statuS Quaestionis
language text culture

Emma Pasquali
eCampus University of Novedrate

The Corpus of *Early Modern English Trials* (1650-1700): Building of the Corpus and Hypotheses of Normalization

Abstract

The present paper discusses the building stages of the Corpus of *Early Modern English Trials* (1650-1700), henceforth *EMET*, a 1.8 million words highly specialized historical corpus of trial proceedings. The main purpose of the creation of the above-mentioned corpus is to shed light on the pragmatic aspects of Early Modern spoken English, since trial proceedings are considered records of authentic dialogues (Culpeper and Kytö 2010, 17). More specifically, the *EMET* was created in order to investigate the pragmatic influences both on the choice of the second person pronoun, which coexisted in the forms *thou* and *you*, and of any T- and Y-form used during the Restoration: *thee*, *prithee*, *prethee*, *prethy*, *pray thee*, *thy*, *thy self*, *thyself*, *thine*, *you*, *ye*, *your*, *your self*, *yourself*, *yours* and *pray you*.

The initial part of the essay will briefly explore the phase of the archives' consultation, the criteria behind the selection of the trials and the technical stages that are necessary to the uploading of a corpus on #LancsBox and its study. Afterwards, the *EMET* itself will be presented (number of documents, total number of tokens and average number of tokens per text, and types of charges involved).

Then, the essay will focus on editing, normalization and POS tagging. More specifically, it will be illustrated how trials, and historical documents in general, should be edited in order to successfully analyse them with corpus linguistics tools. Then, different hypotheses of normalization of the *EMET* will be compared in detail and discussed. After determining which normalization parameters suit best the corpus, the advantages of such process will be highlighted. Lastly, the issues derived from the normalization process – mainly bound to proper nouns, badly preserved documents (i.e., noisy texts), and Latin (and foreign) terms – will be examined.

1. *Towards a Corpus of* Early Modern English Trials *(1650-1700)*

The Corpus of *Early Modern English Trials* (1650-1700) is a highly specialized historical corpus of trial proceedings, which serves the primary purpose of shedding light on the pragmatic aspects of Early Modern spoken English, as trial proceedings are considered to be authentic records of dialogues (Culpeper and Kytö 2014, 17). The present essay will begin with the discussion of the corpus compilation phase, outlining the steps taken in order to upload the corpus to #Lancsbox and prepare it for analysis. Then, the *EMET* corpus itself will be presented.

The corpus building stage took almost two years, as the quality of the research is heavily dependent on the choices made in this phase. As Gablasova, Brezina and McEnery (2019, 127) underline,

> [t]he properties of a corpus, such as representativeness, structure and amount of evidence, directly affect the ability of researchers to interpret findings and generalise to contexts outside of the corpus (Leech 2007; Gablasova et al. 2019). Decisions made at the corpus-building stage can thus have far-reaching consequences for the quality of research studies based on them; this is especially true of large-scale corpus-building projects, with their products expected to be used in a large number of research studies [...].

Therefore, informed decisions were crucial in the first stages of the research, that is: i) archive consultation and trial selection; ii) editing [phase A]; iii) normalization and editing [phase B]; iv) linguistic annotation.

2. *The Archives Consultation*

The first phase of corpus compilation involved querying several databases, including *Archive.org*, the *Oxford Text Archive*, *Old Bailey Online*, and *Early English Books Online* (*EEBO*). The search was conducted using the Early Modern variants of the word 'trial' (i.e., 'trial', 'triall', 'trial' and 'tryall'), which were determined with the help of *Lexicons of Early Modern English* (LEME[1])

1 *LEME* (*Lexicons of Early Modern English: introduction*) is a historical database comprising various types of useful printed or manuscript sources from about 1475 to 1755 (monolingual, bilingual, and polyglot dictionaries, lexical encyclopaedias, hard-word glossaries, spelling lists, and lexically-valuable treatises).

and of *The Oxford Dictionary of English Etymology* (Onions et al. 1966). Because of the nature of the research question – which involves an analysis of face-to-face interaction – accounts of trials in running prose, which was "the form traditionally used for official records" (Culpeper and Kytö 2014, 50), were excluded, and only trials in the dialogue format were selected because of the high frequency of the second person pronoun[2] (Walker 2007, 12). Indeed, the documents included in the *EMET* are believed to be 'speech-based' since they represent real life face-to-face interaction (Culpeper and Kytö 2014, 16-7); furthermore, they may be considered authentic dialogues, since they are "written records of real speech events taken down at the time of the speech event" (Ibid.: 23). However, they are not 'verbatim' transcriptions since no electronic devices existed and stenography was only at its dawn (Aliprandi and Pigò 1936; Culpeper and Kytö 2014, 17). More specifically, no full systems of shorthand existed and "most speech-based texts [were] reconstructions assisted by notes" (Culpeper and Kytö 2014; Shoemaker 2008, 560). In other words, the process followed by the scribes when reconstructing the oral discourse of the people intervened with the cause cannot be definitively known (Doty 2007, 26). For these reasons, an exact copy of what was said in the courtroom is impossible to obtain.[3] Thus, it is more appropriate to adopt the notion of *faithfulness* and to consider the factors of influence pinpointed by Short, Semino and Wynne (2002), as quoted by Culpeper and Kytö (2014, 79-81):

