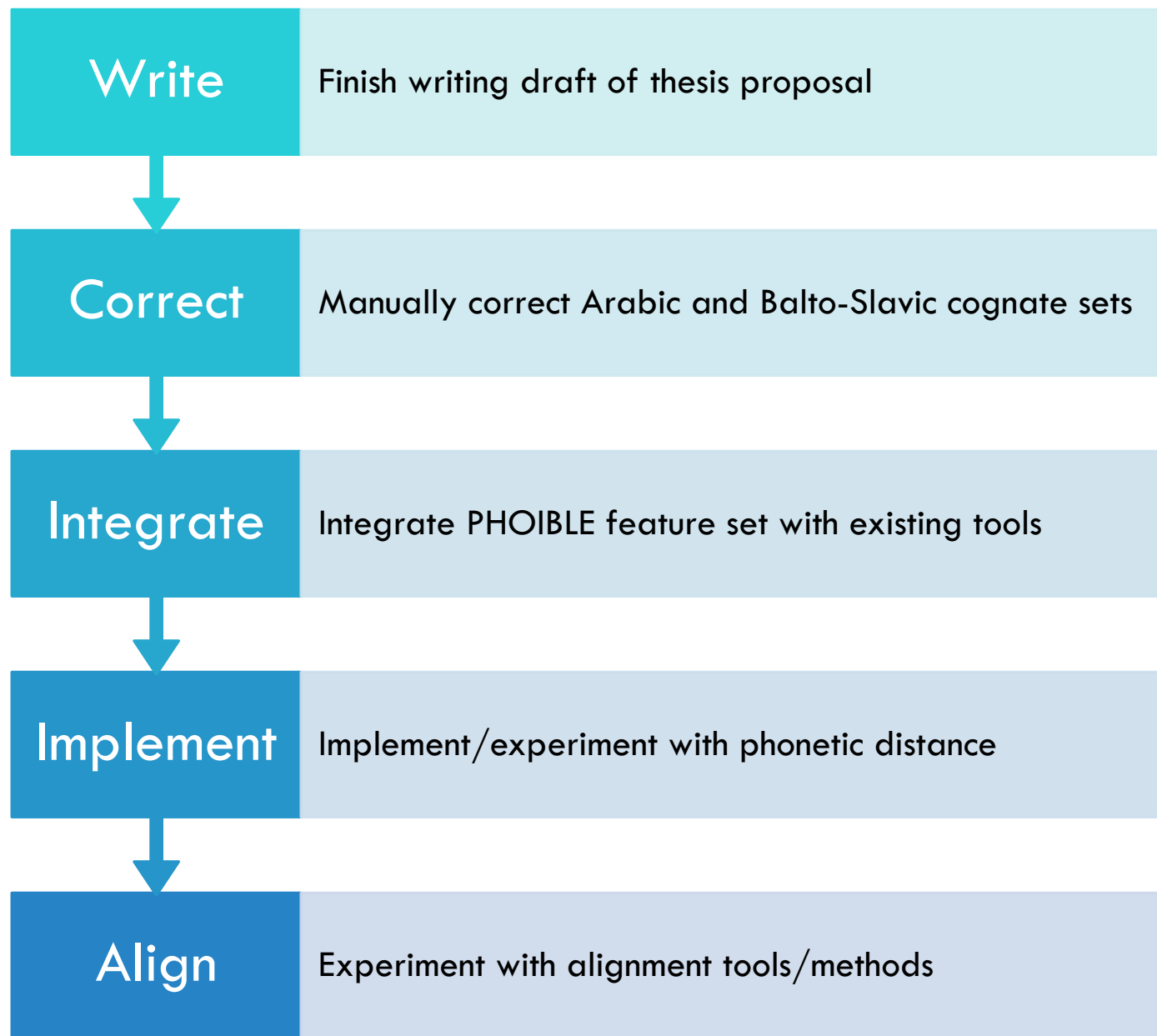




THESIS SEMINAR MEETING

Philip Georgis
July 19, 2021

TASKS FROM LAST TIME



PHOIBLE FEATURES

- Integrated PHOIBLE features into my framework
- Only base segments: e.g. /b, f, t, d, k, l, ɲ, u, e/
- Featural representation of complex/modified segments (e.g. /b^w, f^j, t^h, d̪, k̪p̪, l^y, ɲː, ɹ̥, ẽ/) computed by applying features associated with diacritic to base
- No need to store features for thousands of unique segment/diacritic combinations

PHOIBLE FEATURES

- Small problem: feature table downloaded from <https://github.com/phoible/dev/blob/master/raw-data/FEATURES/phoible-segments-features.tsv> has some errors and doesn't match the features on the PHOIBLE homepage in some cases
 - Looks like it hasn't been updated in a couple years
 - Is it possible to download the feature data somehow directly from the homepage?

