Part-of-Speech Tagging in Shakespeare: Trials, Tribulations and why one might Bother

Jonathan Culpeper, Jane Demmen and Andrew Hardie

@ShakespeareLang
What is POS tagging?

Grammar – not necessarily a sexy topic:

Jack Cade: "It will be proved to thy face that thou hast men about thee that usually talk of a noun and a verb, and such abominable words as no Christian ear can endure to hear." 2H6 IV.vii

Part-of-speech tagging involves adding codes to words which indicate the grammatical category (noun, verb, adverb, adjective, etc.) to which they are deemed to belong.

To_TO be_VBI or_CC not_XX to_TO be_VBI that_DD1 is_VBZ the_AT question_NN1

The tagging process can be automated by programs such as CLAWS. There is no such thing as 100% accuracy; the aim is to be good enough.
(1) It reveals patterns of grammatical usage, and this can help us:

- Write descriptive grammar books
- Teach English grammar
- Write dictionaries
- Study variation in English, e.g. dialects
- Study change in English (i.e. the history of the English language)
- Study the ‘style’ of a particular period or author
Why one might bother

(2) It improves the accuracy of other automated processes (e.g. lemmatization, semantic tagging)

What is lemmatization?
A process that involves grouping word variants or word-forms into ‘lemmas’ (which are like dictionary headwords).

- Dictionary headword/lemma: \(do = 1\)
- Modern (morphological) word-forms: \(do, does, doing, did, done = 5\)
- Early modern (morphological) word-forms: \(do, does, do(e)st, doth, doing, did, didst, done = 8\)
Why one might bother

- POS-tagging improves lemmatization prospects, typically by disambiguating words, e.g. homonyms such as *leaves* (in Shakespeare two-thirds of cases are a noun, one third a verb)

Semantic tagging

Semantic tagging involves adding codes to words which indicate the semantic/meaning category (e.g. people, time, power, being, food, thought, colour, liking, avarice, relationship) to which they are deemed to belong. Cf. *WMatrix*
Why one might bother

(3) Some of the specific POS tags may be of interest, e.g.

- NP1/2 – Singular/plural proper noun (e.g. London, Jane, Frederick)
- NNB – Preceding noun of title (e.g. Mr., Prof., Lord)
- FW – Foreign word
Trials and tribulations

• Grammatical phenomena which are marginal today, but may require addressing by the tagger for EModE
  – 2\textsuperscript{nd} person singular
• Words not in the modern lexicon
• Words whose possible classifications have changed
  – \textit{what}
• Words whose probability profile has changed
  – \textit{prostitute}
• Extra cliticisations (\textit{me}=thinks, \textit{me}=thought)

• But we don’t have the time for a radical system overhaul
Solution

(1) Development work: Patch CLAWS three ways:

