

LANCASTER
UNIVERSITY

uclan



InfoLab21

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

- Dee Gardner “Validating the construct of *word* in applied corpus-based vocabulary research: a critical survey”, *Applied Linguistics* 28:2, 2008.
- Charles Alderson “Judging the frequency of English Words”, *Applied Linguistics* 28:3, 2007

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.
- EEBO: Early English Books Online. <http://eebo.chadwyck.com/>
- The Innsbruck Letter corpus, part of the Innsbruck Computer-Archive of Machine-Readable English Texts (ICAMET) corpus (Markus, 1999).
- 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/>

Corpora

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

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.

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

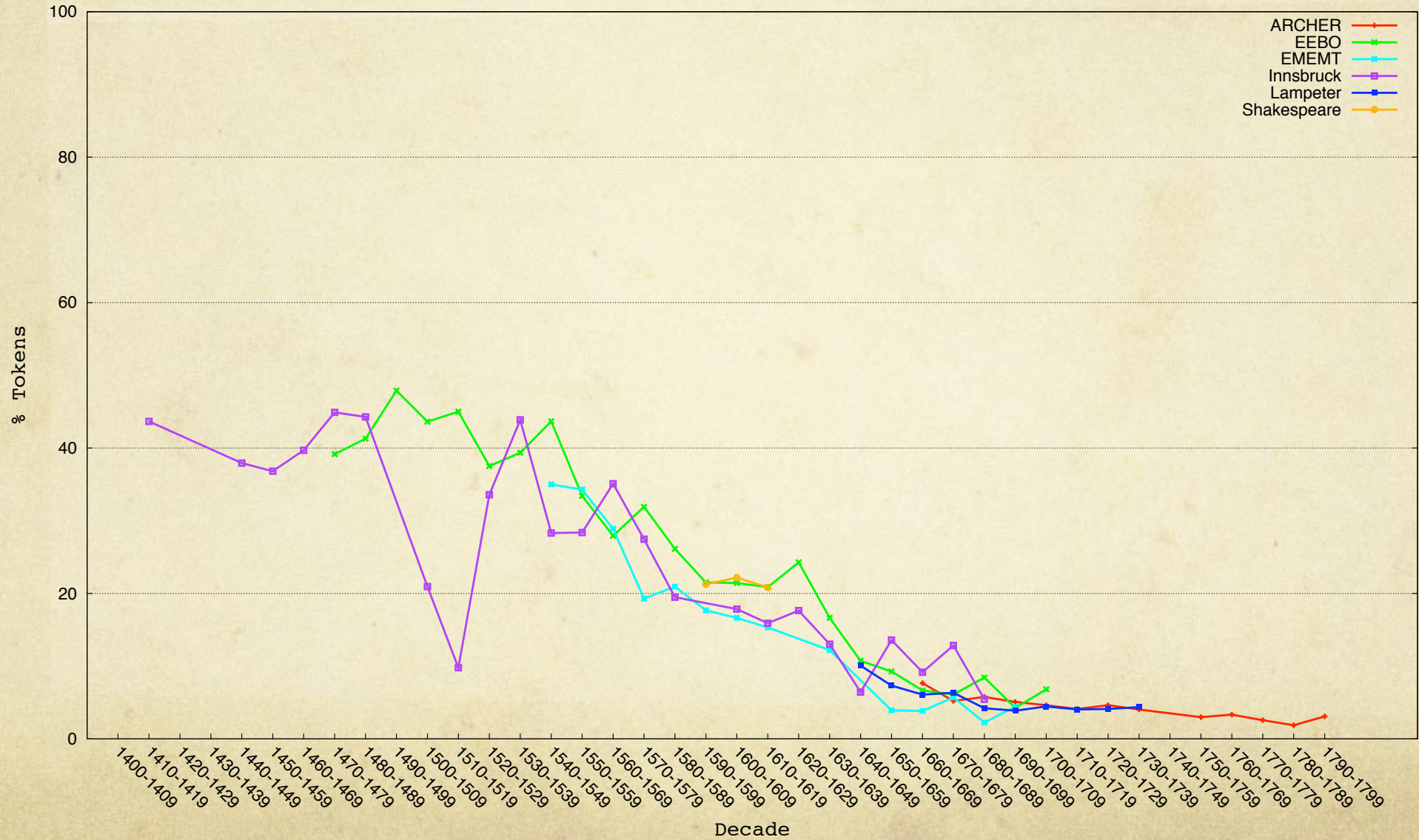
Study details

- Each word in a given historical sample was compared to a modern word list. This was derived from:
 - The Spell Checking Oriented Word Lists (SCOWL).
<http://wordlist.sourceforge.net/scowl-readme>
 - 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



Results - Tokens



Results - Notes

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.

