Quantifying Early Modern English spelling variation: Change over time and genre

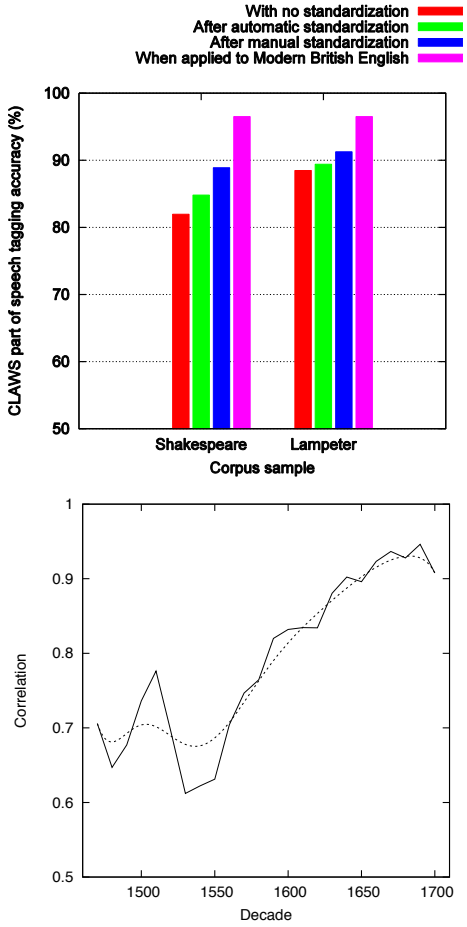
Alistair Baron and Paul Rayson
Lancaster University

Dawn Archer
University of Central Lancashire

New Methods in Historical Corpora Conference
University of Manchester, 29th - 30th April 2011

EModE spelling variation

- Marked degree of spelling variation in Early Modern English texts despite the gradual standardisation between 1500-1700 (Vallins & Scragg, 1965; Görlach, 1991; Nevalainen, 2006).
- Spelling variation has a negative effect on the accuracy of automatic corpus linguistic methods. This has been shown to be the case for:
  - Semantic analysis (Archer *et al.*, 2003)
  - POS tagging (Rayson *et al.*, 2007)
  - Key word analysis (Baron *et al.*, 2009)

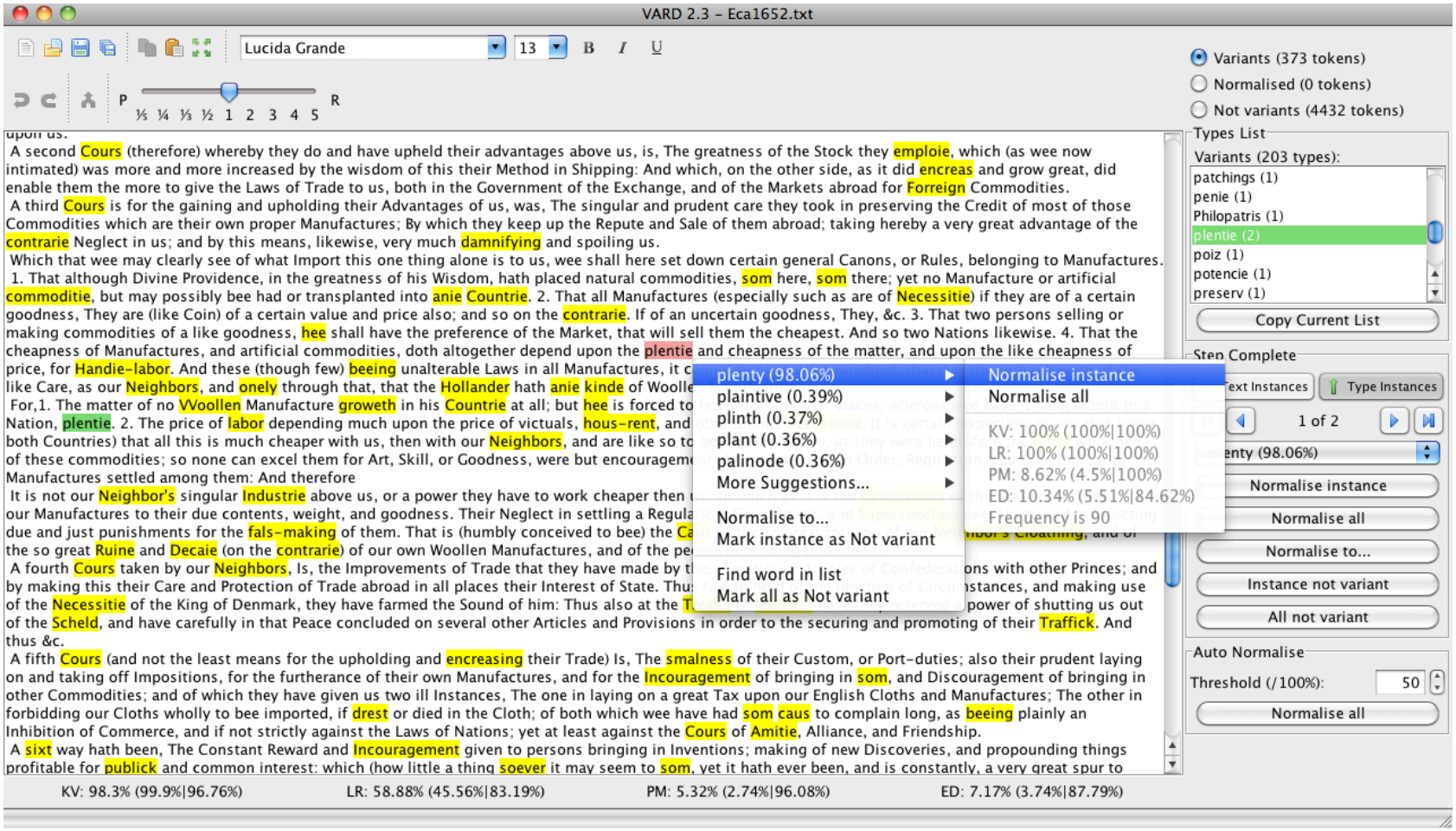


VARD 2

- A tool for normalising spelling variation in historical corpora both manually and automatically.
- Variants are detected by finding those that do not occur in a modern word list.
- A ranked list of normalisation candidates for each variant is produced using four main methods:
  - A manually created list of variant/normalisation pairs.
  - Phonetic matching using a modified Soundex algorithm.
  - A set of letter replacement rules.
  - The Levenshtein Edit Distance algorithm.
- Normalisations are chosen by the user or automatically by the system and replaced in the text with the original spelling retained in an xml tag.

