Bulté, B., Housen, A. and Pallotti, G. (2024), Complexity and Difficulty in Second Language Acquisition: A Theoretical and Methodological Overview. *Language Learning*.
https://doi.org/10.1111/lang.12669

Complexity and Difficulty in Second Language Acquisition: A Theoretical and Methodological Overview

Bram Bulté,[a] Alex Housen,[a] Gabriele Pallotti [b]
[a]Brussels Centre for Language Studies, Vrije Universiteit Brussel, [b]Department of Education and Human Sciences, University of Modena and Reggio Emilia

CRediT author statement–All authors contributed equally to this work.
A one-page Accessible Summary of this article in nontechnical language is freely available in the Supporting Information online and at https://oasis-database.org
Correspondence concerning this article should be addressed to Alex Housen, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. Email: Alex.Housen@vub.ac.be
The handling editor for this article was Scott Jarvis.

<ABS>
This article presents a theoretical review of and methodological guidelines for the study of two key notions in second language acquisition research, complexity and difficulty. The term *complexity* has gained considerable currency over the past decades and has taken on a wide range of meanings. We argue for a more restricted interpretation, focusing exclusively on formal, structural properties of linguistic items. The less employed term *difficulty* in our account refers to the cognitive costs associated with learning and using such items. On the basis of our theoretical definitions, we critically review measures operationalizing these constructs and discuss their strengths, limitations, and potential applicability to second language research, in order to establish a small set of measures to be used routinely in the interest of replicability and knowledge accumulation in the field. In addition, we discuss the relationship between complexity and difficulty and the associated notions of proficiency and development.
<KWG>Keywords complexity; difficulty; language development; language proficiency

<A>Introduction

<TXT>
The study of language complexity has both theoretical and applied relevance. Theoretical approaches to language complexity are typically motivated by the attempt to evaluate and refine theories of language structure, language evolution, or the human language processor, or to describe and compare linguistic structures across languages and varieties. Applied approaches to language complexity (e.g., in clinical and developmental linguistics, language testing and pedagogy) have been primarily motivated by the attempt to develop accurate, reliable, and objective metrics of language performance, proficiency, and development across contexts or conditions. Within the second language acquisition (SLA) literature, a constant complaint is that research on complexity often produces inconclusive results. Some of the most frequently cited reasons include conceptual ambiguity of the key notions and the

proliferation of measures, with little attention to issues such as their construct validity and redundancy, which in turn impedes knowledge accumulation in the field (Bulté & Housen, 2012; Norris & Ortega, 2009; Pallotti, 2015). In an attempt to promote both conceptual clarity and empirical rigor and, ultimately, foster knowledge accumulation, we present in this article a theoretical overview and a critical discussion of how complexity and difficulty have been defined and operationalized in applied linguistics (AL) and SLA research together with some concrete proposals for a more coherent systematization.

Some terminological clarifications are in order before we start our exposition. First, we will discuss the complexity and difficulty of linguistic objects. Linguistic objects can be conceived of at different levels of abstraction. On the one hand, there are the more abstract form-function pairings that make up the (theorized) language system (Saussurean *langue*) either as it is established in a linguistic community and thus possibly represented in published grammars and vocabularies or as the idiolect of an individual first language (L1) or additional language (Ln) user. We will refer to these abstract objects as "structures" or "items" interchangeably and depending on established conventions. For instance, it is more customary to speak of "syntactic structures" and of "lexical items." A (simplified) example of a syntactic structure would be the English passive, consisting of the pattern NP-Subject $_{\text{PATIENT}}$ + BE + PAST PARTICIPLE (+ by NP $_{\text{AGENT}}$). An example of a lexical item would be the pairing of the phonological form /bɔɪ/ (together with its graphological representation <*boy*>) and the meaning [+human, +male, –adult]. On the other hand, there are the concrete instantiations of these abstract structures and items in oral and written language production. For these, we will use the terms "forms" or "units", again interchangeably. An example of a syntactic form is the concrete passive sentence *The symphony was written by Mozart* as it appears in a written text. The form counterpart of the above-mentioned lexical item would be the graphic string *boy* as it occurs in writing or the phonetic sequence [bɔɪ] as it is uttered in spoken language. Finally, we use "text" to refer to oral and written language productions containing such forms and units. Although it is certainly possible (though not always easy) to observe and count different forms in a text produced by a language learner, establishing exactly what functions or meanings they express is often problematic. This is why our discussion on how to analyze learners' texts will be concerned with formal aspects only, as will be further explained and argued below. Second, we will be concerned with the complexity and difficulty of only individual linguistic forms and of the texts containing them, deliberately ignoring the question of the complexity of a whole linguistic system. In the last decades, there has been a lively debate in typological linguistics on the comparability and measurability of complexity in different languages (Baerman et al., 2017; Sampson et al., 2009). This approach is viable only in those cases where a system can be identified and circumscribed, for example when an exhaustive grammatical description is available. However, with unstable, idiosyncratic systems, such as learner varieties, it is virtually impossible to establish what the system is, as this can be inferred and approximated only on the basis of concrete linguistic productions (or performance on receptive tasks). Thus, in what follows we will be concerned with describing the complexity and difficulty realized in actual texts, not the theoretical complexity or difficulty entailed in principle by a linguistic (sub)system, such as German or verb inflection in German. Finally, for the sake of descriptive clarity, we propose to study complexity and difficulty at the level of different layers of language (i.e., the lexicon, syntax, morphology), as is customary in the second language (L2) literature, although we are aware that, in some cases, these layers are not always easy to tease apart. We also incorporate a brief discussion of more recent measures that target complexity and difficulty at the lexico-grammatical interface. Other levels and interfaces, of course, exist or can be conceived of, for instance, the morpho-phonemic or

syntax-pragmatic interfaces or linguistic phenomena at the suprasentential or discourse level, but for reasons of space we will not discuss these here.

This paper is structured as follows: The next section presents the state of the art of complexity and difficulty in SLA research, specifically in relation to the description of L2 productions. This is followed by a discussion of conceptual issues and theoretical definitions of the key constructs. In the section Operationalizing Linguistic Units of Analysis, we pay attention to the linguistic units used in the description and assessment of complexity and difficulty. This is followed by a critical discussion of how these two constructs may be practically operationalized, leading to our proposal for a small selection of complexity and difficulty measures to be routinely used in SLA research. The section Interactions Among Constructs offers some notes on the relationship between complexity and difficulty on the one hand and L2 development and proficiency on the other. The final section contains recommendations for SLA researchers and conclusions. Given our personal academic backgrounds, the discussion will mainly focus on L2 research, although much of what will be said is also relevant for L1 acquisition, and we will draw on the L1 acquisition literature whenever this is relevant.

<A>State of the Art

<TXT>

The term *complexity* was sporadically used in the early years of SLA research to characterize L2 development (e.g., Larsen-Freeman, 1978). It rapidly gained ground in the 1990s in association with the notions of fluency and accuracy to form the Complexity, Accuracy, Fluency (CAF) triad (Housen & Kuiken, 2009). This reflected the field's awareness that language development could not be described simply in terms of increasing accuracy (more target-like uses). Firstly, the addition of fluency accounted for the fact that, over time, language processes become faster, due to their higher automatization. Secondly, the interlanguage perspective (Selinker, 1972) implied that linguistic systems may grow in ways that do not necessarily result in more grammatical correctness, but that nonetheless display a wider, though not always target-like, range of options. Complexity was a suitable term to describe this further dimension, and the number of studies including it among their key variables grew exponentially in the following years. Wolfe-Quintero et al. (1998) provided an extensive overview of the numerous measures of lexical and grammatical complexity used thus far, as was done by Ortega (2003) for syntactic complexity, with a specific focus on academic writing. More recent overviews of syntactic and lexical complexity measurement practices can be found in Johnson's (2017) meta-analysis of research on cognitive task complexity and Crossley's (2020) overview of measures for analyzing writing quality and development.

The relationship between the definition and operationalization of complexity was critically discussed in a number of theoretical contributions. Many of these argued that linguistic complexity itself is not a unidimensional notion and that its use as an umbrella term covering all aspects of language development other than accuracy and fluency may be problematic (Biber et al., 2011, 2020; Bulté & Housen, 2012, 2014; Lambert & Kormos, 2014; Norris & Ortega, 2009). A further issue that was repeatedly pointed out is that the term complexity can have at least two main meanings (Bulté & Housen, 2012, 2014; Kusters, 2008; Miestamo, 2008; Pallotti, 2009, 2015). One refers to the structural properties of an object, phenomenon or system as such and is also called objective complexity. In a general sense, objective complexity has to do with the number and variety of constituent components and the elaborateness of their interrelational structure (see Rescher, 1998). When applied to language, this type of complexity refers to an inherent property of a linguistic structure or system or a text. The second sense concerns the interaction of these linguistic objects with an agent, most typically a human agent. It is in fact sometimes deemed subjective or agent-

related complexity as it has to do with the processing costs and demands for the individual, that is, with what in many cases would be called difficulty.

In language acquisition studies, sometimes the sense of "complex" is further broadened to include everything that is learned later. This is done on the assumption that what is learned first is easier and what is learned later on is more difficult, or what typically occurs later in language production is indicative of more advanced language use or, in short, development (DeKeyser, 2005; Housen, 2021). This attitude is evident in the suggestion that the validity of complexity measures should be argued on the basis of their ability to show that development has taken place over time (e.g., Larsen-Freeman, 1978; Ortega, 2012; Wolfe-Quintero et al., 1998). Yet another sense or use of the term complex(ity) is to describe, characterize, measure, and explain learners' or users' language proficiency according to the assumption that more proficient language knowledge and use is somehow more complex and, conversely, more complex language knowledge and use entail higher proficiency (Bulté & Housen, 2014).

Thus, the term "complex" has been employed to cover at least four distinct notions: (a) structurally elaborate; (b) cognitively demanding; (c) acquired later; (d) typical of (more) proficient language use.

Although umbrella terms may be useful in some cases, when they come to cover too many conceptually and empirically separate meanings, they may cause terminological confusion and hinder the progress of a coherent and shared research agenda. Different studies may in fact attribute different or multiple meanings to the word "complexity," or they may fail to define its exact meaning and boundaries of application altogether. This makes it difficult to compare and accumulate research findings. The situation gets even more complicated with the proliferation of complexity measures, some of which actually represent different constructs and others are substantially overlapping and redundant (Norris & Ortega, 2009). This situation may be partly driven by an increased reliance on computational methods and tools (see, e.g., Lu, 2017, for syntactic complexity; Kyle et al., 2018, for fine-grained indices of syntactic complexity; Kyle & Crossley, 2015, for lexical sophistication) that provide dozens, perhaps hundreds of measures, many of which have been interpreted in terms of complexity. The risk is to select them randomly, without careful reflection on the relationship between empirical measures and theoretical constructs (Bulté & Housen, 2012). Even more problematic is p-hacking, that is, the inclusion of a wide range of measures to then simply select or focus on those that produce statistically significant results. The design of an empirical study crucially includes the identification of the variables of interest and of the measures operationalizing them. All this needs to be specified in advance on the basis of conceptual rigor and theoretical consistency, rather than on the quest for ex-post statistical significance, which may sometimes lead one to select questionable or redundant measures. This is why in later sections we will propose a small set of measures that should be routinely used in order to facilitate knowledge accumulation, regardless of whether they produce significant results.

