Automatic Standardization of Spelling for Historical Text Mining

Alistair Baron, Paul Rayson and Dawn Archer

ucrel.lancs.ac.uk/VariantSpelling/
www.comp.lancs.ac.uk/~barona/vard2/
juilland.comp.lancs.ac.uk/dicer/

Motivation

Early Modern English spelling variation

Book production increased sharply during the Early Modern English period due to the introduction of the printing press in 1476 and increasing literacy levels. As such, Early Modern English is the earliest period of the English language from which a reasonably large corpus can be constructed and studied in detail.

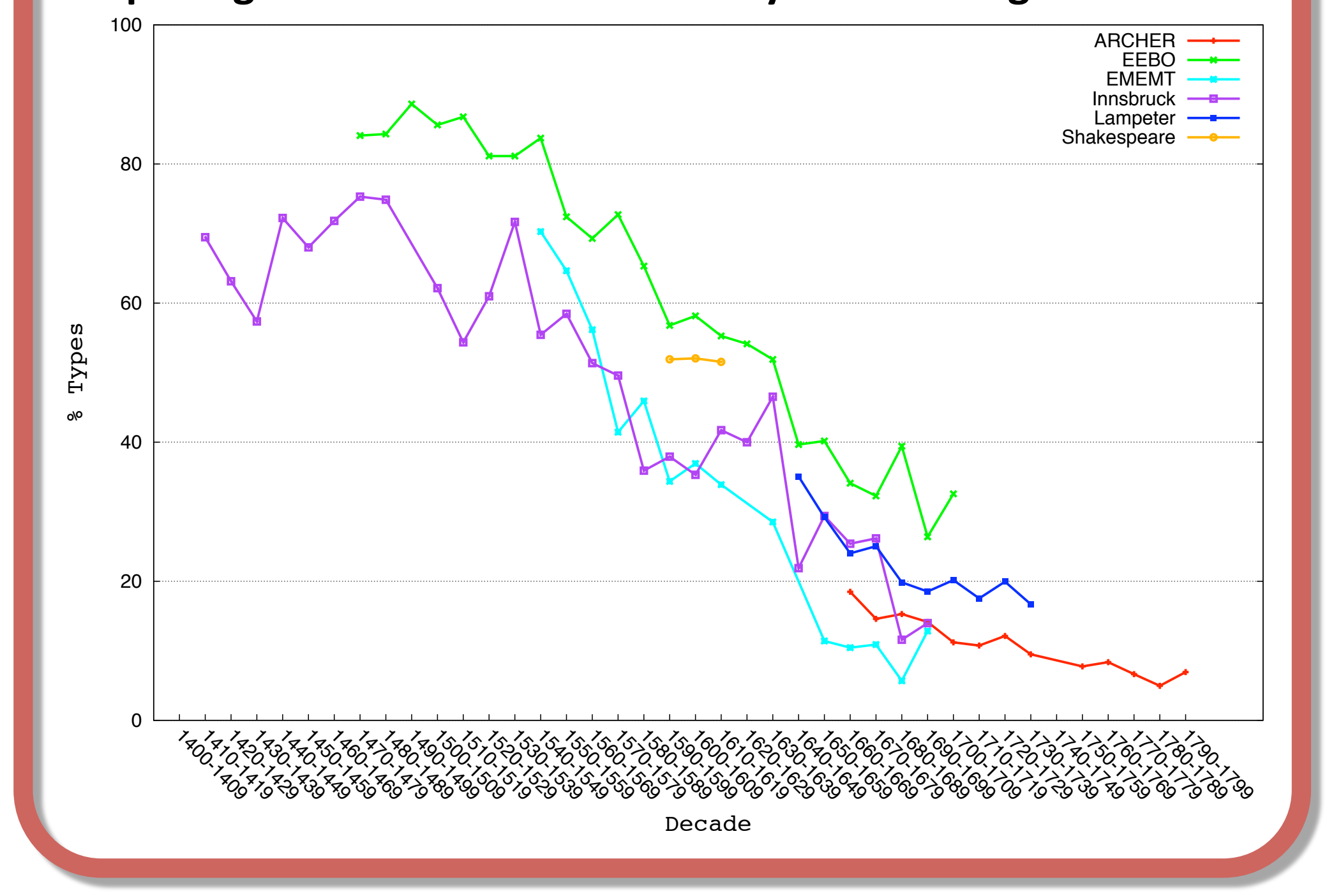
Spelling variation is a major feature of Early Modern English texts, we have recently quantified the extent of this spelling variation and verified previous claims that the amount of spelling variation was decreasing throughout the period until around 1700, this is shown in the graph below (see Baron et al, 2009 for more details).

