LLODifying linguistic glosses

Christian Chiarcos, Maxim Ionov, Monika Rind-Pawlowski, Christian Fäth,
Jesse Wichers Schreur, and Irina Nevskaya

Goethe-Universität Frankfurt am Main, Germany
{chiarcos|ionov|faeth|nevskaya}@em.uni-frankfurt.de
{Rind-Pawlowski|Wichers-Schreur}@lingua.uni-frankfurt.de,
WWW home page: http://www.acoli.informatik.uni-frankfurt.de/liodi/

Abstract. Interlinear glossed text (IGT) is a notation used in various fields of linguistics to provide readers with a way to understand the linguistic phenomena. We describe the representation of IGT data in RDF, the conversion from two popular tools, and their automated linking with resources from the Linguistic Linked Open Data (LLOD) cloud. We argue that such an LLOD edition of IGT data facilitates their reusability, their infrastructural support and their integration with external data sources.
Our converters are available under an open source license, two data sets will be published along with the final version of this paper. To our best knowledge, this is the first attempt to publish IGT data sets as Linguistic Linked Open Data we are aware of.

Keywords: Linguistic Linked Open Data (LLOD), interlinear glossed text (IGT), empirical linguistics, data modeling

1 Background

Interlinear glossed text (IGT) is a notation frequently used in linguistics to provide readers with a way to understand the linguistic phenomena in languages they do not know. This notation provides a description for each morpheme of each word between the original text and translation, with one layer (line of text) for every level of description as in (1).

(1) bän söl-ir-ïm ora-nïn dil-n-dä ...
1sg say-aorist-1sg their-gen language-poss.3sg-loc ...
'I speak their language, ...' (Axiska corpus, cf. Sect. 2.2)

Here, we describe the publication of IGT data as a part of the Linguistic Linked Open Data (LLOD) cloud. Based on popular frameworks used for creating and exchanging IGT annotations, FLEx[1] and Toolbox,[2] and the structure of their respective formats, we propose a shallow RDF(S) data model and describe the

[1] http://fieldworks.sil.org/flex
[2] http://www-01.sil.org/computing/toolbox

conversion and linking of two representative data sets. The converters and data will be published under an open license with the final publication of this paper.

These data sets represent the first pieces of IGT data that will become available within the LLOD cloud. In order to do this, Toolbox and FLEx data are converted into an RDF representation *of their original data structures*. This shallow and direct conversion does not provide the rich semantics of more advanced vocabularies for language resources, but guarantees data structures that are transparent and familiar to their user community. The main contribution of this paper, however, are initial steps towards the development of an RDF vocabulary for IGT data. For this initial model, we follow strictly the structure of the original XML- and text-based formats of FLEx and Toolbox, in the longer perspective, these will represent the nucleus for developing an RDF-native data model that allows to generalize to other use cases in linguistics, as well.

Publishing interlinear glosses as LLOD facilitates their reusability and interoperability, demonstrated here for data from two representative tools. This can be partially achieved by the *linkability* inherent to the approach, i.e., the potential to integrate IGT data with external ressources, e.g., to resolve abbreviations of grammatical categories against ontologies, or to link IGT data with existing dictionaries.

The research described in this paper is conducted as part of the BMBF-funded Research Group "Linked Open Dictionaries (LiODi)" (2015-2020) at the Goethe-Universität Frankfurt, Germany, and our acitivities focus on uses of Linked Data to facilitate the integration of data across different dictionaries, or between dictionaries and corpora. LiODi is a joint effort of the Applied Computational Linguistics (ACoLi) lab at the Institute of Computer Science and the Institute of Empirical Linguistics at Goethe University Frankfurt, Germany, with a focus on Turkic languages (pilot phase, 2015-2016), resp. languages of the Caucasus (main phase, 2017-2020) and selected contact languages.

One main type of data in the project are dictionaries [1], but IGT annotations are of particular relevance: The IGT tools addressed here (FLEx and Toolbox) provide a workflow that integrates dictionary (glossary) development with IGT annotation and grammar engineering: For a given expression, resp. its morphological segmentation, in a transcript, possible meanings are automatically looked up in an internal dictionary. If none can be found, the linguist manually assigns glosses, and these are then stored in the internal dictionary.

IGT data comes in different flavors, depending on the tools used for annotation (e.g., FLEx, Toolbox) or publication (e.g., Microsoft Word, LaTeX, PDF). One goal of our efforts is to provide them in an interoperable fashion, regardless of their original format. In the longer perspective, the RDF data and the vocabulary provided by us represent the basis to develop LLOD-native specifications that (a) generalize beyond these various source formats, and that (b) provide queriable and explicit links between IGT data, lexical and other linguistic resources. As argued in Sect. 6, popular RDF vocabularies currently applied to web annotation [2,3] or NLP pipelines for Semantic Web applications [4] are not directly appropriate to represent IGT data, so, instead, we build our efforts on

established conventions in the scientific community that produces and uses this kind of data.

