An ontology of linguistic annotations

Christian Chiarcos

Angaben zur Veröffentlichung / Publication details:

Chiarcos, Christian. 2008. "An ontology of linguistic annotations." *Journal for Language Technology and Computational Linguistics* 23 (1): 1–16.
https://doi.org/10.21248/jlcl.23.2008.98.

Christian Chiarcos

An ontology of linguistic annotations

This paper describes development and design of an ontology of linguistic annotations, primarily word classes and morphosyntactic features, based on existing standardization approaches (e.g. EAGLES), a set of annotation schemes (e.g. for German, STTS and morphological annotations), and existing terminological resources (e.g. GOLD).
The ontology is intended to be a platform for terminological integration, integrated representation and ontology-based search across existing linguistic resources with terminologically heterogeneous annotations. Further, it can be applied to augment the semantic analysis of a given text with an ontological interpretation of its morphosyntactic analysis.

1 Background and motivation

This paper describes the development and the design of an ontology of linguistic annotations. The ontology is primarily intended as a platform for the terminological integration, integrated representation and access to existing linguistic resources with terminologically heterogeneous annotations. This means that existing annotations are mapped onto ontological representations, according to the underlying semantics a certain tag is assigned.

Beyond this, the ontology can also be applied to the ontological representation of linguistic information in a hybrid model of automated text analysis covering both semantic and morphosyntactic information. Cimiano and Reyle (2003) developed the idea that both semantic and syntactic analysis must integrated within a hybrid system using both types of information. Further, de Cea et al. (2004) proposed to model the dependencies between these two modules at the same level of conceptual representation, i.e. a system of multiple ontologies covering both the semantic concepts of an analyzed text, and the semantics of its linguistic (morphosyntactic) annotations.

Thus, the ontology-based integration of linguistic annotation terminology can be used in two different ways:

Annotation mining perspective The ontology specifies a reference inventory of terms and definitions to which different annotations refer. But also, the ontology assembles and formalizes the available annotation documentation which a user has to consult

to explore a corpus. The annotation mining perspective is basically that of a linguist searching for examples in a corpus.

NLP perspective The ontology specifies a framework for tag-set independent representation and semantic interpretation of linguistic annotations as produced, for example, by a statistical tagger. In this function, an ontology provides a semantic interpretation of linguistic annotations.

The annotation mining perspective is particularly relevant to typological and corpus linguistic research. Attempts for the standardization of morphosyntactic annotation have been made, basically presented by the lists of terms and abbreviations, e.g. the EUROTYP guidelines (König et al., 1993), but also as terminological networks and ontologies, e.g. the Generalized Ontology of Linguistic Description (GOLD) (Farrar and Langendoen, 2003).

Related research on the NLP perspective has mostly relied on the specification of a standard repertoire of linguistic terms which may be used by or must be supported by standard-conformant tag sets, the most prominent example being the EAGLES recommendations for morphosyntax (Leech and Wilson, 1996). An ontology for the linguistic annotations produced by different parsers for Spanish has been described by de Cea et al. (2004).

The classical domain of an ontology besides the annotation mining perspective and the NLP perspective is the **terminological perspective**. In this function, an ontology is employed to specify the linguistic terminology as used in an existing body of literature, a line of research currently explored by Schneider (2007), but not specifically tailored to annotation-relevant terminology.

The ontology presented here, however, is designed with a primary focus on the annotation mining perspective. It is developed in the context of the project "Sustainability of Linguistic Data" to enhance the terminological integration of the resources assembled by three German Collaborative Research Centers, CRC 441 (Tübingen, "Linguistic Data Structures") CRC 538 (Hamburg, "Multilingualism"), and CRC 632 (Potsdam/Berlin,"Information Structure"). Furthermore, the ontologies are applied for tag-set independent, ontology-based corpus querying.

This search functionality represents one of the most important fields of applications for the ontology described here (see Chiarcos (2006) for more details). Still, in the context of this volume, I concentrate on the description of the ontologies themselves, and in particular, in their function as a means for conservation and systematization of annotation documentation. Also, their potential application for the purpose of NLP applications will be shortly sketched, as the ontology also deals with annotation schemes for German, English and Russian which are technically relevant.

