Encyclopaedia of Shakespeare's Language

Lancaster University

Part-of-speech tagging in Shakespeare: Trials, tribulations and preliminary results

Jane Demmen, Andrew Hardie and Jonathan Culpeper, Lancaster University, UK

AHRC project AH/N002415/1

@ShakespeareLang

http://wp.lancs.ac.uk/shakespearelang/

Arts & Humanities Research Council

CASS Corpus Approaches to Social Science

The Queen's Anniversary Prizes For Higher and Further Education 2015

UCREL

Encylopaedia of Shakespeare's Language: Corpus-based & comparative

Lancaster University

- Vol 1 = a dictionary
- Vol 2 = a compendium of themes, character profiles, play profiles, etc.
- Corpora:
  - Shakespeare corpus: 38 plays (c.1589-1613)
  - Comparative corpus: 46 plays by 24 other playwrights (1584-1626)
    (both just over 1 million words)
  - Wider EModE written text: 5697 texts (approx. 269 million words) from EEBO-TCP (1560-1639)

What is part-of-speech (POS) tagging?

Lancaster University

- POS tagging = adding codes to words which indicate the grammatical category 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

- This can be automated by programs such as CLAWS (Constituent Likelihood Automatic Word-tagging System).

  But there's no such thing as 100% accuracy.

Why would we bother?

Lancaster University

(1) It reveals patterns of grammatical usage, and this can help us:

- Write descriptive grammar books; teach English grammar
- Write dictionaries
- Study variation/change in English language, e.g. dialects
- Study the 'style' of a particular period or author

(2) Disambiguation, e.g. *leaves* (plural noun or present tense verb), which also improves the accuracy of other automated processes (e.g. lemmatization, semantic tagging)

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

- NNB – Preceding noun of title (e.g. Mr., Lady); FW – Foreign word

Trials and tribulations

- Grammatical phenomena which are marginal today, but may require addressing by the tagger for EModE
  - 2nd person singular
- Words not in the modern lexicon
- Words whose probability profile has changed
  - *prostitute*
- Extra cliticisations (*me=thinks*, *me=thought*)

- But we don't have the time for a radical system overhaul, so …

Solution

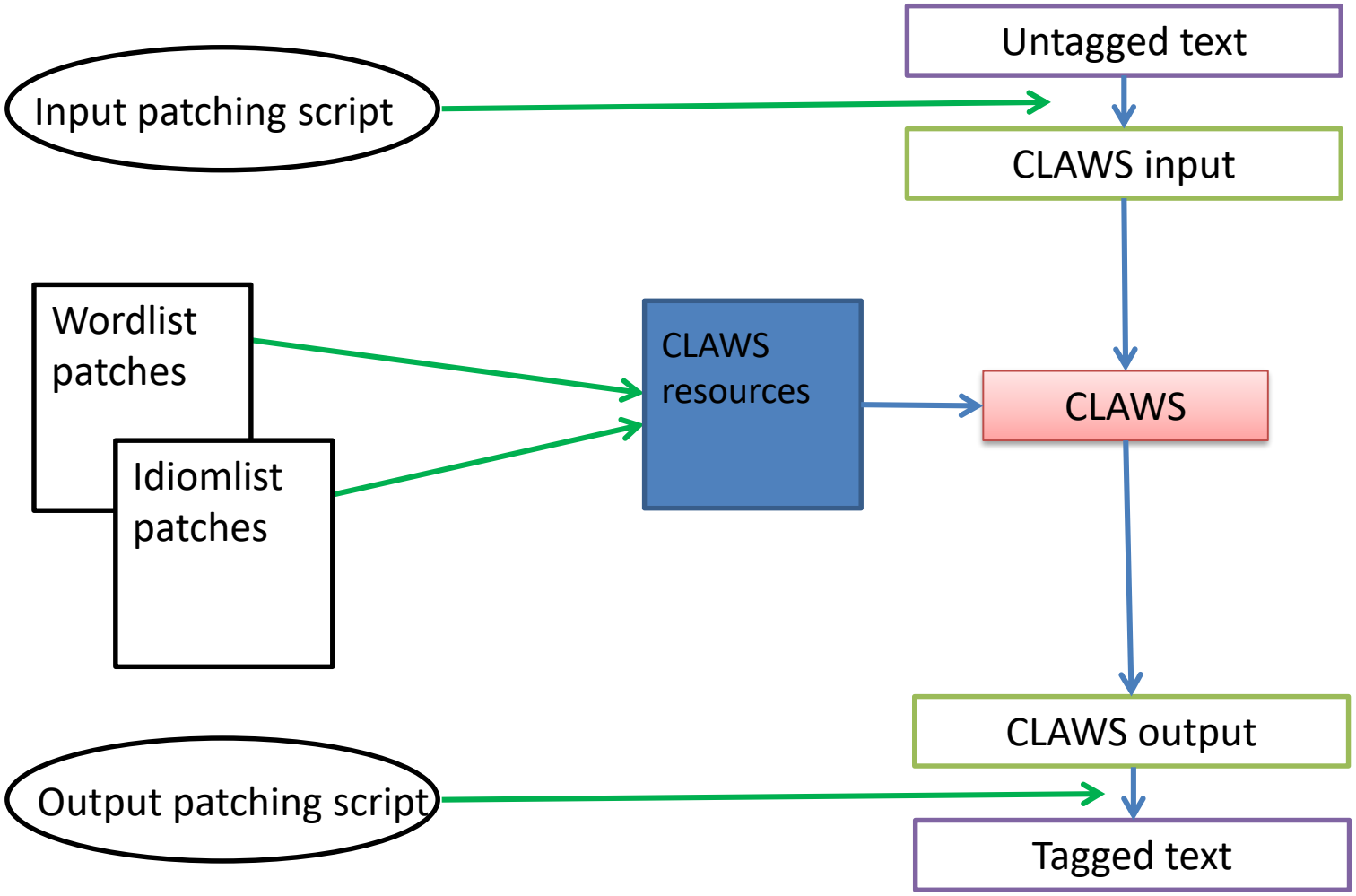
(1) Development work: Patch CLAWS three ways:

- Resource patching
- Input patching
- Output patching

(2) Manual post-editing (Shakespeare corpus only)

… all of which took about a year!

Implementation of development work



Resource patching: lexicon

```
   1 prithee UH
   2 hark VV0
   3 claudio NP1
   4 signior NN1
   5 timon NP1
   6 valour NN1
   7 falstaff NP1
   8 troilus NP1
   9 jove NP1
  10 martius NP1
  11 alarum NN1
  12 angelo NP1
  13 cassius NP1
  14 bardolph NP1
  15 palamon NP1
  16 alack UH
  17 iago NP1
  18 arcite NP1
  19 bolingbroke NP1
  20 proteus NP1
  21 silvia NP1
  22 griefs NN2
  23 benedick NP1
  24 oaths NN2
  25 caius NP1
  26 emilia NP1
  27 bianca NP1
  28 demetrius NP1
  29 warlike JJ
```

```
 200 affliction NN1
 201 anointed VVN JJ VVD@
 202 bondage NN1
 203 bootless JJ
 204 corioles NP1
 205 costard NP1
 206 coxcomb NN1
 207 infect VV0
 208 ingratitude NN1
 209 leonatus NP1
 210 mantua NP1
 211 messengers NN2
 212 mischance NN1
 213 polixenes NP1
 214 reignier NP1
 215 rogues NN2
 216 shylock NP1
 217 stephano NP1
 218 trinculo NP1
 219 tush UH
 220 usurp VV0
 221 alcibiades NP1
 222 bachelor NN1
 223 barbarous JJ
 224 barnardine NP1
 225 blushing JJ VVG NN1@
 226 commends VVZ
 227 cymbeline NP1
```

```
1247 embassage NN1
1248 endowments NN2
1249 enfranchised VVD VVN J
1250 esteems VVZ
1251 expostulate VV0
1252 extenuate VV0
1253 fawning VVG NN1 JJ
1254 fines VVZ NN2
1255 firmament NN1
1256 flatteries NN2
1257 flax NN1
1258 flayed VVD VVN JJ
1259 foison NN1
1260 foretell VV0
1261 forfeited VVD VVN JJ
1262 forgetfulness NN1
1263 forwardness NN1
1264 fulsome JJ
1265 furred JJ
1266 gash VV0 NN1
1267 gashes VVZ NN2
1268 gentlemanlike JJ
1269 gibes NN2
1270 goblins NN2
1271 goth NN1
1272 graceless JJ
1273 grafted VVD VVN JJ
1274 groats NN2
1275 grosser JJR
```

```
2623 unbowed JJ
2624 unbuckle VV0
2625 uncapable JJ
2626 unconquered JJ
2627 uncouple VV0
2628 uncrown VV0
2629 undeserver NN1
2630 undrowned JJ
2631 unforced JJ
2632 unguided JJ
2633 unheedful JJ
2634 unkennel VV0
2635 unmake VV0
2636 unmask VV0
2637 unmingled JJ
2638 unpractised JJ
2639 unpregnant JJ
2640 unprizable JJ
2641 unquestioned JJ
2642 unquietness NN1
2643 unsubstantial JJ
2644 untender JJ
```

EModE tags added (via output patching)

