freely available.[2]

Before the 1920-ies, three local languages – Estonian, German and Russian – were widely spoken and understood. Some members of Asutaw Kogu used their constitutional right to give speeches in their native German or Russian in Asutaw Kogu. No translation for these speeches was provided because everyone understood the languages (Raag, 2008, p. 167). These speeches are not included in the corpus.

Corpus text is divided into paragraphs <p> and sentences <s>. The sentences are numbered. Speakers, agenda items, and passages in foreign language are tagged. The original publication presented text in a two-column format. The corpus contains tags <col> indicating the start of a new column; columns are numbered.

4. Deviations from Contemporary Estonian in AK

Over the course of the last 100 years, Estonian has changed. The changes involve pronunciation, morphology, vocabulary and syntax, all having an effect on spelling. A speller of modern Estonian from Filosoft[3] or Giellatekno[4] is able to recognize 57,000 inflectional word forms out of the total 85,000 that make up the word stock of AK, i.e. 67%; this corresponds to 1,600,000 tokens out of the total 1,930,000 corpus tokens, i.e. 83%.

- The most notable change has been purely orthographic – using *v* instead of *w*. This change started in 1905 and ended in the 1930-ies, mainly happening during 1920-ies. Ignoring *v/w* differences, a modern speller would recognize 76,000 or 89% of the vocabulary; 1,870,000 or 97% of the corpus tokens. In other words, it would miss 9000 (11%) of vocabulary items, corresponding to 60,000 (3%) corpus tokens.
- Another purely orthographic change is using *š* and *ž* instead of *sh*.
- Spelling change corresponding to change in pronunciation has been: using *f* instead of *w*, e.g. *wormaalsus* → *formaalsus* ‘formality’, using a double letter instead of a single one (to indicate a long sound), or, vice versa, a single one instead of a double, e.g. *bilans* → *bilanss* ‘balance’, *mehanism* → *mehhanism* ‘mechanism’, *inseneer* → *insener* ‘engineer’, *kommisjon* → *komisjon* ‘commission’.
- Most frequent changes in suffixes have been the following:
  - *ismus* → *ism*, e.g. *radikalismus* → *radikalism* ‘radicality’
  - *iker* → *ik*, e.g. *tehniker* → *tehnik* ‘technician’, *poliitiker* → *poliitik* ‘polititian’
  - *lik*, *line* → *lik*, *line*, *ne*, e.g. *sõjalik* → *sõjaline* ‘military’, *sümpaatlik* → *sümpaatne* ‘likeable’, *loodusline* → *looduslik* ‘natural’, *üleilmline* → *üleilmne* ‘world-wide’
  - *iteet* → *sus*, e.g. *agressiviteet* → *agressiivsus* ‘agressiveness’, *stabiliteet* → *stabiilsus* ‘stability’
  - *mata* → *matu*, e.g. *kõlbmata* → *kõlbmatu* ‘useless
- Some inflectional classes have changed, meaning that all the words that belong to certain class are nowadays inflected (partly) differently than 100 years ago.
  - Plural partitive allomorph of suffix *-kond* has changed from *e* to *i*, e.g. *osakonde* → *osakondi* ‘departments’, plural partitive allomorph *a* has been substituted by *e*, e.g. *kasulikka* → *kasulikke* ‘useful’, *punkta* → *punkte* ‘points’
  - Plurality allomorph of words ending with *-ik* has changed from *ui* to *e*, e.g. *kodanikuile* → *kodanikele* ‘to citizens’

[1] https://www.riigikogu.ee/en/introduction-and-history/history-riigikogu/constituent-assembly/

[2] https://www.cl.ut.ee/korpused/baaskorpus/akp/

[3] https://github.com/Filosoft/vabamorf

[4] https://github.com/giellalt/lang-est-x-utee/

  - Plurality allomorphs *i*, *u*, and *e* have been substituted by *de* as the default allomorph for several inflectional classes, e.g. *rühmis* → *rühmades* 'in groups'
  - Tens of individual words have moved from one inflectional class into another, e.g. *liige* 'member' plural comitative case form has changed from *liigetega* to *liikmetega* 'with members'.

