

Semantic Anchoring and Rhetorical Inference Load in L2 Listening: An Empirical Mixed-Effects Study of Discourse-Level Processing Demands

Mohamed Mekheimer

mohamed.abdelgawad3@edu.bsu.edu.eg

Beni-Suef University

Article

Keywords: L2 listening, semantic anchoring, rhetorical inference load, mixed-effects modeling, discourse annotation, cognitive load, listening comprehension

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11082186/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

TITLE PAGE

Title:

**Semantic Anchoring and Rhetorical Inference Load in L2 Listening: An Empirical
Mixed-Effects Study of Discourse-Level Processing Demands**

Short Title / Running Head:

Anchoring and Rhetorical Load in L2 Listening

Author Name & Affiliation:

Mohamed Mekheimer^{1*}

¹ Faculty of Education, Beni-Suef University, Beni-Suef, Egypt

*** Corresponding Author:**

Mohamed Mekheimer, PhD

Professor of Applied Linguistics (TEFL)

Faculty of Education, Beni-Suef University, Beni-Suef, Egypt

Email: mohamed.abdelgawad3@edu.bsu.edu.eg

ORCID: <https://orcid.org/0000-0001-6784-0993>

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Abstract

Second-language (L2) listening materials are commonly characterized by lexical, syntactic, acoustic, and topic-based variables, yet these descriptors do not fully capture how discourse organization shapes real-time comprehension. This empirical study tests two complementary discourse-level constructs: semantic anchoring, the degree to which a segment supplies explicit referents, relations, definitions, examples, and conceptual continuity, and rhetorical inference load, the degree to which interpretation depends on metaphor, stance, patterned form, cultural frames, and perspective shifts. Twenty-four listening segments instantiated a 2×2 design crossing high versus low semantic anchoring with high versus low rhetorical inference load. Ninety-six EFL listeners (32 each at CEFR B1, B2, and C1) responded to all 24 segments, yielding 2,304 listener-by-segment observations; three annotators independently coded the discourse features. Inter-annotator reliability was acceptable to strong (ordinal Krippendorff's $\alpha = .723-.800$). In crossed random-intercept mixed-effects models, semantic anchoring predicted higher accuracy ($b = .1623$), shorter response time ($b = -259.74$ ms), lower cognitive load ($b = -1.1042$), and less form-monitoring ($b = -9.3559$), all $p < .001$. Rhetorical inference load predicted the opposite pattern: lower accuracy ($b = -.0838$), longer response time ($b = 375.54$ ms), higher cognitive load ($b = .9722$), and greater form-monitoring ($b = 20.1198$), all $p < .001$. The anchoring $\times$ rhetorical-load interaction was significant only for accuracy ($b = .1224$, $p = .009$), indicating that strong anchoring attenuated the accuracy cost of rhetorically demanding input. Proficiency improved all modeled outcomes and moderated several rhetorical-load effects. At the segment level, semantic anchoring was strongly related to accuracy ($r = .722$), rhetorical density to response time ($r = .785$) and form-monitoring ($r = .848$), whereas lexical entropy was unrelated to the outcomes. The findings support a multidimensional account of L2 listening difficulty in which discourse support and rhetorical inference make separable contributions beyond conventional complexity descriptors.

Keywords: L2 listening; semantic anchoring; rhetorical inference load; mixed-effects modeling; discourse annotation; cognitive load; listening comprehension

1 Introduction

Listening research increasingly depends on reusable audio materials, transcripts, assessment items, learner responses, and metadata. In corpora, experiments, and assessments, those materials are commonly described by genre, topic, duration, lexical frequency, syntactic complexity, speech rate, accent, recording quality, or target proficiency. Such descriptors are indispensable, but they do not fully explain why passages that appear comparable on these dimensions can produce markedly different comprehension demands. One passage may state its topic, referents, causal relations, and event structure explicitly, whereas another may require listeners to infer stance, resolve metaphor, activate a cultural frame, or track rapid shifts in perspective (Field, 2008; Goh, 2000; Rost, 2016).

The problem is partly one of resource description. Recent learner-resource initiatives emphasize rich metadata, transparent annotation, and reusable documentation because interpretability depends on knowing not only who produced the speech and under what conditions, but also what properties of the material are being compared (Knill et al., 2025; Paquot et al., 2024). For listening resources, the missing layer is often discourse organization. A task nominally targeted at the same proficiency level may require literal recall, causal integration, stance recognition, metaphor interpretation, or culture-specific inference; unless those demands are documented, difficulty can be attributed to the listener when it partly originates in an unreported property of the stimulus or task.

Accounts of listening difficulty often distinguish discourse that makes propositional relations explicit from discourse whose communicative force is carried more strongly by figurative, evaluative, rhythmic, symbolic, or culturally situated cues. Treating these properties as opposite ends of a single continuum, however, is conceptually unstable. A political speech can be both factually explicit and rhetorically elaborate, whereas a poorly organized explanation can be low in both explicit anchoring and purposeful rhetoric. The framework therefore represents them as two orthogonal dimensions: semantic anchoring and rhetorical inference load.

Semantic anchoring describes how much stable conceptual support the input supplies for constructing and updating a coherent situation model. Rhetorical inference load describes how much intended meaning depends on interpretive reconstruction beyond literal propositional content. The two dimensions can co-occur at any level, and their effects are expected to vary with listener proficiency, prior knowledge, cultural familiarity, task goal, working-memory

resources, and acoustic conditions. This two-dimensional formulation allows rhetorically rich input to be strongly scaffolded without assuming that rhetorical richness necessarily entails informational weakness.

The article makes four contributions. First, it integrates cognitive-linguistic, discourse-comprehension, psycholinguistic, pragmatic, L2-listening, annotation, and information-theoretic work into a parsimonious account of stimulus-level processing demands. Second, it operationalizes semantic anchoring and rhetorical inference load through explicit coding rules, diagnostic subfeatures, and a staged annotation workflow. Third, it separates stimulus features from listener outcomes and evaluates the reliability and empirical behavior of the resulting descriptors. Fourth, it tests the framework with 96 EFL listeners, 24 systematically collected listening segments, and 2,304 repeated listener-by-segment observations, allowing discourse-level effects, proficiency moderation, and feature-outcome associations to be examined within a single empirical design.

2 Study Orientation and Research Questions

This is an empirical, theory-guided study rather than a review article. The literature synthesis functions as the conceptual foundation for the empirical design: it identifies discourse-level mechanisms that may explain variation in L2 listening performance, converts those mechanisms into annotatable stimulus features, and derives testable predictions about listener outcomes. The framework draws on cognitive grammar and frame semantics, situation-model and construction-integration accounts, prediction and good-enough processing, pragmatics and figurative-language research, L2 listening and individual-difference research, and language-resource annotation and information-theoretic description (Bird & Liberman, 2001; Fillmore, 1982; Kintsch, 1988; Langacker, 2008; van den Heuvel et al., 2008).

The conceptual sources are used to specify two focal constructs rather than to estimate a pooled literature effect. Semantic anchoring captures the explicit conceptual support supplied by a segment, whereas rhetorical inference load captures the extent to which intended meaning must be reconstructed from figurative, evaluative, cultural, perspectival, or patterned-form cues. The empirical component then tests whether these independently manipulated dimensions are reliably annotatable and whether they account for variation in accuracy, response time, cognitive load, confidence, and form-monitoring after listener and segment dependence are taken into account.

