OWL/DL formalization of the MULTEXT-East morphosyntactic specifications

Christian Chiarcos
University of Potsdam, Germany
chiarcos@uni-potsdam.de

Tomaž Erjavec
Jožef Stefan Institute, Slovenia
tomaz.erjavec@ijs.si

Abstract

This paper describes the modeling of the morphosyntactic annotations of the MULTEXT-East corpora and lexicons as an OWL/DL ontology. Formalizing annotation schemes in OWL/DL has the advantages of enabling formally specifying interrelationships between the various features and making logical inferences based on the relationships between them. We show that this approach provides us with a top-down perspective on a large set of morphosyntactic specifications for multiple languages, and that this perspective helps to identify and to resolve conceptual problems in the original specifications. Furthermore, the ontological modeling allows us to link the MULTEXT-East specifications with repositories of annotation terminology such as the General Ontology of Linguistics Descriptions or the ISO TC37/SC4 Data Category Registry.

1 Introduction

In the last 15 years, the heterogeneity of linguistic annotations has been identified as a key problem limiting the interoperability and reusability of NLP tools and linguistic data collections. The multitude of linguistic tagsets complicates the combination of NLP modules within a single pipeline; similar problems exist in language documentation, typology and corpus linguistics, where researchers are interested to access and query data collections on a homogeneous terminological basis.

One way to enhance the consistency of linguistic annotations is to provide explicit semantics for tags by grounding annotations in terminology repositories such as the General Ontology of Linguistics Descriptions (Farrar and Langendoen, 2003, GOLD) or the ISO TC37/SC4 Data Category Registry (Kemps-Snijders et al., 2009, ISOcat). Reference definitions provide an interlingua that allows the mapping of linguistic annotations from annotation scheme A to scheme B. This application requires linking annotation schemes with the terminological repository. This relation can be formalized within the Linked Data paradigm (Berners-Lee, 2006), which requires the use of uniform resource identifiers (URIs), the hypertext transfer protocol (HTTP), standard representation formats (such as RDF) and links to other URIs. Here, we propose a formalization of this linking in OWL/DL, a notational variant of the Description Logic $\mathcal{SHOIN}(\mathcal{D})$ that builds on RDF and Linked Data.

Another way to enhance the consistency of linguistic annotations is to make use of cross-linguistic meta schemes or annotation standards, such as EAGLES (Leech and Wilson, 1996). The problem is that these enforce the use of the same categories across multiple languages, and this may be inappropriate for historically and geographically unrelated languages. For specific linguistic and historical regions, the application of standardization approaches has, however, been performed with great success, e.g., for Western (Leech and Wilson, 1996) and Eastern Europe (Erjavec et al., 2003) or the Indian subcontinent (Baskaran et al., 2008).

Proceedings of the Fifth Law Workshop (LAW V), pages 11–20,
Portland, Oregon, 23-24 June 2011.

In this paper, we illustrate differences and commonalities of both approaches by creating an OWL/DL terminology repository from the MULTEXT-East (MTE) specifications (Erjavec et al., 2003; Erjavec, 2010), which define features for the morphosyntactic level of linguistic description, instantiate them for 16 languages and provide morphosyntactic tagsets for these languages. The specifications are a part of the MTE resources, which also include lexicons and an annotated parallel corpus that use these morphosyntactic tagsets.

The encoding of the MTE specifications follows the Text Encoding Initiative Guidelines, TEI P5 (TEI Consortium, 2007), and this paper concentrates on developing a semi-automatic procedure for converting them from TEI XML to OWL. While TEI is more appropriate for authoring the specifications and displaying them in a book-oriented format, the OWL encoding has the advantages of enabling formally specifying interrelationships between the various features (concepts, or classes) and making logical inferences based on the relationships between them, useful in mediating between different tagsets and tools (Chiarcos, 2008).

