Master Thesis Proposal:
Phonetic and Information-Theoretic Distance Measures for Automated Linguistic Phylogenetic Inference

Philip Allan Georgis, 2576684

August 3, 2021

Abstract:

Recent decades have seen a wave of advances in automated methods for linguistic phylogenetic inference. Modern techniques are capable of automatically detecting cognate word forms and computing likely phylogenetic trees with relative accuracy, representing a significant reduction in labor compared to the manual identification of sound correspondences and potential cognates via the comparative method. Although most frequently a character-based approach is adopted, distance-based methods have increasingly been applied for phylogenetic inference as well. This thesis proposes the development and comparison of several lexical distance measures grounded in principles of phonology and information theory, to evaluate their applicability for tasks in distance-based computational historical linguistics, including automated cognate detection and inference of phylogenetic tree structures.

1. Introduction

For over a century, the comparative method was the primary technique for inferring the historical development of groups of languages from a common source (Campbell, 2013). Possibly related word forms from different languages were systematically compared with one another to identify consistent sound correspondences. The ancestral word forms from which they derived could then be reconstructed, attesting to the languages' genetic relationship. The application of the comparative method has allowed historical linguists to trace the origins of many of the world's language families. A written record of the ancestral language exists for certain taxa – such as Latin and Sanskrit, respectively, for the modern Romance and Indo-Aryan languages – thus, enabling linguists to confirm the accuracy of the comparative method.

Although effective, the comparative method is extremely laborious and time-consuming. Computational approaches using probabilistic and statistical models have been developed to approximate this process, enabling the automatic detection of cognate word forms. Character-based phylogenetic methods such as maximum likelihood and maximum parsimony then compute the phylogenetic tree (e.g. Figure 1) which best explains the observed distribution of cognates according to an evolutionary model, thus inferring the historical development of language families (Pagel, 2017). However, these approaches likewise remain very computationally expensive, requiring the evaluation of vast numbers of possible phylogenies before arriving at the most likely tree topology. In addition, such methods disregard a large amount of data as they only consider the presence or absence of related forms, rather than assessing the degree of similarity among them.

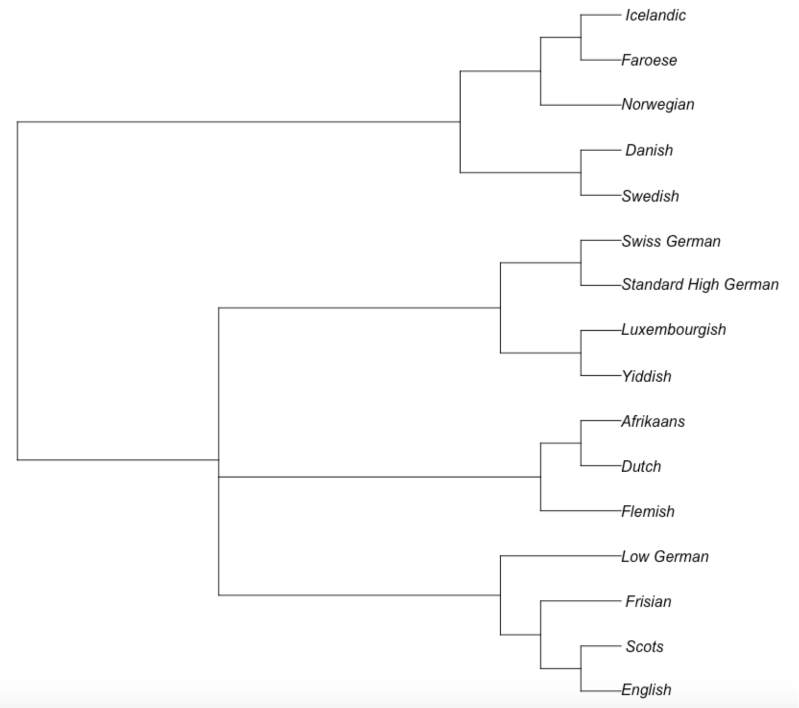

Figure 1. Phylogenetic tree of modern Germanic varieties (Hammarström et al., 2021)

Distance-based methods offer an alternative, more efficient approach. Instead of computing the likeliest tree structure according to an evolutionary model, languages can be clustered into a tree-like structure by their similarity to one another according to selected criteria (Wichmann et al., 2010). Evidence from dialectological studies suggests that both phonetic proximity and predictability of word forms are strong correlates of linguistic distance among related languages (Gooskens, 2007; Moberg et al., 2007; Stenger et al., 2017). Given that the comparative method is also based upon these two principles, phonetic proximity and predictability stand out as suitable criteria for clustering languages in distance-based phylogenetic inference.

Lexical distances calculated by evaluating word pairs for surface similarity or consistency of sound correspondences have been successfully employed for computational historical linguistics tasks such as automatic cognate detection (e.g. Kondrak, 2001; List & Forkel, 2021) and phylogenetic tree building (e.g. Kessler, 1995; Wichmann et al., 2010; Jäger, 2018). Thus far, however, few studies have delved beyond linguistically naïve string comparison techniques such as the Levenshtein distance for measuring the phonetic distance between word forms. Many techniques additionally rely on abstract sound classes, which preclude a more precise measurement of phonetic distance, or else use somewhat arbitrary feature values without a clear foundation in phonological theory. Information theoretic measures of predictability such as word-adaptation surprisal have not yet been applied for phylogenetic inference tasks. This thesis, thus, proposes the development, evaluation, and comparison of several phonetic and information theory-based distance measures to assess their suitability for distance-based linguistic phylogenetic inference tasks including automatic cognate detection and tree building.

2. Background and Motivation

The traditional approach in linguistic phylogenetic inference involves the binary coding of lexical data in the form of glossed wordlists into discrete cognate classes. That is, given a pair of languages, all pairs of words for the same concept are labeled as either cognate (related) or non-cognate (unrelated) words. Word pairs are deemed to be related if they purportedly derive from the same word in a common ancestral language. For example, the English word *eye* and the German word *Auge* would be labeled as cognate as they both derive from the Proto-Germanic word **augon*. The Dutch *oog* and Swedish *öga* also belong to this same cognate class, and thus these four words form a cognate set for the concept EYE. The English *tree* and the German *Baum*, however, derive from separate Germanic roots and would, thus, be labeled as non-cognates to one another, although they are each cognate, respectively, with the Swedish and Dutch equivalents. Cognate classes may also be composed of a single word form, as in the case of English *dog*. Table 1 shows an example of cognate set coding for three concepts into five distinct cognate classes in these four Germanic languages.

Table 1. Example of cognate set coding for three lexical items in four Germanic languages.

	English	Dutch	German	Swedish
EYE-1	*eye*	*oog*	*Auge*	*öga*
TREE-1	*tree*			*träd*
TREE-2		*boom*	*Baum*	
DOG-1		*hond*	*Hund*	*hund*
DOG-2	*dog*			

This method of binary cognate coding into discrete cognate classes is a character-based method, which directly models the evolutionary process of descent with modification according to a set of discrete features, or “characters” (Rama & Kolachina, 2013; Jäger, 2018). In this case, the characters are a finite number of cognate classes, whose values are binarily taken from the presence (1) or absence (0) of each cognate class in the language. An optimal phylogenetic tree is generated from these data by maximizing the likelihood of the distribution of observed feature states according to a particular evolutionary model (Pagel, 2017). Important to note is that character-based methods using cognate classes as features effectively disregard all variation in the phonetic form of the words within a cognate class and treat all members as equally related to one another, discarding a significant amount of data in the process.

Distance-based methods instead calculate a continuous distance measure for each language pair according to various criteria, which are fed into a clustering algorithm (Wichmann et al., 2010; Rama & Kolachina, 2013). Pairwise distance matrices are used to generate dendrograms through agglomerative clustering, whereby the most similar language pairs are combined into clusters, and clusters are iteratively merged into larger subgroupings (Kessler, 1995). Unlike character-based methods, distance-based methods do not directly model the evolutionary process, and the resulting dendrograms, thus, reflect similarity rather than genetic relatedness (Jäger, 2018). However, in many cases there is a significant overlap between similarity and relatedness, and distance-based methods have the advantage of being far more computationally efficient.

