RiCL Research in Corpus Linguistics

The burden of legacy: Producing the *Tagged Corpus of Early English Correspondence Extension* (TCEECE)

Lassi Saario[a] – Tanja Säily[a] – Samuli Kaislaniemi[b] – Terttu Nevalainen[a]
University of Helsinki[a] / Finland
University of Eastern Finland[b] / Finland

Abstract – This paper discusses the process of part-of-speech tagging the *Corpus of Early English Correspondence Extension* (CEECE), as well as the end result. The process involved normalisation of historical spelling variation, conversion from a legacy format into TEI-XML, and finally, tokenisation and tagging by the CLAWS software. At each stage, we had to face and work around problems such as whether to retain original spelling variants in corpus markup, how to implement overlapping hierarchies in XML, and how to calculate the accuracy of tagging in a way that acknowledges errors in tokenisation. The final tagged corpus is estimated to have an accuracy of 94.5 per cent (in the C7 tagset), which is circa two percentage points (pp) lower than that of present-day corpora but respectable for Late Modern English. The most accurate tag groups include pronouns and numerals, whereas adjectives and adverbs are among the least accurate. Normalisation increased the overall accuracy of tagging by circa 3.7pp. The combination of POS tagging and social metadata will make the corpus attractive to linguists interested in the interplay between language-internal and -external factors affecting variation and change.

Keywords – corpus annotation; corpus markup; spelling normalisation; TEI-XML; part-of-speech tagging; Late Modern English

1. Introduction[1]

1.1. Legacy corpora

Many of the corpora used to study the history of English have a history of their own. That is especially true of the pioneering corpora from the early 1990s that are still used nowadays, such as the *Helsinki Corpus of English Texts* (HC) and *A Representative*

[1] People who worked in the tagging project besides the authors deserve to be mentioned here. Mikko Hakala, Emanuela Costea, Anne Kingma and Anna-Lina Wallraff were responsible for a large part of the semi-manual normalisation. We would also like to thank Paul Rayson, Jukka Suomela, Turo Hiltunen, Arja Nurmi and Gerold Schneider for their assistance and advice. This work was supported in part by the Academy of Finland, Grants 293009 and 323390.

Research in Corpus Linguistics 9/1: 104–131 (2021). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.09.01.07

Corpus of Historical English Registers (ARCHER). They were originally encoded in a format that was state-of-the-art at the time, but as the years have gone by, the original format has become outdated and incompatible with new tools. This is a common problem among old corpora and the reason why they are called ‘legacy corpora’. While most of them are small in size by present-day standards, the vast amount of qualitative work invested in them still makes them valuable compared to today’s big data corpora which put quantity before quality (see Hundt and Leech 2012; Davies 2019). Legacy corpora deserve to be rescued, then, but how?

The solution is, of course, to convert them into a new format (as has been done to the HC and ARCHER that have been converted into TEI-XML), but that solution is bound to cause new problems. The original compilers of legacy corpora cannot have foreseen the needs of their successors, and the choices made by them in the past (such as the markup schemes chosen or the features omitted from the texts) limit the options available in the present. If the corpus has been based on secondary sources such as printed editions of original manuscripts, the interpretative work done by the editors also sets certain preconditions. In a sense, it might seem easier to start the markup process from scratch. If these problems can be solved, however, the conversion can bring an old corpus back to life again. Not only may the new format be richer than the old one, but it may also allow for further enrichment (e.g. new kinds of annotation) and so broaden the scope of possible research questions that the corpus can shed light on, making it even more valuable than it was before.

In this article, we present a case study of one legacy corpus and the problems related to converting and enriching it, some of which are general while others are specific to legacy corpora. Our case in point is part-of-speech tagging the *Corpus of Early English Correspondence Extension* (CEECE). The CEECE could be classified as a second-generation legacy corpus, as it follows the markup conventions of the HC but comes equipped with substantially richer metadata. It will serve as an example of a legacy corpus that has been successfully ‘rescued’ and enriched with annotation.

1.2. Enrichment of the CEECE

The CEECE opens a window into the sociohistorical variation and change of Late Modern English (LModE) through personal letters, sampled and digitised from published editions (see Kaislaniemi 2018). Plenty of successful studies have been conducted on the corpus since the initiation of its compilation in 2000 (see e.g. Nevalainen *et al.* 2018). Until now, however, the letter texts have remained in largely unstructured form. More sophisticated queries require more structured data where the linguistic features of interest, such as parts of speech, are explicitly annotated. We hope the need for richer data will now be satisfied as we present the new POS tagged version of the corpus, known as the *Tagged Corpus of Early English Correspondence* (TCEECE).

In the original CEECE, text files are accompanied by an external database that contains structured metadata about the letters and the correspondents, whereas the actual letter bodies consist of mostly unstructured text. While the tagging project did not increase the amount of data (defined as the word count), it did enrich the data by both structuring the unstructured and adding more structure on top of the pre-existing. First, the texts were converted into TEI-XML so as to make their internal structure more transparent and well-formed. Second, the texts were tokenised into word elements and, third, each token was assigned a POS tag. From this point of view, POS tagging the CEECE illustrates how a small corpus can be made more valuable —perhaps even more valuable than a bigger corpus which is not as rich.

On the other hand, the enrichment caused complications that had to do with, for example, normalising spelling variation, converting the legacy format and calculating the accuracy of the tagging. We believe these to be common problems among corpus annotators, especially those who are working with historical material or trying to update legacy corpora to the ‘third generation’ (see Hiltunen *et al.* 2017: §3). We hope our experiences will be of use to colleagues wrestling with similar difficulties. We would like the production of the TCEECE to set an example, not only of how heavy the burden of legacy can be but also of how that burden can eventually be overcome.

We will begin with an overview of the history of the corpus, the POS tagging project and the technologies behind it (Section 2). We will then outline the workflow of the project and reflect on critical points (Section 3), followed by a discussion where we look at our choices in retrospective, trying to learn from our mistakes and to come up with suggestions on better policies for others to follow (Section 4). We will conclude with a summary of what we have done and what remains to be done (Section 5).

2. Background

2.1. The CEEC family of corpora

The *Corpora of Early English Correspondence* constitute a digitised corpus family compiled by the Helsinki-based *Sociolinguistics and Language History* team to facilitate systematic sociolinguistic research into the history of the English language. It has grown over the years from the original core corpus of 2.6 million words to a family of subcorpora twice that size.

The original version, the *Corpus of Early English Correspondence* (CEEC), was completed in 1998 and covers the period from circa 1410 to 1681. A half-a-million-word *Sampler* version of the corpus (CEECS) was published in 1999, and the corpus at large in 2006. Due to copyright restrictions, this grammatically annotated published version, the *Parsed Corpus of Early English Correspondence* (PCEEC), is slightly smaller than the original one, comprising 2.2 million words. The original version was supplemented by circa 400,000 words of additional material from 1402 to 1663, packaged as the CEEC *Supplement* (CEECSU, unpublished). Later, the corpus team also extended the CEEC into the eighteenth century, creating the CEEC *Extension* (CEECE), a 2.2-million-word subcorpus, which stretches the timeline covered to 1800, earning the corpus family the acronym CEEC-400 as it covers four centuries (see Table 1).

