HULAT-HAI @ SimpleText 2026 Task 1.1: Prompting Based on Linguistic Analysis for Text Simplification

SimpleText at CLEF 2026

Nuria Pirvu[1], Rémi Cardon[1]

[1]HULAT / Department of Computer Science and Engineering, Universidad Carlos III de Madrid, Spain

Abstract

We describe the HULAT-HAI system for SimpleText 2026 Task 1.1, which targets automatic text simplification across the languages of the evaluation set. Our approach combines rule-based syntactic analysis with prompt-based generation. For five languages with mature spaCy support (English, French, Spanish, German, and Portuguese), each source sentence is first annotated with a set of linguistic complexity features grounded in plain language research on administrative writing; these annotations guide a large language model (Gemma 3 27B) toward targeted, phenomenon-specific simplification via a specialised prompt per feature. For the remaining evaluation languages, a single prompt listing all guidelines is applied instead. On the English test set, our two submitted runs achieve SARI scores of 43.86 and 44.20, with the run that exposes the original sentence to the combination prompt producing marginally higher n-gram overlap but less aggressive simplification.

Keywords

text simplification, prompting, linguistic analysis,

1. Introduction

Automatic text simplification aims to reduce the linguistic complexity of a text while preserving its meaning, so as to make content accessible to a wider range of readers. This paper describes the HULAT-HAI system submitted to Task 1.1 (sentence-level simplification) of the SimpleText shared task at CLEF 2026 [1].

Our approach builds on the observation that textual complexity arises from a small number of recurring linguistic phenomena (e.g. long sentences, passive voice, subordinate clauses, impersonal and negative constructions) that plain language research has identified. We therefore pair a rule-based analysis stage, which annotates source sentences with the specific complexity features it exhibits, with a prompt-based generation stage in which a large language model addresses each detected phenomenon through a targeted instruction.

The system was designed and tuned on English during the preparation phase, for which the organisers released English-only data. The evaluation phase, however, spanned eleven typologically diverse languages. Because our analysis stage relies on language-specific resources and on complexity criteria whose validity we could establish only for a subset of these languages, we apply the full linguistically informed pipeline where those conditions hold and fall back on a simpler, language-agnostic strategy elsewhere. Section 3 makes this distinction precise. The linguistically informed pipeline is our primary contribution and the focus of our evaluation.

On the English test set, our two submitted runs achieve SARI scores of 43.86 and 44.20. On the training data, the pipeline improves the CEFR reading level of more than half of the sentences, outperforming a stronger baseline on this readability dimension despite lower n-gram overlap with the reference. The remainder of the paper is organised as follows. Section 2 motivates our tag set from plain language research. Section 3 describes the linguistic tagging pipeline and the languages it covers. Section 4 details the prompting strategy. Section 5 reports our experimental results.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

nurialuciapirvu@gmail.com (N. Pirvu); rcardon@inf.uc3m.es (R. Cardon)

0000-0002-7858-3504 (R. Cardon)

2. Background: Plain Language and Syntactic Complexity

The linguistic features targeted by our system are grounded in an empirical study on plain language in French-speaking administrations. Müller and François [2] analysed three Belgian plain language guides against a broader set of 21 francophone guides, and verified their recommendations on a corpus of 102 authentic administrative texts. The syntactic features consistently flagged across these guides are: long sentences, passive voice, negative constructions, subordinate clauses, and impersonal constructions. All three guides also recommend personalising texts via direct address (first and second person), and avoiding unexplained acronyms. Corpus analysis confirmed that these features remain pervasive in practice, particularly in legal texts (e.g. 99.38% of sentences with at least one subordinate clause, 15.66% passive).

2.1. From plain language research to the tag set

The seven features targeted by our tagging pipeline directly reflect the empirical findings described above. Long_Sentence corresponds to the universal advice to limit sentence length. Passive, Negation, Impersonal, and Subordination capture the syntactic constructions most consistently flagged by guides and most resistant to human simplification. Personal_Pronoun operationalises the recommendation to address the reader directly. Acronym targets the lexical recommendation to avoid or explain abbreviations. Although the empirical grounding of these features comes primarily from French-language research on administrative and legal texts, we extend the tag set to four additional languages (English, Spanish, German, Portuguese) on the assumption that these categories identify general patterns of syntactic complexity that apply cross-linguistically. The SimpleText task data is biomedical rather than administrative, but the syntactic phenomena we target are characteristic of specialised writing more broadly and are not specific to any particular domain.

