Some problems in corpus-oriented English historical syntax

David Denison
English Research Association of Hiroshima
University of Hiroshima
3 December 2016

Why make historical corpora?
- Mainly for linguistic research (of many kinds)
- May also arise from – or be of use in – some other research domain (archival, historical, sociological, literary, stylistic, …)
  - cf. Image to Text (Hamilton Papers)
- Ideally, balanced and/or representative or complete
- but there may be practical limitations
- Reliable
- Useful

Plan of talk: two sections
1. Problems for corpus makers
2. Problems for corpus users
case studies in the history of English

Register / genre / text-type
- Relevant to lexis, but also to syntax
  - e.g. information packaging in News genre (VP > NP)
- Boundaries often fuzzy whichever criteria chosen
  - purpose, audience, medium, linguistic properties
  - e.g. is Medical sub-type of Scientific?
- ARCHER 3.1 > 3.2: distinguish Journals from Diaries
  - historical drift of register variation shows in preposition placement in possible stranding constructions
- 3.1 > 3.2 > 3.3: dialogue vs authorial voice (fiction) or stage directions (drama)

Representativeness & balance
- ARCHER 3.1
  - 1650-1999, 8 genres, 1.8m words
  - broadly similar word counts for periods
  - many gaps, especially in American English
- ARCHER 3.2
  - 1600-1999, 12 genres, 3.2m words
  - some gaps filled (and more will be filled in v3 3.3), but word counts more uneven: need to normalise frequencies
  - only 3 genres 1600-1649, one of them not used 1650-
- Balance vs. size/comprehensiveness!
**Foreign insertions**

I think if I can work that incident up a little it will form a very fitting dénoument to my unhappy ‘Mme de V.’ wh: (en passant) I may mention is likely to be fair copied about the A.D. 1900. This must stand, mon cher, for the Sunday edition & entreats an answer.

(1890 Ernest Dowson, lModE Prose)

**How foreign? How incorrect?**

inshort it has broke into all my plans of occupation & has unhinged me quite -- I feel dissipated, without an object or a pursuit which is worse than death to me

(1786 HAM/1/15/1/17)

**Correct the text?**

There are a set of Savages that are employed in making the new Road who are Strangers & earn prodigious Wages & live in on extraordinary eating raw bacon & undressed meat & drinking such Quantities of liquor as is scarce credible

(1813 HAM/1/2/47)

**Correct reading?**

Since my arrival here I have been in a bit of a passion but as I am quiet now I will endeavour w~ Composure to tell you that there is not a Soul in this town who ever heard of Oakford -- I have been to the post Office -- Never heard of the place -- I have spoken to ?? in the street -- knew no such place -- The Landlord says he has a post Boy who must know it but he will not return till later so here have I been fuming & admiring all my Relations as the most discreet sensible people in the World -- to invite me to come 200 miles to pay them a visit to direct me to go to Town & from there to take a Post Chaise & not know where to direct the Driver to direct his horse

(1813 HAM/1/2/43)

**Correct reading?**

• Gunmen? Gentlemen? Yeomen?

• Gemmen = vulgar pronunciation of gentlemen [OED]
Original spelling in Image to Text
- Contemporary spellings retained but usually tagged: 
  - *chuse*, *expence*, *wishd*, etc. [but *wou’d* not tagged!]
  - **XML tag**: `<orig>chuse</orig>`
- Tag doesn’t show on screen or in plain text.
- ‘Errors’ retained but corrected on mouseover:
  - *endeavour'd* — *endeavoured*
  - Standard PDE past tense or past ptcp. = *quit*, but *OED* has quotations with *quitted* to mid-20C.
- **How correct *quited***?!

Normalisation for PoS tagging
- Corpora such as ARCHER (1600-1999) usually normalise text to PDE spelling for tagger and parser:
  - e.g. with VARD2
  - but contradicts original spelling retention policy
- 2 sets of normalisation tags, or 2 versions of corpus?
- **Ideally (?), search** using normalised PDE spelling:
  - so that *wou’d*, *woud* found automatically with *would*
  - but **display** hits as written in original.

