ОНТОЛОГИЯ ЛИНГВИСТИЧЕСКОЙ АННОТАЦИИ: КЛАССЫ СЛОВОФОРМ И МОРФОЛОГИЯ

AN ONTOLOGY OF LINGUISTIC ANNOTATION:WORD CLASSES AND MORPHOLOGY[1]

Christian Chiarcos *(Chiarcos@uni-potsdam.de)*, *University of Potsdam*

В данной статье описывается концепциональная и техническая структура онтологии лингвистической терминологии. Поскольку онтология содержит связи с существующими аннотационными схемами для нескольких языков, она может быть использована для формулировки независимых от языка запросов к корпусам. Помимо технической значимости, онтология распологает стандартизированным репертуаром формальной спецификации аннотационных схем в целом. Благодаря модулярной архитектуре, интеграция дальнейших аннотационных схем возможна с затратой минимума усилий, что влечет за собой дополнительную область применения: создание портативных, независимых от аннотационной разметки средств по обработке языка. В первую очередь, онтология предназначена для обобщенного представления и доступа к терминологически гетерогенным ресурсам. Она будет применяться как часть архива лингвистических ресурсов в рамках проекта «Сохранение лингвистического материала для долгосрочного доступа и применения», который является совместной инициативой трех немецких исследовательских центров. Проект располагает богатым корпусным материалом, включая как хорошо документированные языки, как, например, немецкий, английский, русский, так и несколько африканских языков, исторические корпуса и т.д.

1. Motivation

For researchers unfamiliar with the specific usage and origins of terms that have been applied in the creation of a data source such as a corpus, the variety of abbreviations, terms, tags and possibly conflicting definitions can be confusing and time-consuming.

In a worst case scenario, the effort necessary for a closer examination of the data will prevent later generations of researchers from working with a data collection. The problem becomes even more apparent for very large collections of heterogeneous corpora. That is why it is an urgent task for the unified treatment of such collections to identify and to document commonalities as well as differences in the terminology used: the integration of information on the linguistic terminology can be seen as a core aspect of sustainable maintenance of linguistic data.

As an example, the developers of the ACT Old-Church Slavonic database (cf. Ribarov 2004) complain that "ACT database data are manuscript data annotated using older systems and older annotation strategies. Therefore the POS tags of the lemmas are not uniform and unfortunately not each word form is assigned morphological information. The morphological tag currently present is the very basic one."[2] To overcome these problems it is necessary to provide a consistent terminology and to refer to this terminological backbone in the definition of annotation structures.

In this paper, a general ontology-based framework is presented as a possible solution to querying heterogeneously annotated resources. Especially, our ontology-based approach is intended to provide a unified access to heterogeneously annotated resources in a way that is tag set neutral, tag set independent, theory neutral, and scalable.

2. Research Background

The framework presented here is developed in the context of the project "Sustainability of Linguistic Data", a collaborative initiative formed by three collaborative research centers, CRC 441 (Tübingen, "Linguistic Data Structures"), CRC 538 (Hamburg, "Multilingualism") and CRC 632 (Potsdam/Berlin, "Information Structure") to provide means to guarantee the long-time availability and accessibility of the collected resources. The project is intended to develop sus-

[1] Christian Chiarcos' work is part of research in the Collaborative Research Center 441 ("Linguistic Data Structures"), Tübingen, project C2 "Sustainability of Linguistic Data", funded by the German Research Foundation (DFG).

[2] ACT project, http://prometheus.ms.mff.cuni.cz/act/www/ (06/10/14).

tainable solutions for creation, maintenance, accessibility and distribution of linguistic resources (Schmidt et al. 2006).

One of our primary aims is to provide the means to ensure the long-term availability of the data collections. On one hand this involves technical aspects (Dipper et al. 2006), the collection of documentation and meta data concerning the resources, and the clarification of legal issues (Lehmberg et al. 2007). On the other hand, a substantial degree of terminological integration is necessary in order to provide non-specialised users with an easy access to the extremely heterogeneous resources hosted by our project. This involves two aspects, that is, elements of annotation must refer to a common terminological resource whose specifics are well-documented and covering the major aspects of the underlying annotation schemes, and in addition to this, it should be possible to search for annotations by using a set of standardised terms rather than cryptic abbreviations which are specific to a single tag set. In our approach, an ontology is employed to cover both tasks.

