

Learning Computational Models of Non-Standard Language

Maria Ryskina

Carnegie Mellon University



Carnegie Mellon University
Language
Technologies
Institute



BERG LAB



LLab
Language Technologies Institute
Carnegie Mellon University

Non-standard language & NLP

From standardized language...

```
(S (NP-SBJ (NN Compound)
     (NNS yields))
  (VP (VBP assume)
    (UCP (NP (NP (NN reinvestment))
              (PP (IN of)
                  (NP (NNS dividends))))
    (CC and)
   (SBAR (IN that)
     (S (NP-SBJ (DT the)
           (JJ current)
           (NN yield))
      (VP (VBZ continues)
        (PP-TMP (IN for)
          (NP (DT a)
            (NN year))))))))
```

(. .)

)

e.g. PTB:

- Newswire
- Finance-related
- Formal

...To creative language

Twitter @Twitter

Happeh Fridai! U can nao change ur language 2 lolcat in Settings. Hope u liek it. twitter.com/settings

New **New** New York Times
@NYT_first_said

friendscape

5:03 AM · Jun 1, 2021

GLOSSARY OF NADSAT LANGUAGE

Words that do not appear to be of Russian origin are distinguished by asterisks. (For help with the Russian, I am indebted to the kindness of my colleague Nora Montesinos and a number of correspondents.)

male secondary *
"inactivated" (as if chorned). The of paired names lior with Kitty of his Practi mornings, on ration of beets, and a breath tation of seberries, and her

*appy polly loggy—apology
baboochka—old woman
*baddiwad—bad

cluve—beak
collcoll—bell
*crack—to break up or "burst"
*crak—to yell!
crast—to steal or rob; robbery

Linguistic innovation

- Non-standard, novel linguistic items...
 - Lexical: new word forms (*brony*)
 - Morphological: new morphemes (-gate) or derivatives (*prolifeness*)
 - Orthographic: non-standard spellings (*2nite*)
- ... before they become attested (*tweet*)
- People can infer their meaning, but NLP systems largely treat them as noise

Linguistic innovation

- **Q1: How do people process non-standard items?**
 - Shared knowledge or perception: $2 = \text{'two'} = /tu/$
 - Compositionality: $2nite = \text{'two'} + \text{'nite'} = /tu/ + /naɪt/ \approx /tənaɪt/$
 $\text{antivehicleness} = \text{'anti'} + \text{'vehicle'} + \text{'ness'}$
- **Q2: How can we get our NLP systems to that level?**
 - Text normalization: $2nite \rightarrow tonight$ (Baldwin et al., 2015: W-NUT shared task)
 - Improving robustness to noise & ‘noise’ (Li et al., 2019: WMT shared task)
 - Maybe we can encode creative reasoning into them?

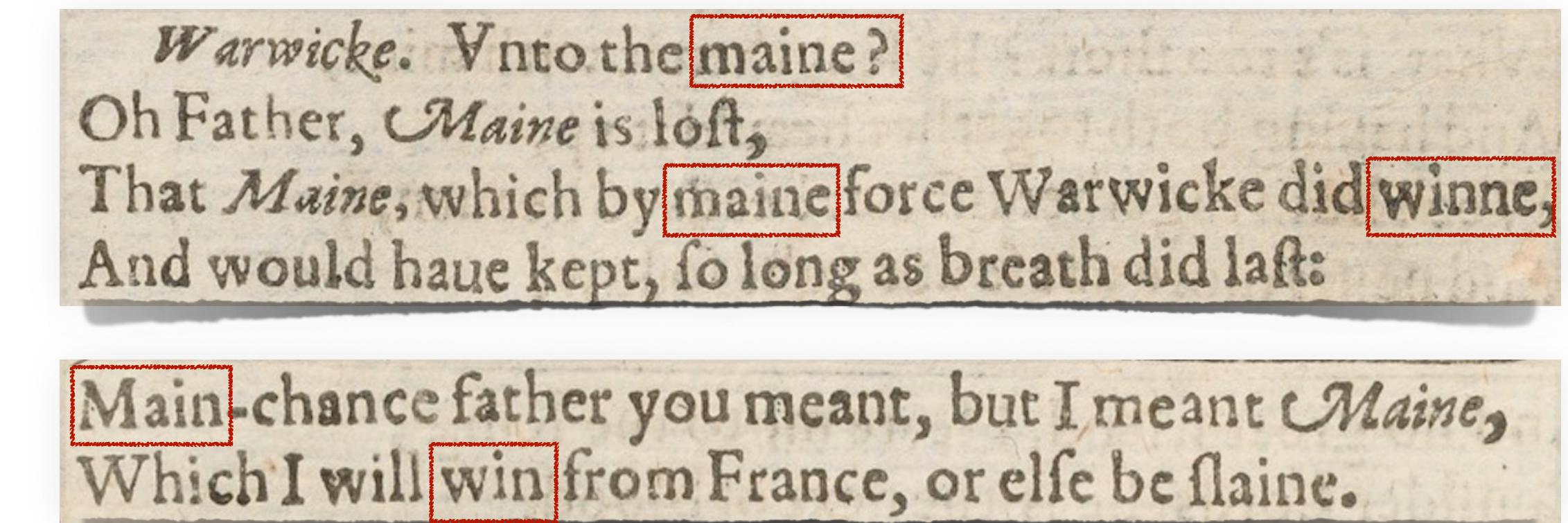
Non-standard(ized) orthographies

- Not having (or not following) a prescribed way of spelling

WARWICK

Unto the **main**! O father, Maine is lost;
That Maine which by **main** force Warwick did **win**,
And would have kept so long as breath did last!
Main chance, father, you meant; but I meant Maine,
Which I will **win** from France, or else be slain,

Shakespeare, Henry VI, Part 2



Warwicke. Vnto the **maine?**
Oh Father, **Maine** is lost,
That **Maine**, which by **maine** force Warwicke did **winne**,
And would haue kept, so long as breath did last:

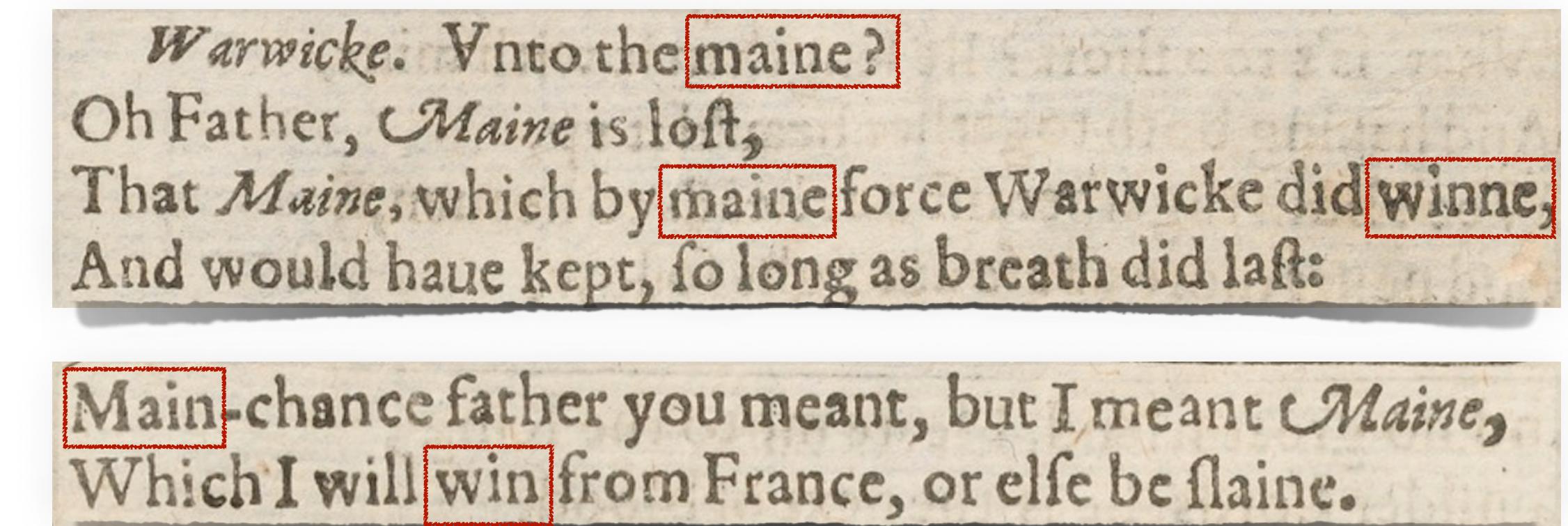
Main chance father you meant, but I meant **Maine**,
Which I will **win** from France, or else be slaine.

Non-standard(ized) orthographies

- Not having (or not following) a prescribed way of spelling
- **Idiosyncratic:** people make different spelling choices
- **Shared:** spellings reflect underlying pronunciation or ‘standard’ spelling

WARWICK

Unto the **main!** O father, Maine is lost;
That Maine which by **main** force Warwick did **win**,
And would have kept so long as breath did last!
Main chance, father, you meant; but I meant Maine,
Which I will **win** from France, or else be slain,



Shakespeare, Henry VI, Part 2

Non-standard(ized) orthographies

- Not having (or not following) a prescribed way of spelling
- **Idiosyncratic:** people make different spelling choices
- **Shared:** spellings reflect underlying pronunciation or ‘standard’ spelling

mariamC
@mariamelmalt

Sba7 el 5eir good luck in your exams<3

Amina Ziad
@Amina_Z

sba7 el kheir friends

8:45 AM · May 9, 2019

De dochter van Y.
@Smeerjamming

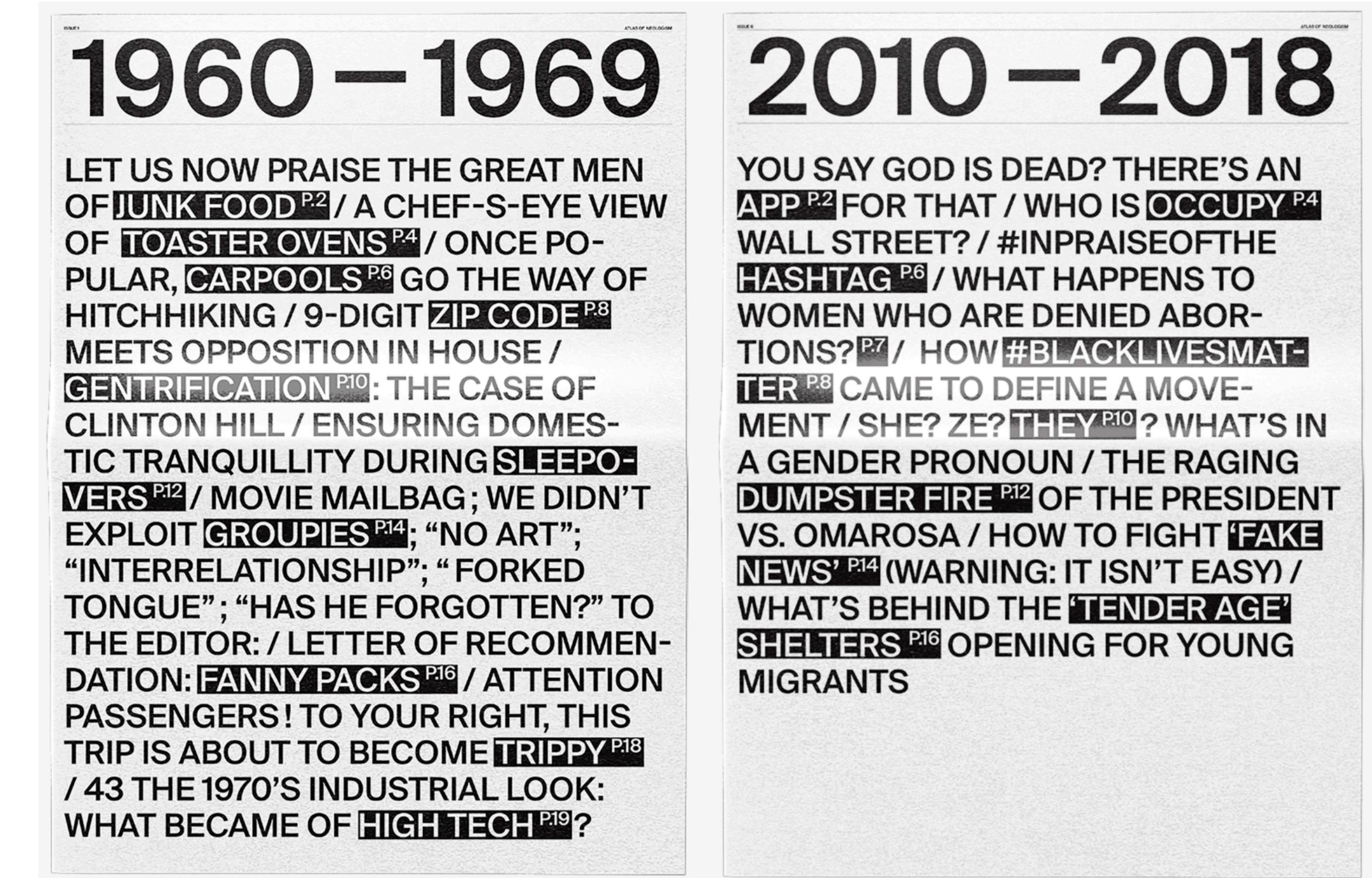
Sbah el khair lovely peeps!

1:18 AM · Jun 17, 2014

صباح الخير
'Good morning'

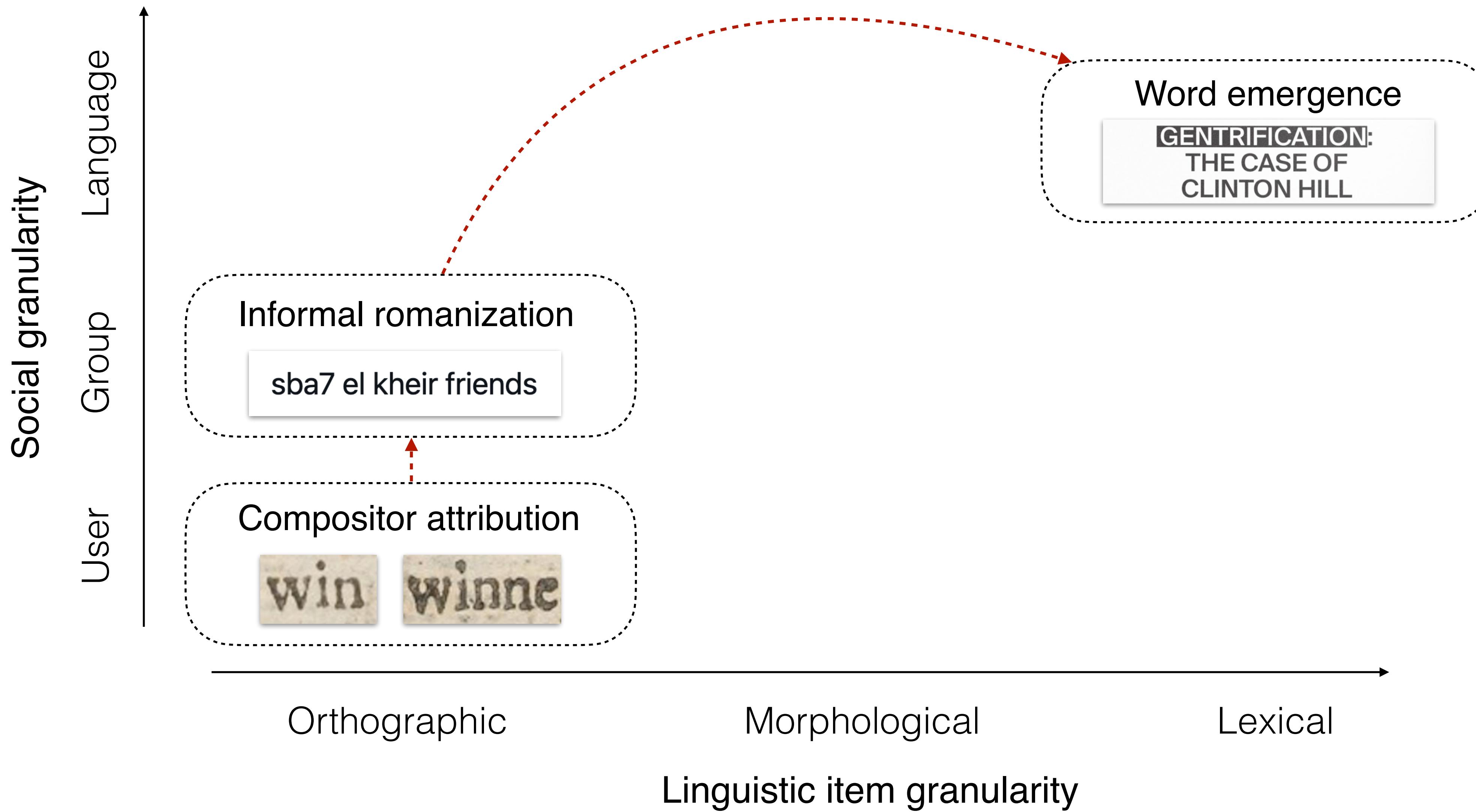
Neology

- Coining of novel words
- Idiosyncratic: words introduced in individual creative acts
- Shared: to survive and spread, words must match the needs of the linguistic community