Here, I concentrate on part of speech (POS) and morphological annotation. Our research centers create and use morphosyntactically annotated corpora for about 42

meta tag sets and multilingual tag sets		language-specific tag sets	languages	granularity
	n/a	Tibetan tag set	Tibetan	$\geq$ 36 tags
EAGLES	generalization over existing tag sets for European languages	Susanne	English	$\approx$ 420 tags
		STTS, 3 variants	German	54 (718) tags
		Menota	Old Norse	$\approx$ 13055 tags
MULTEXT-East	adaptation of EAGLES	Russian tag set	Russian	$\geq$ 877 tags
CRC632 annotation standard	designed for typological research	n/a	$>$ 26 languages	$\approx$ 79 tags
CRC538/E2 tag set	reduced tag set for acquisition studies	n/a	German, Romance, Basque	$\geq$ 8 tags

Table 1: Tag sets and meta-tag sets for part of speech (POS) annotation in the CRCs.

languages or language stages from practically all parts of the world, cf. tab. 1. With respect to annotation schemes applied, Susanne (Sampson, 1995, English), STTS (Schiller et al., 1995, German) and the Uppsala tag set (Russian) are also technically relevant, as they are used by existing POS taggers.

The scenario of the sustainability project is that a linguist can assess the value of a given resource without being too familiar with the annotation scheme. Here, the user may encounter even greater problems hindering the direct access to the data or proper interpretation of tags: tag names are cryptic and appear in idiosyncratic variants, researchers from different communities use tags with the same names, but different definitions, tag definitions can be extremely complex, or be missing completely, or be of differing granularity.

As an example, the dialects of STTS show some degree of variation in the tag used for pronominal adverbs (`PAV`, `PROAV`, `PROP`). Such seemingly marginal variations can lead to false conclusions about the distribution of grammatical categories if they remain undetected, especially in queries with regular expressions. Further, tag sets tend to apply surface ambiguity as a criterion for the assignment of POS tags. As an example, the STTS tag `VAFIN`, intuitively interpreted as "auxiliary verb", applies to all uses of German *haben* and *sein*, in both auxiliary function ("to have, to be") and lexical use ("to own, to exist"). An ontology-based approach provides a natural base for the handling of both problems, it allows abstracting from the concrete surface form of a tag. Also, the possibility to formulate complex relationships between concepts can be used to make contra-intuitive definitions explicit.

Especially, if annotation documentation is generated from such ontological representation, sincere pitfalls of corpus research can be avoided. A widespread strategy to quickly find the tag one is looking for is to search for an appropriate example word and look up its part-of-speech tags in the corpus. For the case of `VAFIN` in STTS, this strategy is particularly treacherous, as the auxiliary use of *haben* and *sein* is not only explicated

by the abbreviation, but it also occurs more frequently than the lexical use. Using this corpus-based strategy of annotation exploration, inclusion of lexical verbs under `VAFIN` will often remain undetected. Using reference definitions to explore annotation schemes helps to avoid such problems.

2 Toward an ontology of linguistic annotations

One appealing solution to the problem of terminological heterogeneity is the standardization approach as employed by the Expert Advisory Group on Language Engineering Standards (EAGLES), an initiative of the European Commission concerned with the development of standards for large-scale language resources. In this context, Leech and Wilson (1996) formulated recommendations for morphosyntactic annotation, further referred to as the "EAGLES meta scheme". In a bottom-up approach, existing tag sets for several European languages have been considered, and commonly used terms and categories have been identified.

This surface-oriented approach, however, faced several problems. First, the outcome of the bottom-up process was merely a list of terms illustrated with examples, but not a fully developed terminological resource with concise definitions. As a consequence, incompatible interpretations of the common terminology occurred among standard-conformant tag sets, contradicting any effort of standardization (Hughes et al., 1995). Further, the standardization approach relies on a direct mapping between concrete tag sets and the meta scheme, that is, every obligatory category in the meta scheme must be implemented by a standard-conformant tag set, and every recommended feature should be implemented. This direct mapping results in a projection of complexity between tag sets and meta scheme. For example, in order to define a standard-conformant tag set for, say, Russian, the tag set needs to provide a tag for articles, which are, however, inexistent in Russian. This problem escalates as the number of languages (a standard is applied to) increases, and in fact, it has been questioned whether universal, or 'obligatory' categories exist at all (Broschart, 1997). Thus, any standardization approach is inherently restricted to a limited set of languages, and is not a general solution for a project also working with data from typological research.

As ontologies provide means for well-defined, structured terminological resources, it seems that these problems can be most easily overcome by the application of an ontology similar to the GOLD approach (Farrar and Langendoen, 2003). Instead of providing a generalization of tag sets for a fixed range of languages, it aimed to cover the full typological variety as far as possible. Finally, it took a different starting point than the EAGLES recommendation due to its orientation towards the documentation of endangered languages. As opposed to this, our joint initiative aims to achieve a unified representation and access to existing resources, which – in their quantitative majority – deal with European languages. Accordingly, we develop an ontology based

on a harmonization between EAGLES, GOLD, and the annotation schemes assembled in section 1.

