Encyclopaedia of Shakespeare's Language

Lancaster University

A series of mini-presentations focussing on method

27th July 2016
FASS Building
@ShakespeareLang

Arts & Humanities Research Council

CASS
Corpus Approaches to Social Science

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

UCREL

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.***

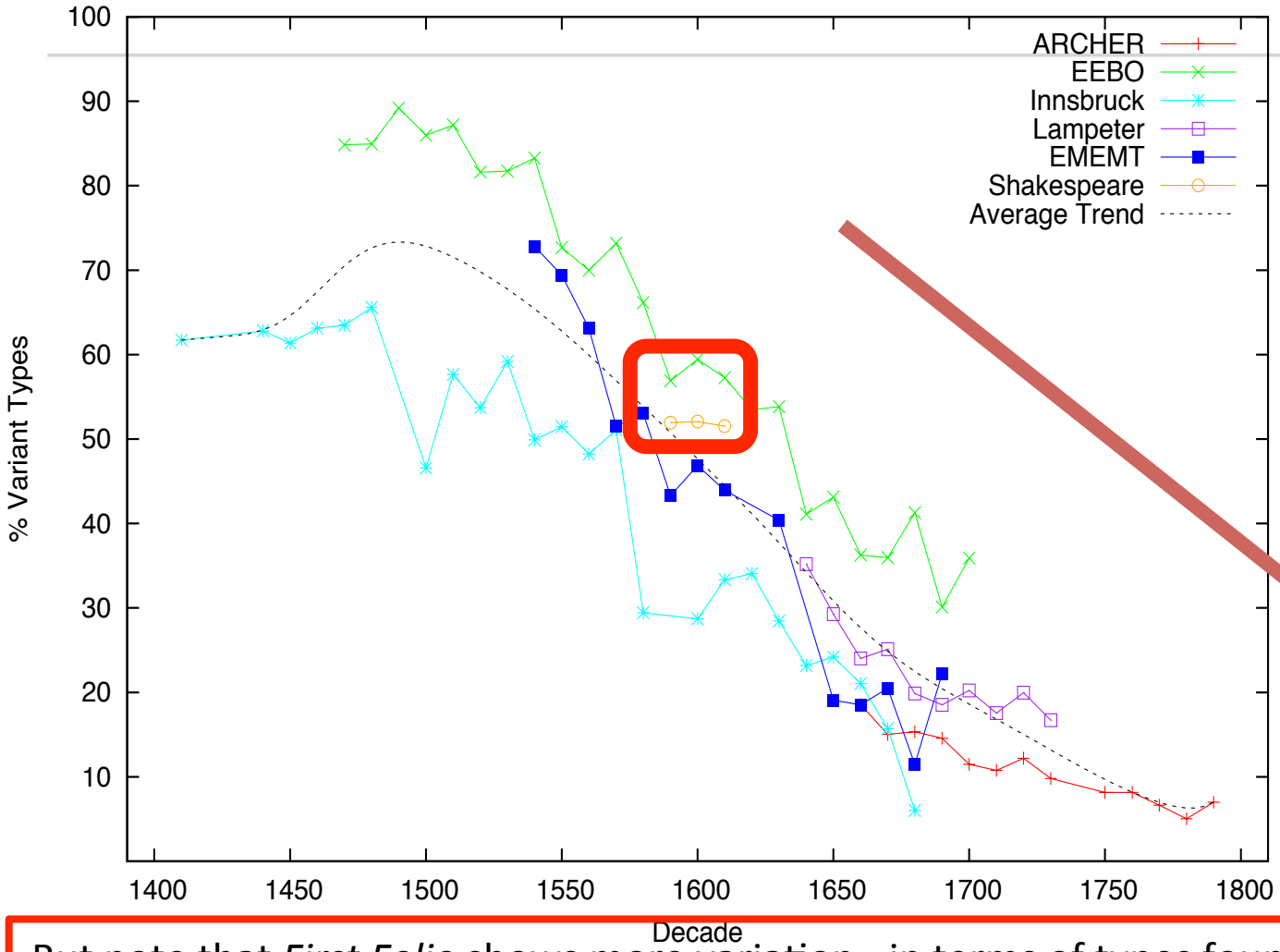
VARD and DICER

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



Downward trend in amount of spelling variation confirms previous hypotheses of, e.g., Görlach (1991) and Howard-Hill (2006) amongst others

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)

Lancaster University

Spelling pattern	No of types showing pattern (in first 500)	Example variants	Total occurrences of pattern (i.e., tokens)
Extra letter e			958
Multiple			265
u – v			681
' – e			144
ie – y			55
Fused form			148
y- i			34
Morphological			75
Hyphenated compound			17
Missing letter			11
Doubling of consonant			26
v – u			17
i – j			25

VARD automatically highlighted 2,114 spelling variants, which equates to more than 50% of the total word count (of 4,195 words).

