



*Encyclopaedia of
Shakespeare's
Language*



A series of mini-presentations focussing on method

27th July 2016

FASS Building

@ShakespeareLang



Arts & Humanities
Research Council



THE QUEEN'S
ANNIVERSARY PRIZES
FOR HIGHER AND FURTHER EDUCATION
2015



Spelling regularisation

Dawn Archer, Paul Rayson and Alistair Baron

Spelling variation

Until recently, diachronic studies of spelling variation have tended to (i) be qualitative in nature, and (ii) focus on most obvious spelling patterns for the period(s).

See, e.g., Smith (2005: 222), who comments on interchangeability of **<u> / <v>** (depending on their initial/medial positioning), use of **<i> to represent <j>** and use of **<vv> for <w>** in respect to Shakespearean English (see also Blake, 1996. 2002; Scragg, 1974).

Predictable focus – given such patterns will “jump out” at the researcher as they read a text – ***but there are issues with this type of approach – not least that patterns below the level of consciousness – due to being more subtle or because they only emerge across many texts – can easily go unnoticed.***

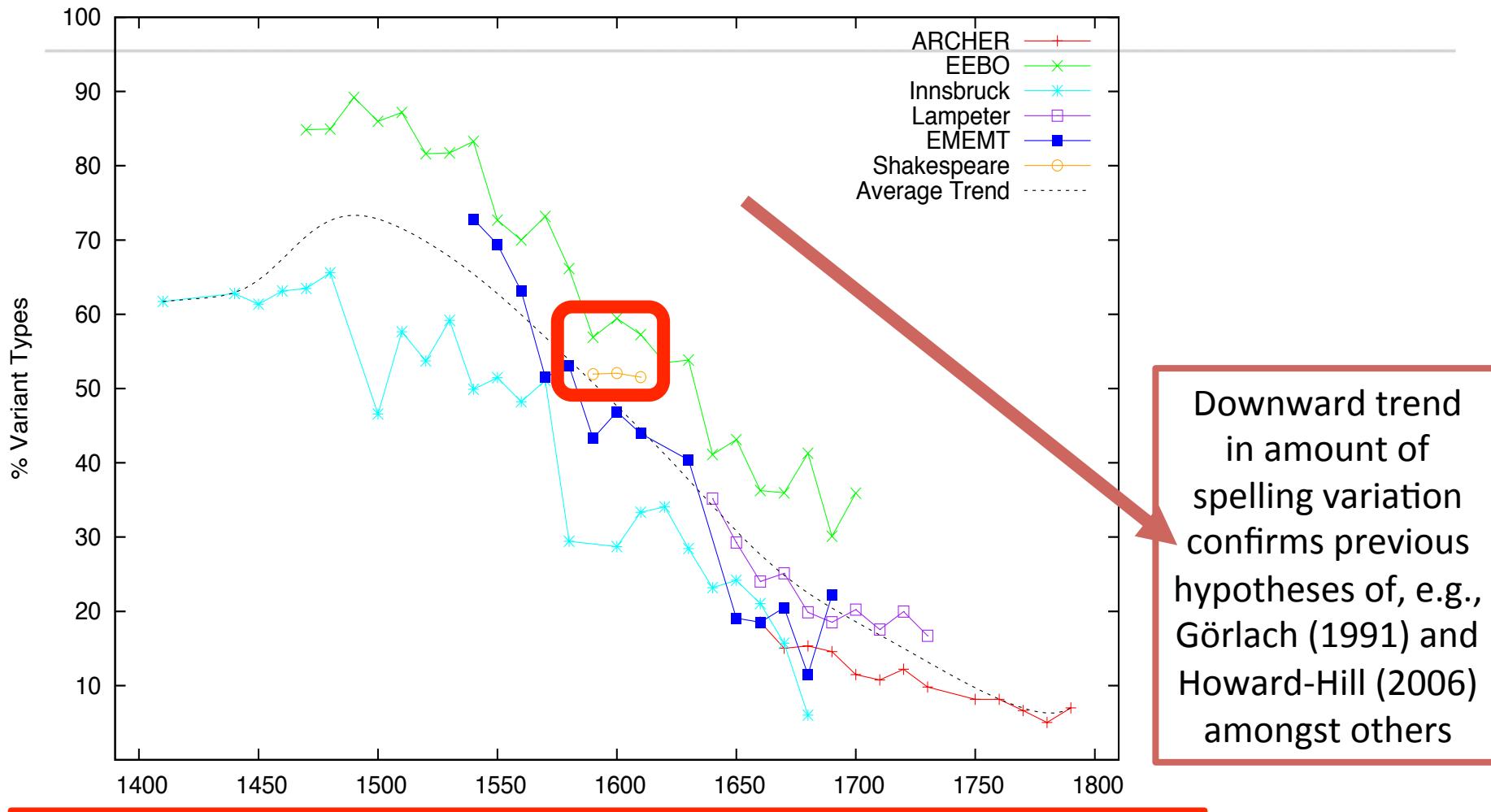
VARD and DICER

When combined, provide researchers with the means of exploring spelling variability more subtly and systematically.

