Squib

How to Distinguish Languages and Dialects

Søren Wichmann
Leiden University Centre for
Linguistics, Kazan Federal University,
and Beijing Advanced Innovation Center
for Language Resources
wichmannsoeren@gmail.com

The terms "language" and "dialect" are ingrained, but linguists nevertheless tend to agree that it is impossible to apply a non-arbitrary distinction such that two speech varieties can be identified as either distinct languages or two dialects of one and the same language. A database of lexical information for more than 7,500 speech varieties, however, unveils a strong tendency for linguistic distances to be bimodally distributed. For a given language group the linguistic distances pertaining to either cluster can be teased apart, identifying a mixture of normal distributions within the data and then separating them fitting curves and finding the point where they cross. The thresholds identified are remarkably consistent across data sets, qualifying their mean as a universal criterion for distinguishing between language and dialect pairs. The mean of the thresholds identified translates into a temporal distance of around one to one-and-a-half millennia (1,075–1,635 years).

1. Two Approaches

Most linguists would agree that it is difficult and often controversial to distinguish languages from dialects. Many, however, would also agree that the notions of language and dialect are still useful, even for the linguist who is aware of the problems of definition that they entail (Agard 1984). The distinction is useful for many different purposes, such as cataloguing languages, assigning ISO 639-3 codes, preparing maps of languages, planning revitalization efforts, or for doing statistics on language distributions (e.g., calculating diversity or density indices) (Korjakov 2017). More importantly, perhaps: If such a distinction is a feature of the way that language varieties are distributed rather than just a distinction we impose in some arbitrary way, then this would be important for the understanding of the sociology of language at large.

There are two main directions to go in order to establish a quantitative distinction. One direction is to measure mutual intelligibility; another is to apply some consistent and objective measure of differences between two variants with regard to phonology, morphology, syntax, lexicon, or some combination.

Early applications of mutual intelligibility testing are detailed in Casad (1974), and more recent work in this area includes Whaley, Grenoble, and Li (1999), Szeto (2000), Gooskens and Schneider (2016), and Gooskens et al. (2018).

Submission received: 18 February 2019; revised version received: 11 July 2019; accepted for publication: 15 September 2019.

https://doi.org/10.1162/COLI_a_00366

© 2019 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International
(CC BY-NC-ND 4.0) license

Glottolog (Hammarström, Forkel, and Haspelmath 2017) adopts the criterion of mutual intelligibility, positing that a language variant that is not mutually intelligible with any other language variant should be counted as a separate language.[1] By this criterion, Glottolog 4.0 contains 7,592 spoken L1 (mother-tongue) languages, excluding sign languages. There are, however, two problems with this criterion. The more serious problem is that intelligibility is often not symmetrical. Thus language variant A can be more intelligible to speakers of language variant B than language variant B is to speakers of language variant A. Such a situation may arise when A is the larger, more influential language, causing speakers of B to have more exposure to A than the other way around. However, the amount of exposure that speakers have to other language variants is entirely determined by historical and sociological factors, and this or other extraneous factors[2] should not affect a linguistically based classification. In some situations the factor of exposure can be circumvented, narrowing in on "inherent intelligibility" (Gooskens and van Heuven 2019), but this is not an easy task. The more practical problem with the criterion of mutual intelligibility is that measurements are usually simply not available.

The second approach was referred to by Voegelin and Harris (1951) as "count sameness." While recognizing that "sameness" can be measured for different areas of linguistic structure, they place emphasis on the then recent approach of Swadesh (1950), who had presented counts of cognates for different varieties of Salishan languages—an approach that represented the birth of glottochronology and lexicostatistics.

In this paper I will use a phonologial distance coming from lexical data, and I will not discuss measures from other types of linguistic data; the fact is that we presently only have sufficient coverage for the lexical domain. I will also leave the issue of mutual intelligibility measures, but it is worth mentioning that such measures actually have been shown to correlate well with counts of cognates on standardized word lists (Biggs 1957; Ladefoged, Glick, and Criper 1972; Bender and Cooper 1971).

