The extent of spelling variation in Early Modern English

Alistair Baron (Computing, Lancaster)
Paul Rayson (Computing, Lancaster)
Dawn Archer (Journalism, Media & Communication, UCLAN)
Motivation: “Word Frequency”

Why is it so important?
- doesn’t really need explaining to the ICAME audience!! 😊
- Sinclair (1991: 30) noted that "anyone studying a text is likely to need to know how often each different word form occurs in it"
- Geoff Leech @ TALC8 “Frequency is important – and challenging” (July 2008)

But we have to be careful about how we calculate it
Research questions

- Already shown that spelling variation in Early Modern English affects accuracy of:
  - key word analysis (Baron et al, 2009)
  - POS tagging (Rayson et al, 2007)
  - semantic tagging (Archer et al, 2003)

(general) What problems occur when counting words in historical corpora?

(specific) How much difference does spelling variation make to frequency results?
Our study

The extent of spelling variation

- How much spelling variation occurs in Early Modern English?
- How does the date of the text relate to the amount of spelling variation?
- How does the amount of spelling variation contrast from corpus to corpus?
Corpora

- ARCHER: A Representative Corpus of Historical English Registers.


- The Lampeter corpus of Early Modern English Tracts (Schmied, 1994).

- The Early Modern English Medical Texts (EMEMT) corpus (Taavitsainen et al., forthcoming; Taavitsainen and Pahta, 1997 and forthcoming).

- Shakespeare’s First Folio, sourced from the Oxford Text Archive. [http://ota.ahds.ac.uk/](http://ota.ahds.ac.uk/)
## Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Genre and Type</th>
<th>Years Eligible</th>
<th>Texts Eligible</th>
<th>Tokens Eligible</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCHER</td>
<td>General / Mixed</td>
<td>1660-1799</td>
<td>364</td>
<td>632,639</td>
</tr>
<tr>
<td>EEBO</td>
<td>General / Mixed</td>
<td>1470-1709</td>
<td>12,265</td>
<td>535,910,150</td>
</tr>
<tr>
<td>Innsbruck</td>
<td>Letters</td>
<td>1410-1689</td>
<td>436</td>
<td>170,538</td>
</tr>
<tr>
<td>Lampeter</td>
<td>Religion, Politics, Economy &amp; Trade, Science, Law, and Miscellaneous tracts and pamphlets</td>
<td>1640-1739</td>
<td>120</td>
<td>1,124,131</td>
</tr>
<tr>
<td>EMEMT</td>
<td>Medical texts</td>
<td>1540-1699</td>
<td>51</td>
<td>491,384</td>
</tr>
<tr>
<td>Shakespeare</td>
<td>All plays (Comedies, Histories and Tragedies) from the First Folio.</td>
<td>c1590-c1613</td>
<td>36</td>
<td>821,123</td>
</tr>
</tbody>
</table>
Extent of Spelling Variation

The aim for the analysis was to discover, quantitatively, the extent of spelling variation in the Early Modern English (EModE) period.

Previous research has commented on the levels of spelling variation without quantifying it (see, e.g., Vallins and Scragg (1965); Görlach (1991)).

Schneider (2002), in his attempts to develop a normalised version of the Zurich English Newspaper (ZEN) Corpus (1670-1799), produced an overview of the spelling variations contained within.

- 3.99% of the tokens and 38.02% of the types within the corpus were unrecognised by the ENGCG tagger, and hence could be considered spelling variants.
- The percentage of unrecognized tokens and types reduced in each subsequent time period, from 4.66% tokens and 36.57% types in the 1670-1709 sub-corpus to 2.85% tokens and 26.06% types in the 1770-1799 sub-corpus.

A more thorough quantitative study of the spelling variation within the entire Early Modern English period is required.
Sampling

The corpora were sampled at 10 year periods.

Samples were chosen from randomly selected texts from each decade, with the sample from each text beginning at a randomly selected index.

All results were normalised to a percentage in order to compare corpora with different sample sizes.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Decade Sample Size</th>
<th>Minimum Texts</th>
<th>Decades not included due to a lack of texts and/or words</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCHER</td>
<td>4,000</td>
<td>10</td>
<td>1740</td>
</tr>
<tr>
<td>EEBO</td>
<td>80,000</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Innsbruck</td>
<td>1,200</td>
<td>4</td>
<td>1420, 1430, 1490, 1590</td>
</tr>
<tr>
<td>Lampeter</td>
<td>40,000</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>EMEMT</td>
<td>Total Possible</td>
<td>2</td>
<td>1620, 1640</td>
</tr>
<tr>
<td>Shakespeare</td>
<td>60,000</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>
Study details

- Each word in a given historical sample was compared to a modern word list. This was derived from:
  - Words with a frequency greater than 5 in the British National Corpus (BNC) (Leech et al., 2001).

- If a word was not found in the modern word lists it was classed as a spelling variant.

- This analysis provided a percentage of variant types and tokens per corpus and per decade sample.
Results - Types

Decade

Types

ARCHER
EEBO
EMEMT
Innsbruck
Lampeter
Shakespeare
A definite downwards trend in the amount of spelling variation occurring throughout the EModE period.

- Corroborates Schneider's (2002) quantitative analysis for the latter part of the EModE period (1670-1799).
- Quantifies the trend over the entire EModE period, verifying many scholar's claims (see, e.g., Görlach, 1991:8-9; Lass, 1999b: 56, Rissanen, 1999: 187).

The rate of reduction in variation slows from around 1700. This backs up Görlach's (1991: 11) claim that, by 1700, the language had achieved “considerable homogeneity.”