- VARD can be used manually or automatically to detect spelling variants and suggest modern equivalents.
- DICER enables users to explore the spelling patterns found within the VARD-ed data.

Proven to be particularly useful when seeking to overcome the difficulties occasioned by attempting to identify a large number of variants across many texts, genres and/or centuries.

Baron *et al's* (2009) investigation of six corpora representative of EmodE period



But note that *First Folio* shows more variation - in terms of types found – than all but EEBO, in decades in question

Tallies with Rayson et al's (2007) VARD-based study of *The Winter's Tale* (1623)

Spelling pattern	No of types showing pattern (in first 500)	Example variants	Total occurrences of pattern (i.e., tokens)
Extra letter e		VARD automatically highlighted 2,114 spelling variants, which equates to more than 50% of the total word count (of 4,195 words).	958
Multiple u – v			265
‘ – e			681
ie – y			144
Fused form y- i			
Morphological		The results confirm many of the spelling inconsistencies highlighted by, e.g., Blake (2002), Scragg (1975), Smith (2005) ... whilst also indicating <i>how often these patterns occur</i> ...	55
Hyphenated compound			148
Missing letter			34
Doubling of consonant			75
v – u			17
i – j		The authors go on to discuss the most common spelling types based on the first 500 spelling variants identified by VARD	11
			26
			17
			25

Spelling pattern	No of types showing pattern (in first 500)	Example variants	Total occurrences of pattern (i.e., tokens)
Extra letter e	147	Drinke; eare; bulke; wisdome	958
u – v	55	Seruices; haue; euer	681
‘ – e	46	Accurs’d; fill’d; steep’d; th’; dear’st; do’s	See Archer & Rayson (2004)
v – u	9	Vnderstanding; vtterance	cf. Smith (2005)
i – j	6	Coniure; ioy; iustly; subiect	cf. Smith (2005)

‘d > ed particularly prevalent into 18th C (1,287 occurrences per mill words)
 Also occurred 177 times per mill words in 19th C data ...

VARD (VARiant Detector)

The screenshot shows the VARD 2.5 application window. The main text area displays a historical text from 1652, with various words highlighted in yellow and green. A context menu is open over the word 'themselves', showing options like 'Normalise instance', 'Normalise all', and 'Frequency is 237'. The 'Sidebar' panel on the right contains a 'Types List' with items like 'themselves (21)', 'som (19)', and 'beeing (16)'. The 'Text toolbar' at the top includes buttons for 'Text', 'Batch', 'Train', 'Setup', and 'Rules'.

THE ADVOCATE. ----- LONDON, Printed by William Du-Gard, Printer to the Council of State; and are to bee sold by Nicolas Bourn in Cornhill, at the Corner shop, at the entrance into the Exchange. 1652.

TO THE RIGHT HONORABLE THE Council of State. Right Honorable! I Am often in very great doubt (if I may so speak), that the Goodness & Wisdom of God, & his thoughts of these, are very rarely met with in the Paths, which the **scantling** of Man's Reason and Judgment walk in; And as I dare not but own the Belief of the Coming of his Appearance, and the breaking forth, very shortly, of his **Glorie**: So I **believe** likewise, this will bee a sight very strange, and very **unexpected** to men; and not **only** greatly above, but in **som** measure even **contrarie** (and perhaps, very **unwelcom**) unto the most enlarged and raised thoughts wee have yet prepared our **selvs** with, to **receiv** it.

The sign of which Coming, will bee the Detection, by little and little, of all Imposture, and the laying of all things low, naked, and mean before him; the stripping men of that **Honor**, Credit, and Repute, that they had by several means beene gaining with **themselves**, (and by **themselves**) with others, either through a **Not-knowledg**, or through an artificial concealing of **themselves**.

In these very things (if well observed) lying, and within these indeed, **beeing** spent and consumed, the whole **Indeavors**, Practice, **Studie** and **Wisdom** (if not **Religion**) of All States, Ages, Nations and Men, viz.

