21

Measuring Lexicogrammar

Magali Paquot, Stefan Th. Gries & Monique Yoder

Background

Lexicogrammar is a level of linguistic structure where lexis and grammar are not seen as independent, but rather as mutually dependent. The idea that lexis and grammar are interrelated has been promoted by a number of linguistic theories and approaches. In corpus linguistics, more particularly, empirical investigations of large corpora have shown that lexicogrammatical co-selection phenomena are ubiquitous in language (e.g., verbs such as *arrest*, *elect*, *name*, and *estimate* are twice as frequent in the passive as in the active form; the adjective *mere* is attested only in attributive position; Biber et al., 1999; see also Römer, 2009). The study of lexicogrammatical patterns has also exhibited particular growth with the popularization of usage-based or constructionist/Construction Grammar approaches, in which lexical items and grammatical constructions are considered to not be qualitatively different from each other, but differ only with regard to the degree of abstractness within one unified construction (Goldberg, 2006).

Lexis and grammar, however, have traditionally been studied separately in second language acquisition (SLA) and measured as two separate constructs in language testing. Historically, grammar used to be the most central area of linguistic enquiry in SLA (Lardière, 2014) and, only over time, the lexicon has come more at the forefront as well. Today, vocabulary is also largely considered an independent component of language competence that covers all aspects (i.e., form, meaning, and use) of what is involved in knowing a word (see Chapter 20, this volume). Within vocabulary research, much interest has been placed on the learner's development of phraseology or formulaic language (e.g., Schmitt, 2004; Siyanova-Chanturia & Pellicer-Sánchez, 2018), and on how this development relates to notions such as Pawley and Syder's (1983) "native-like selection" or Sinclair's (1991) "idiom principle," according to which "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (p. 110).

In language testing, the separation between lexis and grammar "can be seen both in models of language ability that are used to inform test development and validation, and in many frequently used rating scales that are used to score performance assessments" (Römer, 2017, p. 477-478). This is certainly the case in international standards for describing language ability and major international language tests (e.g., the writing band descriptors of the International English Language Testing System (IELTS) distinguish between "lexical resource" and "grammatical range and accuracy"). Römer (2017) attributed this separation to the history of the

construct of "language proficiency" in language testing, with early language testing researchers prior to the 1970s promoting a skills/component model of language reflective of pedagogical methods rooted in behavioral psychology and structural linguistics, which assumed that language could be subdivided into its separate components. Measurement theory and practice as outlined by Lado's (1961) seminal work on foreign language testing advocated task types that facilitated test reliability over authenticity. Thus, discrete-point task types in which grammar and vocabulary were clearly distinguished and measured as separate constructs that demonstrate learners' knowledge of sentence-level grammar and lexis were the norm in test development for decades. In the 1970s and 1980s, John Oller's theories on the pragmatics of knowledge, beliefs, and expectations of language users in the contexts in which they experience language made way for discourse-based testing procedures known as integrative language tests. He perceived discrete-point and integrative tests to be in contrast, where "discrete items take language skill apart, integrative tests put it back together" (Oller, 1979, p. 37), and he identified cloze, dictation, oral interviews, and written compositions as integrative language test task types that require access to language meaning through deeper linguistics resources. However, as authenticity (i.e., how a test reflects real-world language tasks) became a criterion of test validity and communicative competence became a component of language proficiency (Canale & Swain, 1980), cloze and dictation tasks fell out of favor. As a result, tasks that emulate what a test-taker will have to do with the language yet measure lexis and grammar indirectly through spoken and written language performances have also gained popularity in both large-scale and classroom-based assessments.

