

LLM-Based Metadiscourse Analysis of Demographic Variations in ELL Hedge Use

Xinyu Xu Tianyu Xu

Teachers College, Columbia University
New York, NY, USA

xx2503@tc.columbia.edu tx2264@tc.columbia.edu

Abstract

Using an LLM-assisted pipeline, this study examines sociodemographic variation in English Language Learners' hedge use across 6,500 essays. Regression analyses are expected to show that gender, race, and SES predict hedge density and category-specific variation, which subsequently predict writing scores. Beyond scaling annotation, the study evaluates whether LLM-derived metadiscourse indicators provide interpretable evidence about rhetorical development in ELL writing.

1 Background

Writing is a fundamental social act requiring students to position themselves in relation to their readers, acknowledge uncertainty, and project authority through rhetorical features known as metadiscourse markers (Bazerman, 2009; Hyland, 2005). Within Hyland (2005)'s interactional framework, hedges (e.g., *perhaps*, *it seems*) express tentativeness, while boosters (e.g., *clearly*, *in fact*) signal certainty. Mastering these epistemic modality markers is crucial for argumentative writing competence, yet it is challenging for English Language Learners (ELLs) to acquire. According to Language Socialization Theory, how a writer manages certainty and interpersonal relationships is heavily shaped by their sociolinguistic background (Ochs and Schieffelin, 2011). This means that demographic variables like grade level, gender, race, and socioeconomic status (SES) can serve as crucial proxies for these diverse rhetorical socialization histories.

2 Aims

Despite their importance, existing literature on metadiscourse markers suffers from several key limitations. Most prior work has relied on small, manually coded datasets and focused almost exclusively on adult, university-level English as a Foreign Language (EFL) learners abroad rather than adolescent ELL students within U.S. schools (Ahmed and Zhang, 2023; El-Dakhs, 2020). Consequently, there remains a critical research gap regarding how sociodemographic factors shape the specific use of epistemic markers within domestic ELL populations, and how these markers subsequently impact automated or human writing evaluations. Because hedges and boosters anchor the two opposing poles of authorial certainty, this

study intentionally isolates hedging, a decision also supported by practical computational constraints, to deliver a highly focused and technically efficient analysis of how language learners acquire nuanced pragmatic skills. Using the large-scale ELLIPSE corpus of argumentative essays written by U.S. students in grades 8–12 (Crossley et al., 2023), this study addresses these gaps through three primary research questions:

- **RQ1:** To what extent do demographic characteristics, including grade level, gender, race, and SES, predict variation in hedge use density among ELL student writers in the ELLIPSE corpus?
- **RQ2:** Are certain demographic groups more likely to use specific types of hedges (Epistemic Verbs, Epistemic Adverbs, Epistemic Adjectives, Epistemic Modal Verbs, and Approximation)?
- **RQ3:** Does hedge density predict overall writing quality and analytic subscale scores after accounting for essay-level linguistic features and student demographic background?

This study offers significant theoretical, empirical, and methodological contributions to the fields of learning analytics and educational linguistics. Methodologically, it advances the field by combining traditional corpus linguistics with an LLM-assisted hedge annotation pipeline and pilot human validation to efficiently identify nuanced, context-specific metadiscourse markers across a large dataset. Empirically, it provides a detailed investigation of how hedge density and hedge-type distributions vary across sociodemographic groups and relate to writing quality. More importantly, the study moves beyond large-scale annotation by examining whether these AI-derived discourse features function as interpretable indicators of rhetorical development rather than simple frequency-based measures.