PHONETIC FEATURES

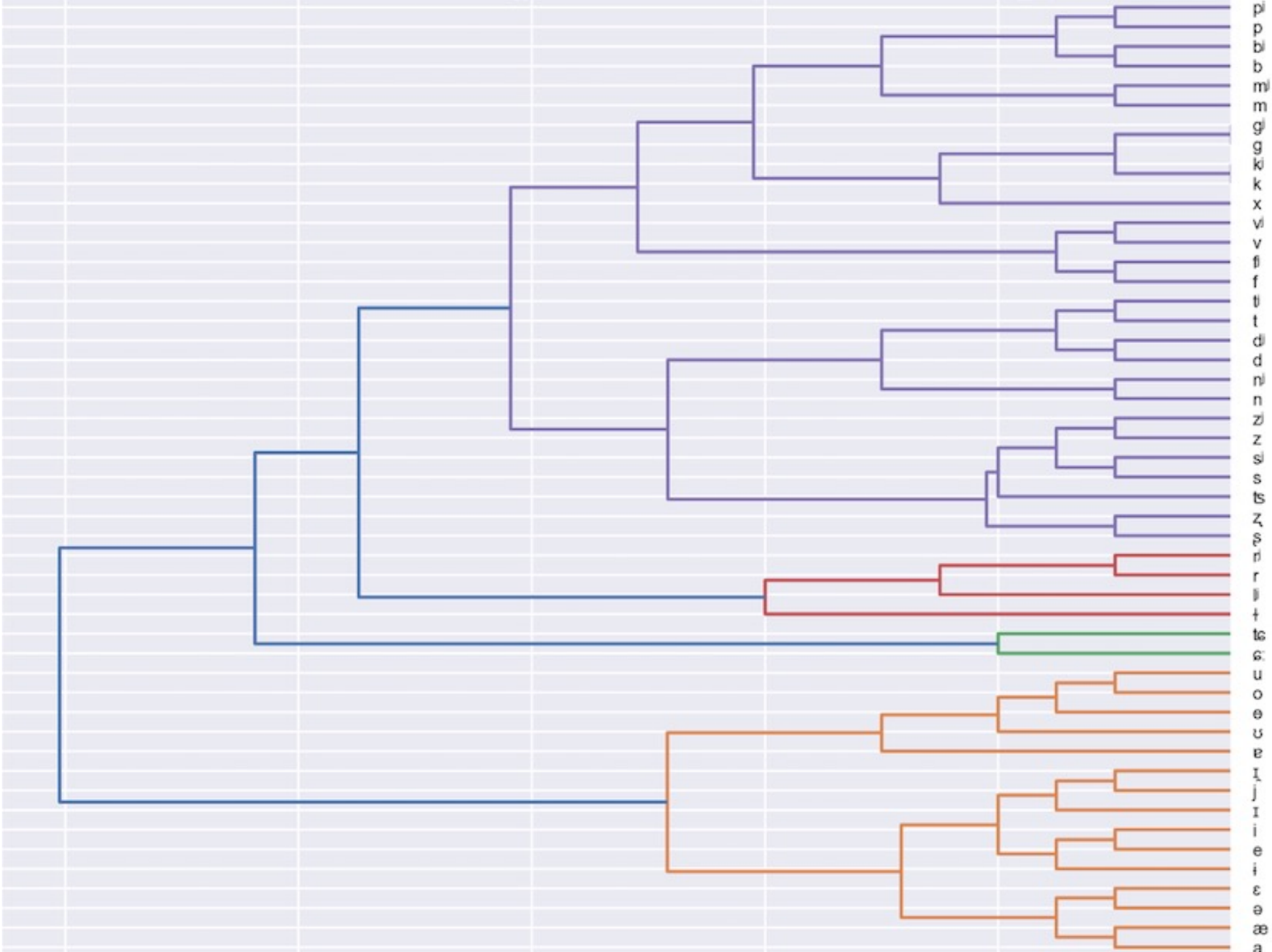
- Found reference to justify re-encoding '0' values as '-'
- Odden, 2005: *Introducing Phonology*, "Feature Theory"
- Therefore features can truly have binary values rather than three-way values

