Identifying and Mapping the Spread of Emerging Words

Jack Grieve
j.grieve1@aston.ac.uk
@JWGrieve
Centre for Forensic Linguistics
Aston University

2 December 2016
BAULT 2016
University of Helsinki

Emerging Words

Very large corpora of social media provide us with new opportunities to track newly emerging words as they enter into the general lexicon for the first time.

This study is based on a ~9 billion word corpus of geocoded American Tweets (~1 billion Tweets) collected between October 2013 and November 2014.

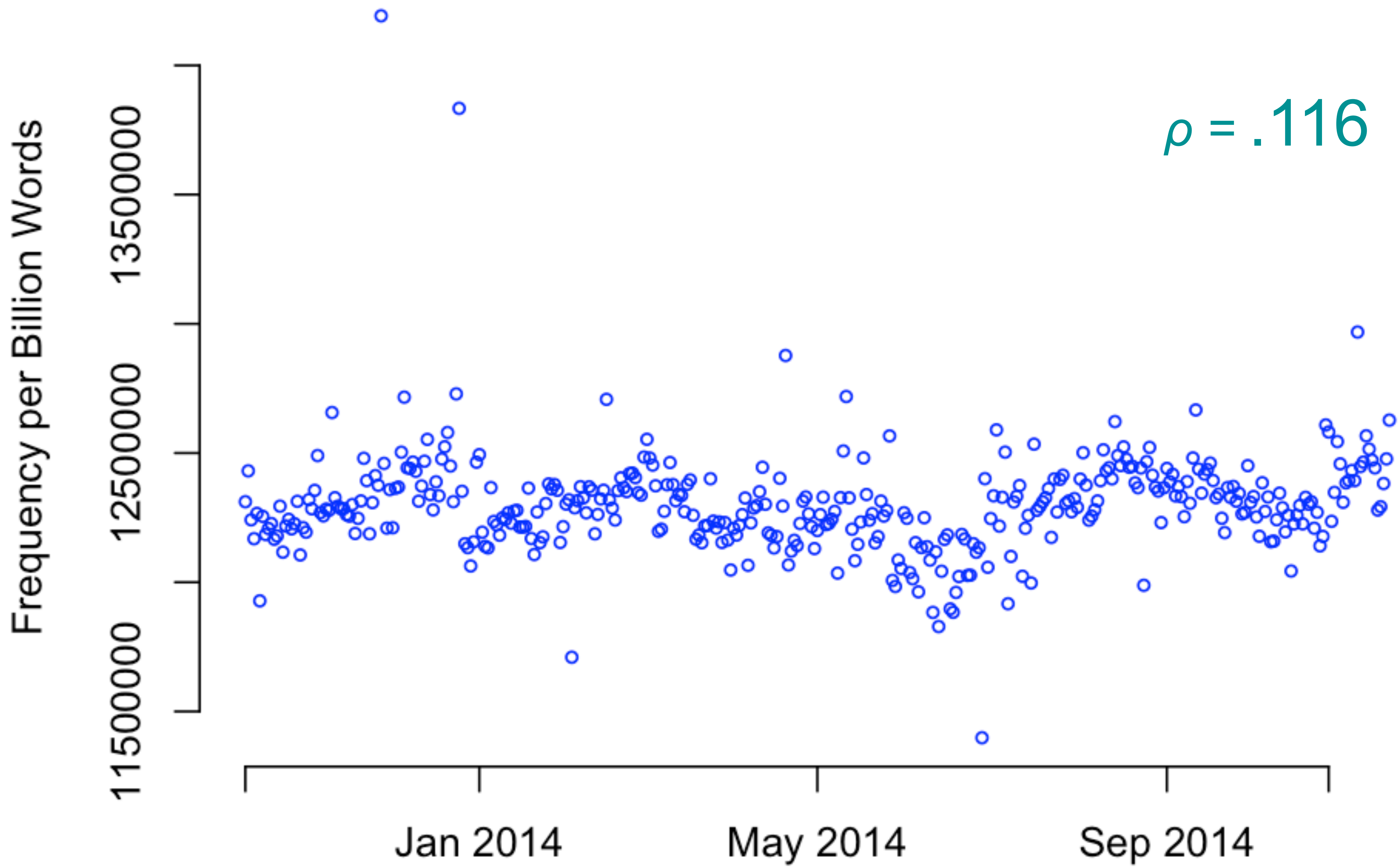
Emerging words are identified and their properties are considered and their usage mapped to identify common sources of lexical innovation in American English.

Rising Words

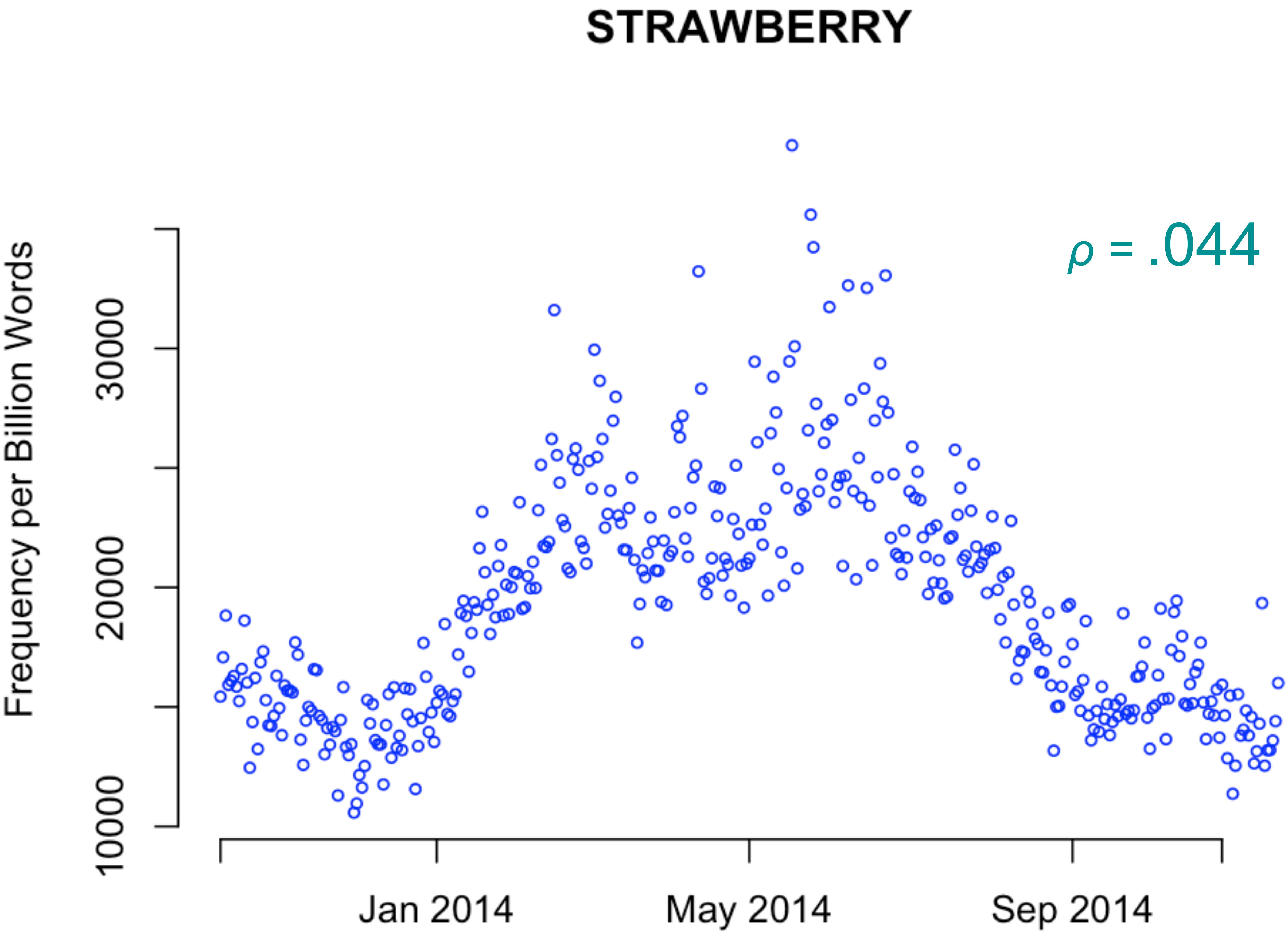
To find newly emerging words, the 97,246 words types that occur at least 500 times in the corpus were extracted from the corpus.

The word relative frequency per day was then compared to day of the year (across 399 days) using a Spearman's rank correlation coefficient to assess the degree to which the usage of each word in the corpus had been rising over the 13 month period.

AND



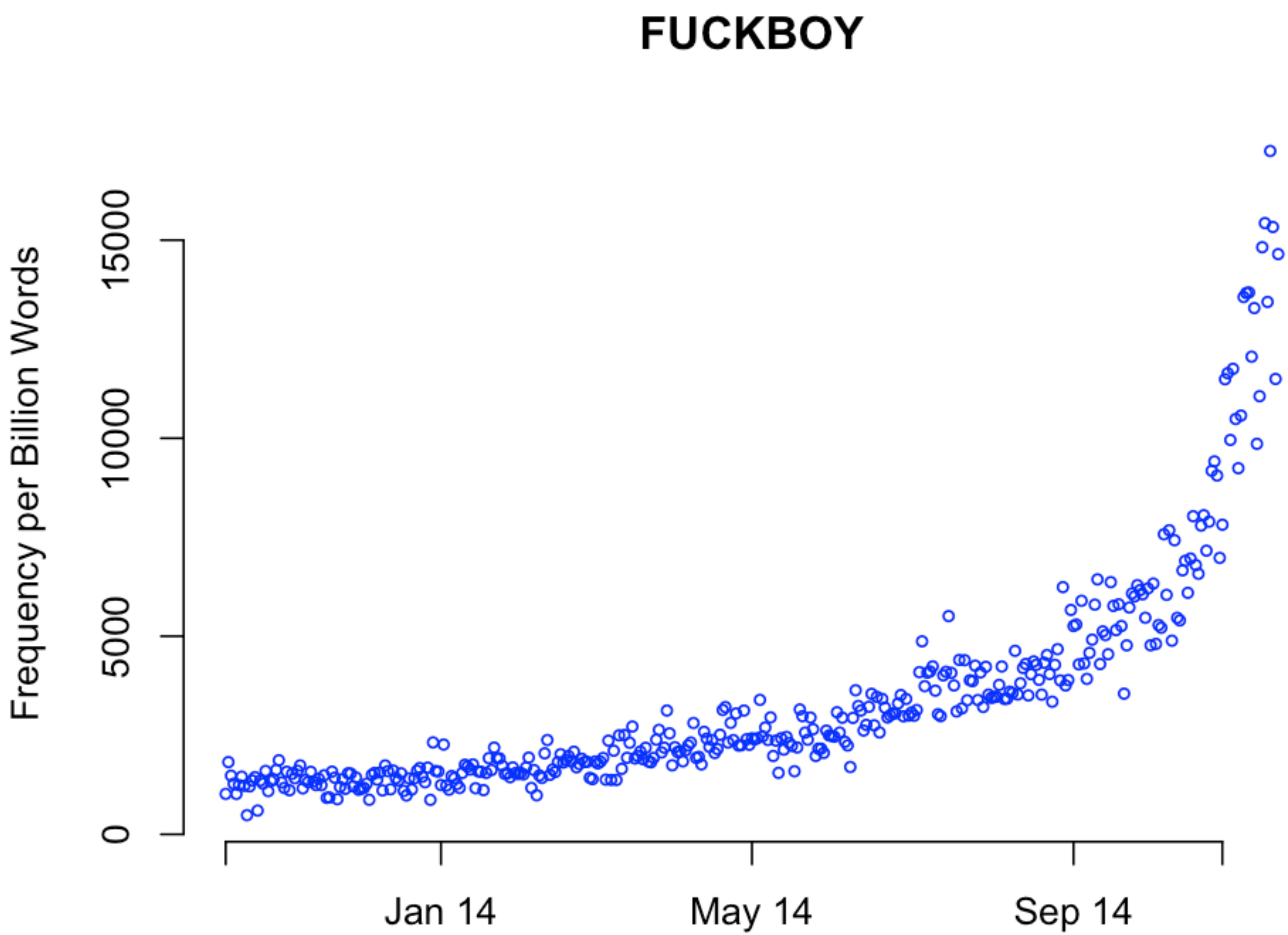
AND
ρ = .116
Frequency per Billion Words
115000000
125000000
135000000
Jan 2014
May 2014
Sep 2014

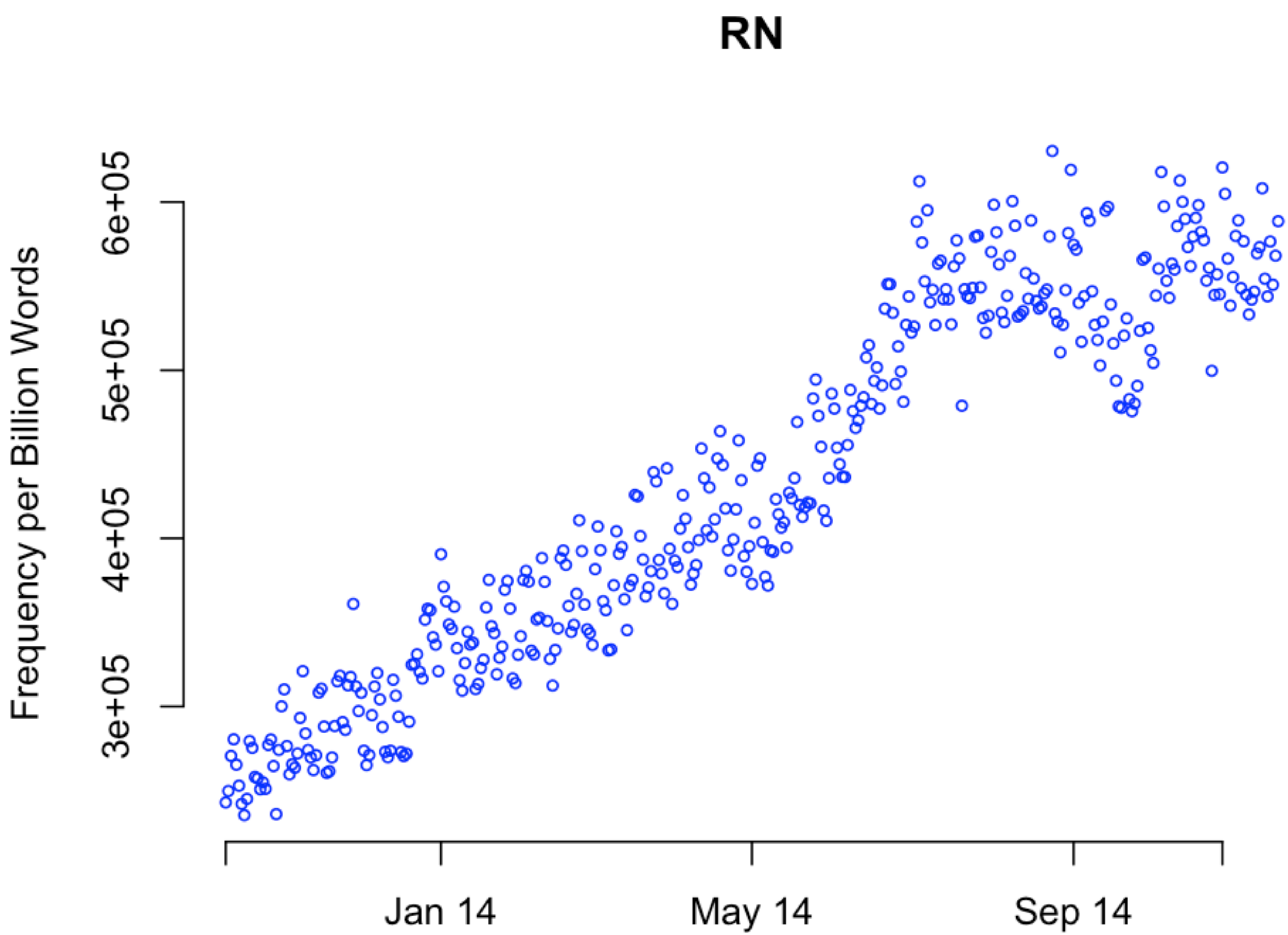


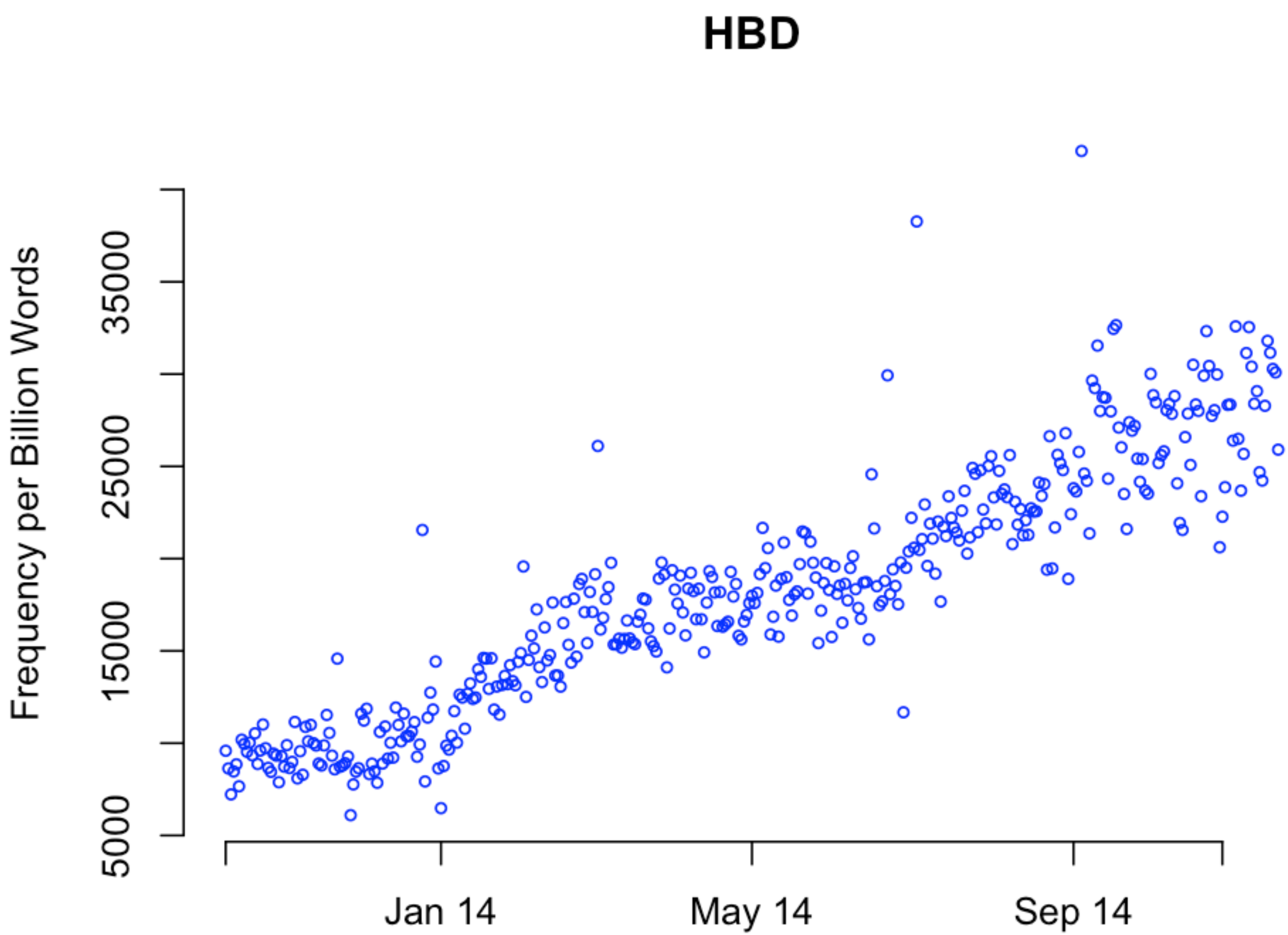
STRAWBERRY
ρ = .044
Frequency per Billion Words
10000
20000
30000
Jan 2014
May 2014
Sep 2014

Top 10 Rising Words on Twitter 2014

Word	ρ	**Definition**
fuckboy	0.947	Asshole, Jerk, Poser, Tool, etc.
rn	0.938	Right Now (Top Riser 2013 too)
hbd	0.928	Happy Birthday
fw	0.927	Fuck with
unbothered	0.926	Unconcerned & Disengaged
ft	0.925	Face time
gmfu	0.924	Get me fucked up
sm	0.919	So Much
squad	0.919	Group of friends
asf	0.918	As fuck


