Distance-based algorithms in the sub-grouping of Dravidian language family

Taraka Rama\textsuperscript{1} and Sudheer Kolachina\textsuperscript{2}

\textsuperscript{1}Språkbanken
University of Gothenburg

\textsuperscript{2}LTRC
IIIT – Hyderabad

Workshop on comparing approaches to measuring linguistic differences, Gothenburg
Outline

Introduction

Extant classifications

Related work

Datasets

Results and Discussion

Conclusions and Future work

References
Introduction I

- Distance-based phylogenetic inference algorithms
  - Subgrouping of Dravidian languages
  - Address issue of ternary vs. binary branching at the highest level in the tree
Introduction II

- Compare subgrouping returned by distance-based algorithms across four datasets
  1. DEDR-based
  2. Lexical Reconstructions (Krishnamurti 2003)
  3. Comparative features (Krishnamurti 2003)
  4. **Automated Similarity Judgment Program** (ASJP; Wichmann et al. (2010))
Traditional subgrouping begins with compilation of cognate sets for a set of related languages.

Traditional lexicostatistics uses Swadesh word lists in a data-poor scenario (Wichmann 2010).

DEDR allows us to go beyond Swadesh lists for Dravidian languages.

Distance-based algorithms for data-driven inference of linguistic phylogeny.
Introduction IV

- Tree algorithms:
  - Neighbor Joining
    - Unrooted tree rooted using Mid-point rooting algorithm
  - UPGMA
- Neighbor Network (Huson & Bryant 2006)
Contributions of this work:

- Results of subgrouping Dravidian languages applying distance-based methods to different datasets
- Answer to the question of ternary vs. binary branching of Proto-Dravidian
Outline

Introduction

Extant classifications

Related work

Datasets

Results and Discussion

Conclusions and Future work

References
Figure: Dravidian family tree

- Position of the Nilgiri languages (Toda, Kota, Iruḷa, Baḍaga and Kuṟumba) in relation to Tamil and Kannaḍa
- Position of Tūḷu
- Placement of Koraga
- Relation between Toda and Kota
- Central Dravidian: Position of Naikṛi

Figure: South Dravidian I family tree
Issues: WALS (Haspelmath et al. 2008)

- Excludes four languages present in Krishnamurti (2003) - Iruļa, Koraga, Naiki and, Ollari.
- A two-level classification with genus and constituent languages.

Figure: WALS distribution of Dravidian language family.
Issues:
Ethnologue (Lewis 2009)

- Proto-North Dravidian is polytomous (more than two children).
- South Dravidian I subgroup’s internal node is polytomous.

**Figure:** Ethnologue (Lewis 2009) tree
Outline

Introduction

Extant classifications

Related work

Datasets

Results and Discussion

Conclusions and Future work

References
Related work I

- Andronov (1964)
  - Collected 100-word Swadesh lists for nineteen Dravidian languages
  - Applied glottochronological method
  - Reviewed by Krishnamurti (2003)
Related work II

- Krishnamurti (1978)
  - Framework of lexical diffusion
  - Example of gradual sound change: Apical displacement
  - Compiled cognate sets for six South-Central Dravidian (SCD) languages qualified for apical displacement
  - Language ‘proximity’ measured as the number of shared cognates-with-change
  - MDS algorithm
  - Resultant plot ‘in agreement’ with standard tree
Related work III

- Krishnamurti et al. (1983)
  - Sequel to Krishnamurti (1978)
  - Lexical diffusion dataset
  - Identified 63 cognate sets in SCD qualified for apical displacement
  - u–o–c distribution pattern
  - Enumerated all possible trees for the six languages
  - Each tree scored based on the number of changes required to explain each cognate set
  - Tree with the least cumulative score over all cognate sets is the best tree
  - Resultant tree agrees with the standard tree
McMahon & McMahon (2007)

- Prolonged extensive contact in South Asia
- Evolution not necessarily tree-like
- Therefore, network models for linguistic phylogeny
Related work V

- Rama et al. (2009)
  - Apply Maximum Parsimony (MP), Bayesian Analysis and distance-based algorithms to Krishnamurti et al.’s (1983) dataset
  - Noted that Krishnamurti et al.’s (1983) method is a special case of MP
Related work VI

- Kolachina et al. (2011)
  - Krishnamurti (2003) used 27 comparative features for supporting ternary branching over binary branching
  - 1/0/? (presence, absence or missing)
  - Apply MP to address question of ternary branching vs. binary branching
  - Conclusion: Branch lengths returned by MP do not support ternary branching
Outline

Introduction

Extant classifications

Related work

Datasets

Results and Discussion

Conclusions and Future work

References
Complete *DEDR* (CD):

- 6027 cognate sets for 28 languages
- 5548 cognate sets with unique entry number
- A cognate set was removed if:
  - Possible borrowing from Dravidian to Indo-Aryan
  - Doubtful cognacy judgement
  - Cross-referencing with another cognate set
- Final dataset has 4169 cognate sets
- Character-based dataset
Reconstructions *DEDR* (RD):

- Krishnamurti (2003) provides 656 lexical reconstructions along with *DEDR* entry numbers
- Post cleanup – 348 items
- Character-based dataset
- Can be used to evaluate approaches that automate reconstruction
Comparative features:

- Character-based dataset from Kolachina et al. (2011)
- Naikṛi and Naiki of Chanda treated as a single language

ASJP lists:

- Consists of only those languages which could be mapped with DEDR or Krishnamurti (2003)
- 20 languages from all the four major subgroups
Outline

Introduction

Extant classifications

Related work

Datasets

Results and Discussion

Conclusions and Future work

References
Exploring CD and RD I

- Smallest cognate set size is two
- Largest cognate set size is 24
- About half of the cognate sets have a size of two
- Cognate set size is inversely proportional to frequency of occurrence
Five languages are over-represented.

