Interoperability of Corpora and Annotations

Christian Chiarcos

Abstract This paper describes the application of OWL and RDF to address the interoperability of linguistic corpora and linguistic annotations within such corpora. Interoperability of linguistic corpora involves two aspects: Structural interoperability (annotations of different origin are represented using the same formalism) and conceptual interoperability (annotations of different origin are linked to a common vocabulary).

Building on an infrastructure developed to represent, to store, to query and to visualize multi-layer corpora with any kind of text-oriented annotation (Chiarcos et al, 2008), this paper proposes to address both aspects by means of OWL/RDF-based formalisms. Key advantages of this approach include the existence of a rich technological ecosystem developed around RDF and OWL, the conceptual similarity of generic data models for linguistic annotations and RDF (both based on labeled directed graphs), and the application of OWL/DL reasoners that can be applied to validate the consistency of linguistic corpora and their annotations and to infer additional information that is relevant, for example, for their appropriate visualization. Additionally, representing corpora in OWL and RDF allows to interlink resources freely, e.g., different annotation layers of a multi-layer corpus, translated texts in parallel corpora, or linguistic corpora and lexical-semantic resources. Modeled in this way, corpora can be fully integrated in a Linked Open Data (sub-)cloud of linguistic resources, along with lexical-semantic resources and knowledge bases of information about languages and linguistic terminology.

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

1 Motivation and Background

In recent years, the interoperability of linguistic resources has become a major topic in the fields of computational linguistics and Natural Language Processing (Ide and Pustejovsky, 2010): Within the last 30 years, the maturation of language technology and the increasing importance of corpora in linguistic research produced a growing number of linguistic corpora with increasingly diverse annotations. While the earliest annotations focused mostly on part-of-speech and syntax annotation, later NLP research included also on semantic, anaphoric and discourse annotations, and with the rise of statistic MT, a large number of parallel corpora became available. In parallel, specialized technologies were developed to represent these annotations, to perform the annotation task, to query and to visualize them. Yet, the tools and representation formalisms applied were often specific to a particular type of annotation, and they offered limited possibilities to combine information from different annotation layers applied to the same piece of text. Such multi-layer corpora became increasingly popular,[1] and, more importantly, they represent a valuable source to study interdependencies between different types of annotation. For example, the development of a semantic parser usually takes a syntactic analysis as its input, and higher levels of linguistic analysis, e.g., coreference resolution or discourse structure, may take both types of information into consideration. Such studies, however, require that all types of annotation applied to a particular document are integrated into a common representation that provides lossless and comfortable access to the linguistic information conveyed in the annotation without requiring too laborious conversion steps in advance.

This has been one motivation for research on interoperability between different types of annotation. Another motivation was that different NLP tools for the same linguistic phenomenon, say, syntactic parsers, produce different output formats, and that comparative evaluations as well as ensemble combination architectures require a interoperable representation of the respective analyses (Pareja-Lora and Aguado de Cea, 2010; Chiarcos, 2010b). And with the development of complex NLP pipeline systems, people became interested in interchangeable pipeline modules, where one module can be replaced by another, equivalent module, for example, if domain-specific parsers are to be used if texts from their particular domain are to be analyzed. This would be possible, however, only if these modules make use of the same input and output representations, and if they refer to a common vocabulary of linguistic categories (Buyko et al, 2008).

In this paper, I focus on **interoperability of linguistic corpora**. This is closely related to interoperability in NLP pipelines (cf. Hellmann et al, this vol.), but differs in two crucial aspects: (1) In an NLP pipelines, annotations can be created on-the-fly, and do not necessarily have to be preserved throughout the entire pipeline. It is therefore not necessary to formally distinguish different layers of annotation, to

[1] For example, parts of the Penn Treebank (Marcus et al, 1993), originally annotated for parts-of-speech and syntax, were later annotated with nominal semantics, semantic roles, time and event semantics, discourse structure and anaphoric coreference (Pustejovsky et al, 2005).

represent the macrostructure of linguistic corpora and metadata. (2) NLP pipelines are usually created for one particular task or a set of tasks. The number of possible annotation layers is thus limited by plausibility considerations.[2] For linguistic corpora, however, the number and the types of annotations applied to a particular text are in principle unlimited, because researchers may have an interest to preserve and to integrate all available legacy annotations created throughout the entire lifetime of a particular corpus. In that sense, corpus interoperability is a more general and harder problem than NLP interoperability.

At the moment, state-of-the-art approaches on **structural interoperability** of linguistic corpora build on standoff-XML (Carletta et al, 2005; Ide and Suderman, 2007; Bouda and Cysouw, this vol.) and relational data bases (Chiarcos et al, 2008; Eckart et al, this vol.). The underlying data models are, however, graph-based, and Sect. 2 pursues the idea that RDF and RDF data bases can be applied for the task to represent all possible annotations of a corpus in an interoperable way, to integrate their information without any restrictions (as imposed, for example, by conflicting hierarchies or overlapping segments in an XML-based format), and to provide means to store and to query this information regardless of the annotation layer from which it originates. Using OWL/DL defined data types as the basis of this RDF representation allows to specify and to verify formal constraints on the correct representation of linguistic corpora in RDF. POWLA, the approach presented here, formalizes data models for generic linguistic data structures for linguistic corpora as OWL/DL concepts and definitions (POWLA TBox) and represents the data as OWL/DL individuals in RDF (POWLA ABox).