- Resource patching
- Input patching
- Output patching

(2) Manual post-editing (if necessary)
Implementation of development work

- Input patching script
- Wordlist patches
- Idiomlist patches
- Output patching script
- Untagged text
- CLAWS input
- CLAWS
- CLAWS output
- Tagged text
| 1  | prithee UH                          | 200 | affliction NN1                             |
| 2  | hark VVO                            | 201 | anointed VVN JJ VVD@                       |
| 3  | claudio NP1                         | 202 | bondage NN1                                |
| 4  | signior NN1                         | 203 | bootless JJ                                |
| 5  | timon NP1                           | 204 | corioles NP1                               |
| 6  | valour NN1                          | 205 | costard NP1                                |
| 7  | falstaff NP1                        | 206 | coxcomb NN1                               |
| 8  | troilus NP1                         | 207 | infect VVO                                |
| 9  | jove NP1                            | 208 | ingratitude NN1                            |
|10  | martius NP1                         | 209 | leonatus NP1                               |
|11  | alarum NN1                          | 210 | mantua NP1                                 |
|12  | angelo NP1                          | 211 | messengers NN2                             |
|13  | cassius NP1                         | 212 | mischance NN1                              |
|14  | bardolph NP1                        | 213 | polixenes NP1                              |
|15  | palamon NP1                         | 214 | reignier NP1                               |
|16  | alack UH                            | 215 | rogues NN2                                 |
|17  | iago NP1                            | 216 | shylock NP1                                |
|18  | arcite NP1                          | 217 | stephano NP1                               |
|19  | bolingbroke NP1                     | 218 | trinculo NP1                               |
|20  | proteus NP1                         | 219 | tush UH                                   |
|21  | silvia NP1                          | 220 | usurp VVO                                 |
|22  | griefs NN2                          | 221 | alcibiades NP1                             |
|23  | benedick NP1                        | 222 | bachelor NN1                               |
|24  | oaths NN2                           | 223 | barbarous JJ                               |
|25  | caius NP1                           | 224 | barnardine NP1                             |
|26  | emilia NP1                          | 225 | blushing JJ VVG NN1®                       |
|27  | bianca NP1                          | 226 | commends VVZ                              |
|28  | demetrius NP1                       | 227 | cymbeline NP1                              |
|29  | warlike JJ                          | 1247 | embassage NN1                             |
|30  |                              | 1248 | endowments NN2                             |
|31  |                              | 1249 | enfranchised VVD VVN JJ                   |
|32  |                              | 1250 | esteems VVZ                                |
|33  |                              | 1251 | expostulate VVO                            |
|34  |                              | 1252 | extenuate VVO                             |
|35  |                              | 1253 | fawning VVG NN1 JJ                        |
|36  |                              | 1254 | fines VVZ NN2                             |
|37  |                              | 1255 | firmament NN1                              |
|38  |                              | 1256 | flatteries NN2                             |
|39  |                              | 1257 | flax NN1                                  |
|40  |                              | 1258 | flayed VVD VVN JJ                         |
|41  |                              | 1259 | foison NN1                                |
|42  |                              | 1260 | forrestell VVO                            |
|43  |                              | 1261 | forfeited VVD VVN JJ                      |
|44  |                              | 1262 | forgetfulness NN1                         |
|45  |                              | 1263 | forwardness NN1                           |
|46  |                              | 1264 | fulsome JJ                                |
|47  |                              | 1265 | furred JJ                                 |
|48  |                              | 1266 | gash VVO NN1                              |
|49  |                              | 1267 | gashes VVZ NN2                            |
|50  |                              | 1268 | gentlemanlike JJ                          |
|51  |                              | 1269 | gibes NN2                                 |
|52  |                              | 1270 | goblins NN2                               |
|53  |                              | 1271 | goth NN1                                  |
|54  |                              | 1272 | graceless JJ                              |
|55  |                              | 1273 | grafted VVD VVN JJ                       |
