Apparent-time variation and spatial diffusion in large social-media corpora
Tweetolectology

• three-year ESRC project ‘Investigating the diffusion of morphosyntactic innovations using social media’, 1 September 2017–30 September 2020

• David Willis (PI, Cambridge), Adrian Leemann (CI, Cambridge/Bern), Tam Blaxter (CI, Cambridge), Deepthi Gopal (RA, Cambridge), Jonathan Mackenzie (IT, Cambridge)

• investigates geospatial patterns of grammatical variation in British English, Welsh, Norwegian/mainland Scandinavian and Turkish
Using Twitter as a dialectological source

- Large-scale social-media data is increasingly of interest to linguists (Eisenstein et al. 2010; Eisenstein et al. 2014; Russ 2012; Doyle 2014a, 2014b; Bamman et al. 2014; Gonçalves & Sánchez 2014; Grieve et al. 2013; Grieve et al. 2017; Grieve et al. 2018; Grieve et al. 2019).

- Much existing work shares the property of dealing with frequent (often lexical) patterns appearing in very large languages (American English, World Spanish), and the granularity of geographic information is often coarse.

- Our aim is to examine smaller languages and use more fine-grained geographic information, with the aim of tracing patterns of language change in detail.

- Existing work examines change only over the timescales covered by social media corpora: here we attempt to extend the timescale which it is possible to study by investigating change in apparent time.
Our corpus

- British English data: corpus of tweets collected between October 2017 and June 2019 which we have manually determined to originate from users in the UK and Ireland; total 104,657,500 tweets from 1,734,260 accounts.

- Keyword-localisation used to map users to best guess based on: user location field, bio, and tweet text searched for Ordnance Survey place names, informal names, and names of regions, weighted by number of occurrences, context and distance between places mentioned.

- Of this set of users, 196,312 (11.3%) were assigned a precise age based on mentions of year of birth or age at time of posting and another 247,999 (14.3%) were assigned an age category on the basis of keywords relating to family relationships and employment.
Distribution of ages

- Young people overrepresented in the birthyear data compared with the age category data because young people are much more likely to mention their age explicitly (sharp falloff for people born before 1996 / 22yos) so for older users we’re more reliant on mentions of relationships and employment.
- by either measure, relatively few people in the dataset over the age of 60 / born before 1959
The variables

- Preposition drop: zero preposition where other varieties would have *to* in e.g. “let’s go pub”
- Leveling in the preterite of *to be* to *was* (“you was”, “they was”) or *were* (“I were”, “she were”)
Preposition drop: background

• A number of varieties of the north west and south east of England allow constructions with no overt preposition where other varieties would have ‘to’:

  Lancashire: Today I’m going the library (Myler 2011, 2013, Biggs 2014)
  Liverpool: Swim the end and back (Biggs 2014)
  Manchester: She went the pub (Haddican 2010)
  London and Kent: I haven’t been Shoreditch in ages (Bailey 2018; also MLE)

• This variation seems to be below the level of consciousness (speakers are not aware that these constructions would be ungrammatical in other varieties, Biggs 2014:15) and is a good candidate for a variable currently undergoing change.
Authors agree that there appear to be syntactic differences between preposition drop found in different dialects.

In the south east, the following restrictions appear to apply (Bailey 2018):

(i) the determiner is obligatorily absent
(ii) the verb must be directional *go* or *come*
(iii) the noun must be interpreted as a directional Goal
(iv) the noun must denote a familiar or anaphoric location or an institution
(v) modification (e.g. with *straight, right*) is not possible

(i) does not hold in the north west (*I’m going the pub*, Haddican 2010, *John came the pub with me*, Biggs 2014)

(ii) does not apply in Lancashire, where a wider range of verbs (at least *come, go, run, drive, nip, jog*) are possible (Myler 2011); in Liverpool, the set of verbs is much wider still (*Joe plodded the pub, swim the end and back*, Biggs 2014)

(iii) does not appear to apply in Liverpool, so that the omitted preposition can be ‘at’ as well as ‘to’ (*I’m working the library tonight*, Biggs 2014); however (iii) does hold elsewhere in the north west

(v) does not apply in Liverpool (*I’m going straight the pub after this*, Biggs 2014) but does apply in Lancashire (Myler 2011)
Preposition drop: background

• This past work is based on relatively small samples of speakers, so it is hard to be absolutely confident that these syntactic differences systematically divide regionally-defined dialects rather than reflecting a high level of interspeaker variation within preposition-drop varieties.

• However, assuming that these really are systematic regional differences, that has implications for diachrony.