The results confirm many of the spelling inconsistencies highlighted by, e.g., Blake (2002), Scragg (1975), Smith (2005) … whilst also indicating *how often these patterns occur …*

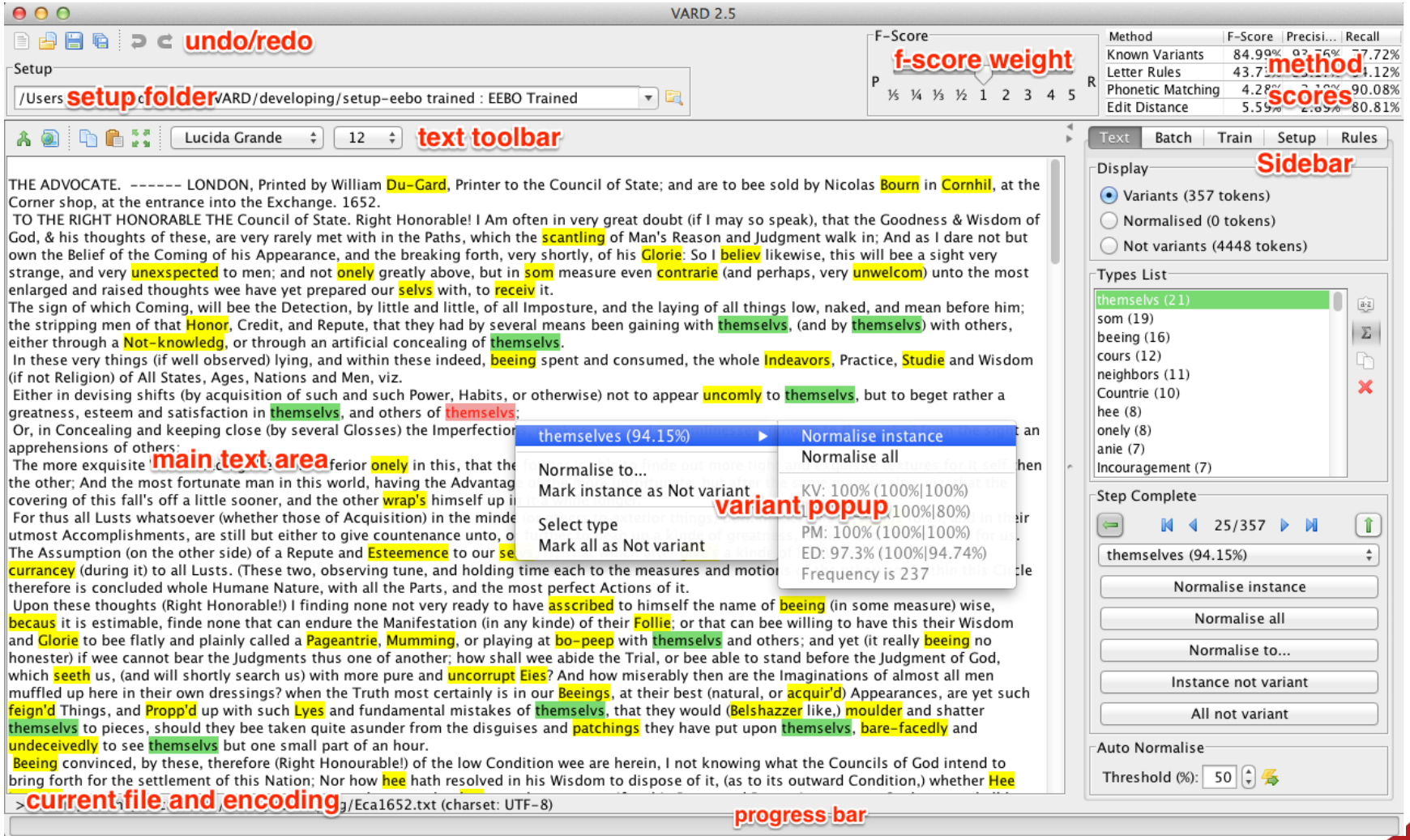
The authors go on to discuss the most common spelling types based on the first 500 spelling variants identified by

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)

http://ucrel.lancs.ac.uk/vard/



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).