- *Anterior discourse accessibility*: if no recordings are available, spoken language is accessible only at the moment of the utterance and, thus, the collected data is to be considered only partially accessible.
- *Posterior discourse accessibility*: report and reported speech are to be compared in order to verify the level of faithfulness.
- *The importance of (the wording of) what is being reported*: the exact word uttered are of fundamental relevance in witchcraft, libel and slander cases.
- *The memorability of the original*: replicability is a fundamental notion within trial proceedings and it is strictly bound to the notion of memorability: if a deposition cannot be repeated because, for instance, the witness is dying,

2 The pragmatic influences on the second person pronoun are the object of the research that required the building of the *EMET*.

3 An additional reason lies in the diamesic variation from speech to writing (Culpeper and Kytö 2014, 79).

their words become more memorable and more efforts will be made in order to remember them precisely.
- *The status, social role and personality of the producer of the original discourse*: it is believed that, when reporting the utterances of powerful personalities, the scribes may have been more punctilious.
- *The social role, personality and attitude of the reporter*: even though information about the scribe is often not available, his attitude towards what happens in the courtroom inevitably influences what is reported.
- *Text-type or speech context*: despite being courtroom speech considered extremely faithful (Culpeper and Kytö 2014, 80), it must be argued that historical trial proceedings cannot be considered as faithful as the contemporary ones.
- *The part of text in which reporting occurs*: utterances between inverted commas or texts in the dialogue format are believed to be more faithful to the exact words that were uttered.

As Shoemaker (2008, 560-2) points out, when discussing the *Proceedings of the Old Bailey* and considering the pivotal researches of Langbein (1978), the published trials constituted abbreviations of what happened in the courtroom: in fact, the scribes (i.e., shorthand writers) as well as the publishers, had the power to decide which parts to include and exclude, and thus to shape the content of the publications, even according to their need to sell copies to a wide audience who desired entertainment (Shoemaker 2008, 564). Therefore, with respect to the scribes, it can be affirmed that despite having a limited explicit role in the dialogues of the trial proceedings[4] (Culpeper and Kytö 2014, 23), their implicit role was extremely influ-

4 The role of the scribe was limited to the identification of the speakers, eventual statements that a certain witness appears or is sworn in court, brief descriptions or comments about non-verbal communication or comments about the tone used during some utterances (Walker 2007, 13). In contrast with trial proceedings, witness depositions, which were often in third person, display a more prominent role of the reporter because of the presence of legal formulae and information about the deponent (e.g., age, domicile, occupation/marital status) (Culpeper and Kytö 2014, 24; Walker 2007). The second person pronouns are rarer in this type of document since they are to be found only when the witness "reports an earlier speech event, and the scribe renders the words quoted as direct speech" (Walker 2007, 13). For the above-mentioned reason, depositions were not included in the *EMET*.

ential.[5] Other influential factors were certainly noise and problems concerning stationery; in fact, courtrooms "were crowded and noisy, making it difficult to hear what was being said" (Culpeper and Kytö 2014, 52) and the writing equipment constantly needed maintenance: ink had to be re-applied, pens had to be resharpened etc. Despite the difficulties, it is believed that scribes aimed at reporting as faithfully as possible the words uttered during the trials (Walker 2007, 15) and occasionally provided explanatory comments about the trials and non-verbal communication (Ibid.: 12). Anyway, the selection of the documents resulted in the *EMET*, a highly specialized historical corpus of trial proceedings containing 59 trials and over 1.8 million words.[6]

As in most small-scale projects, the focus of the research is on a specialized type of discourse used by a relatively restricted group of speakers (Paquot and Gries 2020, 3). The chosen trials are believed to be 'samples', which statistically can be defined as 'a group of cases' representative of a population; because of representativeness, the results concerning the above-mentioned samples can be generalized to the population living in the period of the Republic and Restauration (McEnery and Hardie 2012, 250). Metadata about the speakers have been collected, but it is worth noting that most of the speakers in the *EMET* are from higher ranks of society, likely because trials about personalities were easier to sell and often showed the power of the monarch. High treason is the most common accusation in the *EMET*; and, while acquittals in such cases were rare, they were common in ordinary criminal prosecutions (Langbein 1978, 267).