3 Sample

3.1 Data Source and Variables

Data for this study come from the English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) corpus (Crossley et al., 2023), a publicly available dataset of 6,482 argumentative essays written by ELL students in grades 8 through 12 across the United States. The corpus includes rich demographic metadata (e.g., grade

level, gender, race, and SES), essay-level linguistic features and human-assigned holistic and analytic writing quality scores. Writing quality scores were assigned by trained raters on a scale of 1.0 to 5.0 across seven dimensions: Overall, Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Conventions. Additionally, essay-level linguistic variables include essay length (number of words), Type-Token Ratio (TTR, i.e., a bounded measure of linguistic diversity calculating the ratio of unique words to total words, which naturally decreases as text length increases) and Measure of Textual Lexical Diversity (MTLD, i.e., how diverse a writer's vocabulary is without being biased by the length of the text). The corpus spans grades 8 through 12, with the largest representation at grades 11 and 12. Full distributional details are in Figure 1.

4 Methodology

4.1 Metadiscourse Marker Detection for Hedge

To balance broad candidate coverage with contextual precision, hedge identification was conducted using a two-stage pipeline that combined dictionary-based candidate extraction with LLM validation across 6,482 essays. First, essays were segmented into sentences using NLTK tokenization, and candidate occurrences were flagged via a dictionary baseline drawn from a theoretically grounded taxonomy of five hedge categories: epistemic modal verbs, epistemic verbs, epistemic adverbs, epistemic adjectives, and approximations, including Hyland (2005), Aull and Lancaster (2014), Islam et al. (2020), Biber et al. (1999), Holmes (1988), and Hinkel (2004, 2005), with the full compiled list reported in Appendix A.

Second, because many candidate expressions are multifunctional, each occurrence was evaluated in its sentence context through prompt-based LLM validation. The prompt instructed the model to retain a candidate only when it reduced certainty, precision, or authorial commitment in the writer's own proposition. The model returned structured JSON outputs identifying whether the candidate functioned as a hedge and, if retained, its hedge category.

Prior to full-sample application, a pilot comparison with manual human coding was conducted to refine the prompt and examine whether LLM classifications aligned with the study's operational definition of hedging. This pilot served primarily as a formative validation step rather than a full reliability study; therefore, the resulting annotations should be interpreted as human-informed LLM classifications rather than fully established gold-standard labels. Future work will extend this validation using a larger independently coded sample and formal agreement metrics.

After applying this pipeline, 6,473 of 6,481 essays had at least one candidate, and of the 146,833 candidates, 58,811 were confirmed as actual hedges by the LLM (~40%), from which metrics such as total count and density per 100 words were extracted.

4.2 Statistical Analysis

Three sets of regression models were estimated to address the research questions. For RQ1, ordinary least squares (OLS) regression models were estimated with hedge density as separate outcome variables and demographic characteristics as predictors. Models were built using a stepwise entry approach, with text-level controls including word count, type-token ratio, and MTLD entered prior to demographic predictors, allowing the independent contribution of demographic variables to be assessed after accounting for essay-level linguistic features.

To address RQ2, we used multivariate analysis of variance (MANOVA) to examine whether demographic variables predicted students' hedge-type profiles across the five hedge categories. Category-specific hedge densities served as the multivariate outcome. Because MANOVA evaluates group differences across multiple dependent variables simultaneously, it was appropriate for testing whether demographic factors were associated with different distributions of hedge types rather than simply different total frequencies.

For RQ3, a series of nested OLS regression models were estimated predicting writing quality scores from hedge density respectively. Each marker was first entered alone as a predictor, then with essay-level text controls, and finally with the full set of demographic predictors added. This stepwise structure allows the persistence of the hedge effects across model specifications to be evaluated. Models were estimated separately for the overall holistic score and for each of the six analytic subscales: Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Conventions. All analyses were conducted in Python. Effect sizes are reported as unstandardized regression coefficients with standard errors, and statistical significance is evaluated at $\alpha = 0.05$.

5 Anticipated Results

Preliminary analyses suggest that overall hedge density is not strongly or consistently explained by students' broad demographic characteristics (Figure 2), where demographic factors (grade, gender, SES, race) account for only 0.9% of the total variance in hedge density. Descriptive comparisons showed relatively similar rates of hedging across gender, SES, grade level, and race groups (Figure 3), indicating that the total frequency of hedging may be shaped more by writing proficiency, lexical resources, prompt demands, genre familiarity, or individual rhetorical style than by demographic background alone. This pattern suggests that hedge density should be interpreted cautiously as an AI-derived writing indicator: higher or lower hedge frequency does not necessarily correspond to stronger or weaker writing, nor should it be treated as a simple demographic marker.

