Historical linguistics and big data: The “reading programme” of the future

Mark Liberman
http://ling.upenn.edu/~myl
The “Reading Programme”: Foundation of the OED since 1857

There are some other systematic surveys of English linguistic history

With two extra dimensions:
- Using counts to estimate frequencies
- Durable annotation of a corpus (“Treebanks”)

And two missing ones:
- No notion of “sense” or “lemma” beyond P.O.S.
- Smaller scale – a few million words

More recently: Efforts with larger scale but less analysis

All are now “programs” in OED sense 9 as well as 4
- 4. “a plan or scheme of any intended proceedings”
- 9. “A series of coded instructions and definitions which when fed into a computer automatically directs its operation in performing a particular task”
An example: Research on the history of “do-support”

Quarto 1 (1603):

Madam, how do you like this play?
The Lady protests too much.

Quarto 2 (1604):

Madam, how like you this play?
The Lady doth protest too much mee thinks.

Folio 1 (1623):

Madam, how like you this Play?
The Lady protests too much me thinkes.
The OED’s (1897) explanation – from **do, v.**:

*** As a *Periphrastic Auxiliary* of the present and past Indicative, and Imperative. (Formerly sometimes of the Infinitive.) (For a detailed treatment of this, see ‘Das Umschreibende Do in der Neuenglischen Prosa’ by Hugo Dietze, Jena, 1895.)


a. Originally, simply periphrastic, and equivalent to the simple tense. Found in Old English, frequent in Middle English, very frequent 1500–1700, dying out in normal prose in 18th c.; but still retained in s.w. dialects; also as an archaism in liturgical and legal use, and as a metrical resource in verse.


The periphrastic form with *do, did*, is now the normal form. Its use allows the pronoun to be placed between the auxiliary and main verb, instead of coming after the latter: e.g. ‘Did he recognize her?’ instead of ‘Recognized he her?’
Alvar Ellegård, "The Auxiliary 'Do': The Establishment and Regulation of Its Use in English", 1953:

<table>
<thead>
<tr>
<th>PERIOD</th>
<th>NEG. QUES.</th>
<th>AFF. QUES.</th>
<th>NEG. DECL.</th>
<th>AFF. DECL.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1390-1400</td>
<td>0 00.0</td>
<td>0 00.0</td>
<td>0 00.0</td>
<td>0 00.0</td>
</tr>
<tr>
<td>1400-1425</td>
<td>2 11.8</td>
<td>0 00.0</td>
<td>0 00.0</td>
<td>11 00.2</td>
</tr>
<tr>
<td>1425-1475</td>
<td>2 08.0</td>
<td>6 04.2</td>
<td>11 01.2</td>
<td>121 00.3</td>
</tr>
<tr>
<td>1475-1500</td>
<td>3 11.1</td>
<td>10 07.0</td>
<td>33 04.8</td>
<td>1059 01.8</td>
</tr>
<tr>
<td>1500-1525</td>
<td>46 59.0</td>
<td>41 22.7</td>
<td>47 07.8</td>
<td>396 01.4</td>
</tr>
<tr>
<td>1525-1535</td>
<td>34 60.7</td>
<td>33 32.4</td>
<td>89 13.7</td>
<td>494 02.6</td>
</tr>
<tr>
<td>1535-1550</td>
<td>63 75.0</td>
<td>93 44.9</td>
<td>205 27.9</td>
<td>1564 08.1</td>
</tr>
<tr>
<td>1550-1575</td>
<td>41 85.4</td>
<td>72 56.3</td>
<td>119 38.0</td>
<td>1360 09.3</td>
</tr>
<tr>
<td>1575-1600</td>
<td>83 64.8</td>
<td>228 60.3</td>
<td>150 23.8</td>
<td>1142 06.3</td>
</tr>
<tr>
<td>1600-1625</td>
<td>89 93.7</td>
<td>406 69.2</td>
<td>102 36.7</td>
<td>240 03.0</td>
</tr>
<tr>
<td>1625-1650</td>
<td>32 84.2</td>
<td>116 82.9</td>
<td>109 31.7</td>
<td>212 02.9</td>
</tr>
<tr>
<td>1650-1700</td>
<td>48 92.3</td>
<td>164 79.2</td>
<td>126 46.0</td>
<td>140 01.8</td>
</tr>
</tbody>
</table>

Table 1. The frequency of do by environment. [Ellegård 1953:166]
In graphical form:
Improving on Ellegård 1953 –
Aaron Ecay, “On the graduated evolution of do-support in English”, 2010:
What are “PPCEME” and “PCEEC”???