2 IGT data

The *Leipzig Glossing Rules* [5] define a glossed example as a set of lines, each containing a representation or an analytic description of (a part of) the example. Every line has a particular type (or 'marker'), e.g.,

- The original orthography
- A morpheme-by-morpheme gloss
- A free translation into the description language

These are types of lines for (1), IGTs may, however, contain more information, for example, a phonetic transcription, etc.

An important aspect of IGTs is that lines tend to be positionally aligned: For example, word *dilndä* and its morpheme-by-morpheme gloss *language-poss.3sg-loc* refer to the same segment, more elaborate IGTs may also align morphemes with individual glosses. etc.

Considering morpheme-by-morpheme glosses, we cannot assume that there is a one-to-one correspondence between morpheme and grammatical value. In the Leipzig Rules, this is reflected by different separators, IGT tools such as Toolbox provide space-based subsegmentation and alignment.

Along with the segmentation issue, a second interoperability problem exists when it comes to the abbreviations (tags) used for glossing: Although there is a list of standard abbreviations for the grammatical categories in the rules, it is often extended or modified to reflect specifics of the language or theory adopted. In some cases even the definition of the grammatical categories varies across research teams and methodologies. This is often the case with the language description of less-studied languages. All these details increase variations between glosses produced by various researchers and thus decrease the reuse potential of collections of glossed texts drastically.

A third dimension of variation lies in differences of formats and tools used to produce and publish glosses in electronic form:

- Often, glosses are written with Office tools, primarily in word processing formats. The data produced this way is often disseminated as PDF, plagued by insufficent formalization, which practically leads to an inability to reuse this data.
- A second approach is glossing with tools originally developed for other annotation tasks, e.g., as ELAN or Exmaralda.[3]
- The third (and recommended) approach is to use tools developed specifically for creating and managing IGTs: Most widely known are Toolbox (formerly

[3] `http://annotation.exmaralda.org/index.php?title=Advanced_Glossing`

Shoebox)[4] and its successor FLEx;[5] both allow linguists to enter and store IGTs, perform analyses, and extract dictionaries.

As for data from the first group, [6] created an IGT mining service and provide their data in a structured XML form similar to the third group. As for the second group, these tools have wider application beyond IGTs (esp. multimedia annotation), and therefore their data structures are less transparent to the linguists that use them.

We thus focus on the third group: (1) We derive an initial RDF(S) vocabulary from the data structures of the XML-based export format of FLEx, illustrated for a small Megrelian corpus. (2) We demonstrate its applicability to Toolbox data from a corpus of Axiska (Meskhet).

2.1 Megrelian FLEx data

Megrelian is a Kartvelian language spoken by approximately 500 000 people in Western Georgia. It is genealogically related to Georgian, the official language of the country.

The Megrelian data is a small FLEx corpus collected and processed by Jesse Wichers-Schreur in August 2016 during a short fieldwork stay in the village Orsantia. The data are translations from Georgian sentences, from three different speakers, recorded with a zoom recorder. The recorded files were transcribed in ELAN and exported into FLEx. The Megrelian corpus contains 5 889 tokens and 19 007 glossed morphemes but will be extended in subsequent fieldwork campaigns.

The goal of this effort is to document the specifics of the language contact situation Every speaker of Megrelian is bilingual and speaks Georgian from an early age onwards. Georgian is the language of bureaucracy, education and virtually all written communication. Hence, a lot of influence of this standard language on Megrelian is to be expected. However, this influence has not been studied yet, an automatically supported comparison with dictionaries of Georgian and historical contact languages may provide a quantitative basis for such an enterprise, and a LLOD edition of the Megrelian data together with its dictionary links will facilitate its subsequent reproducibility and transparency.

2.2 Axiska (Meskhet) Toolbox data

The project "Interaction of Turkic Languages and Cultures in Post-Soviet Kazakhstan" (Irina Nevskaya and Claus Schönig; Volkswagen foundation, 2014-2017)) focused on the interaction of languages and cultures of 25 Turkic peoples in Kazakhstan. A nation-wide survey was carried out in Kazakhstan in order to collect information on self-identification, language use, the attitude towards education in the mother-tongue, the historic aspects of the settlement (such as deportation), interethnic relationships, the maintenance of the specific culture, and the

[4] `http://www-01.sil.org/computing/catalog/show_software.asp?id=79`
[5] `http://fieldworks.sil.org/flex`

role of religion. Speech samples of these Turkic languages were recorded and analyzed, focusing on oral literature and autobiographic narrations.

Axiska (Meskhet) is one of the Turkic languages in the scope of the project, spoken by a population that immigrated from Georgia in the 1940s. Their language contains a large number of loan words from an old layer, transmitted mainly via Persian and Georgian influence, as well as new layers of loan words in the 20th c., transmitted via Russian and Kazakh influence. The Axiska corpus contains 1,642 glossed sentences with 13,626 tokens and 21,104 glossed morphemes).

At the moment, the linguistic status of Axiska in relation to Azerbaijani and Turkish is debated, and it has been suggested that Axiska represents an intermediate state in a dialect continuum between both languages. Comparing this corpus data with dictionaries of Turkish, Turkish dialects, Osmanic Turkish, Azerbaijani and other varieties of Azeri may help to confirm this hypothesis. Using a Linked Data edition of the corpus glosses and morphemes can be directly linked with the lexical-semantic resources, thereby facilitating the scientific transparency and reproducibility of the comparison.

