Deutsches Textarchiv: From Diverse Text Resources to One Homogeneous High-Quality Text Corpus

DTA Extensions (DTAE)

The module DTAE offers the infrastructure and tools to integrate texts drawn from external sources into the DTA. After the acquisition of text transcriptions together with the respective source images all texts are converted into the DTA ›base format‹, are made available within the DTA corpora and in CLARIN-D.

Scientific Use Case

The Deutsches Textarchiv project (DTA) aims to provide the basis for a reference corpus for the historical (non-standardized) New High German. There are about 1,300 texts to be digitized by the DTA’s internal digitization workflow. In addition, the module DTAE aims at extending the corpus by integrating texts from external sources, e.g. large text compilations like Wikisource, or single documents transcribed by individual scholars.

DTA and CLARIN-D: Towards Sustainable Resources

CLARIN-D is the German section of the CLARIN European research infrastructure federation. It aims at realizing a sustainable, integrated, interoperable and scalable platform for interdisciplinary research in the social sciences and humanities. CLARIN-D develops a web and centres-based research infrastructure, offering language resources, tools and services for German language processing to support language-centered studies. It ensures longterm access to and archiving of domain-specific resources, and applies resp. develops international standards and best practices in order to stay interoperable and sustainable. CLARIN-D is building on the expertise of currently nine service centres in major research institutions:

- University of Tubingen: annotated corpora, linguistic knowledge components and web services
- University of Leipzig: web services and specific reference corpora
- BBAW Berlin: German language, lexica, diachronic corpora
- University of Stuttgart: corpora and corpus tools, adaptable tools and web services
- IDS Mannheim: German language, large corpora for German time (1600–1900) and text types (fiction, functional texts, scientific texts).
- LMU Munich: German speech and multimodal data, phonetic tools and services
- MPI Nijmegen: minority languages, multimedia and multimodal data
- University of Hamburg: multilingual speech corpora, transcription tools
- Saarland University: multilingual corpora and corpus tools
- Saarland University: multilingual corpora and corpus tools

Linguistic Analyses with CAB

All texts integrated into the DTA are analysed by the Cascaded Analysis Broker CAB (cf. Bryan Jurish, JLCL 2010, 25/1). CAB provides a normalization of the historical orthography in order to allow for spelling-tolerant, lemma based corpus searches.

Text Corpora in DTAE

About 255,000 text pages drawn from various external sources are available through the DTA, yet. Further texts are be added continually, e.g. via the CLARIN-D F-AG 1-1 curation project.

Web-based Quality Assurance Platform (DTAQ)

Quality assurance for all new texts added to the DTA corpus takes place within the web-based quality assurance platform DTAQ. In DTAQ texts may be proofread page by page by comparison to their source images. This way, errors can be detected which may have occurred during the transcription and annotation process. During the transcription and annotation of historical texts errors may occur due to different error sources:

- University of Tubingen: annotated corpora, linguistic knowledge components and web services
- University of Leipzig: web services and specific reference corpora
- BBAW Berlin: German language, lexica, diachronic corpora
- University of Stuttgart: corpora and corpus tools, adaptable tools and web services
- IDS Mannheim: German language, large corpora for German time (1600–1900) and text types (fiction, functional texts, scientific texts).
- LMU Munich: German speech and multimodal data, phonetic tools and services
- MPI Nijmegen: minority languages, multimedia and multimodal data
- University of Hamburg: multilingual speech corpora, transcription tools
- Saarland University: multilingual corpora and corpus tools

Data Availability

All DTA texts are freely available (CC by-nc) for download in different formats:

- the original XML/TEI texts,
- an HTML rendered version,
- the raw text transcription (licence CC-0).

Moreover, DC and CMDI metadata comprising TEI header information may be harvested via OAI-PMH.

Facts (April 2013)

- DTA: 786 books, encompassing 288,855 pages (ca. 465M characters, 65M token, 1.6M types)
- DTAE: 12 cooperation partners, ca. 255,000 pages (ca. 515M characters)
- DTAQ: 28,217 pages with ca. 30M characters proofread, 48,886 tickets reported, 34,692 errors corrected

Links and Contact

DTA team: Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Christian Thomas, Frank Wiegand, Kai Zimmer
dta@bbaw.de www.deutschestextarchiv.de