The ontology is created using a three-step methodology: (i) derive an ontology from EAGLES, (ii) integrate other non-EAGLES conformant tag sets, and finally (iii) harmonize this ontology with GOLD. After an ontology for word classes, resp. part of speech tags, had been completed, this procedure was repeated for morphological features.

The result of this process is the "Reference Model", an ontology of terminology used for linguistic annotations. The basic structure of the Reference Model is derived from EAGLES, but augmented and partly redefined with reference to specific annotation schemes, formalized as "Annotation Models", and the GOLD ontology.

2.1 Building the Reference Model

As an illustration, we consider the special case of nouns. The original definition in the EAGLES recommendations (Leech and Wilson, 1996) is given as:

Nouns (N)					
1. Type:	1. Common	2. Proper			
2. Gender:	1. Masculine	2. Feminine	3. Neuter		
3. Number:	1. Singular	2. Plural			
4. Case:	1. Nominative	2. Genitive	3. Dative	4. Accusative	5. Vocative

Concentrating on the 'Type' feature as a major subclassification among two distinctive parts of speech, we can derive a rudimentary taxonomy of nouns with the concept Noun and two sub-concepts CommonNoun and ProperNoun. The initial, weak ontological representation of the EAGLES meta scheme constructed from such implicitly hierarchical structures is further refined by references to annotation schemes which introduce additional concepts that are usually not assumed for European languages. Examples for such extensions are adverbial participles in Russian, verbal nouns in Cushitic languages, and noun classifiers in Asian languages.

These categories were then aligned with the corresponding categories in the GOLD ontology (Farrar and Langendoen, 2003), which proves especially helpful for the handling of concepts whose interpretation is varying in different tag sets, such as understanding of possessive pronouns which are either regarded as determiners (because of their syntactic function), or as pronouns (because of their semantic function).

For the case of nouns, however, the linking with GOLD introduces another possible perspective on the subclassification of nouns. The concept Noun probably corresponds to Noun_G: "a broad classification of parts of speech which include substantives and nominals". The concept ProperNoun is reserved explicitly for names, and thus covers a sub-class of Substantive_G ("names of physical, concrete, relatively unchanging experiences"). As opposed to this, CommonNoun possibly represents a more general concept than Nominal_G ("whose members differ grammatically from a substantive but which

functions as one"). Especially, CommonNoun covers certain instances of Substantive$_G$ as well. As evident from this example, the GOLD definitions are based on other conceptualizations than those applied in traditional Latin-based grammars underlying most European tag sets. Hence, the Reference Model combines both sub-classifications of nouns.

2.2 Building Annotation Models

The focus of the approach is to integrate existing, heterogeneous terminologies used in existing annotations. In order to achieve sustainability of existing annotations, this also entails the premise to preserve and to systematize the information conveyed in the original annotation documentation.

Therefore, any annotation scheme is formalized within one self-contained ontology, the *Annotation Model*. The Annotation Model is created on the basis of an exhaustive collection of the available annotation documentation. However, besides the information directly formalized in the ontology, the descriptions and a selection of representative examples found in the annotation documentation are preserved and added as comments to concepts and properties in the ontology. For documentation purposes, a hypertext is created from the Annotation Model which conveys both the structure of the Annotation Model and these comments.

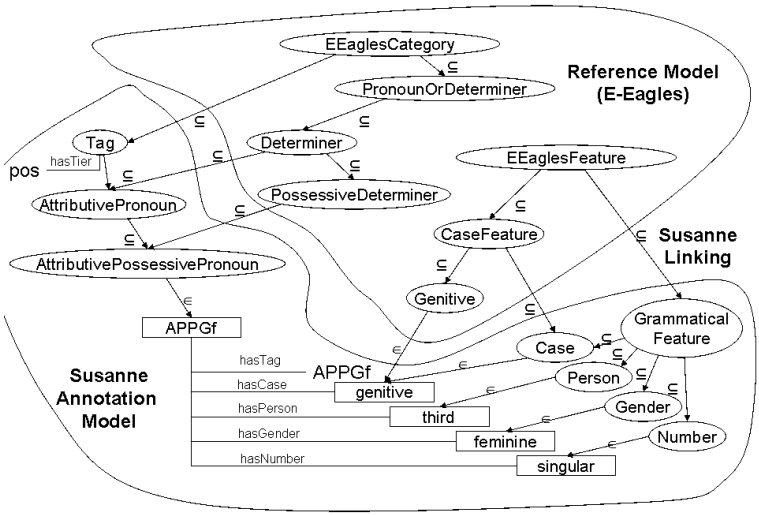
Considering the German tag set STTS as an example, a hierarchically structured Annotation Model can be derived in a similar way as described above. Unlike the EAGLES recommendations, STTS guidelines give detailed enumerations of use-cases, prototypical examples, and critical cases. Further, the aspect of hierarchical structuring is explicitly emphasized. So, the EAGLES-based Reference Model concepts Noun and the sub-concepts CommonNoun and ProperNoun can be aligned easily with the (partial) tags `N` (subsuming `NN` and `NE`), `NN` (concrete and abstract nouns, nominalizations, etc.) and `NE` (surnames, place names, etc.).

