



Castledown

 OPEN ACCESS

Technology in Language Teaching & Learning

ISSN 2652-1687

<https://www.castledown.com/journals/tlt/>

Technology in Language Teaching & Learning, 8, 103327 (2026)
<https://doi.org/10.29140/tltl.2026.103327>

Linguistic Control in AI Text Generation: An Accessible Prompt-Based Approach Targeting L2 Spanish Absolute Beginners



RAÚL GETINO-DIEZ^a 

MIKEL GARCÍA-MADARIAGA^b 

^a*Xi'an Jiaotong-Liverpool University, China*
Raul.GetinoDiez@xjtlu.edu.cn

^b*Xi'an Jiaotong-Liverpool University, China*
Mikel.Garcia@xjtlu.edu.cn

Abstract

Generative artificial intelligence (AI) offers strong potential for developing customized second language learning materials and tools. However, generating texts for absolute beginners which require strict lexical and grammatical control remains a challenge. Although controlled text generation (CTG) techniques exist, they often require technical expertise and infrastructure, limiting accessibility for educators. This study evaluates, in the context of Spanish, a prompt-based approach that leverages large language models (LLMs) without fine-tuning or specialized tools. Prompts enforce linguistic constraints defined in two attachments: a categorized Spanish vocabulary list, and a set of example sentences illustrating approved Spanish grammatical structures organized by communicative function. Three variables were manipulated: *AI model* (ChatGPT-4o vs. Claude 3.5 Sonnet), *prompt type* (standard vs. extended, with constraint-enhancing techniques), and *attachment format* (rich-heavyweight vs. lightweight JSON). A secondary variable, *text type* (city descriptions, personal introductions, and dialogues), was also examined. A total of 720 texts were generated, 30 per condition. Measures included proportions of non-compliant lexical and grammatical items, user-perceived latency, and errors in vocabulary, grammar, and coherence. Model choice was the primary driver of constraint adherence, with Claude 3.5 Sonnet outperforming ChatGPT-4o. Extended prompts improved adherence across models. Attachment format showed no systematic effect on adherence, but JSON significantly reduced latency and response-time variability. Text type also influenced adherence, and error rates remained low. Findings offer educators a scalable, low-barrier solution for generating tailored beginner-level Spanish materials and AI-powered tools using LLMs, along with insights into how different design choices affect performance. This approach, transferable to other languages, provides a practical alternative to resource-intensive CTG techniques, addressing a critical gap in AI-assisted language education.

Copyright: © 2026 Raúl Getino-Diez & Mikel García-Madariaga. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. **Data Availability Statement:** All relevant data are within this paper.

Keywords: Absolute beginners; artificial intelligence in language education; Controllable Text Generation (CTG); technology enhanced language; second language teaching, prompt engineering; Spanish

Introduction

Generative artificial intelligence (AI) presents significant opportunities for second language (L2) educators to develop customized learning materials aligned with learners' needs, proficiency levels, and curricular goals. Recent studies highlight the potential of AI-generated texts in terms of suitability (Young & Shishido, 2023), naturalness (Shin et al., 2025), and learner appeal (Kim & Park, 2023). However, linguistic control remains crucial, since learners benefit most from $i+1$, input just beyond their current level (Krashen, 1985). This control is especially important for absolute beginners, whose vocabulary and grammar knowledge are extremely limited, particularly in contexts with minimal exposure outside the classroom. AI-generated texts for absolute beginners often exceed instructional pacing, frustrating learners and increasing the burden on teachers who must manually adapt texts for classroom use.

Current Approaches and Limitations

Research on controllable text generation (CTG) for L2 contexts has explored several dimensions: proficiency or difficulty control via Common European Framework of Reference (CEFR) levels (Imperial & Madabushi, 2023; Malik et al., 2024; Uchida, 2025), grammar control, often mapped to proficiency levels (Glandorf & Meurers, 2024; Glandorf et al., 2025; Stowe et al., 2022), lexical complexity or predefined vocabulary lists (Jin et al., 2025; Li et al., 2024; Nie et al., 2023; Iso, 2024), conversation-level comprehensibility for absolute beginners (Jin et al., 2025), and curriculum alignment (Li et al., 2024).

However, achieving fine-grained control remains challenging. Simple prompt engineering alone rarely achieves high degrees of compliance (Glandorf & Meurers, 2024; Glandorf et al., 2025; Imperial & Madabushi, 2023; Jin et al., 2025; Malik et al., 2024; Uchida, 2025). To improve controllability, researchers have adopted more advanced but resource-intensive methods, including fine-tuning (Glandorf et al., 2025; Li et al., 2024; Malik et al., 2024; Nie et al., 2023; Stowe et al., 2022), reinforcement learning (Malik et al., 2024), and post-processing control (Glandorf & Meurers, 2024; Glandorf et al., 2025; Jin et al., 2025; Nie et al., 2023). Hybrid approaches, such as AutoTemplate (Iso, 2024), combine supervised fine-tuning with deterministic post-generation lexicalization. To our knowledge, no study has attempted to retrain base LLM architectures for L2 control, likely due to the extremely resource-intensive nature of such approaches (Zhang et al., 2023).

Although none of these studies target the strict lexical and grammatical control needed for absolute beginners (e.g., exclusion of any linguistic item not included in an approved list), existing CTG techniques could be adapted for this purpose. However, they remain inaccessible to most educators and institutions due to their technical complexity, resource demands, and scalability limitations.

Study Aims

To address this gap, this study proposes and evaluates a prompt-based, teacher-deployable approach combining enhanced prompting techniques with structured attachments to enforce Spanish linguistic constraints, without requiring fine-tuning or specialized infrastructure.

This approach offers several practical advantages: it requires minimal computational resources, can be implemented immediately without highly specialized training, supports iterative improvement based on classroom feedback, and is inherently scalable; prompts and knowledge bases can be adjusted as courses progress, without model retraining. Additionally, prompt optimization and attachment design can inform the development of custom LLM-powered tools (e.g., chatbots) via low-code/no-code platforms, broadening accessibility and adaptability in instructional contexts (see Wiboolyasarin et al., 2025, for a systematic review of AI-driven chatbots in second language education).

Specifically, this study investigates how different design factors—*AI model*, *prompt type*, and *attachment format*—influence linguistically constrained Spanish text generation for absolute beginners. *Text type* is included as a secondary variable, as communicative goals, genres, and topics may influence constraint adherence.

Prior work suggests that model architecture (Glandorf & Meurers, 2024; Glandorf et al., 2025; Imperial & Madabushi, 2023; Li et al., 2024; Malik et al., 2024) and prompt design (Imperial & Madabushi, 2023; Malik et al., 2024) affect constraint adherence, while lightweight files like JSON show faster processing than heavyweight alternatives (Zunke & D’Souza, 2014). Research also indicates that including examples in prompts improves constraint compliance (Malik et al., 2024; Uchida, 2025); accordingly, attachments will provide the approved grammatical structures through illustrative examples.

Rather than identifying a single optimal method, this study examines how different design choices influence constraint adherence, responsiveness, and linguistic accuracy, enabling educators to make informed decisions based on available resources. Although focused on Spanish, these insights can benefit other languages where lexical and grammatical scaffolding are similarly important.