2. Using the Normalized Levenshtein Distance (LDN)

The ASJP database (Wichmann, Holman, and Brown 2018) contains word lists for a 40-item subset of the Swadesh 100-item list from 7,655 **doculects** (language varieties as defined by the source in which they are documented). Stating how many languages this corresponds to would beg the question that interests us here, but if a unique ISO 639-3 code represents a unique language then the database can be said to represent around two-thirds of the world's languages. Only word lists are used that are 70% complete (i.e., having at least 28 words out of the 40) and represent languages recorded within the last few centuries. Creoles and pidgins are excluded. This leaves a sample of 5,800 doculects. Although the word lists are short, it has been shown by Jäger (2015), who also used the 70% completeness criterion for his selection of word lists from the ASJP database, that reliability as measured by Cronbach's alpha, following Heeringa et al. (2006), is sufficient for phylogenetic purposes. The word lists are transcribed in a simplified system called ASJPcode. The pros and cons of this system are discussed in Brown, Wichmann, and Holman (2013).

A linguistic distance measure that can be applied to the ASJP data is a version of the Levenshtein (or edit) distance, averaged over word pairs. The Levenshtein distance

[1] See `http://glottolog.org/glottolog/glottologinformation` (accessed 2019-07-01).

[2] Even a factor such as differences in ethnic background may affect perceived intelligibility, despite two people speaking the same language (Rubin 1992).

is the number of substitutions, insertions, or deletions required to transform one word into another. Given two word lists, one can measure the Levenshtein distance for each word pair divided by the length of the longer of the two words. The mean of these individual word distances may be called the LDN (Levenshtein distance normalized). A further modification is to divide the LDN by the average LDNs for words in the two lists that do not refer to the same concepts. This has been called LDND (Levenshtein distance normalized divided) (Wichmann et al. 2010). The second modification is intended primarily for comparisons of languages that are regarded as being unrelated, since it controls for accidental similarities due to similar sound inventories. For the present purposes of using ASJP for distinguishing languages and dialects, we are only interested in comparing genealogically related language varieties. Thus, we can resort to the simpler and faster LDN measure. Nevertheless, LDND measures will also be cited because they translate into years of separation of two related speech varieties (Holman et al. 2011). Both the LDN and the LDND are implemented in Wichmann (2019), which is used here for the following experiments.

3. Distinguishing Languages and Dialects by LDN

Before trying to find a value of LDN that might serve as a criterion for distinguishing languages and dialects, it is of interest to look at the distribution of LDNs for putative language vs. dialect pairs using the ISO 639-3 codes of Ethnologue (Simons and Fennig 2017). Figure 1 is a comparison of two boxplots, the one to the left showing the distribution of LDNs for doculects in ASJP belonging to the same ISO 639-3 code language and the one to the right showing the distribution of LDNs for doculects in ASJP not belonging to the same ISO 639-3 code language but belonging to the same genus (group of relatively closely related languages using the scheme of WALS [Dryer and Haspelmath 2013]).

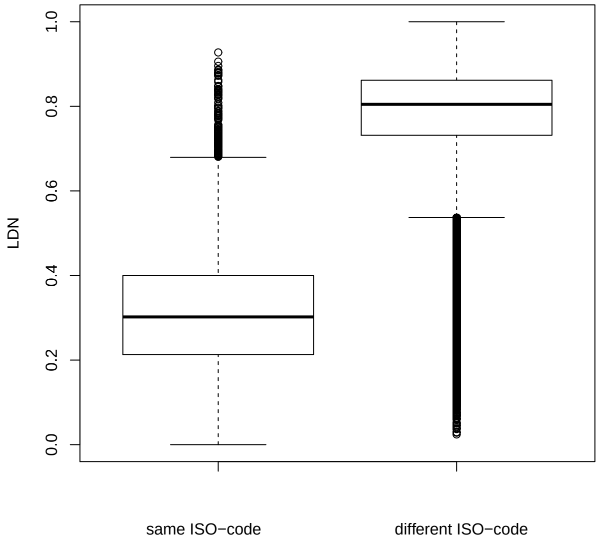

Figure 1
Boxplots of LDNs for ASJP doculects belonging to same vs. different ISO 639-3 codes (but same genera).

