Issue 3 / Часть 3

Articles / Статьи

Ilia Afanasev
HSE University, Moscow, Russia / MTS AI, Moscow, Russia; ilia.afanasev.1997@gmail.com

Cipher, transform, get lost: an anti-transparent system for distance measurement in East Slavic lects

Recent advances in computational historical linguistics have inspired a discussion on newly implemented quantitative methods — mainly, it is about their lack of transparency, and the ways to overcome it. This paper aims to demonstrate the advantages of transparency for such tools.

The study compares two types of language distance measurement systems used in classification. Black-box systems transform the input data (such as the Swadesh list) into output data (language distance) with human- and machine-unexplainable decision-making. Language-agnostic systems (such as string similarity measures) analyse the input data and produce output data transparently, but do not consider the specifics of each language. For a proper comparison, I propose a new anti-transparent system based on hashing algorithms, vectorisation and language contact emulation.

For my purposes, I use material from two test groups — East Slavic and Taa, both lexical and grammatical. East Slavic data are extracted from the corpora of Belogornoje, Megra, and Khislavichi and feature lists of Mokshenskaja, Kritskovschina and Pestschanka. Taa material consists of previously published Swadesh lists for the closely related !Xóõ (!Xoong), Kakia (Masarwa) and N|u‖en. An important new contribution of this work is the publication of new Swadesh wordlists for three East Slavic dialects.

Keywords: black-box methods; East Slavic languages; South Khoisan languages; Tuu languages; language-agnostic methods; automatic language distance measurement; automatic classification; string similarity measures; basic lexicon.

Background

Recent innovations in neural network methods (Sutskever et al. 2013; Cho et al. 2014; Vaswani et al. 2017; Lewis et al. 2019) and training data have critically facilitated a plethora of tasks from grammatical error correction (Syvokon et al. 2023) to sentiment analysis (Barić et al. 2023). However, these results came at the disturbing cost of interpretability (Bastings et al. 2022). The need for explanatory techniques for machine learning (ML) systems has become more viable than ever (Munn & Pitman 2022).

Historical comparative linguistics was among the first fields to raise suspicions about these cutting-edge technologies (Jäger 2019). It was among the first to adopt computational phylogenetic methods, right after dialectometry (Nerbonne and Heeringa 1997; Wichmann et al. 2011; Snoek 2013; List 2014; Rama & Borin 2015; Wichmann & Rama 2018). At the same time, it remains rightfully aware of the new methods becoming less and less interpretable (Carvalho 2020).

In this fashion, Prokić & Moran 2013 require the methods of automatic language distance measurement to be transparent and linguistically explainable. Their otherwise valid point, however, depends on the Levenshtein distance being a black box method (Prokić & Moran 2013: 442), which it is not. It is a transparent method, even if it is language-agnostic.

Black box and language-agnostic approaches are distinctly different in the following aspects: transparency, implementation, reproducibility, overall efficiency, and application scope.

The most important one is explainability, the possibility for a human researcher to trace the inner workings of a model, either manually or with the help of automatic tools (Munn & Pitman 2022). Table 1 demonstrates the key differences between these two types of systems.

System property	Black-box model	Language-agnostic
Transparency	As non-transparent as possible	As transparent as possible
Explainability	Non-explainable by definition	Explainable, often inherently
Implementation	Implemented based on the researcher's idea of what may efficiently transform input to output, with little attention to the inherent properties of the studied object	Implemented based on the inherent properties of the studied object (sequentiality of the string, whether a genetic code sequence or a word in a list)
Reproducibility	Almost impossible to reproduce reliably	Easily reproducible
Transferability	Designed for a specific task, attempts at applying it to other tasks lead to unpredictable fluctuations in results	Designed for a specific type of task
Efficiency	Vary in efficiency	Achieve high enough score

Table 1. The crucial differences in system properties between black-box and language-agnostic models.

This paper intends to illustrate these differences by building a system that matches the definition of a black box as closest to completely nonsensical decision-making mechanisms.

I hypothesise the following:

H1. A true black-box method heavily differs from language-agnostic methods in the degree to which one may linguistically explain its functionality.

H2. Using a true black-box method leads to building a less reliable classification than an application of a language-agnostic method.

H3. A black-box method is much more sensitive to farther degrees of relationship than a language-agnostic method.

I compare a black-box method and a language-agnostic method with different human classifications of specific lects under consideration as well as to a set of language-aware computational historical linguistics methods.

Necessity of cross-evaluation

The introduction of a completely new method requires some necessary steps from a researcher, lack of which may undermine the whole intent of the process. I unite these actions under an umbrella of cross-evaluation, a set of practices that help to highlight the possible restrictions of the introduced method. Cross-evaluation may be implemented for both method and data.

Data cross-evaluation shows the advantages and disadvantages of an introduced method by using different datasets.

Method cross-evaluation demonstrates the advantages and disadvantages of an introduced method by comparing its results to the results acquired by other previously established, and proven efficient, methods. In the case of genetic classification, this includes a comparison between automatic methods and human classification. There are cases, when one may not rely on human classification as a gold standard, due to the lack of consensus between the researchers.

Mostly these are either poorly studied families, not to mention macrofamilies (see examples in Starostin 2011, Vajda 2012, Zhivlov 2021). However, internal reconstruction of some well-established families, such as Slavic, also may lead to this issue (for instance, Feld and Maxwell 2019; Ryko and Spiricheva 2022). It does not mean that automatic classification is more objective or better, yet human data cannot be employed as an indisputable gold standard in the automatic classification task in numerous cases that demonstrate a lack of consensus on the genetic classification even by the human researchers. I propose an approach akin to the Natural Language Generation (NLG). This approach requires us to be cautious of the human data, especially in vague cases of long-distance relationship, or close lects classification, and demanding of the metrics (Novikova et al. 2017).

Data

I utilise three datasets, two in the form of a Swadesh list, and one in the form of a lect feature listing.

The main dataset consists of two corpora of East Slavic small territorial lects. These are the Saratov dialect corpus (Kryuchkova & Goldin 2011) and the Khislavichi corpus (Ryko & Spiricheva 2020). The Saratov dialect corpus represents (among others) two Russian lects: Megra and Belogornoje. The Khislavichi corpus represents the single East Slavic lect of disputable genetic attribution between the Belarusian and Russian continua (Ryko & Spiricheva 2022; Afanasev 2023).

Megra is a northern (Kryuchkova & Goldin 2011) Russian dialect, spoken in the Vologda Region. The Megra corpus consists of transcribed interviews (mostly slice-of-life stories) from Saratov State University field trips (1980–2019). Belogornoje is a central (Barannikova 2005) Russian dialect. The Belogornoje corpus consists of transcribed folklore tales and interviews from Saratov State University field trips (1980–2019). Khislavichi is spoken in the Smolensk Region (Russia). Ryko and Spiricheva (2022) treat it as a Northern Belarusian dialect, intensely Russified during the XXth century. Khislavichi material is a collection of slice-of-life stories, gathered during a 2019 field trip.

There is no ready-made Swadesh list for any of these lects and there are no dictionaries. I collect a basic 40-word list (Holman et al. 2008), generally following the guidelines for East Slavic languages (Kassian et al. 2010). Within classical lexicostatistics framework, all the items would be clear matches and comparison would be meaningless. However, as the metrics applied in this article are more sensitive, using a 40-word list may yield meaningful results, as proven earlier in computational dialectometry (Nerbonne & Heeringa 1997; Gooskens & Heeringa 2004) and computational phylogenetic linguistics (Holman et al. 2008).