Methods

This study adopted an experimental, quantitative design to investigate the effects of key intervention variables on the generation of highly controlled AI texts intended for absolute beginners.

Areas of Intervention: Independent Variables

AI Model

To examine potential model-driven differences in adherence to linguistic constraints, two widely recognized AI models were selected: ChatGPT-4o (2024-11-20) and Claude 3.5 Sonnet (2024-10-22), both known for their advanced natural language generation capabilities.

Prompt Type

Two prompt types were designed to investigate prompt engineering effects on model constraint adherence: standard and extended prompts (see Appendix). Both followed a two-part structure: (a) a *generic section* defining the overall task and linguistic constraints, and (b) a *specific section* providing instructions adapted to each text type (city description, personal introduction, or dialogue), with sufficient contextual detail to support authenticity and relevance for absolute beginners. Both prompt types shared the specific section, differing only in the generic section design.

The standard prompt featured a clear structure with explicit instructions, well-specified lexical and grammatical constraints, and direct references to the attached knowledge files.

The extended prompt retained all elements of the standard version but incorporated additional constraint-enhancing strategies:

- (a) *Persona Pattern*: Assigned the model a specific role (e.g., “*You are a skilled language model tasked with generating Spanish texts for beginners based on limited constraints*”).
- (b) *Context Priming*: Included repeated references to “*for Spanish beginners*” to lower the output’s expected proficiency level and steer the model toward beginner-appropriate language. This technique leverages the premise that general LLMs can produce beginner-level output when appropriately primed, even if they do not strictly follow specific linguistic constraints.
- (c) *Formatting*: Used markdown-style formatting (e.g., **bold**) to highlight headings and key instructions, reinforcing structural hierarchy and increasing critical information’s salience.
- (d) *Rule Reinforcement via Redundancy*: Paired affirmative instructions (e.g., “*Use only the words provided in the vocabulary list*”) with corresponding negations for each constraint (e.g., “*No additional words are allowed*”).
- (e) *Negation of Common Structures Reinforced via Redundancy*: Explicitly discouraged specific non-compliant forms when they were strongly associated with basic communicative functions, along with their corresponding affirmative instruction (e.g., “*Verbs can only be conjugated in the present tense (indicative mood). Avoid other tenses and moods such as past tense or subjunctive*”).
- (f) *Final Reinforcement Section*: Restated all critical constraints at the end of the prompt, also reiterating rule reinforcement via redundancy.

This structured design ensured consistency across experimental conditions—*prompt type* and *text type*—while allowing for customization, enabling controlled comparison between two high-quality prompts differing only in targeted enhancement strategies.

Attachment Format

Two knowledge files were created to serve as prompt attachments, supporting customization, scalability, and potential integration into custom bots or instructional AI agents: (a) a *Vocabulary List* in XLSX format, containing a categorized inventory of lexical items from an entry-level Spanish course through Week 11 of instruction; and (b) a *Functions and Examples* document in DOCX format, comprising example sentences organized by communicative functions and covering all grammatical structures planned for the same instructional period.

To examine attachment format effects on model constraint alignment, both files were converted to JSON, a highly structured, simple, and lightweight format. Files such as DOCX and PDF, in contrast, are considered unstructured, rich, and heavyweight; XLSX, though structured, is also regarded as rich and heavyweight due to its complexity and metadata load.

Attachment format was therefore manipulated across two conditions: (a) the original file formats—XLSX and DOCX for ChatGPT-4o, and PDF for Claude 3.5 Sonnet (as XLSX was not supported by that model at the time of data generation); and (b) JSON, used with both models.

Text Type

Three text types were generated in Spanish: city descriptions, personal introductions, and dialogues, each characterized by distinct communicative goals, genres, and topics. Word count requirements were set at 150 words for descriptions and 250 words for introductions, while dialogues were not subject to a fixed word-count requirement.

Although treated as an independent variable, *text type* was considered a secondary factor. These categories reflect common instructional formats for absolute beginners but do not represent the full range of beginner-level text types. Accordingly, *text type* was analyzed to explore potential differences in model adherence to linguistic constraints based on variations in communicative goals, genres, and topics.

Measuring AI Models' Performance: Dependent Variables

Performance in adhering to linguistic constraints was evaluated using two primary variables: (a) *OOD-W*, the proportion of words not included in the *Vocabulary List* (i.e., out-of-domain words); and (b) *OOD-G*, the proportion of grammatical structures not represented in the *Functions and Examples* document (i.e., out-of-domain structures). A "structure" was operationalized as a discrete construction identified by its core function and form (e.g., noun determination, verb tense or mood, and object-pronoun constructions). Both variables were calculated by dividing the number of out-of-domain items (words or grammatical structures, respectively) by the total word count of each text.

Additional variables included: (c) *user-perceived latency*, recorded for dialogues; (d) *error metrics*, including grammar, vocabulary, and coherence errors; and (e) *word count*, to observe alignment with prompt-specific length requirements.

Out-of-domain words were calculated using *AntWordProfiler* software (Anthony, 2024), while out-of-domain grammar structures and errors were recorded via manual annotation by a single rater. To ensure reliability, a second reviewer independently examined all annotations, and any discrepancies were discussed and resolved collaboratively.

User-perceived latency was approximated by manual timing in seconds and was operationalized as the Time to First Token (TTFT), defined as the interval between prompt submission (keypress) and the appearance of the first character of the model's response on screen. Because the focus is on user experience, this measure captures full end-to-end latency including queuing, prefill time, and network delays, rather than isolating model compute time alone. Although manual timing introduces sub-second imprecision due to human reaction time, this error is nonsystematic and consistent across trials, functioning as classical measurement noise. Means and standard deviations are reported, and results are interpreted with this limitation in mind.

Several exploratory variables were excluded as they did not yield additional insights. These included *OOD-W-Unique* and *OOD-G-Unique* (accounting for repeated items), *OOD-G-w/oFS* (excluding fixed expressions such as farewells), and *OOD-G-Unique-w/oFS* (excluding both repeated structures and fixed expressions).

Data Collection: Text Generation

Text generation was conducted using ChatGPT and Claude web-based interfaces. Generation parameters, including temperature and other decoding settings, are not exposed or configurable through these interfaces. Consequently, all outputs reflect the default configurations enforced by each platform during the data collection period (November-December 2024), which are not publicly disclosed and could not be controlled in this study.

Thirty texts were generated for each of the 24 experimental conditions (2 AI models \times 2 prompt types \times 2 attachment formats \times 3 text types), yielding 720 texts. Each text was produced in a separate, memory-disabled chat, and the attached knowledge files were uploaded and re-indexed for every

run. These procedures reduce the likelihood of contextual carryover and make residual dependencies between texts, including those potentially caused by short-lived server processes, highly unlikely.

Hypotheses

The study specifically tested the following hypotheses:

- (a) **H1:** Different AI models will exhibit different adherence to linguistic constraints (OOD–W, OOD–G) under the same prompt and attachment type conditions.
- (b) **H2:** Extended prompts will improve model adherence to linguistic constraints.
- (c) **H3:** JSON attachments will be processed more efficiently, leading to (**H3a**) improved model adherence and (**H3b**) reduced user-perceived latency.
- (d) **H4:** Different text types will affect linguistic constraint adherence, as certain genres or topics can facilitate or hinder control.