3. Linguistic Tagging Pipeline

A central component of our approach is a set of rule-based syntactic taggers that annotate each source sentence before it is passed to the prompt-based simplification system. The intuition is that making linguistic complexity features explicit allows the prompt to provide more targeted simplification instructions.

The evaluation covers eleven languages: English, French, Spanish, German, Portuguese, Persian, Hindi, Japanese, Korean, Thai, and Simplified Chinese. We developed one tagger for English, French, Spanish, German, and Portuguese, selected on two grounds. First, spaCy provides large pretrained models with full morphosyntactic annotation (POS tagging, morphology, and dependency parsing) for these languages. Second, the syntactic categories our tags target are well defined and broadly aligned with the plain language literature in each of them. For the remaining six languages, where we could not assess whether these syntactic criteria constitute meaningful complexity markers, we bypass the tagging pipeline entirely and apply a single prompt listing all simplification guidelines globally, without feature filtering.

3.1. Architecture

Each tagger is built on top of the large spaCy models for the corresponding language (`en_core_web_lg`, `fr_core_news_lg`, `es_core_news_lg`, `de_core_news_lg`, `pt_core_news_lg`). Given a sentence, the tagger runs the full spaCy pipeline (tokenisation, POS tagging, morphological analysis, dependency parsing) and applies a set of binary detection functions, each targeting one linguistic feature. The result is a set of tags assigned to the sentence, which may be empty if none of the features are detected.

3.2. Tag Set

We define seven tags, described below.

SUBORDINATION. A sentence is tagged as SUBORDINATION when it contains more than one finite verb form, operationalised as tokens whose POS tag is VERB, whose morphological feature `Tense` is non-empty, and whose `VerbForm` is `Fin`.

PASSIVE. Passive voice is detected through a combination of morphological, syntactic, and lexical cues that vary by language. For English, French, and Portuguese, a sentence is flagged when a verbal token carries the morphological feature `Voice=Pass` or bears a passive dependency label (`nsubjpass` for English; `nsubjpass`, `auxpass`, or `aux:pass` for French and Portuguese). Portuguese additionally treats the reflexive marker *se* in an expletive or marker dependency (`expl`, `mark`) as a passive signal. German detects the periphrastic passive by requiring both the auxiliary *werden* (lemma *werden*, POS AUX) and a past participle (a VERB token with `VerbForm=Part`) in the same sentence. Spanish combines two constructions: the periphrastic passive with *ser* (auxiliary *ser* plus a participle) and the reflexive passive *se*, the latter detected either through an expletive dependency or when *se* attaches to a verb that also carries a nominal subject.

PERSONAL_PRONOUN. This tag targets first- and second-person pronominal subjects and objects. Detection is based on two conditions: the token's POS is PRON and it occupies a subject or object dependency role (`nsubj`, `obj`, `iobj`), and its morphological `Person` feature is `1` or `2`. The English tagger additionally filters against an explicit list of surface forms (*I, you, we*). The German tagger departs from this scheme: rather than checking the `Person` feature or dependency role, it flags any pronoun (PRON) whose lemma belongs to a fixed list of personal pronouns, which also includes third-person forms (*er, sie, es, ihn, ihm, ihnen*).

IMPERSONAL. Impersonal constructions are identified via expletive dependency relations (`expl*`), supplemented by language-specific cues: expletive *it* and *there* in English, impersonal *se* in Spanish and Portuguese, together with existential verbs (*haber*/*hay* in Spanish, *haver*/*há* in Portuguese), and expletive *es* or the verb *geben* in German.

NEGATION. A sentence is tagged as containing a negative construction if any token bears the `neg` dependency label, carries the morphological feature `Polarity=Neg`, or (for Spanish, German, and Portuguese) matches a language-specific list of negative adverbs and pronouns (*no*, *nunca*, *jamás*, *tampoco*, *nada* in Spanish, *nicht*, *nie*, *niemals*, *nichts*, and the determiner *kein* in German, *não*, *nunca*, *jamais*, *ninguém*, *nada*, *nem* in Portuguese).

