# Learning Computational Models of Non-Standard Language

Maria Ryskina


Carnegie Mellon University
Language Technologies Institute

# 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


- Variety of genres
- Variety of domains
- Spectrum of formality



https://readandsurvive.com/2014/07/09/how-to-stop-evil-malchiks/

# 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 = 'two'* = /tu/
  - Compositionality: *2nite = 'two' + 'nite'* = /tu/ + /naɪt/ ≈ /tənaɪt/
    *antivehicleness = 'anti' + 'vehicle' + 'ness'*
- Q2: How can we get our NLP systems to that level?
  - Text normalization: *2nite* → *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?

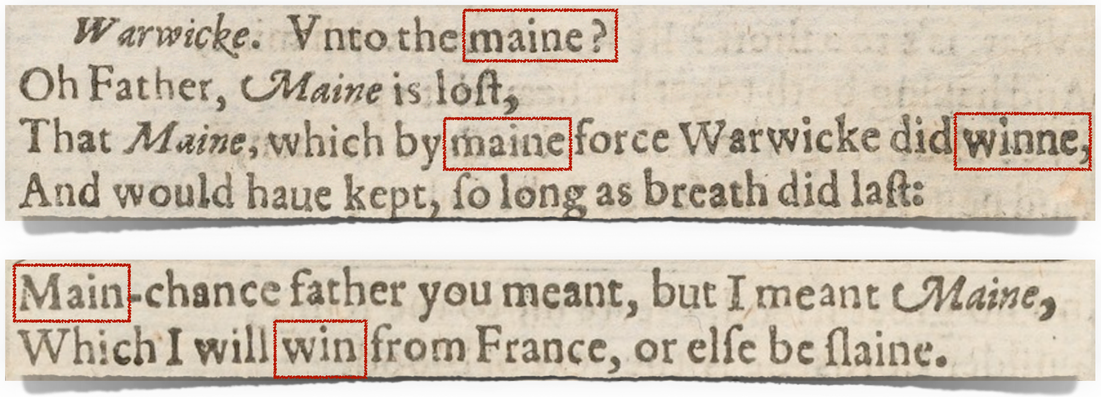
# Non-standard(ized) orthographies

- 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.

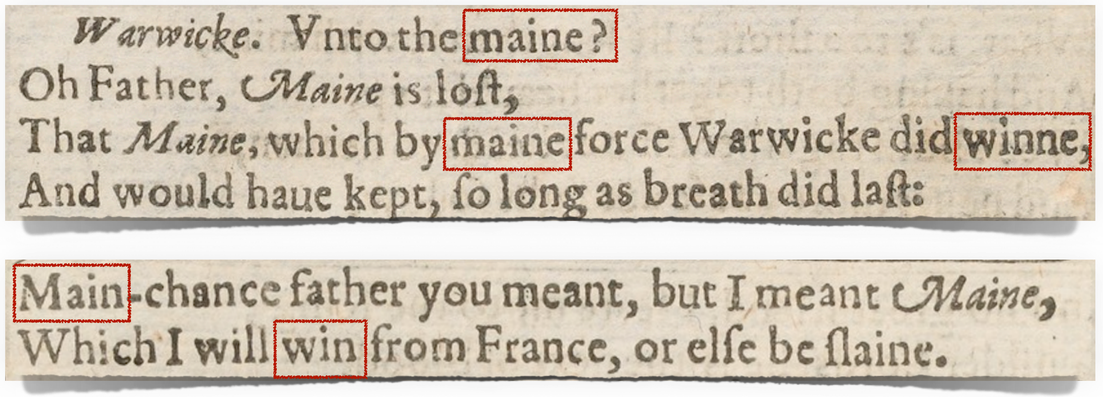
# Non-standard(ized) orthographies

- 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*


*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

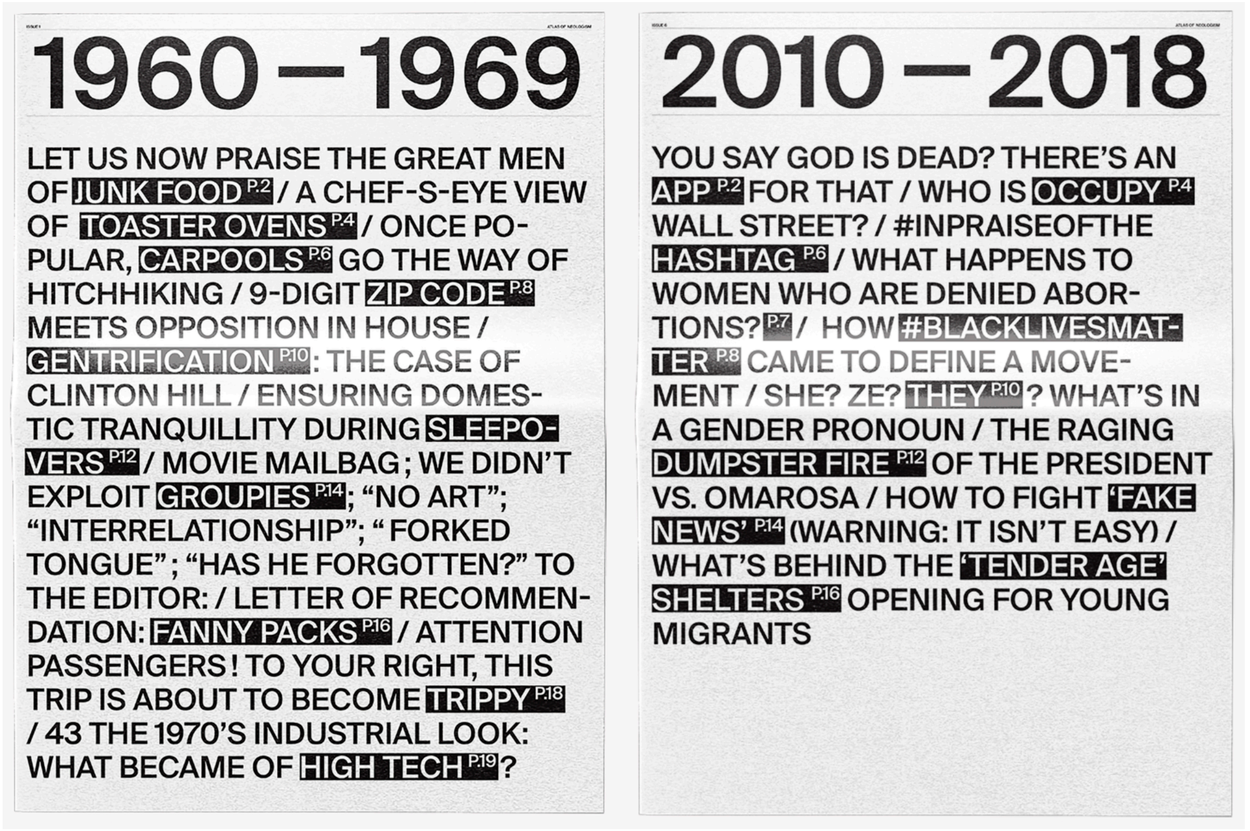
- Not having (or not following) a prescribed way of spelling
- Idiosyncratic: people make different spelling choices
- Shared: spellings reflect underlying pronunciation or ‘standard’ spelling

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



1960—1969

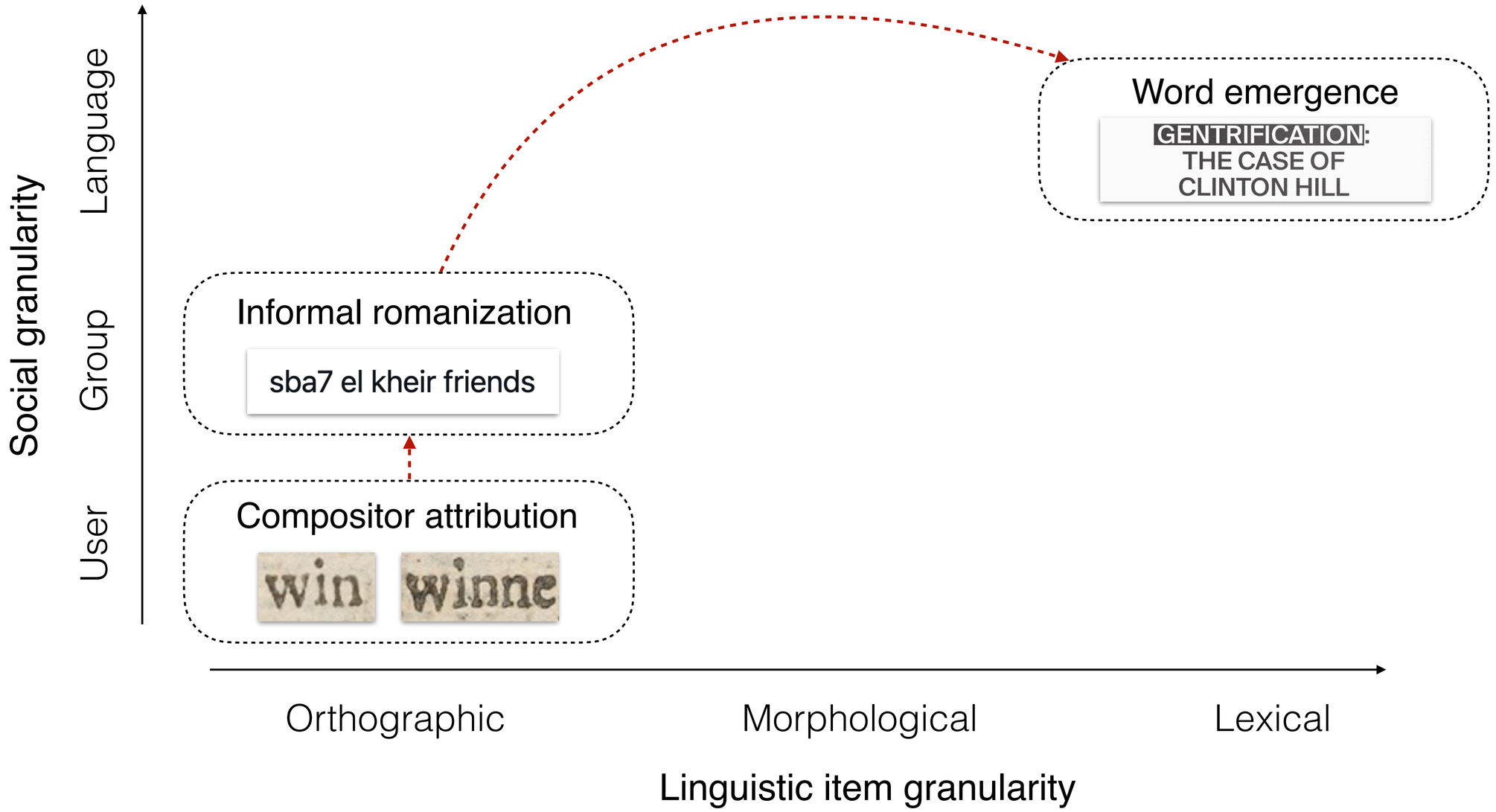
LET US NOW PRAISE THE GREAT MEN OF JUNK FOOD P.2 / A CHEF-S-EYE VIEW OF TOASTER OVENS P.4 / ONCE POPULAR, CARPOOLS P.6 GO THE WAY OF HITCHHIKING / 9-DIGIT ZIP CODE P.8 MEETS OPPOSITION IN HOUSE / GENTRIFICATION P.10: THE CASE OF CLINTON HILL / ENSURING DOMESTIC TRANQUILLITY DURING SLEEPOVERS P.12 / MOVIE MAILBAG; WE DIDN'T EXPLOIT GROUPIES P.14; "NO ART"; "INTERRELATIONSHIP"; "FORKED TONGUE"; "HAS HE FORGOTTEN?" TO THE EDITOR: / LETTER OF RECOMMENDATION: FANNY PACKS P.16 / ATTENTION PASSENGERS! TO YOUR RIGHT, THIS TRIP IS ABOUT TO BECOME TRIPPY P.18 / 43 THE 1970'S INDUSTRIAL LOOK: WHAT BECAME OF HIGH TECH P.19?

2010—2018

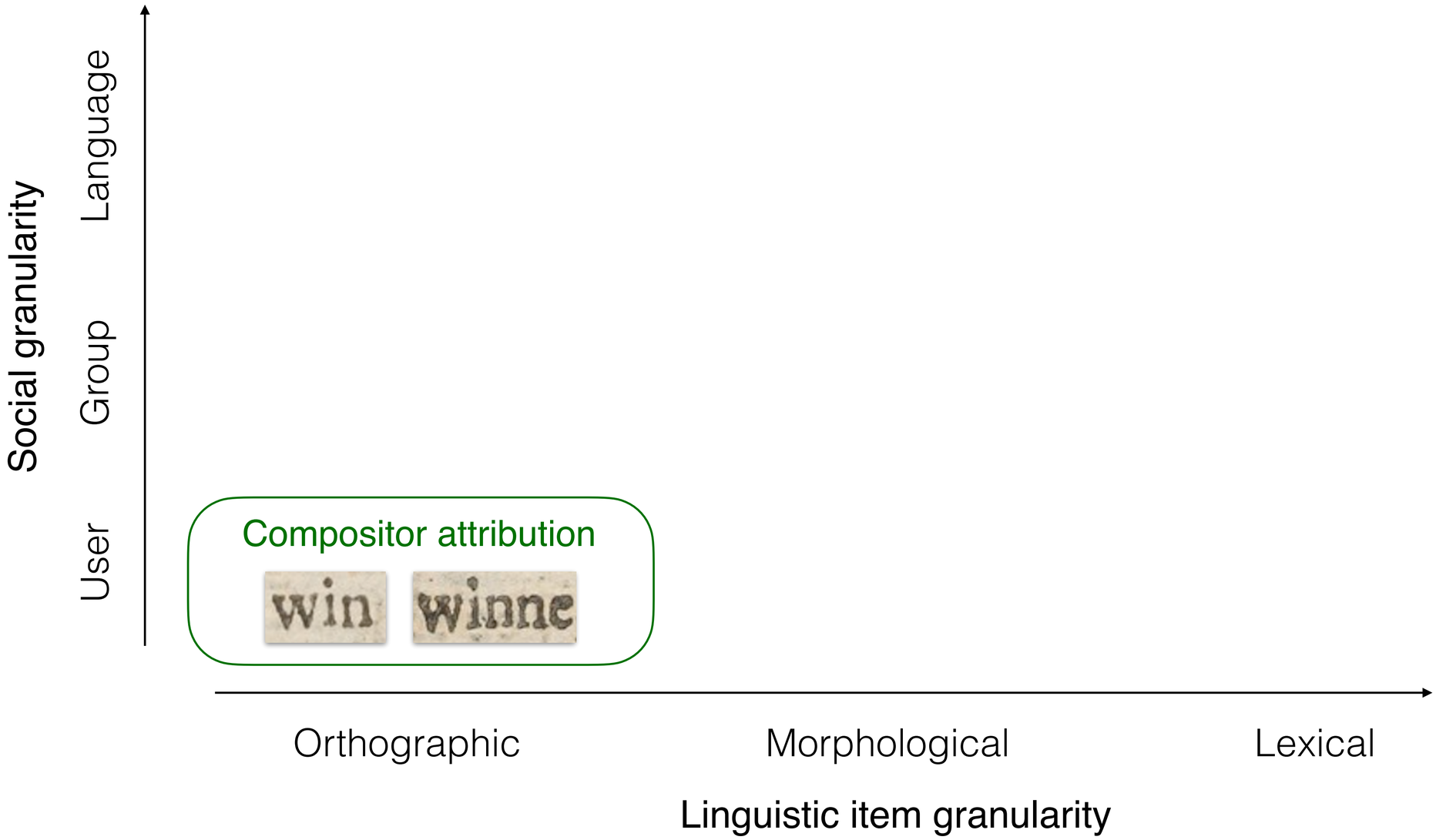
YOU SAY GOD IS DEAD? THERE'S AN APP P.2 FOR THAT / WHO IS OCCUPY P.4 WALL STREET? / #INPRAISEOFTHE HASHTAG P.6 / WHAT HAPPENS TO WOMEN WHO ARE DENIED ABORTIONS? P.7 / HOW #BLACKLIVESMATTER P.8 CAME TO DEFINE A MOVEMENT / SHE? ZE? THEY P.10? WHAT'S IN A GENDER PRONOUN / THE RAGING DUMPSTER FIRE P.12 OF THE PRESIDENT VS. OMAROSA / HOW TO FIGHT 'FAKE NEWS' P.14 (WARNING: IT ISN'T EASY) / WHAT'S BEHIND THE 'TENDER AGE' SHELTERS P.16 OPENING FOR YOUNG MIGRANTS

