JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 11 Number 2 | December 2026: 289-303
DOI: 10.15575/join.v11i2.1792

Text Simplification Research: A Systematic Literature Review on Languages, Datasets, and Modeling Trends

Lisnawita[1,2], **Juhaida Abu Bakar**[2], **Ruziana Mohamad Rasli**[3]
[1]Faculty of Computer Science, Universitas Lancang Kuning, Indonesia
[2]Data Science Research Lab, School of Computing, Universiti Utara Malaysia, Malaysia
[3]School of Multimedia Technology and Communication, Universiti Utara Malaysia, Malaysia

Article Info

Article history:

Received August 6, 2025
Revised January 2, 2026
Accepted January 7, 2026
Published September 18, 2026

Keywords:

Lexical Simplification
Low-Resource Language
Readability
Systematic Literature Review
Text Simplification
Transformer

ABSTRACT

Text simplification (TS) is a Natural Language Processing (NLP) task that reduces linguistic complexity while preserving meaning and supports accessibility for different reader groups (e.g., children, second-language learners, and readers with cognitive or reading difficulties). TS studies are diverse in languages, datasets, and methods. Therefore, a consolidated overview is needed to summarize trends and guide future research. This study conducts a Systematic Literature Review (SLR) of TS research published from 2015 to 31 July 2025 to map trends in languages, target audiences, datasets/corpora, techniques, models/algorithms, and frameworks. Relevant literature was searched using IEEE Xplore, ScienceDirect (Elsevier), SpringerLink, and Google Scholar, applied predefined inclusion–exclusion criteria, and screened studies using a PRISMA-style process. The final set includes 75 primary studies. The results show that English dominates TS research, commonly used resources include Wikipedia-family datasets and Newsela, and modeling trends shift from rule-based and classical machine learning toward transformer-based architectures (e.g., BERT and Seq2Seq) and large language models (LLMs). This review offers: (1) a structured overview of TS research dimensions, (2) an evidence-based summary of key resources and modeling trends, and (3) recommendations for future research, particularly for low-resource languages and audience-specific evaluation.

Corresponding Author:

Lisnawita
Faculty of Computer Science, Universitas Lancang Kuning, Indonesia
Jl. Yos Sudarso, KM. 8 Rumbai
Email: lisnawita@unilak.ac.id

1. INTRODUCTION

Text Simplification (TS) is an important subfield of Natural Language Processing (NLP) that aims to reduce linguistic complexity to improve readability and comprehension while preserving the original meaning [1],[2], TS is beneficial for diverse audiences, including second-language learners and non-native speakers[3], readers with lower literacy levels [4], dyslexic children [5], individuals with aphasia and autistic readers [6],[7]. Despite appearing straightforward, TS involves linguistic and

p-ISSN: 2528-1682
e-ISSN: 2527-9165

computational challenges such as preserving semantic adequacy, ensuring grammaticality, and adapting outputs to user-specific needs.

As illustrated in Figure 1, TS-related publications remained substantial from 2015 to 31 July 2025, covering general text simplification as well as its two core sub-tasks (lexical and syntactic simplification). This trend indicates sustained research attention over the last decade; however, the distribution across sub-streams suggests that progress is not uniform, motivating a structured mapping of languages, methods, datasets, and target audiences through a systematic review. These observations are consistent with previous bibliometric analyses that highlight uneven research distribution and the dominance of certain languages and research topics within lexical simplification studies [8]

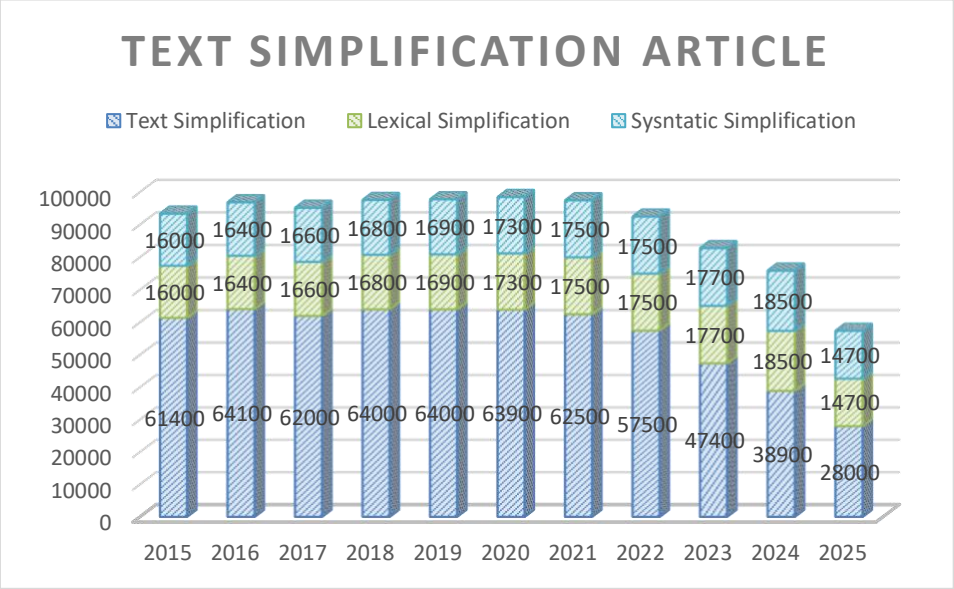


Figure 1. Yearly distribution of TS-related articles (2015–2025) across Text Simplification, Lexical Simplification, and Syntactic Simplification

Existing TS studies can be grouped into three major streams. First, language-coverage and low-resource TS research explores TS beyond English by developing or adapting methods for languages such as Urdu, Italian, Indonesian, and Malay, often facing limitations in corpus availability and evaluation resources [9],[10],[11],[12]. Second, methodological approaches range from rule-based simplification and word sense disambiguation to embedding-based machine learning and neural models. Recent studies increasingly adopt transformer architectures (e.g., BERT and Seq2Seq variants), and more recently, large language models (LLMs), reflecting a shift toward stronger contextual modeling and generation capabilities [13],[14]. Third, audience-oriented TS emphasizes accessibility for specific user groups (e.g., children, L2 learners, and readers with disabilities), yet such studies frequently differ in evaluation protocols, readability criteria, and reporting completeness, making cross-study comparison difficult.

Although TS publications remain consistently high, current research still shows several gaps: (1) imbalanced language coverage with limited attention to non-English and low-resource languages; (2) dependency on a small number of datasets/corpora (e.g., Wikipedia-family datasets and Newsela), which may not represent diverse domains and audiences; and (3) inconsistent reporting of search scope, resources, and experimental settings across studies, which reduces reproducibility and systematic comparison.

Therefore, this paper presents a Systematic Literature Review (SLR) of TS research published from 2015 to 31 July 2025. The contributions of this SLR are: (i) mapping TS studies by language coverage and target audience; (ii) summarizing dominant datasets/corpora and techniques used in TS; (iii) analyzing trends in models/algorithms, including the shift toward transformers and LLMs; and (iv)

identifying research gaps and proposing actionable recommendations for future TS research, particularly for low-resource languages and audience-specific applications.

2. METHOD

2.1 Review Method and Stages

This study applies a Systematic Literature Review (SLR) with three stages: planning, conducting, and reporting[15]. In the planning stage, the review objectives and research questions (RQ1–RQ6) were defined, the databases were selected, and the search strategy and inclusion/exclusion criteria were prepared. In the conducting stage, the searches using predefined keywords were performed, the records were merged, duplicates were removed, studies were screened at the title, abstract, and full-text levels, study quality was assessed using a checklist, and key information was extracted using a standardized form. Finally, the findings were summarized and presented in the selection steps using a PRISMA flow diagram (Figure 2).