The heterogeneity of linguistic annotations has long been recognized as a key problem limiting the reusability of NLP tools and linguistic data collections, and it is generally agreed that repositories of linguistic annotation terminology represent a key element in the establishment of **conceptual interoperability**. With a terminological reference repository, it is possible to overcome the heterogeneity of annotation schemes: Reference definitions provide an interlingua that allows mapping linguistic annotations from annotation scheme *A* to annotations in accordance with scheme *B*. Several repositories of linguistic annotation terminology have been developed by the NLP/computational linguistics community (Leech and Wilson, 1996; Aguado de Cea et al, 2004; Pareja-Lora, this vol.) as well as in the field of language documentation/typology (Bickel and Nichols, 2002; Saulwick et al, 2005), and their continuous application is expected to enhance the consistency of linguistic metadata and annotations. The General Ontology of Linguistic Description (Farrar and Langendoen, 2003, GOLD) and the ISO TC37/SC4 Data Category Registry (Kemps-Snijders et al, 2009; Windhouwer and Wright, this vol., ISOcat) address both communities.

At the moment, however, a problems for the practical application of these terminology repositories persists with the fact that different communities develop and maintain terminology repositories – e.g., GOLD and ISOcat – independently, and these repositories are not always compatible with respect to the definitions they pro-

[2] In the general case, for example, one would expect that there is only a single syntactic analysis performed, but not several parses whose input is integrated.

vide,[3] with respect to the technologies employed,[4] or with respect to the underlying philosophy.[5] Researchers are aware of the problem and actively addressing it, e.g., by providing ISOcat in OWL (like GOLD, cf. Windhouwer and Wright, this vol.) or by integrating GOLD categories as a separate profile in ISOcat (Kemps-Snijders, 2010). Section 3 describes the Ontologies of Linguistic Annotation (OLiA ontologies) that introduce an intermediate level of representation between ISOcat, GOLD and other repositories of linguistic reference terminology in order to facilitate the development of applications and resources that take benefit of a well-defined terminological backbone even **before** the GOLD and ISOcat repositories have converged into a uniform and generally accepted reference terminology.

A novel element in this approach is that conceptual interoperability and structural interoperability are addressed with the same formalism (OWL/RDF). In earlier approaches, e.g., Chiarcos et al (2008), these were treated independently, and conceptual interoperability was established through a software-mediated mapping between annotations and ontologies. This paper describes a fully declarative approach.

Moreover, by using RDF as representation format, both resources described here, POWLA corpora and OLiA ontologies, can be integrated with resources already available as Linked Data, e.g., meta data repositories such as Glottolog/Langdoc (Nordhoff, this vol.), general-purpose knowledge bases like the DBpedia (Hellmann et al, this vol.) or full-fledged lexical-semantic resources such as the Wikipedia and WordNet (McCrae et al, this vol.).

2 Addressing Structural Interoperability

POWLA is an OWL/DL-based formalism to represent linguistic corpora in an interoperable way. As compared to earlier approaches in this direction (Burchardt et al, 2008; Hellmann et al, 2010), POWLA is not tied to a specific selection of annotation layers, or a specific annotation scheme. Instead, it is designed to support any kind of text-oriented annotation.

[3] As one example, the original GOLD Numeral was a Determiner (Numeral $\sqsubseteq$ Quantifier $\sqsubseteq$ Determiner, http://linguistics-ontology.org/gold/2009/Numeral), whereas a ISOcat Numeral (DC-1334) is defined on the basis of its semantic function, without any references to syntactic categories. Thus, *two* in *two of them* may be a ISOcat Numeral but not a GOLD Numeral. Following a suggestion of the author, the current GOLD version (http://linguistics-ontology.org/gold/Numeral) adopted the ISOcat modeling, but as GOLD and ISOcat implement community processes on different communities, the establishment of parallel definitions takes a considerable time, if it is possible at all.

[4] GOLD is based on OWL/RDF, a formalization in OWL/DL is possible, it thereby supports the full power of description logics (i.e. a decidable fragment of first-order predicate logic). ISOcat, however, is designed as a semistructured list of concepts, with only optional elements of hierarchical organization.

[5] GOLD aims to provide a holistic and unified representation of linguistic reference concepts. ISOcat is an extensible collection of annotation categories.

The idea underlying POWLA is to represent linguistic annotations by means of RDF, to employ OWL/DL to define data types and consistency constraints for these RDF data, and to adopt these data types and constraints from an existing representation formalism applied for the loss-less representation of arbitrary kinds of text-oriented linguistic annotation within a generic exchange format. POWLA is designed as an OWL/DL implementation of the PAULA data model (Dipper, 2005; Chiarcos et al, 2008, 2011) developed at the Collaborative Research Center (SFB) 632 “Information Structure” at the University of Potsdam, Germany. At the moment, the standard linearization of PAULA is an XML standoff format that originates from early drafts of the Linguistic Annotation Framework (Ide and Romary, 2004), and it is thus closely related to the later ISO TC37/SC4 format GrAF (Ide and Suderman, 2007). With POWLA as an OWL/DL linearization of the PAULA data model, all annotations currently covered by PAULA (presumably every kind of text-oriented linguistic annotation) can be represented as part of the Linguistic Linked Open Data cloud.

2.1 PAULA Data Types

PAULA implements the insight that any kind of linguistic annotation can be represented by means of **directed (acyclic) graphs** (Bird and Liberman, 2001; Ide and Suderman, 2007), i.e. the basic triple structure underlying RDF: Aside from the primary data (text), linguistic annotations consist of three principal components, i.e. segments (spans of text, e.g. a phrase, modeled as nodes), relations between segments (e.g. dominance relation between two phrases, modeled as edges) and annotations that describe different types of segments or relations (modeled as labels).

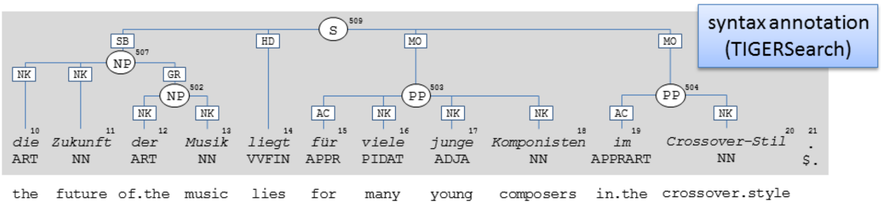

Fig. 1 Constituent syntax: Example from the NEGRA corpus (Skut et al, 1998), visualization based on TIGERSearch (Lezius, 2002).