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

# Spectrum of phenomena

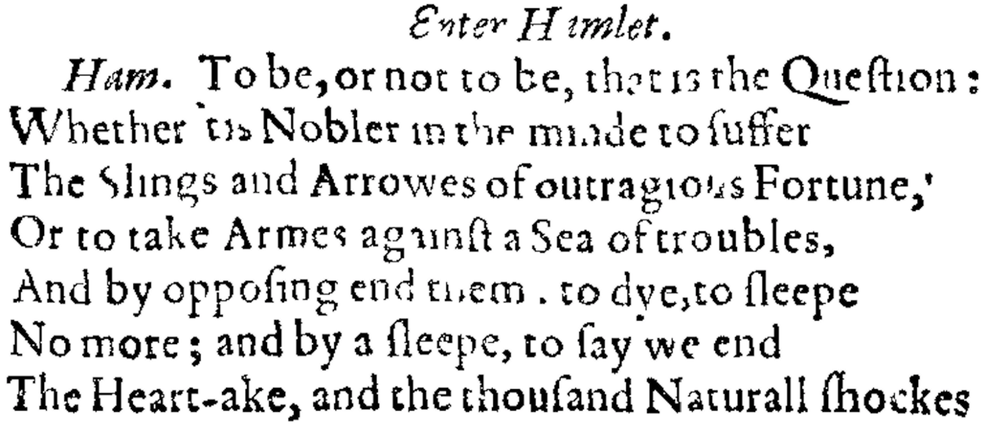


# Spectrum of phenomena



# 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

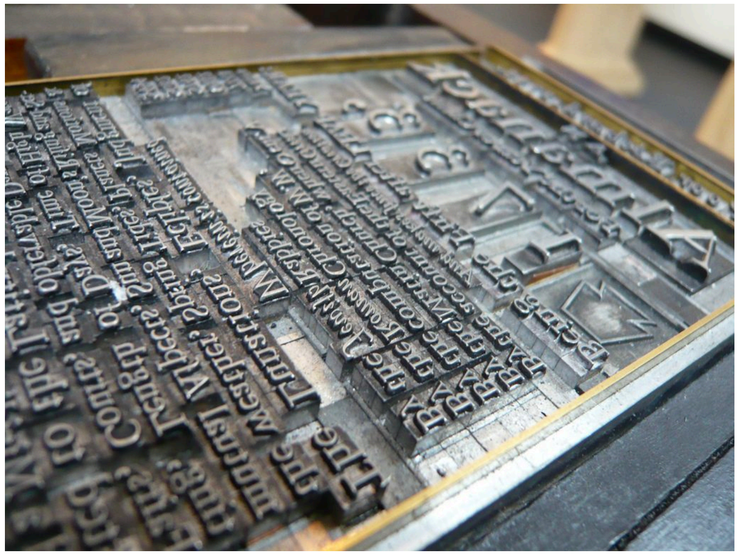


*Enter Hamlet.*

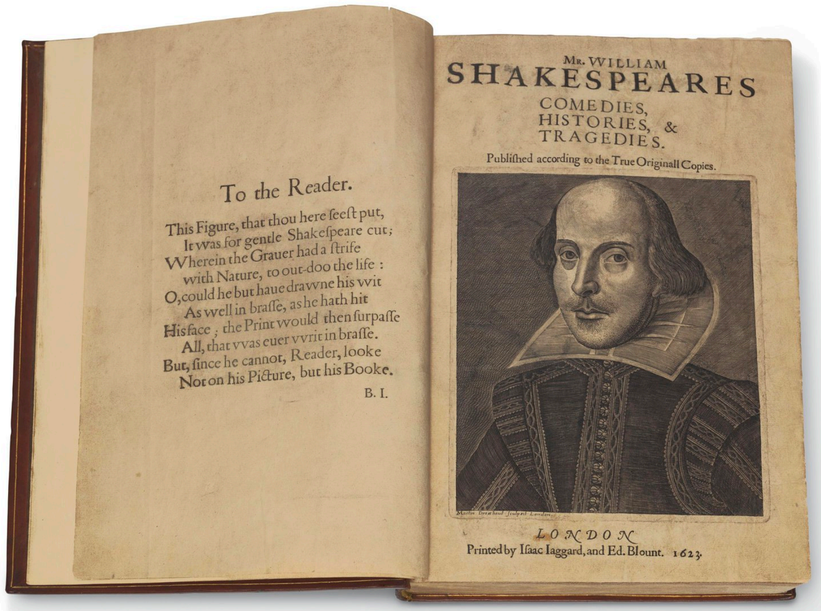
*Ham.* To be, or not to be, that is the Queſtion:
Whether 'tis Nobler in the minde to ſuffer
The Slings and Arrowes of outragious Fortune,'
Or to take Armes againſt a Sea of troubles,
And by oppoſing end them. to dye, to ſleepe
No more; and by a ſleepe, to ſay we end
The Heart-ake, and the thouſand Naturall 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
- *Compositor attribution:* grouping pages set by the same typesetter



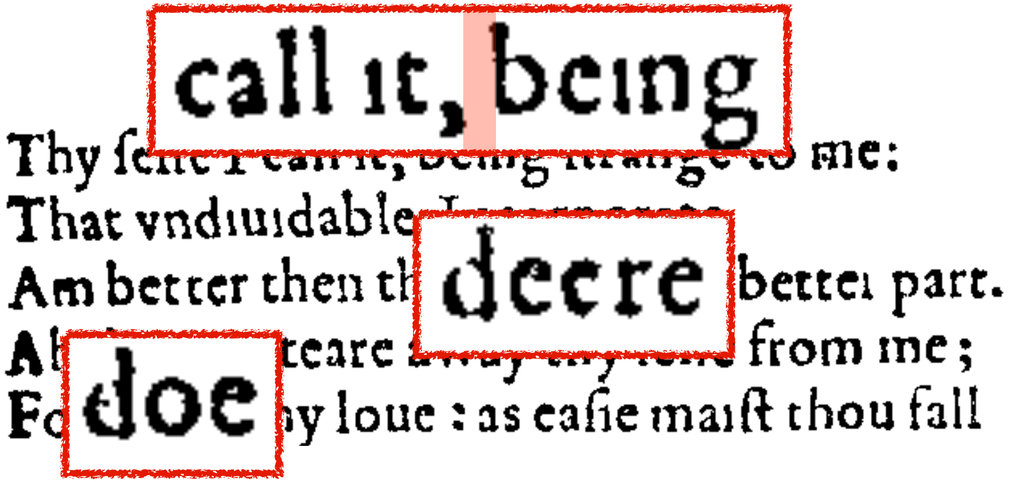
https://www.flickr.com/photos/purdman1/2875431305/



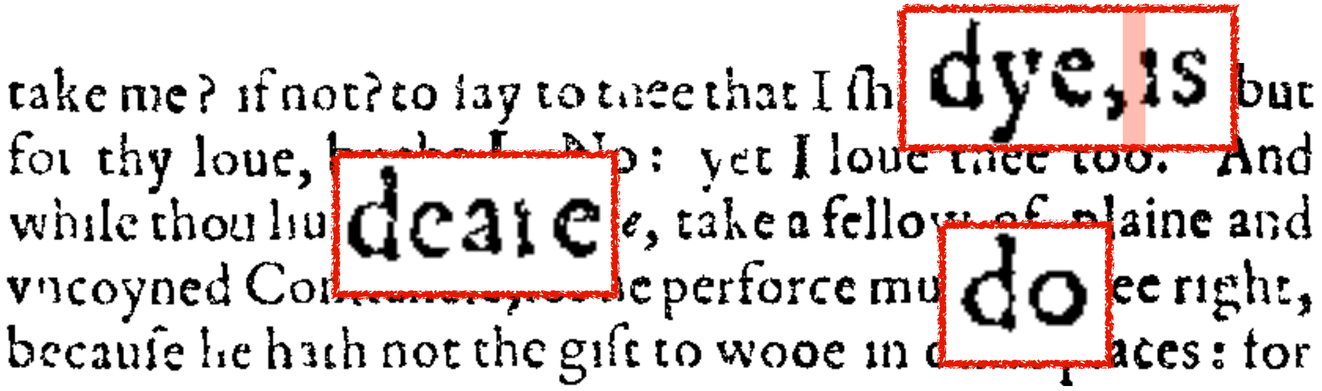

https://www.smithsonianmag.com/smart-news/shakespeares-first-folio-sells-ten-million-dollars-180976074/