In earlier publications, we concentrated on motivations and background of an ontology on part of speech tags (Chiarcos 2006a) and aspects of its application in corpus search (Chiarcos 2006b). Here, the representation of morphological features is considered with greater level of detail and the advantages of the modular structure of the ontology are highlighted.

3. Building an Ontology of Linguistic Annotation

A classical solution to the problem is the "standardisation approach" as employed by the EAGLES project (Leech and Wilson 1996). There, recommendations for morphosyntactic annotation sets have been formulated. In a bottom-up approach, existing tag sets for several European languages have been considered, and commonly used terms and categories have been identified. As a result, 13 obligatory categories were postulated. For each category, a list of features has been assembled that a standard-conformant tag set should respect. Accordingly, the "EAGLES meta tag set" is constituted as the set of reasonable combinations of categories (main tags) and features. The methodology applied in the EAGLES project was later extended to Eastern European languages in the MULTEXT-East project (Erjavec 2004) and recently recovered by the Frontiers in Linguistically Annotated Corpora working group (Meyers 2006).

The standardisation approach faces several disadvantages, the most important being that language-specific conceptualisations have to be integrated into the meta-scheme. As a consequence, the complexity of every standard-conformant scheme is projected onto the meta-scheme. Further, the outcome of the bottom-up process in the case of EAGLES was not a full terminological resource, but only a list of terms. As long as no definitions are included in the description of the standard, community-specific usage of terms can lead to contradictory interpretations of the corresponding tags. This certainly contradicts any effort of standardisation.

Another strand of terminological integration comes from more theoretically oriented approaches in typology and language documentation, with GOLD, the General Ontology for Linguistic Description, as its currently most prominent instantiation (Farrar and Langendoen 2003). In contrast to the EAGLES initiative, which was dedicated to European languages exclusively, in the E-MELD project GOLD aspects of universality and scalability were emphasized from the very beginning. Instead of providing a generalisation 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 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 developed an ontology based on established meta-schemes such as EAGLES and MULTEXT-East. For standard-conformant tag sets, then, the linking with this ontology becomes trivial. Still, as these meta-schemes suffer from the problems of standardisation approaches in general, we further work on a harmonisation between our EAGLES-based ontology and GOLD. Accordingly, the terms used in EAGLES are provided with a formal definition retrievable from the mapping between EAGLES and GOLD. In addition to this, conceptions from other non-EAGLES conformant tag sets are integrated into the ontology. These additions, together with the need to overcome inherent incompatibilities between different conceptualization in GOLD and EAGLES (e.g. the sub-classification of nouns, the differentiation between pronoun, quantifier and numeral, or the differentiation between pronoun and determiner), lead to a reformulation of the original EAGLES-based ontology and suggestions for the modification of the current version of the GOLD ontology which we refer to as E(xtended)-EAGLES and E(xtended)-GOLD.

Thus, our terminological backbone was created in a three-step methodology:

1. derive an ontology from EAGLES,
2. integrate other non-EAGLES conformant tag sets, and finally
3. harmonise this ontology with GOLD.

The result of this procedure, the so-called E(xtended)-EAGLES ontology, is described elsewhere with greater level of detail (Chiarcos 2006a).

4. A structured ontology

Our ontology consists of three major components, i.e.

- a number of *domain models* which are ontologies that each represent one annotation scheme or tag set,
- the *interface model*, i.e. the E-EAGLES ontology, which includes reference definitions and thus serves as a terminological backbone by reference to which domain model concepts are defined in a standardised manner, and
- the *linking* between a domain model and the interface model which is specified apart from both models.

This tripartite structure can be augmented by the optional linking of the interface model with additional *upper models*. As a result, these upper models can be applied for the formulation of search queries as an alternative to the reference terminology specified in the interface model. Reference definitions retrievable from upper models to domain models are thus mediated by the interface model.[3]

We claim that this modular approach is more flexible as it allows alternative specifications of linking and the inclusion of alternative upper models as well as additional domain models. In present-day annotation technology, it finds a close pendant in the *standoff paradigm* according to which different levels of annotation and the primary data have to be separated from each other in order to allow for distributed maintenance and concurrent modification. In addition to these advantages, it allows for user-specific modifications (such as the specification of alternative upper models) without compromising the ontology as a whole.