2 The MULTEXT-East (MTE) Morphosyntactic Specifications

The MTE morphosyntactic specifications define attributes and values used for word-level syntactic annotation, i.e., they provide a formal grammar for the morphosyntactic properties of the languages covered. The specifications also contain commentary, bibliography, notes, etc. Following the original MULTEXT proposal (Ide and Véronis, 1994), the specifications define 14 categories (parts of speech), and for each its attributes, their values, and the languages that every attribute-value pair is appropriate for. The morphosyntactic specifications also define the mapping between the feature structures and morphosyntactic descriptions (MSDs). MSDs are compact strings used as tags for corpus annotation and in the morphosyntactic lexicons. For example, the MSD Ncmsn is equivalent to the feature structure consisting of the attribute-value pairs Noun, Type=common, Gender=masculine, Number=singular, Case=nominative.

The specifications currently cover 16 languages, in particular: Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Resian, Romanian, Russian, Serbian, Slovak, Slovene, and Ukrainian. For a number of these languages the specifications have become a de-facto standard and, for some, the MTE lexicons and corpora are still the only publicly available datasets for this level of linguistic description.[1]

Table 1 lists the defined categories and gives the number of distinct attributes, attribute-value pairs and the number of MTE languages which distinguish the category. The feature-set is quite large, as many of the languages covered have very rich inflection, are typologically different (inflectional, agglutinating), but also have independent traditions of linguistic description; this also leads to similar phenomena sometimes being expressed by different means (see Sect. 4.3).

Category	Code	Atts	Att-Vals	Langs
Noun	N	14	68	16
Verb	V	17	74	16
Adjective	A	17	79	16
Pronoun	P	19	97	16
Determiner	D	10	32	3
Article	T	6	23	3
Adverb	R	7	28	16
Adposition	S	4	12	16
Conjunction	C	7	21	16
Numeral	M	13	81	16
Particle	Q	3	17	12
Interjection	I	2	4	16
Abbreviation	Y	5	35	16
Residual	X	1	3	16

Table 1: MULTEXT categories with the number of MULTEXT-East defined attributes, attribute-value pairs and languages.

The specifications are encoded as a TEI document, consisting of an introductory part, the Common and the Language Specific Specifications, the latter two organized into tables by the

[1]The MTE specifications, as well as the other MTE resources, are available from the Web page of the project at `http://nl.ijs.si/ME/`.

```
<table n="msd.cat" xml:lang="en">
  <head>Common specifications for Noun</head>
  <row role="type">
    <cell role="position">0</cell>
    <cell role="name">CATEGORY</cell>
    <cell role="value">Noun</cell>
    <cell role="code">N</cell>
    <cell role="lang">en</cell>
    <cell role="lang">ro</cell>
    <cell role="lang">sl</cell>
    ...
  </row>
  <row role="attribute">
    <cell role="position">1</cell>
    <cell role="name">Type</cell>
    <cell>
      <table>
        <row role="value">
          <cell role="name">common</cell>
          <cell role="code">c</cell>
          <cell role="lang">en</cell>
          ...
```

Figure 1: Common table for Noun

14 defined categories.

Figure 1 gives the start of the Common table for Noun. It first gives the category, the languages that distinguish it, and then its attributes with their values; the meaning of a particular row or cell is given by its role attribute. As with the category, each attribute-value is also qualified by the languages that make use of the feature. Note that MTE is a positional tagset that specifies the position of the attribute in the MSD string, and the one-letter code of its value, so that Nc would correspond to Noun, Type=common.

The language-specific sections also contain tables for each category, which are similar to the common tables in that they repeat the attributes and their values, although only those appropriate for the language. The language-specific tables can also contain localization information, i.e., the names of the categories, attributes, their values and codes in the particular language, in addition to English. This enables expressing the feature structures and MSDs either in English or in the language in question. Furthermore, each language-specific section can also contain an index listing all valid MSDs. This index is augmented with frequency information and examples of usage drawn for a corpus.

In addition to the source TEI P5 XML, the MTE specifications are delivered in various derived formats, in particular HTML for reading and as tabular files, which map the MSD tagset into various feature decompositions.

