Porting Elements of the Austrian Baroque Corpus onto the Linguistic Linked Open Data Format

Ulrike Czeitschner
Institute for Corpus Linguistics and Text Technology
Austrian Academy of Sciences
Ulrike.Czeitschner@oeaw.ac.at

Thierry Declerck
German Research Center for Artificial Intelligence, GmbH
thierry.declerck@dfki.de

Claudia Resch
Institute for Corpus Linguistics and Text Technology
Austrian Academy of Sciences
Claudia.Resch@oeaw.ac.at

Abstract

We describe work on porting linguistic and semantic annotation applied to the Austrian Baroque Corpus (ABaC:us) to a format supporting its publication in the Linked Open Data Framework. This work includes several aspects, like a derived lexicon of old forms used in the texts and their mapping to modern German lemmas, the description of morpho-syntactic features and the building of domain-specific controlled vocabularies for covering the semantic aspects of this historical corpus. As a central and recurrent topic in the texts is death and dying, a first step in our work was geared towards the establishment of a death-related taxonomy. In order to provide for linguistic information to their textual content, labels of the taxonomy are pointing to linked data in the field of language resources.

1 Introduction

ABaC:us[1] is a project conducted at ICLTT[2] focusing on the creation of a thematic research collection of texts based on the prevalence of sacred literature during the Baroque era, in particular the years from 1650 to 1750. Books of religious instruction and works concerning death and dying were a focal point of Baroque culture. Therefore, the ABaC:us collection holds several texts specific to this genre including sermons, devotional books and works related to the dance-of-death theme. The corpus comprises complete versions, not just samples, of first editions[3] yielding some 165.000 running words. An interdisciplinary approach has been adopted for the creation of this digital corpus, which is designed to meet the needs of both literary/historical and linguistic/lexicographic research.

In order to guarantee easy data-interchange and reusability, the corpus was encoded in TEI (P5).[4] In addition, applied PoS tags and lemma information[5], taken from modern German language, allow for complex search queries and more sophisticated research questions.[6] While starting work on the semantic annotation of the corpus, we saw the need to develop a specific taxonomy, which would also ease the task of semi-automated semantic annotation of the morpho-syntactically annotated corpus and other related texts (Declerck et al., 2011, Mörth et al., 2012). Following a bottom-up strategy, we identified all death-related lexical units such as nom-

[1] Partly supported by funds of the Österreichische Nationalbank, Anniversary Fund (project number 14783), the ABaC:us project started in spring 2012. See http://www.oeaw.ac.at/icltt/abacus and http://www.oeaw.ac.at/icltt/abacus-project for more details.

[2] The Institute for Corpus Linguistics and Text Technology (http://www.oeaw.ac.at/icltt/) of the Austrian Academy of Sciences in Vienna pursues corpus-based linguistic and literary research, focusing on the creation and adaptation of corpora and dictionaries as well as technologies for building, accessing and exploiting such data.

[3] The majority of the selected works can be ascribed to the Baroque Catholic writer Abraham a Sancta Clara (1644-1709): e.g. *Mercks Wienn* (1680), *Lösch Wienn* (1680), *Grosse Todten Bruderschafft* (1681), *Augustini Feuriges Hertz* (1693), and *Todten-Capelle* (1710). For detailed information about the author see Eybl (1992) and Knittel (2012).
The ABaC:us collection combines high quality digital texts with image scans of facsimiles of the earliest known prints housed in different libraries such as the Austrian National Library, the Vienna City Library, the Melk Abbey, and the Library of the University of Illinois.

[4] See http://www.tei-c.org/Guidelines/P5/ for details.

[5] PoS tagging has been realized using Tree Tagger, an open standard developed at the University of Stuttgart. See http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ for more information.

[6] All ABaC:us texts, which represent a non-canonical variety, were tagged using automated tools adapted to the needs of historic language and were afterwards verified by domain-experts.

inal simplicia, compound nouns and multi-word expressions for the personification of death. In addition, all terms and phrases dealing with the "end of life", "dying" and "killing" were identified. In total, more than 1.700 occurrences could be discovered in *Mercks Wienn, Grosse Todten Bruderschafft* and *Todten-Capelle,* the three most important works of our corpus.