Statistical Methods

To select the appropriate statistical tests, the assumptions of normality and homoscedasticity were evaluated for all dependent variables using the Shapiro–Wilk test and Levene’s test, respectively.

The Shapiro–Wilk test indicated significant deviations from normality in most conditions ($p < .001$), with OOD–W and OOD–G scores often clustered near zero, reflecting cases of perfect or near-perfect constraint adherence. Despite these deviations, parametric methods remained justified due to large sample sizes ($n \geq 30$ per condition) and the robustness of t -tests and ANOVAs to moderate non-normality, as supported by the Central Limit Theorem.

Levene’s test revealed significant violations of the homogeneity of variance assumption in most group comparisons ($p < .001$). Accordingly, Welch’s t -tests and Welch’s ANOVAs were selected for unequal variances. When the assumption was met, standard t -tests were applied, and Levene’s test results were reported alongside for reference.

Outliers were identified but retained, as they represented legitimate variability in model behavior rather than data entry or measurement errors. Non-parametric tests including the Mann–Whitney U test and the Kruskal–Wallis test were employed for methodological triangulation, as they are more robust to assumption violations and outliers.

Given that the analyses tested pre-specified hypotheses, corrections for multiple comparisons were not applied. As detailed in the Results section, the consistent pattern of very low p -values ($p < .001$) and moderate-to-large to very large effect sizes across the primary outcomes supports the robustness and reliability of the findings.

Results

AI Model Comparison

Claude 3.5 Sonnet demonstrated significantly lower means of OOD–W and OOD–G compared to ChatGPT-4o based on a total of $N = 720$ texts, with 360 texts generated per model (see Table 1).

Claude 3.5 Sonnet produced texts with a mean OOD–W of .0069 ($SD = .0066$), meaning that, on average, fewer than one word per 100 (0.69%) fell outside the allowed vocabulary. In contrast, ChatGPT-4o

showed a mean OOD–W of .0736 ($SD = .0523$), approximately 7.4 words per 100; more than ten times higher. This difference was statistically significant, Welch’s $t(370.45) = 23.96$, $p < .001$, with a large effect size ($d = 1.79$, 95% CI [1.61, 1.96]). This result was confirmed by a Mann–Whitney U test ($z = -21.90$, $p < .001$).

Table 1 Effect of AI Model on Lexical and Grammatical Constraint Adherence

Measure	Model	n	M	SD	t	df	d	95% CI		z
								LL	UL	
OOD–W	ChatGPT-4o	360	.0736	.0523	23.96	370.45	1.79	1.61	1.96	-21.90
	Claude 3.5 Sonnet	360	.0069	.0066						
OOD–G	ChatGPT-4o	360	.0094	.0116	14.96	365.97	1.12	0.96	1.27	-16.44
	Claude 3.5 Sonnet	360	.0002	.0011						

Note. All p -values $< .001$ for both parametric (Welch’s t -tests) and non-parametric (Mann–Whitney U tests) comparisons. All p -values are two-tailed. M = mean; SD = standard deviation; t = Welch’s t -statistic; d = Cohen’s d ; CI = confidence interval for d ; LL = lower limit; UL = upper limit; z = standardized Mann–Whitney U statistic.

Claude also produced significantly lower OOD–G, with a mean of .0002 ($SD = .0011$), equivalent to approximately 2 out-of-domain structures per 10,000 words on average (0.02%). ChatGPT-4o’s mean was .0094 ($SD = .0116$), or approximately 1 per 100 words (0.94%); about 47 times higher. This difference was statistically significant, $t(365.97) = 14.96$, $p < .001$, with a large effect size ($d = 1.12$, 95% CI [0.96, 1.27]). A Mann–Whitney U test yielded consistent results ($z = -16.44$, $p < .001$).

These findings confirm that Claude 3.5 Sonnet produced not only more accurate but also more consistent outputs, with substantially smaller standard deviations across both lexical and grammatical measures. Its performance demonstrates superior constraint adherence, which is critical for generating pedagogically appropriate input for absolute beginners.

Prompt Type Comparison

Extended prompts led to significantly lower means of OOD–W and OOD–G compared to standard prompts across both AI models, with consistently significant effects in both the full dataset and model-specific analyses (see Table 2).

Across the full dataset (ChatGPT-4o + Claude 3.5 Sonnet), extended prompts reduced OOD–W from $M = .0562$ to .0243, more than halving the average proportion of out-of-domain words. This difference was statistically significant, $t(502.14) = 9.05$, $p < .001$, with a moderate-to-large effect size ($d = 0.67$). A similar pattern emerged for OOD–G, which decreased from $M = .0074$ to .0022, $t(517.43) = 7.60$, $p < .001$, also with a moderate effect size ($d = 0.57$).

When analyzed separately by model, both ChatGPT-4o and Claude 3.5 Sonnet showed improved performance with extended prompts. For ChatGPT-4o, mean OOD–W fell by more than half, from $M = .1032$ to .0439, $t(263.67) = 13.00$, $p < .001$, with a large effect size ($d = 1.37$). OOD–G decreased threefold, from $M = .0143$ to .0044, $t(280.95) = 8.94$, $p < .001$, also with a large effect ($d = 0.94$). In Claude 3.5 Sonnet, the reductions were smaller but still robust: OOD–W decreased from $M = .0093$ to .0046, $t(304.75) = 7.30$, $p < .001$, with a moderate-to-large effect size ($d = 0.77$), and OOD–G dropped from $M = .0004$ to .0000, $t(179.00) = 3.27$, $p < .001$, with a small-to-moderate effect size ($d = 0.34$).

Table 2 *Effect of Prompt Type on Lexical and Grammatical Constraint Adherence*

Measure	Prompt Type	<i>n</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>d</i>	95% CI		<i>z</i>
								LL	UL	
ChatGPT-4o + Claude 3.5 Sonnet										
OOD–W	Standard	360	.0562	.0610	9.05	502.14	0.67	0.52	0.82	-7.11
	Extended	360	.0243	.0278						
OOD–G	Standard	360	.0074	.0116	7.60	517.43	0.57	0.42	0.72	-6.47
	Extended	360	.0022	.0056						
ChatGPT-4o										
OOD–W	Standard	180	.1032	.0546	13.00	263.67	1.37	1.14	1.60	-11.04
	Extended	180	.0439	.0274						
OOD–G	Standard	180	.0143	.0130	8.94	280.95	0.94	0.72	1.16	-8.32
	Extended	180	.0044	.0073						
Claude 3.5 Sonnet										
OOD–W	Standard	180	.0093	.0074	7.30	304.75	0.77	0.55	0.98	-6.51
	Extended	180	.0046	.0047						
OOD–G	Standard	180	.0004	.0016	3.27	179.00	0.34	0.14	0.55	-3.52
	Extended	180	.0000	.0000						

Note. All *p*-values < .001 for both parametric (Welch's *t*-tests) and non-parametric (Mann–Whitney U tests) comparisons. All *p*-values are two-tailed.