A variety of linguistic distance measures have been proposed for different linguistic domains, including the lexicon, phonology, or morphosyntax. For example, Séguy (1973) derived distances between Gascon dialects using a combination of binary lexical, phonetic, and morphological features; Kessler (1995) measured distances between Irish Gaelic dialects using phonetic transcriptions of cognates; Gamallo et al. (2017) used perplexity of character *n*-gram language models measured on corpora to determine distances between European languages; and Heeringa et al. (2018) proposed methods for measuring syntactic distances between Germanic languages using aligned sentences of parallel corpora.

In his 2018 global phylogenetic inference study, Jäger found that a lexical distance measure calculated using pointwise mutual information (PMI) between ASJP[1] sound classes produced more accurate trees than the traditional character-based approach for small and medium-sized families. On larger families the character-based inference method using maximum likelihood slightly outperformed the PMI distance, but the difference was not significant. These results are quite promising and suggest that distance-based methods using information theoretic distances may indeed prove useful for automated linguistic phylogenetic inference. It is worth exploring whether similar information theoretical measures such as word-adaptation surprisal (Stenger et al., 2017) might likewise serve as effective distance measures for phylogenetic inference.

Table 2. Example of binary meaning-ASJP sound class coding for EYE in four Germanic languages.

	English *<eye>*	**Dutch** *<oog>*	**German** *<Auge>*	**Swedish** *<öga>*
IPA Transcription	/aɪ̯/	/oːχ/	/aʊ̯gə/	/øːga/
ASJP Transcription[2]	ai	oX	aug3	ega
EYE-a	1	0	1	1
EYE-i	1	0	0	0
EYE-o	0	1	0	0
EYE-X	0	1	0	0
EYE-u	0	0	1	0
EYE-g	0	0	1	1
EYE-3	0	0	1	0
EYE-e	0	0	0	1

There is also evidence that the phonetic form of words carries a phylogenetic signal and should not be disregarded. Jäger (2018) introduced a second character type combining both meaning and sound classes as a method for capturing differences in the phonetic forms of words within the same cognate class. This method performed even better than the PMI distance and achieved the best fit with the gold standard Glottolog trees. Using the previous Germanic example, Table 2 shows the resulting binary coding according to these meaning/sound class characters for the concept EYE.

[1] Automated Similarity Judgment Program (Wichmann et al., 2020).

[2] These ASJP transcriptions were yielded from the ipa2asjp() function of the asjp Python library (Sofroniev, 2018).

The feature 'EYE-a', for example, would be coded as 1 if the language's word for EYE contains the ASJP sound class <a> and 0 otherwise.

However, a character-based approach to phonetic features may not be the most appropriate. This method recognizes only exact matches in sound classes, thus failing to evaluate the similarity of Swedish /øː/ to Dutch /oː/. The reduction to sound classes is also problematic because it often erases language-specific distinctions in phonemes (e.g. length, palatalization, aspiration) which may have separate sound correspondences in sister languages. Additionally, this method only checks whether the sound class is present *anywhere* in the word. Thus, English, German, and Swedish are all coded as 1 for the feature 'EYE-a', even though the matched /a/ refers to the first vowel in English and German, but to the final vowel in Swedish – a coincidence. Wichmann et al.'s (2010) method of measuring distances between ASJP-transcribed word forms with the Levenshtein algorithm does not exhibit this final flaw, but still suffers from the others. The Levenshtein distance (Levenshtein, 1966) considers substitutions of all non-identical characters equally different from one other, meaning that similar sounds transcribed with different IPA symbols are treated as differently as completely unrelated sounds. A desideratum would, thus, be a more phonologically informed string distance which can be applied directly to IPA transcriptions and yield more accurate measures of similarity between word pairs for use in phylogenetic inference tasks.

3. Methods

Two main types of distance measures will be implemented, namely those grounded in phonetics and phonology and those based in information theoretic PMI and surprisal.

a. Phonetic distance

Various methods of phonetic similarity or distance have been proposed in previous work, and some have been applied in the field of dialectology, but seldom for phylogenetic inference. This is largely because phonetic proximity of word forms alone is not generally a sufficient criterion for establishing cognacy in historical linguistics, as words may resemble one another due to chance or borrowing (Kessler, 2005). Nevertheless, measuring the phonetic distance of word pairs which are either known or assumed to be cognate provides a more fine-grained measure of linguistic distance than does binary cognate coding, which could prove particularly useful for groups of closely related languages or dialects in which most words tend to be cognate. By comparing multiple points within a single word pair, evaluating the similarity of phonetic forms also avoids the rigorous data reduction problem involved in binary cognate coding.

Previous approaches have most often calculated phonetic distance using the Levenshtein distance (e.g. Wichmann et al., 2010) rather than a gradient measure on the basis of the shared phonetic features of the sounds they represent. Some studies (e.g. Heeringa et al., 2006) have employed a modified Levenshtein distance with smaller costs for consonant-consonant substitutions as opposed to consonant-vowel substitutions, but still disregard the varying degrees of similarity between specific consonant or vowel pairs. The grounds for using the Levenshtein algorithm dates back to Kessler's (1995) study of Irish Gaelic dialects, in which his early feature-based phonetic distance was outperformed by the simpler Levenshtein distance, as compared to a baseline isogloss distance matrix of Irish dialects. Nevertheless, Kessler himself admits that a gradient measure would be preferable to one with uniform substitution costs, and suggests that future work could attempt to improve feature-based distance calculations through less arbitrary feature encoding and feature weighting.

Originally developed as a technique for yielding optimal alignments of phonetic sequences, Kondrak's (2002) ALINE algorithm is one method which has successfully derived feature-based measures of phonetic similarity. The features used multiple ordinal values, e.g. the PLACE feature included 11 values, each representing a distinct point of articulation, spaced at roughly equal intervals in the range [0, 1]. The features were further weighted by their salience. This method has been combined with a measure of semantic similarity and integrated into the COGIT cognate identification program, and was found to outperform orthographic methods such as the Levenshtein distance (Kondrak, 2001). The application of the measure has been restricted to phonetic alignment and cognate detection, however, and has not been extended to distance-based phylogenetic inference.

The proposed phonetic distance method for this study is based primarily around binary phonological distinctive features, similar to the method proposed by Heggarty et al. (2005) for dialect classification. The benefits of such a scheme are that it is more in line with theoretical phonology, and that gradient distances between sounds can be computed without needing to arbitrarily set multi-value features. For example, rather than as a single feature with multiple values as in Kondrak's ALINE algorithm, place of articulation is encoded through the combination of several binary features such as LABIAL, CORONAL, and DORSAL. Such a representation also allows for straightforward application of IPA diacritics, which modify the value of specific features, e.g. <ʷ> representing labialization would add the features +LABIAL, +ROUND to the base segment.

As a basis for the featural representation of sounds, the PHOIBLE database will be used, which encodes all IPA segments and diacritics for 37 distinctive features (Moran & McCloy, 2019). These features take three values, namely +, -, and 0. Features with a value of 0 are said to be irrelevant for the sound in question, as they are sub-features of other prerequisite features which are not present (e.g. ROUND is a sub-feature of LABIAL; in order to be +ROUND sounds must also be +LABIAL). Such values can, thus, be re-encoded as -, yielding true binary values + and - (*cf.* Odden, 2005). Pairwise distances between sounds can then be computed by a distance or similarity measures suitable for binary feature vectors, such as the Hamming distance, Jaccard index, or Dice similarity coefficient (Hamming, 1950; Jaccard, 1912; Dice, 1945), quantifying the proportion of distinctive features which the two sounds share.

More sophisticated techniques could involve weighting schemes and phonologically-informed penalties for deletions. Weighting schemes could concern the distinctive features themselves, weighting them by salience (as in Kondrak, 2002), by the average number of cross-linguistic phonemic contrasts they distinguish (as in Heggarty et al., 2005), or based on a hierarchical feature structure such as feature geometry (Clements, 1985). Individual segments within words can also be assigned weights according to their *n*-gram information content, segment type, or position within the word. Weighting by *n*-gram informativity (e.g. using the calculation for trigram information content proposed by Dellert, 2018) would shift the focus of the comparison to word roots, which are more informative compared with commonly occurring affixes. Weighting by segment type could, for example, give greater weight to consonants, which are less variable and influence comprehension of cognate forms to a greater extent than vowels (Gooskens, 2007). The first consonant of a root, in particular, is considered to be more stable and salient than consonants occurring later in the word (Kessler, 2005), so weighting schemes might also take relative position within a word form into account.

