

Regina Stodden

Automatic German Text Simplification: Data, Evaluation, and Models

Contents

I	Introduction & Foundation	19
1	Introduction	21
1.1	Motivation & Relevance	21
1.2	Research Aims & Contributions of this Thesis	24
1.2.1	Research Questions	24
1.2.2	Background of Complexity and Simplification	25
1.2.3	Building Text Simplification Corpora	25
1.2.4	Resources for Text Simplification	26
1.2.5	Evaluation of Text Simplification	26
1.2.6	Text Simplification Models	27
1.3	Structure of the Thesis	28
2	Complexity and Simplification	31
2.1	Background of Simplification	31
2.1.1	Comprehension & Comprehensibility & Readability	31
2.1.2	Linguistic Complexity	32
2.1.3	Text Simplicity & Text Complexity & Text Difficulty	32
2.1.4	Simplification Operations & Simplification Rules	32
2.1.5	Language & Literacy Levels	33
2.1.6	Clear, Plain, Simple, Easy, and Simplified	34
2.2	Automatic Text Simplification	35
2.2.1	Text Units of Simplification	36
2.2.2	Text Simplification Workflow	36
2.2.3	Subtasks of Text Simplification	40
2.2.4	Simplification Purposes	41
2.2.5	German Simplification	45
2.3	Summary & Outlook	47
3	Building Text Simplification Corpora	49
3.1	One vs. Many Target Simplifications	49
3.2	Comparable vs. Parallel Corpora	50
3.3	Building Process	51
3.4	Finding Suitable Data (Component A & B)	52
3.5	Manual Simplification (Component C)	53
3.6	Alignment (Component C)	53
3.6.1	Alignment	53
3.6.2	Alignment Types	54
3.7	Automatic Alignment (Component C)	55

3.8	Simplification Plans (Component D)	57
3.9	Annotation of Simplification Operations and Quality Assessment (Component F)	57
3.9.1	Simplification Operations	58
3.9.2	Simplification Quality Assessment	59
3.10	Annotation Interfaces	59
3.10.1	Interfaces for Sentence-wise Alignment	60
3.10.2	Interfaces for Annotation of Simplification Operations	60
3.10.3	Interfaces for Annotation of Simplification Quality	60
3.11	Summary & Outlook	61
3.11.1	Challenges & Research Gaps	61
3.11.2	Outlook	62
4	German Simplification Corpora	63
4.1	Corpora with Web Texts	64
4.1.1	German Resources	65
4.1.2	Simple German Web Corpus '13	66
4.1.3	Simple German Web Corpus '20	66
4.1.4	Simple German Web Corpus '23	68
4.1.5	BiSECT	69
4.1.6	Capito Corpus	70
4.1.7	HDA-Leichte-Sprache-Corpus & GEASY & DE-Lite	70
4.1.8	Semi-synthetic Simple German Web Corpus	71
4.1.9	Miscellaneous	72
4.2	Corpora with Wikipedia Texts & Knowledge Acquisition Texts	72
4.2.1	Non-German Corpora	72
4.2.2	German Resources	73
4.2.3	Translated Wikipedia Corpus	74
4.2.4	Translated ASSET	75
4.2.5	Lexica Corpus and Klexikon	75
4.2.6	TextComplexityDE	76
4.2.7	GEolino	77
4.3	Corpora with News Texts	78
4.3.1	Non-German Corpora	78
4.3.2	German Resources	79
4.3.3	German News Corpus	80
4.3.4	APA-LHA	80
4.3.5	APA-RST	82
4.3.6	20Minuten	82
4.4	Corpora with Medical & Health Texts	83
4.4.1	Non-German Corpora	83
4.4.2	German Resources	84
4.4.3	Simple-Patho	84
4.5	Corpora with Political & Legal Texts	85
4.5.1	German Resources	85
4.5.2	ABGB	86
4.5.3	Online Participation	86
4.6	Corpora with Narratives Texts	87
4.6.1	Non-German Corpora	87
4.6.2	German Resources	87
4.6.3	GNATS	88
4.7	Non-Parallel Corpora	88
4.7.1	Lexical Simplification Data	88

