OLiA – Ontologies of Linguistic Annotation

Editor(s): Sebastian Hellmann, AKSW, University of Leipzig, Germany; Steven Moran, LMU Munich, Germany; Martin Brümmer, AKSW, University of Leipzig, Germany; John P. McCrae, CITEC, Bielefeld University, Germany
Solicited review(s): Menzo Windhouwer, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; Steve Cassidy, Macquarie University, Sydney, Australia; one anonymous reviewer

Christian Chiarcos [*] and Maria Sukhareva

*Applied Computational Linguistics (ACoLi), Department of Computer Science and Mathematics,
Goethe-University Frankfurt am Main, Germany, http://acoli.cs.uni-frankfurt.de*

Abstract. This paper describes the Ontologies of Linguistic Annotation (OLiA) as one of the data sets currently available as part of Linguistic Linked Open Data (LLOD) cloud. Within the LLOD cloud, the OLiA ontologies serve as a reference hub for annotation terminology for linguistic phenomena on a great band-width of languages, they have been used to facilitate interoperability and information integration of linguistic annotations in corpora, NLP pipelines, and lexical-semantic resources and mediate their linking with multiple community-maintained terminology repositories.

Keywords: linguistic terminology, grammatical categories, annotation schemes, corpora, NLP, tag sets, 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. Several repositories of linguistic annotation terminology have been developed to facilitate annotation interoperability by means of a joint level of representation, or an ‘interlingua’, the most prominent probably being the General Ontology of Linguistic Description [12, GOLD] and the ISO TC37/SC4 Data Category Registry [21, ISOcat].

Still, these repositories are developed by different communities, and are thus not always compatible with each other, neither with respect to definitions nor technologies (e.g., there is no commonly agreed formalism to link linguistic annotations to terminology repositories).

The Ontologies of Linguistic Annotation (OLiA) have been developed 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 repository of reference terminology: They 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. Architecture

The **Ontologies of Linguistic Annotations** [3] represent a modular architecture of OWL2/DL ontologies that formalize 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 [28], and their primary fields of application include the formalization of annotation schemes and concept-based querying over heterogeneously annotated corpora [16,10].

In the OLiA architecture, four different types of ontologies are distinguished (cf. Fig. 1 for an example):

[*] Corresponding author. E-mail: chiarcos@informatik.uni-frankfurt.de

- 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 OWL2/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`).

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. Figure 1 gives an example for the individual `PDAT` from the STTS Annotation Model, the corresponding STTS concepts, and their linking with Reference Model concepts. Taken together, these allow to interpret the individual (and the part-of-speech tag it represents) as an `olia:Determiner`, etc.

3. Data Set Description

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 found in syntactic and morphosyntactic annotation schemes. Other aspects of lexical semantics are beyond the scope of OLiA: Existing reference resources for lexical semantics available in RDF include WordNet, VerbNet and FrameNet, their linking with OLiA is recommended as part of the lexicon model *lemon* [22], and has been implemented, for example, in *lemonUby* [11].

Since their first presentation [3], the OLiA ontologies have been continuously extended. At the time of writing, the OLiA Reference Model distinguishes 280 `MorphosyntacticCategory`s (word classes), 68 `SyntacticCategory`s (phrase labels), 18 `MorphologicalCategory`s (morphemes), 7 `MorphologicalProcess`es, and 405 different values for 18 `MorphosyntacticFeature`s, 5 `SyntacticFeature`s and 6 `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 [3] for 11 Western European languages, and MULTEXT/East [8] 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, as well as multi-lingual schemes applied to languages of Africa, the Americas, the Pacific and Australia. The OLiA ontologies also cover historical varieties, including Old High German, Old Norse and Old 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 [3], the OntoTag ontologies [1], an ontological remodeling of ISOcat [4], and the Typological Database System (TDS) ontologies [26]. From these, only the TDS ontologies are currently available under an open (CC-BY) license,[1] but these take a focus on typological data bases rather than NLP and annotation interoperability.

