

First year as GSLT student: Summary

Taraka Rama

Språkbanken

University of Gothenburg

CLT Seminar

Outline

ASJP: Phoneme diversity, word length and population size

ASJP: Automatic classification

Language distance estimation from parallel corpora

Dravidian languages

Dravidian languages: Top level branching

References

ASJP database I

- ▶ A much larger sample of languages, 3000+ languages
- ▶ Around **half** of the world's languages
- ▶ 109 out of the world's 121 linguistic families
- ▶ 47 out of 123 isolates
- ▶ 40 out of 122 creoles, mixed languages, and pidgins

All the above language classifications are based on *Ethnologue*

- ▶ Word list admitted if and only if it has 70% of the entries

ASJP database II

[Home](#) | [Contact](#) | [Max Planck Institute](#) | [Imprint](#)



The Automated Similarity Judgment Program

[Home](#)

[Languages processed](#)

[Invitation to contribute](#)

[ASJP World Language Tree](#)

[Published papers](#)

[Papers on language families](#)

[Conferences](#)

[Software](#)

[Talks](#)

The Automated Similarity Judgment Program



The ASJP project aims at achieving a computerized lexicostatistical analysis of ideally all the world's languages.

The two main purposes are to provide a classification of all languages by a single, consistent and objective (if perhaps not ideal) method and to perform various statistical analyses regarding the historical and areal behavior of lexical items.

Special new feature: watch the first ASJP movie on Youtube by clicking [here](#).

ichmann/ASJPHomePage.htm#

Figure: ASJP: Webpage

ASJP database III

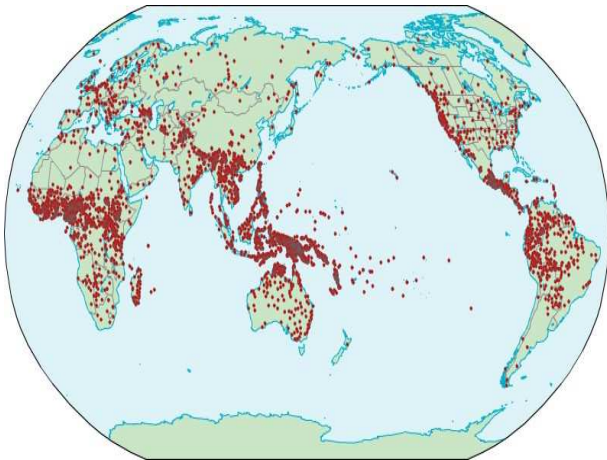


Figure: ASJP: Processed languages

ASJP code I

- ▶ ASJP code is a simple code using QWERTY keyboard
 1. 34 symbols for consonants
 2. 7 symbols for vowels
 3. Two modifiers ~ and \$ for combining the previous segments

ASJP code II

BLOOD, BONE, BREAST, COME, DIE, DOG, DRINK, EAR,
EYE, FIRE, FISH, FULL, HAND, HEAR, HORN, I, KNEE, LEAF,
LIVER, LOUSE, MOUNTAIN, NAME, NEW, NIGHT, NOSE,
ONE, PATH, PERSON, SEE, SKIN, STAR, STONE, SUN,
TONGUE, TOOTH, TREE, TWO, WATER, WE, YOU (SG).

Variables

- ▶ SR – Segments Represented in a word list
- ▶ MWL – Mean Word Length for a word list
- ▶ MMWL – Mean of MWL for a language family
- ▶ logPop – Log of Population Size for a language or family
- ▶ MSR – Mean SR of a family

SR as proxy I

- ▶ Confirm the validity of using segments extracted from the word list (SR)
- ▶ Match the UPSID (Maddieson & Precoda 1990) segment inventory sizes for 392 (out of 451) languages against SR
- ▶ The mean of UPSID/SR is .818 with s.d = .188

SR as proxy II

- ▶ Each UPSID language is matched to ASJP language(s) list based on the following criterion:
 1. Both should pertain to the same geographical dialect
 2. Have similar names
 3. If UPSID covers several word lists in ASJP list, then the ASJP SR is represented by the mean SR of the several ASJP lists

SR as proxy III

- ▶ One might assume that a larger list allows us to represent better all the phonological segments
- ▶ The average length of word list is 35.7 for 3168 languages
- ▶ Very small correlation, $r = .17$ between the number of words attested and SR
- ▶ Very small correlation, $r = -.05$ between word list size and UPSID/SR
- ▶ Further, loanwords are excluded for excluding the rare phonemes

SR as proxy IV

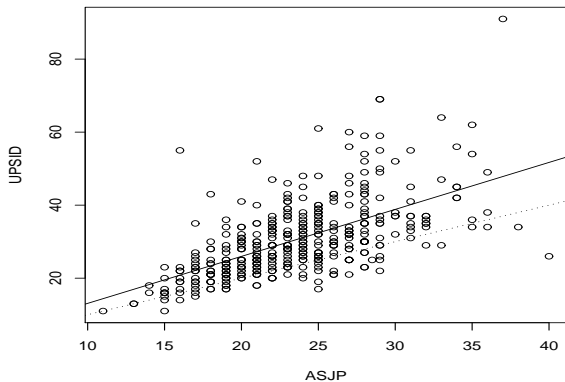


Figure: Pearson's $r = .61$

SR and word length I

- ▶ Word length is inversely correlated with phoneme inventories' size across languages
- ▶ Nettle (1999) used 50 randomly sampled dictionary entries and segment inventory sizes
- ▶ Nettle's (1999) experiment shows a power distribution with high correlation

SR and word length II

What if we use more word lists?

- ▶ Our sample's word lists length ranges from 24 to 40 and has 36.8 words on an average
- ▶ Our word lists are more stable than a randomly sampled dictionary
- ▶ We use MWL extracted from ASJP wordlists vs UPSID segment sizes for 392 languages
- ▶ The correlation is $r = .31$.