• A typical scenario is that an innovation begins in a highly restricted context and then spreads to a progressively wider range of contexts (actualisation) at the same time as it is spreading to new regions (diffusion), with the result that it is most grammatically restricted in the varieties in which it is newest (cf. Willis 2017 for the diffusion and actualisation of 2.sg. pronoun *chdi* in Welsh)—might this be the case here?
Preposition drop: background

- This gives us some predictions we can test against the twitter data:
  - there is ongoing change, with preposition drop increasing over time in British English varieties
  - preposition drop is restricted to varieties of the north west (Lancashire, Liverpool, Manchester) and south east (London, Kent) of England
  - the syntax of the construction differs across these localities, with a wider range of grammatical contexts in the north west than in the south east
  - the construction is oldest / the change is most advanced in the varieties where the fewest grammatical constraints apply
Preposition drop: data collection

- We searched the corpus for forms of *go* and identified the most frequent collocations which fulfilled the semantic criteria in (iv):
  - go Amsterdam, go Asda, go chicken shop, go college, go jail, go London, go Manchester, go Nando’s, go Paris, go prison, go pub, go school, go Tesco
- We then searched for all occurrences of these with and without *to* and *straight* (and where relevant with and without *the*)
- This allows us to see the overall distribution, and to investigate the constraint on the definite article; *straight* modification is vanishingly rare (just one token) so not examining that further here
- We identified a total of 37,709 tokens (of which for 20,843 we could associate some age metadata)

<table>
<thead>
<tr>
<th></th>
<th><em>to</em></th>
<th>preposition drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>no article</td>
<td>27,677</td>
<td>5113</td>
</tr>
<tr>
<td><em>the</em></td>
<td>4492</td>
<td>427</td>
</tr>
</tbody>
</table>
Preposition drop: overall spatial distribution
• This seems largely in line with the literature: preposition drop is strongly associated with the north west (Liverpool, Manchester) and south east (east London, north Kent)
• we can add however that it is also found in the Midlands (West Midlands Conurbation, Coventry, Northampton, Leicester) and north Wales
• it appears to occur at a lower frequency across a much larger area (East Anglia, Devon and Cornwall, etc.) although there are places it does not occur (South Wales, the south coast of England, the north east of England, Scotland, Northern Ireland)
Preposition drop: change over time by year of birth
Here just using the accounts for which we have the most specific metadata (year of birth), we see clear change, with the rate of preposition drop increasing in apparent time.
Preposition drop: spatial change over time by age category

Rate of preposition drop:
- 0-5.0%
- 5.0-8.9%
- 8.9-12.7%
- 12.7-16.5%
- 16.5-20.4%
- 20.4-24.2%
- 24.2-28.1%
- 28.1-31.9%
- 31.9-35.8%
- 35.8-39.6%

Weighted number of accounts:
- 0.0 - 2.3
- 2.3 - 6.0
- 6.0 - 11.1
- 11.1 - 18.2
- 18.2 - 29.1
- 29.1 - 44.3
- 44.3 - 77.1
- 77.1 - 102.9

speaker age: <30
speaker age: 30-60
speaker age: >60
Here using the broader age categories to give us a bigger dataset, we can see how change over time relates to the spatial distribution. Not only does the frequency of preposition drop fall in older age groups, it becomes more geographically restricted, with the core area in the north west (though really Cheshire rather than specifically Liverpool or Manchester).
• Note that with this slightly larger dataset, the differences between regions become somewhat sharper, and we can also see a fifth innovative region in Devon/Cornwall.
Preposition drop: change over time by year of birth and region
Here returning to using year of birth and grouping the data by nearest of the five innovative centres, we find that it is not the north west but the west Midlands that stands out as most innovative.
Preposition drop: change over time by year of birth and region

- We can probably discount the very earliest part of the curve, as it reflects an extremely small number of speakers.
Preposition drop: spatial distribution by presence of definite article
Moving on to grammatical conditioning, here we see the spatial distribution by the presence of the definite article.

As predicted in the literature, this is possible in the north west but not the south east.

It seems that the Midlands patterns with the south east in this respect.
Preposition drop: change over time by presence of definite article
Preposition drop: change over time by presence of definite article

- Assuming again that we should discount the curve before around 1950 since there are too few speakers in this age group, we see sharply divergent age distributions for *go pub* vs. *go the pub*
- the change in apparent time is entirely due to increasing *go pub*, with *go the pub* largely stable
Preposition drop: conclusions

• Predictions from the literature on the spatial distribution of grammatical conditioning (specifically the constraint against preposition drop with an overt definite article) held up extremely well against the twitter data: preposition drop with an overt definite article was possible in the north west, impossible elsewhere

• As predicted, we found ongoing change, with preposition drop increasing in apparent time
Preposition drop: conclusions

• We seem to have two constructions:
  – (1) which allows the definite article, is stable in apparent time, and restricted to north Wales, Liverpool, Merseyside and Cheshire
  – (2) which does not allow the definite article, is increasing in apparent time, and is diffusing in space in apparent time with the west Midlands as the most advanced area

• The prediction that the construction would be most unconstrained in the region in which it is oldest is effectively the claim that (2) is the precursor to (1)—the problem with this is that the area which allows construction (1) is actually relatively conservative w.r.t. construction (2)

