Towards identifying the optimal datasize for lexically-based Bayesian inference of linguistic phylogenies

Taraka Rama[♠] **Søren Wichmann**[♢,♡]
[♠]Department of Informatics, University of Oslo, Norway
[♢]Leiden University Centre for Linguistics, Leiden University, Netherlands
[♡]Laboratory of Quantitative Linguistics, Kazan Federal University, Russia
tarakark@ifi.uio.no, wichmannsoeren@gmail.com

Abstract

Bayesian linguistic phylogenies are standardly based on cognate matrices for words referring to a fix set of meanings—typically around 100-200. To this day there has not been any empirical investigation into which datasize is optimal. Here we determine, across a set of language families, the optimal number of meanings required for the best performance in Bayesian phylogenetic inference. We rank meanings by stability, infer phylogenetic trees using first the most stable meaning, then the two most stable meanings, and so on, computing the quartet distance of the resulting tree to the tree proposed by language family experts at each step of datasize increase. When a gold standard tree is not available we propose to instead compute the quartet distance between the tree based on the n-most stable meaning and the one based on the $n+1$-most stable meanings, increasing n from 1 to $N-1$, where N is the total number of meanings. The assumption here is that the value of n for which the quartet distance begins to stabilize is also the value at which the quality of the tree ceases to improve. We show that this assumption is borne out. The results of the two methods vary across families, and the optimal number of meanings appears to correlate with the number of languages under consideration.

1 Introduction

Phylogenetic methods—both distance-based and character-based methods (specifically Bayesian phylogenetic inference)—from computational biology are being widely used in historical linguistics and typology, for phylogenetic inference, for triangulating homelands of proto-languages (Wichmann et al., 2010b), for investigating rates of lexical change (Greenhill et al., 2017), dating and spread of language families (Gray et al., 2009; Holman et al., 2011; Bouckaert et al., 2012), reconstruction of proto-languages (Bouchard-Côté et al., 2013), and for investigating typological universals (Dunn et al., 2011).

Most of the above studies require cognate matrices based on wordlists annotated by experts for cognacy. Each column in such a matrix represents a *cognate class*, with 1's and 0's indicating whether or not a form in the shared ancestral language has a reflex in the descendant languages with a given predefined meaning pertaining to the fixed list of meanings. Such cognate matrices are fed to Bayesian phylogenetic software such as MrBayes (Ronquist et al., 2012b), BEAST (Drummond et al., 2012) or BayesPhylogenies (Pagel and Meade, 2004). An alternative approach is to apply distance-based methods (Wichmann et al., 2010a; Jäger, 2013; Rama and Borin, 2015). Such methods have the advantage that they do not necessarily require labor-intensive cognate identification or the assumption which goes with cognate identification, namely that the languages analyzed are actually related. Instead, an aggregate phonetic distance can be computed. As in biology, the most popular algorithm for inferring phylogenies from linguistic distance data has been Neighbor-Joining (Saitou and Nei, 1987). All the above methods have typically been applied using variants of Swadesh meaning lists (Swadesh, 1952), sometimes adapted for the language family in focus.

Some information is already available concerning the optimal word list size for distance-based linguistic phylogenetic inference. The pertinent literature is reviewed in the next section. As for

Proceedings of the 27th International Conference on Computational Linguistics, pages 1578–1590
Santa Fe, New Mexico, USA, August 20-26, 2018.

character-based methods, such as Bayesian ones, there is no study investigating the optimal meaning list length when inferring linguistic phylogenies. One reason for this could be the computational costs involved in inferring millions of trees for each of the cumulative meaning lists. To the best of our knowledge the present paper represents the first known computationally intensive study in this direction. We infer trees for eleven language families (a total of 313 languages) using Bayesian phylogenetic inference procedure and apply two methods to determine whether the quality of trees stops changing as the size of the word list increases. The first method requires expert phylogenetic trees for validation. It is similar in nature to one applied by Holman et al. (2008), described in the next section. The second method is entirely novel and does not require expert trees. Both methods are applied to the five language families in our sample for which reasonably reliable expert trees were available, and we show that the results are similar. Thus support is found for also including results for the second method, which does not require expert trees. Thereby results could be added the remaining six language families and more reliable statistics was obtained.

In this paper, we ask the following research questions:

- Is there any optimal list length after which there is no improvement in the trees inferred from Bayesian phylogenetic inference procedure?
- Is there any procedure to determine the optimal list length when there is sparse historical linguistic research regarding the genetic tree for a language family?

The structure of the paper is as follows. We discuss some work related to determining ranks of item stability in section 2. The problem and our approach are described in section 3. Section 4 presents the datasets and outlines how Bayesian phylogenetic inference is applied. Section 5 presents and discusses the implications of our results and is followed by conclusion in section 6.

