Poio API[1]: a CLARIN-D curation project for language documentation and language typology

Peter Bouda, Centro Interdisciplinar de Documentação Linguística e Social, Minde/Portugal, pbouda@cidles.eu

Poio API is an open source software library written in Python and is being developed as part of a curation project within the working group “Linguistic Fieldwork, Anthropology, Language Typology” of CLARIN-D[2]. The goal of Poio API is to provide unified access to pivot data structures parsed from different file formats that researchers use in language documentation projects. As unified data structures we chose an implementation of the “Graph Annotation Framework” (GrAF) that was standardized as ISO 24612 in 2012 ([ISO 2012]). In our presentation, we will discuss the connections between GrAF and TEI, and present two use cases that demonstrate the innovation and advantage of our approach in comparison to existing methods.

Historically, GrAF was developed as a standoff version of the “Corpus Encoding Standard for XML” (XCES)[3], which is a TEI application for computational linguistics. In contrast to XCES, the development of the GrAF standard took place under the ISO umbrella from the beginning. Poio API uses an implementation of GrAF 1.0[4], for which the American National Corpus developed the schemata (Relax NG, W3C and DTD) with the TEI Roma program and ODD files, so that GrAF and TEI share at least the basic data types in its schema definitions. Furthermore, the feature structures of annotations in GrAF are TEI compliant, although GrAF itself is not a TEI compatible format. The main reason why we chose GrAF in our project is that the format is based on annotation graphs that are general enough to serve as pivot structures for the heterogeneous datasets in our research area ([Bird and Liberman 2001]). GrAF was explicitly developed as an underlying data model for linguistic annotations ([Ide and Suderman 2007]).

The first use case that we will present is the problem of file conversion and analysis of data in language documentation. Language documentation projects have collected a large amount of data, for example in the DoBeS corpus maintained by The Language Archive at the Max-Planck-Institute in Nijmegen[5]. The DoBeS corpus contains several different file formats for annotated video and audio files, as each project used its own combination of software tools (among others: Elan, Praat, Toolbox and FLEx), while applying different tier hierarchies and annotations schemes in their annotation workflows. As a result, cross-project queries in analysis is impracticable at the moment. To support ongoing activities in resolving this issue, Poio API makes it extremely easy to map between GrAF and arbitrary file formats. We will demonstrate our implementation of a generic converter class for a specific subset of GrAF annotation graphs, that maps the notions of „tiers“ and „annotations“ onto „nodes“ and „edges“ and vice versa. In the future, the same converter can support to read and write TEI compatible formats and build a bridge between GrAF and the TEI community.

1https://github.com/cidles/poio-api

2http://de.clarin.eu/en/discipline-specific-working-groups/wg-3-linguistic-fieldwork-anthropology-language-typology/curation-project-1.html

3http://www.xces.org/

4http://www.xces.org/ns/GrAF/1.0/

5http://tla.mpi.nl/

In addition, Poio API already contains a filter class for annotation graphs, so that users can search and analyse data sources as heterogeneous as in language documentation projects.

The second use case involves what we call “retro-digitization” of dictionaries. Starting with scanned PDF documents of several native South-American dictionaries, our workflow produces a set of GrAF files that contain the entries of each dictionaries as basic data (text file) and all information that we could extract from the dictionaries as standoff GrAF annotations. The latter consist of two types of annotations: the visible part of a dictionary, like “newline”, “tab” or “bold face”; and an interpretation regarding the meaning of the visible annotations, like “head word”, “translation” or “example sentence” (the workflow is described in more detail in [Bouda and Cysouw 2012]). The project uses Poio API and its GrAF implementation from the CLARIN curation project, but defines another subset of GrAF annotation graphs that is more suited for dictionary data. A very similar workflow is described in [Gómez Guinovart and Simões 2013], where the product is a TEI representation of the original dictionary. In our case we did decide not to publish TEI files as end product, as GrAF is much more flexible regarding the ad-hoc creation of additional annotation layers, the addition of links between random annotations and their later analysis. We will show a conversion of GrAF to a general network and how to apply graph-based methods on the result. Another important reason against TEI were the problems of embedded markup as discussed in [Cayless and Soroka 2010] and [Bański and Przepiórkowski 2009], so that we decided to primarily publish in a radical standoff format. We definitely plan to create a general converter from our subset of “dictionary GrAF” to a TEI representation for the near future and we hope to gain new insights on how to proceed through the discussions at the workshop.

References

[Bański and Przepiórkowski 2009] Bański, Piotr and Przepiórkowski, Adam. 2009. Stand-off TEI annotation: the case of the national corpus of polish. In: *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pp. 64–67, Singapore.

[Bird and Liberman 2001] Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Commun.* 33, 1-2 (January 2001), 23-60. DOI=10.1016/S0167-6393(00)00068-6 http://dx.doi.org/10.1016/S0167-6393(00)00068-6

[Bouda and Cysouw 2012] Bouda, Peter and Cysouw, Michael. 2012. Treating Dictionaries as a Linked-Data Corpus. In: *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, 1, 15–23.

[Cayless and Soroka 2010] Cayless, Hugh A. and Soroka, Adam. 2010. On implementing string-range() for TEI. In: *Proceedings of Balisage: The Markup Conference 2010* (URL: http://www.balisage.net/Proceedings/vol5/html/Cayless01/BalisageVol5-Cayless01.html, accessed 27.8.2012)

[Gómez Guinovart and Simões 2013] Gómez Guinovart, Xavier e Simões, Alberto. 2013. Retreading Dictionaries for the 21st Century. In: *2nd Symposium on Languages, Applications and Technologies*, pp. 115-126. OASIcs: Open Access Series in Informatics, vol. 29. Dagstuhl Publishing: Saarbrücken.

[Ide and Suderman 2007] Ide, Nancy and Suderman, Keith. 2007. GrAF: A graph-based format for linguistic annotations. In: *Proceedings of the Linguistic Annotation Workshop*, pp. 1–8, Prague, Czech Republic, June. Association for Computational Linguistics.

[ISO 2012] http://www.iso.org/iso/catalogue_detail.htm?csnumber=37326, accessed 21.3.2013