Sampling Error in Lexicostatistical Measurements: A Slavic Case Study

Jan Feld & Alexander Maxwell

Abstract:

In this article, we discuss two sources of error in determining which languages are more closely related: measurement error and sampling error. Measurement error is the difficulty of recognizing which words are cognates, sampling error arises because Swadesh lists may not be representative of a language's total lexicon. Examining the relative distances from Upper Sorbian to Czech and Polish as measured in extant lexicostatistical studies, we find that measurement error is manageable, but that sampling error undermines the reliability of lexicostatistical results: sample sizes are too small.

In diesem Artikel behandeln wir zwei Fehlerquellen bei der Bestimmung, welche Sprachen näher verwandt sind: Messfehler und Stichprobenfehler. Messfehler sind Fehler im Erkennen welche Wörter Kognaten sind. Stichprobenfehler treten auf, da Swadesh-Listen möglicherweise nicht repräsentativ für das gesamte Lexikon einer Sprache sind. Nach einer Betrachtung der relativen Entfernungen von Obersorbisch zu Tschechisch und Polnisch, finden wir, dass Messfehler relative unproblematisch sind. Stichprobenfehler aber untergraben die Zuverlässigkeit von lexikostatistischen Ergebnissen: Stichprobengrößen sind zu klein.

Keywords: Lexicostatistics; linguistic distance; sampling error; Slavic; Upper Sorbian

1. Introduction

A key aim of historical linguistics is to classify languages by their historical relationship. One popular method of classification rests on lexicostatistical cognate counting: linguists compare word lists and count the number of cognates; a higher percentage of cognates implies a closer relationship (McMahon & McMahon, 2005). Cognate counting easily determines the relative distance between languages in certain unambiguous cases: German is closer to Dutch than to Russian. In such unambiguous cases, however, relative distances are obvious even without lexicostatistical data. Relative distances between a group of closely related languages, however, are less obvious. Can lexicostatistics help? Since methodological problems are best illustrated with a specific example, we consider a naïve question about three Slavic languages: is Sorbian closer to Czech, or to Polish?

The question's premise, of course, is easily problematized. Neither "Czech," nor "Polish," nor "Sorbian" are homogenous: the diversity of west Slavic varieties might be best analysed as a dialect continuum. The existence of distinct Upper Sorbian and Lower Sorbian literary standards demonstrates that the literary standards codified in dictionaries have undergone a process of abstraction and homogenization. Acknowledging the internal complexities of Sorbian, Czech, and Polish, however, simply leads to another question: is the Upper Sorbian literary standard, say, closer to the Czech literary standard, or to the Polish literary standard? The new question is more precise, more prolix, but still naïve. Its answer is not immediately obvious. Can lexicostatistical estimates provide an answer?

In this study, we examine nine lexicostatistical measurements of the Upper Sorbian-Czech distance and the Upper Sorbian-Polish distance, drawn from six different studies. The earliest study, from Hungarian linguist István Fodor (1961: 316-317), gave four different cognate percentages. Czech scholars Petra Novotná and Václav Blažek (2005) provide not only their own estimates, but also two sets of unpublished cognate percentages: one from Russian lexicostatistician Sergei Starostin, and another from Czech scholar Mirek Čejka (1972). We also use an earlier study Čejka performed with Arnošt Lamprecht (Čejka & Lamprecht 1963), suggesting different figures. We found a final estimate from a much-cited "lexicostatistical experiment" on 84 Indo-European languages, conducted by Isidore Dyen, Joseph Kruskal and Paul Black (1992). We give more information about these studies below.

The various lexicostatistical estimates differ from each other, which raises questions about their accuracy and reliability. We distinguish two challenges in determining the relative distance between language pairs. The first challenge, which we call "measurement error," concerns the difficulty of deciding whether word pairs are or are not cognate.

Lexicostatisticians have already considered measurement error at length. The second challenge, which we call "sampling error," has been comparatively neglected. Sampling error arises because the word lists used in lexicostatistical studies may not be representative of the language as a whole.

In this article, we suggest that lexicostatisticians should seek to quantify the magnitude of sampling error using the McNemar test. While none of the lexicostatistical studies we examined provide enough data to perform the test directly, we find that reported cognate percentages nevertheless allow us to infer which McNemar test results are possible. We find that none of these studies is likely statistically significant at the conventional 5% significance level because their sample sizes are too small. To be more transparent about this shortcoming, we suggest that researchers include p-values with their lexicostatistical estimates. We also suggest that the branches of lexicostatistical dendograms should show whether language differences are statistically significant.

We are not the first to consider the impact of sampling error on lexicostatistical estimates. Robert Lee (1953) and Sheila Embleton (1986), for example, suggested methods for estimating confidence intervals for cognate percentages. In a careful discussion of the statistical nature of lexicostatistical data, Gary Simons (1977) devised tables allowing scholars to conclude whether two cognate percentages are significantly different. Paul Black (1976) even provided a formula scholars can use to test whether two cognate percentages are significantly different from each other. While our method of dealing with sampling error differs from those suggested by our predecessors, we commend these scholars for having recognised sampling error as a problem. Their tools for quantifying sampling error, however, have mostly been ignored. None of the case studies discussed in this article, for example, measure sampling error. Perhaps the McNemar test, which we propose as a method of measuring sampling error, will enjoy more widespread use.

2. Lexicostatistics and Linguistic Distances

In the 1950s, the linguistic subfield of lexicostatistics rose to prominence promising objective and easily replicable methods for establishing linguistic relationships. Lexicostatistics has been defined as "quantifying linguistic similarity," which is equivalent to quantifying linguistic difference (Grimes & Grimes, 1987). In this article, we distinguish

lexicostatistics from glottochronology, even though many pioneering lexicostatisticians are remembered primarily as glottochronologists. Most linguists now consider glottochronological results unreliable, though a few persevere. Nevertheless, several critics characterize lexicostatistics as the baby to glottochronology's bathwater (Fields, 2004: 248; Fodor, 1961: 332; Heggarty, 2010: 307; McMahon & McMahon, 2005: 34). We consider lexicostatistical data originally collected for glottochronological purposes, but ignore glottochronology as such.

