Addressing Variability in Interlinear Glossed Texts with Linguistic Linked Data

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

Natalia Patiño Mazzotti
Goethe University Frankfurt, Germany
nataliapatinomazzotti@gmail.com

Abstract

In this paper, we identify types of uncertainty in interlinear glossed text (IGT) annotation, a common notation for language data in linguistic research. Using the Linked Data paradigm, we provide guidelines for encoding IGT to address these uncertainties, enhancing interpretability and interoperability without compromising expressivity. Finally, we present *ligtsearch*, a command-line tool with Python bindings provided as part of *ligttools* suite, that uses these guidelines to offer searching and filtering capabilities across multiple datasets in an interoperable way.

1 Introduction

1.1 Background

Interlinear glossed text (IGT) is a notation commonly used to represent language examples in descriptive and typological linguistics. It is designed to provide an intuitive way of showing language material so that it could be understood without needing to know the language. IGT data may consist of any number of layers added under the original text (hence *interlinear*): word-by-word translation, grammatical meaning of morphemes, transliteration, free translation, etc. Some layers have morpheme-by-morpheme alignment between each other, e.g. morpheme segmentation and grammatical meaning of morphemes. Consider the following example from Tundra Yukaghir:

(1) Ieruuče lalime-le me=köjle-s-um.
hunter sledge-ACC PF=break-CAUS-TR.3SG
'The hunter broke the sledge.'
(Schmalz, 2013, p. 66)

This example consists of three layers: morphological segmentation, glosses aligned with the transcription layer, and free translation. The second word is divided into two elements: a root glossed as 'sledge' and a morph *-le*, glossed as the accusative case. The next word[1] consists of the clitic *me=* attached to the verb *kölje* 'break' followed by the causative suffix *-s* and *-um* glossed as TR.3SG, that is, transitive and third person singular.

Generally, datasets and published works that contain IGT follow the Leipzig Glossing Rules, LGR (Comrie et al., 2008), a set of guidelines and recommended glosses for common grammatical categories, such as PL to annotate plural grammatical meaning or ACC for accusative case.

Additionally to these guidelines, a list of abbreviations (markers) for less common grammatical categories is usually included with the data, especially in cases in which a grammatical category is relevant in a given language but not necessarily cross-linguistically.

1.2 Variability in IGT

Since the Leipzig Glossing Rules are guidelines, great variability is allowed to annotate data. The flexibility that these guidelines provide allows them to adapt according to the language, distinguishing several subcategories of a particular grammatical category, when needed. Example (2) introduces a very specific gloss BEFORE.UU, which in the context of the Ese Ejja language is used for subordinated clauses coding coreferentiality between the two (unique)[2] arguments of the main and dependent clauses:

(2) poki-ximawa, eya kya-eno pwaje
go-**BEFORE.UU** 1ABS APF-sad be.FUT
'Before (I) leave, I will be sad.'
(Vuillermet, 2014, p. 358)

[1] The term 'word' is used here as a simplification to refer to a visually separated unit of annotation. The strict definition does not impact the annotations since only morphs and complete examples have corresponding translations. For more on the concept of word, see Schiering et al. (2010); Haspelmath (2023).

[2] According to Vuillermet, Unique arguments are the only arguments of intransitive verbs.

Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025), pages 279–284

Generally, coreference of subjects or lack thereof (a grammatical category known as switch-reference) is marked via the glosses SS (same subject) and DS (different subject). In Ese Ejja, marking the specific syntactic function of the coreferent argument is crucial, since it triggers different marking. In this example, both arguments involved in the coreference are subjects of intransitive clauses, which the author specifies as unique arguments.

In cases like (2), using a non-standard gloss is important since it provides additional information about the grammatical category (i.e. the type of clause and the co-reference of specific arguments).

