Considering the classification of some East Slavic lects with the automatic search of linguistic SNP markers

Ilia Afanasev

A lifetime of coevolution

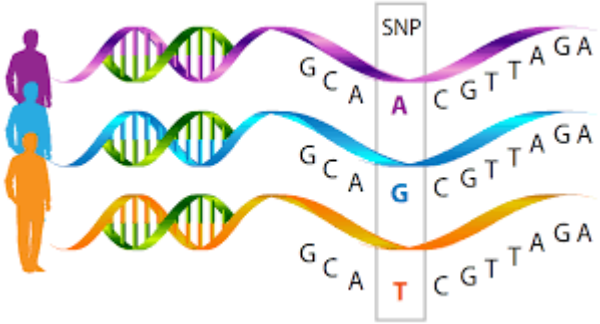
- Developed and emerged as a discipline simultaneously in the 19th century
- Some important concepts in comparative linguistics are explicitly inspired by evolutionary biology (Schleicher, 1863)
- With appearance of computational methods in evolutionary biology, comparative linguistics followed the way, forming the discipline of computational historical linguistics (Jäger, 2019)

Common challenges and common solutions

- Nowadays, disciplines demonstrate the convergent evolution (Starostin, 2022)
- Language is no longer a living system, but an evolvable one (Ladoukakis et al., 2022)
- Prediction of evolution direction becomes more important for evolutionary biology and comparative linguistics (Hejnol and Martindale, 2008; Sims-Williams, 2022; Marlo et al., 2022)
- Statistical methods are important for both the disciplines (Flego, 2022; Engelman, 2023)

Single nucleotide polymorphism

- In genetics, a variation in single position in a DNA sequence among the individuals
- In linguistics, there is no (yet) analogue for DNA
  - (Rama et al., 2014: 3) hypothesise it to be a sequence of *characters* (??)
- Thus, currently it is at least very complicated to find linguistic SNPs (ling-SNPs), changes in the language inner structure
- Language change, however, definitely happens, and there are events that signal about change in a sequence of characters that define the language features
- These events leave visible traces in the units of the language

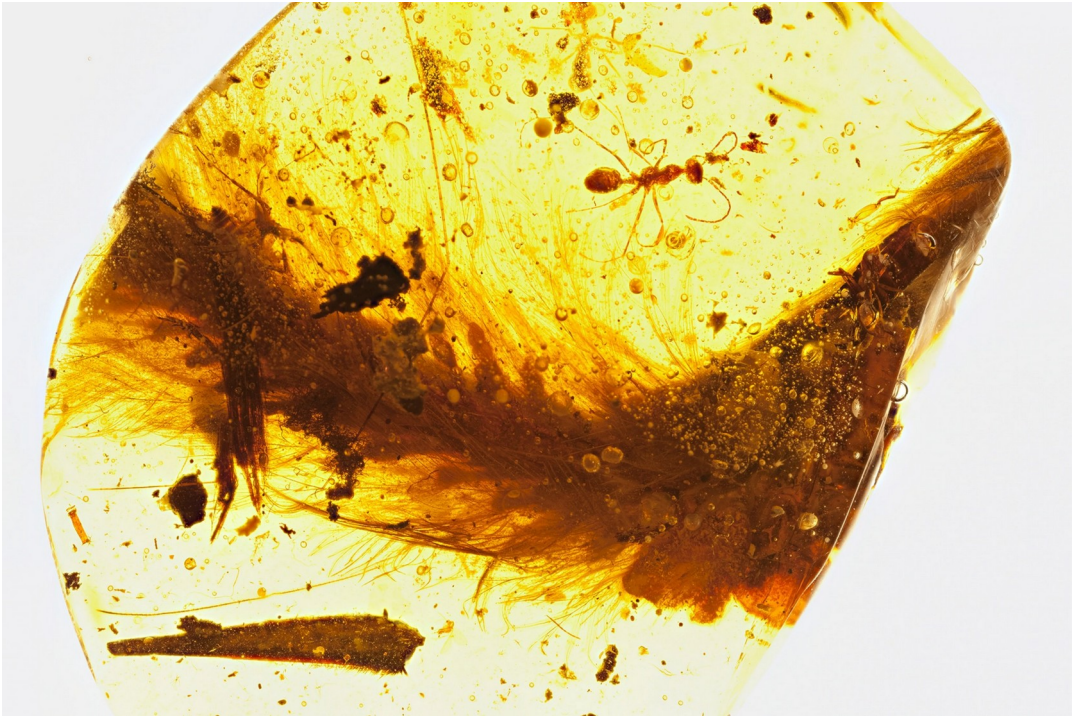


Linguistic single nucleotide polymorphism (ling-SNP) markers

- Entities, directly observed in the empirical data (word list or corpus), that contain the traces of the change in a language
- Do not necessarily map to particular units of language
- Examples include:
  - Stress change
  - Morphological change
  - Lexical swap