Lexicostatisticians base their comparisons on a variety of source material, but most work with word lists. Lexicostatistical word lists are generically known as "Swadesh lists," after glottochronologist Morris Swadesh, who devoted considerable thought to their construction. Swadesh started with a list of 225 words, but later published lists of 215 words, 200 words, and 100 words (McMahon & McMahon, 2005; Oswalt, 1971; Swadesh, 1952, 1955). Most lexicostatisticans use one of Swadesh's lists, though a few have devised other lists. Nakhleh, Ringe, & Warnow (2005: 392) started with a "basic dataset" of 294 words, but often worked from incomplete lists: for Lycian, an ancient language known only from coins and stone inscriptions, Nakhleh et al. satisfied themselves with 44 words, 15% of their word list. Dolgopol'sky (1986) used a list with only 15 words he thought particularly resilient to borrowing. Uri Tadmor, Martin Haspelmath and Bradley Taylor (2010) used computational linguistics to propose the Leipzig-Jakarta list, a list of 100 words resistant to borrowing. Other lexicostatisticians examine texts instead of word lists: Mańczak (2009), for example, compares Biblical passages. Since languages are more than vocabulary items, lexicostatisticians have also tried measuring grammatical features alongside, or instead of, word pairs (Bakker et al., 2009).

Lexicostatisticians deploy diverse methods of comparison. Most take a standardized word list and calculate the cognate percentage, also known as the "cognate ratio." All of the studies examined here compared cognates. Others measure word similarity using the Levenshtein distance, also known as edit distance, following a famous algorithm developed by Russian computer scientist Vladimir Levenshtein (Levenshtein, 1965, 1966). Linguists typically "normalize" Levenshtein distances to a percentage by dividing the edit distance by the number of characters in the longer word (Maguire & McMahon, 2011: 108). Twenty-first century lexicostatistics relies upon computers, often using programs originally designed for analyzing genetic data (Gray & Atkinson 2003; Marris 2006; Nicholls & Gray 2006). In short, lexicostatisticians take many different approaches when selecting a textual corpus to analyze, compare different features of the chosen corpora, and process raw data using different algorithms (Embleton, 1986; Tischler, 1973).

All lexicostatistics, however, involves calculating some putatively objective numbers that represent linguistic differences. These numbers usually take the form of percentages. Sarah Gudschinsky (1956: 206) proposed transforming percentages into "degrees of linguistic difference" (or "dips"), while glottochronologists typically transformed percentages into divergence dates, often presenting both percentages and dates. Since the scholars we examine all used percentages, however, we ignore other measures.

In theory, lexicostatistic measures should provide evidence showing whether Upper Sorbian is closer to Czech or to Polish. One measures linguistic differences, calculates, and then compares the numbers. Logically, the number representing the distance between Upper Sorbian and Czech must be larger than, smaller than, or equal to the number representing the distance between Upper Sorbian and Polish. Yet insofar as lexicostatistical results diverge, they are subject to some error. Results are more reliable when the measured lexicostatistical distances are larger, and the errors smaller.

3. Measurement Error and Sampling Error

Measurement error has been widely discussed in the lexicostatistical literature. The difficulty of recognizing cognates has attracted particular attention. Serva and Petroni (2008) found recognizing cognates "often a matter of sensibility and personal knowledge" in which "subjectivity plays a relevant role." Synonyms are another source of measurement error: lexicostatisticians may struggle to decide which two words to compare, particularly since words shift their meaning slightly over time. Fodor, for example, noted that Upper Sorbian has several words for "cloud," *kuŕawa*, *chmura*, and *mrak*, and emphasized the somewhat arbitrary nature of his decision to use the first word for his cognate comparisons (Fodor, 1961a: 316; Fodor 1961b: 328). Some lexicostatisticians prefer to omit any word pairs that contain such ambiguities. Others compare multiple lexical pairs simultaneously while devising methods for weighing the importance of each pair (Heggarty 2010: 315-16).

Scholars considering lexicostatistics take diametrically opposed views about the importance of measurement error. Pereltsvaig and Lewis (2015: 89) thought uncertainties in cognate measurement "have a grave effect on the construction of language trees" that "cannot be ignored." By contrast, Dyen et al. (1992: 8) argued that error "in determining cognates will change the percentages only slightly, so that the overall classification itself is expected to survive such changes with little or no modification." Alfred Kroeber (1958: 457), pondering

the problem of "the adjustment of cognateness or noncognateness of pairs of words," unhelpfully concluded: "it is possible to be hypercritically negativistic or laxly optimistic."

A few lexicostatisticians have indicated the presence of measurement error when presenting their results. In a study of Bantu languages, David Olmsted (1957: 839-40) classified word pairs as "cognate," "non-cognate," or "ambiguous." By counting and then discarding ambiguous cases, they calculated not a single cognate percentage, but a cognancy range with a minimum and a maximum. For Swahili and Herero, for example, they estimated a cognancy range between 25% and 41%, thus an error width of 16 percentage points. Olmsted measured 44 linguistic distances, finding an average error width of 9.72 percentage points. We will see below that Fodor also included and excluded cognates, though his error widths were smaller: for the Swadesh 100, the Upper Sorbian-Czech and Upper Sorbian-Polish error widths were 4 and 6 percentage points; for the Swadesh 225, the same error widths were 6 and 5 percentage points.

Lexicostatisticans particularly fear that loanwords will create measurement error. Michael Dunn (2014: 196), for example, feared that "even one or two loanwords in a wordlist give a strong false-positive signal." Dyen et al. (1992: 31) also worried that "some borrowed words" might be "erroneously judged to be cognate with source words and with true cognates," but claimed the ability to discover loanwords through "examination of the cognation *pattern*," alternatively "by the behavior of the percentages." In practice, Dyen et al. (1992: 30) appear simply to have discarded any percentages they judged too high, a procedure they called "excluding all deviant percentages." Unfortunately, they neither define "percentage deviancy" nor provide a test for detecting it.

Most lexicostatisticians hope to avoid loanwords, or other problematic word pairs, by restricting their attention to what two lexicostatisticians discussed below called "the basic vocabulary" (Fodor 1961a: 296; Čejka 1972: 40), which is thought resistant to borrowing (Tadmor, 2009: 68-75). Restricting attention to basic vocabulary, however, limits the sample size (Heggarty 2010: 316-18). Hardly any studies examine more than 225 words in a given language, and most consider only 100 words or fewer. Smaller word lists lead to larger sampling error. If two languages are more closely related, furthermore, a larger sample size is necessary to reliably measure their lexicostatistical difference.

Swadesh lists thus present lexicostatisticians with an insolvable dilemma. The normal procedure for reducing sampling error is to increase the sample. In lexicostatistics, unfortunately, a larger sample would increase the measurement error by increasing the share of loanwords, neologisms, or other problematic items. The difficulty of balancing measurement

error with sampling error sometimes leads lexicostatisticians to apparent contradiction. Brett Kessler and Annukka Lehtonen (2006: 35), for example, wrote in the same paragraph both that "increasing the amount of data increases the accuracy and significance of tests," and also that "there is a point post which adding more words just waters down the data." Elsewhere, Kessler (2001: 67) acknowledged that, "all things being equal, it pays to have more words in the sample," but all things are not equal, and even scholars who cite Kessler approvingly have sometimes preferred sample sizes as low as 40 (e.g. Holman et al., 2008). Swadesh's preference for progressively smaller lists suggests that he ultimately prioritized low measurement error. Sheila Embleton, a mathematically sophisticated lexicostatistician experimenting with simulated data, compared lists of 100, 200, and 500 words. She found n = 200 noticeably better than n = 100, but little advantage in n = 500 (Embleton 1986: 89-92; see also Heggarty 2010: 314-318).