However, this might hinder its interpretability and interoperability given that different sources might contain different glossing to encode the same grammatical category. The following examples show this variability for the category of evidentiality in Shipibo-Konibo (Panoan):

(3) a. Jawen jema-ronki ani iki.
POS3 village:ABS-**HSY** large COP
'Her village is very large.'
(Valenzuela, 2003, p. 534)

b. Jawen jema-ronki ani iki.
POSS3 village:ABS-**REP** large COP
'Her village is very large.'
(Valenzuela, 2008, p. 34)

In (3), the morpheme *-ronki* which encodes reportative evidential, has been glossed differently in two different instances. Note, that it is not immediately clear from the examples alone if the analyses of this morph in these two cases are identical or this is the case of different granularity for these two markers. The same example shows a more trivial but common case of variability in glossing, which shows the glosses POSS and POS referring to the same grammatical category. In this case, it is immediately clear that this is, in fact, the same category, but this can still cause problems for search or automatic methods.

In some cases, a morph can be analyzed in several ways, once again leading to inconsistent glossing. In the following example, the clausal clitic *=ti* in Yurakaré, that initially was thought to be a different-subject marker (DS), has been alternatively analysed as a nominalizer (NMZR) in more recent literature:

(4) a. më lëtëmë=chi mala-m=**ti**
2SG.PRN jungle=DIR go.SG-2SG.S=**DS**
sëë mi-n-nënë-ni
1SG.PRN 2SG-IO-cook-INTL:1SG.S
'While you go to the jungle, I'll cook.'
(Van Gijn, 2006, p. 312)

b. ta-ka-n-toro=**ti**
1PL.OBJ-3SG.OBJ-BEN-finish=**NMZR**
baytu tishi ta-sibbë=chi
go.1PL.EXH now 1PL.POSS-house=DIR
'When we finish it, let's go to our house immediately.'
(Gipper and Yap, 2019, p. 366)

These three examples demonstrate different cases of annotation inconsistency and variability:

- Multiple labels for the same category (3);
- Difference in granularity of labels (or overlap) (2);
- Alternative analyses (4).

Note, that this does not stem from an "incorrect" use of LGR, but is, in fact, an expected property described in the rules. However, it poses challenges for understanding the data and aggregating over it, both for people and algorithms. In simplest cases, like with glosses POS and POSS, this can be solved by cleaning the data, selecting a single label and normalising the annotation, but for the most part, modifying the glosses would lead to information loss, e.g. in case of (4), where the choice of a marker depends on the function of a morpheme that the author (annotator) wants to highlight. IGT annotations provide an interpretation of the data by a linguist that depends on many factors, and replacing one marker with with a seemingly similar one might change this interpretation. A better solution would be to preserve the original annotations but *explain* them, i.e. add semantics: establish relationships between annotations, group alternative labels, link to external databases of grammatical categories. In the next sections we show how to combine all that by employing the Linked Data paradigm.

The rest of the paper is organized as follows: Section 2 introduces the Linked Data paradigm and describes Ligt, a Linked Data vocabulary for representing IGT. In Section 3 we use Ligt to address each of the aforementioned issues with IGT annotation. Section 4 presents *ligt-search*, a tool that allow to search across Ligt datasets with different annotations.

Finally, Section 5 concludes the paper and outlines directions for future research.

2 Linguistic Linked Data and Ligt

2.1 Linked Data Paradigm

Linked Data is a set of best practices for publishing and connecting structured data on the Web using open standards (Berners-Lee, 2008). It is built around four key principles: using Universal Resource Identifiers (URIs) to uniquely identify entities, making them accessible via HTTP, providing structured descriptions using open standards such as RDF and SPARQL, and providing links to related resources via URIs. This approach allows for the creation of a machine-readable, semantically interconnected web of data, enabling data interoperability and reuse across domains in line with FAIR principles.

Linguistic Linked Data, LLD (Chiarcos et al., 2012; Cimiano et al., 2020) applies these principles specifically to linguistic resources such as lexicons and corpora. By representing linguistic entities with URIs, describing them in RDF, and linking them to external datasets, LLD facilitates semantic interoperability and integration across linguistic and NLP applications. The result is a distributed, reusable, and extensible ecosystem of linguistic data that supports advanced querying, cross-lingual research, and long-term data sustainability.

