Maximum parsimony method in the subgrouping of Dravidian languages

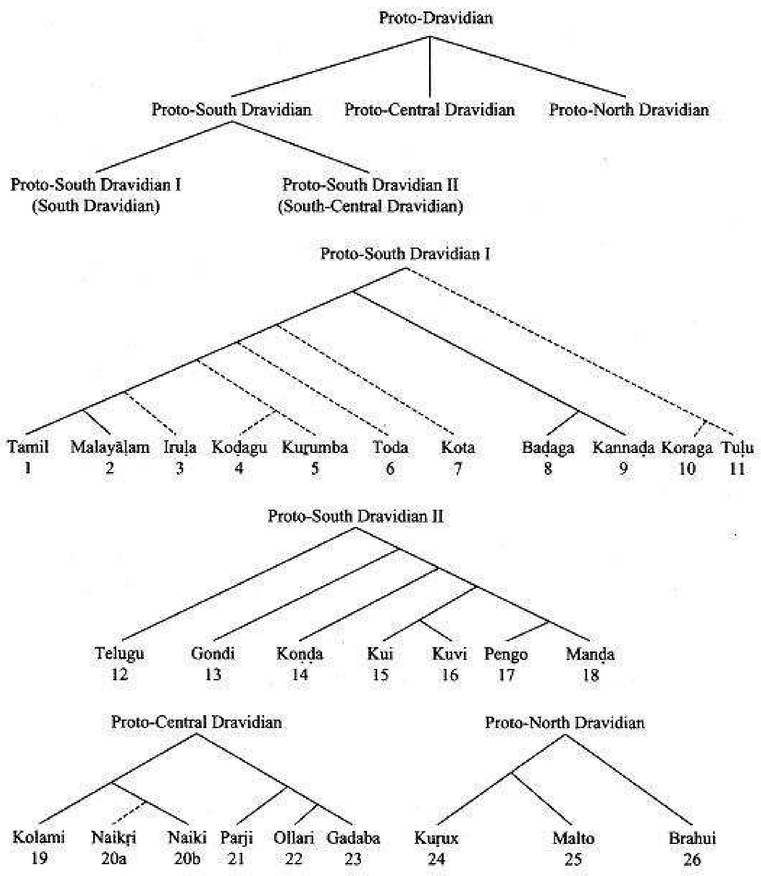
Sudheer Kolachina[1], *Taraka Rama*[2] *and Lakshmi Bai*[1]

[1] **Language Technologies Research Centre, IIIT-Hyderabad**

[2] **Språkbanken, Department of Swedish Language, University of Gothenburg**

Introduction

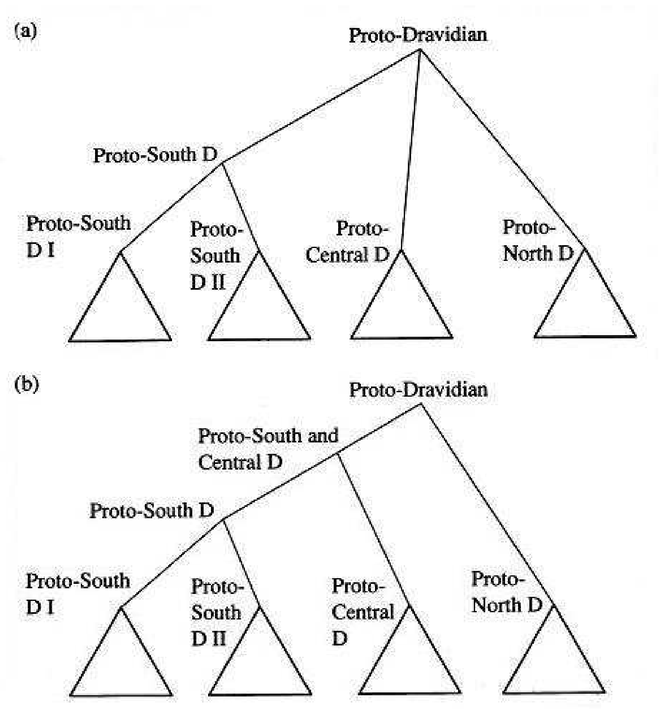
- **Subgrouping**: internal classification of languages within a language family
- **Dravidian language family**
  - consists of 26 languages spoken by over 200 million people in South Asia
  - family tree shown in Figure 1 (taken from (Krishnamurti 2003))



Broken lines reflect uncertainty as to a language's position within the group.