	CEEC	CEECS	PCEEC	CEECE	CEECSU	CEEC-400[2]
Words	2,597,957	450,082	2,159,132	2,218,520	441,304	5,221,349
Collections	96	23	84	77	19	191
Letters	6,053	1,124	4,970	4,923	857	11,714
Writers	778	194	666	308	95	1,125
Time span	c. 1410–1681	1418–1680	1410–1681	1653–1800	1402–1663	1402–1800

Table 1: The CEEC corpus family

2.2. Choice of tagger

The system of grammatical annotation of the CEEC has a history of its own, which is longer and more complex than that of the CEEC corpus family itself. The PCEEC

[2] CEECS and PCEEC are not counted in the numbers of CEEC-400, being subsets of CEEC. Of the two versions of the Plumpton collection, the newer one (in CEECSU) has been excluded from the total counts.

annotation was carried out by the CEEC team in collaboration with researchers from the University of York, with Arja Nurmi in Helsinki being responsible for the part-of-speech tagging and Ann Taylor at York for the syntactic parsing. To ensure compatibility of diachronic corpora that cover largely the same time period, the same annotation system was chosen as had been used earlier to tag and parse the grammatically annotated versions of the HC, that is, the *Penn-Helsinki Parsed Corpus of Middle English* (PPCME2; Kroch *et al*. 2000) and the *Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME; Kroch *et al*. 2004), which both followed the guidelines of the *Penn Treebank*.[3]

The question of annotation system arose again when plans were made to provide the CEECE with POS tagging. One relevant alternative was to adopt the *Brill* tagger and the *Penn Treebank* tagset used in the *Penn Parsed Corpora of Historical English* and, by doing so, to provide continuity with the POS tagging of the PCEEC. The other alternative, originally also experimented with the HC (Kytö 1996: 5), was to opt for the *Constituent-Likelihood Automatic Word-Tagging System* (CLAWS). This had become the *de facto* standard for corpora made available through the widely used Lancaster University *CQPweb* interface (Hardie 2012), including many Present-day English (PDE) corpora as well as the *Early English Books Online* corpus and the *Corpus of English Dialogues*.[4]

The choice between the two systems depended on a number of factors. As LModE is in many ways close to PDE, comparability between the TCEECE and CLAWS-tagged PDE corpora such as the *British National Corpus* (BNC) and the Brown family of corpora[5] was thought to be advantageous; other LModE corpora had been tagged using various annotation systems, so there was no one model to follow there (Hundt 2014: 2). In terms of tagger performance, the accuracy of the *Brill* tagger on the PCEEC was circa 80–90 per cent (Arja Nurmi, personal communication), which is similar to that of CLAWS on Early Modern English (EModE), although automatic spelling normalisation as a pre-processing step has been shown to improve the CLAWS output (Rayson *et al.* 2007; Hiltunen and Tyrkkö 2013). When applied to present-day corpora, both annotation systems are reported to reach comparable levels of accuracy (c. 96–97%).[6]

Our final decision was reached by considering one more factor, namely the annotation scheme. The *Penn* tagset employed in the PCEEC, designed to be used

[3] See https://www.ling.upenn.edu/hist-corpora/, https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html
[4] https://cqpweb.lancs.ac.uk/
[5] https://varieng.helsinki.fi/CoRD/corpora/BROWN/
[6] See http://ucrel.lancs.ac.uk/claws/. For a comparison of the two tagsets, see Lu (2014: 42–47).

throughout the long diachrony of English, has significant drawbacks compared to CLAWS for the study of more modern forms of English. Analysing noun ratios in the PCEEC, Säily *et al.* (2011) found, for example, that the adverb *likewise* was tagged conservatively as a combination of an adjective and a noun (ADJ+N), identically to the noun *gentleman*. Moreover, the annotation scheme follows Huddleston and Pullum's (2002) analysis of prepositions, collating subordinators and prepositions into a single category, which precludes studying them separately unless the corpus is syntactically parsed (Säily *et al.* 2017: 46). As no syntactic parsing was being planned for the CEECE and, unlike in the PCEEC project, checking all the annotation manually was not an option, CLAWS was chosen as the basis for producing the TCEECE.

2.3. Other technological choices

Once CLAWS had been chosen as the tagger, we had yet to choose from the various tagsets that were available for CLAWS. The prominent options at the time were C5 (62 tags), C7 (137–152 tags) and C8 (170 tags).[7] C7 was an enriched version of C5, and C8 likewise of C7. The native output of CLAWS followed C7 and could automatically be mapped to C5, while enrichment into C8 would have required post-processing by a separate software, *Template Tagger* (Fligelstone *et al.* 1997). For that reason, as well as the fact that the BNC had been tagged using C5 and the BNC sampler using C7, we ended up choosing between C5 and C7.

We found it an advantage of C7 that there was a distinct tag for almost every personal pronoun, while C5 only had one tag for all of them (cf. Säily *et al.* 2017: 46). On the other hand, C7 had unnecessarily fine-grained noun categorisation. We decided to provide the tagged corpus in both tagsets, as the C5 tagging could be derived from the C7 tagging without any cost. Neither did we need to check the accuracy of the two taggings separately, for the checking of C7 could also be directly translated into that of C5. Since the BNC Sampler had been tagged in C7, we could largely rely on the same guidelines in checking the accuracy (see Section 4.1 for a comparison of accuracy between the tagsets).

The original markup of the CEECE (as the CEEC-400 in general) is based on that of the HC (Kytö 1996: §3.3.2; Nurmi 1998: §2), which ultimately dates back to the COCOA program that was used on punched cards and magnetic tapes in the 1960s and

[7] See http://ucrel.lancs.ac.uk/claws/

1970s (see e.g. Russell 1965; Corcoran 1974). Before the corpus could be tagged by CLAWS, it had to be converted into XML. The HC had already been converted into TEI P5 XML (Marttila 2011), so it was only natural that we converted the CEECE into a similar schema (see Section 3.2). The BNC, too, had been converted into TEI-XML and made available on *CQPweb*, which encouraged us to import the TCEECE into *CQPweb* as well.

3. Workflow: Problems and solutions

A thorough documentation of the TCEECE project has been published in the *Corpus Resource Database* (Saario and Säily 2020). Figure 1 illustrates the production process. Instead of redocumenting the process in every detail, we will here focus on the central problems we faced, the solutions we came up with and the lessons we learned from them. Many critical choices had to be made, some of which turned out to have a significant effect on the later working stages and the use of the final corpus. Those choices and their effects, as well as the alternative paths that might (and maybe should) have been taken, will be discussed in more depth in Section 4.

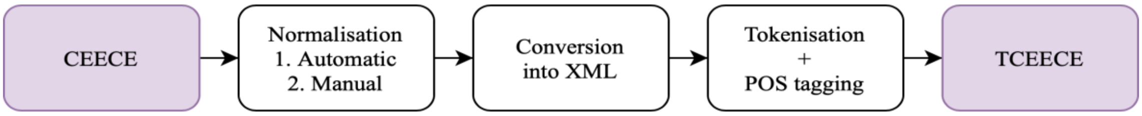

Figure 1: A visualisation of the workflow

3.1. Normalisation

The historical spelling variation in the CEECE was normalised to better comply with present-day standards, so as to make it easier for CLAWS to tag. The first stage of normalisation was performed semi-automatically with UCREL's *Variant Detector* (VARD; Baron 2011a, 2011b) as a part of an earlier project, the creation of the *Standardised-spelling Corpora of Early English Correspondence* (SCEEC). The tagged output of VARD includes both original and normalised-spelling variants inside XML-like tags. The normalised form appears in between the XML tags, while the original variant is kept inside an `orig` attribute:

(1)
```
I would desire you to send me an Oxford <normalised
orig="almanack" auto="true">almanac</normalised>
```

However, it was the untagged output that was moved on to the next stage of normalisation, so that only the normalised forms remained. While omitting the tags did make the text easier to process, losing the original spellings actually ‘impoverished’ the data rather than enriched it, conflicting with the ideal expressed in the introduction. Our choice to produce a POS-tagged version of the corpus that was silently normalised (but retained text-level markup elsewhere) was a compromise between maximal annotation and ease of use. The latter consideration applied both to the people involved in the production process —many of whom were research assistants with no knowledge of XML— and to the end-users of the corpus. Leaving out the original spelling was also not seen as a major issue because the CEEC family of corpora was never designed for the study of orthographic variation. The compilers used original-spelling editions to ensure that the linguistic content would be reliable for morphosyntactic studies, but even these editions frequently normalise features such as *u*/*v* variation, capitalisation or punctuation. Recent work has shown that the CEEC compilers’ reservations towards using the corpora to study spelling, capitalisation, punctuation or word division were largely warranted (Sairio *et al.* 2018; but see Kaislaniemi *et al.* 2017). In any case, the original (editorial) spelling is preserved in the original version of the corpus, so with access to both versions, users are still able to check the spelling, albeit with some difficulty (see Section 4.2 below).