I present the concepts and their realisations for each lect in the form of phonetic transcriptions. The symbolic representations of sounds are taken from IPA. The transcriptions are phonematic and preserve key features of the lects, such as *okanje*, distinguishing between [o] and [a] allophones of phoneme <o> in the first unstressed position; but not the individual features of speakers. The transcription also preserves all the irregularities (such as *okanje* in *voda* ‘water’ ‘fire’ in Khislavichi, a dialect more prone to *akanje*, a coincidence of [o] and [a] allophones of phoneme <o> in the first unstressed position). As word stress does not contribute to the differences between the given lects and the distance measurement methods, discussed in the article, do not utilise it (Holman et al. 2008), the transcriptions do not represent it.

Concept	Megra (Northern Russian, Vologda Region, Russia)	Belogornoje (Central Russian, Saratov Region, Russia)	Khislavichi (Northern Belarusian, Smolensk Region, Russia)
eye	gl^{ɣ}as	gl^{ɣ}as	ɣl^{ɣ}as
ear	uxo	uxo	vuxa
nose	nos	nos	nos
tongue	jazɨk	jazɨk	jazɨk
tooth	zup	zup	zup
hand	ruka	ruka	ruka
knee	kol^{ʲ}eno	kol^{ʲ}eno*	kal^{ʲ}ena
blood	krof	krof^{ʲ}	krow
bone	kos^{ʲ}t^{ʲ}*	kos^{ʲ}t^{ʲ}*	kos^{ʲ}t^{ʲ}*
breast (woman's)	grut^{ʲ}	grut^{ʲ}	ɣrut͡s^{ʲ}
liver	p^{ʲ}et͡ʃen^{ʲ}	p^{ʲ}et͡ʂen^{ʲ}	NA
skin	kozɻa	kozɻa	kozɻa
louse	voʂ	bl^{ɣ}oxa*	bl^{ɣ}axa*
dog	sobaka	sobaka	sabaka
fish (noun)	rɨba	rɨba	rɨba
horn (animal part)	rok*	rok*	rox*
tree	d^{ʲ}er^{ʲ}evo	d^{ʲ}er^{ʲ}eva	d͡z^{ʲ}er^{ʲ}eva
leaf	l^{ʲ}ist	l^{ʲ}ist	l^{ʲ}ist
person	t͡ʃelovek	t͡ʃelovek	t͡ʃelavek
name (noun)	im^{ʲ}a	im^{ʲ}a	im^{ʲ}a
sun	sol^{ɣ}nɨʂko**	son͡tse	son͡tse
star	zv^{ʲ}ozdot͡ʃka*/**	zv^{ʲ}ezda	zv^{ʲ}ezda
water	voda	voda	voda
fire	ogon^{ʲ}	ogon^{ʲ}	aɣon^{ʲ}
stone	kamen^{ʲ}	kamen^{ʲ}	kamen^{ʲ}
path	doroga	doroga	daroɣa
mountain	gora	gora	ɣara
night (dark time)	not͡ʃ	not͡ʃ	not͡ʂ
drink (verb)	p^{ʲ}it^{ʲ}	p^{ʲ}it^{ʲ}	p^{ʲ}it^{ʲ}
die	um^{ʲ}irat^{ʲ}	um^{ʲ}irat^{ʲ}	um^{ʲ}irat͡s^{ʲ}
see	v^{ʲ}id^{ʲ}et^{ʲ}	v^{ʲ}id^{ʲ}et^{ʲ}	v^{ʲ}id^{ʲ}et^{ʲ}
hear	sl^{ɣ}yʂat^{ʲ}	sl^{ɣ}yʂat^{ʲ}	sl^{ɣ}yʂat^{ʲ}
come	pr^{ʲ}ijt^{ʲ}i	pr^{ʲ}ijt^{ʲ}i	pr^{ʲ}ijt^{ʲ}it^{ʲ}
new	novɨj	novɨj	novɨj
full	pol^{ɣ}nɨj	pol^{ɣ}nɨj	pownɨj
one	od^{ʲ}in	od^{ʲ}in	ad^{ʲ}in
two	dva	dva	dva
I	ja	ja	ja
you	tɨ	tɨ	tɨ
we	mɨ	mɨ	mɨ

Table 2. Swadesh lists for Megra, Belogornoje and Khislavichi lects. The lists are separately published in an open-source repository[1].

[1] https://huggingface.co/datasets/djulian13/east-slavic-swadesh-lists

I gather these lists from raw corpora material. The words for which I was not able to find a word in its base form are marked with asterisk (*) in the table. In such cases, I manually transformed the word into its base form. I surmise its paradigm from the data I have gathered during my study of the corpus.

Gathering lists from raw corpora material, without possibility to access the lexicographical data, adds other complications, such as impossibility to reliably check whether a particular word is indeed the best match for a particular concept within the lect from historical point of view, or only a contextual one. This is the case of *blʲaxa*/*blʲoxa* 'louse (lit. flea)' in Khislavichi and Belogornoje, cf. example from Khislavichi: *Блохи . Клопы , меня уже что ?* 'Lice (lit. fleas). Bedbugs, [that bit] me already, what?' (Ryko & Spiricheva 2020). However, since both the Khislavichi and Belogornoje corpora contain the word in this meaning, rather than **voʂ* 'louse', as in Megra, the final dataset has no other choice than to include it.

The Megra Swadesh list includes diminutives (marked with a double asterisk) as the names of astronomical entities (*sun* and *star*). In the Megra lect, they are more frequent than their historically non-diminutive counterparts, more stylistically neutral (used in different context types) and semantically narrow, a relatively frequent phenomenon in Slavic languages[2]. One might also argue that they may be not as historically stable as non-diminutives. However, there is no sufficient data to support this claim. There are almost no data prior to 1980 on the Megra lect, and data from 1980 to 2019 consistently favour the hypothesis I present.

There is no attested word for *liver* in the Khislavichi corpus, henceforth *NA* value in the table.

The purpose of the second Swadesh list dataset is cross-evaluation on a set of lects that would be completely different both genetically and typologically. For these goals, I use three Tuu (Taa subgroup) lects from the Khoisan linguistic area: !Xóõ, N|u‖en and Kakia (Masarwa). For Taa, there is an existing wordlist, compiled and annotated by G. Starostin (2021; 2022) from previously published sources. Table 3 reproduces the list. Since Tuu is not the focus of this study, I give the wordlist itself for reference, but do not thoroughly discuss it, or the relationship between the lects themselves.

The last dataset I use is a set of phonetic features across East Slavic dialects taken from Marchenko & al. 2023. I treat features as sets $\{Realisation1, Realisation2, \dots, RealisationN\}_{feature}$ (Archangeli & Pulleybank 2022: 32). Each realisation receives a specific letter (I use A for the realisation I found first in the first analysed lect, B for the realisation I found first in the second analysed lect, and so forth). If there is no realisation of a feature in a lect, initially I insert an absence sign. After the lists of features for each lect are ready, the preprocessing turns them into a string (cf., *CCCC-------CCCCC--CCCCCCC-AC-C--CC-CC-CC-C-C-CCCC------A---C-BA--B--A-AA----AACCC-ACCCCC-A-AA---------AC-C-CBBBB--A-AAB---BC----CCC-----*).

There is no information of such kind for Khislavichi, Megra and Belogornoje. Yet there is information for the lects that seems to be of the same East Slavic continuum areas. I take Mokshenskaja (Northern Russian, Arkhangelsk Region, close relative of Megra), Piestchanka (Southern Russian, Saratov Region, close relative of Belogornoje), and Kritskovschina (Western Russian, Smolensk Region, close relative of Khislavichi). Picture 1 shows the geographical distribution of all lects.