The spelling variation is particularly difficult to deal with as it is common to find words spelt in a number of different forms in the same text, or even on the same page. This was due to there being no notion of the importance for a single spelling for each word; for example, letters would be added or removed to ease line justification. Some typical spelling variants found in Early Modern English texts are shown in the table below.

Examples of common spelling variants found in Early Modern English texts

Variant	Modern Equivalent	Notes
“goodnesse”	“goodness”	‘e’ often added to end of words.
“brush’d”	“brushed”	Apostrophes used instead of ‘e’.
“encrease”	“increase”	Vowels often interchanged.
“spels”	“spells”	Consonants often doubled or singled.
“deliuering”	“delivering”	Common for ‘u’ and ‘v’ to be interchangeable.
“conuay’d”	“conveyed”	Many combinations of the above.

Spelling variation levels in the Early Modern English Period



Problem

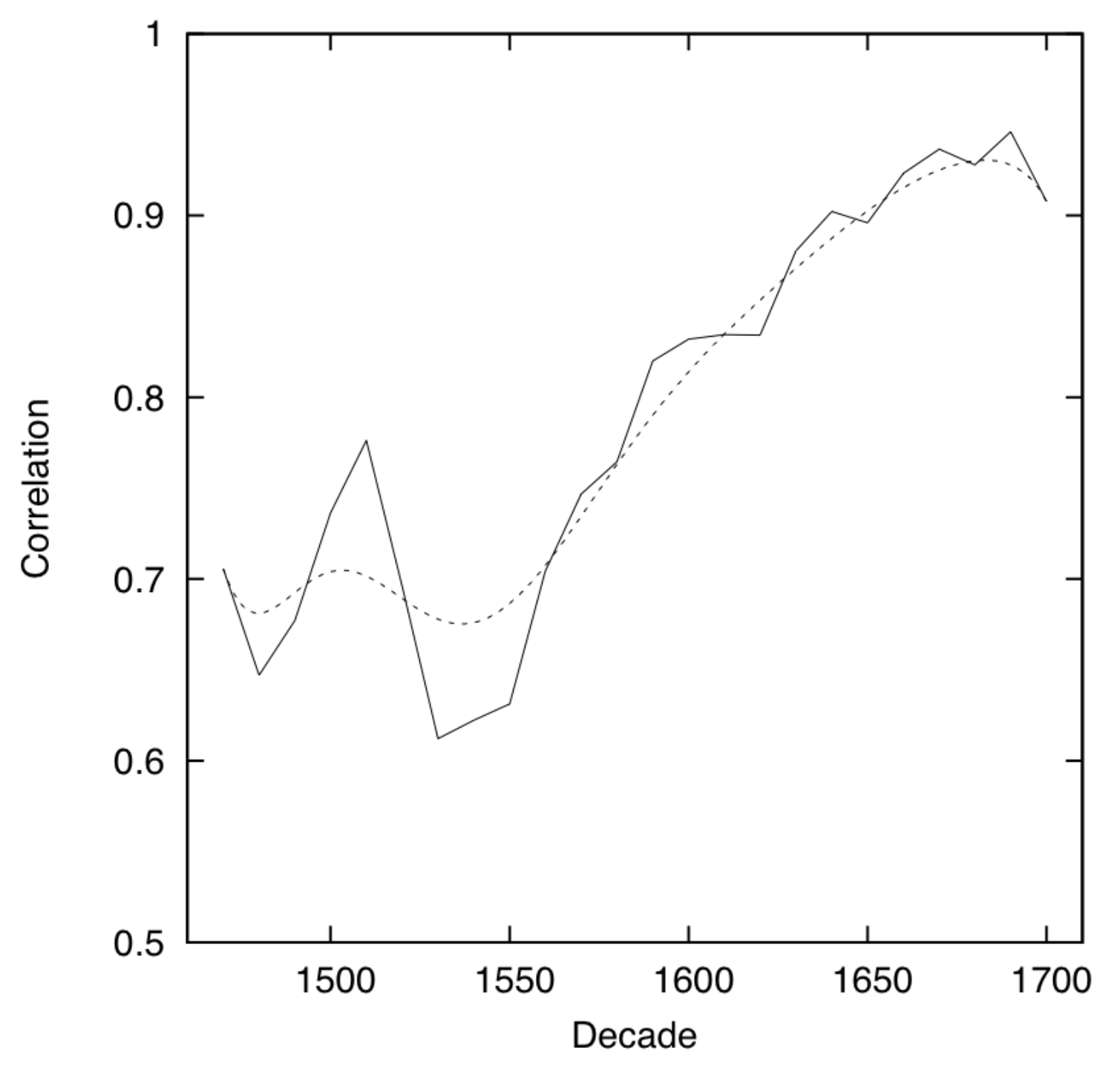
The effect of spelling variation on textual analysis