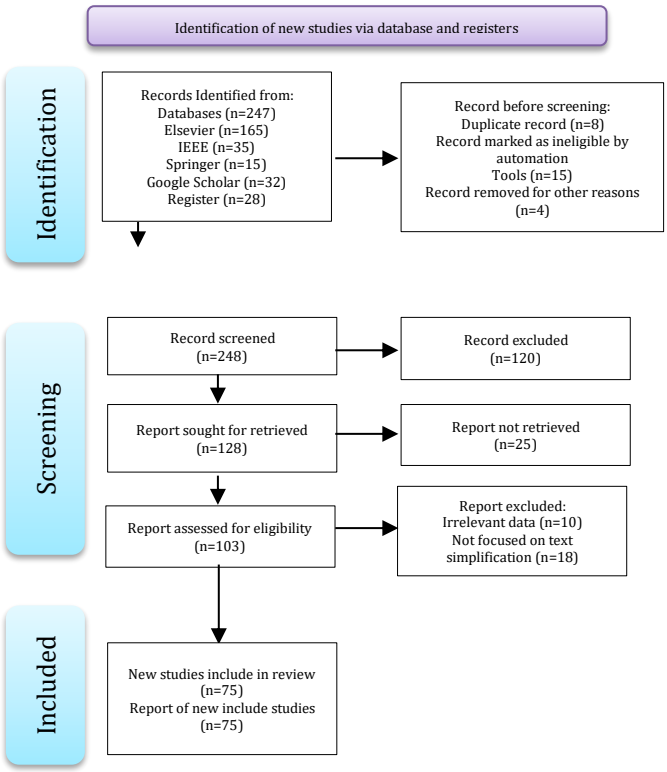


Figure 2. Paper selection process (PRISMA flow diagram)

2.2 Research Questions

Research questions play an important role in scientific research. Details of the research questions to be addressed in this literature review are presented in Table 1.

Table 1. Research Question

No	Question	Main Motivation
RQ1	What languages are most researched in TS?	Identify the language that is most used in TS.
RQ2	What is the Target Audience in TS?	Identification of the target audience used in TS
RQ3	What is a Dataset/Corpus used in TS?	Identification of the Corpus used in TS.
RQ4	What are the Techniques, Approaches, and Linguistic Features used in TS?	Identify the Techniques, Approaches, and Linguistic Features used in TS.
RQ5	What is the Model/Algorithm used in TS Research?	Identify the Model/Algorithm used in TS.
RQ6	What is the tool, system, or framework used in TS Research?	Identify the tool, system, or framework used in TS.

2.3 Data Sources, Time Span, and Article Types

Four digital libraries were searched: ScienceDirect (Elsevier), IEEE Xplore, SpringerLink, and Google Scholar, and additional records from registers. The search covered studies published from 2015 to 31 July 2025. Peer-reviewed journal articles and conference papers with accessible full text that focus on text simplification, lexical simplification, or syntactic simplification were included. Non-research and non-peer-reviewed document types, including book chapters, theses/dissertations, preprints, technical reports/white papers, presentations/slides, posters, tutorials, web pages, and patents, were excluded.

2.4 Search Strategy and Search Strings

Keyword-based searches tailored to each database were used. The core search string was ("text simplification" OR "lexical simplification" OR "syntactic simplification"). All retrieved records were exported, merged, and deduplicated prior to screening.

2.5 Inclusion and Exclusion Criteria

Inclusion criteria (IC):
IC1. Published from 2015 to 31 July 2025
IC2. Peer-reviewed journal or conference paper with accessible full text.
IC3. Focuses on text simplification / lexical simplification / syntactic simplification.
IC4. Provides extractable information relevant to at least one RQ (language, audience, corpus, technique, model, or framework).
Exclusion criteria (EC):
EC1. Not about TS (e.g., readability prediction without simplification objective).
EC2. Non-peer-reviewed or incomplete (abstract-only) items
EC3. Non-article document types (e.g., book chapters, theses/dissertations, preprints, technical reports, posters/slides).
EC4. Duplicate publications of the same study (retain the most complete version).
EC5. Papers without sufficient methodological details for extraction.

2.6 Quality Assessment

Each eligible study was assessed using a quality checklist: QA1. Clear TS objective and scope; QA2. Dataset/corpus and language described; QA3. Method/model explained; QA4. Evaluation/analysis reported; QA5. Transparent reporting. Only studies meeting the minimum quality threshold were included.

2.7 Data Extraction and Synthesis

For each included study, publication year, language(s), target audience, dataset/corpus, technique/approach, model/algorithm, framework/method, evaluation metrics, and main contributions were extracted. The results were synthesized using descriptive statistics and thematic grouping to answer RQ1–RQ6.

3. RESULT AND DISCUSSION

This section answers RQ1–RQ6 and summarizes the main contributions of this SLR. Specifically, an evidence-based mapping of (i) language coverage and target audiences, (ii) datasets/corpora usage, (iii) techniques and models, and (iv) dominant frameworks was provided. Based on these findings, research gaps and practical recommendations for future TS systems, particularly for low-resource languages and audience-specific TS evaluation, were provided.

RQ1. The most researched language for TS

Most research on text simplification has been conducted in English, with a significantly higher number of publications than in other languages, at 34 studies. Furthermore, several other languages

have also been the focus of research, although in much smaller numbers. Languages in the middle group include French (5 studies), German (6 studies), Indonesian (7 studies), Italian (8 studies), and Japanese (9 studies). Meanwhile, Malay appears in 10 studies, followed by Portuguese (11 studies), Russian (12 studies), and Spanish (13 studies). Other languages, such as Swedish, Urdu, Lithuanian, Hindi, Greek, and Turkish, are each only included in 14 to 19 studies. Arabic, Chinese, and Danish rank lowest with only 1 to 3 studies each. This finding indicates an imbalance in research focus on text simplification, with English receiving significant attention, while research on other languages, particularly non-European languages, remains very limited. This indicates the need for efforts to expand text simplification research into various languages so that its benefits are more evenly distributed globally.

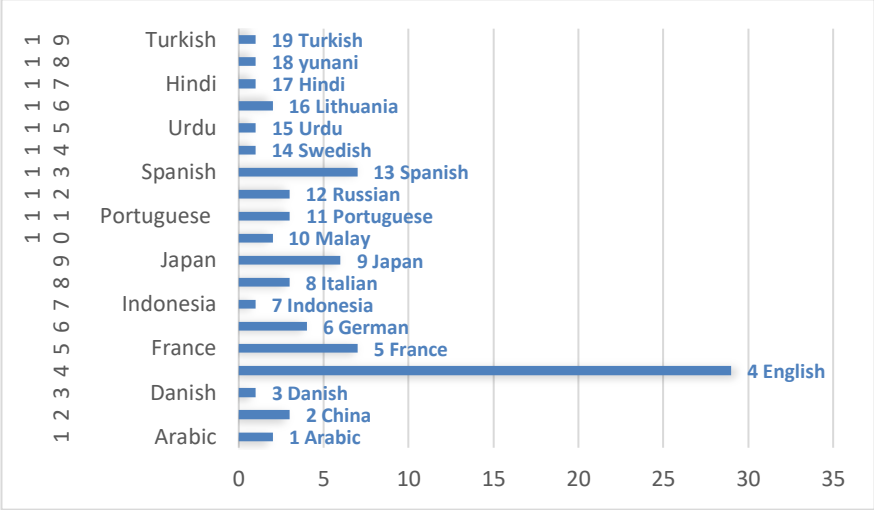
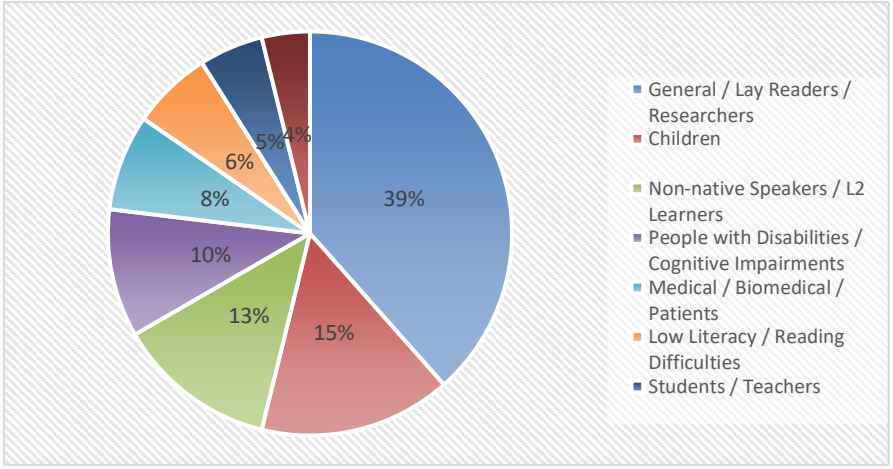