LONG_SENTENCE. A sentence is flagged as long when its token count exceeds a threshold of 25. This threshold was set arbitrarily and is uniform across all five languages.

ACRONYM. A sentence is tagged as containing an acronym when at least one token consists entirely of uppercase letters and has a length of at least two characters.

4. Prompting Strategy

Our prompting strategy is designed to apply targeted simplification operations. The pipeline consists of two stages: a set of *specialised prompts*, one per linguistic phenomenon, and a *combination prompt* that merges their outputs into a single coherent sentence.

4.1. Specialised prompts

For each tag in our set, we define a dedicated prompt that instructs the language model to address exactly one simplification operation on the source sentence. When a sentence is assigned one or more

Table 1
Specialised prompts per tag

Tag	Target operation
Subordination	Split into shorter, coordinate clauses
Passive	Convert to active voice
Personal_Pronoun	Maintain or reinforce direct address
Impersonal	Replace with a personal construction
Negation	Rephrase using a positive formulation
Long_Sentence	Shorten to at most 20 words
Acronym	Spell out or gloss the abbreviation

tags by the tagging pipeline, it is independently passed through each corresponding prompt. Table 1 lists the seven tags and the simplification operation targeted by their respective prompt.

Each specialised prompt receives the original sentence and the instruction to perform only the corresponding operation, leaving other aspects of the sentence unchanged as much as possible. This design keeps each model call focused and avoids the risk that a single, multi-objective prompt conflates competing transformations.

4.2. Combination prompt

When a sentence triggers $n > 1$ specialised prompts, the pipeline collects the n intermediate outputs and feeds them to a combination prompt. This final prompt presents the intermediate rewritings and instructs the model to produce a single, fluent output that integrates the relevant simplifications. If a sentence triggers only one tag, the output of the corresponding specialised prompt is used directly. If no tag is detected, the sentence is passed through a prompt listing all seven guidelines globally.

4.3. Implementation

All models were run locally on an NVIDIA L4 GPU without external API calls. The backbone model is **Gemma 3 27B** [3]; the selection rationale is detailed in Section 5.3.

4.4. Run configurations

We submitted two runs, differing only in the input provided to the combination prompt.

Run 1 (`no-orig`). The combination prompt receives only the outputs of the specialised prompts, without the original complex sentence. The model is thus constrained to synthesise the intermediate rewritings without direct access to the source.

Run 2 (`with-orig`). The combination prompt additionally receives the original complex sentence alongside the specialised outputs. This gives the model an explicit reference point and may help preserve content that the specialised prompts altered or dropped.

All other parameters (model weights, prompt templates, tagging pipeline) are identical across the two runs.

5. Experiments

5.1. Data

We used two distinct datasets across our experiments. For **model selection** (Section 5.3), we relied on last year's sentence-level training set from the Cochrane-auto corpus [4], which originally contained

Table 2
Pipeline evaluation on the full training set. Baseline is the unmodified complex sentence.

System	BLEU	SARI	FKGL	BS-F1	CEFR Δ	Improved	Same	Worse
Baseline (complex)	30.44	–	6.38	0.415	–	–	–	–
GPT-OSS:20B, no orig.	9.45	38.68	11.82	0.316	0.44	40.18%	54.18%	5.64%
GPT-OSS:20B, with orig.	9.68	37.02	12.45	0.292	0.29	32.13%	59.13%	8.74%
Gemma3:27B, no orig.	6.19	36.71	11.53	0.285	**0.62**	**57.62%**	32.25%	10.13%
Gemma3:27B, with orig.	5.90	36.02	12.11	0.264	0.41	46.43%	41.33%	**12.23%**

11,724 lines. After removing entries with embedded newlines, 6,932 sentences remained. The initial ten-model comparison was conducted on the first 50 of these sentences. The subsequent choice between the two best-performing models used the full 6,932-sentence set. For the **pipeline evaluation** (Section 5.4), we used the new English training data released by the SimpleText 2026 organisers.