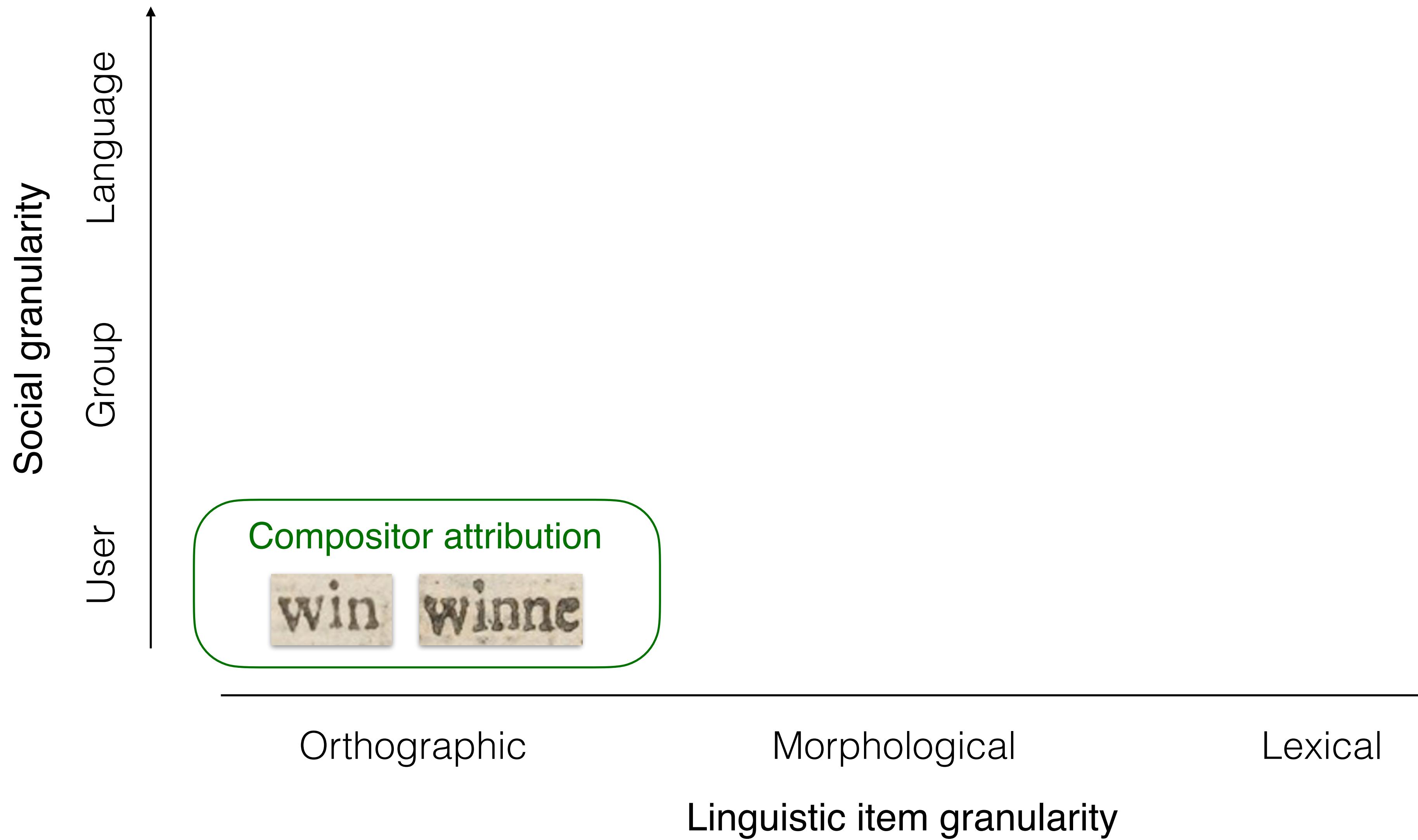


Irini Gleglakou: Atlas of Neologism
[https://www.behance.net/gallery/80427631/Atlas-of-Neologism-\(newspapers\)](https://www.behance.net/gallery/80427631/Atlas-of-Neologism-(newspapers))

Spectrum of phenomena



Spectrum of phenomena



Early Modern English

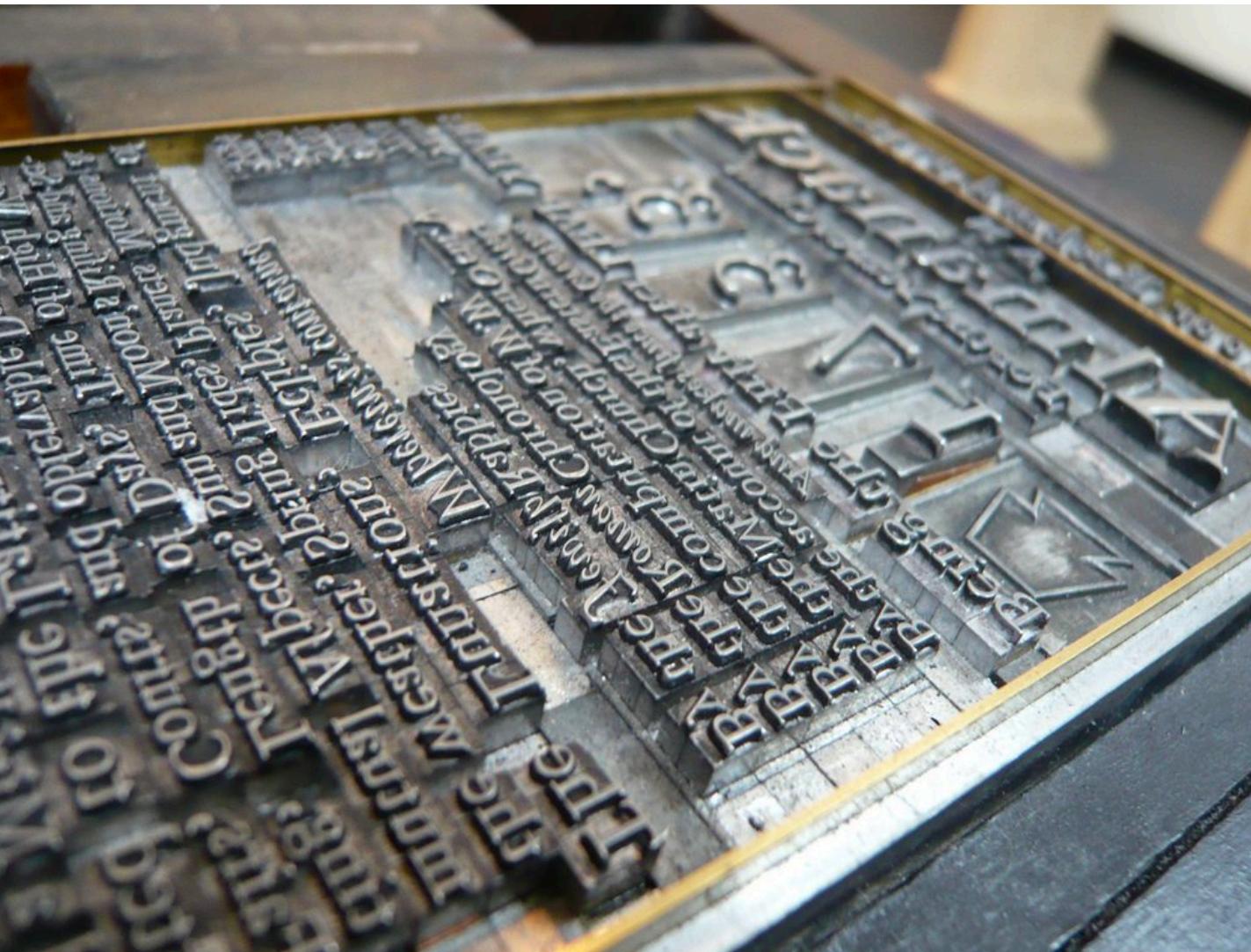
- Word spellings differed from modern ones (*dye, sleepe, naturall*)
- Many words had more than one accepted spelling (*heart, hart, harte*)
- Spelling variant choices could be used to distinguish people

Enter Hamlet.

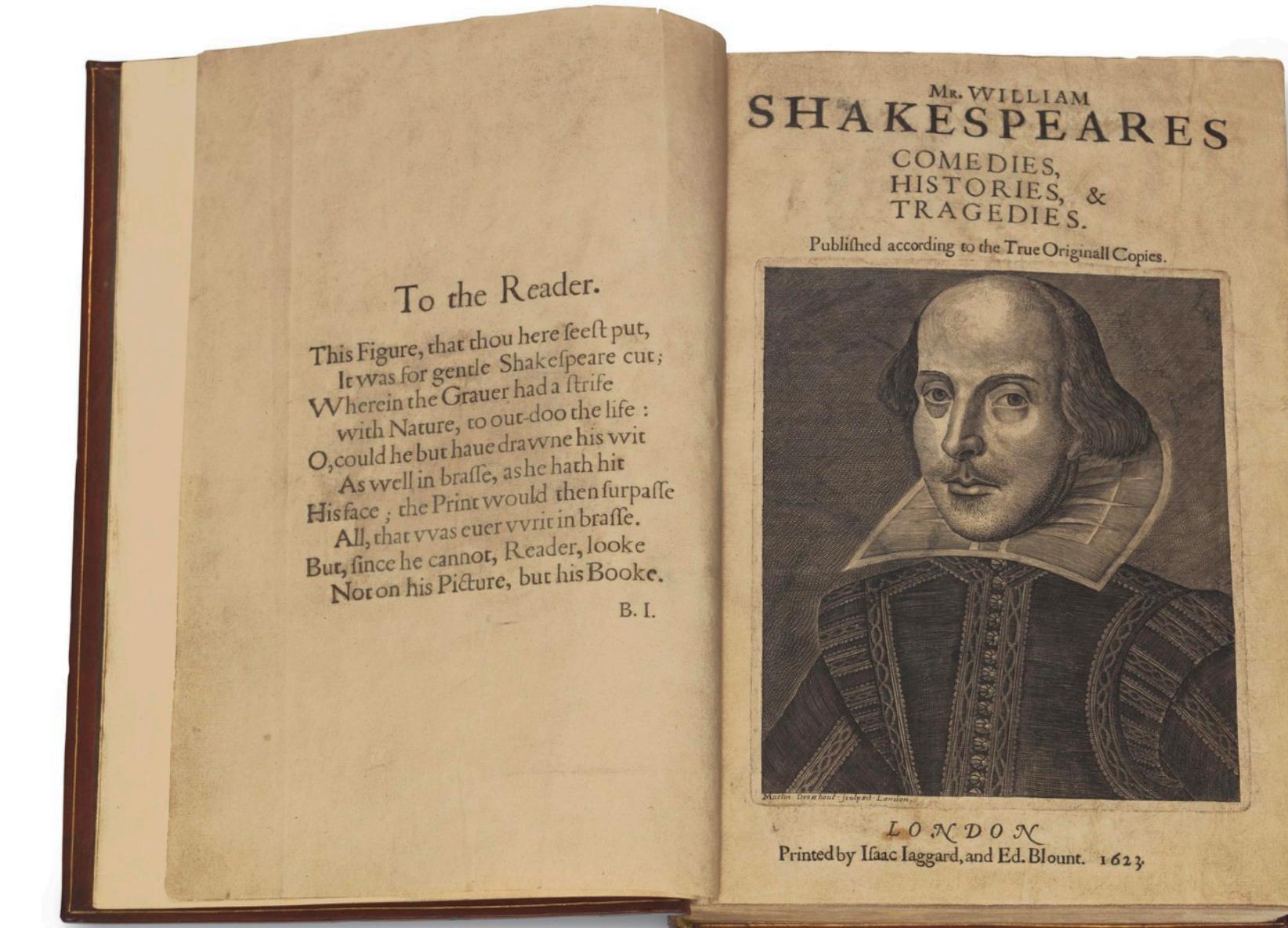
Ham. To be, or not to be, that is the Question :
Whether 'tis Nobler in the mind to suffer
The Slings and Arrowes of outrageous Fortune,
Or to take Armes against a Sea of troubles,
And by opposing end them . to dye, to sleepe
No more ; and by a sleepe, to say we end
The Heart-ake, and the thousand Naturall shockes

Historical printed books

- *Compositors* manually arranged letters on the printing press
 - Different pages in the same book were set by different people
 - Each typesetter has their own spelling preferences
- *Composer attribution*: grouping pages set by the same typesetter



<https://www.flickr.com/photos/purdman1/2875431305/>



<https://www.smithsonianmag.com/smart-news/shakespeares-first-folio-sells-ten-million-dollars-180976074/>

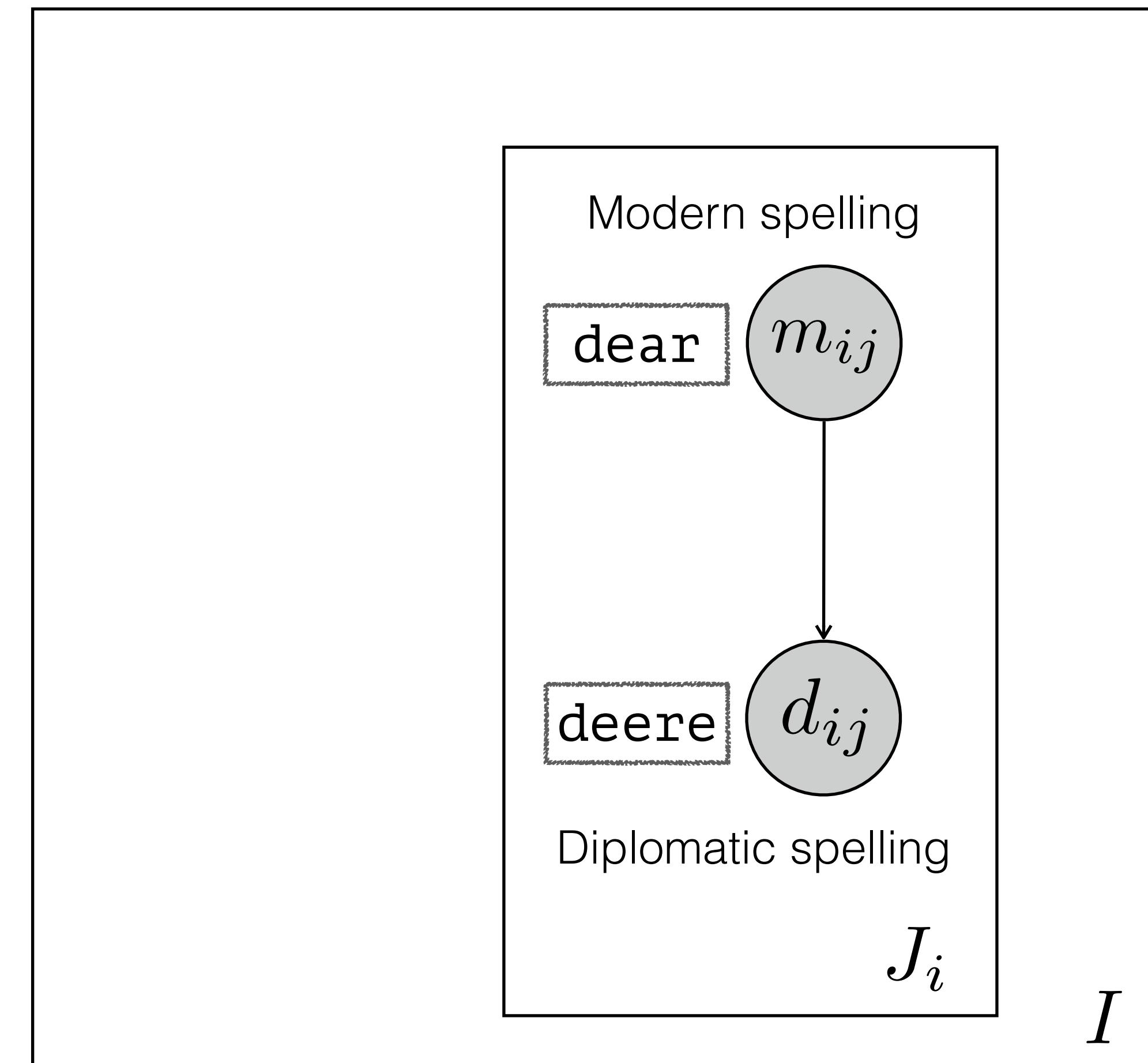
Compositor attribution

- Individual compositors have spelling variant preferences
- Individual typecases have different comma space lengths

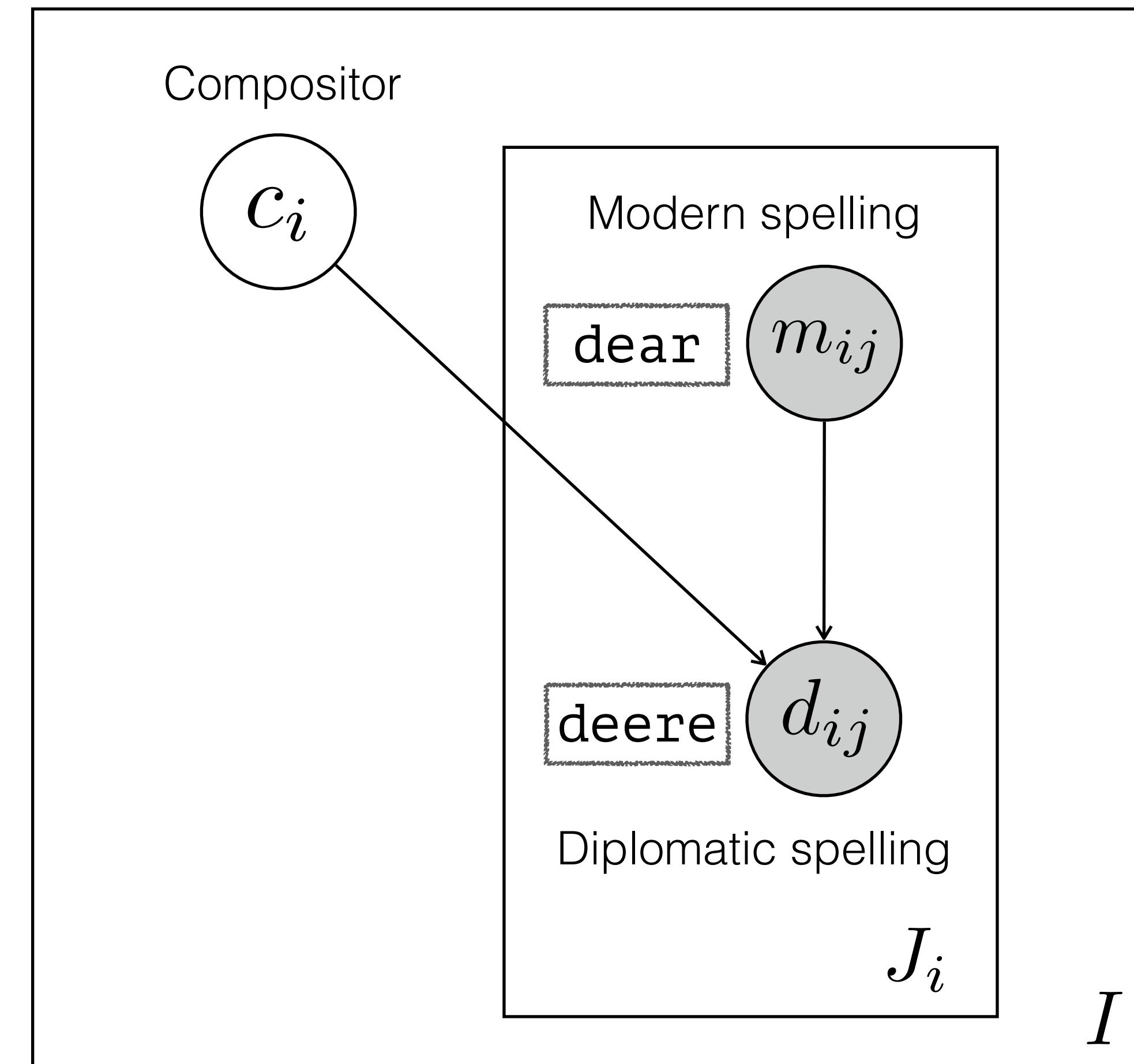
call it, being
Thy sent I can it, being mangle to me:
That vndiuidable, I am
Am better then th~~e~~ decre better part.
All teare away, ... gone from me;
For doe by loue : as easie maist thou fall

take me? if not? to lay to thee that I sh~~e~~ dye, is but
for thy loue, ~~but~~ No: yet I loue thee too. And
while thou liu~~e~~, take a fellowe of plaine and
vicoyned Cor~~e~~, he perforce mu~~d~~ see right,
because he hath not the gis~~t~~ to woee in d~~o~~aces: for

