Comparative Error Analysis in Neural and Finite-state Models for Unsupervised Character-level Transduction

Maria Ryskina, Eduard Hovy, Taylor Berg-Kirkpatrick, Matthew R. Gormley

Carnegie Mellon University
Language
Technologies
Institute

ML
MACHINE LEARNING
DEPARTMENT

UCSDCSE
Computer Science and Engineering

Character-level transduction

- Many NLP tasks can be formalized on character level

Pittsburgh → Питтсбург

cipher → S AY F ER

ехать +2.PL.PRS → едем

- Traditionally solved with structured finite-state approaches
- Recently, powerful neural sequence-to-sequence models became dominant

Model classes

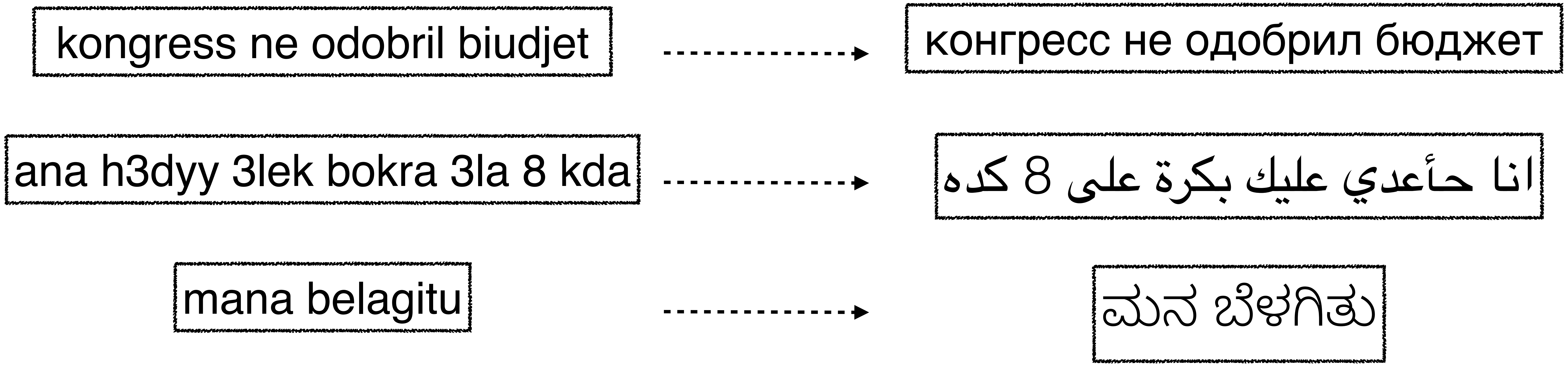
	WFST	Seq2seq
Language model	❌ Character n-gram LM	✅ Stronger RNN LM
Controllability	✅ Easy to encode constraints	❌ Learns orthogonal patterns
Search procedure	✅ Exact maximization	❌ Search errors

Outline

- We compare the two model classes on two unsupervised tasks
    - Deciphering informal romanization
    - Translating between related languages
- We perform error analysis to draw comparisons between models
- We explore simple test-time model combinations
    - Reranking
    - Product of experts

Testbed tasks

1. Converting romanized text to native script (Russian, Arabic, Kannada)



kongress ne odobril biudjet → конгресс не одобрил бюджет

ana h3dyy 3lek bokra 3la 8 kda → انا حاأعدي عليك بكرة على 8 كده

mana belagitu → ಮನ ಬೆಳಗಿತು

Informal romanization

- Informal rendering of non-Latin-script languages in Latin alphabet
- Idiosyncratic: character subsititutions up to the user

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

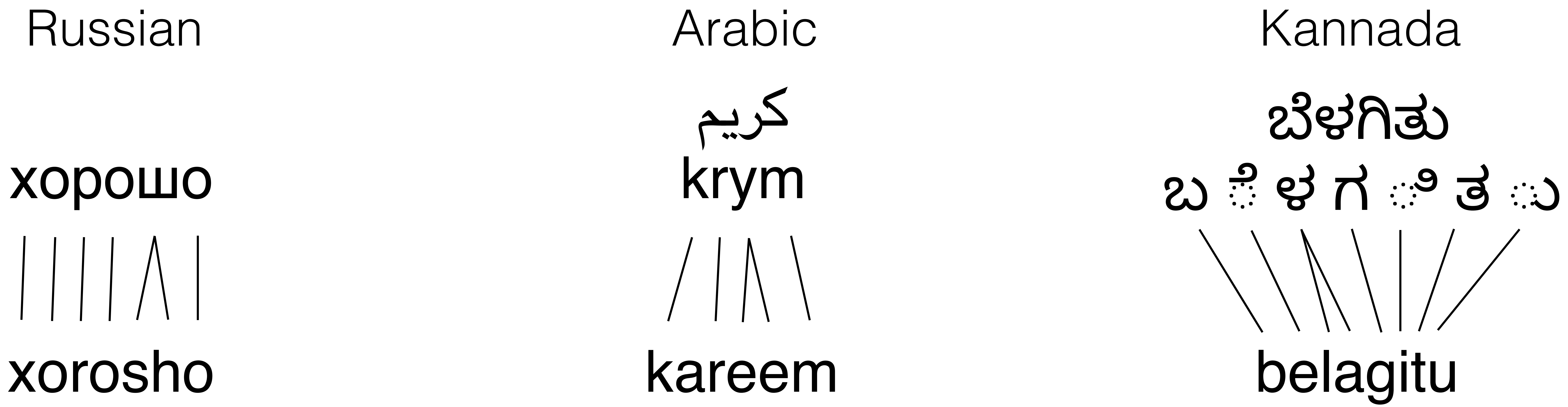
Informal romanization

- Informal rendering of non-Latin-script languages in Latin alphabet
- Idiosyncratic: character subsititutions up to the user
- Character substitution encode similarity (phonetic or visual)

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

Informal romanization

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



Russian

хорошо

xorosho

Arabic

كريم

krym

kareem

Kannada

ಬೆಳಗಿತು

ಬ ೆ ಳ ಗ ಿ ತ ು

belagitu

~ one-to-one

~ one-to-one + null

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

Testbed tasks

1. Converting romanized text to native script (Russian, Arabic, Kannada)

kongress ne odobril biudjet → конгресс не одобрил бюджет

ana h3dyy 3lek bokra 3la 8 kda → انا حأعدي عليك بكرة على 8 كده

mana belagitu → ಮನ ಬೆಳಗಿತು

2. Translating between closely related languages (Serbian and Bosnian)

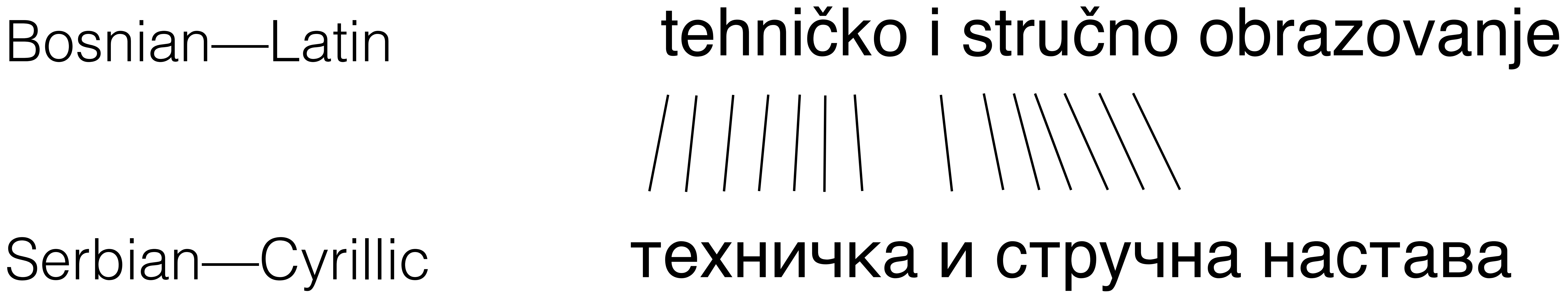
свако има право на живот,
слободу и безбедност личности

↔

svako ima pravo na život,
slobodu i osobnu sigurnost

Translation

- Related languages can have a nearly character-level correspondence…



Bosnian—Latin

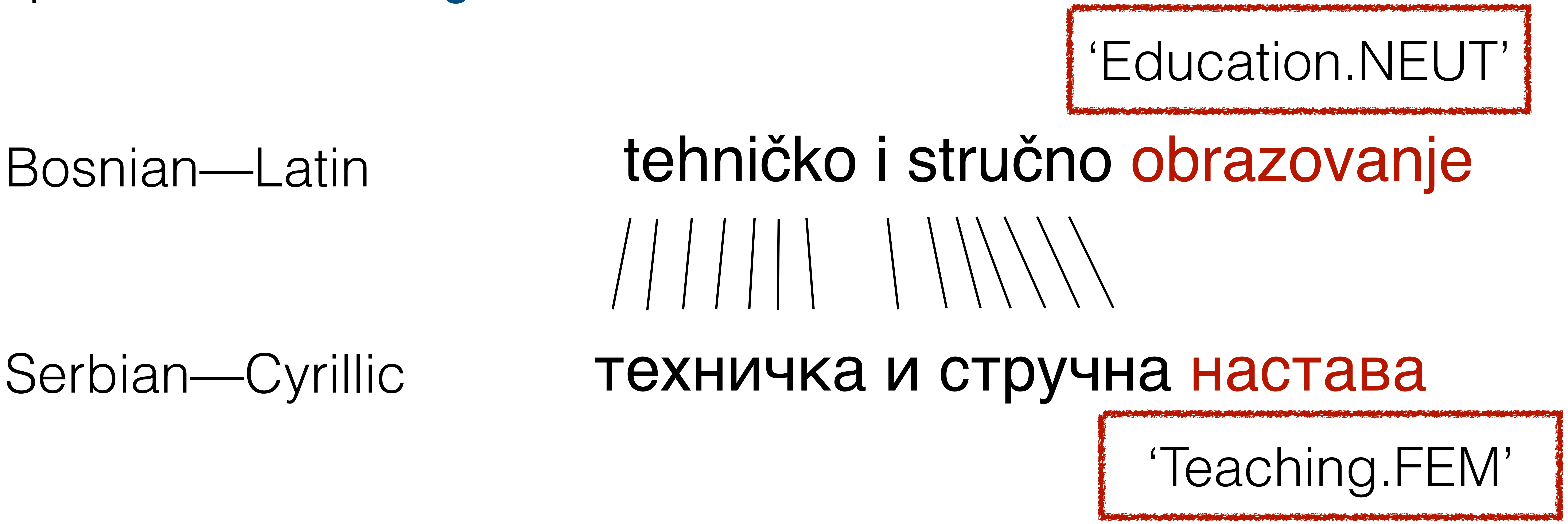
tehničko i stručno obrazovanje

Serbian—Cyrillic

техничка и стручна настава

Translation

- Related languages can have a nearly character-level correspondence…
- …Except for lexical and grammatical differences