Encyclopaedia of Shakespeare's Language

Lancaster University

Part of speech tagging (CLAWS)

Paul Rayson (School of Computing and Communications)
@perayson

Arts & Humanities Research Council

CASS
Corpus Approaches to Social Science

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

UCREL

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.

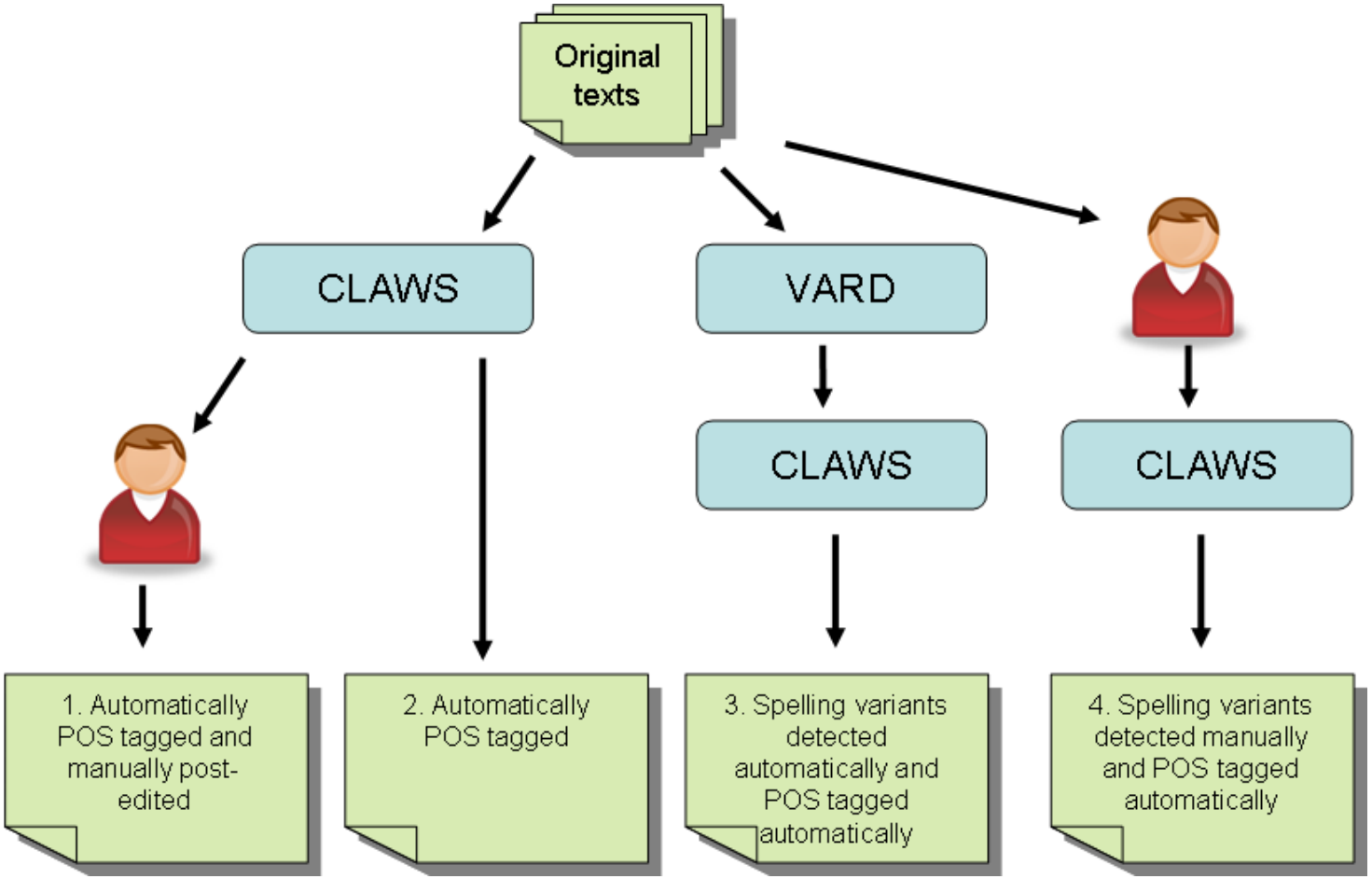
- Origin_NN of_IN state_NN automobile_NN practices_NNS ._. The_DT practice_NN of_IN state-owned_JJ vehicles_NNS for_IN use_NN of_IN employees_NNS on_IN business_NN dates_VVZ back_RP over_IN forty_CD years_NNS ._.