Either in devising shifts (by acquisition of such and such Power, Habits, or otherwise) not to appear **uncomly** to **themselves**, but to beget rather a greatness, esteem and satisfaction in **themselves**, and others of **themselves**;

Or, in Concealing and keeping close (by several Glosses) the Imperfection apprehensions of others;

The more exquisite **main text area** ferior **only** in this, that the other; And the most fortunate man in this world, having the Advantage covering of this fall's off a little sooner, and the other **wrap's** himself up in For thus all Lusts whatsoever (whether those of Acquisition) in the mind utmost Accomplishments, are still but either to give countenance unto, or The Assumption (on the other side) of a Repute and **Esteemence** to our **se** currancy (during it) to all Lusts. (These two, observing tune, and holding time each to the measures and motion therefore is concluded whole Humane Nature, with all the Parts, and the most perfect Actions of it.

Upon these thoughts (Right Honorable!) I finding none not very ready to have **asscribed** to himselfe the name of **beeing** (in some measure) wise, because it is estimable, finde none that can endure the Manifestation (in any kinde) of their **Follie**; or that can bee willing to have this their **Wisdom** and **Glorie** to bee flatly and plainly called a **Pageantrie**, **Mumming**, or playing at **bo-peep** with **themselves** and others; and yet (it really **beeing** no honester) if wee cannot bear the Judgments thus one of another; how shall wee abide the Trial, or bee able to stand before the Judgment of God, which **seeth** us, (and will shortly search us) with more pure and **uncorrect** **Eies**? And how miserably then are the Imaginations of almost all men muddled up here in their own dressings? when the Truth most certainly is in our **Beeings**, at their best (natural, or **acquir'd**) **Appearances**, are yet such feign'd Things, and **Propp'd** up with such **Lyes** and fundamental mistakes of **themselves**, that they would (**Belshazzar** like,) moulder and shatter **themselves** to pieces, should they bee taken quite asunder from the disguises and **patchings** they have put upon **themselves**, **bare-facedly** and **undeceivedly** to see **themselves** but one small part of an hour.

Beeing convinced, by these, therefore (Right Honourable!) of the low Condition wee are herein, I not knowing what the Councils of God intend to bring forth for the settlement of this Nation; Nor how **hee** hath resolved in his **Wisdom** to dispose of it, (as to its outward Condition,) whether **Hee**

current file and encoding: Eca1652.txt (charset: UTF-8)

progress bar

How VARD works ...

- Methods from modern spellchecking used to find historical spelling variants and offer/select appropriate (mod) equivalents.
- Original spelling retained in-text with an xml tag surrounding replacement - <normalized orig="charitie">charity</normalised>
- Allows for use of standard CL/NLP tools without any modification.
- Used to normalise released historical (and other) corpora, e.g. EMEMT (Lehto et al., 2010), CEEC (Palander-Collin & Hakala, 2011).



Part of speech tagging (CLAWS)

