Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

Language Education & Assessment, 9, 104760 (2026)
https://doi.org/10.29140/lea.2026.104760

Assessing AI-Generated Academic Writing in Spanish: A Multivariate Analysis of Syntactic Variation Across ChatGPT, Gemini, and Claude

ESTEBAN VÁZQUEZ-CANO

Faculty of Education. Department of Didactics and School Organization. Universidad Nacional de Educación a Distancia (UNED). Madrid, Spain
evazquez@edu.uned.es

Abstract

Research on large language models has increasingly examined their capacity to produce academically acceptable texts, yet the structural properties of AI-generated academic writing remain underexplored, particularly in Spanish. This study examines whether three proprietary large language models, ChatGPT, Gemini, and Claude, generate distinguishable syntactic configurations in academic Spanish and whether these configurations provide interpretable evidence for language assessment contexts. Using a controlled intra-task design, a corpus of 90 texts was generated under identical prompts, the same source text and task instructions as supporting material, and standardized production conditions. The design did not include a human-written baseline corpus; therefore, the study addresses inter-model syntactic variation rather than human-AI differentiation. Syntactic structure was operationalized through the Universal Dependencies framework and extracted automatically with Stanza. Normalized frequencies of coordination and subordination relations, together with mean sentence length as a structural control variable, were analyzed through Welch's ANOVA, MANOVA, PERMANOVA, principal component analysis, linear discriminant analysis, and Mahalanobis distances with bootstrap confidence intervals. Results showed significant differences across all syntactic variables, with the strongest effects concentrated in coordination and more moderate differences in subordinate constructions. Multivariate analyses confirmed robust structural separation across models, and linear discriminant analysis achieved 91.1% classification accuracy. These findings indicate that the models generate academic Spanish through differentiated syntactic configurations. For language assessment, the results suggest that dependency-based syntactic profiling can support the

Data Availability Statement: *All relevant data are within this paper.*

interpretation of AI-assisted academic writing, inform rubric refinement for academic Spanish, and provide transparent features for automated writing evaluation and model-attribution procedures under controlled conditions.

Keywords: AI-generated writing, language assessment, academic writing assessment, automated writing evaluation, Spanish syntax, large language models, syntactic profiling

Introduction

The rapid diffusion of large language models (LLMs) in educational environments has intensified research on their capacity to generate academically acceptable written discourse. In higher education contexts, these systems increasingly produce extended argumentative texts that resemble the rhetorical and stylistic conventions expected in university writing. As a result, much of the existing research has focused on evaluating global performance indicators of AI-generated texts, including task accuracy, coherence, and reliability, often in relation to established frameworks used in writing assessment (Chang et al., 2024; de Wynter et al., 2023). While this line of work provides important insights into the pedagogical potential of LLMs, it largely emphasizes outcome-oriented measures of textual quality and devotes comparatively less attention to the structural linguistic properties through which model-generated discourse is constructed.

At the same time, emerging studies suggest that AI-generated texts may exhibit identifiable linguistic regularities when compared with human-authored writing. Rather than producing stylistically homogeneous outputs, LLMs appear to generate recurring structural patterns that can function as computationally detectable "fingerprints." This observation is methodologically relevant because it indicates that AI-generated writing can be analyzed through measurable linguistic features rather than relying exclusively on qualitative evaluations of fluency or coherence (Mizumoto et al., 2024). In this respect, syntactic organization represents a particularly informative analytical dimension, as the distribution of coordination and subordination structures plays a central role in the organization of complex academic discourse.

Despite these advances, systematic structural analyses of AI-generated academic writing remain limited, particularly for languages other than English. In Spanish-language educational contexts, existing research has mainly examined issues related to textual adequacy, complexity, or pedagogical usability, while more detailed structural analyses of model-generated discourse remain scarce (Clementi & Jiménez-López, 2025). This limitation is especially relevant in higher education, where analytical and argumentative writing constitutes a core component of academic assessment. Understanding how LLMs structurally realize this type of discourse is therefore necessary for evaluating their implications for writing assessment, academic integrity, and AI-supported learning environments.

To address this gap, the present study investigates the generation of academic Spanish by three widely used large language models (ChatGPT, Gemini, and Claude) from an assessment-oriented perspective. Syntactic structure is operationalized through the Universal Dependencies framework, which provides cross-linguistically consistent representations of grammatical relations central to coordination and subordination (de Marneffe et al., 2021). By combining dependency-based syntactic metrics with multivariate statistical analysis, the study examines whether AI-generated academic texts exhibit model-specific structural configurations that may affect how writing evidence is interpreted in educational and language assessment contexts.

The study is guided by the following research questions: (1) Are there statistically significant differences in the distribution of syntactic structures, particularly coordination and subordination, across the analyzed models? (2) Do these differences configure statistically robust multivariate syntactic profiles? (3) Is it possible to automatically discriminate the generating model from its syntactic profile? And (4) Does geometric separation between multivariate centroids confirm the existence of differentiated structural signatures in AI-generated academic Spanish?

By addressing these questions, the study contributes to language education and assessment by linking interpretable syntactic evidence with the evaluation of AI-assisted academic writing in Spanish. Its contribution is not to propose a universal detector of AI-generated text, but to examine whether model-specific structural tendencies may constitute relevant evidence when assessors interpret academic writing, revise rubrics, or design automated writing evaluation systems.

Assessing AI-Generated Academic Writing: Structural Perspectives

Assessing AI-generated academic writing has evolved from a peripheral issue in technology-enhanced learning into a problem of construct validity in writing assessment. Contemporary large language models (LLMs) can produce extended, register-appropriate prose that meets many surface expectations of university writing, raising questions about how writing quality should be evaluated when texts may be partially or fully generated by AI systems. From an assessment perspective, the central concern is no longer whether LLMs can produce plausible academic texts, but whether such outputs can be evaluated within principled scoring frameworks, whether human and automated judgments converge, and whether model-specific linguistic patterns introduce construct-irrelevant variance or new sources of bias in writing evaluation (Dwivedi et al., 2023; Zhong et al., 2024; Vázquez-Cano et al., 2023; Weber-Wulff et al., 2023).

One important development in this literature has been the use of established assessment frameworks to benchmark AI-generated writing. Zhong et al. (2024), for example, evaluate multiple LLM families using the GRE Analytical Writing framework, combining trained human ratings with the ETS e-rater system to compare model outputs under controlled conditions. Their results illustrate that LLM-generated texts can achieve high scores when evaluated against standardized constructs such as critical reasoning and argumentative coherence, while also highlighting challenges for AI detection, as detectability varies depending on whether detection systems are trained within or across models (Zhong et al., 2024; Yan et al., 2023).

A related line of research examines whether LLMs can function as evaluators of writing through automated scoring or rubric-based feedback. Classroom-oriented studies suggest that AI systems can classify or assess student texts with non-trivial accuracy, although human raters often remain more reliable when evaluation depends on nuanced discourse interpretation (Atasoy & Moslemi Nezhad Arani, 2025). Recent studies point in the same direction: AI-powered evaluation can approximate human assessment under controlled conditions (Es-sarghini & Boumahdi, 2026), and AI-generated feedback can support gains in IELTS writing while still requiring teacher mediation (Jitpaisarnwattana, 2026). This research is particularly relevant in multilingual educational contexts, where rhetorical conventions and syntactic realizations may influence how writing quality is interpreted and assessed (Atasoy & Moslemi Nezhad Arani, 2025; Graham & Rijlaarsdam, 2016).

