A Generic Formalism to Represent Linguistic Corpora in RDF and OWL/DL

Christian Chiarcos

Information Science Institute, University of Southern California
chiarcos@daad-alumni.de

Abstract

This paper describes POWLA, a generic formalism to represent linguistic corpora by means of RDF and OWL/DL. Unlike earlier approaches in this direction, POWLA is not tied to a specific selection of annotation layers, but rather, it is designed to support any kind of text-oriented annotation. POWLA inherits its generic character from the underlying data model PAULA (Dipper, 2005; Chiarcos et al., 2009) that is based on early sketches of the ISO TC37/SC4 Linguistic Annotation Framework (Ide and Romary, 2004). As opposed to existing standoff XML linearizations for such generic data models, it uses RDF as representation formalism and OWL/DL for validation. The paper discusses advantages of this approach, in particular with respect to interoperability and queriability, which are illustrated for the MASC corpus, an open multi-layer corpus of American English (Ide et al., 2008).

Keywords: corpus representation, multi-layer corpora, interoperability, RDF/OWL

1. Motivation and Background

The increasing complexity and diversity of linguistic resources that have become available throughout the last decades lead to growing interest of linguistics and NLP communities in the sustainability and interoperability of NLP tools, annotated corpora and lexical-semantic resources.

This paper describes an approach towards the interoperability of annotated corpora on the basis of formalisms developed by the Semantic Web community: An established generic data model for annotated corpora multi-layer corpora, PAULA (Dipper, 2005; Chiarcos et al., 2009), is reconstructed using RDF and OWL/DL. The primary objective behind this approach is to make use of the convenient means of storing and querying that are available for RDF data, and the formalism developed for this approach, POWLA, is added to the set of alternative linearizations of the PAULA data model that have been developed for different application scenarios, including PAULA XML (Dipper, 2005, XML standoff format for lossless representation and exchange of PAULA data), PAULA inline (Dipper et al., 2007, XML inline representation for XML data bases), relANNIS (Zeldes et al., 2009, PAULA linearization for relational data bases), and Salt (Zipser and Romary, 2010, JAVA implementation of the PAULA data model as part of a converter framework).

POWLA allows to process, to store and to query PAULA data with standard tools for RDF, in particular, RDF data bases. The development of a novel and easy-to-query data base for PAULA data arises from the need for a data base suitable for NLP applications and advanced statistical analyses of PAULA data:[1] POWLA is a straight-forward implementation of the original data model (Chiarcos, 2012), i.e., very transparent to NLP engineers familiar with PAULA XML. As compared to the existing data base solutions, RDF data bases support a standard query language for labeled directed graphs, SPARQL (Prud'Hommeaux and Seaborne, 2008), and they allow for interactive manipulation of data (Gearon et al., 2012).[2] This paper emphasizes another advantage of RDF-based corpus modeling, namely the improved structural interoperability of PAULA data with other linguistic resources, such as terminology repositories and lexical-semantic resources.

Interoperability of linguistic resources involves primarily two aspects (Ide and Pustejovsky, 2010): Conceptual ('semantic') interoperability (represent linguistic annotations using well-defined vocabularies grounded in community-maintained terminology repositories) and structural ('syntactic') interoperability (represent and access linguistic resources using standard formats and protocols). If both conditions are met, it is possible to integrate information from different linguistic resources seamlessly, e.g., annotation layers within a corpus, annotations of different corpora, possibly produced by different tools, or corpora and other linguistic resources.

A minimal requirement for **structural interoperability** is that different linguistic resources do not require complicated conversion routines in order to integrate their information. For this purpose, the NLP and linguistics communities have developed a number of representation formalisms that address this problem either in a fully generic approach (Bird and Liberman, 2001; Carletta et al., 2003; Ide and Suderman, 2007; Chiarcos et al., 2008), or for the full band-width of annotations for one particular phenomenon (e.g., syntax annotation, Declerck, 2006; Romary

[1] The data base ANNIS (Zeldes et al., 2009) and its data model relANNIS do currently not allow for interactive manipulation of PAULA data, but require data export, script-based data manipulation and reimport (Chiarcos and Ritz, 2010). Also, relANNIS is highly optimized for the query language AQL, which is, however, directed to end user of the ANNIS corpus information system and thus crippled with respect to its expressivity (e.g., no all-quantification) in order to guarantee system performance. For example, dominance relations are not encoded directly, but by means of pre- and post-order indices (Trißl and Leser, 2007). For developers familiar with PAULA XML, thus, relANNIS data structures are relatively intransparent and are currently not directly usable for the development of NLP applications.

[2] This can be employed to implement application-specific optimizations of RDF data *when needed*, e.g., to precompile complex queries. A-priori optimization for selected types of queries as in relANNIS is thus not necessary.

et al., 2011). Under the umbrella of the Linguistic Annotation Framework of the ISO TC37/SC4, these approaches gradually converge towards the establishment of standard data models and formalisms (Ide and Romary, 2004; Ide and Romary, 2006), but this is still an on-going process.