FUCKBOY
Frequency per Billion Words
0
5000
10000
15000
Jan 14
May 14
Sep 14


RN
Frequency per Billion Words
3e+05
4e+05
5e+05
6e+05
Jan 14
May 14
Sep 14



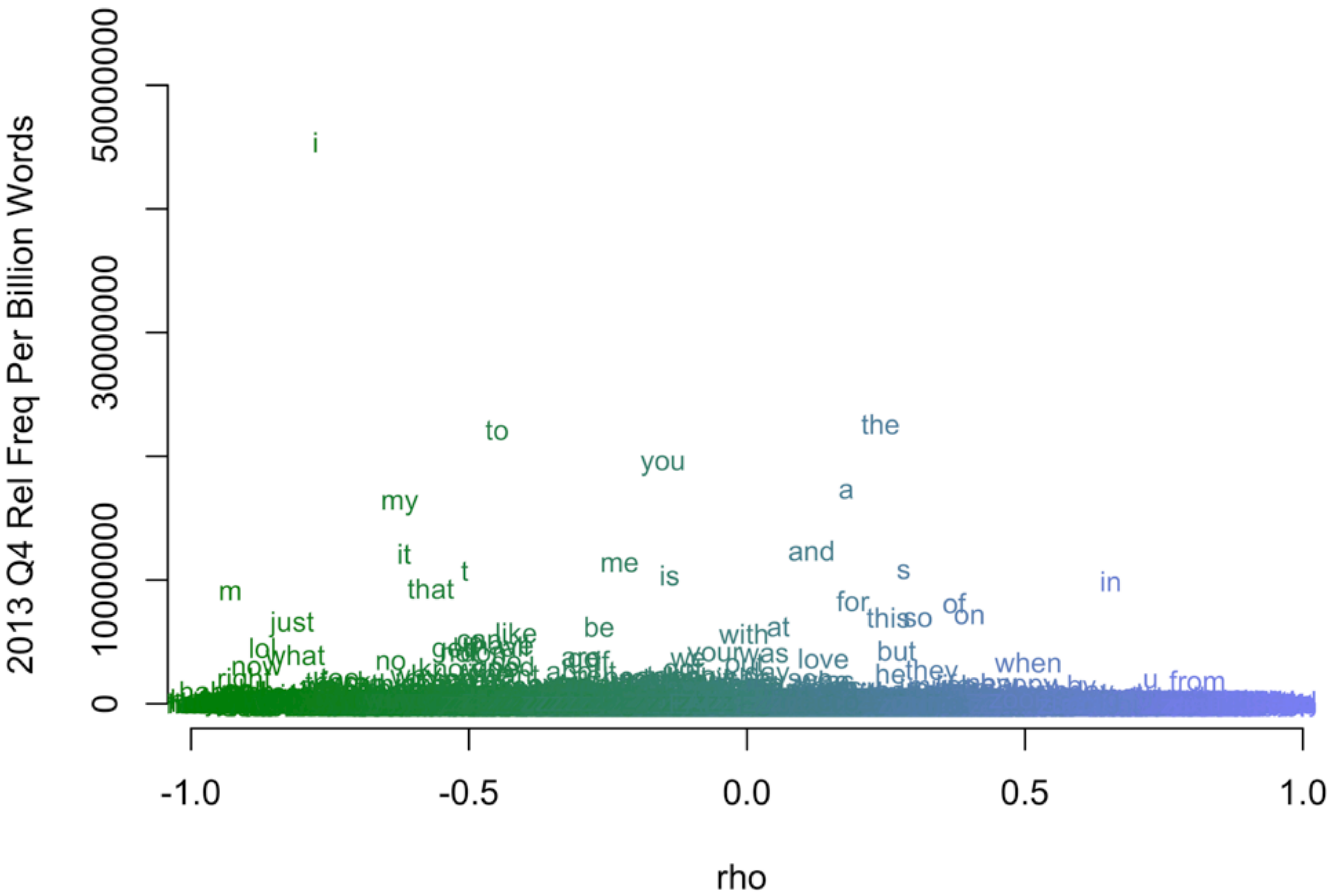
HBD
Frequency per Billion Words
5000
15000
25000
35000
Jan 14
May 14
Sep 14

Emerging Words

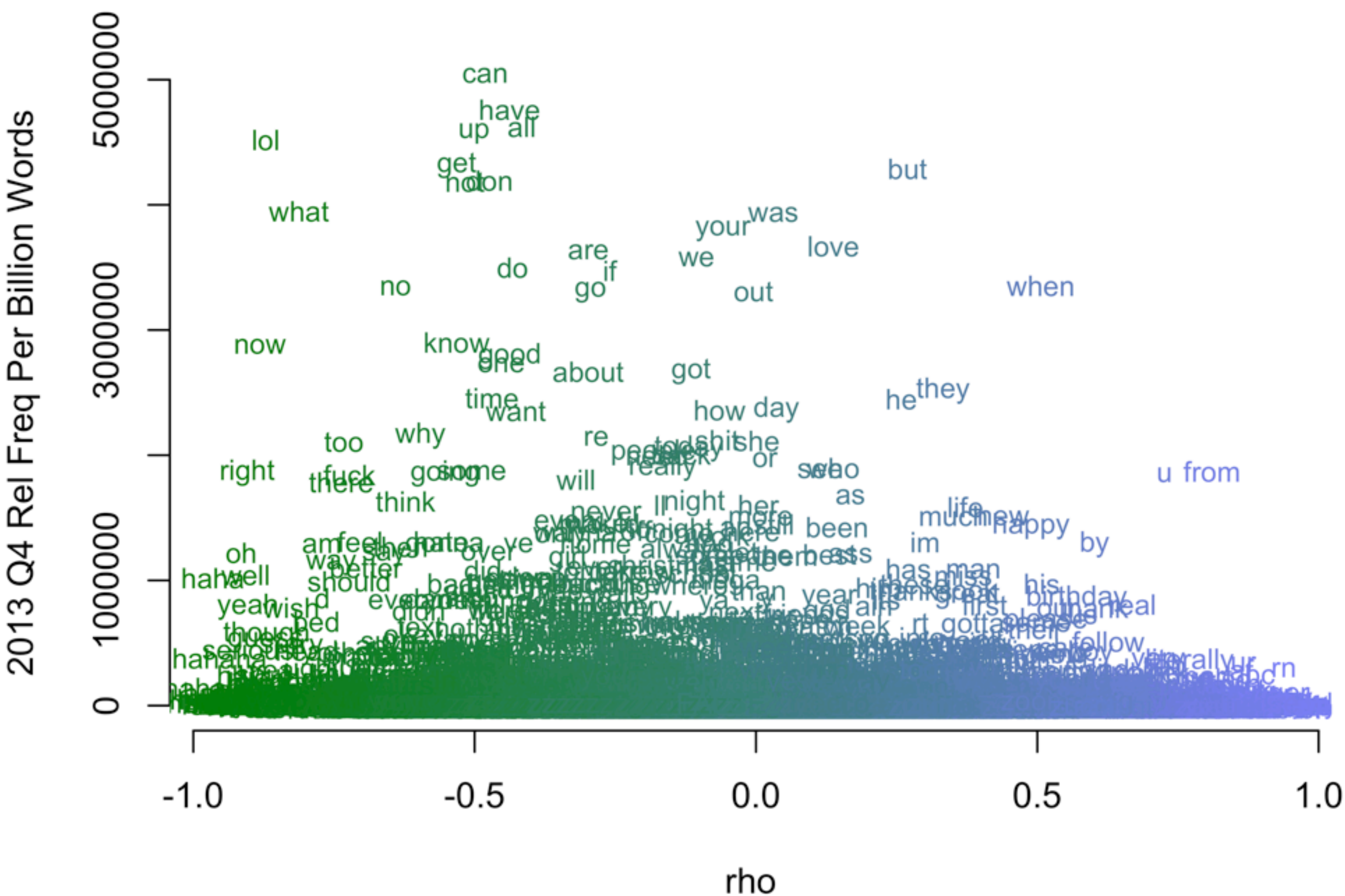
Although measuring correlations allows for rising words to be identified, most are far too common at the start of the time period to be considered emerging words.

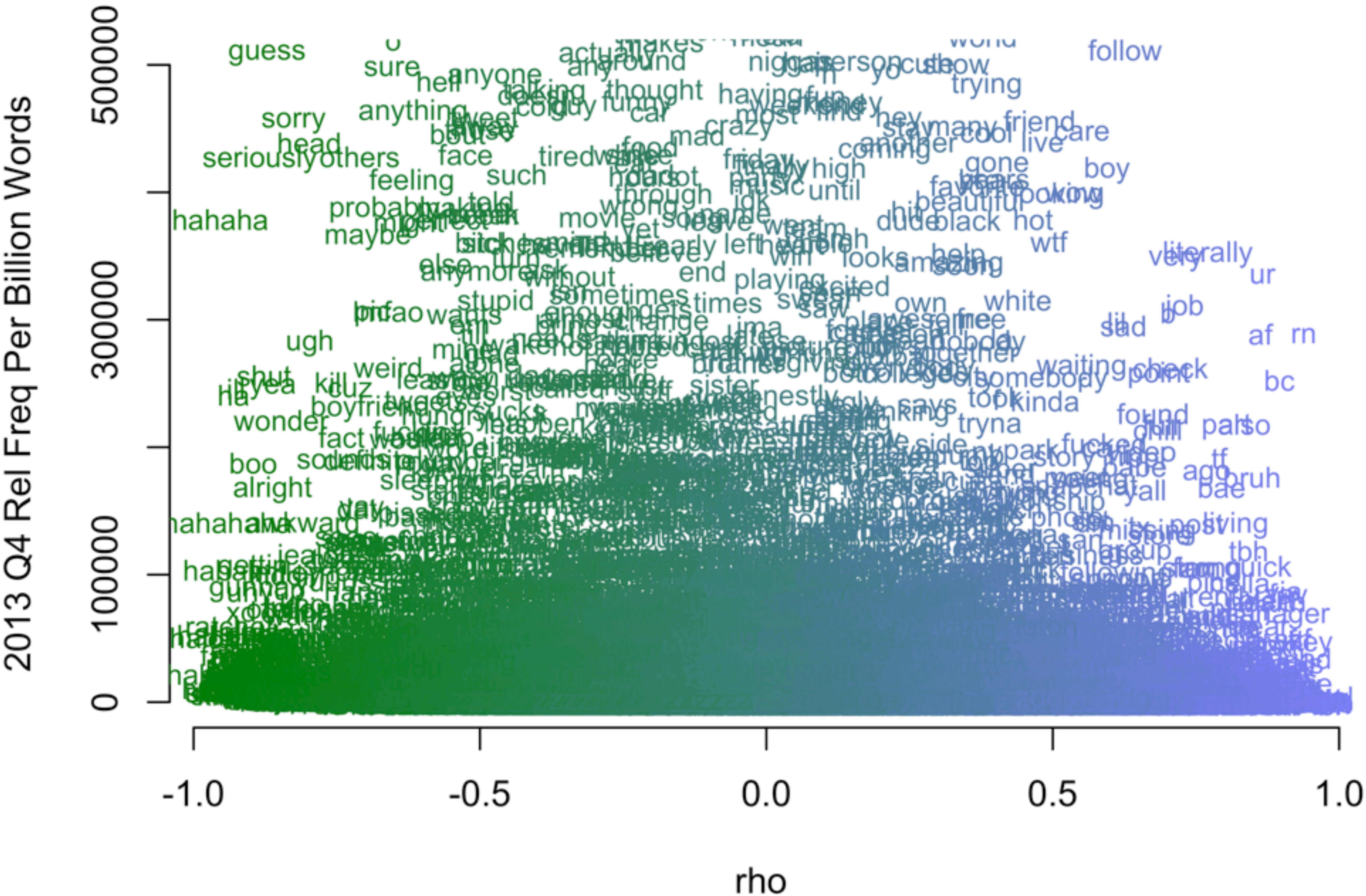
Emerging words were identified by cross-referencing the list of rising words ($\rho > .70$), against a list of rare words (relative frequency < 1 per million words in 4th quarter of 2013), against a list of non-dictionary words.

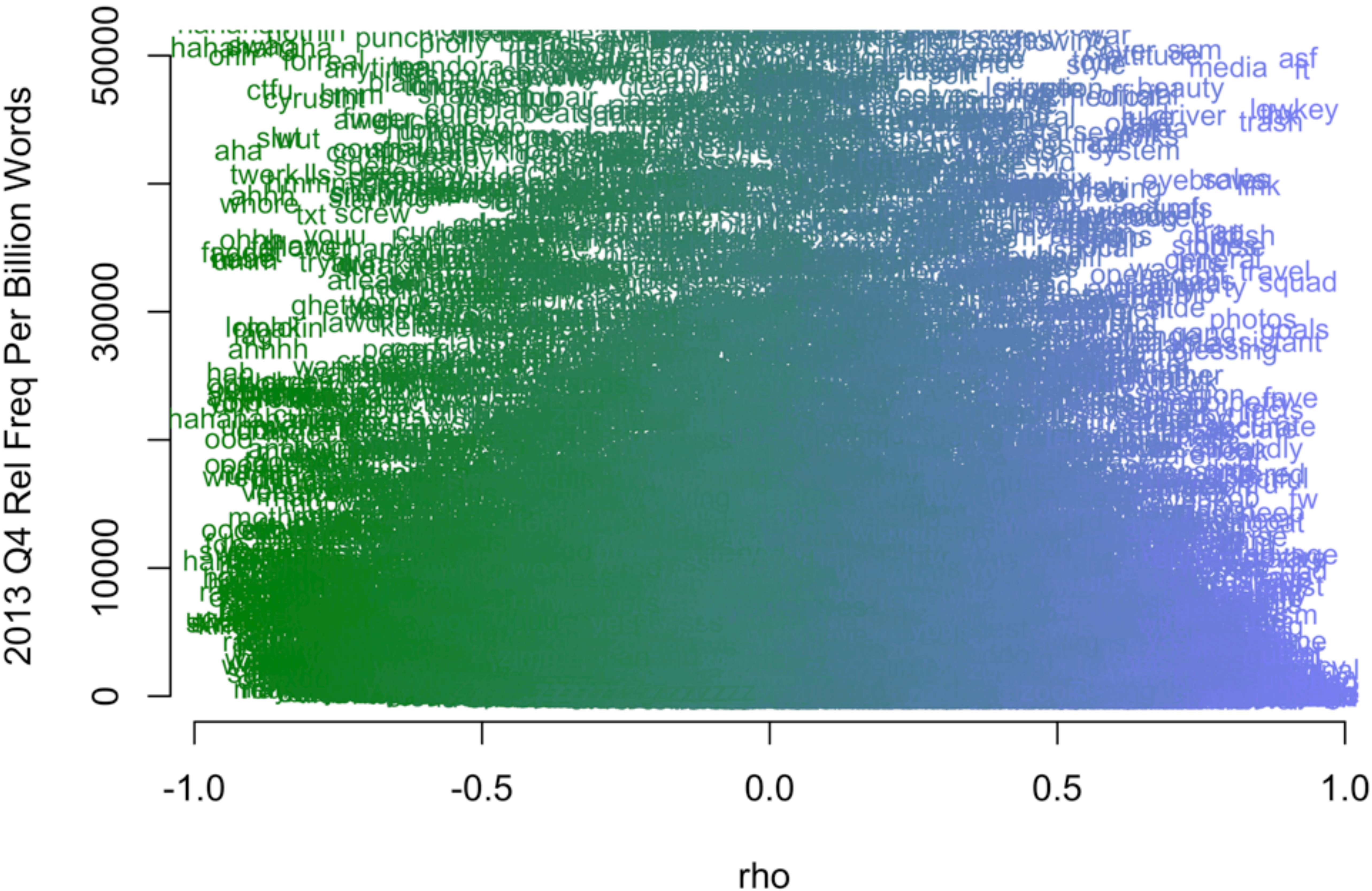
Proper nouns and medical terms associated with a rise of medical job adds over this period were also excluded.

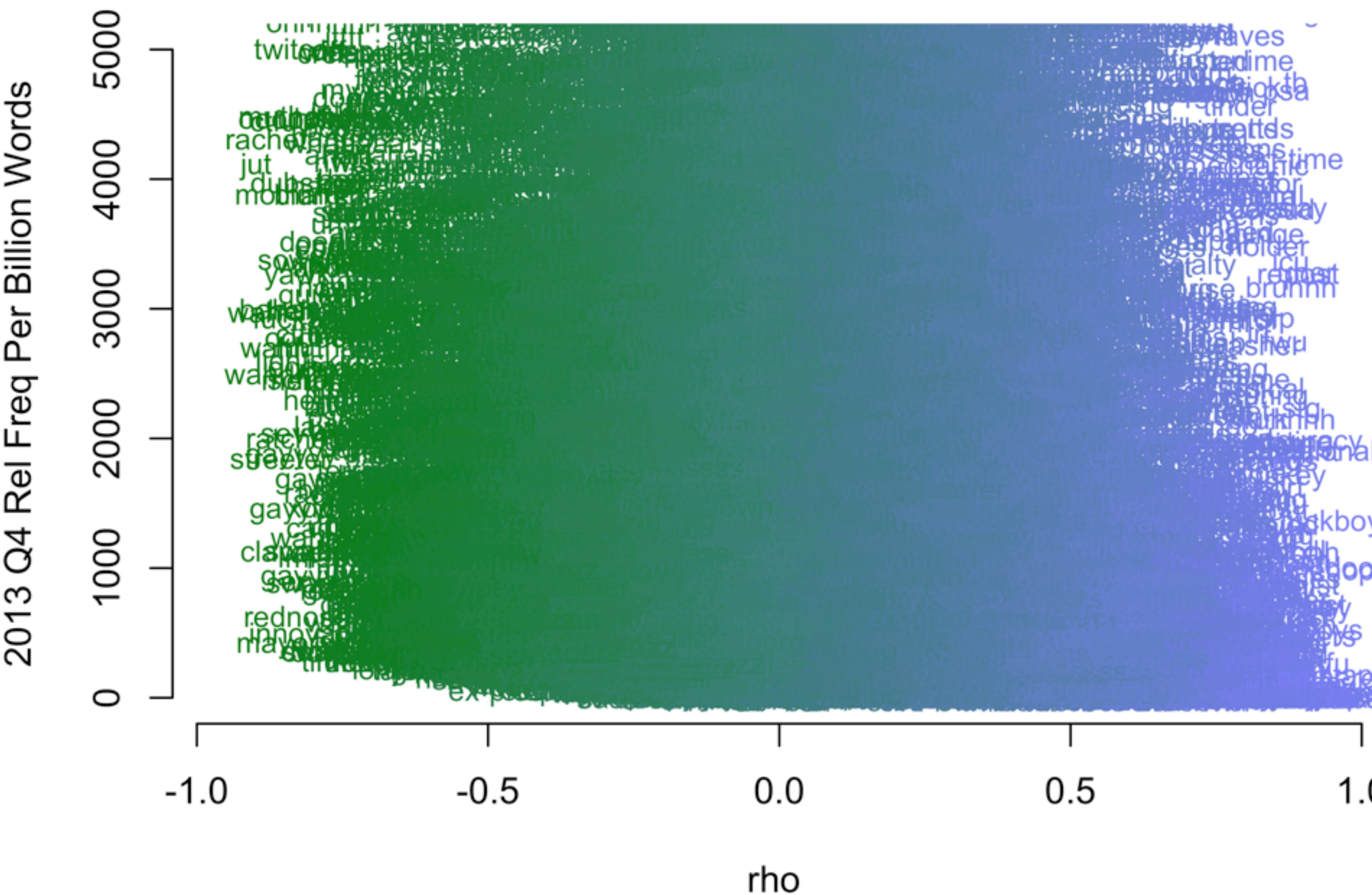


2013 Q4 Rel Freq Per Billion Words
0
10000000
30000000
50000000
rho
-1.0
-0.5
0.0
0.5
1.0
i
to
the
you
a
my
it
and
me
t
s
is
in
m
that
for
of
on
just
so
be
with
at
lol
what
but
love
when
no
u
from

