https://doi.org/10.5817/LB2024-37185

Alexander Maxwell – Louise McMillan

ERROR BARS FOR LEXICOSTATISTICAL ESTIMATES, WITH A CASE STUDY COMPARING THE DIVERSITY OF CHINESE AND ROMANCE

Abstract

This paper applies statistical techniques for measuring sampling error to lexicostatistics, a field in which error has often been discussed, but only rarely measured. We specifically calculate a margin of error for lexicostatistical comparisons based on Swadesh-type vocabulary lists, and use chi-squared tests to estimate a minimum threshold for when two lexicostatistical measurements will be statistically significantly different from one another. The article includes charts which mathematically unsophisticated scholars can easily use to check margins or error. We use margin of error calculations to test the claim that the relative internal diversity of Romance "languages" and Chinese "dialects" is equivalent, finding that no result is possible with extant lexicostatistical studies. We end by suggesting that lexicostatistical dendrograms depict uncertainty with "fat branches," that is, branches whose width corresponds to statistical uncertainty.

Keywords

lexicostatistics; margin of error; dendrograms; fat branches; Romance; Chinese

1 Introduction

This article proposes accounting for the uncertainty inherent in lexicostatistical comparisons with margins of error and significance thresholds. While the statistical theory we suggest is well-understood by mathematicians, we supplement our findings with reference charts, such that even scholars unfamiliar with statistical analyses can easily calculate a margin of error. Scholars wishing to indicate the uncertainty in a single lexicostatistical measurement can look up the appropriate margin of error in the charts provided. Scholars wishing to compare two lexicostatistical measurements take the larger of the two measurements as the baseline and consult the appropriate chart to find the significance threshold. We then illustrate our technique with a case study: drawing on extant lexicostatistical studies, we test whether the relative internal diversity of Romance "languages" is comparable to the internal diversity of Chinese "dialects." We end with suggestions for depicting statistical uncertainty in dendrograms.

Lexicostatistics is fundamentally a technique for quantifying linguistic similarity, which implies the quantification of linguistic distance. Quantitative measures of linguistic similarity/difference have many possible applications. Sheila Embleton (2015, 23) suggests lexicostatistical measurements can show "how closely two languages are related," or alternatively "whether they are related at all." Lexicostatisticians can use measurements of related varieties to construct dendrograms or "tree diagrams," (Geisler – List 2010; 2013); George Starostin (2013, 126–27) once opined that "the main result of every lexicostatistical analysis is a phylogenetic tree (or network)." Lexicostatistical dendrograms in turn shed light on human prehistory, e.g. by revealing migration patterns (e.g. Dyen 1962; Gray – Atkinson 2003). Numerous scholars have proposed lexicostatistical techniques for sorting varieties into the categories "languages" and "dialects" (Swadesh 1954; Wurm – Laycock 1961; Dyen et al. 1992, 9; Koryakov 2017; Wichmann 2020); or indeed into series of other collectives: families, stocks, phyla, and so forth (Swadesh 1954, 321; McElhanon 1971, 134; Crowley 1992, 170). Missionaries affiliated with the Summer Institute of Linguistics (SIL) routinely gather lexicostatistical data to maximize the potential audience of Bible translations, using lexicostatistical thresholds as estimates of mutual intelligibility.

Most of these applications involve comparing several lexicostatistical measurements. The lexicostatistical similarity/distance between two varieties A and B holds little interest in isolation. Investigators are instead asking how the lexicostatistical similarity/distance between varieties A and B compares to that between varieties A and C, or between varieties C and D, and so forth. Which varieties are more closely related?

The centrality of comparison in lexicostatistical analysis raises the possibility of statistical error. If two lexicostatistical measurements are similar, how can we have confidence that a greater measurement actually implies a greater distance? Might one measurement merely appear greater as the result of a statistical fluke? Constructing dendrograms, furthermore, requires that scholars have great confidence in their data. Is such confidence justified? Understanding the many possible sources of error requires a brief summary of lexicostatistics as a technique.

The basic idea of lexicostatistics is to compare linguistic features and count similarities and differences. Any sort of linguistic feature can theoretically be compared. In practice, however, most lexicostatistical studies examine word lists, because phonetic or grammatical features important in some varieties may not exist in others.

How do lexicostatisticians construct word lists? Historically, most have used the so-called "Swadesh lists," developed in the 1950s by lexicostatistician and glottochronologist Morris Swadesh. Swadesh initially proposed a list of 225 words, but later developed shorter lists of 215 words, 200 words, and 100 words (Swadesh 1952; 1955; see also Hymes 1960, 3–5; Oswalt 1971, 421–34; McMahon – McMahon 2005, 34–44). Computational linguists Uri Tadmor, Martin Haspelmath and Bradley Taylor (Tadmor et al. 2010) subsequently introduced the "Leipzig-Jakarta list" of 100 words supposedly less resistant to borrowing. Russian sinologist Sergei Yakhontov, whose work is mostly known though the mediation of Sergei Starostin, chose a list of 35 words on the basis of "stability;" 32 of his words also appear in the Swadesh 100-word list (Starostin 1995, 90–91; Zhuravlev 1994, 35–36). Aharun Dolgopol'sky (1986, 620–628) used a list of 15 words he thought particularly resilient to borrowing. A few lexicostatisticians, finally, examine texts instead of word lists: Witold Mańczak (2009), for example, compared Biblical passages.

