Normalising the *Corpus of English Dialogues (1560-1760)* using VARD2: Decisions and Justifications

Dawn Archer¹, Merja Kytö², Alistair Baron³, Paul Rayson³

University of Central Lancashire¹
Uppsala University²
Lancaster University³
The extent of spelling variation in EmodE corpora

• And its effect on corpus methods such as keywords
• Searching for words can be problematic: *would, wolde, woolde, wuld, wulde, wud, wald, vvould, vvold*, etc.

• Frequencies split by multiple spellings.

• Knock-on effect on key words (Baron et al., 2009), key word clusters (Palander-Collin & Hakala, 2011) and collocates.
The need for normalisation ...

• Automatic semantic analysis of EmodE corpora

• Automatic POS tagging of historical corpora
Development of VARD ...

• Use of existing spell checking techniques

• Hybrid methods
VARD (VARiant Detector)

http://ucrel.lancs.ac.uk/vard/
• Freely available for academic use: [http://ucrel.lancs.ac.uk/vard](http://ucrel.lancs.ac.uk/vard)

• Designed to assist researchers in standardising spelling variation in historical corpora both manually and automatically.

• Uses methods from modern spellchecking to find spelling variants and offer/select appropriate modern equivalents.

• The original spelling is always retained in the text with an xml tag surrounding the replacement.

  – <normalised orig="charitie">charity</normalised>

• Allows for the use of standard corpus linguistics tools without any modification.

• Used to normalise released historical (and other) corpora, e.g. EMEMT (Lehto *et al.*, 2010) and CEEC (Palander-Collin & Hakala, 2011).
Wider aim
(re spelling normalization)

• Determining the feasibility of developing normalisation guidelines that are generalisable to other historical corpora such as ARCHER (A Representative Corpus of Historical English Registers) and EEBO (Early English Books Online).

--- Hence we will illustrate some comparisons with the normalisation decisions made in respect to Early Modern English Medical Texts (see Lehto et al. 2010).
Samuels project (wider context)

• SAMUELS: Semantic Annotation and Mark-Up for Enhancing Lexical Searches
  – funded by the Arts and Humanities Research Council in conjunction with the Economic and Social Research Council (grant reference AH/L010062/1)
  – January 2014 to April 2015

• Aim
  – deliver a system for automatically annotating words in texts with their precise meanings, disambiguating between possible meanings of the same word
  – will provide for each word in a text the Historical Thesaurus of English reference code for that concept.

• Project team:
  – University of Glasgow (lead institution), Lancaster University, University of Huddersfield, University of Central Lancashire, University of Strathclyde, Oxford University Press
  – international partners: Brigham Young University (Utah), Åbo Akademi University (Finland), and the University of Oulu (Finland).

http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/
Manual training process

... INVOLVES THE USER:

• Reading a given text, via the VARD interface.
• Distinguishing variants within the text – via the tool’s recommended list of (ranked) candidate replacements – or personally – by highlighting variant forms manually.
• Choosing the most appropriate normalized form for each variant found – where relevant, being guided by the VARD’s known variant list or f-score calculation (derived from, e.g., letter replacement rules, edit distance measures and/or phonetic matching algorithms).
• Replacing the variant with the normalised form – but in such a way that the original spelling is retained in an XML tag (Baron and Rayson, 2008).
Argument for – and against – normalisation (a summary)

Helps improve automated techniques (e.g., POS and keyword analysis), thereby allowing existing linguistic tools to be used unmodified (see, e.g., Archer et al. 2003; Rayson et al. 2007a/b; Rayson et al. 2009; Hiltunen and Tyrkkö 2013).

We can still get to/retrieve original spellings [in the XML tag – see (iv)]

POTENTIAL LOSS OF IMPORTANT MORPHOSYNTACTIC OR ORTHOGRAPHIC INFORMATION

OPTION OF MAINTAINING SOME ORIGINAL SPELLINGS

= Normalisation process has to be handled sensitively ...
A Corpus of English Dialogues (1560-1760)

• Compiled by Merja Kytö and Jonathan Culpeper, in collaboration with Dawn Archer and Terry Walker

• Contains speech-related texts representative of five genres – trial proceedings, witness depositions, comedy dramas, prose fiction and handbooks – plus a miscellaneous category

• Published 2006; total of 1,157,720 words; 870,240 words coded for direct speech

• Grants: Swedish Research Council, Arts and Humanities Research Board, British Academy
Training set

- CED: 1,157,720 words = 177 files
- A cross cut of CED: target 30,000 words
  - 25 files totalling 213,256 words
  - 25 x 1200 words = \textcolor{red}{30,213} words
  - = 5 files from each of the five 40-year subperiods
    - trials, depositions, drama comedy, fiction, handbooks (one of each per a subperiod)
- Cf. EMEMT: training data 36,000 words (2 mwd, 450 texts/samples, 1500-1700)
Decisions made in respect to:

- Leave as is ... (with caveats)
- Modify ... (to modern form)

  ... Names
  ... Archaisms/rare/obsolete terms
  ... Foreign terms
  ... Dialect terms
  ... Personal pronouns

  ... Genitive
  ... Auxiliaries
  ... Verbs
  ... Compounds
  ... Contractions
  ... Tilde (& other graphemes)
Decision = “leave as is” (with caveats)

- Names
  - e.g., Darbye, North Baiely

- Archaic/obsolete/rare terms – normalized to one variant form
  - e.g., afore, cozen/ed, oft, morrow

- Latinate/foreign terms/dialect terms – standardized
  - e.g., birlady > byr’lady

- Personal pronouns – standardized (cf. modernized)
  - e.g., thyne > thine
Genitive