[2] https://starlingdb.org/cgi-bin/response.cgi?root=new100&basename=new100\ier\slv.

Concept	!Xóõ	Kakia (Masarwa)	N|u‖en
eye	!'ǔĩ	‖x'wĩ	ǂ'ũ
ear	ǂùã^{h}	ǃwa:	ǂu-ša
nose	ʃù^{h}-ɲa	ʃu-ča	ʃu-ša
tongue	ʔ|nà^{ʕ}n	ʃa:n	|'a:ni
tooth	‖q^{h}ǎ:	‖xũ	‖'ʌn-te
hand	|x'à:	|x'a	|x'a
knee	‖xǘ:-ʃàn	‖õ-ʃaŋ	‖ũ ʃi
blood	ǃà:^{ʕ}	ǃa:^{ʕ}a	ǃa:^{ʕ}a
bone	ǂâ:	‖a:	ǂã
breast (woman's)	‖ú:	ǁam	‖u #
liver	ǁàm	*NA*	ǁʌm
skin	tù^{ʕ}m	t'üm	t'ùm
louse	θǘ:^{ʕ}	*NA*	*NA*
dog	ǂq^{h}ài	ǂxai	ǂ^{h}i
fish (noun)	*NA*	*NA*	*NA*
horn (animal part)	‖ãẽ	‖ʌn-ša	‖ã
tree	ʔθnàye	θoe:	θ'a:
leaf	‖āna	‖a:na	ʃabu
person	tâ:	tu	tu
name (noun)	|àũ	|x'ãũ	|ã
sun	‖'ân	‖'ʌn	‖'ʌn
star	‖ōna	‖wana-te	‖'ana-te
water	!q^{h}à:	!^{h}á	!^{h}a
fire	|'ǎ:	|'ã:	|'ã
stone	ʃū-le	ʃü-le	!'um
path	ǂólo	dau #	dau #
mountain	!ù^{h}m	ʃu:-n	‖um
night (dark time)	ʃúe^{ʕ}	ʃòe^{ʕ}	ʃòe
drink (verb)	x'ā̄:^{h}	x'ã	x'a-a
die	|'â:	|'a	|'a:
see	ʃâ	ʃa	ʃe:
hear	tǻ:^{ʕ}	tãa	sa
come	sî:	si	si
new	‖qu^{ʕ}V	‖xwe	*NA*
full	‖ù^{h}m	‖úm	‖um
one	ǂ'ǔã	!k'we	!'oe
two	ǂûm	ǃum	ǃum
I	n̄	n	ŋ
you	*NA*	*NA*	*NA*
we	ī^{h}	i	i

Table 3: Swadesh lists for !Xóõ, N|u‖en and Kakia (Masarwa). *NA* denotes concepts absent in the material.

Figure 1. East Slavic lects areas on the map of Eastern Europe.

Method

I implement a new method that is likely to match the definition of a black box more closely — though by no means perfectly. This method (which I further refer to as Cipher-RF, meaning Random Forest classifier of ciphered data) employs hashing algorithms, vectorisation and language contact emulation. The algorithm consists of four steps.

For the first step, the algorithm takes a word and then applies a hash function, transforming it into a string of a particular size. There are different hash functions, and there is no preferable one for my research. I use one of the most frequently implemented, SHA256. This step is essentially ciphering of the input for human understanding; the machine, on the other hand, still perceives input the same way and manages to decipher it back. The algorithm repeats the step for each given word in a dataset.

The second step aims at an irreversible change of input. I perform byte-pair encoding (BPE) tokenisation (Gage 1994) of a hash string. I use BPE tokenisation because hash strings are not likely to contain typical words of any given human language but may contain some common character n-grams — at least, numbers (Kanjirangat et al. 2023). I employ the GPT-2 tokeniser (Radford et al. 2019).

After the tokeniser transforms the input, I finish preprocessing by vectorising the acquired "token" arrays. At this stage, the connection between an original word and its new form completely breaks apart for a human. I use sci-kit-learn CountVectorizer (Pedregosa et al. 2011) to keep transformations simple, if not comprehensible.

In this fashion, the word *glʸaz* 'eye' (Megra/Belogornoje lects) becomes *4daa4a19c8a0858f4f4058b750854b5f79d8a9f2b81cce1bfc191bbe06f6b1b3* (after hash string transformation), then *4 da a 4 a 19 c 8 a 08 58 f 4 f 40 58 b 750 8 54 b 5 f 79 d 8 a 9 f 2 b 81 c ce 1 b fc 191 b be 06 f 6 b 1 b 3* (after tokenisation), and finally *[0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 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 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0]* after vectorisation.

The next step is to introduce a machine-learning method. I picked a Random Forest classifier (Ho 1995). It trains for a small amount of time and at the same time is one of the least explainable classical machine learning systems (Munn & Pitman 2022).

I train a Random Forest classifier for a language classification task and evaluate it. For evaluation I use the micro-F1 score as it is a widespread method of evaluating lect classification and identification (Kuparinen & Scherrer 2023). I count the micro-F1 score between 5-fold via cross-validation (as the Swadesh list dataset is small and a strict train/test split may significantly affect the result).

For each pair of lects, I swap some concepts to see how this distortion contributes to the classification result. I either swap a restricted number of random concepts between two lects or replace words from one lect with words from another lect, emulating an intense stream of borrowings. The first type of simulation results in a transfer like $a, b \rightarrow b$, a, the second — in a transfer like $a, b \rightarrow a, a$. Then I retrain and re-evaluate the classifier. I acquire a squared error between the original and the resulting F1-score. I repeat the process for a certain number of runs to reduce the element of randomness. I use UPGMA (Sokal & Michener 1958) to build genealogical trees based on acquired data, a tree for a run (for 100 runs I get 100 trees).

The groups under consideration are closely related, providing a similar evolution rate, and monophyletic, having a single ancestor (Ryko & Spiricheva 2020; Starostin 2021; Starostin 2022). Thus, UPGMA is preferable to NJ, which in rare cases gives a negative distance value between groups.

The resulting method is as close to a black box as possible. Table 4 compares Cipher-RF with a Levenshtein distance method by the criteria described in Section 1.

Criterion	Cipher-RF	Levenshtein distance
Transparency	Opaque	Transparent
Explainability	Unexplainable	Explainable
Object-awareness	Object-unaware	Object-aware
Reliability of transfer	Unreliably transferrable	Reliably transferrable
Predictability	Unpredictable	Predictable

Table 4. Comparison between Cipher-RF and Levenshtein distance.

I use two measurements for evaluation, the probability of a correct tree and the average split distance difference. As the datasets I use consist of only three lects, these metrics transform into correct outgroup identification probability and average inner split distance error.

Correct outgroup identification probability is a measure defined by the division of several runs when a model successfully predicted which lect was the most distant from the other two in the group, by an overall number of runs. Correct outgroup identification probability relies on existing classifications, and one should carefully apply it when the relationship between lects is debatable.

Average inner split distance loss is a measure that computes the difference between some pre-existing data on how early the two remaining lects split, and the results of the method under consideration. I score an average inner split distance only when the outgroup identification is correct. Average inner split distance loss requires another automatic language distance measurement method. This may cause some issues for the data that are traditionally problematic for automatic language distance measurement methods (Wichmann & Rama 2018), so one should apply the metric with extreme caution.

Cross-evaluation of the method includes three string similarity measures: two fully language-agnostic and one language-aware.