The use of a finite word list inevitably creates sampling error, which can significantly affect lexicostatistical estimates (Feld – Maxwell 2019). Comparing different wordlists, for example, leads to different lexicostatistical estimates. István Fodor's study of Slavic, to give a specific example, found a maximum 93% similarity between Czech and Polish using the Swadesh 100-word list and including possible synonyms, but a minimum of 88% similarity using the Swadesh 215 list and excluding possible synonyms (Fodor 1961, 303, 304). Over 22 measurements, Fodor found an average discrepancy between the two lists of 1.4%.

How do lexicostatisticians compare word lists, once created? There are many different techniques (Embleton 1986; Heggarty et al. 2011). Lexicostatisticians initially compared the percentage of shared cognates, also known as the "cognancy ratio." Recognizing cognates is not an exact science: Serva – Petroni (2008) found it "often a matter of sensibility and personal knowledge" in which "subjectivity plays a relevant role." Such subjective processes, obviously, generate statistical error. In a study of Bantu languages, for example, David Olmsted (1957, 839–40) classified word pairs as "cognate," "non-cognate," or "ambiguous." By including or excluding the ambiguous cases, Olmstead measured distances not as a single percentage, but as a range with a minimum and a maximum estimate. He found, for example, that Herero and Swahili shared somewhere between 29% and 41% of their vocabulary.

With the advent of computers, many lexicostatisticians have abandoned cognate recognition and employed instead the normalized Levenshtein distance. The Levenshtein distance, also known as the edit distance, based on an algorithm developed by Russian computer scientist Vladimir Levenshtein (1965; 1966), is the minimum number of changes to transform one line of characters into another.

It can be normalized to a percentage by dividing the edit distance by the number of characters in the longer word (Maguire – McMahon 2011, 108; Wichmann 2020).

Applying the Levenshtein algorithm in practice still requires judgment calls. Greenhill noted, for example, that “differences in orthographies might play a crucial role in the accuracy of the Levenshtein distance.” Greenhill argued that “standardized orthography” increased the reliability of Levenshtein classifications from 41.3% “to around 65%,” suggesting that orthographic effects indeed have statistically significant consequences (Greenhill 2011, 692). Several scholars transcribe their word lists into the international phonetic alphabet (IPA), which distinguishes over a hundred phonemes, most commonly through a computer-friendly version developed by John Wells (1994). A team of scholars based in Leipzig, furthermore, collapse related sounds into a single symbol, using e.g. the same character for any “high and mid central vowel, rounded and unrounded (IPA ɨ, ɘ, ə, ɜ, ʉ, ɞ, ɵ)”. The Leipzig team has also done calculations based exclusively on consonants, equating e.g. the character strings “buk” and “bek” on the grounds that both share the pattern “b (vowel) k” (Brown et al. 2008, 288, 289, 306, 307).

In short, lexicostatistics has many possible sources of error. Serious lexicostatisticians have extensively pondered the possible sources of lexicostatistical error. Scholars have also proposed many different techniques to make lexicostatistical data more robust (Gudschinsky 1956; Hymes 1960; Embleton 1986; Heggarty 2010).

We suggest, however, that lexicostatisticians have not always paid sufficient attention to the *quantification* of error. Swadesh (1955, 124) admitted that lexicostatistics lacked “the accuracy of a precision instrument” and claimed for it only “considerable approximate validity.” Paul Heggarty (2010, 307) characterized lexicostatistics as “generally viable, if rather blunt.” Warren Maguire and April McMahon (Maguire – McMahon 2011, 16–17) judged Levenshtein calculations “rather crude,” but found them “effective for measuring the distance between varieties.” Dyen et al. (1992, 8) declared that various statistical uncertainties would “change the percentages only slightly,” while Pereltsvaig – Lewis (2015, 89) thought such uncertainties would “have a grave effect” that “cannot be ignored.” Lexicostatistical data have also been declared “approximately correct,” and “correct in a rough and ready sense (that is, on the average)” with “a certain rough and ready validity” (Dobson et al. 1972, 207). We find these qualitative descriptions unsatisfactory. How approximate is approximate? How grave is “grave”? How rough is ready? Only when a source of error is quantified, we suggest, can we know whether or not it can be ignored.

In this discussion, we pay particular attention to quantifying the sampling error, as distinguished from measurement error. Measurement error concerns the comparison of two lexical items, and encompasses things like the difficulty of recognizing cognates, or the difficulty of standardizing orthography to apply the Levenshtein algorithm. Sampling error arises from the selection of words to compare. One can theoretically avoid sampling error by comparing all possible items in the sample, but in practice gathering all words from two different varieties is impossible. No word list can pose as a complete lexicon, since counting the total number of words poses insurmountable difficulties of definition (Kornai 2002). Furthermore, some words in one variety have no exact equivalents in other varieties.

Sampling error can be minimized. The normal procedure for reducing sampling error is to increase the sample size. Lexicostatisticians, however, fear that increasing the sample size would increase the measurement error, since it would increase the share of loanwords or neologisms. A longer word list that minimizes sampling error might therefore increase the total error. Lexicostatisticians have not found a solution to this dilemma.

If sampling error is inevitable, however, the question arises: how large is it? This paper provides an easy technique for estimating sampling error. We have calculated the margin of error for different thresholds of statistical reliability, and given the result in a series of charts so that scholars may visually display sampling error with error bars. Future scholars may have further ideas about how to quantify measurement error, or other sorts of error. Our primary aim, however, is to suggest that lexicostatisticians should not content themselves with efforts to minimize error. They should consider how to quantify it, and how to depict uncertainty in their results.