5.2. Evaluation metrics

We report four complementary metrics. **BLEU** [5] measures n-gram overlap with a reference simplification (higher is better). **SARI** [6] measures simplification quality by jointly evaluating addition, deletion, and preservation operations relative to both the reference and the source sentence (higher is better). **FKGL** (Flesch-Kincaid Grade Level) estimates surface readability from word and sentence length, lower values indicate simpler text. **BERTScore** [7] measures semantic similarity between the system output and the reference using contextual embeddings (higher is better). Additionally, we report a **CEFR**-based metric that classifies each sentence on the A1–C2 scale and measures the mean shift toward simpler levels compared to the source. We use the classifier released as part of the TSAR 2025 shared task [8]. All reference-based metrics are computed against the human-written simplifications.

5.3. System development

Model selection. We evaluated ten open-weight models served locally via Ollama on a development subset of 50 sentences: Gemma 3 (4B, 12B, 27B), GPT-OSS 20B, DeepSeek-R1 8B, Llama 3 8B, Llama 3.1 8B, Llama 3.3 70B, Mistral 7B, and Qwen3 8B. All inference runs exclusively on an NVIDIA L4 GPU, which ruled out models above 30B parameters in practice. Among the viable candidates, GPT-OSS:20B achieved the highest BLEU and SARI on a generic prompt, while Gemma 3:27B, despite slightly lower surface metrics, produced the greatest CEFR-level improvement: 57.62% of sentences shifted to a simpler level, against 40.18% for GPT-OSS:20B. Since our primary goal is readability improvement rather than reference-matching, we selected **Gemma 3:27B**.

Prompt design. We experimented with prompt strategies of increasing complexity, from a generic simplification instruction to prompts that listed all guidelines globally or appended tag-specific constraints inline. A consistent pattern emerged: loading a single prompt with multiple simultaneous constraints improved n-gram overlap with the reference (SARI, BLEU) but steadily increased FKGL, with the all-guidelines variant reaching FKGL 14.31 compared to 11.16 for the generic prompt. The model appeared to produce more elaborate paraphrases when asked to address several phenomena at once, counteracting the readability goal. This observation motivated the specialised pipeline described in Section 4: by isolating each simplification operation in a separate call, we allow the model to focus on a single transformation at a time and prevent the trade-off between constraint satisfaction and output length.

5.4. Pipeline evaluation on training data

We evaluated the final tagged pipeline (prompt 4) with two models (Gemma 3:27B and GPT-OSS:20B) and two combination-prompt variants (with and without the original sentence) on the full training set.

Table 3
Tag rates in the source and in the two Gemma 3:27B pipeline outputs (full training set).

Tag	Source	No-orig	With-orig
Acronym	38.26%	8.46%	10.06%
Passive	16.40%	11.79%	13.52%
Impersonal	7.09%	2.30%	2.42%
Long sentence	45.54%	34.16%	44.22%
Personal pronoun	16.59%	22.62%	23.68%
Negation	7.86%	22.96%	23.64%
Subordination	10.87%	25.35%	28.32%

Results are reported in Table 2.

Although GPT-OSS:20B achieves better scores on traditional metrics, Gemma 3:27B yields a substantially higher CEFR improvement (mean $\Delta = 0.62$ vs. 0.44 without the original sentence), improving 57.62% of sentences against 40.18%. This difference indicates that Gemma 3:27B produces outputs that are simpler in terms of lexical and syntactic complexity, even when surface-level n-gram overlap with the reference is lower. The "no original" variant clearly outperforms the "with original" variant on both EASSE and CEFR for Gemma 3:27B (CEFR mean $\Delta = 0.62$ vs. 0.41).

Table 3 shows how the rate of each feature changes between the source sentences and the two Gemma 3:27B pipeline variants.

Acronyms, impersonal constructions, and the passive are reliably reduced in both runs. Long sentences are reduced in the no-original run (from 45.54% to 34.16%) but barely change when the original is provided (44.22%), consistent with the more conservative behaviour of that variant. Three features, however, increase: personal pronouns, negation, and subordination. The rise in negation (from 7.86% to roughly 23% in both runs) likely reflects the model rephrasing passive or impersonal constructions using negative wordings as an intermediate step. The increase in subordination (from 10.87% to 25–28%) stems from a bug in the English pipeline: although subordinate clauses were correctly *detected* in the source (10.87% of sentences), the specialised subordination prompt was never applied to English inputs, so these constructions were detected but never simplified. Sentence splitting was therefore never triggered, and the combination prompt tended to introduce additional subordinate clauses rather than remove them. Splitting subordinate clauses was an intended component of the method. We noticed the bug after the official runs had been submitted, so the actual contribution of the subordination criterion to simplification quality could not be assessed here and remains to be evaluated in future work. Together these observations show that the specialised prompts for different phenomena can interact in unintended ways, and constitute a direction for future improvement.