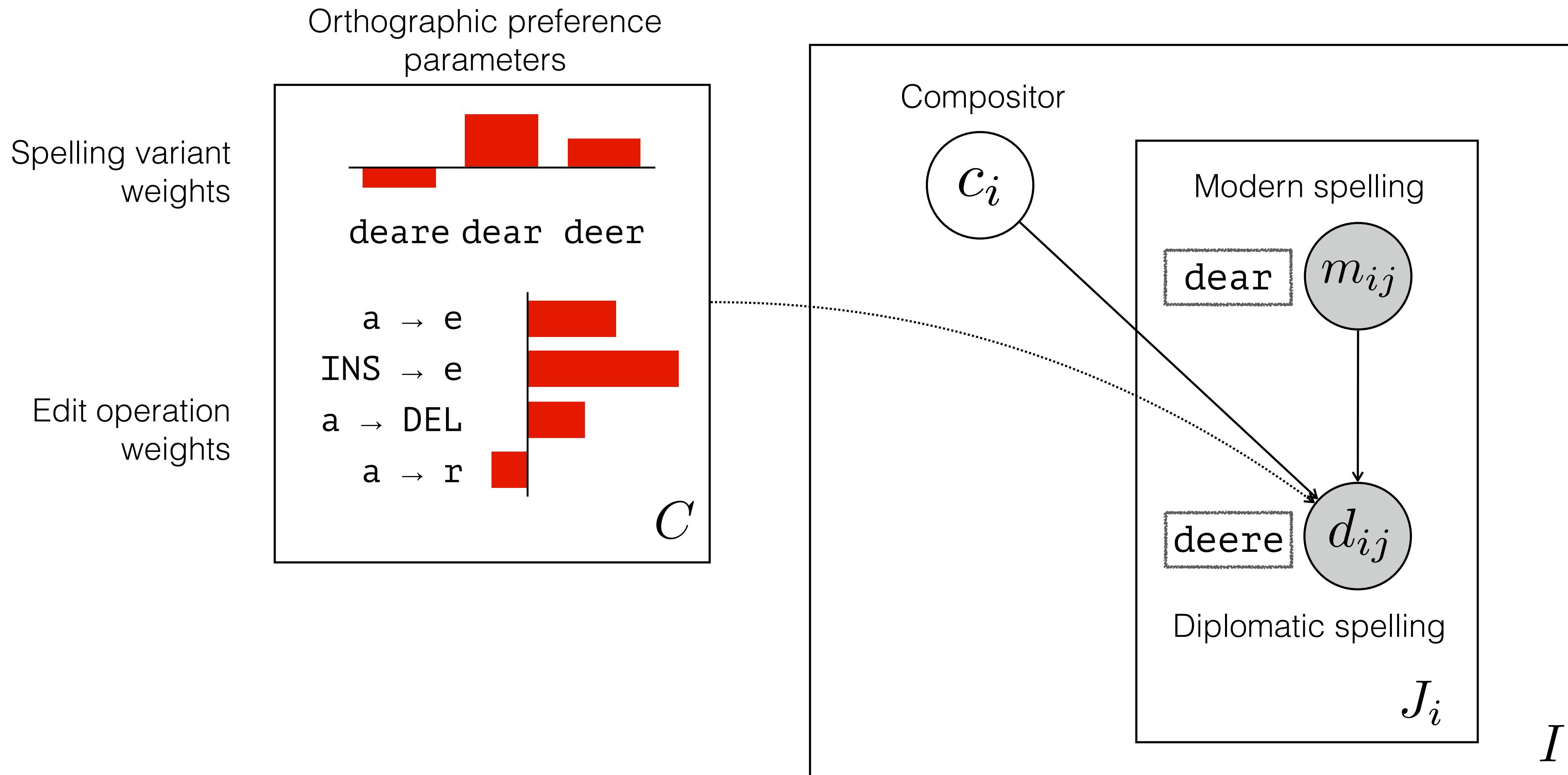
Compositor attribution model



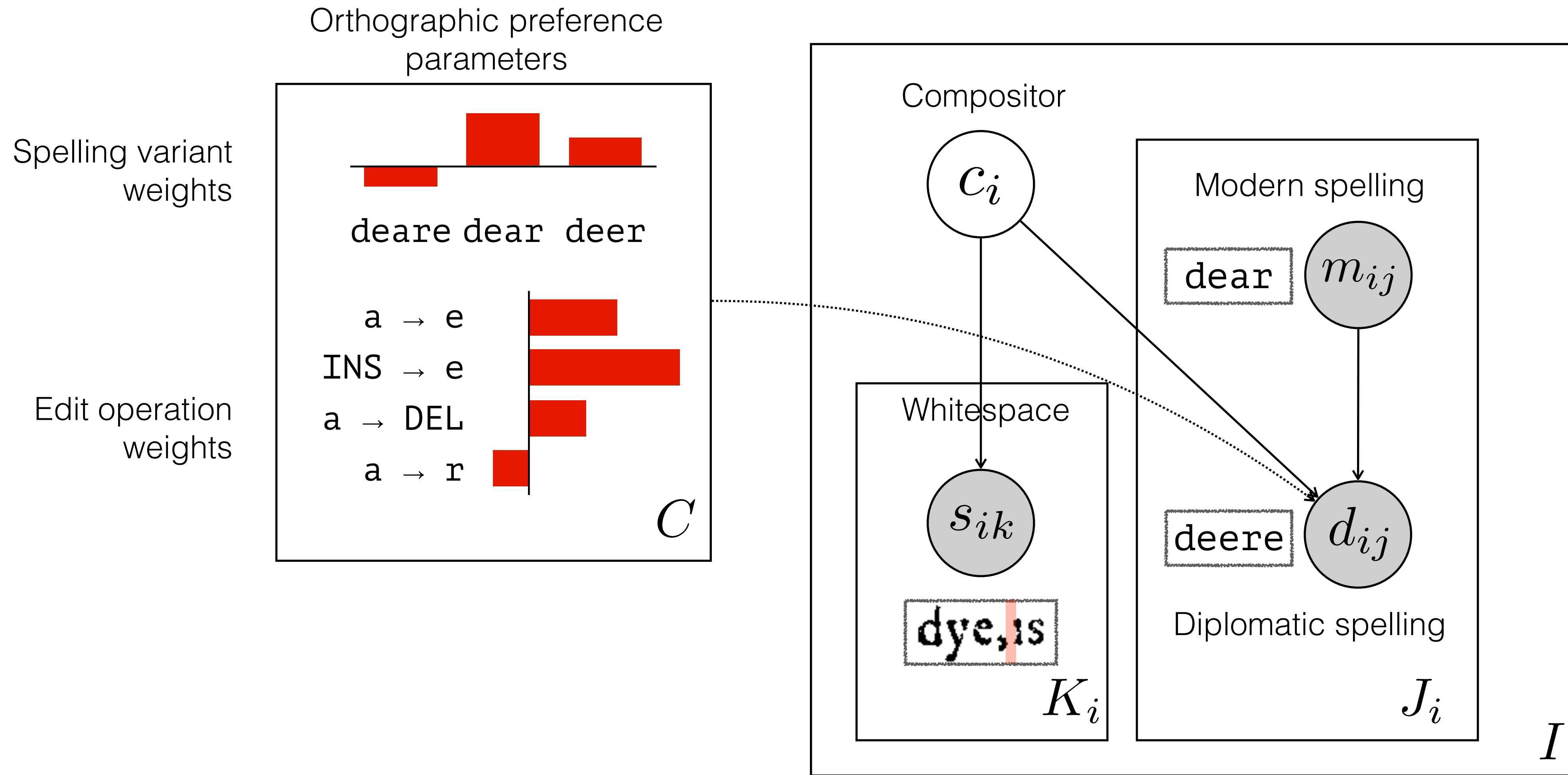
Compositor attribution model



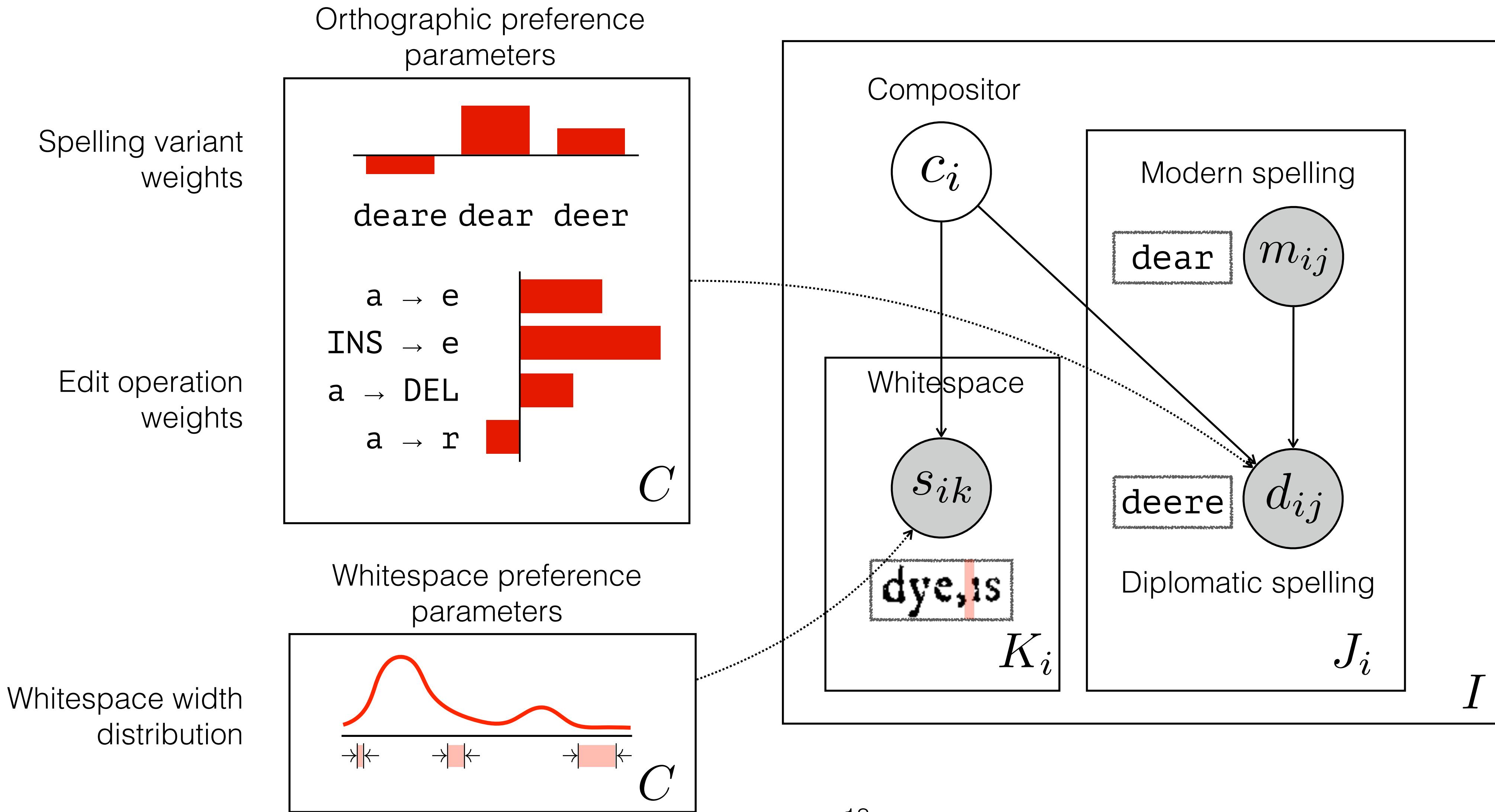
Compositor attribution model



Compositor attribution model



Compositor attribution model

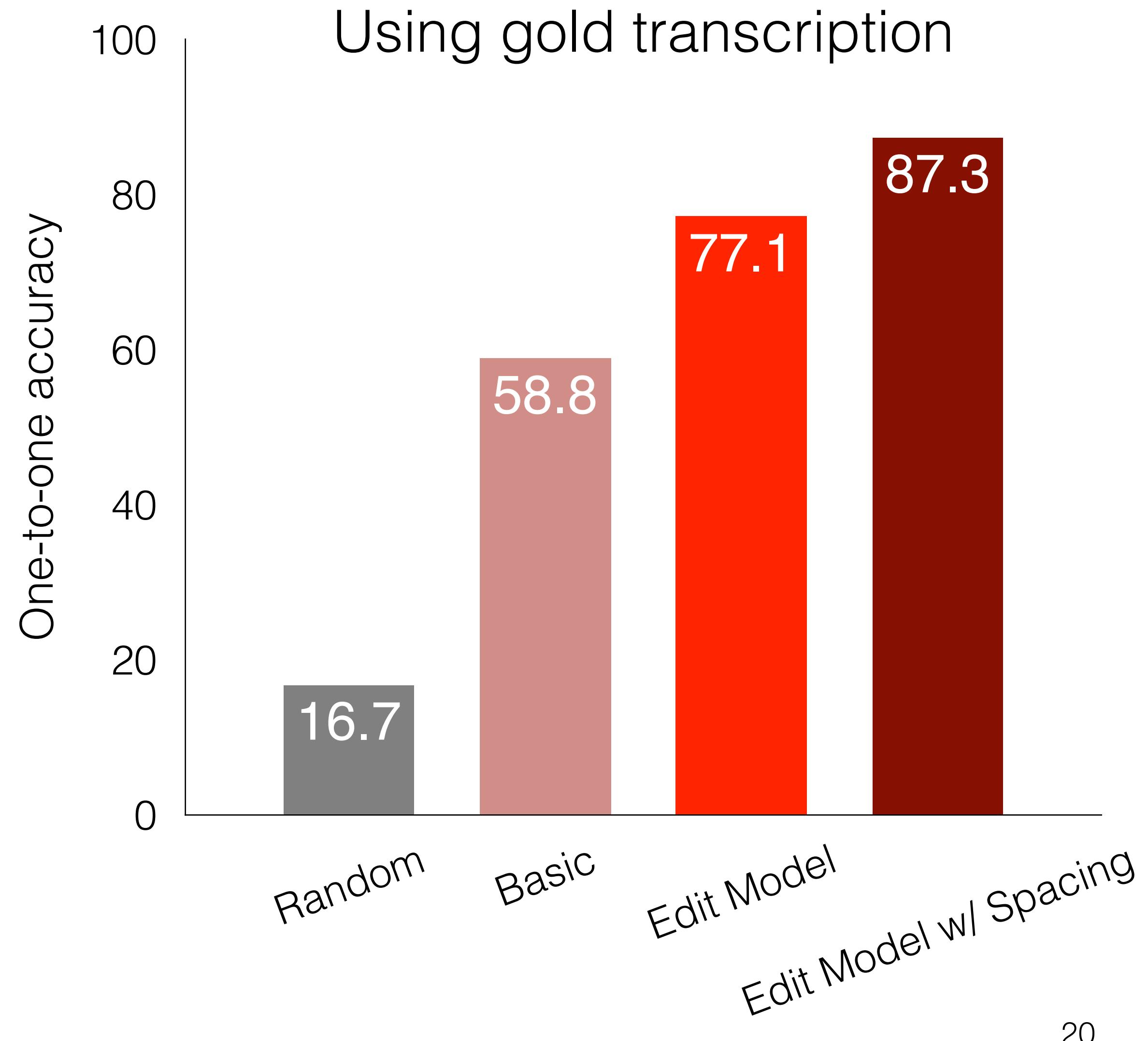


Training and evaluation

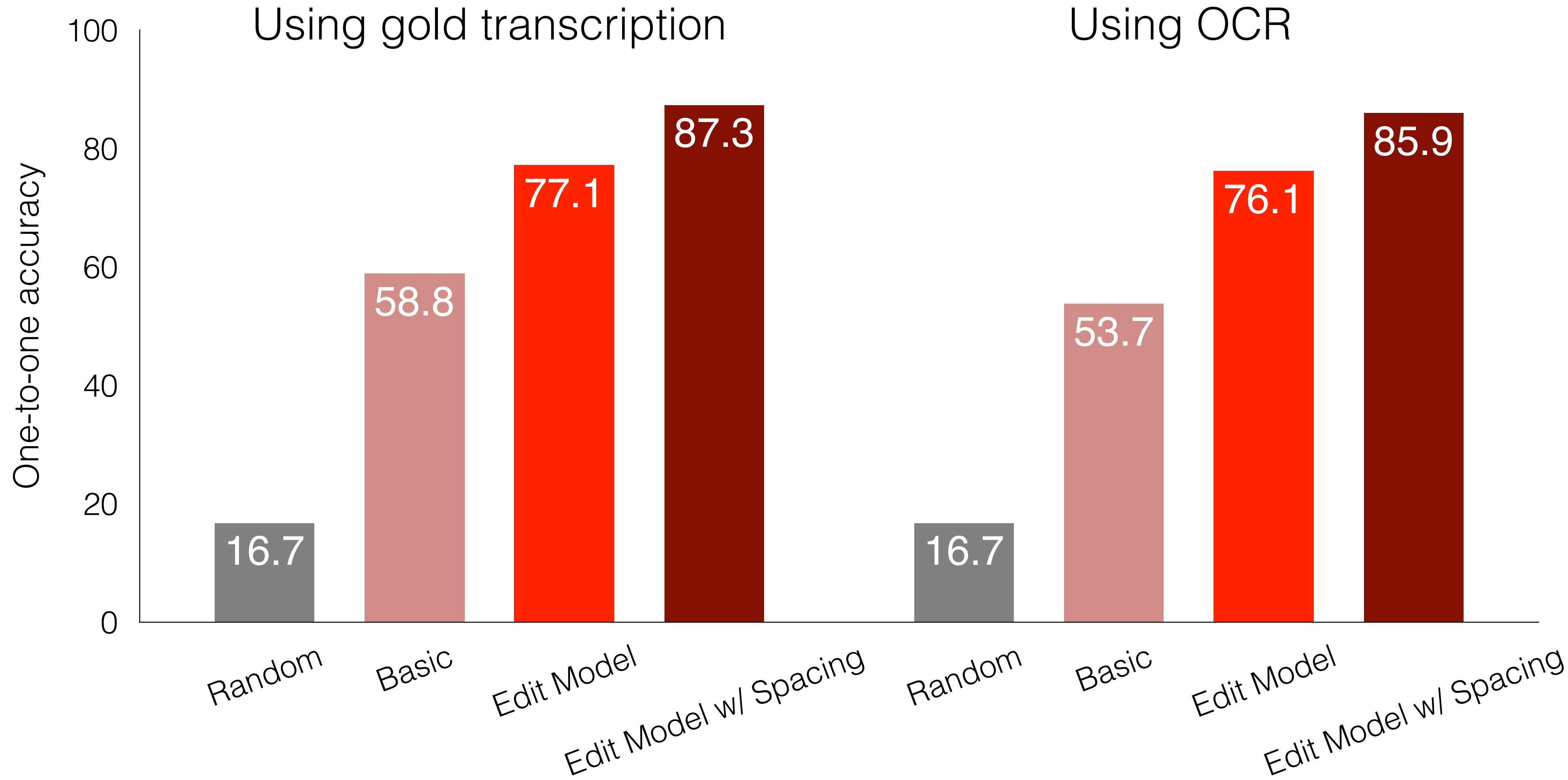
- We learn compositor preferences with EM algorithm
 - Setting the maximum number of composers in advance
 - Inference performed by independent argmax for each page
- Evaluate by comparing against the authoritative attribution
 - Match recovered clusters to ground-truth ones using Hungarian algorithm
 - Measure one-to-one and many-to-one accuracy

M Ryskina, H Alpert-Abrams, D Garrette, T Berg-Kirkpatrick. Automatic Compositor Attribution in the First Folio of Shakespeare. ACL 2017.