As expected, Figure 1 shows that same-ISO-code LDNs tend to be smaller than different-ISO-code LDNs. But we also see that there are many outliers where same-ISO-code LDNs are extremely large and different-ISO-code LDNs extremely small. No doubt, some of these outliers are due to ISO 639-3 codes that were misassigned either in the original sources used by ASJP transcribers or by these transcribers themselves. Even when the outliers are ignored, however, there is an overlap.

To be more consistent than Ethnologue, one could apply a certain cut-off value of LDN to distinguish languages and dialects. The obvious question then arises: Is there some way in which a non-arbitrary cut-off point can be found? In Wichmann (2010) it was suggested that language distances within families may have a multimodal distribution where distances typically belonging to dialects form a roughly normal distribution different from the also roughly normal distribution of distances between two different languages. In that paper the distances used were LDNDs and the example used for illustration came from the Uto-Aztecan family. Here I follow up on the idea by plotting LDNs for 15 language genera for which 10% or more of ASJP doculect pairs have the same ISO 639-3 code. Figure 2 shows a histogram of LDNs for each of the 15 genera. Overlaid on each histogram is a density curve (in black) and two curves (in red and green) fitting the data to a mixture of two normal distributions. These graphs were produced using the R package mixtools (Benaglia et al. 2009), specifically, the plotting method associated with the output of the *normalmixEM()* function. According to the documentation, this function implements "the standard EM algorithm for normal mixtures that maximizes the conditional expected complete-data log-likelihood at each M-step of the algorithm." The curves show that a bimodal distribution can either be manifested very distinctly (e.g., curves for Eleman Proper, Huitoto, Iranian, Mayan, Sama-Bajaw, South Sulawesi) or merely show up as a skew in the left tail of the distribution (e.g., Atayalic, Japanese), and some other curves are more difficult to interpret.

Although Figure 2 visually suggests that the vast majority of cases can be interpreted as a mixture of two normal distributions, we would like to verify this more exactly. Using the *boot.comp()* function of mixtools package, this is achieved by producing 100 bootstrap realizations of the likelihood ratio statistic for testing the null hypothesis of a k-component fit versus the alternative hypothesis of a $(k+1)$-component fit to a model of mixed normal distributions, applying the $p < 0.05$ threshold. The existence of from 1 to 4 components was tested in this way. The column carrying the header 'k' in Table 1 contains the results, which show a strong tendency for the preferred number of components to be two (12 out of 15 cases), or, in a few cases (3 out of 15), three components. In spite of the three cases where $k = 3$ yielded the best fit we can treat all 15 cases in a uniform way toward the objective of finding the LDN that separates the members of two distributions by looking only at the distances within the two distributions containing the lower values. Again the *normalmixEM()* function of Benaglia et al. (2009) is used. This outputs the parameters of the normal distributions, which allow one to identify the LDN value where the two normal distributions cross. Table 1 shows these LDN cut-offs rounded off to four decimals. It also shows the corresponding LDND value. LDND values corresponding to LDN were found through a linear regression using all 639,727 doculect pair distances analyzed in the present study. LDN and LDND are highly correlated ($r = .985$) and the formula for deriving LDND from LDN has the slope 1.00158 and intersect 0.08459. LDND values will become useful for interpreting the cut-off in terms of time depths (cf. next section).

The LDN cut-offs in Table 1 are relatively narrowly distributed. Calculating a 95% confidence interval around the mean of 0.5138 produces ± 0.0707. The rounded-off value of LDN = 0.51 is proposed here as a universal cut-off that may be used to distinguish pairs of dialects from pairs of languages. This distinguishing criterion is easily applied and it was arrived at by an entirely objective procedure that can be both replicated and revised on an evolving data set.