The linking between Annotation Models and the Reference Model is implemented by means of conceptual subsumption (`rdfs:subClassOf`), resulting in a complex ontological structure, see 1. An important difference as compared to the standardization approach, the linking does not only allow for underspecification and disjunction, but it also supports formulating complex linking relations with any combination of set operators.

2.3 Integrating morphological features

So far, I concentrated on the construction of a weak ontology of part of speech tags. In a second step, also grammatical features recommended by Leech and Wilson (1996) were integrated into the Reference Model. While word classes are realized as OWL classes

Figure 1: The Susanne tag `APPGf`, its representation in the Annotation Model and (partial) linking with the Reference Model.



in the ontology, grammatical features are encoded as object properties, relating word classes with concepts describing the corresponding grammatical features. Similarly, grammatical information in the corresponding Annotation Models is specified and linked to the Reference Model specifications. The linking between grammatical feature values is modeled by `rdfs:subClassOf`, the linking between object properties is modeled by `rdfs:subPropertyOf`.

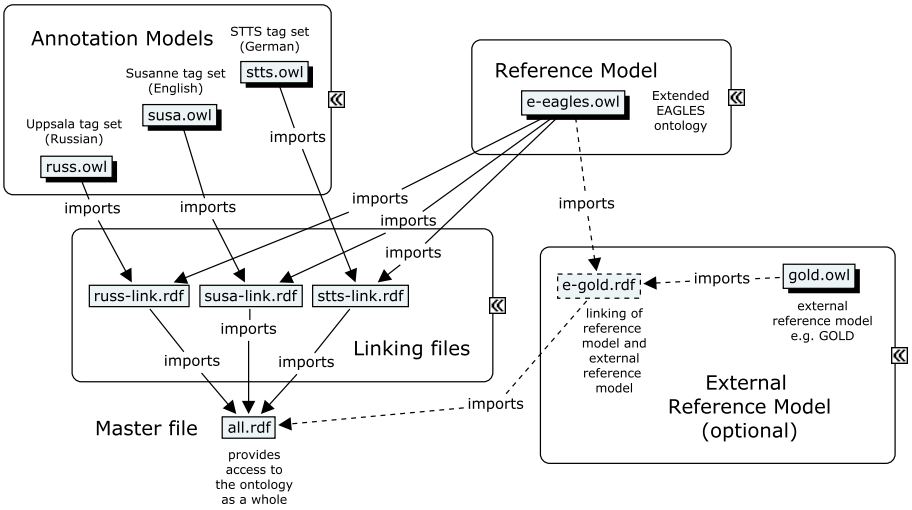
In addition to the formalization of POS tag sets enumerated in table 1, morphological information from the Susanne tag set (English), the Uppsala tag set (Russian), the TIGER annotation scheme (Brants and Hansen, 2002, German) and the CRC632 glossing guidelines are implemented in the corresponding Annotation Models.

For the ontological representation of one example tag from the Susanne tag set, `APPGf`, used for *her* as a possessive, the corresponding inheritance structure of the word class and the case property is presented in fig. 1. Using these inheritance structures, the Susanne tag `APPGf` can be rendered in terms of the Reference Model:

```
PossessiveDeterminer and hasCase(Genitive) and hasPerson(Third) and hasGender(Feminine)
and hasNumber(Singular)
```

The important difference between this description and the (similar) description in terms of the Annotation Model is that this description is tag-set neutral, and does not

Figure 2: The structured ontology.



only apply to the English *her* as a possessive, but also to the corresponding tags in other annotation schemes. The same ontological definition also applies for German *ihr* with the STTS pos tag `PPOSAT` in combination with the morphological description `*.Sg.Fem`, and in the application of the ontology for tag-set neutral corpus querying, this description may be used to retrieve the corresponding tags within different annotation schemes.

3 A structured ontology

As for the technical realization of the ontology, the ontology is broken into multiple OWL files cf. fig. 2) which respectively encode (i) the Reference Model, (ii) several Annotation Model, and (iii) the linking between a Reference Model and each particular Annotation Models.

