Cognate Identification to Improve Phylogenetics for Indian Languages

Diptesh Kanojia[1,2,3], **Pushpak Bhattacharyya**[1], **Malhar Kulkarni**[1], **Gholamreza Haffari**[1,2]

[1]IIT Bombay, [2]Monash University,
[3]IITB-Monash Research Academy,
{diptesh, pb}@cse.iitb.ac.in
malhar@iitb.ac.in
gholamreza.haffari@monash.edu

Introduction and Motivation

- Cognates are words derived from the same origin into one or more languages *i.e.,* they have the same etymological origin.
- 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.
- 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.
- Cognate detection helps phylogenetic inference by helping isolate diachronic sound changes and thus detect the words of a common origin.
- 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.
- 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".

Contributions

- We detect cognates among a few Indian languages namely Hindi, Marathi, Punjabi, and Sanskrit for helping build cognate sets for phylogenetic inference.
- 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.

Background and 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 borrow many words from Sanskrit.
- String similarity based methods are used as the baseline in the cognate detection papers (Melamed, 1999). We have also incorporated XDice (Brew et al., 1996), which is a set based similarity measure.
- Research in automatic cognate identification using phonetic aspects involve computation of similarity by decomposing phonetically transcribed words (Kondrak, 2000), acoustic models (Mielke et al., 2012), phonetic encodings (Rama et al., 2015), aligned segments of transcribed phonemes (List, 2012).
- IndoWordNet (Bhattacharyya, 2010) is a linked wordnet comprising of wordnets for major Indian languages listed in Table 1.
- These wordnets have been created using the expansion approach with Hindi WordNet as a pivot, which is partially linked to English WordNet.

Methodology

Cognate Identification

- 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 the wordlist extracted from all the individual wordnet databases publicly available[1], but maintain them within the ID space.
- We use the baseline measure XDice and string similarity based measures to first prepare cognate sets from every individiual language pair and show the results in Table 2.
- We construct more cognate sets with the use of Orthographic cognate detection methods such as alignment of substrings.
- We use the phonetic aspects of the words decomposing them phonetically and aligning them according to phonemes.
- 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.

Statistics and Results

- These wordnets have, on an average, approximately 32,000 synsets, with Marathi and Hindi having the minimum and the maximum number of synsets respectively.
- The number of synsets in Hindi is maximum due to the fact that work on IndoWordNet started with the Hindi language.

	Noun	Verb	Adjectives	Adverbs	Total
Hindi	29807	3687	6336	541	40371
Marathi	23271	3146	5269	539	32226
Punjabi	23255	2836	5830	443	32364
Sanskrit	32385	1246	4006	265	37907

