Beyond Lexical and Syntactic: Using Psycholinguistics for Better Text Simplicity Analysis

Anonymous ACL submission

Abstract

Automatic text simplification measures correlate poorly with human judgments and are disconnected from psycholinguistic research on comprehension. We propose a multidimensional analysis framework grounded in Cognitive Load Theory and the Construction-Integration model, organizing over 400 linguistic features into six simplicity dimensions targeting distinct comprehension stages. Using this framework, we analyze four human-annotated and six aligned simplification datasets. We find that (1) our dimensions predict human simplicity judgments better than established measures on most datasets; (2) current datasets overwhelmingly target lexical and syntactic dimensions while under-targeting personalization the dimension psychological research identifies as most impactful for comprehension; and (3) existing measures overall do not capture any specific dimension. These findings provide actionable guidance for building measures and datasets better aligned with how readers actually process text.

1 Introduction

Despite rising global literacy, many texts remain inaccessible to general audiences due to their complexity. Recent approaches in Automatic Text Simplification (ATS) address this using Large Language Models (LLMs) prompted or fine-tuned on corpora of aligned complex–simple text pairs (Saggion, 2017; Ermakova et al., 2026). These models are then evaluated on similar datasets, using various measures that compare the generated text to reference simplified versions such as SARI (the most widely used ATS measure) (Xu et al., 2016) BLEU (Papineni et al., 2002), or without references like FKGL (Kincaid et al., 1975). However, these measures have repeatedly shown low correlation with human ratings of *simplicity* (Alfear et al., 2024; Alva-Manchego et al., 2021), but also with *meaning preservation* (Agrawal and Carpuat, 2024) or *hallucination* (Vendeville et al., 2025). They are also often disconnected from findings in psycholinguistic and cognitive research, which highlight that simplicity depends on multiple interacting dimensions such as cohesion, coherence or even personalization (Strohmaier et al., 2023). Moreover, both measures and datasets tend to focus on one or two dimensions of simplicity (typically lexical and syntactic simplicity) (Davari et al., 2024a), producing single isolated scores. These offer limited insight into what is actually happening during simplification, making it difficult to reliably evaluate ATS.

We argue that effective evaluation requires a psycholinguistically grounded, multidimensional framework. Drawing on Cognitive Load Theory (CLT) (Ozuru et al., 2009) and the Construction-Integration (CI) model (Kintsch, 1998), we propose such a framework and use it to reveal shortcomings in current measures and datasets. By identifying which dimensions of simplicity are captured and which are missed, we show how measures and datasets diverge from psycholinguistic evidence. We also show such analysis can provide high correlation with human simplicity judgments, beating SOTA in many scenarios. This framework provides actionable insight for building measures that better reflect human comprehension and for designing datasets that target the dimensions most impactful for readers. Specifically, we address three research questions: **RQ1**: Do psycholinguistically-grounded simplicity dimensions predict human simplicity judgments, and does their relative importance vary across datasets? We train per-dimension regressors on 4 annotated datasets and compare their predictive power to established measures. **RQ2**: Which simplicity dimensions do different simplification datasets actually transform during simplification and does the magnitude change between datasets? We measure per-dimension effect sizes ($|\Delta z|$) on 6 aligned datasets, validated by per-dimension binary classification.

RQ3: Which simplicity dimensions do existing evaluation measures capture? We correlate measure scores with dimension-level predictions across datasets to reveal measure biases. We will release the datasets as well as the complete code used for scoring and analysis.

2 Related Work

2.1 Limitations of Current ATS Measures

Effective evaluation of text simplification requires measures that are both predictive of human judgments and interpretable enough to guide system development, but current approaches fall short on both counts. Indeed, most measures produce a single score that conflates distinct aspects of simplicity (often just lexical difficulty and syntactic complexity) into one number, making it impossible to diagnose *why* a simplification succeeds or fails. This is true of classical readability formulas (Kincaid et al., 1975; Flesch, 1948), which reduce comprehension difficulty to word and sentence length; of reference-based measures like SARI (Xu et al., 2016), BLEU (Papineni et al., 2002), and ROUGE (Lin, 2004), which measure textual similarity without distinguishing what kind of linguistic transformation was applied. Similarly, neural approaches like LENS (Maddela et al., 2023) and BERTScore (Zhang et al., 2020), improve predictive power but remain opaque. Across all these classes, correlation with human ratings of simplicity remains low (Alfear et al., 2024; Alva-Manchego et al., 2021), Beyond simplicity, Agrawal and Carpuat (2024) and Davari et al. (2024b) demonstrate that standard measures also fail to capture meaning preservation across multiple datasets. Vendeville et al. (2025) show that faithfulness measures (FactCC, QAGS, FEQA) also perform poorly on simplification data, and propose a detailed error taxonomy in the SALTED benchmark. These failures point to a structural problem rather than a calibration one. Simplicity is not a single construct: psycholinguistic research shows that comprehension depends on the interaction of multiple dimensions such as lexical, syntactic, but also others such as discourse, and prior knowledge (Kintsch, 1998; Ozuru et al., 2009). Despite that, no existing measure decomposes its score along these lines. Without such decomposition, measures cannot explain domain-specific failures or provide actionable feedback for improving simplification systems.

2.2 Toward Multidimensional Assessment

A few approaches have moved beyond single-score evaluation. TextEvaluator (Sheehan et al., 2014) combines 43 features across lexical, syntactic, and discourse categories, but it relies heavily on closed-source tools such as word dictionaries. BATS (Kreutz et al., 2024) implements 37 binary simplicity characteristics from the literature and analyzes their predictive importance across domains and audiences, revealing that feature relevance varies substantially by context.

However, these approaches share critical limitations. They identify predictive features without organizing them into theoretically coherent dimensions. BATS observes domain-dependent variation but does not explain *why* certain features matter more in certain contexts. Neither framework includes cohesion and coherence measures, despite psycholinguistic evidence that these are distinct, interacting components of comprehension (McNamara and Kintsch, 1996; Ozuru et al., 2009). Most importantly, they lack grounding in cognitive theory that could provide interpretability by connecting observable text features to the processing stages and cognitive loads they affect during reading. Our framework addresses these gaps by deriving simplicity dimensions from the Construction-Integration model (Kintsch, 1998) and Cognitive Load Theory (Ozuru et al., 2009), grounding each dimension in a specific comprehension mechanism.

3 Methodology

We ground our analysis in psycholinguistic theory (Section 3.1) to derive six simplicity dimensions (Section 3.3), which we operationalize using 402 text features. Using this framework, we examine whether these dimensions predict simplicity ratings (**RQ1**), which dimensions are transformed during simplification and by how much (**RQ2**), and which dimensions existing evaluation measures capture (**RQ3**). We evaluate across 6 text simplification datasets spanning different domains, granularities, and target audiences. Specifically, we include Simplicity-DA (Alva-Manchego et al., 2021) (SDA) and D-Wiki (Sun et al., 2021), which are datasets built from the Simple English Wikipedia dataset, focusing on sentence-level and document level simplifications, respectively. We also include the SimpleText dataset (Ermakova et al., 2024), which is built on manually written simplifications of scientific abstracts, and includes a sentence level

Dataset	Samples	Domain	Align	Level	Raters
SDA	600	Wiki	*C,S,R*	*s*	15
$ST_{sent, para}$	1.6K, 271	Science	*C,S,R*	*s,m*	N/A
D-Wiki	601	Wiki	*C,S,R*	*m*	3
ASSET	4,500	Wiki	*C,S,R*	*s*	-
OneStopEnglish	2663	Edu.	*E,I,A*	*d*	-
Cochrane	3967	Medical	*C,S*	*d*	-
$Cochrane_{auto}$	849	Medical	*C,S*	*d*	-

Table 1: Dataset properties. Align. shows available texts: *C*=Complex, *S*=Simple, *R*=Reference. Level shows text granularity: *s*=sentence, *m*=multi-sentence. Number of raters not provided for SimpleText data, and no ratings for OSE and Cochrane datasets.