The sample of genera listed in Table 1 was chosen so as to ensure that each contains a good number of close varieties, applying the selection criterion that at least 10% of the doculect pairs should represent the same ISO 639-3 code. As the next section shows, this does not imply that the ISO standard comes to determine the results; it is just a way of filtering away genera that would be unsuitable for the present investigation. The selection subtly introduces another potential bias, however: For all of the genera some data are included that come from dialect surveys. It may well be that the researcher carrying out the survey, consciously or not, aimed at a certain resolution, and this resolution would have a major impact on the parameters of the normal distribution identified here as belonging to dialects. To control for such a potential bias a resampling

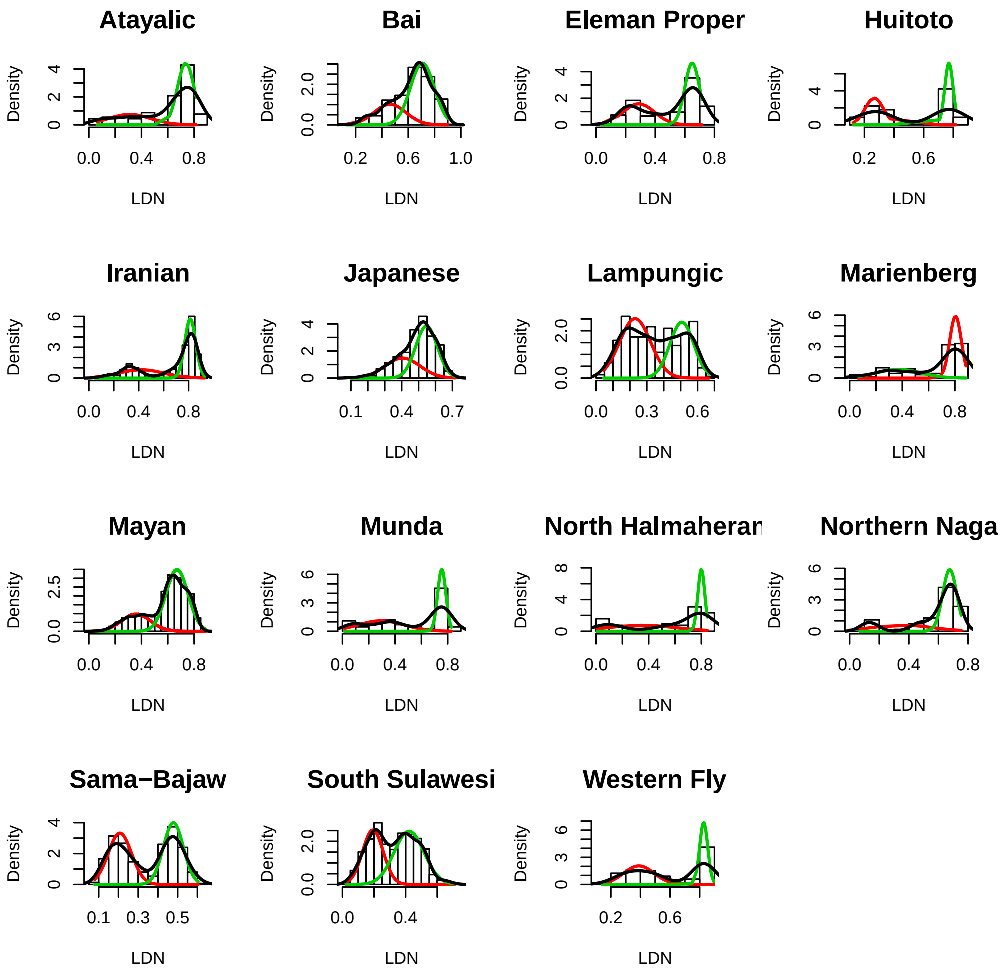

Figure 2
Density plots of LDNs for 15 genera for which 10% or more of doculect pairs' members pertain to one and the same ISO 639-3 code.