State-of-the-art approaches on structural interoperability of annotated corpora are built on the assumption that *all kinds of linguistic annotations* that can be attached to textual data can be represented by means of **labeled directed graphs** (Bird and Liberman, 2001; Ide and Suderman, 2007). A labeled directed graph $G = \langle N, E, l_N, l_E \rangle$ is a 4-tuple consisting of a set N of nodes, a set $E \subseteq N \times N$ of edges, and relations $l_N : N \mapsto \Sigma^*$ and $l_E : E \mapsto \Sigma^*$ that map nodes and edges to their respective labels (represented as strings here). In the following, we deviate from the classical definition by assuming that the relations l_N and l_E are not right-unique functions, but general relations that can assign the same node (or edge) multiple labels.

On this basis, generic data structures for linguistic annotations can be defined:[3]

segment a node in a graph, $n \in N$

relation an edge in the graph, $e \in E$

annotation (attribute-value pairs) represented by their string representation as labels using the relations l_N and l_E.

layer like a segment, a layer can be represented within a document as a node $n \in N$, but with a special label, and connected to the elements (segments, or relations) it contains by means of a special relation, in order to mark its special status as compared to a node.

privileged segmentation layer a layer with specific well-formedness conditions (minimal addressable units, totally ordered, covering the entire text covered by annotations), and with a special label

In this way, linguistic annotations can be represented within a graph, and this graph can then be anchored to the primary data. It should be noted here that graphs do not provide a sufficiently restrictive data model to represent linguistic annotations, but that additional constraints apply, manifested in these definitions by *naming conventions* for specific labels (e.g., for the privileged segmentation layer). To check the consistency of annotations represented as directed graphs, thus additional means of validation are required.

Technically, state-of-the-art approaches represent these data structures using **standoff XML**, i.e., a bundle of XML files that represent a particular document and the annotations applied to it as separate files heavily interlinked by means of XLink/XPointer. Figure 1 shows an example clause drawn from the Manually Annotated Sub-Corpus (MASC) of the American National Corpus (Ide et al., 2010),[4] with annotations for syntax (Bies et al., 1995) and frame semantics (Baker et al., 1998) as represented in GrAF. Standoff formats are based on the physical separation between primary data and different annotation layers, usually in different files, that are interconnected with XLink/XPointer. In Fig. 1, this is indicated by the names of the XML files that contain the different layers in GrAF. Naturally, the efforts to parse, to validate and to process standoff annotations are relatively high, and there are no efficient means for storing and querying general standoff XML data available (Eckart, 2008), so that it is necessary to convert standoff XML to other representations in order to process it efficiently. Thus, any approach using standoff XML in a comparably massive way as GrAF requires the parallel development of multiple linearizations of the data model, synchronization of this development, and permanent conversion between both formats.

But standoff XML is not the only option to encode graph-based data structures. The Semantic Web community, for example, developed the **Resource Description Framework (RDF)**, a W3C standard that implements a data model based on labeled directed (multi-)graphs, and that can be linearized in different ways (including XML representations). Unlike standoff-XML formats developed specifically for the linguistics/NLP communities, it comes with a rich infrastructure of APIs, tools, data bases, and query languages, and with an interdisciplinary, comparably large and active user community.

Specialized sub-languages have been developed to define more specialized data structures by creating a *reserved vocabulary* and *structural constraints*. The **Web Ontology Language (OWL)** defines the datatypes necessary for the representation of ontologies as an extension of RDF, i.e., *classes* (concepts), *instances* (individuals) and *properties* (relations). OWL/DL is a dialect of OWL that is restricted such that the language corresponds to a description logic (decidable fragment of first-order predicate logic). Exploiting this restriction, a number of reasoners have been developed that allow to verify consistency constraints (*axioms*) and to draw inferences from the ontology.

This paper combines these developments with a state-of-the-art approach on corpus interoperability, PAULA (Dipper, 2005; Chiarcos et al., 2009), that provides a data model whose RDF/OWL description is described here: OWL/DL is used to define data types for annotated corpora and constraints over these. On this basis, existing reasoners can be applied to check the consistency of RDF corpora, e.g., that segments (nodes) and relations (edges) are disjoint, that every edge has one source and one target node, that every node has at most one layer, etc.

2. POWLA

As this paper focuses on the application of POWLA to annotated corpora, it describes its components briefly only,

[3] This paper focuses on data structures *for annotations*. Data structures necessary for corpus organization are discussed with greater level of detail elsewhere (Chiarcos, 2012).

[4] This clause is part of the sentence *While Byzantine land was being divided, there was no one in control of the seas, so pirates raided towns on many of the islands*, taken from the file `HistoryGreek`, written section of MASC v.1.0.3, `http://www.anc.org/MASC`, for the semantic annotations of this sentence see Baker and Fellbaum (2009).

