Distance-based Phylogenetic Inference Algorithms in the Subgrouping of Dravidian Languages

Taraka Rama and Sudheer Kolachina

1. Introduction[1]

Historical linguistics has as one of its main aims the classification of languages into language families. The internal classification of languages within a language family is known as *subgrouping*. Subgrouping is concerned with the way daughter languages within a single family are related to one another and, therefore, with the branching structure of the family tree (Campbell 2004). In much of the literature on the subject, *shared innovations* are discussed as the only acceptable criteria while establishing subgroups within a language family. Within the framework of lexical diffusion, it has been shown that it is possible to infer subrelations among a set of related languages from the distributional pattern of changed (innovations) versus unchanged (retentions) cognates across these languages even with respect to a single sound change (Krishnamurti 1983).

The origins of quantitative methods in historical linguistics can be traced back to the lexicostatistical methods and glottochronology of Swadesh (1952, 1955). Although Swadesh's methods are criticized to this day as being fraught with untenable assumptions, it is indisputable that his work marks the beginning of a search for alternatives to the traditional comparative method. See McMahon and McMahon (2005) for a historical overview of the use of quantitative methods for language classification. In particular, recent years have seen a rapid increase in interest in the application of phylogenetic inference methods, most of which come from computational biology, to diachronic language data leading to the emergence of a distinct research area, increasingly being referred to as Computational historical linguistics (CHL, henceforth). The basic intuition

in such research is that these methods, which were developed to infer (genetic) phylogeny from gene sequences, can do so from language data too, which also consist of sequences. Interestingly, this is not the first time that a cross-pollination of ideas between the fields of biology and linguistics has taken place (Atkinson and Gray 2004).

Phylogenetic inference methods that have been used for estimating linguistic phylogeny in recent CHL literature are either *character-based* or *distance-based*. Character-based methods such as Maximum Parsimony (MP) (Felsenstein 2003) and Bayesian inference (Felsenstein 2003) estimate phylogeny of a set of related languages from character-based data. A *character* can represent any aspect of language evolution lexical, phonological, morphological or syntactic change. For example, a lexical character encodes information about the presence or absence of a cognate across the languages that are to be subgrouped. Each language is assigned a state with respect to this character based on the presence or absence of that cognate in the language. Two languages would have the same state if and only if the cognate represented by this lexical character is either present or absent in both languages. Similarly, phonological, morphological and syntactic characters encode information about the presence or absence of corresponding types of language change. Thus, in a character-based dataset, each language is represented as a sequence consisting of states of that language with respect to the different characters considered. Distance-based methods such as Unweighted Pair Group Method with Arithmetic means (UPGMA) and Neighbor Joining (NJ) (Felsenstein 2003) estimate linguistic phylogeny from a distance matrix containing pairwise inter-language *phylogenetic distances*. Different measures have been discussed in the literature as estimates of phylogenetic distance between languages. One common practice is to estimate pairwise phylogenetic distance from character-based data as the Hamming distance between character sequences representing languages. All the above methods, whether character-based (MP, Bayesian inference) or distance-based (UPGMA and NJ), assume linguistic phylogeny to be tree-like. However, this assumption is problematic in the context of linguistic areas where shared linguistic traits could also be the result of convergence due to extensive language contact. Some recent works such as Nakhleh et al. (2005), Huson and Bryant (2006) propose the use of phylogenetic networks to address the limitations of the tree model of language evolution. See the tutorial on linguistic phylogeny by Nichols and Warnow (2008) for a comprehensive and detailed discussion about network-based phylogenetic inference methods.

Diachronic datasets used in recent literature on inference of linguistic phylogeny are lexical datasets, usually derived from Swadesh lists. A Swadesh list is a short (of length 40 – 200), culturally universal list of meanings that are supposed to be highly resistant to borrowing. Such lists are most often compiled from etymological dictionaries. A character-based dataset can be obtained from the Swadesh lists by grouping the lexical items corresponding to a meaning slot in different languages into cognate classes based on the cognacy judgments available in the etymological dictionary. As mentioned previously, languages with cognates belonging to the same cognate class are coded as being in the same state with respect to that lexical character. Similarly, phylogenetic distances required for the application of distance-based phylogenetic inference methods can also be estimated from Swadesh lists. The ASJP project (Brown et al. 2008), a notable recent work on the application of distance-based methods for language classification, estimates inter-language distances as the aggregate sum of the degree of cognateness between pairs of strings in parallel Swadesh lists of two languages, where degree of cognateness is measured using a metric based on Levenshtein distance.[2] In fact, even the traditional lexicostatistical method is a distance-based method that treats the percentage of shared cognates for each language pair as an estimate of the phylogenetic distance between them. It must be noted that while the use of Swadesh lists or rather lexical datasets is quite convenient when etymological dictionaries are available, lexical data is relatively more prone to borrowing compared to datasets containing phonological, morphological or syntactic features.

Over the past few years, phylogenetic inference methods have been applied to data from well-studied large-scale language families such as Indo-European, Austronesian among others to address interesting questions about their time depth (Gray and Atkinson 2003), spatial spread (Holman et al. 2008) and prehistoric migration patterns (Gray et al. 2009). As discussed above, the availability of diachronic datasets in electronic formats – Swadesh lists with cognate judgments, comparative feature datasets and typological databases, is a prerequisite for such studies. The focus of our work in this paper is the Dravidian language family. Although Dravidian languages are one of the few instances of the successful application of the comparative method to reconstruct the proto-language (Campbell 2004), there is very little work on the application of phylogenetic inference methods to these languages. Such an application will be interesting not only to compare the automatically inferred phylogenetic trees against the

manually constructed family tree but also to look for possible solutions to unresolved questions about the subgrouping of Dravidian languages. The lack of diachronic datasets is the main hurdle that needs to be overcome to take up such studies. In this paper, we present two new diachronic datasets for Dravidian languages created from existing resources which can be used in different kinds of quantitative studies on these linguistic phylogeny of these languages. We also explore the application of distance-based phylogenetic inference methods to the task of subgrouping Dravidian languages. In particular, we study the performance of this class of methods with respect to a specific subgrouping question discussed in recent literature. In addition to subgrouping, these diachronic datasets can also be used to study possible correlations between the current spatial distribution of these languages and the genetic subrelations among them. The geographical discontiguity of Dravidian languages is an interesting puzzle which, if solved, will open up new directions in the study of the linguistic prehistory of the subcontinent.

The paper is organized as follows. In section 2, we present a few details about Dravidian languages by way of providing the reader with a background about these languages. We briefly review three extant classifications and point out the differences among them. In section 2.3, we describe the specific subgrouping issue of ternary versus binary branching of Proto-Dravidian. Section 3 summarizes previous work on the application of quantitative methods to study the diachrony of Dravidian languages. In section 4, we describe the four diachronic datasets created from existing resources. In section 5, we describe the various distance-based methods used in our experiments, the method used to obtain distance matrices from character-based data and also, the strategy we followed for rooting the unrooted trees returned by the phylogenetic inference methods. In section 6, we present the trees and networks resulting from our experiments. We summarize our findings and conclude in section 7.

2. Dravidian Languages

This section is divided into three subsections. In the first subsection, we present some general details about the Dravidian language family by way of providing the readers with a background about these languages. In the next subsection, we briefly review three subgrouping schemes from different sources. In the last subsection, we discuss a specific problem in the subgrouping of Dravidian languages, namely ternary versus binary

branching of Proto-Dravidian, which we will attempt to address in our experiments.

2.1. Dravidian language family

The Dravidian language family is the world's fifth largest language family with over 200 million speakers in South Asia (Krishnamurti 2003). The majority of the languages are geographically located in the southern and central parts of the Indian sub-continent with a few scattered pockets in Northern India (Kurux, Malto) and Nepal (Kurux) and a lone population scattered across Pakistan and Afghanistan (Brahui). There are four major languages with long literary, written traditions – Tamil, Malayalam, Kannada and, Telugu. They are written, if at all, using scripts of neighbouring languages.

Krishnamurti (2003) is a compendious work covering various aspects of the Dravidian languages. There exists a voluminous body of literature in the area of Dravidian linguistics owing to the efforts of many scholars. One resource that needs mention for its potential value to research on various aspects of the Dravidian language family is the Dravidian Etymological Dictionary (DEDR)[3] (Burrow and Emeneau 1984).

2.2. Subgrouping of Dravidian languages

In this section, we briefly the describe three main subgrouping schemes for Dravidian languages discussed in the literature.

2.2.1. Krishnamurti (2003)

Figure 1 shows the family tree of the Dravidian languages discussed in Krishnamurti (2003). This tree can be considered as the "gold standard" tree as it is the one widely accepted. However, there still remains some unresolved issues (Krishnamurti 2003) in this classification, such as the following:

– The position of the Nilgiri languages (Toda, Kota, Irula, Badaga and Kurumba) in relation to Tamil and Kannada is not clear.

– The position of Tulu in the family tree is doubtful.

– The placement of Koraga in the subgrouping scheme is undecided.

– The position of Naikri in Central Dravidian subgroup is doubtful.

The above uncertainties are indicated using broken lines in figure 1. It is interesting to note that most of these uncertainties pertain to the South Dravidian I subgroup.

