POWLA: Modeling linguistic corpora in OWL/DL

Christian Chiarcos

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

Abstract. This paper describes POWLA, a generic formalism to represent linguistic annotations in an interoperable way by means of OWL/DL. Unlike other approaches in this direction, POWLA is not tied to a specific selection of annotation layers, but it is designed to support any kind of text-oriented annotation.

1 Background

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 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.

At the moment, state-of-the-art approaches on corpus interoperability build on standoff-XML [5, 26] and relational data bases [12, 17]. The underlying data models are, however, graph-based, and this paper pursues the idea that RDF and

[1] For example, parts of the Penn Treebank [30], 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 [31].

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 described 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).

POWLA takes its conceptual point of departure from the assumption that any linguistic annotation can be represented by means of directed graphs [3, 26]: Aside from the primary data (text), linguistic annotations consist of three principal components, i.e., segments (spans of text, e.g., a phrase), relations between segments (e.g., dominance relation between two phrases) and annotations that describe different types of segments or relations.

In graph-theoretical terms, segments can be formalized as **nodes**, relations as **directed edges** and annotations as **labels** attached to nodes and/or edges. These structures can then be connected to the primary data by means of pointers. A number of generic formats were proposed on the basis of such a mapping from annotations to graphs, including ATLAS [3] and GrAF [26]. Below, this is illustrated for the PAULA data model, that is underlying the POWLA format. Traditionally, PAULA is serialized as an XML standoff format, it is specifically designed to support multi-layer corpora [12], and it has been successfully applied to develop an NLP pipeline architecture for Text Summarization [36], and for the development of the corpus query engine ANNIS [39]. See Fig. 1 for an example for the mapping of syntax annotations to the PAULA data model.

RDF also formalizes directed (multi-)graphs, so, an RDF linearization of the PAULA data model yields a generic RDF representation of text-based linguistic annotations, and corpora in general. The idea underlying POWLA is to represent linguistic annotations by means of RDF, and to employ OWL/DL to define data types and consistency constraints for these RDF data.

2 POWLA

This section first summarizes the data types in PAULA, then their formalization in POWLA, and then the formalization of linguistic corpora with OWL/DL.

2.1 PAULA data types

The data model underlying PAULA is derived from labeled directed acyclic (hyper)graphs (DAGs). Its most important data types are thus different types of nodes, edges and labels [14]:

