OPEN ACCESS

Conceptual Foundations in Language Research

Conceptual Foundations in Language Research, 1, 260105 (2026)
https://doi.org/10.65553/CFLR.260105

Toward a Working Memory–Chunking Interface Framework of Language Processing and Learning

Zhisheng (Edward) Wen*

Division of Languages and Communication, College of Professional and Continuing Education, The Hong Kong Polytechnic University, Hong Kong SAR, China

Abstract

Working memory (WM) is routinely invoked to explain why some learners comprehend, speak, and learn a language more successfully than others. Yet capacity-based accounts leave a basic question unanswered: capacity for what? Four places in memory hold very different amounts of language, depending on whether each holds a sound, a word, or a whole multiword expression. In this position paper, I argue that chunking supplies the missing link between a limited WM and open-ended language use. I propose the Working Memory - Chunking Interface Framework (WMCIF), in which WM constrains how linguistic units are formed and maintained online, while chunking, in turn, reduces the effective load on WM by reorganising many elements into fewer, experience-dependent units. The framework distinguishes nominal capacity from effective processing load, specifies what counts as a chunk, and separates three timescales on which WM and chunking interact. Five propositions follow, with testable predictions for L1 - L2 processing, acceptability, and language aptitude. For teachers and learners, the framework implies that instruction may not enlarge memory, but it can change what occupies it. WMCIF shifts the foundational question from *how* much capacity learners have to *what counts as a unit* for a particular learner in a particular task.

Keywords: working memory; chunking; effective processing load; second language acquisition; language aptitude

Submission date: October 30, 2026; Revision date: November 12, 2026; Online: November 18, 2026

CC BY NC

(corresponding author: edward.wen@cpce-polyu.edu.hk)

GLODAL

Introduction: Capacity for What?

Human working memory (WM) is limited, yet human language is combinatorially rich and, in principle, open-ended. Within a few seconds, language users have to hold on to sounds and words, resolve dependencies, build propositions, keep track of context, plan and articulate utterances, and revise earlier interpretations. The standard answer has been that WM provides the temporary workspace for all this, so learners with more storage or stronger executive control should cope better with demanding language tasks (Baddeley et al., 1998; Engle, 2002). I have worked within this tradition for many years (Wen, 2012, 2016; Schwieter & Wen, 2022), and I still regard it as indispensable. But I now think it is incomplete.

The gap can be restated in one simple question: *capacity for what?* To say that WM can hold a handful of units does not tell us whether those units are phonemes, words, multiword expressions, constructions, propositions, or discourse routines. Cowan's (2001) estimate of about four chunks in the "focus of attention" already presupposes that information has been organised into units. Yet in language research and second language acquisition (SLA), WM is usually measured as a learner trait, separately from the linguistic material it works on (Teng, 2026; Wen, 2016). A span score is correlated with comprehension, production, or learning, while the size of the units being held stays in the theoretical background.

The empirical record supports this concern. WM-language relationships are real but modest and uneven: a meta-analysis of L2 reading found a significant but modest association that varied with measurement (In'nami et al., 2022), and in L2 writing, proficiency rather than WM predicted several dimensions of performance (Manchón et al., 2023). Such variation is puzzling if WM is a fixed container, but it is expected if the influence of WM depends on how a learner's knowledge has already packaged the material to be held.

This paper addresses that gap conceptually. My central claim is that chunking changes the functional burden placed on limited memory. An everyday analogy helps: we cannot easily make a suitcase bigger, but we can pack it better, folding items together so that more fits into the same space. In the same way, WM constrains how chunks are formed and how readily they are available online; once formed, chunks in turn change what WM has to handle. I call this interface the **Working Memory–Chunking Interface Framework (WMCIF)**. It does not replace established WM, formulaic-language, or usage-based accounts; rather, it specifies an interface among them that has so far been missing. Its contribution is to shift the foundational question of WM-language research from *how much capacity* learners have to *what counts as a unit* for a particular learner, language, and task.

