German Eltec Corpus

Contributors

- Leonard Konle
- Carolin Odebrecht
- Fotis Jannidis

and very valuable and helpful feedback by Lou Burnard

How was the work organized

- Fotis selected the texts
- Leo prepared the ELTEC TEI in collaboration with Carolin and feedback by Lou

Current state of the collection

				AUTHORSHIP				LENGTH			TIME SLOT					REPRINT COUNT		
Language	Last update	Texts	Words	Male	Female	1-title	3-title	Short	Medium	Long	1840-59	1860-79	1880-99	1900-20	range	Frequent	Rare	E5C
cze	2020-11-16	16	366626	14	2	12	0	16	0	0	5	6	5	0	6	0	15	33.85
deu	2020-11-15	98	12086096	65	33	36	8	20	37	41	24	24	25	25	1	46	46	93.85
eng	2020-11-21	99	12198190	49	50	69	10	26	27	46	21	22	31	25	10	32	67	97.69
fra	2020-11-15	100	8712219	66	34	58	10	32	38	30	25	25	25	25	0	44	56	101.54
gre	2019-09-22	11	42524	10	1	11	0	11	0	0	0	1	6	4	6	3	4	37.83

How were the texts processed?

- Based on a corpus of digitized German novels bought from a company which sold CD-ROMs with the texts in the 1990ies: ca. 400 novels Source: www.zeno.org
- The collection was encoded in a proprietary XML format, which encoded structural and visual aspects.
- It consists of two groups of texts:
  - a) Canonical texts (all novels by Goethe). These texts are ‚normalized' to modern spelling.
  - b) Texts by women. Most of these texts are not normalized, because only first prints were available.
- The whole collection was converted to TEI a few years ago.
- Conversion to the Eltec format showed some conversion problems which were fixed for the Eltec collection.

How were texts selected?

- We tried to satisfy as many constraints as possible based on the available texts

How did you learn how to use TEI

- ~1993 in a hands-on course by Michael Sperberg-McQueen and Lou (?) in Tübingen
- Leo Konle: In a seminar in the DH program in Würzburg

tools: Oxygen, lxml in Python

What challenges did you encounter?

- Differences in spelling. We still need to do a historical normalization
- Single texts - because the digitization campaign was focusing on complete works
- Reasons for the delay: We halted our work on creating a larger corpus of German novels from the 18th to the 20th Century. We are concentrating on other texts (dime novels) at the moment, because I am more and more sceptical that our methods in distant reading work with more complex literary texts in a satisfying way yet.