(ST_{sent}) and a paragraph level (ST_{para}) split. All of those datasets are provided with human ratings of simplicity for each sample. In the case of ST, an unspecified number of writers explicitly provided evaluations of *lexical simplicity* and *syntax simplicity*, which we averaged to get a single simplicity score. We also include 3 datasets that do not contain any simplicity ratings: Cochrane (Devaraj et al., 2021), a dataset of medical abstract simplified by experts, $Cochrane_{auto}$ (Bakker and Kamps, 2024), a re-aligned version of Cochrane, and finally OneStopEnglish (Vajjala and Lucic, 2018) (OSE), a dataset composed of educational texts written at three levels Elementary, Intermediate, and Advanced (respectively OSE_E, OSE_I and OSE_A). Table 1 summarizes those datasets.

3.1 Cognitive Models

We use two frameworks: the Construction-Integration (CI) model (Kintsch, 1998), which describes the phases of text comprehension, and Cognitive Load Theory (CLT) (Ozuru et al., 2009), which models how different readers allocate cognitive resources during reading. Figure 1 summarizes how these map to our simplicity dimensions.

In the CI model, reading comprehension is presented as a three-stage process: *construction*, where readers identify concepts from the text, *integration*, where readers link identified concepts into a textbase graph representing the text, and *situation model construction* where the reader integrates the text representation with prior knowledge into a coherent mental model called a *situational model*.

The CLT models task difficulty as a function of working-memory effort. Working memory is the limited-capacity system used to process and maintain information during cognitive tasks. Tasks impose cognitive loads that consume this capacity. It distinguishes three forms of cognitive load: *intrinsic load*, determined by the inherent complexity of the content or task; *extraneous load*, which is the load imposed by how information is presented (e.g., linguistic form, discourse structure, layout); and *germane load*, in contrast, represents the desirable cognitive effort dedicated to constructing and refining mental representations, and is therefore essential for meaningful comprehension.

3.2 Text Comprehension Stages

In Strohmaier et al. (2023), a meta-analysis of 45 studies with 6,477 learners, the authors examine how linguistic modifications impact comprehension and learning from STEM texts. They organize linguistic modifications into three categories, each targeting a specific processing stage and cognitive load type. Figure 1 summarizes the multidimensional nature of comprehension and illustrates how different cognitive stages and loads map to potential ATS interventions.

During the *construction* stage, readers decode words and linguistic units, constructing propositions from lexical and syntactic forms. This process is heavily influenced by text presentation and induces high *extraneous load*. Reducing word-level complexity (word frequency, scientific vocabulary...) or syntactic complexity (sentence length, syntax simplification...) can minimize this load (Berglund-Barraza et al., 2019). However, Strohmaier et al. (2023) shows that such surface-level complexity reduction shows negligible or inconsistent effects on comprehension and learning.

During the *integration* stage, readers link concepts into a connected textbase representing the text. This process contributes moderately to *germane load*. Increasing cohesion, referential overlap, and connectives aim to reduce *extraneous load* by making semantic connections explicit (Jake Follmer and Sperling, 2018). Yet cohesion modifications alone show weak or null effects, suggesting that just freeing working memory capacity provides minimal benefit.

Importantly, prior knowledge plays an important part here. For expert readers, reducing extraneous load at this stage might bring no improvement, or even hinder comprehension (Rey and Buchwald, 2011), in what is called *expertise reversal effect*. This effect is not symmetrical: the meta-analysis finds that additional assistance helps novices more than it hurts experts (Tetzlaff et al., 2025).

During the *situation model construction*, readers

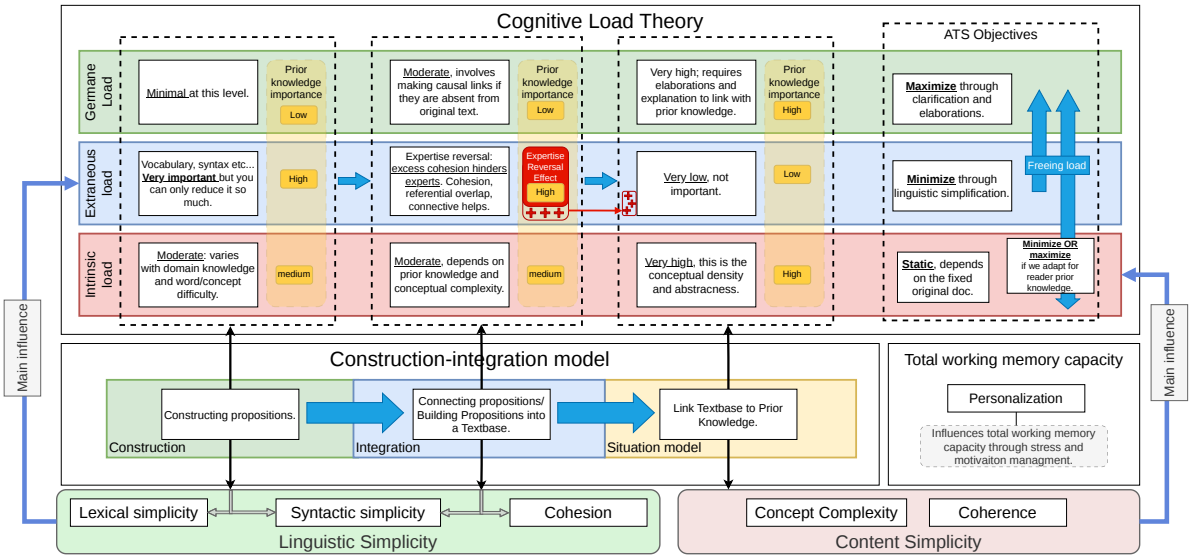

Figure 1: Unified Simplicity Framework. A mapping of Linguistic and Content simplicity dimensions to the Construction-Integration (CI) model and Cognitive Load Theory (CLT) illustrating how specific text features target distinct comprehension stages to modulate Extraneous and Intrinsic cognitive loads.

integrate the text representation with prior knowledge to build a coherent mental model (O'reilly and Mcnamara, 2007; Graesser et al., 2011). This stage depends heavily on both *germane load* (productive effort) and *intrinsic load* (inherent content difficulty) (Strohmaier et al., 2023). Increasing clarity and elaboration (explaining concepts, specifying causal relations, adding summaries, personalization) supports germane load. Notably, these modifications demonstrate substantial positive effects. Additionally, conceptual density (number of concepts mentioned per unit of text) and concept abstractness constitute the intrinsic load and significantly influence comprehension difficulty.

We therefore distinguish two main categories: *Linguistic simplicity* and *Content simplicity*. Linguistic simplicity is composed of modifications that mainly reduce extraneous load across comprehension stages by transforming the actual words being used and their order. Content simplicity, in contrast, addresses mainly intrinsic load by managing the concepts being used and their interactions. Together, these dimensions reflect how comprehension is structured into simplicity dimensions, influencing each processing stages and cognitive load types. Each of these dimensions can be further decomposed into specific features.

3.3 Simplicity Dimensions

We derive six simplicity dimensions (summarized in Table 2) from the literature described above, and operationalize each using subsets of the 402 linguistic features extracted publicly available tools and resources listed in Appendix.

3.3.1 Linguistic Simplicity

Lexical Simplicity captures word-level difficulty during the construction stage, where readers decode individual words and terms. Unfamiliar words, technical jargon, and uncommon vocabulary increase extraneous load at this stage (Berglund-Barraza et al., 2019). We operationalize this dimension using psycholinguistic norms (age of acquisition, concreteness, imageability from MRC (Wilson, 1988)), vocabulary sophistication (academic word coverage (Coxhead, 2000), CEFR profiling (Dürlich and François, 2018)), and word-level statistics (frequency, length, syllable counts).

Syntactic Simplicity captures sentence-level structural complexity during the construction and integration stages. Complex syntax (deep embedding, passive constructions, long dependency chains,...) increases extraneous load by requiring readers to maintain multiple open syntactic relations in working memory. We operationalize this using dependency tree properties (arc length, depth, branching) (Yngve, 1960), clause counts, passive constructions, and sentence length measures.

