Searching for a Minimal Model of Spread of Lexemes across Languages

Keywords: Lexical similarity; Borrowing; Language change; Hierarchical clustering; Indic languages

Lexical elements tend to spread across languages and as a result the lexicon of a language may contain words of different genealogies whose parentage is difficult to trace on the synchronic level (Poplack and Sankoff, 1984). There are a vast number of words borrowed from Latin across languages and language families. But Latin has not been an extensive borrower of words from other languages. However, we can find multiple French-origin words in English and English-origin words in the contemporary French lexicon. This shows that there exists a disparity in donor-recipient relations which is obvious given the linguistic histories, areal and genetic factors, social implications, and so on to list a few.

If we consider the fact that the need for a concept is one of the motivations for languages to borrow (Campbell, 2020) there could be an access to illustrating the relations between languages in a linguistic area and capturing the languages carrying the potential of being lexical donors, borrowers or even languages resistant to participate in lexical transfer. We propose to achieve this through agglomerative hierarchical clustering methods. This stems from the technical unavailability of annotated lexical databases for low-resource languages that implicitly illustrate the donor languages for the individual lexical items.

A synchronic clustering of languages based on the contemporary vocabulary set will illustrate how the languages might cluster differently than the genetic phylogenies. This shift can be seen as a result of extensive borrowing where the languages now share words with genealogically unrelated languages and may cluster closer to the language group through which it received a higher flux of non-native vocabulary. For this study, we consider the languages spoken in the Indian subcontinent due to the rich linguistic diversity albeit the languages are low-resourced. We exploit CogNet (Batsuren et al., 2019) as our data resource and we consider the Indo-Aryan (Assamese, Bengali, Gujarati, Hindi, Kashmiri, Konkani, Nepali[1], Oriya, Punjabi, Sanskrit, Urdu), Dravidian (Kannada, Malayalam, Tamil, Telugu), and Sino-Tibetan (Bodo) languages.

We perform a synchronic clustering of languages to draw insights into the evolutionary pressures or conflicting signals created by the process of borrowing or lexeme transfer. This is an attempt to better understand how the borrowing process functions by extracting informative signals from the modern-day state of language lexicons due to the shared lexical concepts.

We achieve this through agglomerative hierarchical clustering and show that it is a promising data structure for visualizing lexeme transfers (Figure 1).

[1] We consider Nepali due to its orthographic similarity with Hindi and the geographic and genetic closeness to the Indo-Aryan languages primarily spoken in India.

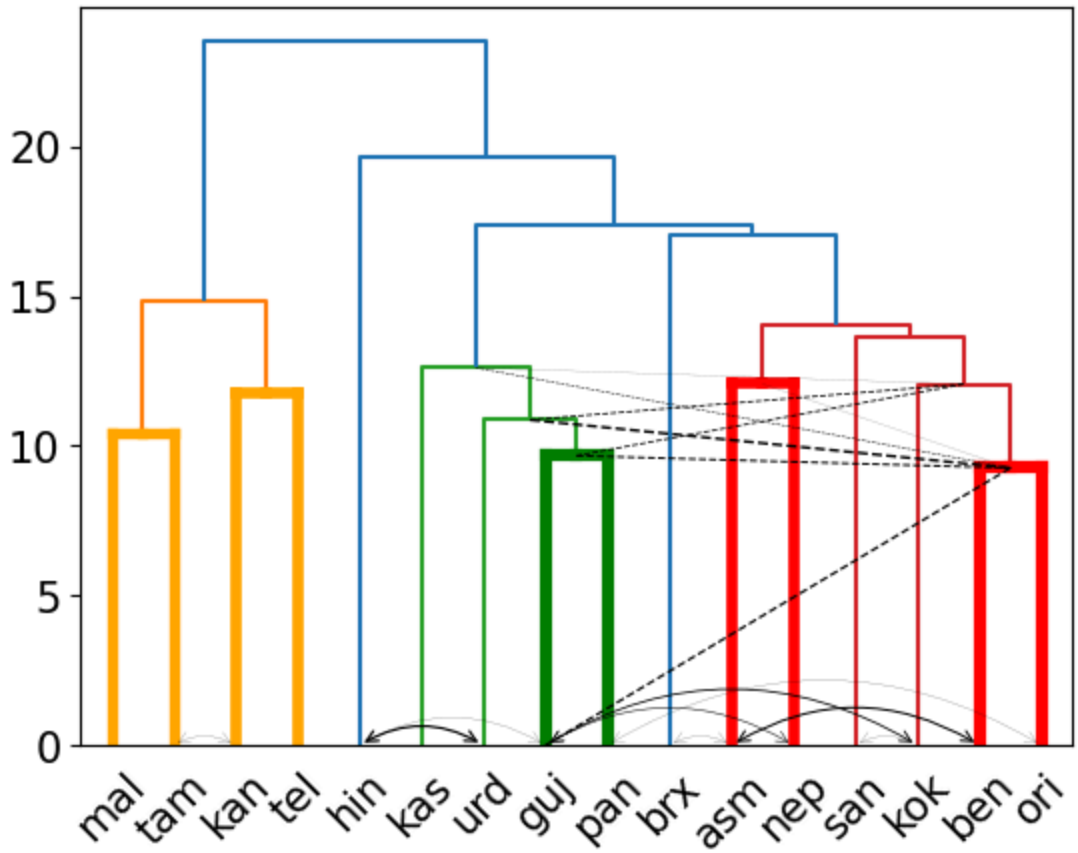

Figure 1: The dendrogram visualization showing the inheritance and the horizontal links. The language codes used are ISO 639-2.

References

Shana Poplack and David Sankoff. 1984. Borrowing: The synchrony of integration. Linguistics, 22(1):99–136.

Lyle Campbell. 2020. Historical Linguistics: An Introduction. Edinburgh University Press.

Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. 2019. CogNet: A large-scale cognate database. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3136–3145, Florence, Italy. Association for Computational Linguistics.