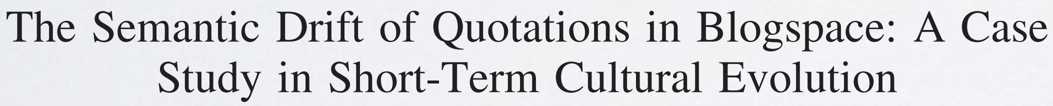
LEXICAL TRANSFORMATIONS IN BLOGSPACE

A CASE STUDY IN SHORT-TERM CULTURAL EVOLUTION

from



The Semantic Drift of Quotations in Blogspace: A Case Study in Short-Term Cultural Evolution

COGNITIVE SCIENCE
A Multidisciplinary Journal
(2017) 1–32

Sébastien Lerique
(EHESS / Centre Marc Bloch Berlin)

Camille Roth
(Sciences Po / Centre Marc Bloch Berlin)

EMPIRICAL STUDY OF CULTURAL EVOLUTION

IN VIVO

- *using historical data:*
  e.g.,
  - Morin 2013
  - Miton et al. 2015

Young Titian
(born c. 1485)
from 1500 to 1540
averted gaze
direct gaze

EMPIRICAL STUDY OF CULTURAL EVOLUTION

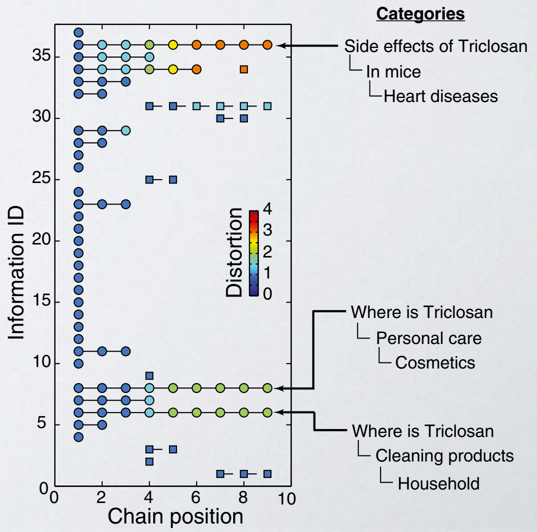
IN VIVO

- *using historical data:* e.g.,
  - Morin 2013
  - Miton et al. 2015

Young Titian
(born c. 1485)
from 1500 to 1540
averted gaze
direct gaze

IN VITRO

- *using transmission chains:* e.g.,
  - Claidière et al. 2014
  - Moussaïd et al. 2015

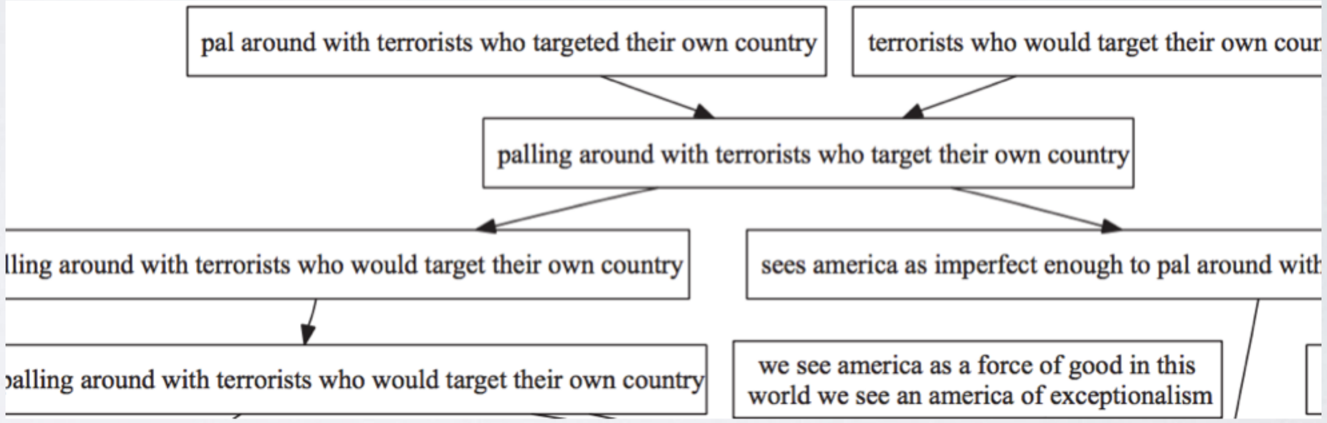


IN VIVO ONLINE DATA

(Leskovec, Backstrom, Kleinberg, 2009)

- Corpus of quotations from a large corpus of (8.5m) blog posts

(Aug'08-Apr'09)


pal around with terrorists who targeted their own country
terrorists who would target their own cour
palling around with terrorists who target their own country
lling around with terrorists who would target their own country
sees america as imperfect enough to pal around wit
alling around with terrorists who would target their own country
we see america as a force of good in this world we see an america of exceptionalism

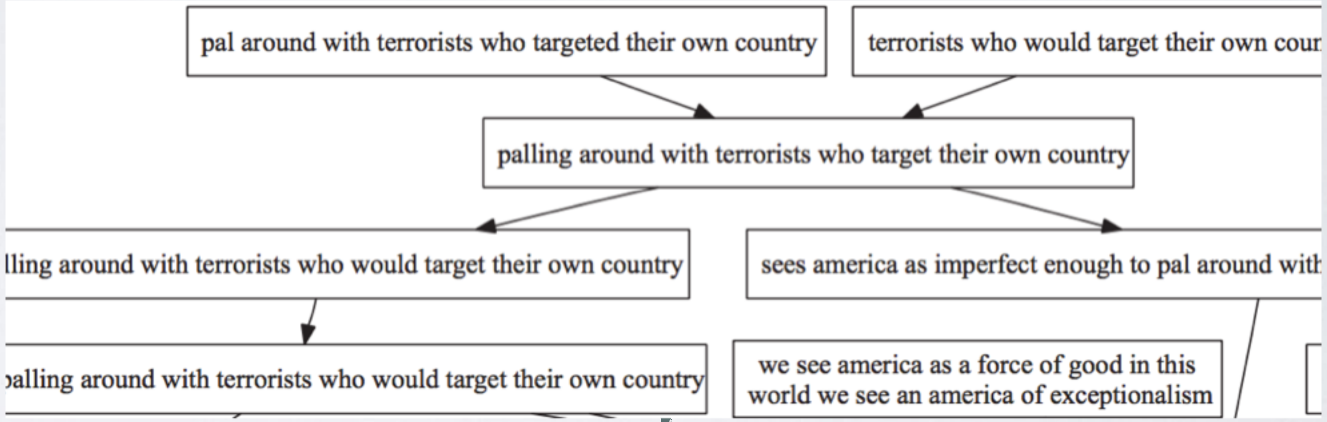
IN VIVO ONLINE DATA

(Leskovec, Backstrom, Kleinberg, 2009)

- Corpus of quotations from a large corpus of (8.5m) blog posts

  (Aug'08-Apr'09)

