Building a Searchable Bibliographic Database from 26 Printed Volumes

“Turkologischer Anzeiger / Turkology Annual” (TA):
An indispensable systematic bibliography for Turkology and Ottoman Studies

- contributions by experts from all over the world
- funded by several institutions including UNESCO
- edited by the Department of Oriental Studies of the University of Vienna
- 26 printed volumes (around 6,500 pages with about 50,000 entries)
- publicly available in print only; digital pre-print files exist for only 8 volumes

Properties:
- structured by broad categories
- references to books, articles, reviews, and conferences in more than 20 mostly non-Western languages like Russian, Turkish, Arabic, Japanese
- titles in less common languages (e.g. Hungarian, Arabic) are translated
- even single entries may contain chunks in several different languages

Our aims and objectives: Contributing to a modern research infrastructure

- Open Access to scientific knowledge by digitizing all TA volumes and re-publishing the entries in an online database
- added value by offering new, efficient search options in a multi-lingual user interface enhanced with an editing environment

Our challenges and how we address them:
Applied research in computational language processing

- Optical Character Recognition (OCR) software can produce high quality results, if it is provided with lexica and parameterized for specific languages
- but even at 99% OCR precision, a typical TA entry would still contain one or two wrong characters: database searches would not be reliable
- due to the multi-lingual nature of many TA entries and the names and special terms they contain we have to face a serious drop in precision compared to OCR of longer and more homogeneous texts
⇒ develop computational linguistics methods for automatic language identification and language-aware OCR correction

Turkologischer Anzeiger Online: A pilot project

- demonstrating the chances of an Open Access bibliography
- promoting IT usage in the humanities
- developing tailored computational linguistics methods for eHumanities: tasks like OCR, automatic language recognition, named entity recognition are key methods in advanced language technology
- establishing a workflow for future digitizations of multi-lingual bibliographies

http://www.asia-europe.uni-heidelberg.de/research/heidelberg-research-architecture/hra-projects-1/turkologischer-anzeiger-online

properties:
- structured by broad categories
- references to books, articles, reviews, and conferences in more than 20 mostly non-Western languages like Russian, Turkish, Arabic, Japanese
- titles in less common languages (e.g. Hungarian, Arabic) are translated
- even single entries may contain chunks in several different languages

open access
more information:
http://www.open-access.net

Department of Computational Linguistics
Department of Languages and Cultures of the Near East