Prerequisites to a corpus-based analysis of EEBO-TCP

Alistair Baron & Andrew Hardie
Lancaster University

LANCASTER UNIVERSITY

UCREL

Corpus Linguistics with EEBO-TCP

- The transcribed texts from EEBO-TCP provide an unparalleled resource for the corpus linguistic analysis of the Early Modern English period.

- Our current version (Phase 1 and Phase 2 up until June 2011) contains 39,595 texts, totalling approx. 900,000,000 words.

- In order to perform corpus linguistic analysis of EEBO:
  - Metadata needs to be extracted to allow analyses of a particular set of texts and to compare sub-corpora defined by metadata-based filtering.
  - Some “cleanup” is needed to allow for more accurate tokenisation.
  - Spelling variation needs to be normalised, i.e. VARD.
  - Powerful corpus analysis tools are required to deal with the large amount of data, i.e. CQPweb.

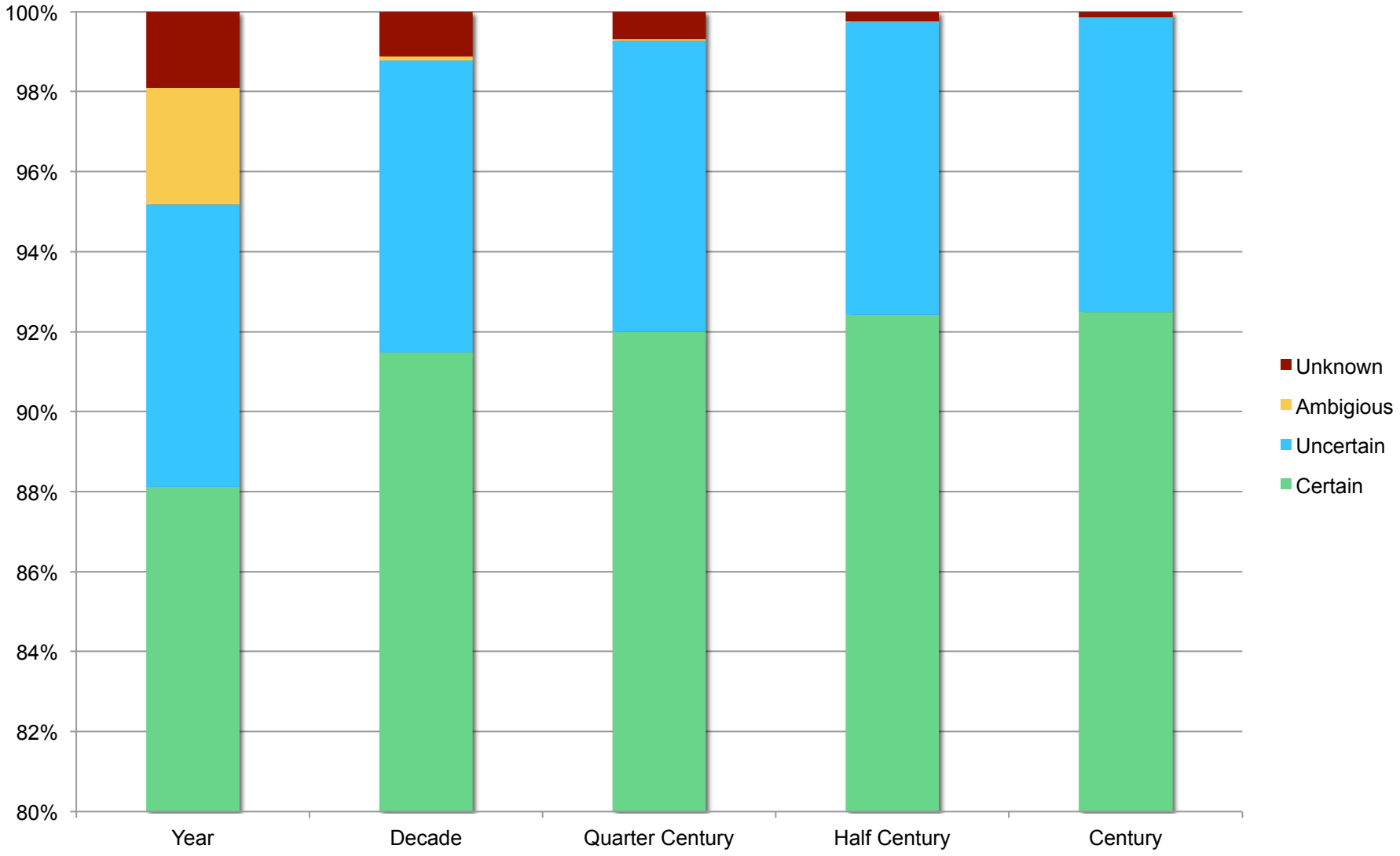
Metadata extraction

- MARC records initially unavailable, so extracted metadata for each text from the header files and from the texts themselves.
- Publication date, converted into year (98.1%), decade (98.9%), quarter century (99.3%), half century (99.8%) & century (99.9%).
- Title (70.2%), Alternative Title(s) (21%) & Additional Title(s) (2.6%)
- Author(s) (84.3%). 1 author (58.7%), 2 (19.1%), 3 (4.6%), 4+ (1.9%).
- Publisher (99.7%) (0.3% of which have 2 publisher fields).
- Publication place (97.89% (3 not present & 832 unknown). 103 have multiple publication place fields, 199 have multiple publication places present.
  - Notes (99.9%, 87.4% have multiple).
  - IDNO's (97.3%, 220 have 2).
  - Bib Names, e.g. STC, Wing (99.8%, 27% have multiple).
  - Key Word Terms (63.8%, 39.4% have multiple).

