Ontologies of Linguistic Annotation: Survey and Perspectives

Christian Chiarcos

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

Abstract

This paper announces the release of the Ontologies of Linguistic Annotation (OLiA). The OLiA ontologies represent a repository of annotation terminology for various linguistic phenomena on a great band-width of languages. This paper summarizes the results of five years of research, it describes recent developments and directions for further research.

Keywords: ontologies, linguistic annotation, conceptual interoperability

1. Background

The heterogeneity of linguistic annotations has been recognized as a key problem limiting the interoperability and reusability of NLP tools and linguistic data collections.

In Natural Language Processing, standard architectures such as UiMA (Egner et al., 2007) and GATE (Cunningham, 2002) address the interoperability of linguistic data structures by providing wrappers around existing NLP components that make use of formalisms that represent input and output of these modules in a tool-independent way. While this approach indeed yields interoperable data *structures*, and thereby establishes **structural interoperability**, it is limited insofar as the annotations itself, their content and their meaning, are not standardized in the same way. This problem, the establishment of **conceptual interoperability** between different linguistic annotations, is addressed here.

This problem has long been recognized and numerous initiatives have addressed the problem to represent linguistic annotations in an interoperable way. By now, it is generally agreed upon that **repositories of linguistic annotation terminology** represent a key element in the establishment of conceptual interoperability. With a terminological reference repository, it is possible to abstract from the heterogeneity of annotation schemes: Reference definitions provide an interlingua that allows to map linguistic annotations from annotation scheme A to annotations in accordance with scheme B. Several repositories of linguistic annotation terminology have been developed by the NLP/computational linguistics community (Aguado de Cea et al., 2004) as well as in the field of language documentation/typology (Saulwick et al., 2005), and their continuous application is expected to enhance the consistency of linguistic metadata and annotations. The General Ontology of Linguistic Description (Farrar and Langendoen, 2010, GOLD) and the ISO TC37/SC4 Data Category Registry (Kemps-Snijders et al., 2009, ISOcat) address both communities.

At the moment, however, two problems for the practical application of any of these terminology repositories persist:

- Different communities develop and maintain independent terminology repositories (e.g., GOLD and ISOcat), and these repositories are not always compatible with respect to the definitions they provide, with respect to the technologies employed, or with respect to the underlying philosophy. These problems are actively addressed by the GOLD and ISOcat communities, e.g., in the context of the RELISH project (Kemps-Snijders, 2010). The possible integration between GOLD and ISOcat is, however, expected to by a longer process.
- There is no commonly agreed formalism to link linguistic annotations to terminology repositories. For GOLD, concrete annotations are linked to reference concepts by means of hand-crafted mapping scripts (Simons et al., 2004). For ISOcat, RDF has been suggested as a means of addressing data categories only recently (Windhouwer and Wright, 2012).

The Ontologies of Linguistic Annotation (OLiA) have been developed to address both problems in order to facilitate the development of applications that take benefit of a well-defined terminological backbone even before the GOLD and ISOcat repositories have converged into a generally accepted reference terminology. The OLiA ontologies introduce an intermediate level of representation between ISOcat, GOLD and other repositories of linguistic reference terminology and are interconnected with these resources, and they provide not only means to formalize reference categories, but also annotation schemes, and the way that these are linked with reference categories.

2. The Ontologies of Linguistic Annotation

2.1. Linking Annotations with Reference Categories

The classic approach to link annotations with reference concepts is to specify rules that define a direct mapping (Teufel, 1995). It is, however, not always possible to find a 1:1 mapping.

One problem is **conceptual overlap**: A common noun may occur as a part of a proper name, e.g., German *Palais* 'baroque-style palace' in *Neues Palais* lit. 'new palace', a Prussian royal palace in Potsdam/Germany. *Palais* is thus both a proper noun (in its *function*), and a common noun (in its *form*). Such conceptual overlap is sometimes represented with a specialized tag, e.g., in the TIGER scheme (Brants et al., 2004). ISOcat does currently not provide the corresponding hybrid category, so that *Palais* is to be linked to both `properNoun`/DC-1371 and `commonNoun`/DC-1256

if the information carried by the original annotation is to be preserved. **Cliticization** and **fusion** are similar in that multiple word classes can be assigned to (different parts of) one token as represented, for example, by the English *gonna*, that can be annotated in the PennTreebank tagset as both `VBG` (gerund, *going*) and `TO` (*to*) (Santorini, 1990).

A somewhat different problem is the representation of **ambiguity**: The SUSANNE (Sampson, 1995) tag `ICSt` applies to English *after* both as a preposition and as a subordinating conjunction. The corresponding ISOcat category is thus *either* `preposition`/DC-1366 or `subordinatingConjunction`/DC-1393. Without additional disambiguation, `ICSt` is to be linked to both data categories.

