Grappling with Shakespeare's words: maximizing historical corpus-based approaches

Jonathan Culpeper and Amelia Joulain-Jay
Lancaster University

@ShakespeareLang
http://wp.lancs.ac.uk/shakespearelang/
The Encyclopedia of Shakespeare’s Language project: A corpus-based approach to Shakespeare

- For Shakespeare’s language this means:
  - All ‘words’ treated equally (not just ‘hard’ words).
  - Meanings based on usage in context (not etymology, not narrowly-defined semantic meaning).
  - Context includes linguistic aspects (e.g. collocations) and non-linguistic aspects (e.g. registers, social properties of speaker/character).
  - A comparative approach.
What will be in the encyclopaedia?

- Volume 1 (a kind of dictionary)
  - The use and meanings of each of Shakespeare's words, both in the context of what he wrote and in the context in which he wrote.
  - Every word is, for example, compared with a 321 million word corpus comprising the work of Shakespeare's contemporaries.
  - Establishes both what is unique about Shakespeare's language and what Shakespeare's language meant to his contemporaries.
What will be in the encyclopaedia?

- Volume 2 (a kind of compendium of semantic patterns)

  ➢ Plays and characters
    - Major characters (> 5% of total word count)
    - Play profiles

  ➢ Gender and social stratification
    - Male/female
    - Social status

  ➢ Genre

  ➢ Major themes in Shakespeare
    - Love and marriage
    - War and conflict
    - etc.
Our core data: Shakespeare texts

- Core data: plays generally agreed to be part of the Shakespeare canon
  - The largest near-contemporary body of work attributed to Shakespeare, i.e. the First Folio (1623), plus *Pericles* and *The Two Noble Kinsmen*.
- Quartos constitute a secondary dataset.
- Poetry constitutes a third Shakespeare dataset.
Spelling variation

• If untreated, frequency counts for a word would be split across several variants
  – e.g. in our core Shakespeare data, *would* is also spelled *vvould, wold, wad*
• Addressing this improves the prospects for matching words in target and reference corpora
• Easier for the present-day reader/user
• Our solution: use VARiant Detector (VARD 2) software (Baron & Rayson 2008; http://ucrel.lancs.ac.uk/vard/)
THE ADVOCATE. ------ LONDON. Printed by William Du-Card, Printer to the Council of State; and are to bee sold by Nicolas Bourn in Cornhill, at the Corner shop, at the entrance into the Exchange. 1652.

TO THE RIGHT HONORABLE THE Council of State. Right Honorable! I am often in very great doubt (if I may so speake), that the Goodness & Wisdom of God, & his thoughts of these, are very rarely met with in the Paths, which the scantling of Man's Reason and Judgment walk in; And as I dare not but own the Belief of the Coming of his Appearance, and the breaking forth, very shortly, of his Gloire. So I believe likewise, this will bee a sight very strange, and very unexpected to men, and not only greatly above, but in some measure even contrary (and perhaps, very unwelcome) unto the most enlarged and raised thoughts wee have yet prepared our selves with, to receiue it.

The sign of which Coming, will bee the Detection, by little and little, of all Imposture, and the laying of all things low, naked, and mean before him; the stripping men of that Honor, Credit, and Repute, that they had by several means been gaining with themselves, and (by themselves) with others, either through a Not-knowled, or through an artificial concealing of the menselv.

In these very things (if well observed) lying, and within these indeed, being spent and consumed, the whole Indexers, Practice, Studie and Wisdom (if not Religion) of All States, Ages, Nations and Men, viz. Either in devising shifts (by acquisition of such and such Power, Habits, or otherwise) not to appear uncomely to themselves, but to beget rather a greatness, esteem and satisfaction in themselves, and others of themselves.

Or, in Concealing and keeping close (by several Glasses) the Imperfections and apprehensions of others.

The more exquisite themselves, but one small part of the whole. The more fortunate man in this world, having the Advantage of not rising above the fall of his fall sooner, and the other wrapps himself up in it.

For thus all Lusts whatsoever (whether those of Acquisition) in the mind, and all Accomplishments, are still but either to give content unto, or to beget the meanest Lusters, and to beget the meanest Lusters.

The Currency (during it) to all Lusts. (These two, observing tune, and holding time each to the measures and motion therefore is concluded whole Humane Nature, with all the Parts, and the most perfect Actions of it.

