

Sudheer Kolachina, Taraka Rama, Lakshmi Bai B.  
LTRC, IIITHyderabad and Språkbanken, Department of Swedish Language, University  
of Gothenburg  
[sudheer.kpg08@research.iiit.ac.in](mailto:sudheer.kpg08@research.iiit.ac.in), [taraka.rama.kasicheyanaula@svenska.gu.se](mailto:taraka.rama.kasicheyanaula@svenska.gu.se),  
[lakshmi@iiit.ac.in](mailto:lakshmi@iiit.ac.in)

## **Maximum parsimony method in the subgrouping of Dravidian languages**

### **Introduction**

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 2003).

In recent years, there has been a rapid increase in interest in the application of phylogenetic inference methods to diachronic language data leading to the emergence of Computational historical linguistics as a distinct field within historical linguistics. The basic intuition underlying such research is that these methods which can infer phylogeny from gene sequences can do so from language data too which also consist of sequences (Atkinson & Gray 2004).

### **Subgrouping of Dravidian languages**

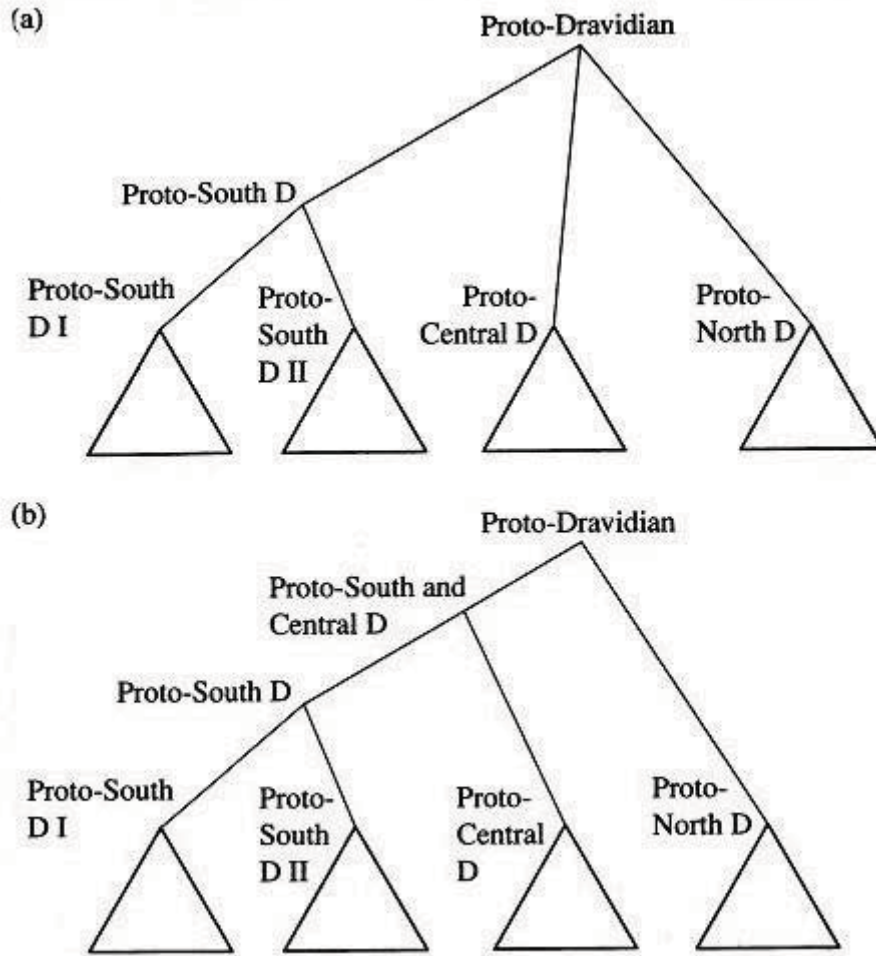
The Dravidian language family consists of 26<sup>1</sup> languages spoken by over 200 million people in South Asia making it the world's fifth largest language family (Krishnamurti 2003). Most of these languages are geographically located in the southern and the central parts of the India with a few scattered pockets in Northern India (Kurux, Malto) and Nepal (Kurux) and a lone population in Pakistan (Brahui).

Krishnamurti (2003) is a compendious work which extensively covers various aspects of Dravidian languages. In this work, two alternative subgroupings of the Dravidian languages are proposed. These alternatives numbered (a) and (b) are shown in Figure 1. The subgrouping adopted in Krishnamurti (2003) is the alternative (a). As can be seen from Figure 1 (a), this subgrouping alternative is a ternary branching structure. In other words, 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 subgrouping is established on the basis of isogloss maps constructed using 27 features from comparative phonology and morphosyntax. It is possible to conceive of a binary division of Proto-Dravidian (shown as alternative (b) in Figure 1) into Proto-North Dravidian (ND) and Proto-South and Central Dravidian (SCD). 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 lean evidence to set up a common stage of South and Central Dravidian.