2 Related work

Holman et al. (2008) used 245 100-item word lists from languages of 69 families previously compiled for another paper (Brown et al., 2008) and then devised a stability measure quantifying the resistance of a meaning to undergo lexical replacement. The authors ranked the 100 meanings by the order of stability and then created lists consisting of the 5, 6 ... 100 most stable meanings. They then proceeded to infer linguistic trees from the meaning lists, computing the fit of the trees to expert trees. The authors found that the quality of the trees ceased to improve when using more than the 40 most stable meanings (and subsequently 40-item lists were collected for as many languages of the world as possible, ultimate leading to the so-called ASJP database of Wichmann et al. (2018), covering more than two thirds of the world's languages). Evaluation studies (Wichmann et al., 2010a; Pompei et al., 2011; Jäger, 2013; Rama and Borin, 2015) have suggested that the quality of the phylogenetic trees inferred from 40-item word lists is quite high when compared to gold standard phylogenetic trees.

Another investigation of the effect of the number of meanings on the quality of a phylogeny is represented by Petroni and Serva (2010). The authors rank 200-item meaning lists of 50 Austronesian and 50 Indo-European languages through a stability measure derived from the Levenshtein distance between synonyms across word lists. They then infer a distance-based tree for each cumulative meaning list and compare the tree with the Levenshtein distance tree inferred from the full dataset using the Robinson-Foulds distance (Robinson and Foulds, 1979). The authors observe that the Robinson-Foulds distance drops as the length of meaning list changes from 1 to 200. The authors do not compare the inferred trees from subsets to the gold standard family tree. Neither do the authors use the cognate information to infer trees. Instead, the distance between two languages is based on aggregate Levenshtein distance computed between synonyms in the word lists.

In another work, Rama and Borin (2013) computed phoneme n-gram entropy for synonyms belonging the 100-item meaning lists from 245 languages used by Holman et al. (2008). Then they rank the meanings in order of increasing entropy. The meaning with the lowest entropy is the most stable item.

The authors find that the entropy-ordered meaning lists' ranks correlate highly with the order inferred by Holman et al. (2008).

In a related work, which uses Bayesian inference, Pagel and Meade (2006) employ a Bayesian phylogenetic inference model where each meaning is treated as a separate partition within the complete dataset (cf. table 2). Each partition or meaning has a separate weight that contributes to the likelihood calculation. The weight for each meaning is inferred through an MCMC procedure (cf. section 4.2 for details on Bayesian phylogenetic inference procedure). Then the authors correlated the mean weight for each meaning estimated from the posterior samples with the number of cognate classes in the meaning for 87 Indo-European languages and 95 Bantu languages. The authors find a positive correlation ($r > 0.6$) between the mean weights and the number of cognate classes for both language families.

In a follow-up paper, the same authors Pagel et al. (2007) interpreted the inferred weight from posterior samples as an index for lexical replacement and found a negative correlation with corpus-based estimates of the frequency of words representing the different meanings. In essence, meanings with low frequency undergo lexical replacement more easily than meanings with high frequency. In the present paper, we also apply Bayesian inference procedure to cumulative datasets ranked by the diachronic lexical stability.

3 Problem and approach

To investigate the influence of the datasize on the quality of a Bayesian phylogeny, word lists of different sizes must be produced. In doing so, it is necessary to take the differential stabilities of the words corresponding to different meanings into account, because more stable items are expected to lead to better phylogenies and the influence of stability can thus mask the influence of the number of items. It is well known that the words for some meanings tend to be more quickly replaced than words for some other meanings.

Different methods and datasets discussed in section 2 lead to somewhat different assessments of relative stabilities, but not to such a degree that any major controversies have arisen. The measure of stability used in the present paper is based on the simple and quite old idea idea (cf. Thomas (1960), Kroeber (1963), and Oswalt (1970)) that words for more stable items can be identified by their greater tendency to yield cognates within groups of closely related languages than words for less stable items. Here we use the number of cognate classes as a measure of stability: the smaller the number of cognate classes found for a given meaning is, the fewer word substitutions there would have been, and thus the greater the stability. Having ranked the meanings in this way for a given language family we can go on to define word lists of different sizes, n, each consisting of words for the n-most stable meanings. As a note, in case of ties—two or more equally stable meanings—these meanings will be arbitrarily ranked.

We propose two tests for comparing the quality of trees based on two meaning lists differing by the number of items they contain. In the first test, for each of the two meaning lists, we report the average of the generalized quartet distances (GQD) between the posterior tree distribution and a gold standard tree. We hypothesize that after reaching a certain size, n, of meaning list, the difference between the average GQD distance between the list having n meanings and the one having $n + 1$ meanings will stabilize. This first test is designed for cases where the gold standard family tree is reasonably well established and uncontroversial. This is not always the case, since the amount of historical linguistic work dedicated to language families throughout the world varies greatly (Campbell and Poser, 2008). For cases where expert classifications are lacking or not very trustworthy we propose a second test that does not depend on the gold standard tree. This test consists in computing the GQD's between trees inferred from meaning lists of size n and $n + 1$, where n ranges from 1 to the next to largest number of meanings covered in the dataset used, and then plot the average quartet distance for each such pair. Our hypothesis is that after a point, the difference of average quartet distances will begin to stabilize.

4 Materials and Methods

In this section, we describe the datasets, tree inference procedure, and the evaluation procedure.