[1] `http://languagelink.let.uu.nl/tds/ontology/LinguisticOntology.owl`

GOLD and ISOcat *should* be available under an open license,[2] but can currently not be redistributed because of an uncertain licensing situation (no explicit license information). The OntoTag ontologies are available to the author but have not been publicly released.

In this context, the function of the OLiA Reference Model is not to provide a novel and independent view on linguistic terminology, but rather to serve as a stable intermediate representation between (ontological models of) annotation schemes and these terminology repositories. This allows any concept that can be expressed in terms of the OLiA Reference Model also to be interpreted in the context of ISOcat, GOLD, OntoTag or TDS. OLiA serves to aggregate annotation terminology as found in linguistic resources and provides a middle ground between these and the External Reference Models linked to it. We would like to emphasize that OLiA is not meant as a substitute for any of these repositories, but rather, that it serves to facilitate their further harmonization and interoperability, as they are maintained by different communities and remain for the foreseeable future in a continuous state of enrichment and specialization. Initial efforts towards their gradual convergence include the support of linking mechanisms to external knowledge bases in GOLD and ISOcat. Within a GOLD context, for example, OLiA may be referred to as a Community-of-Practice Extension for the NLP community. From the perspective of ISOcat, it may be seen as an ontological view on annotation terminology among the otherwise unstructured data categories. Along with ontologies for other ISOcat profiles, e.g., metadata [35], OLiA may provide a seed for populating RELcat [29], an ongoing effort to provide structured views on ISOcat data.

As compared to a direct linking between annotation models and these terminology repositories, the modular structure limits the number of linkings that need to be defined (if a new Annotation Model is linked to the Reference Model, it inherits its linking with ISOcat, GOLD, OntoTag and TDS), and also, it provides stability (GOLD and ISOcat are developed in community processes with occasional revisions), a clear and non-redundant taxonomical organization (similar to GOLD, TDS and OntoTag, but very different from the semi-structured ISOcat) and establishes interoperability between GOLD and ISOcat (that – despite ongoing harmonization efforts [20] – are maintained by different communities and developed independently). 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.

4. Application

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 [28]. From the ontologies, dynamic HTML can be generated,[3] and tags in the annotation can be represented as hyperlinks pointing to the corresponding definition [10].

In earlier **corpus query systems**, e.g., ANNIS [9], OLiA was used to formulate interoperable corpus queries: Assume we wanted to retrieve noun phrases from German newspaper corpora; instead of querying for `cat="NX"` on TüBa-D/Z [33] *or* `cat="NP"` on NEGRA [30], a query for `cat in {olia:NounPhrase}` was expanded into a disjunction of possible tags and formatted according to the query language under consideration. Only if corpora are represented as Linked Data (which is exceptional at the moment), they can be directly linked with OLiA Annotation Models, and queried without a query preprocessor [6]. When dealing with non-RDF corpora, ontology-based query rewriting using OLiA can be applied as sketched above, it was implemented, for example, in a generic query framework for linguistic corpora in heterogeneous XML-formats [16].

In a similar vein, OLiA can be employed in **NLP pipeline systems** for tagset-independent, interoperable information processing [1]. In this function, OLiA is part of the NLP Interchange Format (NIF) specification[4] to formalize linguistic annotations in a conceptually interoperable way. Using OLiA, the NLP2RDF platform developed on this basis unifies various NLP result outputs and maps them into RDF, as currently implemented for English [18] and Korean [17].

[2] Actually, the developers are sympathetic to the idea of releasing this data under an open license, Helen Aristar-Dry (for GOLD) and Menzo Windhouwer (for ISOcat), pers. communication, June 2012.