- Groups (and dynamics) of sentences



pal around with terrorists who targeted their own country

terrorists who would target their own coun

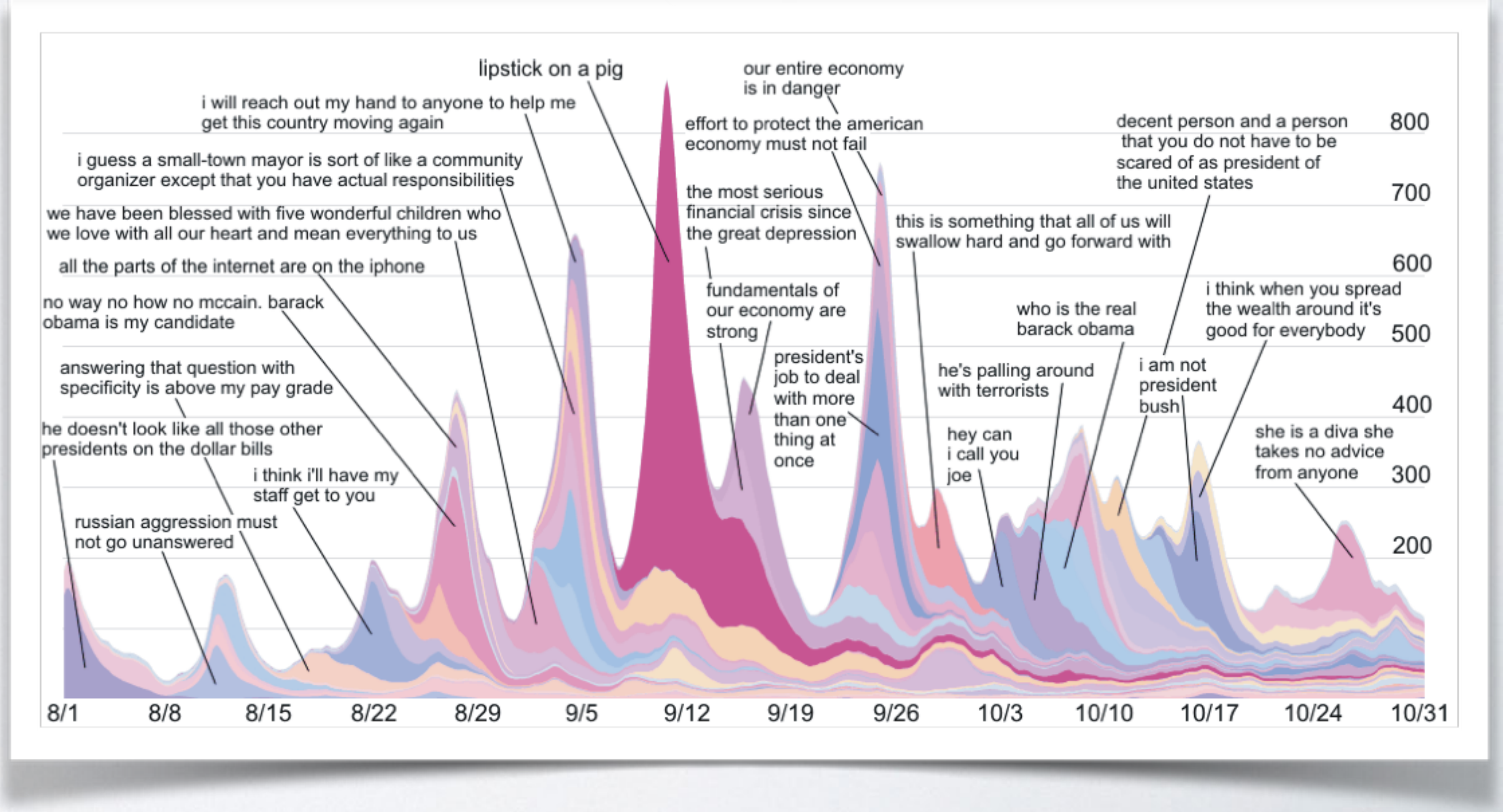
palling around with terrorists who target their own country

lling around with terrorists who would target their own country

sees america as imperfect enough to pal around with

palling around with terrorists who would target their own country

we see america as a force of good in this world we see an america of exceptionalism



SENTENCE REFORMULATION

- *Pakistani President Asif Ali Zardari:*
    - “we will not be scared of these cowards”
    - ▷ “we will not be **afraid** of these cowards.”

- *US Senator McCain:*
    - “I admire Senator Obama and his accomplishments”
    - ▷ “I **respect** Senator Obama and his accomplishments.”

SENTENCE REFORMULATION

- *Pakistani President Asif Ali Zardari:*
  - “we will not be scared of these cowards”
  - “we will not be **afraid** of these cowards.”

- *US Senator McCain:*
  - “I admire Senator Obama and his accomplishments”
  - “I **respect** Senator Obama and his accomplishments.”

- Task similar to word (list) recall

SENTENCE REFORMULATION

- *Pakistani President Asif Ali Zardari:*
  - “we will not be scared of these cowards”
  - ▹ “we will not be **afraid** of these cowards.”

- *US Senator McCain:*
  - “I admire Senator Obama and his accomplishments”
  - ▹ “I **respect** Senator Obama and his accomplishments.”

- Task similar to word (list) recall
- Lexical features expected to influence the likelihood of substitution

SENTENCE REFORMULATION

- *Pakistani President Asif Ali Zardari:*
  - “we will not be scared of these cowards”
  - ▷ “we will not be **afraid** of these cowards.”

- *US Senator McCain:*
  - “I admire Senator Obama and his accomplishments”
  - ▷ “I **respect** Senator Obama and his accomplishments.”

- Task similar to word (list) recall
- Lexical features expected to influence the likelihood of substitution
  - *for instance:* word frequency, age of acquisition, number of phonemes, phonological neighborhood density, position in a semantic network...

SENTENCE REFORMULATION

- *Pakistani President Asif Ali Zardari:*
    - “we will not be scared of these cowards”
    - ▷ “we will not be **afraid** of these cowards.”

- *US Senator McCain:*
    - “I admire Senator Obama and his accomplishments”
    - ▷ “I **respect** Senator Obama and his accomplishments.”

- Task similar to word (list) recall
- Lexical features expected to influence the likelihood of substitution
    - *for instance:* word frequency, age of acquisition, number of phonemes, phonological neighborhood density, position in a semantic network...
- Address e.g., the "word-frequency paradox" (Mandler et al. 1982)

SENTENCE REFORMULATION

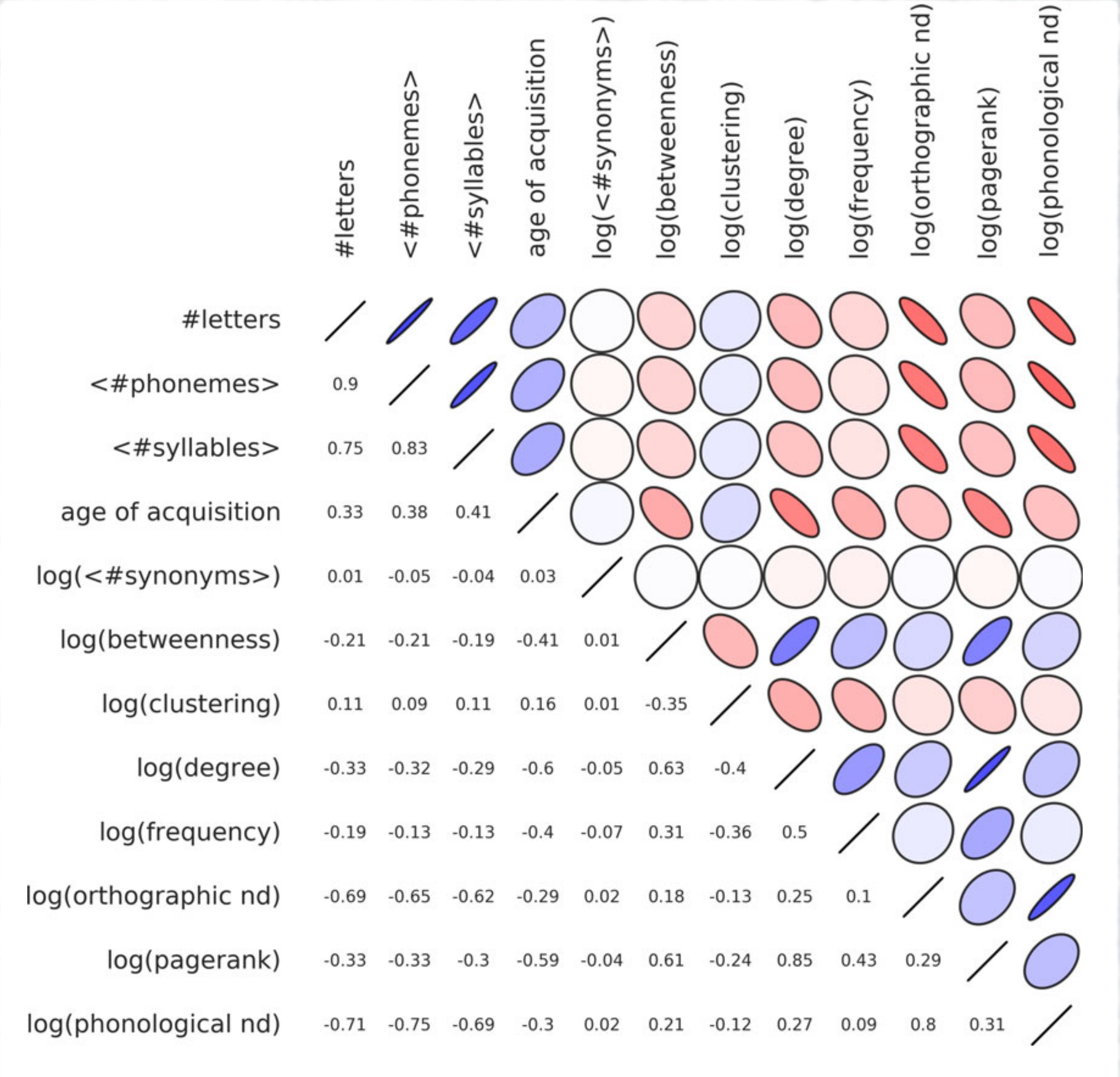
- *Pakistani President Asif Ali Zardari:*
  - “we will not be scared of these cowards”
  - ▶ “we will not be **afraid** of these cowards.”