However, the distribution of hedge types appears to provide a more sensitive measure of students' rhetorical development. Students relied most heavily on epistemic modal verbs, such as *may*, *might*,

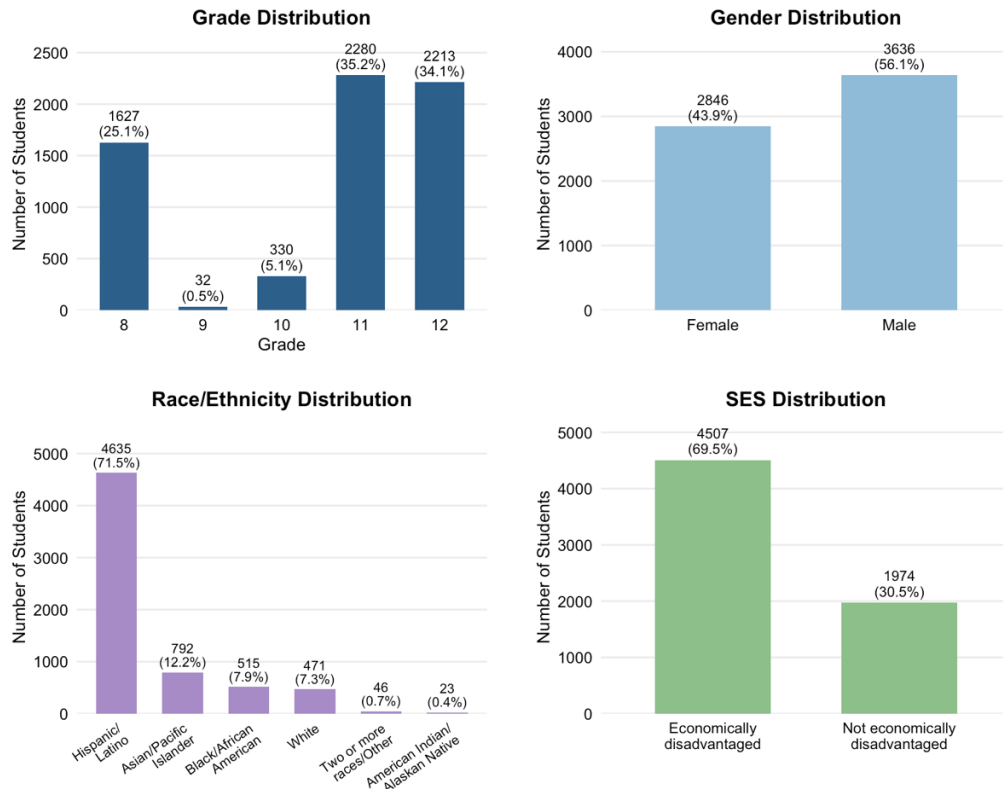


Figure 1: ELLIPSE corpus sample characteristics ($N = 6,482$).

