<LRH>On Umbrellas and Omnibuses

<RRH>On Umbrellas and Omnibuses

<ArtType>IN PERSPECTIVE

<AT>On Umbrellas and Omnibuses: A Response to Open Peer Commentaries

<AU>Bram Bulté[a], Alex Housen[a], Gabriele Pallotti[b]

<AF>[a]Vrije Universiteit Brussel, [b]University of Modena and Reggio Emilia

<AN>

[optional additional author notes here]

Correspondence concerning this article should be addressed to Alex Housen, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. Email: Alex.Housen@vub.be

The handling editor for this manuscript was Scott Jarvis.

<KWG>Keywords complexity; difficulty; language development; language proficiency

Abstract

This response to the commentaries on our conceptual review article on structural complexity and learning difficulty in second language acquisition (SLA) clarifies the scope and objectives of our framework, the challenges of defining and measuring complexity and difficulty, and the broader relevance of our proposal for both empirical research and theory-building in SLA, including its relationships to frameworks such as CAF.

The eight insightful commentaries on our position paper provide a valuable opportunity to clarify our conceptualization of structural complexity and learning difficulty in second language acquisition (SLA). Encouragingly, there appears to be broad agreement among the

commentators on the value of and need for striving for terminological and conceptual clarity, distinguishing constructs, and selecting valid measures. Since the range of perspectives and comments raised in the commentaries exceeds what we can fully address here, we will focus on some key themes that emerged recurrently across the responses. These include clarifying the scope and objectives of our framework, the challenges of precisely defining, distinguishing and operationalizing complexity and difficulty, and the broader relevance of our proposal for both empirical research and theory-building in SLA, including its relationships to other frameworks such as CAF. We begin by addressing theoretical issues before turning to questions of measurement and operationalization.

A new theory of SLA?

Granfeldt (2024) wonders to what extent our proposal may contribute to SLA theorizing, given that we do not commit to any specific theories of linguistics or learning in our position paper. While the authors each have their own linguistic and learning theories of choice, we clearly stated that our approach does not depend on any specific theory. By using terms like "property/transition theory", or referencing authors such as Gregg and O'Grady, we did not imply wholesale acceptance of the theoretical frameworks these authors advocate. Rather, we attempted to express our ideas in ways that show their compatibility with several theoretical approaches. There are, however, several ways in which we believe our discussion may be relevant for theory construction and assessment, three of which are discussed below: (a) striving for conceptual clarity, (b) recommending more careful linguistic description, and (c) calling for a research programme for a relatively new theoretical construct, difficulty.

Conceptual clarification: how many umbrellas and how large should they be?

A central argument of our position paper is that “complexity” has been employed in SLA research as an umbrella term encompassing an overly broad range of phenomena. To address this issue, we introduced “difficulty” to refer to what others have referred to as “cognitive/relative complexity”, or even complexity in general—a terminological choice that can understandably lead to confusion. Encouragingly, most commentators appear to support this fundamental distinction, recognizing it as an important step toward greater theoretical clarity. Clearly delineating these constructs and assigning them distinct labels enables, amongst others, more precise investigation of the relationship between linguistic structural complexity and learning and processing difficulty, a key theoretical concern.

We do not contend that the use of umbrella terms is inherently problematic; grouping and categorizing related phenomena is essential for theory-building. Indeed, certain constructs are valuable precisely because they unify multiple dimensions under a coherent principle—an idea reflected in the notion of “composite constructs” in the philosophy of science. However, such categorization implies striking the right balance: too narrow constructs run the risk of overlooking commonalities, while overly broad characterizations may result in incoherent constructs. Determining the appropriate scope and granularity of such constructs — the size of the umbrella — involves a careful evaluation of whether a given construct represents a specific instantiation of a broader concept or constitutes a distinct, independent dimension.

In our mind, difficulty is a dimension that cannot be reduced to, or subsumed under, complexity: one is a potential effect, the other a potential cause; one has to do with human cognitive processing, the other with the description of linguistic systems. Referring to the Complexity-Accuracy-Fluency (CAF) framework, we suggested adding appropriateness as an additional, independent dimension, yielding CAFDA. Appropriateness has to do with how linguistic structures are interpreted within communicative contexts — a structure may be more or less appropriate regardless of whether it is more or less complex or difficult. In other

words, complexity has to do with language per se, difficulty with the cognitive processing of language, and appropriateness with its social interpretations and implications.