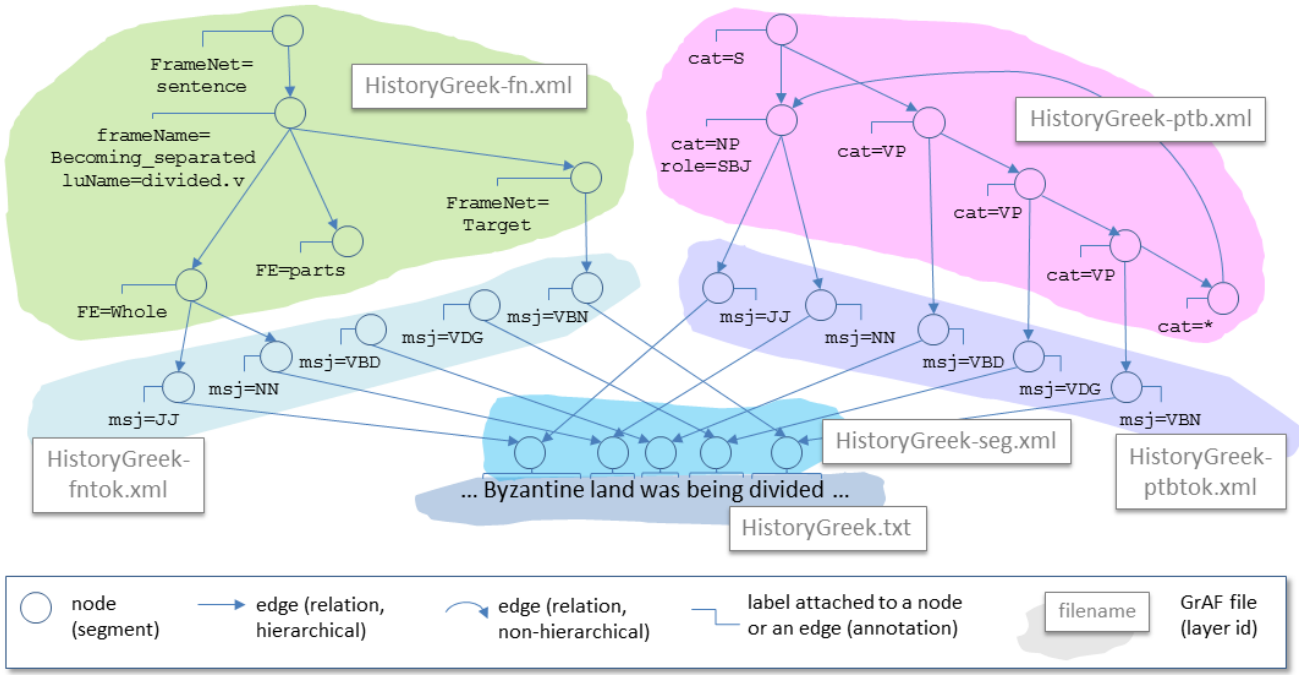

Figure 1: Representing and integrating annotations for syntax and frame-semantics in GrAF

and with a focus on the RDF formalization of *annotations*. The derivation of POWLA from the PAULA data model is described with greater level of detail by Chiarcos (2012). The POWLA ontology, converters, a SPARQL query preprocessor, data from two converted corpora and further documentation available from `http://purl.org/powla`.

2.1. POWLA TBox: Data Structures

The **POWLA TBox** represents a straight-forward implementation of the data types of PAULA in an OWL/DL ontology. All POWLA concepts are subconcepts of `POWLAElement`: `Node` and `Relation` that are used to represent linguistic annotations, `Document` and `Layer` are concerned with corpus organization (not discussed here).

A `Node` is a `POWLAElement` that covers a (possibly empty) 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. PAULA distinguishes between dominance (hierarchy-building) and pointing (general) relations. In POWLA, this difference is represented by the `hasChild` property that connects source and target node of dominance relations. It is thus not necessary to distinguish pointing relations and dominance relations as separate concepts in the POWLA ontology.

`Node`s and `Relation`s can be assigned one or multiple labels that correspond to the string value of the linguistic annotation. The corresponding property `hasAnnotation`is, however, not to be used directly, but rather, subproperties are to be created that express the attribute name, e.g., `has_pos` for part-of-speech annotation, or `has_cat` for phrase labels in the syntax annotation.

Two basic subclasses of `Node` are distinguished: A `Terminal` is a `Node` that does not have a `hasChild` property. `Terminal`s constitute the privileged segmentation layer mentioned above, i.e., they represent the minimal unit of annotation, they are totally ordered and they cover the entire stretch of annotated text. A `Nonterminal` is a `Node` that has at least one `hasChild` property.