With this development came an increased awareness of the difficulty in assessing vocabulary and grammar independently. Ruegg, Fritz, and Holland (2011), for example, showed that the lexis scores for test-takers' written performances on the Kanda English Proficiency Test (KEPT) were predicted by the scores on the grammar scale much more than range, frequency, or lexical accuracy. Because of the difficulty in distinguishing lexical vs. grammatical competence, they recommended collapsing the scales into a single "lexicogrammar" scale as already done for the speaking section of the KEPT (cf. also Alderson & Kremmel, 2013 for a similar discussion in a study that re-examined the content validation of a grammar test). The controversial issue here is whether it is enough to merge the dimensions of grammar and lexis into one lexicogrammatical dimension to claim that we are now measuring lexicogrammar as defined above. We believe that it is not. However, there is a shortage of research to date on how lexicogrammatical competence can and should be measured in language testing and SLA research. In what follows, we sketch out key issues that should be addressed in future research.

Key Concepts

***Lexicogrammar*:** The lexicogrammatical dimension of language deals with (preferred) co-occurrences between words and their grammatical environments, or between grammatical structures and their lexical environments (cf. Biber, Conrad, & Reppen, 1998, p. 84).

***Phraseology*:** Loosely defined as "the study of the structure, meaning and use of word combinations" (Cowie 1994, p. 3168). From a distributional or frequency-based perspective on phraseology, Gries (2008) defined phraseological units or phraseologisms as co-occurring forms or lemmas of lexical items with other linguistic elements. These elements may be other lexical items, such as deterministic co-occurrences (e.g., *kith and kin* / *strong tea*), or grammatical patterns when these are not grammatical relations (i.e., lexical items combined with grammatical constructions, such as the verb *hem* frequently occurring in the passive form).

***Constructions*:** Form-meaning pairs that exist at all levels of linguistic representation, forming the core of usage-based theories of second language acquisition. As Goldberg (2006, p. 5) stated:

> Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency.

Key Issues

Indirect Measurement of Lexicogrammar

By its very nature, lexicogrammar lends itself well to indirect measurement in learners' written and spoken performances, and corpus linguistic techniques provide effective ways to do so. Corpus linguists have long studied the degree to which co-occurrence patterns of linguistic elements help to identify and understand differences of functional characteristics between two or more such elements such as active vs. passive or *will* vs. *shall* vs. *going-to* future. A range of theoretical frameworks and techniques is available to explore lexicogrammar in corpora, including Pattern Grammar (Hunston & Francis, 2000), lexical bundles or *n*-gram analysis (Biber, 2007), and phrase-frames (Römer, 2017), but one type of approach that follows squarely from the above statement that lexis and grammar are deeply intertwined is collostructional analysis (Stefanowitsch & Gries, 2003; Gries & Stefanowitsch, 2004a, 2004b). Collostructional studies explore how the co-occurrence patterns of lexical items in/and grammatical constructions help to identify, quantify, and interpret the following kinds of functional characteristics of linguistic elements:

- characteristics of individual constructions in collexeme analysis (Stefanowitsch & Gries, 2003), e.g., what do the verbs occurring in the ditransitive (*give*, *tell*, *show*, …) reveal about the construction's function(s)?
- differences of two or more constructions in distinctive collexeme analysis (Gries & Stefanowitsch, 2014a), e.g, how do verbs occurring in the ditransitive and the prepositional dative distinguish the two constructions' functions?

While the methods of collostructional analysis mentioned above differ slightly from each other, they are both based on 2×2 co-occurrence frequency tables as represented in Figure 21.1.

	Constr: *X*	Constr: other	Sum
Word: *y*	*A*	*B*	*a+b*
Word: other	*C*	*D*	*c+d*
Sum	*a+c*	*b+d*	*a+b+c+d*

	Constr: *x*	Constr: *y*	Sum
Word: *w*	*a*	*b*	*a+b*
Word: other	*c*	*d*	*c+d*
Sum	*a+c*	*b+d*	*a+b+c+d*

Figure 21.1 Schematic Representation of Co-Occurrence Tables for Collexeme Analysis (Left) and Distinctive Collexeme Analysis (Right).