3 FLEx IGT $\mapsto$ RDF

Figure 1 shows selected glosses in the FLEx graphical user interface and Fig. 2 provides the corresponding fragment from the XML export for the word *koʒir* 'he saw'.

The FLEx distribution includes a (non-validating) XSD schema that illustrates the basic data structure of FLEx files, illustrated in Fig. 3. Although the schema is non-validating, we use it as a basis for FLEx concepts and properties. One characteristic is that the XML format clearly separates structures and annotations, the latter being kept in separate `item` elements whose `@type` carries the corresponding line type ('marker'). Another important characteristic (esp. in comparison to Toolbox) is that FLEx defines explicit datatypes for different types of segments and clarifies their nesting, we thus distinguish `paragraph`, `phrase`, `word` and `morph` – in Toolbox, such segmentation is implicitly expressed via spaces and indents.

The actual conversion is implemented with a simple XSLT script using the following procedure:

- Every XML element in FLEx IGT XML has an ID, we use this as the local name of the URI, the base URI is provided by the user at conversion time and should identify the original corpus. For FLEx concepts and properties, we use the FLEx IGT namespace `http://fieldworks.sil.org/flex/interlinear/`.
- Aggregators (`paragraphs`, `phrases`, `words`, `morphemes` aggregate multiple `paragraph`, `phrase`, `word`, `morph` elements, also cf. `languages` and `media-files` for `language` and `media`) and property wrappers (`item`) are omitted, their information is reorganized (see below). Language `constraints` are currently omitted.

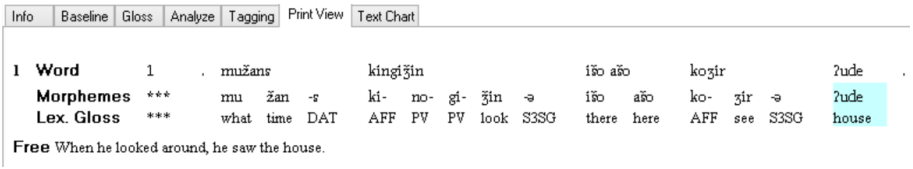


Fig. 1. Megrelian IGT sample, FLEx print view

```
<phrases>
  <phrase guid="6ddb5ab4-58b0-4a14-b0ba-de2a6ee50bd5" begin-time-offset="161450" e
    <item type="segnum" lang="en">1</item>
    <words>
      ...
      <word guid="d3fe4bf6-7150-4ab1-8346-dfae5a840930">
        <item type="txt" lang="xmf-Latn-GE-x-megrelia">kogir</item>
        <morphemes>
          <morph type="prefix" guid="d7f713db-e8cf-11d3-9764-00c04f186933">
            <item type="txt" lang="xmf-Latn-GE-x-megrelia">ko-</item>
            <item type="gls" lang="en">AFF</item>
          </morph>
          <morph type="stem" guid="d7f713e8-e8cf-11d3-9764-00c04f186933">
            <item type="txt" lang="xmf-Latn-GE-x-megrelia">gir</item>
            <item type="gls" lang="en">see</item>
          </morph>
          <morph type="suffix" guid="d7f713dd-e8cf-11d3-9764-00c04f186933">
            <item type="txt" lang="xmf-Latn-GE-x-megrelia">-ə</item>
            <item type="gls" lang="en">S3SG</item>
          </morph>
        </morphemes>
      </word>
      ...
    </words>
    <item type="gls" lang="en">When he looked around, he saw the house.</item>
  </phrase>
</phrases>
```

Fig. 2. Megrelian IGT sample, FLEx XML export

- `media` and `language` with CDATA content T are captured as properties: A `flex:media` T, and A `flex:language` T for the closest non-omitted ancestor A.
- Every child X of an aggregator X`(eme)s` is connected with its following X sibling by the property `flex:next_`X.
- Every non-omitted element X is connected with its closest non-omitted ancestor by the property `flex:has_`X.
- If an aggregated element has a `@type` X, we define it as `rdf:type flex:`X.
- If an `item` has `@type` X and CDATA content T, we create a property A `flex:`X T for its parent A.
- If an `item` has CDATA content, we preserve the value of the attribute `@lang` content as the language type of the associated property.
- If an `item` has further attributes, the associated RDF property is reified and the item information attached to the `rdfs:Statement`.

