Linking Linguistic Resources: Examples from the Open Linguistics Working Group

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff

Abstract The contributions of this part have described recent activities of the OWLG as a whole and of individual OWLG members aiming to provide linguistic resources as Linked Data. Here, we describe how linguistic resources can be linked with each other, and we illustrate possible use cases of information integration from various sources with example queries for the major types of linguistic resources: Using DBpedia (Hellmann et al, this vol.) to represent lexical-semantic resource, the German NEGRA corpus in its POWLA representation (Chiarcos, this vol.) to represent linguistic corpora, the OLiA ontologies to represent repositories of linguistic terminology, and languoid definitions in Glottolog/Langdoc (Nordhoff, this vol.) to represent linguistic knowledge bases and metadata repositories.
We use data from German for illustration purposes, the NEGRA corpus, a linguistically annotated collection of newspaper articles from the Frankfurter Rundschau. The architecture described here, is, however, not specific to German, but can also be applied to other languages. In fact, the Glottolog/Langdoc resources have been developed primarily for language documentation and typological studies, but their applications as described below naturally extends for less-resourced languages.

Christian Chiarcos
Information Sciences Institute, University of Southern California, 4676 Admiralty Way # 1001, Marina del Rey, CA 90292 e-mail: chiarcos@daad-alumni.de

Sebastian Hellmann
Universität Leipzig, Fakultät für Mathematik und Informatik, Abt. Betriebliche Informationssysteme, Johannisgasse 26, 04103 Leipzig, Germany e-mail: hellmann@informatik.uni-leipzig.de

Sebastian Nordhoff
Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany e-mail: sebastian_nordhoff@eva.mpg.de

1 The NEGRA Corpus

This contribution illustrates possible use cases of information integration from various sources, taking data from the NEGRA Corpus as an example.[1] The NEGRA corpus consists of 355,096 tokens (20,602 sentences) of German newspaper text from the Frankfurter Rundschau. The corpus was annotated for parts-of-speech and syntax (Skut et al, 1998), and subsequently, also for coreference and entity types (Schiehlen, 2004). It can thus serve to illustrate the problems of multi-layer annotations in linguistic corpora, further, it can be linked to the German DBpedia, and added to Glottolog/Langdoc as a linguistic resource.

Figure 1 illustrates the first sentence[2] in its original representation, a tab-separated text format. (A graphical visualization of this sentence is provided in Chiarcos, this vol..)

For terminal nodes of the annotation, the first column contains the word, the second a part-of-speech tag, the third morphological annotations, the fourth the label of the edge that connects the terminal with its parent, and the last the ID of the nonterminal parent node. For nonterminal nodes of the annotation, the first column contains the ID, the second the category label, the third is empty, the fourth contains the edge label and the fifth the id of the parent (resp. 0 for the root node).

Traditionally, NEGRA annotations are modified, queried and visualized with specialized tools, e.g., Annotate[3] and Synpathy[4] for syntax annotation, and TIGER-Search (Lezius, 2002) for querying and visualization of syntax annotations. These tools do not, however, allow to annotate and to query over additional layers of annotations, e.g., alignment with parallel corpora, semantic annotations or coreference. For these types of data, further special-purpose tools have been developed, e.g., the Stockholm TreeAligner[5] for aligning parallel corpora annotated with TIGER XML (the XML-based successor of the NEGRA format, König and Lezius, 2000), or SALTO (Burchardt et al, 2006) for the annotation of semantic relations (and also applied to coreference, see Eckart et al, this vol.). These tools are accompanied by other special-purpose formats, e.g., SALSA (Erk and Pado, 2004) for semantic annotations.

As an example for such formats, the coreference annotation of the NEGRA corpus (Schiehlen, 2004) is illustrated in Fig. 3. This is a space-separated text format, where the first column provides the NEGRA sentence id, the second column the token position of a terminal within the the sentence, resp. the id of a NEGRA non-

[1] `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html`

[2] The examples given in this section are taken from public sample of the NEGRA corpus, the syntax sample is available from `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/corpus-sample.export`, the coreference sample is available from `http://www.ims.uni-stuttgart.de/~mike/annotated-negra.txt`.