These results, replicated by Mann–Whitney U tests, confirm that extended prompts reliably improve adherence to both lexical and grammatical constraints. Effects were especially pronounced in ChatGPT-4o, while Claude 3.5 Sonnet achieved perfect grammatical control when given extended prompts.

Attachment Format Comparison

This section examines whether *attachment format* (XLSX–DOCX/PDF vs. JSON) influenced adherence to lexical and grammatical constraints. Since the rich, heavyweight formats differed by model—XLSX and DOCX for ChatGPT-4o, and PDF for Claude 3.5 Sonnet—the variable *attachment format* was partially confounded with model. Therefore, comparisons were conducted within each model only.

While some statistically significant differences emerged, they were isolated; most effect sizes were small, and several results were inconsistent across parametric and non-parametric tests (see Table 3).

For ChatGPT-4o, JSON files yielded slightly lower OOD–W values than XLSX–DOCX ($M = .0671$ vs. $.0800$). Welch's $t(292.51) = 2.35$, $p = .020$, indicated a statistically significant difference with a small effect size ($d = 0.25$), but significance was not confirmed by the Mann–Whitney U test ($z = -0.37$, $p = .710$). Restricting the analysis to extended prompts to isolate file-type effects from prompt quality, XLSX–DOCX files produced lower OOD–G values than JSON ($M = .0033$ vs. $.0056$), Welch's $t(163.82) = -2.12$, $p = .035$, with a small-to-moderate effect size ($d = -0.32$). The corresponding Mann–Whitney test was non-significant but approached significance ($p = .051$).

Table 3 *Effect of Attachment Format on Lexical and Grammatical Constraint Adherence*

Measure	File Type	<i>n</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>d</i>	95% CI		<i>z</i>	<i>p</i> (<i>MW</i>)
									LL	UL		
ChatGPT-4o												
OOD-W	XLSX-DOCX	180	.0800	.0631	2.35	292.51	.020	0.25	0.04	0.46	-0.37	.710
	JSON	180	.0671	.0378								
OOD-G	XLSX-DOCX	180	.0089	.0125	-0.85(a)	358	.399	-0.09	-0.30	0.12	-1.92	.055
	JSON	180	.0099	.0107								
ChatGPT-4o (Extended Prompt)												
OOD-W	XLSX-DOCX	90	.0458	.0290	0.92(b)	178	.360	0.14	-0.16	0.43	-0.67	.503
	JSON	90	.0421	.0257								
OOD-G	XLSX-DOCX	90	.0033	.0060	-2.12	163.82	.035	-0.32	-0.61	-0.02	-1.96	.051
	JSON	90	.0056	.0082								
Claude 3.5 Sonnet												
OOD-W	PDF	180	.0057	.0064	-3.52(c)	358	<.001	-0.37	-0.58	-0.16	-4.12	<.001
	JSON	180	.0081	.0066								
OOD-G	PDF	180	.0003	.0015	2.05	253.10	.042	0.22	0.01	0.42	-2.33	.020
	JSON	180	.0001	.0007								
Claude 3.5 Sonnet (Extended Prompt)												
OOD-W	PDF	90	.0036	.0041	-2.85(d)	178	.005	-0.43	-0.72	-0.13	-2.56	.010
	JSON	90	.0055	.0051								
OOD-G	PDF	90	.0000	.0000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	JSON	90	.0000	.0000								

Note. All *p*-values are two-tailed. Significant *p*-values are bolded for clarity. Letters in parentheses with *t* values (a, b, c, d) indicate that standard independent-samples *t*-tests were conducted (instead of Welch's *t*-test) since Levene's test indicated equal variances: (a) Levene's $F = 0.82$, $p = .366$; (b) Levene's $F = 1.77$, $p = .186$; (c) Levene's $F = 0.63$, $p = .427$; (d) Levene's $F = 3.37$, $p = .068$. MW = Mann-Whitney U test.

For Claude 3.5 Sonnet, attachment format effects were statistically significant but pointed in opposite directions for vocabulary and grammar. PDF outperformed JSON for OOD-W ($M = .0057$ vs. $.0081$), $t(358) = -3.52$, $p < .001$, with a small-to-moderate effect size ($d = -0.37$). Conversely, JSON outperformed PDF for OOD-G ($M = .0001$ vs. $.0003$), Welch's $t(253.10) = 2.05$, $p = .042$, with a small effect size ($d = 0.22$). Both findings were supported by Mann-Whitney tests. When restricting the analysis to extended prompts, PDF again yielded lower OOD-W ($M = .0036$ vs. $.0055$), $t(178) = -2.85$, $p = .005$, with a small-to-moderate effect size ($d = -0.43$). This result was also corroborated by the Mann-Whitney U test. For grammar, Claude produced no out-of-domain grammatical structures with extended prompts, reflecting perfect adherence regardless of the attachment format.

Overall, evidence for file-type effects is limited. Significant findings are isolated rather than systematic, most effect sizes are small, and several contrasts yield inconsistent conclusions across parametric and non-parametric tests. In the ChatGPT analyses, the variable reaching significance reversed depending on whether the analyses included both prompt types or only extended prompts. In the Claude analyses, attachment format yielded opposite-direction effects for vocabulary and grammar. Although these analyses followed pre-specified hypotheses (and therefore no multiple-comparison correction was applied), the combination of isolated significances, small effects, and occasional inconsistencies suggests that any practical impact of *attachment format* on constraint adherence is likely modest and sensitive to analytic choices.

Attachment Format and User-Perceived Latency

User-perceived latency was measured for all dialogues ($n = 240$) in seconds (s). ChatGPT-4o responded significantly faster when using JSON files ($n = 60$, $M = 2.05$ s, $SD = 0.22$) than with XLSX–DOCX files ($M = 3.90$ s, $SD = 3.21$), Welch’s $t(59.55) = 4.45$, $p < .001$, with a large effect size ($d = 0.81$, 95% CI [0.44, 1.18]). A Mann–Whitney U test ($z = -4.83$, $p < .001$) supported this finding.

Claude 3.5 Sonnet showed a similar pattern. Perceived latency was significantly shorter with JSON files ($M = 3.97$ s, $SD = 0.80$) than with PDF files ($M = 7.80$ s, $SD = 5.83$), Welch’s $t(61.23) = 5.05$, $p < .001$, again with a large effect size ($d = 0.92$, 95% CI [0.54, 1.30]). This result was corroborated by the Mann–Whitney U test ($z = -8.95$, $p < .001$).

Examination of standard deviations highlights differences in performance consistency. XLSX–DOCX (ChatGPT) and PDF (Claude) conditions exhibited high variability in latency ($SD = 3.21$ and 5.83 , respectively), indicating inconsistent and occasionally delayed response times. In contrast, JSON yielded markedly more stable latencies, with substantially lower variability across both models ($SD = 0.22$ for ChatGPT and 0.80 for Claude). These patterns suggest that JSON attachments not only enabled faster generation but also enhanced the stability of perceived responsiveness.

