Debunking myths about Shakespeare’s language with corpus methods

Jonathan Culpeper,
Lancaster University, UK

@ShakespeareLang
Acknowledgements

• Jane Demmen

• Sheryl Banas

• And the Encyclopaedia of Shakespeare’s Language project folks: http://wp.lancs.ac.uk/shakespearelang/people/
Coming up ...

- Setting the scene: Encyclopaedia of Shakespeare’s Language project
- What is a language myth?
- Myths about Shakespeare’s language:
  1. Shakespeare’s language is (wholly) Shakespeare’s language (Gary Taylor, authorship attribution and corpus linguistics)
  2. Shakespeare had a larger vocabulary than any other writer (Ward & Valenza, and Craig)
  3. Shakespeare coined more words than any other writer (with a discussion of word counting)
  4. Shakespeare’s language transcends time and space (with a discussion of the word ‘good’)
- A brief coda: Shakespeare -- the creative genius
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.
What is a myth about language?
What is a myth about language?

Language myths are:
Beliefs about language that are produced and reproduced within particular communities, and become part of a cultural ideology or “folk wisdom” that is used to evaluate language and account for how it is. They are inconsistent with empirically observed linguistic “facts”.

Myth 1: Shakespeare’s language is (wholly) Shakespeare’s language

- No clear authorial oversight. 36 plays were put together and published in 1623 as the First Folio – Shakespeare had been dead for 7 years.
Myth 1: Shakespeare’s language is (wholly) Shakespeare’s language
Myth 1: Shakespeare’s language is (wholly) Shakespeare’s language

- No clear authorial oversight. 36 plays were put together and published in 1623 as the First Folio – Shakespeare had been dead for 7 years.

- 18(?) plays had been previously published as Quartos, but some are considered “bad” (perhaps reconstructed from memory)

- An early play-text was a bundle of manuscript fragments written for performance, rather than a unitary whole written for publication
Myth 1: Shakespeare’s language is (wholly) Shakespeare’s language

• Collaboration amongst playwrights is known to have been very common.

• Plagiarism is a modern notion; re-using portions of text from elsewhere could be construed as complimentary.

• So, “Shakespeare’s language” = “surviving written texts that purport to represent, for the most part, the language that Shakespeare produced for his ‘literary’ works”
Myth 1: Shakespeare’s language is (wholly) Shakespeare’s language

- The contribution made by other authors to “Shakespeare’s” works? And vice versa.

- Authorship attribution.

- Gary Taylor & Gabriel Egan (2016). *The New Oxford Shakespeare*. Christopher Marlowe credited as co-author of *Henry VI* plays, Thomas Middleton as co-author of *All’s Well That Ends Well*; *Arden of Faversham* added to Shakespeare’s 'canon'.

- What was the basis for these decisions?

- N-grams.
Myth 1: Shakespeare’s language is (wholly) Shakespeare’s language

“I will finish this lecture shortly”

I will
will finish
finish this
this lecture
lecture shortly

I will finish
will finish this
finish this lecture
this lecture shortly
 Myth1: Shakespeare’s language is (wholly) Shakespeare’s language

- Does the disputed text have more n-gram types (i.e. different n-grams) in common with Shakespeare’s works or Marlowe’s?

- Does the disputed text favour the n-grams in the way that Shakespeare’s works do or Marlowe’s?

- Etc.

But such authorship attribution studies do not account for meaning.
# N-grams in Shakespeare and beyond

<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>do you think</td>
</tr>
<tr>
<td>I am not</td>
<td>I have a</td>
<td>do you want</td>
</tr>
<tr>
<td>my good lord</td>
<td>I will not</td>
<td>I don’t think</td>
</tr>
<tr>
<td>there is no</td>
<td>in the world</td>
<td>to do with</td>
</tr>
<tr>
<td>I would not</td>
<td>I tell you</td>
<td>do you know</td>
</tr>
<tr>
<td>it is a</td>
<td>I know not</td>
<td>going to be</td>
</tr>
<tr>
<td>and I will</td>
<td>I warrant you</td>
<td>don’t want to</td>
</tr>
</tbody>
</table>

Three-word N-grams in order of frequency (coloured items appear in another column)

Data in 2\textsuperscript{nd} and 3\textsuperscript{rd} columns draw from Culpeper and Kytö (2010)
N-grams in Shakespeare and beyond
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. [...]

Willy Russell, Educating Rita, 1981, p.8
N-grams in Shakespeare and beyond

Purpose-built outdoor theatres:
The Theatre (1576),
The Curtain (1577),
The Rose (1587),
The Swan (1595),
The Globe (1599), and
The Fortune (1600).
Myth 2: Shakespeare had a larger vocabulary than any other writer

An educated adult monolingual today: most studies approx.