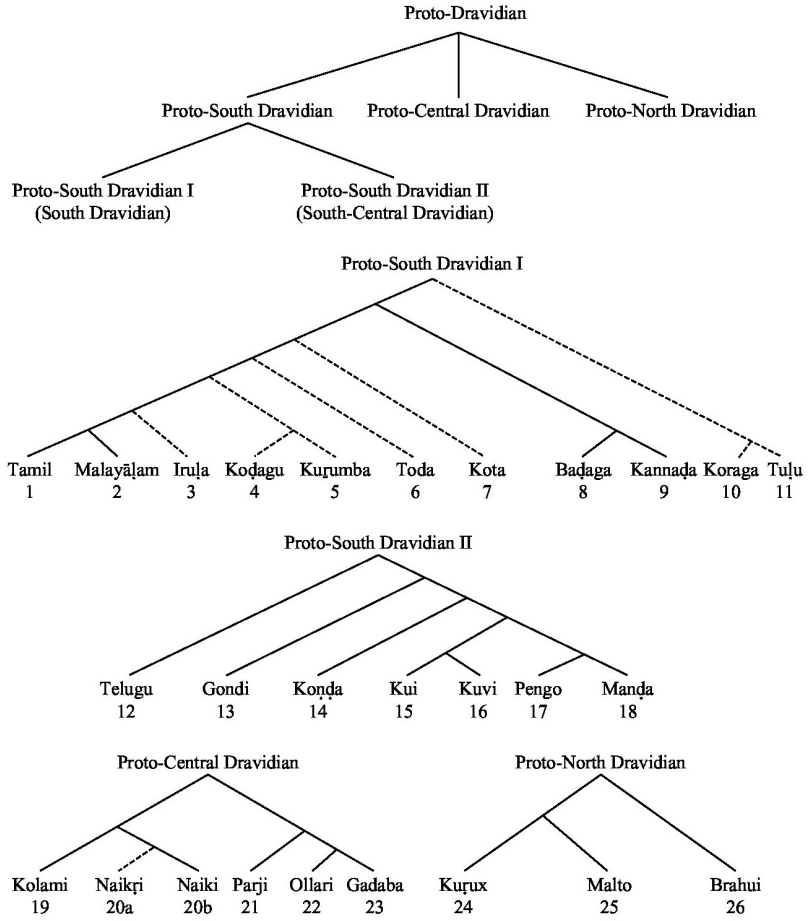

Figure 1. Family tree of Dravidian languages given in Krishnamurti (2003). Broken lines depict uncertain relationships.

2.2.2. WALS

The *World Atlas of Language Structures* (Haspelmath et al. 2008) provides a two level classification of the Dravidian languages (23 languages) with the following subgroups (in order of geographical contiguity):

– South Dravidian: Badaga, Betta Kurumba, Kannada, Kodava, Kota, Malayalam, Tamil, Tamil (spoken), Toda, and, Tulu
– South-Central Dravidian: Gondi, Konda, Koya, Kui, Kuwi (Kuvi; a name variant), Pengo, and, Telugu
– Central Dravidian: Gadaba, Kolami and Parji
– Northern Dravidian: Brahui, Kurukh and Malto

The WALS classification does not include Irula, Koraga, Naiki and, Ollari (present in Krishnamurti's classification) since, WALS does not include them.

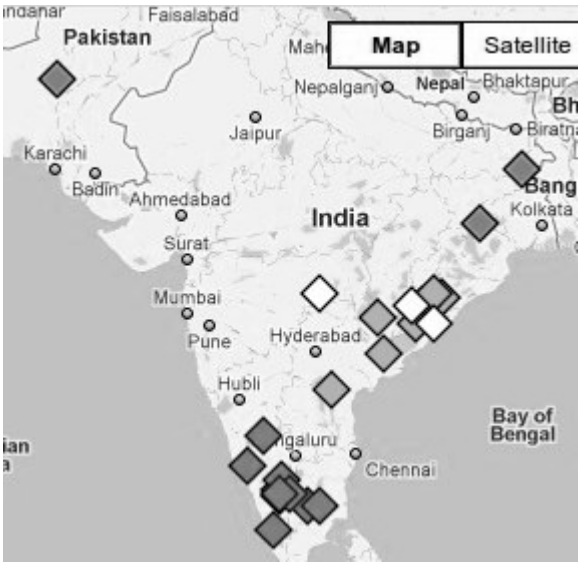

Figure 2. Geographical distribution of Dravidian languages in WALS database. Dark diamonds above the white diamonds are North Dravidian; White

diamonds are Central Dravidian; Light-dark diamonds are South-Central Dravidian; The darkest diamonds are South Dravidian.

2.2.3. Ethnologue

Ethnologue (Lewis 2009) lists a far larger number of languages and dialects (85) than WALS or Krishnamurti (2003). Figure 3 displays the *Ethnologue* classification for only those languages present in the ASJP database.[4] The *Ethnologue* tree for Dravidian languages shows four subgroups attached to the root of the tree and the highest level subgrouping is unresolved.

Concerning the internal classification within each subgroup :

1. Proto-North Dravidian is polytomous (more than two children).
2. South Dravidian I subgroup's ancestral node is polytomous as well.

We can conclude that the Ethnologue tree is at least not as resolved as the tree given by the comparative method at the highest level subgrouping of the Dravidian language family (Krishnamurti 2003) and that there are quite a number of nodes which are polytomous. The Ethnologue tree shows the same subgroups as the tree given by Krishnamurti (2003). It differs largely in the placement of languages in the SDI subgroup, where the two trees differ in the placement of Koromfe and Kodava in SDI subgroup.

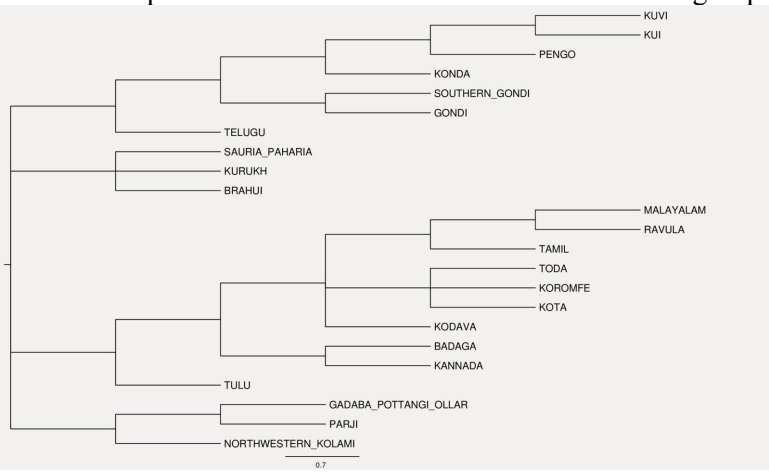

Figure 3. Ethnologue Tree for the Dravidian languages present in ASJP database

2.3. Ternary vs. binary branching

There are two prevailing thoughts about the main subdivision of the Dravidian languages: ternary vs. binary (Krishnamurti 2003). According to the ternary hypothesis, Proto-Dravidian (PD) has three branches: Proto-North Dravidian (ND), Proto-Central Dravidian (CD) and Proto-South Dravidian (SD), which is further split into South Dravidian I (SD I) and South Dravidian II (SD II). This is the subgrouping adopted in Krishnamurti (2003). This subgrouping is established on the basis of isogloss maps constructed using 27 features from comparative phonology and morpho-syntax. An alternate subgrouping option is to have a binary division of Proto-Dravidian into Proto-North Dravidian (ND) and Proto-South-Central Dravidian (SCD). Proto-South-Central Dravidian further splits into Proto-South Dravidian and Proto-Central Dravidian. In this regard, Krishnamurti (2003) notes that although in general a binary division of a speech community is more likely than a ternary, there is scant evidence to set up a common stage of South and Central Dravidian. In this paper, we explore the application of distance-based methods to the datasets described previously in our search for a solution to this subgrouping problem.

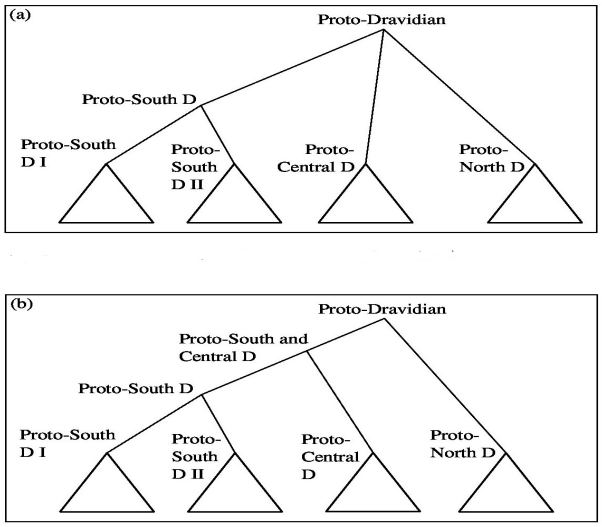

Figure 4. Alternative splits of Proto-Dravidian (Krishnamurti, 2003)

3. Background

The aim of this section is to provide a brief overview of the previous attempts at applying quantitative/computational methods to the internal classification of Dravidian languages.

The first attempt to apply quantitative methods to Dravidian languages was made by Andronov (1964). He applied the method of glottochronology proposed by Swadesh (1952, 1955) to word lists of nineteen Dravidian languages. Although this work is several decades old, it remained largely unnoticed until it was reviewed by Krishnamurti (2003). This early glottochronological study was followed by two other similar lexicostatistical studies: Kameswari (1969) and Namboodiri (1976). All the three works received critical review in Krishnamurti (2003). Krishnamurti's main objection relates to the wide variation in the divergence times predicted by these studies. According to Andronov (1964), the Tamil and Telugu sub-branching took place around 10th century BC, whereas according to Kameswari (1969), it was in 4th century BC to 4th century AD. Krishnamurti (2003) takes this divergence times as an evidence of the unreliability of the glottochronological/lexicostatistic technique.

