DEplain: A German Parallel Corpus with Intralingual Translations into Plain Language for Sentence and Document Simplification

Regina Stodden, Omar Momen, Laura Kallmeyer[1]

[1]Heinrich Heine University Düsseldorf, Germany

regina.stodden@hhu.de

Text Simplification

Aim

- Adaption of a given text to improve the text comprehension for a target group:
  - Non-native (German) speakers,
  - People with reading problems.

How to Simplify

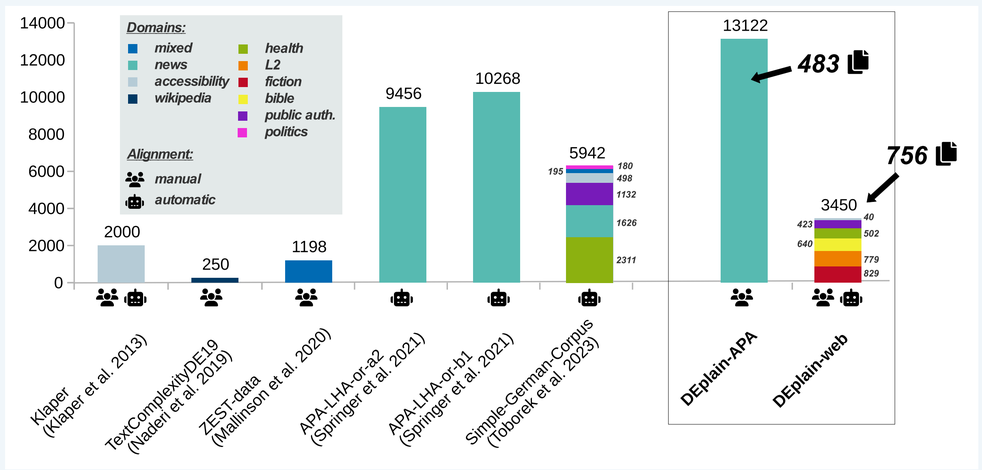
- Sentence Split,
- Complex Word Substitution,
- Rephrasing,
- Compound Segmentation, etc.

Evaluation

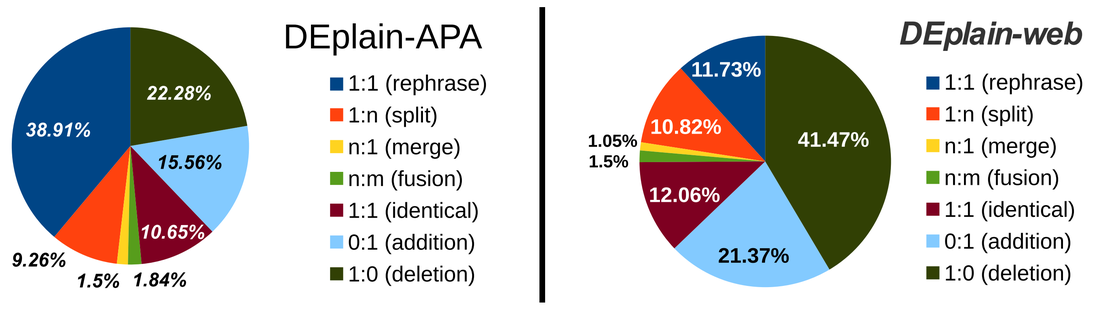
Manual	Automatic
Simplification	SARI & BERT-Score
Readability	FRE
Meaning Preservation	BLEU

German Text Simplification Corpora

Sentence Level



Alignment Statistics



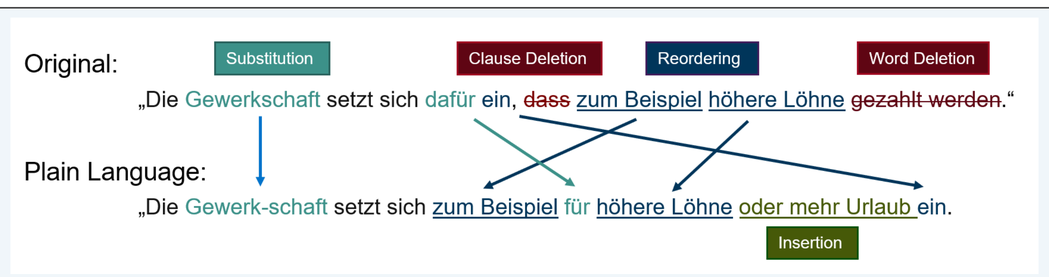
Use Cases of DEplain

Automatic Alignment Evaluation

Results of the alignment methods with 1:1 (upper part) and n:m capabilities (lower part)

