Improving Access to Text
IMPACT

D-EE3.6 CORE GENERAL LEXICON FOR GERMAN - EXPANDED | EE3

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	Christoph Ringlstetter, Annette Gotscharek	13 July 2011	Created
1.0	Final	Annette Gotscharek	5 December 2011	Incorporating final comments

Approvals

This document requires the following approvals:

Version	Date of approval	Name	Role in project	Signature
0.1	1 November 2011	Katrien Depuydt	Workpackage leader	OK
0.1	5 December 2011	Michael Kranewitter	Tool patron	OK
0.1	2 December 2011	Tomaz Erjavec	Technical reviewer	OK
1.0	8 December 2011	Max Kaiser	Subproject Leader EE	OK
1.0	9 December 2011	Hildelies Balk	Project Director	OK

Distribution

This document was sent to:

Version	Date of sending	Name	Role in project
0.1	1 November 2011	Katrien Depuydt, Michael Kranewitter, Tomaz Erjavec	See above (internal review)
0.1	5 December 2011	Max Kaiser	Subproject Leader EE
1.0	5 December 2011	Hildelies Balk	Project Director
1.0	9 December 2011	Liina Munari	EC Project Officer
		All staff	All project members (Sharepoint)

KB IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands

Technical documentation: D-EE3.6 German Core Lexicon - expanded

1.) Partner

LMU

2.) Deliverable

D-EE3.6 German Core Lexicon - expanded

3.) Background

The deliverable D-EE3.6 German Core Lexicon is delivered as an SQL Database respectively as an XML document according to the IMPACT lexicon format. It is intended to improve both OCR and retrieval for historical German documents. The deliverable D-EE3.6 is an expansion of the lexicon delivered in D-EE3.3. In this second release, the additionally keyed corpus material of Early New High German collaboratively developed by LMU/BSB is included in the background corpus.

The Core General lexicon for German currently relies on the following data:

1. A corpus composed during the first phase of the project from materials of different sources, mainly IDS Mannheim with 2,693,966 tokens and 288,709 types. The corpus ranges from the year 1500 to 1900, thus providing a core around which more specific lexicon data based on selected corpora can be developed.
2. Additionally data for the 16th century had to be keyed because for this period only little groundtruth materials were electronically available in the initially composed general corpus. The delivery of the data was finished by month 25. In collaboration with BSB, we compiled a selection of documents for Early New High German. The BSB-LMU corpus consists of 101 works with 1,766 pages, adding up to approximately 858,000 tokens. The materials were randomly selected from digitized images of the 16th and 17th century collection of the BSB. The materials of the corpus (1) merged with the new extension establish the current historical corpus for lexicon acquisition. It contains 3,552,690 tokens (words in running text) and 369,730 types (unique words). The figure below shows the gains of tokens for each period realized by the additionally acquired corpus materials.

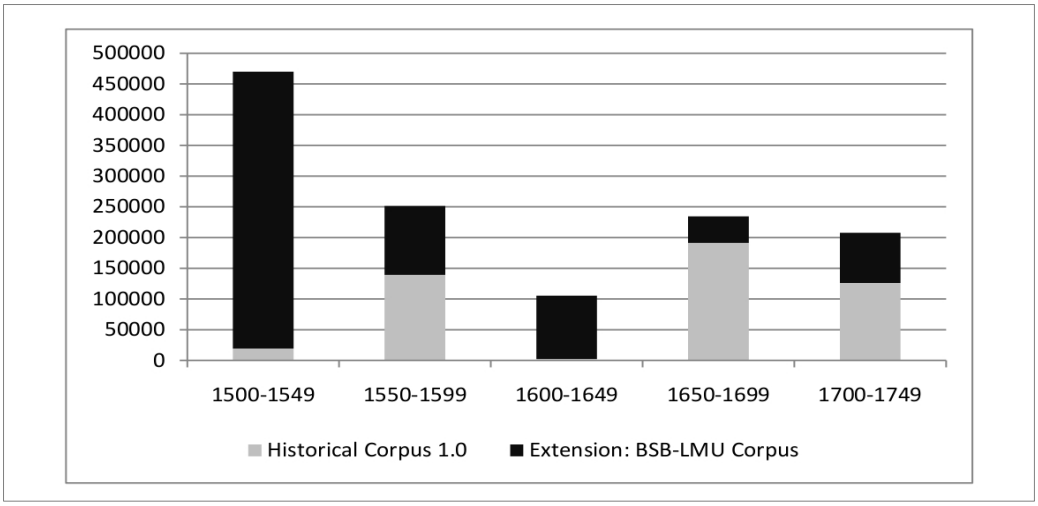

Figure 1: Gains of tokens by the extension of the historical corpus.

4.) Outline of the Functionality

Lexicon development is conducted with the same tool as described in D-EE3.3 available at LMU. This tool supports the processing of both, entries that can be derived automatically as a historical variant and such that do not have an easy pattern induced relation to the modern word. Additionally, functionalities for new corpus integration (i. e. the additional corpus for Early New High German) have been implemented.

For the corpus materials, based on frequency, all non-modern words (recognized at string level with a modern lexicon) have been processed through the following steps:

1. Alignment of historical wordform and modern wordform supported by a matching algorithm with rewrite rules and verified by a trained historical linguist
2. Lemmatization with CISLEX (a modern German lexicon)
3. Attestations to verify corpus occurrence and to resolve lemmatization ambiguities

With this approach until month 42 a historical lexicon of 22,800 non modern entries with attestations has been built on the available corpus materials. The lexicon contains 20,700 different historical strings, this means we found attestations for approximately 1.1 different readings of a string. All together 36,800 readings have been manually marked as feasible but 14,000 of them could not be verified in the corpus. From all processed 36,800 readings 31,700 are pattern based, 5,100 are “irregular”. These 36,800 readings point to 19,200 lemmata.

The deliverable will be composed of the 22,800 entries that have corpus attestations which were manually checked by a linguist. These 22,800 entries point to 11,600 lemmata.

It should be noted that this lexicon refers to enrichment in Information Retrieval were each historical wordform comes assigned with a modern lemma. For the use in OCR a wordlist with 270,000 historical wordforms has been compiled from the historical corpus and is available for use of the technical partners.

Output	Delivery method
Lexicon Data	XML export in human readable and database version

Delivery 3.6 : the lexicon is delivered in a format according to the database schema as developed in EE2.

5.) Evaluation

The lexicon in SQL and in XML version has been checked for errors. The structure meets requirements formulated in IMPACT deliverable database structure D.EE2.1.

6.) License and IPR protection

The licensing follows the consortium agreement.

The product will be integrated as a linguistic resource into the OCR workflow, wordlists from the complete historical corpus will be integrated as well into products developed in TR3 and TR5.

So far no resources available by IMPACT are used for the development of the product.