Further, he also notes that the glottochronological/lexicostatistic approach is incapable of dealing with differing rates of lexical replacement in different languages. For example, Brahui, as a result of heavy borrowing from Balochi and Indo-Aryan languages, has retained only 15 percent of the native lexemes. Due to such a high degree of cognate loss, the glottochronological method estimates that Brahui separated from the rest 5000 BP (5000 years before present), a date which is untenable in the light of other kinds of evidence (Elfenbein 1987).

In addition to these early studies, there are two other interesting applications of computational methods for sub-grouping Dravidian languages before the advent of phylogenetic inference methods from computational biology. These two studies are Krishnamurti (1978) and Krishnamurti et al. (1983). Both these works attempt to show that within the framework of lexical diffusion, data pertaining to just one sound change is sufficient to discover the internal classification of a set of related languages.

Krishnamurti (1978) aims at showing that a single sound change in progress is sufficient for the internal classification of six South-Central Dravidian languages.[5] Krishnamurti (1978) compiles a list of cognate sets from six South-Central Dravidian languages qualified for a particular sound change (apical displacement) using the *DED*.[6] Since the sound change

considered was one in progress, in each language, some of the items would have changed while others would have remained unaffected. The number of changed items that a language shares with a sister language is treated as a measure of their "proximity". A multi-dimensional scaling algorithm was applied to a matrix containing such pairwise proximity values (numbers of shared cognates-with-change) for all six languages. The resultant scatter plot makes the following predictions: Kui and Kuvi are closest to each other and similarly, Pengo-Manda form another cluster and Konda is closer than Gondi to the rest. All these predictions are in agreement with the relations obtained from the manually constructed standard tree for this sub-family.

In a sequel to this work, Krishnamurti et al. (1983) apply another interesting quantitative method to a subset of the earlier lexical diffusion data to setup subrelations for the same set of languages. The dataset used in this study contains 63 cognate sets which are all qualified for the apical displacement sound change. This work claims that genetic subrelations can be inferred from the distribution pattern of just one sound change in progress. Their method for doing this can be described briefly as containing the following steps: Encode the status of the sound change in each cognate set (either changed [1] or unchanged [0]) for a language. Enumerate all possible binary branching trees for the six languages. The number of changes required to explain a cognate set is taken to be the score of a tree for that cognate set. The tree with the lowest score (least number of accumulated changes) over all the cognate sets is selected as the best tree explaining the data. Krishnamurti et al. (1983) find that the best tree obtained thus is "identical" to the standard tree manually constructed using the comparative method. It must be noted that had the number of languages been greater than six, the authors would have encountered the tree combinatorial explosion problem. Although the authors claim their approach to be novel, Embleton (1986) notes that this tree-scoring criterion had in fact already been explored in the historical linguistic literature and was being independently rediscovered in this work for the third time.

McMahon and McMahon (2007) note that the general linguistic scenario in South Asia where contact between four language families – Indo-Aryan (Indo-Iranian sub-family of Indo-European family), Sino-Tibetan, Munda (of Austroasiatic) and Dravidian – over several millenia is well-attested and attempt to cast new light on the genetic classification of South Asian languages by applying network building programs rather than phylogenetic tree inference methods. According to the authors, the

rationale behind doing this is that in extensive contact situations such as evidenced in South Asia, evolution of languages cannot be tree-like. The tree model is incapable of handling the wide-spread intra-family borrowing which is highly likely in the South Asian context. McMahon and McMahon (2007) create two sets of data the thirty most conservative and the thirty least conservative items for Indo-Aryan languages taken from the older Dyen et al. (1992) database of Indo-European Swadesh lists and apply the Neighbour Network method to each of these datasets. They observe that the resulting networks do not differ significantly, which can be taken to suggest that wide-spread family-internal borrowing affects the most conservative vocabulary items as much as it affects the least conservative ones. The main drawback of this work, in our opinion, is that the datasets used are not large enough for the results to be of general interest. It would be interesting to repeat this study using a larger dataset containing data for more Indo-Aryan languages.

Rama et al. (2009) apply different phylogenetic inference methods such as Maximum Parsimony, UPGMA, Neighbor-joining and Bayesian phylogenetic inference to the datasets of Krishnamurti (1983) to infer the phylogeny of six South-Central Dravidian languages. In this exploratory study, they report the output trees and discuss the similarities and differences with the standard tree. They also point out that the approach discussed in Krishnamurti et al. (1983) is a restricted case of the well-known maximum parsimony method known as Dollo's parsimony.

Kolachina et al. (2010) apply the maximum parsimony method (MP) to address a specific problem pertaining to the subgrouping of Dravidian languages. Krishnamurti (2003) discusses two subgrouping alternatives – one with ternary branching of Proto-Dravidian and another with binary branching, finally adopting the ternary branching alternative for the highest order subgrouping, based on isoglosses of 27 features from comparative phonology, morphology and syntax. Kolachina et al. (2010) convert this feature data into character sequences of 1/0 bits and apply the maximum parsimony method for internal classification. Since MP returns an unrooted tree, they root the output tree using ND as the outgroup. This is done because both the subgrouping alternatives have in common North Dravidian as the outgroup. The authors observe that branch lengths returned by MP do not support a ternary branching at the highest level and thus select the binary branching alternative.

4. Datasets

In this section, we give a brief description of four datasets to be used in our subgrouping experiments. We have created two of these datasets (1 and 2), one based on the DEDR and the other based on Krishnamurti (2003).

– We created a new character-based dataset using the *DEDR*. The *DEDR* is a compilation of 6027 cognate sets for 28 Dravidian languages (29 if Pālu Kuṟumba and Ālu Kuṟumba are counted as separate languages)[6] belonging to the Dravidian language family. Each cognate set in DEDR is identified by an entry number. In a few instances, cognate sets share the same entry number. It is not clear why two widely differing cognate sets are listed under the same number. There are 5548 cognate sets with unique entry number. With regard to each pair of cognate sets with the same identification number, we included the first cognate set in our dataset. Further, we also excluded those cognate sets which are probable borrowings from Indo-Aryan to Dravidian. For a cognate set to appear in the database, it is not necessary that the cognate set has corresponding entries in all 28 languages. The language entry consists of the lexical item along with its possible variants and its meaning in that language. It is worth pointing out here that in the *DEDR* cognate sets, the meanings of the corresponding items across individual languages is not necessarily the same. This characteristic feature of our database distinguishes this dataset from the datasets used in previous works on phylogenetic inference (e.g., the well-known IE dataset compiled by Dyen et al. (1992)), where a lexical item has the same meaning across all the languages. Furthermore, doubtful cognate judgments are indicated in DEDR, by a "?", was removed. There are cases similarly, where a cognate can belong to more than one cognate set and is cross-referenced. We also excluded such items from our dataset. The final DEDR based character dataset consists of 4169 characters containing data from 28 Dravidian languages. Each cognate set is represented as a binary character with the presence or absence of a language coded as 1/0. The state 0 for a character in a language could be either due to the real absence of the lexical item in that language or, simply, due to missing data. At this point, there is no way of differentiating between these two possibilities. Our dataset, we refer to as CDR (based on complete DEDR).
– The second database which we built is based on Krishnamurti (2003). Krishnamurti (2003) provides reconstructions for 656 cognate sets in an

appendix along with the reconstructed proto-forms. Each reconstruction is given along with its *DEDR* entry number. Here again, as in the case of *DEDR*, the entry numbers are not unique. After removing cognate sets with duplicate entry numbers, the dataset is left with 348 characters each representing a lexical item. This dataset can be used not only to infer linguistic phylogeny but also to evaluate approaches that claim to automate reconstruction of proto-forms. We refer to this as ADR (based on the appendix in Krishnamurti 2003).

– Apart from providing cognate sets and their reconstructed proto-forms, Krishnamurti (2003) also provides a list of phonological, morphological and syntactic features which form the basis of subgrouping discussed in that work (figure 1). Kolachina et al. (2010) encode these features as characters and create a character-based dataset. A character can have one of three states: 1 indicating presence of a feature, 0 indicating absence and ? indicating unknown. It must be noted that character-based datasets 1, 2 and 3 need to be converted into distance matrices in order for distance-based phylogenetic inference methods to be applicable. We convert these character-based datasets to a distance matrix by computing pair-wise length-normalized Hamming distance between the languages represented as character sequences. We refer to this as the Comparative Features database.
– The fourth database is based on Swadesh lists for Dravidian languages from the ASJP database. Estimation of inter-language distances from the Swadesh word lists is another direction in which a number of recent efforts (Serva & Petroni 2008; Holman et al. 2008) have been directed. One such notable effort is the ASJP (Holman et al. 2008) project which estimates inter-language distance as the aggregate sum of the degree of cognateness between pairs of strings in parallel Swadesh lists of two languages; where degree of cognateness is measured using a metric based on Levenshtein distance (See note 1). As part of this effort, a database of Swadesh lists for 4817 languages was compiled and a distance matrix containing all possible pairwise inter-language distances was constructed. The ASJP database contains 40-item Swadesh lists for 23 Dravidian languages. In this work, we explore the ASJP approach too for subgrouping Dravidian languages using inter-language distances obtained from this database.

5. Methods

In this section, we describe two distance-based algorithms (UPGMA and NJ), conversion of character data matrices to distance matrices, significance testing of trees, calculation of ASJP inter-language distances, and a rooting strategy for addressing the specific question of ternary vs. binary split of Proto-Dravidian.