- Haplology has been lost in impersonal forms of *da/ta*-ending verbs, e.g. from *parandakse* to *parandatakse* 'is being repaired'.
- Words have shortened, in particular losing *-eer-*, e.g. *kontroleeri* → *kontrolli* 'control', *analüseeri* → *analüüsi* 'analyze' and *-ne-*, e.g. *tutvunema* → *tutvuma* 'get acquainted', *valminema* → *valmima* 'ripen'.
- Tens of high-frequency words have changed, e.g. a productive first component of compound words has changed from *nõnda* to *nii* 'so', e.g. *nõndapalju* → *niipalju* 'so much'; *siiamaalne* changed to *senine* 'until now', *otstarb* to *otstarve* 'purpose', *ütelus* to *ütlus* 'saying', *kaebtus* to *kaebus* 'complaint'.

5. Operationalization with FST

Mapping historical variants to modern ones is the task of mapping strings, and as such similar to mapping a spelling variant to a norm or an erroneous word to a correct one. Ordering string manipulation rules according to their likelihood, allowing/prohibiting certain rule configurations, and checking the correctness of resulting strings would be a complex task for a programmer; an example of such an approach to process non-standard Estonian is presented by (Kaalep and Muischnek, 2011). Managing this combinatorial complexity becomes much easier with FSTs, because they offer a natural way of avoiding the need to explicitly deal with all of the combinatorics. The implementation described in the current article is available at TartuNLP repository[5]

It is natural to treat a string mapping FST as a composition of two transducers $E \circ V$, where E is an edit transducer and V is a vocabulary that acts as a filter that restricts the set of output strings, e.g. the vocabulary of contemporary Estonian. E is essentially a collection of letter and string pairs representing various ways in which one could modify the input string. The smaller the weight attached to a letter or string pair, the more plausible this change is. In the notation of the HFST toolkit (Lindén et al., 2011) used in this work, an example excerpt from E would be: `[w (->) v::10] .o. [ri (->) n::300]`. The example is a composition `.o.` of two optional mappings `(->)` with vastly different weights `10` and `300`. Composing the mappings means that they are all applied, one by one, from left to right. Given an input string *wabariema* containing substrings *w* and *ri*, this example transducer would yield four output variants, with weights 0, 10, 300 and 310. Composition of this example with V representing modern Estonian vocabulary would result in a FST that filters out all the rest but the last one – *vabanema* 'get free' with weight 310 – as a valid modern Estonian word.

The task of converting the vocabulary of AK to modern Estonian can be formulated as a task of building a transducer that maps every word in AK corpus to its modern counterpart: $A \circ E \circ V$, with A being the vocabulary of AK. Its simplest final form would be a two-column table with every vocabulary item as one row; it is trivial to convert this to the LEXC format (e.g. `wabariema:vabanema # ;`) which in turn is the preferred format for (morphological) lexicons to be compiled into FSTs.

We know A, i.e. the left or input side of the resulting transducer – it is the vocabulary of AK, but we do not have complete knowledge of V, i.e. contemporary vocabulary with lexical coverage that would include modern equivalents for all vocabulary items of AK.

The challenge here is to achieve the necessary lexical coverage, in addition to building the edit transducer E. Fortunately, one can download open source Estonian morphology files from the Giellalt repository[6]), build the morphological analyzer FST, extract its input (i.e. surface) side (discarding the output, i.e. lexical side containing lemma and grammatical category info), and thus get a weighted transducer S that accepts all the word forms a speller is able to recognize.