| | p | p ^y | p ^y , p ^w | p ^w | p ^{w̥} | p ^ʔ |
|-------|---|----------------|---------------------------------|----------------|-----------------|----------------|
| hi | - | + | + | + | + | - |
| back | - | - | + | + | - | + |
| low | - | - | - | - | - | + |
| round | - | - | - | + | + | - |

| | | Coronal Obstruents [+cons, -son, +cor] | | | | | | | | | | | | Palatal Obstruents [+cor + dors] | | | | Non-coronal Obstruents [+cons, -son, -cor] | | | | | | | | | | Laryngeals [-cons, -son] | | | | | |
|----------------|---------|---|----------------|----|---|---|---|---|---|---|---|---|---|-------------------------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|-----------------------------|---|---|---|---|---|
| | | t' | t ^h | t̥ | t | d | s | z | ʈ | ʂ | ʈ | ʈ | ʈ | ʈ | p | b | f | v | ɸ | β | k | g | x | ɣ | q | ɢ | χ | ʁ | ħ | ʕ | h | ɦ | ʔ |
| Class features | cons | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | - | - | - |
| | son | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| | syll | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Place features | labial | - | - | - | - | - | - | - | - | - | - | - | - | - | + | + | + | + | + | + | - | - | - | - | - | - | - | - | - | - | - | - | - |
| | round | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | - | - | - | - | - | - | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | coronal | + | + | + | + | + | + | + | + | + | + | + | + | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| | ant | + | + | + | + | + | + | + | + | + | + | + | + | + | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | dist | - | - | + | - | - | - | - | - | - | - | - | + | + | + | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | dorsal | - | - | - | - | - | - | - | - | - | - | - | + | + | + | - | - | - | - | - | - | + | + | + | + | + | + | + | + | - | - | - | - |
| | high | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | + | + | + | 0 | 0 | 0 | 0 | 0 | 0 | + | + | + | + | - | - | - | - | 0 | 0 | 0 | 0 |
| low | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | - | - | 0 | 0 | 0 | 0 | 0 | 0 | - | - | - | - | - | - | - | - | 0 | 0 | 0 | 0 | |
| back | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | - | - | - | 0 | 0 | 0 | 0 | 0 | 0 | + | + | + | + | + | + | + | + | 0 | 0 | 0 | 0 | |
| tense | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | - | - | - | 0 | 0 | 0 | 0 | 0 | 0 | - | - | - | - | - | - | - | - | 0 | 0 | 0 | 0 | |