UPGMA is a simple hierarchical clustering algorithm which works in the following fashion. In the first iteration, UPGMA combines the least distant language pair, A and B into a language group AB. Then, UPGMA recomputes the distance between AB and any language C by computing the average of the distance between AC and BC and recreates the distance matrix. The algorithm repeats the above steps of combining the two closest languages or language groups and recomputing the distance matrix until the distance matrix is left with a single language group. UPGMA assumes an evolutionary clock model which states that each unit branch length corresponds to a unit time. UPGMA returns a rooted tree due to the assumption of an evolutionary clock.

NJ is a fast, greedy and heuristic tree building algorithm which yields an unrooted tree. NJ builds the tree by beginning with a star-like phylogeny, where each taxon is connected to a single node, and iteratively computes the branch lengths until the phylogeny is resolved. NJ does not assume an evolutionary clock and returns an unrooted tree. The sum of the branch lengths along the path connecting any two languages in a NJ tree indicates the lexical distance between the two languages and does not represent divergence time between the corresponding pair of languages.[8]

The character matrices are converted into distance matrices before they are input to the UPGMA and NJ implementations in the Splitstree package.[9] Three of the four datasets described in section 3.2 are character-based datasets (character matrices). Each of these character matrices is converted into a distance matrix by computing the pair-wise length-normalized Hamming distance between languages represented as character sequences. Hamming distance, between two character sequences, is defined as the total number of positions at which the corresponding characters differ. This pair-wise distance is a length normalized (transformed into a value between 0 and 1) by dividing the distance by the length of a sequence.

Since NJ is a heuristic tree building program, phylogenetic trees obtained through the application of NJ to distance matrices derived from character data matrices have to be tested for statistical significance by

running a bootstrap analysis with a large number of iterations. Note that the original character-based data is required to perform the bootstrap analysis. In each bootstrap iteration, a new dataset of the same size as the original character dataset is created through random selection of data points (characters) from the original data. The same character can be drawn more than once which means that some of the columns of the random data matrix can be duplicated.[10] In each bootstrap iteration, the newly created character dataset is converted to a distance matrix and a new NJ tree is inferred from the distance matrix. This process is repeated for 10, 000 iterations yielding a set of 10, 000 trees.

The next step is to get an estimate of the confidence in the phylogenetic analysis returned by NJ. The confidence score of a node in the phylogenetic tree can be estimated as the count of its occurrence in the set of 10, 000 bootstrap trees. This confidence estimate is also known as the *support value* of that node. A support value greater than 95%, for a node, implies a high statistical significance and that the node was not constructed by chance.

Splitstree also assigns another confidence measure to a tree (both NJ and UPGMA trees) which is the *least squares fit* of the tree. This is computed as the sum of the squares of difference between the true distances (from the distance matrix) and the total branch length (from the inferred tree) for a pair of languages. Unlike the bootstrap analysis which assigns a confidence score to each node in the tree, the least squares fit is a general goodness measure of the constructed tree and measures the amount of tree signal in the data. A least squares fit score of 95% implies that the constructed tree explains 95% of the true inter-language distances given by the distance matrix.

UPGMA and NJ are tree building programs and, therefore, impose a tree structure regardless of the underlying structure of the data. However, it is well-known that evolution of languages need not be tree-like, especially in cases of extensive contact situations. As noted earlier, the Dravidian language family is one such situation. In such a situation, a tree structure is not sufficient to display the relationship between languages and a network can be used to display the relations. There are two kinds of networks: *explicit* and *implicit* (Nichols and Warnow 2008). In explicit networks, the borrowing between related languages can be shown using directed dotted lines from one branch to another branch in a tree. In implicit networks, the branches are not resolved when there is a conflict and the language relations are shown by parallel edges or a web-like structure. The parallel

edges could be collapsed to obtain a tree structure. Alternatively, the parallel edges could be interpreted as an indication of reticulation in the language group. As Nichols and Warnow (2008) note, the interpretation of phylogenetic networks is an open problem. In our experiments, we use the Neighbor Network program available in the Splitstree package to produce implicit networks.

The ASJP word lists contain entries for a subset of the 200-word Swadesh list. The word lists' composition and length is determined empirically, by Brown et al. (2008). The ASJP database and Krishnamurti's (2003) list of the Dravidian language family differ not only in the number of languages but also use different names for the same language. Our dataset consisting of ASJP lists includes only those languages which could be mapped with DEDR or Krishnamurti (2003) based on similarity of name and the proximity of geographical region of the speakers. This criterion yields an ASJP subset of 20 languages from all the four subgroups and allows for a meaningful comparison with the standard tree. We use the LDND (Levenshtein distance corrected for chance similarity) implementation available on the ASJP website[11] for computing the distance matrix suitable for input to a phylogenetic program (specifically MEGA).[12]

Note that while UPGMA returns a rooted tree, NJ, returns an unrooted tree. A NJ tree needs to be rooted for a meaningful comparison with the standard tree. The issue of rooting the unrooted trees inferred by different phylogenetic inference methods has been the subject of much lively debate in recent literature. Kolachina et al. (2011) treat the North Dravidian (ND) clade as outgroup to root the unrooted phylogenetic trees returned by the Maximum parsimony method since both the subgrouping alternatives for Dravidian evaluated in that work agree upon ND being the first outgroup to diverge (cf. figure 4). In our experiments, we follow the same rooting strategy as Kolachina et al. (2011).

Whenever a tree does not group all the ND languages under a single node, that tree is rooted using only those ND languages which are grouped together. We follow this outgrouping strategy consistently in our experiments. One might argue for a simpler solution by adopting Brahui as the outgroup for rooting the trees. It has to be noted that the choice of Brahui as an outgroup is not without problems since Brahui has undergone substantial lexical replacement. It is also unclear whether Brahui was the first language to diverge in the ND subgroup.

Finally, we describe our procedure for evaluating the trees returned through the application of the above methods. We qualitatively evaluate the

NJ and UPGMA tree of each dataset using the *minimum compatibility* criterion of Nichols and Warnow (2008) – the criterion tells that the constructed phylogenetic tree should return all the established subgroups (all the four major subgroups in Dravidian language family) – in the next section. One might argue for a quantitative evaluation of tree quality by application of a tree distance measure such as Robinson-Fould's distance (Felsenstein 2003) for comparing each NJ and UPGMA tree with the standard tree. However, it has to be kept in mind that the trees returned by the tree building methods such as NJ and UPGMA trees are binary trees whereas the standard tree is not only polytomous but also unresolved in the SD I subgroup. In such a scenario, Robinson-Fould's distance is not really helpful in gauging the quality of the inferred trees.

6. Experiments and Results

This section is divided into five subsections. The first subsection gives an analysis of the composition of character-based datasets derived from *DEDR* in terms of number of languages and size of each cognate set (defined below). Each of the remaining four subsections presents and qualitatively evaluates the trees inferred from the application of NJ and UPGMA algorithms to the four datasets.

6.1. Composition of CDR

In this subsection, we analyze the composition of the character-based dataset derived from the CDR.

In the first step, we plot the distribution of cognate size in CDR. The size of a cognate set – henceforth, referred to as cognate set size – is the number of languages attested in that cognate set. We make the following observations about cognate set size:

- The minimum cognate set size is two; the maximum size is twenty- four.
- No cognate set has all the twenty-eight languages from the Dravidian language family.

Figure 5 displays the bar-plot of cognate set sizes for CDR. The figure shows that half of the cognate sets are of size two. The bar-plot suggests that there is an inverse relation between cognate set size and frequency of occurrence, commonly known as Zipf's law. It is not clear if the Zipf's

law-like distribution exhibited between cognate set size and frequency of occurrence is an intrinsic property of cognate set size or the effect of cognate set sampling.

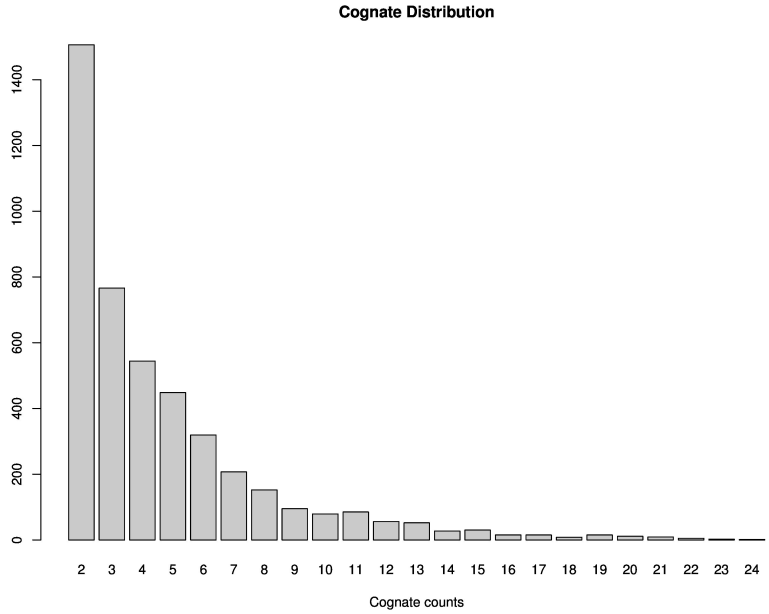

Figure 5. Barplot of the cognate set sizes from DEDR

We now turn to an examination of the distribution of languages in DEDR-based datasets. Figure 6 displays the dot-plot of number of cognate sets for each language. The dot-plot shows a clear division between literary, to the right, and non-literary languages, to the left. One might argue that this distribution is expected due to the vast amounts of information available on literary languages. We successively removed cognate sets of size ranging from 2 to 6 and observed a similar distribution. We further observe that the same distribution holds for the much smaller ADR dataset. The dot-plots suggest that there is a representational bias towards literary (and semi-literary) languages in DEDR. Irula, Kuruba, Kurumba, and Belari have the least cognate set counts.

