A generic formalism to represent linguistic corpora in RDF and OWL/DL

October 21, 2011

Abstract

This paper introduces POWLA, as formalism to represent linguistic corpora by means of semantic web formalisms, in particular, 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 [13, 9] that is based on early sketches of the ISO TC37/SC4 Linguistic Annotation Framework [17]. As opposed to existing standoff XML linearizations for such generic data models (e.g., NXT [6], PAULA XML [13] or GrAF [18]), it uses RDF as representation formalism and OWL/DL for validation. The paper discusses advantages of this approach.

1 From generic data models to RDF and OWL

POWLA is an OWL/DL serialization of PAULA, the data model the generic interchange format PAULA XML that was developed at the Collaborative Research Center (SFB) 632 "Information Structure" [13, 8, 9]. PAULA itself originates from early drafts of the Linguistic Annotation Framework [17], PAULA XML is thus closely related to the later ISO TC37/SC4 format GrAF [18].

PAULA is the input format of the linguistic information system ANNIS [30, 8, 32] that was developed with a focus on multi-layer annotations. PAULA was thus developed to support the loss-less representation of arbitrary kinds of text-oriented linguistic annotation, and in particular the merging of annotations produced by different tools, including annotations for morphology, dependency syntax, constituent syntax, coreference, and discourse structure.[1]

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 the PAULA data model. Consequently, all annotations currently covered by PAULA can also be represented by means of Semantic Web standards.

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, this approach offers three crucial advantages:

1. The increasing number of RDF data bases provides us with convenient means for the management of linguistic data collections.

[1] For example, PAULA was applied to represent multiple independent syntax annotations of the same text [7], or syntax, coreference and discourse structure annotation at the same time, [10]. At the moment, a wide range of input formats is supported, including TIGER XML [19], EXMARaLDA [27], MMAX2 [22], Toolbox [5], as well as tab-separated text (CSV), generic XML, and annotations produced by special-purpose tools such as the RSTTool [23, for discourse structure annotations] and ConAno [29, for the annotation of discourse connectives]. Using existing converters to one of these source formats, an even broader band-width of tools and formats is supported, e.g., via TIGER XML the Penn Treebank bracketing notation [21], and via EXMARaLDA ELAN [16] and Praat [3].

2. 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.

3. Resources can be freely interconnected with each other and with lexical-semantic resources that make use of the same representation formalism, e.g., those available from the Linked Open Data cloud.[2]

2 POWLA

The Resource Description Framework (RDF) formalizes relations between entities by means of a directed (hyper)graph where both the nodes (resources) connected and edges (relations) between them can be assigned labels. Since [2], this data structure (or, more specifically, directed acyclic graphs) have also been identified as a means to develop generic formats for linguistic annotations. Connecting these independent developments, we propose an RDF formalization of an established, graph-based generic format, PAULA, and the application of OWL/DL to define datatypes and constraints on corpora of RDF data.

POWLA consists of two parts: the POWLA TBox (or 'POWLA ontology') defines data categories, the POWLA ABox contains the actual corpus data.

The **POWLA TBox** represents a straight-forward implementation of the data types of PAULA in an OWL/DL ontology[3]

All POWLA concepts are subconcepts of `POWLAElement`: A `POWLAElement` is anything that can carry a label (property `hasLabel`). For the subconcepts `Node` and `Relation` (see below) that are used to represent linguistic annotations and discussed further below, this label corresponds to the String value of the linguistic annotation (subproperty `hasAnnotation`). The properties `hasLabel` and `hasAnnotation` 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 `hasCat` for phrase labels in the syntax annotation.

Aside from `Node` and `Relation`, the remaining subconcepts of `POWLAElement` are `Document` and `Layer`. These are concerned with corpus organization and 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. 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, i.e., the minimal unit of annotation. A `Nonterminal` is a `Node` that has at least one `hasChild` property.

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`). `Roots` play an important role in the structuring of annotation projects, they can be used to define the relevant context to be extracted for a query match, and they play an important role in the visualization of tree annotations in existing multi-layer data bases. Based on our experience with the ANNIS data base, where top-level nodes of trees are currently calculated at run time, we decided to represent `Root` explicitly in the data model.

