The Encyclopaedia of Shakespeare’s Language Project and Corpus Methods

Jonathan Culpeper,
Lancaster University, UK

@ShakespeareLang
The project’s rogues’ gallery

• http://wp.lancs.ac.uk/shakespearelang/people/
What the project aims to do ....

- Produce the first systematic and comprehensive (?) account of Shakespeare’s language using methods derived from corpus linguistics – an approach that uses computers in large-scale language analysis.
What will be in the encyclopedia?

**Volume 1** (a kind of dictionary)
Focuses on 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.
What will be in the encyclopedia?

**Volume 2** (a compendium of semantic patterns)
Focuses on patterns of words in Shakespeare's writings. It describes how these patterns create the 'linguistic thumbprints' of characters, different genders, themes, plays and dramatic genres. It also considers clusters of words that relate to concepts (e.g. love, death).

**Volume 3** (a kind of grammar)
Focuses on grammatical words and patterns.
Problem:
- Modern editions of Shakespeare are edited collations of the Folio and Quartos, mixed with a liberal dose of editorial license.
- Words are standardized to modern forms.
- Original morphology is (variously) stripped out.
- Even what counts as a word is variable, cf. compounds (e.g. hour glass).

Solution: Have as our base the First Folio with original spelling, and, specifically, the ‘diplomatic’ transcription (i.e. a faithful warts and all transcription) produced by Shakespeare Internet Editions (http://internetshakespeare.uvic.ca/Foyer/plays/).
Preliminary methodological issues

Spelling variation:

**Problem:** You decide to study the use of the word *would* in a corpus. You type it into your search program ... and look at the result.

But you miss: *wold, wolde, woolde, wuld, wulde, wud, wald, wvould, vvold*, etc., etc.

**Solution:** *Variant Detector* (VARD) program, primarily devised by generations of scholars at Lancaster, but most recently given a significant boost by Alistair Baron.
Preliminary methodological issues

**Further 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 (–(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 form that had EModE currency.

- Prioritize the most frequent spelling in Shakespeare

But: Very occasionally reader accessibility would have a bearing, e.g. Shakespeare *powr’st*; becomes *pourest* or *pour’st* in Arden; *pour’st* is not used in EModE (EBBO); we chose *pourest*. 
A glance at the First Folio and spelling variation in English (Baron et al’s 2009)
Preliminary methodological issues

The comparative corpus

**Problem:** Size matters

- Any pattern is a matter of frequency.
- Linguistics is centrally focussed on patterns in language.
- Historical linguistics work is often hampered by low frequencies, because the historical record is not complete.
- Corpus-based methods and concepts (e.g. collocates) are centrally driven by frequencies and statistical operations.

**Solution:** Various new corpora and electronic texts, but especially *Early English Books Online* (EEBO-TCP) – 1520-1679, and at least 723 million words.
Shakespeare and numbers: Neologisms and survivals

Myths about Shakespeare and the English language:

What can we ‘learn’ from the internet?

- Shakespeare coined more words than other writers, around 1700 words ...
- or is that 3,000 ...
- or did he invent half the words in the English language ...

N.B. The issues are twofold: neologisms and survivals
Shakespeare and numbers: Neologisms and survivals

Work on neologisms (with Sheryl Banas):
• 1,502 words recorded in the Oxford English Dictionary as first citations in Shakespeare
• We are checking these in EEBO-TCP

Preliminary findings:
• If the current pattern continues, less than a quarter of those 1,502 words can reasonably be attributed to Shakespeare.
Shakespeare and numbers: Neologisms and survivals

Issues

• How do we know that Shakespeare coined it as opposed to recorded it? Cf. *down staires* vs. *incarna[r]dine* (v.)

• What about borrowings, such as Latin *acerb[ic]*, that appear in mixed Latin-English texts before Shakespeare?

• Is it actually just a nonce word rather than neologism? Cf. *dropsied* vs. *domineering*

Do Shakespeare’s coinages survive into today’s English?

