Unsupervised Decipherment of Informal Romanization

Maria Ryskina

Carnegie Mellon University
Language Technologies Institute

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, …*

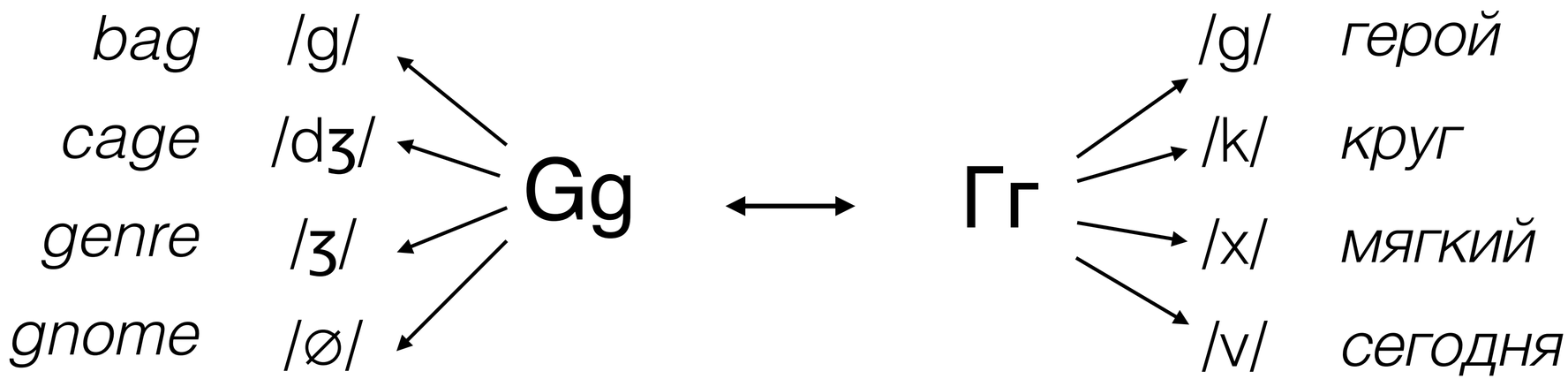
Informal romanization

- 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, …*

Phonetic romanization

- What does it mean for two characters to be phonetically similar?



- This is just in one language each!

Phonetic romanization

- What does it mean for two characters to be phonetically similar?
- Out-of-context grapheme-phoneme association: ɾ~/g/→g

Every letter makes a sound:
‘A’ says /eɪ/!*

*and /a/

https://www.youtube.com/watch?v=FFhH6576psw

Phonetic romanization

- What does it mean for two characters to be phonetically similar?
- Out-of-context grapheme-phoneme association: ɾ~/g/→g
- Phoneme produced in context: انتي /enti/→enty, صباح /sabaħ/→saba7

Visual romanization

- Broad similarity between glyph shapes
- Single characters can map to bi-/trigraphs
- Can be conditioned on a transformation
- Can be applied to a part of a glyph

а~/a/→a, г~/g/→r

ы→bl, ж→}|{

ɛ →3, л→v

ӈ →2

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

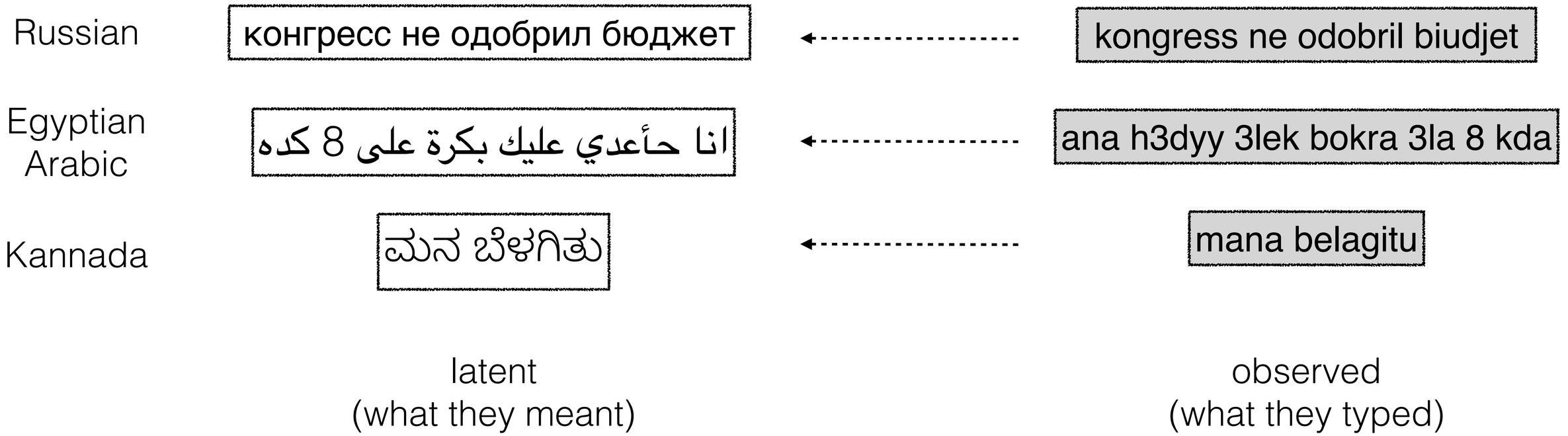
Unicode: ಬ ೆ ಳ ಗ ಿ ತ ು

belagitu

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

Task framing

- Convert romanized text to the conventional orthography of the language



Task framing

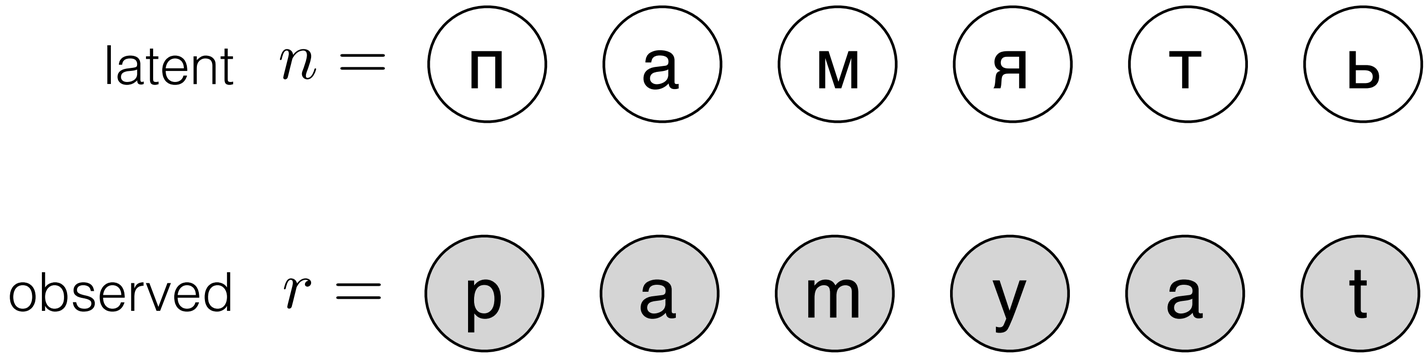
- Parallel data does not occur naturally ⇒ unsupervised learning
- Perceptions of similarity are shared across users and even languages!
- **Hypothesis**: inductive bias encoding these similarity notions provides signal that can somewhat approximate human supervision
  - We rely on manually-curated resources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



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

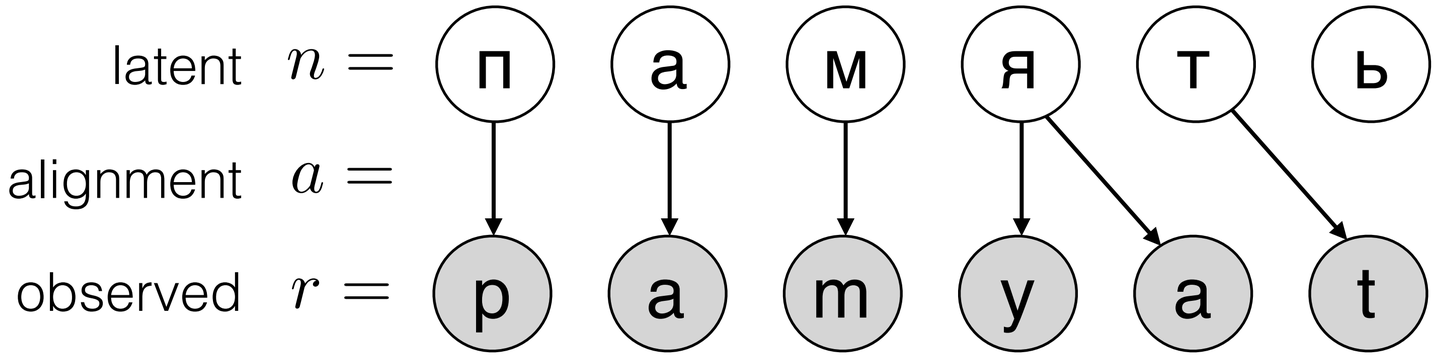
all possible native script sequences

transition probabilities

emission probabilities

prior on parameters

Noisy-channel model



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

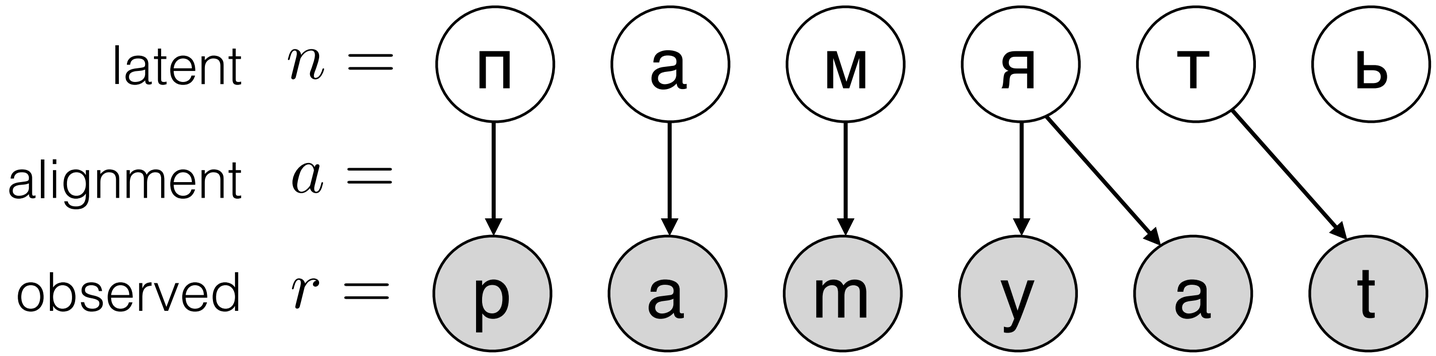
all possible native script sequences and alignments