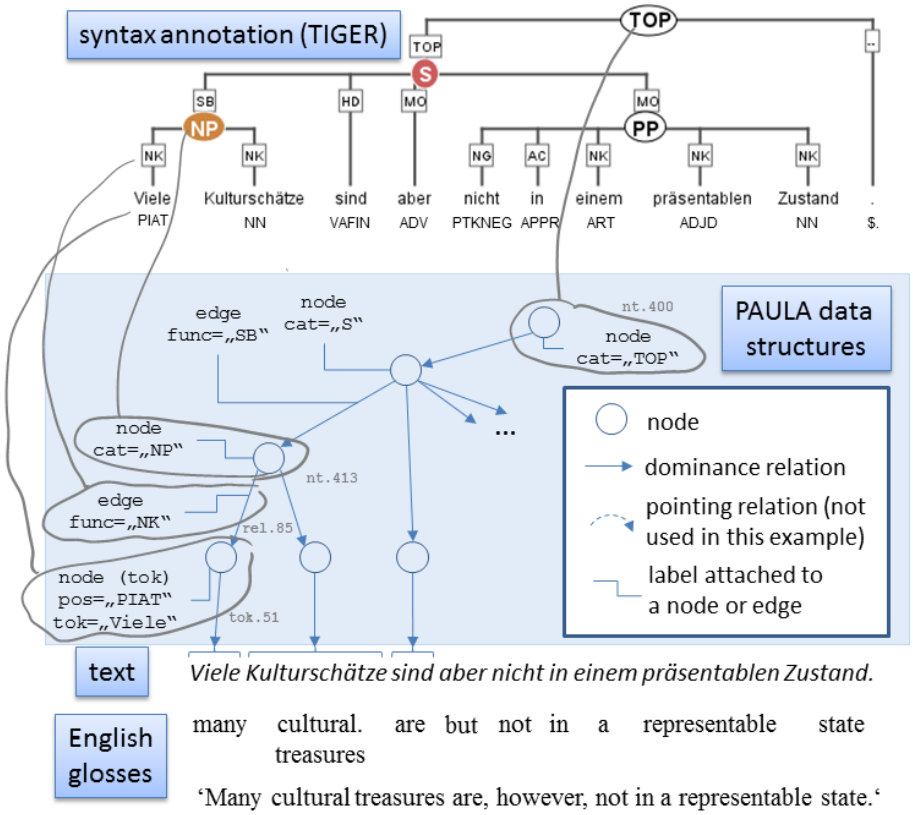

Fig. 1. Using PAULA data structures for constituent syntax (German example sentence taken from the Potsdam Commentary Corpus, [35])

node	(structural units of annotation)
terminal	character spans in the primary data
markable	span of terminals (data structure of flat, layer-based annotations defined e.g., by their position)
struct	hierarchical structures (forming trees or DAGs), establishes parent-child relations between a (parent) struct and child nodes (of any type)
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)
feature	linguistic annotation

A unique feature of PAULA as compared to other generic formats is that it introduces a clear distinction between two types of edges that differ with respect to their relationship to the primary data. For hierarchical structures, e.g., phrase structure trees, a notion of **coverage inheritance** is necessary, i.e., the text covered by a child node is always covered by the parent node, as well. 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 need 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. Edges without coverage inheritance are referred to in PAULA 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.

Additionally, PAULA includes specifications for the organization of annotations, i.e.

layer (grouping together nodes and relations that represent a single annotation layer, in PAULA represented by a namespace prefixed to a label, e.g., `tiger:...` for original TIGER XML)
document (or 'annoset', grouping together all annotations of one single resource of textual data)
collection (an annoset that comprises not only annotations, but also other annosets, e.g., constituting a subcorpus)
corpus (a collection not being part of another collection)

Also, layers and documents can be assigned labels, that correspond to metadata, rather than annotations, e.g., date of creation or name of the annotator.

2.2 POWLA TBox: The POWLA ontology

The POWLA ontology represents a straight-forward implementation of the PAULA data types in OWL/DL. `Node`, `Relation`, `Layer` and `Document` correspond to PAULA node, edge, layer and document, respectively, and they are defined as subclasses of `POWLAElement`.

A `POWLAElement` is anything that can carry a label (property `hasLabel`). For `Document` and `Layer`, these annotations contain metadata (subproperty `hasMetadata`), for `Node` and `Relation`, they contain string values of the linguistic annotation (subproperty `hasAnnotation`). The properties `hasAnnotation` and `hasMetadata` are, however, not to be used directly, but rather, subproperties are to be created for every annotation phenomenon, e.g., `hasPos` for part-of-speech annotation, or `hasCreationDate` for the date of creation.

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. 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 PAULA terminal. A `Nonterminal` is a `Node` with 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 a `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. In this case, using OWL/DL to model linguistic data types allows us to *infer* the relevant distinction, the data model can thus be pruned from artifacts necessary for visualization, etc.

The concept `Root` was introduced for organizational reasons. It corresponds to a `Nonterminal` that does not have a parent (and may be either a `Terminal` or a `Nonterminal`). `Root`s play an important role in structuring annosets: A `DocumentLayer` (a `Layer` defined for one specific `Document`) can be defined as a collection of `Root`s, so that it is no longer necessary to link every `Node` with the corresponding `Layer`, but only the top-most `Node`s. In ANNIS, `Root`s are currently calculated during runtime.

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 `nextTerminal` properties. This is, however, a preliminary solution and may be revised. Further, `Terminal`s may be linked to the primary data (strings) in accordance to the currently developed NLP Interchange Format (NIF).[2]

The `POWLAElement Layer` corresponds to a layer in PAULA. It is characterized by an ID, and can be annotated with metadata. `Layer` refers to a *phenomenon*, however, not to one specific layer within a document (annoset). Within a document, the subconcept `DocumentLayer` is to be used, that is assigned all `Root` nodes associated with this particular layer (property `rootOfDocument`). A `Root` may have at most one `Layer`.