4.1 The interface model: E-EAGLES

By now, the first version of the E-EAGLES ontology has been implemented using OWL/DL with Protege. Currently, it covers all the obligatory and recommended features from the EAGLES recommendations for morpho-syntactic annotation (Wilson and Leech 1996) plus several categories from non-EAGLES conformant tag sets (e.g. noun classifiers).

The classes in the interface model are retrieved from the EAGLES recommendations in the following way:

- obligatory features (i.e. main word classes, such as noun, verb, etc.) specify top-level categories
- recommended features which specify distinctions that are *not* purely inflectional specify more fine-grained subcategories of top-level categories (e.g. the type distinction of nouns: proper nouns vs. common nouns)
- recommended features which specify *inflectional* distinctions are modelled as properties

As the project data includes a MULTEXT-East-based annotation scheme for Russian, the Uppsala scheme, the relevant definitions of MULTEXT-East have been integrated as well.

The hierarchy of verbal classes in E-Eagles is given in Fig. 1. Note that compared to the original EAGLES recommendations, AuxillaryVerb and VerbalNoun are redefined in order to account for non-EAGLES conformant tag sets.

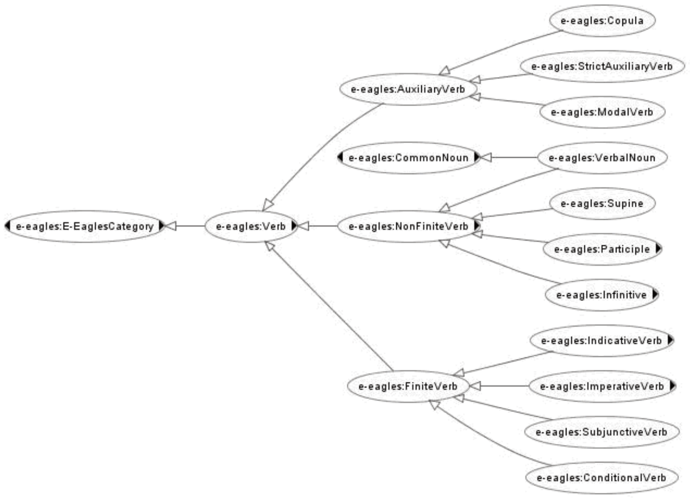

Figure 1: Fragment of upper model: Sub-classification of verbal categories in E-Eagles

[3] Originally, the interface model was termed "upper model" as well, cf. Chiarcos 2006b. However, it is not our specific goal to provide a reference ontology of linguistic terminology in general, but only to specify definitions which hold for the languages and annotation schemes considered by the project. This shift in terminology signals that the function of our specific ontology is not to provide a prescriptive definition of linguistic terminology, but only to gather terms from a restricted set of languages and linguistic phenomena, i.e. currently those hosted by the project.
However, we find currently more than 40 languages and more than 10 part of speech / morphological annotation schemes in the resources hosted by the project, which seems sufficient for a certain degree of generalization, thus making our interface model a very likely starting point to develop a general upper model of linguistic terminology.

In addition to this hierarchy of classes, verbs can be further specified by properties such as hasTense, hasAspect, hasPerson, hasNumber, hasVoice, hasSeparability, hasReflexivity and hasGender.

4.2 A domain model: Uppsala

Then, domain models are built in a similar manner. Usually, annotation guidelines have a document structure which specifies an otherwise implicitly assumed hierarchical organization, thus, a similar hierarchical structuring of concepts can be achieved.

For the tagset applied to the Uppsala corpus, the corresponding structuring of the domain model ontology is given in Fig. 2.

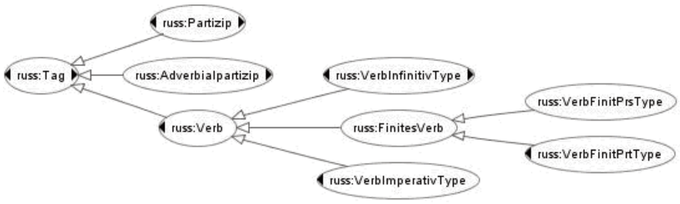

Figure 2: Fragment of domain model: verbal categories in the Uppsala tag set.