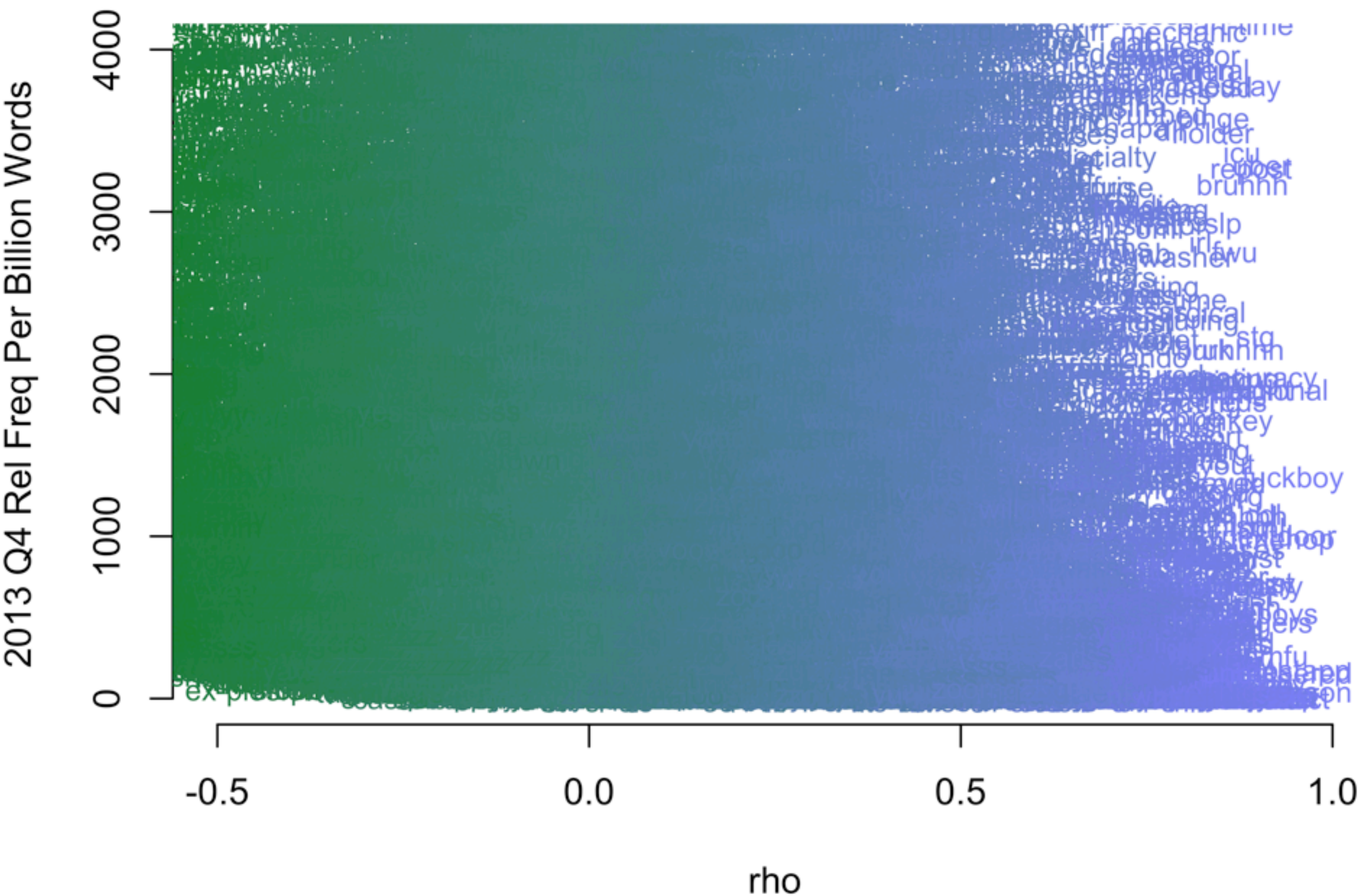

2013 Q4 Rel Freq Per Billion Words
0
10000000
30000000
50000000
rho
-1.0
-0.5
0.0
0.5
1.0
can
have
up
all
lol
get
but
what
was
your
love
are
we
do
if
no
go
out
when
now
know
good
about
got
time
they
he
want
how
day
too
why
re
she
right
u
from
will
think
as
been
happy
by
oh
well
haha
yeah
should


2013 Q4 Rel Freq Per Billion Words
0
1000000
3000000
5000000
-1.0
-0.5
0.0
0.5
1.0
rho

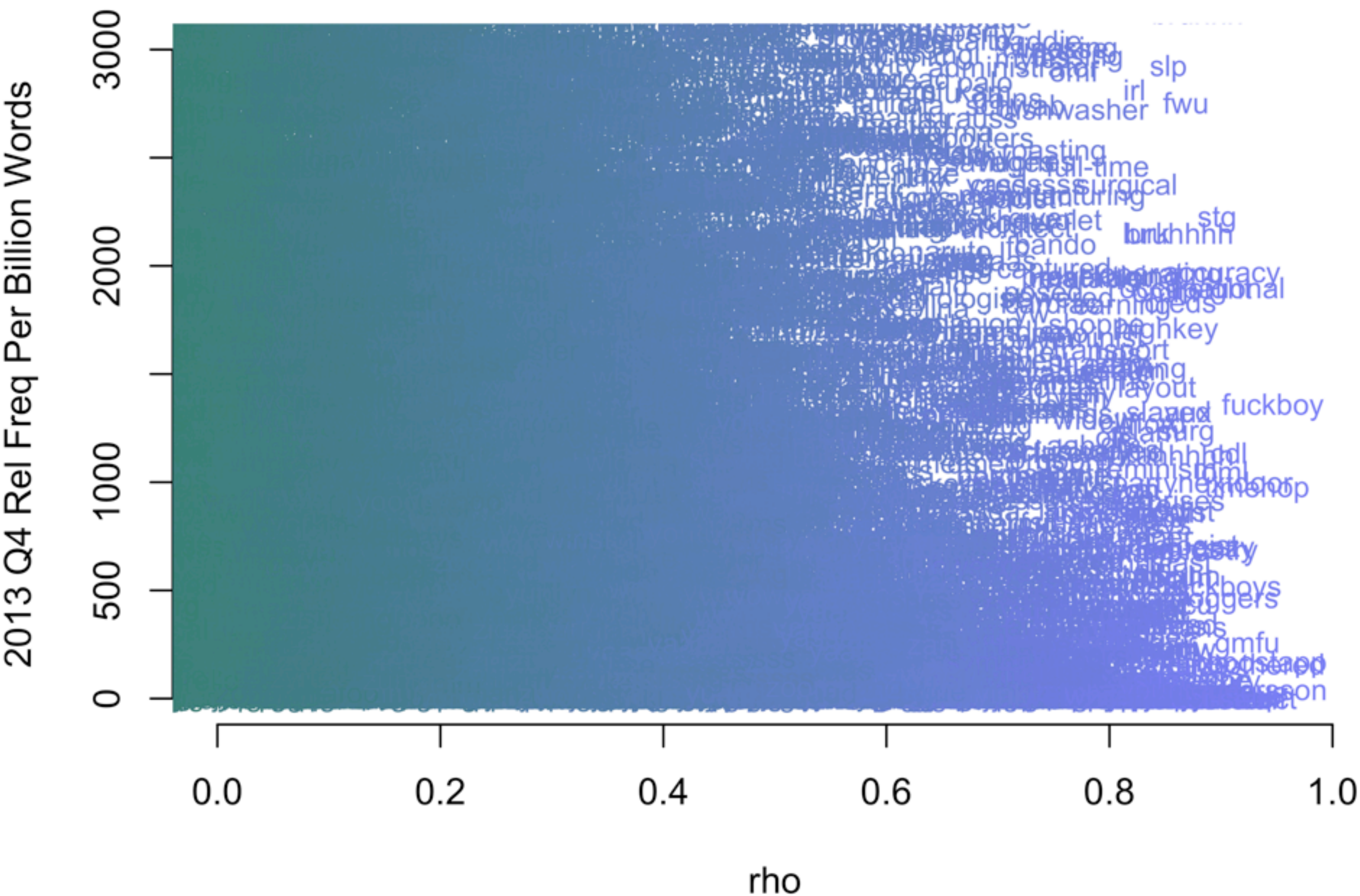

2013 Q4 Rel Freq Per Billion Words
0
10000
30000
50000
rho
-1.0
-0.5
0.0
0.5
1.0
media
asf
trash
system
squad
photos
goals
txt screw
aha
whore



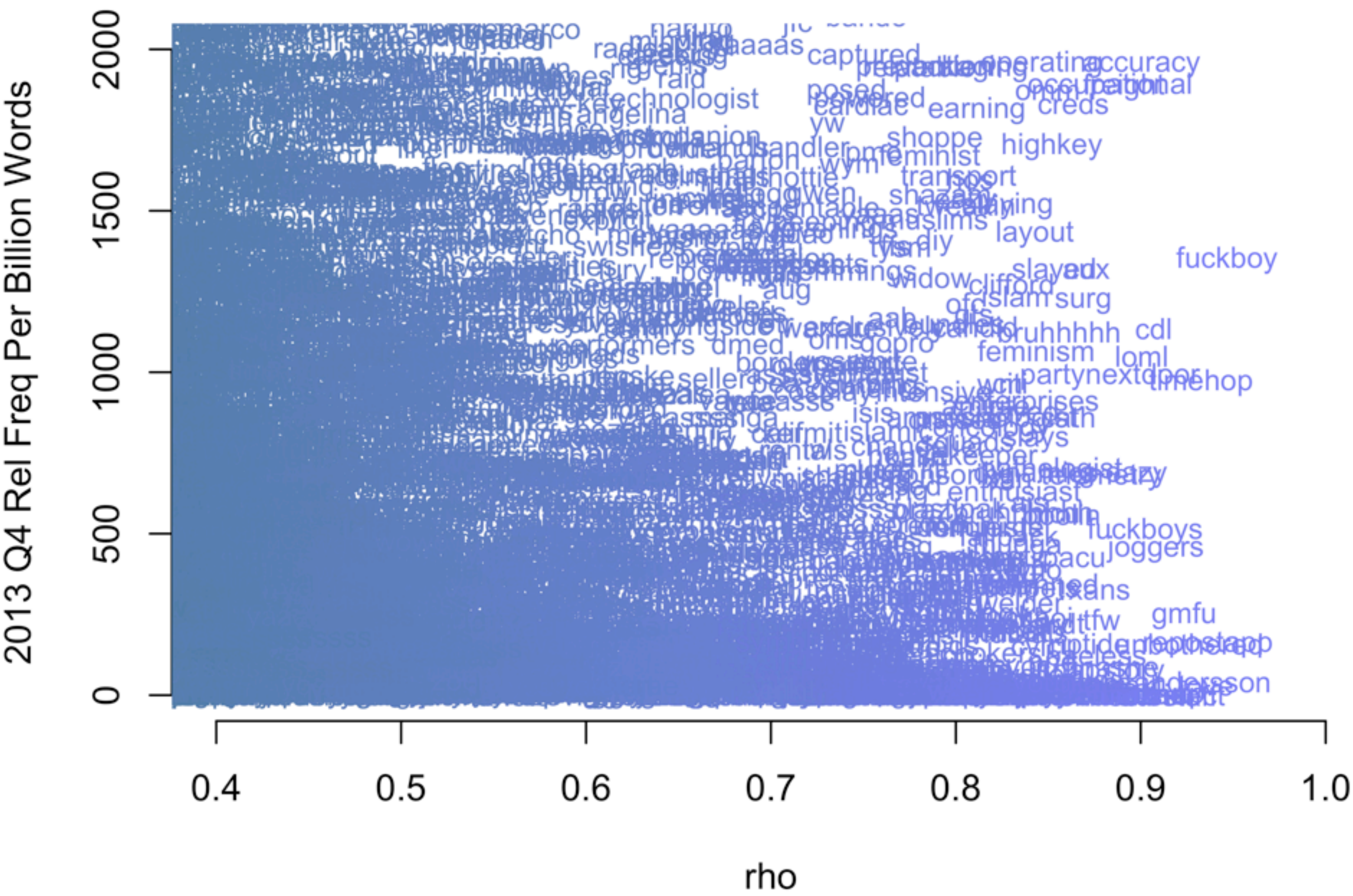
2013 Q4 Rel Freq Per Billion Words
0
1000
2000
3000
4000
5000
-1.0
-0.5
0.0
0.5
1.0
rho

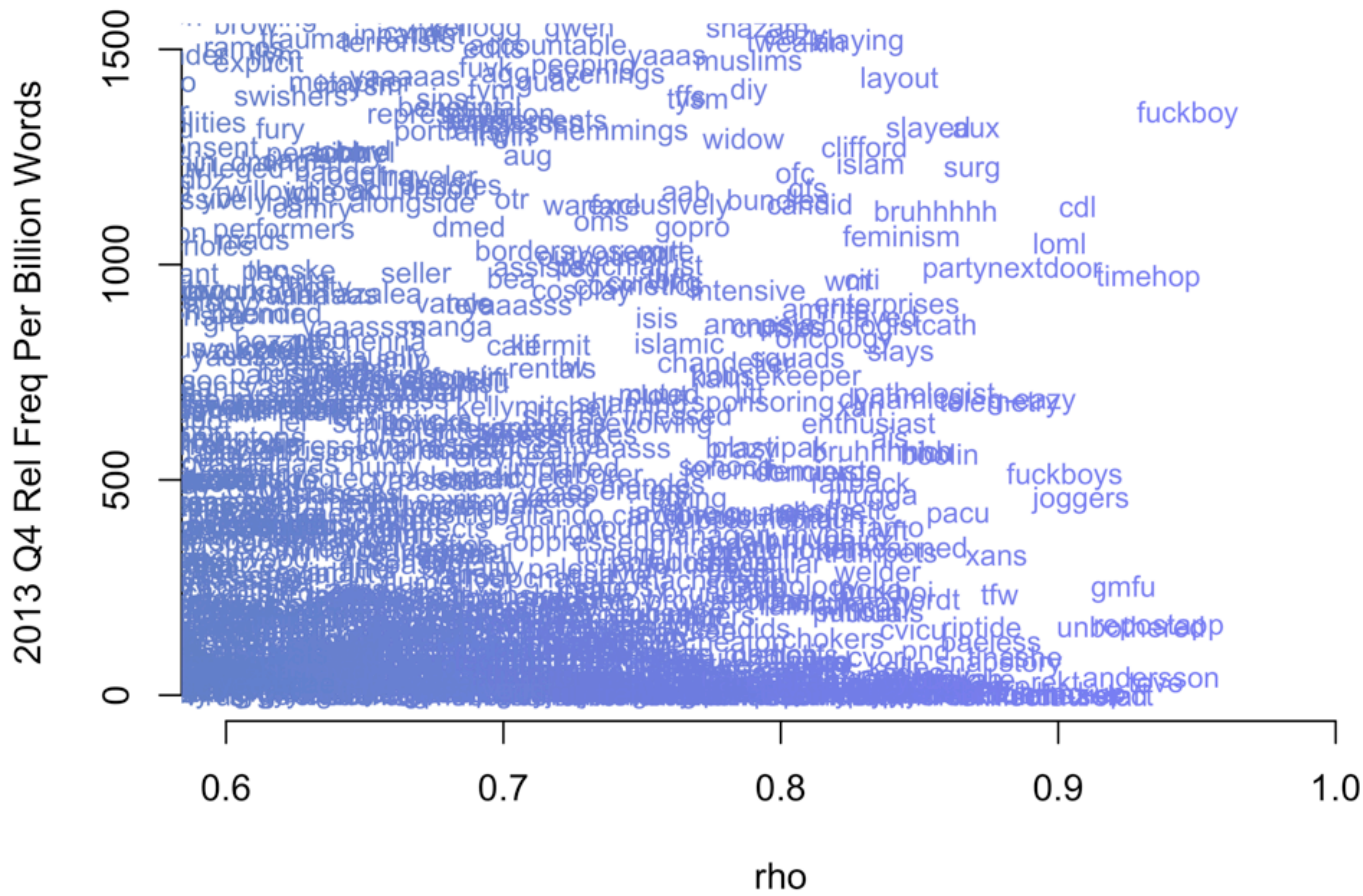


2013 Q4 Rel Freq Per Billion Words
4000
3000
2000
1000
0
-0.5
0.0
0.5
1.0
rho
mechanic
brunhhh



2013 Q4 Rel Freq Per Billion Words
3000
2000
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
rho
slp
irl
fwu
surgical
full-time
stg
bando
highkey
layout
fuckboy
gmfu


2013 Q4 Rel Freq Per Billion Words
0
500
1000
1500
2000
rho
0.4
0.5
0.6
0.7
0.8
0.9
1.0
accuracy
creds
earning
highkey
shoppe
layout
fuckboy
surg
cdl
feminism
loml
timehop
enthusiast
fuckboys
joggers
gmfu
technologist
angelina
performers
captured
posed
yw
transport
widow
isis
rentals
bruhhhhh
tfw



2013 Q4 Rel Freq Per Billion Words
0
500
1000
1500
rho
0.6
0.7
0.8
0.9
1.0
fuckboy
layout
diy
muslims
slayed
aux
widow
clifford
islam
surg
aug
fury
swishers
explicit
cdl
bruhhhhh
feminism
loml
partynextdoor
timehop
dmed
gopro
oms
seller
performers
alongside
otr
intensive
isis
islamic
kermit
enterprises
oncology
slays
squads
housekeeper
enthusiast
fuckboys
joggers
pacu
xans
welder
tfw
gmfu
riptide
chokers
andersson

Emerging Words

Through this process 81 emerging words were identified, including 27 spelling variants or inflected forms, leaving 54 unique emerging words for further analysis.

In general these words can be characterized as everyday slang; most are not restricted to Twitter or to CMC more generally (including acronyms).

Common topical domains include family and friends, sex and relationships, intoxication, technology, and Japanese culture.