Further normalisation was performed partly manually and partly automatically. Given the variability of historical spelling, even after being processed with VARD, the CEECE texts contained great numbers of tokens not found in PDE. As we did not have the resources to manually normalise all remaining non-standard items, it was decided to focus on the most frequent types, and ones that were easy to identify. The bulk of these were abbreviations, which are commonly marked by punctuation (`Ld.` for ‘Lord’; `desir'd` for ‘desired’), superscripts (coded in CEECE with equal signs: `w=ch=` for w^{ch} ‘which’), or special characters (changed in CEECE to tildes: `com~and` for ‘command’; `lr~es` for ‘letters’; `p~mit` for ‘permit’). Some of the abbreviations in CEECE are still current in PDE, such as *Mrs*, but with formatting that makes them opaque to CLAWS, such as `M=rs=`. In the case of abbreviations not found in PDE, `Sep=br=`, `Sep=t=`, `Septem` and `7=br=` may be intelligible to human readers, but not to CLAWS. And the same applies to otiose abbreviations, mostly marked with superscripts, such as `you=r=` ‘your’ and the ubiquitous `y=e=` ‘the’ —which was particularly tricky when occurring without superscripts, as it needed to be disambiguated from the plural pronoun *ye*. This variability in the spelling and formatting of abbreviations in the CEECE partly reflects manuscript reality, but also the practices of different editors and printers.

In the first cycle of post-VARD normalisation, a concordancer was used to find such items. These were then manually reviewed in a spreadsheet, and those chosen for normalisation were given normalised forms in a separate column. Finally, Python scripts were used to replace the original variants in the texts with the normalised forms. Nearly 8,000 abbreviated words or otherwise non-standard variants were normalised in this way. In the second cycle, the same process was repeated by a different method: a sample of the twice-normalised texts from across the CEECE was run through CLAWS, and problematic items were identified. Scripts were then used to capture and normalise such cases in the whole corpus, to a number of roughly 9,200. Aside from abbreviations, other frequent features requiring such manual attention included punctuation as well as word division in indefinite pronouns (*every body* > *everybody*) and reflexive pronouns (*my self* > *myself*) (see Saario and Säily 2020: §3). The total number of (semi-)manual replacements came to 17,024.

More information about the original text was, of course, lost at this stage, as the variants to be normalised were simply replaced with PDE forms without leaving any trace of the original variants. All text-level encoding that was involved in the original variant was also lost in the process, so that, for example, `fin[{is{]h'd`, where `[{is{]` marks an emendation, was normalised into `finished` where there is no sign of the original spelling nor the emendation. Again, getting rid of that information did streamline the pipeline but it also had unfortunate consequences for the use of the end product, which will be discussed in Section 4.2.

3.2. XML conversion

Throughout the normalisation process, the corpus remained in the ancient COCOA format. The parameter lines that preceded each letter were not a problem, but the letter bodies involved a great deal of custom text-level coding that CLAWS would not have understood (see Saario and Säily 2020: §4.4). Apart from paragraph shifts that were only implicitly indicated, there were 'P-lines' to mark page shifts and various code brackets to mark comments, emendations, etc.[8] The easiest solution would have been to remove all

[8] Special characters (e.g. the pound sign) had already been converted into XML in the normalisation.

text-level coding, which would have lost still more information and further impoverished the data. We wanted to avoid that outcome and decided to convert all the coding into XML in order for it to survive through POS tagging.

Our approach to XML could be characterised as 'modest' in the sense of Hardie (2014). While we did model our XML schema after that of the HC (Marttila 2011) which, in turn, is based on the TEI guidelines, we did not even try to implement all of their potential but only the bare minimum that was required to preserve the encoded information. We also prioritised effectiveness over tidiness and sought to automatise the conversion as far as possible. Despite the modesty of our intentions, several problems arose along the way, the most symptomatic two of which are treated here.[9]

3.2.1. Separating 'proper comments' from 'emendation comments'

Following the HC, editors' comments in the CEECE were originally annotated with the code `[\...\]` and compilers' comments with `[^...^]`. One issue was that both codes were used for two different types of annotations. The same code might be used in, for example, the following two instances:

(2) `reminding him of his obligations and his [\ONE WORD MISSING\]`

(3) `she walked about [\her\] Chamber`

The difference is that in the first instance the comment is a meta-level remark about the body text, whereas in the second instance it is an editorial addition that is meant to be read as a part of the text like an emendation (which are encoded as `[{...{]`). We call the two uses a 'proper comment' and an 'emendation comment', respectively.

The two uses of the same code had to be recognised and separated in order for CLAWS to ignore proper comments and only tag emendation comments, which are in effect normalisations. The task was performed by an algorithm, based on the observation that proper comments, unlike emendation comments, generally involved several

[9] Soon after completing the first version of the TCEECE, we got funding for converting the entire CEEC-400 into XML (see Saario 2020). This allowed us to further develop our converter program and update the underlying XML format of the TCEECE accordingly. We here describe the updated format.

consecutive capital letters. The latter were placed between XML tags, while the former were hidden inside XML attributes, as follows:

(4) `<note resp="editor" value="ONE WORD MISSING" />`

(5) `<note resp="editor">her</note>`

We acknowledge that our algorithm is not perfect, as it assumes the original encoders to have been more consistent in their application of the codes than they probably were, but it does succeed frequently enough to justify itself. It is more important to extract relevant structure than to avoid casual errors. Hardly any such errors have shown up yet, and they can be manually corrected whenever they do.

3.2.2. Dealing with 'trans-token' codes

In addition to editors' comments, compilers' comments and emendations, there were separate codes for headings, typeface changes and foreign language, encoded as `[}...}]`, `(^...^)` and `(\...\)`, respectively. Whenever a code covered a single token or a sequence of tokens, it could be converted directly into XML, as in the examples above. Problems arose when a code transcended the token division, as in (6).

(6) `thank you for the unus[{ual plea{]sure it has given me.`

The obvious XML translation would have been `unus<supplied>ual plea</supplied>sure`, but CLAWS only tags whole words (cf. the nesting problem in Section 3.3.1). If the information about the exact range of the code was to be kept, it had to be done indirectly. We initially decided to extend the corresponding XML code into the closest sequence of whole words and keep the original encoded sequence inside an 'orig' attribute (cf. the treatment of 'split' words in Rodríguez-Puente *et al.* 2019: 73), as shown in (7).

(7)
```
<supplied orig="unus[{ual plea{]sure">unusual
pleasure</supplied>
```

Later, following a suggestion by our colleagues in Lancaster, we added a 'range' attribute to record the start and end indices of the code. The characters in each sequence were indexed starting from zero (skipping whitespaces). This approach also generalised into cases where there are several code ranges in one sequence, as in (8).