All the five languages are literary (semi-literary: Tuḷu).

Iruḷa, Kuruba, Kuṟumba and Belari are represented the least.

Similar distribution observed for RD.
Exploring CD and RD III

Figure: Cognate set distribution for individual languages in RD
Experiments: CD I

- **Binary branching:**
  - literary and non-literary

- **Literary branch:**
  - Tamil & Malayalam; Telugu & Kannada

- **SDr II, except Telugu**

- **NDr:** Kurukh & Malto; Brahui placed with Nilgiri languages

- **CDr:** Naikri & Kolami

**Figure:** NJ tree
Experiments: CD II

- Toda & Kota; Parji & Gadaba
- Krishnamurti (2003) makes distinction between Baḍaga and Kannada, DEDR lists both as Kannada
- Naikri and Naiki of Chanda are related?
- Koragu (Koraga) & Bellari; Kuruba and two other Nilgiri languages (Iruļa and Kuṟumba)
- Languages from CDr and SDr I mixed
Experiments: CD III

- UPGMA tree similar to NJ tree

**Figure:** UPGMA tree
Figure: Neighbor Network
Experiments: CD V

- Literary & non-literary languages separated by a long parallel edge
- Literary languages: Tamil & Malayalam, Kannada & Telugu; Tulu is the earliest to diverge
- Telugu & Kannada: despite supposed contact due to geographical proximity, no reticulated structure
- NDr on the right side of non-literary languages
- Six SDr II languages at top left
- Toda & Kota as in other trees
- Nilgiri languages show highest reticulation
- CDr: Naikṟi, Kolami, Gadba and Parji grouped together next to NDr languages
Experiments: RD I

- Kodagu, Kota and Toda, added to literary languages cluster
- SDr II languages not a single group

Figure: NJ tree
Experiments: RD II

- Better resolved than NJ tree
- SDr II (except Telugu) grouped together
- Nilgiri languages and NDr languages grouped under a single node
- Telugu earliest to diverge among literary languages

Figure: UPGMA tree
Experiments: RD III

Figure: Neighbor Network
Experiments: RD IV

- Different from network of CD dataset
- Clear gap between literary languages and non-literary languages
- SDr II (except Telugu) placed together at the bottom
- Substructure showing Belari, Kuruba, Kurumba, Iruła, Koraga and, Brahui highly unresolved
- NDr: Kurux & Malto
- CDr: Gadaba & Parji; Naikṛi & Kolami are placed together
- Brahui, Koraga show clear divergence; structure of other four languages unresolved
- Naikṛi & Naiki of Chanda are placed next to each other
Experiments: Comparative features I

- Not quite unexpected
- Binary tree and resolves the four major subgroups
- Common with previous trees: Kota & Toda; Naiki & Kolami; Kui & Kuvi; Malayalam & Tamil
- Internal branch lengths are non-existent in many subgroups

Figure: NJ tree
Experiments: Comparative features II

Figure: UPGMA tree

- UPGMA tree same as NJ tree
Experiments: ASJP I

- SD II languages except Telugu, under a single group
- CDr languages grouped together
- SDr I languages placed under a single node
- Brahui, Kurukh and Telugu diverge at the outset

Figure: NJ tree
Experiments: ASJP II

Figure: UPGMA tree

- None of the major subgroups clearly resolved
Outline

Introduction

Extant classifications

Related work

Datasets

Results and Discussion

Conclusions and Future work

References
Conclusions and Future work I

- New datasets: complete DEDR and Krishnamurti’s reconstructions
- Non-literary languages under-represented in both datasets
- Trees inferred using these datasets alone unreliable
- Little resemblance to the standard tree
- Food for thought: How to use such sparse datasets?
- Interesting direction: Combine these datasets with ASJP lists and QITL dataset which are not so sparse
Conclusions and Future work II

- Support for binary branching at highest level comes only from results on QITL dataset (NJ and UPGMA trees)
- All four subgroups present only in trees from the QITL dataset
- NJ tree from ASJP lists gets almost all subgroups (exceptions: Telugu and North-Dravidian)
- Positions of Telugu, Brahui unresolved in subgroupings from ASJP lists (lexical replacement in Swadesh lists?)
- UPGMA tree much less resolved than NJ tree on ASJP lists
- Interesting direction: Combine ASJP lists and QITL dataset
- Food for thought: How to investigate family-internal borrowing?
Outline

Introduction

Extant classifications

Related work

Datasets

Results and Discussion

Conclusions and Future work

References
References


Lewis, M. P., ed. (2009), *Ethnologue: Languages of the World*, sixteenth edn, SIL International, Dallas, TX, USA.