Another strand of research focuses on the quality of AI-generated feedback as an assessment output. Studies show that even when feedback is generated using detailed rubric-aligned prompts, it may remain generic and insufficiently grounded in specific textual evidence, suggesting that rubric alignment alone does not guarantee valid assessment practices (Lo et al., 2026; Steiss et al., 2024).

These findings reinforce the need to examine not only the quality of AI-generated texts but also the evaluative processes through which they are interpreted.

In parallel, several studies have identified measurable linguistic differences between AI-generated and human-authored writing. Machine learning approaches can distinguish between the two with high accuracy using features such as paragraph complexity, sentence-length diversity, punctuation patterns, and lexical distributions (Desaire et al., 2023). Similarly, analyses of educational discourse show that AI-generated texts may exhibit distinctive patterns of repetition, phrase structure, or sentence organization that vary across models (Durak et al., 2025). These results suggest that AI-generated writing may contain identifiable structural signatures that affect both detectability and the interpretation of writing evidence (Desaire et al., 2023; Durak et al., 2025; Zhong et al., 2024).

Taken together, this literature indicates that research on AI-generated academic writing must address three complementary assessment targets: the quality of AI-produced texts, the reliability of AI-generated evaluative outputs such as scores and feedback, and the validity of detection or attribution mechanisms used in academic-integrity frameworks (Walters, 2023; Weber-Wulff et al., 2023; Zhong et al., 2024; Lo et al., 2026; Steiss et al., 2024; Vázquez-Cano et al., 2021). Research in L2 writing assessment also shows that automated textual features can contribute to writing-quality modelling and task classification, provided that such features are interpreted in relation to task demands and human scoring evidence (Tywoniw & Crossley, 2019). For multilingual academic writing in particular, these challenges highlight the need for analytically transparent approaches capable of linking linguistic structure with computational evidence. In this context, examining the syntactic organization of AI-generated discourse offers a promising avenue for understanding whether different models produce identifiable structural patterns in academic writing.

Syntax matters for assessment because it offers observable evidence about how academic writing is organized. In writing assessment, syntactic organization helps represent the construct of academic discourse because it shows how writers link clauses, package information, and control rhetorical progression. If AI-generated texts show stable model-specific patterns, those patterns may influence how assessors interpret writing samples, support academic-integrity decisions, and select transparent features for automated essay scoring (AES) or automated writing evaluation (AWE) systems. This does not convert syntactic profiling into a stand-alone judgment of authorship or quality; rather, it positions it as one component of a broader validity argument.

Syntactic Structure as an Indicator of Academic Writing Quality in Spanish

Research on writing development and assessment has consistently treated syntactic structure as an important dimension of written performance. However, this literature does not assume that more syntactic complexity automatically means better writing. Syntactic complexity must instead be interpreted in relation to the theoretical framework, the unit of analysis, the discourse genre, and the communicative demands of the writing task (Crespo Allende et al., 2011; Beers & Nagy, 2009). From this perspective, syntax is not a superficial feature of written discourse. It is an observable domain in which researchers can examine how academic writing organizes information, establishes interclausal relations, and manages textual density.

This perspective is especially relevant for academic writing in Spanish. Bustos Gisbert (2018) argues that analyses of written syntax should move beyond lists of sentence types and examine how writers use syntactic resources in discourse. The relevant issue, therefore, is not merely whether a text contains subordinate or coordinated structures, but how those structures help organize written academic discourse. For the present study, this distinction is essential: the aim is not to count isolated grammatical

forms, but to determine whether different language models show differentiated structural tendencies in academic Spanish.

Within the Spanish-language tradition, research on syntactic complexity has moved beyond purely quantitative approaches toward more differentiated models of analysis. Crespo Allende et al. (2011) identify traditionalist, generativist, and functional-discursive perspectives, showing that the meaning of syntactic complexity changes depending on whether the analyst prioritizes length, embedding, transformation, or discourse organization. This observation is methodologically important for the present study because it cautions against treating coordination, subordination, and related dependency relations as interchangeable indicators of a single construct. Instead, these relations should be understood as distinct structural resources that may reveal different ways of organizing academic prose.

Research on university writing in Spanish supports this view. Muse and Delicia (2013), in a study of expository-argumentative texts written by university students, show that syntactic complexity is associated with grammatical competence, but that the most informative indicators are those linked to interclausal integration, particularly clauses per T-unit and embedding intensity. Their findings suggest that recursive and subordinate organization may be especially relevant for characterizing academic writing development. At the same time, Beers and Nagy (2009) demonstrate that the relationship between syntactic complexity and writing quality is both genre-dependent and measure-dependent. In their study, some measures were positively associated with writing quality in essays, while others were not. This reinforces a central caution for the present article: higher frequencies of coordination or subordination in AI-generated academic Spanish should not be interpreted as inherently better or worse, but as structurally meaningful only in relation to the discourse demands of academic exposition and argumentation.

A functional-discursive perspective refines this argument further by shifting attention from sentence-level counts to clause combining in discourse. Katzenberger (2004) conceptualizes clause packages as minimal discourse units delimited by syntactic, thematic, and discursive criteria, while Nir and Berman (2010) and Berman and Nir-Sagiv (2009) treat complex syntax as the organization of interclausal relations in discourse, distinguishing alternative architectures of clause linkage. For the purposes of the present study, the main value of this tradition lies in its insistence that coordination and subordination should not be reduced to formal counts alone. They are alternative strategies for structuring information, managing dependency, and advancing rhetorical progression in text (Katzenberger, 2004; Nir & Berman, 2010; Berman & Nir-Sagiv, 2009).

From this perspective, coordination and subordination become theoretically interpretable as different modes of discourse construction. Paratactic relations can support additive progression or looser sequencing of propositions, whereas hypotactic and embedding structures may contribute to specification, qualification, and conceptual hierarchy. In academic Spanish, where explanatory precision and argumentative development are central, these structural options are consequential rather than stylistically neutral. Bustos Gisbert (2018) is particularly relevant here because his account of written discourse syntax emphasizes that the analysis of Spanish writing must address how syntactic relations operate in actual texts, not only which constructions are available in the grammar. This is closely aligned with the present study, which seeks to identify model-specific structural profiles in academic discourse rather than isolated instances of grammatical complexity.

Taken together, this literature supports a precise claim: syntactic structure is an observable component of academic writing quality because it shows how writers link clauses, integrate discourse, and manage information density under genre-specific constraints (Beers & Nagy, 2009; Muse & Delicia, 2013; Crespo Allende et al., 2011). Syntax alone does not determine writing quality, which also depends on

argumentation, coherence, lexical precision, and disciplinary adequacy. It does, however, provide a theoretically grounded basis for analyzing how academic discourse is organized.

This framework connects directly with the study's research questions. It supports treating coordination and subordination as meaningful dimensions of academic discourse, interpreting syntax as a multivariate profile rather than as a single scalar index, and using structural separation across models as evidence of differentiated discourse organization. In this sense, the study contributes not only to research on AI-generated writing, but also to a broader line of inquiry in which syntactic variation is examined as evidence of how academic writing in Spanish is organized.