- “[...] astonishing vocabulary of some 25,000 words” (Greenblatt, S. 2008, The Norton Shakespeare, p.65)

- “Twice as large as an educated person today ... 30,000” (McCrum, R. et al. (2002, The Story of English, p.102)
Myth 2: Shakespeare had a larger vocabulary than any other writer

- Vocabulary of 20,000 words “was a large vocabulary in its day” (Crystal, D., 2008, Think on my Words: Exploring Shakespeare’s Language, p.2)

- “Of Shakespeare it may be said without fear of exaggeration that his contribution to our phraseology is ten times greater than that of any writer to any language in the history of the world” (Weekley, E. 1952 [1928] The English Language, p. 55).
Myth 2: Shakespeare had a larger vocabulary than any other writer

Problem: Counting assumptions

<table>
<thead>
<tr>
<th>Author</th>
<th>Total different words (approx.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shakespeare</td>
<td>20,000</td>
</tr>
</tbody>
</table>
Myth 2: Shakespeare had a larger vocabulary than any other writer

Problem: Counting assumptions

<table>
<thead>
<tr>
<th>Author</th>
<th>Total different words (approx.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shakespeare</td>
<td>20,000</td>
</tr>
<tr>
<td>Jonson</td>
<td>18,500</td>
</tr>
<tr>
<td>Peele</td>
<td>6,000</td>
</tr>
<tr>
<td>Author</td>
<td>Segments</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
</tr>
<tr>
<td>Webster</td>
<td>3</td>
</tr>
<tr>
<td>Dekker</td>
<td>5</td>
</tr>
<tr>
<td>Peele</td>
<td>4</td>
</tr>
<tr>
<td>Marlowe</td>
<td>5</td>
</tr>
<tr>
<td>Jonson</td>
<td>12</td>
</tr>
<tr>
<td>Greene</td>
<td>4</td>
</tr>
<tr>
<td>Shakespeare</td>
<td>28</td>
</tr>
<tr>
<td>Lyly</td>
<td>8</td>
</tr>
<tr>
<td>Chapman</td>
<td>10</td>
</tr>
<tr>
<td>Heywood</td>
<td>5</td>
</tr>
<tr>
<td>Middleton</td>
<td>12</td>
</tr>
<tr>
<td>Fletcher</td>
<td>7</td>
</tr>
<tr>
<td>Wilson</td>
<td>3</td>
</tr>
</tbody>
</table>

1 Analysis is by author, arranged by average count, largest value first.
Myth 3: Shakespeare coined more words than any other writer

- How deep and apparently inexhaustible were the wells of his memory and invention, and how marvellous his aptitude for word-coining, are evident from the addition of 302 peculiar words in writing Troilus and Cressida (Alfred Hard, 1943, The Review of English Studies, 19(75): 254)
Myth 3: Shakespeare coined more words than any other writer (contd.)

What can we ‘learn’ from the internet?

- “Shakespeare coined more words than other writers, around 1700 words ...”
- “The English language owes a great debt to Shakespeare. He invented over 1700 of our common words ...”
- “Shakespeare introduced nearly 3,000 words ...”
- “Shakespeare is credited by the Oxford English Dictionary with the introduction of nearly 3,000 words into the language”
- “Shakespeare invented a quarter of our language”
- “Shakespeare invented half the words in the English language”
- “Shakespeare is our language”
Myth 3: Shakespeare coined more words than any other writer (contd.)

The problem with counting words

a) Defining a word  
b) Textual starting point  
c) Word-forms and lexemes  
d) Spelling variation
Myth 3: Shakespeare coined more words than any other writer (contd.)

Defining a word
a) A phoneme or phonemes bounded by pauses?

Tybalt: Gentlemen, **good den**, a word with one of you.  
*Romeo and Juliet, III.1*

b) A unit of meaning?

- *The plane landed* = 3 words?
- *The plane took off* = 3 words? (cf. phrasal verbs)
- *He kicked the bucket* = 2 words? (cf. idioms)

Polonius: **God buy you**; fare you well.  
*Hamlet II.1*
Myth 3: Shakespeare coined more words than any other writer (contd.)

*Goodbye* < god be with ye

*Good buy* (i.e. redeem, save) you

c) A minimal free form?

- Leonard Bloomfield (1887-1949): words = minimal free forms
- minimal free forms = the smallest units that can meaningfully stand on their own (i.e. be a complete utterance)
- *The, of, to, at, by, etc.?*

Present-day *gonna* < *going to* (BNC “gon-na”)

*Alarum* – French à l’arme; Italian all’arme

*a larum > alarum > larum*
Myth 3: Shakespeare coined more words than any other writer (contd.)

d) A letter or letters bounded by spaces?

*orthographic word* = ‘a string of uninterrupted non-punctuation characters with white space or punctuation at each end’ (Leech et al. 2001: 13-14)

A ‘natural’ division?
Myth 3: Shakespeare coined more words than any other writer (contd.)

To my dearest

I love you and I love you

Love from

Emily
Myth 3: Shakespeare coined more words than any other writer (contd.)

hourglass

hour-glass

hour glass
Myth 3: Shakespeare coined more words than any other writer (contd.)

**Shakespeare: The textual starting point**

**Problem:** Modern editions of Shakespeare are edited loose collations of the Folio and Quartos, mixed with a liberal dose of editorial licence.

**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/](http://internetshakespeare.uvic.ca/Foyer/plays/)).

**Problem:** A faithful transcription is more of an ideal than a reality.
Myth 3: Shakespeare coined more words than any other writer (contd.)