SR and word length III

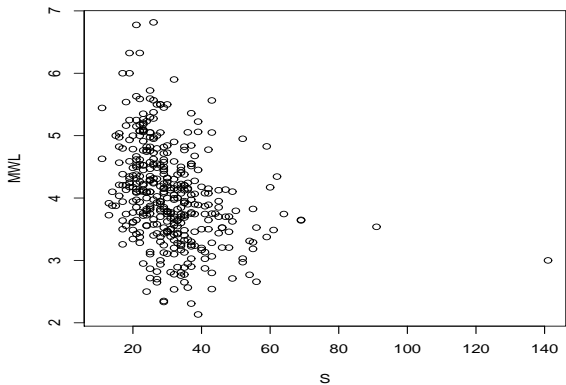


Figure: MWL vs UPSID. Archi (Caucasian) and !XU are the outliers

SR and word length IV

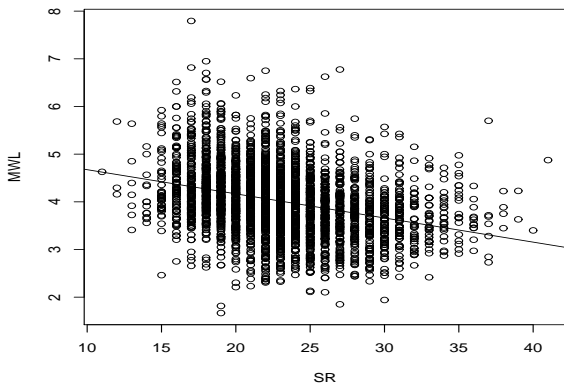


Figure: MWL vs SR for 3168 languages. $r = -.31$

SR and word length V

In contrast with Nettle (1999)

- ▶ Our sample uses wordlists wherever available
- ▶ All living language families
- ▶ Our sample can be tested for significance
 - ▶ Genealogical: Languages in the same family cannot be treated as independent units
 - ▶ Areal: Genetically unrelated languages in contact
- ▶ Histograms for SR and MWL are normal for larger families

SR and word length VI

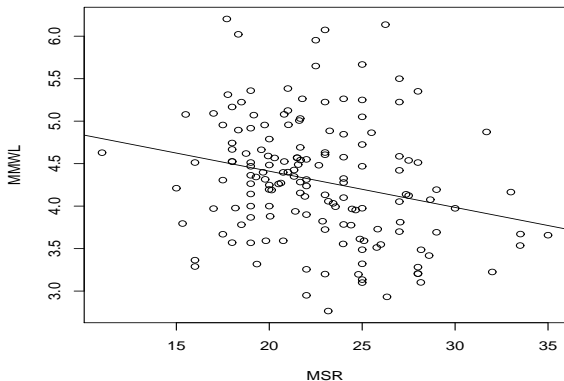


Figure: MMWL vs SR for 157 families. $r = -.23$

SR and word length VII

- ▶ Language contact, migration increases similarity between (genetically) related and unrelated language
- ▶ There is a little variation between major geographic macro-areas
- ▶ Area of majority of the language family is its macro-area

SR and word length VIII

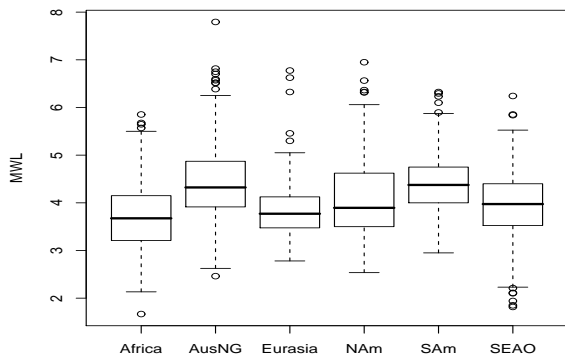


Figure: Box plots of MWL for different families across six macro-areas

SR and word length IX

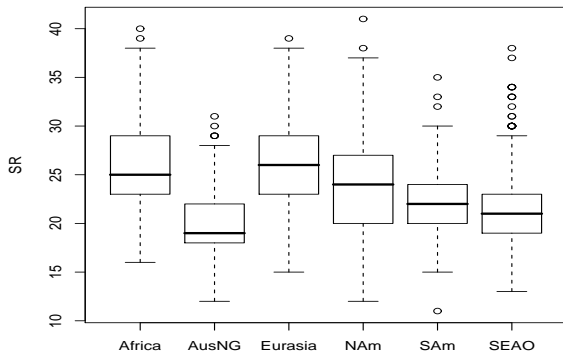


Figure: Box plots of SR for different families across six macro-areas

SR and word length X

- ▶ A linear mixed model is used to assess the p -value of the regression
- ▶ Macro-areas are treated as a random effects
- ▶ The correlation is significant at p -value of 0.009
- ▶ Correlation increases by removing smaller families(> 2)
- ▶ A better correlation for 91 families with $r = -.31$ and $p = 0.0008$
- ▶ Finally, the correlations support Nettle's findings

SR vs population sizes I

- ▶ Population sizes and phoneme inventory sizes are related (Nettle 1999)
- ▶ Smaller population sizes can undergo more innovations
- ▶ Direction of the pull?
- ▶ Hay & Bauer (2007) report a Spearman's $\rho = .37$ between logarithm of population size and phoneme inventory sizes

Cannot to be trusted due to interdependence of data points

SR vs population sizes II

- ▶ Atkinson (2011) finds a negative correlation between phoneme inventory sizes and the distance from Africa
- ▶ Atkinson (2011) uses WALS' phoneme inventory sizes – is ordinal and not numeric

Not exact results

SR vs population sizes III

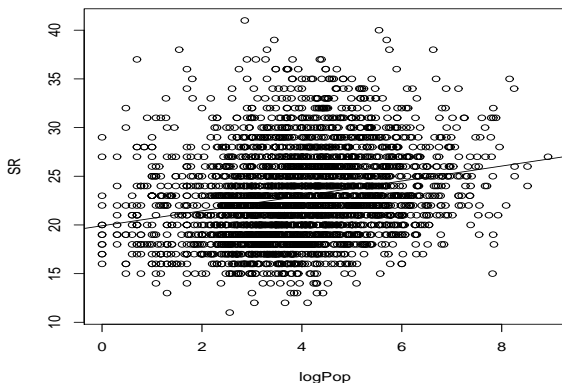


Figure: Uncontrolled correlation between 3153 languages and population sizes (> 1). $r = .236$