Technically, such problems can be solved with a 1:n mapping between annotations and reference concepts. Yet, overlap/contraction and ambiguity differ in their underlying meaning: While overlapping/contracted categories are in the intersection ($\sqcap$) of reference categories, ambiguous categories are are in their join ($\sqcup$). This difference is relevant for subsequent processing, e.g., to decide whether disambiguation is necessary. A standard mapping approach, however, fails to distinguish $\sqcap$ or $\sqcup$.

Being based on a decidable fragment of first-order predicate logic, **OWL/DL** represents a formalism that supports the necessary operators and flexibility: With reference concepts and annotation concepts are formalized as OWL classes, the linking between them can be represented by `rdfs:subClassOf` ($\sqsubseteq$). OWL/DL provides operators such as `owl:intersectionOf` ($\sqcap$), `owl:unionOf` ($\sqcup$) and `owl:complementOf` ($\neg$), and it allows to define properties and restrictions on the respective concepts.

An OWL/DL-based formalization has the additional advantage that it can employ existing terminological repositories, e.g., GOLD (native OWL/DL) and ISOcat (with an OWL/DL conversion as described by (Chiarcos, 2010a)). GOLD and ISOcat are, however, under development. The efforts to maintain the linking between annotations and the terminological repository can be reduced if another ontology is introduced that mediates between terminological repositories and annotation schemes: If a major revision of the repository occurs, only the linking between the intermediate ontology and the repository is to be revised, but the linking with not every single tagset.

Moreover, this intermediate ontology allows linking annotations to multiple terminological repositories at the same time. The OLiA ontologies implement the idea of an architecture of modular OWL/DL ontologies with an ontology mediating between terminological repositories and annotation schemes.

2.2. A Modular Architecture of OWL/DL Ontologies

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

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

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

- The OLiA REFERENCE MODEL specifies the common terminology that different annotation schemes can refer to. It is derived from existing repositories of annotation terminology and extended in accordance with the annotation schemes that it was applied to.
- Multiple OLiA ANNOTATION MODELs formalize annotation schemes and tagsets. Annotation Models are based on the original documentation, so that they provide an interpretation-independent representation of the annotation scheme.
- For every Annotation Model, a LINKING MODEL defines $\sqsubseteq$ relationships between concepts/properties in the respective Annotation Model and the Reference Model. Linking Models are interpretations of Annotation Model concepts and properties in terms of the Reference Model.
- Existing terminology repositories can be integrated as EXTERNAL REFERENCE MODELs, if they are represented in OWL/DL. Then, Linking Models specify $\sqsubseteq$ relationships between Reference Model concepts and External Reference Model concepts.

The OLiA Reference Model specifies classes for linguistic categories (e.g., `olia:Determiner`) and grammatical features (e.g., `olia:Accusative`), as well as properties that define relations between these (e.g., `olia:hasCase`). Far from being yet another annotation terminology ontology, the OLiA Reference Model does not introduce its own view on the linguistic world, but rather, it is a derivative of EAGLES (Leech and Wilson, 1996), MULTEXT/East (Erjavec, 2004), and GOLD (Farrar and Langendoen, 2010) that was introduced as a technical means to interpret linguistic annotations with respect to these terminological repositories, and further enriched with information drawn from the annotation schemes it was applied to.

Conceptually, Annotation Models differ from the Reference Model in that they include not only concepts and properties, but also individuals: Individuals represent concrete tags, while classes represent abstract concepts similar to those of the Reference Model.

2.3. Current Status

The OLiA ontologies are available from `http://purl.org/olia` under a Creative Commons Attribution license (CC-BY).

The OLiA ontologies cover different grammatical phenomena, including inflectional morphology, word classes, phrase and edge labels of different syntax annotations, as well as prototypes for discourse annotations (coreference, discourse relations, discourse structure and information structure). Annotations for lexical semantics are only

covered to the extent that they are encoded in syntactic and morphosyntactic annotation schemes. For lexical semantic annotations in general, a number of reference resources is already available, including RDF versions of WordNet and FrameNet.

In recent years, the OLiA ontologies have been substantially extended. At the time of writing, the OLiA Reference Model distinguishes 14 `MorphologicalCategory`s (morphemes), 263 `MorphosyntacticCategory`s (word classes), 83 `SyntacticCategory`s (phrase labels), and 326 different values for 16 `MorphosyntacticFeature`s, 4 `MorphologicalFeature`s, 4 `SyntacticFeature`s and 4 `SemanticFeature`s (for glosses, part-of-speech annotation and for edge labels in syntax annotation).