For collexeme analyses, one determines for every word $y_{1\text{-}n}$ occurring in a slot of construction *x* its frequency in *x* (*a*), its overall frequency (*a*+*b*), the frequency of the construction in the corpus (*a*+*c*), and the corpus size (*a*+*b*+*c*+*d*) to then compute a statistic that expresses how much y_n "(dis)likes to occur" in *x*. For distinctive collexeme analyses, one determines for every word $w_{1\text{-}n}$ occurring in a slot of either construction *x* or *y* its frequency in *x* (*a*), its frequency in *y* (*b*), and the frequency of each construction in the corpus (*a*+*c* and *b*+*d*) to then compute a statistic that expresses which of the two constructions w_n "(dis)likes to occur" and how much. Both of these measures are essentially basic association measures that capture the contingency between a word (often a verb) and a construction. For the collexeme analysis of the ditransitive in Stefanowitsch and Gries (2003, p 229), for example, *a* is the frequency of *give* in the ditransitive (461), *b* is the frequency of *give* elsewhere (699), *c* is the frequency of ditransitives without *give* (497), and *d* is the frequency of all other verbs in all other constructions (137007).

While the initial studies based on these methods involved native-speaker (NS) data only, soon first connections to second and foreign language acquisition/learning appeared, to determine to what degree the constructional preferences that NSs exhibit with certain words are also characteristic of (differently proficient) non-native speakers (NNSs). This approach implies that lexicogrammatical accuracy in learner language cannot simply be reduced to a rule-based binary concept or right vs. wrong: The *standard* of native speakers' usage is nowhere near monolithic, but is instead inherently probabilistic and unpredictable from general rules (see Pawley and Syder's (1983) discussion of "nativelike selection"). Modifying Wulff and Gries's (2011) definition of accuracy slightly, one might define lexicogrammatical accuracy in L2 production as the nativelike selection of constructions (in the above Goldbergian sense of the term) given a certain context, where context can be defined as broadly as necessary. This in turn requires a quantitative/probabilistic perspective to measure the degree of convergence between native and non-native language use and patterning.

One application of collostructional methods to learner data was performed by Ellis, Römer, and O'Donnell (2016). Among other things, they analyzed native speaker corpus data (the British National Corpus) on the use of a variety of verb-argument constructions (VACs) and computed bidirectional measures that reflect the degree to which verbs "like to occur" in constructions and vice versa. These results were then also correlated with learner verb generation in experimental studies (see below) and, more importantly, with longitudinal corpus data from the European Science Foundation (ESF) project involving learners of English with Italian (four speakers) and Punjabi (three speakers). For three verb-argument constructions—verb-locative, verb-object-locative, and double-object constructions—they found "that NS collexeme strength (Fisher-Yates) is a very strong predictor of NNS language acquisition, as is Δ*P* (Construction→Word)" (p. 231). Ellis et al. (2016)'s research is also noteworthy for how the authors extended the traditional corpus-linguistic co-occurrence approach involved in collostructional analyses to investigate the effect of entrenchment (as measured by lemma frequencies of the verbs in the constructions), contingency (again measured by ΔP Construction→Word), and semantic prototypicality (as measured by betweenness centrality in the semantic network of all verbs per construction) on VACs in learner language.

Another way to approach lexicogrammatical competence in learner performance is to frame it in terms of lexicogrammatical complexity. This follows from Paquot's (2019) proposal to add the construct of phraseological complexity to the toolbox of interlanguage complexity research on the ground that traditional measures of syntactic complexity and lexical complexity cannot account for how words naturally combine to form conventional patterns of meaning and use. Building on previous research on complexity, Paquot (2019) distinguished two main dimensions of phraseological complexity, i.e., variety and sophistication (cf. Ortega, 2003), and approached phraseological complexity via relational co-occurrences, i.e., co-occurring words that appear in a specific structural or syntactic relation (e.g., adjective + noun, adverbial modifier + verb, verb +