3 Linking annotation schemes with terminology repositories

3.1 Linguistic terminology initiatives

There have been, by now, several approaches to develop terminology repositories and data category registries for language resources, systems for mapping between diverse (morphosyntactic) vocabularies and for integrating annotations from different tools and tagsets, ranging from early texts on annotation standards (Bakker et al., 1993; Leech and Wilson, 1996) over relational models and concept hierarchies (Bickel and Nichols, 2002; Rosen, 2010) to more formal specifications in OWL/RDF (or with OWL/RDF export), e.g., the already mentioned GOLD and ISOcat, OntoTag (Aguado de Cea et al., 2002) or the Typological Database System ontology (Saulwick et al., 2005).

Despite their common level of representation these efforts have not yet converged into a unified and generally accepted ontology of linguistic annotation terminology and there is still a considerable amount of disagreement between their definitions. As these repositories nevertheless play an important role in their respective communities, it is desirable to link the MTE specifications with the most representative of them, notably with GOLD and the morphosyntactic profile of ISOcat. As we argue below, different design decisions in the terminology repositories make it necessary to use a linking formalism that is capable of expressing both disjunctions and conjunctions of concepts. For this reason, we propose the application of OWL/DL.

By representing the MTE specifications, the repositories, and the linking between them as separate OWL/DL models, we follow the architectural concept of the OLiA architecture (Chiarcos, 2008), see Sect. 5.

3.2 Annotation mapping

The classic approach to link annotations with reference concepts is to specify rules that define a direct mapping (Zeman, 2008). 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 and Hansen, 2002). ISOcat (like other terminological repositories) 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. **Contraction**s pose similar problems: English *gonna* combines *going* (PTB tag `VBG`, Marcus et al., 1994) and *to* (`TO`). If whitespace tokenization is applied, both tags need to be assigned to the same token.

A related 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 `subordinating Conjunction`/DC-1393. Without additional disambiguation, `ICSt` needs 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 meaning: While overlapping/contracted categories are in the intersection ($\sqcap$) of reference categories, ambiguous categories are in their join ($\sqcup$). This difference is relevant for subsequent processing, e.g., to decide whether disambiguation is necessary. A mapping approach, however, fails to distinguish $\sqcap$ and $\sqcup$.

The linking between reference categories and annotations requires a formalism that can distinguish intersection and join operators. A less expressive linking formalism that makes use of a 1:1 (or 1:n) mapping between annotation concepts and reference concepts can lead to inconsistencies when mapping annotation concepts from an annotation scheme A to an annotation scheme B if these use the same terms with slightly deviating definitions, as noted, for example, by Garabík et al. (2009) for MTE.

3.3 Annotation linking with OWL/DL

OWL/DL is a formalism that supports the necessary operators and flexibility. Reference concepts and annotation concepts are formalized as OWL classes and the linking between them can be represented by `rdfs:subClassOf` ($\sqsubseteq$). OWL/DL provides `owl:intersectionOf` ($\sqcap$), `owl:unionOf` ($\sqcup$) and `owl:complementOf` ($\neg$) operators and it allows the definition of properties and restrictions on the respective concepts. As an example, the MTE Definiteness=definite refers to either a clitic determiner or ($\sqcup$) to the 'definite conjunction' of Hungarian verbs. More precisely, it is in the intersection between these and ($\sqcap$) a category for ambiguous feature values (Sect. 4.3).

An OWL/DL-based formalization has the additional advantage that it can be linked with existing terminology repositories that are available in OWL or RDF, e.g., GOLD or ISOcat (Chiarcos, 2010). The linking to other terminology repositories will be subject of subsequent research. In this paper, we focus on the development of an OWL/DL representation of MTE morphosyntactic specifications that represents a necessary precondition for OWL/DL-based annotation linking.

4 Building the MTE ontology

We built the MTE ontology[2] in a three-step scenario: first, a preliminary OWL/DL model of the common MTE specifications was created (Sect. 4.1); we then built language-specific subontologies and linked them to the common ontology (Sect. 4.2); finally, the outcome of this process

[2] All MTE ontologies are available under `http://nl.ijs.si/ME/owl/` under a Creative Commons Attribution licence (CC BY 3.0).

was discussed with a group of experts and revised (Sect. 4.3).