4.7.2	Syntactical Simplification Data	89
4.7.3	Monolingual Data	89
4.8	Data Augmentation	90
4.8.1	Word Replacement	91
4.8.2	Translation & Round-Trip Translation	92
4.8.3	Monolingual Data	92
4.9	Summary & Outlook	93
4.9.1	Challenges & Research Gaps	93
4.9.2	Outlook	99
5	Text Simplification Evaluation	101
5.1	Manual Evaluation	102
5.1.1	Intrinsic vs. Extrinsic Evaluation	102
5.1.2	Evaluation Aspects (Intrinsic)	103
5.1.3	Overall Quality of Simplicity	109
5.1.4	Datasets with Human Judgments	110
5.2	Automatic Evaluation	112
5.2.1	Evaluation Aspects	113
5.2.2	Overall Quality of Simplicity	123
5.2.3	Reference-less Metrics	124
5.2.4	EASSE: Evaluation Framework	124
5.3	German Evaluation Studies	124
5.3.1	Manual Evaluation	125
5.3.2	German Datasets with Human Judgments	128
5.3.3	Automatic Evaluation	128
5.4	Summary & Outlook	130
5.4.1	Challenges	131
5.4.2	Outlook	135
6	German Text Simplification Models	137
6.1	Chronicle of English TS Models	138
6.2	Rule-based Models	139
6.2.1	Rule-based Model by Suter et al. (2016)	139
6.2.2	hda-etr	139
6.2.3	DISSIM	140
6.3	Training Sequence-to-sequence Models – Sockeye	140
6.3.1	Sockeye-benchmarking	141
6.3.2	Sockeye-APA-LHA	141
6.4	Fine-tuning Sequence-to-Sequence Models	142
6.4.1	mBART	143
6.4.2	mT5	146
6.5	Prompting with Zero- & Few-shot Learning on Auto-regressive Models	148
6.5.1	ZEST	149
6.5.2	GUTS	150
6.5.3	BLOOM-zero, BLOOM-sim-10, & BLOOM-random-10	150
6.5.4	BLOOM-BiSECT	152
6.5.5	ChatGPT	152
6.6	(Fine-tuning) Auto-regressive Language Models	154
6.6.1	customer-decoder-ats	154
6.6.2	GPT-2 & LeoLM	155
6.7	Proprietary TS Models & Real World Application	155
6.7.1	Proprietary TS Models	155

6.7.2	Use Cases of Text Simplification in Real World Applications	156
6.8	Summary & Outlook	157
6.8.1	Challenges & Research Gaps	157
6.8.2	Outlook	162
II	Publications	165
7	Publications	167
7.1	Overview of the Chapter	167
7.2	Complexity & Simplification	168
7.2.1	A multi-lingual and cross-domain analysis of features for text simplification.	169
7.2.2	RS_GV at SemEval 2021 Task 1: Sense Relative Lexical Complexity Prediction	170
7.3	Building Text Simplification Corpora	171
7.3.1	Creation of a parallel simplification corpus – Using the annotation tool TS-anno	172
7.3.2	TS-ANNO: An annotation tool to build, annotate and evaluate text simplification corpora	173
7.4	German Simplification Corpora	174
7.4.1	Accessibility and comprehensibility of user-generated content: Challenges and changes for easy-to-read languages	175
7.4.2	DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification	176
7.5	Text Simplification Evaluation	177
7.5.1	When the scale is unclear – analysis of the interpretation of rating scales in human evaluation of text simplification	178
7.5.2	HHUplexity at text complexity DE challenge 2022	179
7.5.3	EASSE-DE & EASSE-multi: Easier Automatic Sentence Simplification Evaluation for German & Multiple Languages	180
7.5.4	Overview of the GermEval 2024 Shared Task on Statement Segmentation in German Easy Language (StaGE)	181
7.6	German Text Simplification Models	182
7.6.1	Reproduction & Benchmarking of German Text Simplification Systems	183
7.6.2	Can Text Simplification Help to Increase the Acceptance of E-Participation?	184
III	Discussion & Conclusion	185
8	Discussion & Future Works	187
8.1	Overview of the Chapter	187
8.2	Complexity & Simplification	187
8.2.1	RQ 2-1: German Simplification Operations	189
8.2.2	RQ 2-2: Identification and Explanation of Complex Texts	191
8.3	Building Text Simplification Corpora	192
8.3.1	RQ 3-1: Corpus Creation Challenges	192
8.3.2	RQ 3-2: Characteristics of new Corpora and RQ 3-3: Quality & Representativeness of Corpora	198
8.4	German Simplification Corpora	201
8.4.1	RQ 4-1: Missing Domains	202
8.4.2	RQ 4-2: New Data	203

8.5	Text Simplification Evaluation	213
8.5.1	RQ 5-1: Robust Evaluation	213
8.5.2	RQ 5-2: Multi-lingual Evaluation	214
8.5.3	RQ 5-3: New Aspects for Evaluation	216
8.5.4	Recommendations	218
8.6	German Text Simplification Models	219
8.6.1	RQ 6-1: Document & Sentence Simplification	221
8.6.2	RQ 6-2: Effect of Data & Models	221
8.6.3	RQ 6-3: Effect on Real-World Application	228
9	Limitations	231
9.1	Automatic Text Simplification	231
9.2	Simplification of Written Texts	231
9.3	Supporting the Target Group via Post Editing	232
9.4	Simplification on Document and Sentence Level	233
9.5	Evaluation of Automatic Text Simplification	233
10	Conclusion	235
11	Ethics & Impact Statement	239
	References	241
	Plain Text Summary	275
	Appendices	