2 Estimating Lexicostatistical Sampling Error

We are not the first scholars attempting to quantify lexicostatistical sampling error. Taking the advice of “A. T. James of Yale University, a mathematical statistician,” the prolific lexicostatistician Isidore Dyen (1962, 42) applied “a combined Chi-squared test at the 5% level” to pairs of lexicostatistical measurements, finding that “a difference of about 10% between two percentages is necessary to produce a satisfactory result.” He warned that the 10% threshold could only “be regarded as a rule-of-thumb.” Indeed, in his subsequent work, Dyen (1975, 113) proposed “as a rule of thumb” that lexicostatistical percentages were only “significantly different if their difference is 9.5% or greater.”

Dyen’s “rule of thumb” method, however, is not a satisfactory method for estimating statistical significance in lexicostatistics. Indeed, Dyen himself noted in a footnote that his rule of thumb is not a substitute for the chi-squared test. The main problem is that the margin of error for a given percentage measurement varies: the error differs significantly depending on how close that percentage is to 50%.

We can illustrate the change in error with a pair of examples. Let us say that we have three varieties, A, B and C and that we wish to compare the margin of error at the 95% confidence level. For this measurement, the margin of error can be calculated using the standard formula for proportions: $\pm 1.96 \times \sqrt{\hat{p}\,(1 - \hat{p})\,/\,n}$, where $\hat{p}$ is the estimated percentage, and n is the sample size. If we compare A and B using the Swadesh 100-word list and find that if 95% of the vocabulary is cognate, then $\hat{p}$ is 95%, and n is 100, so the margin of error will be $\pm 1.96 \times \sqrt{.95(1 - .95)\,/\,100}$, = 4.3%. But if we then compare A and C using the Swadesh 100-word list and find that 85% of vocabulary is cognate, then the margin of error will be 7.0%.

Now let us consider the process of comparing pairs of measurements, to determine whether they are significantly different from one another. For the A–B and A–C cognate measurements, we can compare them using the chi-squared test, as Dyen proposed, by treating them as a contingency table.

	Variety B	Variety C
Cognates with Variety A (%)	95	85
Non-cognate with Variety A (%)	5	15

We can apply the chi-squared test to this table using standard statistical software. The test rests on the assumption that the true proportions are the same in both columns and that any differences we see are just random variation in the observed values. In this case, the initial assumption for the test would be that Variety B and Variety C are both 90% cognate to Variety A, and that the observed 95% and 85% measurements are just due to random variation in the data. The test compares the observed values to what we would expect under that assumption. If the observed values are sufficiently different from that expectation, we conclude that the difference observed cannot actually be due to random variation and there is some underlying difference in the two measurements. In this case, if we apply the chi-squared test at the 5% significance level (which is equivalent to 95% confidence for the margin of error), we find that the observed values are unlikely to be due to chance, and that there is a statistically significant difference between how close A is to B and how close A is to C. For this case, Dyen's rule of thumb holds: a difference of 10% between the two measurements is statistically significant.

Now consider a further pair of varieties, D and E. We compare A and D using the Swadesh 100-word list and find that 55% of the vocabulary is cognate; we then compare A and E using the Swadesh 100-word list and find that 45% of the vocabulary is cognate. Using a similar contingency table and chi-squared test as before, we find that at the 5% level there is no significant difference between how close A is to D and how close to A is to E, even though these two measurements are also 10% apart. In this case, Dyen's rule of thumb fails.

Both of these situations can be illustrated with concrete examples by substituting A, B, C, D with some sample Indo-European varieties. Using the Swadesh

100-word list and excluding all doubtful cases, Fodor (1961, 301) found that Czech and Slovak share 95% of their vocabulary. With 95% confidence, we can say that a plausible value for the actual percentage of cognates for Czech and Slovak is any value in the range between 90.5% and 99.5% Using the same method, Fodor (1961, 302) also found that Russian and Czech shared 85% of their vocabulary, so plausible values for the actual percentage of cognates for Russian and Czech would fall in a range between 78% and 92%. Fodor found a smaller distance between Czech and Slovak than between Czech and Russian, but applying the chi-squared test shows with 95% confidence that the smaller distance is actually statistically significant. Dyen et al. (1992, 107), meanwhile, measured the similarity between Pennsylvania Dutch and English as 55.3%, between Pennsylvania Dutch and Swedish as 45.3%. In this case, the smaller difference is not statistically significant.

It is feasible to calculate the smallest difference between A–D and A–E cognancy measurements that would be statistically different. For each value of A–D, this minimum significant distance defines a threshold value at which A–E becomes significantly different. The threshold values depend on the confidence ratio and the sample size. We have therefore prepared different charts for five different confidence ratios. We have also prepared charts for two different sample sizes, 100 and 200, facilitating measurements made using the Swadesh 200-word list, the Swadesh 100-word list, and the Leipzig-Jakarta list (which also has 100 items).

We have designed our charts for scholars without a deep understanding of statistics. For a margin of error chart, scholars take the measurement they are assessing, and read down the leftmost column to find the appropriate row. Select the chart value closest to the lexicostatistical measurement in question. An approximate margin of error is given.

For the significance threshold chart, scholars take the larger of the two measurements as the baseline and read down the leftmost column to find the appropriate row. If the higher measurement is *x*, any value lower than *y* will be significantly different from *x*. For entries in the significance thresholds table marked as "no significant differences," which often occur at the bottom of the chart, the larger measurement is small and the smaller measurement therefore even smaller. In such cases, the two measurements will never be significantly different from each other.

