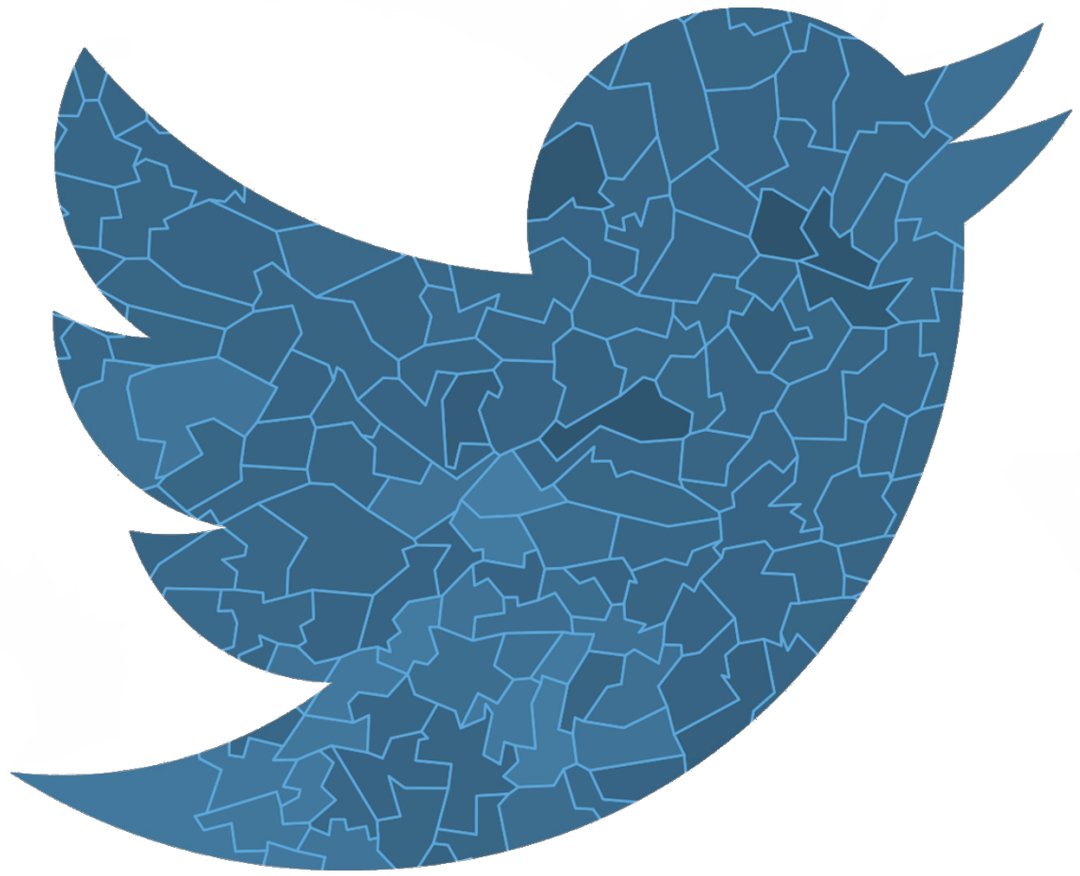
Tweetin’ in your own voice

Parallels between written and spoken (ing)

George Bailey
University of York



NWAV 49, University of Texas at Austin

23 October 2021

This study

cross-modal variation

phonetics

orthography

Language variation and change on Twitter

A range of topics have been addressed using Twitter as a source of sociolinguistic data

Tracing lexical innovations

(Grieve, Nini & Guo 2017)

Morphosyntactic variation

(Stevenson 2016; Willis 2020)

Stylistic variation in orthography

(Ilbury 2019)

Written form of sociophonetic variables

(Eisenstein 2015; Tatman 2016)

Language variation and change on Twitter

Tracing lexical innovations

(Grieve, Nini & Guo 2017)

Morphosyntactic variation

(Stevenson 2016; Willis 2020)

Stylistic variation in orthography

(Ilbury 2019)

Written form of sociophonetic variables

(Eisenstein 2015; Tatman 2016)

Dialect writing

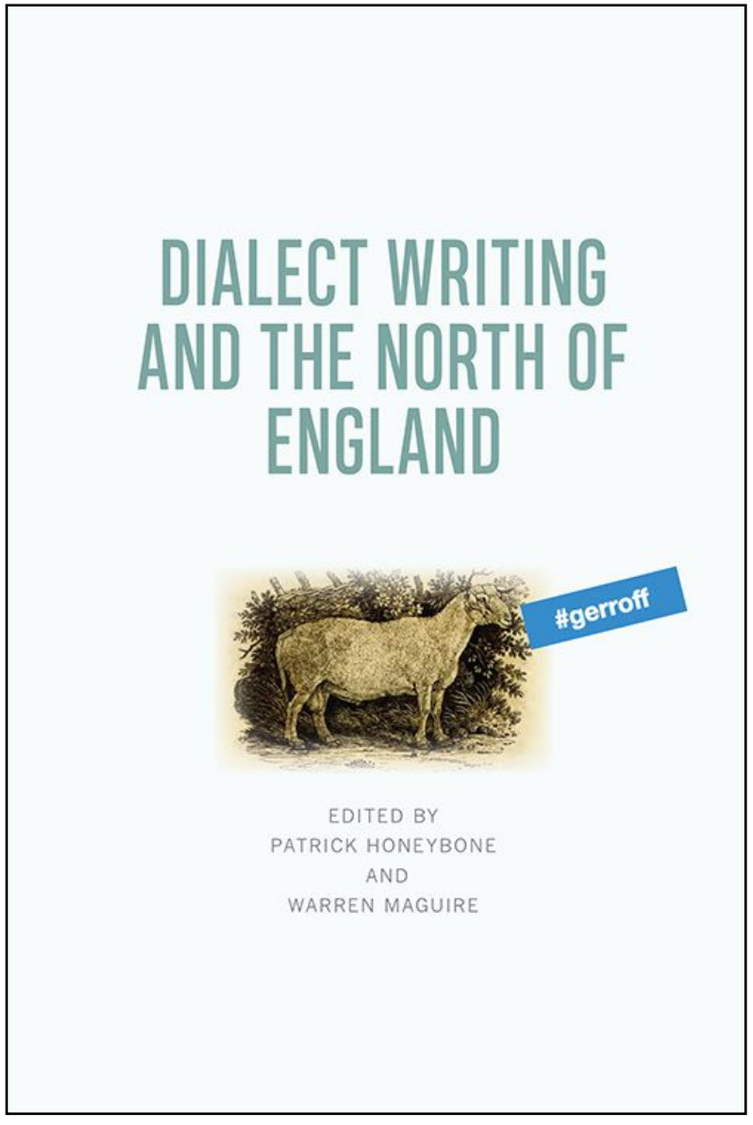
- *Dialect writing*: The representation of a ‘non-standard’ dialect in written form, involving one or more of the following:
  - dialectal lexis
  - dialectal morphosyntactic structures
  - dialectal ‘respellings’ to reflect phonological features

(Honeybone & Maguire 2020)

- Found across a range of texts, including poetry, novels, cartoons, tourist souvenirs, and **tweets**
- Studies of dialect writing lend insights into cultural salience of linguistic features, identity construction, and dialect enregisterment (Agha 2007)

Wider context

Already established parallels in regional patterns of sociophonetic variation and phonetically-motivated orthography



The graphical representation of phonetic dialect features of the North of England on social media, in collaboration with:

- Andrea Nini (University of Manchester)
- Diansheng Guo (University of South Carolina)
- Jack Grieve (University of Birmingham)

some examples...