‘Education.NEUT’

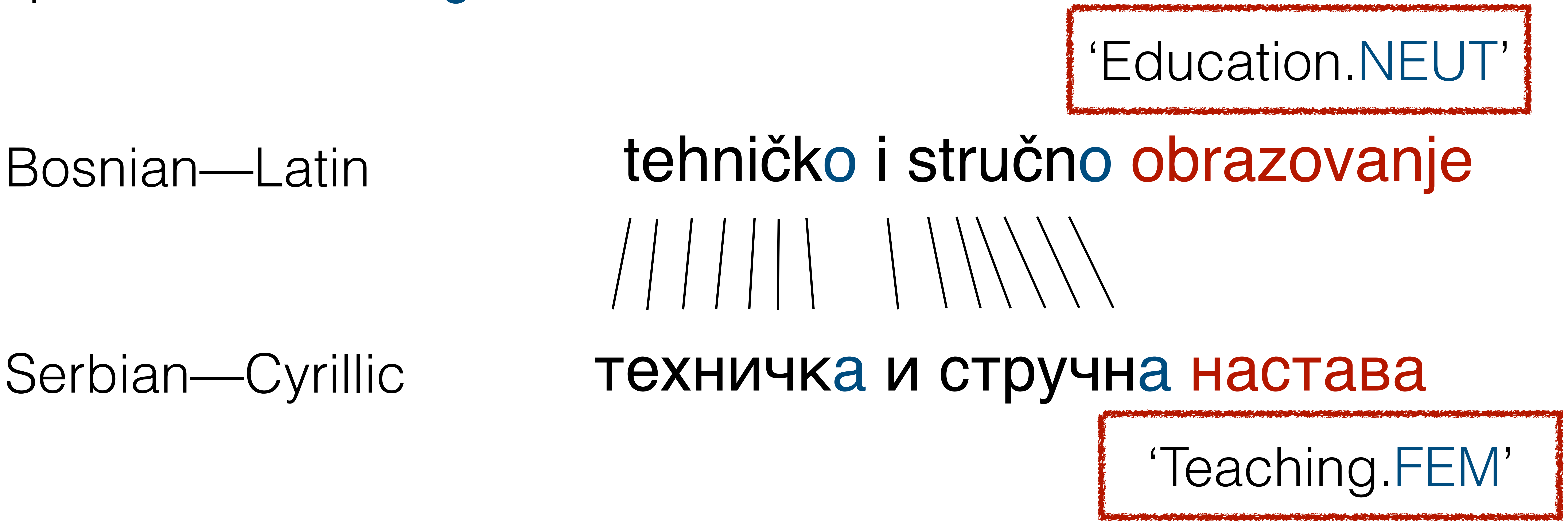
Bosnian—Latin tehničko i stručno obrazovanje

Serbian—Cyrillic техничка и стручна настава

‘Teaching.FEM’

Translation

- Related languages can have a nearly character-level correspondence…
- …Except for lexical and grammatical differences



‘Education.NEUT’

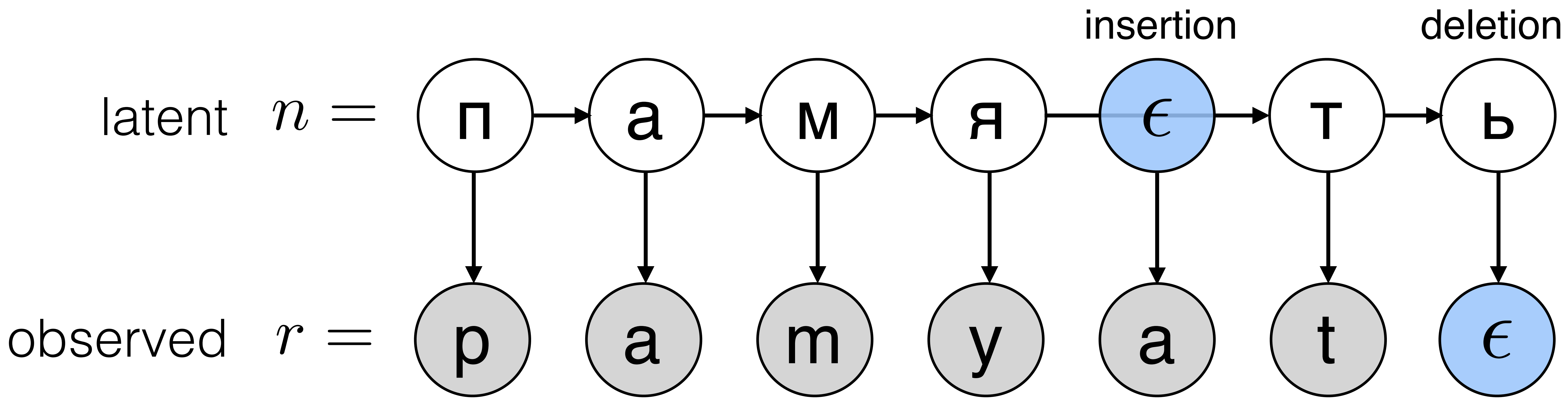
Bosnian—Latin tehničko i stručno obrazovanje

Serbian—Cyrillic техничка и стручна настава

‘Teaching.FEM’

FST: Parameterization

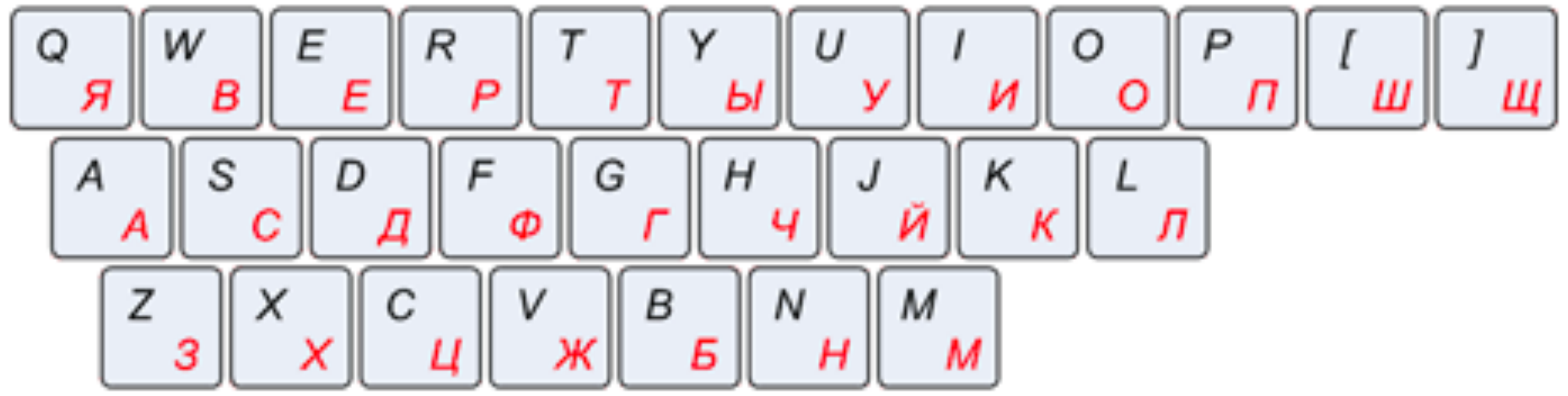
- Noisy channel parameterization (Ryskina et al., 2020)
- Representing character alignment via insertions and deletions

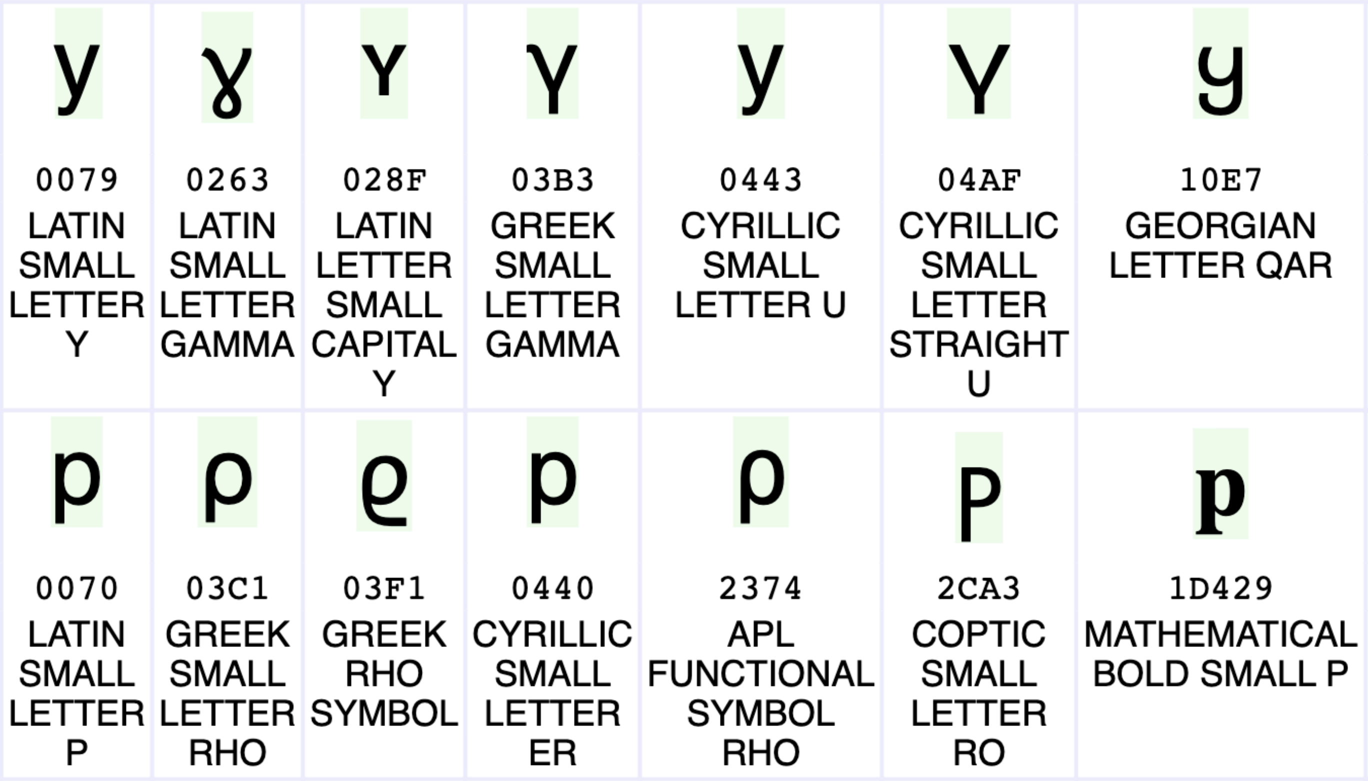


$$p(r) = \sum_n \underbrace{p(n;\gamma)}_{\text{transition probabilities}} \cdot \underbrace{p(r|n;\theta)}_{\text{emission probabilities}} \cdot \underbrace{p_{\text{prior}}(\theta;\alpha)}_{\text{prior on parameters}}$$