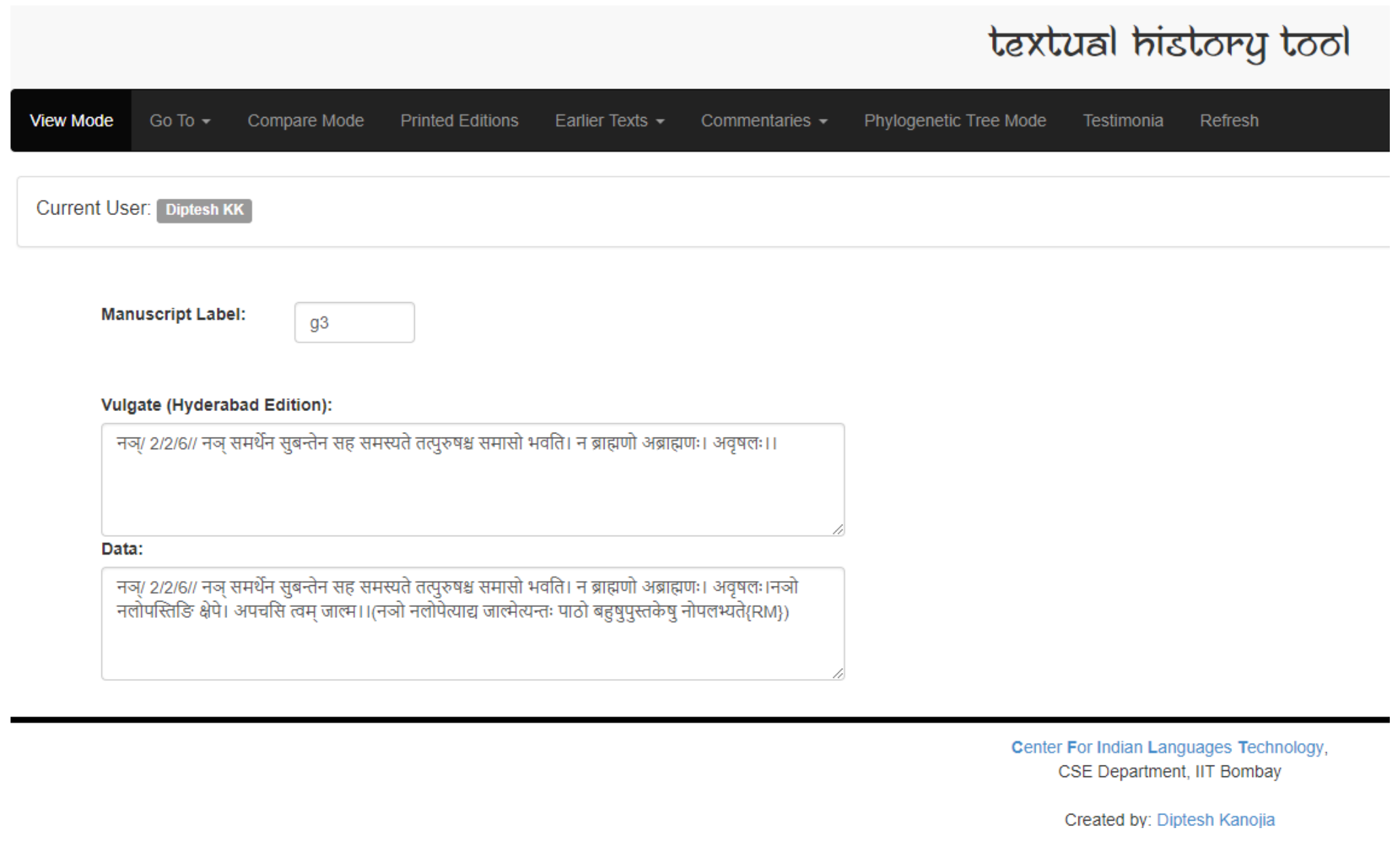
Table 1: *Number of synsets in different wordnets*

	Hindi-Punjabi	**Hindi-Marathi**	**Hindi-Sanskrit**
True Cognates	497	**621**	378
False Cognates	301	284	211

Table 2: *Number of True and False Cognates detected for each language pair*

Textual History Tool

- 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.



