5 Configurable Language-Specific Tokenization for CLDF Databases

Johannes Dellert[1], Verena Blaschke[2]

[1] Institute of Linguistics, University of Tübingen

[2] Center for Information and Language Processing, LMU Munich

In any workflow for computational historical linguistics, tokenization of IPA sequences is a crucial preprocessing step, as it shapes the alignments which provide the input of algorithms for cognate detection and proto-form reconstruction. This is also true for EtInEn (Dellert 2019), our forthcoming integrated development environment for etymological theories. An EtInEn project can be created from any CLDF database such as the ones that have been aggregated and unified by the Lexibank initiative (List ea. 2022). Whereas the tools for preparing CLDF databases (Forkel & List 2020) encourage the application of a uniform tokenization across all languages in a dataset, our view is that in many contexts, it is more natural to tokenize phonetic sequences in ways that differ between languages. To provide a simple example, many geminates in Italian need to be aligned to consonant clusters in other Romance languages (e.g. notte vs. Romanian noapte "night"), which is much easier if they are tokenized into two instances of the same consonant, whereas geminates in Swedish are best treated as cognate to their shortened counterparts in other Germanic languages.

To provide comprehensive support for such cases, EtInEn includes configurable language-specific tokenizers as an additional abstraction layer that allows to reshape forms after the import, and also serves as a generic way to bridge phonetic surface forms and the underlying forms that historical linguists are primarily interested in. Each tokenizer is defined by a token alphabet which is used for greedy tokenization, a list of allophone sets which can be used to abstract over irrelevant subphonemic distinctions, and a list of non-IPA symbols that are defined in terms of phonetic features. The initial state of each tokenizer is based on an analysis of the tokens used by the imported CLDF database. Tokenizer definitions are stored in a human-editable plain-text format which we would like to propose as a new standard.

In EtInEn, tokenizer definitions are manipulated through a graphical editor in which the potential tokens for each language are arranged in the familiar layout of consonant and vowel charts, enhanced by additional panels for diphthongs and tones. Currently defined tokens are highlighted, and allophone sets are summarized under their canonical symbols. Basic edit operations serve to group several sounds into an allophone set, and to join or split a multi-symbol sequence, such as a diphthong or a sound with a coarticulation. More complex operations support workflows for parallel configuration of multiple tokenizers.

Additional non-IPA symbols can be given semantics in terms of a combination of phonetic features, and declared to be part of the token set for any language. On the representational level, this provides the option to use non-IPA symbols for form display, whereas underlyingly, the system will interpret the symbols in terms of their features. On the conceptual level, underspecified definitions provide support for metasymbols. In addition to some predefined metasymbols (such as V for vowels and C for consonants), the user can assign additional symbols to arbitrary classes of sounds. These are then available throughout EtInEn for various purposes, such as concisely representing the conditioning environments for a soundlaw, or summarizing the probabilistic output of an automated reconstruction module.

In addition to configurable tokenizers, EtInEn provides the option to define form-specific tokenization overrides, allowing to substitute the result of automated tokenization with any sequence over the current token alphabet for the relevant language. This is currently our strategy for handling otherwise challenging phenomena such as metathesis or root-pattern morphology, which we normalize into alignable and concatenative representations. This forms a bridge to existing standards for representing morphology in the CLDF framework (e.g. Schweikhard & List 2020), which currently only support the annotation of morpheme boundaries in terms of simple splits in phonetic IPA sequences.

Dellert, Johannes (2019): "Interactive Etymological Inference via Statistical Relational Learning." Workshop on Computer-Assisted Language Comparison at SLE-2019.

Forkel, Robert and Johann-Mattis List (2020): "CLDFBench. Give your Cross-Linguistic data a lift." Proceedings of LREC 2020, 6997-7004.

List, Johann-Mattis, Robert Forkel, S. J. Greenhill, Christoph Rzymski, Johannes Englisch & Russell Gray (2022): "Lexibank, A public repository of standardized wordlists with computed phonological and lexical features." Scientific Data 9.316, 1-31.

Schweikhard, Nathanael E. and Johann-Mattis List (2020): “Developing an annotation framework for word formation processes in comparative linguistics.“ SKASE Journal of Theoretical Linguistics 17(1), 2-26.