Table 1
Language groups, threshold for LDN assuming two normal distributions, corresponding LDND value, the number of components (*k*) as found by a bootstrap test, percent of pairs of language varieties belonging to same ISO-639-3 code, and the number of doculects in the data for each group (N).

Group	LDN cut-off	LDND	*k*	% same ISO-code	*n*
Atayalic	0.5770	0.6625	2	28.57	14
Bai	0.5551	0.6406	2	31.48	28
Eleman Proper	0.5195	0.6049	2	25.74	17
Huitoto	0.6108	0.6964	2	13.33	10
Iranian	0.5405	0.6259	3	12.25	73
Japanese	0.4534	0.5387	2	12.90	31
Lampungic	0.3882	0.4734	2	40.58	24
Marienberg	0.6845	0.7702	2	24.18	14
Mayan	0.5148	0.6002	3	13.98	106
Munda	0.6658	0.7514	2	12.00	25
North Halmaheran	0.2686	0.3536	3	24.17	16
Northern Naga	0.5627	0.6482	2	12.73	11
Sama-Bajaw	0.3511	0.4362	2	45.33	25
South Sulawesi	0.2870	0.3720	2	10.80	39
Western Fly	0.7275	0.8132	2	52.94	18

experiment was carried out where distances for 100 same-ISO-code pairs and 900 differerent-ISO-code pairs were sampled randomly (with replacement) from the total pool of distances pertaining to all genera. This was done 100 times. Each of these 100 "resampled genera" was subjected to the same analysis as the original genera, finding a cut-off between the two normal distributions having the smallest means. The result was a mean LDN threshold of 0.5686 ± 0.0072, that is, a range of 0.5614–0.5758. This is not inconsistent with the range 0.4431–0.5845 found for the unscrambled data, but is more narrow and lies toward its upper end. I interpret this result as suggesting that a future, more extensive sample of genera might lead to a somewhat higher and more narrowly defined threshold, perhaps around 0.57. In the interest of transparency and replicability, I still propose the directly measured threshold of 0.51, but with the qualification that this is a conservative estimate.

The question remains whether the resampling experiment really does away with any sampling bias. Do the findings perhaps reveal more about shared perceptions among linguists about where to draw borders between languages and dialects when sampling data than about real distributions? This point of critique is somewhat speculative and therefore hard to counter, but it may perhaps be addressed in future research through computer simulations free of sampling biases, using a framework such as that of Wichmann and Holman (2017).

4. Discussion

Going back to the Ethnologue classification (cf. Figure 1), we may wonder whether this classification tends to over- or underdifferentiate, assuming that LDN $= 0.51$ is a sensible cut-off point. The 5,800 doculects in our sample lead to 635,419 pairs whose members both belong to the same WALS genus and both of which carry ISO-codes. Out of all these pairs, 0.1% have the same-ISO-code and LDN > 0.51, while there are 3.1% pairs with different ISO-codes and LDN < 0.51. This means that if we pick a random pair of doculects from a given genus, the chance that Ethnologue will overdifferentiate, treating a dialect pair as a language pair, is more than 30 times greater than the chance that it will underdifferentiate, treating a language pair as a dialect pair. In short, Ethnologue tends to overdifferentiate, so the number of languages counted in this catalogue would be too high.

To make the results more palpable, some examples of pairs of speech varieties whose status as dialects or languages probably tend to be contested are supplied in Table 2. The examples are mere illustrations that are meant to help the reader interpret different LDN values, including the 0.51 cut-off.

Besides a possible solution to the perennial problem of distinguishing languages and dialects, this paper has yielded a result of potentially deeper importance for the understanding of language dynamics. We found (cf. Figure 2) that, given a sufficiently balanced sample of data from very closely related speech varieties and more distantly related ones, it is normally possible to discern a mixture of different distributions yielding peaks corresponding to characteristic means of what can be interpreted as, respectively, dialects and languages. The valleys between these peaks are a highly interesting phenomenon: They would seem to suggest that values around our LDN = 0.51 cut-off are atypical. This corresponds to a situation where we have a chain of dialects of one language and then a relatively abrupt transition to a neighboring chain of dialects of some other language. Such a situation characterizes national languages like German and Dutch, for instance, but, as seen in Figure 2, we also find it for minority languages around the world.