In their commentary, Verspoor and Rosmawati (2024) suggest adding idiomaticity and cohesion/coherence to the CAF triad (yielding CAFIC). We agree that all these dimensions concern important aspects of language acquisition and use, yet, following Ockham's recommendation not to multiply entities without necessity, we prefer keeping the acronym as short as possible, and add new dimensions only if they are not traceable to extant ones. For example, as also pointed out by Verspoor and Rosmawati, idiomaticity could be conceived of as a subdimension of appropriateness, and we would add that coherence, too, can be seen as contributing to a text's appropriateness. Therefore, in this response we propose a slight revision to our original suggestion by further shortening CAFDA to CAFD (Complexity, Appropriateness, Fluency, Difficulty). One could argue that Appropriateness is a more general term subsuming Accuracy among other dimensions — being grammatically accurate (and what is 'accurate' depends itself on the context and what one implicitly or explicitly sets as the norm) is just one way of producing appropriate texts, like selecting certain sociolinguistic variants or complexity levels.

Another important conceptual distinction concerns the notion of development. Many scholars, including we, are interested in how language features develop over time, and one of the main reasons for assessing constructs such as complexity and fluency is to track their development from the initial to the most advanced stages of language learning. Again, in order to study the relationship between two constructs, it is important to define them independently from one another. Therefore, any definition of complexity that subsumes or makes reference to development is, in our view, flawed. Paquot (2024), for example, argues that various measures grouped under the term "sophistication" are useful for describing L2 production and assessing L2 development and proficiency and should therefore be included

under the umbrella. We concur that such measures, including lexical frequency and multi-word unit usage, effectively track interlanguage development. However, we caution against equating later-acquired features with greater complexity. While empirical research demonstrates that more complex structures are often mastered at advanced stages of acquisition, this does not justify labelling all later-acquired elements as inherently more complex. Furthermore, we find the use of the term "sophistication" in SLA to be insufficiently precise. It is often ambiguously defined and serves as a catch-all category for measures that do not fit under elaboration or diversity but are somehow linked to "advanced", technical, academic, or native-like formal language use. In such cases, terms like "more advanced" or "more appropriate" offer greater clarity.

Finally, Szmrecsanyi and Dubois (2024) point out the importance of studying optionality and variation in their relation to complexity and difficulty, and we agree that these are important phenomena, too. The complexity of some structures does indeed vary across genres, registers, developmental levels, languages and dialects, and some of this variation is optional (what Pallotti, 2015, calls "stylistic variation"), some is mandatory, as is the complexity of noun and verb inflection in certain languages as opposed to others.

Recommending more careful linguistic description

Contrary to Biber et al. (2024)'s claim, we advocated for rigorous linguistic analysis and precise definition of linguistic units, despite not adhering to a specific theoretical framework or analytic approach. Our decision to forgo a particular grammatical theory was deliberate, aiming to ensure the broad applicability of our arguments across theoretical and descriptive systems. While this approach limited the level of detail of our descriptions of linguistic units and measures, our focus was on general principles rather than theory-specific units and measures.

We also want to clarify that the exclusion of certain units, features, dimensions etc. from our overview of "core measures" does not imply that these are flawed, useless or should be avoided. Rather, in our opinion, they do not have to do with complexity. That is, complexity analysis does not equal linguistic description at large, nor should any kind of linguistic description be equated with an analysis of complexity, a point we will return to later. Finally, we acknowledge Lu's (2024) observation that the selection and operationalization of linguistic units also depend on the language under study. While we aimed to include units and measures applicable across linguistic systems, we recognize that some operationalizations may be more suited to certain (particularly European) languages than to others, such as Chinese. This limitation reflects both gaps in our knowledge and the socio-historical context of prior research.

A research programme for a new theoretical construct: difficulty

A final way in which our contribution could aid theory-building in SLA is by calling for the development of an organic research programme on the causes and effects of difficulty in (second) language acquisition and use. The potential of the difficulty construct as a starting point for a new research agenda lies in that it enables the formulation of clear, directional predictions and the testing of falsifiable hypotheses regarding the relationships among the various causes and effects of difficulty. While our concept of difficulty has received less commentary than complexity, the critiques it has attracted are more fundamental. Several commentators (e.g., Paquot, Granfeldt, Verspoor & Rosmawati) acknowledge the importance of distinguishing difficulty from complexity to enhance clarity and coherence in the field, but also point to challenges in its current conceptualisation. We recognize that difficulty remains a relatively new and theoretically underdeveloped construct in SLA research compared to complexity, necessitating further refinement through continued research and debate.