TH-stopping

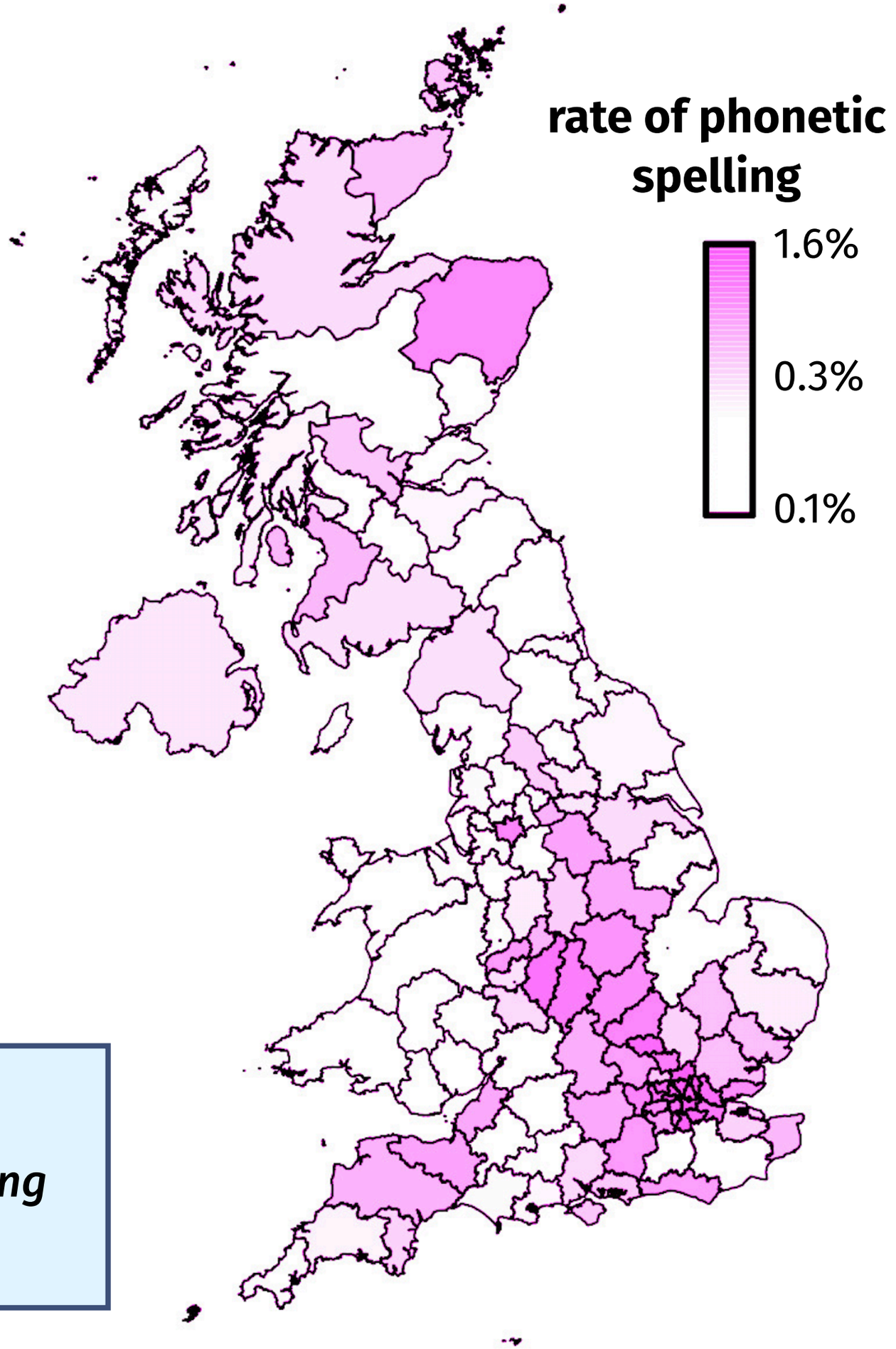
e.g. *then* [**ð**ɛn] ~ *den* [**d**ɛn]

- Strong ties to Multicultural London English but more strongly associated with performance of ethnic rather than regional identity (Drummond 2018)

- **On Twitter**: most frequent in London but also areas of the Midlands and Manchester

Example tweets:

- *fam **dis** trip every day is jus a long **ting***
- ***dese** man jus vexing my life*

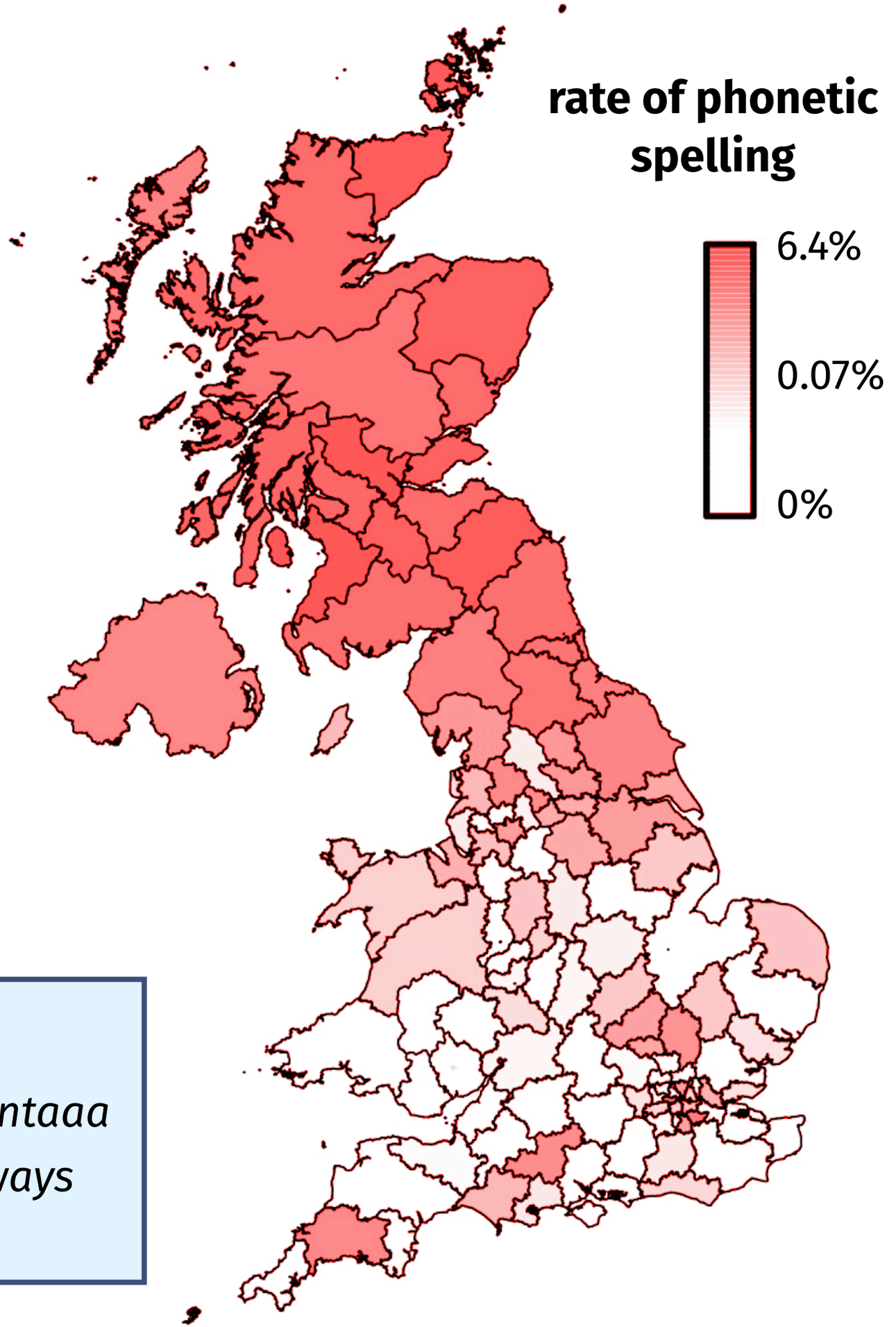


MOUTH as /uː/

e.g. *down* [daʊn] ~ *doon* [duːn]

- Retention of /uː/ in MOUTH characteristic of Tyneside English (Hughes et al. 2012) and Scots (Johnston 1997)

- **On Twitter**: highest rates found in North East England, and Scotland



Example tweets:

- *I miss you too and the **doon toon** bantaaa*
- *Was going to go for a nap but as always **cooncil** are cutting the grass*

‘Dialect writing’ on Twitter

Investigated regional patterns for the following variables:

T-to-R

HAPP**Y**-laxing

(h)-dropping

MOUTH as /u:/

LETT**ER**-backing

(th)-fronting

FOOT-STRUT split

g-dropping

(th)-stopping

vowel reduction

consonant reduction

This study

g-dropping

(ing)

[ɪŋ] → <ing>

[ɪn] → <in> <in’>

To what extent do the factors influencing (ing) variation **in speech** also play a role in this **orthographic** (ing) variation?