Both `Terminal`s and `Nonterminal`s are characterized by a string value (property `hasString`), and a particular position (properties `hasStart` and `hasEnd`) with respect to the primary data. `Terminal`s are further connected with each other by means of `next` properties. This is, however, a preliminary solution. In forthcoming versions of POWLA, `Terminal`s may be linked to strings in accordance to the currently developed NLP Interchange Format (NIF).[5]

The POWLA TBox specifies a number of constraints, for example, that `Nonterminal` and `Terminal` are disjoint, hence OWL/DL is necessary for this ontology. Using OWL/DL has a number of advantages, for example, we can *infer* whether a `Node` is a `Nonterminal` or a `Terminal`. This can also be exploited to distinguish different classes of `Nonterminal`s: PAULA requires a formal distinction between markables (flat, layer-based annotations) and structs (hierarchical annotations, e.g., trees). In POWLA, this information can be expressed as a property of an annotation layer, i.e., `Layer` (informally, a set of `Node`s and `Relation`s): If all nodes from a `Layer` dominate only `Terminal`s and none of them uses a labeled `Relation` to one of its children, this `Layer` is a `MarkableLayer`, otherwise, it is a `StructLayer`. The differentiation is, however, a technical issue only relevant for visualization,[6] but not for

[5] `http://nlp2rdf.org/nif-1-0#toc-nif-recipe-offset-based-uris`

[6] `StructLayer`s can be visualized as multi-rooted trees, `MarkableLayer`s can be visualized as rows in a table, cf.

querying or other purposes, and it may not be necessary to provide this information *unless needed*. The POWLA TBox allows us to infer this differentiation automatically from the data, so it does not have to be specified explicitly.

2.2. POWLA ABox: Corpus Data

A corpus can be represented as a set of individuals that instantiate the concepts defined in the POWLA TBox.
The POWLA ontology defines data types that can now be used to represent linguistic annotations. Considering the syntactic annotation of the phrase *Byzantine land* from Fig. 1, the following `Node`s are created:[7]

```
<Nonterminal rdf:about="#ptb-n03094">
  <has_cat>NP</has_cat>
  <has_role>SBJ</has_role>
  <hasChild rdf:about="#ptb-n03095">
  <hasChild rdf:about="#ptb-n03096">
  ...
</Nonterminal>

<Nonterminal rdf:about="#ptb-n03095">
  <has_msj>JJ</has_msj>
  <hasChild rdf:about="#seg-r3149">
  ...
</Nonterminal>

<Terminal rdf:about="#seg-r3149">
  <hasString>Byzantine</hasString>
  <next rdf:about="#seg-r3151">
  ...
</Terminal>
```

The Nonterminal `ptb-n03094` represents the NP *Byzantine land*, its ID is inherited from the original GrAF XML, the properties `powla:has_cat` and `powla:has_role` were created as subproperties of `powla:hasAnnotation` to reflect the original GrAF attributes `cat` and `role`. The property `powla:has_msj` of its child node `ptb-n03095` reflects the original GrAF attribute `msj` that contained part-of-speech annotation. As in GrAF, this node is distinguished from the `Terminal seg-r3149` that is part of the privileged segmentation layer (`HistoryGreek-seg.xml`) in GrAF.

3. Benefits of POWLA

As for benefits of RDF representations of linguistic corpora, this paper focuses on interoperability issues, in particular, structural interoperability among annotation layers in multi-layer corpora, interoperability between lexical-semantic resources and conceptual interoperability between annotations of related types but different annotation schemes.

3.1. Querying Multi-Layer Corpora with POWLA

In accordance with the definition given above, structural interoperability can be said to be successfully achieved, if different corpora are represented within the same data format, they can be stored within the same data base, and be successfully queried with a query language based on the underlying generic data model. As querying different corpora for the same type of annotations is trivial, structural interoperability can thus be shown by formulating queries across different layers of annotation.
Two experiments were conducted:

- Two multi-layer corpora were converted to POWLA, the German newspaper corpus NEGRA corpus (Skut et al., 1998) with the coreference annotations by (Schiehlen, 2004), and the MASC corpus, v. 1.0.3. The corpora were loaded into the RDF data base OpenLink Virtuoso (OpenLink Software, 2009), and could be queried with SPARQL.
- A set of SPARQL macros was implemented that emulate the PAULA-based ANNIS Query Language AQL (Chiarcos et al., 2008). We showed that every operator in AQL can be rendered in terms of SPARQL.

Details on these experiments can be found unter `http://purl.org/powla`. The first experiment showed that it is *possible* to generate, to store and to query POWLA data, the second experiment showed that this conversion was *useful*, i.e., that no linguistically relevant information was lost (assuming that AQL corresponds to what linguists search in a corpus). The most important result, however, is, how little resources were necessary for this task: In total, both took us about 3 man-weeks.
With RDF specifications for syntax and FrameNet annotations in MASC, it is possible to combines information from both layers of annotation. The following query can be used to find out which grammatical role the `Whole` argument of the frame `Becoming_separated` is assigned in a corpus:

```
SELECT ?gr
WHERE {
  ?frame a powla:Node.
  ?frame has_frameName "Becoming_separated".
  ?frame powla:hasChild ?wholeArg.
  ?wholeArt has_FE "Whole".
  ?wholeArg _=_ ?phrase.
  ?phrase has_cat "NP".
  ?phrase has_role ?gr.
}
```