SR vs population sizes IV

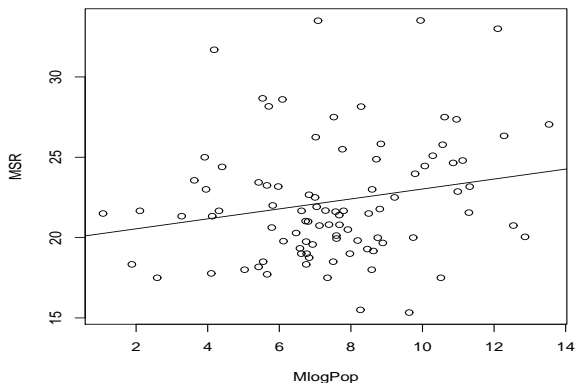


Figure: Genealogically controlled correlation between 91 language families and population sizes (> 2). $r = .18$ with $p = .0485$

SR vs population sizes V

A slight gain obtained by raising the cutoff for language family size. No clear cut threshold.

SR and geography I

- ▶ Atkinson (2011) claims: phoneme inventory size decreases as one moves away from Africa
- ▶ We investigate the claim using ASJP data
- ▶ We do not take the centroid point as a **homeland** for a family
- ▶ We use the coordinates calculated by Wichmann, Müller & Velupillai (2010) for worlds' language families
- ▶ Homeland is defined as the place which has the maximum language diversity

SR and geography II

- ▶ Addis Ababa was taken as the homeland from Africa since it is equidistant from other African families
 1. Afro-Asiatic 3367 km
 2. Khoisan 3862 km
 3. Niger-Congo 3676 km
 4. Nilo-Saharan 1099 km
- ▶ We use great circle distance and continental waypoints to constrain the migration paths
- ▶ Waypoints are Cairo, Istanbul, Phnom Penh, Bering Strait and Panama

SR and geography III

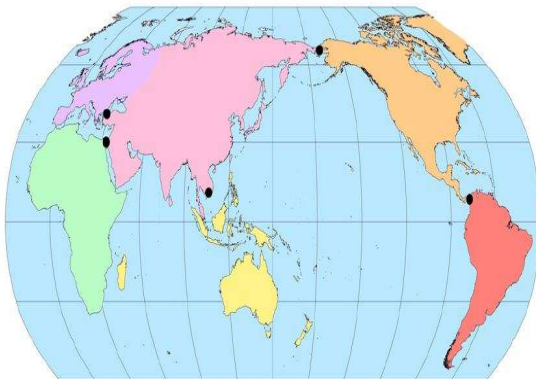


Figure: World map showing waypoints (Atkinson 2011)

SR and geography IV

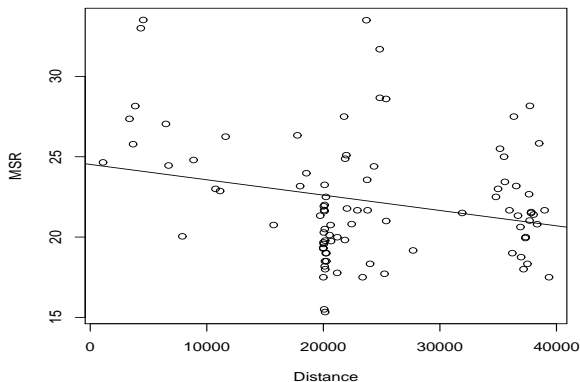


Figure: MSR vs distance from Addis Ababa shows $r = .23$, $p = .015$

Further Regressions I

- ▶ A simple regression of mean of logPop and distance from Africa gives $r = -.34$, $p = .0005$ for families as a unit of analysis
- ▶ Also there is a inverse correlation between mean of logPop and distance for 5602 languages $r = -.45$
- ▶ There seems to be a **conspiracy** between logPop and MSR

Further Regressions II

- ▶ A multiple regression with $MSR \sim Dist + MlogPop$ gives $R^2 = .059$ and $p = .025$
 1. $p = .060$ for distance
 2. $p = .23$ for log population
- ▶ Effect of distance is not significant when MlogPop is controlled

Further Regressions III

- ▶ link?
- ▶ A multiple regression with $MSR \sim MMWL + Dist + MlogPop$ gives $R^2 = .11$ and $p = .0044$
 1. $p = .0165$ for MMWL
 2. $p = .1413$ for Dist
 3. $p = .5952$ for MlogPop
- ▶ A significant correlation between MSR and MMWL.
- ▶ Suggests that MMWL mediates between MSR and MlogPop

Further Regressions IV

- ▶ A multiple regression with $\text{MMWL} \sim \text{MSR} + \text{Dist} + \text{MlogPop}$ gives $R^2 = .173$ and $p = .0002$
 1. $p = .0165$ for MSR
 2. $p = .2391$ for Dist
 3. $p = .018$ for MlogPop
- ▶ This is a new relation between population and MMWL suggesting that MMWL mediates between MSR and population

Discussion

- ▶ SR and MWL can be explained in terms of competing forces for perception (Chinese)
- ▶ Nettle (1999) indicates that languages with larger populations have shorter words
- ▶ This shows up in the multiple regression
- ▶ Atkinson (2011) suggests prehistoric bottlenecks to account for SR vs Dist
- ▶ Languages of North America has large phoneme inventory sizes

Conclusion

- ▶ Regression analysis reveals a chain of three effects:
 1. Distance from Africa is associated with smaller populations
 2. which is related with longer words
 3. which in turn is related with fewer phonemes

Outline

ASJP: Phoneme diversity, word length and population size

ASJP: Automatic classification

Language distance estimation from parallel corpora

Dravidian languages

Dravidian languages: Top level branching

References

Similarity between lexical items

- ▶ String Similarities can be used for this purpose
- ▶ Levenshtein Distance is one of the most widely used measures
 - ▶ Defined as minimum number of additions, deletions and substitutions required to transform one word into another
- ▶ ASJP ¹ Consortium (Holman et al. 2008a, Holman et al. 2008b) use a modified version for calculating the distance

¹Automated Similarity Judgement Program

More about ASJP methodology

- ▶ Levenshtein distance is normalized by the maximum of the lengths of the two words
 - ▶ Accounts for length bias
 - ▶ Transforms the distance in the range of $[0,1]$
 - ▶ Allows for comparison across word pairs
- ▶ Distance between two languages is the mean of the Levenshtein distance between the word pairs having the same meaning, LDN (Levenshtein Distance Normalized).
- ▶ LDN normalized by the mean of the $n(n-1)/2$ words yields LDND (Levenshtein Distance Normalized Divided)
 - ▶ Corrects chance similarity between the compared languages

Our approach

- ▶ Petroni & Serva (2010) claim that LDN is better than LDND for distinguishing related languages from unrelated languages
- ▶ Wichmann, Holman, Bakker & Brown (2010) empirically show that LDND is better than LDN.