[3] `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html`

[4] `http://www.lat-mpi.eu/tools/synpathy/`

[5] `http://kitt.cl.uzh.ch/kitt/treealigner`

```
%% word        tag     morph           edge parent gloss
...
die            ART     Def.Fem.Nom.Sg  NK   507    the
Zukunft        NN      Fem.Nom.Sg.⋆    NK   507    future
der            ART     Def.Fem.Gen.Sg  NK   502    of
Musik          NN      Fem.Gen.Sg.⋆    NK   502    music
liegt          VVFIN   3.Sg.Pres.Ind   HD   509    lies
für            APPR    Akk             AC   503    for
viele          PIDAT   ⋆.Akk.Pl        NK   503    many
junge          ADJA    Pos.⋆.Akk.Pl.St NK   503    young
Komponisten    NN      Masc.Akk.Pl.⋆   NK   503    composers
im             APPRART Dat.Masc        AC   504    in.the
Crossover-Stil NN      Masc.Dat.Sg.⋆   NK   504    crossover.style
.              $.      --              --   0
#502           NP      --              GR   507
#503           PP      --              MO   509
#504           PP      --              MO   509
#507           NP      --              SB   509
#509           S       --              --   0
...
```

Fig. 1 First sentence of the NEGRA corpus, original NEGRA format with English glosses added. Translation: '... many young composers believe that the future of music lies in a crossover style.'

```
%% word      tag   morph           edge parent gloss
Sie          PPER  3.Pl.⋆.Nom      SB   504    they
gehen        VVFIN 3.Pl.Pres.Ind   HD   504    enter
gewagte      ADJA  Pos.⋆.Akk.Pl.St NK 500    adventurous
Verbindungen NN    Fem.Akk.Pl.⋆    NK   500    associations
und          KON   --              CD   502    and
Risiken      NN    Neut.Akk.Pl.⋆   CJ   502    risks
ein          PTKVZ --              SVP  504    in
,            $,    --              --   0
```

Fig. 2 Second sentence of the NEGRA corpus, original NEGRA format with English glosses added. Translation: 'They experiment with adventurous associations and take risks, ...'

terminal, the third column information about the discourse status of the referent (e.g., `R` for referring expressions, `R1` first mention of a referring expression) and its id (e.g., `Komponist` for the NEGRA nonterminal corresponding to the phrase *für viele Komponisten* in Fig. 1.

```
%%sentence  negraid    coref
1           503        %R1=Komponist
2           1          %R=Komponist
```

Fig. 3 Coreference annotation of the first two sentences of the NEGRA corpus, original format

With comfortable tools for querying and visualization available for syntax only, the question arises how the additional information about coreference can be integrated with other annotation layers, and how queries can be performed on this data.

2 Multi-Layer Corpora as Linked Data (NEGRA/POWLA Coreference ↦ NEGRA/POWLA Syntax)

In POWLA, NEGRA syntax annotation and coreference annotation can be combined easily as different annotation layers (Chiarcos, this vol.):

```
<!-- syntax "für viele Komponisten" ("for many composers") -->
<powla:Nonterminal rdf:about="s1_503">
   <powla:hasLayer rdf:resource="syntax"/>
   <powla:has_cat>PP</powla:has_cat>
   <powla:hasChild rdf:resource="s1_18"/>
   ...
</powla:Nonterminal>

<!-- syntax "Sie" ("they", sentence 2) -->
<powla:Terminal rdf:about="s2_1">
   <powla:hasLayer rdf:resource="syntax"/>
   <powla:hasString>Sie</powla:hasString>
   <powla:has_pos>PPER</powla:has_pos>
   ...
</powla:Terminal>

<!-- coreference -->
<powla:Relation rdf:about="s2_1_to_s2_530">
   <powla:hasLayer rdf:resource="coref"/>
   <powla:hasSource rdf:resource="s2_1"/>
   <powla:hasTarget rdf:resource="s1_530"/>
</powla:Relation>
```

Using OWL/RDF-based technologies like POWLA, the integration of multiple annotation layers in multi-layer corpora is straight-forward, as previously noticed by Burchardt et al (2008) for the specific case of syntactic and semantic annotations in the SALSA/TIGER corpus. Formally, different layers can be (but do not have to be) stored in different files and actually at different locations, and can thus be viewed as Linked Data.

Using SPARQL, it is thus possible to query across multiple layers at the same time (which would not have been possible with the original formats and the original tools). For example, we can query for personal pronouns (as defined on the syntactic annotation layer) that take prepositional phrases (defined on syntax, again) as their anaphoric antecedent (coreference layer) using the following query:

```
PREFIX powla:<http://purl.org/powla/powla.owl#>
PREFIX negra:<http://purl.org/powla/negra-sample.owl#>
SELECT ?anaphor
WHERE {
    ?anaphor a powla:Node.
    ?anaphor powla:has_pos "PPER".
    ?relation a powla:Relation.
    ?relation powla:hasSource ?anaphor.
    ?relation powla:hasTarget ?antecedent.
    ?relation powla:hasLayer negra:coref.
    ?antecedent powla:has_cat "PP"
}
```

3 Linking Corpora to Metadata Repositories (POWLA ↦ Glottolog/Langdoc)

For the linking of linguistic corpora, we take Glottolog language specifications as an example (Nordhoff, this vol.).

