A Web-Based Interactive Tool for Creating, Inspecting, Editing, and Publishing Etymological Datasets

Johann-Mattis List
Max Planck Institute for the Science of Human History
Kahlaische Straße 10
07745 Jena
list@shh.mpg.de

Abstract

The paper presents the Etymological DICtionary ediTOR (EDICTOR), a free, interactive, web-based tool designed to aid historical linguists in creating, editing, analysing, and publishing etymological datasets. The EDICTOR offers interactive solutions for important tasks in historical linguistics, including facilitated input and segmentation of phonetic transcriptions, quantitative and qualitative analyses of phonetic and morphological data, enhanced interfaces for cognate class assignment and multiple word alignment, and automated evaluation of regular sound correspondences. As a web-based tool written in JavaScript, the EDICTOR can be used in standard web browsers across all major platforms.

1 Introduction

The amount of large digitally available datasets for various language families is constantly increasing. In order to analyse these data, linguists turn more and more to automatic approaches. Phylogenetic methods from biology are now regularly used to create evolutionary trees of language families (Gray and Atkinson, 2003). Methods for the comparison of biological sequences have been adapted and allow to automatically search for cognate words in multilingual word lists (List, 2014) and to automatically align them (List, 2014). Complex workflows are used to search for deep genealogical signals between established language families (Jäger, 2015).

In contrast to the large arsenal of software for automatic analyses, the number of tools helping to *manually* prepare, edit, and correct lexical datasets in historical linguistics is extremely rare. This is surprising, since automatic approaches still lag behind expert analyses (List et al., 2017). Tools for data preparation and evaluation would allow experts to directly interact with computational approaches by manually checking and correcting their automatically produced results. Furthermore, since the majority of phylogenetic approaches makes use of manually submitted expert judgments (Gray and Atkinson, 2003), it seems indispensable to have tools which ease these tasks.

2 The EDICTOR Tool

The Etymological DICtionary ediTOR (EDICTOR) is a free, interactive, web-based tool that was specifically designed to serve as an interface between quantitative and qualitative tasks in historical linguistics. Inspired by powerful features of STARLING (Starostin, 2000) and RefLex (Segerer and Flavier, 2015), expanded by innovative new features, and based on a very simple data model that allows for a direct integration with quantitative software packages like LingPy (List and Forkel, 2016), the EDICTOR is a lightweight but powerful toolkit for computer-assisted applications in historical linguistics.

2.1 File Formats and Data Structure

The EDICTOR was designed as a lightweight file-based tool that takes a text file as input, allowing to modify and save it. The input format is a plain tab-separated value (TSV) file, with a *header* indicating the value of the columns. This format is essentially identical with the format used in LingPy. Although the EDICTOR accepts all regular TSV files as input, its primary target are *multi-lingual word lists*, that is, datasets in which a given number of *concepts* has been translated into a certain range of *target languages*.

Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017, pages 9–12

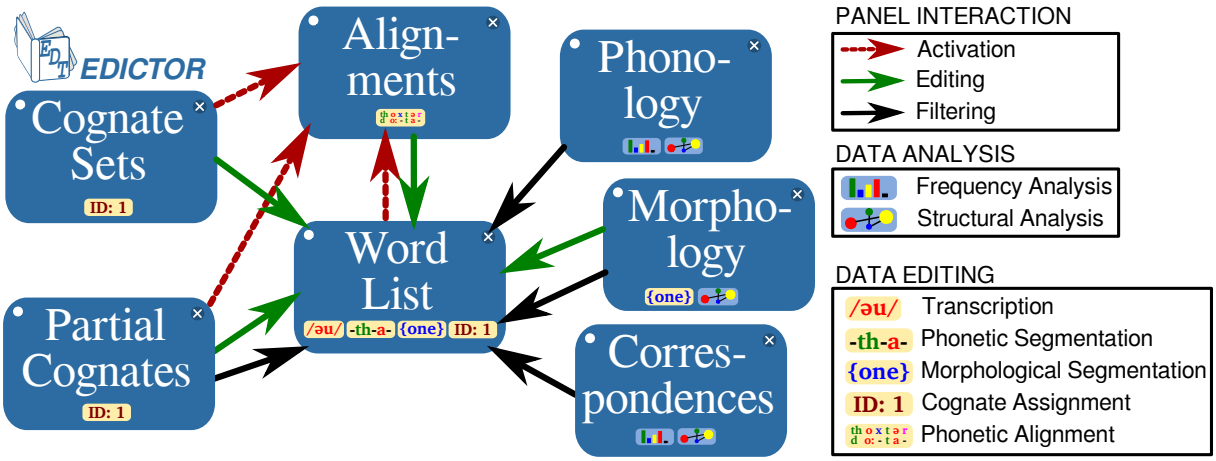

Figure 1: Basic panel structure of the EDICTOR.