Computational Profiling of AI-generated Academic Writing in Spanish

The computational analysis of AI-generated writing must be understood within the broader development of contemporary natural language processing. In current large language models, text generation operates through probabilistic prediction based on large-scale training data rather than through explicit, human-readable grammatical rules (Moreno Sandoval, 2024). As a consequence, the internal mechanisms that produce model outputs are not directly interpretable. For writing research, this implies that analysis must rely on observable linguistic traces that can be systematically extracted and compared across texts. Parser-based pipelines therefore provide an important methodological bridge, enabling researchers to move from impressionistic observations about "AI style" toward reproducible measurements of syntactic structure.

In the case of Spanish, however, computational analysis cannot be treated simply as a transfer of methods developed for English. Recent work highlights structural asymmetries in the development of language technologies for Spanish, including uneven data availability, unequal representation of linguistic varieties, and dependence on infrastructures largely shaped by English-language resources. Muñoz-Basols et al. (2024) conceptualize this situation as Digital Linguistic Bias, arguing that language models may reproduce both interlinguistic asymmetries, due to the predominance of English in training architectures, and intralinguistic asymmetries related to the uneven representation of Spanish varieties. These dynamics imply that AI-generated Spanish should not be interpreted as a homogeneous linguistic output, but rather as the product of specific training distributions and representational priorities.

Empirical evidence supports this concern. Kawasaki (2026), in a large-scale analysis of lexical variation across Spanish-speaking countries, shows that language models vary in their ability to represent regional varieties and that such variation cannot be explained solely by differences in data volume. These findings suggest that linguistic regularities observed in model outputs should be interpreted cautiously, not as neutral reflections of a unified language system but as model-dependent realizations shaped by the data and architectures underlying each system. For studies of AI-generated academic writing, this reinforces the need to examine structural patterns rather than relying exclusively on general impressions of fluency.

From a computational-linguistic perspective, the analysis of such patterns is well established. Dunn (2019) demonstrates that multivariate models of syntactic variation can provide robust representations of linguistic differences by modeling structural feature spaces rather than relying solely on lexical information. Importantly, this approach allows classification methods not only to attribute texts to categories but also to characterize the structural similarity and separation between groups. Although Dunn's work addresses linguistic varieties rather than language models, the methodological logic is directly transferable: if a structural feature space captures systematic variation, multivariate modeling can reveal whether different generating systems occupy differentiated regions within that space.

This perspective aligns with the multivariate procedures used in the present study, including dimensionality reduction and discriminant analysis.

At the same time, research on AI-generated text detection indicates that purely forensic approaches provide an incomplete picture. Detection systems are highly sensitive to domain, model family, prompting conditions, and post-editing, and their performance often varies across contexts (Han et al., 2025; Weber-Wulff et al., 2023). Additional work confirms that machine-generated texts can often be distinguished from human writing but also shows that detection reliability differs across models and genres (Chaka, 2023; Cingillioglu, 2023; Corizzo & Leal-Arenas, 2023). Educational studies similarly report measurable linguistic differences between AI-generated and human-generated discourse while emphasizing that such differences vary by model and cannot be reduced to a single universal signal (Durak et al., 2025). Together, these findings suggest that computational analysis should move beyond the binary question of whether a text is AI-generated and instead examine whether different systems produce identifiable structural profiles.

This shift from detection to profiling is particularly relevant for Spanish, where dedicated datasets for synthetic text research remain limited. García-Campos et al. (2026) highlight this limitation and propose controlled generation pipelines that ensure traceability between prompts, generation settings, and textual outputs. Such methodological control is essential because structural differences between models can only be interpreted reliably when prompts, generation conditions, and preprocessing decisions are standardized.

Within this framework, the use of dependency-based syntactic analysis provides a linguistically interpretable approach to computational profiling. The Universal Dependencies framework offers a cross-linguistically consistent representation of grammatical relations, allowing clause linkage and coordination structures to be operationalized through clearly defined dependency relations. Relations such as cc, conj, ccomp, xcomp, advcl, and acl:relcl therefore provide measurable indicators of how propositions are linked and structured in discourse. Because these relations can be normalized, aggregated, and analyzed in multivariate space, they allow the structural tendencies of different models to be examined systematically.

From this perspective, computational analysis is not merely a technical procedure applied after textual generation. Rather, it constitutes the methodological mechanism through which syntactic structure, previously established as a relevant dimension of academic writing, can be empirically modeled in AI-generated discourse. By combining dependency-based metrics with multivariate statistical analysis, the present study seeks to determine whether ChatGPT, Gemini, and Claude exhibit differentiated structural profiles in academic Spanish and whether these profiles can be statistically identified and geometrically separated within a shared syntactic space.

Method

A comparative intra-task design was implemented to identify differentiated syntactic profiles in Spanish academic writing generated by large language models. The experiment followed a controlled text generation approach in which prompts, supporting materials, and production conditions were standardized to ensure that observed variation could be attributed to the generating model rather than to differences in input, editing, or conversational context. From a language assessment perspective, this design allows syntactic variation to be examined as controlled textual evidence rather than as an artefact of task drift or unequal input. Such control is recommended in experimental studies of LLM-generated texts to ensure the interpretability of linguistic differences across systems (Chang et al., 2024; de Wynter et al., 2023).

The following experimental conditions were established. First, an identical prompt was used across the 90 generations, including explicit instructions regarding academic register, integration of cited authors, and the exclusion of meta-commentary. Second, the same supporting material was supplied to each model: a common source text containing the academic content to be integrated in the response, the same bibliographic-author references, and the same task instructions. Third, each generation was produced in an independent session, ensuring that no conversational history influenced subsequent outputs. Fourth, default system parameters were maintained for each model without manual adjustment of generation settings such as temperature or top-p. Fifth, all texts were generated using the publicly available versions of the models during the data collection period in February 2026. Finally, no human intervention was applied after generation, and the texts were analyzed in their original form without editing or rewriting.

For temporal reproducibility, the models are identified by their public-facing system names and the data collection period: ChatGPT, Gemini, and Claude, accessed through their publicly available interfaces in February 2026. Because public interfaces do not always expose stable backend identifiers or generation parameters to end users, the exact internal model identifiers could not be independently verified from the stored outputs. Accordingly, the findings should be interpreted as temporally situated evidence for the public versions available at the time of generation rather than as claims about all later releases of these systems.

These controls minimized common confounding factors in studies of LLM outputs, such as contextual carryover across turns or prompt drift, thereby enabling a reliable comparison of syntactic configurations across models.

Corpus and Unit of Analysis

The corpus consisted of N = 90 texts, with 30 generated by each of the three models analyzed: ChatGPT, Gemini, and Claude. The unit of analysis was the complete text produced in each generation, such that one model response corresponded to one observation. Each output was assigned a unique identifier (model_interaction_01–30) to preserve traceability across models and generation instances.

All responses were stored in TXT format without post-processing modifications, preserving the original segmentation of the generated outputs. This procedure ensured consistency during preprocessing and facilitated direct integration with the computational analysis pipeline.

To enable cross-model comparability, syntactic metrics were normalized by text length, using rates calculated per 1,000 tokens for dependency relations. Normalization controlled for variation in token counts across generated texts and allowed more reliable comparison of grammatical distributions between models.

The corpus was deliberately restricted to AI-generated texts. Consequently, the inferential target is inter-model variance among LLM outputs, not the detection of AI writing against human-produced Spanish academic texts. This boundary was maintained to avoid conflating model attribution with human-AI discrimination and to keep the validity argument aligned with the controlled comparative design.

