Debunking myths about Shakespeare’s language with corpus methods

Jonathan Culpeper,
Lancaster University, UK

@ShakespeareLang
Acknowledgements

- The Encyclopaedia of Shakespeare’s Language project folks: http://wp.lancs.ac.uk/shakespearelang/people/
• 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 digression on word counting)
  4. Shakespeare’s language transcends time and space (with a brief discussion of the word ‘good’)
What will the project produce?

The key output: The encyclopedia

Two volumes:
(1) a kind of dictionary, and
(2) a compendium of word patterns relating to characters, character groups, plays, play-genres, themes, etc..

To be published by Bloomsbury in paper and electronically.
What is innovative about it?

• It will be the first systematic and comprehensive account of Shakespeare’s language using methods derived from corpus linguistics.

• It will be comparative, looking at use in Shakespeare *and* his contemporaries.
Why now?

• Corpus-based techniques have developed sufficiently to handle non-standardised spellings, provide evidenced accounts of meanings, contribute to the description of style, etc.

• Comparative historical resources have developed considerably, notably, *Early English Books Online* (EEBO-TCP), amounting to some 1.2 billion words.
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 observed linguistic “facts”.
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.
- 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.
Myth 1: Shakespeare’s language is (wholly) Shakespeare’s language

So, Shakespeare’s language =

surviving written texts that purport to represent, for the most part, the language that Shakespeare produced for his ‘literary’ works
Myth1: Shakespeare’s language is (wholly) Shakespeare’s language

But what exactly was the contribution made by other authors to “Shakespeare’s works”? And vice versa.

• Authorship attribution (cf. forensic linguistics).


• 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
Myth 1: 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 2nd and 3rd columns draw from Culpeper and Kytö (2010)
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

The vocabulary size of an educated adult monolingual today:

Most studies approx **9,000 – 18,000** words

Myth 2: Shakespeare had a larger vocabulary than any other writer

Shakespeare:

• “[...] astonishing vocabulary of some \textbf{25,000 words}” (Greenblatt, S. 2008, \textit{The Norton Shakespeare}, p.65)

• “Twice as large as an educated person today ... \textbf{30,000}” (McCrum, R. et al. 2002, \textit{The Story of English}, p.102)

• Vocabulary of \textbf{20,000} words “\textit{was} a large vocabulary in its day” (Crystal, D., 2008, \textit{Think on my Words: Exploring Shakespeare’s Language}, p.2)
Myth 2: Shakespeare had a larger vocabulary than any other writer

Problems:

1. What counts as a word?

2. What assumptions are made in calculating vocabulary size?
Myth 2: Shakespeare had a larger vocabulary than any other writer

Problems: Counting assumptions

<table>
<thead>
<tr>
<th>Author</th>
<th>Total different words (approx.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shakespeare</td>
<td>21,000</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></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.

<table>
<thead>
<tr>
<th>Author</th>
<th>Segments</th>
<th>No. of different words</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lowest</td>
<td>Highest</td>
<td>Average</td>
</tr>
<tr>
<td>Webster</td>
<td>3</td>
<td>1,694</td>
<td>1,957</td>
<td>1,827.0</td>
</tr>
<tr>
<td>Dekker</td>
<td>5</td>
<td>1,502</td>
<td>2,011</td>
<td>1,772.0</td>
</tr>
<tr>
<td>Peele</td>
<td>4</td>
<td>1,649</td>
<td>1,845</td>
<td>1,749.5</td>
</tr>
<tr>
<td>Marlowe</td>
<td>5</td>
<td>1,554</td>
<td>1,924</td>
<td>1,740.8</td>
</tr>
<tr>
<td>Jonson</td>
<td>12</td>
<td>1,494</td>
<td>1,964</td>
<td>1,727.1</td>
</tr>
<tr>
<td>Greene</td>
<td>4</td>
<td>1,512</td>
<td>1,829</td>
<td>1,675.5</td>
</tr>
<tr>
<td>Shakespeare</td>
<td>28</td>
<td>1,426</td>
<td>1,930</td>
<td>1,663.5</td>
</tr>
<tr>
<td>Lyly</td>
<td>8</td>
<td>1,497</td>
<td>1,688</td>
<td>1,604.5</td>
</tr>
<tr>
<td>Chapman</td>
<td>10</td>
<td>1,281</td>
<td>1,783</td>
<td>1,581.3</td>
</tr>
<tr>
<td>Heywood</td>
<td>5</td>
<td>1,389</td>
<td>1,769</td>
<td>1,554.6</td>
</tr>
<tr>
<td>Middleton</td>
<td>12</td>
<td>1,308</td>
<td>1,655</td>
<td>1,550.8</td>
</tr>
<tr>
<td>Fletcher</td>
<td>7</td>
<td>1,394</td>
<td>1,682</td>
<td>1,525.5</td>
</tr>
<tr>
<td>Wilson</td>
<td>3</td>
<td>1,333</td>
<td>1,710</td>
<td>1,499.0</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