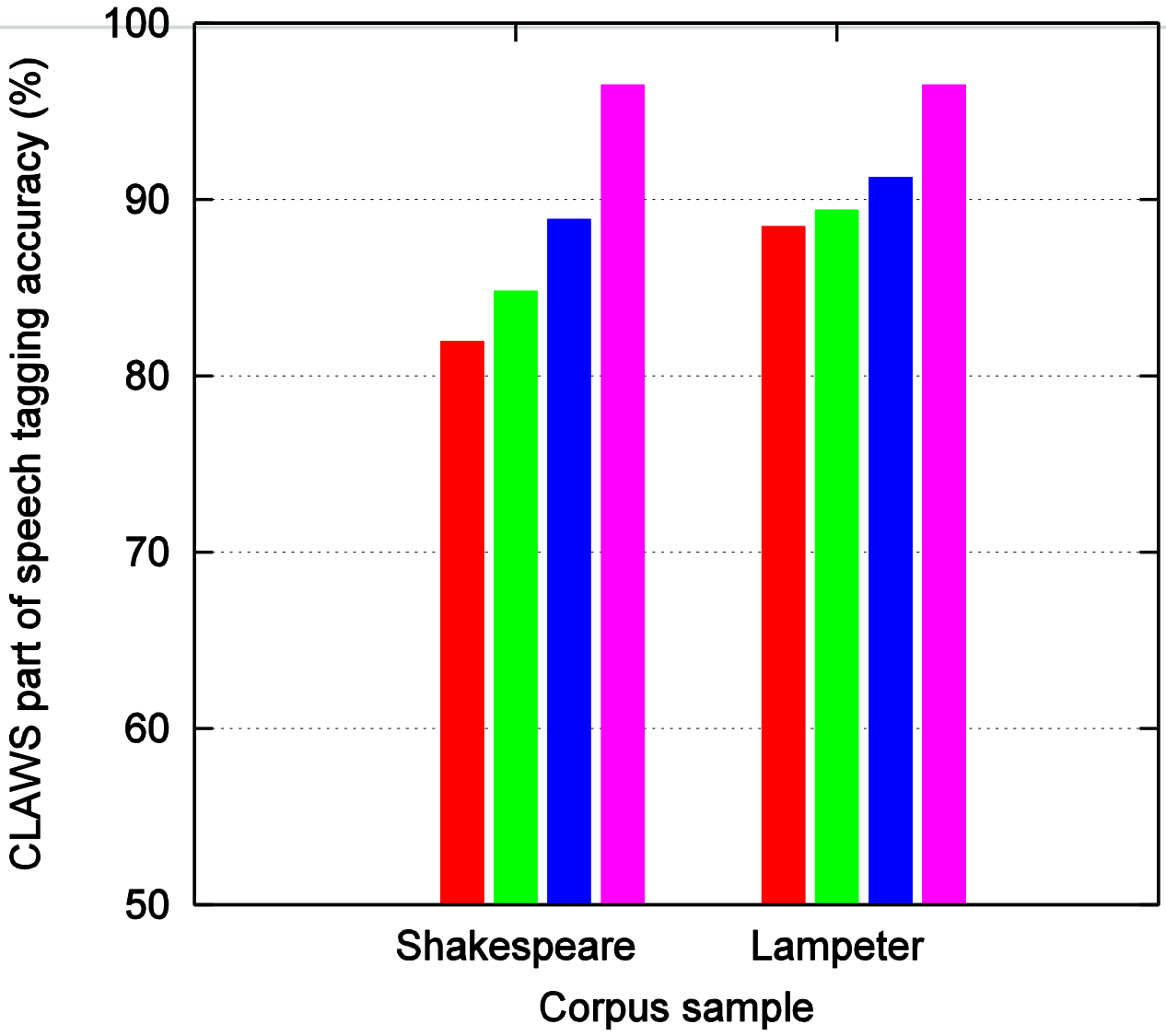
CLAWS overview

= **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
CLAWS part of speech tagging accuracy (%)
100
90
80
70
60
50
Shakespeare
Lampeter
Corpus sample