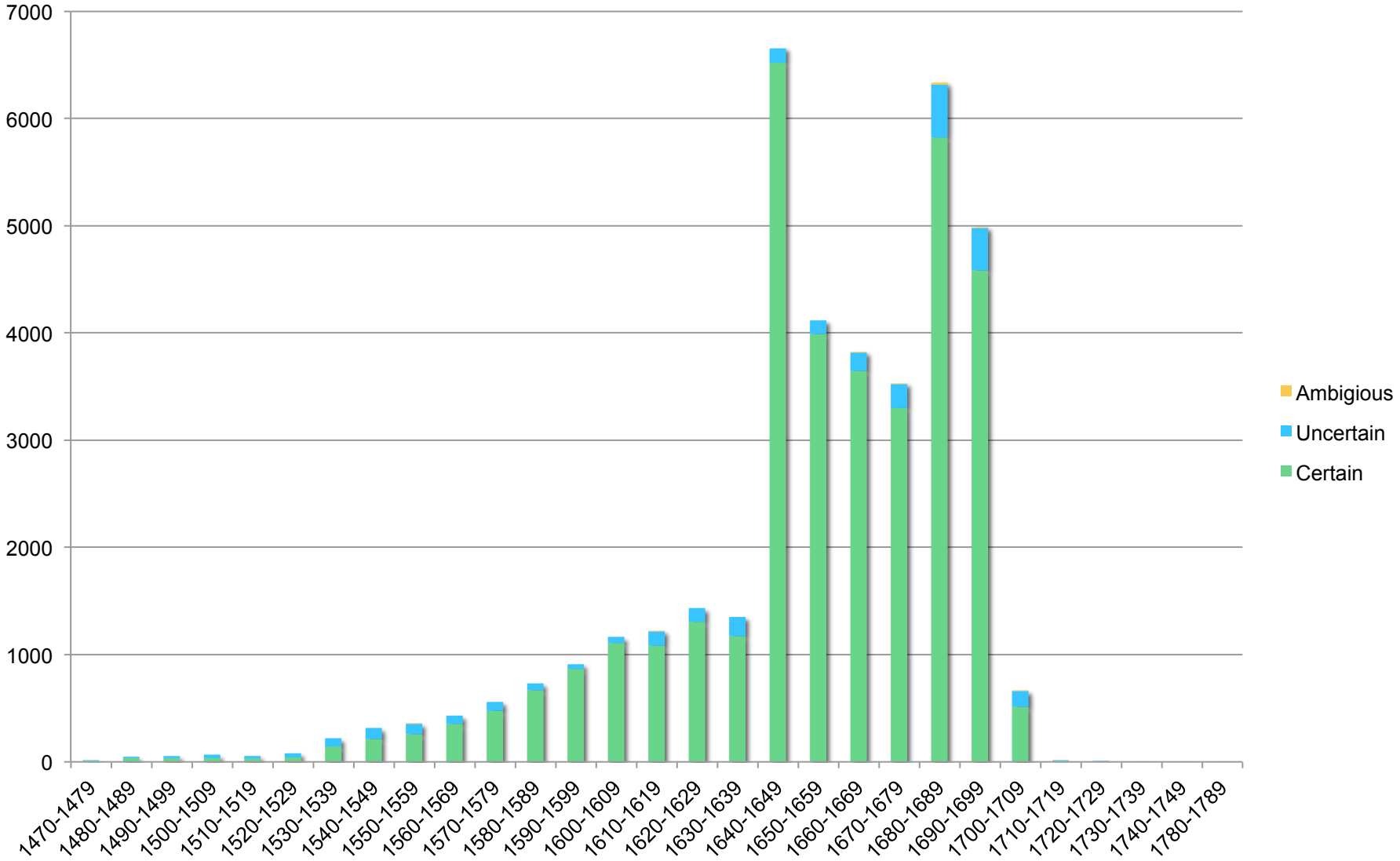
Languages

- 3,734 (9.4%) of texts have no language indicated.
- 35,592 (89.9%) of texts have one language indicated.
- 247 texts have 2 languages indicated. 16 texts have 3 languages. 6 texts have 4 or more languages (most = 8).
- Of the texts which are indicated as one language, the vast majority (98.5%) are marked as English. Others include:
  - Latin: 279 texts (0.72%)
  - Welsh: 125 texts (0.67%)
  - French: 43 texts (0.12%)
  - Plus a few each of Dutch, Middle French, German, Hebrew, Italian, Portuguese, Scots & Spanish.