Five research questions organize the study. RQ1 asks whether semantic anchoring, rhetorical inference load, form foregrounding, and profile shifts can be coded with acceptable inter-annotator reliability. RQ2 asks how semantic anchoring and rhetorical inference load relate to listening accuracy, response time, cognitive load, and form-monitoring. RQ3 asks whether semantic anchoring moderates the costs associated with rhetorical inference load. RQ4 asks whether listener proficiency changes these relationships. RQ5 asks whether human-coded and computational segment indicators show differentiated associations with listener outcomes, rather than collapsing into a single global information-density or difficulty dimension.

3 From Discourse Density to Two Orthogonal Resource Dimensions

3.1 Semantic anchoring

Semantic anchoring is the degree to which a discourse segment provides stable, explicit conceptual support for constructing and updating a situation model. Anchoring increases when a passage names its domain, maintains referential chains, states causal or temporal relations, defines unfamiliar terms, supplies examples, marks transitions, and summarizes central relations. The construct draws on frame semantics and situation-model theory, both of which treat comprehension as the integration of linguistic material with structured knowledge about entities, events, relations, and goals (Fillmore, 1982; Kintsch, 1988; van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998).

High semantic anchoring is not synonymous with low linguistic complexity or low information content. A technically dense lecture can be highly anchored if it identifies the domain, defines central concepts, and signals how propositions relate to one another. Conversely, a grammatically simple passage can be weakly anchored when pronouns are ambiguous, causal relations are merely implied, or the topic must be reconstructed from scattered cues. Anchoring is therefore a discourse-support property rather than a readability label.

3.2 Rhetorical inference load

Rhetorical inference load is the degree to which successful interpretation depends on reconstructing meaning from cues whose communicative force exceeds literal propositional content. Relevant cues include metaphor and analogy, evaluative or epistemic stance, contrast, parallelism, repetition used for emphasis, symbolic imagery, irony, audience positioning, culture-

bound allusion, and shifts in viewpoint or conceptual profile. Cognitive-linguistic and pragmatic accounts treat these phenomena as meaning making rather than decorative ornamentation (Fauconnier & Turner, 2002; Lakoff & Johnson, 1980; Langacker, 2008; Sperber & Wilson, 1995).

High rhetorical inference load is not intrinsically difficult or undesirable. Conventional metaphors can support comprehension, repetition can strengthen memory, and parallelism can increase predictability. Difficulty is more likely when rhetorical interpretation competes with phonological decoding, lexical access, syntactic integration, and memory under time pressure, particularly when the relevant frames are unfamiliar. The construct therefore describes an interpretive requirement of the stimulus, not a negative value judgment about rhetorical language.

3.3 Orthogonality, moderation, and processing priority

The central conceptual claim is that semantic anchoring and rhetorical inference load are orthogonal. A segment can be high on both dimensions: a skilled speaker can state a clear causal argument while also using metaphor, stance, and patterned repetition. A segment can also be low on both: it may be vague without being rhetorically elaborate. Treating the dimensions independently avoids the assumption that explicit information and rhetoric necessarily trade off against each other and yields a four-cell profile that can be annotated and tested empirically. The two-dimensional model is presented in Fig. 1, and the resulting four-cell discourse profile is summarized in Table 3.

Neither dimension alone determines observed difficulty. Proficiency, lexical and grammatical knowledge, prior knowledge, cultural familiarity, working memory, metacognitive regulation, affect, task goal, and acoustic quality can change how the same passage is processed (Baddeley, 2000; Karalık & Merç, 2025; Saito et al., 2025; Satori, 2021; Vandergrift & Baker, 2015; Vafae & Suzuki, 2020). A stance-identification task can make rhetorical cues useful, whereas a literal-recall task may render the same cues peripheral. The resource record should therefore keep source features, moderator metadata, and task outcomes analytically separate.

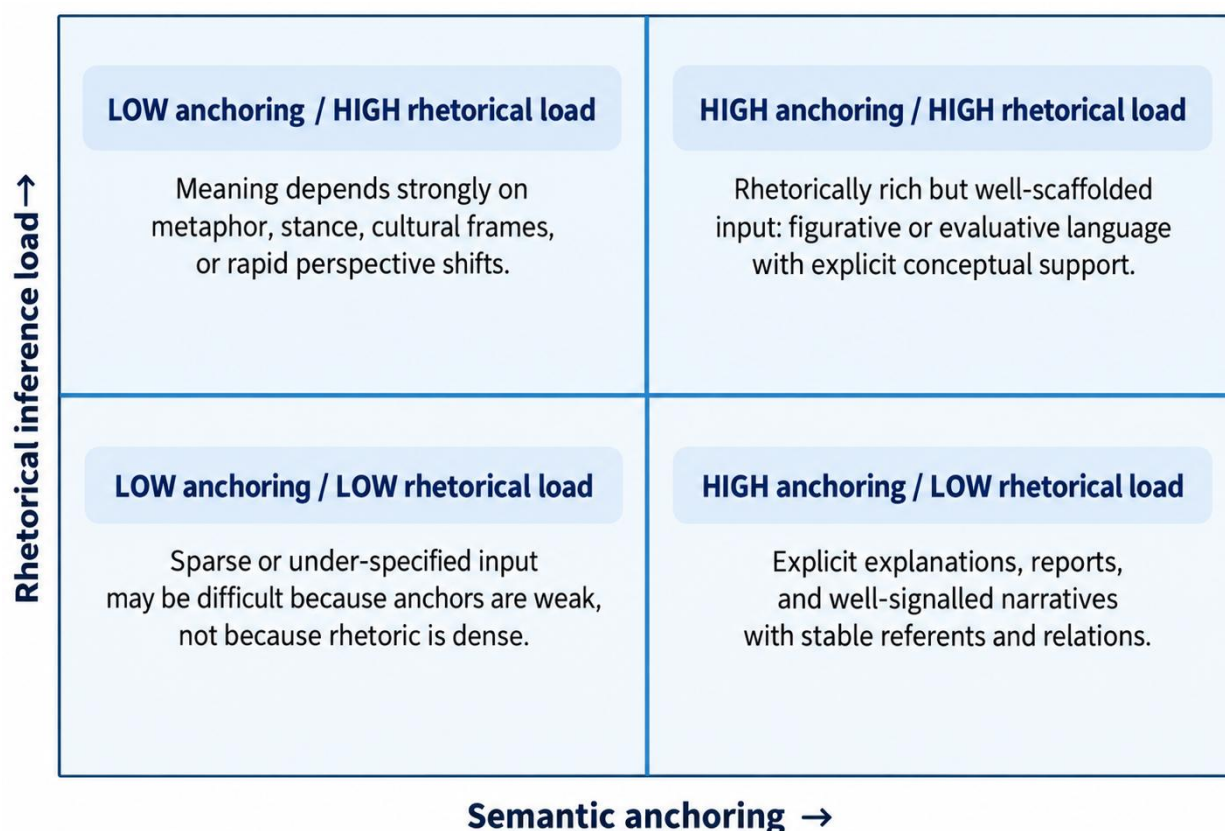
This distinction also clarifies the role of grammar. Grammar remains central to listening because it organizes participant roles, temporal relations, modality, causality, and information structure, but its cognitive visibility may change with discourse conditions. When semantic

anchoring is strong, grammatical form can operate relatively quietly as a scaffold for integrating content. When meaning is indirect, evaluative, or rhetorically patterned, form itself may become salient because word order, modality, parallelism, fronting, or tense-aspect contrast contributes to stance and inference. The framework consequently treats processing priority as dynamic rather than as a grammar-versus-meaning dichotomy. The core annotation scales used to operationalize the two dimensions are presented in Table 1, and the diagnostic subfeatures used to justify each rating are listed in Table 2.