Annotation Framework, Extraction Procedure, and Validation

The structural characterization of the corpus followed the Universal Dependencies (UD) framework, a typologically consistent scheme for syntactic dependency annotation widely used in computational

linguistics and multilingual comparative research (de Marneffe et al., 2021). UD provides standardized labels for grammatical relations, allowing coordination and subordination to be operationalized through comparable dependency structures across languages.

All preprocessing and extraction procedures were implemented in Python 3.14 using the Stanza natural language processing library (Qi et al., 2020). Stanza provides neural models trained on UD treebanks for tokenization, sentence segmentation, part-of-speech tagging, and dependency parsing. The official Spanish UD-compatible model distributed with Stanza was used with default parameters to ensure homogeneous processing conditions across all generated texts.

The computational pipeline consisted of four steps. First, each generated text stored in TXT format was automatically loaded. Second, the Stanza pipeline segmented the texts into sentences and tokens and produced dependency parses. Third, the selected dependency relations were extracted programmatically from the parsed output. Fourth, their frequencies were computed for each text and exported to a structured dataset in CSV format, where each row represented a text and each column a syntactic variable. All procedures were implemented in a single reproducible Python script, ensuring traceability from raw outputs to the statistical dataset.

The variables analyzed consisted of normalized rates per 1,000 tokens for selected UD dependency relations associated with coordination and subordination. The subordinating relations included ccomp, xcomp, advcl, and acl:relcl, while coordination was represented by cc and conj. In addition, mean sentence length was included as a structural control variable, operationalized as the ratio between tokens and sentences (tokens_per_sentence), to account for potential differences in verbosity across models.

The reliability of the automatic annotation was evaluated through manual validation on a stratified subset of 30 texts (10 per model). For each text, all occurrences identified by the parser for the six dependency relations were manually reviewed. Interrater agreement was calculated using Cohen's kappa, yielding an overall coefficient of $\kappa = .84$, indicating substantial agreement. Coordination relations (cc and conj) showed the highest reliability ($\kappa > .90$), while adverbial subordinate clauses (advcl) and relative clauses (acl:relcl) presented slightly lower but still substantial agreement levels ($\kappa = .76–.79$).

Parser performance ranged from .84 for advcl to .95 for cc, with an overall precision of .89. These results support the operational reliability of the automatic procedure for identifying aggregated syntactic patterns. Because the objective of the study is the multivariate modeling of structural tendencies across texts rather than exhaustive manual annotation, this level of agreement was considered adequate for subsequent statistical analysis.

Data Analysis

The statistical analysis was organized at three complementary levels: univariate, multivariate, and geometric. At the univariate level, Welch's ANOVA was applied to test differences between models under heterogeneity of variances (Welch, 1951). When distributional assumptions were not satisfied, Kruskal–Wallis tests were additionally conducted. Multiple comparisons were adjusted using the Holm sequential correction (Holm, 1979). Effect sizes were estimated using omega squared and Hedges' g (Hedges, 1981).

At the multivariate level, MANOVA was used to evaluate global differences across the set of syntactic variables. To complement this parametric test, a PERMANOVA with 10,000 permutations was performed to assess the robustness of group differences under fewer distributional assumptions

(Anderson, 2001). The multivariate effect size was estimated using R^2. To examine the structure of the syntactic feature space, principal component analysis (PCA) was applied to standardized variables, enabling dimensionality reduction and visualization of model distributions. Linear discriminant analysis (LDA) was then used to evaluate the capacity to classify the generating model based on its syntactic profile. Classification performance was assessed using stratified fivefold cross-validation.

Finally, Mahalanobis distances between group centroids were calculated to quantify geometric separation among models. Confidence intervals were estimated through nonparametric bootstrap resampling (Efron, 1979). All analyses were conducted in a reproducible computational environment, where the workflow automated the generation of statistical tables and graphical outputs, ensuring analytical transparency and facilitating full replication of the procedures.

Results

To examine whether the models differed in the distribution of the syntactic relations analyzed, Welch's ANOVA was applied to each variable. The analysis was complemented with Kruskal–Wallis tests and the estimation of effect sizes (ω^2). As shown in Table 1, the results indicate statistically significant differences across all syntactic variables ($p < .001$). Effect sizes ranged from moderate to large, indicating substantial variation in the frequency of the analyzed dependency relations across models.

In addition to the dependency-based variables, mean sentence length was analyzed as a structural control variable to determine whether differences in syntactic relations could be explained by variation in sentence expansion across models (Table 2).

Mean sentence length differed significantly across models (Welch $F = 17.25$, $p < .001$, $\omega^2 = .235$). ChatGPT produced longer sentences on average ($M = 35.93$, $SD = 2.65$) than Claude ($M = 32.42$, $SD = 3.60$) and Gemini ($M = 32.37$, $SD = 2.30$). However, this effect is smaller than those observed for the coordination variables *cc* and *conj*, which show the largest effect sizes in the univariate analysis.

Table 1. *Univariate Welch tests and effect sizes (ω^2)*

Variable (UD)	Welch F	p (Welch)	H (Kruskal)	p (Kruskal)	ω^2
cc_per1000	48.32	< .001	39.41	< .001	.52
conj_per1000	44.87	< .001	36.02	< .001	.49
ccomp_per1000	22.61	< .001	19.87	< .001	.31
xcomp_per1000	15.34	< .001	13.92	< .001	.24
advcl_per1000	18.92	< .001	16.77	< .001	.27
acl_relcl_per1000	17.45	< .001	15.64	< .001	.25

Note. ω^2 represents the approximate effect size derived from classical ANOVA. Values greater than .14 are typically interpreted as large effects in social science research.

The variables correspond to dependency relations defined in the Universal Dependencies framework (de Marneffe et al., 2021). The relation cc refers to the coordination marker, typically realized as coordinating conjunctions. The relation conj indicates the second coordinated element dependent on a coordinated structure. The relation ccomp represents a clausal complement with a finite verb. The relation xcomp corresponds to a nonfinite clausal complement without its own explicit subject. The relation advcl identifies adverbial subordinate clauses, including causal, conditional, or concessive constructions. Finally, acl:relcl refers to relative clauses functioning as adjectival modifiers of a noun.

All metrics were normalized per 1,000 tokens in order to control for differences in text length across generated outputs.

Table 2. *Mean sentence length across models*

Model	Mean tokens per sentence	SD
ChatGPT	35.93	2.65
Claude	32.42	3.60
Gemini	32.37	2.30

Note. Mean sentence length corresponds to the ratio between tokens and sentences (tokens_per_sentence) automatically extracted during the parsing pipeline.

Overall, the results indicate systematic differences in the distribution of both coordination and subordination relations across models. Coordination structures (cc, conj) show the largest effect sizes ($\omega^2 \approx .50$), whereas subordination relations (ccomp, advcl, acl:relcl, xcomp) present moderate effects ($\omega^2 = .24–.31$). The convergence between the Welch and Kruskal–Wallis tests further supports the robustness of these findings. Research Question 1 can therefore be answered affirmatively: statistically significant differences exist in the distribution of syntactic structures across the analyzed models.

The second research question examined whether these univariate differences form coherent multivariate syntactic profiles. To address this question, a MANOVA and a complementary PERMANOVA were conducted to evaluate the global effect of the model factor on the set of normalized syntactic variables. Table 3 reports the results of the multivariate analysis of variance examining the global effect of the generating model on the set of syntactic variables.