Date certainty / ambiguity



Texts per decade

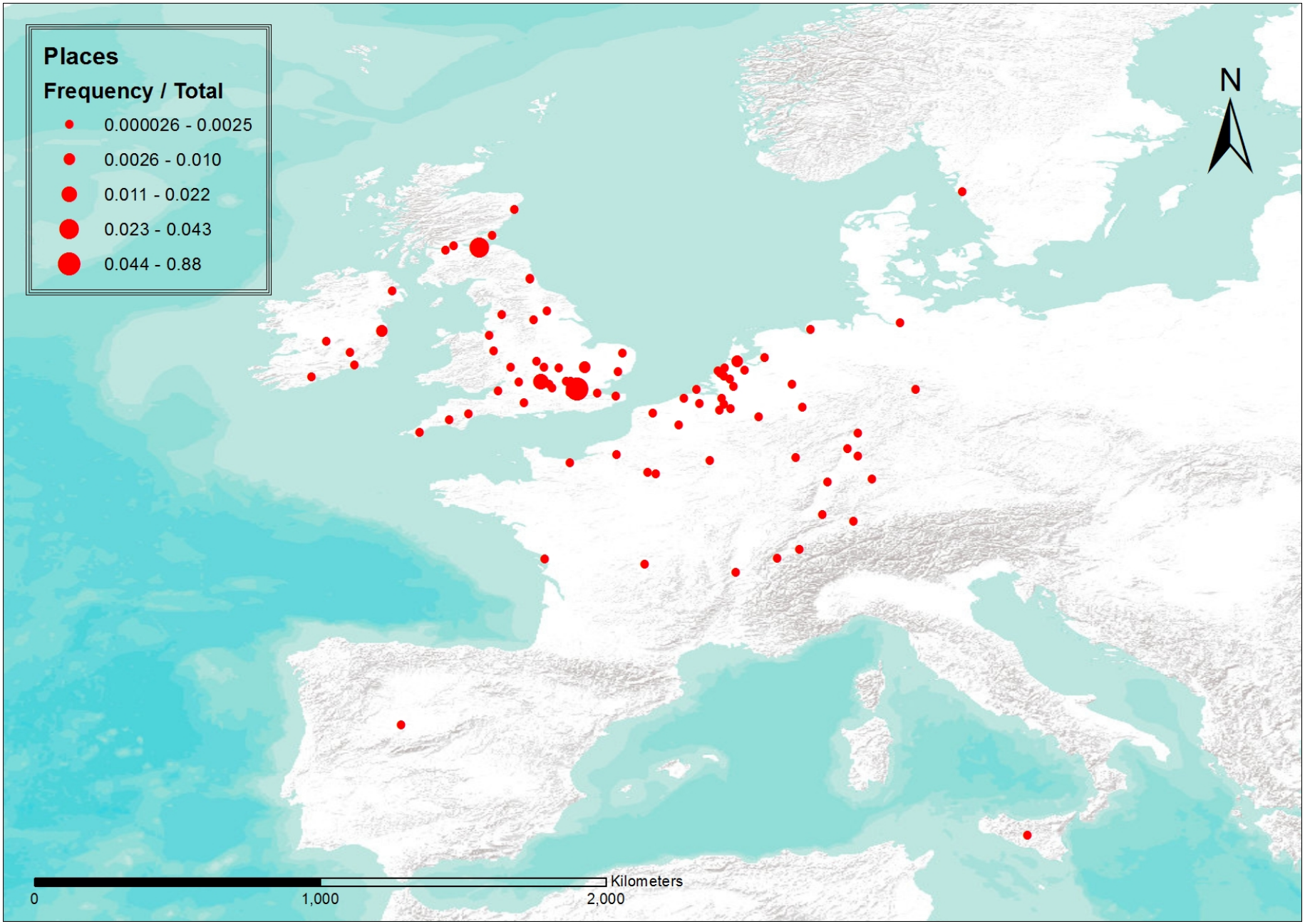


Publication Place

- Publication place cleaned up to only include place name. (e.g. “printed in”, etc. removed).

- For each text with a single publication place present (38,458), the place name was manually normalised to the modern spelling. For 30 cases, the place name could not be found.

- Top 5 places: London (88%), Edinburgh (4.3%), Oxford (2.2%), Dublin (1%) & Cambridge, England (0.7%)


Places
Frequency / Total
0.000026 - 0.0025
0.0026 - 0.010
0.011 - 0.022
0.023 - 0.043
0.044 - 0.88
N
0
1,000
2,000
Kilometers

Pre-processing texts

- Some problems in the texts which will have an impact on accurate tokenisation:

- | and ¦ removed.