Figure 3. Language in use TS

RQ2. Target audiences used for TS

In the context of research on Text Simplification, based on the analysis of the main studies, seven categories of target audiences are identified, which are General, Non-Native, Dyslexic, Children, Disability, Medical, and Second Language (L2).

Text simplification was targeted at the General/Lay Readers/Researchers group at 39%. Furthermore, Children were the second most targeted group at 15%. Non-native Speakers/L2 Learners came in third with 13%, followed by People with Disabilities/Cognitive Impairments at 10%. Furthermore, the Medical/Biomedical/Patient group received attention at 8%, Low Literacy/Reading Difficulties at 6%, Students/Teachers at 5%, and the least targeted were other groups at 4%. This data shows that while the primary focus of text simplification remains on general readers and researchers, there is a significant trend to extend its benefits to vulnerable groups such as children, second language learners, and individuals with cognitive disabilities and reading difficulties.



p-ISSN: 2528-1682
e-ISSN: 2527-9165

Figure 4. Target audience used for TS

RQ3. Dataset/Corpus used for TS

Figure 5 shows that Wikipedia (including WikiLarge, WikiSmall, etc.) was the most frequently used dataset source in text simplification research, with 14 instances. This was followed by Newsela, with seven instances used as a corpus. Furthermore, Wikinews/News/News Article was used in four studies, while Parallel Corpus was used three times. Other datasets, such as BenchLS, were used in two studies. Several other corpora, such as LexMTurk5, NNSEval, MWLS1, Cambridge Corpus, ParaPhraser Plus/Opusparcus, Arabic Corpus, French Corpus, German pathology report, TurkCorpus, and SIMPITIKI Corpus, were only used once in the analyzed studies. Several language-specific corpora, such as SIMPLEX-PB, Chinese Corpus, Malay Corpus, Japanese Corpus, and MultiLS/HanLS, were also used two to three times each. These results demonstrate that text simplification research still relies heavily on big data sources like Wikipedia and Newsela. Meanwhile, the use of datasets from various languages and domains is still very limited, indicating the need for corpus diversification for research development in this area.

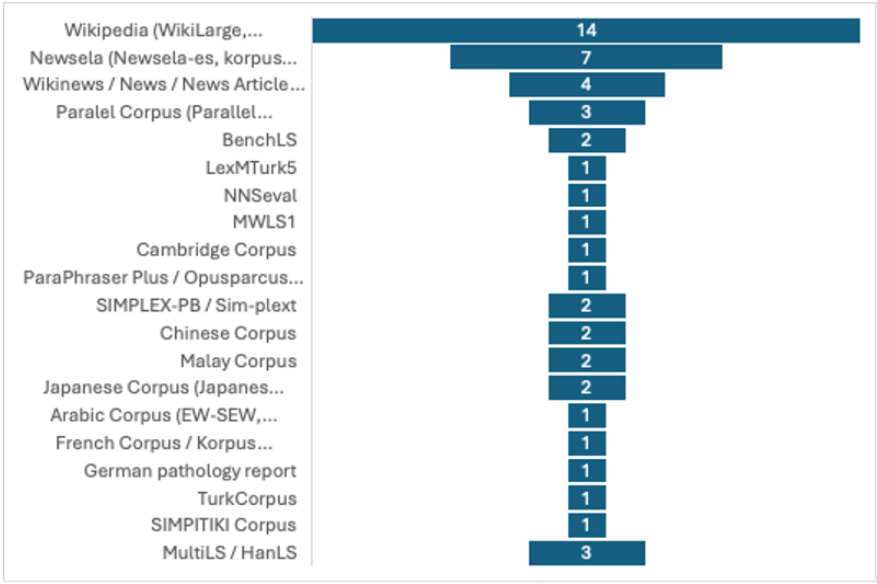


Figure 5. Dataset/Corpusused for TS

For completeness, Table 2 summarizes the included primary studies with respect to target audiences and datasets/corpora.

Table 2. The list of primary studies in the field of text simplification targets and the dataset/corpus

Language	Author	Target audience	Dataset/Corpus
Spanish	[16]	General/Researcher	IrekiaLF_es Corpus
	[17]	Disability	Wikipedia
	[18]	Children	Wikipedia
	[19]	General/Researcher	-
	[20]	Disability	-
	[21]	General/Researcher	Simplext and Newsela-es
English	[22]	Children	BenchLS
	[23]	General/Researcher	-
	[24]	Children	Wikipedia)
	[25]	General/Researcher	Semantic Corpus
	[14]	General/Researcher	LexMTurk5, BenchLS, NNSeval
	[26]	General/Researcher	Wikinews News

Language	Author	Target audience	Dataset/Corpus
	[27]	General/Researcher	Digital source
	[28]	General/Researcher	Wikipedia
	[29]	General/Researcher	MWLS1
	[30]	General/Researcher	-
	[31]	General/Researcher	WikiSmall And WikiLarge
	[32]	General/Researcher	Wikipedia
	[33]	General/Researcher	Cambridge Corpus
	[34]	General/Researcher	Newsela Corpus/GUG
	[35]	General/Researcher	Newsela Corpus
	[36]	General/Researcher	Space, Amazon
	[37]	Teacher	Teaching Artefact
	[38]	General/Researcher	Wikipedia, Wikidoc
	[39]	Non-native speakers, Children	OneStopQA
	[40]	General Patients	SimpleDC, Med-EASi
	[41]	Deaf, Hard-of-Hearing	Article
	[42]	Non-native speakers, Children	-
	[43]	lay readers	INFOLOSSQA
	[44]	Medical, General	various corpora
	[45]	Children, Non-native speakers	WikiLarge, ASSET
	[46]	Orthopedic patients	Orthoaedic Patient Education Materials
	[21]	General/Researcher	Wiki-Large and Newsela.
Russian	[47]	Children	Wikilarge/ Corpus of Russian paraphrases
	[48]	General	Wikilarge/Corpus of Russian paraphrases
	[49]	Non-Native	ParaPhraser Plus/Opusparcus/RuAdapt/ RuSimpleSentEval
Portuguese	[50]	Children	SIMPLEX-PB
	[51]	Low literacy	MultiLS
	[24]	Children	Wikipedia
China	[52]	Children	Chines Corpus
	[53]	Children	Chines Corpus
	[54]	Students, Foreign speakers, Individuals with cognitive impairments	HanLS
Malay	[3]	Non-Native	Malay Corpus
	[55]	Non-Native	Malay Corpus
Indonesia	[11]	Children	Wikipedia
Japan	[56]	Children	Wikipedia, News articles
	[57]	General/Researcher	Japanese Corpus
	[58]	Non-Native	Jades: News Article
	[59]	General/Researcher	Japanese Corpus
	[60]	General/Researcher	Paralel Corpus
	[61]	General/Researcher	Web Corpus 2010 (NWC 2010)
Italian	[10]	General/Researcher	Newsela Corpus
	[62]	Readers with reading difficulties, second language (L2) learners, and individuals with low literacy.	Multi-LS
	[21]	General/Researcher	SIMPITIKI Corpus, Newsela-es
Arabic	[63]	General/Researcher	Arabic Corpus
	[64]	General/Researcher	Arabic EW-SEW, Arabic WikiLarge
France	[65]	Dyslexia	Corpus for dyslexic
	[66]	Medical	New article
	[1]	General/Researcher	TurkCorpus
	[67]	People with language disorders due to a genetic disease or aphasia	ImageCLEF 2024
	[68]	Dyslexia	Parallel Corpus
	[69]	Medical/Patient	French Radiology Corpus
German	[70]	General/Researcher	Web
	[71]	General/Researcher	Newsela Corpus
	[72]	Disability	Wikipedia
	[73]	Medical and Patient	German Pathology Report
Urdu	[74]	General/Researcher	News, Magazine, Book
Hindi	[46]	General/Researcher	Wikipedia