- *US Senator McCain:*
  - “I admire Senator Obama and his accomplishments”
  - ▶ “I **respect** Senator Obama and his accomplishments.”

- Task similar to word (list) recall

- Lexical features expected to influence the likelihood of substitution
  - *for instance:* word frequency, age of acquisition, number of phonemes, phonological neighborhood density, position in a semantic network...

- Address e.g., the "word-frequency paradox" (Mandler et al. 1982)



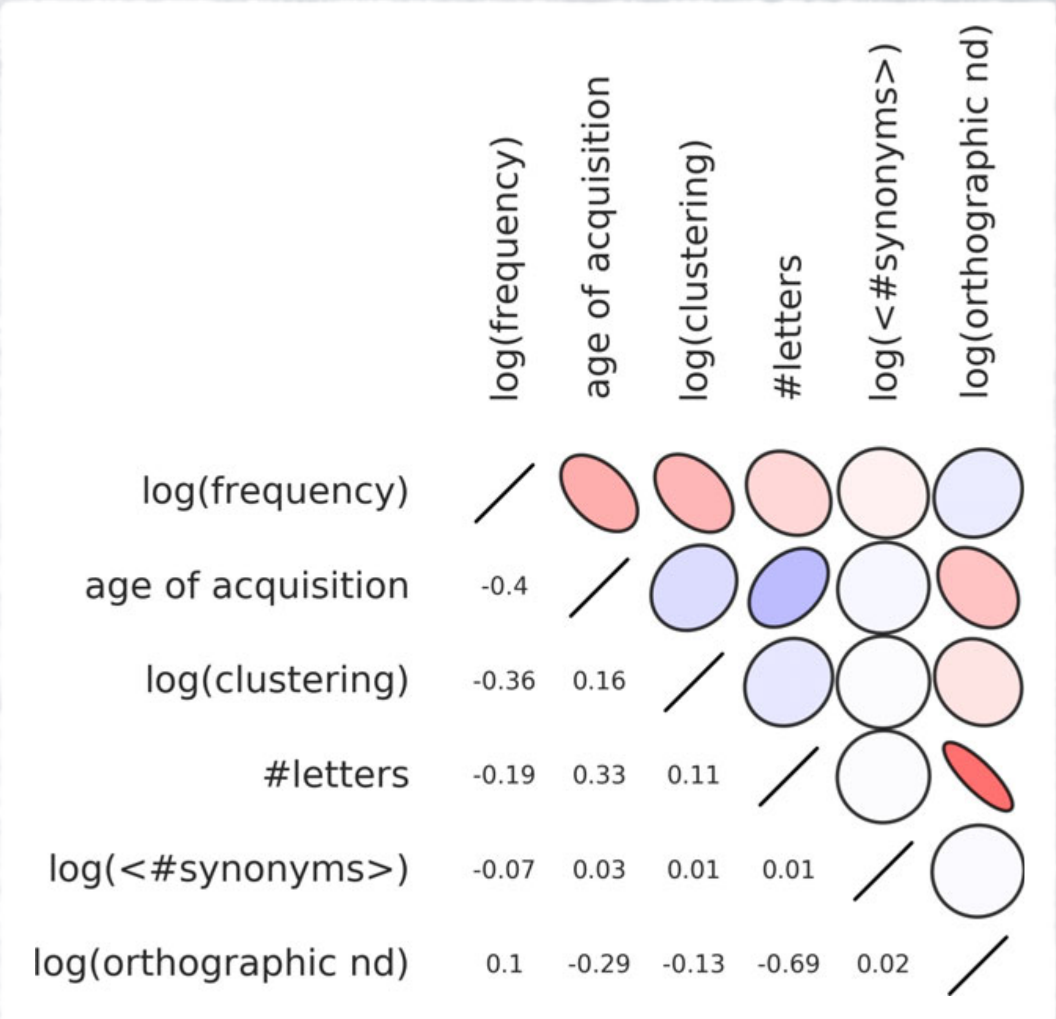
	#letters	<#phonemes>	<#syllables>	age of acquisition	log(<#synonyms>)	log(betweenness)	log(clustering)	log(degree)	log(frequency)	log(orthographic nd)	log(pagerank)	log(phonological nd)
#letters												
<#phonemes>	0.9											
<#syllables>	0.75	0.83										
age of acquisition	0.33	0.38	0.41									
log(<#synonyms>)	0.01	-0.05	-0.04	0.03								
log(betweenness)	-0.21	-0.21	-0.19	-0.41	0.01							
log(clustering)	0.11	0.09	0.11	0.16	0.01	-0.35						
log(degree)	-0.33	-0.32	-0.29	-0.6	-0.05	0.63	-0.4					
log(frequency)	-0.19	-0.13	-0.13	-0.4	-0.07	0.31	-0.36	0.5				
log(orthographic nd)	-0.69	-0.65	-0.62	-0.29	0.02	0.18	-0.13	0.25	0.1			
log(pagerank)	-0.33	-0.33	-0.3	-0.59	-0.04	0.61	-0.24	0.85	0.43	0.29		
log(phonological nd)	-0.71	-0.75	-0.69	-0.3	0.02	0.21	-0.12	0.27	0.09	0.8	0.31	

Fig. 1. Spearman correlations in the initial set of features.

SENTENCE REFORMULATION

- *Pakistani President Asif Ali Zardari:*
  - “we will not be scared of these cowards”
  - “we will not be **afraid** of these cowards.”

- *US Senator McCain:*
  - “I admire Senator Obama and his accomplishments”
  - “I **respect** Senator Obama and his accomplishments.”



	log(frequency)	age of acquisition	log(clustering)	#letters	log(<#synonyms>)	log(orthographic nd)
log(frequency)						
age of acquisition	-0.4					
log(clustering)	-0.36	0.16				
#letters	-0.19	0.33	0.11			
log(<#synonyms>)	-0.07	0.03	0.01	0.01		
log(orthographic nd)	0.1	-0.29	-0.13	-0.69	0.02	

Fig. 2. Spearman correlations in the filtered set of feature.

- Task similar to word (list) recall
- Lexical features expected to influence the likelihood of substitution
  - *for instance:* word frequency, age of acquisition, number of phonemes, phonological neighborhood density, position in a semantic network...
- Address e.g., the "word-frequency paradox" (Mandler et al. 1982)