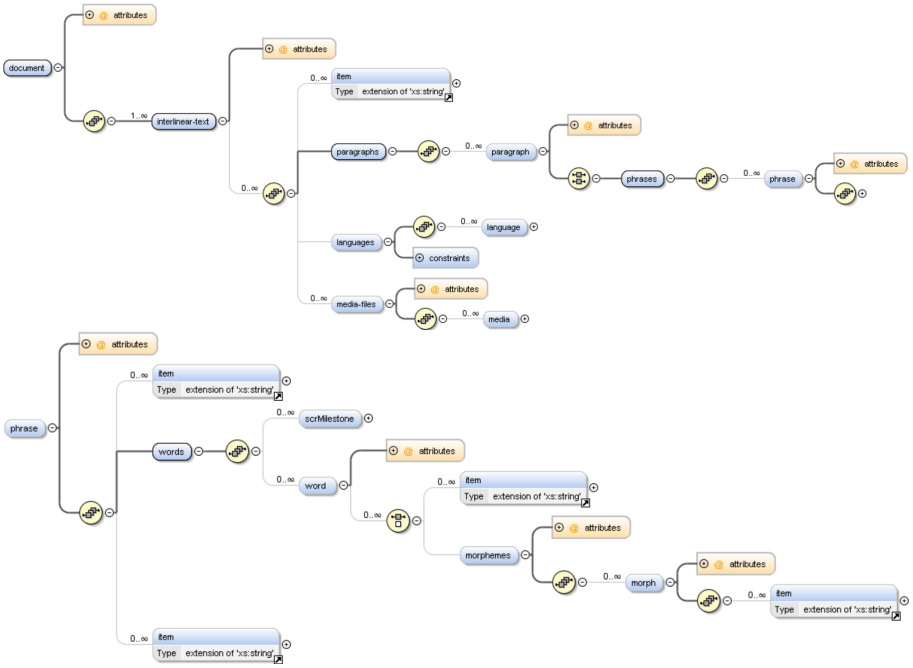


Fig. 3. Graphical visualization of the FLEx XSD schema (Oxygen)

– Every other attribute a is preserved as datatype property `flex:`a with its original value.

The result of the conversion process is illustrated in Fig. 4. Two aspects are to be noted: (1) This converter is generic in that it is not constrained to a fixed set of datatype properties (`item/@type`, line types, 'markers'). (2) We do *not* develop a novel data model for IGT – for which LLOD-native vocabularies are otherwise lacking –, but that we merely transpose the FLEx data model to the LLOD world by creating an isomorphic reconstruction of the FLEx data model in RDF(S). In both aspects, we do not provide explicit semantics, but rather mirror the structures in the original XML file as defined in the accompanying XSD schema. Such 'syntactically defined' semantics are, however, insufficient in terms of the Semantic Web. Nevertheless, the FLEx RDF data model represents a first, empirically grounded step to develop a full-fledged vocabulary for IGTs in the LLOD cloud.

Figure 5 illustrates the RDFS data model of the Megrelian corpus – with classes, their attributes (datatype properties) and object properties that can hold between them. Note that this schema is empirically deduced from the converted data, not pre-defined. It is expected to be exhaustive with respect to FLEx IGT concepts, but other corpora may introduce resource-specific properties not observed in our data.

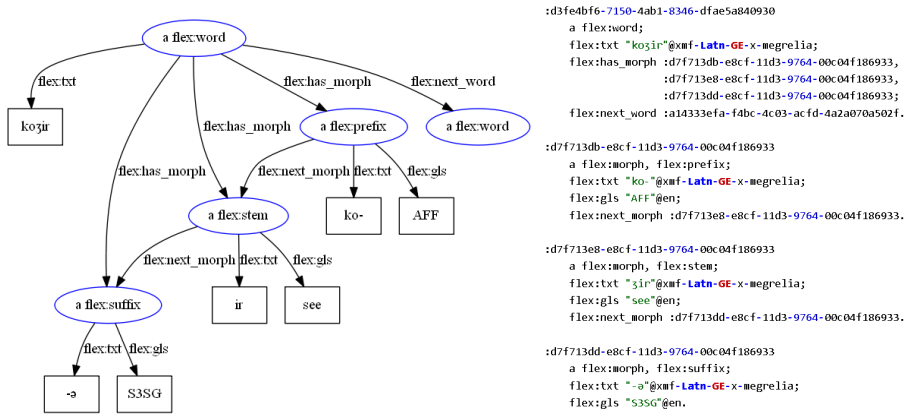


Fig. 4. Megrelian IGT sample, generated RDF graph

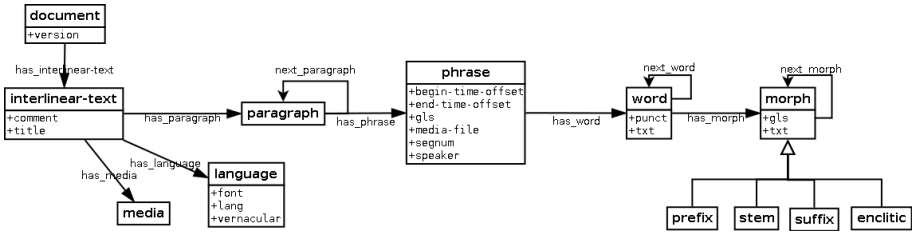


Fig. 5. RDF schema fragment for the Megrelian corpus

4 Toolbox IGT ↦ RDF

FLEx allows to import Toolbox IGT data, but differs from Toolbox in that it relies on predefined markers for different IGT lines. Importing from Toolbox, which is more flexible in this regard, thus requires a manual mapping from annotation-specific markers used in the original glosses and the FLEx identifiers. Such a mapping is, however, not always possible in a loss-less way, e.g., if problem-specific glossing layers are to be added that represent, for example, multiple orthographies (e.g., original non-Latin, transliterated into Latin alphabet, normalized/tonalized, internal scientific transcription, scientific transcription as required by a particular publisher). To preserve information which exceeds the capabilities of FLEx and to facilitate fully automated conversion of Toolbox data, we developed a second processing pipeline for converting Toolbox data directly (rather than via FLEx). We employ the FLEx data model, but as Toolbox is unrestricted in its markers, they (resp., the datatype properties generated from them) are kept apart in a separate `toolbox` namespace.

