Introducing the Austrian Baroque Corpus: Annotation and Application of a Thematic Research Collection

Claudia Resch, Ulrike Czeitschner, Eva Wohlfarter, Barbara Krautgartner

Austrian Academy of Sciences

claudia.resch@oeaw.ac.at, ulrike.czeitschner@oeaw.ac.at,
eva.wohlfarter@oeaw.ac.at, barbara.krautgartner@oeaw.ac.at

Abstract

This paper gives an overview of a relatively new thematic corpus based on German sacred literature of the Baroque period. At present, the digital collection consists of several texts specific to the *memento mori* genre. All texts in the Austrian Baroque Corpus (ABaC:us) have been enriched with different layers of structural information and tagged using automated tools adapted to the specific needs of the language of the period. One important achievement of the project is that each occurring historic word form has been electronically mapped to its corresponding lemma in High German and corrected or verified by domain experts. In all phases of the workflow, the interdisciplinary team (literary, linguistic, and text technology specialists) insisted on high quality linguistic and semantic annotation, and worked towards creating a sound basis that would allow for more sophisticated research questions. The current version of the interface can be seen as a case example showing how the ABaC:us team provides improved access to these rare pieces of macabre literature that give fascinating evidence of Baroque culture and attitudes towards Life and Death.

1 Introduction

The acronym ABaC:us stands for Austrian Baroque Corpus, a digital thematic collection of German texts published during the Baroque Era, in particular the years from 1650 to 1750. The present corpus was compiled between 2010 and 2015 at the Institute for Corpus Linguistics and Text Technology (ICLTT) and at the newly founded Austrian Centre for Digital Humanities (ACDH) of the Austrian Academy of Sciences, alongside two other associated research projects[1].

[1] The generated textual data of both the project "Text-Technological Methods for the Analysis of Austrian Baroque Literature" (March 2012 – September 2014, supported by funds of the Österreichische Nationalbank, Anniversary Fund) and of the project "Mortuary Cult in 17th Century Vienna: Confraternity Studies in the Digital Age" (June 2014 – May 2015, supported by funds of the City of Vienna) as well are part of the existing ABaC:us collection. The first project mentioned started with building the corpus from scratch and aimed at identifying rhetorical patterns and regularities on a lexical and syntactic level. The second project focuses on rare literary and non-literary texts that were produced by a Viennese confraternity. The digital analysis of the printed pamphlets (calendars, statutes and instructions) will lead to evidence-based assumptions on the role of confraternal associations in Counter Reformation Vienna.

Digital textual data in German for the Early Modern period[2] remain underdeveloped and underexplored. Therefore it was decided that a historical corpus should contain significant and thematically connected examples of Baroque literature. Accordingly, the ABaC:us collection is based on the prevalence of sacred literature and contains mainly textual sources concerning death and dying. The Baroque transience topos and its *memento mori* appeal shaped and permeated different literary and non-literary genres. Richly illustrated emblem books[3] in prose and verse were a focal point of Baroque culture and were frequently printed. They were meant to remind people of the fragility of their existence and the inevitability of death. By reading these moralizing texts, people were supposed to be directed towards holding themselves in steady expectation of their own demise and admonished to live a life of virtue in order to be prepared for death at all times.

Figure 1: Range of digitized (title) pages being part of the ABaC:us collection, including works of Abraham a Sancta Clara, Jacob Balde, Abraham Megerle, Johann Valentin Neiner, and Florentius Schilling.

[2] Examples for other resources from that time period are some texts of the „Bonner Frühneuhochdeutschkorpus“ http://www.korpora.org/Fnhd/, the “GerManC project: A representative historical corpus of German“ at the University of Manchester http://www.llc.manchester.ac.uk/research/projects/germanc/, and the earlier texts of the huge collection „Deutsches Textarchiv“ at the Berlin Brandenburg Academy of Sciences and Humanities www.deutschestextarchiv.de/.

[3] See also Peter Boot. *Mesotext. Digitised Emblems, Modelled Annotations and Humanities Scholarship.* Amsterdam: Pallas Publications – Amsterdam University Press 2009.

At present, the thematic ABaC:us collection holds 20 religious writings motivated by the fear of sudden death and the central *memento mori* theme including sermons, obituaries, devotional books, compilations of prayers, songs and works related to the dance-of-death theme. When building up the corpus from scratch, only original textual data in full length were chosen. As a matter of philological principle, only early and if possible the first known editions (editio princeps) and rare specimens from different monastic and public libraries were selected for the digitalization procedure.