SUBSTITUTION MODEL

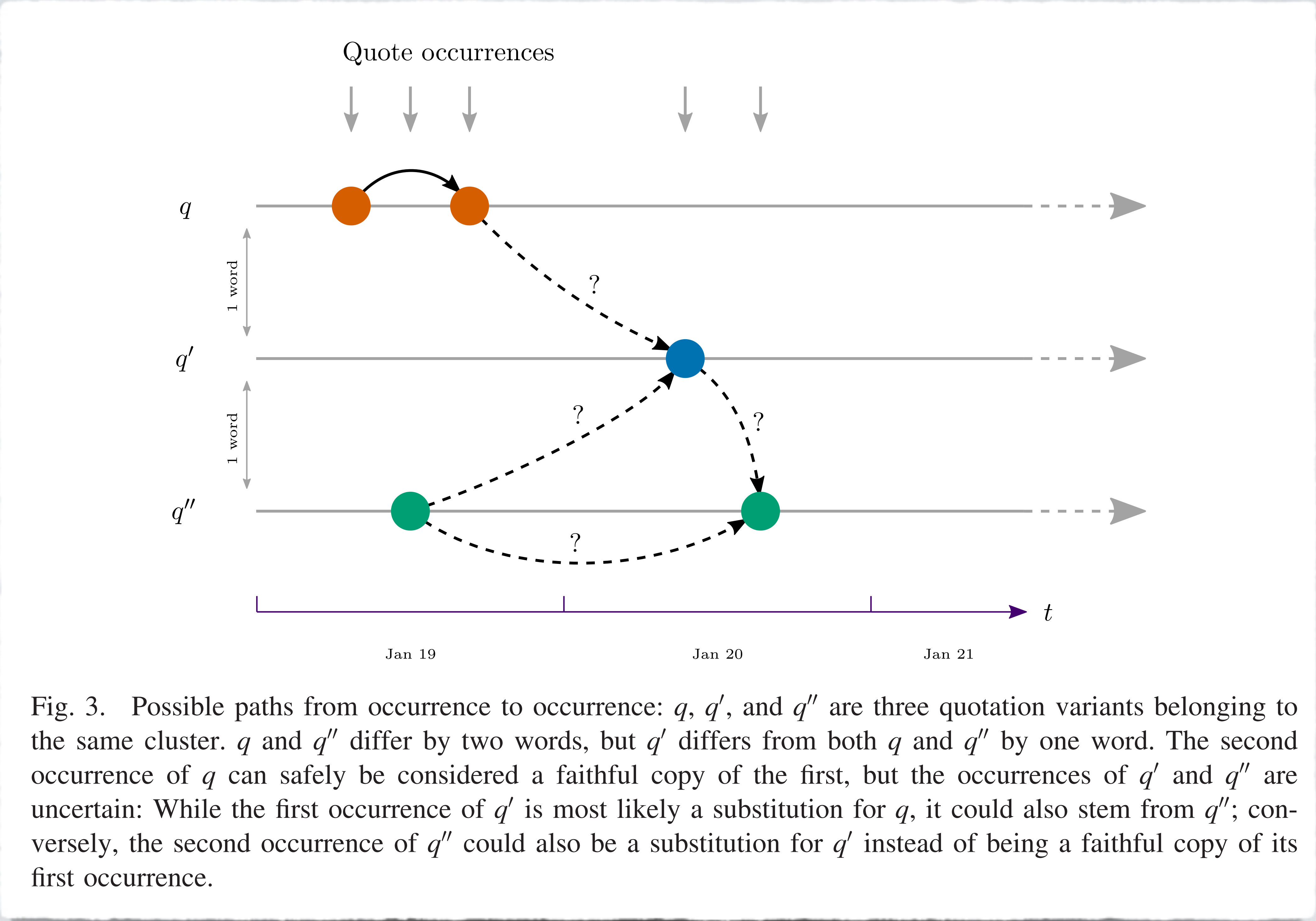


Fig. 3. Possible paths from occurrence to occurrence: q, q', and q'' are three quotation variants belonging to the same cluster. q and q'' differ by two words, but q' differs from both q and q'' by one word. The second occurrence of q can safely be considered a faithful copy of the first, but the occurrences of q' and q'' are uncertain: While the first occurrence of q' is most likely a substitution for q, it could also stem from q''; conversely, the second occurrence of q'' could also be a substitution for q' instead of being a faithful copy of its first occurrence.

SUSCEPTIBILITY

$$\sigma_g = \frac{s_g}{s_g^0}$$

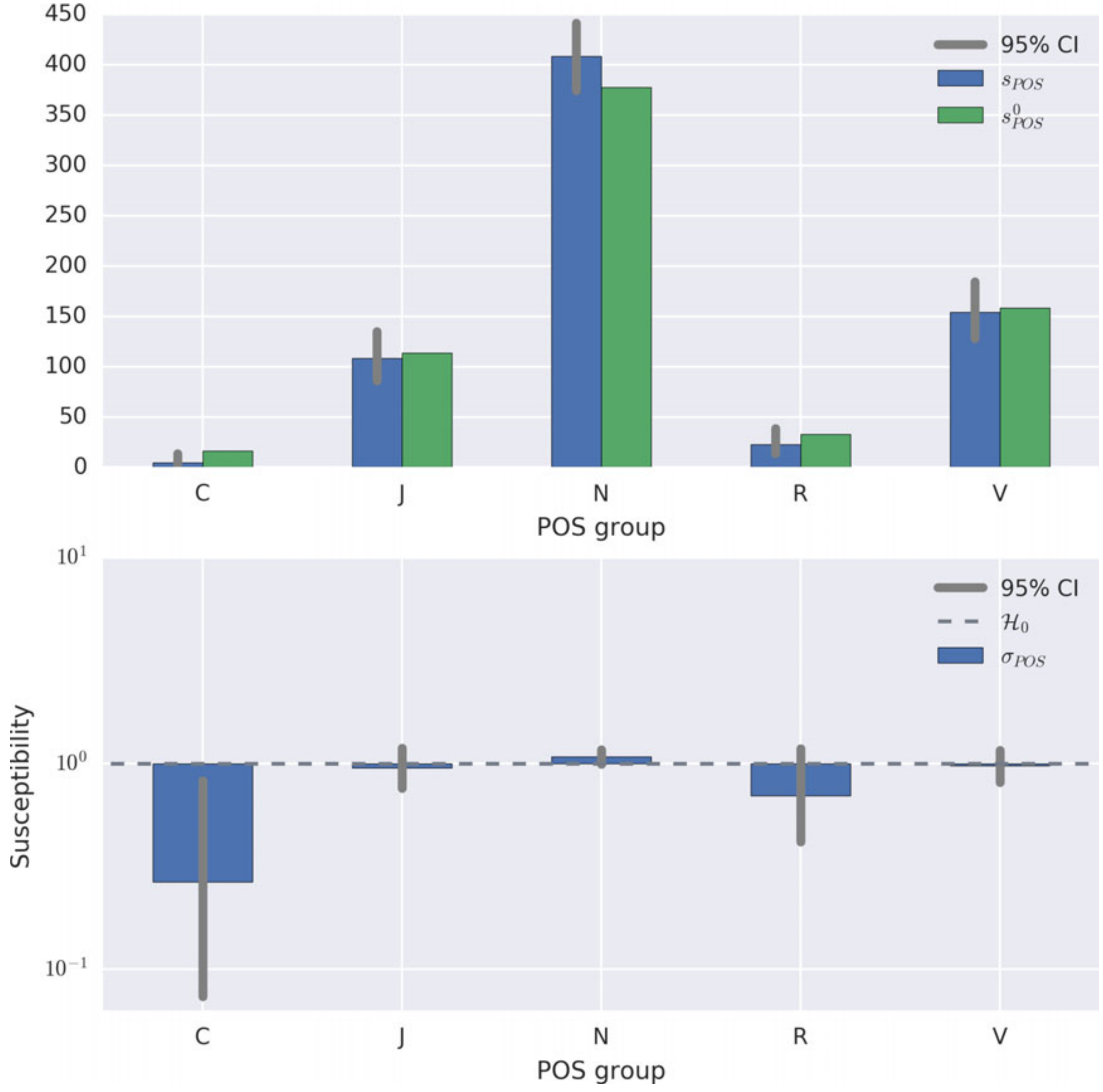
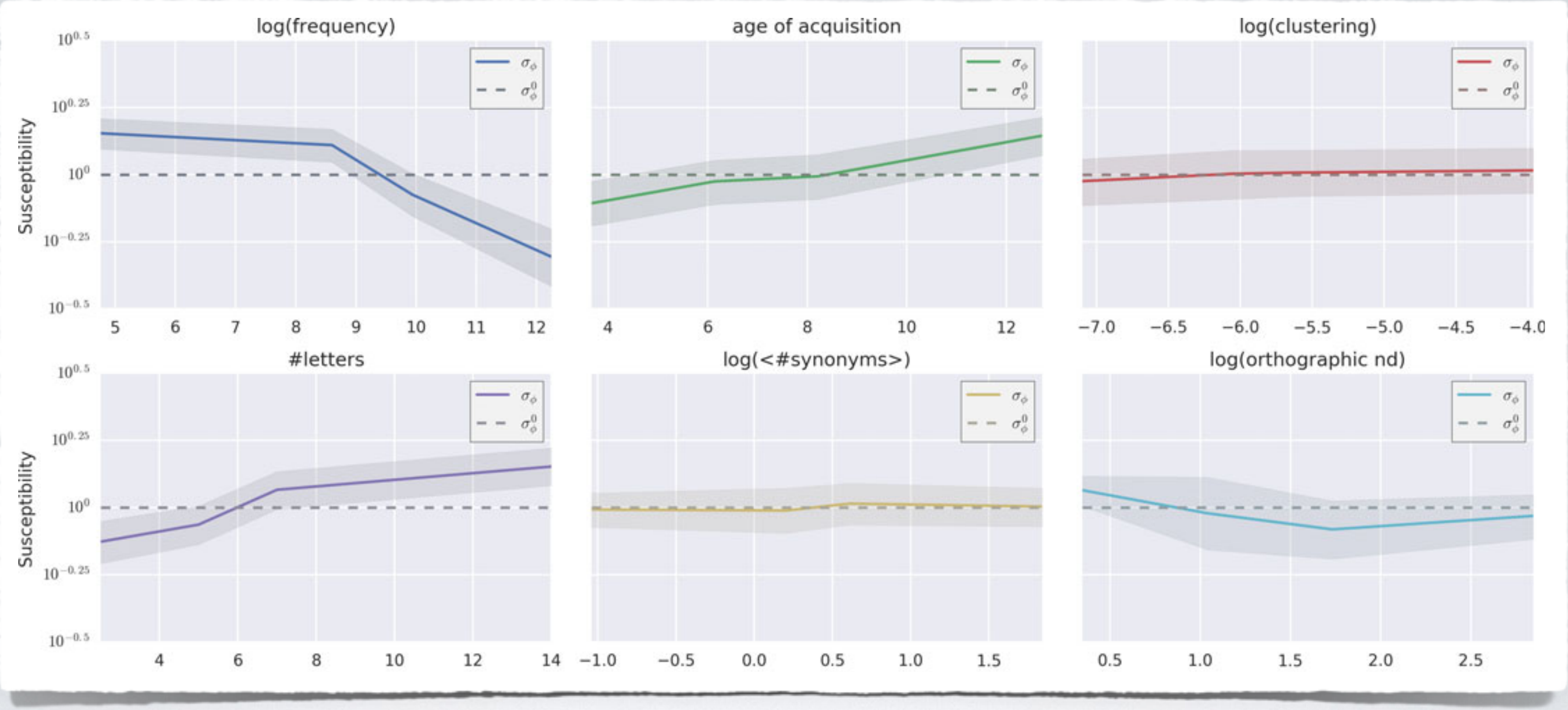