To our knowledge, there has been no study regarding the comparison of various orthographic similarity measures as applied to the ASJP dataset. We test other orthographic measures DICE, LCSR for this purpose as defined in Inkpen, Frunza & Kondrak (2005)

DICE, LCSR

- ▶ DICE is defined as
 - ▶ Twice the number of common bigrams divided by the total number of bigrams in the two words
- ▶ LCSR is defined as
 - ▶ Longest common subsequence between the two words divided by the maximum of the lengths of the words
- ▶ Both LCSR and DICE are similarity measures
- ▶ Converted into distance measures by subtracting it from 1

Extensions of DICE, LCSR

DICED (Dice distance) is defined in the same spirit as LDN.

DICEDN (Dice distance normalized) defined similar to LDND

LCSR converted to a distance measure is defined as LCSR_D

Dataset

- ▶ The dataset for these experiments are the ASJP's version 12 database.
- ▶ It consists of a 40-item Swadesh list for 4169 languages
- ▶ Removed pidgins, creoles, artificial languages, languages extinct before 1700 CE and language families with less than 10 languages.
- ▶ Final size of the dataset is 3730 languages. There are 49 language families.

Experiments

- ▶ Comparing our results with ASJP:
 - ▶ Replicated LDN, LDND on the same dataset
 - ▶ Observed that 60 languages didnot have English glosses.
These were not included for the analysis
- ▶ The experiments are computationally expensive, taking days for computing a single measure over the whole dataset
- ▶ The row computations are parallelized using **ppss**

Results and Evaluation

LDN	LDND	DICED	DICEDN	LCSR
6.145	8.143	10.988	11.964	4.937

- ▶ The results were evaluated using the distinctiveness measure.
- ▶ Distinctive measure is defined as $(d_{out} - d_{in})/sd_{out}$
 - ▶ d_{out} is the mean of the distance of languages from the family to outside the family.
 - ▶ d_{in} is the mean of the distance of languages within the family.
 - ▶ sd_{out} is the standard deviation of d_{out}
- ▶ It reflects how well the measure performs in grouping related languages.

Outline

ASJP: Phoneme diversity, word length and population size

ASJP: Automatic classification

Language distance estimation from parallel corpora

Dravidian languages

Dravidian languages: Top level branching

References

What we do...

- ▶ Automatically identify the genetic relationships between languages from parallel corpus
- ▶ Estimate the distance matrix between the languages
- ▶ Use a clustering algorithm to infer the family tree based on distance matrix

Cognates

- ▶ Cognates are words which are genetically related.
 - ▶ English/German: **hound/Hund**
 - ▶ English/Armenian: **two/erku**
- ▶ Loanwords from other languages are not considered as cognates
 - ▶ English/Sanskrit : **avatar**

Automatic identification of cognates I

- ▶ *Cognates* in Computational linguistics
 - ▶ *false positives* : loanwords and chance resemblances
 - ▶ *false negatives* : genetically related but different forms

Automatic identification of cognates II

- ▶ Orthographic measures are generally used for judging the similarity between the word pairs
- ▶ Methods such as HMMs require initial training data

Previous Work

- ▶ Koehn (2005) trains pair-wise statistical translation systems
- ▶ Clusters the languages based on BLEU scores
- ▶ **Assumption:** ease of translation correlates with genetic closeness

Automatic Identification of Cognates

We require a method with high precision.

- ▶ Automatically identify word pairs which are translations of each other
- ▶ Use an orthographic measure for computing the similarity between each word pair
- ▶ Remove word pairs which are below a particular *cutoff*
- ▶ We use *longest common subsequence ratio* as the orthographic measure
- ▶ The *cutoff* is fixed at 0.58 to account for length bias

Distance Estimation

- ▶ Distance between two languages l_a, l_b can be expressed using the following equation:

$$Dist(l_a, l_b) = 1 - \frac{\sum_i sim(l_a^i, l_b^i)}{N} \quad (1)$$

- ▶ $sim(l_a^i, l_b^i)$ is the similarity between the i th cognate pair and is in the range of $[0, 1]$
- ▶ N is the length of the cognate list

String Similarity Measures

- ▶ Levenshtein distance is defined as the minimum number of insertions, deletions and subtractions to transform a string into other
- ▶ Dice is defined as twice the overlap of the number of bigrams divided by the total number of bigrams
- ▶ LCSR is defined as the length of the longest common subsequence divided by the maximum length of the two strings

Dataset: Europarl Corpus

- ▶ The corpus is from English to ten languages
- ▶ 45 pairs of parallel corpora were created by using English as bridge language
- ▶ The first 100,000 sentences were included
- ▶ Every language except Finnish is a Indo-European language
- ▶ All the other languages fall into different branches of Indo-European language family, Germanic and Romance

Experiments

- ▶ MOSES (Koehn et al. 2007) was used for aligning the words
- ▶ Cognate pairs extracted from word alignments
- ▶ For every language pair, word pairs with **LCSR** less than cutoff were removed
- ▶ We experimented with three string similarity measures
Levenshtein Distance, Dice and LCSR
- ▶ UPGMA as implemented in PHYLIP was used to cluster these distances

Probable Cognates

Language	# Probable Cognates
English	1458
German	1043
Dutch	1489
Swedish	2624
Danish	2149
French	955
Spanish	823
Portuguese	831
Italian	1333

Table: The number of probable cognates of each language with Finnish.