- “<SEG REND="decorInit">W</SEG>Ere it as” (94,492 cases)
  - → “WEre it as”

- “y^e” (430,384 cases)
  - → “y^t”

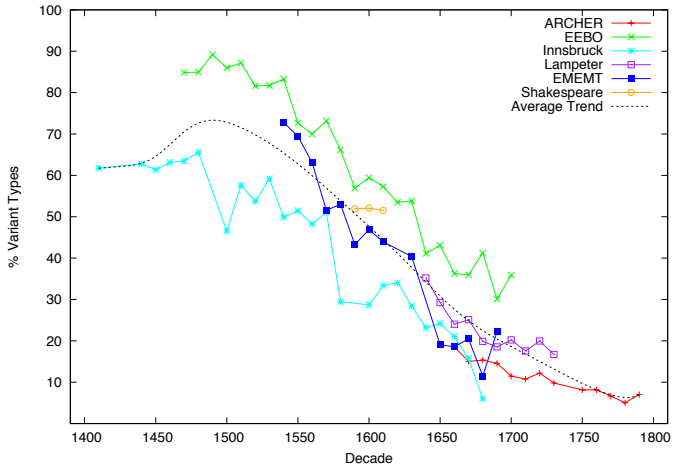
- “wor<GAP […] EXTENT="1 letter” […]/>d” (1,646,975 cases)
  - → “wor~d”

Spelling variation

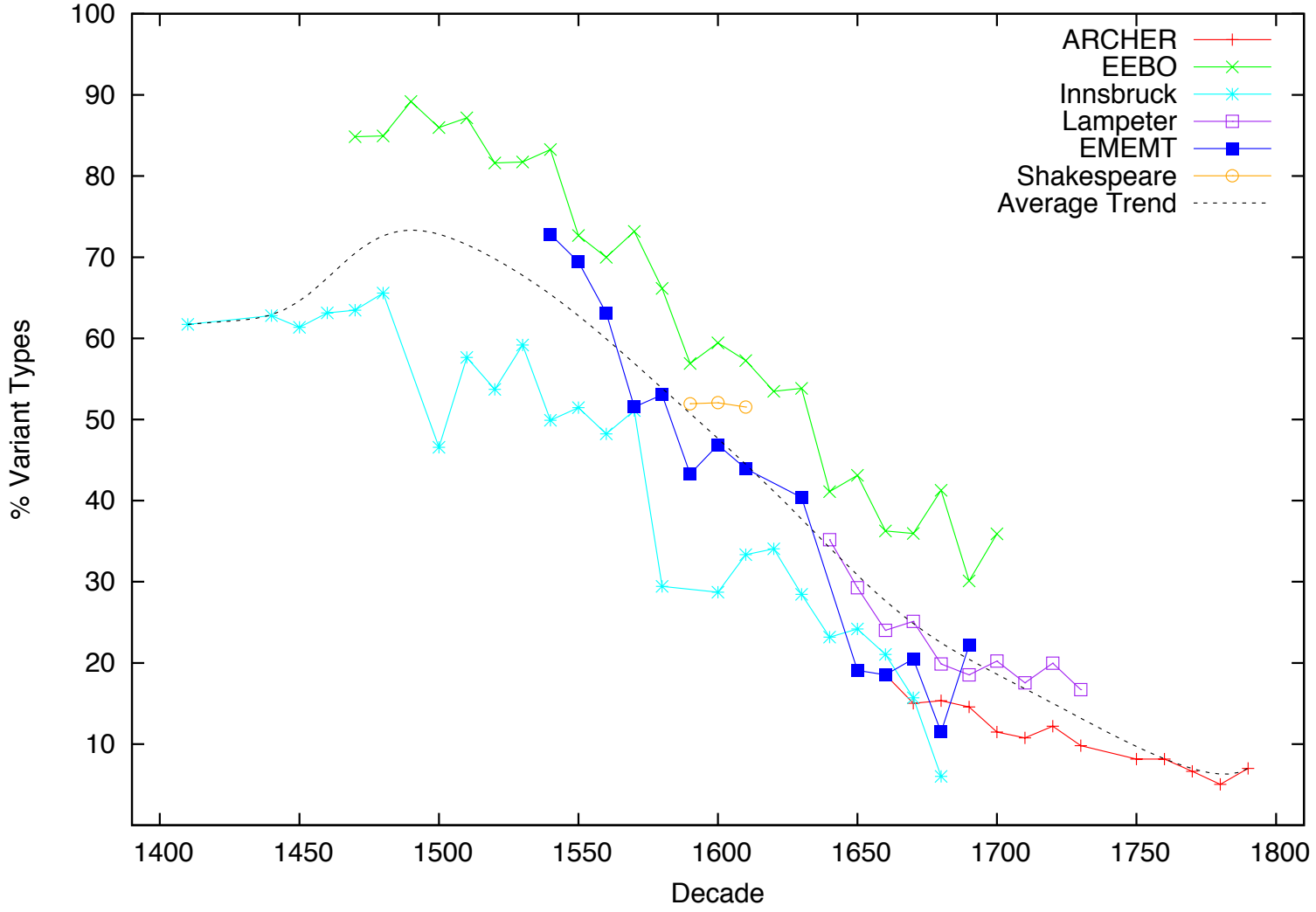
- Large amount spelling variation in Early Modern English texts.
  - No notion of the importance of having a single spelling for each word.
  - Letters would be added or removed to ease line justification.
- Spelling variation became less frequent over the period (Baron *et al.*, 2009).
  - Spread of London and Chancery English.
  - Dictionaries (Samuel Johnson, 1755)

Though I **speake** with the tongues of men & of Angels, and **haue** not charity, I am become as sounding **brasse** or a tinkling cymbal. And though I **haue** the gift of **prophesie**, and **vnderstand** all mysteries and all knowledge: and though I **haue** all faith, so that I could **remooue mountaines**, and **haue** no **charitie**, I am nothing...

