## Phonotactics for Loan Word Detection

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

- 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
- 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.
- 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
- Turkic language with 450000 speakers.
- Spoken primarily in Siberia.
- Agglutinating and verb final (SOV).
- Suffixes largely follow vowel harmony and consonant assimilation rules
- 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 Oull $->$ a
- eiüo->e
- I-harmony (backness + rounding)
- ei->i
- $\mathrm{a}_{1->1}$
- ü ö -> $\ddot{u}$
- $\mathrm{OH}->\mathrm{u}$


## Sakha Vowel harmony

- Sakha vowels are similar to Turkish, but in addition, Sakha has four diphthongs (ie,1a, 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 $\ddot{u}$ 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 $[\mathrm{i}]$ | э $[\mathrm{e}]$ | ы $[1]$ | а $[\mathrm{a}]$ |  |
| Rounded | Y [ü] | ө $[\ddot{\mathrm{o}}]$ | у $[\mathrm{u}]$ | о $[\mathrm{o}]$ |  |

Таблица: Sakha vowels according to their features.

## Sakha Vowel harmony

улуус-тар-ыгар [uluus-tar-1gar]
district-PL-DAT
'To their district'

көр-сүөх-хэ [kör-süöx-xe]
see-REFL-COH
'Let's see each other'
ahaa-ты-быт [ahaa-tı-bıt]
eat-PAST-1P.PL
'We ate'

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 аһыыр-сиир [ahır-siir] 'eats' in the present tense..

- 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 кинигэ [kinige] (book).


## Data

- We base our calculations on a corpus collected by Leontiev ( 2015)
- 21000 newspaper articles, 21 million words.
- Some Latin text, and some OCR-read text.
- OCR errors were corrected where possible.
- 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.
- 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

|  |  | Conf. |  | Non-Conf |  |
| :--- | ---: | ---: | ---: | ---: | ---: |
| Data | Sum | $\#$ | $\%$ | $\#$ | $\%$ |
| 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.

- 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 [0, yo] and
- уларыйыытыгар, [ularıyıtıgar] 'to her/his change' becomes [y, а, ы, Ыы, Ы, а].


