From 500 passages to 50,000 books: Creating and using a large-enough historical corpus

Mark Liberman
http://ling.upenn.edu/~myl
From 500 passages to 500,000 books:
Creating and using a large-enough historical corpus

Mark Liberman
http://ling.upenn.edu/~myl
Outline

• Existing shared historical corpora are not big enough
• Much larger datasets are in reach
• But there are problems (metadata, OCR, parsing, ...)
• Social and technical solutions are possible
• These solutions are needed for other reasons
• So we should join others in reaching for the moon
Why a few million words is not (always) enough...
Example #1: \( V \) (that) \( S \)
COHA Counts for “suggested (that) he|she|they”:

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DEL</td>
<td>0 0 1 1 1 1 6 4 9 7 10 25 30 47 39 58 83 87 101</td>
</tr>
<tr>
<td>NO DEL</td>
<td>1 4 8 12 27 32 38 51 59 78 86 83 83 93 71 74 86 47 66</td>
</tr>
<tr>
<td>WORDS (M)</td>
<td>6.9 13.8 16 16.5 17.1 18.6 20.9 21.2 22.5 22.7 25.6 24.4 24.1 24.4 23.9 23.8 25.2 27.9 29.5</td>
</tr>
</tbody>
</table>

Conclusion:
For this investigation, 
~20 million words per decade is marginal.
Example #2:  Contraction of *will not* and *do not*
How about contractions in a set of sources from last week?

Specifically, real-estate listings from trulia.com, e.g.

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 in 2013 Real Estate Listings vs. COHA

Percentage "would" contraction

Percentage "don't" contraction
### Contraction counts from trulia.com real estate listings:

<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:

100k words per source is marginal for estimating source effect here  
10 sources are not enough to get a stable estimate of the overall pattern
Size of some available collections

Penn-Helsinki Parsed Corpus of Early Modern English:
1.7 million words from 448 texts over 210 years (1500-1710)
Curated (text & metadata), annotated, published

Early English Books Online / Text Creation Partnership (EEBO-TCP):
125,000 texts over ~220 years (1483-1700)
Partly curated (40,000 done), not annotated, not published yet (though accessible online)
“Phase I” (25,363 texts) to be made available in 2015
“Phase II” (45k more texts) to be made available ~ 2017+

Corpus of Historical American English:
400 mw from 100k texts over 200 years (1810-2009)
Semi-curated, semi-annotated, not published (and will not be)

Eighteenth Century Collections Online (ECCO)
200k texts over 100 years (~1700-1800)
Curated, not annotated, release prospects unclear (to me?)

Hathi Trust:
>10M texts over ~400 years (~31% public domain)
Lightly curated, not annotated, not published (but some can be downloaded)

Internet Archive:
??? Texts(12k on line) over ~400 years
Not curated, not annotated, can be downloaded

... etc. ...
Sometimes we’re stuck

• For Old English, Classical Latin, etc., the extant text is limited and all of it is available for historical study

• But for English since ~1500, and for many other languages, there are 100s or 1000s of books per year now available in digital form
So what’s the problem?

1. Bad OCR (optical character recognition)
2. Problematic metadata
   (editions, genres, authors)
3. Lack of annotation
   – Headings, captions, marginalia, ...
   – Quotations, dialogue, other languages, ...
   – Tagging, parsing, . . .
Thus in page 55", he fays " the Society haxe ma-
*^ nifefted a sufficient forwardnefs to encourage an4
" increafe fmall difaffe<5ted parties in our towns,
" upon an application to theni." And in the 5'7th
page he repreffents the Society as hpping that thefe
miali parties will by their influence gradually bring
on a general submiffion to an epifcopal sovereignt ; and
affirms that this has long been the formal deGgi^ of the Society, and is the true plan and grand
mytery of their operations in New-England." In his 1 06th page he tells us that the " affair of
Bishops in America, has been a favourite obje^ with the Society," and in the next page, that
the Society fpare neither endeavours, applications, nor expence, in order to effe6l their grand defign
" of epifcopizing all New-England," and a few FinesJ further, " The Society have long had 2, formal deftgn
" to diffolve and root out all our New-England " churches. — ^l^his (he fays) fully and clearly ac-*
" counts for their being fo ready to encourage fmall
*! epifcopal parties all over New-England, by fend-
* ing them miflionarics."
Choice of editions and sources:

Thomas Jefferson, *Notes on the State of Virginia*

Written in 1781, updated and enlarged in 1782 and 1783;
printed anonymously in Paris in 1785;
first public edition in 1787 in London.

Hathi Trust:
9 versions, published 1801, 1802, 1803, 1829, 1832 (2), 1853 (3), 1894.