Conceptual Framework: The Working Memory–Chunking Interface

Nominal Capacity versus Effective Processing Load

Nominal capacity refers to the resources estimated by a WM task, while **effective processing load** refers to the number, complexity, and accessibility of the functional units that must be maintained and integrated during a particular activity. The two are related but not interchangeable. Figure 1 illustrates the difference.

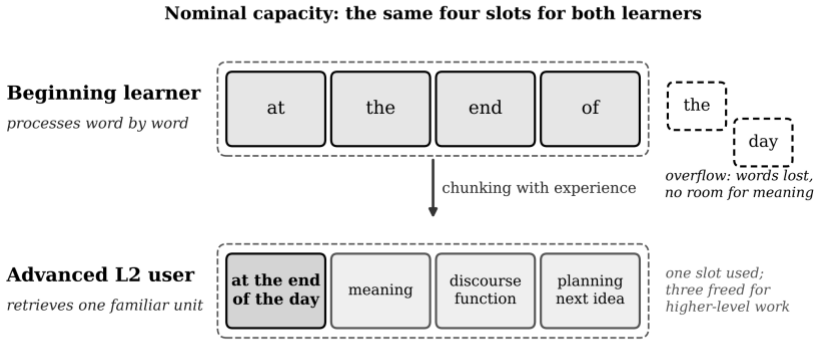


Figure 1. Nominal capacity versus effective processing load. Both learners have the same nominal capacity (here, four slots; cf. Cowan, 2001). For the beginning learner, "at the end of the day" needs more slots than are available; for the advanced user it occupies one, leaving capacity free for meaning, discourse function, and planning.

Consider the English multiword expression "at the end of the day." A beginning learner may initially encounter six separate words, whereas a more advanced L2 user may process the sequence as a highly familiar expression, retrieve it with less internal assembly, and allocate attention to its discourse function. The Chinese idiom 不可思议 (bù kě sī yì, "unbelievable") works the same way: a learner may first process it morpheme by morpheme but later access it as one lexical unit. In both cases the input is unchanged, but the functional unit has changed.

The same distinction applies above the lexical level. "The reporter filed the story" places modest demands on memory, whereas "The reporter who the senator that the president criticised contacted filed the story" is still grammatical but requires several incomplete dependencies to be held open at once. Grammaticality has not changed; the effective load has.

This distinction explains why correlational designs of the form WM score → language outcome are informative but incomplete. A more adequate formulation is **WM capacity × representational organisation → effective processing demand → performance.** Under this formulation, weak or variable correlations are exactly what we should expect. Expertise research points the same way: skilled performers use retrieval structures and chunks to access large amounts of domain knowledge without any expansion of raw short-term storage (Ericsson & Kintsch, 1995; Gobet et al., 2001). Language is a particularly important case, because every new linguistic experience changes the units that can be recognised, predicted, and retrieved.

What Counts as a Chunk

The term *chunk* is useful because it is broad, but that breadth risks circularity: if every frequent sequence is called a chunk and every processing advantage is attributed to chunking, the construct explains nothing. For present purposes, I define a chunk as **an experience-dependent representation in which multiple informational elements become sufficiently integrated to function as a coordinated unit for a particular cognitive operation** (Ellis, 2017). Four properties follow: a chunk is learned rather than given; it integrates several elements; it is defined by function rather than form; and its status depends

on the task. Chunks thus include, but are broader than, formulaic sequences (Wray, 2002), whether these are described as frequency-driven units (Ellis, 2012) or fluency-supporting routines (Wood, 2010). Chunks can occur at different linguistic scales–phonological sequences, morphemes, lexical forms, collocations, multiword expressions, constructional frames, and discourse routines–and chunking can refer to a process as well as its product.

To avoid circularity, WMCIF specifies what counts as evidence for chunking rather than mere co-activation. A chunk is inferred when (a) recall or processing of constituent elements is conditional on sequence-level retrieval, (b) the unit can be selectively accessed or adapted without full decomposition, and (c) behavioural or neural markers distinguish it from item-by-item processing. WMCIF also separates *online chunking*, the temporary grouping of incoming elements even when no stored unit yet exists, from *stored chunk retrieval*. With repeated successful grouping, temporary groupings consolidate into stored units. Chunking is therefore not only the retrieval of ready-made formulas but a developmental process linking momentary grouping, consolidation, and later retrieval.