Encyclopaedia of Shakespeare's Language

Lancaster University

Semantic tagging (USAS and HTST)

Paul Rayson (School of Computing and Communications)
@perayson

Arts & Humanities Research Council

CASS
Corpus Approaches to Social Science

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

UCREL

UCREL Semantic Analysis System (USAS)

Lancaster University

- 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

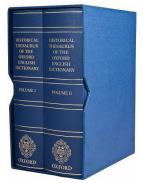
Lancaster University

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

Lancaster University

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 *Cripple-Gate*, for I gueſs 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 Houſe 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

Lancaster University

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

Lancaster University

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) (Sharpes, 1987: 38; Coward, 1988).

A comparative corpus of plays by Shakespeare's contemporaries

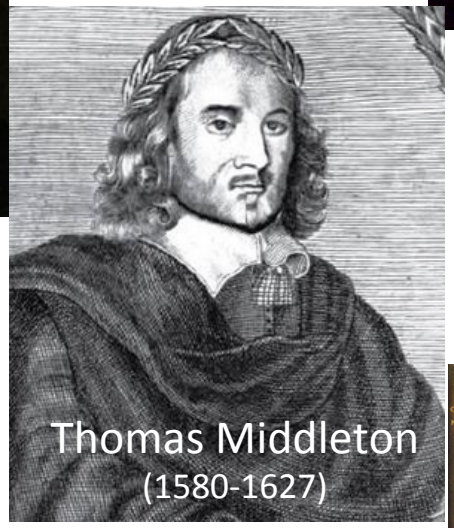
Jane Demmen

Thomas Drue
(c.1586-1627)

Thomas Kyd
(1558-1594)

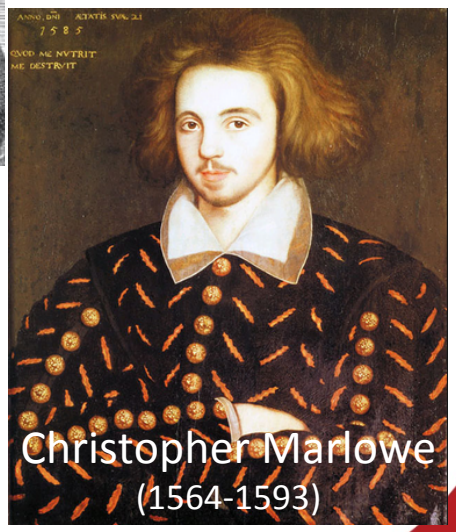
Thomas Dekker
(c.1570-1632)

John Lyly
(1554-1606)



Philip Massinger
(1583-1640)

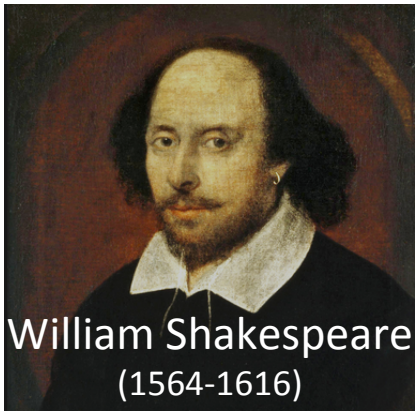
Thomas Heywood
(c.1574-1641)



Thomas Drue
(c.1586-1627)

Anthony Munday
(1553-1633)

George Wilkins
(c.1576-1618)



John Webster
(c.1578-1634)

George Chapman
(c.1559-c.1634)

Robert Greene
(1558-1592)

George Peele
(1556-1596)

Henry Porter
(d.1599)

John Marston
(c.1575-1634)

William Rowley
(1585-1637)

Why compare?