5.5. CEFR level distribution

To characterise the readability improvement in more detail, we classify every source and system sentence on the A1–C2 scale using the TSAR 2025 CEFR classifier [8] and compare the resulting level distributions on the English training set. Figure 1 shows the per-level counts for the source and for the two Gemma 3:27B pipeline variants.

The source is concentrated at the upper levels (B2–C1 account for 65% of the sentences), whereas both variants move substantial mass toward B1 and A2. To summarise these shifts numerically, we represent the six levels as consecutive integers from A1=0 to C2=5, so that a sentence's *delta* (Δ) is the number of levels by which the predicted output is simpler than its source (a positive value denotes simplification). The shift is stronger for the *no-orig* variant, as summarised in Table 4: its mean level drops from the source value of 2.99 to 2.37 (against 2.58 for *with-orig*), and it records a mean improvement of $\Delta = 0.62$, moving 57.62% of sentences to a simpler level against 46.43% for *with-orig*. Exposing the combination prompt to the original sentence therefore yields a markedly more conservative rewriting, in line with the higher compression ratio observed on the test set (Section 4.4).

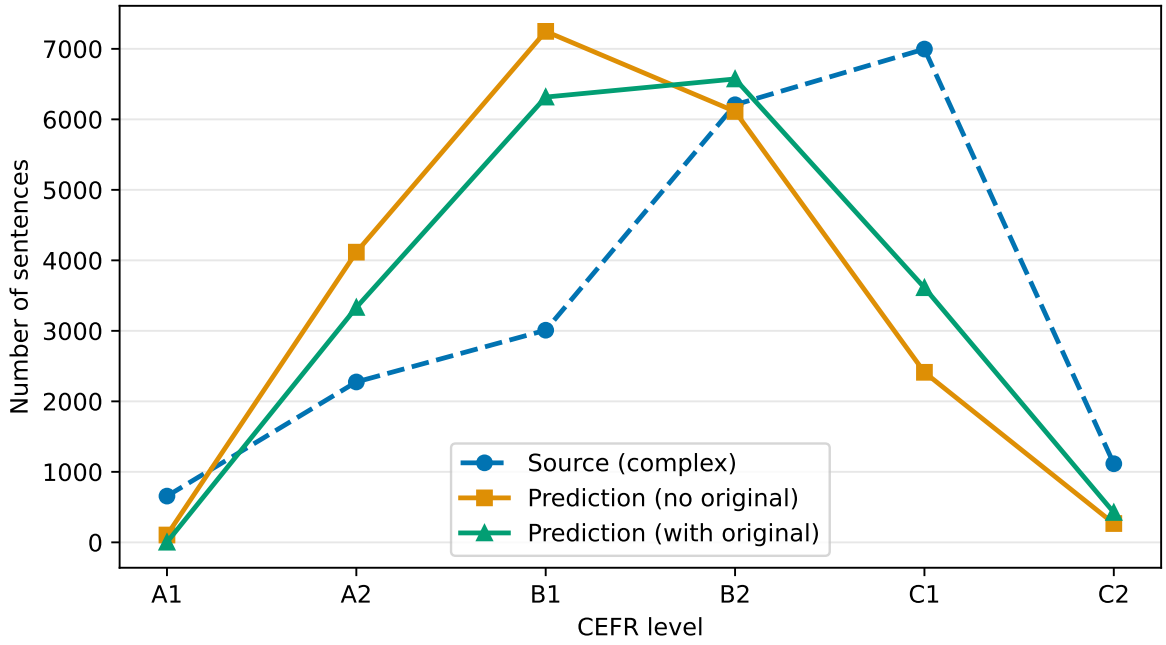

Figure 1: CEFR level distribution on the English training set: the source sentences (dashed) and the two Gemma 3:27B pipeline variants. Both variants shift probability mass from the upper levels (C1–C2) toward B1–A2, the effect being stronger for the no-original variant.