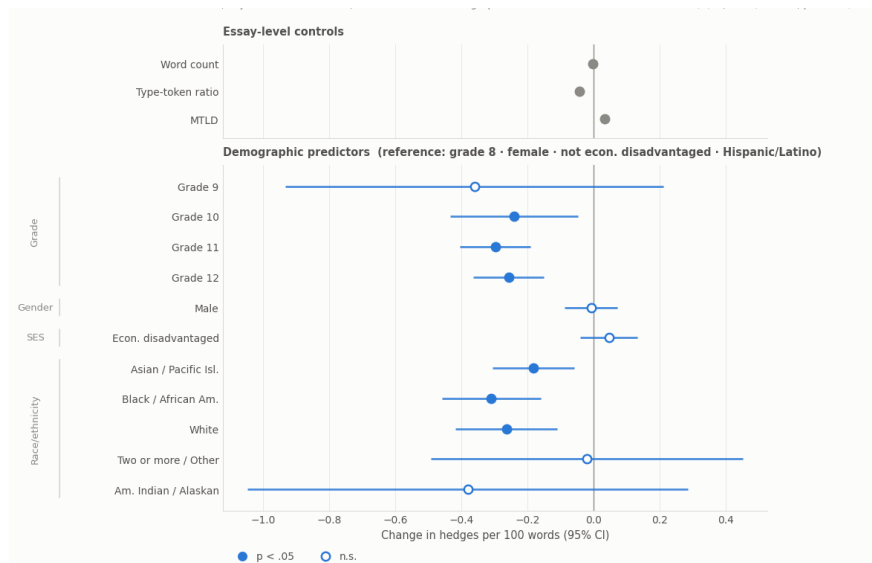


Figure 2: OLS regression coefficients for demographic predictors of overall hedge density (adjusted for text-level controls).

can, *could*, and *would*, and on approximation markers, such as *some*, *many*, *most*, *about*, and *often*, which accounts for 83% of all hedge tokens in total (Figure 4). In contrast, more lexically specific stance markers, including epistemic verbs, epistemic adverbs, and epistemic adjectives, appeared less frequently. This pattern suggests that many ELL writers may have access to basic hedging strategies but may use a relatively narrow repertoire of stance-marking resources.

The MANOVA results further suggest that grade

level may be associated with students' hedge-type profiles even when overall hedge density remains relatively stable. Figure 5 shows that grade level has the largest multivariate effect size (partial $\eta^2 = 0.0087$, $F = 11.30$, $p < 0.001$) compared to all other demographics, though all effects are significant. In other words, developmental differences may appear less as a simple increase or decrease in the amount of hedging and more as a shift in the kinds of hedges students use. Follow-up analyses in Figure 6 point especially to differences in

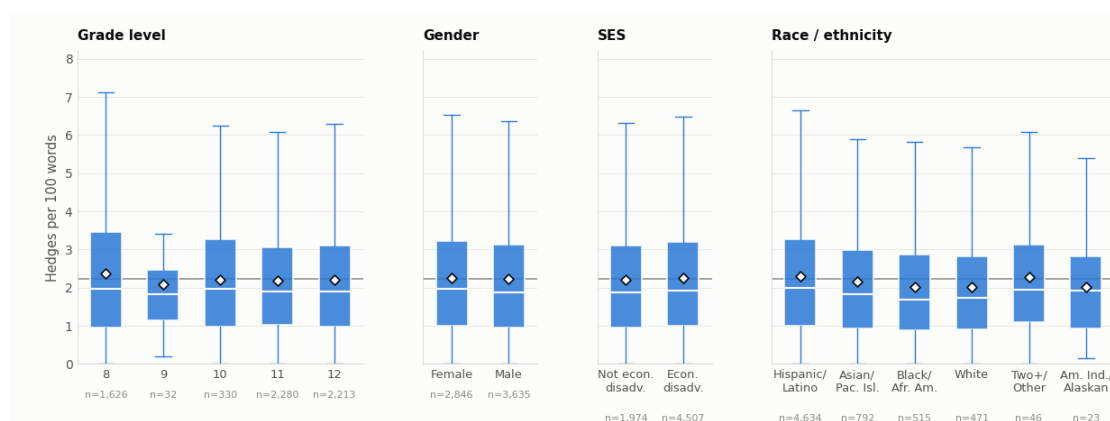


Figure 3: Distribution of total hedge density by group.

