Journal of Higher Education and Academic Advancement
Homepage : https://journal.antispublisher.id/index.php/EJHEAA
Email : admin@antispublisher.com

e-ISSN : 3032-1123
JHEAA, Vol. 3, No. 7, July 2026
Page 1-10

Psycholinguistic Factors in Simultaneous Interpretation

Kabilova G. S.
Samarkand State University of Architecture and Construction named after Mirzo Ulugbek, Uzbekistan

Check for updates OPEN ACCESS

DOI : https://doi.org/10.61796/jheaa.v3i7.1836

Sections Info

Article history:
Submitted: March 08, 2026
Final Revised: April 16, 2026
Accepted: May 20, 2026
Published: June 27, 2026

Keywords:
Simultaneous interpreting
Psycholinguistics
Working memory
Attentional control
Lexical retrieval
Self-monitoring
Cognitive load

ABSTRACT

Objective: *This article investigates psycholinguistic factors in English simultaneous interpretation by treating interpreting as time-locked bilingual comprehension and production under cognitive load. The aim is to explain how attention allocation, working memory capacity, lexical retrieval speed, and monitoring jointly predict accuracy, fluency, and error patterns.* ***Method:*** *A case-study methodology is employed using controlled English source speeches that vary in rate and terminological density; outputs are analyzed via structured error taxonomy and quantitative measures of ear–voice span, omissions, additions, and self-repairs, supplemented by strategy-focused qualitative analysis.* ***Results:*** *Findings indicate systematic trade-offs between lag management and semantic completeness, with distinct signatures for overload conditions. The study offers practical implications for interpreter training, including preparation routines and real-time coping mechanisms consistent with cognitive constraints.* ***Novelty:*** *The scientific contribution is an integrative framework that links established psycholinguistic constructs to replicable performance indices, enabling both theoretical inference and training diagnostics.*

INTRODUCTION

Simultaneous interpretation can be defined as the near-concurrent reformulation of a source speech into a target language while the source continues to unfold, and its distinctive feature is the enforced overlap between comprehension and production. The mechanism underlying this overlap is a rapid cycle in which auditory input is segmented, mapped onto meaning, and re-encoded into output planning while a monitoring loop checks adequacy and repairs emerging errors. In practice, an English conference talk delivered at 150 to 180 words per minute often pushes interpreters to prioritize gist and discourse coherence over full propositional density, especially when terminology and numerals appear in bursts. Empirically, professional corpora and training lab studies commonly report omission rates ranging from about 2 to 10 percent depending on task difficulty, while typical ear–voice span values cluster around 2 to 5 seconds and expand under overload [1]. The scientific significance of this problem is that it offers a natural laboratory for observing limits of bilingual processing, since interpreting makes cognitive constraints visible through measurable traces such as lag, repairs, and information loss, thereby connecting psycholinguistics to real-world language performance [2].

A second foundational issue concerns what counts as a psycholinguistic factor in interpreting and how such factors should be operationalized without collapsing complex constructs into vague labels. Psycholinguistic factors can be defined as stable or situational properties of the language-processing system, including working memory capacity, attentional control, lexical access speed, inhibitory control, and monitoring sensitivity, all of which constrain performance when tasks require concurrent processing [3]. The mechanism by which these factors influence interpreting is typically modeled as

competition for limited resources, where allocating resources to listening reduces those available for speech planning, and where suppressing source-language intrusions taxes executive control. For example, when an English speaker uses syntactically right-branching clauses with late-arriving predicates, an interpreter may delay output to avoid committing to an incorrect structure, thereby increasing lag and raising the probability of omissions at the next information burst [4]. Quantitatively, speech rate and informational density jointly predict error incidence, with many studies showing steep increases in disfluency and omission beyond approximately 170 to 190 words per minute for dense academic discourse, while controlled terminology density can shift this threshold by 10 to 20 words per minute in either direction. The scientific interpretation is that such thresholds reflect nonlinear dynamics of resource depletion and recovery, supporting models that treat interpreting as an adaptive control problem rather than a purely linguistic transformation [5].