Family	Languages	Meanings	Sources
Austronesian	96	210	Gray et al. (2009)
Mayan	30	100	Wichmann and Holman (2013)
Mixe-Zoque	10	100	Cysouw et al. (2006)
Indo-European	52	208	Dunn (2012)
Uto-Aztecan	31	100	Miller (1984)
Afro-Asiatic	25	100	Militarev (2000)
Bai	9	110	Wang (2006)
Chinese	18	180	Dàxué (1964)
Mon-Khmer	16	100	Peiros (1998)
Ob-Ugrian	21	110	Zhivlov (2011)
Tujia	5	109	Starostin (2013)

Table 1: Number of languages, meanings, and sources for each dataset.

4.1 Datasets

The experiments involving expert trees are performed on datasets belonging to five different language families: Austronesian, Mayan, Mixe-Zoque, Indo-European, and Uto-Aztecan. In addition to this dataset we draw on another six language families or subfamilies: Afro-Asiatic, Bai, Chinese, Mon-Khmer, Ob-Ugrian, and Tujia. The number of languages, number of meanings, and sources for each dataset is given in Table 1. In case of Austronesian and Indo-European, we use a subset of languages provided in Jäger et al. (2017).

Each of the above datasets is a multilingual word list with expert annotated cognate judgments concerning the words representing each meaning. An excerpt of a multilingual word list from Indo-European is shown in table 2a. The meanings are ordered by the number of cognate classes. We also show the corresponding binary matrix in table 2b. The binary matrix has 2 columns and 3 columns for the meanings ALL and AND since there are, respectively, two and three cognate classes for the meanings ALL and AND. In contrast, the meaning NAIL, for instance, has only one cognate class since all the daughter languages show words that can be traced back to the Proto-Indo-European stage. The meanings are accumulated at each rank and converted to a binary matrix of dimensions $L \times C$ where L is the number of languages and C is the number of cognate sets in the list of meanings accumulated. We explain the Bayesian phylogenetic inference and the tree evaluation algorithms below.

Language	ALL	AND	...
English	ɔːl^1	aɛnd^1	...
German	alə^1	ʊnt^1	...
French	tu^2	e^2	...
Spanish	toðo^2	i^2	...
Swedish	ˈalːa^1	ɔkɾ^3	...

(a) Forms and cognate classes

Language	ALL		AND		
English	1	0	1	0	0
German	1	0	1	0	0
French	0	1	0	1	0
Spanish	0	1	0	1	0
Swedish	1	0	0	0	1

(b) Binary Matrix

Table 2: Excerpt from meaning list showing cognate classes (table 2a) and the binary cognate matrix (table 2b) for ALL and AND in Germanic and Romance subfamilies of Indo-European. The superscript indicates words that are cognate.

The meanings ranked by the number of cognate classes in each of the five families for which we make use of expert trees are provided below. A number in a square bracket introduces meanings for which there are as many cognate classes as the number indicates. The lists have some interest in their own right since they contribute to the available empirical data on word stabilities across different language families. For reasons of page limitations we do not include the remaining six families.

Austronesian : [1] fifty, [2] twenty, six, [3] stickwood, seven, louse, five, [4] to yawn, breast, ten, eye, [5] two, that, to die/be dead, three, fruit, eight, nine, [6] one thousand, father, [7] liver, [8] we, to vomit, they, new, four, i, to pound/beat, [9] thou, who, one hundred, [10] name, to dig, to choose, stone, he/she, branch, when, to come, above, [11] thin, what, wife, mother, to hide, to grow, one, to buy, [12] dust, worm/earthworm, root, star, salt, toeat, [13] person/human being, in/inside, this, to drink, leaf, woman/female, feather, needle, to dream, [14] fog, head, where, lake, bird, rain, ear, thick, to hear, if, to fear, bad/evil, road/path, [15] fire, to sleep, right, blood, hair, tongue, to climb, sky, moon, to chew, [16] to kill, you, to shoot, to burn, man/male, to throw, child, and, to steal, thatch/roof, house, ash, [17] shoulder, how, night, tail, to cry, bone, to blow, nose, spider, other, shy/ashamed, tooth, to open/uncover, year, [18] flower, to swell, to think, to say, water, below, sea, to bite, to hunt, to count, to swim, to walk, to plant, [19] to suck, fish, old, back, egg, rope, dry, to stand, narrow, to flow, heavy, day, all, [20] to see, far, mosquito, to cook, painful/sick, wide, black, dull/blunt, snake, to spit, fat/grease, skin, green, [21] rat, sand, hand, intestines, to sew, at, to live/be alive, to turn, cold, white, to fall, to split, [22] correct/true, warm, wet, sharp, to squeeze, to hold, smoke, leg/foot, grass, [23] to sniff/smell, to lie down, to scratch, left, near, woods/forest, to stab/pierce, earth/soil, husband, [24] short, no/not, to hit, to cut/hack, wing, mouth, dog, [25] to work, lightning, meat/flesh, wind, small, to laugh, to fly, cloud, [26] to tie up/fasten, red, yellow, [27] big, neck, belly, to breathe, to sit, [28] rotten, long, [29] to know/be knowledgeable, good, [30] dirty, thunder.