(Authorised Version of the Bible, 1611)



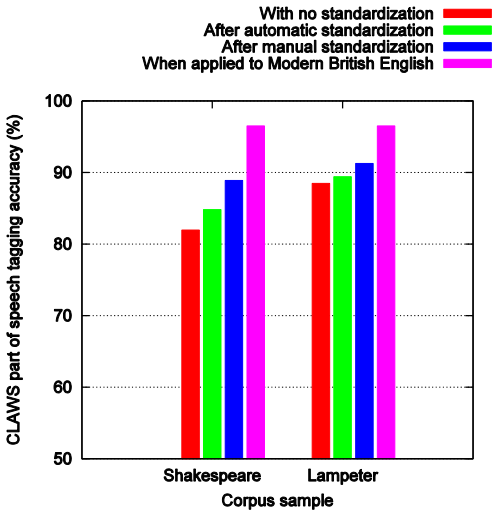
EModE Spelling Variation



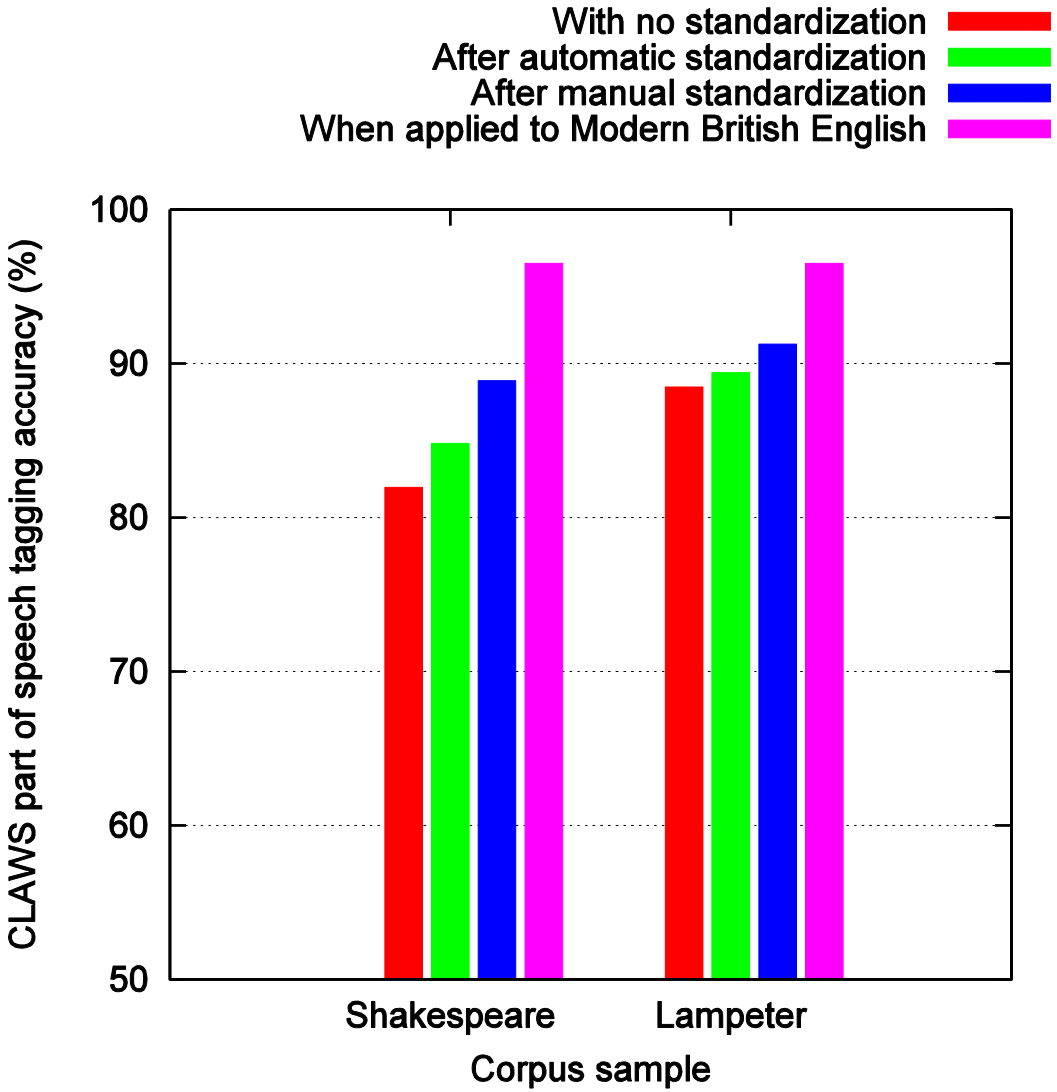
(Baron *et al.*, 2009)

Effect on corpus linguistics

- Searching for words can be problematic: *would, wolde, woolde, wuld, wulde, wud, wald, vvould, vvold*, etc.
- Frequencies split by multiple spellings.
- Knock-on effect on key words (Baron *et al.*, 2009), key word clusters (Palander-Collin & Hakala, 2011) and collocates.
- Automatic annotation will also be affected, e.g. Part of speech tagging (Rayson *et al.*, 2007) and Semantic annotation (Archer *et al.*, 2003).



