# Cognate Identification to improve Phylogenetic trees for Indian Languages

Diptesh Kanojia*
IITB-Monash Research Academy
IIT Bombay
Monash University
diptesh.kanojia@monash.edu

Malhar Kulkarni
IIT Bombay
Mumbai, India
malhar@iitb.ac.in

Pushpak Bhattacharyya
IIT Bombay
Mumbai, Maharashtra, India
pb@cse.iitb.ac.in

Gholamreza Haffari
Monash University
Melbourne, Victoria, Australia
gholamreza.haffari@monash.edu

## ABSTRACT

Cognates are present in multiple variants of the same text across different languages. Computational Phylogenetics uses algorithms and techniques to analyze these variants and infer phylogenetic trees for a hypothesized accurate representation based on the output of the computational algorithm used. In our work, we detect cognates among a few Indian languages namely Hindi, Marathi, Punjabi, and Sanskrit for helping build cognate sets for phylogenetic inference. Cognate detection helps phylogenetic inference by helping isolate diachronic sound changes and thus detect the words of a common origin. A cognate set manually annotated with the help of a lexicographer is generally used to automatically infer phylogenetic trees. Our work creates cognate sets of each language pair and infers phylogenetic trees based on a bayesian framework using the Maximum likelihood method.

We also implement our work to an online interface and infer phylogenetic trees based on automatically detected cognate sets. The online interface helps create phylogenetic trees based on the textual data provided as an input. It helps a lexicographer provide manual input of data, edit the data based on their expert opinion and eventually create phylogenetic trees based on various algorithms including our work on automatically creating cognate sets. We go on to discuss the nuances in detection cognates with respect to these Indian languages and also discuss the categorization of Cognate words *i.e., "Tatasama"* and *"Tadbhava"* words.

## CCS CONCEPTS

• **Computing methodologies → Information extraction**.

*This is the corresponding author


*CoDS-COMAD '19, January 3–5, 2019, Kolkata, India*

ACM ISBN 978-1-4503-6207-8/19/01…$15.00
https://doi.org/10.1145/3297001.3297045

## KEYWORDS

Natural Language Processing, Cognate Identification, Cognate Detection, Computational Phylogenetics, Phylogenetics, Phylogenetic Tree Generation, Indian Languages, Historical Linguistics

**ACM Reference Format:**
Diptesh Kanojia, Malhar Kulkarni, Pushpak Bhattacharyya, and Gholamreza Haffari. 2019. Cognate Identification to improve Phylogenetic trees for Indian Languages. In *6th ACM IKDD CoDS and 24th COMAD (CoDS-COMAD '19), January 3–5, 2019, Kolkata, India.* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3297001.3297045

## 1 INTRODUCTION AND MOTIVATION

Cognates are words derived from the same origin into one or more languages *i.e.,* they have the same etymological origin. The study of cognates plays a crucial role in applying comparative approaches for historical linguistics, in particular, solving language relatedness and tracking the interaction and evolvement of multiple languages over time. A cognate instance in Indian languages is given as the word group: *putra* (Sanskrit), *putra* (Hindi), *putra* (Marathi) and *puttar* (Punjabi), all of which mean the word "Son".

Previous studies on cognate detection try to distinguish between a pair of words help decipher whether a pair of words are cognates or non-cognates [3, 8]. These studies do not approach the problem of predicting the possible cognate of the target language, if the cognate of the source language is given. Cognate detection as a problem for searching cognates for a particular word in a wordlist has also been explored and applied it to the problem of ranking for information retrieval [12]. Identifying sound correspondences and cognates can be a costly process in terms of both time and cognitive load, requiring the expert knowledge of a lexicographer. Many langugaes recieve little attention due to the effort involved and many dialects are ignored from various studies due to the same reason. Finding automatic methods for performing or bootstrapping these processes would be a great benefit to historical linguists and has been a major motivation for research on cognate identification. Achieving good performance on automatic cognate identification can also benefit machine translation when dealing with two languages that share a certain quantity of cognates, as cognates are usually translations and serve as anchors when aligning. Cognates borrowed among Indian languages are categorized in two parts: *Tatasam* word and

*Tadbhava* words. Tatsama means "same as that" and Tadbhava means "arising from that". For *e.g.,* The Sanskrit word "*putra*" is borrowed "as-is" in Hindi and retians it's orthographic form in the word "*putra*", meaning "Son". In case of the Sanskrit word "*Satya*", the Hindi word takes an intermediary form first namely "*sacchh*", and later is borrowed in Hindi as "*sach*" meaning "Truth", in it's *Tadbhava* form.