POWLA corpora can be linked with glottolog languoid specifications using the Dublin Core `language` feature.[6] If it is redefined as an `owl:ObjectProperty`, a POWLA Document, a Terminal or a Nonterminal can be defined as being defined for a particular language:[7]

```
<powla:Document rdf:about="http://purl.org/powla/negra-
                                                sample.owl">
  <dcterms:language
     rdf:resource="http://glottolog.livingsources.org/resource/
                                        languoid/id/10077"/>
</powla:Document>
```

On this basis, we can query for POWLA Documents in German using a simple SPARQL query:

```
PREFIX dcterms: <http://purl.org/dc/terms/>.
SELECT ?doc
WHERE {
   ?doc dcterms:language glottolog:10077
}
```

Alternatively, if we don't know the languoid id, we can query for the label:

[6] The description here is somewhat shortened as `dcterms:language` ranges over `dcterms:linguisticSystem`, which is not immediately connected to `glottolog:languoid`. The intermediate steps of the query are not self-evident and will be glossed over here for reasons of space.

[7] Besides Glottolog, other language taxonomies could be applied, e.g., as specified by ISO 639. Glottolog is, however, much more fine-grained than these and captures differentiations that are highly relevant relevant to linguistics, e.g., different dialectal and historical variants of German which are not represented by ISO 639.

```
SELECT ?doc
WHERE {
   ?doc dcterms:language ?languoid.
   ?languoid rdfs:label "German"
}
```

If we know the languoid's ISO 639/3 code,[8] we can query:

```
PREFIX dcterms: <http://purl.org/dc/terms/>.
PREFIX lexvo:   <http://lexvo.org/ontology/>.
SELECT ?doc
WHERE {
  ?doc dcterms:language ?languoid.
  ?languoid lexvo:iso639P3Code "deu"
}
```

In a similar way, other types of metadata can be linked to a linguistic resource, e.g., geographical or historical information.

Glottolog has not been specifically developed for German, instead, it takes a focus on less-resourced languages. However, modeling and querying for corpus resources on, say, the dialects of the indigenous languages of Taiwan documented in the Formosan Languages Archive[9] is analoguous to the treatment of German in this case.

A concrete application of such information for less-resourced languages can be seen, for example, in the context of annotation projection experiments, a flourishing field of Natural Language Processing, where annotated corpora are created on the basis of translated (parallel) text and the assumption that linguistic annotations assigned to word *A* in the source language are also applicable to the target language word *B* that *A* is aligned with. Of course, such experiments benefit from genetic proximity between the language pairs considered, and genetic proximity can be measured by their relative distance in Glottolog/Langdoc classifications or the ASJP tree (Nordhoff, this vol.).[10]

Moreover, Langdoc provides information about the availability of the necessary resources, i.e., translated text (e.g., the Bible), and annotated corpora. By providing this information, and linking to linguistic corpora, Glottolog/Langdoc can thus support the development of NLP resources for languages where such resources are currently not available.

Before starting on an NLP project of Language X, Glottolog/Langdoc allows to check whether sisters (cousins, grandcousins, ...) of language X have already been studied with regard to NLP, and thus jump-start the development.

[8] If the languoid does not have an ISO 639 code, information from super- or sublanguoids can be consulted.

[9] `http://formosan.sinica.edu.tw`

[10] Typological similarity might also be a good indicator for some syntactic questions. This could be retrieved from WALS (Dryer, 2005). The integration of WALS data in the Linguistic Linked Data cloud is currently underway, but not discussed here for reasons of space.

4 Linking Corpora to Terminology Repositories (NEGRA/POWLA ↦ OLiA)

Jsut like linguistic resources can be linked with metadata, annotations can be linked with terminology repositories, so that the semantics of the annotations are represented in an interoperable way. For this purpose, we take the OLiA ontologies as an example (Chiarcos, this vol.). OLiA ontologies provide OWL/DL representations of annotation schemes (OLiA Annotation Model), conventional linguistic terminology (OLiA Reference Model), and that formalize the linking between both.

For the NEGRA corpus, we consider the Annotation Models `stts` for part-of-speech annotation and `tiger-syntax` for syntax annotations according to the TIGER and NEGRA schemes. The Annotation Model defines a taxonomy of linguistic concepts where concrete tags are represented by individuals. For every individual, the property `hasTag`[11] defines its string representation.