Table 3. *MANOVA (Pillai's Trace) for the set of syntactic variables*

Effect	Pillai's Trace	F	df1	df2	p
Modelo	0.78	6.94	12	164	< .001

Note. Dependent variables: ccomp, xcomp, advcl, acl:relcl, cc, conj, normalized as rates per 1,000 tokens.

The MANOVA results indicate a significant multivariate effect of the generating model (Pillai's Trace = .78, $p < .001$). This result suggests that the differences observed in the univariate analyses emerge as a coherent structural pattern across the joint syntactic feature space.

To complement the parametric analysis, a PERMANOVA based on Euclidean distances in the standardized syntactic space was conducted using 10,000 permutations. This procedure provides an additional test of the robustness of the multivariate effect. Table 4 reports the results of the permutation-based multivariate analysis.

Table 4. *PERMANOVA for the multivariate syntactic space*

Observed F	p_perm	R^2	Permutations
25.40	< .001	.369	10,000

The PERMANOVA results confirm a significant multivariate effect of the generating model ($F = 25.40$, $p\_perm < .001$). The effect size ($R^2 = .369$) indicates that the model accounts for 36.9% of the variance in the joint syntactic space. The convergence between MANOVA and PERMANOVA supports the robustness of the multivariate differences identified.

To examine the geometric organization of the syntactic feature space, a principal component analysis (PCA) was conducted on the standardized variables (Figure 1).

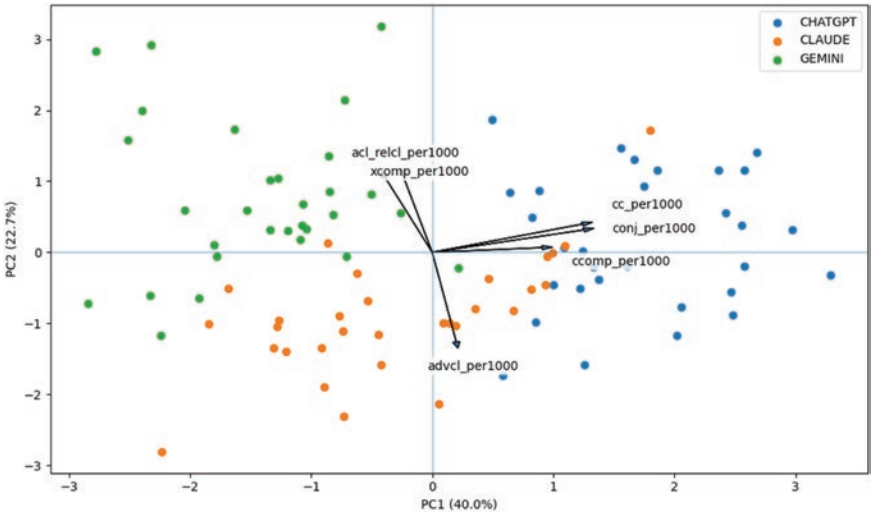

Figure 1. *Biplot of the principal component analysis (PCA) based on normalized syntactic variables.*

Note. Points represent individual texts projected in the PC1–PC2 space, which explains 62.7 percent of the total variance. Arrows indicate the loadings of the Universal Dependencies relations ccomp, xcomp, advcl, acl:relcl, cc, and conj. All variables were standardized prior to the analysis.

The PCA reveals a clear structural differentiation across models. The first principal component (PC1 = 40.0%) is primarily associated with coordination density, whereas the second component (PC2 = 22.7%) differentiates types of subordination. The distribution of texts across the PCA space indicates that the models occupy distinct regions within the syntactic feature space, consistent with the multivariate results reported above. Research Question 2 can therefore be answered affirmatively: the observed differences extend beyond isolated variables and form coherent multivariate syntactic profiles.

The third research question examined whether these profiles allow automatic discrimination of the generating model. To address this question, linear discriminant analysis (LDA) was applied to the standardized syntactic variables using stratified fivefold cross-validation. Table 5 reports the classification performance of the linear discriminant analysis (LDA) model.

Table 5. *Classification accuracy (LDA, fivefold cross-validation)*

Metric	Value
Overall accuracy	0.911
Folds	5
Validation method	Stratified

Note. Predictor variables: ccomp, xcomp, advcl, acl:relcl, cc, and conj, standardized prior to analysis.

The LDA model achieved an overall classification accuracy of 91.1%, indicating that the syntactic profile correctly identifies the generating model in approximately nine out of ten cases. This performance substantially exceeds the expected accuracy under random classification in a three-class problem (~33%). Table 6 presents the confusion matrix aggregated across the cross-validation folds.

Table 6. *Confusion matrix, aggregated across cross-validation*

	Pred. ChatGPT	Pred. Claude	Pred. Gemini
ChatGPT	27	2	1
Claude	3	26	1
Gemini	1	2	27

Note. Each row represents the true class and each column the predicted class.

The confusion matrix shows a high rate of correct classifications across the three models. Misclassifications occur primarily between ChatGPT and Claude, whereas Gemini exhibits a more distinctive classification pattern, consistent with its greater separation in the multivariate syntactic space.

To visualize the discriminant structure identified by the LDA model, texts were projected into the space defined by the linear discriminant functions (Figure 2).

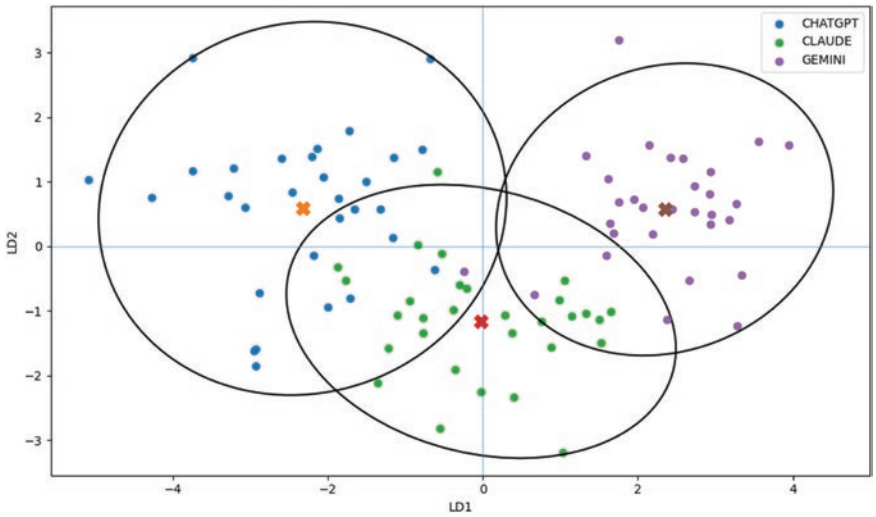

Figure 2. *LDA projection with centroids and 95% confidence ellipses.*

Note. Each point represents an individual text projected in the discriminant space defined by LD1 and LD2. The X markers indicate the centroids of each model. The ellipses correspond to approximate 95 percent confidence regions based on within-model covariance.

The projection shows a clear separation between model centroids, with partial overlaps consistent with the classification errors observed in the confusion matrix. The first discriminant axis (LD1) is primarily associated with coordination features, whereas LD2 captures residual variation related to subordination. These results indicate that the syntactic feature set allows reliable automatic discrimination of the generating model.

The fourth research question examined whether this differentiation corresponds to a stable geometric separation in the multivariate syntactic space. To address this question, Mahalanobis distances between model centroids were calculated with confidence intervals estimated through bootstrap resampling (5,000 iterations). Table 7 reports the Mahalanobis distances between the syntactic centroids of the three models.