[2] Representative lexico-semantic resources include RDF versions of WordNet (e.g., `http://thedatahub.org/dataset/vu-wordnet`), FrameNet (`http://www.haphan.co.uk/owl/download/5vj7INLRuv.owl`, previously available from `http://wiki.loa-cnr.it/index.php?title=LoaWiki:OFN`) and the Wikipedia (i.e., the DBpedia [1]).

[3] The POWLA ontology, tools and further documentation are available from `http://purl.org/powla`.

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 more efficiently by means of pre- and post-order as defined by [31], and Terminals may be linked to strings in accordance to the currently developed NLP Interchange Format (NIF).[4]

The POWLA TBox posits 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, a Terminal or a Root. This can also be exploited to differentiate between markables and structs, that are different subtypes of node in PAULA that differ in their subsequent visualization: Markables are refor flat, layer-based annotations. In POWLA, this information can be expressed as a property of an annotation layer, i.e., Layer (informally, this is a set of Nodes and Relations): If all nodes from a Layer dominate only Terminals and node 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,[5] but not for querying or other purposes. The POWLA TBox allows us to infer this differentiation automatically from the data, so it does not have to be specified explicitly in the corpus.

A corpus can be represented as an **POWLA ABox** associated with the POWLA TBox, i.e., represented as a set of individuals that instantiate the concepts defined in the POWLA ontology.

Considering the phrase *viele Kulturschätze* 'many cultural treasures' from the German sentence analyzed in Fig. 1, Terminals, Nonterminals and Relations are created as shown in Fig. 2:

Terminals tok.51 and tok.52 are the tokens *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 have been 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 explicitly. A reasoner would also allow us to verify whether the axioms specified in the POWLA TBox are respected.

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 modeled in OWL.

3 Querying multi-layer corpora with POWLA

For this paper, we chose corpus querying as an example application to show how corpora represented in POWLA can be processed.

Due to space limitations, it is not possible to describe the approach in detail here. In brief, we conducted the following experiments:

- We implemented a set of SPARQL macros that emulate the PAULA-based ANNIS Query Language AQL [8]. We showed that every operator in AQL can be rendered in terms of SPARQL.

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

[5] StructLayers can be visualized as multi-rooted trees, MarkableLayers can be visualized as rows in a table, cf. [8].

- We converted the German NEGRA corpus [28] with the coreference annotations by [26] to POWLA. The corpus was loaded into the RDF data base OpenLink Virtuoso and could be queried with SPARQL.

Details on both experiments can be found unter `http://purl.org/powla`. The first experiment showed that POWLA represents the information of linguistic corpora in a *useful* way, i.e., a standard query language for multi-layer corpora could be successfully emulated. The second experiment showed that it is possible to *implement a query system* on the basis of RDF. The most important result, however, is, how little resources were necessary for this task: In total, it took us about 3 man-weeks.

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. Out 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.

Moreover, SPARQL actually provides us with even more powerful means of querying than ANNIS QL. An important restriction is that ANNIS QL does not support queries for the absence of a particular annotation (e.g., an NP not dominating a pronoun). In SPARQL, this can be easily expressed.

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, e.g., [4] or [15], the approach described here is not tied to one particular type of annotation, but rather, applicable to any kind of text-based linguistic annotation, because it takes its point of departure from an existing XML standoff format capable to represent any kind of linguistic annotation.

A pilot study was conducted for a German corpus with multi-layer annotations (syntax, coreference), it was shown how the original annotations can be converted to OWL, linearized in RDF, loaded into an RDF database and queried with SPARQL. We have shown how SPARQL macros applied to POWLA data can be employed to emulate the ANNIS query language AQL, a query language for multi-layer corpora. This shows that POWLA provides the information of the corpus in a similarly usable fashion as the underlying PAULA data model.