4.1 Common specifications

Following the methodology described by Chiarcos (2008), the structure of the MTE ontology was derived from the original documentation. The initial ontology skeleton was created automatically (the organization of the specifications was exploited to develop an XSLT script that mapped TEI XML to OWL), but subsequently manually augmented with descriptions and examples found in the individual languages.

1. Two top-level concepts `MorphosyntacticCategory` and `MorphosyntacticFeature` represent root elements of the MTE ontology. An object property `hasFeature` maps a `MorphosyntacticCategory` onto one or multiple `MorphosyntacticFeature` values.

2. All MSD categories are subconcepts of `MorphosyntacticCategory`, e.g., `Noun`, `Verb`, `Adjective`, etc.

3. For every category, the MTE attribute Type was used to infer subcategories, e.g., the concept `ExclamativePronoun` ($\sqsubseteq$ `Pronoun`) for Pronoun/Type=exclamative.

4. From more specialized type attributes (e.g., Wh_Type, Coord_Type, Sub_Type, and Referent_Type), additional subcategories were induced at the next deeper level, e.g., `SimpleCoordinatingConjunction` ($\sqsubseteq$ `CoordinatingConjunction`) from Conjunction/Type=coordinating, Coord_Type=simple.

5. All remaining attributes are subconcepts of `MorphosyntacticFeature`, e.g., `Aspect`, `Case`, etc.

6. For every subconcept of `MorphosyntacticFeature` (e.g., `Aspect`) a corresponding `hasFeature` subproperty (e.g., `hasAspect`) was introduced, with the morphosyntactic feature as its range and the join of morphosyntactic categories it can cooccur with as its domain. An additional constraint restricts its cardinality to at most 1.

7. All attribute values are represented as subclasses of the corresponding attribute concept, e.g., `AbessiveCase` (for Case=abessive) as a subconcept of `Case`.[3]

8. Every concept was automatically augmented with a list of up to 10 examples for every language which were drawn from the language-specific MSD index.

4.2 Language-specific subontologies

Having represented the common MTE specifications in OWL, we decided to represent the annotation scheme for every language in a separate OWL model, and to make use of the OWL import mechanism to link it with the common specifications. The language-specific subontologies do not specify their own taxonomy, but rather inherit the concepts and properties of the common model. Unlike the common model, they include individuals that provide information about the tags (MSDs) used for this particular language.

Every individual corresponds to an MSD tag. We use data properties of the OLiA system ontology[4] to indicate its string realization (e.g., `system:hasTag 'Ncmsn'`) and the designator of its annotation layer (e.g., `system:hasTier 'pos'`). Additionally, `rdfs:comment` elements contain all examples of the original MSD specifications.

In accordance to the specified annotation values, every individual is defined as an instance of the corresponding `MorphosyntacticCategory` (e.g., `Noun`) and `MorphosyntacticFeature` (e.g., `SingularNumber`) from the common specifications. Additionally, for every `MorphosyntacticFeature` (e.g., `Number`, the superconcept of `SingularNumber`), it is assigned

[3] This ontology does not contain individuals. In our approach, individuals represent feature bundles in the language-specific subontologies, corresponding to the individual MSD tags. (or, in other application scenarios, the token that the tag is applied to).

[4] `http://nachhalt.sfb632.uni-potsdam.de/owl/system.owl`, prefix `system`

```
<mte:Noun rdf:ID="Ncmsn_sl">
  <system:hasTag>Ncmsn</system:hasTag>
  <system:hasTier>pos</system:hasTier>
  <rdf:type
    rdf:resource="...#CommonNoun"/>
  <rdf:type
    rdf:resource="...#MasculineGender"/>
  <rdf:type
    rdf:resource="...#SingularNumber"/>
  <rdf:type
    rdf:resource="...#NominativeCase"/>
  <mte:hasGender rdf:resource="#Ncmsg_sl"/>
  <mte:hasNumber rdf:resource="#Ncmsg_sl"/>
  <mte:hasCase rdf:resource="#Ncmsg_sl"/>
  <rdfs:comment>e.g., cas, svet, denar, ...
</mte:Noun>
```

Figure 2: MSD Ncmsn in the Slovene subontology

itself as target of the corresponding object property (e.g., `hasNumber`).