Variant percentages are approximate values:

- “Real-word errors” will not be detected (i.e. those historical variants which match other modern words e.g bee/be, doe/do, then/than
- Some valid words will be marked as variants incorrectly.
An analysis of two small manually standardised samples used in a previous study (see Rayson et al., 2007) indicates the likely error rates.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total words</th>
<th>% of words which required normalisation</th>
<th>% of variants which are real-word errors</th>
<th>% of words erroneously marked as variants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Types</td>
<td>Tokens</td>
<td>Types</td>
<td>Tokens</td>
</tr>
<tr>
<td>Lampeter</td>
<td>839</td>
<td>2,726</td>
<td>19.19%</td>
<td>9.61%</td>
</tr>
<tr>
<td></td>
<td>Types</td>
<td>Tokens</td>
<td>Types</td>
<td>Tokens</td>
</tr>
<tr>
<td></td>
<td>4.35%</td>
<td>2.67%</td>
<td>12.04%</td>
<td>4.37%</td>
</tr>
<tr>
<td>Shakespeare</td>
<td>897</td>
<td>3,991</td>
<td>63.88%</td>
<td>24.03%</td>
</tr>
<tr>
<td></td>
<td>Types</td>
<td>Tokens</td>
<td>Types</td>
<td>Tokens</td>
</tr>
<tr>
<td></td>
<td>8.55%</td>
<td>5.11%</td>
<td>7.80%</td>
<td>3.38%</td>
</tr>
</tbody>
</table>

“Real-word error” rates are lower than generally found in modern spelling errors.

- Peterson (1986) found that between 2% and 16% of typing errors would be undetected depending on the size of the word list used.
- Mitton (1987) found much larger rates; 40% of the spelling errors found in his study were real-word errors.
- In our own study on a manually processed corpus of child language spelling errors we found 24.07% of variant types and 18.31% of variant tokens were real-word errors.
Approximation – Innsbruck
Approximation – Innsbruck

% Types

- 'Real Word Errors Split Words'
- '- Split Words'
+ Not In Dictionary

Decade

0 20 40 60 80 100


% Types

0 20 40 60 80 100


Decade
VARD 2

- An interactive piece of software designed to assist users in dealing with spelling variation.
- Uses techniques from modern spell checkers as well as a manually derived list of known variant replacements.
- Can be used to manually and automatically standardize spelling variation.
- Free to use for academic research. Available from http://www.comp.lancs.ac.uk/~barona/vard2
VARD 2

FILE | EDIT | STYLE | ADVANCED
--- | --- | --- | ---

Enter passage:

1. Lord, He saues my labor by his owne approach.
Du. Sen. Why how now Monsieur, what a life is this
That your poor friends must woe your companie.
What, you looke merrily.

Iag. A Foolie, a foolie: I met a foolie th' Forrest,
A motley Foolie (a miserable world)
As I do live by foolie, I met a foolie.
Who laid him downe, and bask'd him in the Sun,
And rall'd on Lady Fortune in good terms,
In good set terms, and yet a motley foolie.

Good mornow foolie (quoth I) no Sir, quoth he.
Call me not foolie, till Heaven hath sent me fortune.
And then he drew a dial from his poake,
And looking on it, with jadge-lustre eye.
Says, very wisely, it is ten a clocke.
Thus we may se quoth he how the world wagge
This but an hour agoe, since it was nine.
And after one hour more, 'twill be eleuen.
And so from hour to hour, we ripe, and ripe,
And then from hour to hour, we rot, and rot,
And thereby hangs a tale. When I did heare
The motley Foolie, thus morall on the time,
My Lungs began to crow like Chanticleere.

That Foolies should be so deepe contemptuature.
And I did laugh, sans intermission.
An hour by his diall. Oh noble foolie.
A worthy foolie Motley's the onely weare.
Du. Sen. What foolie is this?

Iag. O worthie foolie: One that hath bin a Courtier
And sayes, if Ladies be but yong, and faire,
They haue the gift to know it: and in his braine,
Which is as drie as the remainder bisket
After a voyage. He hath strange places cram'd
With shrewdness, the which he soe.
Previous papers describe VARD 2 in more detail:
- Rayson et al. (2008)
- Rayson et al. (2007)

Upcoming papers will evaluate VARD 2’s ability in dealing with spelling variation, particularly the effect of training the tool:
- Baron, A. and Rayson, P. (forthcoming)
- Baron, A., Rayson, P. and Archer, D. (forthcoming)
Summary

- We have quantified the extent of spelling variation in Early Modern English.

- The trends identified match the expected rapid decline in spelling variation until around 1700.

- Further details are available in a journal paper (Baron et al, 2009).

- In that paper we also show that spelling variation does have an effect on key word analysis and researchers should be aware of the reduced accuracy when studying historical corpora.
Current and Future Work

- The development of VARD 2 is ongoing. VARD 2.2 was released in December 2008. [http://www.comp.lancs.ac.uk/~barona/vard2/](http://www.comp.lancs.ac.uk/~barona/vard2/)

- Further investigation is needed into the extent of “real-word errors”
  - Contextual information will help to detect such variants.

- Variation across genres

- Training VARD 2 to deal with different Early Modern English corpora and other language varieties containing spelling variation.
  - How much training data is needed? (Corpus Linguistics 2009)
  - Letter replacement rules from DICER.
Thanks for listening

Any questions?

Acknowledgements

Our thanks to Irma Taavitsainen and Manfred Markus for providing access to the EMEMT and Innsbruck corpora
References


References


References


