Double modals in YouTube videos from North America and the British Isles

Steven Coats
English Philology, University of Oulu, Finland
steven.coats@oulu.fi

Corpus-based and Computational Approaches to Variation Workshop, Helsinki
April 27th, 2022
Outline

1. CoNASE and CoBISE
2. YouTube ASR captions files, data collection and geocoding
3. Methods: Frequency analysis (frequent features), manual inspection/annotation (rare features)
4. Double modals in North America and in the British Isles
5. Caveats, summary

Slides for the presentation are on my homepage at https://cc.oulu.fi/~scoats
Introduction

- Available corpora of transcribed spoken English (Anderwald & Wagner 2007; Corbett 2014; Corrigan et al. 2012; Du Bois et al. 2000-2005; Kallen & Kirk 2007) are small or lack a broad geographic focus; size may make it difficult to find some features
- CoNASE: 1.25b token corpus of 301,846 word-timed, part-of-speech-tagged Automatic Speech Recognition (ASR) transcripts (Coats forthcoming a)
- CoBISE: 112m token corpus of 38,680 ASR transcripts (Coats forthcoming a)
- Correspond to more than 166,000 hours of video from more than 3,000 YouTube channels of local councils and other government entities in locations in the US, Canada, England, Scotland, Wales, Northern Ireland, and the Republic of Ireland
- Freely available for research use; download from the Harvard Dataverse here and here
YouTube captions files

- Videos can have multiple captions files: user-uploaded captions, auto-generated captions created using automatic speech recognition (ASR), or both, or neither
- User-uploaded captions can be manually created or generated automatically by 3rd-party ASR software
- Auto-generated captions are generated by YT's speech-to-text service
- CoNASE and CoBISE: target YT ASR captions
Focus on regional and local council channels

Many recordings of meetings of elected councillors: advantages in terms of representativeness and comparability

- Speaker place of residence (cf. videos collected based on place-name search alone)
- Topical contents and communicative contexts comparable
Data collection and processing

- Identification of relevant channels (YouTube API, searches of public-facing server, lists of councils with YT channels)
- Inspection of returned channels to remove false positives
- Retrieval of ASR transcripts using YouTube-DL
- VPN or Tor to circumvent IP blocking
- Geocoding: String containing council name + channel name + country location to Google's geocoding service
- PoS tagging with SpaCy (Honnibal et al. 2019)
Transcript accuracy

- ASR transcripts contain errors (WER ~22%)
- High-frequency phenomena: signal of correct transcriptions will be stronger (Agarwal et al. 2009) → classifiers
- Low-frequency phenomena: manually inspect corpus hits
Example analysis: Double modals

- Non-standard rare syntactic feature in the British Isles, North America, and elsewhere (Montgomery & Nagle 1994; Coats 2022)
- Will you can help me with this?
- Occurs only in the American Southeast and in Scotland/Northern Ireland/Northern England?
- Most studies based on non-naturalistic data with limited geographical scope (LAMSAS, LAGS, Murray 1873; Wright 1898-1905; Anderwald & Wagner 2007; Kallen & Kirk 2007; Smith et al. 2019)
Script: Generating a table for manual inspection of double modals