• Examples of phrases first recorded in Shakespeare and their more recent life.
Four phrases first recorded in Shakespeare and their use in printed material over the last 200 years (Google’s N-Gram Viewer)
Shakespearean ‘dictionaries’ and present-day corpus-based dictionaries

Some typical differences in approach:

• Words for inclusion: ‘hard’ words vs. all words in the corpus
• Word-meanings: etymological meanings and etymological organization vs. meanings based on usage in context and organised according to frequency

Note:
No Shakespearean dictionary has treated Shakespeare’s language as relative, i.e. put Shakespeare’s usage in the context of that of his contemporaries.
Case study: ‘horrid’ today

Examples from the BNC (random):
one day could take over from Morgan. A horrid man.
really glad to be on there to dispense with all those horrid people.
the horrid male instructor drills you as if you're in the Green Berets)
Smith being beaten by spotty, horrid little Nails tickled Nutty's imagination.
the tramp! He's horrid!" Shirley's cheeks had turned pale at the thought
will be giving the editor of New Scientist the full horrid details without delay.
recent research suggests that lead isn't as horrid in its effects as the

Top-40 rank-ordered most frequently occurring nouns within 5 words to the right of ‘horrid’ in the BNC:
things, man, thing, creature, stuff, truth, people, feeling, word, beast, phrase,
teeth, girls, flat, day, child, place, state, time, blighters, imprecations, defilement,
deodorants, cruelties, malady, apparitions, weasels, double-glazing, panoply,
sunflowers, bungling, separateness, puns, premonition, shrieks, jingle, hairstyle,
imaginations, blasphemy
Case study: ‘horrid’ (contd.)

Philological approach:

*Oxford English English Dictionary*

**horrid** (ˈhɔrɪd), a. (adv.) Also 7 **horred, horride**.

[ad. L. *horrid-us* bristling, rough, shaggy; rude, savage, unpolished; terrible, frightful, f. *horrere*: see horre v. Cf. It. *orrido*.]

A. adj.
1. Bristling, shaggy, rough. (Chiefly *poetic.*)

1590 Spenser *F.Q.* i. vii. 31 His haughtie Helmet, horrid all with gold.

1621 Burton *Anat. Mel.* i. ii. iii. xiv. (1651) 125 A rugged attire, hirsute head, horrid beard.
Case study: ‘horrid’ (contd.)

2. Causing horror or aversion; revolting to sight, hearing, or contemplation; terrible, dreadful, frightful; abominable, detestable.

   In earlier use nearly synonymous with *horrible*; in modern use somewhat less strong, and tending to pass into the weakened colloquial sense (3).


   [Shakespeare dictionaries concur with sense 2]

3. *colloq.* in weakened sense. Offensive, disagreeable, detested; very bad or objectionable.

   Noted in *N.E.D.* as especially frequent as a feminine term of strong aversion.

   1666 J. Davies *Hist. Caribby Isls* 281 Making horrid complaints that treated them ill.
Case study: ‘horrid’ in Shakespeare

Appeare in formes more horrid) yet my Duty, As doth a Rocke Vp Sword, and know thou a more horrid hent When he is drunke And cleaue the generall eare with horrid speech: Make mad the guilty heard and seene, Recounts most horrid sights seene by the Watch. shall breake his winde With feare and horrid flight. 1.Sen. Noble, To. I wil meditate the while vpon some horrid message for a Challenge. armes. Macd. Not in the Legions Of horrid Hell, can come a Diuell deformitie seemes not in the Fiend So horrid as in woman. all the sparkes of Nature To quit this horrid acte. Reg. Out treacherous Such sheets of Fire, such bursts of horrid Thunder, Such groanes of Curriors of the Ayre, Shall blow the horrid deed in euery eye, on is Of thy deere Husband. Then that horrid Act Of the diuorce, to themselves Beene deathes most horrid Agents, humaine grace I yeeld to that suggestion, Whose horrid Image doth vnfixe my Heire
Case study: ‘horrid’ in Shakespeare

The beginnings of a contextualised dictionary entry:

**Headword**: HORRID. Adj..

**Sense**: Something that is *horrid* causes fear; typically, it refers to supernatural or unnatural acts, sights and sounds. E.G. ‘Whose horrid Image doth vnfixe my Heire’ (Mac.)