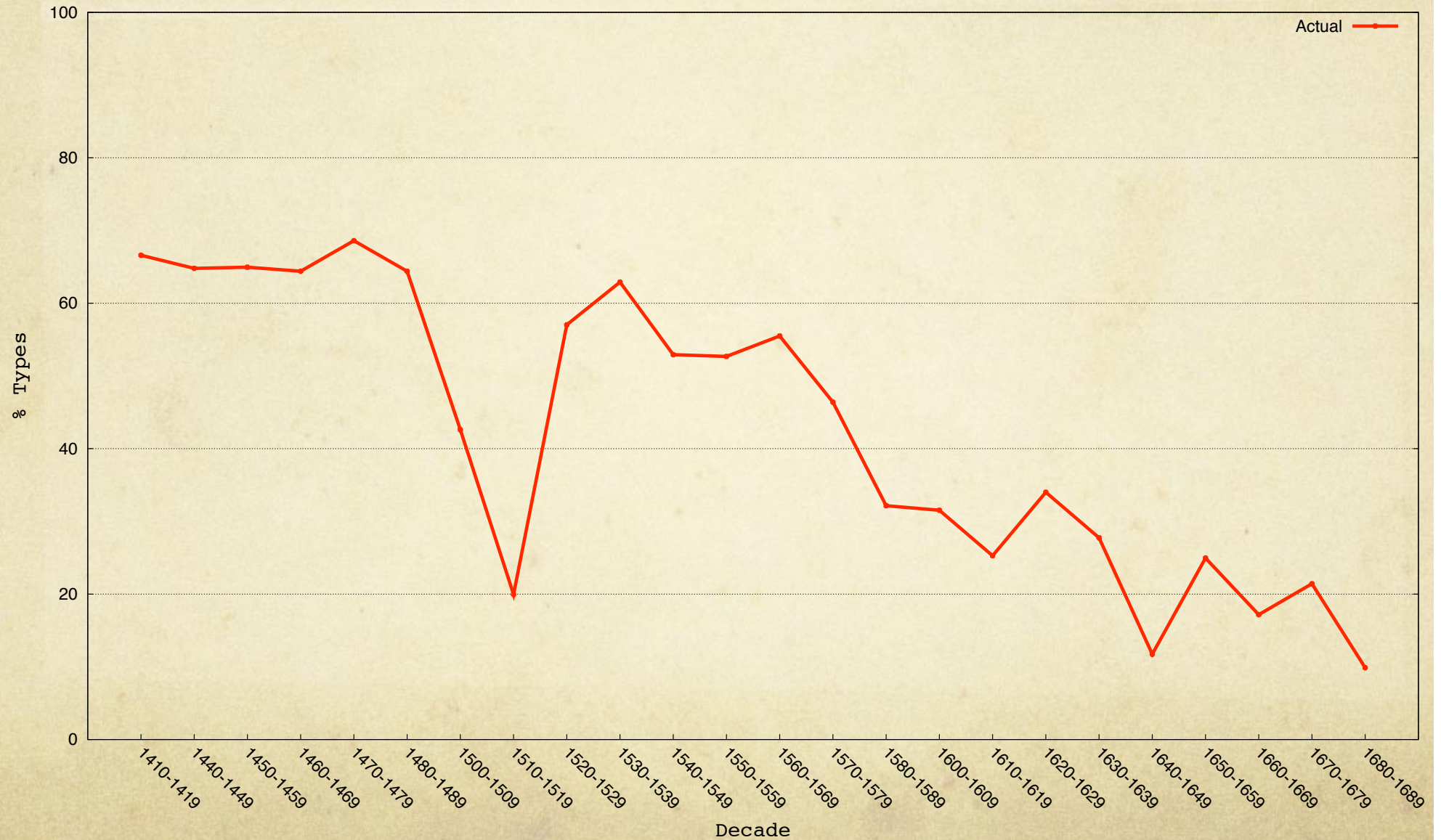
Approximate Percentages

- An analysis of two small manually standardised samples used in a previous study (see Rayson et al., 2007) indicates the likely error rates.