Some typical transcription errors

• $\text{y}^t$ ("that")

\[1. \text{Lo. G. Oh my sweet Lord y you wil stay behind vs.}\]

(AW, II:i)

• Space-saving formatting conventions

\[\text{In the vnpartiall judging of this Businessse. (ted Kine. Two equall men: The Queene shall be acquain-}\]
\[\text{Forthwith for what you come. Where's Gardiner? ted}}\]

(H8, II:ii)

<u norm="1 Lord" label="1. Lo. G"> Oh my sweet Lord CyC you , wil stay behind vs.</u>

<u norm="King Henry" label="Kin."> Two equall men: The Queene shall be acquainForthwith<lb type="inline"/> for what you come. Where's Gardiner? ted<lb/></u>
Myth 3: Shakespeare coined more words than any other writer (contd.)

Some typical transcription errors

- Mis-tagged text (1) 

  “And of all Christian Soules, I pray God. God buy ye.”

  Exeunt Ophelia


<lb/><lb/> And of all Christian Soules, I pray God. </u>

  <stage>God buy ye. Exeunt Ophelia</stage>

"God buy ye" is part of the soldier’s dialogue, not stage direction.
Myth 3: Shakespeare coined more words than any other writer (contd.)

Some typical transcription errors

• Mis-tagged text (2)

\[ \text{Duke. It is a rupture that you may easily heale: and the } \]
\[ \text{cure of it not onely saues your brother, but keepes you } \]
\[ \text{from dishonor in doing it.} \]

\[ \text{Isab. Shew me how (good Father.)} \]

\[ \text{Duk. This fore-named Maid hath yet in her the con-} \]

\[ <\text{u norm="Duke" label="Isab."}> \text{Shew me how (good Father.)</u}> \]

The speaker tag should be “Isabella”, as the Folio label indicates.
Myth 3: Shakespeare coined more words than any other writer (contd.)

Word-forms and lexemes

• Dictionary headword: 
  do

• Modern (morphological) word-forms: 
  do, does, doing, did, done

• Early modern (morphological) word-forms: 
  do, does, do(e)st, doth, doing, did, didst, done
Myth 3: Shakespeare coined more words than any other writer (contd.)

Word-forms and lexemes

Dictionary headword:

\[ do = 1 \]

Modern (morphological) word-forms:

\[ do, does, doing, did, done = 5 \]

Early modern (morphological) word-forms:

\[ do, does, do(e)st, doth, doing, did, didst, done = 8 \]
Myth 3: Shakespeare coined more words than any other writer (contd.)

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, vwould, vvold*, etc., etc.

One orthographic word today; many in EModE.

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

**Further problem**: What do you regularize the spelling to? There is no standardised regular form in the way that there is today.
Myth 3: Shakespeare coined more words than any other writer (contd.)

**Solution:** Our policy was to

- **Preserve the morphology,** e.g. $2^{nd}$ + $3^{rd}$ 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 spelling that had EModE currency.

- Prioritize the **most frequent spelling in Shakespeare**
Myth 3: Shakespeare coined more words than any other writer (contd.)

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 *Early English Books Online* (EEBO-TCP): 125,000 titles of printed material, 1473-1700. About 1.2 billion words.

Preliminary findings:
• If the current pattern continues, less than a quarter of those 1,502 words can reasonably be attributed to Shakespeare.
Myth 3: Shakespeare coined more words than any other writer (contd.)

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*
Myth 3: Shakespeare coined more words than any other writer (contd.)

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

• A glimpse 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)
Myth 4: Shakespeare’s language transcends time and space
Myth 4: Shakespeare’s language transcends time and space

- Universal characters, emotions, themes .... language??

“Shakespeare has given us a universal language medium in which are crystalized the battle hymns, the intellectual conceptions and the spiritual aspirations of the Anglo-Saxons.”

(Rutherford, N.J. and Bennett, E.H., 1918-1922, English Speaking World, Vol.2 (8): 14)

- No Shakespearean dictionary has treated Shakespeare’s language as relative, i.e. put Shakespeare’s usage in the context of that of his contemporaries.
Myth 4: Shakespeare’s language transcends time and space - good


(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.

+ phrases and compounds
Myth 4: Shakespeare’s language transcends time and space - **good**


1. [intensifying use] real, genuine (‘love no man in good earnest’).
2. **kind**, benevolent, generous.
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/troth', 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.
A brief coda: Shakespeare -- the creative genius

Antony: Let Rome in Tiber melt, [...]  

(Antony and Cleopatra I.i)
Concluding reflections

Shakespeare and the myths. Why?

“Shakespeare has given us a universal language medium in which are crystalized the battle hymns, the intellectual conceptions and the spiritual aspirations of the Anglo-Saxons.” (Rutherford, N.J. and Bennett, E.H., 1918-1922, English Speaking World, Vol.2 (8): 14)
Concluding reflections

Beyond the myths:
• Texts and their production (including transcription)
• Spelling variation
• Words and lexemes
• N-grams and collocations
• Counting assumptions
• The nature of corpus linguistics
• Corpus linguistics and historical linguistics
• Corpus linguistics and lexicography