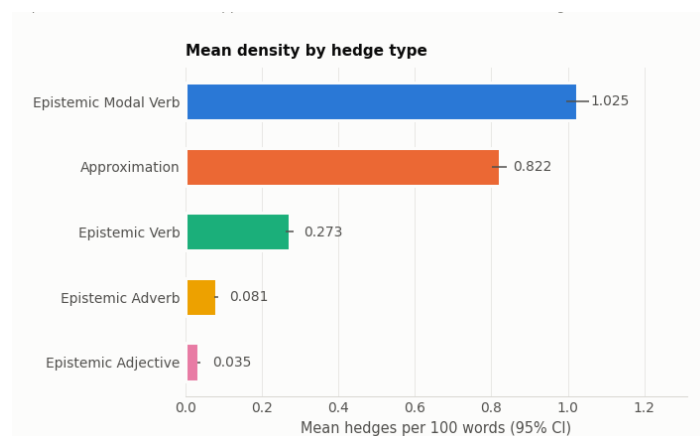


Figure 4: Mean hedges per 100 words by hedge type.

modal verbs and approximation markers, suggesting that students across grade levels may vary in how they balance familiar, high-frequency forms of uncertainty marking with more precise or lexically nuanced forms of epistemic stance.

Findings for gender, SES, and race were substantially weaker and should be interpreted cautiously. While some model-based estimates suggest possible group differences in hedge density or category use, these effects are not consistent across all descriptive and inferential tests (Figures 7 and 8). Rather than treating demographic differences as evidence of deficit, the study interprets them as potential indicators of varied rhetorical socialization, language-learning histories, and access to academic writing resources.

For writing quality, we anticipate that hedge density alone will have limited explanatory power after controlling for essay length, lexical diversity, and demographic background. If hedge measures remain associated with holistic or analytic writing scores, this would provide evidence that AI-derived metadiscourse indicators capture a meaningful aspect of argumentative writing. However, the expected pattern is not that “more hedging is better.” Instead, writing quality is likely to depend on whether students use hedges flexibly and appropriately to qualify claims, align certainty with evidence, and construct a credible academic stance.

Overall, the anticipated results suggest that

hedge-type distribution and hedge variety may be more instructionally and psychometrically useful than raw hedge frequency. For AI-integrated assessment, this means that automated metadiscourse indicators should not be interpreted as simple counts. Instead, they should be evaluated as evidence of students’ developing rhetorical range, pragmatic control, and ability to manage certainty in academic argumentation.

6 Anticipated Conclusions & Contributions

If supported by the full analyses, the findings suggest that epistemic stance should receive more explicit attention in ELL writing instruction. Students may need support not only with vocabulary and grammar, but also with how to express uncertainty, qualification, and evidence-based confidence in academic arguments. Instruction should avoid simply encouraging students to use more or fewer hedges. Instead, teachers can help students select stance markers according to evidence strength and rhetorical purpose. For example, students can compare expressions such as “This proves that,” “This suggests that,” and “This may indicate that” to examine different degrees of certainty.

For assessment and learning analytics, the study suggests that hedge density alone should not be treated as a direct indicator of writing quality. If category-specific patterns or hedge variety remain

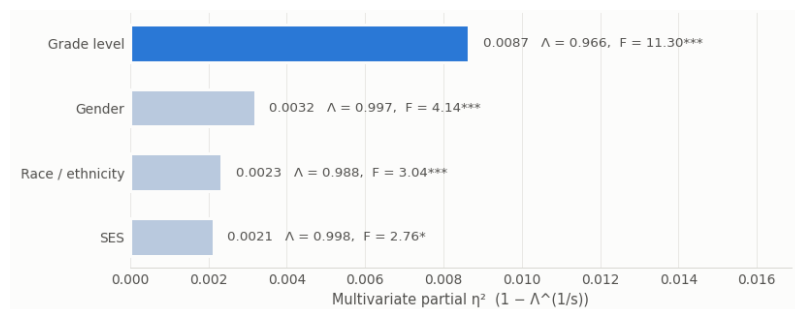


Figure 5: MANOVA on the hedge type. Note that *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