## 2 RELATED WORK

Previous studies on cognate identification do not study Indian languages. Most of the Indian languages borrow cognates or "loan words" from Sanskrit. Indian languages like Hindi, Bengali, Sinhala, Oriya and even Dravidian languages like Malayalam, Tamil, Telugu and Kannada borrow many words from Sanskrit. Identification of Cognates for helping Information Retrieval has already been explored for Indian languages [9]. String similarity based methods are often used a baseline methods for cognate detection and the most commonly used among them is Edit distance based similarity measure. It is used as the baseline in the cognate detection papers [10]. This computes the number of operations required to transform from source to target cognate. We have also incorporated XDice [2], which is a set based similarity measure. Research in automatic cognate identification using phonetic aspects involve computation of similarity by decomposing phonetically transcribed words [7], acoustic models [11], phonetic encodings [13], aligned segments of transcribed phonemes [8]. We study Rama's research (2015), which employs a Siamese convolutional neural network to learn the phonetic features jointly with language relatedness for cognate identification, which was achieved through phoneme encodings. Although it performs well on accuracy, it shows poor results with MRR, possibly the reason as same as SVM performance. Papers related Orthographic cognate detection usually take alignment of substrings which in classifier like support vector machines [4, 5] or hidden markov models [1]. We also consider the method of Alina et al as the baseline (2014), which employs the dynamic programming based methods for sequence alignment. Among Cognate sets common overlap set measures like set intersection, Jaccard [6], XDice [2] or TF-IDF [14] could be used to measure similarities and validate the members of the set. **The key contribution of our work is:**

'*We create cognate sets for Indian language pairs and apply them for phylogenetic inferece on our tool inferece. We begin this pilot study in the detection of cognates for Sanskrit, Hindi, Marathi, and Punjabi for phylogenetic inference and hope to include other Indian languages in our dataset soon. We also release this dataset publicly.*'

## 3 EXPERIMENT DESIGN AND SETUP

### 3.1 Dataset Creation

We create the dataset by extracting word list for Hindi, Marathi, Sanskrit , and Punjabi WordNets. We transliterate the words in the Punjabi wordlist using Google Transliterate. We use the unique words from wordlist extracted from all the individual wordnet databases but maintain them within the ID space. We extract 15000 unique words from every wordnet and create an aligned wordlist for every language with Sanskrit, the pairs being Sanskrit - Hindi, Sanskrit - Marathi, and Sanskrit - Punjabi.

### 3.2 Setup

We design our dataset by first creating wordlists for every language pair involved. We extract unique words from wordnet data publicly available[1]. We align words from every language pair in a comma separated form for each concept ID thus ensuring a high probablity of detecting cognates. We, also, use the baseline measure XDice and string similarity based measures to first prepare cognate sets from every individiual language pair. Later, we construct more cognate sets with the use of Orthographic cognate detection methods such as alignment of substrings which uses support vector machines and hidden markov models. Among other methodologies, we use the phonetic aspects of the words decomposing them phonetically and aligning them according to phonemes. For validating our cognate sets, we use string similarity measures and use the threshold value of 0.75 arrived at by empirical measures. We use Jaccard, XDice and TF-IDF are used to validate our cognate sets.

We also implement our work with Textual History Tool[2] and verify the impact of our cognate sets on the creation of phylogenetic trees. The tool was created to facilitate the input of manuscript data and its variants digitally, and facilitates the creation of phylogenetic trees based on Maximum Likelihood and String similarity based measures. We verify that inducing cognate words along with the manuscript variants indeed helps in the creation of better phylogenetic trees. In the Textual History Tool, the phylogenetic tree creation mode allows a lexicographer to choose the variants they want to use to build a phylogenetic tree and provides the functionality of building the tree using various methods. The tree mode allows a lexicographer to use distance matrix based methods to generate baseline phylogenetic trees. The tree mode also allows a lexicographer to move the nodes manually if they find the output to be inaccurate based on the gold data. It also allows one to save the ouput as both an image and a PDF file. We present the screenshots of Textual history tool in figures 1 and 2.

## 4 DISCUSSION

During the validation of cognate sets created by various measures, we decided the threshold of matching at 0.75 for a pair to be cognate words. While arriving at this value, we observed that we could easily form pairs of cognate words which are *Tatsama* words. On the other hand, *Tadbhava* words were hardly detected among the cognate words unless phonetic methodologies were not used. This poses a new challenge as *Tadbhava* word form a large set of cognate words among the Indian languages. This can also be verified intuitively as the former retain their orthographic form and are easy to detect via the string similarity measure and the orthographic measure but the latter need phonetic measures.