---

<sup>1</sup> 27 if Naikri is treated as distinct from Naiki.

FIGURE 1. DRAVIDIAN SUBGROUPING ALTERNATIVES.



The aim of this present work is to address this specific question of ternary versus binary branching of Proto-Dravidian through the application of the maximum parsimony method for phylogenetic inference to the comparative feature dataset used by Krishnamurti (2003) in the subgrouping. Evidence from this feature data in support of the binary division is claimed to be lean but that is only when the traditional method of subgrouping based on isogloss maps is followed. It is possible to hypothesize that the application of a different subgrouping method (which relies on a different kind of ‘evidence’) to the same feature data can result in the setting up of a Proto-South and Central Dravidian stage.

### Maximum Parsimony method for phylogenetic inference

The Maximum Parsimony (MP) method is a well-known discrete character-based method which takes as input character sequences. MP is an optimization problem which seeks a tree on which a minimum number of character state changes occurs (Nakhleh et al. 2005b). MP is an NP-hard problem and therefore, exact solutions cannot be guaranteed within polynomial time. As such, heuristics need to be applied to find good (though not provably optimal) solutions. There can be many equally good solutions and a single solution is obtained by applying a consensus method to the Kbest output. Maximum parsimony methods have been claimed to be the most efficient for inferring

the phylogenetic tree that is closest to the traditional standard tree (Ringe et al. 2006; Nakhleh et al. 2005b; Barbancon et al. 2007).

Since the aim of our study is to address the specific issue of binary versus ternary branching of Proto-Dravidian, a method to be qualified for use in our study must be one that at least searches over the sets of binary and ternary branching trees to find the most parsimonious tree. After going through implementations of maximum parsimony available in the PAUP\* (Swofford 2002) and Phylip (Felsenstein 2003) packages, we found that only the *pars* program in the Phylip package meets the requirements of our study as it searches over a tree space consisting of both bifurcating and multifurcating trees.

## Experimental Results and Discussion

Krishnamurti (2003) contains four sets of features from comparative phonology, morphology and syntax that support the subgrouping adopted in that work (Figure 2(a)). We use the same datasets in our experiments<sup>2</sup>. In order to apply any phylogenetic inference method to these datasets, the feature data are encoded as character data (summarized in Table 1 below). Following Maddison (1993), we make a distinction between missing characters (feature not relevant) as opposed to missing data and code them using two distinct characters.

TABLE 1. FEATURES USED IN OUR EXPERIMENTS.

Feature type	Phonology	Nominal morphology	Verbal morphology	Syntax
# features	13	9	13	5

In order to guard against statistical bias, bootstrapping procedure was run for 10000 times with ‘sampling with replacement’ using the *seqboot* program in PHYLIP. The *pars* program was applied to these multiple datasets to find the most parsimonious trees from each set. The consensus tree was estimated from all these parsimonious trees using majority consensus (*consense*). The consensus tree obtained is the single most parsimonious tree containing edges annotated with their support values. Next, branch lengths on the consensus tree were re-estimated using the *pars* program. Finally, the unrooted tree returned by *pars* was rooted using the North Dravidian (ND) clade as the outgroup. The rationale behind using ND as the outgroup is that both the subgrouping alternatives agree on ND being the first to diverge from Proto-Dravidian. The phylogenetic tree inferred from the character sequences after rooting is shown below in Figure 2. The branches in the tree are annotated with the branch lengths returned by the *pars* program. The internal nodes are labeled with names of known subgroups they represent.

<sup>2</sup> The datasets are available at the following link: <https://docs.google.com/leaf?id=0B6U29M4CJtXXODJkZWM4MGQtOThiMC00NDZiLTgxZGYtMTRiMGRyMTE0YjVj&hl=en>



- Nakhleh, L., Warnow, T., Ringe, D. and Evans, S.N. 2005b. A Comparison of Phylogenetic reconstruction Methods on an Indo-European Dataset. *Transactions of the Philological Society* 3(2), 171-192.
- Nichols, J. and Warnow, T. 2008. Tutorial on computational linguistic phylogeny. *Linguistics and Language Compass* 2(5), 760-820.
- Ringe, D., Warnow, T. and Evans, S. 2006. Polymorphic characters in Indo-European. In *Indo-European Languages, Languages and Genes*, September 2006.
- Swofford, D.L. 2002. PAUP\*. *Phylogenetic analysis using parsimony (and other methods)*. Version 4. Sunderland, MA: Sinauer Associates.
-