A new corpus of spoken English
– extending the diachronic scope of the DCPSE

1. Introduction
My research focus: The progressive’s use and development in spoken British English

# Is the frequency increase of the progressive still under way in the 21st century?
# If yes, in which genres?
# What are the driving forces behind a possible frequency increase?
# In what way is the progressive’s constructional network changing?

What corpus resources are available?

2. The Diachronic Corpus of Present-Day Spoken English

- A diachronic corpus of spoken British English
- Released by the Survey of English Usage in 2006 (Aarts and Wallis 2006)
- Consists of two sub-corpora: London-Lund Corpus (LLC; ~400 000 words) and spoken part of ICE-GB (~400 000 words).
- ~ 800 000 words
- Time period covered: 1950s – 1970s (LLC) and early 1990s (ICE-GB)

Most recent data is already 25 years old. Need for a new corpus of spoken British English.
2. The Diachronic Corpus of Present-Day Spoken English

- Why not using the new BNC2014?
  - Not available before mid. 2017
  - Only face-to-face conversations

- Why not using the Hansard Corpus? (Davies 2016)
  - Only parliamentary language
  - Most recent data from 2005
  - Quality of Hansard transcripts (Mollin 2007)

What is needed is a diachronic extension of the DCPSE.

3. A new corpus of spoken English

- Extending the diachronic scope of the DCPSE by ~25 years
- Data from ~2012-2016 (~ 400 000 words)
- Altogether, the LLC, ICE-GB and the new corpus data will cover a period of about 65 years.
- Other corpora such as the Hansard and BNC2014 can be used to further investigate specific aspects of progressive use that require a larger data base.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Data Source</th>
<th>Word No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face-to-Face Conversation</td>
<td>The Listening Project</td>
<td>~ 100 000</td>
</tr>
<tr>
<td>Broadcast Discussions</td>
<td>Ouch Transcripts</td>
<td>~ 50 000</td>
</tr>
<tr>
<td>Broadcast Interviews</td>
<td>In Touch and The Andrew Marr Show Transcripts (relevant parts)</td>
<td>~ 50 000</td>
</tr>
<tr>
<td>Spontaneous Commentary</td>
<td>Radio and TV Commentary</td>
<td>~ 50 000</td>
</tr>
<tr>
<td>Parliamentary Language</td>
<td>Hansard – House of Commons Debates (corrected)</td>
<td>~ 50 000</td>
</tr>
<tr>
<td>Prepared Speech</td>
<td>Transcripts of Speeches, Lectures and Sermons (corrected)</td>
<td>~ 100 000</td>
</tr>
</tbody>
</table>
3. A new corpus of spoken English

- Work stages:
  - **Stage 1**: Data collection (Internet) and transcription (ExpressScribe)
  - **Stage 2**: Data selection, correction and editing / standardisation (TextWrangler)
  - **Stage 3**: Tokenisation, lemmatisation, tagging and merging (WebLicht and R)
  - **Stage 4**: XML encoding (and improved tagging)
  - **Stage 5**: Selecting a platform (e.g. FreiDoc or CQPweb), resolving legal issues

---

4. The progressive in two non-conversational genres

- Hypothesis #1:
  The frequency increase of the progressive in the 21st century is still under way, and it is genre-specific.
  It is levelling off in conversational genres and taking hold in less conversational, more conservative genres.

  - **Prepared Speech and Parliamentary Language**
    - high register
    - expository genres
    - contain scripted speech
    - non-conversational spoken language
  
- Hypothesis #2:
  The frequency increase of the progressive in the 20th and 21st century is based on further entrenchment of prototypical core uses rather than on structural or functional innovations.
  Structural and functional innovations can be observed but have little impact on the corpus data.
  
  (cf. Leech et al. 2009: 143)
4. The progressive in two non-conversational genres

syntactic contexts: the increase of the present progressive active

Significant for p < 0.001

Summary

- The corpus extends the DCPSE’s diachronic scope by ~ 25 years.
- Six genres
- DCPSE + new corpus researching ongoing language change
- Well suited for the study of (frequent) lexical and grammatical phenomena such as the progressive
- The corpus will be readily compiled by the end of this year.
- As matters stand, a publicly available version of the corpus will not be ready before 2018/2019.

References


BBC. The Andrew Marr Show. Transcripts available online at: http://www.bbc.co.uk/programmes/articles/34afHfHMD6EM63q6G3SDFP/transcripts.