Trees I

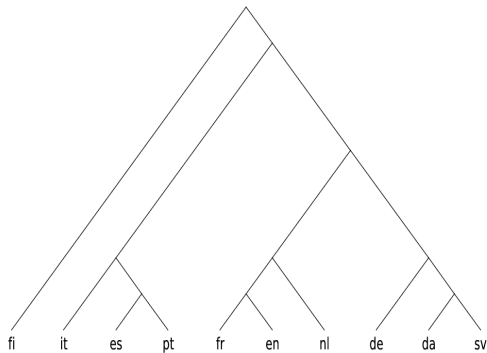


Figure: **Levenshtein Distance**

Trees II

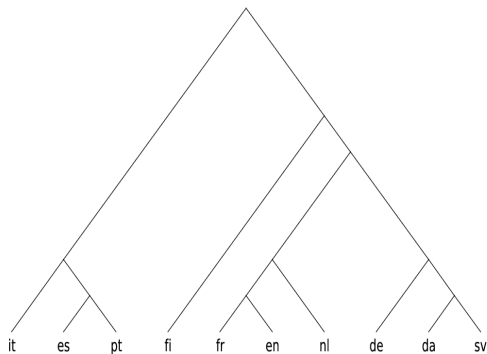


Figure: **Dice**

Trees III

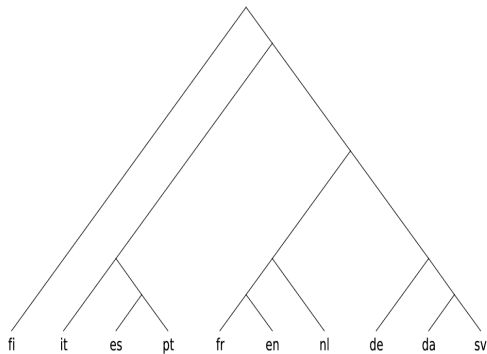


Figure: **LCSR**

Discussion

- ▶ Working with corpus avoids the subjectivity involved in collecting the Swadesh list
- ▶ It also brings in automation which is not available in Dyen et al. (1992)
- ▶ The tree on the whole agrees with the commonly accepted subgrouping
- ▶ Comparing with Koehn (2005) it returns lower order relationships better than the higher order

Conclusion

- ▶ Use different string similarity measures for estimation of genetic distances from word pairs automatically extracted from parallel corpus
- ▶ Indicates a parallel corpus could be used for this kind of study

Future Work

- ▶ Automatically distinguish cognates from loanwords
- ▶ Incorporate syntactic and semantic features in the future
- ▶ Use POS tags and context vectors for estimating the similarity

Outline

ASJP: Phoneme diversity, word length and population size

ASJP: Automatic classification

Language distance estimation from parallel corpora

Dravidian languages

Dravidian languages: Top level branching

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. **A**utomated **S**imilarity **J**udgment **P**rogram (ASJP; Wichmann et al.(2010))

Introduction III

- ▶ 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)

Introduction V

- ▶ Contributions of this work:
 - ▶ Create two new diachronic datasets for Dravidian from **D**ravidian **E**tymological **D**ictionary **R**evised (*DEDR*; Burrow & Emeneau (1984)) and Krishnamurti (2003)
 - ▶ Results of subgrouping Dravidian languages applying distance-based methods to different datasets
 - ▶ Answer to the question of ternary vs. binary branching of Proto-Dravidian

Krishnamurti (2003)

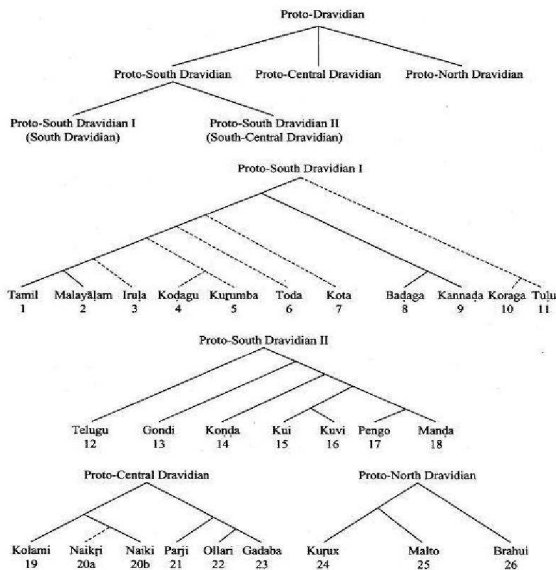


Figure: Dravidian family tree

Issues: Krishnamurti (2003)

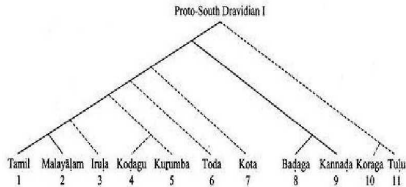


Figure: South Dravidian I family tree

- ▶ Position of the Nilgiri languages (Toda, Kota, Irula, Badaga and Kurumba) in relation to Tamil and Kannada
- ▶ Position of Tuḷu
- ▶ Placement of Koraga
- ▶ Relation between Toda and Kota
- ▶ Central Dravidian: Position of Naikri

Issues: WALS (Haspelmath et al. 2011)

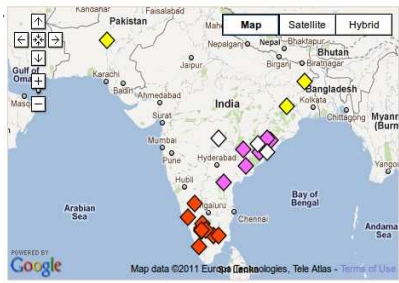
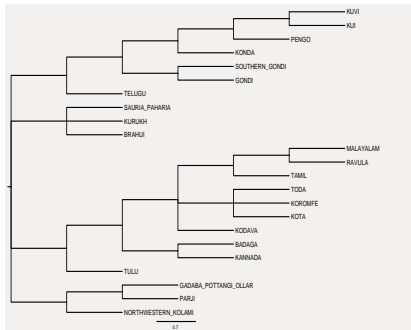


