Project Approval

Project Title: Linguistic analysis of Indo-European Languages

Project Guide: Mr. Shreekanth M Prabhu

Project Team:

Roshan U	[01FB15ECS246],
Sanath Bhimsen	[01FB15ECS260],
Mukesh M Karanth	[01FB15ECS361].

Problem Statement

- The current model of the Indo-European languages is a predominant-tree like structure which implies that all languages developed strictly divergently with little frequency of borrowing.
- This might be a biased model due to the limited considerations and the restricted visualisation of the languages.
- We want to broaden the considerations by including possibilities of word transfers and mutual growth and come up with a better, more realistic network model of the Indo-European Languages.

Literature Survey

Indo-European languages

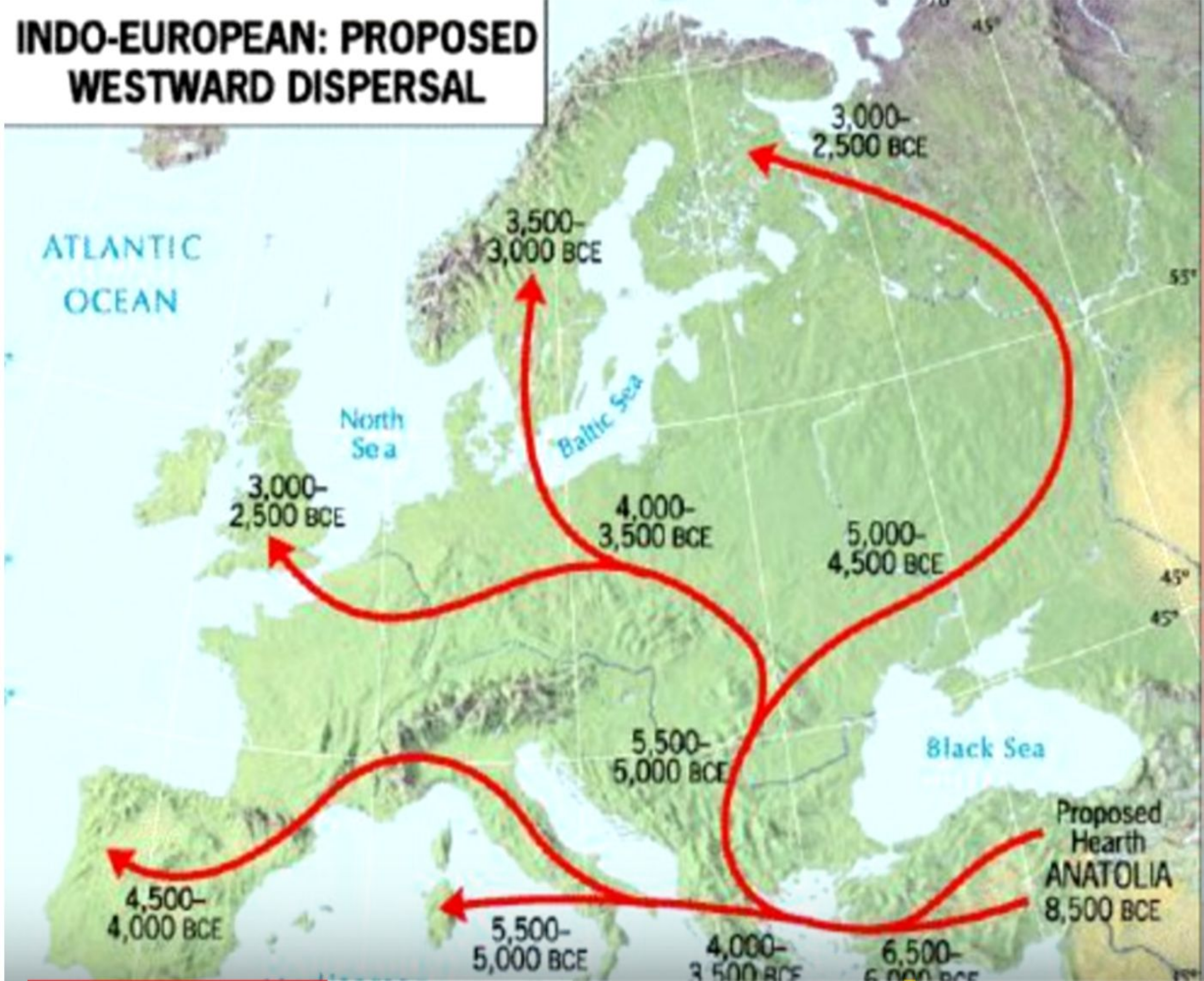
The Indo-European languages are a language family of several hundred related languages and dialects.