Spelling variation effect on POS-Tagging



VARD 2

- Freely available for academic use: http://ucrel.lancs.ac.uk/vard
- Designed to assist researchers in standardising spelling variation in historical corpora both manually and automatically.
- Uses methods from modern spellchecking to find spelling variants and offer/select appropriate modern equivalents.
- The original spelling is always retained in the text with an xml tag surrounding the replacement.
  - <normalised orig=”reuenge”>revenge</normalised>
- Allows for the use of standard corpus linguistics tools without any modification.
- Used to normalise released historical (and other) corpora, e.g. EMEMT (Lehto *et al.*, 2010) and CEEC (Palander-Collin & Hakala, 2011).

VARDing EEBO

- VARD trained with manually normalised texts:
  - Innsbruck Letters Corpus (fully normalised) (Markus, 1999)
  - EMEMT samples (Lehto *et al.*, 2010)
  - CEEC samples (Palander-Collin & Hakala, 2011)

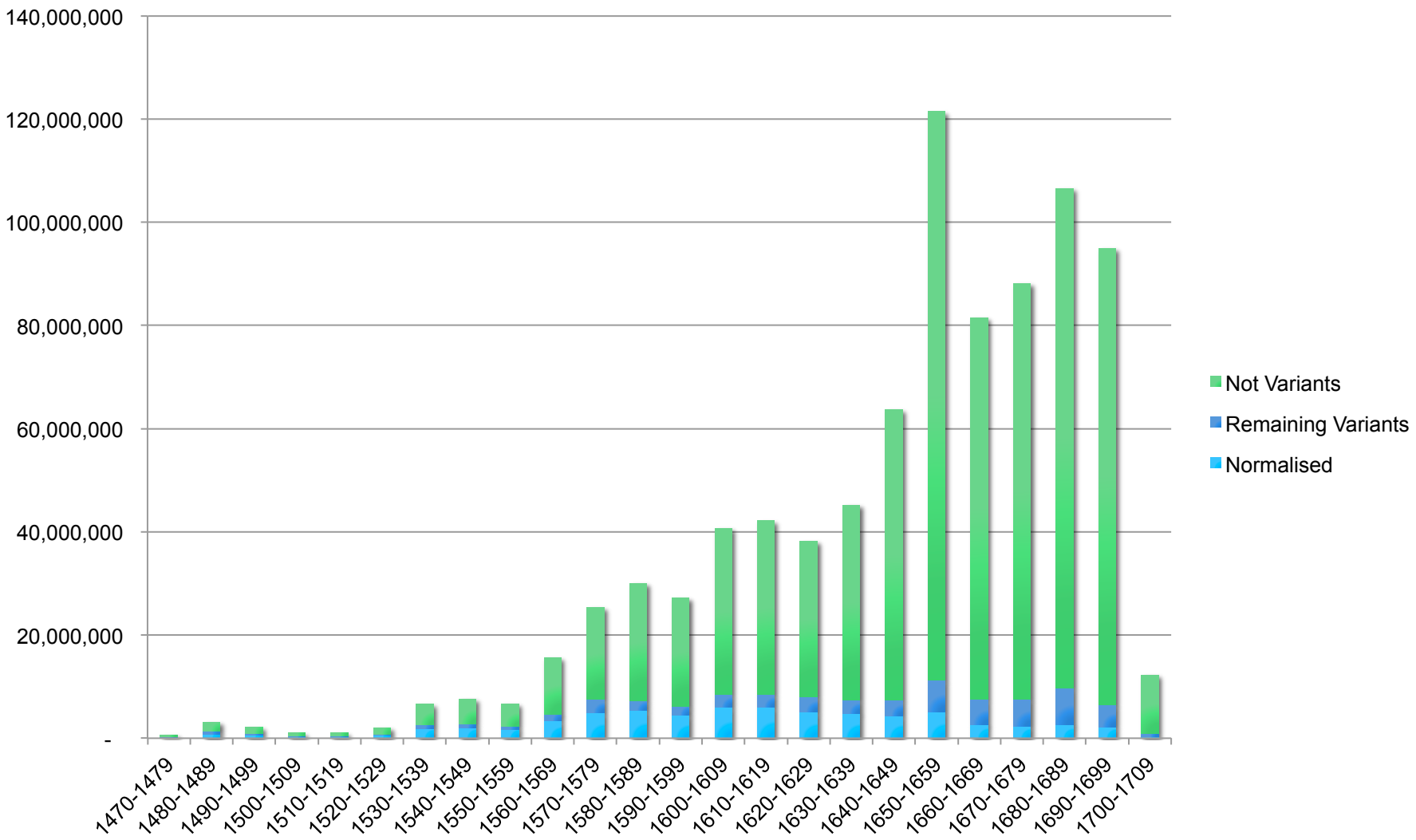
- All texts automatically normalised with a confidence score threshold of 50% set.