transition probabilities

emission probabilities

prior on parameters

Noisy-channel model



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

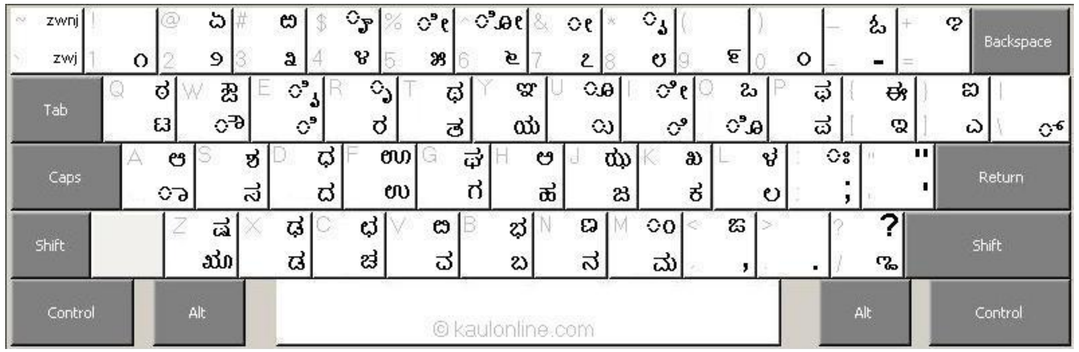
transition probabilities

emission probabilities

prior on parameters

Phonetic bias

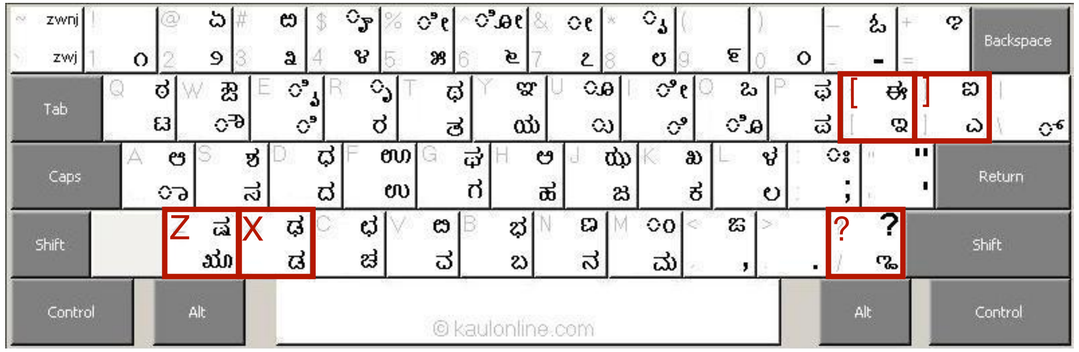
- Phonetic priors: mappings off phonetic keyboard layouts



https://en.wikipedia.org/wiki/Phonetic_keyboard_layout,
https://arabic.omaralzabir.com/, http://kaulonline.com/uninagari/kannada/

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/, http://kaulonline.com/uninagari/kannada/

Visual bias

- Visual priors: mappings off the Unicode confusables list

- Designed to combat spoofing attacks

y	ɣ	ʏ	γ	у	Ү	ყ
0079 LATIN SMALL LETTER Y	0263 LATIN SMALL LETTER GAMMA	028F LATIN LETTER SMALL CAPITAL Y	03B3 GREEK SMALL LETTER GAMMA	0443 CYRILLIC SMALL LETTER U	04AF CYRILLIC SMALL LETTER STRAIGHT U	10E7 GEORGIAN LETTER QAR
p	ρ	ϱ	р	⍴	⍴	𝐩
0070 LATIN SMALL LETTER P	03C1 GREEK SMALL LETTER RHO	03F1 GREEK RHO SYMBOL	0440 CYRILLIC SMALL LETTER ER	2374 APL FUNCTIONAL SYMBOL RHO	2CA3 COPTIC SMALL LETTER RO	1D429 MATHEMATICAL BOLD SMALL P

sigtyp.io

sigtyp.io

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

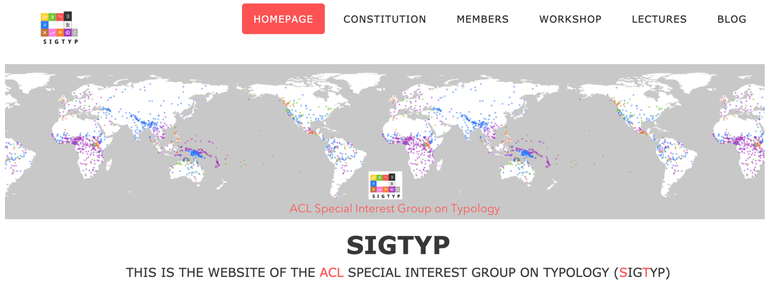
Visual bias

- Visual priors: mappings off the Unicode confusables list

- Designed to combat spoofing attacks

y	ɣ	ʏ	γ	у	Ү	ყ
0079 LATIN SMALL LETTER Y	0263 LATIN SMALL LETTER GAMMA	028F LATIN LETTER SMALL CAPITAL Y	03B3 GREEK SMALL LETTER GAMMA	0443 CYRILLIC SMALL LETTER U	04AF CYRILLIC SMALL LETTER STRAIGHT U	10E7 GEORGIAN LETTER QAR
p	ρ	ϱ	р	⍴	ⲣ	𝗽
0070 LATIN SMALL LETTER P	03C1 GREEK SMALL LETTER RHO	03F1 GREEK RHO SYMBOL	0440 CYRILLIC SMALL LETTER ER	2374 APL FUNCTIONAL SYMBOL RHO	2CA3 COPTIC SMALL LETTER RO	1D429 MATHEMATICAL BOLD SMALL P

sigtyp.io



sigtyp.io

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 and Kannada!

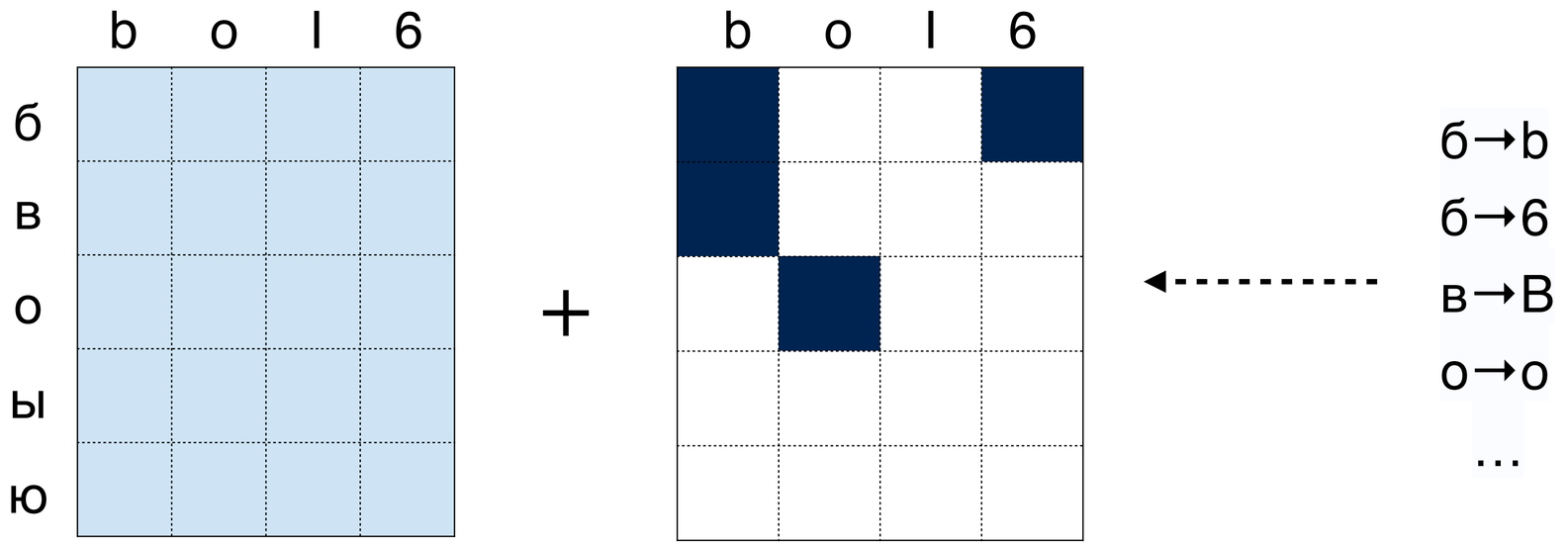
ʏ	ɣ	ʏ	Ɣ	у	Ү	ყ
0079 LATIN SMALL LETTER Y	0263 LATIN SMALL LETTER GAMMA	028F LATIN LETTER SMALL CAPITAL Y	03B3 GREEK SMALL LETTER GAMMA	0443 CYRILLIC SMALL LETTER U	04AF CYRILLIC SMALL LETTER STRAIGHT U	10E7 GEORGIAN LETTER QAR
ᴘ	ρ	ϱ	р	⍴	⍴	𝐩
0070 LATIN SMALL LETTER P	03C1 GREEK SMALL LETTER RHO	03F1 GREEK RHO SYMBOL	0440 CYRILLIC SMALL LETTER ER	2374 APL FUNCTIONAL SYMBOL RHO	2CA3 COPTIC SMALL LETTER RO	1D429 MATHEMATICAL BOLD SMALL P