(ing)

- Very well-studied in speech, in both American and British varieties of English
- The ‘staple’ sociolinguistic variable (Hazen 2006) with many conditioning factors:

part of speech

- ‘Nominal-verbal continuum’ widely reported
- General trend: more nominal categories favour ***-ing***, more verbal categories favour ***-in***
- Attested in American English (Labov 2001, Forrest 2017), British English (Houston 1985, Tagliamonte 2004), Australian English (Shnukal 1982) and New Zealand English (Bell & Holmes 1992)

(ing)

- Very well-studied in speech, in both American and British varieties of English
- The ‘staple’ sociolinguistic variable (Hazen 2006) with many conditioning factors:

region

- ***-in*** said to be favoured in the North of England and Scotland (Labov 2001: 90)
- Regional pattern mirrors isogloss for regions that first underwent *-inde* → *-ynge* replacement in OE/ME period (Houston 1985, 1991; Moore et al. 1935)

(ing)

- Very well-studied in speech, in both American and British varieties of English
- The ‘staple’ sociolinguistic variable (Hazen 2006) with many conditioning factors:

phonological environment

- Regressive assimilation: ***-in*** favoured before alveolar consonants, ***-ing*** favoured before velar consonants
- Progressive dissimilation: ***-in*** favoured after velar consonants, ***-ing*** favoured after alveolar consonants
- Fully (or at least partially) attested in many varieties (Houston 1985; Shuy et al. 1968; Cofer 1972; Watts 2005; Schleef et al. 2011; Bailey 2018)

(ing)

- Very well-studied in speech, in both American and British varieties of English
- The ‘staple’ sociolinguistic variable (Hazen 2006) with many conditioning factors:

lexical frequency

- Evidence that ***-in*** is favoured in ‘everyday’ words’, disfavoured in ‘specialised’ or ‘learned’ words (Wald & Shopen 1981; Tagliamonte 2004)
- No effect of frequency in Philadelphia (Abramowicz 2007), Edinburgh or London (Schleef et al. 2011), but significant effect reported in Raleigh, NC (Forrest 2017)

Methodology

Corpus creation

- Tweets collected in a 4-month period in 2016 using the Streaming API
  - sample of all tweets sent in real time, no filter on content
  - only geotagged tweets sent from within the UK
  - filtered out tweets sent from bot accounts (e.g. automated weather forecasts, traffic updates etc)
  - **16 million tweets ~ 183 million words**
- Ran each tweet through automated POS tagger
  - **twitie-tagger** (Derczynski et al. 2013) uses the Penn Treebank tagset
  - 91% accuracy rate, can deal with Twitter-specific 'words' such as https://hyperlinks, @usernames and #hashtags

Data collection

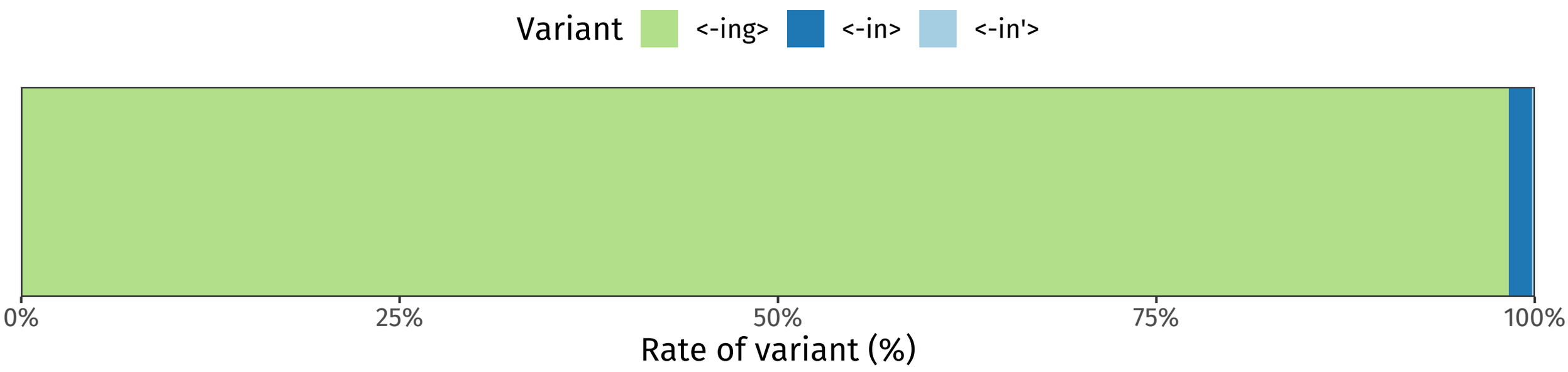
- Python script to extract (ing) tokens: search for all words ending with **<in>** or **<ing>**
- Cross-referenced with CMU pronouncing dictionary to check phonemic transcription ends in IH0/AH0 /ɪ, ə/ then N/NG /n, ŋ/
- Still lots of false positives:
  - names (e.g. *Dustin*, *Turing*) and polysemy (e.g. *puffin*, the bird!)
- Semi-automated cleaning:
  - removed individual words, and tokens in tweet-medial position where only the initial character is capitalised
- Each token coded for POS, audience (open vs. @-reply), preceding/following 'segment' (based on CMU pronouncing dictionary), and lexical frequency (based on SUBTLEX-UK corpus)

Results

Overview

4.4 million tokens of (ing)