direct object). Phraseological diversity was operationalized as root type-token ratios. Two methods were tested to measure phraseological sophistication. First, sophisticated word combinations were defined as academic collocations that appear in the *Academic Collocation List* (Ackermann & Chen, 2013). Second, it was approximated with average pointwise mutual information (*MI*) scores as this measure has been shown to bring out word combinations made up of closely associated medium to low-frequency (i.e., advanced or sophisticated) words. Interestingly, the proposed measures of phraseological sophistication have the potential to tap into the (lexicogrammatical) specificities of different registers, tasks, and modalities as they rely on corpus-derived lists of word combinations and their associations (see Paquot, 2019). Results revealed that, unlike traditional measures of syntactic and lexical complexity, measures of phraseological sophistication, and the *MI*-based measures more particularly, distinguish L2 performance at the B2, C1, and C2 levels of the Common European Framework of References (Council of Europe, 2001). This suggests that essential aspects of language development from upper-intermediate to very advanced proficiency levels are situated in the phraseological dimension (see also Paquot, 2018; Paquot, Naets, & Gries, in press).

Paquot's (2019) proposal can be directly applied to measuring lexicogrammatical complexity, which is correspondingly defined here as the range of lexicogrammatical patterns surfacing in language production and their degree of sophistication. Thus, a learner text with a wide range of (target-like) lexicogrammatical patterns and a high proportion of sophisticated patterns will be said to be more complex than one where the same few basic patterns are often repeated. Although there are currently no instruments available to measure lexicogrammatical complexity:

- specialists in vocabulary assessment will already be familiar with the constructs of range and sophistication, and how to measure them;
- recent lexical diversity measures such as the ones discussed in Jarvis and Daller (2013) could be tweaked to work with word combinations, lexicogrammatical patterns, and the like;
- there are already a few general or register-specific frequency-based lists of phrases that could serve as the basis for the measurement of lexicogrammatical sophistication (e.g., Martinez & Schmitt, 2012) but we definitely need more resources, especially lists of sequences that do not have non-compositionality as their main defining criterion, developed for production purposes, and for languages other than English;
- sophistication of lexicogrammatical patterns can also be approached via collostructional analysis (see above), for which Gries provides an R script and its documentation (www.stgries.info/teaching/groningen/index.html).

Note that the corpus techniques described here are largely language independent (except for frequency-based lists) and should also help establish whether the construct is theoretically valid across languages (cf. Rubin et al., in press, for a first study of lexicogrammatical complexity in L2 Dutch). Although they are certainly more computationally expensive than *n*-gram models, they also hold the potential to inform automated assessment.

Direct Measurement of Lexicogrammar

For practicality reasons, language testers and SLA researchers may sometimes want/have to measure lexicogrammatical competence directly. As mentioned above, the lexicogrammatical dimension partly overlaps with phraseology or formulaic language. As a result, any attempt at measuring lexicogrammar directly will typically run into the same issues as encountered in the measurement of formulaic sequences. Among the characteristics identified by Gyllstad and Schmitt (2019) as factors that render the testing of formulaic language highly challenging, the following are particularly relevant when it comes to measuring lexicogrammar:

- *Construct definition and operationalization*: The lexicogrammatical dimension of language entails a range of categories with their own characteristics, e.g., constructions (conventionalized form-meaning pairings that will be identified with statistical association measures), and lexical bundles or phrases (recurring word sequences typically identified with the criteria of frequency and dispersion).
- *The scope of the construct*: There are a very large number of lexicogrammatical items, but lexicogrammar is not rule-based: There is simply no "rule" that the learners should apply to demonstrate good knowledge of the construct. As with vocabulary items, checking learners' knowledge means each lexicogrammatical item would need to be tested separately.

As put by Gyllstad and Schmitt (2019, p. 175),

> These two factors work against both the identification of the target population and the representative sampling of items from that population. This leaves researchers in a very challenging position, and it is probably next to impossible to develop a definite list of all the existing [lexicogrammatical patterns] in a language, and then to develop a test for these sequences.

