Linguistic Ontologies

Interface Model and Domain Model for Russian word classes and morphology

Interface Model
E(xtended)-EAGLES

Domain Model
Uppsala tagset

E-EaglesFeatures
Tense
Degree
Aspect
Definiteness
Gender
Number
Case
Inflection
Voice
Person
Nominative
Dative
Oblique
Vocative
Genitive
E-EaglesCategory
Classifier
Adposition
Circumposition
Postposition
Preposition
Unique
PronounOrDeterminer
Determiner
Article
FusedPrepArt
DefiniteArticle
IndefiniteArticle
PartitiveArticle
PartitiveDeterminer
WHDeterminer
InterrogativeDeterminer
ExclamatoryDeterminer
RelativeDeterminer
IndefiniteDeterminer
PossessiveDeterminer
DemonstrativeDeterminer
Quantifier
Numeral
CardinalNumber
OrdinalNumber
Interjection
Adverb
PronominalAdverb
DegreeAdverb
WHTypeAdverbs
ExclamatoryAdverb
RelativeAdverb
InterrogativeAdverb
Verb
AuxiliaryVerb
StrictAuxiliaryVerb
ModalVerb
Copula
FiniteVerb
SubjunctiveVerb
ConditionalVerb
IndicativeVerb
ImperativeVerb
NonFiniteVerb
Supine
Gerund
Infinitive
ing
Participle
VerbalNoun
Noun
ProperNoun
MassNoun
CountableNoun
CommonNoun
Adjective
AttributiveAdjective
PredicativeAdjective
Conjunction
CoordinatingConjunction
SubordinatingConjunction
Residual
Symbol
Abbreviation
Acronym
Foreign
Formula
Punctuation
SentenceMedial Punctuation
LeftParenthetical Punctuation
RightParenthetical Punctuation
SentenceFinal Punctuation
Pronoun
WHPronoun
ExclamatoryPronoun
InterrogativePronoun
RelativePronoun
PersReflPronoun
ReflexivePronoun
PersonalPronoun
DemonstrativePronoun
PossessivePronoun
IndefinitePronoun
ReciprocalPronoun
russ-Feature
Degree
Voice
Tense
Aspect
Mood
Person
Animacy
animate
inanimate
hasAnimacy
Case
dative
hasCase
Number
singular
plural
hasNumber
Gender
masculine
feminine
neuter
hasGender
russ-Tag
Abbreviation
Other
DateDigit
KompDisj
Punctuation
Conjunction
Disjunctive Conjunction
Coordinative Conjunction
Subordinative Conjunction
Preposition
Particle
Interjection
Predicative
Adjective
PossessiveAdjective
Noun
substantiv_a_pl_gen_bel
substantiv_a_sg_gen_bel
substantiv_fem_pl_akk_bel
substantiv_fem_pl_akk_unb
substantiv_fem_pl_dat_bel
substantiv_fem_pl_dat_unb
...
Name
name_fem_sg_akk_bel
name_fem_sg_akk_unb
name_fem_sg_dat_bel
name_fem_sg_dat_unb
...
Adverb
ComplexAdverb
NonderivedAdverb
DeadjectivalAdverb
Numeral
CardinalNumeral
OrdinalNumeral
CollectiveCardinal Numeral
IndefiniteCardinal Numeral
DemonstrativeCardinal Numeral
DigitalCardinalNumeral
AdverbialParticiple
Verb
Participle
Infinitive
ImperativeVerb
FiniteVerb
FiniteVerbPast
FiniteVerbPresent
Pronoun
DemonstrativePronoun
SubstitutiveDemonstrative Pronoun
AttributiveDemonstrative Pronoun
ReciprocalPronoun
PersonalPronoun
RelativePronoun
SubstitutiveRelative Pronoun
AttributiveRelative Pronoun
ReflexivePronoun
ReflexivePossessive Pronoun
PossessivePronoun
RelativePossessive Pronoun
IndefinitePronoun
IndefinitePossessive Pronoun
AttributiveIndefinite Pronoun
SubstitutiveIndefinite Pronoun
InterrogativePronoun
SubstitutiveInterrogative Pronoun
AttributiveInterrogative Pronoun
InterrogativePossessive Pronoun
GeneralPronoun
SubstitutiveGeneral Pronoun
AttributiveGeneral Pronoun
NegativePronoun
AttributiveNegative Pronoun
SubstitutiveNegative Pronoun

An ontology-based approach towards terminological integration

Crucial problems hindering exchange and re-usability of linguistic resources are differences in theoretical assumption guiding the design of annotation schemes, the annotation process itself and the interpretation of linguistic annotations. Here, a fragment of an ontology-based proposal to the terminological integration and access to terminologically heterogeneous resources is illustrated for the case of the Uppsala tag set, an annotation scheme for parts of speech and morphology for Russian.
The Uppsala tag set is formalized as an ontology, the so-called domain model. Conceptual classes (word classes) and grammatical features (morphological information) are rendered as ontological classes and properties. The tags specified by the tag set are instances of these classes. As an example, the tag substantiv_fem_pl_dat_unb is analysed with greater level of detail.

Similar to this application-specific ontology, other domain models for different annotation schemes, different languages and different purposes can be formalized. As a generalization over these task-, language- or project-specific domain models, an interface model has been developed which represents a general terminological reference. In the interface model, a minimal inventory of linguistic terminology applied in different annotation projects is formalized. It thus serves as a common terminological reference point.

This reference function is formalized by the linking between interface models and domain model. More concrete, any domain model class, and any property are specified as sub-classes and sub-properties of domain model classes and domain model properties.