[Fig. 1 should appear here]

Figure 1

Two-Dimensional Model of Semantic Anchoring and Rhetorical Inference Load



Note. The dimensions are independent resource descriptors. Listener proficiency, prior knowledge, cultural familiarity, task goal, and acoustic conditions are moderators rather than components of either dimension.

[Table 1 should appear here]

Table 1

Core Annotation Scales Used in the Supplementary Data

Variable	Scale	Operational anchor
Semantic anchoring	0–4	0 = no explicit domain/referents; 1 = topic recoverable but relations mostly implicit; 2 = identifiable domain with some explicit referents; 3 = named domain, explicit referents, and stated relation(s); 4 = level 3 plus definition, example, or summary.
Rhetorical inference load	0–4	0 = literal/propositional; 1 = isolated evaluative or conventional cue; 2 = figurative/evaluative framing supports meaning; 3 = meaning partly depends on metaphor, stance, or repetition; 4 = meaning is strongly distributed across metaphor, symbolism, stance, and/or culture-bound frames.
Profile shifts	Count	Number of changes in focal entity, stance, audience, time frame, value, or conceptual base within the segment.
Form foregrounding	0–4	Degree to which grammatical form itself—for example, parallelism, fronting, modality, marked word order, or tense-aspect contrast—becomes a salient rhetorical cue.
Annotator confidence	1–3	1 = low/ambiguous; 2 = moderate; 3 = high. Ambiguous segments are retained rather than forced into certainty.

Note. These coding scales correspond to those used in the supplementary statistical materials. They are operational descriptors for annotation and empirical analysis, not established interval-level psychometric measures.

[Table 2 should appear here]

Table 2

Diagnostic Subfeatures for Segment-Level Annotation

Dimension	Feature	Annotation question
Semantic anchoring	Domain explicitness	Is the topic or domain explicitly named and maintained?
Semantic anchoring	Referential clarity	Can participants, objects, and pronouns be tracked without repair?
Semantic anchoring	Relational explicitness	Are causal, temporal, spatial, or explanatory relations overtly signaled?
Semantic anchoring	Explanatory support	Are definitions, examples, paraphrases, or summaries supplied where needed?

Dimension	Feature	Annotation question
Semantic anchoring	Discourse continuity	Does the segment maintain a stable topic or frame with clear transitions?
Rhetorical inference load	Figurative dependence	Must metaphor, analogy, irony, or symbolic imagery be interpreted to recover intended meaning?
Rhetorical inference load	Stance/evaluation	How much meaning depends on evaluative or epistemic positioning?
Rhetorical inference load	Patterned form	Do repetition, parallelism, rhythm, fronting, or contrast carry rhetorical force?
Rhetorical inference load	Cultural-frame dependence	Is culture-specific knowledge needed to recover the intended interpretation?
Rhetorical inference load	Perspective/profile shifts	Does interpretation require tracking changes in viewpoint, audience, time frame, or profiled value?

Note. The subfeatures are diagnostic cues used to justify the two core 0–4 ratings.

[Table 3 should appear here]

Table 3

Four-Cell Discourse Profile by the Two Dimensions

Semantic anchoring	Rhetorical inference load	Illustrative profile	Interpretive implication
High	Low	Explicit report or explanation	Stable referents and overt relations; little figurative reconstruction. Predicted burden falls mainly on decoding and semantic integration.
High	High	Scaffolded rhetorical discourse	Clear domain and relations combined with metaphor, stance, or patterned emphasis. Rhetorical interpretation is required but supported by strong anchors.
Low	Low	Under-specified discourse	Few stable cues and little purposeful rhetorical structure. Difficulty may arise from vagueness, ellipsis, or weak cohesion rather than rhetorical density.

Semantic anchoring	Rhetorical inference load	Illustrative profile	Interpretive implication
Low	High	Inference-heavy rhetorical discourse	Meaning depends strongly on metaphor, stance, cultural frames, or perspective shifts while explicit anchors are limited; inferential burden is expected to be greatest for unfamiliar listeners.

Note. These are discourse profiles, not genre labels. Any genre can contain segments from more than one cell.

4 Theoretical Foundations

Cognitive grammar provides one foundation because it treats grammar as a symbolic system in which form and meaning are inseparable. Grammatical constructions profile particular participants, relations, temporal configurations, and viewpoints against broader conceptual bases; the same event can therefore be construed in different ways depending on what is foregrounded (Langacker, 2008). For listening-resource design, the implication is that grammatical complexity should not be recorded only as a surface count. Researchers should also consider whether grammatical patterning supplies a background scaffold or becomes a salient cue that listeners must actively exploit.

Frame semantics adds a second foundation. Words and expressions evoke structured background knowledge, so comprehension depends partly on whether the discourse supplies enough cues for the relevant frame to be activated and maintained (Fillmore, 1982). Semantic anchoring operationalizes this requirement by asking whether the domain, referents, causal or temporal relations, examples, and explanatory links are recoverable from the discourse itself. Culture-bound rhetoric places a stronger burden on frame access because the intended interpretation may depend on assumptions that are not stated explicitly.

Situation-model and construction-integration theories converge with this account by emphasizing coherent representations of entities, events, time, space, causality, and goals (Kintsch, 1988; van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). A listener must integrate successive propositions while speech continues to unfold, and unstable reference or implicit causal relations can force repair. Strong anchoring should therefore reduce the amount of reconstruction required to maintain coherence, even when the vocabulary or syntax is demanding.

Prediction research explains why the distribution of information also matters. Comprehenders use contextual constraints and prior knowledge to anticipate likely continuations, but those predictions are probabilistic and can be disrupted by unexpected frame shifts or weak discourse cues (Huettig, 2015; Kuperberg & Jaeger, 2016; Levy, 2008; Pickering & Garrod, 2013). Information-density accounts similarly model the cost of encountering relatively unexpected linguistic material. The present framework extends that logic from local predictability to resource-level documentation by asking where information is placed, how strongly it is anchored, and whether meaning depends on rhetorical reconstruction.

Cognitive-load and good-enough accounts explain why the same resource feature can have different consequences across listeners and tasks. Working-memory capacity constrains the simultaneous maintenance and integration of speech, while listeners often settle for representations that are sufficient for the immediate goal rather than exhaustively parsed (Baddeley, 2000; Ferreira & Patson, 2007; Satori, 2021). Explicit anchoring can therefore compensate for incomplete parsing during gist-oriented tasks, whereas stance, irony, metaphor, or audience alignment may require more attention to patterned form. Resource evaluation must report the task construct alongside the text construct rather than assigning a passage an intrinsic difficulty value.

Pragmatics clarifies why rhetorical cues are not decorative extras. Relevance-oriented accounts propose that hearers seek interpretations that justify processing effort through contextual effects, and figurative language can prompt cross-domain mapping, evaluation, and audience positioning (Fauconnier & Turner, 2002; Lakoff & Johnson, 1980; Sperber & Wilson, 1995). In a listening resource, rhetorical inference load should therefore be coded as a property of meaning construction. Its influence may be beneficial, neutral, or costly depending on conventionality, familiarity, proficiency, and task demands.

