LIGT: TOWARDS AN ECOSYSTEM FOR MANAGING Interlinear Glossed Texts with Linguistic Linked Data Maxim Ionov

max.ionov@gmail.com University of Zaragoza, Spain

ligt:hasUtterances

(sub dc:Text)

ligt:hasMorphs

ligt:MorphTier

ligt:Word

ligt:Morph

ligt:item

Interlinear Glossed Text

- Interlinear Glossed Text (IGT) is a notation commonly used in typological and descriptive linguistics.
- It is an intuitive way to show language material without needing to know the language.
- Underdefined by design, a lot of variability.
- Annotation consists of a number of layers under the original text, e.g.:
- Morphological segmentation
- Grammatical meaning of morphemes
- Free translation
- me=köjle-s-um. leruuče lalime-le hunter sledge-acc pf=break-caus-tr.3sg 'The hunter broke the sledge.' Tundra Yukagir (Schmalz, 2013, p.66)

Linked Data Publication and Ligt

- · Linked Data is a set of best practices for publishing and connecting structured data on the Web using open standards:
- URIs to identify entities
- Accessible via HTTP
- Structured descriptions with RDF
- Linking related resources via URIs
- Data and annotations represented as Linked Data form a multigraph.
- Annotations as Linked Data are objects, not strings.
- Dataset interoperability due to adhering to ontologies and vocabularies.
- Ligt is a vocabulary to represent IGT as Linked Data.

Ligt Status

According to a commonly used methodology for publishing multilingual Linked Data (Vila-Suero 2014), the process consists of the following steps:

- 1. Specification: Analyzing and describing data (data sources and RDF data) characteristics;
- 2. **Modelling**: Creating/selecting vocabularies to describe the RDF resources;
- 3. Generation: Transforming the data sources to RDF;
- 4. Linking: Connecting the RDF dataset;
- 5. Publication: Making the dataset available and discoverable on the Web.

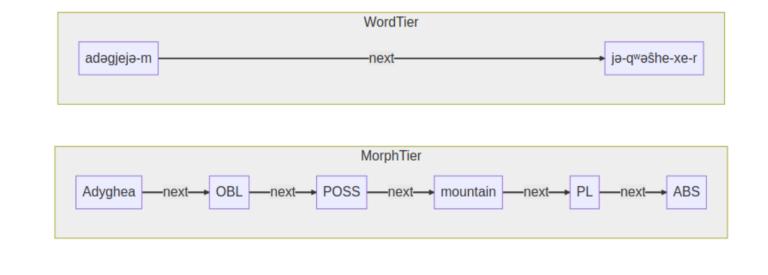
Current focus is on generation, linking and publication.

Ligt Vocabulary

Recent changes:

Decoupled from NIF ligt:Document • New property: ligt:utterance, can link either a (sub dc:Dataset) ligt:hasText ligt:InterlinearCollection ligt:InterlinearText (sub dc:Text)

- text or a set of examples to individual utterances • ligt:Tier is now a subclass or rdfs:Seq, which
- allows data providers to explicitly set the order of elements
- The alignment between elements can be set using DCMI properties dct:hasPart and dct:isPartOf



Converters

- CLDF: Mainly used for typological databases. Can be converted via Metadata URL
- SIL FLExText: Mainly used for language documentation projects. XML-based

ligt:Utterance

ligt:Tier

ligt:Item

ligt:hasTier

ligt:next

ligt:item

• SIL Toolbox: Mainly used for lamguage documentation projects. Text-based. Supports any number of layers.

Publication

- The standard way to publish RDF is to upload it to a SPARQL endpoint
- However, this creates unnecessary obstacles to use RDF/SPARQL
- An alternative to this is to use data dumps (Turtle or any other RDF serialization) and a SPARQL engine, e.g. Apache Jena ARQ or Comunica
- This eliminates the need for maintaining a server and managing the data
- But increases the overhead required to load the data in memory

To compare the performance between the endpoint and a SPARQL engine, we selected 3 queries:

- Q1 Search for all surface forms with a gloss
- Q2 Search for all entries of Causative and Past tense in the same utterance
- Q3 List all case markers used in a language Additionally, we tested on-the-fly conversion to determine the overhead from conversion. Each tested query was on low-to-medium-sized dataset (15k sentences, 1M triples) in 2 different scenarios:
 - on an Apache Jena Fuseki SPARQL endpoint, and
 - locally, using Comunica SPARQL engine.

Evaluation Results

	Endpoint	Locally	On-the-fly
Q1	1.1	3.4	23.9
Q2	1.2	5.4	25.8
Q3	2.7	4.6	25.3



Further developments

Search

- A simple command-line interface to search for Ligt examples across local and remote datasets and SPARQL endpoints
- Supports providing additional triples containing local annotations and a table of mappings from string labels to external ontologies

Serve (enligten)

- A REST API for extracting and converting IGT data from supported formats to RDF on-the-fly with ligt-convert
- In the future support for on-the-fly scraping IGTs from websites and PDFs is planned

Manage (Ligt Workbench)

- Developing a GUI environment for managing and searching in IGT collections.
- Gathering real-life use-cases of data aggregation for extending the toolchain.
- Offline, single-user, cross-platform