The weights of S have the following explanation. A surface word form may be ambiguous as to its grammatical categories, and the morphological analyzer FST attaches a weight to every possible reading. The weight is calculated by morphological properties: the weight of the lexeme (based on its rank in a frequency dictionary), plus weights of the inflectional categories (singular, plural, nominative, genitive etc.), plus a weight for every derivational affix, plus a penalty weight for any component stem in case the word is a compound word. This complex calculation scheme gives a context-ignorant likelihood for every morphological analysis variant of a word form. This is what an FST morphological transducer outputs as a default. Having discarded the lemma and grammatical info, we are not inter-

[5] https://github.com/TartuNLP/ak_kaasaegseks_LT4HALA/

[6] https://github.com/giellalt/lang-est-x-utee/

ested in all the readings a word form may have, and choose the variant with the smallest weight, saying that this is the weight of the contemporary word form itself.

6. Correcting errors in the AK

Re-typing in has resulted in errors because of the poor quality of the print, and/or the fact that the typists were students of Estonian as a foreign language, not native speakers.

In addition to keyboarding errors (substitutions, omissions, insertions, transpositions), the corpus contains errors similar to those one encounters with OCR, because the typists have misread typographically confusable letters like *l*-*I*-*i*-*t*, or *rn*-*m*. Accented letters *ä*, *ö*, *õ* and *ü* have been often confused with their un-accented counterparts.

Prior to contemporizing, an attempt to correct the corpus semi-automatically was made. How do you know that a token is actually a spelling error (and thus you cannot transform it into a contemporary word)?

The following assumptions seem justified: 1. Spelling errors are rare; the same error does not happen twice in the same word form. 2. A word form that has an error is unique exactly because of this error; the correct form is not unique, i.e. has a frequency greater than one.

However, being unique does not mean that the word form is erroneous.

So, a unique word form might either contain an error or be a genuinely rare word, and one needs to find out which alternative to choose.

It is actually very difficult to decide whether a word form is part of the vocabulary of a language, and thus could possibly be also part of the vocabulary of a given corpus. If a word is common, it is usually listed in some dictionary, and once one has access to this dictionary, it is easy to look it up. However, no dictionary contains all the words of a language.

Approximately half of the vocabulary of a text corpus normally consists of words that occur only once, i.e. hapax legomena; the exact proportion is dependent on the corpus size (Baayen, 2001). This number of unique words in a text corpus is also approximately equal to the number of unknown text tokens, i.e. words that are missing from any lexicon built from some other corpus, e.g. a speller's word list.

To sum it up: taken together, rare words form a large part of the vocabulary of a corpus. One should be cautious when suggesting that some rare word form is actually an error. Not wanting to introduce any new errors in the corpus (by wrongly assuming that a rare word form should be replaced by a more frequent one), one needs to assess what is more likely: is this a rare word in that corpus, or a mis-spelled form of some other word.

Notice that I rely only on type frequencies and context-independent transformations of a single word form.

For assessment, I create two transducers $E \circ S$ and $E \circ A2$ by composing an edit FST E with S, an FST representing the vocabulary of a speller, and with $A2$, an FST containing the set of non-unique word forms of the AK corpus (with their weights equal to zero). Those two transducers will yield different outputs with weights to the same input string. The candidate with the smaller weight wins.

I allow maximum two changes per word that is at least 4 characters long, and one change per shorter words.

The list of hapax legomena comprises 42,000 words. 1070 of them can be converted to some non-hapax word from, by changing up to 2 letters, i.e. can be recognized by $E \circ A2$. However, 470 of them would be recognized by a modern speller in their original form already (i.e. by S), so these are excluded.

Additionally, 160 are recognized by $E \circ S$, while having a smaller weight than those recognized by $E \circ A2$. This means that the unique corpus word is more similar to a modern word than to a more frequent corpus word and thus it is too risky to assume this is an error.

Half of these 160 words are instances where the unique corpus word is an inflectional form that is missing from this corpus, while a more frequent similar form of the same lexeme is present. E.g. for a lexeme *elav* 'living', the corpus contains a unique singular allative case form *elawale*, its contemporary form would be *elavale*, but the most similar corpus non-unique form is plural genitive *elawate*. The other half are rare words which look almost like some other word, e.g. unique word *wälkus* 'flickered', non-unique word *waikus* 'silence'.