A number of solutions to the above issues are available. Item selection can rely on the combined use of native and learner corpora. Techniques such as collostructional analysis (see above) or *n*-gram extraction can first be used to extract the most typical constructions or lexical bundles in a given register. Learner corpora can then be used to identify lexicogrammatical items which learners "tend not to use (absence) or use infrequently (underuse) or too frequently (overuse) in speech or writing" (Granger, 2015, p. 487), thus potentially informing the selection, description and sequencing of test items (cf. also Granger, 2009). In vocabulary research, Gyllstad and Schmitt (2019) also recommend to "focus on much more constrained, and thus identifiable, subsets of FLs, in order to make the resulting scores more meaningful" (p. 181) and to use adaptive tests "to achieve a better and more focused sampling" (p. 187).

A different approach is adopted in the development of "Language in Use" tests that aim to test lexicogrammar in context, such as the integrative multiple-choice gap-fill, banked gap-fill, word formation gap-fill, and editing tasks that can be found in the Austrian secondary school-leaving examination for foreign languages. As authenticity is central to communicative language testing, the language elements that make up such tests are not selected beforehand but emerge from the texts that are selected for the design of such integrative tasks (Weiler, 2018, p. 65). One obvious advantage of these tasks is that they follow Schmitt and Schmitt's recommendation:

> But the more test writers wish to measure learners' ability to actually use words in real world situations, the further the tests need to move toward the embedded, comprehensive, and context-dependent ends of the continuums (Schmitt & Schmitt, in press).

The downside, however, is that this is done at the expense of prototypical use. It is also not clear how scores from these tests would relate to the underlying population of items that make up the lexicogrammar construct (see Gyllstad, 2019, for a similar discussion on the measurement of formulaic language).

In SLA studies using experiments, by contrast, researchers have typically resorted to discrete-point measures to explore foreign language learners' knowledge of a delimited set of conventionalized lexicogrammatical patterns carefully selected from corpus data. Gries and Wulff (2009), for example, made use of a sentence completion task and an acceptability rating task to investigate whether the gerund and infinitival complement constructions in English (*She tried rocking the baby* vs. *She tried to rock the baby*) are also stored as constructions by German foreign language learners of English. Ellis et al. (2016) relied on a range of psycholinguistic experiments to

explore the role of VAC frequency, type-token frequency distribution, contingency, and semantic prototypicality upon VAC free-association processing in L2 speakers.

Recommendations for Practice

In rating scales (rubrics), a lexicogrammatical scale should tap into more than just the sum of lexical and grammatical knowledge—it is just a different construct. Unlike Ruegg et al. (2011), and despite our theoretical commitment to the lexis/grammar continuum, we do not think that it is necessarily a good idea to replace the grammar and vocabulary scales by a lexicogrammatical scale in foreign language tests. This is especially true if this is mainly done for lack of agreement between raters on what can be considered lexical as opposed to grammatical.

Apart from them just being different constructs, one of the main practical reasons for maintaining a distinction between grammar, lexis, and lexicogrammar is that the three scales may not be similarly useful at all proficiency levels. Research has shown that unlike traditional measures of syntactic and lexical complexity, measures of phraseological/lexicogrammatical complexity can discriminate between learners' performances across the more advanced proficiency levels (Paquot, 2019; see also Crossley et al., 2012 for a study that reported strong correlations between *n*-gram indices and essay quality). This is because the phraseological and lexicogrammatical dimensions of language develop slowly and remain an area of difficulty even for advanced learners (e.g., Paquot & Granger, 2012). A lexicogrammatical scale will thus probably be particularly relevant at the (upper) intermediate and advanced levels. By contrast, vocabulary and grammatical scales may still remain essential scales to assess beginner and intermediate learners. More research is definitely needed in this area.