- Base modals will, would, can, could, might, may, must, should, shall, used to, 'll, ought to, oughta
- Script to generate regex of two-tier combinations, plus forms with intervening pronouns, auxiliary verbs, negations

```python
import re
hits = []
for i, x in cobise_df.iterrows():
    pat1 = re.compile("((\w+_S+\S+)\S+\S+\s){3}'+x[0]+'_\w+_S+'n?_\w+_S+(\w+_S+_S+\S+\s){3})",re.IGNORECASE)
    if pat1.search(x["text_pos"]):
        finds1 = pat1.findall(x["text_pos"][0])
        seq = " ".join([x.split("_")[-1] for x in finds1[0].split()])
        time = finds1[0].split()[0].split("_")[-1]
        hits.append((x["country"],x["channel_title"],seq,"https://youtu.be/"+x["video_id"]+"?t="+str(round(float(time)-3))))
pd.DataFrame(hits)
```
<table>
<thead>
<tr>
<th>Country</th>
<th>Channel</th>
<th>DM</th>
<th>Regex hit</th>
<th>link</th>
<th>type</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>England</td>
<td>Fylde Council</td>
<td>would will</td>
<td>would not will not</td>
<td><a href="https://youtu.be/VogFB5X_1UM?t=490">https://youtu.be/VogFB5X_1UM?t=490</a></td>
<td>fp o1</td>
<td></td>
</tr>
<tr>
<td>England</td>
<td>Fylde Council</td>
<td>may would</td>
<td>may would</td>
<td><a href="https://youtu.be/6UudWle_wYM?t=1733">https://youtu.be/6UudWle_wYM?t=1733</a></td>
<td>t</td>
<td>&quot;the wording, which it may would be better to read&quot;</td>
</tr>
<tr>
<td>England</td>
<td>Fylde Council</td>
<td>could can</td>
<td>could he can</td>
<td><a href="https://youtu.be/udvH08tQ2ls?t=960">https://youtu.be/udvH08tQ2ls?t=960</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>England</td>
<td>Fylde Council</td>
<td>may would</td>
<td>may I would</td>
<td><a href="https://youtu.be/V3YFSetBxgM?t=2184">https://youtu.be/V3YFSetBxgM?t=2184</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>England</td>
<td>Fylde Council</td>
<td>should can</td>
<td>should you can</td>
<td><a href="https://youtu.be/6erR7ZuYjtc?t=97">https://youtu.be/6erR7ZuYjtc?t=97</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>England</td>
<td>Fylde Council</td>
<td>'ll can</td>
<td>'ll can</td>
<td><a href="https://youtu.be/zA7LkTk0Vt4?t=1447">https://youtu.be/zA7LkTk0Vt4?t=1447</a></td>
<td>t</td>
<td>&quot;and we'll can see how&quot;</td>
</tr>
<tr>
<td>England</td>
<td>Fylde Council</td>
<td>would can</td>
<td>would we can 't</td>
<td><a href="https://youtu.be/f1_x5C1ttCk?t=1702">https://youtu.be/f1_x5C1ttCk?t=1702</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>England</td>
<td>Fylde Council</td>
<td>should can</td>
<td>should you can</td>
<td><a href="https://youtu.be/7AsXVW1vako?t=119">https://youtu.be/7AsXVW1vako?t=119</a></td>
<td>sr d</td>
<td></td>
</tr>
<tr>
<td>England</td>
<td>Fylde Council</td>
<td>can will</td>
<td>can will</td>
<td><a href="https://youtu.be/APNxUQP3Zok?t=1718">https://youtu.be/APNxUQP3Zok?t=1718</a></td>
<td>t</td>
<td>&quot;that business rating can will neither increase nor&quot; Scottish accent</td>
</tr>
<tr>
<td>England</td>
<td>Fylde Council</td>
<td>could would</td>
<td>could would</td>
<td><a href="https://youtu.be/X3r_Y_Qdk5kA?t=492">https://youtu.be/X3r_Y_Qdk5kA?t=492</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>England</td>
<td>Fylde Council</td>
<td>could can</td>
<td>could we can</td>
<td><a href="https://youtu.be/ded7W7m7id4?t=15">https://youtu.be/ded7W7m7id4?t=15</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>England</td>
<td>Fylde Council</td>
<td>would can</td>
<td>would can</td>
<td><a href="https://youtu.be/PiSnXSul8tQ?t=447">https://youtu.be/PiSnXSul8tQ?t=447</a></td>
<td>fp a1 o1</td>
<td></td>
</tr>
</tbody>
</table>
Double modals

- Regular-expression-search and manual annotation approach
- Double modals can be found in the US North and West and in Canada; in Scotland, N. Ireland, and N. England, but also in the English Midlands and South and in Wales (Coats in Review)
A few caveats

- Meetings of local government not representative of speech in general
- ASR errors, quality of transcript related to quality of audio as well as dialect features (Tatman 2017; Meyer et al. 2020; Markl & Lai 2021)
Summary and outlook

- Large corpora of ASR transcripts from YouTube channels of local governments in the US, Canada, Britain, and Ireland (coming soon: Australia/NZ, 190m tokens, Germany, 56m tokens)
- Useful for corpus studies of spoken language, dialectology, pragmatics
- Double modals are more widespread than has previously been documented
Thank you!
References


Coats, S. In review. Double modals in contemporary British and Irish Speech.


References II