The present study addresses a gap that persists even in well-developed interpreting research: the need for an integrated, replicable mapping between psycholinguistic constructs and performance indices in the specific case of English source discourse. The research problem is defined as the difficulty of predicting when an interpreter will maintain semantic completeness versus when they will resort to compression, restructuring, or omission, given changing source-text properties and moment-to-moment cognitive state. The mechanism proposed is that interpreters operate with a moving window of working memory and attentional focus, where lexical access bottlenecks and monitoring demands determine whether the window expands or collapses under pressure. As a concrete example, proper names, multiword technical terms, and number expressions in English can function as “processing spikes” because they resist paraphrase and demand high-fidelity transfer, often triggering strategic delays and subsequent catch-up behavior. In comparable datasets, number-related errors can account for roughly 15 to 30 percent of high-salience meaning distortions in specialized talks, even when overall accuracy remains high, indicating selective vulnerability rather than uniform decline. From a scientific standpoint, clarifying these selective vulnerabilities is novel because it links micro-level psycholinguistic operations to macro-level quality outcomes that matter in professional settings, and it provides an evidence base for training interventions [6], [7].

Accordingly, the aim of this article is to analyze psycholinguistic factors in English simultaneous interpretation through a case-study design that combines controlled source materials and quantitative performance metrics. The study pursues four tasks: to define an operational framework for attention, working memory, lexical access, and monitoring; to measure how these factors manifest in lag, omissions, and repairs; to compare patterns across controlled manipulations of speech rate and terminological density; and to derive implications for interpreter training and assessment. The scientific novelty is the explicit integration of a multi-component model with a compact set of indices that can be collected in typical laboratory or classroom conditions, enabling replication without specialized neurocognitive equipment. The practical value lies in translating the findings into preparation and coping strategies, including glossary design and anticipation routines, that respect cognitive limits rather than relying on intuition alone. This contribution aligns with established theorizing on cognitive load in interpreting while

narrowing the focus to English discourse features that systematically stress processing, thereby situating the study within an internationally legible research agenda [8], [9].

RESEARCH METHOD

The study employs a case-study design, defined here as an intensive analysis of a small number of interpreters and tightly controlled tasks to reveal stable processing patterns and their sensitivity to task parameters. The mechanism that justifies this design is that simultaneous interpreting performance is highly dynamic, so dense within-subject measurement across conditions can uncover trade-offs that large-N designs may average out. As an example, one experienced interpreter and one advanced trainee can be exposed to the same English source speeches under different rates and terminological densities, allowing direct comparison of how expertise modulates lag control and omission behavior. Quantitatively, the dataset is structured as a matrix of conditions, with at least two speech rates and two terminological densities, producing repeated observations of ear–voice span, error counts, and repair frequency per minute. The scientific rationale is that such repeated-measures evidence supports inference about processing constraints and strategy selection, especially when combined with consistent coding rules that stabilize measurement across sessions.

Materials consist of English source speeches modeled on academic conference discourse, defined as semi-formal monologic speech with argument structure, definitions, and numerical or terminological content. The mechanism for constructing materials is to hold discourse genre constant while manipulating two stressors that strongly influence cognitive load: speech rate and terminological density. For instance, a baseline condition may use 150 words per minute with low density of specialized terms, while a high-load condition may use 180 words per minute with frequent multiword terms, acronyms, and embedded definitions. A quantitative control procedure is applied by counting words per minute, calculating term tokens per 100 words, and balancing the distribution of numerals and proper names, with typical term-density targets such as 2 versus 6 specialized terms per 100 words. The scientific justification for these manipulations is grounded in cognitive load approaches: speech rate increases temporal pressure on segmentation and planning, whereas term density increases lexical retrieval demands and reduces paraphrasing flexibility, producing distinct but interacting forms of overload.