**Contexts**: *Horrid* has a much closer association with Shakespeare's tragedies than either histories or comedies, and is used slightly more frequently by male characters than female. Shakespeare used it considerably more than his contemporary playwrights did. Generally, it is most characteristic of Early Modern plays and, perhaps surprisingly, scholarly literature.

**Distribution**: All = 16 (1.8); T = 10 (3.9), C = 2 (0.6), H = 4 (1.5); M = 14 (1.9), F = 2 (1.4).

**Comparisons**: Pla = 187 (0.17), Fic = 0, Tr = 0, Ha = 0, Sc = 1 (0.14).

• Frequency limitations
Case study (2): I

How was the 1\textsuperscript{st} person singular pronoun written?

- Always I
- But the 1\textsuperscript{st} person pronoun did not have a monopoly: it competed with the affirmative ay(e), e.g.

Ros. Did your brother tell you how I counterfeyted to sound, when he shew 'd me your handkercher?
Orl. I, and greater wonders then that. (AYL)

But it was dominant in the First Folio (1623):
20,293 instances of I (1\textsuperscript{st} pers. pronoun) vs. 302 instances of I (= aye) [(\textbackslash. | \textbackslash:) I (\textbackslash. | \textbackslash.\textbackslash.)]
Case study (2): I

Short digression on ay(e)/I:
The Oxford English Dictionary suggests the following under the entry for "aye | ay, int. (and adv.) and n."

Appears suddenly about 1575, and is exceedingly common about 1600; origin unknown. The suggestion that it is the same as AY adv. ‘ever, always,’ seems set aside by the fact that it was at first always written I, a spelling never found with AY adv.

• Not true: ay(e) dates back at least as far as 1584, well before the spelling I peaks about 1600.
• always > in all cases > by all means > certainly > yes
Case study (2): I

Shakespearean dictionaries:

- Words such as this typically omitted from Shakespearean dictionaries (e.g. Crystal and Crystal 2002; Onions 1986), presumably on the assumption that frequent and / or grammatical words:
  
  (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.
Top 25 collocates one to the right (Log-ratio):
am, thanke, prethee, warrant, protest, pray, humbly, prythee, beseech, hope, dare, saw, thinke, know, knew, could, owe, perceive, will, wil, meane, have, would, can, have, feele, told, doubt, have

“I am”: A case of I-dentity:
Were I the Moor I would not be Iago
In following him I follow but myself...
... I am not what I am. (Othello 1.1.57
Expressing personal states: am

Expressing thoughts and feelings: hope, dare, saw, thinke, know, knew, perceive, feele, doubt

Doing relational work: thanke, prethee, pray, humbly, prythee, beseech, owe, protest

Securing meaning: warrant, meane,

Narrative (speech presentation): told

Other: can, could, will/wil, would, have, had, would
## A glance at Vol.2: Character

### Desdemona:

<table>
<thead>
<tr>
<th>TOTAL</th>
<th>2753</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>132</td>
</tr>
<tr>
<td>my</td>
<td>79</td>
</tr>
<tr>
<td>and</td>
<td>61</td>
</tr>
<tr>
<td>you</td>
<td>60</td>
</tr>
<tr>
<td>to</td>
<td>57</td>
</tr>
<tr>
<td>not</td>
<td>48</td>
</tr>
<tr>
<td>me</td>
<td>47</td>
</tr>
<tr>
<td>do</td>
<td>44</td>
</tr>
<tr>
<td>the</td>
<td>41</td>
</tr>
<tr>
<td>him</td>
<td>41</td>
</tr>
<tr>
<td>lord</td>
<td>39</td>
</tr>
<tr>
<td>that</td>
<td>38</td>
</tr>
</tbody>
</table>
I and Desdemona