Compared to complexity, difficulty is not a popular term in AL and SLA research and, even more than linguistic complexity, linguistic difficulty has rarely been thoroughly theorized (Housen & Simoens, 2016). Studies have often simply posited certain L2 structures to be "hard," "problematic," "challenging," or "difficult" for L2 learners, without providing any further argumentation. Although the term difficulty as such may not have been employed frequently in the literature, the notion has played a considerable role as a descriptive or explanatory variable in empirical studies in various research strands, including on text readability, psycholinguistic processing, L1 and L2 acquisition, and language attrition. It is related to the notion of processing cost proposed in psycholinguistics and various areas of theoretical and applied linguistics. It has been measured in terms of, for instance, the

iconicity of structures (Steger & Schneider, 2012), communicative efficiency (Gibson et al., 2019; Hawkins, 2014), or semantic transparency (Seuren & Wekker, 1986). Difficulty has also figured in several theoretical models of SLA, albeit indirectly and under the guise of other terms and related concepts, such as learnability (Izumi & Lakshmanan, 1998; White, 1990), processability (Pienemann, 1998; Pienemann & Lenzing, 2020), or markedness (Callies, 2013; Eckman, 2008).

Our understanding of difficulty, which will be further elaborated below, is largely grounded in theories seeing language acquisition and use in terms of cognitive processes and skills, for example, skill acquisition theory (DeKeyser, 2020) or emergentism (O'Grady, 2022). According to these approaches, the acquisition of any linguistic skill (e.g., the ability to produce and comprehend a certain word, syntactic pattern, morphological process) occurs as a chain of cognitive processes and mechanisms that operate at different levels or stages of cognitive processing. First, at the level of input, new linguistic elements are attended to (i.e., perceived, detected, selected) and stored in working/short-term memory as intake (input processing). They may then be further analyzed and subsequently stored as new linguistic representations (knowledge) and integrated in a dynamic system of representations in long-term memory (intake processing). Finally, these representations may be further activated, strengthened, and, possibly, proceduralized/automatized for use in reception/comprehension and for generating output (Suzuki, 2023; VanPatten, 2020). It goes beyond the scope of this paper to discuss these mechanisms and processes in detail, or the extent to which they overlap with the cognitive processes involved in language comprehension and production. Important for our discussion is that they have different costs in terms of the cognitive resources, that is, the effort and energy that their execution takes, which is what we call difficulty.

<A>Conceptual Issues and Definitions

<TXT>

The aim of this article is primarily to bring conceptual and terminological clarity by proposing a key conceptual and terminological distinction between complexity and difficulty and, secondly, by discussing how these two constructs relate to the constructs of development and proficiency. We see this endeavor as a continuation of previous approaches and practices. Adding complexity as a further dimension besides accuracy and fluency has led to a more perspicuous representation of language development, but we believe that, after three decades of systematic investigation of this dimension, a critical reappraisal seems in order, to differentiate it from other related, yet distinct, constructs. These constructs should not be conceived of as subcategories of complexity; rather they are on a par with it, which allows us to provide a more articulate picture of the multidimensional nature of language development. Thus, rather than continuing to use complexity as an umbrella term covering many different dimensions, we suggest restricting its scope so that it refers to a clear, relatively homogeneous notion. We also advocate using different terms for different constructs, which can then be (empirically) related to complexity, instead of being seen as its manifestations. We realize that this proposal may be controversial, but we hope that it will at least stimulate discussion surrounding the issues that are presented here.

In our framework, complexity refers exclusively to the structural characteristics of linguistic items/structures and texts. By difficulty we mean the cognitive demands that these structures or texts place on human users, whether in production, reception, or acquisition. Development refers to changes in the learner's language system over time, most notably the temporal order in which linguistic structures appear and/or are mastered. Proficiency, finally, refers to a language learner/user's ability to use linguistic structures (alongside paralinguistic and nonlinguistic devices) for a range of communicative goals. In what follows, we offer conceptual definitions of linguistic complexity, which strive to be as much as possible

language-independent and theory-neutral, after which we will examine their relationships with the constructs of development and proficiency.

<B>Defining Complexity

<TXT>

Depending on the field and object of study, complexity in contemporary science has been associated with various related, but also quite distinct, measurable properties, such as entropy, information, perplexity, intricacy, and description length. Many of these modes, or "standards," as Rescher (1998) called them, have also been used to study language systems, specific linguistic structures or texts. In this contribution, we take a more restrictive approach, which allows one to study the relationships between complexity and other constructs in a more perspicuous manner.

We thus define language complexity as structural complexity, that is, the quantity and variety of linguistic components and of the relationships between them (Bulté & Housen, 2012; Pallotti, 2015). These components are linguistic items resulting from linguistic description or analysis. More specifically, our definition of complexity largely overlaps with Rescher's (1998) ontological modes of compositional and structural complexity, which comprise constitutional (i.e., number of constituents) and taxonomical complexity (i.e., variety of constituents) on the one hand and hierarchical (i.e., number of subordination relationships) and organizational complexity (i.e., variety of arrangements and interrelationships) on the other. Our definition is also largely in line with Biber et al.'s (2020) characterization of complexity as the structural elaboration of linguistic units. In contrast to, for example, Ortega (2003), our definition does not refer to the notion of sophistication, which is often defined in terms of frequency, as frequency is not part of the structural makeup of a linguistic item (yet see the discussion in the section Individual-Structure Difficulty).

In fact, complexity, narrowly defined, can be used as an explanatory variable accounting for difficulty, development, and proficiency, whereas interpretation and explanation become more problematic if these notions are included in the definition of complexity itself. Our definition also does not make reference to any notion of communicative or register/genre-based adequacy or appropriacy (see, e.g., Biber et al., 2011). Finally, we also define complexity independently of the cognitive effort that may be associated with processing linguistic structures, that is, difficulty, which we discuss in the next section.

<B>Defining Difficulty

<TXT>

The notion of difficulty proposed in this article is akin to the category of computational complexity in Rescher's (1998) taxonomy, defined as the effort involved in resolving a problem either by a human being or by a computer executing a program or an algorithm. In our case, resolving a problem means producing, comprehending, or learning a particular linguistic structure. A linguistic structure is thus said to be more difficult if its processing (production, comprehension) and/or learning requires more cognitive resources (activity, energy, effort) from a language learner or user in a particular context.

In this paper, we will be mainly concerned with the difficulty of learning linguistic structures, that is, the cognitive processes involved in decoding linguistic input and integrating this information in phonological, lexical, morphological, syntactic, and semantic representations stored in memory, which can then be used in the production or comprehension of linguistic messages. Three main types of causes of learning difficulty have been identified in the literature: structure-related, context-related, and learner-related difficulty (DeKeyser, 2005; Housen & Simoens, 2016). Structure-related difficulty arises from the properties of the target linguistic phenomenon itself, such as its linguistic/structural complexity, as previously discussed, or its transparency, that is, its interpretability. A second

potential set of causes determining a structure's difficulty has to do with its use in the learning context. One of the key variables in this respect is frequency in the input (DeKeyser, 2005, 2016). The third set of causes concerns the individual contribution that each language learner/user brings to the L2 learning and processing task. A relevant variable in this respect is linguistic knowledge, including knowledge of previously learned languages or general metalinguistic abilities. In this article, we will not develop the discussion of learner-related (or individual) difficulty, as we are more interested in identifying ways of establishing difficulty across learners (i.e., interindividual difficulty; Housen & Simoens, 2016).

<B>Relationships Among Constructs

<TXT>

The approach proposed in this paper has several implications for the epistemological status of and relationship between the various constructs, whose illustration will allow us to be more explicit. Firstly, the study of complexity, as defined here, relates to the "what" of acquisition and can contribute to a property theory of SLA as it can provide an account of some (not all) quantifiable nontrivial aspects of the underlying interlanguage systems, and the notion of difficulty relates to the "how" and "why" of acquisition and can be part of a transition theory of SLA (Gregg, 2003). Secondly, in order to establish the degree of complexity of a language structure or text, it is not necessary to collect any further data from human participants. Rather, complexity is an observable attribute (Kane, 2001), and the validity of its measurement rests on the descriptive adequacy, internal coherence, and perspicuity of the linguistic account that is adopted. Difficulty, on the other hand, is more like a theoretical construct (Kane, 2001). In order to assert that something is difficult, one must invoke a theory explaining the causes of difficulty and gather empirical evidence proving that this is indeed the case. Thirdly, difficulty may be the (ontological) cause of development or a certain developmental order, and this developmental order may be taken as evidence for the structure's difficulty, but one thing does not coincide with the other. Finally, proficiency is a much broader, general notion that has been widely discussed especially in the language assessment literature (for reviews, see Harsch, 2014; Harsch & Malone, 2020). Some of its many characterizations also include aspects related to complexity and difficulty alongside many more constructs, most notably accuracy and fluency (see the CAF framework), but also notions such as adequacy, efficiency, appropriateness, quality, idiomaticity, native-likeness, and so forth.

Complexity, difficulty, development, and proficiency are thus four distinct constructs that must be called by different names in order to study, among other things, the possible relations among them. These relations are not circular but have a clear directionality. The gradual development of linguistic structures over time is an observable fact, and it has been amply demonstrated that at least some of them emerge in predictable and systematic orders in first and additional language acquisition. The structures' greater or lesser difficulty may be one possible explanation for these developmental orders, and, in turn, more structurally complex items may be argued to be more difficult to process and learn. The causal chain is thus as follows: complexity > difficulty > development. This causality is not absolute or exclusive, in that complexity is not the only cause of difficulty, nor is difficulty the only cause of developmental trajectories.

The relationship between complexity, difficulty, and proficiency is of a different nature. Firstly, complexity, difficulty, and even typical developmental timing are properties of linguistic structures (i.e., linguistic structures are more or less complex, difficult, or late acquired), whereas proficiency is a property of persons who use linguistic structures. Secondly, the complexity, difficulty, and developmental order of the linguistic structures that language users can process may contribute to perceptions, evaluations, or characterizations of their proficiency, but the relationship is not directly causal, and it depends on how

proficiency is defined and operationalized. Defining whether and to what extent (communicative linguistic) proficiency should include the ability to process complex, difficult, and advanced language implies a whole discussion of the proficiency construct, which has engaged applied linguists for decades. Although in this article we cannot develop this point much further, we will return to it in the section Interactions Among Constructs.

<A>Operationalizing Linguistic Units of Analysis

<TXT>

Before turning to the operationalization and measurement of complexity and difficulty, it is important to identify and delimit the different linguistic units involved, which is a crucial, though often neglected, step.