The components as well as the ontology as a whole are defined in OWL/DL, thus enabling the processing with OWL/DL reasoners.

Reference Model The Reference Model represents the 'terminological backbone' of the structured ontology. As the skeleton of the Reference Model is originally derived from the EAGLES meta scheme as described above, it is associated with the name space `e-eagles`, i.e. extended EAGLES.

Annotation Model Annotation Models represent self-contained ontologies covering the documentation available about a particular annotation scheme. POS tags are modeled as instances, with every tag corresponding to one single instance. The surface form of this instance is defined by means of the property `hasTag`.[1]

Further, different annotation schemes employ different classifications of levels of annotation. Morphological information can be annotated on an independent annotation layer `morph`, it may be integrated with POS annotation, or together with semantic annotations on the annotation layer `gloss`. Thus, property `hasTier` specifies the name of the annotation layer where the corresponding annotation is to be found in accordance with the annotation guidelines.

Again, OWL name spaces are introduced to separate different Annotation Models (`stts, susa, russ, ...`) and Reference Model (`e-eagles`).

Linking Annotation Model and Reference Model Reference Model and Annotation Model are independent ontologies of linguistic terms. Thus, the linking between them has to be made explicit. For this purpose, we apply separate owl files which import Reference Model and Annotation Model. For every Annotation Model, say `stts.owl`, a corresponding link file `stts-link.rdf` exists. In this link file, the relationship between the STTS Annotation Model concepts and Reference Model concepts is represented in a declarative way, by means of `rdfs:Descriptions` pertaining `rdfs:subClassOf`-statements.

As both Annotation Model and Reference Model can have independent hierarchical structure, it is not necessary to assign every single tag to a concept of the Reference Model by its own. Rather, explicit references between Annotation Model concepts and Reference Model concepts are possible, thus making instances of Annotation Model concepts indirect instances of Reference Model concepts.

Linking Reference Model and external Reference Model The same mechanism as applied for the linking between Annotation Model and Reference Model may be used to relate Reference Model concepts with external ontological resources. A possible external reference model is GOLD, resp. its modified variant, with which the current Reference Model is linked to. This differentiation allows a user to differentiate between the modeling of linguistic terminologies in general (or by a specific community, that is the primary function of the external reference model) and the formalization and generalization over specific annotation guidelines. Only the latter is the primary function of the (internal)

[1] For more complex tag sets, which involve also information about morphology (such as the Uppsala tag set for Russian with 877 known tags) or semantic classes (such as the Susanne scheme for English with 420 tags), however, it is reasonable not to require a 1-to-1 mapping between instances and tags, but to rather assemble multiple tags under one instance. Thus, the property `hasTag` can be replaced by `hasTagStartingWith`, `hasTagEndingWith`, or `hasTagContaining`.

Reference Model. However, by specifying a linking between the internal Reference Model and some external reference model, the external reference model is indirectly related to the Annotation Models as well. The internal Reference Model thus serves to mediate between Annotation Models and external reference models. In this sense, the internal Reference Model provides an *interface* to the Annotation Models it is associated with.

The master file Finally, one additional file is needed which represents an interface to the ontology as a whole. Basically, this is an OWL file importing all relevant linking files (these are importing Reference Model and Annotation Models). When loading this master file, the whole ontology with all the parts becomes available to the importing program.

4 Application and evaluation

4.1 Fields of application

At the moment, the ontology focuses on the annotation mining perspective, with an application to ontology-based corpus querying and annotation documentation.

For this purpose, we have developed a problem-specific HTML visualization,[2] which enables a user to browse the ontology, in order to find out definitions of tags and concepts within an Annotation Model and their relationship to the Reference Model. As the ontology contains the comments from the original annotation documentation, it is not to be misunderstood as an ontology of linguistic terminology *in general*, as the ontologies developed by Schneider (2007) and Farrar and Langendoen (2003). Rather, the ontologies described here only concern the documentation of existing annotations, without making any claims about the use of terms beyond this. Still, it would be a great achievement to relate these or similar approaches to each other, thus directly relating terms discussed in grammatical theory with concrete linguistic annotations.

Moreover, the OntoClient was implemented, a JAVA-based pre-processor for corpus queries, which supports annotation-independent search queries by using concepts and definitions in the Reference Model. In essence, it is a specialized OWL reasoner, which translates ontological descriptions of concepts and properties into a disjunction of instances from which, then, the form of tag and the annotation can be retrieved using the `hasTag` and `hasTier` properties. Using the OntoClient, the Reference Model definitions can be applied for the formulation of tagset-neutral corpus queries, cf. Rehm et al. (2007).