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

Informative priors

- Use mappings of similar characters as priors on emission parameters

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

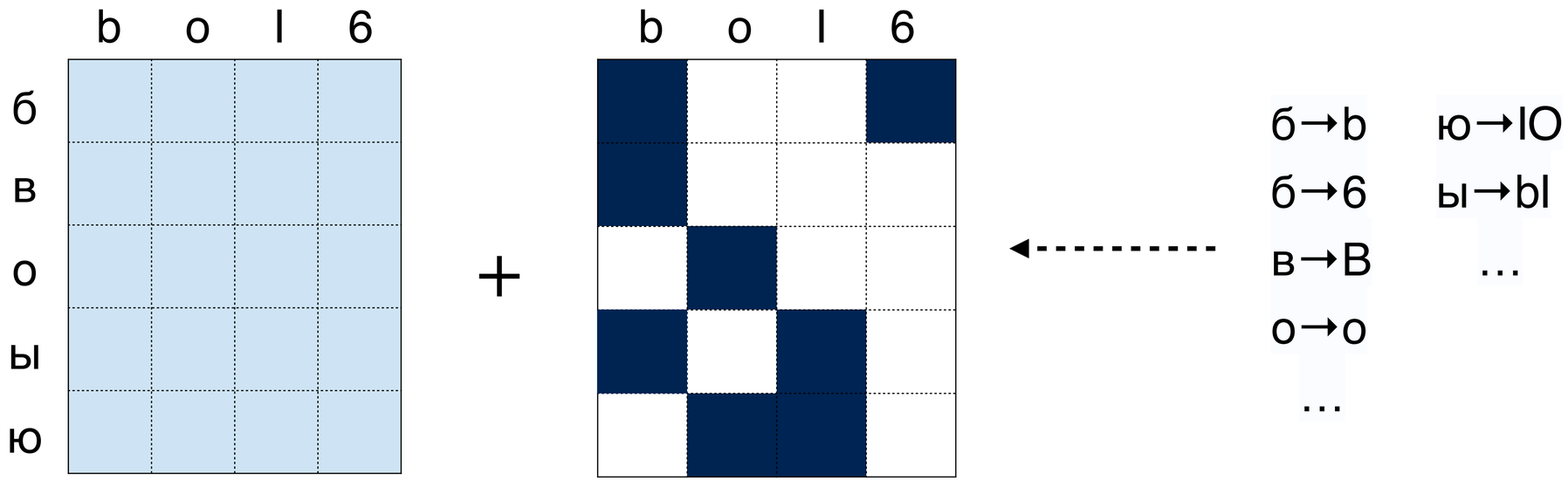


Informative priors

- Use mappings of similar characters as priors on emission parameters

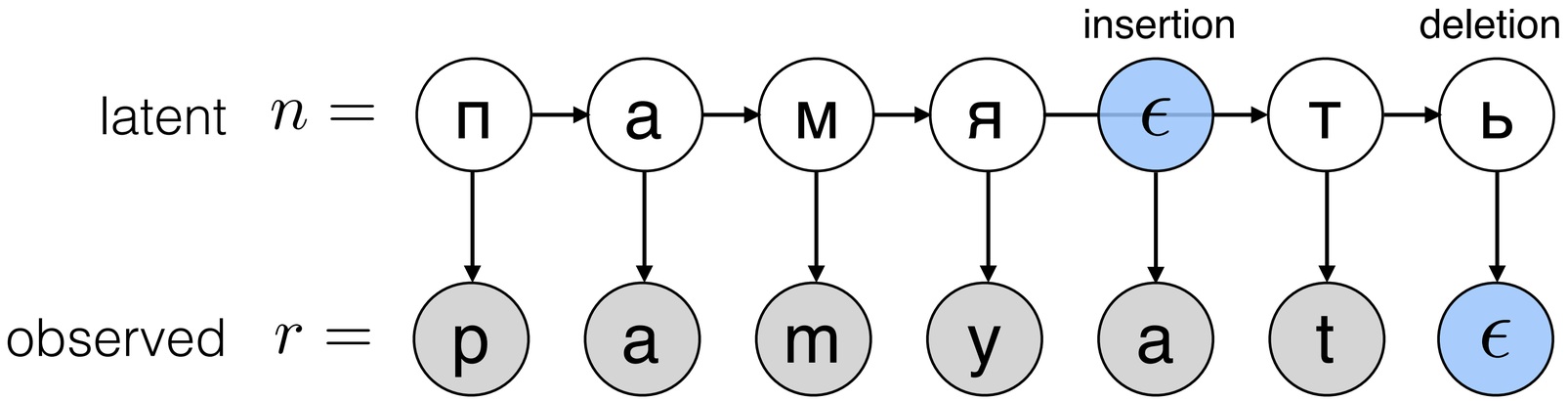
$$c_r | c_n \sim \mathrm{Mult}(\theta_{c_n})$$

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



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

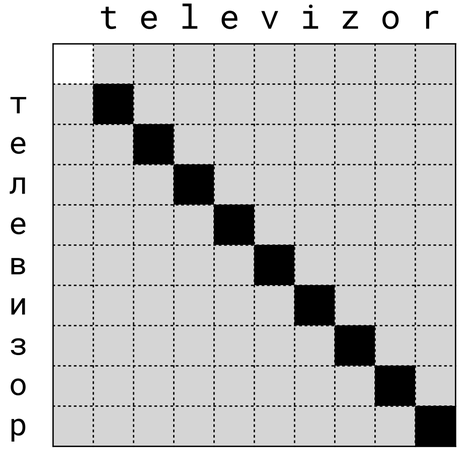
prior on parameters

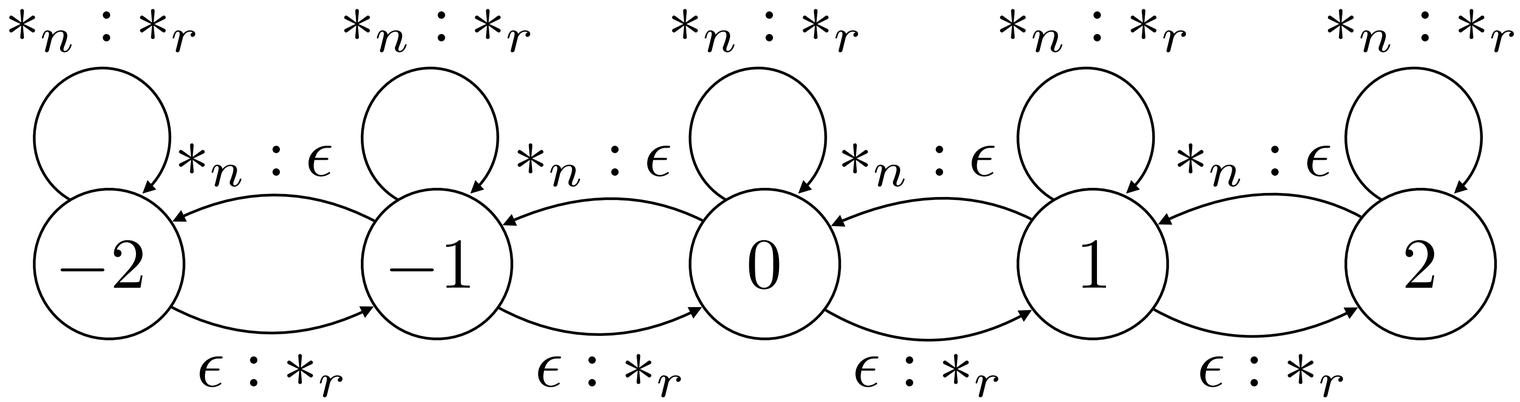
WFST cascade

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

Emission model

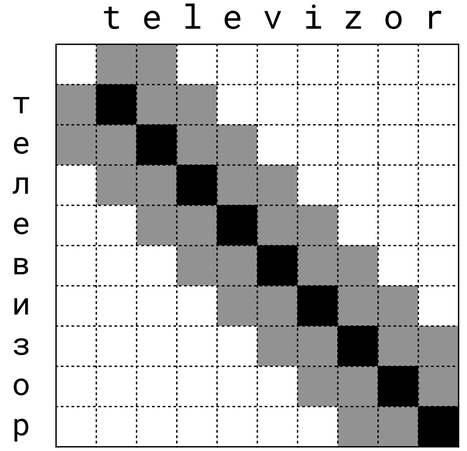
- Needs to support substitutions, insertions and deletions
- Fixed limit on delay: |# of insertions - # of deletions|

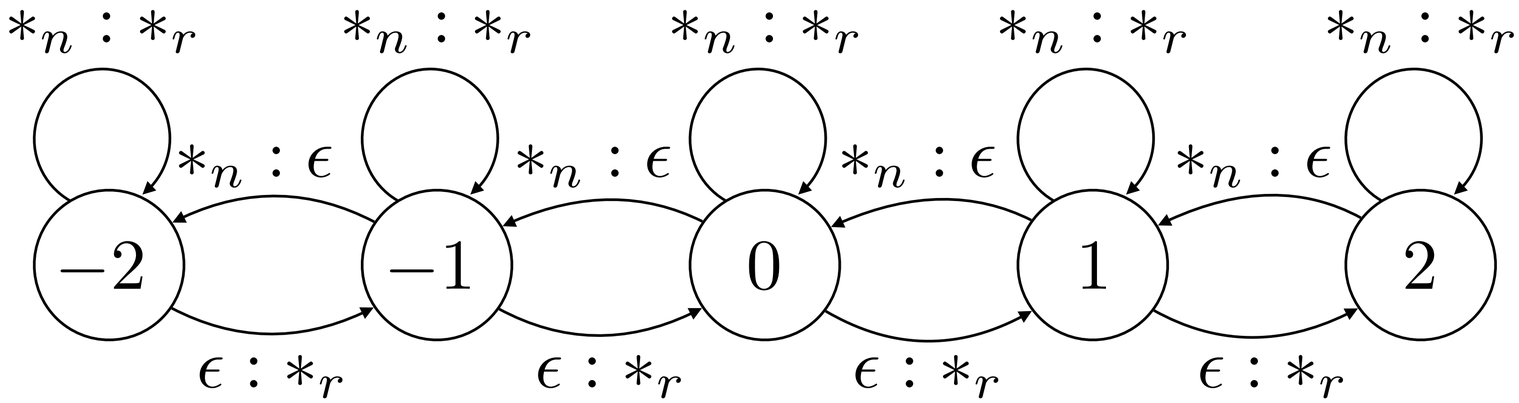




Emission model

- 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)

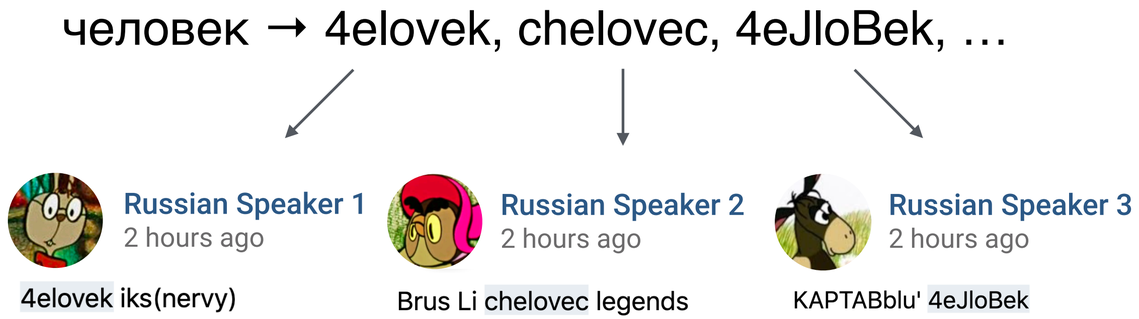
Saba7 el 5eir!
Ezayeeky?