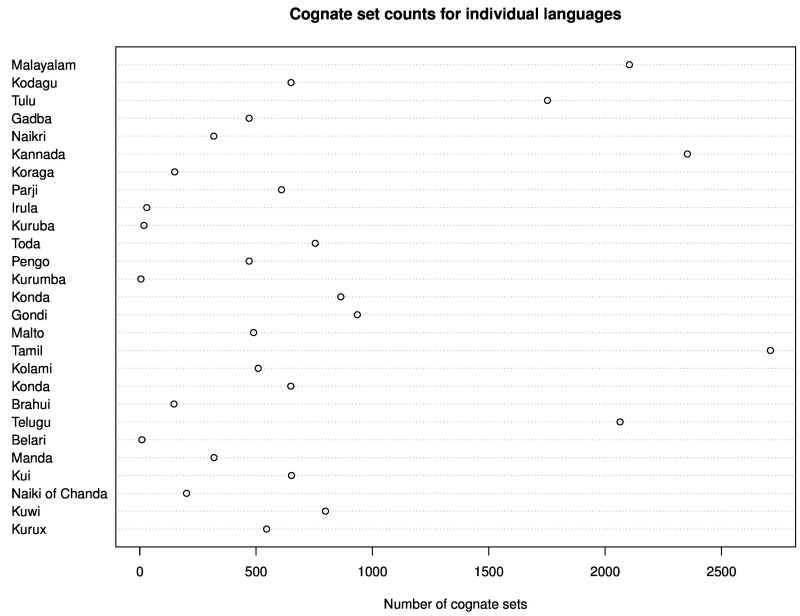

Figure 6. Dotplot of distribution of languages in CDR

Before we proceed to present and describe the trees, we note that each of the trees is a phylogram which provides details not only about the topology but also the branch lengths where branch length represents the amount of linguistic change that took place along the branch.

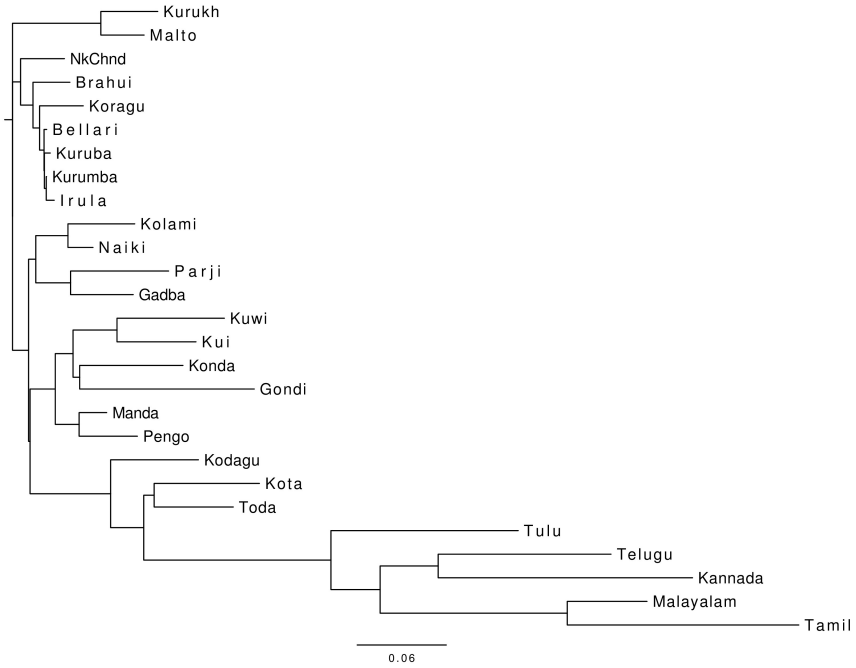

Figure 7. NJ tree rooted using Kurukh-Malto

6.2. CDR

The NJ tree displayed in figure 7 is rooted using Kurukh-Malto as the outgroup. The tree does not return all the major subgroups. The following observations can be made about the internal grouping of the languages.

– The tree returns Tamil-Malayalam.
– The NJ tree groups Toda and Kota together whereas the closeness of Toda and Kota is viewed as suspicious in Krishnamurti (2003), who does not place Toda and Kota together.
– The standard tree groups Kodagu with other Nilgiri languages whereas NJ tree classifies Kodagu closer to Toda and Kota.
– All languages in SD II (except for Telugu) are placed under a single node. The internal classification of SD II is not identical to the standard

tree. Kui-Kuwi, Pengo-Manda are grouped together just as in the standard tree. Konda and Gondi are incorrectly placed together.

– In CD languages: Kolami-Naikri, Parji-Gadba are placed together. Krishnamurti (2003) groups Naikri with Naiki of Chanda whereas Bhattacharya and Burrow considered Naikri to be a dialect of Kolami.
– The remaining languages: Naiki of Chanda (CD), Brahui (ND), Bellari (SD I), Kurumba (SD I), Kuruba (SD I), Koragu (SD I), and Irula (SD I) belonging to different subgroups are placed under a single node. Of these languages, Irula and Kurumba are Nilgiri languages. All these languages are placed close to the root and the support for the branch connecting these languages to the root is not statistically significant (58.1; figure 17).
– The bootstrapped NJ tree (figure 17) annotated with support value is given in the appendix. The support value suggests that the branch connecting Kurukh-Malto to rest of the tree is well supported. The internal branches connecting the CD languages and SD II languages to the rest of the tree have support values of 34.6 and 71 respectively which are statistically non-significant.
– Finally, the NJ tree shows a ternary branching at the root.

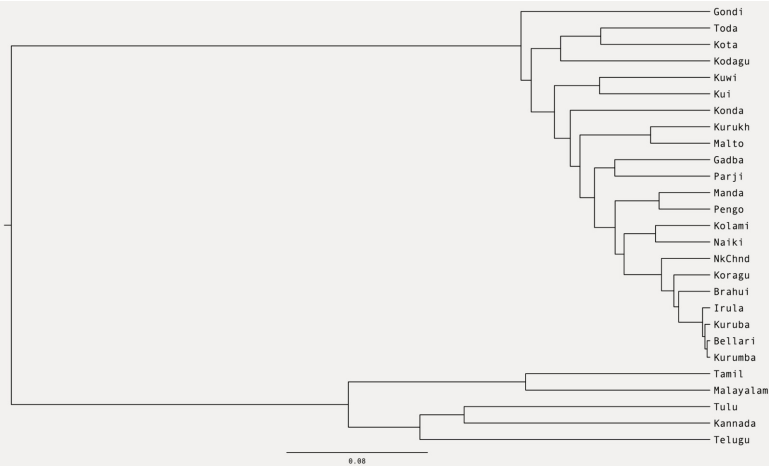

Figure 8. UPGMA tree

The UPGMA tree (figure 8) does not return any of the major subgroups. The tree mixes languages from the established different subgroups. The tree

returns the following language pairs correctly: Toda-Kota, Kui-Kuwi, Kurukh-Malto, Gadba-Parji, Pengo-Manda, Kolami-Naiki, Tamil-Malayalam. Konda and Gondi are neither grouped together nor do they share an immediate common ancestor with the remaining SD II group's languages. The tree groups Kannada with Tulu incorrectly. The tree shows a binary branching at the root.

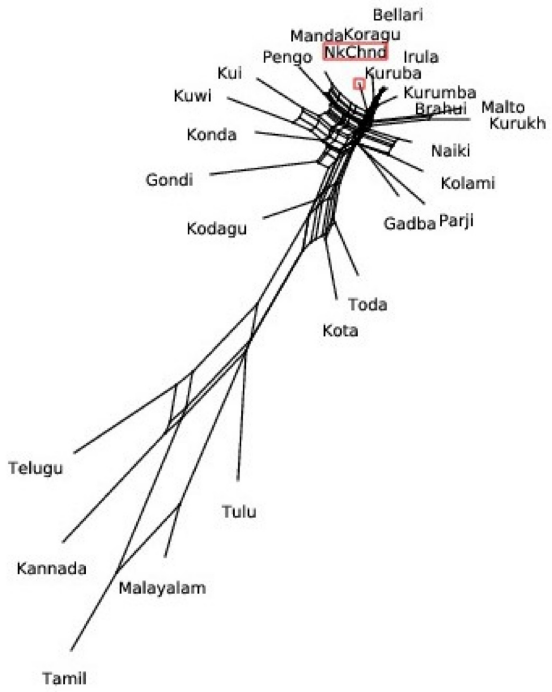

Figure 9. Network derived from CDR dataset

Since the interpretation of a network is an open question (Nichols and Warnow 2008), we describe the network in figure 9 conservatively. We observe that the literary and non-literary languages are separated by a long parallel edge. The network returns the Tamil–Malayalam language pair correctly and places Tulu as the most distant group of the literary languages.

Among the non-literary languages:

– Among SD I languages, Toda and Kota are grouped together. The remaining Nilgiri languages (Irula, Kuruba and Kurumba) are grouped together with the non-literary SD I languages. These Nilgiri languages show a highly undecipherable reticulation.
– SD II languages (except Telugu) are grouped to the left hand side of the structure.
– Among CD languages, Naiki-Kolami, Gadba, and Parji are grouped together. These CD languages are placed next to ND languages.
– All the ND languages are placed together on the right hand side of the network.