Mayan: [1] green, white, tail, two, water, [2] tongue, path, die, yellow, bone, rain, tree, red, [3] name, tooth, mouth, mountain, hand, one, louse, sleep, foot, dry, [4] person, fire, ash, [5] claw, flesh, heart, blood, dog, sun, [6] neck, black, ear, fish, moon, I, [7] smoke, walk, see, earth, seed, breasts, man, skin, drink, [8] we, grease, kill, night, nose, cloud, leaf, hair, [9] head, star, stone, say, horn, woman, come, give, [10] sand, hot, cold, hear, full, [11] eat, egg, you, feather, new, bird, fly, stand, [12] bark, big, bite, eye, [13] not, [14] liver, this, burn, belly, good, swim, long, [15] know, knee, [16] round, who, [17] many, [18] all, root, [19] that, [20] lie, small, [21] what, [24] sit.

Mixe-Zoque: [1] ear, tooth, green, see, star, yellow, earth, stone, what, blood, bone, rain, white, leaf, thou, hair, drink, road/path, dry, water, red, [2] name, black, eat, head, person, bark, ashes, tongue, feather, walk, mountain, die, night, liver, new, moon, I, hand, tree, louse, hear, one, come, sun, tail, skin, good, no, clawnail, stand, [3] we, neck, sand, fire, smoke, mouth, kill, horn, knee, seed, round, give, eye, who, two, meat, root, [4] egg, fish, big, know, heart, say, bird, this, nose, fly, woman, sleep, man, [5] warm, sit, breast, belly, cloud, long, [6] cold, bite, that, dog, full, small, swim, many, [7] fat, foot, [8] lie, all, burn.

Indo-European: [1] name, four, three, fingernail, five, two, [2] we, what, I, who, tongue, when, ant, [3] how, new, egg, sun, tooth, one, [4] where, give, heart, thou, mother, full, drink, [5] sit, die, night, ear, not, star, horn, salt, day, [6] you, spit, root, snow, father, knee, [7] know, fish, this, seed, flower, eye, sew, water, long, foot, [8] warm, there, nose, other, dry, hear, sleep, [9] fly, sand, hand, bone, wind, leaf, stand, he, feather, right side, [10] eat, ice, far, old, that, liver, louse, sing, swim, with, smoke, wing, sea, moon, live, suck, [11] head, they, fire, all, rain, small, flow, blow, some, wash, dog, skin, grass, here, [12] lake, freeze, blood, bird, see, hold, red, cold, bark, laugh, green, and, breathe, lie, snake, year, meat, [13] rub, dig, sharp, tie, man, person, heavy, fall, sky, yellow, worm, cloud, straight, [14] burn, thin, short, stone, round, river, black, mouth, animal, think, fruit, in, come, [15] few, bite, neck, leg, breast, tree, play, left, guts, earth, woman, white, [16] fog, count, big, rotten, float, dust, hair, wide, thick, narrow, near, say, mountain, ashes, [17] wet, smooth, fear, pull, smell, cut, vomit, [18] right, turn, at, split, many, [19] road, belly, hunt, good, [20] fat, back, kill, hit, wipe, if, [21] stab, scratch, walk, dull, wife, [24] tail, swell, throw, woods, stick, [25] rope, fight, [26] child, [28] squeeze, husband, [29] dirty, [32] bad, push, [37] because.

Uto-Aztecan: [1] tooth, bite, [2] path, liver, water, [3] ear, moon, breasts, horn, sun, tail, eye, heavy, two, dry, [4] eat, tongue, smoke, stone, salt, nose, drink, [5] name, neck, die, knee, blood, hand, bone, [6] star, night, earth, one, sleep, clawnail, stand, [7] fire, mouth, kill, sky, dog, louse, hear, wind, root, [8] fish, grease, heart, no, cold, mountain, seed, [9] person, snow, see, bellybutton, give, ash, [10] head, know, sit, snake, egg, belly, bark, hair, leaf, meat, long, [11] black, foot, sand, hot, big, tree, white, [12]

lie, feather, vomit, sing, man, many, red, [13] skin, walk, rain, [14] burn, green, yellow, new, fly, [15] year, cloud, small, [16] bird, [17] woman, good, [18] old, come, [19] rope.

4.2 Phylogenetic inference

The Bayesian phylogenetic inference (Ronquist and Huelsenbeck, 2003) is based on the following Bayes rule:

$$f(\tau, \mathbf{T}, \theta|X) = \frac{f(X|\tau, \mathbf{T}, \theta) f(\tau, \mathbf{T}, \theta)}{f(X)}, \tag{1}$$

where X is the binary data matrix of dimensions $L \times C$, τ is the tree topology, θ is the substitution model parameters, $\mathbf{T}$ is the branch length vector of the tree. The posterior distribution $f(\tau, \mathbf{T}, \theta|X)$ is difficult to calculate analytically since one has to sum over all the possible rooted topologies ($\frac{(2L-3)!}{2^{L-2}(L-2)!}$). Therefore, Markov Chain Monte Carlo (MCMC) methods are used to calculate the posterior probability of the parameters $\tau, \mathbf{T}, \theta$. The Metropolis-Hastings algorithm (a MCMC algorithm) is used to sample the parameters from the posterior distribution. This algorithm constructs a Markov chain by proposing change to one or a block of parameters and then accepts the proposal with the following probability:

$$r = \frac{f(X|\tau, t^*, \theta)}{f(X|\tau, t, \theta)} \frac{f(t^*)}{f(t)} \frac{q(t|t^*)}{q(t^*|t)} \tag{2}$$

We assume that the parameters $\tau, \mathbf{T}, \theta$ are independent of each other. Therefore, the joint prior probability $f(\tau, \mathbf{T}, \theta)$ can be decomposed into $f(\tau) f(\mathbf{T}) f(\theta)$. In equation 2, the inference program proposes a change to a branch length t to generate a new branch length t^*. Subsequently, the likelihood of the data to the new parameters is computed using the pruning algorithm (Felsenstein, 2004), which is a special case of the Sum-Product algorithm (Jordan, 2004).