FST: Inductive bias

- Phonetic priors: mappings off the phonetic keyboard layouts
- Visual priors: mappings off the Unicode confusables list
- Encoded as priors on emission parameters



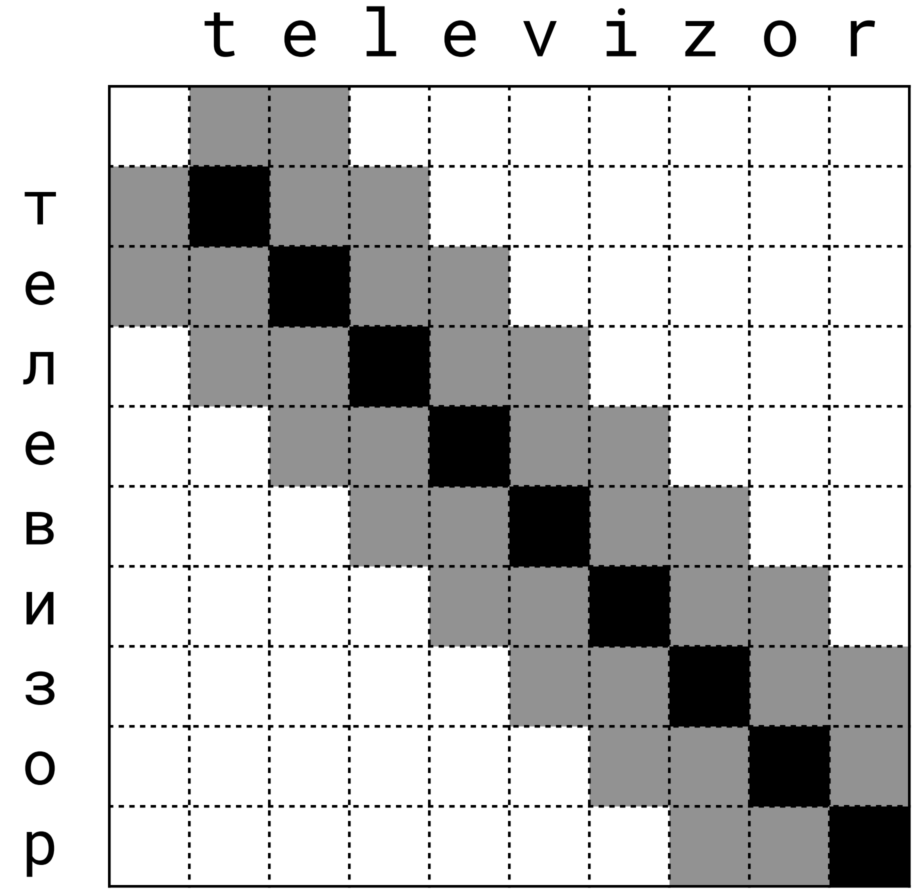


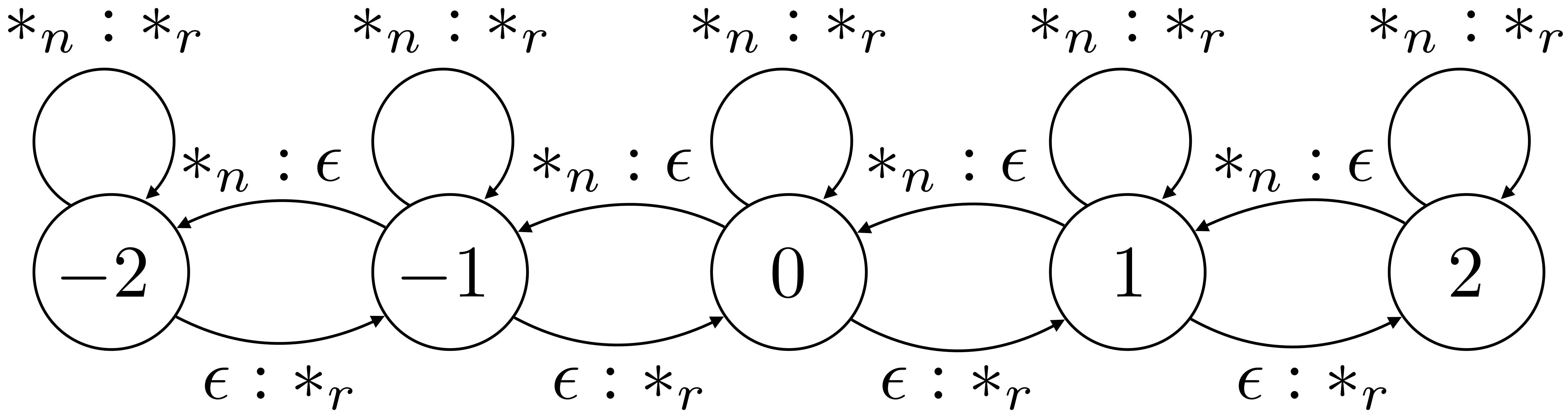
y	ɣ	ʏ	γ	у	ү	ყ
0079	0263	028F	03B3	0443	04AF	10E7
LATIN SMALL LETTER Y	LATIN SMALL LETTER GAMMA	LATIN LETTER SMALL CAPITAL Y	GREEK SMALL LETTER GAMMA	CYRILLIC SMALL LETTER U	CYRILLIC SMALL LETTER STRAIGHT U	GEORGIAN LETTER QAR
p	ρ	ϱ	р	⍴	ⲣ	𝐩
0070	03C1	03F1	0440	2374	2CA3	1D429
LATIN SMALL LETTER P	GREEK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

https://en.wikipedia.org/wiki/Phonetic_keyboard_layout,
https://util.unicode.org/UnicodeJsps/confusables.jsp

FST: Implementation

- Transition WFSA
    - 6-gram character-level language model
- Emission WFST
    - Supports all substitutions, insertions and deletions
    - Fixed limit on delay: |# of insertions - # of deletions|



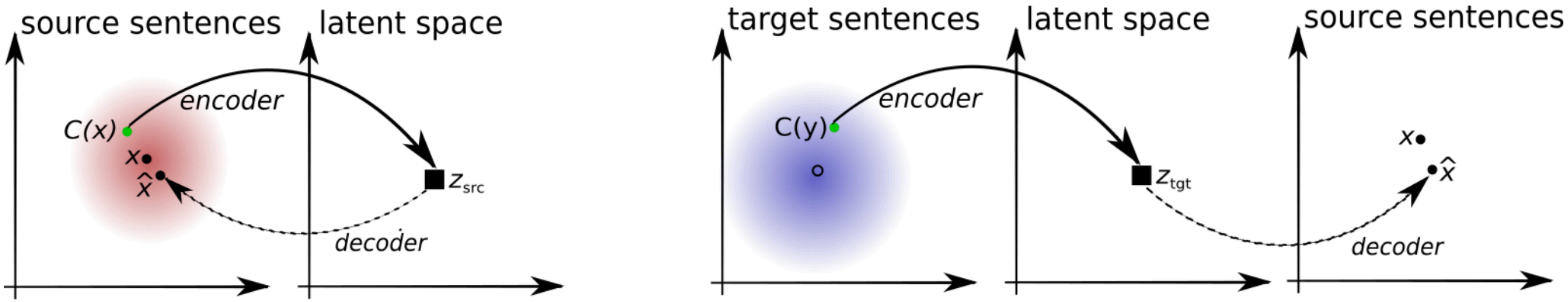


FST: Implementation

- Transition WFSA
  - 6-gram character-level language model
- Emission WFST
  - Supports all substitutions, insertions and deletions
  - Fixed limit on delay: |# of insertions - # of deletions|
- Trained with ‘hard’ EM algorithm
  - OpenFst (Allauzen et al., 2007)
  - Only a subset of shortest sequences used for training!