Key word analysis

In Baron et al (2009) we showed that spelling variation has an effect on the accuracy of key word analysis. Two rank correlation coefficients were used to compare key word lists from an original corpus and its manually standardized equivalent. Kendall’s Tau Coefficient was found to be 0.53 and Spearman’s Coefficient was 0.7, proving that spelling variation has a considerable effect on key word analysis.

Furthermore, we showed that the decreasing amount of spelling variation in the period has a direct association to the effect on key word analysis. The graph to the right shows the correlation between keyword lists generated for a series of EEBO decade samples both before and after automatic (partial) standardization. A correlation of 1 indicates the key word lists have identical rankings.

Kendall’s Tau Rank Correlation Coefficient



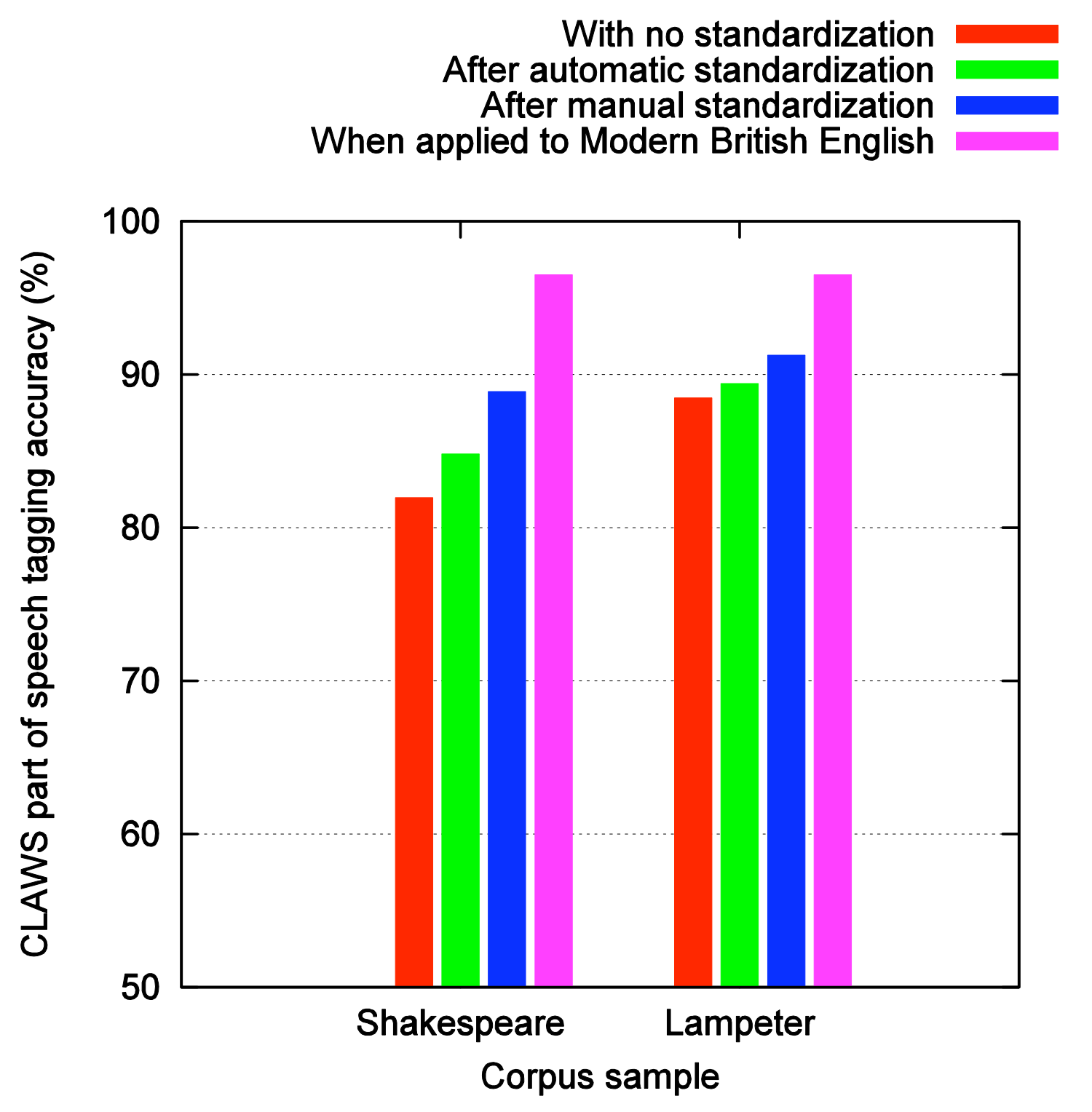
Annotation

The impact of spelling variation on part of speech and semantic tagging has been described in previous papers.

In Rayson et al (2007) the CLAWS part of speech tagger was evaluated on two Early Modern English samples, the results (summarized in the chart to the right) show than the accuracy of 96-97% found when annotating modern standard British English can not be expected when dealing with Early Modern English texts. However it was found that dealing with spelling variation significantly increased the tagger’s accuracy.

Archer et al (2003) discuss developing the USAS semantic tagger for Early Modern English. It was found that when dealing with relatively contemporary texts from 1640, dealing in part with spelling variation reduced error rates from 2.9% to 1.2% in one text and from 4.0% to 1.4% in the other text evaluation was carried out upon.