# Compositor attribution

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

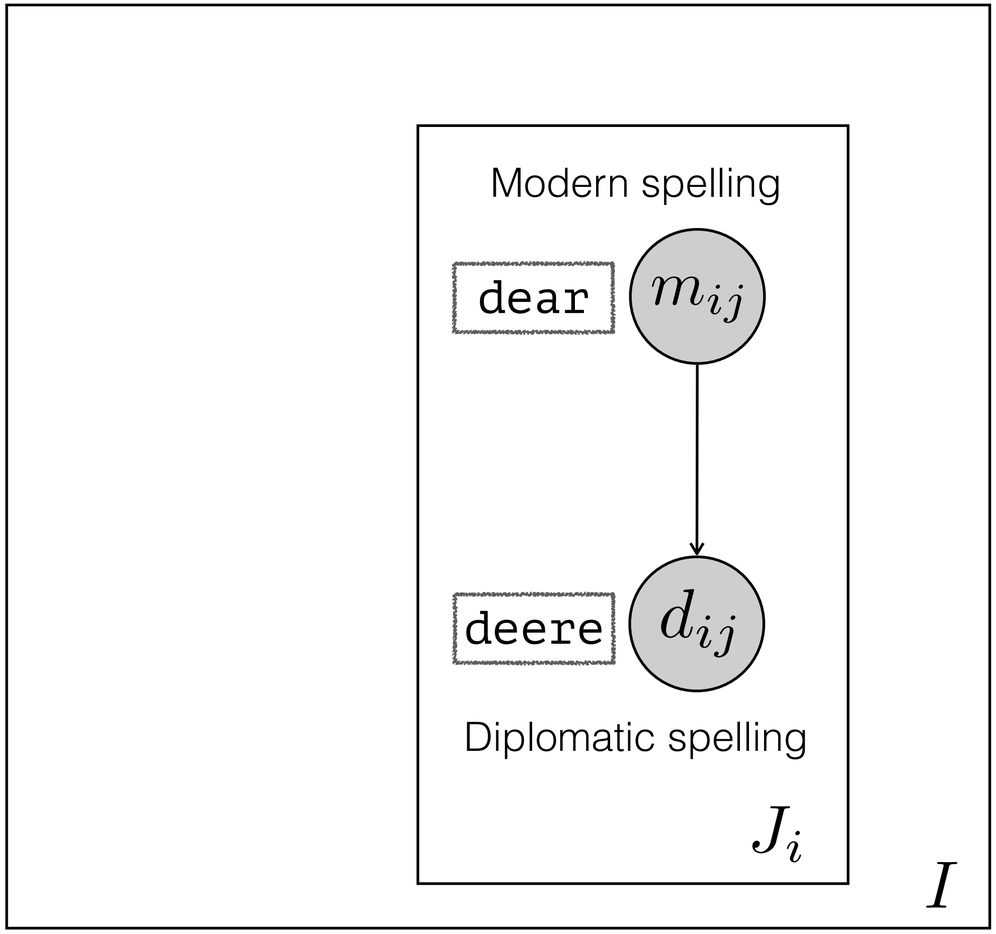



*Comedies, p. 89*

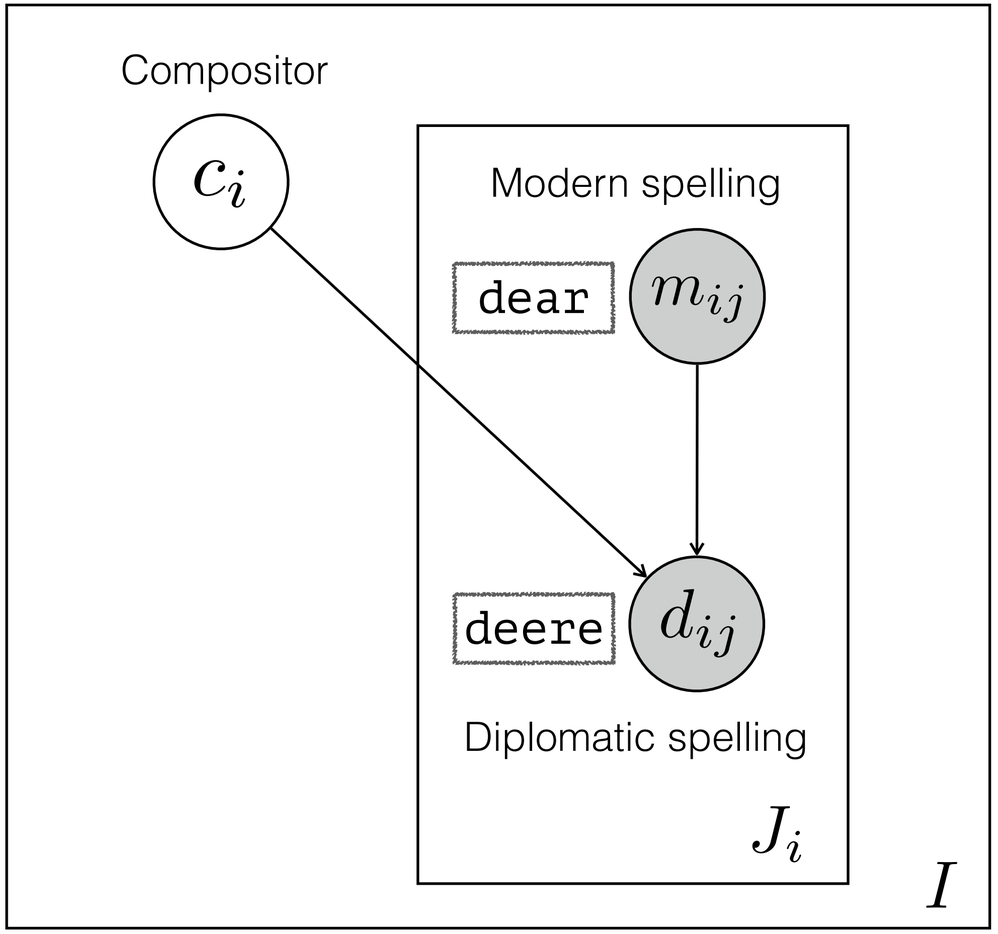



*Histories, p. 93*

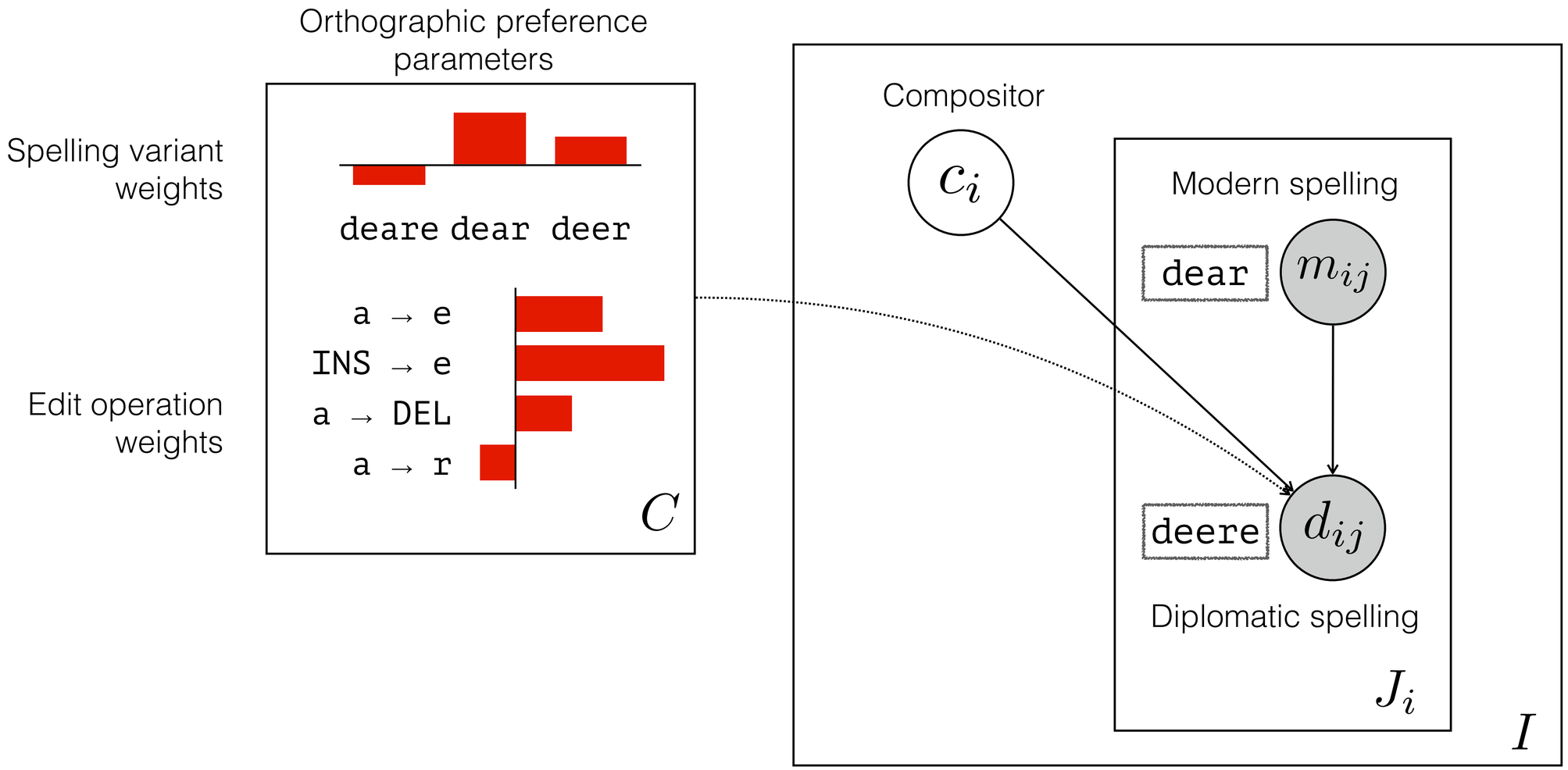
# Compositor attribution model



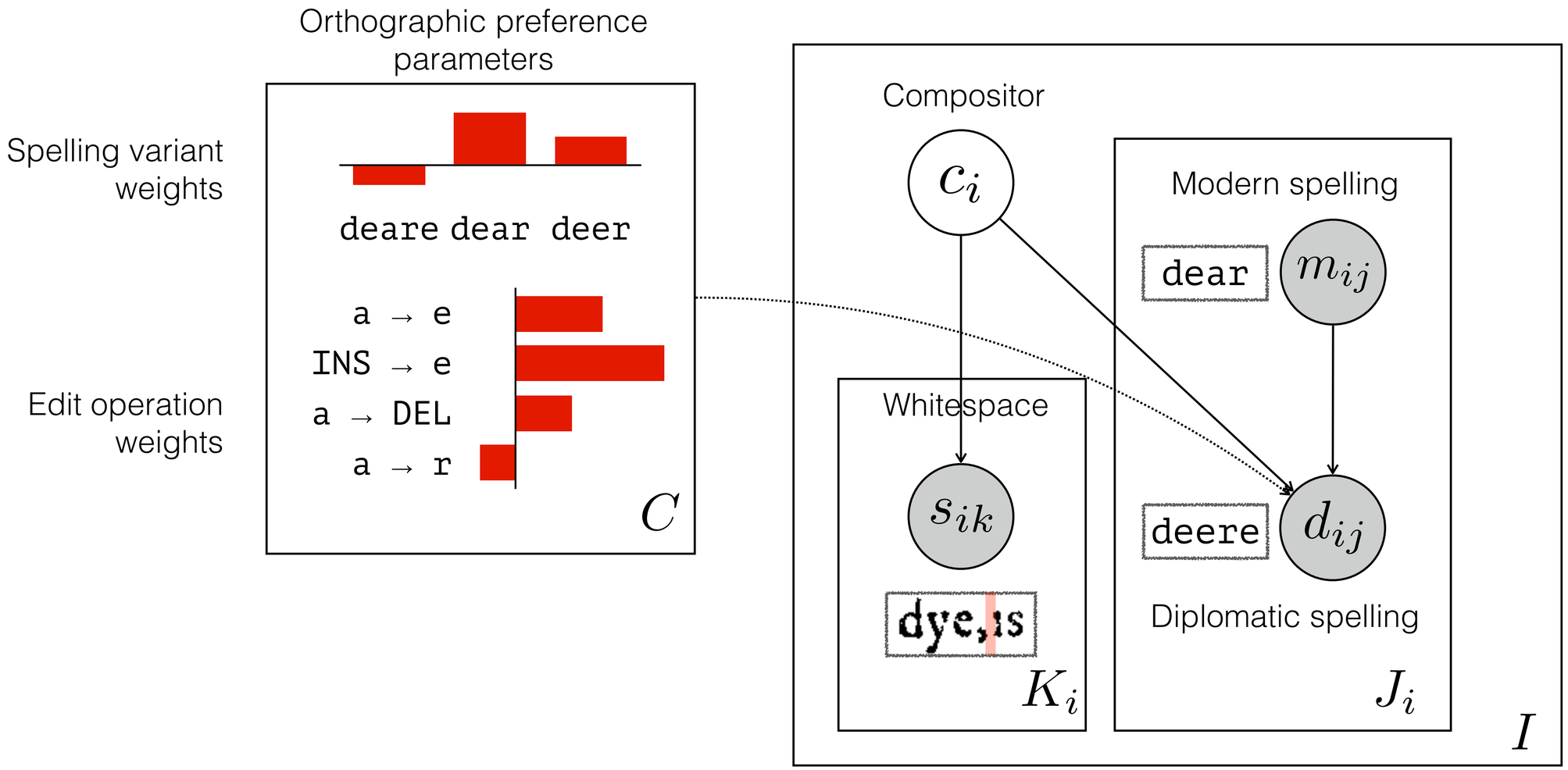
# Compositor attribution model



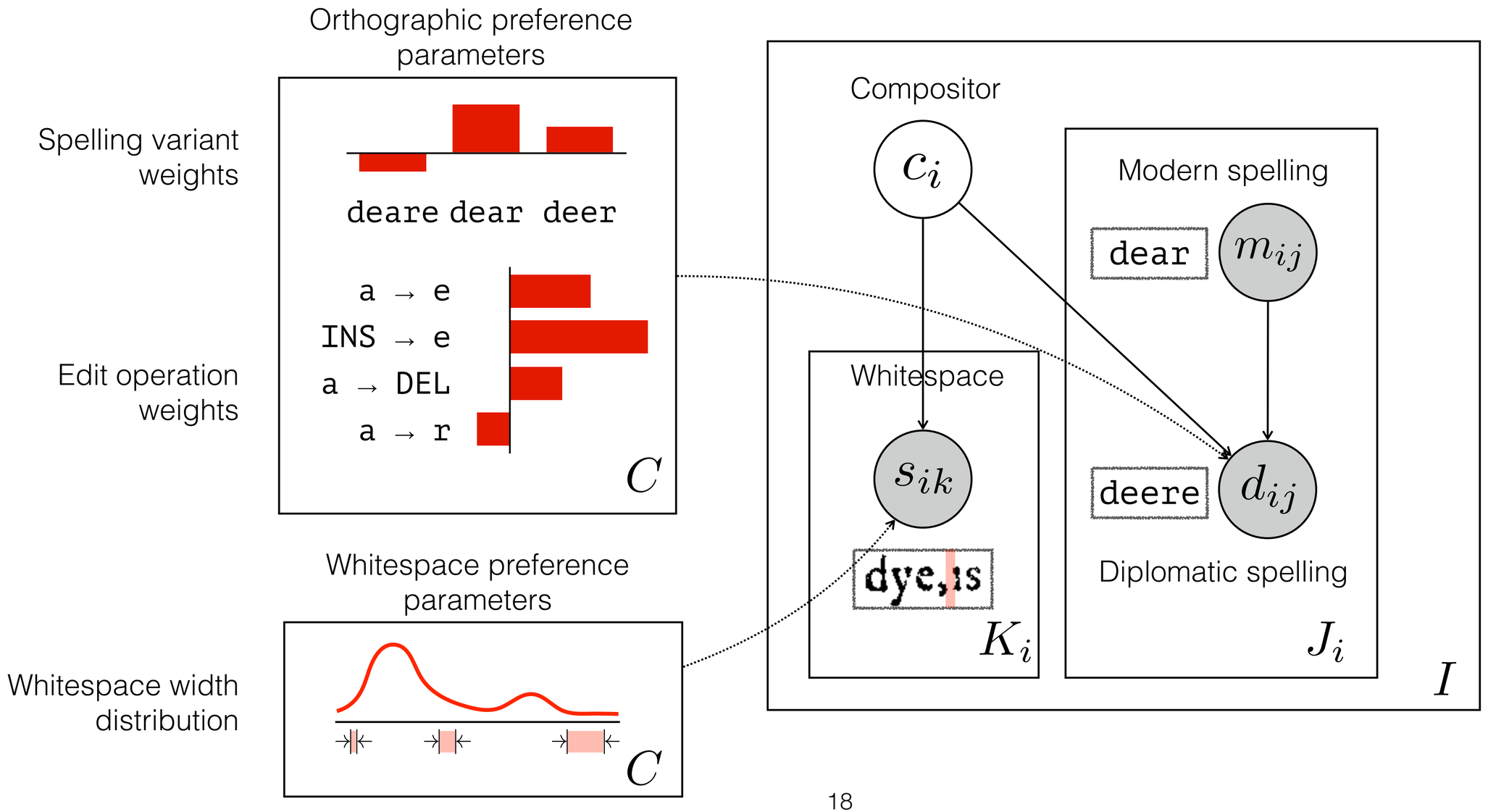
# Compositor attribution model



# Compositor attribution model



# Compositor attribution model

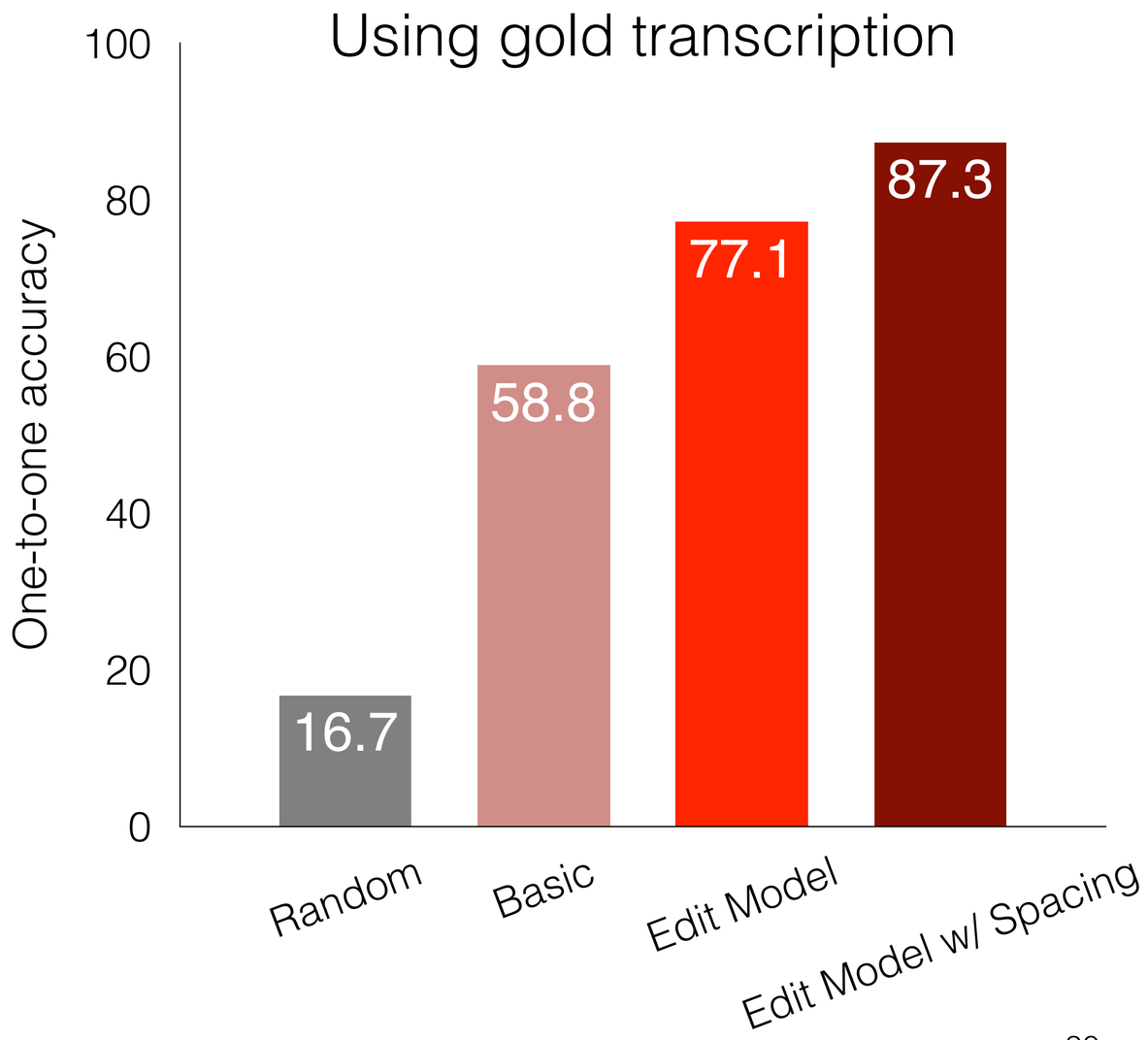


# Training and evaluation

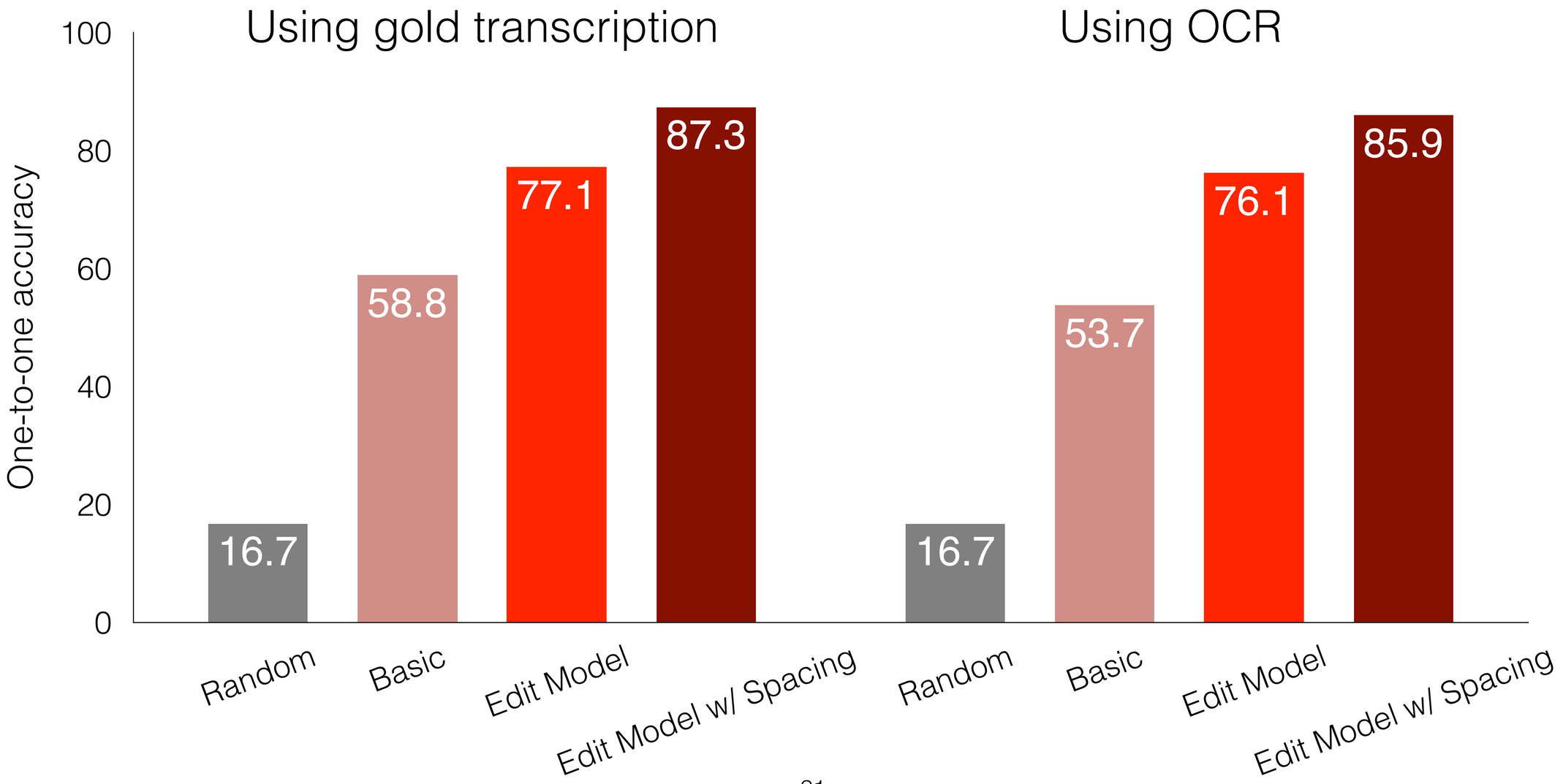
- We learn compositor preferences with EM algorithm
  - Setting the maximum number of compositors in advance
  - Inference performed by independent argmax for each page
- Evaluate by comparing against the authoritative attribution
  - Match recovered clusters to ground-truth ones using Hungarian algorithm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