At the level of the lexicon, the practical measurement of complexity and difficulty has traditionally started from the identification of lexical units, "words" in everyday parlance. However, as many linguists have pointed out (see, e.g., Ramat, 2019), it is not possible to provide crosslinguistically valid criteria for sharply demarcating what is a word and what is not. Well-known cases such as clitics, compounds or idioms, or isolating languages such as Chinese, or polysynthetic ones such as Inuktitut, challenge the intuitive notion of a word that seems so obvious to (literate) speakers of many European languages. This implies that any empirical study that includes among its measures those based on word counts should specify, relative to the language or languages in question, how the word construct is defined, providing examples of what is included and what is excluded. At a more practical level, for the purposes of measuring lexical, as opposed to morphological, complexity and difficulty, one should remove the morphological variation manifested by words in a text. This can be achieved by lemmatizing the different word forms (i.e., reducing them to one single lexical base) prior to the calculation of lexical complexity and difficulty measures (Vermeer, 2004; Nation, 2006), and/or by grouping them into word families (i.e., by merging derivational forms; see Jarvis & Hashimoto, 2021). This raises the question as to how derivational morphological variants should be treated. Are these different words (i.e., different entries in the mental lexicon), or rather, are they linked to one entry through processes of derivational morphology? There are no straightforward theoretical answers to these questions, and decisions on how to treat such variants of words may vary from language to language and according to one's theoretical orientations. In addition, analyses of lexical complexity should be explicit about whether so-called function words are included as part of the lexicon or excluded on the grounds that they belong to grammar.

Turning to morphology, to overcome the difficulties inherent in the word construct, 20th century linguistics introduced the more technical notion of morpheme. This too has been discussed and criticized, so much so that many authors propose to abandon it altogether (for a review, see Leu, 2020). Many morphologists find it preferable, which is also the orientation of this article, to speak of morphological processes, patterns, or operations. These include both concatenative processes, such as the addition of phonological segments, and other processes, such as phoneme change, contraction, reduplication, or even the absence of any change at all (Haspelmath & Sims, 2010). The different forms that lexemes can take as a result of inflectional processes are called *exponents* (Matthews, 1974).

The scope and nature of syntax is still widely debated in theoretical linguistics as well. Simply, and somewhat simplistically, syntax refers to the principles determining how units of language are combined to signal a range of linguistic meanings and functions. These combinations may be described in terms of constituencies (smaller units forming increasingly larger units), dependencies (hierarchical relationships between individual words that make up linguistic expressions), or constructions (patterns of cooccurrence going from relatively fixed expressions like idioms to templates, such as *X enjoy Y*, or abstract patterns like NP-Subject + Transitive Verb + NP-Object). The study of syntactic complexity and difficulty in

mainstream applied linguistics has often been conducted with scant attention to syntactic theorizing in general or without pledging explicit allegiance to any specific syntactic theory. There are, however, some notable exceptions to this, such as Biber and colleagues (working with his corpus-based grammar), and scholars working within generative grammar, such as Slabakova (2014), who identified the structures in an L2 that are harder to process and acquire on the basis of their inherent properties as defined by minimalist generative theory.

For practical purposes, the sentence is often chosen as the main syntactic unit of analysis, especially when automated analyses are performed, although this raises several problems. Firstly, there is no clear agreed upon linguistic definition of a sentence, which is most typically operationalized on the basis of punctuation. This in turn makes it unusable for analyzing oral data or productions by writers with limited punctuation skills. Furthermore, the traditional sentence encompasses both coordination and subordination as clause linking mechanisms, which is rather questionable from a complexity measurement perspective. In fact, juxtaposition with or without coordinating conjunctions leads to a lower level of integration than subordination (of which there are different types with varying degrees of integration; see Lehmann, 1988). In this respect, we recommend counting independent coordinated clauses as separate syntactic units. This means that a main clause together with its dependent clauses (i.e., a T-unit; Hunt, 1965; or AS-unit; Foster et al., 2000) is the largest syntactic unit of analysis that we consider (see Pallotti, 2015).

Similarly, clauses have also been defined and operationalized in different ways, with some studies using finite clauses only, others including nonfinite clauses, which leads to varying results when the same metric is used (Bulté & Housen, 2012). In addition, some studies have included subclausal units (e.g., without a verb) in the calculation of syntactic measures (Foster et al., 2000), whereas others discard any utterances that do not consist of a verb (+ predicate) structure. Finally, Biber et al. (2020) have argued that instead of using coarse-grained units like the clause or the phrase, one should distinguish between different types of finite and nonfinite dependent clauses and dependent phrases at different levels of granularity. We believe that this is complementary with a more holistic (or "omnibus") approach, in that it can offer a more detailed picture of the specific type of (syntactic) complexification that occurs in a text.

Recent approaches have looked beyond traditional linguistic units by focusing on the interfaces between previously identified levels of linguistic structure, in particular the lexico-grammatical interface (Bestgen & Granger, 2014; Kyle et al., 2021; Paquot, 2019; Stefanowitsch & Gries, 2003). Typically, lexico-grammatical units consist of combinations of two (or more) items (Kyle & Eguchi, 2023), either defined on the basis of mere sequential cooccurrence (i.e., n-grams) or on the basis of syntactic dependency relationships (e.g., noun–direct object, or verb–adverb; see Paquot, 2019). Some researchers have also studied the combination of a specific lexical unit (e.g., the verb *walk*) and a grammatical construction in which it can occur (e.g., a transitive or intransitive construction; Stefanowitsch & Gries, 2003). A wide range of potential units of analysis has thus been proposed in this context, some of which could be argued to be very closely related to the traditional lexicon (e.g., an adjective–noun combination such as *tight grip*), whereas others are closer to what is often called syntax (e.g., the template of a ditransitive verb-argument construction). The way in which lexico-grammatical units are identified also varies across studies. In some cases, this is done by looking at their collocational properties, meaning that the strength of the association between the items (i.e., how often they tend to occur together) is used to determine whether they qualify as a unit or not; in other cases, units are identified based on the researchers' or informants' intuitions (Gablasova et al., 2017). All this variety at the theoretical and methodological level testifies to the liveliness of this research area, although it makes it

difficult at the current state of knowledge to arrive at shared measures targeting the complexity and difficulty of these units.

After considering how to define and delimit the relevant linguistic units of analysis, we proceed by discussing how complexity and difficulty can be operationalized and measured.

<A>Operationalizing Complexity

<TXT>

On a theoretical level, we have defined complexity as the quantity and variety of constituents and relationships between constituents. Opting for such a narrow definition has consequences for its measurement. For example, frequency-related measures do not fall within the scope of the complexity construct as defined here, which is not in line with most SLA research (including some of our own previous studies; e.g., Bulté et al., 2008; Bulté & Housen, 2014). It can be argued that a learner who produces more infrequent words most likely has a larger vocabulary size (see Jarvis, 2013), so frequency-based measures could be used as a proxy for the complexity (in terms of number of different elements) of the language system of the person who produced that text. However, a text containing more infrequent words is in itself not necessarily more structurally complex (e.g., varied) than a text with more frequent words. This being said, the fact that frequency-related measures do not fall within the scope of complexity as defined and operationalized here does not make them a less valuable tool for SLA research in general, for example, for the purpose of studying difficulty (see also the section Individual-Structure Difficulty).

For the analysis of learner texts, we also do not recommend using measures calculating the number of meanings or functions expressed by linguistic forms or the complexity of form-meaning-function relationships, even though this is a common approach to calculating the complexity of a single lexical item or a morphological process (e.g., Goldschneider & DeKeyser, 2001). Even in a standard language with published grammars, it is not easy to count how many meanings are expressed by a linguistic item. What seems practically impossible, though, is to determine which and how many meanings are expressed by a linguistic item in an evolving linguistic system. We can certainly record a learner's production of the form *speaks*, but how can we be sure that the ending *-s* expresses the entire set of morphological properties that it has in the standard language or only a fraction of it and, if so, which? The conclusion is that one can calculate the semantic complexity of structures in the target language input, and this may have an impact on the difficulty of learning them, but it is often highly problematic to determine the semantic complexity of these structures in an interlanguage system or (learner) texts.

In the following sections, we present a number of concrete complexity measures. Our aim is not to be exhaustive or to propose new measures but rather to provide illustrations and bring clarity as to what these measures actually measure. This review of measures is subdivided into those targeting complexity at the level of individual linguistic forms and those assessing texts.

<B>Individual-Form Complexity

<TXT>

A first category of measures targets the constitutional complexity dimension, that is, the quantity of constituents in individual linguistic forms. The most obvious way in which this complexity of individual forms can be measured is arguably by calculating their length in terms of formal constituent components. Deciding what these constituent components are is not straightforward. At the level of syntax, the most common operationalization consists in calculating the length in words of various syntactic units, such as T-units or AS-units, (finite) clauses and phrases. Note that these word-based measures are compatible with different syntactic frameworks. Most typically, they have been operationalized from a constituency

grammar perspective (e.g., Lu, 2010), but also dependency parsers can be queried to calculate them (e.g., Brunato et al., 2020). Their main appeal is that word-based length measures are relatively easy to implement and interpret. However, some of these measures, especially those targeting higher-order syntactic units (e.g., T-units or AS-units), are hybrid or omnibus measures since it can be argued that the larger the syntactic unit the wider the range of syntactic phenomena that may contribute to its length (for instance, a long sentence or T-unit may be made of a few long clauses or many short ones; Biber et al., 2020; Pallotti, 2015). The hierarchical nature of syntactic units means that these different syntactic length measures, in part, capture the same information and are thus not independent of one another (e.g., mean length of clause contributes to mean length of sentence). To mitigate this issue, rather than looking at the number of words per syntactic unit, we recommend using a different denominator for each unit (words per phrase, phrases per clause, clauses per T-unit or AS-unit; see Pallotti, 2015)[1]. An alternative way of quantifying the number of syntactic elements within a syntactic unit is by counting the number of nodes (in the parse tree) that are dominated by this unit (Hawkins, 1994) or the number of dependents per unit, which, of course, depends on the specific syntactic framework that is used for the analysis. Both counting the number of nodes and counting the number of dependents per hierarchically superior unit can be said to tap into hierarchical complexity as well.

Length-based measures are less commonly used for measuring constitutional complexity at the level of the lexicon, where the length of lexical forms can be measured in terms of letters/phonemes, syllables, or (concatenative) derivational morphs, which may be counted more or less easily in different languages. Such lexical length measures figure prominently in research on language and text processing (e.g., studies on readability, memory, and lexical decision) but have occasionally also been employed in language acquisition studies (Verspoor et al., 2008). Operationalizing constitutional complexity at the level of morphology is less straightforward. One possible understanding of individual structure in the context of morphological analysis is the single morphological operation, such as adding the *-s* ending to an English verb. In this respect, a periphrastic process like the English present perfect, consisting of auxiliary + ending on the lexical base, can be said to be structurally more complex than the simple addition of an ending, as in the simple past. Although the picture is rather clear in the case of periphrastic morphological processes, it is less clear how different degrees of formal complexity may be established for different morphological operations (such as concatenation, ablauting, reduplication, or stem alternation), and we are not aware of any previous attempt in this direction. The applicability of these measures thus seems to be restricted to specific languages and studies.