An important difference as compared to current standardization initiatives in the NLP community (e.g., ISO TC37/SC4) is that POWLA makes use of established standards maintained by their own community rather than to pursue the development of standards for linguistic data in particular. One advantage, in particular if compared to existing formats based on standoff XML [6, 13, 18, 25] is that POWLA can make use of an ecosystem of existing formalisms and technologies, including APIs, parsers and means for validation (OWL/DL reasoner). Like these formats, POWLA establishes structural interoperability between linguistic annotations produced by different tools or for different corpora, but it also makes it particularly easy to link annotated corpora with other linguistic resources, be it other corpora (e.g., the Penn Treebank [21] may be linked with its Czech translation [11]), lexico-semantic resources (e.g., resources from the Linked Open Data cloud) or reference repositories for annotation terminology (e.g., the General Ontology of Linguistic Description [14]) or metadata (e.g., Lexvo, a repository of language identifiers [12]).

Most existing data bases for multi-layer formats are based on relational data bases [32] or XML data bases [24] whose optimization for graph-based data structures is a particularly labor-intense task. As opposed to this, RDF data bases provide us with a query language and with a data model that is sufficiently general for linguistic annotations in general. The most important result to be reported here is that surprisingly few resources have been necessary to develop a data base solution that provides the functionality required for querying multi-layer corpora, which is an encouraging result for the prospective development of linguistic data bases on the basis of RDF.

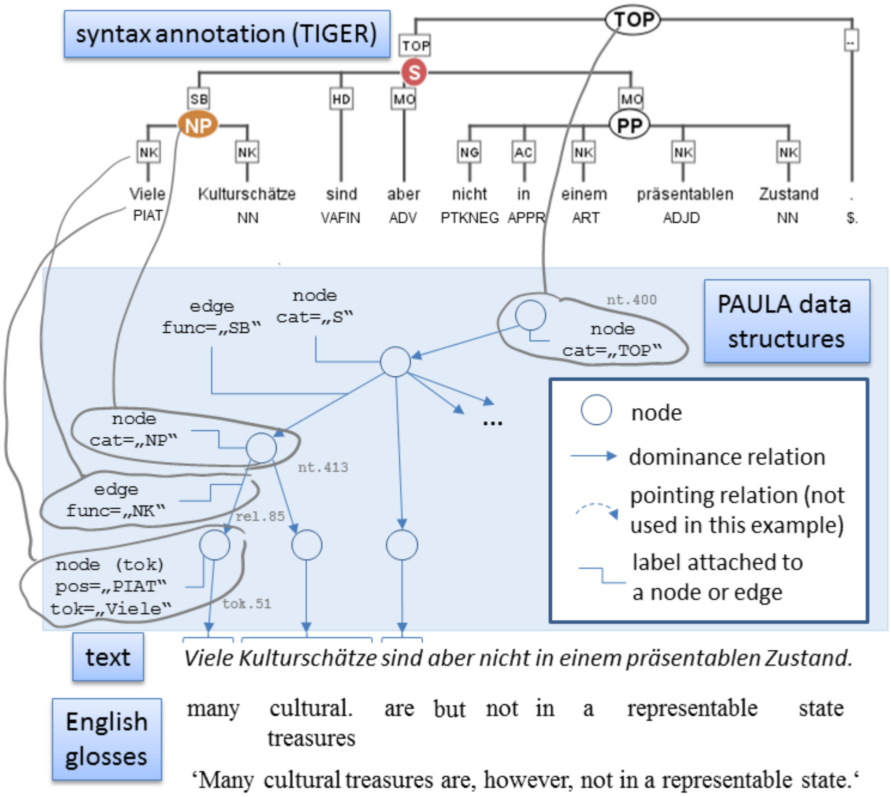


Figure 1: Using PAULA data structures for constituent syntax

References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2008.