Paul Rayson (School of Computing and Communications)
@perayson



Arts & Humanities
Research Council



THE QUEEN'S
ANNIVERSARY PRIZES
FOR HIGHER AND FURTHER EDUCATION
2015



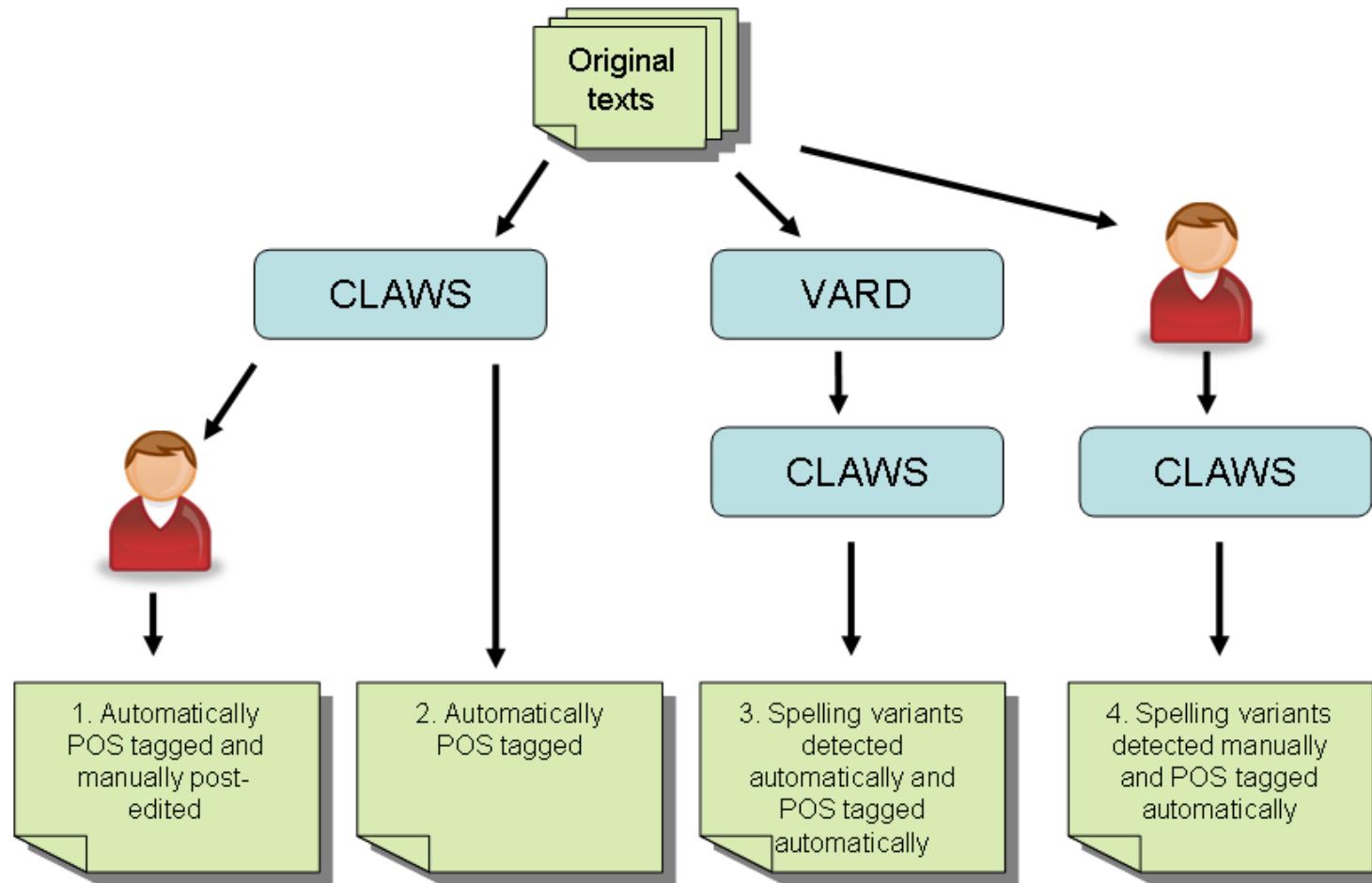
Part of speech tagging (using CLAWS)

- Origin of state automobile practices. The practice of state-owned vehicles for use of employees on business dates back over forty years.
- Origin_NN of_IN state_NN automobile_NN practices_NNS ._. The_DT practice_NN of_IN state-owned_JJ vehicles_NNS for_IN use_NN of_IN employees_NNS on_IN business_NN dates_VVZ back_RP over_IN forty_CD years_NNS ._.