The first language-agnostic method I use is the Levenshtein distance normalised divided (LDND), a classic method for language distance measurement (Nerbonne & Heeringa 1997; Gooskens & Heeringa 2004; Holman et al. 2008). It consists of three steps. First, one scores additions, deletions, and substitutions (Levenshtein distance measurement) between two sequences. Then, one normalises them, dividing them by the size of the longest string in comparison. These two steps repeat for each concept in the Swadesh list, thus yielding a list of LDNs (Levenshtein distance normalised scores). After this, I score their mean value and thus acquire LDND.

The second implemented language-agnostic method is the weighted Jaro-Winkler distance normalised divided (WJWDND). Jaro-Winkler distance is a string similarity measure that resembles the Levenshtein distance, though it prioritises strings that have similar beginnings (Jaro 1989; Winkler 1990). The weighted Jaro-Winkler distance is a multiplication of the Jaro-Winkler distance by the Levenshtein distance, to get the best of two metrics (Gueddah et al. 2015). The normalisation and division parts remain the same as for LDND.

The language-aware method is the phonetics-aware Damerau-Levenshtein distance normalised divided (PADLDND). It works similar to LDND and WJWDND, with two key differences. It scores transpositions, which may be useful for metathesis detection, like Russian ***vsj****akʲij* — Croatian ***sv****akʲi* 'everyone among set'. PADLDND also multiplies each Damerau-Levenshtein distance score by a difference in the vectors of phonetic features of the symbols under consideration. There are different implementations of this metric with proven efficiency, but they are generally closed-source (Normanskaya 2020). I use the open-source implementation[3].

Data cross-evaluation tactics include using two different datasets of Swadesh list items, one of them consisting of understudied closely related lects and the other one of somewhat more distantly related, but relatively better studied ones.

The former, consisting of East Slavic Swadesh wordlists, is the main one. For this dataset, I use every method that I have mentioned in this section up to this point: Cipher-RF, LDND, WJWDND, and PADLDND.

The latter dataset consists of Tuu (Taa subgroup) Swadesh wordlists. This dataset allows me to test whether the efficiency of language-agnostic methods gets closer to the one of black-box methods with the language distance increase. Here I use only Cipher-RF, LDND, and WJWND. It feels safe — to some degree — to compare the automatic classification results for Taa with the current state-of-the-art human classification provided by Starostin (2022).

I also use a combined method and data cross-evaluation. This is a phonetics-aware Hamming distance (PAHD), a full-fledged alternative for a Swadesh list-based classification, based on calculating the Hamming distance between strings of phonetic features. PAHD does not analyse language units directly but deals with the results of human analysis.

The code that implements all the methods is available on GitHub[4].

Experiments and Analysis

The experiments run in four stages. I start by testing the black-box method via the correct outgroup identification measurement on the East Slavic material. The method is a Random Forest classifier of vectorised ciphered data (Cipher-RF). The next stage is to cross-evaluate via language-agnostic and language-aware string similarity measures. I also combine method and data cross-evaluation with the phonetics-aware Hamming distance (PAHD). The basis for comparison is the average inner split distance error. For the final stage, we repeat the first and the second stages on Taa lects.

Black-box method and East Slavic lects

As described in the section dedicated to methodology, using Cipher-RF includes 10 subsequent instances of lects re-classification after swapping concepts, which differ by the number of

[3] https://github.com/Stoneberry/phonetic-algorithmIPA.

[4] https://github.com/The-One-Who-Speaks-and-Depicts/black-box

swapped concepts (3, 5, 8, 11, 14), and the presence of borrowing processes emulation (when the concepts from one lect replace their counterparts in the second lect). Each setup, for statistical correctness, goes through 100 runs of random swaps. The results of the experiments are in Tables 5 and 6.

Number of swaps	3	5	8	11	14
Present borrowing	0.67	0.62	0.6	0.62	0.63
Non-present borrowing	0.69	0.62	0.52	0.5	0.48

Table 5. Correct outgroup identification probability for the East Slavic lects by Cipher-RF

Number of swaps	3	5	8	11	14
Present borrowing	0.001	0.001	0.002	0.002	0.002
Non-present borrowing	0.001	0.002	0.002	0.002	0.002

Table 6. Average correct inner split distance for East Slavic lects, Cipher-RF (only runs with correctly identified outgroup)

The correct outgroup identification here means that metric joins Belogornoje and Megra into a single group, while leaving Khislavichi as an outgroup, as presented in figure 2.

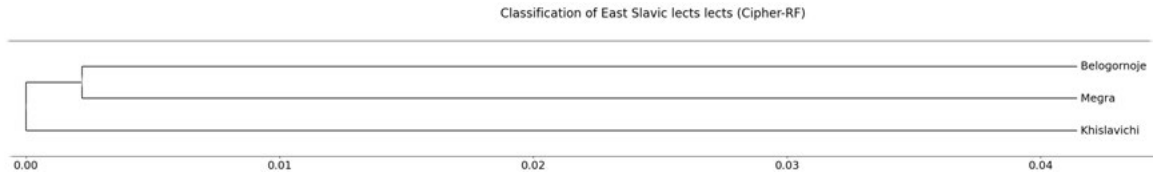

Figure 2. Correct classification of Belogornoje, Megra and Khislavichi with Cipher-RF.

While one may evaluate average inner split distance results only later, when compared to the other metrics, correct outgroup identification probability shows an overall low efficiency of Cipher-RF. Only seven setups out of 10 lead to significantly higher than a 50% chance of correct classification. There is no correlation between either the number of the swapped concepts or the presence of borrowings and the quality of Cipher-RF performance. It can be seen that the presence of borrowing makes the results more stable, even though lower for the lesser number of swaps. However, there is no way to explain it, as Cipher-RF is a black box. One just does not know what drives it and which data transformations are the most crucial.

String similarity measures and East Slavic lects

The next step is to test string similarity measures against the same dataset. The Levenshtein distance normalised divided (LDND)-based classification and the weighted Jaro-Winkler distance normalised divided (WJWDND)-based classification are in figures 3 and 4 respectively.

Both language-agnostic string similarity measures demonstrate the correct outgroup prediction, placing Khislavichi quite far from the last common ancestor (LCA; Brower & Schuh 2021) of Megra and Belogornoje. One can also see the difference in the inner split: it is 0.039 to 0.04 for LDND and 0.031 to 0.032 for WJWDND. For Cipher-RF this value is generally almost ten times lower. The comparison is shown in Tables 7 and 8.

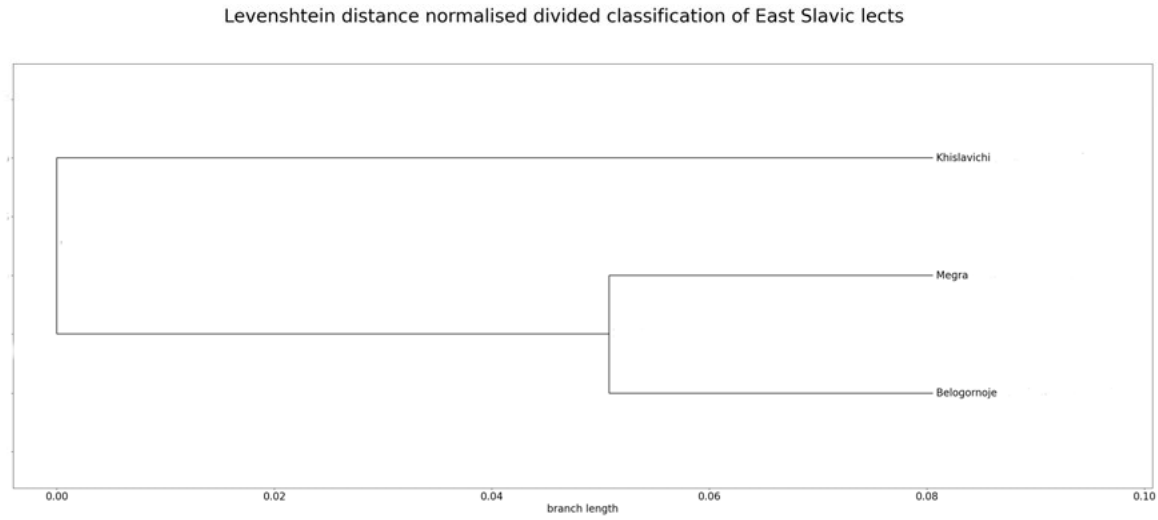

Figure 3. Classification of Belogornoje, Megra and Khislavichi with LDND.