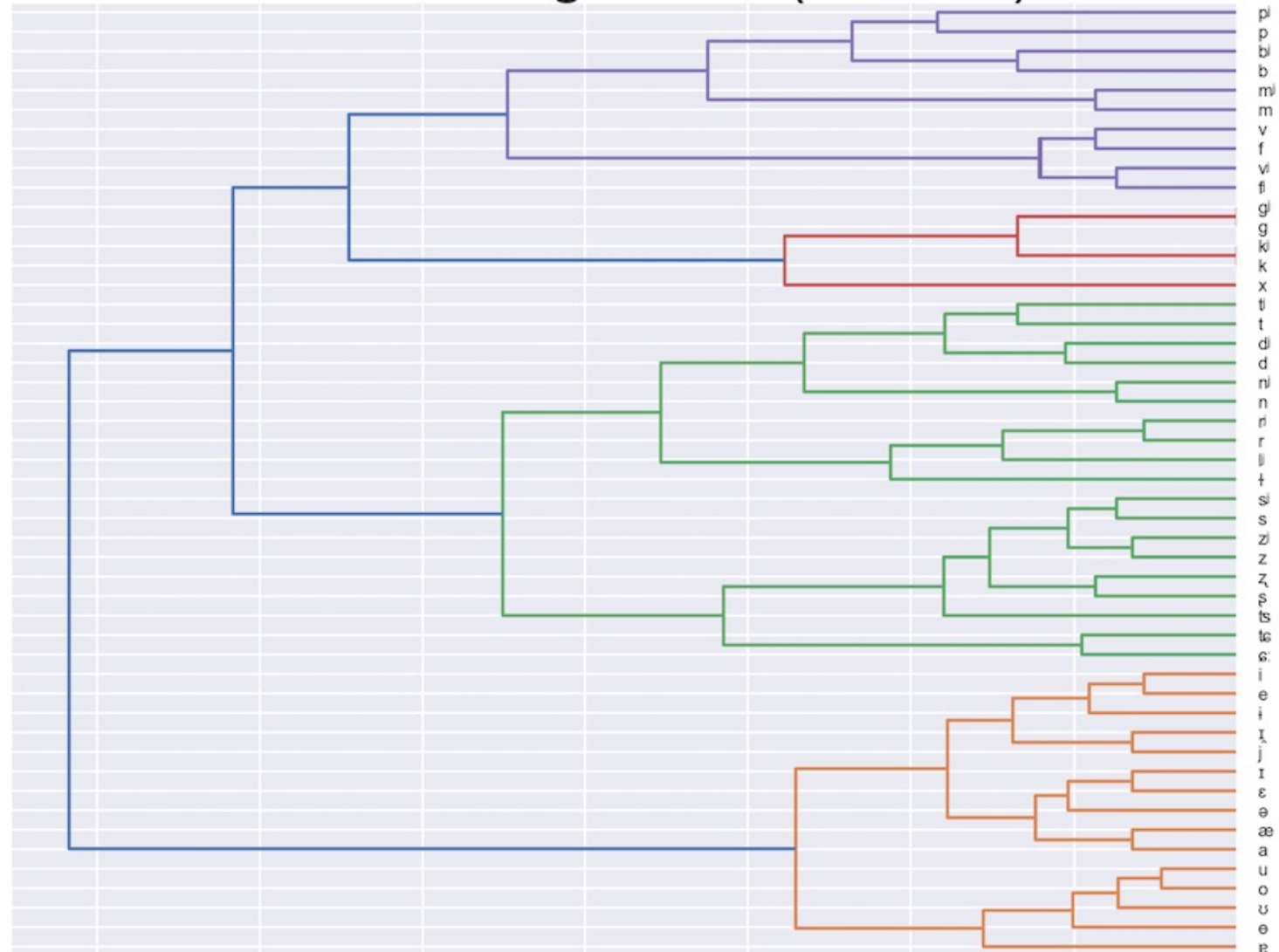
- Hamming distance
- Cosine similarity
- Jaccard index
- Dice dissimilarity