Part of Speech tagging
- Highly useful for syntactic research.
- Prerequisite for parsing – makes more subtle searches possible.
- Tagging done by software, ideally then corrected.
  - ARCHER uses CLAWS tagger, then ‘Template Tagger’ (as BNC, but unlike COHA, COCA).
  - Small corpora such as Penn parsed corpora can even be hand-corrected.

Consistency vs. accuracy
Mr. Kenyon Parker, Q.C. [...] was run over by a Hansom cab yesterday afternoon in Chancery-lane, and seriously injured. (1866 ARCHER)
- Tagging by Nick Smith treats *run over* as a phrasal verb here, as in PDE (cf. the cab *run him over*):
  - [VBDZ was] [VBN run] [RP over] by a Hansom cab
- **Over** is tagged as an adverbial particle
Fun and key in ARCHER

- These and many others in recent decades are N at first, later can be Adj as well.
- Cf. so fun, very key, it was key, etc. in recent corpora.
- Transition is stepwise.
- For speakers with both N and Adj possibilities, some instances are underspecified or vague.

After the first few times this was fun. (1939)
It was fun in bed (1971)
Farm work was fun in the spring [...] (1993)

Fun and key in ARCHER

The key verse in this first section is verse4: it is a crucial one. (1959)
But the key foreign and defense portfolios remained unchanged. (1980)
This is a key post for an experienced worker (1995)

- Three relevant periods, and accurate tagging must be different in each (therefore not consistent):
  - when word could only be N
  - when it might be N–Adj or Adj
  - when it can be unequivocally Adj

Pooled data

- Search hits come from multiple speakers/writers.
- Stepwise nature of some PoS changes implies layers of advanced vs. conservative speakers at any one time:
  - Differential acceptability of new patterns and the stepwise nature of change point to grammatical variation within a population [...] corpus techniques which rely on pooled data can only give limited insight.
- Individual speakers tracked in NY Times Annotated Corpus, Corpus of English Novels, Hansard.

Large corpora

- Hand-correction by corpus makers impossible
- Low-frequency features can be found by users.
- Do greater absolute frequencies mean that false positives or negatives will not distort results too badly?
- While megacorpora like COHA and COCA are reasonably balanced within (limited) genres and dates, gigacorpora like GLoWbE and Google Books are less controlled.

2. Problems for corpus users

- case studies in the history of English

Date

- User must be sensitive to chronological layers within ‘synchronic’ data, even within individual usage.
- Linguistic fossils
  - My dear Sherwood, How goes it (1927 ARCHER)
- Linguistic playfulness
  - Geniuser than the geniusest of the geniusest, dahlink! (2003)
  - Speaker/writer’s date of birth may be as pertinent as date of utterance/publication.
    - cf. work done using CEEC
**That-clause complements**

1. **Hundt’s study** (2009), which **advocates** that the subjunctive is in fact replacing the periphrastic [...]

2. **this highlights** once more that [...]

3. with Puousa **criticising** that the French influence was sporadic

4. **Sweet defines** that “grammar may be regarded either from a theoretical or practical point of view. [...]

5. This study has **displayed** that older participants have more stable and confident results than [...]

6. **Follet** (1966) [...] **poses** that the informality of try and leads to [...]

7. **Steven Pinker**, (1994) **puts forward** that chimps often just imitate the messages of the trainer

8. which can be reinforced by Milroy et al, who **utters** that “In other locations [...]

**V + that-clause in student work**

1. **Hundt’s study** (2009), which **advocates that** the subjunctive is in fact replacing the periphrastic [...]

2. **this highlights** once more that [...]

3. with Puousa **criticising** that the French influence was sporadic

4. **Sweet defines** that “grammar may be regarded either from a theoretical or practical point of view. [...]

5. This study has **displayed that** older participants have more stable and confident results than [...]

6. **Follet** (1966) [...] **poses that** the informality of try and leads to [...]