2 The Core of ABaC:us

The Austrian Baroque Corpus currently contains a total of more than 210,000 running words. The main part, approximately 180,000 tokens (85%), can be attributed to the Augustinian monk Abraham a Sancta Clara (1644-1709)[4], a very popular preacher and widely read author[5]. Due to his literary talent and his distinctive style, Abraham a Sancta Clara's books reached a wide audience; they were frequently reprinted and distributed across the German-speaking lands so that Abraham a Sancta Clara remained very popular even after his death: "The fact that Pater Abraham's books sold well, led several publishers to the idea of combining parts of his already published works with texts of other authors, ascribing the literary hybrid to the Augustinian preacher."[6] The 300th anniversary of his death was the motivation to start digitization in part – also including works whose authorship is in doubt – hence some parts of the collection also derive from the preacher's field of influence, for example from other friars in his religious order, or his publishers and imitators.

Five of his most popular texts dealing with death and dying – *MERCKS WIENN* (1680), *LÖSCH WIENN* (1680), *DIE GROSSE TODTENBRUDERSCHAFT* (1681), *AUGUSTINI FEURIGES HERTZ* (1691) and *BESONDERS MEUBLIERT- UND GEZIERTE TODTEN-CAPELLE* (1710) – have a strong association with the city of Vienna. They constitute the core of ABaC:us and are the first to be made available online.

[4] Eybl, Franz M. *Abraham a Sancta Clara*. In: *Killy Literaturlexikon* Volume 1. Berlin: de Gruyter, 2008, p. 10-14 and Eybl, Franz M. *Abraham a Sancta Clara. Vom Prediger zum Schriftsteller.* Tübingen: Max Niemeyer Verlag 1992.

[5] The remaining texts are corresponding to Jacob Balde, Abraham Megerle, Johann Carl Megerle, Johann Valentin Neiner, Franz Peikhart, Emmerich Pfendtner, and Florentius Schilling.

[6] Šajda, Peter: *Abraham a Sancta Clara: An Aphoristic Encyclopedia of Christian Wisdom*. In: *Kierkegaard and the Renaissance and Modern Traditions – Theology*. Ashgate 2009. p. 3.

Figure 2: Bar graph of the core corpus: *MERCKS WIENN* 424 pages/57.945 tokens (blue), *LÖSCH WIENN* 250 pages/24.843 tokens (green), *TODTEN BRUDERSCHAFT* 76 pages/12.514 tokens (pink), *AUGUSTINI FEURIGES HERTZ* 112 pages/19.659 tokens (violet), *TODTEN-CAPELLE* 546 pages/67.034 tokens (turquoise).

3 Annotation Process

Text capture using OCR of early modern editions is well-known to be problematic. When possible, existing digitized texts have been used as a basis for ingestion into the corpus (as it was only the case with *MERCKS WIENN*, freely available on Zeno.org[7]). For texts that had to be captured using OCR, ABBYY FineReader 7 was the best solution as it is capable of scanning both Roman and Black Letter typefaces (Fraktur).

All original primary sources have been fully digitized, transcribed, and encoded as TEI P5 conformant files[8]. Considerable attention has been paid to the structural and typographic features of the texts. The manual tagging covered names of historical, mythological and biblical interest as well as place names:

[7] See http://www.zeno.org/Literatur/M/Abraham+a+Sancta+Clara/Satirischer+Traktat/Mercks+Wienn

[8] See http://www.tei-c.org/index.xml

```
<seg rend="antiqua">
    <persName type="hist" key="juliusCäsar">
        <w lemma="Julius" type="NE" xml:id="MW_dlel33376">Julius</w>
        <seg type="whitespace"> </seg>
        <lb/>
        <w lemma="Cäsar" type="NE" xml:id="MW_dlel33381">Cæsar</w>
    </persName>
    <pc type="comma" xml:id="MW_dlel33383">,</pc>
    <seg type="whitespace"> </seg>
    <persName type="hist" key="antoniusPius">
        <w lemma="Antonius" type="NE" xml:id="MW_dlel33388">Antonius</w>
        <seg type="whitespace"> </seg>
        <w lemma="Pius" type="NE" xml:id="MW_dlel33392">Pius</w>
    </persName>
    <pc type="comma" xml:id="MW_dlel33394">,</pc>
    <seg type="whitespace"> </seg>
    <persName type="hist" key="hadrianus">
        <w lemma="Hadrianus" type="NE" xml:id="MW_dlel33399">Hadrianus</w>
    </persName>
    <pc type="comma" xml:id="MW_dlel33401">,</pc>
    <seg type="whitespace"> </seg>
    <lb/>
    <persName type="hist" key="carolusMagnus">
        <w lemma="Carolus" type="NE" xml:id="MW_dlel33407">Carolus</w>
        <seg type="whitespace"> </seg>
        <w lemma="Magnus" type="NE" xml:id="MW_dlel33411">Magnus</w>
    </persName>
```