A second type of measures targets the hierarchical complexity of forms. This is mainly relevant for syntax. The most popular measure used thus far that taps into this dimension is the subordination ratio, which is the proportion of subordinated clauses relative to the total number of clauses. This measure, however, counts only the number of subordinated clauses, disregarding the degree of embedding. Another criticism that has been leveled at this measure is that it lumps different types of embedded clauses (relative, complement, adverbial) together (Biber et al., 2011, 2020).

A more fine-grained, yet in SLA research rarely used measure, gauges the maximum depth of the syntactic parse tree (or the length of the longest dependency path; Ouyang et al., 2022). Alternatively, also the average vertical distance between each node in a parse tree and the root node (hierarchical distance) has been used as a measure of syntactic complexity (Liu et al., 2017). Finally, mean dependency distance, or the average linear distance in terms of number of words between a word and the one on which it depends, is a syntactic measure that falls within our definition of complexity, as it quantifies the number of elements separating

two words that are hierarchically dependent from on one another; however, it has been rarely applied in SLA research (Liu et al., 2022).

<B>Text-Level Complexity

<TXT>

A common approach to calculating text-level complexity is by computing the average of measures assessing the complexity of forms. This has most typically been done for syntax, with measures such as mean length of unit, average number of dependents per unit, and average tree depth. In principle, it would also be possible to do the same with measures gauging the complexity of lexical or morphological forms, for example, by calculating average word length or the average complexity of morphological operations. A valid alternative approach consists in calculating the normalized rate of occurrence of linguistic forms that contribute to complexity (e.g., number of subordinated phrases or clauses per 100 words). Normalized rates of occurrence have been argued to have certain desirable properties compared to length and ratio measures (Biber et al., 2013), even though we are of the opinion that meaningful ratio measures have their own merits. It would lead us too far, however, to discuss these in detail here.

A third type of text-level complexity measure targets the diversity of linguistic forms. Diversity is in itself a multidimensional concept, as demonstrated by Jarvis (2013), who distinguished seven subconstructs (i.e., size, richness, effective number of types, evenness, disparity, importance, and dispersion). In line with our general definition of complexity, we adhere to a narrower interpretation, which does not incorporate the frequency of items in the language as a whole (as in Jarvis' subconstruct of importance). The lexical and morphological measures discussed here target taxonomical complexity (i.e., the variety of constituents), whereas the syntactic measures tap into organizational complexity (i.e., the variety of relationships).

Diversity has most commonly been measured for lexical forms. The most basic operationalization of diversity consists in calculating the ratio between the number of different forms in a text (i.e., types) and the total number of forms (i.e., tokens; type-token ratio or TTR). A number of computationally more complex diversity measures have been proposed with the aim of reducing unwanted text length effects, including the hypergeometric distribution measure (HD-D), the measure of textual linguistic diversity (MTLD; McCarthy & Jarvis, 2010), and the mean segmental type-token ratio (MSTTR; Johnson, 1944) and its variant, the moving-average type-token ratio (MATTR; Covington & McFall, 2010). We follow Zenker and Kyle's (2021) recommendation to use MATTR, which has been shown to be stable even with relatively short texts. TTR-based diversity measures can also be applied to the lexico-grammatical units in a text (Paquot, 2019).

In analogy with calculations of lexical diversity indices, the diversity of morphological processes can also be measured by looking at the variety of morphological processes, which may be understood as a series of operations on lexical bases. The Morphological Complexity Index (MCI; Pallotti, 2015) relies on the precise definition of these operations in order to identify a series of inflectional types, the diversity of which is calculated on the basis of standardized samples that include a constant number of forms of the same word class (e.g., 10 verbs or 10 nouns). The index can be calculated with an automated online tool (Brezina & Pallotti, 2019). Brezina and Pallotti (2019) justified why this measure is a better alternative to the Inflectional Diversity (ID) measure previously proposed by Malvern et al. (2004) and the (Normalized) Mean Size of Paradigm proposed by Xanthos and Gillis (2010).

In contrast to lexical and morphological complexity, the diversity of syntactic forms has only rarely been investigated in SLA research. Bi and Jiang (2020), for example, used dependency labels obtained from a Universal Dependencies parser to calculate a (mean

segmental) syntactic TTR. A different approach was taken by De Clercq and Housen (2017), who classified AS-units according to their internal clausal structure (e.g., main clause + finite adverbial clause). In this method, each AS-unit (or T-unit) gets one (composite) label, and the diversity of these syntactic labels is then calculated. A related type of syntactic diversity measure is based on calculating the (dis)similarity between the internal structure of syntactic forms (e.g., as implemented in the Coh-Metrix tool; McNamara et al., 2014).

Many more measures have been developed that tap into text properties that are compatible with our definition of complexity. These include measures based on the entropy, uncertainty, or quantity of information of texts, derived from information theory (e.g., Tanaka-Ishii & Aihara, 2015; Gries & Ellis, 2015). Their formulas and outcomes, however, are less straightforward than those of TTRs. Other measures, such as Yule's K (Yule, 1944), more directly target the degree of recurrence of words in a text or, alternatively, text constancy, properties which are related to the dispersion dimension of diversity (Jarvis, 2013; Tanaka-Ishii & Aihara, 2015).

Other approaches, based on, for instance, the Kolmogorov complexity algorithm, measure the compressibility of texts after they are "distorted" (by deleting words or characters) in order to isolate the differential contribution of morphological and syntactic complexity (Ehret & Szmrecsanyi, 2019). However, they do not work well with shorter texts, and their results are difficult to interpret and to relate to more standard linguistic analyses. Finally, lexical density, operationally defined as the proportion of content words relative to the total number of words, has sometimes been proposed in the SLA literature as a measure of lexical complexity. Although lexical density may be a useful metric in the context of, for example, stylometry and genre analysis, we do not consider it to be a dimension of lexical complexity, as there is no formal criterion by which content words can be considered more complex than function words.

<A>Operationalizing Difficulty

<TXT>

The operationalization of difficulty is less straightforward and more tentative than that of complexity. In contrast to complexity, which pertains to linguistic forms in themselves, difficulty looks at linguistic structures through the lens of the language learner/user. As with complexity measures, the operationalizations proposed in this article have to do with the difficulty of structures as they appear in texts. But in order to establish these difficulty levels, several types of evidence have to be invoked, most of which depend on behaviors taking place outside the production of a specific text. These include empirical observations of the overall time needed by learners to acquire individual linguistic structures and learners' performance on psycholinguistic tasks tapping into their cognitive processing of specific linguistic structures. Furthermore, we argue that if a sufficiently strong empirical link is established between a linguistic dimension potentially causing difficulty and its observable effects on learning or processing costs, this dimension itself can also be used as an indicator of difficulty. Postulating these theoretical links is useful given that exhaustive empirical evidence for many structures or forms is often unavailable for most languages. In such cases, one might still use measures based on dimensions that have been shown to produce difficulty, assuming that evidence for these dimensions gathered in other contexts may be applicable to less investigated structures and languages (for example, as discussed below, the effects of form-to-function transparency have been empirically demonstrated in the acquisition of English morphology; transparency may thus be assumed to be a general cause of difficulty in other languages, and possibly for other levels of linguistic analysis, too). Thus, we divide difficulty measures into those targeting dimensions that have been demonstrated to cause difficulty and those that target specific linguistic structures whose difficulty has been empirically established. We first discuss these two types of difficulty measures for individual

structure difficulty, and then for text-level difficulty. As will become clear in these two sections, assessing difficulty in learner language is novel territory, and there are few final answers.

<B>Individual-Structure Difficulty

<C>Measures Targeting Causes of Difficulty

<TXT>

As outlined in the section Defining Difficulty, we first discuss measures targeting structure-related causes of difficulty, followed by measures targeting context-related causes. A first cause of structure-related difficulty is complexity, as defined in previous sections. With regard to the lexicon, longer words and derivationally more complex words are more difficult to process and learn than shorter, derivationally simpler words, all other things being equal (Barclay & Pellicer-Sánchez, 2021; Laufer, 1997; Schmitt, 2010). The same seems to hold for morphology where, for example, in L1 acquisition, discontinuous inflectional forms (e.g., auxiliary + suffix, or prefix + suffix) are learned later than those involving a single operation (Clark, 2017). Likewise, syntactic units that are longer, or with a more intricate constituency/dependency structure, take longer to process and thus require more mental effort (Gibson, 1998; Gibson et al., 2019; Resnik, 1992).

Salience is a second contributor to the difficulty of acquiring linguistic structures, though it is not yet clear in which ways due to its multidimensional nature (Ellis, 2016; Gass et al., 2017). Some structures and items have phonological and graphological realizations that are perceptually more salient and are more easily perceived and attended to by the human mind and are therefore "more likely than others to enter into subsequent cognitive processing and learning" (Ellis, 2017. p. 71). Goldschneider and DeKeyser (2001), using the same criteria as Brown (1973), operationalized one type of salience, perceptual salience, on the basis of three quantitative dimensions: number of phones, syllabicity (1 for functors containing a vowel, 0 for others), and sonority of phones (on the basis of a sonority scale of 1 to 9). Most typically, perceptual salience has been investigated for lexical items and morphological structures. For example, children tend to acquire peripheral morphological markers (i.e., operations taking place at a word's margins) before markers modifying the word's internal structure (Dressler, 2012). L2 learners more easily learn morphological structures composed of more phones, more sonorous phones, with syllabic characters, and that are clearly segmentable (Collins et al., 2009; Goldschneider & DeKeyser, 2001). This is an example of how, at times, different sources of difficulty may go in opposite directions—in this case, more structural complexity (i.e., more phones) produces more saliency and thus a decrease in difficulty. A third causal variable often mentioned is transparency, which has to do with the consistency and multiplicity of form-meaning mappings in linguistic structures. The most natural, canonical, and, therefore, easiest condition for the learner (Audring, 2019) is the use of a single form to express a single semantic/syntactic function. This principle seems to apply to both L1 (Clark, 2017) and Ln acquisition (DeKeyser, 2005). Phenomena such as allomorphy, suppletion, syncretism, and cumulative exponence constitute violations to this principle and are known to pose more problems for learners, who tend to overregularize in order to bring these "anomalies" back to the canonical one-to-one mapping (Godfroid, 2016). Likewise, morphological processes with a clear semantic motivation, for example, number marking on nouns or tense marking on verbs, tend to be acquired earlier than processes with unclear or no semantic content, such as gender, verb mood, and prepositional regencies (Hawkins & Casillas, 2008; Tsimpli & Dimitrakopoulou, 2007).