As for morphological, morphosyntactic and syntactic annotations, the OLiA ontologies include 32 Annotation Models for about 70 different languages, including several multi-lingual annotation schemes, e.g., EAGLES (Chiarcos, 2008) for 11 Western European languages, and MULTEXT/East (Chiarcos and Erjavec, 2011) for 15 (mostly) Eastern European languages. As for non-(Indo-)European languages, the OLiA ontologies include morphosyntactic annotation schemes for languages of the Indian subcontinent, for Arabic, Basque, Chinese, Estonian, Finnish, Hausa, Hungarian and Turkish. Other languages, including languages of Africa, the Americas, the Pacific and Australia are covered by Annotation Models developed for typology and language documentation. The OLiA ontologies also cover historical language stages, including Old High German, Old Norse and Old/Classical Tibetan. Additionally, 7 Annotation Models for different resources with discourse annotations have been developed.

External reference models currently linked to the OLiA Reference Model include GOLD (Chiarcos, 2008), the OntoTag ontologies (Buyko et al., 2008), and ISOcat (Chiarcos, 2010a). Thereby, the OLiA Reference Model provides a stable intermediate representation between existing terminology repositories and ontological models of annotation schemes. This allows any concept that can be expressed in terms of the OLiA Reference Model also to be interpreted in the context of ISOcat or GOLD. Using the OLiA Reference Model, it is thus possible to develop applications that are interoperable in terms of GOLD *and* ISOcat even though both are still under development and both differ in their conceptualizations. Such applications are briefly described in the following section.

3. Fields of Application

3.1. Corpus Linguistics

Initially, the OLiA ontologies have been intended to serve a **documentation function**, i.e., as a formal means to specify the semantics of annotation schemes (Schmidt et al., 2006). From the ontologies, dynamic HTML can be generated,[1] and tags in the annotation can be represented as hyperlinks pointing to the corresponding definition (Chiarcos et al., 2008). Figure 1 shows a screenshot of the HTML version of

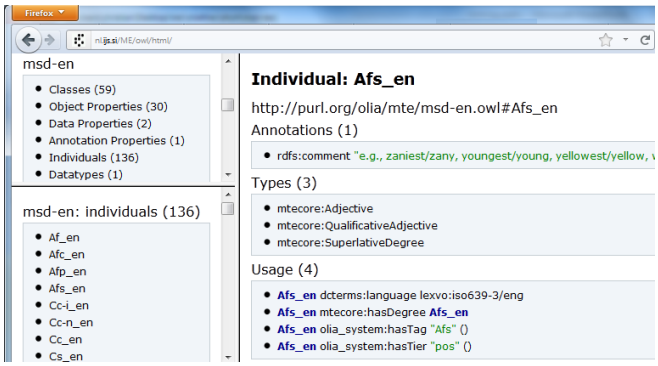

Figure 1: HTML version of the OLiA Annotation Model for the MULTEXT/East morphosyntactic specifications for English, `http://purl.org/olia/mte`.

the OLiA Annotation Models of the MULTEXT/East morphosyntactic specifications (Chiarcos and Erjavec, 2011).

OLiA has been integrated in **corpus query systems**, e.g., ANNIS (Chiarcos and Götze, 2007), and SPLICR (Rehm et al., 2007), so that corpus queries could be formulated on the basis of Reference Model concepts. It is thus possible to explore corpora with unfamiliar tagsets, e.g., to reduce the initial bias to evaluate their appropriateness for a given problem. Moreover, ontology-based corpus queries allow to abstract from individual annotation schemes and make it possible to query across heterogeneously annotated corpora. Technically, this was implemented as a query preprocessor: Instead of querying for `cat="NX"` to retrieve noun phrases from the TüBa-D/Z corpus (Telljohann et al., 2003) or `cat="NP"` on the NEGRA corpus (Skut et al., 1998, both are corpora of German newspaper text), we can just query for `cat in {olia:NounPhrase}` and this is then expanded into a disjunction of possible tags (Chiarcos and Götze, 2007). Due to the possibility to create extremely large disjunctions (by just querying for top-level concepts), this approach was relatively inefficient and hence not incorporated in the release versions of the systems mentioned above.

One alternative would be linking OLiA Annotation Models and annotations in a corpus directly. For reasons of interoperability, this approach should respect existing standards to implement such a linking. Despite on-going efforts of the linguistics and NLP communities to develop such formalisms (Nancy Ide, p.c.), I am not aware of any existing recommendation to interlink corpora in traditional NLP formats directly with terminology repository. However, such a possibility is provided by Semantic Web formalisms, namely RDF/OWL. In recent years, several researchers have developed schemes to convert specific corpora (Burchardt et al., 2008), specific types of annotation (Hellmann, 2010; Rubiera et al., accepted), or generic corpus representation formalisms (Cassidy, 2010; Chiarcos, this vol) to RDF/OWL. With both corpora and terminology repositories represented in RDF, linking them is reduced to an application of the Linked Data Paradigm (Berners-Lee, 2006), i.e., the main function of RDF in the Semantic Web context, and hence, well-supported by RDF-based technologies.