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.)

Academic views:

• 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).

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

Work on neologisms (with Shery Banas and Poppy Plumb):

- 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)
A digression on the problem of counting words

a) Defining a word
b) Word-forms and lexemes
c) Spelling variation
A digression on the problem of counting words (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 semantic 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*
A digression on the problem of counting words (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 larpum > alarum > larum*
A digression on the problem of counting words (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?
A digression on the problem of counting words (contd.)

tomypaddob

I like you and I love you

Love from

Emily
A digression on the problem of counting words (contd.)

hourglass

hour-glass

hour glass
A digression on the problem of counting words (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
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
A digression on the problem of counting words (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, vvould, 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.
Solution: Our policy was to

1) Prioritise the orthographic definition of words. But deploy the semantic definition in addition where necessary (e.g. some compounds).

2) Distinguish lexemes (lemmatisation). We generally preserved the morphology, e.g. 2nd + 3rd 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).

3) Regularise spellings. Only use a spelling that had EModE currency. Prioritize the most frequent spelling in Shakespeare.
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, hiborn, 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.
(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.
Encyclopaedia of Shakespeare's Language

Definition preview: definition 104 for entry good_ADJ

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.
Does Shakespeare’s usage reflect that of his contemporaries? A glance at collocates in EEBO 1560-1640

<table>
<thead>
<tr>
<th><strong>Good</strong> (= 577,864)</th>
<th><strong>Bad</strong> (= 28,137)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Works</td>
<td>Bad</td>
</tr>
<tr>
<td>Cheer</td>
<td>Witch</td>
</tr>
<tr>
<td>Evil</td>
<td>Adieu</td>
</tr>
<tr>
<td>Advisement</td>
<td>Worse</td>
</tr>
<tr>
<td>Hap</td>
<td>Good</td>
</tr>
<tr>
<td>Deeds</td>
<td>Dealing</td>
</tr>
<tr>
<td>Sooth</td>
<td>Success</td>
</tr>
<tr>
<td>Wallace</td>
<td>Humours</td>
</tr>
<tr>
<td>Readers</td>
<td>Usage</td>
</tr>
<tr>
<td>Christian</td>
<td>News</td>
</tr>
</tbody>
</table>
Does Shakespeare’s usage reflect that of his contemporaries? A glance at collocates in EEBO 1560-1640

<table>
<thead>
<tr>
<th>Good (= 577,864)</th>
<th>Bad (= 28,137)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Works</strong></td>
<td>Bad</td>
</tr>
<tr>
<td>Cheer</td>
<td>Witch</td>
</tr>
<tr>
<td><strong>Evil</strong></td>
<td>Adieu</td>
</tr>
<tr>
<td>Advisement</td>
<td>Worse</td>
</tr>
<tr>
<td>Hap</td>
<td><strong>Good</strong></td>
</tr>
<tr>
<td><strong>Deeds</strong></td>
<td>Dealing</td>
</tr>
<tr>
<td>Sooth</td>
<td>Success</td>
</tr>
<tr>
<td>Wallace</td>
<td>Humours</td>
</tr>
<tr>
<td><strong>Readers</strong></td>
<td>Usage</td>
</tr>
<tr>
<td><strong>Christian</strong></td>
<td>News</td>
</tr>
</tbody>
</table>
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

Aside from the myths:

• The notion of a word
• Spelling variation
• Words and lexemes
• N-grams
• Counting assumptions
• Meanings and collocations
• Corpus linguistics and lexicography
• Corpus linguistics and historical linguistics
• Etc.