Accordingly, tags in different domain models can be retrieved from the same expression over interface model concepts. Classes and properties of the interface model thus present a **unified access to terminologically heterogeneous resources**.

This basic architecture is formalized in separate OWL files, that is, interface model, domain models and the linking between are stored in individual files.

The application of multiple, linked ontologies leads to a flexible, modular, and thus scalable structure which allows the integration of additional ontological resources (e.g. GOLD, OntoTag, DOLCE, SUMO) and the adaption towards specific needs of a user.

Background and Related Research

The ontology-based approach towards the terminological integration arose in the context of the project “Sustainability of linguistic resources”, a collaborative project of three German research centres, SFB 441 “Linguistic Data Structures”, Tübingen, SFB 538 “Multilingualism”, Hamburg, and SFB 632 “Information Structure”, Potsdam/Berlin (Dipper et al. 2006). The projects aims to build a sustainability archive for the resources of the participating research centres, and to develop general solutions for sustainable data creation, maintenance, and usability. The development of an ontology in order to integrate the terminology applied in different annotation standards guarantees a consistent documentation of annotations, but moreover, it is an important contribution to the development of a tagset-neutral and theory-independent querying interface which allows the access to resources even to people not familiar with certain tag sets applied. Accordingly, ontology-based querying reduces the bias for non-specialists to access the assembled data.

Attempts towards the terminological integration of linguistic resources, however, have been made since the 1990s. As such, the EAGLES recommendations for morpho-syntax (Leech and Wilson 1996) aimed to overcome the terminological heterogeneity of different tag sets applied to European languages by the introduction of a minimal standard inventory of linguistic categories and features. This standardization approach, as his successors (Erjavec 2004, Meyers 2006), has crucial disadvantages. As long as it operates by means of an equation of terminological concepts, i.e. a 1-to-1-mapping between standard and standard-conformant tag sets, it will result in a projection of complexity both from the tag sets to the standard (the standard has to include specifics of a certain family of languages in order to express their commonalities), and from the standard to the tag sets (a standard-conformant tag set has to implement the standard, regardless of the grammar of the language).

As opposed to this, the application of ontologies allows for a more detaillled and scalable management of linguistic terminology. An ontology is a formalization of conceptional structures by means of hierarchically related concepts and relations (properties) between them. Whereas the EAGLES recommendations resulted in a plain list of terms, an ontology allows for the concrete specification of definitions. Furthermore, an ontology-based reference differs from a standardization approach in that is does not dictate a certain degree of granularity a standard-conformant tag set has to represent. Instead, it supports a natural handling of underspecification. Thus, as compared to a standardization solution, an ontology-based approach enhances both explicitness and scalability.

The interface model developed here is partly a formalization of the EAGLES recommendations (cf. de Cea et al. 2004 for a similar approach). However, it is intended to cover non-European languages (Tibetan, Haussa, Austronesian, etc.) as well, and thus, another important inspiration was the General Ontology of Linguistic Description (GOLD) developed by the E-MELD project (Farrar and Langendoen 2003). However, the interface model developed here does not attempt to replace these existing solutions. Instead, the flexible architecture allows to integrate these or other ontological resources into the system by specifying a linking between interface model and such external reference ontologies. Thus, the main benefit of the interface model is not that it represents a formalization of linguistic terminology in general, but that it provides a mechanism for the unified handling of its specific domain models, both for querying and the linking with external upper models.

Literature and References

Chiarcos, Ch. (2006a), *An Ontology for Heterogeneous Data Collections*, to appear in Proceedings of Ontologies in Text Technology (OTT'06)
Chiarcos, Ch. (2006b), An Ontology for Heterogeneous Data Collections, in Proc. Corpus Linguistics 2006, October 10–14, 2006, St.-Petersburg, St.-Petersburg University Press, 373-380.
de Cea, Guadalupe Aguado, Asunción Gómez-Pérez, Inmaculada Álvarez de Mon, Antonio Pareja-Lora (2004): OntoTag's Linguistic Ontologies: Improving Semantic Web Annotations for a Better Language Understanding in Machines. Proc. *International Conference on Information Technology: Coding and Computing* (ITCC'04), vol. 2, 04 05 - 04, 2004, Las Vegas, Nevada, p. 124-128
Dipper, St., E. Hinrichs, Th. Schmidt, A. Wagner, and A. Witt (2006): Sustainability of linguistic resources. In: Proceedings of the LREC Workshop on merging and layering linguistic information. Genoa.
Erjavec, Tomaž (2004): *MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora*. In: Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, ELRA, Paris.
Farrar, S. and D. T. Langendoen: A Linguistic Ontology for the Semantic Web. GLOT International (3) (2003), 97–100
Götze, M., Dipper, S. (2006): ANNIS – Complex Multilevel Annotations in a Linguistic Database. In: Proceedings of the EACL Workshop on Multi-dimensional Markup in Natural Language Processing NLPXML-2006. Triento, Italy (2006) 61–64
Ide, Nancy, Laurent Romary, Eric de la Clergerie (2004), International Standard for a Linguistic Annotation Framework, *Natural Language Engineering*, 10: 211-225.
Leech, G. and Wilson, A.: EAGLES Recommendations for the Morphosyntactic Annotation of Corpora, Version of Mar, 1996, http://www.ilc.cnr.it/EAGLES/annotate/annotate.html.
Meyers, A (ed.) (2006), Annotation Compatibility Working Group Report. Proc. of the ACL/Coling '06 Workshop "WS6: Frontiers in Linguistically Annotated Corpora 2006„
Schmidt, Th., Ch. Chiarcos, T. Lehmberg, G. Rehm, A. Witt and E. Hinrichs (2006): Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources, paper presented at the 6th E-MELD workshop, Ypsilanti (2006)