5 Relation to Existing Complexity, Cohesion, and Information-Density Metrics

The proposed dimensions are complementary to, rather than replacements for, established measures of linguistic complexity and cohesion. Complexity-accuracy-fluency research has developed indices of syntactic elaboration, subordination, and lexical sophistication, while Coh-Metrix and related systems estimate multilevel properties such as referential and causal cohesion (Housen et al., 2019; McNamara et al., 2014). Entity-based coherence models track how

discourse entities are distributed across adjacent units, and information-theoretic approaches quantify uncertainty, redundancy, or compression (Barzilay & Lapata, 2008; Ehret & Szmrecsanyi, 2016, 2019; Jaeger, 2010). These families of measures answer important questions, but they do not directly identify whether intended meaning requires metaphor resolution, stance inference, or culture-specific framing.

Information-theoretic work is particularly relevant because it demonstrates that information is not a single scalar property. Lexical dispersion, grammatical divergence, structural regularity, compression, and positional concentration can move differently across proficiency levels or representational layers (Alzahrani, 2024; Ehret et al., 2023; Mekheimer, 2026; Sun & Wang, 2021; Wang et al., 2022). The methodological lesson for listening is that a high-entropy segment is not automatically difficult, and a low-entropy segment is not automatically easy. Interpretation depends on what the index captures, the preprocessing decisions used to compute it, and its relationship to the discourse function and listener task.

Direct behavioral evidence distinguishing explicitly informational from rhetorically elaborated L2 listening input remains limited. Mekheimer and Fageeh (2025) reported different comprehension and processing patterns across two types of EFL listening input, but their restricted passage set warrants a hypothesis-generating rather than universal interpretation. The present framework therefore specifies constructs that can be independently annotated and tested across resources, genres, languages, tasks, and learner groups.

The critical distinction is between formal predictability and interpretive requirement. A passage can be highly cohesive at the lexical or entity level while still demanding substantial rhetorical inference, and it can be lexically simple while leaving its domain or referents underspecified. For this reason, automated measures should be treated as candidate correlates or auditing tools rather than as definitions of semantic anchoring or rhetorical inference load. The framework is strongest when computational features and human discourse annotation are analyzed together.

6 Annotation Framework and Reliability

The framework is designed for segment-level annotation because whole-text labels obscure internal variation. A three-minute lecture can open with a rhetorical hook, move into an explicitly anchored explanation, and end with an evaluative summary. Depending on the

resource, the unit may be an intonation unit, clause, turn, sentence-like transcription unit, or discourse move. The segmentation rule should be chosen before coding, justified in the documentation, and applied consistently across the resource.

Annotation should proceed in stages. First, annotators segment the audio-transcript pair and label the dominant discourse move, such as definition, example, claim, narrative event, evaluation, appeal, or transition. Second, they code semantic anchoring without consulting listener performance. Third, they code rhetorical inference load, profile shifts, and form foregrounding. Fourth, they record uncertainty or mixed cases; only after independent source coding is complete are the annotations linked to task properties and listener outcomes. This ordering reduces criterion contamination because a segment is not labeled rhetorically demanding merely because learners performed poorly on it. The recommended annotation workflow is summarized in Table 4.

A codebook should include definitions, positive examples, non-examples, borderline cases, and decision rules. Annotators should train on a calibration subset and then independently code a held-out reliability subset before adjudication. Ordinal Krippendorff's α is appropriate for the 0–4 ratings because it can accommodate multiple annotators and ordinal disagreement (Artstein & Poesio, 2008; Krippendorff, 2018). Agreement should be reported construct by construct, because a resource may achieve good reliability for referential support yet remain unstable for metaphor or perspective shifts.

Uncertainty should be represented rather than suppressed. Ambiguous segments can receive a lower confidence rating, and systematic low-confidence cases should trigger codebook revision. If a feature cannot attain acceptable agreement after repeated clarification, it should not be retained simply because it is theoretically attractive. This principle is essential for converting an interpretive framework into a reproducible language resource.

[Table 4 should appear here]

Table 4

Recommended Annotation Workflow

Stage	Action	Required output
-------	--------	-----------------

Stage	Action	Required output
1. Prepare	Select audio, transcript, metadata, and a segmentation unit; document transcription conventions.	Primary resource and segmentation specification
2. Calibrate	Train annotators on examples, non-examples, and borderline cases; revise decision rules.	Versioned codebook
3. Annotate source features	Code anchoring, rhetorical inference, profile shifts, and form foregrounding independently of learner outcomes.	Segment-level annotation file
4. Assess reliability	Compute feature-level agreement on a held-out subset before adjudication.	Reliability report
5. Adjudicate	Resolve disagreements while retaining provenance of the original independent labels.	Adjudicated layer plus provenance
6. Link tasks and outcomes	Attach task type, proficiency, accuracy, response time, recall, confidence, load, or strategy data.	Evaluation layer
7. Release and report	Publish documentation, codebook, derived features, scripts, and licensing constraints where possible.	Reusable resource package

Note. Primary data and annotation should remain separable so that annotation can be corrected or extended without altering the original audio and transcript.

7 Automated and Information-Theoretic Indicators

Some framework features can be approximated computationally, but automated indicators and discourse annotations should not be treated as interchangeable. Speech rate, pause distribution, lexical frequency, lexical entropy, repetition counts, POS-sequence divergence, compression ratio, and positional measures can be extracted automatically or semi-automatically. Such indicators may help audit annotation or describe complementary layers of the resource, especially lexical dispersion, local structural regularity, and redundancy (Alzahrani, 2024; Ehret & Szmrecsanyi, 2016, 2019; Jaeger, 2010; Wang et al., 2022). Candidate automated indicators and the cautions that apply to their evaluation are listed in Table 5.

Other features depend on interpretation. A metaphor can be conventional or novel; repetition can signal disfluency, repair, cohesion, or rhetorical intensification; and a proper noun can function as a simple referent or activate a culturally dense frame. Automated counts are therefore candidate correlates, not definitions, of semantic anchoring or rhetorical inference load.

The same caution applies to automatically generated labels: their agreement with trained human judgments is an empirical question that must be measured rather than assumed.

Information density also has position and structure. Recent work on spoken multimodal discourse shows that information can be concentrated differently across positions within an utterance or exchange (Trujillo & Holler, 2025). Consequently, whole-passage averages can hide local peaks of uncertainty, salience, or rhetorical organization. Whenever technically feasible, computational indicators should therefore be calculated at the same segmental resolution as the human discourse annotations.

[Table 5 should appear here]

Table 5

Candidate Indicators and Evaluation Cautions

Indicator family	What it can estimate	Implementation caution
Lexical entropy or dispersion	Breadth and evenness of lexical choice in a transcript or segment.	High dispersion is not automatically beneficial or difficult; interpret it with domain familiarity and task goal.
POS n-gram divergence	Distance between a segment and a reference distribution of local grammatical sequencing.	Requires stable tagging; learner or nonstandard speech may need manual audit.
Compression ratio	Global redundancy and recurring organization in standardized text strings.	Sensitive to preprocessing; report normalization and window length.
Positional concentration	Whether information is front-loaded or distributed across the segment.	Use rate-based comparison when segment lengths differ; avoid interpreting raw ratios as a total difficulty score.
Rhetorical feature density	Concentration of metaphor, repetition, stance, contrast, symbolic imagery, and evaluative framing.	Human annotation may remain necessary; explicit codebook examples are essential.
Semantic anchoring index	Degree of explicit domain marking, reference, causal/temporal relations, definitions, and examples.	Validate against independent judgments and listener-task outcomes rather than assuming that anchoring equals comprehensibility.

Indicator family	What it can estimate	Implementation caution
Outcome measures	Accuracy, response time, recall, confidence, cognitive load, and strategy/form-monitoring measures.	Keep stimulus features separate from task-specific outcomes.