- Kannada: Dakshina dataset (Roark et al., 2020)
  - Kannada Wikipedia, romanizations elicited from native speakers

- Russian:
  - Romanized: collected and partly annotated data from social media
  - Native: Taiga corpus (Shavrina & Shapovalova, 2017), comments in political forums

Russian data

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



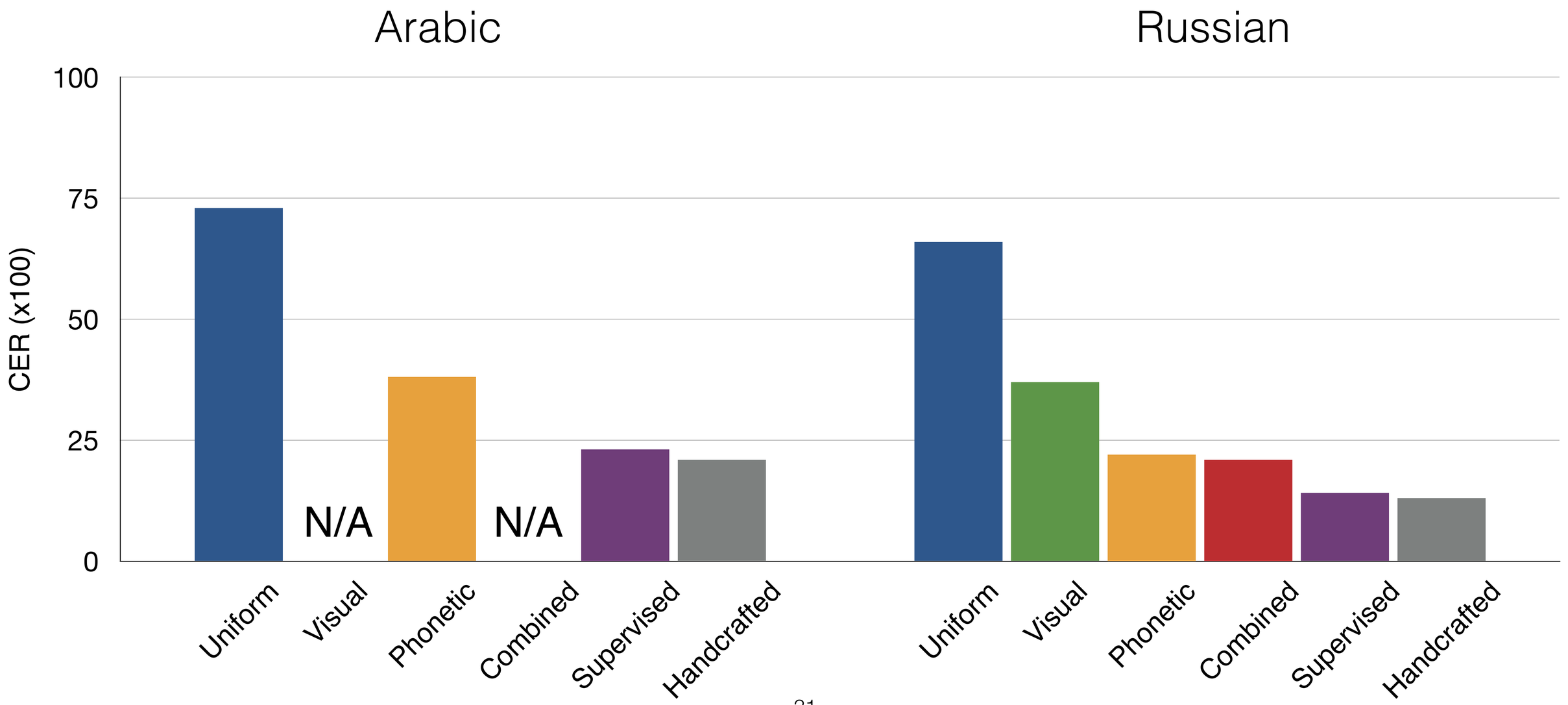
- Annotated validation and test with minor error correction

Source: proishodit s prirodoy 4to to very very bad

Filtered: proishodit s prirodoy 4to to <…>

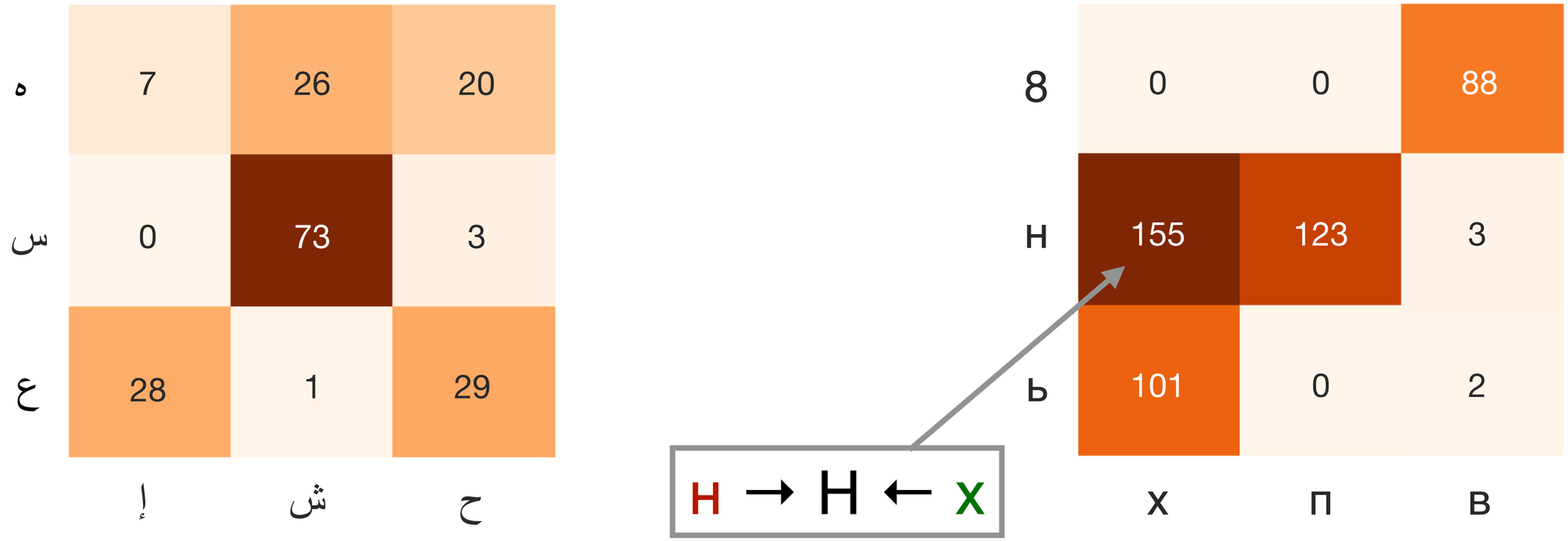
Target: происходит с природой что-то <…>

WFST results



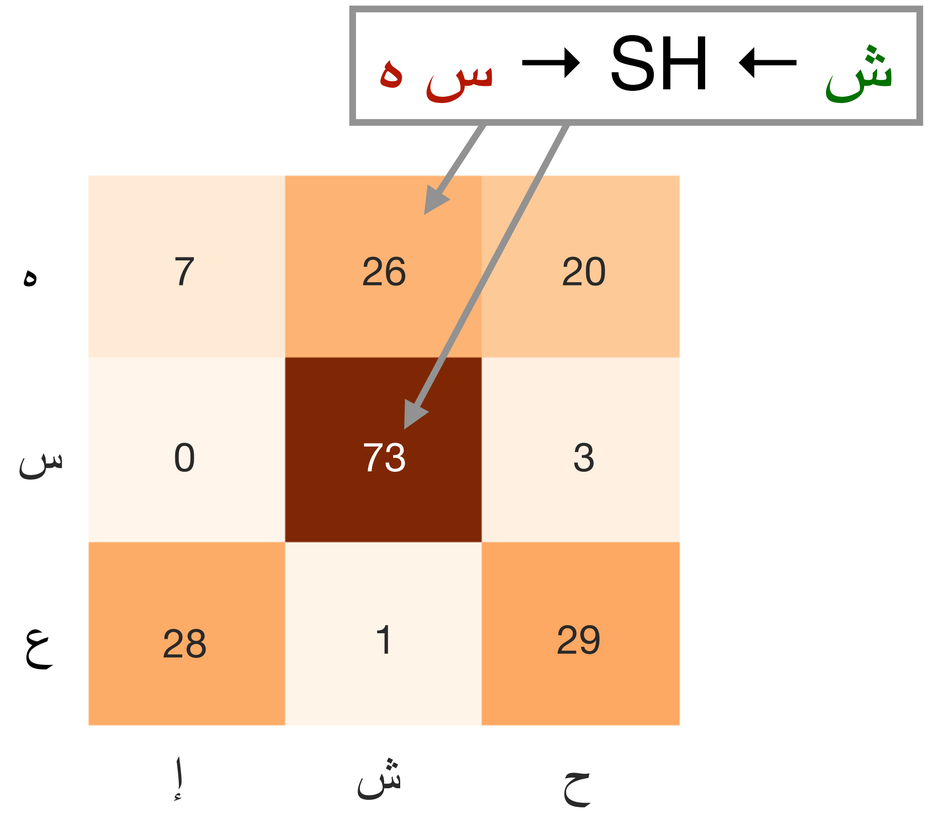
WFST error analysis

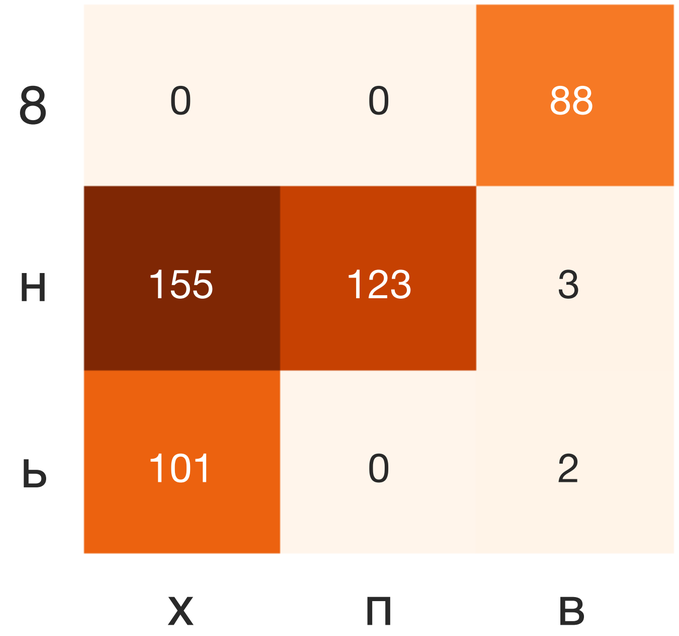
- Incorrect choice of plausible de-romanization (e.g. visual instead of phonetic)