All our Bayesian analyses use binary datasets with states 0 and 1 (cf. table 2b). We employ the Generalized Time Reversible Model (Yang, 2014) for computing the transition probabilities between individual states. The rate variation across sites is modeled using a four category discrete Γ distribution (Yang, 1994). We follow Lewis (2001) and Felsenstein (1992) in correcting the likelihood calculation for ascertainment bias resulting from unobserved 0 patterns. We used a uniform tree prior (Ronquist et al., 2012a) in all our analyses which constructs a rooted tree and draws internal node heights from a uniform distribution. All our experiments are performed using MrBayes 3.2.6 (Zhang et al., 2015).

A Markov chain is initiated at a random starting point which consists of random tree, random branch lengths, and random substitution model parameters. In our experiments, we run two independent chains and sample every thousandth state in the chain and write it to the file. The runs are stopped until the average standard deviation of split (unique bipartition in a tree) frequencies does not differ beyond a threshold of 0.01. The size of the dataset influences the number of states needed for convergence. The tree space required to explore grows factorially with the number of languages whereas the cost of likelihood calculation increases linearly with the number of languages. In the case of Austronesian, we run the analysis for fifty million states; in the case of Indo-European and Uto-Aztecan, we run the analysis for ten million and five million states respectively. For the rest of the datasets, we run the Markov chains for two million states.

The Bayesian phylogenetic inference is performed on each meaning list's binary matrix. Each meaning list is processed cumulatively and the phylogenetic trees are stored to disk. In summary, each language family has binary matrices equal to the number of meanings in the family. Once the inference is finished, we proceed to evaluate the phylogenetic trees as described in the next subsection. For the evaluation we discarded the initial 25% of the trees as part of burnin and used the rest of the trees for evaluation.

4.3 Evaluation of inferred trees

The quality of inferred phylogenetic trees is evaluated using the quartet distance (Christiansen et al., 2006). The quartet distance measures the distance between two trees in terms of the number of different quartets. A quartet is a subtree consisting of four languages. A quartet for languages a, b, c, d can either be a *star* quartet when there is no internal branch separating the leaves or a *butterfly* quartet when there is an internal branch separating any two leaves from the two other leaves. The quartet distance measures the total number of different butterflies and star quartets divided by the total number of possible quartets in the tree. However, this definition of quartet distance does not work well when the gold standard tree is unresolved and has non-binary internal nodes. Expert linguistic phylogenies often have unresolved internal nodes due to lack of detailed knowledge causing the splits. In such a case the quartet distance penalizes the inferred tree because a butterfly quartet in the inferred tree is a star quartet in the gold standard tree and the former will be counted as erroneous. To remedy this, Pompei et al. (2011) proposed an extension to quartet distance known as Generalized Quartet Distance (GQD), which measures the distance between a binary tree and the gold standard tree as the number of different butterflies divided by the number of butterflies in the gold standard tree. We extract the gold standard trees from Glottolog (Hammarström et al., 2017)—a publicly available repository of language references and phylogenetic trees.

5 Results and Discussion

5.1 Comparison with gold standard trees

The results of the first test are given in Figure 1. Here, we plot the GQD as a function of number of meanings in each meaning list. Each point in the graph presents the GQD score for each meaning list. We fit a LOESS curve for each language family.[1] For all families but Austronesian the most drastic effect of list length size is seen for the first 50 or so items. For Mayan and Uto-Aztecan the GQDs appear to fluctuate not far after this point, whereas for Indo-European the drop continues to somewhat beyond 100 items. For Austronesian the regime with the most drastic drop is for the first 100 or so items, and then a slower drop sets in, continuing for the remainder of the curve.

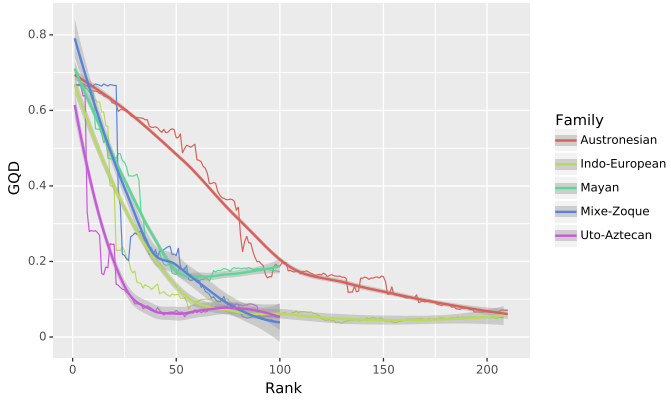

Figure 1: Lineplot of GQD of meaning lists sorted by the number of cognate classes when compared with gold standard trees. The trendlines are drawn using LOESS smoothing.