Emerging Words on Twitter 2014

Words	**Definition**
amirite	Am I right?
baeless	Single
baeritto	Bae (i.e. significant other)
balayage	Hair style
boolin	Hanging out
brazy	Crazy
bruuh	Bro
candids	Candid public picture
celfie	Selfie
cosplay	Costume role playing
dwk	Driving While Kissing
fallback (game)	Skillful at talking one's way out of trouble
famo	Family and friend
faved	Favorited

Emerging Words on Twitter 2014

Words	Definition
fhritp	Fuck Her Right In The Pussy
figgity	Intoxicated; Very
(on) fleek	(On) point
fuckboys	Assholes
gainz	Earnings
gmfu	Get Me Fucked Up
goalz	Goals (i.e. life goals)
idgt	I Don’t Get Tired
lfie	Life
lifestyleeee	Lifestyle
litt	Lit (i.e. intoxicated, good)
litty	Lit (i.e. intoxicated, good)
lituation	A lit situation
lordt	Lord (esp. as exclamation)

Emerging Words on Twitter 2014

Words	**Definition**
lw	Light Weight
mce	Man Crush Everyday
mmmmmmmuah	Laughter
mutuals	Mutual friends
nahfr	Nah For Real
notifs	Notifications (esp. online)
pcd	Post Concert Depression
pullout (game)	Skillful at *coitus interruptus*
rekt	Wrecked (i.e. intoxicated; defeated esp. in
rq	Real Quick
scute	Cute
senpai	Father figure
shordy	Shorty (i.e. a young woman)
slayin	Slaying

Emerging Words on Twitter 2014

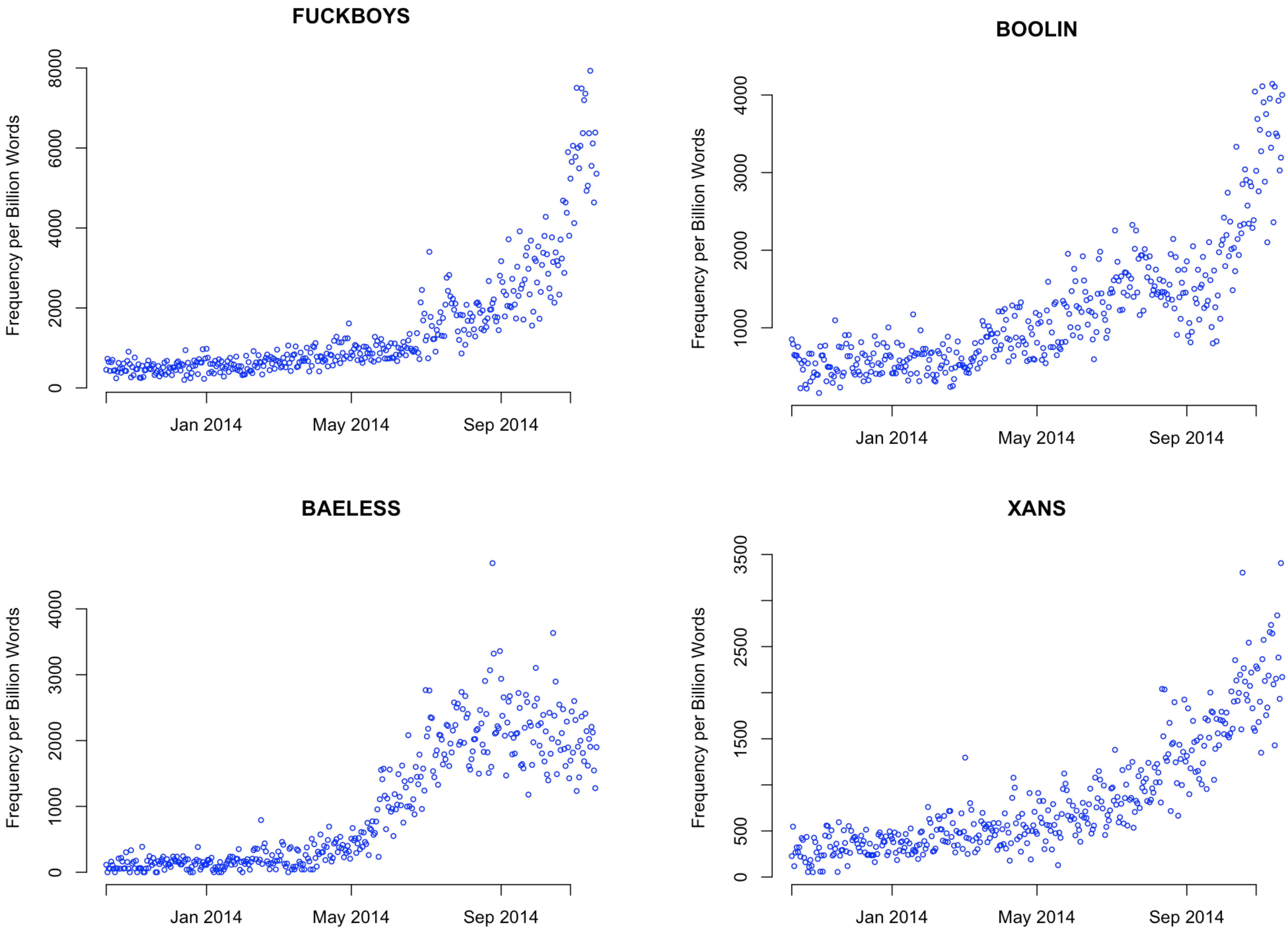
Words	**Definition**
sqaud	Squad (i.e. a crew)
tbfh	To Be Fucking Honest
tfw	That Feel When
thotful	Thoughtful
thottin	Looking for thots (i.e. promiscuous
tookah	Marijuana
traphouse	Drug house
unbae	Break up with
waifu	Wife
wce	Woman Crush Everyday
xans	Benzodiazapane pills
yaas	Yes

General Findings

By providing a relatively large and unbiased set of emerging words, generalizations can be made about the process of lexical emergence:

1. Most of the emerging words were formed using standard word formation processes, with truncation, compounding and acronymazation being most common.

2. The relative frequencies of most newly emerging words are characterized by s-shaped curve time series (or partial s-shaped curves).

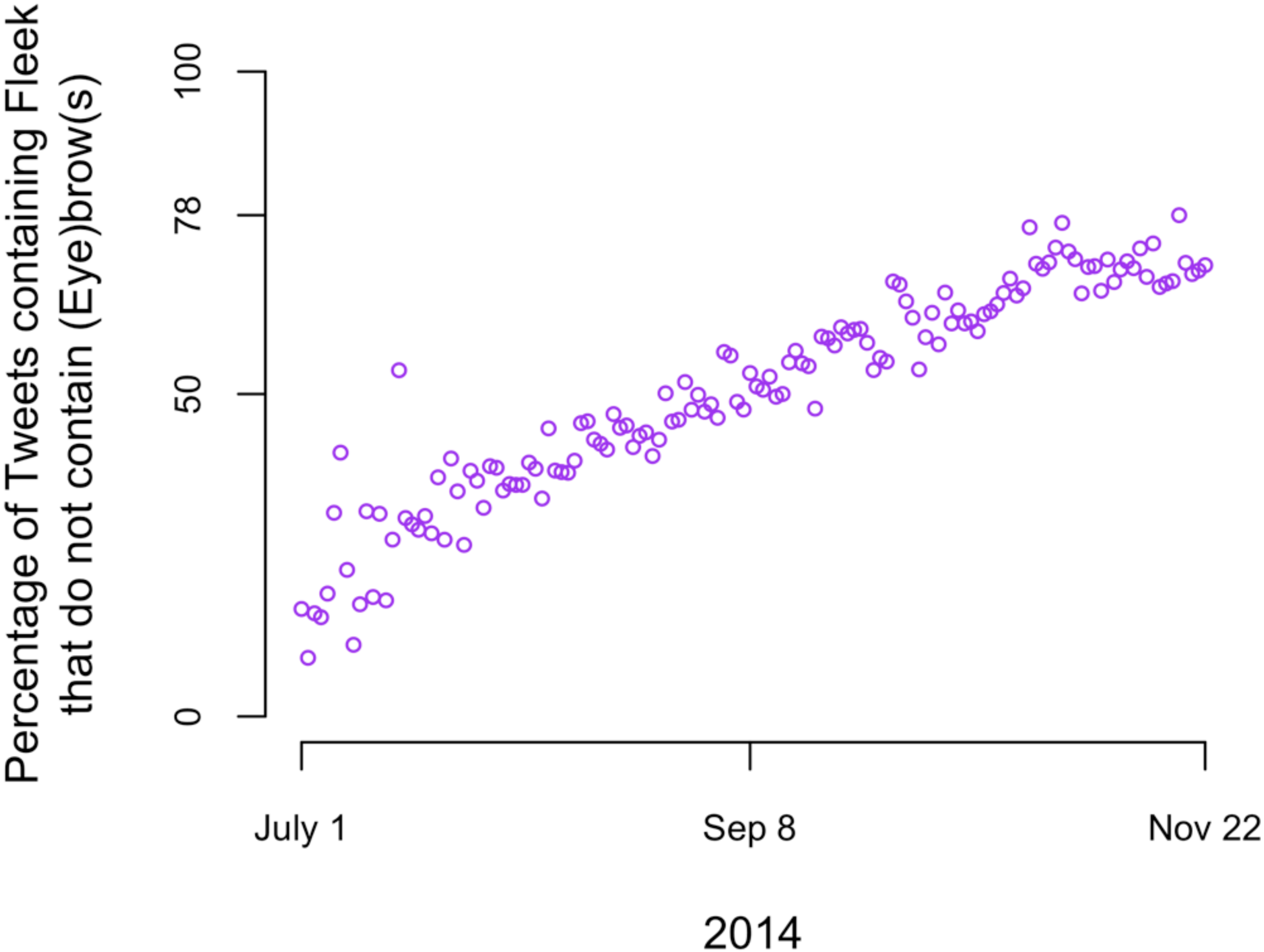

FUCKBOYS
Frequency per Billion Words
0
2000
4000
6000
8000
Jan 2014
May 2014
Sep 2014
BOOLIN
Frequency per Billion Words
1000
2000
3000
4000
Jan 2014
May 2014
Sep 2014
BAELESS
Frequency per Billion Words
0
1000
2000
3000
4000
Jan 2014
May 2014
Sep 2014
XANS
Frequency per Billion Words
0
500
1500
2500
3500
Jan 2014
May 2014
Sep 2014

General Findings

3. Numerous words emerge into the language each year; some persist, whereas others quickly die.

4. Most of the newly emerging words are not really new; rather, they have been laying dormant since the 2000s.

5. The meanings of a number of the newly emerging words appear to generalize over time.



Percentage of Tweets containing Fleek
that do not contain (Eye)brow(s)
100
78
50
0
July 1
Sep 8
Nov 22
2014

Publication and Future Research

For more information on these results please see

Grieve, Nini & Guo. 2016. Analyzing lexical emergence in Modern American English online. *Forthcoming in English Language and Linguistics*. (Open access pre-print available on publisher's website).

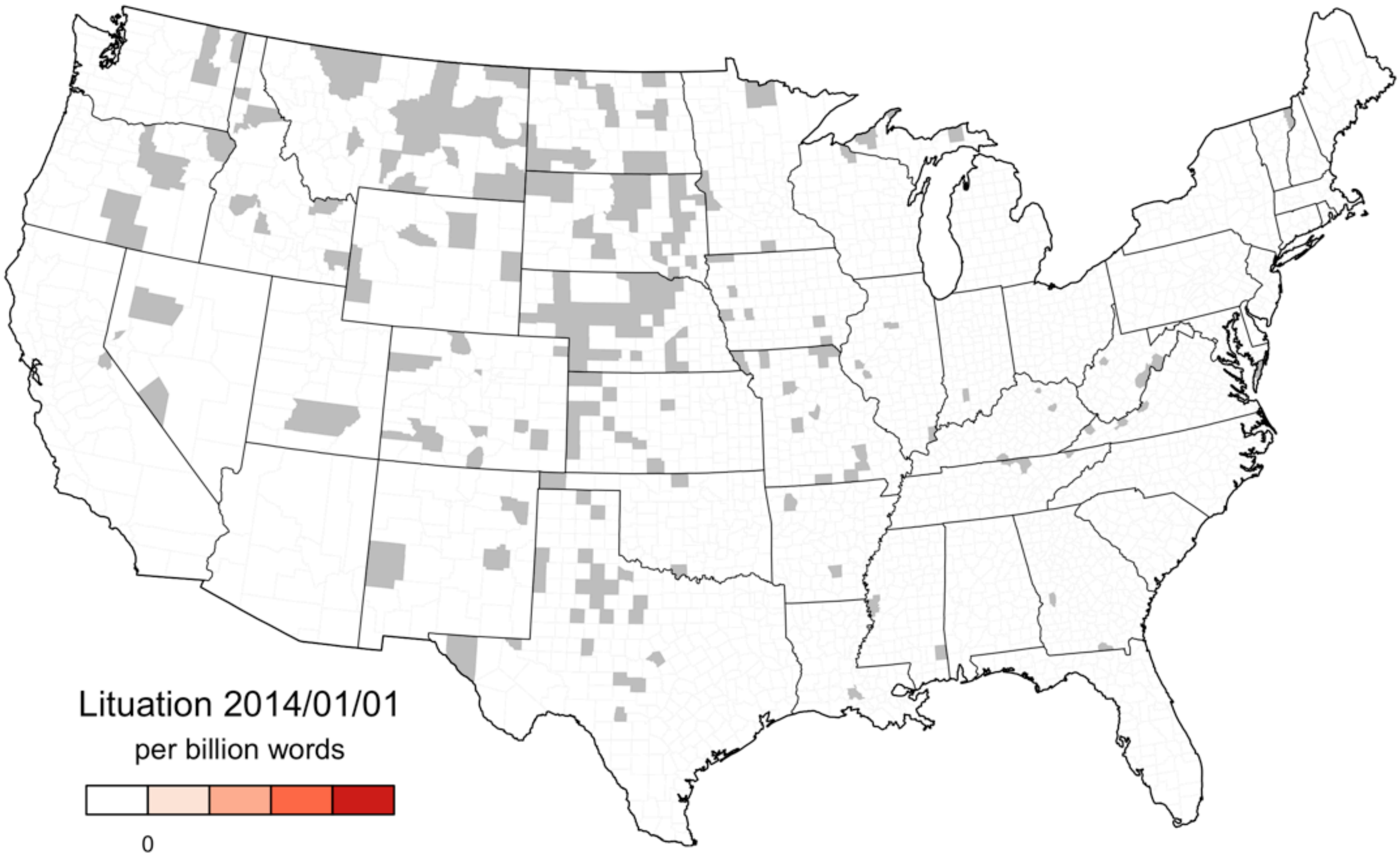
Future research: What triggers the emergence of a new word? What predicts the success or failure of new word? How can you model the rise in the frequency of a new word over time?

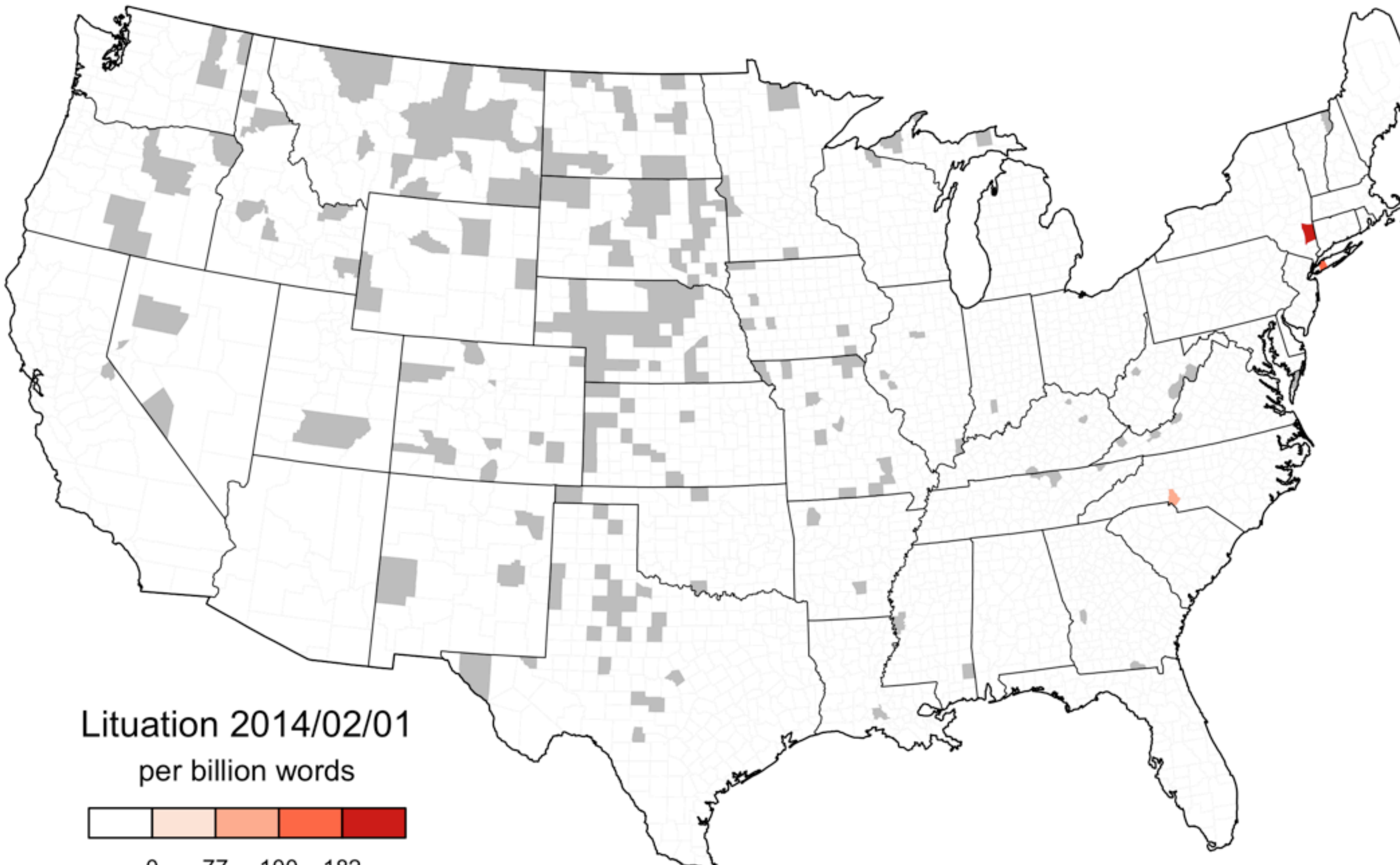
Mapping Lexical Emergence

In addition to identifying and charting the spread of new words over time, because the Twitter corpus contains precise geocoding information, it is also possible to analyze lexical emergence from a regional perspective, including identifying regional hubs of lexical innovation.

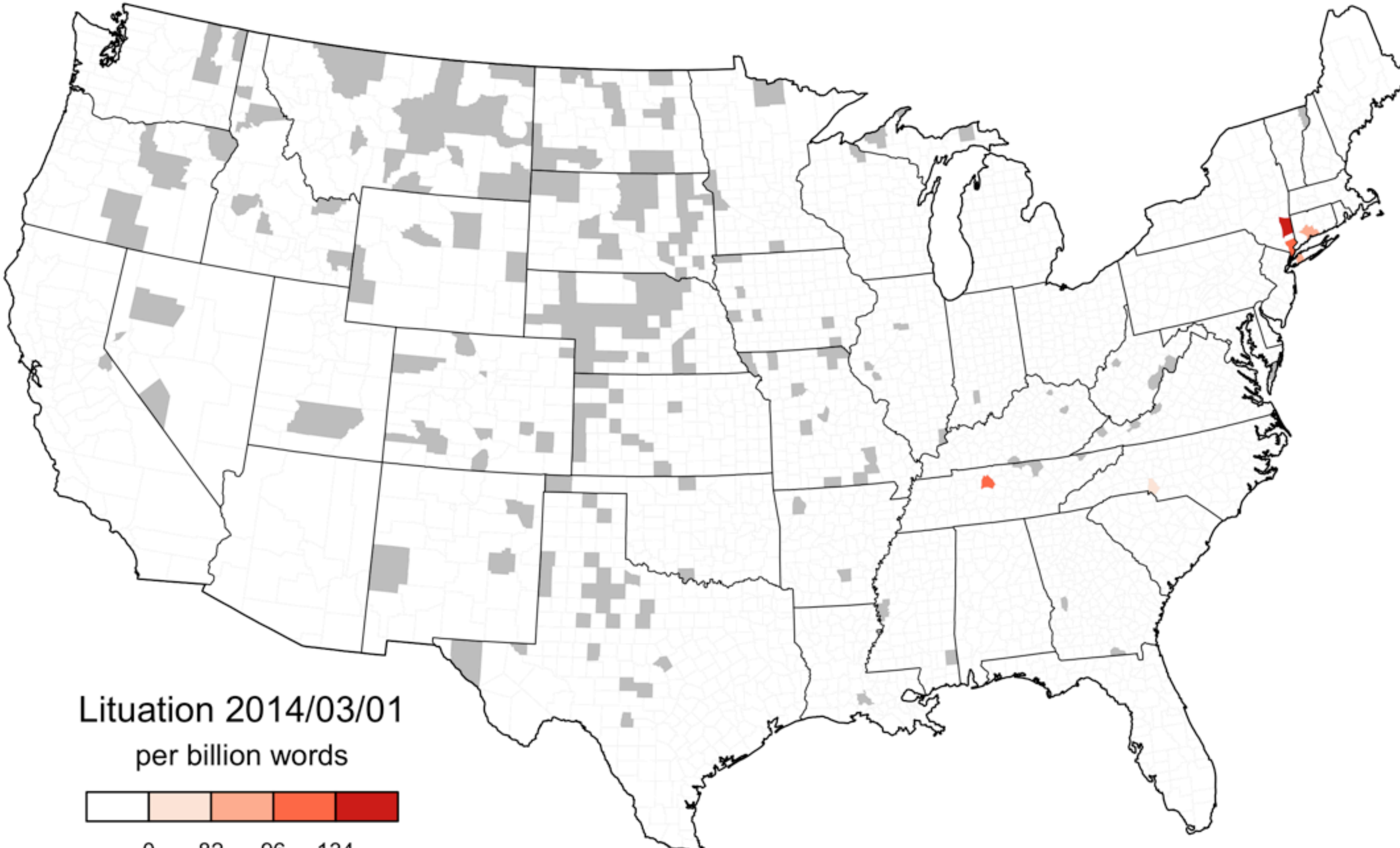
An initial problem, however, is how to map the spread of an emerging word.