Our charts enable scholars to find a margin of error for five different levels of statistical reliability: 5% significance, the two-sigma standard, the three-sigma standard, the four-sigma standard and, most stringently, the five-sigma standard physicists use as the standard for "discovery claims" at the Large Hadron Collider (Lyons 2013). Some linguists draw inspiration from physics; indeed, Chomsky (1988, 172) once expressed the hope that linguistics "can be reduced to physics." Lexicostatisticians who take physics as their model can consult the five-sigma column of the chart, but the results are not very encouraging. Dyen, Kruskal and Black (Dyen et al. 1992, 107), using the Swadesh 200-word list, measured the distance between standard (Stockholm) Swedish and German as 69.5%%, and the distance between standard Swedish and Danish as 87.4%. At the five-sigma confidence level (which corresponds to a confidence level of 99.99997), the lower distance measurement would have to be lower than 65.9% to be statistically significant. At the five-sigma confidence level, in short Dyen, Kruskal and Black's lexicostatistical data cannot decide whether Swedish is closer to Danish or to German. The width of the error bars illustrates just how far lexicostatistical data fall short of the accuracy standards in particle physics.

Scholars prepared to accept a lower standard of significance will naturally enjoy smaller error bars. The 95% confidence level for margin of error, i.e. a 5% significance threshold, is popular in several academic disciplines. At this lower confidence level, the figures of Dyen et al. (1992, 107) suggest that the difference between standard Swedish and German is significantly larger than that between standard Swedish and Danish Swedish. Nevertheless, even at this less rigorous confidence level, the lower distance measurement must be lower than 79.7% to be significant. The measured distances from Swedish to (Riksmål) Norwegian is 84.2%, and from Swedish to Faroese as 80.0%. So even at the relatively undemanding 95% confidence level, Dyen, Kruskal and Black's data cannot say whether standard Swedish is closer to Danish, Norwegian, or Faroese.

3 Case study: Romance and Chinese

To illustrate a more complex use of this chart, consider whether the "dialects" of Chinese are more or less internally diverse than the Romance "language family." The terms "dialect" and "language family," of course, anticipate a particular answer: they imply a greater diversity among Romance and a greater similarity among Chinese. To avoid pre-judging the results, we will speak only of comparing "Chinese varieties" and "Romance varieties." So, which are more diverse, Chinese varieties or Romance varieties?

Several scholars have pondered this naïve question, typically declaring Chinese and Romance varieties to be equivalently diverse. Chomsky, for example, has twice made the comparison in order to ridicule the language-dialect dichotomy. In a 1977 conversation with French scholar Mitsou Ronat, he reasoned as follows:

> "Why is 'Chinese' called a language and the Romance languages, different languages? The reasons are political, not linguistic. On purely linguistic grounds, there would be no reason to say that Cantonese and Mandarin are dialects of one language while Italian and French are different languages." (Chomsky 1977, 195; English translation from Chomsky 1979, 190).

A decade later, Chomsky's *Knowledge of Language: Its Nature, Origin and Use* contained a similar *reductio ad absurdum*:

> We speak of Chinese as 'a language,' although the various 'Chinese dialects' are as diverse as the several Romance languages. We speak of Dutch and German as two separate languages, although some dialects of German are very close to dialects that we call 'Dutch' and not mutually intelligible with others that we call 'German.' …. That any coherent account can be given of 'language' in this sense is doubtful" Chomsky 1986, 27).

Since Chomsky twice invoked the Chinese-Romance example, it evidently played some role in his thinking. In neither passage, however, did he adduce any actual evidence: his argument rests on simple assertion. Chomsky provided neither measurements of "diversity," nor a definition of how it might be measured. Nevertheless, we can attempt to fact-check Chomsky's assertion using extant lexicostatistical evidence.

To the best of our knowledge, no single lexicostatistical study has ever included both Chinese varieties and Romance varieties. Nevertheless, several studies have analyzed Chinese and Romance using the Swadesh lists. Indeed, for both Romance and Chinese, we have found studies using both the Swadesh 100-word list and the Swadesh 200-word list.

Yude Wang (1960, 91, 103) used the Swadesh 200-word list to calculate lexicostatistical distances for five varieties of Chinese, Mandarin, Wu, Min, Yüeh, and Hakka. Wang specifically proposed that "the dialects of Peking [Beijing 北京] (=P), Su-chou [Suzhou 苏州] (=S), Amoy [Xiamen 厦门] (=A), Canton [Guangzhou 廣州] (=C) and Moiyan [Meixian 梅州] (=M) can be considered respectively as a standard dialect" capable of meaningful comparison. Aware that cognate recognition is not an exact science, Wang provided a lower and upper bound of similarity by both excluding and including doubtful cognates. About 30 years later, Xu (1991, 422) used the Swadesh 100-word list to calculate lexicostatistical distances for seven varieties of Chinese; his 21 distance measurements include 10 that correspond to Wang's. Xu did not acknowledge any doubtful cases, his figures are straightforward integer percentages. Table 1 compares Wang's lower and upper measurements with Xu's measurements between five selected Chinese varieties.