Figure: WALS distribution of Dravidian language family

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

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

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

Related work IV

- ▶ 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

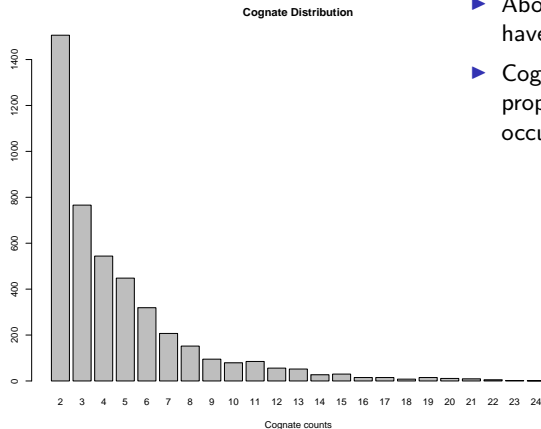
- ▶ Kolachina et al. (2011)
 - ▶ Krishnamurti (2003) used 27 comparative features for supporting ternary branching over binary branching
 - ▶ 1/0/? (presence, absence or missing)
 - ▶ Apply MP to address question of ternary branching vs. binary branching
 - ▶ Conclusion: Branch lengths returned by MP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

- ▶ 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

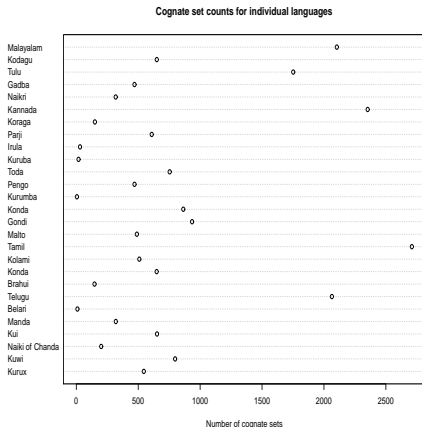
- ▶ Comparative features:
 - ▶ Character-based dataset from Kolachina et al. (2011)
 - ▶ Naikṛi and Naiki of Chanda treated as a single language
- ▶ ASJP lists:
 - ▶ Consists of only those languages which could be mapped with *DEDR* or Krishnamurti (2003)
 - ▶ 20 languages from all the four major subgroups

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

Exploring CD and RD II



- ▶ Five languages are over-represented
- ▶ All the five languages are literary (semi-literary: Tulu).
- ▶ Irula, Kuruba, Kurumba and, Belari are represented the least
- ▶ Similar distribution observed for RD

Exploring CD and RD III

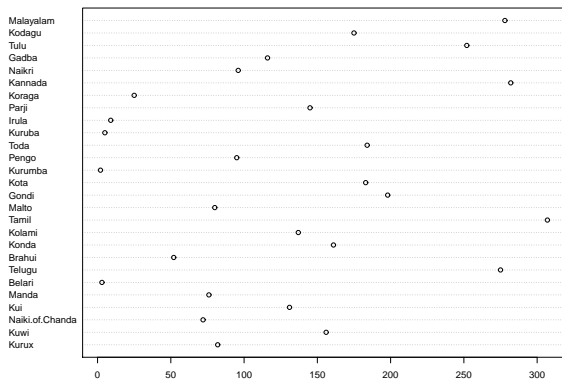
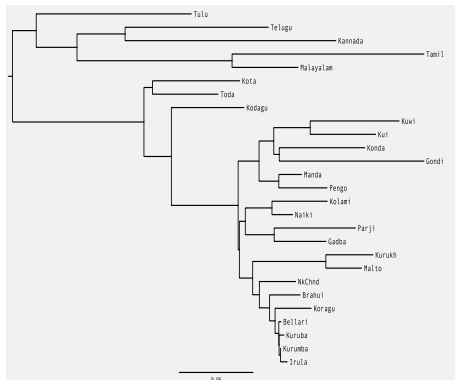


Figure: Cognate set distribution for individual languages in RD

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 Kannaḍa, DEDR lists both as Kannaḍa
- ▶ Naikṛi and Naiki of Chanda are **related**?
- ▶ Koragu (Koraga) & Bellari; Kuruba and two other Nilgiri languages (Iruḷa and Kuṛumba)
- ▶ Languages from CDr and SDr I mixed

Experiments: CD III

- UPGMA tree similar to NJ tree

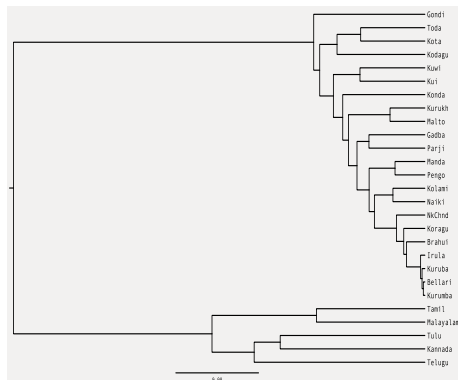


Figure: UPGMA tree

Experiments: CD IV

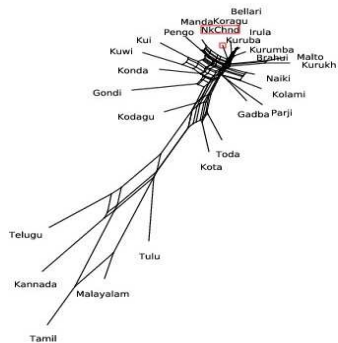
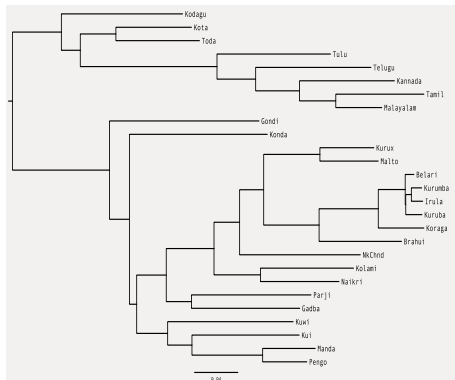