Just as not all substitutions should be treated equally, neither should all deletions. Many processes of sound change involve lenition, or weakening, of consonants, eventually resulting the in deletion of the consonant altogether (Heggarty, 2000). The likelihood of deletion of individual consonant segments can be modeled by their sonority, whereby more sonorous consonants are more likely to be elided. Elision is also more common in certain phonotactic positions, such as at the right edge of words or in consonant clusters (List, 2012). Deletions in such cases can thus be penalized to a lesser degree as they are more phonologically predictable. Alignment 1 shows an example of such a progression, with the Latin /t/ undergoing lenition in Italian and Spanish before being elided entirely in Portuguese and French, with the French form also deleting the final vowel.

Alignment 1

LA *patrem*, IT *padre*, ES *padre*, PT *pai*, FR *père* ‘FATHER’

LA	p	a	t	r	ɛ̃
IT	p	a	d	r	e
ES	p	a	ð	ɾ	e
PT	p	a			i
FR	p	ɛ		ʁ	

Alignment 2

IT *cane,* CA *ca,* FR *chien* ‘DOG’

IT	k		a	n	e
CA	k		a		
FR	ʃ	j	ɛ̃		

Context-dependent penalties for deleted segments may also prove useful. In addition to penalties differing by the sonority and position of the deleted segment, context-dependent penalties can account for phonetic features which surface differently in related forms. For example, a deleted nasal consonant would incur a smaller penalty if the preceding vowel is nasalized. An example (Alignment 2) would be the /n/ in Italian /kane/ ‘DOG’, which is deleted in both Catalan /ka/ and French /ʃjɛ̃/, though in French the nasal gesture is simply transferred to the preceding vowel and thus could be penalized less severely than the total deletion of the nasal in Catalan.

b. Information theoretic distances

A final, yet crucial, consideration in the proposed cognate evaluation techniques is a measure of consistency. The identification of genuine cognates which have developed from a common source, as opposed to chance resemblances or foreign borrowings, depends upon the adherence to recurring sound correspondences (Campbell, 2013). In most cases, sound correspondences in related languages involve pairs of relatively similar sounds. However, some correspondences are phonologically rather unexpected – for example, the correspondence between /k/ in Hawaiian and

/t/ in most other Polynesian languages[3] (Blust, 2004). Other correspondences appear surprising simply because a number of intermediate steps are hidden. An example is Spanish /x/, which corresponds with /ʎ/ in many other Romance languages, including closely related Portuguese[4]: /ʎ/ was first fricativized to /ʒ/ in medieval Spanish and subsequently devoiced to /ʃ/ along with other voiced sibilants, and finally its place of articulation was retracted to /x/ in modern Spanish (Mackenzie, 1999-2020). Despite their phonetically distant surface forms, these correspondences remain consistent and predictable throughout the lexicons of these languages.

Under the assumption that more closely related languages exhibit consistency in sound correspondences to a greater degree than do less closely related languages, distance methods which can capture the degree of consistency might be included alongside or in addition to phonetic distance methods. Jäger's (2018) distance using pointwise mutual information (PMI) is one such method, measuring co-occurrence of sound classes. Perhaps due to the fact that Jäger's study used short wordlists of only about 40 concepts, PMI between sound classes was measured globally and then scored for individual language pairs, rather than being calculated for each language pair. This approach is able to capture cross-linguistically common sound correspondences, but would penalize the more unusual examples cited above, which are nonetheless highly consistent within their respective contexts. One simple adjustment would be to calculate PMI for each language pair individually given long enough wordlists.

Two additional information theoretic measures are surprisal and entropy. Whereas PMI measures the degree of co-occurrence of two events, surprisal quantifies, in bits, the (un)expectedness of an outcome, with less predictable outcomes yielding higher surprisal values (Shannon, 1948). Surprisal is closely related to information theoretic entropy, which expresses the total uncertainty of a random variable, equivalent to the sum of surprisal values for all possible outcomes, multiplied by their respective probabilities (Moberg et al., 2007). Conditional entropy and surprisal measure the complexity of correspondences, yielding an uncertainty measure of zero for perfectly consistent correspondences, and increasing for more complex mappings. Conditional entropy of phoneme mappings has been shown to correlate with measures of intelligibility among Germanic languages, itself a correlate of linguistic distance (Moberg et al., 2007; Gooskens, 2007). Likewise, Stenger et al. (2017) introduce the measure of word adaptation surprisal as a strong predictor of intelligibility in Slavic languages, calculated as the length-normalized sum of conditional surprisal values of a word in Language A given its aligned equivalent in Language B. Word adaptation surprisal has the distinct advantage that it can be computed for specific word pairs, making it suitable as a distance measure to model the predictability of individual words of one language given equivalent words in another language. Although Stenger et al. (2017) measure word adaptation surprisal using orthographic texts, it is a versatile measure which can easily be applied to phonetic transcriptions, as well as extended to larger *n*-gram units beyond single phoneme correspondences.

[3] Compare Hawaiian *kolu* /kolu/ 'THREE' and *kanaka* /kanaka/ 'PERSON', with Māori *toru* /toru/ and *tangata* /taŋata/, Samoan *e tolu* /etolu/ and *tagata* /taŋata/, Ra'ivavae Austral /toɢu/ and /taʔata/, etc. (Walworth, 2018).

[4] Compare Spanish *oreja* /oɾexa/ 'EAR', *hoja* /oxa/ 'LEAF', *mujer* /muxeɾ/ 'WOMAN', with Portuguese equivalents *orelha* /oɾeʎɐ/, *folha* /foʎɐ/, and *mulher* /muʎɛɾ/ (Mackenzie, 1999-2000; Saenko, 2015).

4. Resources

The proposed methods will be applied to nine multilingual datasets (detailed in Table 3), representing a wide range of linguistic diversity and varying time depths. Collectively, the sampling includes approximately 230 varieties, belonging to six recognized language families and one proposed language family (Hokan), and representing four of the six linguistic macroareas defined by Hammarström & Donohue (2014), missing only South America and Australia.

Table 3. Datasets included in thesis study

Family, Subgroup	Varieties	Macroareas	Sources	References
Afro-Asiatic, Arabic	16	Africa, Eurasia	Varieties of Arabic Swadesh lists The glottometrics of Arabic	Wiktionary (2021) Ratcliffe (2020)
Austronesian, Polynesian	31	Papunesia	Polynesian Segmented Data	Walworth (2018)
Dravidian	20	Eurasia	DravLex	Kolipakam et al. (2018)
Hokan	20	North America	Global Lexicostatistical Database	Zhivlov (2011-2015)
Indo-European, Balto-Slavic	16	Eurasia	NorthEuraLex Global Lexicostatistical Database	Dellert et al. (2019) Kassian (2014) Saenko (2016-17)
Indo-European, Italic	58	Eurasia	Global Lexicostatistical Database	Saenko (2015)
Sino-Tibetan, Sinitic	19	Eurasia	Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects	Líu et al. (2007)
Turkic	31	Eurasia	Basic vocabulary datasets for the Turkic languages	Savelyev & Robbeets (2020)
Uralic	23	Eurasia	UraLex 2.0 NorthEuraLex	De Heer et al.; Syrjänen [*submitted manuscript*] Dellert et al. (2019)

In preparation for analysis, all selected datasets have been automatically extracted, preprocessed, and converted to a standardized format in accordance with the CLDF style (Forkel et al., 2018). In some cases, datasets have been assembled from multiple sources. General preprocessing steps included the removal of annotations (e.g. grammatical notes and morphemic segmentation characters) from phonetic transcriptions, conversion of non-IPA characters to proper IPA equivalents, correction of transcriptions for certain languages, standardization of glossing information to their equivalent concept labels in the Concepticon database (List et al., 2021), and cross-indexation of varieties with their Glottolog names, Glottocodes, and ISO 639-3 codes. Noteworthy additional details concerning individual datasets, including any exceptional preprocessing steps, are given in the following sections.