While lexicostatisticians have thus often considered sampling error and how to reduce it, they have rarely integrated it into a lexicostatistical analysis. One exception is Mead and Mead (1991), who made sampling error central to their analysis of Sulawesian dialects. They reported not cognate percentages but "significance groups," defined through both internal and external criteria: a significance group contains languages which are not statistically significantly different from each other, but which are statistically significantly different from all other significance groups. Their approach followed a procedure first suggested by Gary Simons, who sought to "reduce a matrix of lexicostatistical percentages to display only its meaningful differences" (Simons, 1977: 83).

In his study of Malayopolynesian, Dyen also dealt with sampling error, using an approach similar to the one we suggest below. Comparing languages with the Swadesh 200 list, he applied chi-square tests to determine whether the differences between two cognate percentages were statistically significant. There are many different chi-square tests and Dyen does not explicitly state which used. However, we suspect he used Pearson's chi-square test, since "the chi-square test" is a common shorthand for "Pearson's chi-square test." When he conducted his study in 1962, sixteen hours were required for his computer calculations; perhaps the limits of computational capacity in the 1960s influenced Dyen's proposed rule of thumb that, at the 5% significance level, cognate percentages were only "significantly different if their difference is 9.5% or greater" (Dyen 1963: 42, 60). The figure of 9.5% resembles a formula proposed without explanation in subsequent study by Black. Black suggested that the percentage point difference required for two language pairs to be significantly different at the 5% level is 139 divided by the square root of the length of the word list (Black, 1976: 50). For

word lists of 200, Black's threshold for statistical significance would be 9.82, which is close to Dyen's rule of thumb. Nevertheless, Dyen emphasized that his rule of thumb could not replace the actual application of the chi-square test.

We commend Dyen for considering sampling error, but suggest that the Pearson's chi-square test is not the best tool for testing whether the distance between two language pairs is statistically significant. Dyen himself acknowledged the limitations of the chi-square test by conceding that "…it assumes what is not true, namely that the probability of cognation is the same for each pair of words in a pair of lists" (Dyen 1962: 41).

To satisfy the assumption that all measurements are independent, furthermore, cognate probabilities should be unrelated between the language pairs (Mchugh 2013: 144). Assume we want to know whether language A (e.g. Upper Sorbian) is more closely related to B (Czech) or to C (Polish). If language A has idiosyncratic words, the set of non-cognates between A and B and the set of non-cognates between A and C will overlap more than one would expect by chance. Such overlap violates the independence assumption necessary for the Pearson chi-square test.

4. Quantifying Sampling Error with the McNemar Test

To quantify the role of sampling error and test whether the difference between two language pairs is statistically significant, we propose to use the McNemar test, which is easily performed using most statistical software, and many web-based applications.[1] The McNemar test belongs to the family of chi-square tests, but requires less strict assumptions than the Pearson's chi-square test. In particular, the McNemar test requires less strict assumptions about the independence of measurements between language pairs. If non-cognates overlap more than would be expected by chance, Pearson's chi-square test will not be valid, but the McNemar test will be.

In its general approach, the McNemar test resembles many statistical tests. Assume we measure the cognate percentage between languages A and B, and between languages A and C. Further assume that the cognate percentage A-B is larger than the cognate percentage A-C. We start from the null hypothesis that both language distances (A-B, A-C) are actually equidistant, and the differences in measured cognate percentages result from sampling error, i.e. differences between our sample and the total population. We then ask whether the differences we observe in the sample are large enough to conclude that language A is actually closer to B than to C.

[1] For example: https://www.graphpad.com/quickcalcs/McNemar1.cfm

Larger differences in the sample would provide stronger evidence that our null hypothesis is false.

Table 1 shows a two-by-two contingency table which forms the basis of the McNemar test. The four cells of the table show all possible combinations of cognate and non-cognate words for two language pairs. Two of these cells show the number of words for which there are no differences between the language pairs and both pair's words are either cognate (cell *a*) or non-cognate (cell *d*). The McNemar test only uses information where both language-pairs differ to estimate whether the overall difference is statistically significant. These combinations, known as discordant pairs, are shown in cells *b* and *c*.

Table 1: Contingency Table for McNemar test

		Second Language Pair	
		Cognates	Non-Cognates
First Lang. Pair	Cognates	**a**	**b**
	Non-Cognates	**c**	**d**

To implement the McNemar test, we calculate the value of the χ^2 (chi-square) statistic using the formula shown in Equation (1):

$$\chi^2 = \frac{(b\text{-}c)^2}{b + c} \tag{1}$$

and look up its associated p-value in the χ^2 table with one degree of freedom. We then interpret the p-value as the probability of observing the χ^2 statistic if there were no difference between these two language pairs, that is, if the observed differences in cognate percentages were solely due to sampling error. Larger χ^2 statistics lead to lower p-values, which in turn provide stronger evidence that the language pairs are actually different from each other.

For the McNemar test to be valid, three assumptions need to hold (Sheskin, 2003). First, the outcomes of each observation should only take one of two possible values. Our study easily meets this first criteria: each word is classified as either cognate or non-cognate. Secondly, each of the words on the Swadesh list should be statistically independent of each other. We believe that this second assumption holds, though perhaps some linguists may disagree. Scholars who believe that this second assumption is violated, however, could use alternate versions of the McNemar test designed for clustered data (Yang, Sun, & Hardin, 2010).

A third necessary assumption proves more problematic: all the words in the sample should have been randomly selected. Swadesh, famously, did not select his words randomly, but according to particular criteria: he wanted universal words he thought would be resistant to borrowing. Perhaps words selected according to these criteria are "as-good-as-random" for statistical purposes. Other scholars may have more insight into this problem. In this study, however, we assume that all three assumptions of the McNemar test hold.

5. Case Study: Is Upper Sorbian Closer to Czech or to Polish?

How do different ways of dealing with measurement error and the existence of sampling error affect our ability to determine which languages are more closely related? The answer to this question depends, of course, on the specific estimates and studies that they are based on. We illustrate how to answer this question with the example of the specific question of whether Upper Sorbian is closer to Czech or to Polish.