Text Type Comparison

An analysis of OOD–W across the three text types: city descriptions, personal introductions, and dialogues, revealed statistically significant differences. Dialogues showed lower OOD–W ($n = 240$, $M = .0244$, $SD = .0284$) than both city descriptions ($M = .0438$, $SD = .0472$) and personal introductions ($M = .0525$, $SD = .0638$). A Welch’s ANOVA confirmed a significant main effect, $F(2, 430.48) = 28.40$, $p < .001$, with a moderate effect size ($\eta^2 = .06$). Games–Howell post hoc comparisons indicated that dialogues contained significantly lower OOD–W than both city descriptions and personal introductions ($p < .001$ for both), while the latter two did not differ significantly ($p = .210$). A Kruskal–Wallis test further supported these findings, $\chi^2(2) = 13.77$, $p = .001$, and Bonferroni-adjusted pairwise comparisons corroborated the parametric results: dialogues vs. city descriptions ($p = .035$), dialogues vs. personal introductions ($p < .001$), and city descriptions vs. personal introductions ($p = .816$).

For OOD–G, city descriptions yielded the lowest values ($M = .0012$, $SD = .0034$), followed by dialogues ($M = .0041$, $SD = .0071$) and personal introductions ($M = .0091$, $SD = .0132$). A Welch’s ANOVA revealed a statistically significant effect, $F(2, 391.71) = 52.11$, $p < .001$, with a moderate-to-large effect ($\eta^2 = .13$). Games–Howell post hoc comparisons showed all pairwise differences to be statistically significant ($p < .001$ for all). These findings were corroborated by a Kruskal–Wallis test, $\chi^2(2) = 79.25$, $p < .001$, and Bonferroni-corrected pairwise comparisons ($p < .001$ for all).

Error Metrics

Errors in Vocabulary

Only two vocabulary errors were identified in 151,589 words generated across 720 texts—an exceptionally low rate (0.0013%)—both in personal introductions:

- (a) ChatGPT-4o; extended prompt; XLSX–DOCX: “[...] y *visitar [ir a] un concierto.” The verb *visitar* was inappropriate in this context.
- (b) Claude 3.5 Sonnet; standard prompt; PDF: “Lo hablo bien porque es mi lengua *oficial [materna].” The adjective *oficial* was semantically incorrect in this context.

Given the extremely low number of cases, no generalizations can be drawn about their source.

Errors in Grammar

Grammatical errors were equally rare, appearing in only two texts generated by ChatGPT-4o:

- (a) Personal introduction; extended prompt; XLSX–DOCX: “[...] y **ver* [veo] *películas en francés.*” The verb *ver* was incorrectly given in the infinitive instead of the first-person form *veo*.
- (b) Dialogue; standard prompt; JSON: “*¿*De qué apellido te llamas?* [¿*Cómo te llamas?*].” This error is more consequential as it affects a core communicative function taught early in Spanish learning. Nevertheless, it was the only instance of its kind in the dataset.

These occurrences also appear to be outliers, and no systematic patterns can be inferred.

Errors in Coherence

Coherence issues were infrequent but revealed notable patterns, particularly in the personal introductions:

- (a) Claude 3.5 Sonnet, when using extended prompts, produced the same structural coherence error six times: “*Hablo [LANGUAGE] porque *vivo [viví] en [PLACE] por [NUMBER] años.*” This formulation occurred three times with JSON and three times with PDF files. The model attempted to justify current language ability by referencing a past living experience, but to avoid using the restricted past tense, defaulted to an inappropriate present-tense structure that violated logical coherence.
- (b) ChatGPT-4o, when using standard prompts and XLSX–DOCX files, shifted from first person to third person (from “I” to “he/she”) in eight different personal introductions. This pattern did not occur in texts generated with extended prompts or JSON files.

Additional coherence errors occurred within dialogues, though these were isolated instances. Using standard prompts with JSON files, ChatGPT twice produced “*¿*Cuáles son tus trabajos?* [¿*Cuál es tu trabajo?*]” (incorrect plural form). In another instance, a dialogue character asked for someone’s name after it had already been provided: “*Mucho gusto, Carlos. *¿Cómo te llamas?*”

Coherence errors, though relatively infrequent overall, occurred more frequently than vocabulary and grammatical errors. These patterns carry important implications for prompt engineering and AI tool design, as discussed in the following section.

Word Count

A descriptive analysis of word count patterns revealed that models adhered more closely to length requirements in shorter tasks. City descriptions, targeted at 150 words, averaged 151.83 ($n = 240$, $SD = 29.48$), indicating high compliance. By contrast, the longer personal introductions, targeted at 250 words, yielded a mean word count of 217.58 ($SD = 27.82$), falling short of the target. For dialogues, where no specific constraint was imposed, outputs were longer overall ($M = 262.21$, $SD = 34.32$), likely reflecting the open-ended and interactive nature of this text type.

A model-level breakdown reveals different patterns of adherence. ChatGPT-4o consistently undershot word count targets in both city descriptions ($n = 120$, $M = 126.68$, $SD = 17.98$) and personal introductions ($M = 200.27$, $SD = 23.32$). Claude 3.5 Sonnet, in contrast, exceeded the target for city descriptions ($M = 176.98$, $SD = 12.08$) while falling slightly short for personal introductions ($M = 234.90$, $SD = 20.13$). For dialogues, Claude generated longer outputs ($M = 279.75$, $SD = 21.06$) than ChatGPT ($M = 244.67$, $SD = 36.08$).

Discussion

Model Selection as a Primary Driver of Constraint Adherence

Across 720 texts, Claude 3.5 Sonnet consistently outperformed ChatGPT-4o on every constraint metric, confirming that model choice is a first-order design decision when generating controlled L2 materials. Claude demonstrated superior adherence to both lexical and grammatical constraints. On average, it produced only 0.69% OOD-W per text, compared to 7.36% from ChatGPT, and just 0.02 OOD-G per 100 words compared to 0.94. Moreover, Claude exhibited substantially smaller standard deviations (0.66% vs. 5.23% for OOD-W; 0.11 vs. 1.16 for OOD-G), suggesting not only greater accuracy but also higher reliability.

The robustness of these differences is supported by consistently low p -values ($p < .001$) and large effect sizes, providing strong empirical support for hypothesis **H1**: Different AI models will exhibit different adherence to linguistic constraints (OOD-W, OOD-G) under the same prompt and attachment type conditions.

These findings align with previous research showing the influence of LLMs on constraint adherence (e.g., Glandorf & Meurers, 2024; Glandorf et al., 2025; Imperial & Madabushi, 2023; Li et al., 2024; Malik et al., 2024). Notably, Malik et al. (2024) found that even a simple prompt used with GPT-4 outperformed more elaborate prompts applied to open-source models such as LLaMa-2-7b-chat and Mistral-7b-instruct, demonstrating that model architecture can outweigh prompt complexity.

Prompt Engineering as a Primary Strategy for Linguistic Control

Extended prompts significantly improved adherence to linguistic constraints in both models, reducing non-compliance across both vocabulary and grammar. Vocabulary non-compliance fell from 5.62% to 2.43% and grammatical non-compliance from 0.74 to 0.22 per 100 words. The largest gains occurred in ChatGPT, where non-compliant word averages decreased from 10.32% to 4.39%, and non-compliant grammatical structure rates dropped from 1.43 to 0.44. Although Claude already showed near-perfect adherence, it still improved, with lexical violations declining from 0.93% to 0.46%, and grammatical violations from 0.04 to zero.