Subgrouping of the Dravidian languages

- Two possible subgroupings of the Dravidian languages according to (Krishnamurti 2003)



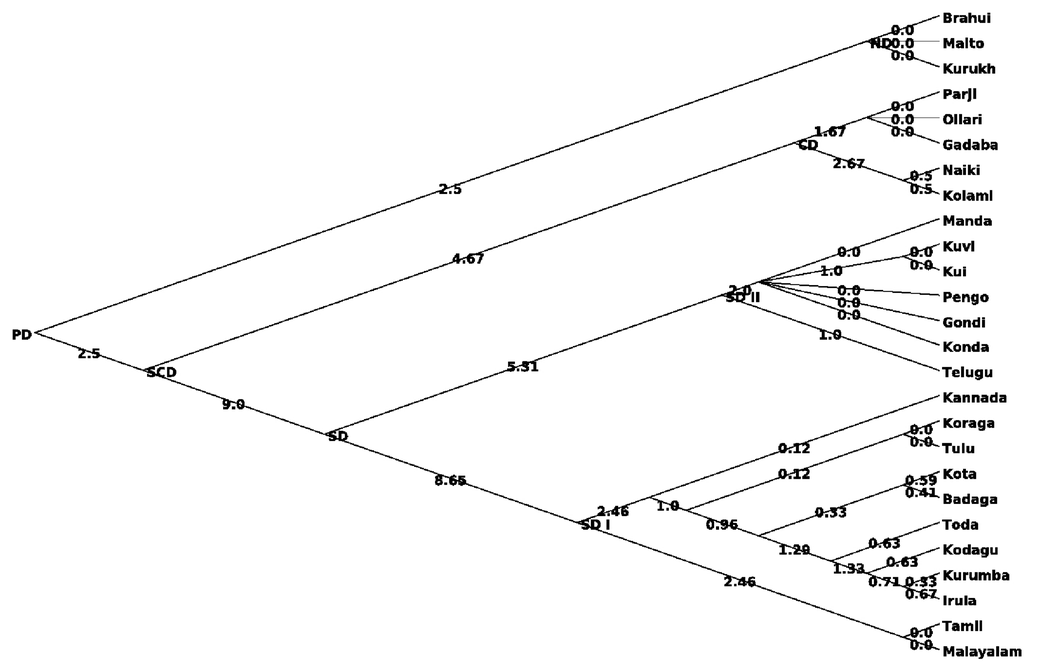
- **Aim**: To address this specific question of ternary versus binary branching of Proto-Dravidian via application of the Maximum Parsimony method (MP) to the Dravidian data
- **Dataset**: Features from comparative phonology, morphology and syntax used for subgrouping (Krishnamurti 2003) (available on request)
- **Intuition**: Binary branching of speech communities more likely than ternary
- **Procedure**: Apply MP to the same dataset and compare inferred tree to the tree constructed using traditional methodology

Maximum Parsimony (MP) method

- MP infers phylogeny by searching for the phylogeny with the minimum number of evolutionary events
- MP shown to be the most efficient for inferring the phylogenetic tree that is closest to the traditional standard tree (Nakhleh et al. 2005)
- Implementation of MP used in our experiments: *pars* program in PHYLIP
- Reason: *pars* searches over the space of both bifurcating and multifurcating trees

Experimental setup

- Bootstrapping procedure run for 10000 times with 'sampling with replacement'
- *pars* applied to the bootstrapped datasets to get multiple parsimonious trees
- Consensus tree *a*) estimated using majority consensus *b*) rooted using the North Dravidian (ND) clade as the outgroup
- *pars* applied to the dataset again, this time giving the rooted consensus tree as additional input
- Branch lengths on the consensus tree re-estimated using *pars*



Results

- Phylogenetic tree inferred using MP shown above
- Notes on interpreting the inferred tree
  - Ternary branching can show up as binary branching with zero branch length
  - A binary branching internal node can be eliminated if the number of state changes (indicated by branch lengths) along its two branches is equal
  - Difference in branch lengths between SCD and SD, and SCD and CD is 4.33 and hence, SCD cannot be eliminated

Conclusions and Future Work

- Main conclusion: MP Tree inferred clearly shows binary branching of Proto-Dravidian and not ternary as suggested in (Krishnamurti 2003)
- Features shared by CD and SD II ignored in the subgrouping using the traditional method (Figure 2(a))
- It treats these similarities between CD and SD II as a result of a common stage in their evolution: Proto South-Central Dravidian (SCD)
- Additional outcomes: MP resolves other uncertainties such as position of Nilgiri languages
- In future,
  - Experiment with weighted Maximum Parsimony (WMP) by weighting different kinds of features
  - Experiment with a much larger set of features by including lexical features
  - Explore network-based methods to address borrowing and homoplasy

References

Krishnamurti, B. (2003), *The Dravidian languages*, Cambridge Univ Press.

Nakhleh, L., Warnow, T., Ringe, D. & Evans, S. (2005), 'A comparison of phylogenetic reconstruction methods on an Indo-European dataset', *Transactions of the Philological Society* **103**(2), 171–192.

Acknowledgements: Special thanks to Prof. Bh. Krishnamurti first, for entertaining us and second, for the extremely valuable discussion. We would also like to thank Dr. Harald Hammarström for his helpful comments and suggestions. The second author is supported by Graduate School in Language Technology, Sweden.