The `POWLAElement Document` corresponds to a PAULA document, i.e., an annoset, or annotation project that assembles all annotations of a body of text and its parts. An annoset may contain other annotation projects (`hasSubDocument`), if it does so, it represents a collection of documents (e.g., a subcorpus, or a pair of texts in a parallel corpus), otherwise, it contains the annotations of one particular text. In this case, it is a collection of different `DocumentLayers` (property

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

`hasDocument`). A `Corpus` is a `Document` that is not a subdocument of another `Document`.

A diagram showing core components of the ontology is shown in Fig. 2.

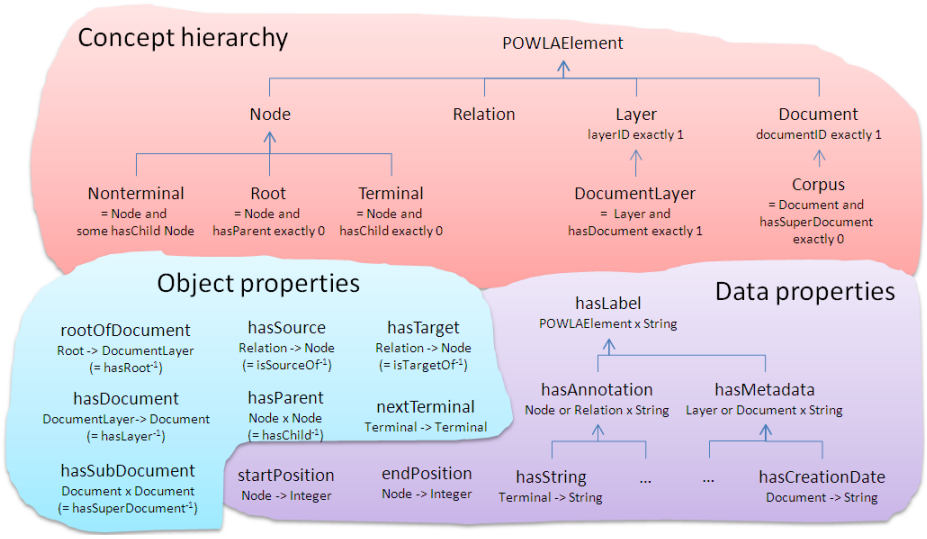

Fig. 2. The POWLA ontology (fragment)

2.3 POWLA ABox: Modelling linguistic annotations in POWLA

The POWLA ontology defines data types that can now be used to represent linguistic annotations. Considering the phrase *viele Kulturschätze* 'many cultural treasures' from the German sentence analyzed in Fig. 1, `Terminal`s, `Nonterminal`s and `Relation`s are created as shown in Figs. 3 and 4.

`Terminal`s `tok.51` and `tok.52` are the terminals *Viele* and *Kulturschätze*. The `Nonterminal` `nt.413` is the NP dominating both, the `Relation` `rel.85` is the relation between `nt.413` and `tok.51`. The properties `hasPos`, hasCat and hasFunc are subproperties of `hasAnnotation` that were created to reflect the `pos`, `cat` and `func` labels of nodes and edges in Fig. 1. `Relation` `rel.85` is marked as a dominance relation by the accompanying `hasChild` relation between its source and target.

As for corpus organization, the `Root` of the tree dominating `nt.413` is `nt.400` (the node with the label TOP in Fig. 1), and it is part of a `DocumentLayer` with the ID `tiger`. This `DocumentLayer` is part of a `Document`, etc., but for reasons of brevity, this is not shown here.