Internet Archive:
12 versions, published 1787, 1801 (3), 1802, 1803, 1829, 1832 (2), 1853, 1955 (2)

Images of the original mss. online at the Massachusetts Historical Society.

Careful e-text version of 1787 edition at UVa Electronic Text Center
"Boundaries of Virginia"
An exact description of the limits and boundaries of the state of Virginia.

Limits
Virginia is bounded on the East by the Atlantic: on the North by a line of latitude, crossing the Eastern Shore through Watkins's Point, being about 37°.57'. North latitude from thence by a straight line to Cinquac, near the mouth of Patowmac; thence by the Patowmac, which is common to Virginia and Maryland, to the first fountain of its northern branch; thence by a meridian line, passing through that fountain till it intersects a line running East and West, in latitude 39°.43'.42.4" which divides Maryland from Pennsylvania, and which was marked by Messrs. Mason and Dixon; thence by that line, and a continuation of it westwardly to the completion of five degrees of longitude from the eastern boundary of Pennsylvania, in the same latitude, and thence by a meridian line to the Ohio: On the West by the Ohio and Mississippi, to latitude 36°.30'. North: and on the South by the line of latitude last-mentioned.
What we need

Organized effort to

- Select
  - texts
  - sources

- Correct
  - metadata
  - texts

- Annotate
  - text structure
  - linguistic structure
Luckily,

- Lots of texts in decent shape already exist (EEBO, ECCO, various smaller collections)
- Although OCR for older books sucks,
  
  It can be improved by better font training and better language models!
- Tagging, parsing etc. are good and improving, and there are ideas for making them MUCH better!
- Crowdsourcing often works
- There are other applications and customers for the improvements we need
OCR improvements

It’s easy to make BIG improvements in OCR for older texts, using adaptive language modeling techniques from ASR and similar areas.

... if you’re interested in this and you know something about language modeling and machine learning --
-- or you know someone with skills and interest --

Please contact me!
Because OCR will never be perfect...

TypeWright – web app for crowdsourcing OCR correction, at http://18thconnect.org
Current parsers are pretty good at short sentences

Thomas Jefferson, *Notes on the State of Virginia*:

I knew a single instance of gold found in this state.

( (S (NP (PRP I)) (VP (VBD knew) (SBAR (S (NP (NP (DT a) (JJ single) (NN instance)) (PP (IN of) (NP (NN gold))))) (VP (VBD found) (PP (IN in) (NP (DT this) (NN state)))))))) (. .)) )
Longer sentences are more likely to go wrong...

It was interspersed in small specks through a lump of ore, of about four pounds weight, which yielded seventeen pennyweight of gold, of extraordinary ductility.
The metal is mixed, sometimes with earth, and sometimes with rock, which requires the force of gunpowder to open it; and is accompanied with a portion of silver, too small to be worth separation under any process hitherto attempted there.
So what can we do NOW?
Claim #1:

• Taggers and parsers are already good enough to support *sampling methods* of analysis:
  – In this approach,
    a search yields a random unbiased (?) sample of relevant examples,
    mostly classified correctly,
    mixed with a not-excessive amount of junk
  – Researchers can efficiently check the output,
    correcting the classification,
    and rejected the junk

• As taggers and parsers improve, this gets easier
Claim #2:

• Given a few cognitively-easy human judgments per sentence, parser output can be as good as human treebanking.

• These judgments might be pre-annotation
  – e.g. marking scope of conjunction, clause boundaries, etc.

  or post-annotation
  – e.g. checking and correcting things the parser is “unsure” about.

• Literate native speakers can learn in a few hours to make these judgments at a rate of >1,000 words/hour.

• At this rate, we could create a >100-million-word Treebank for the cost of the original 1-million-word Penn Treebank.

• Selected material relevant to some research topic could easily and efficiently be parsed or otherwise analyzed.
. . . And now for something completely different . . .
(Meta-)Claim

It’s our fault that grammar is no longer taught in American grammar schools, because:

– Religious wars among linguists leave educators confused and depressed.
– Each grammatical sect changes its theology in fundamental ways every decade or so.
– So by the time a textbook is published and sold, its analyses will no longer be supported even by its authors...
BUT ... 

There’s one framework of syntactic analysis that
  – has been essentially stable for 25 years
  – has been used for 5 or 6 very different languages
  – is documented in excruciating detail in many manuals
  – is used by many parsers, both proprietary and open-source
  – is used in many parsed corpora, supporting research from a variety of perspectives

... namely the Penn Treebank framework.
Conclusion

• PTB treebanking could be taught (in some form) to students in from grammar school to college
  – without offending too many linguists
  – without having to re-train the teachers every ten years

• This would be good for the students and for the field

• And it would result in a VERY big crowd to source from
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