For the linking between annotations and OLiA Annotation Models, two strategies are possible: Either we define a property that expresses the linking between annotations and annotation models, or we copy the features of the individual in the Annotation Model to the annotated resource, thereby declaring it an instance of Annotation Model concepts. It should be noted, however, that an annotation may match several individuals, e.g., if tags are composed of multiple independent components, `hasTagMatching` is used in place of `hasTag`. (For example, the `morph` attribute in the NEGRA corpus, every morphological feature – Case, Number, Tense, etc. – is represented by another individual that corresponds to a different substring of the annotation.) In that case, the strategy to copy properties from individuals, rather than to link these individuals, naturally yields an accumulation of all the information expressed in the annotation.

Accordingly, we can define the NEGRA/POWLA entities in terms of OLiA:

```
<powla:Nonterminal rdf:about="s1_503">
  <rdf:type
    rdf:resource="http://purl.org/olia/tiger-syntax.owl#
                                       PrepositionalPhrase"/>
</powla:Nonterminal>

<rdf:Description rdf:about="s1_18">
  <rdf:type
    rdf:resource="http://purl.org/olia/stts.owl#CommonNoun"/>
</rdf:Description>

<rdf:Description rdf:about="s2_1">
  <rdf:type
    rdf:resource="http://purl.org/olia/stts.owl#PersonalPronoun"/>
</rdf:Description>
```

The SPARQL query for PP antecedents of personal pronouns can thus be rephrased as follows:

[11] Aside from `hasTag`, there are also properties for the partial matching of strings, e.g., `hasTagContaining` (partial string), `hasTagMatching` (regular expression), etc.

```
PREFIX negra: <http://purl.org/powla/negra-sample.owl#>.
PREFIX powla: <http://purl.org/powla/powla.owl#>.
PREFIX stts:  <http://purl.org/olia/stts.owl#>.
PREFIX tiger: <http://purl.org/olia/tiger-syntax.owl#>.
SELECT ?anaphor
WHERE {
  ?anaphor a stts:PersonalPronoun.
  ?relation a powla:Relation.
  ?relation powla:hasSource ?anaphor.
  ?relation powla:hasTarget ?antecedent.
  ?relation powla:hasLayer negra:coref.
  ?antecedent a tiger:PrepositionalPhrase.
}
```

As concepts of OLiA Annotation Models are linked by `rdfs:subClassOf` properties to concepts in the OLiA Reference Model, we can infer the corresponding `rdf:type` properties for the annotated POWLA Nodes:

```
PREFIX negra: <http://purl.org/powla/negra-sample.owl#>.
PREFIX powla: <http://purl.org/powla/powla.owl#>.
PREFIX olia:  <http://purl.org/olia/olia.owl#>.
SELECT ?anaphor
WHERE {
  ?anaphor a olia:PersonalPronoun.
  ?relation a powla:Relation.
  ?relation powla:hasSource ?anaphor.
  ?relation powla:hasTarget ?antecedent.
  ?relation powla:hasLayer negra:coref.
  ?antecedent a olia:NounPhrase.
}
```

Despite its name, the OLiA Reference Model does not attempt to establish a terminological consensus; it only serves as an interface between various Annotation Models and other terminology repositories from which it is derived, for example, the GOLD ontology (Farrar and Langendoen, 2003; Chiarcos, 2008) and the ISO TC37/SC4 Data Category Registry (Kemps-Snijders et al, 2008; Chiarcos, 2010). So, this query can also be formulated in terms of these repositories, e.g., by replacing `olia:PersonalPronoun` with `<http://www.isocat.org/datcat/DC-1463>` or `<http://linguistics-ontology.org/gold/PersonalPronoun>`.

In this way, interoperable corpus queries can be formulated, which can be applied to corpora with differing annotation schemes.

5 Linking Terminology Repositories to Metadata Repositories (OLiA ↦ Glottolog/Langdoc)

Like linguistic corpora, also other knowledge bases can be augmented with linguistically relevant metadata: Many OLiA Annotation Models, for example the `stts` and `tiger-syntax` models mentioned above, are specific to a particular corpus, a stage or a particular language, in this case New High German.