3. *Editing (Phase A)*

The texts underwent editing prior to the word count; more specifically, information about retrieval, which was often automatically included in the files,[7]

5 It should also be noticed that information about the scribe is rarely available; thus, it is not known whether the scribe was a professional (Ibid.: 15). (Culpeper and Kytö 2014; Walker 2007)

6 The exact number of tokens after editing and normalisation is 1,847,699.

7 The files frequently included the URL, an abstract, page numbering (often including the word 'page'), information about publication, author and manuscript (or book); furthermore, the title was often listed twice.

dedications,[8] advertising,[9] warnings, and disclaimers[10] were deleted, as well as letters and depositions[11] that were not read during the trials. Furthermore, extra documents, which publishing houses often added at the end of the texts for entertainment purposes, were also omitted. Any deletions were indicated within the text using square brackets ([...]), and were excluded from the final word count.

Afterwards, each document was converted into plain text (.txt) and normalized due to the significant spelling variation found in Early Modern Eng-

8 The dedication that was present at the beginning of the *The Proceedings and Tryal in the Case of the Most Reverend Father in God, William, Lord Archbishop of Canterbury and the Right Reverend Fathers in God, William, Lord Bishop of St. Asaph, Francis, Lord Bishop of Ely, John, Lord Bishop of Chichester, Thomas, Lord Bishop of Bath and Wells, Thomas, Lord Bishop of Peterborough, and Jonathan, Lord Bishop of Bristol* is partially reproduced here: "To his Most Illustrious Highness William Henry, Prince of Orange. May it please Your Highness, how deeply the Design was laid, and with what Violence carry'd on by those who lately Steer'd the Helm of this State, for the Subversion of the Establish'd Religion and Government of these Three Kingdoms, is already sufficiently well known to Your Highness. [...]".

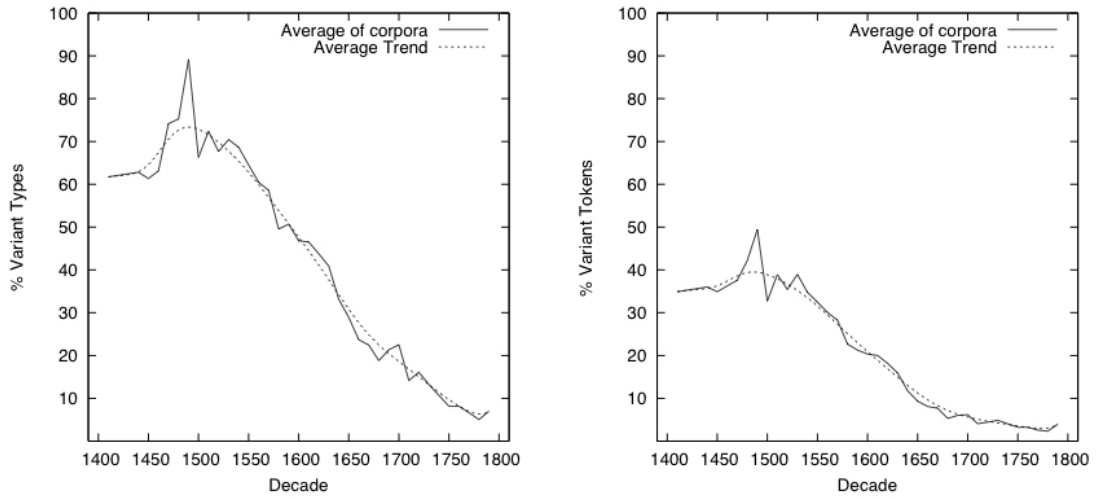
9 Early Modern Courts were venues for entertainment, especially if notorious individuals were involved in the trials (Culpeper and Kytö 2014, 119); for the same reason, accounts of trials and 'verbatim' records were a form of written entertainment. Thus, it was customary to include information concerning the next publications at the end of the pamphlet or book. For instance, the following advertisement, which was placed at the end of *The Tryal and Condemnation of Dr. Oliver Plunket*, was deleted: "Advertisement. Some Passages of the Life and Death of John Earl of Rochester, who died the 26. of July, 1680. By Gilbert Burnet, D. D. Are to be sold by Eliphal Dobson Bookseller on Cork-Hill, 1681".

10 For instance, here is partially reported a disclaimer that was part of *The Tryal of John Giles*: "To the Reader. Certain it is, that by the Fall of Adam the General Peace establish'd through the whole Creation betwixt Man and Man, and even among the Beasts themselves, was universally-broken. Nature could never restore that Peace to the Brute Animals, but that they still devour and prey one upon another. But Heaven provided for Rational Man a Sacred Means to regain and preserve that Blessed Unity, which would have always accompany'd his State of Innocency, which was the Observance of Religion; which as it binds us to God, so ought it to tie us one to another in the strict bonds of Heavenly Example. To this intent, at length Christ himself brought down from Heaven a Gospel of Love and Charity; so that, as it is the True Character of a True Religion to Unite and Preserve, so it is the most certain Sign of a False and Counterfeit Religion, to disunite and destroy Mankind. [...]".