(8)
```
<note resp="editor" range="1,4;5,7"
orig="m[\ist\]r[\es\]s">mistress</note>
```

If matters were not complicated enough, sometimes the consecutive codes were of different kinds —and not only could there be several consecutive codes in one sequence, but there could also be codes inside codes, and more codes inside those codes. In the end, we did find a way to contain all this variation and convert it systematically into XML, but it required a robust algorithm and an elaborate conceptualisation of the hierarchy of codes (Saario 2021).

3.3. Tokenisation and POS tagging

The XML edition of the CEECE was tokenised and POS tagged by CLAWS, using the C7 tagset, and post-processed by a simple script that switched the POS tags inside foreign language passages (encoded as `<foreign>...</foreign>` in XML) to the proper tag for foreign words. The final output was then converted into various formats in both C7 and C5. The accuracy of C7 tagging was checked from a sample and mapped to that of C5.

3.3.1. Final format

The direct output of CLAWS is called ‘vertical’ as there is one line for each token. Long tokens and XML tags with whitespaces have been moved to an associated supplement file and must be manually retrieved from there when the output is converted back to XML. The conversion was performed by a separate program written by Paul Rayson.[10]

We would have liked to enclose sentence tokens in `s` elements and word tokens in `w` elements, as in, for example, the BNC XML edition. Unfortunately, the text-level codes that had been translated into XML before tagging turned out to be incompatible with `s` elements. Whenever the converter reached an opening XML tag inside a sentence, it closed the `s` element before the tag even if the sentence continued after it, for example:

[10] https://github.com/UCREL/convert

(9)
```
<s>
  <w id="410.1" pos="PPH1">It</w>
  <w id="410.2" pos="VM">will</w>
  <w id="410.3" pos="VBI">be</w>
</s>
<supplied range="4,9" orig="some[{thing{]">
  <w id="410.4" pos="PN1">something</w>
</supplied>
<w id="410.5" pos="JJ">chargeable</w>
```

In the above example, there is no reason why the `s` element could not continue over the `supplied` element, but there are other cases where a supplied passage (or another text-level code) continues across a sentence break. In those cases, closing the former sentence and opening the latter one at the breaking point would not have been well-formed XML, as the supplied passage would have been nested in neither sentence. One solution would have been to split the `supplied` code at the sentence break, which would have required some robust post-editing. We took the easiest path and simply omitted `s` elements. The sentence tokenisation is, nevertheless, implicitly encoded in the `id` attributes of `w` elements, where the first number identifies the sentence and the second identifies the word in that sentence. We acknowledge that our solution is not optimal and hope that a better one will be found in the future (cf. the problem of overlapping hierarchies in Marttila 2014: 200–201).

3.3.2. Checking the accuracy

The value of a tagged corpus largely depends on the accuracy of tagging. Typically, the accuracy is estimated by calculating certain key figures from a representative sample. The overall accuracy rate is the number of correctly tagged tokens divided by the total number of tokens. Each tag also has two measures: ‘precision’ is the number of correct assignments divided by the total number of assignments of the tag, and ‘recall’ is the number of correct assignments divided by the total number of tokens for which the tag in question is correct (see Figure 2 for illustration). All it takes to calculate those figures is to go through the tokens one by one and determine for each whether the tag assigned by the tagger is correct, and if it is not, what the correct alternative is.

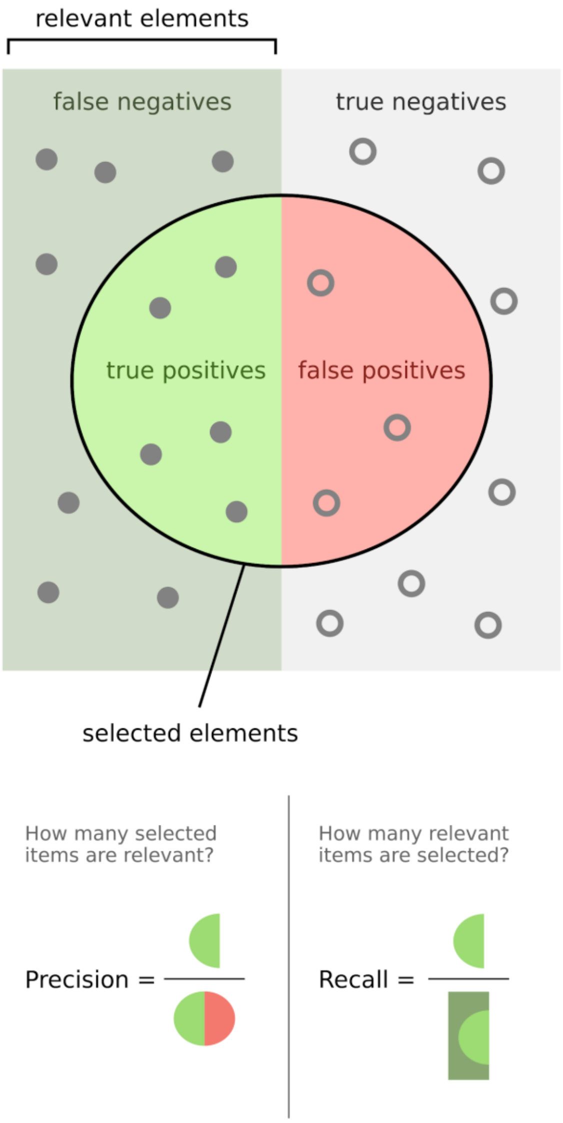

Figure 2: Precision and recall[11]

What was said above presupposes that the underlying tokenisation itself is correct, which in the case of the TCEECE was not true. While our XML converter did merge lines of text into paragraphs, tokenisation into sentences and words was left to the tagger. Even if we would have liked to check the tokenisation before tagging, we were unable to do that, as the tagger only produced one output where the text had been both tokenised and tagged. As a result, there are many incorrect tags due to incorrect tokenisation, as in (10).

(10) `and_CC when_RRQ twill_NN1 be_VBI better_JJR`

[11] Source: https://commons.wikimedia.org/wiki/File:Precisionrecall.svg, published by 'Walber' under the licence CC BY-SA 4.0.

Here, we have an obsolete contraction *'twill* that remains non-normalised due to a lack of resources and because it is not marked by any of the features listed in Section 3.1. The tagger has interpreted it as one token while, in fact, there are two tokens (*it* and *will*) that ought to be tagged separately. Similarly, there are cases where one token has been mistaken for two. In these cases, it is senseless to ask what the correct tags for the incorrectly tagged tokens would be, as there are no correctly tagged tokens in the first place.

We solved the problem as follows. Whenever there is one token that should have been two, it is counted as one incorrectly tagged token. Whenever there are two tokens that should have been one, they are counted as two incorrectly tagged tokens (unless the other is a punctuation mark, in which case only that one counts as incorrect). The correct alternatives for the incorrect tags in either case are classified as 'excluded'.

This workaround allowed us to calculate the overall accuracy as well as precisions and recalls for particular tags in a way that does not distort the figures too much. We could, of course, have chosen the opposite way and counted the true tokens (*it* and *will*) instead of those given by the tagger (*twill*), which might have been closer to the presupposition of perfect tokenisation and the idea of having a baseline or 'gold standard' of tagging, against which the actual tagging is measured (Rayson *et al.* 2007: 8). A third alternative would have been to exclude the incorrect tokens from our sample, but we preferred our figures to reflect the errors in tokenisation as well as in tagging. The results of our calculations are presented in the next section.

4. Discussion

In this section, we will reflect on our choices and their effects, trying to assess the use value of the end product and come up with alternative or additional actions that might have improved it. We will first look at the accuracy of tagging from various points of view and compare it with, for example, other tagged corpora. We will then discuss on a more general level the management of corpus projects and outline suggestions based on the lessons we learned.