Dendrogram of Russian phones according to Hamming distance



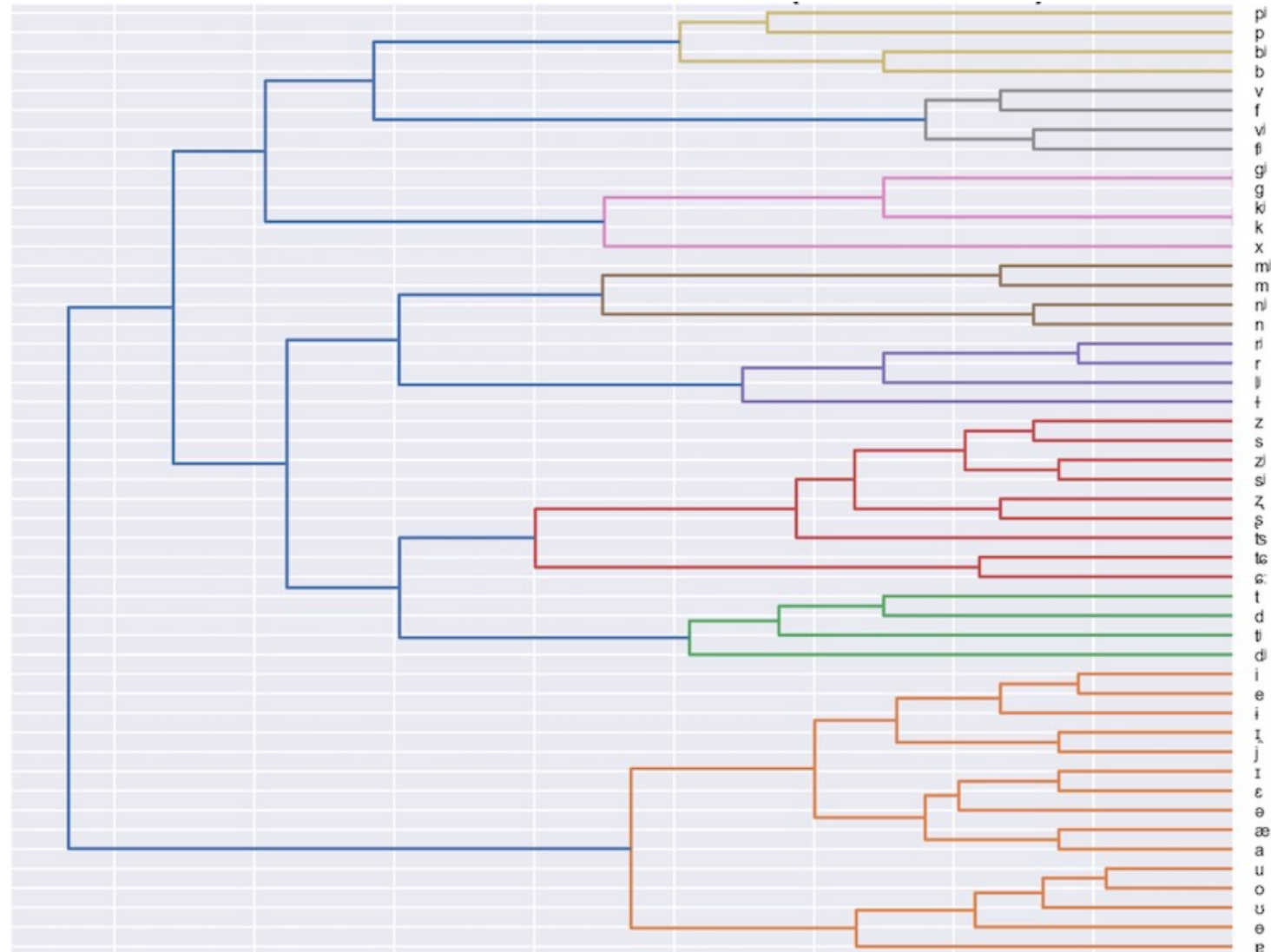
- Which distance/similarity measure to use?
 - Hamming distance
 - Cosine similarity
 - Jaccard index
 - Dice dissimilarity