Finally, 440 corpus words get from $E \circ A2$ a weight smaller than they get from $E \circ S$, e.g. unique word *riiglkodanikka*, contemporary *riigikodanikke*, corpus non-unique *riigikodanikka* 'citizens'. These 440 are thus good candidates for being substituted by a more frequent corpus word. However, manual inspection reveals that one group of these candidates is susceptible to being erroneous.

For 380 of these, the changes needed for getting a contemporary word are the same that one needs to get a non-unique corpus word, plus possibly a change from *w* to *v*. So the modern speller and the corpus agree on what the correct form of these words should be. There is no error in this group.

For another 30, there is no way to get a contemporary word form, because they are either family names (which are normally not a part of a lexicon), or word forms very different from contempo-

rary language. E.g. unique form *Strändman*, corpus non-unique form *Strandman*, or unique form *kakskümmendvüs*, corpus non-unique form *kaksümmendviis*, the contemporary form containing a space: *kakskümmand viis* 'twenty five'. This group is very error-prone: 6 instances, i.e. 20% of candidates should not be substitued by a similar-looking corpus non-unique word form, because the unique word form is really a rare instance of a correct word, not an error.

The rest 30 are cases where the possible contemporary word contains more changes or less likely ones than the non-unique corpus form, e.g. unique word *Tonis*, contemporary *Topis* 'stuffed', corpus non-unique *Tõnis* (proper name); unique word *rnaterjaali*, contemporary *materjali*, corpus non-unique *materjaali* 'material'. There are no errors in this group.

Why is the second group so much less reliable? Paradoxically, the fact that a word form can be transformed into a correct contemporary word form, lends credibility to the idea that it might actually be just a typing error. Proper names and word forms, which differ greatly from contemporary language, violate the initial assumption that they are just a few key strokes away from a more frequent, correct word.

7. Creating a Conversion FST

Contemporization in our case means that words are substituted by modern ones, and everything else (word order, layout, mark-up) remains unchanged.

As mentioned earlier, the FST for mapping AK words to contemporary ones would be $A \circ E \circ V$. As the first approximation, I take $V = S$.

If one is unable to find a contemporary version of a word, keep the old version. This is more likely to happen with proper names, acronyms and abbreviations, and typographical errors.

The question is thus how to arrive at the FST V containing a contemporary version for every vocabulary item of AK corpus.

The steps would be the following.

1. Create a frequency list of AK corpus words, stripping them of punctuation, and splitting also at hyphens, but without downcasing. The resulting lexicon has 85,000 items.

2.1. Split compound words which are nowadays written with a space, e.g. **kõigeõigem* -> *kõige õigem* 'most correct'.

2.2. Contemporize AK words with $E \circ S$. The final total weight of a word form (after it has passed through this transducer) is a sum of two components: summary of weights of letter changes from E plus weight of the contemporary word form from S. The larger the weight, the less trustworthy the result is (because it is a combination of rarer events).

3. Can I be sure that the output of the transducer $E \circ S$ is really a contemporary version of the original, and not some wildly different word? One might make a lot of changes, to eventually arrive at some acceptable word with a large weight. Obviously, it would be sensible to discard words above some threshold.

However, a word with a large weight might also be a rare inflectional form of a rare compound, the weight from S indicating the unlikely process of its morphological creation, and not the complex way of changing the original by E to arrive at it. It would be nice if I knew that this word is a legitimate word in our corpus; the contemporization weight would then reflect solely the likelihood of changes. So I should include contemporized versions of words that surely exist in AK corpus (S_{AK}) in V, thus $V = S \cup S_{AK}$. I assume that a word shorter than 3 letters, or a contemporized word with a small weight is correct, and thus belong to S_{AK} (with a weight of zero).

4. Proper names should be contemporized, but more cautiously than other words: only the most common orthographical changes (*w* -> *v*, *sh* -> *š*) are allowed (E_{min}). The proper name transducer would be $P_A \circ E_{min}$, where P_A denotes proper names in AK.