In order to evaluate the performance in the tree inference task, gold standard trees for each of the families have been extracted from the Glottolog database in Newick format (Hammarström et al., 2021). The trees have been automatically modified to prune varieties not present in the datasets, to add varieties not present in the Glottolog trees, and to modify the structure such that all varieties are listed as tips of the tree rather than internal nodes.

a. Arabic

Originally the Ratcliffe (2020) dataset of 14 Arabic dialects had been selected as source material. Several problems emerged while preprocessing this dataset, however. Many transcriptions were given only in the form of triconsonantal roots rather than full word forms, or else included characters whose phonetic value was unclear and could not be determined. Some varieties were missing large amounts of data, or were rather obscure and could not be classified within the Glottolog tree. Furthermore, incomplete cognate coding was provided – only listing cognacy with respect to Classical Arabic for each dialect and certain word forms.

Instead, the "Varieties of Arabic Swadesh lists" collection available from Wiktionary was chosen as a more suitable alternative. The Wiktionary dataset includes 16 dialects with sufficient coverage, all of which were successfully matched with their equivalents in the Glottolog tree. Moreover, explicit transcription conventions were provided with IPA equivalents, and all transcriptions were given in full word forms rather than triconsonantal roots. The exception is Maltese, for which orthographic forms were provided instead. Although Maltese orthography is largely transparent, its vowel length distinction is not indicated in the orthography and thus could not be transcribed automatically. Maltese IPA transcriptions were, thus, extracted automatically from the corresponding Wiktionary entries.

After extraction and standardization, the Wiktionary transcriptions were compared with those in the Ratcliffe dataset in order to assess the quality of the dataset. Very few discrepancies were found in terms of the word forms included. Minor differences in IPA transcription conventions can likely be attributed to the narrower transcriptions given in Wiktionary, and to the fact that Arabic dialects are not standardized, and thus are liable to some degree of variability.

Cognate set coding was performed semi-automatically, first using the LexStat cognate detection tool implemented in the LingPy Python library (List & Forkel, 2021), then corrected manually with reference to the limited cognate coding available in Ratcliffe's dataset.

b. Hokan

Hokan is a proposed language family of North America, comprising three smaller families (Cochimi-Yuman, Pomoan, Tequistlatecan) and five isolates (Chimariko, Karuk, Seri, Shasta, Yana), indigenous to California, Arizona, and Mexico. The data used here are taken from the Global Lexicostatistical Database (Zhivlov 2011-15) and are organized according to the constituent groups. Although it is an isolate, the Yana dataset includes data from three dialects of the language. Conversely, Highland Chontal is not an isolate, but it is the sole Tequistlatecan language for which data were available. Gold cognate sets are annotated separately for each of the constituent groups with more than one variety (Cochimi-Yuman, Pomoan, and Yana).

As opposed to the other families in this study which are already known to be related, the validity of Hokan as a family is still debated. Evidence in favor of or against the Hokan family is not expected to come from this study, but it will nonetheless be an interesting application of these methods toward a group without a consensus, and the distance measures derived among the Hokan sub-families and isolates can be compared with those of branches of other known families. Cognate sets and trees will be evaluated only with respect to each constituent group individually.

c. Balto-Slavic

The Slavic data have been drawn primarily from the NorthEuraLex dataset, which contains transcriptions of approximately 1000 concepts in over 100 languages of northern Eurasia and North America (Dellert et al., 2019). The NorthEuraLex owes its breadth, in large part, to the fact that much of it was compiled semi-automatically by non-experts in the target languages, and primarily transcribed using automatic grapheme-to-phoneme conversion tools. The contributors admit that many corrections would likely be necessary due to the novel methods of data collection and transcription.

Accordingly, adjustments been made to the data in all nine of the NorthEuraLex Slavic languages to correct clearly mistaken transcriptions or translations.[5] For Slovenian and Štokavian[6], the issues were simple enough to correct through modifying the existing transcriptions, whereas many of the transcription mistakes in the other Slavic languages were too complex. Instead, new grapheme-to-phoneme (G2P) conversion tools have been developed to automatically re-transcribe the data for these languages from the orthographic forms, ensuring that they address the shortcomings of the original automatic transcriptions. As Belarusian, Ukrainian, and Bulgarian all exhibit vowel reduction in unstressed syllables (Bird & Litvin, 2020; Pompino-Marschall et al., 2017; Ternes & Vladimirova-Buhtz, 1990), which is unmarked in the standard orthography, explicit stress annotation was added before applying the G2P conversion. Russian transcriptions were instead automatically extracted from Wiktionary entries as Russian orthography is less transparent and, thus, unsuitable for automatic G2P conversion.

Data for the Baltic languages, as well as for several additional Slavic varieties, have been extracted from the Global Lexicostatistical Database (Kassian, 2014; Saenko, 2016-17). These transcriptions were of sufficiently high quality that no corrections were necessary beyond the standard procedure of converting non-IPA symbols from the local transcription scheme to standard IPA symbols.

As no gold cognate set annotation was supplied for the NorthEuraLex dataset, word forms were first automatically organized into preliminary cognate sets using the LexStat cognate detection tool (List & Forkel, 2021) and then manually corrected with reference to the cognate coding in the Balto-Slavic portion of the IE-CoR lexical corpus.[7]

[5] Appendix A provides a list of common transcription problems with their corrections, along with the sources which these are based upon. Appendix B lists the translations which were removed or replaced by more suitable translations.

[6] In order to avoid the politicized issue surrounding the name(s) of the language(s) formerly known as Serbo-Croatian, and because the Kajkavian and Čakavian dialects are also included, the Standard Croatian data from NorthEuraLex will be referred to instead as Štokavian – the name of the dialect group upon which the standard Bosnian, Croatian, Montenegrin, and Serbian languages are all based (Alexander, 2006).

[7] Many thanks to Cormac Anderson and the IE-CoR consortium for making their gold Balto-Slavic cognate set data available as a reference for this study.

d. Turkic

The Turkic data were taken from the annotated dataset of 31 modern and historical Turkic varieties compiled by Savelyev & Robbeets (2020) for a recent Bayesian phylogenetic study of the family. A considerable amount of data cleaning was necessary, as in many cases orthographic forms, common Turkic transcriptions, and phonetic transcriptions were used interchangeably. In some cases it is not possible to distinguish between these without familiarity with the language in question, so some transcription errors may persist. Additionally, most transcriptions are very broad, possibly omitting relevant phonetic detail. In cases of doubt for how particular non-IPA symbols should be rendered in IPA, the transcriptions in comparable Turkic datasets (e.g. Straughn's (2017) online Turkic Database and the Turkic portion of NorthEuraLex) and relevant phonological descriptions were consulted. A few notable corrections are detailed in Appendix A.

e. Uralic

The UraLex 2.0 dataset (De Heer et al.; Syrjänen) contains 26 modern languages from all major branches of the Uralic family, plus Proto-Uralic, and includes both gold cognate set annotation and marking of loanwords. Of these original 27 varieties, 23 of the modern varieties have been included in the present study.

Several languages had no IPA transcriptions – instead they were given either in the Uralic Phonetic Alphabet or in the native orthography. Among these, Karelian, Livonian, North Saami, Skolt Saami, South Saami, Tundra Nenets, and Veps were also included in the NorthEuraLex dataset and thus IPA transcriptions were extracted semi-automatically from the latter. Only entries with matching Concepticon glosses in both datasets *and* whose orthographic form in UraLex was identical to that in NorthEuraLex were automatically combined, in order to ensure that the cognate set and borrowing data included in the UraLex entries would remain relevant. The remaining entries with matching Concepticon glosses but non-identical forms were checked manually. In most cases, genuine matches were obvious and only concerned a difference of a few characters.[8] Over 200 transcriptions were successfully extracted from NorthEuraLex for each of these languages.

The Võro language had only one IPA transcription available in the UraLex dataset, and was not included in NorthEuraLex. Instead, a grapheme-to-phoneme (G2P) converter was written to automatically transcribe the given orthographic forms into IPA as the orthography is transparent.

Only the four remaining UraLex languages without IPA transcriptions (Ume Saami, Pite Saami, Inari Saami, and Proto-Uralic), which were not included in NorthEuraLex, were excluded from the present study. As four other Saami varieties are already included, and all major branches of modern Uralic languages remain represented in the dataset, this omission seems acceptable. Other Uralic languages from NorthEuraLex not represented in UraLex (e.g. Moksha, Olonets Karelian, Forest Enets) were likewise excluded because cognate coding was not available for them.