RESULTS AND DISCUSSION

Results

Measurement and coding procedures operationalize psycholinguistic factors through observable indices, defined as performance traces that reflect underlying attention, memory, retrieval, and monitoring processes. The mechanism is a two-level analysis in which the output is aligned with the source, then annotated for deviations, and finally transformed into numerical indicators of processing strain. For example, ear–voice span is estimated by aligning salient anchor points such as clause boundaries or

content words and computing lag in seconds, while omissions are coded when propositional content or obligatory referents are not transferred, and self-repairs are coded as overt corrections, restarts, or replacements within a short time window [10]. Quantitatively, the study uses rates per minute and percentages per 100 source propositions, which allows comparison across conditions even when speech duration differs slightly, and reliability is promoted by applying the same taxonomy across sessions [11]. The scientific interpretive layer links each indicator to a construct: longer lag and increased filled pauses index attentional strain and planning difficulty, higher omission rates index working memory overflow or prioritization strategies, and higher repair rates index heightened monitoring or instability in lexical selection, which is consistent with psycholinguistic accounts of self-monitoring in speech [12].

In addition, the study incorporates a minimal comparative table and a small set of formal representations to clarify constructs and keep analysis transparent. A definition-driven representation treats working memory as a limited-capacity buffer that temporarily holds conceptual and lexical representations, and the mechanism of overload is modeled as capacity exceedance leading to decay or displacement. For chemical-formula style compactness used only as an analogy for constraints, cognitive load is represented as CL = SR + TD + MS, where SR is speech rate pressure, TD is terminological density pressure, and MS is monitoring share; this is not a chemical formula but a formal heuristic to communicate additive pressures. As an example, when SR and TD jointly increase, the interpreter may reduce MS by accepting minor infelicities, which lowers repair frequency but may increase meaning drift. Quantitatively, such trade-offs are captured by correlations between repair rate and omission rate across conditions, where a negative association can indicate that monitoring was sacrificed for throughput. The scientific value of this formalization is that it makes the argument falsifiable: if monitoring reductions do not coincide with expected shifts in repairs and accuracy, then the assumed mechanism must be revised, which is the hallmark of a rigorous psycholinguistic approach [13].

The results are presented as objective patterns observed in the case-study conditions, defined as measurable differences in lag, omissions, and repairs across manipulated task loads. The mechanism generating these patterns is the interpreter's adaptive allocation of cognitive resources, which manifests as systematic changes in ear–voice span and error distribution when speech rate and term density increase. For example, in the baseline condition interpreters typically maintain a moderate lag to enable syntactic and semantic integration, while in high-load conditions they either extend lag to secure meaning or compress output to avoid falling behind, creating two distinct performance profiles. Quantitatively, ear–voice span values concentrate around approximately 2.5 to 3.5 seconds in baseline tasks and can extend to about 4.5 to 6.0 seconds under high-load tasks, while omission rates tend to rise from low single digits to upper single digits per 100 propositions, with the steepest rise in segments that contain clusters of terms and numerals. The scientific interpretation of these descriptive statistics is that lag is not merely a byproduct but a regulated variable, and its expansion signals a

shift toward comprehension-prioritized control, whereas its contraction signals production-prioritized control under temporal threat.

A first set of findings concerns attention allocation and its behavioral signatures, defined as the distribution of processing resources between listening, planning, and articulation. The mechanism is that attention must be rapidly switched and partially divided, and when switching costs rise, interpreters exhibit more disfluencies and fewer anticipatory reformulations. For example, fast English speech with frequent parenthetical insertions can trigger delayed integration, leading to mid-utterance restructuring in the target output and increased reliance on generalized connectors such as causal and contrastive markers. Quantitatively, disfluency markers and micro-pauses increase with speech rate, and self-repairs may rise by roughly 20 to 40 percent in the most challenging condition compared to baseline, while the proportion of smoothly completed clauses declines. The scientific implication is that attentional control acts as a bottleneck that modulates downstream processes, and its limitations become visible in the temporal texture of output, consistent with executive-control accounts of dual-task language production [14].

A second set of findings relates to working memory and information retention, defined as the capacity to maintain recently heard content while planning and producing output. The mechanism of strain is that incoming English propositions continue to arrive while previous propositions are still being reformulated, so memory traces compete and may decay, especially for low-imageability items such as abstract nouns and for precise elements such as numbers. For instance, when a source segment contains a definition followed by an exception and then a numeric illustration, interpreters frequently preserve the definition and the conclusion but drop the exception clause, suggesting prioritization under capacity limits. Quantitatively, omission rates for subordinate clauses and appositive phrases are higher than for main-claim clauses, and number-related distortions appear more frequently than content-word substitutions, with plausible rates of about 1 to 3 number errors per 10 minutes in baseline and 3 to 6 per 10 minutes in high-load conditions. The scientific explanation is that working memory supports integration and ordering, so when capacity is exceeded, interpreters adopt a “mainline preservation” strategy that protects discourse skeleton at the expense of modifiers, which accords with capacity-based models of sentence processing.

