APiCS-Ligt: Towards Semantic Enrichment of Interlinear Glossed Text

Maxim Ionov
Applied Computational Linguistics Lab, Goethe University Frankfurt, Germany

Abstract

This paper presents APiCS-Ligt, an LLOD version of a collection of interlinear glossed linguistic examples from APiCS, the Atlas of Pidgin and Creole Language Structures. Interlinear glossed text (IGT) plays an important role in typological and theoretical linguistic research, especially with understudied and endangered languages: It provides a way to understand linguistic phenomena without necessarily knowing the source language which is crucial for these languages since native speakers are not always easily accessible.

Previously, we presented Ligt, RDF vocabulary created for representing interlinear glosses in text segments. In this paper, we present our conversion of the APiCS IGT dataset into this model and describe our efforts in linking linguistic annotations to an external ontology to add semantic representation.

2012 ACM Subject Classification Information systems → Graph-based database models; Computing methodologies → Language resources; Computing methodologies → Knowledge representation and reasoning

Keywords and phrases Linguistic Linked Open Data (LLOD), less-resourced languages in the (multilingual) Semantic Web, interlinear glossed text (IGT), data modeling

Digital Object Identifier 10.4230/OASIcs.LDK.2021.27

Supplementary Material *Software (Source Code and Dataset)*: `https://github.com/acoli-repo/ligt/tree/master/stable/apics/`
Dataset: `https://doi.org/10.5281/zenodo.5155753`

Funding This work was funded by the project "Prêt-à-LLOD" within the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825182, as well as the project "Linked Open Dictionaries" (LiODi), funded within the eHumanities program of the German Ministry of Education and Science (BMBF, 2015-2021).

1 Background

Linguistic examples with interlinear glossing, be that texts or elicitations, are crucial for linguistic research since they provide a way to understand linguistic structures in languages researchers do not know. Both for exploring language material and to provide proof of a claim, they accompany linguistic research on all stages.

This data may consist of any number of layers: free translation, word-by-word translation, grammatical meaning of morphemes, transliteration, etc. Some layers has morpheme-by-morpheme correspondence between each other, e.g. morpheme segmentation and grammatical meaning of morphemes. Consider the following example in Gurindji Kriol language:[1]

(1) Jambala dei meikim nyawanginyima.

Jambala dei meik-im nyawa-nginyi-ma.
somebody 3PL.SBJ make-TR this-ABL-TOP

"Some people make it out of this one."

[1] For source data, attribution and more information see `https://apics-online.info/sentences/72-35`.

In this example, there are two layers without morpheme alignment ("baseline" and "free translation") and the other two are aligned. The list of layers is not restricted, but there are guidelines, Leipzig Glossing Rules (LGR) [4].

In our previous work [2], we presented Ligt, RDF-native vocabulary capable of representing *structure* of IGT and demonstrated how it could be used to model data produced by widely-used tools for field linguistics: Toolbox and FLEx (based on our research before that [3]). Since then, the vocabulary was also used to represent a massive typologically diverse dataset based on language archive data [14].

While providing a shared model for different source formats increase interoperability between *formats*, i.e. allowing to query over data produced with different tool sets, it does not save against variability of annotations. LGR provide a list of commonly used abbreviations for grammatical categories (e.g. ABL for Ablative case), but this list is neither full nor universally used, and both these reasons lead to mismatches in tags across different datasets. Usually there is a list of abbreviations either in a book or attached to the dataset,[2] and this could be used for disambiguating the labels. However, these are still *labels* (strings), not *categories*.[3] In order to provide semantics for these labels, we create a mapping linking the labels with external ontologies.

The rest of the paper is structured as follows: in Section 2 we describe the source dataset, Section 3 briefly presents the Ligt data model. Sections 4 and 5 are devoted to the conversion and the linking respectively. Finally, Section 6 concludes the paper also pointing out future work.

2 APiCS

The Atlas of Pidgin and Creole Languages [13] is an online database[4] with linguistic information on 76 pidgin and creole languages of the world. This information includes grammatical and lexical features of these languages, collection of references, grammatical surveys. Most importantly for this paper, this database contains a collection of linguistic examples with interlinear glossing (18 526 in total). These examples are of different nature: naturalistic spoken, written or translated, constructed by a linguist, a native speaker, etc. Some of these examples are augmented with speech recordings.

APiCS data model is based on Cross-Linguistic Data Formats (CLDF) [7] which is employed by several typological databases due to its convenience in installation and usage. The model is based on the W3C standard "Model for Tabular Data on the Web" [10] which, in turn, is a dialect of JSON-LD, which lead to the database structure having semantic annotations. Examples are connected to additional information, such as presence of certain grammatical features, but their internal structure is stored as strings, without connections to structured information, e.g. tables of features, meaning that it is not possible to query these examples for grammatical categories in an easy way.