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

- “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



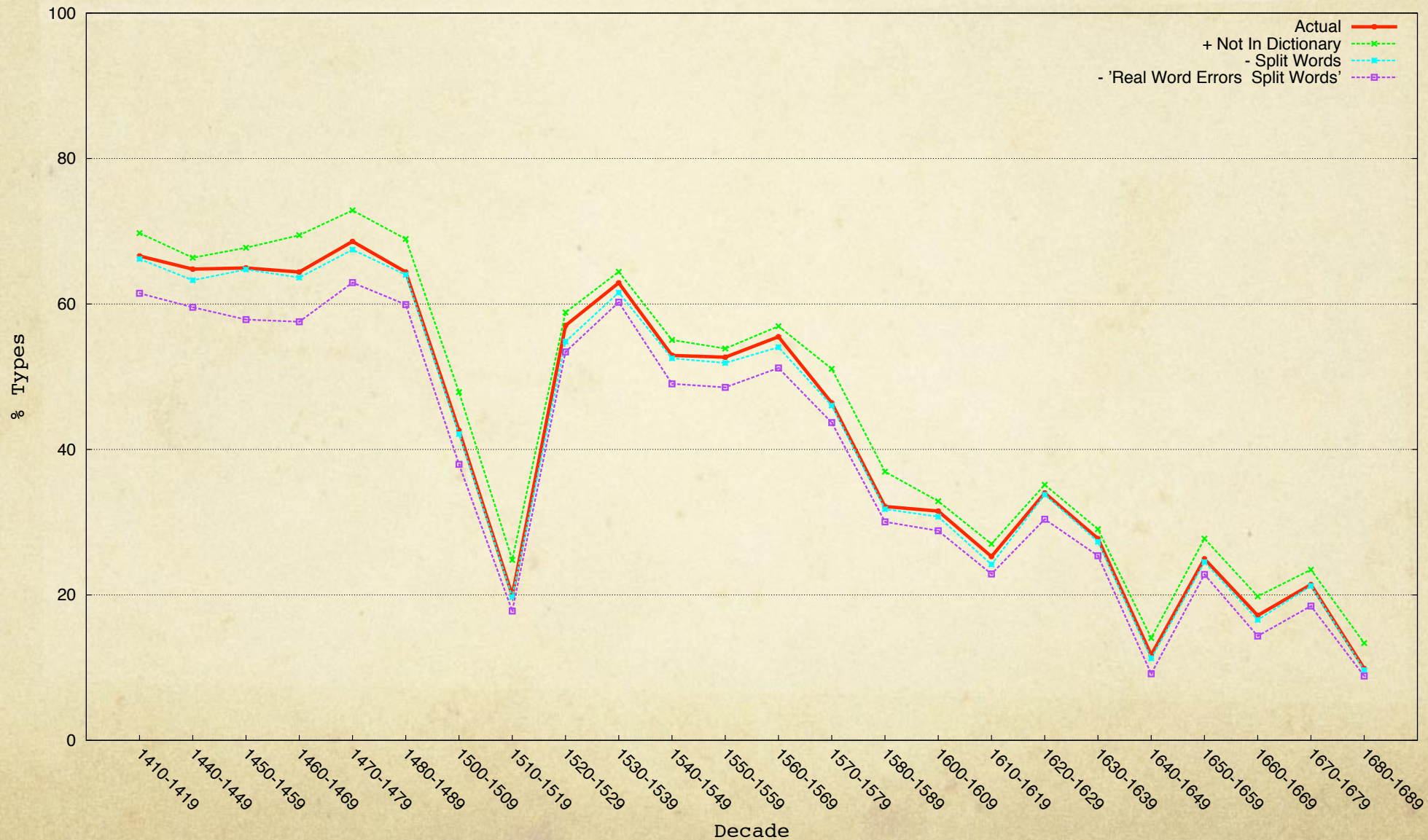
Approximation – Innsbruck



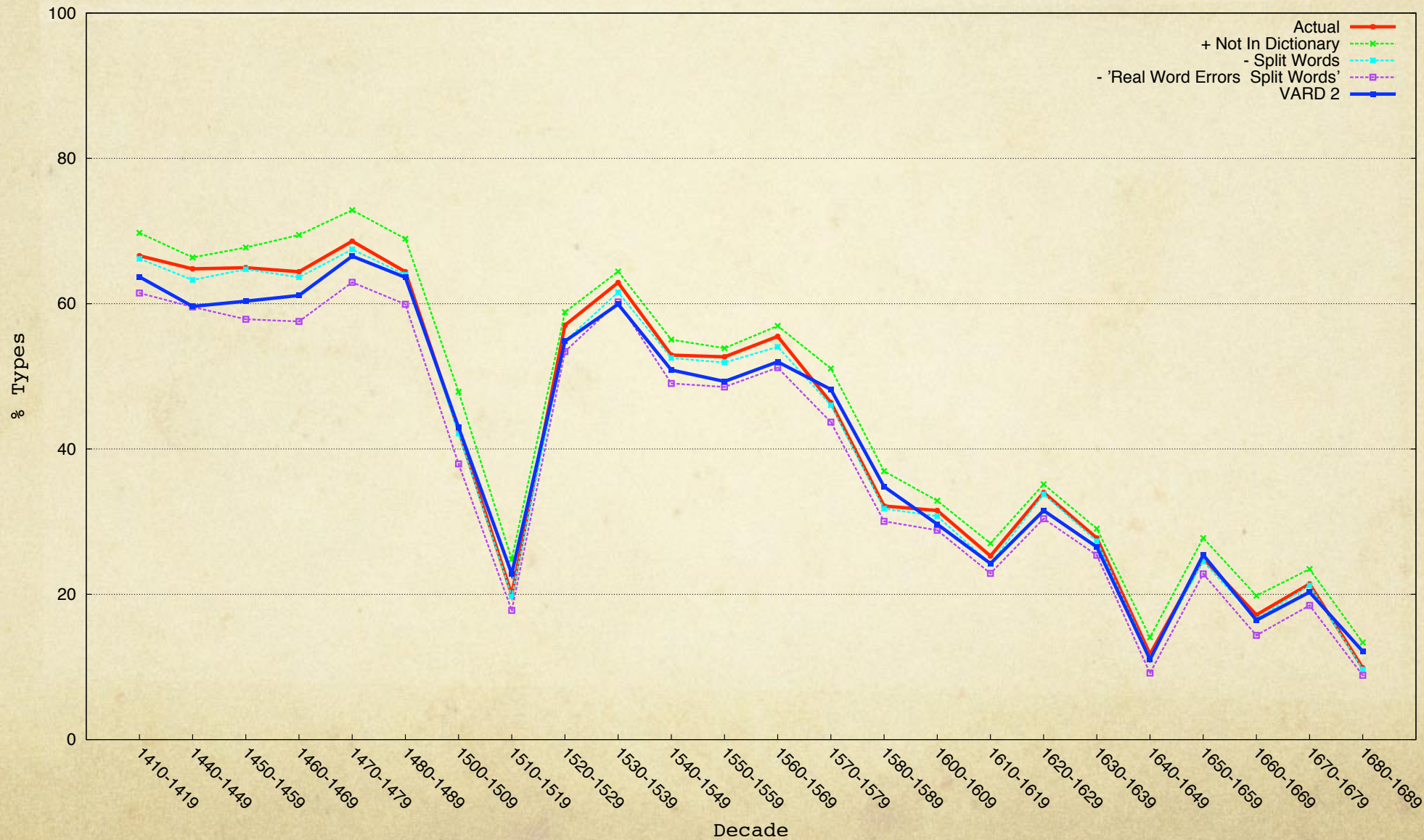
Approximation – Innsbruck



Approximation – Innsbruck



Approximation – Innsbruck



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

The screenshot shows the VARD 2.2 software interface. The title bar reads "VARD 2.2 - As You Like It.txt". The menu bar includes "File", "Edit", "Style", and "Advanced". The toolbar contains icons for undo, redo, and font settings. The font is set to "Lucida Grande" and the size is "13". The text area contains the following text with highlighted words:

Enter **laques**. [980]
1. Lord. He **saues** my **labor** by his **owne** approach.
Du.Sen. Why how now Monsieur, what a **life** is this
That your **poore** friends must woe your **companie**,
What, you **looke** merrily.
laq. A **Foole**, a **foole**: I met a **foole** i'th **Forrest**,
A motley **Foole** (a miserable world:)
As I do **liue** by **foode**, I met a **foole**,
Who laid him **downe**, and **bask'd** him in the Sun,