L2 learning offers a revealing test. Adult learners bring entrenched L1 patterns that can bias segmentation, and they typically have smaller and less stable repertoires of multiword units (Wang & Christiansen, 2024). Wei et al. (2026) showed that, under divided attention, L1 speakers of Chinese kept binding words into complete multiword expressions, whereas L2 learners recalled them as smaller pieces. Crucially, proficiency was associated with chunk *size*, while WM was associated with chunk *number*. WM thus affects how many units are available; experience affects how much each unit contains.

The WMCIF Cycle and Its Three Timescales

WMCIF describes a cycle (Figure 2). Language experience supplies recurring patterns; attention and prior knowledge determine which regularities are registered; chunk formation and retrieval reorganise them into functional units; these units lower the effective load of subsequent processing; the freed resources can be directed to meaning, monitoring, or more complex constructions; and new use generates further experience. Individual WM capacity constrains every pass through the cycle, but it does not determine the representational grain size in isolation.

The coupling between WM and chunking operates on three timescales, and keeping them apart is what makes the framework precise. On the *online* timescale (milliseconds to seconds), WM limits the temporary grouping of incoming input, which must be recoded rapidly before it is lost (Christiansen & Chater, 2016). On the *task* timescale (minutes), task demands determine which chunks are recruited and which must be built on the spot (Skehan, 2022). On the *developmental* timescale (weeks to years), repeated online grouping consolidates chunks in long-term memory, changing the load of future processing (Gobet et al., 2001). The relationship is therefore not simply WM → performance but **WM ↔ chunking → language processing and learning**: capacity constrains learning, while learning reorganises the effective burden on capacity.

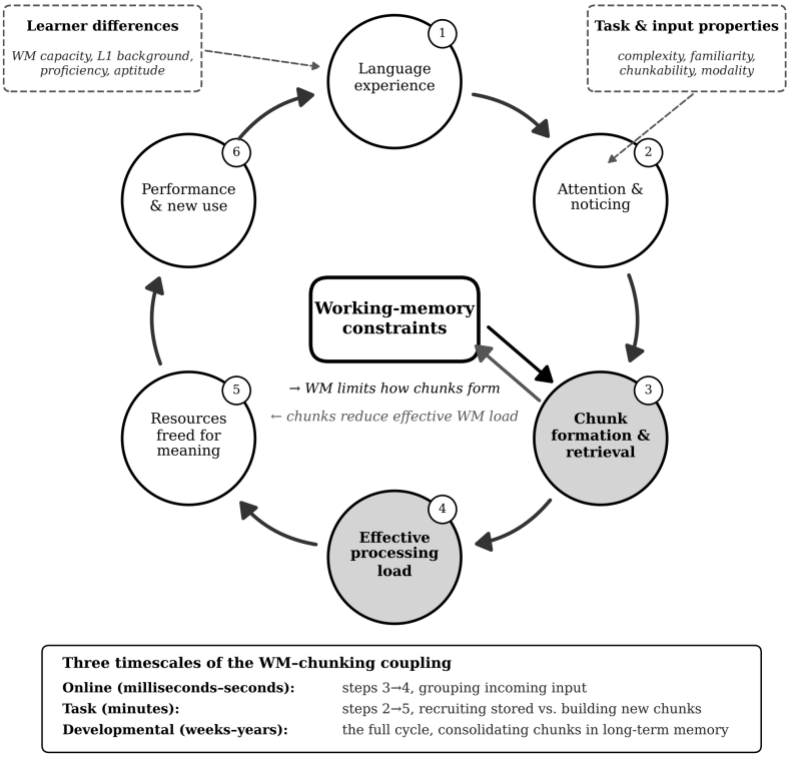


Figure 2. The Working Memory–Chunking Interface Framework (WMCIF). Experience supports chunk formation and retrieval; chunking lowers effective processing load; and WM constraints, in turn, shape how chunks are formed. Learner differences and task/input properties moderate the cycle, which operates on online, task, and developmental timescales.