**的重要性 of distinguishing genitive from plural

my sonne sonnes  >  my son’s son
then may you well say, seeing my race is so profitably increased, that good fat oxe, and that same large eard asse are my sonne sonnes, that caulfe with a white face is his faire daughter, (D1CCHAPM)

my mistres eyes  >  my mistress's eyes
• [$ (^Lab.^) $] Talke not to me of creame, for such vaine meate I do despise as food, my stomach dies drowned in the cream boules of my mistres eyes.
  (D1CCHAPM)
Other uses of apostrophe

giue's  >  give us

Ille  >  I'll
Auxiliaries and verbs

- **t** > **ed**
- **'d** > **ed**
- **th** > **to change** (except in case of doth/hath [as plural])
- **st** > **-** (e.g., wouldst, wouldest, would'st > **would**)

**laught** > **laughed**

at this the King **laught**,

> at this the King <normalised orig="laught" auto="false">laughed</normalised> (D2FARMIN)
CED examples (cont.)

Then she desired the following Witnesses might be call'd in her Defence.

> Then she desired the following Witnesses might be <normalised orig="call'd" auto="false">called</normalised> in her Defence.

(D5WBLAND)

but thus you see the Duke<br>
confesseth the receipt of the Letter

> but thus you see the Duke <normalised orig="confesseth" auto="false">confesses</normalised> the receipt of the Letter

(D1TNORFO)
Auxiliaries and verbs (cont.)

- shew/s/ed > show / shows / showed
- didst > did
- dost > do
Compounds

** split or divide as in PDE; but leave problematic cases

my self > *myself*

any way > *anyway*

Pray don't trouble *your self* on my Account.

> Pray don't trouble <join original="your self">*yourself*</join> on my Account.  

(D5HGBEIL)
CED examples (cont.)

to morrow > tomorrow
And, if you please, to morrow we shall begin.
> And, if you please, <join original="to morrow">tomorrow</join> we shall begin.

(D4HEMIEG)

an other > another
It shalbe then for an other tyme.
> It <normalised orig="shalbe" auto="false">shall be</normalised> then for <join original="an other">another</join> <normalised orig="tyme" auto="false">time</normalised>.

(D1HEBELL)
## Contractions

**normalise where we know the full form**

<table>
<thead>
<tr>
<th>Short Form</th>
<th>Normalised Form</th>
<th>Original Form</th>
<th>Normalised Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘em</td>
<td>them</td>
<td>tis or 'tis</td>
<td>it’s</td>
</tr>
<tr>
<td>for’it</td>
<td>for it</td>
<td>twas, t’was</td>
<td>it was</td>
</tr>
<tr>
<td>igad</td>
<td>i’gad</td>
<td>twill, t’will</td>
<td>it'ill</td>
</tr>
<tr>
<td>on’t</td>
<td>on it</td>
<td>qd</td>
<td>quod</td>
</tr>
<tr>
<td>sblood</td>
<td>s’blood</td>
<td>weel(e)</td>
<td>we’ll</td>
</tr>
<tr>
<td>sha’n’t</td>
<td>shan’t</td>
<td>wy</td>
<td>with you</td>
</tr>
<tr>
<td>tho</td>
<td>though</td>
<td>y’are</td>
<td>you’re</td>
</tr>
<tr>
<td>til</td>
<td>till</td>
<td>yfaith, yfayth, ifaith</td>
<td>i'faith</td>
</tr>
</tbody>
</table>
Tilde

• the~ > to full form (according to context)
• dispositio~ > disposition (etc.)

Let vs begin the~.

Let <normalised orig="vs" auto="false">us</normalised> begin
<normalised orig="the~" auto="false">then</normalised>.

(D1HEBELL)

But you dealt all to the~.

But you dealt all to <normalised orig="the~" auto="false">them</normalised>.

(D1HEBELL)
Importance of context when making decisions

Context list
• bee/be
• doe/do
• the/thee
• then/than
• too/to
• y=t=/that
CED examples (cont.)

bee > be
the more it is to bee feared?
> the more it is to <normalised orig="bee" auto="false">be</normalised> feared? (D2FARMIN)

doe > do
What to doe?
> What to <normalised orig="doe" auto="false">do</normalised>? (D1HEBELL)
the > thee
and make the spend all thie meanes.

> and make <normalised orig="the" auto="false">thee</normalised> spend <normalised orig="thie" auto="false">thy</normalised> whole estate" (D2WDIOCE)
CED examples (cont.)

then > than
Excuse me, Sir, I understand it more then I do high German.
> Excuse me, Sir, I understand it more <normalised orig="then" auto="false"> than </normalised> I do high German." (D3HFFEST)

too > to
in good faith you are too blame
> in good faith you are <normalised orig="too" auto="false"> to </normalised> blame [...]" (D1CHAPM)
CED examples (cont.)

y=t= > that
hir husbande said diuers times y=t= he would cut it of,

> <normalised orig="hir" auto="false">her</normalised>
<normalised orig="husbande" auto="false">husband</normalised> said <normalised orig="diuers" auto="false">divers</normalised> times <normalised orig="y=t=" auto="false">that</normalised> he would cut it <normalised orig="of" auto="false">off</normalised>,
To conclude

• Normalisation guidelines are a compromise ... 😊

• Important to combine automatic processing and manual screening

• VARD2 to be applied to EEBO shortly (in the Samuels project)

• Acknowledgements
  – Samuels project; AHRC grant reference AH/L010062/1
    http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/
  – University of Uppsala travel grant
  – Thanks to Terry Walker and Gerold Schneider for acting as VARDers on the CED in June 2013
References


A Corpus of English Dialogues 1560-1760. (2006). Compiled under the supervision of Merja Kytö (Uppsala University) and Jonathan Culpeper (Lancaster University).