WFST error analysis

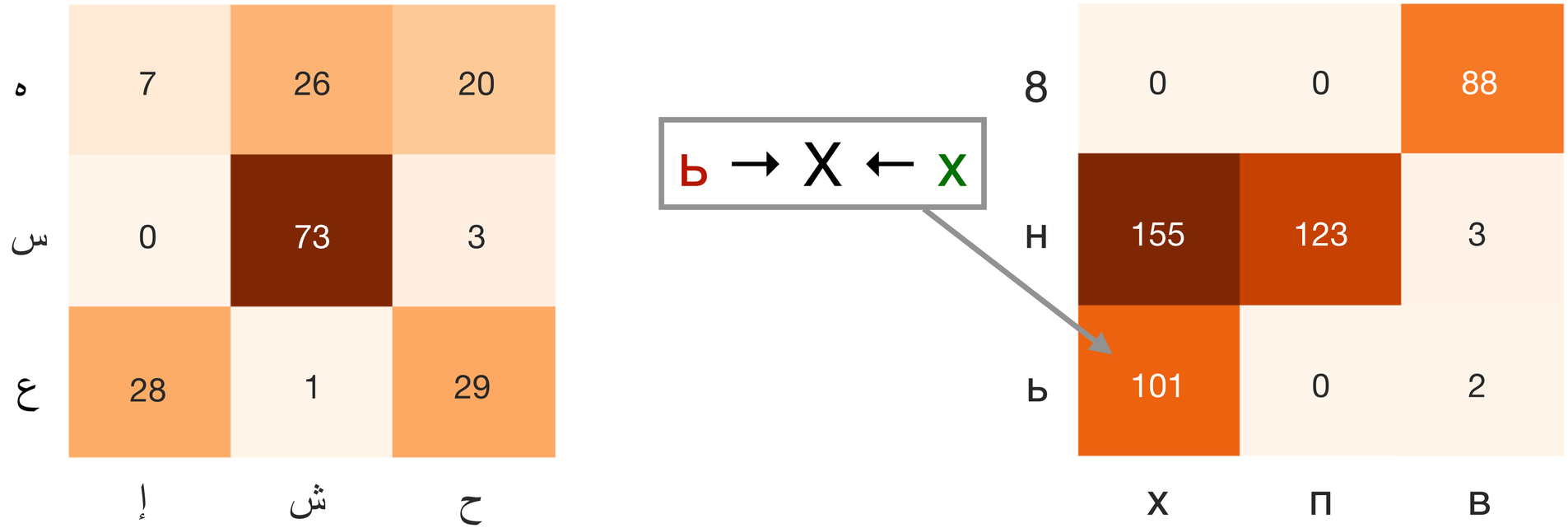
- Inability to handle digraphs like SH





WFST error analysis

- Distracted by spurious mappings in priors



Model classes

WFSTs are structured

- ✓ Easy to encode constraints
- ✓ Can learn from small data
- ✗ Slow exact maximization
- ✗ Weak n-gram language model

Seq2seqs are powerful

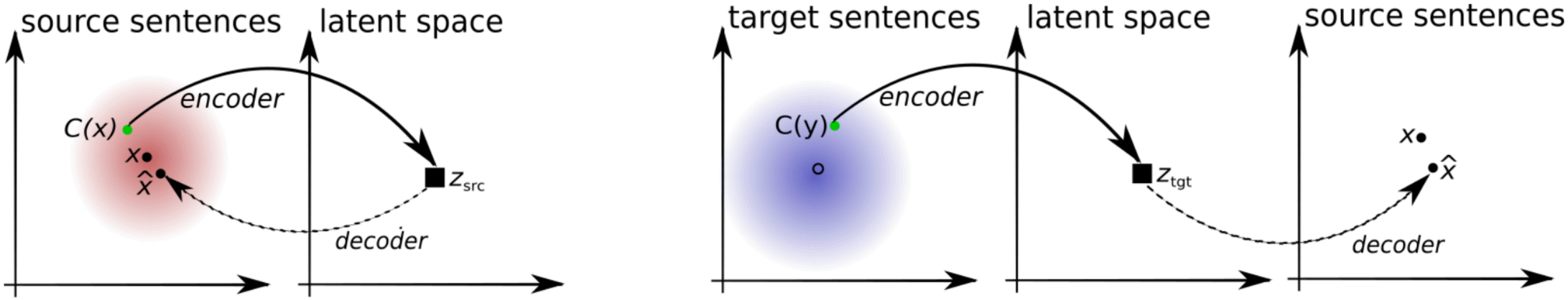
- ✓ Strong language model
- ✓ Faster batch processing
- ✗ Need large training data
- ✗ Hallucinations and search errors

In our case, both are trained unsupervised!

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.

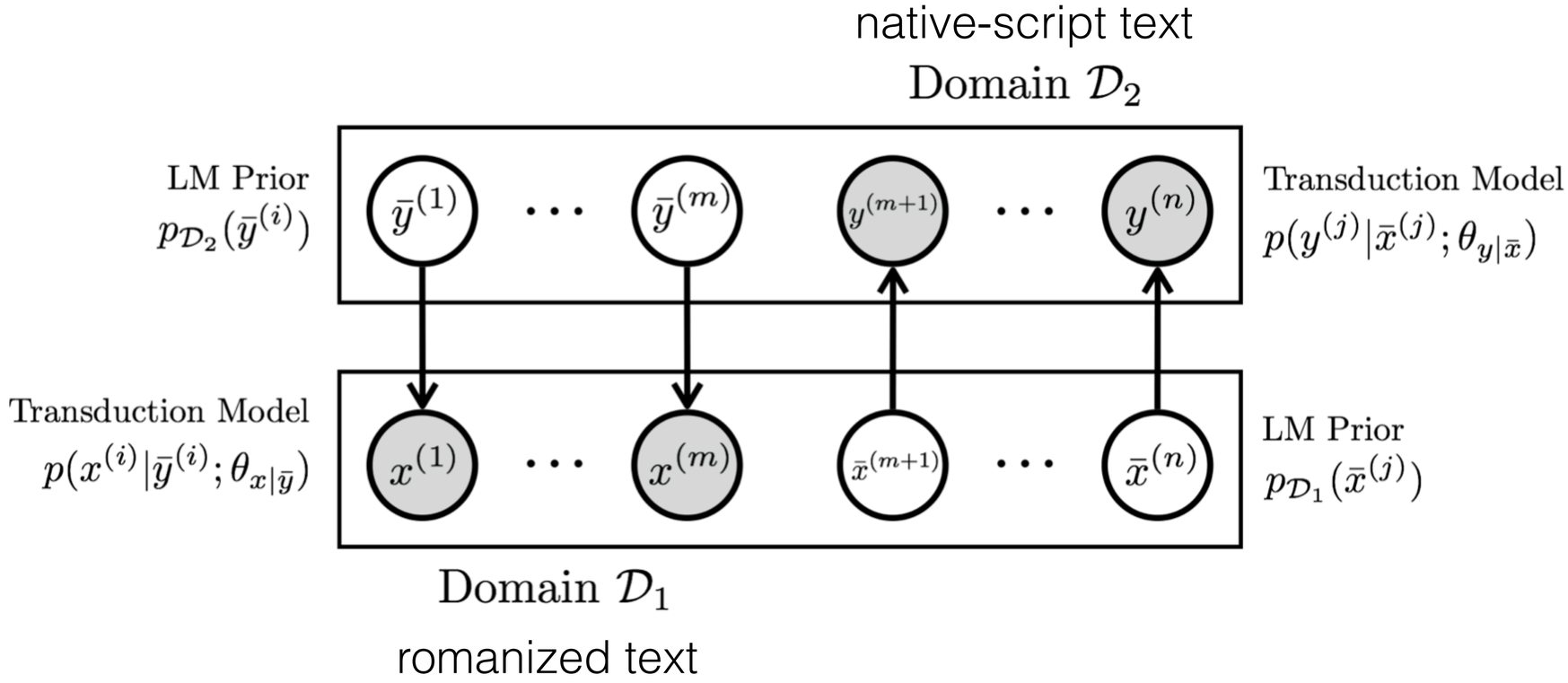
Unsupervised seq2seq

- Unsupervised neural machine translation (UNMT; Lample et al., 2018)
  - Auto-encoding: reconstructing a sentence from its noisy version
  - Back-translation: round trip through the latent space
  - Adversarial: discriminating between sentences in two domains



Unsupervised seq2seq

- Probabilistic formulation of UNMT: deep latent sequence model (He et al., 2020)