The next step consisted in organizing the identified vocabulary in a taxonomy, which is encoded in the SKOS format (Simple Knowledge Organization System)[7]. Based on the Resource Description Framework (RDF)[8], SKOS "provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary." [9] We chose it because SKOS concepts can be (1) "semantically related to each other in informal hierarchies and association networks", (2) "the SKOS vocabulary itself can be extended to suit the needs of particular communities of practice" and finally, because it (3) "can also be seen as a bridging technology, providing the missing link between the rigorous logical formalism of ontology languages such as OWL and the chaotic, informal and weakly-structured world of Web-based collaboration tools."[10] With the use of SKOS (and RDF), we are also in the position to make our resource compatible with the Linked Data Framework[11].

The following sections provide an overview of the ABaC:us taxonomy and describe the way the language data contained in its labels are linked to web resources in the Linguistic Linked Open Data (LLOD) cloud[12].

2 The ABaC:us Taxonomy

Currently the scheme of the ABaC:us taxonomy consists of 7 concepts comprising 362 terms or phrases, which are encoded in SKOS labels. In addition, 137 compounds and associated terms have been integrated in 4 more temporary concepts, which still await a further processing. The terms included in the labels (both preferred and alternative ones) have been manually excerpted from the original texts and partly normalized. The majority of texts are written in German, some parts in Latin, therefore all lexical labels belong to one of these languages.

Table 1 lists concepts and definitions. Row 3 and 4 show selected examples for preferred and alternative terms in German and Latin—for better readability, a rudimentary English translation has been added. The reader can see how the death as "end of life" (concept/1) and the personalized death (concept/2) are distinguished.

Labels are related to each other by means of the following properties: *abacus:hasTranslation* and inverse *abacus:isTranslationOf,* used for German and corresponding Latin terms, *abacus:hasVariant* and inverse *abacus:isvariantOf* indicate spelling variants.

In order to systemize concept 4 (dealing with "manners of death") we use the annotation property *skos:comment*: "death by accident or circumstances", "death by disease", "death by foreign hand", and "death as a murderer" (i.e. personification of death)[13]. We refrained from creating concepts (labeled *skos:broader*) in this case, as this kind of terms does not represent corpus text. Next, we will link for this purpose to corresponding concepts included in external knowledge sources, allowing thus to distinguish between concepts and terms directly related to our corpus and other knowledge sources that can be used for additional interpretation and classification. This can be seen as the most important difference of the ABaC:us taxonomy to other vocabularies, which are often characterized by strict hierarchical formalisms making them little useful for literary sciences[14].

3 Lexicalization of the Taxonomy

In order to be able to use the taxonomy in the context of NLP applications, there is the need to lexicalize the content of its labels, enriching them with linguistic information. This includes tokenization, lemmatization, PoS tagging, and possibly other levels of natural language (NL) processing. Labels enriched with this information can be better compared to text, which has also been submitted to NL processing tools. If a certain amount of linguistic similarity is found in a text passage with a lexicalized label, this text segment can then be semantically annotated with the concepts the label is associated with.

[7] http://www.w3.org/2004/02/skos/
[8] http://www.w3.org/RDF/
[9] http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/
[10] Ibid.
[11] See http://linkeddata.org/
[12] http://linguistics.okfn.org/resources/llod/

[13] Those comments are not displayed in Table 1.
[14] Recently Bradley and Pasin (2012 and 2013) described how informal semantic annotations could become more compatible with computer ontologies and the Semantic Web.