Figure 2 shows the subontology entry for the tag `Ncmsn` in the Slovene subontology. The individual could thus be retrieved with the following queries for "singular noun":

```
(1) Noun and hasNumber some
    SingularNumber
(2) Noun and SingularNumber
```

The language-specific subontologies were fully automatically created from the TEI XML using XSLT scripts. During the revision of the common specifications, these scripts were updated and reapplied.

4.3 Revision of the initial OWL model

After the automatic conversion from XML to OWL the resulting ontology skeleton of the common specifications was manually augmented with descriptions, explanations and selected examples from the language-specific MTE specifications. Furthermore, concept names with abbreviated or redundant names were adjusted, e.g., the concept `CorrelatCoordConjunction` (Coord_Type=correlat) was expanded to `CorrelativeCoordinatingConjunction`, and `DefiniteDefiniteness` (Definiteness=definite) was simplified to `Definite`. Finally, if one attribute value represents a specialization of another, the former was recast as a subconcept of the latter (e.g., `CliticProximalDeterminer` $\sqsubseteq$ `CliticDefiniteDeterminer`).

Moreover, a number of potential problems were identified. Some of them could be addressed by consulting MTE-related publications (Qasemizadeh and Rahimi, 2006; Dimitrova et al., 2009; Derzhanski and Kotsyba, 2009), but most were solved with the help of the original authors of the MTE specifications and an open discussion with these experts over a mailing list.

The problems fall in two general classes: (a) terminological problems, and (b) conceptual problems. By terminological problems we mean that a term required a more precise definition than provided in the MTE specifications; conceptual problems pertain to design decisions in a positional tagset (overload: the same annotation refers to two different phenomena in different languages) and to artifacts of the creation process of the MTE specifications (redundancies: the same phenomenon is represented in different ways for different languages). Figure 3 shows a fragment of the MTE ontology that showed all types of conceptual problems as described below.

Terminological problems include the use of non-standard or language-specific terminology (e.g., Clitic=burkinostka for conventional collocations in Polish, or Case=essive-formal for Hungarian), and the need to understand design decisions that were necessary for language-specific phenomena (e.g., Numeral/Class=definite34 for Czech and Polish quantifiers with the same patterns of agreement as the numerals 3 and 4).

In the course of the revision, most non-standard terms were replaced with conventional, language-independent concept names, and language-specific phenomena were documented by adding relevant excerpts from discussions or literature as `owl:versionInfo`.

For a few concepts, no language-independent characterization could be found. For example, Numeral/Form=m_form refers to numerals with the suffix *-ma* in Bulgarian (a special form of the numerals '2' to '7' for persons of masculine gender). In the ontology, the concept `MFormNumeral` is preserved, but it is constrained so that every instance matches the fol-

lowing OWL/DL expression:

(3) `CardinalNumber and hasAnimacy some Animate and hasGender some Masculine`

Attribute overload means that one attribute groups together unrelated phenomena from different languages. In a positional tagset, attribute overload is a natural strategy to achieve compact and yet expressive tags. As every attribute requires its own position in the tag, the length of MSD tags grows with the number of attributes. Overload thus reduces tag complexity. To an ontological model, however, these complexity considerations do not apply, whereas proper conceptual differentiations are strongly encouraged.

We thus decided to disentangle the various senses of overloaded attributes. For example, the `MorphosyntacticFeature Definiteness`, is split up in three subconcepts (cf. Fig. 3).

`CliticDeterminerType`: presence of a postfixed article of Romanian, Bulgarian and Persian nouns and adjectives.