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Language	Author	Target audience	Dataset/Corpus
Danish	[74]	General/Researcher	Wikipedia
Swedish	[75]	General/Researcher	Corpora Pseudo-Parallel
Lithuania	[76]	People with cognitive disabilities, Non-native speakers, and children	Parallel Corpus1,2
	[77]	General/Researcher	-
Spanish	[78]	Biomedical	SympTEMIST dataset
Turkish	[79]	General/Researcher	news, books/magazines, and Wikipedia
Greece	[39]	General	Wikipedia

RQ4. Techniques, Approaches, and Linguistic Features used for TS

Based on the techniques, approaches, and linguistic features identified in the reviewed studies, word embedding methods (including GloVe, fastText, and related models) are the most frequently used element in text simplification research, appearing six times, which indicates their dominant role in representing semantic similarity. Rule-based reasoning follows with five occurrences, showing that handcrafted linguistic rules remain relevant, while hybrid approaches that combine rule-based and machine learning methods are used four times. Lexical complexity prediction, back-translation, and word sense disambiguation (WSD) are each employed three times, mainly to support complex word identification and substitution. In contrast, morphology or morphosyntax, personalized approaches, graph-based approaches, and word frequency or character frequency features appear less frequently, with two occurrences each. Other elements, including natural language inference, end-to-end prompting, critique-refine strategies, candidate generation, retrieval-based interpretation augmentation, knowledge distillation, part-of-speech tagging, sememe-based approaches, synonym-based approaches, and word length, are each observed only once, suggesting that although a wide range of methods has been explored, text simplification research still predominantly focuses on embedding based, rule-based, and hybrid approaches, while newer strategies remain relatively underexplored.

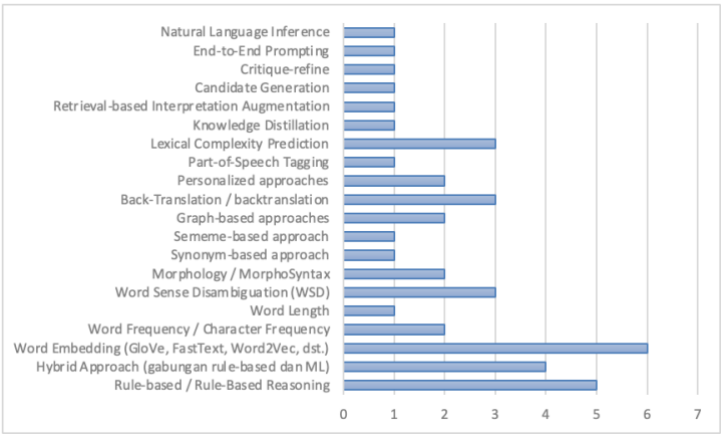


Figure 6. Techniques, Approaches, and Linguistic Features that are often used for TS

RQ5. Model /Algorithm used for TS

Based on the distribution of models shown in Figure 7, large language models (LLMs) emerge as the most dominant category, being used in six studies, which indicates a strong recent trend toward generative and prompt-based approaches in text simplification research. BERT-based models (including BERT, BERT-WWM, and BETO) and Seq2Seq architectures are each employed in five studies, highlighting their continued importance for contextual representation learning and encoder–decoder-based simplification. Transformer-based models and GPT-family models (GPT-2, GPT-4, GPT-4o, and ChatGPT) appear in three studies each, demonstrating a gradual transition from traditional transformer architectures toward more advanced pretrained and instruction-following models. Classical neural approaches, such as BiLSTM/LSTM, are also used in three studies, while conventional machine learning methods, including Support Vector Machine (SVM) and tree-based models (Decision Tree, Random

Forest, Naive Bayes, and Boosted Trees), appear with equal frequency, mainly for tasks such as complex word identification. In contrast, models such as BART, mBART/mT5/T5, RoBERTa, Feed-Forward Neural Networks (FFNN), and Helsinki-NLP/opus-mt-ROMANCE-en are used relatively infrequently, typically only once or twice, suggesting that although diverse modeling strategies exist, recent research increasingly prioritizes large-scale pretrained and generative models for text simplification.

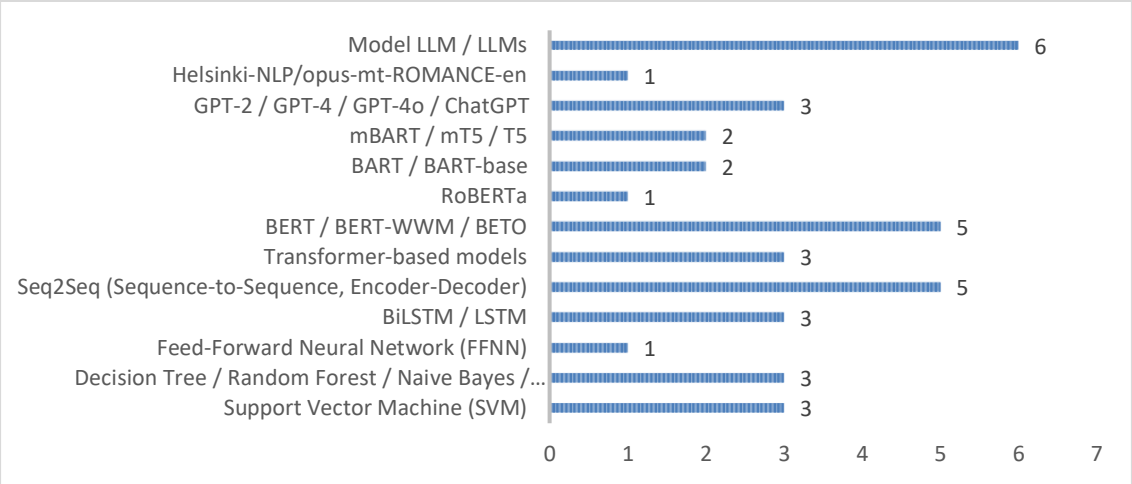


Figure 7. Model/ Algorithm used for TS

Details of the primary studies in the field of text simplification methods or approaches are presented in Table 3.

Table 3. The list of primary studies in the field of text simplification methods or approaches.