Part of speech tagging accuracy with Early Modern English texts



Solution

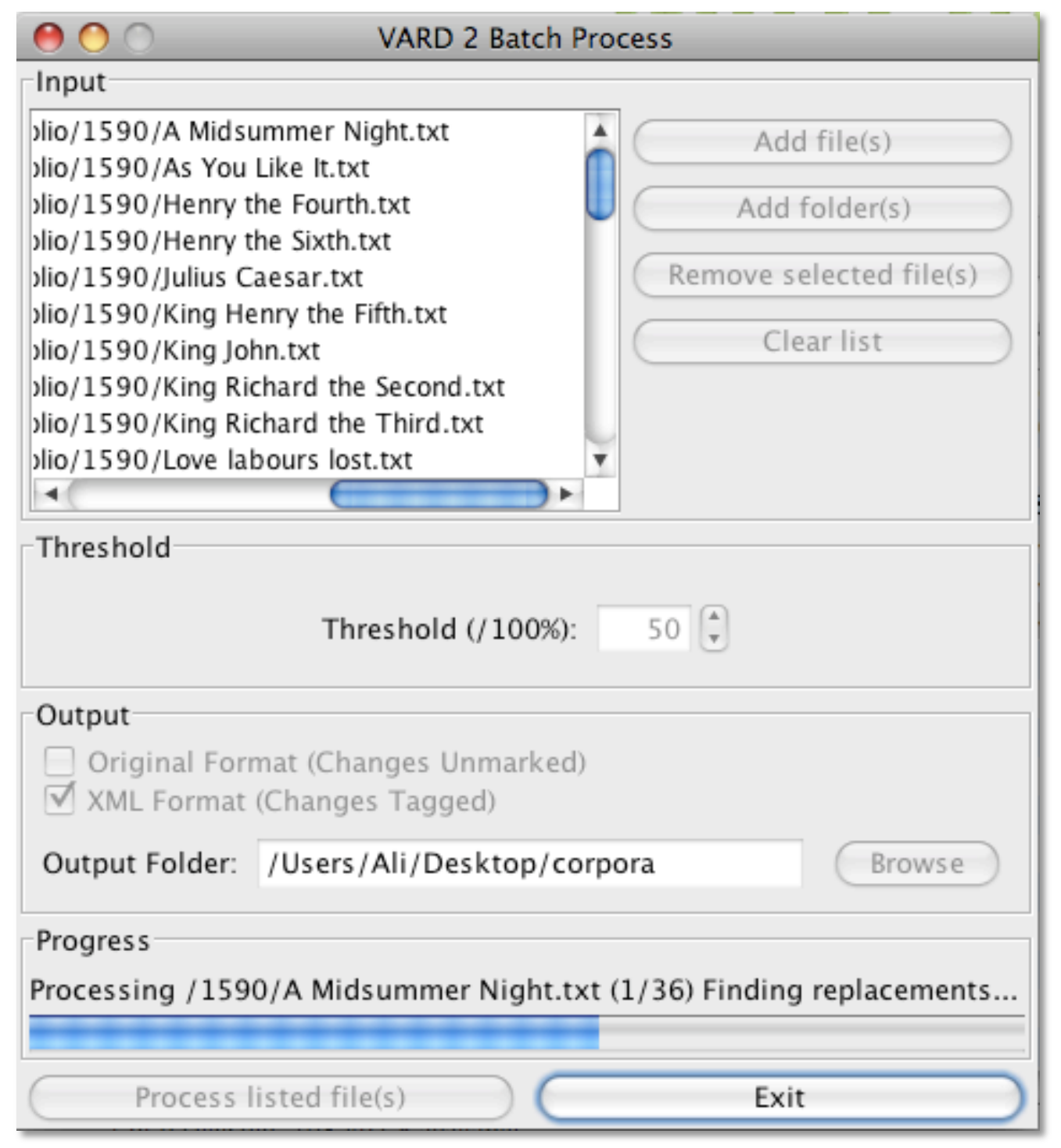
Automatic standardization with VARD 2

VARD 2 is a piece of software designed to assist researchers in standardizing historical corpora both manually and automatically. Variants are detected by comparing each word in a text with a modern word list. Replacements are found for these variants with a manually created list of variant replacements as well as employing methods from modern spell checking software; such as phonetic matching, letter replacement heuristics and Edit Distance. The original spelling is retained in the text with an xml tag surrounding the replacement.

For Early Modern English texts VARD 2’s batch mode (shown to the right) can be used with no training to automatically standardize the spelling variation in an entire corpus. However, for better results the user can train the software on a particular corpus by using the interactive mode (see below) to manually process samples of the corpus. This will improve the tool’s ability to deal with a corpus by learning which of its methods are most successful, editing its dictionary and adding specific variant replacements.

Shakespeare’s First Folio was automatically standardized to test VARD 2. With no training, the tool can deal with **70.33%** of tokens deemed by the tool to be variants. Using a 5,000-word sample (6% of the entire corpus) as training data, this increases to **73.75%**, a respectable improvement considering the small size of the sample.

VARD 2 – Batch Mode



olio/1590/A Midsummer Night.txt
olio/1590/As You Like It.txt
olio/1590/Henry the Fourth.txt
olio/1590/Henry the Sixth.txt
olio/1590/Julius Caesar.txt
olio/1590/King Henry the Fifth.txt
olio/1590/King John.txt
olio/1590/King Richard the Second.txt
olio/1590/King Richard the Third.txt
olio/1590/Love labours lost.txt