Experimental results

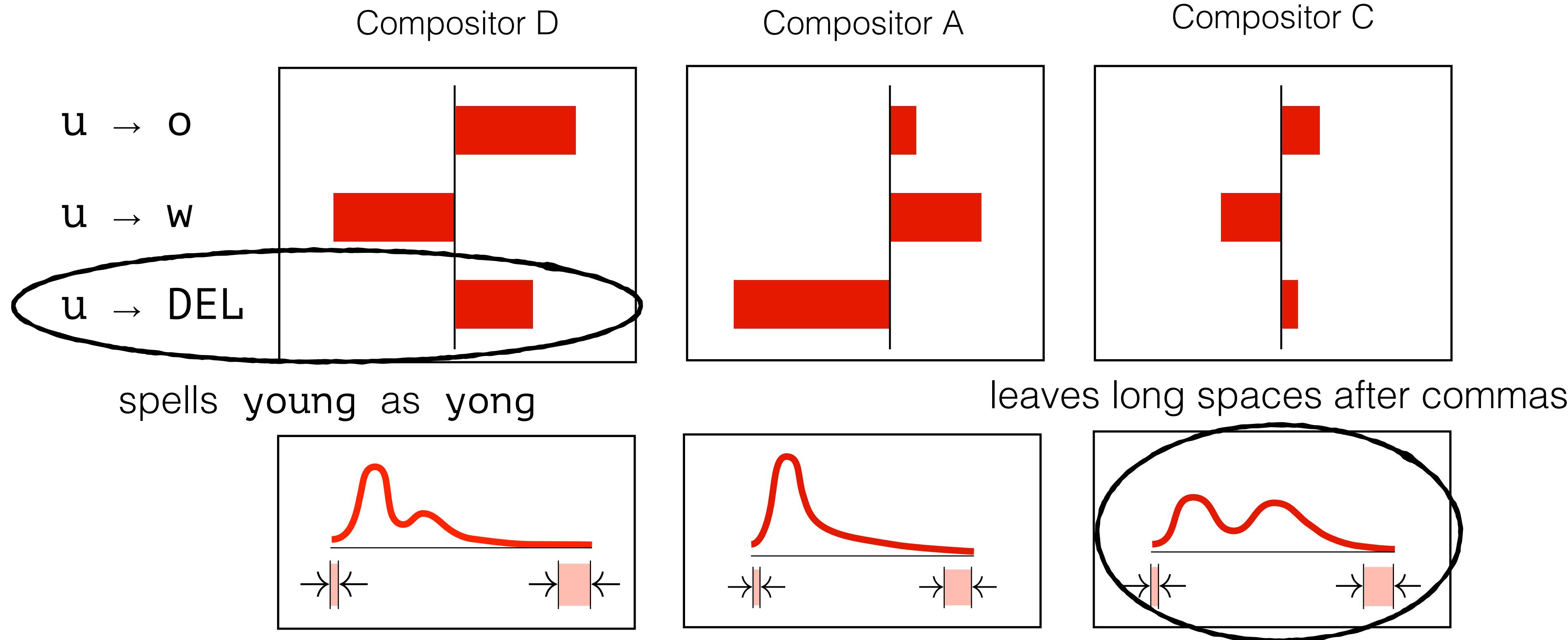


Experimental results



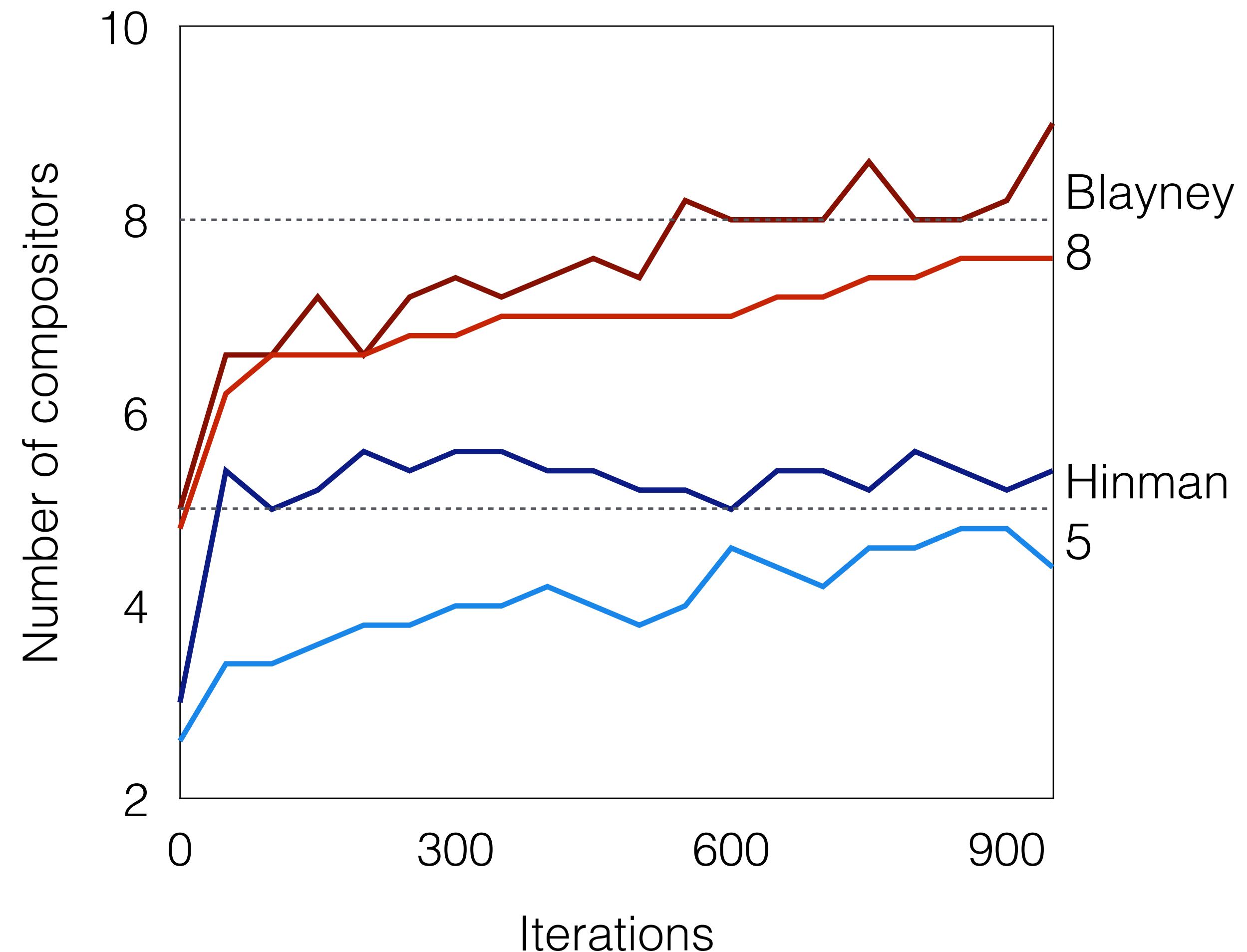
Learned behaviors

- Patterns discovered by our model match the scholars' observations!

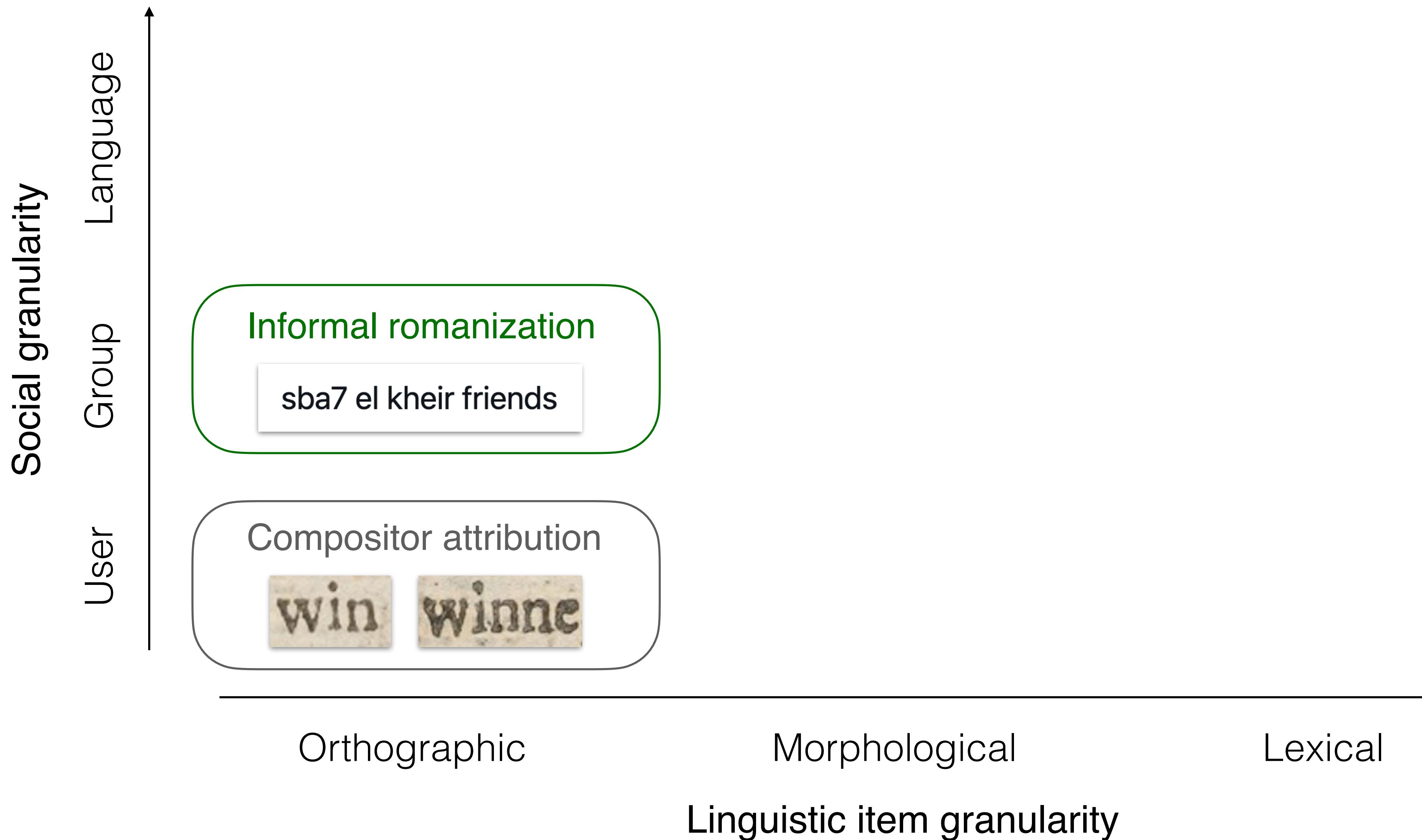


Number of composers

- Bibliographers posited different numbers of composers
 - Analyses based on different word lists
- We extend our model to non-parametric clustering
 - Features restricted to the ones used by each scholar
 - Predictions agree with the corresponding scholar's judgment!



Spectrum of phenomena



Informal romanization

- *Romanization*: rendering non-Latin-script languages in Latin alphabet
- *Informal*: used online, arises out of Unicode/keyboard issues

Russian	человек	<i>chelovek, 4elovek, ceJloBek, ...</i>
Arabic	صباح	<i>saba7, sba7, sabah, ...</i>
Greek	ξένος	<i>xenos, ksenos, 3enos, ...</i>

Informal romanization

- Idiosyncratic representation: character substitutions up to the user

Russian	человек	<i>chelovek, 4elovek, ceJloBek, ...</i>
Arabic	صباح	<i>saba7, sba7, sabah, ...</i>
Greek	ξένος	<i>xenos, ksenos, 3enos, ...</i>

Informal romanization

- Idiosyncratic representation: character substitutions up to the user
- Most substitutions are based on **phonetic** or **visual** similarity

Russian	человек	<i>chelovek, 4elovek, ceJloBek, ...</i>
Arabic	صباح	<i>saba7, sba7, sabah, ...</i>
Greek	ξένος	<i>xenos, ksenos, 3enos, ...</i>

Character similarity

- Phonetic: similarity between sounds associated with characters
 - Out-of-context grapheme-phoneme association: $\Gamma \sim /g/ \rightarrow g$
 - Phoneme produced in context: انتي /enti/ \rightarrow enty, صباح /sabaħ/ \rightarrow saba7
- Visual: similarity between glyph shapes
 - Characters expressed by same or similar glyph: $a \sim /a/ \rightarrow a$, $\Gamma \sim /g/ \rightarrow r$
 - Characters can map to bi-/trigraphs: ы \rightarrow bl
 - Can be conditioned on a transformation: ل \rightarrow v, ز \leftarrow ع
 - Can be applied to a part of a glyph: ۲ \leftarrow ۱

Character alignment

- Monotonic alignment that depends on the writing system of the language

Alphabet

хорошо

|||||

xorosho

~ one-to-one

Abjad
(consonantal)

كريم

krym

/|\\|

kareem

~ one-to-one + null

Abugida
(alphasyllabary)

బెలగితు

/\\|\\|

belagitu

~ one-to-many

Character alignment

- Monotonic alignment that depends on the writing system of the language

Alphabet

хорошо

|||||

xorosho

~ one-to-one

Abjad
(consonantal)

كريم

krym

/|\\|

kareem

~ one-to-one + null

Abugida
(alphasyllabary)

బెలగితు

Unicode: బ ల గ త ట ఱ

\|\|/\|/\|/\|

belagitu

~ one-to-one + one-to-many

Task framing

- Convert romanized text to the conventional orthography of the language

Russian

конгресс не одобрил бюджет



kongress ne odobril biudjet

Egyptian
Arabic

انا حأعدك على 8 كده



ana h3dyy 3lek bokra 3la 8 kda

latent
(what they meant)

observed
(what they typed)

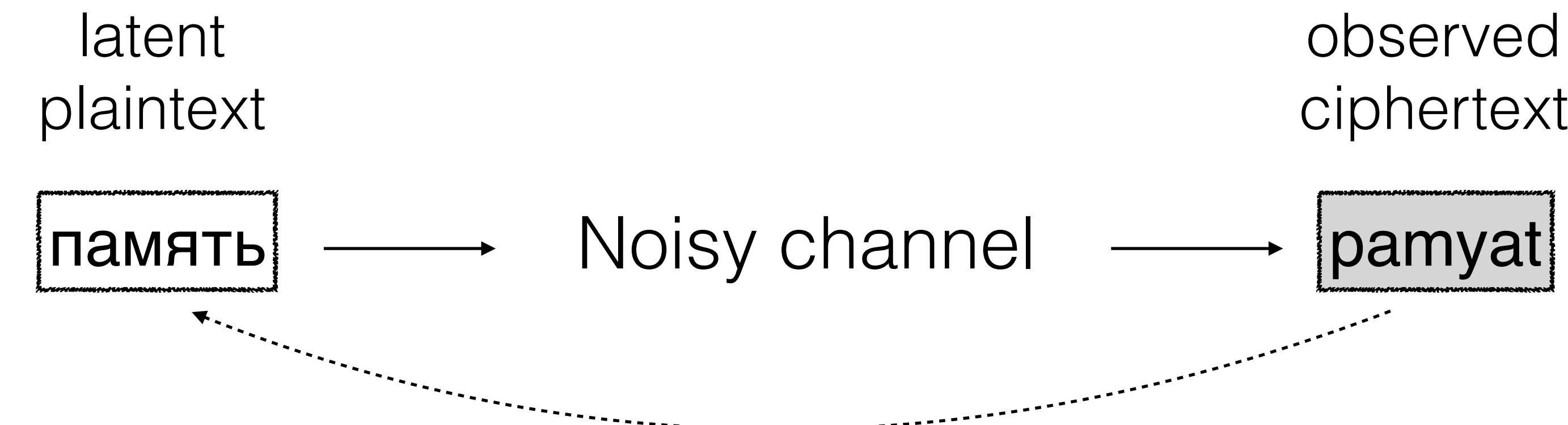
Task framing

- Parallel data does not occur naturally ⇒ **unsupervised** learning
- Perceptions of similarity are shared across users and even languages!
- **Hypothesis:** **inductive bias** encoding these similarity notions provides signal that can somewhat **approximate human supervision**
 - We rely on **manually-curated resources** to operationalize it

M Ryskina, MR Gormley, T Berg-Kirkpatrick. Phonetic and Visual Priors for Decipherment of Informal Romanization. ACL 2020.