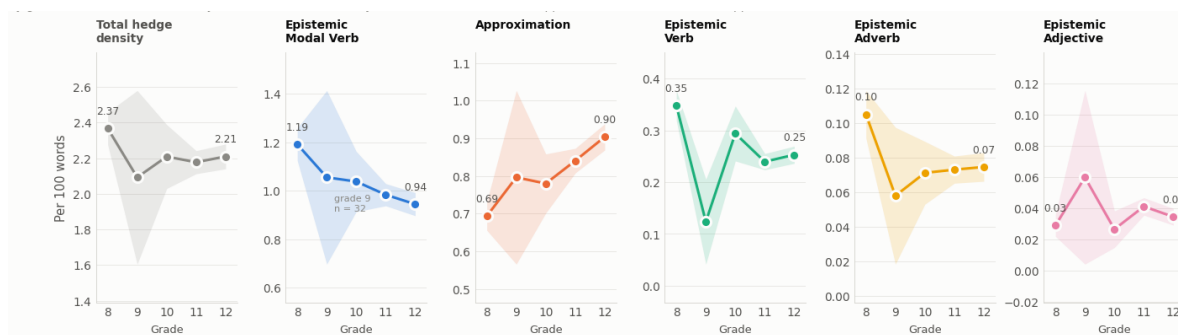


Figure 6: Mean density of total hedges and specific hedge categories across Grades 8–12.

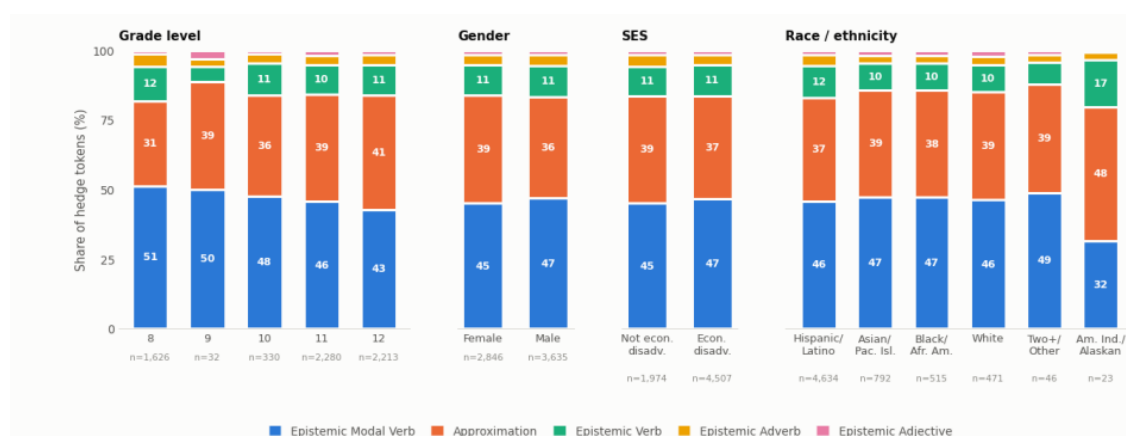


Figure 7: Proportional distribution of hedge categories across demographic subgroups.

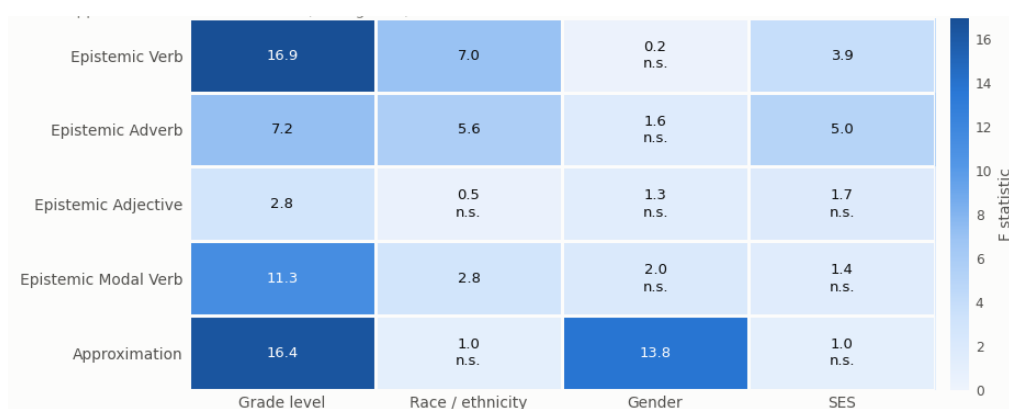


Figure 8: Univariate Type II ANOVA F-statistics for demographic predictors across hedge categories. Darker shading indicates larger test statistics; cells marked n.s. denote non-significant effects ($p \geq 0.05$).