[8] e.g. Karelian <jeätyö> in UraLex vs. <jiätyö> /jiætyœ/ 'FREEZE' in NorthEuraLex;
North Saami < soadjá ~ soadji> in UraLex vs. <soadji> /sɔ̯ɑcːi/ 'WING' in NorthEuraLex;
Tundra Nenets <ŋodya> in UraLex vs. <ӈодя> /ŋoːdʲɑ/ 'BERRY' in NorthEuraLex

5. Structure of Thesis Study

As a preceding step, the datasets to be analyzed have been cleaned, preprocessed, and converted into a standardized format. Gold cognate sets and Glottolog trees have been extracted for the families to be examined. A common subset of the most frequent 110 concepts has been selected for the purposes of comparing results across datasets.

The proposed phonetic, PMI, and surprisal-based distance measures will be developed and implemented in Python prior to analysis. A composite distance measure combining aspects of both phonetic and information theoretic distances will also be implemented. These measures all require aligned phonetic sequences as input, so a prerequisite is a working phonetic sequence alignment algorithm. The accuracy of the alignment algorithm will be evaluated by its performance on a database of gold phonetic alignments from multiple language families (List & Prokić, 2014).

An automated cognate detection tool will then be implemented, in which each distance measure will be taken as the criterion for judging cognacy of individual word pairs, modeled after the method outlined by Jäger (2018). Given that all words in a cognate set should be labeled as cognate with all other members of the set, individual cognate pairs should be additionally clustered into cognate sets in a following step. The automatically generated cognate sets according to each criterion will be evaluated for accuracy against the gold cognate sets available for each family.

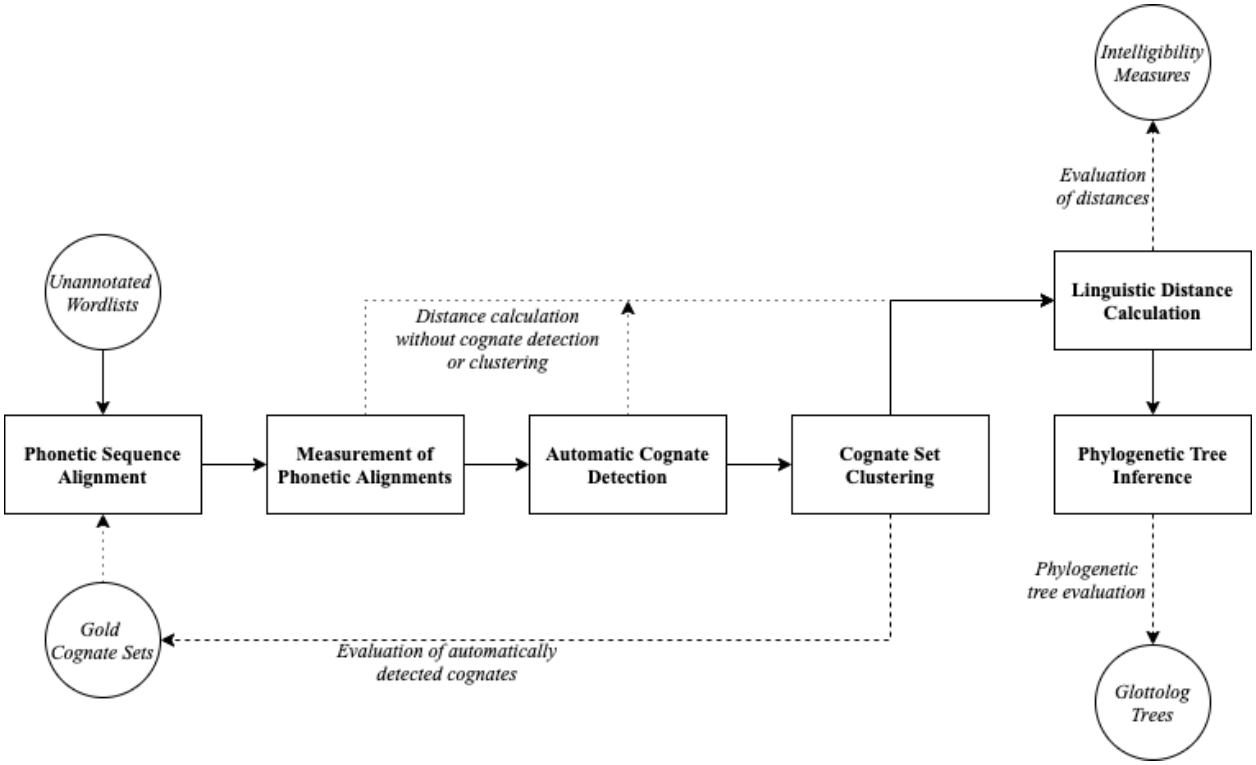
Intelligibility Measures
Evaluation of distances
Unannotated Wordlists
Distance calculation without cognate detection or clustering
Linguistic Distance Calculation
Phonetic Sequence Alignment
Measurement of Phonetic Alignments
Automatic Cognate Detection
Cognate Set Clustering
Phylogenetic Tree Inference
Phylogenetic tree evaluation
Gold Cognate Sets
Evaluation of automatically detected cognates
Glottolog Trees

Figure 2. Workflow of phylogenetic inference steps and evaluation

The phylogenetic tree inference task can proceed either following or independent of the automatic cognate detection task, as illustrated in Figure 2. Wichmann et al.'s (2010) and Jäger's (2018) pairwise linguistic distances are calculated without first performing cognate detection – that is, both cognate and non-cognate word pairs receive a score. While this is possible, a more logical approach would be to first perform cognate detection and keep only the scores of cognate pairs, with detected non-cognate pairs receiving a similarity score of zero instead. Particularly for phonetic distance measures, it makes little sense to speak of the distance between unrelated word forms. However, it may turn out that certain measures prove effective for distance-based phylogenetic inference through measuring cognate distances but ineffective for cognate detection in the preceding step. To control for the differing cognate sets which each distance measure may yield, they should also be applied to gold cognate sets so that their performances can be compared more directly.

The suitability of these methods as distance measures for distance-based phylogenetic inference will be assessed by comparing generated trees against gold family trees from Glottolog. The degree to which phylogenetic trees overlap can be measured using the Generalized Quartet Distance, which measures the number of matching clusters within the two trees, and by TreeDist, an information-theoretic measure which evaluates the mutual information expressed the splits of the two trees (Smith, 2020).

Trees are not necessarily the optimal representation of language families, however. Factors such as horizontal transmission and sustained contact, in which sister or cousin languages continue to influence one another after diverging, as well as the theory that innovations gradually spread across groups of languages in waves, mean that a tree model with binary splits is often insufficient for describing the relationships among a group of languages (François, 2014). Instead, a network-like structure derived from pairwise distances may be a more appropriate representation for certain families. Although such data are not available for all language families examined in this study, correlations between the computed inter-linguistic distances and empirical measures of intelligibility will be measured where available.[9] This comparison will provide a valuable additional perspective to the evaluation of these methods, as intelligibility measures are more directly comparable with the linguistic distances which will be calculated and likewise support visualization of language groups as networks.

6. Conclusion

In this work several lexical distance measures based in phonology and information theory will be developed for tasks in computational historical linguistics, including automatic cognate detection and phylogenetic tree building. These measures will be applied to multilingual datasets encompassing over 200 languages from at least seven language families. The performance will be assessed with respect to expert-produced gold cognate sets and phylogenetic trees, and the distance measures will be correlated with empirical measures of intelligibility.

The comparison of these methods in parallel will shed light onto what types of criteria are most useful for distance-based phylogenetic inference. The methods are expected to produce relatively accurate phylogenetic trees using a less computationally intensive approach, compared with

[9] For example, for Sinitic (Tang & van Heuven, 2015), Arabic dialects (Čéplö et al., 2016; Trentman & Shiri, 2020), Slavic (Golubović & Gooskens, 2015), and Romance (Gooskens et al., 2018).

character-based methods. They will make the most of the available data by taking a more fine-grained, linguistically nuanced approach to evaluating word pairs, thus avoiding issues associated with rigorous data reduction through binary cognacy annotation or sound classes. If successful, these methods may represent useful new techniques for language classification.

7. Works Cited

Alexander, R. (2006). *Bosnian, Croatian, Serbian: A Grammar with Sociolinguistic Commentary*. University of Wisconsin Press.

