Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

An Extended Concept List of Vietnamese

Annika Tjuka
Department of Linguistic and Cultural Evolution
Max Planck Institute for Evolutionary Anthropology

As part of our ongoing endeavour to expand the possible mappings in Concepticon, I introduce an extended concept list of Vietnamese based on the Intercontinental Dictionary Series (IDS). The list includes elicitation glosses for 1,310 concepts that can be used as a reference to add more data or for comparative analyses. Here, I present the creation of the list and its content.

1 Introduction

The Concepticon includes 3,924 concept sets in its current version (v3.2, List et al. 2024, https://concepticon.clld.org/) and concept lists with elicitation glosses in a variety of languages such as English, Spanish, German and Russian are mapped to concept sets such as 906 TREE or 3626 KNOW. Two of the largest lists in terms of the number of concepts and languages, for which elicitation glosses are available, are the Intercontinental Dictionary Series (IDS, Key & Comrie 2023, https://ids.clld.org/) and the World Loanword Database (WOLD, Haspelmath & Tadmor 2009, https://wold.clld.org/). Due to their large scope, the lists are often used as a reference point to map glosses from other concept lists to their respective Concepticon concept set and they contribute to the decisions made by the automated mapping algorithm established in the cross-linguistic database of Norms, Ratings, and Relations of Words and Concepts (NoRaRe, Tjuka et al. 2022; 2023, https://norare.clld.org/). However, IDS and WOLD offer only partial coverage of Vietnamese glosses and the lists are not exhaustive.

Following List (2020), I used the 1,310 concepts from IDS (Key & Comrie 2023) as a starting point to build an elicitation list for Vietnamese. When deciding on the Vietnamese elicitation glosses for the concepts, I had to carefully consider both linguistic accuracy and context. I used two English-Vietnamese online dictionaries that contain not only translations but also example sentences and are commonly accepted in the Vietnamese community: The Free Vietnamese Dictionary Project (Duc et al. 2004,

Computer-Assisted Language Comparison in Practice
Volume 7, Number 1
URL: https://calc.hypotheses.org/7208 DOI: 10.15475/calcip.2024.1.6
Published under a Creative Commons Attributions 4.0 LICENSE
Published on 19/06/2024

https://www.informatik.uni-leipzig.de/~duc/Dict/) and the Laban Dictionary (https://dict.laban.vn/). This allowed me to ensure that each elicitation gloss reflects a given concept. To increase accuracy, I compared my glosses with the Vietnamese dictionary by Ilya Peiros provided in IDS v4.3. However, this list contained fewer items (774 elicitation glosses) and some glosses which I did not find in the dictionaries or for which I could not find context examples. I therefore added the missing glosses and sometimes used alternative glosses where it seemed appropriate. In addition, part of the glosses were double-checked with native Vietnamese speakers. However, not all elicitation glosses have been checked and the remaining errors are the sole responsibility of the author.

2 Structure and Content of the Extended Vietnamese Concept List

I aimed to streamline the elicitation glosses so that in most cases only one gloss is given for a corresponding concept. There are a few cases where this was not possible, for example, the concept SHOW corresponds to *cho xem* or *cho thấy* depending on the context. Examples (1) and (2) illustrate the different uses of the two word forms.

(1) Bức ảnh cho thấy cô ta mặc đồ đen.
The photo shows her dressed in black.
(2) Anh cho tôi xem những bức ảnh của anh.
He showed me his pictures.

Vietnamese also has a complex pronoun system that is more detailed than its English or German counterparts. The concept YOU (SINGULAR) corresponds to *anh* 'older brother', *chị* 'older sister', *ông* 'grandfather', *bà* 'grandmother', and other Vietnamese kinship terms used as personal pronouns (Sidnell & Shohet 2013). Vietnamese also has an extensive classifier system, for example, *con* for animals, *quả* for fruits and round things or *cái* for tools and furniture. I made an effort to use the classifiers consistently. For some glosses, I used commas and parentheses. Commas indicate that two glosses mean the same concept. For example, the concept DITCH can be expressed with *hào* or *mương*. Parentheses mean that the part in parentheses can also be omitted. It is often the case that a morpheme of a word is omitted from the sentence if the context allows it. For example, the concept KNEEL can be expressed with *quỳ xuống* or just *quỳ*.

The extended Vietnamese concept list has been published on Zenodo and the resource is freely accessible here: https://doi.org/10.5281/zenodo.11482543

3 Conclusion and Outlook

The concept list for Vietnamese will be integrated into Concepticon (List et al. 2016; Tjuka et al. 2023). This will broaden the scope of available mappings, offering researchers an enriched dataset for language comparison. By adding Vietnamese elicitation glosses to Concepticon, we are expanding the multilingual mapping of the resource and enabling an additional language for the automated mapping in NoRaRe (Tjuka et al. 2022; 2023). The next step would be to incorporate phonetic transcriptions and create a word list that can be integrated into Lexibank (List et al. 2022).

The approach presented here and in List (2020) can be used to add a more diverse set of glossing languages to Concepticon in the future.

References

Duc, Ho Ngoc & Free Vietnamese Dictionary Project. 2004. Free Vietnamese Dictionary Project. Leipzig: Leipzig University. https://www.informatik.uni-leipzig.de/~duc/Dict/install.html.

Haspelmath, Martin & Uri Tadmor. 2009. World Loanword Database. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org.

Key, Mary Ritchie & Comrie, Bernard (eds.) 2023. The Intercontinental Dictionary Series. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://ids.clld.org

List, Johann Mattis, Annika Tjuka, Mathilda van Zantwijk, Frederic Blum, Carlos Barrientos Ugarte, Christoph Rzymski, Simon Greenhill & Robert Forkel (eds.). 2024. Concepticon. A Resource for the Linking of Concept Lists (Version 3.2). Leipzig: Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org/.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch & Russell D. Gray. 2022. Lexibank, a Public Repository of Standardized Wordlists with Computed Phonological and Lexical Features. Scientific Data 9(1). 1-16. https://doi.org/10.1038/s41597-022-01432-0.

List, Johann-Mattis. 2020. Towards a refined wordlist of German in the Intercontinental Dictionary Series. Computer-Assisted Language Comparison in Practice 3(10). 1-2. https://calc.hypotheses.org/2545.

List, Johann-Mattis, Michael Cysouw & Robert Forkel. 2016. Concepticon: A Resource for the Linking of Concept Lists. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), Proceedings of the 10th International Conference on Language Resources and Evaluation, 2393–2400. Portorož, Slovenia: European Language Resources Association. https://aclanthology.org/L16-1379/.

Sidnell, Jack & Merav Shohet. 2013. The Problem of Peers in Vietnamese Interaction. Journal of the Royal Anthropological Institute 19(3). 618–638. https://doi.org/10.1111/1467-9655.12053.

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2023. Curating and Extending Data for Language Comparison in Concepticon and NoRaRe. Open Research Europe 2(141). 1–13. https://doi.org/10.12688/openreseurope.15380.3.

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2022. Linking Norms, Ratings, and Relations of Words and Concepts Across Multiple Language Varieties. Behavior Research Methods 54. 864–884. https://doi.org/10.3758/s13428-021-01650-1.

Supplementary Material
Tjuka, Annika (2024). Concept list of Vietnamese based on the Intercontinental Dictionary Series (IDS) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.11482543