PAULA data types relevant for linguistic annotations are the following (see Fig. 2 for the German phrase *für viele junge Komponisten* ‘for many young composers’ as an illustration, for the original syntax visualization see Fig. 1):

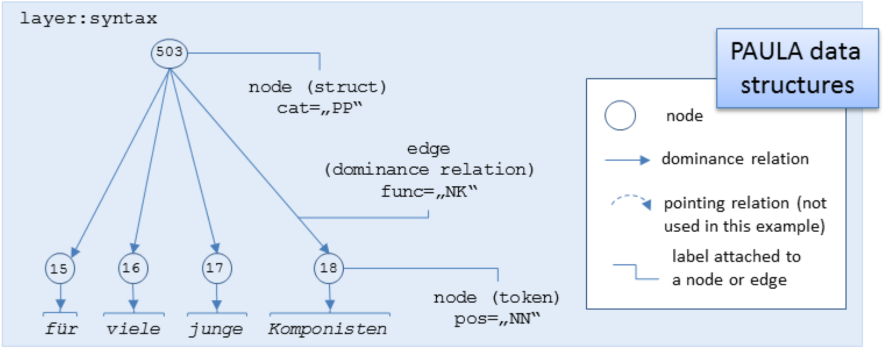

Fig. 2 Constituent syntax: Fragment from Fig. 1 with PAULA data structures.

node (structural units of annotation)
token character spans in the primary data
markable span of tokens (data structure of flat, layer-based annotations defined with respect to, e.g. a timeline)
struct hierarchical data structure forming DAGs (e.g. trees) by establishing parent-child relations between a struct (parent) and tokens, markables or other structs.
edge (relational unit of annotation, connecting nodes)
dominance relation directed edge between a struct and its children, coverage inheritance (see below)
pointing relation general directed edge, no coverage inheritance
label (attached to nodes or edges)
features linguistic annotations

A unique feature of PAULA is the differentiation of two types of edges with respect to their relationship to the primary data. For hierarchical structures, e.g. phrase structure trees, a notion of **coverage inheritance** is necessary (the text covered by a child node is always covered by the parent node). In PAULA, such edges are referred to as **dominance relations**. For other kinds of relational annotation, no constraints on the coverage of the elements connected needs to be postulated (e.g. anaphoric relations, alignment in parallel corpora, dependency analyses), and source and target of a relation may or may not overlap at all. In PAULA, edges without coverage inheritance are referred to as **pointing relations**. This distinction does not constrain the generic character of PAULA (a general directed graph would just use pointing relations), but it captures a fundamental distinction of linguistic data types. As such, it was essential for the development of convenient means of visualization and querying of PAULA data: For example, the appropriate visualization (hierarchical or relational) within a corpus management system can be chosen on the basis of the data structures alone, and it does not require any external specifications (Chiarcos et al, 2008).

Additionally, PAULA includes specifications for the organization of annotations, which is, however, skipped here for reasons of space.

2.2 The POWLA Ontology (POWLA TBox)

The POWLA ontology represents a straight-forward implementation of the PAULA data types in OWL/DL. `Document`, `Relation`; `Node` and `Layer` are defined as subclasses of `POWLAElement`. Here, we concentrate on `Node` and `Relation`, `Document` and `Layer` are more important for corpus organization.

A `POWLAElement` is anything that can carry a label (property `hasLabel`). For `Node` and `Relation`, `hasLabel` contains string values of linguistic annotation (subproperty `hasAnnotation`). The property `hasAnnotation` are, however, not to be used directly, but rather, subproperties are to be created for every annotation phenomenon, e.g. `has_pos` for part-of-speech annotation.

A `Node` is a `POWLAElement` that covers a stretch of primary data. It can carry `hasChild` properties (and the inverse `hasParent`) that express coverage inheritance. A `Relation` is another `POWLAElement` that is used for every edge that carries an annotation. The properties `hasSource` and `hasTarget` (resp. the inverse `isSourceOf` and `isTargetOf`) assign a `Relation` source and target node. Dominance relations are relations whose source and target are connected by `hasChild`, pointing relations are relations where source and target are not connected by `hasChild`. It is thus not necessary to distinguish pointing relations and dominance relations as separate concepts in the POWLA ontology.

Two basic subclasses of `Node` are distinguished: A `Terminal` is a `Node` that does not have a `hasChild` property. It corresponds to a “token” in PAULA. A `Nonterminal` is a `Node` that has at least one `hasChild` property. The differentiation between PAULA struct and markable can be inferred and is therefore not explicitly represented in the ontology: A struct is `Nonterminal` that has another `Nonterminal` as its child, or that is connected to at least one of its children by means of a (dominance) `Relation`, any other `Nonterminal` corresponds to a PAULA markable.

The concept `Root` was introduced for organizational reasons. It corresponds to a `Nonterminal` that does not have a parent, i.e. the top-level node within an annotation layer (and may be either a `Terminal` or a `Nonterminal`).

Both `Terminals` and `Nonterminals` are characterized by a string value (property `hasString`), and a particular position (properties `hasStart` and `hasEnd`) with respect to the primary data. `Terminals` are further connected with each other by means of `nextTerminal` properties. This is, however, a preliminary solution. Forthcoming versions of POWLA may address `Nonterminals` by means of pre- and post-order as defined by Trißl and Leser (2007), and `Terminals` may be linked to strings in accordance to the NLP Interchange Format NIF (cf. Hellmann et al, this vol.).