[2] `http://nachhalt.sfb632.uni-potsdam.de/OntoBrowser`

4.2 Application in NLP contexts

In addition to this kind of technical application, the ontologies can be used for semantic interpretation of linguistic annotations independently of the underlying tag set. One domain where technical applications can benefit from an ontological interpretation of linguistic annotations is their natural handling of underspecification. More precisely, the accuracy and robustness of tools like taggers or parsers can be improved by the application of tool-specific ontologies.

As an example, consider the tagging of the substitutive demonstrative pronouns *der, die, das* in German. These are homonymous with the definite article and the relative pronoun, and thus, for correct identification of these pronouns, a (partial) syntactic analysis is needed. Schmid (1994)'s TreeTagger achieved a precision of 89.2% and a recall of 92.4% for the corresponding STTS tag `PDS` on the morphosyntactically analyzed Potsdam Commentary Corpus (Stede, 2004). `PDS` was misleadingly chosen for manually annotated `PDAT` (attributive demonstrative pronoun) in 5.0% of the cases and for `PRELS` (substitutive relative pronoun) in 5.8% of the cases. In terms of the ontology, this could be expressed by assigning the tag not to one ontological concept, but rather to a disjunction of the ontological concepts.[3] In this way, tool-specific underspecified ontologies for annotation schemes can be derived from an Annotation Model using a manually annotated reference corpus and the output of the corresponding tool.

On ontological, tag-set independent representation also allows to combine information from different linguistic tools such as another tagger. Considering two currently used POS tag sets for German, STTS and the Morphy tag set (Lezius et al., 1998), we find differentiations in both tag sets that are absent in the other. As such, Morphy distinguishes definite and indefinite articles, both tagged as `ART` in STTS. On the basis of an ontological representation, however, both analyses can be represented not only in parallel, but also as a conjunction, and, for this case, they can also be simplified.

$$\texttt{DefiniteArticle} \cap \texttt{Article} = \texttt{DefiniteArticle}$$

In a similar way, it is possible to enrich linguistic analyses with semantic analyses and vice versa, e.g. in the resolution of underspecification at both levels, as suggested by Cimiano and Reyle (2003). Following de Cea et al. (2004), such dependencies can benefit from the use of ontologies as a common elementary representation for both linguistic and semantic features within a text.

[3] It seems reasonable not to require the ontological interpretation of the tagger to cover any possible exception but only systematic errors. By demanding a minimal precision of 95% of the output, then, the ontological representation could be defined as the disjunction of the most frequent concepts, i.e. `PDS` and `PRELS`. A possible underspecified ontological interpretation of this disjunction was `SubstitutivePronoun`. As opposed to the original tagger output, a minimal precision of 95% percent is guaranteed for this interpretation.

The ontology presented above represents an elementary component for such a hybrid system, in particular with reference to the Annotation Models for German, English and Russian, which are employed by automatic tools.

4.3 Evaluation

The ontologies developed so far are comparably small[4] and are based only on annotation documentation, i.e. a limited selection of documents, as their source.

The hierarchical structure in the Reference Model and the Annotation Models follows from the hierarchical structure reflected in the annotation documentation resp. in the EAGLES recommendations, that is, usually one single document, for which an ontology construction procedure has been described above. For reasons of size and great homogeneity of the textual base of the ontology, an evaluation of *structural* properties of the ontologies, such as detection of cycles, seems unnecessary.

The linking was developed in co-operation with specialists for the corresponding domain and literature on the language under consideration. For non-European languages, thus, any expert knowledge that was available was dedicated to the refinement and precision of the linking, rather than its evaluation. For better-known European languages, the linking was adapted from the EAGLES recommendations, and only modified were more precise definitions of terms were provided.

As for the qualitative evaluation of the Reference Model, the implementation of several linkings with external Reference Models revealed that the conceptualizations and the definitions adopted in the Reference Model are compatible with these external Reference Models, confirming its validity. In particular, the morphosyntactic module of the OntoTag ontologies (de Cea et al., 2004) and the Data Category Registry (Ide et al., 2005) are important in this respect, as these had not been consulted during the design of the Reference Model. The morphosyntactic module of the OntoTag ontologies was developed on the basis of the EAGLES recommendations, but specifically for Spanish. The ontology differs from the Reference Model, in that the following characteristics were specified: `exhaustive`, `disjoint`, `partition` and `partOf`. These were excluded from the Reference Model in order to guarantee applicability to languages which require introduction of additional concepts. Yet, the concepts identified, the hierarchical structure and the grammatical features could be mapped onto each other, most exceptions being extensions of the OntoTag ontologies specific to Spanish.