Again, inflectional differentiations are specified by properties in the ontology, i.e. hasGender, hasMood, hasVoice, hasPerson, hasNumber, hasFiniteness, hasAspect and hasTense.

In addition to these abstract conceptualizations, concrete tags are integrated as instances into the domain model ontology. Informally, the definition of the Uppsala tag verb_finit_prt_0_sg_neut_refl_pfipf in the ontology can thus be given as:

verb_finit_prt_0_sg_neut_refl_pfipf-is-a VerbFinitPrtType and hasTense(past) and hasVoice(reflexive) and hasFiniteness(finite) and hasGender(neuter) and hasMood(indicative) and hasNumber(singular)

4.3 Linking domain model and interface model

While domain model and interface model are specified as self-contained ontologies in individual owl files, the linking between both is implemented in a separate file by the rdf:description mechanism.

Basically, the linking file contains a specification of domain model classes (not instances) in terms of interface model classes and properties, making up a complex inheritance structure as in Fig. 3 (restricted to subclass relationships). Note that in addition to the primary classes of word types, also properties and property values from the domain model are specified as sub-properties, instances or sub-classes of properties and classes in the interface model.

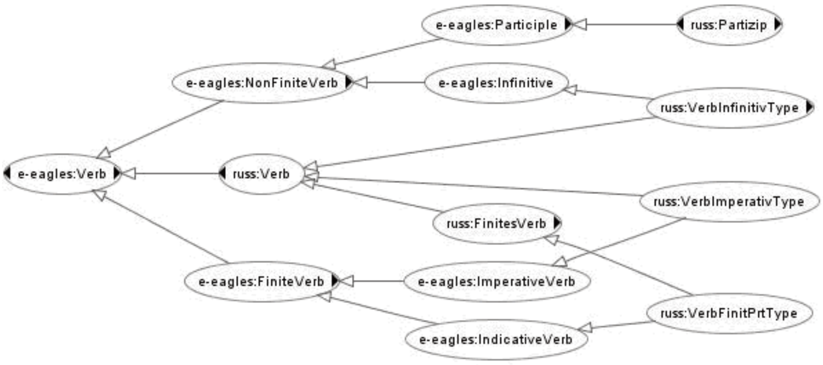

Figure 3: Linking domain model and interface model. The case of verbal categories in the Uppsala tagset.

4.4 A sample query

The linking file imports both the interface model and the corresponding domain model, and thus, it represents an integrating ontology comprising both. If multiple domain models (tag sets) are considered, the corresponding linking files (and the ontologies they import) have to be imported by another file, the so-called *master file* which represents the ontology as a whole.

In the querying scenario, then, expressions based upon classes and properties in the interface model are expanded according to the inheritance structure within and between interface model and domain models, and then evaluated.

As an example, if we're searching for past-tense reflexive verbs, a specification like *Verb and hasTense(Past) and hasVoice(Reflexive)* mentions the interface model classes *e-eagles:Verb*, *e-eagles:Past* and *e-eagles:Reflexive* and the properties *e-eagles:hasTense* and *e-eagles:hasVoice*. According to the interitance structure depicted in Fig. 3, *e-eagles:Verb* expands to *russ:Verb* and further to *russ:VerbFinitPrtType*, etc. Similarly, *e-eagles:hasTense* expands to *russ:hasTense* etc. Thus, amongst other instances, the instance verb_finit_prt_0_sg_neut_refl_pfipf is returned.

The ontology-based query preprocessor, OntoClient, then replaces the ontology-sensitive part of a search query by a disjunction of the tags corresponding to the respective instances, and this modified search query can be further processed by a corpus querying tool.

4.5 Alternative Upper Models

The very same mechanism that was used to link domain model concepts with interface model concepts can be employed to establish a linking between the interface model and an additional upper model which provides independent conceptualizations of linguistic terms. Candidates for such upper models are the OntoTag ontologies (an EAGLES-based ontology of linguistic terms with a special application to English and Iberian languages, cf. de Cea 2004), the Data Category Registry currently developed in the context of the Linguistic Annotation Framework (Ide et al. 2004), or GOLD.