Decipherment

- Can be viewed as a decipherment task (Knight et al., 2006)



Noisy-channel model

latent $n = \text{п а м я т ъ}$

observed $r = \text{p а m y a t}$

$$p(r) = \sum p(n; \gamma) \cdot p(r|n; \theta) \cdot p_{\text{prior}}(\theta; \alpha)$$

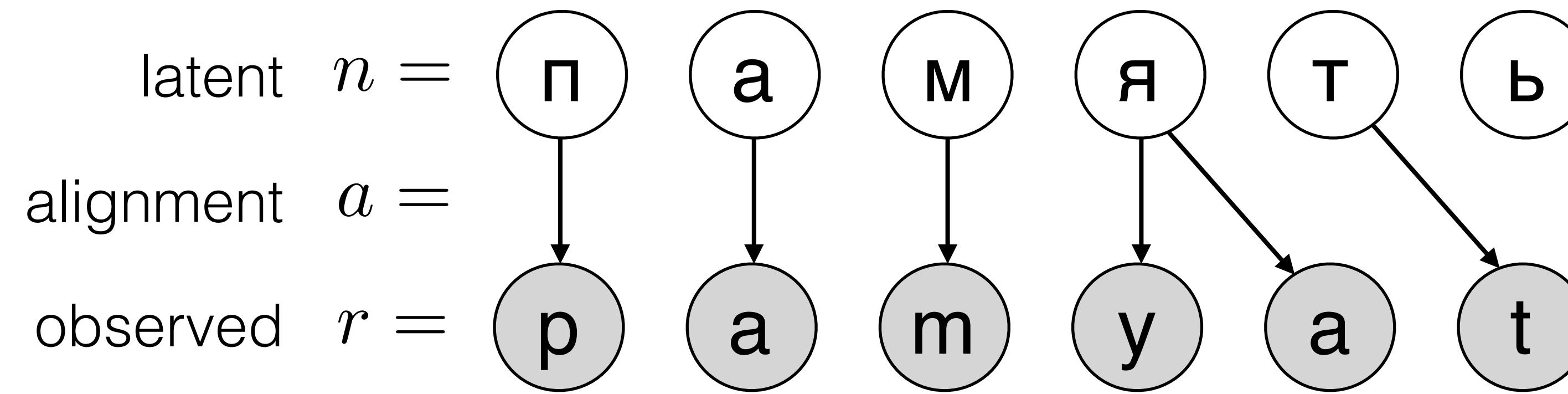
all possible
native script
sequences

n
transition probabilities

emission probabilities

θ
prior on parameters

Noisy-channel model



$$p(r) = \sum_{n,a} p(n; \gamma) \cdot p(r|n, a; \theta) \cdot p_{\text{prior}}(\theta; \alpha)$$

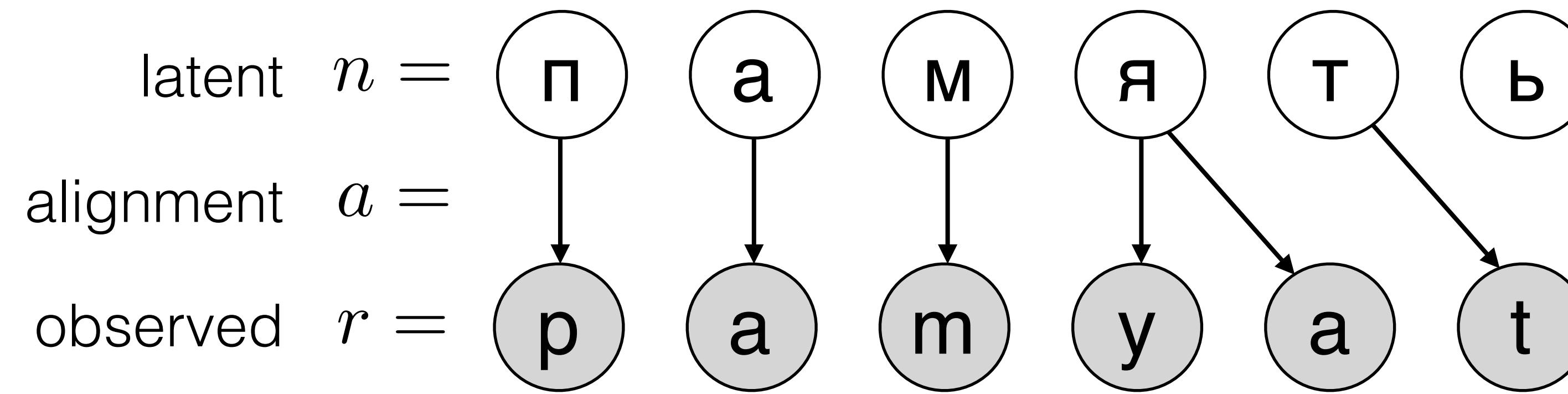
all possible
native script
sequences and
alignments

n, a
transition probabilities

emission probabilities

prior on parameters

Noisy-channel model



$$p(r) = \sum_{n,a} p(n; \gamma) \cdot p(r|n, a; \theta) \cdot p_{\text{prior}}(\theta; \alpha)$$

/ |
transition probabilities emission probabilities
prior on parameters

Phonetic bias

- ‘Phonetic’ priors: mappings off **phonetic keyboard layouts**



Phonetic bias

- ‘Phonetic’ priors: mappings off **phonetic keyboard layouts**
 - One-to-one mapping constraints lead to spurious mappings



https://twitter.com/arabic_bad/



Visual bias

- ‘Visual’ priors: mappings off the **Unicode confusables list**
 - Designed to combat spoofing attacks

o	o	σ	o	o	o	ດ	ດ
006F LATIN SMALL LETTER O	03BF GREEK SMALL LETTER OMICRON	03C3 GREEK SMALL LETTER SIGMA	043E CYRILLIC SMALL LETTER O	0585 ARMENIAN SMALL LETTER OH	05E1 HEBREW LETTER SAMEKH		0647 ARABIC LETTER HEH
c	c	c	c	c	c	c	c
0063 LATIN SMALL LETTER C	03F2 GREEK LUNATE SIGMA SYMBOL	0441 CYRILLIC SMALL LETTER ES	1D04 LATIN LETTER SMALL CAPITAL C	217D SMALL ROMAN NUMERAL ONE HUNDRED	2CA5 COPTIC SMALL LETTER SIMA		ABA5 CHEROKEE SMALL LETTER TLI

utoronto.ca

utoronto.ca

Visual bias

- ‘Visual’ priors: mappings off the **Unicode confusables list**
- Designed to combat spoofing attacks

006F LATIN SMALL LETTER O	03BF GREEK SMALL LETTER OMICRON	03C3 GREEK SMALL LETTER SIGMA	043E CYRILLIC SMALL LETTER O	0585 ARMENIAN SMALL LETTER OH	05E1 HEBREW LETTER SAMEKH	0647 ARABIC LETTER HEH
0063 LATIN SMALL LETTER C	03F2 GREEK LUNATE SIGMA SYMBOL	0441 CYRILLIC SMALL LETTER ES	1D04 LATIN LETTER SMALL CAPITAL C	217D SMALL ROMAN NUMERAL ONE HUNDRED	2CA5 COPTIC SMALL LETTER SIMA	A8AF CHEROKEE SMALL LETTER TLI

utoronto.ca



utoronto.ca

The site you just tried to visit looks fake. Attackers sometimes mimic sites by making small, hard-to-see changes to the URL.

Visual bias

- ‘Visual’ priors: mappings off the **Unicode confusables list**
 - Designed to combat spoofing attacks
 - Hardly any mappings for Arabic!

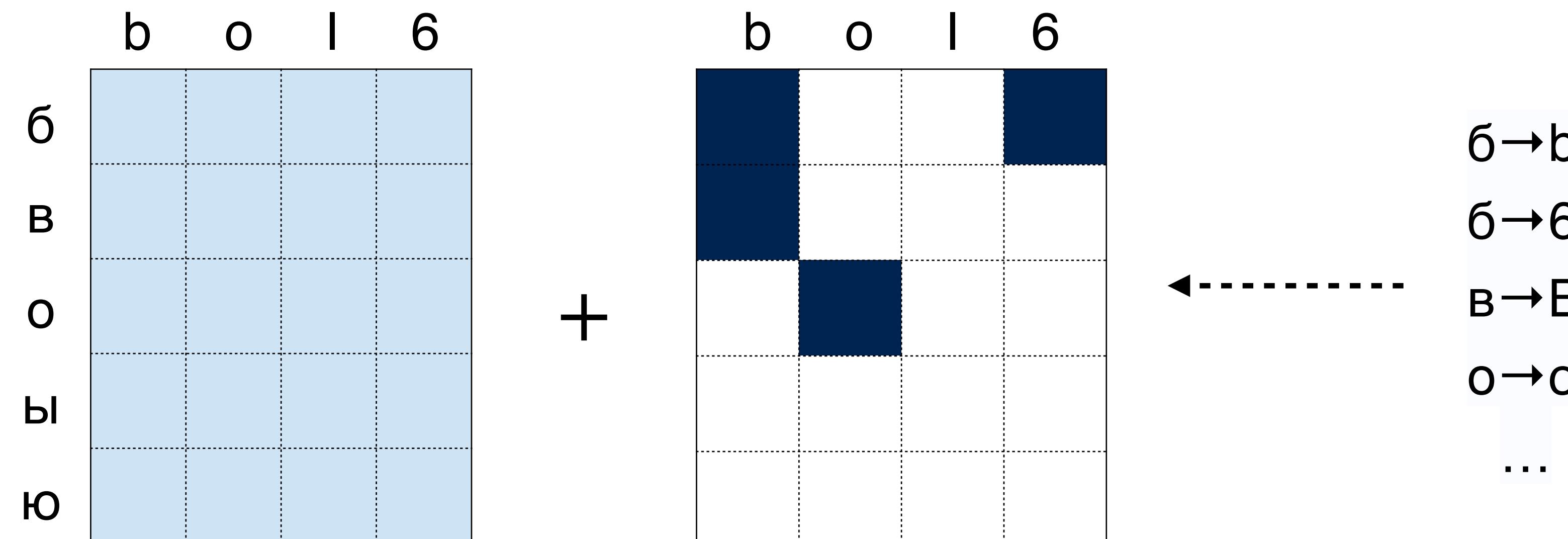
o	o	σ	o	o	o	ດ	ດ
006F LATIN SMALL LETTER O	03BF GREEK SMALL LETTER OMICRON	03C3 GREEK SMALL LETTER SIGMA	043E CYRILLIC SMALL LETTER O	0585 ARMENIAN SMALL LETTER OH	05E1 HEBREW LETTER SAMEKH		0647 ARABIC LETTER HEH
c	c	c	c	c	c	c	c
0063 LATIN SMALL LETTER C	03F2 GREEK LUNATE SIGMA SYMBOL	0441 CYRILLIC SMALL LETTER ES	1D04 LATIN LETTER SMALL CAPITAL C	217D SMALL ROMAN NUMERAL ONE HUNDRED	2CA5 COPTIC SMALL LETTER SIMA		ABA5 CHEROKEE SMALL LETTER TLI

Informative priors

- Use mappings of similar characters as **priors on emission parameters**

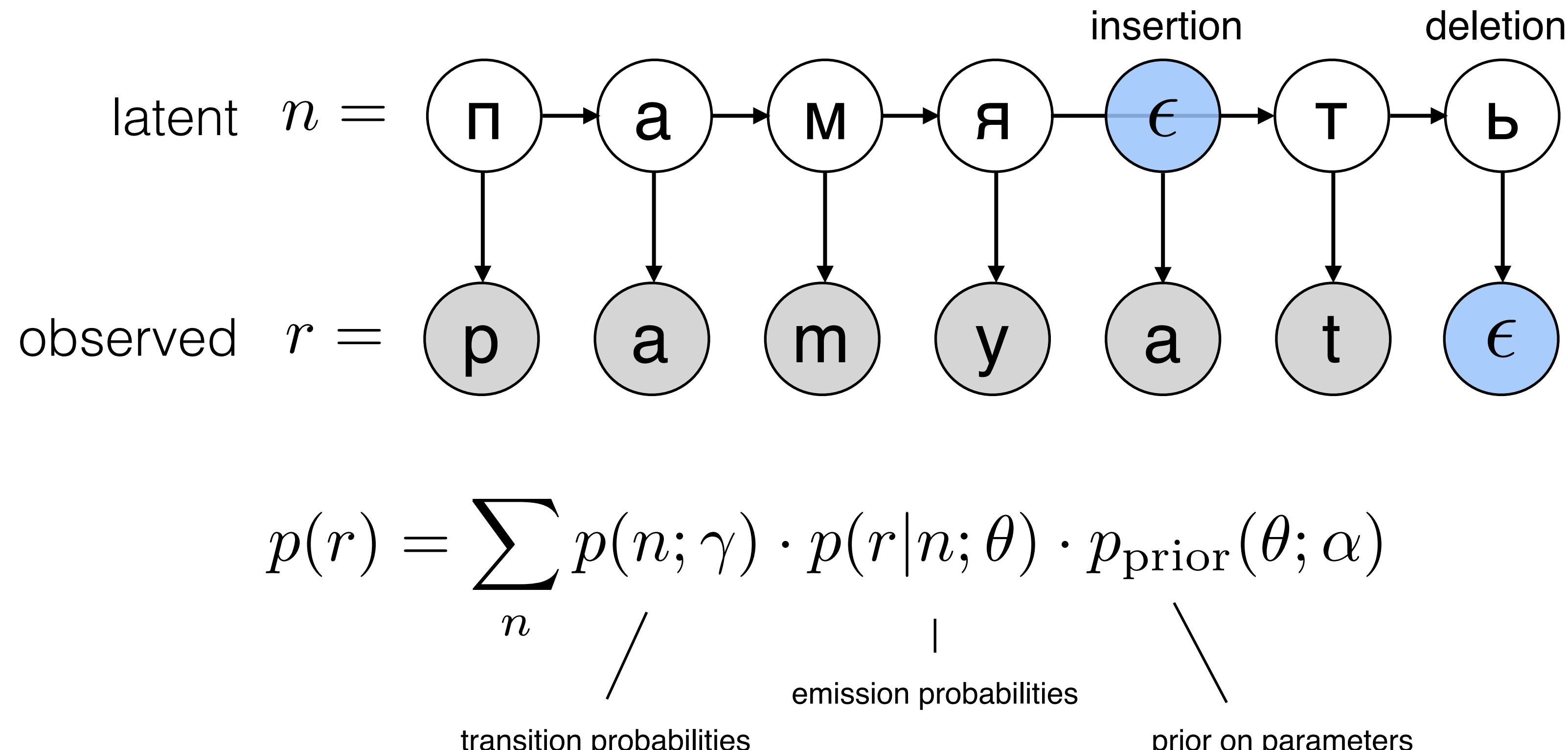
$$c_r | c_n \sim \text{Mult}(\theta_{c_n})$$

$$\theta \sim \text{Dir}(\alpha)$$



Noisy-channel model

- Representing latent alignments via **insertions and deletions**

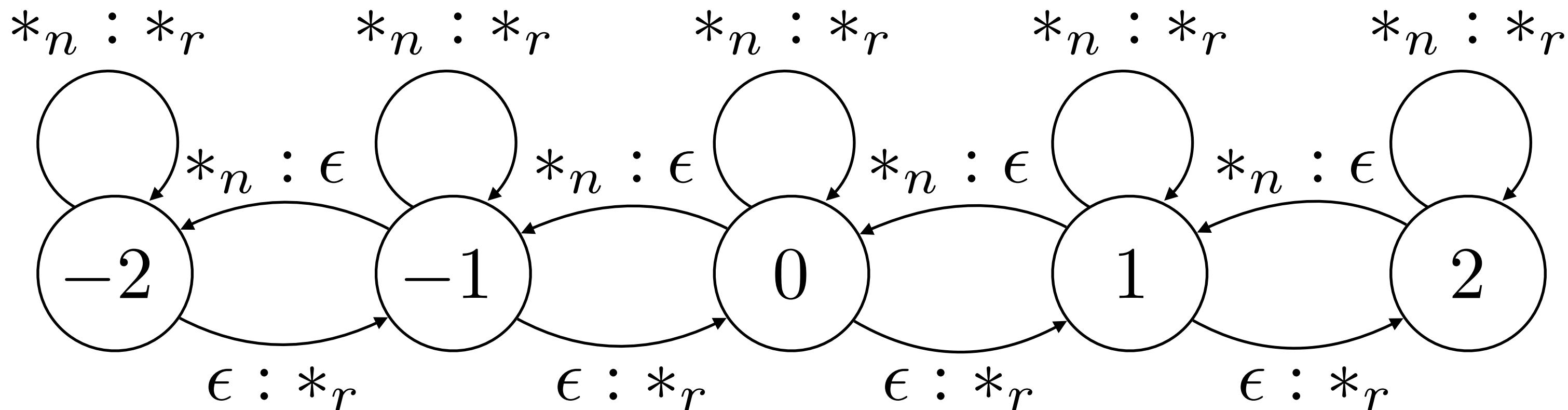
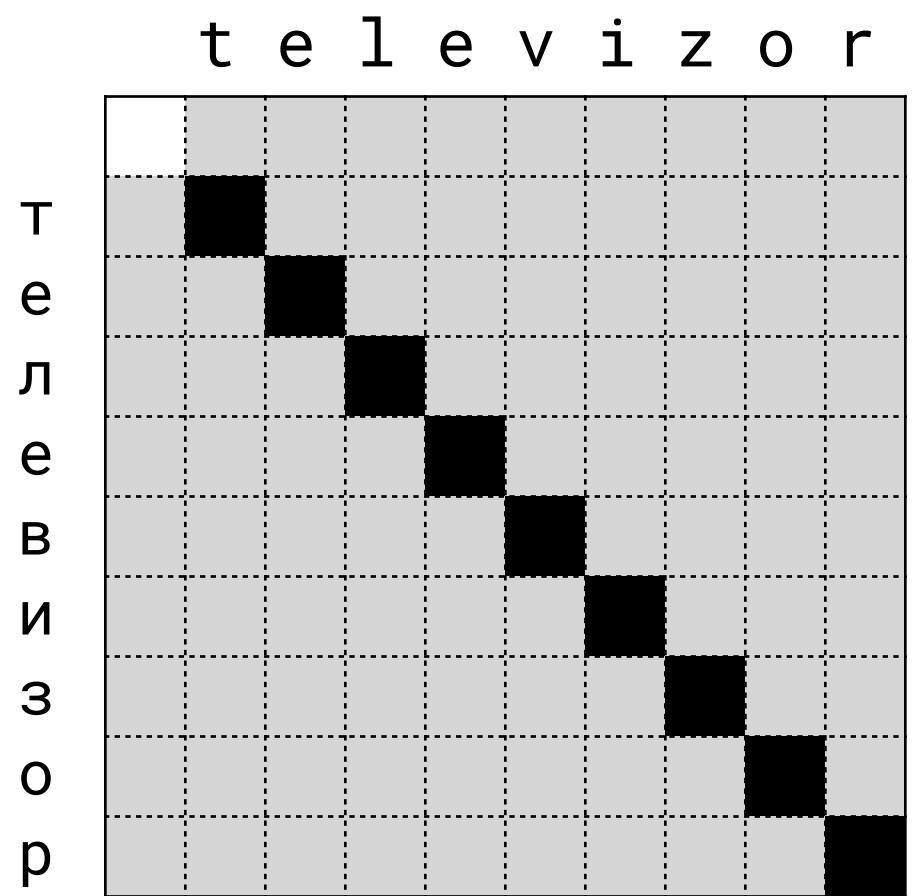


WFST cascade

- Transition WFSA
 - 6-gram LM built with OpenGrm (Roark et al., 2012)
- Emission WFST
 - Supports all substitutions, insertions and deletions