[1] `http://code.google.com/p/co-ode-owl-plugins/wiki/OWLDoc`

As a first example application, Hellmann et al. (2010) developed the TIGER Corpus Navigator, a tool to explore the RDF representation of the TIGER corpus (Brants et al., 2004). Given a user's input, a query is automatically generated and run against the repository. Query generation was performed using ontology-based learning techniques (Lehmann et al., 2011); to represent linguistic annotations, the OLiA ontologies were employed.

The TIGER Corpus Navigator was an experimental pilot study focusing on one particular corpus. More recently, OLiA was combined with POWLA (Chiarcos, 2012), a generic formalism to represent corpora in RDF. So far, two corpora have been converted to POWLA, the MASC corpus (Ide et al., 2008), a genre-balanced multi-layer corpus for American English, and the NEGRA corpus (Skut et al., 1998), a German newspaper corpus,[2] and their morphosyntactic and syntactic annotations have been converted to links to the corresponding OLiA Annotation Models. It is thus possible to query these corpora and the OLiA ontologies using the RDF query language SPARQL.

3.2. Interoperability

Linking annotations to terminology repositories is also essential in terms of interoperability of NLP resources. Interoperability involves two core aspects (Ide and Pustejovsky, 2010), structural ('syntactic') interoperability (different resources make use of the same representation formalism), and conceptual ('semantic') interoperability (resources make use of a shared, well-defined vocabulary).

OLiA can be used to establish conceptual interoperability between linguistic corpora, in that the same queries can be applied to corpora with different annotation schemes. In a similar vein, Buyko et al. (2008) suggested to employ OLiA in UiMA (Ferrucci and Lally, 2004), a pipeline architecture for NLP. However, similar problems with the integration between different formalisms persist as observed above for corpus queries, e.g., directly integrating ontologies led to a drop in performance because optimizations based on the use of annotation type system cannot be applied any more (but cf. Verspoor et al., 2009).

More recently, it has been suggested to develop NLP pipeline systems using Semantic Web technologies, with the objective to integrate the output of NLP tools in ontology-based machine learning algorithms (Hellmann, 2010). Along with other ontologies, that specify units of analysis (NLP Interchange Format, NIF),[3] the OLiA ontologies represent one fundamental part of this framework: Using RDF as a representation format, and RDF-based wrappers around existing NLP components, linguistic annotations are formalized as direct links to the corresponding OLiA Annotation Models. Subsequent processing modules can make use of this information, and also chose the appropriate level of abstraction by working either directly with the Annotation Models (that precisely reflect the information found in the original annotation and its documentation), with the OLiA Reference Model (that involves an interpretation of the original descriptions as defined in the

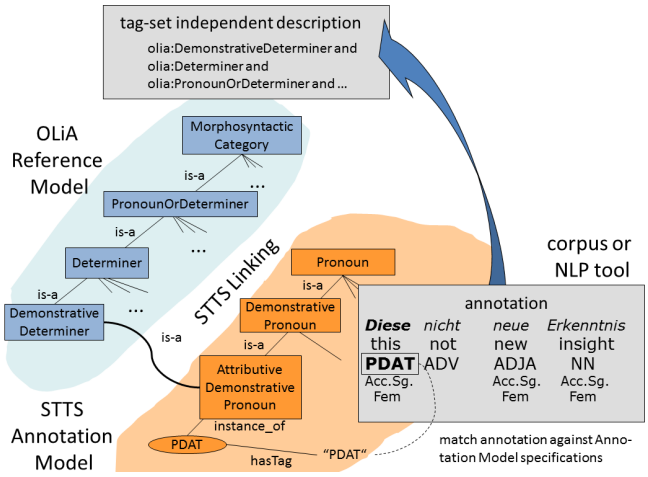

Figure 2: Interpreting annotations in terms of the OLiA Reference Model

Linking Model), or with one or several External Reference Models (that involve another step in interpretation).

Figure 2 illustrates how annotations can be mapped onto Reference Model concepts for the German phrase *Diese nicht neue Erkenntnis* 'this well-known (lit. not new) insight' from the Potsdam Commentary Corpus (Stede, 2004, file 4794), with part-of-speech annotations according to the STTS scheme (Schiller et al., 1999): The tag `PDAT` matches the surface string of the individual `stts:PDAT` from the STTS Annotation Model.[4] The superconcept `stts:AttributiveDemonstrativePronoun` is a subconcept of `olia:DemonstrativeDeterminer` (STTS Linking Model).[5] The word *diese* 'this' from the example can thus be described in terms of the OLiA Reference Model as `olia:DemonstrativeDeterminer`, etc.