Five Propositions

Taken together, WMCIF rests on five propositions. Each is a claim about the WM-language relationship, and each generates testable predictions.

Proposition 1: WM capacity is unit relative. WM capacity cannot be interpreted independently of the units over which it is estimated. Four phonemes, four words, four constructions, and four propositions represent very different amounts of linguistic information. A limit expressed in chunks therefore does not fix the amount of linguistic information that can be processed: two learners who hold the same number of units may differ greatly in effective capacity because their units differ in size and accessibility. In this sense, learning need not enlarge working memory in order to enlarge what working memory can accomplish. It follows that if chunking and raw storage are partly separable, improvements in chunking should be possible without corresponding increases in measures of raw storage.

Proposition 2: Chunking regulates effective WM load. The same surface input imposes different WM demands depending on how efficiently it can be chunked. I use *effective linguistic WM* for the amount of linguistically meaningful information a user can maintain and integrate under current task conditions. It is an analytical term, not a new memory module.

Unlike Engle's (2002) domain-general executive attention, it is domain-specific; unlike Cowan's (2001) focus of attention, it allows the content of each chunk to vary with experience; and unlike Just and Carpenter's (1992) capacity theory, it locates differences in representational organisation rather than in the capacity pool alone. It is also not a store or interface like the awareness buffer of the multicomponent model (Allen et al., 2026), and, unlike Jackendoff's (2002) treatment of WM within phonological, syntactic, and conceptual processing, it focuses specifically on how experience-dependent chunking alters processing load.

Proposition 3: Chunk formation is experience dependent. Chunks emerge through experience with frequency, transitional relations, semantic coherence, syntactic regularity, pragmatic function, and contextual recurrence. Exposure alone is not sufficient, however. A sequence becomes useful only when it can be retrieved or constructed reliably and adapted to communicative goals. WMCIF gives primacy to no single cue; what matters is their convergence. It predicts that semantic coherence should dominate when pragmatic function is clear, and frequency should dominate when semantic transparency is low. L1 entrenchment can therefore either support or obstruct L2 chunking (Wang & Christiansen, 2024).

Proposition 4: WM and chunking are dynamically coupled. WM limits what can be held while a chunk is being formed (Baddeley et al., 1998): a learner must retain enough of a recurring sequence to detect relations among its parts. Once consolidated, the chunk lowers later WM demand. If chunking is the channel through which experience reorganises WM load, chunking measures should partially mediate the WM–performance relationship, which should weaken, but not vanish, once chunking is controlled.

Proposition 5: Individual differences reflect capacity and representational efficiency. Two learners with similar WM scores may perform differently because one has developed more efficient chunks; conversely, a lower-capacity learner may match a higher-capacity one by retrieving larger units (Table 1).

Table 1. *A 2 × 2 matrix of nominal working-memory capacity and chunking efficiency as partially independent sources of language performance.*

	Lower chunking efficiency	Higher chunking efficiency
Lower nominal WM	High processing vulnerability	Potential compensation through integrated units
Higher nominal WM	Capacity-dependent performance	Most efficient performance under familiar conditions

The matrix yields a falsifiable prediction that, to my knowledge, has not been tested directly: WM should predict performance more strongly for low-chunkability than for high-chunkability material. This extends stage-based accounts such as the P/E model (Wen, 2016; Wen & Skehan, 2021). If WM predicts equally well across chunkability conditions, Proposition 5 is undermined.

What WMCIF Adds to Existing Accounts

WMCIF is best read as a bridge rather than a rival. Capacity-based models, from the multicomponent model (Allen et al., 2026) to Cowan's (2001) embedded-processes account, explain *why* processing is limited but are largely silent on what fills the limited space.

Usage-based and formulaic-language accounts (Ellis, 2012; Wray, 2002) explain how experience builds larger units but rarely specify the memory limits under which those units form. Stage-based aptitude models such as the P/E model (Wen, 2016; Wen & Skehan, 2021) link WM components to phases of acquisition but treat the input itself as constant. WMCIF connects these strands with one mechanism: limited resources constrain processing, while distributional experience reorganises the units through which those constraints are expressed. Proficiency, on this view, is partly representational change.