2.3 Modelling Linguistic Annotations in POWLA (POWLA ABox)

The POWLA ontology defines data types that can now be used to represent linguistic annotations. Figure 3 shows the `Nonterminal` created for the phrase *für viele junge Komponisten* ‘for many young composers’, the `Terminal` for *Komponisten* ‘composers’, and the `Relation` between them.

```
<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>

<powla:Relation rdf:about="s1_503_to_s1_18">
    <powla:hasLayer rdf:resource="syntax"/>
    <powla:has_func>NK</powla:has_func>
    <powla:hasSource rdf:resource="s1_503"/>
    <powla:hasTarget rdf:resource="s1_18"/>
</powla:Relation>

<powla:Terminal rdf:about="s1_18">
    <powla:hasLayer rdf:resource="syntax"/>
    <powla:hasString>Komponisten</powla:hasString>
    <powla:has_pos>NN</powla:has_pos>
    <powla:nextTerminal rdf:resource="s1_19"/>
    <powla:startPosition>103</powla:startPosition>
    <powla:endPosition>113</powla:endPosition>
</powla:Terminal>
```

Fig. 3 Constituent syntax: A `Nonterminal`, a `Terminal`, and the `Relation` between them in POWLA, fragment of Fig. 2.

The `Nodes` are taken from the German sentence analyzed in Fig. 1. The `Node` IDs preserve the original identifiers used in the NEGRA corpus (Skut et al, 1998), with `s1_18` corresponding to the 18th word, and `s1_503` corresponding to the phrase with ID 503 in sentence 1.[6] The `Relation` ID is derived from the IDs of the source and the target node. The properties `has_pos`, `has_cat` and `has_func` are subproperties of `hasAnnotation` that have been created to reflect the `pos`, `cat` and `func` labels of nodes and edges in Fig. 1. `Relation s1_503_to_s1_18` is marked as a dominance relation by the accompanying `hasChild` relation between its source and target.

It should be noted that the RDF representation given in Fig. 3 is by no means complete. Inverse properties, for example, are missing, e.g., the `hasParent` rela-

[6] The data was not directly converted from the NEGRA format, but through TIGER XML. These naming conventions, as well as those of the attribute names (`func` instead of `edge`, `pos` instead of `tag`) originate from the converter integrated in the TIGERRegistry (Lezius, 2002).

tion between `s1_18` and `s1_503`. Using a reasoner, however, the missing RDF triples can be inferred from the information provided explicitly. This inference mechanism can be also be applied to infer $\sqsubseteq$ (`rdfs:subClassOf`) relationships (e.g., that `s1_18` is not only a `Terminal`, but also a `Node`), and the transitive closure of relations (if the corresponding transitivity axioms are added). Further, functional differentiations can be inferred, e.g., the difference between PAULA markables and structs. This is not relevant for processing, but for visualization only, and therefore explicitly represented in PAULA XML. In POWLA, it can be inferred whether a `Layer` requires a spreadsheet-like visualization ('grid view', applicable to PAULA markable layers, i.e., a POWLA `Layer` with no recursive structures and no labeled dominance relations), or as a directed acyclic graph (e.g., a tree, applicable to PAULA struct layers, i.e., a POWLA `Layer` with recursion or labeled dominance relations). Moreover, a reasoner would also allow us to verify whether the necessary cardinality constraints are respected, e.g., every `Relation` has exactly one `hasSource` and one `hasTarget` relation etc.

Although illustrated here for syntax annotations only, other annotation layers represented in PAULA XML can be equally easily transformed to POWLA. Using existing converters that generate PAULA XML, e.g., Salt'N'Pepper (Zipser and Romary, 2010), a broad variety of input formats from various tools are supported, including TIGER XML (König and Lezius, 2000, constituent and dependency syntax), MMAX2 (Müller and Strube, 2006, coreference annotation), EXMARaLDA (Schmidt, 2004, transcriptions and layer-based annotation), RSTTool (O'Donnell, 2000, discourse structure annotation), and Toolbox (Busemann and Busemann, 2008, typological glosses).[7]

Like PAULA XML, POWLA represents these different kinds of annotations in an interoperable way, but here on the basis of OWL and RDF. One important difference between PAULA XML and POWLA is that POWLA data can be directly fed into an RDF triple store, whereas current data base solutions for PAULA involve relational data bases and the transformation into a proprietary table format (Zeldes et al, 2009). Another important difference is that all IDs used in the POWLA representation are URIs, with an XML base name as specified in the file they are contained in. Accordingly, these URIs can be referred to from external resources, e.g., from scientific papers (as stable references to corpus examples) or from lexical-semantic resource (as examples to illustrate a specific semantic role). And of course, POWLA corpora can be augmented with references to other resources available as Linked Data, e.g., terminology repositories as described in the following section, or meta data repositories as, for example, described by Nordhoff (this vol.).

[7] Through existing converters that produce these formats, numerous additional formats are supported, e.g., Penn Treebank syntax annotations (Marcus et al, 1994, through TIGER XML), Praat (Boersma, 2002, through EXMARaLDA), ELAN (Hellwig et al, 2008, through EXMARaLDA), etc. Additionally, PAULA XML wrappers for a number of NLP tools exist, which cover part-of-speech tagging, parsing, anaphor resolution, connective classification and discourse parsing, as applied, for example, in the text summarization pipeline described by Stede et al (2006).

3 Addressing Conceptual Interoperability

The Ontologies of Linguistic Annotation (OLiA) are a repository of annotation terminology for various linguistic phenomena on a great band-width of languages. In combination with RDF-based formats like POWLA (Sect. 2) and NIF (Hellmann et al, this vol.), the OLiA ontologies allow to represent linguistic annotations and their meaning within the Linguistic Linked Open Data cloud in an interoperable way.

3.1 Towards Conceptual Interoperability of Linguistic Annotations

The OLiA ontologies introduce an intermediate level of representation between ISOcat, GOLD and other repositories of linguistic reference terminology and are interconnected with these resources, and they provide not only a means to formalize reference categories, but also annotation schemes, and the way that these are linked with reference categories.