Applied to the example, this query retrieves `ptb-n03094` as value for the variable `?phrase` and thus `SBJ` as value for the variable `?gr`. It should be noted that the operator `_=_` is *not* a SPARQL expression, but rather one of the AQL macros mentioned above. It serves to retrieve elements that are co-extensional, i.e., that cover the same `Terminal`s. To implement this operator, two additional properties were added to the POWLA TBox, `firstTerminal` and `lastTerminal`. For the example, `seg-r3149` is the `firstTerminal` of `ptb-n03094`.
Using a JAVA-based query preprocessor, the AQL macro α `_=_` β is expanded into SPARQL as follows:

α `firstTerminal` γ. α `lastTerminal` δ.
β `firstTerminal` γ. β `lastTerminal` δ.

Chiarcos et al. (2008).

[7] For readability, the `powla:` namespace is omitted.

In this way, all AQL query operators can reconstructed as SPARQL macros:

Dominance relations can be queried with $>$ (parent-child relationship), with optional constraints for the features attached to it, e.g., α `>[role="SBJ"]` β for β being the grammatical subject of α.

Pointing relations can be queried with the operator $->$ with a type attribute, e.g., α `->anaphor_antecedent` β.

Extensionality operators retrieve pairs of nodes that cover the same stretch of primary data, e.g., co-extensionality (α `_=_` β) or inclusion (α `_i_` β, i.e., α covers all tokens covered by β).

It should be noted that, to our best knowledge, ANNIS is the *only* corpus information system that can query over unrestricted combinations of hierarchically and relationally structured annotations. Our pilot study showed how easily RDF data bases can be employed for this task, and thus, how easily corpus query systems on the basis of RDF data bases, POWLA and SPARQL can be built.

While emulating AQL in SPARQL is possible, SPARQL differs from AQL in both its flexibility and expressivity. Because it is part of a corpus information system, AQL always returns values for all variables in the query, whereas SPARQL allows to specify which variables are returned (`SELECT`), to retrieve the broader context of a match (`DESCRIBE`), etc. AQL does not allow queries for the absence of an annotation (this is an implicit all-quantification), but in SPARQL, this can be easily expressed (albeit with potentially huge runtime). I would like to emphasize that is does not reflect insufficiencies of AQL, but merely the purpose that AQL was developed for, i.e., to allow linguists to query corpora and to visualize the results using an expressive, but efficient query language. For advanced statistical analyses or NLP applications, more powerful means to access PAULA data are provided by POWLA and SPARQL.

3.2. Interoperability with Lexical-Semantic Resources

Aside from syntactic and frame-semantic annotations, MASC also contains annotations for WordNet senses, e.g., the word *Byzantine* from the example is assigned the sense key `byzantine%3:01:00::`. With standoff XML, such annotations can be represented and processed, and, given a query language like AQL, they can be retrieved. However, these annotations would be represented as strings (as in the corpus) and it would not be possible to access information from outside the corpus, such as the WordNet specifications for `byzantine%3:01:00::`.

Using RDF representations of WordNet and MASC, however, this can be easily achieved, if sense key annotations are transformed into properties that point to the corresponding URI in an RDF version of WordNet, e.g., `http://wordnet.rkbexplorer.com/id/synset-Byzantine-adjective-2`. It is thus possible to formulate a SPARQL query to retrieve, say, all sentences that contain hypernyms of `byzantine%3:01:00::` in a POWLA corpus.

Actually, such queries are even possible if the lexical-semanantic resource and the corpus are stored in physically separated repositories, as RDF identifiers are URIs (thus globally unique) and SPARQL supports federated search (Hartig et al., 2009).

3.3. Conceptual Interoperability

Similar to the interlinking and the joint querying across corpora and lexical-semantic resources, it is also possible to establish ties between annotations in a corpus and terminology repositories, and thereby address the problem of conceptual interoperability, i.e., that linguistic annotations are anchored in the same reference vocabulary.

For this task, the Ontologies of Linguistic Annotation (Chiarcos, this vol, OLiA) can be employed that formalize annotation schemes as well as reference concepts that are grounded in community-maintained terminology repositories like GOLD (Farrar and Langendoen, 2010) and ISOcat (Chiarcos, 2010).

As an example, consider the annotation `cat="NP"` from the MASC example above. If information about the annotation scheme is provided, the `has_cat` property of the `Node ptb-n03094` can be matched against the corresponding individual `penn-syntax:NP` in the Penn Syntax Annotation Model:[8] `penn-syntax:NP hasTag "NP"`. The property `hasTag` defines the surface form of the tag, and in case of a match, the corresponding POWLA `Node` can be declared an instance of the corresponding superconcept `penn-syntax:NounPhrase`. The OLiA ontologies further specify that `penn-syntax:NounPhrase rdfs:subClassOf olia:NounPhrase`,[9] and they provide further information about `olia:NounPhrase`.[10] Accordingly, the triple `?phrase has_cat "NP"` from the query above can be replaced by `?phrase a olia:NounPhrase`. In this formulation, it is, however, independent from the annotation scheme used, and may also be applied, e.g., to retrieve noun phrases from the German newspaper corpus TüBa-D/Z (Telljohann et al., 2003), even though the corresponding annotation is `NX` rather than `NP`.

