Choices over time

Some methodological issues in research into current change

Bas Aarts, Jo Close and Sean Wallis

Survey of English Usage
University College London

{b.aarts, j.close, s.wallis}@ucl.ac.uk
Introducing DCPSE

• The *Diachronic Corpus of Present-day Spoken English*
  – *orthographically transcribed* spoken BrE
  – *fully parsed*, searchable with ICECUP and FTFs
  – *400,000 words each from*
    • LLC (‘Survey Corpus’)
    • ICE-GB
  – balanced by *text category*
  – *not* evenly distributed by *year*
    • LLC: samples from 1958-1977
    • ICE-GB: 1990-1992
What can a parsed corpus tell us?

- Parsed corpora contain tree diagrams
  - Use Fuzzy Tree Fragment (FTF) queries to get data
  - An FTF:

- A matching case in a tree:
will vs. shall

- Barber (1964)
  - “[T]he distinctions formerly made between shall and will are being lost, and will is coming increasingly to be used instead of shall.”

- Mair and Leech (2006)
  - lexical counts in Brown family of corpora (written)
    - BrE and AmE: shall falls (~50%) with time

<table>
<thead>
<tr>
<th></th>
<th>1960s BrE</th>
<th>1990s BrE</th>
<th>1960s AmE</th>
<th>1990s AmE</th>
</tr>
</thead>
<tbody>
<tr>
<td>will</td>
<td>2,798</td>
<td>2,723</td>
<td>2,702</td>
<td>2,402</td>
</tr>
<tr>
<td>shall</td>
<td>355</td>
<td>200</td>
<td>267</td>
<td>150</td>
</tr>
</tbody>
</table>
will vs. shall

- Barber (1964)
  - “[T]he distinctions formerly made between shall and will are being lost, and will is coming increasingly to be used instead of shall.”

- Mair and Leech (2006)
  - lexical counts in Brown family of corpora (written)
    - BrE and AmE: shall falls (~50%) with time
    - Transatlantic convergence: AmE and BrE are distinct in 1960s but not distinct in the 1990s

<table>
<thead>
<tr>
<th></th>
<th>1960s BrE</th>
<th>1990s BrE</th>
<th>1960s AmE</th>
<th>1990s AmE</th>
</tr>
</thead>
<tbody>
<tr>
<td>will</td>
<td>2,798</td>
<td>2,723</td>
<td>2,702</td>
<td>2,402</td>
</tr>
<tr>
<td>shall</td>
<td>355</td>
<td>200</td>
<td>267</td>
<td>150</td>
</tr>
</tbody>
</table>
will vs. shall

Questions...

- Are will and shall true alternates in each case?
  - what about will not, shall not, won’t, shan’t and interrogative forms?
  - do we include ‘ll?
  - Mair and Leech cite log-likelihood of words
    - a kind of $\chi^2$ for $[\{x, x\prime\}, \{N-x, N'-x\prime\}]$
    - (x, x’ = frequency of item, N, N’ = corpus size)
    - it tells us that shall is less frequent in the later corpus
    - it does not tell us whether will is replacing shall

<table>
<thead>
<tr>
<th></th>
<th>1960s BrE</th>
<th>1990s BrE</th>
<th>1960s BrE</th>
<th>1990s BrE</th>
</tr>
</thead>
<tbody>
<tr>
<td>will</td>
<td>2,798</td>
<td>2,723</td>
<td>shall</td>
<td>355</td>
</tr>
<tr>
<td>N-will</td>
<td>997,202</td>
<td>997,277</td>
<td>N-shall</td>
<td>999,645</td>
</tr>
</tbody>
</table>

N = 1M
will vs. shall

• Questions...
  – Are will and shall true alternates in each case?
    • what about will not, shall not, won’t, shan’t and interrogative forms?
    • do we include ‘ll?
  • Mair and Leech cite log-likelihood of words
    – a kind of $\chi^2$ for [{\(x, x'\)}, {\(N-x, N'-x'\)}]
      (\(x, x'\) = frequency of item, \(N, N'\) = corpus size)
    – it tells us that shall is less frequent in the later corpus
    – it does not tell us whether will is replacing shall
  • we’ve reanalysed data using $\chi^2$ for [{\(x, x'\)}, {\(y, y'\)}]

<table>
<thead>
<tr>
<th></th>
<th>1960s BrE</th>
<th>1990s BrE</th>
<th>1960s AmE</th>
<th>1990s AmE</th>
</tr>
</thead>
<tbody>
<tr>
<td>will</td>
<td>2,798</td>
<td>2,723</td>
<td>2,702</td>
<td>2,402</td>
</tr>
<tr>
<td>shall</td>
<td>355</td>
<td>200</td>
<td>267</td>
<td>150</td>
</tr>
</tbody>
</table>
will vs. shall

• Questions...
  – Are will and shall true alternates in each case?
    • what about will not, shall not, won’t, shan’t and interrogative forms?
    • do we include ‘ll?
  – Mair and Leech cite log-likelihood of words
    – a kind of $\chi^2$ for $\{x, x'\}, \{N-x, N'-x'\}$
      $(x, x' = \text{frequency of item, } N, N' = \text{corpus size})$
    – it tells us that shall is less frequent in the later corpus
    – it does not tell us whether will is replacing shall
  • we’ve reanalysed data using $\chi^2$ for $\{x, x'\}, \{y, y'\}$

– Can we show a change in use in speech?
– Can we show change over this period?
**will** vs. **shall** vs. **‘ll** (DCPSE)

- Use parsing to find plausible alternates
  1. Create FTFs like this for **shall**, **will** and **‘ll**

![Diagram showing the structure of sentences with **shall** and **will**]