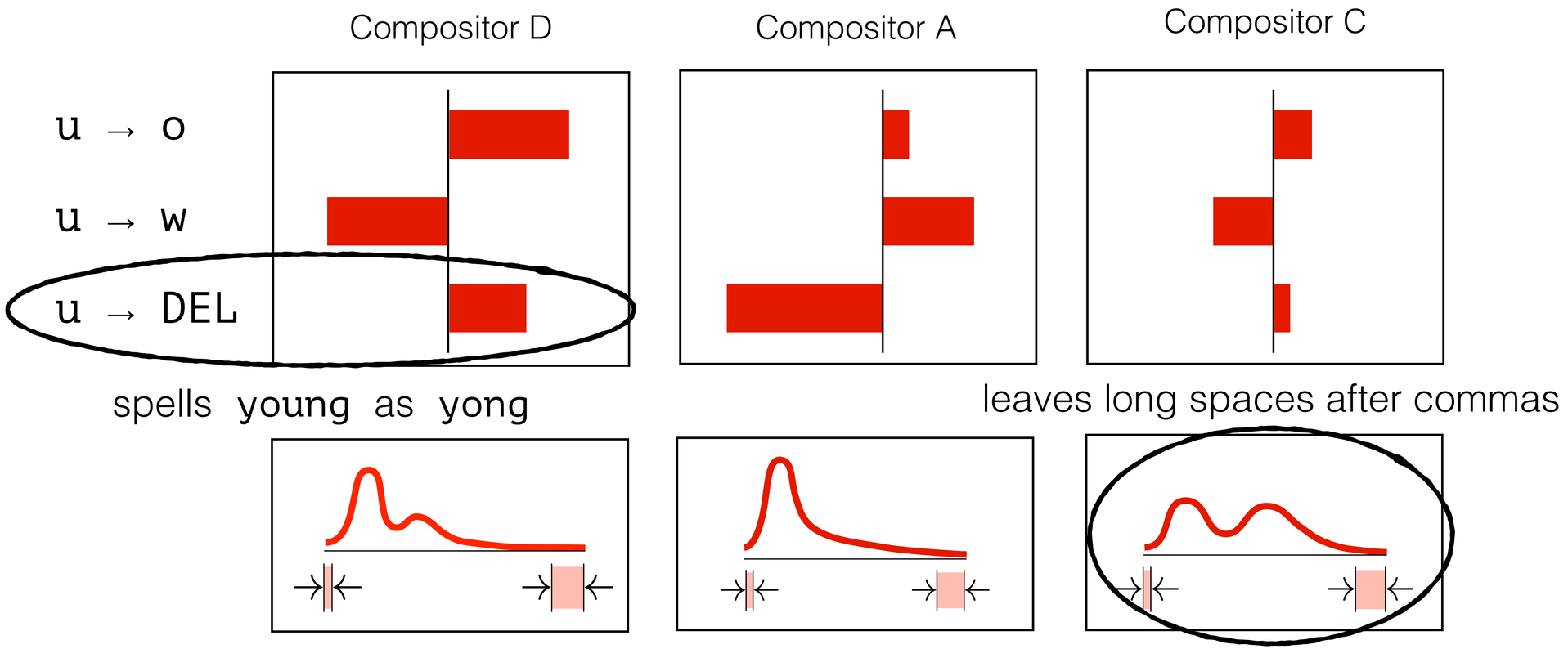


# Experimental results



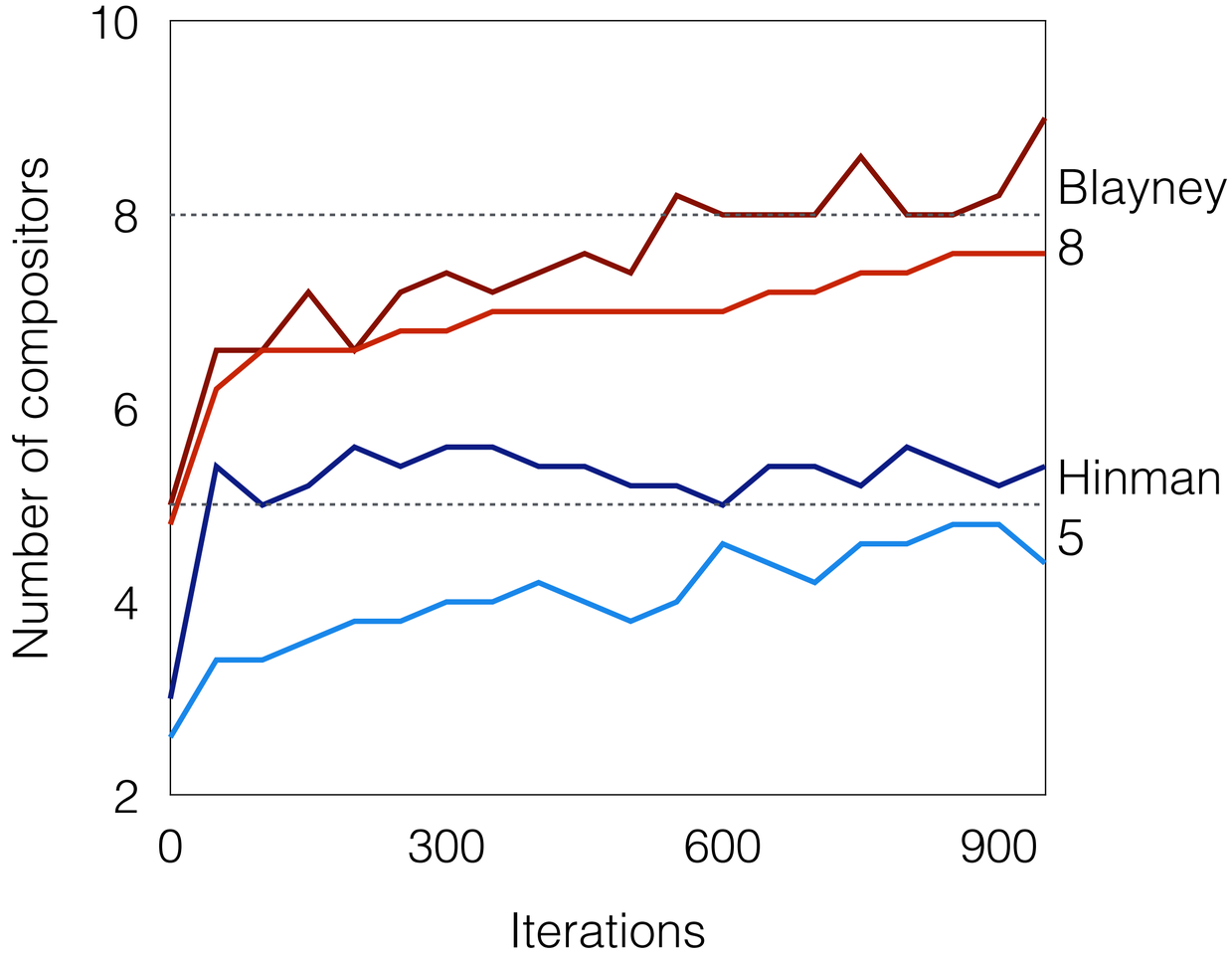
# Learned behaviors

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

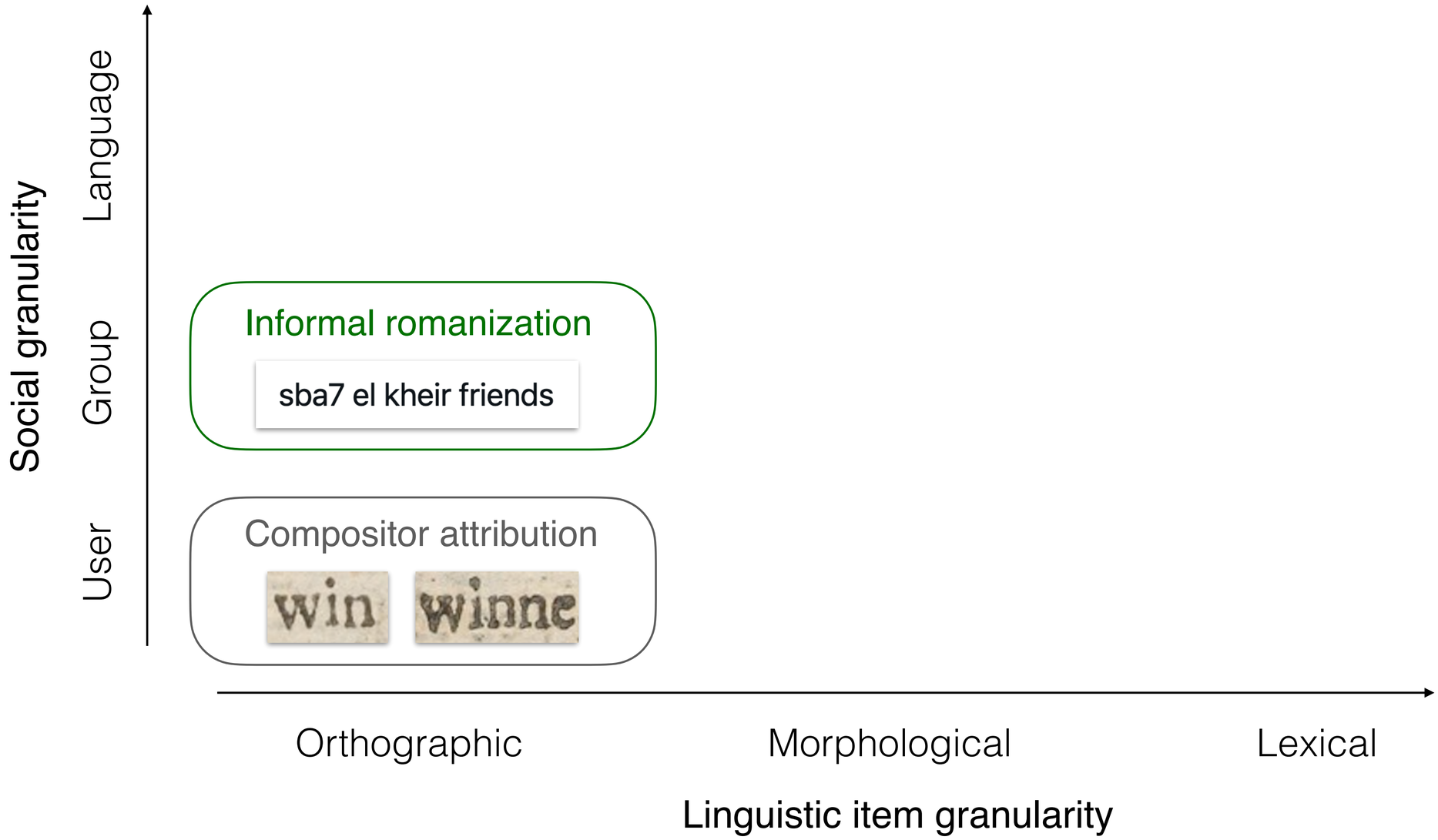


# Number of compositors

- Bibliographers posited different numbers of compositors
  - 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* | *chelovek, 4elovek, ceJIoBek, …* |
| Arabic | صباح | *saba7, sba7, sabah, …* |
| Greek | *ξένος* | *xenos, ksenos, 3enos, …* |

# Informal romanization

- Idiosyncratic representation: character subsititutions up to the user

| | | |
|---|---|---|
| Russian | *человек* | *chelovek, 4elovek, ceJloBek, …* |
| Arabic | صباح | *saba7, sba7, sabah, …* |
| Greek | *ξένος* | *xenos, ksenos, 3enos, …* |

# Informal romanization

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

| | | |
|---|---|---|
| Russian | *человек* | *chelovek, 4elovek, ceJloBek, …* |
| Arabic | صباح | *saba7, sba7, sabah, …* |
| Greek | *ξένος* | *xenos, ksenos, 3enos, …* |

# Character similarity

- Phonetic: similarity between sounds associated with characters
  - Out-of-context grapheme-phoneme association: г~/g/→g
  - Phoneme produced in context: انتي /enti/→enty, صباح /sabaħ/→saba7
- Visual: similarity between glyph shapes
  - Characters expressed by same or similar glyph: a~/a/→a, г~/g/→r
  - Characters can map to bi-/trigraphs: ы→bl
  - Can be conditioned on a transformation: л→v, 3←ع
  - Can be applied to a part of a glyph: 2←أ

# Character alignment

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

| Alphabet | Abjad (consonantal) | Abugida (alphasyllabary) |
|---|---|---|
| хорошо | كريم | ಬೆಳಗಿತು |
|  |  |  |
| xorosho | kareem | belagitu |
| ~ one-to-one | ~ one-to-one + null | ~ 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)

ಬೆಳಗಿತು

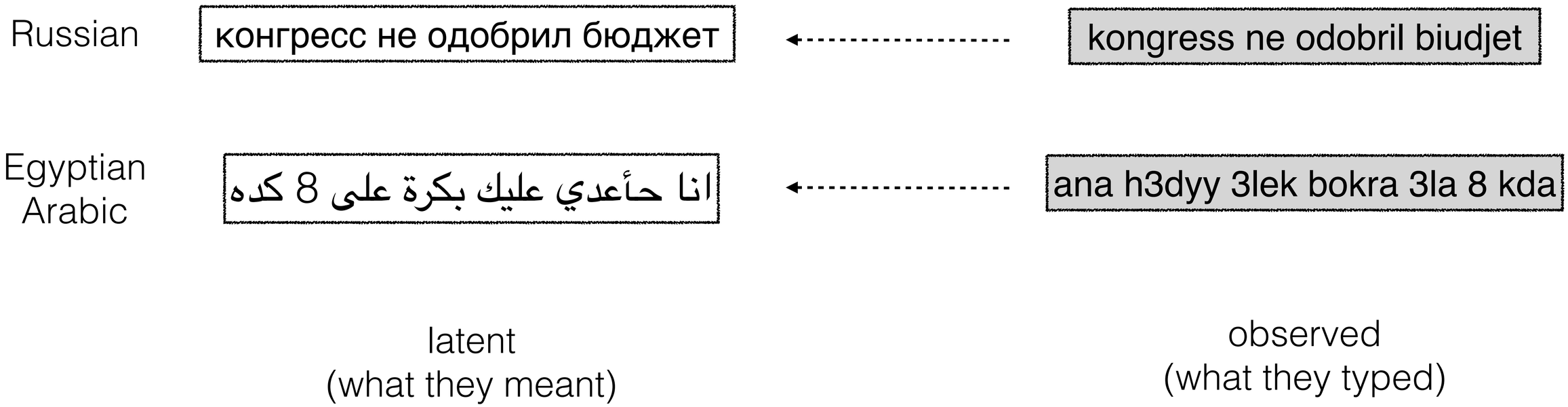
*Unicode:* ಬ ೆ ಳ ಗ ಿ ತ ು

belagitu

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

# Task framing

- Convert romanized text to the conventional orthography of the language



# Task framing

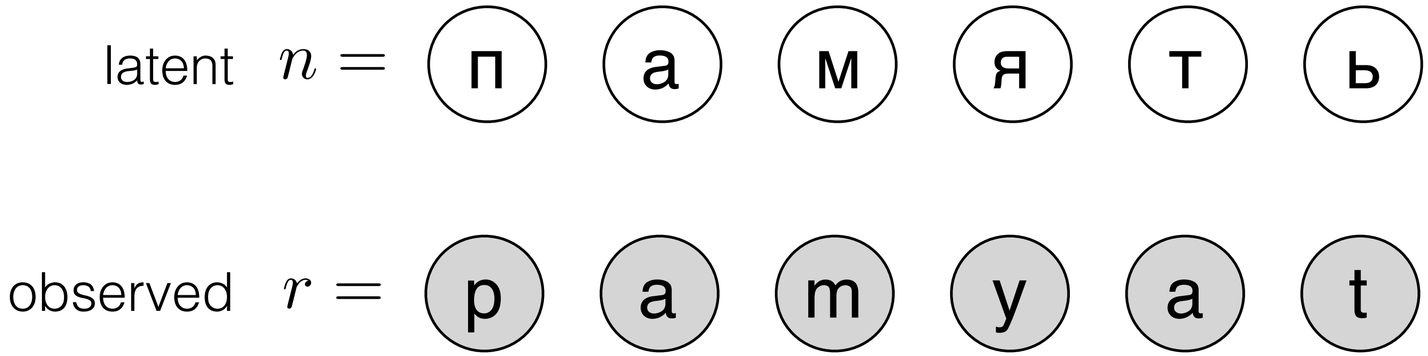
- Parallel data does not occur naturally ⇒ unsupervised learning
- Perceptions of similarity are shared across users and even languages!
- Hypothesis: inductive bias encoding these similarity notions provides signal that can somewhat approximate human supervision
  - We rely on manually-curated resources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