Fig. 5. Part-of-Speech-related results: Categories are simplified from the TreeTagger tag set: *C* means *Closed class-like* (see main text for details), *J* means adjective, *N* noun, *R* adverb, and *V* means verb. The top panel shows the actual s_{POS} and s^0_{POS} counts. The bottom panel shows the substitution susceptibility σ_{POS}, which is the ratio between the two previous counts. Confidence intervals are computed with the Goodman (1965) method for multinomial proportions.

SUSCEPTIBILITY

$$\sigma_g = \frac{s_g}{s_g^0}$$



log(frequency)

age of acquisition

log(clustering)

#letters

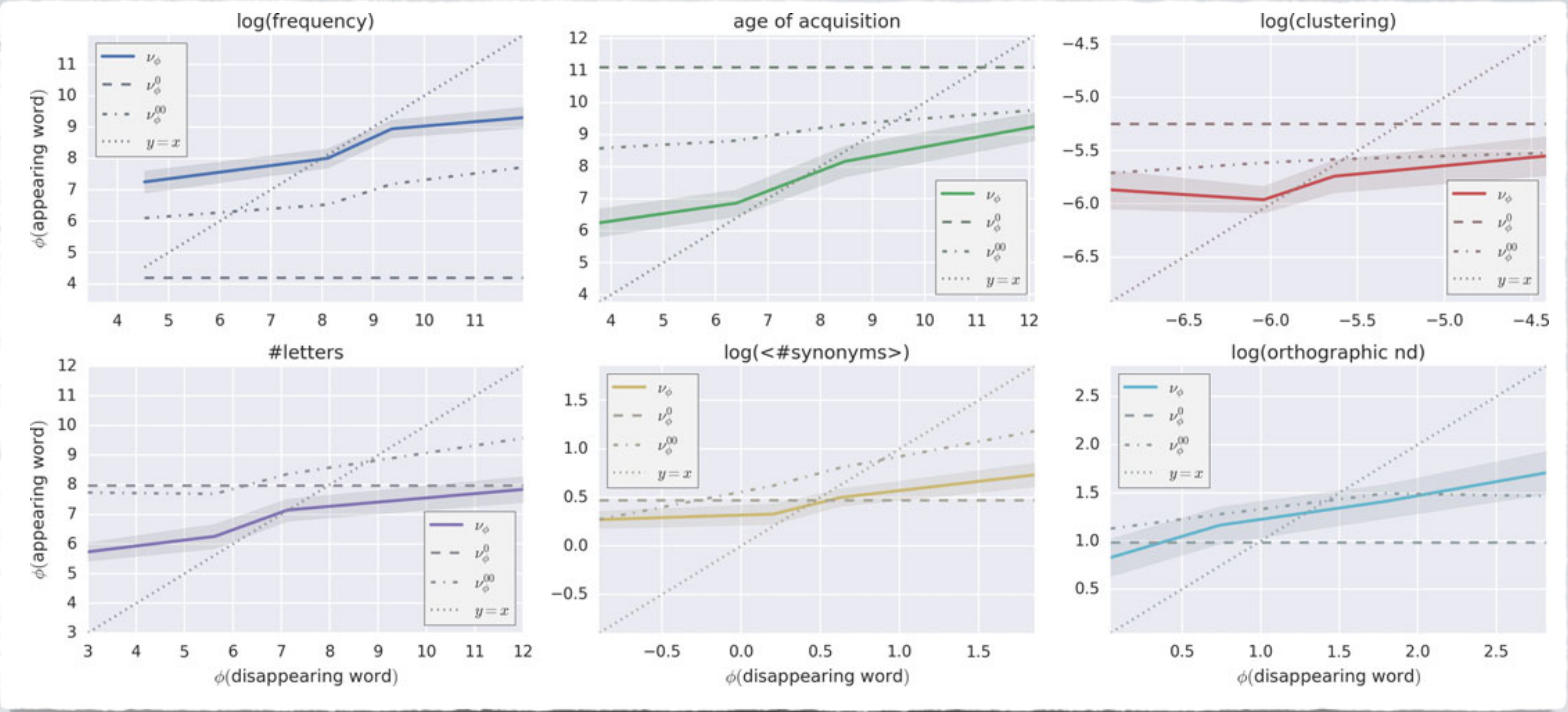
log(<#synonyms>)

log(orthographic nd)

On the whole, the trends observed are consistent with known effects of word frequency, age of acquisition, and number of letters, indicating that the triggering of a substitution could behave quite similarly to word recall in standard tasks.