2.2 Ligt

Ligt is an RDF vocabulary designed for modelling IGT as Linked Data (Chiarcos and Ionov, 2019). It was developed as a generalisation over shallow RDF representations of traditional formats of storing IGT annotations, namely, Toolbox, FLEx and Xigt (Chiarcos et al. (2017) has a detailed description of the formats, their limitations, and these shallow representations). Since its inception, the vocabulary has been applied to multiple datasets, covering language data from hundreds of languages (Nordhoff, 2020b,a; Nordhoff and Krämer, 2022; Ionov, 2021) showing significantly increasing interoperability of collections of IGT coming from different sources stored in different formats.

The most commonly used components of the model are presented on Fig. 1: A dataset consists of texts or collections of IGT, both of which contain a number of `ligt:Utterances`. Utterances, in turn, consist of tiers of annotation which contain the smallest units of annotation — `ligt:Items`. The tiers can be either word-level or morph-level, with the property `ligt:correspondsTo` creating alignment between tiers.[3]

An important but underused feature of Ligt is that it allows having multiple tiers of the same type and multiple annotations for the same unit. Surprisingly, this is lacking in many common formats,[4] but as we show in Section 3.3, it is incredibly important for encoding parallel annotations.

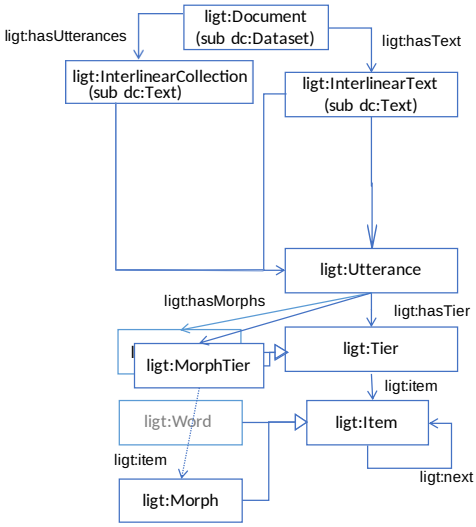

Figure 1: A simplified Ligt data model

3 Addressing Types of Annotation Variability in IGT

3.1 Multiple Labels

Probably the most straightforward issue leading to variation in annotation of IGT across datasets is having multiple labels referring to the same category. This can happen due to personal preferences of the annotator, convenience, or linguistic tradition. An example of this can be found in (3) with the markers REP and HSY both coding the hearsay type of evidentiality.

To address both cases, a user could provide a mapping from the label to a definition of the grammatical category in an external knowledge base. In practice, it is not strictly necessary to use a knowledge base for that, and the annotations can be mapped to an RDF entity defined ad-hoc in the dataset, however this solution lacks interoperability and will require a mapping from properties in each dataset that the user wants to query. With the mappings to a knowledge base, as long as all the datasets map to the same one, the data is interoperable.

[3] Full model description can be found at `https://ligt-dev.github.io/ligt/`.

[4] As far as we know, only Xigt representation allows this.

For example, the following triples map both evidentiality markers from (3) to hearsay evidentiality in the Ontology of Linguistic Annotation (OLiA) (Chiarcos and Sukhareva, 2015), specifically, to its module based on the UniMorph initiative (Batsuren et al., 2022):[5]

```
<http://purl.org/olia/unimorph.owl#HRSY>
                                skos:notation "HSY"@en .
<http://purl.org/olia/unimorph.owl#HRSY>
                                skos:notation "REP"@en .
```

Written like this, the mappings can be added to the triple store alongside with the data or used by SPARQL engines to add the new relations during runtime. The following SPARQL fragment selects morphs annotated as both HSY and REP:

```
...
?morph ligt:gloss ?label .
?meaning skos:notation ?label .
FILTER(?meaning = unimorph:HRSY)
...
```

This example is quite simple, and the same could have been achieved with a simple correspondence table between tagset-specific and universal tags. However, using RDF technologies provides several advantages: First, extending the mappings to several different knowledge bases is trivial. Second, while Ligt is designed to model the *syntax* of IGT, external mappings provide *semantics*: tags are not mere strings, but RDF entities which contain (depending on a knowledge base) additional information, including paradigmatic relationships with other tags.