Note. No single indicator should be treated as a total difficulty score. Interpretation depends on the representational layer, preprocessing choices, and listener-task context.

8 Method

8.1 Design and data structure

The empirical study used a fully crossed repeated-measures design in which semantic anchoring (high vs. low) and rhetorical inference load (high vs. low) formed a 2×2 stimulus structure. The dataset contains 96 EFL listeners, divided evenly across CEFR B1, B2, and C1 levels ($n = 32$ per level). Each listener contributed responses to all 24 listening segments, producing 2,304 listener-by-segment observations. The observation-level records include number correct out of four items, proportional accuracy, response time in milliseconds, cognitive load on a 1–9 scale, confidence on a 1–7 scale, and form-monitoring on a 0–100 scale.

8.2 Listening segments and experimental conditions

The study used 24 listening segments, with six segments in each of the four semantic-anchoring-by-rhetorical-inference-load cells. The conditions were designed to vary semantic anchoring and rhetorical inference load independently while matching segments on duration, speech rate, vocabulary level, and word count. Segment-level records also included a semantic-anchoring index, rhetorical-density score, profile-shift count, form-foregrounding rating, lexical entropy, POS n-gram divergence, compression ratio, and positional concentration. This layered structure allowed the focal discourse constructs to be examined alongside complementary computational indicators without treating any single measure as a total difficulty score.

8.3 Annotation and outcome measures

Three independent annotators coded the segment-level discourse features using the study codebook. Semantic anchoring, rhetorical inference load, and form foregrounding were rated on 0–4 ordinal scales; profile shifts were counted, and annotator uncertainty was recorded separately. Stimulus features were coded independently of listener performance so that poor

performance could not be used to define a segment as intrinsically difficult. Inter-annotator reliability was evaluated with ordinal Krippendorff's α . Listener outcomes were linked only after the stimulus-feature layer had been established.

8.4 Ethical approval

The study protocol received ethical approval from the Faculty of Education, Beni-Suef University (Approval No. BSU-FOE-004-01-02-26). All study procedures were conducted in accordance with the applicable local institutional ethical guidelines and the ethical principles of the Declaration of Helsinki.

8.5 Statistical analysis

Analysis proceeded at both listener-by-segment and segment levels. Descriptive cell means were computed from participant-level condition means, and paired by-subject contrasts were summarized with Cohen's d . The primary inferential analyses used linear mixed-effects models with crossed random intercepts for listener and segment. Semantic anchoring and rhetorical inference load were contrast-coded +0.5 (high) and -0.5 (low), and proficiency was coded -1, 0, and +1 for B1, B2, and C1. Models included the two discourse factors, proficiency, all two-way interactions, and the three-way interaction. Finally, Pearson correlations at the segment level ($N = 24$) examined how human-coded and computational indicators related to mean accuracy, response time, cognitive load, and form-monitoring. These segment-level correlations were interpreted as exploratory because of the small number of segments.

9 Findings

9.1 Annotation reliability

The annotation layer showed acceptable to strong agreement across the three annotators and 24 segments. Ordinal Krippendorff's α was .800 for semantic anchoring, .768 for rhetorical inference load, .765 for form foregrounding, and .723 for profile shifts. Thus, all four constructs cleared the .667 benchmark commonly used for tentative conclusions in content analysis, and semantic anchoring reached the more stringent .800 benchmark. The pattern is important because the constructs differ in interpretive complexity: explicit anchoring cues were most reproducible, whereas profile shifts, which require more interpretive judgment, were somewhat less stable but

remained usable. The study design, sample, and annotation reliability estimates are summarized in Table 6.

[Table 6 should appear here]

Table 6

Design, Sample, and Annotation Reliability of the Empirical Study

Component	Specification
Segment design	24 segments; 2×2 design crossing semantic anchoring and rhetorical inference load; six segments per cell; duration, speech rate, vocabulary level, and word count matched across cells.
Annotation layer	Ratings from three independent annotators for semantic anchoring, rhetorical inference load, form foregrounding, and profile shifts.
Krippendorff's α	Semantic anchoring = .800; rhetorical inference load = .768; form foregrounding = .765; profile shifts = .723.
Listeners	96 EFL listeners: 32 each at CEFR B1, B2, and C1.
Listener-segment observations	2,304 observations including accuracy, response time, cognitive load, confidence, and form-monitoring.

Note. The supplementary materials contain the coding records, segment-level features, listener roster, observation-level data, correlations, descriptive statistics, fixed effects, effect sizes, and reliability calculations.

9.2 Descriptive outcomes and effect sizes

The descriptive results show a consistent two-dimensional pattern rather than a single easy-difficult continuum. Performance was most favorable in the high-anchoring/low-rhetorical-load condition: mean accuracy was .83, response time was 2,610.96 ms, cognitive load was 2.64, confidence was 5.56, and form-monitoring was 22.68. The low-anchoring/high-rhetorical-load condition produced the least favorable profile, with accuracy of .59, response time of 3,246.24 ms, cognitive load of 4.71, confidence of 4.28, and form-monitoring of 52.16. The two mixed conditions fell between these extremes, indicating that strong semantic support and rhetorical inferential demand contributed differently to comprehension and processing effort. Outcome means and standard deviations for each condition are reported in Table 7.

[Table 7 should appear here]

Table 7

Outcome Means and Standard Deviations by Semantic-Anchoring and Rhetorical-Inference Condition

Measure	High A / High R	High A / Low R	Low A / High R	Low A / Low R
Accuracy (0–1)	.81 (.16)	.83 (.12)	.59 (.26)	.73 (.18)
Response time (ms)	2,990.96 (440.37)	2,610.96 (332.73)	3,246.24 (442.70)	2,875.16 (300.32)
Cognitive load (1–9)	3.62 (1.03)	2.64 (0.76)	4.71 (1.05)	3.75 (0.73)
Confidence (1–7)	4.96 (0.59)	5.56 (0.56)	4.28 (0.60)	4.90 (0.62)
Form-monitoring (0–100)	42.86 (5.98)	22.68 (5.63)	52.16 (6.02)	32.09 (6.48)

Note. Values are participant-level condition means aggregated across the six segments in each cell; SDs are across listeners. A = semantic anchoring; R = rhetorical inference load.

By-subject main-effect contrasts were large. For accuracy, semantic anchoring yielded $d = 1.274$ and rhetorical inference load $d = -0.885$. For response time, the corresponding effects were $d = -2.273$ and $d = 2.014$, and for cognitive load they were $d = -3.492$ and $d = 2.080$. The direction of these effects is theoretically coherent: stronger anchoring was associated with better and more efficient comprehension, whereas greater rhetorical inference load was associated with reduced accuracy and greater processing effort.

9.3 Mixed-effects model findings

The crossed random-intercept models confirmed the descriptive pattern while accounting for repeated observations from both listeners and segments. Semantic anchoring significantly predicted higher accuracy ($b = .1623, p < .001$), shorter response time ($b = -259.74 \text{ ms}, p < .001$), lower cognitive load ($b = -1.1042, p < .001$), and lower form-monitoring ($b = -9.3559, p < .001$). Rhetorical inference load significantly predicted the reverse pattern: lower accuracy ($b = -.0838, p < .001$), longer response time ($b = 375.54 \text{ ms}, p < .001$), higher cognitive load ($b = .9722, p < .001$), and substantially greater form-monitoring ($b = 20.1198, p < .001$). These effects demonstrate that the two discourse dimensions make separable contributions to outcome variation. Selected fixed effects from the mixed-effects models are presented in Table 8.