We use six lexicostatistical studies that provide cognate percentages for both Upper Sorbian-Czech and Upper Sorbian-Polish (see Table 2). Fodor (1961), the first lexicostatistician to consider Slavic languages, calculated four different cognate percentages. He used both the Swadesh 100 and the Swadesh 215, and while he claimed he had "introduced certain changes," the word lists when compared seemed identical. After calculating cognate percentages for both lists, Fodor then "excluded the possible synonyms from the computations" and recalculated. In essence, therefore, Fodor calculated a minimum-maximum range for two Swadesh lists. The distances from Upper Sorbian to Czech ranged 90 – 94% and 86 – 92%, and the distances from Upper Sorbian to Polish ranged 83-89% and 84 – 89%. By comparison, the distances from Upper to Lower Sorbian ranged from 97 – 98% and 96 – 96%.

Čejka and Lamprecht criticized Fodor's lists for "obvious errors and omissions" and proposed slightly different Slavic translations of the Swadesh 100. In practice, however, their disagreements arise from ambiguities inherent to lexicostatistics, rather than Fodor's alleged incompetence. For example, they thought the Czech word *muž* "adult male" a better translation

of Swadesh's "man" than Fodor's *člověk* (Hungarian *ember*), "person of either gender." Other lexicostatisticians, however, have preferred Fodor's interpretation of an ambiguous English word (e.g. Pereltsvaig & Lewis 2015: 72). In the end, Čejka and Lamprecht measured larger lexicostatistical distances than Fodor.

In 1972, Čejka published another glottochronological study based on the Swadesh 100 list. Čejka's article did not include any lexicostatistical percentages; Čejka unhelpfully explained that the raw data could be found "in a lecture given at the School of Slavonic and East European Studies in London, 1968." When contacted, the School of Slavonic and East European Studies had no record of Čejka's talk. Fortunately, Novotná and Blažek (2005) published Čejka's revised data in their literature review. Assuming Novotná and Blažek reproduced Čejka's data faithfully, Čejka's 1972 figures were lower than his 1963 figures.

In 1992, Dyen, Kruskal and Black published the results of an extensive "lexicostatistical experiment," for which they calculated distances between 84 Indo-European languages. Numerous scholars have subsequently used the Dyen, Kruskal, Black database. Dyen et al. took their Upper Sorbian word list directly from Fodor, though they used other sources for Czech and Polish because Fodor's results "seemed … to have higher percentages within Slavic than Slavic lists from other sources." They did not cite these "other sources." Black, who performed the actual calculations, estimated the similarity between Upper and Lower Sorbian as 95.8%, a figure close to Čejka's. Black's other calculations, however, suggested lower cognate percentages: his estimated similarity to Czech was only 83.0%; to Polish, only 76.8%.

The aforementioned literature review from Novotná and Blažek also reproduced a 2004 taxonomy that Russian scholar Sergei Starostin presented at the Santa Fe Institute in New Mexico. The Santa Fe Institute, when contacted, had no record of the talk, Starostin has since died, and Blažek, when contacted, had no original copy of Starostin's data. Yet the dendrogram Novotná and Blažek constructed from the Santa Fe data closely resembles a dendrogram that Starostin (1992: 78) published without lexicostatistical data attached. If Novotná and Blažek reproduced the Starostin data faithfully, Starostin's percentages are higher than Čejka and Lamprecht's figures, Čejka's 1972 figures and Dyen et al.'s figures, but at the low end of Fodor's ranges.

Novotná and Blažek, finally, calculated lexicostatistical percentages of their own. They declared themselves influenced by Starostin's statistical methods, but their cognate percentages were noticeably higher than Starostin's. Novotná and Blažek estimated the cognate percentage between Upper Sorbian and Polish as 92.9%; Starostin had estimated 85%. Novotná and Blažek are also the only lexicostatisticians to measure Upper Sorbian closer to Polish than to Czech.

None of these studies permit skeptical readers to replicate their calculations. Čejka's revised data and Starostin's data are unavailable. Fodor and Čejka and Lamprecht both provided their word lists, but neither provided their cognate judgements: if a scholar recalculating cognate percentages from one of their word lists found a different result, there would be no way to know where or why the disagreement arose. The remaining studies simply proclaim their cognate percentages without providing word lists.

Table 2 summarizes the estimated Upper Sorbian-Czech and Upper Sorbian-Polish language distances. Five estimates are based on the Swadesh list with 100 words, two used the Swadesh 225 word list, and one used the Swadesh 200 list. For one estimate we could not clearly determine the word list used, but suspect the Swadesh 100 list. All the studies shown in Table 2 are based on cognate counting, even if Novotná and Blažek compared "all semantically identical pairs" instead of single words, which is why they produced a non-integer cognate percentage using a list with 100 words (Novotná & Blažek 2005: 63). All the studies based their Slavic word lists on the English-language Swadesh lists.

Table 2: Lexicostatistical Distances: Upper Sorbian to Czech and Polish

(1)	(2)	(3)	(4)	(5)
Source study	Word list used	Proximity to Czech (%)	Proximity to Polish (%)	Relative distance (3)-(4)
Fodor (1961)	Swadesh 100 with synonyms	94	89	5
Fodor (1961)	Swadesh 100 no synonyms	90	83	7
Čejka & Lamprecht (1963)	Swadesh 100	88	83	5
Čejka (1972)	Swadesh 100	87	80	7
Novotná & Blažek (2005)	Swadesh 100	91.9	92.9	-1
Fodor (1961)	Swadesh 225 with synonyms	92	89	3
Fodor (1961)	Swadesh 225 no synonyms	86	84	2
Dyen et al. (1992)	Swadesh 200	83	76.8	6.2
Starostin (1992/2004)	? Swadesh 100 ?	89	85	4
Mean		89.0	84.7	4.2

As concerns the relative distance between Upper Sorbian-Czech and Upper Sorbian-Polish, the studies reported in Table 2 give a clear result. Only one out of nine measurements finds Upper Sorbian closer to Polish; of studies using the Swadesh 100, four of five studies find Upper Sorbian closer to Czech. Though lexicostatistical measurements vary, Table 2 gives some confidence that results are robust with respect to measurement error.

The studies reported in Table 2, however, are also subject to sampling error. To quantify the role of sampling error, one could simply perform the McNemar test we proposed above. Unfortunately, we do not have access to data from any of the studies in Table 2 and thus cannot perform the McNemar test with actual data. Instead, we will show possible outcomes of the McNemar test with data that could have generated the results reported in Table 2. This procedure allows us to estimate whether reported differences are statistically significant or not.

We do this by calculating for each study the highest and lowest possible p-values as well as the median possible p-value, which is our best guess of significance. For brevity, we explain the procedure in greater detail in the Appendix. Here we briefly illustrate our procedure with the estimate of Fodor (1961) with synonyms using the Swadesh 100 list, which suggests that Upper Sorbian is 5 percentage points closer to Czech than to Polish.