7. **Steven Pinker**, (1994) **puts forward that** chimps often just imitate the messages of the trainer

8. which can be reinforced by Milroy et al, who **utters that** “In other locations [...]

**V + that-clause or V + N-that?**

You have to **accept that** this could happen again. (2015, COCA)

If you just **accept the fact that** there’s no self [...] (2007, COCA)

*The aforementioned authors **espouse that** students from the age of four to eight are aware of racial difference (2011, COCA)

Poland also **espoused the idea that** the COMECON Members should [...] (1990, COCA)

**Factual and suasive verbs**

<table>
<thead>
<tr>
<th>verb</th>
<th>N that</th>
<th>that</th>
</tr>
</thead>
<tbody>
<tr>
<td>accept, acknowledge, add, affirm, allege, allow, announce, assert, assume, believe, (claim), concede, confess, consider, consent, (decide), (determine), demonstrate, deny, disclose, discover, doubt, emphasise, establish, explain, (find), forget, guarantee, hold, imagine, (indicate), infer, (judge), maintain, maintain, observe, (point out), (predict), (presume), (pronounce), propose, prove, recite, repair, report, see, (show), state, (stipulate), stress, submit, suppose, (suggest), support, understand</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>verb</th>
<th>N that</th>
<th>that</th>
</tr>
</thead>
<tbody>
<tr>
<td>advance, articulate, back up, challenge, communicate, contradict, convey, define, discuss, dispute, encourage, endorse, entitle upon, espouse, express, oppose, promote, put across, put forward, question, rule, support, sustain, underline, underscore, utter</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Error vs. innovation**

- **Distinction crucial to Kachru’s concentric circles model.**

- In historical linguistics, some errors turn out (with hindsight) to be innovations.

- The sporadic occurrence of ‘new’ V-that patterns has affinities with learner English and with new Englishes.

- These are **native speakers** using (misusing?) words and patterns **in writing** that would be rare or non-existent in their everyday conversation.

**Erroneous usage?**

*... communication verbs controlling that-clauses (apart from say) are most frequent in academic prose*’

- Such verbs are needed to avoid risk of plagiarism.

- Word processor thesaurus for ‘elegant variation’?

- Students in question tend to be relatively unskilled writers, insecure about written expression.

- Asking them about grammaticality not helpful.

- Once written down, usage can get entrenched.

**Denison (2011)**

**Biber et al. (1999: 668)**

**V + that-clause in student work**

1. **Hundt’s study** (2009), which **advocates that** the subjunctive is in fact replacing the periphrastic [...]

2. **this highlights** once more that [...]

3. with Puousa **criticising** that the French influence was sporadic

4. **Sweet defines** that “grammar may be regarded either from a theoretical or practical point of view. [...]

5. This study has **displayed that** older participants have more stable and confident results than [...]

6. **Follet** (1966) [...] **poses that** the informality of try and leads to [...]

7. **Steven Pinker**, (1994) **puts forward that** chimps often just imitate the messages of the trainer

8. which can be reinforced by Milroy et al, who **utters that** “In other locations [...]

**V + that-clause or V + N-that?**

You have to **accept that** this could happen again. (2015, COCA)

If you just **accept the fact that** there’s no self [...] (2007, COCA)

*The aforementioned authors **espouse that** students from the age of four to eight are aware of racial difference (2011, COCA)

Poland also **espoused the idea that** the COMECON Members should [...] (1990, COCA)

**Factual and suasive verbs**

<table>
<thead>
<tr>
<th>verb</th>
<th>N that</th>
<th>that</th>
</tr>
</thead>
<tbody>
<tr>
<td>accept, acknowledge, add, affirm, allege, allow, announce, assert, assume, believe, (claim), concede, confess, consider, consent, (decide), (determine), demonstrate, deny, disclose, discover, doubt, emphasise, establish, explain, (find), forget, guarantee, hold, imagine, (indicate), infer, (judge), maintain, maintain, observe, (point out), (predict), (presume), (pronounce), propose, prove, recite, repair, report, see, (show), state, (stipulate), stress, submit, suppose, (suggest), support, understand</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>verb</th>
<th>N that</th>
<th>that</th>
</tr>
</thead>
<tbody>
<tr>
<td>advance, articulate, back up, challenge, communicate, contradict, convey, define, discuss, dispute, encourage, endorse, entitle upon, espouse, express, oppose, promote, put across, put forward, question, rule, support, sustain, underline, underscore, utter</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Error vs. innovation**

- **Distinction crucial to Kachru’s concentric circles model.**

- In historical linguistics, some errors turn out (with hindsight) to be innovations.

