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

Petter Mæhlum
Department of Informatics
University of Oslo
pettemae@ifi.uio.no

Sardana Ivanova
Department of Computer Science
University of Helsinki
sardana.ivanova@helsinki.fi

Abstract

Obtaining information about loan words and irregular morphological patterns can be difficult for low-resource languages. Using Sakha as an example, we show that it is possible to exploit known phonemic regularities such as vowel harmony and consonant distributions to identify loan words and irregular patterns, which can be helpful in rule-based downstream tasks such as parsing and POS-tagging. We evaluate phonemically inspired methods for loanword detection, combined with bigram vowel transition probabilities to inspect irregularities in the morphology of loanwords. We show that both these techniques can be useful for the detection of such patterns. Finally, we inspect the plural suffix -ЛАр [-LAr] to observe some of the variation in morphology between native and foreign words.

1 Introduction

Sakha is a Turkic language, with around half a million native speakers (Eberhard et al., 2022), primarily residing in the Sakha Republic. The Sakha Republic is located in Northeast Asia, and is part of the Russian Far East. Sakha belongs to the Lena group of the Turkic language family. Like other Turkic languages, Sakha is agglutinative (Ubryatova et al., 1982). It has complex, four-way vowel harmony, and the Subject-Object-Verb word order, which we want to use to identify loan words. Its lexicon consists of Turkic words, borrowings from Mongolic and Tungusic languages, loanwords from Russian, and words of unclear (possibly Paleo-Asiatic) origin (Kharitonov, 1987). Note that in this project we do not draw any distinction between different types of borrowing or degree of naturalization. Where not specified, "loanword" should be understood to mean non-nativized loanword. Words should be understood as types, and we do not account for homography. Sakha words are transliterated using the Turkish ortography, expressed in brackets []. While the corpus cannot be re-distributed freely, functions and code details will be made available [1].

2 Earlier Research and Motivation

As the tools available to Sakha, as for many other low-resource languages are rule-based, spelling inconsistencies can affect down-stream tasks. An example is the errors in inflection of loanwords during analysis of errors made both by systems submitted for SIGMORPHON 2021 Shared Task on Morphological Reinflection (Pimentel et al., 2021) and forms generated by a morphological analyser created for Sakha (Ivanova et al., 2022) which was considered as the ground truth. The authors experienced that in some cases several native speakers could not agree on what should be the correct spelling. This is one of the indications of inconsistencies when it comes to vowel harmony in loanwords. For example both forms автомобилэ [avtomobile] and автомобила [avtomobila] were found for the original Russian автомобиль [avtomobil'] 'car'.

Other attempts at loanword identification for Turkic languages include (Mi et al., 2018) for Uyghur, using word embeddings. An example of using phonemic information is Mao and Hulden (2016), who map Japanese and English loan pairs to inspect their phonology.

3 Sakha Phonotactics and Vowel Harmony

In addition to looking at letters used only in Russian, we will exploit certain phonotactic regular-

[1] Available at https://github.com/Tyriflis/sakha_phonotactics

ities in Sakha, namely restrictions on consonant distributions, and vowel harmony.

Consonant Distributions While letters such as г [g] and д [d] and ь [soft sign; indicates palatalization] are present in Sakha words, г and д are never found word-finally, due to the fact that all voiced sounds are disallowed in this position(Ubryatova et al., 1982). While ь is found in the digraphs дь [c] and нь [ñ], its presence after any other consonant indicates a borrowing. Sakha is also typically much more restrictive with consonant combinations than Russian (Ubryatova et al., 1982) Consequently we can classify all illegal consonant bigrams as loans, as well as all words containing consonant trigrams, as no consonant trigrams are allowed in Sakha.

We will use Sakha consonantal features mainly to be able to create a rule-based classifier of words into *native-like* and *foreign*. Looking at the features outlined above, we can classify with reasonable certainty a word as foreign, but we cannot class a word as native with equal possibility, as the presence of a foreign consonant or a specific pattern can quite confidently mark a word as non-native, the opposite is not true. Many unnaturalized loanwords conform to Sakha spelling by chance. The consonant-related features we will be looking for are the following: 1) presence of foreign letters 2) illegal consonant positions 3) bigrams 4) trigrams.

Vowel Harmony Vowel harmony is a phenomenon where the use of a vowel is dependent on vowels in its context. Sakha exhibits a relatively strict, four-way vowel harmony. The vowels are shown in table 1. Sleptsov (2018) classifies this harmony into *velar-palatal*, corresponding to a back-front harmony, and *labial*, corresponding to rounding harmony. Together these two types of vowel harmony creates four different sets of vowels that harmonize. Vowel harmony is most pronounced in suffixes, but also governs which vowels can be found within a root. If a front vowel (eg. и [i] or э [e]) or a back vowel (eg.a [a] or о [o]) appears in a word, all following vowels must be of the same velar-palatal class. The case is the same for labial harmony (Sleptsov, 2018), with the exception of the two close, rounded vowels ү [ü] and у [u], along with the corresponding diphthongs үө [üö] and уо [uo], which all harmonize as if they were unrounded.

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

Table 1: Sakha vowels according to their features.

In example 1 we see that the low vowel ы [ɪ] requires that ы [ɪ] and а [a] are used in the suffixes as well. In example 2 we see that the round vowel in the root көр [kör] triggers the round vowel ü, here as the diphthong үө [üö] . As both these are high vowels, the final vowel (which can be э [e] or а [a]) is э [e]. In example 3 we see that although we would expect -тор- [-tor] - looking at rounding and height, we get -tar here, as u does not follow rounding for suffixes.

(1) *аһаа-ты-быт* *[ahaa-tɪ-bɪt]*
eat-PAST-1P.PL
'We ate'

(2) *көр-сүөх-хэ* *[kör-süöx-xe]*
see-REFL-COH
'Let's see each other'

(3) *улуус-тар-ыгар* *[uluus-tar-ɪgar]*
district–PL-DAT
'To their district'

3.1 Exceptions

Two classes of tokens do not follow vowel harmony. The first class is loanwords, the main focus of this paper. The second class is a collection of certain compounds that in writing are typically joined by a hyphen. While they do harmonize in terms of suffixes, as compunds the different roots in the compound do not necessarily harmonize with each other. Some examples are от-мас [ot-mas] 'grass-tree', i.e 'plants' and аһаа-сиэ [ahaa-sie] 'eat (intrans.)-eat (trans.)', i.e. 'eat'. The first example does not follow rounding harmony, while the second does not follow height-harmony. We see their endings harmonize in cases such as аһаа-сиэ [ahaa-sie] which is аһыыр-сиир [ahɪɪr-siir] 'eats'.

Of all words believed to be compounds, 47.6% had vowel harmony conforming in both roots. 50.4% had at least one non-conforming root. 1.5% had more than three hyphens and were excluded. 0.5% of the words had hyphens but with some tokenization error. These were also excluded.

3.2 Consonant Assimilation

Sakha also exhibits several cases of consonant assimilation rules, where for example voiced consonants have to match other voiced consonants, and some consonants are assimilated with others.

3.3 Suffix Conventions

The majority of Sakha suffixes follow both vowel harmony and consonant assimilation rules. We will follow convention and use capital letters to indicate phonemes that are affected by consonant or vowel harmony. Some examples include the plural suffix ЛАр [-LAr], the dative suffix ГА [GA], the interrogative suffix Ый [Iy] and the commitative suffix ДЫн [-DIn].

4 Data

We base our calculations on a corpus collected by Leontiev (2015). This corpus contains 21 000 newspaper articles, gathered from 2006 to 2015. The corpus contains a total of 21 million words. These texts also contain some OCR-read text, as well as Latin-letter text. Predictable OCR errors are corrected on reading, and Latin words are removed before further processing. The resulting list of lowered, normalized tokens counts 454 190 items.

4.1 Annotation

The result of foreign-word classification was doubly annotated by two native Russian speakers. The annotators agreed on 80% of the words that were supposed to be loanwords as being loanwords, with a kappa score of 0.63, indicating some disagreement, but indicating that our functions are reasonably successful in identifying foreign lexemes. Almost half of the disagreements seem to be on proper names. A third annotator annotated the validity of the plural extraction, showing that 90% of these were indeed plural forms.

4.2 Loanword Identification

A large portion of loanwords in Sakha come through Russian, and although both Sakha and Russian uses the Cyrillic alphabet, the Sakha alphabet contains certain extensions that can be used to class a word as non-Sakha. The letters ш, ж, я, з, е, ю and ё are not used in native Sakha words.

4.3 Vowel Transition Probabilities

We calculated the transition probabilities for each possible vowel pair in Sakha. We consider all tokens, first of all because the derivational and inflectional endings are important to us, as they are one of the clearest places where vowel harmony comes into play. First all words are reduced to a vowel representation. This was done using a function that took a token as its input, and then identifying all vowel-marking letters and adding them to a list. Long vowels are treated separately. For example, остуол [ostuol] ‘table’ becomes [o, yo] and уларыйыытыгар, [ularıyıtıgar] ‘to her/his change’ becomes [у,а,ы,ыы,ы,а]. We then created a bigram representation of each vowel set, and use these to accumulate the frequencies for each vowel given the previous vowel. These frequencies were then converted to transition probabilities. We calculated transition probabilities for the entire corpus and for four sub-groupings: foreign words, native words, hyphenated native words and non-hyphenated native words. The Russian-specific vowels я [ya] е [ye] ю [yu] and ё [yo] are treated as their corresponding vowels in Sakha, respectively: а [a], э [e], у [u] and о [o].

4.4 Degree of Conforming to Vowel Harmony

Using the above-mentioned methods, we split the data into three main groups: native words, foreign words and a combined group. We also looked at hyphenated and non-hyphenated words, which are subgroups of native words. Their statistics are reported in Table 2. We note that the percentages of conforming vs. non-conforming types is striking: A significantly higher portion of the expected native words conform, at 93.18% , while only 32.26% of foreign words conform. We also see that if we remove hyphenated words from the native set, we reach a conform percentage of 96.29%.

5 Analysis

5.1 Transition Probabilities

For all sets but the foreign one, there is a clear connection between transition probabilities and the expected harmonies. In Figure 1 we see that the probabilities are markably larger on the diagonal than the remaining areas, except for the foreign words. The reason why the harmonies between RB and UB and RF and UF are consistently a bit darker is due to the beforementioned special

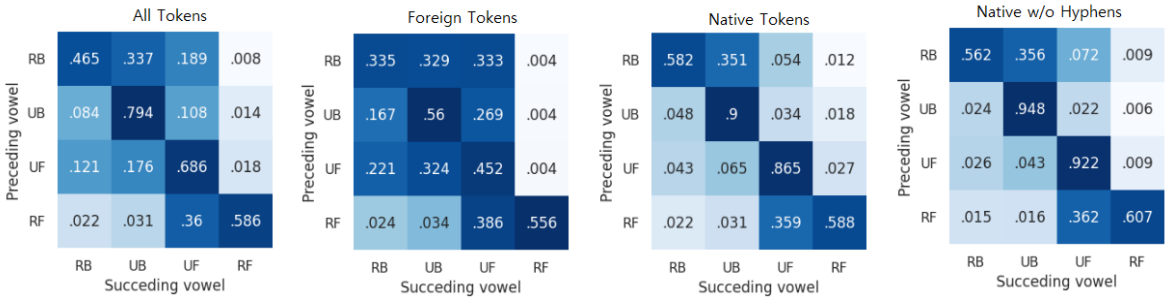

Figure 1: Transition tables for four different sets. R= rounded, U=unrounded, B=back, F=front.

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

Table 2: The total number of types, and whether they conform to vowel harmony or not.

cases of high, rounded vowels. If we inspect the 20 most common non-conforming transitions, we see that apart from the three cases өө-a [öö-a], өө-э [öö-e], and өө-ы [öö-ɪ], all transitions contain an overwhelming number of foreign-classified words. We also see that when removing hyphenated words from the non-foreign words, the non-conforming noise is largely reduced, indicating that these words, if not dealt with, contribute to vowel harmony noise. Closer inspection shows that almost half of the compound words conform to vowel harmony.

5.2 Suffix Analysis and Variation

In order to inspect vowel alternations in practice, we chose to focus on the plural suffix -LAr, as it is a quite frequent suffix, and it is a bit long, making it easier to identify, compared to single-letter or two-letter suffixes. With consonant and vowel alternations, there is a total of 16 allomorphs for -LAr. We first inspected all words in the corpus ending in this sequence, before ruling out words ending in letters that would not fit the first letter in the ending. We accounted for the apparent de-voicing of Russian voiced letters. Of a total of 30 280 words ending in the selected sequences, 26 602 were judged to be plural forms. Note that plural suffixes that do not come last were not counted. 23 779 of these were vowel harmony compliant in terms of the last vowel of the word and the vowel of the suffix, while 2 823 were not. Then, we inspected the variance. We looked at any word stem that appeared with more than one vowel in the set. The highest number of varying vowels were 2, and only 60 words were found with this alternation. 44 of these were foreign words. We see that the majority of confusion is between a-ə [a-e] (both directions) with 76.7% of cases, and with o-a [o-a] (both directions) also being common, with 20% of all cases.

6 Conclusion and Future Work

We have seen that phonotactic rules can be useful for loanword identification in Sakha, and that by using this information, we can gain insight on the morphological treatment of these words. We have shown that when vowel harmony is strict, it is also a good indicator of loanwords, and we have showed how this can be used to illustrate alternations in morphology. We expect the results here to be relevant for any language with vowel harmony or similar phenomena. We would also like to stress that these rule-based methods are simple and efficient, and allow large amounts of lexicographic work and preprocessing be done on languages where preprocessing tools or lexical lists are unavailable, but some raw data exists. However, we only inspected one of many Sakha suffixes, and believe that further investigations can shed light on the actual state of vowel-oriented morphological variation in Sakha. We also note that although the rule-based function work well, good lemmatization techniques would be able to remove some ambiguity in our analyses.

Acknowledgments

We would like to thank Elena Klyachko, Daniil Larionov and Karina Sheifer for their swift and detailed annotation efforts.

References

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2022. *Ethnologue: Languages of the World*, online, twenty-fifth edition. SIL International, Dallas, Texas.

Sardana Ivanova, Jonathan N. Washington, and Francis M. Tyers. 2022. A free/open-source morphological analyser and generator for sakha. In *Proceedings of the 13th Language Resources and Evaluation Conference*.

Leonid Kharitonov. 1987. *(Yakut language tutorial)*. Yakutsk Publishing.

Nyurgun Leontiev. 2015. Turkic languages (turklang 2015) the newspaper corpus of the yakut language.

Lingshuang Jack Mao and Mans Hulden. 2016. How regular is japanese loanword adaptation? a computational study. In *International Conference on Computational Linguistics*.

Chenggang Mi, Yating Yang, Lei Wang, Xi Zhou, and Tonghai Jiang. 2018. Toward better loanword identification in Uyghur using cross-lingual word embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3027–3037, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Petr Sleptsov. 2018. *Саха тылын быһаарыылаах улахан тылдьыта: Большой толковый словарь якутского языка. [Large explanatory dictionary of the Yakut language: in 15 volumes]*. Novosibirsk, Nauka.

E. I. Ubryatova, E. I Korkina, L. N. Haritonov, and H.E Petrov, editors. 1982. *Грамматика современного якутского литературного языка: Фонетика и морфология [E. I. Ubryatova et al. Grammar of the modern Yakut literary language: Phonetics and morphology]*. Mosvka, Nauka.

6.1 Appendix

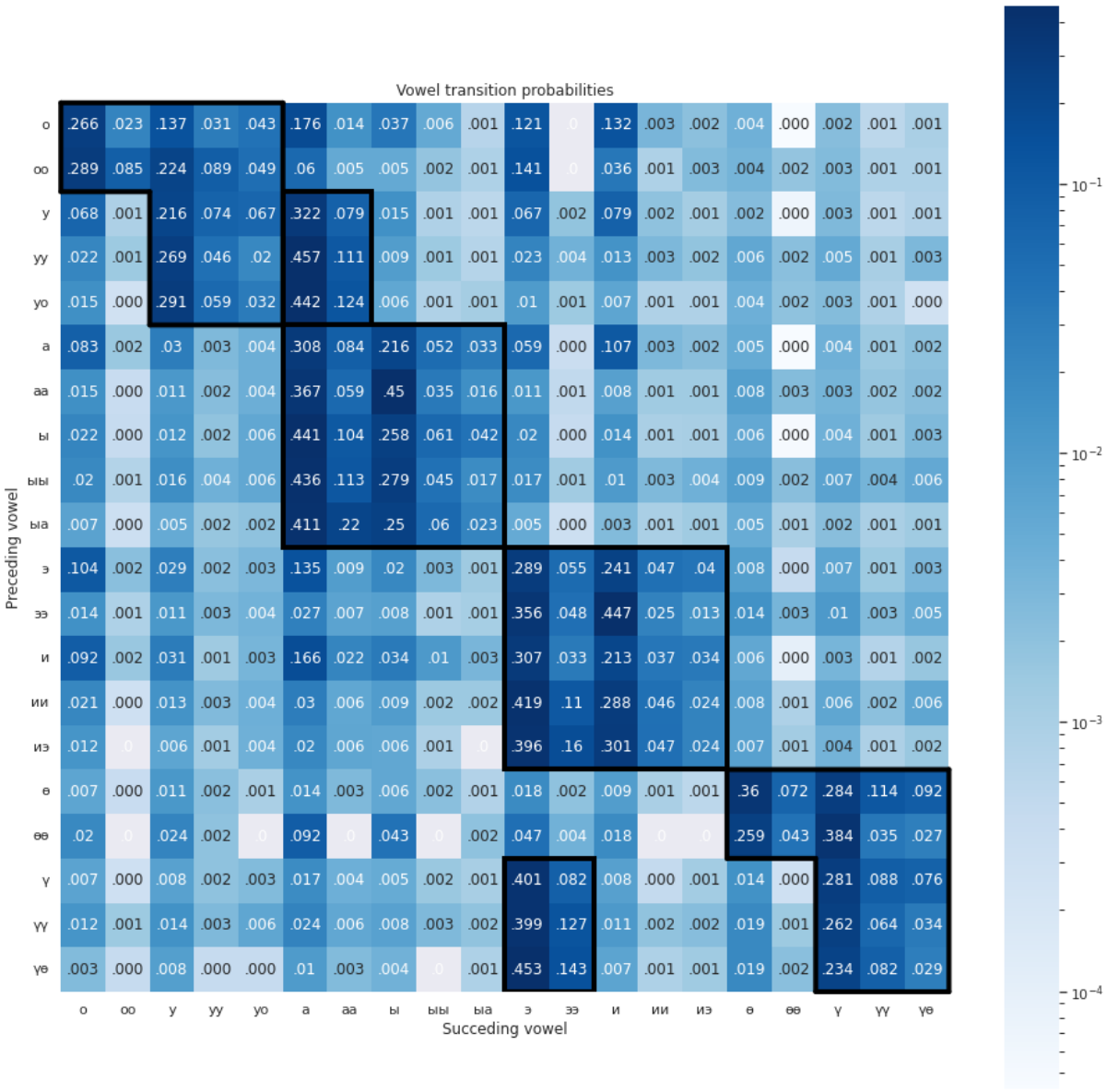

Figure 2: Transitions for all tokens. Note the irregular areas outside

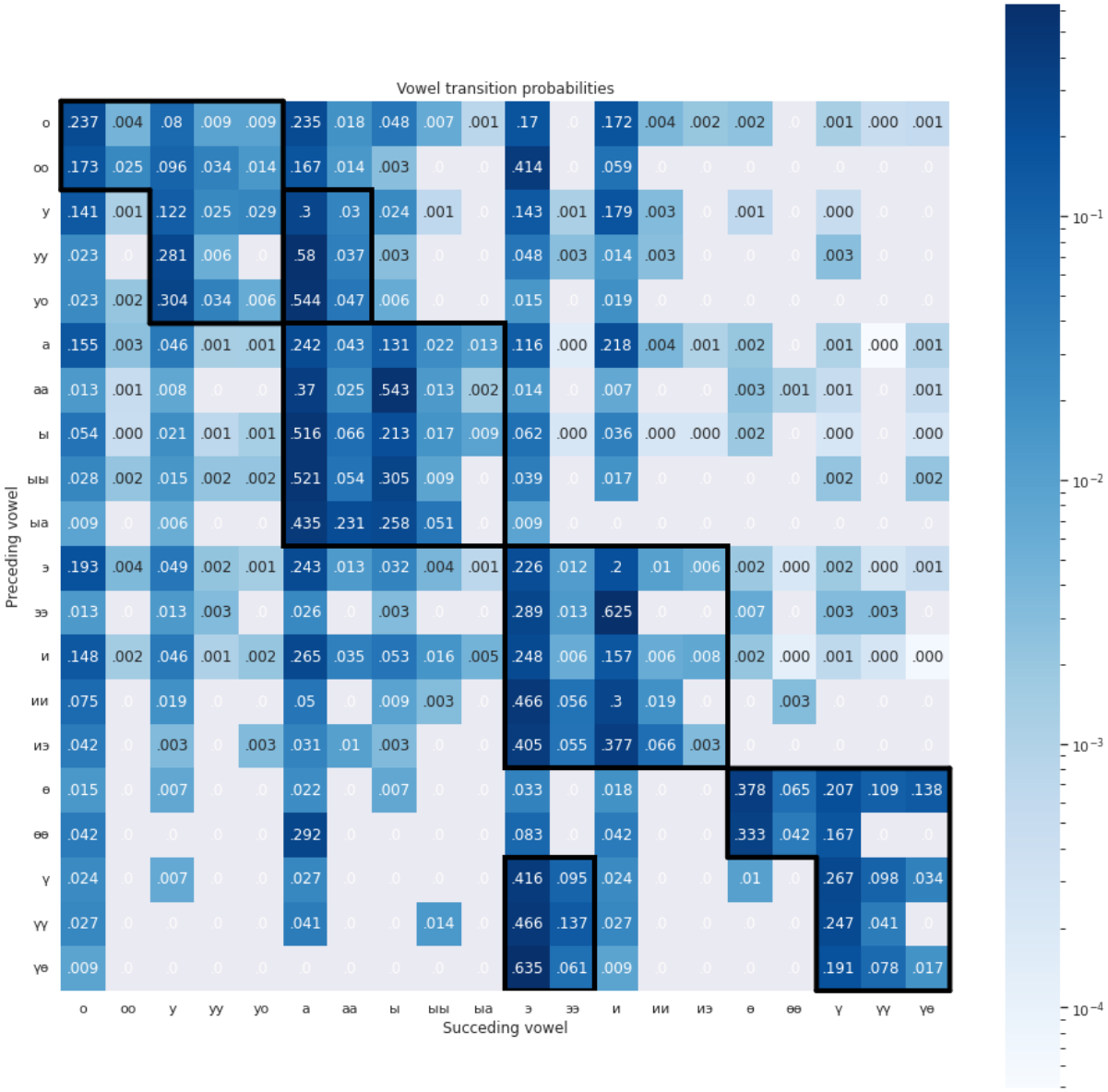

Figure 3: Transitions for words labeled as foreign. Note how there are very little data on the native Sakha vowels not found in Russian.

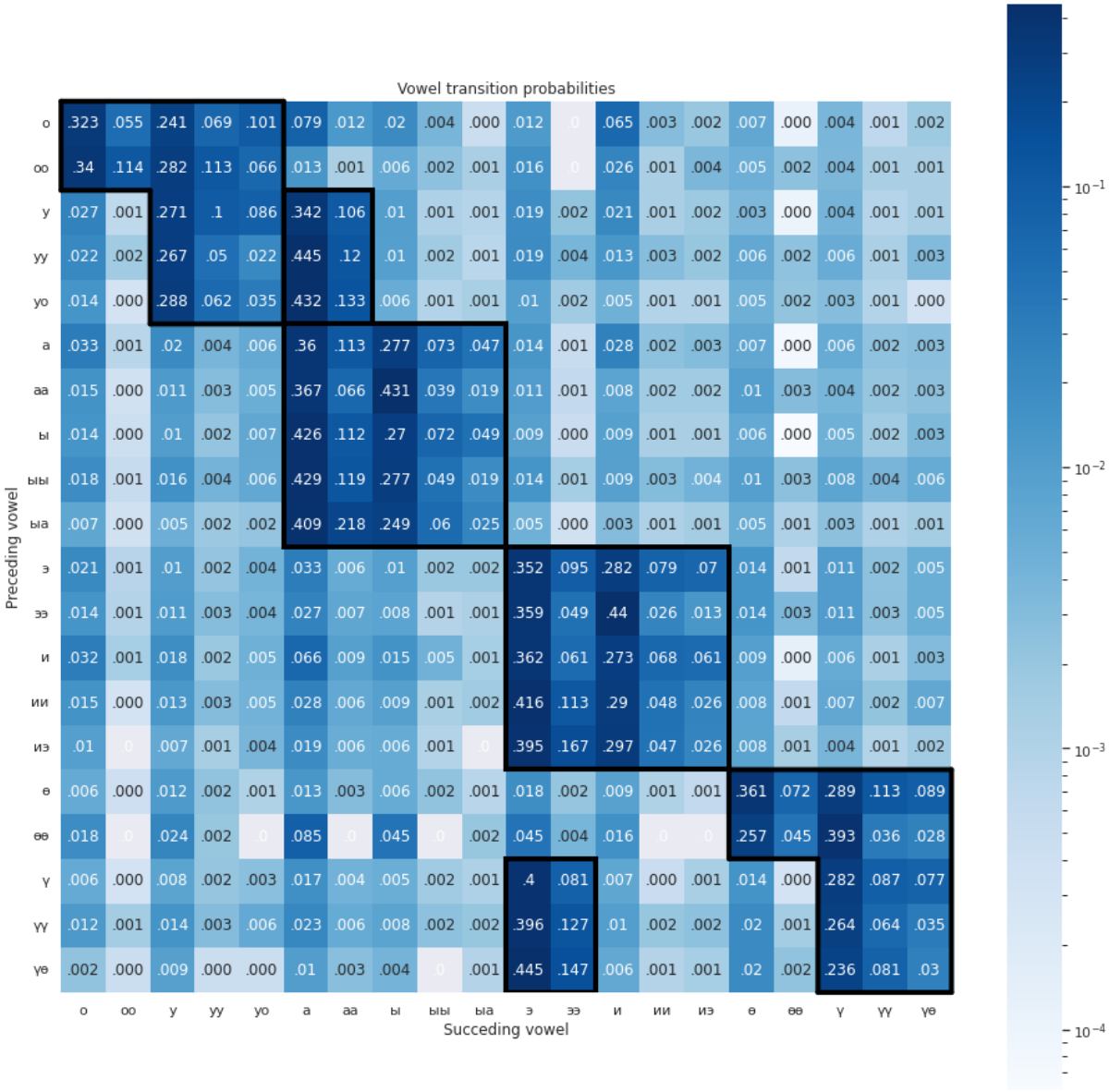

Figure 4: Native transitions. Note the clear difference between conforming and non-conforming transitions.

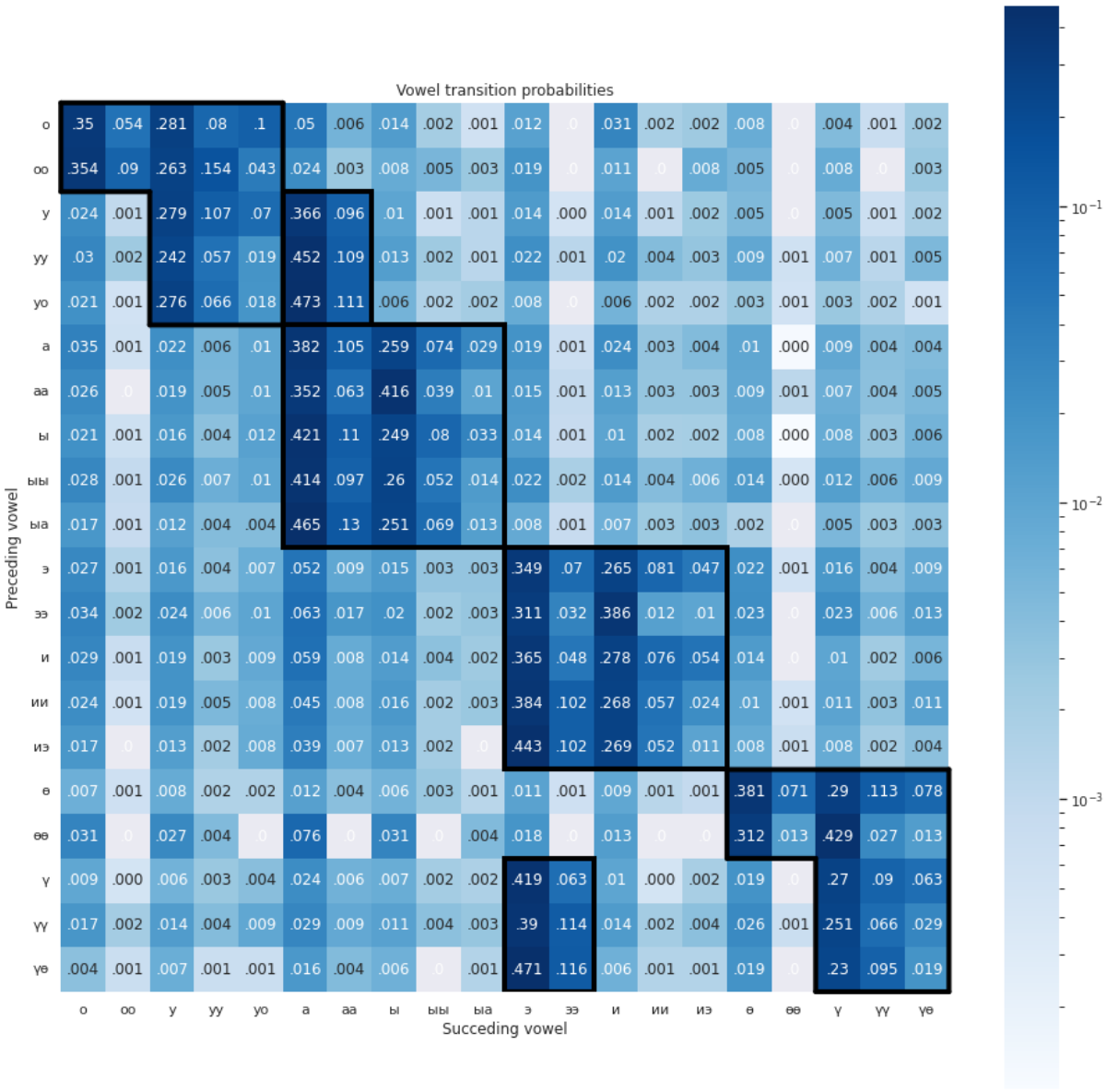

Figure 5: All words that have hyphens in them. Similar to native, but a bit more variation.

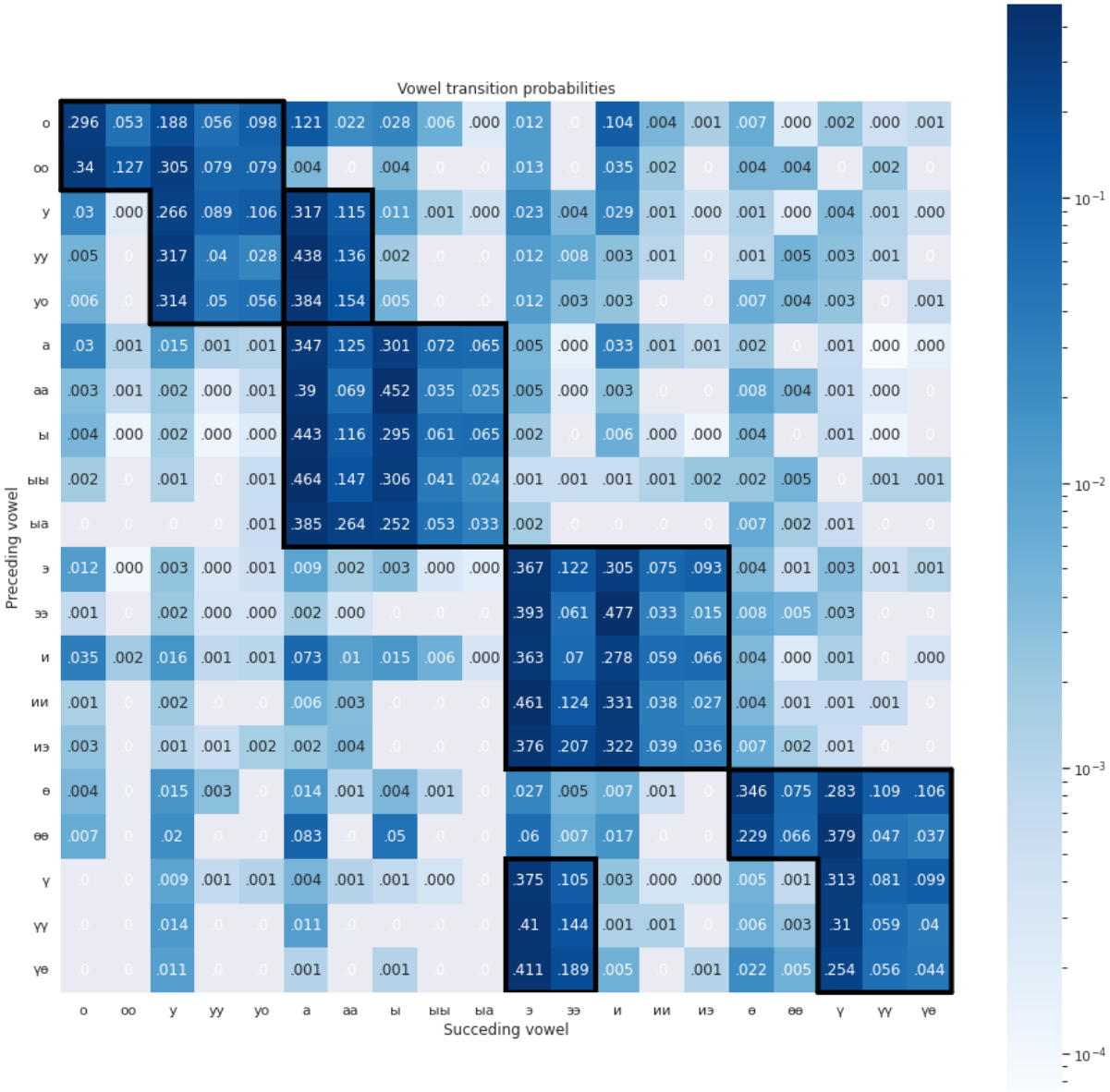

Figure 6: Native words excluding all words with hyphens. Notice how much the likelihood of “illegal" transitions is reduced.