Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 288–297.

Sowmya Vajjala Balakrishna. 2015. *Analyzing text complexity and text simplification: Connecting linguistics, processing and educational applications*. Ph.D. thesis, University of Tübingen.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 12–22.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2872–2881.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 584–594.

A Details of Sentence Selection

Dependence on external factors makes the sentence-level-assessment problem ill-formed. This phenomenon was noticed in (Jacob and Uitdenbogerd, 2019): linguistic features that are typically well-correlated with document readability were poorly correlated with it in tweets, which inevitably depend on external factors. To avoid this problem, we carefully selected stand-alone sentences for annotation.

For Wiki-Auto, we excluded the first paragraphs of an article to avoid dictionary-definition-like sentences, *e.g.*, ‘X is the capital of country Y’. While we excluded sentences containing named entities recognised by Stanza, we allowed named entities of types of DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, and CARDINAL, as well as those in a list that we manually prepared containing names of well-known regions, countries, and cities (*e.g.*, Europe, France, and Paris), and common personal names (*e.g.*, William). Finally, we regularised spellings to the American forms using the localspelling library.[17]

B Details of Corpus Splitting

First, we computed the cosine distances between all pairs of sentence embeddings obtained using a pretrained Sentence-BERT model (Reimers and Gurevych, 2019).[18] Next, the average cosine distance for each sentence was calculated. The sentences were then allocated to the test, validation, and training sets according to the descending order of their average cosine distances. Thus, sentences with the least similarity to other sentences were allocated to the test and validation sets, and the rest to the training set.

C Hyperparameter Settings

For all models, the loss weighting factor α was searched in the range $[0.1, 1.0]$ with 0.1 interval. For neural network models, the learning rate was searched in the range $[1e-5, 7e-5]$ with $1e-5$ interval. For the BoW baseline using support vector machines, the kernel was chosen from linear or radial basis function networks, and the regularisation parameter γ was searched in the range $[0.01, 100]$ by loguniform sampling of 40 points. Table 7 presents the hyperparameter settings of the proposed and BERT baseline models, Table 8 those of the BoW baseline.

[17] https://github.com/fastdatascience/localspelling

[18] Specifically, we used `all-mpnet-base-v2`, which had the highest performance at https://www.sbert.net/docs/pretrained_models.html.

		Learning Rate	α
BERT baseline	w/o lossW	$6.0e-5$	–
		$3.0e-5$	0.4
Proposed	w/o lossW	$3.0e-5$	–
	w/o init	$1.0e-5$	0.2
		$1.0e-5$	0.2

Table 7: Hyperparameter settings of the proposed and BERT baseline models

		Kernel	γ	α
BoW	w/o lossW	linear	4.6	–
		linear	0.7	0.3

Table 8: Hyper-parameter settings of the Bag-of-Words baseline

D Hyperlinks to Libraries

Here we list hyperlinks to the libraries used in implementation.

PyTorch `https://pytorch.org/`

Lightning `https://www.pytorchlightning.ai/`

Transformers `https://huggingface.co/docs/transformers/index`

scikit-learn `https://scikit-learn.org/`

Optuna `https://optuna.readthedocs.io/`