Figure: Neighbor Network

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: Naikṛi, Kolami, Gadba and Parji grouped together next to NDr languages

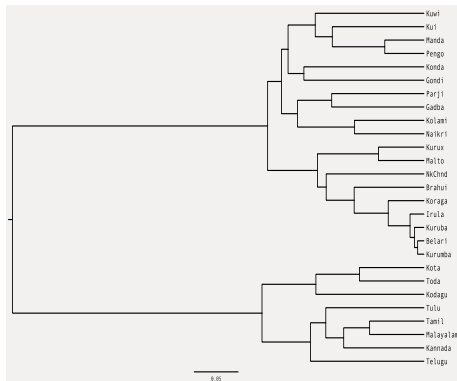
Experiments: RD I



- ▶ 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

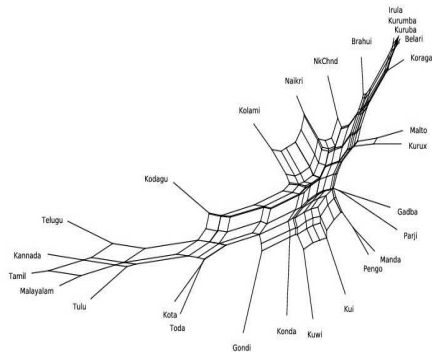


Figure: Neighbor Network

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, Kuṛumba, Iruḷa, 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

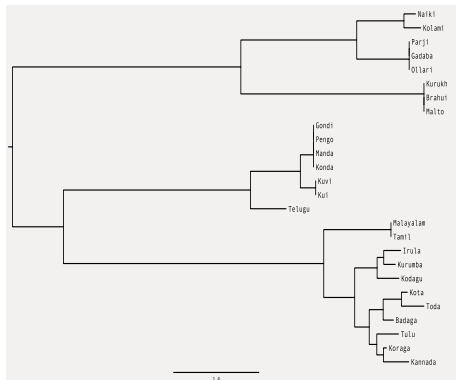


Figure: NJ tree

- ▶ 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

Experiments: Comparative features II

- UPGMA tree
same as NJ tree

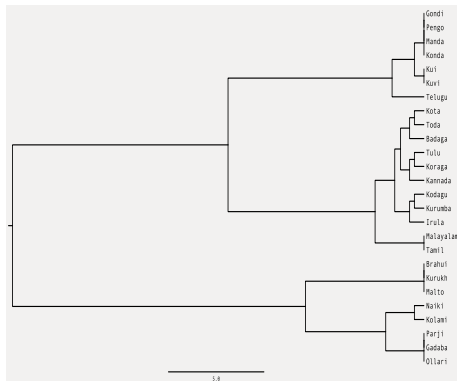


Figure: UPGMA tree

Experiments: ASJP I

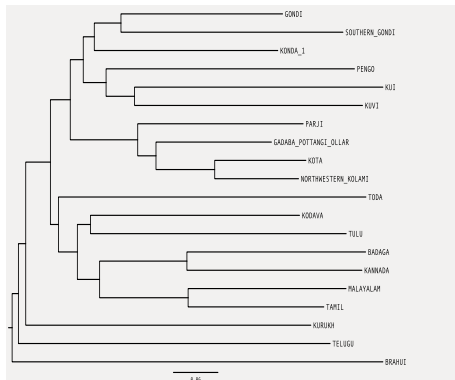


Figure: NJ tree

- ▶ 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

Experiments: ASJP II

- ▶ None of the major subgroups clearly resolved

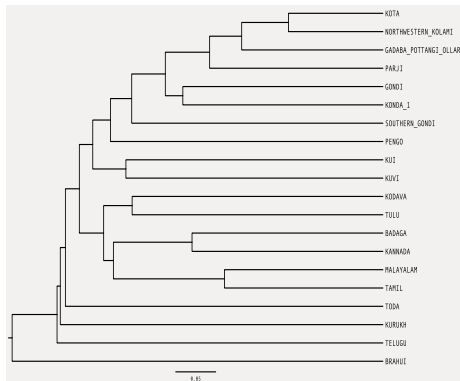


Figure: UPGMA tree

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

ASJP: Phoneme diversity, word length and population size

ASJP: Automatic classification

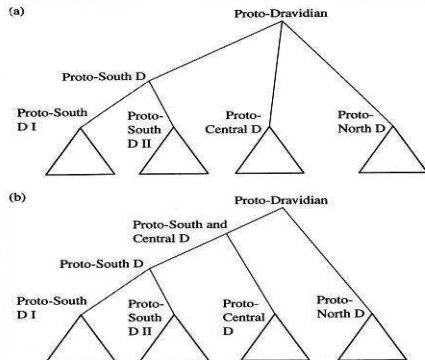
Language distance estimation from parallel corpora

Dravidian languages

Dravidian languages: Top level branching

References

Subgrouping of the Dravidian languages I



Subgrouping of the Dravidian languages II

- ▶ Two possible subgroupings of the Dravidian languages according to (Krishnamurti 2003)
- ▶ Alternative (a): ternary branching of Proto-Dravidian (PD)
- ▶ Alternative (b): binary branching of PD
- ▶ Alternative (a) adopted in (Krishnamurti 2003)
- ▶ Aim of the current work: 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

Subgrouping of the Dravidian languages III

- ▶ Dataset: Features for comparative phonology, morphology and syntax used for subgrouping (Krishnamurti 2003)²
- ▶ 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
- ▶ Application of MP can also shed light on other uncertainties in classification

²Available on request

Maximum Parsimony (MP) method I

- ▶ A well-known discrete character-based phylogenetic inference method
- ▶ MP infers phylogeny from character sequences representing taxa
- ▶ MP in a nutshell: Search among all possible phylogenies for the one or ones with the minimum number of evolutionary events

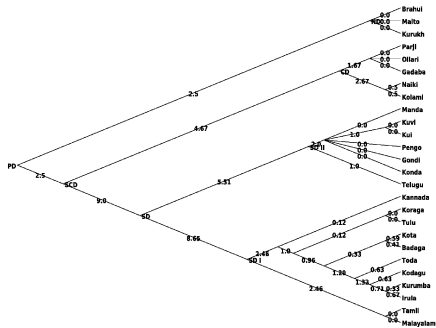
Maximum Parsimony (MP) method II

- ▶ Choice of MP
 - ▶ 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 both bifurcating and multifurcating trees