Granfeldt (2024) particularly questions the usefulness of difficulty as a theoretical construct, given that SLA research has already addressed specific notions such as salience, frequency and transparency, and has posited, and in some cases even demonstrated, their impact on the development of linguistic structures. In Granfeldt's view, grouping them under the broad umbrella term *difficulty* adds little value as it itself lacks explanatory power and risks becoming as conceptually ambiguous as complexity, which would contradict our own stated objectives. We argue that this view is too strong. In our conceptualization, difficulty is not a spurious, incoherent construct, as it has a unitary definition with one single meaning, namely the (cumulative) amount of cognitive activity involved in the acquisition and development of a linguistic element, from its first perception in the input to, ultimately, its storage and retrieval as a robust representation in declarative or procedural memory. Importantly, salience, transparency, acquisitional timing etc. are not equivalent to, nor dimensions of, difficulty in the way that, for example, length constitutes a dimension of complexity. Difficulty can nonetheless still be understood as a multidimensional construct, encompassing various sources or causes, effects, and operationalizations that can be meaningfully grouped as they all contribute to the overarching notion of difficulty by either (i) directly influencing the cognitive effort required for processing and acquiring linguistic items or (ii) reflecting their impact on the rate and outcome of language development. We therefore propose difficulty as a unifying framework for coherently interpreting these causal factors and their effects on language learning and development.

In order to avoid circular reasoning, we defined difficulty narrowly as the cognitive resources required to process a given linguistic structure, without including developmental timing. The question is whether this narrow definition is both desirable and viable or if a slightly broader conceptualization is necessary. For example, the relationship between frequency and difficulty, as defined by us, is admittedly tenuous. Items that are less frequent have fewer

opportunities to be encountered in the input and thus to be processed by a language user, but the amount of cognitive processing needed to learn each of these items is not necessarily greater. Thus, frequency and other constructs such as dispersion/spacing potentially have an impact on developmental timing that is independent from processing difficulty. It may also be claimed, and has indeed been shown (e.g. Ellis, 2002; Gries & Divjak, 2012; for a critical review, see Baayen et al., 2016), that rarity and dispersion in the input and output impact the processing cost of individual items, so that these factors may be seen as having an indirect effect on difficulty *qua* processing.

Several commentators also criticized the exclusion of individual differences (including the learner's first language background) from our treatment of difficulty. We excluded these factors for the purpose of our position paper, but definitely not in principle. Clearly, many more potential causal or contributing factors — context-related and individual-learner related — should be included when investigating overall difficulty (see Housen & Simoens, 2016, for a taxonomy of factors influencing overall learning difficulty), particularly for studying learning difficulty in one specific learning context by a specific learner or group of learners. In our paper, we only focused on those sources of difficulty that pertain to linguistic items and that potentially hold across language learners (all other things being equal). An important task for the research agenda we advocate will be to establish the relationships and interactions between linguistic-item related, context-related and learner-related sources/causes of difficulty. It may turn out that some of the factors that we discussed (e.g., salience, frequency, complexity) are less important overall than, or are ultimately overruled by, other factors that we did not discuss in our position paper.

From theory to measurement

Several commentators raised issues regarding the measurement of complexity and difficulty as discussed in our paper. In response, we emphasize that our selection of measures was based on multiple criteria, including construct validity, feasibility and ease of operationalization, and cross-linguistic applicability. In this section, we focus on four key issues specifically related to the construct validity and scope of complexity and difficulty measures.

Relevant complexity measures

Yasuda (2024) and Paquot (2024) argue that the relevance or usefulness of complexity measures varies depending on proficiency levels and genres. The key issue here is to clarify the specific aims for which a set of measures is considered relevant or useful. If the objective is to track the development of the complexity of learner productions within a particular register or genre, or to differentiate between learners of varying proficiency levels, then some measures may serve these practical purposes better than others. Our article, however, pursued a different goal, one oriented toward fundamental research rather than practical applications. We examined various measures theoretically aligned with our definitions of complexity and difficulty, ultimately identifying a small set suitable for cross-context comparison. Using different complexity measures in different contexts (registers, tasks, and proficiency levels) obscures comparability, making it difficult to determine whether and to what extent complexity varies across contexts.