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

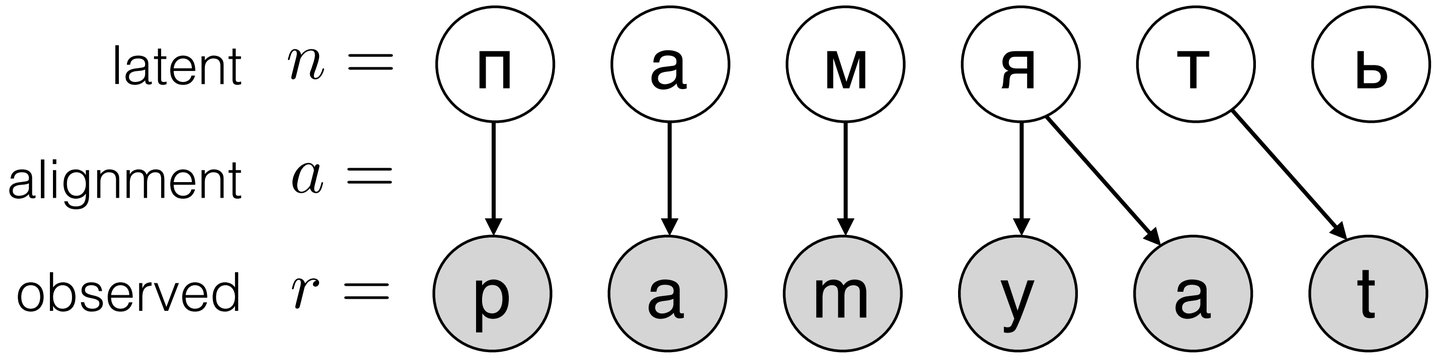
all possible native script sequences

transition probabilities

emission probabilities

prior on parameters

# Noisy-channel model




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

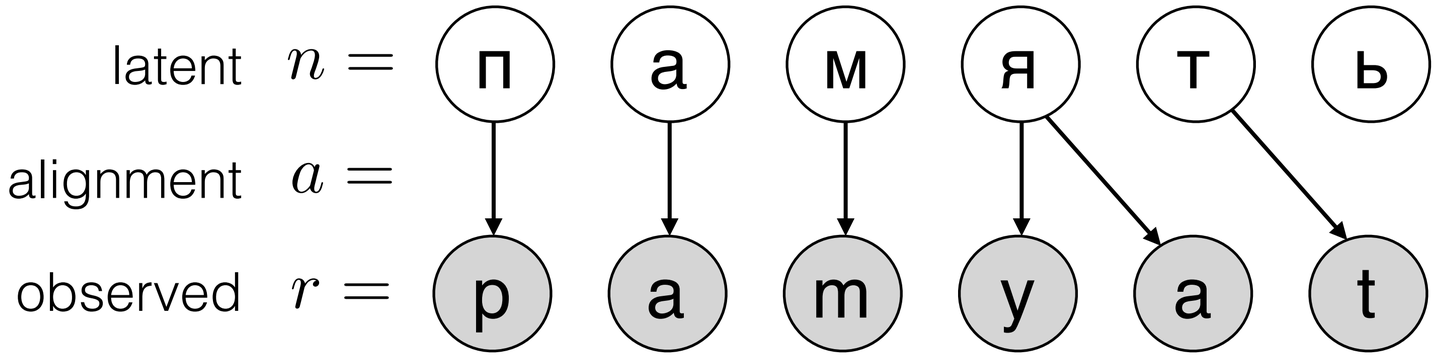
all possible native script sequences and alignments

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



https://en.wikipedia.org/wiki/Phonetic_keyboard_layout

https://arabic.omaralzabir.com/

# Phonetic bias

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



https://en.wikipedia.org/wiki/Phonetic_keyboard_layout

https://arabic.omaralzabir.com/

# Visual bias

- ‘Visual’ priors: mappings off the Unicode confusables list

- Designed to combat spoofing attacks

| | | | | | | |
|---|---|---|---|---|---|---|
| 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 | ABAF CHEROKEE SMALL LETTER TLI |
| o | o | σ | o | o | ᗤ | ه |
| 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 |

mcqll.org mcqll.org

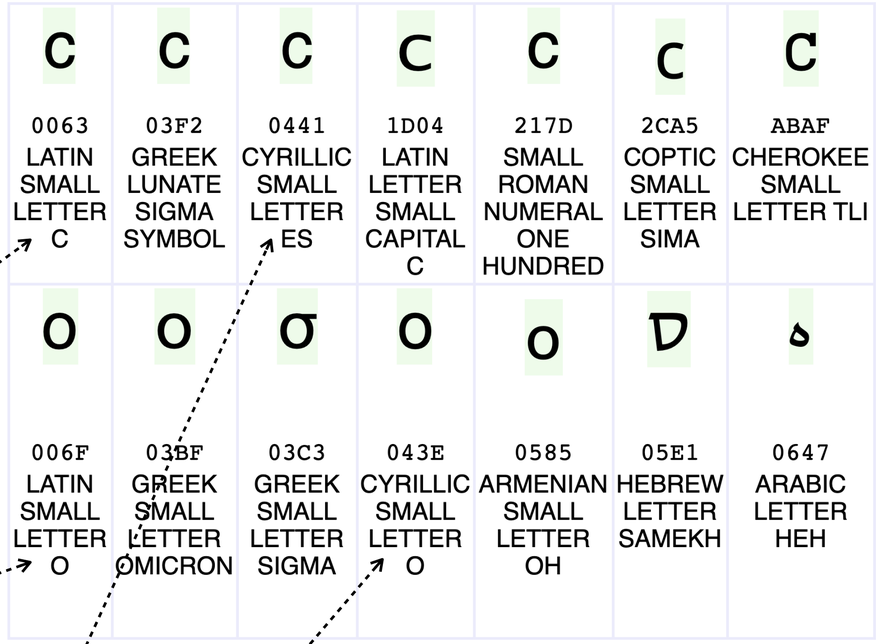


https://util.unicode.org/UnicodeJsps/confusables.jsp

# Visual bias

- ‘Visual’ priors: mappings off the Unicode confusables list

- Designed to combat spoofing attacks



| | | | | | | |
|---|---|---|---|---|---|---|
| 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 | ABAF CHEROKEE SMALL LETTER TLI |
| o | o | σ | o | o | ᗡ | ه |
| 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 |

mcqll.org


mcqll.org

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



https://util.unicode.org/UnicodeJsps/confusables.jsp

# Visual bias

- ‘Visual’ priors: mappings off the Unicode confusables list

- Designed to combat spoofing attacks

- Hardly any mappings for Arabic!

| | | | | | | |
|---|---|---|---|---|---|---|
| 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 | ABAF CHEROKEE SMALL LETTER TLI |
| o | o | σ | o | o | ᗤ | ه |
| 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 |

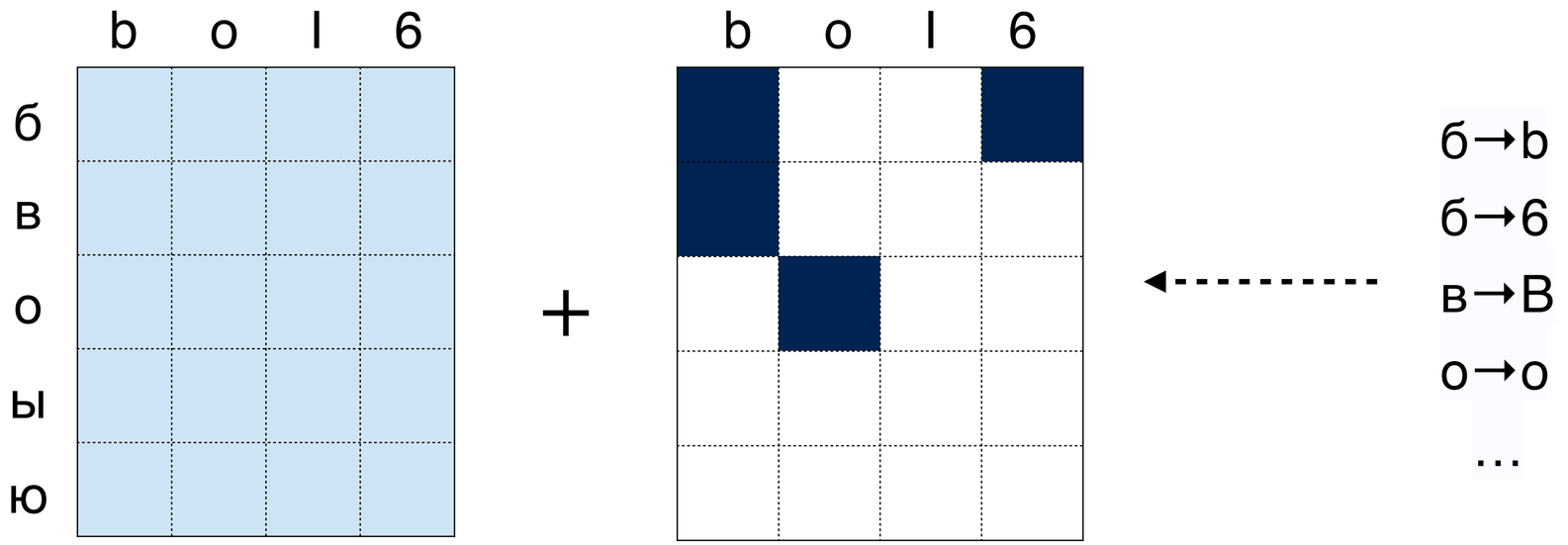


https://util.unicode.org/UnicodeJsps/confusables.jsp

# Informative priors

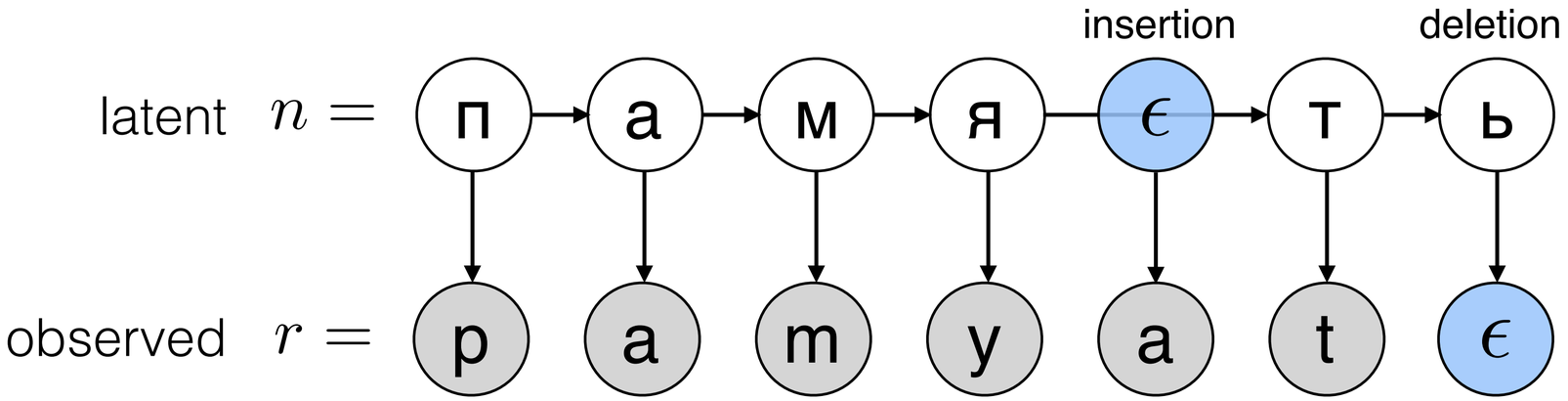
- Use mappings of similar characters as priors on emission parameters

$$c_r | c_n \sim \mathrm{Mult}(\theta_{c_n}) \qquad \theta \sim \mathrm{Dir}(\alpha)$$



# Noisy-channel model

- Representing latent alignments via insertions and deletions




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

transition probabilities

emission probabilities

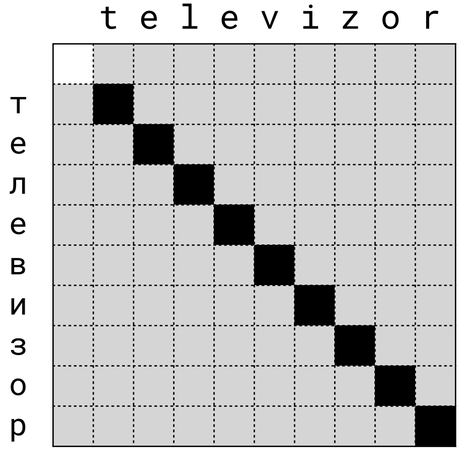
prior on parameters

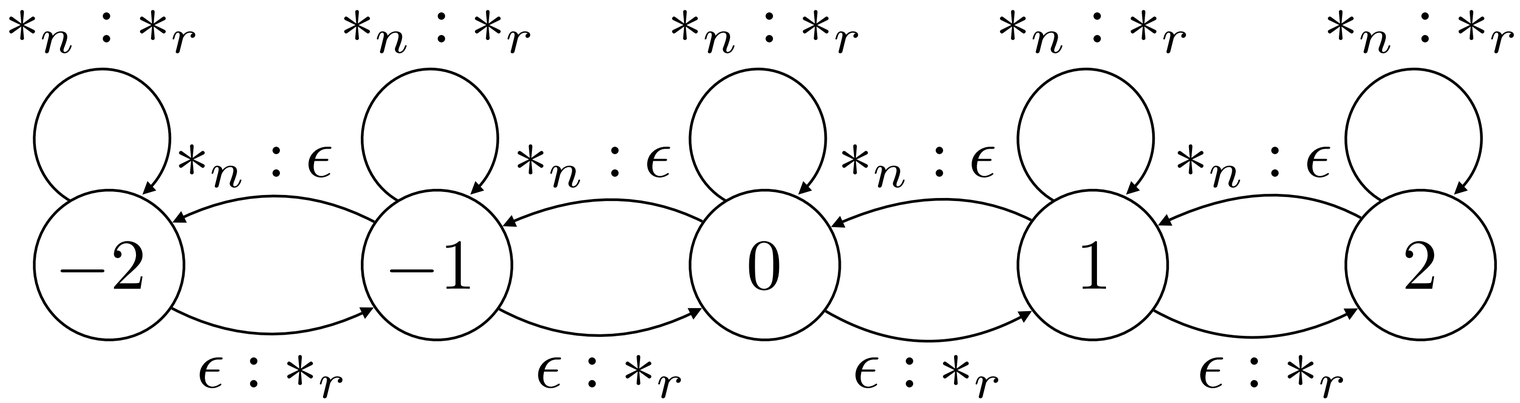
# WFST cascade

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

# Emission model

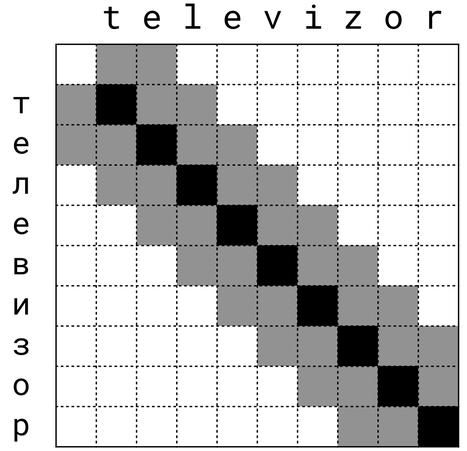
- Needs to support substitutions, insertions and deletions
- Fixed limit on delay: |# of insertions - # of deletions|