<-ing>	<-in>	<-in'>
~4,300,000	~69,000	~7,000
98.27%	1.56%	0.16%

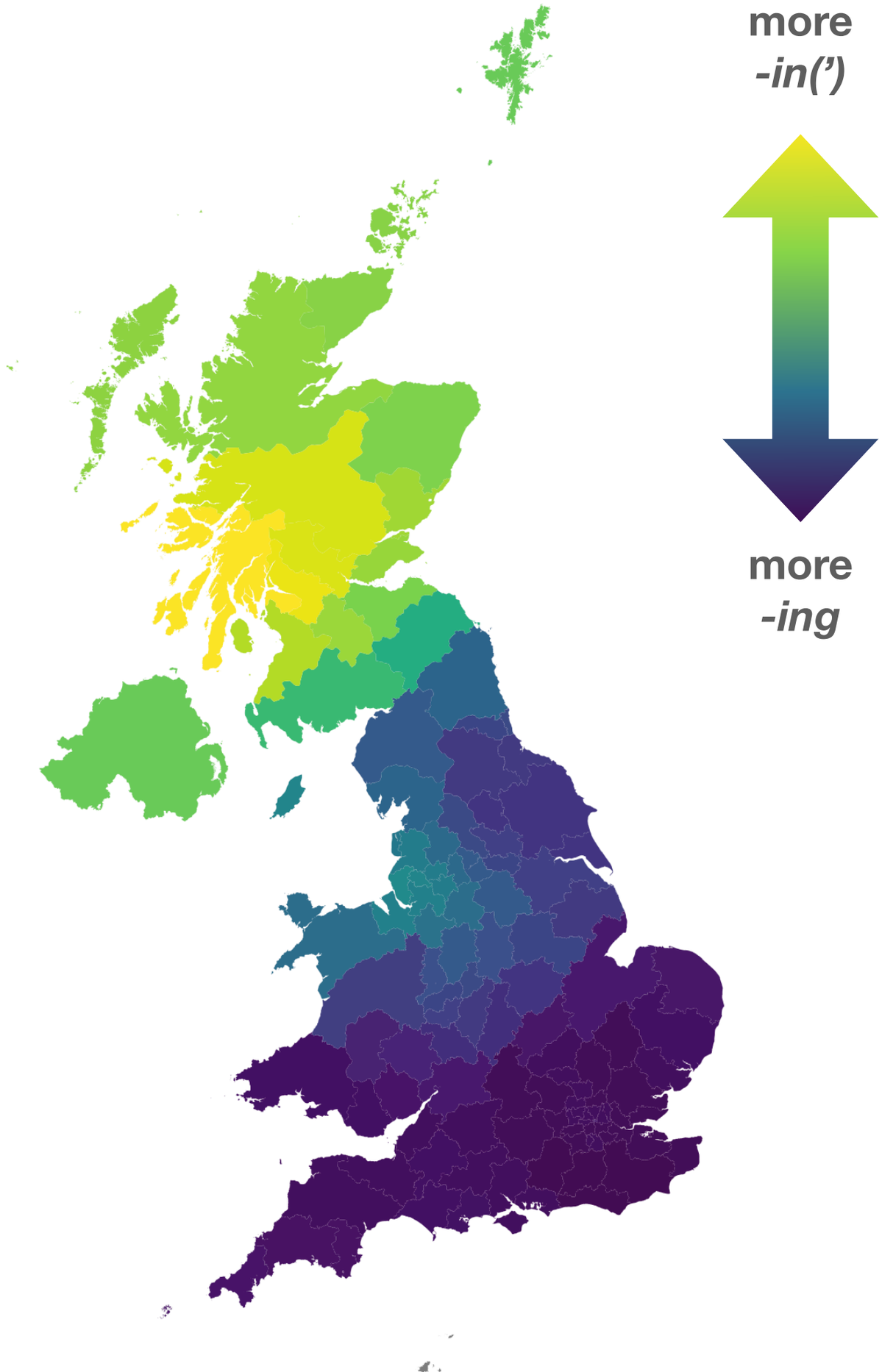


Region

- Hotspot analysis based on Getis-Ord *Gi** local spatial autocorrelation

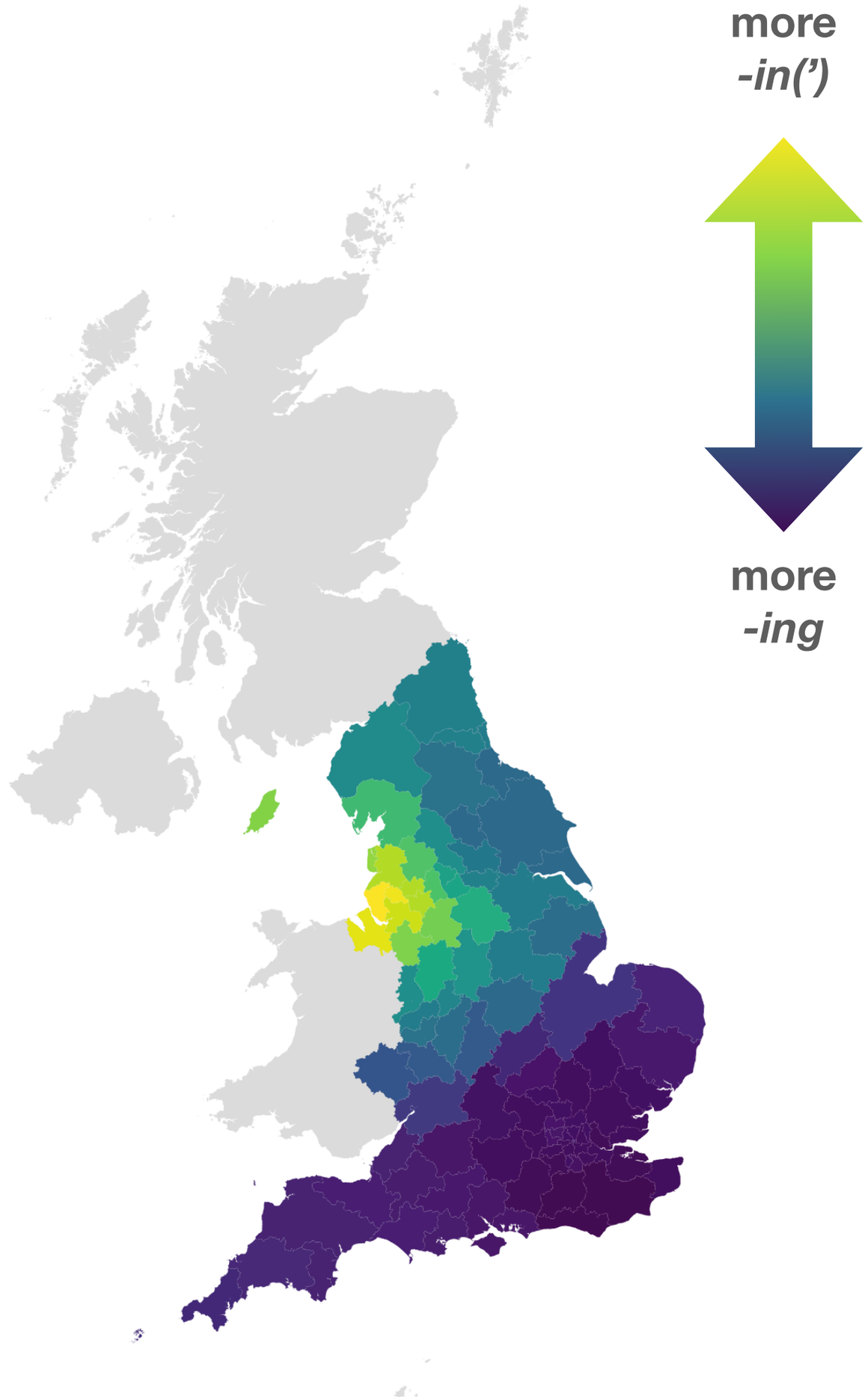
- Regional pattern mirrors that of spoken (ing)