3.2 A Modular Architecture of OWL/DL Ontologies

The Ontologies of Linguistic Annotations – briefly, OLiA ontologies (Chiarcos, 2008) – represent a modular architecture of OWL/DL ontologies that formalize several intermediate steps of the mapping between annotations, a 'Reference Model' and existing terminology repositories ('External Reference Models').

The OLiA ontologies were developed as part of an infrastructure for the sustainable maintenance of linguistic resources (Schmidt et al, 2006), and their primary fields of application include the formalization of annotation schemes and concept-based querying over heterogeneously annotated corpora (Rehm et al, 2007; Chiarcos et al, 2008).

In the OLiA architecture, four different types of ontologies are distinguished:

- The OLiA REFERENCE MODEL specifies the common terminology that different annotation schemes can refer to. It is derived from existing repositories of annotation terminology and extended in accordance with the annotation schemes that it was applied to.
- Multiple OLiA ANNOTATION MODELs formalize annotation schemes and tagsets. Annotation Models are based on the original documentation of an annotation scheme, they provide an interpretation-independent representation.
- For every Annotation Model, a LINKING MODEL defines `rdfs:subClassOf` ($\sqsubseteq$) relationships between concepts/properties in the respective Annotation Model and the Reference Model. Linking Models are interpretations of Annotation Model concepts and properties in terms of the Reference Model.

- Existing terminology repositories can be integrated as External Reference Models, if they are represented in OWL/DL. Then, Linking Models specify $\sqsubseteq$ relationships between Reference Model concepts and External Reference Model concepts.

The OLiA Reference Model specifies classes for linguistic categories (e.g. `olia:Determiner`) and grammatical features (e.g. `olia:Accusative`), as well as properties that define relations between these (e.g. `olia:hasCase`). Far from being yet another annotation terminology ontology, the OLiA Reference Model does not introduce its own view on the linguistic world, but rather, it is a derivative of the EAGLES recommendations(Leech and Wilson, 1996), MULTEXT/East (Erjavec, 2004), and GOLD (Farrar and Langendoen, 2003) that was introduced as a technical means to allow to interpret linguistic annotations with respect to these terminological repositories and extended with respect to the annotation schemes linked with it. These extensions are also further communicated to the communities behind GOLD and ISOcat. The Reference Model specifies for example that `olia:PrepositionalPhrase` $\sqsubseteq$ `olia:NounHeadedPhrase`.[8]

Annotation Models differ conceptually from the Reference Model in that they include not only concepts and properties, but also individuals: Individuals represent concrete tags, while classes represent abstract concepts similar to those of the Reference Model. As an example, consider the tag `PP` from the syntax annotation the NEGRA corpus Skut et al (1997) and the corresponding individual `tiger:prepositionalPhrase` in the Annotation Model http://purl.org/olia/tiger-syntax.owl:[9]

```
      tiger:prepositionalPhrase system:hasTag 'PP'
  tiger:prepositionalPhrase a tiger:PrepositionalPhrase
```

Linking Models then import an Annotation Model and the Reference Model and specify relations between their concepts: `tiger:PrepositionalPhrase` $\sqsubseteq$ `olia:PrepositionalPhrase`. The Linking with External Reference Models like ISOcat is analogous: `olia:PrepositionalPhrase` $\sqsubseteq$ `isocat:DC-2257`. In consequence, it is true that `tiger:prepositionalPhrase rdf:type isocat:DC-2257`

[8] `olia:NounHeadedPhrase` was introduced as a generalization over prepositional phrase and noun phrase to account for the constituent representation of certain dependency parsers where this differentiation is not made. However, there is little agreement whether the noun or the preposition is the head of a prepositional phrase, but this debate is independent of the modeling of the OLiA Reference Model: The *linking* with external reference models has to specify that the `olia:NounHeadedPhrase` corresponds to a single, well-defined category in an independently developed community-maintained terminology repository, or that it does not (then, a disjunction would be necessary).

[9] Because the annotation scheme of the NEGRA corpus is closely related with the annotation scheme of the TIGER corpus (Brants et al, 2004), both are represented together in the Annotation Model developed for TIGER-style syntax.

Within an application, the German phrase *für viele junge Komponisten* considered above can then be circumscribed by means of the concepts it is associated with:[10]

```
olia:PrepositionalPhrase and
powla:hasChild some (olia:CommonNoun and
                     olia:hasGender some olia:Masculine and
                     olia:hasNumber some olia:Plural) and
...
```

This description is concept-based and thus independent from any particular tagset, and applied to another Annotation Model, a query for `olia:PrepositionalPhrase` would retrieve another set of individuals that represent the same meaning with different annotations, e.g., the phrase *von den geschäftlichen Revisoren* 'by the business auditors' from the TüBa-D/Z corpus (document 1, sentence 19)[11] – even though it carries the label `PX` and not `PP` like in the NEGRA/TIGER annotation scheme.[12]

3.3 Current Status of the OLiA Ontologies

The OLiA ontologies are available from http://purl.org/olia. At the moment, they have not been officially released, they will be released under a Creative Commons Attribution license in mid-2012.

The OLiA ontologies cover different grammatical phenomena, including inflectional morphology, word classes, phrase and edge labels of different syntax annotations, as well as discourse annotations (coreference, discourse relations, discourse structure and information structure). Annotations for lexical semantics are only covered by the OLiA ontologies to the extent that they are encoded in syntactic and morphosyntactic annotation schemes (e.g. as grammatical roles). For lexical semantic annotations in general, a number of reference resources are already available as Linked Data, including RDF versions of WordNet,[13] FrameNet,[14] and the Wikipedia.[15]

The OLiA Reference Model comprises 14 `MorphologicalCategory`s (morphemes), 263 `MorphosyntacticCategory`s (word classes/part-of-speech tags), 83 `SyntacticCategory`s (phrase labels), and 326 different values for 16 `MorphosyntacticFeature`s, 4 `MorphologicalFeature`s, 4 `SyntacticFeature`s and 4 `SemanticFeature`s.