Cohesion captures the explicit linguistic cues that help readers connect ideas during the integration stage, where propositions are linked into a

coherent textbase. Things such as referential overlap, causal connectives, and semantic similarity reduce extraneous load by making relationships between sentences explicit (Jake Follmer and Sperling, 2018). We operationalize this primarily using TAACO features (Crossley et al., 2016): lexical overlap at sentence and paragraph level, semantic similarity (LSA, LDA, word2vec), connective counts, and givenness indicators.

Personalization and Motivation captures the use of direct address (you, we) and interpersonal language to make the text feel more personal and motivate the reader. While not targeting a specific cognitive load type, personalization has been shown to influence total working memory capacity through motivation and engagement (Strohmaier et al., 2023). We operationalize this using pronoun distributions (personal, possessive), and semantic category indicators for communicative, motivational, and affective language from the Harvard IV-4 dictionary (Stone et al., 1966). While these may seem like shallow indicators, this does follows the analysis of Sheehan et al. (2014).

3.3.2 Content Simplicity

Concept Simplicity captures the inherent difficulty of the ideas themselves, constituting the intrinsic load throughout all CI stages (Sweller, 1988). Abstraction level, domain-specificity, and information density (number of concepts per unit of text) determine how much cognitive effort is required regardless of how the text is written. (entities per sentence, entity density, unique entity counts), concreteness and imageability norms from MRC (Wilson, 1988), and academic sophistication indicators from TextEvaluator (Sheehan et al., 2014) (nominalization rates, abstract noun ratios, academic word coverage).

Discourse Coherence captures the logical and semantic organization of ideas in the text. Unlike cohesion, which relies on explicit linguistic markers, coherence refers to the underlying conceptual structure. We look at whether ideas follow a logical progression (Kintsch, 1998). Strong coherence supports situation model construction by reducing germane load. We operationalize this using CoCo features (Stajner et al., 2020) capturing entity-based coherence (activation dynamics, persistence, semantic connectivity, local and global coherence scores), Brownian Bridge coherence scoring (Sheng et al., 2025), CoUDA distance-based coherence (Zhu et al., 2024), and TAACO paragraph-level semantic similarity (LSA, LDA, word2vec).

Dimension	#Features Raw	#Features Clean	Primary sources
Linguistic			
Lexical	86	23	CEFR, AWL
Syntactic	64	18	Syntax Trees, TAACO
Cohesion	182	22	TAACO, Semantic Cat.
Content			
Coherence	47	11	TAACO, CoUDA, BBscore
Concept	42	7	CoCo, AWL
Person.	24	12	Semantic Cat.
Total	402	92	

Table 2: Dimension feature set summary.

3.4 Feature Processing

Word count correlates substantially with simplicity ratings, and many linguistic features are correlated with length. To isolate dimension-specific contributions beyond brevity, we apply the following normalization pipeline to each dataset independently: we first residualize all features with respect to word count via OLS regression, removing the linear effect of length. We then discard features with near-zero variance across all datasets and, for each pair with high collinearity ($|r| > 0.85$), randomly remove one. Finally, we z-score all remaining features using the pooled source–simplified distribution per dataset. After cleaning and selection, 92 features remain (from 402 initial).

3.5 Validating Dimensions Against Ratings

To verify that our dimensions capture meaningful aspects of perceived simplicity, we examine their correlation with human ratings. For each dataset with simplicity ratings, we train Linear Regression and Random Forest regressors to predict ratings using features from each simplicity dimension separately. We use 5-fold CV with randomized hyperparameter search and nested loops. We report the absolute Pearson correlation between predicted and actual ratings. This allows us to determine, for example, whether Lexical Simplicity predicts significantly better than Coherence on scientific text. To test whether dimensions differ significantly in predictive power within each dataset, we include Williams significance tests in appendix.

3.6 Effect Size of Simplification

For each feature, we compute the standardized mean difference $|\Delta z|$ between simplified and

source texts, where features are z-scored per dataset using the pooled distribution. Averaging $|\Delta z|$ across features within each simplicity dimension yields a cross-dataset comparable measure of simplification magnitude per dimension. As validation, we train per-dimension binary classifiers (Random Forest) to discriminate source from simplified texts. Classification accuracy above random chance confirms that a dimension captures genuine information about transformations during simplifications.

3.7 Mapping Measures to Dimensions

Finally, we examine which dimensions existing evaluation measures capture. We compute scores from SARI, BLEU, ROUGE-L, FKGL, BERTScore, LENS-SALSA, and LLM-as-judge (Claude Haiku) and correlate measure scores with dimension scores across all texts in each dataset. Strong correlation indicates sensitivity to that aspect of simplicity. This analysis reveals measure biases and explains domain-specific performance: a measure capturing only lexical features will underperform when ratings are driven by coherence.

4 Results

4.1 RQ1: Correlation with Human Judgment

Table 3 shows the absolute Pearson correlation with human judgments of simplicity on the included datasets. We trained a Random Forest (RF) and a linear regression (LR) model on the entire set of features and on each group of features separately. We also trained stacked models by taking the dimension-level models as base learners, then training a meta-learner on their outputs. We provide baselines and established measures for reference, including SARI, D-SARI, BLEU, ROUGE-L, and LLM-based measures (SLE, Claude). On sentence-level datasets, our RF model trained on all features consistently outperforms established measures: on SDA it edges out Claude (0.74 vs 0.73), and on ST_{sent} it surpasses the best baseline, BLEU (0.42 vs. 0.35). On D-Wiki, it likewise outperforms all baselines (0.60 vs. Claude's 0.52). However, on ST_{para}, all our models yield low correlations ($|r| \leq 0.13$), falling well below SARI (0.32). This suggests that the features, while predictive at the sentence level, do not transfer straightforwardly to paragraph-level evaluation on this dataset.

An important caveat is that our models are trained in cross-validation per dataset, so we do not guarantee transferability of learned weights to new datasets. Annotator instructions, backgrounds, and source text characteristics may all differ across datasets. However, a cross-dataset transfer analysis (Section D) suggests that dimension-level representations partially generalize even when learned weights do not. Dimension-level regressors trained out of domain, combined with a locally trained meta-learner recover 80% of in-domain performance, indicating that the dimensions capture transferable properties of simplicity. Still, these results show that the selected features are salient enough to predict human ratings of simplicity better than established measures on three out of four datasets (*RQ1*). Regarding of individual dimensions, taking *All Features* generally outperforms any single one, with the exception of ST_{para} where *Personalization / Motivation* (RF) achieves the highest correlation (0.13 vs. 0.10 for All Features). Stacking does not consistently improve over the base RF model, though RF→LR is competitive across datasets.

The low performance on ST_{para} likely stems from its synthetic construction: paragraphs were formed by concatenating ST_{sent} sentences with averaged annotations (Davari et al., 2024a), creating systematic mismatch between paragraph-level feature computation and aggregated sentence-level labels. Discourse features (cohesion, coherence) measure properties absent in ratings. The good performance of the D-Wiki (0.60) document-level dataset confirms this limitation is specific to the aggregation procedure, not to longer text spans.

4.2 RQ2: Analyzing Effect Size of Dimensions

Table 4 reports the average standardized mean difference ($|\Delta z|$) per dimension and dataset. Since features are z-scored per dataset using the pooled source–simplified distribution, $|\Delta z|$ represents the average shift in pooled standard deviations (SD).