6.3. ADR

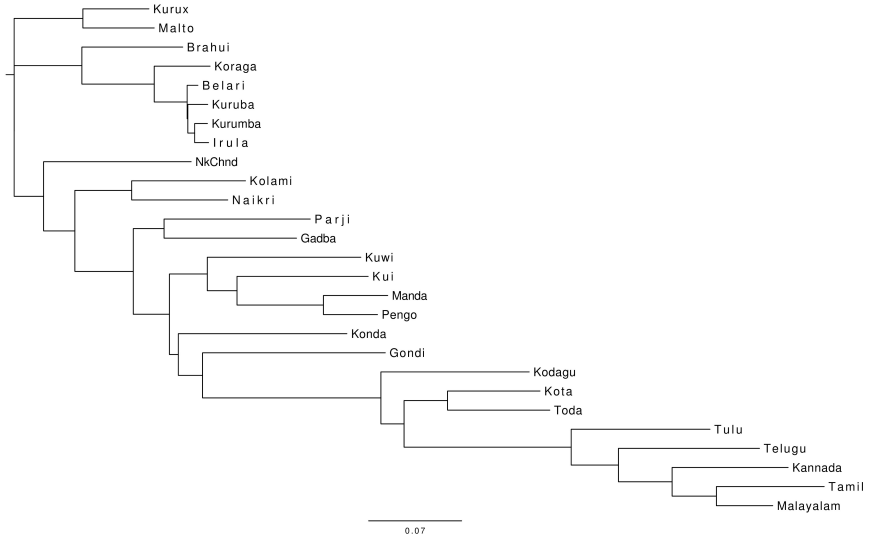

Figure 10. NJ tree rooted using Kurux-Malto

The NJ tree given in figure 10 is different from the NJ tree from the CDR dataset (figure 7). The tree does not return any of the major subgroups given in the standard tree. The tree differs from figure 7 in the following aspects:

– For the first time, Kannada and Telugu are not placed together. Rather, Telugu is placed apart from Tulu-Tamil-Malayalam-Kannada.
– Kui-Kuwi, Gondi-Konda do not occur together.
– Naiki of Chanda is placed closer to the remaining CD languages (Naikri-Kolami, Parji-Gadba).
– The bootstrapped NJ tree displayed in figure 18, in the appendix, suggests that the internal branches connecting the different language groups are not well supported.

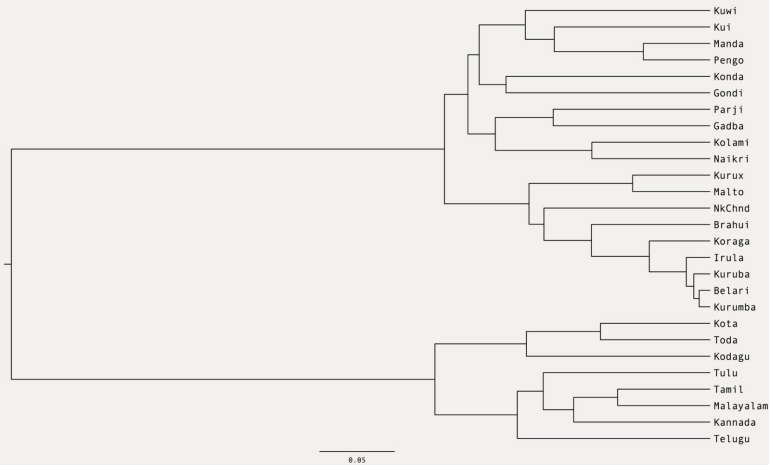

Figure 11. UPGMA tree

The UPGMA tree (figure 11) does not return all the four established subgroups. Surprisingly, the tree makes fewer mistakes than its NJ counterpart (figure 10) and is closer to the standard tree in internal classification at lower level subgroups. Comparing with the standard tree:

– Three Nilgiri languages – Kota, Toda, Kodagu – are classified under a single node.
– The SD II languages (except Telugu) are grouped together. Manda-Pengo, Konda-Gondi are placed together. Kui and Kuwi are shown to diverge separately from a common node.
– In Central Dravidian languages: Naikri and Naiki of Chanda are not grouped together.

– The ND languages Kurukh-Malto are grouped together.
– As usual, languages from different subgroups, Brahui, Naiki of Chanda, Koraga, Irula, Kuruba, Belari, and Kurumba are grouped together.
– The UPGMA tree shows a binary branching.

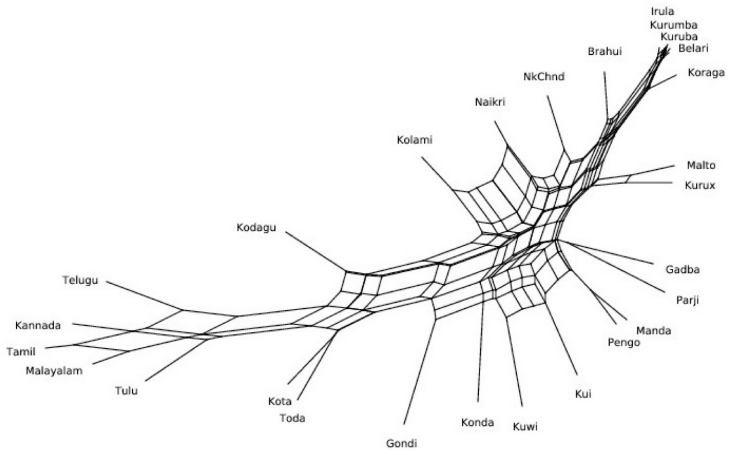

Figure 12. Neighbor Network

The network (figure 12) for the ADR dataset is visibly different from the network derived from CDR dataset (figure 11). There is a clear distinction between literary languages and non-literary languages. The SD II languages, except Telugu, are placed together at the bottom of the network. Kurux and Malto are placed together. The CD languages Gadba-Parji, Naikri-Kolami occur together. The substructure in the far right of the network grouping Belari, Kuruba, Kurumba, Irula, Koraga, and Brahui is highly reticulated. Brahui and Koraga clearly diverge whereas the structure of the remaining four languages is highly unresolved. The network places Naikri-Kolami and Naiki of Chanda next to each other.

6.4. Comparative features

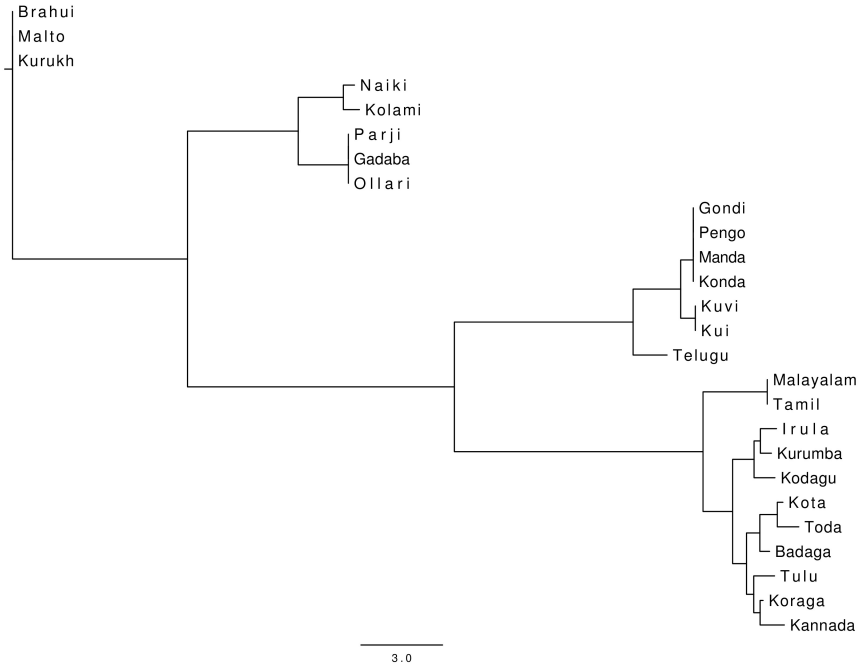

Figure 13. NJ tree using ND as outgroup

The NJ tree is displayed in figure 13. The tree is rooted using the ND clade as the outgroup. The tree returns all the four subgroups intact. The tree returns the following language groups.

– Among the SD I languages, Tamil-Malayalam, Toda-Kota are placed together. Koraga-Kannada and Tulu are placed together under a single node. The standard tree lists Badaga and Kannada as related whereas the NJ tree places them in different subgroups. Among the Nilgiri languages, Toda-Kota and Irula-Kurumba are grouped together. Kodagu occurs with the Irula-Kurumba language group.
– Telugu is the earliest diverging language in the SD II subgroup. Kui-Kuvi are grouped together. The node depicting Gondi, Pengo, Manda, and Konda is polytomous.
– The dataset treats Naikri and Naiki of Chanda as a single language. All the CD languages are grouped together. The tree classifies Naikri-Kolami

under a single node. The Parji-Gadba-Ollari language group's ancestral node is polytomous.

– The tree shows a clear binary split at the root with ND and CD groups placed under one branch and SDI and SDII placed under the other branch.