There are about 445 living Indo-European languages, according to the estimate by *Ethnologue*, with over two thirds (313) of them belonging to the Indo-Iranian branch. The most widely spoken Indo-European languages by native speakers are Hindustani (Hindi-Urdu), Spanish, English, Portuguese, Bengali, Punjabi, and Russian, each with over 100 million speakers, with German, French, Marathi, Italian, and Persian also having more than 50 million. Today, nearly 42% of the human population (3.2 billion) speaks an Indo-European language as a first language, by far the highest of any language family.

Background

Indo-European Languages developed from Proto Indo European language which was spoken about 6500 years ago.

Domestication of horses and agriculture were one of the key reasons which led to the migration of people from Europe to India and other countries thus leading to birth of languages which share a connection which each other.

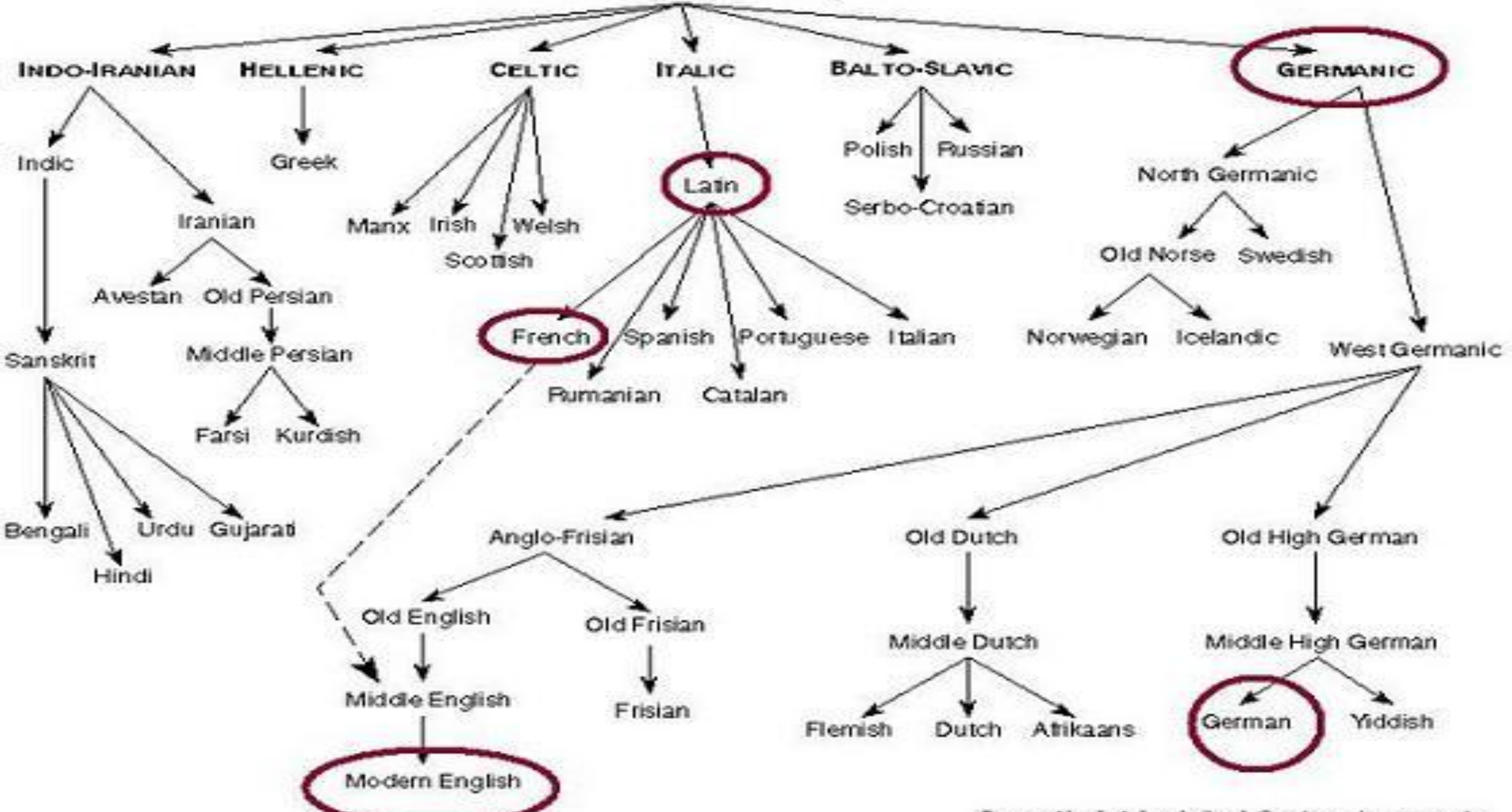


INDO-EUROPEAN: PROPOSED WESTWARD DISPERSAL