Theoretical and Practical Implications

Theoretical implications

Reading WM evidence differently. Span scores partly conflate capacity with knowledge-dependent chunking, especially across languages and proficiency levels. Individual-difference models should therefore include chunk repertoire and task demand alongside WM components. Chunking efficiency can be captured by three kinds of measure, each with a trade-off: recall of multiword expressions under divided attention (Wei et al., 2026) separates chunk size from chunk number but mainly captures receptive familiarity; corpus-based measures of phraseological breadth estimate chunk repertoire from production but conflate frequency with representation; and online indices index processing pressure rather than the chunk itself. A study testing WMCIF should combine one measure of each kind and treat convergence as evidence that chunking has been captured rather than assumed.

L1 and L2 processing. L1 and L2 users may share a cognitive architecture while differing in representational history. WMCIF predicts that L1–L2 differences should appear not only in accuracy or span but in the grain size and stability of the units recruited online. Longitudinal research should therefore ask whether changes in chunk size mediate relations among exposure, proficiency, and fluency.

Formulaic language and proficiency. Formulaic sequences let larger stretches of language be handled with less internal assembly, so formulaicity may moderate WM effects (Skehan, 2022). Yet their benefit depends on flexibility. A memorised expression that cannot be adapted may reduce load in one context while offering little generative value elsewhere. Research should therefore distinguish among frequency, familiarity, retrieval fluency, internal predictability, and flexible recombination, rather than using formulaicity as a single explanation.

Grammaticality and acceptability. Acceptability judgements reflect both grammatical knowledge and the cost of building a representation. Two online indices operationalise distinct aspects of this cost: Momentary Chunk Number, the number of chunks active at a given processing moment (Lu & Wen, 2022), captures chunking efficiency, while Mean Dependency Distance, the average linear distance between syntactically dependent words (Liu, 2008), captures integration cost. WMCIF therefore predicts that acceptability judgements should correlate more strongly with WM for high-load structures, such as long dependencies, than for local ones, a prediction that distinguishes it from purely competence-based accounts.

Language aptitude. WMCIF extends WM-based views of aptitude (Wen, 2016; Wen & Skehan, 2021; Williams et al., in press). Aptitude may include not only the capacity to hold and manipulate material but also the efficiency with which recurring input becomes chunked. Phonological WM should matter most when input is unfamiliar and little can be chunked; executive WM and chunking efficiency should matter more as larger units become available. WM-based aptitude measures should thus predict learning more strongly for low-chunkability

input, and measures of chunking efficiency should add predictive value beyond WM scores; if they do not, the claim that aptitude includes representational efficiency is undermined. This view also cautions against deficit interpretations: a lower span is not a fixed ceiling.

Human–AI boundary conditions. Large language models show chunk-like sensitivity to distributional patterns without human WM architecture, so similar outputs need not imply similar mechanisms. Because AI can redistribute rather than simply reduce cognitive demands (Deng, 2026), WMCIF predicts that AI support helps learners most when it removes extraneous load without replacing the repeated processing through which chunks are formed. Where AI does replace that processing, it may impair learning even as it improves immediate performance. Human-AI comparison should therefore be used to locate boundary conditions, not as a proxy for human cognition.

Practical implications for Teachers and Learners

I keep the practical implications deliberately modest (cf. Skehan, 2012). Instruction cannot be assumed to expand general WM capacity, but it can change what occupies it. Four principles follow.

- **Teach in units, not only in words.** Present useful multiword sequences (e.g., *on the other hand, I was wondering if*) as wholes, and help learners notice where one unit ends and the next begins, particularly where L1 segmentation habits mislead.
- **Recycle for consolidation.** Because chunks consolidate through repeated successful use, the same sequences should reappear across tasks, moving learners gradually from supported analysis toward fluent retrieval and recombination.
- **Manage load without removing difficulty.** When material is new and hard to chunk, learners rely heavily on raw WM, so pre-task planning, repetition, and familiar content can free capacity. As chunks develop, support can be withdrawn so that desirable difficulty remains.
- **Assess flexible use, not reproduction.** Assessment should examine whether learners can adapt chunks to new contexts, not merely reproduce memorised expressions (Wen, 2018).