Figure 3: Example of the annotation of historical names

```
<w lemma="zu" type="APPR" xml:id="MW_dle23895">zu</w>
<seg type="whitespace"> </seg>
<lb/>
<placeName key="prag" subtype="dire" type="city">
    <w lemma="Prag" type="NE::top" xml:id="MW_dle23902">Prag</w>
</placeName>
<seg type="whitespace"> </seg>
<pc type="internal" xml:id="MW_dle23906">/</pc>
<seg type="whitespace"> </seg>
<w lemma="in" type="APPR" xml:id="MW_dle23910">in</w>
<seg type="whitespace"> </seg>
<placeName key="thur" subtype="dire" type="dist">
    <w lemma="Thüringen" type="NE::top" xml:id="MW_dle23915">Thüringen</w>
</placeName>
<seg type="whitespace"> </seg>
<pc type="internal" xml:id="MW_dle23919">/</pc>
<seg type="whitespace"> </seg>
<placeName key="neth" subtype="dire" type="coun">
    <w lemma="Niederlande" type="NE::top" xml:id="MW_dle23925">Niederland</w>
</placeName>
```

Figure 4: Example of the annotation of place names with NE::top (as city / district / country)

Errors and mistakes in the original texts have been allowed to stand; editorial interventions and suggestions are recorded in the mark-up.

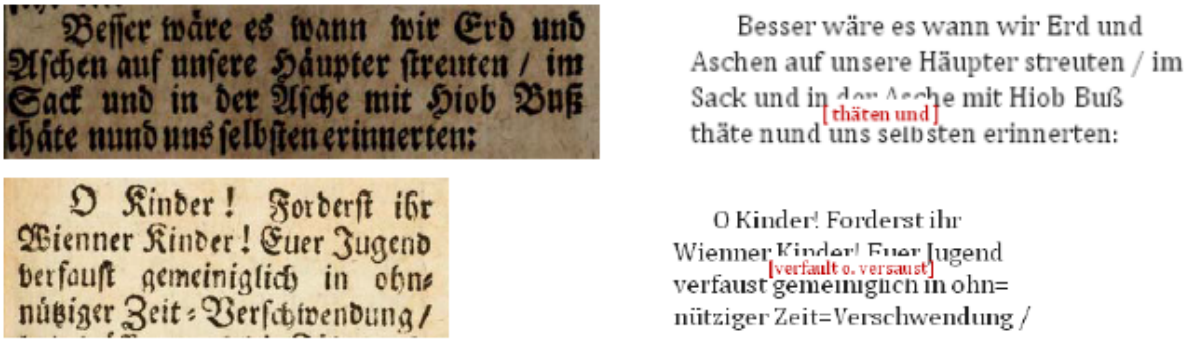

Figure 5: Corrections of obvious typographical errors: Original passage and electronic text

The five core texts ascribed to Abraham a Sancta Clara have been linguistically annotated with Part-of-Speech tags (PoS) and lemmatized. Thus each token was automatically mapped to a word class by *TreeTagger*[9] according to

[9] See http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

the guidelines (1999) of the 54-part *Stuttgart-Tübingen-TagSet* (STTS)[10], a standardized tagset for German language resources. In addition, every word was automatically supplied with its lemma.

As the tagger was developed for modern day German language and was used out of domain – namely for a historical variety of Early Modern German, which at that time was not fully standardized – the project group had to cope with many (expected) erroneous mappings and mismatches: Even the slightest deviations in orthographic conventions (e.g. variational suffixes: *behutsamb/behutsam* (cautious), doubling of consonants: *Freundschafft/Freundschaft* (friendship), elimination of long vowels *Bott/Bote* (messenger), etc.) caused wrong annotations and had to be verified and manually corrected.[11] Problems arose with so-called multi-word lexemes and with historical conventions for separating words, such as *Stephans Domkirchen* or *Haupt Statt* (according to modern German orthographic norms: *Stephansdomkirche* (St. Stephen's Cathedral) and *Hauptstadt* (Capital) or contracted forms (see examples below)). Because of the differences between Modern and Early Modern German the *Stuttgart-Tübingen-TagSet* had to be adapted and extended with additional categories to deal with contracted forms such as *wirstu* (old form with clitic instead of *wirst du*), *mans* (*man es*) or *machs* (*mach es*). For this purpose multiple tags have been linked by an underscore: tag_tag. Subsequently, lemmatization had to be adjusted as well.

Token	PoS	Lemma
wirstu	VVFIN_PPER (finite content verb + irreflexive personal pronoun)	werden_du become_you
mans	PIS_PPER (substituting indefinite pronoun + irreflexive personal pronoun)	man_es one_it
machs	VVIMP_PPER (imperative content verb + irreflexive personal pronoun)	machen_es make_it

Table 1: Examples for the extended tagset consisting of PoS and lemma information

Historical corpora particularly require a comprehensible lemmatization. Word occurrences such as *Fegefeuer*, meaning "purgatory", which appear in several orthographic versions (*Feeg=Feuer, Feegfeuer, Feg=Feuer, Fegefeuer, Fegfeuer, Fegfeur, Fegfewer*), can only be found easily through their lemma or base form. The manual control and mapping of the data refers to two different common dictionaries: the modern standard dictionary Duden (http://www.duden.de/) and the German dictionary of the Grimm Brothers (http://woerterbuchnetz.de/DWB/). Words labelled FM („fremdsprachliches Material", elements from a foreign language) had to be lemmatized entirely by domain experts. Most FM words stem from phrases and passages in Latin, for this reason, Stowasser (1998) was used in order to assign the right lemma to each Latin word. Only a minority of words tagged FM were not Latin, but French or Italian.