Tag	Words
PPYS1	thou, th'
PPYO1	thee
VBT	art, beest
VBDT	wast, wert
VDT	dost, doest
VDDT	didst
VHT	hast
VHDT	hadst
VMT	wilt, wouldst, canst, couldst, shalt, shouldst etc.
VMTK	oughtest, usedest
VVT	givest, workest
VVDT	gavest, workedst

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 <lb/>                        ERROR?        01
35298  0001365 200                                        the 93 AT
35299  >
35300  0001365 210                                       News 93 NN1
35301  >
35302  0001365 220                                         in 93 [II/100] RP@/0
35303  0001365 230                                       Rome 93 NP1
35304  0001365 231                                          : 03 :
35305  0001365 240                                          I 93 [PPIS1/100] MC1%/0 ZZ1%/0
35306  >
35307  0001365 250                                       have 93 VHO
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 <lb/>                        ERROR?        01
35317  0001365 310                                     state 93 [NN1/95] VVO/5
35318  0001365 320                                        to 97 TO
35319  >
35320  0001365 330                                      find 97 VVI
35321  >
35322  0001365 340                                       you 93 PPY
35323  0001365 350                                       out 93 [RP/100] II%/0
35324  0001365 360                                     there 93 [RL/100] EX/0
35325  0001365 361                                          . 03 .
35326  0001365 362 ----------------------------------------------------
35327  0001365 370                                       You 93 PPY
```

Results: main POS categories (Shakespeare & other playwrights)

Shakespeare corpus		
	Raw freq	PMW
VERB	187583	180627
NOUN	177060	170494
PRON	147273	141812
PREP	81808	78774
ADJ	79383	76439
ADV	67588	65082
CONJ	59445	57241
ART	47573	45809
INTERJ	8179	7876

Comparative corpus		
	Raw freq	PMW
NOUN	201987	185016
VERB	199251	182510
PRON	150763	138096
ADJ	86171	78931
PREP	85683	78484
ADV	72483	66393
CONJ	60534	55448
ART	48882	44775
INTERJ	7624	6983

PMW = per million words

More nouns than verbs

	Genre	Verbs PMW	Nouns PMW
1H4	H	168450	176860
1H6	H	176475	190354
2H4	H	167245	179879
2H6	H	177551	192276
3H6	H	173545	188757
H5	H	160365	183862
H8	H	171075	175172
KJ	H	170058	189236
LLL	C	171609	181892
MND	C	175693	177631
R2	H	169881	189394
R3	H	169600	180051
TC	T	174641	180474
Tit	T	178775	184470

10 histories; 2 comedies; 2 tragedies

More verbs than nouns

	Genre	Verbs PMW	Nouns PMW
AC	T	195033	169105
AW	C	190424	157568
AYL	C	183209	154658
CE	C	184341	161767
Cor	T	190736	157731
Cym	C	184866	160206
Ham	T	172262	168493
JC	T	191996	167700