Appendix:Varieties of Arabic Swadesh lists. (2021). Wiktionary, The Free Dictionary. https://en.wiktionary.org/w/index.php?title=Appendix:Varieties_of_Arabic_Swadesh_lists&oldid=62250595

Bird, S., & Litvin, N. (2020). Belarusian. *Journal of the International Phonetic Association*, 1–18. https://doi.org/doi:10.1017/S0025100319000288

Blust, R. (2004). t to k: An Austronesian Sound Change Revisited. *Oceanic Linguistics*, *43*(2), 365–410. https://doi.org/DOI:10.1353/ol.2005.0001

Campbell, L. (2013). *Historical Linguistics: An Introduction* (3rd ed.). Edinburgh University Press.

Čéplö, S., Bátora, J., Benkato, A., Milička, J., Pereira, C., & Zemánek, P. (2016). Mutual intelligibility of spoken Maltese, Libyan Arabic, and Tunisian Arabic functionally tested: A pilot study. *Folia Linguistica*, *50*(2), 583–628.

Clements, G. N. (1985). The Geometry of Phonological Features. *Phonology Yearbook*, *2*, 225–252.

de Heer, M., Blokland, R., Dunn, M., & Vesakoski, O. (submitted manuscript). *Loanwords in basic vocabulary as an indicator of borrowing profiles*.

de Heer, M., Heikkilä, M., Syrjänen, K., Lehtinen, J., Vesakoski, O., Suutari, T., Dunn, M., Määttä, U., & Leino, U.-P. (2021). *UraLex 2.0 Uralic basic vocabulary with cognate and loanword information*. Zenodo. https://zenodo.org/record/4777568#.YMXcH5MzbjA

Dellert, J. (2018). Combining Information-Weighted Sequence Alignment and Sound Correspondence Models for Improved Cognate Detection. *Proceedings of the 27th International Conference on Computational Linguistics*, 3123–3133.

Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., Grigorjew, I., Balabel, M., Boga, H. I., Baysarova, Z., Mühlenbernd, R., Wahle, J., & Jäger, G. (2020). NorthEuraLex: A wide-coverage lexical database of Northern Eurasia. *Lang Resources & Evaluation*, *54*, 273–301. https://doi.org/10.1007/s10579-019-09480-6

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297–302. https://doi.org/doi:10.2307/1932409

Forkel, R., List, J.-M., Greenhill, S., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G., & Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, *5*(180205). https://doi.org/10.1038/sdata.2018.205

François, A. (2014). Trees, waves and linkages: Models of language diversification. In C. Bowern & B. Evans (Eds.), *The Routledge Handbook of Historical Linguistics* (pp. 161–189). Routledge.

Gamallo, P., Pichel, J. R., & Alegria, I. (2017). From language identification to language distance. *Physica A*, *484*, 152–162. http://dx.doi.org/10.1016/j.physa.2017.05.011

Golubović, J., & Gooskens, C. (2015). Mutual intelligibility between West and South Slavic languages. *Russian Linguistics*, *39*(3), 351–373.

Gooskens, C. (2007). The Contribution of Linguistic Factors to the Intelligibility of Closely Related Languages. *Journal of Multilingual and Multicultural Development*, *28*(6), 445–467. https://doi.org/10.2167/jmmd511.0

Gooskens, C., van Heuven, V. J., Golubović, J., Schüppert, A., Swarte, F., & Voigt, S. (2018). Mutual intelligibility between closely related languages in Europe. *International Journal of Multilingualism*, *15*(2), 169–193.

Hamann, S. (2004). Retroflex fricatives in Slavic languages. *Journal of the International Phonetic Association*, *34*(1), 53–67. https://doi.org/DOI:10.1017/S0025100304001604

Hammarström, H., & Donohue, M. (2014). Some Principles on the Use of Macro-Areas in Typological Comparison. *Language Dynamics and Change*, *4*(1), 167–187. https://doi.org/DOI:10.1163/22105832-00401001

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2021). *Glottolog 4.4*. Leipzig: Max Planck Institute for Evolutionary Anthropology; https://doi.org/10.5281/zenodo.4761960. http://glottolog.org

Hanulíková, A., & Hamann, S. (2010). Slovak. *Journal of the International Phonetic Association*, *40*(3), 373–378. https://doi.org/doi:10.1017/S0025100310000162

Heeringa, W., Kleiweg, P., Gooskens, C., & Nerbonne, J. (2006). Evaluation of String Distance Algorithms for Dialectology. *Proceedings of the Workshop on Linguistic Distances*, 51–62.

Heeringa, W., Swarte, F., Schüppert, A., & Gooskens, C. (2018). Measuring syntactical variation in Germanic texts. *Digital Scholarship in the Humanities*, *33*(2), 279–296.

Heggarty, P. (2000). Quantifying change over time in phonetics. In C. Renfrew, A. McMahon, & L. Trask (Eds.), *Time Depth in Historical Linguistics* (pp. 531–562). McDonald Institute for Archaeological Research.

Heggarty, P., McMahon, A., & McMahon, R. (2005). From phonetic similarity to dialect classification: A principled approach. In N. Delbecque, D. Geeraerts, & J. van der Auwera (Eds.), *Perspectives on Variation: Sociolinguistic, Historical, Comparative* (pp. 43–91). Mouton de Gruyter.

Hoey, E. M. (2013). *Grammatical sketch of Turkmen* [Master of Arts]. University of California, Santa Barbara.

Jaccard, P. (1912). The Distribution of the Flora in the Alpine Zone. *The New Phytologist*, *11*(2), 37–50.

Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, *5*(180189), 1–16. https://doi.org/10.1038/sdata.2018.189

Jassem, W. (2003). Polish. *Journal of the International Phonetic Association*, *33*(1), 103–107. https://doi.org/DOI:10.1017/S0025100303001191

Kapović, M. (2007). Hrvatski standard – evolucija ili revolucija? Problem hrvatskoga pravopisa i pravogovora. *Jezikoslovlje*, *8*(1), 61–76.

Kassian, A. (2014). Annotated Swadesh wordlist for Dihovo Macedonian (Slavic group, Indo-European family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow: Higher School of Economics, & Santa Fe: Santa Fe Institute. http://starling.rinet.ru/new100/

Kessler, B. (2005). Phonetic Comparison Algorithms. *Transactions of the Philological Society*, *103*(2), 243–260.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics*, 60–66. https://doi.org/10.3115/976973.976983

Kolipakam, V., Dunn, M., Jordan, F. M., & Verkerk, A. (2018). *DravLex: A Dravidian lexical database*. Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. http://doi.org/10.5281/zenodo.1117644

Kondrak, G. (2001). *Identifying Cognates by Phonetic and Semantic Similarity*. Second Meeting of the North American Chapter of the Association for Computational Linguistics. https://aclanthology.org/N01-1014

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

List, J.-M. (2012). SCA: Phonetic alignment based on sound classes. In D. Lassiter & M. Slavkovik (Eds.), *New directions in logic, language, and computation* (pp. 32–51). Springer. http://rd.springer.com/chapter/10.1007/978-3-642-31467-4_3

List, J.-M., & Forkel, R. (2021). *LingPy. A Python library for historical linguistics* (2.6.8) [Python]. Max Planck Institute for Evolutionary Anthropology. https://lingpy.org

List, J.-M., & Prokić, J. (2014). A Benchmark Database of Phonetic Alignments in Historical Linguistics and Dialectology. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 288–294.

List, J.-M., Rzymski, C., Greenhill, S., Schweikhard, N., Pianykh, K., Tjuka, A., Hundt, C., & Forkel, R. (Eds.). (2021). *CLLD Concepticon 2.5.0 [Data set]*. Zenodo. https://doi.org/10.5281/zenodo.4911605

Líu, L., Wáng, H., & Bǎi, Y. (2007). *Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí 现代汉语方言核心词·特征词集 [Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]*. Nánjīng 南京: Fènghuáng 凤凰.

Mackenzie, I. (1999, 2020). *History of Spanish Consonants*. The Linguistics of Spanish. https://www.staff.ncl.ac.uk/i.e.mackenzie/cons.htm

Moberg, J., Gooskens, C., Nerbonne, J., & Vaillette, N. (2007). Conditional Entropy Measures Intelligibility among Related Languages. *Proceedings of Computational Linguistics in the Netherlands*.