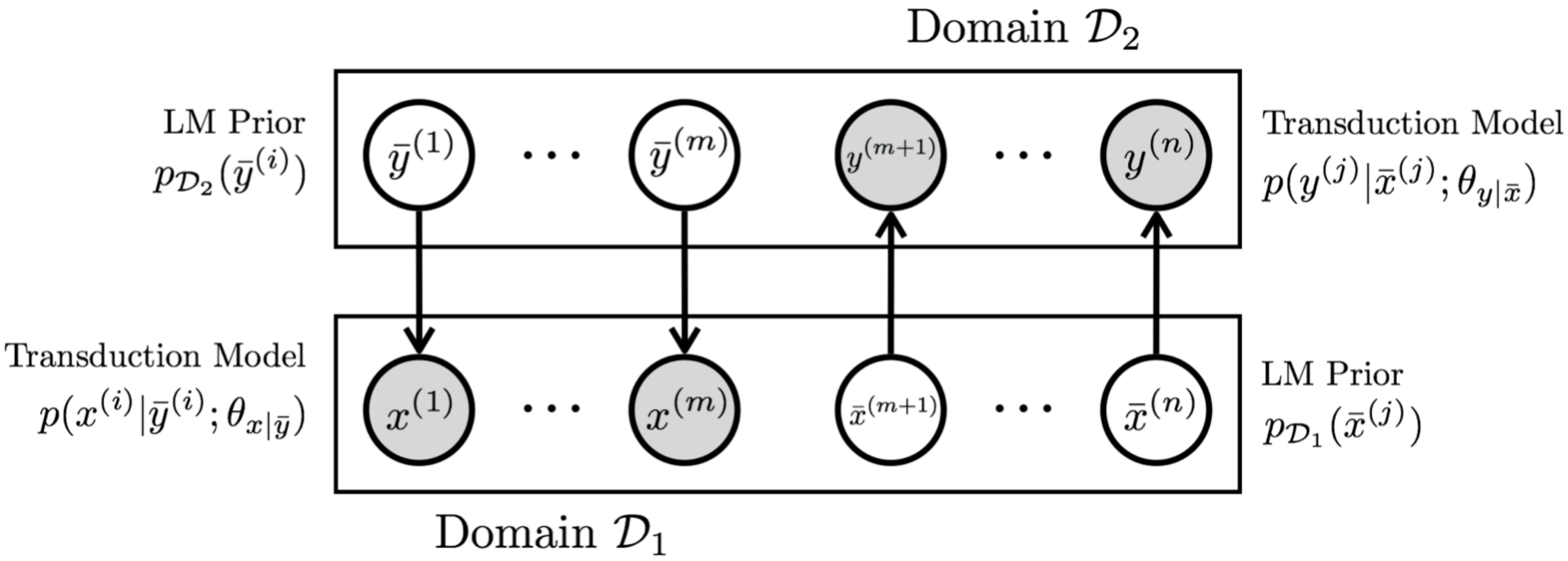
Seq2seq model

- 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



Seq2seq model

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

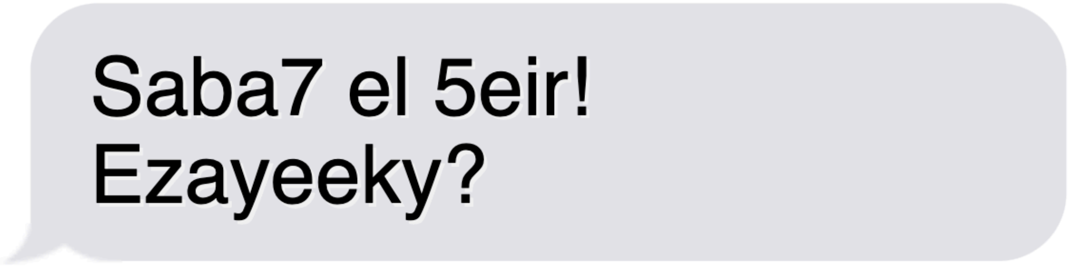


Model combinations

- 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

Romanization data

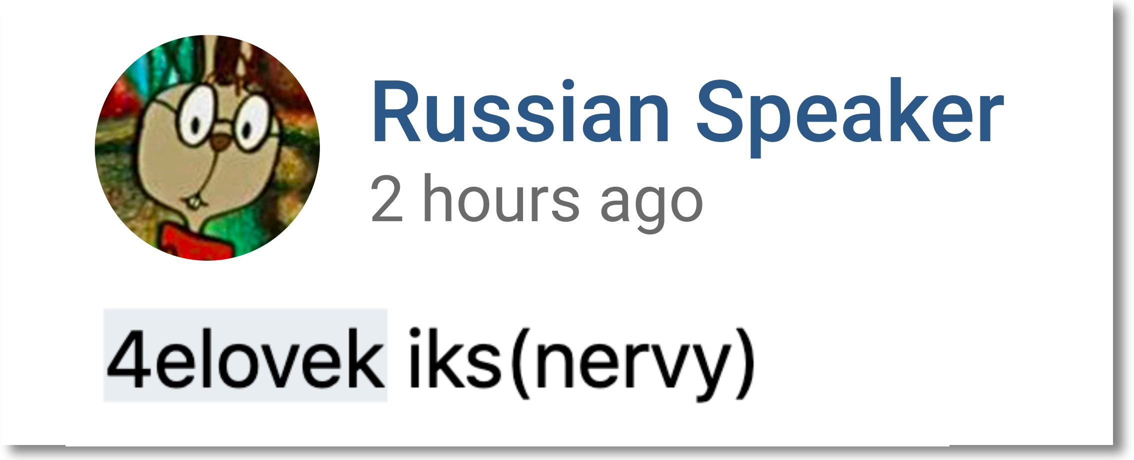
- Arabic: LDC BOLT dataset (Bies et al., 2014)
  - Arabizi SMS/chat dialogs



Saba7 el 5eir!
Ezayeeky?

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

- Russian:
  - Romanized: vk.com comments (Ryskina et al., 2020)
  - Native: Taiga (Shavrina & Shapovalova, 2017), vk.com comments from political groups



Russian Speaker
2 hours ago

4elovek iks(nervy)

Translation data

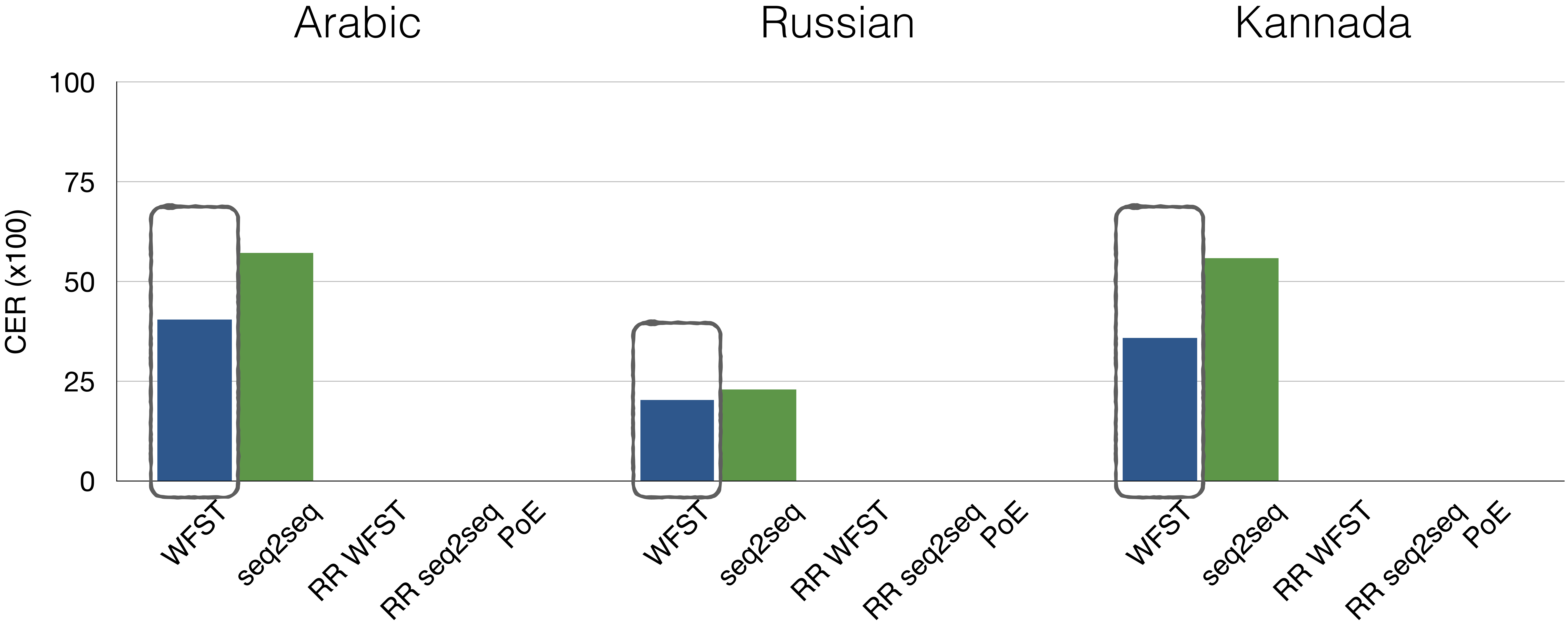
- Monolingual data: Leipzig corpora (Goldhahn et al., 2012)

- Parallel validation data: synthetic (Yang et al., 2018)
    - Machine-translated portions of Leipzig corpora

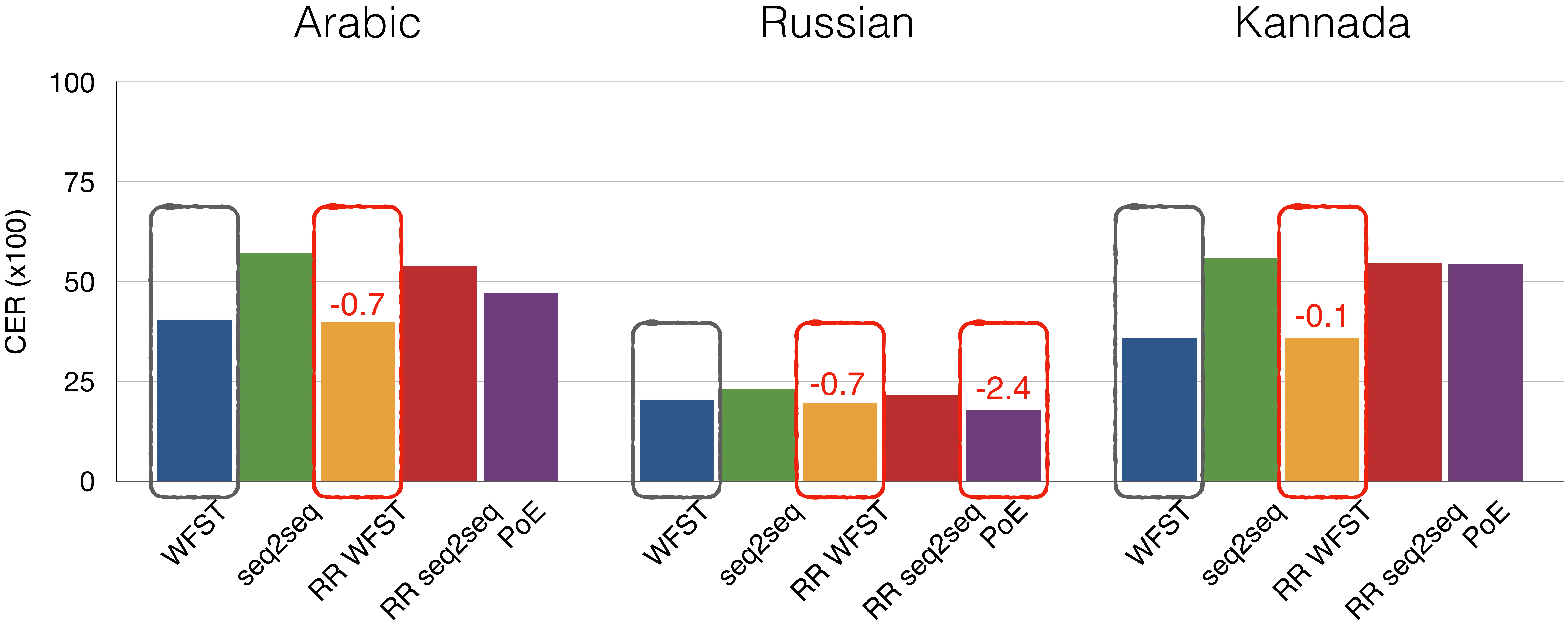
- Parallel test data: Universal Declaration of Human Rights