Overall simplification intensity varies substantially across datasets, from 0.32 SD (SDA) and 0.40 SD (ASSET) to 0.82 SD (Cochrane). The OSE datasets show a clean gradient where the effect size scales with the proficiency gap between levels (E→A: 0.74 > E→I: 0.56 > I→A: 0.52). Across datasets, *Lexical Simplicity* consistently shows the largest effect sizes, followed by *Syntactic Simplicity* in line with the traditional emphasis on these two dimensions in ATS research. *Concept Simplicity* follows closely, on par with or slightly below lexical and syntactic effects. Notably, *Personalization* shows the smallest effect size across most datasets, OSE being the excep-

	Measure	SDA	ST_{sent}	ST_{para}	D-Wiki
Baselines	FKGL	0.07	0.32	0.01	0.09
	Word Count	0.03	0.34	0.02	0.34
	Avg word len.	0.07	0.00	0.17	0.06
	SLE	0.03	0.34	0.02	0.09
	BLEU	0.26	**0.35**	0.23	0.06
	ROUGE-L	0.28	0.33	0.15	0.21
	SARI	0.33	0.22	**0.32***	0.37
	D-SARI	0.33	0.23	0.17	0.37
	Claude	**0.73**	0.18	0.19	**0.52**
RandomForest	All Features	**0.74***	**0.42**	0.10	**0.60***
	Lexical	0.73	0.37	0.03	0.55
	Syntactic	0.71	0.29	0.09	0.53
	Cohesion	0.21	0.32	0.05	0.38
	Concept	0.23	0.31	0.01	0.26
	Disc. Coherence	0.14	0.30	0.08	0.38
	Personalization	0.07	0.19	**0.13**	0.18
LinearRegression	All Features	0.56	0.02	**0.12**	0.27
	Lexical	**0.60**	**0.16**	0.08	**0.42**
	Syntactic	0.59	0.00	0.06	0.36
	Cohesion	0.19	0.00	0.07	0.12
	Concept	0.14	0.12	0.03	0.08
	Disc. Coherence	0.07	0.08	0.06	0.05
	Personalization	0.03	0.05	0.06	0.13
Stack	RF→RF	0.71	**0.46***	0.18	**0.59**
	RF→LR	**0.73**	0.45	**0.25**	**0.59**
	LR→LR	0.60	0.14	0.12	0.40
	LR→RF	0.62	0.27	0.07	0.43

Table 3: Absolute Pearson correlation with human ratings. RF = RandomForest, LR = LinearRegression. **Bold**: best in category per dataset. *Best on dataset. Document-level datasets underlined.

tion. The effect size ranges from small (0.25, SDA) to moderate (0.59, $\text{Cochrane}_{\text{auto}}$), being high only on Cochrane (0.69) and $\text{OSE}_{\text{E}\rightarrow\text{A}}$, despite being among the strongest predictors of human simplicity judgments in psychological research. This gap between scientifically-observed importance, and effect size suggests that current simplification practices under-target this dimension. *Cohesion* and *Discourse Coherence* similarly show below-average effect sizes, with the exception of the Cochrane dataset where Cohesion effects are large (0.90) equal with Lexical and Syntactic Simplicity. Likewise, *discourse coherence* shows higher effect sizes (0.38 - 0.61 SD) on document level datasets compared to sentence level ones (0.34-0.53 SD). This may reflect the specific training and guidelines of the Cochrane PLS writers. Table 5 reports per-dimension classification accuracy (F1 macro) for discriminating source from simplified texts. Using all features, Random Forest achieves near-perfect separation on most datasets (F1 ≥ 0.87), confirming that the feature set captures systematic differences introduced by simplification.

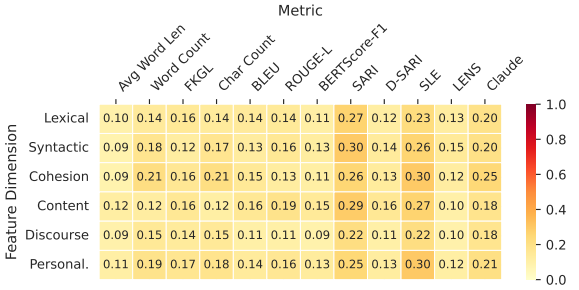

Figure 2: Mean absolute Pearson correlation ($|r|$) between dimension and measures, averaged across 4 datasets (D-Wiki, SDA, ST-para, ST-sent).

The OSE datasets are harder to classify (0.61–0.78), consistent with their lower effect sizes (Table 4). Per-dimension classifiers confirm that each dimension captures genuine simplification signal, with one notable exception: *Personalization* achieves near-chance performance on several datasets (F1 = 0.53–0.63). This indicates that personalization changes are too small in magnitude relative to between-text variability to support population-level discrimination.

4.3 RQ3: measure correlation

Figure 2 shows the average correlation between measure scores and dimension-level predictions across four datasets. Correlations are low across the board: no traditional measure exceeds $|r| = 0.30$ on any dimension. More importantly, all measures show nearly flat profiles across dimensions and none is selectively sensitive to any specific simplicity dimension. Among traditional measures, SARI shows the strongest correlations (0.22–0.30) but without meaningful differentiation between dimensions. Reference-based measures (BLEU, ROUGE-L, BERTScore) perform on par with length baselines, suggesting they respond to general text similarity rather than simplicity. Among neural approaches, SLE is the only measure showing a mild dimensional preference, with slightly higher sensitivity to Cohesion and Personalization. These flat profiles are notable given that dimensions vary substantially in their predictive power (RQ1) and transformation magnitude (RQ2). This means existing measures are not merely weak, they are responding similarly to dimensions that matter most for comprehension and those that matter least. This limits their diagnostic value and may partly explain their poor correlation with simplicity ratings, which as shown in RQ1 are driven by different

Dimension	ASSET	$OSE_{E\to A}$	$OSE_{E\to I}$	$OSE_{I\to A}$	SDA	ST_{sent}	ST_{para}	D-Wiki	Cochrane	$Cochrane_{auto}$
FKGL (resid.)	0.80	0.89	0.70	0.72	0.65	0.91	0.63	0.70	0.88	0.74
Word Count	0.60	0.88	0.54	0.58	0.45	0.76	0.81	0.77	0.98	1.10
Lexical	**0.54**	**0.93**	**0.66**	**0.64**	**0.46**	**0.56**	0.61	**0.69**	**0.90**	0.76
Syntactic	0.44	0.74	0.56	0.53	0.37	0.46	**0.64**	0.62	**0.90**	0.77
Cohesion	0.37	0.75	0.57	0.53	0.27	0.40	0.59	0.56	**0.90**	**0.78**
Concept	0.46	0.82	0.61	0.57	0.42	0.50	0.58	0.62	0.88	0.75
Disc. Coherence	0.37	0.53	0.42	0.38	0.34	0.53	0.67	0.61	0.57	0.58
Personalization	0.31	0.73	0.55	0.51	0.25	0.33	0.44	0.47	0.69	0.59
Avg. All Features	0.40	0.74	0.56	0.52	0.32	0.43	0.58	0.58	0.82	0.71

Table 4: Average absolute effect size ($|\Delta z|$) per dimension and dataset. **Bold**: highest effect size dimension per dataset. Document-level datasets underlined. FKGL and Word Count included as baselines