11 Cusack (as quoted in Culpeper and Kytö 2014, 54) affirms that "[t]he regular procedure was for depositions to be read aloud in court, the witness being present to confirm his or her evidence and to answer any questions that might arise".

lish texts, despite the gradual standardization that occurred between 1500 and 1700 (Görlach 1991; Nevalainen 2006). As the following graphs, based on the average variant percentage from six corpora (*ARCHER*, *EEBO*, *Innsbruck*, *Lampeter*, *EMEMT*, and *Shakespeare*), illustrate (Baron 2011, 55), spelling variation decreased significantly between 1400 and 1800 but was still present. Thus, researchers must address the problem, otherwise the search in any corpus of Early Modern English texts would be particularly problematic because:

> using a simple search algorithm would only return the occurrences of the word when it is spelt exactly the same as the search query – spelling variants of a word would not be returned. One option is to search for both the word and its variants, however, it is often difficult to know all of the possible spelling variants for a word and the lists can be very long, substantially increasing processing time. (Baron 2011, 18)



(a) Types %

(b) Tokens %

Fig. 1. Comparison of variant counts in EEBO corpus samples with (=original) and without initial capital words (Baron 2011, 55).

The trials examined in this study, as well as most Early Modern texts, can be classified as 'noisy' texts due to the significant variability in spelling they exhibit. As Baron (Ibid.: 2) points out, the spelling of a word could change depending on the author, scribe, or publisher, among other factors.

Given the aforementioned challenge and the need to analyse Early Modern corpora, normalization becomes a crucial step in any corpus linguistics research involving historical texts. In other words, spelling variation poses a hindrance to corpus linguistics analysis and must be addressed meticulously, either by using normalized texts when available or by normalizing the documents in question (Ibid.: 17).

4. *Normalisation and Editing (Phase B)*

The software utilized for normalization in this study is the Variant Detector (VARD, version 2.5.4), which is a customizable and trainable tool that allows for tuning the normalization process to "a specific corpus and the unique properties of its spelling variation" (Baron 2011, 140). VARD was developed through the manual compilation of a large 'Early Modern regularization list', created by manually inspecting words tagged as Z99 by the UCREL Semantic Analysis System (USAS). Since USAS relies on a modern dictionary, any untagged word could potentially be a spelling variant (Ibid.: 28).

VARD is designed to standardize spelling variation in historical corpora, facilitating analysis with computational linguistics tools. It can be used for manual or automated processing ('batch normalization') of texts. The tool helps identify spelling variants and, when used manually, suggests "appropriate modern equivalents"; when used automatically, it selects "the appropriate equivalents" (Archer et al. 2015, 11). Scholars using VARD have the option to either retain the original spelling within the corpus, signalling it with an XML tag surrounding the replacement (e.g., ["h]ence: <normalized orig="charitie">charity</normalized>) (Ibid.), or use the plain version of the corpus without any indication of the normalization that took place.

Manual normalization of a corpus containing 1.8 million words is time-consuming, so batch normalization was performed. As the *EMET* was specifically designed to investigate pragmatic differences in the use of Y- and T-forms, every term under investigation was searched in various hypotheses of normalization, as well as in the non-normalized corpus. After quickly comparing multiple versions of the batch normalized corpus, two different combinations of parameters were selected and thoroughly compared to determine the most appropriate option:

a. F-score weight: 1.0; Threshold: 50%;
b. F-score weight: 1.0; Threshold: 75%.

The parameter called 'f-score weight' is closely related to confidence scores for methods and replacements, and it is calculated by considering both precision and recall scores. In the *EMET*, precision and recall are considered to be equally relevant, so the f-score weight was set to 1, avoiding any bias towards precision (f-score weight < 1) or recall (f-score weight > 1) (VARD User Guide 2013).

Setting the threshold for normalization required a slower process, as it is closely tied to the concept of confidence score. For each potential normalization of a given variant, a confidence score is assigned, and when using batch processing mode, the tool automatically selects the normalization with the highest confidence score to replace the variant (Ibid.). Therefore, the threshold is crucial in determining the minimum confidence score required for a normalization to be accepted, and if the threshold is not met by the top normalization suggestion, the word is retained as a variant (Ibid.).

Upon observing the following table, it becomes immediately apparent that the corpus (and consequently the research question itself) greatly benefits from the normalization process.