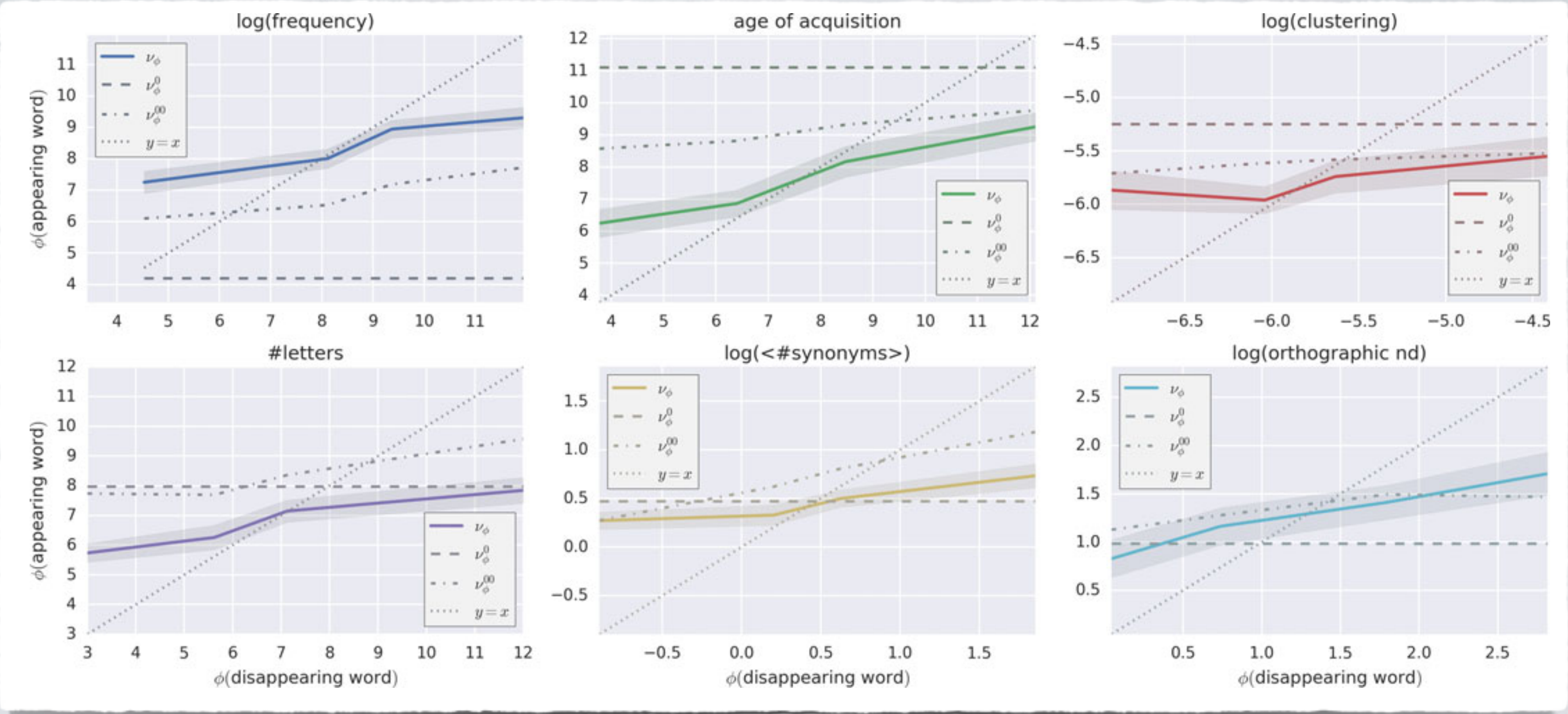
FEATURE VARIATION

$$\nu_\phi(f) = \langle \phi(w') \rangle_{\{w \to w' | \phi(w) = f\}}$$



FEATURE VARIATION

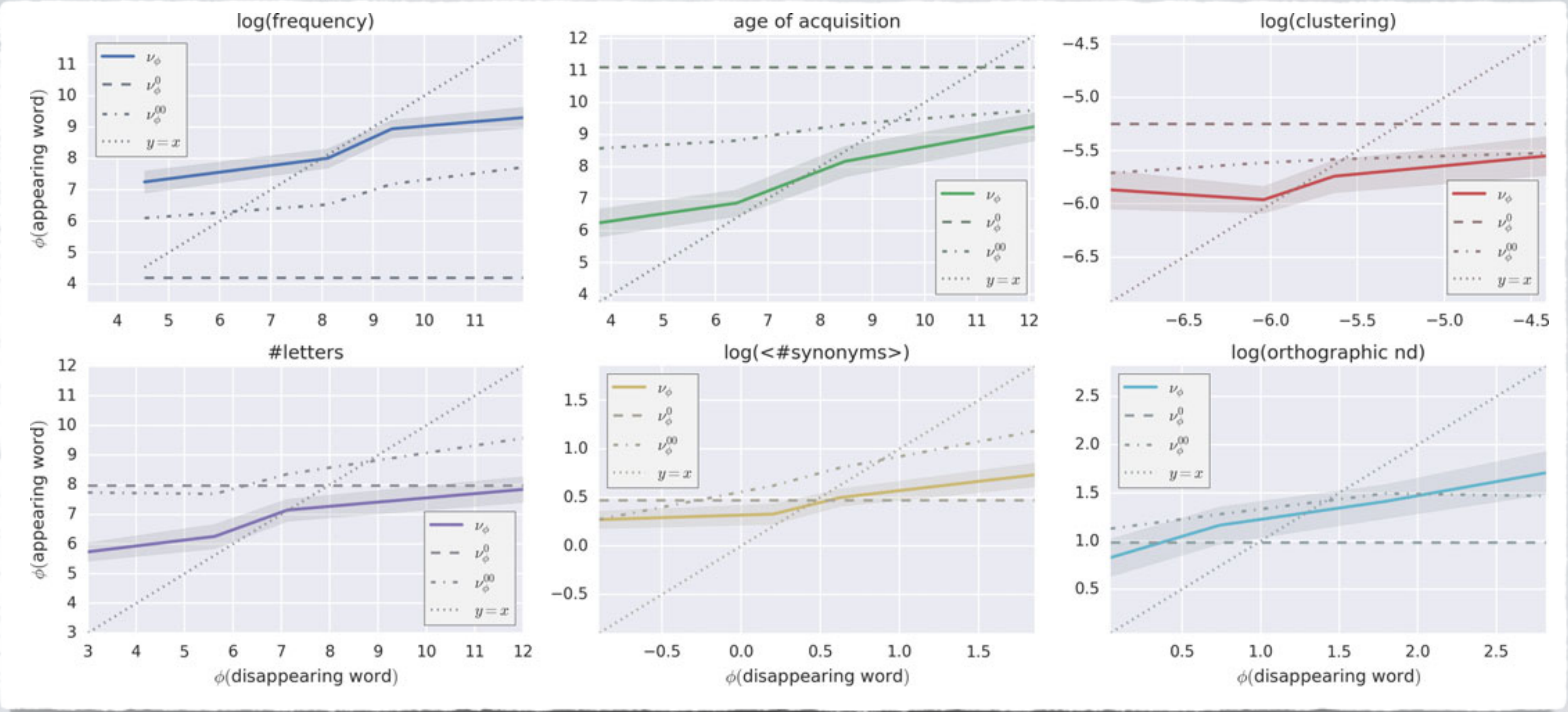
$$\nu_\phi(f) = \langle \phi(w') \rangle_{\{w \to w' | \phi(w) = f\}}$$



First, there is a single intersection of ν_φ with y=x and the slope of ν_φ remains smaller than 1:
the substitution process exhibits a single attractor

FEATURE VARIATION

$$\nu_\phi(f) = \langle \phi(w') \rangle_{\{w \to w' | \phi(w) = f\}}$$

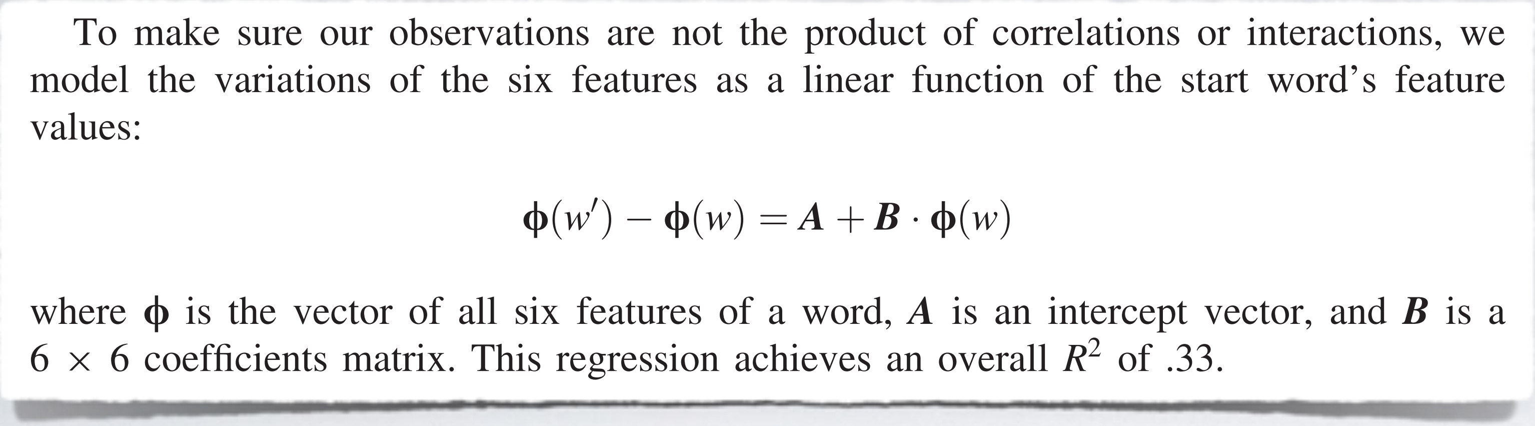


First, there is a single intersection of ν_φ with $y=x$ and the slope of ν_φ remains smaller than 1:
the substitution process exhibits a single attractor

Second, the comparison with ν^0_ϕ and ν^{00}_ϕ shows that there are two classes of attractors, depending on whether:

1. there is a triple intersection (of $y = x$, ν_ϕ, and ν^0_ϕ or ν^{00}_ϕ);
2. or ν_ϕ always remains above or below ν^0_ϕ and ν^{00}_ϕ.