When making a list of proper names of AK, I took a position to favor precision over recall. A proper name is a word with an initial capital letter, which occurs sentence-internally, and which has no lowercased version in the corpus. The latter condition is necessary because an orthographic sentence may contain a part starting with a capitalized word, e.g. direct speech, a comment in parenthesis etc. I identified 3100 proper names this way.

5. So the final transducer is $((A - P_A) \circ E \circ (S \cup S_{AK})) \cup (P_A \circ E_{min})$

Using this final transducer revealed that 3000 entries of the 85,000 lexicon were not converted to a plausible contemporary word. When making the final 85,000-entry 2-column table prior to converting it to LEXC format and compiling into a transducer for contemporizing the corpus texts, these 3000 entries were included as simply pairs of identical strings.

They can be grouped in the following way: abbreviations and acronyms, typos (often caused by an inserted or a deleted space character), old inflectional or derivational forms (not covered by mappings of the edit transducer E), loanwords with outdated orthography, proper names which had not been classified as such by the method described above, and some lexemes which contemporized form is missing from the modern speller lexicon.

It should be noted that I have found only one old lexeme that has diverged into more than one modern lexeme – *näitus* 'exhibit' into *näitus* 'exhibition'

and *näide* ‘example’. At first glance, it seems impossible to have a simple 1-to-1 conversion in this case. However, on closer examination, it appears that different inflectional forms of *näitus* in the AK corpus are used in different senses, e.g. *näituseks* ‘for example’ would be contemporized as *näiteks*; *näitusel* ‘at an exhibition’ would be contemporized as *näitusel*.

8. Evaluation

For evaluation, the minutes of 5 sittings (9, 11, 12, 31 and 74) were chosen randomly and checked manually. They contain 53,000 tokens (2.8% of the corpus text volume), with a vocabulary of 10,000 (11.75% of the corpus vocabulary). Both type-in errors and contemporarization errors were checked by comparing with the original PDF-files.

The final transducer contains a few instances where a word with a type-in error gets correctly contemporized, e.g. *wõtrna*, contemporary *võtma* ‘to take’. However, in the checked files, a type-in error always resulted in failure to get a correct contemporary word.

The number of letter type-in errors was found to exceed the number of contemporization errors very slightly. There were 14 word-internal type-in errors – 7 involving a letter, and 7 being failures to delete a hyphenation mark induced by end of line; versus 11 contemporization errors. These 11 errors involved 5 different words, among them three proper names *Westholm*, *Wiedemann* and *Kurs-Olesk* that should not be changed. It is noteworthy that re-typing contributed as much to the overall error rate as did making the words contemporary. In addition, there were 60 type-in errors of changing, adding or deleting a punctuation mark.

The evaluation shows that 0.05% of the resulting vocabulary is populated by incorrectly contemporized items, and the same is true for 0.02% of the text tokens. The large amount of punctuation mark errors came as a surprise. However, this had no impact on the procedure of contemporizing, because the method is word based, and ignores punctuation marks. It will probably have an impact on some follow-up steps of the corpus processing workflow that operates on the contemporized texts, e.g. POS-tagging and lemma disambiguation, but this has not been evaluated.

Remember that 67% of the vocabulary and 83% of the corpus tokens are similar to modern Estonian. This means that only 3300 vocabulary items and 9000 text tokens of the evaluation set are in need of contemporization; thus 0.15% of vocabulary items needing contemporization end up being incorrect, and 0.12% of the corpus tokens.

9. Contemporizing with Office 365 Copilot

A natural idea for transforming one text into another is to use a generative AI model. I made a small experiment with Office 365 Copilot. I gave Copilot two 50-sentence-long extracts. One extract is a part of a continuous speech by a member of AK. The other extract is part of a procedure for adopting a law, where the items are read out loud, commented on briefly, and voted on. I prompted Copilot with "*Alljärgnev tekst on vanas eesti keeles. Teisenda see tänapäevaseks eesti keeleks.*" ‘The following text is in old Estonian. Transform it into modern day Estonian’.