Romanization results

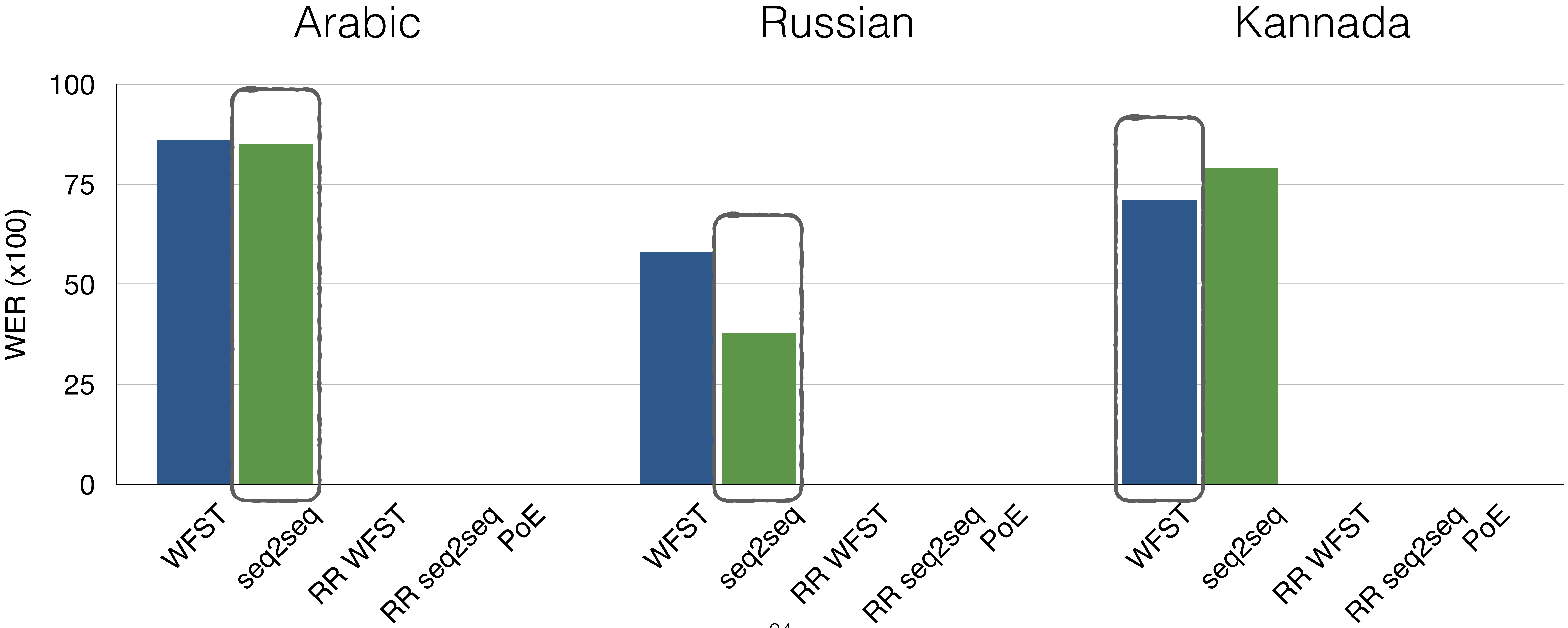
Base models are trained on different amounts of data!