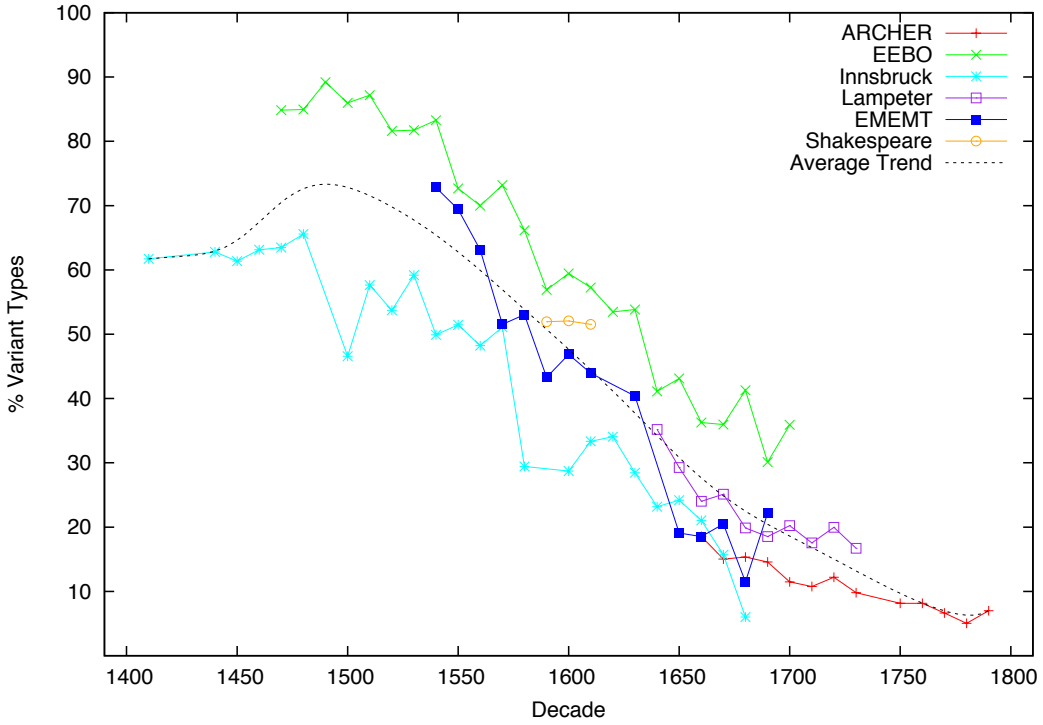
(Baron & Rayson, 2009)

VARD 2.3


VARD 2.3 - Eca1652.txt
Lucida Grande
13
Variants (373 tokens)
Normalised (0 tokens)
Not variants (4432 tokens)
Types List
Variants (203 types):
patchings (1)
penie (1)
Philopatris (1)
plentie (2)
poiz (1)
potencie (1)
preserv (1)
Copy Current List
Step Complete
Text Instances
Type Instances
1 of 2
plenty (98.06%)
Normalise instance
Normalise all
Normalise to...
Instance not variant
All not variant
Auto Normalise
Threshold (/100%):
50
Normalise all
plenty (98.06%)
plaintive (0.39%)
plinth (0.37%)
plant (0.36%)
palinode (0.36%)
More Suggestions...
Normalise to...
Mark instance as Not variant
Find word in list
Mark all as Not variant
Normalise instance
Normalise all
KV: 100% (100%|100%)
LR: 100% (100%|100%)
PM: 8.62% (4.5%|100%)
ED: 10.34% (5.51%|84.62%)
Frequency is 90
KV: 98.3% (99.9%|96.76%)
LR: 58.88% (45.56%|83.19%)
PM: 5.32% (2.74%|96.08%)
ED: 7.17% (3.74%|87.79%)

Quantifying spelling variation

- VARD allows for the study of spelling variation in EModE texts, and its effects.
- A large-scale study of the spelling variation in different EModE corpora quantified the steady decline in the ratio of spelling variants to modern spellings. (Baron *et al.*, 2009)



DICER

- Discovery and Investigation of Character Edit Rules
- Examines variant / normalisation pairs found in the XML output from VARD.
- Determines what letter replacement rules are required to convert the variant form into the normalised form. For example:

Variant	Normalisation	Rules
anie	any	ie → y
publick	public	remove k
ioynte	joint	i → j y → i remove e

- Frequencies are calculated for each rule indicating how often each rule occurs, which position of the variant it should be applied and with which surrounding letters.
- Meta-data is also stored to allow for the analysis of spelling rule trends over time, genre or any other meta-data present.

DICER

DICER: Discovery and Investigation of Character Edit Rules

Innsbruck Letters

Change... Go!

User

Logged in as *a.baron@comp.lancs.ac.uk.*

Logout

Summary | Rules | Search | Profile

Display: Tokens | **Types**

Link to this page

Edit Distance:

Edits	Frequency
1	5988 (44.35%)
2	4490 (33.25%)
3	2034 (15.06%)
4	676 (5.01%)
5	224 (1.66%)
6	64 (0.47%)
7	19 (0.14%)
8	6 (0.04%)
9	1 (0.01%)
14	1 (0.01%)
Total	**13503**

Positions:

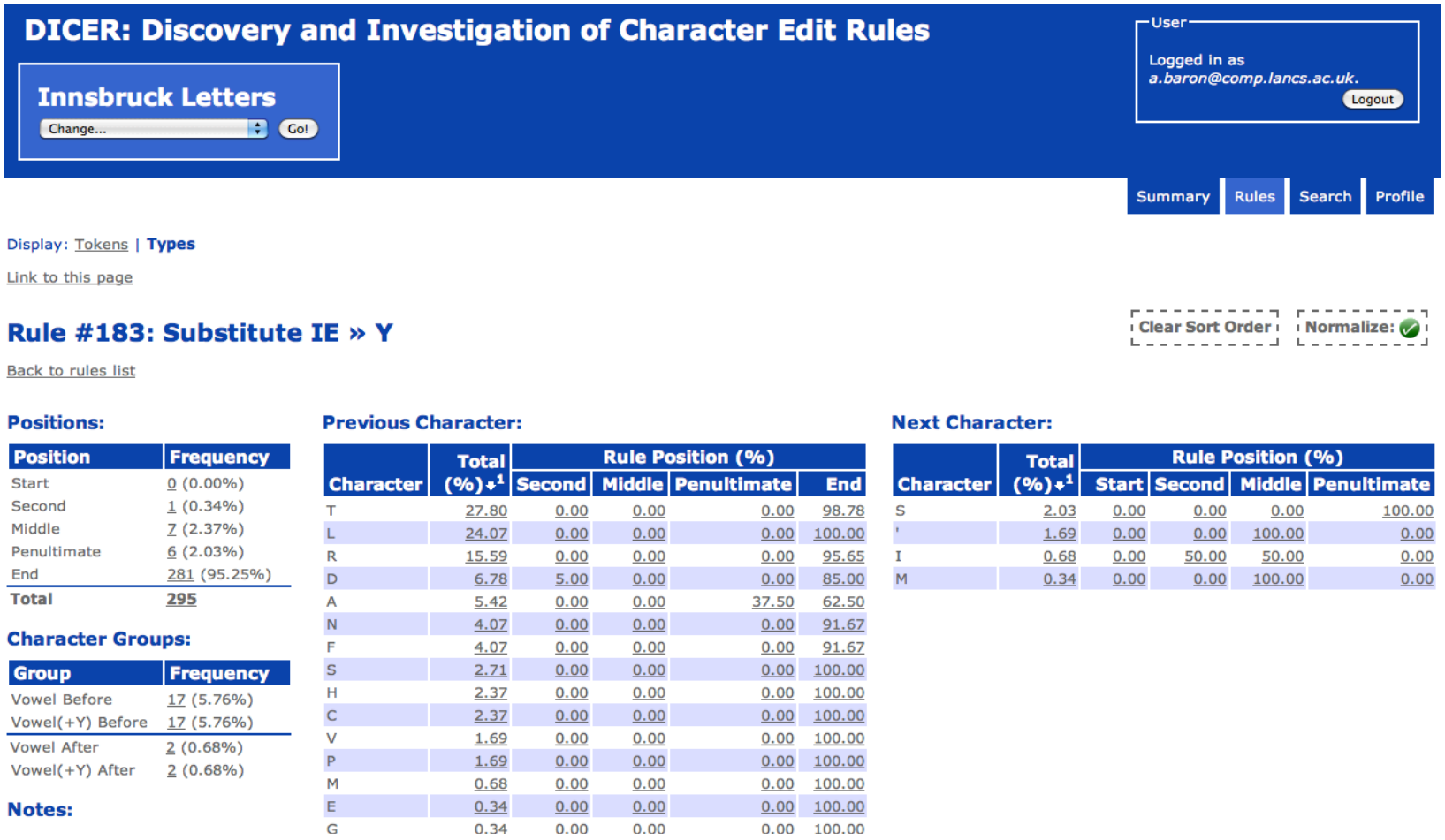
Position	Frequency
Start	1043 (5.10%)
Second	2238 (10.94%)
Middle	8088 (39.54%)
Penultimate	3046 (14.89%)
End	6042 (29.54%)
Total	**20457**

Rules:

Clear Sort Order | Normalize

#	ID	Rule	Variant	Standard	Total (/1000)[1]	Position (%)				
						Start	Second	Middle	Penultimate	End
1	2	Deletion	E		144.40	0.10	0.34	11.78	13.47	74.31
2	3	Substitution	Y	I	89.90	1.36	23.22	69.98	5.44	0.00
3	7	Insertion		E	43.46	1.01	5.06	28.57	15.52	49.83
4	4	Substitution	LL	L	27.33	0.18	7.16	19.14	7.51	66.01
5	130	Substitution	Y	E	22.93	1.28	8.10	22.60	66.31	1.71
6	33	Insertion		A	19.11	6.39	6.14	78.01	8.70	0.77
7	45	Substitution	E	I	17.99	11.68	24.46	57.34	6.52	0.00
8	43	Substitution	U	V	15.54	0.63	7.55	78.62	13.21	0.00
9	183	Substitution	IE	Y	14.42	0.00	0.34	2.37	2.03	95.25
10	32	Deletion	U		14.37	0.00	6.46	84.69	8.50	0.34
11	16	Substitution	I	E	13.59	29.86	13.67	21.94	33.45	1.08
12	207	Substitution	TT	T	12.95	0.00	0.38	30.57	16.60	52.45
13	14	Deletion	Y		11.78	2.90	0.41	36.93	58.09	1.66
14	53	Insertion		U	10.31	0.00	4.74	81.52	13.27	0.47
15	66	Substitution	E	A	10.22	6.70	32.06	45.45	14.35	1.44
16	19	Substitution	E	EE	10.17	0.00	27.88	45.67	20.19	6.25
17	55	Insertion		I	10.12	1.93	8.21	84.54	4.35	0.97
18	229	Substitution	S	SS	9.87	3.96	3.96	18.32	8.42	65.35
19	105	Deletion	_		8.95	0.00	8.20	91.80	0.00	0.00

DICER



DICER: Discovery and Investigation of Character Edit Rules

Innsbruck Letters

Change... Go!

User

Logged in as *a.baron@comp.lancs.ac.uk.*

Logout

Summary | Rules | Search | Profile

Display: Tokens | **Types**

Link to this page

Rule #183: Substitute IE » Y

Clear Sort Order | Normalize:

Back to rules list

Positions:

Position	Frequency
Start	0 (0.00%)
Second	1 (0.34%)
Middle	7 (2.37%)
Penultimate	6 (2.03%)
End	281 (95.25%)
Total	**295**

Character Groups:

Group	Frequency
Vowel Before	17 (5.76%)
Vowel(+Y) Before	17 (5.76%)
Vowel After	2 (0.68%)
Vowel(+Y) After	2 (0.68%)

Notes:

Previous Character:

Character	Total (%)[1]	Rule Position (%)			
		Second	Middle	Penultimate	End
T	27.80	0.00	0.00	0.00	98.78
L	24.07	0.00	0.00	0.00	100.00
R	15.59	0.00	0.00	0.00	95.65
D	6.78	5.00	0.00	0.00	85.00
A	5.42	0.00	0.00	37.50	62.50
N	4.07	0.00	0.00	0.00	91.67
F	4.07	0.00	0.00	0.00	91.67
S	2.71	0.00	0.00	0.00	100.00
H	2.37	0.00	0.00	0.00	100.00
C	2.37	0.00	0.00	0.00	100.00
V	1.69	0.00	0.00	0.00	100.00
P	1.69	0.00	0.00	0.00	100.00
M	0.68	0.00	0.00	0.00	100.00
E	0.34	0.00	0.00	0.00	100.00
G	0.34	0.00	0.00	0.00	100.00

Next Character:

Character	Total (%)[1]	Rule Position (%)			
		Start	Second	Middle	Penultimate
S	2.03	0.00	0.00	0.00	100.00
'	1.69	0.00	0.00	100.00	0.00
I	0.68	0.00	50.00	50.00	0.00
M	0.34	0.00	0.00	100.00	0.00