- Then create FTFs for **shall not** and **will not**
  - Subtract from first set of results (a different experiment)
  - These counts exclude
    - negative forms: **shall not**, **shan’t**, **will not**, **won’t**
    - subject-auxiliary inversion
**will vs. shall vs. ’ll (DCPSE)**

- Consider the three-way alternation

- Most variation is for *shall*

<table>
<thead>
<tr>
<th></th>
<th>shall</th>
<th>will</th>
<th>’ll</th>
<th>TOTAL</th>
<th>(\chi^2)(shall)</th>
<th>(\chi^2)(will)</th>
<th>(\chi^2)(’ll)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC</td>
<td>124</td>
<td>501</td>
<td>663</td>
<td>1,288</td>
<td>15.71</td>
<td>2.16</td>
<td>0.01</td>
</tr>
<tr>
<td>ICE-GB</td>
<td>46</td>
<td>544</td>
<td>638</td>
<td>1,228</td>
<td>16.48</td>
<td>2.26</td>
<td>0.01</td>
</tr>
<tr>
<td>TOTAL</td>
<td>170</td>
<td>1,045</td>
<td>1,301</td>
<td>2,516</td>
<td>(\chi^2)</td>
<td>36.63s</td>
<td></td>
</tr>
</tbody>
</table>
**Will vs. Shall vs. 'll (DCPSE)**

- If *will* and *'ll* behave similarly, group them

<table>
<thead>
<tr>
<th></th>
<th>Shall</th>
<th>Will</th>
<th>'ll</th>
<th>Total</th>
<th>$\chi^2($will$)$</th>
<th>$\chi^2($'ll$)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC</td>
<td>124</td>
<td>501</td>
<td>663</td>
<td>1,164</td>
<td>0.58</td>
<td>0.47</td>
</tr>
<tr>
<td>ICE-GB</td>
<td>46</td>
<td>544</td>
<td>638</td>
<td>1,182</td>
<td>0.58</td>
<td>0.47</td>
</tr>
<tr>
<td>Total</td>
<td>170</td>
<td>1,045</td>
<td>1,301</td>
<td>2,346</td>
<td>$\chi^2$</td>
<td>2.11 ns</td>
</tr>
</tbody>
</table>
**will vs. shall vs. 'll (DCPSE)**

- If *will* and *'ll* behave similarly, group them:

<table>
<thead>
<tr>
<th></th>
<th>shall</th>
<th>will+’ll</th>
<th>TOTAL</th>
<th>$\chi^2$(shall)</th>
<th>$\chi^2$(will+’ll)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC</td>
<td>124</td>
<td>1,164</td>
<td>1,288</td>
<td>15.71</td>
<td>1.14</td>
</tr>
<tr>
<td>ICE-GB</td>
<td>46</td>
<td>1,182</td>
<td>1,228</td>
<td>16.48</td>
<td>1.19</td>
</tr>
<tr>
<td>TOTAL</td>
<td>170</td>
<td>2,346</td>
<td>2,516</td>
<td>$\chi^2$</td>
<td>34.52s</td>
</tr>
</tbody>
</table>
**will vs. shall vs. 'll (DCPSE)**

- If *will* and *'ll* behave similarly, group them.

<table>
<thead>
<tr>
<th></th>
<th>shall</th>
<th>will+’ll</th>
<th>TOTAL</th>
<th>$\chi^2$(shall)</th>
<th>$\chi^2$(will+’ll)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC</td>
<td>124</td>
<td>1,164</td>
<td>1,288</td>
<td><strong>15.71</strong></td>
<td>1.14</td>
</tr>
<tr>
<td>ICE-GB</td>
<td>46</td>
<td>1,182</td>
<td>1,228</td>
<td><strong>16.48</strong></td>
<td>1.19</td>
</tr>
<tr>
<td>TOTAL</td>
<td>170</td>
<td>2,346</td>
<td>2,516</td>
<td>$\chi^2$</td>
<td><strong>34.52s</strong></td>
</tr>
</tbody>
</table>
shall over time (DCPSE)

- Proportion of alternates that are *shall*, by year

\[ p(\text{shall} \mid \{\text{shall, will, 'll}\}) \]
shall over time (DCPSE)

• Proportion of alternates that are shall, by year

\[ p(\text{shall} \mid \{\text{shall}, \text{will}, \,'ll\}) \]

Error bars based on Poisson
Focusing on true alternation

- Aim: to focus on true alternation
  - minimise other sources of variation

- Consider changing use of the progressive
The progressive (DCPSE)

- FTF to retrieve progressives from DCPSE

- Identifying the alternates
  (see Smitterberg 2005; Aarts, Close & Wallis forthcoming)
  - VP(prog)
    - Exclude *be going to* future (automatic)
  - VP(¬prog)
    - Exclude imperatives, infinitives, (benefits of using a parsed corpus)
The progressive over time (DCPSE)

- The rise of the English progressive in spoken English (as a proportion of alternates)

\[ p(\text{VP}(\text{prog}) \mid \{\text{VP}(\text{prog}), \text{VP}(\neg\text{prog})\}) \]

[Graph showing the rise of the English progressive over time]
Conclusions

- We focus on **true alternation** to investigate if **replacement** is occurring by considering:
  - variation (over time) *where there is a choice*
  - hierarchies of alternates
    - as with {shall, {will, 'll}}

- This can be difficult
  - Requires a linguistic argument
  - May require careful examination of cases

- It is extensible to other types of experiment, e.g. interaction between choices
References

- Aarts, Bas, Jo Close and Sean Wallis (forthcoming) Recent changes in the use of the progressive construction in English. In: Bert Cappelle and Naoaki Wada (eds.) Festschrift for (secret).