## Transitions

Vowel transition probabilities


## Transitions

Vowel transition probabilities


## Transitions

Vowel transition probabilities


## Transitions

Vowel transition probabilities

| 0 | 296 | . 053 | 188 | . 056 | . 098 | .121 | . 022 | . 028 | . 006 | . 000 | . 012 |  | 104 | . 004 | . 001 | 007 | . 000 | . 002 | . 000 | . 001 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\infty$ | 34 | . 127 | 305 | . 079 | . 079 | . 004 |  | . 004 |  |  | 013 |  | . 035 | . 002 |  | . 004 | . 004 |  | . 002 |  |
| y | 03 | . 000 | 266 | . 089 | 106 | .317 | 115 | . 011 | . 001 | . 000 | . 023 | . 004 | . 029 | . 001 | . 000 | . 001 | . 000 | . 004 | . 001 | . 000 |
| yy | . 005 |  | 317 | . 04 | . 028 | . 438 | 136 | . 002 |  |  | . 012 | . 008 | . 003 | . 001 |  | . 001 | . 005 | . 003 | . 001 |  |
| yo | 006 |  | . 314 | 05 | . 056 | . 384 | 154 | . 005 |  |  | . 012 | . 003 | . 003 |  |  | . 007 | . 004 | . 003 |  | . 001 |
| a | . 03 | . 001 | . 015 | . 001 | . 001 | . 347 | .125 | . 301 | . 072 | . 065 | . 005 | . 000 | . 033 | . 001 | . 001 | . 002 |  | . 001 | . 000 | . 000 |
| aa | 003 | . 001 | . 002 | . 000 | . 001 | . 39 | . 069 | . 452 | . 035 | . 025 | . 005 | . 000 | . 003 |  |  | . 008 | . 004 | . 001 | . 000 |  |
| ы | . 004 | . 000 | . 002 | . 000 | . 000 | 443 | 116 | 295 | . 061 | 065 | . 002 |  | 006 | . 000 | . 000 | . 004 |  | . 001 | . 000 |  |
| ыы | . 002 |  | . 001 |  | . 001 | . 464 | . 147 | . 306 | . 041 | . 024 | . 001 | . 001 | . 001 | . 001 | . 002 | . 002 | . 005 |  | . 001 | . 001 |
| ыа |  |  |  |  | .001 | . 385 | 264 | . 252 | . 053 | . 033 | . 002 |  |  |  |  | 007 | . 002 | . 001 |  |  |
| 3 | 012 | . 000 | . 003 | . 000 | . 001 | . 009 | . 002 | . 003 | . 000 | . 000 | 367 | . 122 | 305 | . 075 | . 093 | . 004 | . 001 | . 003 | . 001 | . 001 |
| 33 | 001 |  | . 002 | . 000 | . 000 | . 002 | . 000 |  |  |  | 393 | . 061 | 477 | . 033 | . 015 | . 008 | . 005 | . 003 |  |  |
| и | . 035 | . 002 | . 016 | . 001 | . 001 | . 073 | . 01 | 015 | . 006 | . 000 | 363 | 07 | 278 | . 059 | . 066 | . 004 | . 000 | . 001 |  | . 000 |
| ии | . 001 |  | . 002 |  |  | . 006 | . 003 |  |  |  | 461 | . 124 | 331 | . 038 | . 027 | . 004 | . 001 | . 001 | . 001 |  |
| иэ | . 003 |  | . 001 | . 001 | . 002 | . 002 | . 004 |  |  |  | 376 | 207 | . 322 | . 039 | . 036 | . 007 | . 002 | . 001 |  |  |
| $\theta$ | . 004 |  | 015 | . 003 |  | . 014 | . 001 | . 004 | . 001 |  | . 027 | . 005 | 007 | . 001 |  | . 346 | . 075 | 283 | 109 | 106 |
| $\theta$ | 007 |  | 02 |  |  | . 083 |  | . 05 |  |  | 06 | . 007 | . 017 |  |  | 229 | . 066 | . 379 | . 047 | . 037 |
| Y |  |  | 009 | . 001 | . 001 | . 004 | . 001 | . 001 | . 000 |  | 375 | . 105 | . 003 | . 000 | . 000 | 005 | . 001 | 313 | . 081 | . 099 |
| W |  |  | . 014 |  |  | . 011 |  |  |  |  | 41 | . 144 | . 001 | . 001 |  | . 006 | . 003 | 31 | . 059 | . 04 |
| ye |  |  | . 011 |  |  | . 001 |  | . 001 |  |  | 411 | . 189 | 005 |  | . 001 | . 022 | 005 | 254 | . 056 | . 044 |

## Transitions

Vowel transition probabilities

| 0 | . 35 | . 054 | 281 | . 08 | 1 | 05 | . 006 | . 014 | . 002 | . 001 | . 012 |  | . 031 | . 002 | . 002 | . 008 |  | . 004 | . 001 | . 002 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\infty$ | 354 | . 09 | 263 | . 154 | . 043 | . 024 | . 003 | . 008 | . 005 | . 003 | . 019 |  | . 011 |  | 008 | . 005 |  | . 008 |  | . 003 |
| y | . 024 | . 001 | 279 | . 107 | 07 | 366 | 096 | 01 | . 001 | . 001 | . 014 | . 000 | 014 | . 001 | . 002 | . 005 |  | 005 | . 001 | . 002 |
| yy | 03 | . 002 | 242 | . 057 | . 019 | . 452 | 109 | . 013 | . 002 | . 001 | . 022 | . 001 | . 02 | . 004 | . 003 | 009 | . 001 | .007 | . 001 | . 005 |
| yo | 021 | . 001 | 276 | . 066 | 018 | . 473 | 111 | . 006 | . 002 | . 002 | . 008 |  | . 006 | . 002 | . 002 | . 003 | . 001 | . 003 | . 002 | . 001 |
| a | . 035 | . 001 | . 022 | . 006 | . 01 | . 382 | . 105 | .259 | . 074 | . 029 | . 019 | . 001 | . 024 | . 003 | . 004 | 01 | . 000 | . 009 | . 004 | . 004 |
| aa | 026 |  | . 019 | . 005 | 01 | . 352 | . 063 | . 416 | . 039 | 01 | . 015 | . 001 | . 013 | . 003 | . 003 | . 009 | . 001 | . 007 | . 004 | . 005 |
| ы | . 021 | . 001 | . 016 | . 004 | 012 | . 421 | 11 | 249 | . 08 | . 033 | . 014 | . 001 | 01 | . 002 | . 002 | . 008 | . 000 | . 008 | . 003 | . 006 |
| ыы | . 028 | . 001 | . 026 | . 007 | 01 | . 414 | . 097 | 26 | . 052 | . 014 | . 022 | . 002 | 014 | . 004 | . 006 | 014 | . 000 | . 012 | . 006 | . 009 |
| ыа | 017 | . 001 | 012 | . 004 | . 004 | . 465 | 13 | . 251 | . 069 | 013 | . 008 | . 001 | 007 | . 003 | . 003 | . 002 |  | . 005 | . 003 | . 003 |
| 3 | . 027 | . 001 | . 016 | . 004 | . 007 | . 052 | . 009 | . 015 | . 003 | . 003 | 349 | . 07 | 265 | . 081 | . 047 | . 022 | . 001 | 016 | . 004 | . 009 |
| 33 | 034 | . 002 | . 024 | . 006 | 01 | . 063 | . 017 | . 02 | . 002 | . 003 | 311 | . 032 | 386 | 012 | 01 | 023 |  | . 023 | . 006 | . 013 |
| и | . 029 | . 001 | . 019 | . 003 | .009 | . 059 | . 008 | . 014 | . 004 | . 002 | 365 | . 048 | 278 | . 076 | . 054 | . 014 |  | 01 | . 002 | . 006 |
| ии | . 024 | . 001 | . 019 | . 005 | .008 | 045 | 008 | . 016 | . 002 | . 003 | 384 | . 102 | 268 | . 057 | 024 | 01 | . 001 | 011 | . 003 | . 011 |
| иэ | . 017 |  | . 013 | . 002 | . 008 | . 039 | . 007 | . 013 | . 002 |  | 443 | .102 | 269 | . 052 | . 011 | . 008 | . 001 | . 008 | . 002 | . 004 |
| $\theta$ | 007 | . 001 | . 008 | . 002 | . 002 | 012 | . 004 | . 006 | . 003 | . 001 | 011 | . 001 | 009 | . 001 | . 001 | . 381 | . 071 | 29 | 113 | . 078 |
| $\theta$ | . 031 |  | . 027 | . 004 |  | 076 |  | . 031 |  | . 004 | . 018 |  | 013 |  |  | 312 | . 013 | 429 | . 027 | . 013 |
| Y | 0.09 | . 000 | 006 | . 003 | . 004 | 024 | . 006 | . 007 | . 002 | . 002 | 419 | . 063 | 01 | . 000 | . 002 | . 019 |  | 27 | . 09 | . 063 |
| W | 017 | . 002 | . 014 | . 004 | .009 | . 029 | . 009 | . 011 | . 004 | . 003 | 39 | . 114 | . 014 | . 002 | . 004 | . 026 | . 001 | 251 | . 066 | . 029 |
| ye | . 004 | . 001 | . 007 | . 001 | . 001 | 016 | . 004 | .006 |  | . 001 | 471 | 116 | 006 | . 001 | . 001 | . 019 |  | 23 | . 095 | . 019 |

- 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 30280 words ending in -LAr, we found 26602 to formally fit the plural criteria.
- 23779 of these were vowel harmony conforming, while 2823 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-a [e-a] | 10 | $16.7 \%$ |
| о-a [o-a] | 7 | $11.7 \%$ |
| a-o [a-o] | 5 | $8.3 \%$ |
| о-э [o-e] | 1 | $1.7 \%$ |
| а-ө [a-̈] | 1 | $1.7 \%$ |

Таблица: Vowel alternations in -LAr plural suffix use.

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

Thank you for your attention! Махтал!
Maxtal!