These ontology-based descriptions are comparable across different corpora and/or NLP tools, across different languages, and even across different types of language resources: McCrae et al. (2011) describe the application of the OLiA ontologies to represent grammatical specifications of machine-readable dictionaries, that are thus interoperable with OLiA-linked corpora. Moreover, through the linking with External Reference Models like GOLD and ISOcat, OLiA-linked resources are also conceptually interoperable with resources directly grounded in either GOLD or ISOcat.

3.3. Ontology-Based NLP

Using Semantic Web formalisms to represent corpora and annotations also provides us with the possibility to develop novel NLP algorithms. RDF/OWL as common representation formalism allows, for example, to integrate analyses of different levels of description, say, syntax and semantics, in order to disambiguate each other (Cimiano and Reyle, 2003).

But even within traditional fields of NLP, possible applications can be found, e.g., in ensemble combination architectures: Ensemble combination means that different NLP

[2] For details and links, see `http://purl.org/powla`.
[3] `http://nlp2rdf.org/nif-1-0`
[4] `http://purl.org/olia/stts.owl`
[5] `http://purl.org/olia/stts-link.rdf`

modules (say, part-of-speech taggers) are applied in parallel, that they produce annotations for one particular phenomenon, and that these annotations are then integrated, e.g., by choosing one possible analysis based on the evaluation of the agreement between the different modules. If modules are combined that use different approaches, e.g., different machine learning paradigms, it has been frequently observed that the combination of tools yields an increase in accuracy and robustness (Brill and Wu, 1998; Halteren et al., 2001).

So far, however, these approaches were limited to combine tools trained on the *same* tagset. If tools with different tagsets are combined, we should not only expect an increase in accuracy and robustness, but also an increase in the *level of detail*: If the ensemble combination has decided to adopt one particular analysis, then, information from another module with more detailed analyses can be merged with this candidate if it is compatible with the favored analysis. Chiarcos (2010b) described such an architecture and tested it with 7 tools for the morphosyntactic analysis of German that used 4 different annotation schemes. Figure 2 shows how the original morphosyntactic annotations were translated into ontological descriptions for the word *diese*, and Fig. 3 for the corresponding annotations created by the Connexor dependency parser (Tapanainen and Järvinen, 1997) and the RFTagger (Schmid and Laws, 2008). Both analyses in Fig. 3 vary with respect to case and with respect to the part-of-speech (`olia:Pronoun` or `olia:Determiner`). These conjunctions are transformed into multisets of conjuncts, and conjunct frequencies establish a simple confidence ranking among these. To this ranking, a pruning routine was applied that filtered out every conjunct that was *inconsistent* with a higher-ranked conjunct. Consistency can be defined within the ontology. The resulting set of conjuncts was then compared against manual annotations. For evaluation, data from three German newspaper corpora was considered, the NEGRA corpus (Skut et al., 1998), the TIGER corpus (Brants et al., 2004) and the Potsdam Commentary Corpus (Stede, 2004): It could be shown that recall[6] monotonically increased with the number of tools combined. Combining part-of-speech annotations from all (7) tools outperformed the best-performing individual tool.[7]

This experiment serves as a proof-of-concept implementation for ontology-based ensemble combination approaches; with more elaborate voting techniques, more significant improvements can be expected. Comparable results have been reported for the analysis of an ambiguous particle in Spanish (Pareja-Lora and Aguado de Cea, 2010). Taken together, both studies support our observation that the ontology-based integration of morphosyntactic analyses enhances both the robustness and the level of detail of morphosyntactic and morphological analyses.

[6] An evaluation in terms of precision would be inappropriate, because some NLP annotations are more fine-grained than the original manual annotation.

[7] The only exception was RFTagger on the NEGRA corpus, albeit only due to the fact that it was *trained* on NEGRA.

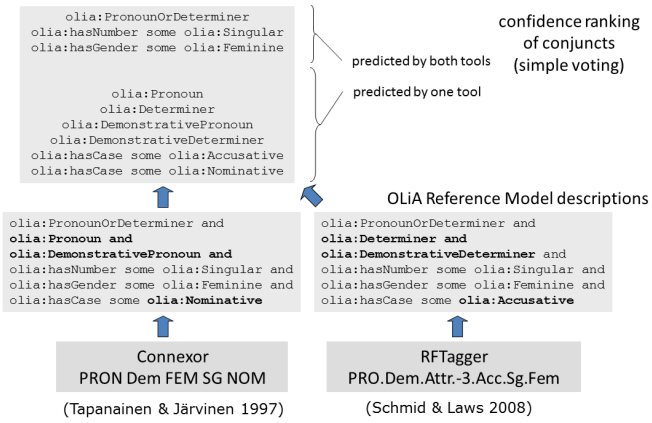

Figure 3: Ontology-based ensemble combination

4. Discussion and Outlook

This paper summarized the development of the OLiA ontologies since 2006, their current status, and a number of applications that have been developed on this basis.