[10] The linking between corpus data and the OLiA ontologies can be accomplished, for example, by copying all properties of an OLiA Annotation Model individual to the POWLA individuals with the corresponding annotations, as specified by the `hasTag` property.

[11] http://www.sfs.uni-tuebingen.de/tuebadz.shtml

[12] See http://purl.org/olia/tueba.owl#PX and http://purl.org/olia/tueba-link.rdf.

[13] http://thedatahub.org/dataset/w3c-wordnet, also see McCrae et al (this vol.).

[14] http://www.loa.istc.cnr.it/codeps/owl/ofntb.owl, cf. Nuzzolese et al (2011).

[15] http://dbpedia.org, cf. Hellmann et al (this vol.).

As for morphological, morphosyntactic and syntactic annotations, the OLiA ontologies include 32 Annotation Models for about 70 different languages, including several multi-lingual annotation schemes, e.g. the EAGLES recommendations (Chiarcos, 2008) for 11 Western European languages, and Multext-East (Chiarcos and Erjavec, 2011) for 15 (mostly) Eastern European languages. As for non-(Indo-)European languages, the OLiA ontologies include morphosyntactic annotation schemes for languages of the Indian subcontinent, for Arabic, Basque, Chinese, Estonian, Finnish, Hausa, Hungarian and Turkish. Other languages, including languages of Africa, the Americas, the Pacific and Australia are covered by Annotation Models developed for glosses as produced in typology and language documentation. The OLiA ontologies also cover historical language stages, including Old High German, Old Norse and Old/Classical Tibetan.

As mentioned above, application of modular OWL/DL ontologies allows to link annotations with terminological repositories: Annotation schemes and reference terminology are formalized as OWL/DL ontologies, and the linking is specified by `rdfs:subClassOf` descriptions. This mechanism has also been applied to link the OLiA Reference Model with existing terminology repositories, including GOLD (Chiarcos, 2008), the OntoTag ontologies (Buyko et al, 2008, cf. Pareja-Lora, this vol.) and ISOcat (Chiarcos, 2010a, cf. Windhouwer and Wright, this vol.). Thereby, the OLiA Reference Model provides a stable intermediate representation between existing terminology repositories and ontological models of annotation schemes. This allows any concept that can be expressed in terms of the OLiA Reference Model also to be interpreted in the context of ISOcat, GOLD or the OntoTag ontologies. Using the OLiA Reference Model, it is thus possible to develop applications that are interoperable in terms of GOLD *and* ISOcat even though both are still under development and both differ in their conceptualizations (see footnote 3).

Example applications of the OLiA ontologies include the specification of grammatical features in lexical resources (McCrae et al, 2011, this vol.), the development of tagset independent NLP architectures (Chiarcos, 2010b; Hellmann et al, this vol.), and tagset independent corpus queries, e.g. the combination of OLiA reference concepts with SPARQL queries on POWLA data, as shown in the following section.

4 A Use-Case for Corpus Interoperability: Developing Resource-Independent Corpus Queries

A number of different applications of OWL/RDF-encoded corpora are possible. From the perspective of corpus linguistics and NLP, it is essential that the data structure provides exhaustive access to the information conveyed in the corpus. This can be shown, for example, by implementing a **corpus querying engine**. A pilot experiment has been performed where a multi-layer corpus, the syntax-annotated German NEGRA corpus (Skut et al, 1998) with the coreference annotations by Schiehlen (2004) was converted to POWLA. The RDF data was loaded in an RDF database, OpenLink Virtuoso, and could thus be queried with SPARQL. To illustrate that SPARQL queries can retrieve all relevant information from POWLA data, the operators of the ANNIS query language (AQL) were reimplemented as SPARQL macros, the query language used in the corpus querying system ANNIS (Chiarcos et al, 2008). With ANNIS, a query engine for PAULA data is available, and was successfully applied in a number of linguistic and philological projects, so, it can be assumed that AQL represents the minimal power necessary to explore any kind of linguistic corpora. The portability of PAULA to the Semantic Web could thus be shown with the reconstruction of all AQL operators as SPARQL macros.[16] Although detailed results of this evaluation are described elsewhere, we can state here that multi-layer corpora in POWLA can be queried with SPARQL macros in basically the same way as with AQL.

SPARQL actually provides us with an even more powerful means of querying: For example, AQL does not support queries for the absence of a particular annotation (e.g. an NP not dominating a pronoun), which can be easily expressed in SPARQL:

```
PREFIX negra: <http://purl.org/powla/negra-sample.owl#>.
PREFIX powla: <http://purl.org/powla/powla.owl#>.
SELECT DISTINCT ?np
WHERE {
    ?np a powla:Nonterminal.
    ?np negra:has_cat "NP".
    OPTIONAL { ?np powla:hasChild ?pronoun.
               ?pronoun negra:has_pos "PPER" }
    FILTER !bound(?pronoun)
}
```

POWLA can thus be used to create linguistic information systems. An additional advantage of the OWL/RDF formalization is that it represents a standardized to represent heterogeneous data collections: With RDF, a standardized representation formalism for different corpora is available, and with data types being defined in OWL/DL, the validity of corpora can be automatically checked (according to the consistency constraints posited by the POWLA ontology). POWLA represents a possible solution to the **structural interoperability** challenge for linguistic corpora (Ide and Pustejovsky, 2010). Unlike other formalisms developed in this direction (e.g., PAULA, or GrAF Ide and Suderman, 2007), it does not involve a special-purpose XML standoff format, but it builds on established standards with broad technical support and an active and comparably large community. Standard formats specifically designed for linguistic annotations as developed in the context of ISO TC37/SC4 (e.g. GrAF Ide and Suderman, 2007 and TIGER2 Romary et al, 2011), are, however, still under development.