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

[4] `http://persistence.uni-leipzig.org/nlp2rdf/`

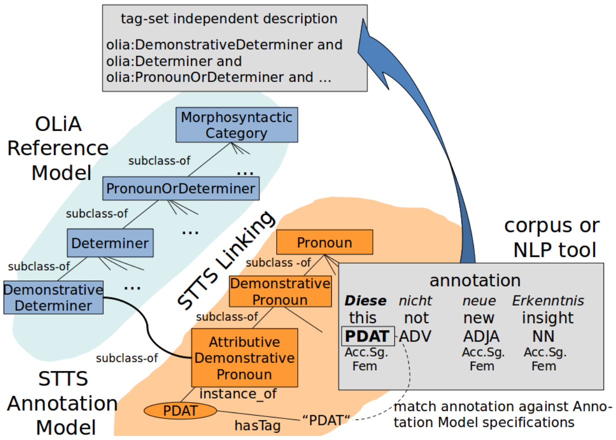

Fig. 1. Interpreting annotations in terms of the OLiA Reference Model

Figure 1 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 [31, file 4794]: Given the information that its part-of-speech annotations follow the STTS scheme [27], we may consult the corresponding Annotation Model,[5] and find that the tag `PDAT` matches the string value of the property `hasTag` of the individual `stts:PDAT`. The associated class `stts:AttributiveDemonstrativePronoun` is a subconcept of `olia:DemonstrativeDeterminer`.[6] 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: Recently, the OLiA ontologies have also been applied to represent grammatical specifications of machine-readable dictionaries, that are thus interoperable with OLiA-linked corpora [22,11]. Moreover, through the linking with External Reference Models, OLiA-linked resources are also interoperable with resources directly grounded in GOLD, ISOcat, etc.

Using Semantic Web formalisms to represent corpora and annotations also provides us with the possibility to develop novel, **ontology-based NLP** algorithms. One application are ensemble combination architectures, where different NLP modules (say, part-of-speech taggers) are applied in parallel, so that they produce annotations for one particular phenomenon, and that these annotations are then integrated. Using OLiA Reference Model specifications to integrate the analyses of multiple NLP tools for German, [5] showed that a simple majority-based combination increased both the robustness and the level of detail of morphosyntactic and morphological analyses: Despite imposing rigid ontological consistency constraints, abstraction from tool-specific representations and integration of different annotations on this basis resulted in an increase of recall. Similar results have been obtained with the OntoTag ontologies for Spanish [24].

We see possible applications of this technology in situations where multiple, domain-specific NLP tools are available. In a monolingual setting, this may be the case where rule-based morphologies [34] or parsers [32] are to be combined with robust statistical part-of-speech taggers, whose coarse-grained tagsets cannot be trivially mapped onto the detailed annotations provided by deep, rule-based systems. Here, OLiA representations leverage tools with different granularity. Currently, we experiment with multilingual annotation projection, where annotations are projected from *multiple* source languages onto parallel (translated) text in a less-resourced language. Using conceptual representations instead of complex strings as the basis of projection, different types of information can be drawn from different sources, e.g., aspect from Russian verbal morphology, and definiteness from English morphosyntax. For a (hypothetical) language that exhibits both features, information present in both sets of projected annotations may be adopted, merged and, most importantly, pruned using OLiA specifications.

5. Discussion

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 to establish interoperability of one annotation scheme with multiple terminology repositories. Further, annotation schemes and their linking to the Reference Model are formalized as

[5] `http://purl.org/olia/stts.owl`
[6] `http://purl.org/olia/stts-link.rdf`

separate OWL2/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 as apparently favored by GOLD and ISOcat, OLiA allows to recover information about sources of mismatches between Reference Model concepts and Annotation Model concepts, and its declarative linking supports 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. When confronted with conceptual overlap/fusion or ambiguity, many tagsets for, say, part-of-speech annotation introduce hybrid categories (e.g., `IN` for preposition *or* subordinating conjunction, or `TO` for *all* functions of English *to* [25]) or introduce tagset-specific notations, e.g., `|` for ambiguities [25], or + for cliticization/fusion [14]. As opposed to such ad-hoc solutions that may or may not be transparent to tagset users, OWL2/DL provides a W3C-standardized vocabulary to formalize these relationships, that also extends beyond individual tagsets: Ambiguity can be modeled as disjunction ($\sqcup$), conceptual overlap/fusion as conjunction ($\sqcap$).

