ThoTDB: A Database of the Tonal Languages of the World[1]

Kirill Maslinsky (INALCO), Valentin Vydrin (INALCO-LLACAN)

The aim of the ThoTDB is to represent various types of data on tonal languages of the world for the broad audience of linguists.

There are two major types of information provided by the ThoTDB: (a) the distribution of tonal languages in the world, and (b) the parameters of tonal systems.

The ThoTDB is currently in the beta development stage and is available for users at https://thot.huma-num.fr/db.

Distribution of tonal languages in the world

The list of all the world languages is imported into ThoTDB from the Glottolog (Hammarström et al. 2024). For each language, its tonal status is indicated: tonal, toneless, unknown. In this respect, ThoTDB integrates all existing sources of data: databases compiled and kindly shared with us by Harald Hammerström and Larry Hyman, along with data from published databases: LapSyD (Maddieson et al. 2014-2016), and WALS (Maddieson 2013). These data have been complemented by our own research.

At the moment of writing, from the 7674 world languages included in the database 2140 are marked as tonal in ThoTDB, and 2742 as toneless. For 2792 languages the tonal status still remains unknown. Taking into account the distribution of tonal and toneless languages across families, we estimate that 42.5% of languages of the world are tonal. The work to verify tonal status of the languages continues.

Parameters of tonal systems

Out of the total number of 2140 tonal languages, we are planning to select up to 250 for our sample. The sample is intended to represent different genetic families of the world with the focus on typological and geographical diversity. The tonal systems of the languages included into the sample will be analyzed in detail, according to the structured standard model developed by ourselves.

The parameters of a tonal system represented in the database are the following:

- Type of tone-bearing unit (TBU) — syllable or mora.
- Number of tonal levels.
- Presence of downdrift and (non-automatic) downstep.
- Inventory of tonemes. For each toneme, the criteria that served to identify the toneme as a meaningful unit in the language is question are listed.
- The minimal and the maximal length of a tonal span (in TBUs).
- Correlation between the tonal span and segmental units: mora, syllable, prosodic foot, morpheme, word.
- Availability of toneless syllables and toneless morphemes.
- Availability of stress and its relation to tone.
- Tonal rules. Each rule is described both in a formalized form and verbally.
- Grammatical tones.
- Tonal classes of words.

For each language of the sample, deeply annotated texts will be provided. The markup conventions adopted for ThoTDB require to mark every toneme with the

1 The work on the database is supported by the ERC advanced project grant “Theory of Tone” to Valentin Vydrin.

boundaries of its tonal span, as well as boundaries of morphemes, words, syllables, morae (if relevant), and prosodic feet (if relevant). Texts annotated in this way allow to quantitatively estimate the parameters of the respective tonal systems in terms of relative frequency of units (boundaries), and their cross-correlations.

The following statistical indicators are automatically generated for each text:

- Syllabic Tonal Density Index (i.e., the number of tonemes per 100 syllables).
- Moraic Tonal Density Index (i.e., the number of tonemes per 100 morae).
- Morphemic Tonal Density Index (i.e., the number of tonemes per 100 morphemes).
- Podal Tonal Density Index (i.e., the number of tonemes per 100 prosodic feet, for the languages where the notion of prosodic feet is relevant).
- Tonal Density Index per word (i.e., the number of tonemes per 100 words).
- Share of occurrences of each toneme (per total the number of tonemes).
- The number of occurrences of grammatical tonal morphemes.
- An average length of a tonal span (in TBUs), in general and for each individual toneme.

All these data are projected on a zoomable map; a search combining different parameters will be possible.

Key terms and notions

Tonal level is a distinctive pitch range relevant for the tonal system.

Toneme is a meaningful tone, i.e. a tone which can (potentially) distinguish lexical and/or grammatical meanings. Identificitaion of tonemes in a language is based on a set of distributional **criteria** based on axiomatically defined toneme properties and principles for tonal processes. Criteria for the tonemic status are cross-linguistically standard.

Tonal span is a speech/text segment associated with a toneme on the surface level.

Tonal Density Index is a number of tonemes per 100 segmental units (primarily, syllables or morae; it can be also calculated with relation to other units, in particular: words, morphemes, prosodic feet).

Bibliography

Hammarström et al. 2024 — Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2024. Glottolog 5.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at https://glottolog.org/). DOI: 10.5281/zenodo.10804357

Maddieson 2013 — Maddieson, Ian. 2013. Tone (v2020.4). In Dryer, Matthew S. & Haspelmath, Martin (eds.). The world atlas of language structures online. Zenodo. https://doi.org/10.5281/zenodo.13950591.

Maddieson et al. 2014-2016. — Maddieson I., Flavier S., Marsico E., Pellegrino F., 2014-2016. LAPSyD: Lyon-Albuquerque Phonological Systems Databases, Version 1.0. https://lapsyd.huma-num.fr/lapsyd/.