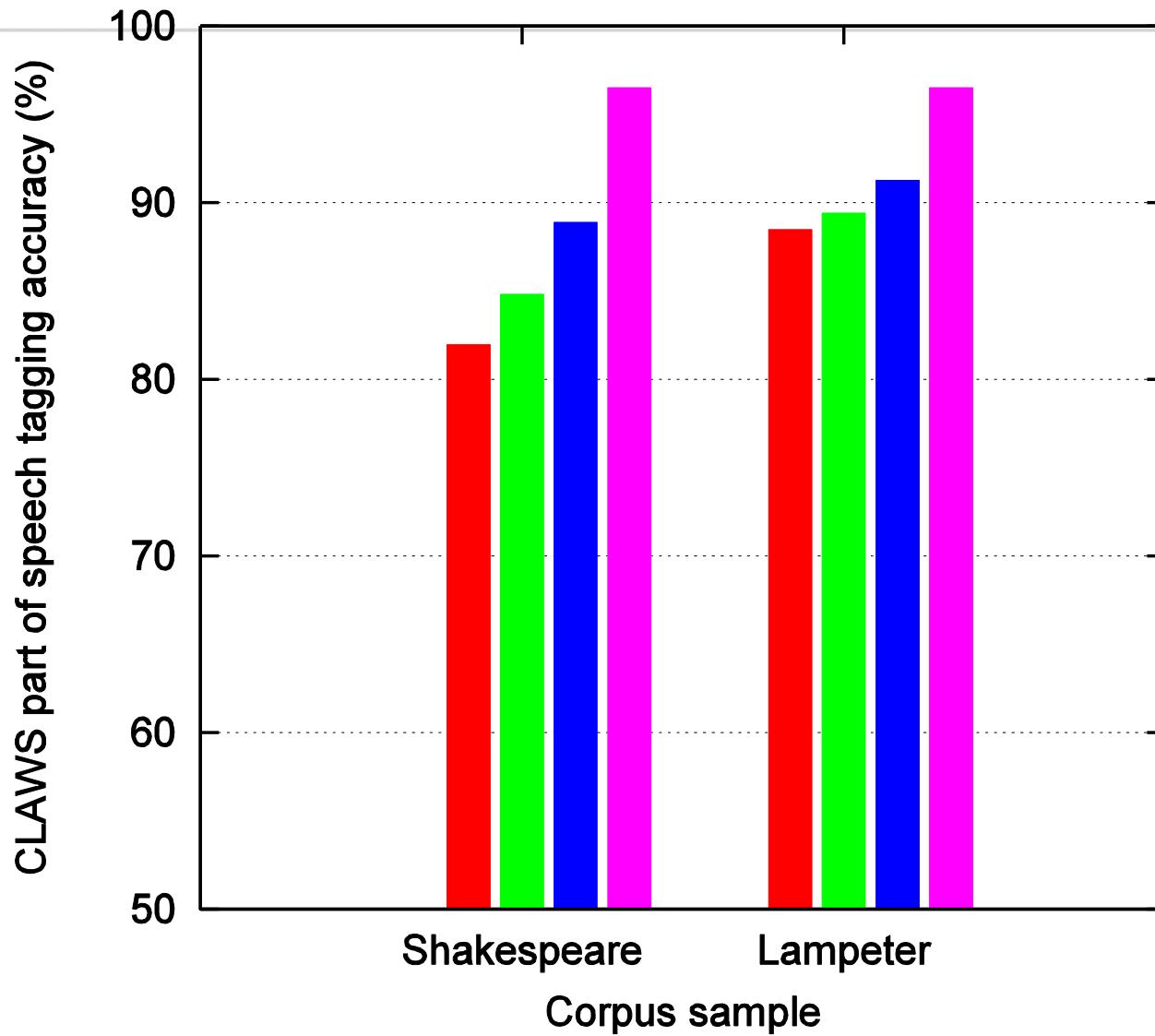
CLAWS overview

- = **Constituent Likelihood Automatic Word-tagging System**, made up of **Lexicons** (i.e., words and MWUs) + **matrix containing sequence probabilities** (e.g. likelihood Noun will follow Adjective)
- Applied to each sentence to disambiguate words, which could potentially be several parts-of-speech
- Trained predominantly on standard English (written & spoken) modern corpora, but some exposure to non-standard English and EModE through our research ...
- CLAWS achieves around 97% accuracy re modern General English

An experiment



With no standardization
After automatic standardization
After manual standardization
When applied to Modern British English





Semantic tagging (USAS and HTST)

Paul Rayson (School of Computing and Communications)
@perayson



Arts & Humanities
Research Council



THE QUEEN'S
ANNIVERSARY PRIZES
FOR HIGHER AND FURTHER EDUCATION
2015



UCREL Semantic Analysis System (USAS)

- Semantic field annotation has applications for conceptual or topic tagging:

There_Z5 's_Z5 been_A3+ more_N5++ violence_E3- in_Z5 the_Z5
Basque_Z2 country_M7 in_Z5 northern_M6 Spain_Z2 :_PUNC
one_N1 policeman_G2.1/S2m has_Z5 been_Z5 killed_L1-,_PUNC
and_Z5 two_N1 have_Z5 been_Z5 injured_B2- in_Z5 a_Z5
grenade_G3 and_Z5 machine-gun_G3 attack_G3 on_Z5 their_Z8
patrol-car_M3/G2.1 ._PUNC

- E3 = emotional states; Z2 = geographical names; M7 = places; M6 = location and direction; G3 = warfare; M3 = land transportation

How USAS works ...

- Lexicon of 56,316 items
e.g., **presentation NN1 Q2.2 A8 S1.1.1 K4**
- MWE list of 18,971 items
e.g., **travel_NN1 card*_NN* M3/Q1.2**
- A small wildcard lexicon
e.g., ***kg NNU N3.5**
- Unknown words using WordNet synonym lookup
- A set of six disambiguation techniques
- Accuracy of around 91% re modern General English

A General and abstract terms	B The body and the individual	C Arts and crafts	E Emotion
F Food and farming	G Government and public	H Architecture, housing and the home	I Money and commerce in industry
K Entertainment, sports and games	L Life and living things	M Movement, location, travel and transport	N Numbers and measurement
O Substances, materials, objects and equipment	P Education	Q Language and communication	S Social actions, states and processes
T Time	W World and environment	X Psychological actions, states and processes	Y Science and technology
Z Names and grammar			

Introducing the HTST

- HTOED
 - = comprehensive analysis of English, as found in OED (2nd ed.)
 - = 793,742 word forms arranged into 225,131 semantic categories
- HT semantic categories recently mapped to 4,028 thematic-level categories as part of SAMUELS project.
- Enables:
 - context sensitive tagging (OED sense mapping, sense definitions (14.5M tokens) and example sentences (50.2M tokens))
 - Time sensitive tagging
 - Accuracy of 84%+



Social tagging

Dawn Archer

Record. He did not go out of your Company at all?