Swadesh list: a chronicle of amber

- There are spans of text, where one is more likely to find ling-SNP markers
- Swadesh list items are an example of such a span



Example: East Slavic ‘mother’

- A member of 40-item Swadesh list (Holman et al., 2008)
- Standard Belarusian: *ма**ці***, standard Russian: *ма**ть***, Khislavichi: *ма**ць***
- SNP marker: *ці/ть/ць*
- Possible language structure changes: vocalisation of a reduced front vowel (Khislavichi, Russian - Belarusian), palatalisation of a dental consonant before front vowel (Khislavichi, Belarusian - Russian)
- Levenshtein distances:
  - Khislavichi - Russian: 1.0, normalised: 0.25
  - Khislavichi - Belarusian: 1.0, normalised: 0.25
  - Belarusian - Russian: 2.0, normalised: 0.5

Case study: Khislavichi lects on the East Slavic tree

- The position of Khislavichi lects between the Russian and Belarusian lects is controversial (Karski, 1903; Durnovo et al., 1915; Zakharova and Orlova, 2004; Ryko and Spiricheva, 2022)
- Traditional comparison by Swadesh list is not sensitive enough, as the lects are closely related (Nerbonne et al., 1999)

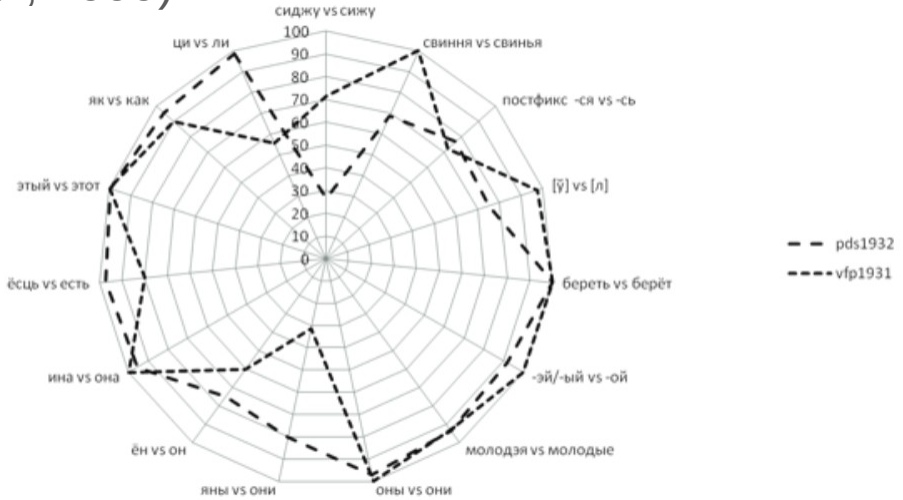


Figure from (Ryko and Spiricheva, 2022)

Method

- Extraction of linguistic SNP markers
- Distance measurement: Levenshtein distance normalised divided (LDND) over the spans of texts that contain l-SNP markers
- Classification: rooted tree over distance-tree matrix

Automatic search of ling-SNP markers

- There are no ready word lists for Khislavichi, and there is a corpus of Khislavichi texts (approximately 100 000 tokens)
- There are ready word lists for Slavic languages, and big corpora for Slavic languages
- Solution:
  - automatic extraction of a particular type of lexical items (i. e., Swadesh list items) that may contain ling-SNP markers from Khislavichi corpus by the model trained on the various Slavic languages
  - currently, manual check and manual extraction of ling-SNP markers (automatic extraction is not yet possible)

Automatic search for Swadesh list items: methods

- Statistical methods
  - **Hidden Markov Model (HMM)**
  - **Conditional Random Fields (CRF)**
- Neural networks
  - **DistilBERT for Named Entity Recognition (DistilBERT-NER)**
  - VERNet (for Grammatical Error Detection/Correction)
- Data augmentation
  - Token 3-grams

Training data

Dataset	Language	Script	Size
CLTT	Czech	Latin	36 000
SET	Croatian	Latin	199 000
Belarusian-HSE	Belarusian	Cyrillic	305 000
Taiga	Russian	Cyrillic	197 000

Performance on the training dataset evaluation part

Model	Precision	Recall	F1-score
HMM	0.64	0.99	0.77
HMM-augmented	0.64	0.99	0.78
CRF	**1**	0.87	0.93
CRF-augmented	0.99	**1**	**0.99**
DistilBERT-NER	0.23	0.66	0.34
DistilBERT-NER-augmented	0.85	0.11	0.2