Dendrogram of Russian phones according to cosine similarity



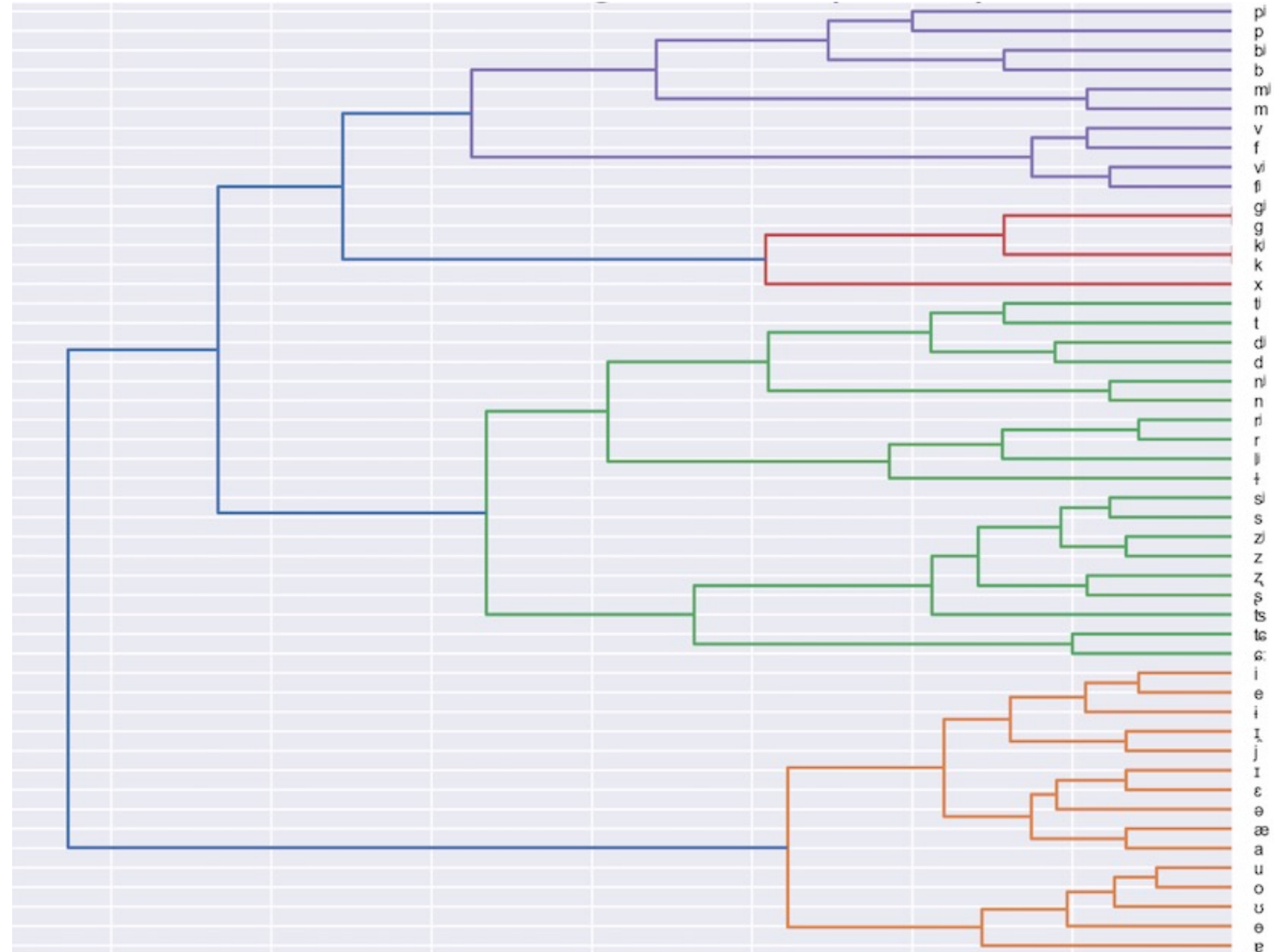
- Which distance/similarity measure to use?
 - Hamming distance
 - Cosine similarity
 - Jaccard index
 - Dice dissimilarity

Dendrogram of Russian phones according to Jaccard index



- Which distance/similarity measure to use?
 - Hamming distance
 - Cosine similarity
 - Jaccard index
 - Dice dissimilarity

Dendrogram of Russian phones according to Dice dissimilarity



PHONETIC SEQUENCE ALIGNMENT

- Alignment yielded using only pairwise phone distances as substitution costs won't produce correct alignments in cases where genuine phoneme correspondences are between non-similar segments and/or where many gaps are needed

Maori /taŋata/ - Tahitian /taʔata/ 'PERSON'

*

| | | | | | | | |
|------------|---|---|---|---|---|---|---|
| MRI | t | a | ŋ | | a | t | a |
| TAH | t | a | | ʔ | a | t | a |

Spanish /oxo/ - French /œj/ 'EYE'

*

| | | | |
|------------|---|---|---|
| SPA | o | x | o |
| FRA | œ | | j |

$\Lambda > ʒ > ʃ > x$
 $\Lambda > j$

Croatian /topao/ - Russian /tjɵpɫɪ/ 'WARM'

*

| | | | | | | |
|------------|----|---|---|---|---|---|
| HRV | t | o | p | | a | o |
| RUS | tj | ɵ | p | ɫ | ɪ | ɪ |

| | | | | | | | |
|------------|----|---|---|---|---|---|---|
| HRV | t | o | p | a | o | | |
| RUS | tj | ɵ | p | | ɫ | ɪ | ɪ |