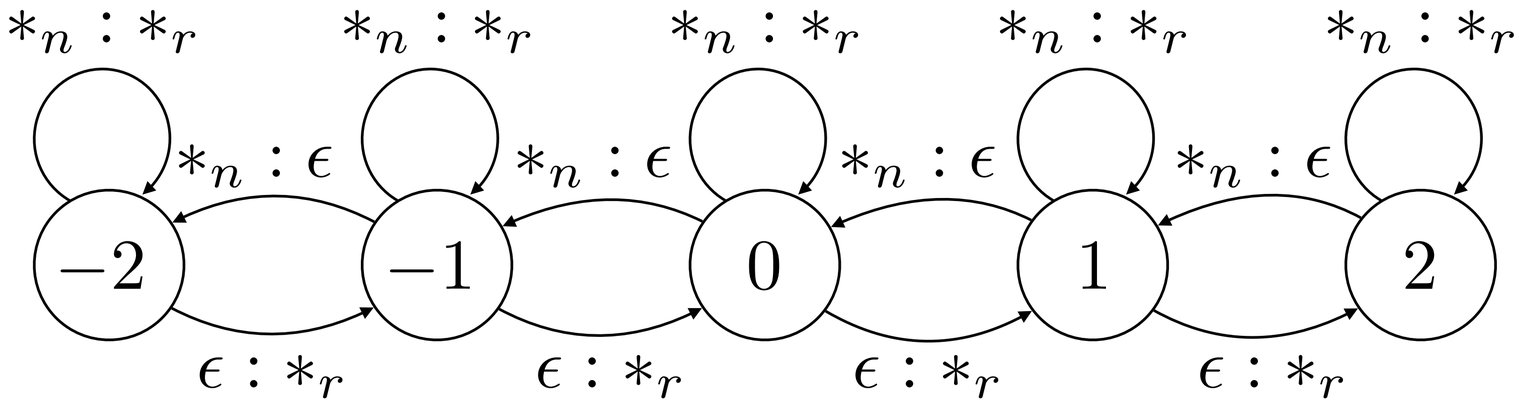


# Emission model

- Needs to support substitutions, insertions and deletions
- Fixed limit on delay: |# of insertions - # of deletions|



# WFST cascade

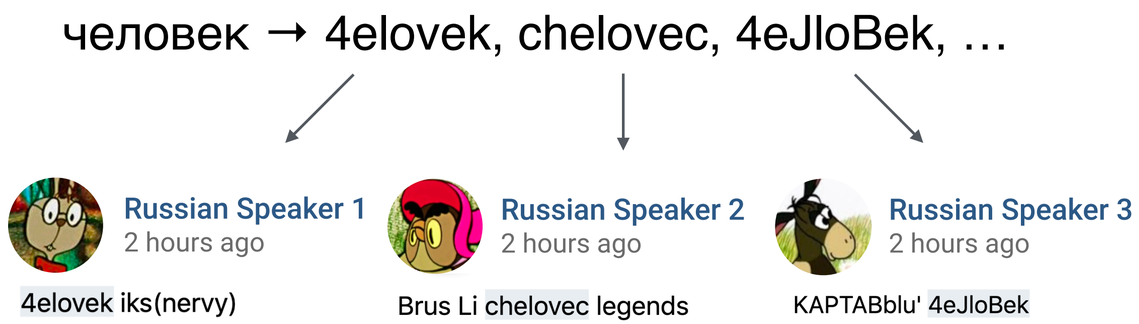
- Transition WFSA
  - 6-gram LM built with OpenGrm (Roark et al., 2012)
- Emission WFST
  - Supports all substitutions, insertions and deletions
- Trained with EM algorithm
  - OpenFst (Allauzen et al., 2007)
  - Speedup tricks: stepwise training, curriculum learning, pruning…

# Datasets

- Arabic: LDC BOLT dataset (Bies et al., 2014)
  - Arabizi SMS/chat dialogs, converted to CODA (Habash et al., 2012)

Saba7 el 5eir!
Ezayeeky?

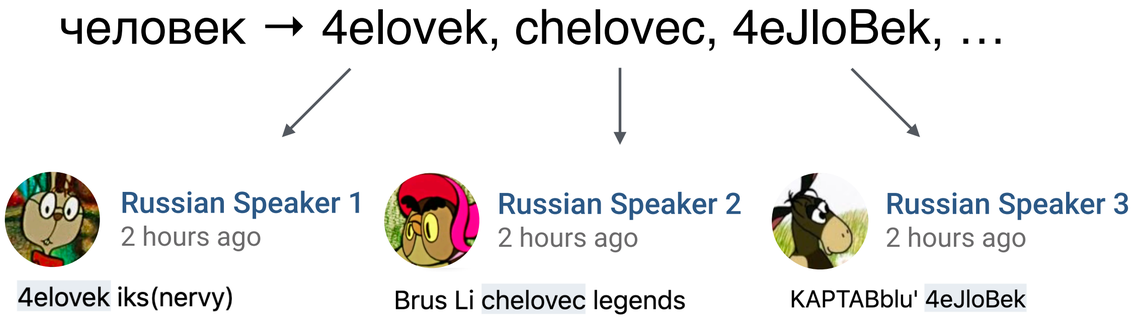
- Russian:
  - Romanized: collected and partly annotated data from social media




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

# Data collection

- Romanizations of common words used as queries (Darwish, 2014)



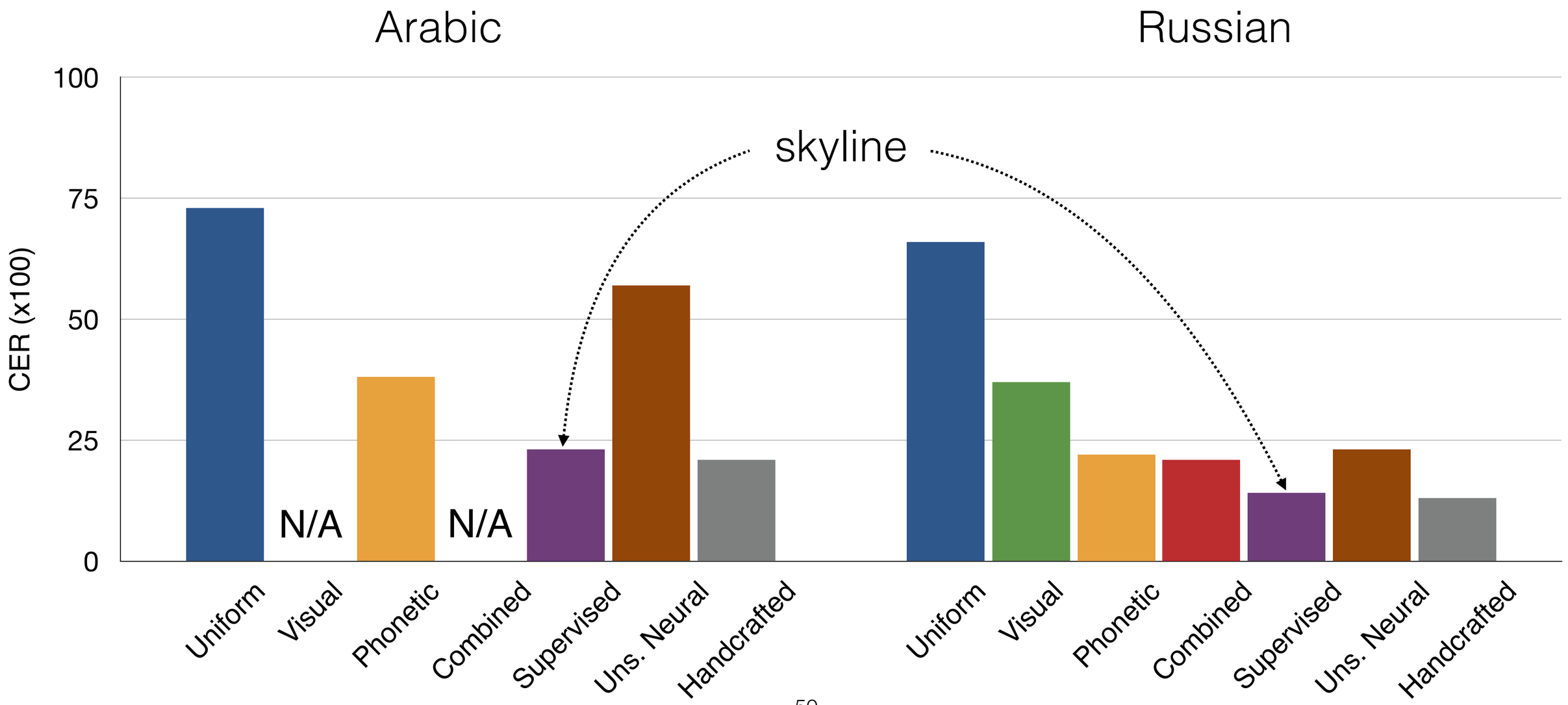

- Annotated validation and test with minor error correction

Source: proishodit s prirodoy 4to to very very bad

Filtered: proishodit s prirodoy 4to to <…>

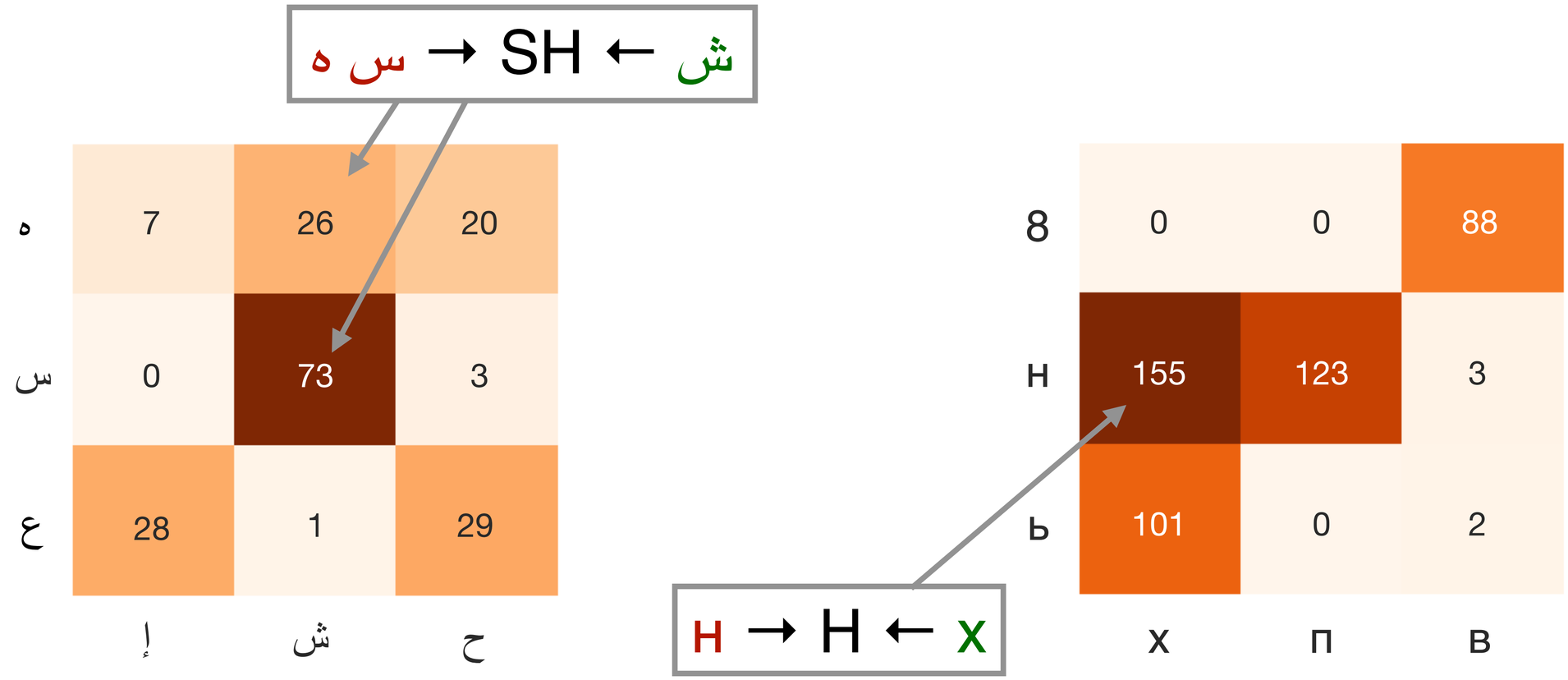
Target: происходит с природой что-то <…>

