Phonotactics as an Aid in Low Resource Loan Word Detection and Morphological Analysis in Sakha

Petter Mæhlum and Sardana Ivanova

University of Oslo, University of Helsinki

30 March, 2023, LTG Seminar

Motivation

- Loanwords behave differently and can cause problems for tasks such as POS-tagging and parsing
- Large-scale annotations or more advanced techniques might not be readily available for low-resource languages
- When the loanword source language differs significantly from the target language, phonotactics can be useful
- We suggest that phonotactics can be a useful pre-prosessing step when other methods are not available
- We explore vowel harmony and consonant restrictions to investigate loanwords and morphology

Phonotactics

- The study of how phonemes combine, "syntax of phonemes"
- All languages have restrictions on which sounds go together, both consonants and vowels.
- This explains, along with the phoneme inventory of a language, why languages that have similar sounding phonemes still cannot arrange them in the same way, e.g. why Norwegian has /h/, /a/ // but a word like /ha/ is ok, but /ah/ is not.
- Change over time, but can be relatively reliable at a given point in time.
- Many rules are quick to establish if not already present
- It helps if the orthography reflects pronunciation.

Vowel harmony

- Governs which vowels can follow which, both inside words and suffixes
- Found in all Turkic languages except Uzbek
- Also found in Finnish, Mongolian and Korean, among others
- Varies in degree, and does not necessarily apply to loanwords

Sakha

- Turkic language with 450 000 speakers.
- Spoken primarily in Siberia.
- Agglutinating and verb final (SOV).
- Suffixes largely follow vowel harmony and consonant assimilation rules

Vowel harmony examples

- Finnish a, o, u vs. ä, ö, y, with e and i being neutral.
- Korean u harmonizes with eo, and o harmonizes with a.
- Turkish has two-way harmony: A-harmony and I-harmony.
- A-harmony (back vs front)
  - a o u ɪ -> a
  - e i ü o -> e
- I-harmony (backness + rounding)
  - e i -> i
  - a ɪ -> ɪ
  - ü ö -> ü
  - o u -> u

Sakha Vowel harmony

- Sakha vowels are similar to Turkish, but in addition, Sakha has four diphthongs (ie,ıa, uo, üö), and a distinction between long and short vowels. Vowels harmonize with their close counterpart.
- The system is quite symmetric, but the two close vowels ü and u harmonize as if they were not rounded, and the same goes for their diphthongs üö and uo.
- The group corresponding to the Turkish A-group is more fine-grained.

	Front		Back	
	close	open	close	open
Unrounded	и [i]	э [e]	ы [ɪ]	а [a]
Rounded	ү [ü]	ө [ö]	у [u]	о [o]

Таблица: Sakha vowels according to their features.

Sakha Vowel harmony

улуус-тар-ыгар [uluus-tar-ıgar]

district–PL-DAT

‘To their district’

көр-сүөх-хэ [kör-süöx-xe]

see-REFL-COH

‘Let’s see each other’

аһаа-ты-быт [ahaa-tı-bıt]

eat-PAST-1P.PL

‘We ate’

Hyphenated Compounds

Sakha has several hyphenated compounds, consisting some times of nouns or verbs that have similar but different meanings, to mean something more general.

1. **от-мас** [ot-mas] ‘grass-tree’, i.e ‘plants’
2. **ahaa-сиэ** [ahaa-sie] ‘eat (intrans.)-eat (trans.)’, i.e. ‘eat’.

We see their endings harmonize for each component in cases such as **ahaa-сиэ** [ahaa-sie] which is **aһыыр-сиир** [ahıır-siir] ‘eats’ in the present tense..

Consonant Restrictions

- We look at phonotactic rules but also at letters not used in native Sakha words.
- The letters **ш, ж, я, з, е, ю and ё**
- Sakha does not allow clusters with three or more consonants(str, vdr, mgl, etc.)
- Sakha does not allow any voiced consonants at the end of words.
- Also, the number of two-consonant clusters is very limited
- A detailed description can be found in Ubryatova et al (1982).
- Combined, these rules allow us to classify many words as foreign, or at least unnaturalized, as there are many naturalized loanwords in Sakha, such as **ыскамыайка** (bench) [ıskamıayka]and **книгэ** [kinige] (book).

Data

- We base our calculations on a corpus collected by Leontiev (2015)
- 21 000 newspaper articles, 21 million words.
- Some Latin text, and some OCR-read text.
- OCR errors were corrected where possible.

Annotation

- Two native Russian speakers annotated loanwords based on our rule-based functions.
- Another annotator had knowledge of Sakha and annotated whether a given list of plural forms were plural or not.
- For loanwords, annotators agree on 80% of 300 words, with a kappa score of 0.63. Most of the disagreement comes from place names.
- Of the 300 suggested plural forms, 90% were judged to be correct. Most of the errors were due to missing Russian letters, and were fixed.

Conformity

- A vowel conforms to vowel harmony if it harmonizes with the previous vowel according to the rules described above.
- We expect loanwords to conform less, and native words to conform more, but note that it is perfectly possible for loanwords to conform both by chance and by naturalization.

Data counts

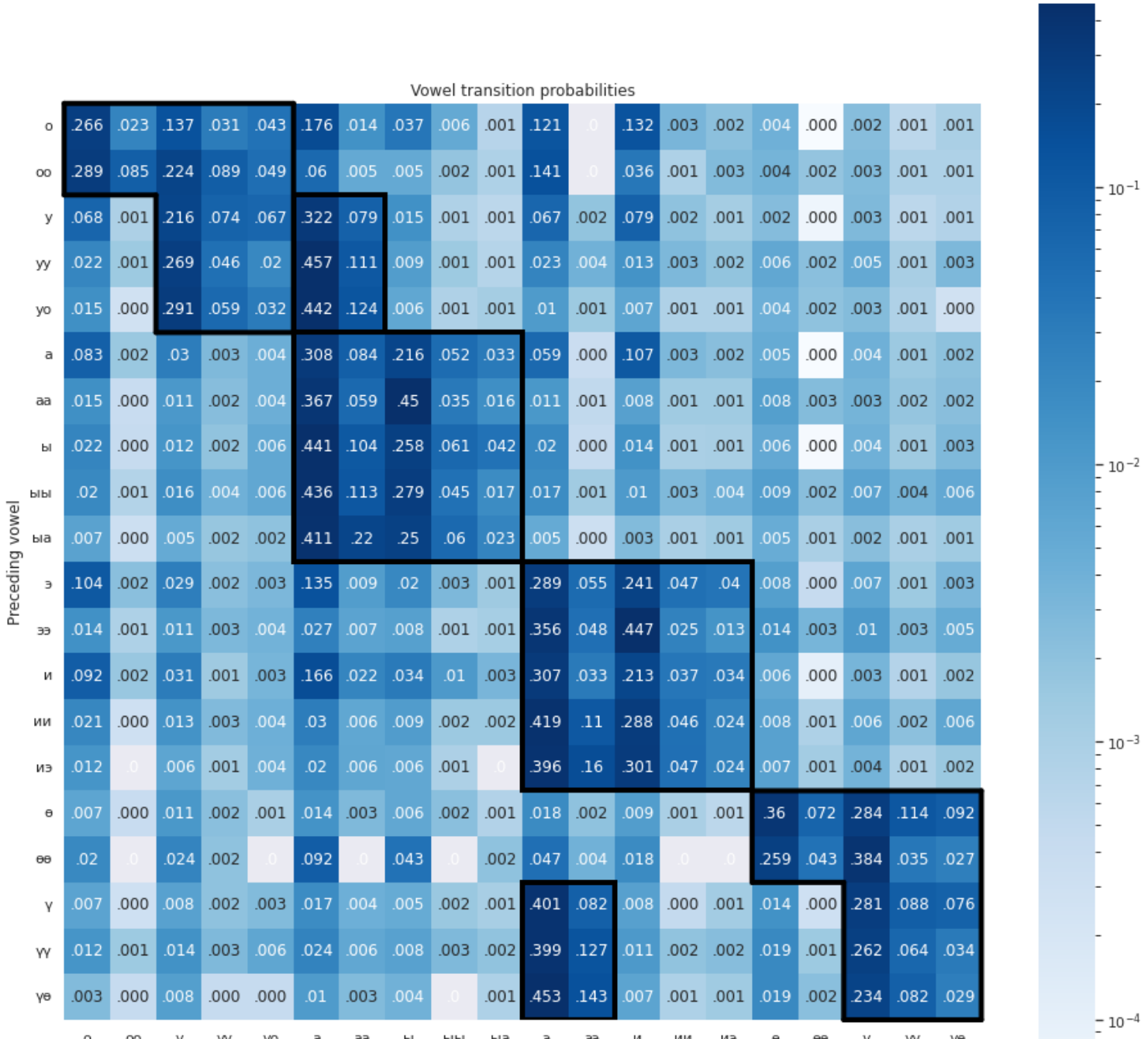
Data	Sum	Conf. #	Conf. %	Non-Conf #	Non-Conf %
All	453072	95849	21.16	357223	78.84
Foreign	106603	72208	67.74	34395	32.26
Native	346469	23641	6.82	322828	93.18
Hyph	34933	12085	34.59	22848	65.41
N-hyph	311536	11556	3.71	299980	96.29

Таблица: Counts for the total number of types counted, and for foreign, native and hyphenated and non-hyphenated words, and whether they conform or not.

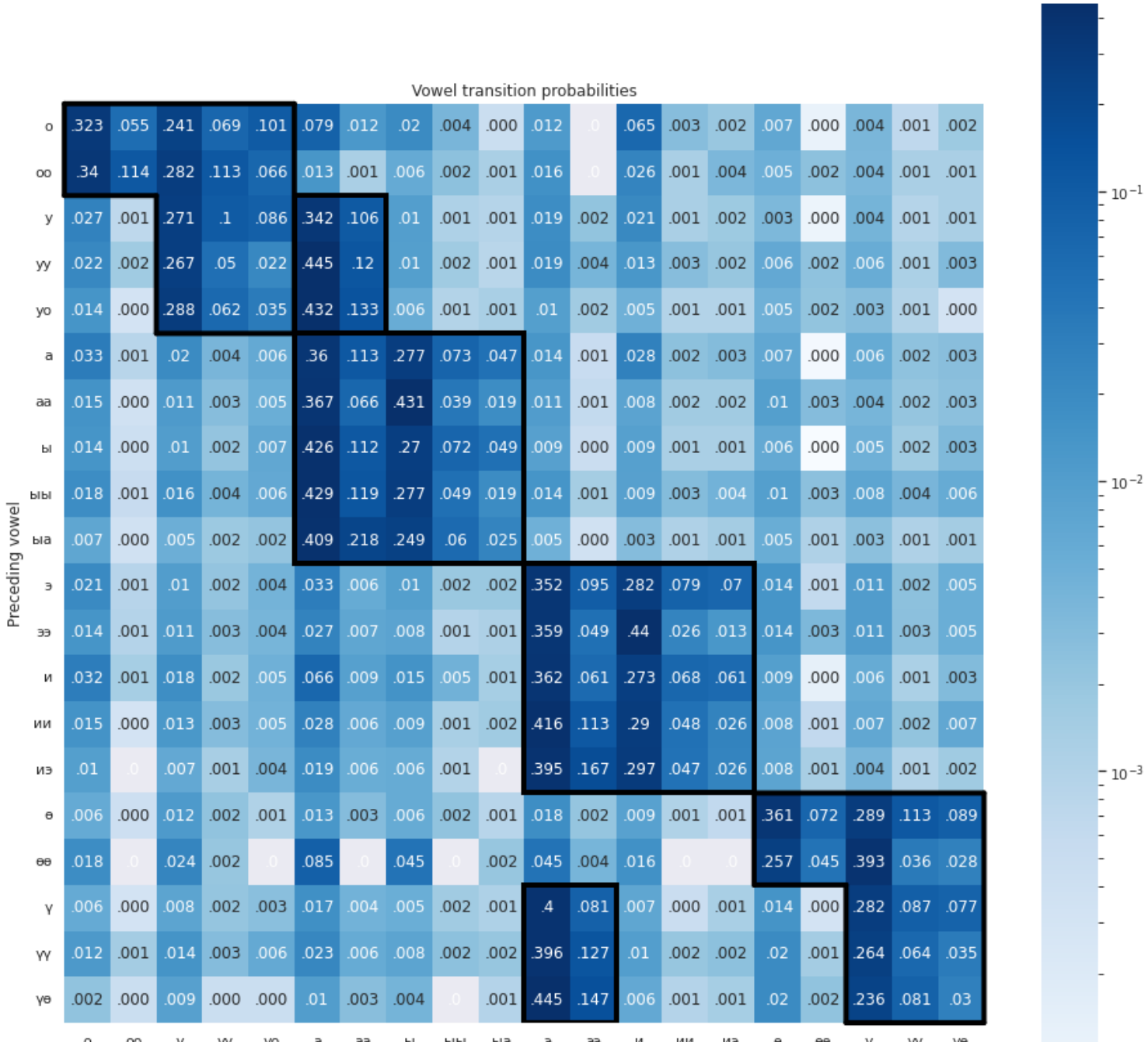
Transition Probabilities

- One goal was to use transition probabilities to say something more general, as well as to explore more specific phenomena.
- We reduce words to a list of vowels, and then create the probabilities based on vowel bigrams.
  - **остуол** [ostuol] 'table' becomes [**o, yo**] and
  - **уларыйыытыгар**, [ularıyııtıgar] 'to her/his change' becomes [**y, a, ы, ыы, ы, a**].

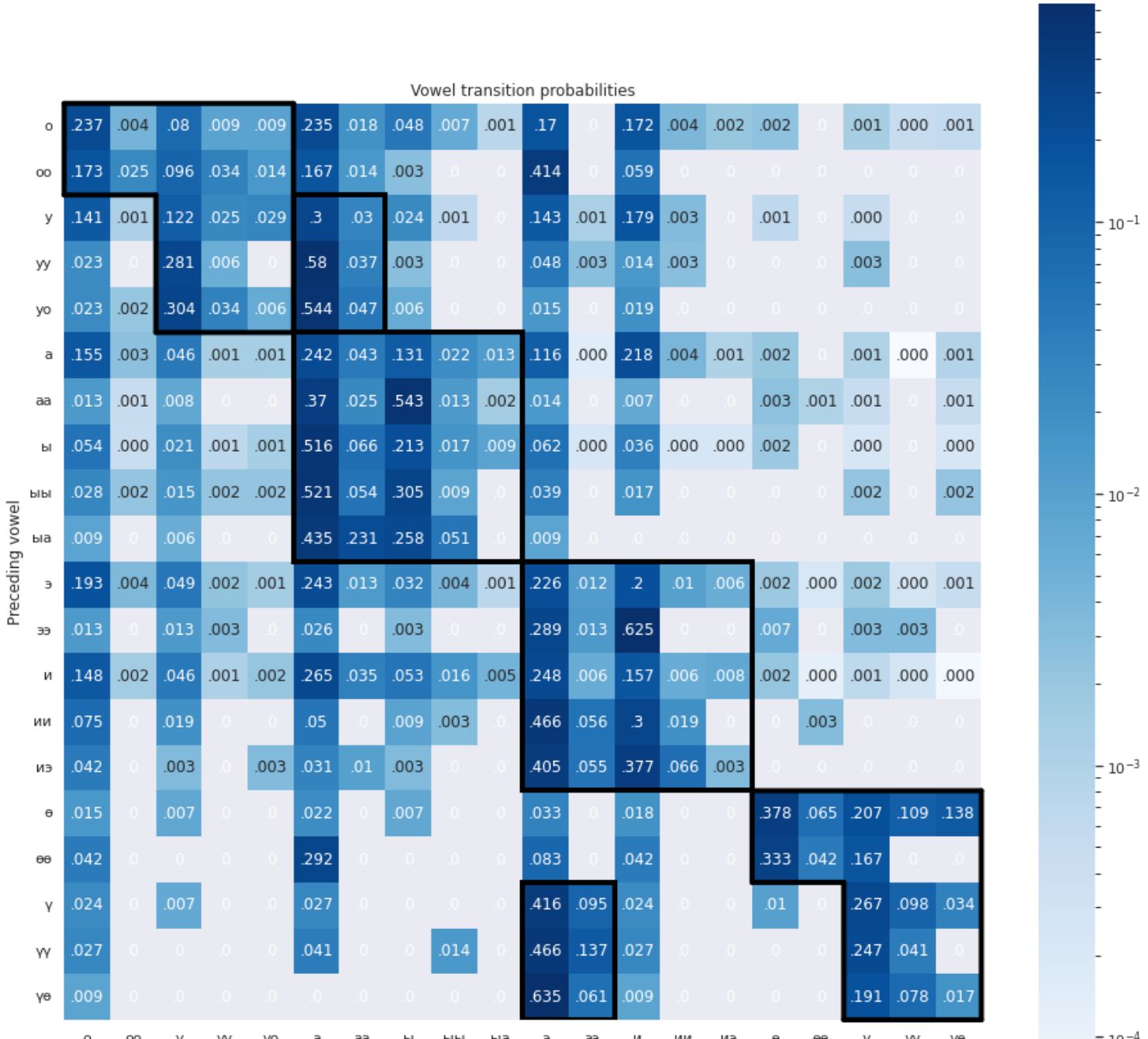
Transitions



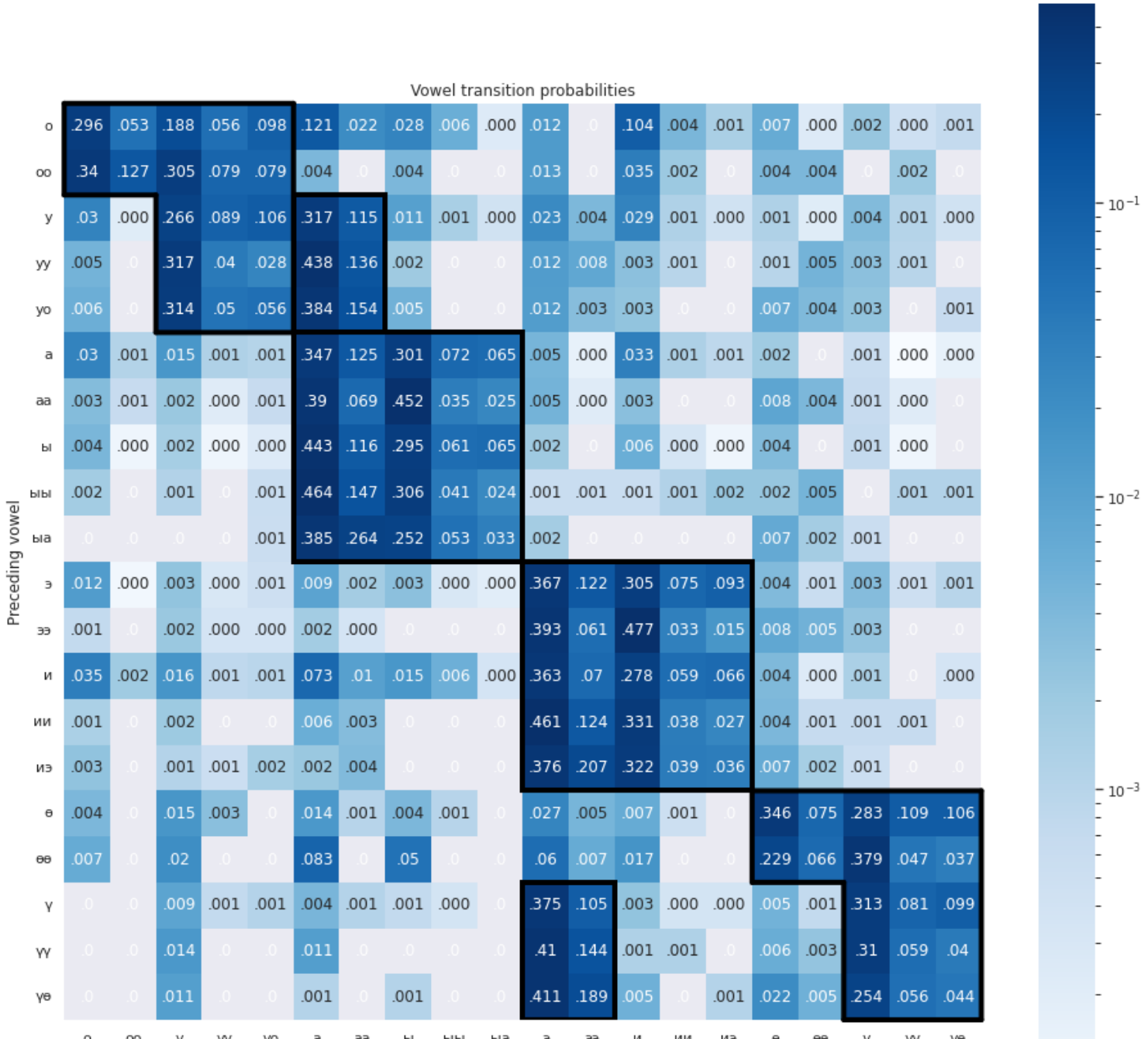
Transitions



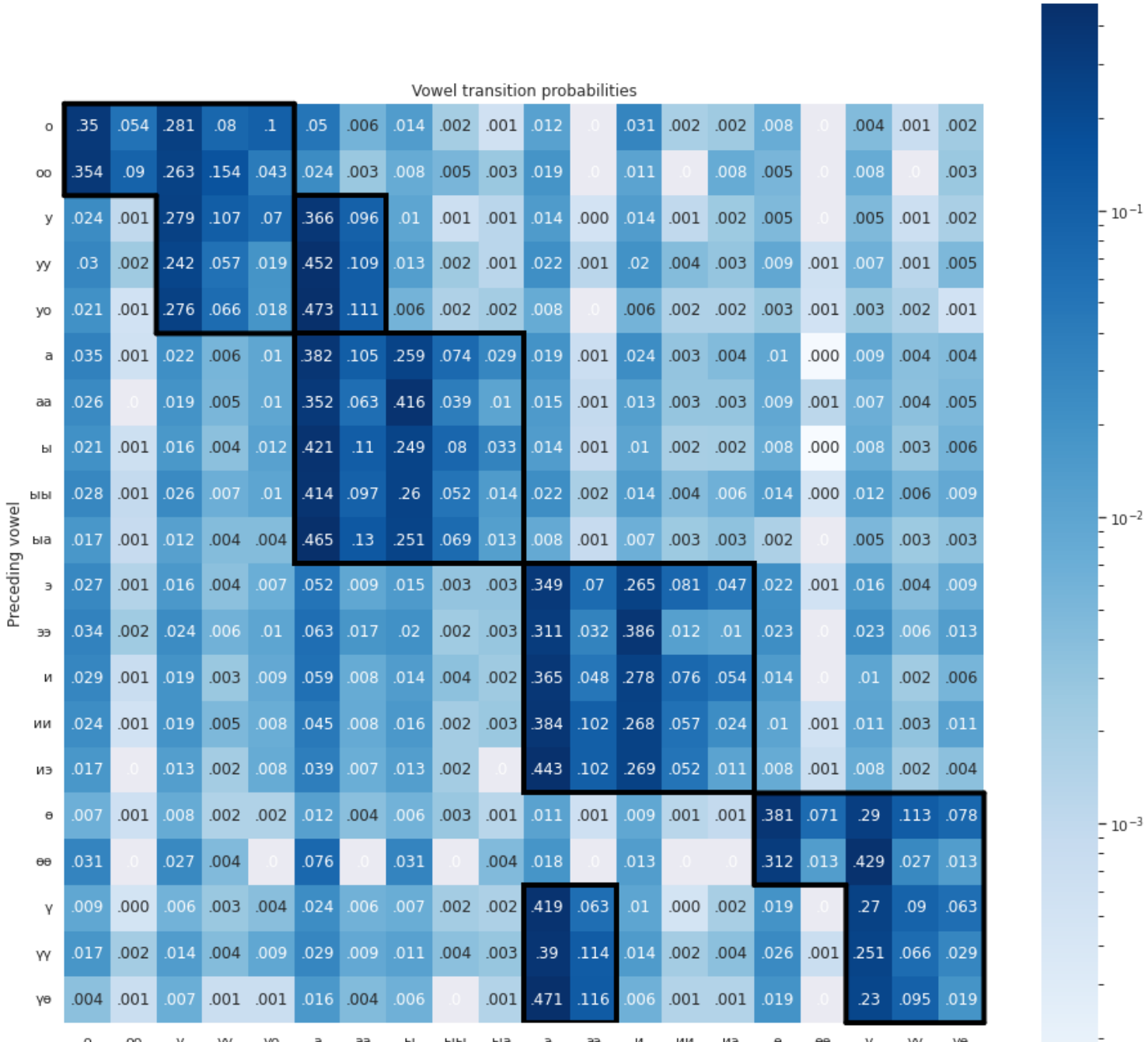
Transitions



Transitions



Transitions



Case study with -LAr

- Finally we applied this to the plural suffix -LAr.
- Due to the combination of consonant assimilation rules and vowel harmony, the suffix has 16 allomoprhs.

Cons. feature	A	E	O	Ö
Nasals	**нар** [nar]	**нэр** [ner]	**нор** [nor]	**нөр** [nör]
R and Y	**дар** [dar]	**дэр** [der]	**дор** [dor]	**дөр** [dör]
Unvoiced cons.	**тар** [tar]	**тэр** [ter]	**тор** [tor]	**төр** [tör]
Vowels and L	**лар** [lar]	**лэр** [ler]	**лор** [lor]	**лөр** [lör]

Таблица: Allomorphs of -LAr

-LAr analysis

- Of 30 280 words ending in -LAr, we found 26 602 to formally fit the plural criteria.
- 23 779 of these were vowel harmony conforming, while 2 823 were not.
- When looking at variation, we found that no words occur with more than 2 different forms.

Alternation	Count	Percent
а-э [a-e]	36	60.0%
э-а [e-a]	10	16.7%
о-а [o-a]	7	11.7%
а-о [a-o]	5	8.3%
о-э [o-e]	1	1.7%
а-ө [a-ö]	1	1.7%

Таблица: Vowel alternations in -LAr plural suffix use.

Maxtal!

We want to thank Karina Sheifer, Daniil Larionov and Elena Klyachko for their swift and detailed annotations.

Thank you for your attention!

Махтал!

Maxtal!