When comparing the AI generated texts with the ones where only words had been contemporized, I found that 1) Copilot preserved the contents of the original, 2) Copilot converted individual words into modern Estonian correctly, 3) Copilot paraphrased a lot, substituting words and expressions with modern ones, reordering text fragments, and changing punctuation and text layout.

As a result, the output by Copilot feels like it has made a modern edited version of the minutes. The text retains the meaning of the original, presenting it in a more formal way, not keeping the original way the speakers expressed themselves.

Several runs with the same texts and prompt were made. The runs were sometimes immediately after each other, sometimes days apart. The resulting texts were very different from each other, but always kept the original meaning and expressed it in the form of modern meeting minutes. It is noteworthy that not a single instance of misrepresenting the original content was found during these runs.

Assuming future AI experiments will confirm the preservation of content when changing the presentational form to a modern one, one could envisage that contemporization would boil down to creating another corpus with AI, parallel to the original AK, and linked to it. This heavily edited new version would be usable for someone who intends to read the original closely but needs help navigating the vast corpus to find the part that needs close reading.

10. Conclusion

A freely available speller lexicon, hand-crafted transformation rules, and various corpus-based word lists were combined into a work flow for contemporizing a corpus of Estonian spoken and written 100 years ago. Evaluation showed that 0.02% of text tokens ended up with an incorrect contemporary form. However, if we count only the tokens that actually need to be changed in the contemporization process, we see that 0.12% end up being

incorrect.

11. Acknowledgements

This work has been supported by the Estonian Research Council grant PRG2006 and by the national program "Estonian Language and Culture in the Digital Age" project EKKD-TA10.

12. Bibliographical References

R Harald Baayen. 2001. *Word frequency distributions*, volume 18. Springer Science & Business Media.

Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. 2022. Automatic Normalisation of Early Modern French. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 3354–3366, Marseille, France. European Language Resources Association.

Kenneth R Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI, Stanford.

Yannic Bracke. 2025. A lexical normalizer for historical spelling variants using a transformer architecture. https://github.com/ybracke/transnormer. GitHub repository.

Anton Ehrmanntraut. 2024. Historical german text normalization using type- and token-based language modeling. *arXiv preprint arXiv:2409.02841*.

Izaskun Etxeberria, Iñaki Alegria, Larraitz Uria, and Mans Hulden. 2016. Evaluating the noisy channel model for the normalization of historical texts: Basque, Spanish and Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1064–1069, Portorož, Slovenia. European Language Resources Association (ELRA).

Heiki-Jaan Kaalep and Kadri Muischnek. 2011. Morphological analysis of a non-standard language variety. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 130–137.

Asutaw Kogu. 1919-1920. *Asutawa Kogu I-V istungjärk. [Sessions 1 to 5 of the Constitutional Assembly]*. Täht, Tallinn, Estonia.

Kimmo Matti Koskenniemi and Pirkko Kuutti. 2017. Indexing old literary finnish text. In *K + K = 120: Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays*. Research Institute for Linguistics, Hungarian Academy of Sciences.

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A. Pirinen, and Miikka Silfverberg. 2011. Hfst—framework for compiling and applying morphologies. In *Systems and Frameworks for Computational Morphology*, pages 67–85, Berlin, Heidelberg. Springer Berlin Heidelberg.

Peter Makarov and Simon Clematide. 2020. Semi-supervised contextual historical text normalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7284–7295, Online. Association for Computational Linguistics.

Jordi Porta, José-Luis Sancho, and Javier Gómez. 2013. Edit transducers for spelling variation in old spanish. In *Proceedings of the workshop on computational historical linguistics at NODALIDA*, pages 22–24.

Raimo Raag. 2008. *Talurahva keelest riigikeeleks [From Peasant Language to State Language]*. Atlex, Tartu, Estonia.