Desdemona’s keywords

<table>
<thead>
<tr>
<th></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
## Multi-word units

<table>
<thead>
<tr>
<th>Shakespeare</th>
<th>EModE Plays</th>
<th>Present-day Plays</th>
</tr>
</thead>
<tbody>
<tr>
<td>I pray you</td>
<td>it is a</td>
<td>I don’t know</td>
</tr>
<tr>
<td>I will not</td>
<td>what do you</td>
<td>what do you</td>
</tr>
<tr>
<td>I know not</td>
<td>and I will</td>
<td>I don’t want</td>
</tr>
<tr>
<td>I am a</td>
<td>it is not</td>
<td>I don’t want to</td>
</tr>
<tr>
<td>I am not</td>
<td>I have a</td>
<td>do you think</td>
</tr>
<tr>
<td>my good lord</td>
<td>I will not</td>
<td>do you want</td>
</tr>
<tr>
<td>there is no</td>
<td>in the world</td>
<td>I don’t think</td>
</tr>
<tr>
<td>I would not</td>
<td>I tell you</td>
<td>to do with</td>
</tr>
<tr>
<td>it is a</td>
<td>I know not</td>
<td>do you know</td>
</tr>
<tr>
<td>and I will</td>
<td>I warrant you</td>
<td>going to be</td>
</tr>
<tr>
<td></td>
<td></td>
<td>don’t want to</td>
</tr>
</tbody>
</table>

### Three-word lexical bundles in order of frequency (coloured items appear in another column)

Data in 2\(^{nd}\) and 3\(^{rd}\) columns draw from Culpeper and Kytö (2010)
Theatrical context: Stage and staging today
Frank: *What I want to know* is what is it that’s suddenly led you to this?
Rita: What? Comin’ here?
Frank: Yes.
Rita: It’s not sudden.
Frank: Ah.
Rita: I’ve been realizin’ for ages that I was, y’ know, slightly out of step. I’m twenty-six. I should have had a baby by now; everyone expects it. I’m sure me husband thinks I’m sterile. [...]
Theatrical context: EModE stage and staging

Purpose-built outdoor theatres:
The Theatre (1576),
The Curtain (1577),
The Rose (1587),
The Swan (1595),
The Globe (1599), and
The Fortune (1600).
Multi-word units

• A trend in the Early Modern data is for the lexical bundle to begin with a first person pronoun

• Especially notable trend for Shakespeare, where it combines with verbs relating to states, desires and knowledge. *I pray you* is most distinctive.

• Perhaps reflects a tendency for characters to present themselves (and others) relatively directly (including via soliloquies and asides).
The language of emotion in Shakespeare’s plays

+ Alison Findlay, Beth Cortese and Mike Thelwall

• “Sentiment analysis” and commercial goals
• What is it analysing? Emotion words, whether they are positive or negative (valence), and their strength.
• SentiStrength (Thelwall; http://sentistrength.wlv.ac.uk/)
• Lexicon adjusted for EModE and Shakespeare in particular.
• Checked against a human rater.
Overall negative sentiment across Shakespeare’s plays (average negative sentiment subtract average positive sentiment)
Concluding thoughts

A corpus approach to Shakespeare’s language means:

• All ‘words’ treated equally (e.g. not just ‘hard’ words).
• Meanings based on usage in context (e.g. not etymology, not narrowly-defined semantic meaning).
• The context includes linguistic aspects (e.g. collocations) and non-linguistic aspects (e.g. registers, social properties of the speaker/character).

A corpus/computational approach to literary texts means:

• Makes a kind of “distant reading” possible through the identification of linguistic patterns.
Problems and limitations

• The methodology is not (entirely) suitable for items below a certain frequency.

• Grammatical and semantic annotation need further development (manual correction), if they are to be deployed.

• It is never automatic – the human is needed to (1) devise/train the software, (2) select the data and prepare it; and (3) interpret the results.