Language	Author	Method/Approach/Model/System/Framework
Spanish	[16]	Combination of rule-based and Machine learning approaches
	[17]	Word embedding, Length, Frequency, Support Vector Machine (SVM)
	[18]	Word Sense Disambiguation /Bert BETO
	[20]	SVM/Word2Vec
	[21]	Seq2seq
English	[22]	Personalized
	[24]	Word Sense Disambiguation/BERT-WWM
	[14]	Word Embedding/ BERT
	[26]	Frequency word / LSTM/BERT
	[27]	Rule-based Approaches
	[28]	Word Embeddings /Transformer Embedding perceptron, SVM, Random Forest
	[80]	Seq2Seq
	[32]	LSTM
	[33]	Graph-based
	[34]	BERT
	[35]	Graph-based
	[36]	TTSG
	[37]	Eng-Editor
	[38]	Transformer-based Sequence-to-Sequence Model, Feed-Forward Neural Network (FFNN)
	[39]	Supervised Model, Unsupervised model, Black-box model
	[40]	LLM
	[41]	Automatic text simplification
	[42]	LLMs
	[43]	End-to-end prompting, Natural Language Inference
	[44]	Language Models (LMs): BERT, RoBERTa.
	[45]	BART-base (fine-tuning)
	[46]	LLM
	[21]	Seq2seq
Russian	[48]	Mbart
	[49]	mBART, T5
Portuguese	[24]	Word Sense Disambiguation
	[50]	Back-Translation

Language	Author	Method/Approach/Model/System/Framework
	[51]	LLM
China	[53]	Word2vec
	[52]	Synonym-based approach, Word embedding-based approach, BERT-based approach, Sememe-based approach, and a Hybrid approach
	[54]	Retrieval-based Interpretation Augmentation, Knowledge Distillation
Malay	[3]	Boosted Tree, Random Forest Naive Bayes, Decision Tree/supervised and unsupervised machine learning algorithms
	[55]	Boosted Tree, Random Forest Naive Bayes, Decision Tree (Word Stemming)/ Hybrid approach
Indonesia	[11]	Word embeddings-based and rule-based reasoning method
Japan	[56]	Transformer dan BART
	[57]	Transformer and Encoder decoder
	[58]	N-gram embedding with unsupervised learning
Italian	[10]	Seq2Seq
	[21]	Seq2Seq
Arabic	[63]	Morphology
	[64]	Hybrid approach
France	[80]	Rule-Based Reasoning
	[66]	Rule-Based Reasoning
	[1]	Rule-Based Reasoning
	[67]	GPT-2, Helsinki-NLP/opus-mt-ROMANCE-en
	[68]	MorphoSyntax
German	[70]	Back-Translation
	[71]	Back-Translation
Urdu	[9]	Word vector / Word2vec
Turkish	[79]	BiLSTM
Lituana	[76]	LLM
	[77]	LLM

RQ6. Tools/systems/frameworks that are often used for TS

Based on the distribution of tools, systems, and frameworks commonly used for Text Simplification (TS) as shown in Figure 8, the reviewed studies employ four main types with equal frequency. Eng-Editor, automatic text simplification (end-to-end) systems, Plainifier, and JDM Network each represent 25% of the total usage, indicating that no single tool or system dominates TS research. This equal distribution reflects the diversity of implementation strategies in TS, ranging from editor-assisted tools and fully automated end-to-end systems to lexicon-driven tools and network-based simplification systems. The results suggest that TS research does not rely on a single standardized tool or framework, but rather explores multiple system-level solutions depending on research objectives, language resources, and application scenarios.

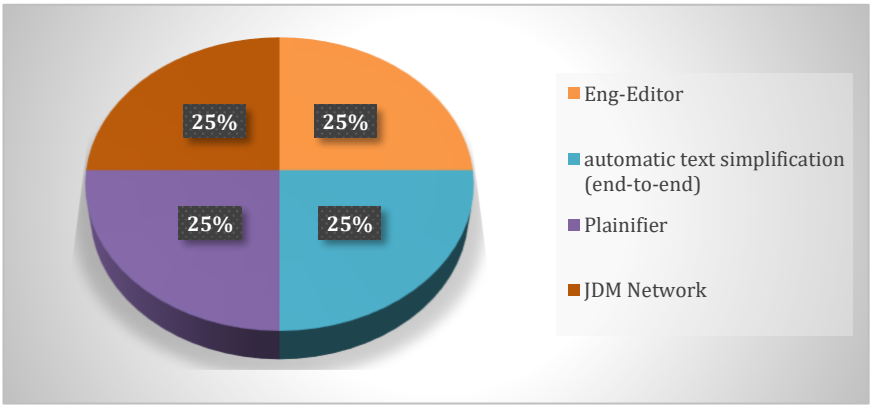


Figure 8. Tools/Systems/Frameworks that are often used for TS

3.1 Discussion

Main contributions of this SLR are as follows: (1) a structured mapping of TS research from 2015 to 31 July 2025 across languages, target audiences, datasets/corpora, techniques, models/algorithms, and tools/frameworks were provided; (2) the dominant resources and modeling trends using evidence from the included primary studies were summarized; and (3) the research gaps and translate them into actionable recommendations, particularly for low-resource languages and audience-specific evaluation were identified.

Overall, the results indicate three major trends that also clarify current challenges in TS research. First, TS research remains language-imbalanced, with English dominating and many non-English languages receiving limited attention. This indicates a practical need for new datasets, benchmarks, and evaluation resources in low-resource settings. Second, studies rely heavily on a small number of datasets/corpora (e.g., Wikipedia-family datasets and Newsela), which may limit domain diversity and audience representativeness. This motivates the development of domain- and audience-specific corpora beyond general-domain resources. Third, approaches increasingly shift toward transformer-based models and LLMs, showing strong interest in context-aware generation; however, audience-specific requirements (e.g., dyslexia, children, L2 learners) are still not consistently addressed through tailored evaluation protocols and readability constraints.

Based on these findings, the future solutions in four directions were recommended: (1) develop and release language- and audience-specific corpora (education, medical, accessibility) with clear annotation guidelines; (2) adopt transparent reporting of data sources, selection criteria, and evaluation protocols to improve reproducibility; (3) explore hybrid and controllable simplification that combines linguistic constraints (e.g., morphology and syntactic rules) with neural generation; and (4) strengthen evaluation by combining automatic metrics (e.g., SARI, BLEU) with human-centered readability assessment aligned with the target audience.

4. CONCLUSION

This paper presented a Systematic Literature Review (SLR) of text simplification research published from 2015 to 31 July 2025. Based on 75 primary studies, TS research across languages was mapped, target audiences, datasets/corpora, techniques, models/algorithms, and frameworks. The results show that TS research is highly concentrated in English and depends heavily on a small number of widely used corpora, while low-resource languages remain underrepresented. Importantly, the methodological trend has shifted from traditional rule-based and classical machine-learning approaches toward transformer-based models and, more recently, large language models (LLMs), reflecting the growing focus on context-aware neural generation.

Future research should focus on three priorities: (1) creating and sharing simplification datasets for underrepresented languages (e.g., Arabic, Urdu, Indonesian, Malay, Danish) and for specific domains such as education and healthcare; (2) developing clearer TS guidelines and evaluation protocols for different audiences, including children, dyslexic readers, L2 learners, and users with cognitive impairments; and (3) building controllable simplification systems that combine linguistic rules (morphology and syntax) with neural generation to produce outputs that are accurate, easier to read, and practical for real-world use.

ACKNOWLEDGEMENTS

Non-funded

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] R. Hijazi, B. Espinasse, and N. Gala, "GRASS: A Syntactic Text Simplification System based on Semantic Representations," *Computer Science & Information Technology (CS & IT)*, vol. 12, no. 15, pp. 221–236, 2022, doi: 10.5121/csit.2022.121518.