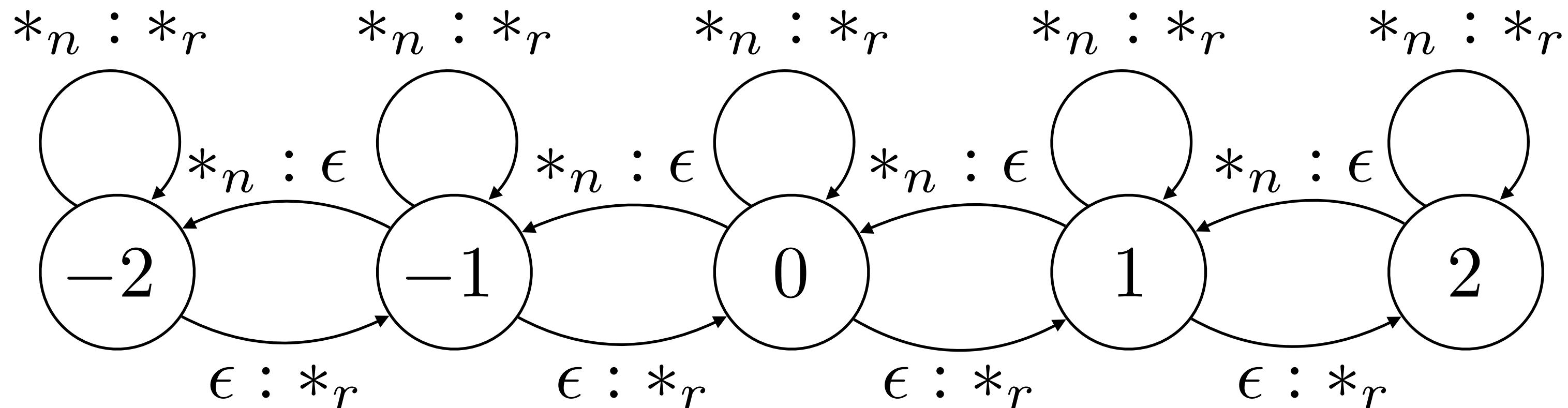
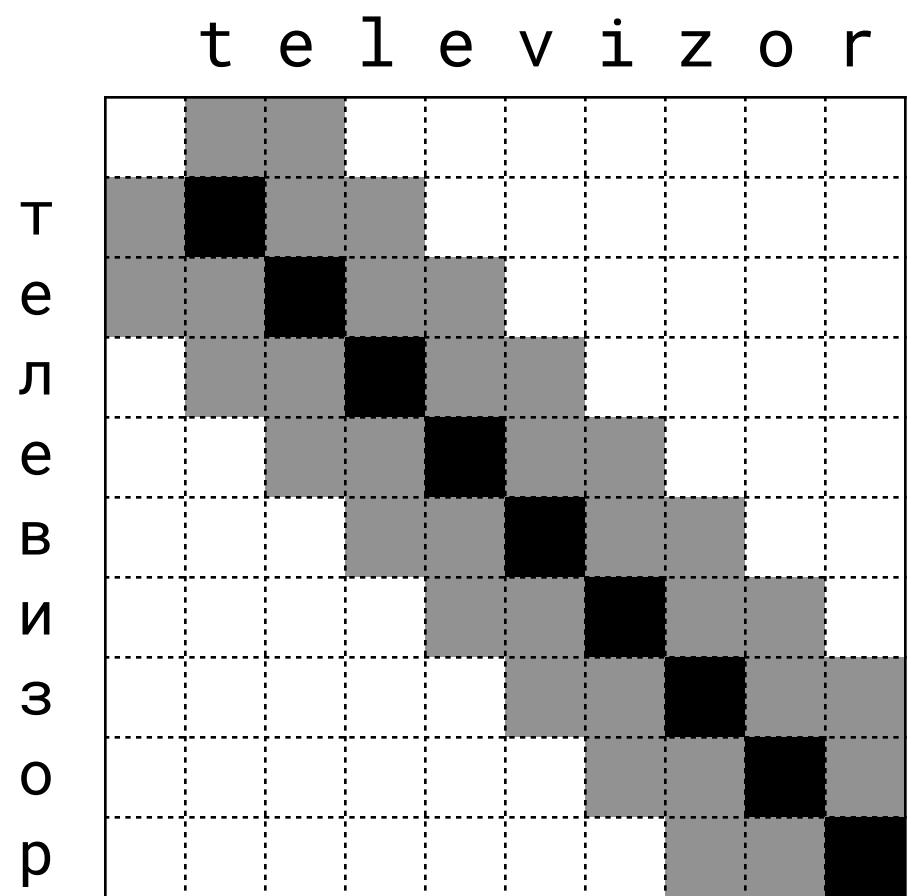
Emission model

- Needs to support substitutions, insertions and deletions
- Fixed limit on delay: $| \# \text{ of insertions} - \# \text{ of deletions} |$



Emission model

- Needs to support substitutions, insertions and deletions
- Fixed limit on delay: $| \# \text{ of insertions} - \# \text{ of deletions} |$



WFST cascade

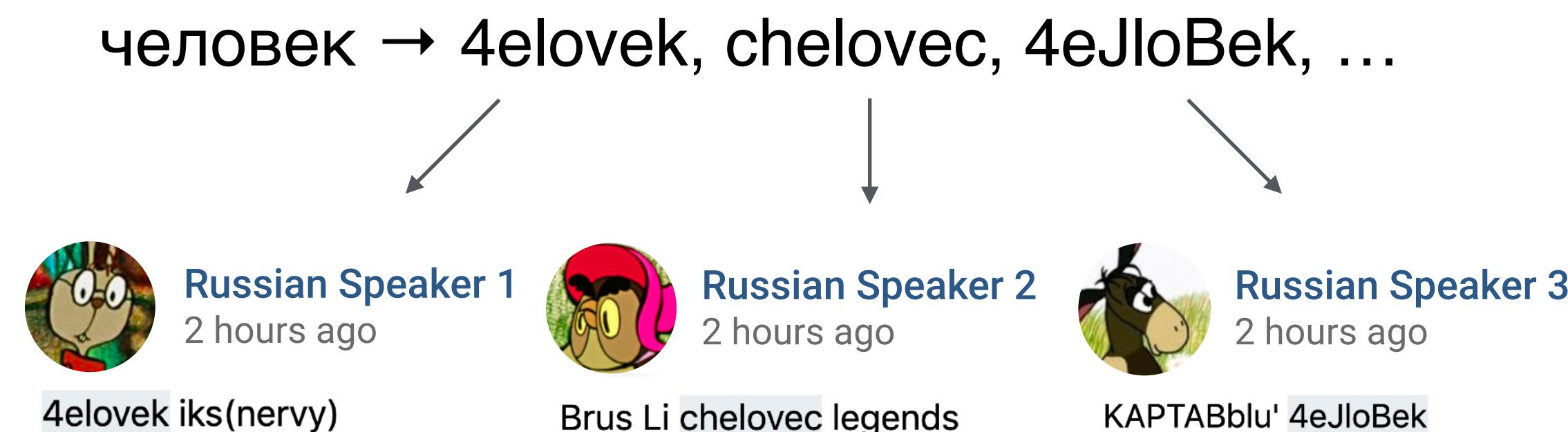
- Transition WFSA
 - 6-gram LM built with OpenGrm (Roark et al., 2012)
- Emission WFST
 - Supports all substitutions, insertions and deletions
- Trained with EM algorithm
 - OpenFst (Allauzen et al., 2007)
 - Speedup tricks: stepwise training, curriculum learning, pruning...

Datasets

- Arabic:
- Arabizi SMS/chat dialogs, converted to CODA (Habash et al., 2012)

Saba7 el 5eir!
Ezayeeky?

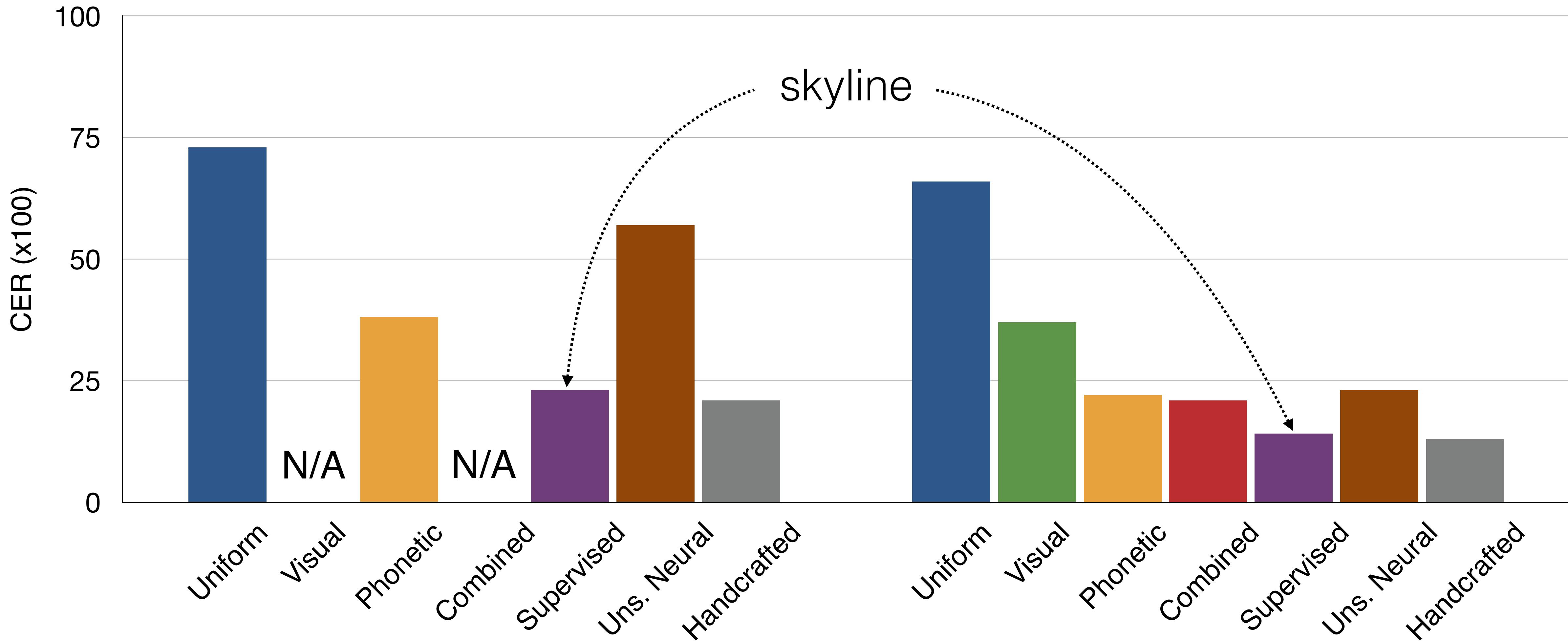
- Russian:
- Romanized: collected and partly annotated data from social media



- Native: Taiga corpus (Shavrina & Shapovalova, 2017), scraped from the same platform

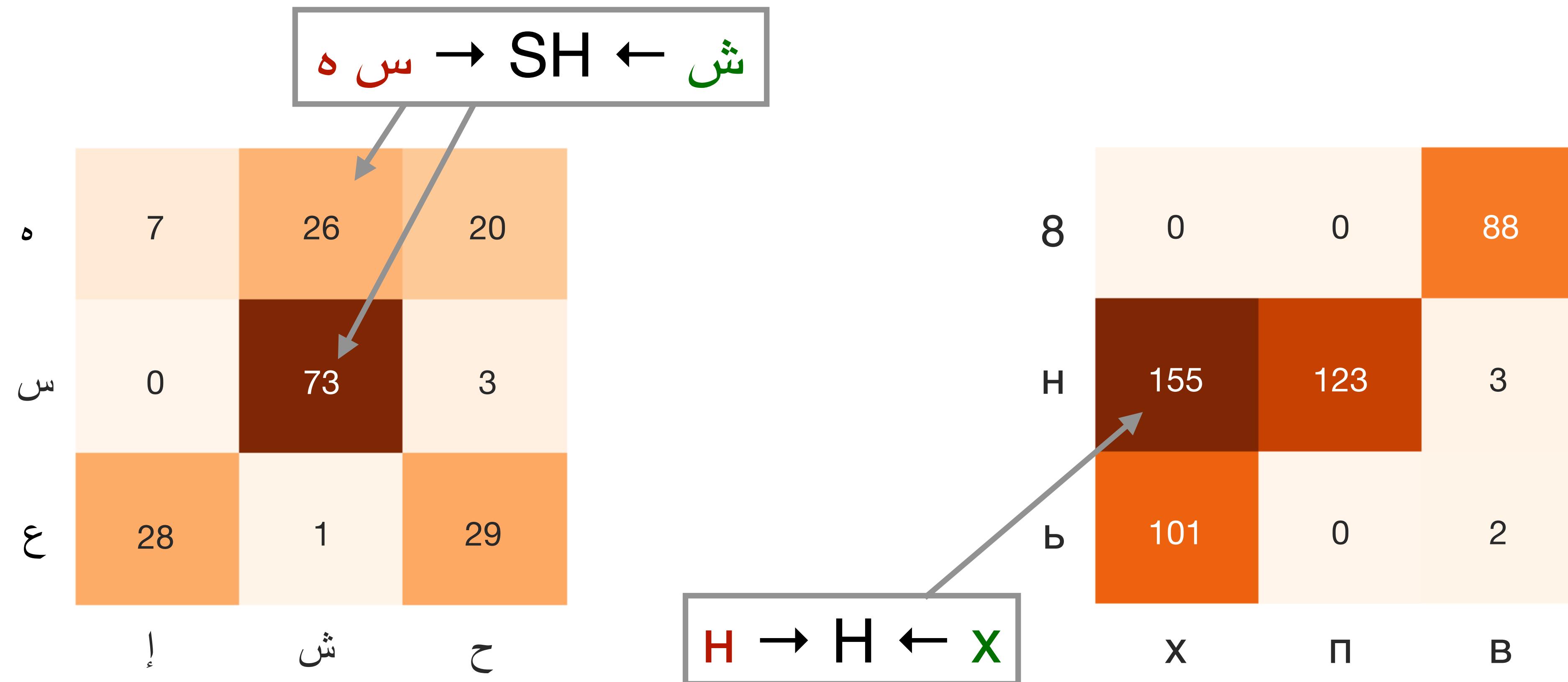
Experimental results

Arabic



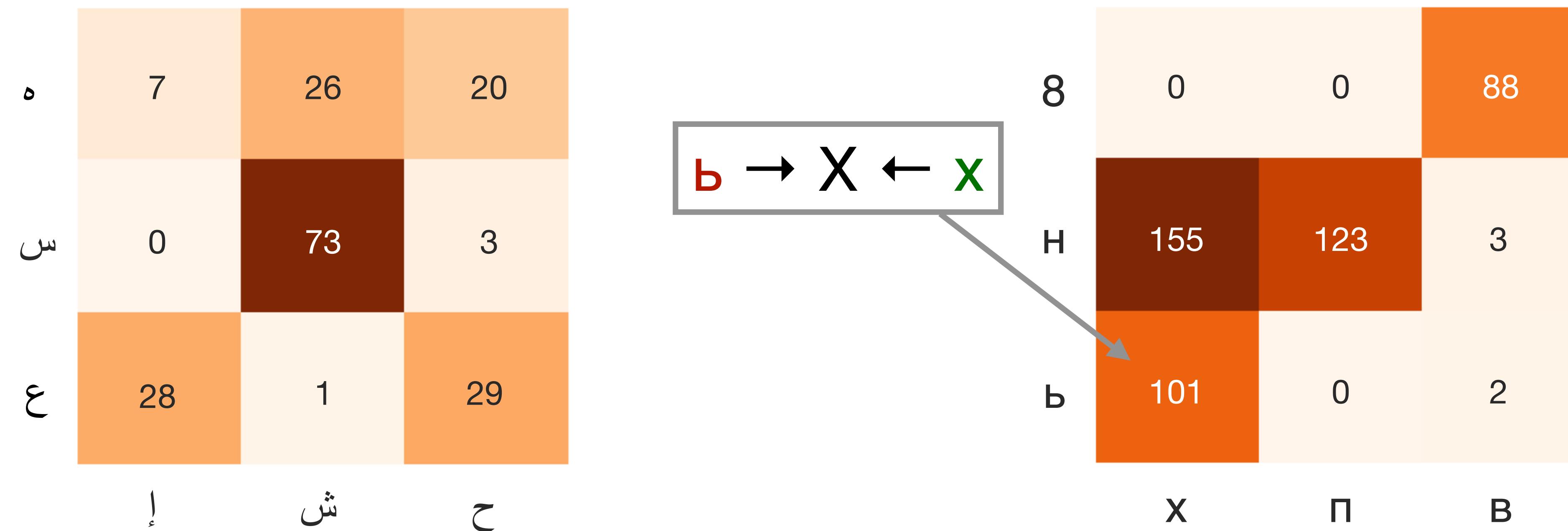
Error analysis

- Many errors are likely due to a weak language model



Error analysis

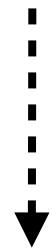
- Some errors are caused by spurious mappings in priors



Error analysis

- Additional experiments with romanized Kannada
- Error analysis for unsupervised finite-state and neural models
 - Finite-state models make more repetitive errors
 - Neural models are more sensitive to distributional shift
 - Character tokenization boosts performance of the neural model

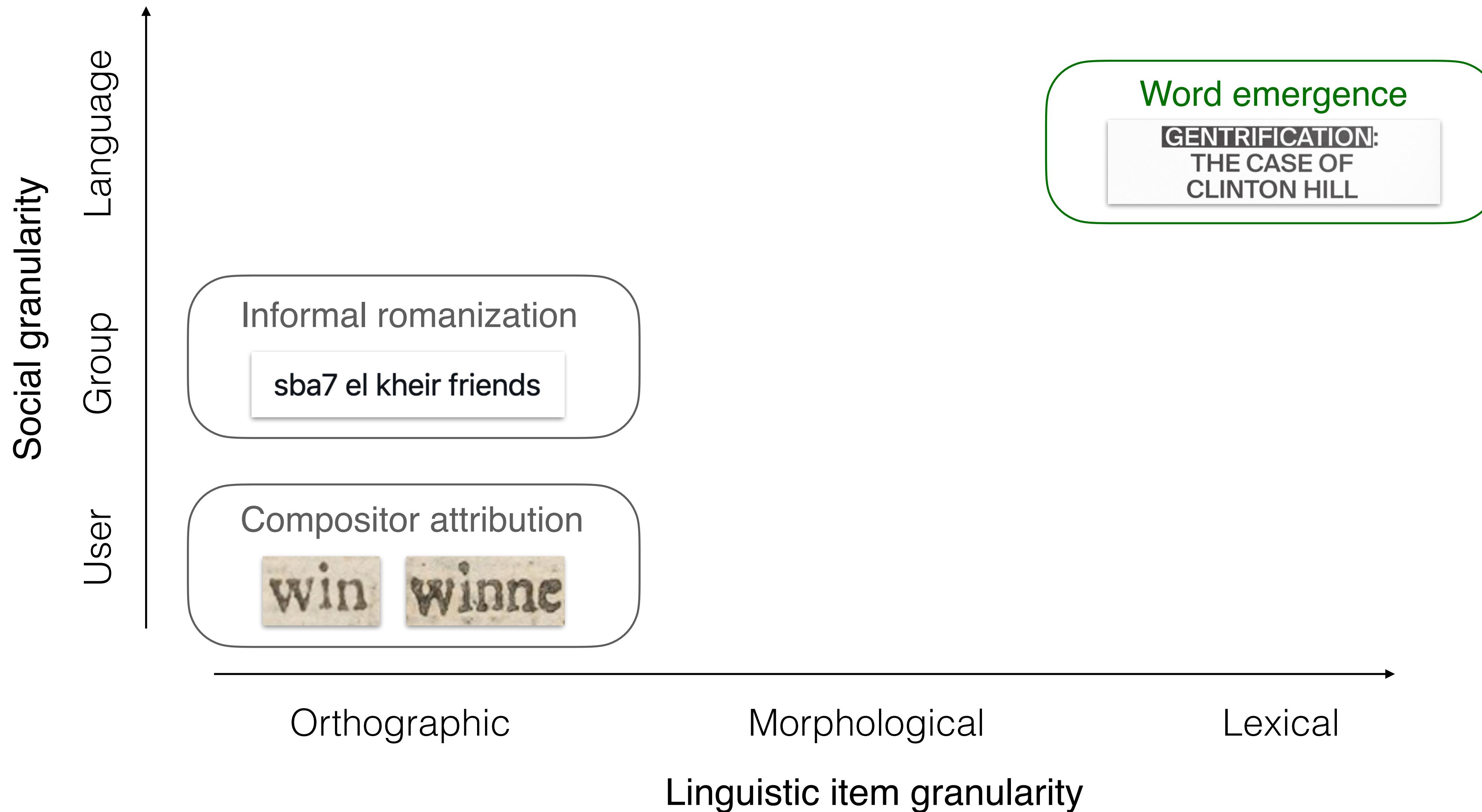
mana belagitu



ಮನ ಬೆಳಗಿತು

M Ryskina, E Hovy, T Berg-Kirkpatrick, MR Gormley. Comparative Error Analysis in Neural and Finite-state Models for Unsupervised Character-level Transduction. SIGMORPHON 2021.