As for the linking with the Data Category Registry categories, an OWL representation of the data categories specified by Monachini et al. (2005) was developed. The linking between this DCR ontology and the Reference Model could be established only on

[4]The Reference Model consists of 18 object properties (grammatical features), 161 classes (word classes, grammatical categories), and has a maximum inheritance depth of 5. The Uppsala Annotation Model consists of 18 object properties (grammatical features) and 79 classes (word classes and grammatical categories) with a maximum inheritance depth of 4. Further, it contains 906 instances (tags and values of grammatical features).

the basis of similarity of concept names, as Monachini et al. (2005) did not provide definitions. For word classes, however, 85.71% of top-level concepts could be linked to morphological feature types in DCR, indicating that with the exception of specifics of the typologically-oriented annotation schemes considered, the Reference Model formalizes a sub-set of DCR categories.

In this sense, the validity of the Reference Model with respect to two external knowledge sources has been shown. The high level of agreement between these is most likely due to the influence of the EAGLES recommendations that played a crucial role in the design of the Reference Model as well as in the design of the OntoTag ontologies and the DCR. More interesting, however, are the differences, which reflect different orientations of the ontologies. Those concepts that were missing in the Reference Model were either language-specific (OntoTag) or were not considered in either EAGLES or in the annotation schemes relevant to the sustainability project. The Reference Model concepts that did not find a counterpart in OntoTag or the DCR mostly originated in typologically-oriented annotation schemes, annotation schemes used by historical linguists, or the Russian Uppsala tag set, indicating that OntoTag and the DCR seem to have a stronger focus on Western European languages, or more generally, languages for which a broader range of linguistic tools exists.

5 Summary and discussion

In this paper, I have described design principles and implementation of a structured ontology of linguistic annotation. It is currently applied for purposes of annotation documentation, tag-set neutral corpus search and can also be applied in NLP contexts.

For the ontology, sustainability considerations entail the premise to preserve and to systemize existing annotations and relevant annotation documentation. In line with this conservation perspective, a structured ontology was developed which involves several self-contained ontologies, which are linked in a declarative way. Hence, a clear separation between the information drawn from the annotation documentation and its interpretation with respect to the reference terminology is established, as required by the ethics of conservation:

> The principal goal should be the stabilisation of the object or specimen. All conservation procedures should be documented and as reversible as possible, and all alterations should be clearly distinguishable from the original object or specimen. (ICOM, 2006, §2.24).

The structured ontology consists of a Reference Model specifying conventional linguistic terminology, and several Annotation Models, each representing a formalization of the annotation documentation of a given annotation scheme. Both Reference Model and the respective Annotation Models are self-contained ontologies. Between these,

however, a linking is specified which describes any Annotation Model concept in terms of the Reference Model.

As compared to related approaches, which operate on the direct mapping of annotations to an ontology of reference terms, e.g. Farrar and Langendoen (2003), de Cea et al. (2004), this structured ontology involves a high level of redundancy. The modular representation of Reference Model and Annotation Model, however, allows to view Annotation Models as a form of annotation documentation, as annotation-relevant comments are clearly separated from interpretation-relevant comments. In particular, these annotation-relevant comments are supposed to cover excerpts and examples from the original documentation which provide an informal, non-ontological definition and description of the respective concepts and properties. Also, a hypertext visualization of Annotation Models, the Reference Model and the linking has been implemented which allows a user to assess both the ontological information and these comments and thus, use the ontology as a key to annotation documentation.

Further, this modular structure is highly flexible, as it allows a user to replace any component of the system by his own specifications, that is, the linking may be altered independently from the participating Reference and Annotation Models. Similarly, an Annotation Model may be exchanged. Further, this design supports an open, extensible architecture, that is, new Annotation Models can be developed and linked to the Reference Model. Finally, a non-redundant ontological representation can be automatically retrieved from the structured ontology by unifying concepts from the Reference Model with the Annotation Model concepts that are defined as sub-concepts in the linking.

The Reference Model itself may be linked by the same mechanism to external Reference Models of linguistic terminology in general. Such external Reference Models may evolve from approaches like Farrar and Langendoen (2003) or Schneider (2007). These external Reference Models, then, must not be related to any existing Annotation Model, but instead, the linking with the Annotation Models is mediated by the (internal) Reference Model.