skos:concept	skos:definition	skos:prefLabel	skos:altLabel
concept/1	"Das Ende des Lebens" *"the end of life"*	"Tod" @de "mors" @la *"death"*	"End" "Garauß" "Hintritt" "Todsfall" "Verlust deß Lebens"
concept/2	"Der Tod als Subjekt" *"death as a subject"*	"Tod" @de "mors" @la *"death"*	"dürrer Rippen-Kramer" "General Haut und Bein" "ohngeschliffener Schnitter" "Reuter auf dem fahlen Pferd" "Verbeinter Gesell"
concept/3	"aufhören zu leben" *"the process of dying"*	"sterben" @de "mori" @la *"dying"*	"ad Patres gehen" "das Valete von der Welt nehmen" "dem Tod vnter die Sensen gerathen" "den Todten-Tantz antretten" "in Gott entschlaffen"
concept/4 *(Comment: This concept is about* "Todesarten", *"manners of death")*	"einen bestimmten Tod erleiden" *"specific ways of dying"*	"getötet werden" @de *"to be killed"*	"aufgehängt werden" "erbärmlich hingerichtet werden" "ermort werden" "mit solchen vergifften Pfeil getroffen werden" "zu todt gebissen werden"
concept/5	"Verstorbene, Leichen" *"dead bodies"*	"Toter" @de "mortuus" @la *"corpses"*	"christliche Leiche" "Leichnam" "seelig-verstorbener" "todter Cörper" "Todter"
concept/6	"tot sein" *"to be dead"*	"tot" @de "mortuus" @la *"dead"*	"abgestorben" "der Geist ist hinaus" "leblos" "verblichen" "verstorben"
concept/7	"töten, ermorden" *"to kill someone"*	"töten" @de *"killing"*	"erwürgen" "morden" "todt schlagen" "tödten" "Vergifften"

Table 1: ABaC:us Taxonomy

The model we adopt for the representation of the results of lexicalized labels is the one described by *lemon*[15], developed in the context of the Monnet project[16]. *lemon* is also available as an ontology[17], which has been imported in our taxonomy, so that we can make direct use of all classes and properties of this model.

3.1 Tokenization and Sense Disambiguation

All tokens in ABaC:us have been semi-automatically annotated with lemma and PoS information, following the STTS tag-set (Mörth et al., 2012)[18], so that all parts of the texts selected as relevant terms for inclusion in the labels come already with this information. Thus, our task consists mainly in applying *lemon* ontology elements for annotating the labels of the taxonomy with this linguistic information.

As can be seen in Example **1** below, for the term included in the alternative label "rasender Tod@de" (*raging death*), we make use of the *lemon* property *decomposition* for encoding the results of tokenization. And we use the *lemon* property *altRef*, which has as rdfs:range an entity that is encoded as an instance of the *lemon* class *lexicalSense*[19], for linking to the concept the alternative label is an expression of.

3.2 Linking to external Lexical and Linguistic Resources

We still need to associate the tokens, which are now each encoded as value of the *lemon* property *decomposition*, with morpho-syntactic information. As mentioned earlier, we already have all the information about the corresponding modern German lemmas and PoS (in the STTS format) for all tokens of the corpus.

But, instead of using directly the *lemon* class *lexical entry* and the *lemon* properties *canonical form* and *lexical property* for including the linguistic information we have for every token in the corpus, we are for now linking the values of

[15] *lemon* stands for "Lexicon Model for Ontologies". See http://lemon-model.net/ and McCrae et al. (2012)
[16] See www.monnet-project.eu
[17] See http://www.monnet-project.eu/lemon
[18] The STTS tag-set is described, among others. here: http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html
[19] See http://lemon-model.net/lemon.rdf for the whole list of properties and classes of *lemon*.

the *lemon* property *decomposition* to already existing lexical entries that are encoded in the LOD format. We choose for this the actual DBpedia instantiation of Wiktionary[20]. There we get also the information that "rasender" is an adjective with lemma "rasend" and that "Tod" is a noun with lemma "Tod" (see Example **1** below)[21]. The two meanings we have distinguished in the ABaC:us taxonomy for "Tod" (*death*), as the "end of life" and as "a subject", are also present in this external resource[22]. Depending on the specific Wiktionary entries, we have a variable number of sense-specific translations at our disposal. The word "Tod", with the meaning "end of life", is provided with 44 translations. We can automatically add those labels to our taxonomy and link them to the German labels via the *abacus:isTranslationOf* property, and so support cross-lingual access to our semantically annotated corpus. It was more difficult to find an English equivalent for the second meaning of "death", "death as a subject"[23], since no direct translation for English is given in this instantiation of Wiktionary. The same can be said of the ambiguous German lemma "rasend" (*raging*).