Moreover, negation ($\neg$) is available in OWL2/DL. This is of particular importance 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` pertains to determiners only, so, 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 existing ontology browsers, and corrected independently from the interpretation-invariant Annotation Model and Reference Model. Furthermore, multiple *alternative* 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 $\sqcup$), 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, `PIAT` covers both uses).

In ISOcat, the problem of conflicting interpretations of data categories is currently *not* addressed, and the definitions provided are not always sufficient to distinguish 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 MULTEXT/East [15], which, however, conflates different uses of 'definite': (1) postfixed Determiner in Romanian, Bulgarian and Persian nouns or adjectives, (2) difference between 'full' and 'reduced' adjectives in Slavic (diachronically, full forms reflect a clitic pronominal), (3) a pattern of quantifier agreement in Slavic, and (4) the so-called 'definite conjunction' of Hungarian verbs. Even though the generic definition captures most of these different meanings, they remain incompatible with each other. 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*.[7] 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' [8]. The OLiA Reference Model provides such a fully developed taxonomy of linguistic categories. Recent activities to aug-

[7] Providing a top-down perspective does not automatically disclose such inconsistencies, but the resulting dialog between tagset provider and ontology developer may facilitate their detection, as in the example given above.

ment ISOcat with a relation category registry [29] may eventually lead to a comparable global perspective, so that the problem of conflicting interpretations of data categories may become more obvious to the ISOcat community, but these are still on-going developments.

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 following a request by the first 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 [13], 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 OWL2/DL allows one to formulate specific conditions for the linking, including the use of negation and disjunction. This complements the concept of Community-of-Practice Extensions in GOLD, that presuppose GOLD as an upper model providing the top-level categorizations for dependent ontologies, whereas OLiA remains agnostic about which External Reference Model to be adopted.

Conceptually, the OLiA ontologies are closer related to the OntoTag ontologies [2], that were also applied to develop NLP applications on the basis of ontological representations of linguistic annotations [24]. One important difference is that the OntoTag ontologies are considering only Iberian Romance languages (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 [1].

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 (LLOD) cloud.[8] In this context, the OLiA ontologies are used to provide linguistic reference terminology for lexical-semantic resources such as *lemon* [23] and Uby [11] as well as for linguistic corpora such as the Manually Annotated Sub-Corpus of the Open American National Corpus [19].[9]

Acknowledgments

We would like to thank Menzo Windhouwer, Steve Cassidy and four anonymous reviewers for valuable feedback and comments to this paper and its immediate predecessor [7]; beyond this, we thank OLiA users, contributors and funders. OLiA has originally been developed at the Collaborative Research Center (SFB) 441 "Linguistic Data Structures" (University of Tübingen, Germany) in the context of the project "Sustainability of Linguistic Resources" in cooperation with SFB 632 "Information Structure" (University of Potsdam, Humboldt-University Berlin, Germany) and SFB 538 "Multilingualism" (University of Hamburg, Germany) from 2006 to 2008. From 2007 to 2011, is has been maintained and further developed at SFB 632 in the context of the project "Linguistic Data Base". In 2012, the first author continued his research on OLiA in the context of PostDoc fellowship at the Information Sciences Institute of the University of Southern California funded by the German Academic Exchange Service (DAAD). The work of the second author was conducted in the context of the LOEWE cluster "Digital Humanities" at the Goethe-University Frankfurt (2011-2014).

In parts, this data set description is based on [7], shortened, updated and thoroughly revised.

