Standardizing the *Corpus of Early English Correspondence* (CEEC)

Minna Palander-Collin
Department of Modern Languages
minna.palander-collin@helsinki.fi

Mikko Hakala
Department of Modern Languages
mikko.hakala@helsinki.fi

Introduction

•Spelling variation in Early Modern English poses considerable problems for the accuracy of corpus linguistic methods. Private writings like personal letters in particular show extensive spelling variation reflecting regional and social differences, but it is not uncommon for a single writer to exhibit spelling variation.

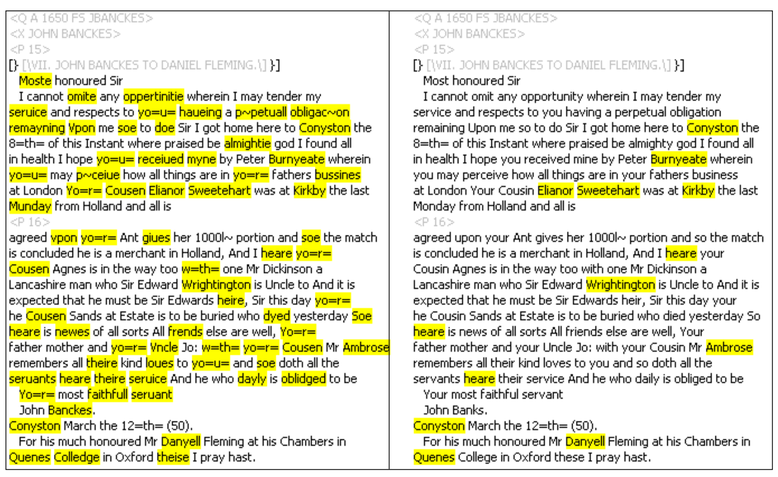
•Here we report on the standardization process of CEEC, compare the CEEC results with a public genre and analyze the impact of standardization on cluster and keyword analyses.

Standardization

• Spelling in CEEC standardized with VARD 2 (Variant Detector), software developed specifically for dealing with spelling variation in Early Modern English texts.[1]

• Two-stage process

1) *Manual training*: Extracts from the corpus are standardized manually with VARD 2 to prepare the software for automatic processing by adding to its inventory of potential replacements for different variant forms and improving its precision in choosing correct replacements.
2) *Automatic standardization*: With enough training, VARD 2 can process all the texts in a corpus automatically to produce a standardized version.



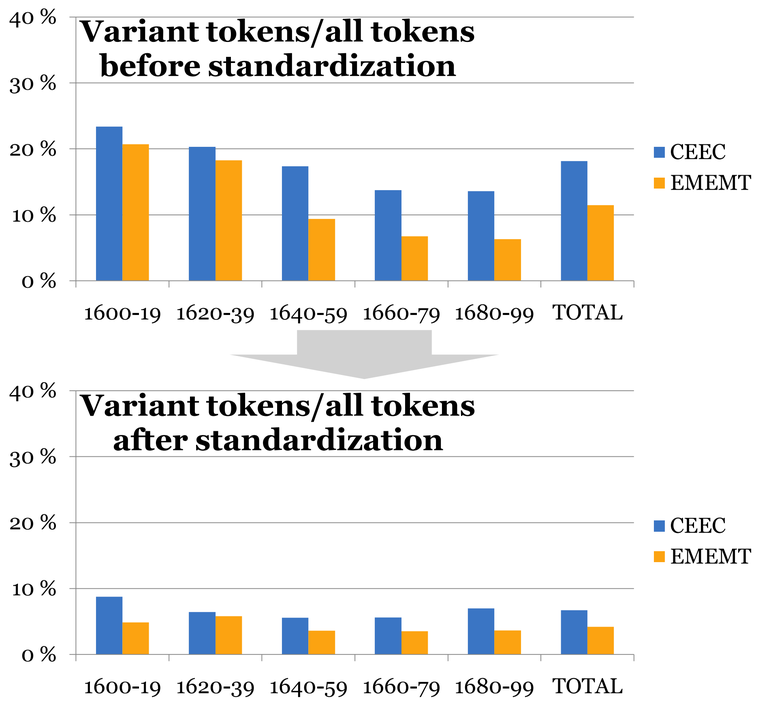
<Q A 1650 FS JBANCKES> <X JOHN BANCKES> <P 15> [} [VII. JOHN BANCKES TO DANIEL FLEMING.\] }] Moste honoured Sir I cannot omite any oppertinitie wherein I may tender my seruice and respects to yo=u= haueing a p~petuall obligac~on remayning Vpon me soe to doe Sir I got home here to Conyston the 8=th= of this Instant where praised be almightie god I found all in health I hope yo=u= receiued myne by Peter Burnyeate wherein yo=u= may p~ceiue how all things are in yo=r= fathers bussines at London Yo=r= Cousen Elianor Sweetehart was at Kirkby the last Munday from Holland and all is <P 16> agreed vpon yo=r= Ant giues her 1000l~ portion and soe the match is concluded he is a merchant in Holland, And I heare yo=r= Cousen Agnes is in the way too w=th= one Mr Dickinson a Lancashire man who Sir Edward Wrightington is Uncle to And it is expected that he must be Sir Edwards heire, Sir this day yo=r= he Cousen Sands at Estate is to be buried who dyed yesterday Soe heare is newes of all sorts All frends else are well, Yo=r= father mother and yo=r= Vncle Jo: w=th= yo=r= Cousen Mr Ambrose remembers all theire kind loues to yo=u= and soe doth all the seruants heare theire seruice And he who dayly is oblidged to be Yo=r= most faithfull seruant John Banckes. Conyston March the 12=th= (50). For his much honoured Mr Danyell Fleming at his Chambers in Quenes Colledge in Oxford theise I pray hast.	<Q A 1650 FS JBANCKES> <X JOHN BANCKES> <P 15> [} [VII. JOHN BANCKES TO DANIEL FLEMING.\] }] Most honoured Sir I cannot omit any opportunity wherein I may tender my service and respects to you having a perpetual obligation remaining Upon me so to do Sir I got home here to Conyston the 8=th= of this Instant where praised be almighty god I found all in health I hope you received mine by Peter Burnyeate wherein you may perceive how all things are in your fathers business at London Your Cousin Elianor Sweetehart was at Kirkby the last Monday from Holland and all is <P 16> agreed upon your Ant gives her 1000l~ portion and so the match is concluded he is a merchant in Holland, And I heare your Cousin Agnes is in the way too with one Mr Dickinson a Lancashire man who Sir Edward Wrightington is Uncle to And it is expected that he must be Sir Edwards heir, Sir this day your he Cousin Sands at Estate is to be buried who died yesterday So heare is news of all sorts All friends else are well, Your father mother and your Uncle Jo: with your Cousin Mr Ambrose remembers all their kind loves to you and so doth all the servants heare their service And he who daily is obliged to be Your most faithful servant John Banks. Conyston March the 12=th= (50). For his much honoured Mr Danyell Fleming at his Chambers in Quenes College in Oxford these I pray hast.

