Ch. Chiarcos

AN ONTOLOGY FOR HETEROGENEOUS DATA COLLECTIONS[1]

In this paper, I describe derivation and practical application of an ontology of word classes manually derived from four different sources:

- the EAGLES recommendations for the morphosyntactic annotation of corpora,
- several language-specific, or task-specific tag sets for part-of-speech tagging,
- the typologically-oriented SFB632 guidelines for part-of-speech tagging, and
- the General Ontology for Linguistic Description (GOLD).

The resulting ontology is intended to provide integrated representation and access to terminologically heterogeneous resources. It will be applied as part of a sustainable archive of linguistic resources to be developed by the project «Sustainability of Linguistic Data», a just-started joint initiative by three German collaborative research centers (SFB).

While in the first phase, the focus of the ontology development has been put on terminology for part-of-speech (POS) tagging which requires hand-crafted methods, a possible extension towards the semi-automatic integration of syntactic annotation will be sketched as an outlook.

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.

---

[1] University of Potsdam, Collaborative Research Center 632 project D1, co-project «Sustainability of Linguistic Data».

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[1] 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]

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
    - It gives access to tag sets employing different tag names.
- tag set independent
    - It gives access to tag sets developed in different theoretical traditions, for different purposes and different languages.
- theory neutral
    - It gives access to heterogeneous tag sets that apply different and incompatible conceptualizations.

[1] *Ribarov K.* The Latest Prague Contributions to Written Cultural Heritage Processing // International Journal Information Theories and Applications 11(3) 2004: P. 224–231.

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

- scalable
    - It gives access to tag sets of different level of detail and coverage.

As we apply an ontology as central data repository which allows for the explicit specification of complex relationships between different terms, the reductionism forced by previous tag set standardization projects can be avoided.

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, SFB 441 (Tübingen, «Linguistic Data Structures»), SFB 538 (Hamburg, «Multilingualism») and SFB 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 sustainable solutions for creation, maintainance, accessibility and distribution of linguistic resources.[1]

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	adaption of EAGLES	Russian tag set	Russian	$\geq$ 877 tags
SFB632 annotation standard	designed for typological research	n/a	26 languages	$\approx$ 79 tags
SFB538/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 in the SFBs

[1] *Schmidt Th., Chiarcos Ch., Lehmberg T., Rehm G., Witt A., Hinrichs E.* 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).

One of our primary aims is to provide the means to ensure the long-term availability of the data collections. Along with technical aspects, as discussed by Dipper et al.[1], this goal involves creating a thorough documentation for the corpora in order to provide easy access for non-specialised users. This includes meta data about the corpora themselves, such as type of data, formats, standards and levels of annotation. Furthermore, the terminology relevant for the annotations has to take into account sustainability considerations.

Focusing on the tag sets used for part-of-speech (POS) tagging in Tübingen, Hamburg and Potsdam/Berlin, we find that our research centers create and use POS-annotated corpora for 29 languages or language stages, annotated according to nine tag sets or tag set variants, cf. Tab. 1.

With this amount of data, several problems can be identified that hinder the direct access to data by using these tag sets: (i) tag names are cryptic, arbitrary and appear in idiosyncratic variants, (ii) different community – or project-specific definitions of tags with the same names, (iii) tag definitions can be extremely complex or missing, (iv) tag sets are of differing granularity.

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.

3. Towards an Ontology of Part-of-Speech Tags

A classical solution to the problem is the «standardisation approach» as employed by the EAGLES recommendations[2]. There, standards for POS tag sets have been formulated – further referred to as the «EAGLES meta scheme» –, which are intended to increase

[1] *Dipper St., Hinrichs E., Schmidt Th., Wagner A., Witt A.*: Sustainability of linguistic resources // Proceedings of the LREC Workshop on merging and layering linguistic information. Genoa, 2006.

[2] *Leech G., Wilson A.* EAGLES Recommendations for the Morphosyntactic Annotation of Corpora, Version of Mar, 1996, URL: **http://www.ilc.cnr.it/EAGLES/annotate/annotate.html**.

tagging accuracy and comparability of automatic taggers and tag sets for most European languages. 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 standardisation approach faces several disadvantages: 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.