References

[1] Buyko E, Chiarcos C, Pareja-Lora A (2008) Ontology-based interface specifications for a NLP pipeline architecture. In: Chair) NCC, Choukri K, Maegaard B, Mariani J, Odijk J,

[8] `http://linguistics.okfn.org/llod`

[9] `http://datahub.io/dataset/masc`

Piperidis S, Tapias D (eds) Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco

[2] Aguado de Cea G, Gomez-Perez A, Alvarez de Mon I, Pareja-Lora A (2004) OntoTag's linguistic ontologies: Improving semantic web annotations for a better language understanding in machines. In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04), Volume 2, IEEE Computer Society Washington, Washington, DC, pp 124–128

[3] Chiarcos C (2008) An ontology of linguistic annotations. GLDV-Journal for Language Technology and Computational Linguistics 23(1):1–16, URL `http://www.jlcl.org/2008_Heft1/Chiarcos.pdf`

[4] Chiarcos C (2010) Grounding an ontology of linguistic annotations in the Data Category Registry. In: Budin G, Declerck LRT, Wittenburg P (eds) Proceedings of the LREC 2010 Workshop Language Resource and Language Technology Standards (LT&LTS). State of the Art, Emerging Needs, and Future Developments, Valetta, Malta, pp 37–40

[5] Chiarcos C (2010) Towards robust multi-tool tagging. An OWL/DL-based approach. In: Hajic J, Carberry S, Clark S (eds) Proceedings of 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Association for Computational Linguistics, Uppsala, Sweden, pp 659–670

[6] Chiarcos C (2012) Interoperability of Corpora and Annotations. In: Chiarcos C, Nordhoff S, Hellmann S (eds) Linked Data in Linguistics, Springer, Heidelberg, pp 161–179

[7] Chiarcos C (2012) Ontologies of linguistic annotation: Survey and perspectives. In: Chair) NCC, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, pp 303–310

[8] Chiarcos C, Erjavec T (2011) OWL/DL formalization of the MULTEXT-East morphosyntactic specifications. In: Ide N, Meyers A, Pradhan S, Tomanek K (eds) Proc. 5th Linguistic Annotation Workshop (LAW 2011), Association for Computational Linguistics, Portland, Oregon, pp 11–20

[9] Chiarcos C, Götze M (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

[10] Chiarcos C, Dipper S, Götze M, et al (2008) A flexible framework for integrating annotations from different tools and tag sets. TAL (Traitement Automatique des Langues), Volume 49 49(2):217–246

[11] Eckle-Kohler J, McCrae J, Chiarcos C (2013) *lemonUby* - a large, interlinked, syntactically-rich resource for ontologies. Semantic Web Journal Special Issue on Multilingual Linked Open Data

[12] Farrar S, Langendoen DT (2010) An OWL-DL implementation of GOLD: An ontology for the Semantic Web. In: Witt AW, Metzing D (eds) Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology, Springer, Dordrecht, Germany

[13] Fillmore CJ (1968) The case for case. In: Bach E, Harms R (eds) Universals in Linguistic Theory, Holt, Rinehart, and Winston, New York, pp 1–88

[14] Francis WN, Kucera H (1979) Brown corpus manual. Tech. rep., Department of Linguistics, Brown University, Providence, Rhode Island, URL `http://icame.uib.no/brown/bcm.html`

[15] Francopoulo G, Declerck T, Sornlertlamvanich V, et al (2008) Data Category Registry: Morpho-syntactic and syntactic profiles. In: Witt A, Sasaki F, Teich E, Calzolari N, Wittenburg P (eds) Proc. LREC-2008 Workshop on Uses and Usage of Language Resource-Related Standards, Marrakech, Morocco, pp 31–40

[16] Georg Rehm CC Richard Eckart, Dellert J (2008) Ontology-based xquery'ing of xml-encoded language resources on multiple annotation layers. In: Chair) NCC, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Tapias D (eds) Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, http://www.lrec-conf.org/proceedings/lrec2008/

[17] Hahm Y, Lim K, Park J, Yoon Y, Choi KS (2012) Korean NLP2RDF resources. In: Weerasinghe R, Hussain S, Sornlertlamvanich V, Roxas REO (eds) Proc. 10th Workshop on Asian Language Resources (ALR 2012), The COLING 2012 Organizing Committee, Mumbai, India, pp 1–10