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

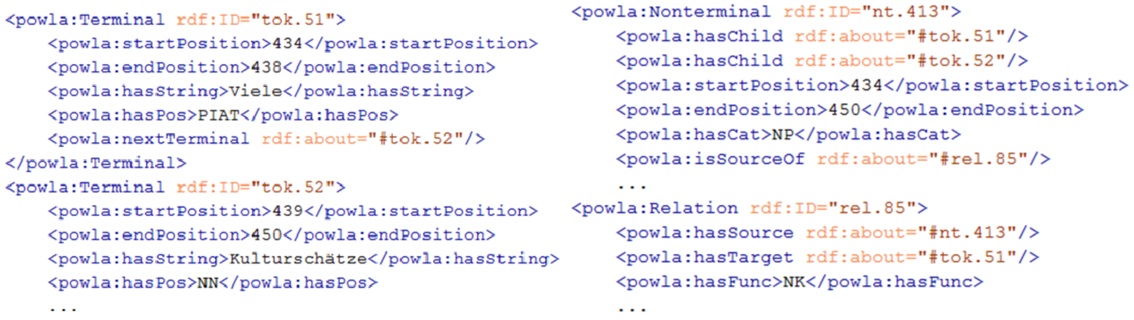
[3] P. Boersma. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345, 2002.

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

[6] J. Carletta, J. Kilgour, T. O'Donnell, S. Evert, and H. Voormann. The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets. In *Proceedings of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML)*, 2003.

[7] C. Chiarcos. Towards robust multi-tool tagging. An OWL/DL-based approach. In *ACL 2010*, pages 659–670, Uppsala, Sweden, July 2010.



```
<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>
    ...
```

```
<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>
    ...
```

Figure 2: Examples of Terminals, Nonterminals and Relations in POWLA

[8] C. Chiarcos, S. Dipper, M. Gtze, U. Leser, A. Ldeling, J. Ritz, and M. Stede. A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *Traitement Automatique des Langues*, 49(2), 2009.

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

[10] C. Chiarcos, J. Ritz, and M. Stede. Querying and visualizing coreference annotation in multi-layer corpora. In I. Hendrickx, A. Branco, S. L. Devi, and R. Mitkov, editors, *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)*, pages 80–92, Faro, Algarve, Portugal, Oct 2011. Edicoes Colibri.

[11] M. Čmejrek, J. Hajič, and V. Kuboň. Prague czech-english dependency treebank: Syntactically annotated resources for machine translation. In *In Proceedings of EAMT 10th Annual Conference*. Citeseer, 2004.

[12] G. de Melo and G. Weikum. Towards universal multilingual knowledge bases. In P. Bhattacharyya, C. Fellbaum, and P. Vossen, editors, *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the 5th Global WordNet Conference (GWC 2010)*, pages 149–156, New Delhi, India, 2010. Narosa Publishing.

[13] 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.

[14] S. Farrar and D. T. Langendoen. A Linguistic Ontology for the Semantic Web. *GLOT International*, 7:97–100, 2003.

[15] 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.

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

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

[18] 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.

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

[20] M. Lux, J. Laußmann, A. Mehler, and C. Menßen. An online platform for visualizing lexical networks. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, volume 1, pages 495–496. IEEE, 2001.

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

[22] C. Müller and M. Strube. Multi-level annotation of linguistic data with mmax2. In S. Braun, K. Kohn, and J. Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Peter Lang, Frankfurt, Germany, 2006.

[23] M. O'Donnell. Rsttool 2.4 – a markup tool for Rhetorical Structure Theory. In *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, pages 253–256, Mitzpe Ramon, Israel, 2000.

[24] G. Rehm, O. Schonefeld, A. Witt, E. Hinrichs, and M. Reis. Sustainability of annotated resources in linguistics: A web-platform for exploring, querying, and distributing linguistic corpora and other resources. *Literary and Linguistic Computing*, 2009.

[25] L. Romary, A. Zeldes, and F. Zipser. [tiger2/]-serialising the iso synaf syntactic object model. *Arxiv preprint arXiv:1108.0631*, 2011.

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

[27] T. Schmidt. Transcribing and annotating spoken language with E X M A Ra L D A. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*, Paris, 2004. ELRA.

[28] 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*, Saarbrcken, Germany, 1998.

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

[30] M. G. Stefanie Dipper. ANNIS: Complex Multilevel Annotations in a Linguistic Database. In *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*, Trento, Italy., 2006.

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

[32] 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.