FINAL REPORT

LANGUAGE PHYLOGENES

Samudra Banerjee (109294061; sabanerjee@cs.sunysb.edu)
Saipreethy (109231628; smanickavall@cs.sunysb.edu)
Phaneendra Reddy (109346762; pyreddy@cs.sunysb.edu)

INTRODUCTION

Phylogenic trees have been widely used to study the evolution of various organisms. Since the last decade or so, there has been a lot of work on the application of these trees in studying the evolution of languages [1]. There are broadly two ways to construct phylogenic trees - feature or character based approach and distance based approach. In the first case, we create a feature matrix where each row is the list of languages and each column is a feature we would like to consider. Thus, we would have a feature list for each language which we need to encode and then feed this matrix to the tree creation program. Selecting which features to use is very important and has been one of the main issues in critiquing a phylogenetic analysis [1]. Coding of these features is critically important as well and it is here that even trained linguists make mistakes [1]. Once our feature matrix is ready, we need an algorithm to find the best tree (for instance, one with minimum number of evolutionary changes – Maximum Parsimony) which matches these features. The problem of finding such a "best tree" is an NP-hard problem and since finding the global optimal solution takes exponential time, most of the softwares available today for generating phylogenic trees employ the greedy heuristic 'hill-climbing' technique [1]. Distance based approaches simply convert this feature matrix into a distance matrix (i.e. similarity between languages) and constructs the phylogenic tree out of that. Some of the recent works [3][4] use Levenshtein distance (edit-distance) approach in which they solely consider words from the Swadesh list [5] and simply ignore words which do not exist in a language. Techniques like these do not consider a host of features and hence there is a difference in accuracy. However, the need to consider a rich feature set is eliminated.

In this project, we produce a Phylogenic tree of Wikipedia Languages from a mixture of the following two sources –

1. Automated Similarity Judgement Program (ASJP) database of words in languages – for lexical data
2. World Atlas of Languages (WALS) – for typological data

We choose a set of 122 languages to build our tree on. Our results show that a good mix of data from both the sources results in a good tree.

CHOICE OF SOURCES

We choose to use data from WALS [10] and ASJP [6], because data from these sources have been used in previous works [4][7] and have provided good results. Following is a brief description of these sources:

The World Atlas of Languages (WALS)

WALS contains a set of typological features and their corresponding values for a subset of the world's languages. Each feature can have a certain number of states. For instance, feature 10A (Vowel Nasalization) can take only two states – present or absent and its value is 1 (present) for French, but 2 (absent) for English. However, feature 30A (Number of Genders) takes 5 states (None, Two, Three, Four and Five or more). Since few features contain values for all or most of the languages, the WALS matrix is very sparse. Only around 16% of the cells have actual values and many of the features are mutually dependent [7]. As such, there is need for a set of the most stable WALS features. Thus, like many of the previous works, we consider the Wichmann and Holman's most stable feature set [2]. WALS also provides its data for download in CSV format, which can be used to computationally extract the features we need.

The Automated Similarity Judgement Program (ASJP) database

For lexical data, we use the ASJP database. The ASJP database consists of lexical data for most of the world's languages. These data are the transliteration of the standard Swadesh Word List [5], augmented with some phonetic characteristics. ASJP data is available for download as a single text file for all the languages. For our project, we extract the data for the languages in consideration in put them in separate files for ease of computation.

CHOICE OF LANGUAGES

Our choice of languages was based solely on availability of data in the above sources. We initially began with the list of all Wikipedia languages [12]. There are 287 languages in Wikipedia. Out of them, data for only 225 of them were available on WALS. We built our initial tree on this set of languages. However, since evaluating our tree was necessary, we had to restrict ourselves to a smaller subset – only those languages present in the ASJP World Language Tree of Lexical Similarity [13]. This reduced our list to 161 languages. To improve the accuracy of our tree, we had to cut down on languages which were not that well represented in WALS (had values for 10 or less out of the 62 stable features). This left us with 122 languages. Following is the final list of the languages we considered:

fijian,samoan,tahitian,hawaiian,tongan,maori,chamorro,malagasy,sundanese,minangkabau,indonesian,ilokano,pangasinan,kapampangan,cebuano,tagalog,vietnamese,khmer,burmese,chinese,cantonese,hakka,nahuatl2,aymara,thai,kabyle,hausa,egyptian_arabic,arabic_modern_standard,arabic_moroccoan,maltese,hebrew,amharic,kalmyk,albanian,telugu,kannada,tamil,malayalam,igbo,wolof,greek,basque,chuvash,tuvan,sakha,bashkir,tatar,karachay_balkar,uzbek,karakalpak,uyghur,turkish,turkmen,azerbaijani,korean,navajo,urdu,nepali,hindi,gujarati,marathi,bengali,kashmiri,sorani,tajik,persian,armenian_eastern,armenian_western,latvian,lithuanian,ukrainian,russian,polish,czech,bulgarian,slovenian,serbo_croatian,breton,welsh,irish,scottish_gaelic,romanian,italian,catalan,portuguese,spanish,french,english,swedish,norwegian_nynorsk,danish,icelandic,pennsylvania_german,german,afar,somali,cherokee,kabardian_circassian,choctaw,chechen,lak,avar,bambara,fula6,georgian,kanuri,hungarian,erzya,udmurt,meadow_mari,finnish,estonian,japanese,greenlandic,swahili,shona,ndonga,luganda,kinyarwanda,xhosa,zulu

METHODOLOGY

We use both the feature-based approach and distance-based approaches for constructing the tree. The latter allows room for inclusion of ASJP data and hence gives better results.

Feature Based Approach

The sequence of steps we followed to create the feature matrix of the Wikipedia Languages from WALS features is as follows:

1. Fetch the WALS codes for the languages in [12]. Using PYTHON scripts, we computationally achieve this step. Some languages needed manual verification as their names in Wikipedia did not exactly match those in WALS. Out of 287 languages in the list, 225 are present in WALS.
2. We programmatically compute the feature matrix (of all WALS features) for these 225 languages. We notice that the feature matrix becomes much denser if we take only the Wikipedia languages.
3. We reduce our matrix further to consider only the most stable WALS features. Here again, we remove the overlapping features for feature 81A (Order of Subject, Object, Verb). Since features 83A (Order of Object and Verb) and 82A (Order of Subject and Verb) overlap with 81A, we consider only 81A.

Since all of these features have values for some (in fact many) of the Wikipedia languages (as found computationally), we do not reduce our matrix further. A part of our resultant matrix (as seen in CSV format) is as follows:

wals_code	31A	118A	30A	119A	29A	85A	28A
eng	2	2	3	2	2	2	2
dut		2		2		2	
ger	2		3		2	2	3
swe		2		2		2	
fre	2	2	2	2	2	2	3
ita						2	
rus	2	2	3	2	3	2	3
spa	2	2	2	1	2	2	3
pol		2		2		2	
wwy							
ceb		1		1		2	

PHYLIP is a software suite and has a program, “pars” [15], which generates the phylogenic tree using Maximum Parsimony. However, here we have the restriction of having no more than 8 states per feature. We therefore remove features 51A and 33A which have 9 character states, assuming that removal of two features should not affect our analysis. Also, since we would be comparing our results with that of an already computed tree, we reduce our taxa to those languages which are present in that tree

Using Maximum Parsimony over the set of languages and using the WALS features does not yield a pretty good result (as discussed later). We therefore shift our approach to combine some lexical features from ASJP in our data.

Distance Based Approach

We collect data from ASJP database [11] and code our distance matrix based on Levenshtein Distance between the words in the languages. The distance between two languages is the average of the distances between their corresponding words. We call this distance the *ASJP Distance*.

We later enhance our distance score by augmenting it with *WALS distance* score from the feature matrix above. We calculate the WALS distance score as follows:

$$WALS\ Distance = \frac{Number\ of\ features\ in\ common}{Total\ number\ of\ features}$$

The net distance between languages L1 and L2 is therefore given by:

$$DIST_{(L1,L2)} = A \times WALSDistance_{(L1,L2)} + B \times ASJPDistance_{(L1,L2)}$$

Where A and B are less than 1 and A+B=1. From our experiments, discussed later, we find that we get an optimal tree when we choose A and B to be **0.4 and 0.6** respectively, i.e. we give 40% weightage to WALS and 60% to ASJP.