Romance languages have also attracted lexicostatistical attention. John Rea (1958) used the Swadesh 100-word list to calculate 27 selected differences between eight Romance varieties: Catalan, French, Italian, Portuguese, Sardinian, Spanish, Rhaeto-Romance and Romanian. His 27 measurements do not actually cover all 28 total possibilities; he neglected to calculate a distance between Rhaeto-Romance and Catalan. Isidore Dyen, Joseph Kruskal, and Paul Black (Dyen et al. 1992, 19, 33, 103), using the Swadesh 200-word list, also included Romance varieties in a book-length study comparing no less than 84 Indo-European varieties, giving results for all 3,486 possible pairs. Rea listed simple integer percentages; Dyen, Kruskal and

Black give percentages with one digit after the decimal point. Table 2 summarizes their measurements for five selected Romance varieties: Portuguese, Spanish, French, Italian and Romanian.

Table 1. Lexicostatistical similarity between selected Chinese varieties

	WANG (lower-upper) Swadesh 200 (1960)	XU Swadesh 100 (1991)
Xiamen-Beijing	48.88 – 51.56	56
Xiamen-Suzhou	51.40 – 54.12	59
Xiamen-Guangzhou	55.31 – 56.77	63
Xiamen-Meixian	58.56 – 59.90	58
Meixian-Suzhou	63.10 – 64.43	73
Meixian-Beijing	63.78 – 65.10	69
Meixian-Guangzhou	69.70 – 70.53	79
Guanzhou-Beijing	70.16 – 70.77	74
Guangzhou-Suzhou	70.27 – 71.05	77
Beijing-Suzhou	72.73 – 73.47	73
Average distance between Chinese varieties	**62.4 – 63.8**	**68.1**

Table 2. Lexicostatistical similarity between selected Romance varieties

	DYEN et al. Swadesh 200 (1992)	REA Swadesh 100 (1958)
French-Spanish	73.4	75
French-Italian	80.3	89
French-Romanian	57.9	75
French-Portuguese	70.9	75
Italian-Portuguese	77.3	85
Italian-Spanish	78.8	82
Italian-Romanian	66	77
Spanish-Romanian	59.4	71
Spanish-Portuguese	87.7	89
Romanian-Portuguese	62.9	72
Average distance between Romance varieties	**71.4**	**79.0**

The various distances, of course, permit cherry picking: the lexicostatistical difference Xu measured between Guangzhou and Suzhou (77%) is identical to the distance Rea measured between Italian and Romanian. On the other hand, the lexicostatistical difference Xu measured between Xiamen and Beijing (56%) differs dramatically from the distance Rea measured between Spanish and Portuguese (89%). Chomsky's argument, however, adduced the internal diversity of Romance and Chinese as a whole. We suggest that the average distances given at the bottom of the charts are more relevant than any individual measurements.

Do Tables 1 and 2 support Chomsky's assertion? At first glance, the average distances appear to contradict Chomsky's claims in a way that might actually have strengthened Chomsky's argument. Chomsky assumed that the diversity between Romance varieties and between Chinese varieties was essentially the same, concluding that while consensus opinion among linguists views Romance varieties as "languages" and Chinese varieties as "dialects," linguists actually ought to classify them identically. According to the lexicostatistical measurements shown in Figures 1 and 2, however, Chinese "dialects" are actually *more* different from each other than the Romance "languages." The scholarly consensus is at even greater variance with lexicostatistical measurements than Chomsky assumed: if anything, linguists ought to speak about Romance "dialects" and Chinese "languages."

All lexicostatistical measurements, however, are subject to sampling error. Consider Rea and Xu, who both used the Swadesh 100-word list. The average of Rea's Romance measurements is 79%, and the average of Xu's Chinese measurements is 68%. Taking 79% as the higher measurement and rounding up, we see from the chart that at the five-sigma confidence threshold, the difference between Xu's measurements and Rea's measurements would be statistically significant only if Xu's average measurement were below 44.7%. Even if we satisfy ourselves with a two-sigma confidence level, the statistically significant difference appears only when Xu's average measurement is 66.5% or lower. Put another way, the significance threshold at the five-sigma confidence level is 36.3 percentage points, and the significance threshold at the two-sigma confidence level is 14.5 percentage points. The difference between Rea and Xu's measurements is below the significance threshold at either level of confidence. Thus there is no real way to say whether Chinese varieties are more or less diverse than Romance varieties.

4 Fat Branches: Depicting Error in Lexicostatistical Dendrograms

Double-digit significance thresholds for lexicostatistical differences raise questions about the validity of lexicostatistical dendrograms (tree diagrams). Lexicostatisticians who construct dendrograms have admittedly already taken steps to highlight statistical uncertainty. In their 2003 article, for example, Gray and Atkinson expressed confidence percentages on the branches of their dendrogram (Gray – Atkinson 2003). Feld and Maxwell have proposed three different methods for signaling statistical insignificance on dendrograms (Feld – Maxwell 2019, 114).

Scholars presenting dendrograms nevertheless rely too heavily on written caveats in explanatory text. Nicholls – Gray (2006, 168), for example, presented dendrograms with the caveat that "inferring a single tree will be misleading," since "there will always be uncertainty in the topology and branch lengths." They do not, however, depict error in their dendrograms themselves. They present one of their dendrograms with an explanatory caption warning that it "does not constitute our 'result'." Readers of scholarly journals, however, do not always read captions: publishing a dendrogram runs an inevitable risk that readers will mistake it for a result. We are also unsure why Nicholls and Gray would choose to publish a dendrogram that they did not consider a result.

Our main observation, however, is that even error-sensitive lexicostatisticians tend to depict the branches of dendrograms as widthless lines. The inevitability of error, we suggest, calls for visual representation. We suggest that lexicostatistical error can be depicted by adding width to dendrogram branches. The greater the error, the fatter the branch. Drawing on Dyen, Kruskal, and Black's lexicostatistical data for Indo-European languages, we have constructed some sample dendrograms with fat branches.