Tables 3 and 4 show the contingency tables of data that could be the basis of this estimate. We show which underlying data would have led to the highest possible p-value (Table 3), and the lowest possible p-value (Table 4). To get the lowest possible p-value, assume that Fodor's 5 percentage point difference results from 11 words cognate for Upper Sorbian-Czech but non-cognate for Sorbian-Polish (cell *b*) and 6 words cognate for Upper Sorbian-Polish but

non-cognate for Upper Sorbian-Czech (cell *c*). Such data would lead to the highest possible p-value of 0.2253. Table 4 similarly shows the data that would lead to the lowest possible p-value of 0.0253 ($b = 5$, $c = 0$).

Table 3: Contingency Table for Highest Possible P-value

Upper Sorbian – Polish

U. Sorbian – Czech	Cognates	Non-Cog.	Sum
Cog.	*a* **83**	*b* **11**	94
Non-Cog.	*c* **6**	*d* **0**	6
Sum	89	11	100

Note: This contingency table leads to the highest possible p-value for the Fodor (1961) estimate (with synonyms using the Swadesh 100 list). The data reported in the table lead to a chi-square statistic of 1.47 ($\chi^2 = \frac{(11-6)^2}{11+6} = 1.47$) and a p-value of 0.225.

Table 4: Contingency Table for Lowest Possible P-value

Upper Sorbian – Polish

U. Sorbian – Czech	Cognates	Non-Cog.	Sum
Cog.	*a* **89**	*b* **5**	94
Non-Cog.	*c* **0**	*d* **6**	6
Sum	89	11	100

Note: This contingency table leads to the lowest possible p-value of the Fodor (1961) estimate (with synonyms using the Swadesh 100 list). The data reported in the table lead to a chi-square statistic of 5 ($\chi^2 = \frac{(5-0)^2}{5+0} = 5$) and a p-value of 0.0253.

Overall, there are seven possible combinations of *b* and *c* (5,0; 6,1; 7,2; 8,3; 9,4; 10,6; 11,6) that could have led to the 5 percentage point difference reported in Fodor (1961). Our best guess for the actual p-value is the p-value of the median of these combinations, which is 0.1317 (χ^2 = 2.27, *b* = 8, *c* = 3). This result would not be statistically significant at the 5% level.

Table 5 shows the median of all possible p-values (Column 5), the lowest possible p-value (Column 6) and the highest possible p-value (Column 7) for all estimates shown in Table 2 for which we could identify the length of the word list used. The possible p-values have quite a large range and we can only rule out statistical significance at the 5% level for the Novotná and Blažek (2005) estimate, since even their minimum possible p-value is higher than 0.05. However, when we take the median of these p-values as our best guess for the actual p-value, as reported in Column 5, we find that no single study is likely statistically significant at the 5% level, since the median possible values are all greater than 0.05. This exercise suggests that the currently used sample sizes of up to 225 words are too small to reliably detect whether Upper Sorbian is closer to Czech or to Polish.

Table 5: Possible P-values of Studies Reported in Table 1

(1) Estimate	(2) Word list length	(3) # of cognate U.Sorb.-Czech	(4) # of cognate U.Sorb.-Polish	(5) Median possible p-value	(6) Min. possible p-value	(7) Max. possible p-value
Fodor (1961) with syn.	100	94	89	0.132	0.025	0.225
Fodor (1961) w/o syn.	100	90	83	0.090	0.008	0.178
Čejka & Lamprecht (1963)	100	88	83	0.225	0.025	0.353
Čejka (1972)	100	87	80	0.117	0.008	0.223
Novotná & Blažek (2005)	100	92	93	0.722	0.317	0.796
Fodor (1961) with syn.	225	207	200	0.162	0.008	0.286
Fodor (1961) w/o syn.	225	194	189	0.405	0.025	0.541
Dyen et al. (1992)	200	166	154	0.077	0.001	0.180

Note: This table shows the median, min. and max. possible p-values of the estimates reported in Table 1 for which we could identify the length of the word list, i.e. all estimates except Starostin (1992/2004).

6. Conclusion

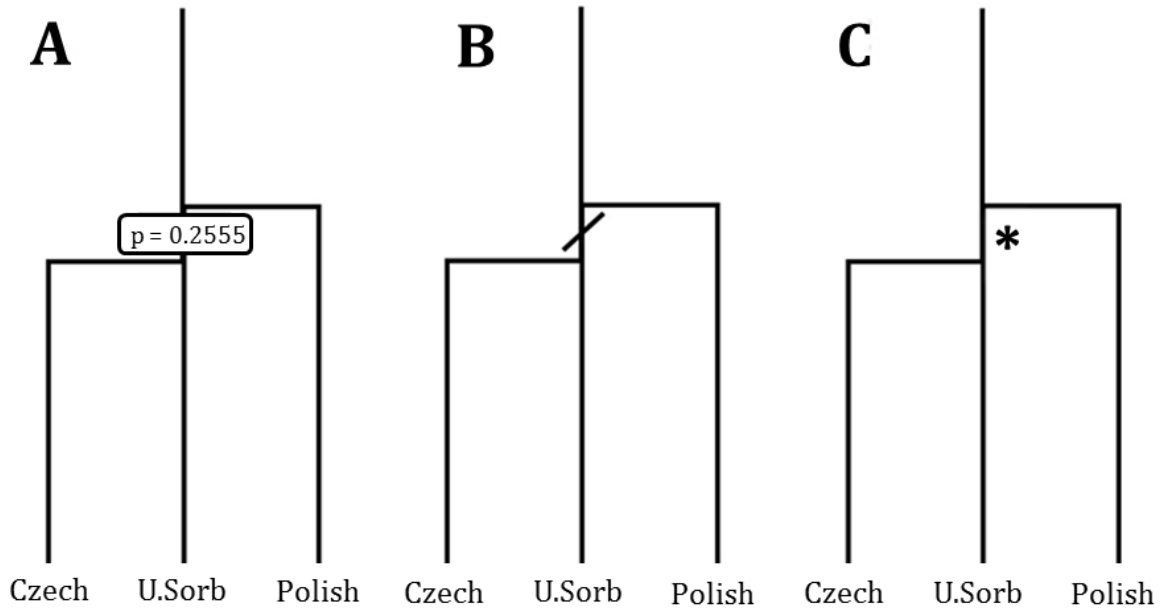
We have distinguished measurement error from sampling error and discussed how both types of error affect the ability to determine which languages are more closely related. Considering the specific example of Upper Sorbian, Czech and Polish, we saw that four out of five estimates using the same source material found Upper Sorbian closer to Czech than to Polish. An 80% consensus suggests that this finding is robust even though different lexicostatistical studies approach measurement error in different ways.