`ReductionFeature`: the difference between full and reduced adjectives in many Slavic languages.

`PersonOfObject`: the so-called 'definite conjugation' of Hungarian verbs.

Value overload has a similar meaning to attribute overload. Definiteness=definite, for example, can refer to a clitic definite determiner (a `CliticDeterminerType` in Romanian and Bulgarian), to a clitic determiner that expresses specificity (a `CliticDeterminerType` in Persian), or to a verb with a definite 3rd-person direct object (a `PersonOfObject` in Hungarian).

In the ontology, this is represented by defining `Definite` as a subconcept of the `owl:join` (⊔) of `CliticDefiniteDeterminer`, `CliticSpecificDeterminer` and `PersonOfObject`. Additional concepts, e.g., `AmbiguousDefinitenessFeature`, were created to anchor ambiguous concepts like `Definite` in the taxonomy (see Fig. 3).

Redundancy: For many languages, the MTE specifications were created in a bottom-up fashion, where existing NLP tools and lexicons were integrated with a pre-existing taxonomy of annotation categories. Language-specific features were introduced when necessary, but sometimes in different ways for the same phenomenon in closely related languages. The MTE specifications thus comprise a certain degree of redundancy.

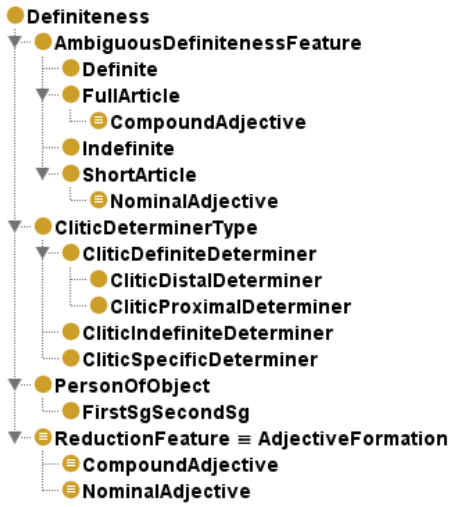

Figure 3: Definiteness in the MTE ontology

For example, the distinction between full and reduced adjectives in Slavic languages is expressed differently: For Czech, reduced adjectives are marked by Formation=nominal, but for Polish by Definiteness=short-art.

In the ontology, such redundancies are resolved by `owl:equivalentClass` statements, marked by ≡ in Fig. 3.

5 Summary and Discussion

We have described the semi-automatic creation of an ontological model of the MTE morphosyntactic specifications for 16 different languages. Such a model may be fruitfully applied in various ways, e.g., within an NLP pipeline that uses ontological specifications of annotations rather than their string representations (Buyko et al., 2008; Hellmann, 2010). The ontological modeling may serve also as a first step towards an ontology-based documentation of the annotations within a corpus query system (Rehm et al., 2007; Chiarcos et al., 2008),

or even the ontological modeling of entire corpora (Burchardt et al., 2008; Hellmann et al., 2010) and lexicons (Martin et al., 2009). As an interesting side-effect of the OWL conversion of the entire body of MTE resources, they could be easily integrated with existing lexical-semantic resources as Linked Data, e.g., OWL/RDF versions of WordNet (Gangemi et al., 2003), which are currently being assembled by various initiatives, e.g., in the context of the LOD2 project (`http://lod2.eu`) and by the Open Linguistics Working Group at the OpenKnowledge Foundation (`http://linguistics.okfn.org`).

Another very important element is that the ontological modeling of the MTE annotations allows it to be interpreted in terms of existing repositories of annotation terminology such as ISOcat and GOLD. A bridge between these terminology repositories and the MTE ontology may be developed, for example, by integrating the ontology in an architecture of modular ontologies such as the Ontologies of Linguistic Annotations (Chiarcos, 2008, OLiA), where the linking between annotations and terminology repositories is mediated by a so-called 'Reference Model' that serves as an interface between different levels of representation.