- 126,059,275 (14.2%) words detected as variants.

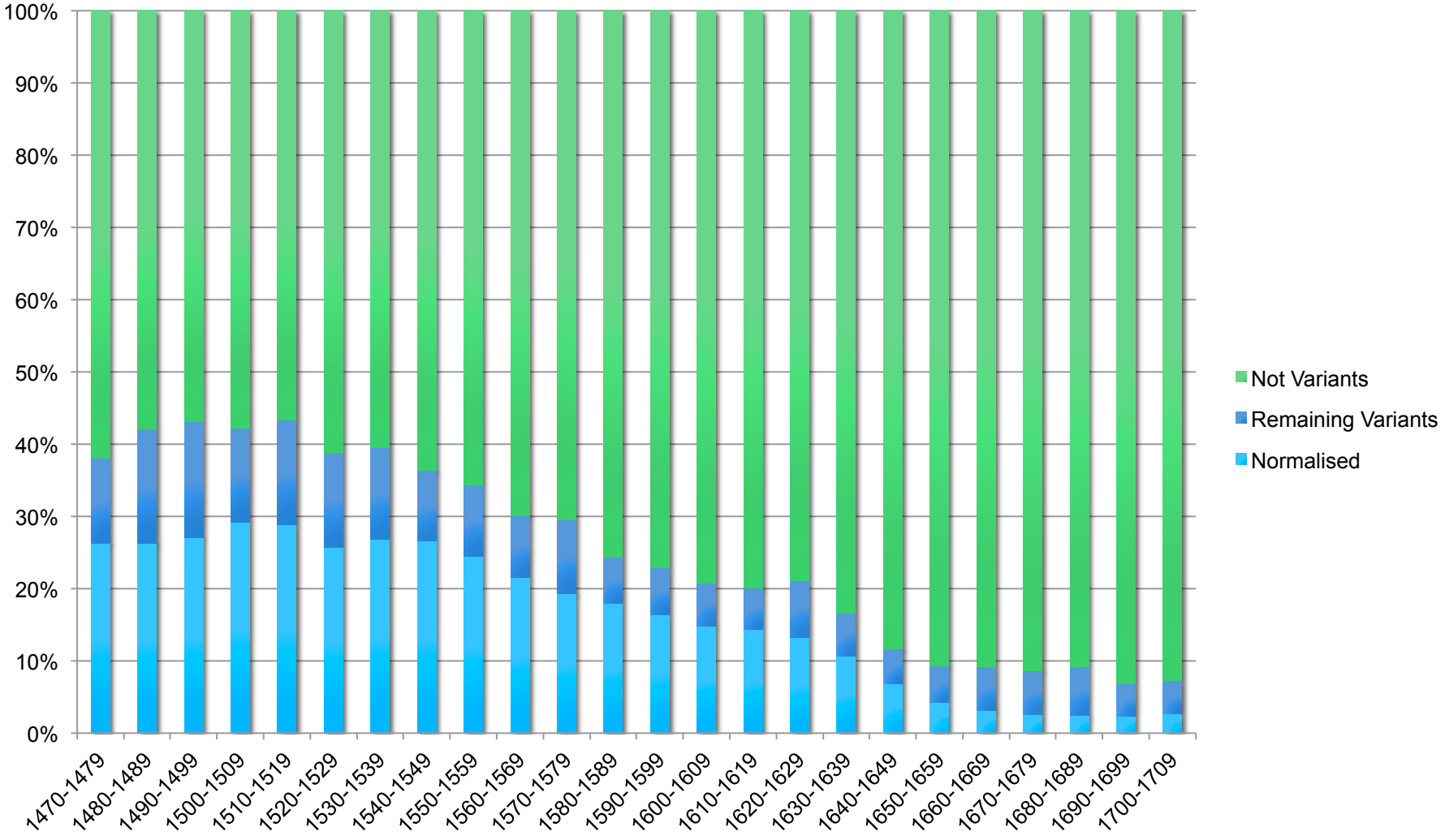
- 71,044,697 (56.4%) of these were automatically normalised.

- Leaving 55,014,578 (6.2%) of words left as detected variants.