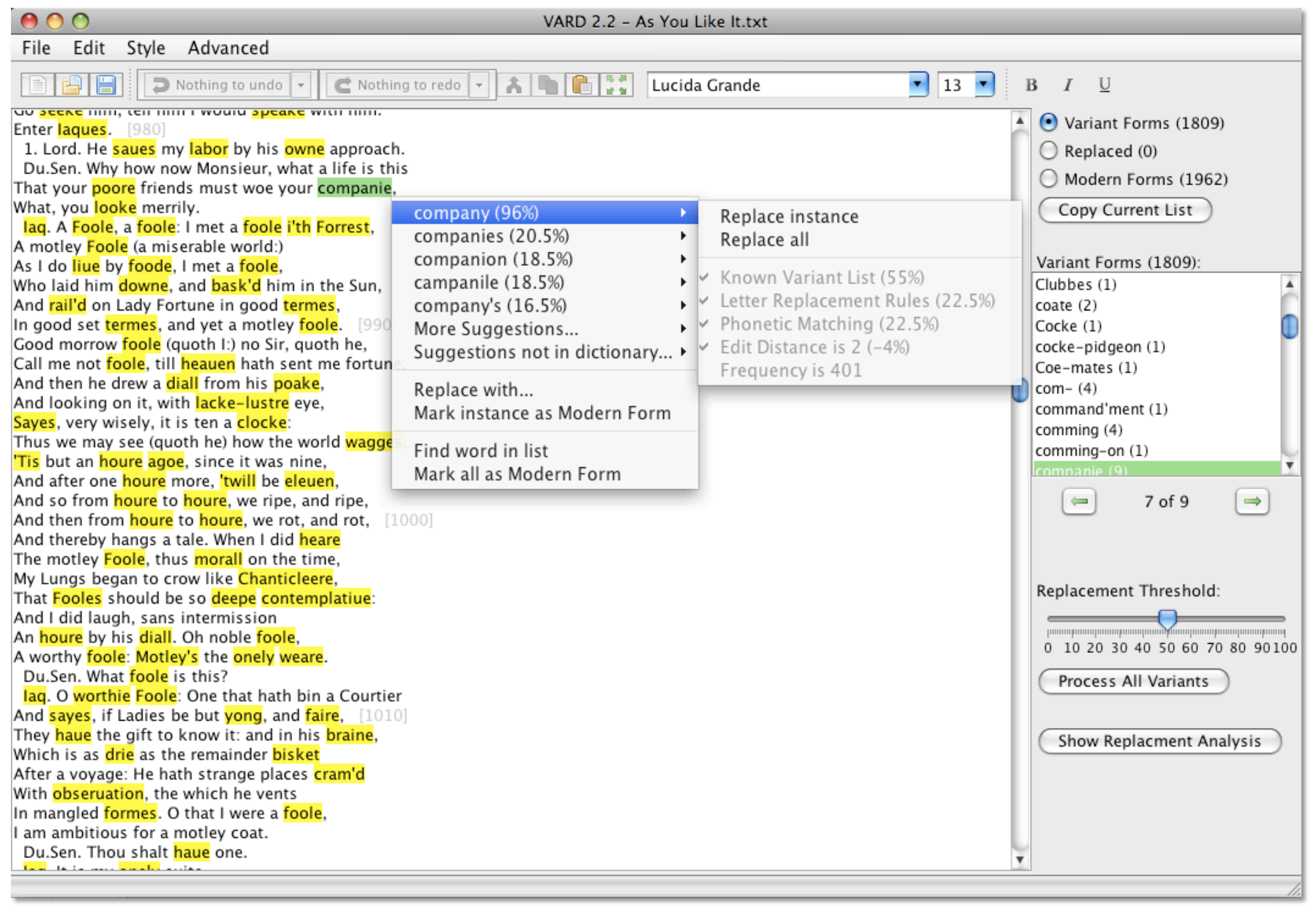
Threshold (/100%): 50

□ Original Format (Changes Unmarked)
☑ XML Format (Changes Tagged)

Output Folder: /Users/Ali/Desktop/corpora

Processing /1590/A Midsummer Night.txt (1/36) Finding replacements...

VARD 2 – Interactive Mode



Go seeke him, tell him I would speake with him.
Enter Iaques.
 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.
 Iaq. 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.
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 wagges:
'Tis but an houre agoe, since it was nine,
And after one houre more, 'twill be eleuen,
And so from houre to houre, we ripe, and ripe,
And then from houre to houre, we rot, and rot,
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 houre by his diall. Oh noble foole,
A worthy foole: Motley's the onely weare.
 Du.Sen. What foole is this?
 Iaq. O worthie Foole: 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 obseruation, the which he vents
In mangled formes. O that I were a foole,
I am ambitious for a motley coat.
 Du.Sen. Thou shalt haue one.