We thus have a distance matrix, which is fed to “fitch” [16] – a program in PHYLIP to compute the Phylogenic Tree based on distance. A snapshot of the distance matrix is as follows.

```
lda       0.0000 0.8973 0.9169 0.9577 0.8764 0.9001
rom       0.8973 0.0000 0.9252 0.9326 0.7075 0.5895
kbl       0.9169 0.9252 0.0000 0.9410 0.9039 0.9349
tkm       0.9577 0.9326 0.9410 0.0000 0.8938 0.9261
alb       0.8764 0.7075 0.9039 0.8938 0.0000 0.7667
```

PARS or FITCH generate the tree in parenthesis format. Our resultant tree in parenthesis format generated as such by FITCH is as under (branch lengths have been stripped off).

```
(kin,((swa,shn),(((fni,(((((igb,(khm,(tha,vie))),((((mal,((min,i
nd),sun)),(((pnn,ilo),kpm),(tag,ceb))),cha),(fij,(((mao,tah),(sa
m,tng)),haw)))),hau),wlf),((knr,bam),((som,nav),(((((((lat,lit),
(((((cze,(slo,scr)),pol),bul),ukr),rus)),(((eng,(dut,ger)),(ice,
(dsh,(swe,nor)))),(alb,(grk,(fre,((rom,ctl),(por,(spa,ita)))))))
),((gae,iri),(bre,wel))),((taj,prs),krd)),(nhn,(((((((chv,(tur,(
(((tkm,aze),(uyg,(kkp,(uzb,krc)))),(bsk,tvo)),(ykt,tuv)))),kmk),
((moe,(udm,mme)),(hun,(est,fin)))),((bsq,geo),(arm,arw))),(((kab
,qaf),(((lak,ava),chc),aym)),(((knd,(mym,tml)),tel),(kas,((((urd
,hin),guj),(ben,nep)),mhi))))),(cct,che)),(grw,(((cnt,hak),mnd),
(kor,(jpn,brm))))))),((amh,((mlt,(heb,aeg)),(amr,ams))),kbl)))))
),ndo),(zul,xho))),lda);
```

For visualization, we use the DRAWGRAM [17] program in PHYLIP which takes the parenthesis tree as input and generates the visual tree. For ease of visualization, we replace the language codes above with the corresponding names to get the new parenthesis tree:

```
(kinyarwanda,((swahili,shona),(((fula6,(((((igbo,(khmer,(thai,vi
etnamese))),((((malagasy,((minangkabau,indonesian),sundanese)),(
((pangasinan,ilokano),kapampangan),(tagalog,cebuano))),chamorro)
,(fijian,(((maori,tahitian),(samoan,tongan)),hawaiian)))),hausa)
,wolof),((kanuri,bambara),((somali,navajo),(((((((latvian,lithua
nian),(((((czech,(slovenian,serbo_croatian)),polish),bulgarian),
ukrainian),russian)),(((english,(pennsylvania_german,german)),(i
celandic,(danish,(swedish,norwegian_nynorsk)))),(albanian,(greek
,(french,((romanian,catalan),(portuguese,(spanish,italian)))))))
),((scottish_gaelic,irish),(breton,welsh))),((tajik,persian),sor
ani)),(nahuatl2,(((((((chuvash,(turkish,((((turkmen,azerbaijani)
,(uyghur,(karakalpak,(uzbek,karachay_balkar)))),(bashkir,tatar))
,(sakha,tuvan)))),kalmyk),((erzya,(udmurt,meadow_mari)),(hungari
an,(estonian,finnish)))),((basque,georgian),(armenian_eastern,ar
menian_western))),(((kabardian_circassian,afar),(((lak,avar),che
chen),aymara)),(((kannada,(malayalam,tamil)),telugu),(kashmiri,(
(((urdu,hindi),gujarati),(bengali,nepali)),marathi))))),(choctaw
,cherokee)),(greenlandic,(((cantonese,hakka),chinese),(korean,(j
apanese,burmese))))))),((amharic,((maltese,(hebrew,egyptian_arab
ic)),(arabic_moroccoan,arabic_modern_standard))),kabyle)))))),nd
onga),(zulu,xhosa))),luganda);
```