PPCEME: *Penn-Helsinki Parsed Corpus of Early Modern English*
1.7 million words from 448 texts over 210 years (1500-1710)

Authors: Anthony Kroch, Beatrice Santorini, Ariel Diertani.
Distributed by the University of Pennsylvania

PCEEC: *The Parsed Corpus of Early English Correspondence*
2.2 million words from 4970 letters over 285 years (1410-1695)

Authors: Ann Taylor, Anthony Warner, Susan Pintzuk, Arja Nurmi, Terttu Nevalainen
Distributed by the Oxford Text Archive
(From Sir Thomas More’s History of Richard III, 1513)
Alvar Ellegård’s research:
    a decade reading texts and writing examples on slips of paper.

Aaron Ecay’s research:
    an hour writing and running search scripts on his laptop.

And Aaron was able to look at the phenomenon in finer detail, e.g.
A closer examination of the Affirmative Declarative case:
So we’re done, right?

Any question about the history of English usage can be answered with a few minutes of digital corpus search!

Alas, no.

Three big problems remain:

We need analysis – beyond “words as letter strings”

We need larger samples – most linguistic events are rare

We need more sources – to allow for geographical, social, and individual variation

Let’s look at a simple example of the last two issues:

Contraction across time and space
Some more recent morphosyntactic change . . .
But there’s obviously a lot of stylistic variation – including from sources you might not expect, like geographical variation in U.S. real-estate listings.

I harvested 6 million words from 10 cities on the site of nation-wide listing service.

Here’s what such listings look like:

Truly, "A Diamond in the Desert". From the lovely dbl. door glass entry out to the gorgeous yard with oversized pool, you will not be disappointed. All rooms are open and airy. The large kit. has a cntr.island w/cooktop. The raised bar area has a sink and area for sitting. There is a lg. room upstairs w/wet bar-perfect for a media or hobby room. The master ste. is large with a beaut.sitting area for relaxing + fab.WI closet.

This pre-war one bedroom home is located on a high floor on a beautiful tree lined block. Open northern exposures from each room with beautiful city views, and a spectacular view of The Cathedral of Saint John the Divine. The apartment has recently been completely renovated without removing the original pre-war charm. Features include beamed ceilings, original oak wood flooring, stainless steel appliances, and baseboard moldings. Part-time doorman 3 PM-7 AM, live in super, bike room, storage, and laundry. Pied a terre's, gifting, co-purchasing are permitted with board approval.
This documents geographical diversity in socio-cultural preoccupations:
Katherine Seelye,
“And With a Roof, They’d Cost Even More: Two Boston Parking Spots Sell for $560,000 at Auction”,
*New York Times* 6/14/2013:

Jaws dropped in 2009 when someone paid $300,000 for a parking space, which was thought to be a record.

But now, even that has been shattered. At an auction on Thursday, the bidding for a tandem spot — space for two cars, one behind the other — started out at $42,000. It ended 15 minutes later at $560,000.
And diversity in (simulated?) enthusiasm in relation to price:
Or the financial stratification of adjectives:

<table>
<thead>
<tr>
<th></th>
<th>Rate per MW (top 50%)</th>
<th>Rate per MW (bottom 50%)</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>exquisite</td>
<td>235</td>
<td>53</td>
<td>4.5/1</td>
</tr>
<tr>
<td>dramatic</td>
<td>247</td>
<td>57</td>
<td>4.3/1</td>
</tr>
<tr>
<td>soaring</td>
<td>215</td>
<td>54</td>
<td>4.0/1</td>
</tr>
<tr>
<td>expansive</td>
<td>361</td>
<td>97</td>
<td>3.7/1</td>
</tr>
<tr>
<td>sophisticated</td>
<td>149</td>
<td>40</td>
<td>3.7/1</td>
</tr>
<tr>
<td>luxurious</td>
<td>512</td>
<td>143</td>
<td>3.6/1</td>
</tr>
<tr>
<td>lush</td>
<td>183</td>
<td>60</td>
<td>3.0/1</td>
</tr>
<tr>
<td>breathtaking</td>
<td>256</td>
<td>90</td>
<td>2.8/1</td>
</tr>
<tr>
<td>prestigious</td>
<td>236</td>
<td>90</td>
<td>2.6/1</td>
</tr>
<tr>
<td>charming</td>
<td>262</td>
<td>419</td>
<td>1/1.6</td>
</tr>
<tr>
<td>excellent</td>
<td>518</td>
<td>873</td>
<td>1/1.7</td>
</tr>
<tr>
<td>lovely</td>
<td>301</td>
<td>551</td>
<td>1/1.8</td>
</tr>
<tr>
<td>fresh</td>
<td>65</td>
<td>164</td>
<td>1/2.5</td>
</tr>
<tr>
<td>convenient</td>
<td>147</td>
<td>391</td>
<td>1/2.7</td>
</tr>
<tr>
<td>clean</td>
<td>100</td>
<td>292</td>
<td>1/2.9</td>
</tr>
<tr>
<td>good</td>
<td>205</td>
<td>819</td>
<td>1/4.0</td>
</tr>
<tr>
<td>nice</td>
<td>173</td>
<td>1196</td>
<td>1/6.9</td>
</tr>
<tr>
<td>cute</td>
<td>9</td>
<td>75</td>
<td>1/8.3</td>
</tr>
</tbody>
</table>
These listings have variable contraction:

You **will not** want to miss this wonderful home in sought after Martin Manor. Classic 1920’s Brick Bungalow in Historic West End with energy features that **will not** drain your pockets! Seller **will not** turn on utilities for inspections.

Great price, **do not** miss!
Please **do not** enter the property site without an appointment.
... the master closet has the laundry room, which most units in Foxcroft **do not** have!

Hurry! This one **won't** last long!
You **won't** find a street like this anywhere in Buckhead!
Don't wait. An investment you **won't** regret.

Completion May 2013, but **don't** wait so builder can customize.
You Will Hate Yourself For The Rest Of Your Life If You **Don't** Buy This Home!
We **don't** work with multiple offers and the buyer must be prepared to wait until bank approval.

10 Cities: Atlanta, Boston, Chicago, Denver, Houston, L.A., Miami, N.Y.C., Philadelphia, Las Vegas
Contraction counts from the listings corpus:

<table>
<thead>
<tr>
<th></th>
<th>Atlanta</th>
<th>Boston</th>
<th>Chicago</th>
<th>Denver</th>
<th>Houston</th>
<th>LA</th>
<th>Miami</th>
<th>NYC</th>
<th>Phila</th>
<th>Vegas</th>
</tr>
</thead>
<tbody>
<tr>
<td>do not</td>
<td>26</td>
<td>6</td>
<td>92</td>
<td>16</td>
<td>170</td>
<td>374</td>
<td>575</td>
<td>100</td>
<td>74</td>
<td>21</td>
</tr>
<tr>
<td>don’t</td>
<td>92</td>
<td>22</td>
<td>133</td>
<td>99</td>
<td>246</td>
<td>95</td>
<td>453</td>
<td>205</td>
<td>334</td>
<td>25</td>
</tr>
<tr>
<td>will not</td>
<td>38</td>
<td>15</td>
<td>67</td>
<td>47</td>
<td>133</td>
<td>98</td>
<td>323</td>
<td>161</td>
<td>144</td>
<td>18</td>
</tr>
<tr>
<td>won’t</td>
<td>51</td>
<td>5</td>
<td>56</td>
<td>55</td>
<td>152</td>
<td>44</td>
<td>245</td>
<td>55</td>
<td>130</td>
<td>28</td>
</tr>
<tr>
<td>TOTAL WRDS</td>
<td>248K</td>
<td>96K</td>
<td>571K</td>
<td>223K</td>
<td>733K</td>
<td>421K</td>
<td>1.4M</td>
<td>1.7M</td>
<td>754K</td>
<td>147K</td>
</tr>
</tbody>
</table>

Conclusions:

500k words per source is marginal for estimating source effects -- although the words involved are among the most common -- and 10 sources are not enough to get a stable estimate of the overall pattern.
A modest proposal

500-1000 documents per year
  from 1500 to 1922
  = about half a million documents
< 5% of the Hathi Trust holdings
with good meta-data and accurate OCR
  and automatic tagging and parsing

Plus searching and checking software
  for sample-based research,
  with fixes folded back into the dataset
“A Dictionary on Historical Principles”

Strengths: Large scale; Senses/usages, not just letter strings
Weaknesses: No counts; No syntactic analysis; Datasets mostly not available

Historical treebanks:

Strengths: Counts; Syntactic analysis; Datasets are published
Weaknesses: Small scale; “Words” as letter strings

Google books:

Strengths: Very large scale; Counts
Weaknesses: Bad texts; Worse metadata; “words” as letter strings;
Datasets mostly not published

In 100 years, scholarly resources will combine all these strengths
and lack all these weaknesses

Why not start in that direction now?
The key missing ingredient:

Stable standard lemma & sense IDs across history, geography, topic, register

No one is better able to provide these than the OED!

Could “Oxford Codes” be the basis of future English corpus linguistics?
Thank you!