|56  |                              | 1274 | groats NN2                                |
|57  |                              | 1275 | grosser JJR                               |
|58  |                                | 2623 | unbowed JJI                               |
|59  |                                | 2624 | unbucket VVO                              |
|60  |                                | 2625 | uncapable JJ                              |
|61  |                                | 2626 | unconquered JJ                            |
|62  |                                | 2627 | uncouple VVO                              |
|63  |                                | 2628 | uncrown VVO                               |
|64  |                                | 2629 | undeserver NN1                            |
|65  |                                | 2630 | undrowned JJ                              |
|66  |                                | 2631 | unforced JJ                               |
|67  |                                | 2632 | unguided JJ                               |
|68  |                                | 2633 | unheedful JJ                              |
|69  |                                | 2634 | unkennel VVO                              |
|70  |                                | 2635 | unmake VVO                                |
|71  |                                | 2636 | unmask VVO                                |
|72  |                                | 2637 | unmingled JJ                              |
|73  |                                | 2638 | unpractised JJ                            |
|74  |                                | 2639 | unpregnant JJ                             |
|75  |                                | 2640 | unprizable JJ                             |
|76  |                                | 2641 | unquestioned JJ                           |
|77  |                                | 2642 | unquietness NN1                           |
|78  |                                | 2643 | unsubstantial JJ                         |
|79  |                                | 2644 | untender JJ                               |
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;
```
Output patching

/pronouns
thou PPY PPYS1
thee PPY PPYS1
th PPY PPYS1
th' PPY PPYS1

/verbs
art VBR VBT
beest VBR VBT
wast VBDR VBDT
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
showedest VVD VVDT
showedest VVD VVDT
showedest VVD VVDT
showest VVO VVT
showest VVO VVT
showest VVO VVT
showest VVO VVT
startedest VVD VVDT
startest VVO VVT
startst VVO VVT
startst VVO VVT
takest VVO VVT
takst VVO VVT
takedest VVD VVDT
takedest VVD VVDT

owst VVO VVT
scaldst VVO VVT
scoldst VVO VVT
scornst VVO VVT
scorndst 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
Manual Post-editing

35293 0001365 181  , 03 .
35294 0001365 182  ---------------------------------------------
35295 0001365 190  >  What 93 DDQ
35296 0001365 191  <  's 96 VBZ
35297 0001365 192  <1b/>  ERROR?  01
35298 0001365 200  the 93 AT
35299  >  
35300 0001365 210  News 93 NN1
35301  >  
35302 0001365 220  in 93 [II/100] RP0/0
35303 0001365 230  Rome 93 NP1
35304 0001365 231  : 03 :
35305 0001365 240  I 93 [PPIS1/100] MC1½/0 ZZ1½/0
35306  >  
35307 0001365 250  have 93 VHU
35308  >  
35309 0001365 260  a 93 AT1
35310 0001365 270  Note 93 [NN1/100] VVO/0
35311 0001365 280  from 93 II
35312 0001365 290  the 93 AT
35313  >  
35314 0001365 300  Volscian 93 JJ
35315  >  
35316 0001365 302  <1b/>  ERROR?  01
35317 0001365 310  state 93 [NN1/95] VVO/5
to 97 TO
35318 0001365 320  >  
35319 0001365 330  find 97 VVI
35320  >  
35321 0001365 340  you 93 PPY
35322 0001365 350  out 93 [RF/100] II½/0
35323 0001365 360  there 93 [RL/100] EX/0
35324 0001365 361  . 03 .
35325 0001365 362  ---------------------------------------------
35326 0001365 362  You 93 PPY
35327 0001365 370
Methods for manual post-editing

1. Search on the word itself and check, e.g. *what*
   “And *what* was he?” = determiner
   “*What* ho Brabantio” = interjection

2. Search on tags which we anticipate may need fixing, e.g. tag ZZ1 (singular letter of the alphabet) wrongly applied to pronoun *I*.

3. Use regular expressions, e.g. to find and check words ending in –ing (disambiguating nouns, verbs and adjectives)
   \S\S\Sing \d\d

4. Check by reading ...
Recurrent issues (1)

• Familiar words in unfamiliar grammatical roles
  – e.g. *marry* can be an interjection as well as a verb
    “Marry [VV0/86] UH/14 for justice she is so employed” (Tit 4:3)
  – *fright* can be a verb as well as a noun
    “What shall they seek the Lion in his den, And fright him there?” (KJ)

• Unfamiliar phrases
  – e.g. *go to* (*to* is an adverb not a preposition here)
    "Go to, you're a dry fool:" (TN 1:5)

• Some names of characters not recognised as proper nouns
  – e.g. [Mistress] *Quickly* (not an adverb), [Lord] *Say* (not a verb)
Recurrent issues (2)

• Wordplay sometimes confuses CLAWS

  – “A mark, O mark but that mark: a mark says my Lady” (LLL 4_1)
    
    CLAWS tags all four cases as nouns; the second is a verb

  – “Marry this is Miching Mallico, that means Mischief” (Ham 3_2)
    
    marry wrongly tagged as verb
    miching mallico tagged as verb, proper noun, but more likely to be adjective, noun (meaning unclear)