In order to preserve the original annotations, but add internal structure, we decided to use APiCS sentence identifiers as identifiers in Ligt annotation and add `owl:sameAs` links from Ligt sentence fragments back to APiCS.

[2] `https://github.com/cldf-datasets/apics/blob/master/cldf/glossabbreviations.csv`
[3] See also [14, Section 4.3].
[4] `https://apics-online.info/`

3 Ligt

Ligt vocabulary is grounded in three well-established vocabularies: Dublin Core [17], NIF [9] and WebAnnotation [15]. Since this paper focuses on the application of the vocabulary, and not on its definition, we will list only its key aspects here. For a more in-depth description, related work and a survey on alternative representations for IGT, see [2]. Below are some key aspects and new additions:

- The central element is `ligt:Document`, a subclass of `dc:Dataset`. Objects of this class can have multiple pointers to texts and sentences. Previously, the model was limited to collections of texts via the `ligt:hasText` property with an object of type `ligt:InterlinearText` or (`dc:Text`). Since a large amount of IGT data, including APiCS database consists of elicitations or at least of sentences not organized into bigger elements, we have introduced a new property: `hasUtterances` with an object of type `ligt:InterlinearCollection`.
- `ligt:InterlinearCollection` consists of one or more utterances. Some datasets logically consist of independent (or weakly dependent) parts which can be modeled with a single document having multiple `hasUtterances` properties pointing to different `ligt:InterlinearCollection` instances.
- As with text, Using NIF predicate `nif:subString` it is possible to split a text or a interlinear collection into smaller parts: `ligt:Paragraph` or `ligt:Utterance`. `ligt:Utterance` roughly corresponds to a sentence or an elicitation.[5]
- To represent layers, we introduced a class `ligt:Tier` and two subclasses: `ligt:WordTier` and `ligt:MorphTier` which should correspond to sequences of words and morphs, respectively. Tiers in Ligt must consist of elements on the same level of granularity (e.g. words with words vs. morphemes with morphemes).
- Both `ligt:Word` and `ligt:Morph` are subclasses of `ligt:Item` and are objects of a property `ligt:item` for the word and morph tiers, respectively.
- An instance of a tier (a sentence or a word) should have a property `ligt:item` that points to its smaller components. Components within one tier must be connected by a property `ligt:next`.
- Data properties can be added to an item depending on the data (e.g. translation)
- Finally, for compatibility with FLEx data, we keep subclasses of `ligt:Morph` for representing prefixes, suffixes, stems and enclitics.

The data model (excluding metadata) is illustrated in Fig. 1.

4 Conversion

4.1 Conversion details

APiCS example sentences are stored in a CSV table which conforms to a schema[6] describing which layers can be in the data, whether they are required and if there is a separator symbol for the data (e.g. morpheme line).[7] Each row corresponds to a separate sentence so the

[5] Splitting `ligt:InterlinearCollection` into paragraphs might seem strange, but this can, in fact, lead to a nicer modeling: if a group of elicitations is not big enough to be a `ligt:InterlinearCollection` but there need to be some grouping (e.g. subsection in a grammar or a group of examples related to a single phenomenon).

[6] `http://cldf.clld.org/v1.0/terms.rdf#ExampleTable`

[7] Separators were not present in the previous release of the data so initially we split the data during conversion heuristically.

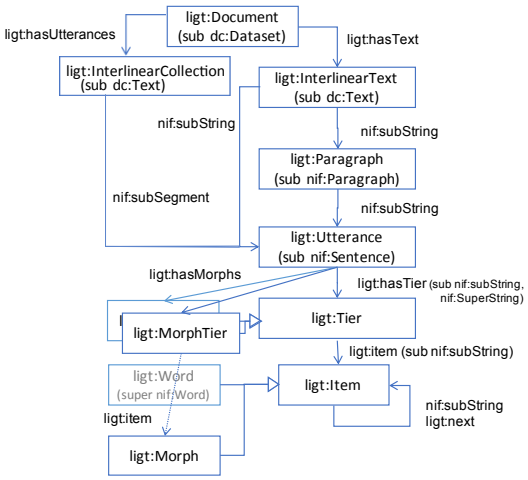

Figure 1 Ligt data model.

conversion process was limited to creating triples with the dataset metadata, adding triples for each sentence, and creating layers for sequences of words for each sentence and for sequences of morphs for each word. The resulting structure is the following:

- Dataset-specific metadata: bibliographic citation
- One `ligt:Document` for all the sentence
- A single `ligt:InterlinearCollection` for all the sentences
- Metadata for each `ligt:Utterance` (sentence): language code, comment, `owl:sameAs` with a link to APiCS[8]
- 3 tiers for each sentence: phrase, words, and morphs
- Original text as an `rdfs:label`, translation as an object of `ligt:translation`,[9] and a comment as an `rdfs:comment`
- For every morph: original text in a `rdfs:label`, gloss marker in a `ligt:gloss`

An excerpt from the converted data is illustrated on Fig. 2.[10]

4.2 Querying

Even after purely structural conversion, without adding semantic information to linguistic categories, it is possible to perform qualitative and quantitative analysis on the dataset. Doing corpus analysis on RDF datasets is beyond the scope of this paper, so we will just demonstrate some exploratory queries.

First query returns grammatical markers which are found in the most number of languages:

[8] In this version of the conversion, we do not add attribution and provenance information for each sentence, but it is easily retrievable since there is a link to the original example record in APiCS.

[9] Here we do not model free translation as a separate `Tier`, but creating a separate tier for it would be a possible design decision, and in fact, a single SPARQL update can be used to convert between the two.

[10] Full resolution and more diagrams can be found at `https://github.com/acoli-repo/ligt/tree/master/stable/apics/diagrams/`

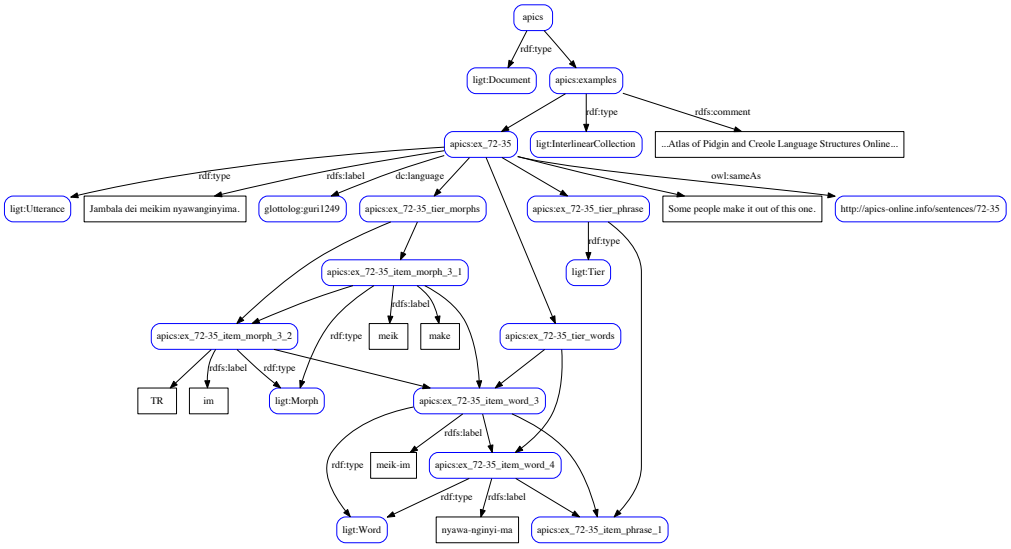

Figure 2 APiCS-Ligt example.

```
PREFIX ligt: <http://purl.org/ligt/ligt-0.2#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT (COUNT(?lang) as ?n_lang) ?val
WHERE {
  ?morph ligt:gloss ?val ;
         rdfs:label ?label .

  BIND(LANG(?label) as ?lang)
  FILTER(?val = UCASE(?val) && ?lang != '')
} GROUP BY ?val ORDER BY DESC(?n_lang)
```

Marker	#
3SG	2650
1SG	2397
NEG	1400
2SG	1306
PST	1099

We can also look for more typologically interesting questions. For example, it might be possible to see the morphosyntactic alignment strategies that exist in languages of the dataset.[11] An easy approximation of this would be to look at the presence of Accusative and Ergative grammatical markers in language data:

```
PREFIX ligt: <http://purl.org/ligt/ligt-0.2#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?case ?lang
WHERE {
VALUES ?case { "ACC"@en "ERG"@en }
?morph ligt:gloss ?case;
       rdfs:label ?label .
  BIND (LANG(?label) AS ?lang)

  FILTER(?lang != '')
}
```

Case	Language
ACC	idb-x-dama1278
ACC	mcm-x-mala1533
ACC	pga-x-suda1237
ACC	sci-x-sril1245
ACC	mue-x-medi1245
ERG	gjr-x-guri1249

This query points to an obvious problem: it is necessary to list all the labels for grammatical cases, we can not query for all possible sets of them. In order to be able to do so, we need to map the labels to some external source, to augment string labels with linguistic categories.

