MANCHESTER 1824

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

2

Plan of talk: two sections

1. Problems for corpus makers
2. Problems for corpus users
   case studies in the history of English

3

1. Problems for corpus makers

4

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)

Biber (2001), Yáñez-Bouza (2015, 2016)

5

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 vsn 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!

6

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)

probably foreign

probably foreign

probably English

9

Foreign or not?

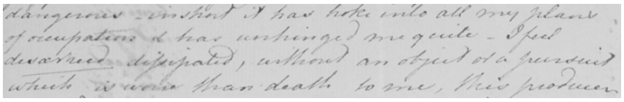
I was appointed Lectrice to the society and every morning read a French Drama or story of some kind, loud to a very attentive audience, from which I generally drew tears, for the choice of the lecture being left to me, you may imagine it was of a serious or affecting nature (1786 HAM/1/15/1/17)

```
<hi rend="underlined"><foreign xml:lang="fr">lectrice
</foreign></hi>
<hi rend="underlined"><foreign xml:lang="fr">lecture
</foreign></hi>
```

10

How foreign? How incorrect?

inshort it has broke into all my plans of occupation & has unhinged me quite -- I feel desœuvreed dissipated, without an object or a pursuit which is worse than death to me (1787 HAM/1/15/1/20)



```
<hi rend="underlined"><foreign xml:lang="fr"><sic
corr="désœuvrée">desœuvre</sic></foreign>ed</hi>
```

11

Correct the text?

There are a set a 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)

- *a set a savages*: should we correct to *a set of savages*?
- No: in *OED* not a form of *of* but form in its own right, exactly as here on page

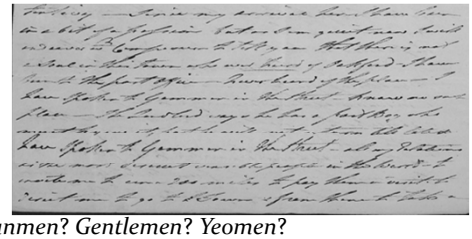
12

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^h^- 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)

13

Correct reading?



- *Gunmen*? *Gentlemen*? *Yeomen*?
- <gap>
- *Gemmen* = vulgar pronunciation of *gentlemen* [*OED*]

14

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:
  would, performed, endeavourd > *endeavoured*
- We hesitantly retained *quitted* but corrected *quited*
  - Standard PDE past tense or past ptcp. = *quit*, but *OED* has quotations with *quitted* to mid-20C.
  - How correct *quited*?! > *quitted*

15

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.

Baron & Rayson (2008)

16

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.

17

Part of Speech tagging

- Some tagsets include 'ambiguity tags'.✓
- If mark multiwords,✓exact extent may be problematic:
  - *on behalf of* in pre-20C texts: compound prep. or not?
  - idiom *a nasty piece of work* for semantic tagging
    - *piece of work*
    - *a piece of work*
    - *a* ([optional intensifier]) [pejorative adjective] *piece of work*
    - *a* (...) *nasty piece of work*
- Problems of consistency over a long time-span:
  - PoS can change, multiword sequences can lexicalise.

Denison (2007, 2013a)

18

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 ran him over*):

[$_{VABDZ}$ was] [$_{VVN}$ run] [$_{RP}$ over] by a Hansom cab

- *Over* is tagged as an adverbial particle

19

Consistency vs. accuracy

- But *run over* was clearly a prepositional verb in origin:

don't you start till we are nearly at the bottom, or you will run over us and break our necks (1873 ARCHER)

- *Over* must be a preposition here.
- Clear phrasal verb *run over* only emerges in mid-20C.
- Tagging of underdetermined cases cannot be both consistent and accurate if period includes 1840-1960.
- Passive of *run over* was **vague** (not ambiguous) in syntax from mid-19C to mid-20C. (?)

Denison (in press)

20

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)

Denison (2013b, in press)

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.