5.2 Comparison between successive trees

The second method is designed for related languages where the gold standard tree is not determined to the fullest satisfaction of the language family experts. Following this method, we compute the GQD between the inferred trees from successive meaning lists and then plot the GQD. The results, shown in Figure 2 are remarkably similar to the ones making reference to a gold standard tree, displayed in Figure

[1] All the plots are generated using `plotnine` python library available at `http://plotnine.readthedocs.io/en/stable/index.html`.

1. The visual impression of similarities in the shape of the curves is supported by Pearson correlations, which all lie in the range $0.890 > R^2 > 0.967$.

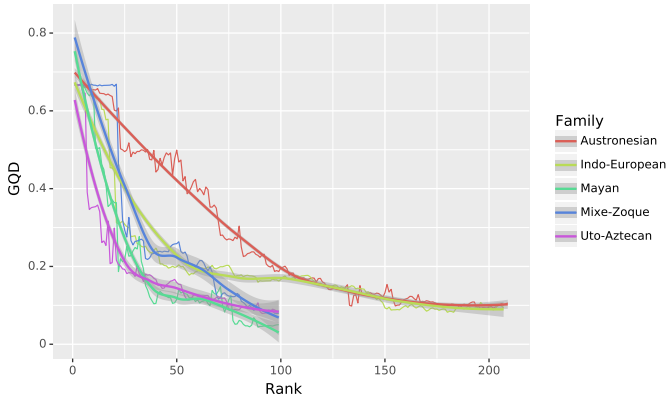

Figure 2: Lineplot of GQD between two successive meaning lists. We show the LOESS fit for each family.

Finally, we provide Figure 3, which shows plots for the remaining six language groups in our sample for which expert trees are generally absent or less reliable.

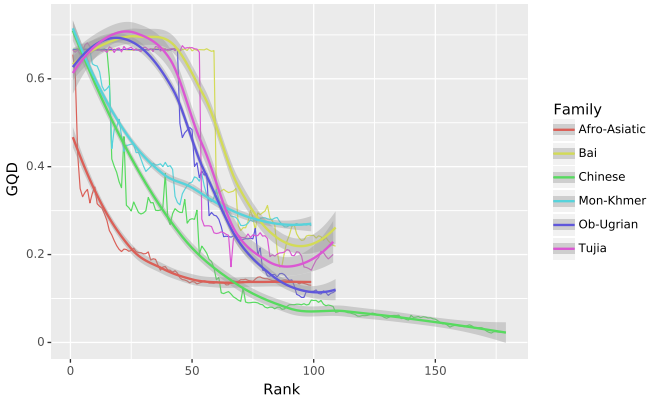

Figure 3: Lineplot of GQD between two successive meaning lists for the remaining language groups in our sample. We show the LOESS fit for each family.

5.3 Analyzing the results

In order to replace eye-balling with a more consistent identification of the cut-off between the approximately monotonic regime of each curve and the regime (if any) where fluctuations set in, we tested two different approaches. The first approach, inspired by Grieve et al. (2017), was as follows. We compute the Spearman rank correlation between the mean GQDs in each segment of the left tail with the corresponding numbers of meaning items, increasing the size of the segment in steps of one datapoint from 2 to the maximal number of meanings. For instance, for a segment of the curve corresponding to rank 1-10 we correlate the 10 GQD values with the series of numbers 1-10. As long as the result continues to be a strong negative correlation one could consider the curve to be approximately monotonically falling. Here we experimented with different cut-offs, including $\rho = -0.96$. If the increase of a segment of the curve by one meaning results in a weaker correlation than the cut-off, this point could then be considered to pertain to the fluctuating regime. As it turned out, there was no sensible way of defining a cut-off that would work across all eleven cases. Moreover, this method was too sensible to local fluctuations. Instead we applied a second and much simpler approach, which is to assume that the relevant point identifying the point where the GQD ceases to drop as simply the minimum. Table 3 summarizes the results.

Family	Lgs.	Meanings	Min (gold)	CC (gold)	Min (consec.)	CC (consec.)
Austronesian	96	210	203	3251	197	3091
Mayan	30	100	61	310	90	655
Mixe-Zoque	10	100	100+	300+	99	292
Indo-European	52	208	130	1078	181	1883
Uto-Aztecan	31	100	100+	846+	92	711
Afro-Asiatic	25	100			100+	1273+
Bai	9	110			87	137
Chinese	18	180			174	1115
Mon-Khmer	16	100			90	606
Ob-Ugrian	21	110			96	172
Tujia	5	109			100	152

Table 3: The information on languages and total meanings from Table 1 is repeated, and in addition the last four columns indicate the point where the curves reach their minimum ("Min"), and the corresponding number of cognate classes ("CC") for respectively the method using a gold standard ("gold") and the one using comparisons between consecutive size-increased meaning lists ("consec."). A plus sign following a number indicates that the minimum is reached when all items are used, such that we can suspect, although we cannot know for certain, that adding some more items (also implying more cognate classes) would continue to bring improvements.

