A series of mini-presentations focussing on method

27th July 2016
FASS Building
@ShakespeareLang
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

Downward trend in amount of spelling variation confirms previous hypotheses of, e.g., Görlach (1991) and Howard-Hill (2006) amongst others
Tallies with Rayson et al’s (2007) VARD-based study of *The Winter’s Tale* (1623)

<table>
<thead>
<tr>
<th>Spelling pattern</th>
<th>No of types showing pattern (in first 500)</th>
<th>Example variants</th>
<th>Total occurrences of pattern (i.e., tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extra letter e</td>
<td></td>
<td>VARD automatically highlighted 2,114 spelling variants, which equates to more than 50% of the total word count (of 4,195 words).</td>
<td>958</td>
</tr>
<tr>
<td>Multiple</td>
<td></td>
<td></td>
<td>265</td>
</tr>
<tr>
<td>u – v</td>
<td></td>
<td></td>
<td>681</td>
</tr>
<tr>
<td>‘ – e</td>
<td></td>
<td></td>
<td>144</td>
</tr>
<tr>
<td>ie – y</td>
<td></td>
<td></td>
<td>55</td>
</tr>
<tr>
<td>Fused form</td>
<td></td>
<td></td>
<td>148</td>
</tr>
<tr>
<td>y- i</td>
<td></td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>Morphological</td>
<td></td>
<td></td>
<td>75</td>
</tr>
<tr>
<td>Hyphenated compound</td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>Missing letter</td>
<td></td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>Doubling of consonant</td>
<td></td>
<td></td>
<td>26</td>
</tr>
<tr>
<td>v – u</td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>i – j</td>
<td></td>
<td></td>
<td>25</td>
</tr>
</tbody>
</table>

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

’d > ed particularly prevalent into 18th C (1,287 occurrences per mill words)
Also occurred 177 times per mill words in 19th C data ...
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</normalized>`
- 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
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

Original texts

CLAWS

VARD

1. Automatically POS tagged and manually post-edited
2. Automatically POS tagged
3. Spelling variants detected automatically and POS tagged automatically
4. Spelling variants detected manually and POS tagged automatically
With no standardization
After automatic standardization
After manual standardization
When applied to Modern British English

CLAWS part of speech tagging accuracy (%)

Shakespeare  Lampeter
Corpus sample

Lancaster University
Semantic tagging (USAS and HTST)

Paul Rayson (School of Computing and Communications)
@perayson
UCREL Semantic Analysis System (USAS)

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

  There’s been more violence in the Basque country in northern Spain: one policeman has been killed, and two have been injured in a grenade and machine-gun attack on their patrol-car.

- 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.,  
  \texttt{presentation} \texttt{NN1} \texttt{Q2.2 A8 S1.1.1 K4}

• MWE list of 18,971 items
e.g.,  
  \texttt{travel\_NN1 card\_*\_NN*} \texttt{M3/Q1.2}

• A small wildcard lexicon
e.g.,  
  \texttt{*kg} \texttt{NNU N3.5}

• Unknown words using WordNet synonym lookup

• A set of six disambiguation techniques

• Accuracy of around 91\% re modern General English
<table>
<thead>
<tr>
<th>Column</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>General and abstract terms</td>
</tr>
<tr>
<td>B</td>
<td>The body and the individual</td>
</tr>
<tr>
<td>C</td>
<td>Arts and crafts</td>
</tr>
<tr>
<td>E</td>
<td>Emotion</td>
</tr>
<tr>
<td>F</td>
<td>Food and farming</td>
</tr>
<tr>
<td>G</td>
<td>Government and public</td>
</tr>
<tr>
<td>H</td>
<td>Architecture, housing and the home</td>
</tr>
<tr>
<td>I</td>
<td>Money and commerce in industry</td>
</tr>
<tr>
<td>K</td>
<td>Entertainment, sports and games</td>
</tr>
<tr>
<td>L</td>
<td>Life and living things</td>
</tr>
<tr>
<td>M</td>
<td>Movement, location, travel and transport</td>
</tr>
<tr>
<td>N</td>
<td>Numbers and measurement</td>
</tr>
<tr>
<td>O</td>
<td>Substances, materials, objects and equipment</td>
</tr>
<tr>
<td>P</td>
<td>Education</td>
</tr>
<tr>
<td>Q</td>
<td>Language and communication</td>
</tr>
<tr>
<td>S</td>
<td>Social actions, states and processes</td>
</tr>
<tr>
<td>T</td>
<td>Time</td>
</tr>
<tr>
<td>W</td>
<td>World and environment</td>
</tr>
<tr>
<td>X</td>
<td>Psychological actions, states and processes</td>
</tr>
<tr>
<td>Y</td>
<td>Science and technology</td>
</tr>
<tr>
<td>Z</td>
<td>Names and grammar</td>
</tr>
</tbody>
</table>
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 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 sometime: 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.

See Archer and Culpeper (2003)
## Our approach in this project

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

**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

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

**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
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 against light, lewd, and lascivious dancing: wherein are refuted all those reasons, which the common people use to bring in defence thereof.

Compiled and made by Christopher Fetherston.

Eccle. 9. 4.
Use not the company of a woman, that is a singer and a dancer, least thou be entrapped in her snares.

Imprinted at London, by Thomas Dawson. 1582.
<table>
<thead>
<tr>
<th>Styles</th>
<th>Domains</th>
<th>Genres</th>
<th>Sub-genres (examples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literary</td>
<td>Imaginative</td>
<td>Plays, Poetry, Verse &amp; Song, Fiction, General</td>
<td>Comedy, History, Tragedy, Masque, Ballads</td>
</tr>
<tr>
<td>Formal – Spiritual</td>
<td>Religion</td>
<td>Bible, Catholicism, Protestantism, Doctrine, Theology and Governance, General</td>
<td>Anti-Catholicism, Church of England, Sin and Repentance, Sermons</td>
</tr>
<tr>
<td>Formal - Statutory</td>
<td>Government</td>
<td>Royal, Parliamentary, Legal, General</td>
<td>Proceedings, Reports, Trials, Speeches</td>
</tr>
<tr>
<td>Formal - Instructional</td>
<td>Didactic</td>
<td>Philosophy, Science, Mathematics, Medicine, General</td>
<td>Experiments, Anatomy, Alchemy</td>
</tr>
<tr>
<td>Informational</td>
<td>Factual</td>
<td>Biography, Essay, Letters, Pamphlets, General</td>
<td>Dialogue, Food and Cookery</td>
</tr>
<tr>
<td>Other</td>
<td>French, Latin, Unclassified</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Corpus data, corpus affordances:
Methodology to support the Encyclopaedia

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

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

- 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/shakttextff