All estimates based on limited word lists, however, are also subject to sampling error. We proposed that sampling error in cognate counting can be quantified using the McNemar test. The Swadesh 100 list enjoys the greatest popularity, and hardly any lexicostatisticians use a word list longer than the Swadesh 225. Yet the sampling error with a 225 word list swamps the signal, at least for the specific example of Upper Sorbian, Czech, and Polish. The Upper Sorbian-Czech distance and the Upper Sorbian-Polish distance are evidently too similar to distinguish reliably with word samples of the size normally used in lexicostatistical studies.

We conclude by suggesting that p-values ought to feature more prominently in lexicostatistical results. Only one of the six lexicostatistical studies discussed in Table 2 provided any measure of uncertainty: Fodor calculated a range with a lower and upper bound. The other lexicostatisticians presented linguistic distances as a single number, and indeed Dyen et al. and Novotná and Blažek gave numbers with three significant figures, implicitly claiming accuracy to a tenth of a percentage point! Lexicostatistical estimates are not that precise.

Lexicostatistical dendrograms, furthermore, might depict visually whether scholars should have confidence in the dendogram's branches. Figure 1 shows two examples of how this could be done. First, one could simply show the p-value next to the line that depicts differences between language pairs (Figure 1 A). This way of depicting confidence in the dendogram's branches resembles the approach by Gray and Atkinson (2003), which shows the degree of confidence in a particular branch by depicting the Bayesian posterior probabilities. Second, one could indicate with a dash through the branch that the given difference is not statistically significant (Figure 1 B) or a star if it is (Figure 1 C).

Figure 1: Possible Dendrograms Showing Statistical Significance



Overall, we suggest that statistical uncertainties should be highlighted, rather than conceded and ignored. Even if there exists a "true linguistic difference," expressible as a single number, actual lexicostatistical measurements are estimates that suffer from inevitable inaccuracy. Lexicostatisticians already understand the necessity of minimizing measurement error. However, we suggest that lexicostatisticians ought to pay more attention to quantifying sampling error, and depicting it prominently when presenting their results.

References

Bakker, D., Müller, A., Velupillai, V., Wichmann, S., Brown, C. H., Brown, P., … Holman, E. W. (2009). Adding Typology to Lexicostatistics: A Combined Approach to Language Classification. *Linguistic Typology*, 13(1), 169–181. https://doi.org/10.1515/LITY.2009.009

Black, P. (1976). Multidimensional Scaling Applied to Linguistic Relationships. *Cahiers de l'Institut de Linguistique de Louvain*, 3, n5-6.

Čejka, M & Lamprecht, A. (1963). K otázce vzniku a diferenciace slovanských jazyků. *Sborník prací Filozofické Fakulty Brněnské University, A, Řada jazykovědná* 12, 5-20.

Čejka, M. (1972). Lexicostatistic Dating and Slavonic Languages. *Sborník Prací Filozofické Fakulty Brněnské University, A, Řada jazykovědná* 21, 39-52.

Dolgopol'sky, A.. 1986. A Probabilistic Hypothesis Concerning the Oldest Relationships Among the Language Families of Northern Eurasia. In V. Shevoroshkin & T. L. Markey (eds.) *Typology, Relationship and Time: A collection of Papers on Language Change and Relationship by Soviet Linguists*, 27-50. Ann Arbor: Karoma.

Dunn, Michael. 2014. Language Phylogenies. In: Claire Bowern, Bethwyn Evans (eds). *The Routledge Handbook of Historical Linguistics*, 190-211. London: Routledge.

Dyen, I. (1962). The Lexicostatistical Classification of the Malayopolynesian Languages. *Language*, 38(1), 38-46.

Dyen, I., Kruskal, J., & Black, P. (1992). *An Indoeuropean Classification: A Lexicostatistical Experiment.* Philadelphia: American Philosophical Society.

Embleton, S. (1986). *Statistics in Historical Linguistics*. Bochum: Brockmeyer.

Fields, E. L. (2004). Before "Baga": Settlement Chronologies of the Coastal Rio Nunez Region, Earliest Times to C.1000 CE. *The International Journal of African Historical Studies*, 37(2), 229. https://doi.org/10.2307/4129008

Fodor, I. (1961a). A glottochronologia ervenyessege a szlav nyelvek anyaga alapjan. *Nyelvtudományi Közlemények*, 63(2), 308-344.

Fodor, I. (1961b). The Validity of Glottochronology on the Basis of the Slavonic languages. *Studia Slavica*, 7, 295–346.

Gray, R., & Atkinson, Q., (2003). Language-tree Divergence Times Support the Anatolian theory of Indo-European origin. *Nature* 426, 435-439.

Grimes, C., & Grimes, B. (1987). *Languages of South Sulawesi*. Canberra: Research School of Pacific and Asian Studies, ANU.

Gudschinsky, S. C. (1956). The ABC'S of Lexicostatistics (Glottochronology). *WORD*, 12(2), 175–210. https://doi.org/10.1080/00437956.1956.11659599

Heggarty, P. (2010). Beyond lexicostatistics: How to get More out of "Word List" Comparisons. *Diachronica*, 27(2), 301–324. https://doi.org/10.1075/dia.27.2.07heg

Holman, E., Wichmann, S., Brown, C., Velupillai, V., Müller, A. (2008). Explorations in Automated Language Classification. *Folia Linguistica*, 42(3-4), 331-354.

Kessler, B. (2001). *The Significance of Word Lists*. CSLI Publications.

Kessler, B. & Lehtonen, A. (2006). Multilateral Comparison and Significance Testing of the Indo-Uralic Question. In: Colin Renfew, Peter Forster (eds), *Phylogenic Methods and the Prehistory of Languages*, 33-42. Cambridge: McDonald Institute.

Kroeber, A. (1958). Romance History and Glottochronology. *Language*, 34(4), 454-457.

Lees, R. (1953). The Basis of Glottochronology. *Language,* 29(2), 113-127. doi:10.2307/410164

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

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10(8), 707–710. https://doi.org/citeulike-article-id:311174

Maguire, W., & McMahon, A. (2011). Quantifying Relations between Dialects. In W. Maguire & A. McMahon (eds). *Analysing Variation in English*. Cambridge: Cambridge University Press.

Mańczak, W. (2009). The Original Homeland of the Slavs. *The Original Homeland of the Slavs*, 12, 135–145.

Marris, E. (2006). Language: The Language Barrier. *Nature*, 453, 446-448

Mchugh, M. L. (2013). The Chi-square Test of Independence: Lessons in Biostatistics. *Biochemia Medica*, 23(2), 143–9. https://doi.org/10.11613/BM.2013.018

McMahon, A., & McMahon, R. (2005). How Do Linguists Classify Languages? In: McMahon, A., & McMahon, R., *Language Classification by Numbers*, 20–49. Oxford: Oxford University Press.