[10] See http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

[11] Hinrichs and Zastrow have already noticed that – compared to other texts – Abraham a Sancta Clara's style exhibits by far the highest average sentence length, which might also be a reason why the author's test data had "the highest number of tagging errors", see Erhard Hinrichs, Thomas Zastrow. *Linguistic Annotations for a Diachronic Corpus of German*. In: *Linguistic Issues in Language Technology*, Volume 7, issue 7 (2012), p. 11.

Words not listed in any of the dictionaries were particularly challenging. The main part of these "out of-vocabulary words" were compounds invented by Abraham a Sancta Clara or his imitators to add a playful element to the texts, for instance *vernunftselig* (rationally blessed), *Tigergemüt* (temper of a tiger), or *Felsenzucht* (breeding of rocks). In other cases of creative language use, we decided to use two lemmata to capture their full meaning in the linguistic annotation. Multiple meanings and ambiguous forms (occurring particularly in puns) were separated by a vertical bar: for instance *Kümmernis* | *Kümmer=Nuß* (grievance, containing the German word for "nut").

```
<w lemma="Kümmernis | Kummernuss" type="NN" xml:id="MW_dlell5681">Kummer=Nuß</w>

<w lemma="Friesländer | Frissländer" ana="#oov" type="NN" xml:id="MW_dle5845">Frißländer</w>

 <w lemma="müssen | Mus" type="NN" xml:id="TB_dle44882">Muß</w>
```

Figure 6:
Kummer=Nuß: Lemma *Kümmernis* | *Kümmer=Nuß* (grievance *or* nut of grief)
Frißländer: Lemma *Friesländer* | *Frissländer* (inhabitant of Frisia *or* somebody who likes to eat)
Muß: Lemma *müssen* | *Mus* (he has to *or* mush)

All words not found in any of the above-mentioned dictionaries were marked by an asterisk and annotated with a notional modern lemma that comes closest to the particular word form.

- Venus |6|
- Venusbild |1|
- Venusfeuer* |1|
- Venusgenosse* |1|
- Venuskonfekt* |2|
- Venusluder* |1|

Figure 7: Lemmata with asterisk (fire of Venus, companion of Venus, sweets of Venus, slut of Venus)

In order to identify, process and remove all tagging errors and mismatches, an updated version of the *token_editor* developed at the ICLTT was used. The tool loads the PoS and lemma data of the corpus and displays them in vertical lists: It allows researchers to read the text, create lists of particular items on the basis of regular expressions and assign new data to these datasets. Although the tagger determines a PoS-tag for each word and the *token_editor* facilitates the evaluation of the automatic assignment of word labels and allows for the verification of the tagger's suggestion and if necessary the correction of the results by human annotators in an accelerated way, the semi-automatic linguistic annotation was still a time-consuming process.

After having fully worked with the first three texts, we had a sound basis for the following ones and made first experiments in adapting the specific historical language material for further annotation procedures. Using our manually-corrected annotations as additional training material had a very positive effect for domain adaptation and

improved the performance of the applied tools significantly[12]. A wordlist generated from the first texts was instrumental in reducing errors in the ongoing process of tagging.

In case of *MERCKS WIENN* and *LÖSCH WIENN* improvement was evident:

LÖSCH WIENN	**without** the lexicon of *MERCKS WIENN*	**with** the lexicon of *MERCKS WIENN*
PoS accuracy	71,2%	82,7%
Lemma accuracy	57,5%	72,4%

Table 2: Results of the tagger evaluation

Manually-corrected data[13] provided relevant lexical information to positively influence the performance of the tagger and not only improved the PoS results but also the lemmatization (by almost 15%).

Our method of annotating more texts of the same time period and genre was an incremental bottom-up process and resulted in high quality data. Using the described bootstrapping approach together with meticulous revision has made ABaC:us a thoroughly validated and reliable corpus which can be utilized for:

- the annotation of other distinct texts
- evaluating the quality of automatically generated lexical data from corpora
- the training of different taggers
- reusing the corpus for creating lexical data from that time period (ABaC:us could serve as a fundament for generating an expandable computational lexicon)
- and more sophisticated and complex linguistic research questions (such as identifying stylistic features and rhetorical patterns described in the next paragraph).