Threshold	Thou	Thee	Prithee	Prethee	Thy (includes thy self)	Thy self | Thyself	Thine	You (singular and plural)	Ye	Your (includes your self)	Yourself | Your self	Yours	Pray you
Non normalized	523	146	13	17	350	38 | 0	0	29 090	57	8 606	11 | 480	68	30
50%	523	155	25	6	355	38 | 0	0	29 096	62	8 607	11 | 482	69	30
75%	523	155	25	6	350	38 | 0	0	29 095	62	8 607	11 | 482	68	30

Table 1. Comparison among different versions of the *EMET*.

Upon setting the parameters as follows: a) f-score weight: 1.0 and threshold: 50%, the results were unusual. The forms of the second person singular pronoun 'thy' and 'thy self' were found to be 355, whereas in the non-normalized corpus they were 350. This difference can be attributed to the normalization process: when the parameters threshold 50% and f-score weight 1 were used, some forms that lacked the final letter(s) in the files due to incomplete readability of the manuscripts were erroneously emended by the normalization software.

Setting the b) parameters (f-score weight: 1.0 and threshold: 75%) resolved the issue;[12] however, it introduced another problem. Specifically, in *The Trial of Thomas White alias Whitebread* (1679), a form was incorrectly left as "prithe", and thus it was manually amended to "prithee":

12 VARD 2 was utilized on several corpora, including the *Corpus of English Dialogues* (*CED*), the corpus of *Early Modern English Medical Texts* (*EMEMT*), the *Lampeter Corpus of Early Modern English Tracts* and the *Corpus of Early English Correspondence* (*CEEC*), significantly diminishing spelling variation (see Lehto, Baron, Ratia and Rayson 2010; Baron, Rayson and Archer 2011; Archer, Kytö, Baron and Rayson 2015; Palander-Collin and Hakala 2011). Notably, the use of VARD on the *CED* and the *Lampeter Corpus* is particularly interesting.
The *CED (1560-1760)* comprises a variety of documents, such as trials, witness depositions, prose, handbooks, comedy drama and miscellaneous materials (Culpeper and Kytö 2014). Baron, Rayson and Archer (2011) specifically selected trials and witness depositions in order to test VARD 2.4 and DICER, a software that "[d]etermines what letter replacement rules are required to convert the variant form into the normalised form" (Baron, Rayson and Archer 2011) and that was under maintenance at the time of writing this essay. After dividing the documents into two periods (1560-1639 and 1640-1749), the scholars trained VARD for each sub-corpus with a sample of 10 000 randomly selected words. They then decided to adopt a 75% replacement threshold.
Similarly, the *Lampeter Corpus of Early Modern English Tracts*, which contains tracts and pamphlets about religion, politics, science, law, economy, trade and miscellaneous, dated between 1640 and 1740 (Schmied 1994) underwent an analogous process. Baron, Rayson and Archer (2014), in this case as well, selected law texts and, after training VARD on "10 randomly selected 1,000 words samples", they opted for a 75% threshold. Thus, it seems that the correct threshold to normalize Early Modern English texts is 75%, since such normalization threshold was used for the *CED*, the *Lampeter Corpus* and the *EMET*.

NORMALIZED VERSION	MANUALLY EMENDED VERSION
No Simpson said I, well said he *prithe* come to us. So I was with him walking a little while, and then this Blunt and one Henry Howard were playing one with an-other, throwing stones at one anothers Shins.	No Simpson said I, well said he *prithee* come to us. So I was with him walking a little while, and then this Blunt and one Henry Howard were playing one with an-other, throwing stones at one anothers Shins.

Table 2. Manual emendation of the word "prithe"
in *The Trial of Thomas White alias Whitebread* (1679).

Setting a high threshold ensured that terms were normalized only when the software had a "high confidence" in its top-ranked candidate normalization (Baron 2011, 141), resulting in a higher precision level.

There are several challenges in the normalization process, including terms that may not be present in the dictionary, such as proper nouns,[13] encoded words, words in other languages (e.g., Latin, which is frequent in the *EMET*), and words that are not part of the modern list, such as "betwixt" and "how-

13 For this reason, the names of the files are composed by the year when the trial was held and name of the (main) defendant(s), with the exception of the documents (re) printed in Dublin (for instance, "1679 RDU Thomas White alias Whitebread".txt). In fact, in the above-mentioned documents, the name of the (main) defendant(s) is preceded by the acronym (R)DU, which stands for '(reprinted in) Dublin'. The choice was made in order to possibly divide the corpus into two sub-corpora, depending on the place where the record was printed. In other words, the file names are not constituted by an acronym, as it is common in other corpora. For instance, in the *Corpus of English Dialogues*, "[t]he name of the text file has eight or fewer characters. In the file names, the first character is D for 'dialogues corpus'. The second is the subperiod number. The third character, or third and fourth characters, is the code for text type [T (Trial), W (Witness Deposition), C (Drama Comedy), HO (Didactic Work, other than Language Teaching handbook), HF (Language Teaching handbook, with French as the target language), HE (Language Teaching handbook, with English as the target language), HG (Language Teaching handbook, with German as the target language), F (Fiction), M (Miscellaneous). The remaining characters consist of the first five letters (or initials) of the name of the author, the defendant (or initials of defendants), the place of the speech event, or a keyword from the short title" (Kytö and Walker 2006, 33); for instance. D1TNORFO.