[2] G. Smolenska, *Complex Word Identification for Swedish*, M.S. thesis, Department of Linguistics and Philology, Uppsala University, Sweden, 2018

[3] S. Omar, J. A. Bakar, M. M. Nadzir, N. H. Harun, and N. Yusoff, "Malay lexical simplification model for non-native speaker," in *2022 International Conference on Intelligent Systems and Computer Vision, ISCV 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ISCV54655.2022.9806133.

[4] G. Lo Bosco, G. Pilato, and D. Schicchi, "A Neural Network model for the Evaluation of Text Complexity in Italian Language: A Representation Point of View," *Procedia Comput. Sci.*, vol. 145, pp. 464–470, 2018, doi: 10.1016/j.procs.2018.11.108.

[5] L. Rello, R. Baeza-Yates, S. Bott, and H. Saggion, "Simplify or help? Text simplification strategies for people with dyslexia," in *Proc. 10th Int. Cross-Disciplinary Conf. Web Accessibility (W4A 2013)*, Rio de Janeiro, Brazil, 2013, Art. no. 15, 10 pp., doi: 10.1145/2461121.2461126.

[6] R. Evans, C. Orăsan, and I. Dornescu, "An evaluation of syntactic simplification rules for people with autism," in *Proc. 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, Gothenburg, Sweden, 2014, pp. 131–140, doi: 10.3115/v1/W14-1215.

[7] E. Barbu, M. T. Martín-Valdivia, and L. A. Ureña-López, "Open Book: A tool for helping ASD users' semantic comprehension," in *Proc. Workshop Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, Atlanta, GA, USA, 2013, pp. 11–19.

[8] L. Lisnawita and J. A. Bakar, "Prospects for Lexical Simplification: Bibliometric Analysis and New Directions of Research," in *Proc. 2nd Int. Conf. Environmental, Energy, and Earth Science (ICEEES 2023)*, Pekanbaru, Indonesia, Oct. 30–31, 2023, EAI, 2024, doi: 10.4108/eai.30-10-2023.2343091.

[9] N. H. Qasmi, H. Bin Zia, A. Athar, and A. A. Raza, "SimplifyUR: Unsupervised lexical text simplification for Urdu," in *Proc. 12th Language Resources and Evaluation Conf. (LREC)*, Marseille, France, 2020, pp. 3484–3489.

[10] A. L. Megna, D. Schicchi, G. Lo Bosco, and G. Pilato, "A Controllable Text Simplification System for the Italian Language," in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, Laguna Hills, CA, USA, Jan. 27–29, 2021, pp. 191–194, doi: 10.1109/ICSC50631.2021.00040.

[11] M. S. Wibowo, A. Romadhony, and S. Sa'Adah, "Lexical and Syntactic Simplification for Indonesian Text," *2019 2nd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2019*, pp. 64–68, 2019, doi: 10.1109/ISRITI48646.2019.9034582.

[12] J. A. Bakar, N. Yusoff, N. H. Harun, M. M. Nadzir, and S. Omar, "Text Simplification using Hybrid Semantic Compression and Support Vector Machine for Troll Threat Sentences," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, pp. 322–330, 2023, doi: 10.14569/IJACSA.2023.0141035.

[13] G. Venugopal and D. Pramod, "Bibliometric Analysis of Studies on Lexical Simplification," *Lecture Notes in Networks and Systems*, vol. 647 LNNS, pp. 3–12, 2023, doi: 10.1007/978-3-031-27409-1_1.

[14] J. Qiang, Y. Li, Y. Zhu, Y. Yuan, and X. Wu, "Lexical Simplification with Pretrained Encoders," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8649–8656, 2020, doi: 10.1609/aaai.v34i05.6389

[15] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering-A systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009

[16] I. Gonzalez-Dios, I. Gutiérrez-Fandiño, O. M. Cumbicus-Pineda, and A. Soroa, "IrekiaLFes: A new open benchmark and baseline systems for Spanish automatic text simplification," in *Proc. Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, Abu Dhabi, UAE (Virtual), 2022, pp. 86–97, doi: 10.18653/v1/2022.tsar-1.8.

[17] R. Alarcon, L. Moreno, I. Segura-Bedmar, and P. Martínez, "Lexical simplification approach using easy-to-read resources," *Procesamiento del Lenguaje Natural*, vol. 63, pp. 95–102, 2019, doi: 10.26342/2019-63-10.

[18] S. Štajner, I. Calixto, and H. Saggion, "Automatic text simplification for Spanish: Comparative evaluation of various simplification strategies," in *Proc. Int. Conf. Recent Advances in Natural Language Processing (RANLP)*, Hissar, Bulgaria, 2015, pp. 618–626.

[19] H. Saggion, S. Bott, S. Szasz, N. Pérez, S. Calderón, and M. Solís, "Lexical Complexity Prediction and Lexical Simplification for Catalan and Spanish: Resource Creation, Quality Assessment, and Ethical Considerations," in *TSAR 2024 - 3rd Workshop on Text Simplification, Accessibility and Readability, Proceedings of the Workshop*, M. Shardlow, H. Saggion, F. Alva-Manchego, M. Zampieri, K. North, S. Stajner, and R. Stodden, Eds., Association for Computational Linguistics (ACL), 2024, pp. 82–94. doi: 10.18653/v1/2024.tsar-1.9.

[20] R. Alarcon, "Lexical Simplification System to Improve Web Accessibility," *IEEE Access*, vol. 9, pp. 58755–58767, 2021, doi: 10.1109/ACCESS.2021.3072697.

[21] O. M. Cumbicus-Pineda, I. Gonzalez-Dios, and A. Soroa, "A syntax-aware edit-based system for text simplification," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, Held Online, Sep. 1–3, 2021, pp. 324–334, doi: 10.26615/978-954-452-072-4_038.

[22] J. Lee and C. Y. Yeung, "Personalizing lexical simplification," in *Proc. 27th Int. Conf. Computational Linguistics (COLING 2018)*, Santa Fe, NM, USA, Aug. 2018, pp. 224–232.

[23] N. Srikanth and J. J. Li, "Elaborative simplification: Content addition and explanation generation in text simplification," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online, Aug. 2021, pp. 5123–5137, doi: 10.18653/v1/2021.findings-acl.455.

[24] S. Štajner, D. Ferrés, M. Shardlow, K. North, M. Zampieri, and H. Saggion, "Lexical simplification benchmarks for English, Portuguese, and Spanish," *Front. Artif. Intell.*, vol. 5, pp. 1–32, 2022, doi: 10.3389/frai.2022.991242.

[25] E. Sulem, O. Abend, and A. Rappoport, "Simple and effective text simplification using semantic and neural methods," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL 2018), vol. 1: Long Papers*, Melbourne, Australia, Jul. 2018, pp. 162–173, doi: 10.18653/v1/P18-1016.

[26] H. Mahajan, P. Koul, and S. Singh, "Lexical analyser web extension for text simplification," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 8, no. 4, pp. 24–30, 2022.

[27] A. Garain, A. Basu, R. Dawn, and S. K. Naskar, "Sentence Simplification using Syntactic Parse trees," *2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019*, pp. 672–676, 2019, doi: 10.1109/ISCON47742.2019.9036207.

[28] C.-O. Truică, A.-I. Stan, and E.-S. Apostol, "SimpLex: a lexical text simplification architecture," *Neural Comput. Appl.*, vol. 35, no. 8, pp. 6265–6280, 2023, doi: 10.1007/s00521-022-07905-y.