Spectrum of phenomena



Neology

- New words (*neologisms*) appear in our languages all the time
 - Coined from whole cloth (*grok*)
 - Borrowed from other languages (*tahini*)
 - Derived from existing words (*unfriend*)
 - Existing words taking on new senses (*web* → ‘*the Internet*’)
- **What are the semantic factors that characterize word emergence?**
 - What is the trade-off between language-internal and language-external factors?

Hypotheses

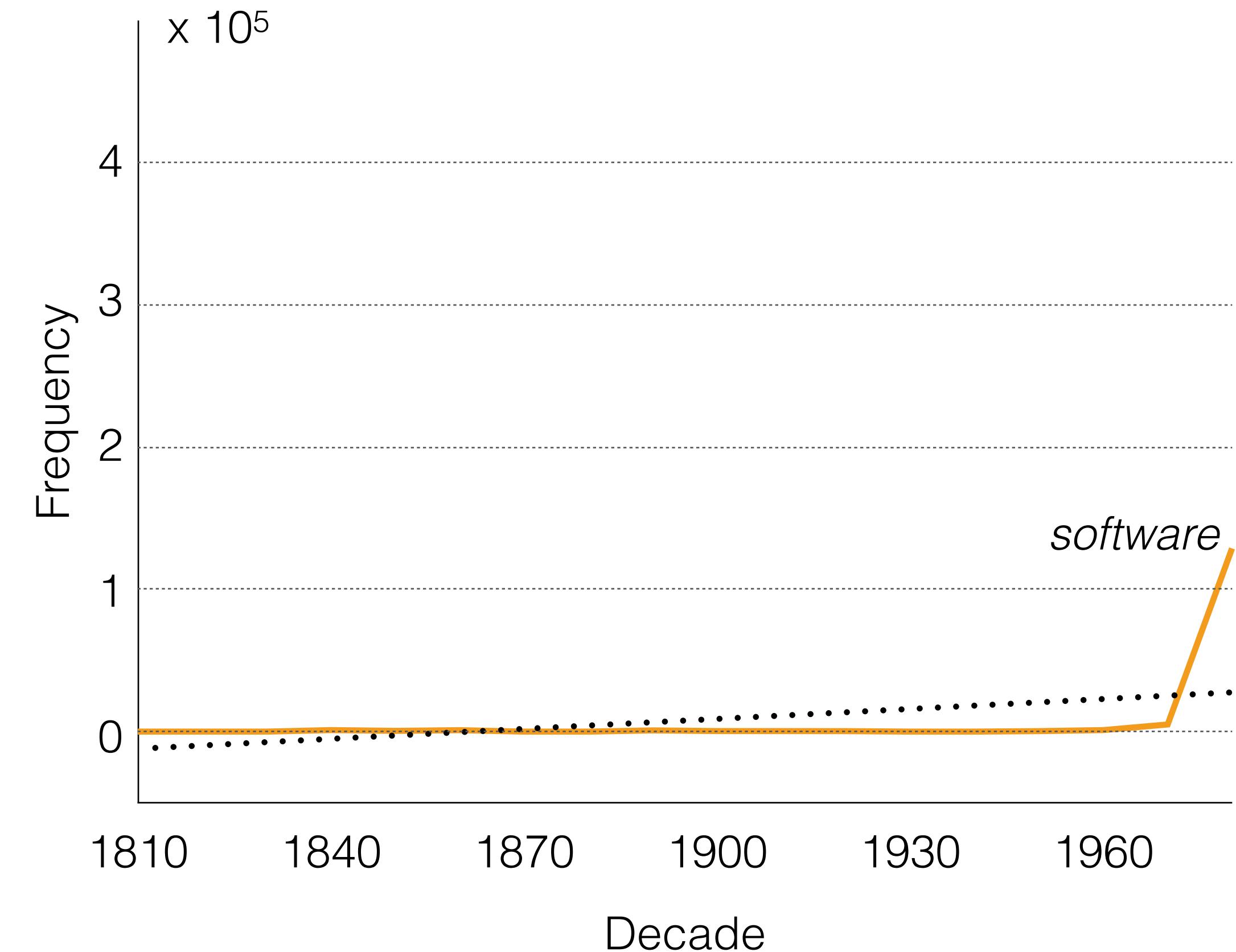
- **Supply:** neologisms are more likely to emerge in **sparser areas of the semantic space**
 - Semantic space tends towards uniformity (Bréal, 1904)
 - New words emerge to fill in ‘semantic gaps’
- **Demand:** neologisms are more likely to emerge in **semantic neighborhoods of growing popularity**
 - Word frequency growth correlates with growing importance in discourse
 - The more important the domain is, the more new concepts (and words) emerge

Diachronic corpora

- Historical (COHA; Davies, 2002) and modern (COCA; Davies, 2008) American English corpora

Diachronic corpora

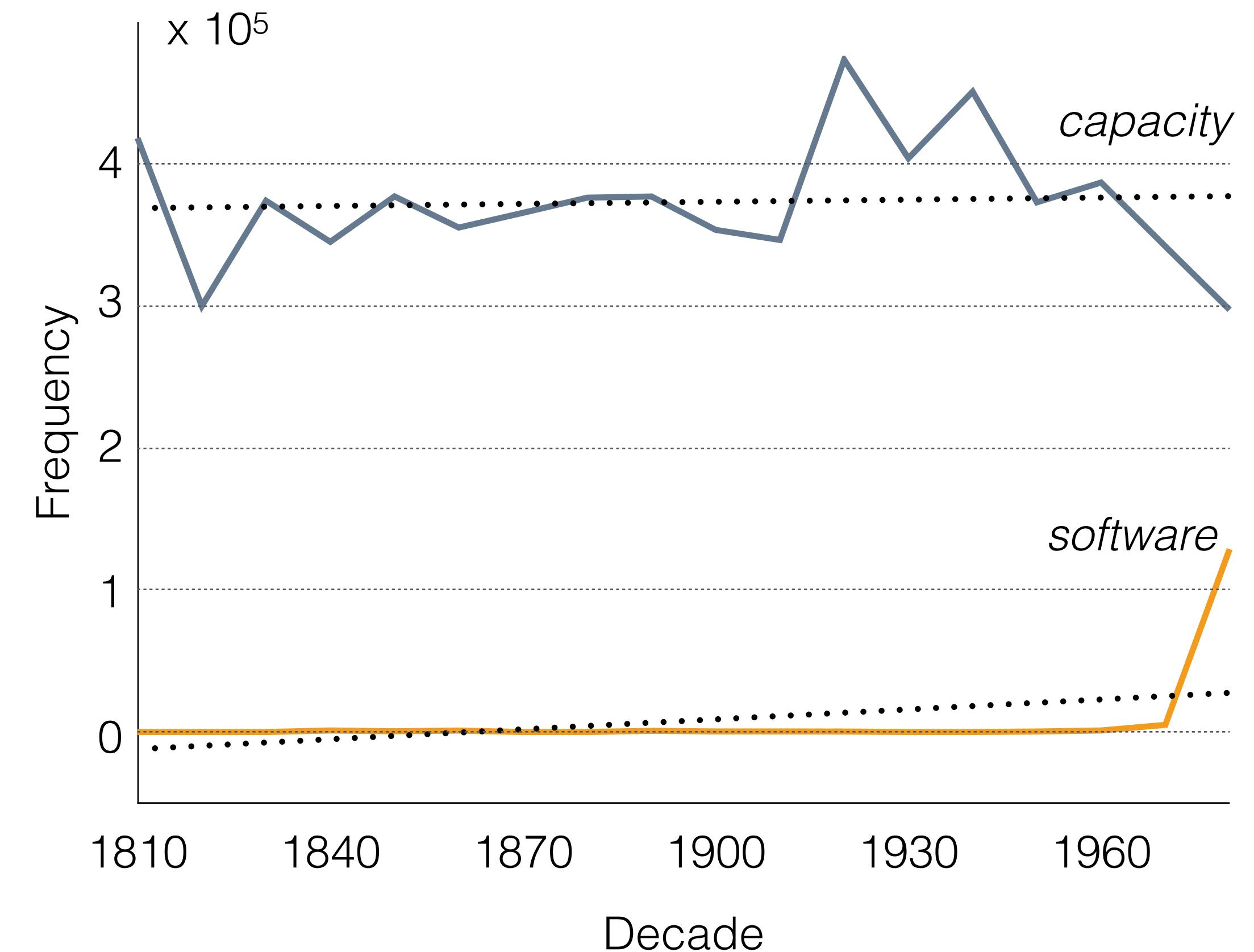
- Historical (COHA) and modern (COCA) American English corpora
 - Neologisms by frequency ratio between modern and historical data



M Ryskina, E Rabinovich, T Berg-Kirkpatrick, DR Mortensen, Y Tsvetkov. Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods. SCiL 2020.

Diachronic corpora

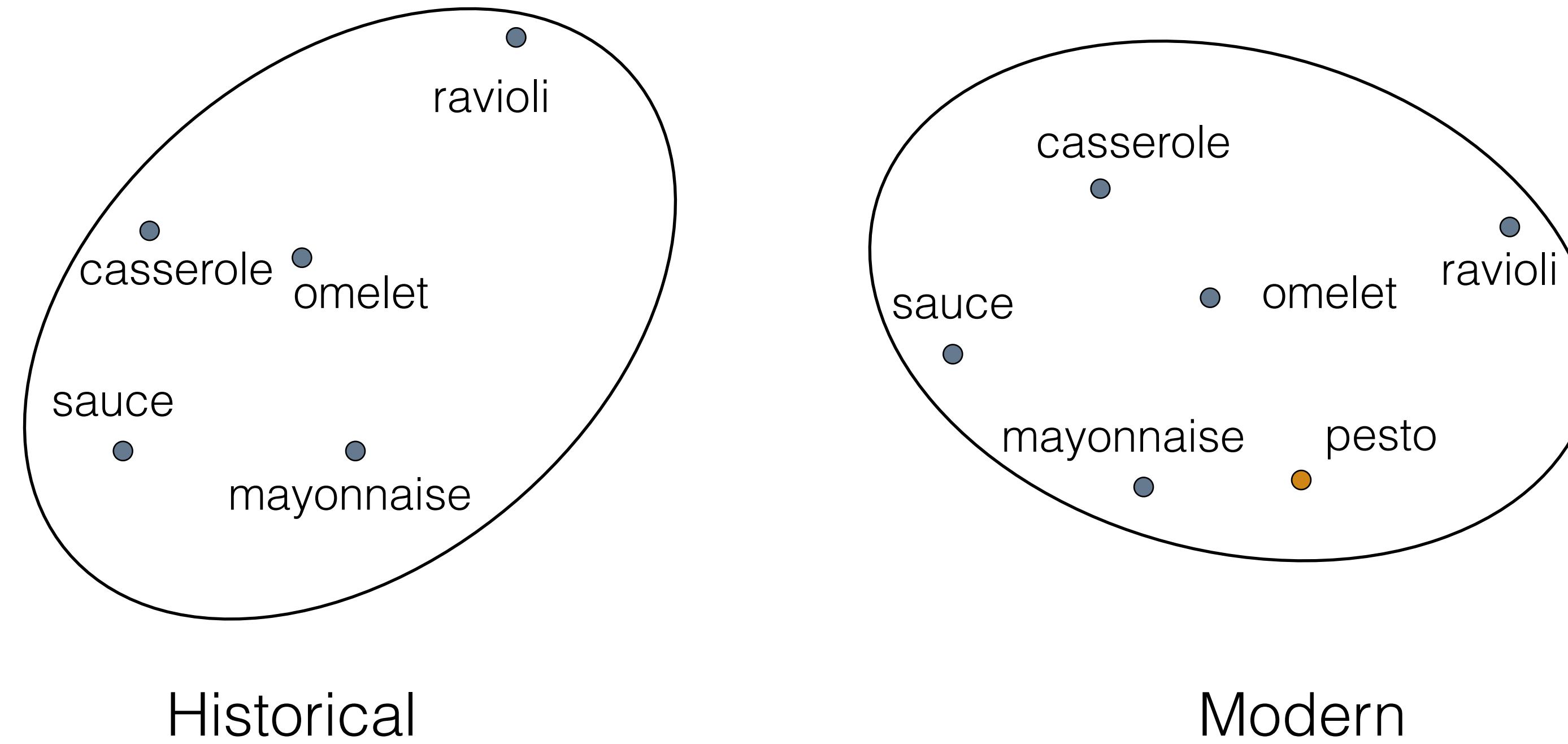
- Historical (COHA) and modern (COCA) American English corpora
 - Neologisms by frequency ratio between modern and historical data
- Each neologism paired with a non-neologism control word
 - Controlling for frequency, length, frequency stability



M Ryskina, E Rabinovich, T Berg-Kirkpatrick, DR Mortensen, Y Tsvetkov. Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods. SCIL 2020.