List of Figures

1.1	Workflow diagram of text simplification including the corpus building process. . .	23
1.2	Contributions of my thesis including corresponding chapters and publications. . .	28
2.1	Workflow diagram of text simplification with the corpus building process (same as Figure 1.1).	37
3.1	Corpus building process (Cutout of the whole TS workflow of Figure 1.1).	51
3.2	Crossing alignment in a news document pair of the Austrian Press Agency (Title: “News from June 21, 20191”; ID = 403).	55
4.1	Common German sentence simplification corpora including alignment types and domains.	95
4.2	Corpus sizes of German sentence simplification corpora.	97
4.3	Target groups of all German TS corpora.	98
7.1	Contributions of this thesis including chapters and publications (same as Figure 1.2).	167
8.1	Text simplification workflow including contributions of this thesis (extended version of Figure 1.1).	188
8.2	Typology of German simplification operations.	190
8.3	Alignment types in different German TS corpora.	195
8.4	Alignments between complex document (left) and simple document (right). The full document texts can be found in the Appendix.	196
8.5	Corpus building process including the DEplain corpora.	205
8.6	Comparison between previous corpora and our DEplain corpora (extension of Figure 4.1).	206
8.7	Ratings of meaning preservation and grammaticality per domain.	210
8.8	Ratings of simplicity type and coherence per domain.	210
8.9	Simplification operations per simple-complex pair in DEplain-APA and DEplain-web (in %).	212
8.10	Simplicity ratings in five English test sets with different rating scales.	214
10.1	Text simplification workflow including contributions of this thesis (same as Figure 8.1).	236

List of Tables

2.1	Language varieties.	42
2.2	Simplification purposes. The length of the bars indicates the degree of simplification. All URLs have lastly been accessed at July 24, 2024.	46
3.1	Automatic alignment methods. Extended Table of Table 1 in Spring et al. (2023).	56
4.1	Web crawler of websites with texts in simplified German. The crawler marked with † only extract simplified texts and no parallel text pairs. Last part shows own contributions. All URLs have lastly been accessed at July 24, 2024.	65
4.2	Resources for German web TS corpora without own contributions. The line separates German Plain (PL) and Easy Language (EL). OG = Old German, SG = Standard German. All URLs have lastly been accessed at July 24, 2024.	67
4.3	Characteristics of the document simplification corpus GEASY. The Table is based on Table 3 in Hansen-Schirra et al. (2020b).	71
4.4	Characteristics of the document simplification corpus Klexikon. Scores are based on Table 4 in Aumiller and Gertz (2022). Word length in characters.	76
4.5	Characteristics of the document simplification corpus Lexica-Corpus. Scores are based on Table 1 in Hewett and Stede (2021).	76
4.6	Characteristics of the sentence simplification corpus TextComplexityDE. Own calculation. Sentence length in Tokens. Word length in syllables.	77
4.7	Characteristics of the sentence simplification corpus GEOLino. Own calculation. Sentence length in tokens. Word length in syllables.	78
4.8	Characteristics of the sentence simplification corpus APA-LHA OR-A2. Own calculation. Sentence length in tokens. Word length in syllables.	81
4.9	Characteristics of the sentence simplification corpus APA-LHA OR-B1. Own calculation. Sentence length in tokens. Word length in syllables.	82
4.10	Characteristics of the sentence simplification corpus APA-RST. Extended version of Table 3 in Hewett (2023) with additional own calculations.	82
4.11	Characteristics of the document simplification corpus 20Minuten. Own calculation. Sentence length in tokens. Word length in syllables.	83
4.12	Characteristics of the paragraph simplification corpus Simple-Patho. The values are copied from Trienes et al. (2022).	85
4.13	Characteristics of the paragraph simplification corpus Online Participation Corpus. Own calculation.	85
4.14	Corpora with non-parallel simplified German data.	89
4.15	Characteristics of simplified German resources per web crawler. PL = German Plain Language, EL = German Easy Language. All URLs have lastly been accessed at July 24, 2024.	91