The fundamental idea of the OLiA architecture is that annotation schemes are linked to community-maintained terminology repositories through an intermediate 'Reference Model', thereby minimizing the number of mappings necessary establish interoperability of one annotation scheme with multiple terminology repositories. Further, annotation schemes and their linking to the Reference Model are formalized as separate OWL/DL ontologies, so that interpretation-independent conceptualization (annotation documentation) and its interpretation in terms of the Reference Model (linking) are properly distinguished.

The OLiA ontologies differ from related approaches in that they take a focus on modeling annotation schemes and their linking with reference categories rather than merely providing reference categories. The differentiation of Annotation Models, the OLiA Reference Model and External Reference Models (community-maintained terminology repositories) represents increasing levels of abstraction, and, possibly, loss of information. However, no information about the original annotation is lost, and tools may chose the appropriate level of abstraction. Unlike a direct mapping approach, OLiA allows to recover information about sources of mismatches between Reference Model concepts and Annotation Model concepts, because a declarative linking is provided that allows inspection and refinement using standard RDF/OWL tools.

The relationship between annotations and reference concept is not only represented in a transparent way, but also, conceptual *mis*matches can be represented. Many tagsets for part-of-speech annotation, for example, introduce hybrid categories to represent either conceptual overlap/fusion or ambiguity using OWL/DL constructs to represent conjunction ($\sqcap$) or disjunction ($\sqcup$). As compared to tagset-specific solutions (Santorini, 1990, | for ambiguities, and + for cliticization/fusion), OWL/DL provides a W3C-standardized vocabulary to express these relationships, that also extends beyond individual tagsets. Another difference is that negation (`owl:complementOf`) is available in the linking. This is of particular impor-

tance for the linking between External Reference Models and the OLiA Reference Model. For example, an `olia:ProQuantifier` (pronominal quantifier, can substitute for an independent noun phrase, e.g., *someone*) can be defined as subclass of `gold:Quantifier`. According to its definition, however, `gold:Quantifier` primarily pertains to determiners, so that a more appropriate superclass would be `gold:Quantifier` $\sqcap \neg$`gold:Determiner`.

The physical separation of Linking Models from Annotation Models and Reference Model introduces a clear distinction between externally provided information and the ontology engineer's interpretation. Annotation Models formalize annotation documentation, and the Reference Model is based on a generalization of a broad band-width of resources. However, there may be different terminological traditions involved, so that apparently similar concepts found in Reference Model and Annotation Model are in fact unrelated. If nevertheless an incorrect identification takes place, the linking can be inspected by standard ontology browsers, and corrected independently from the interpretation-invariant Annotation Model and Reference Model. Furthermore, *multiple* linkings between an Annotation Model and the Reference Model can be implemented, e.g., to accommodate for systematic tagger errors (i.e., more extensive usage of `owl:join`), or for multiple dialects of the same tagset (e.g., the STTS tagset distinguishes indefinite attributive pronouns in indefinite noun phrases [`PIAT`] and in definite noun phrases [`PIDAT`], but in the TüBa-D/Z corpus (Telljohann et al., 2003), `PIAT` covers both uses).

In ISOcat, the problem of conflicting interpretations of data categories is currently *not* addressed. There are definitions provided, but they may not be sufficient to distinguish different classes, e.g., the category `definite`/DC-2004 is defined as 'value referring to the capacity of identification of an entity'. The concept is (at least partially) grounded in the MULTEXT/East morphosyntactic specifications (Francopoulo et al., 2008), but there, different uses of 'definite' were conflated: (1) postfixed determiner in Romanian, Bulgarian and Persian nouns or adjectives, (2) difference between 'full' and 'reduced' adjectives in Slavic languages (diachronically, full forms reflect cliticization with pronominal elements), (3) a specific pattern of quantifier agreement in Slavic, and (4) the so-called 'definite conjunction' of Hungarian verbs (indicating the presence of a definite object argument). These definitions are (mostly) compatible with the generic definition, but they are not necessarily compatible *with each other*; linking such different conceptions to the same reference category provides little improvement in terms of conceptual interoperability. However, without modeling relations between different language-specific annotation schemes and a data category registry from a global perspective, it is possible that such ill-defined data categories and/or links remain *undetected*. Within MULTEXT/East, for example, only the ontological modeling of language-specific annotation schemes and the common morphosyntactic specifications led to the proper differentiation between these different conceptions of 'definite' (Chiarcos and Erjavec, 2011).