It should be noted that this representation in OWL/RDF is by no means complete. Inverse properties, for example, are missing. Using a reasoner, however, the missing RDF triples can be inferred from the information provided

```
<powla:Terminal rdf:ID="tok.51">
    <powla:startPosition>434</powla:startPosition>
    <powla:endPosition>438</powla:endPosition>
    <powla:hasString>Viele</powla:hasString>
    <powla:hasPos>PIAT</powla:hasPos>
    <powla:nextTerminal rdf:about="#tok.52"/>
</powla:Terminal>
<powla:Terminal rdf:ID="tok.52">
    <powla:startPosition>439</powla:startPosition>
    <powla:endPosition>450</powla:endPosition>
    <powla:hasString>Kulturschätze</powla:hasString>
    <powla:hasPos>NN</powla:hasPos>
    ...
```

Fig. 3. Examples of `Terminal`s in POWLA

```
<powla:Nonterminal rdf:ID="nt.413">
    <powla:hasChild rdf:about="#tok.51"/>
    <powla:hasChild rdf:about="#tok.52"/>
    <powla:startPosition>434</powla:startPosition>
    <powla:endPosition>450</powla:endPosition>
    <powla:hasCat>NP</powla:hasCat>
    <powla:isSourceOf rdf:about="#rel.85"/>
    ...
<powla:Relation rdf:ID="rel.85">
    <powla:hasSource rdf:about="#nt.413"/>
    <powla:hasTarget rdf:about="#tok.51"/>
    <powla:hasFunc>NK</powla:hasFunc>
    ...
```

Fig. 4. Examples of `Nonterminal`s and `Relation`s in POWLA

explicitly. A reasoner would also allow us to verify whether the necessary cardinality constraints are respected, e.g., every `Root` assigned to a `DocumentLayer`, etc.

Although illustrated here for syntax annotations only, the conversion of other annotation layers from PAULA to POWLA is similarly straight-forward. As sketched above, all PAULA data types can be modelled in OWL, and by `Root` and `DocumentLayer`, also PAULA namespaces ("tiger" for the example in Fig. 1) can be represented.

3 Corpora as Linked Data

With POWLA specifications as sketched above, linguistic annotations can be represented in RDF, with OWL/DL-defined data types. From the perspective of computational linguistics, this offers a number of advantages as compared to state-of-the-art solutions using standoff XML (i.e., a bundle of separate XML files that are densely interconnected with XLink and XPointer) as representation formalism and relational data bases as means for querying (e.g., [12] for PAULA XML, or [26, 17] for GrAF):

1. Using OWL/DL reasoners, RDF data can be validated. (The semantics of XLink/XPointer references in standoff XML cannot be validated with standard tools, because XML references are *untyped*.)
2. Using RDF as representation formalism, multi-layer corpora can be directly processed with off-the-shelf data bases and queried with standard query languages. (XML data bases do not provide efficient standoff XML querying [18], relational data bases require an additional conversion step.)
3. RDF allows to combine information from different types of linguistic resources, e.g., corpora and lexical-semantic resources. They can thus be queried with the same query language, e.g., SPARQL. (To formulate similar queries using representation formalisms that are specific to either corpora or lexical-semantic resources like GrAF, or LMF [20], novel means of querying would yet have to be developed.)
4. RDF allows to connect linguistic corpora directly with repositories of reference terminology, thereby supporting the interoperability of corpora. (Within GrAF, references to the ISOcat data category registry [28] should be used for this purpose, but this does not make use of mechanisms that already have been standardized.)

The first benefit is sufficiently obvious not to require an in-depth discussion here, the second and the fourth are described in [11] and [10], respectively. Here, I focus on the third aspect, which can be more generally described as treating linguistic corpora as linked data.

The application of RDF to model linguistic corpora is sufficiently motivated from benefits (1) and (2), and this has been the motivation of several RDF/OWL formalizations of linguistic corpora [4, 22, 32, 7]. It is, however, not only a way to represent linguistic data, but also, other forms of data, and in particular,

to establish links between such resources. This is captured in the **linked data paradigm** [2] that consists of four rules: Referred entities should be designated by URIs, these URIs should be resolvable over http, data should be represented by means of standards such as RDF, and a resource should include links to other resources. With these rules, it is possible to follow links between existing resources, and thereby, to find other, related, data. If published as Linked Data, corpora represented in RDF can be linked with other resources already available in the Linked Open Data (LOD) cloud.[3]

To this end, integrating corpora into the LOD cloud has not been suggested, probably mostly because of the gap between the linguistics and the Semantic Web communities. Recently, however, some interdisciplinary efforts have been brought forward in the context of the Open Linguistics Working Group of the Open Knowledge Foundation [13], an initiative of experts from different fields concerned with linguistic data, whose activities – to a certain extent – converge towards the creation of a Linguistic Linked Open Data (LLOD) (sub-)cloud that will comprise different types of linguistic resources, unlike the current LOD cloud also linguistic corpora. The following subsections describe ways in which linguistic corpora may be linked with other LOD (resp. LLOD) resources.