associated with writing performance after linguistic controls are introduced, these features may provide

more interpretable evidence of students' rhetorical repertoire than raw frequency alone. Learning

analytics tools could therefore support formative feedback by identifying overreliance on a limited set of stance markers and encouraging a broader range of context-appropriate expressions.

The demographic analyses should also be interpreted cautiously. Differences across grade, gender, race, or SES do not in themselves demonstrate assessment bias or fairness. Rather, they provide fairness-relevant evidence by showing whether AI-derived discourse indicators vary systematically across student groups and by identifying patterns that may call for further investigation in future validation studies. In this sense, the present study contributes to fairness research by informing how such indicators should be interpreted, rather than by establishing that the assessment system itself is fair or biased.

Limitations

A key limitation of this study is its exclusive focus on hedges rather than examining both hedges and boosters. Because these features represent different forms of stance calibration, future work should examine a broader range of metadiscourse markers to provide a more complete account of how ELL students construct academic stances.

References

- A. M. Ahmed and X. Zhang. 2023. Students' voice in L2 English writing: A systematic review of literature. *Ampersand*, 10:100114. doi:10.1016/j.amper.2023.100114.
- S. Alanazi, A. H. Al-Hoorie, and A. Alghamdi. 2024. The use of interactional metadiscourse markers by Saudi EFL male and female college students: The case of a gender-sensitive topic. *Humanities and Social Sciences Communications*, 11(1):988. doi:10.1057/s41599-024-03506-3.
- L. L. Aull and Z. Lancaster. 2014. Linguistic markers of stance in early and advanced academic writing: A corpus-based comparison. *Written Communication*, 31(2):151–183.
- C. Bazerman. 2009. Genre and cognitive development: Beyond writing to learn. *Pratiques. Linguistique, littérature, didactique*, 143–144:127–138.
- D. Biber. 2004. Historical patterns for the grammatical marking of stance: A cross-register comparison. *Journal of Historical Pragmatics*, 5(1):107–136.
- D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman Grammar of Spoken and Written English*. Longman.
- C. Castillo and N. Ávila Reyes. 2026. Students' sociodemographic characteristics and writing performance: A systematic literature review. *Reading and Writing*, 39(6):1085–1121. doi:10.1007/s11145-025-10658-4.
- S. Crossley, Y. Tian, P. Baffour, A. Franklin, Y. Kim, W. Morris, and U. Boser. 2023. The English language learner insight, proficiency and skills evaluation (ELIPSE) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.
- D. A. S. El-Dakhs. 2020. Why are hedges used differently in research articles and opinion columns? A contrastive genre analysis. *Journal of Pragmatics*, 156:63–75. doi:10.1016/j.pragma.2019.09.009.
- M. Fuoli, W. Huang, J. Littlemore, S. Turner, and E. Wilding. 2025. Metaphor identification using Large Language Models. *arXiv preprint arXiv:2509.24866*.
- E. Hinkel. 2004. *Teaching Academic ESL Writing: Practical Techniques in Vocabulary and Grammar*. Lawrence Erlbaum Associates.
- E. Hinkel. 2005. Hedging, inflating, and persuading in L2 academic writing. *Applied Language Learning*, 15(1/2):29.
- J. Holmes. 1988. Doubt and certainty in ESL textbooks. *Applied Linguistics*, 9(1):20–44.
- K. Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins.
- K. Hyland. 2005. *Metadiscourse: Exploring Interaction in Writing*. Continuum.
- J. Islam, L. Xiao, and R. E. Mercer. 2020. A lexicon-based approach for detecting hedges in informal text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC '20)*, pages 3109–3113. European Language Resources Association.
- J. J. Lee and L. Deakin. 2016. Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of Second Language Writing*, 33:21–34. doi:10.1016/j.jslw.2016.06.004.
- G. Lorenz. 1999. *Adjective Intensification—Learners versus Native Speakers: A Corpus Study of Argumentative Writing*. Rodopi.
- E. Ochs and B. B. Schieffelin. 2011. The theory of language socialization. In *The Handbook of Language Socialization*, pages 1–21.
- S. Park and S. Y. Oh. 2018. Korean EFL learners' metadiscourse use as an index of L2 writing proficiency. *SNU Journal of Education Research*, 27(2):65–89.
- W. S. Pearson and E. Abdollahzadeh. 2023. Metadiscourse in academic writing: A systematic review. *Lingua*, 293:103561.
- P. Uccelli, C. L. Dobbs, and J. Scott. 2013. Mastering academic language: Organization and stance in the persuasive writing of high school students. *Written Communication*, 30(1):36–62.
- M. T. Yeganeh and S. M. Ghoreyshi. 2015. Exploring gender differences in the use of discourse markers in Iranian academic research articles. *Procedia - Social and Behavioral Sciences*, 192:684–689.