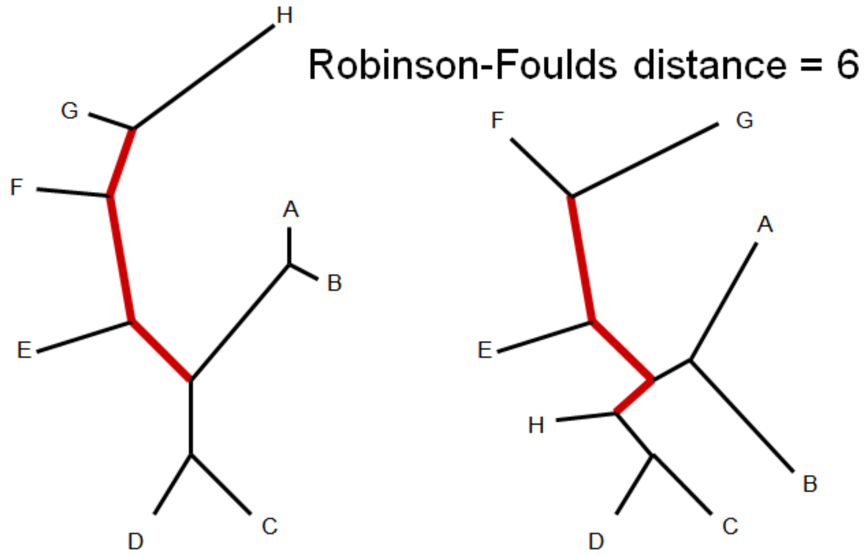
The visual tree is available in APPENDIX A.

EVALUATING OUR TREE

Using the TREEDIST [14] program in PHYLIP, we find the distance between our tree and the ASJP World Language Tree [13]. We manually construct the parenthesis tree for the ASJP World Language Tree since we could not find any parenthesis tree of the world readily available. We initially construct the tree for 161 languages. Later we programmatically reduce the tree to 122 languages for evaluation of the above tree. Following is the reduced parenthesis tree of [13].

```
(((((((fij,(sam,(tah,(haw,(tng,mao))))),((cha,((mal),(((sun),(mi
n,(ind))),(((ilo,pnn),kpm),(((ceb),tag)))))),(((vie,khm),((brm,(
(mnd,(cnt,hak)))),(nhn,(((aym),(tha)),(((kbl,hau),((((((aeg,ams)
,amr),mlt),(heb))),amh)),(kmk,(alb,(((tel,(knd,(tml,mym))),((igb
,(wlf,(((grk),bsq),(((chv,((tuv,ykt),((bsk,tvo),(((((krc),uzb),(
kkp)),uyg),(((tur),tkm),aze))))),(kor,nav)),((((((((((urd,(nep,h
in)),guj),mhi),ben),kas),(krd,(taj,prs))),(arm,arw)),((lat,lit),
(ukr,(rus,((pol),((cze),((bul),(slo,(scr))))))))),((bre,wel),((i
ri,gae)))),((((((rom,ita),ctl),por),(spa)),(fre)),(((eng)),(((sw
e,(nor,dsh)),(ice)),(((dut)),(((ger)))))))))))))),((qaf,(som)),(
che,((kab,cct),(chc,(lak,ava))))))))))))))))),((bam,(fni,(geo)))
,(knr,(hun,(moe,(((udm),mme),(fin,est))))))),jpn),grw),((((swa,s
hn),ndo),(lda,kin)),(xho,zul)));
```

We use the “Symmetric Difference Distance of Robinson and Foulds” measure to compute the distance since we could not find a proper World Tree with branch sores available. The Symmetric Difference is a count of how many partitions there are, among the two trees, that are on one tree and not on the other [14][18].