A brief illustration shows how these principles combine. Before an opinion-sharing speaking task, a teacher might pre-teach five stance chunks (*in my view, the thing is, to be fair, that said, at the end of the day*), have students practise them in a short, low-pressure exchange, and then repeat the main task with a new topic a week later. In WMCIF terms, the first round relies on online chunking under heavy load; repetition moves the chunks toward stored retrieval; and the freed capacity can be spent on ideas and argument rather than on assembling each phrase word by word. Feedback can then target whether students *adapt* the chunks, for example *that said, I still think…*, rather than whether they recite them.

For learners, the message is encouraging: a modest memory span is not destiny. Building a repertoire of well-practised chunks, and using AI tools to support rather than replace that practice, enlarges what the same memory can accomplish.

Conclusion

WM remains indispensable to language research, but after more than three decades, capacity alone is no longer a sufficient explanation (Wen et al., 2022). The primary contribution of this paper is conceptual. By distinguishing nominal capacity from effective processing load,

defining what counts as a chunk, and separating three timescales of interaction, the WMCIF captures a dynamic coupling in which WM constrains the online formation of linguistic units,

while repeated successful grouping consolidates them into chunks that, in turn, reorganise the burden placed on WM. Language proficiency may therefore depend less on expanding the workspace than on changing what occupies a place within it.

Future research should test four predictions that set WMCIF against capacity-only accounts:

- **Mediation.** Chunking should partially mediate the relationship between WM and language performance, whereas capacity-only accounts predict no mediation.
- **Interaction.** WM should predict performance more strongly for low-chunkability than for high-chunkability material, whereas capacity-only accounts predict a uniform main effect.
- **Segmentation.** L1 and L2 users should segment the same input into units of different grain size, with differences visible in online measures rather than only in accuracy or span.
- **Dissociation.** Training-induced changes in chunking should be neurally dissociable from changes in raw storage, whereas capacity-only accounts predict a common substrate.

Testing these predictions will require longitudinal, intervention, and neurocognitive designs that measure chunking directly rather than assume it. Findings contrary to these predictions would count against WMCIF, with the mediation and interaction predictions carrying the most weight. A further question is whether the same pressures, over longer timescales, favour linguistic patterns that are easy to chunk. On this view, the framework is not a replacement for capacity-based accounts but a specification of the units over which capacity is exercised.

Declarations

Declaration of competing interest. The authors declare no competing interests.

CRediT authorship contribution statement.

Conceptualization, writing, review, and editing

Declaration of generative AI use.

Gemini 3.5 has been used for checking language flow for easier understanding.

References

Allen, R. J., Baddeley, A. D., & Hitch, G. J. (2026). Awareness as the heart of working memory. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2026.04.001

Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*(1), 158–173. https://doi.org/10.1037/0033-295X.105.1.158

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39*, e62. https://doi.org/10.1017/S0140525X1500031X

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Deng, X. (2026). Working memory in technology-enhanced language learning: A systematic review from interactive to AI-mediated contexts. *Frontiers in Psychology, 17*, 1758104. https://doi.org/10.3389/fpsyg.2026.1758104

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics, 32*, 17–44. https://doi.org/10.1017/S0267190512000025

Ellis, N. C. (2017). Chunking in language usage, learning and change: "I don't know." In M. Hundt, S. Mollin, & S. E. Pfenninger (Eds.), *The changing English language: Psycholinguistic perspectives* (pp. 113–147). Cambridge University Press. https://doi.org/10.1017/9781316091746.006

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2), 211–245. https://doi.org/10.1037/0033-295X.102.2.211

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*(6), 236–243. https://doi.org/10.1016/S1364-6613(00)01662-4

In'nami, Y., Hijikata, Y., & Koizumi, R. (2022). Working memory capacity and L2 reading: A meta-analysis. *Studies in Second Language Acquisition, 44*(2), 381–406. https://doi.org/10.1017/S0272263121000267

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122

Liu, H. T. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9*(2), 159–191. https://doi.org/10.17791/jcs.2008.9.2.159

Lu, B., & Wen, Z. (2022). Short-term and working memory capacity and the language device: Chunking and parsing complexity. In J. W. Schwieter & Z. Wen (Eds.), *The Cambridge handbook of working memory and language* (pp. 393–418). Cambridge University Press.