3.1 Grounding the POWLA ontology in existing schemes

POWLA is grounded in Dublin Core (corpus organization), and closely related to the NLP Interchange Format NIF (elements of annotation).

In terms of Dublin Core, POWLA `Document` is a `dctype:Collection` (it aggregates either different `DocumentLayer`s or further `Document`s), a POWLA `Layer` is a `dctype:Dataset`, in that it provides data encoded in a defined structure. POWLA represents the primary data only in the values of `hasString` properties, hence, there is no `dctype:Text` represented here. Extending `Terminal`s with string references as specified by NIF would allow us to point directly to the primary data (`dctype:Text`).

With respect to NIF, POWLA is more general (but also, less compact). Many NIF data structures can be regarded as specializations of POWLA categories, others are equivalent. For example, a NIF `String` corresponds to a POWLA `Node`, however, with more specific semantics, as it is tied to a stretch of text, whereas a POWLA `Node` may also be an empty element. The POWLA property `hasString` corresponds to NIF `anchorOf`, yet `hasString` is restricted to `Terminal`s, whereas `anchorOf` is obligatory for all NIF *String*s. Hence, both are not equivalent, however, it is possible to construct a generalization over NIF and POWLA that allows to define both data models as specializations of a common underlying model for NLP analyses and corpus annotations. The development of such a generalization and a transduction from NIF to POWLA is currently in preparation. NIF and POWLA are developed in close synchronization, albeit optimized for different application scenarios.

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

A key difference between POWLA and NIF is the representation of `Relation`s, that correspond to object properties in NIF. Modeling edges as properties yields a compact representation in NIF (one triple per edge). In POWLA, it should be possible to assign a `Relation` to a `DocumentLayer`, i.e., a property with a `Relation` as subject. OWL/DL conformity requires to model `Relation`s to be concepts (with `hasSource` and `hasTarget` at least 3 triples per edge). For the transduction from NIF to POWLA, such incompatibilities require more extensive modifications. At the moment, the details of such a transduction are actively explored by POWLA and NIF developers.

3.2 Linking corpora with lexical-semantic resources

So far, two resources have been converted using POWLA, including the NEGRA corpus, a German newspaper corpus with annotations for morphology and syntax [34], as well as coreference [33], and the MASC corpus, a manually annotated subcorpus of the Open American Corpus annotated for a great band-width of phenomena [23]. MASC is represented in GrAF, and a GrAF converter has been developed [11].

MASC includes semantic annotations with FrameNet and WordNet senses [1]. WordNet senses are represented by sense keys as string literals. As sense keys are stable across different WordNet versions, this annotation can be trivially rendered in URIs references pointing to an RDF version of WordNet. (However, the corresponding WordNet version 3.1 is not yet available in RDF.)

FrameNet annotations in MASC make use of feature structures (attribute-value pairs where the value can be another attribute-value pair), which are not yet fully supported by the GrAF converter. However, reducing feature structures to simple attribute-value pairs is possible. The values are represented in POWLA as literals, but can likewise be transduced to properties pointing to URIs, if the corresponding FrameNet version is available. An OWL/DL version of FrameNet has been announced at the FrameNet site, it is, however, available only after registration, and hence, not strictly speaking an open resource.

With this kind of resources being made publicly available, it would be possible to develop queries that combine elements of both the POWLA corpus and lexical-semantic resources. For example, one may query for sentences about *land*, i.e., 'retrieve every (POWLA) sentence that contains a (WordNet-)synonym of *land*'. Such queries can be applied, for example, to develop semantics-sensitive corpus querying engines for linguistic corpora.

3.3 Meta data and terminology repositories

In a similar way, corpora can also be linked to other resources in the LOD cloud that provide identifiers that can be used to formalize corpus meta data, e.g., provenance information. Lexvo [15] for example, provides identifiers for languages, GeoNames [37] provides codes for geographic regions. ISOcat [29] is another repository of meta data (and other) categories maintained by ISO TC37/SC4, for which an RDF interface has recently been proposed [38].