[29] P. Przybyła and M. Shardlow, "Multi-word lexical simplification," in *Proc. 28th Int. Conf. Computational Linguistics (COLING 2020)*, Barcelona, Spain (Online), Dec. 2020, pp. 1435–1446, doi: 10.18653/v1/2020.coling-main.123.

[30] S. Seneviratne, E. Daskalaki, and H. Suominen, "CILS at TSAR-2022 Shared Task: Investigating the Applicability of Lexical Substitution Methods for Lexical Simplification," in *Proc. Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, Abu Dhabi, UAE (Virtual), 2022, pp. 207–212, doi: 10.18653/v1/2022.tsar-1.21.

[31] T. Li, Y. Li, J. Qiang, and Y.-H. Yuan, "Text simplification with self-attention-based pointer-generator networks," in *Neural Information Processing*, Cham, Switzerland: Springer, 2018, pp. 537–545, doi: 10.1007/978-3-030-04221-9_48.

[32] A. Dmitrieva and J. Tiedemann, "A multi-task learning approach to text simplification," in *Recent Trends in Analysis of Images, Social Networks and Texts*, vol. 1357, Communications in Computer and Information Science. Cham, Switzerland: Springer, 2021, pp. 78–89, doi: 10.1007/978-3-030-71214-3_7.

[33] R. Xu, W. Pan, C. Chen, X. Chen, S. Lin, and X. Li, "Graph-based model using text simplification for readability assessment," in *Proc. 2022 Int. Conf. Asian Language Processing (IALP)*, 2022, pp. 401–406, doi: 10.1109/IALP57159.2022.9961261.

[34] A. Nakamachi, T. Kajiwara, and Y. Arase, "Text simplification with reinforcement learning using supervised rewards on grammaticality, meaning preservation, and simplicity," in *Proc. 1st Conf. Asia-Pacific Chapter of the Association for Computational Linguistics and 10th Int. Joint Conf. Natural Language Processing: Student Research Workshop (AACL-IJCNLP 2020 SRW)*, Suzhou, China, Dec. 2020, pp. 153–159, doi:10.18653/v1/2020.aacl-srw.22.

[35] M. Schwarzer, T. Tanprasert, and D. Kauchak, "Improving human text simplification with sentence fusion," in *Proc. 15th Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, Mexico City, Mexico, Jun. 2021, pp. 106–114, doi: 10.18653/v1/2021.textgraphs-1.10.

[36] R. Wang, T. Lan, Z. F. Wu, and L. Y. Liu, "Unsupervised extractive opinion summarization based on text simplification and sentiment guidance," *Expert Syst. Appl.*, vol. 272, 2025, doi: 10.1016/j.eswa.2025.126760.

[37] F. K. Liu, Y. S. Jiang, C. Lai, and T. Jin, "Teacher engagement with automated text simplification for differentiated instruction," *LANGUAGE LEARNING & TECHNOLOGY*, vol. 28, no. 2, pp. 163–182, 2024, doi: 10.10125/73576.

[38] S. A. Bahrainian, J. Dou, and C. Eickhoff, "Text simplification via adaptive teaching," in *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, Aug. 2024, pp. 6574–6584, doi: 10.18653/v1/2024.findings-acl.392.

[39] S. Agrawal and M. Carpuat, "Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 432–448, 2024, doi: 10.1162/tacl_a_00653.

[40] M. M. Rahman, M. S. Irbaz, K. North, M. S. Williams, M. Zampieri, and K. Lybarger, "Health text simplification: An annotated corpus for digestive cancer education and novel strategies for reinforcement learning," *J. Biomed. Inform.*, vol. 158, 2024, doi: 10.1016/j.jbi.2024.104727.

[41] O. Alonzo, S. Lee, A. Al Amin, M. Maddela, W. Xu, and M. Huenerfauth, "Design and evaluation of an automatic text simplification prototype with Deaf and hard-of-hearing readers," in *Proc. 26th Int. ACM SIGACCESS Conf. Computers and Accessibility (ASSETS 2024)*, St. John's, NL, Canada, 2024, Art. no. 40, 18 pp., doi: 10.1145/3663548.3675645.

[42] N. Freyer, H. Kempt, and L. Klöser, "Easy-read and large language models: on the ethical dimensions of LLM-based text simplification," *Ethics Inf. Technol.*, vol. 26, no. 3, 2024, doi: 10.1007/s10676-024-09792-4.

[43] J. Trienes, S. Joseph, J. Schlötterer, C. Seifert, K. Lo, W. Xu, B. Wallace, and J. J. Li, "InfoLossQA: Characterizing and recovering information loss in text simplification," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL 2024), vol. 1: Long Papers*, Bangkok, Thailand, 2024, pp. 4263–4294, doi: 10.18653/v1/2024.acl-long.234.

[44] D. Davari, L. Ermakova, and R. Krestel, "Comparative analysis of evaluation measures for scientific text simplification," in *Linking Theory and Practice of Digital Libraries: 28th International Conference on Theory and Practice of Digital Libraries (TPDL 2024), Lecture Notes in Computer Science*, vol. 15177, Ljubljana, Slovenia, 2024, pp. 76–91, doi: 10.1007/978-3-031-72437-4_5.

[45] Z. H. Li and M. Shardlow, "How do control tokens affect natural language generation tasks like text simplification," *Nat. Lang. Eng.*, 2024, doi: 10.1017/S1351324923000566.

[46] S. Andalib, S. S. Solomon, B. G. Picton, A. C. Spina, J. A. Scolaro, and A. M. Nelson, "Source characteristics influence AI-enabled orthopaedic text simplification: Recommendations for the future," *JBJS Open Access*, vol. 10, no. 1, Art. no. e24.00007, 2025, doi: 10.2106/JBJS.OA.24.00007.

[47] A. Dmitrieva and J. Tiedemann, "Creating an aligned Russian text simplification dataset from language learner data," in *Proc. 8th Workshop on Balto-Slavic Natural Language Processing (BSNLP)*, Kiyv, Ukraine, 2021, pp. 73–79.

[48] F. Galeev, M. Leushina, and V. Ivanov, "ruBTS: Russian Sentence Simplification Using Back-translation," *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, vol. 2021-June, no. 20, pp. 259–267, 2021, doi: 10.28995/2075-7182-2021-20-259-267.

[49] A. Dmitrieva, "Automatic text simplification of Russian texts using control tokens," in *Proc. 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, Dubrovnik, Croatia, 2023, pp. 70–77, doi: 10.18653/v1/2023.bsnlp-1.9

[50] N. S. Hartmann, G. H. Paetzold, and S. M. Aluísio, "A dataset for the evaluation of lexical simplification in Portuguese for children," in *Computational Processing of the Portuguese Language (PROPOR 2020)*, Lecture Notes in Computer Science, vol. 12037, Cham, Switzerland: Springer, 2020, pp. 55–64, doi: 10.1007/978-3-030-41505-1_6.

[51] K. North, T. Ranasinghe, M. Shardlow, and M. Zampieri, "MultiLS: An End-to-End Lexical Simplification Framework," in *TSAR 2024 - 3rd Workshop on Text Simplification, Accessibility and Readability, Proceedings of the Workshop*, M. Shardlow, H. Saggion, F. Alva-Manchego, M. Zampieri, K. North, S. Stajner, and R. Stodden, Eds., Association for Computational Linguistics (ACL), 2024, pp. 1–11. doi: 10.18653/v1/2024.tsar-1.1.

[52] J. Qiang, X. Lu, Y. Li, Y.-H. Yuan, and X. Wu, "Chinese lexical simplification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1819–1828, 2021, doi: 10.1109/TASLP.2021.3078361.