beit" (Ibid.: 56). The large normalization ranges observed[14] are believed to be caused by the inclusion of a significant number of proper nouns in the trials. As mentioned in section 2, only trials in the dialogue format, which resemble plays, were selected for this research, and nouns tend to have a high frequency of occurrences in such formats. However, it should be noted that while the normalized corpus produced is a viable substitute for corpus analysis, a fully normalized and manually checked corpus would be the ideal choice for publication (Ibid.: 170).

4. *Linguistic Annotation*

Despite the existing various types of annotation, everyone of each capable of enhancing the value of a corpus (Aijmer and Rühlemann 2015, 6; Paquot and Gries 2020, 25), it was decided that in the first stages of the research only linguistic annotation would be added. The main reason of this choice is practical. In fact, the automatic process of tagging[15] is particularly helpful when managing large corpora such as the *EMET*. Probably, the "annotated [*EMET*] corpus is unlikely to meet all the expectations of a researcher in terms of its categories of annotation, [but] it can still be an invaluable resource" (Paquot and Gries 2020, 26) and, thus, a great help.

14

WORDS	Total Words		Variant Forms		+	Normalised		=	Originally Variants		Variants Normalised		Not Variants	
SUM	129362		13509			9717			23226				106136	
AVERAGE	2192,58		229	10,44%		164,7	7,51%		393,66	17,95%	41,84%		1798,9	82,05%
MAX	5492		1042	28,13%		403	11,42%		1371	37,01%	63,82%		4313	92,26%
MIN	235		20	4,33%		9	3,34%		29	7,74%	24,00%		206	62,99%

Table 2. Statistics about words.

	Total tokens		Variant Forms		+	Normalised		=	Originally Variants		Variants Normalised		Not Variants	
SUM	1691415		53981			31068			85049				2E+06	
AVERAGE	28668,05		914,9	3,19%		526,6	1,84%		1441,5	5,03%	36,53%		27227	94,97%
MAX	123899		4490	7,19%		2163	4,39%		5709	10,57%	66,60%		119506	96,99%
MIN	503		31	1,52%		15	0,74%		46	3,01%	21,35%		457	89,43%

Table 3. Statistics about tokens.

15 Lately, corpus linguistics is focusing on the development of new methods to automatically annotate corpora (Paquot and Gries 2020, 26).

The process was conducted automatically; the role of VARD 2.5.4 has been fundamental since it is a 'pre-processor to other corpus linguistic tools' [among them Parts Of Speech (POS) tagging], which aims to improve their accuracy (VARD User Guide 2013). In short, POS tagging consists in labelling (or tagging) "each word of a corpus with information about the grammatical category of the word at issue (e.g., noun, verb, adjective, etc.)" (Ibid.). However, POS tagging is strictly bound to tokenization and, thus, to the concept of token. The term 'token' is sometimes considered a synonym of 'word'. Nevertheless, this simplification may induce to think that tokenization is an absolute concept and that there is one and one only tokenization possible. Instead, depending on the decisions, the results vary, and the term should not be considered a synonym of 'word'. Indeed, a token can be defined as "an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing" (Manning et al. 2008, 22). For instance, 'aren't' could be tokenized in the following ways (Ibid.): 1) aren't, 2) arent, 3) are / n't, 4) aren / t.

The discussion around tokenization is not the focus of the present essay and of the study that required the development of the *EMET*. Consequently, since #LancBox allows an easy change of tokenization if the corpus is reloaded on the software, the standard parameters were set and POS tags were automatically added to the corpus.

TOKEN DELIMITERS	\t\n \r
LEMMA	Include POS groups
POS	Tagging
PUNCTUATION	. , : ; ? " ! 。 ， ； ： ？ ！ ， ¿ ¡ … ’ ” ‘ ’ ` “ ” „ () <=> [] {} 《》<> 《》 ---—*
SENTENCE DELIMITERS	(?s) . * [\. | ! | \? | 。 | ？ | ！]. *

Table 4. Standard tokenization parameters in #LancsBox.

5. *Concluding Remarks*

The present essay has highlighted that it is essential to normalize corpora of documents written in Early Modern English. Indeed, the standardization between 1500 and 1700 was gradual: the English language throughout the period still presented a marked spelling variation and was still undergoing major

changes. Thus, normalization ensures the consistency and standardization of data, eliminating (most) spelling variation and allowing for an accurate and reliable analysis of the corpus object of study, which appears more homogeneous, thus, POS tagging can also be applied. Furthermore, it enables the comparison of the language in a corpus across different documents and speakers, as well as with other corpora.