Similar phenomena, when applied to syntax, have also been described in terms of transparency (Schwartz et al., 1987) or canonicity (Bettoni & Di Biase, 2015). These have to do with the syntactic encoding of thematic roles (Lidz, 2022) so that, for example, the coincidence of agent, topic, and subject in clause-initial position is an easier to process

configuration than cases where these discursive, semantic, and syntactic roles do not match. As was the case with salience, quantifying the transparency of linguistic structures is not straightforward, if only because the construct has an important semantic dimension. It is not immediately clear how structures can be ranked or scored in terms of their transparency other than, for example, by counting the number of meanings expressed by a form, which is notoriously difficult (e.g., Goldschneider & DeKeyser, 2001).

When it comes to lexical items, a range of properties has been found to contribute to their processing difficulty, recently brought together under the heading of "lexical sophistication" (Kim et al., 2018; Kyle & Crossley, 2015). Some of these properties partially overlap with the constructs of salience and especially transparency, but they are rarely labeled or categorized as such. These properties include the semantic criteria of abstractness/concreteness and imageability (typically estimated by means of large-scale subjective evaluations) as well as degree of polysemy and hypernymy. This type of data, however, is available only for a limited number of lexical items in a limited number of languages and, once again, counting the number of meanings expressed by a lexical form is bound to be controversial. More formal properties, such as neighborhood density (i.e., the number of similar words) can be measured in a more straightforward way, and they have been found to impact the processing or learning difficulty of words (Hashimoto & Egbert, 2019).

Frequency (in the input) is a first source of contextual difficulty. It is, arguably, the most commonly used measure in the AL and SLA literature of the difficulty (often called sophistication in this domain) of individual lexical items, as it has been shown to strongly affect the ease or difficulty of productive and receptive word processing and learning in both L1 and L2 (Desai et al., 2020). It has also been shown that frequency plays a role in the acquisition of morphology, although this is less straightforward than with lexicon, as it has been repeatedly noted that some morphological phenomena are impervious to learning even after massive exposures in the input (Ellis, 2022; Slabakova, 2019). This suggests that frequency is perhaps not one of the strongest explanatory variables in this domain. The impact on difficulty of the frequency of syntactic structures appears to be less thoroughly investigated in SLA (yet see theoretical claims in Ellis, 2002; Gass & Mackey, 2002) in contrast to child language acquisition research, which has yielded ample evidence indicating that frequency effects extend to syntactic structures, such as interrogatives, relatives, and passives (see Ambridge et al., 2015, for a review). Frequency can be quantified either by calculating the (logarithm of the) number of occurrences of an item or by computing its probability of occurrence (i.e., occurrences divided by the total number of words in a corpus). Its measurement requires one to specify a reference corpus that is assumed to approximate the language input that learners are exposed to. The choice of reference corpus is therefore of paramount importance.

In addition to frequency, also dispersion (i.e., how equally distributed items are, e.g., across different contexts of use) has been shown to contribute to word processing costs and can thus be used as an additional determinant of difficulty (Gries & Ellis, 2015; Jarvis, 2013). Similarly, a distinction must be made between frequency and productivity of morphological operations. For example, an irregular or suppletive inflectional form may be very frequent at the level of token frequency (i.e., be repeated many times in the input), but it may have a low type frequency (i.e., it may appear on only a few lexical items). Both types of frequency condition the learnability of morphological structures, and it is important to bear in mind which one is referred to (Collins et al, 2009; Dressler, 2012). The importance of considering both token and type frequency has also been demonstrated with regard to the L1 learning of syntactic and lexico-grammatical structures (such as verb-argument constructions; Ambridge et al., 2015).

For lexico-grammatical items, next to their absolute frequency, also their association strength, or how often two (or more) items occur together relative to how often they appear in total, has been found to impact their processing (Yi & Zhong, 2024). This may have to do with learning-related phenomena, such as prototype formation, and cue validity and reliability (Stefanowitsch & Gries, 2003). In addition to learning individual lexical items, learners also have to learn which words or constructions these lexical items (usually or exclusively) combine with. Both frequency and strength of association appear to play an important role in learning (Ellis & Ferreira-Junior, 2009; Yi & Zhong, 2024), but the exact nature and directionality of their combined effects are still under investigation (Kyle & Eguchi, 2023). Various measures are used to quantify strength of association, which are based on the observed and expected frequencies of the items individually and in combination (Gries & Ellis, 2015). Each of these measures has specific properties: Some, such as mutual information, attribute higher scores to rare combinations, whereas others, such as the t-score, attribute higher values to combinations that occur frequently (Gablasova et al., 2017). The choice of measure depends on the intended application.

Several of these causes of difficulty are brought up in discussions of the *markedness* of language structures, which in some early accounts of L2 acquisition was seen as a general explanatory variable for predicting learning orders (for a review, see Callies, 2013; Eckman, 2008). However, the notion of markedness turns out to be ambiguous and polysemic, since it can mean the difficulty of a structure, its structural complexity, its rarity, its (ab)normality relative to others, and so on. For these reasons, Haspelmath (2006) proposes to abandon the term and replace it with others that refer to the specific dimensions being discussed.

<C>*Difficulty Measures Based on Processing and Acquisitional Evidence*

<TXT>

In certain cases, evidence of processing and learning difficulty is available for particular linguistic structures without reference to potential causes of their difficulty. A first type of evidence comes from studies on acquisitional timing, that is, when in the developmental trajectory a given linguistic structure or item is acquired, relative to other structures and items (for a review, see Ellis, 2008). Although acquisition may be operationalized variously in different studies (e.g., as emergence or mastery, on the basis of interlanguage regularities or target-like accuracy), acquisition timing is probably the most frequently invoked source of evidence to claim that a given structure is (more) difficult to learn. One could thus operationalize the acquisitional difficulty of a structure as the developmental stage at which it appears. To this end, a sequence of stages should be first established (such as those in processability theory; Pienemann, 1998; Pienemann & Lenzing, 2020; or, in other theoretical frameworks, the sequences proposed for French by Bartning & Schlyter, 2004 or for English by Biber et al., 2011), making explicit how they were defined with particular regard to the acquisition criterion. That done, each structure would receive a score based on its acquisitional level. With regard to lexical items, for certain languages, lists of words have been established according to their average age of acquisition in the L1 (e.g., Kuperman et al., 2012). A related indicator of difficulty is performance accuracy by both typically and atypically developing L1 and L2 populations. If users systematically fail to comprehend and/or produce some linguistic structures correctly, as evidenced, for example, by grammaticality judgment tasks, sentence-completion tasks, imitation tasks, or production tasks, these structures are said to be more difficult (which has in turn been taken as a proxy for acquisition time, notably in some cross-sectional studies).

A third type of evidence comes from online processing studies. This includes measurements of time spent on task (e.g., reaction times; Hamrick, 2023), psycho-physiological measures of brain activity (e.g., event-related potentials, ERP, or functional magnetic resonance imaging, fMRI; Morgan-Short, 2014; Uddén et al, 2022), or eye

movement and pupil dilation (e.g., eye tracking; Godfroid, 2020; pupillometry; Schmidtke, 2018). With regard to syntax, it has been shown that, for example, center embedding is more difficult than noncenter embedding, that (some types of) clefts are more difficult than noncleft structures, that object relative clauses are more difficult to process than subject relative clauses, that relative clauses appear to be difficult to process in general, and that passive structures are more difficult than active structures (see, e.g., Gibson et al., 2019; Juffs & Rodriguez, 2014; Levy et al., 2013; Traxler et al., 2002). This type of research has also been conducted within the framework of specific syntactic theories (sometimes called experimental syntax), which provides evidence on the processing difficulty of a further range of theory-specific syntactic structures (e.g., different types of island effects or movement constraints in generative grammar; see a review by Sprouse & Villata, 2021). Similarly, for some languages (especially English) words have been ranked according to the average reaction times of native speakers on lexical decision tasks (Balota et al., 2007), with longer reaction times being indicative of higher processing difficulty.

It is clear that the evidence on the processing costs of individual linguistic structures gathered thus far is at best partial, in that it is restricted to certain languages and certain structures only. In other words, we do not currently dispose of lists containing all (or a wide selection of) linguistic items in even a single language with their associated processing costs (in terms of, e.g., reaction times or pupil dilation). This is, at least in part, due to the fact that it is difficult to combine evidence across studies, especially when it comes to the raw values of measurements. Alternatively, measures could be based on a binary logic, for example, singling out and counting those specific structures that have been found to be (more) difficult to process. Another approach would be to rank structures according to an ordinal difficulty scale on the basis of accumulated evidence from multiple studies, similar to establishing the developmental level at which structures tend to appear.

Finally, an alternative approach consists in using subjective difficulty measures, such as ratings and rankings of selected structures and phenomena by experts (applied linguists, teachers) or L2 learners themselves. Examples of this approach include Silva and Roehr-Brackin (2016), who related the perceived difficulty of structures to learners' actual learning and performance of those structures, and Cerezo et al. (2016), who used introspective methods (think-aloud protocols). A more novel, yet more indirect, source of evidence of the difficulty of individual syntactic structures may come from research in computational linguistics measuring the processing cost of individual syntactic structures (often defined in terms of specific syntactic theories) by automated parsers, which has been shown to correlate with processing costs experienced/displayed by humans (see, e.g., Caucheteux & King, 2022). However, at the present state of research, this type of measure does not seem to have wide applicability in mainstream SLA and AL studies but seems restricted to computational applications.

<B>Text-Level Difficulty

<TXT>

Once a principled way is found to assign difficulty scores to the relevant individual structures, then the difficulty of texts could be computed by averaging the difficulty scores of all the relevant different structures in a text. This exercise in effect amounts to establishing a difficulty profile of a text. A related approach consists in calculating the (normalized) rates of occurrence of specific structures that have been empirically demonstrated to be difficult because they appear late in acquisition and/or are difficult to process (e.g., number of passives or finite causative adverbial clauses per 100 words).

We have mainly focused on difficulty in L2 production, but in the context of measuring text-level difficulty, we should also mention difficulty measures of L2 comprehension. These could be found in research on text readability. Such measures often

encompass individual word and syntactic structure difficulty either as part of readability formulas or as inputs to machine-learning algorithms (François & Miltsakaki, 2012; Vajjala & Meurers, 2012).

<A>A Small Set of Core Complexity and Difficulty Measures

<TXT>

In the previous sections, we classified complexity and difficulty measures according to the constructs and subdimensions that they target and discussed some of their strengths and limitations. In this section, we propose a restricted set of "core" measures that we recommend be routinely used in SLA studies. Their identification is grounded in the discussion of the previous sections and takes into account their construct validity, feasibility, and/or relatively wide acceptance in the field. Moreover, in our selection, we strive for parsimony (i.e., avoiding overlap and multicollinearity) in the interest of replicability and knowledge accumulation. Since researchers analyzing L2 productions are rarely interested in measures targeting single linguistic structures, our list of core measures (see Table 1) contains only text-level measures (which, as explained in the previous section, are often based on averaging structure-level measures). We also recognize that individual studies may require additional "noncore" measures, which in Table 1 are indicated by italics[2]. These are not necessarily inappropriate or less valid but rather are either targeted to answer very (theory-)specific research questions or (as yet) based on limited evidence or hard to implement in practice. To provide some examples, noncore measures such as mean words per (finite) clause, T-unit, or AS-unit may be used for the sake of comparability with previous research, and maximum depth of syntactic tree may be used in studies grounded in paradigms positing trees and their depth. Likewise, acquisitional timing for certain linguistic structures would make an ideal measure of difficulty, but still we deem it to be a noncore measure, as this type of information is either lacking or incomplete for many languages.