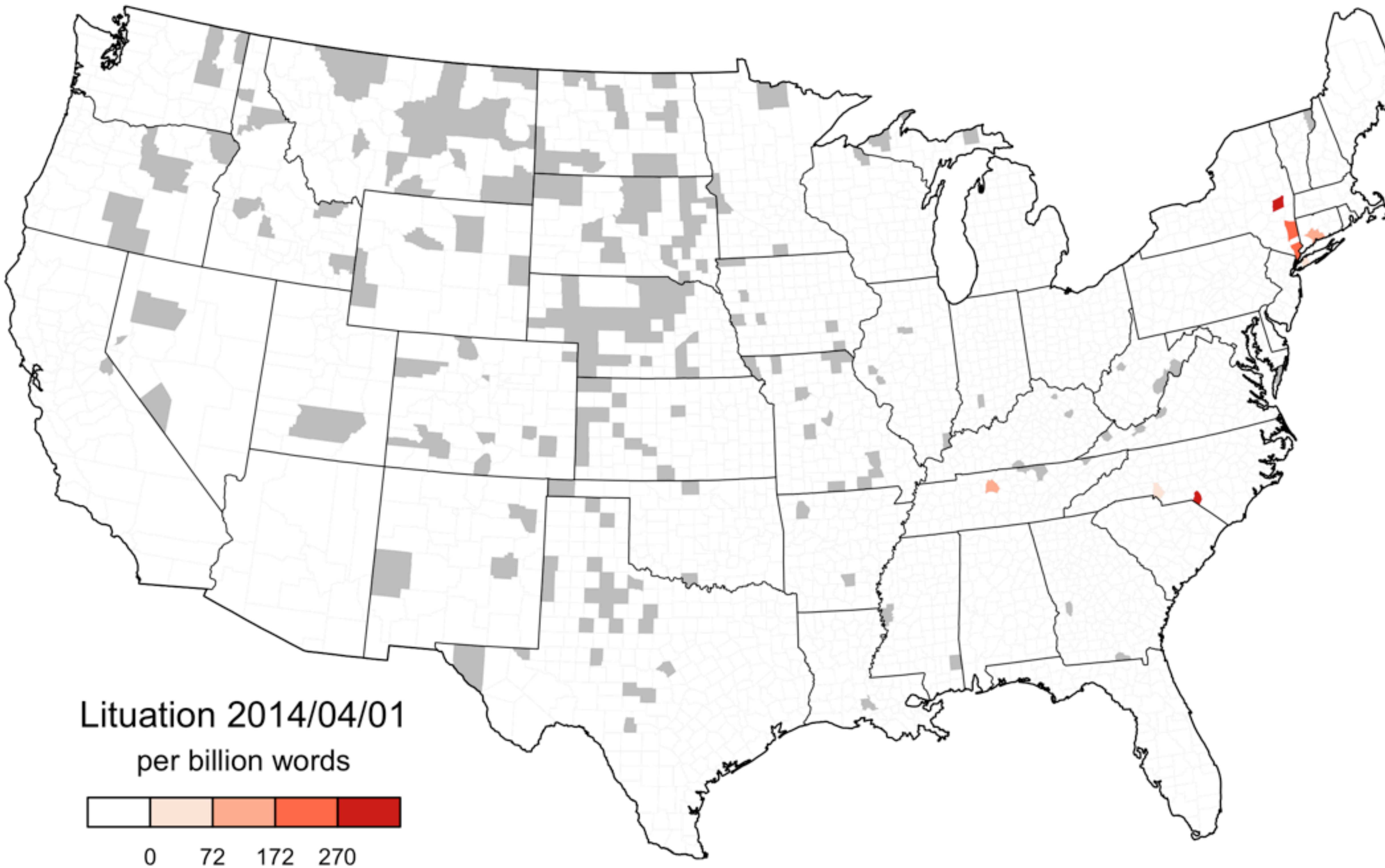
A basic approach is to map the cumulative relative frequency of the word across regionally defined sub-corpora (e.g. counties) over a series of points in time.

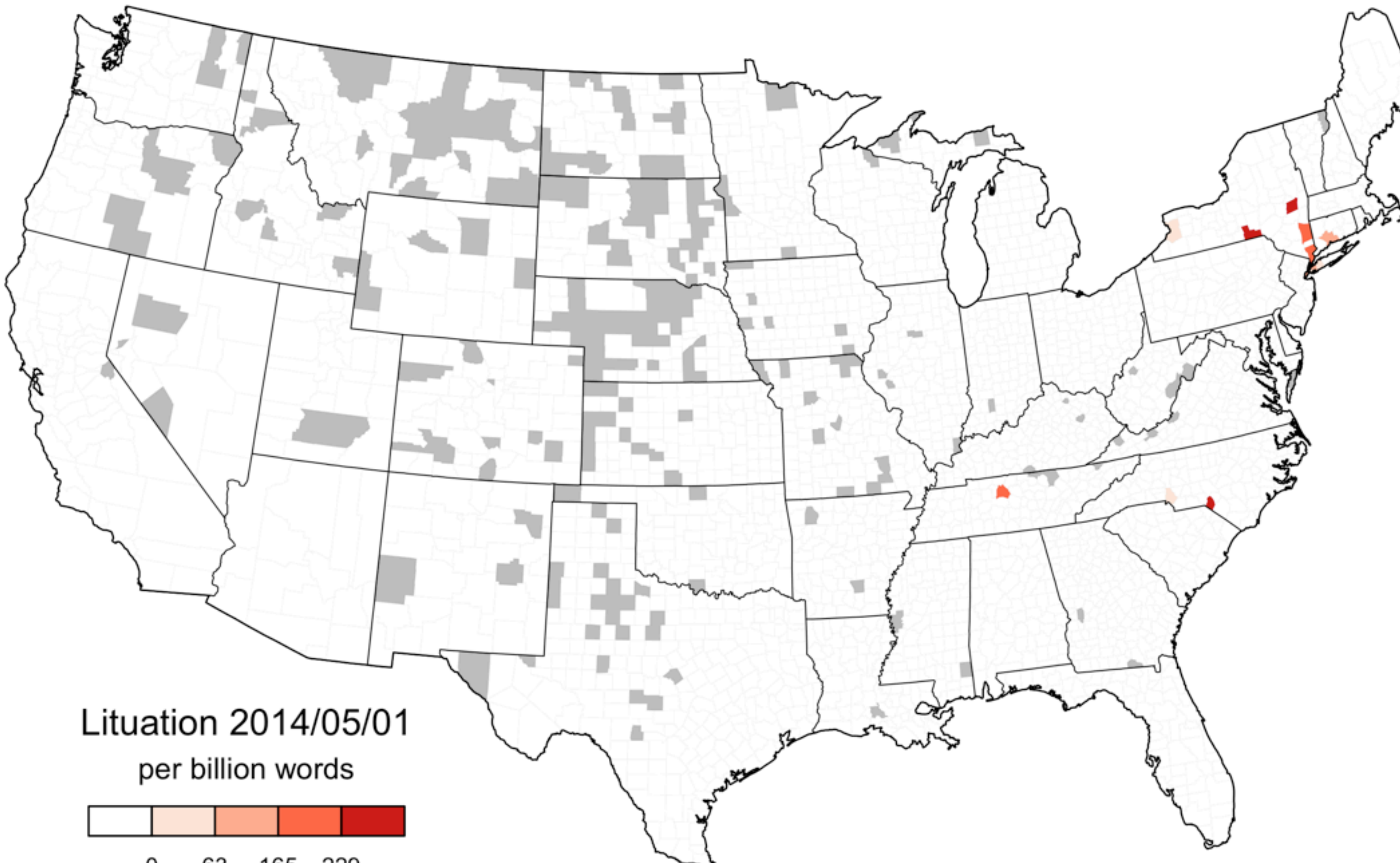

Lituation 2014/01/01
per billion words
0



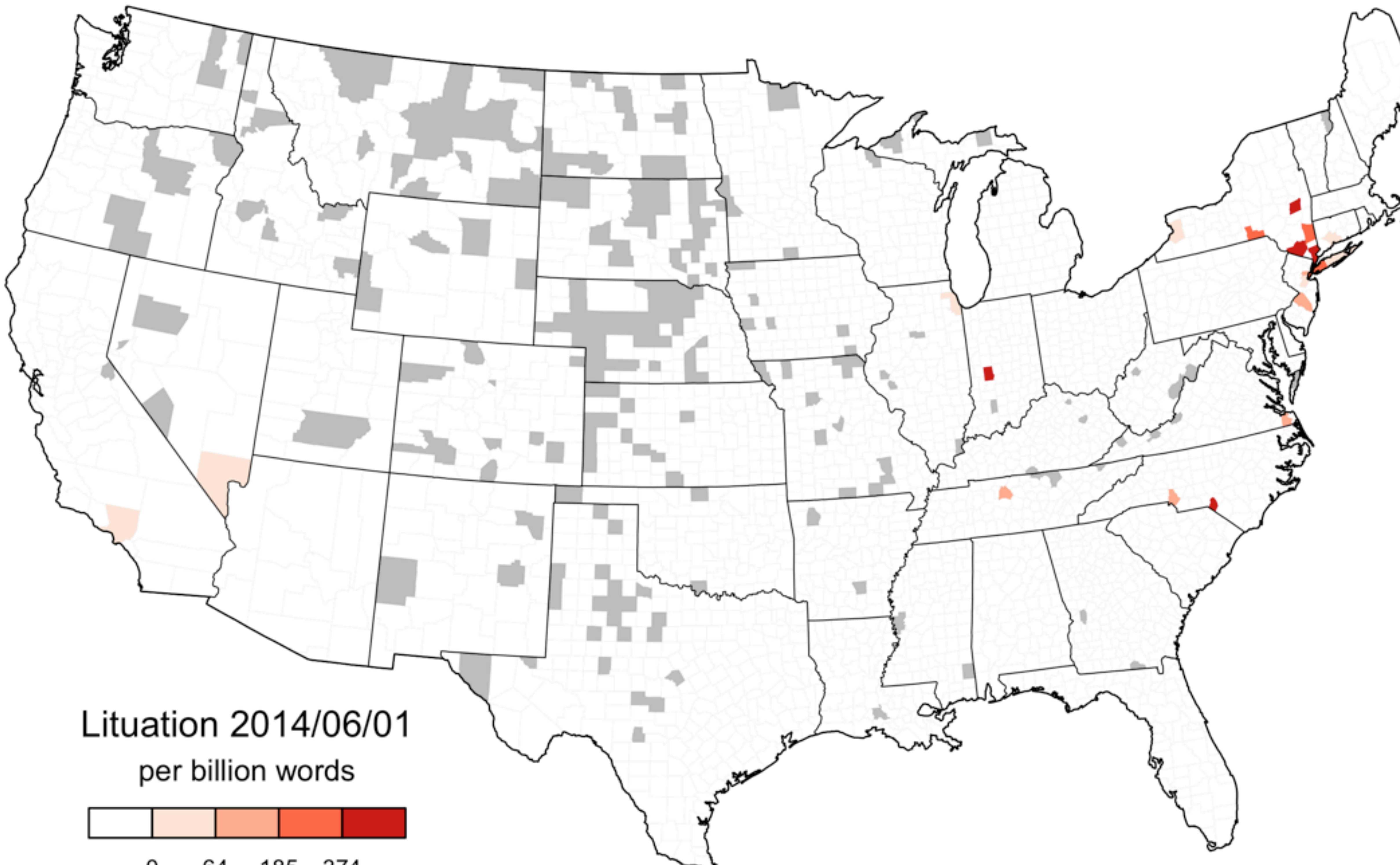
Lituation 2014/02/01
per billion words
0
77
100
182

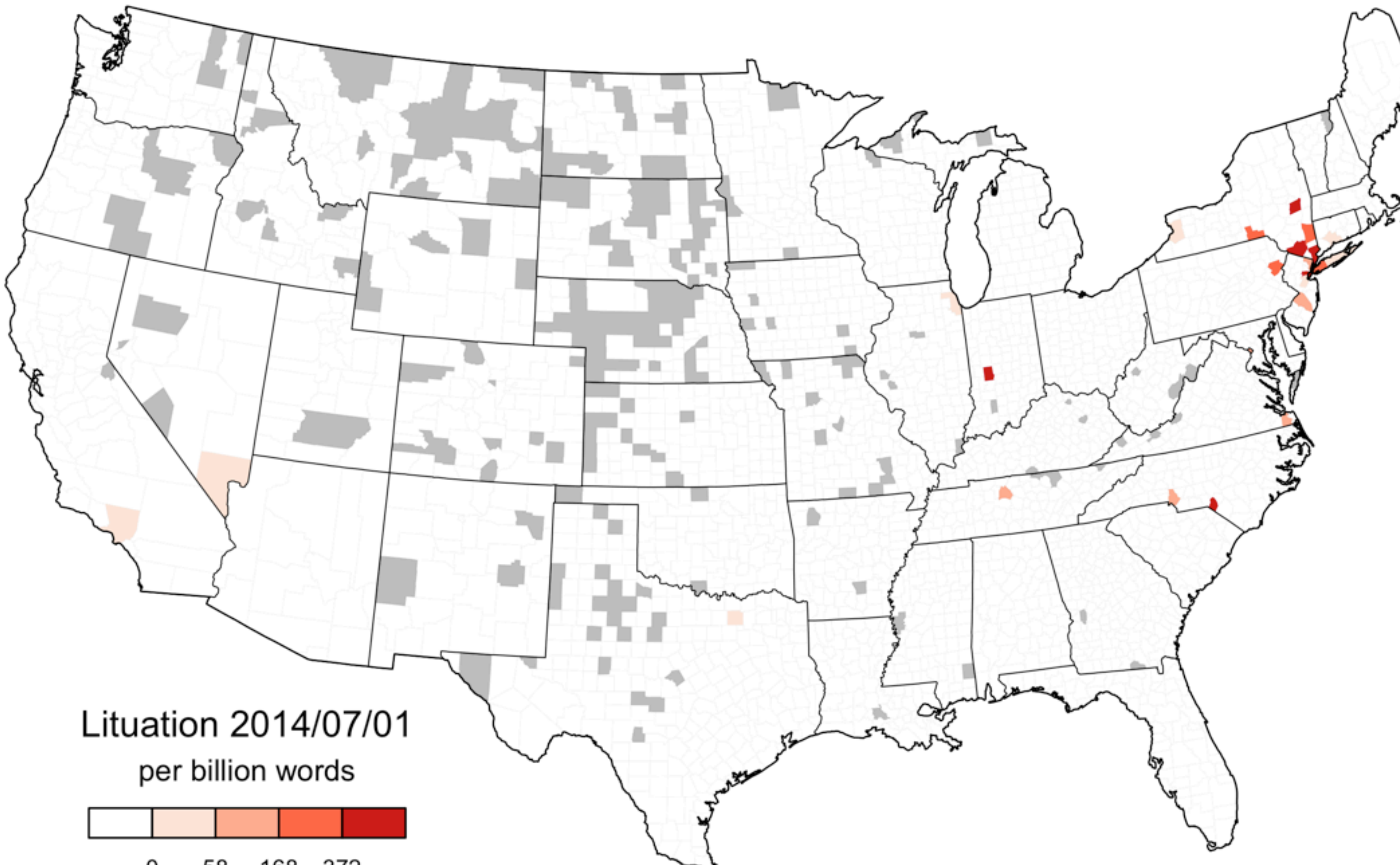

Lituation 2014/03/01
per billion words
0
82
96
134


Lituation 2014/04/01
per billion words
0
72
172
270

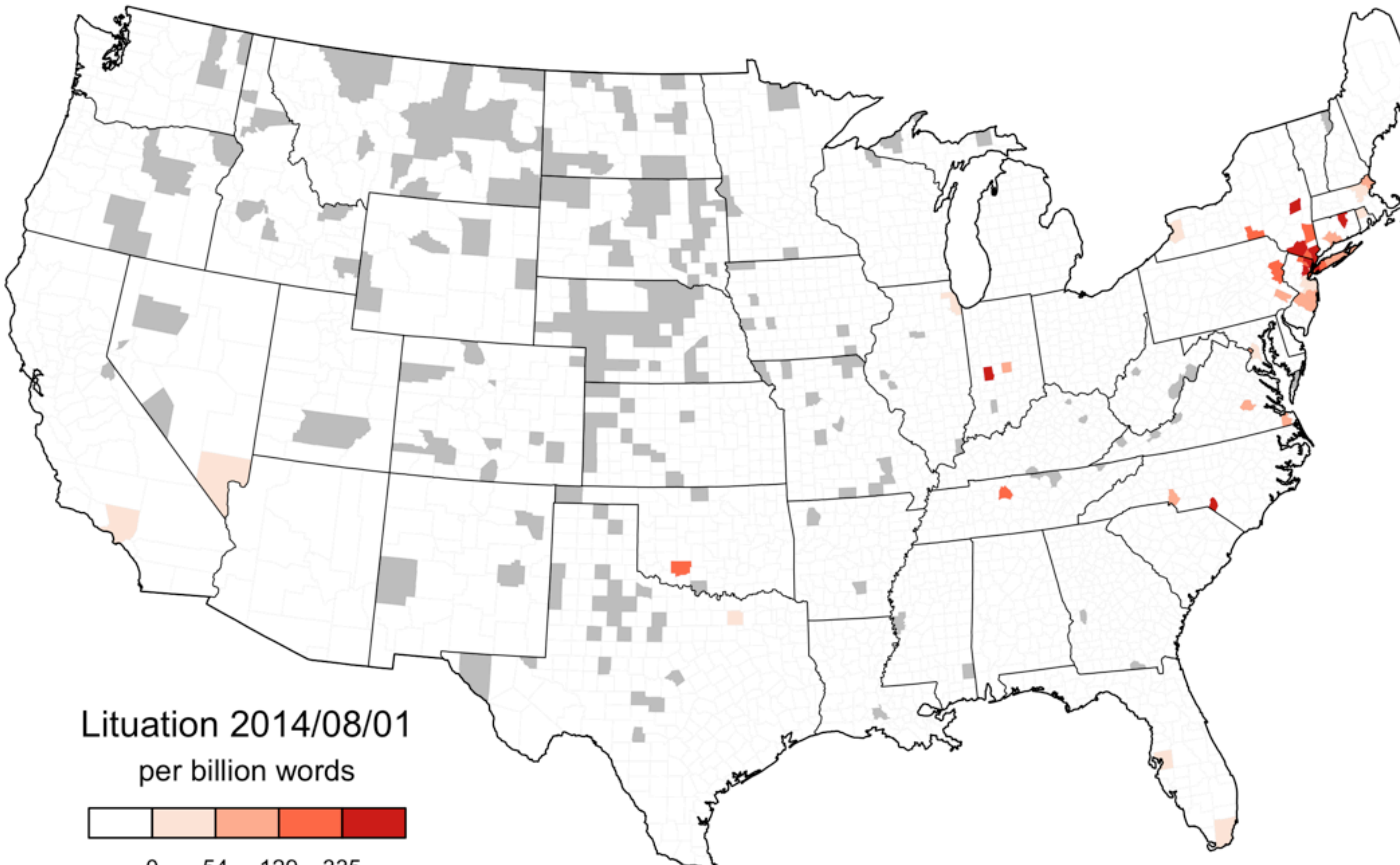


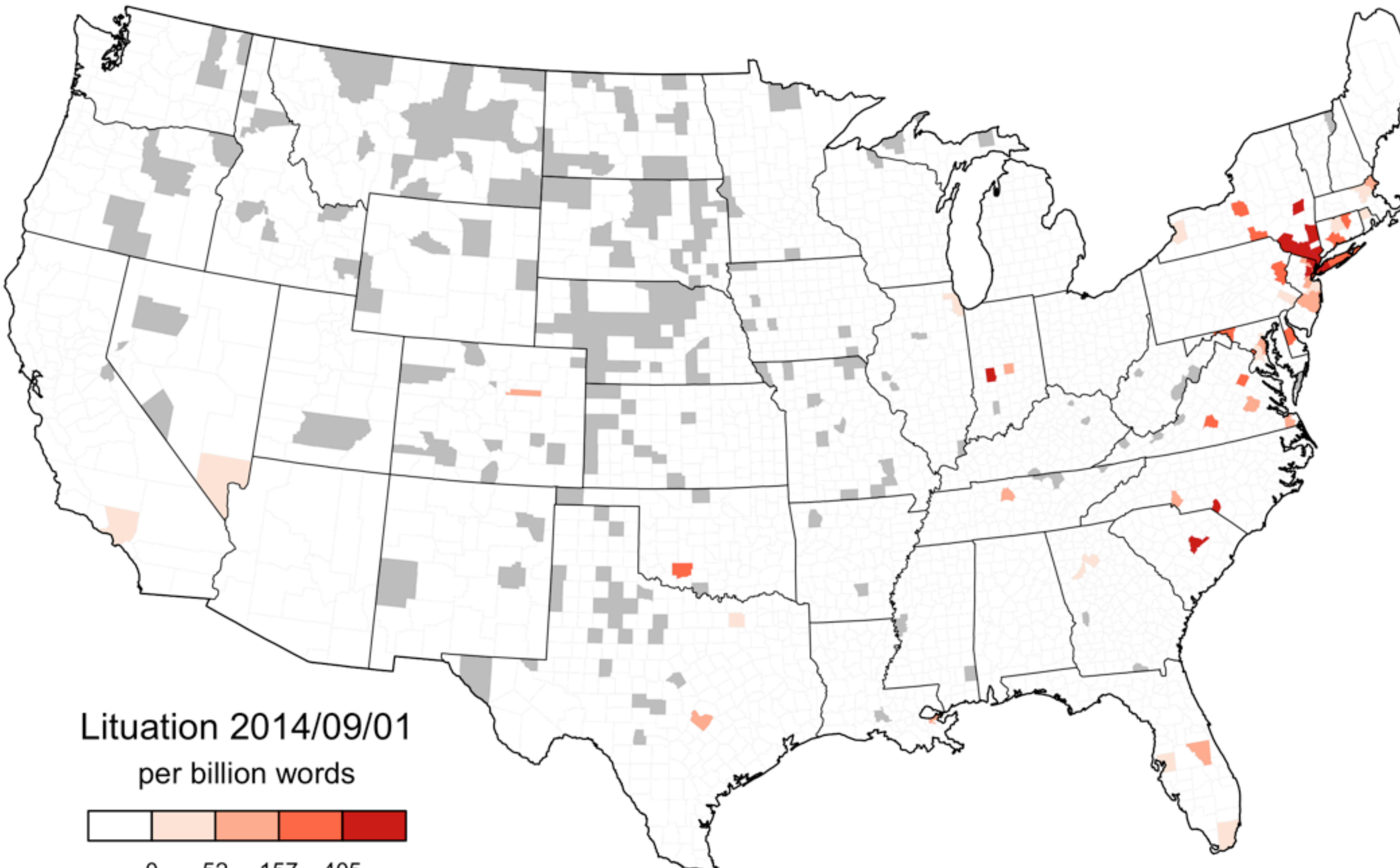
Lituation 2014/05/01
per billion words
0
63
165
229