- The sporadic occurrence of ‘new’ V-that patterns has affinities with learner English and with new Englishes.

- These are **native speakers** using (misusing?) words and patterns **in writing** that would be rare or non-existent in their everyday conversation.

**Erroneous usage?**

*... communication verbs controlling that-clauses (apart from say) are most frequent in academic prose*’

- Such verbs are needed to avoid risk of plagiarism.

- Word processor thesaurus for ‘elegant variation’?

- Students in question tend to be relatively unskilled writers, insecure about written expression.

- Asking them about grammaticality not helpful.

- Once written down, usage can get entrenched.
Systematic research in corpora

- Tagged and lemmatised corpus distinguishes e.g. advance, display V ~ advance, display N
- that CONJ ~ that D
- Search COHA for \([\text{display}],[v^*] \text{ that}\).\[\text{cst}\]
- Accuracy 5/41 = 12% \(\text{display} = N \times 18\), \text{ that} = D \times 18\)
- Perhaps tagger trained on data without marginal examples
- And then only 1/5 somewhat relevant!
  but all this accomplished was to display that the poor creature’s teeth have been yanked out (logCOHA)

Take long

Word class

- Straightforward uses of long are
  - Adj
    - The road is very long
    - a long night
  - Adv
    - He won’t live long
    - It lasted too long
  - What about
    - It won’t be long.  Adv or Adj?
    - I won’t be long.  Adv or Adj?
    - It won’t take long.  Adv or N or Adj? (cf. a long time)
    - I won’t take long.  Adv or N or Adj?
    - the whole night long  Adv or P? (cf. the whole night through)

Penn parsed corpora

- All instances of long are tagged as Adj in Penn parsing scheme.
  - except 14 instances in PPCME2 inexplicably tagged as Adv
- But YCOE differs and maintains conventional Adj ~ Adv distinction for lang/long, etc.
- Consistent tag scheme for long not tenable over whole span of English OE-1914.
- PPCME2, PPCME and PPCMBE distinguish uses of long not by tagging but by parsing.

long in earliest Penn corpora

<table>
<thead>
<tr>
<th>corpus</th>
<th>tagged as Adj</th>
<th>tagged as Adv</th>
<th>totals</th>
<th>Adj</th>
<th>Adv</th>
<th>unclear</th>
<th>totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>YCOE</td>
<td>341</td>
<td>594</td>
<td>935</td>
<td>329</td>
<td>590</td>
<td>16</td>
<td>935</td>
</tr>
<tr>
<td>PPCME2</td>
<td>735</td>
<td>14</td>
<td>749</td>
<td>262</td>
<td>428</td>
<td>59</td>
<td>749</td>
</tr>
<tr>
<td>totals</td>
<td>1076</td>
<td>608</td>
<td>1684</td>
<td>591</td>
<td>1018</td>
<td>75</td>
<td>1684</td>
</tr>
</tbody>
</table>

An unreal conditional
The Madonna of George III

Indirect quotation
IN THIS third collection of excerpts from his diaries [...] Alan Bennett complains that people see him as “cosy and essentially harmless”. Even if he stabbed Dame Judi Dench with a pitchfork he would, he hazards, still be reckoned a teddy bear.

(Sunday Times, 16 Oct 2016)

The Madonna of King George

Indirect quotation

Even if he stabbed Dame Judi Dench with a pitchfork

• If-clause protasis, past stabbed

‘Direct quotation’
“I am in the pigeonhole marked ‘no threat’ and were I to stab Judi Dench with a pitchfork I should still be a teddy bear,” he writes at the end of 2005. He worries that his work – that he – is considered cosy.