Tokens found in Khislavichi by CRF-augmented

Type	Swadesh list items	Non-Swadesh lexical items that may contain ling-SNP markers	Errors
Tokens	цябе ‘you-SG.GEN’, людзя́м ‘people-PL.DAT’, пришоł ‘come-PAST.SG.M’, ноччу ‘night-SG.INS’, лисця ‘leaf-PL.NOM’, hарах ‘mountain-PL.LOC’, дзярэ́ўях ‘tree-PL.LOC’, и́мя ‘name-SG.NOM’	фатаhрафираваць ‘take_a_picture-INF’, вот ‘here’	маць ‘mother-SG.NOM’, hлаза ‘eye-PL.NOM’ *inter alia*

Orthographic normalisation

- Getting Khislavichi tokens closer to their actual form: ночью > ноччу ‘night-SG.INS’, тебя > цябе ‘you-SG.GEN’
- Adding graphemes to better depict unique Khislavichi sounds: полный > пoўный ‘full-SG.NOM.M’
- Unification of akanje manifestation: гора > гара ‘mountain-SG.NOM’ (Russian)
- Differentiation of voiced velar fricative/velar plosive: гара > hara ‘mountain-SG.NOM’ (Belarusian)
- Putting accents where necessary: імя > имя́ ‘name-SG.NOM’ (Belarusian)
- Unification of different graphical manifestations of the same sound: лісцця > лисцця ‘leaf-PL.NOM’ (Belarusian)

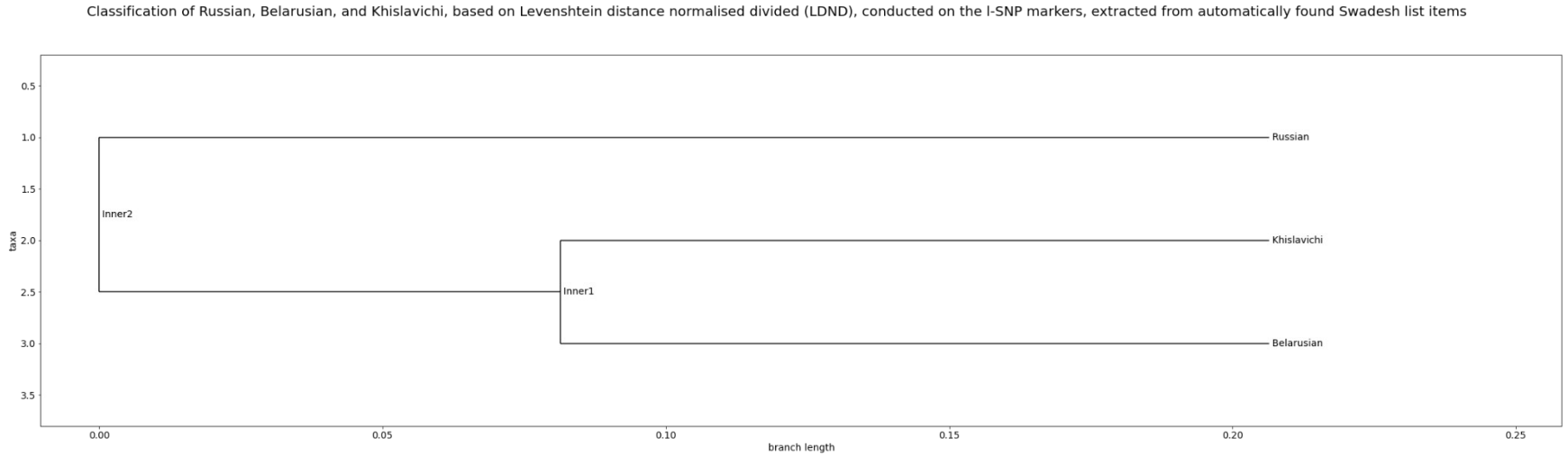
Detected l-SNP markers (Khislavichi/Russian/Belarusian)

- **Phonetic segmental:** оłн/олн/оўн
- **Phonetic suprasegmental:** юдзя́м/ю́дям/юдзя́м
- **Morphological segmental:** ahрафирав/аграфирав/ahрафав
- **Morphological suprasegmental:** дзяревья/деревья/дрэвы
- **Lexical swap:** видзиць/видит/бачыць

Classification basis: triangular matrix of pairwise LDND

Lect	Belarusian	Russian	Khislavichi
Belarusian	0	#	#
Russian	0.5	0	#
Khislavichi	0.25	0.33	0

Classification results



Current achievements and desired enhancements

+ Ling-SNP markers show different types of language variation, from phonetic to lexical, while rely on the verifiably efficient for the genetic classification Swadesh list items
+ Non-Swadesh list items found by model provide additional insights into the differences between lects
+ Hyper-sensitivity of LDND is helpful for the East Slavic material, problematic for traditional lexicostatistics methods (Starostin, 1989)
+ A very special place of Khislavichi lects in the East Slavic system is shown