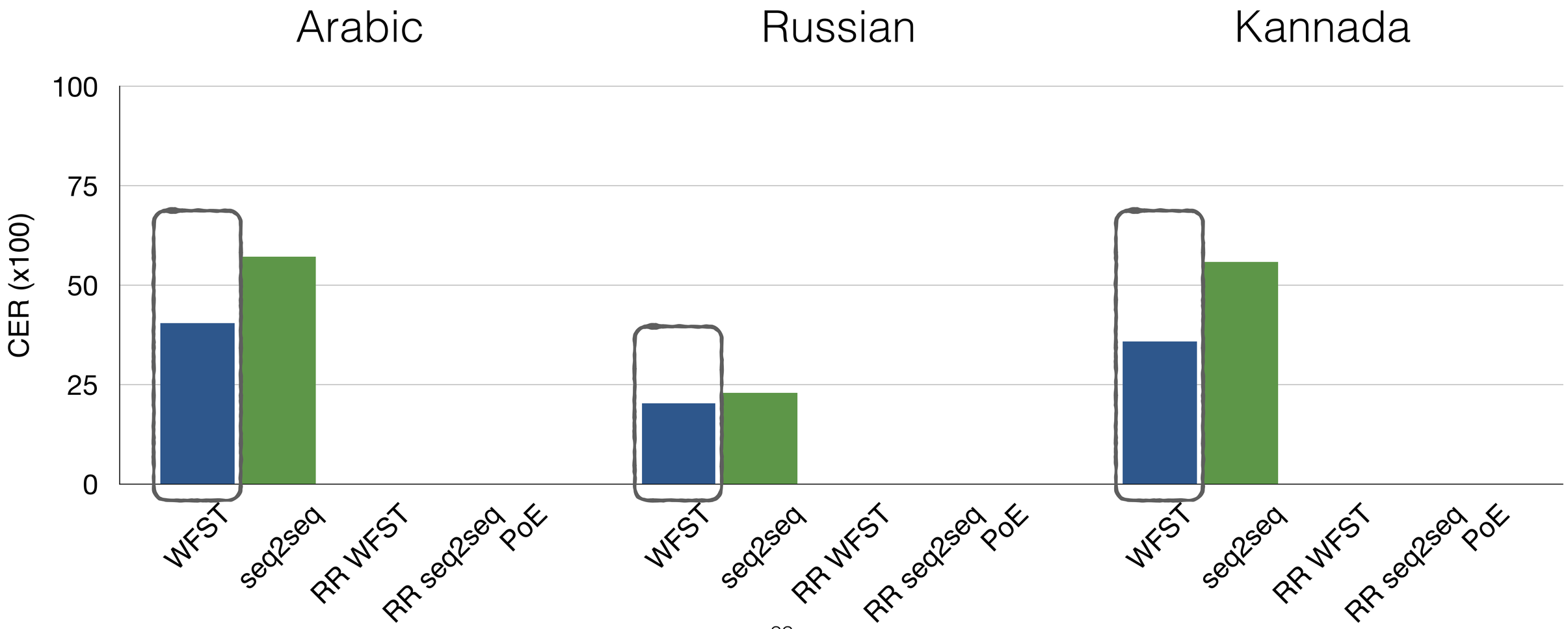


Model combinations

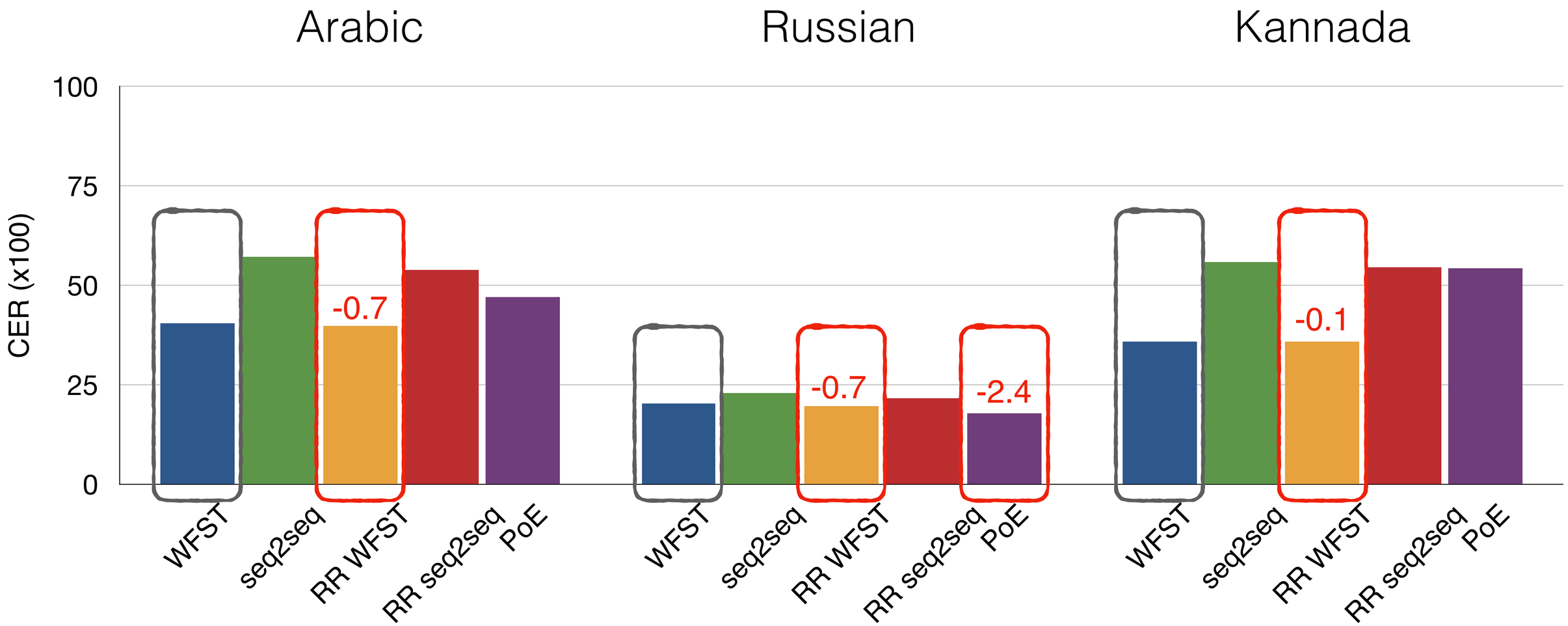
- Reranking
  - M1 generates top k candidate outputs
  - M2 selects the highest-scoring candidate
- Product of experts
  - Beam search on the WFST lattice
  - WFST arcs reweighted with Seq2seq softmax at the corresponding timstep
  - Deletions of input characters are not reweighted
  - Candidates are grouped by consumed input length
- We train the models separately and combine at test time

Results

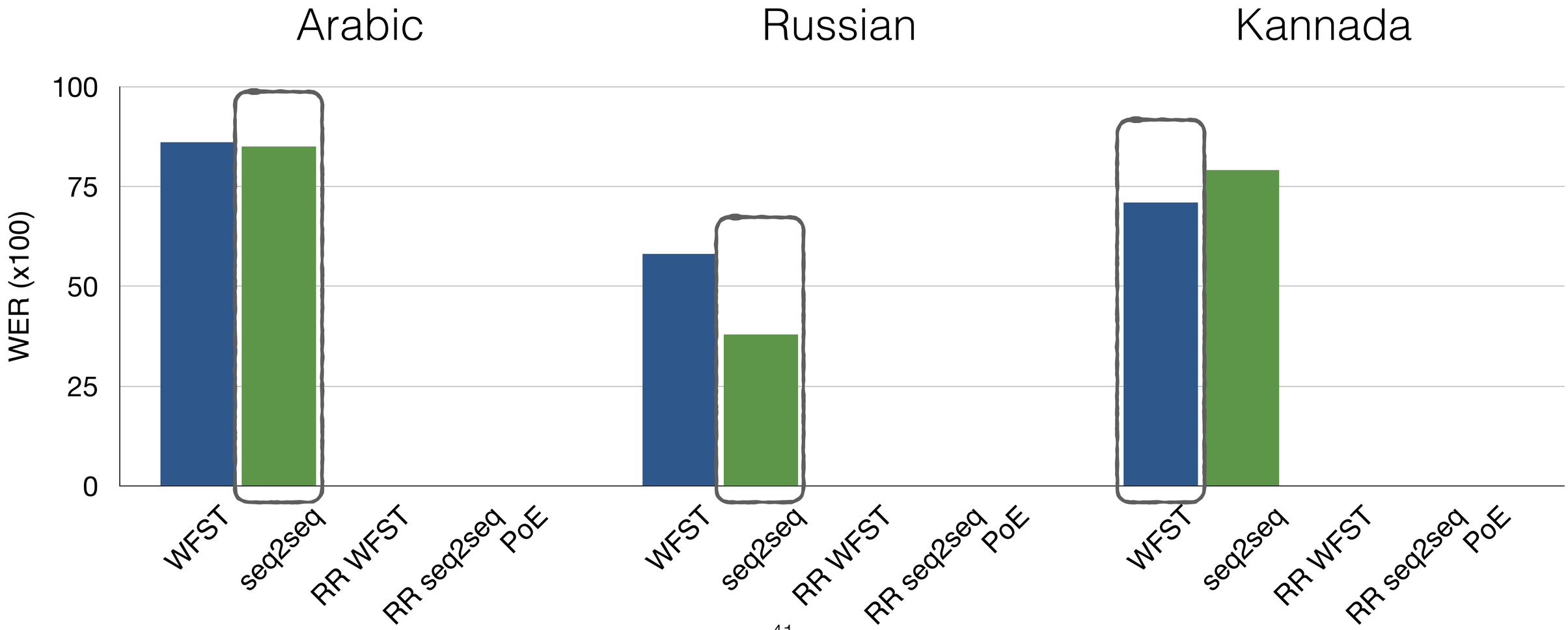
Base models are trained on different amounts of data!



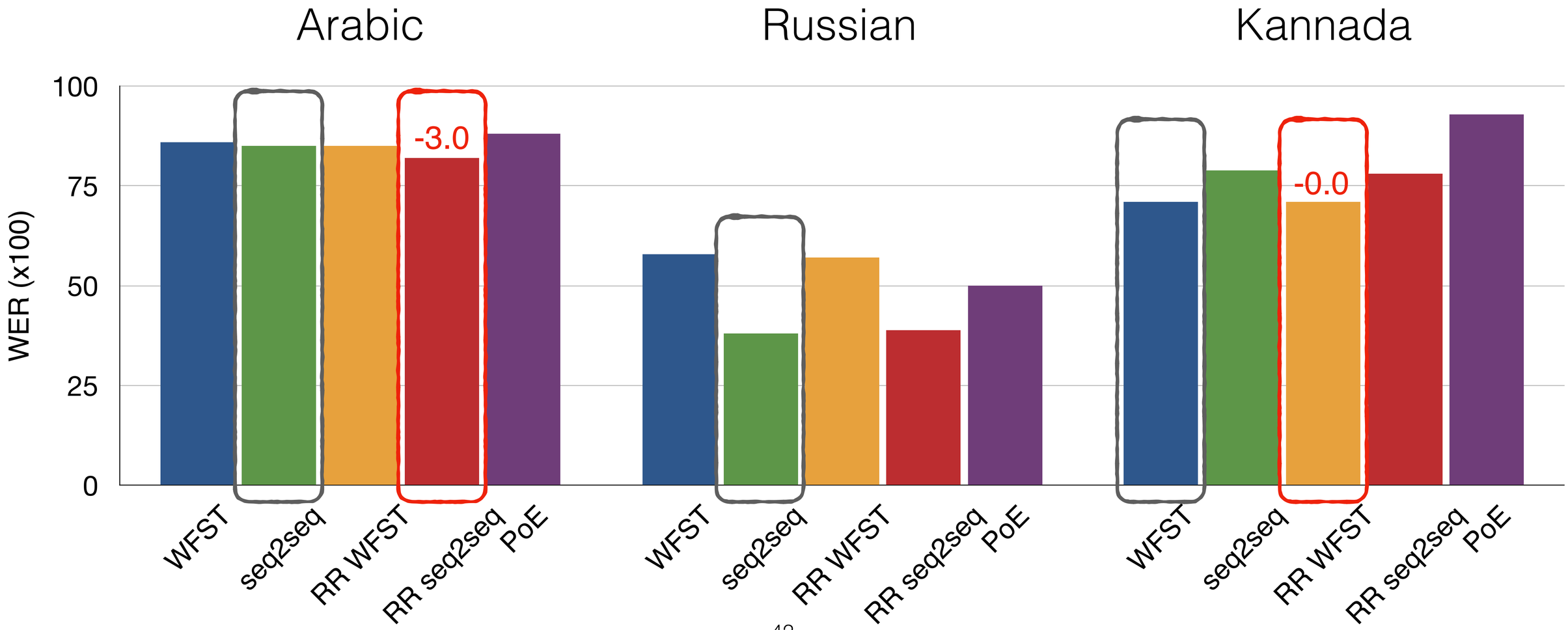
Results


Arabic
Russian
Kannada
CER (x100)
100
75
50
25
0
-0.7
-0.7
-2.4
-0.1
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE

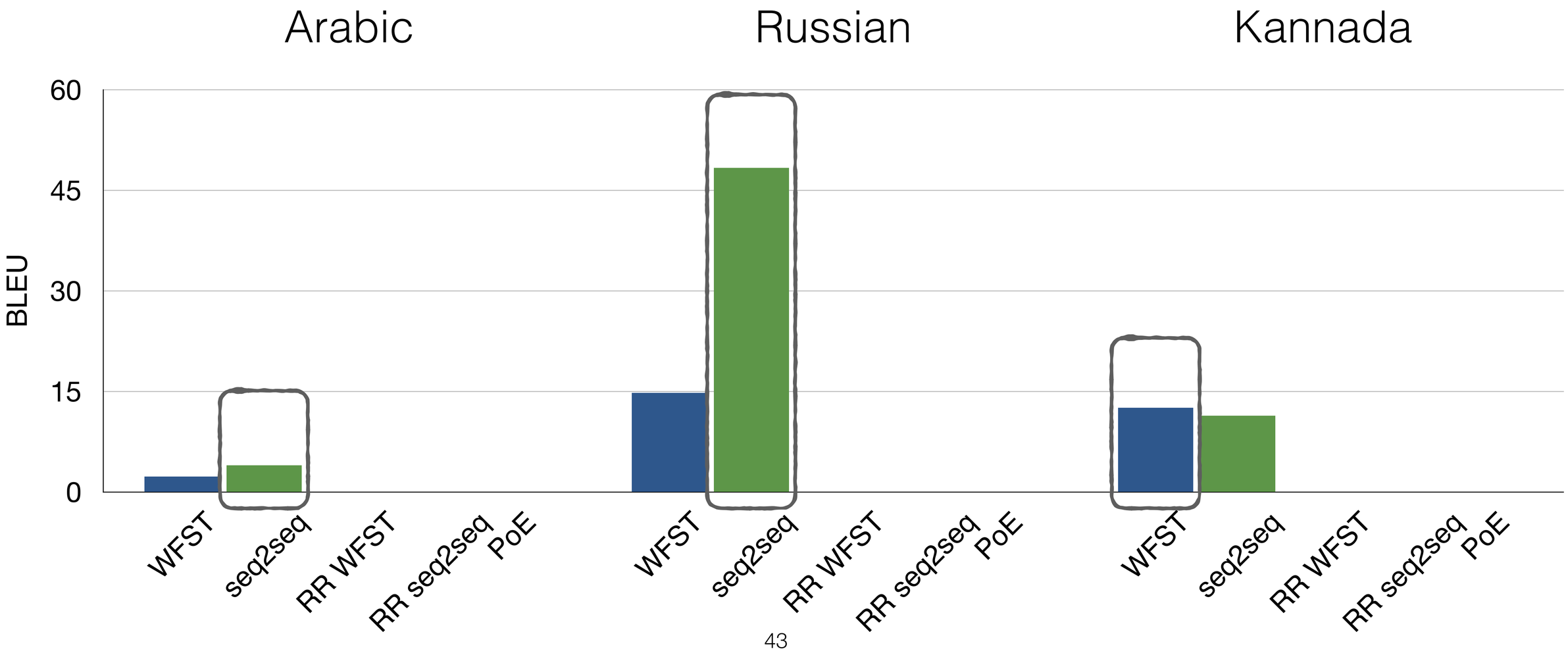
Results


Arabic
Russian
Kannada
WER (x100)
100
75
50
25
0
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE

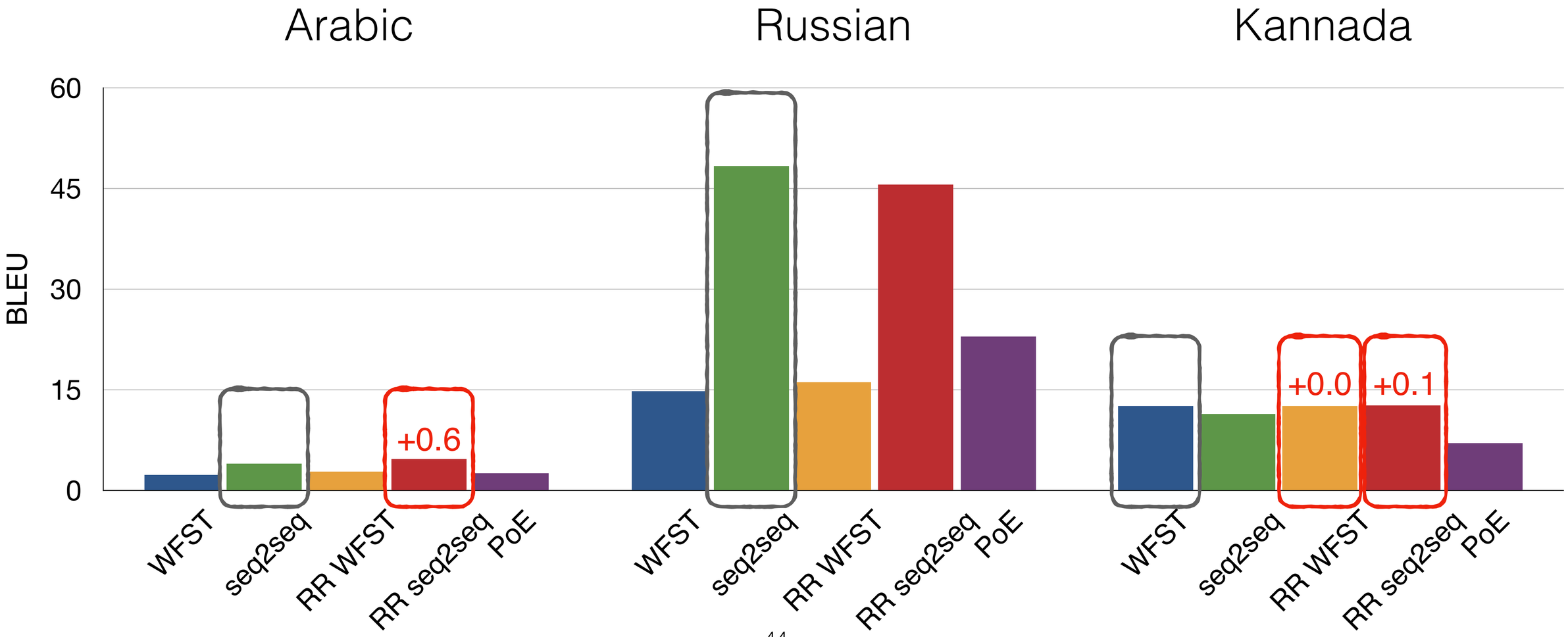
Results


Arabic
Russian
Kannada
WER (x100)
100
75
50
25
0
-3.0
-0.0
WFST
seq2seq
RR WFST
RR seq2seq
PoE

Results


Arabic
Russian
Kannada
BLEU
60
45
30
15
0
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE

Results


Arabic
Russian
Kannada
BLEU
60
45
30
15
0
+0.6
+0.0
+0.1
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE
WFST
seq2seq
RR WFST
RR seq2seq
PoE

Error analysis

Input	kongress ne odobril biudjet dlya osuchestvleniye "bor'bi s kommunizmom" v yuzhniy amerike.	
Ground truth	конгресс не одобрил бюджет для осуществления "борьбы с коммунизмом" в южной америке.	kongress ne odobril bjudžet dlja osuščestvlenija "bor'by s kommunizmom" v južnoj amerike.
WFST	конгресс не одобрил виудет дла осусчествлениые "бор#би с коммунизмом" в уузнани америке.	kongress ne odobril viudet dla osusčestvleniye "bor#bi s kommunizmom" v uuznani amerike.
Reranked WFST	конгресс не одобрил видет дела осусчествлениые "бор#би с коммунизмом" в уузнани америке.	kongress ne odobril videt dela osusčestvleniye "bor#bi s kommunizmom" v uuznani amerike.
Seq2Seq	конгресс не одобрил бы удивительно с коммунизмом" в южный америке.	kongress ne odobril by udivitel'no s kommunizmom" v južnyj amerike.
Reranked Seq2Seq	конгресс не одобрил бюджет для осуществление "борьбы с коммунизмом" в южный америке.	kongress ne odobril bjudžet dlja osuščestvlenie "bor'by s kommunizmom" v južnyj amerike.
Product of experts	конгресс не одобрил бидет для а осуществлениые "борьбы с коммунизмом" в уузник амери	kongress ne odobril bidet dlja a osuščestvleniye "bor'by s kommunizmom" v uuznik ameri

#=UNK

Error analysis

Input	kongress ne odobril biudjet dlya osuchestvleniye "bor'bi s kommunizmom" v yuzhniy amerike.	
Ground truth	конгресс не одобрил бюджет для осуществления "борьбы с коммунизмом" в южной америке.	kongress ne odobril bjudžet dlja osuščestvlenija "bor'by s kommunizmom" v južnoj amerike.
WFST	конгресс не одобрил виудет дла осусчествлениые "бор#би с коммунизмом" в уузнани америке.	kongress ne odobril viudet dla osusčestvleniye "bor#bi s kommunizmom" v uuznani amerike.
Reranked WFST	конгресс не одобрил видет дела осусчествлениые "бор#би с коммунизмом" в уузнани америке.	kongress ne odobril videt dela osusčestvleniye "bor#bi s kommunizmom" v uuznani amerike.
Seq2Seq	конгресс не одобрил бы удивительно с коммунизмом" в южный америке.	kongress ne odobril by udivitel'no s kommunizmom" v južnyj amerike.
Reranked Seq2Seq	конгресс не одобрил бюджет для осуществление "борьбы с коммунизмом" в южный америке.	kongress ne odobril bjudžet dlja osuščestvlenie "bor'by s kommunizmom" v južnyj amerike.
Product of experts	конгресс не одобрил бидет для а осуществлениые "борьбы с коммунизмом" в уузник амери	kongress ne odobril bidet dlja a osuščestvleniye "bor'by s kommunizmom" v uuznik ameri

Hallucination

Incorrect but faithful

High-level takeaways

- Model combinations still suffer from search issues

Source: eto uzhe (strashno skazat') stariy rolik.