Mokari, P. G., & Werner, S. (2017). Azerbaijani. *Journal of the International Phonetic Association*, *47*(2), 207–212.

Moran, S., & McCloy, D. (Eds.). (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History. http://phoible.org

Murdarov, V. (2012). *Официален правописен речник на българския език [Official orthographic dictionary of the Bulgarian language]* (M. Buneva, Ed.). Просвета [Prosveta].

Odden, D. (2005). Feature Theory. In *Introducing Phonology* (pp. 129–168). Cambridge University Press.

Pagel, M. (2017). Darwinian perspectives on the evolution of human languages. *Psychonomic Bulletin & Review*, *24*(1), 151–157. https://doi.org/DOI:10.3758/s13423-016-1072-z

Pompino-Marschall, B., Steriopolo, E., & Żygis, M. (2017). Ukrainian. *Journal of the International Phonetic Association*, *47*(3), 349–357.

Rama, T., & Kolachina, S. (2013). Distance-based Phylogenetic Inference Algorithms in the Subgrouping of Dravidian Languages. In L. Borin & A. Saxena (Eds.), *Approaches to Measuring Linguistic Differences* (pp. 121–158).

Ratcliffe, R. R. (2020). The glottometrics of Arabic: Quantifying linguistic diversity and correlating it with diachronic change. *Language Dynamics and Change*, *11*(1), 1–29.

Sabev, M. (2000). *The Sound System of Standard Bulgarian*. http://www.personal.rdg.ac.uk/~llsroach/phon2/b_phon/b_phon.htm

Saenko, M. (2015). Annotated Swadesh wordlists for the Romance group (Indo-European family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow: Higher School of Economics, & Santa Fe: Santa Fe Institute. http://starling.rinet.ru/new100/

Saenko, M. (2016). Annotated Swadesh wordlists for the Slavic group (Indo-European family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow: Higher School of Economics, & Santa Fe: Santa Fe Institute. http://starling.rinet.ru/new100/

Saenko, M. (2017). Annotated Swadesh wordlists for the Baltic group (Indo-European family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow: Higher School of Economics, & Santa Fe: Santa Fe Institute. http://starling.rinet.ru/new100/

Savelyev, A., & Robbeets, M. (2020). Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family. *Journal of Language Evolution*, *5*(1), 39–53.

Séguy, J. (1973). La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane*, *37*, 1–24. https://doi.org/10.5169/seals-658403

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Šimáčková, Š., Podlipský, V. J., & Kateřina, C. (2012). Czech spoken in Bohemia and Moravia. *Journal of the International Phonetic Association*, *42*(2), 225–232. https://doi.org/doi:10.1017/S0025100312000102

Smith, M. R. (2020). *TreeDist: Distances between Phylogenetic Trees* (R package version 2.1.1.) [Computer software]. doi: 10.5281/zenodo.3528124

Sofroniev, P. (2018). *Asjp* (0.0.2) [Computer software]. https://pypi.org/project/asjp/

Stenger, I., Avgustinova, T., & Marti, R. (2017). Levenshtein distance and word adaptation surprisal as methods of measuring mutual intelligibility in reading comprehension of Slavic languages. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017."* Dialogue 2017, Moscow.

Stolarski, Ł. (2010). *Palatalization of consonants in Polish before /i /and /j/.*

Straughn, C. A. (2017). *Turkic Database*. http://turkic.elegantlexicon.com/

Syrjänen, K., Maurits, L., Leino, U., Honkola, T., Rota, J., & Vesakoski, O. (submitted manuscript). *Crouching TIGER, Hidden Structure: Exploring the nature of linguistic data using TIGER values.*

Tang, C., & van Heuven, V. J. (2015). Predicting mutual intelligibility of Chinese dialects from multiple objective linguistic distance measures. *Linguistics*, *53*(2), 285–311.

Ternes, E., & Vladimirova-Buhtz, T. (1990). Bulgarian. *Journal of the International Phonetic Association*, *20*(1), 45–47.

Trentman, E., & Shiri, S. (2020). The Mutual Intelligibility of Arabic Dialects: Implications for the Language Classroom. *Critical Multilingualism Studies*, *8*(1), 104–134.

Ünal-Logacev, Ö., Żygis, M., & Fuchs, S. (2019). Phonetics and phonology of soft 'g' in Turkish. *Journal of the International Phonetic Association*, *49*(2), 183–206. https://doi.org/doi:10.1017/S0025100317000317

Vakulenko, M. O. (2018). Ukrainian vowel phones in the IPA context. *Govor*, *35*(2), 189–214.

Walworth, M. (2018). *Polynesian Segmented Data (Version 1) [Data set]*. Zenodo. http://doi.org/10.5281/zenodo.1689909

Wichmann, S., Holman, E. W., Bakker, D., & Brown, C. H. (2010). Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and Its Applications*, *389*, 3632–3639. https://doi.org/doi:10.1016/j.physa.2010.05.011

Wichmann, S., Holman, E. W., & Brown, C. H. (Eds.). (2020). *The ASJP Database (version 19)*.

Zaleska, J., & Nevins, A. (2014, September 4). *Polish nasal vowels are autosegmental: Word game evidence*. 2014 Annual Meeting of the Linguistics Association of Great Britain, Oxford.

Zhivlov, M. (2011a). Annotated Swadesh wordlists for the Karuk group (Hokan family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow: Higher School of Economics, & Santa Fe: Santa Fe Institute. http://starling.rinet.ru/new100/

Zhivlov, M. (2011b). Annotated Swadesh wordlists for the Seri group (Hokan family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow: Higher School of Economics, & Santa Fe: Santa Fe Institute. http://starling.rinet.ru/new100/

Zhivlov, M. (2012a). Annotated Swadesh wordlists for the Tequistlatecan group (Hokan family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow: Higher School of Economics, & Santa Fe: Santa Fe Institute. http://starling.rinet.ru/new100/

Zhivlov, M. (2012b). Annotated Swadesh wordlists for the Shastan group (Hokan family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow: Higher School of Economics, & Santa Fe: Santa Fe Institute. http://starling.rinet.ru/new100/

Zhivlov, M. (2013a). Annotated Swadesh wordlists for the Yana group (Hokan family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow: Higher School of Economics, & Santa Fe: Santa Fe Institute. http://starling.rinet.ru/new100/

Zhivlov, M. (2013b). Annotated Swadesh wordlists for the Pomo group (Hokan family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow: Higher School of Economics, & Santa Fe: Santa Fe Institute. http://starling.rinet.ru/new100/

Zhivlov, M. (2015). Annotated Swadesh wordlists for the Yuman group (Hokan family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow: Higher School of Economics, & Santa Fe: Santa Fe Institute. http://starling.rinet.ru/new100/

Zimmer, K., & Orgun, O. (1992). Turkish. *Journal of the International Phonetic Association*, *22*(1), 43–45.