The results in Table 3 suggest the existence of a linear correlation between the optimal number of meanings and the number of languages to be classified for the gold standard method. The correlation is strong ($R^2 = 0.786$) but only barely below the 0.05 significance level ($p = 0.0452$). As just mentioned, however, the real impact should come not from the mere number of items, but from the number of cognate classes they produce. Thus, not unexpectedly, the correlation is stronger yet between the optimal number of cognate classes and the number of languages ($R^2 = 0.923$) and also significant ($p = 0.0094$). For the method using consecutive datasets for eleven language families and groups the correlation between the number of meanings and the number of languages is weak albeit significant ($R^2 = 0.537$, $p = 0.010$); but again the situation radically improves when the number of languages is correlated with the number of cognate classes ($R^2 = 0.876$, $p < 0.0001$).

The linear model for the optimal number of cognate classes as a function of the number of languages intercepts at -11.24. Since to classify zero languages zero cognate classes are needed the intercept can be set at zero, something which only requires a small adjustment, and the slope of the resulting function will then be 32.397, indicating that ∼33 cognate classes per language classified is optimal. The gold standard results from the smaller set of five families gives us a slope of 29.297 when the intercept is at zero, indicating that ∼30 cognate classes is optimal. We choose to place more weight on the more conservative of these estimates.

Our result should be highly useful in the design of future studies in Bayesian phylogenetics. The initial decision on how many items to include in a lexical dataset can be difficult since the amount of resulting cognate classes may be unknown, but across our case studies the number of items required was always at least 87. Thus, for a small or medium-sized family (∼30 languages or less) the researcher could initially aim at 100 items, but for larger families this will clearly not suffice. For around 31-100 languages 200 items may be needed, and for yet larger families more than that. Counting cognate classes will yield a much more precise estimate of the adequacy of the sample. As an example, a recent study of Dravidian languages (Kolipakam et al., 2018) reports that 778 cognate classes were used for inferring a phylogeny for 20 languages. Since $20 \times 33 = 660$, we now have reasons to believe that this is an adequate sample. The sample of 20 languages, however, only contains around one fourth of the total set of Dravidian languages (Hammarström et al., 2017), so a future investigation may be directed at extending the phylogeny. In that case it can be anticipated that it will not suffice to add data for the list

of meanings used in Kolipakam et al. (2018)—rather, new items may have to be added.

6 Conclusion

This paper has, for the first time, addressed the question of the size of meaning lists required for the optimal Bayesian-based inferencing of linguistic phylogenies. In our search for this optimum, that is, the point where Bayesian inference stops improving, we find the following:

- Bayesian inference does not necessarily improve when given more cognate classes. In fact, more data can reduce the fit with expert trees, as is seen in the case of the Mayan family. Such a situation can be explained by a high level of borrowing where less stable words are exchanged more among lineages than more stable words, such that including these less stable words is not necessarily advantageous. (Indeed, Mayan languages are known to borrow heavily from one another, cf. Wichmann and Brown (2003)).

- The curves representing fits with gold standard trees and the ones representing fits between trees produced by successively adding items in descending order of stability are highly similar, so the latter can be used as a proxy for the former in order to increase the number of families sampled for the purpose of determining the optimal list size.

- The optimal list size for Bayesian phylogenetic inference depends on the number of cognate classes represented by the words corresponding to the items on the meaning list, and the number of cognate classes, in turn, is strongly correlated with the number of languages under investigation. Our results indicate that the optimal list size is one that produces around 33 cognate classes times the number of languages to be classified.

Acknowledgments

The first author is supported by BIGMED project (a Norwegian Research Council LightHouse grant, see bigmed.no). The second author is supported by a subsidy of the Russian Government to support the Programme of Competitive Development of Kazan Federal University. The experiments were performed when both authors took part in the ERC Advanced Grant 324246 EVOLAEMP project led by Gerhard Jäger. All these sources of support are gratefully acknowledged.

References

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

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

Lyle Campbell and William J. Poser. 2008. *Language Classification: History and Method*. Cambridge University Press, Cambridge.

Chris Christiansen, Thomas Mailund, Christian N. S. Pedersen, Martin Randers, and Martin Stig Stissing. 2006. Fast calculation of the quartet distance between trees of arbitrary degrees. *Algorithms for Molecular Biology*, 1(1).

Michael Cysouw, Søren Wichmann, and David Kamholz. 2006. A critique of the separation base method for genealogical subgrouping, with data from Mixe-Zoquean. *Journal of Quantitative Linguistics*, 13(2-3):225–264.

Alexei J. Drummond, Marc A. Suchard, Dong Xie, and Andrew Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973.

Michael Dunn, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

Michael Dunn. 2012. Indo-European lexical cognacy database (IELex). http://ielex.mpi.nl/.

Běijīng Dàxué, editor. 1964. *Hànyǔ fāngyán cíhuì [Chinese dialect vocabularies]*. Wénzì Gǎigé.

Joseph Felsenstein. 1992. Phylogenies from restriction sites: A maximum-likelihood approach. *Evolution*, 46(1):159–173.

Joseph Felsenstein. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Russell D. Gray, Alexei J. Drummond, and Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, 323(5913):479–483.

Simon J. Greenhill, Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C. Levinson, and Russell D. Gray. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, 114(42):E8822–E8829.