These gains extended to reliability in both models. For Claude, standard deviations fell from 0.74% to 0.47% for vocabulary, and from 0.16 to 0 for grammar. ChatGPT's improvements were especially notable, with standard deviations decreasing from 5.46% to 2.74% for vocabulary, and from 1.30 to 0.73 for grammar. These results indicate that extended prompts not only reduce non-compliance rates but also enhance output consistency; an essential factor in instructional settings, where control and predictability are critical.

The statistical significance of these improvements was confirmed by low p -values ($p < .001$), providing strong support for **H2**: Extended prompts will improve model adherence to linguistic constraints. Effects were particularly large for ChatGPT, suggesting that models with lower constraint adherence benefit the most from extended prompts.

These results align with prior research in CTG (e.g., Imperial & Madabushi, 2023; Malik et al., 2024). For example, Malik et al. (2024) found that prompts incorporating CEFR-level descriptions or concrete linguistic examples improved output accuracy, further validating enhanced prompting as a means of linguistic control.

Non-Compliance and the $i+1$ Threshold in Generated Output

The non-compliance observed in this study aligns with Krashen's $i+1$ input hypothesis under certain controlled conditions. While Krashen did not explicitly quantify how much new language constitutes the "+1," empirical research provides practical thresholds. Learners generally require at least 95% known-word coverage for minimal reading comprehension (Laufer, 1989), and 98% for independent comprehension (Hu & Nation, 2000; Schmitt et al., 2011). For listening comprehension, estimates range from 90 to 95% for basic understanding (Van Zeeland & Schmitt, 2013). Although these thresholds were not developed specifically for absolute beginners, they offer a reasonable approximation until research targets this population directly.

Based on these estimates, L2 instructional texts for beginners should ideally include no more than 2–5 unfamiliar words per 100 words. Claude 3.5 Sonnet meets this criterion under both prompting conditions, with vocabulary non-compliance rates of 0.93% under the standard prompt and 0.46% with the extended prompt. Its consistently low standard deviations further confirm its reliability, enabling educators to incorporate pre-planned unfamiliar words into the vocabulary lists without exceeding comprehensibility thresholds.

By contrast, ChatGPT-4o exceeds the acceptable threshold when using the standard prompt, with a non-compliance rate of 10.32% and high variability ($SD = 5.46\%$), making it unsuitable for absolute beginners in that condition. However, when paired with the extended prompt, its non-compliance drops to 4.39%—just below the upper bound for minimal comprehension—and its standard deviation decreases to 2.74%, bringing it within acceptable parameters for beginner instruction. Even so, it lacks the precision needed to support the introduction of planned unknown vocabulary.

While no established thresholds exist for grammatical structures comparable to those for vocabulary, the observed OOD-G rates suggest reasonable levels of control. ChatGPT's highest non-compliance rate (1.43 per 100 words with the standard prompt) drops to 0.44 under extended prompting, while Claude's rates range from 0.04 to full compliance. These results suggest that both models, particularly when used with extended prompts, can generate texts that sufficiently adhere to grammatical constraints.

Taken together, these findings reinforce earlier conclusions: prompt engineering is a first-order technique, especially critical when deploying models with lower baseline adherence to linguistic constraints. With stronger models, educators can introduce planned unfamiliar vocabulary with greater flexibility and control. In contrast, with less-performant models, stricter prompting becomes essential, as the "+1" is already occupied by incidental out-of-domain words that escape control during generation.

Attachment Format Does Not Consistently Affect Adherence, but Influences Perceived Latency and Response-Time Stability

Despite some observed differences across file formats, the results do not support a systematic effect of attachment format on adherence to lexical or grammatical constraints. Statistically significant differences appeared in isolated cases, but these effects were modest in magnitude and failed to replicate consistently across models, prompt types, or statistical tests.

Claude's slightly lower vocabulary violation rates with PDF files compared to JSON warrant brief consideration, as the effect remained statistically significant across analyses. One possible interpretation is that document-based formats may support more reliable enforcement of lexical constraints than data-structured formats, although the mechanisms underlying this pattern are unclear. This interpretation remains speculative given the limited magnitude of effect sizes and the inconsistency across constraint types; JSON

performed better for grammatical constraints in the combined analysis integrating both prompt types. Clarifying these effects would require further research that systematically compares document-based and data-structured formats, rather than simple-lightweight versus rich-heavyweight formats.

In contrast, attachment format exerted a robust and consistent influence on user-perceived latency and response stability. JSON files produced significantly faster response times than XLSX/DOCX and PDF formats, with consistently low p -values ($p < .001$) and large effect sizes for both ChatGPT-4o and Claude 3.5 Sonnet. On average, ChatGPT's latency nearly doubled when using XLSX–DOCX files (3.90s) compared to JSON (2.05s). Claude showed a similar pattern, with average latency decreasing from 7.80s with PDF files to 3.97s with JSON.

More notably, JSON files substantially reduced response-time variability, indicating more stable performance. For ChatGPT, the standard deviation dropped from 3.21s (XLSX–DOCX) to 0.22s (JSON). For Claude, variability decreased even more dramatically, from 5.83s to 0.80s.

Taken together, these findings suggest that any potential effect of attachment format on constraint adherence is marginal compared with the pronounced advantages JSON offers in response speed and predictability. These benefits are especially relevant in real-time educational applications, such as AI-powered tutors, conversational agents, or instructional content generators, where responsiveness and reliability are critical.

These results provide partial support for hypothesis **H3**. While JSON attachments were indeed processed more efficiently, the data do not support **H3a**, as no file format systematically enhanced adherence to linguistic constraints. By contrast, the data strongly support **H3b**, as JSON files significantly reduced user-perceived latency and, additionally, improved response predictability—both important dimensions of efficiency in applied settings.

Although user-perceived latency was recorded manually, the magnitude of the observed differences in response times and variability far exceeded any plausible timing error, suggesting that the findings are robust and practically meaningful. The performance advantages of JSON on latency align with prior research showing that its minimal metadata overhead and efficient structure yield faster processing times and a reduced memory footprint compared with more heavyweight formats like XML (Zunke & D'Souza, 2014).

Text Type Affects Adherence—and Generalization Requires Caution

The influence of text type, examined as a secondary focus, revealed several noteworthy insights. Although all three text types were generated under identical conditions, statistically significant differences in constraint adherence emerged. Across all generated texts—regardless of model or prompt type—dialogues produced the lowest rate of non-compliant words (2.44%), compared to both city descriptions (4.38%) and personal introductions (5.25%). In contrast, city descriptions exhibited the fewest grammatical violations (0.12 non-compliant structures per 100 words), followed by dialogues (0.41) and personal introductions (0.91).

These results support the hypothesis **H4**: Different text types will affect linguistic constraint adherence, as certain genres or topics can facilitate or hinder control. These findings carry important implications for practical application. While these genres and topics reflect common beginner-level tasks, they do not capture the full range of text types typically used with absolute beginners. This highlights the need for caution when generalizing these results to other tasks, which may exhibit different patterns of constraint adherence. To ensure that generated texts meet learners' needs, it is essential to examine model behavior across the specific task types relevant to classroom contexts.