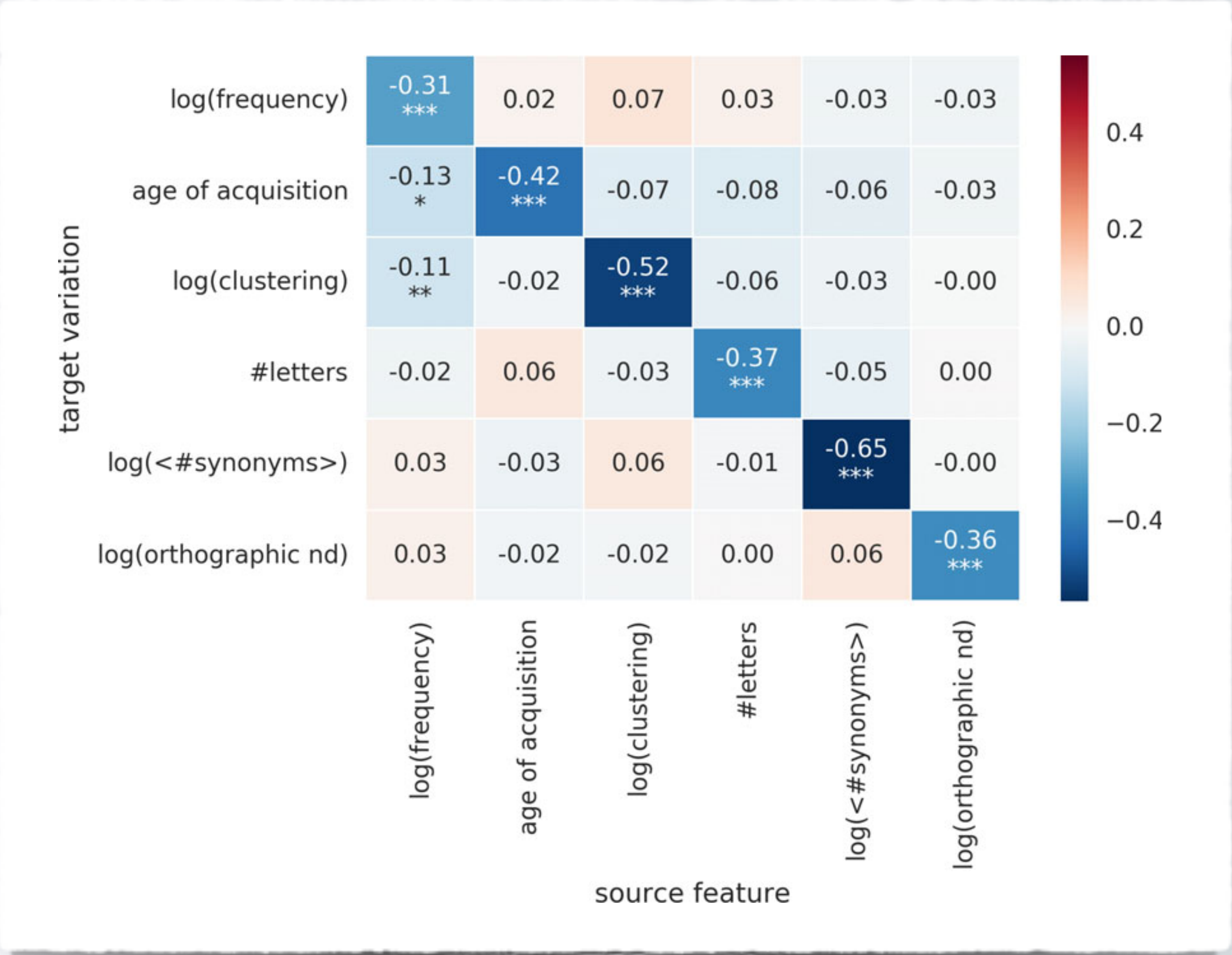
COMBINED EFFECTS



To make sure our observations are not the product of correlations or interactions, we model the variations of the six features as a linear function of the start word's feature values:

$$\phi(w') - \phi(w) = \boldsymbol{A} + \boldsymbol{B} \cdot \phi(w)$$

where ϕ is the vector of all six features of a word, $\boldsymbol{A}$ is an intercept vector, and $\boldsymbol{B}$ is a 6×6 coefficients matrix. This regression achieves an overall R^2 of .33.



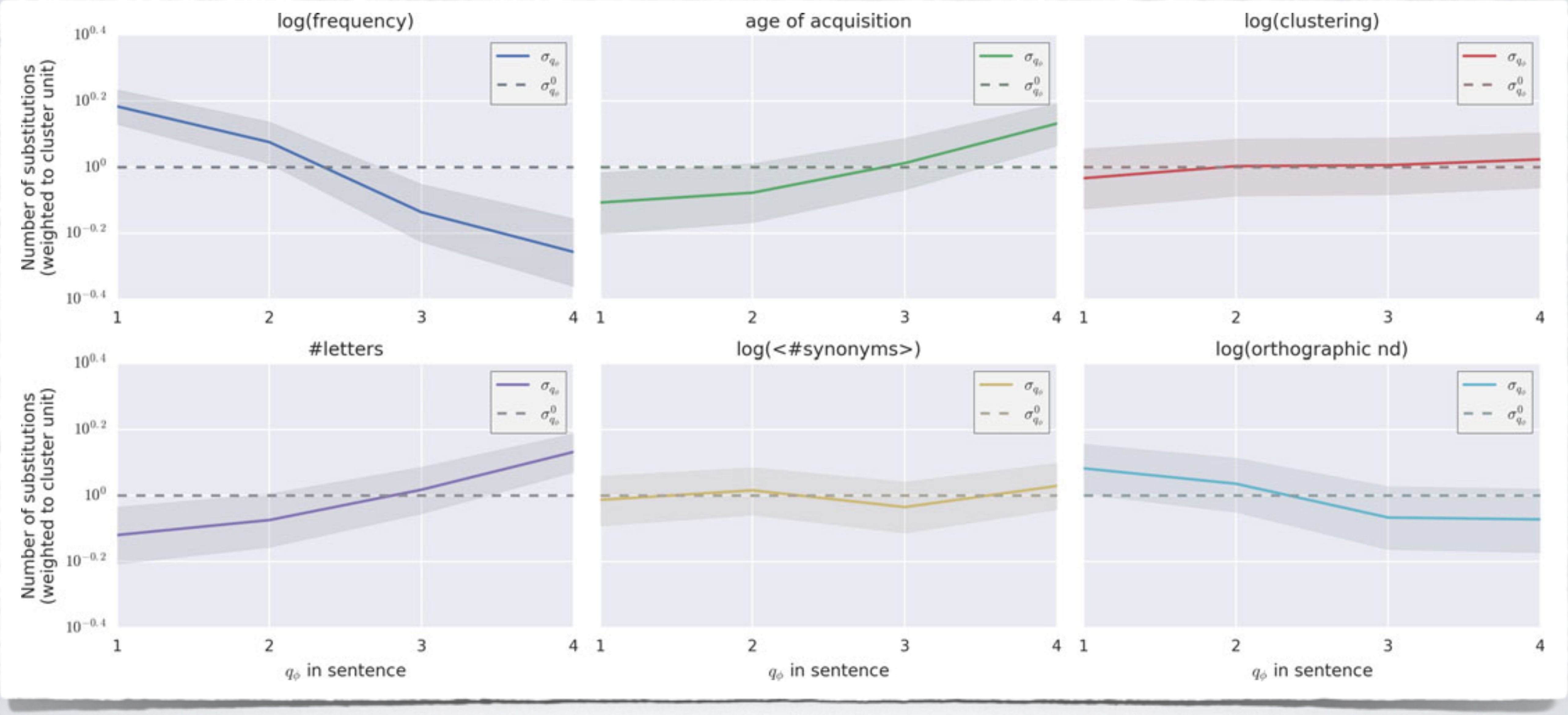
target variation \ source feature	log(frequency)	age of acquisition	log(clustering)	#letters	log(<#synonyms>)	log(orthographic nd)
log(frequency)	-0.31 ***	0.02	0.07	0.03	-0.03	-0.03
age of acquisition	-0.13 *	-0.42 ***	-0.07	-0.08	-0.06	-0.03
log(clustering)	-0.11 **	-0.02	-0.52 ***	-0.06	-0.03	-0.00
#letters	-0.02	0.06	-0.03	-0.37 ***	-0.05	0.00
log(<#synonyms>)	0.03	-0.03	0.06	-0.01	-0.65 ***	-0.00
log(orthographic nd)	0.03	-0.02	-0.02	0.00	0.06	-0.36 ***