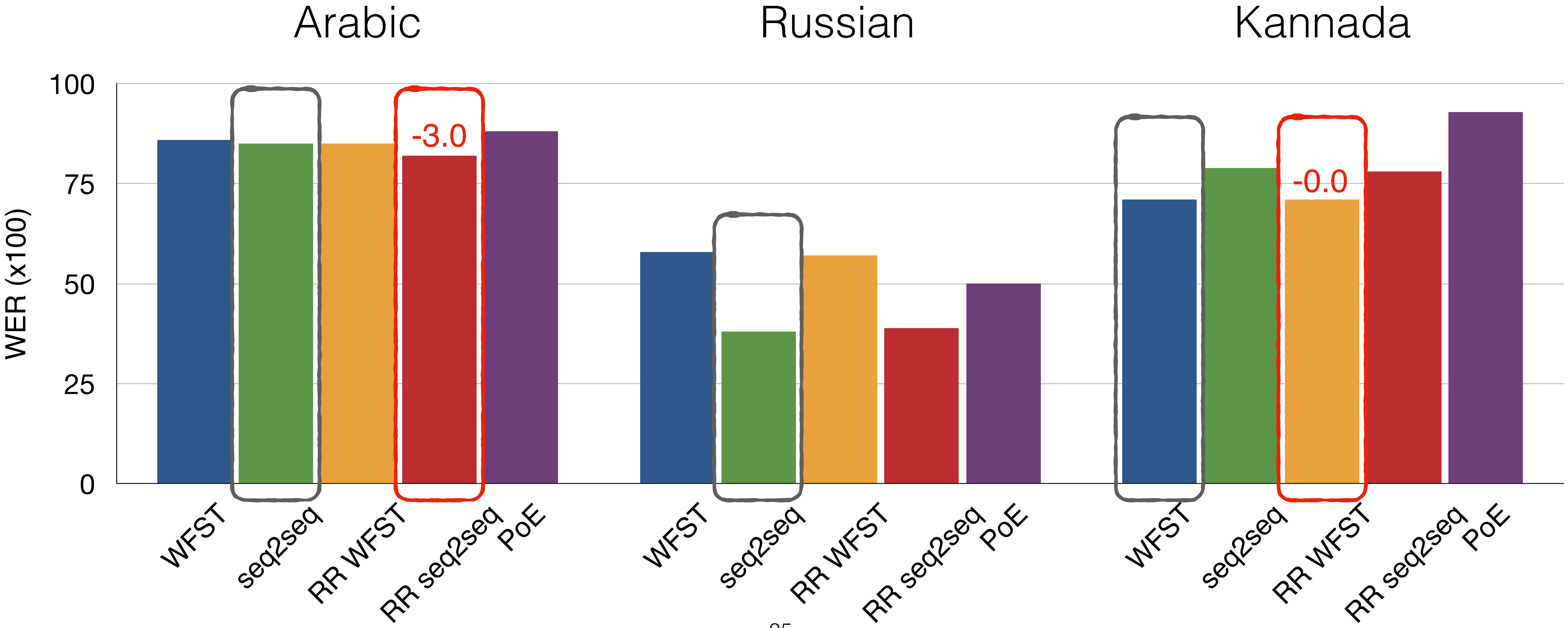
Romanization results



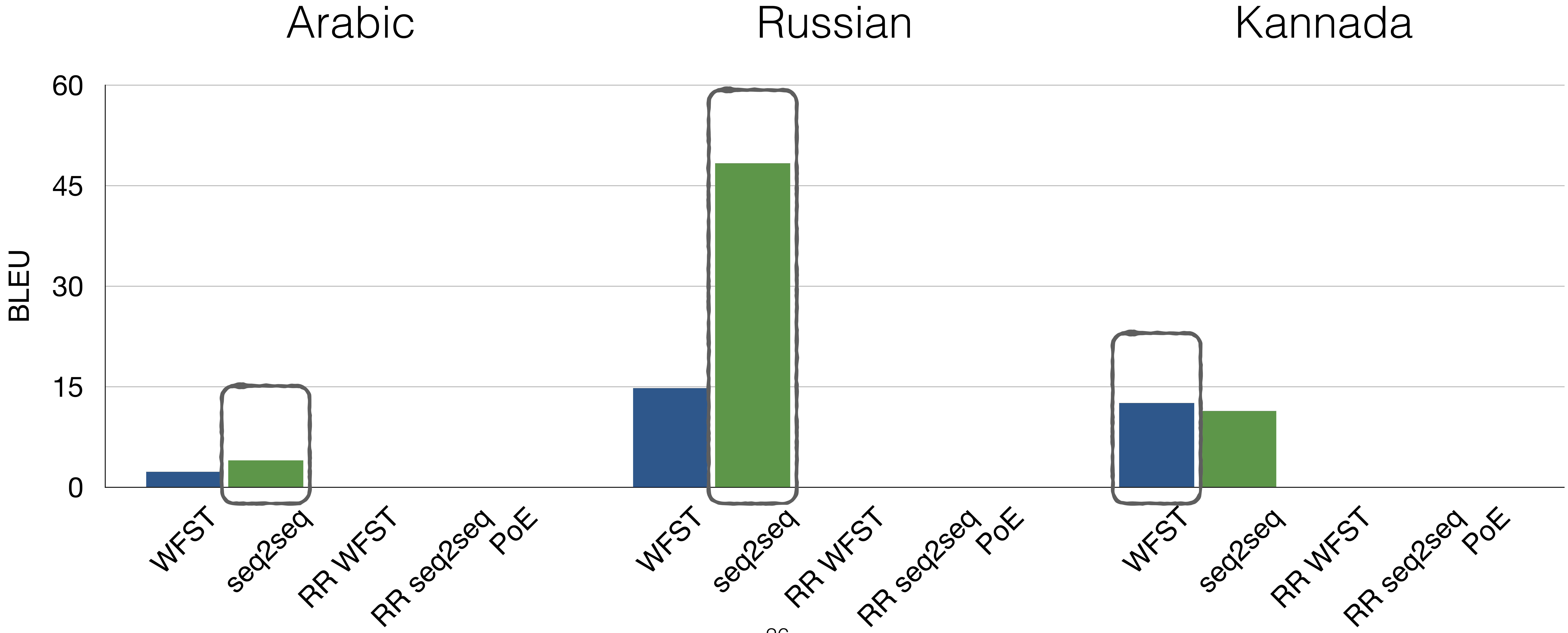
Romanization results



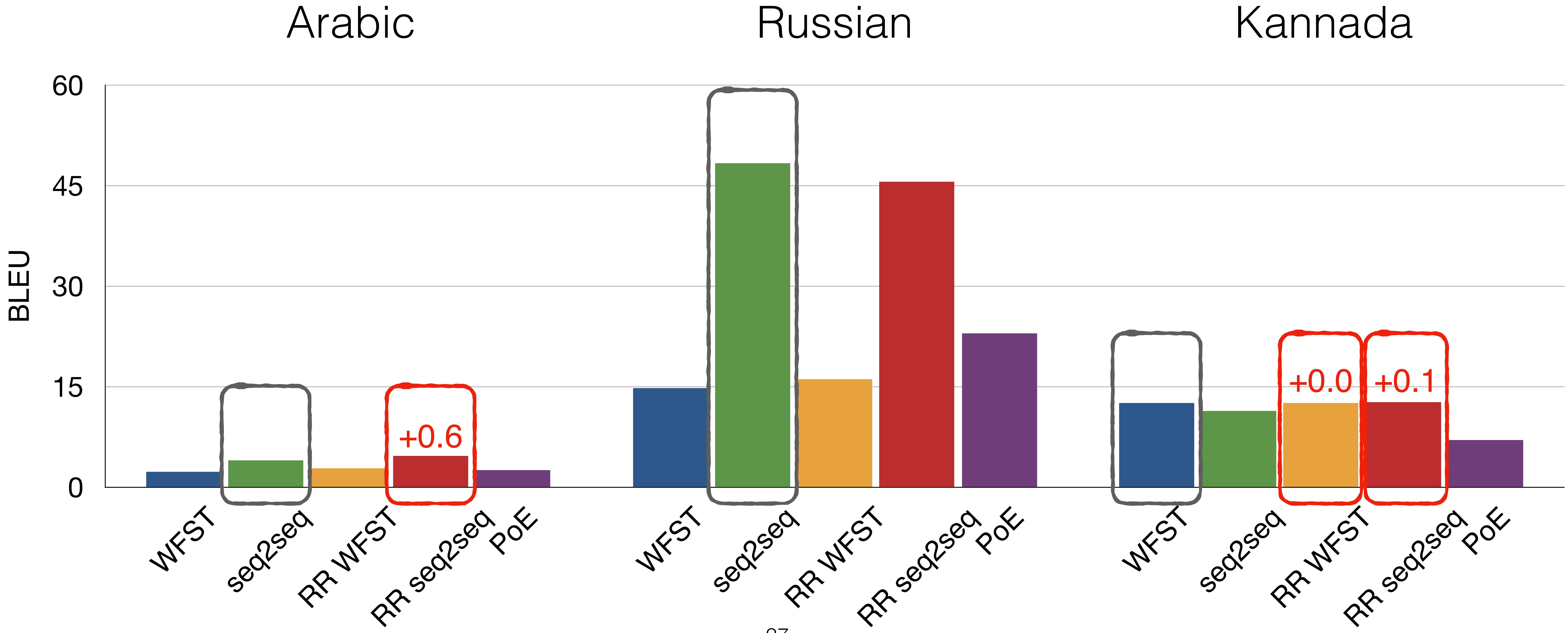
Romanization results



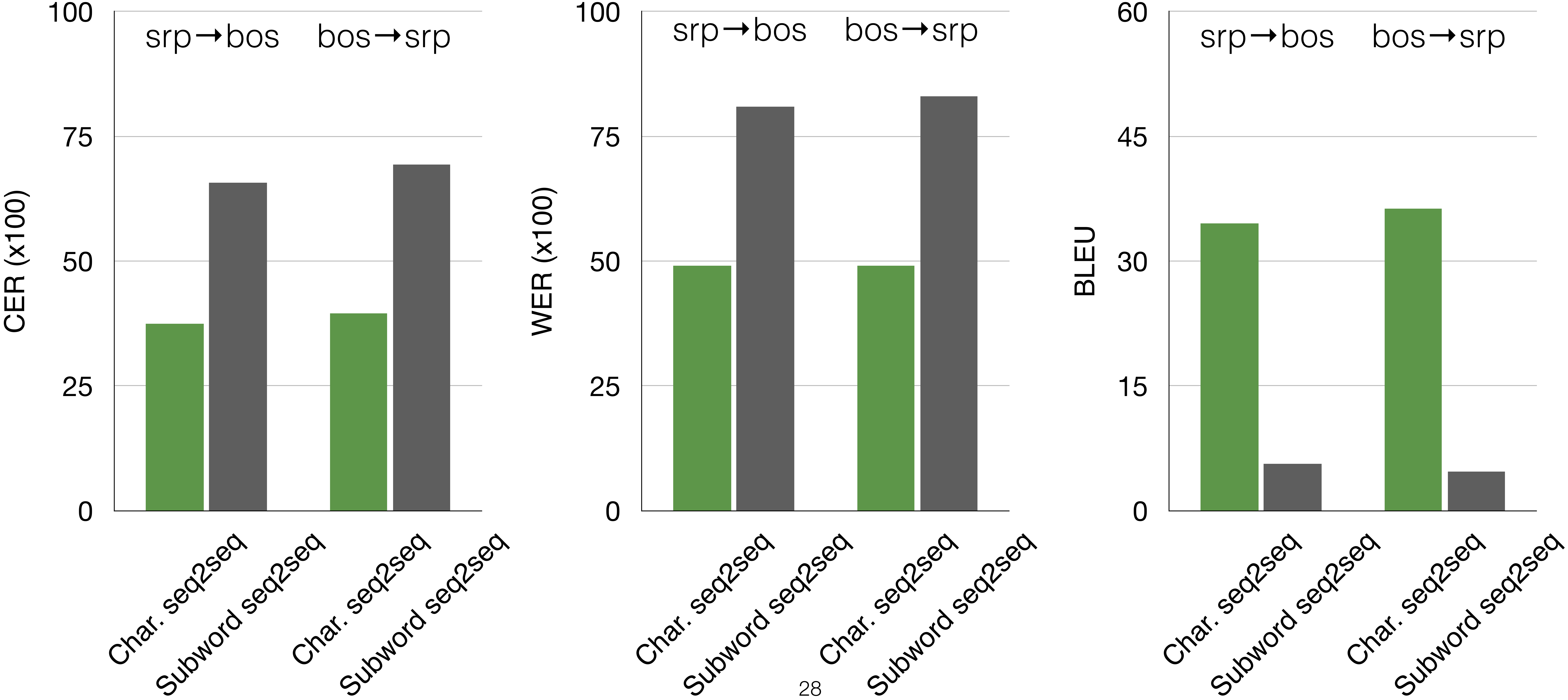
Romanization results



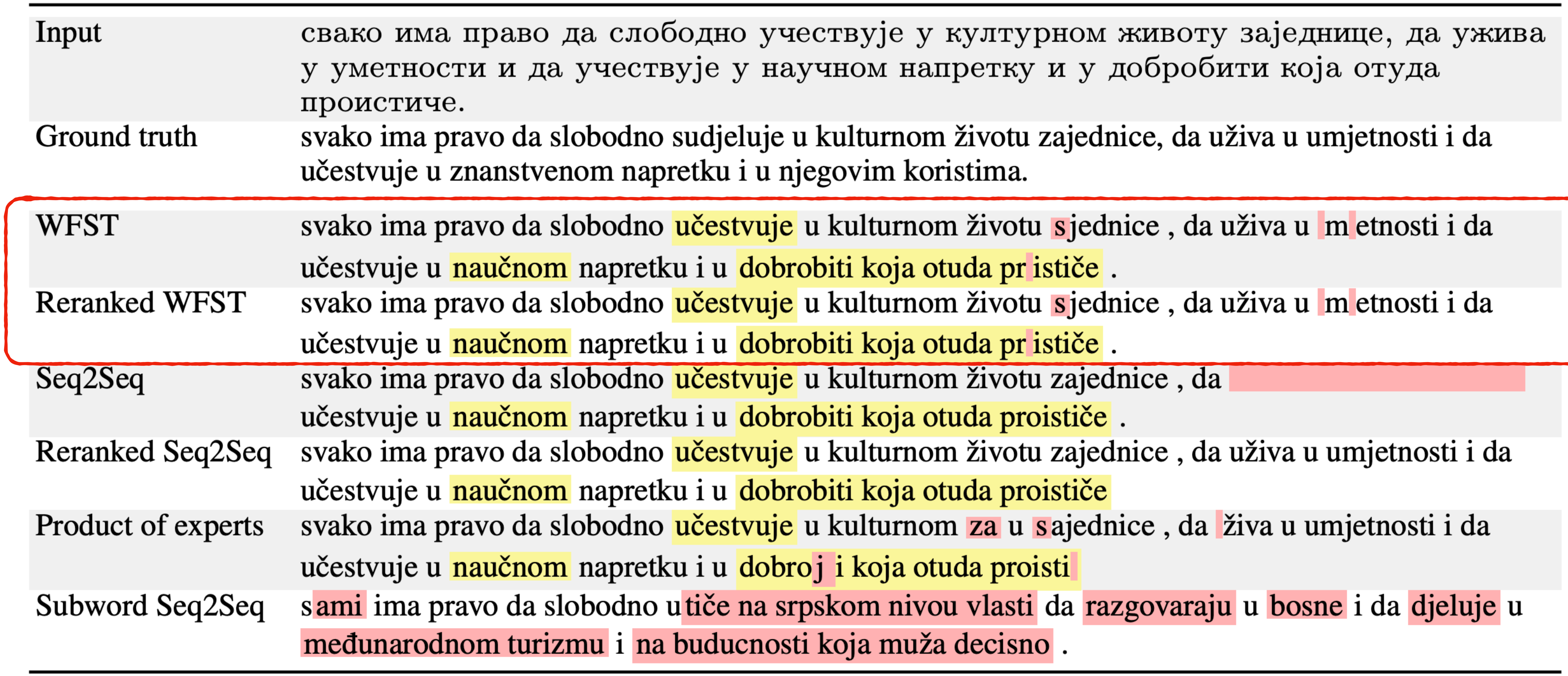
Romanization results



Translation results



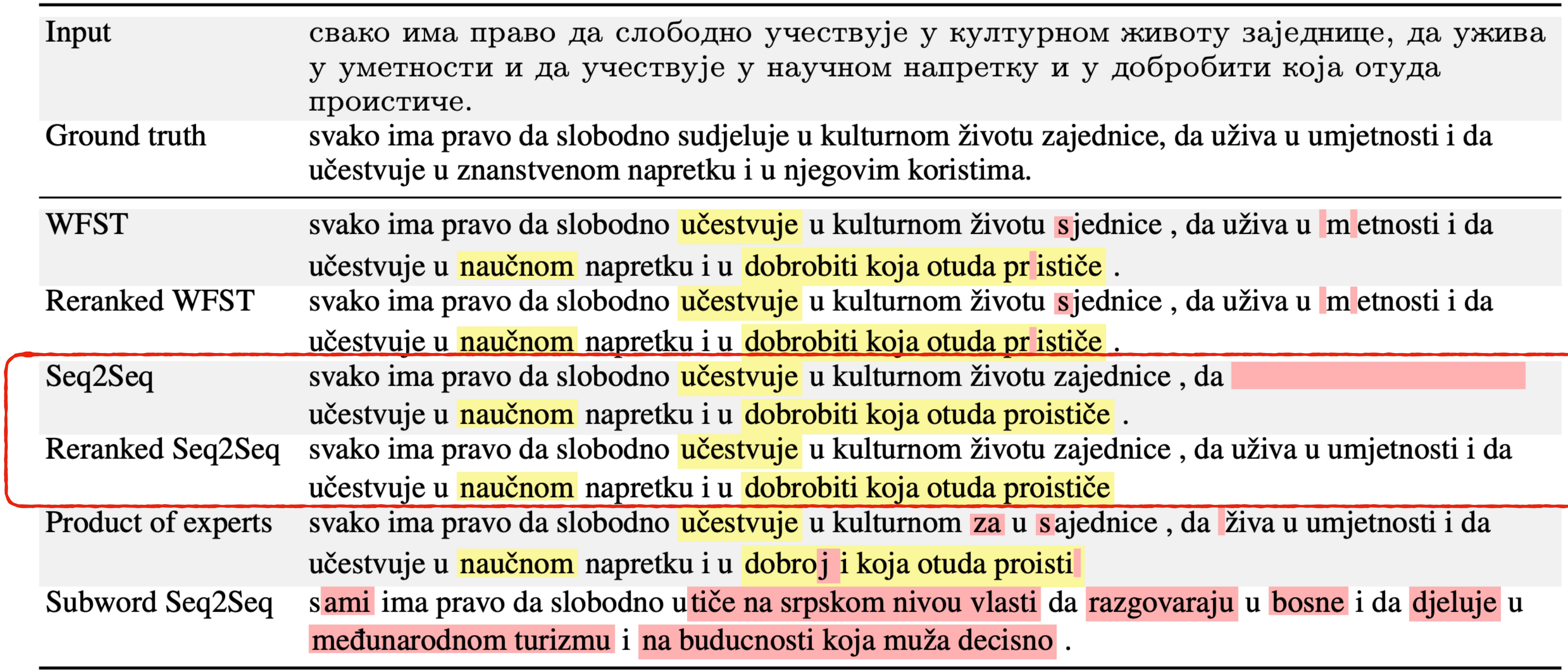
Error analysis



Input	свако има право да слободно учествује у културном животу заједнице, да ужива у уметности и да учествује у научном напретку и у добробити која отуда проистиче.
Ground truth	svako ima pravo da slobodno sudjeluje u kulturnom životu zajednice, da uživa u umjetnosti i da učestvuje u znanstvenom napretku i u njegovim koristima.
WFST	svako ima pravo da slobodno učestvuje u kulturnom životu sjednice , da uživa u metnosti i da učestvuje u naučnom napretku i u dobrobiti koja otuda pr ističe .
Reranked WFST	svako ima pravo da slobodno učestvuje u kulturnom životu sjednice , da uživa u metnosti i da učestvuje u naučnom napretku i u dobrobiti koja otuda pr ističe .
Seq2Seq	svako ima pravo da slobodno učestvuje u kulturnom životu zajednice , da učestvuje u naučnom napretku i u dobrobiti koja otuda proističe .
Reranked Seq2Seq	svako ima pravo da slobodno učestvuje u kulturnom životu zajednice , da uživa u umjetnosti i da učestvuje u naučnom napretku i u dobrobiti koja otuda proističe
Product of experts	svako ima pravo da slobodno učestvuje u kulturnom za u sajednice , da živa u umjetnosti i da učestvuje u naučnom napretku i u dobro j i koja otuda proisti
Subword Seq2Seq	sami ima pravo da slobodno utiče na srpskom nivou vlasti da razgovaraju u bosne i da djeluje u međunarodnom turizmu i na buducnosti koja muža decisno .

Character-level mistakes

Error analysis

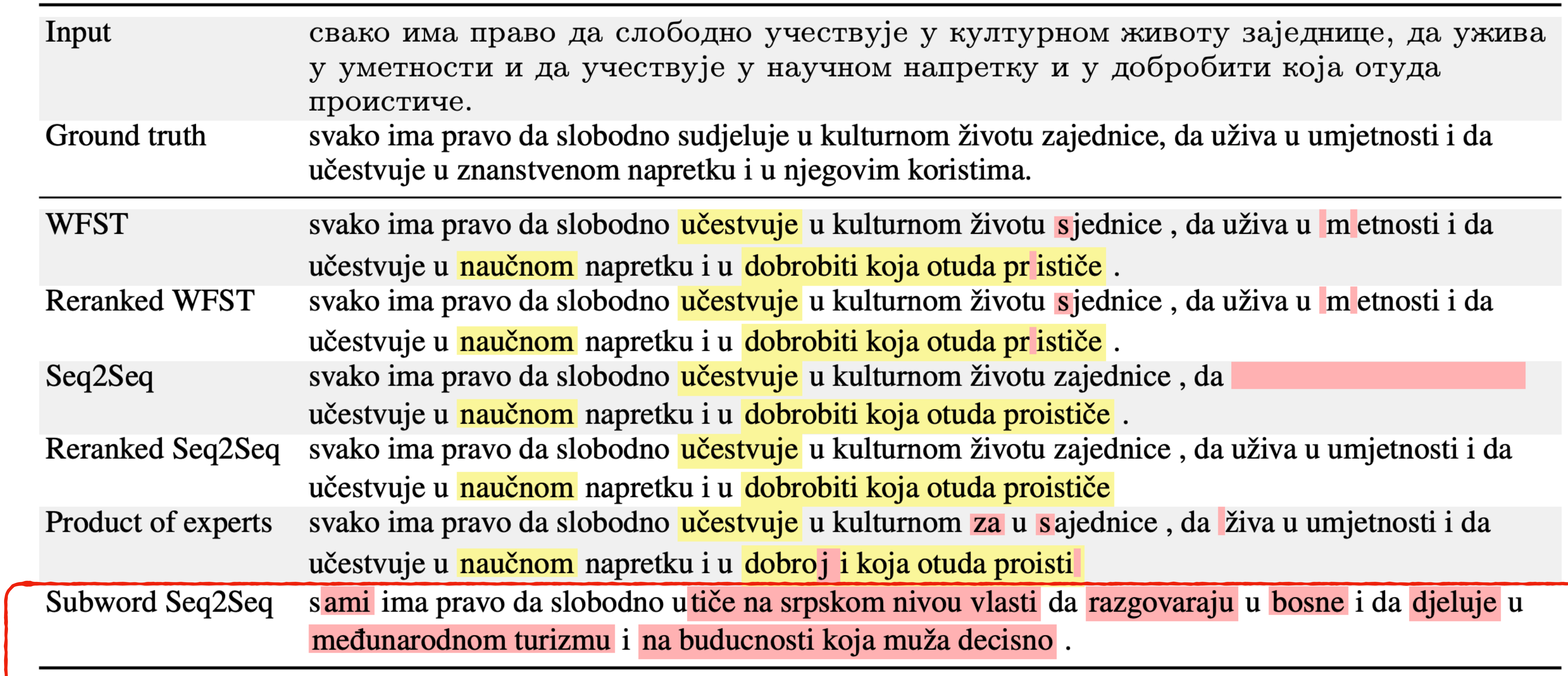


Input	свако има право да слободно учествује у културном животу заједнице, да ужива у уметности и да учествује у научном напретку и у добробити која отуда проистиче.
Ground truth	svako ima pravo da slobodno sudjeluje u kulturnom životu zajednice, da uživa u umjetnosti i da učestvuje u znanstvenom napretku i u njegovim koristima.
WFST	svako ima pravo da slobodno učestvuje u kulturnom životu s jednice , da uživa u m etnosti i da učestvuje u naučnom napretku i u dobrobiti koja otuda pr ističe .