Table 7. *Mahalanobis distances between syntactic centroids, bootstrap 5,000*

Model 1	Model 2	Distance	CI 95% (2.5–97.5)
ChatGPT	Claude	1.73	[1.47, 2.17]
ChatGPT	Gemini	2.17	[2.01, 2.48]
Claude	Gemini	1.74	[1.50, 2.12]

Note. Mahalanobis distances between syntactic centroids, bootstrap 5,000.

All pairwise distances are clearly greater than zero, and their confidence intervals do not include values near the absence of separation. This indicates that the syntactic centroids of the models are stably differentiated within the multivariate feature space. The largest distance corresponds to the ChatGPT–Gemini pair (D = 2.17), whereas the distances involving Claude show slightly smaller but comparable magnitudes. Figure 3 provides a visual representation of these distances.

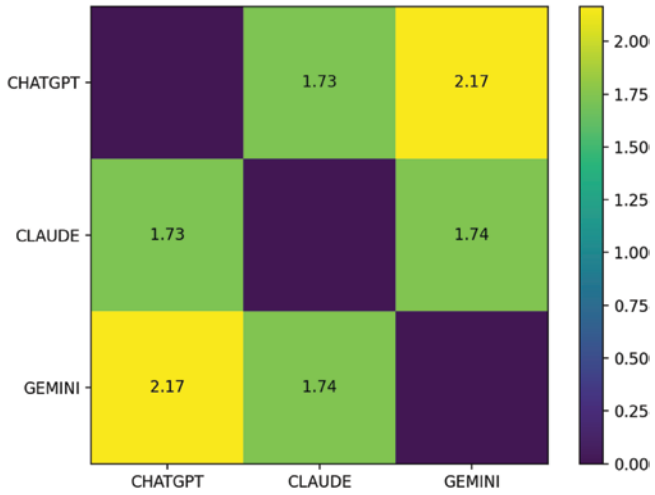

Figure 3. *Heatmap of Mahalanobis distances between syntactic centroids.*

The heatmap highlights the relative magnitude of the distances and facilitates direct comparison across model pairs. Consistent with the previous analyses, Gemini shows the greatest separation from the other systems, whereas ChatGPT and Claude appear relatively closer in the syntactic space. These results confirm that the models occupy distinct regions within the multivariate syntactic space. Accordingly, Research Question 4 can be answered affirmatively: the observed differences correspond to a stable geometric separation of syntactic profiles across models.

Discussion

The present study examined whether ChatGPT, Gemini, and Claude generate distinguishable syntactic patterns in academic Spanish and whether those differences form statistically robust structural profiles. The results support all four research questions. All dependency relations differed significantly across models, and the strongest effects appeared in coordination. These differences also formed integrated multivariate profiles, allowed high classification accuracy, and produced stable geometric separation among model centroids, particularly between ChatGPT and Gemini. In assessment terms, the central

implication is clear: AI-generated academic Spanish is not structurally interchangeable. Model-specific syntactic preferences may therefore form part of the evidence considered when academic writing is evaluated, attributed, or interpreted in AI-mediated contexts.

The most substantive finding concerns the role of coordination as a principal axis of differentiation among models. Within the framework developed in the theoretical section, this result is particularly meaningful. Research on syntactic complexity in Spanish has consistently emphasized that structural analysis should focus on how writers organize information and establish interclausal relations rather than simply accumulating formal complexity (Beers & Nagy, 2009; Bustos Gisbert, 2018; Crespo Allende et al., 2011; Muse & Delicia, 2013). From this perspective, the strong effects observed for cc and conj suggest that the primary differences between models lie in how propositions are linked and developed. In practical terms, the models appear to diverge in their preferred strategies for additive progression and paratactic expansion.

This interpretation aligns with the clause-packaging tradition developed by Katzenberger (2004), Nir and Berman (2010), and Berman and Nir-Sagiv (2009), which conceptualizes syntax not only as sentence-level form but as a discourse-organizing mechanism through which propositions are hierarchically structured and advanced across text. Within this framework, coordination and subordination represent distinct strategies of discourse organization rather than interchangeable measures of complexity. The present results support this view: the models do not simply differ in the amount of syntactic complexity they produce but in the distribution of paratactic and hypotactic structures within the same academic register. One system appears to favor more additive coordination, whereas the others show lower paratactic density combined with different distributions of subordinate constructions. This distinction is especially relevant in academic Spanish, where argumentative development and conceptual specification depend on the structured linking of propositions rather than on sentence length alone.

The inclusion of mean sentence length further clarifies this interpretation. Although sentence length differed significantly across models, its effect size was substantially smaller than those observed for coordination relations. This indicates that the structural differentiation identified cannot be reduced to verbosity or sentence expansion alone. Instead, the primary differences emerge from the internal organization of propositions within sentences. From a methodological perspective, this finding strengthens the interpretation that the observed variation reflects genuine syntactic structuring rather than superficial textual length effects.

Subordination also contributes meaningfully to model differentiation. The significant effects observed for ccomp, xcomp, advcl, and acl:relcl indicate that hypotactic organization remains an important dimension of variation, even if it is less discriminative than coordination. This pattern is consistent with the Spanish-language literature reviewed in the theoretical framework. Crespo Allende et al. (2011) and Muse and Delicia (2013) highlight the multidimensional nature of syntactic complexity in academic discourse, while Beers and Nagy (2009) demonstrate that the relationship between complexity and writing quality varies by genre and metric. The present findings support this perspective. Rather than reflecting a simple continuum of complexity, the models instantiate different balances between paratactic and hypotactic organization within the same discourse domain.

The multivariate analyses reinforce this interpretation by demonstrating that these differences form coherent structural profiles. The combined evidence from MANOVA, PERMANOVA, PCA, and Mahalanobis distances indicates that the models occupy differentiated regions within the same syntactic feature space. This approach addresses a common limitation in the literature on AI-generated writing, where heterogeneous feature inventories are often analyzed without demonstrating whether they jointly constitute meaningful structural configurations. By treating syntax as a multivariate profile

rather than as a list of independent features, the present study aligns with Dunn's (2019) argument that structural modeling can reveal systematic group differences and support robust classification. The triangular configuration observed in the syntactic space further suggests that the models do not lie along a single continuum of elaboration but instead exhibit distinct orientations within the same academic discourse domain.

These findings are also consistent with broader research on Spanish-language language models. Moreno Sandoval (2024) emphasizes that contemporary LLMs generate text through probabilistic prediction rather than explicit grammatical rules, making output profiling essential for understanding how linguistic structures emerge in model-generated text. Moreover, Muñoz-Basols et al. (2024) show that Spanish-language LLM development is shaped by digital linguistic bias, including uneven representation of linguistic varieties and dependence on English-centered infrastructures. Kawasaki (2026) similarly demonstrates that LLMs capture Spanish variation unevenly across the Spanish-speaking world. Although the present study focuses on academic register rather than dialectal variation, these findings help explain why model-specific structural tendencies should emerge in Spanish-language outputs. Differences in training data composition and representational priorities are likely to produce distinct structural preferences across models.