• We could perhaps imagine that construction (1) is indeed older and stable, and the innovation of (2) was due to dialect contact between varieties which allow (1) and those which don’t—this would explain why the most innovative regions for (2) are not the regions which allow (1) but instead are regions neighbouring them
was/were-leveling: background

- Long history of variation in the past tense paradigm of to be in English
- Northern Middle English varieties showed leveling to was...
  - regularly in the second person singular
  - in plural contexts with full noun subjects
  - less frequently in plural contexts with pronoun subjects (Mossé 1952; Forsström 1948)
- Fast forwarding to the 20th c., the SED gives us data on the 1sg, 1pl, pronominal 3sg and 3pl, but no data on the 2nd person or on full noun subjects
Leveling in the past tense of *to be* with pronominal subjects in the SED.
was/were-leveling: background

- Sociolinguistic studies show that non-standard patterning of *was* and *were* in the past tense of *to be* is found in most varieties of English.
- Non-standard *was* generally more frequent than non-standard *were* (Tagliamonte 1998:157).
- Subject to grammatical constraints (Tagliamonte 1998:154, 157-8):
  - for non-standard *was*, there is a hierarchy of subject types: existential NP > you > plural NP > we, they
  - non-standard *were* is particularly associated with negatives, particularly with negative tag questions.
- Both kinds of leveling typically associated with class and education (non-standard variants used at higher rates by working class speakers and/or those with lower degree of education) (Tagliamonte 1998:154).
- Kerswill & Williams (2002) report finding that non-standard *was* is particularly high frequency in Hull, of the towns they look at, suggesting some more complex spatial distribution than ‘vernacular universal’ (and a different spatial distribution to that seen in the SED).
- As a very old non-standard variable strongly associated with class, we would expect this to be a variable of which speakers are conscious: as such it is more of an open question whether speakers represent this variation in writing, on Twitter.
was/were-leveling: data collection

- For this initial study, we searched for past tense forms of *to be* with singular pronoun subjects.
- As expected, the rate of non-standard usage is relatively low, especially for *were*-leveling.

<table>
<thead>
<tr>
<th>subject</th>
<th>was</th>
<th>were</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>124,520</td>
<td>61</td>
</tr>
<tr>
<td>he</td>
<td>356,816</td>
<td>421</td>
</tr>
<tr>
<td>she</td>
<td>46,047</td>
<td>21</td>
</tr>
<tr>
<td>you, yous</td>
<td>7774</td>
<td>75,895</td>
</tr>
</tbody>
</table>
was/were-leveling: background

- Aim in this initial study is to test whether it is possible to investigate this variable using twitter data...
- ...and to establish an overall geography of was- and were-leveling and see whether this matches what we might expect from the historical literature and/or from the SED
- Next steps after this will be to widen the range of subjects and syntactic contexts examined so as to be able to explore some of the grammatical constraints previously reported
was-leveling: overall spatial distribution
- Obvious mismatch with the SED distribution...
- But matches Kerswill & Williams (2002)'s observation that non-standard was is characteristic of Hull

**was-levelling:** overall spatial distribution
was-leveling: overall spatial distribution

- Though other hot-spots for was-leveling on Twitter are relatively consistent with the historical picture
was-leveling: change over time by year of birth
was-leveling: change over time by year of birth

- Unlike in the case of preposition drop, it seems unlikely that this apparent-time pattern reflects ongoing change and more likely that it reflects different relationships with the written standard for users of different ages.
was-leveling: spatial change over time by age category
were-leveling: overall spatial distribution
Finding that were-leveling is very highly localised is relatively unexpected, given the historical situation. But the hotspot for were-leveling on Twitter we see here is indeed within the historical were-leveling area.
were-leveling: change over time by year of birth
were-leveling: spatial change over time by age category
were-leveling: spatial change over time

- Note mismatch between overall apparent-time distribution (n-shaped curve) and the apparent-time distribution we see here (increase in apparent time)
were-leveling: change over time by year of birth and region

Dividing the data shows that there are quite different apparent-time patterns in West Yorkshire vs. the rest of the dataset.
were-leveling: change over time by year of birth and region

...but overall, given the extremely small size of the dataset, we probably shouldn’t take these apparent time effects very seriously.
Conclusions

• For a relatively new variable like preposition drop, generally below the level of speaker consciousness, twitter offers us an effective tool to investigate spatial and time distributions at scale
  - It also allows us to see the differing syntax of the construction in different regions
  - Next steps for this variable will be to examine other reported regional syntactic differences: possibility of prep drop with *at*; range of verbs

• For an older variable like leveling in the preterite paradigm of *to be*, where the nonstandard variants are clearly proscribed against, twitter is a less effective tool
  - It’s not clear that the spatial or temporal patterns are plausible reflections of speech, where presumably use of the nonstandard variants is higher frequency and more widespread


References


Thanks very much for listening!

David Willis
Deepthi Gopal
Tam Blaxter
University of Cambridge

Adrian Leemann

Universität Bern