4. Discussion and Outlook

This paper described POWLA, a generic formalism to represent linguistic corpora by means of Semantic Web formalism. POWLA implements the generic data model PAULA, a state-of-the-art approach capable to represent text-oriented linguistic annotations in an interoperable way. POWLA employs RDF to represent linguistic annotations, and OWL/DL to define the necessary data structures. The primary objective of the efforts described here was to employ the rich technological infrastructure of tools, in particular, data base implementations to store, to manipulate and to query graph-based data structures that have been developed in the context of the Semantic Web. It should be noted, however, that POWLA is not intended to replace existing solutions like PAULA XML or GrAF with RDF, but that it adds RDF as an application-specific linearization of the PAULA data model, that can be used, for example, to store and to query PAULA data.

[8] `http://purl.org/olia/penn-syntax.owl`

[9] `http://purl.org/olia/penn-syntax-link.rdf`

[10] `http://purl.org/olia/olia.owl`

Related approaches include early applications of RDF/OWL to define data categories of corpora represented in XML (Langendoen et al., 2002; Sasaki et al., 2004; Chiarcos et al., 2008). More recently, the joint modeling of both data categories and corpus data has been applied to enable the querying of multi-layer corpora, albeit mostly with a focus on individual corpora (Burchardt et al., 2008; Hellmann et al., 2010; Mazziotta, 2010), or specific types of corpora, e.g., speech corpora (Grönroos and Miettinen, 2004), or typological data collections (Schalley, 2012). Further, it has been suggested to develop NLP pipelines on the basis of RDF/OWL (Aguado de Cea et al., 2002; Hellmann, 2010; Rubiera et al., 2012), albeit with a focus on the phenomena and tools supported by the pipeline. Unlike these approaches, the approach described here is not tied to one particular resource, or a restricted inventory of annotations, but rather, it is applicable to any kind of text-based linguistic annotation, because it takes its point of departure from an existing XML standoff format that is assumed to be a generic representation formalism for the representation linguistic annotation applicable to textual data.

With the notable exception of Cassidy (2010), I am not aware of any approach that aims for a comparable level of genericity. Cassidy's triplification of GrAF, however, focuses on linguistic annotations, it does not provide an explicit model of corpus organization (`Layer`, `Document`, etc.), which is instead implicitly expressed through the organization of *files* (that is inherited from GrAF). Also, Cassidy's model does not represent information about the primary data in RDF, hence it is not possible to query for surface strings using SPARQL. As in GrAF, references to the primary data are represented as string values in specialized properties (`graf:anchors`), but no information about the interpretation of these anchors is provided. This approach is more light-weight than POWLA, and allows to represent GrAF data in RDF, yet, it does not lead to a *self-contained* representation of GrAF data in RDF, because its interpretation depends on externally provided information.[11]

So far, two corpora have been converted to POWLA, the NEGRA corpus, a German newspaper corpus annotated for syntax and coreference annotations (Chiarcos, 2012), and the MASC corpus, v. 1.0.3, as sketched here. In the longer perspective, any other annotations for which converters to PAULA have been implemented may also be linearized in POWLA. This includes TIGER XML (König and Lezius, 2000, syntax), EXMARaLDA (Schmidt, 2004, layer-based annotations), MMAX2 (Müller and Strube, 2006, relational annotations), Toolbox (Busemann and Busemann, 2008, typological glosses), as well as tab-separated text, inline XML, and annotations produced by special-purpose tools such as the RSTTool (O'Donnell, 2000, discourse structure) and ConAno (Stede and Heintze, 2004, discourse connectives). Using existing converters to one of these source formats, an even broader band-width of tools and formats is supported, e.g., the Penn Treebank bracketing notation (Marcus et al., 1994, via TIGER XML), and ELAN (Hellwig et al., 2008, via EXMARaLDA).

POWLA allows to store PAULA corpora in RDF data bases and query them using SPARQL. For multi-layer corpora, POWLA preserves the structural interoperability established by PAULA. Beyond this, I have sketched how an RDF representation enhances the interoperability between corpora and lexical-semantic resources for the example of WordNet annotations in MASC, and how existing terminology repositories can be employed to establish conceptual interoperability between annotations in different corpora and/or produced by different tools. Technically, both aspects represent standard applications of the Linked Data paradigm (Berners-Lee, 2006).

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

1. The increasing number of RDF data bases provides us with convenient means for the management of linguistic data collections. (Unrestricted standoff XML data cannot be efficiently processed with off-the-shelf XML data bases, Eckart, 2008)

2. By augmenting an RDF representation of linguistic corpora with an OWL/DL specification of data types and constraints for these, existing reasoners can be applied to check the consistency of this representation. (Standoff XML formats need to provide their own means of validation, because XLink/XPointer references are untyped.)