Somewhat similarly, Biber et al. (2024) argue that the complexity measures we included fail to capture the full complexity of the English language and grammar. In turn, we fail to see how the individual rates of occurrence of specific syntactic forms/functions they advocate capture the complexity of a text (be it in terms of their elaboration or diversity). Such rates of occurrence may, when somehow considered together, provide a comprehensive linguistic

description of a text at the level of syntax, but their relationship to the construct of complexity, as defined by us and agreed on by Biber et al., is unclear. In this sense, it seems to us that Biber et al. are still using 'complexity' as an umbrella term for linguistic description at large. Is a text containing more finite adverbial clauses (e.g., *John left after Pete arrived*) per n words necessarily more complex than a text containing fewer of these clauses but with more prepositional phrases functioning as adverbial modifiers per n words (e.g., *John left after Pete's arrival*)? These texts are surely different, linguistically speaking, and they may be typical for certain genres or L2 proficiency levels, but we can't see how one may be said to be more or less complex than the other in purely structural terms.

The measurability of difficulty

Some commentators question the utility of the difficulty construct, arguing that it lacks direct measurability. The problem, which we acknowledge, is not that difficulty is a hidden substance that we are not yet able to measure, but rather that difficulty is a composite construct with several causes and effects. These causes and effect can, indeed, be measured individually, and in our paper we have reviewed several operationalizations of, for example, saliency, frequency, and transparency (as potential causes) and psycho-physiological processes or learning time (as potential effects). What remains to be determined in future investigations is whether and how it may be possible to aggregate these dimensions into a single, unified measure of difficulty. However, this would prove just as challenging as providing a single, unified measure of linguistic complexity, and more theoretical reflection and empirical investigations will be needed to address such fundamental questions.

Granularity

Biber et al. (2024) agree with our narrow definition of complexity but reiterate their earlier criticism of the omnibus measures we advocate. We maintain that coarse-grained, high-level holistic measures complement fine-grained measures —a position reinforced by Lu (2024), who remarks that this debate is fundamentally about granularity. In this context, Lu rightly observes that the classification of forms and functions proposed by Biber and colleagues is not maximally specific either. The distinction between holistic and fine-grained measures is essentially one of purpose, not validity or applicability. Holistic measures serve different analytical objectives than fine-grained measures but are not inherently flawed. Omnibuses are cheap and social, taxis are expensive and private, and each have their advantages and disadvantages. To extend Biber et al.'s biological analogy of measuring the complexity of a forest, biologists such as Dutilleul et al. (2022) have in fact demonstrated that mean tree length, mean branch length, or mean number of branches per tree can indeed provide valuable insights to the biologist or forester, including about the complexity of a forest. However, we do not claim that some measures can capture all relevant information across all levels of granularity; achieving a balance between specificity and comprehensiveness remains a key issue in any measurement framework.

An important point in this respect is that Biber et al.'s commentary focuses on measures of constitutional complexity at the syntactic level only[1]. They acknowledge that our proposed measures of syntactic organizational complexity, targeting the variety of syntactic structure and relations (e.g., MATTR of dependency relations or syntactic structures), effectively capture the number of different syntactic functions within a text, albeit holistically. While we did not expand on this in our position paper, developing more targeted measures for specific syntactic structures, functions, or relationships is entirely feasible (e.g., as a first step, by distinguishing between the nominal and clausal level). We encourage researchers to apply diversity measures (such as MATTR) to theoretically motivated subsets of syntactic forms

and functions, including those identified by Biber et al. or derived from any well-grounded grammatical analysis, regardless of the underlying syntactic framework. This would provide a more nuanced view of syntactic diversity within texts.

A matter of semantics?

The choice to exclude meaning-based measures from our proposed core set of complexity measures sounds controversial (Lu, 2024). Once again, we would like to stress that this does not mean that such measures should be avoided because they are flawed or useless. However, given our conservative and minimalistic approach, we excluded them among our core measures, as their application is often problematic—particularly in early learner varieties. In such cases, accurately determining the number of meanings conveyed by a word or grammatical structure is virtually impossible, and even assigning a form to a specific word class (noun, verb, adverb...) is not always straightforward (Klein, 1986). Thus, such measures would be applicable only to certain studies involving certain types of learners, notably more advanced ones, and possibly in rather controlled elicitation conditions, which would limit their generalizability across tasks and proficiency levels. However, we agree with Lu (2024) that sense-aware lexical measures, in as far as word sense disambiguation is concerned, are feasible in certain cases, in that it is "possible to pinpoint the meanings of most polysemous words in learner texts" (p. 4), with the added caveat that the texts analyzed by Lu and Hu (2022) were produced by upper-intermediate learners.