	Measure	ASSET	$OSE_{E\to A}$	$OSE_{E\to I}$	$OSE_{I\to A}$	SDA	ST_{sent}	ST_{para}	D-Wiki	Cochrane	$Cochrane_{auto}$
Baselines	Random	0.49	0.50	0.50	0.50	0.48	0.50	0.52	0.49	0.50	0.48
	Shuffle	0.50	0.51	0.51	0.51	0.53	0.51	0.52	0.53	0.50	0.52
	Avg Word Len.	0.60	0.63	0.52	0.55	0.62	0.60	0.69	0.62	0.53	0.54
	Word Count	0.58	0.62	0.57	0.56	0.58	0.54	0.46	0.65	**0.68**	**0.72**
	FKGL (resid.)	0.64	0.65	0.58	0.57	0.61	**0.65**	0.66	0.67	0.56	0.51
	FKGL	**0.69**	**0.70**	**0.62**	**0.59**	**0.64**	**0.65**	**0.70**	**0.73**	0.59	0.57
RandomForest	All Features	**0.98***	**0.78**	**0.67***	**0.61**	0.93	**1.00***	**0.99***	**0.95**	**0.91***	**0.87***
	Lexical	0.96	0.70	0.61	0.58	0.73	0.92	0.97	0.89	0.74	0.67
	Syntactic	**0.98***	0.71	0.62	0.58	0.79	0.92	**0.99***	0.94	0.86	0.82
	Cohesion	0.94	0.70	0.60	0.58	0.70	0.88	**0.99***	0.92	0.82	0.78
	Concept	0.97	0.70	0.61	0.57	0.75	**0.93**	**0.99***	0.92	0.79	0.73
	Disc. Coherence	0.94	0.66	0.60	0.58	**0.96***	**0.99***	0.98	0.93	0.81	0.68
	Personalization	0.95	0.63	0.56	0.53	0.68	0.88	**0.99***	0.89	0.77	0.67
LogisticRegression	All Features	**0.92**	**0.77***	**0.67***	**0.62***	**0.73**	**0.99**	**0.96**	**0.84**	**0.87**	**0.77**
	Lexical	0.61	0.66	0.60	0.56	0.58	0.59	0.68	0.59	0.72	0.66
	Syntactic	0.72	0.66	0.60	0.57	0.68	0.65	0.87	0.73	0.76	0.68
	Cohesion	0.69	0.63	0.57	0.56	0.61	0.66	0.81	0.60	0.78	0.68
	Concept	0.63	0.67	0.60	0.57	0.61	0.61	0.68	0.63	0.75	0.68
	Disc. Coherence	0.59	0.60	0.57	0.53	0.56	0.98	0.69	0.58	0.70	0.66
	Personalization	0.56	0.57	0.54	0.53	0.56	0.53	0.65	0.59	0.66	0.59
Stack	LR→RF	**0.98***	0.71	0.62	0.55	0.73	0.99	0.95	0.86	0.83	0.73
	RF→LR	**0.98***	**0.77**	**0.66**	**0.61**	**0.97**	**1.00***	**0.99***	**0.98***	**0.90**	**0.86**

Table 5: F1 Macro classification scores. RF = RandomForest, LR = LogisticRegression. **Bold**: best in category per dataset. *Best on dataset. Document-level datasets underlined.

dimensions depending on the dataset. Because dimension regressors are trained on separate feature sets, differential correlations across dimensions reflect genuine measure sensitivity rather than shared prediction targets. Unsupervised composites confirm these patterns (Section E).

5 Conclusions

We showed that psycholinguistically-grounded simplicity dimensions predict simplicity ratings, outperforming established measures on most datasets (**RQ1**). We also showed that simplification datasets vary substantially in their transformation profiles: lexical and syntactic dimensions dominate, while personalization (despite being a strong predictor of perceived simplicity) is consistently under-targeted (**RQ2**). Finally, we showed that existing measures capture little of this dimensional structure, correlating weakly and indiscriminately across all dimensions (**RQ3**). This work bridges psycholinguistic research on text comprehension with ATS evaluation. Our results demonstrate that grounding simplicity analysis in cognitive theory is not just theoretically motivated but empirically productive: it yields stronger correlations with ratings, reveals blind spots invisible to current methods, and provides explanatory power that single-score measures lack. We believe this establishes cognitive grounding as a valuable new direction for ATS evaluation, and offer this framework as a foundation for dimension-aware measures and more balanced simplification datasets.

Limitations

Our claims are subject to several methodological constraints that we outline here.

Circularity in Observed Dimensional Importance. While we argue (based on relevant psychology research) that dimensions beyond lexical and syntactic simplicity matter for comprehension, our empirical evidence for their importance is necessarily limited by the datasets available. The ATS field has historically focused on lexical and syntactic transformations, meaning existing datasets largely reflect this emphasis. As a result, dimensions such as personalization, cohesion, and coherence exhibit limited variance in these datasets, which may attenuate both their observed effect sizes (RQ2) and their predictive power for human judgments (RQ1). The lower influence we observe for these dimensions may therefore reflect a data artifact from dataset design and implementation, rather than genuine unimportance. This circularity can only be perfectly resolved with datasets explicitly designed to vary along all dimensions.

Evaluation of Under-Represented Dimensions. Relatedly, we lack human-annotated datasets where raters were specifically trained to evaluate constructs such as cohesion or coherence in ATS. It is unclear whether human annotators can reliably assess these dimensions on a numerical scale, as opposed to holistic simplicity (annotators may not have a clear intuition of what an "8" or a "2" in cohesion means). Rather than relying on direct numerical assessment, one potential solution would be to evaluate along a distribution of texts known to span a wide difficulty range: from texts aimed at young children to scientific papers, through known popular outlets. Building such benchmarks remains an open challenge.

Feature and Dimensional Overlap. Several features load on multiple dimensions (marked with * in Table 6), meaning the dimensional separation is not perfectly clean. While we train per-dimension models on disjoint feature *subsets* after assignment, the shared features introduce some conceptual overlap, notably between *Lexical* and *Concept* (e.g., both use features based on entity complexity). This limits the interpretability of per-dimension comparisons, but remains a reasonable trade-off given that these dimensions are not fundamentally independent constructs, and that our evaluators are not perfect. For example, entity complexity will inherently be correlated with the lexical complexity of the terms used to describe it, despite being seemingly independent conceptually.

Language Specificity. Our operationalization relies on English-specific resources: MRC psycholinguistic norms, CEFR profiling, the Academic Word List, and the Harvard IV-4 dictionary. The framework's dimensional structure is mostly language-agnostic in principle, as it derives from universal cognitive models (CI and CLT), but the feature-level implementation does not transfer to other languages without substantial adaptation. Languages that differ significantly from English, such as logographic writing systems (e.g., Chinese) or languages with free word order (e.g., Turkish, Finnish), may require rethinking which features are appropriate for each dimension (especially syntax or cohesion).

Directionality of Changes. In RQ2, we measure the magnitude of change per dimension using $|\Delta z|$ but do not assess whether changes move *toward* simplicity. We assume that transformations in aligned datasets are predominantly simplifying, which is reasonable given their construction, but individual features may shift in the direction of greater complexity (e.g., increased nominalization despite overall simplification). Moreover, directionality is not always straightforward at the feature level: for instance, a decrease in `efllex_ratio_B1` may indicate simplification if compensated by an increase in `efllex_ratio_A1` (shifting vocabulary toward a lower CEFR level), but not if `efllex_ratio_C1` increases instead. A more fine-grained directional analysis would therefore require feature-level polarity annotations that account for such interactions, which we leave to future work.

Acknowledgments

The authors acknowledge the use of *ChatGPT*, *Claude*, and *Grammarly* during the preparation of this work for the following purposes: assistance with code development, grammar and spelling correction, and rewording of text. All output generated with the aid of these tools was reviewed and revised by the authors, who bear full responsibility for the content of this publication.

References

Sweta Agrawal and Marine Carpuat. 2024. Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension. *Preprint*, arXiv:2312.10126.

Noof Abdullah Alfear, Dimitar Kazakov, and Hend Al-Khalifa. 2024. Meta-Evaluation of Sentence Simplification Metrics. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11229–11235, Torino, Italia. ELRA and ICCL.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

Jan Bakker and Jaap Kamps. 2024. Cochrane-auto: An Aligned Dataset for the Simplification of Biomedical Abstracts. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 41–51, Miami, Florida, USA. Association for Computational Linguistics.

Amy Berglund-Barraza, Fenghua Tian, Chandramalika Basak, and Julia L. Evans. 2019. Word Frequency Is Associated With Cognitive Effort During Verbal Working Memory: A Functional Near Infrared Spectroscopy (fNIRS) Study. *Frontiers in Human Neuroscience*, 13:433.

Averil Coxhead. 2000. A New Academic Word List. *TESOL Quarterly*, 34(2):213.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4):1227–1237.

Dennis Davari, Liana Ermakova, and Ralf Krestel. 2024a. Comparative Analysis of Evaluation Measures for Scientific Text Simplification. In *Linking Theory and Practice of Digital Libraries*, pages 76–91, Cham. Springer Nature Switzerland.