It should also be noted that, when consulting archives, the best choice is to consider every spelling variant of the keyword(s), which allows the search of every document including such term(s). In fact, since most archives do not contain normalized versions of the files or of their titles, querying only the contemporary spelling of the keyword(s) would allow the identification and selection of only a restricted group of documents. In order to identify the Early Modern variants of a word, the consultation of both the *LEME* and an etymological dictionary ensures accuracy.

Moreover, the comparison between the normalization parameters of the *Corpus of English Dialogues 1560-1760*, the *Lampeter Corpus of Early Modern English Tracts* and the corpus of *Early Modern English Trials* seems to suggest that the best decision for Early Modern English texts is to adopt a 75% replacement threshold. Indeed, despite the documents included in the above-mentioned corpora are of different genres, among them trials, witness depositions, prose, handbooks, comedy drama, and pamphlets about religion, politics, science, law, economy and trade, it is with such threshold that the best results are obtained.

In conclusion, the analysis of corpora of Early Modern texts might be troublesome because of spelling variation. Thus, normalization is the key to (try to) overcome such issues since it facilitates effective searching, and thus, the analysis of the corpus itself.

Bibliography

Aijmer, Karin and Rühlemann, Christoph. 2015. *Corpus Pragmatics: A Handbook*. Cambridge: Cambridge University Press.

Aliprandi, Giuseppe and Pigò, Artiode. 1936. "Stenografia." In *Enciclopedia Italiana*.

Archer, Dawn, Kytö, Merja, Baron, Alistair and Rayson, Paul. 2015. "Guidelines for Normalising Early Modern English Corpora: Decisions and Justifications." *ICAME Journal* 39, no. 1: 5-24.

Barber, Charles. 1976. *Early Modern English*. London: Deutsch.

Baron, Alistair and Rayson, Paul. 2008. "VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora." In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Birmingham, UK, Aston University.

Baron, Alistair. 2011. *Dealing with Spelling Variation in Early Modern English Texts*. PhD Dissertation, Lancaster University.

Baron, Alistair, Rayson, Paul and Archer, Dawn. 2011. "Innovators of Early Modern English Spelling Change: Using DICER to Investigate Spelling Variation Trends." *Helsinki Corpus Festival*.

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Brezina, Vaclav, Timperley, Matthew and McEnery, Tony. 2018. #LancsBox v. 4.x [software]. Available at: http://corpora.lancs.ac.uk/lancsbox.

CED = A Corpus of English Dialogues 1560-1760. 2005. Compiled by Merja Kytö (Uppsala University, Sweden) and Jonathan Culpeper (Lancaster University, England).

Culpeper, Jonathan and Kytö, Merja. 1997. "Towards a Corpus of Dialogues, 1550-1750." In *Language in Time and Space: Studies in Honour of Wolfgang Viereck on the Occasion of his 60th Birthday*, edited by Heinrich Ramisch and Kenneth Wynne, 60-73. Stuttgart: Franz Steiner Verlag.

Culpeper, Jonathan and Kytö, Merja. 2014. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge University Press.

Doty, Kathleen L. 2007. "Telling Tales: The Role of Scribes in Constructing the Discourse of the Salem Witchcraft Trials." *Journal of Historical Pragmatics* 8, no. 1: 25-41.

Gablasova, Dana, Brezina, Vaclav and McEnery, Tony. 2019. "The Trinity Lancaster Corpus: Development, Description and Application." *International Journal of Learner Corpus Research* 5, no. 2: 126-58.

Görlach, Manfred. 1991. *Introduction to Early Modern English*. Cambridge: Cambridge University Press.

Kytö, Merja and Walker, Terry. 2003. "The Linguistic Study of Early Modern English Speech-Related Texts: How 'Bad' can 'Bad' Data Be?" *Journal of English Linguistics* 31, no. 3: 221-48.

Kytö, Merja and Walker, Terry. 2006. *Guide to A Corpus of English Dialogues 1560-1760.* Uppsala: Acta Universitatis Upsaliensis.

Lass, Roger. 1999. *The Cambridge History of the English Language*, vol 3. Cambridge: Cambridge University Press.

Langbein, John H. 1978. "The Criminal Trial before the Lawyers." *The University of Chicago Law Review* 45, no. 2: 263-316.

Lehto, Anu, Baron, Alistair, Ratia, Maura and Rayson, Paul. 2010. "Improving the Precision of Corpus Methods: The Standardized Version of Early Modern English Medical Texts." In *Early Modern English Medical Texts: Corpus Description and Studies*, edited by Irma Taavitsainen and Päivi Pahta, 279-90. Amsterdam: John Benjamins.