Internally, Toolbox uses a text-based format, with markers identifying different lines of IGT as well as metadata fields. Format and visualization in Toolbox is illustrated in Fig. 6. As a first glance, we recognize units of analysis that corre-

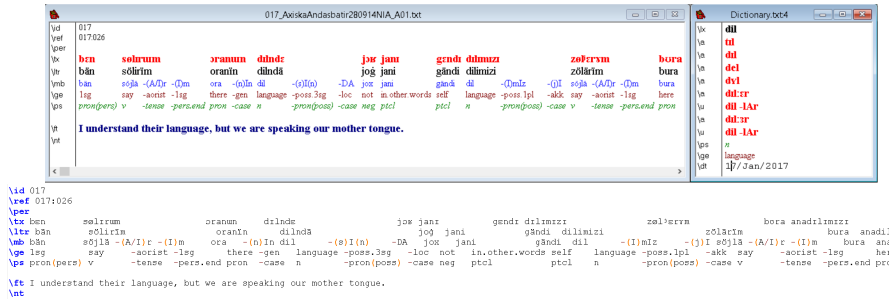


Fig. 6. Axiska sample data (Toolbox view and text format)

spond to `flex:phrase` (i.e., one IGT, cf. \ft), `flex:word` ([multi-]columns containing orthographic words, cf. \tx and \lt), and `flex:morph` (columns aligned with the morphological segmentation, cf. \mb, \ge and \ps). A clear counterpart of `flex:paragraph` cannot be found, these are informally represented by IGT-like structures with special markers that occur *between* IGTs.

It should be noted that detecting layers and morph segmentation in Toolbox data can only be partially automatized, as it is not systematically formalized by the tool. Accordingly, our converter applies a number of heuristics, and extraction errors are possible. The flexibility of Toolbox poses three challenges, in particular:

(a) distinguish IGT markers (that constitute a IGT) and metadata markers (which apply to a document or parts of it),
(b) clarify which markers constitute a single IGT (empty lines may serve both as IGT separators but also occur within IGTs),
(c) capture alignment between different annotation layers (in Toolbox represented by 'exact' positioning using spaces)

As an approximative solution to (a) and (b), we provide the converter with an 'anchor' parameter that identifies the first and unique element in an IGT (normally, an identifier). Everything following it is considered part of the same IGT. All information preceding the first such 'anchor' marker is considered document metadata. Based on these conventions, paragraphs cannot be identified; in Toolbox data, we thus allow `flex:has_phrase` to hold directly between `flex:interlinear-glosses` and `flex:phrase`.

From Toolbox data, the alignment between word and morph segmentation can only be guessed,[6] as the 'exact' positioning using spaces is specific on the font. Using the number of space-separated segments per line, we apply the following heuristics:

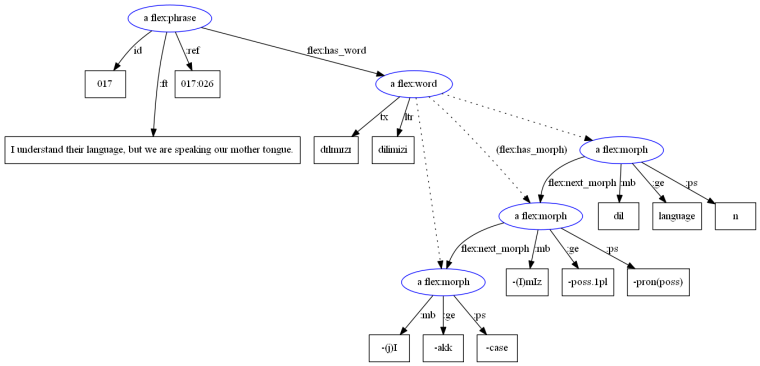
– if a line contains multiple spaces, or if it precedes or follows another line with the same number of segments, it is a segmented line

[6] In fact, the FLEx importer also omits such data.

– unsegmented lines are cast as datatype properties of the `flex:phrase`, their type attribute defines their local name in the `toolbox` namespace
– the first segmented line in an IGT constitutes `flex:word`s
– a segmented line following another segmented line that has the same number of segments uses its URIs
– a segmented line following another segmented line that has a lower number of segments constitutes `flex:morph`s

This extraction process is heuristic, and likely to fail if, for example, multiple whitespaces have been used in a comment. When extraction errors occur, it is recommended to check and to adjust the use of whitespaces in the original Toolbox data rather than to work the generated RDF data. In this way, the LLOD converter contributes to validating Toolbox data.