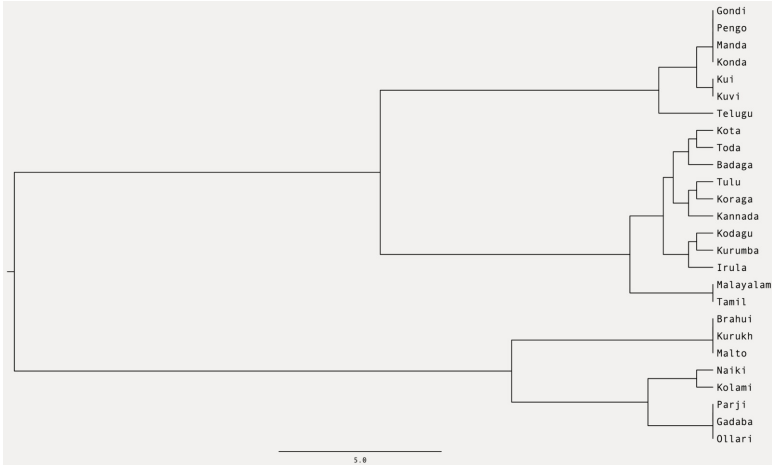

Figure 14. UPGMA tree

The UPGMA tree (figure 14) returns all the four subgroups. The UPGMA tree is topologically similar to the NJ tree (figure 13). Thus, we do not describe the internal classification of each subgroup. The tree shows Proto-Dravidian splitting into ND-CD and SD I-SD II. It is interesting to note that this branching structure does not occur as an alternative in figure 4.

6.5. ASJP

In this subsection, we describe the NJ and UPGMA trees inferred from the ASJP distance matrix of 20 Dravidian languages.

The NJ tree, displayed in figure 15, is rooted using Kurukh as outgroup. The NJ tree is unresolved and returns the following language groups.

– Telugu is placed outside the SD II subgroup. The dialects of Gondi (Gondi and Southern Gondi) are placed next to each other and sharing an immediate common ancestor with Konda_1 (Konda).
– The tree groups the CD languages – Parji, Gadba Pottangi Ollari and Northwestern Kolami – with Kota (belonging to the SD I subgroup).
– All the SD I languages are placed under a single node. The tree returns the Badaga-Kannada and Tamil-Malayalam language pairs correctly.
– Brahui (a ND language) and Telugu are placed next to each other and are shown to diverge at the outset of the tree. Unlike the other datasets, ADR, CDR or Comparative features, the technique of bootstrapping is not applicable to the ASJP dataset. Hence, the support for the branch joining Brahui-Telugu to the root cannot be determined conclusively.
– The tree shows a ternary branching at the root.

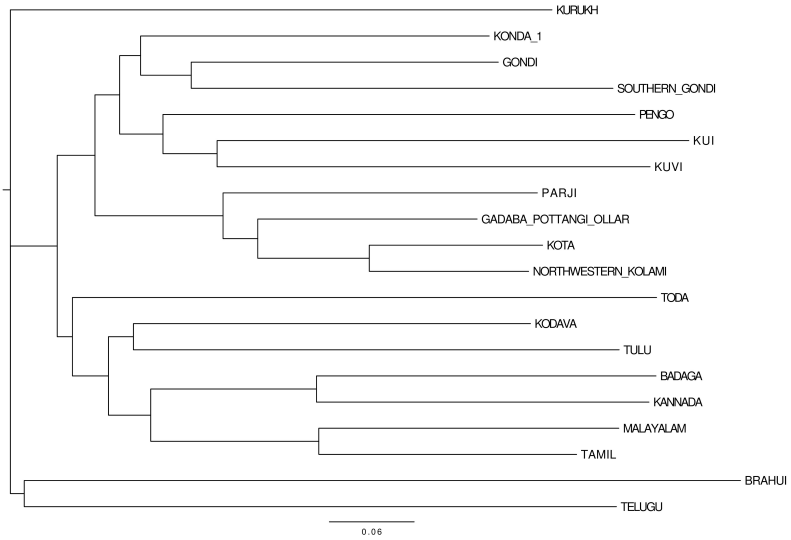

Figure 15. NJ tree rooted using Kurukh as the outgroup

The UPGMA tree inferred from ASJP data is displayed in figure 16. The tree does not return any of the four major subgroups. The tree returns the following language groups correctly:

– SD I: Kannada-Badaga, Tamil-Malayalam.

– SD II: Kui-Kuvi, Konda_1-Gondi.
– The root of the UPGMA tree shows a binary branching.

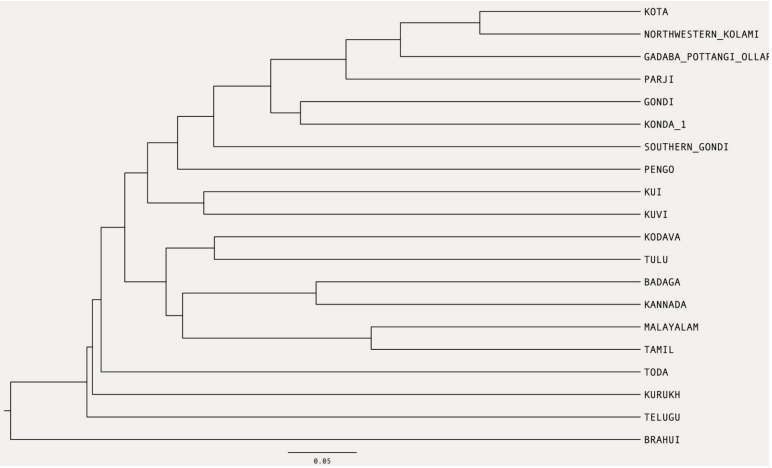

Figure 16. UPGMA tree

Table 1 presents the least squares fit (LSF) value for each of the trees described above. The least squares fit is highly significant for the NJ tree in each dataset. The NJ tree beats UPGMA by a large margin in all the datasets except ASJP.

Table 1. Least squares fit for NJ and UPGMA trees for four different datasets

Dataset	NJ	UPGMA
CDR	99.372	65.663
ADR	98.218	68.222
Comp. feat.	99.857	76.818
ASJP	96.857	99.372

7. Conclusion

We have pointed to the relevance of creating new datasets for subgrouping Dravidian languages for the purpose of throwing new light on the prehistory of the Indian subcontinent. We summarized three extant classifications of the Dravidian language family. We created two new diachronic datasets from *DEDR* which can not only be character encoded but also be used for computing lexical and semantic distances among Dravidian languages. The non-literary languages are underrepresented in both the datasets. We summarized two other datasets based on comparative features and Swadesh lists collected from different sources. In this work, we applied two distance methods NJ and UPGMA, and a network method for subgrouping Dravidian languages. The quality of the resolution of subgrouping for each dataset is summarized in table 2.

Table 2. Summary of subgrouping from different datasets

Dataset	Subgrouping resolution
CDR	No
ADR	No
Comp. Feat.	Yes
ASJP	No

The trees inferred using these datasets are unreliable. There is a little resemblance to the standard tree (Krishnamurti 2003). The NJ tree from the ASJP lists gets almost all the subgroups right with the exception of Telugu and North-Dravidian. Although the UPGMA tree has a higher LSF than the NJ tree on the ASJP list, it is much less resolved than the NJ tree. It is unclear why the trees are different for the CDR and ADR datasets when both datasets are derived from DEDR. The language group consisting of Naiki of Chanda, Brahui, Koragu, Belari, Kuruba, Kurumba, and Irula recurs across all the trees based on CDR and ADR. One possible explanation is the under-representation of these languages in CDR and ADR. The support for binary branching at highest level comes from the results on the Comparative features dataset (both NJ and UPGMA trees).

The unreliability of the trees based on NJ and UPGMA points to the need for the application of character-based methods such as Bayesian

Inference (Huelsenbeck and Ronquist 2001) and Maximum Parsimony (Felsenstein 2003) to the CDR and ADR datasets for subgrouping the Dravidian languages. There is a need for quantitative work in determining the direction of family-internal borrowing. We conclude that the subgrouping of Dravidian languages is an open problem which requires future work.