Upon these thoughts (Right Honorable) I finding none not very ready to have ascribed to himself the name of being (in some measure wise).

because it is estimable, finde none that can endure the Manifestation (in any kind) of their Folie; or that can bee willing to have this their Wisdom and Gloire to bee flatly and plainly called a Puerpitude, Humming, or playing at bo-peep, with themselves and others; and yet (it really being no honest) if wee cannot bear the judgments thus one of another, how shall wee abide the Trial, or bee able to stand before the Judgment of God, which seeth us, (and will shortly search us) with more pure and uncorrupt Lies! And how miserably then are the Imaginations of almost all men muffled up here in their own dressings? when the Truth most certainly is in our Beings, at their best (natural, or acquire) Appearances, are yet such fignor Things, and Prop'd up with such Lyes and fundamental mistakes of themselves, that they would (beelshazer like) moulder and shatter themselves to pieces, should they bee taken quite asunder from the disguises and patchings they have put upon themselves, bare-facedly and undecievably to see themselves by.

Beeing convinced, by these, therefore (Right Honourable) of the low Condition wee are herein, I not knowing what the Counsils of God intend to bring forth for the settlement of this Nation: Nor how he hath resolved in his Wisdom to dispose of it, (as to its outward Condition,) whether Hee...
**Spelling variation**

**Problem:** What do you regularize the spelling to? There is no standardised regular form in the way that there is today.

**Solution:** Our policy was to

- Preserve the morphology, e.g. 2\textsuperscript{nd} + 3\textsuperscript{rd} person verb inflections (\textasciitilde(e)st, -(e)th), past tense forms (e.g. holp), past participle forms (e.g. holpen), plural forms (e.g. shooen), non-standard superlatives (e.g. horrider), and you/thou,
- Only use a spelling that had EModE currency.
- Prioritize the most frequent spelling in Shakespeare
Part of speech tagging and EModE (cf. Andrew Hardie)

CLAWS performs at 85% accuracy for Shakespearean texts (Rayson et al. 2007). Not good enough!

- Grammatical phenomena which are marginal today, but may require addressing by the tagger for EModE (e.g. 2nd person singular)
- Words not in the modern lexicon (e.g. hent)
- Words whose possible classifications have changed (e.g. faith)
- Words whose probability profile has changed (e.g. prostitute)
- Extra cliticisations (e.g. me=thinks, me=thought)
- No time for a radical system overhaul
- Solution: Patches for CLAWS
Implementation

Input patching script

Wordlist patches
Idiomlist patches

Output patching script

Untagged text

CLAWS input

CLAWS resources

CLAWS

CLAWS output

Tagged text
Data and genre: *Early English Books Online (TCP)*
1560-1639 (379 million words; 5,750 texts categorized by genre, domain and style)

<table>
<thead>
<tr>
<th>Styles</th>
<th>Domains</th>
<th>Genres</th>
<th>Sub-genres (examples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literary</td>
<td>Imaginative</td>
<td>Plays, Poetry, Verse &amp; Song, Fiction, General</td>
<td>Comedy, History, Tragedy, Masque</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ballads</td>
</tr>
<tr>
<td>Formal – Spiritual</td>
<td>Religion</td>
<td>Bible, Catholicism, Protestantism, Doctrine, Theology and Governance, General</td>
<td>Anti-Catholicism, Church of England, Sin and Repentance, Sermons</td>
</tr>
<tr>
<td>Formal - Statutory</td>
<td>Government</td>
<td>Royal, Parliamentary, Legal, General</td>
<td>Proceedings, Reports, Trials, Speeches</td>
</tr>
<tr>
<td>Formal - Instructional</td>
<td>Didactic</td>
<td>Philosophy, Science, Mathematics, Medicine, General</td>
<td>Experiments, Anatomy, Alchemy</td>
</tr>
<tr>
<td>Informational</td>
<td>Factual</td>
<td>Biography, Essay, Letters, Pamphlets, General</td>
<td>Dialogue</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Food and Cookery</td>
</tr>
</tbody>
</table>
The challenge

Multiple pieces of information:

- spellings, parts-of-speech, collocates, genre distribution, social distribution (e.g. male/female; high rank/low rank)

In multiple information sets:

- Shakespeare's plays, his poetry, the Folios, the Quartos, our comparative corpus of playwrights and the EEBO-TCP.
Lexicography interface: background

- We have CPQweb... why have another interface?
  - automation of repetitive tasks
  - dictionary writing system

- What does the system involve?
  - a MySQL database with two parts:
    - fixed data (number of occurrences, etc.)
    - modifiable data (definitions, etc.)
  - a user interface which:
    - provides access to the data
    - allows for the creation of modifiable data
Lexicography interface: set-up