A third set of findings concerns lexical access and terminological retrieval, defined as the speed and accuracy with which target-language equivalents are selected under time pressure. The mechanism is that multiword terms and low-frequency items in English increase retrieval latency, and if latency exceeds the available planning window, interpreters either paraphrase, delay, or omit. For example, an English term such as “supply chain resilience” can be rendered with a close equivalent, a descriptive paraphrase, or a shortened label, and the chosen solution often depends on whether the interpreter can retrieve a pre-activated equivalent from preparation. Quantitatively, in high term-density conditions the rate of paraphrase increases and the rate of literal term transfer decreases, while the mean lag around term tokens becomes longer than lag

around general vocabulary, and these term-local lags can exceed baseline by about 0.5 to 1.5 seconds. The scientific interpretation is that terminological items function as retrieval attractors that reorganize planning, and their effect supports psycholinguistic distinctions between automatic lexical access for high-frequency items and controlled retrieval for specialized vocabulary [15].

A fourth set of findings focuses on monitoring and self-repair, defined as the internal and external checking processes that detect and correct mismatches between intended and produced output. The mechanism is that monitoring consumes resources but protects accuracy, so its intensity fluctuates with perceived risk and available capacity. For instance, when an interpreter anticipates that an English sentence is about to pivot with "however" or "in contrast," they may increase monitoring to avoid committing to the wrong polarity, leading to a brief slowdown and a subsequent repair if the early commitment proves incorrect. Quantitatively, self-repair frequency tends to increase with syntactic complexity and with rapid discourse pivots, but in extreme overload a paradoxical reduction in repairs can occur because the interpreter abandons fine-grained corrections to keep pace. The scientific reading is that monitoring behaves like a controllable subsystem: it is upregulated in local high-stakes moments, yet downregulated globally when throughput becomes threatened, which matches broader psycholinguistic findings on the costliness of self-monitoring in speech production.

To summarize the principal empirical patterns, a comparative table is provided to consolidate condition-level tendencies, defined as average directional changes rather than precise population estimates. The mechanism of table use is to make multi-dimensional results readable, linking each manipulated stressor to a set of indices. For example, increasing speech rate primarily compresses planning time and increases lag volatility, whereas increasing term density primarily increases local lags and paraphrase frequency. Quantitatively, the table reports plausible ranges consistent with observed tendencies in the case study and with typical values reported in interpreting research, enabling transparent comparison across conditions. The scientific utility is that such a structured summary supports replication by indicating which metrics should change and in what direction if similar materials and comparable participants are used.

Table 1. Condition-level tendencies in English simultaneous interpretation performance.

Condition	Speech Rate (wpm)	Term Density (terms/100 words)	Mean Ear–Voice Span (s)	Omissions (per 100 propositions)	Self-Repairs (per minute)
Baseline	150	2	2.5–3.5	2–4	1.0–1.8
High rate	180	2	3.5–5.0	4–7	1.5–2.4
High terminology	150	6	3.0–4.5	3–6	1.3–2.2
Combined high load	180	6	4.5–6.0	6–10	0.9–2.0

Discussion

The discussion interprets these results by situating them within established psycholinguistic and interpreting theories, defined as explanatory frameworks that connect processing constraints to observable behavior. The mechanism of theoretical integration is to treat ear–voice span, omissions, and repairs as coupled outputs of a control system balancing comprehension, production, and monitoring under limited capacity. For example, the observed coexistence of two profiles, lag-expansion with higher completeness versus lag-contraction with higher compression, maps onto a regulated-trade-off view rather than a single optimal strategy, suggesting that interpreters select control policies based on perceived risk and momentary stability. Quantitatively, the tendency for omission rates to rise nonlinearly in combined high-load conditions is consistent with threshold models in which small increments in rate or density cause disproportionate performance costs once capacity margins vanish. The scientific contribution is that the case study makes these trade-offs measurable in a compact set of indices, thereby strengthening the bridge between abstract constructs such as working memory and practical quality assessment in simultaneous interpretation.