To overcome the deficits of the standardisation approach and the definition of meta tag sets, the application of an ontology similar to the GOLD approach[1] can be considered. 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 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

[1] *Farrar S., Langendoen D.T.* A Linguistic Ontology for the Semantic Web. GLOT International. 3. 2003. P. 97–100.

EAGLES. 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 apply 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. Finally, conceptualizations from other non-EAGLES conformant tag sets are integrated into the ontology.

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)-GOLD ontology, has been described elsewhere with greater level of detail[1].

4. Application of the Developed Ontology

By now, the first prototype of the E-GOLD ontology has been implemented using OWL/DL. Currently, it covers all obligatory and recommended features from EAGLES except those which are purely inflectional. These will be added in a later version.

The terminological backbone derived from EAGLES is implemented serves as an *upper model*. The tag sets enumerated in Tab. 1 are stored as independent ontologies, the so-called *domain models*. The concepts of the domain models, then, are *linked* to concepts in the upper model in terms of conceptual subsumption. Accordingly, all tags from different tag sets that are linked to a certain upper model concept or its subconcepts can be retrieved as indirect instances of the corresponding upper model concept.

[1] *Chiarcos Ch.* An Ontology for Heterogeneous Data Collections. To appear in Proceedings of Ontologies in Text Technology (OTT–06).

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 upper model, the domain models and the linking between them from separate OWL/RDF files. As the upper model does *not* specify the ultimate repository of linguistic terminology, but only a rather traditional view of generally accepted categories, individual modifications of the upper model can be placed in this master file. As a user can define his own conceptualisations based on this upper model, the main benefit 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 upper model provides a unified access to different tag sets* for both querying and redefinition.

5. Current Research Activities

In this paper, an ontology-based approach towards the integration of linguistic terminology was presented, the derivation of an ontology for word classes, aspects of its technical realization, and general conceptual considerations affecting its design.

Having implemented a first version of the ontology covering the tag sets mentioned in Tab. 1, we're currently developing the ONTOCLIENT, a JAVA package which can be used as a pre-processor for ontology-based corpus queries. Confronted with upper model concepts, these are expanded into disjunctions of POS tags as shown in Fig. 1.

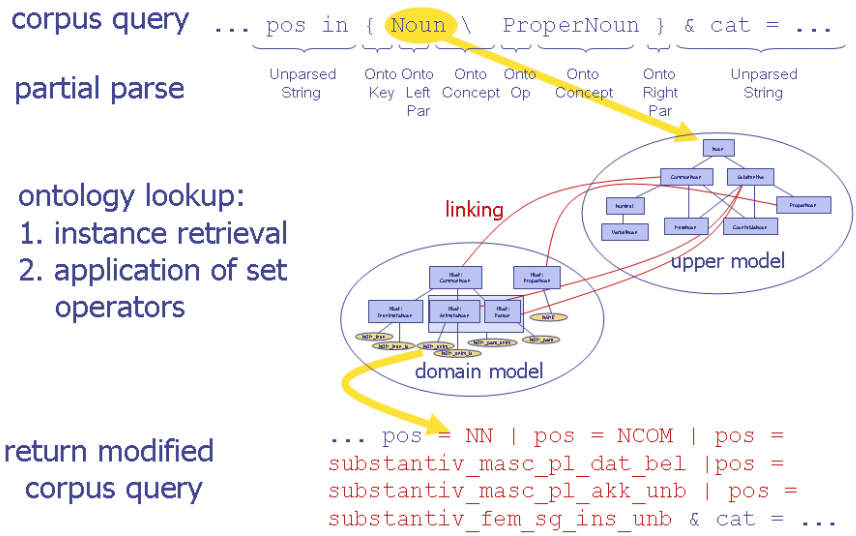

Fig. 1. Ontology-based corpus querying

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[1], and in this context, the ONTOCLIENT will be evaluated for tag set-neutral/cross-tag set querying of data collections with different part of speech tag sets.

[1] *Götze M., Dipper S.* ANNIS – Complex Multilevel Annotations in a Linguistic Database // Proceedings of the EACL Workshop on Multi-dimensional Markup in Natural Language Processing NLPXML–2006. Trient, Italy, 2006. P. 61–64.