[11] Typologically, there are tendencies to have certain combination of case markers on subjects and objects. Most notably, Nominative-accusative and Ergative-absolutive types. [5]

5 Linking

5.1 Mapping to ontologies

There has been a debate regarding universality and cross-applicability of linguistic categories [8]. While this is, undoubtedly, an important topic, and we carefully agree with the premise that sparked the debate, having linguistic categories such as parts of speech as an approximation is extremely helpful for practical reasons. Nevertheless, we find it important not to overgeneralize and this was one of the concerns in choosing the source we could map APiCS annotation to.

There is a variety of community-maintained repositories of annotation terminology evolved during the 2000s, which aimed to replace annotation standards by collecting and defining categories without requiring them to be disjoint. Exemplary repositories developed at this time include ISOcat [11], developed with a specific focus on language technology, and the General Ontology of Language Description [6], developed with a specific focus on language documentation.

Another repository designed to be flexible and non-reductionist is OLiA, Ontologies of Linguistic Annotation [1]. In its conception, OLiA aimed to address what could be called the "standardization gap" of linguistic annotation. That means that a consistent standardization of linguistic annotation would either have to neglect language specific characteristics (cf. Universal Dependencies tagset), or constantly grow in complexity with every new language added to it. OLiA is modular, and allows users formalize their annotation schemes and to link them with reference concepts. This approach suits our task very well given that:

- The set of markers in APiCS is quite extensive – matched against standard Leipzig Glossing Rules list of abbreviations, we got less than a quarter of the markers matched (23.54%). The list of glosses that is distributed with the dataset has 267 abbreviations.
- Annotations in the dataset are *morpheme markers*, they does not necessarily correspond to grammatical categories: reduplication, oblique stem and agreement are present in the dataset as morpheme values, but they do not directly correspond to a grammatical category (which could be, e.g. an intensifier in case of reduplication).
- One of the modules is a morpheme inventory converted from the UniMorph project [16] which links OLiA Reference Model classes to UniMorph morpheme inventory.

Additionally to OLiA, we decided to map morpheme labels to the morpheme inventory in the MMoOn Core ontology [12]. This ontology was created to provide a shared semantic model for morphological information, which is precisely our goal. In the core of this ontology there is a language-independent collection of morphemes with their labels and description, which we also referenced to enrich our morpheme annotations.

By matching tags and their description we were able to map 123 unique labels, 81 with OLiA ontologies and 91 with MMoOn. For each mapping, we added an additional statement to the dataset:

```
<http://mmoon.org/core/Ablative> apics:hasValue "ABL"@en .
<http://mmoon.org/core/Absolutive> apics:hasValue "ABS"@en .
<http://mmoon.org/core/Accusative> apics:hasValue "ACC"@en .
<http://mmoon.org/core/Active> apics:hasValue "ACT"@en .
<http://mmoon.org/core/Adjective> apics:hasValue "ADJ"@en .

<http://purl.org/olia/unimorph.owl#ABL> apics:hasValue "ABL"@en .
<http://purl.org/olia/unimorph.owl#ABS> apics:hasValue "ABS"@en .
<http://purl.org/olia/unimorph.owl#ACC> apics:hasValue "ACC"@en .
<http://purl.org/olia/unimorph.owl#ACT> apics:hasValue "ACT"@en .
<http://purl.org/olia/unimorph.owl#ADJ> apics:hasValue "ADJ"@en .
```

In our future work we will analyze which labels did not map and whether it is possible to find mappings for them.

5.2 Querying

Now that we have semantic value behind some of the tags, we can query using this additional knowledge. Query below groups all the case markers encountered in sentences in each language:[12]

```
PREFIX ligt: <http://purl.org/ligt/ligt-0.2#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX apics: <http://purl.org/liodi/ligt/apics/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX mmcore: <http://mmoon.org/core/>
PREFIX unimorph: <http://purl.org/olia/unimorph.owl#>

SELECT (GROUP_CONCAT(DISTINCT ?case; SEPARATOR=", ")
        AS ?cases) ?lang
WHERE {
  ?morph ligt:gloss ?case ;
         rdfs:label ?label .
  ?tag apics:hasValue ?case .
  { ?tag rdfs:subClassOf+ mmcore:Case . }
  UNION
  { ?tag rdfs:subClassOf+ unimorph:Case . }


  BIND(LANG(?label) as ?lang)
  FILTER(?lang != '')
} GROUP BY ?lang
```

Cases	Language code
LOC, COM, INS, DAT, TEMP	rop-x-krio1252
LOC, INS, ABL, BEN, ALL, ACC, GEN	mue-x-medi1245
LOC, COM, INS, MOD, ABL, ALL, DAT, ERG	gjr-x-guri1249
INS	gcf-x-guad1242
LOC, VOC, GEN	kcn-x-nubi1253
LOC, VOC, MOD	jam-x-jama1262
COM, INS, VOC	pov-x-uppe1455
LOC, VOC, MOD	srm-x-sara1340
LOC	fpe-x-fern1234
LOC, COM, INS	bah-x-baha1260
VOC	lou-x-loui1240
...	