Lancaster University

- 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

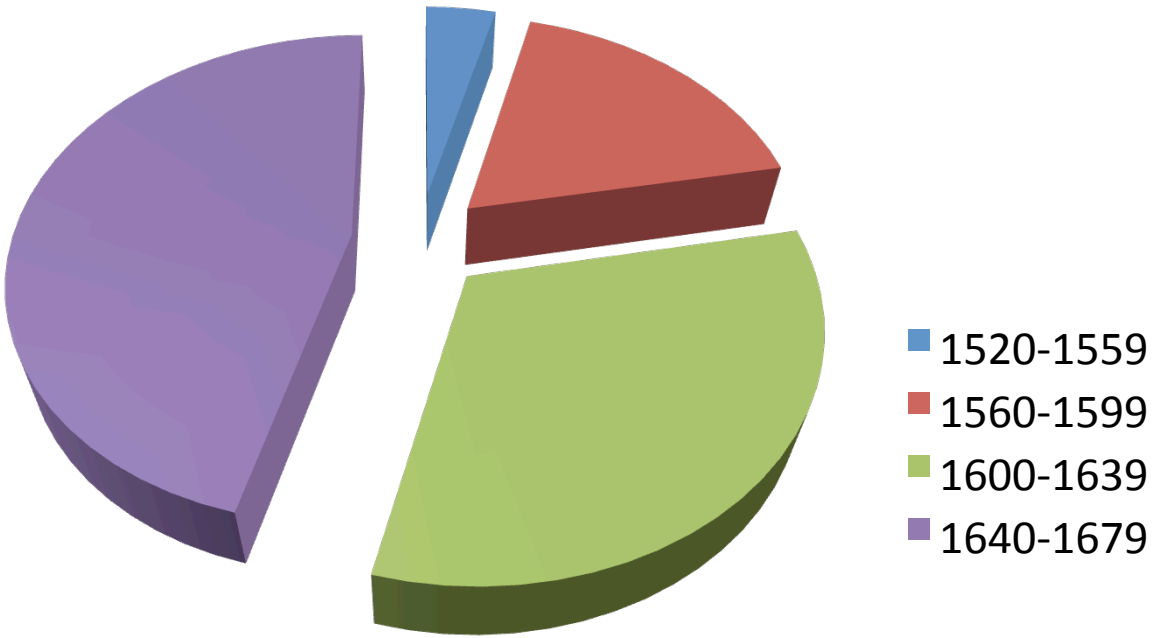
- Shakespeare corpus: 38 plays, with first production dates from c.1589-1613
- Comparative corpus: 46 plays by 24 other playwrights, with first production dates from 1584-1626
- Both about 1 million words in size