However, there will probably not be such a thing as a one-size-fits-all lexicogrammatical test. Lexicogrammatical patterns differ from register to register (e.g., Biber et al., 1999). Their selection is influenced by a variety of linguistic and extra-linguistic factors (including crosslinguistic influence for foreign language learners; e.g., Ellis et al., 2016), which makes lexicogrammar a highly context-dependent phenomenon. As a result, a lexicogrammatical test developed in one specific context will probably not fit many other different contexts. If the fields of language testing and SLA want to assess lexicogrammatical competence reliably and validly, they will likely need to develop tests for different purposes (e.g., tests for learners with different L1 backgrounds, tests for different situational contexts). However, before such developments can occur, there is a need for more research into how lexicogrammatical competence interacts with foreign language proficiency and development across modes, tasks, registers, and other linguistic, contextual, and/or situational characteristics. One approach that will suit these purposes particularly well is the MuPDAR approach (short for Multifactorial Prediction and Deviation Analysis using Regressions; Gries & Deshors, 2014) approach. This technique imputes the choices and/or judgments of the speakers of the reference/standard language from corpus data annotated with regard to many linguistic and contextual characteristics with a first statistical model/classifier, which allows the analyst to then explore non-nativelike choices with a second statistical model/classifier. Wulff and Gries (2015) put forth an application of this protocol to the question of prenominal adjective order by native speakers of English, and Chinese and German learners of English. Their first regression was run on adjective-adjective-noun sequences from the British National Corpus, which were annotated for a variety of predictors governing adjective order from multiple levels of linguistic analysis. That first model was used to impute native speaker choices for the non-native speaker data, which were then compared to the non-native speaker choices to run a second analysis on whether the learners made nativelike choices or not, a protocol that leads to extremely fine-grained results.

While it is probably not realistic to expect language testers to conduct MuPDAR studies to, for example, select valid test items at different proficiency levels, findings of such sophisticated and

time-consuming research designs have important implications for the field. They provide empirical evidence in support of Schmitt's (2010, p. 173) claim that "context-dependent formats will obviously provide a better way of tapping into the 'contextualized' facets of word knowledge like collocation and register," but also demonstrate very clearly that context needs to be approached in all its complexity when it comes to test item selection and description. In that sense, it is recommended that, when developing instruments for measuring lexicogrammatical knowledge, language testers and SLA researchers focus as much on authentic as on *prototypical* linguistic context of use, which is likely to differ from one situational context to the next and will best be described based on corpus data.

Testing Tips

- When setting out to measure lexicogrammar, choose and make explicit your choice of construct definition.
- Establish for what purposes you want to test lexicogrammatical competence.
- Identify language elements that make up the construct (e.g., constructions, phrases).
- Determine which dimension (i.e., accuracy, range, or sophistication) of lexicogrammatical knowledge you want to test and develop or choose your instruments accordingly.
- To select and/or describe lexicogrammatical items, use large corpora ideally representing the registers that best correspond to the objectives of your language test or research instrument.
- Make use of the relevant metrics to analyze corpus data and extract/describe lexicogrammatical items: Frequency should never be used as a single criterion and you need to consider dispersion too. Strength of association is a useful way to operationalize lexicogrammatical sophistication depending on the association measure used.
- Never underestimate the impact of the selected corpus on the selection and description of lexicogrammatical items. Evaluating the reliability of item selection/description in a second corpus is highly recommended.
- Favor context-dependent instruments.
- Pilot and refine your instruments sufficiently—make sure you are tapping into the construct of lexicogrammar and can distinguish it from lexis and grammar if these two constructs are part of your test or research instrument too.

Recommended Readings

Gyllstad, H. (2019). Measuring knowledge of multiword items. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 387–405). Routledge.

This chapter offers a review of different approaches to the measurement of multiword items from the perspective of vocabulary research.

Kremmel, B., Brunfaut, T., & Alderson, J. C. (2017). Exploring the role of phraseological knowledge in foreign language reading. *Applied Linguistics*, *38*(6), 848–870. https://doi:10.1093/applin/amv070

This study found that measures of phraseological knowledge outperformed traditional syntactic and vocabulary measures in predicting reading comprehension variance.

Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, *15*(1), 29–43. https://doi.org/10.1080/15434303.2017.1405421

This article demonstrates with the help of learner corpus data the practical relevance of the phraseological dimension of language for writing assessment in higher education.