Lituation 2014/06/01
per billion words
0
64
185
374

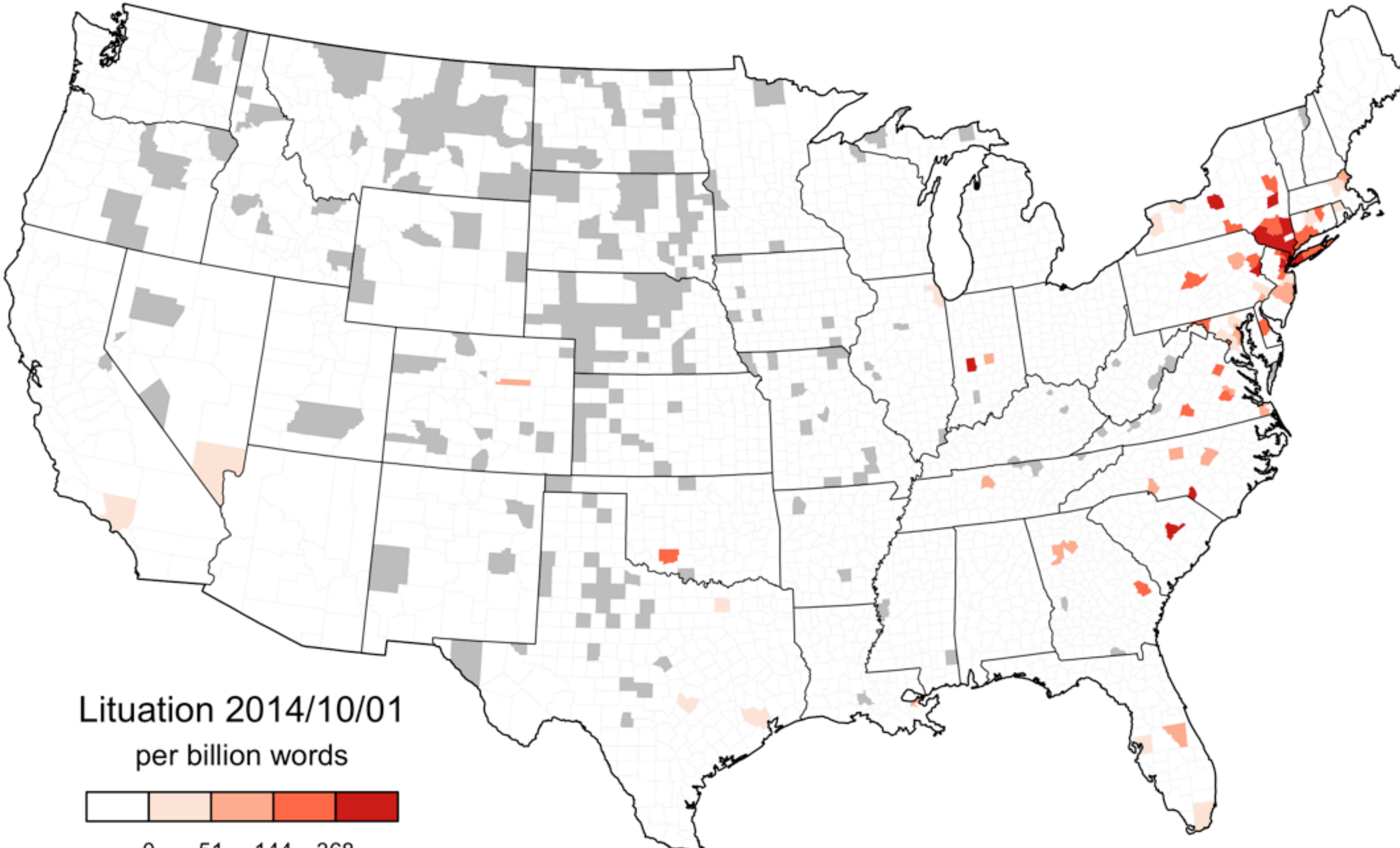


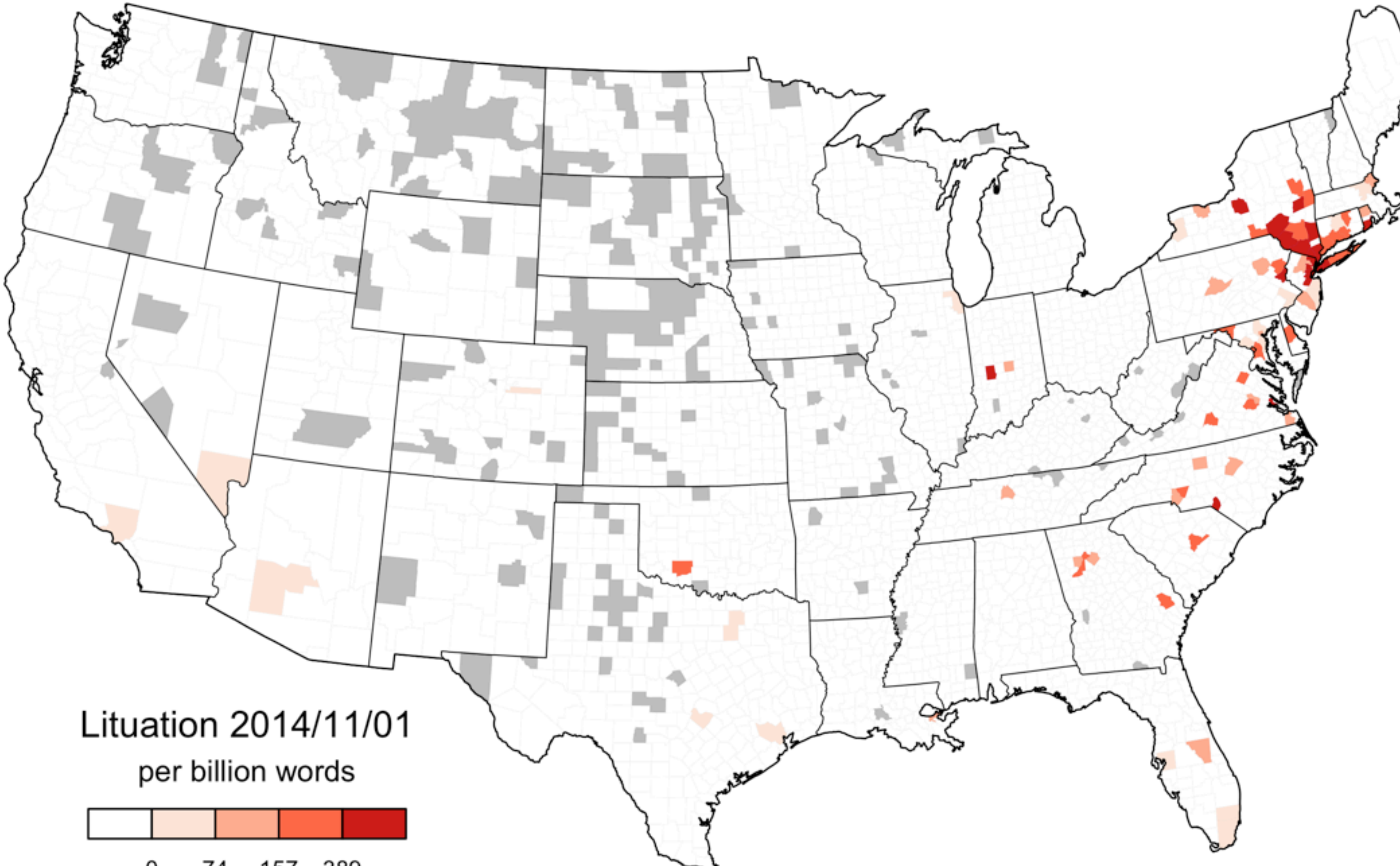
Lituation 2014/07/01
per billion words
0
58
168
372


Lituation 2014/08/01
per billion words
0
54
129
335



Lituation 2014/09/01
per billion words
0
52
157
405


Lituation 2014/10/01
per billion words
0
51
144
368

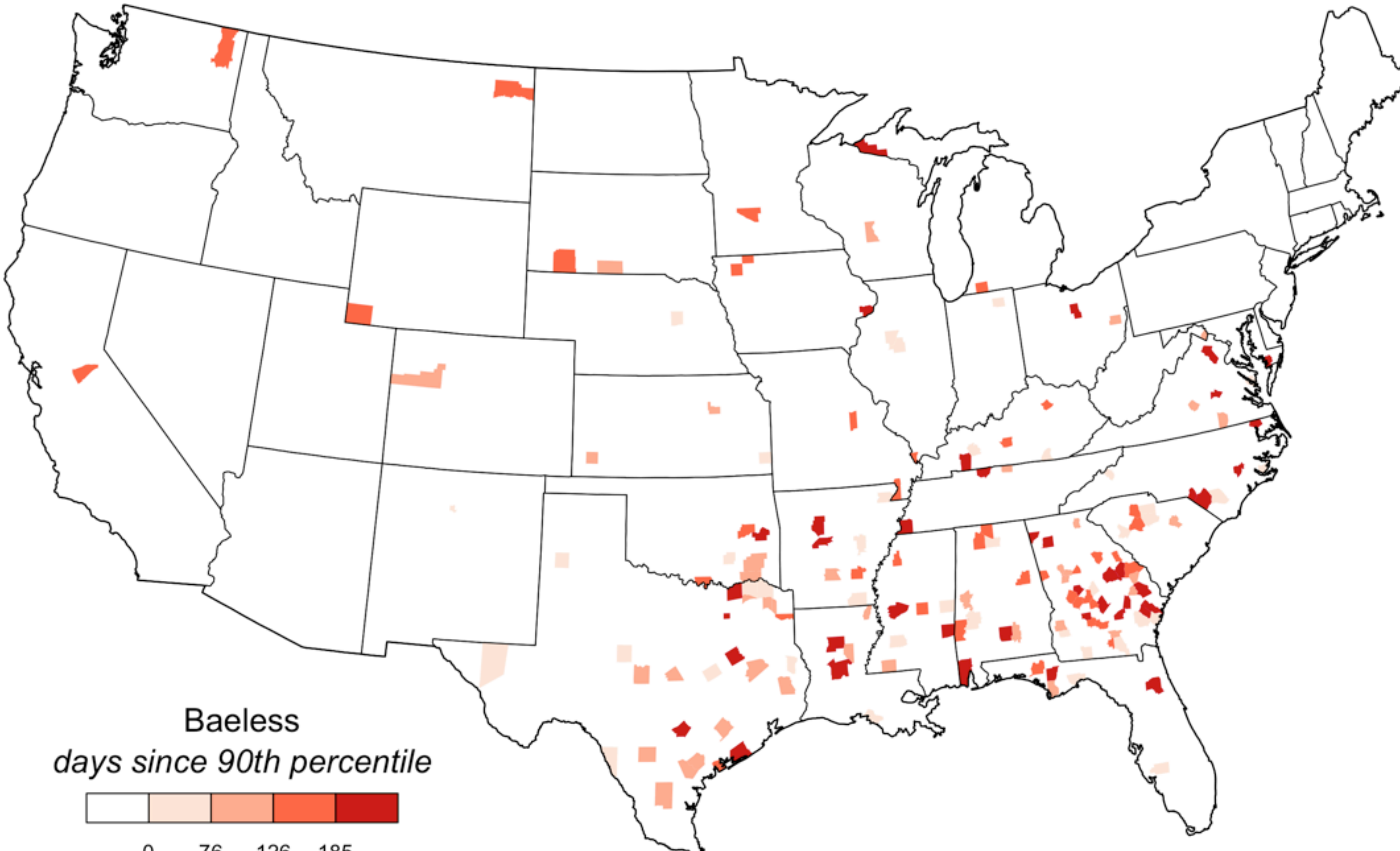

Lituation 2014/11/01
per billion words
0
74
157
389

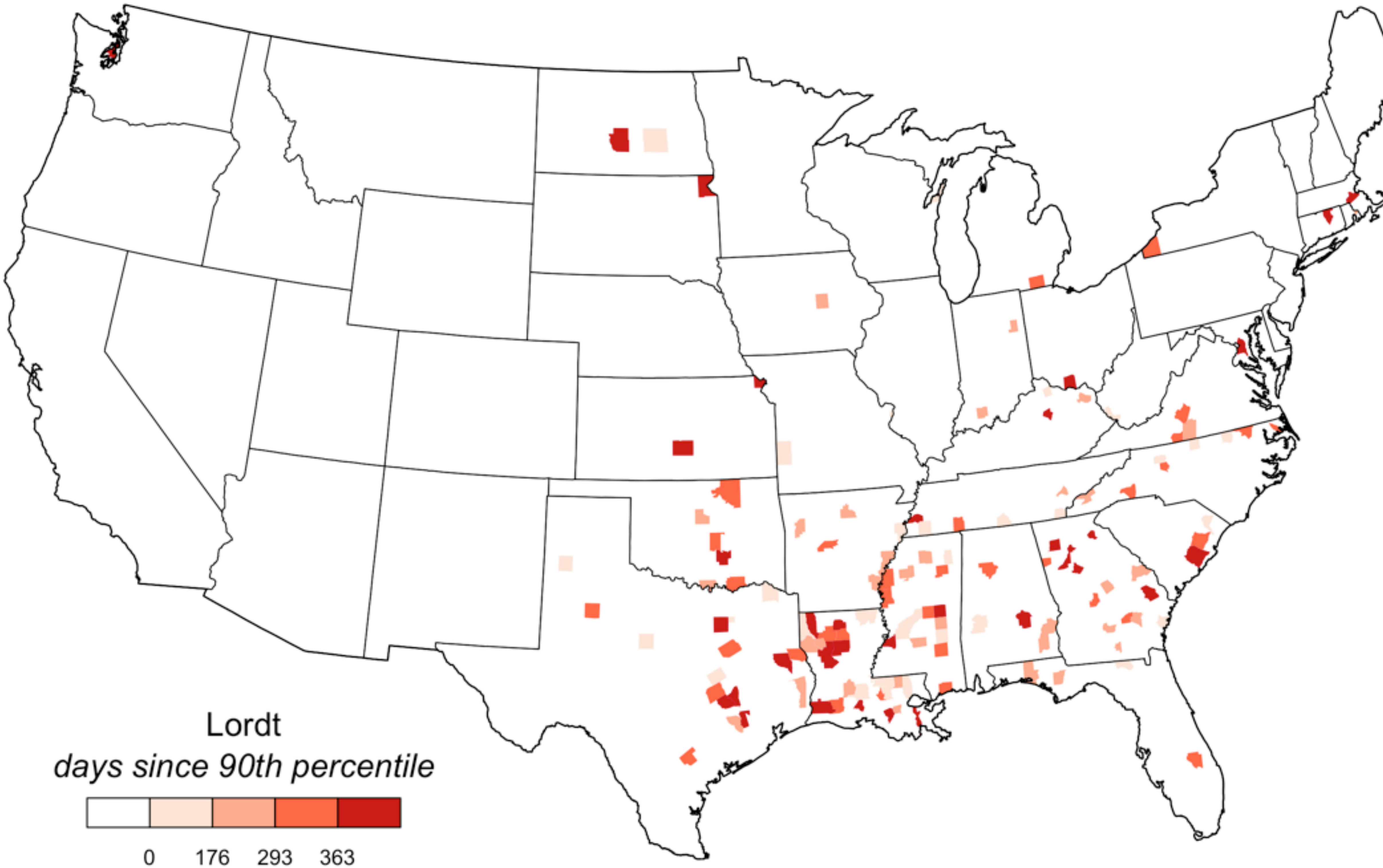
Mapping Lexical Emergence

A problem with this approach is that it is unclear (at least to me) how to identify common regional patterns based on a series of maps for each word.

To solve this problem spatial time series for each word was reduced to a single map by measuring the number of days since the word reached a specific relative frequency (e.g. 90th percentile overall) by county.

This controls for variation in amount of data across counties (as opposed to mapping days since first occurrence, which follows population density).