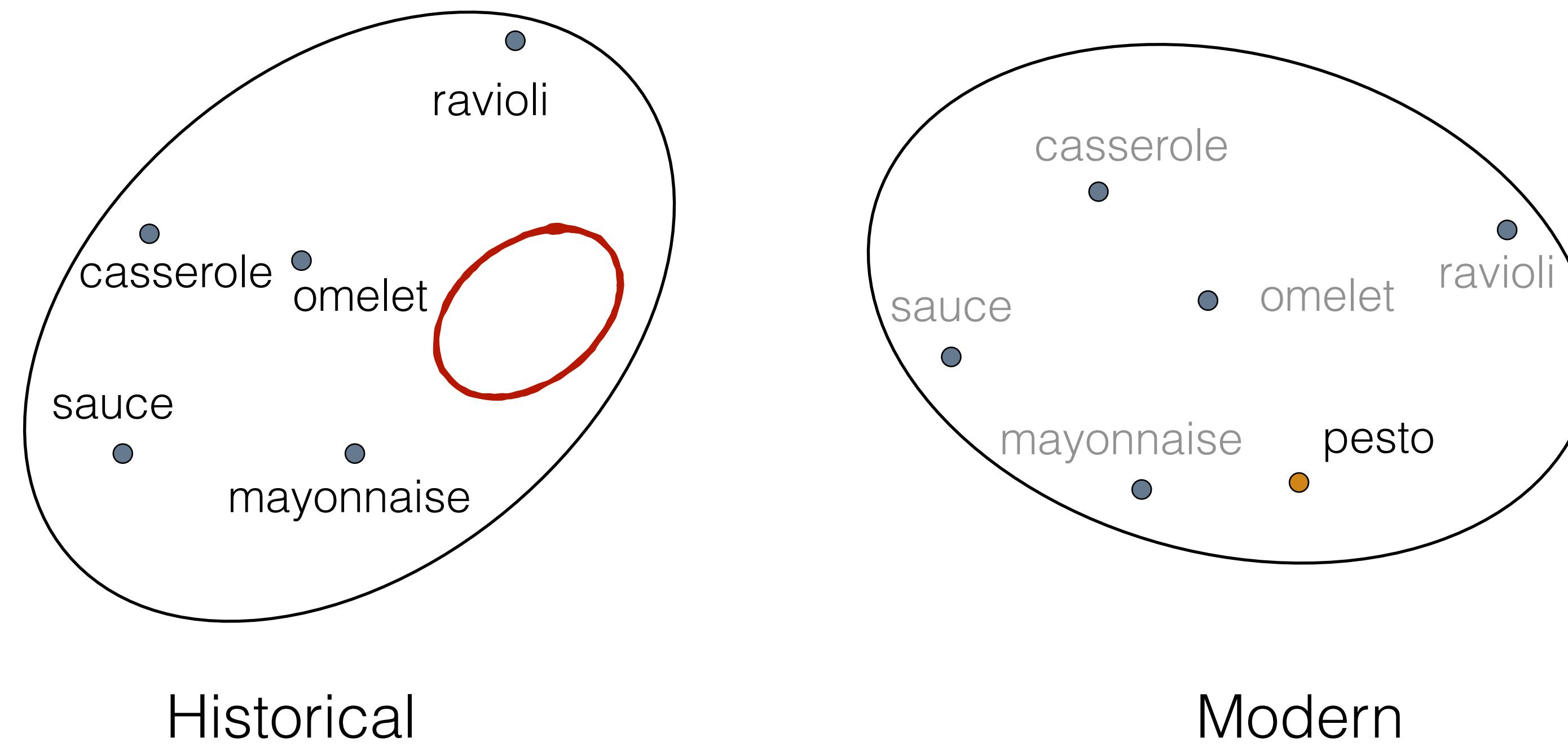
Distributional semantics

- Embeddings learned separately from historical and modern data



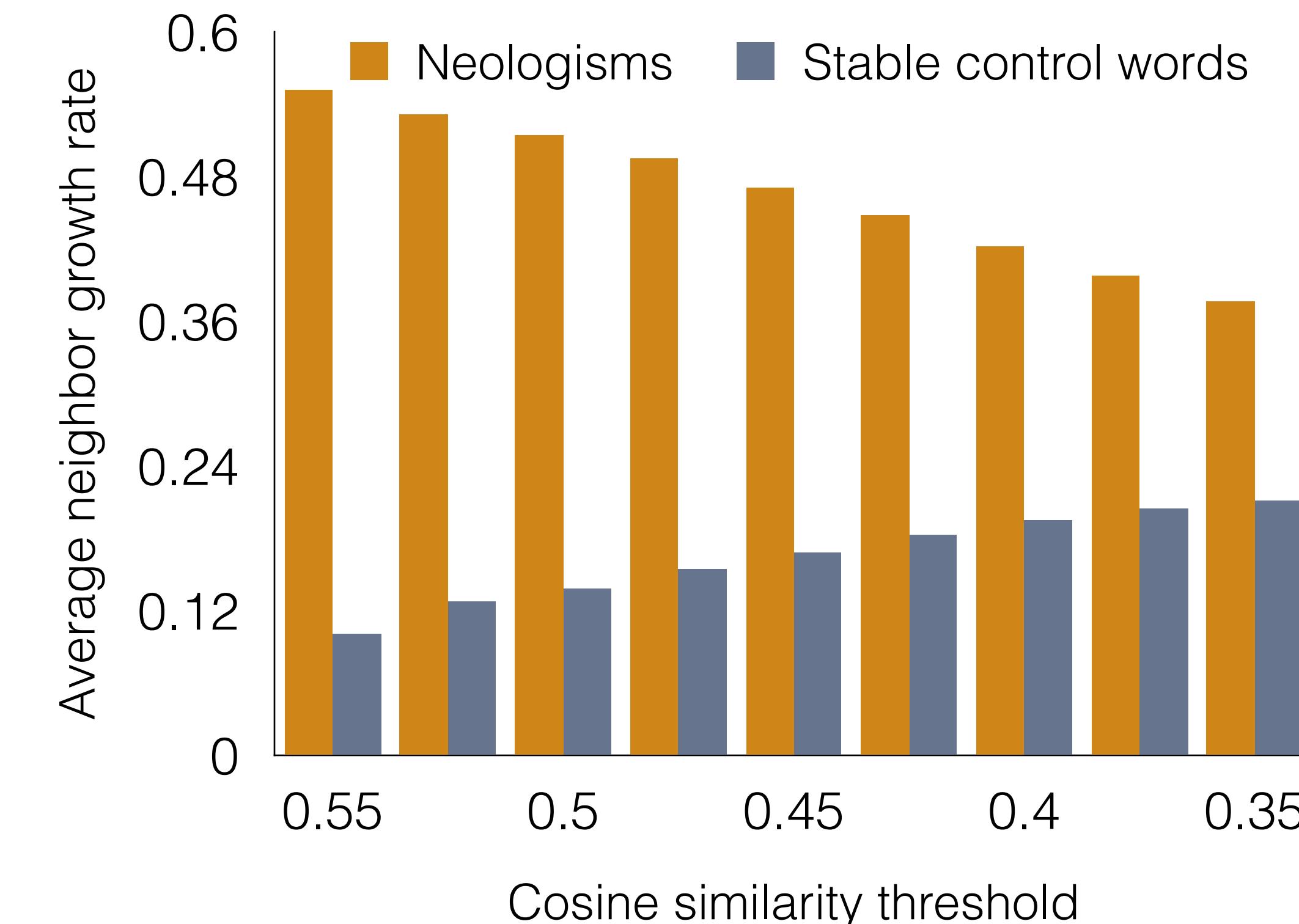
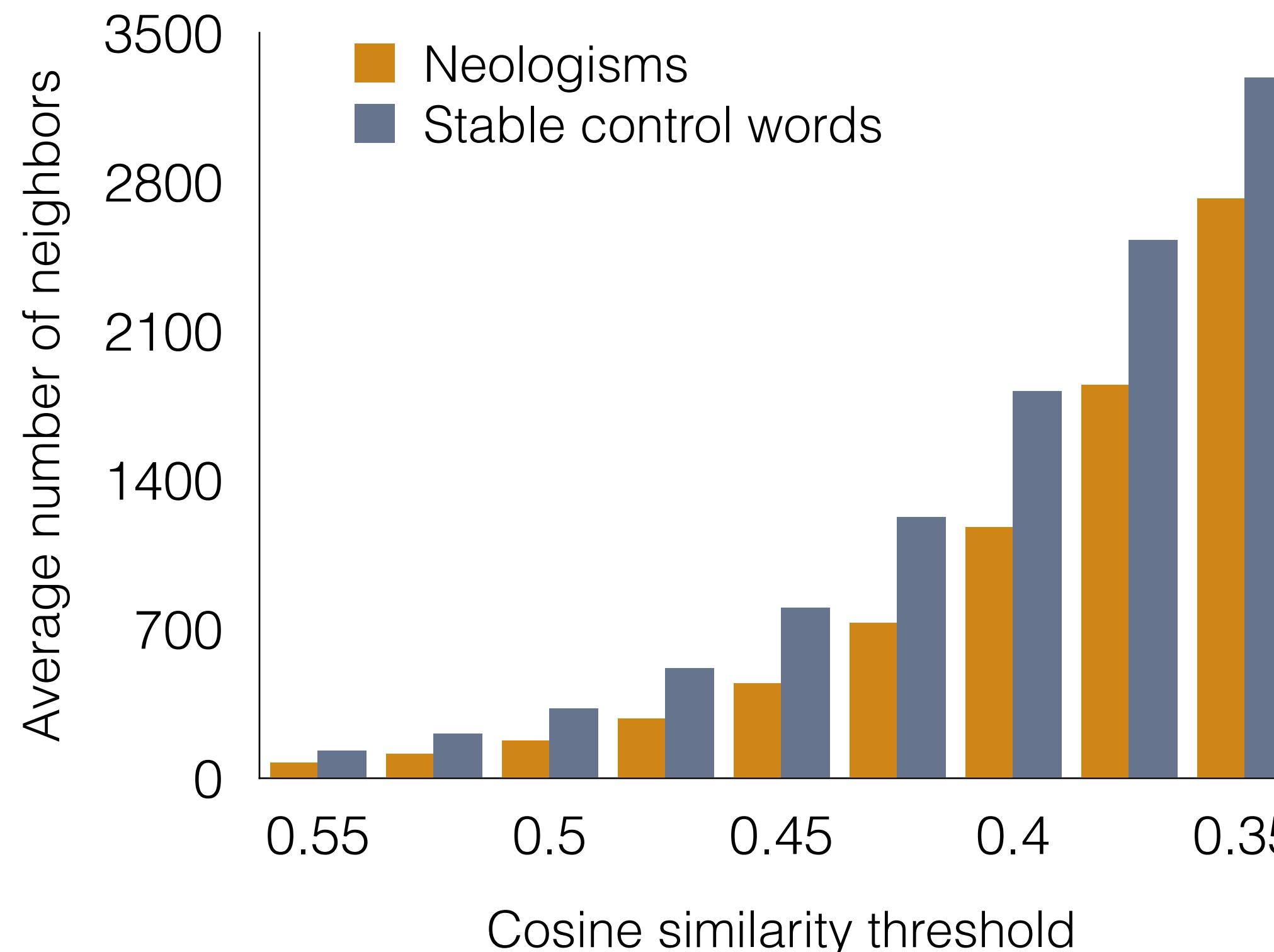
Distributional semantics

- Embeddings learned separately from historical and modern data
 - Aligned using anchor words, neologisms projected into historical space
- Measure neighborhood density and average word frequency growth rate



Experimental results

- Both density and frequency growth are predictive of neology
 - Only frequency growth is always statistically significant in GLM analysis

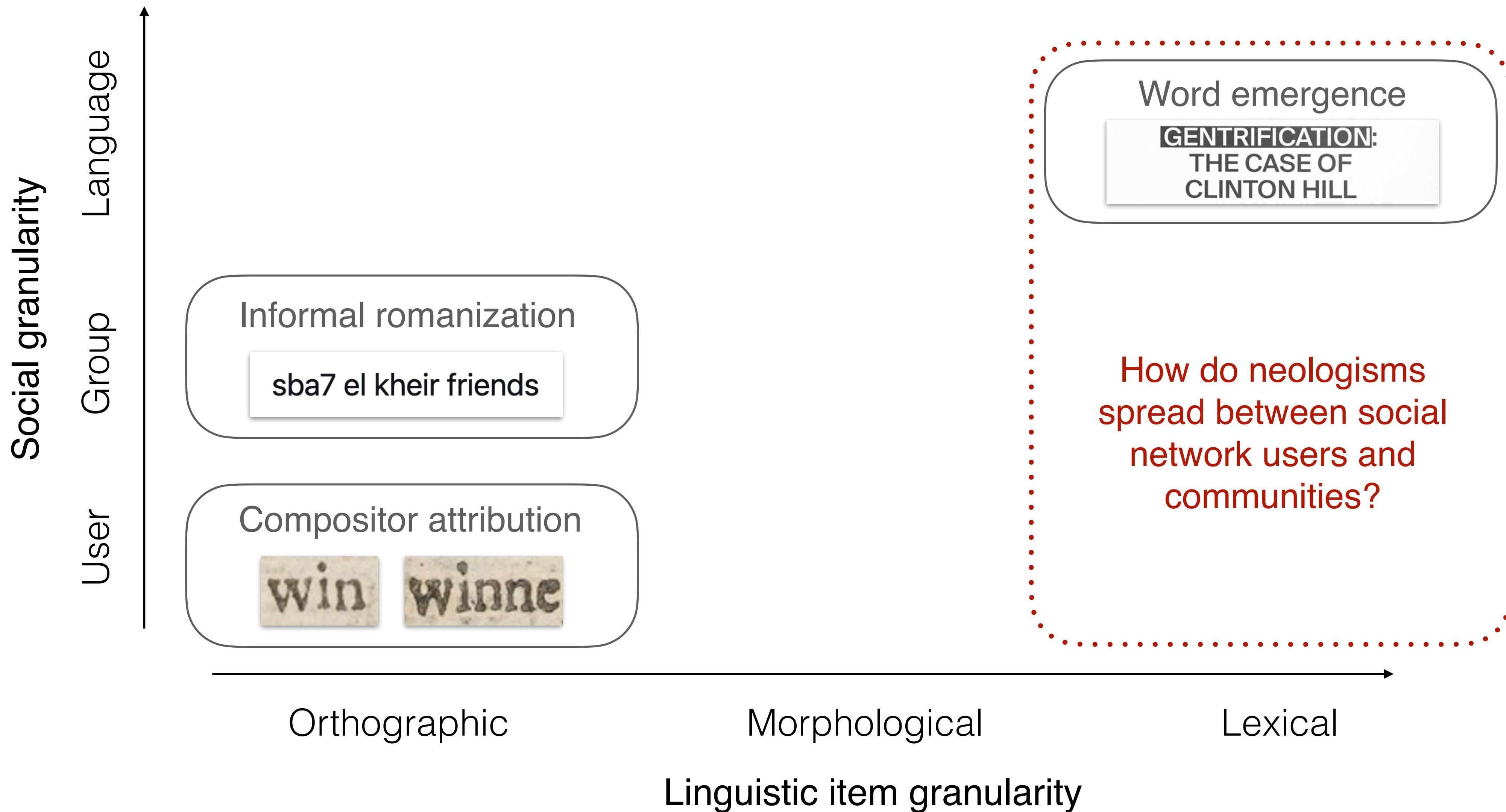


Qualitative analysis

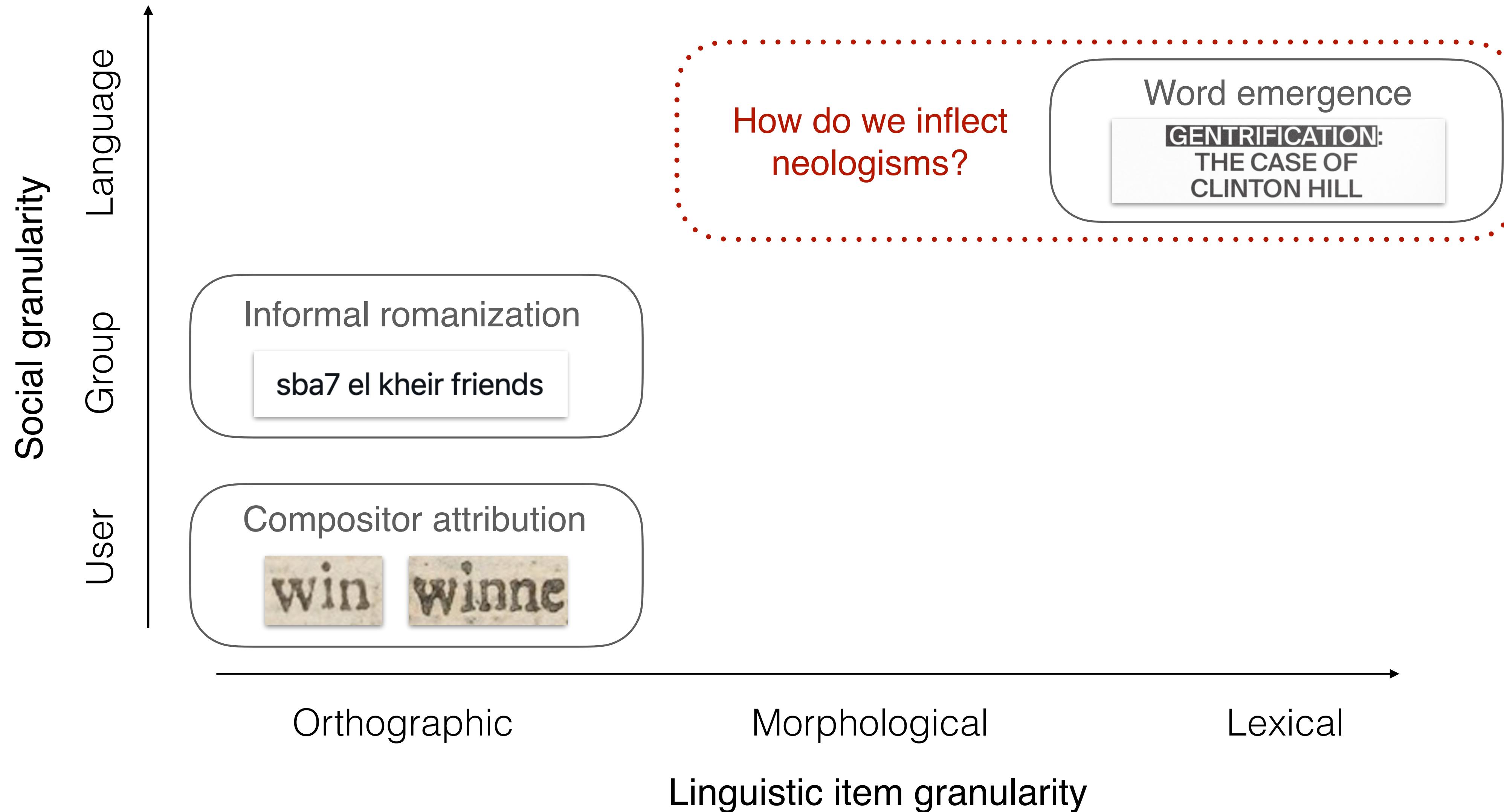
- Example of nearest historical neighbors of projected neologisms:

Neologism	Nearest neighbors	
email	telegram	letter
pager	beeper	phone
blogger	journalist	columnist
spokeswoman	spokesman	director
sushi	caviar	risotto
e-book	paperback	hardcover
hip-hop	jazz	rock-n-roll
daycare	day-care	childcare
vibe	ambience	ambiance
chemo	chemotherapy	dialysis

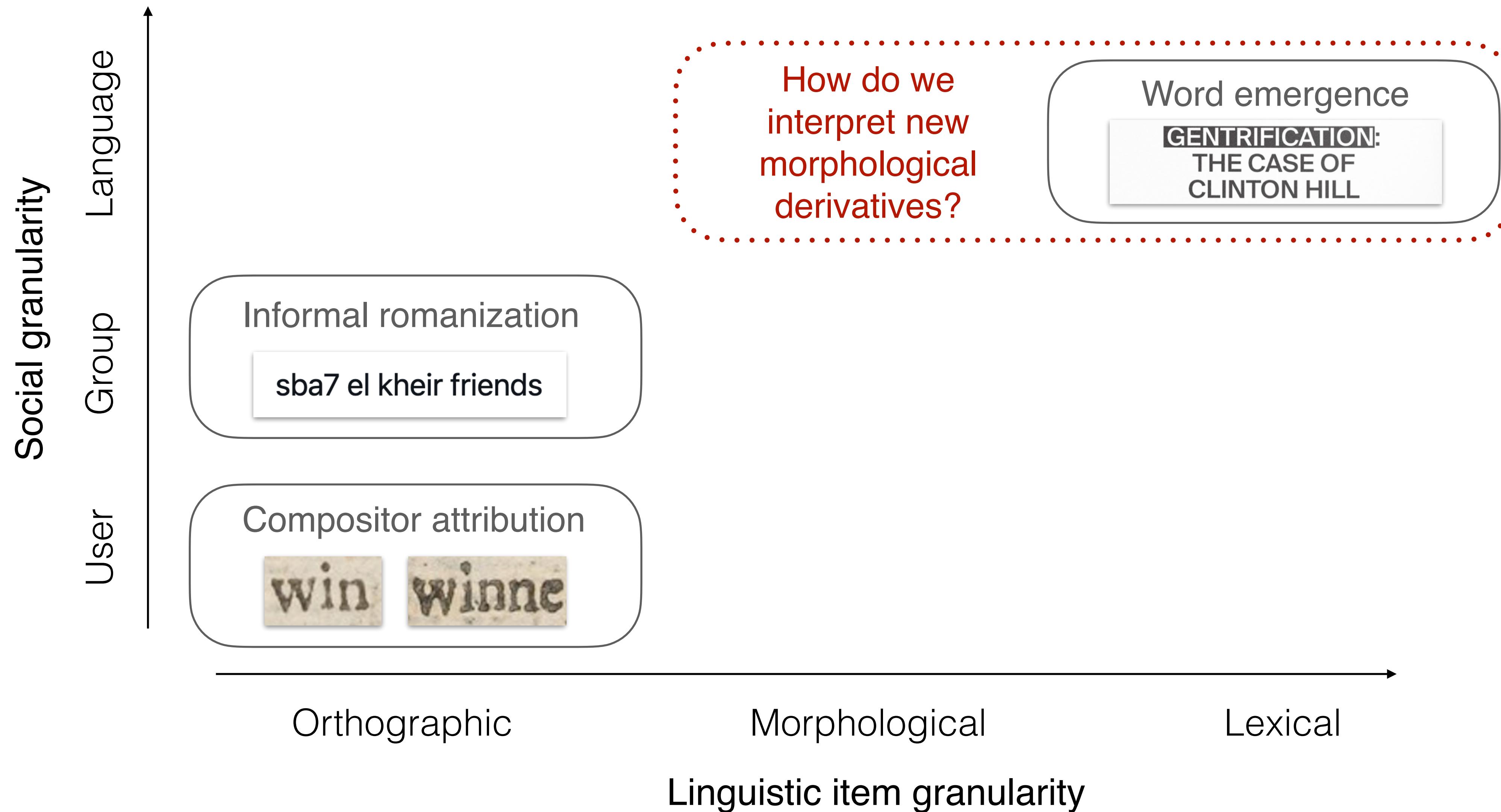
Future work: Multiscale studies



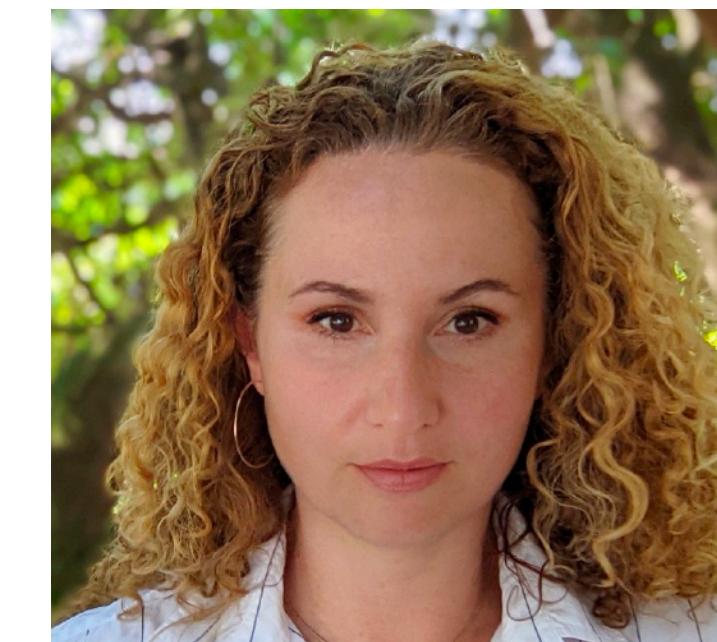
Future work: Morphology



Future work: Morphology



Thank you!



Get in touch:

www.cs.cmu.edu/~mryskina/

mryskina@cs.cmu.edu

[@maria_ryskina](https://twitter.com/maria_ryskina)