Screenshots of a letter in VARD 2 before (left) and after (right) automatic standardization

Comparison to EMEMT

•The corpus of *Early Modern English Medical Texts* (EMEMT) is another corpus standardized with VARD 2 at the VARIENG research unit.[2,3]

•How does text type affect the results of standardization?

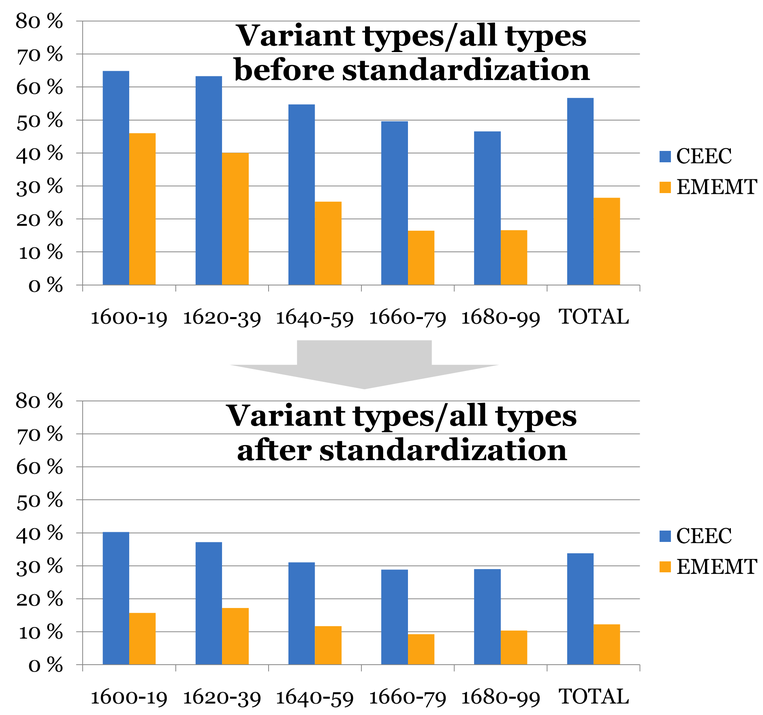


•Word tokens with variant spelling prior to standardization: CEEC 18.2%, EMEMT 11.5%.

•Spelling variation decreases over time in both corpora, but is consistently higher in CEEC.

•After standardization, the frequency of variant tokens is 6.7% in CEEC and 4.2% in EMEMT (62.9% standardization rate for CEEC and 63.4% for EMEMT).

•Initial frequency of variant tokens seems to have very little impact on standardization results.