Ann. Yes about Ten a Clock.

Record. Woman you must be mistaken, he came to Town at Twelve or One, and might be in thy company, but it is plain he went to a Brokers in *Long-lane*, and so to the *Artillery-Ground* at *Cripplegate*, for I guesst it might be so: Then they went to *Whetstones-Park*, and spent Six-Pence, and after that they went into *Drury-lane*.

Giles, My Lord, she don't say she was with us all the while, but we came to an Housse where she was, and several other People our Neighbours.

Record. She says you did go out some time: Now see whether I mistake you.

Ann. Yes you do mistake me.

Record. He went out, did he?

Ann. Yes he went out after we came into the City, he and some others, and then they came back to me again in two or three hours.

Record. Then you were two or three hours at Dinner. Now I ask you, After they came back, was you with him all the while?

Ann. Yes that I was.

Record. Where was it?

Ann. At the *Peacock*.

Record. That is the place in *Drury-lane*.

Ann. No, indeed, it is in *Covent Garden*.

Mr. Darnal. When did he go to Bed, do you know that upon your Oath?

Ann. We were in the Inn between Nine and Ten a Clock, nearer Ten then Nine, and I saw him sitting taking a Pipe of Tobacco.

Mr. Darnal. What time was that?

Ann. A little after Ten I believe.

Mr. Thomp. He sat there till he was call'd away to do his business.

<P 37>

[\$ (^Record.^) \$]<u speaker="s" spid="s4tgiles001" spsex="m" sprole1="re" spstatus="1" spage="8" addressee="s" adid="s4tgiles027" adsex="f" adrole1="w" adstatus="5" adage="x">He did not go out of your Company at all? </u>

[\$ (^Ann.^) \$]<u speaker="s" spid="s4tgiles027" spsex="f" sprole1="w" spstatus="5" spage="8" addressee="s" adid="s4tgiles001" adsex="m" adrole1="re" adstatus="1" adage="x">Yes about Ten a Clock.</u>

See Archer and Culpeper (2003)

Our approach in this project

Field	Feature marked	Sign	Possible values
1	speaker(s)	speaker- =	singular (s) or multiple (m)
2	speaker ID tag	spid=	already undertaken for us
3	gender of speaker	spsex=	male (m), female (f), assumed male (am) , assumed female (af) , neither (n)
4	status/social rank of speaker	spstatus=	nobility (0), gentry (1), professionals (2), other middling groups (3), ordinary commoners (4), lowest groups (5)
5	speaker age	spage=	young (6), adult (8), older adult (9)

Status/social rank categories are based on rank, estate or sort, in order to reflect (i) pre-industrialised nature of EmoDE society, and (ii) way in which EmoDE contemporaries spoke about status (Harrison, 1577; Holmes, 1982; Wrightson 1982, 1991; Sharpe, 1987; Nevalainen & Raumolin-Brunberg, 1996).

Category relating to **gender of speaker** reworked to enable assumed genders to be marked specifically

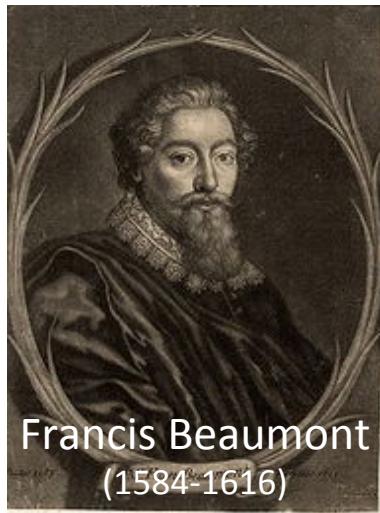
Our approach in this project

Field	Feature marked	Sign	Possible values
1	speaker(s)	speaker- =	singular (s) or multiple (m)
2	speaker ID tag	spid=	already undertaken for us
3	gender of speaker	spsex=	male (m), female (f), assumed male (am) , assumed female (af) , neither (n)
4	status/social rank of speaker	spstatus=	nobility (0), gentry (1), professionals (2), other middling groups (3), ordinary commoners (4), lowest groups (5)
5	speaker age	spage=	young (6), adult (8), older adult (9)

Young indicates 0-14, **adult**, 15-44, **older adult**, 45+. A nominal age range rather than a specific age so as to reflect socio-historical situation – i.e., to correspond with significant milestones such as age of first marriage (Sharpe, 1987: 40; Coward, 1988: 20; Wrightson, 1982), commencement/completion of apprenticeships (O'Day, 2000: 20-24; Holmes, 1982), significant advancement within profession (Foss, 1870; Simpson, 1984), and average expectation of life at *birth* (i.e., upper 30s/early 40s) (Sharpe, 1987: 38; Coward, 1988).