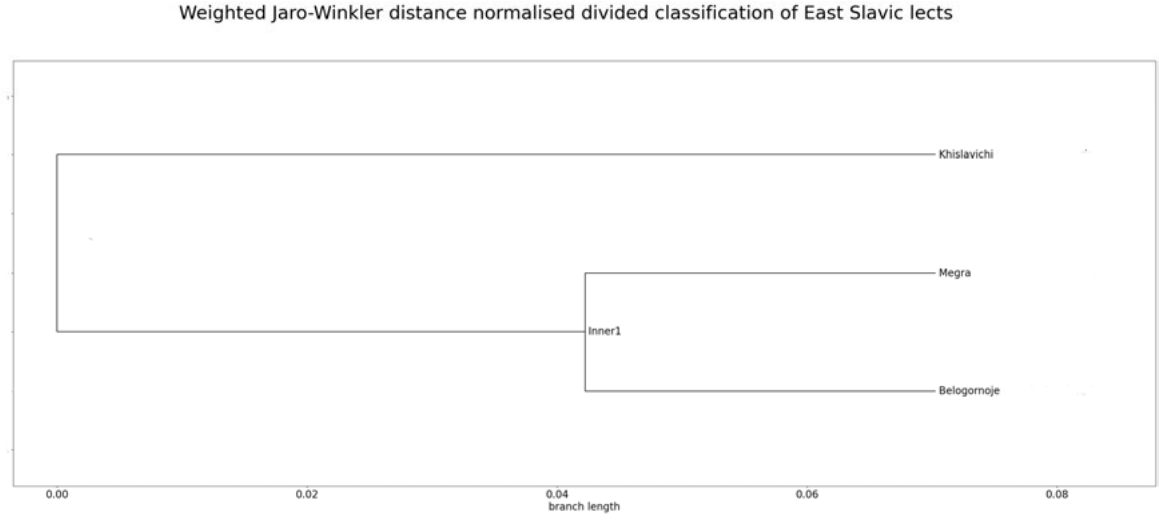

Figure 4. Classification of Belogornoje, Megra and Khislavichi with WJWDND.

Number of swaps	3	5	8	11	14
Present borrowing	0.04	0.04	0.039	0.039	0.039
Non-present borrowing	0.04	0.039	0.039	0.039	0.039

Table 7. Average inner split distance loss between Cipher-RF (only runs with correctly identified outgroup) and LDND.

Number of swaps	3	5	8	11	14
Present borrowing	0.032	0.032	0.032	0.031	0.032
Non-present borrowing	0.032	0.032	0.032	0.031	0.032

Table 8. Average inner split distance loss between Cipher-RF (only runs with correctly identified outgroup) and WJWDND.

For the difference in scores between Cipher-RF and string similarity measures, there are two different reasons. The first is the difference of scale: LDND and WJWDND may have values of 0 to 1, while Cipher-RF possesses a restriction of the squared difference between the gold score and the set up score (in this series of experiments, it is approximately 0.45). The second reason is that the distance between the LCA of Khislavichi, Belogornoje and Megra, and the LCA of Belogornoje and Megra is lesser for Cipher-RF classification than for LDND and WJWDND classification.

Method cross-evaluation: language-aware string similarity measure and East Slavic lects

Figure 5 shows the results of experiments based on a language-aware string similarity measure.

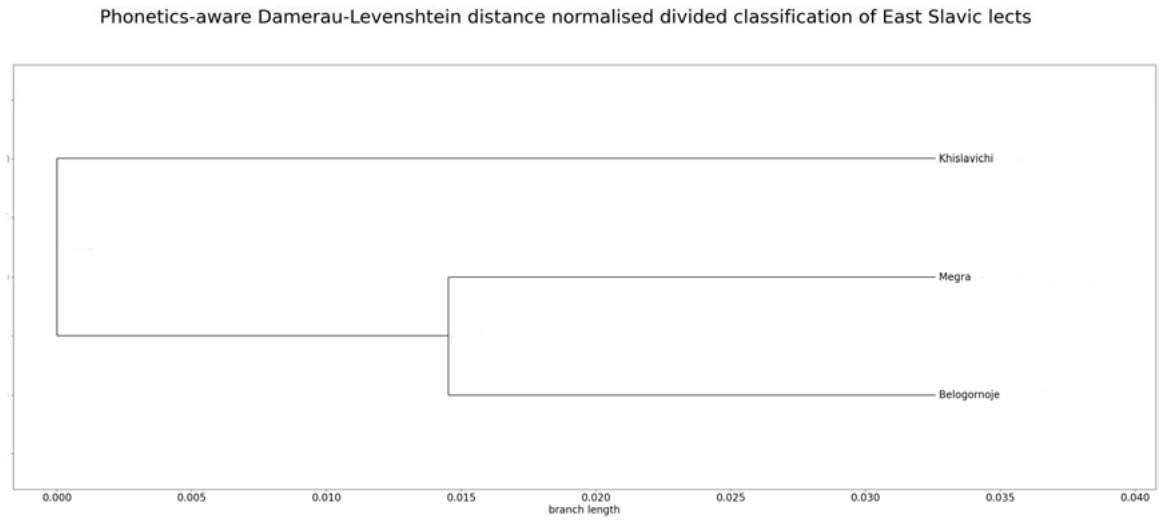

Figure 5. Classification of Belogornoje, Megra and Khislavichi with PADLDND.

The phonetics-aware string similarity measure demonstrates correct outgroup prediction in the same manner as the language-agnostic string similarity measures do. However, there are two differences. The first one is branch length: it is approximately two times smaller in LDND and WJWND than in PADLDND. The second difference is more linguistically explainable. PADLDND considers Megra and Belogornoje to be much more closely related to each other than LDND and WJWDND both do. This correlates with coincidences and disagreements of some East Slavic phonetic phenomena manifestations between lects (strong presence of *okanje* in both Megra and Belogornoje, contrasted with *akanje* in Khislavichi), and proves that language awareness indeed helps string similarity measures to be more precise. However, the average inner split distance loss between PADLDND (0.011) and Cipher-RF is still significant, as one may see in Table 9.

Number of swaps	3	5	8	11	14
Present borrowing	0.012	0.011	0.011	0.011	0.011
Non-present borrowing	0.011	0.011	0.011	0.011	0.011

Table 9. Average inner split distance loss between Cipher-RF (only runs with correctly identified outgroup) and PADLDND.

Overall, these experiments show that there is a crucial difference between black-box and language-agnostic methods. Language-agnostic methods are much closer in their results to the language-aware methods of the same type, though they lack some necessary linguistic insight. However, their advantages and shortcomings are easily explainable, and any researcher with sufficient skills may attempt to maximise the former and minimise the latter by introducing language-aware features.