4 Identifying rhetorical patterns based on PoS tags

The Baroque literature of the corpus is full of stylistic patterns and a start has been made to identify those typical "abrahamic" features in the texts by using the applied PoS annotation as a basis. The figures below document how the search for sequences of PoS tags was conducted in order to get results that can be described as often recurring stylistic patterns such as ostensive comparisons (Figure 8, e.g. *wie ein Lambel von den Wölffen* / as a lamb among

[12] See Resch, Claudia, Declerck, Thierry, Krautgartner, Barbara and Czeitschner, Ulrike. 2014. *ABaC:us revisited – Extracting and Linking Lexical Data from a historical Corpus of Sacred Literature.* In: Atwell, Eric, Brierley, Claire and Sawalha, Majdi (eds.): *Proceedings of the 2nd Workshop on Language Resources and Evaluation for Religious Texts / LREC 2014,* p. 36-41, particularly chapter 2.1. "Improvement through reliable data".

[13] Two research assistants have worked independently on the manual correction and a senior researcher supervised their decisions, not only, but especially in cases of doubt.

wolves or *[zittern] wie ein Laub von der Espen* / tremble like an aspen leaf) and pairs of words consisting of a foreign term, a conjunction, and a noun (Figure 9, e.g. *Epilogus vnd Weltschluss* / Epilogus and end of the world or *Fratrum vnd Lay-Brüder* / Fratrum and lay friar).

Although it will not be possible to divide the several works where the authorship of Abraham a Sancta Clara is still in question from those where the author is certain, we hope to uncover the characteristics of Abraham a Sancta Clara's often imitated literary style by analyzing, measuring and counting these significant features. Stylometric methodologies could become relevant by automatically finding and counting distinctive patterns of "abrahamic style" and can enhance our knowledge about the identified features.

1. MW	jener armer Tropff der etlich 30 . Jahr	als ein verlassener Krippel bey dem Schwem=Teuch	zu Jerusalem lage : nichts als W . W .
1. MW	würdiger als die Societet Jesu ? welche	wie ein strahlende Sōn in der Catholischen Kirchen	glantzet / dahero kein Wunder / daß neidige
1. MW	Menschlichen Vrtel nach so wenig reimbte /	als ein Faust auff ein Aug	/ als er von Christo gefragt worden / was
1. MW	Wiederkämpfung der feindlichen Anstoß nicht	wie der Butter an der Sonn	möchten bestehen / auch kan wohl seyn /
1. MW	Scherganten vnd Hebreischen Lothers Knechten /	wie ein Lambel von den Wölffen	/ feindlich angegriffen worden / vnd dise
1. MW	kommet / vnd das eytle Welt=Wesen abflieget /	wie die Mucken auß einer kalten Kuchl	/ so dann wüntscht ihm mancher / sein Kopff
1. MW	nicht also nach dem Brunnquell trachte /	wie ein Weib nach der Schönheit	. Die Heilige Schrifft thut dißfalls
1. MW	dem Hauß zu / vnd erzehlen gantz zitterend	wie ein Laub von der Espen	/ was ihnen begegnet ; was in dem fall
1. MW	/ daß anjetzo die Zähn hervor blecken /	wie einem murrenden Hund an der Ketten	; Kombt herzu / schaut das jenige / was
1. MW	ein Schwindsüchtiger / ein armer Beutel /	wie das arme Teubel für ein Hauß	/ vmb willens ein Gnad zu fischen / vmb
1. MW	Wunderwerck gemacht / daß selbiges Eysen	wie ein Bimbsen auff dem Wasser	geschwommen : Wann man schon einem Advocaten
1. MW	wie das Saltz im Wasser vnd verschwinden /	wie der Schatten an der Sonnen=Uhr	/ wans Abend ist . O / wie mancher
1. MW	Wir elende Adams=Kinder seynd gar offt	wie die Wein=Trauben vnter der Preß	/ wie ein Rosen vnter den Dörner / wie
1. MW	offt wie die Wein=Trauben vnter der Preß /	wie ein Rosen vnter den Dörner	/ wie ein Uhr mit dem schwären Gewicht
1. MW	Preß / wie ein Rosen vnter den Dörner /	wie ein Uhr mit dem schwären Gewicht	/ wie ein Bürckenbaum mit lauter Ruthen
1. MW	müheseeligen Tropffen / der so viel Jahr	als ein verlassener Krippl bey dem Schwemm=Teich	zu Jerusalem kein anders Liedl stimbte
1. MW	der du alldort in dem zerlassene Metall /	wie ein Arbes in einem siedenten Hafen	empor strudlest / was hat dise deine Verdamnuß
1. MW	gäntzlich in die heiligste Wunden JEsu /	wie ein Tauben in die offne Ritzen	deß Felsen verschlossen / hat gleichwohl