To illustrate the idea at its most basic, Figure 1 shows the divergence between German and Italian at three levels of confidence. Dyen, Kruskal and Black measured the cognancy ratio between German and Italian as 26.5% (Dyen et al. 1992, 106). Using the charts for a single lexicostatistical measurement, the chart depicts an error bar at the 95% confidence level, the three-sigma confidence level, and the five-sigma confidence level. The width of the branches is the width of the margin of error: the higher the confidence level, the fatter the branches.

Fat branching illustrates the advantages of a larger sample size. For the German/Italian measurement, Dyen, Kruskal and Black used the Swadesh 200-word list, and the error bar at the five-sigma confidence level is ± 15.0. Had Dyen, Kruskal and Black calculated their measurements from the Swadesh 100-word list instead, the margin of error at five-sigma confidence would have been ± 21.7%, nearly a quarter of the total chart.

Figure 1 only depicts sampling error: it does not account for any error arising from the processes of cognate recognition, Levenshtein orthographic transcription, and so forth. Indeed, the chart may underestimate sampling error. Dyen, who from the team Dyen, Kruskal and Black took responsibility for cognate recognition, excluded lexical items in any case "when a single dialect offers two different words for one meaning and these words are members of different cognate sets." While Dyen, Kruskal and Black assured readers that "this was not common," they did not quantify how common or uncommon, nor indicate which varieties were affected (Dyen et al. 1992, 20). If either German or Italian were affected, then the sample was less than 200 items and our chart slightly under-estimates the sampling error.

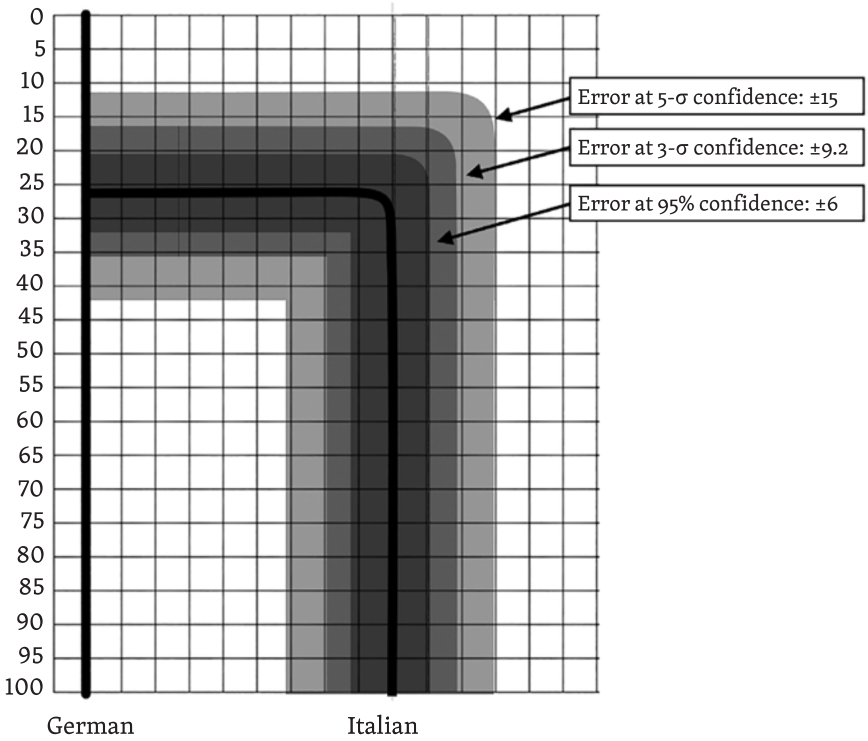

Figure 1.

Figure 2 displays five Slavic varieties as measured from Russian. Dyen, Kruskal and Black measured the cognancy ratio between Russian and Slovene, "Serbo-Croat," Czech and Ukrainian as 61.4%, 67.5%, 74.5% and 77.9%, respectively (Dyen et al. 1992, 112). A traditional dendrogram with widthless lines is visible in white. Three different confidence levels are depicted with fat branches in three different shades of greyscale: darkest at the 95% confidence level, medium grey at the four-sigma confidence level of 99.997%, and light grey at the five-sigma confidence level of 99.99997%. At the 95% confidence level, all these varieties can just barely be distinguished. At four sigma, the fat branches begin to overlap: South Slavic can be distinguished from Northern Slavic, but the margin of error cannot separate Slovene from Serbo-Croat, nor Russian from Ukrainian, nor Russian from Czech. At five sigma, all Slavic varieties are indistinguishable.