(Miranda Sawyer, The Guardian, 2 Oct 2016)
(clause also London Review Bookshop website, 7 Oct 2016)
‘Direct quotation’

\[
\text{were I to stab Judi Dench with a pitchfork}
\]

- Inverted protasis, past subjunctive were

\[
\text{Original}

(Book of the Week, BBC Radio 4, 27 Oct 2016)

I shall still be thought to be kindly, cozy and essentially harmless. I am in the pigeon-hole marked ‘no threat’ and did I stab Judi Dench with a pitchfork I should still be a teddy bear.

(2016 Alan Bennett, Keeping on Keeping On, entry for 20 Dec 2007)

\[
\text{Original}

\text{did I stab Judi Dench with a pitchfork} \text{ I should still be a teddy bear.}
\]

- Inverted protasis, past did

\[
\text{Inversion protases in PDE}

\bullet \text{Overall frequency has declined through ModE period}
\bullet \text{Fewer verbs now invert there: had, were, should, ...}
\bullet \text{What about did?}
  \bullet \text{Biber et al. (1999: 851-3, 919): No}
  \bullet \text{Quirk et al. (1985: 1084): not mentioned, implicit No}
  \bullet \text{Huddleston & Pullum (2002: 793, 976): not mentioned}
  \bullet \text{journalist or subeditor on The Guardian: ?No}
  \bullet \text{Denison (1998: 299-300): examples up to 1993}
  \bullet \text{Visser (1963-73: 767): 'nowadays only in literary style'}

\[
\text{Is did I protasis grammatical?}
\]

- Depends not just on speaker but on register. ‘Passive’ grammaticality not same as ability to use.
- Almost as if Alan Bennett is wilfully retaining or reviving a dying usage. Does same for could:
  
  \text{Could I slip into the seat behind, I would put a hand on my young shoulder and say, ‘It’s going to be all right’. (ibid. 2014)}

  \bullet \text{Would he use these in everyday conversation?}
  \bullet \text{Grammaticality not either-or}
  \bullet \text{Quoted speech in journalism frequently not accurate}

\[
\text{Concluding remarks}
\]
Marginal problems?
- Problems for corpus maker sometimes not so serious
  - situations infrequent
  - usefulness more important than own ‘nerdiness’
- Problems for corpus user more often serious, but
  - esp. for historical corpora, corpus ≠ language of time
  - tagging only practical aid, not publishable analysis
  - awareness of corpus limitations is halfway to solution
- Historical linguists hope to discover change.
  - Not surprising if change not yet reflected in mark-up.

Prospects / desiderata
- Use of TEI, XML and XSLT to present alternative views of material according to needs of viewer
- Multiple mark-up schemes for different purposes, e.g. by using stand-off mark-up
- Tags that explicitly signal indeterminacy between two categories; could be like an ambiguity tag in form
  - Ajo-VVG and VVG-Ajo
  - Ajo-NN and NN-Ajo (BNC)
- More spoken corpora, transcribed and searchable
  - audio aligned with transcription
  - already many decades of recordings, so diachronic too

JSPS Fellowship
- I gratefully acknowledge funding from the JSPS Invitation Fellowship Programme for Research in Japan (short term) which has made this lecture possible
- and I warmly thank Professor Fujio Nakamura for organising the JSPS application and resultant visit.

Last slide!
- Presentation can be downloaded from http://tinyurl.com/DD-download
- Comments welcome!
  Domo arigato gozaimashita
References
(but not general introductions to or surveys of corpus linguistics)


Denison, David. 2007. Playing tag with category boundaries. VARIENG e-Series 1, Annotating variation and change (Proceedings of ICAME 27 Annotation Workshop). Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG). http://www.helsinki.fi/varieng/journal/volumes/01/denison/


Corpora cited


I18C Prose = Denison, David & Linda van Bergen. 2002. A Corpus of late 18c Prose. Manchester. [see http://personalpages.manchester.ac.uk/staff/david.denison/late18c]