Similarly, references to terminology repositories may be used instead of string-based annotations. For example, the OLiA ontologies [8] formalize numerous annotation schemes for morphosyntax, syntax and higher levels of linguistic description, and provide a linking to the morphosyntactic profile of ISOcat [9] with the General Ontology of Linguistic Description [19], and other terminology repositories. By comparing OLiA annotation model specifications with tags used in a particular layer in a particular layer annotated according to the corresponding annotation scheme, the transduction from string-based annotation to references to community-maintained category repository is eased. Using such a resource to describe the annotations in a given corpus, it is possible to abstract from the surface form a particular tag and to interpret linguistic annotations on a conceptual basis.

Linking corpora with terminology and metadata repositories is thus a way to achieve **conceptual interoperability** between linguistic corpora and other resources.

4 Results and discussion

This paper presented preliminaries for the development of a generic OWL/DL-based formalism for the representation of linguistic corpora. As compared to related approaches [4, 22, 32], the approach described here is not tied a restricted set of annotations, but applicable to any kind of text-based linguistic annotation, because it takes its point of departure from a generic data model known to be capable to represent any kind of linguistic annotation.

One concrete advantage of the OWL/RDF formalization is that it represents a standardized to represent heterogeneous data collections (whereas standard formats developed within the linguistic community are still under development): With RDF, a standardized representation formalism for different corpora is available, and with datatypes 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 [24]. In comparison to other formalisms developed in this direction (including ATLAS [3], NXT [6], GrAF and PAULA), it does, however, not propose a special-purpose XML standoff format, but rather, it employs existing and established standards with broad technical support (schemes, parsers, data bases, query language, editors/browsers, reasoners) and an active and comparably large community. Standard formats specifically designed for linguistic annotations as developed in the context of the ISO TC37/SC4 (e.g., GrAF), are, however, still under development.

As mentioned above, the development of POWLA as a representation formalism for annotated linguistic corpora is coordinated with the development of the NLP Interchange Format NIF [21]. Both formats are designed to be mappable, but they are optimized for different fields of application: POWLA is developed to represent annotated corpora with a high degree of genericity, whereas NIF is a compact and NLP-specific format for a restricted set of annotations. At the

moment, NIF is capable to represent morphosyntactic and syntactic annotations only, the representation of more complex forms of annotation, e.g., alignment in a parallel corpus, has not been addressed so far. Another important difference is that NIF lacks any formalization of corpus structure. NIF is thus more compact, but the POWLA representation is more precise and more expressive, and both are designed to be mappable. This means that NIF annotations can be converted to POWLA representations, and then, for example, combined with other annotation layers.

PAULA is closely related to other standards: It is based on early drafts for the Linguistic Annotation Framework [25, LAF] developed by the ISO TC37/SC4. Although it predates the official LAF linearization GrAF [27] by several years [16], it shares its basic design as an XML standoff format and the underlying graph-based data model. One important difference is, however, the treatment of segmentation [14]. While PAULA provides formalized terminal elements with XLink/XPointer references to spans in the primary data, GrAF describes segments by a sequence of numerical 'anchors'. Although the resolution of GrAF anchors is comparable to that of `Terminal`s in PAULA, the key difference is that anchor resolution is not formalized within the GrAF data model.

This has implications for the RDF linearizations of GrAF data: The RDF linearization of GrAF recently developed by [7] represents anchors as literal strings consisting of two numerical, space-separated IDs (character offsets) like in GrAF. This approach, however, provides no information how these IDs should be interpreted (the reference to the primary data is not expressed). In POWLA, `Terminal`s are modeled as independent resources and information about the surface string and the original order of tokens is provided. Another difference is that this RDF linearization of GrAF is based on single GrAF files (i.e., single annotation layers), and that it does not build up a representation of the entire annotation project, but that corpus organization is expressed implicitly through the file structure which is inherited from the underlying standoff XML. It is thus not directly possible to formulate SPARQL queries that refer to the same annotation layer in different documents or corpora.

