Compiling Digital Fragments of *ALZ*, a Major Textual Resource for German Romanticism Research, into a Single Comprehensive Text Corpus

Tinghui Duan[1,2] and **Udo Hahn**[2]

[1] Graduate School "Romanticism as a Model. Variation - Scope - Relevance"
[2] Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany

MODELL ROMANTIK
Variation • Reichweite • Aktualität

DFG Deutsche Forschungsgemeinschaft

Jena University Language & Information Engineering Lab

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

What is *ALZ*?

- *Allgemeine Literatur-Zeitung* (General Literature Gazette, ALZ)
- Published between 1785 and 1849 in Jena/Halle, Germany
- Consists of anonymous reviews of contemporary publications
- Most comprehensive review organ around 1800 in Germany
- Invaluable historical resource for literary research on German romanticism

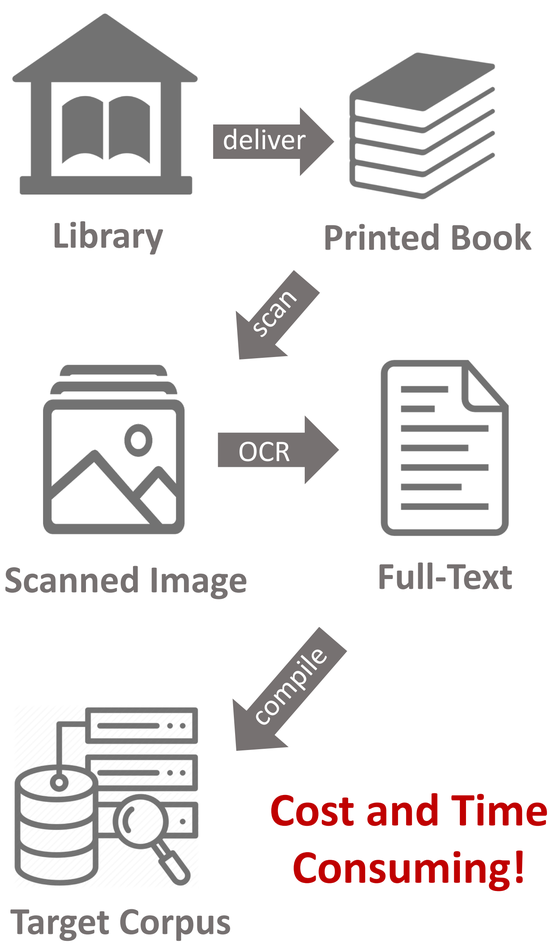
Task

- Create a comprehensive full-text corpus of *ALZ* for research on German romanticism

Scattered Digital Fragments of *ALZ* Hosted by Various Digital Libraries

Collecting

Traditional Workflow



Cost and Time Consuming!

Overview of All Available Full-Text Resources

	Google Books			Internet Archive			Bavarian State Library		
Library	Volumes	Pages	Tokens	Volumes	Pages	Tokens	Volumes	Pages	Tokens
Bavarian State Library	1	285	289,269	-	-	-	152	78,675	76,568,508
University of Lausanne	4	2,008	1,969,395	-	-	-	-	-	-
Harvard University	4	3,056	2,645,386	4	3,054	2,520,832	-	-	-
Indiana University	67	33,721	32,255,992	49	24,923	25,146,053	-	-	-
New York Public Library	103	48,949	47,909,730	94	42,636	42,970,913	-	-	-
Princeton University	38	36,854	35,245,420	29	29,096	30,036,319	-	-	-
Stanford University	2	887	869,272	3	1,102	1,216,273	-	-	-
University of California	45	28,69	28,032,026	54	33,248	33,179,736	-	-	-
University of Illinois	-	-	-	4	3,192	2,820,205	-	-	-
University of Michigan	144	71,681	68,554,109	308	147,559	146,025,563	-	-	-
University of Oxford	46	20,995	20,506,977	56	26,181	25,185,788	-	-	-
Total	**454**	**247,126**	**238,277,576**	**601**	**310,991**	**309,101,682**	**152**	**78,675**	**76,568,508**

Deduplicating

Compiled Corpus (TEI-Format)

Library	Volumes	Pages	Tokens
Bavarian State Library	152	78,675	76,568,508
University of Lausanne	-	-	-
Harvard University	2	1,528	1,322,693
Indiana University	3	2,641	2,547,872
New York Public Library	4	2,08	2,043,980
Princeton University	15	8,038	7,267,803
Stanford University	-	-	-
University of California	9	3,354	3,162,025
University of Illinois	2	1,512	1,299,100
University of Michigan	64	25,538	23,276,002
University of Oxford	8	3,246	2,881,022
Total (≈ 82% of all)	**261**	**126,612**	**120,369,005**

Encoding

Compiling

Deduplicationsformel*

$L \sim Library$
$E \sim OCR\ Engine$
$S \sim Sources \subseteq L \times E$
$F = Fulltexts$
$F \sim \{F_1, F_2, \ldots, F_n\}$
$F_i \sim Duplicates\ of\ text\ i$
$R \subseteq F \times S \sim Text\ exists\ in\ source$
$f^i_{best} = \arg\max_{f \in F_i} \{ocrQuality(s): fRs\}$

Minimum Edit Distance

Manually Generated Ground Truth

Future Work

- Adding Missing Volumes (≈18%)
- OCR Error Correction
- Named Entity Recognition
- Social Network Analysis
- Topic Modeling

Reference: Udo Hahn, Tinghui Duan. Corpus Assembly as Text Data Integration from Digital Libraries and the Web. In: *JCDL '19. Proceedings of the 2019 ACM/IEEE Joint Conference on Digital Libraries*, Urbana-Champaign, IL, USA, June 02–06, 2019. Pages 25-28.
Corpus Available under: https://github.com/julielab/alz
* The authors would like to thank Erik Fässler for his idea to formalize the step of deduplication.

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD
CL 2019
International Corpus Linguistics Conference
Cardiff, Wales, UK
July 22-26, 2019