4.1. Accuracy of tagging

To check the accuracy of tagging, we compiled a sample of 15 letters, comprising 5,245 running words (c. 0.24% of the total word count) that had been tagged using the C7 tagset. The sample is representative in the sense that the average length of letters and the distributions of letter-writers' genders and ranks as well as times of writing somewhat correspond to those in the entire corpus. Post-processed passages of foreign language were excluded from the sample to avoid bias (see Saario and Säily 2020: §6.1). The tagging of the sample was checked, and the accuracy of tagging calculated following the procedure explained in Section 3.3.2. The results are provided in what follows (see also *ibid.*: §§6.4–6.6).

4.1.1. Overall accuracy

The sample had been tokenised by CLAWS into 5,889 tokens, 5,566 of which we classified as accurately tagged. The overall accuracy of the sample is therefore 94.5 per cent. There is a great deal of variation among the letters, however: the lowest accuracy is 90.6 per cent while the highest is 97.2 per cent. Of the 323 inaccurately tagged tokens in the sample, 32 (9.9%) were due to incorrect tokenisation, which allows us to conclude that the accuracy of tokenisation is 99.5 per cent.

Having combined the results with metadata on the letters and their writers, we learned that the accuracy is 95.4 per cent for letters by men and 92.8 per cent for letters by women, which might be explained by the fact that women typically had less access to education than men. Neither is it unexpected that the accuracy is 93.5 per cent for letters from the seventeenth century and 94.7 per cent for letters from the eighteenth century, given that spelling in English became increasingly standardised over that time. There is no observable difference in the overall accuracy between the upper and lower social ranks, but that is understandable as the sample only has three letters from the lowest rank. In general, the lower social ranks are underrepresented in our corpus for obvious reasons and those who are represented are often the most literate ones, which is bound to cause some bias.

The letter with the lowest tagging accuracy (90.6%) was written by Joanna Clift, a domestic servant with no formal education. We have reason to believe that the Clift letter collection is, in fact, one of the worst collections in terms of tagging accuracy, as it contains relatively many letters from poorly literate writers (see Saario and Säily 2020: §6.6). We had no time to manually correct its tagging because of its size, but we did correct the smaller Pauper collection which we also expected to have been tagged rather inaccurately for the same reason (*ibid*.). We learned that the accuracy of the uncorrected Pauper collection was 87.9 per cent, which may be considered the approximate lower bound for the tagging accuracies of all CEECE collections.

4.1.2. Accuracy by tags

In addition to the overall accuracies, end-users of the tagged corpus will want to know the accuracies of particular tags, especially those they intend to use in their research. Precision and recall were calculated for each C7 tag (see Saario and Säily 2020: Appendix 2) and are summed up into groups in Table 2.

Tag group		Selected assignments (a)	Relevant assignments (b)	True assignments (c)	Precision (c / a)	Recall (c / b)
	Punctuation marks	570	562	562	98.6%	100.0%
A-	Articles	443	439	438	98.9%	99.8%
C-	Conjunctions	384	393	359	93.5%	91.3%
D-	Determiners	179	168	158	88.3%	94.0%
I-	Prepositions	535	527	511	95.5%	97.0%
J-	Adjectives	259	265	237	91.5%	89.4%
M-	Numbers	103	96	95	92.2%	99.0%
N-	Nouns	1,032	1,017	944	91.5%	92.8%
P-	Pronouns	639	644	636	99.5%	98.8%
R-	Adverbs	385	412	358	93.0%	86.9%
TO	Infinitive marker	106	108	106	100.0%	98.1%
V-	Verbs	1,140	1,123	1,063	93.2%	94.7%
	Miscellaneous[12]	114	103	99	86.8%	96.1%

Table 2: The precisions and recalls of C7 tags grouped into categories[13]

[12] Includes, for example, negation, genitive marker, letters of the alphabet and existential *there*.

[13] Note that for each group of tags, the values (a)–(c) are sums of those of the tags in the group. The value (c) might be greater were the particular tags mapped onto the level of the groups, which would consequently improve precision and recall. For example, in the group of nouns, there are 73 false negatives, 46 of which

The most accurate tag groups in terms of both precision and recall are articles, punctuation marks, pronouns and infinitive markers. Numbers also have a high recall even if their precision is relatively low. Determiners and miscellaneous tags are worst in precision, adjectives and adverbs in recall.

What this means in practice is that queries for precise tags print concordances where most lines truly represent the tags in question, but users cannot trust that most true instances are included unless the recall is high, too. On the other hand, concordances for tags with a high recall but low precision include many false instances but, at least, users may suppose most true instances are included and they just have to eliminate the false ones, which is often easier than to dig up missing instances from outside the search results. We might go as far as to say that in corpus linguistics, recall is generally more important than precision (cf. Hoffmann 2005: 21).

To help users to deal with tags that have low recall, we have also calculated the accuracies by pairs of true and false tags (Saario and Säily 2020: Appendix 2). If one were interested in, for example, the tag JJ (general adjective), one would not only know that the recall is 88.9 per cent, and thus 11.1 per cent of all true JJs in the sample have been tagged as something else; one would also know that 3.8 per cent of them have been tagged as VVN, 2.6 per cent as NN1 and so on, which would help to trace the missing JJs.

The specification by tag pairs reveals that some tags have often been confused with other tags under the same group: this is the case, for instance, when a general adverb (RR) has been mistaken for a degree adverb (RG). That is, of course, more forgivable an error than misplacement in a completely wrong category. If the tagset would not distinguish between the two kinds of adverbs, the given case would not cause an error. It is therefore instructive to see what happens to precision and recall when the C7 tagging is mapped into the more coarse-grained C5.

Surprisingly enough, the mapping does not increase the overall accuracy by more than 0.2 percentage points (pp). Of all the 323 errors, only 11 go away. While a transition into C5 significantly impoverishes the annotation, it barely improves the accuracy; specifically, it does not suffice to overcome the poor recall of adjectives and adverbs, even if the latter is slightly increased (by 1.0pp). That does not prevent the tagging from

are confusions between different noun tags. The number of true group assignments (as opposed to particular tag assignments) is therefore 944 + 46 = 990.

being useful, however, as long as the users recognise it does not represent the "God's truth" (Rissanen 1989: 17).

4.1.3. Accuracy by version

Next, we shall compare the accuracy of tagging across the stages of normalisation to find out how much the tagging was improved by each stage. Let us call the original CEECE 'Version 0', the VARD-processed (or 'VARDed') corpus 'Version 1' and the further normalised corpus 'Version 2'. Above, we have already discussed the accuracy of Version 2, having converted it into XML and tagged by CLAWS. The accuracies of the other two versions were determined in the same way: a sample was compiled from the earlier versions of the same letters, converted and tagged, and the tagging was then checked following the same principles as with the final corpus. The results are summarised in Figure 3, where maximum and minimum are the accuracies of the most and least accurately tagged letters, respectively.