A comparative corpus of plays by Shakespeare's contemporaries

Jane Demmen



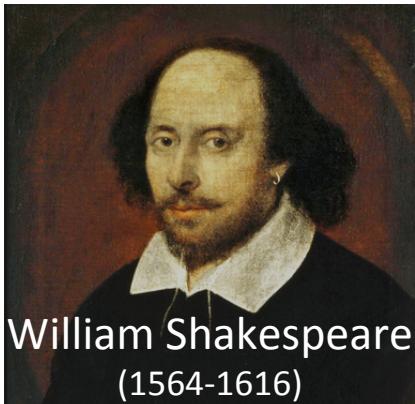
Francis Beaumont
(1584-1616)



John Fletcher (1579-1625)

Anthony Munday
(1553-1633)

George Wilkins
(c.1576-1618)



William Shakespeare
(1564-1616)

Thomas Drue
(c.1586-1627)

John Webster
(c.1578-1634)

George Chapman
(c.1559-c.1634)

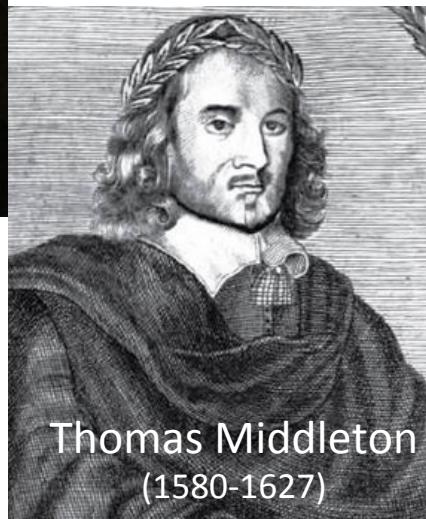
Robert Greene
(1558-1592)

George Peele
(1556-1596)

Thomas Kyd
(1558-1594)

Thomas Dekker
(c.1570-1632)

John Lyly
(1554-1606)



Thomas Middleton
(1580-1627)

Henry Porter
(d.1599)

John Marston
(c.1575-1634)

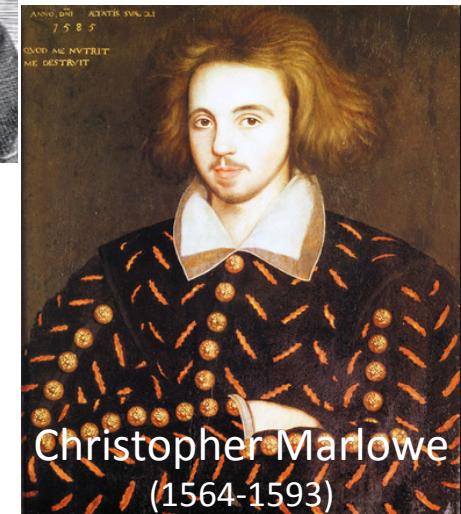
William Rowley
(1585-1637)



Ben Jonson (1572-1637)

Philip Massinger
(1583-1640)

Thomas Heywood
(c.1574-1641)



Christopher Marlowe
(1564-1593)

Why compare?

- To contextualise Shakespeare's language
(relative to that of a group of his peers)
-> we can see language style features which are typical of plays more widely in this period (not just Shakespeare's)

A comparative corpus for Shakespeare's plays

- Shakespeare corpus: 38 plays, with first production dates from c.1589-1613
- Comparative corpus: 46 plays by 24 other playwrights, with first production dates from 1584-1626
- Both about 1 million words in size

Early English Books Online and genre classification

Sean Murphy

Early English Books Online

1520-1679: 732 million words



1560-1640
368 million words

Bible
Catholicism
Essays
Law
Letters
Parliament
Philosophy
Plays
Poetry
Protestantism
Royalty
Science



A
**Dialogue agaynst
 light, lewde, and la-
 sciuous dauncing: wher-
 in are refuted all those rea-
 sons, which the com-
 mon people vse to
 bring in defence
 thereof.**

Compiled and made by
Christopher Fetherston.