The MTE ontology will be integrated in this model as an annotation model, i.e., its concepts will be defined as subconcepts of concepts of the OLiA Reference Model and thereby inherit the linking with GOLD (Chiarcos et al., 2008) and ISOcat (Chiarcos, 2010). The linking with these standard repositories increases the comparability of MTE annotations and it serves an important documentation function.

More important than merely *potential* applications of the MTE ontology, however, is that its creation provides us with a new, global perspective on the MTE specifications. A number of internal inconsistencies could be identified and strategies for their resolution (or formalization) were developed. Redundancies and overload were documented, and we further added expert definitions of controversial or non-standard concepts. When used as a documentation, these specifications may prevent misunderstandings with respect to the meaning of the actual annotations. For later versions of the MTE morphosyntactic specifications, they may even guide the refactoring of the annotation scheme.

The result of the development process described above is a prototype, that has to be augmented with definitions for non-controversial and well-understood concepts, which can be derived from the linking with OLiA, GOLD and ISOcat.

As for its language type, our strategy to resolve overload requires OWL/DL (`owl:join`). Without value overload and redundancy, the ontology would be OWL/Lite, as were the initial ontologies (Sect. 4.1 and Sect. 4.2). However, the current modeling is still sufficiently restricted to allow the application of reasoners, thereby opening up the possibility to use SemanticWeb technologies on MTE data, to connect it with other sources of information and to draw inferences from such Linked Data.

We would also like to point out that the conversion of the MTE specifications to OWL required relatively little effort. The total time required for conversion (without the revision phase) took approximately four days of work for a computational linguist familiar with OWL and part-of-speech tagsets in general (the most labor-intense part were discussions and literature consultation during the revision phase). Given the complexity of the MTE specifications (a highly elaborate set of morphosyntactic specifications for 16 typologically diverse languages and with more than thousand tags for many of the languages), this may be regarded an upper limit for the time necessary to create OWL models for annotation schemes.

We have thus not only shown that the ontological modeling of annotation schemes is possible and that it allows us to use our data in novel ways and to perform consistency control, but also that this was achievable with relatively low efforts in time and personnel.

Acknowledgements

The authors would like to thank the members of the mocky-l mailing list for their invaluable input; all errors in the paper remain our own. The research on linguistic ontologies described in this paper was partially funded by the German Research Foundation (DFG) in the context of the Collaborative Research Center (SFB) 632.

References

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

Dik Bakker, Osten Dahl, Martin Haspelmath, Maria Koptjevskaja-Tamm, Christian Lehmann, and Anna Siewierska. 1993. EUROTYP guidelines. Technical report, European Science Foundation Programme in Language Typology.

S. Kalika Bali Baskaran, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Monojit Choudhury, Girish Nath Jha, S. Rajendran, K. Saravanan, L. Sobha, and KVS Subbarao. 2008. Designing a common POS-tagset framework for Indian languages. In *6th Workshop on Asian Language Resources*, pages 89–92, Hyderabad, India.

Tim Berners-Lee. 2006. Design issues: Linked data. `http://www.w3.org/DesignIssues/LinkedData.html` (May 11, 2011).

Balthasar Bickel and Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In *Proceedings of the LREC 2002 Workshop on Resources and Tools in Field Linguistics*, Las Palmas, Spain, May.

Sabine Brants and Silvia Hansen. 2002. Developments in the TIGER annotation scheme and their realization in the corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1643–1649, Las Palmas, Spain, May.

Aljoscha Burchardt, Sebastian Padó, Dennis Spohr, Anette Frank, and Ulrich Heid. 2008. Formalising Multi-layer Corpora in OWL/DL – Lexicon Modelling, Querying and Consistency Control. In *Proceedings of the 3rd International Joint Conference on NLP (IJCNLP 2008)*, Hyderabad, India, January.

Ekaterina Buyko, Christian Chiarcos, and Antonio Pareja-Lora. 2008. Ontology-based interface specifications for a NLP pipeline architecture. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

Christian Chiarcos, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. 2008. A flexible framework for integrating annotations from different tools and tag sets. *Traitement Automatique des Langues (TAL)*, 49(2).