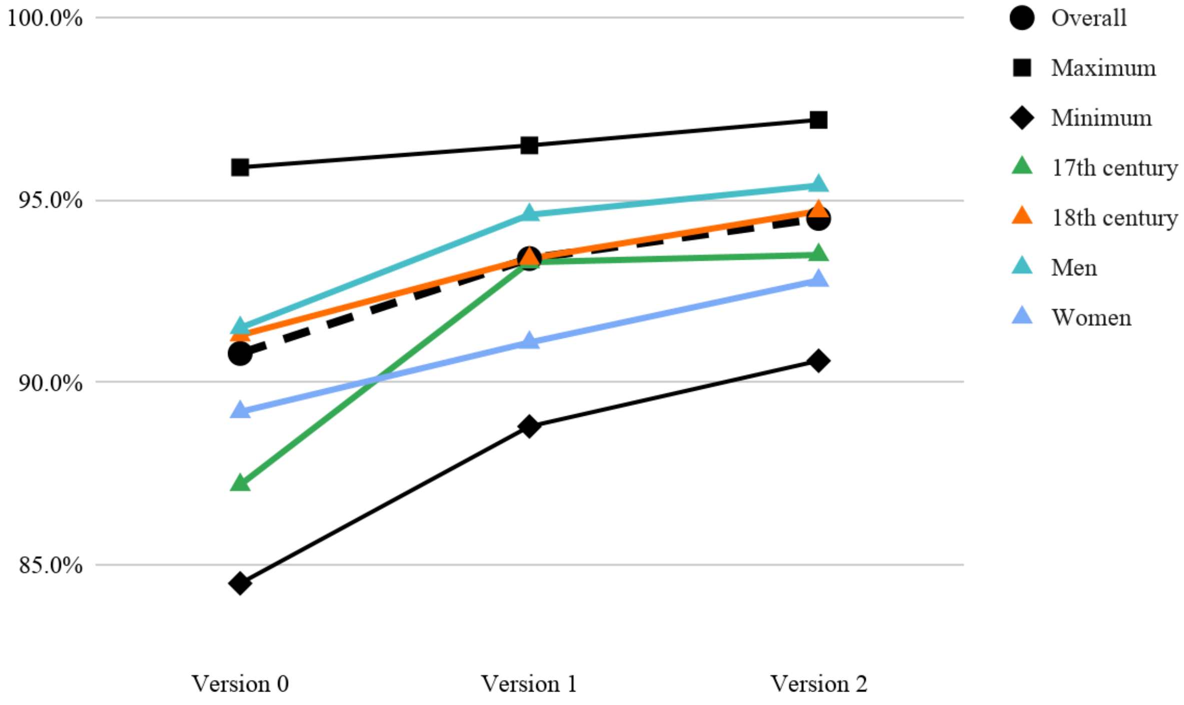

Figure 3: The tagging accuracy of the corpus versions 0, 1 and 2

The overall accuracy of the original corpus (Version 0) was 90.8 per cent. Normalisation in VARD increased the accuracy by 2.6pp to 93.4 per cent and further normalisation by

1.1pp to 94.5 per cent. The difference between maximum and minimum accuracy (the 'range') decreased from 11.4 per cent to 7.7 per cent and finally to 6.6 per cent.

By contrast, Rayson *et al.* (2007) tested the effect of normalisation on CLAWS tagging with two samples from EModE: one from Shakespeare's plays and one from the *Lampeter Corpus of Early Modern English Tracts*.[14] They observed that Shakespeare's initial accuracy of 81.94 per cent was increased to 84.81 per cent (+2.9pp) in VARD and to 88.88 per cent (+4.1pp) in full manual normalisation. For the *Lampeter* sample, which dates from a later period (1640s) and is stylistically closer to the kind of data CLAWS is familiar with, the figures were 88.46 per cent, 89.39 per cent (+0.9pp) and 93.22 per cent[15] (+3.8pp), respectively. The accuracies are lower than those of the TCEECE because of the earlier language form, but the changes in accuracy are more comparable. Differences in the effect of VARDing might have something to do with genre, as the speech-related genres of plays and correspondence probably involve more spelling variation than tracts and pamphlets. Differences in further normalisation are interesting, as the TCEECE was not fully normalised like Shakespeare and *Lampeter*; still, the 1.1pp improvement in the former is relatively good compared to the circa 4pp improvement in the latter.

VARDing the CEECE increased the seventeenth-century accuracy by 6.1pp and the 18th-century accuracy by 2.1pp, bringing the former to almost the same level as the latter. That is not surprising, given that VARD has been designed for EModE in particular. Further normalisation did not improve the seventeenth century by more than 0.2pp, but it did improve the eighteenth century by 1.3pp and so compensated for the bias of the earlier stage. Yet, the sample only has two letters from the seventeenth century, so one must be careful not to generalise too much.

Another (albeit slighter) difference between the two stages of normalisation concerns the gender of writers. The accuracy of men's letters was increased in VARD by 3.1pp and women's by 1.9pp. Further normalisation, in turn, increased men's accuracy by 0.8pp and women's accuracy by 1.7pp. This might imply that men's letters are easier to normalise (semi-)automatically, based on general patterns of variation, whereas women's letters require closer attention to the idiosyncrasies of individual writers.

[14] http://korpus.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM

[15] In the calculation of this figure, a passage of Latin which CLAWS had failed to tag as FWs (foreign words) was excluded from the sample, just like we did with our sample. The figure without exclusion is 91.24 per cent.

4.1.4. Comparison with ARCHER

Schneider *et al.* (2016) employed CLAWS to tag a sample of ARCHER that had been normalised by VARD. They originally used the C5 tagset and mapped it to the *Penn* tagset which only has 39 tags. The accuracy of the final tagging is reported to be 87.8 per cent in the seventeenth century and 93.2 per cent in the eighteenth century, which is 5.8pp and 1.7pp lower than the respective accuracies of the TCEECE in C5. Were the accuracies for the TCEECE calculated in the *Penn* tagset, the difference with respect to ARCHER would be even larger.

The difference between ARCHER and the TCEECE in the seventeenth century is expected, as the TCEECE only covers the end of the century. One must also bear in mind that ARCHER was not further normalised beyond VARDing like the TCEECE. A closer comparison requires that we determine the accuracy of the eighteenth-century part of the TCEECE as it was after VARDing and before further normalisation, using the C5 tagset, and compare it to the eighteenth-century part of ARCHER. We get the result that the TCEECE accuracy is 93.7 per cent, that is, 0.5pp higher than ARCHER. This is surprising, since private spelling as represented by the TCEECE is typically more variable than the spelling of published texts, which is what ARCHER mostly represents. Perhaps, the ARCHER genres have presented the tagger with challenges of a different sort, such as mathematical formulae.

Comparison with other corpora presupposes that the accuracy figures have been calculated in the same way. We have tried to be as transparent as possible about our principles of calculation (see Section 3.3.2), but earlier research has been somewhat vague on the matter. In addition to the treatment of tokenisation errors, one ambiguity that should be resolved is the role of punctuation marks. In checking the tagging of ARCHER, punctuation marks were counted as tagged tokens (Gerold Schneider, personal communication). While we have also followed this convention for comparability, we wonder if it really is wise to equalise punctuation marks with other tokens, given that their tagging tends to be correct by default and is not very interesting anyway.

Even if the tagging of the TCEECE is relatively accurate, it is useful to know what more could have been done to improve it. In the corpus manual, we have listed plenty of known issues, some of which could have been prevented by additional normalisation whereas others are more difficult to avoid before tagging (Saario and Säily 2020: §7).

Comprehensive post-processing by UCREL's *Template Tagger* and manual correction of more collections are, of course, options that we may consider in the future.

4.2. Ideal of gradual enrichment

As we have already noted, the ideal of enrichment did not actualise throughout the process. In the spelling normalisation, information was lost on the original variants as well as text-level coding in normalised variants. Secondly, as already noted, the corpus was not tokenised until it was tagged by CLAWS, so the token identifiers of the TCEECE cannot be used to refer back to earlier versions of the CEECE. Thirdly, because of the problems noted, the TCEECE is largely unsynchronised with the non-tagged, non-tokenised, non-converted or non-normalised versions of the CEECE which will still be used and developed alongside the tagged corpus. For instance, when users find an interesting passage in the TCEECE and want to check its original spelling, they are unable to directly identify the same tokens in the original CEECE. They do have the letter identifier that helps them to find the original letter, but from there onwards, they are on their own, trying to discern the corresponding tokens from the unstructured text that may look a lot different than its normalised, reformatted, tokenised and tagged counterpart.