- Highest rates of ***-in*** found in Scotland and the North of England



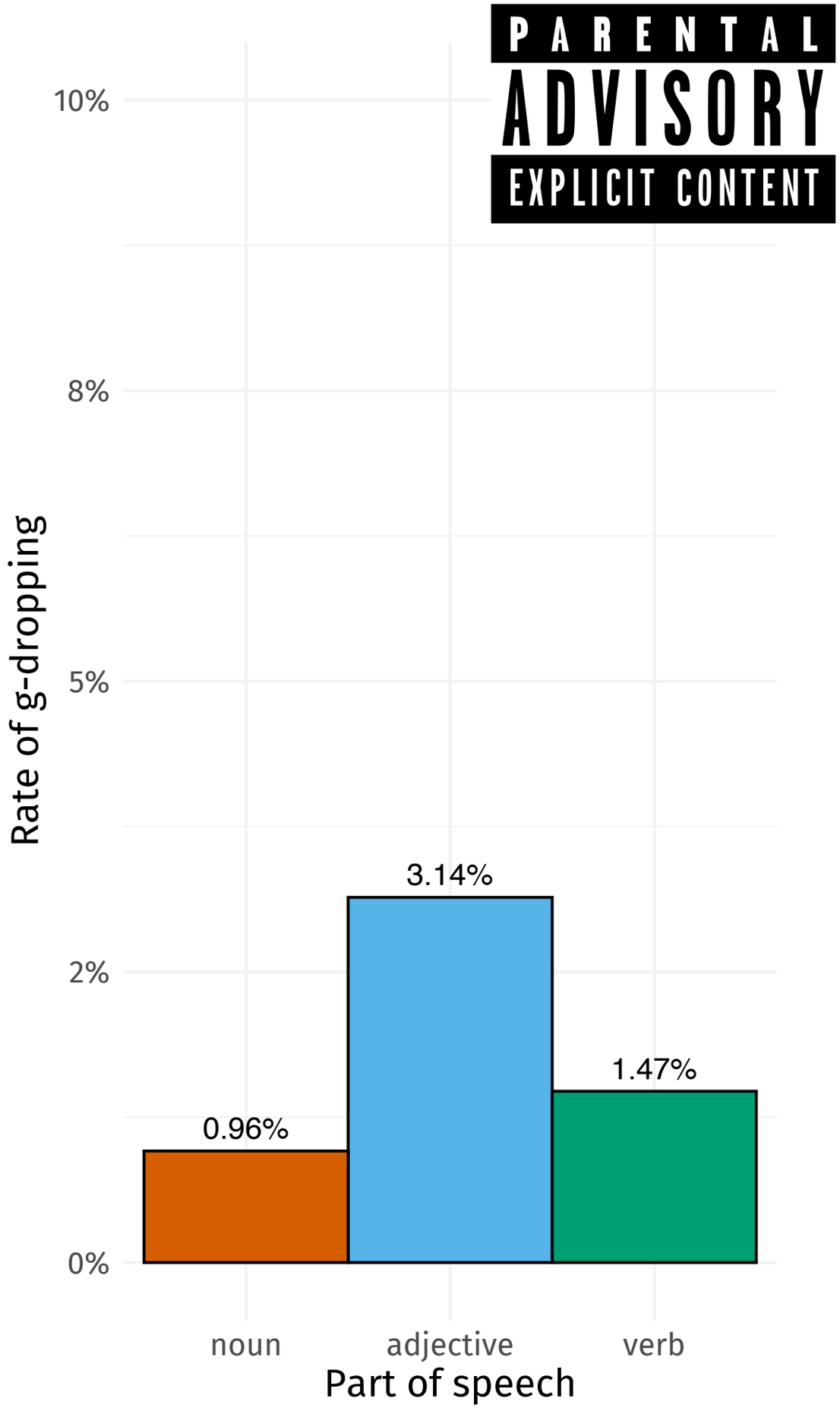
Region

- Hotspot analysis based on Getis-Ord *Gi** local spatial autocorrelation
- Regional pattern mirrors that of spoken (ing)
- Highest rates of ***-in*** found in Scotland and the North of England



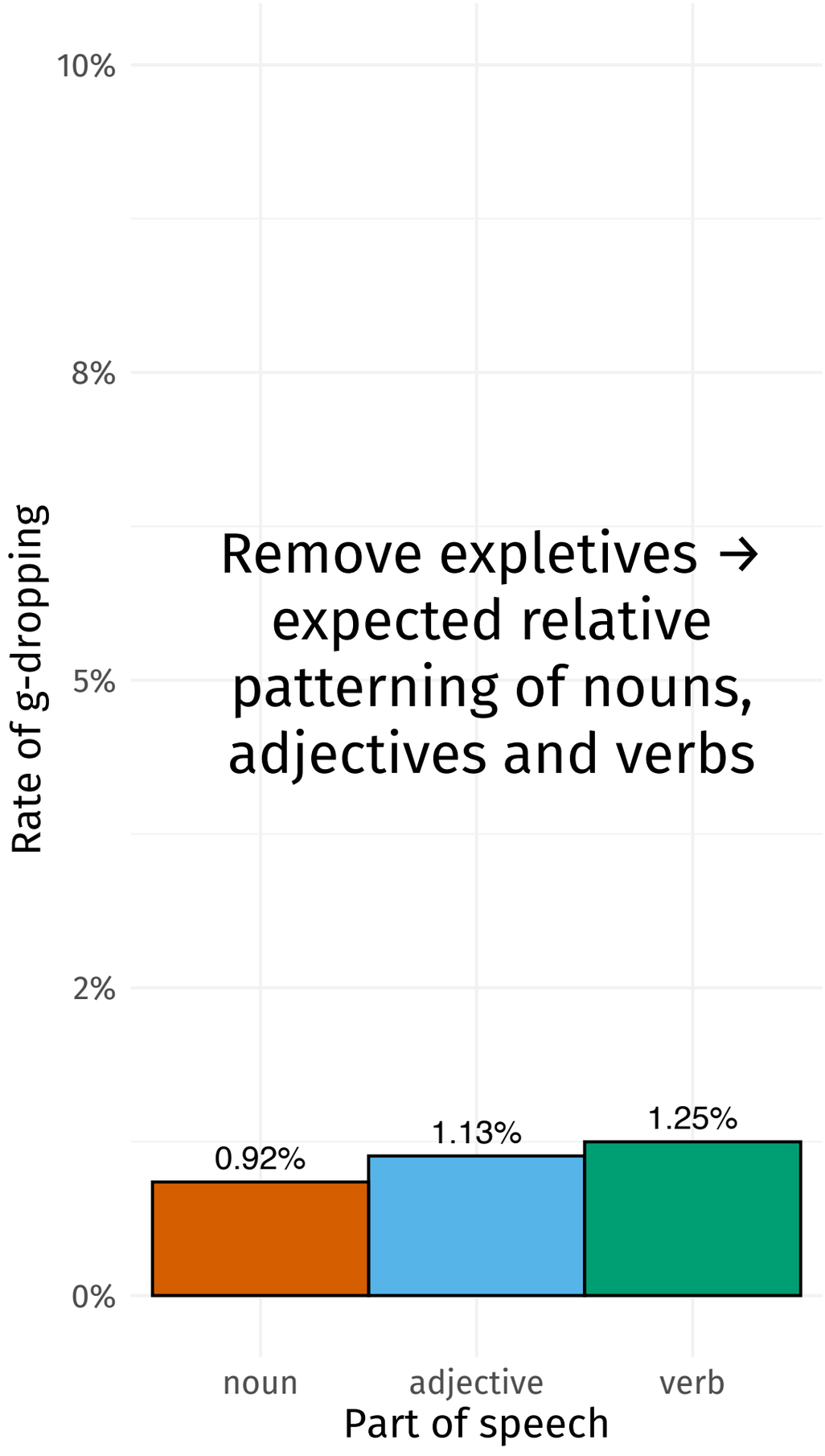
Part of speech

- More ***-in*** found in verbs (N=3.4m) than nouns (N=600k), mirroring the spoken variable
- But adjectives (N=300k) show a surprisingly high rate of g-dropping
- Driven by expletives tagged as adjectives: *fucking, motherfucking, freaking, frigging, fricking*