3.2 Difference in Granularity

A more challenging issue in compatibility of glosses is partial overlap or difference in granularity between the two labels. For example, the aforementioned tag BEFORE.UU in (2) indicates a special case of switch reference, and could be mapped to the same category as the marker SS (same subject). However, with that we lose additional information, encoded in the gloss: a temporal relation between the dependent and the main clauses (BEFORE) and the type of coreference with regards to the semantic roles (unique-to-unique).

In order to create a mapping, we need to provide all the values that it expresses and map them to the string label with the property `skos:notation`, like in the previous section. However, this gloss corresponds to heterogeneous set of values: it combines grammatical categories with syntactic and semantic roles. While it is possible to find a suitable vocabulary to represent syntactic roles and clausal relationship — with OLiA discourse extension (Chiarcos, 2014), we have to create a property for the coreference type ourselves.[6]

```
:uu a owl:Class ;
      rdfs:label "Unique-to-Unique Coreference"@en ;
      rdfs:comment "A coreferent configuration where both
      referring expressions are the only arguments of an
      intransitive verb."@en .
:before_uu a owl:Class ;
    owl:intersectionOf (olia:PrecedenceRelation :uu) ;
    skos:notation "BEFORE.UU"@en .
```

With this, we can introduce the mapping between the gloss and the class as in the previous section:

```
:before_uu skos:notation "BEFORE.UU"@en .
```

Since the gloss is dataset-specific, we create the corresponding class ad-hoc. Despite that, we still have access to additional information about its components according to the relationships established for the ad-hoc class. For example, the following SPARQL fragment extracts labels of all the components of the class that corresponds to the label BEFORE.UU:

```
SELECT ?component ?label WHERE {
  ?compositeClass skos:notation "BEFORE.UU"@en ;
                  owl:intersectionOf ?list .

  ?list rdf:rest*/rdf:first ?component .
  OPTIONAL { ?component rdfs:label ?label }
}
```

3.3 Parallel Analyses

The final issue concerns alternative analyses. In (4), we see an example of that: clitic *=ti* is glossed differently in the same context in two different publications. Unlike the first issue, not only the label is different, but the underlying value as well: DS, a marker indicating switch-reference, was changed to NMZR, a nominalizer, which is a marker indicating a *process* of nominalisation.[7]

The previous solutions were applied to the marker itself, not to its instance, since those issues concerned every usage of a marker. In this case, we cannot apply the same method, since the change is in a specific annotation. However, Ligt provides native support for multiple analyses for both individual words and whole tiers. In this case, we only need to add an additional `ligt:Item` (a subclass

[5] This is just one of possible data sources that the annotations can be mapped to, and the same principle would work with any other repository of grammatical categories. More information on this can be found in (Ionov, 2021).

[6] While this is not necessary, this might be beneficial, especially if the new property would be created as a subclass of an existing context.

[7] As a side note, this is yet another demonstration of heterogeneity of IGT annotations: while switch-reference is a grammatical category, nominalisation is a process. So it is not only a change in the value, but in a type of the annotation.

of `ligt:Analysis`) in the appropriate part of the tier with morphs:[8]

```
:morphs a ligt:MorphTier ;
       ligt:item :m3_1, :m3_2, m3_3, m3_3_alt .
:w3 a ligt:Word ; rdfs:label "mala-m=ti" .
:m3_1 a ligt:Morph ; ligt:correspondsTo :w3 ;
      rdfs:label "mala" ; ligt:gloss "go.SG" ;
      ligt:next :m3_2 .
:m3_2 a ligt:Morph ; ligt:correspondsTo :w3 ;
      rdfs:label "-m" ; ligt:gloss "2SG.S" ;
      ligt:next :m3_3, m3_3_alt .
:m3_3 a ligt:Morph ; ligt:correspondsTo :w3 ;
      rdfs:label "=ti" ; ligt:gloss "DS" .
:m3_3_alt a ligt:Morph ; ligt:correspondsTo :w3 ;
          rdfs:label "=ti" ; ligt:gloss "NMZR" .
```

4 Searching and filtering IGT with *ligt-search*

Following this analysis, we developed *ligt-search*, a tool which allows users to search across local and remote Ligt datasets. Integrated into a package *ligttools*,[9] it can be used either as a standalone command-line utility or called from Python code. In order to allow users combine datasets with different annotations, the tool accepts mappings and additional annotations for each dataset. This way, it addresses the issues discussed in this paper. As a result, not only it allows using datasets from different sources, it provides an opportunity to use opinionated annotations stored locally for the data that is being accessed remotely.