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL): A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*(4), 535–556. https://doi.org/10.1177/0265532213489568

Biber, D. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *3*, 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–14. https://doi.org//10.1093/applin/I.1.1

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Cowie, A. P. (1994). Phraseology. In R. E. Asher (Ed.), *The encyclopaedia of language and linguistics* (pp. 3168–3171). Pergamon.

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012). Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In P. M. McCarthy & G. M. Youngblood (Eds.), *Proceedings of the 25th international Florida artificial intelligence research society (FLAIRS) conference* (pp. 214–219). AAAI Press.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of Construction Grammar*. Wiley.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In A. Aijmer (Ed.), *Corpora and langauge teaching* (pp. 13–32). John Benjamins.

Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp.485–510). Cambridge University Press.

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3–25). John Benjamins.

Gries, S. T., & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*, *9*(1), 109–136. https://doi.org/10.3366/cor.2014.0053

Gries, S. T., & Stefanowitsch, A. (2004a). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, *9*(1), 97–129. https://doi.org/10.1075/ijcl.9.1.06gri

Gries, S. T., & Stefanowitsch, A. (2004b). Co-varying collexemes in the *into*-causative. In M. Achard & S. Kemmer (Eds.), *Language, culture, and mind* (pp. 225–236). CSLI.

Gries, S. T., & Wulff, S. (2009). Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics*, *7*, 163–186. https://doi.org/10.1075/arcl.7.07gri

Gyllstad, H. (2019). Measuring knowledge of multiword items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 387–405). Routledge.

Gyllstad, H., & Schmitt, N. (2019). Testing formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 174–191). Routledge.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins.

Jarvis, S., & Daller, M. (2013). *Vocabulary knowledge: Human ratings and automated measures*. Benjamins.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. McGraw-Hill.

Lardière, D. (2014). Linguistic approaches to second language morphosyntax. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 163–176). Routledge.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, *33*(3), 299–320. https://doi.org/10.1093/applin/ams010

Oller, J. (1979). *Language tests at school: A pragmatic approach*. Longman.

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

Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, *15*(1), 29–43. https://doi.org/10.1080/15434303.2017.1405421

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

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149. https://doi.org/10.1017/S0267190512000098

Paquot, M., Hubert, N., & Gries, S. T. (in press). Using syntactic co-occurrences to trace phraseological complexity development in learner writing: Verb + object structures in LONGDALE. In B. L. Bruyn & M. Paquot (Eds.), *Second language acquisition and learner corpora*. Cambridge University Press.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Native-like selection and native-like fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). Longman.

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*, 140–162. https://doi.org/10.1075/arcl.7.06rom

Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing*, *34*(4), 477–492. https://doi.org/10.1177/0265532217711431

Rubin, R., Housen, A., & Paquot, M. (in press). Phraseological complexity as an index of L2 Dutch writing proficiency: A partial replication study. In S. Granger (Ed.), *Perspectives on the second language phrasicon: The view from learner corpora*. Multilingual Matters.

Ruegg, R., Fritz, E., & Holland, J. (2011). Rater sensitivity to qualities of lexis in writing. *TESOL Quarterly*, *45*(1), 63–80. https://doi.org/10.5054/tq.2011.240860

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing, and use*. John Benjamins.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Schmitt, N., & Schmitt, D. (in press). *Vocabulary in language teaching* (2nd ed.). Cambridge University Press.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Siyanova-Chanturia, A., & Pellicer-Sánchez, A. (2018). *Understanding formulaic language: A second language acquisition perspective*. Routledge.

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

Weiler, T. (2018). *Investigating the construct tested through four item types used to assess lexicogrammatical competence in English as a foreign language* (Unpublished doctoral dissertation). Lancaster University, UK.

Wulff, S., & Gries, S. T. (2011). Corpus-driven methods for assessing accuracy in learner production. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 61–87). John Benjamins.

Wulff, S., & Gries, S. T. (2015). Prenominal adjective order preferences in Chinese and German L2 English: A multifactorial corpus study. *Linguistic Approaches to Bilingualism*, *5*(1), 122–150. https://doi.org/10.1075/lab.5.1.05wul