# Experimental results



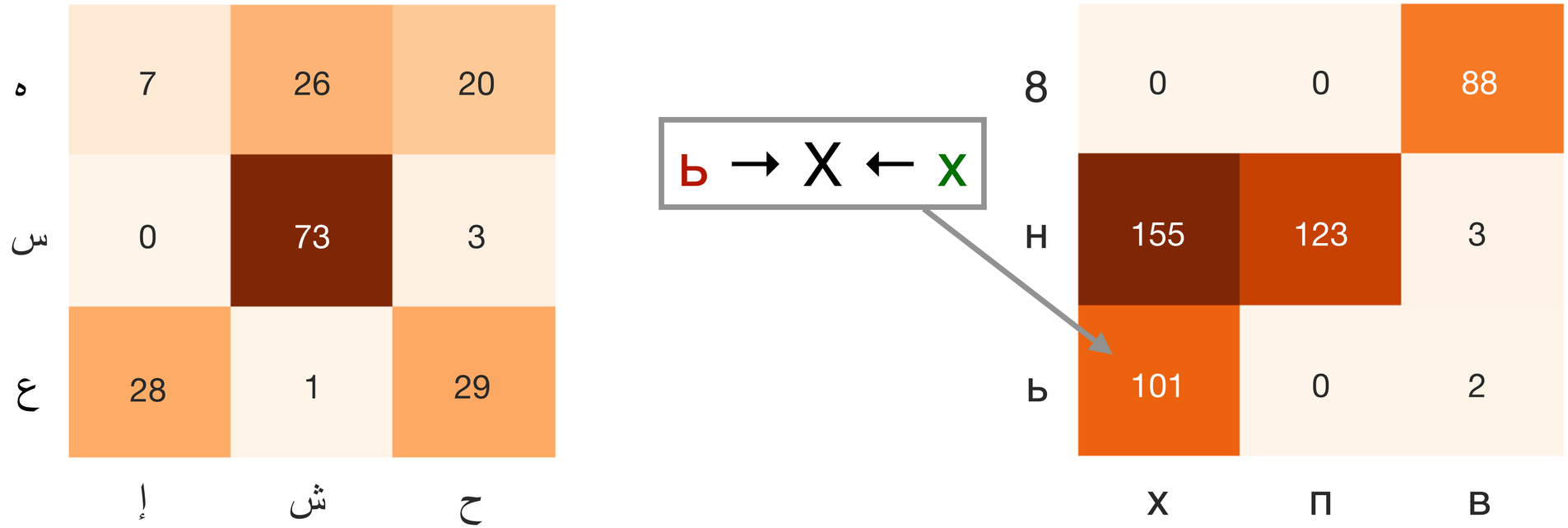
# Error analysis

- Many errors are likely due to a weak language model



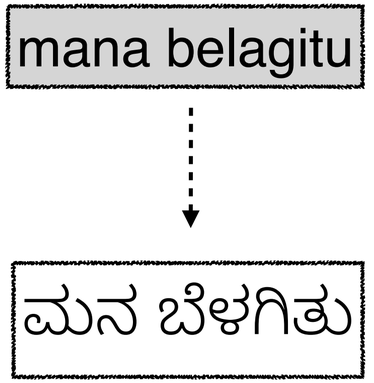
# Error analysis

- Some errors are caused by spurious mappings in priors



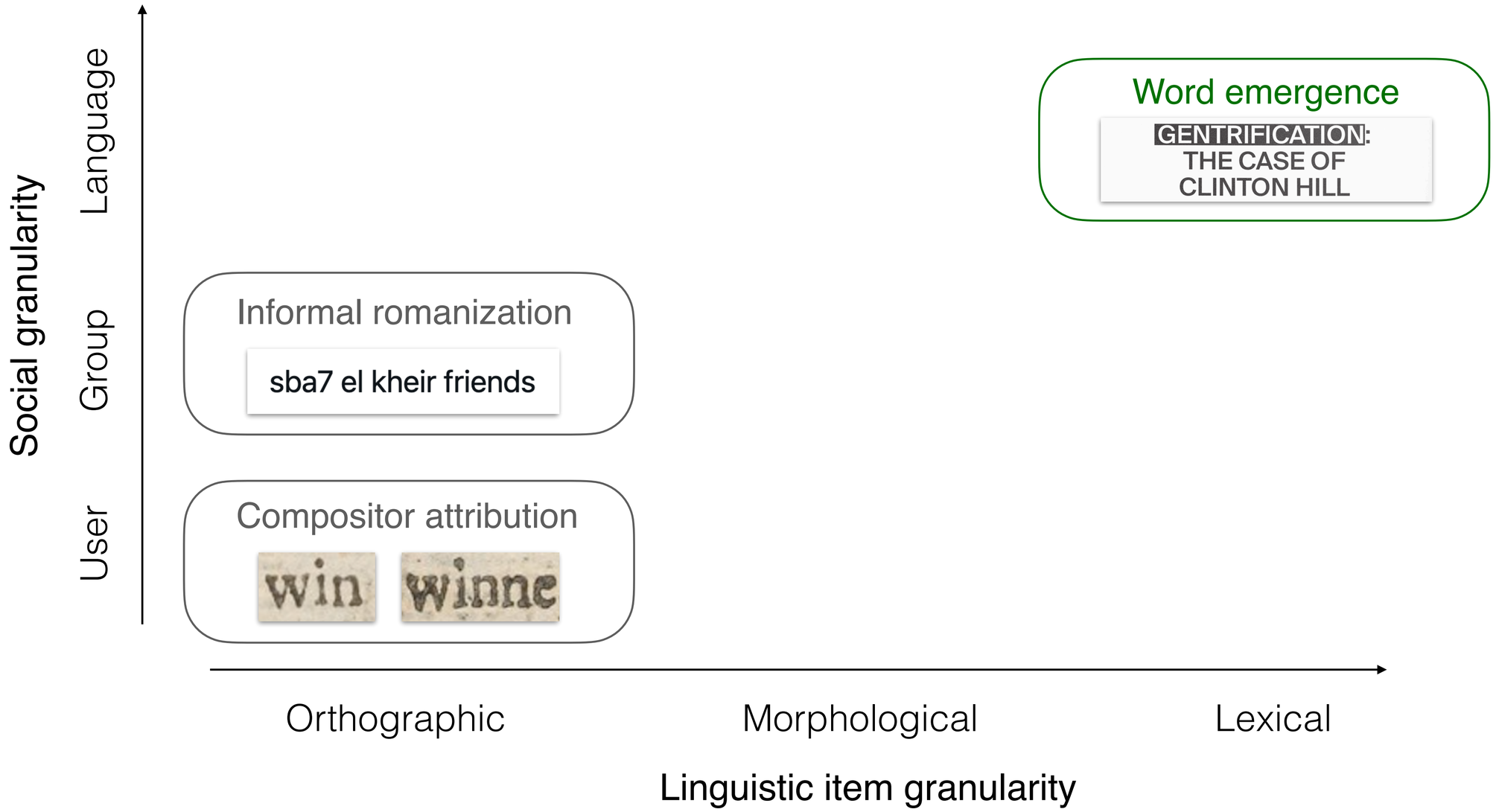
# Error analysis

- 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




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
- What are the semantic factors that characterize word emergence?

**WORD OF THE YEAR**

- **antiracism**: the practice of actively working to prevent or combat racism
- **Before Times**: the time before the beginning of the pandemic
- **BIPOC**: acronym for *Black, Indigenous and People of Color*
- **Blursday**: humorous indication of difficulty in determining what day of the week it is
- **covidiot**: a person who foolishly ignores COVID-19 protocols
- **doomscrolling**: obsessively scanning social media and websites for bad news

https://www.americandialect.org/nominations-for-words-of-the-year-2020

# Hypotheses

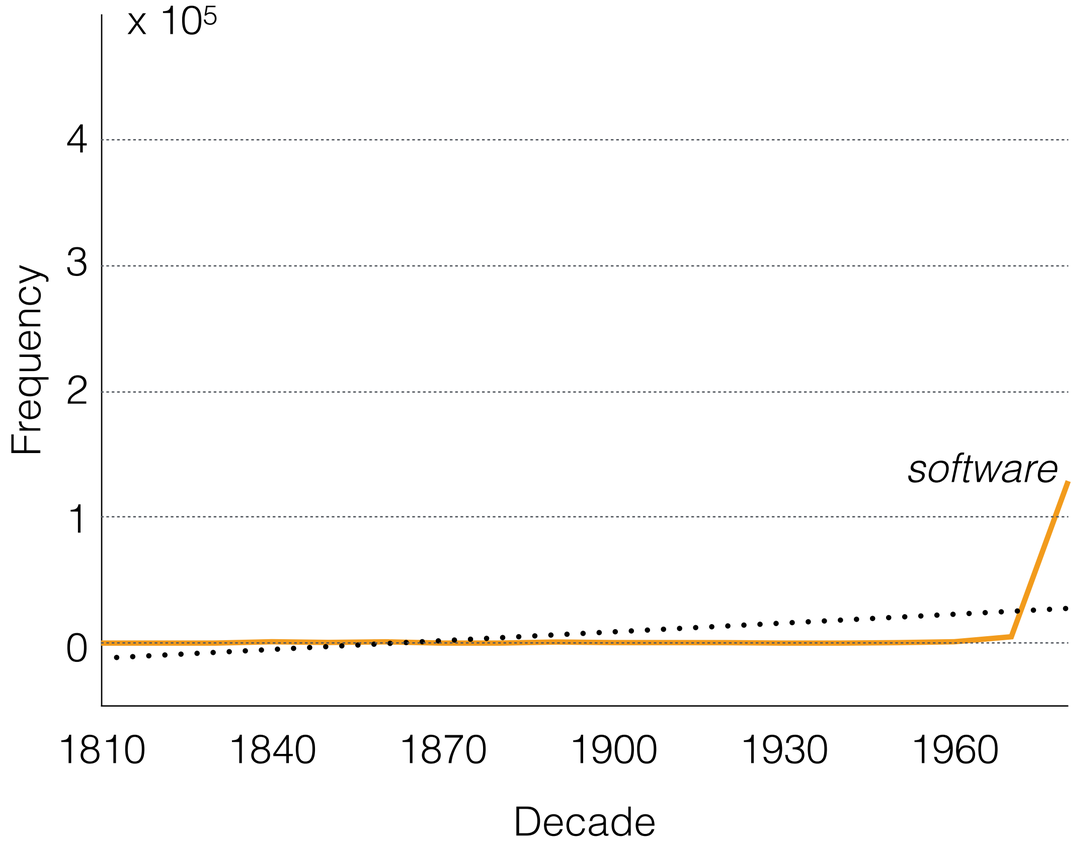
- Supply: neologisms are more likely to emerge in sparser areas of the semantic space
  - Semantic space tends towards uniformity (Bréal, 1904)
  - New words emerge to fill in ‘semantic gaps’
- Demand: neologisms are more likely to emerge in semantic neighborhoods of growing popularity
  - Word frequency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

# Diachronic corpora

- Historical (COHA) and modern (COCA) American English corpora

# Diachronic corpora

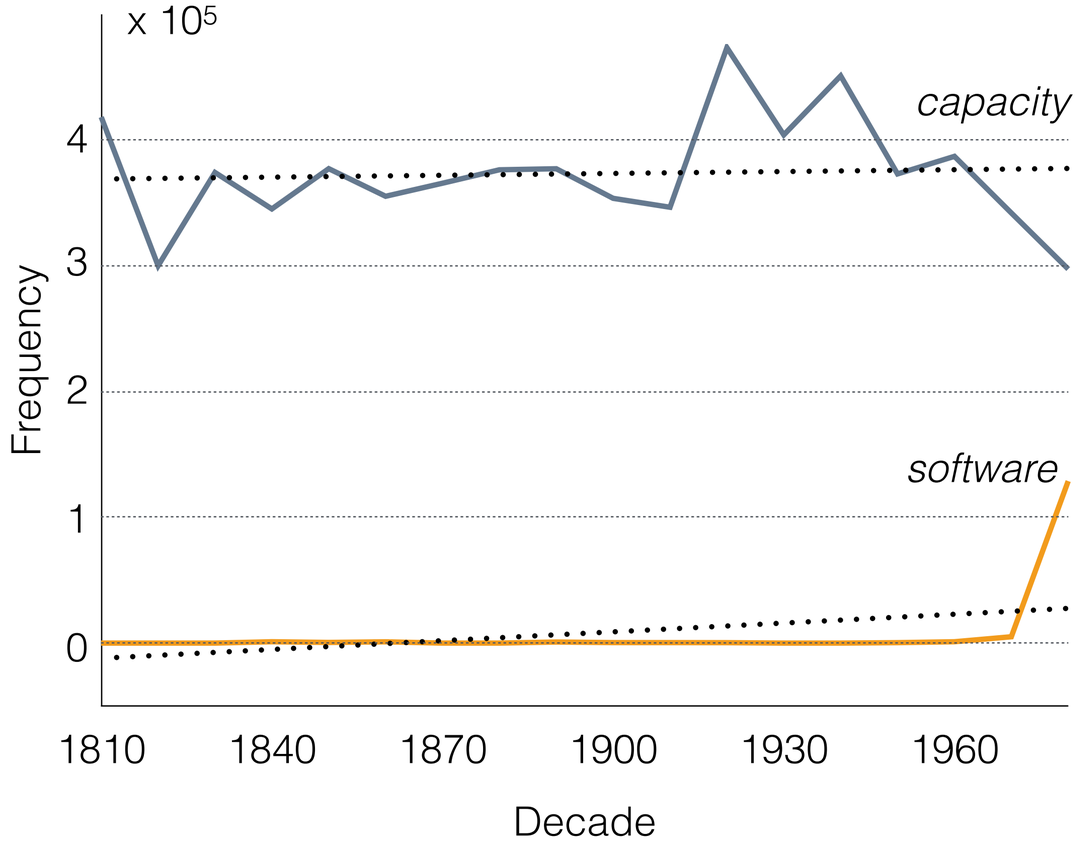
- Historical (COHA) and modern (COCA) American English corpora
  - Neologisms by frequency ratio between modern and historical data




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

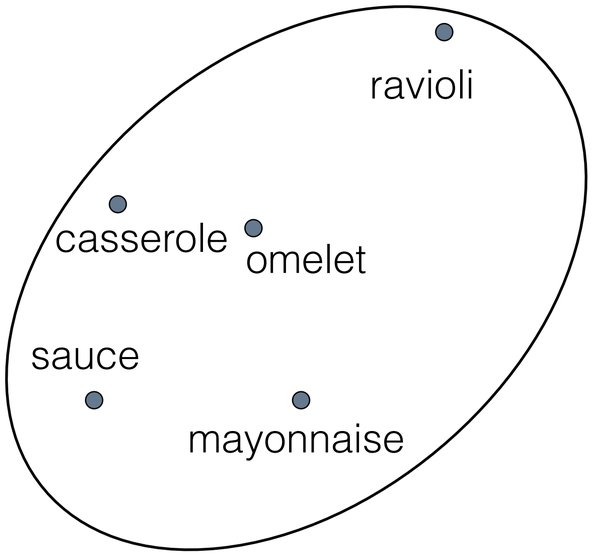
- Historical (COHA) and modern (COCA) American English corpora
  - Neologisms by frequency ratio between modern and historical data
- Each neologism paired with a non-neologism control word
  - Controlling for frequency, length, frequency stability



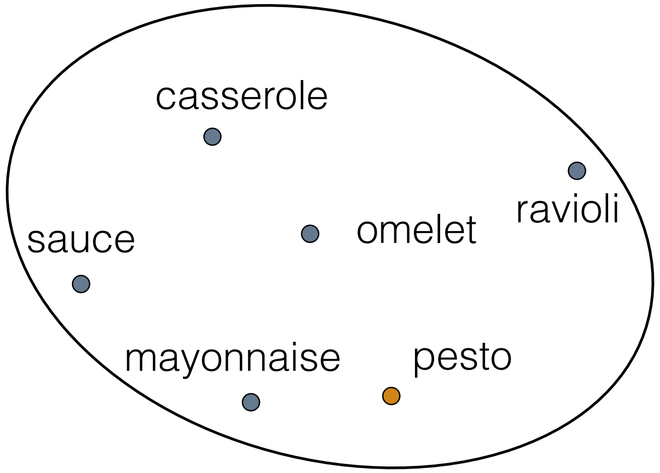

M Ryskina, E Rabinovich, T Berg-Kirkpatrick, DR Mortensen, Y Tsvetkov. Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods. SCiL 2020.