McMahon, A., & McMahon, R. (2005). *Language Classification by Numbers*. Oxford: Oxford University Press.

Mead, D., & Mead, M. (1991). Survey of the Pamona Dialects of Kecamatan Bungku Tengah. *Workpapers in Indonesian Languages and Cultures*, 11, 121-142.

Nakhleh, L., Ringe, D., & Warnow, T. (2005). Perfect Phylogenetic Networks: a New Methodology for Reconstructing the Evolutionary History of Natural Languages. *Language*, 81(2), 382–420. https://doi.org/10.1353/lan.2005.0078

Nicholls, G, Gray, R. (2006). Quantifying Uncertainty in a Stochastic Model of Vocabulary Evolution. In: P. Forster, C. Renfrew (eds), *Phylogenetic Methods and the Prehistory of Language*, 161-171. Cambridge: McDonald Institute.

Olmsted, David. 1957. Three Tests of Glottochronological Theory. *American Anthropologist*, 59(5), 839-842.

Oswalt, R. L. (1971). Towards the Construction of a Standard Lexicostatistic List. *Anthropological Linguistics*, 13(9), 421–434. Retrieved from http://www.jstor.org/stable/30029088

Pereltsvaig, A. & Martin L. (2015). *The Indo-European Controversy: Facts and Fallacies in Historical Linguistics*. Cambridge: Cambridge University Press.

Serva, M., & Petroni, F. (2008). Indo-European Languages Tree by Levenshtein Distance. *EPL (Europhysics Letters)*, 81(6), 68005. https://doi.org/10.1209/0295-5075/81/68005

Sheskin, D. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures* (3rd ed.). Chapman & Hall Crc. https://doi.org/10.1198/tech.2004.s209

Starostin, S. (1992). Methodology of Long-Range Comparison. In: Vitalii Shevoroshkin (ed.). *Nostratic, Dene-Caucasian, Austric and Amerind*, 75-59. Bochum: Brockmeyer.

Simons, G. (1977). Tables of Significance for Lexicostatistics. In: Richard Loving, Gary Simons (eds). *Language Variation and Survey Techniques: Workpapers in Papua New Guinea languages* 21, pp. 75-106. Ukarumpa: Summer Institute of Linguistics.

Swadesh, M. (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), 452–463. https://doi.org/10.2307/3143802

Swadesh, M. (1955). Towards Greater Accuracy in Lexicostatistics Dating. *International Journal of American Linguistics*, 21(2), 121–137. Retrieved from http://www.jstor.org/stable/1263939

Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. In: Martin Haspelmath, Uri Tadmor, eds., *Loanwords in the World's languages: A Comparative Handbook*, pp. 55-75. de Gruyter

Tadmor, U., Haspelmath, M. and Taylor, B. (2010). Borrowability and the notion of basic vocabulary. *Diachronica*, 27(2), 226-246.

Tischler, J. (1973). *Glottochronologie und Lexikostatistik*. Innsbruck: Kowatch.

Yang, Z., Sun, X., & Hardin, J. W. (2010). A Note on the Tests for Clustered Matched-pair Binary Data. *Biometrical Journal*, 52(5), 638–652. https://doi.org/10.1002/bimj.201000035

Appendix: Calculating Possible P-Values with McNemar Tests

Here we explain in greater detail how we infer possible results of McNemar tests that are consistent with the cognate percentages reported in Table 2. We first provide a general explanation that outlines the logic of our approach. We then give a step-by-step description of the calculations.

To conclude whether Upper Sorbian is closer to Czech or to Polish, we seek to determine whether the differences in cognate percentages, as reported in Column 5 of Table 2, are statistically significant. Unfortunately, none of these studies provide the data necessary to perform the McNemar test. We therefore "reverse engineer" results of the McNemar test consistent with the results reported in the original studies.

As in the main text, we illustrate our explanation with the example of Fodor's estimate for the Swadesh 100 list with synonyms (hereafter: Fodor_100_S). In this table and throughout the remainder of this appendix, *a* refers to the number of words that are cognate for both Upper Sorbian-Czech and Upper Sorbian-Polish, *b* refers to the number of words that are cognate for Upper Sorbian-Czech and non-cognate for Upper Sorbian-Polish, *c* refers to the number of words that are non-cognate for Upper Sorbian-Czech and cognate for Upper Sorbian-Polish, and *d* refers to the number of words that are non-cognate for both Upper Sorbian-Czech and Upper Sorbian-Polish (see Table A1).

Table A1: Contingency Table Sorbian-Czech and Sorbian-Polish

Upper Sorbian - Polish

U.Sorbian - Czech

	Cognates	Non-Cog.	Sum
Cognates	**a**	**b**	94
Non-Cognates	**c**	**d**	6
Sum	89	11	100

To be able to perform the McNemar test we need to calculate the χ^2 statistic

$$\chi^2 = \frac{(b-c)^2}{b+c} \quad \text{(A1)}$$

and look up its associated p-value in the chi-square distribution table with one degree of freedom. This table is widely available. The magnitude of the χ^2 statistic depends on the numerator and denominator shown in Equation (A1). Given the results reported for each estimate, we can calculate the numerator of this equation, which is equal to square of the difference in number of cognate words between the language pairs. To see why, consider the Fodor_100_S estimate shown in Table A1. For this estimate, the number of cognate words for Sorbian-Czech is equal to 94 ($a + b$) and the number of cognate words for Upper Sorbian-Polish is equal to 89 ($a + c$). The difference between the number of cognate words between these two language pairs is therefore 5 ($b - c$). Its square, and the numerator of the χ^2 statistic, is 25.

While the numerator is easy to calculate, many different denominators of Equation (A1) are consistent with the results reported in Table 2. The larger $b + c$, the larger the actual value of the χ^2 statistic, and the smaller the p-value. To constrain the possible values of the χ^2 statistic, therefore, we look for the highest and lowest possible values for $b + c$.