A subsequent problem pertains to the `flex:has_morph` property holding between `flex:word`s and `flex:morph`s. As there is no principled way to align both layers automatically without enforcing additional constraints on markers, we formalize `toolbox:has_morph` as a direct relation between `flex:phrase` and `flex:morph`s. Subsequent, problem-specific processing, however, should replace this property with the regular `flex:has_morph` property. Technically, this can be implemented using SPARQL update scripts.




```
_:interlinear-text flex:has_phrase :s26.

:s26 a flex:phrase; flex:next_phrase :s27.      # phrase: structural information
:s26 toolbox:id "017"; toolbox:ref "017:026".   # phrase: metadata and data
:s26 toolbox:ft "I understand their language, but we are speaking our mother tongue.".

:s26 flex:has_word :s26_tx8.                              # word
:s26_tx8 a flex:word; flex:next_word :s26_tx9.            # structual information
:s26_tx8 toolbox:tx "dɨlɨmɨzɨ"; toolbox:ltr "dilimizi". # annotations

:s26 toolbox:has_morph :s26_mb13, :s26_mb14, :s26_mb15.  # morph(eme)s
:s26_mb13 a flex:morph; flex:next_morph :s26_mb14.       # structural information
:s26_mb14 a flex:morph; flex:next_morph :s26_mb15.
:s26_mb15 a flex:morph; flex:next_morph :s26_mb16.
                                                         # annotations
:s26_mb13 toolbox:mb "dil";      toolbox:ge "language";  toolbox:ps "n".
:s26_mb14 toolbox:mb "-(I)mIz"; toolbox:ge "-poss.1pl"; toolbox:ps "-pron(poss)".
:s26_mb15 toolbox:mb "-(j)I";    toolbox:ge "-akk";      toolbox:ps "-case".
```

Fig. 7. RDF fragment for the Axiska IGT corpus

The revised FLEx data scheme as applied to the Axiska Toolbox data is shown in Fig. 7.[7] Despite minor systematic differences (different namespaces for datatype properties, omission of `flex:paragraph` in Toolbox, word-morph alignment in Toolbox), the basic data structures for both corpora resemble each other closely and will eventually lead to a common generalization: `toolbox:has_morph` and `flex:has_morph` refer to superstructures of morphemes, and are conceptually comparable (but not adequately represented by) `nif:subString`. The open set of Toolbox properties (markers), on the other hand, can be represented by means of property subsumption, with FLEx properties defining superproperties. In comparison to the mapping-based approach currently provided by the FLEx import (which may easily lead to data loss in the FLEx import if multiple Toolbox markers must be mapped to a single FLEx marker), this generalization is *lossless* – as neither the conversion from FLEx to Toolbox nor vice versa –, thereby illustrating the added value of an RDF-based data model for IGT data in the current landscape of tools and infrastructures.

5 Resource integration: Linking with dictionaries

IGT annotation with FLEx or Toolbox is connected with maintenance of and reference to an internal dictionary. An important extension beyond FLEx is that an RDF edition of IGT data can also be easily integrate **externally provided** knowledge bases from the web of data. In particular, we may integrate further, externally provided dictionaries. Here, we employ DBnary [7] to facilitate the cross-linguistic interpretability of gloss labels. DBnary is an LLOD edition of Wiktionary data for 16 languages, it provides definitions and – more importantly – translations. By linking the English gloss labels in our data that refer to lexical items with DBnary, we can thus easily lookup their translation in other languages. Glosses with different description languages can thus more easily compared with each other.

Fig. 8 shows a SPARQL update script that links glossed FLEx concepts with the English DBnary via `ontolex:lexicalForm`.[8] While this update script requires a local copy of DBnary data, it should be noted that we can also access DBnary data *directly* via its SPARQL end point with the `SERVICE` keyword. In this way, it is also possible to provide external access to FLEx dictionaries (and corpus data), it will thus become possible to develop glosses and dictionaries in a distributed fashion, synchronized over shared dictionaries.

[7] The dotted lines are yet to be inferred from `toolbox:has_morph`.

[8] As defined in the lemon/ontolex vocabulary, this implicitly casts `flex:morph`s as `ontolex:Form` (expected object of `ontolex:lexicalForm`). To satsify the `ontolex:Form` definition, we may add `flex:gls rdfs:subPropertyOf ontolex:representation`. Note that the rendering of a gloss as a lexical entry mirrors the way glosses are treated in FLEx and Toolbox: During annotation, a dictionary comprising all glossed forms is created. In many cases, this dictionary (and the accompanying grammar) represents the main outcome of IGT annotation.

```
# preprocessing for Axiska: "translate" toolbox:ge to flex:gls
# as used in the FLex/Axiska data
INSERT { ?glossedTerm flex:gls   ?gloss }
WHERE  { ?glossedTerm toolbox:ge ?gloss };

# generate ontolex:lexicalForm links with DBnary lexical entries
INSERT { ?lexeme ontolex:lexicalForm ?glossedTerm }
WHERE  { ?glossedTerm flex:gls    ?gloss.
          GRAPH <http://kaiko.getalp.org/dbnary/eng> {
            ?lexeme dbnary:writtenForm ?form.
            FILTER(str(?form)=?gloss)
       }};

# remove ambiguous ontolex:lexicalForm links
# define ?glossedTerm as an ontolex:LexicalForm
DELETE { ?lexeme ontolex:lexicalForm ?glossedTerm. }
WHERE  { ?lexeme ontolex:lexicalForm ?glossedTerm.
         ?other  ontolex:lexicalForm ?glossedTerm.
         FILTER(?other!=?lexeme)
};
```