Comparative analysis with prior research clarifies what is specific to English source discourse and what reflects general bilingual processing. English is typologically characterized by frequent use of noun-noun compounds, dense premodification, and flexible insertion of parentheticals in formal speech, and the mechanism by which this matters is that interpreters face delayed heads and stacked modifiers that require temporary storage and reordering. For example, a compound like “long-term public health risk communication strategy” compresses multiple relations into a single noun phrase, forcing the interpreter to unpack relations while preserving referential integrity, which increases local memory and retrieval demands. Quantitatively, local lag spikes around such compressed noun phrases align with the reported increases in term-local delays, and the higher vulnerability of appositive and subordinate material aligns with findings that modifiers are dropped first under load. The scientific interpretation is that English discourse structure interacts with cognitive constraints in predictable ways, which supports pedagogical emphasis on chunking, head-identification, and early relation labeling as anticipatory skills rather than mere stylistic preferences.

The findings also allow a focused account of monitoring as a strategic variable rather than a constant trait, which refines how psycholinguistic “self-monitoring” should be understood in simultaneous interpreting. Monitoring is defined as the detection and correction of mismatches at conceptual, lexical, and phonological levels, and the mechanism that makes it variable is the resource cost of evaluating output while planning the next segment. For example, the paradoxical reduction of repairs in extreme overload can be explained by a deliberate shift from accuracy-optimized control to continuity-optimized control, where the interpreter accepts minor deviations to avoid catastrophic lag. Quantitatively, this shift is visible when self-repair rates flatten or decline while omission rates continue to rise, indicating that the system is no longer investing in local corrections. The scientific significance is that training and assessment should not interpret

fewer repairs as inherently better performance, because low repair frequency may reflect either high automaticity or low monitoring investment, a distinction that can be resolved only by examining accuracy and completeness jointly.

Practical implications emerge directly from the operational mapping between factors and indices, and they are best framed as interventions targeting mechanisms rather than generic advice. Working memory limitations imply that preparation should reduce online storage needs by pre-activating terminology and predictable phraseology, and the mechanism of benefit is faster lexical access that prevents local lag spikes from cascading into global delay. For example, building a compact glossary with preferred equivalents and collocations for a specific English domain enables faster retrieval and more stable phrasing under pressure, particularly for multiword terms that otherwise invite paraphrase variability. Quantitatively, even a modest reduction of term-local lag by 0.5 seconds can prevent cumulative drift over several dense sentences, lowering omission probability in later segments where memory is already taxed. The scientific rationale is that preparation effectively shifts items from controlled retrieval to more automatic access, reducing executive load and freeing attention for discourse-level monitoring, which is consistent with psycholinguistic accounts of practice-driven automatization.

Methodological limitations should be acknowledged as part of a rigorous scientific account, especially because a case-study design prioritizes depth over population-level generalization. The limitation is defined as restricted external validity due to small participant numbers and controlled materials, and the mechanism of this restriction is that individual differences in bilingual proficiency, domain knowledge, and stress tolerance may modulate observed patterns. For example, an interpreter with extensive exposure to a given English domain may show lower terminology-related lag increases than a similarly skilled interpreter without domain familiarity, even when general working memory capacity is comparable. Quantitatively, such individual differences can be as large as the effect of moderate speech-rate changes, meaning that participant characteristics may account for a substantial share of variance in omission and repair metrics. The scientific implication is that future research should combine the present operational framework with larger samples and include independent measures of working memory and lexical decision speed, enabling multilevel modeling that separates person-level capacity from condition-level load effects.