Data cross-evaluation: the Taa lects

The question, however, remains: does this difference prevail on a bigger diachronic scale? I took the Taa lects as a representative of somewhat more distantly related lects. Figure 6 reproduces a classification of Taa lects by Starostin (2021; 2022)[5].

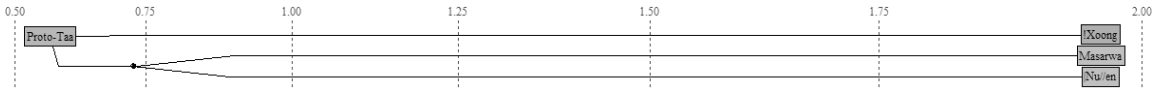

Figure 6. Classification of Taa lects.

The results of measuring language distance between Taa lects with Cipher-RF are in Tables 9 and 10.

Number of swaps	3	5	8	11	14
Present borrowing	0.44	0.43	0.32	0.41	0.36
Non-present borrowing	0.34	0.35	0.31	0.39	0.26

Table 9. Correct outgroup identification probability for the Taa lects by Cipher-RF.

Number of swaps	3	5	8	11	14
Present borrowing	0.005	0.006	0.006	0.005	0.006
Non-present borrowing	0.005	0.006	0.006	0.006	0.007

Table 10. Average correct inner split distance for the Taa lects, Cipher-RF (only runs with correctly identified outgroup).

Thus, Cipher-RF results are becoming worse with a language distance increase between the analysed lects. There are at least some (though unknown) pieces of linguistic information that influence the model, even after all the data transformations.

The average inner split distance is the same, which means that Cipher-RF implementation blocks our attempts to transfer language distance information into a precise absolute timing (i.e., how many years ago) of the split. As Figure 7 shows, the branch length for Taa is the same as for East Slavic. It may seem that the black-box method-produced graph is closer to the gold one than the language-agnostic string similarity measures-produced one. Cipher-RF, however, detects much fewer differences between lects than any other method.

Language-agnostic string similarity measures predictions remain correct. The length of branches also grows (0.087 and 0.078 for LDND and WJWDND), depending on the time passed, as is visible in Figures 8 and 9. It makes these methods more suitable for diachronic studies of language variation.

[5] Image taken from https://starlingdb.org/images/xoo.png.

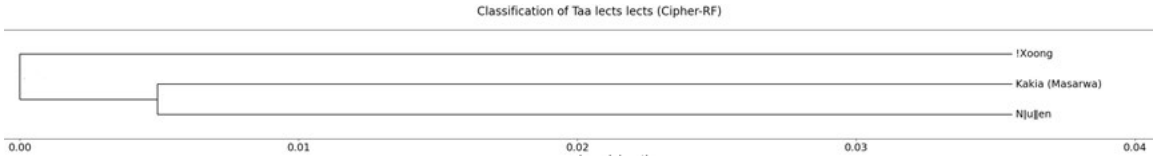

Figure 7. Correct classification of !Xóõ, Kakia (Masarwa) and N|u‖en with Cipher-RF.

The bigger distance also highlights the difference between LDND and WJWDND. LDND augments the differences between the lects under consideration, while WJWDND tends to smooth it by normalisation. It helps to distinguish the possible spheres of application for both methods: LDND for comparing closely related lects and WJWDND — for the distantly related ones.

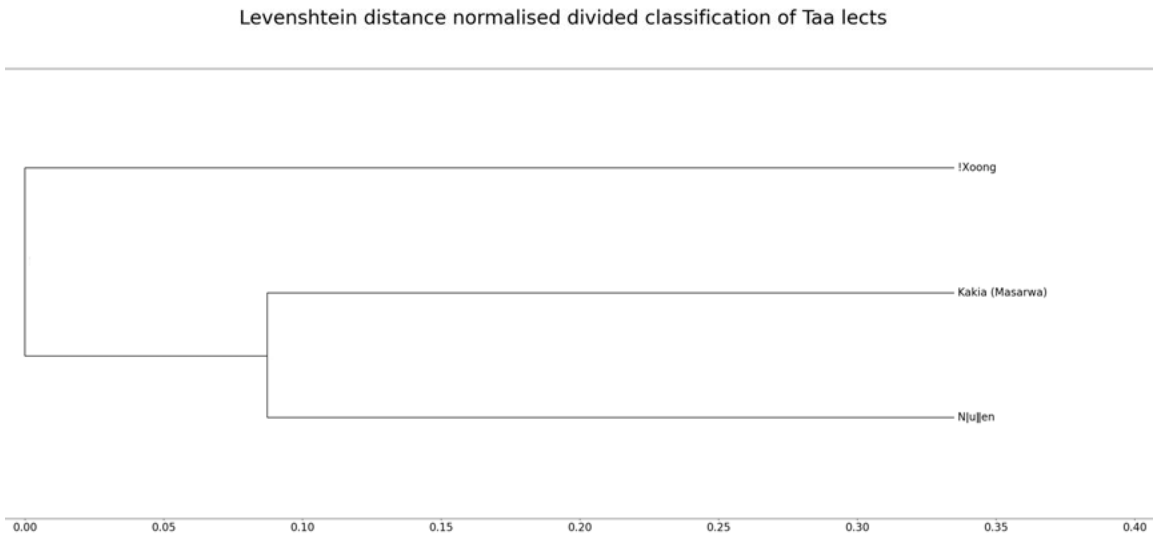

Figure 8. Classification of !Xóõ, Kakia (Masarwa) and N|u‖en with LDND.

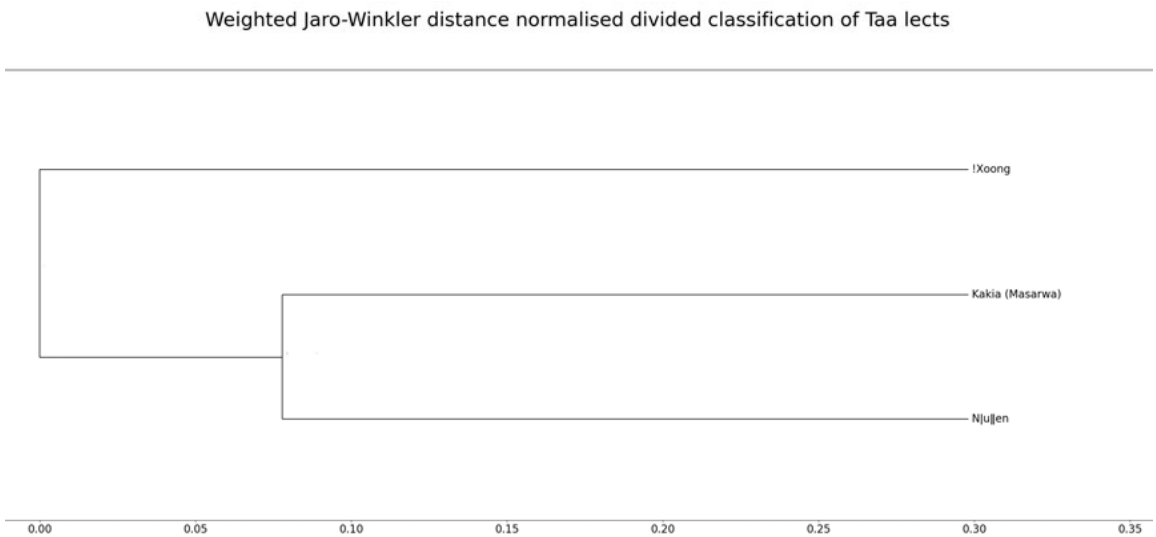

Figure 9. Classification of !Xóõ, Kakia (Masarwa) and N|u‖en with WJWDND.