Lexicons of Early Modern English. 2014. Ed. Ian Lancashire. Toronto, ON: University of Toronto Library and University of Toronto Press. leme.library.utoronto.ca.

Manning, Christopher D., Raghavan, Prabhakar and Schütze, Hinrich. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

McEnery, Tony and Hardie, Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, Tony and Wilson, Andrew. 2001. *Corpus Linguistics*, 2nd ed. Edinburgh: Edinburgh University Press.

McEnery, Tony, Xiao, Richard and Tono, Yukio. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. Milton Park: Taylor & Francis.

Nevalainen, Terttu. 1999. "Making the Best Use of 'Bad' Data: Evidence for Sociolinguistic Variation in Early Modern English." *Neuphilologische Mitteilungen* 100, no. 4: 499-533.

Onions, Charles Talbut, Friedrichsen, George Washington Salisbury and Burchfield, Robert William. 1996. *The Oxford Dictionary of English Etymology*, vol. 178. Oxford: Clarendon Press.

Palander-Colin, Minna and Hakala, Mikko. 2011. "Standardising the *Corpus of Early English Correspondence* (CEEC)." Poster presented at *ICAME 32*, Oslo, 1-5 June 2011.

Nevalainen, Terttu. 2006. *An Introduction to Early Modern English*. Edinburgh: Edinburgh University Press.

Paquot, Magali and Gries, Stefan T. 2020. *A Practical Handbook of Corpus Linguistics*. Cham: Springer.

Schmied, Joseph. 1994. "The Lampeter Corpus of Early Modern English Tracts." In *Corpora Across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, 25-27 March 1993* no. 11, edited by Kytö, Merja, Rissanen, Matti and Wright, Susan, 81-90. Amsterdam: Rodopi.

Short, Mick, Semino, Elena and Wynne, Martin. 2002. "Revisiting the Notion of Faithfulness in Discourse Presentation Using a Corpus Approach." *Language and Literature* 11, no. 4: 325-55.

The Proceedings and Trial in the Case of the Most Reverend Father in God William Lord Archbishop of Canterbury, And the Right Reverend Fathers in God, William Lord Bishop of St. Asaph, Francis Lord Bishop of Ely, Iohn Lord Bishop of Chichester, Thomas Lord Bishop of Bath and Wells, Thomas Lord Bishop of Peterborough, And Ionathan Lord Bishop of Bristol. In the Court of Kings-Bench at Westminster, in Trinity-Term in the Fourth Year of the Reign of King

Iames the Second, Annoque Dom. 1688. 1689. London: Thomas Bassett, https://www.proquest.com/books/proceedings-tryal-case-most-reverend-father-god/docview/2240860766/se-2 (accessed 1 November 2023).

The Trial and Condemnation of Dr Oliver Plunket Titular Primate of Ireland, for High-Treason, At the Barr of the Court of King's Bench, at Westminster, in Trinity Term 1681. 1681. Dublin: Joseph Ray for Eliphal Dobson, https://www.proquest.com/books/tryal-condemnation-dr-oliver-plunket-titular/docview/2264215874/se-2 (accessed 1 November 2023).

The Trial of John Giles at The Sessions-House In The Old Bailey: Held by Adjournment from the 7th Day of July, 1680, until the 14th Day of the same Month: The Adjournment being appointed on purpose for the said Giles his Trial, for a Barbarous and Inhumane Attempt, to Assassinate and Murder John Arnold. 1681. London: Thomas James for Randal Taylor, https://www.proquest.com/books/tryal-john-giles-at-sessions-house-old-bayly-held/docview/2248508327/se-2 (accessed 1 November 2023).

UCREL Semantic Analysis System, n.a., https://ucrel.lancs.ac.uk/usas/ (accessed 13 April 2023).

"VARD User Guide 2013," http://ucrel.lancs.ac.uk/vard/userguide/ (accessed 13 April 2023).

Walker, Terry. 2007. Thou *and* You *in Early Modern English Dialogues: Trials, Depositions, and Drama Comedy*. Amsterdam: John Benjamins Publishing Company.

Emma Pasquali is adjunct professor of English Language and Translation and English Literature at the eCampus University of Novedrate. She was awarded a PhD from the University of Naples "L'Orientale". Her research interests lie in corpus linguistics, stylistics, cognitive poetics and historical pragmatics. Her recent work includes studies on stylistics and cognitive poetics (2022, "Contamination between Stylistics and Cognitive Poetics: An Analysis of Lord Randal". *Testo e Senso* and 2022, "The Paralysing Sea: A Cognitive Analysis of the Discourse Worlds in James Joyce's 'Eveline'". *Ticontre. Teoria Testo e Traduzione*).