- Corpus is pre-processed
- Corpus is uploaded to CQPweb
- Data is extracted from CQPweb and stored in the fixed part of the database
- Interface displays fixed data
- Users interact with the interface to create data which gets stored in the modifiable part of the database
- Interface displays fixed and modifiable data
Lexicography interface: demo

https://corpora.lancs.ac.uk/shakencyc/
A more grammatical word: I

Shakespearean dictionaries:

• Words such as this omitted from Shakespearean dictionaries (e.g. Crystal and Crystal 2002; Onions 1986), presumably on the assumption that they:
  (a) have obvious meanings (because they are considered more or less the same as those of today), and
  (b) do not contribute much to understanding Shakespeare.
A more grammatical word: I

Encyclopaedia of Shakespeare's Language

Definition preview: definition 122 for entry i_PRON

i***** pron. (I, me):
As now. Most regularly co-occurs with verbs, relating to mental states/processes (e.g. HOPE, THINK, KNOW, REMEMBER, FORGET, DREAM, UNDERSTAND, DOUBT, WISH) and speech act devices (e.g. PRAY, THANK, SWEAR, BESEECH, ASSURE, PROMISE, BEG, ENTREAT, MEAN). Co-occurs frequently with verbs relating to communication (e.g. TELL, SAY, READ, WRITE), movement (e.g. GIVE, TAKE, GO), emotion state/processes (e.g. FEAR, LOVE, HATE) and perception (e.g. SEE, HEAR, WATCH). Aside from modal verbs (e.g. CAN, WILL, MAY) and adverbs, especially involving negation (e.g. NEVER, NOT, NO), it co-occurs regularly with the pronouns THOU, YOU, MY and IT.
### Desdemona’s keywords

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Raw freq.</th>
<th>Log-L.</th>
<th>LogRatio</th>
</tr>
</thead>
<tbody>
<tr>
<td>prithee</td>
<td>8</td>
<td>16.47</td>
<td>3.24</td>
</tr>
<tr>
<td>lord</td>
<td>39</td>
<td>64.82</td>
<td>2.74</td>
</tr>
<tr>
<td>lost</td>
<td>7</td>
<td>10.4</td>
<td>2.53</td>
</tr>
<tr>
<td>alas</td>
<td>8</td>
<td>8.7</td>
<td>2.04</td>
</tr>
<tr>
<td>him</td>
<td>41</td>
<td>24.75</td>
<td>1.41</td>
</tr>
<tr>
<td>do</td>
<td>44</td>
<td>19.64</td>
<td>1.18</td>
</tr>
<tr>
<td>my</td>
<td>79</td>
<td>28.03</td>
<td>1.03</td>
</tr>
<tr>
<td>me</td>
<td>47</td>
<td>11.61</td>
<td>0.84</td>
</tr>
<tr>
<td>i</td>
<td>132</td>
<td>26.85</td>
<td>0.76</td>
</tr>
</tbody>
</table>

For Othello: 'I' is ranked 109, 'me' 70 and 'my' 74
A more lexical word: *good*

Dictionaries (in brief):

- *Onions* (1911): (1) Conventional epithet to titles of high rank, (2) comely, (3) Financially sound; (hence) wealthy, substantial.

- *Crystal & Crystal* (2004): (1) [intensifying use] real, genuine (‘love no man in good earnest’). (2) kind, benevolent, generous. (3) kind, friendly, sympathetic. (4) amenable, tractable, manageable. (5) honest, virtuous, honourable. (6) seasonable, appropriate proper. (7) just, right, commendable. (8) intended, right, proper. (9) high-ranking, highborn, distinguished. (10) rich, wealthy, substantial.
good**** adj. (good, better, best):
1. A polite address: '(my) good Lord/friend/Sir/Master/Lady/Madam/etc.'. Typically used when meeting or parting, thanking or making suggestions. But (good my Lord) do it so cunningly TGV, III. 1.
2. Honest, truthful, principled; of high moral standards. (This sense also shapes the discourse markers '(in) good faith/sooth/truth', which mean truly or honestly). a man of good repute, carriage, bearing, & estimation LLL, I. 1.
3. Positive rather than negative. Typically, contrasted with 'bad'. Is thy news good or bad? ROM, II. 5.
4. In one's favour, especially favourable wishes or blessings. The Gods be good to us COR, V. 4.
5. A welcoming, cheerful manner. Therefore for Gods sake entertain good comfort, And cheer his Grace with quick and merry eyes R3, I. 3.
good will As now.
good morrow Good morning.
good night As now.
Questions?
Input patching

```c
10 // splits that CLAWS does for us already:
11 // twas
12 // tis
13 // twould
14 // twere
15
16 $splits_raw = <<<END
17 me|thinks
18 me|thought
19 me|thoughts
20 me|seems
21 me|seemeth
22 me|seemed
23 t|as
24 t|will
25 END;
```
### EModE tags