ID	DOCULECT	CONCEPT	TRANSCRIPTION	...
1	German	Woldemort	valdəmar	...
2	English	Woldemort	wɔldəmɔrt	...
3	Chinese	Woldemort	fu^{51}ti^{51}mɔ^{35}	...
4	Russian	Woldemort	vladimir	...
...	...	...	...	...
10	German	Harry	haralt	...
11	English	Harry	hæri	...
12	Russian	Harry	gali	...
...	...	...	...	...

Figure 2: Basic file format in the EDICTOR

2.2 User Interface

The EDICTOR is divided into different *panels* which allow to edit or analyse the data in different ways. The core module is the *Word List panel* which displays the data in its original form and can be edited and analysed as one knows it from spreadsheet applications. For more complex tasks of data editing and analysis, such as cognate assignment or phonological analysis, additional panels are provided. Specific modes of interaction between the different panels allow for a flexible interaction between different tasks. Using drag-and-drop, users can arrange the panels individually or hide them completely. Figure 1 illustrates how the major panels of the EDICTOR interact with each other.

2.3 Technical Aspects

The EDICTOR application is written in plain JavaScript and was tested in Google Chrome, Firefox, and Safari across different operating systems (Windows, MacOS, Linux). For the purpose of offline usage, users can download the source code. For direct online usage, the tool can be accessed via its project website.

3 Data Editing in the EDICTOR

3.1 Editing Word List Data

Editing data in the Word List panel of the EDICTOR is straightforward by inserting values in text-fields which appear when clicking on a given field or when browsing the data using the arrow keys of the keyboard. Additional keyboard shortcuts allow for quick browsing. For specific data types, automatic operations are available which facilitate the input or test what the user inserts. Transcription, for example supports SAMPA-input. The segmentation of phonetic entries into meaningful sound units is also carried out automatically. Sound segments are highlighted with specific background colors based on their underlying sound class and sounds which are not recognized as valid IPA symbols are highlighted in warning colors (see the illustration in Figure 3). The users can decide themselves in which fields they wish to receive automatic support, and even Chinese input using an automatic Pīnyīn converter is provided.

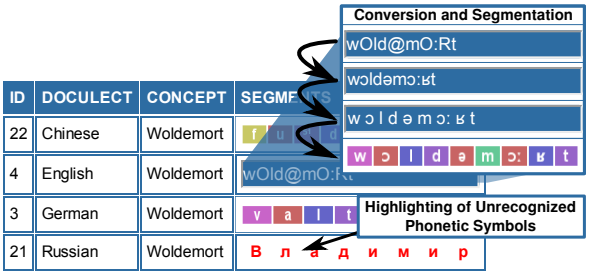

Figure 3: Editing word lists in the EDICTOR

3.2 Cognate Assessment

Defining which words in multilingual word lists are cognate is still a notoriously difficult task for machines (List, 2014). Given that the majority of datasets are based on manually edited cognate judgments, it is important to have tools which facilitate this task while at the same time controlling for typical errors. The EDICTOR offers two ways to edit cognate information, the first assuming complete cognacy of the words in their entirety, and the second allowing to assign only specific parts of words to the same cognate set. In order to carry out partial cognate assignment, the data needs to be morphologically segmented in a first stage, for example with help of the Morphology panel of the EDICTOR (see Section 4.2). For both tasks, simple and intuitive interfaces are offered which allow to browse through the data and to assign words to the same cognate set.

Figure 4: Aligning words in the EDICTOR

3.3 Phonetic Alignment

Since historical-comparative linguistics is essentially based on sequence comparison (List, 2014), alignment analyses, in which words are arranged in a matrix in such a way that corresponding sounds are placed in the same column, are underlying all cognate sets. Unfortunately they are rarely made explicit in classical etymological dictionaries. In order to increase explicitness, the EDICTOR offers an alignment panel. The alignment panel is essentially realized as a pop-up window showing the sounds of all sequences which belong to the same cognate set. Users can edit the alignments by moving sound segments with the mouse. Columns of the alignment which contain unalignable parts (like suffixes or prefixes) can be explicitly marked as such. In addition to manual alignments, the EDICTOR offers a simple alignment algorithm which can be used to pre-analyse the alignments. Figure 4 shows an example for the alignment of four fictive cognates in the EDICTOR.

4 Data Analysis in the EDICTOR

4.1 Analysing Phonetic Data

Errors are inevitable in large datasets, and this holds also and especially for phonetic transcriptions. Many errors, however, can be easily spotted by applying simple sanity checks to the data. A straightforward way to check the consistency of the phonetic transcriptions in a given dataset is provided in the Phonology panel of the EDICTOR. Here all sound segments which occur in the segmented transcriptions of one language are counted and automatically compared with an internal set of IPA segments. Counting the frequency of segments is very helpful to spot simple typing errors, since segments which occur only one time in the whole data are very likely to be errors. The internal segment inventory adds a structural perspective: If segments are found in the internal inventory, additional phonetic information (manner, place, etc.) is shown, if segments are missing, this is highlighted. The results can be viewed in tabular form and in form of a classical IPA chart.

