|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| https://lh4.googleusercontent.com/proxy/YA9Xoqs7jhpeuwrEjwhdi_EVSCDwUdpr72V-2YHZ2lz2y1FaqityK8c8RlZRTvUDEw3Y2TekyGNi07wcREil5Ez3ii80dA-DE8G6HAQjEmJVz8W32Wy2uaDAWwuZs6uPZtJp2zrUJ_Qps2T1CUmSpuPR8dk2XA=w128-h144-k-no | | | **Document Ref.:** | | |  | |
| **Version No.:** | | | 1.0 | |
| **Date:** | | | **22-03-2019** | |
| **Copy No.:** | | | **1** | |
| **Project Name:** | Linguistic Analysis of Indo-European Languages | | | | | | |
| **Project Code:** | PW19SMP003 | | | | | | |
| **Status:** | ~~Draft~~ / **Current** / ~~Superceded~~ | | | | | | |
| **Document Type:** | **Controlled** / ~~Uncontrolled~~ | | | | | | |
| **Linguistic Analysis of Indo-European Languages** | | | | | | | |
| **This project is a research oriented project which deals with**  **linguistic analysis of Indo-European Languages using Social**  **Network analysis.** | | | | | | | |
| **Prepared By:** | | | | **Reviewed By:** | | | |
| **Name** | | **Date** | | **Name** | | | **Date** |
| Roshan U | | 22-03-2019 | |  | | |  |
|  | | |  |
| Sanath Bhimsen | | 22-03-2019 | | **Approved By:** | | | |
| **Name** | | | **Date** |
| Mukesh M Karanth | | 22-03-2019 | |  | | |  |
|  | | |  |
| **Distribution List** | | | | | | | |
| **Project Representative(s)** | | | | | **Guide Representative(s)** | | |
| 1. Roshan U 01FB15ECS246 2. Sanath Bhimsen 01FB15ECS260 3. Mukesh M Karanth 01FB15ECS361 | | | | | Prof.Shreekanth M Prabhu | | |

TABLE OF CONTENTS

[Definitions, Acronyms and Abbreviations 3](#_Toc3195224)

[References 3](#_Toc3195225)

[Change History 4](#_Toc3195226)

[1.0 Introduction 5](#_Toc3195227)

[1.1 Overview 5](#_Toc3195228)

[1.2 Scope 5](#_Toc3195229)

[1.3 Objective 5](#_Toc3195230)

[1.4 Outcomes 5](#_Toc3195231)

[2.0 Proposed Approach 5](#_Toc3195232)

[3.0 Interim Exploration and Results Chapters 5](#_Toc3195233)

[4.0 Further Exploration Plans and Timelines 5](#_Toc3195234)

Definitions, Acronyms and Abbreviations

* Customer Specification Requirements : **CSR**
* User Requirements Document : **URD**
* Indo-European : **IE**

References

1. Lethbridge, Timothy C., Janice Singer, and Andrew Forward. "How software engineers use documentation: The state of the practice." *IEEE software* 20.6 (2003): 35-39.
2. Briand, Lionel C. "Software documentation: how much is enough?." *Seventh European Conference onSoftware Maintenance and Reengineering, 2003. Proceedings.*. IEEE, 2003.
3. 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.
4. Shreekanth M Prabhu, Evolving a Framework to interpret the Vedas.
5. Clustering Semantically Equivalent Words into Cognate Sets in Multilingual Lists http://www.aclweb.org/anthology/I11-1097
6. The Origins of Indo-European Languages Colin Renfrew Scientific American Vol. 261, No. 4 (OCTOBER 1989), pp. 106-115
7. Mapping the Origins and Expansion of the Indo-European Language Family Remco Bouckaert, Philippe Lemey, Science 24 Aug 2012:Vol. 337, Issue 6097, pp. 957-960 DOI: 10.1126/science.1219669

Change History

This section describes the details of changes that have resulted in the current High-Level Design document.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Date** | **Document Version No.** | **Change Description** | **Reason for Change** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

# Introduction

## Overview

This document is designed for the purpose of understanding the additional explorations and research that has been done by the team in order to advance the project and to improve upon the fields mentioned in the previous reports and reviews.

It is to be used by the panel to understand what we have done so far and what we have explored and searched to obtain the intermediate progress we have achieved.

## Scope

The scope of the project is subject to the project being a minor project with heavy time constraints, hence we are making use of transliterated words, the number of words is limited to a maximum of 200 and we are using a select number of centrality and similarity

Measures like Closeness Centrality, Between-ness Centrality and similarity measures like Levenshtien Edit distance, cosine similarity etc.

## Objective

**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.

**What is Indo-European?**

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.

**Motivation**

Now that we have an idea that Indo-European languages share a connection, we intend to analyse the similarities between words of different Indo-European languages and see the degree to which they cognate by modelling the languages. When modelling 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.

## Outcomes

The outcome of the project will be considered fruitful or successful after we obtain a visualization for all the considered languages at hand, for their similarities and ties between each other; represented in the form of a network with distinct central language and words that link other languages together.

# Proposed Approach

* Collect the dataset of Indo-European languages from Langfocus website and other similar websites.
* Select a few key languages and pre-process 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.
* The data being used is a collective dataset that contains around 200+ words in English which comprise of stop words, nouns and verbs which are then transliterated into the other languages that we are using like French, Sanskrit, etc to use as our core database for evaluations and analytics. The words in English themselves have been obtained from ‘Langinfo’ website, and mostly with the help of Google Dictionary. The transliterated words have been solely obtained via Google Translate API.

# Interim Exploration and Results Chapters

* + - 1. Data Collection:
         1. After obtaining a list of all required English words, it was saved in excel and more columns for the other languages were created as null attributes.
         2. A python script was then created and in it google translate api was called for each of the words in the English column
         3. For each translation, its pronunciation was checked and appended if it exited, otherwise the vanilla translated word was appended to the respective column to which the word belonged to.
      2. Pre-Processing:
         1. Once the dataset was created, each word was replaced with its phonetic pronunciation using manual or automatic conversion wherever applicable.
         2. The Persian language was deleted and in its place Russian was added due to translation errors.
      3. Similarity Computations:
         1. A row from the processed dataset was taken and then converted into a word matrix
         2. Then, the edit distance between every two words in the matrix was computed and stored in the matrix
         3. Finally, a link was established between two words if their similarity exceeded a particular threshold
         4. The threshold was determined by exhaustive exploration.
         5. The results were put into a new dataset that had the word, its parent language and its cluster number
      4. Visualization:
         1. The resultant dataset from the computations were fed into a data frame and was converted into nodes of language and cluster.
         2. The nodes were then converted into a network and the underlying communities were highlighted to show the result of our computations.

# Further Exploration Plans and Timelines

* By the following reviews, we plan on implementing other similarity measures to compute cognates between words of different languages instead of just one measure.
* We plan on coming up with more visualizations that are more descriptive and less graphical
* We plan on improving the dataset by adding more words from nouns and verbs and getting a standard phonetic translation.

Timelines:

1. Advanced Similarity measures : 23rd & 24th March
2. Visualization : 30th & 31st March
3. Dataset improvement : 6th & 7th April