3. Resources can be freely interconnected with each other and with lexical-semantic resources that make use of the same representation formalism.

In the long perspective, the relationship between RDF and linguistics-specific standoff XML approaches like PAULA XML or GrAF should probably not be seen as competitive, but as complementary. We might expect, for example, an evolution from linguistic-specific formats to meta models that can be linearized in different ways. This may be compared, for example with the current status of the Lexical Markup Framework (LMF), which is regarded a meta model for the representation of lexical-semantic resources, where LMF-XML is only one of several possible linearizations (Francopoulo et al., 2009), but alternative linearizations have also been suggested, in particular in RDF/OWL (Francopoulo, 2007).

With PAULA and the closely related GrAF (both originate from early drafts of the Linguistic Annotation Framework, Ide and Romary, 2004) understood as meta models in this sense, standoff XML may continue to be used for representation and exchange of linguistic annotations, but RDF-based formalisms like POWLA can be regarded an alternative linearization for specific applications. In particular, RDF provides excellent means of querying graph data, it may thus be a format specifically applied for this purpose,

[11] A related problem is that this approach focuses on the conversion of individual XML files, and that IDs are thus not globally disambiguated. GrAF IDs are unambiguous only within an annotation layer and all annotation layers it depends on, independent annotation layers can thus contain identical IDs.

although applications beyond this limited domain (e.g., in NLP) can be imagined, in particular in the context of the evolving Linguistic Linked Open Data cloud (Chiarcos et al., this vol).

Acknowledgements

The work described in this paper was partially conducted at the University of Potsdam, Germany, in the context of the Collaborative Research Center (SFB) 632 "Information Structure", project D1 "Linguistic Database". I would like to thank Nancy Ide and Keith Suderman for sharing Steve Cassidy's RDF implementation of GrAF with me, and the anonymous reviewers for helpful comments and feedback.

5. References

G. Aguado de Cea, Á. I. de Mon-Rego, A. Pareja-Lora, and R. Plaza-Arteche. 2002. OntoTag: A semantic web page linguistic annotation model. In *Proceedings of the ECAI 2002 Workshop on Semantic Authoring, Annotation and Knowledge Markup*, Lyon, France, July.

C. F. Baker and C. Fellbaum. 2009. WordNet and FrameNet as Complementary Resources for Annotation. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 125–129, August.

C.F. Baker, C.J. Fillmore, and J.B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 86–90.

T. Berners-Lee. 2006. Design issues: Linked data. `http://www.w3.org/DesignIssues/LinkedData.html`.

A. Bies, M. Ferguson, K. Katz, and R. MacIntyre. 1995. Bracketing guidelines for Treebank II Style Penn Treebank Project. `ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/root.ps.gz`. version of January 1995.

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

A. Burchardt, S. Padó, D. Spohr, et al. 2008. Formalising Multi-layer Corpora in OWL/DL – Lexicon Modelling, Querying and Consistency Control. In *Proc. 3rd International Joint Conference on NLP (IJCNLP 2008)*, Hyderabad, India.

A. Busemann and K. Busemann. 2008. Toolbox self-training. Technical report, `http://www.sil.org`. Version 1.5.4, Oct 2008.

J. Carletta, S. Evert, U. Heid, et al. 2003. The NITE XML Toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3):353–363.

S. Cassidy. 2010. An RDF realisation of LAF in the DADA Annotation Server. In *Proc. 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-5)*, Hong Kong, January.

C. Chiarcos and J. Ritz. 2010. Qualitative and quantitative error analysis in context. In I. Rehbein, S. Schulte im Walde, A. Storrer, and M. Pinkal, editors, *Semantic Approaches in Natural Language Processing: Proceedings of the 10th Conference on Natural Language Processing (KONVENS-2010)*, pages 111–117, Saarbrücken, Germany, Sep. Universaar.

C. Chiarcos, S. Dipper, M. Götze, et al. 2008. A flexible framework for integrating annotations from different tools and tag sets. *TAL (Traitement Automatique des Langues)*, 49(2).

C. Chiarcos, J. Ritz, and M. Stede. 2009. By all these lovely tokens... Merging conflicting tokenizations. In *Proc. Third Linguistic Annotation Workshop (LAW-III), held in conjunction with ACL-IJCNLP 2009*, pages 35–43, Singapore, August.

C. Chiarcos, S. Hellmann, S. Nordhoff, et al. this vol. The Open Linguistics Working Group.

C. Chiarcos. 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)*, pages 37–40, Valetta, Malta, May.

C. Chiarcos. 2012. POWLA: Modeling linguistic corpora in OWL/DL. In *Proc. 9th Extended Semantic Web Conference (ESWC 2012)*, Heraklion, Crete, May.