Current User: Diptesh KK

Manuscript Label: g3

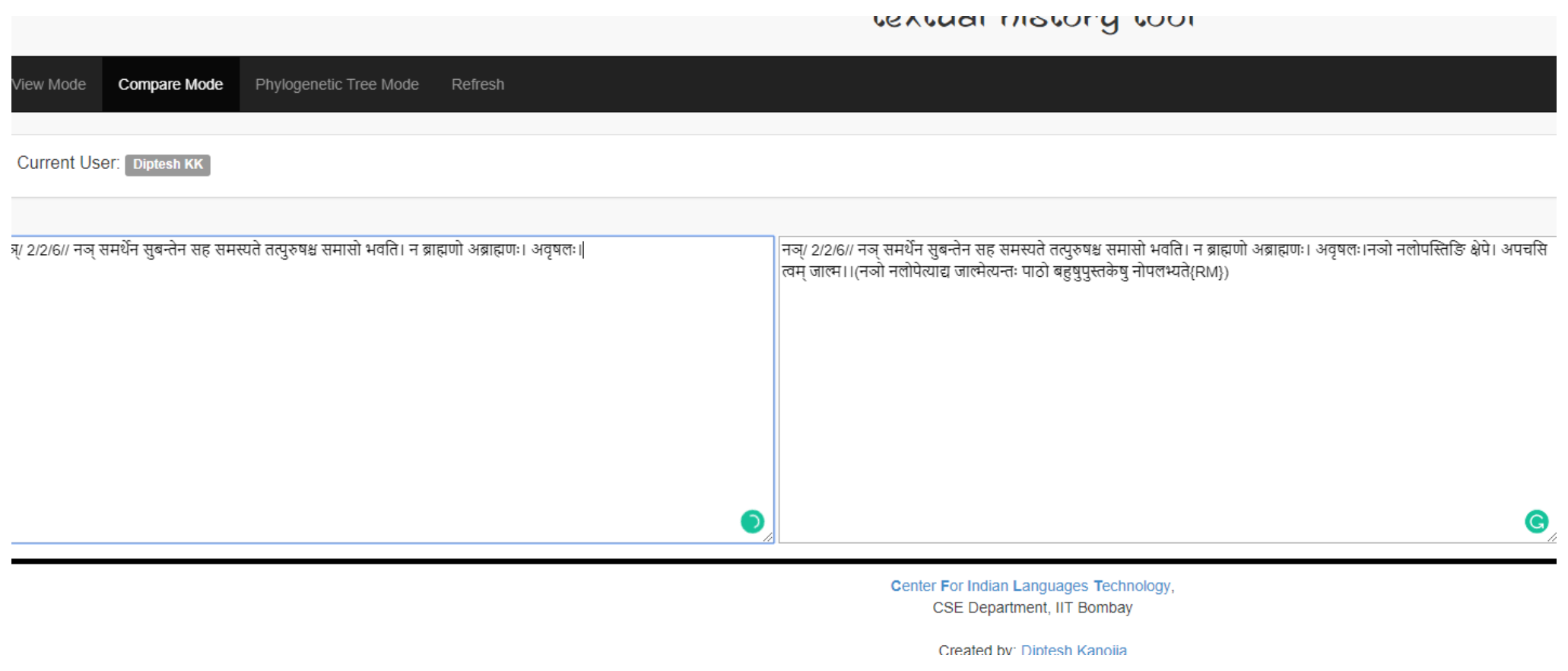
Vulgate (Hyderabad Edition):
नञ्/ 2/2/6// नञ् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। न ब्राह्मणो अब्राह्मणः। अवृषलः।।

Data:
नञ्/ 2/2/6// नञ् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। न ब्राह्मणो अब्राह्मणः। अवृषलः।नञो नलोपस्तिङि क्षेपे। अपचसि त्वम् जाल्म।।(नञो नलोपेत्याद्य जाल्मेत्यन्तः पाठो बहुषुपुस्तकेषु नोपलभ्यते{RM})

Center For Indian Languages Technology,
CSE Department, IIT Bombay

Created by: Diptesh Kanojia

Figure 1: *Textual History Tool: View Mode*



Current User: Diptesh KK

ञ्/ 2/2/6// नञ् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। न ब्राह्मणो अब्राह्मणः। अवृषलः।|

नञ्/ 2/2/6// नञ् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। न ब्राह्मणो अब्राह्मणः। अवृषलः।नञो नलोपस्तिङि क्षेपे। अपचसि त्वम् जाल्म।।(नञो नलोपेत्याद्य जाल्मेत्यन्तः पाठो बहुषुपुस्तकेषु नोपलभ्यते{RM})

Center For Indian Languages Technology,
CSE Department, IIT Bombay

Created by: Diptesh Kanojia

Figure 2: *Textutal History Tool: Compare Mode*

Important Insights

- 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.

Conclusion and Future Work

- We describe our work on cognate detection for Indian language pairs Sanskrit - Hindi, Sanskrit - Marathi, and Sanskrit - Punjabi.
- 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

Bhattacharyya, P. (2010). Indowordnet. In *In Proc. of LREC-10*. Citeseer.

Brew, C., McKelvie, D., et al. (1996). Word-pair extraction for lexicography. In *Proceedings of the 2nd International Conference on New Methods in Language Processing*, pages 45–55.

Kondrak, G. (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*, pages 288–295. Association for Computational Linguistics.

List, J.-M. (2012). Lexstat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125. Association for Computational Linguistics.

Melamed, I. D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

Mielke, M. M., Roberts, R. O., Savica, R., Cha, R., Drubach, D. I., Christianson, T., Pankratz, V. S., Geda, Y. E., Machulda, M. M., Ivnik, R. J., 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):929–937.

Rama, T., Borin, L., Mikros, G., and Macutek, J. (2015). Comparative evaluation of string similarity measures for automatic language classification.

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

Presented at The ACM India Joint International Conference on Data Science Management of Data CoDS-COMAD 2019, 3-5 January 2019, Kolkata (India)