On the other hand, some people may consider it a relief that not all layers of annotation are piled on top of each other in one file. If they were, users of the corpus might feel overloaded with information, struggling to discern the relevant parts from the thick jungle of code. It is for this reason that, for example, the compilers of CHELAR decided to keep the POS tagged and TEI-XML versions of their corpus separate (Rodríguez-Puente *et al.* 2019: 79–80). Indeed, it seems as if there were an upper bound on how far the enrichment of a corpus should go. Corpus developers should be careful about enriching one corpus version too much. If you add too much annotation, the corpus will become unusable and lose its value.

As much as we sympathise with the underlying concern, we believe there is something to be done other than just settling for many imperfect versions. The layers of annotation can be separated to distinct files by means of stand-off markup that preserves their linking to the 'primary text' (see Marttila 2014: 195ff). On the other hand, even if it is not optimal for an end-user to have all the data in one place, that does not mean there could not be such a place where the plainer versions come from. A distinction should be made between an all-inclusive ‘master’ version, maintained by developers of the corpus, and simplified subalternate versions that are actually used by researchers. If the master corpus is encoded in XML, tailor-made versions can be derived from it with XSL transformations to suit each user’s individual needs (see e.g. the BNC stylesheets).[16]

There is an additional reason for separating maintenance and use. Our experiences with the TCEECE and other CEEC corpora have taught us that the maintenance of many parallel versions of one corpus becomes excessively laborious as time goes by. The more versions there are, the likelier it is that changes are made to some versions while others fall out of sync (for a cautionary example, see Saario 2020: §1). It would be preferable to have all the data kept up to date in one branch and have a version control system (e.g. *Git*)[17] keep track of the changes.

Of course, that still leaves open the question of how exactly the master corpus is to be constructed, organised and encoded. Incorporating multiple overlapping layers of annotation into one format is a challenge that will not be solved here (see Marttila 2014: §5.6). In an ideal world, we would have been able to envision an all-inclusive format right from the start, allowing us to do things in a more logical order. First, we would have converted the original CEECE into XML. Second, we would have tokenised the unstructured text, which would have assigned identifiers to the data points that could then have been referred back to from later stages. Third, we would have normalised the spelling, keeping the original tokens in store alongside the normalised ones. Finally, we would have POS tagged the normalised tokens. Each stage would have built on the earlier ones, and no information would have been lost in the process. See Figure 4 for illustration (and cf. Figure 1).

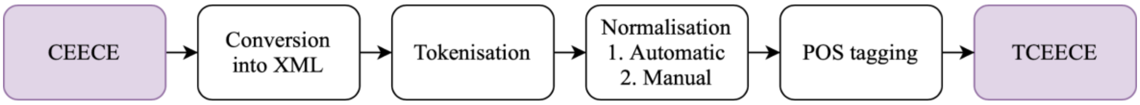

Figure 4: A visualisation of the ideal workflow

That said, we acknowledge our vision will be inaccessible to many corpus compilers for similar reasons as it was to us. Linguists often do not have the expertise to do things like implementing layered annotation, writing XSL transformations or using a version control system, and nor do many of the temporary research assistants who do a major part of the

[16] http://www.natcorp.ox.ac.uk/using/index.xml?ID=stylesheets
[17] https://git-scm.com

actual work. Even if they did, the third-party software they use (e.g. VARD or CLAWS) might not support the ideal workflow without adjustments. In the real world, people will have to come up with easy workarounds past difficult problems, just like we did. Still, it is useful to evaluate different workarounds against the ideal way of doing things, as it may help to avoid the worst pitfalls.

5. CONCLUSION

The long legacy of the CEECE is still present in the TCEECE. The markup has changed, but the content of the letters is still based on the source editions from which they have been compiled. Traces of the legacy format remain in, for example, attributes and headers. Yet the new format is well-formed and valid XML that largely complies with the TEI guidelines and is compatible with modern tools, which we hope is enough to prolong the life of the corpus by decades.

The value of the TCEECE may be measured along various axes. Extensive automatisation throughout the production process has resulted in errors that should be manually corrected. The accuracy of tagging seems sufficient, even if it could be improved by more ambitious post-processing. The end product is not as rich as it could be, which some users may find a good thing, while the maintainers will have to face the fact that we now have one more parallel version of the same corpus. Yet the corpus is primarily intended to provide a resource that can be easily used by linguists, including those with little technical know-how. All in all, what we have accomplished so far may very well be a good enough compromise between the desiderata of effectiveness, correctness, richness, usability and maintenance.

At the time of writing, the TCEECE is being imported to our *CQPweb* server and will hopefully soon be available to researchers and visitors in our unit. Preliminary research on neologisms has already been tried on the corpus; other prospective topics include large-scale investigations of variation and change in POS frequencies (cf. Säily *et al.* 2011, 2017) as well as keyness and collocation analyses that take word class into account. The combination of POS tagging and social metadata in a relatively large and representative historical corpus of private writing will make the corpus attractive to many linguists interested in the interplay between language-internal and -external factors affecting language variation and change (excluding orthography). In the future, we may consider further enriching the corpus by adding, for example, lemmatisation to word tokens or distinct markup to the formulaic elements of letters.

References

ARCHER = *A Representative Corpus of Historical English Registers*. 1990–1993/2002/2007/2010/2013. Originally compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. https://www.projects.alc.manchester.ac.uk/archer/ (25 February, 2020.)

Baron, Alistair. 2011a. VARD 2. Computer program. http://ucrel.lancs.ac.uk/vard/ (25 February, 2020.)

Baron, Alistair. 2011b. *Dealing with Spelling Variation in Early Modern English Texts*. Lancaster: Lancaster University dissertation. https://eprints.lancs.ac.uk/id/eprint/84887/ (25 February, 2020.)

BNC = *The British National Corpus*, version 3 (BNC XML edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk (25 February, 2020.)

CEEC-400 = *Corpora of Early English Correspondence*. 2020. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Jukka Keränen, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily and Anni Sairio at the Department of Modern Languages, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/CEEC/ (19 June, 2021.)

CEECE = *Corpus of Early English Correspondence Extension*. 2012. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily and Anni Sairio at the Department of Modern Languages, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/CEEC/ (19 June, 2021.)

CLAWS. Computer program. Developed by UCREL at Lancaster University. http://ucrel.lancs.ac.uk/claws/ (25 February, 2020.)

Corcoran, Paul E. 1974. COCOA: A FORTRAN program for concordance and word-count processing of natural language texts. *Behavior Research Methods & Instrumentation* 6/6: 566.

Davies, Mark. 2019. Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design. In Carla Suhr, Terttu Nevalainen and Irma Taavitsainen eds. *From Data to Evidence in English Language Research* (Language and Computers 83). Leiden: Brill, 66–87.

Fligelstone, Steve, Mike Pacey and Paul Rayson. 1997. How to generalize the task of annotation. In Roger Garside, Geoffrey Leech and Anthony McEnery eds. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, 122–136. http://ucrel.lancs.ac.uk/papers/CAB_CH08.pdf (25 February, 2020.)

Hardie, Andrew. 2012. *CQPweb* – Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17/3: 380–409.

Hardie, Andrew. 2014. Modest XML for corpora: Not a standard, but a suggestion. *ICAME Journal* 38: 73–103.

HC = *The Helsinki Corpus of English Texts*. 1991. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old

English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). Department of Modern Languages, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/ (19 June, 2021.)