Early English Books Online and genre classification

Sean Murphy

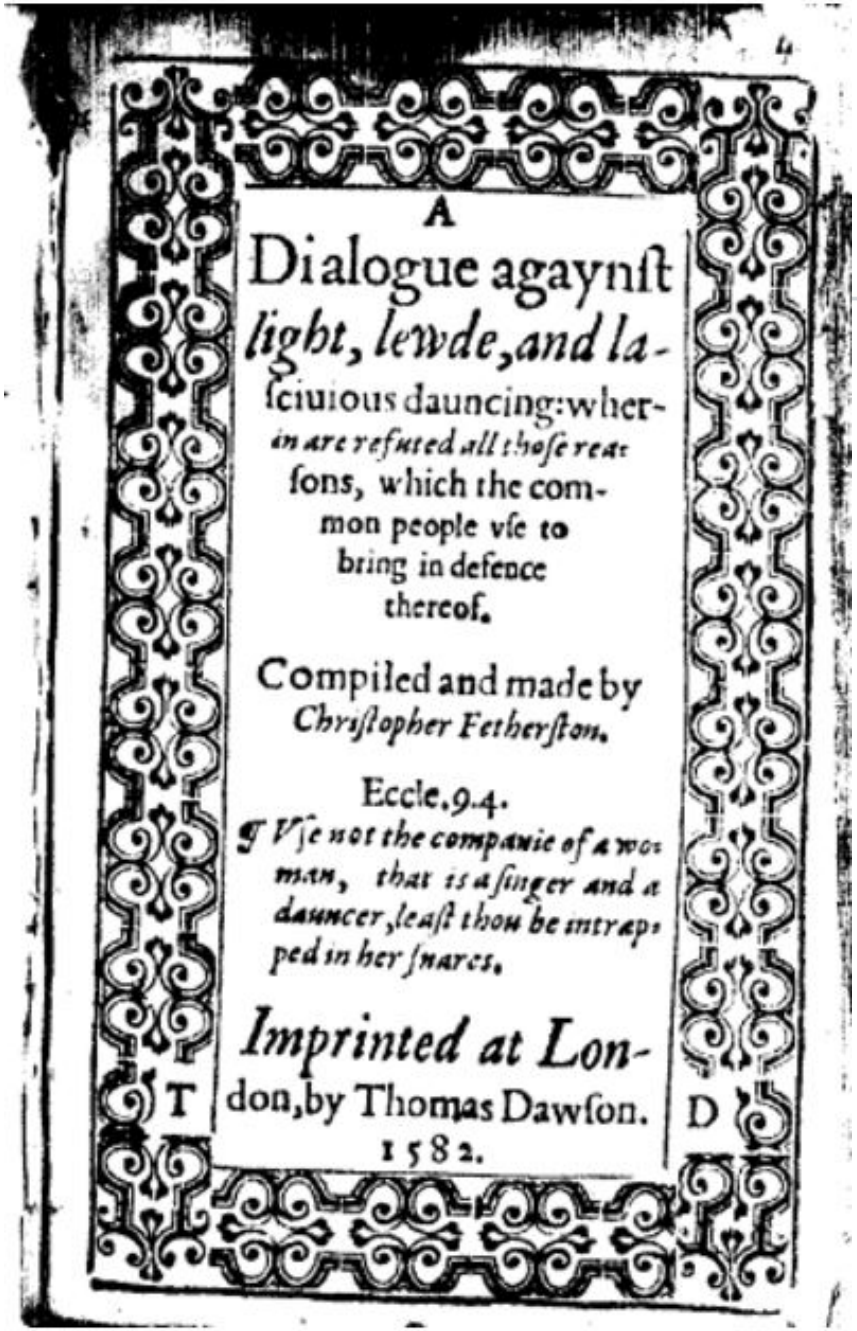
Lancaster University

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-
ſciuious 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.
¶ Vſe not the companie of a wo-
man, that is a ſinger and a
dauncer, leaſt thou be intrap-
ped in her ſnares.

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

T D

Lancaster University

A
Dialogue against
light , lewd , and la-
scivious 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.

1582.

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	

Encyclopaedia of Shakespeare's Language

Lancaster University

Corpus data, corpus affordances:

Methodology to support the Encyclopaedia

Andrew Hardie
Lancaster University
a.hardie@lancaster.ac.uk | @HardieResearch | *web:* cass.lancs.ac.uk

Arts & Humanities Research Council

CASS
Corpus Approaches to Social Science

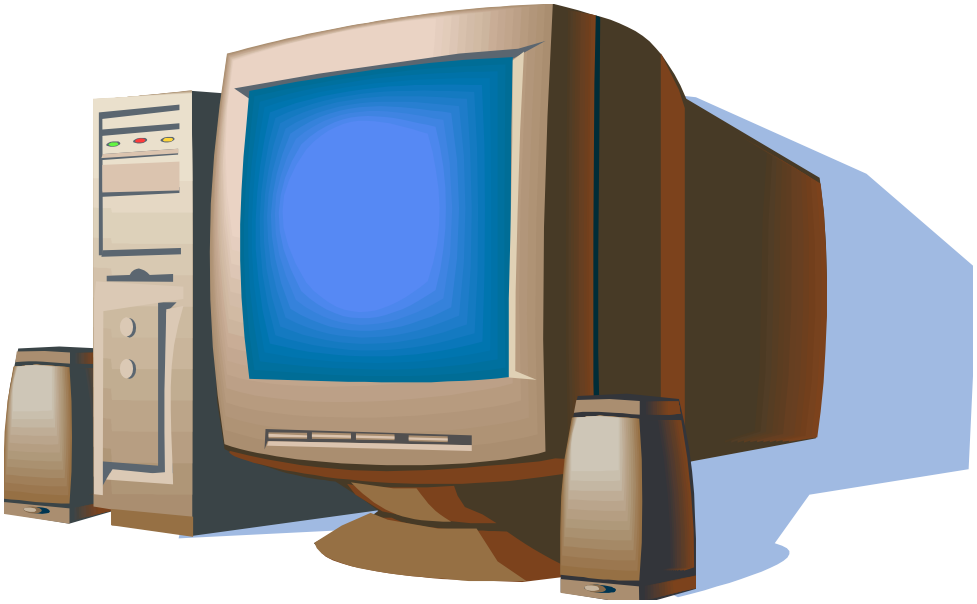
The Queen's Anniversary Prizes
For Higher and Further Education
2015

UCREL

Corpus linguistics as methodology

Lancaster University

- 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***

CQPweb

Lancaster University

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

CQPweb

Lancaster University

- 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