[53] S. Yang and F. Yang, "Chinese automatic text simplification based on unsupervised learning," in *Proc. 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, China, 2021, pp. 1666–1672, doi: 10.1109/IAEAC50856.2021.9390937.

[54] Z. Xiao, J. Gong, S. Wang, and W. Song, "Optimizing Chinese Lexical Simplification Across Word Types: A Hybrid Approach," in *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, Y. Al-Onaizan, M. Bansal, and C. Y.-N., Eds., Association for Computational Linguistics (ACL), 2024, pp. 15227–15239. doi: 10.18653/v1/2024.emnlp-main.849.

[55] S. Omar, J. A. Bakar, M. M. Nadzir, N. H. Harun, and N. Yusoff, "Text simplification for Malay corpus: A review," in *Proc. 2021 Int. Conf. Comput. Inf. Sci. (ICCOINS)*, 2021, doi: 10.1109/ICCOINS49721.2021.9497167.

[56] T. Kato, R. Miyata, and S. Sato, "BERT-Based Simplification of Japanese Sentence-Ending Predicates in Descriptive Text," in *Proc. 13th International Conference on Natural Language Generation (INLG 2020)*, Dublin, Ireland, 2020, pp. 242–251, doi: 10.18653/v1/2020.inlg-1.31

[57] D. Nishihara and T. Kajiwara, "Word complexity estimation for Japanese lexical simplification," in *Proc. 12th Language Resources and Evaluation Conf. (LREC)*, Marseille, France, 2020, pp. 3114–3120.

[58] A. Hayakawa, T. Kajiwara, H. Ouchi, and T. Watanabe, "{JADES}: New Text Simplification Dataset in {J}apanese Targeted at Non-Native Speakers," *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pp. 179–187, 2022. doi: 10.18653/v1/2022.tsar-1.17

[59] K. Hatagaki, T. Kajiwara, and T. Ninomiya, "Parallel Corpus Filtering for {J}apanese Text Simplification," *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pp. 12–18, 2022. doi: 10.18653/v1/2022.tsar-1.2.

[60] T. Maruyama and K. Yamamoto, "Extremely low-resource text simplification with pre-trained transformer language model," *International Journal of Asian Language Processing*, vol. 30, no. 1, Art. no. 2050001, 2020, doi: 10.1142/S2717554520500010.

[61] A. Katsuta and K. Yamamoto, "Improving text simplification by corpus expansion with unsupervised learning," in *Proc. 2019 Int. Conf. Asian Language Processing (IALP)*, Shanghai, China, 2019, pp. 216–221, doi: 10.1109/IALP48816.2019.9037567.

[62] L. Occhipinti, "Introducing MultiLS-IT: A dataset for lexical simplification in Italian," in *Proc. 10th Italian Conf. Computational Linguistics (CLiC-it)*, Pisa, Italy, 2024, pp. 662–669.

[63] R. Hazim, H. Saddiki, B. Alhafni, M. Al Khalil, and N. Habash, "Arabic word-level readability visualization for assisted text simplification," in *Proc. 2022 Conf. Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, Abu Dhabi, UAE, 2022, pp. 242–249, doi: 10.18653/v1/2022.emnlp-demos.24.

[64] S. S. Al-Thanyyan and A. M. Azmi, "Simplification of Arabic text: A hybrid approach integrating machine translation and transformer-based lexical model," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 8, p. 101662, 2023, doi: 10.1016/j.jksuci.2023.101662.

[65] R. Wilkens, B. Oberle, and A. Todirascu, "Coreference-Based Text Simplification," in *Proc. 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, Marseille, France, 2020, pp. 93–100.

[66] A. Koptient and N. Grabar, "Fine-grained text simplification in French: Steps towards a better grammaticality," in *Proc. 18th Int. Symp. Health Inf. Manage. Res. (ISHIMR)*, 2020, doi: 10.15626/ishimr.2020.07.

[67] A. Elliah, A. Narayanan P., B. S., and P. Mirunalini, "Text-To-Picto using lexical simplification," in *CLEF 2024 Working Notes*, 2024, pp. 1579–1584.

[68] N. Gala, A. Tack, L. Javourey-Drevet, T. François, and J. C. Ziegler, "Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers," in *Proc. 12th Language Resources and Evaluation Conf. (LREC)*, Marseille, France, 2020, pp. 1353–1361.

[69] L. Ramadier and M. Lafourcade, "Radiological text simplification using a general knowledge base," in *Computational Linguistics and Intelligent Text Processing*, vol. 10762, Cham, Switzerland: Springer, 2018, pp. 617–627, doi: 10.1007/978-3-319-77116-8_46.

[70] A. Battisti, D. Pfütze, A. Säuberli, M. Kostrzewa, and S. Ebling, "A corpus for automatic readability assessment and text simplification of German," in *Proc. 12th Language Resources and Evaluation Conf. (LREC)*, Marseille, France, 2020, pp. 3302–3311.

[71] A. Säuberli, S. Ebling, and M. Volk, "Benchmarking data-driven automatic text simplification for German," in *Proc. 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, Marseille, France, 2020, pp. 41–48

[72] S. Strübbe, I. Sidorenko, S. Roy, A. T. D. Grünwald, and R. Lampe, "Development and verification of a user-friendly software for German text simplification focused on patients with cerebral palsy," *Natural Language Processing Journal*, vol. 3, Art. no. 100012, 2023, doi: 10.1016/j.nlp.2023.100012.

[73] J. Trienes, J. Schlötterer, H.-U. Schildhaus, and C. Seifert, "Patient-friendly clinical notes: Towards a new text simplification dataset," in *Proc. Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, Abu Dhabi, UAE, 2022, pp. 19–27, doi: 10.18653/v1/2022.tsar-1.3.

[74] J. Bingel, G. H. Paetzold, and A. Søgaard, "Lexi: A tool for adaptive, personalized text simplification," in *Proc. 27th Int. Conf. Computational Linguistics (COLING)*, Santa Fe, NM, USA, 2018, pp. 245–258

[75] D. Holmer and E. Rennes, "Constructing pseudo-parallel Swedish sentence corpora for automatic text simplification," in *Proc. 24th Nordic Conf. Computational Linguistics (NoDaLiDa)*, Tórshavn, Faroe Islands, 2023, pp. 113–123.

[76] J. Mandravickaitė, E. Rimkienė, D. K. Kapkan, D. Kalinauskaitė, and T. Krilavičius, "Automatic Simplification of Lithuanian Administrative Texts," *Algorithms*, vol. 17, no. 11, 2024, doi: 10.3390/a17110533.

[77] J. Mandravickaitė, E. Rimkienė, D. K. Kapkan, D. Kalinauskaitė, A. Čenys, and T. Krilavičius, "Automatic Text Simplification for Lithuanian: Transforming Administrative Texts into Plain Language," *Mathematics*, vol. 13, no. 3, pp. 1–28, 2025, doi: 10.3390/math13030465.

[78] F. Borchert, I. Llorca, and M. P. Schapranow, "Improving biomedical entity linking for complex entity mentions with LLM-based text simplification," *Database-The Journal Of Biological Databases And Curation*, vol. 2024, 2024, doi: 10.1093/database/baae067.

[79] B. Ilgen and G. Hattab, "A Lexical Simplification Framework for Turkish," *NLPIR 2024 - 2024 8th International Conference on Natural Language Processing and Information Retrieval*, pp. 218–224, 2025, doi: 10.1145/3711542.3711571.

[80] A. Todirascu, R. Wilkens, E. Rolin, T. François, D. Bernhard, and N. Gala, "HECTOR: A hybrid text simplification tool for raw texts in French," in *Proc. 13th Language Resources and Evaluation Conf. (LREC)*, Marseille, France, 2022, pp. 4620–4630.