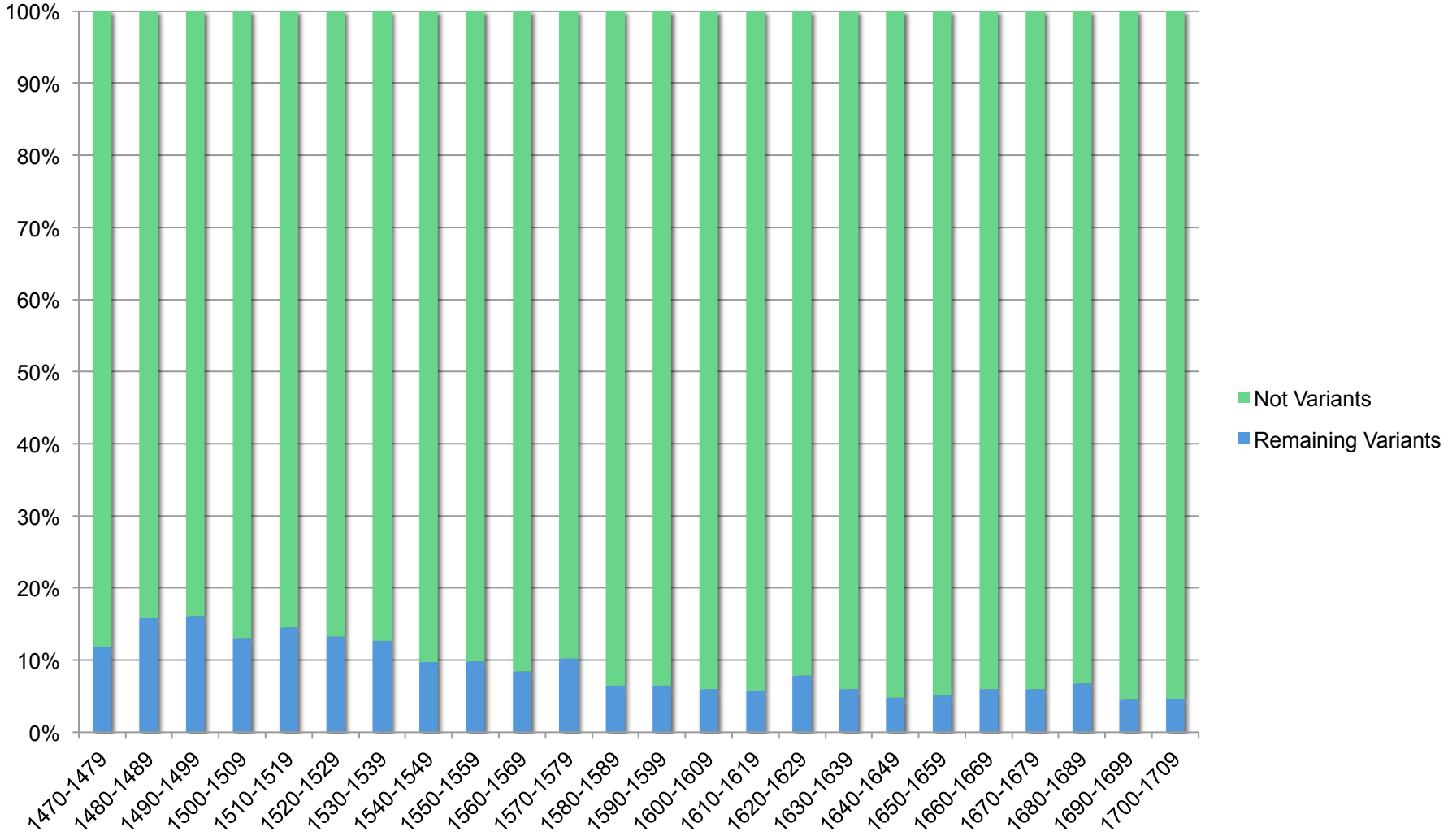
Number of words per decade



Proportion of spelling variants per decade



Variant levels after normalisation



Archer, D., McEnery, T., Rayson, P. & Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In D. Archer, P. Rayson, A. Wilson & T. Mcenery, eds., Proceedings of Corpus Linguistics 2003, 22–31, Lancaster University, Lancaster, UK.

Baron, A., Rayson, P. and Archer, D. (2009). Word frequency and key word statistics in historical corpus linguistics. Anglistik: International Journal of English Studies, 20 (1), pp. 41–67.

Lehto, A., Baron, A., Ratia, M. and Rayson, P. (2010). Improving the precision of corpus methods: The standardized version of Early Modern English Medical Texts. In Taavitsainen, I. and Pahta, P. (eds.) Early Modern English Medical Texts: Corpus description and studies, pp. 279–290. John Benjamins, Amsterdam.

Markus, M. (1999). Innsbruck Computer-Archive of Machine-Readable English Texts. In Innsbrucker Beitraege zur Kulturwissenschaft, Anglistische Reihe, vol. 7, Leopold-Franzens-Universitaet Innsbruck, Institut fuer Anglistik, Innsbruck.

Palander-Colin, M. and Hakala, M. (2011). Standardizing the Corpus of Early English Correspondence (CEEC). Poster presented at ICAME 32, Oslo, 1-5 June 2011.

Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In Davies, M., Rayson, P., Hunston, S. and Danielsson, P. (eds.) Proceedings of the Corpus Linguistics Conference: CL2007, University of Birmingham, UK, 27-30 July 2007.

Acknowledgements

- Thank you to Paul Schaffner (University of Michigan) for sending the latest version of the EEBO-TCP texts for us to play with.

- Thank you to Patricia Murrieta-Flores (Spatial Humanities Project, History, Lancaster) for producing the publication places map.

- Thank you to Yehia El-Khatib (SCC, Lancaster) for providing one of the many servers used to complete the VARDing of the EEBO texts.