The OLiA Reference Model provides such a fully developed taxonomy of linguistic categories rather than a semi-structured set, whereas, ISOcat lacks a comparable global, or top-down perspective at the moment. Recent activities to develop a relation category registry (Schuurman and Windhouwer, 2011) on top of ISOcat may eventually provide such a perspective and the problem of conflicting interpretations of data categories may become more obvious to ISOcat developers, but still, no standard way to represent interpretations of annotation concepts in terms of data categories has been established.

In comparison to GOLD, OLiA is more focused on NLP and corpus interoperability, whereas GOLD originates from the language documentation community. Therefore, a number of data categories commonly assumed in NLP were not originally represented in GOLD. For example, `gold:CommonNoun` was added only recently (between 2006 and 2008), following a suggestion by the author. While the GOLD community process will eventually lead to a compensation of such coverage issues, a more fundamental problem is that the views of academic linguists and NLP engineers may deviate with respect to the overarching taxonomy of concepts. GOLD, for example, seems to conflate both semantic roles ('case' in the sense of (Fillmore, 1968), e.g., `gold:BenefactiveCase`) and syntactic roles under `gold:CaseProperty`. Therefore, OLiA adopts a relatively agnostic view on the taxonomical order of concepts. While the taxonomy is modeled in a specific way (mostly following established annotation schemes), it is not assumed that this way of modeling is the only possibility. In fact, alternative taxonomies can be formulated as External Reference Models, and OWL/DL-based allows to formulate specific conditions for the linking, including the use of negation and disjunction. Consequently, mismatches can be represented. (As opposed to this, GOLD Community of Practice Extensions are assumed to adopt the GOLD hierarchy and only to extend it, not to redefine it.)

Conceptually, the OLiA ontologies are closer related to the OntoTag ontologies (Aguado de Cea et al., 2004), that were also applied to develop NLP applications on the basis of ontological representations of linguistic annotations (Pareja-Lora and Aguado de Cea, 2010). One important difference is that the OntoTag ontologies are considering only the languages of the Iberian peninsula (in particular Spanish), that they are partially designed with a top-down perspective (whereas the development of the OLiA Reference Model is guided by the annotation schemes it is applied to) and are thus richer in consistency constraints (that are, however, often language-specific), and that the OntoTag ontologies are not publicly available at the moment. Within the OLiA architecture, the morphosyntactic layer of the OntoTag ontologies is integrated as an External Reference Model (Buyko et al., 2008).

It should be noted, however, that OLiA is not intended to serve as an alternative to either GOLD, ISOcat or the OntoTag ontologies. In fact, it is linked to (and, partially, derived from) all of them and helps establishing interoperability between GOLD-linked resources, ISOcat-linked resources, etc. In parts, integration efforts between these resources have already begun, as manifested, for example, in the RELISH project (Kemps-Snijders, 2010), but until

concrete results in this direction have been achieved, the OLiA ontologies already bridge between these repositories and concrete annotations with a level of detail and information that renders them useful for the development of various applications. In fact, the linking between the External Terminology Repositories that was developed as part of the OLiA ontologies may represent a seed for links between GOLD and ISOcat as currently envisioned in the relation category registry (Schuurman and Windhouwer, 2011).
The OLiA ontologies may play an important role in NLP, corpus and annotation interoperability in that they relate these activities to initiatives in different linguistic communities to establish reference repositories for linguistic annotation terminology, e.g., recent developments towards the creation of a Linguistic Linked Open Data cloud (Chiarcos et al., this vol).

Acknowledgements

The OLiA ontologies have been developed at the Collaborative Research Center (SFB) 441 "Linguistic Data Structures" (University of Tübingen) in the context of the project "Sustainability of Linguistic Resources" in cooperation with SFB 632 "Information Structure" (University of Potsdam, Humboldt-University Berlin) and SFB 538 "Multilingualism" (University of Hamburg) from 2006 to 2008. From 2007 to 2011, they have been maintained and further developed in the context of SFB 632 in the context of the project "Linguistic Data Base".

5. References

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

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

S. Brants, S. Dipper, P. Eisenberg, et al. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

E. Brill and J. Wu. 1998. Classifier combination for improved lexical disambiguation. In *Proc. COLING-ACL 1998*, pages 191–195, Montréal, Canada, August.

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.

E. Buyko, C. Chiarcos, and A. Pareja-Lora. 2008. Ontology-based interface specifications for a NLP pipeline architecture. In *Proc. LREC 2008*, Marrakech, Morocco.