Alternate theory

PIE began several thousand years earlier in Anatolia, and spread with the expansion of agriculture.

Proto-Indo-European



INDO-IRANIAN
Indic
Sanskrit
Bengali
Hindi
Urdu
Gujarati
Iranian
Avestan
Old Persian
Middle Persian
Farsi
Kurdish

HELLENIC
Greek

CELTIC
Manx
Irish
Scottish
Welsh

ITALIC
Latin
French
Spanish
Portuguese
Italian
Rumanian
Catalan

BALTO-SLAVIC
Polish
Russian
Serbo-Croatian

GERMANIC
North Germanic
Old Norse
Swedish
Norwegian
Icelandic
West Germanic
Anglo-Frisian
Old English
Middle English
Modern English
Old Frisian
Frisian
Old Dutch
Middle Dutch
Flemish
Dutch
Afrikaans
Old High German
Middle High German
German
Yiddish

Prepared by Jack Lynch, jlynch@andromeda.rutgers.edu

Motivation

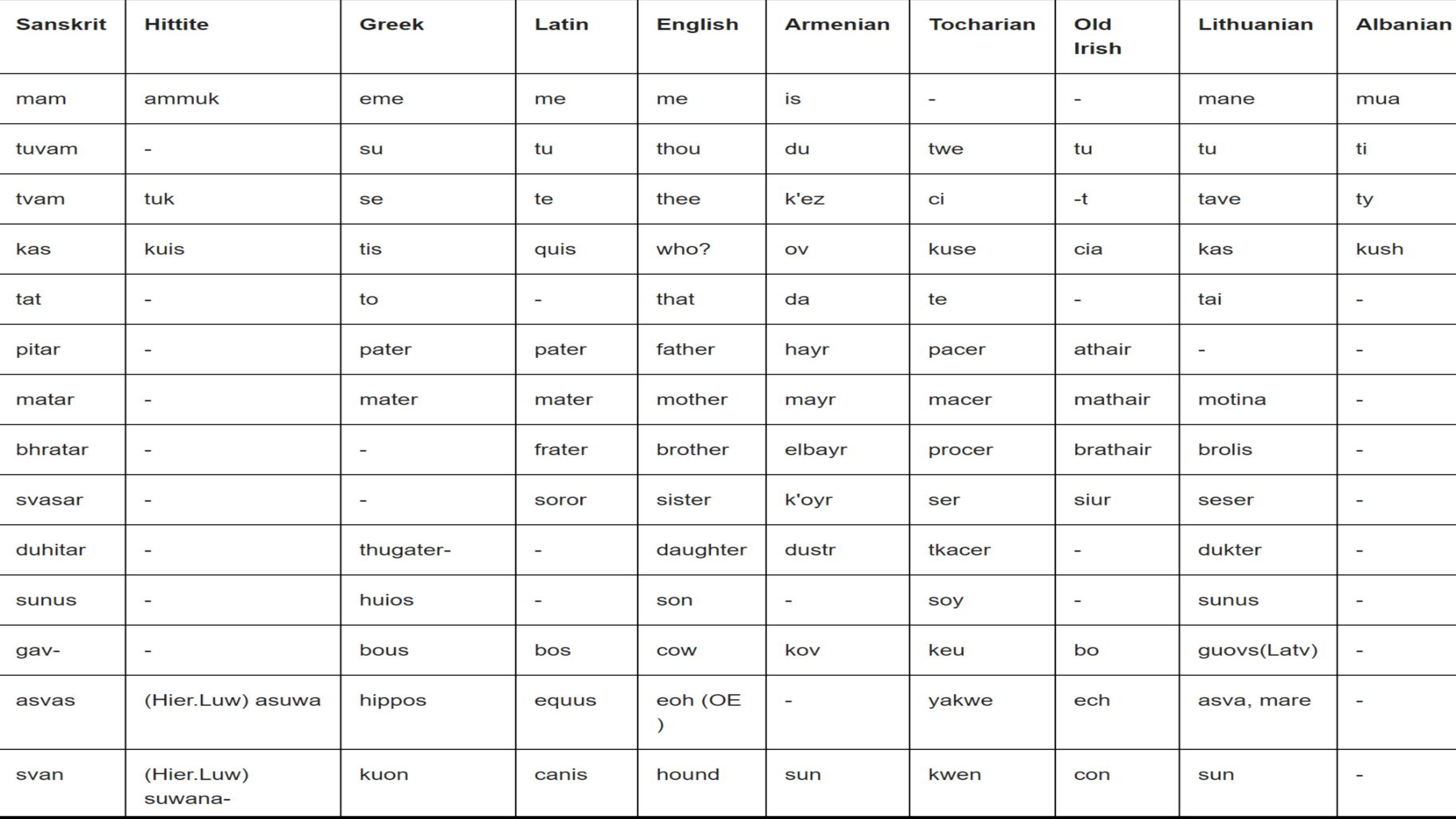
Now that we have an idea that indo-european languages share a connection, we intend to analyze the similarities betweens words of different Indo-European languages and see the degree to which they cognate by modeling the languages.