•Word types with variant spelling prior to standardization: CEEC 56.7%, EMEMT 26.4%.

•The high frequency of variant types in CEEC shows that there are more idiosyncrasies in spelling, possibly due to the higher number of writers, than in EMEMT, and this makes it more difficult to standardize type variation.

•With 33.8% of all word types in CEEC still reflecting variant spelling after standardization (compared to 12.2% in EMEMT), the standardization rate of variant types is considerably lower for CEEC (40.4%) than for EMEMT (53.7%).

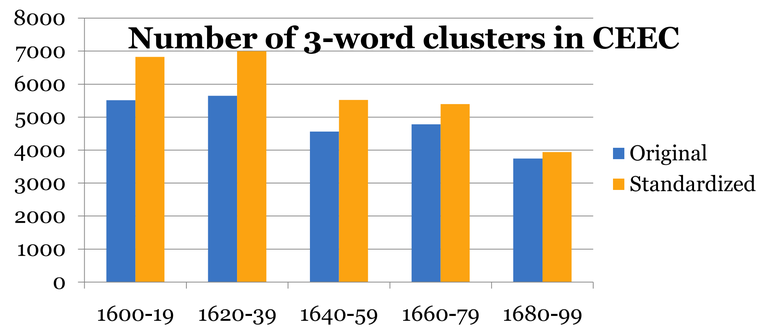
•However, the remaining type variation should not have a significant impact on keyword and cluster analyses as the number of tokens for these types is also quite low.

Effects on clusters and keywords

•Standardization increases the accuracy of cluster and keyword analyses in CEEC.

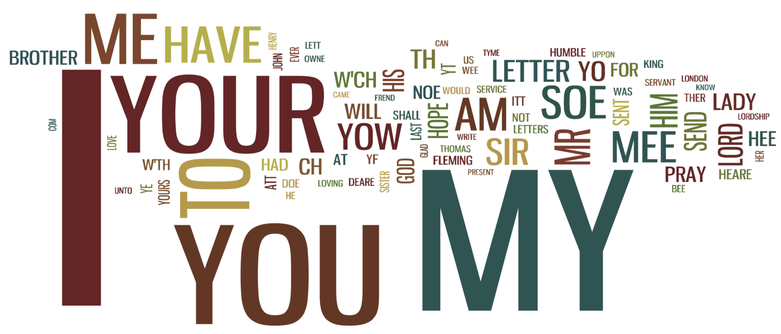
•The impact of standardization on clusters is less dramatic towards the end of the 17th century as spelling conventions become more uniform.

•The number of 3-word clusters extracted with WordSmith[4] increases in the standardized corpus by over 20% until 1659, but only by 5% during 1680-99 in comparison to the original version.

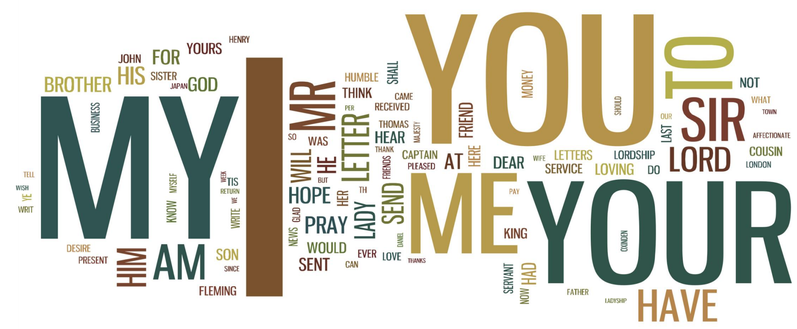


•Most clearly, standardization eradicates the impact of variant spellings: Frequent clusters are even more frequent; keyword lists no longer highlight deviant spellings (e.g. *soe*, *itt*, *yf*).

•However, genre-characteristic words like pronouns also show in the keyword analysis of the original CEEC vs. EMEMT.



Keywords of CEEC (original) vs. EMEMT (original)



Keywords of CEEC (standardized) vs. EMEMT (standardized)

REFERENCES

[1] Baron, Alistair and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham: Aston University.

[2] Taavitsainen, Irma, Päivi Pahta, Turo Hiltunen, Martti Mäkinen, Ville Marttila, Maura Ratia, Carla Suhr and Jukka Tyrkkö (eds). 2010. *Early Modern English Medical Texts*. CD-ROM. Amsterdam: John Benjamins.

[3] Lehto, Anu, Alistair Baron, Maura Ratia and Paul Rayson. 2010. Improving the precision of corpus methods: The standardized version of *Early Modern English Medical Texts*. In Irma Taavitsainen and Päivi Pahta (eds), *Early Modern English Medical Texts: Corpus Description and Studies*. Amsterdam: John Benjamins. 279–289.

[4] Scott, Mike. 2004–2007. *Oxford WordSmith Tools* (4.0). Oxford: Oxford University Press.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS