Comparing phonological systems and syllable structure of Botlikh and Zilo Andi: a data-driven analysis

George Moroz
Linguistic Convergence Laboratory, NRU HSE, Moscow, Russia

24 October 2019

“Caucasian Languages: Typology and Diachrony” in honor of M. E. Alekseev,
Institute of Linguistics RAS, Moscow

Presentation is available here: https://tinyurl.com/y5xoks3l
Phonological description: data-driven analysis

<table>
<thead>
<tr>
<th>Traditional analysis</th>
<th>Data-driven analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Done by trained linguist</td>
<td>Evaluated by trained linguist</td>
</tr>
<tr>
<td>Can be done from scratch</td>
<td>Previous description needed (or at least prior expectations)</td>
</tr>
<tr>
<td>Doesn't care about amount of data</td>
<td>Care more about amount of data</td>
</tr>
<tr>
<td>Less reproducible</td>
<td>More reproducible</td>
</tr>
<tr>
<td>Can not be automated</td>
<td>Can be automated</td>
</tr>
</tbody>
</table>

Data-driven approach to phonological description and syllable structure analysis:

• was proposed in [Moroz 2018]
• was applied to syllable structure in [Moroz 2019] to Adyghe data
• was applied to syllable structure in [Romanova 2019] to Russian and Macedonian data

I will present an application of this method to Botlikh and Zilo Andi data

2
## Phonological description: data-driven analysis

<table>
<thead>
<tr>
<th>Traditional analysis</th>
<th>Data-driven analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Done by trained linguist</td>
<td>Evaluated by trained linguist</td>
</tr>
<tr>
<td>2. Can be done from scratch</td>
<td>Previous description needed</td>
</tr>
<tr>
<td></td>
<td>(or at least prior expectations)</td>
</tr>
<tr>
<td>3. Doesn’t care about amount of data</td>
<td>Care more about amount of data</td>
</tr>
<tr>
<td>4. Less reproducible</td>
<td>More reproducible</td>
</tr>
<tr>
<td>5. Can not be automated</td>
<td>Can be automated</td>
</tr>
</tbody>
</table>

Data-driven approach to phonological description and syllable structure analysis:

- was proposed in [Moroz 2018](#)
- was applied to syllable structure in [Moroz 2019](#) to Adyghe data
- was applied to syllable structure in [Romanova 2019](#) to Russian and Macedonian data

I will present an application of this method to Botlikh and Zilo Andi data.
## Phonological description: data-driven analysis

<table>
<thead>
<tr>
<th>Traditional analysis</th>
<th>Data-driven analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Done by trained linguist</td>
<td>Evaluated by trained linguist</td>
</tr>
<tr>
<td>2. Can be done from scratch</td>
<td>Previous description needed</td>
</tr>
<tr>
<td></td>
<td>(or at least prior expectations)</td>
</tr>
<tr>
<td>3. Doesn’t care about amount of data</td>
<td>Care more about amount of data</td>
</tr>
<tr>
<td>4. Less reproducible</td>
<td>More reproducible</td>
</tr>
<tr>
<td>5. Can not be automated</td>
<td>Can be automated</td>
</tr>
</tbody>
</table>

Data-driven approach to phonological description and syllable structure analysis:

- was proposed in [Moroz 2018]
- was applied to syllable structure in [Moroz 2019] to Adyghe data
- was applied to syllable structure in [Romanova 2019] to Russian and Macedonian data
Phonological description: data-driven analysis

<table>
<thead>
<tr>
<th>Traditional analysis</th>
<th>Data-driven analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Done by trained linguist</td>
<td>Evaluated by trained linguist</td>
</tr>
<tr>
<td>2. Can be done from scratch</td>
<td>Previous description needed</td>
</tr>
<tr>
<td></td>
<td>(or at least prior expectations)</td>
</tr>
<tr>
<td>3. Doesn’t care about amount of data</td>
<td>Care more about amount of data</td>
</tr>
<tr>
<td>4. Less reproducible</td>
<td>More reproducible</td>
</tr>
<tr>
<td>5. Can not be automated</td>
<td>Can be automated</td>
</tr>
</tbody>
</table>

Data-driven approach to phonological description and syllable structure analysis:

- was proposed in [Moroz 2018]
- was applied to syllable structure in [Moroz 2019] to Adyghe data
- was applied to syllable structure in [Romanova 2019] to Russian and Macedonian data

I will present an application of this method to Botlikh and Zilo Andi data
Andi and Botlikh villages, created with lingtypology package [Moroz 2017]
<table>
<thead>
<tr>
<th>Andic</th>
<th>Andic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Botlikh</td>
<td>Andi</td>
</tr>
<tr>
<td>Unwritten (can be written with extended Cyrillic script for Avar)</td>
<td>Unwritten (can be written with extended Cyrillic script for Avar)</td>
</tr>
<tr>
<td>~5,000–8,000 speakers</td>
<td>~16,500 speakers</td>
</tr>
<tr>
<td>Mostly spoken in 3 villages in northwestern Daghestan (Russian Federation): Botlikh, Miarso, Ashino, (Ankho); minor dialectal differences</td>
<td>About 14 villages; There are two main dialect groups: Lower Andi (Muni, Kvankhidatli) and Upper Andi (the rest)</td>
</tr>
<tr>
<td>Two dictionaries: [Saidova and Abusov 2012], [Alekseev and Azaev 2019]</td>
<td>No dictionary except [Kibrik and Kodzasov 1988]</td>
</tr>
</tbody>
</table>
Comparing two Botlikh dictionaries

[Saidova and Abusov 2012]

- Compiled in the 2000s by a native speaker (M. G. Abusov) and an experienced linguist (P. A. Saidova)
- Mostly Botlikh with some notes on Miarso
Comparing two Botlikh dictionaries

[Saidova and Abusov 2012]

- Compiled in the 2000s by a native speaker (M. G. Abusov) and an experienced linguist (P. A. Saidova)
- Mostly Botlikh with some notes on Miarso

[Alekseev and Azaev 2019]

- Compiled in the 1960s / 1970s by a native speaker / philologist (X. G. Azaev) and later (in the 2000s) systematized by an experienced linguist (M. E. Alekseev)
- Subsequently edited by T. A. Maisak and scheduled for posthumous publication later this year
- Botlikh only
Comparing two Botlikh dictionaries

Summary:

- Dictionaries were compiled independently of each other
- with no metadata on the speakers consulted
- data collection was separated with several decades break
• Automatically merge two .doc file into one unified .xls file, ...
- Automatically merge two .doc file into one unified .xls file, ...
- Manually check for similarities (S. Verhees, C. Naccarato and me)
Comparing two Botlikh dictionaries

Alekseev, Azaev 2019
6821 lexemes

Saidova, Abusov 2012
8464 lexemes

- 3931 lexemes (31.7%)
- 2890 lexemes (23.3%)
- 5574 lexemes (45%)
Comparing two Botlikh dictionaries

- If we remove the stress sign, there are only 2495 lexemes which look phonetically the same, and 395 are different (14%)
- If we don’t remove the stress sign, there are 2027 lexemes which look phonetically the same, and 863 are different (30%)

⇒ 16% of lexemes have different stress pattern?
Comparing two Botlikh dictionaries

- If we remove the stress sign, there are only 2495 lexemes which look phonetically the same, and 395 are different (14%)
- If we don’t remove the stress sign, there are 2027 lexemes which look phonetically the same, and 863 are different (30%)
⇒ 16% of lexemes have different stress pattern? Including cases where stress is present in one dictionary and absent in the other.

- Multiple cases where stress is present in one dictionary and absent in the other.

- Stressed pattern differences in 317 lexemes (about 11%)
- Multiple cases where there is a small difference that could be explained either as a typo or in terms of phonological variation.
  - čuhí 'to run' [Alekseev and Azaev 2019] vs. čũhí [Saidova and Abusov 2012],
  - kusu 'cherry plum' [Alekseev and Azaev 2019] vs. kusːu [Saidova and Abusov 2012],

- Multiple cases where Russian borrowings were adopted differently.
  - awtobus 'bus' [Alekseev and Azaev 2019] vs. abtabus [Saidova and Abusov 2012],
  - biton 'milk can' [Alekseev and Azaev 2019] vs. bitun [Saidova and Abusov 2012],
  - apteka 'pharmacy' [Alekseev and Azaev 2019] vs. abteka [Saidova and Abusov 2012],

- Morphological preferences.
  - dinija=w 'pious' [Alekseev and Azaev 2019] vs. dinija=b [Saidova and Abusov 2012]
Comparing two Botlikh dictionaries

- If we remove the stress sign, there are only 2495 lexemes which look phonetically the same, and 395 are different (14%)
- If we don’t remove the stress sign, there are 2027 lexemes which look phonetically the same, and 863 are different (30%)

⇒ 16% of lexemes have different stress pattern? Including cases where stress is present in one dictionary and absent in the other.

- What causes the difference between dictionaries?
  - Stress pattern differences in 317 lexemes (about 11%)
  - Multiple cases where there is a small difference that could be explained either as a typo or in terms phonological variation
    - čuhí ‘to run’ [Alekseev and Azaev 2019] vs. čůhí [Saidova and Abusov 2012],
    - Multiple cases where Russian borrowings were adopted differently
      - awtobus ‘bus’ [Alekseev and Azaev 2019] vs. abtabus [Saidova and Abusov 2012],
      - biton ‘milk can’ [Alekseev and Azaev 2019] vs. bitun [Saidova and Abusov 2012],
      - apteka ‘pharmacy’ [Alekseev and Azaev 2019] vs. abteka [Saidova and Abusov 2012]
    - Morphological preferences
      - dinija=w ‘pious [Alekseev and Azaev 2019] vs. dinija=b [Saidova and Abusov 2012]
Comparing two Botlikh dictionaries

About 25 cases:

<table>
<thead>
<tr>
<th>[Alekseev and Azaev 2019]</th>
<th>[Saidova and Abusov 2012]</th>
</tr>
</thead>
<tbody>
<tr>
<td>āhajr</td>
<td>āhar</td>
</tr>
<tr>
<td>bezajr</td>
<td>bezir</td>
</tr>
<tr>
<td>mik’kujr</td>
<td>mik’ːur</td>
</tr>
<tr>
<td>reqχujr</td>
<td>reqχʷir</td>
</tr>
<tr>
<td>reskujr</td>
<td>reskur</td>
</tr>
<tr>
<td>rikʷajr</td>
<td>rikʷar</td>
</tr>
<tr>
<td>χwardar</td>
<td>χwardir</td>
</tr>
<tr>
<td>miʔar</td>
<td>miʔar</td>
</tr>
</tbody>
</table>

...  ...  ...

About 6 cases:

<table>
<thead>
<tr>
<th>ʃːalaj</th>
<th>ʃːallaj</th>
</tr>
</thead>
<tbody>
<tr>
<td>inuʕala</td>
<td>inuʕalla</td>
</tr>
<tr>
<td>ʃila</td>
<td>ʃilla</td>
</tr>
</tbody>
</table>

...  ...  ...

'message'
'roasting'
'swallowing'
'fight'
'overnight stay'
'lighting'
'digging'
'nose'
'silt'
'everywhere'
'reason'
Comparing two Botlikh dictionaries (without stress patterns)
Zilo Andi data

Dictionary data for Zilo were collected during fieldtrips to Zilo in 2016–2019 with N. Rochant, S. Verhees, A. Martynova and A. Zakirova who contributed to the same FieldWorks project.

- Contain morphological affixes
- Doesn’t contain additional affixes in a lemma form
- Contain different stems of the same lexeme (e.g. SG.ABS, SG.OBL, PL.ABS, PL.OBL, PST, NPST)
- No information about stress
Comparing [Saidova, Abusov 2012] and Zilo data

Number of utterances in Andi (log scale) vs. Number of utterances in [Saidova, Abusov 2012] (log scale)

Color scale ranges from blue (-0.05) to yellow (0.05).
Discussion:

- Botlikh dictionaries were specially selected for shared meaning, the same procedure for the Andic dictionary was not done
Discussion:

- Botlikh dictionaries were specially selected for shared meaning, the same procedure for the Andic dictionary was not done.
- Botlikh dictionaries contain a lot of borrowings, this is not true for the Andic dictionary.
Discussion:

- Botlikh dictionaries were specially selected for shared meaning, the same procedure for the Andic dictionary was not done.
- Botlikh dictionaries contain a lot of borrowings, this is not true for the Andic dictionary.
- Lemmata are not the same as wordforms, so the model should be checked with the wordform material.
Discussion:

- Botlikh dictionaries were specially selected for shared meaning, the same procedure for the Andic dictionary was not done.
- Botlikh dictionaries contain a lot of borrowings, this is not true for the Andic dictionary.
- Lemmata are not the same as wordforms, so the model should be checked with the wordform material.
- Lemmata can contain some affix that will shift all frequencies (e.g. INF, PL or =CL) for some types of phonological units.
Discussion:

- Botlikh dictionaries were specially selected for shared meaning, the same procedure for the Andic dictionary was not done.
- Botlikh dictionaries contain a lot of borrowings, this is not true for the Andic dictionary.
- Lemmata are not the same as wordforms, so the model should be checked with the wordform material.
- Lemmata can contain some affix that will shift all frequencies (e.g. INF, PL or =CL) for some types of phonological units.
- It would be nice to compare the obtained models with the models built on corpora data.
Discussion:

- Botlikh dictionaries were specially selected for shared meaning, the same procedure for the Andic dictionary was not done.
- Botlikh dictionaries contain a lot of borrowings, this is not true for the Andic dictionary.
- Lemmata are not the same as wordforms, so the model should be checked with the wordform material.
- Lemmata can contain some affix that will shift all frequencies (e.g., INF, PL or =CL) for some types of phonological units.
- It would be nice to compare the obtained models with the models built on corpora data, when/if it will be available.
Discussion:

- Botlikh dictionaries were specially selected for shared meaning, the same procedure for the Andic dictionary was not done.
- Botlikh dictionaries contain a lot of borrowings, this is not true for the Andic dictionary.
- Lemmata are not the same as wordforms, so the model should be checked with the wordform material.
- Lemmata can contain some affix that will shift all frequencies (e.g. INF, PL or =CL) for some types of phonological units.
- It would be nice to compare the obtained models with the models built on corpora data, when/if it will be available.
- Model computed using row frequencies should be extended with Markov Chains and vector models (like in a Distributional semantics).
Discussion:

- Botlikh dictionaries were specially selected for shared meaning, the same procedure for the Andic dictionary was not done
- Botlikh dictionaries contain a lot of borrowings, this is not true for the Andic dictionary
- Lemmata are not the same as wordforms, so the model should be checked with the wordform material
- Lemmata can contain some affix that will shift all frequencies (e.g. \textsc{inf}, \textsc{pl} or \textsc{=-cl}) for some types of phonological units
- It would be nice to compare the obtained models with the models built on corpora data, when/if it will be available
- Model computed using row frequencies should be extended with Markov Chains and vector models (like in a Distributional semantics): only in this way it will be possible to compare \textbf{paradigmatic} and \textbf{syntagmatic} relations within the phonological systems and across languages
Send me a letter!
agricolamzgmail.com

Presentation is available here:
tinyurl.com/y3wtkcbq
References