As illustration, we're concentrating on GOLD here, as it is a freely available and well-recipied ontological resource with a good coverage of non-European languages. At the moment, any concept in the E-EAGLES ontology is augmented with a reference to the (E-)GOLD ontology.

Nevertheless, it seems reasonable to keep the interface model ontology and the upper model apart. As the development of GOLD is still ongoing, updated versions of GOLD could compromise the linking with the domain models if the domain models are mapped onto the upper model directly. If both upper model and interface model are separated, a modification of the upper model might force an adaption of the linking between upper model and interface model, but not necessarily between the upper model and any other existing domain models.

As the upper model is linked with the interface model in the same way as the interface model and domain model, the corresponding upper model expressions can be used for the formulation of ontology-sensitive corpus queries.

5. Advantages of the structured approach

The crucial advantage of a structured modular ontology is its highly flexible and user-adaptive character. As illustrated in Fig. 4, the different components of the ontology are stored separately from each other, and as the import mechanism relies on rdf mechanisms, the concrete location of the corresponding files does not affect the validity of the references. As an example, a user may prefer to use a local variant of a certain domain model, for example because his version of the underlying annotation scheme had slightly different naming conventions than the "official" domain model for this annotation scheme, for a typical example see the numerous variants of STTS which have different tags for pronominal adverbs, e.g. PAV, PROAV and PROP. In this case, only some instances in the domain model have to be renamed, whereas the linking can stay as it is. However, in this case the user has to use a local copy of the linking as well which does not differ from the "official" linking in any other ways than the source of the domain model to be imported.

A user thus may introduce an external upper model, he may redefine the linking between an existing domain model and the interface model without affecting either of them, and he may integrate additional domain models. However, he may *not* modify the interface model. As it is the central reference point for any linking file, this could affect the linking of other domain models and produce inconsistencies.

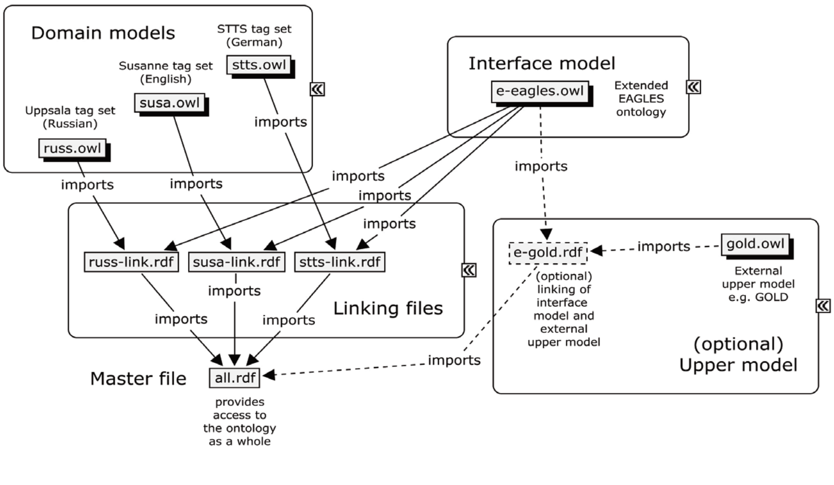

Figure 4. Structured modular ontology

This modular structure is thus highly flexible and user-adaptive. A user might even decide to disagree with the conceptualizations in the interface model and develop his very own alternative, but as long as he provides a linking between his conceptualizations and those of the interface model (i.e. he implements his alternative as an upper model in our sense), he does not have to reconsider the linking to all existing domain models.

Especially in the long run, ongoing maintenance of the ontology might require the integration of additional upper models in order to keep touch with the continuous process of terminological evolution, but not the redesign of the interface model. The effort to have an intelligible interface to the resources associated with certain domain models is thus reduced to the task to maintain the linking between interface model and upper model.

Our implementation provides a *modular view* on the ontology. The ontology consists of three principal components, the upper model presenting a central registry of relevant terminology, several domain models, each covering the tags of one specific POS tag set, and the respective linking between upper model and domain model, which are each stored in independent files.