Appendix

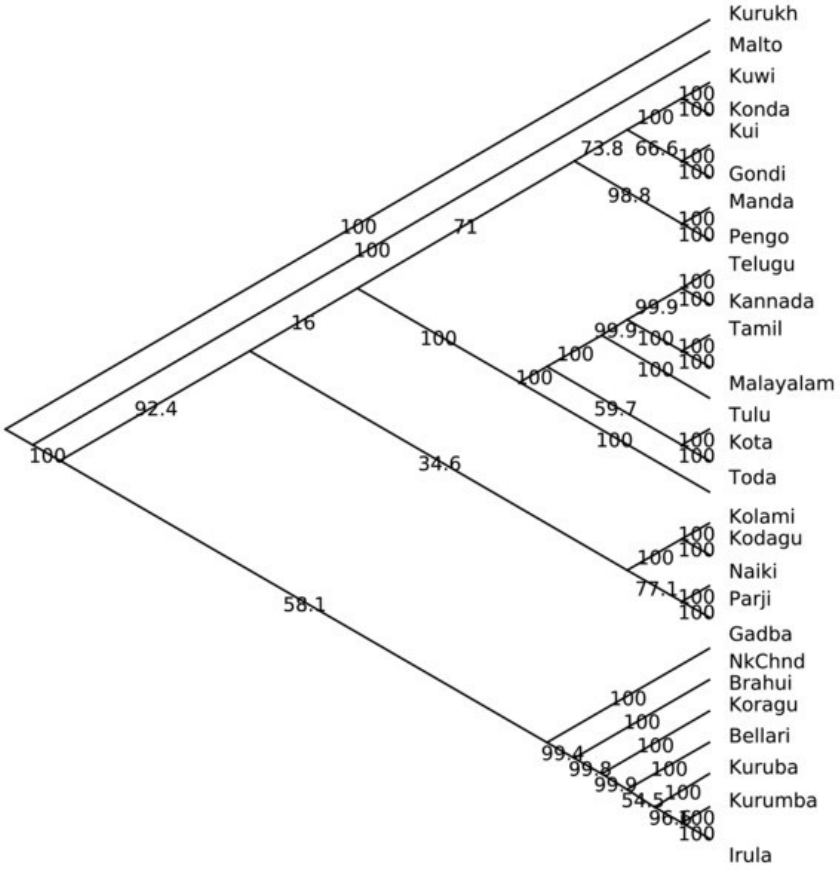

Figure 17. Bootstrapped NJ tree of CDR

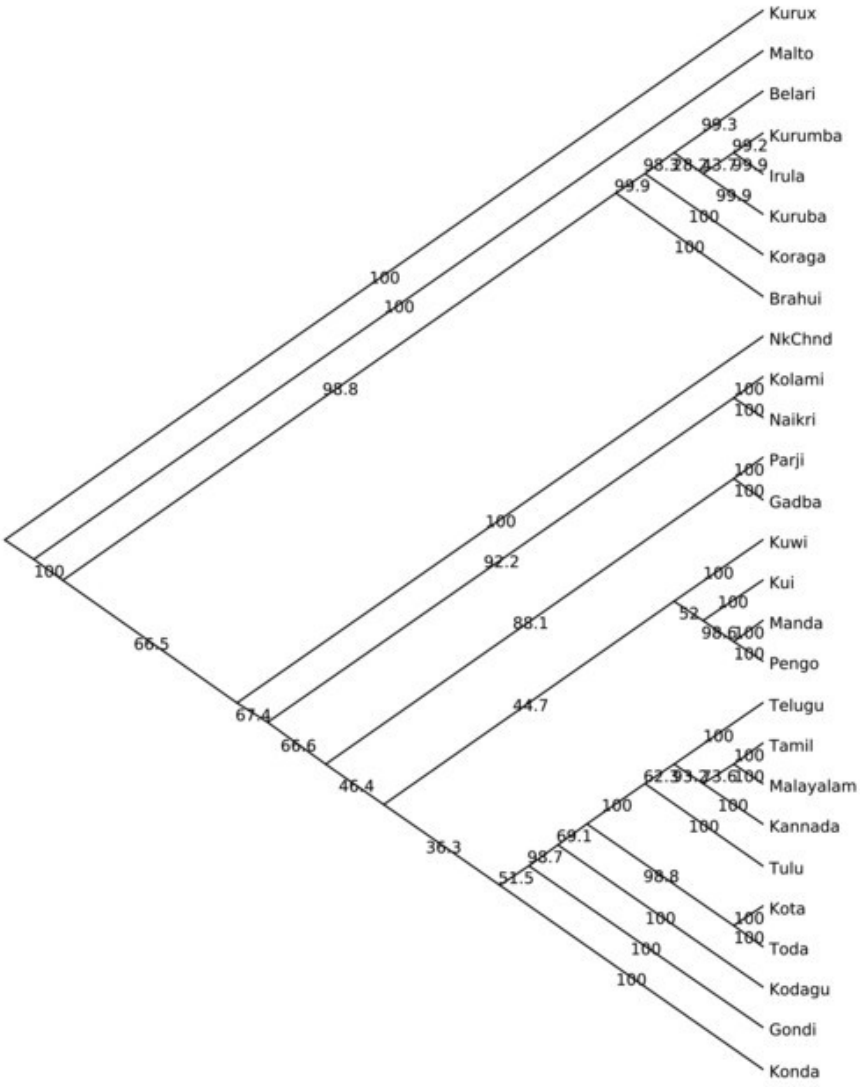

Figure 18. Bootstrapped NJ tree of ADR

Notes

1. We are grateful to Anju Saxena and Lars Borin for the useful comments on the earlier draft of the paper. We thank the reviewer for the highly useful comments in preparing the revised version of the paper.
2. Levenshtein distance is defined as the minimum number of basic edit operations required to convert a string of characters to another string.
3. An electronic version of the dictionary is available online at http://dsal.uchicago.edu/dictionaries/burrow/
4. Paul Huff's program on the ASJP website was used to generate the Ethnologue tree.
5. Konda, Gondi, Kui, Kuvi, Pengo and Manda
6. Dravidian Etymological Dictionary (An earlier version of DEDR).
7. DEDR includes Belari (its name variants, Bellari, Bellary, Belary), but Belari does not appear in Krishnamurti's (2003) classification.
8. Huff and Lonsdale (2011) provide an excellent step-by-step explanation to both UPGMA and NJ algorithms in the context of inferring linguistic phylogeny.
9. Downloadable at http://www.splitstree.org/
10. The tree constructed from a new random matrix might be different from the tree constructed using the original data matrix.
11. http://wwwstaff.eva.mpg.de/~wichmann/ASJP_Distances.zip
12. We use MEGA for the sake of replicability.

References

Andronov, M. S.
1964 Lexicostatistic analysis of the chronology of disintegration of Proto-Dravidian. *Indo-Iranian Journal* 7 (2): 170–186.

Atkinson, Quentin D. and Russell D. Gray.
2005 Curious parallels and curious connections: Phylogenetic thinking in biology and historical linguistics. *Systematic Biology* 54 (4): 513–526.

Brown, Cecil H., 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 – Language Typology and Universals* 61 (4): 285–308.

Burrow, Thomas and Murray B. Emeneau
1984 *A Dravidian Etymological Dictionary [DEDR]*. Second Edition. Oxford: Clarendon Press.
Campbell, Lyle
2004 *Historical Linguistics: An Introduction.* Second Edition. MIT Press.
Dyen, Isidore, Joseph B. Kruskal, and Paul Black
1992 An Indo-European classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82: 1–132.
Elfenbein, J.
1987 A periplus of the 'Brahui problem'. *Indo-Iranica* 16. 215–233.
Felsenstein, Joseph
2003 *Inferring Phylogenies.* Sunderland, MA: Sinauer Associates.
2004 PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genetics, University of Washington, Seattle.
Gray, Russell D. and Quentin D. Atkinson
2003 Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426: 435–439.
Gray, Russell D., Andrew J. Drummond, and Simon J. Greenhill
2009 Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323: 479–483.
Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie
2011 Wals online, Munich: Max Planck Digital Library.
Holman, Eric W., 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.
Huelsenbeck, John P. and Fredrik Ronquist
2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17 (8): 754–755.
Huff, Paul and Deryle Lonsdale
2011 Positing language relationships using ALINE. *Language Dynamics and Change* 1 (1): 128–162.
Huson, Daniel H. and David Bryant
2006 Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23 (2): 254.
Kameswari, T. M.
1969 The chronology of Dravidian languages – a lexico-statistic analysis. In Agesthialingom and Kumaraswami Raja (eds.), 269–274.
Kolachina, Sudheer, Taraka Rama, and B. Lakshmi Bai
2011 Maximum parsimony method in the subgrouping of Dravidian languages. In *Quantitative Investigations in Theoretical Linguistics,*

Amir Zeldes, and Anke Lüdeling (eds.), 52–56. Berlin: Humboldt-Universität.

Krishnamurti, Bhadriraju

1978 Areal and lexical diffusion of sound change: Evidence from Dravidian. *Language* 54 (1): 1–20.

2003 *The Dravidian Languages.* Cambridge: Cambridge University Press.

Krishnamurti, Bhadriraju, Lincoln Moses, and Douglas G. Danforth

1983 Unchanged cognates as a criterion in linguistic subgrouping. *Language* 59 (4): 541–568.

Lewis, M. Paul (ed.)

2009 *Ethnologue: Languages of the World.* 16th edition. Dallas: SIL International. Online version: http://www.ethnologue.com.

Maddison, Wayne P.

1993 Missing data versus missing characters in phylogenetic analysis. *Systematic Biology* 42 (4): 576–581.

McMahon, April and Robert McMahon

2005 *Language Classification by Numbers*. Oxford: Oxford University Press.

2007 Language families and quantitative methods in South Asia and elsewhere. In *The Evolution and History of Human Populations in South Asia,* Michael D. Petraglia*,* Bridget Allchin (eds.), 363–384. Netherlands: Springer Press.

Nakhleh, Luay, Donald A. Ringe, Jr., and Tandy Warnow

2005 Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81 (2): 382–420.

Namboodiri, E. V. N.

1976 *Glottochronology (as applied to four Dravidian languages).* Trivandrum: Sangma.

Nichols, Johanna and Tandy Warnow

2008 Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2 (5): 760–820.

Rama, Taraka, Sudheer Kolachina, and B. Lakshmi Bai

2009 Quantitative methods for phylogenetic inference in historical linguistics: An experimental case study of South Central Dravidian. *Indian Linguistics* 70.

Ringe, Don, Tandy Warnow and Ann Taylor

2002 Indo-European and computational cladistics. *Transactions of the Philological Society* 100 (1): 59–129.

Serva, Maurizio and Filippo Petroni

2008 Indo-European languages tree by Levenshtein distance. *Europhysics Letters* 81 (6): 68005.

Swadesh, Morris

1952 Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96 (4): 452–463.

1955 Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21 (2): 121–137.

Tamura, Koichiro, Joel Dudley, Masatoshi Nei, and Sudhir Kumar

2007 MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0 *Molecular Biology and Evolution* 24 (8): 1596–1599.

Wichmann, Søren

2010a Internal language classification. In *The Continuum Companion to Historical Linguistics*, Luraghi, Silvia, and Vit Bubenik (eds.), 70–86. London/New York: Continuum Books.

Wichmann, Søren, Eric W. Holman, Dik Bakker, and Cecil H. Brown

2010b Evaluating linguistic distance measures. *Physica A* 389: 3632–3639.