$\mathfrak{t} > w > u > o$

ALIGNMENT SOLUTIONS

Multiple alignment: combining pairwise alignments might result in better alignments

| | | | | | |
|------------|---|---|---|---|---|
| HRV | t | o | p | a | o |
| BUL | t | o | p | ə | l |

| | | | | | | | |
|------------|----|---|---|---|---|---|----|
| BUL | t | o | p | ə | l | | |
| RUS | tʲ | ə | p | | l | i | ɪ̯ |

| | | | | | | |
|------------|----|---|---|---|---|---|
| HRV | t | o | p | a | o | |
| POL | tɛ | ɛ | p | | w | i |

| | | | | | | |
|------------|----|---|---|---|---|----|
| POL | tɛ | ɛ | p | w | i | |
| RUS | tʲ | ə | p | l | i | ɪ̯ |



| | | | | | | | |
|------------|----|---|---|---|---|---|----|
| POL | tɛ | ɛ | p | | w | i | |
| HRV | t | o | p | a | o | | |
| BUL | t | o | p | ə | l | | |
| RUS | tʲ | ə | p | | l | i | ɪ̯ |

ALIGNMENT SOLUTIONS

Multiple alignment: combining pairwise alignments might result in better alignments

| | | | | | |
|------------|---|---|---|---|---|
| HRV | t | o | p | a | o |
| BUL | t | o | p | e | l |

| | | | | | | | |
|------------|----|---|---|---|---|---|---|
| BUL | t | o | p | e | l | | |
| RUS | tʲ | ə | p | | l | i | ɪ |

| | | | | | | |
|------------|----|---|---|---|---|---|
| HRV | t | o | p | a | o | |
| POL | tɛ | ɛ | p | | w | i |

| | | | | | | |
|------------|----|---|---|---|---|---|
| POL | tɛ | ɛ | p | w | i | |
| RUS | tʲ | ə | p | l | i | ɪ |

*LingPy's multiple alignment tool with default
setting $GOP = -1$*




| | | | | | | | |
|---|------------|----|---|---|---|---|---|
| * | POL | tɛ | ɛ | p | w | i | |
| | HRV | t | o | p | | a | o |
| | BUL | t | o | p | | e | l |
| | RUS | tʲ | ə | p | l | i | ɪ |

*Computationally expensive, and still
incorrect...*

ALIGNMENT SOLUTIONS

- Adding trigram information content might help to some extent




| | | | | | | |
|------------|----|---|---|---|---|----|
| HRV | t | o | p | | a | o |
| RUS | tʃ | ə | p | ɫ | ɪ | ɪ̯ |

| | | | | | | | |
|------------|----|---|---|---|---|---|----|
| HRV | t | o | p | | a | o | |
| RUS | tʃ | ə | p | ɫ | | ɪ | ɪ̯ |

still not correct alignment, but closer...

Correct alignments:

| | | | | | | | |
|------------|----|---|---|---|---|---|----|
| HRV | t | o | p | a | o | | |
| RUS | tʃ | ə | p | | ɫ | ɪ | ɪ̯ |




| | | | |
|------------|---|---|---|
| SPA | o | x | o |
| FRA | œ | | j |

| | | | | |
|------------|---|---|---|---|
| SPA | o | x | | o |
| FRA | œ | | j | |

| | | | |
|------------|---|---|---|
| SPA | o | x | o |
| FRA | œ | j | |

ALIGNMENT SOLUTIONS

- Adding trigram information content might help to some extent




| | | | | | | |
|------------|----|---|---|----|---|----|
| HRV | t | o | p | | a | o |
| RUS | tj | ə | p | l̩ | i | ɪ̃ |

| | | | | | | | |
|------------|----|---|---|----|---|---|----|
| HRV | t | o | p | | a | o | |
| RUS | tj | ə | p | l̩ | | i | ɪ̃ |

still not correct alignment, but closer...

Correct alignments:

| | | | | | | | |
|------------|----|---|---|---|----|---|----|
| HRV | t | o | p | a | o | | |
| RUS | tj | ə | p | | l̩ | i | ɪ̃ |



| | | | | | | |
|------------|---|---|---|---|---|---|
| SPA | o | r | e | x | a | |
| FRA | ɔ | | | ʁ | ɛ | j |

| | | | | | | | |
|------------|---|---|---|---|---|---|---|
| SPA | o | r | | e | | x | a |
| FRA | ɔ | | ʁ | ɛ | j | | |

| | | | | | |
|------------|---|---|---|---|---|
| SPA | o | r | e | x | a |
| FRA | ɔ | ʁ | ɛ | j | |