## 5 CONCLUSION AND FUTURE WORK

We describe our work on cognate detection for Indian language pairs Sanskrit - Hindi, Sanskrit - Marathi, and Sanskrit - Punjabi. We create a wordlist of 15000 unique words from every individual

[1] http://www.cfilt.iitb.ac.in/indowordnet/
[2] http://www.cfilt.iitb.ac.in/~yogyata/4/admin/login.php

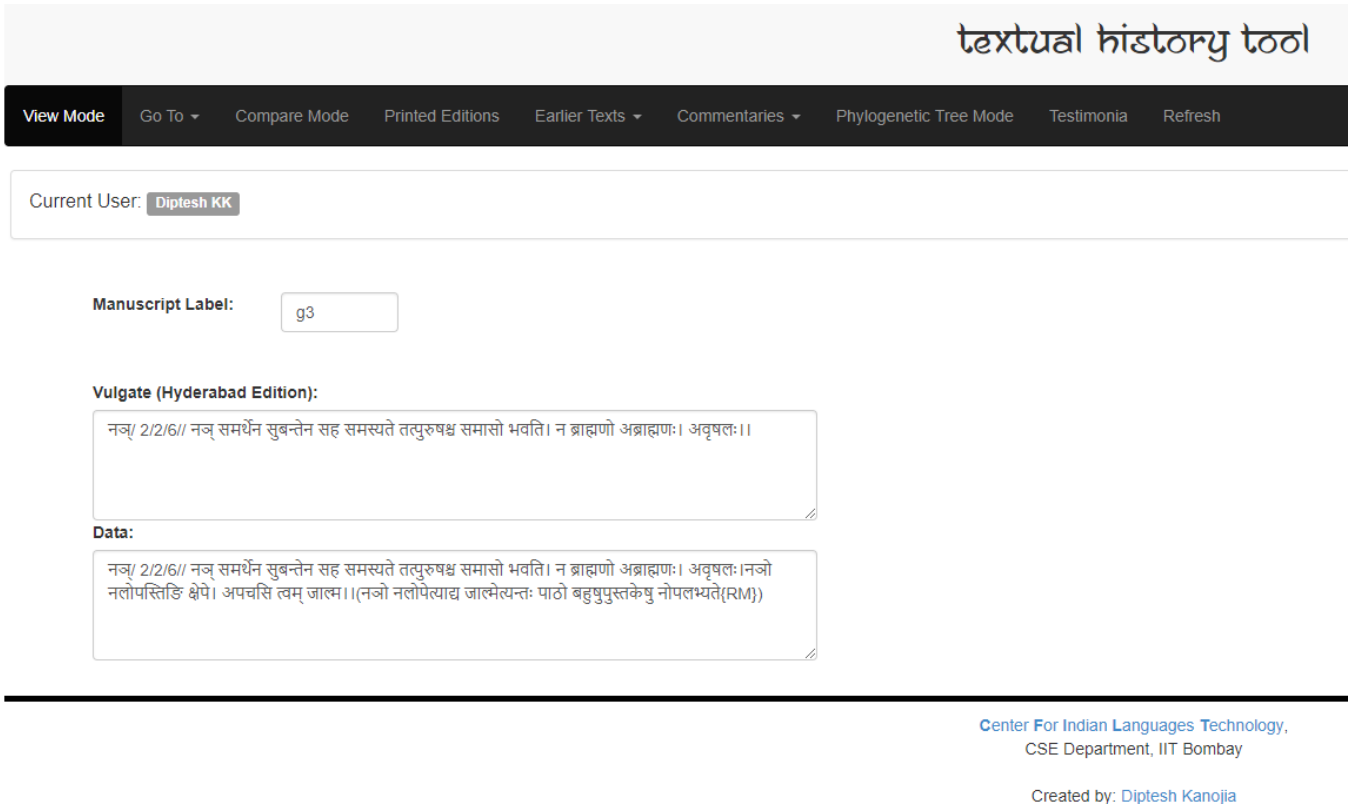


**Figure 1: Screenshot 1**

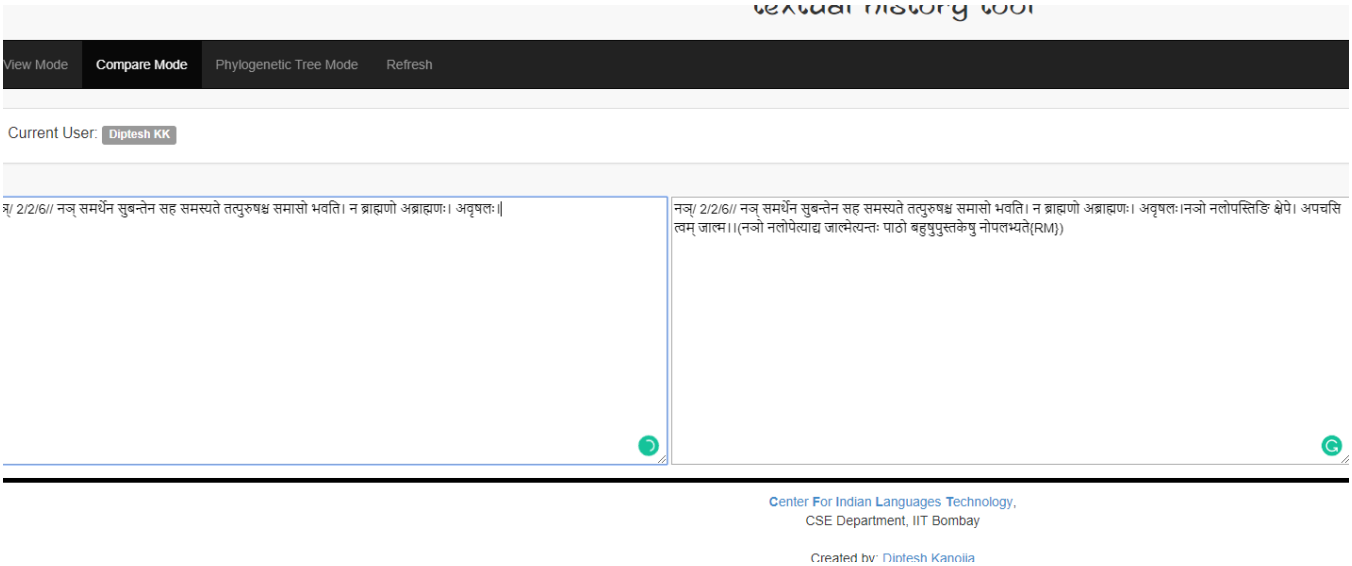


**Figure 2: Screenshot 1**

wordnet data and also create cognate word sets for these three language sets; and create the Textual History Tool. In the phylogenetic tree creation mode, we verify that cognate sets help in better phylogenetic tree creation. We also release this cognate set dataset publicly. In this pilot study, we create cognate categorization and the nuances of cognate detection for Indian languages.

In future, we aim to expand our dataset to multiple Indian languages as wordlists in their root form are available publicly via the Indowordnet website. We also aim to experiment with corpus instead of wordlists in their root form as morphological inflection would be a tougher challenge to tackle for detection of cognates in a corpus.

## REFERENCES


[1] Aditya Bhargava and Grzegorz Kondrak. 2009. Multiple word alignment with profile hidden Markov models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*. Association for Computational Linguistics, 43–48.

[2] Chris Brew, David McKelvie, et al. 1996. Word-pair extraction for lexicography. In *Proceedings of the 2nd International Conference on New Methods in Language Processing*. 45–55.

[3] Alina Maria Ciobanu, Anca Dinu, and Liviu Dinu. 2014. Predicting Romanian Stress Assignment. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*. 64–68.