The anchoring $\times$ rhetorical-load interaction was significant for accuracy ($b = .1224$, $p = .009$) but not for response time ($p = .763$), cognitive load ($p = .957$), or form-monitoring ($p = .881$). The cell means clarify this interaction: high rhetorical load reduced accuracy by only about .02 when anchoring was high (.83 to .81), but by about .15 when anchoring was low (.73 to .59). Strong semantic anchoring therefore buffered the accuracy cost of rhetorically demanding input, even though it did not eliminate the additional time, load, or form-monitoring associated with rhetorical processing.

Proficiency produced robust main effects on every modeled outcome: higher proficiency predicted greater accuracy ($b = .1720$, $p < .001$), shorter response time ($b = -406.45$ ms, $p < .001$), lower cognitive load ($b = -.9837$, $p < .001$), and lower form-monitoring ($b = -6.1732$, $p < .001$). Proficiency also moderated the discourse effects selectively. The anchoring $\times$ proficiency interaction was significant for accuracy only, whereas the rhetorical-load $\times$ proficiency interaction was significant for accuracy, response time, and cognitive load. The model-implied average rhetorical penalty on accuracy attenuated from approximately -.147 at B1 to -.084 at B2 and -.020 at C1; the corresponding response-time cost decreased from about 549 ms to 376 ms and 202 ms, and the cognitive-load increase from about 1.35 to .97 and .59 points. No three-way interaction reached significance.

[Table 8 should appear here]

Table 8

Selected Fixed Effects From the Mixed-Effects Models

Term	Accuracy (0–1)	Response time (ms)	Cognitive load (1–9)	Form-monitoring (0–100)
Semantic anchoring	.1623 ($p < .001$)	-259.74 ($p < .001$)	-1.1042 ($p < .001$)	-9.3559 ($p < .001$)
Rhetorical inference load	-.0838 ($p < .001$)	375.54 ($p < .001$)	.9722 ($p < .001$)	20.1198 ($p < .001$)
Anchoring $\times$ rhetorical load	.1224 ($p = .009$)	8.92 ($p = .763$)	.0104 ($p = .957$)	.1146 ($p = .881$)
Proficiency (linear)	.1720 ($p < .001$)	-406.45 ($p < .001$)	-.9837 ($p < .001$)	-6.1732 ($p < .001$)

Note. Values are unstandardized fixed-effect coefficients. Models use crossed random intercepts for listener and segment. The full interaction output is reported in the supplementary statistical materials; coefficients describe the behavior of the listeners in the dataset.

9.4 Indicator-outcome correlations

Segment-level correlations further showed that discourse difficulty was not captured by a single information-density statistic. The semantic anchoring index correlated strongly with accuracy ($r = .722, p < .001$) and negatively with cognitive load ($r = -.639, p = .001$). Rhetorical density correlated strongly with response time ($r = .785, p < .001$), cognitive load ($r = .624, p = .001$), and form-monitoring ($r = .848, p < .001$). Form foregrounding showed a similar pattern for response time ($r = .737, p < .001$) and form-monitoring ($r = .815, p < .001$), while positional concentration was inversely related to response time ($r = -.737, p < .001$), cognitive load ($r = -.623, p = .001$), and form-monitoring ($r = -.813, p < .001$). The full set of segment-level correlations is reported in Table 9.

By contrast, lexical entropy was essentially unrelated to all four mean outcomes ($|r| \leq .061, p \geq .776$). POS n-gram divergence, compression ratio, profile shifts, and other layer-specific indicators showed selective rather than uniform relationships. Because these correlations are based on only 24 segments and the p values are unadjusted for multiplicity, they should be read as exploratory evidence about convergent patterns, not as estimates of independent causal effects. Their main value is in showing that theoretically distinct discourse layers can carry information that a global lexical-entropy measure misses.

[Table 9 should appear here]

Table 9

Segment-Level Correlations Between Indicators and Mean Outcomes

Indicator	Accuracy	RT (ms)	Cognitive load	Form-monitoring
Semantic anchoring index	.722 (< .001)	-.432 (.035)	-.639 (.001)	-.293 (.164)
Rhetorical density	-.408 (.048)	.785 (< .001)	.624 (.001)	.848 (< .001)
Profile shifts	-.172 (.423)	.478 (.018)	.350 (.093)	.548 (.006)
Form foregrounding	-.262 (.216)	.737 (< .001)	.582 (.003)	.815 (< .001)
Lexical entropy	.028 (.895)	.054 (.801)	.026 (.905)	.061 (.776)
POS n-gram divergence	-.681 (< .001)	.466 (.022)	.656 (.001)	.341 (.102)
Compression ratio	-.258 (.223)	.581 (.003)	.444 (.030)	.679 (< .001)

Indicator	Accuracy	RT (ms)	Cognitive load	Form-monitoring
Positional concentration	.381 (.066)	-.737 (< .001)	-.623 (.001)	-.813 (< .001)

Note. Entries are Pearson r values with p values in parentheses ($N = 24$ segments). RT = response time. The p values are unadjusted; these correlations are exploratory.

9.5 Synthesis of findings

Taken together, the findings answer the five research questions coherently. The discourse features were codable with acceptable reliability; semantic anchoring consistently supported accuracy and processing efficiency; rhetorical inference load consistently increased processing demands and reduced accuracy; semantic anchoring buffered the accuracy cost of rhetorical load; and proficiency reduced several of the costs associated with rhetorically demanding input. The segment-level indicators also showed differentiated associations with outcomes, supporting the decision to model discourse architecture as multidimensional rather than as one undifferentiated density score.

The most important empirical pattern is therefore a partial dissociation. Semantic anchoring and rhetorical inference load were not simply opposite ends of the same continuum: both could be high, and their combination produced relatively high accuracy alongside elevated response time, cognitive load, and form-monitoring. This distinction provides the basis for interpreting the results theoretically rather than merely ranking the four conditions from easy to difficult.

10 Discussion

The findings provide empirical support for treating semantic anchoring as a discourse-level comprehension resource. Explicit domains, stable referents, stated relations, definitions, examples, and summaries appear to reduce the reconstruction required to maintain a coherent situation model while speech unfolds. The positive effect on accuracy and the simultaneous reductions in response time and cognitive load are consistent with construction-integration and situation-model accounts, in which comprehension becomes more efficient when entities and causal or temporal relations can be integrated without repeated repair (Kintsch, 1988; van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). The result also supports the framework's central

claim that discourse support is not reducible to lexical simplicity: the materials were matched on major surface properties, and lexical entropy itself was unrelated to the observed outcomes.

Rhetorical inference load showed a different empirical signature. Higher rhetorical load was associated with lower accuracy, slower responses, greater cognitive load, and an especially large increase in form-monitoring. This pattern suggests that metaphor, stance, parallelism, fronting, symbolic framing, cultural cues, and perspective shifts do not merely add stylistic ornament; they can redirect attention toward the relationship between linguistic form, speaker positioning, and intended meaning. That interpretation is consistent with pragmatic and cognitive-linguistic accounts in which figurative and evaluative cues require contextual integration and perspective construction (Fauconnier & Turner, 2002; Langacker, 2008; Sperber & Wilson, 1995). Importantly, the result should not be read as evidence that rhetoric is intrinsically harmful. Rather, rhetorical features appear to increase the amount or kind of interpretation required under the present task conditions.