Target: это уже (страшно сказать) старый ролик

Gloss: ‘By now this is (I’m almost afraid to say it) an old video’

Final beam hypotheses and reranker scores:

456.7, единая россия уже #страшно сказать) старый
502.0, единоросы уже #страшно сказать) старый рол
482.0, единороссы уже #страшно сказать) старый ро
456.8, единую россию уже #страшно сказать) старый
449.8, единой россии уже #страшно сказать) старый

High-level takeaways

- Model combinations still suffer from search issues

Source: eto uzhe (strashno skazat') stariy rolik.

Target: это уже (страшно сказать) старый ролик

Gloss: ‘This’ w this is (I’m almost afraid to say it) an old video’

Final beam hypotheses and reranker scores:

456.7, единая россия уже #страшно сказать) старый
502.0, единоросы уже #страшно сказать) старый рол
482.0, единороссы уже #страшно сказать) старый ро
456.8, единую россию уже #страшно сказать) старый
449.8, единой россии уже #страшно сказать) старый

‘United Russia’

High-level takeaways

- Model combinations still suffer from search issues

Source: eto uzhe (strashno skazat') stariy rolik.

Target: это уже (страшно сказать) старый ролик

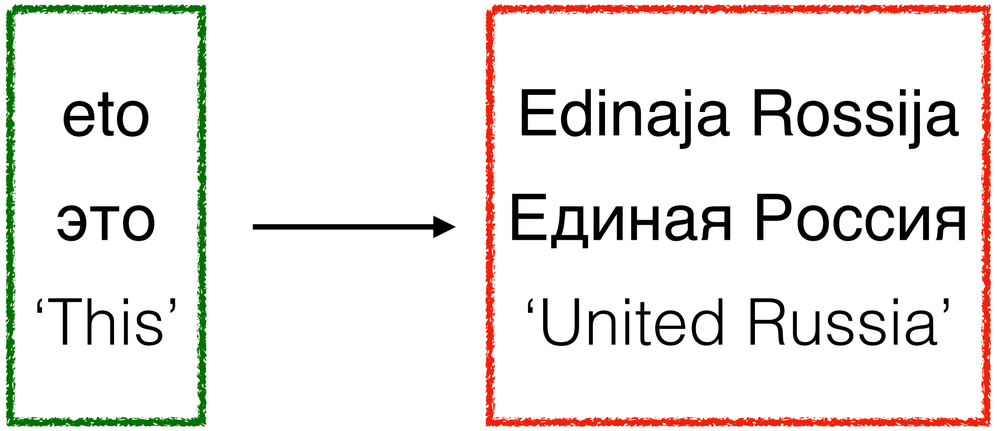
Gloss: ‘By now this is (I’m almost afraid to say it) an old video’

Final beam hypotheses and reranker scores:

456.7, единая россия уже #страшно сказать) старый
502.0, единоросы уже #страшно сказать) старый рол
482.0, единороссы уже #страшно сказать) старый ро
456.8, единую россию уже #страшно сказать) старый
449.8, единой россии уже #страшно сказать) старый

High-level takeaways

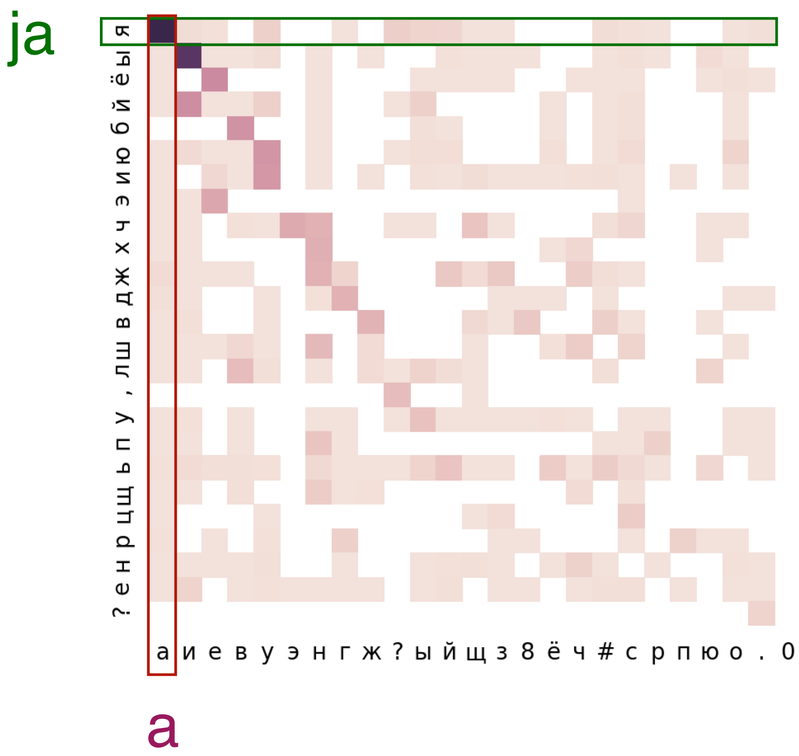
- Seq2seq is more sensitive to distributional shifts
  - Remember that our Cyrillic data comes from political discussion groups
  - 25% of common word-level errors in seq2seq are of this type!



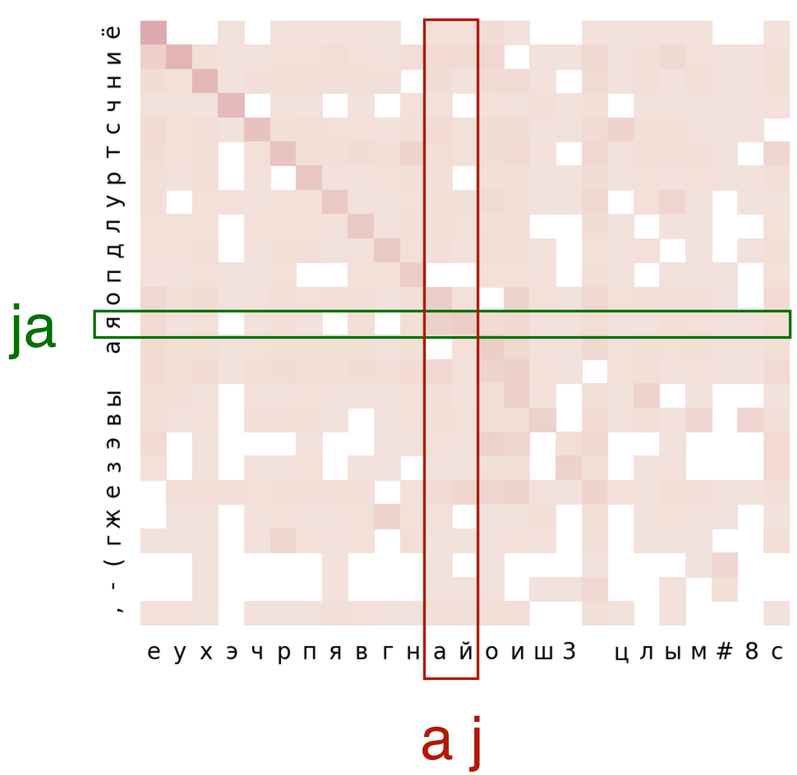
High-level takeaways

- WFST makes more repetitive errors
  - Suggests that WFST outputs might be easier to correct with rule-based postprocessing

WFST



seq2seq



Future work

- Combining unsupervised finite-state and neural models at training time
  - Joint training of two separate models
  - Holistic structural combinations
- Analysis of user preferences
  - Users tend to be consistent in their preferences
  - Substitution choices can reveal user background
  - Might want to obscure these attributes for privacy reasons
- Readability analysis instead of automated metrics

References

C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, M. Mohri. OpenFst: A general and efficient weighted finite-state transducer library. CIAA 2007.

A. Bies, Z. Song, M. Maamouri, S. Grimes, H. Lee, J. Wright, S. Strassel, N. Habash, R. Eskander, O. Rambow. Transliteration of Arabizi into Arabic orthography: Developing a parallel annotated Arabizi-Arabic script SMS/chat corpus. ANLP 2014.

K. Darwish. Arabizi detection and conversion to Arabic. ANLP 2014.

N. Habash, M. Diab, O. Rambow. Conventional orthography for dialectal Arabic. LREC 2012.

J. He, X. Wang, G. Neubig, T. Berg-Kirkpatrick. A probabilistic formulation of unsupervised text style transfer. ICLR 2020.

K. Knight, A. Nair, N. Rathod, K. Yamada. Unsupervised analysis for decipherment problems. COLING/ACL 2006.

G. Lample, A. Conneau, L. Denoyer, M. Ranzato. Unsupervised machine translation using monolingual corpora only. ICLR 2018.

B. Roark, R. Sproat, C. Allauzen, M. Riley, J. Sorensen, T. Tai. The OpenGrm open-source finite-state grammar software libraries. ACL 2012.

B. Roark, L. Wolf-Sonkin, C. Kirov, S. J. Mielke, C. Johny, I. Demirsahin, K. Hall. Processing South Asian languages written in the Latin script: The Dakshina dataset. LREC 2020

M. Ryskina, M. R. Gormley, T. Berg-Kirkpatrick. Phonetic and visual priors for decipherment of informal romanization. ACL 2020.

M. Ryskina, E. Hovy, T. Berg-Kirkpatrick, M. R. Gormley. Comparative error analysis in neural and finite-state models for unsupervised character-level transduction. To appear at SIGMORPHON 2021.

T. Shavrina, O. Shapovalova. To the methodology of corpus construction for machine learning: Taiga syntax tree corpus and parser. CORPORA 2017.