For lexical complexity, we list one core measure that targets lexical diversity, MATTR, as well as three noncore measures: mean word length, measures based on entropy, and Yule's K. For morphology, we propose to use the Morphological Complexity Index (MCI). We list the measure mean number of morphological operations (or morphemes under some accounts) as a noncore measure, as it requires an explicit operationalization of these operations, which is not always easy to achieve. For syntactic complexity, we recommend measures that target three different subdimensions of complexity: mean words per phrase, mean phrases per clause, and mean clauses per T-unit or AS-unit for constitutional complexity, the normalized rate of occurrence of dependent syntactic structures for hierarchical complexity, and the MATTR of dependency relations or syntactic structures for organizational complexity. Table 1 also contains a number of noncore measures of syntactic complexity whose applicability may be limited or debated for reasons explained in the section Individual-Form Complexity. Furthermore, some of these noncore measures are analytically correlated with other constitutional and hierarchical complexity measures, which may produce multicollinearity issues. Finally, for lexico-grammatical complexity, we recommend calculating the MATTR of lexico-grammatical units, however defined.

<COMP: Place Table 1 near here>

Table 2 lists core and noncore (in italics) measures of linguistic difficulty, distinguishing between those based on the causes of difficulty and those based on empirical evidence for difficulty. Given that difficulty research is a relatively new research domain, the list of core difficulty measures is limited and the details for many of the noncore measures still need to be worked out.

<COMP: Place Table 2 near here>

For lexical difficulty, we recommend widely used measures such as the mean length of lexical units and their frequency in a reference corpus. The only rather uncontroversial measure of morphological difficulty is stage of acquisition, as determined by empirical observation or as predicted by specific theories. Transparency/regularity and saliency of different morphological structures are certainly valid indicators of difficulty, but their operationalizations and computation in language samples remains more an objective for future research than an accomplished result. The core measures of structural syntactic complexity listed earlier are also good candidates for measuring the difficulty of syntactic structures, as are measures based on the notion of more or less canonical word orders (as defined in various theoretical frameworks) and acquisitional timing. At the current state of knowledge, we do not think that any measure of lexico-grammatical difficulty has yet received enough empirical support to be considered a core option for capturing this construct, even though measures of association strength and/or frequency are promising candidates.

<A>Interactions Among Constructs: Complexity, Difficulty, Development, Proficiency <TXT>

Following the definition of complexity and difficulty presented above, one may track their development over time using the measures presented in the previous sections and empirically investigate their relationship with constructs such as L2 development and proficiency. As regards L2 development, several studies have shown that as learners progress, they become able to produce and comprehend more complex structures (e.g., Barrot & Agdeppa, 2021; Bulté & Housen, 2018; Crossley et al., 2011). However, this does not imply that complexity grows steadily and endlessly: Some complexity levels are more appropriate to certain communicative situations or genres, and in some cases, the developmental trend goes from more to less complexity (for a discussion, see Pallotti, 2023). Moreover, longitudinal studies have shown that scores on these measures often develop in nonlinear ways, especially at the level of single learners, and that there is considerable interindividual variation (Bulté & Housen, 2018; Kyle et al., 2021; Lowie & Verspoor, 2019). The relationship between learning difficulty and acquisitional timing is tighter, and we have taken learning time as one of the key indicators of a structure's difficulty. Nonetheless, we do not think that the late appearance of difficult structures is an analytic truth or an unfalsifiable tautology, if only because there are multiple causes of difficulty, and a structure may be difficult in one respect (e.g., transparency) and easy in another (e.g., salience). Thus, there are probably structures that are easy in all or most respects and others that are unequivocally hard to learn, but more research is needed to unravel all causes of difficulty and empirically establish different degrees of difficulty for different structures in different languages.

In this context, it is important to once more stress that language development also occurs along dimensions that are neither complexity nor difficulty (nor do they concern accuracy or fluency), such as appropriateness and adequacy (e.g., Durrant & Durrant, 2022; Kuiken & Vedder, 2017).

Similar remarks may be made for using complexity and difficulty measures in the context of proficiency assessment. Here, too, the ability to produce and comprehend complex and difficult linguistic structures may be considered to be part of a more general proficiency construct, and several studies have shown that higher proficiency levels, as established, for example, by standardized proficiency tests (e.g., Test of English as a Foreign Language, TOEFL; the International English Language Testing System, IELTS), tend to be associated with more complexity in learners' productions (e.g., Bi & Jiang, 2020; Bulté & Roothooft, 2020; Ortega, 2003). However, once again, the relationship is not always linear and does not hold across the board. Especially at more advanced levels, proficiency may entail using relatively simple language or language with appropriate complexity (Pallotti, 2023), where appropriateness has to do both with the quantity and the type of complex structures.

We hope that our contribution can bring more clarity to the notions of difficulty and complexity, which are often brought to bear on the definition of proficiency along with accuracy, fluency, communicative adequacy, and linguistic development. This in turn has an impact on the identification of developmental profiles, some of which will be common to many or most individuals, whereas others may be idiosyncratic. Profiling has a number of practical applications for formative assessment, course placement, or the development of didactic materials and language tests.

One of the implications of this approach is that the well-known CAF triad could be expanded to include more dimensions in order to give a fuller picture of language proficiency and its development. One could thus consider including the constructs of difficulty (the ability to comprehend and produce difficult linguistic structures) and appropriateness (the ability to choose within one's repertoire the alternatives that are most adequate for a given communicative context), so that the acronym may become CAFDA.

<A>Conclusions and Recommendations

<TXT>

The main point of this article is that different constructs should be given different names and be carefully defined, both theoretically and operationally. This allows one to investigate their relationships and to answer some fundamental questions of SLA research, such as how interlanguage systems develop over time, how these constructs contribute to general language proficiency, and how language acquisition and use are affected by a number of internal and external variables. In particular, we advised against using complexity as an umbrella term covering several of these aspects together, as this does not help the comparison across studies and the accumulation of reliable knowledge. In order to arrive at a more coherent and fruitful research program, we advocate separating complexity from difficulty. Furthermore, we recommend that a relatively small set of measures for both constructs be recurrently used across studies on the basis of a few well-defined units of analysis.

Tools for automatic text analysis have become increasingly popular, and we believe that they are a very valuable resource. However, they also bring some potential threats, such as the indiscriminate calculation of as many measures as possible, possibly with the aim of selecting those producing significant results, with no clear theoretical rationale. It is also possible that different tools operationalize the same variables differently, thus producing inconsistent results when compared with each other or with manual coding. Establishing the accuracy of automated measurement is an important endeavor in its own right (Châu & Bulté, 2023). Our recommendation to the developers and users of tools for automated text analysis is that everything should be explicitly defined and nothing be taken for granted, beginning with the definition and delimitation of units of analysis. Different measures could be labeled and grouped according to the taxonomy proposed here to clarify what constructs they refer to, whether it is complexity, difficulty, or other aspects of linguistic description.

We are of the opinion that terminological and methodological clarity is indispensable to promote an organic research program on some of the key issues of SLA studies. In particular, we believe that the notion of difficulty should have a more central role, as it is at the core of many fundamental issues, such as language use and processing, acquisitional timing, task effects, and language pedagogy. We acknowledge that research in this area is still limited, evidence is sparse and controversial, and results hard to interpret and integrate, mostly because what we call difficulty has been labeled in many different ways, including an overextension of the term complexity. We hope that our reflections and suggestions will contribute to a more coordinated and effective endeavor to unravel some of the most fascinating and challenging issues in applied linguistics research.

Final revised version accepted 25 June 2024

<A>Notes

1 In our account, any measure that counts the (mean) number of constituent components (words, phrases, clauses) of a larger syntactic unit is a length measure. A ratio, on the other hand, shows the proportion of elements belonging to a specific category relative to a more general category and can thus be expressed as a percentage (e.g., percentage of subordinated clauses compared to all clauses, unique words out of the total number of words, etc.). Thus, for instance, the (mean) number of phrases per clause or clauses per T-Unit are to be interpreted as length measures (based on units other than the word) and not as ratio measures, as is often done in the literature.

2 Space restrictions prevent us from discussing in detail the rationale behind every individual measure designated as core or noncore in this section.

<A>References

<TXT>Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language, 42*(2), 239–273. https://doi.org/10.1017/S030500091400049X

<TXT>Audring, J. (2019). Canonical, complex, complicated? In B. Olsson & B. Wälchli (Eds.), *Grammatical gender and linguistic complexity* (pp. 15–52). Language Science Press. https://doi.org/10.5281/ZENODO.3462756

<TXT>Baerman, M., Brown, D., & Corbett, G. G. (2017). *Morphological complexity*. Cambridge University Press. https://doi.org/10.1017/S0332586519000015

<TXT>Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459. https://doi.org/10.3758/bf03193014

<TXT>Barclay, S., & Pellicer-Sanchez, A. (2021). Exploring the learning burden and decay of foreign language vocabulary knowledge: The effect of part of speech and word length. *International Journal of Applied Linguistics, 172*(2), 259–289. https://doi.org/10.1075/itl.20011.bar

<TXT>Barrot, J. S., & Agdeppa, J. Y. (2021). Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assessing Writing, 47*, Article 100510. https://doi.org/10.1016/j.asw.2020.100510

<TXT>Bartning, I., & Schlyter, S. (2004). Itinéraires acquisitionnels et stades de développement en français L2. *Journal of French Language Studies, 14*, 281–299. https://doi.org/10.1017/S0959269504001802

<TXT>Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing, 26*, 28-41. https://doi.org/10.1016/j.jslw.2014.09.004

<TXT>Bettoni, C., & Di Biase, B. (2015). *Grammatical development in second languages: Exploring the boundaries of processability theory*. European Second Language Association. http://www.eurosla.org/eurosla-monograph-series-2/eurosla-monographs-03/

<TXT>Bi, P., & Jiang, J. (2020). Syntactic complexity in assessing young adolescent EFL learners' writings: Syntactic elaboration and diversity. *System, 91*, Article 102248. https://doi.org/10.1016/j.system.2020.102248

<TXT>Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5–35. https://doi.org/10.5054/tq.2011.244483

<TXT>Biber, D., Gray, B., & Poonpon, K. (2013). Pay attention to the phrasal structures: Going beyond T-units—A response to WeiWei Yang. *TESOL Quarterly, 47*(1), 192–201. https://doi.org/10.1002/tesq.84

<TXT>Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes, 46*, Article 100869. https://doi.org/10.1016/j.jeap.2020.100869

<TXT>Brezina, V., & Pallotti, G. (2019). Morphological complexity in written L2 texts. *Second Language Research, 35*(1), 99–119. https://doi.org/10.1177/0267658316643125

<TXT>Brown, R. (1973). Development of the first language in the human species. *American Psychologist, 28*(2), 97–106. https://doi.org/10.1037/h0034209

<TXT>Brunato, D., Cimino, A., Dell'Orletta, F., Venturi, G., & Montemagni, S. (2020). Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 7145–7151). European Language Resources Association.