The maximum values b and c are constrained by two numbers: the number of non-cognate words for each language pair, that is, $c + d$ for Upper Sorbian-Czech, and $b + d$ for Upper Sorbian-Polish. To find the highest possible values for b and c, we set $d = 0$ and solve for b and c. With these values for b and c, we can then calculate the highest possible p-value that is consistent with our observed results. For the Fodor_100_S estimate shown in Table A1, $b + d = 11$ and $c + d = 6$, so $b = 11$ and $c = 6$. We then plug in the values of b and c into Equation (A1) to calculate a χ^2 statistic of 1.47, for which the associated p-value is 0.2253.[2]

The minimum values for b and c are constrained by the number of non-cognate words in each language pair. To find the lowest possible values for b and c, we set d equal to its highest possible value, which is the number of non-cognate words for the language pair with the lowest number of non-cognate words. We then solve for b and c as in the previous paragraph. The resulting values for b and c enable us to calculate the lowest possible p-value consistent with our results. For Fodor_100_S estimate, Upper Sorbian-Czech has the fewest non-cognate words, namely 6, so we set $d = 6$. We can then solve for b and c as follows: $b + d = 11$ and $c + d = 6$, so $b = 5$ and $c = 0$. We can then plug in the values of b and c into Equation (A1) to calculate a χ^2 statistic of 5, for which the associated p-value is 0.0253. For the

[2] To check whether this is the contingency table with the lowest possible χ^2 statistic, consider increasing b and c by one, so that $b = 12$ and $c = 7$. This is not possible because we know that the total number of non-cognate words for Upper Sorbian-Polish ($b + d$) is 11 and the total number of non-cognate words for Upper Sorbian-Czech ($c+d$) is 6.

Fodor_100_S estimate, therefore, the p-value lies between a maximum of 0.2253 and a minimum of 0.0253.

Since we now know the highest and lowest possible p-values, we can calculate the median possible p-value for each result in Table 1. Since the cognate percentages were calculated using a finite number of words, the number of possible p-values is also finite. The median possible p-value is the p-value of the McNemar test with the middle of the ordered possible combinations of *b* and *c*, or, if there are two middle combinations, the average of the p-values of the middle two combinations. For the Fodor_100_S estimate, the possible combinations of *b* and *c* are (5,0), (6,1), (7,2), (8,3), (9,4), (10,6) and (11,6). Because the number of possible combinations is odd, we simply calculate the p-value for the middle combination of (8,3), which is 0.1317 (χ^2 = 2.27).

Step-by-Step Calculation of Possible P-values

Below we provide more detailed instructions for recreating our calculations. These instructions supplement the calculations in the Excel file "Feld.Maxwell_2018", which is available online as supplementary material to this article.

Step 1: Calculating the Number of Words that are Cognate, Non-Cognate and the Difference between Number of Cognate Words

To calculate the number of cognate words for the language pairs Upper Sorbian-Czech and Upper Sorbian-Polish, we multiply the cognate percentages with the length of the word list and round any non-integers. The remaining words are non-cognate and the difference between number of cognate words between the two language pairs is self-explanatory. Because we could not identify the length of the word list used in Starostin (1992/2004), we exclude his estimate from this exercise.

Step 2: Calculating the Highest and Lowest Possible Values for b and c

To find the highest possible value for *b* and *c*, we set *d* = 0. The highest possible value for *b* is the number of non-cognate words between Upper Sorbian and Polish. Similarly, the highest possible value for *c* is the number of non-cognate words between Upper Sorbian and Czech.

To find the lowest possible value for *b* and *c*, we set *d* equal to the number of non-cognate words for the language pair with the lowest number of non-cognate words. To find the lowest possible value for *b*, we subtract the highest possible value for *d* from the number

of non-cognate words between Upper Sorbian and Polish. To find the lowest possible value for *c*, we subtract the highest possible value for *d* from the number of non-cognate words between Upper Sorbian and Czech.

Step 3: Calculating the Highest Possible P-value

To find the highest possible p-value, we need the highest possible values for *b* and *c,* as calculated in Step 2, to calculate the chi-square statistic using Equation (A1). To find the p-value associated with this chi-square statistic, which is the highest possible p-value consistent with the results reported in the original studies, we look up the associated p-value of the chi-square statistic with one degree of freedom.

Step 4: Calculating the Lowest Possible P-value

To find the lowest possible p-value, we need the lowest possible values for *b* and *c*, as calculated in Step 2, to calculate the chi-square statistic using Equation (A1). To find the p-value associated with this chi-square statistic, the lowest possible p-value consistent with the results reported in the original studies, we look up the associated p-value of the chi-square statistic with one degree of freedom.

Step 5: Calculating Median Possible P-value

Recall that that the highest possible value for *b* leads to the highest possible p-value (see Step 3) and the lowest possible value for *b* leads to the lowest possible p-value (see Step 4). To find the median possible p-value, we seek the middle value or values of *b*. If the number of possible *b* values is odd, there is only one middle *b*. If the number of possible *b* values is even, there are two middle *b* values (i.e. the two values of *b* in the middle of an ordered list). We find the middle *b* value or values through the following procedure:

i) Take the difference between the lowest and highest possible *b*,

ii) Divide the difference by two,

iii) Add the resulting figure to the lowest *b*

If there is an odd number of possible *b* values, steps i) – iii) will result in the single median *b*. If there are an even number of possible *b* values, however, we need two extra steps:

iv) Round the result up, and

v) Round the result down.

Step iv) yields the lowest middle *b* value, step v) the highest middle *b* value. In our Excel file, we performed these steps for all values. The results are the same, because rounding up and down from a single median *b* will result in the same value.

To calculate the values for *c*, we subtract the median *b* or each of the middle *b* values from *b* – *c*, the difference between the number of cognates. Using the values for *b* and *c*, we then calculate the chi-square statistic and look up the p-values associated with the highest and lowest middle *b* values. Finally, we take the average of both of these p-values. This average shows the median possible p-value as reported in Table 5.

Example of median possible p-value with an odd number of possible p-values

For the Fodor_100_S estimate, the possible values of *b* are 5, 6, 7, 8, 9, 10, and 11. Using the formula described in the previous paragraph, we get: (11-5)/2 + 5 = 8, which remains 8 whether we round up or down. To calculate *c*, we subtract the difference in the number of cognates, 8 – 5, which gives *c* = 3. We then calculate the chi-square statistic for *b* = 8 and *c* = 3 which is 2.273, for which the associated p-value is 0.132. This is the median possible p-value reported in Table 5.

Example of median possible p-value with an even number of possible p-values

For the estimate of Čejka (1972) using the Swadesh 100 list, all possible values of *b* are 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, and 20. Using the formula described above, we get: (20-7)/2+7 = 13.5, which becomes 13 when we round down and 14 when we round up. This means that we have two middle *b* values: 13 and 14. To calculate *c*, we subtract the difference in the number of cognates from both middle *b* values, thus 13 – 7 and 14 - 7, which gives us *c* = 6 and 7. We calculate the chi-square statistics for *b* = 13 and *c* = 6, which is 2.579, for which the associated p-value is 0.108. We then calculate the chi-square statistics for *b* = 14 and *c* = 7 which is 2.33, for which the associated p-value is 0.127. The median possible p-value is the average of these two p-values: (0.127 + 0.108)/2 = 0.117. This is the median possible p-value reported in Table 5.