# Distributional semantics

- Embeddings learned separately from historical and modern data



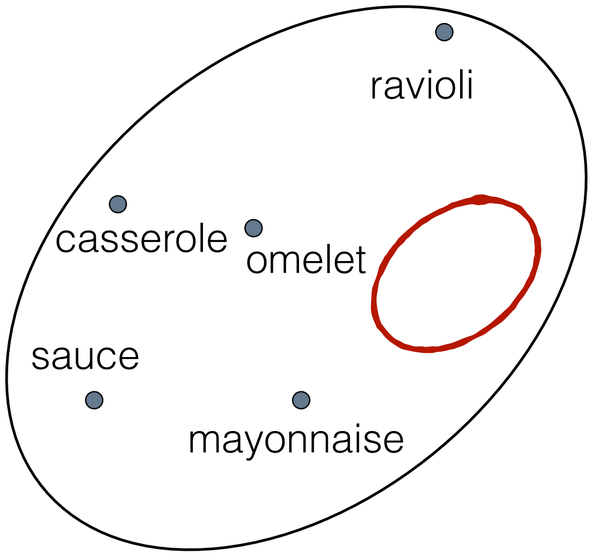

Historical



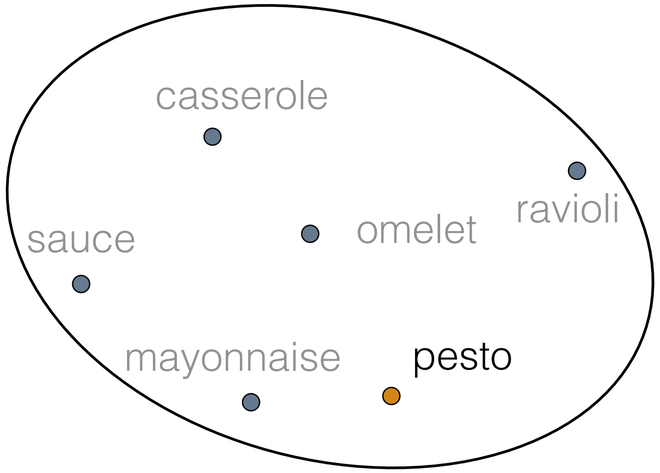

Modern

# 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



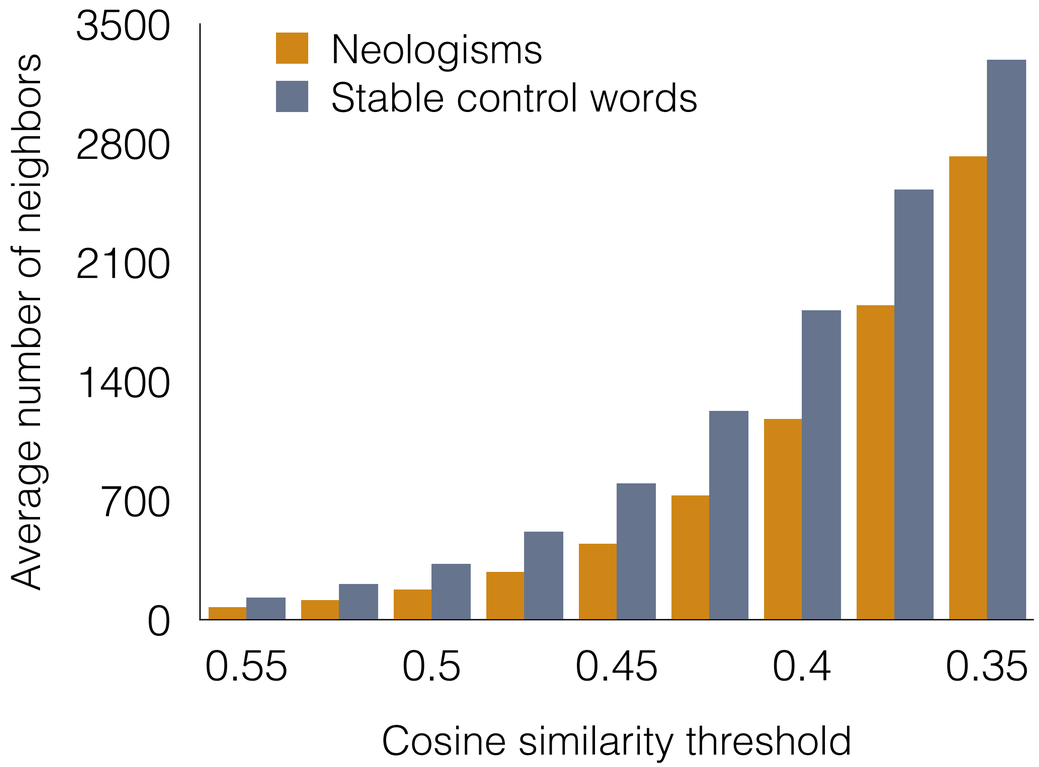

Historical

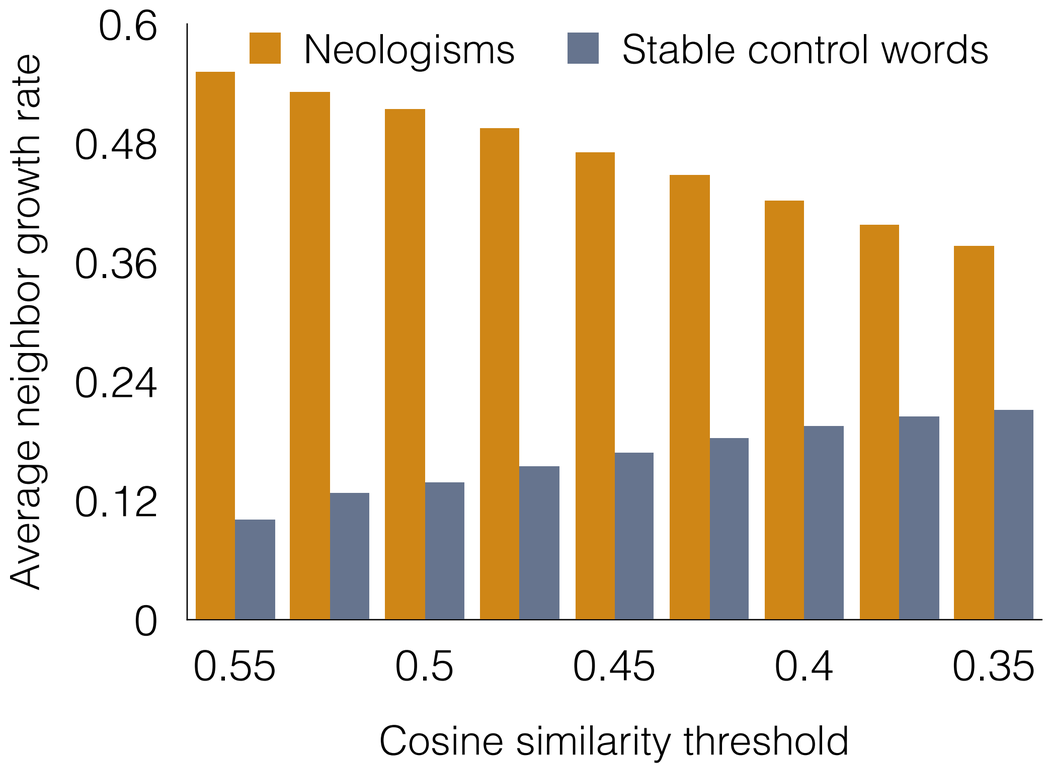



Modern

# Experimental results

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



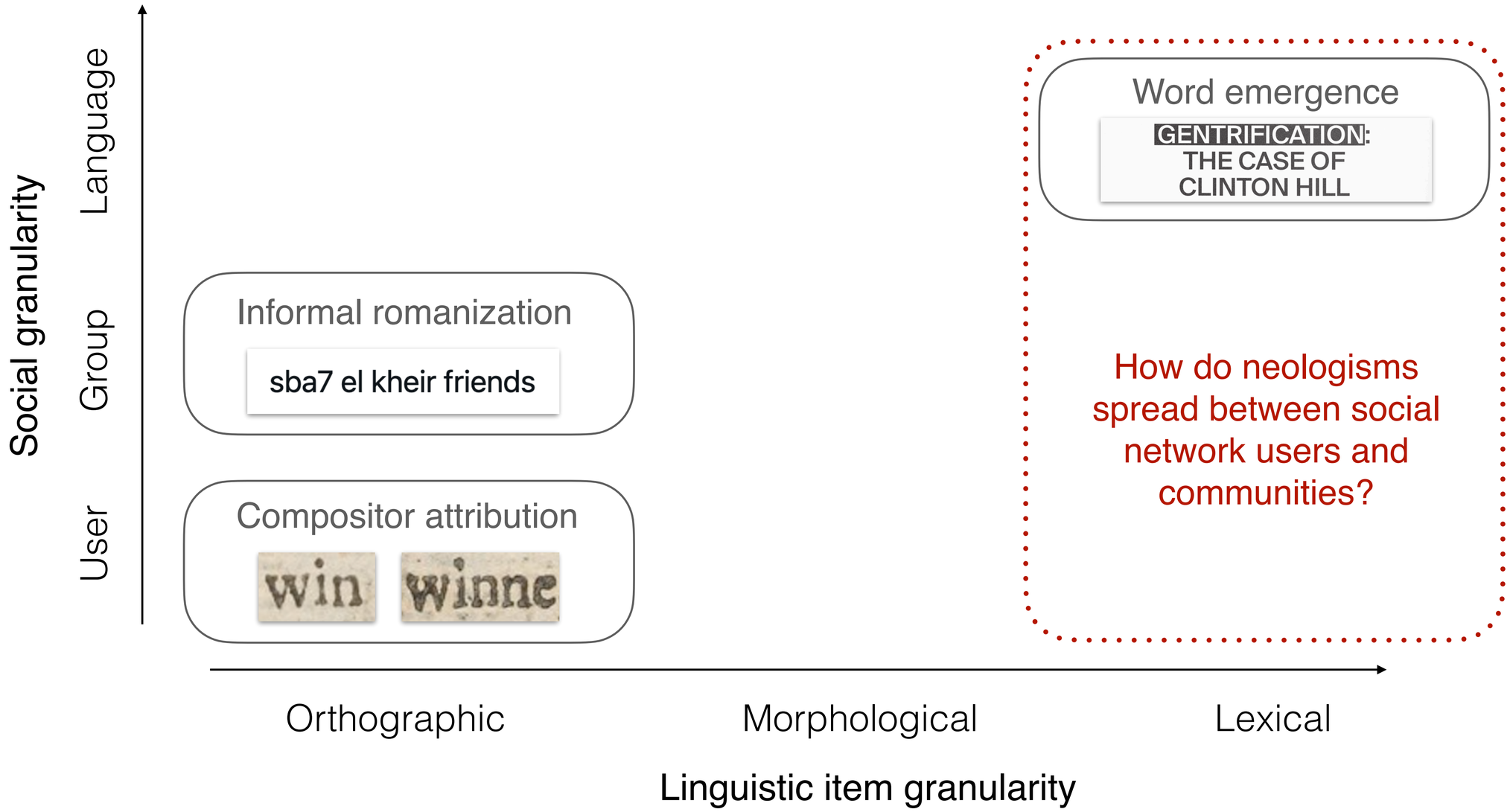



# Qualitative analysis

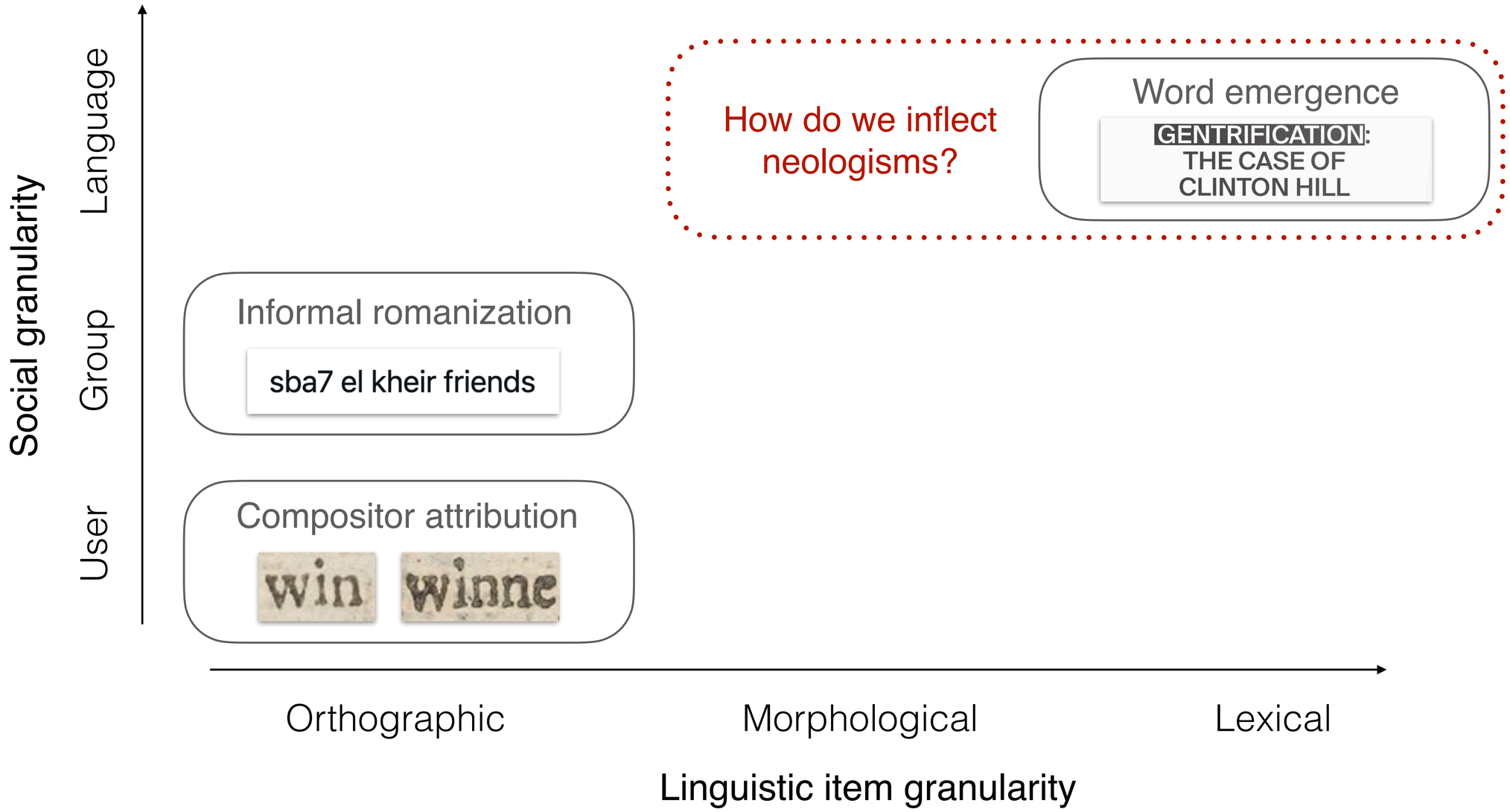
- Example of nearest historical neighbors of projected neologisms:

| 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 |

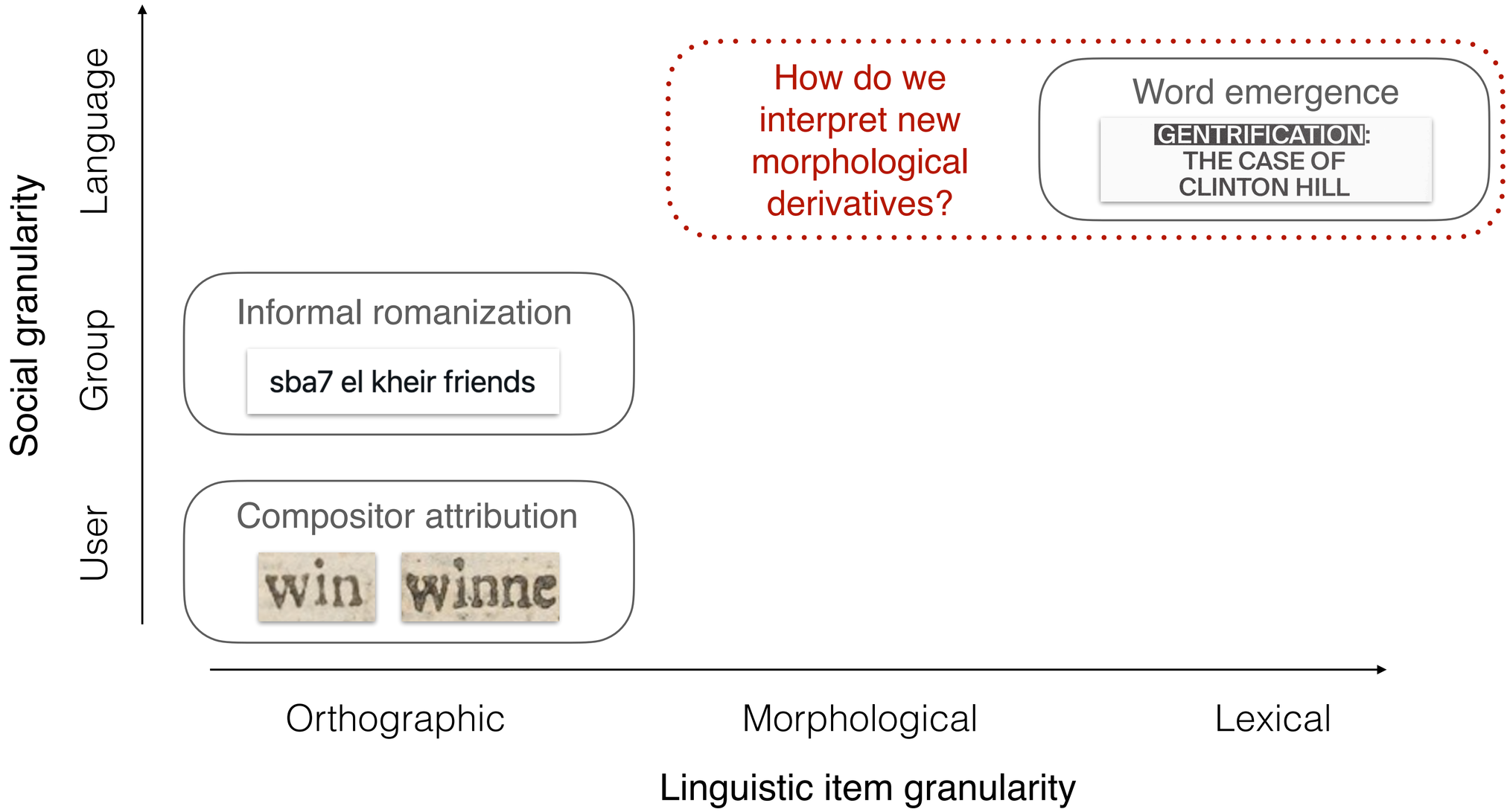
# Future work: Multiscale studies



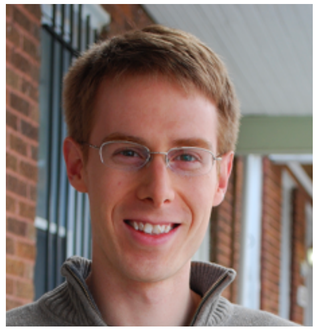
# Future work: Morphology



# Future work: Morphology



# Thank you!





Get in touch:

www.cs.cmu.edu/~mryskina/

mryskina@cs.cmu.edu

@maria_ryskina