		1:1				n:m			
Name	**Description**	**P**	**R**	**F_1**	**$F_{0.5}$**	**P**	**R**	**F_1**	**$F_{0.5}$**
LHA	Hierarchical alignment using sentence embeddings similarity	.94	.41	.57	.747	-	-	-	-
Sent-LaBSE	Similar embeddings of Language-agnostic BERT transformer	**.961**	.444	.608	**.780**	-	-	-	-
Sent-RoBERTa	Similar embeddings of Cross English & German RoBERTa	.960	.444	.607	.779	-	-	-	-
CATS-C3G	Different similarity measures e.g n-grams (C3G)/word vectors	.247	**.553**	.342	.278	-	-	-	-
VecAlign	Multilingual aligner based on mulitlingual sentence embeddings	.271	.404	.323	.290	.260	.465	.333	.285
BERTalign	Allows sentence-transformer methods produce n:m alignments	.743	.465	.572	.664	.387	.561	.458	.412
MASSalign	A vicinity-driven approach with a TF-IDF similarity matrix	.846	.477	**.610**	.733	**.819**	.509	**.628**	**.730**

Text Simplification Example



English: The trade union advocates that higher wages be paid, for example.
Plain English: The trade union advocates for higher wages or more vacation, for example.

Automatic Text Simplification

Document Level

Results on Document Simplification using finetuned long-mBART.
n corresponds to the length of the training data.

train data	n	SARI ↑	BLEU ↑	BS-P ↑	FRE ↑
DEplain-APA	387	**44.56**	**38.136**	**0.598**	**65.4**
DEplain-web	481	35.02	12.913	0.475	59.55
DEplain-APA+web	868	42.862	36.449	0.589	65.4
src2Src-baseline		17.637	34.247	0.583	58.85

Table 5: DEplain-APA test (n=48)

train data	n	SARI ↑	BLEU ↑	BS-P ↑	FRE ↑
DEplain-APA	387	43.087	21.9	0.377	**64.7**
DEplain-web	481	49.584	23.282	**0.462**	63.5
DEplain-APA+web	868	**49.745**	**23.37**	0.445	57.95
src2Src-baseline		12.848	23.132	0.432	59.4

Table 7: DEplain-web test (n=147)

Sentence Level

Results on Sentence Simplification using finetuned mBART.
n corresponds to the length of the training data.

train data	n	SARI ↑	BLEU ↑	BS-P ↑	FRE ↑
DEplain-APA	10660	34.818	28.25	0.639	**63.072**
DEplain-APA+web	11941	**34.904**	**28.506**	**0.64**	62.669
src2src-baseline		15.249	26.893	0.627	59.23

Table 9: DEplain-APA test (n=1231)

train data	n	SARI ↑	BLEU ↑	BS-P ↑	FRE ↑
DEplain-APA	10660	30.867	15.727	0.413	64.516
DEplain-APA+web	11941	**34.828**	**17.88**	**0.436**	**65.249**
src2src-baseline		11.931	20.85	0.423	60.825

Table 11: DEplain-web test (n=1846)

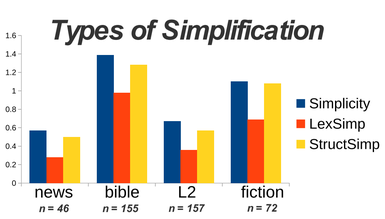
DEplain Statistics

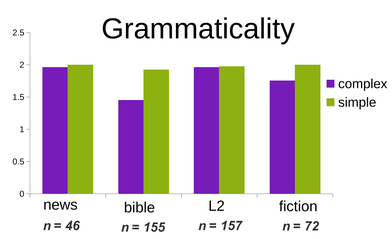
Corpus Size

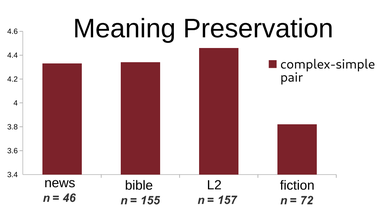
Name	License	# Doc. Pairs	# Original Sents	# Simple Sents.	Alignment	# Sent. Pairs
DEplain-APA	upon request	483	25,607	26,471	manual	13,122
	open	147	6,138	6,402	manual	1,846
DEplain-web	open	249	7,087	7,760	auto	652
	closed	360	12,847	18,068	auto	942
In total	mixed	1,239	51,681	58,701	mixed	16,562

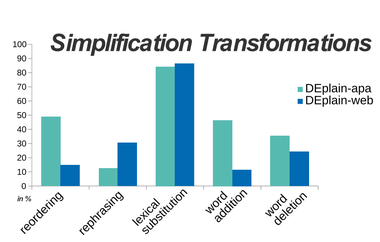
Rating Statistics

We manually rated some of the sentence-wise aligned pairs to get insights into the corpus.









Future Works

- Human evaluation of the ATS outputs
- Improve automatic alignment methods
- Scale up the web crawler

Limitations

- Different data sources licenses
- Unreliable n:m alignment methods
- Absence of alternative references in test sets