Closer to our conceptualization is [4] who used OWL/DL to model a multilayer corpus with annotations for syntax and semantics. The advantages of OWL/DL for the representation of linguistic corpora were carefully worked out by the authors. Similar to our approach, [4] employed an RDF query language for querying. However, this approach was specific to a selected resource and its particular annotations, whereas POWLA, is a generic formalism for linguistic corpora based on established data models developed to the interoperable formalization of arbitrary linguistic annotations assigned to textual data.

As emphasized above, a key advantage of the representation of linguistic resources in OWL/RDF is that they can be published as Linked Data [2], i.e., that different corpus providers can provide their annotations at different sites, and link them to the underlying corpus. For example, the Prague Czech-English Dependency Treebank[4] which is an annotated translation of parts of the Penn

[4] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T25`

Treebank, could be linked to the original Penn Treebank. Consequently, the various and rich annotations applied to the Penn Treebank [31] can be projected onto Czech.[5] Similarly, existing linkings between corpora and lexical-semantic resources, represented so far by string literals, can be transduced to URI references if the corresponding lexical-semantic resources are provided as linked data.An important aspect here is that corpora can be linked to other resources from the Linked Open Data cloud *using the same formalism*.

Finally, linked data resources can be used to formalize meta data or linguistic annotations. This allows, for example, to use information from terminology repositories to query a corpus. As such, the corpus can be linked to terminology repositories like the OLiA ontologies, ISOcat or GOLD, and these community-defined data categories can be used to formulate queries that are independent from the annotation scheme, but use an abstract, and well-defined vocabulary. In this way, linguistic annotations in POWLA are not only structurally interoperable (they use the same representation formalism), but also conceptually interoperable (they use the same vocabulary).

References

1. C. F. Baker and C. Fellbaum. WordNet and FrameNet as Complementary Resources for Annotation. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 125–129, August 2009.
2. T. Berners-Lee. Design issues: Linked data. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006.
3. S. Bird and M. Liberman. A formal framework for linguistic annotation. *Speech Communication*, 33(1-2):23–60, 2001.
4. A. Burchardt, S. Padó, D. Spohr, A. Frank, and U. Heid. Formalising Multi-layer Corpora in OWL/DL – Lexicon Modelling, Querying and Consistency Control. In *Proceedings of the 3rd International Joint Conf on NLP (IJCNLP 2008)*, Hyderabad, 2008.
5. J. Carletta, S. Evert, U. Heid, and J. Kilgour. The NITE XML Toolkit: data model and query. *Language Resources and Evaluation Journal (LREJ)*, 39(4):313–334, 2005.
6. J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. The NITE XML Toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3):353–363, 2003.
7. S. Cassidy. An rdf realisation of laf in the dada annotation server. *Proceedings of ISA-5, Hong Kong*, 2010.
8. C. Chiarcos. An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16, 2008.
9. C. Chiarcos. Grounding an ontology of linguistic annotations in the Data Category Registry. In *Workshop on Language Resource and Language Technology Standards (LR&LTS 2010), held in conjunction with LREC 2010*, Valetta, Malta, May 2010.

[5] Unlike existing annotation projection approaches, however, this would not require that English annotations are directly applied to the Czech data – which introduces additional noise –, but instead, SPARQL allows us to follow the entire path from Czech to English to its annotations, with the noisy part (the Czech-English alignment) clearly separated from the secure information (the annotations).