Dennis Davari, Liana Ermakova, and Ralf Krestel. 2024b. 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, Ljubljana, Slovenia, September 24-27, 2024, Proceedings, Part I*, volume 15177 of *Lecture Notes in Computer Science*, pages 76–91. Springer.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level Simplification of Medical Texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

Luise Dürlich and Thomas François. 2018. EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Liana Ermakova, Hosein Azarbonyad, Jan Bakker, Benjamin Vendeville, and Jaap Kamps. 2026. Overview of the CLEF 2025 SimpleText Track: Simplify Scientific Text (and Nothing More). In Jorge Carrillo-de-Albornoz, Alba García Seco De Herrera, Julio Gonzalo, Laura Plaza, Josiane Mothe, Florina Piroi, Paolo Rosso, Damiano Spina, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 16089, pages 436–463. Springer Nature Switzerland, Cham.

Liana Ermakova, Eric SanJuan, Stéphane Huet, Hosein Azarbonyad, Giorgio Maria Di Nunzio, Federica Vezzani, Jennifer D'Souza, and Jaap Kamps. 2024. Overview of the CLEF 2024 SimpleText Track. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 283–307, Cham. Springer Nature Switzerland.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Arthur C. Graesser, Danielle S. McNamara, and Jonna M. Kulikowich. 2011. Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, 40(5):223–234.

D. Jake Follmer and Rayne A. Sperling. 2018. Interactions between reader and text: Contributions of cognitive processes, strategy use, and text cohesion to comprehension of expository science text. *Learning and Individual Differences*, 67:177–187.

J. Kincaid, Robert Fishburne, Richard Rogers, and Brad Chissom. 1975. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. *Institute for Simulation and Training*.

Walter Kintsch. 1998. *Comprehension: A Paradigm for Cognition*. Comprehension: A Paradigm for Cognition. Cambridge University Press, New York, NY, US.

Christin Kreutz, Fabian Haak, Björn Engelmann, and Philipp Schaer. 2024. BATS: BenchmArking Text Simplicity. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11968–11989, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A Learnable Evaluation Metric for Text Simplification. *Preprint*, arXiv:2212.09739.

Danielle S. McNamara and Walter Kintsch. 1996. Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, 22(3):247–288.

Tenaha O'reilly and Danielle S. Mcnamara. 2007. Reversing the Reverse Cohesion Effect: Good Texts Can Be Better for Strategic, High-Knowledge Readers. *Discourse Processes*, 43(2):121–152.

Yasuhiro Ozuru, Kyle Dempsey, and Danielle S. McNamara. 2009. Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, 19(3):228–242.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Günter Daniel Rey and Florian Buchwald. 2011. The expertise reversal effect: Cognitive load and motivational explanations. *Journal of Experimental Psychology. Applied*, 17(1):33–48.

Horacio Saggion. 2017. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Kathleen M. Sheehan, Irene Kostin, Diane Napolitano, and Michael Flor. 2014. The TextEvaluator Tool: Helping Teachers and Test Developers Select Texts for Use in Instruction and Assessment. *The Elementary School Journal*, 115(2):184–209.

Zhecheng Sheng, Tianhao Zhang, Chen Jiang, and Dongyeop Kang. 2025. BBScore: A Brownian Bridge Based Metric for Assessing Text Coherence. *Preprint*, arXiv:2312.16893.

Sanja Stajner, Sergiu Nisioi, and Ioana Hulpuș. 2020. CoCo: A Tool for Automatically Assessing Conceptual Complexity of Texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7179–7186, Marseille, France. European Language Resources Association.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer (Harvard IV-4 Dictionary)*. MIT Press, Cambridge, MA. Harvard IV-4 word lists accessed via the General Inquirer online repository.

Anselm R. Strohmaier, Timo Ehmke, Hendrik Härtig, and Dominik Leiss. 2023. On the role of linguistic features for comprehension and learning from STEM texts. A meta-analysis. *Educational Research Review*, 39:100533.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-Level Text Simplification: Dataset, Criteria and Baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Sweller. 1988. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, 12(2):257–285.

Leonard Tetzlaff, Bianca Simonsmeier, Tabea Peters, and Garvin Brod. 2025. A cornerstone of adaptivity – A meta-analysis of the expertise reversal effect. *Learning and Instruction*, 98:102142.

Sowmya Vajjala and Ivana Lucic. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Benjamin Vendeville, Liana Ermakova, and Pierre De Loor. 2025. Resource for Error Analysis in Text Simplification: New Taxonomy and Test Collection. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pages 3723–3732, New York, NY, USA. Association for Computing Machinery.

Michael Wilson. 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Victor H. Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *Preprint*, arXiv:1904.09675.

Dawei Zhu, Wenhao Wu, Yifan Song, Fangwei Zhu, Ziqiang Cao, and Sujian Li. 2024. CoUDA: Coherence Evaluation via Unified Data Augmentation. *Preprint*, arXiv:2404.00681.

A Feature List per Dimension

B Length Residualization

B.1 Motivation

Many linguistic features are inherently correlated with text length (e.g., longer texts naturally have more clauses, more discourse markers, higher type-token ratios at certain scales). If left uncorrected,

Table 6: Features retained after preprocessing, grouped by dimension. Features marked with * load on multiple dimensions.

Dimension	Features	#
Lexical Simplicity	avg_age_of_acquisition, dale_chall, smog, total_syllables efllex_ratio_{a1, a2, b1, b2, c1} proportion_long_words_{letters_9, syllables_3, syllables_4} ratio_infrequent_words_bats taaco_{adj, adv, bigram_lemma, lemma, noun, prp, verb}_ttr te_awl_ratio*, te_ratio_nominalizations_token* unieval_fluency*	23
Syntactic Simplicity	avg_words_before_main_verb, per_word_noun_phrases, token_count ratio_{appositions, conditionals, modifiers, negations} ratio_passive_voice_ext, ratio_past_tense_ext* ratio_past_perfect_{bats, ext} ratio_relative_clauses_{bats, ext} ratio_subordinate_clauses te_avg_{dependent_clauses, yngve_depth}, te_freq_long_sentences_25w unieval_fluency*	18
Cohesion	avg_distance_same_entities*, coco_avgTokenDistance* taaco_addition, taaco_opposition, taaco_order, taaco_sentence_linking taaco_adjacent_overlap_pronoun_sent taaco_all_{causal, demonstratives, logical, negative, positive, temporal} taaco_{coordinating_conjuncts, determiners, lexical_subordinators} taaco_lsa_1_all_sent taaco_positive_{causal, intentional, logical} te_ratio_causal_conjuncts, unique_entity_total_entity_ratio_text	22
Discourse Coherence	avg_distance_same_entities*, bbscore, couda_global, paragraph_count coco_{activationDecayRate, activationMax, activationMin, activationStability} coco_avgNodeDegree, coco_avgTokenDistance*, coco_complexityScore*	11
Concept Simplicity	coco_complexityScore*, coco_entityDensity, entities_per_sentence ratio_{cog_orient, institution}_gi te_awl_ratio*, te_ratio_nominalizations_token*	7
Personalization & Engagement	ratio_{active, communicating, motivation, negativ, our}_gi ratio_{ovrst, passive, positiv, self, undrst, you}_gi te_ratio_pronouns_3rd_singular*	12
Total unique features		**93**

a model trained on these features may learn to predict simplicity primarily through length rather than through genuine linguistic simplification properties. We use length residualization so we can study the effect of the variance not attributable to text length, which can easily be studied separately.

B.2 Procedure

The procedure is as follows:

1. **Identifying length-correlated features**: For each feature f in the feature matrix, the Pearson correlation r between f and the length proxy (word count) is computed. A feature is marked for residualization if $|r| > 0.1$ (a deliberately low threshold to be conservative so even weak length confounds are removed).

2. **Fitting an OLS regression**: For every feature f flagged in Step 1, an Ordinary Least-Squares model is fitted:

$$f = \beta_0 + \beta_1 \cdot \text{word_count} + \varepsilon$$

where β_0 is the intercept, β_1 is the regression coefficient capturing the linear relationship between the feature and word count, and ε represents the residuals. We use OLS via the 'statsmodels.OLS' library, with an intercept term.