<table>
<thead>
<tr>
<th>Tag</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPYS1</td>
<td>thou, th’</td>
</tr>
<tr>
<td>PPYO1</td>
<td>thee</td>
</tr>
<tr>
<td>VBT</td>
<td>art, beest</td>
</tr>
<tr>
<td>VBDT</td>
<td>wast, wert</td>
</tr>
<tr>
<td>VDT</td>
<td>dost, doest</td>
</tr>
<tr>
<td>VDDT</td>
<td>didst</td>
</tr>
<tr>
<td>VHT</td>
<td>hast</td>
</tr>
<tr>
<td>VHDT</td>
<td>hadst</td>
</tr>
<tr>
<td>VMT</td>
<td>wilt, wouldst, canst, couldst, shalt, shouldst etc.</td>
</tr>
<tr>
<td>VMTK</td>
<td>oughtest, usedest</td>
</tr>
<tr>
<td>VVT</td>
<td>givest, workest</td>
</tr>
<tr>
<td>VVDT</td>
<td>gavest, workedst</td>
</tr>
</tbody>
</table>
Resource patching: lexicon

28 hall VM
29 wilt VM VVO NN1%
30 wouldst VM
31 canst VM
32 couldst VM

11 art VBR NN1@ NP1:
12 beeth VBZ
13 beest VBR
14 wast VBDR
15 wert VBDR

76 calledest VVD
77 calledest VVD
78 callst VVO
79 calleth WZ
80 callst VVO
81 camst VVD
82 camst VVD
83 comest VVO
84 cometh VVZ
85 comst VVO
86 cutteth WZ

404 / some extra nonlex adverbs
405 theneart RR
406 herean RR
407 therewithal RR
408 herewithal RR
409 thitherward RL
410 hitherward RL
411 thitherwards RL
412 hitherwards RL

358 unworthier JJR
359 verier JJR
360 violenter JJT
361 welcomest JJT
362 wholesomest JJT
363 willingest JJT
364 wiselier JJR
365 woefullest JJT
366 worser JJR

289 bid NN1 VVO VVN VVD
290 bids NN2 VWZ
291 fee NN1 VVO
292 fees NN2 VWZ
293 feed VVO NN1 VVD@ VVN@
294 prostitutes VWZ NN2@
295 prostitute VVO NN1@
296 passing VVG JJ RR NN1@
297 marry VVO UH
298 wanton VVO JJ NN1
299 wantons VWZ NN2

420 for IF CS
421 but CCB RR II
422 an AT1 CS%

436 / interjections, disc particles
437 o UH ZZ1@
438 fie UH
439 faith NN1 UH
440 infaith UH
441 in'faith UH
443 inprimis RR
443 iwis RR
444 prithee UH VVO@
Output patching

/ pronouns
thou PPY PPYS1
thee PPY PPY01
th PPY PPYS1
th’ PPY PPYS1

/ verbs
art VBR VBT
beest VBR VBT
wast VBDR VBDM
wert VBDR VBDT
hast VHO VHT
havest VHO VHT
hadst VHD VHDT
haddest VHD VHDT
dost VDO VDT
doest VDO VDT
didst VDD VDDT
wilt VM VMT
wouldst VM VMT
canst VM VMT

seemedst VVD VVDT
seemest VVO VVT
seemst VVO VVT
seest VVO VVT
showedest VVD VVDT
showedst VVD VVDT
showdest VVD VVDT
showest VVO VVT
showst VVO VVT
showst VVO VVT
startedest VVD VVDT
startest VVO VVT
startst VVO VVT
takest VVO VVT
takst VVO VVT
talkedest VVD VVDT
talkedst VVD VVDT

owst VVO VVT
scaldst VVO VVT
scoldst VVO VVT
scornst VVO VVT
scorndest VVD VVDT
seekst VVO VVT
sentst VVD VVDT
servst VVO VVT
setst VVO VVT VVD VVDT
settlest VVO VVT
shakst VVO VVT
shamst VVO VVT
shinst VVO VVT
shrugst VVO VVT