10. C. Chiarcos. Interoperability of corpora and annotations. In C. Chiarcos, S. Nordhoff, and S. Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 161–179, Heidelberg, 2012. Springer.
11. C. Chiarcos. A generic formalism to represent linguistic corpora in RDF and OWL/DL. In *8th International Conference on Language Resources and Evaluation (LREC-2012)*, accepted.
12. C. Chiarcos, S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz, and M. Stede. A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *Traitement Automatique des Langues*, 49(2), 2009.
13. C. Chiarcos, S. Hellmann, and S. Nordhoff. The Open Linguistics Working Group of the Open Knowledge Foundation. In C. Chiarcos, S. Nordhoff, and S. Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 153–160, Heidelberg, 2012. Springer.
14. C. Chiarcos, J. Ritz, and M. Stede. By all these lovely tokens ... Merging conflicting tokenizations. *Journal of Language Resources and Evaluation (LREJ)*, to appear.
15. G. De Melo and G. Weikum. Language as a foundation of the Semantic Web. In *Proceedings of the 7th International Semantic Web Conference (ISWC 2008)*, volume 401, 2008.
16. S. Dipper. XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, Berlin, Germany, 2005.
17. K. Eckart, A. Riester, and K. Schweitzer. A discourse information radio news database for linguistic analysis. In C. Chiarcos, S. Nordhoff, and S. Hellmann, editors, *Linked Data in Linguistics*. Springer, 2012.
18. R. Eckart. Choosing an xml database for linguistically annotated corpora. *Sprache und Datenverarbeitung*, 32(1):7–22, 2008.
19. S. Farrar and D. T. Langendoen. 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, 2010.
20. G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. Multilingual resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation*, 43(1):57–70, 2009.
21. S. Hellmann. The semantic gap of formalized meaning. In *The 7th Extended Semantic Web Conference (ESWC 2010)*, Heraklion, Greece, May 30th – June 3rd 2010.
22. S. Hellmann, J. Unbehauen, C. Chiarcos, and A. Ngonga Ngomo. The TIGER Corpus Navigator. In *9th International Workshop on Treebanks and Linguistic Theories (TLT-9)*, pages 91–102, Tartu, Estonia, 2010.
23. N. Ide, C. Fellbaum, C. Baker, and R. Passonneau. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL-2010*, pages 68–73, 2010.
24. N. Ide and J. Pustejovsky. What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong, China, 2010.
25. N. Ide and L. Romary. International standard for a linguistic annotation framework. *Natural language engineering*, 10(3-4):211–225, 2004.

26. N. Ide and K. Suderman. GrAF: A Graph-based Format for Linguistic Annotations. In *Proceedings of The Linguistic Annotation Workshop (LAW) 2007*, pages 1–8, Prague, Czech Republic, 2007.
27. N. Ide and K. Suderman. GrAF: A graph-based format for linguistic annotations. In *Proceedings of The Linguistic Annotation Workshop (LAW) 2007*, pages 1–8, Prague, Czech Republic, 2007.
28. M. Kemps-Snijders, M. Windhouwer, P. Wittenburg, and S. Wright. ISOcat: Corralling data categories in the wild. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008.
29. M. Kemps-Snijders, M. Windhouwer, P. Wittenburg, and S. Wright. ISOcat: Remodelling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies*, 4(4):261–276, 2009.
30. M. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.
31. J. Pustejovsky, A. Meyers, M. Palmer, and M. Poesio. Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference. In *Proc. of ACL Workshop on Frontiers in Corpus Annotation 2005*, 2005.
32. E. Rubiera, L. Polo, D. Berrueta, and A. El Ghali. TELIX: An RDF-based model for linguistic annotation. In *ESWC 2012*, accepted.
33. M. Schiehlen. Optimizing algorithms for pronoun resolution. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 515–521, Geneva, August 2004.
34. W. Skut, T. Brants, B. Krenn, and H. Uszkoreit. A linguistically interpreted corpus of German newspaper text. In *Proc. ESSLLI Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany, 1998.
35. M. Stede. The Potsdam Commentary Corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain, 2004.
36. M. Stede and H. Bieler. The mots workbench. In A. Mehler, K.-U. Kühnberger, H. Lobin, H. Lüngen, A. Storrer, and A. Witt, editors, *Modeling, Learning, and Processing of Text Technological Data Structures*, volume 370 of *Studies in Computational Intelligence*, pages 15–34. Springer Berlin / Heidelberg, 2012.
37. B. Vatant and M. Wick. GeoNames ontology. `http://www.geonames.org/ontology`, accessed March 19, 2012, Feb 2012. version 3.01.
38. M. Windhouwer and S. E. Wright. Linking to linguistic data categories in ISOcat. In *Linked Data in Linguistics (LDL 2012)*, Frankfurt/M., Germany, Mar accepted.
39. A. Zeldes, J. Ritz, A. L?deling, and C. Chiarcos. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics*, pages 20–23, Liverpool, UK, July 2009.