Nidaba: Lexicon database and configurable segmental analysis tools

Eden, Elizabeth (University College London)

I present a new online database and lexical analysis program, Nidaba. Its core functions are the search and comparison of segmental patterns in transcribed lexicons. It has been designed for configurability in transcription and representation.

To analyse a lexicon with the program, two sets of input data are required: firstly, a list of lexical items in some transcription system, together with any data the researcher would like to tag items with (e.g. English gloss, part of speech, origin of loan items); secondly, a conversion from that system to the IPA. Each lexicon may be associated with multiple conversions, allowing researchers to compare the results of different levels of abstraction, different analyses of the same initial phonetic transcription, or different languages using a preferred convention.

The main analysis tools are based on searching for segmental patterns. Patterns of consonants and vowels are returned alongside the words in which they are found. If corpus frequency data is available, this tool can also give the total frequency of a pattern summed over all items, and similar statistics. From this basic overview, more detailed searches can be conducted. The researcher can specify properties of patterns such as length, number of items, frequency, or sonority profile; place, manner and/or voicing features; and part of speech, gloss or other lexical tags.

There are also set comparison tools. The results of the detailed searches can be automatically compared, making it easy to see which sequences occur e.g. word-initially but not word-finally; in nouns but not in verbs; or in high frequency items but not in numerous ones.

The results of the various searches can be viewed simultaneously across multiple languages, making it easy for researchers to compare inventories, syllable structure, reflexes of cognates, different phonemic analyses of a single lexicon, and more.

Nidaba is designed as an online database to enable better collaboration and typological studies, and is available at http://nidaba.co.uk. Data can be shared with specific users or publicly, with one-click sharing available for transitioning from an editing stage to public collaboration. So far, the database contains publicly accessible data from a wide variety of Indo-European and some Austronesian languages.

The online version presumes a segmental model with a typical set of binary features (e.g. Gussenhoven & Jacobs, Hayes, Odden). However, the source code is also available for more significant reconfigurations. By updating or replacing the representation module, any approach which can be represented as a unicode string can be used for analysis. This configurability could be integrated into the online user interface in a future version.

References

Gussenhoven, C. & Jacobs, H. 2005. *Understanding Phonology* (Understanding Language). Arnold.
Hayes, B. 2008. *Introductory phonology*. Blackwell.
Odden, D. 2005. *Introducing phonology*. Cambridge University Press.