Language-agnostic string similarity measures, yet again, demonstrate a greater ability to deal with language distance than Cipher-RF. The average inner split distance loss becomes more visible, as seen in Tables 11 and 12.

Number of swaps	3	5	8	11	14
Present borrowing	0.082	0.082	0.082	0.082	0.082
Non-present borrowing	0.082	0.081	0.081	0.081	0.081

Table 11. Average inner split distance loss between Cipher-RF (only runs with correctly identified outgroup) and LDND for Taa lects classification.

Number of swaps	3	5	8	11	14
Present borrowing	0.072	0.072	0.072	0.072	0.072
Non-present borrowing	0.072	0.072	0.072	0.072	0.071

Table 12. Average inner split distance loss between Cipher-RF (only runs with correctly identified outgroup) and WJWDND for the Taa lects classification.

Thus, the behaviour and efficiency of black-box and language-agnostic methods differ. While language-agnostic methods are still sensitive to the degree of language distance under consideration, black-box methods do not provide a researcher with any additional information on diachronic scope.

Method and data cross-evaluation: phonetically aware Hamming distance

PAHD experiments (Figure 10) include scoring the Hamming distance on strings of phonetic features for Kritskovschina, Mokshenskaja and Piestchanka East Slavic lects. It is important to note that these lects are not direct representations of respective Khislavichi, Megra and Belogornoje phonetic states (Kritskovschina and Khislavichi belong to different, though phonetically close, dialect continua). Thus, the comparison is approximate.

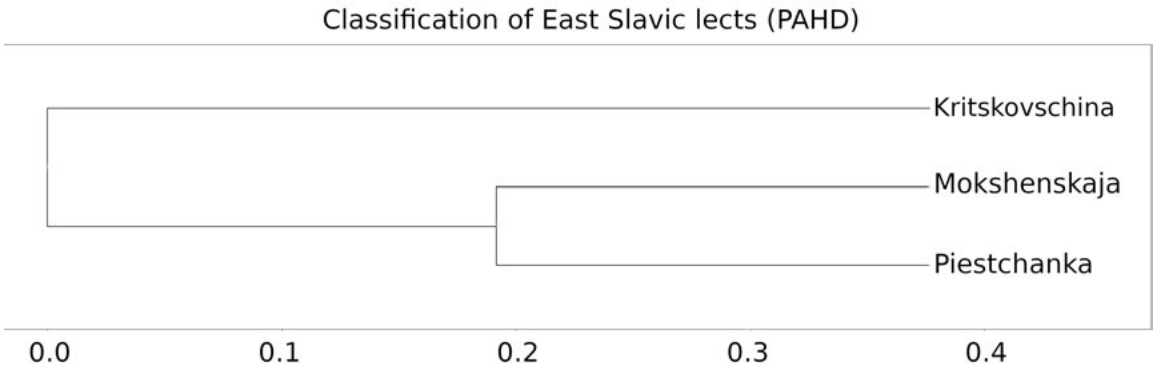

Figure 10. Classification of Kritskovschina, Mokshenskaja and Piestchanka with PAHD.

The difference here is probably the most drastic one, as table 13 shows.

Number of swaps	3	5	8	11	14
Present borrowing	0.191	0.19	0.19	0.19	0.19
Non-present borrowing	0.19	0.19	0.19	0.19	0.19

Table 13. Average inner split distance loss between Cipher-RF (only runs with correctly identified outgroup) and PAHD for East Slavic lects classification.

PAHD separates the lects more strictly by a huge margin, the distances it calculates are larger than the ones of LDND and WJWDND classifications for Taa. Using only phonetic features, thus, magnifies the scale of differences by almost ten times.

This experiment again proves that Cipher-RF is the method that is unlikely to provide us with linguistic insight (or any insight, for that matter) on its decisions. It also shows that restricting the data to historically stable vocabulary smoothes out phonetic differences between lects, making methods more sustainable on a large scale while somewhat harming their sensitivity on a small scale.

PAHD agrees with LDND, WJWDND, and PADLDND on the division between Northern Belarusian/Western Russian and Northern/Southern Russian lects. It confirms the original gold presupposition of Khislavichi being an outgroup for Belogornoje and Megra, and thus reiterates the correctness of Cipher-RF evaluation.

Conclusion

This research presents a new black-box method for historical comparative linguistics, Cipher-RF, based on a combination of hashing algorithms and a Random Forest Classifier. I have applied this method to East Slavic and Taa lects and showed that language-agnostic methods and black-box methods significantly differ in their behaviour and that it is crucial to distinguish between them, contrary to past papers on the topic, such as Prokić and Moran 2013. Black-box methods are less efficient for the purposes of historical comparative linguistics, but they provide a good baseline for other automatic methods to beat while being compared to an ideal classification. One should still use language-agnostic methods with a high degree of caution and always try to interpret their results linguistically. Their transparent structure allows for that.

The paper introduces a new corpus-based dataset of Swadesh-type lists for East Slavic lects of Khislavichi, Megra, and Belogornoje. The dataset consists of lexical units gathered from both open and unpublished corpora. The actual type of Swadesh list is ASJP; in the future, it is going to be expanded into a more classical 110-item one (Kassian et al. 2010).

The next step could be the introduction of new lects and new classification methods, especially when the material is presented by raw corpora. One more possible expansion is the automatic search for Swadesh list items in raw corpora.

Acknowledgements

The author owes his reviewers for their insightful comments and discussion of the material. All the remaining errata are author's. The author is also grateful to the compilers of Saratov dialectological corpus, Khislavichi corpus, Russian dialectology atlas, and Global Lexicostatistical Database, without whom access to the research material would be impossible.

References

Afanasev, Ilia. 2023. The Use of Khislavichi Lect Morphological Tagging to Determine its Position in the East Slavic Group. In: *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*: 174–186.

Archangeli, Diana, Douglas Pulleyblank. 2022. *Emergent phonology*. Berlin: Language Science Press.

Barannikova, Lidija. 2005. Govory territorij pozdnego zaselenija i problema ih klassifikacii. In: Valentin Goldin, Olga Kryuchkova (eds.). *Barannikova L. I. Obshhee jazykoznanie: izbrannye raboty:* 192–203. Moscow: KomKniga.

Bastings, Jasmijn, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, Sarah Wiegreffe. 2022. In: *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

Barić, Ana, Laura Majer, David Dukić, Marijana Grbeša-Zenzerović, Jan Snajder. 2023. Target Two Birds With One SToNe: Entity-Level Sentiment and Tone Analysis in Croatian News Headlines. In: *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*: 78–85.

Brower, Andrew V. Z., Randall T. Schuh. 2021. *Biological Systematics: Principles and Applications*. Ithaca, NY: Cornell University Press.

Burlak, Svetlana. 2021. Ustoichivost' i chastonost': jest' li korrel'acija? [Stability and frequency: is there a correlation?] *Journal of Language Relationship* 19(3-4): 293–307.

Carvalho, Fernando O. de. 2020. Evaluation of cognation judgments undermines computational phylogeny of the Arawakan language family. *Journal of Language Relationship* 18(1-2): 87–110.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP):* 1724–1734.

Gage, Philip. 1994. A New Algorithm for Data Compression. *The C User Journal* 12(2): 23–38.

Gooskens, Charlotte, Wilbert Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16: 189–207.