- Low recall for Swadesh list items require new methods of automatic ling-SNP markers search
- LDND is hyper-sensitive, it is not scalable for the deeper classification (Prokić and Moran, 2013) - this requires other methods
- UPGMA/distance-tree matrix is significantly less informative in comparison to current computational phylogenetic algorithms, such as Naive Bayes, which are to be implemented
- Comparing standard and territorial varieties does not seem exactly fair

Further research

- Quest for the new automatic ling-SNP markers search and extraction methods
- Systemic investigation of Khislavichi/Russian/Belarusian Swadesh and basic vocabulary lists
- Search for other sources of ling-SNP markers
- Including the material of the other East Slavic lects for triangulation/outgroup comparison
- Search for new classification and distance measurement methods
- Tests on the other material and thorough discussion of a method are required

Thank you!

References

Nikolaj N. Durnovo, Nikolaj N. Sokolov, and Dmitrij N. Ushakov. 1915. Opyt dialektologicheskoj karty russkogo jazyka v Evrope s prilozheniem Ocherka russkoj dialektologii.Sinodal'naja tipografija,Moscow.

Russell K. Engelman. A Devonian Fish Tale: A New Method of Body Length Estimation Suggests Much Smaller Sizes for Dunkleosteus terrelli (Placodermi: Arthrodira). *Diversity*, 2023; 15 (3): 318 DOI: 10.3390/d15030318

Flego, S. (2022). The Emergence of Vowel Quality Mutation in Germanic and Dinka-Nuer: Modeling the Role of Information-Theoretic Factors Using Agent-Based Simulation. Indiana University

Hejnol, A.; Martindale, M.Q. (2008). "Acoel development supports a simple planula-like urbilaterian". *Philosophical Transactions of the Royal Society of London B*. **363** (1496): 1493–1501. doi:10.1098/rstb.2007.2239

Holman, E., Wichmann, S., Brown, C., Velupillai, V., Müller, A., Bakker, D. Explorations in automated language classification. Folia Linguistica, v.42, 331-354 (2008).

Jäger, Gerhard. "Computational historical linguistics" *Theoretical Linguistics*, vol. 45, no. 3-4, 2019, pp. 151-182. https://doi.org/10.1515/tl-2019-0011

Yefim F. Karskij. 1903. Belorusy. Vol. I. Vvedenie v izuchenie jazyka i narodnoj slovesnosti. Warsaw.

Ladoukakis, M., Michelioudakis, D., Anagnostopoulou, E. Toward an evolutionary framework for language variation and change // BioEssays, vol. 44, pp. 210 – 216.

Michael R. Marlo, Rebecca Grollemund, Thanh Nguyen, Erik Platner, Sarah Pribe, Alexa Thein. A phylogenetic classification of Luyia language varieties // Sibanda, Galen, Ngonyani, Deo, Choti, Jonathan & Biersteker, Ann (eds.). 2022. Descriptive and theoretical approaches to African linguistics: Selected papers from the 49th Annual Conference on African Linguistics. (Contemporary African Linguistics 6). Berlin: Language Science Press. DOI: 10.5281/zenodo.6358613

Prokić, J., Moran, S. Black box approaches to genealogical classification and their shortcomings. 2013. In: Borin, L., Saxena, A. (eds.) Approaches to Measuring Linguistic Differences (Trends in Linguistics. Studies and Monographs 265. Walter De Gruyter GmbH, Boston/Berlin, 2013.

Rama, Taraka, Kolachina, Sudheer, Bai B, Lakshmi (2014). Quantitative methods for Phylogenetic Inference in Historical Linguistics: an experimental case study of South Central Dravidian // CoRR, arXiv, abs/1401.0708.

Anastasiia Ryko and Margarita V. Spiricheva. 2022. The degree of preservation of dialectal features in different generations (khislavichi district of the smolensk region). RSUH/RGGU Bulletin. "Literary Theory. Linguistics. Cultural Studies" Series, (5):121–141.

Schleicher, A. Die Darwinsche Theorie und die Sprachwissenschaft – offenes Sendschreiben an Herrn Dr. Ernst Haeckel. Weimar, H. Boehlau (1863).

Sims-Williams, H. (2022). Token frequency as a determinant of morphological change. *Journal of Linguistics, 58*(3), 571-607. doi:10.1017/S0022226721000438

Starostin, G. The value of "triangulation" in determining phylogenetic relationship: on the areal and genetic connections of the Bertha languages //Language in Africa, 3(2), pp. 352–367. 2022.

Starostin, S. A. Sravnitel'no-istoricheskoe yazykoznanie i leksikostatistika // Lingvisticheskaya rekonstrukciya i drevnejshaya istoriya Vostoka. Chast' 1. 1989. – S. 407 – 447.

Kapitolina F. Zaharova and Varvara G. Orlova. 2004. Dialektnoe chlenenie russkogo yazyka. URSS, Moscow.