When modeling the languages, everyone assumes that the evolution of languages is strictly divergent and the frequency of borrowing is very low and or non-existent.

As consequence, the results suggest a predominantly tree-like pattern of the Indo-European language evolution.

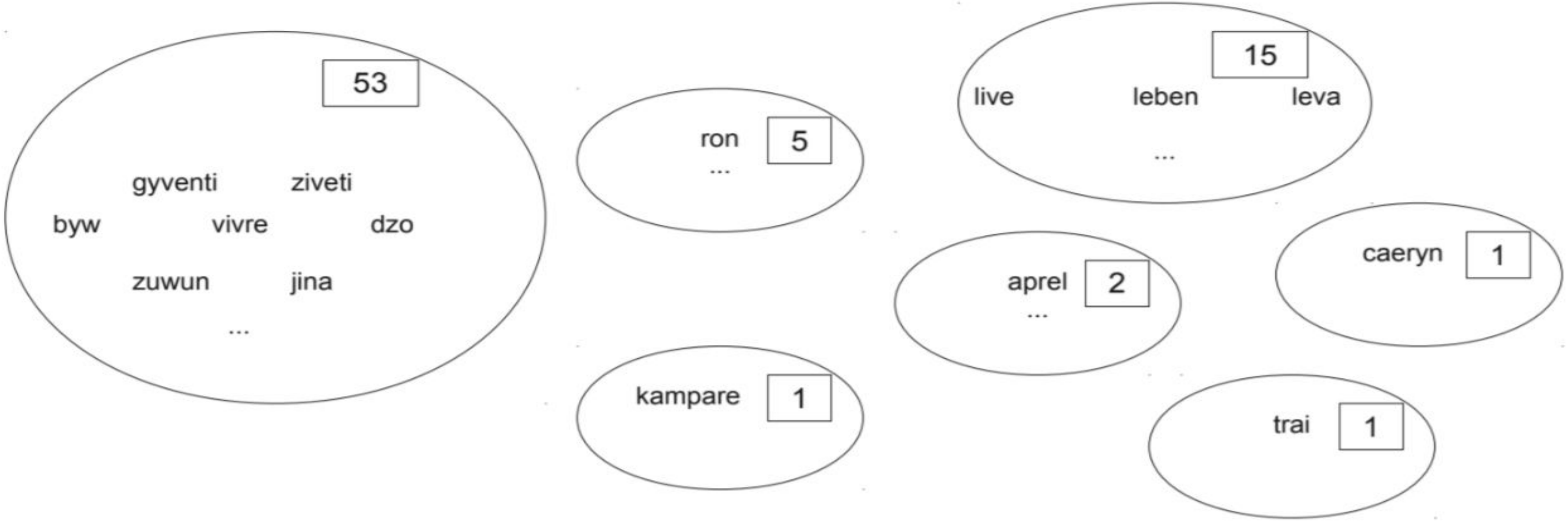
Hence, we want to model the Indo-European Languages as a network, using centrality measures and apply methods to estimate how close a language is to another language.