KL	T	184425	166404
MA	C	192675	158473
Mac	T	185908	169571
MM	C	185785	159629
MV	C	188310	169519
MW	C	182500	175262
Oth	T	185992	150147
Per	C	195035	166674
RJ	T	184937	172148
Tem	C	182306	163119
TGV	C	186168	148407
Tim	T	184185	166459
TN	C	185537	163234
TNK	C	186099	160378
TS	C	186750	164615
WT	C	185071	149455

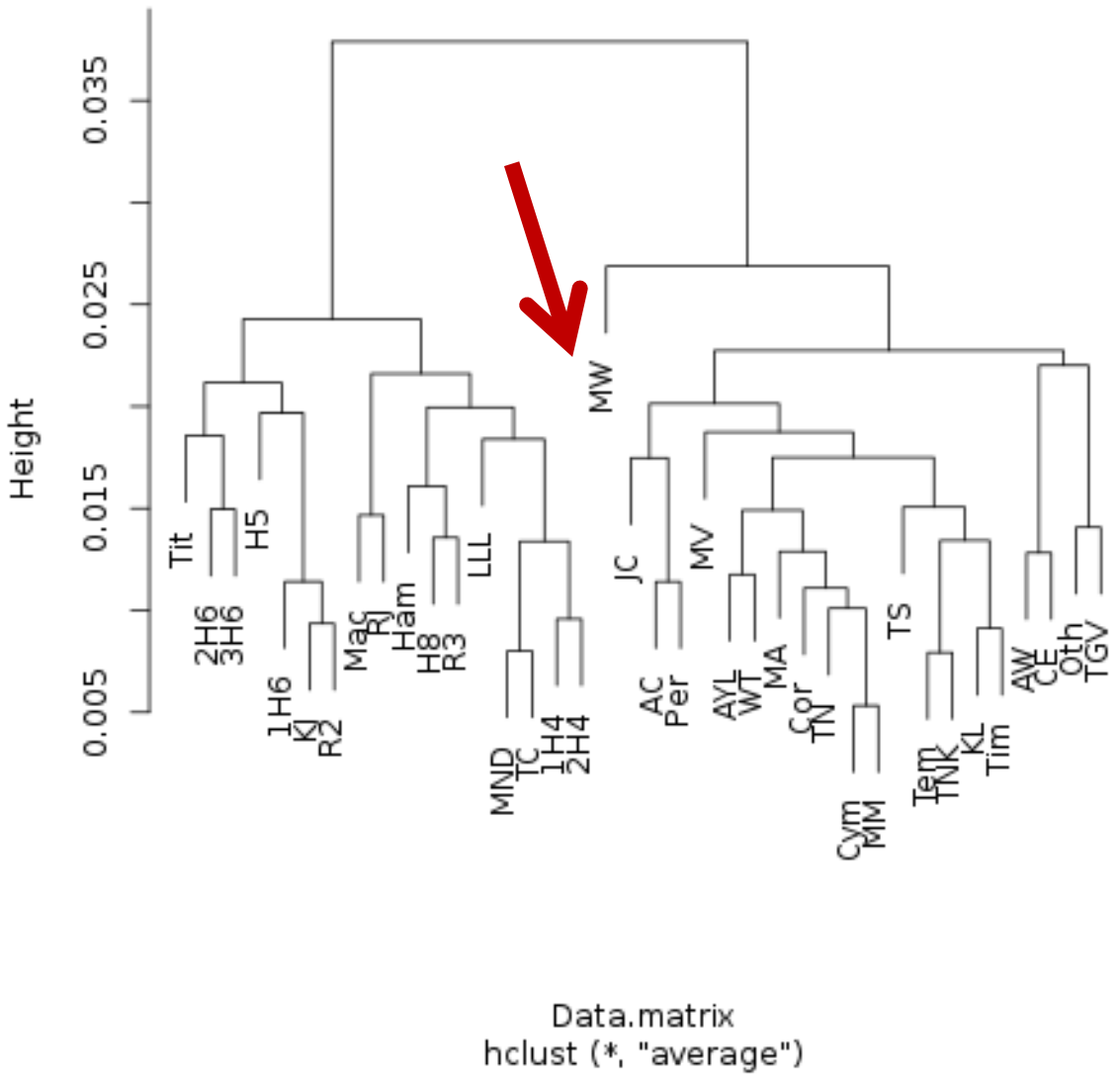
15 comedies; 9 tragedies; 0 histories

Lancaster University

Plotting POS in the Shakespeare corpus

Lancaster University

Cluster plot resulting from hierarchical agglomerative cluster analysis (using Euclidean distance and average linkage)



Conclusions/future plans

Lancaster University

- See how our Shakespeare results compare to existing findings (e.g. Craig 1991, 2000; Craig & Kinney 2009; Hope 2012; Lancashire 1997)
- Investigate possible reasons/explanations, e.g.
  - stylistic effects
  - characterisation
  - playwright idiolect & preferences
  - changes in authorial style over time (and/or the language of playwriting)
  - and so on …
- Obtain full results from comparative corpus
- Looking at grammar beyond just POS

References

Craig, H. (2004) "Stylistic analysis and authorship studies". In Schreibman, S. Siemens, R. & Unsworth, J. (eds.) *A Companion to Digital Humanities*. Oxford: Blackwell, pp. 273-88.

Craig, H. & Kinney, A.F. (2009) "Introduction". In Craig, H. & Kinney, A.F. (eds.) *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: CUP, pp. 1-14.

EEBO-TCP: Early English Books Online – Text Creation Partnership. http://www.textcreationpartnership.org/tcp-eebo/.

Internet Shakespeare Editions: http://internetshakespeare.uvic.ca.

Hardie, A. (2012) CQPweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3), 380-409. http://www.lancaster.ac.uk/staff/hardiea/cqpweb-paper.pdf.

Hope, J. (2012) "Shakespeare's mythic vocabulary – and his invisible grammar". Blog post 14 February 2012. http://winedarksea.org/?p=1487.

Lancashire, I. (1997) Empirically determining Shakespeare's idiolect. *Shakespeare Studies* 25: 171-85.

Leech, G.N., Garside, R. & Bryant, M. (1994) "CLAWS 4: The tagging of the British National Corpus". In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, Kyoto, Japan, 622-628. See http://ucrel.lancs.ac.uk/claws/.

Leech, G. and Smith, N. (2000) *Manual to accompany The British National Corpus (Version 2) with Improved Word-class Tagging*. http://www.natcorp.ox.ac.uk/docs/bnc2postag_manual.htm.

Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. (2007) "Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora". In *Proceedings of Corpus Linguistics 2007*, Birmingham University, U.K., 27-30 July 2007. http://ucrel.lancs.ac.uk/VariantSpelling/.