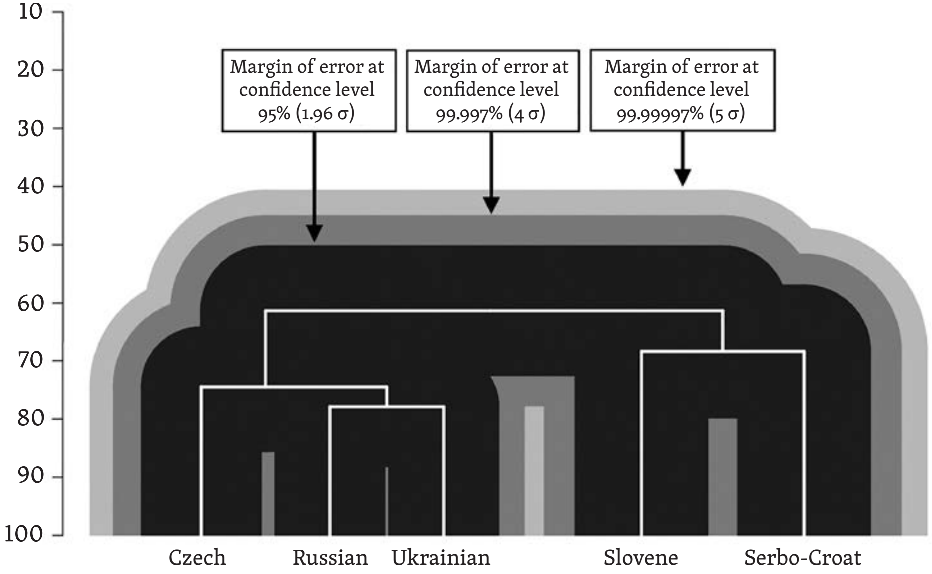

Figure 2.

Some readers may object that fat branches make the chart hard to read, since they conceal the relationship between the branches. We suggest, however, that the benefit of fat branches lies precisely in that they reveal no relationship when the margin of error is too large to draw meaningful conclusions. Fat branches indicate when lexicostatistical data are accurate enough to reveal a linguistic relationship and when they are not. The Slavic varieties are too closely related for lexicostatistical measurements to distinguish any clear relationships at the five-sigma confidence level; even at the three-sigma confidence level several Slavic varieties become indistinguishable. If the data are uncertain, then a lexicostatistical dendrogram should depict that uncertainty: presenting ambiguous results with false precision would be misleading. Fat branches prevent lexicostatisticians from becoming too enamored with imprecise data.

Other scholars may develop better techniques for the visual display of statistical uncertainty. Fat branches may have disadvantages we have not anticipated, perhaps some other technique would be better. Nevertheless, we conclude with the suggestion that dendrograms, and any other sort of diagram, should only depict clear results if the supporting data are statistically reliable. Statistically ambiguous data should produce an ambiguous diagram. Lexicostatisticians, and dialectometricians generally, should consider more carefully how to integrate error and statistical uncertainty into the visual representation of their results.

REFERENCES

Brown, Cecil – Holman, Eric – Wichmann, Søren – Velupillai, Viveka. 2008. Automated classification of the world's languages: A description of the method and preliminary results. *Language Typology and Universals / Sprachtypologie und Universalienforschung* 61(4), pp. 285–308.

Chomsky, Noam. 1977. *Dialogues avec Mitsou Ronat*. Paris: Flammarion.

Chomsky, Noam. 1979. *Language and Responsibility: Based on Conversations with Mitsou Ronat*. New York: Pantheon.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin and Use.* Westport: Praeger.

Chomsky, Noam. 1988. *Language and Politics*. Montreal: Black Rose Books.

Chrétien, C. Douglas. 1962. The mathematical models of glottochronology. *Language* 38, pp. 11–37.

Crowley, Terry. 1992. *An Introduction to Historical Linguistics*. Oxford: Oxford University Press.

Dobson, Annette – Kruskal, Joseph – Sankoff, David – Savage, Leonard. 1972. The mathematics of glottochronology revisited. *Anthropological Linguistics* 14(6), pp. 205–212.

Dolgopol'sky, Aharon. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia. In: Shevoroshkin, Vitaly – Markey, Thomas, eds. *Typology, Relationship and Time: A collection of Papers on Language Change and Relationship by Soviet Linguists*. Ann Arbor: Karoma, pp. 27–50.

Dyen, Isidore. 1962. The lexicostatistical classification of the Malayopolynesian languages. *Language* 38(1), pp. 38–46.

Dyen, Isidore. 1975. *Linguistic Subgrouping and Lexicostatistics*. The Hague: Mouton.

Dyen, Isidore – Kruskal, Joseph B. – Black, Paul. 1992. *An Indoeuropean Classification: A Lexicostatistical Experiment.* Philadelphia: American Philosophical Society.

Embleton, Sheila. 1986. *Statistics in Historical Linguistics*. Bochum: Brockmeyer.

Embleton, Sheila. 2015. Historical linguistics: Numerical methods. In: Wright, James, ed. *International Encyclopedia of the Social and Behavioral Sciences*. Oxford: Elsevier, pp. 23–26.

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

Fodor, István. 1961. The validity of glottochronology on the basis of the Slavonic languages. *Studia Slavica* 7, pp. 295–346.

Geisler, Hans – List, Johannes-Mattis. 2010. Beautiful trees on unstable ground: Notes on the data problem in lexicostatistics. In: Hettrich, Heinrich, ed. *Die Ausbereitung des Indogermanischen: Thesen aus Sprachwissenschaft, Archäologie und Genetik*. Wiesbaden: Reichert, pp. 1–10.

Geisler, Hans – List, Johannes-Mattis. 2013. Do languages grow on trees? The tree metaphor in the history of linguistics. In: Fangerau, Heiner – Geisler, Hans – Halling, Thorsten – Martin, William, eds. *Classification and Evolution in Biology, Linguistics, and the History of Science*. Stuttgart: Franz Steiner, pp. 111–124.

Gray, Russell D. – Atkinson, Quentin D. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426, pp. 435–439.

Greenhill, Simon. 2011. Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics* 37(4), pp. 689–698.

Gudschinsky, Sarah. 1956. The ABCs of lexicostatistics (glottochronology). *Word* 12(2), pp. 175–210.