3. **Replacing the feature with the residuals**: The original feature values are replaced by the OLS residuals ε. These residuals represent the portion of the feature's variance that is *orthogonal to text length* i.e., the feature value after the linear effect of length has been subtracted.

B.3 Grouping Mode

Residualization is performed per-dataset: for paired datasets (e.g., source and generated texts

from the same corpus), both roles are pooled together before fitting the OLS model. This is important because fitting separate regressions per role (source vs. generated) would introduce artificially different intercepts/slopes between the two groups, creating spurious separability in downstream discrimination analyses. By using a single shared regression across both roles of a dataset, genuine linguistic differences between source and simplified texts are preserved.

B.4 Effect

After residualization, the mean absolute correlation between residualized features and word count drops to approximately zero, confirming that the linear length confound has been effectively removed. Features that were not correlated with length ($|r| \leq 0.1$) are left unchanged.

C Williams Significance Tests

Figure 3 reports pairwise Williams significance tests comparing the Pearson correlations of all models with human judgments, for each annotated dataset. A green cell indicates that the row model correlates significantly more than the column model ($p < 0.05$). Models are sorted by descending $|r|$.

Several patterns emerge. First, on SDA and D-Wiki, the top-performing models (All Features RF, Lexical RF, Syntactic RF, and stacked variants) form a cluster with no significant pairwise differences, indicating that Lexical and Syntactic features alone capture most of the predictive signal on those datasets. On STsent, by contrast, stacked models significantly outperform individual dimensions, suggesting that multidimensional combination is most beneficial when no single dimension dominates human judgments. Second, RF significantly outperforms LR on the same feature sets across most datasets, indicating that nonlinear feature interactions contribute meaningfully to prediction — a relevant consideration for future measure design.

Third, Personalization / Style ranks significantly below most other dimensions on SDA, ST_{sent}, and D-Wiki. Combined with the small effect sizes reported in RQ2 (Table 4), this suggests that current datasets not only under-target personalization but that the resulting feature variance is too limited to reliably predict human ratings. Finally, the relative ordering of dimensions shifts across datasets: Cohesion and Discourse Coherence rank significantly below Lexical Simplicity on SDA but are not significantly different from it on ST_{sent}. This confirms that simplicity is not a fixed hierarchy of dimensions but a dataset-dependent structure, as discussed in Section 4.1.

D Out-Of-Domain Generalization

	Out-Of-Domain			In Domain	
Dataset	AF	Stack	Hybrid	AF	Stack
SDA	0.61	0.57	0.64	0.74	0.74
ST-sent	0.12	0.11	0.57	0.74	0.75
ST-para	0.29	0.30	0.34	0.35	0.40
D-Wiki	0.43	0.45	0.50	0.71	0.69
Mean	0.34	0.36	0.51	0.64	0.64

Table 7: Transfer strategy results by dataset. Best single donor as foreign source. AF being "All Features"

We want to evaluate whether dimension-level representations generalize across datasets. The question here is whether the per-dimension regressors learn transferable properties of simplicity, or just overfit to dataset-specific annotation patterns. To study this, we compare five transfer strategies with increasing access to target-domain data. Foreign All-Features and Full Foreign transfer an entire model trained on a foreign dataset, either as a single regressor or as a full stacking pipeline. These establish a lower bound on transferability. Hybrid is the critical condition: it retains the foreign dimension-level base regressors but retrains only the meta-learner on the target dataset via cross-validation. If dimensions capture generalizable signal, then recalibrating their combination locally should recover most of the in-domain performance, even when the base regressors have never seen the target data. Results are shown in Figure 4 and Table 7. This is what we observe: Hybrid recovers 80% of in-domain performance (0.51 vs 0.64), while full foreign transfer remains weak (0.36). The gap confirms that dimension-level representations transfer across datasets but their relative importance does not, consistent with the dataset-dependent patterns observed in RQ1.

E Composite Scores Correlation with Metrics

In the main analysis (Section 4.3), dimension scores are obtained from per-dimension regressors trained on human judgments. While these regres-

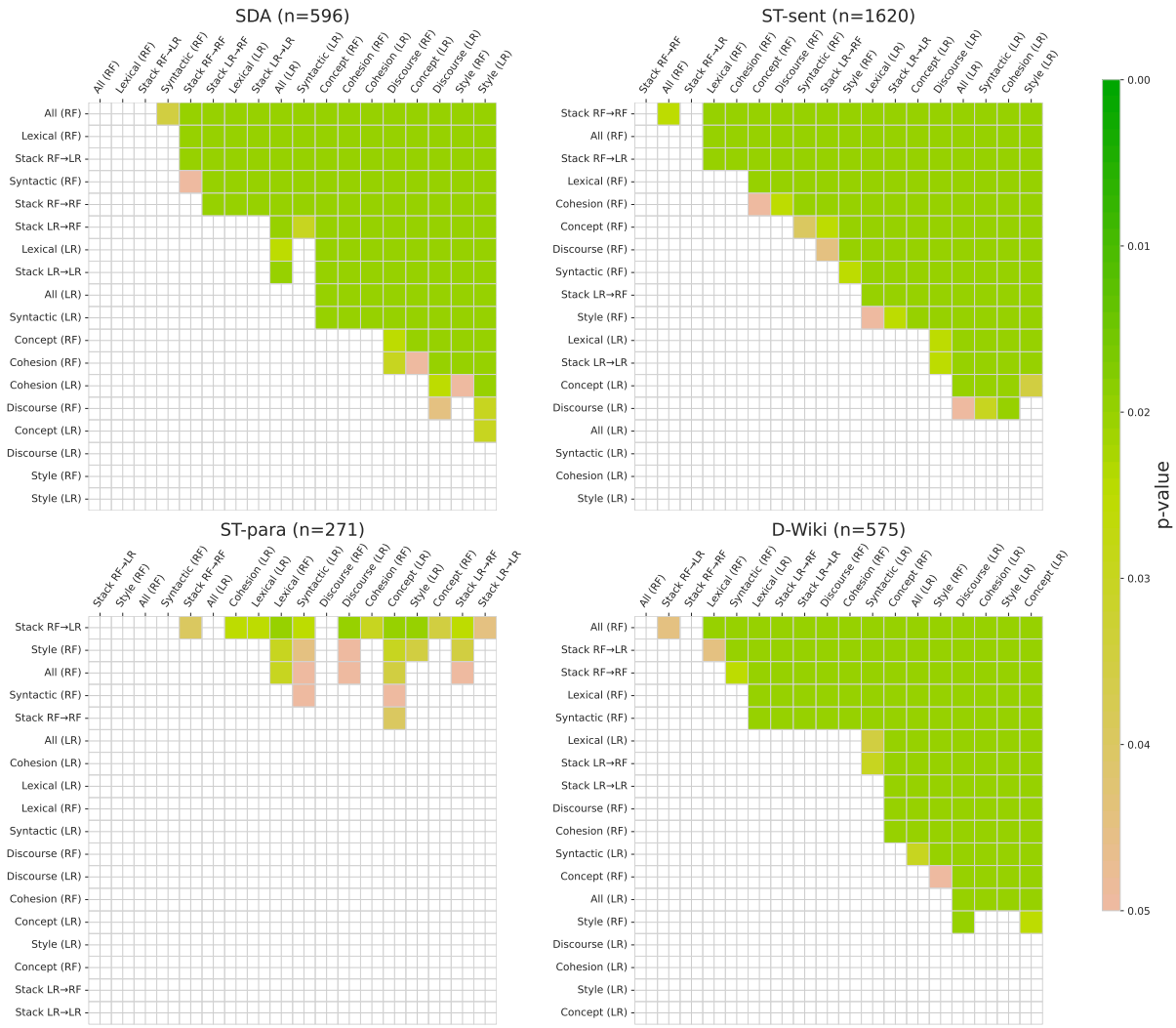

Figure 3: Pairwise Williams significance tests. A green cell indicates the row model correlates significantly more than the column model ($p < 0.05$) with human judgments.

