**Discovering Cognates Using LSTM Networks**

Shantanu Kumar[[1]](#footnote-2), Ashwini Vaidya and Sumeet Agarwal

Indian Institute of Technology, Delhi

Cognates are words across different languages that are known to have a common ancestral origin. For example, the English word *Night* and the German *Nacht*, both meaning *night* are cognates with a common ancestral (Proto-Germanic) origin. Cognates are not always revealingly similar and can change substantially over time such that they do not share form similarity. Automatic cognate identification refers to the application of string similarity algorithms with machine learning to determine whether a given word pair is cognate or not. A cognate pair may have diverged at the surface level over time, but it shares a common ancestor and is likely to have similar meanings. This is especially true in languages that are typologically closer to each other.

Our system uses a character-level model with recurrent neural network architecture and attention. We test its performance on datasets drawn from three different language families. Our results show an improvement in performance as compared to existing models and highlights the usefulness of phonetic and conceptual features. We demonstrate the usefulness of our model by finding similar word pairs with high accuracy from a pair of closely related languages. This work contributes towards the development of technology by demonstrating that a deep learning model augmented with phonetic and conceptual features can discover cognates with high accuracy across diverse language families. Secondly, the discovery of string similarity is useful in improving the performance of NLP tasks like sentence alignment for machine translation. Further, if a language like Hindi has more extensive resources, then similarity information can be used to bootstrap lexical resource creation in a closely related language such as Punjabi or Marathi. This sharing of resources is relevant for languages in South Asia, which are diverse but share areal and typological properties.

**Keywords**: cognates, word similarity, language change, lexical resources

**Type of proposed chapter**: Research, 5000 words

**Corresponding author:** Ashwini Vaidya,avaidya@hss.iitd.ac.in

**Corresponding author’s bionote:** Dr. Ashwini Vaidya is an Assistant Professor in the Department of Humanities and Social Sciences, IIT Delhi. Her PhD was in Linguistics and Cognitive Science from the University of Colorado, Boulder. She was also a DST-CSRI post-doctoral fellow in IIT Delhi, and before that a DAAD Research scholar at the University of Konstanz, Germany. Her areas of research interest are computational linguistics and lexical semantics, particularly for South Asian languages. Her overall research goal is to build computational models for linguistic phenomena in a manner that is informed by theories of language and cognition.

1. At the time of writing this paper, Shantanu Kumar was a student at Indian Institute of Technology, Delhi. He is now working at Sizmek, USA [↑](#footnote-ref-2)