Further Evidence of Validity: Low Error Rates and Practical Considerations

Across the 720 texts analyzed in this study—comprising 151,589 generated words—only marginal instances of vocabulary, grammar, or coherence errors were identified: two vocabulary errors, two grammatical errors, and a small number of coherence issues, mostly involving isolated structures. This exceptionally low incidence rate reinforces the overall validity and pedagogical usability of the texts.

Even rare errors, however, can pose risks. Lexical substitutions such as *lengua *oficial* for *materna*, or ungrammatical formulations like **¿De qué apellido te llamas?* may mislead students, particularly at the absolute beginner level.

Some coherence issues also displayed weak but noticeable patterns. Claude occasionally used present-tense forms for past events (e.g., **vivo en [...]* instead of *viví en [...]*), while ChatGPT occasionally shifted grammatical person to *él/ella* in first-person introductions. Although too infrequent to establish systematic trends, such issues underscore the need for careful validation when using AI-generated instructional content or developing AI learning tools.

Given the small sample of models evaluated, it remains unclear whether other LLMs may exhibit more frequent or systematic errors. Educators and developers should therefore assess model output for error type and frequency before classroom use. When identifiable patterns emerge, such as consistent misuse of a grammatical structure, prompting strategies can be adapted to mitigate the potential risks. For instance, if a model tends to shift to third-person pronouns in first-person introductions, the prompt could specify: “*When generating a first-person introduction, consistently use the pronoun ‘yo’ to refer to yourself.*” Such targeted prompt refinements offer a practical solution to enhance the validity and reliability of AI-generated content.

In sum, while the observed error rate was exceptionally low, AI-generated output should never be assumed error-free. Before deploying such content or AI applications in educational settings, model outputs should be thoroughly tested. Any recurring patterns of error should be addressed through prompt design or manual post-editing, depending on whether the context involves automated applications or instructional materials. Finally, learners should be explicitly informed that AI outputs are not infallible and that critical engagement with these tools is essential for effective language learning.

Text Length is Difficult to Regulate

Without any prompting strategies explicitly aimed at controlling word count, both models produced texts whose lengths diverged from the predefined targets yet remained pedagogically viable. For city descriptions with a 150-word target, ChatGPT generated a mean of 126.68 words per text, while Claude produced 176.98. For personal introductions targeting 250 words, the outputs averaged 200.27 for ChatGPT and 234.90 for Claude. These results indicate that word count remains a difficult parameter to regulate without explicit instruction. If greater precision is needed in specific instructional contexts, such as standardized assessments, future work should explore prompting techniques that better manage text length without compromising adherence to lexical and grammatical constraints.

Conclusions

This study proposed and evaluated a practical, prompt-based approach for generating tightly controlled texts suitable for absolute beginners in second language learning. Prompting was chosen over more complex and resource-intensive CTG methods to ensure scalability and accessibility, enabling educators to adopt the approach without highly technical expertise.

The method relied on strategically designed prompts to guide general LLMs in producing beginner-level texts aligned with predefined linguistic constraints. These constraints were operationalized through attached knowledge base files: a categorized list of allowed vocabulary, as well as a comprehensive set of example sentences illustrating the permitted grammatical structures organized by communicative functions.

To refine the method and provide practical implementation guidance, three key variables were systematically manipulated: (1) the *AI model* (ChatGPT-4o vs. Claude 3.5 Sonnet), (2) the *prompt type* (standard vs. extended, incorporating constraint-enhancing techniques), and (3) the *attachment format* (rich, heavyweight formats—XLSX, DOCX, and PDF—vs. the lightweight, highly structured, and simple JSON).

Key Findings

Model selection exerted the strongest influence on constraint adherence, highlighting the critical importance of evaluating an LLM's compliance with linguistic constraints as a foundational step. Extended prompts incorporating specific constraint-enhancing strategies significantly improved performance, enabling less performant models to achieve pedagogically viable outputs that met the 95% word-coverage threshold essential for minimal text comprehensibility.

While attachment format showed no systematic advantages for constraint adherence, JSON reduced user-perceived latency by approximately 50% and substantially lowered response-time variability, making it preferable for real-time applications. Other structured, simple, and lightweight formats, such as Markdown or CSV, may offer similar benefits. Text type also affected constraint adherence, indicating that educators should test their genre and topic targets before deployment. Error rates were remarkably low across all conditions, although other models may exhibit different error profiles that could require additional prompt refinements.

The best-performing configuration in our sample demonstrated that the proposed approach can deliver remarkable lexical and grammatical control without the technical sophistication and computational cost associated with advanced CTG methods. Specifically, Claude 3.5 Sonnet with the extended prompt achieved fewer than 0.5% non-compliant words per text and zero non-compliant grammatical structures.

Practical Implications

This research offers second-language educators a scalable, immediately implementable framework for generating customized materials or developing AI tools through user-friendly platforms supporting prompt customization and knowledge base integration. The findings provide actionable guidance for selecting AI models, designing prompts, and choosing attachment formats to suit diverse instructional contexts and resource constraints.

Limitations and Future Research Directions

Several limitations merit attention. First, the findings are specific to Spanish, two LLMs, and three text types; replication across typologically diverse languages and a broader range of text genres and topics is therefore essential. Additionally, while the approach targets predetermined vocabulary and grammatical structures aligned with course progression, individual learner knowledge inevitably varies and cannot be fully accommodated without additional personalization mechanisms.

At the linguistic level, further research should investigate other surface-level features, such as lexical and grammatical density, that may influence variability beyond basic constraint adherence. Higher-order dimensions such as textual relevance, authenticity, and naturalness also require systematic evaluation.

Methodologically, although residual dependence is highly unlikely in practice given the independence safeguards described in the Methods section, subtle system-related correlations cannot be theoretically ruled out entirely. Nevertheless, the medium-to-large and large effect sizes and very low p -values observed make it improbable that any minor dependence, if present, would materially affect the conclusions.

Regarding measurement, user-perceived Time to First Token (TTFT) latency was recorded manually, introducing some degree of random noise. While this method realistically captures the user-facing experience, future work could incorporate automated measures (e.g., network interception or JavaScript instrumentation) for greater temporal precision. However, as discussed before, the magnitude of observed differences far exceeds any plausible timing error.

Finally, because LLMs evolve rapidly, findings may require periodic updates, even though the underlying principles of strategic prompt design and structured knowledge representation are likely to remain relevant.

Future research should: (1) apply the proposed framework in authentic educational contexts, including the development of instructional materials and AI-based tools; (2) integrate these principles into adaptive systems that monitor and respond to individual learners' lexical and grammatical development; and (3) empirically evaluate the implementation of generated materials and AI tools with actual learners to assess factors such as user experience, engagement, and impact on learning outcomes.

Despite these limitations, this study shows that prompt engineering can achieve the level of linguistic control necessary for absolute beginner instruction while remaining accessible to educators—a critical balance for widespread adoption in language education.