6 Conclusion

In this paper we presented APiCS-Ligt, an RDF edition of interlinear glossed linguistic examples from the Atlas of Pidgin and Creole Languages. Our conversion remains linked with the original dataset therefore preserving all additional information such as bibliographical references or linguistic features, but at the same time adding for linguistic examples both a structural level and a layer of semantics, providing more interpretability to linguistic annotations and interoperability with another resources with linguistic annotations.

We showed that such semantic linking can be problematic due to both practical (ambiguity of markers) and theoretically-motivated (differences in definitions of linguistic categories) reasons which might be improved if linguists were more involved in the data modeling and data standardization stages.

[12] We give only an excerpt of the results in the table below.

In the future we are planning to go beyond APiCS IGT data to other sources of IGT to see how transferable are the solutions that we came up with. We also plan on publishing a Python module aimed at combining in one place all previously developed procedures for importing and exporting Ligt format and add functionality for working with Ligt data.

The dataset and the code to reproduce the conversion is available at `https://github.com/acoli-repo/ligt/tree/master/stable/apics/`.

References

1 C. Chiarcos and M. Sukhareva. OLiA – Ontologies of Linguistic Annotation. *Semantic Web Journal*, 518:379–386, 2015.

2 Christian Chiarcos and Maxim Ionov. Ligt: An LLOD-Native vocabulary for representing Interlinear Glossed Text as RDF. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

3 Christian Chiarcos, Maxim Ionov, Monika Rind-Pawlowski, Christian Fäth, Jesse Wichers Schreur, and Irina Nevskaya. LLODifying linguistic glosses. In *Proceedings of Language, Data and Knowledge (LDK-2017)*, Galway, Ireland, June 2017.

4 Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. `https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf`, 2008.

5 Robert MW Dixon. *Ergativity*. Cambridge University Press, 1994.

6 S. Farrar and D. T. Langendoen. An OWL-DL implementation of GOLD: An ontology for the Semantic Web. In A. Witt and D. Metzing, editors, *Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology*. Springer, Dordrecht, Netherlands, 2010.

7 Robert Forkel, Johann-Mattis List, Simon J Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A Kaiping, and Russell D Gray. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific data*, 5(1):1–10, 2018.

8 Martin Haspelmath. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic typology*, 11(1), 2007.

9 Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. Integrating NLP using Linked Data. In *Proc. 12th International Semantic Web Conference, 21-25 October 2013*, Sydney, Australia, 2013. also see `http://persistence.uni-leipzig.org/nlp2rdf/`.

10 Gregg Kellogg and Jeni Tennison. Model for tabular data and metadata on the web. W3C recommendation, W3C, 2015. `https://www.w3.org/TR/2015/REC-tabular-data-model-20151217/`.

11 M. Kemps-Snijders, M. Windhouwer, P. Wittenburg, and S. E. Writh. ISOcat: remodelling metadata for language resources. *International Journal of Metdata, Semantics and Ontologies*, 4(4):261–276, 2009.

12 Bettina Klimek, Markus Ackermann, Martin Brümmer, and Sebastian Hellmann. Mmoon core-the multilingual morpheme ontology. *Semantic Web Journal*, 2020.

13 Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. *APiCS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL: `https://apics-online.info/`.

14 Sebastian Nordhoff. Modelling and annotating interlinear glossed text from 280 different endangered languages as linked data with ligt. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 93–104, 2020.

15 Robert Sanderson, Paolo Ciccarese, and Herbert Van de Sompel. Open annotation data model. Technical report, W3C Community Draft, 08 February 2013, 2013.

16 John Sylak-Glassman. The composition and use of the universal morphological feature schema (unimorph schema). *Johns Hopkins University*, 2016.

17 S. Weibel, J. Kunze, C. Lagoze, , and M. Wolf. RFC 2413 - Dublin Core metadata for resource discovery. URL `http://www.ietf.org/rfc/rfc2413.txt` (July 31, 2012), September 1998. Network Working Group.