Heggarty, Paul. 2010. Beyond lexicostatistics: How to get more out of "word list" comparisons. *Diachronica* 27(2), pp. 301–324.

Heggarty, Paul – McMahon, April – McMahon, Robert. 2011. From phonetic similarity to dialect classification: A principled approach. Delbecque, Nicole – Auwera, Johan van der – Geeraerts, Dirk, eds. *Perspectives on Variation: Sociolinguistic, Historical, Comparative*. Berlin: De Gruyter, pp. 43–92.

Hymes, Dell. 1960. Lexicostatistics so far. *Current Anthropology* 1(1), pp. 3–44.

Kornai, András. 2002. How many words are there? *Glottometrics* 4, pp. 61–86.

Koryakov, Yuri. 2017. Language vs. dialect: A lexicostatistic approach. *Voprosy Jazykoznanija* 6, pp. 79–101.

Levenshtein, V. I. 1965. Dvoichnye kody s ispravleniem vypadenij, vstavok i zameshhenij simvolov. *Doklady Akademii Nauk SSSR* 163(4), pp. 845–848.

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

Lyons, Louis. 2013. Discovering the significance of 5 sigma. arXiv preprint, arXiv:1310.1284.

Maguire, Warren – McMahon, April. 2011. Quantifying relations between dialects. In: Maguire, Warren – McMahon, April, eds. *Analysing Variation in English*. Cambridge: Cambridge University Press, pp. 93–120.

Mańczak, Witold. 2009. The original homeland of the Slavs. *Studia Mythologica Slavica* 12, pp. 135–145.

McElhanon, Kenneth A. 1971. Classifying New Guinea languages. *Anthropos* 66(1–2), pp. 120–144.

McMahon, April – McMahon, Robert. 2005. How do linguists classify languages? In: McMahon, April – McMahon, Robert, eds. *Language Classification by Numbers*. Oxford: Oxford University Press, pp. 20–49.

Nicholls, Geoff – Gray, Russell D. 2006. Quantifying uncertainty in a stochastic model of vocabulary evolution. In: Forster, Peter – Renfew, Colin, eds. *Phylogenetic Methods and the Prehistory of Language*. Cambridge: McDonald Institute, pp. 161–172.

Olmsted, David. 1957. Three tests of glottochronological theory. *American Anthropologist* 59(5), pp. 839–842.

Oswalt, Robert. 1971. Towards the construction of a standard lexicostatistic list. *Anthropological Linguistics* 13(9), pp. 421–434.

Pereltsvaig, Asya – Lewis, Martin W. 2015. *The Indo-European Controversy: Facts and Fallacies in Historical Linguistics*. Cambridge: Cambridge University Press.

Rea, John. 1958. Concerning the validity of lexicostatistics. *International Journal of American Linguistics* 24(2), pp. 145–150.

Serva, Maurizio – Petroni, Filippo. 2008. Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)* 81(6), 68005. URL: https://doi.org/10.1209/0295-5075/81/68005

Starostin, George. 2013. Lexicostatistics as a basis for language classification: Increasing the pros, reducing the cons. In: Fangerau, Heiner – Geisler, Hans – Halling, Thorsten – Martin, William, eds. *Classification and Evolution in Biology, Linguistics and the History of Science: Concepts – Methods – Visualization*. Stuttgart: Franz Steiner, pp. 125–146.

Starostin, Sergei. 1995. *Altajskaja problema i proiskhozhdenie japonskogo jazyka*. Moscow: Nauka: Glavnaja redakcija vostochnoj literatury

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society* 96(4), pp. 452–463.

Swadesh, Morris. 1954. Perspectives and problems of Amerindian comparative linguistics. *Word* 10(2–3), pp. 306–332.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating, *International Journal of American Linguistics* 21, pp. 121–137.

Swadesh, Morris. 1972. What is glottochronology? In: Swadesh, Morris, *The Origin and Diversification of Language*. London: Routledge & Kegan Paul, pp. 271–284.

Tadmor, Uri – Haspelmath, Martin – Taylor, Bradley. 2010. Borrowability and the notion of basic vocabulary. *Diachronica* 27(2), pp. 226–246.

Wang, Yude [Yu Te]. 1960. The lexicostatistic estimation of the time depths of the five main Chinese dialects, *Gengo Kenkyu: Journal of the Linguistic Society of Japan* 38, pp. 33–105.

Wells, John C. 1994. Computer-coding the IPA: A proposed extension of SAMPA, *Speech, Hearing and Language, Work in Progress* 8, pp. 271–289.

Wichmann, Søren. (2020). How to distinguish languages and dialects. *Computational Linguistics* 45(4). URL: https://doi.org/10.1162/coli_a_00366

Wurm, Stephen Adolphe – Laycock, Donald C. 1961. The question of language and dialect in New Guinea. *Oceania* 32(2), pp. 128–143.

Xu, Tonquiang. (1991). *Lishi Yuyanxue.* Beijing: Shangwu Yinshuguan.

Zhuravlev, Anatolij. 1994. *Leksikostatisticheskoe modelirovanie sistemy slavjanskogo jazykovogo rodstva*. Moscow: Indrik.

Alexander Maxwell
School of History, Philosophy, Political Science, and International Relations
Victoria University of Wellington
PO Box 600, Wellington 6140
New Zealand
alexander.maxwell@vuw.ac.nz

Louise McMillan
School of Mathematics and Statistics
Victoria University of Wellington
PO Box 600, Wellington 6140
New Zealand
louise.mcmillan@vuw.ac.nz