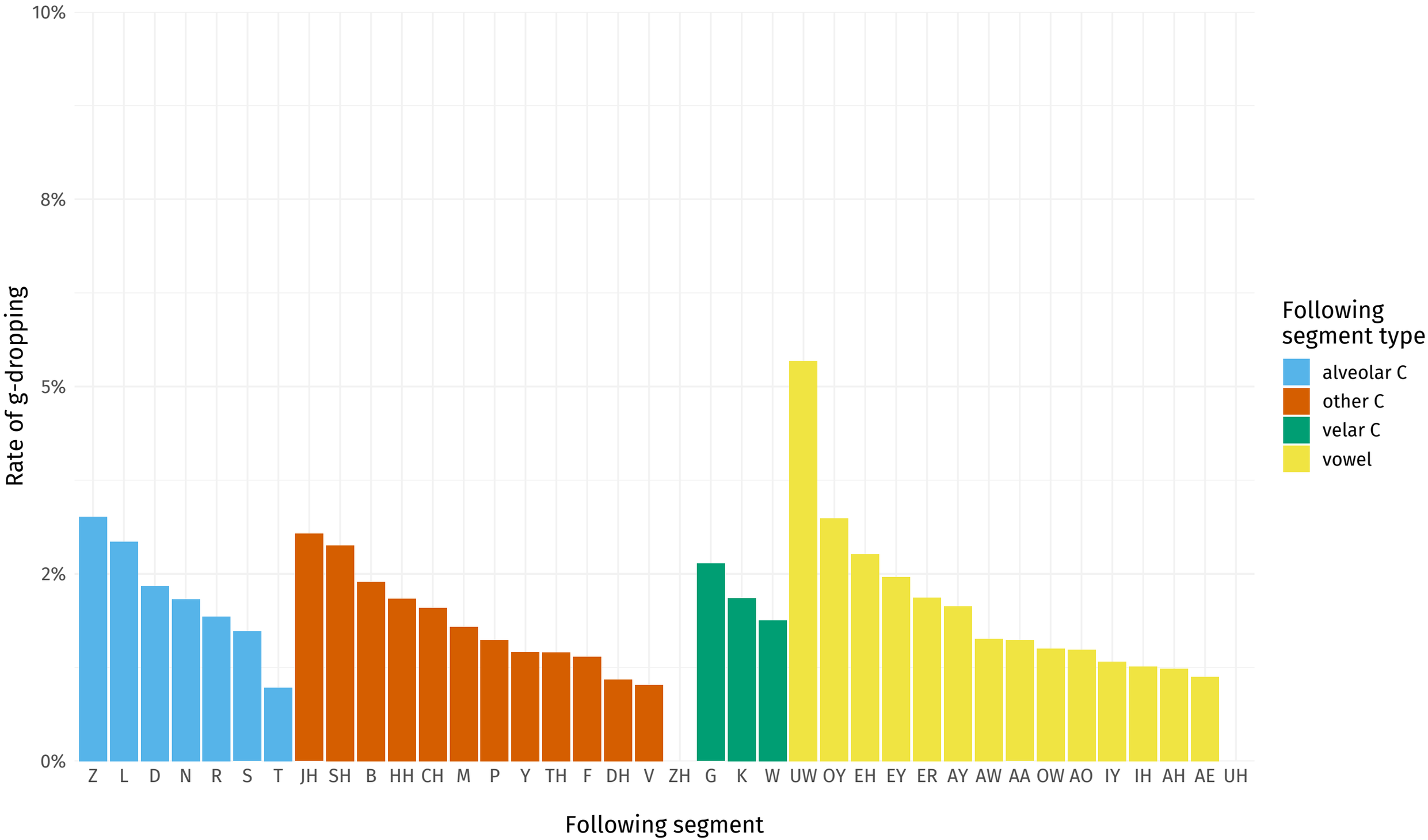


Part of speech (SFW)

- More ***-in*** found in verbs (N=3.4m) than nouns (N=600k), mirroring the spoken variable
- But adjectives (N=300k) show a surprisingly high rate of g-dropping
- Driven by expletives tagged as adjectives: *fucking, motherfucking, freaking, frigging, fricking*
- Small effect, but this mirrors work in BrEng anyway (Watts 2005 in **Colshaw** and **Wilmslow**, Bailey 2018 in **Manchester** and **Blackburn**, Schleef et al. 2011 in **Edinburgh** and **London**)

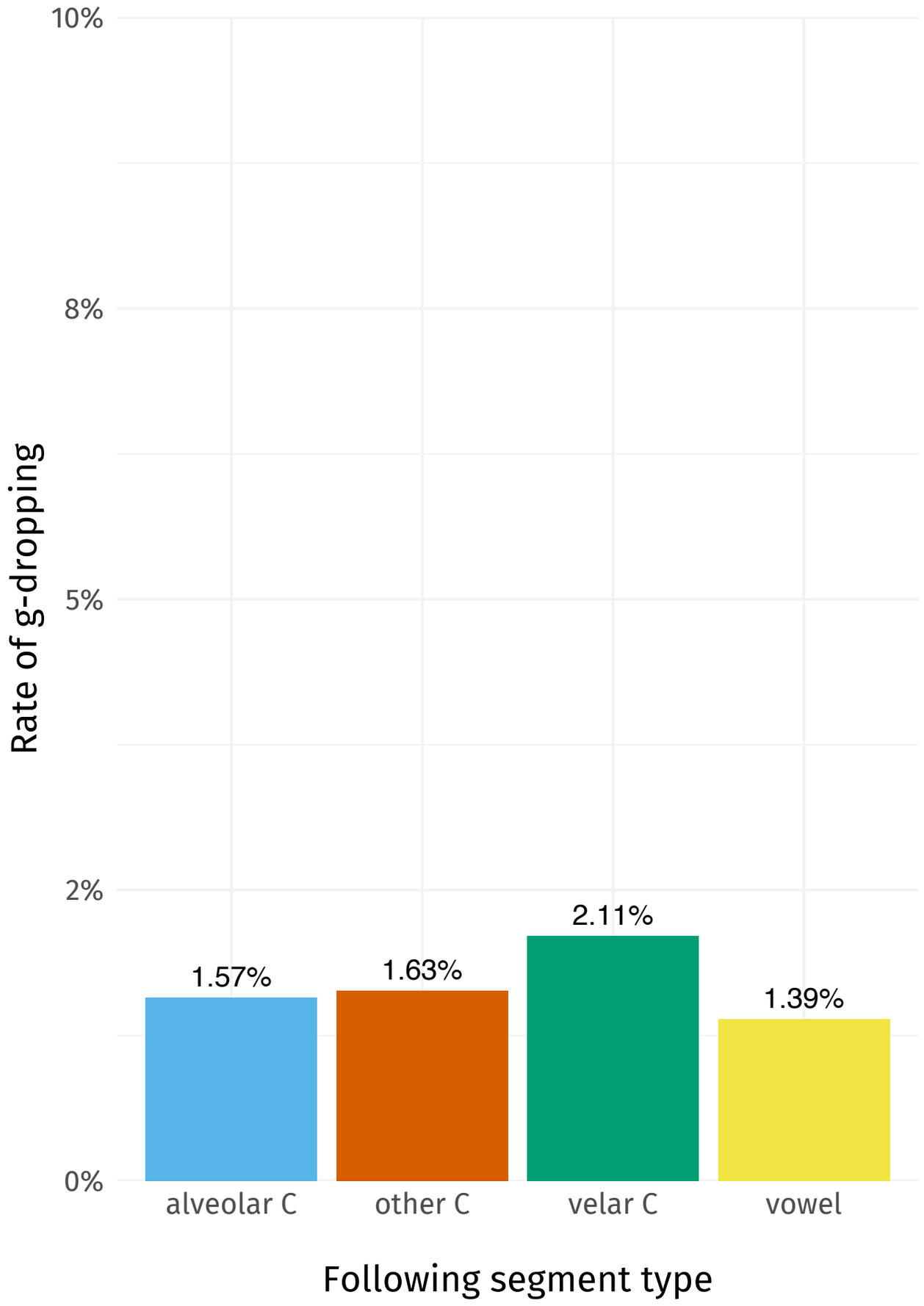


Following ‘segment’