- *Burmese poet Saw Wai (Nov 2008):*
  - "Senior general Than Shwe is foolish with power"
  - "Senior general Than Shwe is **crazy** with power"

"foolish": *8.94* y.o., *675* times, cc of *.0082*

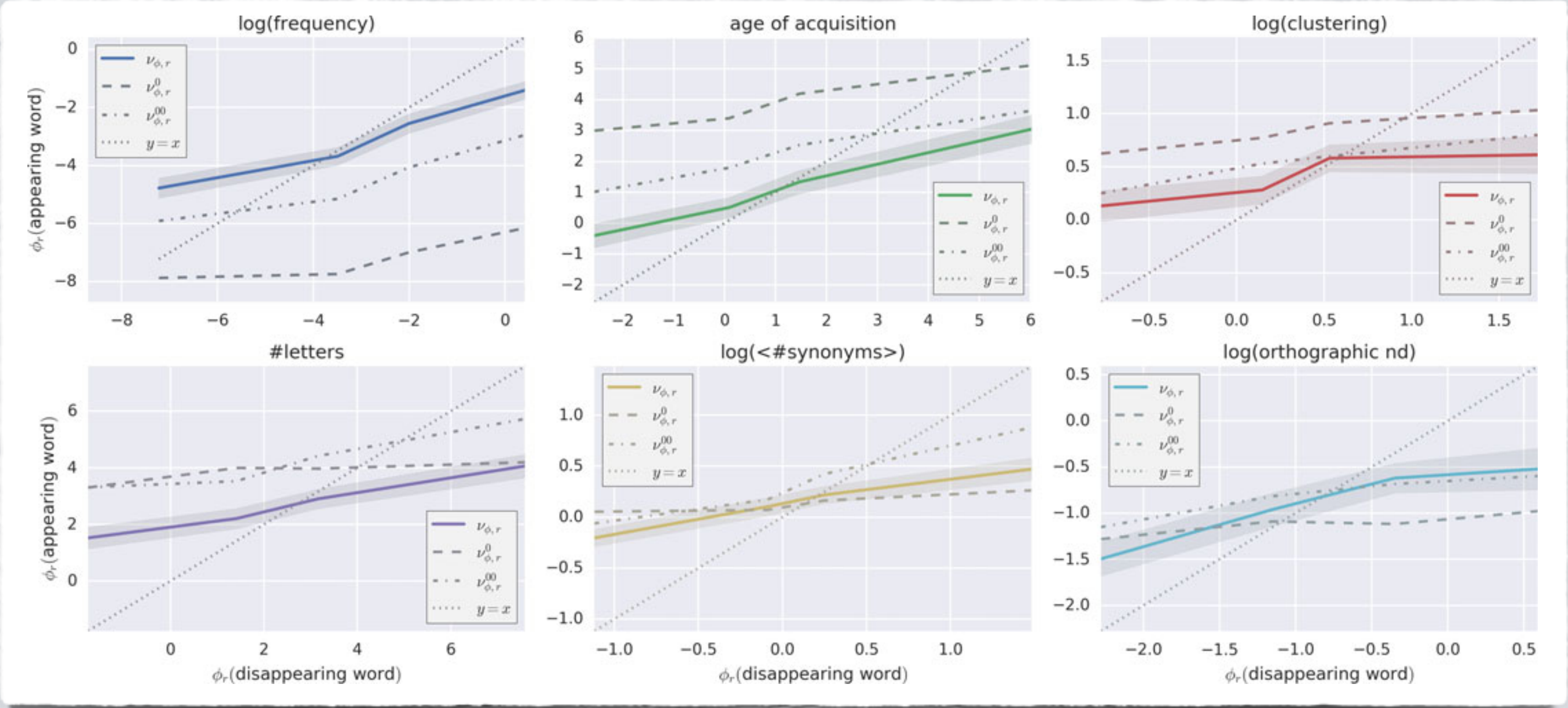
>"crazy": 5.22 y.o., *4100* times, cc of *.0017*

TAKING SENTENCE CONTEXT INTO ACCOUNT



susceptibility based on the position of the word in the sentence (quartiles)

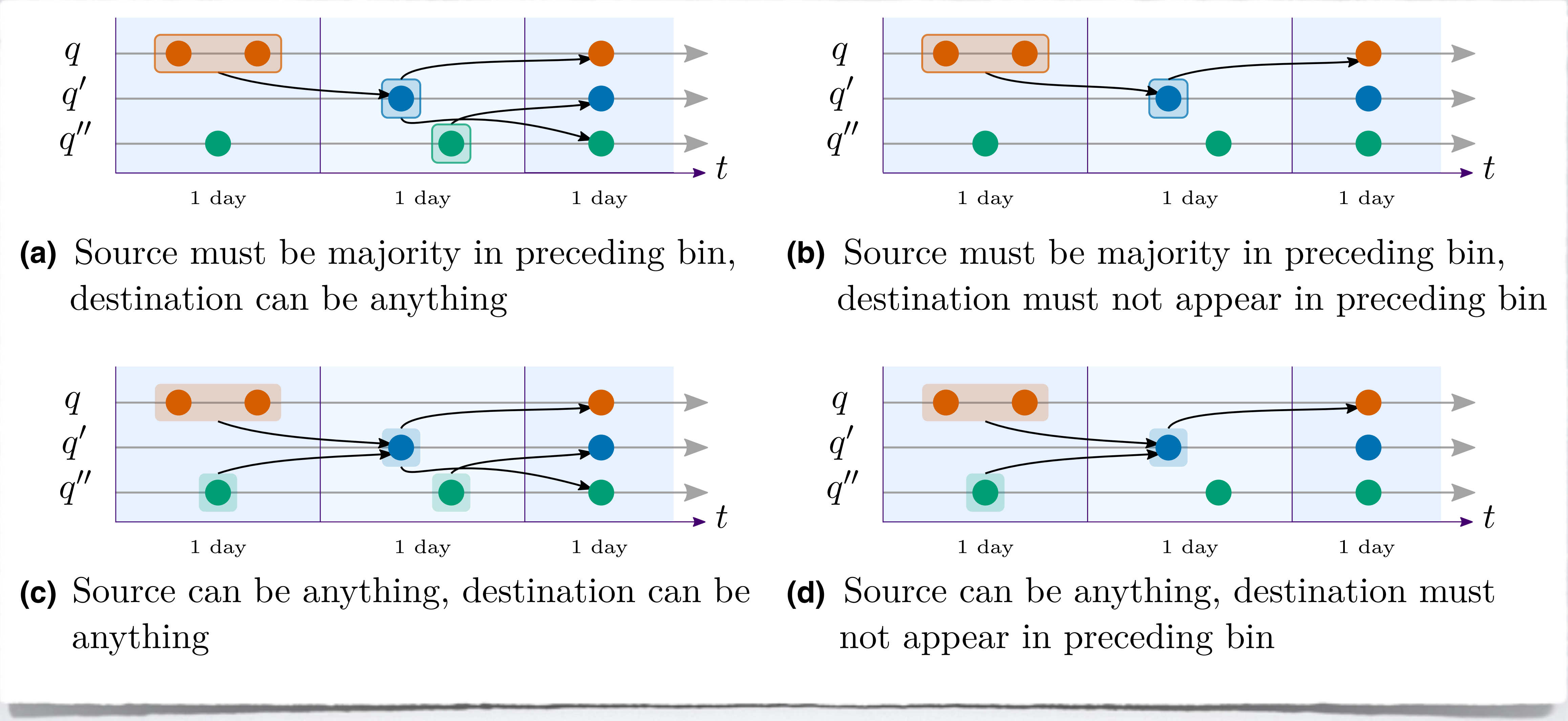
TAKING SENTENCE CONTEXT INTO ACCOUNT



feature variation w.r.t. median feature value in the sentence

The speaker says “Thanks” –> “Danke”

SUBSTITUTION MODEL VARIANTS



(a) Source must be majority in preceding bin, destination can be anything

(b) Source must be majority in preceding bin, destination must not appear in preceding bin

(c) Source can be anything, destination can be anything

(d) Source can be anything, destination must not appear in preceding bin

(i) bin position

(ii) bin length

(iii) candidate sources

(iv) candidate destinations