DICER

	Variant	Standard	ED	Position	Index	Other Rules	Category	Tokens[1]
☐	verie	very	2	End	3			51
☐	Majestie	Majesty	2	End	6			28
☐	Secretarie	Secretary	2	End	8			28
☐	maiestie	majesty	3	End	6	120(2 - Middle)		16
☐	hartie	hearty	3	End	4	7(1 - Second)		14
☐	companie	company	2	End	6			14
☐	anie	any	2	End	2			12
☐	sorie	sorry	2	End	3	59(2 - Penultimate)		11
☐	daylie	daily	3	End	4	3(2 - Middle)		10
☐	trustie	trusty	2	End	5			10
☐	happie	happy	2	End	4			8
☐	mercie	mercy	2	End	4			8
☐	humblie	humbly	2	End	5			7
☐	copie	copy	2	End	3			6
☐	contrarie	contrary	2	End	7			6
☐	almightie	almighty	2	End	7			6
☐	treatie	treaty	2	End	5			6
☐	glorie	glory	2	End	4			6
☐	shortlie	shortly	2	End	6			6
☐	libertie	liberty	2	End	6			6
☐	dutie	duty	2	End	3			5
☐	manie	many	2	End	3			5
☐	Nobilitie	nobility	2	End	7			5
☐	Ladie	Lady	2	End	3			5
☐	certaintie	certainty	2	End	8			4
☐	holie	holy	2	End	3			4
☐	honestie	honesty	2	End	6			4
☐	partie	party	2	End	4			4
☐	satisfie	satisfy	2	End	6			4
☐	adversarie	adversary	2	End	8			4
☐	pittie	pity	3	End	4	207(2 - Middle)		4
☐	armie	army	2	End	3			4

DICER

Search Results | Decade | Year | Century | ½ Century | ¼ Century

¼ Century	Matching types	Total types	%
1375-1399	0	345	0.000
1400-1424	10	776	1.289
1425-1449	10	1999	0.500
1450-1474	3	1693	0.177
1475-1499	0	2368	0.000
1500-1524	33	2047	1.612
1525-1549	64	2687	2.382
1550-1574	62	1354	4.579
1575-1599	51	1263	4.038
1600-1624	65	1790	3.631
1625-1649	64	1676	3.819
1650-1674	54	1208	4.470
1675-1699	8	446	1.794

Corpora – EMEMT

- Contains 2 millions words from texts dated between 1500 and 1700 from the specific domain of science and medicine (Taavitsainen & Pahta, 2010).
- Corpus released with spelling variation automatically normalised using VARD 2 (Lehto *et al.*, 2010).
- VARD 2 was trained by Anu Lehto manually normalising a representative sample of the corpus. This comprised of:
  - 24 text extracts of 1,000 words representing all six categories at each 50-year time period.
  - 24 samples of 500 words generated by randomly selecting small portions of texts from the remaining corpus.
- The manually normalised samples (36,000 words total) contain 5,406 variant tokens and 2,820 variant types for analysis in DICER.

Corpora – Innsbruck Letters

- Part of the Innsbruck Computer-Archive of Machine-Readable English Texts (ICAMET) (Markus, 1999).
- 469 complete letters dated between 1386 and 1688, containing a total of 182,000 words.
- Contains parallel line pairs, one of the original text and one with a normalised version of the first line:

  $I schepyng at thys day, but be the grace of God I am avysyd
  $N shipping at this day, but by the grace of God I am advised

- Converted into XML format so individual spelling variant-normalisation pairs can be analysed:

  <replaced orig="schepyng">shipping</replaced> at <replaced orig="thys">this </replaced> day, but <replaced orig="be">by</replaced> the grace of God I am <replaced orig="avysyd”>advised</replaced>

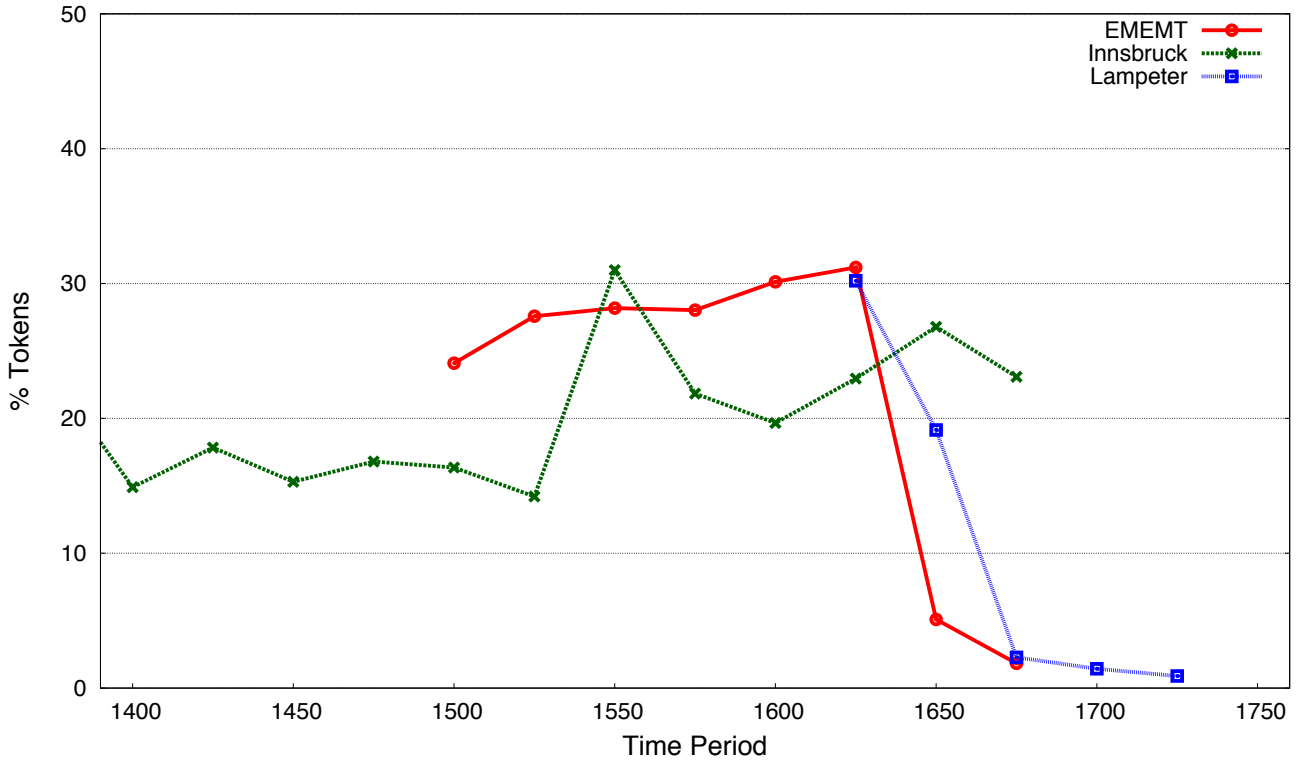
- 43,740 variant tokens and 13,503 variant types to be analysed with DICER.