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

● Variant Forms (1809)
○ Replaced (0)
○ Modern Forms (1962)

Variant Forms (1809):
Clubbes (1)
coate (2)
Cocke (1)
cocke-pidgeon (1)
Coe-mates (1)
com- (4)
command'ment (1)
comming (4)
comming-on (1)
companie (9)

Replacement Threshold: 0 10 20 30 40 50 60 70 80 90 100

Conclusions and Future Work

The initial evaluation results shown here are promising, increasing the size of the training data should improve these figures further. The DICER analysis can also be used to greater effect as the tool can be used to provide probabilities dictating how likely a rule should be applied in a given position. Modifying VARD 2 to use these probabilities will see even greater improvements in performance. With further training it is plausible that VARD 2 could be used with other varieties of non-standard English (e.g. SMS corpora and weblogs).

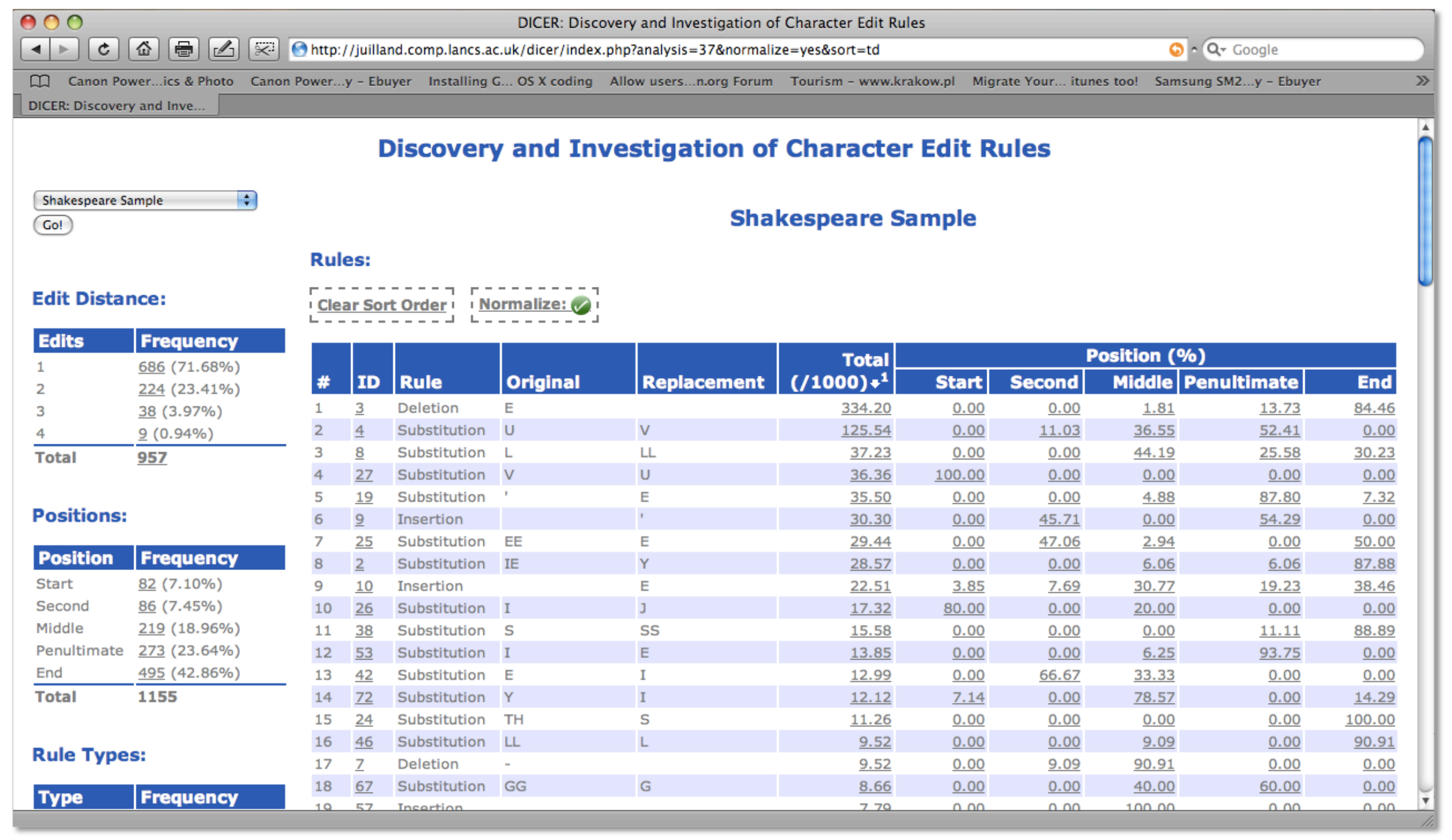
The variant levels described here are estimates as some words will be incorrectly marked as variants (particularly proper names) whilst other variants are not detectable as they occur in the modern lexicon (i.e. *real-word errors*, such as ‘bee’ for ‘be’). To deal with these problems more advanced techniques, e.g. part-of-speech tagging, need to be used in the detection phase to take into account the context of a given word.