4.2 Analysing Morphological Data

The majority of words in all languages consist of more than one morpheme. If historically related words differ regarding their morpheme structure, this poses great problems for automatic approaches to sequence comparison, since the algorithms usually compare words in their entirety. German *Großvater* 'grandfather', for example, is composed of two different morphemes, *groß* 'large' and *Vater* 'father'. In order to analyse multi-morphemic words historically, it is important to carry out a morphological annotation analysis. In order to ease this task, the Morphology panel of the EDICTOR offers a variety of straightforward operations by which morpheme structure can be annotated and analysed at the same time. The core idea behind all operations is a search for similar words or morphemes in the same language. These *colexifications* are then listed and displayed in form of a bipartite *word family network* in which words are linked to morphemes, as illustrated in Figure 5. The morphology analysis in the EDICTOR is no miracle cure for morpheme detection, and morpheme boundaries need still to be annotated by the user. However, the dynamically produced word family networks as well as the explicit listing of words sharing the same subsequence of sounds greatly facilitates this task.

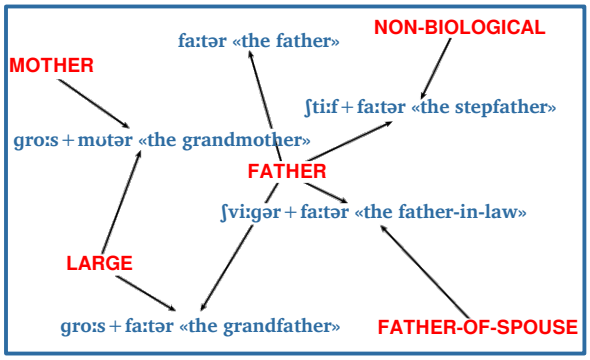

Figure 5: Word family network in the EDICTOR: The morphemes (in red) link the words around German *Großvater* 'grandfather' (in blue).

4.3 Analysing Sound Correspondences

Once cognate sets are identified and aligned, searching for regular sound correspondences in the data is a straightforward task. The Correspondences panel of the EDICTOR allows to analyse sound correspondence patterns across pairs of languages. In addition to a simple frequency count, however, *conditioning context* can be included in the analysis. Context is modeled as a separate string that provides abstract context symbols for each sound segment of a given word. This means essentially that context is handled as an additional *tier* of a sequence. This multi-tiered representation is very flexible and also allows to model suprasegmental context, like tone or stress. If users do not provide their own tiers, the EDICTOR employs a default context model which distinguishes consonants in syllable onsets from consonants in syllable offsets.

5 Customising the EDICTOR

The EDICTOR can be configured in multiple ways, be it while editing a dataset or before loading the data. The latter is handled via URL parameters passed to the URL that loads the application. In order to facilitate the customization procedure, a specific panel for customisation allows the users to define their default settings and creates a URL which users can bookmark to have quick access to their preferred settings.

The EDICTOR can be loaded in read-only mode by specifying a "publish" parameter. Additionally, server-side files can be directly loaded when loading the application. This makes it very simple and straightforward to use the EDICTOR to publish raw etymological datasets in a visually appealing format as can be seen from this exemplary URL: `http://edictor.digling.org?file=Tujia.tsv&publish=true&preview=500`.

6 Conclusion and Outlook

This paper presented a web-based tool for creating, inspecting, editing, and publishing etymological datasets. Although many aspects of the tool are still experimental, and many problems still need to be solved, I am confident that – even in its current form – the tool will be helpful for those working with etymological datasets. In the future, I will develop the tool further, both by adding more useful features and by increasing its consistency.

Acknowledgements

This research was supported by the DFG research fellowship grant 261553824 *Vertical and lateral aspects of Chinese dialect history*. I thank Nathan W. Hill, Guillaume Jacques, and Laurent Sagart for testing the prototype.

References

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439.

Gerhard Jäger. 2015. Support for linguistic macrofamilies from weighted alignment. *PNAS*, 112(41):1275212757.

Johann-Mattis List and Robert Forkel. 2016. *LingPy*. Max Planck Institute for the Science of Human History, Jena. URL: `http://lingpy.org`.

Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE*, 12(1):1–18, 01.

Johann-Mattis List. 2014. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf.

Guillaume Segerer and S. Flavier. 2015. *RefLex*. Laboratoire DDL, Paris and Lyon. URL: `http://reflex.cnrs.fr`.

Sergej Anatol'evič Starostin. 2000. *STARLING*. RGGU, Moscow. URL: `http://starling.rinet.ru`.

Supplementary Material

Supplements for this paper contain a demo video (`https://youtu.be/IyZuf6SmQM4`), the application website (`http://edictor.digling.org`), the source (v. 0.1, `https://zenodo.org/record/48834`), and the development version (`http://github.com/digling/edictor`).