Figure 8: Combinations with KOKOM – comparative conjunction / ART – definite or indefinite article / (ADJA – attributive adjective) / NN – common noun / APPR – preposition / ART / (ADJA) / NN in *MERCKS WIENN*

doc#0	vielen vorkommen / es sey der allgemeine	Epilogus vnd Weltschluß	verhanden / es findet sich nicht ein einige
doc#0	Heiligen Evangelio : Serve nequam : Weit andere	Servos vnd Diener	zehlet diser H . Orden / in welchem da
doc#0	werden : Adio ! behüt euch Gott meine liebe	Patres vnd Ordens-Mitbrüder	/ ist mir leyd / daß ich euch wegen meiner
doc#0	weniger als diß / solches Kraut mit Nahmen	Eringion oder Manns=Treu	ist ein Distel / ein Brach=Distel / voller
doc#0	einest ihren andächtigen Gebrauch nach das	Officium oder Tagzeiten	vnser Lieben Frauen auß dem Büchl eyffrigst
doc#0	man mehrer Federbusch als Schein auff den	Chaßkett vnd Peckelhauben	: die grosse Kriegsstuck pflegt man der
doc#0	hätte wollen vnnd sollen ebenmessig aller	Fratrum vnd Lay-Brüder	der Religio sen gedencken / deren sehr
doc#0	Jdioten seynd ein verworffnes Confect . 170	Scienz vnd Wissenschafft	ist sehr nutzlich . 177 Advocaten Lob .

Figure 9: Combinations with FM – material of a foreign language / KON – coordinating conjunction / NN – common noun

The ABaC:us corpus is also a rich source for the study of semantic aspects. "Death" and "dying", a leitmotif of Baroque texts, can be found in numerous variations: examples for the personification of death are *Aschen=Mann* (ash man), *Dieb der Fröhlichkeit* (thief of happiness), *Reuter auf dem fahlen Pferd* (rider on the pale horse). Together with terms and phrases dealing with the "end of life", "dying" and "killing" more than 1700 death-related lexical units have been identified in *MERCKS WIENN*, *TODTEN BRUDERSCHAFT* and *TODTEN-CAPELLE*. To organize this rich terminology and to provide enhanced access to it we currently are working on the creation of controlled vocabularies in SKOS (Simple Knowledge Organization System). These taxonomies will also ease the task of semi-automated semantic annotation of additional texts. Not only can the project group use the data for

various research interests, such as taxonomy building and semantic enrichment by LOD-sources[14], but other researchers can as well.

5 ABaC:us Future

As texts of this era are usually difficult to access, because of issues of fragility and copies scattered across libraries and institutions, availability is an important issue: The value and uniqueness of the fully annotated corpus precisely consists in its online access and already encoded textual knowledge.

In order to guarantee that the processed materials will be re-usable for different purposes, the project team and the technical task force are working on a web-based interface that can be used for further research. ABaC:us was implemented with the publication framework *cr_xq*[15], a further development of the *Scalable Architecture for Digital Editions* (SADE) of the Berlin-Brandenburg Academy of Sciences and Humanities. We strived to keep the system as generic and flexible as possible. All functions of the edition are parameterized with simple XPath expressions that are held in a project-specific configuration. This dynamic architecture allows to manage the data in a very flexible way and to publish any kind of XML-structured text as a web application. With this format, we represent the textual structure of the works (e.g. chapters, front and back matter etc.) as well as the physical structure of the documents (pages). METS also provides the framework to document the corpus' data structure, references metadata records of the single works it is composed of, and acts as a wrapper for the web application's basic configuration that is necessary for rendering.

The user-friendly application (screenshot below) includes a dual display with digital text and parallel facsimile will enable users to obtain different views of the text, provide different kinds of indices and allow for flexible search strategies on both the linguistic (key words, PoS-labels or lemmata) and semantic level. ABaC:us will be a corpus available for a range of communities: literary and linguistic studies, historical research, theology, but may also meet the interest of a broader public audience.

[14] See Czeitschner, Ulrike, Declerck, Thierry, and Resch, Claudia. 2014. *Porting Elements of the Austrian Baroque Corpus onto the Linguistic Linked Open Data Format.* In: Osenova, Petya, Simov, Kiril, Georgiev, Georgi and Nakov, Preslav (eds.): *Proceedings of the Joint Workshop on NLP&LOD and SWAIE: Semantic Web, Linked Open Data and Information Extraction associated with the 9th International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*. Sofia: p. 12-16.

[15] *cr_xq* is built upon proven standards: The edition is described with a METS-Container, a metadata standard developed by the *Library of Congress* that is widely employed by libraries and other cultural heritage institutions to model compound digital objects.