Reranked WFST	svako ima pravo da slobodno učestvuje u kulturnom životu s jednice , da uživa u m etnosti i da učestvuje u naučnom napretku i u dobrobiti koja otuda pr ističe .
Seq2Seq	svako ima pravo da slobodno učestvuje u kulturnom životu zajednice , da učestvuje u naučnom napretku i u dobrobiti koja otuda proističe .
Reranked Seq2Seq	svako ima pravo da slobodno učestvuje u kulturnom životu zajednice , da uživa u umjetnosti i da učestvuje u naučnom napretku i u dobrobiti koja otuda proističe
Product of experts	svako ima pravo da slobodno učestvuje u kulturnom za u s ajednice , da živa u umjetnosti i da učestvuje u naučnom napretku i u dobro j i koja otuda proisti
Subword Seq2Seq	s ami ima pravo da slobodno u tiče na srpskom nivou vlasti da razgovaraju u bosne i da djeluje u međunarodnom turizmu i na buducnosti koja muža decisno .

Word deletion

Incorrect but faithful

Error analysis



Input	свако има право да слободно учествује у културном животу заједнице, да ужива у уметности и да учествује у научном напретку и у добробити која отуда проистиче.
Ground truth	svako ima pravo da slobodno sudjeluje u kulturnom životu zajednice, da uživa u umjetnosti i da učestvuje u znanstvenom napretku i u njegovim koristima.
WFST	svako ima pravo da slobodno učestvuje u kulturnom životu sjednice , da uživa u metnosti i da učestvuje u naučnom napretku i u dobrobiti koja otuda prističe .
Reranked WFST	svako ima pravo da slobodno učestvuje u kulturnom životu sjednice , da uživa u metnosti i da učestvuje u naučnom napretku i u dobrobiti koja otuda prističe .
Seq2Seq	svako ima pravo da slobodno učestvuje u kulturnom životu zajednice , da učestvuje u naučnom napretku i u dobrobiti koja otuda proističe .
Reranked Seq2Seq	svako ima pravo da slobodno učestvuje u kulturnom životu zajednice , da uživa u umjetnosti i da učestvuje u naučnom napretku i u dobrobiti koja otuda proističe
Product of experts	svako ima pravo da slobodno učestvuje u kulturnom za u sajednice , da živa u umjetnosti i da učestvuje u naučnom napretku i u dobroj i koja otuda proisti
Subword Seq2Seq	sami ima pravo da slobodno utiče na srpskom nivou vlasti da razgovaraju u bosne i da djeluje u međunarodnom turizmu i na buducnosti koja muža decisno .

Effects of tokenization

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’ ow 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

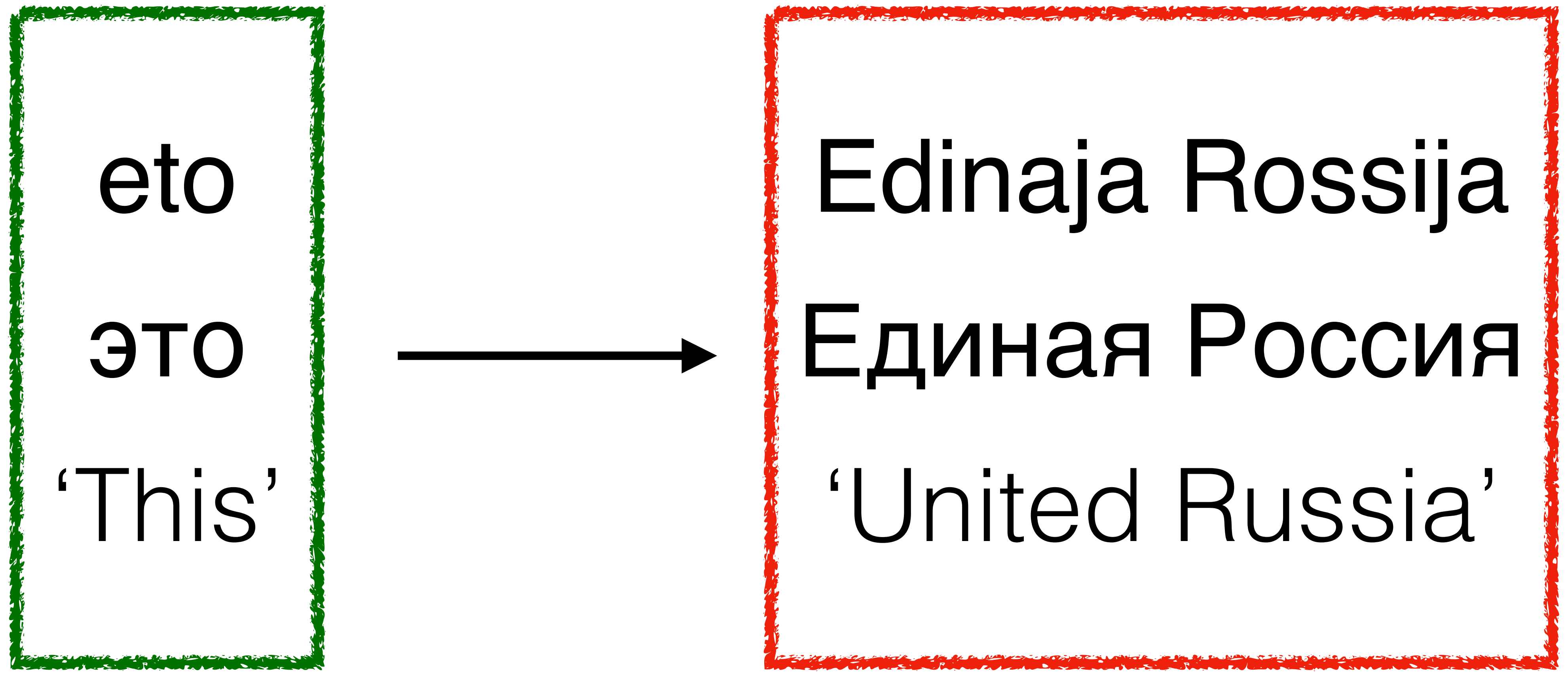
Gloss: ‘By now this is (I’m almost afraid to say it) an old video’

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 most frequent substitions under the seq2seq are caused by domain mismatch, compared to 3% for WFST!



eto

это

‘This’

Edinaja Rossija

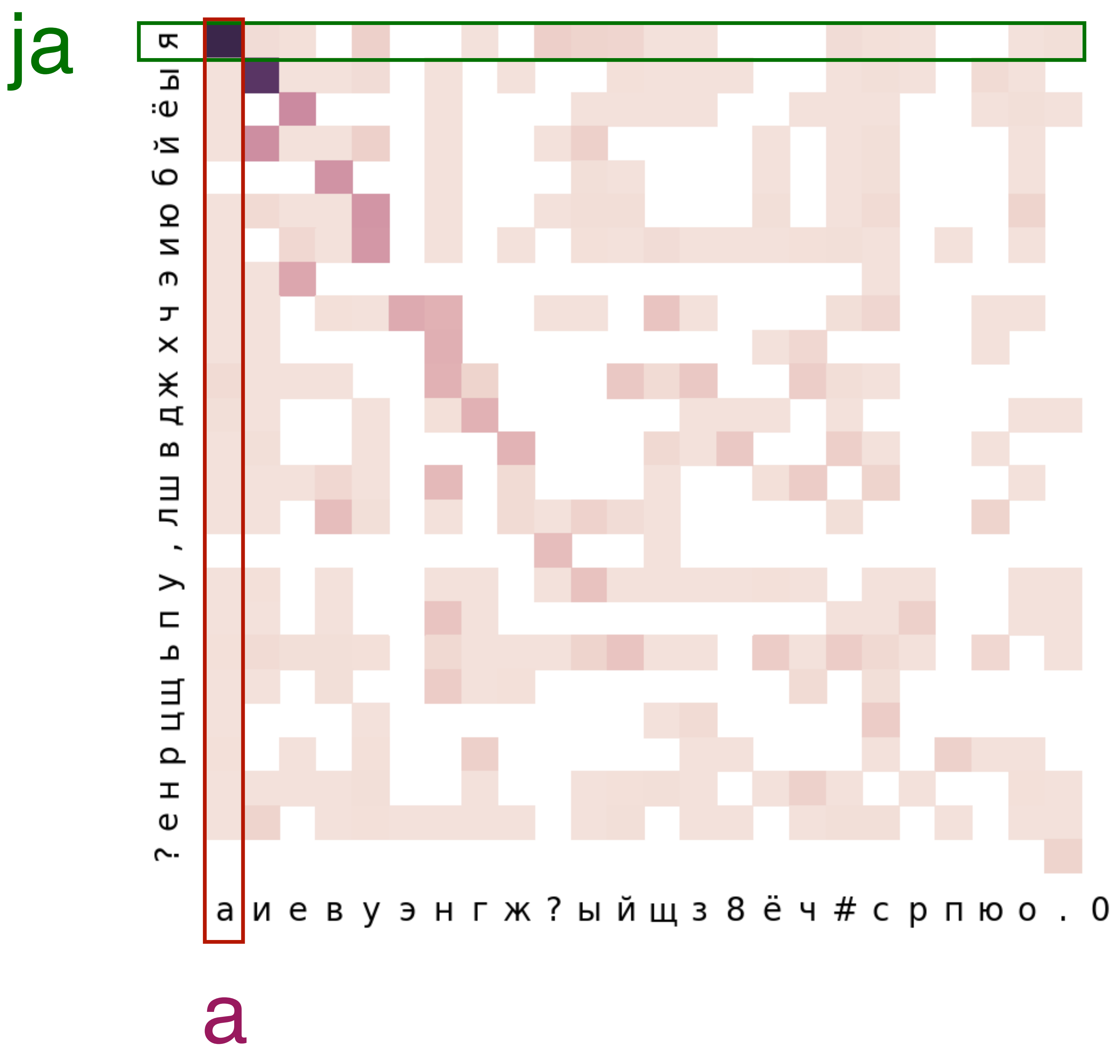
Единая Россия

‘United Russia’

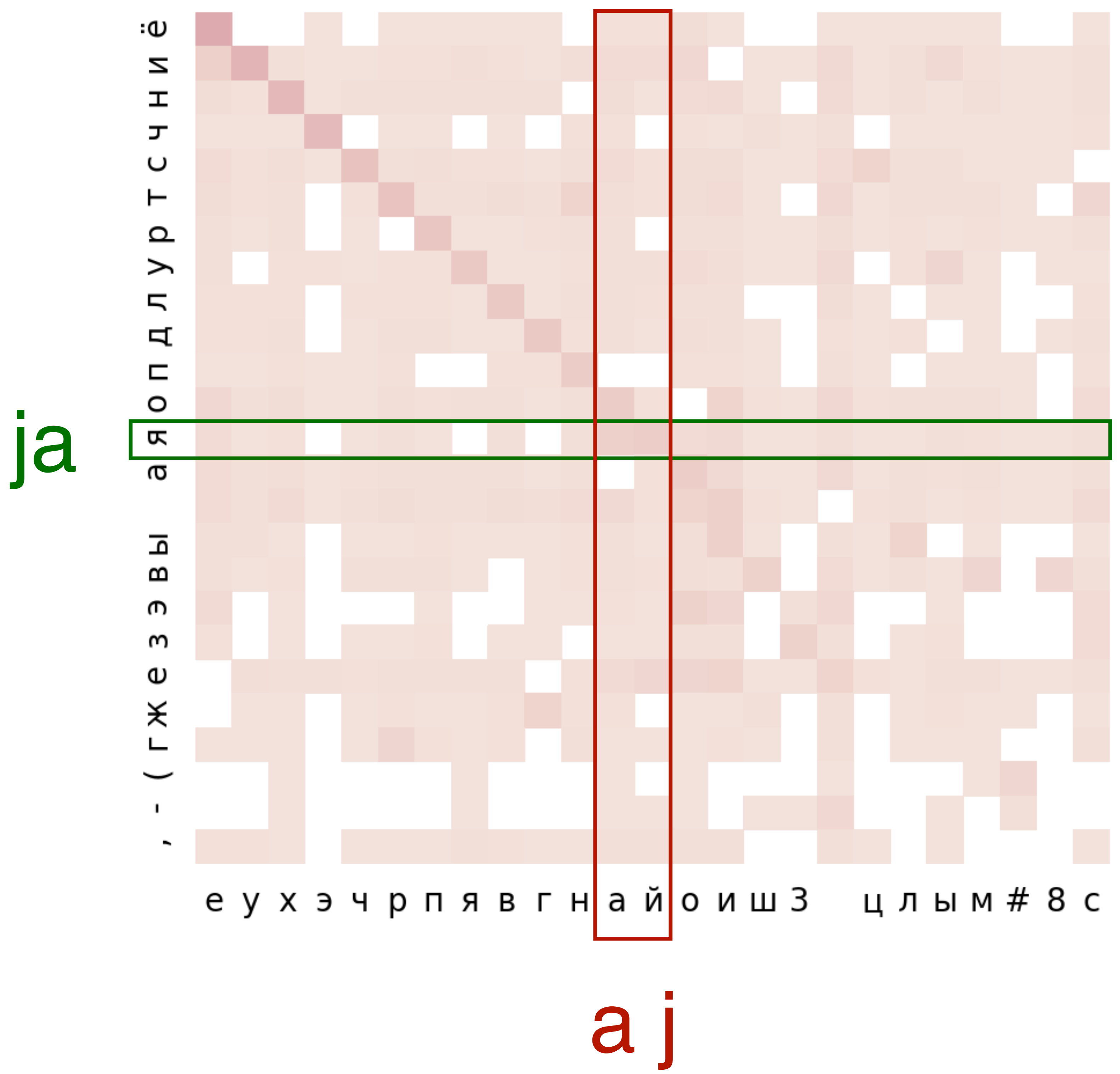
High-level takeaways

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

WFST



seq2seq



Future work

- Can enough ‘power’ replace ‘structure’?
    - Transformer can learn character-level transduction without structural constraints (Wu et al., 2021)
    - But less likely to suffice in unsupervised or low-data settings!
- More promising combinations of unsupervised finite-state and neural models
    - Joint training
    - Holistic structural combinations
    - Biasing one model towards another model’s behavior

Thank you!

Link to paper:

Questions?

✉ mryskina@cs.cmu.edu

@maria_ryskina