[4] Alina Maria Ciobanu and Liviu P Dinu. 2014. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 99–105.

[5] Alina Maria Ciobanu and Liviu P Dinu. 2015. Automatic discrimination between cognates and borrowings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2. 431–437.

[6] Anni Järvelin, Antti Järvelin, and Kalervo Järvelin. 2007. s-grams: Defining generalized n-grams for information retrieval. *Information Processing & Management* 43, 4 (2007), 1005–1019.

[7] Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Association for Computational Linguistics, 288–295.

[8] Johann-Mattis List. 2012. LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Association for Computational Linguistics, 117–125.

[9] Ranbeer Makin, Nikita Pandey, Prasad Pingali, and Vasudeva Varma. 2007. Approximate string matching techniques for effective CLIR among Indian languages. In *International Workshop on Fuzzy Logic and Applications*. Springer, 430–437.

[10] I Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics* 25, 1 (1999), 107–130.

[11] Michelle M Mielke, Rosebud O Roberts, Rodolfo Savica, Ruth Cha, Dina I Drubach, Teresa Christianson, Vernon S Pankratz, Yonas E Geda, Mary M Machulda, Robert J Ivnik, et al. 2012. Assessing the temporal relationship between cognition and gait: slow gait predicts cognitive decline in the Mayo Clinic Study of Aging. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 68, 8 (2012), 929–937.

[12] A Pranav. 2018. Alignment Analysis of Sequential Segmentation of Lexicons to Improve Automatic Cognate Detection. In *Proceedings of ACL 2018, Student Research Workshop*. 134–140.

[13] Taraka Rama, Lars Borin, GK Mikros, and J Macutek. 2015. Comparative evaluation of string similarity measures for automatic language classification.

[14] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)* 26, 3 (2008), 13.