Fig. 8. Linking FLEx morph(eme)s with DBnary concepts

In the same way, we may also access other external resources. As an example, Ontologies for Linguistic Annotation [8, OLiA] represent a hub of annotation terminology in the LLOD cloud. Among other annotation schemes, they also formalize labels for IGT glossing following [9]. As these are based on the Leipzig Glossing Rules, most abbreviations used in FLEx and Toolbox datasets can be resolved and linked. As a result, it becomes possible to compare the grammatical annotations with specification as provided by dictionaries, for example. These may also help disambiguating lexical forms, e.g., verbs and deverbal nouns in English (cf. *help* or *play*) and thus improve dictionary linking.

6 Summary and discussion

We described an approach of representing interlinear glosses in RDF, the conversion of FLEx and Toolbox data, and we proposed an extensible FLEx-based RDFS vocabulary for IGT data. The converters described here are available from https://github.com/acoli-repo/LLODifier under an Apache 2.0 license. The IGT datasets for Axiska and Megrelian will be published with the final publication of this paper will be accompanied with the publication, until then, their linking with LLOD resources is being improved. Despite a considerable previous efforts towards this direction (see below), this is the first attempt to publish IGT data sets as Linguistic Linked Open Data we are aware of.

Our approach is based on a direct transformation of the highly popular XML-based data model of FLEx. The RDFS vocabulary derived in this fashion is semantically shallow, but represents a suitable basis for developing more advanced specifications and the semantic enrichment of IGT data with LLOD resources. We expect several benefits of an LLOD representation of IGT data: explicit, declarative and transparent linking with external dictionaries and terminology bases, the existence of off-the-shelf technology for querying, storing and manipulating RDF data, and the of RDFS and ontologies to formalize a data model that provides a loss-less generalization over the output of Toolbox and FLEx.

Although FLEx and Toolbox are normally considered as equivalent, conversion between them is neither lossless nor fully automated: Converting Toolbox data to FLEx requires a mapping of an unrestricted inventory of markers to a restricted inventory of markers; converting FLEx to Toolbox requires transforming nested data structures (`flex:word` and `flex:morph`) to space-aligned segments as well as recovering a highly complex structure of dependencies and configuration files which is laborsome and manual task. The FLEx-based IGT data model developed here, however, can be used to develop an RDF-native infrastructure for querying and evaluating data from both (and other) sources. While this is already possible with off-the-shelf RDF technology, it requires considerable experience with SPARQL. In order to develop a solution appropriate for field linguists, a user-friendly interface is yet to be designed. A natural choice in this regard is to develop a graphical query editor akin to the FLEx concordancer. Its queries, however, are internally translated to SPARQL and run against an RDF triple store. Yet, this is just one direction for future development.

Another direction is to elaborate on the relationship between FLEx data types and lemon/ontolex. The research described in this paper is conducted in the context on a project on "Linked Open Dictionaries". Among other aspects, it features the prototype of a comparative-linguistic workbench that implements a transitive search across multiple lexical resources, e.g., Chalkan > Russian > English > other Turkic.[9] This transitive search exploits the fact that the dictionaries employed all make use of (different variants of) lemon. Transitive search thus boils down to transitive property paths and can be easily implemented. With IGT data adherent to a lemon/ontolex-based vocabulary, it can be seamlessly added into the existing infrastructure.

A third direction is the development of the FLEx RDF(S) data model towards an LLOD-native vocabulary for IGTs. Such an effort must be cautiously aware of existing vocabularies and harmonized with related efforts. Although the predominant paradigm when developing novel tools for IGT annotation seems to continue to be XML (as, for example, in Xigt,[10] the toolset and native format of the ODIN project [6]), the RDF edition of IGTs data has been discussed early on [10]. For example, [11] "suggest that a unified RDF representation is well-suited for the creation of materials for local communities due to rapidly developing trends in data dissemination technology" and discuss this with application to ELAN, FLEx and Toolbox. To our best knowledge, no implementation nor RDF data has been published as results of these efforts, but nevertheless, the future development of the FLEx RDF specifications need to be coordinated with this community. In different branches of linguistics, similar efforts have begun. This includes TYTO,[12] a collaborative infrastructure for linguistic data grounded in Semantic Web technologies. At the moment, TYTO exists as a fully functional prototype, and it is populated with data comparable to that considered here, but has not been publicly released (A. Schalley – pers. comm., May 2016). Another

[9] `http://dbserver.acoli.cs.uni-frankfurt.de:5000/search/?query=&originLang=&targetLang=trk`

[10] `https://github.com/xigt/xigt/wiki`

relevant community effort is the Cross-Linguistic Linked Data (CLLD) platform [13] that provides LLOD editions of resources such as WALS and Glottolog. IGT data is a natural extension to these data sets and its treatment has been discussed as an open issue, but seems to converge to a JSON variant of Xigt.[11] To our best knowledge, a common data model for IGT data in RDF is thus not in sight. The FLEx-based vocabulary introduced here represents an initial effort towards its creation.