Experimental setup I

- ▶ Bootstrapping procedure run for 10000 times with 'sampling with replacement'
- ▶ Consensus tree estimated using majority consensus
- ▶ Branch lengths on the consensus tree re-estimated using the pars program
- ▶ The consensus tree is rooted using the North Dravidian (ND) clade as the outgroup
- ▶ PHYLIP (Felsenstein 1993) is used for the experiments

Experimental setup II



Results

- ▶ Interpreting inferred trees
 - ▶ Number of state changes that take place along the branches, indicated by the branch lengths
 - ▶ Two internal branches having the same state changes can be eliminated
 - ▶ Difference in branch lengths between SCD and SD, and SCD and CD is 4.33

Conclusion I

- ▶ Main conclusion: Tree inferred using MP is binary 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))
- ▶ Since MP assumes a homoplasy-free scenario, it treats these similarities between CD and SD II as a result of a common stage in their evolution: Proto South-Central Dravidian (SCD)

Conclusion II

- ▶ Additional outcomes: MP tree resolves other uncertainties such as position of Nilgiri languages

Outline

ASJP: Phoneme diversity, word length and population size

ASJP: Automatic classification












Language distance estimation from parallel corpora

Dravidian languages

Dravidian languages: Top level branching

References






References I

-  Andronov, M. (1964), 'Lexicostatistic analysis of the chronology of disintegration of proto-dravidian', *Indo-Iranian Journal* **7**(2), 170–186.
-  Atkinson, Q. D. (2011), 'Phonemic diversity supports a serial founder effect model of language expansion from Africa', *Science* **332**(6027), 346.
-  Burrow, T. H. & Emeneau, M. B. (1984), 'A Dravidian Etymological Dictionary (rev.)'.
-  Dyen, I., Kruskal, J. B. & Black, P. (1992), 'An Indo-European classification: A lexicostatistical experiment', *Transactions of the American Philosophical Society* **82**(5), 1–132.
-  Felsenstein, J. (1993), 'PHYLP (phylogeny inference package) version 3.5 c', *Department of Genetics, University of Washington, Seattle* **1118**.
-  Haspelmath, M., Dryer, M. S., Gil, D. & Comrie, B. (2011), 'WALS online', *Munich: Max Planck Digital Library* .
<http://wals.info>.
-  Hay, J. & Bauer, L. (2007), 'Phoneme inventory size and population size', *Language* **83**(2), 388–400.
-  Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A. & Bakker, D. (2008a), 'Advances in automated language classification', *Quantitative Investigations in Theoretical Linguistics* .
-  Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A. & Bakker, D. (2008b), 'Explorations in automated language classification', *Folia Linguistica* **42**(3–4), 331–354.
-  Huson, D. & Bryant, D. (2006), 'Application of phylogenetic networks in evolutionary studies', *Molecular biology and evolution* **23**(2), 254.
-  Inghen, D., Frunza, O. & Kondrak, G. (2005), Automatic identification of cognates and false friends in French and English, in 'Proceedings of the International Conference Recent Advances in Natural Language Processing', pp. 251–257.

References II

- Koehn, P. (2005), Europarl: A parallel corpus for statistical machine translation, *in* 'MT summit', Vol. 5.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. et al. (2007), Moses: Open source toolkit for statistical machine translation, *in* 'Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions', Association for Computational Linguistics, pp. 177–180.
- Kochina, S., Rama, T. & Lakshmi Bai, B. (2011), 'Maximum parsimony method in the subgrouping of dravidian languages', *QITL-4* **4**, 52.
- Krishnamurti, B. (1978), 'Areal and lexical diffusion of sound change: Evidence from dravidian', *Language* pp. 1–20.
- Krishnamurti, B. (2003), *The Dravidian languages*, Cambridge Univ Pr.
- Krishnamurti, B., Moses, L. & Danforth, D. (1983), 'Unchanged cognates as a criterion in linguistic subgrouping', *Language* **59**(3), 541–568.
- Lewis, P. M., ed. (2009), *Ethnologue: Languages of the World*, sixteenth edn, SIL International, Dallas, TX, USA.
- Maddieson, I. & Precoda, K. (1990), 'UPSID-PC', *The UCLA Phonological Segment Inventory Database*.
- McMahon, A. & McMahon, R. (2007), 'Language families and quantitative methods in south asia and elsewhere', *The Evolution and History of Human Populations in South Asia* pp. 363–384.
- 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.
- Nettle, D. (1999), *Linguistic Diversity*, Oxford University Press, Oxford.
- Petroni, F. & Serva, M. (2010), 'Measures of lexical distance between languages', *Physica A: Statistical Mechanics and its Applications* **389**(11), 2280–2283.

References III

-  Rama, T., Kolachina, S. & Lakshmi Bai, B. (2009), 'Quantitative methods for phylogenetic inference in historical linguistics: An experimental case study of south central dravidian', *Indian Linguistics* **70**.
-  Wichmann, S. (2010), 'Internal language classification', *Continuum Companion to Historical Linguistics* p. 70.
-  Wichmann, S., Holman, E. W., Bakker, D. & Brown, C. H. (2010), 'Evaluating linguistic distance measures', *Physica A: Statistical Mechanics and its Applications* **389**, 3632–3639.
-  Wichmann, S., Müller, A. & Velupillai, V. (2010), 'Homelands of the world's language families: A quantitative approach', *Diachronica* **27**(2), 247–276.
-  Wichmann, S., Müller, A., Velupillai, V., Brown, C. H., Holman, E. W., Brown, P., Urban, M., Sauppe, S., Belyaev, O., Molochieva, Z., Wett, A., Bakker, D., List, J.-M., Egorov, D., Mailhammer, R. & Geyer, H. (2010), 'The ASJP database (version 12)'. <http://email.eva.mpg.de/wichmann/listss12.zip>.