References

- Anthony, L. (2024). *AntWordProfiler* (Version 2.2.1) [Computer software]. Waseda University. <https://www.laurenceanthony.net/software/AntWordProfiler>
- Glandorf, D., & Meurers, D. (2024). Towards fine-grained pedagogical control over English grammar complexity in educational text generation. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 299–308). Association for Computational Linguistics. <https://aclanthology.org/2024.bea-1.24/>
- Glandorf, D., Cui, P., Meurers, D., & Sachan, M. (2025). Grammar control in dialogue response generation for language learning chatbots. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 9820–9839). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.naacl-long.495>
- Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–430. <https://doi.org/10.64152/10125/66973>
- Imperial, J. M., & Madabushi, H. T. (2023). Flesch or fumble? Evaluating readability standard alignment of instruction-tuned language models. In S. Gehrmann, A. Wang, J. Sedoc, E. Clark, K.

- Dhole, K. R. Chandu, E. Santus, & H. Sedghamiz (Eds.), *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)* (pp. 205–223). Association for Computational Linguistics. <https://aclanthology.org/2023.gem-1.18/>
- Iso, H. (2024). AutoTemplate: A simple recipe for lexically constrained text generation. In S. Mahamood, N. L. Minh, & D. Ippolito (Eds.), *Proceedings of the 17th International Natural Language Generation Conference* (pp. 1–12). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.inlg-main.1>
- Jin, M., Dugan, L., & Callison-Burch, C. (2025). *Controlling difficulty of generated text for AI-assisted language learning*. arXiv. <https://doi.org/10.48550/arXiv.2506.04072>
- Kim, S., & Park, S. H. (2023). Young Korean EFL learners' perception of role-playing scripts: ChatGPT vs. textbooks. *Korea Journal of English Language and Linguistics*, 23, 1136–1153. <https://doi.org/10.15738/kjell.23..202312.1136>
- Krashen, S. (1985). *The input hypothesis: Issues and implications* (Vol. 1). Longman
- Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren, & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters.
- Li, Y., Qu, S., Shen, J., Min, S., & Yu, Z. (2024). Curriculum-driven EduBot: A framework for developing language learning chatbots through synthesizing conversational data. In T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, & K. Komatani (Eds.), *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 400–419). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.sigdial-1.35>
- Malik, A., Mayhew, S., Piech, C., & Bicknell, K. (2024). From Tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 15670–15693). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.926>
- Nie, J., Yang, L., Chen, Y., Kong, C., Zhu, J., & Yang, E. (2023). Lexical complexity controlled sentence generation for language learning. In M. Sun, B. Qin, X. Qiu, J. Jing, X. Han, G. Rao, & Y. Chen (Eds.), *Chinese computational linguistics: 22nd China national conference, CCL 2023, Harbin, China, August 3–5, 2023, proceedings* (pp. 106–126). Springer. https://doi.org/10.1007/978-981-99-6207-5_7
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, 95(1), 26–43. <https://doi.org/10.1111/j.1540-4781.2011.01146.x>
- Shin, D., Lee, J., & Kim, K. (2025). An exploratory study on two automated item generators for generating L2 reading test items. *RELC Journal*. <https://doi.org/10.1177/00336882251326284>
- Stowe, K., Ghosh, D., & Zhao, M. (2022). Controlled language generation for language learning items. In Y. Li & A. Lazaridou (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 294–305). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-industry.30>
- Uchida, S. (2025). Generative AI and CEFR levels: Evaluating the accuracy of text generation with ChatGPT-4o through textual features. *Vocabulary Learning and Instruction*, 14(1), Article 2078. <https://doi.org/10.29140/vli.v14n1.2078>
- Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, 34(4), 457–479. <https://doi.org/10.1093/applin/ams074>
- Wiboolyasarini, W., Wiboolyasarini, K., Tiranant, P., Jinowat, N., & Boonyakitanont, P. (2025). AI-driven chatbots in second language education: A systematic review of their efficacy and pedagogical implications. *Ampersand*, 14, Article 100224. <https://doi.org/10.1016/j.amper.2025.100224>

- Young, J. C., & Shishido, M. (2023). Evaluation of the potential usage of ChatGPT for providing easier reading materials for ESL students. In T. Bastiaens (Ed.), *Proceedings of EdMedia + Innovate Learning* (pp. 155–162). Association for the Advancement of Computing in Education (AACE). <https://www.learntechlib.org/primary/p/222496/>
- Zhang, H., Song, H., Li, S., Zhou, M., & Song, D. (2023). A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3), Article 64. <https://doi.org/10.1145/3617680>
- Zunke, S., & D'Souza, V. (2014). JSON vs XML: A comparative performance analysis of data exchange formats. *International Journal of Computer Science and Network*, 3(4), 257–261. <https://ijcsn.org/articles/0304/JSON-vs-XML-A-Comparative-Performance-Analysis-of-Data-Exchange-Formats.html>

Appendix

Prompt Examples

The following examples illustrate the standard and extended prompt types used in this study. Both examples are for the city description task (Barcelona). Prompts for personal introductions and dialogues followed the same structure but with task-specific content adapted to each text type.

Standard Prompt

Generate one coherent text in Spanish that describes Barcelona. The text must be generated in two paragraphs with around 150 words. Include the following aspects:

- Its location.
- General characteristics (including at least two main places of tourist interest).
- Other specific information.
- Climate.

The text must strictly adhere to the following guidelines:

Vocabulary Constraint:

1. Use only the words provided in the vocabulary list (refer to “SPA001_Vocabulary”).
2. Words can vary in gender, number, and person to provide accurate Spanish sentences.
3. Proper nouns such as names of people, surnames, cities, tourist places, and typical dishes can be freely generated and are not restricted to the vocabulary list.

Grammar and Structure Constraint:

Follow the grammatical structures and sentence patterns exactly as given in the example sentences (refer to “SPA001_Functions and Examples”). This includes word order, verb conjugation, use of prepositions, and other syntactical elements.

Extended Prompt

You are a skilled language model tasked with generating Spanish texts for Spanish beginners based on limited constraints. Key resources:

- A list of Spanish vocabulary words grouped by categories (refer to “SPA001_Vocabulary”)
- A set of example sentences organized by communicative functions (refer to “SPA001_Functions and Examples”).

Your task is to create coherent Spanish texts for Spanish beginners that strictly adhere to the following guidelines:

****Vocabulary Constraint:****

1. Use only the words provided in the vocabulary list. No additional words are allowed.
2. Words can vary in gender, number, and person to provide accurate Spanish sentences.
3. Proper nouns such as names of people, surnames, cities, tourist places, and typical dishes can be freely generated and are not restricted to the vocabulary list.

****Grammar and Structure Constraint:****

1. Follow the grammatical structures and sentence patterns exactly as given in the example sentences. This includes word order, verb conjugation, use of prepositions, and other syntactical elements.
2. Verbs can only be conjugated in the present tense (indicative mood). Avoid other tenses and moods such as past tense or subjunctive.
3. Do not introduce any new grammatical structures or sentence forms not present in the examples.

****TASK:****

Generate one coherent text in Spanish that describes Barcelona. The text must be generated in two paragraphs with around 150 words. Include the following aspects:

- Its location.
- General characteristics (including at least two main places of tourist interest).
- Other specific information.
- Climate.

Use only the words from the provided list (“SPA001_Vocabulary”) and follow the grammatical structures and style of the given examples (“SPA001_Functions and Examples”).

****Important Instructions:****

- Do not include words not present in the provided vocabulary list (“SPA001_Vocabulary”).
- Do not include grammatical structures not present in the given examples (“SPA001_Functions and Examples”).