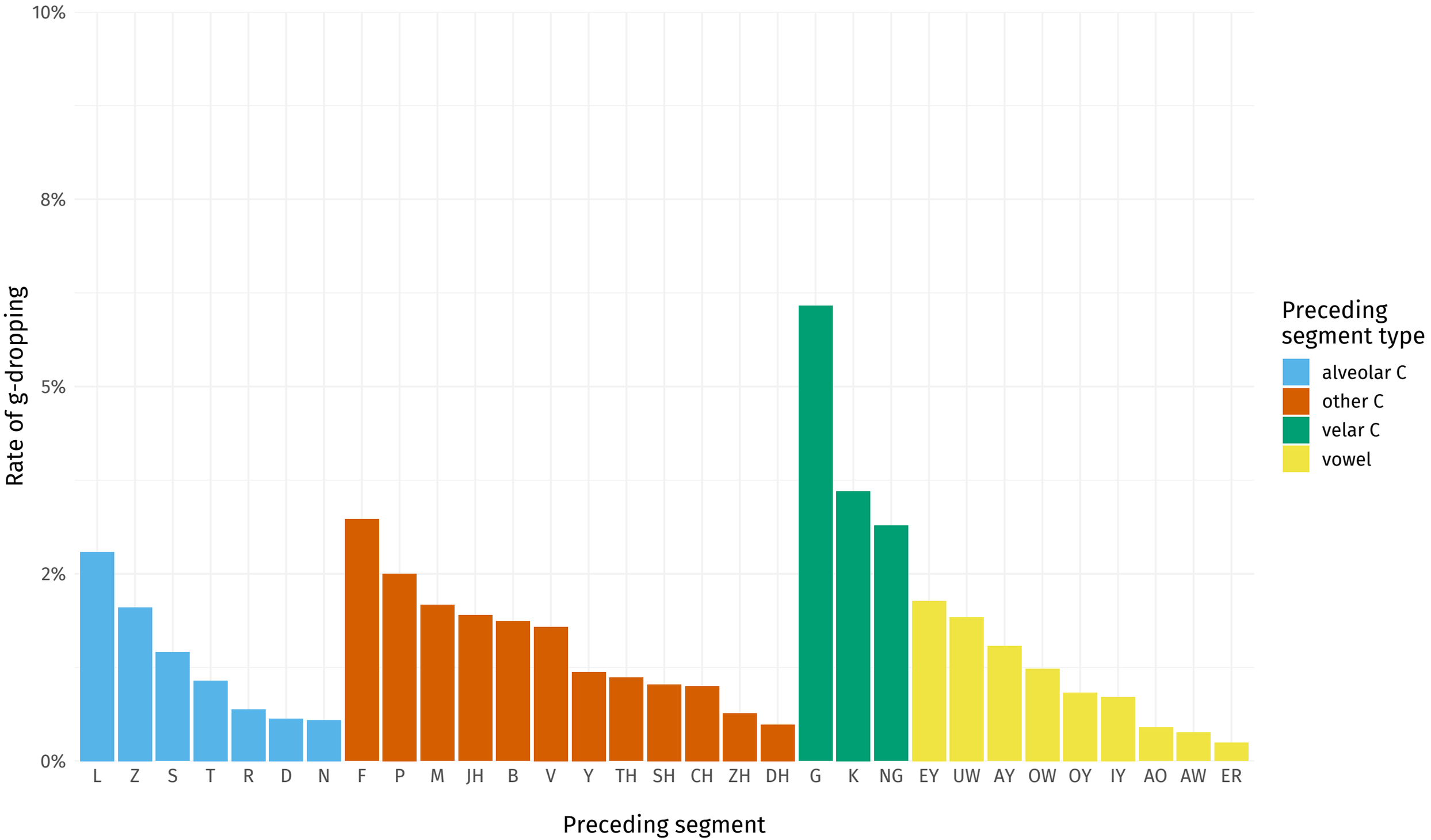


Following ‘segment’

- Not really an effect of following ‘segment’
- Slightly more ***-in*** before velar-initial words, contrary to results in speech

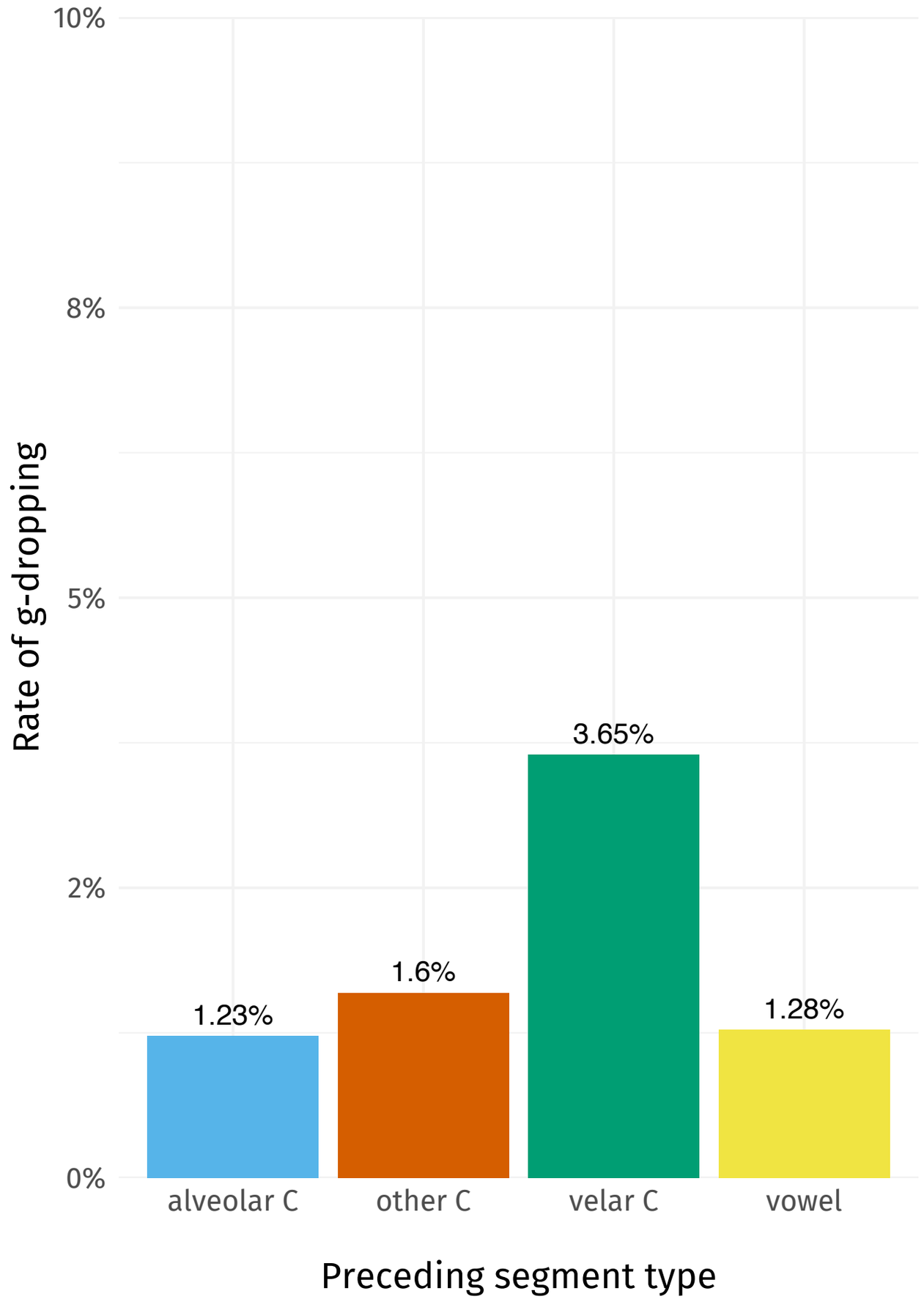


Preceding ‘segment’