In Holman et al. (2011), it was shown that the twice-modified Levenshtein distance LDND translates into a time separation between the language varieties compared. The

Table 2
LDN values for pairs of speech varieties prone to turn up in discussions about how to distinguish languages and dialects. Language names are given as in ASJP. From the top down to and including East and West Greenlandic, the pairs constitute dialects of one and the same language according to the LDN = 0.51 cut-off proposed here. The rest of the pairs constitute different languages.

Speech variety A	Speech variety B	LDN
Indonesian	Malay	0.1199
Bosnian	Croatian	0.1324
Quechua Chachapoyas	Quechua Huaylas Ancash	0.3055
Hindi	Urdu	0.4281
Classical Nahuatl	Pipil	0.4336
Standard German	Bernese German (Switzerland)	0.4638
Russian	Belarusian	0.4647
Danish	Swedish	0.4921
East Greenlandic	West Greenlandic	0.5036
Navajo	Jicarilla Apache	0.5708
Cairo Arabic	Moroccan Arabic	0.5814
Dongshan Chinese	Fuzhou Chinese	0.6013
Catalan	Spanish	0.6589
Japanese	Miyako (Ryukuan)	0.6680

LDND values in Table 1 have a mean of 0.5992 ± 0.0709. Using the formula of Holman et al. (2011) this is equivalent to a range of 1,075–1,635 years. Thus, it takes around one to one-and-a-half millennia for a speech community to diverge into different language groups.

Why can we find those valleys in the distributions in Figure 2? Or, put differently, what is it about the way that speakers interact that allows us to distinguish languages and dialects? A possible explanation is that there is a threshold of mutual intelligibility where language varieties will influence one another if they are below the threshold but will cease to influence one another if they are above it. If mutual intelligibility between variety A and B is impeded completely, speakers may take recourse to just the more prestigious of the two, if not some third language, leaving A and B to drift apart more rapidly than would the case if both A and B were used for communication between the two groups. It awaits future studies to corroborate this idea through modeling and, ideally, through systematic sampling of both lexical data and data on intelligibility from the cells of a large but also fine geographically defined grid or, less ideally, through an analysis of the literature on mutual intelligibility.

5. Conclusion

In this paper the question of how to distinguish languages and dialects was addressed by studying the distribution of lexical distances within groups of uncontroversially related languages (the genera of Dryer and Haspelmath [2013]). Following up on an idea tentatively suggested in Wichmann (2010), it was verified that distances among speech varieties represent mixed distributions, including a cluster that may be said to correspond to dialects and another cluster corresponding to languages. Applying an expectation-maximization algorithm to tease apart the mixture of normal distributions across a sample of 15 language groups, the average cut-off point between the two distributions was found to be LDN = 0.51, where LDN is the normalized Levenshtein distance across word pairs in the ASJP 40-item word lists of Wichmann, Holman, and Brown (2018). The corresponding temporal distance lies around 1,355 years, within the interval 1,075–1,635 years. Thus, we now have a principled way of distinguishing languages and dialects. A tantalizing question for future research is why there seems to be a real distinction, not just a theoretical or arbitrary one. Some suggestions for ways to approach this question were suggested.

Acknowledgments

Thanks go to Eric Holman, Harald Hammarström, Qibin Ran, and four anonymous reviewers for stimulating discussion. The research was carried out under the auspices of the project "The Dictionary/Grammar Reading Machine: Computational Tools for Accessing the World's Linguistic Heritage" (NWO proj. no. 335-54-102) within the European JPI Cultural Heritage and Global Change programme. It was additionally funded by a subsidy from the Russian government to support the Programme of Competitive Development of Kazan Federal University and a grant (KYR17018) from Beijing Language Innovation Center in support of a sub-topic directed by Qibin Ran.