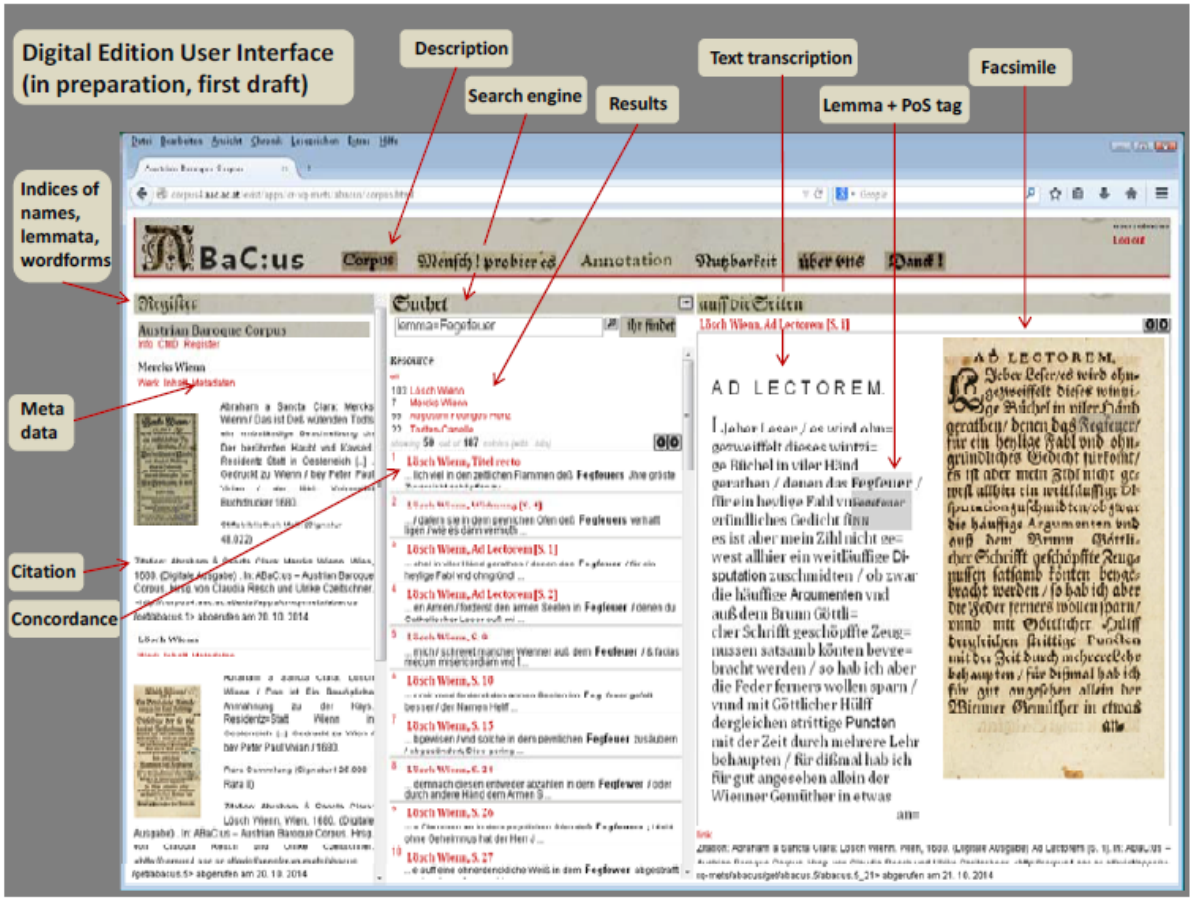

Figure 10: ABaC:us Interface

The annotated ABaC:us data have been selected as an Austrian contribution to CLARIN (Common Language Resources and Technology Infrastructure) and have been released in 2015 as: ABaC:us – Austrian Baroque Corpus, digital edition (2015), edited by Claudia Resch and Ulrike Czeitschner.

6 Conclusion

In this paper, we have given detailed insights into the corpus building process of the Austrian Baroque Corpus ABaC:us, a collection of 20 texts from different authors of the Baroque era (1650-1750). Although German can be seen as a well-documented linguistic variety in terms of language data and tools, data from historical linguistic stages are still scarce and under-explored. Thus, ABaC:us had to be built from scratch, based on thematically connected examples of sacred Baroque literature with the leitmotif of death and dying. The stages of text selection, digitization and transcription have been accurately described. The focus of the paper, however, rests on the annotation process and particularly on the linguistic information. A large part of the corpus – approximately 180,000 tokens (85%), ascribed to the preacher Abraham a Sancta Clara (1644-1709) – contains Part-of-Speech tags and lemma information; challenges during the annotation process are not concealed and our solutions to various problems are illustrated with examples. A showcase analysis of certain rhetorical and semantic features aims to give an idea of the value that the linguistic annotation adds to the corpus. And finally, the brand-new user interface is

introduced, which enables researchers as well as any interested persons to read, examine and analyze the core texts of ABaC:us. The online edition and its accompanying documentation can be found here: https://acdh.oeaw.ac.at/abacus/

References

BOOT, Peter, 2009, Mesotext. Digitised Emblems, Modelled Annotations and Humanities Scholarship. Amsterdam: Pallas Publications – Amsterdam University Press