<TXT>Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity accuracy and fluency in SLA* (pp. 21–46). Benjamins. https://doi.org/10.1075/lllt.32.02bul

<TXT>Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

<TXT>Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics, 28*(1), 147–164. https://doi.org/10.1111/ijal.12196

<TXT>Bulté, B., Housen, A., Pierrard, M., & Van Daele, S. (2008). Investigating lexical proficiency development over time–the case of Dutch-speaking learners of French in Brussels. *Journal of French Language Studies*, 18(3), 277–298.

<TXT>Bulté, B., & Roothooft, H. (2020). Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech. *System, 91*, Article 102246. https://doi.org/10.1016/j.system.2020.102246

<TXT>Callies, M. (2013). Markedness. In P. Robinson (Ed.), *The Routledge encyclopedia of second language acquisition* (pp. 406–409). Routledge.

<TXT>Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology, 5*(1), Article 134. https://doi.org/10.1038/s42003-022-03036-1

<TXT>Cerezo, L., Caras, A., & Leow, R. P. (2016). The effectiveness of guided induction versus deductive instruction on the development of complex Spanish gustar structures: An analysis of learning outcomes and processes. *Studies in Second Language Acquisition, 38*(2), 265–291. https://doi.org/10.1017/S0272263116000139

<TXT>Châu, Q. H., & Bulté, B. (2023). A comparison of automated and manual analyses of syntactic complexity in L2 English writing. *International Journal of Corpus Linguistics, 28*(2), 232–262. https://doi.org/10.1075/ijcl.20181.cha

<TXT>Clark, E.V. (2017). Morphology in language acquisition. In A. Spencer & A.M. Zwicky (Eds.), *The handbook of morphology* (pp. 374–389). Wiley. https://doi.org/10.1002/9781405166348.ch19

<TXT>Collins, L., Trofimovich, P., White, J., Cardoso, W., & Horst, M. (2009). Some input on the easy/difficult grammar question: An empirical study. *Modern Language Journal*, *93*(3), 336–353. https://doi.org/10.1111/j.1540-4781.2009.00894.x

<TXT>Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94–100. https://doi.org/10.1080/09296171003643098

<TXT>Crossley, S.A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research, 11*(3), 415–443. https://doi.org/10.17239/jowr-2020.11.03.01

<TXT>Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*(3), 282–311. https://doi.org/10.1177/0741088311410188

<TXT>De Clercq, B., & Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *Modern Language Journal, 101*(2), 315–334. https://doi.org/10.1111/modl.12396

<TXT>DeKeyser, R. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning, 55(S1)*, 1–25. https://doi.org/10.1111/j.0023-8333.2005.00294.x

<TXT>DeKeyser, R. (2016). Of moving targets and chameleons: Why the concept of difficulty is so hard to pin down. *Studies in Second Language Acquisition, 38*(2), 353–363. https://doi.org/10.1017/S0272263116000024

<TXT>DeKeyser, R. (2020). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–113). Lawrence Erlbaum Associates Publishers.

<TXT>Desai, R. H., Choi, W., & Henderson, J. M. (2020). Word frequency effects in naturalistic reading. *Language, Cognition and Neuroscience, 35*(5), 583–594. https://doi.org/10.1080/23273798.2018.1527376

<TXT>Dressler, W. U. (2012). On the acquisition of inflectional morphology: Introduction. *Morphology*, *22*(1), 1–8. https://doi.org/10.1007/s11525-011-9198-1

<TXT>Durrant, P., & Durrant, A. (2022). Appropriateness as an aspect of lexical richness: What do quantitative measures tell us about children's writing? *Assessing Writing, 51*, Article 100596. https://doi.org/10.1016/j.asw.2021.100596

<TXT>Eckman, F. R. (2008). Typological markedness and second language phonology. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (Vol. 36, pp. 95–115). John Benjamins. https://doi.org/10.1075/sibil.36.06eck

<TXT>Ehret, K., & Szmrecsanyi, B. (2019). Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Research, 35*(1), 23–45. https://doi.org/10.1177/0267658316669559

<TXT>Ellis, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

<TXT>Ellis, N. (2016). Salience, cognition, language complexity, and complex adaptive systems. *Studies in Second Language Acquisition, 38*(2), 341–351. https://doi.org/10.1017/S027226311600005X

<TXT>Ellis, N. (2017). Salience in language usage, learning and change. In M. Hundt, S. Mollin, & S. Pfenninger (Eds.), *The changing English language: Psycholinguistic perspectives* (pp. 71–92). Cambridge University Press. https://doi.org/10.1017/9781316091746.004

<TXT>Ellis, N. (2022). Second language learning of morphology. *Journal of the European Second Language Association*, *6*(1), 34–59. https://doi.org/10.22599/jesla.85

<TXT>Ellis, N., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics, 7*(1), 188–221. https://doi.org/10.1075/arcl.7.08ell

<TXT>Ellis, R. (2008). *The study of second language acquisition* (2nd. Ed.). Oxford University Press.

<TXT>Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics, 21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354

<TXT>François, T., & Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations* (pp. 49–57).

<TXT>Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning, 67(S1)*, 155–179. https://doi.org/10.1111/lang.12225

<TXT>Gass, S., & Mackey, A. (2002). Frequency effects and second language acquisition: A complex picture? *Studies in Second Language Acquisition, 24*(2), 249–260. https://doi.org/10.1017/S0272263102002097

<TXT>Gass, S. M., Spinner, P., & Behney, J. (Eds.). (2017). *Salience in second language acquisition*. Routledge.

<TXT>Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

<TXT>Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences, 23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.09.005

<TXT>Godfroid, A. (2016). The effects of implicit instruction on implicit and explicit knowledge development. *Studies in Second Language Acquisition, 38*(2), 177–215. https://doi.org/10.1017/S0272263115000388

<TXT>Godfroid, A. (2020). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. Routledge. https://doi.org/10.4324/9781315775616

<TXT>Goldschneider, J.M., & Dekeyser, R.M. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning, 55*, 27–77. https://doi.org/10.1111/1467-9922.00147

<TXT>Gregg, K. R. (2003). SLA theory: Construction and assessment. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 831–865). Blackwell.

<TXT>Gries, S. T., & Ellis, N. (2015). Statistical measures for usage-based linguistics. *Language Learning, 65(S1)*, 228–255. https://doi.org/10.1111/lang.12119

<TXT>Hamrick, P. (2023). Conducting reaction time research in second language psycholinguistics. In A. Godfroid & H. Hopp (Eds.), *The Routledge handbook of second language acquisition and psycholinguistics* (pp. 150–163). Routledge.

<TXT>Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly*, *11*(2), 152–169. https://doi.org/10.1080/15434303.2014.902059

<TXT>Harsch, C., & Malone, M. E. (2020). Language proficiency frameworks and scales. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (1st ed., pp. 33–44). Routledge. https://doi.org/10.4324/9781351034784-5

<TXT>Hashimoto, B. J., & Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning, 69*, 839–872. https://doi.org/10.1111/lang.12353

<TXT>Haspelmath, M. (2006). Against markedness (and what to replace it with). *Journal of Linguistics, 42*(1), 25–70. https://doi.org/10.1017/S0022226705003683

<TXT>Haspelmath, M., & Sims, A. (2010). *Understanding morphology*. Routledge. https://doi.org/10.4324/9780203776506

<TXT>Hawkins J. (1994). *A performance theory of order and constituency.* Cambridge University Press.

<TXT>Hawkins, J. A. (2014). *Cross-linguistic variation and efficiency*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199664993.001.0001

<TXT>Hawkins, R., & Casillas, G. (2008). Explaining frequency of verb morphology in early L2 speech. *Lingua, 118*(4), 595–612. https://doi.org/10.1016/j.lingua.2007.01.009

<TXT>Housen, A. (2021). Complexity and difficulty of language features and second language instruction. In C. Chapelle (Ed.), *The concise encyclopedia of applied linguistics* (pp. 388–396). Wiley & Sons. https://doi.org/10.1002/9781405198431.wbeal1443.pub2

<TXT>Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461–473. https://doi.org/10.1093/applin/amp048

<TXT>Housen, A., & Simoens, H. (2016). Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition, 38*(2), 163–175. https://doi.org/10.1017/S0272263116000176

<TXT>Hunt, K. (1965). *Grammatical structures written at three grade levels. NCTE Research report No. 3.* NCTE. https://files.eric.ed.gov/fulltext/ED113735.pdf

<TXT>Izumi, S., & Lakshmanan, U. (1998). Learnability, negative evidence and the L2 acquisition of the English passive. *Second Language Research, 14*, 62–101. https://doi.org/10.1191/0267658986757004

<TXT>Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*, 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

<TXT>Jarvis, S., & Hashimoto, B. (2021). How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research, 7*(1), 163–194. https://doi.org/10.1075/ijlcr.20004.jar

<TXT>Johnson, M.D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing, 37*, 13–38. https://doi.org/10.1016/j.jslw.2017.06.001

<TXT>Johnson, W. (1944). A program of research. *Psychological Monographs, 56*(2), 1–15. https://doi.org/10.1037/h0093508

<TXT>Juffs, A., & Rodríguez, G. A. (2014). *Second language sentence processing*. Routledge.

<TXT>Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

<TXT>Kim, M., Crossley, S.A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *Modern Language Journal, 102*(1), 120–141. https://doi.org/10.1111/modl.12447

<TXT>Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing, 34*(3), 321–336. https://doi.org/10.1177/0265532216663991

<TXT>Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978–990. https://doi.org/10.3758/s13428-012-0210-4

<TXT>Kusters, W. (2008). Complexity in linguistic theory, language learning and language change. In M. Miestamo, K. Sinnemäki & F. Karlsson (Eds.), *Language complexity: Typology, contact, change* (pp. 3–22). John Benjamins.

<TXT>Kyle, K., & Crossley, S.A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*, 757–786. https://doi.org/10.1002/tesq.194

<TXT>Kyle, K., Crossley, S.A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods, 50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

<TXT>Kyle, K., Crossley, S., & Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition, 43*(4), 781–812. https://doi.org/10.1017/S0272263120000546

<TXT>Kyle, K., & Eguchi, M. (2023). Assessing spoken lexical and lexicogrammatical proficiency using features of word, bigram, and dependency bigram use. *Modern Language Journal, 107*(2), 531–564. https://doi.org/10.1111/modl.12845

<TXT>Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition, *Applied Linguistics, 35*(5), 607–614. http://doi.org/10.1093/applin/amu047

<TXT>Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly, 12*(4), 439–448. https://doi.org/10.2307/3586142

<TXT>Laufer, B. (1997). What's in a word that makes it hard or easy? Intralexical factors affecting the difficult of vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp.140–155). Cambridge University Press.