Table 4
CEFR summary on the English training set for the two Gemma 3:27B variants, with CEFR levels as integers A1=0, …, C2=5 (source mean level: 2.99). Δ is the mean sentence-level shift toward a simpler level; *Improved*/*Same*/*Worse* give the share of sentences moving to a simpler, identical, or more complex level.

Variant	Mean level	Δ	Improved	Same	Worse
No-orig	2.37	0.62	57.62%	32.25%	10.13%
With-orig	2.58	0.41	46.43%	41.33%	12.23%

Table 5
Official test set results.

Run	BLEU	SARI	FKGL	Comp.
Run 1 (`no-orig`)	7.16	43.86	11.06	0.760
Run 2 (`with-orig`)	7.80	44.20	12.71	0.884

5.6. Results on the test set

Table 5 reports the official evaluation scores on the English test set provided by the SimpleText 2026 organisers.

Both runs achieve comparable SARI scores (∼44), with Run 2 obtaining a marginally higher BLEU (7.80 vs. 7.16) but also a higher FKGL (12.71 vs. 11.06), indicating slightly less readable output. The compression ratio confirms that Run 2 preserves more of the source content (0.884 vs. 0.760), consistent with the combination prompt having access to the original sentence.

6. Conclusion

We presented a text simplification system that grounds its prompting strategy in plain language research, using rule-based syntactic taggers to identify complexity features sentence by sentence before passing

each one to a specialised LLM prompt. The approach yields SARI scores around 44 on the English test set and, on the training data, improves the CEFR reading level of over 50% of sentences, outperforming a stronger baseline (GPT-OSS:20B) on this dimension despite lower n-gram overlap metrics.

The analysis of tag rates reveals both strengths and limitations of the pipeline. Acronym expansion, impersonal constructions, and passive voice are reliably reduced. Negation, however, consistently increases across outputs, suggesting that prompts targeting different phenomena can introduce unintended interactions. Additionally, as noted in Section 5.4, a bug in the English pipeline meant that subordination, although correctly detected, was never targeted by its specialised prompt. Subordinate clauses were therefore left unsimplified and their rate even rose in the output, which likely limited simplification quality for that language. As this bug was identified only after the official runs had been submitted, the real impact of the subordination criterion remains to be measured.

Future work should include redesigning the combination prompt to explicitly penalise the introduction of new complexity features. The multilingual extension of the tag set to languages beyond our five covered languages, potentially using cross-lingual models, is another natural direction.

Acknowledgments

This work has been supported by grant PID2023-148577OB-C21 (Human-Centered AI: User-Driven Adapted Language Models-HUMAN_AI) by MICIU/AEI/10.13039/501100011033 and by FEDER/UE.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to assist with writing (grammar and style). After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. Bakker, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 1: Simplify Scientific Text, in: E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[2] A. Müller, T. François, Les avancées en rédactologie influencent-elles les guides de rédaction claire en belgique francophone ? Une étude sur corpus, in: I. Clerc (Ed.), La Communication État-citoyens : défis numériques, perspectives rédactologiques, Presses de l'Université Laval, 2022, pp. 177–192.

[3] Gemma Team, Google DeepMind, Gemma 3 technical report, https://ai.google.dev/gemma, 2025.

[4] J. Bakker, J. Kamps, Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts, in: M. Shardlow, H. Saggion, F. Alva-Manchego, M. Zampieri, K. North, S. Štajner, R. Stodden (Eds.), Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 41–51. URL: https://aclanthology.org/2024.tsar-1.5/. doi:10.18653/v1/2024.tsar-1.5.

[5] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318. doi:10.3115/1073083.1073135.

[6] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, in: Transactions of the Association for Computational Linguistics, volume 4, 2016, pp. 401–415. doi:10.1162/tacl_a_00107.

[7] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: Proceedings of the 8th International Conference on Learning Representations, 2020.

[8] F. Alva-Manchego, R. Stodden, J. M. Imperial, A. Barayan, K. North, H. Tayyar Madabushi, Findings of the TSAR 2025 shared task on readability-controlled text simplification, in: M. Shardlow, F. Alva-Manchego, K. North, R. Stodden, H. Saggion, N. Khallaf, A. Hayakawa (Eds.), Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), Association for Computational Linguistics, Suzhou, China, 2025, pp. 116–130. URL: https://aclanthology.org/2025.tsar-1.8/. doi:10.18653/v1/2025.tsar-1.8.