To access to the ontology as a whole, an additional "master file" is necessary which provides unified access to the interface model, the upper model, the domain models and the linking between them from separate OWL/RDF files. As the interface model does *not* specify the ultimate repository of linguistic terminology, additional upper models can be integrated in this master file. As a user can define own conceptualisations by this mechanism, the main benefit of our approach and the development of the interface model lies in the fact that it is no longer necessary to consider every tag set by its own. Instead, later refinements are mediated by the upper model, thus the most important achievement is that *the interface model provides a unified access to different tag sets* for both querying and redefinition.

In addition to its function in tag set neutral corpus queries and in the theory-neutral definition of language-, project- or task-specific annotation schemes by linking the corresponding domain model with the interface model, the ontology can be practically applied in the design of tag set neutral corpus processing scripts (Krasavina et al. 2007), or, more generally, in the field of Semantic Web applications and ontology-based annotation (for a similar approach on a more restricted set of languages cf. de Cea et al. 2004).

6. Conclusions

In this paper, a modular ontology for the integration of linguistic terminology was presented, and discussed in its logical structure and technical realization. Currently, the interface model covers part of speech and morphological information for a broad variety of languages and is linked to domain models representing eight different part of speech tag sets (STTS in four variants, SUSANNE in two variants, the Uppsala tagset, MENOTA, one typologically-oriented tag set, one tag set from an acquisition project, one tag set for Old High German, a tagset for Tibetan) applied to more than 40 languages (German, English, Russian, Old Norse, several African languages, etc.). Actual versions of different tag sets are available under http://nachhalt.sfb632.uni-potsdam.de/owl/index.html.

Further, we implemented the OntoClient, a JAVA package which can be used as a pre-processor for ontology-based corpus queries. The OntoClient has been integrated into ANNIS, a web application providing access to a broad variety of heterogeneously annotated data collections hosted at Potsdam/Berlin (Götze and Dipper 2006), and in this context, the OntoClient is currently evaluated for its applicability in tag set-neutral/cross-tag set querying of data collections with different part of speech tag sets.

References

1. Chiarcos, Ch. An Ontology for Heterogeneous Data Collections // to appear in Proceedings of Ontologies in Text Technology (OTT'06). 2006a.
2. Chiarcos, Ch. An Ontology for Heterogeneous Data Collections // in Proc. Corpus Linguistics 2006, October 10–14, 2006, St.-Petersburg, St.-Petersburg University Press, 373-380. 2006b.
3. de Cea, G. A., A. Gómez-Pérez, I. Álvarez de Mon, A. Pareja-Lora. 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), vol. 2, 04 05 - 04, 2004, Las Vegas, Nevada, p. 124-128. 2004.
4. Dipper, St., E. Hinrichs, Th. Schmidt, A. Wagner, and A. Witt. Sustainability of linguistic resources // In: Proceedings of the LREC Workshop on merging and layering linguistic information. Genoa. 2006.
5. Erjavec, T. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora // in Proceedings of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, ELRA, Paris. 2004.
6. Farrar, S. and D. T. Langendoen, A Linguistic Ontology for the Semantic Web // GLOT International (3), 97–100. 2003.
7. Götze, M., Dipper, S. 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. 2006.

8. Ide, N., L. Romary, E. de la Clergerie. International Standard for a Linguistic Annotation Framework // Natural Language Engineering, 10: 211-225. 2004.

9. Krasavina, O., Ch. Chiarcos, C, D. Zalmanov. Aspects of topicality in the use of demonstrative expressions in German, English and Russian // Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC-2007), Lagos, Portugal, 29-30 March (to appear). 2007.

10. 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. 1996.

11. Lehmberg, T., Ch. Chiarcos, E. Hinrichs, G. Rehm, and A.Witt. Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System // paper to be presented at Digital Humanities 2007, Digital Text Resources for the Humanities, session on Legal Issues. 2007.

12. Meyers, A (ed.). Annotation Compatibility Working Group Report // in Proceedings of the ACL/Coling '06 Workshop "WS6: Frontiers in Linguistically Annotated Corpora 2006". 2006.

13. Ribarov, K. The Latest Prague Contributions to Written Cultural Heritage Processing // International Journal Information Theories and Applications 11(3): 224-231. 2004.

14. Schmidt, Th., Ch. Chiarcos, T. Lehmberg, G. Rehm, A. Witt and E. Hinrichs. 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.