<TXT>Lehmann, C. (1988). Towards a typology of clause linkage. In J. Haiman & S.A. Thompson (Eds.), *Clause combining in grammar and discourse* (pp. 181–225). John Benjamins.

<TXT>Leu, T. (2020). The status of the morpheme. In R. Lieber, S. Arndt-Lappe, A. Fàbregas, C. Gagné & F. Masini (Eds.), *The Oxford encyclopedia of morphology*. Oxford University Press.

<TXT>Levy, R., Fedorenko, E., & Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *Journal of Memory and Language, 69*(4), 461–495. https://doi.org/10.1016/j.jml.2012.10.005

<TXT>Lidz, J. (2022). Parser-grammar transparency and the development of syntactic dependencies. *Language Acquisition*, 1–12. https://doi.org/10.1080/10489223.2022.2147840

<TXT>Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews, 21*, 171–193. https://doi.org/10.1016/j.plrev.2017.03.002

<TXT>Liu, X., Zhu, H. & Lei, L. (2022). Dependency distance minimization: a diachronic exploration of the effects of sentence length and dependency types. *Humanities and*

Social Sciences Communications, 9, Article 420. https://doi.org/10.1057/s41599-022-01447-3

<TXT>Lowie, W., & Verspoor, M. (2019). Individual differences and the ergodicity problem. *Language Learning, 69*, 184–206. https://doi.org/10.1111/lang.12324

<TXT>Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

<TXT>Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing, 34*(4), 493–511. https://doi.org/10.1177/0265532217710675

<TXT>Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Palgrave Macmillan.

<TXT>Matthews, P. H. (1974). *Morphology*. Cambridge University Press.

<TXT>McCarthy, P.M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*, 381–392. https://doi.org/10.3758/BRM.42.2.381

<TXT>McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

<TXT>Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. In M. Miestamo, K. Sinnemäki & F. Karlsson (Eds.), *Language complexity: Typology, contact, change* (pp. 23–41). John Benjamins.

<TXT>Morgan-Short, K. (2014). Electrophysiological approaches to understanding second language acquisition: A field reaching its potential. *Annual Review of Applied Linguistics, 34*, 15–36. https://doi.org/10.1017/S026719051400004X

<TXT>Nation, P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

<TXT>Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555–578. https://doi.org/10.1093/applin/amp044

<TXT>O'Grady, W. (2022). *Natural syntax. An emergentist primer.* (3rd ed.). Retrieved from: https://www.researchgate.net/publication/362967378_NATURAL_SYNTAX_AN_EMERGENTIST_PRIMER_3rd_ed

<TXT>Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

<TXT>Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 127–155). De Gruyter.

<TXT>Ouyang, J., Jiang, J., & Liu, H. (2022). Dependency distance measures in assessing L2 writing proficiency. *Assessing Writing, 51*, Article 100603. https://doi.org/10.1016/j.asw.2021.100603

<TXT>Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590–601. https://doi.org/10.1093/applin/amp045

<TXT>Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research, 31*(1), 117–134. https://doi.org/10.1177/0267658314536435

<TXT>Pallotti, G. (2023). Appropriate complexity. In C. Granget, I. Repiso, & G. Fon Sing (Eds.), *Language, creoles, varieties: From emergence to transmission*. Language Science Press.

<TXT>Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research, 35*(1), 121–145. https://doi.org/10.1177/0267658317694221

<TXT>Pienemann, M. (1998). *Language processing and second language development: Processability theory*. John Benjamins. https://doi.org/10.1075/sibil.15

<TXT>Pienemann, M., & Lenzing, A. (2020). Processability theory. In B. VanPatten, G.D. Keating & S. Wulff. (Eds.), *Theories in second language acquisition* (pp. 162–91). Routledge.

<TXT>Ramat, P. (2019). Morphological units: Words. In R. Lieber (Ed.), *The Oxford encyclopedia of morphology*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.543

<TXT>Rescher, N. (1998). *Complexity: A philosophical overview.* Transaction Publishers.

<TXT>Resnik, P. (1992). Left-corner parsing and psychological plausibility. *Proceedings of COLING-92*, 191–197.

<TXT>Rodríguez Silva, L.H., & Roehr-Brackin, K. (2016). Perceived learning difficulty and actual performance: Explicit and implicit knowledge of L2 English grammar points among instructed adult learners. *Studies in Second Language Acquisition, 38*(2), 317–340. https://doi.org/10.1017/S0272263115000340

<TXT>Sampson, G., Gil, D., & Trudgill, P. (Eds.). (2009). *Language complexity as an evolving variable* (Vol. 13). Oxford University Press.

<TXT>Schmidtke, J. (2018). Pupillometry in linguistic research: An introduction and review for second language researchers. *Studies in Second Language Acquisition, 40*(3), 529–549. https://doi.org/10.1017/S0272263117000195

<TXT>Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan. https://doi.org/10.1057/9780230293977

<TXT>Schwartz, M. F., Linebarger, M. C., Saffran, E. M., & Pate, D. S. (1987). Syntactic transparency and sentence interpretation in aphasia. *Language and Cognitive Processes*, *2*(2), 85–113. https://doi.org/10.1080/01690968708406352

<TXT>Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching, 10*(1–4), 209–232. https://doi.org/10.1515/iral.1972.10.1-4.209

<TXT>Seuren, P. A. M., & Wekker, H. (1986). Semantic transparency as a factor in Creole genesis. In P. Muysken & N. Smith (Eds.), *Substrata versus universals in creole genesis: Papers from the Amsterdam Creole Workshop, April 1985* (pp. 57–70). John Benjamins. https://doi.org/10.1075/cll.1.05seu

<TXT>Slabakova, R. (2014). The bottleneck of second language acquisition. *Foreign Language Teaching and Research, 46*(4), 543–559. https://doi.org/10.1177/0267658318825067

<TXT>Slabakova, R. (2019). The Bottleneck hypothesis updated. In T. Ionin & M. Rispoli (Eds.), *Language acquisition and language disorders* (pp. 319–345). John Benjamins. https://doi.org/10.1075/lald.63.16sla

<TXT>Sprouse, J., & Villata, S. (2021). Island effects. In G. Goodall (Ed.), *The Cambridge handbook of experimental syntax* (pp. 227–257). Cambridge University Press.

<TXT>Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209–243. https://doi.org/10.1075/ijcl.8.2.03ste

<TXT>Steger, M., & Schneider, E. (2012). Complexity as a function of iconicity: The case of complement clause constructions in New Englishes. In B. Kortmann & B.

Szmrecsanyi (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 156–191). De Gruyter. https://doi.org/10.1515/9783110229226.156

<TXT>Suzuki, Y. (2023). Automatization and practice. In A. Godfroid & H. Hopp (Eds.), *The Routledge handbook of second language acquisition and psycholinguistics* (pp. 308–321). Routledge.

<TXT>Tanaka-Ishii, K., & Aihara, S. (2015). Computational constancy measures of texts—Yule's K and Rényi's entropy. *Computational Linguistics, 41*(3), 481–502. https://doi.org/10.1162/COLI_a_00228

<TXT>Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language, 47*(1), 69–90. https://doi.org/10.1006/jmla.2001.2836

<TXT>Tsimpli, I. M., & Dimitrakopoulou, M. (2007). The interpretability hypothesis: Evidence from wh-interrogatives in second language acquisition. *Second Language Research, 23*(2), 215–242. https://doi.org/10.1177/0267658307076546

<TXT>Uddén, J., Hultén, A., Schoffelen, J. M., Lam, N., Harbusch, K., Van den Bosch, A., Kempen, G., Petersson, K. M., & Hagoort, P. (2022). Supramodal sentence processing in the human brain: fMRI evidence for the influence of syntactic complexity in more than 200 participants. *Neurobiology of Language, 3*(4), 575–598. https://doi.org/10.1162/nol_a_00076

<TXT>Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In M. Eskenazi, A. Black & D. Traum (Eds.), *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 163–173). Association for Computational Linguistics.

<TXT>VanPatten, B. (2020). Input processing in adult second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 115–135). Lawrence Erlbaum.

<TXT>Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards, & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 173–189). John Benjamins.

<TXT>Verspoor, M., Lowie, W., & Van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *Modern Language Journal, 92*(2), 214–231. https://doi.org/10.1111/j.1540-4781.2008.00715.x

<TXT>White, L. (1990). Implications of learnability theories for second language learning and teaching. In M.A.K. Halliday, J. Gibbons & H. Nicholas (Eds.), *Learning, keeping and using language* (Vol 1, pp. 271–286). John Benjamins. https://doi.org/10.1075/z.lkul1.20whi

<TXT>Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. University of Hawaii, Second Language Teaching and Curriculum Center.

<TXT>Xanthos, A., & Gillis, S. (2010). Quantifying the development of inflectional diversity. *First Language, 30*(2), 175–198. https://doi.org/10.1177/0142723709359236

<TXT>Yi, W., & Zhong, Y. (2024). The processing advantage of multiword sequences: A meta-analysis. *Studies in Second Language Acquisition*, *46*(2), 427–452. https://doi.org/10.1017/S0272263123000542

<TXT>Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge University Press.

<TXT>Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing, 47*. https://doi.org/10.1016/j.asw.2020.100505

Table 1 Core and noncore measures of complexity (noncore measures in italics)

Complexity dimension/ Language level	Constitutional	Taxonomical	Hierarchical	Organizational
Lexicon	– *Mean word length*	– MATTR of lemmas – *Entropy-based measures*		

		– *Yule's K*		
Morphology	– *Mean number of morphological operations*	– MCI		
Syntax	– Mean words per phrase – Mean phrases per clause – Mean clauses per T-unit or AS-unit – *Number of nodes in tree* – *Number of dependents per syntactic unit* – *Mean words per (finite) clause, T-unit, AS-unit*		– Normalized rate of occurrence of dependent clauses and phrases – *Maximum depth of syntactic tree* – *Dependency distance* – *Hierarchical distance*	– MATTR of dependency relations or syntactic structures – *Mean sentence similarity score*
Lexico-grammatical		-MATTR of lexico-grammatical units		

Note. MATTR = moving-average type-token ratio; MCI = morphological complexity index.

Table 2 Core and noncore measures of difficulty (noncore measures in italics)

Complexity dimension/ Language level	Causes of difficulty		Evidence for difficulty
	Structure-related	Context-related	
Lexicon	– Average word length – *Derivational compositionality* – *Abstractness, imageability*	– Average frequency in reference corpus	– *Age of L1 acquisition*
Morphology	– *Transparency/regularity* – *Saliency*	– *Communicative value*	– Stage of acquisition – *Subjective ratings*
Syntax	– Structural complexity (variously operationalized: words, nodes, dependencies...) – Canonicity (e.g., in extended Processability Theory)		– Stage of acquisition – *Subjective ratings* – *Processing cost for automated parser*
Lexico-grammatical		-*Average frequency in reference corpus* -*Association strength*	

Note. L1 = first language.