As far as difficulty is concerned, we argue that meaning and function could justifiably be incorporated into certain recommended measures of difficulty. Crucially, our paper proposed assessing the difficulty of learner texts not by the actual cognitive effort involved in their production by the learners but, rather, by evaluating "the difficulty of structures as they appear in texts" and "by averaging the difficulty scores of all relevant structures within a

text.” These difficulty scores are derived from multiple sources of evidence, most of which extend beyond the specific production of a given text. This includes measures based on dimensions empirically shown to cause difficulty, also assuming that findings from well-researched linguistic items/structures and languages can inform less-explored items/structures and languages. For instance, our paper highlighted the role of form-to-function transparency, primarily demonstrated in the acquisition of morphology in English and other languages, but likely a general factor influencing difficulty across languages and linguistic levels, including syntax and lexis. While we acknowledge the challenge of scoring and ranking linguistic items/structures based on form-function transparency, a systematic approach akin to frequency-based lexical difficulty assessment could be applied. This would involve calculating a transparency score by, for example, counting the distinct meanings *in the target language* of the lexical items and grammatical structures produced by the learner in a text, using reference dictionaries and grammars. A general form-meaning/function transparency score for a text could then be computed by averaging these scores across all items and structures present in the text.

Conclusion

The diverse perspectives offered by the eight commentaries on our position paper highlight both the challenges and significance of refining the constructs of linguistic complexity and learning and processing difficulty in second language acquisition. While there is broad agreement on the need for conceptual clarity, precise definitions, and valid measurement approaches, the responses also underscore several areas requiring further discussion and empirical investigation.

A key takeaway from these discussions is that complexity and difficulty must remain distinct yet interconnected constructs. Our framework aims to clarify their respective roles and scopes

in SLA research: complexity pertains to linguistic structure, while difficulty concerns cognitive processing by learners. While some commentators express skepticism about the necessity of difficulty as a theoretical construct, we maintain that it provides a coherent and testable framework for unifying multiple contributing factors under a single research agenda. The question of measurement remains central to advancing this agenda. We acknowledge that directly measuring difficulty remains a challenge, much like other composite constructs in science, but argue that meaningful insights can be gained by triangulating its causes and effects. The critiques related to the granularity and the suitability of different complexity measures reinforce our view that SLA research benefits from a combination of comparative, broad-scope measures and fine-grained linguistic descriptions.

Looking ahead, we welcome further theoretical and empirical research that builds upon and refines our proposed framework. Addressing the interplay between linguistic complexity, cognitive difficulty, and development and proficiency in SLA will require systematic cross-linguistic studies, improved operationalizations, and continued theoretical engagement. By fostering clearer distinctions and robust measurement strategies, we hope to contribute to a more precise and comprehensive understanding of (second) language development and its outcomes.

[1] Biber et al.'s (2024) criticism mainly targets length of T-unit, which incidentally does not feature amongst our core complexity measures, and is included among the non-core measures only "for the sake of comparability with previous research". They illustrate that length of T-unit yields identical scores for two example sentences that show widely varying clausal and phrasal structures, both in terms of their form and function. However, they fail to mention that the two sentences yield very different scores using two of our proposed core syntactic measures, namely mean phrases per clause and mean clauses per T-Unit. In contrast, most of the more fine-grained rates of occurrence of specific syntactic forms/functions they advocate, when considered individually, would yield a score of 0 for both of these sentences.

References

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiaology, 30*(11), 1174–1220.

Dutilleul, P., Mudalige, N., & Rivest, L. P. (2022). Learning how a tree branches out: A statistical modeling approach. *PloS one, 17*(9), e0274168. https://doi.org/10.1371/journal.pone.0274168

Ellis, N. C. (2002). Frequency effects in language processing and acquisition. *Studies in Second Language Acquisition, 24*(4), 143–188.

Gries, S.T. & Divjak, D. (2012). (Eds.). *Frequency Effects in Language Learning and Processing*. Mouton de Gruyter.

Housen, A., & Simoens, H. (2016). Introduction: Cognitive Perspectives on Difficulty and Complexity in L2 Acquisition. *Studies in Second Language Acquisition*, *38*(2), 163–175.

Klein, W. (1986). *Second Language Acquisition*. Cambridge University Press.

Lu, X., & Hu, R. (2022). Sense-aware lexical sophistication indices and their relationship to second language writing quality. *Behavior Research Methods, 54*(3), 1444–1460. https://doi.org/10.3758/s13428-021-01675-6