Appendix A: Noteworthy Transcription Changes

Code	Description of Problem or Incorrectly Transcribed Phenomenon	Example Word(s)	Incorrect Transcription	Correct Transcription	Reference
AZ	/k, g/ are palatalized to /c, ɟ/ in syllables with front vowels	iki, göy	iki, gjøj	ici, ɟøj	Mokari & Werner, 2017
AZ	<q> is voiced velar stop /g/, not uvular /ɢ/	balıq	balɯɢ	bɑlɯg	Mokari & Werner, 2017
AZ	<ə> is near-open front unrounded vowel /æ/	mən	mɛn	mæn	Mokari & Werner, 2017
BE	word-final obstruents are devoiced	зуб	zub	zup	Bird & Litvin, 2020
BE	consonants are palatalized before <ь, е, і, ю, я>	кісць	kist͡sʲ	kʲisʲʦʲ	Bird & Litvin, 2020
BE	/a, ɛ, o, u/ are [æ, e, ɵ, ʉ] after palatalized consonants (orthographically <я, е, ё, ю>)	пяро	pʲarɔ	pʲæro	Bird & Litvin, 2020
BE	non-palatalized <г> is uvular /ʁ/	галава	ɣalava	ʁalava	Bird & Litvin, 2020
BE	<дз> is an affricate /ʥ/	дзесяць	dzʲɛsʲat͡sʲ	ʥʲesʲæʦʲ	Bird & Litvin, 2020
BE	obstruents assimilate regressively to voicing and palatalization of following obstruents	гладкі, блізкі	ɣladki, blizki	ʁlatʲkʲi, blʲisʲkʲi	Bird & Litvin, 2020
BG	<a, ъ> /a, ɤ/ are both reduced to [ɐ] in unstressed syllables	глава, нокът	głava, nɔkɤt	głɐvˈa, nˈɔkɐt	Ternes & Vladimirova-Buhtz, 1990
BG	<o> /ɔ/ is reduced to [o] in unstressed syllables	коляно	kɔʎanɔ	kolʲˈano	Ternes & Vladimirova-Buhtz, 1990
BG	obstruents assimilate regressively to voicing of following obstruents	дъжд	dɤʒt	dɤʃt	Sabev, 2000
BG	verbs ending in stressed <a, я> are pronounced with /(j)ɤ/ instead of /(j)a/, as if spelled <ъ>	стоя, спя, лежа	stɔja, spʲa, lɛʒa	stojˈɤ, spjɤ, lɛʒˈɤ	Murdarov, 2012
CZ	/r̝/ does not trigger regressive voicing assimilation	stříbro	zdr̝iːbro	str̝̊iːbro	Šimáčková et al., 2012
CZ	/r̝/ undergoes voicing assimilation in both directions	hořký, přijít	ɦor̝kiː, br̝ijiːt	ɦor̝̊kiː, pr̝̊ijiːt	Šimáčková et al., 2012
CZ	/v/ does not trigger regressive voicing assimilation	světlo	zvjɛtlo	svjɛtlo	Šimáčková et al., 2012
CZ	<d, t> are both palatalized to /ɟ, c/ before <i, í, ě>	dítě	diːcɛ	ɟiːcɛ	Šimáčková et al., 2012
CZ	sequence <sh> /sɦ/ is exception to regressive voicing assimilation, realized as [sx]	shnít	zɦɲiːt	sxɲiːt	Šimáčková et al., 2012
HR	<ije> represents /jeː/, the long version of <je>	zvijezda, korijen	zʋijĕzda, kɔ̂riɛn	zʋjeːˇzda, kôrjeːn	Kapović, 2007
PL	sequence <ni> represents /ɲ(i)/	śnieg	ɕniɛk	ɕɲɛk	Jassem, 2003
PL	<rz> represents /ʐ/, or /ʂ/ if devoiced	twarz	tfaɾs	tfaʂ	Hamann, 2004
PL	<iV> represents either <ʲV>~<ʲjV>~<jV>, depending on the analysis, but not <iV>	światło	ɕfiatwɔ	ɕfʲjatwɔ	Stolarski, 2010
PL	<ę, ą> realized as oral vowel + nasal glide or consonant /Vw̃/~/VN/, depending on context	kąsać	kɔ̃sat͡ɕ	kɔw̃saʨ	Zaleska & Nevins, 2014
PL	word-final <ę> typically loses its nasality	się	ɕɛ̃	ɕɛ	Zaleska & Nevins, 2014
PL	word-final obstruents are devoiced	krew	krɛv	krɛf	Jassem, 2003
SK	<n, t, d> are palatalized to /ɲ, c, ɟ/ before <e, i>	oni, otec, sedem	ɔni, ɔtɛt͡s, sɛdɛm	ɔɲi, ɔcɛʦ, sɛɟɛm	Hanulíková & Hamann, 2010
SK	/v/ does not trigger regressive voicing assimilation	kvet	gvɛt	kʋɛt	Hanulíková & Hamann, 2010
SK	obstruents devoice word-finally and assimilate regressively to voicing of following obstruents	dážď	daːʒɟ	daːʃc	Hanulíková & Hamann, 2010
TK	<s, z> represent /θ, ð/	saç, gyzyl	saʨ, ɢɯzɯl	θɑʧ, gɯðɯl	Hoey, 2013
TK	<g> is voiced velar stop /g/, not uvular /ɢ/	gara	ɢara	gɑrɑ	Hoey, 2013
TR	Standard Turkish does not have phoneme /q/	balık	balɯq	bałɯk	Zimmer & Orgun, 1992
TR	<ğ> realized as lengthening of preceding vowel (never /ɣ/ in standard Turkish)	yağmur, soğuk	jaɣmur, soɣuk	jɑːmur, soːuk	Ünal-Logacev et al., 2019
TR	/k, g/ are palatalized to /c, ɟ/ in syllables with front vowels	gölge, kemik	gølge, kemik	ɟølʲɟe, cemic	Zimmer & Orgun, 1992
TR	<l> is /ɫ/ adjacent to back vowels and /lʲ/ adjacent to front vowels	kol, kül	kol, kyl	koł, cylʲ	Zimmer & Orgun, 1992
UK	non-palatalized <л> is dark /ɫ/	голова	ɦɔlɔwɑ	ɦɔłɔʋɑ	Pompino-Marschall et al., 2017

UK	no transcription of stress, or vowel reduction in unstressed syllables	ягода	jaɦɔda	jˈaɦɔdɐ	Pompino-Marschall et al., 2017; Vakulenko, 2018
UK	<в> is [w] only before rounded back vowels, otherwise [ʋ]~[ʋʲ]	новий	nɔwɪj	nɔʋɪi̯	Pompino-Marschall et al., 2017
UK	consonants are palatalized before <ь, є, і, ю, я>	кістка	kistka	kʲistka	Pompino-Marschall et al., 2017

Appendix B: Replaced or Removed Translations

Code	Concept	Original Word	New Word	Reason	Notes
AZ	NIGHT	kecə	gecə	Misspelling	
BE	COLD	холадны	халодны	Misspelling	
BE	HAND	кісць	рука	Not most common term	
BG	BELLY	корем	търбух	Match cognate set in related languages	Turkic loanword; *търбух* is native Slavic and matches cognate set in other languages
BG	BREAST	бозка	гръд	Not most general term	NIPPLE rather than BREAST
BG	HAND	китка	∅	Incorrect translation	WRIST rather than HAND
BG	SHORT	къс	кратък	Not most common term	
CZ	SEED	osivo	semeno	Incorrect translation	
HR	BREAST	njedra	prsa	Not most common term	Literary; *prsa* is the more common term
HR	BREAST	sisa	grudi	Match cognate set in related languages	*sisa* and *grudi* both refer to BREAST, but *grudi* matches the cognate set in sister languages
HR	DOG	pseto	∅	Incorrect translation	Derogatory term for a dog or detestable person; Correct translation *pas* already included
HR	MOUNTAIN	brdo	planina	Incorrect translation	HILL rather than MOUNTAIN
HR	SEED	sjemenje	sjeme	Incorrect form	Plurale tantum of *sjeme*
HR	SKIN	put	∅	Incorrect translation	COMPLEXION rather than SKIN; more general term *koža* already included
HR	SNAKE	guja	∅	Not most common term	Literary; more general word *zmija* already included
HR	WHITE	bio	∅	Repeated lemma; Incorrect form	Ikavian form of *bijel*, which is already included
PL	HAND	dłoń	ręka	Incorrect translation	PALM OF THE HAND rather than HAND
PL	SEED	siew	nasiono	Incorrect translation	SOWING rather than SEED
RU	BREAST	женская грудь	грудь	Not most general term	Original translation was literally FEMALE BREAST rather than BREAST
RU	FULL	наполненный	∅	Not most general term	More general translation *полный* already included
RU	HAND	кисть	рука	Not most common term	
RU	LONG	длинный	долгий	Match cognate set in related languages	*долгий* and *длинный* both mean LONG, but *долгий* matches the cognate set in sister languages
SK	CLOUD	mračno	mrak	Incorrect form	
SK	EGG	vajíčko	∅	Repeated lemma	Diminutive of *vajce*, which is already included
SK	SEED	osivo	semeno	Incorrect translation	
TR	ABOVE	üştünde	üstünde	Misspelling	
UK	WHITE	сивий	білий	Incorrect translation	GRAY rather than WHITE

Language Codes in Appendices

AZ	Azerbaijani
BE	Belarusian
BG	Bulgarian
CZ	Czech
HR	Croatian (Štokavian)
PL	Polish
SK	Slovak
SL	Slovenian
TK	Turkmen
TR	Turkish
UK	Ukrainian