Integrating grammatical description, text collection and dictionary: Language documentation and description for the digital age

The primary goal of GRAMR is to enable linguists to create interlinked digital representations of the three components of the "Boasian trilogy" – grammar, texts, and dictionary. It also incorporates elements of what has been called the "Himmelmannian" trilogy (Camp et al. 2018) – recordings, morphosyntactic annotation, and metadata. This is achieved with a web app written in the CLLD framework (Forkel et al. 2019), which allows for fast and user-friendly representation of various kinds of linguistic data. The application aims to close the gaps between linguistic documentation and description – the two processes are typically closely interwoven, but their respective end products often are not.

At the core of GRAMR is a collection of texts, which are sequenced into sentences and brought into the traditional interlinear example format, with a morpheme-by-morpheme glossing and a free translation, supplemented with audio. The second component is the grammatical description, which can contain interlinear examples taken directly from the text collection. This means that 1. examples can be looked up in the context in which they appear, 2. the audio can be played directly in the grammar, 3. segmentations and morphological annotations are always up to date. The morphemes in the object line of interlinear examples are all clickable, leading to the respective dictionary entry; links to individual morphemes can also be inserted into the grammatical description. The dictionary is at the moment the least feature-rich component of GRAMR, mainly serving as a link between texts and the grammar. There is a dictionary entry for every morpheme, and entries display all examples in which the morpheme occurs, again with links to the corpus. Entries for grammatical morphemes discussed in the grammatical description can also have back references to the relevant sections.

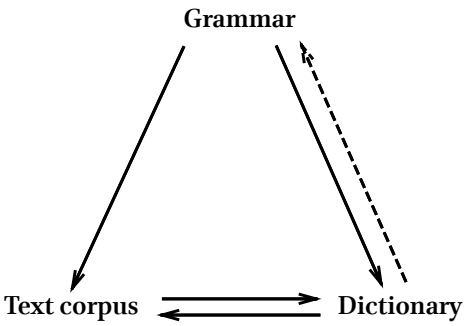
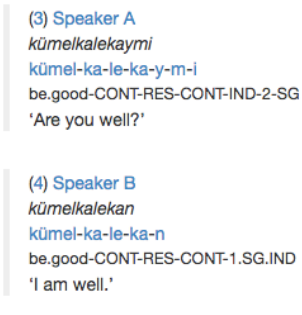


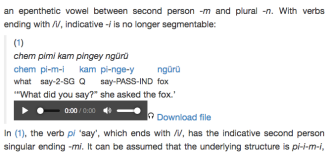
Figure 1: Links between the three components of GRAMR; the dashed line represents optional, non-automatic links.

As of now, data entry is organized around two staple tools of field linguists: ELAN (The Language Archive 2018) and FLEx (Summer Institute of Linguistics 2019). Texts are transcribed and translated in ELAN, then exported to FLEx, where they are parsed into morphemes and annotated. The text collection is exported from FLEx, including ELAN annotation time and speaker information, which enables automatic extraction of audio snippets and linking examples to speakers. From the FLEx export, CLDF (Forkel et al. 2017) representations of the text collection and the dictionary are created. Speaker and text metadata are provided via CSV table files. The input for the grammatical description is provided as a collection of plain text (markdown) files, along with a CSV file detailing the structure and titles of sections. Along with traditional markdown elements like tables, all text in the app

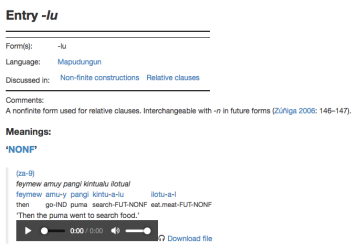
can contain GRAMR-specific markdown; for example `ex:X-2` will render the example with the ID `X-2`; `morph:X` will provide a link to the morpheme with the ID `X`. This is all the linguistic input needed for creating the data-rich, holistic representation of linguistic data and analysis that is GRAMR.



(a) Automatic speaker and morpheme links in texts.



(b) Interlinearized examples from the corpus can be used anywhere.



(c) Dictionary entry with references to grammar and examples from corpus.

Figure 2: Screenshots.

References

Camp, Amber B et al. (2018). "Writing Grammars of Endangered Languages". In: *The Oxford Handbook of Endangered Languages*. Kenneth Rehg & Lyle Campbell (eds.). Oxford: Oxford University Press: 271–304.

Forkel, Robert et al. (2017). *CLDF 1.0*. DOI: 10.5281/zenodo.1117644. URL: https://doi.org/10.5281/zenodo.1117644.

Forkel, Robert et al. (2019). "clld: a toolkit for cross-linguistic databases". DOI: 10.5281/zenodo.3239095. Online: https://doi.org/10.5281/zenodo.3239095.

Summer Institute of Linguistics (2019). *FieldWorks*. Computer program. URL: https://software.sil.org/fieldworks/.

The Language Archive (Apr. 4, 2018). *ELAN (Version 5.2)*. Computer program. Nijmegen. URL: https://tla.mpi.nl/tools/tla-tools/elan/.