The significant anchoring $\times$ rhetorical-load interaction for accuracy is theoretically informative because it qualifies a simple additive difficulty account. Strong anchoring almost eliminated the accuracy penalty associated with high rhetorical load, yet it did not remove the extra response time, cognitive load, or form-monitoring. In other words, explicit conceptual support helped listeners arrive at the correct interpretation, but rhetorically rich input still demanded additional processing. This distinction separates outcome success from processing effort and suggests that a passage can be accurately understood while remaining cognitively expensive. Such a pattern would be obscured by studies that use accuracy alone as the index of listening difficulty.

Proficiency further shaped these relationships. Higher-proficiency listeners were faster, more accurate, and reported less cognitive load, but the moderation pattern is more revealing than the main effect alone. The accuracy benefit of semantic anchoring was largest at B1 and smaller at C1, while the rhetorical-load costs for accuracy, response time, and cognitive load declined markedly as proficiency increased. This is consistent with work showing that more developed lexical, syntactic, phonological, and strategic resources allow advanced listeners to absorb greater inferential demands with less disruption (Saito et al., 2025; Vafaei & Suzuki, 2020; Vandergrift & Baker, 2015). The results therefore suggest that semantic scaffolding may

be especially valuable for less proficient listeners, whereas advanced listeners are more resilient to rhetorical complexity.

The indicator analysis also has methodological implications. Lexical entropy was effectively null across all outcomes, whereas semantic anchoring, rhetorical density, form foregrounding, POS divergence, compression, and positional concentration showed distinct patterns of association. This supports the broader information-theoretic argument that linguistic information is layered rather than scalar (Ehret & Szmrecsanyi, 2016, 2019; Jaeger, 2010; Mekheimer, 2026). For L2 listening research, a single global complexity or entropy index is therefore unlikely to explain why two passages with similar lexical profiles produce different comprehension trajectories. Human discourse annotation and computational indicators are best treated as complementary measurements of different representational levels.

More broadly, the study shifts the framework from conceptual specification to empirical test. The two proposed dimensions were independently manipulated, reliably annotated, and linked to multiple listener outcomes in a crossed data structure. The resulting pattern supports the framework's central distinction between conceptual support and inferential demand while also showing where the theory needs qualification: anchoring buffered accuracy rather than all forms of processing cost, proficiency moderated some but not all effects, and not every candidate computational indicator tracked performance. These boundary patterns strengthen the framework by making it falsifiable rather than merely descriptive.

11 Implications for L2 Listening Resources, Assessment, and Pedagogy

For listening assessment and experimental design, the results indicate that matching materials only on topic, length, vocabulary, syntax, and speech rate is insufficient. Two segments can be similar on those descriptors while differing substantially in how explicitly they anchor referents and relations or how much rhetorical reconstruction they require. Reporting semantic anchoring and rhetorical inference load can therefore improve stimulus balancing, clarify the construct represented by test items, and reduce the risk of attributing stimulus-driven difficulty entirely to learner ability.

For pedagogy, the findings argue for scaffolding rather than eliminating rhetorically demanding discourse. High rhetorical inference load is part of authentic listening: persuasion, evaluation, narrative stance, metaphor, ceremony, and culturally situated speech all depend on it.

Teachers can make the inferential architecture visible by previewing relevant frames, tracing perspective shifts, comparing literal and rhetorical formulations, and revisiting how parallelism, modality, fronting, or repetition contributes to stance. The stronger proficiency moderation at lower levels also suggests that such scaffolding may be most consequential for B1 and developing B2 listeners.

For resource development, annotation should preserve the distinction between source features and learner outcomes. A segment should not be labeled difficult because listeners performed poorly; instead, semantic anchoring, rhetorical inference load, profile shifts, and form foregrounding should be coded independently and then linked to accuracy, response time, load, confidence, or strategy data. Computational measures can be added as complementary layers, but the present correlations caution against substituting a single automatic index for human discourse analysis.

Reproducible listening resources should document primary material, annotation, computational preprocessing, task design, and data provenance as separable layers. Primary audio and transcripts should remain distinct from derived annotations so that coding can be corrected, extended, or reanalyzed without altering the source; time-aligned formats and explicit schemas are useful for maintaining this provenance (Bird & Liberman, 2001; Ide & Romary, 2004; Wittenburg et al., 2006).

A resource report should identify the source or corpus, language variety, speaker metadata where permissible, recording context, duration, audio format, and licensing terms. It should also state the selection and matching criteria, including genre, topic, proficiency target, length, speech rate, vocabulary, syntax, and topic familiarity where these are controlled. Transcription and segmentation conventions should be documented together with time alignment, boundary rules, disfluency treatment, and quality-control procedures.

The codebook should be reported with construct definitions, coding anchors, diagnostic subfeatures, examples and non-examples, borderline cases, uncertainty rules, annotator training, calibration procedures, reliability statistics, and adjudication practices. Automated features should be accompanied by software versions, tokenization and tagging decisions, normalization, windowing or smoothing choices, and manual audit procedures where relevant.

12 Boundary Conditions and Limitations

Several limitations qualify the findings. First, the item sample is small: only 24 listening segments were studied, six in each factorial cell. The controlled construction of the segments strengthens internal comparison but limits ecological generalizability to naturally occurring lectures, dialogues, narratives, or public discourse. Replication with larger and independently sampled sets of authentic, time-aligned materials is therefore essential.

Second, the primary models treated proportional accuracy as a continuous 0–1 outcome and used crossed random intercepts for listener and segment. This specification reflects the study's primary analysis; future work should test robustness with item-level binomial or ordinal generalized mixed models where appropriate and should consider theoretically justified random slopes. Third, the segment-level correlation analysis has $N = 24$ and includes multiple unadjusted tests; those coefficients are exploratory and should not be interpreted as stable estimates of unique predictive effects.

Fourth, the 0–4 annotation scores are operational descriptors, not validated psychometric scales. Krippendorff's α values from .723 to .800 show that trained annotators could apply the coding scheme with reasonable consistency, but reliability alone does not establish dimensionality, invariance, discriminant validity, or generalizability. Additional studies should compare the codes with independent expert judgments and established lexical, syntactic, cohesion, and acoustic measures.

Fifth, the framework and empirical materials are centered on English L2 listening, and the models reported here do not include prior knowledge, cultural familiarity, working-memory capacity, task variation, or acoustic degradation as measured moderators. These factors are likely to influence when anchoring is helpful and when rhetorical cues become costly. Cross-linguistic adaptation should therefore revise examples and decision rules rather than simply translate the existing labels.

Finally, semantic anchoring should not be equated with ease, and rhetorical inference load should not be equated with difficulty. Highly anchored discourse can remain difficult because of vocabulary, syntax, accent, speed, or task demands, while familiar metaphor, repetition, rhythm, and parallelism can support prediction and memory. The present results establish empirical relationships under the controlled study conditions; broader causal and

ecological claims require replication across authentic genres, tasks, learner populations, and languages.

13 Conclusion

This empirical study shows that L2 listening materials differ not only in vocabulary, grammar, duration, speech rate, accent, and topic, but also in how explicitly they anchor conceptual structure and how much rhetorical inference they require. Across 2,304 listener-by-segment observations, stronger semantic anchoring was associated with higher accuracy, faster responses, lower cognitive load, and less form-monitoring, whereas higher rhetorical inference load produced the opposite pattern.

The two dimensions were not redundant. Strong semantic anchoring substantially buffered the accuracy cost of rhetorically demanding input, but it did not eliminate the additional response time, cognitive load, or form-monitoring that rhetorical interpretation required. Proficiency also mattered: rhetorical costs for accuracy, time, and load decreased from B1 toward C1, indicating that advanced listeners were more resilient to inferentially demanding discourse.