Eccle. 9.4.
*Use not the company of a wo-
 man, that is a singer and a
 dauncer, least thou be entrap-
 ped in her snares.*

Imprinted at Lon-
 don, by Thomas Dawson.
 1582.

A
**Dialogue against
 light, lewd, and la-
 sciuous dancing : wher-
 ein are refuted all those rea-
 sons , which the com-
 mon people use to
 bring in defence
 thereof.**

Compiled and made by
Christopher Fetherston.

Eccle. 9. 4 .
 Use not the company of a wo-
 man , that is a singer and a
 dancer , least thou be entrap-
 ped in her snares.

Imprinted at Lon-
 don , by Thomas Dawson.

Styles	Domains	Genres	Sub-genres (examples)
Literary	Imaginative	Plays Poetry, Verse & Song Fiction General	Comedy, History, Tragedy, Masque Ballads
Formal – Spiritual	Religion	Bible Catholicism Protestantism Doctrine, Theology and Governance General	Anti-Catholicism Church of England Sin and Repentance Sermons
Formal - Statutory	Government	Royal Parliamentary Legal General	Proceedings Reports Trials Speeches
Formal - Instructional	Didactic	Philosophy Science Mathematics Medicine General	Experiments Anatomy Alchemy
Informational	Factual	Biography Essay Letters Pamphlets General	Dialogue Food and Cookery
	Other	French, Latin, Unclassified	



Corpus data, corpus affordances: *Methodology to support the Encyclopaedia*

Andrew Hardie

Lancaster University

a.hardie@lancaster.ac.uk | [@HardieResearch](https://twitter.com/HardieResearch) | web: cass.lancs.ac.uk



Arts & Humanities
Research Council

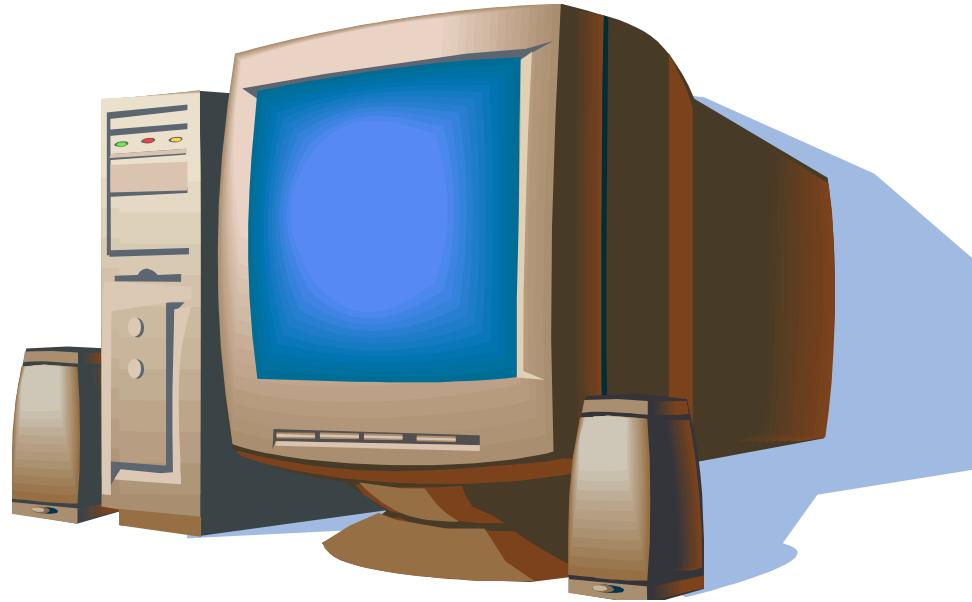


THE QUEEN'S
ANNIVERSARY PRIZES
FOR HIGHER AND FURTHER EDUCATION
2015



Corpus linguistics as methodology

- Scale



- Quantitative / qualitative analysis
- Lexicography

What can the computer do with a corpus?

- **Search**
 - Concordance
- **Count**
 - Frequency data
- Statistical abstraction via comparison
 - Concordance >> Collocation analysis
 - Frequencies >> Keyness analysis
- The four basic methods
 - *>>> Close reading of examples*

- Server-based
- Basic and advanced systems
- Access control
- Flexibility

- Software homepage:
 - <http://cwb.sf.net/cqpweb.php>
- Create an account on the Lancs CQPweb server:
 - <https://cqpweb.lancs.ac.uk> >> “Create account”
- Our “test data”:
 - <https://cqpweb.lancs.ac.uk/shaktextff>