RDF-native corpus formalisms represent another strand of related research. It should be noted, though, that the most popular vocabularies applied to represent linguistic annotations were originally developed for other aspects of metadata and data exchange in the Web of Data, in particular, approaches focusing on representing the output of natural language processing (NLP) tools for Semantic Web applications, and approaches to formalize metadata ('annotations') about web content. The NLP Interchange Format (NIF) is an RDF/OWL-based format that aims to achieve interoperability between tools, language resources and annotations [14]. It provides a URI scheme that identifies (sub-)strings in a particular reference context, and it comprises basic data structures for linguistic annotations such as `nif:Word`, `nif:Phrase` and `nif:Sentence` – all of which are subclasses of `nif:String`. However, NIF does not provide data structures required for IGTs such as annotation units below word level (morphs) or empty elements (zero pronouns, null morphemes). These could be modeled as `nif:String`, but they do not come with a predefined adjacency property (such as `nif:nextSentence` for `nif:Sentence`s). Open Annotation (OA) / Web Annotation (WA) [2,3] was originally developed to represent textual descriptions ('annotations') to web content. OA/WA formalize the representation of annotations as reified `Annotation` properties that point from a body (= annotation value) to a target (= annotated string), can have a type and carry other annotations. While this can be employed for linguistic annotations, the string- ('target') based approach seems difficult when representing overlapping annotations with different segmentation granularity (morph vs. word), but also for the annotation of zero elements, be it elliptical pronouns (on the word level) or null morphology (on the morph(eme) level). Modelling IGTs with either NIF or OA/WA thus requires considerable conceptual work, potentially requiring a revision of the existing core specifications.

Neither of these established formalisms is thus adequate to provide a host vocabulary to which FLEx data structures can be added. A lesson to be learnt from the relative success of NIF and OA/WA is, however, that community standards are more likely to be applied if they follow the needs of their user community closely. Both have been developed in a bottom-up perspective and extended when needed (for their original use case, but not for IGTs). The RDF reconstruction of FLEx serves exactly this purpose, and represents a nucleus from which community-specific formalisms to represent IGT and related corpus formalisms adequately and transparently as Linked Open Data.

[11] `https://github.com/glottobank/cldf/issues/10`

References

1. Abromeit, F., Chiarcos, C., Fäth, C., Ionov, M.: Linking the Tower of Babel: Modelling a massive set of etymological dictionaries as RDF. In McCrae, J., Chiarcos, C., Montiel Ponsoda, E., Declerck, T., Osenova, P., Hellmann, S., eds.: Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources, Portoroz, Slovenia (2016) 11–19
2. Sanderson, R., Ciccarese, P., Van de Sompel, H.: Open annotation data model. Technical report, W3C Community Draft, 08 February 2013 (2013)
3. Sanderson, R., Ciccarese, P., Young, B.: Web annotation data model. Technical report, W3C Recommendation 23 February 2017 (2017)
4. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: 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/`.
5. Comrie, B., Haspelmath, M., Bickel, B.: The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. `https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf` (2008)
6. Lewis, W.D.: ODIN: A model for adapting and enriching legacy infrastructure. In: Second International Conference on e-Science and Grid Technologies (e-Science 2006), 4-6 December 2006, Amsterdam, The Netherlands, IEEE Computer Society (2006) 137
7. Sérasset, G.: DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. Semantic Web Journal **648** (2014) cf. `http://kaiko.getalp.org/about-dbnary/`.
8. Chiarcos, C., Sukhareva, M.: OLiA - Ontologies of Linguistic Annotation. Semantic Web Journal **518** (2015) 379–386
9. Dipper, S., Götze, M., Skopeteas, S.: Information structure in cross-linguistic corpora: Annotation guidelines for phonology, morphology, syntax, semantics, and information structure. Interdisciplinary Studies on Information Structure (ISIS), Working papers of the SFB 632 **7** (2007)
10. Poornima, S., Good, J.: Modeling and encoding traditional wordlists for machine applications. In: Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, Uppsala, Sweden, Association for Computational Linguistics (July 2010) 1–9
11. Nakhimovsky, A., Good, J., Myers, T.: Interoperability of language documentation tools and materials for local communities. In: Digital Humanities (DH2012), Hamburg (July 2012) `http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/interoperability-of-language-documentation-tools-and-materials-for-local-communities.1.html`.
12. Schalley, A.C.: Tyto – A collaborative research tool for linked linguistic data. In Chiarcos, C., Nordhoff, S., Hellmann, S., eds.: Linked Data in Linguistics. Springer, Heidelberg (2012) 139–149
13. Forkel, R.: The Cross-Linguistic Linked Data project. In: 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing, Reykjavik, Iceland (May 2014) 60–66
14. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating nlp using linked data. In: 12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia. (2013)