Baeless
days since 90th percentile
0
76
126
185



Lordt
days since 90th percentile
0
176
293
363

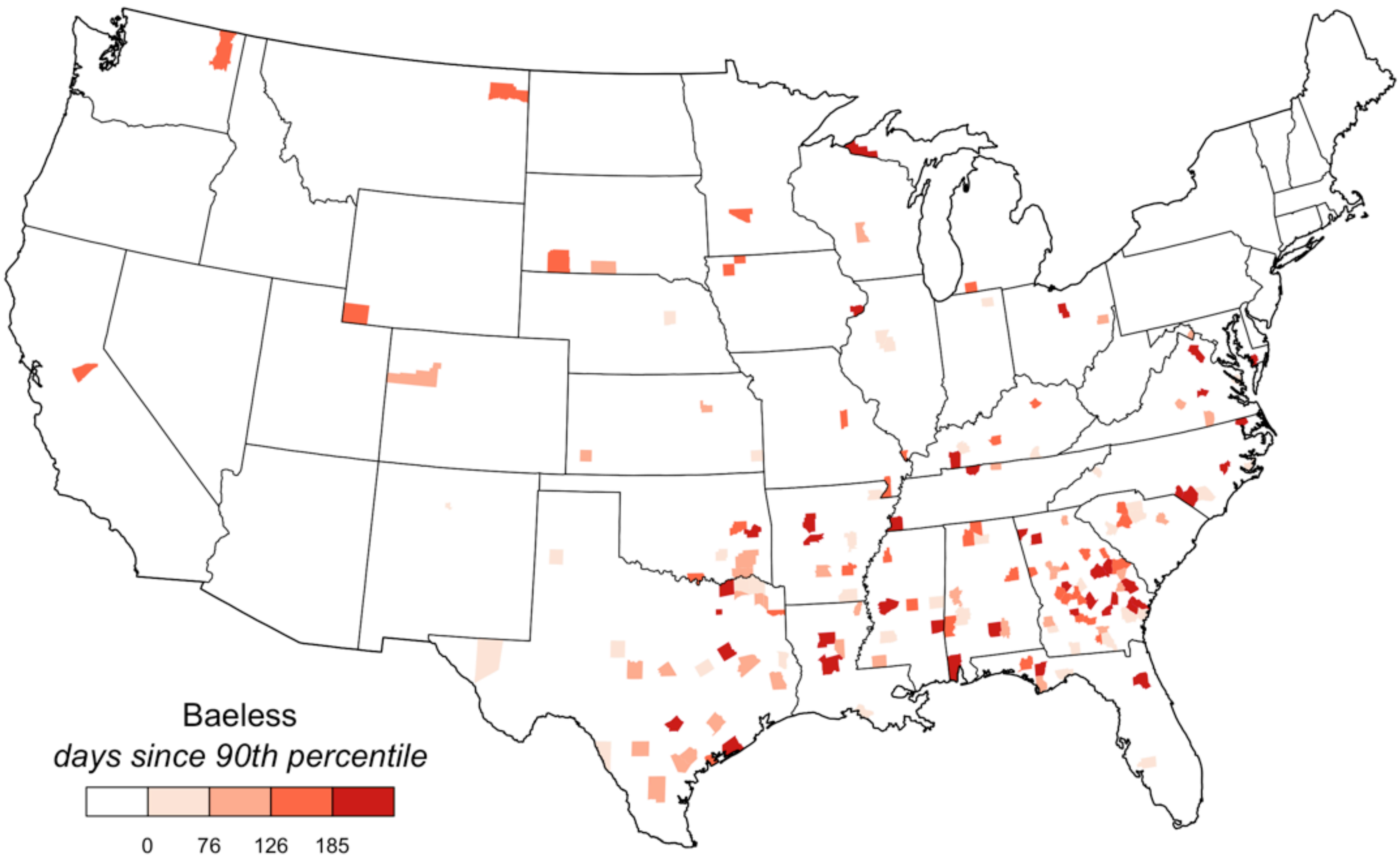
Mapping Lexical Emergence

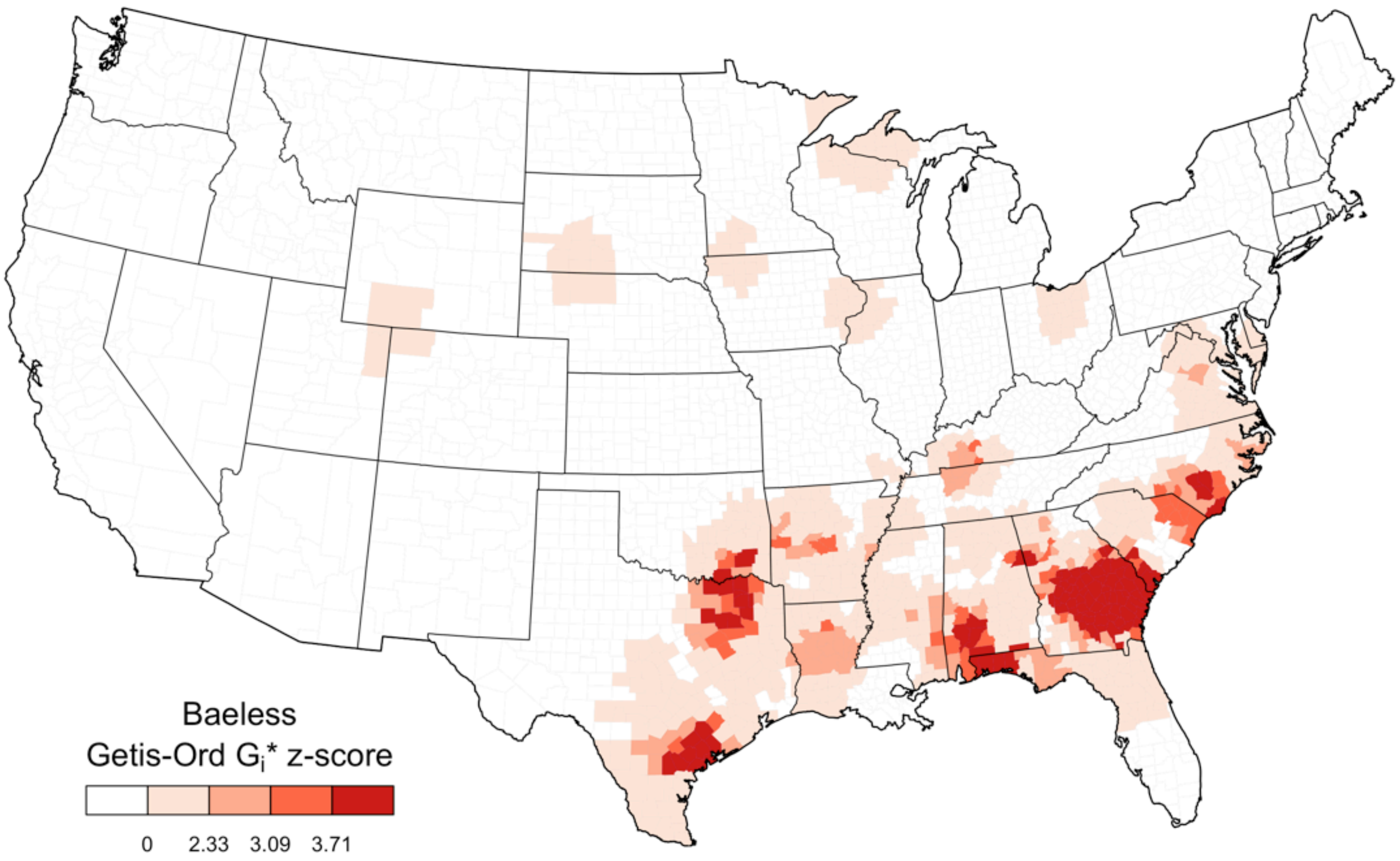
To identify hubs of lexical innovation the 90th percentile relative frequency threshold maps for all 54 words (measured across 3,075 counties) were subjected to a multivariate spatial analysis.

1. All maps are subjected to a Getis-Ord Gi* analysis to identify underlying regional signals.

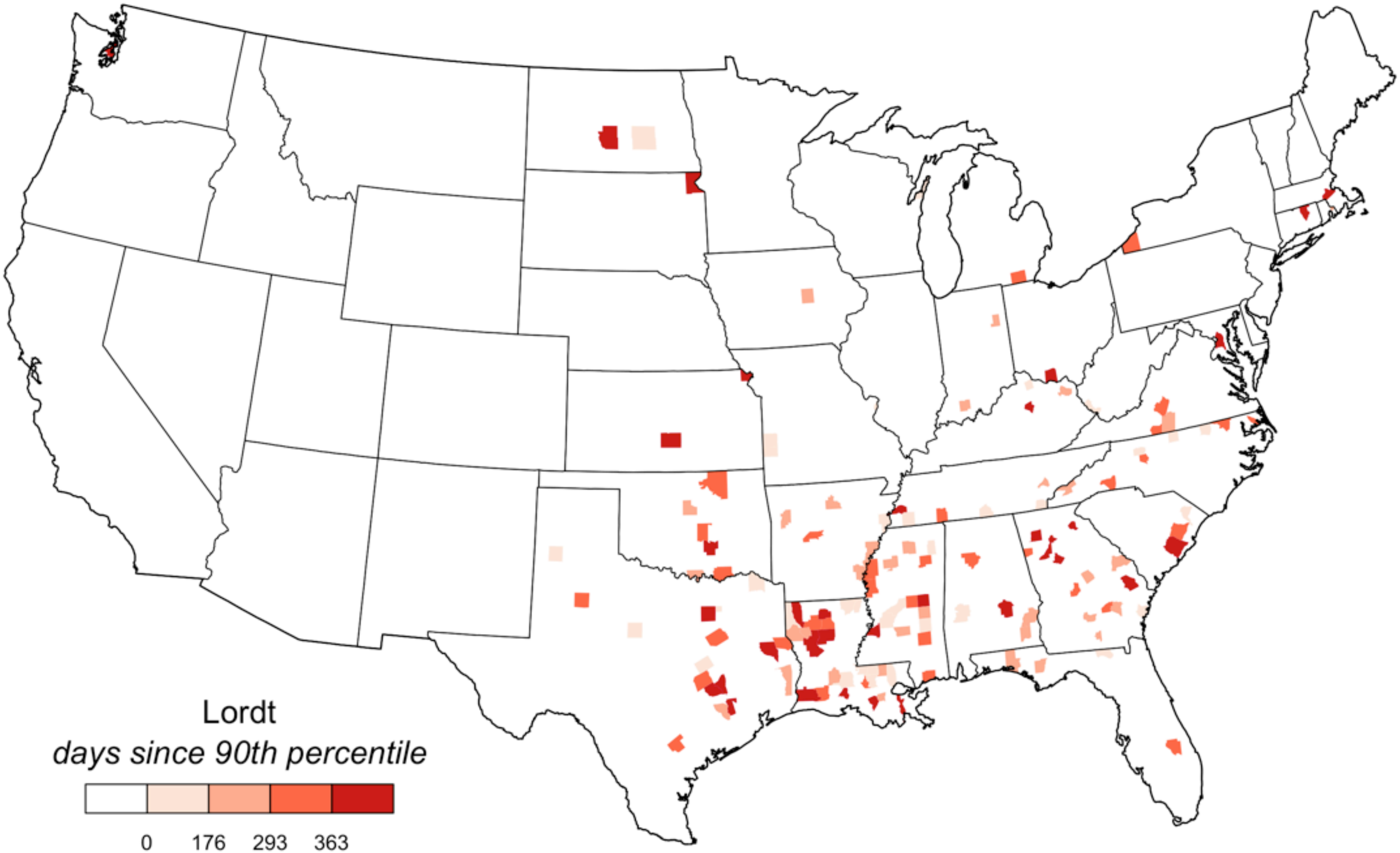
2. The smoothed maps are subjected to a factor analysis to identify common regional patterns.

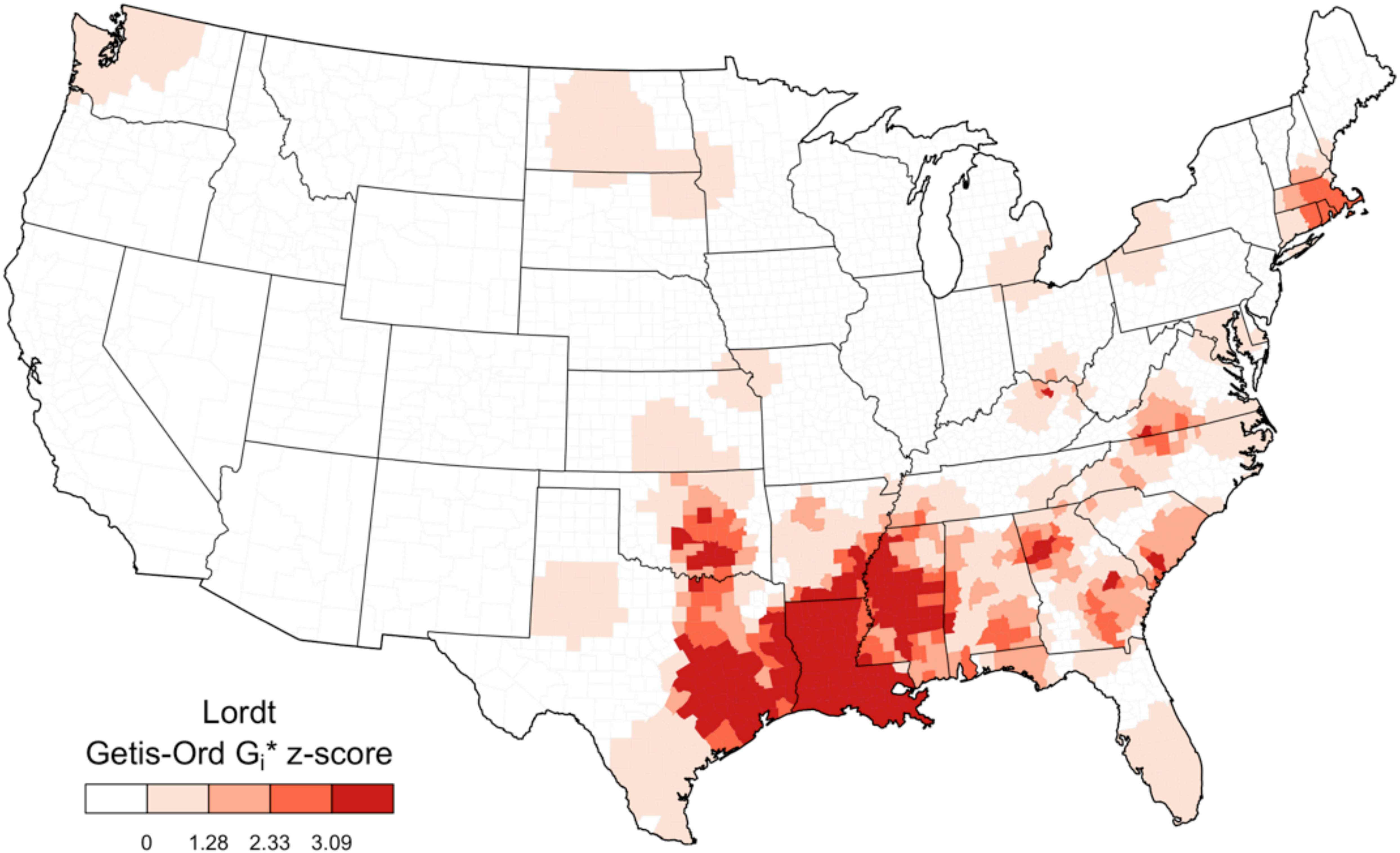
See Grieve. 2016. Regional Variation in Written American English. Cambridge University Press.


Baeless
days since 90th percentile
0
76
126
185



Baeless
Getis-Ord G_i* z-score
0
2.33
3.09
3.71


Lordt
days since 90th percentile
0
176
293
363


Lordt
Getis-Ord G_i^* z-score
0
1.28
2.33
3.09

Mapping Lexical Emergence

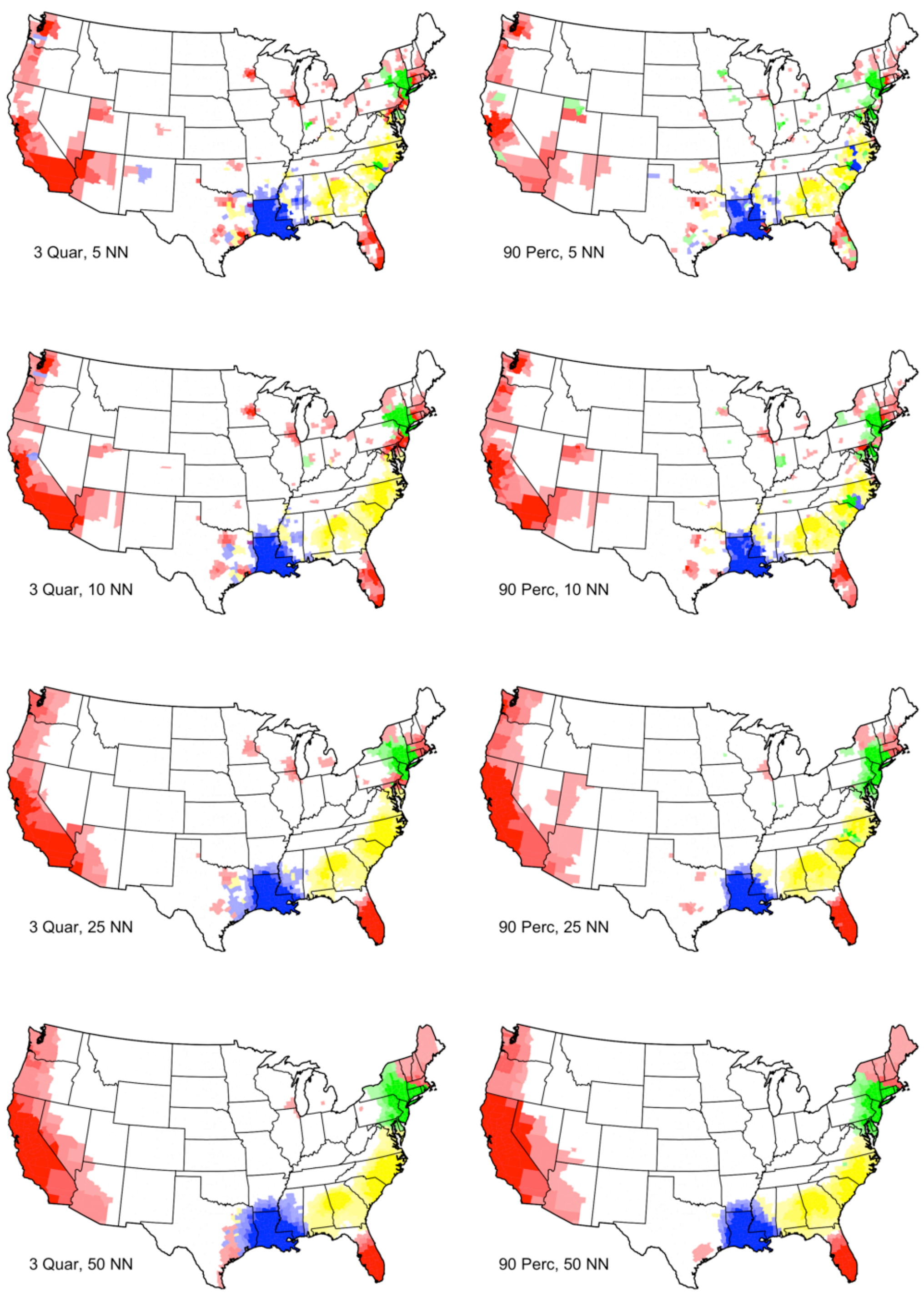
To identify hubs of lexical innovation the 90th percentile relative frequency threshold maps for all 54 words (measured across 3,075 counties) were subjected to a multivariate spatial analysis.

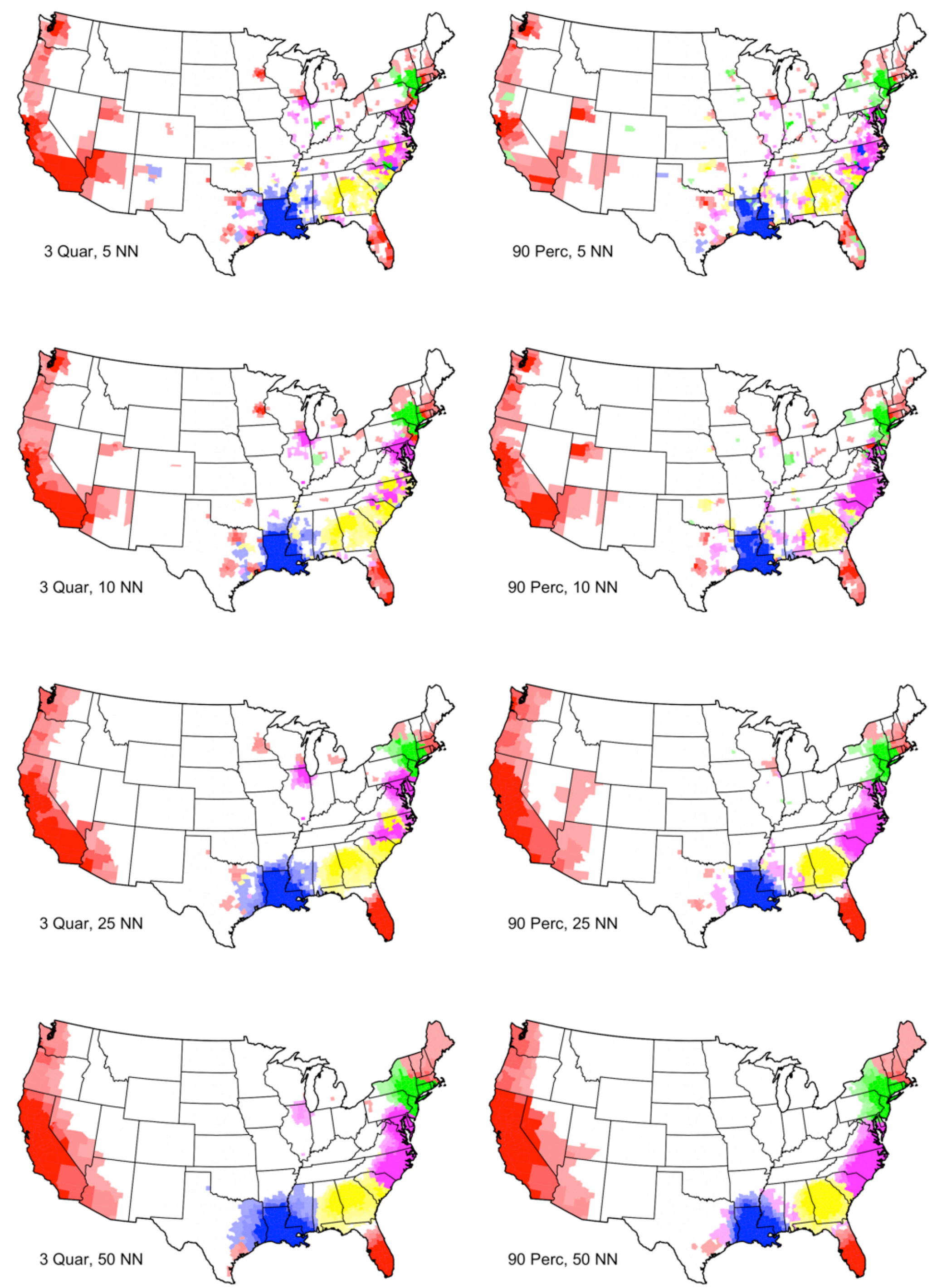
1. All maps are subjected to a Getis-Ord Gi* analysis to identify underlying regional signals.