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 T. Erjavec. 2011. OWL/DL formalization of the MULTEXT-East morphosyntactic specifications. In *Proc. 5th Linguistic Annotation Workshop, held in conjunction with ACL-HTL 2011*, pages 11–20, Portland, June.

C. Chiarcos and M. Götze. 2007. A linguistic database with ontology-sensitive corpus querying. system demonstration at Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Frühjahrstagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV 2007). Tübingen, Germany, April 2007.

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

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

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

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

C. Chiarcos. 2012. 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. Springer.

C. Chiarcos. this vol. A generic formalism to represent linguistic corpora in RDF and OWL/DL.

P. Cimiano and U. Reyle. 2003. Ontology-based semantic construction, underspecification and disambiguation. In *Proc. Lorraine/Saarland Workshop on Prospects and Recent Advances in the Syntax-Semantics Interface*, pages 33–38, Nancy, France.

H. Cunningham. 2002. Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254.

M.T. Egner, M. Lorch, and E. Biddle. 2007. UIMA Grid: Distributed large-scale text analysis. In *Proc. 7th IEEE International Symposium on Cluster Computing and the Grid (CCGRID'07)*, pages 317–326, Rio de Janeiro, Brazil, May.

T. Erjavec. 2004. MULTEXT-East version 3. In *Proc. LREC 2004*, pages 1535–1538, Lisboa, Portugal.

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.

D. Ferrucci and A. Lally. 2004. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3/4):327–348.

C. J. Fillmore. 1968. The case for case. In E. Bach and R.T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart, and Winston, New York.

G. Francopoulo, T. Declerck, V. Sornlertlamvanich, et al.

2008. Data Category Registry: Morpho-syntactic and syntactic profiles. In *Proc. LREC-2008 Workshop on Uses and Usage of Language Resource-Related Standards*, Marrakech, Morocco, May.

H. Halteren, J. Zavrel, and W. Daelemans. 2001. Improving accuracy in word class tagging through the combination of machine learning systems. *Computational Linguistics*, 27(2):199–229.

S. Hellmann, J. Unbehauen, C. Chiarcos, and A. Ngonga Ngomo. 2010. The TIGER Corpus Navigator. In *Proc. 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.

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, 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.

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

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

G. Leech and A. Wilson. 1996. EAGLES recommendations for the morphosyntactic annotation of corpora. Version of March 1996.

J. Lehmann, S. Auer, S. Tramp, et al. 2011. Class expression learning for ontology engineering. *Journal of Web Semantics*, 9(1):71–81.

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

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

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

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

G. Sampson. 1995. *English for the computer: The SUSANNE corpus and analytic scheme*. Oxford University Press.

B. Santorini, 1990. *Part-of-speech tagging guidelines for the Penn Treebank Project*. Department of Computer and Information Science, University of Pennsylvania. Technical report MS-CIS-90-47.

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

A. Schiller, S. Teufel, C. Stöckert, and C. Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universities of Stuttgart and Tübingen.

H. Schmid and F. Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proc. COLING 2008*, Manchester, UK.

T. Schmidt, C. Chiarcos, T. Lehmberg, et al. 2006. Avoiding data graveyards. In *Proc. E-MELD Workshop 2006*, Ypsilanti.

I. Schuurman and M.A. Windhouwer. 2011. Explicit semantics for enriched documents. What do ISOcat, RELcat and SCHEMAcat have to offer? In *Proc. 2nd Supporting Digital Humanities conference (SDH 2011)*, Copenhagen, Denmark, November.

G.F. Simons, W.D. Lewis, S.O. Farrar, et al. 2004. The semantics of markup. In *Proc. 4th Workshop on NLP and XML (NLPXML-2004)*, pages 25–32, Barcelona, Spain, July.

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.

Manfred Stede. 2004. The Potsdam Commentary Corpus. In *Proc. ACL-2004 Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain, July.

P. Tapanainen and T. Järvinen. 1997. A nonprojective dependency parser. In *Proc. 5th Conference on Applied NLP*, pages 64–71, Washington, DC.

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. Teufel. 1995. A support tool for tagset mapping. In *Proc. EACL-1995 SIGDAT Workshop: From Text to Tags*.

K. Verspoor, W. Baumgartner, C. Roeder, and L. Hunter. 2009. Abstracting the types away from a UIMA type system. In C. Chiarcos, R. Eckhart de Castilho, and M. Stede, editors, *From Form to Meaning: Processing Texts Automatically*, pages 249–256. Gunter Narr.

M. Windhouwer and S. E. Wright. 2012. Linking to linguistic data categories in ISOcat. In C. Chiarcos, S. Nordhoff, and S. Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 99–107. Springer, Heidelberg.