So far, three external Reference Models have been linked to the internal Reference Model, i.e. GOLD, the morphosyntactic component of de Cea et al. (2004)'s OntoTag ontologies, and an ontological representation of Ide et al. (2005)'s Data Category Registry. This is particularly interesting for the application of ontologies to the formulation of annotation-independent corpus queries, i.e. expressions formulated in terms of external Reference Models can be translated into queries for specific annotations on the basis of the linkings with the internal Reference Model and the Annotation Models. The internal Reference Model thus represents an *interface* to the Annotation Models, and the annotations.

Currently, Annotation Models pertaining parts of speech and morphosyntax for

German, English and Russian have been implemented. Also, Annotation Models for a typologically oriented annotation scheme has been developed, that applies not only to parts of speech and morphosyntactic annotation, but also to glossing, syntactic phrases and information structure in a broad variety of languages. Finally, several project-specific Annotation Models relevant to the CRCs (concerning historic linguistics, typological research and first language acquisition) have been created.

From these, the Annotation Models pertaining German, Russian and English are particularly relevant to text technology, as these tag sets are also used by existing tools and thus, these ontologies can be used to support the tag-set independent interpretation of automatically derived linguistic analyses. More precisely, the ontology-based approach presents a natural handling of underspecification, and by exploiting this information, the robustness of linguistic analyses in technical contexts may be improved.

References

Brants, S. and Hansen, S. (2002). Developments in the TIGER annotation scheme and their realization in the corpus. In *Proc. 3rd Conference on Language Resources and Evaluation (LREC-02)*, Las Palmas de Gran Canaria, Spain.

Broschart, J. (1997). Why Tongan does it differently: Categorial distinctions in a language without nouns and verbs. *Linguistic Typology*, 1-2:123–166.

Chiarcos, C. (2006). An ontology for heterogeneous data collections. In *Proc. Corpus Linguistics 2006*, pages 373–380, St.-Petersburg. St.-Petersburg University Press.

Cimiano, P. and Reyle, U. (2003). Ontology-based semantic construction, underspecification and disambiguation. In *Proc. Prospects and Advances in the Syntax-Semantic Interface Workshop*.

de Cea, G. A., Gómez-Pérez, A., Álvarez de Mon, I., and Pareja-Lora, A. (2004). OntoTag's linguistic ontologies. In *Proc. Int'l Conference on Information Technology, Coding and Computing (ITCC'04)*, pages 124–128, Las Vegas, Nevada.

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

Hughes, J., Souter, C., and Atwell, E. (1995). Automatic extraction of tag set mappings from parallel annotated corpora. In *From Text to Tags: Issues in Multilingual Language Analysis, Proc. ACL-SIGDAT Workshop*, pages 10–17.

ICOM (2006). ICOM code of ethics for museums. In Hoffman, B. T., editor, *Art and Cultural Heritage. Law, Policy and Practice*. Cambridge University Press.

Ide, N., Romary, L., and de la Clergeri, E. (2005). International standard for a linguistic annotation framework. In *Proc. HLT-NAACL'03 Workshop Software Engineering and Architecture of Language Technology*.

König, E., Bakker, D., Dahl, e., Haspelmath, M., Koptjevskaja-Tamm, M., Lehmann, C., and Siewierska, A. (1993). EUROTYP Guidelines. Technical report, European Science Foundation Programme in Language Typology.

Leech, G. and Wilson, A. (1996). EAGLES recommendations for the morphosyntactic annotation of corpora. Technical report, Expert Advisory Group on Language Engineering Standards.

Lezius, W., Rapp, R., and Wettler, M. (1998). A freely available morphological analyzer, disambiguator, and context sensitive lemmatizer for German. In *Proc. COLING-ACL 1998*, pages 743–747.

Monachini, M., Soria, C., and Ulivieri, M. (2005). Evaluation of existing standards for NLP lexica. draft 1.1. Technical report, LIRICS (Linguistic Infrastructure for Interoperable Resource and Systems).

Rehm, G., Eckart, R., and Chiarcos, C. (2007). An OWL- and XQuery-based mechanism for the retrieval of linguistic patterns from XML-corpora. In *Proc. RANLP 2007: Recent Advances in Natural Language Processing*. Borovets, Bulgaria.

Sampson, G. (1995). *English for the Computer*. Clarendon Press, Oxford.

Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines fur das Tagging deutscher Textkorpora mit STTS. Technical report, University of Stuttgart and Universitat of Tübingen.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester,UK.

Schneider, R. (2007). A database-driven ontology for German grammar. In Rehm, G., Witt, A., and Lemnitzer, L., editors, *Data Structures for Linguistic Resources and Applications*, pages 305–314, Tübingen. Narr.

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proc. ACL-04 Workshop on Discourse Annotation*, pages 96–102, Barcelona.