Appendix A. Hedge Taxonomy

Hedge Category	Definition	Examples	Research Evidence
Epistemic Verb	A lexical verb that presents a proposition as inferred, believed, assumed, suggested, or only partially certain, rather than fully asserted as fact. Do not count a candidate word as a hedge if it only refers to the content of thinking, believing, or opinion-holding, rather than reducing the writer's own commitment in context. Do not count <i>think</i> as a hedge when it only refers to the content of someone's thoughts or beliefs, rather than softening the writer's own assertion.	suggest, believe, appear, indicate, assume, seem, consider, doubt, estimate, expect, feel, guess, imagine, speculate, suppose, think, understand, imply, presume, suspect, postulate, reckon, infer, hope	Biber treats verbs that control stance/complement constructions as a major way English expresses stance, and Varttala-style hedge taxonomies treat these as a core hedge class.
Epistemic Adverb	An adverb that modifies a clause or proposition to signal likelihood, tentativeness, limited precision, or non-absolute commitment.	rather, slightly, barely, strictly, presumably, fairly, theoretically, basically, relatively, possibly, preferably, slenderly, scantily, decidedly, arguably, seemingly, occasionally, partially, partly, practically, roughly, virtually, allegedly	Adverbials are a major grammatical resource for expressing epistemic meaning; Biber distinguishes stance adverbials as a major type of stance marking.
Epistemic Adjective	An adjective that encodes possibility, probability, uncertainty, or appearance.	presumable, possible, probably, likely, apparent, probable, improbable, unlikely, rarely, improbably, unclearly, unsure, sure, chance, unclear	Adjectives like <i>possible</i> and <i>likely</i> as expressing modal-like meanings, and hedge classifications regularly include epistemic adjectives as a core category.
Epistemic Modal Verb (Modal Auxiliaries)	A modal auxiliary used to indicate possibility, weak prediction, or non-categorical commitment rather than certainty. Count modal verbs as hedges only when they express uncertainty, possibility, or limited commitment, not when they express obligation, advice, or normative judgment.	may, might, maybe, shall, should, can, could, would, ought	Modal auxiliaries are one of the most established hedge types, namely can, could, may, might, must, should, will, would.
Approximation	A word or phrase that makes a claim less exact, less categorical, or more general by signaling approximate quantity, frequency, degree, or scope.	usually, approximately, generally, frequently, normally, some, many, most	Hyland's line of work treats accuracy-related hedging as reducing precision, Salager-Meyer's classic taxonomy includes approximators, and Hinkel includes frequency and quantity softeners such as normally, relatively, often, frequently, and usually among hedge-like resources in academic writing.