C. Chiarcos. this vol. Ontologies of Linguistic Annotation: Survey and perspectives.

T. Declerck. 2006. Synaf: Towards a standard for syntactic annotation. In *Proc. of the 5th LREC Conference*.

S. Dipper, M. Götze, U. Küssner, and M. Stede. 2007. Representing and querying standoff XML. In *Proc. GLDV Conference 2007. Biannual Conference of the Society for Computational Linguistics and Language Technology (GLDV)*, Tübingen, Germany.

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

R. Eckart. 2008. Choosing an xml database for linguistically annotated corpora. *Sprache und Datenverarbeitung*, 32(1):7–22.

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

G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. 2009. Multilingual resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation*, 43(1):57–70.

G. Francopoulo. 2007. Strategy for an OWL specification of LMF. unpublished ms., `http://www.tagmatica.fr/lmf/StrategyForLMFInOWL29october2007.pdf`.

P. Gearon, A. Passant, and A. Polleres. 2012. SPARQL 1.1 Update. W3C Working Draft 05 January 2012. `http://www.w3.org/TR/sparql11-update`.

M. Grönroos and M. Miettinen. 2004. Infrastructure for collaborative annotation of speech. In *Proc. LREC 2004*, Genoa, May.

O. Hartig, C. Bizer, and J.C. Freytag. 2009. Executing

SPARQL queries over the web of linked data. *The Semantic Web-ISWC 2009*, pages 293–309.

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

S. Hellmann. 2010. The semantic gap of formalized meaning. In *Proc. 7th Extended Semantic Web Conference (ESWC 2010)*, Heraklion, Greece, May 30th – June 3rd.

B. Hellwig, D. Van Uytvanck, and M. Hulsbosch. 2008. ELAN - Linguistic Annotator. Technical report, `http://www.lat-mpi.eu/tools/elan`. version of 2008-07-31.

N. Ide and J. Pustejovsky. 2010. What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proc. 2nd International Conference on Global Interoperability for Language Resources (ICGL 2010)*.

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

N. Ide and L. Romary. 2006. Representing linguistic corpora and their annotations. In *Proceedings of the Fifth Language Resources and Evaluation Conference*. Citeseer.

N. Ide and K. Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of Linguistic Annotation Workshop (LAW 2007)*, pages 1–8.

N. Ide, C. Baker, C. Fellbaum, C. Fillmore, and R. Passonneau. 2008. MASC: The Manually Annotated Sub-Corpus of American English. In *Proc. 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakesh, Morocco.

N. Ide, C. Fellbaum, C. Baker, and R. Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL-2010*, pages 68–73.

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

D.T. Langendoen, S. Farrar, and W. Lewis. 2002. Bridging the markup gap: Smart search engines for language researchers. In *Proc. LREC-2002 Workshop on Resources and Tools for Field Linguistics*, pages 24–1, Las Palmas, Spain, May.

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

N. Mazziotta. 2010. Building the syntactic reference corpus of medieval french using notabene rdf annotation tool. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 142–146. Association for Computational Linguistics.

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

M. O'Donnell. 2000. RSTTool 2.4 – A markup tool for Rhetorical Structure Theory. In *Proc. International Natural Language Generation Conference (INLG'2000)*, pages 253–256, Mitzpe Ramon, Israel.

OpenLink Software. 2009. Virtuoso Open-Source Edition. `http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main`.

E. Prud'Hommeaux and A. Seaborne. 2008. SPARQL query language for RDF. *W3C working draft*, 4(January).

L. Romary, A. Zeldes, and F. Zipser. 2011. <tiger2/> - Serialising the ISO SynAF Syntactic Object Model. *Arxiv preprint arXiv:1108.0631*.

E. Rubiera, L. Polo, D. Berrueta, and A. El Ghali. 2012. TELIX: An RDF-based model for linguistic annotation. In *Proc. 9th Extended Semantic Web Conference (ESWC 2012)*, May.

F. Sasaki, A. Witt, D. Gibbon, and T. Trippel. 2004. Concept-based queries: Combining and reusing linguistic corpus formats and query languages. In *Proceedings of LREC 2004*.

Andrea C. Schalley. 2012. TYTO – A collaborative research tool for linked linguistic data. In C. Chiarcos, S. Nordhoff, and S. Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*. Springer, Heidelberg. p. 139-149.

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

T. Schmidt. 2004. Transcribing and annotating spoken language with EXMARaLDA. In *Proc. LREC 2004 Workshop on XML based richly annotated corpora*, Lisbon.

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

M. Stede and S. Heintze. 2004. Machine-assisted rhetorical structure annotation. In *Proc. 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.

H. Telljohann, E. W. Hinrichs, and S. Kübler. 2003. Stylebook for the Tübingen treebank of written German (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany.

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

A. Zeldes, J. Ritz, A. Lüdeling, and C. Chiarcos. 2009. ANNIS: A search tool for multi-layer annotated corpora. In *Proc. Corpus Linguistics*, pages 20–23, Liverpool, UK, July.

F. Zipser and L. Romary. 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, May.