Corpora – Lampeter

- Tracts and pamphlets published between 1640 and 1740 (Schmied, 1994).
- Six domains represented (Religion, Politics, Economy & Trade, Science, Law and Miscellaneous) with two texts for each domain per decade.
- Total of 120 complete texts by 120 different authors. 1.1 million words.
- Spelling variants automatically normalised with VARD 2.3 at a 50% threshold after being trained by manually normalising a 3,000 word sample (as used in Rayson *et al.*, 2007).
- 34,304 variant tokens and 7,339 variant types to analyse in DICER.

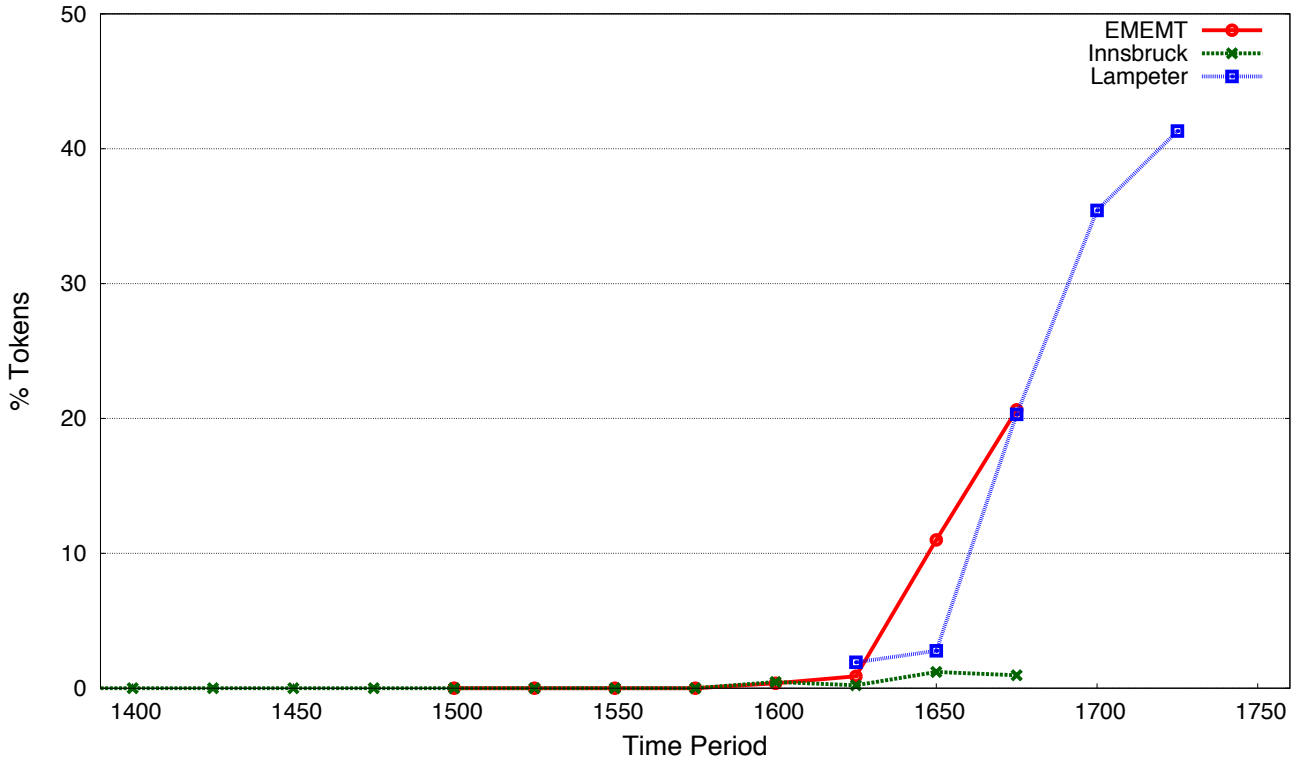
Extra final e removed

- Examples:
  - doe (do)
  - thinke (think)
  - owne (own)
- Most common rule in all three datasets.