Combined with the other tool in the package, *ligt-convert*, which supports conversion from FLEx, ToolBox and CLDF formats at the time of writing, this allows searching across heterogeneous datasets in common IGT formats.

5 Summary and Outlook

In this paper, we identified three types of variability in IGT annotation and, using RDF vocabulary Ligt, proposed ways to address them to make the annotations more comparable and compatible across datasets. We also introduced *ligt-search*, a tool that uses these techniques to allow users search across IGT datasets in a flexible way, allowing them to provide their own mappings and additional annotations. In the future, this should become a basis for a user-friendly tool that could combine local and remote data, regardless of annotation inconsistencies and personal preferences.

[8] A good practice would be to add a metadata object to both analyses to provide provenance, which we skip here since it is not directly related to the issue.

[9] https://github.com/ligt-dev/ligttools

References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, and 1 others. 2022. Unimorph 4.0: Universal morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855.

Tim Berners-Lee. 2008. Linked Data. https://www.w3.org/DesignIssues/LinkedData.html. [Online; accessed 10-April-2025].

C. Chiarcos and Maria Sukhareva. 2015. OLiA - Ontologies of Linguistic Annotation. *Semantic Web*, 6:379–386.

Christian Chiarcos. 2014. Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4569–4577, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christian Chiarcos and Maxim Ionov. 2019. Ligt: An LLOD-Native Vocabulary for Representing Interlinear Glossed Text as RDF. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *Open Access Series in Informatics (OASIcs)*, pages 3:1–3:15, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Christian Chiarcos, Maxim Ionov, Monika Rind-Pawlowski, Christian Fäth, Jesse Wichers Schreur, and Irina Nevskaya. 2017. Llodifying linguistic glosses. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings 1*, pages 89–103. Springer.

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. 2012. *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*. Springer Science & Business Media.

Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer Nature.

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

Sonja Gipper and Foong Ha Yap. 2019. Life of= ti: Use and grammaticalization of a clausal nominalizer in yurakaré. In *Nominalization in Languages of the Americas*, pages 363–390. John Benjamins Publishing Company.

Martin Haspelmath. 2023. Defining the word. *WORD*, 69(3):283–297.

Maxim Ionov. 2021. APiCS-Ligt: Towards Semantic Enrichment of Interlinear Glossed Text. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 27:1–27:8, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Sebastian Nordhoff. 2020a. From the attic to the cloud: mobilization of endangered language resources with linked data. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, pages 10–18, Marseille, France. European Language Resources Association.

Sebastian Nordhoff. 2020b. 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, Barcelona, Spain. Association for Computational Linguistics.

Sebastian Nordhoff and Thomas Krämer. 2022. IMT-Vault: Extracting and enriching low-resource language interlinear glossed text from grammatical descriptions and typological survey articles. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 17–25, Marseille, France. European Language Resources Association.

René Schiering, Balthasar Bickel, and Kristine A. Hildebrandt. 2010. The prosodic word is not universal, but emergent. *Journal of Linguistics*, 46(3):657–709.

Mark Schmalz. 2013. *Aspects of the grammar of Tundra Yukaghir*. Ph.D. thesis, Universiteit van Amsterdam.

Pilar M. Valenzuela. 2003. *Transitivity in Shipibo-Konibo grammar*. Ph.D. thesis, University of Oregon.

Pilar M Valenzuela. 2008. Evidentiality in shipibo-konibo, with a comparative overview of the category in panoan. *Studies in evidentiality*, pages 33–61.

Rik Van Gijn. 2006. *A grammar of Yurakaré*. Ph.D. thesis, Radboud University Nijmegen.

Marine Vuillermet. 2014. The multiple coreference systems in the ese ejja subordinate clauses. In *Information Structure and Reference Tracking in Complex Sentences*, pages 341–371. John Benjamins Publishing Company.