This can be expressed, for example, with an axiom that postulates that all instances of the top-level element in an Annotation Model inherit a particular `dcterms:language` property:

```
<owl:Class rdf:about="http://purl.org/olia/stts.owl#Tag">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty rdf:about="http://purl.org/dc/terms/
                                                       language"/>
      </owl:onProperty>
      <owl:hasValue
        rdf:about="http://glottolog.livingsources.org/resource/
                                              languoid/id/10077"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

With a linking to POWLA data as described above, every POWLA Terminal annotated with `stts` individuals can now be inferred to be New High German.

6 Linking Corpora to Lexical-Semantic Resources (NEGRA/POWLA ↦ DBpedia)

As described by Hellmann et al (this vol.), textual data can be automatically enriched with entity links, e.g., to the (German instantiation of the) German DBpedia, e.g., using a NIF-based NLP pipeline. The development of the NLP Interchange Format is synchronized with the development of POWLA. Although both are optimized for different purposes – POWLA is developed to represent annotated corpora with a high degree of genericity, whereas NIF is a compact and NLP-specific format[12] –, they are designed to be mappable. This means that NIF annotations can be converted to POWLA representations, and then, for example, combined with other annotation layers.

[12] One difference is the representation of labeled relations between two entities: In NIF, these are represented as properties, with the ID reflecting the annotation attached to the relation. The NIF modeling requires only a single triple per relation. In POWLA, however, a labeled relation is an individual that is linked by properties to its source and target, and that is assigned its annotation by another property. The POWLA modeling requires at least four triples per relation. Unlike NIF, however, this modeling allows to attach complex annotations to relations, e.g., a direct linking to the OLiA concept hierarchy.
Another difference is that NIF lacks any formalization of corpus structure and annotation layers. More important is that, at the moment, NIF is capable to represent morphosyntactic and syntactic annotations only, the representation of more complex forms of annotation, e.g., alignment in a parallel corpus, has not been addressed so far.
The NIF representation is thus more compact, but the POWLA representation is more precise and more expressive.

As a result, the POWLA individual `negra:s1_18` (*Komponisten* ‘composers’), for example, can be annotated with the corresponding DBpedia concept:

```
<powla:Terminal rdf:about="s1_18">
  <powla:hasString>Komponisten</powla:hasString>
  <scms:means
    rdf:resource="http://de.dbpedia.org/resource/Komponist"/>
  ...
</powla:Terminal>
```

As generated by DBpedia spotlight, this information is attached to POWLA `Terminals` (words), and it can easily be projected further to the corresponding phrases:

```
PREFIX powla: <http://purl.org/powla/powla.owl#>.
PREFIX olia:  <http://purl.org/olia/olia.owl#>.
PREFIX scms:<http://ns.aksw.org/scms/>.
CONSTRUCT { ?np scms:means ?semClass }
SELECT {
    ?term a olia:Noun.
    ?term scms:means ?semClass.
    ?np powla:hasChild ?term.
    ?np a olia:NounHeadedPhrase
}
```

For pronouns with `NounPhrase` or `PrepositionalPhrase` antecedents[13] this information can be used to assign them a semantic class:

```
PREFIX negra: <http://purl.org/powla/negra-sample.owl#>.
PREFIX powla: <http://purl.org/powla/powla.owl#>.
PREFIX olia:  <http://purl.org/olia/olia.owl#>.
PREFIX scms:  <http://ns.aksw.org/scms/>.
CONSTRUCT { ?np scms:means ?semClass }
WHERE {
    ?relation a powla:Relation.
    ?relation powla:hasLayer negra:coref.
    ?relation powla:hasSource ?pronoun.
    ?pronoun a olia:Pronoun.
    ?relation powla:hasTarget ?antecedent.
    ?antecedent scms:means ?semClass.
}
```

By combining information from POWLA, OLiA and DBpedia, we can thus achieve richer semantic annotations for linguistic corpora, that can then be used, for example, to develop NLP applications on this basis, e.g., an anaphor resolution system that takes DBpedia categories into account (Bryl et al, 2010).

[13] `NounHeadedPhrase` is a generalization over `NounPhrase` and `PrepositionalPhrase` that was introduced to account for annotation schemes where both are not properly distinguished.

7 Enriching Lexical-Semantic Resources with Linguistic Information (DBpedia (↦ POWLA) ↦ OLiA)

Unlike classical lexical-semantic resources, DBpedia offers almost no information about the linguistic realization of the entities it contains. Using corpora with entity links and syntactic annotation, however, this information can be easily obtained. The following SPARQL query identifies possible syntactic realizations for concepts of the German DBpedia:

```
PREFIX powla: <http://purl.org/powla/powla.owl#>.
PREFIX olia:  <http://purl.org/olia/olia.owl#>.
PREFIX scms:  <http://ns.aksw.org/scms/>.
CONSTRUCT { ?semClass <#realizedAs> ?syntClass }
WHERE {
  ?x a powla:Node.
  ?x scms:means ?semClass.
  ?x a ?syntClass
  FILTER(regex(str(?syntClass),"http://purl.org/olia/olia.owl#")).
  ?syntClass rdfs:subClassOf olia:MorphosyntacticCategory.
}
```

With the newly generated triples added to the DBpedia, this information about possible grammatical realizations of an entity can be used, for example, to enhance entity-linking algorithms.

8 Enriching Lexical-Semantic Resources with Metadata (DBpedia (↦ POWLA) ↦ Glottolog)

Similarly, lexical-semantic resources can be enriched with metadata. With fine-grained languoid definitions as provided by Glottolog, and corpora representing different historical stages of a language, for example, the historical development of terms can be extrapolated (when and where was a term recorded).

The following SPARQL query assigns a concept a `dcterms:language` property for every languoid for which it is found in a particular corpus:

```
PREFIX powla:   <http://purl.org/powla/powla.owl#>.
PREFIX dcterms: <http://purl.org/dc/terms/>.
PREFIX scms:    <http://ns.aksw.org/scms/>.
CONSTRUCT {?semClass dcterms:language ?language}
WHERE {
    ?word a powla:Node.
    ?word dcterms:language ?language.
    ?word scms:means ?semClass.
}
```

This query presupposes that entity linking has been performed on the corpus before. On a historical or dialectal corpus, entity linking is possible only if its spelling follows the same conventions as modern standard orthography for the respective language. However, dialectal and historical corpora often include lemmas in modern languages, and on these ‘hyperlemmas’ (Dipper et al, 2004), a standard entity linking routine can be applied.

9 Enriching Metadata Repositories with Linguistic Features (Glottolog ↦ OLiA)

Finally, one may consider also to enrich metadata repositories with linguistic features, e.g., to record which languoid makes use of which linguistic categories and features.

On the basis of the resources described before, this can be extrapolated from annotations in a languoid-annotated corpus.[14] The following query retrieves all syntactic categories that are used for a particular Glottolog languoid (given a set of corpora to which this query is applied):