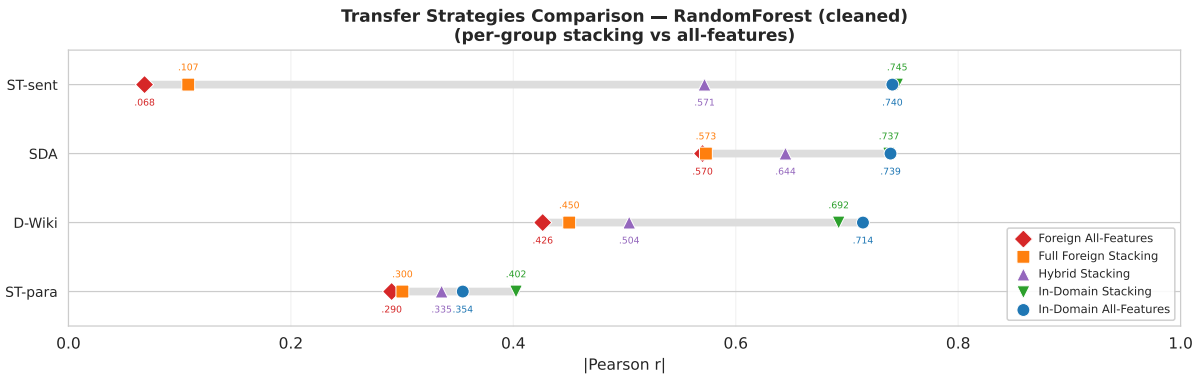

Figure 4: Cross-dataset transfer performance (|Pearson r|) on different test dataset. Multiple setups are used including completely foreign training (Foreign All-Features, full foreign stacking), full in-domain (In-Domain Stacking, In-Domain All-Features) and Hybrid Stacking, where dimension classifiers trained Out-Of-Domain, are used to train an in-Domain Stacked regressor

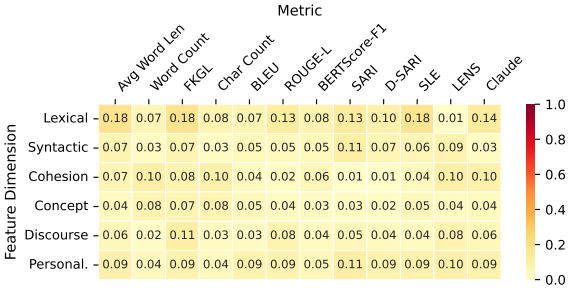

Figure 5: Mean absolute Pearson correlation (|r|) between measures and unsupervised dimension scores (first PCA component per dimension), averaged across 4 datasets (D-Wiki, SDA, ST-para, ST-sent).

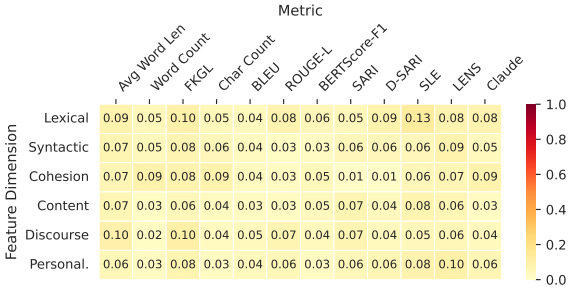

Figure 6: Mean absolute Pearson correlation (|r|) between measures and unsupervised dimension scores (mean of z-scored features per dimension), averaged across 4 datasets (D-Wiki, SDA, ST-para, ST-sent).

sors are trained on disjoint feature subsets, meaning differential correlations across dimensions may reflect genuine measure sensitivity rather than just shared prediction targets. We verify this by replicating the analysis using two unsupervised alternatives that involve no training on human ratings.

The first method aggregates features within each dimension by averaging their z-scores across all features belonging to that dimension, for each text independently. This produces a single unsupervised score per dimension per text, where all features contribute equally. (Figure 6).

The second extracts the first principal component per dimension, capturing the dominant axis of variance within each feature group (Figure 5).

Absolute correlations are lower in both cases ($max(|r|) \approx 0.13$ for z-composite, ≈ 0.18 for PCA, compared to ≈ 0.38 for regressor-based scores), which is expected: unsupervised aggregation weights all features equally or by variance rather than by relevance to perceived simplicity, diluting signal from the most informative features. Despite this, the key patterns from the main analysis hold. First, Lexical Simplicity remains the dimension most strongly associated with the majority of measures, including FKGL, SLE, and Claude-based evaluators. Second, no measure shows strong selective sensitivity to any single dimension — profiles remain relatively flat across dimensions. Third, Discourse Coherence, Concept Simplicity, and Personalization remain poorly captured by most evaluation measures. The convergence of supervised and unsupervised approaches on these patterns strengthens our conclusion that existing measures are weakly and indiscriminately coupled to the dimensional structure of simplicity.

F LLM Evaluation

We use Claude Haiku (claude-haiku-4-5-20251001) as our LLM-as-a-judge. Table 8 summarizes the configuration.

Parameter	Value
Model	claude-haiku-4-5-20251001
Temperature	0
max_tokens	10
Rate limiting	20 requests/second
Max retries	3 (exponential backoff)

Table 8: Details of the claude LLM-as-a-judge evaluation setup

Listing 1: Claude Prompt

```
Given the original text, its simplification,
    ↪ and human reference simplifications,
    ↪ rate the simplification on a scale of
    ↪ 1-10:
- 1 = Poor (doesn't improve readability, loses
    ↪ meaning, or much worse than references)
- 10 = Excellent (greatly improves readability,
    ↪ preserves meaning, comparable to
    ↪ references)
Original: "{orig}"
Simplified: "{simp}"
Human references: "{refs}"
Respond with ONLY a single integer from 1 to 10.
```

G Datasets Samples

SDA - Score: 37.87/100, higher=simpler

Source: Gable also earned an Academy Award nomination when he portrayed Fletcher Christian in 1935's Mutiny on the Bounty.
Simplified: Gable earned an Academy Award nomination he portrayed Fletcher Christian.

ASSET

Source: Prior to the arrival of the storm, the National Park service closed visitor centers and campgrounds along the outer banks.
Simplified: Before the storm came to the National Park service, the National Park service closed places and campgrounds along the outer banks.

ST-sent - Score: 1/14
syntax=1/7, lexical=1/7
lower=simpler

Source: Inappropriate use of medicines increases the risk of hospital admissions for the elderly.
Simplified: misusing medicines harms the elderly.

ST-para - Score: 2.165/14
syntax=1/7, lexical=2.33/7
higher=simpler

Source: As various kinds of output devices emerged, such as highresolution printers or a display of PDA(Personal Digital Assistant), the importance of high-quality resolution conversion has been increasing. This paper proposes a new method for enlarging image with high quality. One of the largest problems on image enlargement is the exaggeration of the jaggy edges. To remedy this problem, we propose a new interpolation method, which uses artificial neural network to determine the optimal values of interpolated pixels. The experimental results are shown and evaluated. The effectiveness of our methods is discussed by comparing with the conventional methods.
Simplified: in this paper, we propose a new method for resolution conversion paper proposes a new method for enlarging image with high quality. image enlargement's exaggeration of jaggy edges is one of the largest problems. to fix this issue, we suggest a new method that uses a computer brain to find the best values for filling in pixels. experimental results are shown and evaluated. efecundity of our methods is discussed by comparing with the conventional methods.

D-Wiki - Score: 4/5, higher=simpler

Source: on 22 january 2019 , floods caused by heavy rainfall struck the indonesian province of south sulawesi . at least 68 people were killed and thousands were displaced . gowa regency was particularly hard-hit , experiencing the majority of the fatalities.the most destructive flood was caused by the overflowing of the jeneberang river and the opening of the watergates of the bili-bili dam in gowa , though other floods also occurred elsewhere in the province .
Simplified: on 22 january 2019 , floods caused by heavy rainfall hit the indonesian province of south sulawesi . at least 68 people were killed and thousands were lost . gowa regency was hard-hit having the majority of the fatalities.the most destructive flood was caused by the overflowing of the jeneberang river and the opening of the watergates of the bili-bili dam in gowa , though other floods also occurred elsewhere in the province .