CZEITSCHNER, Ulrike, DECLERCK, Thierry, MOERTH, Karlheinz and RESCH, Claudia, 2012, Linguistic and Semantic Annotation in Religious Memento Mori Literature. In: ATWELL, Eric, BRIERLEY, Claire and SAWALHA, Majdi (eds.): Proceedings of the LREC 2012 Workshop: Language Resources and Evaluation for Religious Texts. Paris: ELRA, p. 49-52

CZEITSCHNER, Ulrike, DECLERCK, Thierry, and RESCH, Claudia, 2014, Porting Elements of the Austrian Baroque Corpus onto the Linguistic Linked Open Data Format. In: OSENOVA, Petya, SIMOV, Kiril, GEORGIEV, Georgi and NAKOV, Preslav (eds.), 2013, Proceedings of the Joint Workshop on NLP&LOD and SWAIE: Semantic Web, Linked Open Data and Information Extraction associated with the 9th International Conference on Recent Advances in Natural Language Processing (RANLP 2013). Sofia: p. 12-16

EYBL Franz M, 1992, Abraham a Sancta Clara. Vom Prediger zum Schriftsteller. Tübingen: Max Niemeyer Verlag

EYBL, Franz M, 2008, Abraham a Sancta Clara. In: Killy Literaturlexikon Volume 1. Berlin: de Gruyter, p. 10-14

HINRICHS, Erhard, ZASTROW, Thomas, 2012, Linguistic Annotations for a Diachronic Corpus of German. In: Linguistic Issues in Language Technology, Volume 7, issue 7, p. 1-16

RESCH, Claudia, DECLERCK, Thierry, KRAUTGARTNER, Barbara and CZEITSCHNER, Ulrike, 2014, ABaC:us revisited – Extracting and Linking Lexical Data from a Historical Corpus of Sacred Literature. In: ATWELL, Eric, BRIERLEY, Claire and SAWALHA, Majdi (eds.): Proceedings of the 2nd Workshop on Language Resources and Evaluation for Religious Texts / LREC 2014. Reykjavik: p. 36-41

RESCH, Claudia, KRAUTGARTNER Barbara and CZEITSCHNER, Ulrike, (Forthcoming). ABaC:us für LinguistInnen – Morphosyntaktische Annotation im „Austrian Baroque Corpus“. In: RESCH, Claudia and DRESSLER, Wolfgang Ulrich (eds.): Digitale Methoden der Korpusarbeit in Österreich. Tagungsbeiträge der 40. Österreichischen Linguistiktagung. Wien: Verlag der österreichischen Akademie der Wissenschaften

ŠAJDA Peter, 2009, Abraham a Sancta Clara: An Aphoristic Encyclopedia of Christian Wisdom. In: Kierkegaard and the Renaissance and Modern Traditions – Theology. Ashgate 2009. p. 1-20

STOWASSER, 1998, Lateinisch-deutsches Schulwörterbuch von STOWASSER Joseph Maria, PETSCHENIG Michael, SKUTSCH Franz. Auf der Grundlage der Bearbeitung 1979 neu bearbeitet und erweitert (Gesamtredaktion: Fritz Lošek). Zug: HPT-Medien AG

Biography of the authors

The ABaC:us Team consists of several people with different academic backgrounds partly working together since 2010 to build up the Austrian Baroque Corpus:

Claudia Resch is a senior researcher and project leader at the Austrian Centre for Digital Humanities of the Austrian Academy of Sciences. Current research focuses on German literature of the early modern period and the application of literary and linguistic computing in a corpus-based approach to textual issues. Key areas covered are historical linguistics, text stylistics, and annotation problems associated with non-standard varieties of Early Modern German.

Ulrike Czeitschner is a senior researcher and project leader at the Austrian Centre for Digital Humanities. With an academic background in cultural anthropology and a particular interest in the impact of digital technologies on all kinds of humanities studies, she has been focusing on structural and semantic annotations of various genres.

Eva Wohlfarter is a junior researcher at the Austrian Centre for Digital Humanities and a PhD candidate in linguistics at the University of Vienna. She holds a master's degree in Applied Linguistics and her main fields of interest are corpus linguistics, historical linguistics, sociolinguistics, and discourse studies.

Barbara Krautgartner is junior researcher at the Austrian Centre for Digital Humanities. She holds a master's degree in German philology and is currently studying Web- and App-Development in Vienna. Developing scientific web applications and automatic data processing are her special fields of interest.

Acknowledgements

Since the ABaC:us working group insisted on high-quality linguistic and semantic annotation throughout the project, most phases of the project would not have been feasible without external funding:

- Abraham a Sancta Clara and his Danses Macabres (October 2009 – September 2010), research grant by the City of Vienna.
- Text-Technological Methods for the Analysis of Austrian Baroque Literature (March 2012 – September 2014), supported by funds of the Österreichische Nationalbank, Anniversary Fund.
- Mortuary Cult in 17th Century Vienna: Confraternity Studies in the Digital Age (June 2014 – May 2015), supported by funds of the City of Vienna.