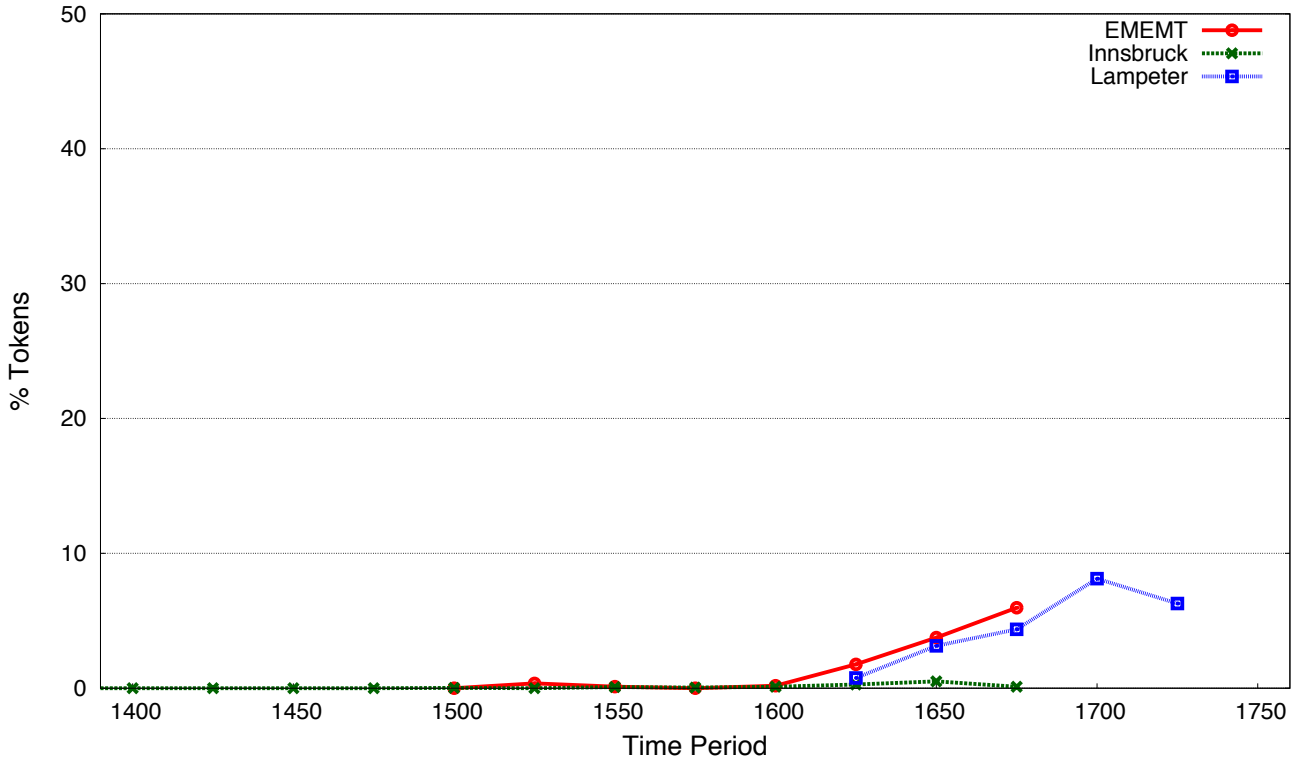
-'d → -ed

- Examples:
  - call'd (called)
  - pleas'd (pleased)
  - prov'd (proved)
- Difference between corpora:
  - 10th in EMEMT.
  - 91st in Innsbruck.
  - 2nd in Lampeter.



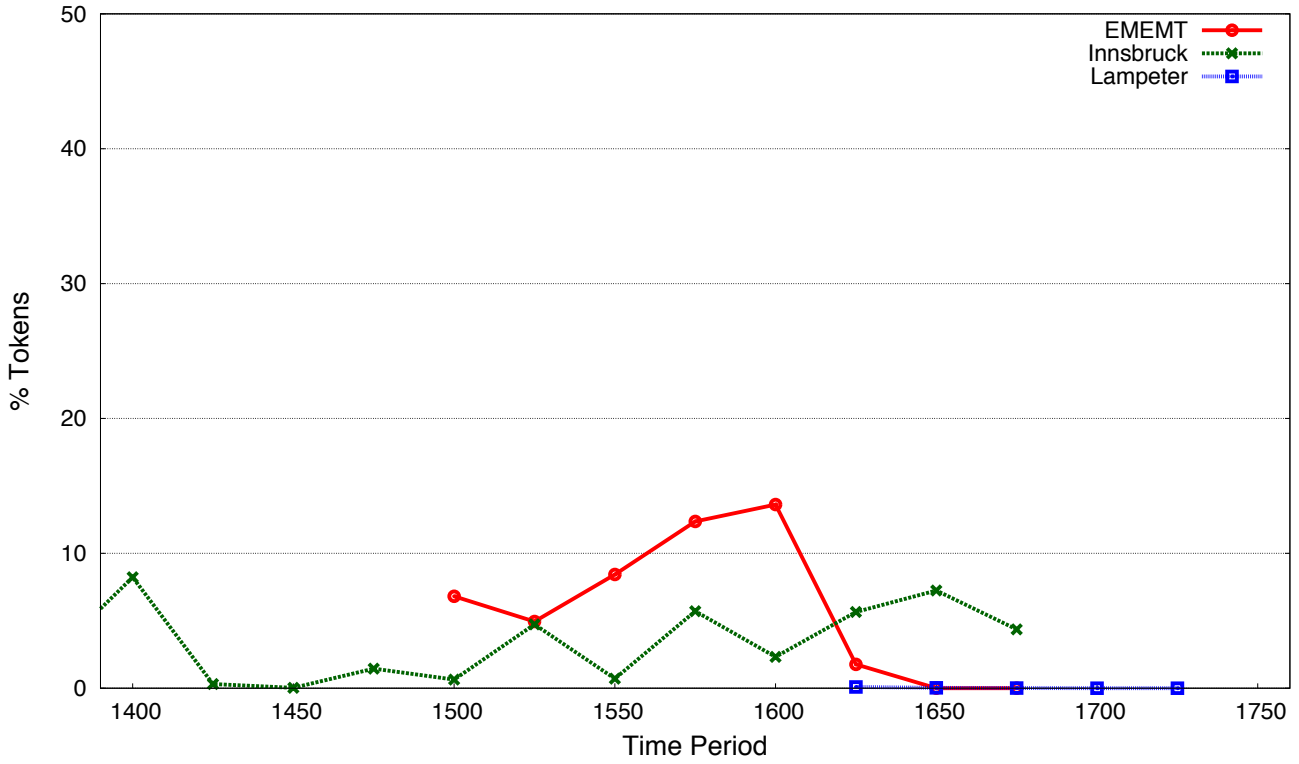
ck → c

- Examples:
  - Physick (Physic)
  - publick (publick)
  - Zodiack (Zodiac)
- Vast majority –ick endings.
- Lower frequency:
  - 21st in EMEMT.
  - 138th in Innsbruck.
  - 5th in Lampeter.