Sample Dataset



Sanskrit	Hittite	Greek	Latin	English	Armenian	Tocharian	Old Irish	Lithuanian	Albanian
mam	ammuk	eme	me	me	is	-	-	mane	mua
tuvam	-	su	tu	thou	du	twe	tu	tu	ti
tvam	tuk	se	te	thee	k'ez	ci	-t	tave	ty
kas	kuis	tis	quis	who?	ov	kuse	cia	kas	kush
tat	-	to	-	that	da	te	-	tai	-
pitar	-	pater	pater	father	hayr	pacer	athair	-	-
matar	-	mater	mater	mother	mayr	macer	mathair	motina	-
bhratar	-	-	frater	brother	elbayr	procer	brathair	brolis	-
svasar	-	-	soror	sister	k'oyr	ser	siur	seser	-
duhitar	-	thugater-	-	daughter	dustr	tkacer	-	dukter	-
sunus	-	huios	-	son	-	soy	-	sunus	-
gav-	-	bous	bos	cow	kov	keu	bo	guovs(Latv)	-
asvas	(Hier.Luw) asuwa	hippos	equus	eoh (OE)	-	yakwe	ech	asva, mare	-
svan	(Hier.Luw) suwana-	kuon	canis	hound	sun	kwen	con	sun	-

Cognate Clusters for the Word ‘Live’



Word similarity features(We intend to Consider to find cognates)

- Minimum edit distance
- The longest common prefix length
- Number of common bigrams
- The length of each word (2 separate features)
- The difference in length between the longer and the shorter word

Proposed Solution

- Collect the dataset of Indo-European languages from Langfocus website and other similar websites.
- Select a few key languages and preprocess the dataset for any discrepancies.
- Perform analysis on the dataset, by getting distance between languages by the closeness of their words using distance measures like Levenshtein distance, etc. and centrality measures.
- Understand and apply Horizontal Gene Transfer Detection Algorithm.
- Combine all the results and visualise the dataset to obtain a new and better model of the layout of Indo-European Languages.

Why our solution is better?

- It models the Indo-European languages in their true, unbiased states.
- The result is not a predominant tree like structure which is purely divergent.
- We get a complex network view of the languages with links to other languages due to word transfers which were previously disregarded.
- Overall improvement of the understanding of how the Indo-European languages came to be.

Tools to be Used (Considerations)

- iGraph tool in R:
    - Has good methods and features to perform network analysis and measures with ease.
    - Has a good visualisation functionality.
    - Language is user friendly.
    - Familiarity with the language and tool.

Project Timelines

1. Week 1 - Week 3:
   - Idea generation and approval from guide.
   - Understanding of the problem and development of problem statement.
   - Feasibility Study and Discussion with guide.
2. Week 4:
   - Discussion of the tools and technologies to be used.
   - Data Collection.
3. Week 5 - Week 7:
   - Development of the model and testing.
   - Fine Tuning the model and visualisations.
4. Week 8:
   - Optimisation.

Expected Contributions

- Literature survey: Mukesh M Karanth, Sanath Bhimsen.
- Data collection and preprocessing: Roshan U.
- Data modeling via similarity measures: Sanath Bhimsen.
- Data visualisation: Mukesh M Karanth.
- Testing & verifying results: Roshan U.
- Optimisations: Sanath Bhimsen, Mukesh M Karanth.
- Documentation: Sanath, Mukesh & Roshan.

References

1. Boc A, Di Sciullo AM, Makarenkov V. Classification of the Indo-European languages using a phylogenetic network approach. In: Locarek-Junge H, Weihs C, editors. Classification as a Tool for Research. Berlin Heidelberg: Springer; 2010. p. 647–55.
2. Shreekanth M Prabhu, Evolving a Framework to interpret the Vedas.
3. Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists http://www.aclweb.org/anthology/I11-1097

Thank You