Denison (2013b, in press), de Smet (2016)

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! (2005)
- Speaker/writer's date of birth may be as pertinent as date of utterance/publication.
  - cf. work done using CEEC

That-clause complements

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 Poussa **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)

Denison (2011)

Factual and suasive verbs

verb	N-*that*	*that*
accept, acknowledge, add, affirm, allege, allow, announce, assert, assume, believe, (claim), concede, confirm, consider, convey, (deduce), (determine), demonstrate, deny, disclose, discover, doubt, emphasise, establish, explain, (find), forget, guarantee, hold, imagine, (indicate), infer, (judge), maintain, mention, observe, (point out), (predict), (presume), ?pronounce, propose, prove, recognise, regret, repeat, report, see, (show), state, (stipulate), stress, submit, suggest, (suppose), suspect, understand	✓	✓
advance, articulate**, back up, challenge, communicate, contradict, convey, **define**, discuss, dispute#, encourage, endorse, enlarge upon, espouse, **express**, oppose, promote, put across, **put forward**, question#, rule, support, sustain, underline, **underscore**, **utter	✓	×
advocate, analyse, bring to the surface, cite, clarify, contest, criticise, deem [OK?], deliberate, depict, describe, display, exemplify, explicate, highlight, identify, illustrate, inform, instigate, interpret, moot, portray, pose, posit [OK?], propound, publicise, quote, reflect, refute, reinforce, reiterate, respect, rule out, solidify, stand, summarise, take into account, uncover, update, view, welcome, yield	?	?

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.

Biber et al. (1999: 668)

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.

Hundt & Mukherjee (2011)

Systematic research in corpora

- Tagged and lemmatised corpus distinguishes e.g.
  - *advance, display* V ~ *advance, display* N
  - *that* CONJ ~ *that* D
- Search COHA for **[display].[v*] that.[cst]**
- Accuracy 5/41 = 12% (*display* = N ×18, *that* = D ×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 (1990 COHA)

33

Take long

34

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*)

35

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, PPCEME and PPCMBE distinguish uses of *long* not by tagging but by parsing.

Santorini (2010)

36

long in earliest Penn corpora

corpus	tagged as Adj	tagged as Adv	totals		Adj	Adv	unclear	totals
YCOE	341	594	935		329	590	16	935
PPCME2	735	14	749		262	428	59	749
totals	1076	608	1684		591	1018	75	1684

37

An unreal conditional

Two 'national treasures'

Alan Bennett

Judi Dench

39

The Madness of George III

40

The Madness of King George

41

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)

42

Indirect quotation

Even if he stabbed Dame Judi Dench with a pitchfork

- *If*-clause protasis, past *stabbed*

43

'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, ?15 Oct 2016)

44

‘Direct quotation’

<u>were I to stab</u> Judi Dench with a pitchfork

- Inverted protasis, past subjunctive *were*

45

Original

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

I shall still be thought to be kindly, cosy and essentially harmless. I am in the pigeon-hole marked “no threat” and **<u>did I stab</u> Judi Dench with a pitchfork** I should still be a teddy bear.

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

46

Original

<u>did I stab</u> Judi Dench with a pitchfork I should still be a teddy bear.

- Inverted protasis, past *did*

47

Inversion protases in PDE

- Overall frequency has declined through ModE period
- Fewer verbs now invert there: *had, were, should, …*
- What about *did*?
  - Biber et al. (1999: 851-3, 919): No
  - Quirk et al. (1985: 1084): not mentioned, implicit No
  - Huddleston & Pullum (2002: 753, 970): not mentioned
  - journalist or subeditor on *The Guardian*: ?No
  - Denison (1998: 299-300): examples up to 1993
  - Visser (1963-73: 767): ‘nowadays only in literary style’

48

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*:

  <u>Could I slip into the seat behind</u>, I would put a hand on my young shoulder and say, ‘It’s going to be all right’. (*ibid.* 2014)
- Would he use these in everyday conversation?
- Grammaticality not either-or
- Quoted speech in journalism frequently **not** accurate

49

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.