We normalize our distance measure to calculate the similarity between the trees. Similarity can be defined as:

$$Similarity = \frac{Distance\ Found}{Maximum\ Distance\ Possible}$$

We compare our tree with the above tree at high granularity (original trees) as well as slightly lower granularity (ignoring lowest level bifurcations). The latter was done so as to take care of errors which might have occurred during construction of the ASJP parenthesis tree. Following are some of our results:

Tree Source	**A**	**B**	**Similarity**	**Similarity (at lower granularity)**
WALS			22%	35%
ASJP			45%	50%
MIX	0.2	0.8	40%	43%
MIX	0.3	0.7	42%	44%
MIX	**0.4**	**0.6**	**48%**	**60%**
MIX	0.5	0.5	37%	40%
MIX	0.7	0.3	24%	27%

The World Tree has been added in APPENDIX B for further manual comparison.

MANUAL ANALYSIS

Our version might as well be in some aspects better than the ASJP World Tree we are comparing with. For instance, we see all Indian languages grouped together in our tree. In the ASJP Tree, the South Indian Languages are separated from the North Indian ones. Though they do differ a lot, our experience suggests that they should not be as far spaced as shown in the ASJP Tree. In our case, they are in different sub-trees, but they share a near common ancestor, which is closer to expectation. Our tree performs well for other Asian languages like Chinese, Japanese, Korean and Burmese, which also share a near ancestor. European languages also perform pretty well in our classification.

A major reason for the difference between our tree and the ASJP World Language tree might be the fact that the latter uses only Lexical features, i.e. the words in the languages.

CONCLUSION

A good mix of lexical and typological features is necessary to get a good tree. Lexical features seem to perform a little better than the typological ones. This might be because typological features have a lot of attributes missing or unknown. A good mix of both yields better results.

REFERENCES

[1] Nichols, J. and Warnow, T. (2008), Tutorial on Computational Linguistic Phylogeny. Language and Linguistics Compass, 2: 760–820. doi: 10.1111/j.1749-818X.2008.00082.x

[2] Schnoebelen, Tyler. "A how-to guide for using phylogenetic tools on linguistic data.(SplitsTree, MrBayes)." *Stanford University*. *Revised on April* 23 (2009): 2009.

[3] Holman, Eric W., et al. "Automated dating of the world's language families based on lexical similarity." *Current Anthropology* 52.6 (2011): 841-875.

[4] Serva, Maurizio, and Filippo Petroni. "Indo-European languages tree by Levenshtein distance." *EPL (Europhysics Letters)* 81.6 (2008): 68005.

[5] Swadesh List. http://en.wikipedia.org/wiki/Swadesh_list

[6] World Atlas of Language Structures. http://wals.info/

[7] Bakker, Dik, et al. "Adding typology to lexicostatistics: a combined approach to language classification." *Linguistic Typology* 13.1 (2009): 169-181.

[8]
http://www.academia.edu/1020464/A_Phylogenetic_Approach_to_Comparative_Linguistics_A_Test_Study_Using_the_Languages_of_Borneo

[9] Maddison, David R., David L. Swofford, and Wayne P. Maddison. "NEXUS: an extensible file format for systematic information." *Systematic Biology* 46.4 (1997): 590-621.

[10] **Recent Work in Computational Linguistic Phylogeny**
Joseph F. Eska and Don Ringe
Language , Vol. 80, No. 3 (Sep., 2004), pp. 569-582
Published by: Linguistic Society of America
Article Stable URL: http://www.jstor.org/stable/4489723

[11] Automated Similarity Judgement Program (ASJP) http://wwwstaff.eva.mpg.de/~wichmann/ASJPHomePage.htm

[12] List of Wikipedias. http://meta.wikimedia.org/wiki/List_of_Wikipedias

[13] Müller, André, et al. "ASJP World Language Tree of Lexical Similarity: Version 3 (July 2010)." (2010).

[14] PHYLIP TREEDIST http://evolution.genetics.washington.edu/phylip/doc/treedist.html

[15] PHYLIP PARS http://evolution.genetics.washington.edu/phylip/doc/pars.html

[16] PHYLIP FITCH http://evolution.genetics.washington.edu/phylip/doc/fitch.html

[17] PHYLIP DRAWGRAM http://evolution.genetics.washington.edu/phylip/doc/drawgram.html

[18] Distance Between Trees http://ibis.tau.ac.il/intro_bioinfo/MolEvolSlides/DistanceBetweenTrees.ppt