Jack Grieve, Andrea Nini, and Diansheng Guo. 2017. Analyzing lexical emergence in Modern American English online. *English Language & Linguistics*, 21(1):99–127.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2017. *Glottolog*. Max Planck Institute for Evolutionary Anthropology, Leipzig. http://glottolog.org/.

Eric W. Holman, Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica*, 42(3-4):331–354.

Eric W. Holman, Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, et al. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6):841–875.

Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Long Papers*, pages 1204–1215, Valencia. Association for Computational Linguistics.

Gerhard Jäger. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, 3(2):245–291.

Michael I. Jordan. 2004. Graphical models. *Statistical Science*, 19(1):140–155.

Vishnupriya Kolipakam, Fiona M. Jordan, Michael Dunn, Simon J. Greenhill, Remco Bouckaert, Russell D. Gray, and Annemarie Verkerk. 2018. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science*, 5:171504.

Alfred Louis Kroeber. 1963. *Yokuts Dialect Survey*, volume 11(3) of *University of California Publications: Anthropological Records*. University of California Press, Berkeley.

Paul O. Lewis. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6):913–925.

Alexander Militarev. 2000. Towards the chronology of Afrasian (Afroasiatic) and its daughter families. In Colin Renfrew, April McMahon, and Larry Trask, editors, *Time Depth in Historical Linguistics*, pages 267–307. McDonald Institute for Archaelogical Research, Cambridge.

Wick R. Miller. 1984. The classification of the Uto-Aztecan languages based on lexical evidence. *International Journal of American Linguistics*, 50(1):1–24.

R. L. Oswalt. 1970. The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior*, 3(3):117–129.

Mark Pagel and Andrew Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, 53:571–581.

Mark Pagel and Andrew Meade. 2006. Estimating rates of lexical replacement on phylogenetic trees of languages. In Peter Forster and Colin Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*, pages 173–182. McDonald Institute Monographs, Cambridge.

Mark Pagel, Quentin D. Atkinson, and Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449(7163):717–720.

Ilia Peiros. 1998. *Comparative Linguistics in Southeast Asia*. Number 142. Pacific Linguistics, Research School of Pacific and Asian Studies, National University of Australia.

Filippo Petroni and Maurizio Serva. 2010. Lexical evolution rates derived from automated stability measures. *Journal of Statistical Mechanics: Theory and Experiment*, 2010:P03015.

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS one*, 6(6):e20109.

Taraka Rama and Lars Borin. 2013. N-gram approaches to the historical dynamics of basic vocabulary. *Journal of Quantitative Linguistics*, 21(1):50–64.

Taraka Rama and Lars Borin. 2015. Comparative evaluation of string similarity measures for automatic language classification. In Ján Mačutek and George K. Mikros, editors, *Sequences in Language and Text*, pages 203–231. Walter de Gruyter.

D. F. Robinson and L. R. Foulds. 1979. Comparison of weighted labelled trees. *Combinatorial Mathematics*, 6:119–126.

Fredrik Ronquist and John P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Fredrik Ronquist, Seraina Klopfstein, Lars Vilhelmsen, Susanne Schulmeister, Debra L. Murray, and Alexandr P. Rasnitsyn. 2012a. A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera. *Systematic Biology*, 61(6):973–999.

Fredrik Ronquist, Maxim Teslenko, Paul van der Mark, Daniel L. Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A. Suchard, and John P. Huelsenbeck. 2012b. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542.

Naruya Saitou and Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

George S Starostin. 2013. Annotated swadesh wordlists for the tujia group. *The Global Lexicostatistical Database, RGGU, Moscow. URL: http://starling. rinet. ru/new100/tuj. xls*.

Morris Swadesh. 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.

David Thomas. 1960. Basic vocabulary in some Mon-Khmer languages. *Anthropological Linguistics*, 2(3):7–11.

Feng Wang. 2006. *Comparison of languages in contact: the distillation method and the case of Bai*, volume 3. Institute of Linguistics, Academia Sinica.

Søren Wichmann and Cecil H. Brown. 2003. Contact among some mayan languages: inferences from loanwords. *Anthropological Linguistics*, 45(1):57–93.

Søren Wichmann and Eric W Holman. 2013. Languages with longer words have more lexical change. In Lars Borin and Anju Saxena, editors, *Approaches to Measuring Linguistic Differences*, pages 249–281. Mouton de Gruyter, Berlin.

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

Søren Wichmann, André Müller, and Viveka Velupillai. 2010b. Homelands of the world's language families: A quantitative approach. *Diachronica*, 27(2):247–276.

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

Ziheng Yang. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular evolution*, 39(3):306–314.

Ziheng Yang. 2014. *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford.

Chi Zhang, Tanja Stadler, Seraina Klopfstein, Tracy A. Heath, and Fredrik Ronquist. 2015. Total-evidence dating under the fossilized birth–death process. *Systematic Biology*, 65(2):228–249.

M Zhivlov. 2011. Annotated swadesh wordlists for the ob-ugrian group (uralic family). *The Global Lexicostatistical Database. RGU*.

A Supplemental Material

The code and data used in this paper are uploaded as a zip file along with this paper. In addition, they are available for download via Zenodo at `https://doi.org/10.5281/zenodo.1287317`