4.16	Summary of German document, paragraph, and sentence simplification corpora without own work. The lines separate the domains of the corpora. EL = German Easy Language, PL = German Plain Language. All URLs have lastly been accessed at July 24, 2024.	94
5.1	Summary of scales, their descriptions and names per rating aspect in TS human evaluation studies.	104
5.2	Evaluation dimensions, scales, raters (CW = crowd workers), and number of sources per dataset. Extended version of Table 1 in Stodden (2021c).	111
5.3	Scoring example of BERTScore of four system outputs (see item 1 to item 6).	116
5.4	Scoring example of SARI of four system outputs (see item 1 to item 6).	121
5.5	Names of evaluation aspects per German TS study. Last part shows own contributions. Studies marked with * do not contain intrinsic evaluation of German TS.	125
5.6	Scale points per German human TS evaluation study. All scales are described as Likert-Scales except for the ones marked with †, for these each scale point is labeled content-wise. Last part shows own contribution.	126
5.7	Statements and questions per evaluation aspect and German human TS evaluation study. Studies marked with * provide German (and English) statements. Last part shows own contributions.	127
5.8	Number of participants and test set size per German human TS evaluation study. The participants in all studies are German native speakers. Last part shows own contributions.	127
5.9	Automatic scores used per German TS study. R-1 = Rouge-1, R-2 = ROUGE-2, R-L = ROUGE-L. Studies marked with * have not been automatically evaluated. Last part shows own contributions.	129
5.10	Summary of automatic metrics. The vertical lines separates the metrics by their aspects they belong to most: 1) meaning preservation, 2) grammaticality, 3) simplicity, 4) overall simplicity, 5) other.	132
6.1	Comparison of SARI scores of TS approaches by Ryan et al. (2023). Evaluated on GermanNews, a small version of TextComplexityDE, and a small version GE-Olino. Best results are highlighted in bold face.	152
6.2	Comparison of BLEU, SARI, and ROUGE-L scores of TS approaches by Rios et al. (2021) and Anschütz et al. (2023). Evaluated on 20min. Best results are highlighted in bold face.	155
6.3	Summary of German TS models (without own work). Each line separates different model approaches. Extended version of Stodden (2024b). All URLs have lastly been accessed at July 24, 2024. Part I (continued on next page).	158
6.4	Summary of German TS models (without own work). Each line separates different model approaches. Extended version of Stodden (2024b). All URLs have lastly been accessed at July 24, 2024. Part II (continued from previous page).	159
8.1	Inter-Annotator agreement per domain including average, standard deviation, number of sentence combinations (# sents), and number of documents (# docs). Copied from Stodden et al. (2023).	197
8.2	Results of the alignment methods with 1:1 (upper part) and $n:m$ capabilities (lower part) on sentence pairs with 1:1 ($n=1750$, left part) and $n:m$ alignments ($n=991$, right part) wrt. precision (P), recall (R), F1 score (harmonic mean of P&R), and $F_{0.5}$ score (more emphasis on P than R). Copied from Stodden et al. (2023).	197
8.3	Results of MASSAlign on DEplain-web and DEplain-APA.	198

8.4	Rating aspects.	200
8.5	Summary of German document, paragraph, and sentence simplification corpora including own work (last part). The lines separate the domains of the corpora. EL = German Easy Language, PL = German Plain Language. All URLs have lastly been accessed at July 24, 2024. Extended version of Table 4.16.	204
8.6	Resources for German web TS corpora including own contributions (last column). The line separates German Plain (PL) and Easy Language (EL). OG = Old German, SG = Standard German. All URLs have lastly been accessed at July 24, 2024. Extended version of Table 4.2.	209
8.7	Scores of identity baseline on three German test sets when using different language settings and tokenizers. Copied from Stodden (2024a).	215
8.8	Results of StaGE shared task subtask 1. Copied from Schomacker et al. (2024).	217
8.9	Overview of test sets for German sentence simplification which are included in EASSE-DE. Extended version of Table 1 in Stodden (2024a).	220
8.10	Summary of German TS models including own work. Each line separates different model approaches. Extended version of Stodden (2024b) and Table 6.3. All URLs have lastly been accessed at July 24, 2024. Part I (continued on next page).	222
8.11	Summary of German TS models including own work (last part). Each line separates different model approaches. Extended version of Stodden (2024b) and Table 6.3. All URLs have lastly been accessed at July 24, 2024. Part II (continued from previous page).	223
8.12	Results on Document Simplification using finetuned long-mBART. n corresponds to the length of the training data. Copied from Table 4 in Stodden et al. (2023).	224
8.13	Results on Document Simplification Testing on 20min with long-mBART. Copied from Table 15 of Stodden et al. (2023).	225
8.14	Comparison of DEplain-mBART models on DEplain test sets.	227
8.15	Evaluation on DEplain-APA.	227