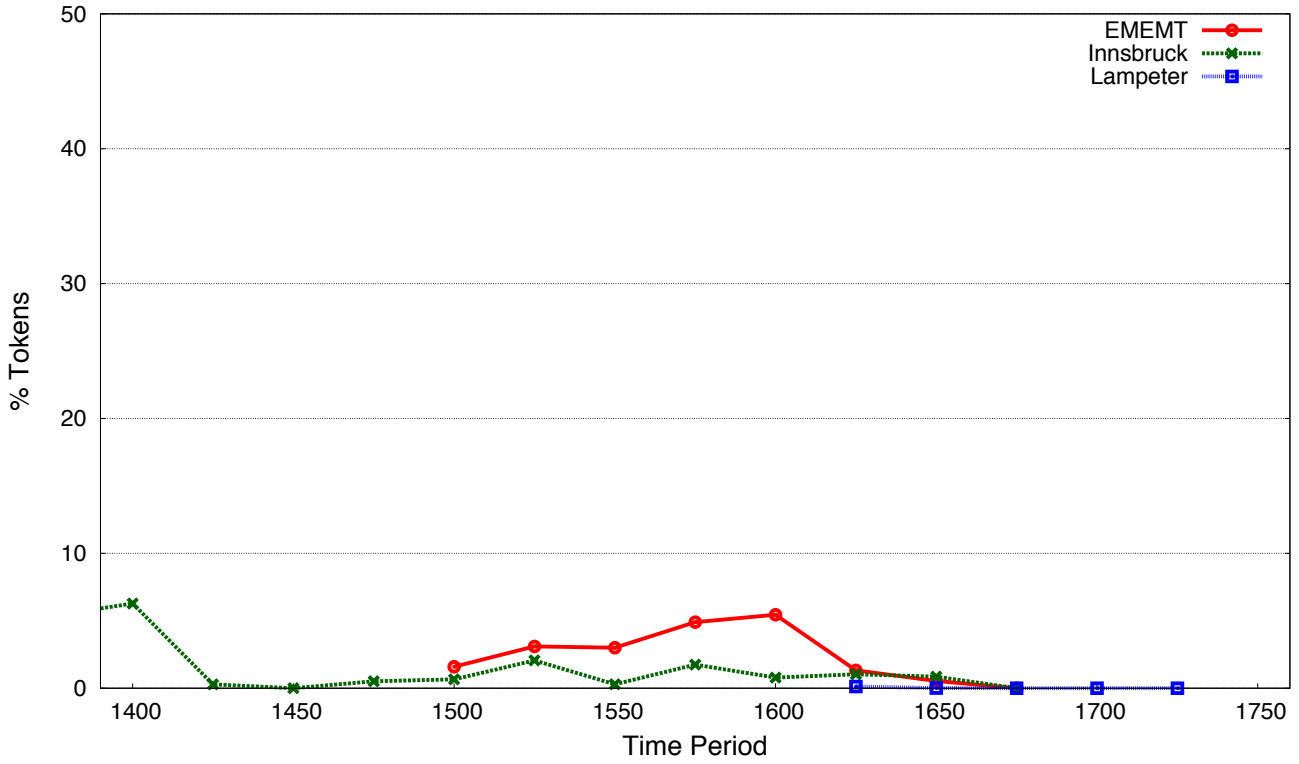
U → V

- Examples:
  - neuer (never)
  - haue (have)
  - Uote (Vote)
- Mainly middle of variant.
- (Mostly) high frequency:
  - 3rd in EMEMT.
  - 4th in Innsbruck.
  - 91st in Lampeter.



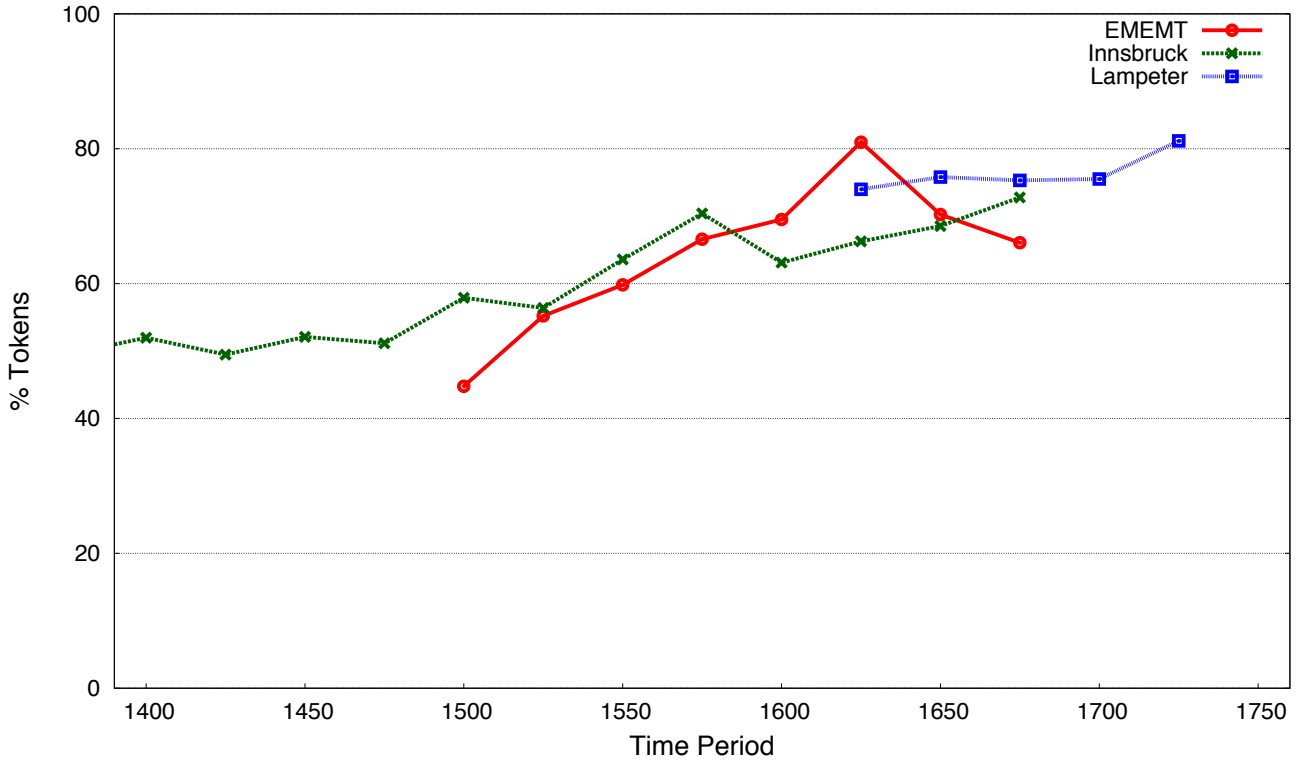
V → U

- Examples:
  - vpon (upon)
  - vs (us)
  - Vnicorn (Unicorn)
- Nearly always first letter.
- Less frequent:
  - 8th in EMEMT.
  - 22nd in Innsbruck.
  - 135th in Lampeter.