Research Question 3 is addressed by the classification results. The LDA model achieved an accuracy of 91.1%, demonstrating that a relatively small set of interpretable dependency relations contains sufficient information to identify the generating system in most cases. This finding highlights the analytic potential of syntactic profiling as a method for model attribution. However, the implications of this result should be interpreted cautiously. The present evidence demonstrates explainable structural profiling within a controlled corpus rather than a universal detection tool. This position is consistent with the broader detection literature, which shows that AI text detection performance varies across domains, prompts, models, and post-editing conditions (Chaka, 2023; Cingillioglu, 2023; Corizzo & Leal-Arenas, 2023; Han et al., 2025; Weber-Wulff et al., 2023). Accordingly, the contribution of this study lies less in forensic detection than in linguistically interpretable model characterization.

Implications for Language Assessment

These findings have direct implications for language assessment. First, syntactic profiles can complement, but not replace, human judgment in academic-integrity contexts. The 91.1% LDA accuracy indicates that dependency relations contain attribution-relevant information under controlled conditions; however, this evidence should be treated as probabilistic and context-bound, especially when texts are post-edited, translated, or produced with different prompts. This cautious stance is consistent with recent Language Education & Assessment research, which treats AI-supported assessment and feedback as tools that must be interpreted alongside human judgment, task criteria, and pedagogical context (Es-sarghini & Boumahdi, 2026; Jitpaisarnwattana, 2026). Second, the features analyzed here may inform rubric refinement for academic Spanish by making explicit how coordination, subordination, and clause linkage contribute to the construct of academic writing. This claim also aligns with evidence that automated textual features can help model L2 writing quality and classify writing tasks (Tywoniw & Crossley, 2019). Third, the same features can be incorporated into transparent AES/AWE systems as explainable indicators, avoiding exclusive dependence on opaque lexical or neural scores.

At the same time, the results should not be interpreted as indicating that AI-generated academic Spanish is deficient simply because it is differentiable. A more balanced conclusion is that LLMs produce academically plausible Spanish through identifiable structural preferences. Because the present study does not compare model outputs with a matched human corpus, its strongest claim concerns inter-model differentiation rather than comparisons with human writing quality or universal AI-text

detection. In addition, the model comparison is temporally situated in February 2026 and should be revisited as ChatGPT, Gemini, and Claude are updated. This interpretive restraint strengthens the contribution of the study by aligning its claims with the scope of the design.

Overall, the findings demonstrate that academic Spanish generated by ChatGPT, Gemini, and Claude contains statistically significant, multivariately stable, and computationally discriminable structural signatures. The main contribution of the study lies in combining linguistically interpretable syntactic analysis with robust multivariate modeling in a Spanish-language context that remains underrepresented in AI writing research. Future work should examine whether these structural profiles remain stable across prompt variation, human post-editing, different academic genres, and direct comparisons between human and AI-generated academic writing in Spanish. In this sense, syntactic profiling should be understood not primarily as a detection mechanism but as a productive framework for analyzing how different LLM architectures instantiate academic discourse and how such instantiations may be used responsibly in language assessment.

Conclusion

This study examined whether large language models generate distinguishable syntactic configurations in academic Spanish by comparing patterns of coordination, subordination, and sentence length across ChatGPT, Gemini, and Claude. The results revealed systematic differences in the distribution of all analyzed dependency relations. Coordination structures (cc, conj) showed the largest effect sizes, indicating that the principal axis of differentiation among models lies in the use of paratactic linking strategies. Subordination relations (ccomp, xcomp, advcl, acl:relcl) also varied significantly, although with more moderate effects, suggesting differences in how models structure hypotactic relations within academic discourse. Mean sentence length differed across systems but with a smaller effect size, indicating that the observed variation cannot be explained by verbosity alone but primarily reflects differences in syntactic organization. These structural patterns formed coherent multivariate profiles, enabling model identification with high classification accuracy and producing stable geometric separation within the syntactic feature space.

Beyond these empirical findings, the study demonstrates that syntactic profiling offers a linguistically interpretable approach for analyzing AI-generated academic writing in language assessment contexts. Rather than relying solely on surface fluency or generic detection tools, examining clause linkage and syntactic organization reveals how different LLM architectures instantiate academic discourse through identifiable structural preferences. These features may support more transparent assessment practices when AI-assisted academic writing is considered as evidence of linguistic performance, authorship, or task completion.

From an educational and assessment perspective, the findings suggest that the evaluation of AI-assisted academic writing should consider structural features in addition to surface textual adequacy. For rubric design, coordination and subordination metrics can help operationalize dimensions of syntactic control in academic Spanish. For AES/AWE development, dependency-based features can provide explainable indicators that complement global scores. For academic integrity, syntactic profiles may contribute to cautious model-attribution procedures under controlled conditions, while remaining insufficient as a stand-alone basis for high-stakes decisions. Future research should examine the stability of these syntactic profiles across prompts, genres, post-editing conditions, and matched human-written corpora.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This work has been developed within the framework of the Teaching Innovation Project at UNED entitled: "Small Language Model Architecture in Educational Chatbots: Development of an Automated Curricular Tool and Its Effect on the Academic Performance of University Students." (Resolution of the Vice-Rectorate for Digitalization and Innovation, UNED. BICI, No. 1, 30 September 2025).

Acknowledgements

Not applicable.

Data Availability Statement

The de-identified dataset supporting this study has been structured following the FAIR data principles and deposited in the Harvard Dataverse repository: https://doi.org/10.7910/DVN/ZYIISQ. The dataset includes the syntactic metrics extracted from the corpus of generated texts, the data dictionary describing all variables, and the analysis scripts used to reproduce the statistical procedures reported in this study.

Ethics Statement

This study did not require ethical approval as it did not involve sensitive personal data or procedures that posed ethical risks to participants.

Author Contributions Statement

The author solely conceived and designed the study, obtained the research funding, and conducted all phases of the research process. This included the development of the theoretical framework, study design, data collection, methodological procedures, and statistical analyses. The author also performed data curation, ensuring compliance with FAIR principles, and prepared all visualisations. The manuscript was written, reviewed, and edited exclusively by the author.

Generative Artificial Intelligence (AI)

The author declares that generative artificial intelligence was used in the preparation of this manuscript. Specifically, ChatGPT (version 5.5) was employed to support the translation and stylistic refinement of selected sections of the text. All substantive content, data analysis, interpretation of results, and final editorial decisions were carried out by the authors, who take full responsibility for the integrity and originality of the work.

References

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology, 26*(1), 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Atasoy, A., & Moslemi Nezhad Arani, S. (2025). ChatGPT: A reliable assistant for the evaluation of students' written texts? *Education and Information Technologies, 30*, 20385–20415. https://doi.org/10.1007/s10639-025-13553-1

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing, 22*, 185–200. https://doi.org/10.1007/s11145-007-9107-5

Berman, R. A., & Katzenberger, I. (2004). Form and function in introducing narrative and expository texts: A developmental perspective. *Discourse Processes, 38*(1), 57–94. https://doi.org/10.1207/s15326950dp3801_3

Berman, R. A., & Nir-Sagiv, B. (2009). Clause-packaging in narratives: A crosslinguistic developmental study. In J. Guo, E. Lieven, S. Ervin-Tripp, N. Budwig, S. Özçalişkan & K. Nakamura (Eds.), Crosslinguistic approaches to the psychology of language: Research in the tradition of Dan I. Slobin (pp. 149–162). Lawrence Erlbaum.