[18] Hellmann S, Lehmann J, Auer S, Brümmer M (2013) Integrating NLP using Linked Data. In: Alani H, Kagal L, Fokoue A, Groth P, Biemann C, Parreira JX, Aroyo L, Noy N, Welty C, Janowicz K (eds) Proc. 12th International Semantic Web Conference (ISWC 2013), Springer, Sydney, Australia, Lecture Notes in Computer Science, vol 8219, pp 98–113

[19] Ide N, Baker C, Fellbaum C, Fillmore C, Passonneau R (2008) Masc: the manually annotated sub-corpus of american english. In: Chair) NCC, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Tapias D (eds) Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, pp 68–73, http://www.lrec-conf.org/proceedings/lrec2008/

[20] Kemps-Snijders M (2010) RELISH: Rendering endangered languages lexicons interoperable through standards harmonisation. In: 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages, Valetta, Malta

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

[22] McCrae J, Spohr D, Cimiano P (2011) Linking lexical resources and ontologies on the semantic web with *lemon*. In: Proc. 8th Extended Semantic Web Conference (ESWC 2011), Heraklion, Greece, pp 245–259

[23] McCrae J, Montiel-Ponsoda E, Cimiano P (2012) Integrating WordNet and Wiktionary with *lemon*. In: Chiarcos C, Nordhoff S, Hellmann S (eds) Linked Data in Linguistics, Springer, Heidelberg, pp 25–34

[24] Pareja-Lora A, de Cea GA (2010) Ontology-based interoperation of linguistic tools for an improved lemma annotation in spanish. In: Chair) NCC, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D (eds) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta

[25] Santorini B (1990) Part-of-speech tagging guidelines for the Penn Treebank Project. Tech. rep., Department of Computer and Information Science, University of Pennsylvania

[26] Saulwick A, Windhouwer M, Dimitriadis A, Goedemans R (2005) Distributed tasking in ontology mediated integration of typological databases for linguistic research. In: Pastor O, e Cunha JF (eds) Advanced Information Systems Engineering, 17th International Conference, CAiSE 2005, Porto, Portugal, June 13-17, 2005, Proceedings, Springer, Porto, Portugal, pp 303–317

[27] Schiller A, Teufel S, Stöckert C, Thielen C (1999) Guidelines für das Tagging deutscher Textcorpora mit STTS. Tech. rep., Universities of Stuttgart and Tübingen, Germany

[28] Schmidt T, Chiarcos C, Lehmberg T, Rehm G, Witt A, Hinrichs E (2006) Avoiding data graveyards: From heterogeneous data collected in multiple research projects to sustainable linguistic resources. In: Proc. E-MELD workshop on Digital Language Documentation, East Lansing, Michigan

[29] Schuurman I, Windhouwer M (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

[30] Skut W, Brants T, Krenn B, Uszkoreit H (1998) A linguistically interpreted corpus of German newspaper text. In: Proc. ESSLLI Workshop on Recent Advances in Corpus Annotation, Saarbrücken, Germany

[31] Stede M (2004) The Potsdam Commentary Corpus. In: Webber B, Byron DK (eds) Proc. ACL-2004 Workshop on Discourse Annotation, ACL, Barcelona, Spain, pp 96–102

[32] Tapanainen P, Järvinen T (1997) A nonprojective dependency parser. In: Proc. 5th Conference on Applied Natural Language Processing (ANLP 1997), Washington, DC, pp 64–71

[33] Telljohann H, Hinrichs EW, Kübler S (2003) Stylebook for the Tübingen treebank of written German (TüBa-D/Z). Tech. rep., Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany

[34] Zielinski A, Simon C (2008) Morphisto: An open-source morphological analyzer for German. In: J P, B W, A Y (eds) Proceedings of the Conference on Finite State Methods in Natural Language Processing (FSMNLP), IOS Press, Ispra, Italy

[35] Zinn C, Hoppermann C, Trippel T (2012) The ISOcat registry reloaded. In: Simperl E, Cimiano P, Polleres A, Corcho O, Presutti V (eds) Proc. 9th Extended Semantic Web Conference (ESWC 2012), Springer, Heraklion, Greece, pp 27–31