And **rail'd** on Lady Fortune in good **termes**,
In good set **termes**, and yet a motley **foole**. [990]
Good morrow **foole** (quoth I: no Sir, quoth he,
Call me not **foole**, till **heauen** hath sent me fortune.
And then he drew a **diall** from his **poake**,
And looking on it, with **lacke-lustre** eye,
Sayes, very wisely, it is ten a **clocke**:
Thus we may see (quoth he) how the world **wagge**
'Tis but an **heure** agoe, since it was nine,
And after one **heure** more, **'twill** be **eleuen**,
And so from **heure** to **heure**, we ripe, and ripe,
And then from **heure** to **heure**, we rot, and rot, [1000]
And thereby hangs a tale. When I did **heare**
The motley **Foole**, thus **morall** on the time,
My Lungs began to crow like **Chanticleere**,
That **Fooles** should be so **deepe** **contemplatiue**:
And I did laugh, sans intermission
An **heure** by his **diall**. Oh noble **foole**,
A worthy **foole**: **Motley's** the **onely** **weare**.
Du.Sen. What **foole** is this?
laq. O **worthie** **Foole**: One that hath bin a Courtier
And **sayes**, if Ladies be but **yong**, and **faire**, [1010]
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 **obscurities**, the which he scents

A context menu is open over the word "company". The menu items are:

- company (96%)
- companies (20.5%)
- companion (18.5%)
- campanile (18.5%)
- company's (16.5%)
- More Suggestions...
- Suggestions not in dictionary...
- Replace with...
- Mark instance as Modern Form
- Find word in list
- Mark all as Modern Form

Sub-menus for "company" and "More Suggestions..." are also visible:

- Replace instance
- Replace all
- Known Variant List (55%)
- Letter Replacement Rules (22.5%)
- Phonetic Matching (22.5%)
- Edit Distance is 2 (-4%)
- Frequency is 401

The right sidebar shows the following options:

- Variant Forms (1809)
- Replaced (0)
- Modern Forms (1962)
- Copy Current List
- Variant Forms (1809):
- 'em (1)
- 'faith (2)
- 'gainst (2)
- 'od's (1)
- 'ods (1)
- 'tis (30)
- 'twas (3)
- 'twill (4)
- ' (5)
- Replacement Threshold: 50
- Process All Variants
- Show Replacment Analysis

VARD 2

- Previous papers describe VARD 2 in more detail:
 - Baron, A. and Rayson, P. (2008)
 - 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/ward2/>
- 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