Gueddah, Hicham, Abdellah Yousfi, Mostafa Belkasmi. 2015. The filtered combination of the weighted edit distance and the Jaro-Winkler distance to improve spellchecking Arabic texts. In: *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA):* 1–6.

Feld, Jan, Alexander Maxwell. 2019. Sampling error in lexicostatistical measurements: A Slavic case study. *Diachronica* 36(1): 100–120.

Ho, Tin Kam. 1995. Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition:* 278–282.

Holman, Eric, Søren Wichmann, Cecil Brown, Viveka Velupillai, André Müller, Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica* 42: 331–354.

Jaro, Matthew A. 1989. Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. *Journal of the American Statistical Association* 84: 414–420.

Jäger, Gerhard. 2019. Computational historical linguistics. *Theoretical Linguistics* 45(3-4): 151–182.

Kanjirangat, Vani, Tanja Samardžić, Ljiljana Dolamic, Fabio Rinaldi. 2023. Optimizing the Size of Subword Vocabularies in Dialect Classification. In: *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023):* 14–30.

Kassian, Alexei, George Starostin, Anna Dybo, Vasiliy Chernov. 2010. The Swadesh wordlist. An attempt at semantic specification. *Journal of Language Relationship* 16: 46–89.

Kryuchkova, Olga, Valentin Goldin. 2011. Corpus of Russian dialect speech: concept and parameters of evaluation. In: *Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2011"*: 359–367.

Kuparinen, Olli, Yves Scherrer. 2023. Dialect Representation Learning with Neural Dialect-to-Standard Normalization. In: *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*: 200–212.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8): 707–710.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics:* 7871–7880.

List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. Walter de Gruyter GmbH & Co KG.

Marchenko, I. A., O. N. Dolgov, A. S. Azanova, M. S. Zambrzhitskaya, E. A. Zalivina, S. A. Zemlyanskaya, D. I. Mochul'skij, E. I. Tsejtina, D. G. Chistyakova, R. V. Ron'ko. Database of the Dialectological Atlas of the Russian Language. Available online at: https://da.ruslang.ru/ [accessed: 18.10.2023].

Munn, Michael, David Pitman. 2022. *Explainable AI for Practitioners*. O'Reilly Media, Inc.

Nerbonne, John, Wilbert Heeringa. 1997. Measuring Dialect Distance Phonetically. *Proceedings of the EACL 1997:* 1–18.

Normanskaya, Julija V. 2020. Komi-jaz'vinskij — dialekt komi-permjackogo ili otdel'nyj jazyk? *Ezhegodnik finno-ugorskih issledovanij* 4: 628–641.

Novikova, Jekaterina, Ondřej Dušek, Amanda Cercas Curry, Verena Rieser. 2017. Why We Need New Evaluation Metrics for NLG. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*: 2241–2252.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever. 2019. *Language Models are Unsupervised Multitask Learners*. Available at: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [Accessed: 18.10.2023].

Rama, Taraka, Lars Borin. 2015. Comparative Evaluation of String Similarity Measures for Automatic Language Classification. In: G. K. Mikros, J. Macutek (eds.). *Sequences in Language and Text:* 171–200. Berlin / München / Boston: De Gruyter Mouton.

Ryko, Anastasiya I., Margarita S. Spiricheva. 2020. Corpus of the Russian dialect spoken in Khislavichi district. Available online at: http://lingconlab.ru/khislavichi/ [accessed: 18.10.2023].

Ryko, Anastasiya I., Margarita S. 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.

Snoek, Connor. 2013. Using semantically restricted word-lists to investigate relationships among Athapaskan languages. In: Lars Borin, Aniu Saxena (eds.). *Approaches to Measuring Linguistic Differences*: 231–248. Boston/Berlin: Walter De Gruyter GmbH.

Sokal, Robert Reuven, Charles Duncan Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38:1409–1438.

Starostin, George. 2011. On Mimi. *Journal of Language Relationship* 6(1): 115–140.

Starostin, George. 2021. Lexicostatistical studies in Khoisan II/1: How to make a Swadesh wordlist for Proto-Tuu. *Journal of Language Relationship* 19(1-2): 99–135.

Starostin, George. 2022. Lexicostatistical studies in Khoisan II/2: Towards a more precise phylogeny for the Tuu family. *Journal of Language Relationship* 20(1-2): 25–70.

Sutskever, Ilya, Orial Vinyals, Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems:* 3104–3112.

Syvokon, Oleksiy, Olena Nahorna, Pavlo Kuchmiichuk, Nastasiia Osidach. 2023. UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language. In: *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP):* 96–102. l Linguistics.

Vajda, Edward. 2012. The Dene-Yeniseian connection: a reply to G. Starostin. *Journal of Language Relationship* 8(1): 138–152.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. 2017. Attention Is All You Need. *Advances in neural information processing systems* 30: 1–11.

Wichmann, Søren, Eric W. Holman, Dietrich Stauffer, Cecil H. Brown. 2011. Similarities among languages of the Americas: An exploration of the WALS evidence. *Journal of Language Relationship* 5: 130–134.

Wichmann, Søren, Taraka Rama. 2018. Jackknifing the Black Sheep: ASJP Classification Performance and Austronesian. In: Ritsuko Kikusawa, Lawrence A. Reid (eds.). *Let's Talk about Trees: Genetic Relationships of Languages and Their Phylogenic Representation:* 39–58. Osaka: National Museum of Ethnology, Japan.

Winkler, William E. 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In: *Proceedings of the Section on Survey Research Methods*: 354–359.

Zhivlov, Mikhail. 2021. Does the comparative method work in New Caledonia? *Proceedings of the 15th International Conference on Austronesian Linguistics, Olomouc*. Available online at: https://www.hse.ru/data/2021/07/04/432013048/Comparative%20 method.pdf [accessed: 01.09.2023].

И. Афанасьев. Потенциал систем — «чёрных ящиков» для автоматического измерения расстояния между восточнославянскими лектами.

Активное развитие новых квантитативных методов в современной исторической лингвистике в 2000-е — 2010-е годы актуализировало проблему невозможности адек-

ватной интерпретации подобных методов и поиск путей преодоления данной проблемы. Задачей данной статьи является демонстрация преимуществ систем, которые по умолчанию обладают прозрачностью для исследователя.

В работе сравниваются два типа систем, измеряющих языковое расстояние и используемых для задач внутренней генетической классификации. Механизм действия систем — «чёрных ящиков» предполагает обработку исходных данных и представление результата максимально непрозрачным как для исследователя, так и для автоматических методов анализа. Напротив, независимые от частных языковых свойств методы (к примеру, меры сходства строк) анализируют данные прозрачным образом, но не учитывают особенностей конкретных языков. Для сравнения систем — «чёрных ящиков» с существующими независимыми от частных языковых свойств методами в данной статье предлагается новая непрозрачная система, основанная на хешировании, векторизации и имитации языкового контакта.

В статье использован восточнославянский материал (лексический и грамматический), а также материал группы таа (койсанский языковой ареал Южной Африки). Восточнославянские данные состоят из корпусов говоров с. Белогорное, д. Мегра и с. Хиславичи, а также списков фонетических особенностей говоров д. Мокшенская, д. Крицковщина и с. Песчанка. Данные таа представлены списками Сводеша для къхонг, масарва и н|у‖ен. Важным вкладом работы является публикация новых списков Сводеша для ряда восточнославянских диалектов.

Ключевые слова: «чёрный ящик»; восточнославянские языки; южнокойсанские языки; языки туу; независимые от частных языковых свойств методы; автоматическое измерение языковой дистанции; автоматическая классификация; меры сходства строк; базисная лексика.