Bustos Gisbert, J. M. (2018). El estudio de las características sintácticas del discurso escrito. *Dicenda. Estudios de Lengua y Literatura Españolas, 36*, 89–114. https://doi.org/10.5209/DICE.62139

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching, 6*(2), 94–104. https://doi.org/10.37074/jalt.2023.6.2.12

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://doi.org/10.1145/3641289

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In Proceedings of the 2014 *Conference on Empirical Methods in Natural Language Processing* (EMNLP) (pp. 740–750). Doha, Qatar. Association for Computational Linguistics.

Cingillioglu, I. (2023). Detecting AI-generated essays: The ChatGPT challenge. International *Journal of Information and Learning Technology, 40*(3), 259–268. https://doi.org/10.1108/IJILT-03-2023-0043

Clementi, L., & Jiménez López, M. D. (2025). Análisis de textos generados por ChatGPT para la enseñanza del español: complejidad y adecuación a niveles de competencia. *Journal of Spanish Language Teaching*, 1–28. https://doi.org/10.1080/23247797.2025.2534203

Corizzo, R., & Leal-Arenas, S. (2023). One-class learning for AI-generated essay detection. *Applied Sciences, 13*(3), 7901. https://doi.org/10.3390/app13137901

Crespo Allende, N., Alfaro Faccio, P., & Góngora Costa, B. (2011). La medición de la sintaxis: evolución de un concepto. *Onomázein, 24*, 155–172. https://doi.org/10.7764/onomazein.24.07.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

de Wynter, A., Wang, X., Sokolov, A., Gu, Q., & Chen, S.-Q. (2023). An evaluation on large language model outputs: Discourse and memorization. *Natural Language Processing Journal*, *4*, 100024. https://doi.org/10.1016/j.nlp.2023.100024

Desaire, H., Chua, A. E., Isom, M., Jarosova, R., & Hua, D. (2023). Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science, 4*, 101426. https://doi.org/10.1016/j.xcrp.2023.101426

Dunn, J. (2019). Global syntactic variation in seven languages: Toward a computational dialectology. *Frontiers in Artificial Intelligence, 2*, Article 15. https://doi.org/10.3389/frai.2019.00015

Durak, H. Y., Eğin, F., & Onan, A. (2025). A comparison of human-written versus AI-generated text in discussions at educational settings: Investigating features for ChatGPT, Gemini and BingAI. *European Journal of Education, 60*, e70014. https://doi.org/10.1111/ejed.70014

Dwivedi, Y. K., Kshetri, N., Hughes, L., et al. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Es-sarghini, A., & Boumahdi, A. (2026). Introducing sentinel assessment: An AI-powered approach for evaluating writing competence in French as a Foreign Language. *Language Education & Assessment, 9*, 103686. https://doi.org/10.29140/lea.v9.103686

García-Campos, J. M., Lara-Romero, A. W., Mayor, V., & Calvillo-Arbizu, J. (2026). Controlled generation of synthetic Spanish texts: A dataset using LLMs with and without contextual retrieval. *Data, 11*(2), 29. https://doi.org/10.3390/data11020029

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202. https://doi.org/10.3758/BF03195564

Graham, S., & Rijlaarsdam, G. (2016). Writing education around the globe: Introduction and call for a new global analysis. *Reading and Writing, 29*, 781–792. https://doi.org/10.1007/s11145-016-9640-1

Han, J., Yang, Y., & Liu, G. (2025). Are teachers assessing work written by students or by AI? A rapid literature review of research on detecting content generated by generative AI. *European Journal of Education, 60*, e70240. https://doi.org/10.1111/ejed.70240

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. http://www.jstor.org/stable/4615733

Jitpaisarnwattana, N. (2026). Generative AI as an IELTS tutor: The effects of AI-generated feedback on EFL students' writing. *Language Education & Assessment, 9*, 104046. https://doi.org/10.29140/lea.2026.104046

Katzenberger, I. (2004). The development of clause packaging in spoken and written texts. *Journal of Pragmatics, 36*, 1921–1948. https://doi.org/10.1016/j.pragma.2004.01.010

Kawasaki, Y. (2026). *Digital linguistic bias in Spanish: Evidence from lexical variation in LLMs* [Preprint]. arXiv:2602.09346. https://arxiv.org/abs/2602.09346

Lo, J., Wong, C., Ng, A., Wong, P., Cheung, D., & Lai, P. (2026). Stretching AI's reach: Assessing an AI-driven feedback system for extended academic writing. *Computers and Education: Artificial Intelligence, 10*, 100511. https://doi.org/10.1016/j.caeai.2025.100511

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*(4), 292–330. https://doi.org/10.1080/01638530902959943

Mizumoto, A., Yasuda, S., & Tamura, Y. (2024). Identifying ChatGPT-generated texts in EFL students' writing through comparative analysis of linguistic fingerprints. *Applied Corpus Linguistics*, *4*(3), 100106. https://doi.org/10.1016/j.acorp.2024.100106

Moreno Sandoval, A. (2024). El español artificial. En *El español en el mundo 2024. Anuario del Instituto Cervantes*. Instituto Cervantes.

Muñoz-Basols, J., Palomares Marín, M. del M., & Moreno Fernández, F. (2024). El sesgo lingüístico digital (SLD) en la inteligencia artificial: Implicaciones para los modelos de lenguaje masivos en español. *Lengua y Sociedad. Revista de Lingüística Teórica y Aplicada, 23*(2), 623–647. https://doi.org/10.15381/lengsoc.v23i2.28665

Muse, C. E., & Delicia, D. D. (2013). Complejidad sintáctica y discurso académico: parámetros gramaticales en la producción escrita de estudiantes universitarios. *Revista Nebrija de Lingüística Aplicada, 14*. https://doi.org/10.26378/rnlael714200

Nir, B., & Berman, R. A. (2010). Complex syntax as a window on contrastive rhetoric. *Journal of Pragmatics, 42*, 744–765. https://doi.org/10.1016/j.pragma.2009.07.006

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Tywoniw, R., & Crossley, S. (2019). The effect of cohesive features in integrated and independent L2 writing quality and text classification. *Language Education & Assessment, 2*(3), 110–134. https://doi.org/10.29140/lea.v2n3.151

Vázquez-Cano, E., Mengual-Andrés, S., & López-Meneses, E. (2021). Chatbot to improve learning punctuation in Spanish and to enhance open and flexible learning environments. International *Journal of Educational Technology in Higher Education, 18*, 33. https://doi.org/10.1186/s41239-021-00269-8

Vázquez-Cano, E., Ramírez-Hurtado, J. M., & Sáez-López, J. M., & López-Meneses, E. (2023). ChatGPT: The brightest student in the class. *Thinking Skills and Creativity, 49*, 101380. https://doi.org/10.1016/j.tsc.2023.101380

Walters, W. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science, 7*(1), 20220158. https://doi.org/10.1515/opis-2022-0158

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltynek, T., Guerrero-Dib, J. G., Popoola, O., et al. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*, 1–39. https://doi.org/10.1007/s40979-023-00146-z

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38*(3–4), 330–336. https://doi.org/10.1093/biomet/38.3-4.330

Yan, D., Fauss, M., Hao, J., & Cui, W. (2023). Detection of AI-generated essays in writing assessment. *Psychological Testing and Assessment Modeling, 65*(2), 125–144.

Zhong, Y., Hao, J., Fauß, M., Li, C., & Wang, Y. (2024). Evaluating AI-generated essays with GRE Analytical Writing assessment. *arXiv* (2410.17439v3). https://arxiv.org/html/2410.17439v3