51

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
  - AJ0-VVG and VVG-AJ0
  - AJ0-NN1 and NN1-AJ0 (BNC)
- More spoken corpora, transcribed and searchable
  - audio aligned with transcription
  - already many decades of recordings, so diachronic too

52

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.

53

Last slide!

- Presentation can be downloaded from

  http://tinyurl.com/DD-download

- Comments welcome!

Domo arigato gozaimashita

54

References

(but not general introductions to or surveys of corpus linguistics)

Baron, Alistair and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. Proceedings of the Postgraduate Conference in Corpus Linguistics, Aston University, Birmingham, UK, 22 May 2008. http://eprints.lancs.ac.uk/41666/1/BaronRaysonAston2008.pdf

Biber, Douglas. 2001. Dimensions of variation among eighteenth-century speech-based and written registers. In Susan Conrad & Douglas Biber (eds.), *Variation in English: Multi-dimensional studies*, 200–14. London: Longman.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson.

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/

Denison, David. 2013a. Grammatical mark-up: Some more demarcation disputes. In Paul Bennett, Martin Durrell, Silke Scheible & Richard J. Whitt (eds.), *New methods in historical corpora* (Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP) 3), 17-35. Tübingen: Narr.

Denison, David. 2013b. Parts of speech: Solid citizens or slippery customers? *Journal of the British Academy* 1, 151-85.

Denison, David. in press. Ambiguity and vagueness in historical change. In Marianne Hundt, Sandra Mollin & Simone Pfenninger (eds.), *The changing English language: Psycholinguistic perspectives* (Studies in English Language). Cambridge: Cambridge University Press.

Denison, David. to appear. Why would anyone *take long*? Word classes and Construction Grammar in the history of *long*. In Kristel Van Goethem, Muriel Norde, Evie Coussé & Gudrun Vanderbauwhede (eds.), *Category change from a constructional perspective* (Constructional Approaches to Language). Amsterdam: John Benjamins.

De Smet, Hendrik. 2016. How gradual change progresses: The interaction between convention and innovation. *Language Variation and Change* 28.1, 83-102.

Hundt, Marianne & Joybrato Mukherjee. 2011. Discussion forum: New Englishes and learner Englishes - *quo vadis?* In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap* (Studies in Corpus Linguistics 44), 209-17. Amsterdam: John Benjamins.

Santorini, Beatrice. 2010. Annotation manual for the Penn Historical Corpora and the PCEEC. http://www.ling.upenn.edu/hist-corpora/annotation/index.html.

Yáñez-Bouza, Nuria. 2011. ARCHER past and present (1990-2010). *ICAME Journal* 35, 205-36.

Yáñez-Bouza, Nuria. 2015. 'Have you ever written a diary or a journal?' Diurnal Prose and Register Variation. *Neuphilologische Mitteilungen* 116.2, 449-74.

Yáñez-Bouza, Nuria. 2016. Daily jottings: Preposition placement in English diaries and travel journals from 1500 to 1900. *Folia Linguistica Historica* 37.1, 281-314.

Corpora cited

ARCHER 3.2 = A Representative Corpus of Historical English Registers version x. 1990–1993/2002/2007/2010/2013. Originally compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. Current member universities are Bamberg, Freiburg, Heidelberg, Helsinki, Lancaster, Leicester, Manchester, Michigan, Northern Arizona, Santiago de Compostela, Southern California, Trier, Uppsala, Zurich. [see http://www.projects.alc.manchester.ac.uk/archer/]

l18C Prose = Denison, David & Linda van Bergen. 2002. A Corpus of late 18c Prose. Manchester. [see http://personalpages.manchester.ac.uk/staff/david.denison/late18c]

lModE Prose = Denison, David. 1994. A Corpus of late Modern English Prose. Manchester. [see http://personalpages.manchester.ac.uk/staff/david.denison/lmode_prose.html]

Image to Text = Denison, David & Nuria Yáñez Bouza. 2016. Image to Text: Mary Hamilton Papers (c.1750-c.1820). Manchester. See http://www.projects.alc.manchester.ac.uk/image-to-text/.