References

Agard, Frederick. 1984. *A Course in Romance Linguistics*. Georgetown University Press.

Benaglia, Tatiana, Didier Chauveau, David R. Hunter, and Derek S. Young. 2009. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

Bender, Marvin L. and Robert L. Cooper. 1971. Mutual intelligibility within Sidamo. *Lingua*, 27:32–52.

Biggs, Bruce. 1957. Testing mutual intelligibility among Yuman languages. *International Journal of American Linguistics*, 23(2):57–62.

Brown, Cecil H., Søren Wichmann, and Eric W. Holman. 2013. Sound correspondences in the world's languages. *Language*, 89(1):4–29.

Casad, Eugene H. 1974. *Dialect Intelligibility Testing*. Summer Institute of Linguistics of the University of Oklahoma, Norman.

Dryer, Matthew S. and Martin Haspelmath, editors. 2013. *The World Atlas of Language Structures Online*. http://wals.info. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Gooskens, Charlotte and Vincent J. van Heuven. 2019. How well can intelligibility of closely related languages in Europe be predicted by linguistic and non-linguistic variables? *Linguistic Approaches to Bilingualism*, Early online publication, https://doi.org/10.1075/lab.17084.goo.

Gooskens, Charlotte, Vincent J. van Heuven, Jelena Golubovic, Anja Schüppert, Femke Swarte, and Stefanie Voigt. 2018. Mutual intelligibility between closely related languages in Europe. *International Journal of Multilingualism*, 15(2):169–193.

Gooskens, Charlotte and Cindy Schneider. 2016. Testing mutual intelligibility between closely related languages in an oral society. *Language Documentation and Conservation*, 10:278–305.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath. 2017. *Glottolog 3.0*. http://glottolog.org. Max Planck Institute for the Science of Human History, Jena.

Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In *Proceedings of the Workshop on Linguistic Distances*, pages 51–62, Sydney.

Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List, and Dmitry Egorov. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6):841–875.

Jäger, Gerhard. 2015. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences of the U.S.A.*, 112(41):12752–12757.

Korjakov, Yurij Borisovich. 2017. Problema "jazyk ili dialekt" i popytka leksikostatisticheskogo podxoda. *Voprosy Jazykoznanija*, 6:79–101.

Ladefoged, Peter, Ruth Glick, and Clive Criper. 1972. *Language in Uganda*. Oxford University Press.

Rubin, Donald L. 1992. Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, 33(4):511–531.

Simons, Gary F. and Charles D. Fennig. 2017. *Ethnologue: Languages of the World, Twentieth Edition*. SIL International, Dallas, TX.

Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics*, 16(4):157–164.

Szeto, Cecilia. 2000. Testing intelligibility among Sinitic dialects. In *Proceedings of ALS2K, the 2000 Conference of the Australian Linguistic Society*, Melbourne.

Voegelin, Charles F. and Zellig S. Harris. 1951. Methods for determining intelligibility among dialects of natural languages. *Proceedings of the American Philosophical Society*, 95(3):322–329.

Whaley, Lindsay J., Lenore A. Grenoble, and Fengxiang Li. 1999. Revisiting Tungusic. *Language*, 75(2):286–321.

Wichmann, Søren. 2010. Internal language classification. In Luraghi, Silvia and Vit Bubenik, editors, *The Continuum Companion to Historical Linguistics*. Continuum Books, London/New York, pages 70–86.

Wichmann, Søren. 2019. Interactive R program for ASJP version 1. https://github.com/Sokiwi/InteractiveASJP01.

Wichmann, Søren and Eric W. Holman. 2017. New evidence from linguistic phylogenetics supports phyletic gradualism. *Systematic Biology*, 66(4):604–610.

Wichmann, Søren, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010. Evaluating linguistic distance measures. *Physica A*, 389:3632–3639.

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown, editors. 2018. The ASJP Database (version 18). http://asjp.clld.org/.