As a result, the term "rasender Tod@de" (*raging death*) is now encoded in our taxonomy (with *lemon* being integrated) this way:

<http://www.oeaw.ac.at/icltt/abacus/term/2.004-de>
rdf:type owl:NamedIndividual , skosxl:Label ;
skosxl:literalForm "rasender Tod"@de ;
<http://www.monnet-project.eu/lemon#decomposition>
<http://wiktionary.dbpedia.org/page/rasend-German-Adjective-1de> ;
<http://www.monnet-project.eu/lemon#decomposition>
<http://wiktionary.dbpedia.org/page/Tod-German-Noun-2de> ;
<http://www.lemon-model.net/lemon#altRef>
<http://www.oeaw.ac.at/icltt/abacus/concept/2> ;
abacus:isVariantOf
<http://www.oeaw.ac.at/icltt/abacus/term/2.003-de> .

Example 1: The simplified entry for the label "rasender Tod" (*raging death*)

[20] See http://dbpedia.org/Wiktionary. There, *lemon* is also used for the description of certain lexical properties.

[21] But in the longer term we will use the *lemon* constructs for linking to the URIs associated to those pieces of information in the DBpedia coverage of Wiktionary.

[22] See wiktionary.dbpedia.org/page/Tod-German-Noun-1de and wiktionary.dbpedia.org/page/Tod-German-Noun-2de for abacus:concept/1 and abacus:concept/2 respectively.

[23] http://wiktionary.dbpedia.org/page/death-English-Noun-2en.

4 Conclusion

The ABaC:us collection contains a wide range of death-related linguistic vocabulary deriving from the Baroque era. Its writers were extremely inventive in paraphrasing experiences with death and dying. Thus, one integral approach was to make those different concepts more easily discernible. The numerous SKOS labels in the ABaC:us taxonomy give evidence of how the culture of death and dying was transmitted in lexical and linguistic patterns. By making those patterns accessible and reusable on the (L)LOD, we complement existing contemporary concepts of the topic and provide a basis for sharing and comparing the concepts, which can be used in NLP applications in the context of eHumanities.

References

John Bradley, Michele Pasin. 2012. Annotation and Ontology in most Humanities research: accommodating a more informal interpretation context. DH2012 NeDiMaH Ontology Workshop.

John Bradley, Michele Pasin. 2013. Fitting Personal Interpretations with the Semantic Web. In: *Proceedings of Digital Humanities 2013.* University of Nebraska-Lincoln:118-120.

Thierry Declerck, Ulrike Czeitschner, Karlheinz Mörth, Claudia Resch, Gerhard Budin. 2011. A Text Technology Infrastructure for Annotating Corpora in the eHumanities. In: *Proceedings of the International Conference on Theory and Practice of Digital Libraries* (TPDL-2011):457-460.

Franz M. Eybl. 1992. Abraham a Sancta Clara. Vom Prediger zum Schriftsteller. Max Niemeyer, Tübingen, D.

Anton Philipp Knittel (Ed.). 2012. Unterhaltender Prediger und gelehrter Stofflieferant. Abraham a Sancta Clara (1644-1709). Beiträge eines Symposions anlässlich seines 300. Todestages. Edition Isele, Eggingen, D.

John McCrae, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, Tobias Wunner. 2012. Interchanging lexical resources on the Semantic Web. In: *Language Resources and Evaluation.* Vol. 46, Issue 4, Springer:701-719.

Karlheinz Mörth, Claudia Resch, Thierry Declerck, Ulrike Czeitschner. 2012. Linguistic and Semantic Annotation in Religious Memento Mori Literature. In: *Proceedings of the LREC'2012 Workshop: Language Resources and Evaluation for Religious Texts (LRE-Rel-12).* ELRA: 49-52.