Manchón, R. M., McBride, S., Mellado Martínez, M. D., & Vasylets, O. (2023). Working memory, L2 proficiency, and task complexity: Independent and interactive effects on L2 written performance. *Studies in Second Language Acquisition, 45*(3), 737–764. https://doi.org/10.1017/S0272263123000141

Schwieter, J. W., & Wen, Z. (Eds.). (2022). *The Cambridge handbook of working memory and language*. Cambridge University Press.

Skehan, P. (2012). Language aptitude. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 38–52). Routledge.

Skehan, P. (2022). Working memory and L2 speaking tasks. In J. W. Schwieter & Z. Wen (Eds.), *The Cambridge handbook of working memory and language* (pp. 635–655). Cambridge University Press.

Teng, M. F. (2026). *Working memory in vocabulary acquisition.* Routledge.

Wang, S. Y., & Christiansen, M. H. (2024). Chunking in the second language: Implications for language learning and teaching. *Language Teaching Research Quarterly, 44*, 84–106. https://doi.org/10.32038/ltrq.2024.44.09

Wei, Y., Wang, J., & MacWhinney, B. (2026). Chunking words into multi-word expressions: Exploring chunking ability in immediate recall among second language learners of Chinese. *Applied Psycholinguistics, 47*, e1, 1–23. https://doi.org/10.1017/S0142716425100441

Wen, Z. (2012). Working memory and second language learning. *International Journal of Applied Linguistics, 22*(1), 1–22. https://doi.org/10.1111/j.1473-4192.2011.00290.x

Wen, Z. (2016). *Working memory and second language learning: Towards an integrated approach*. Multilingual Matters. https://doi.org/10.21832/9781783095735

Wen, Z. (2018). Using formulaic sequences to measure task performance: The role of working memory. In M. J. Ahmadian & M. del P. García Mayo (Eds.), *Recent perspectives on task-based language learning and teaching* (pp. 29–52). De Gruyter Mouton. https://doi.org/10.1515/9781501503399-003

Wen, Z., & Skehan, P. (2021). Stages of acquisition and the P/E model of working memory: Complementary or contrasting approaches to foreign language aptitude? *Annual Review of Applied Linguistics, 41*, 6–24. https://doi.org/10.1017/S0267190521000015

Wen, Z., Teng, M. F., & Han, L. (2022). Working memory models and measures in language and bilingualism research: Integrating cognitive and affective perspectives. *Brain Sciences, 12*(6), 729. https://doi.org/10.3390/brainsci12060729

Williams, J. N., Wen, Z., & Xue, Y. (in press). Working memory in SLA. In S. Gass, A. Mackey, & J. Behney (Eds.), *The Routledge handbook of second language acquisition* (2nd ed.). Routledge.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications*. Continuum.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Author bio

Zhisheng (Edward) Wen is Associate Head of the Division of Languages and Communication at the College of Professional and Continuing Education, The Hong Kong Polytechnic University. His research focuses on working memory, language aptitude, and cognitive individual differences in language learning, alongside translation and translanguaging. He has authored and edited more than ten books, including *Working Memory and Second Language Learning* (Multilingual Matters, 2016), *Language Aptitude: Advancing Theory, Testing, Research and Practice* (Routledge, 2019), *The Cambridge Handbook of Working Memory and Language* (Cambridge University Press, 2022), and *Memory in Science for Society* (Oxford University Press, 2023). His work has appeared in *Language Teaching and Annual Review of Applied Linguistics*. He is Co-founding Editor-in-Chief of *Individual Differences in Language Education: An International Journal.*
Email: edward.wen@cpce-polyu.edu.hk
ORCID: https://orcid.org/0000-0001-9041-6920