- Archer, Dawn, Tony McEnery, Paul Rayson and Andrew Hardie. "Developing an automated semantic analysis system for Early Modern English." *Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16*. Eds. Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery. (2003): 22-31.
- Baron, A., Rayson, P. and Archer, D. (2009). Word frequency and key word statistics in historical corpus linguistics. In Ahrens, R. and Antor, H. (eds.) *Anglistik: International Journal of English Studies*, 20 (1), pp. 41-67.
- Baron, A. and Rayson, P. (forthcoming). Automatic standardization of texts containing spelling variation, how much training data do you need? To appear at *Corpus Linguistics 2009, University of Liverpool, UK, 20-23 July 2009*.
- Baron, A., Rayson, P. and Archer, D. (forthcoming). Automatic Standardization of Spelling for Historical Text Mining. To appear at *Digital Humanities 2009, University of Maryland, USA, 22-25 June 2009*.
- Görlach, Manfred. *Introduction to Early Modern English*. Cambridge: Cambridge University Press, 1991.

References

Lass, Roger. "Phonology and Morphology." *The Cambridge History of the English Language: Volume III, 1476-1776*. Ed. Roger Lass. Cambridge: Cambridge University Press, 1999b.

Leech, Geoffrey, Paul Rayson and Andrew Wilson. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Longman, 2001.

Markus, Manfred. "Manual of ICAMET (Innsbruck Computer-Archive of Machine-Readable English Texts)." *Innsbrucker Beitræge zur Kulturwissenschaft, Anglistische Reihe 7*. Innsbruck: Leopold-Franzens-Universitaet Innsbruck, Institut fuer Anglistik, 1999.

Mitton, Roger. "Spelling Checkers, Spelling Correctors and the Misspelling of Poor Spellers." *Information Processing & Management* 23.5 (1987): 495-505.

Peterson, James L. "A Note on Undetected Typing Errors." *Communications of the ACM* 29.7 (1986): 633-637.

References

- Rayson, Paul, Dawn Archer, Alistair Baron, Jonathan Culpeper and Nicholas Smith. "Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora." *Proceedings of Corpus Linguistics 2007*. University of Birmingham, UK (27-30 July 2007).
- Rayson, P., Archer, D., Baron, A. and Smith, N. (2008). Travelling Through Time with Corpus Annotation Software. In Lewandowska-Tomaszczyk, B. (ed) *Corpus Linguistics, Computer Tools, and Applications – State of the Art. Palc 2007*. Peter Lang, Frankfurt am Main.
- Rissanen, Matti. "Syntax." *The Cambridge History of the English Language: Volume III, 1476-1776*. Ed. Roger Lass. Cambridge: Cambridge University Press, 1999.
- Schmied, Josef. "The Lampeter Corpus of Early Modern English Tracts." *Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora*. Cambridge, March 1993. Eds. Merja Kytö, Matti Rissanen, Susan Wright. Amsterdam: Rodopi, 1994.

References

Schneider, Peter. "Computer Assisted Spelling Normalization of 18th Century English." *New Frontiers of Corpus Research: Papers from the 21st International Conference on English language Research on Computerized Corpora*, Sydney, 2000. Eds. Pam Peters, Peter Collins and Adam Smith. Amsterdam: Rodopi, 2002. 199-211.

Sinclair, John. *Corpus, concordance, collocation*. Oxford: Oxford University Press, 1991.

Taavitsainen, Irma and Päivi Pahta. "Corpus of Early English medical writing 1375-1750." *ICAME Journal* 21 (1997): 71-81.

Taavitsainen, Irma and Päivi Pahta, eds. *Medical Writing in Early Modern English*. Cambridge: Cambridge University Press, forthcoming.

Taavitsainen, Irma, Päivi Pahta, Turo Hiltunen, Martti Mäkinen, Ville Marttila, Maura Ratia, Carla Suhr and Jukka Tyrkkö. *Early Modern Medical Texts*. Forthcoming.

Vallins, George H. and Donald G. Scragg. *Spelling*. London: André Deutsch, 1965.