In comparison of this approach with current initiatives within the linguistics/NLP community, e.g. ISO TC37/SC4, which focus on complex standoff XML formats specifically designed for linguistic data, this approach offers three crucial advantages:

[16] Details for this conversion, a sample data and SPARQL macros for querying this data can be found under http://purl.org/powla.

1. The increasing number of RDF data bases provides us with convenient means for the management of linguistic data collections.
2. Augmenting an RDF representation of linguistic corpora with formally specified consistency conditions, i.e., an OWL/DL specification of data types and constraints, existing reasoners can be applied to check the consistency of this representation.
3. Resources can be freely interconnected with each other and with lexical-semantic resources available from the Linked Open Data cloud.

Additionally, an RDF representation of linguistic corpora allows for their **integration with other RDF resources**. The linking of POWLA corpora with OLiA ontologies as described before, for example, allows reformulating the SPARQL query for NPs not dominating a pronoun, such that the reformulated query can also be applied to corpora with other annotation schemes and is thus, interoperable:

```
WHERE {
    ?np a olia:NounPhrase.
    OPTIONAL { ?np powla:hasChild ?pronoun.
               ?pronoun a olia:PersonalPronoun }
    FILTER !bound(?pronoun)
}
```

Representing metadata and annotations of a corpus by means of references to resources in the LOD cloud, it becomes possible to define metadata filters for linguistic corpora and queries for linguistic annotations that are independent of the string representation of this information in the corpora. Resources represented in this way are thus not only structurally, but also **conceptually interoperable**.

If published as Linked Data, corpora represented in RDF can further be **linked with lexical-semantic resources** already available as Linked Data, e.g., a general knowledge base like DBpedia (Hellmann et al, this vol.), or linguistic resources like WordNet or Wiktionary (McCrae et al, this vol.). Semantic annotations, e.g., those of PropBank (Kingsbury and Palmer, 2002), can therefore be implemented as a linking between corpora and semantic resources, without the need to duplicate or to synchronize resources obtained from different providers.

Another important type of resources available as Linked Data includes **repositories of metadata and terminology**. Lexvo,[17] for example, provides identifiers for languages based on ISO 639; Glottolog (Nordhoff, this vol.) provides an even more fine-grained taxonomy of languoids.

A unique feature of the approach described here is that OWL and RDF provide a solution for both aspects of corpus interoperability – and even interoperability between corpora and lexical-semantic resources – **within the same formalism**. One concrete advantage of such a holistic approach is that novel corpus information systems can be developed that achieve interoperability by simpler means, e.g., using a single data base for both corpus data and annotation mapping, whereas traditional systems that combine XML- or SQL-based corpus representations with linguistic

[17] http://lexvo.org

ontologies (e.g., Rehm et al, 2007; Chiarcos et al, 2008) require a more complex system architecture and thus, greater implementation efforts.[18]

Acknowledgements The research described here was partially supported by the German Research Foundation (DFG) through the Collaborative Research Center (SFB) 441 "Linguistic Data Structures" (University of Tübingen, 2006-2008, OLiA ontologies) and through the Collaborative Research Center (SFB) 632 "Information Structure" (University of Potsdam, 2008-2011, OLiA ontologies, POWLA)

References

Bickel B, Nichols J (2002) Autotypologizing databases and their use in fieldwork. In: Proceedings of the LREC-2002 Workshop on Resources and Tools in Field Linguistics, Las Palmas, Spain

Bird S, Liberman M (2001) A formal framework for linguistic annotation. Speech Communication 33(1-2):23–60

Boersma P (2002) Praat, a system for doing phonetics by computer. Glot international 5(9/10):341–345

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

Brants S, Dipper S, Eisenberg P, Hansen S, önig EK, Lezius W, Rohrer C, Smith G, Uszkoreit H (2004) TIGER: Linguistic interpretation of a German corpus. Research on Language and Computation 2(4):597–620

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

Busemann A, Busemann K (2008) Toolbox self-training. Tech. rep., http://www.sil.org, version 1.5.4, Oct 2008

Buyko E, Chiarcos C, Pareja-Lora A (2008) Ontology-based interface specifications for a NLP pipeline architecture. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco

Carletta J, Evert S, Heid U, Kilgour J (2005) The NITE XML Toolkit: data model and query. Language Resources and Evaluation Journal (LREJ) 39(4):313–334

[18] Another crucial advantage of RDF data bases as compared to relational data bases is that RDF data can be flexibly added or removed. For debugging purposes, the OWL/DL-defined data model (and the RDF triples inferred) can thus be adjusted without without reinitializing the data base, thereby substantially accelerating development cycles.

As compared to XML data bases, RDF data dases support multi-layer corpora with an unrestricted band-width of annotations, whereas XML data bases are optimized for hierarchical annotations (e.g., syntax trees without crossing edges and overlapping segments), but relatively inefficient for non-hierarchical data (e.g., coreference, overlapping hierarchies in multi-layer corpora).

Aguado de Cea G, Gomez-Perez A, Alvarez de Mon I, Pareja-Lora A (2004) OntoTag's linguistic ontologies. In: Proc. Information Technology: Coding and Computing (ITCC'04), Washington, DC, USA

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

Chiarcos C (2010a) 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 (2010b) Towards robust multi-tool tagging. An OWL/DL-based approach. In: ACL 2010, Uppsala, Sweden, pp 659–670

Chiarcos C, Erjavec T (2011) OWL/DL formalization of the MULTEXT-East morphosyntactic specifications. In: 5th Linguistic Annotation Workshop, Portland, pp 11–20

Chiarcos C, Dipper S, Götze M, Leser U, Lüdeling A, Ritz J, Stede M (2008) A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. TAL (Traitement automatique des langues) 49(2)

Chiarcos C, Ritz J, Stede M (2011) By all these lovely tokens ... Merging conflicting tokenizations. Journal of Language Resources and Evaluation (LREJ) 4(45), to appear