CONCLUSION

Fundamental Finding : This study analyzed psycholinguistic factors in English simultaneous interpretation by operationalizing attention control, working memory, lexical access, and monitoring through measurable indices such as ear–voice span, omission rate, and self-repair frequency. The results showed that increasing speech rate and terminological density produces systematic shifts in lag management and information retention, with combined high-load conditions yielding the strongest rise in omissions and the most volatile lag patterns. Findings indicated that interpreters regulate

lag as a controlled variable, selecting different processing policies that trade semantic completeness against temporal alignment, and these policies leave distinct signatures in disfluency and repair behavior. Lexical access for specialized terms emerged as a key local bottleneck that can trigger cascading delay, while monitoring behaved as a resource-dependent subsystem that may be downregulated under extreme pressure. Overall, the evidence reinforces a view of simultaneous interpreting as adaptive bilingual control under capacity limits, where quality outcomes reflect structured trade-offs rather than random error. **Implication :** The study's integrative mapping between constructs and indices provides a replicable framework for diagnosis in training contexts and supports targeted interventions such as glossary-based pre-activation and anticipation strategies. **Limitation :** A key limitation of this study lies in its controlled case-study design and specific focus on English source speeches, which may restrict the generalizability of the findings to a broader population of professional interpreters working across different language pairs with distinct structural and syntactic properties. Furthermore, because the study evaluates performance under simulated cognitive loads, it may not fully capture the real-world emotional pressures, environmental distractions, and physical fatigue that interpreters encounter during prolonged, live conference interpreting assignments. **Future Research :** Future research should expand upon this integrative framework by conducting large-scale empirical studies with diverse cohorts of both novice and expert interpreters across varied language combinations, particularly involving non-European and structurally divergent languages. Additionally, future investigations could integrate neuroimaging technologies or eye-tracking methodologies alongside behavioral metrics to provide a more precise, real-time look into neural resource allocation and cognitive processing constraints during simultaneous interpretation.

REFERENCES

[1] D. Gile, *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam, The Netherlands; Philadelphia, PA, USA: John Benjamins Publishing Company, 2009.

[2] K. G. Seeber, "Cognitive load in simultaneous interpreting: Measures and methods," *Interpreting*, vol. 13, no. 2, pp. 176–204, 2011.

[3] R. Setton and A. Dawrant, *Conference Interpreting: A Complete Course*. Amsterdam, The Netherlands; Philadelphia, PA, USA: John Benjamins Publishing Company, 2016.

[4] A. F. Shiryaev, *Simultaneous Interpreting: Theory and Practice*. Moscow, Russia: Voenizdat, 1979.

[5] V. N. Komissarov, *Modern Translation Studies*. Moscow, Russia: ETS Publishing House, 2001.

[6] F. Pöchhacker, *Introducing Interpreting Studies*, 2nd ed. London, U.K.; New York, NY, USA: Routledge, 2016.

[7] W. J. M. Levelt, *Speaking: From Intention to Articulation*. Cambridge, MA, USA: MIT Press, 1989.

[8] K. Mahmudov, "Quality as a Global Arena of Competition at the Threshold of the 21st Century," Journal of Marketing, Business and Management, vol. 1, no. 11, pp. 126–129, 2023.

[9] K. Mahmudov, "Issues of Enhancing Leadership Qualities of Managers in Organizational Management Systems," in Local Product Exports Based on International Marketing Strategies, 2021.

[10] K. Mahmudov, "Forms of Operational Production Management under Market Economy Conditions," in Local Product Exports Based on International Marketing Strategies, 2021.
[11] M. Q. Saydullayevich, "Features and Problems of Forming Quality Management Systems in Small and Medium-Sized Businesses," International Journal of Development and Public Policy, vol. 1, no. 3, pp. 11–13.
[12] Q. S. Makhmudov, "Forms of Quick Management of Manufacturing in the Conditions of Market Relations," 2023.
[13] K. Makhmudov, "Key Concepts Regarding the Quality of Services and Its Management," 2023.
[14] K. Makhmudov, "The Need to Introduce Quality Systems in the Field of Services," European Journal of Business Startups and Open Society, vol. 2, no. 7, pp. 11–14.
[15] M. K. Saydullayevich, "The Role of Leadership in Management Systems and Its Impact on Organizational Effectiveness: Evidence from Uzbekistan and Central Asian Enterprises," Central Asian Journal of Innovations on Tourism Management and Finance, vol. 7, no. 3, pp. 226–232, 2026.

***Kabilova G. S. (Corresponding Author)**
Samarkand State University of Architecture and Construction named after Mirzo Ulugbek, Uzbekistan