Christian Chiarcos. 2008. An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16. Foundations of Ontologies in Text Technology, Part II: Applications.

Christian Chiarcos. 2010. Grounding an ontology of linguistic annotations in the Data Category Registry. In *Proceedings of the LREC 2010 Workshop on Language Resource and Language Technology Standards (LR&LTS 2010)*, Valetta, Malta, May.

Ivan Derzhanski and Natalia Kotsyba. 2009. Towards a consistent morphological tagset for Slavic languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian. In *Mondilex Third Open Workshop*, pages 9–26, Bratislava, Slovakia, April.

Ludmila Dimitrova, Radovan Garabík, and Daniela Majchráková. 2009. Comparing Bulgarian and Slovak Multext-East morphology tagset. In *Mondilex Second Open Workshop: Organization and Development of Digital Lexical Resources*, pages 38–46, Kyiv, Ukraine, February.

Tomaž Erjavec, Cvetana Krstev, Vladimír Petkevič, Kiril Simov, Marko Tadić, and Duško Vitas. 2003. The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. In *Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages*, pages 25–32.

Tomaž Erjavec. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta, May.

Scott Farrar and D. Terence Langendoen. 2003. A linguistic ontology for the semantic web. *Glot International*, 7(3):97–100.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The OntoWordNet project: Extension and axiomatization of conceptual relations in WordNet. In R. Meersman and Z. Tari, editors, *Proceedings of On the Move to Meaningful Internet Systems (OTM 2003)*, pages 820–838, Catania, Italy, November.

Radovan Garabík, Daniela Majchráková, and Ludmila Dimitrova. 2009. Comparing Bulgarian and Slovak MULTEXT-East morphology tagset. In *Mondilex Second Open Workshop: Organization and Development of Digital Lexical Resources*, pages 38–46, Kyiv, Ukraine. Dovira Publishing House.

Sebastian Hellmann, Jörg Unbehauen, Christian Chiarcos, and Axel-Cyrille Ngonga Ngomo. 2010. The TIGER Corpus Navigator. In *Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories (TLT 2010)*, pages 91–102, Tartu, Estonia, December.

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

Nancy Ide and Jean Véronis. 1994. MULTEXT (Multilingual Tools and Corpora). In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)*, pages 90–96, Kyoto.

Marc Kemps-Snijders, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2009. ISOcat: remodelling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies*, 4(4):261–276.

Geoffrey Leech and Andrew Wilson. 1996. Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Report EAG–TCWG–MAC/R, ILC, Pisa. `http://www.ilc.cnr.it/EAGLES96/annotate/` (May 11, 2011).

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

Fabienne Martin, Dennis Spohr, and Achim Stein. 2009. Representing a resource of formal lexical-semantic descriptions in the Web Ontology Language. *Journal for Language Technology and Computational Linguistics*, 21:1–22.

Behrang Qasemizadeh and Saeed Rahimi. 2006. Persian in MULTEXT-East framework. In Tapio Salakoski, Filip Ginter, Sampo Pyysalo, and Tapio Pahikkala, editors, *Advances in Natural Language Processing, Proceedings of the 5th International Conference on NLP (FinTAL 2006)*, pages 541–551, Turku, Finland, August.

Georg Rehm, Richard Eckart, and Christian Chiarcos. 2007. An OWL-and XQuery-based mechanism for the retrieval of linguistic patterns from XML-corpora. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgaria, September.

Alexandr Rosen. 2010. Mediating between incompatible tagsets. In *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pages 53–62, Tartu, Estonia, December.

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

Adam Saulwick, Menzo Windhouwer, Alexis Dimitriadis, and Rob Goedemans. 2005. Distributed tasking in ontology mediated integration of typological databases for linguistic research. In *Proceedings of the 17th Conference on Advanced Information Systems Engineering (CAiSE 2005)*, Porto, Portugal, June.

TEI Consortium, editor. 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.