Dipper S (2005) XML-based stand-off representation and exploitation of multi-level linguistic annotation. In: Proc. Berliner XML Tage 2005 (BXML 2005), Berlin, Germany, pp 39–50

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

Erjavec T (2004) MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In: Fourth International Conference on Language Resources and Evaluation, (LREC 2004), Lisboa, Portugal, pp 1535–1538

Farrar S, Langendoen D (2003) A linguistic ontology for the semantic web. Glot International 7(3):97–100

Hellmann S, Unbehauen J, Chiarcos C, Ngonga Ngomo A (2010) The TIGER Corpus Navigator. In: 9th International Workshop on Treebanks and Linguistic Theories (TLT-9), Tartu, Estonia, pp 91–102

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

Hellwig B, Uytvanck DV, Hulsbosch M (2008) ELAN - Linguistic Annotator. Tech. rep., http://www.lat-mpi.eu/tools/elan, version of 2008-07-31

Ide N, Pustejovsky J (2010) What does interoperability mean, anyway? Toward an operational definition of interoperability. In: Proc. Second International Conference on Global Interoperability for Language Resources (ICGL 2010), Hong Kong, China

Ide N, Romary L (2004) International standard for a linguistic annotation framework. Natural language engineering 10(3-4):211–225

Ide N, Suderman K (2007) GrAF: A graph-based format for linguistic annotations. In: Proc. Linguistic Annotation Workshop (LAW 2007), Prague, Czech Republic, pp 1–8

Kemps-Snijders M (2010) Relish: Rendering endangered languages lexicons interoperable through standards harmonisation. In: 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages, held in conjunction with LREC 2010, Valetta, Malta

Kemps-Snijders M, Windhouwer M, Wittenburg P, Wright S (2009) ISOcat: Remodelling metadata for language resources. International Journal of Metadata, Semantics and Ontologies 4(4):261–276

Kingsbury P, Palmer M (2002) From treebank to propbank. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Citeseer, pp 1989–1993

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

Leech G, Wilson A (1996) EAGLES recommendations for the morphosyntactic annotation of corpora. http://www.ilc.cnr.it/EAGLES/annotate/annotate.html, version of March 1996

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

Marcus M, Santorini B, Marcinkiewicz MA (1993) Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics 19(2):313–330

Marcus M, Santorini B, Marcinkiewicz M (1994) Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19(2):313–330

McCrae J, Spohr D, Cimiano P (2011) Linking lexical resources and ontologies on the semantic web with Lemon. The Semantic Web: Research and Applications pp 245–259

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

Müller C, Strube M (2006) Multi-level annotation of linguistic data with MMAX2. In: Corpus Technology and Language Pedagogy, Peter Lang, Frankfurt am Main, pp 197–214

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

Nuzzolese A, Gangemi A, Presutti V (2011) Gathering lexical linked data and knowledge patterns from framenet. In: Proceedings of the sixth international conference on Knowledge capture, ACM, pp 41–48

O'Donnell M (2000) Rsttool 2.4 – a markup tool for Rhetorical Structure Theory. In: Proc. International Natural Language Generation Conference (INLG'2000), Mitzpe Ramon, Israel, pp 253–256

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

Pareja-Lora A, Aguado de Cea G (2010) Ontology-based interoperation of linguistic tools for an improved lemma annotation in Spanish. In: Proceedings of LREC 2010, Valetta, Malta

Pustejovsky J, Meyers A, Palmer M, Poesio M (2005) Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference. In: Proc. ACL Workshop on Frontiers in Corpus Annotation 2005, Ann Arbor, MI, USA

Rehm G, Eckart R, Chiarcos C (2007) An OWL-and XQuery-based mechanism for the retrieval of linguistic patterns from XML-corpora. In: Proc. RANLP 2007, Borovets, Bulgaria

Romary L, Zeldes A, Zipser F (2011) [tiger2/]-serialising the iso synaf syntactic object model. Arxiv preprint arXiv:11080631

Saulwick A, Windhouwer M, Dimitriadis A, Goedemans R (2005) Distributed tasking in ontology mediated integration of typological databases for linguistic research. In: Proc. 17th Conf. on Advanced Information Systems Engineering (CAiSE'05), Porto

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

Schmidt T (2004) EXMARaLDA – ein System zur computergest ützten Diskurstranskription. In: Mehler A, Lobin H (eds) Automatische Textanalyse. Systeme und Methoden zur Annotation und Analyse nat ürlichsprachlicher Texte, Verlag f ür Sozialwissenschaften, Wiesbaden, Germany, pp 203–218

Schmidt T, Chiarcos C, Lehmberg T, Rehm G, Witt A, Hinrichs E (2006) Avoiding data graveyards. In: Proceedings of the E-MELD workshop on Digital Language Documentation, East Lansing

Skut W, Krenn B, Brants T, Uszkoreit H (1997) An annotation scheme for free word order languages. In: Proc. 5th Conference on Applied Natural Language Processing (ANLP), Washington, D.C.

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

Stede M, Bieler H, Dipper S, Suriyawongkul A (2006) Summar: Combining linguistics and statistics for text summarization. In: Proc. 17th European Conference on Artificial Intelligence (ECAI-06), Riva del Garda, Italy, pp 827–828

Trißl S, Leser U (2007) Fast and practical indexing and querying of very large graphs. In: Proc. 2007 ACM SIGMOD international conference on Management of data, ACM, pp 845–856

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

Zeldes A, Ritz J, Lüdeling A, Chiarcos C (2009) ANNIS: A search tool for multilayer annotated corpora. In: Proc. Corpus Linguistics, Liverpool, UK, pp 20–23

Zipser F, Romary L (2010) A model oriented approach to the mapping of annotation formats using standards. In: Proc. LREC-2010 Workshop on Language Resource and Language Technology Standards (LR&LTS 2010), Valetta, Malta