```
PREFIX dcterms:  <http://purl.org/dc/terms/>.
PREFIX powla:    <http://purl.org/powla/powla.owl#>.
PREFIX olia:     <http://purl.org/olia/olia.owl#>.
PREFIX rdfs:     <http://www.w3.org/2000/01/rdf-schema#>.
CONSTRUCT { ?languoid <#uses> ?syntacticCategory }
WHERE {
  ?node dcterms:language ?languoid
  FILTER(regex(str(?languoid),"http://glottolog.livingsources.
                                 org/resource/languoid/id/.*")).
  ?node a powla:Node.
  ?node a ?syntacticCategory
  FILTER(regex(str(?syntacticCategory),
               "http://purl.org/olia/olia.owl#.*")).
  ?syntacticCategory rdfs:subClassOf olia:SyntacticCategory.
}
```

On this basis, then, one may study to what extent genealogical relationships correspond to certain syntactic features (as far as reflected in the underlying resources). For instance, one might build a reasoner which asserts the existence of a grammatical category to a `glottolog:superlanguoid` if all its sublanguoids happen to have this particular property. For instance, the category ‘Preposition’ is found in

[14] It should be noted that this approach is *approximative* only, because it considers only information expressed in annotations. If is possible that the underlying schemes make a number of simplifying assumptions, e.g., not to distinguish two functionally different categories that appear superficially and that cannot be unambiguously distinguished by NLP tools or human annotators. Greater precision could probably be achieved if such queries are applied to language-annotated lexicons that make use of a standard vocabulary to represent detailed grammatical information, as created, for example, in the context of the LEGO project (Poornima and Good, 2010) whose lexicons are linked to the GOLD ontology (Farrar and Langendoen, 2003). The queries necessary for this purpose would be, however, almost identical.

corpora of German, Dutch, English, and all other Germanic languages. Such a category can therefore be posited on the family level. Postpositions on the other hand are only found in a subset of the Germanic languages and thus do not 'climb up the tree' as high as their prenominal brethren.

If knowledge bases are provided that provide other metrics of language relatedness (e.g., ASJP, Nordhoff, this vol.), it can be tested whether these metrics correspond to the occurrence of similar grammatical features. The Linked Data approach furthermore allows to map nodes of different trees to each other. Computation of consensus trees from trees based on different datasets is another possibility.

10 Outlook

We illustrated how a Linguistic Linked Open Data cloud can be created, and what possible gains of information would be possible. The resources described in this part and their possible linking are summarized in Fig. 4.

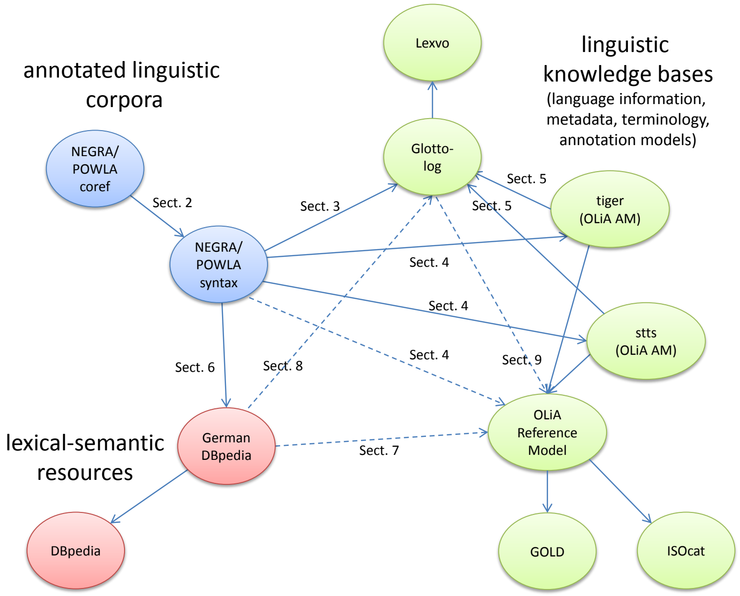

Fig. 4 Linguistic resources and possible links between them as described in this contribution.

It should be noted that the resources considered here only serve illustrative purposes, and they are chosen such that they represent major types of linguistic resources: lexical-semantic resources, linguistic corpora and repositories of metadata and linguistic terminology. Various members of the OWLG are engaged in related efforts, and we expect that these converge in the creation of a Linguistic Linked Open Data cloud as described here. As of December 2011, a number of OWLG members have expressed their willingness to provide data for such a Linguistic Linked Open Data cloud, aside from the resources shown in Fig. 4 this includes RDF formalizations of WordNet and Wiktionary (e.g., those described by McCrae et al, this vol.) and other lexical-semantic resources, typological metadata repositories (e.g., those described by Moran, this vol.), additional corpora, and multi-lingual word lists.

These novel resources are complemented by linguistic resources already present in the Linked Open Data cloud,[15] e.g., meta data repositories such as Lexvo,[16] lingvoj,[17] GeoNames,[18] Project Gutenberg,[19] the OpenLibrary;[20] and lexical-semantic resources such as Cornetto,[21] Freebase,[22] OpenCyc,[23] the Open Data Thesaurus,[24] YAGO,[25] and WordNet.[26] Additionally, the OWLG has compiled an extensive list of resources that represent candidates to be included in the LLOD and that are available under open licenses, but that have not yet been converted to RDF.

It is our hope that the workshop on Linked Data in Linguistics (LDL-2012) and this volume contribute to the on-going formation of an interdisciplinary community actively working towards the application of the Linked Open Data paradigm to all forms of linguistic resources. In the last few years, interested researchers in different sub-communities have begun to organize themselves. One example is the Open Linguistics Working Group,[27] founded in 2009, who organized LDL-2012, another one is the W3C Ontology-Lexica Community Group, founded in 2011.[28] At the same time, established standardization initiatives from the fields of Natural Language Processing and computational lexicography, e.g., the ISO TC37/SC4, adopt ideas and formalisms developed in the context of the Semantic Web (e.g., Windhouwer and Wright, this vol.; Pareja-Lora, this vol.). In the fields of language documentation, typology and in the humanities in general (e.g., Bouda and Cysouw, this vol.; Declerck et al, this vol.; Schalley, this vol.), Linked Data approaches seem to be gaining popularity.