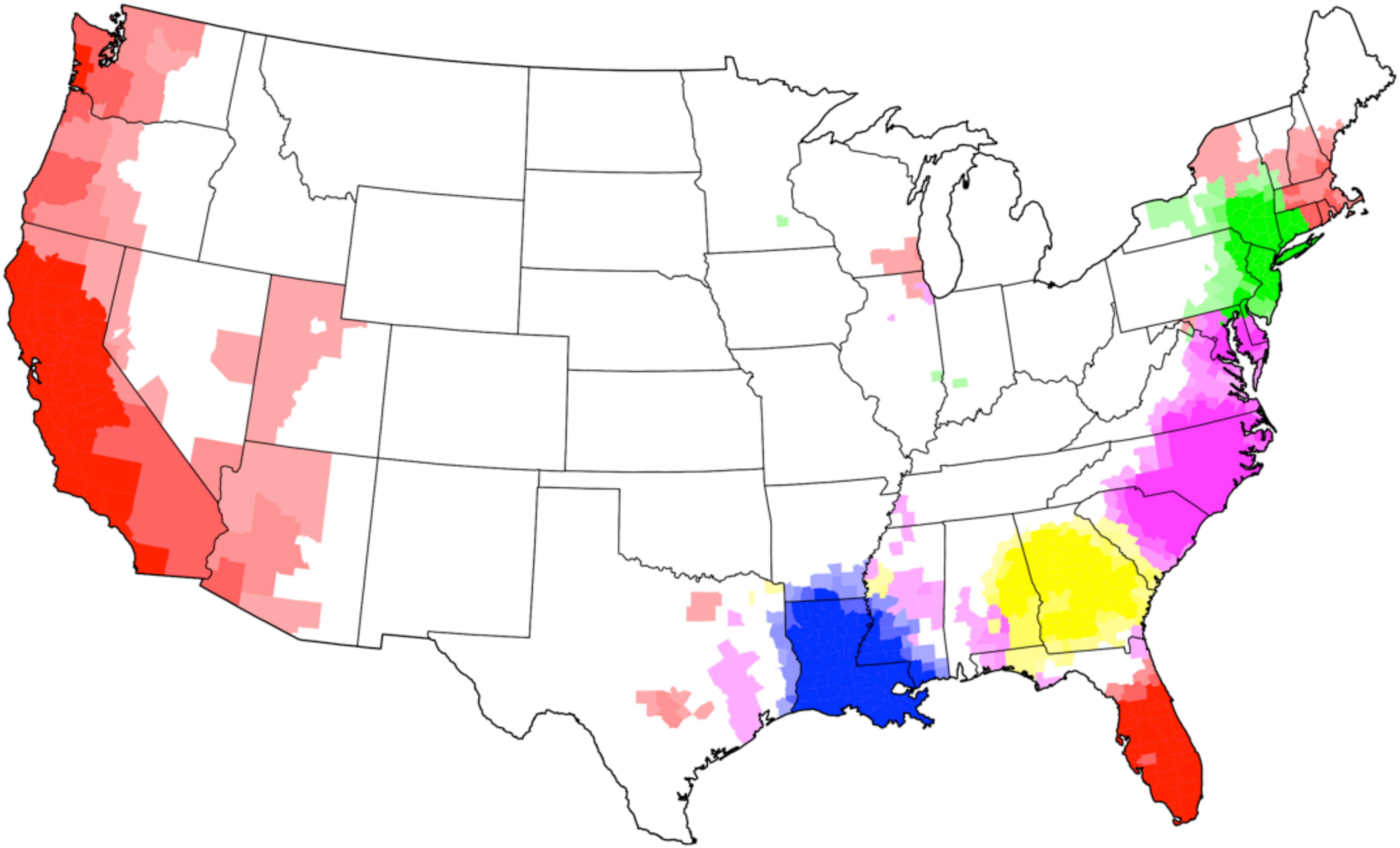
2. The smoothed maps are subjected to a Factor Analysis to identify common regional patterns.

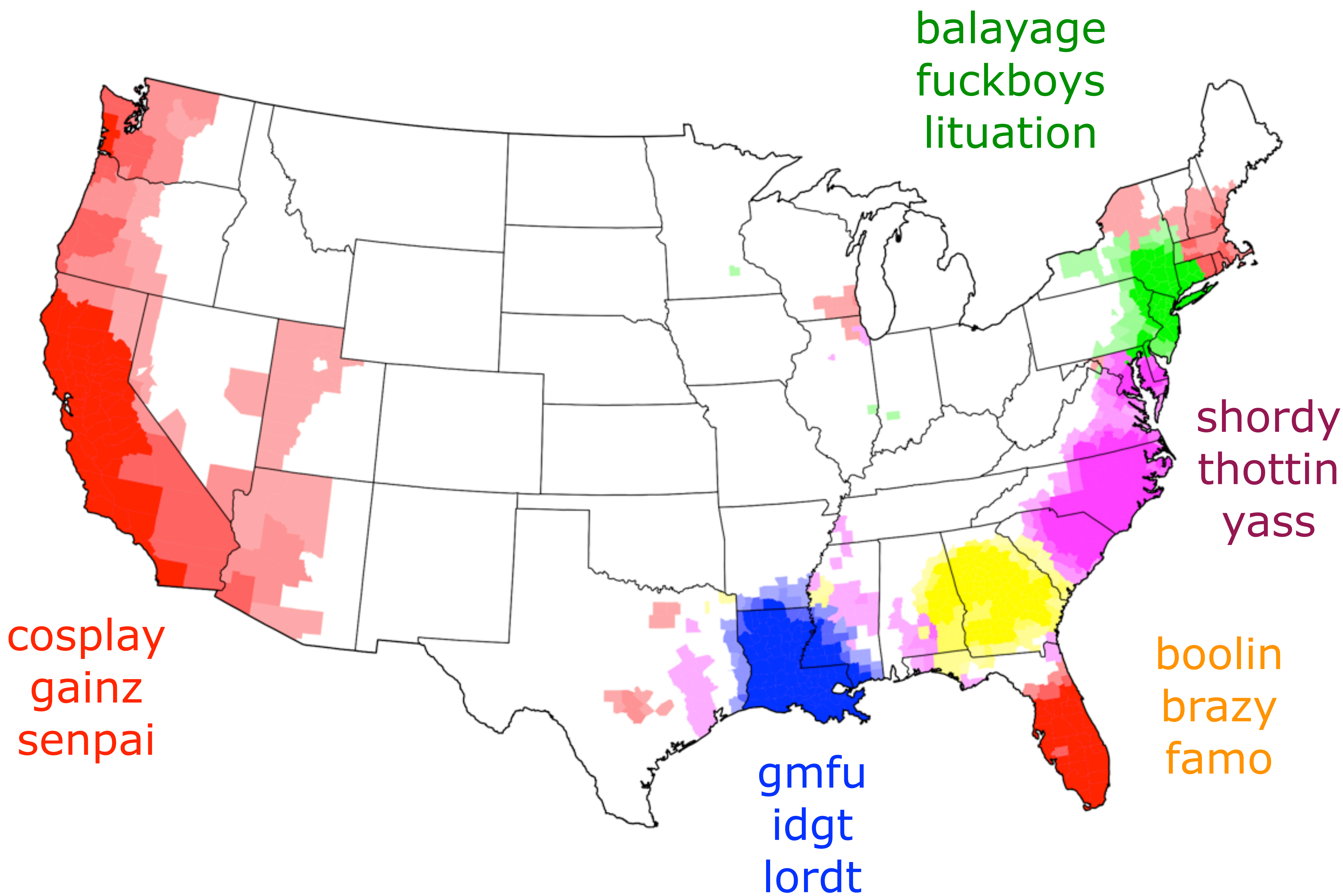
See Grieve. 2016. Regional Variation in Written American English. Cambridge University Press.

Results Across Parameter Settings







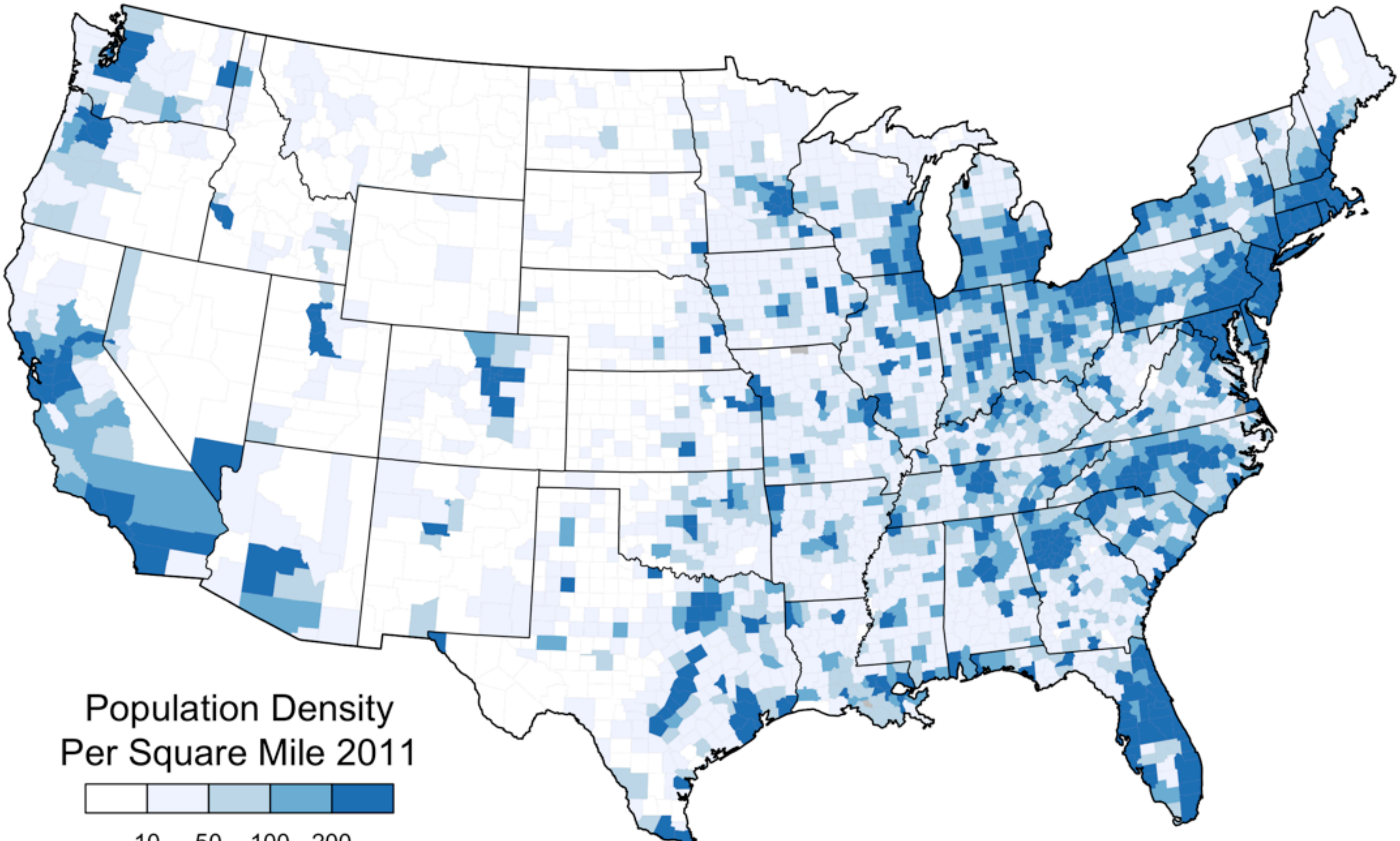

balayage
fuckboys
lituation
shordy
thottin
yass
boolin
brazy
famo
gmfu
idgt
lordt
cosplay
gainz
senpai

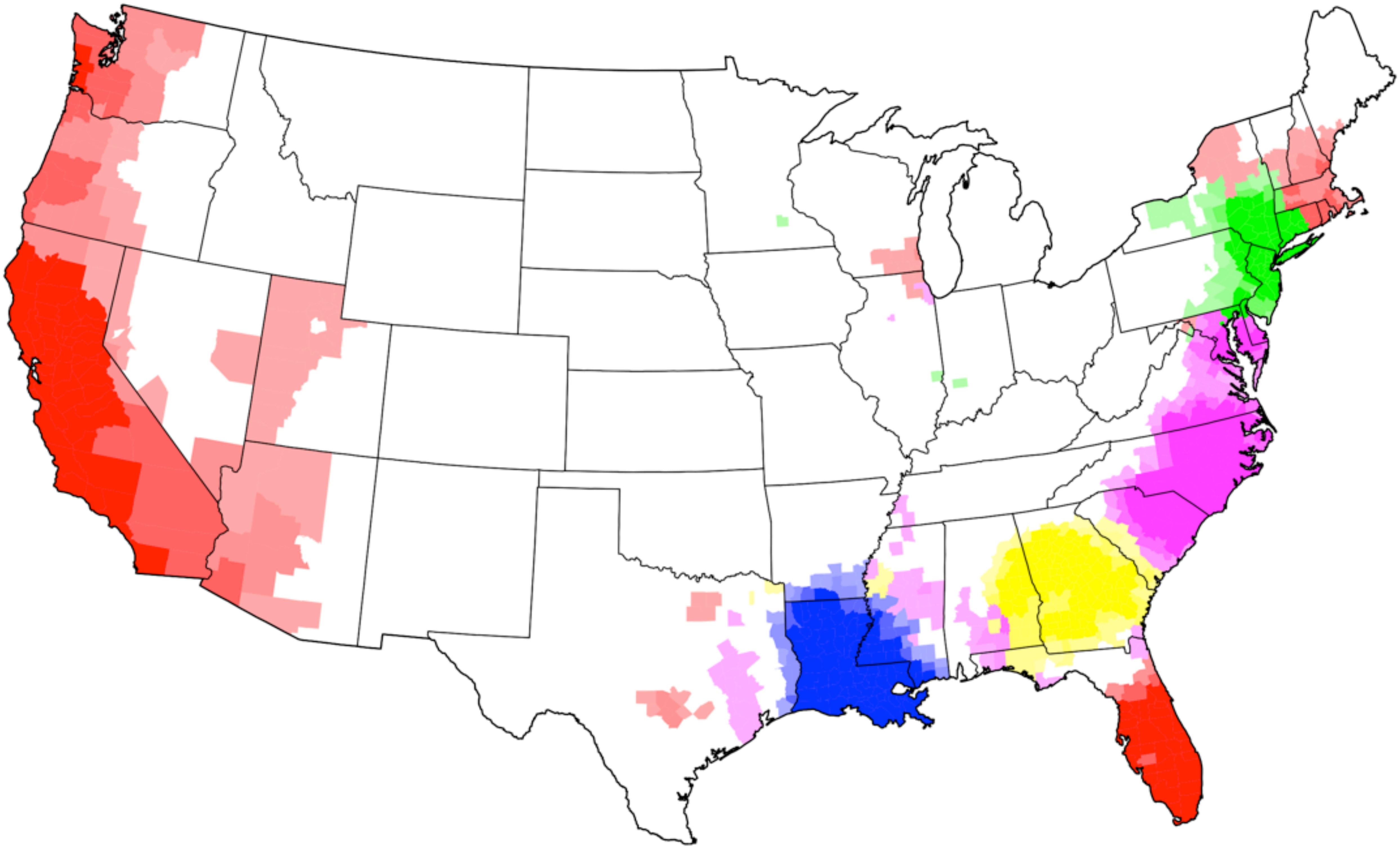
Conclusions

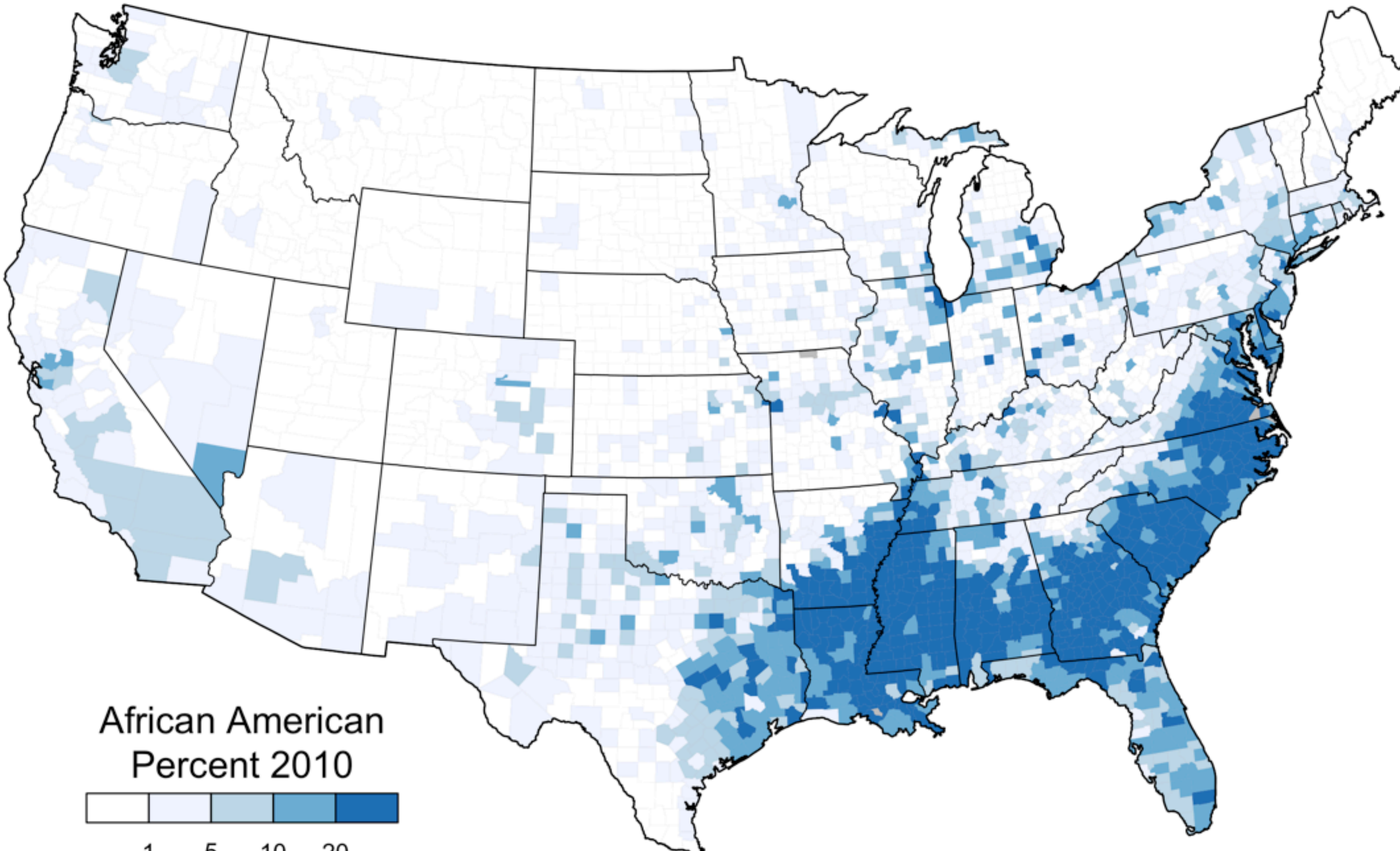
Five main hubs of lexical innovation were identified on American Twitter from 2014.

Newly emerging words do not originate from one location or at random, nor do they tend to simply follow patterns of population density.

In addition to California and New York clusters, three distinct hubs were surprisingly found in the Southeastern United States, attesting to the influence and diversity of African American English (and perhaps revealing AAE dialect regions).


Population Density
Per Square Mile 2011
10
50
100
200




African American
Percent 2010
1
5
10
20

Future Research

How do these patterns change over time and over registers?

How do newly emerging words spread across space (e.g. wave vs. gravity models)?

See Grieve, Nini, Guo. 2016. Mapping lexical innovation on American social media. In review at *Journal of English Linguistics*.

Identifying and Mapping the Spread of New Words

Jack Grieve
j.grieve1@aston.ac.uk
@JWGrieve
Centre for Forensic Linguistics
Aston University

2 December 2016
BAULT 2016
University of Helsinki