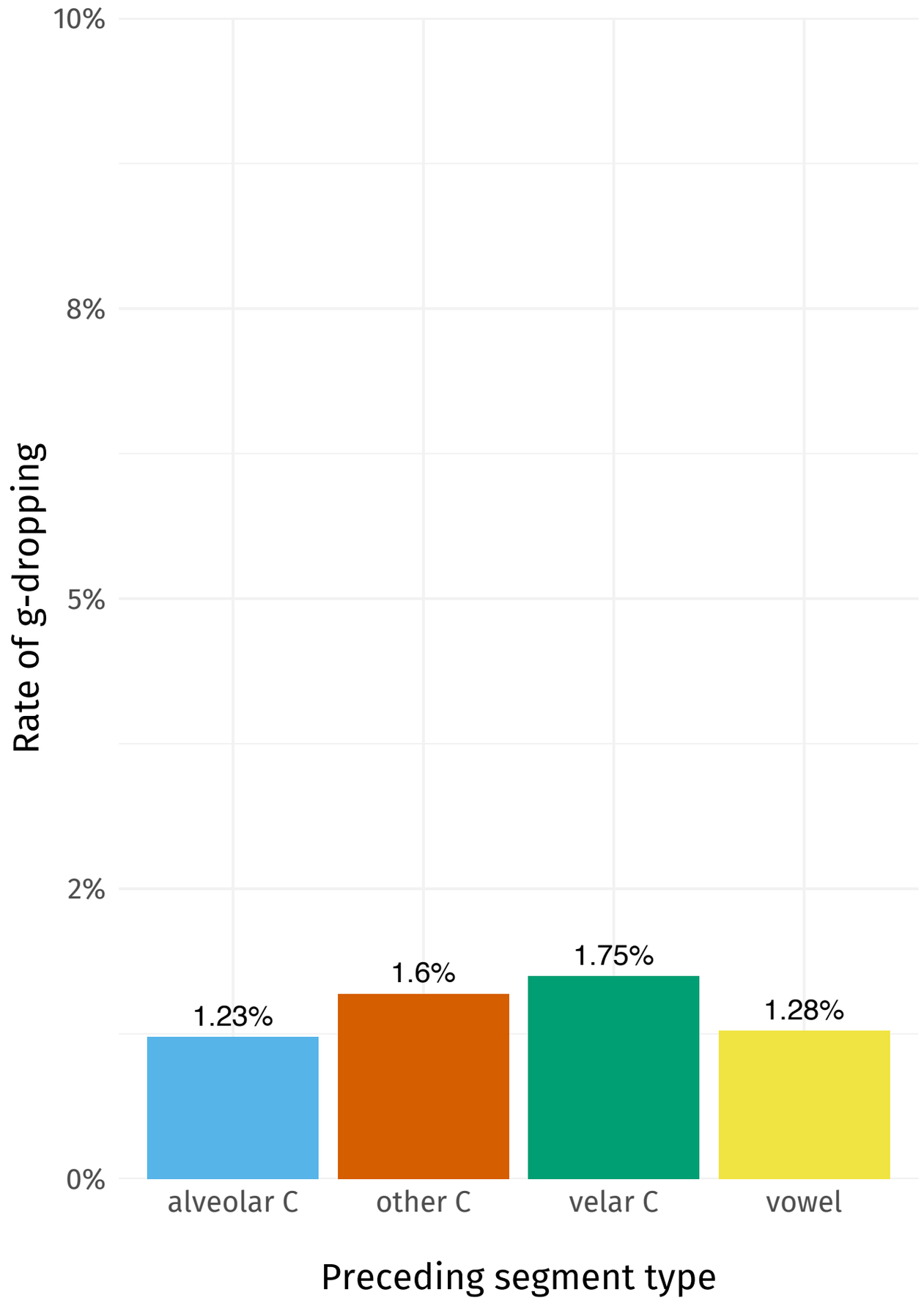
Preceding ‘segment’

- Evidence for a preceding segment effect, with more ***-in*** after velar consonants
- Why would we even expect this effect in written (ing)?
- Stronger effect for *preceding* segment, which is word internal, may reflect phonetically-rich representations at the word level

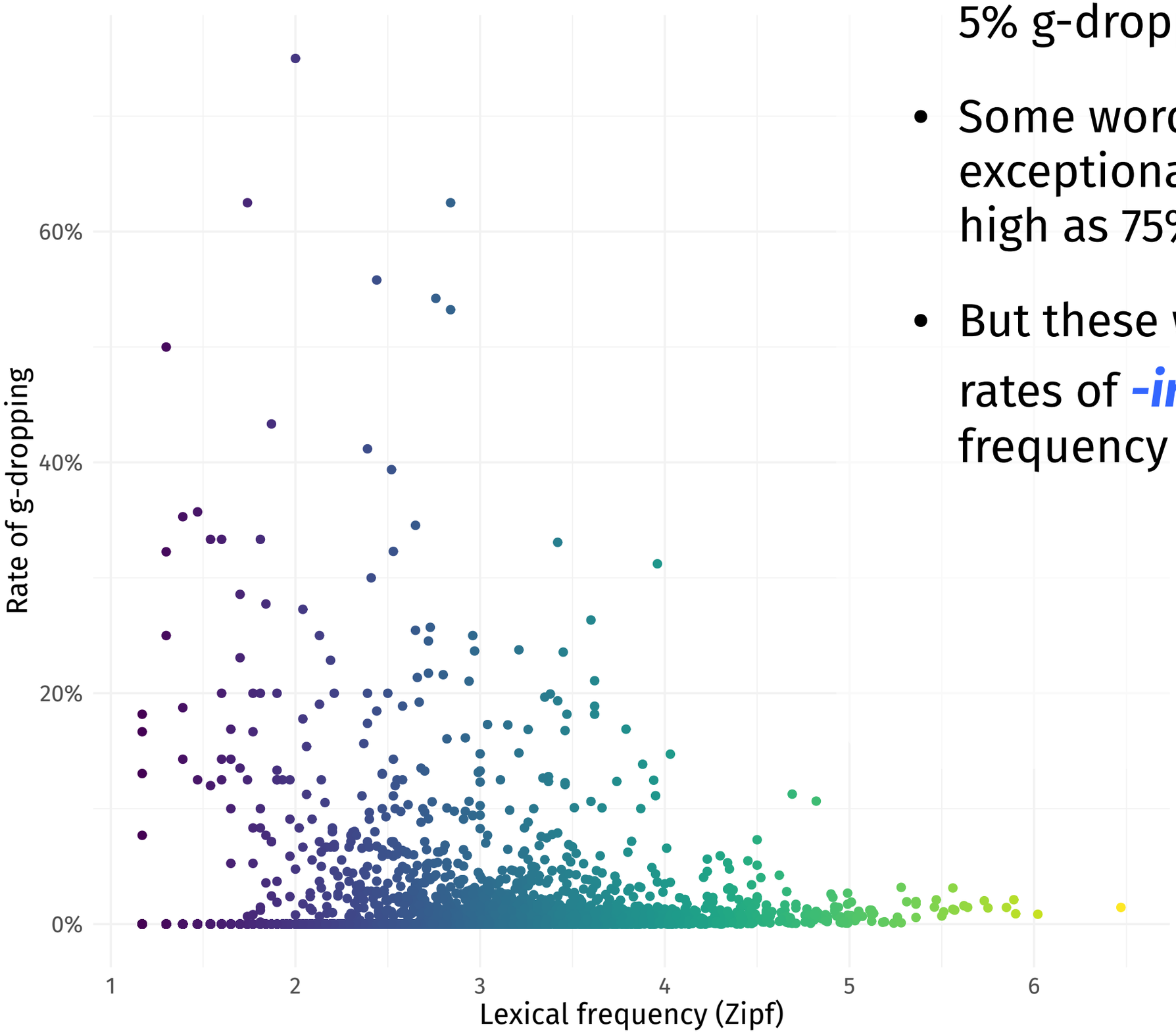


Preceding ‘segment’

- Remove expletives and this effect becomes much smaller in magnitude



Lexical frequency



- Most words have between 0–5% g-dropping overall
- Some words show exceptionally high rates, as high as 75%
- But these words with highest rates of ***-in*** tend to be *lower* frequency

Lexical frequency

- Most words have between 0–5% g-dropping overall
- Some words show exceptionally high rates, as high as 75%
- But these words with highest rates of ***-in*** tend to be *lower* frequency:
  - *nuttin'*
  - *boggin'*
  - *wildin'*
  - *summin'*
  - *reekin'*
  - *ballin'*
  - *bussin'*

Co-variation

- Some of these forms represent co-variation between (ing) and other sociophonetic variables
  - (ing) and **(th)-stopping** in *nothing* → <**nuttin**>

> Got **nuttin** but positive vibes

> hope they lose again, got **nuttin** against them but their fans are annoying

Co-variation

- Some of these forms represent co-variation between (ing) and other sociophonetic variables
  - (ing) and **(th)-dropping** in *something* → <**summin**>

> Gettin told **summin** bad when your at work is not what u want

- Much lower rates of ***-in*** when used as the verb *summing*:

> Woeful defending **summing** up our season. Out of our hands now

Conclusions

- Evidence for systematic (ing) variation on Twitter, but only *some* parallels between phonetic and orthographic variation

- Linguistic constraints play only a minor role, suggesting that social factors might be more central in cross-modal variation

- Stylistic nature of g-dropping on Twitter quite performative compared with its relatively low social profile in British English (Levon & Fox 2014)

- g-dropping rarely used in isolation but rather as part of a wider stylistic repertoire with other phonetic spellings in a socially meaningful way

Thanks for watchin’

Tweet

@grbails

george.bailey@york.ac.uk