[15] http://richard.cyganiak.de/2007/10/lod

[16] http://www.lexvo.org

[17] http://www.lingvoj.org

[18] http://www.geonames.org/ontology

[19] http://www4.wiwiss.fu-berlin.de/gutendata

[20] http://openlibrary.org

[21] http://www2.let.vu.nl/oz/cltl/cornetto

[22] http://freebase.com

[23] http://sw.opencyc.org

[24] http://vocabulary.semantic-web.at/PoolParty/wiki/OpenData

[25] http://mpii.de/yago

[26] http://semanticweb.cs.vu.nl/lod/wn30, http://www.w3.org/TR/wordnet-rdf, http://wordnet.rkbexplorer.com

[27] linguistics.okfn.org

[28] http://www.w3.org/community/ontolex

In this chapter, we have discussed some first nodes of a Linguistic Linked Data Cloud. We have also discussed how links between these nodes can be established. The wide range of topics covered in this volume as well as the commitment shown by scholars from very different subdisciplines of linguistics to render their data interoperable make us very optimistic that this network will quickly grow and that the coverage of the LLD cloud as well as its density will significantly increase in the very near future.

References

Bouda P, Cysouw M (this vol.) Treating dictionaries as a Linked-Data corpus. P. 15-23

Bryl V, Giuliano C, Serafini L, Tymoshenko K (2010) Supporting natural language processing with background knowledge: coreference resolution case. The Semantic Web–ISWC 2010 pp 80–95

Burchardt A, Erk K, Frank A, Kowalski A, Pado S (2006) SALTO: A versatile multi-level annotation tool. In: Proc. LREC-2006, Genoa, Italy

Burchardt A, Padó S, Spohr D, Frank A, Heid U (2008) Formalising Multi-layer Corpora in OWL/DL – Lexicon Modelling, Querying and Consistency Control. In: Proceedings of the 3rd International Joint Conference on NLP (IJCNLP 2008), Hyderabad

Chiarcos C (2008) An ontology of linguistic annotations. LDV Forum 23(1):1–16

Chiarcos C (2010) Grounding an ontology of linguistic annotations in the Data Category Registry. In: LREC 2010 Workshop on Language Resource and Language Technology Standards (LT&LTS), Valetta, Malta, pp 37–40

Chiarcos C (this vol.) Interoperability of corpora and annotations. P. 161-179

Declerck T, Lendvai P, Mörth K, Budin G, Váradi T (this vol.) Towards Linked Language Data for Digital Humanities. P. 109-116

Dipper S, Faulstich L, Leser U, Lüdeling A (2004) Challenges in modelling a richly annotated diachronic corpus of german. In: Workshop on XML-based richly annotated corpora, Lisbon, Portugal, pp 21–29, URL `http://www.informatik.hu-berlin.de/Forschung\_Lehre/wbi/publications/2004/xbrac04\_final.pdf`

Dryer MS (2005) Genealogical language list. In: Comrie B, Dryer MS, Gil D, Haspelmath M (eds) World Atlas of Language Structures, Oxford University Press, pp 584–644

Eckart K, Riester A, Schweitzer K (this vol.) A discourse information radio news database for linguistic analysis. P. 65-75

Erk K, Pado S (2004) A powerful and versatile XML format for representing role-semantic annotation. In: Proc. Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal

Farrar S, Langendoen DT (2003) A Linguistic Ontology for the Semantic Web. GLOT International 7:97–100

Hellmann S, Stadler C, Lehmann J (this vol.) The German DBpedia: A sense repository for linking entities. P. 181-189

Kemps-Snijders M, Windhouwer M, Wittenburg P, Wright S (2008) ISOcat: Corralling data categories in the wild. In: Proc. LREC 2008, Marrakech, Morocco

König E, Lezius W (2000) A description language for syntactically annotated corpora. In: Proc. 18th International Conference on Computational Linguistics (COLING 2000), Saarbrücken, Germany, pp 1056–1060

Lezius W (2002) TIGERSearch. Ein Suchwerkzeug für Baumbanken. In: Proceedings of the 6. Konferenz zur Verarbeitung natürlicher Sprache (6th Conference on Natural Language Processing, KONVENS 2002), Saarbrücken, Germany

McCrae J, Montiel-Ponsoda E, Cimiano P (this vol.) Integrating WordNet and Wiktionary with *lemon*. P. 25-34

Moran S (this vol.) Using Linked Data to create a typological knowledge base. P. 129-138

Nordhoff S (this vol.) Linked Data for linguistic diversity research: Glottolog/Langdoc and ASJP. P. 191-200

Pareja-Lora A (this vol.) OntoLingAnnot's ontologies: Facilitating interoperable linguistic annotations (up to the pragmatic level). P. 117-127

Poornima S, Good J (2010) Modeling and encoding traditional wordlists for machine applications. In: Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, Association for Computational Linguistics, pp 1–9

Schalley AC (this vol.) TYTO – A collaborative research tool for linked linguistic data. P. 139-149

Schiehlen M (2004) Optimizing algorithms for pronoun resolution. In: Proc. 20th International Conference on Computational Linguistics (COLING), Geneva, pp 515–521

Skut W, Brants T, Krenn B, Uszkoreit H (1998) A linguistically interpreted corpus of German newspaper text. In: Proc. ESSLLI Workshop on Recent Advances in Corpus Annotation, Saarbrücken, Germany

Windhouwer M, Wright SE (this vol.) Linking to linguistic data categories in ISOcat. P. 99-107