The segment-level analysis reinforces the same conclusion from a measurement perspective. Human-coded anchoring and rhetorical indicators showed theoretically differentiated relationships with outcomes, while lexical entropy alone was essentially unrelated to performance. Listening difficulty is therefore better represented as a layered interaction among discourse support, rhetorical demand, listener proficiency, and task requirements than as a single scalar property of the text.

The practical objective is not to simplify L2 listening input artificially, but to make its discourse architecture visible, testable, and reportable. Incorporating semantic anchoring and rhetorical inference load into stimulus documentation can improve experimental control, assessment interpretation, pedagogical scaffolding, and resource reuse. Future studies should now test the framework with larger authentic corpora, richer moderator data, distribution-appropriate mixed models, and independent replication across languages and listening tasks.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The author declares that no funds, grants, or other support were received for the conduct of this study or the preparation of this manuscript.

References

- Alzahrani, A. (2024). Utility of Kolmogorov complexity measures: Analysis of L2 groups and L1 backgrounds. *PLOS ONE*, 19(4), e0301806.
<https://doi.org/10.1371/journal.pone.0301806>
- Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596. <https://doi.org/10.1162/coli.07-034-R2>
- Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4(11), 417–423. [https://doi.org/10.1016/S1364-6613\(00\)01538-2](https://doi.org/10.1016/S1364-6613(00)01538-2)
- Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1), 1–34. <https://doi.org/10.1162/coli.2008.34.1.1>
- Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33(1–2), 23–60. [https://doi.org/10.1016/S0167-6393\(00\)00068-6](https://doi.org/10.1016/S0167-6393(00)00068-6)
- Ehret, K., Berdicevskis, A., Bentz, C., & Blumenthal-Dramé, A. (2023). Measuring language complexity: Challenges and opportunities. *Linguistics Vanguard*, 9(s1), 1–8.
<https://doi.org/10.1515/lingvan-2022-0133>
- Ehret, K., & Szmrecsanyi, B. (2016). An information-theoretic approach to assessing linguistic complexity. In R. Baechler & G. Seiler (Eds.), *Complexity, isolation, and variation* (pp. 71–94). De Gruyter. <https://doi.org/10.1515/9783110348965-004>
- Ehret, K., & Szmrecsanyi, B. (2019). Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Research*, 35(1), 23–45. <https://doi.org/10.1177/0267658316669559>
- Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.
- Ferreira, F., & Patson, N. D. (2007). The good enough approach to language comprehension. *Language and Linguistics Compass*, 1(1–2), 71–83. <https://doi.org/10.1111/j.1749-818X.2007.00007.x>
- Field, J. (2008). *Listening in the language classroom*. Cambridge University Press.
- Fillmore, C. J. (1982). Frame semantics. In The Linguistic Society of Korea (Ed.), *Linguistics in the morning calm* (pp. 111–137). Hanshin.

- Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, 28(1), 55–75. [https://doi.org/10.1016/S0346-251X\(99\)00060-3](https://doi.org/10.1016/S0346-251X(99)00060-3)
- Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second Language Research*, 35(1), 3–21. <https://doi.org/10.1177/0267658318809765>
- Huetting, F. (2015). Four central questions about prediction in language processing. *Brain Research*, 1626, 118–135. <https://doi.org/10.1016/j.brainres.2015.02.014>
- Ide, N., & Romary, L. (2004). International standard for a linguistic annotation framework. *Natural Language Engineering*, 10(3–4), 211–225. <https://doi.org/10.1017/S135132490400350X>
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1), 23–62. <https://doi.org/10.1016/j.cogpsych.2010.02.002>
- Karalık, T., & Merç, A. (2025). Linguistic, cognitive, and affective components of L2 listening comprehension: A multidimensional model. *Porta Linguarum*, 43, 273–293. <https://doi.org/10.30827/portalin.vi43.30647>
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), 163–182. <https://doi.org/10.1037/0033-295X.95.2.163>
- Knill, K., Nicholls, D., Gales, M. J. F., Qian, M., & Stroinski, P. (2025). *The Speak & Improve Corpus 2025: An L2 English speech corpus for language assessment and feedback*. Cambridge University Press & Assessment. <https://doi.org/10.17863/CAM.114333>
- Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE.
- Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1), 32–59. <https://doi.org/10.1080/23273798.2015.1102299>
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.
- Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford University Press.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177. <https://doi.org/10.1016/j.cognition.2007.05.006>
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

- Mekheimer, M. (2026). Measuring information density in interlanguage through entropy analysis. *Scientific Reports*, 16, 18672. <https://doi.org/10.1038/s41598-026-56853-3>
- Mekheimer, M. A., & Fageeh, A. I. (2025). Prioritizing information over grammar: A behavioral investigation of information density and rhetorical discourse effects on EFL listening comprehension. *Discover Education*, 4, Article 24. <https://doi.org/10.1007/s44217-025-00411-y>
- Paquot, M., König, A., Stemle, E. W., & Frey, J.-C. (2024). The Core Metadata Schema for Learner Corpora (LC-meta): Collaborative efforts to advance data discoverability, metadata quality, and study comparability in L2 research. *International Journal of Learner Corpus Research*, 10(2), 280–300. <https://doi.org/10.1075/ijlcr.24010.paq>
- Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4), 329–347. <https://doi.org/10.1017/S0140525X12001495>
- Rost, M. (2016). *Teaching and researching listening* (3rd ed.). Routledge.
- Saito, K., Uchihara, T., Takizawa, K., & Suzukida, Y. (2025). Individual differences in L2 listening proficiency revisited: Roles of form, meaning, and use aspects of phonological vocabulary knowledge. *Studies in Second Language Acquisition*, 47(1), 26–52. <https://doi.org/10.1017/S027226312300044X>
- Satori, M. (2021). Effects of working memory on L2 linguistic knowledge and L2 listening comprehension. *Applied Psycholinguistics*, 42(5), 1313–1340. <https://doi.org/10.1017/S0142716421000345>
- Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Blackwell.
- Sun, K., & Wang, R. (2021). Using the relative entropy of linguistic complexity to assess L2 language proficiency development. *Entropy*, 23(8), Article 1080. <https://doi.org/10.3390/e23081080>
- Trujillo, J. P., & Holler, J. (2025). Multimodal information density is highest in question beginnings, and early entropy is associated with fewer but longer visual signals. *Discourse Processes*, 62(2), 69–88. <https://doi.org/10.1080/0163853X.2024.2413314>

- van den Heuvel, H., Iskra, D., Sanders, E., & de Vriend, F. (2008). Validation of spoken language resources: An overview of basic aspects. *Language Resources and Evaluation*, 42, 41–73. <https://doi.org/10.1007/s10579-007-9049-1>
- van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.
- Vafaei, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition*, 42(2), 383–410. <https://doi.org/10.1017/S0272263119000676>
- Vandergrift, L., & Baker, S. C. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, 65(2), 390–416. <https://doi.org/10.1111/lang.12105>
- Wang, G., Wang, H., & Wang, L. (2022). Kolmogorov complexity metrics in assessing L2 proficiency: An information-theoretic approach. *Frontiers in Psychology*, 13, 1024147. <https://doi.org/10.3389/fpsyg.2022.1024147>
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556–1559). European Language Resources Association.
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2), 162–185. <https://doi.org/10.1037/0033-2909.123.2.162>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [L2ListeningAppendicesRawDataAnalysesanonymized.pdf](#)
- [L2ListeningAppendicesRawDataAnalysesanonymized.tex](#)