DICER is another tool under development which allows further training of VARD 2. XML output from VARD 2 is processed with variant replacements collected and analyzed. Character edit rules are produced which could transform the variant into its modern equivalent, these are then collated into a database, which can be viewed through a set of web pages. An example of a main table is shown to the right.

By using DICER to analyze manually standardized samples of a corpus, a list of common character edit rules can be created. VARD 2’s own list of such rules can then be augmented with these rules to improve its performance when finding potential replacements.

Automatic standardization with VARD 2 (after training) of Shakespeare’s First Folio resulted in 10,601 unique variant replacements. VARD 2’s original rule set alone could deal with **70.35%** of these replacements. When the rule list was augmented with additional rules from DICER this increased to **77.66%**.

DICER (Discovery and Investigation of Character Edit Rules) Website


DICER: Discovery and Investigation of Character Edit Rules
http://juilland.comp.lancs.ac.uk/dicer/index.php?analysis=37&normalize=yes&sort=id
Canon Power...ics & Photo
Canon Power...y – Ebuyer
Installing G... OS X coding
Allow users...n.org Forum
Tourism
DICER: Discovery and Inve...
Shakespeare Sample
Go!
Clear Sort Order
Normalize: ✓

Discovery and Investigation of Character Edit Rules

Shakespeare Sample

Edit Distance:

Edits	Frequency
1	686 (71.68%)
2	224 (23.41%)
3	38 (3.97%)
4	9 (0.94%)
Total	957

Positions:

Position	Frequency
Start	82 (7.10%)
Second	86 (7.45%)
Middle	219 (18.96%)
Penultimate	273 (23.64%)
End	495 (42.86%)
Total	1155

Rule Types:

Type	Frequency

Rules:

#	ID	Rule	Original	Replacement	Total (/1000)↓1	Position (%)				
						Start	Second	Middle	Penultimate	End
1	3	Deletion	E		334.20	0.00	0.00	1.81	13.73	84.46
2	4	Substitution	U	V	125.54	0.00	11.03	36.55	52.41	0.00
3	8	Substitution	L	LL	37.23	0.00	0.00	44.19	25.58	30.23
4	27	Substitution	V	U	36.36	100.00	0.00	0.00	0.00	0.00
5	19	Substitution	'	E	35.50	0.00	0.00	4.88	87.80	7.32
6	9	Insertion		'	30.30	0.00	45.71	0.00	54.29	0.00
7	25	Substitution	EE	E	29.44	0.00	47.06	2.94	0.00	50.00
8	2	Substitution	IE	Y	28.57	0.00	0.00	6.06	6.06	87.88
9	10	Insertion		E	22.51	3.85	7.69	30.77	19.23	38.46
10	26	Substitution	I	J	17.32	80.00	0.00	20.00	0.00	0.00
11	38	Substitution	S	SS	15.58	0.00	0.00	0.00	11.11	88.89
12	53	Substitution	I	E	13.85	0.00	0.00	6.25	93.75	0.00
13	42	Substitution	E	I	12.99	0.00	66.67	33.33	0.00	0.00
14	72	Substitution	Y	I	12.12	7.14	0.00	78.57	0.00	14.29
15	24	Substitution	TH	S	11.26	0.00	0.00	0.00	0.00	100.00
16	46	Substitution	LL	L	9.52	0.00	0.00	9.09	0.00	90.91
17	7	Deletion	-		9.52	0.00	9.09	90.91	0.00	0.00
18	67	Substitution	GG	G	8.66	0.00	0.00	40.00	60.00	0.00
19	57	Insertion			7.79	0.00	0.00	100.00	0.00	0.00