Hiltunen, Turo, Joe McVeigh and Tanja Säily. 2017. How to turn linguistic data into evidence? In Turo Hiltunen, Joe McVeigh and Tanja Säily eds. *Big and Rich Data in English Corpus Linguistics: Methods and Explorations* (Studies in Variation, Contacts and Change in English 19). Helsinki: VARIENG. https://varieng.helsinki.fi/series/volumes/19/introduction.html (19 June, 2021.)

Hiltunen, Turo and Jukka Tyrkkö. 2013. Tagging Early Modern English Medical Texts (1500–1700). Presentation at *The First Corpus Analysis with Noise in the Signal Workshop* (CANS 2013), 22 July, Lancaster University, UK. http://ucrel.lancs.ac.uk/cans2013/abstracts/Hiltunen%20Tyrkk%C3%B6.pdf (25 February, 2020.)

Hoffmann, Sebastian. 2005. *Grammaticalization and English Complex Prepositions: A Corpus-based Study*. London: Routledge.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hundt, Marianne ed. 2014. *Late Modern English Syntax*. Cambridge: Cambridge University Press.

Hundt, Marianne and Geoffrey Leech. 2012. “Small is beautiful”: On the value of standard reference corpora for observing recent grammatical change. In Terttu Nevalainen and Elizabeth C. Traugott eds. *The Oxford Handbook of the History of English*. Oxford: Oxford University Press, 175–188.

Kaislaniemi, Samuli. 2018. The *Corpus of Early English Correspondence Extension* (CEECE). In Terttu Nevalainen *et al*. eds., 45–59.

Kaislaniemi, Samuli, Mel Evans, Teo Juvonen and Anni Sairio. 2017. ‘A graphic system which leads its own linguistic life’? Epistolary spelling in English, 1400–1800. In Tanja Säily *et al*. eds., 187–214.

Kroch, Anthony, Ann Taylor and Beatrice Santorini. 2000. *The Penn-Helsinki Parsed Corpus of Middle English.* Department of Linguistics: University of Pennsylvania.

Kroch, Anthony, Beatrice Santorini and Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English.* Department of Linguistics: University of Pennsylvania.

Kytö, Merja. 1996. *Manual to the Diachronic Part of The Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts* (third edition). Helsinki: Department of English, University of Helsinki. http://clu.uni.no/icame/manuals/HC/INDEX.HTM (25 February, 2020.)

Lu, Xiaofei. 2014. *Computational Methods for Corpus Annotation and Analysis*. New York: Springer.

Marttila, Ville. 2011. *Helsinki Corpus TEI XML Edition Documentation*. Helsinki: VARIENG. https://helsinkicorpus.arts.gla.ac.uk/display.py?fs=100&what=manual (25 February, 2020.)

Marttila, Ville. 2014. *Creating Digital Editions for Corpus Linguistics: The Case of Potage Dyvers, a Family of Six Middle English Recipe Collections*. Helsinki: University of Helsinki dissertation. http://urn.fi/URN:ISBN:978-951-51-0060-3 (25 February, 2020.)

Nevalainen, Terttu, Minna Palander-Collin and Tanja Säily eds. 2018. *Patterns of Change in 18th-century English: A Sociolinguistic Approach*. Amsterdam: John Benjamins.

Nurmi, Arja ed. 1998. *Manual for the Corpus of Early English Correspondence Sampler, CEECS*. Helsinki: Department of English, University of Helsinki. http://korpus.uib.no/icame/manuals/CEECS/ (25 February, 2020.)

PCEEC = *Parsed Corpus of Early English Correspondence*. 2006. Annotated by Arja Nurmi, Ann Taylor, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. http://hdl.handle.net/20.500.12024/2510 (25 February, 2020.)

Rayson, Paul, Dawn Archer, Alistair Baron, Jonathan Culpeper and Nicholas Smith. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In Matthew Davies, Paul Rayson, Susan Hunston and Pernilla Danielsson eds. *Proceedings of Corpus Linguistics 2007, 27–30 July, University of Birmingham, UK*, article 192. http://ucrel.lancs.ac.uk/publications/CL2007/ (25 February, 2020.)

Rissanen, Matti. 1989. Three problems connected with the use of diachronic corpora. *ICAME Journal* 13: 16–19.

Rodríguez-Puente, Paula, Cristina Blanco-García and Iván Tamaredo. 2019. Mark-up and annotation in the *Corpus of Historical English Law Reports* (CHELAR): Potential for historical genre analysis. *Journal of the Spanish Association of Anglo-American Studies* 41/2: 63–84.

Russell, D. B. 1965. COCOA —A Word-Count and Concordance Generator. http://www.chilton-computing.org.uk/acl/applications/cocoa/p001.htm (25 February, 2020.)

Saario, Lassi. 2020. *Conversion of the CEEC-400 into XML. A Manual to Accompany the XML Edition*. Helsinki: VARIENG. https://varieng.helsinki.fi/CoRD/corpora/CEEC/xml_doc.html (19 June, 2021.)

Saario, Lassi. 2021. *XmlConverter*. *A Java Application to Process the File Format of the Corpora of Early English Correspondence*. Helsinki: VARIENG. https://version.helsinki.fi/ceec/ceec-tools/XmlConverter (19 June, 2021.)

Saario, Lassi and Tanja Säily. 2020. *POS Tagging the CEECE. A Manual to Accompany the Tagged Corpus of Early English Correspondence (TCEECE)*. Helsinki: VARIENG. https://varieng.helsinki.fi/CoRD/corpora/CEEC/tceece_doc.html (19 June, 2021.)

Säily, Tanja, Terttu Nevalainen and Harri Siirtola. 2011. Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing* 26/2: 167–188.

Säily, Tanja, Turo Vartiainen and Harri Siirtola. 2017. Exploring part-of-speech frequencies in a sociohistorical corpus of English. In Tanja Säily *et al*. eds., 23–52.

Säily, Tanja, Arja Nurmi, Minna Palander-Collin and Anita Auer eds. 2017. *Exploring Future Paths for Historical Sociolinguistics*. Amsterdam: John Benjamins.

Sairio, Anni, Samuli Kaislaniemi, Anna Merikallio and Terttu Nevalainen. 2018. Charting orthographical reliability in a corpus of English historical letters. *ICAME Journal* 42/1: 79–96.

SCEEC = *Standardised-spelling Corpora of Early English Correspondence*. 2012. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Jukka Keränen, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily and Anni Sairio. Standardised by Mikko Hakala, Minna Palander-Collin and Minna Nevala. Department of English / Department of Modern Languages, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/CEEC/ (19 June, 2021.)

Schneider, Gerold, Marianne Hundt and Rahel Oppliger. 2016. Part-of-speech in historical corpora: Tagger evaluation and ensemble systems on ARCHER. In Stefanie Dipper, Friedrich Neubarth and Heike Zinsmeister eds. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)* (Bochumer Linguistische Arbeitsberichte 16). Bochum: Ruhr-Universität Bochum, 256–264. https://www.linguistics.rub.de/konvens16/proceedings.html (25 February, 2020.)

TCEECE = *Tagged Corpus of Early English Correspondence Extension*. 2020. Annotated by Lassi Saario and Tanja Säily. Spelling standardised by Mikko Hakala, Minna Palander-Collin, Minna Nevala, Emanuela Costea, Anne Kingma and Anna-Lina Wallraff. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily and Anni Sairio at the Department of Modern Languages, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/CEEC/ (19 June, 2021.)

TEI Consortium, eds. 2020. *Guidelines for Electronic Text Encoding and Interchange*. Last updated on 13 February, 2020. http://www.tei-c.org/P5/ (25 February, 2020.)

Corresponding author
Lassi Saario
P.O. Box 24
FI-00014
University of Helsinki
Finland
e-mail: lassi.saario@helsinki.fi

received: February 2020
accepted: June 2021