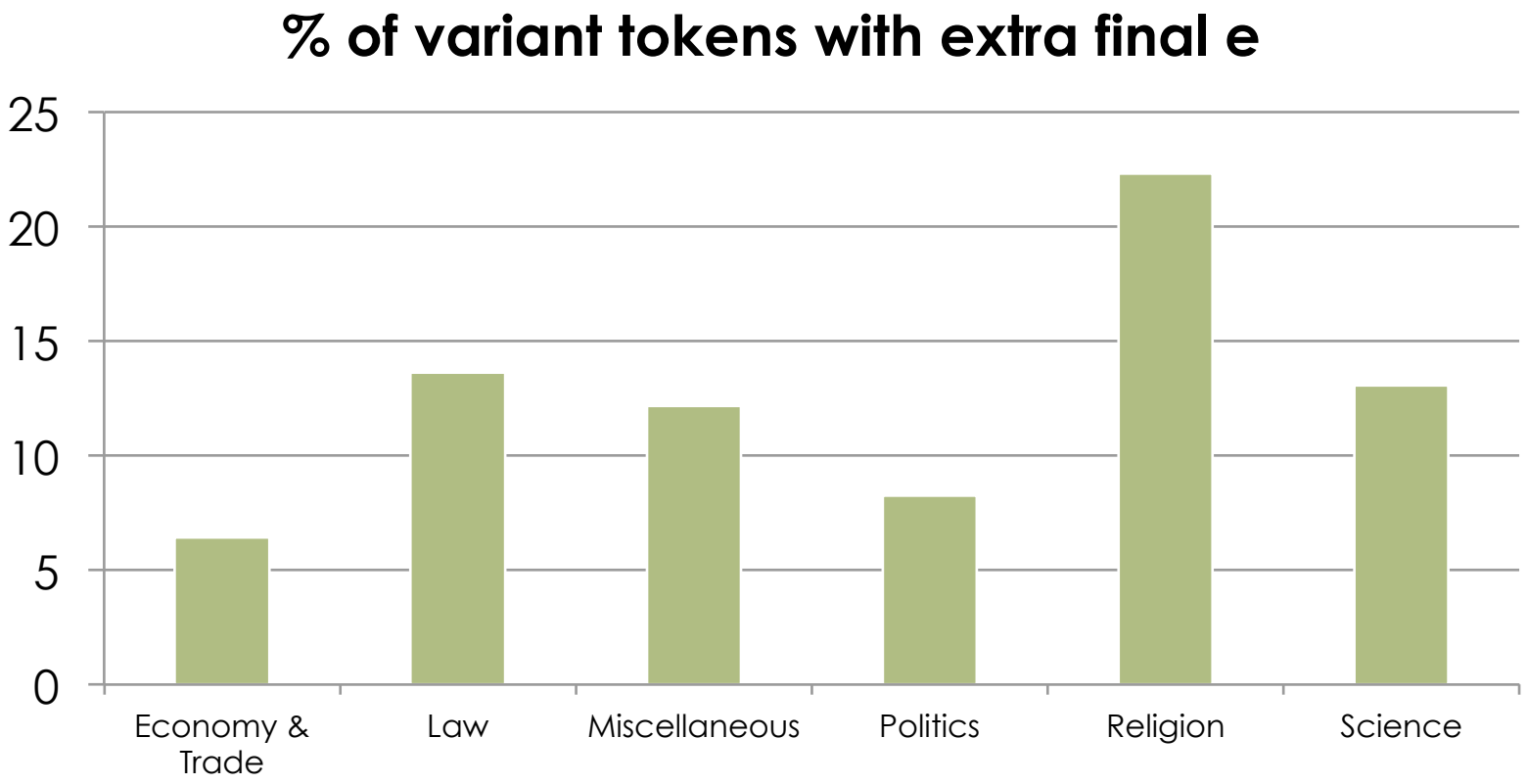


Single edits

- Single edit variants, e.g. one insertion, deletion or substitution from the standard form.
- Generally easier to normalise automatically.
- More variants requiring more than one edit in later texts makes spelling normalisation harder further back in time.

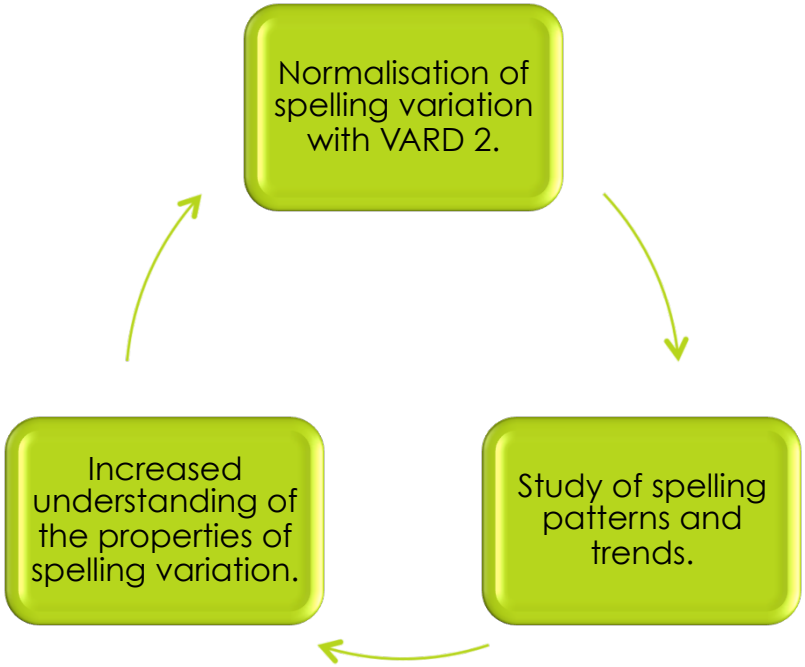


Lampeter Domain



Future work

- Further analyse DICER results to search for (new) trends over time, genre and text types.
- Look at other (larger) datasets, such as Early English Books Online.
- Incorporate DICER into VARD 2 to allow for learning normalisation rules "on the fly".



Thanks for listening

- **Acknowledgements:**
  - Thanks to Irma Taavitsainen and the Helsinki team for providing the EMEMT corpus, particularly Anu Lehto for the manual normalised samples.
  - Thanks to Manfred Markus for providing the Innsbruck Letters corpus with manually checked normalised text.
  - Research funded by EPSRC PhD Plus at Lancaster University.

- **More information:**
  - VARD: http://www.comp.lancs.ac.uk/~barona/vard
  - DICER: http://corpora.lancs.ac.uk/dicer

References

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. (2009). Automatic standardisation of texts containing spelling variation: How much training data do you need? In M. Mahlberg, V. González-Díaz & C. Smith, eds., Proceedings of Corpus Linguistics 2009, University of Liverpool, Liverpool, 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.

Görlach, M. (1991). Introduction to Early Modern English. Cambridge University Press, Cambridge.

References

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

Nevalainen, T. (2006). An Introduction to Early Modern English. Edinburgh Textbooks on the English Language, Edinburgh University Press, Edinburgh.

Rayson, P., Archer, D., Baron, A., Culpeper, J. & Smith, N. (2007). Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In M. Davies, P. Rayson, S. Hunston & P. Danielsson, eds., Proceedings of Corpus Linguistics 2007, UCREL, Lancaster University, Lancaster, UK.

References

Schmied, J. (1994). The Lampeter Corpus of Early Modern English Tracts. In M. Kytö, M. Rissanen & S. Wright, eds., Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, Rodopi, Amsterdam, St. Catherine's College, Cambridge.

Taavitsainen, I. & Pahta, P., eds. (2010). Early Modern English Medical Texts: Corpus description and studies. John Benjamins, Amsterdam.

Vallins, G.H. & Scragg, D.G. (1965). Spelling. André Deutsch, London.