



This article is part of the topic “More Than Words: The Role of Multiword Sequences in Language Learning and Use,” Morten H. Christiansen and Inbal Arnon (Topic Editors). For a full listing of topic papers, see <http://onlinelibrary.wiley.com/doi/10.1111/tops.2017.9.issue-2/issuetoc>

# More Than Words: The Role of Multiword Sequences in Language Learning and Use

Morten H. Christiansen,<sup>a,b</sup> Inbal Arnon<sup>c</sup>

<sup>a</sup>*Department of Psychology, Cornell University*

<sup>b</sup>*Centre for Interacting Minds & School of Communication and Culture, Aarhus University*

<sup>c</sup>*Department of Psychology, Hebrew University of Jerusalem*

Received 3 October 2016; accepted 10 March 2017

---

## Abstract

The ability to convey our thoughts using an infinite number of linguistic expressions is one of the hallmarks of human language. Understanding the nature of the psychological mechanisms and representations that give rise to this unique productivity is a fundamental goal for the cognitive sciences. A long-standing hypothesis is that single words and rules form the basic building blocks of linguistic productivity, with multiword sequences being treated as units only in peripheral cases such as idioms. The new millennium, however, has seen a shift toward construing multiword linguistic units not as linguistic rarities, but as important building blocks for language acquisition and processing. This shift—which originated within theoretical approaches that emphasize language learning and use—has far-reaching implications for theories of language representation, processing, and acquisition. Incorporating multiword units as integral building blocks blurs the distinction between grammar and lexicon; calls for models of production and comprehension that can accommodate and give rise to the effect of multiword information on processing; and highlights the importance of such units to learning. In this special topic, we bring together cutting-edge work on multiword sequences in theoretical linguistics, first-language acquisition, psycholinguistics, computational modeling, and second-language learning to present a comprehensive overview of the prominence and importance of such units in language, their possible role in explaining differences between first- and second-language learning, and the challenges the combined findings pose for theories of language.

*Keywords:* Multiword sequences; First-language learning; Second-language learning; Computational model; Generative grammar; Usage-based approach

---

## 1. Introduction

The infinite productivity of linguistic expressions is one of the most celebrated features of human language. It has long been noted that such unbounded linguistic productivity must rely on speakers' and listeners' ability to "make infinite employment of finite means" (von Humboldt, 1836/1999: p. 91). Within linguistics and cognitive science more generally, single words and rules for combining them are often viewed as the finite means for infinite linguistic expressivity. From this perspective, words and rules are the fundamental building blocks of language (Pinker, 1999). Multiword sequences are seen as either constructed from these basic elements or sidelined to the periphery as largely irrelevant exceptions. However, evidence from corpus analyses suggests that language is more repetitive than previously expected and that recurring multiword sequences are far from being mere rarities. For example, based on a small-scale analysis of a corpus of utterances from the TV show *Wheel of Fortune*, Jackendoff (1997) estimates that multiword sequences are as plentiful as single words in everyday language use and concludes that "they are hardly a marginal part of our use of language" (p. 156). Similar corpus investigations show that multiword sequences constitute a high proportion (up to 50%) of the language produced by native speakers in both written and spoken modalities (e.g., DeCock, Granger, Leech, & McEnery, 1998), a pattern that is found across languages (see Conklin & Schmitt, 2012, for a review). That is, speakers seem to know (and use) many recurring multiword sequences.

Of course, just as the number of words that people know is difficult to assess (Lieberman, 1989), so is the number of multiword sequences speakers are familiar with hard to estimate (Church, 2013). Complicating matters further, researchers from various disciplines often refer to multiword sequences using different terms. The set of terms used in this topic alone includes "multiword construction" (Culicover, Jackendoff, & Audring, 2017; Ellis & Ogden, 2017), "multiword unit" (Arnon & Christiansen, 2017; Geeraert, Newman, & Baayen, 2017), "multiword chunks" (McCauley & Christiansen, 2017), "multiunit string" (Theakston & Lieven, 2017), and "formulaic sequence" (Wray, 2017) (see Table 1 for additional terms). Importantly, the differences in the specific terms used in this topic to refer to multiword sequences illustrate the breadth of theoretical perspectives and backgrounds of the contributing authors.

Although many of these terms stress the notion of multiple *words*, in some languages it may be more appropriate to talk about multi-*morphemic* sequences (Peters, 1983; see also Theakston & Lieven, 2017). For example, in agglutinative and polysynthetic languages a single long word (created by combining a set of morphemes) may express what in English is conveyed using a multiword sequence or even a whole sentence. Accordingly, corpus analyses of Turkish—an agglutinating language—document the existence of multimorphemic sequences that are similar in function to multiword sequences in English

Table 1  
Some commonly used terms for multiword sequences

Multiword lexemes	Prefabricated phrases
Multiword phrases	Frozen phrases
Multiword strings	Lexicalized phrases
Multiword expressions	Collocations
Multiword constructions	Multiword collocations
Multiword chunks	Formulaic sequences
Multiword units	Formulaic expression
Multiword building blocks	Fixed expressions
Multiunit string	Listemes

*Note.* For additional example terms, see Wray (2002), fig. 1.2.

(Durrant, 2013). However, for simplicity we will use “multiword sequences” here as a cover term that is also intended to include such multimorphemic units. Thus, as a first approximation, we can think of a multiword sequence as a continuous or discontinuous string of meaningful elements commonly interpreted together as a single unit, in some cases allowing modifications of specific elements (see Table 2 for alternative definitions).

The overarching goal of this special topic is to take the prevalence of multiword sequences as a starting point and explore their possible implications for language learning and use, including for the nature of linguistic representations and the computational mechanisms that may underlie our hallmark linguistic productivity. Importantly, the research reported here on multiword sequences is not only just of theoretical concern to linguists and psychologists studying language learning and use but also to cognitive scientists more generally. For example, multiword sequences constitute an important research topic in computer science, where they have been famously flagged as “a pain in the neck” (Sag, Baldwin, Bond, Copestake, & Flickinger, 2002), standing in the way of developing large-scale natural language processing systems. The challenge for computer scientists is that multiword sequences are highly idiosyncratic, making them difficult to capture using standard machine learning techniques. Their ubiquity and heterogeneous properties (see Culicover et al., 2017) further exacerbate this problem.<sup>1</sup>

Table 2  
Some definitions of multiword sequences

“a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.” (Wray, 2002: p. 9)
“idiosyncratic interpretations that cross word boundaries (or spaces)” (Sag et al., 2002: p. 2)
“a multi-morphemic unit memorized and recalled as a whole, rather than generated from individual items based on linguistic rules.” (Myles, Hooper, & Mitchell, 1998)
“a multi-morphemic phrase or sentence that . . . has become available to a speaker as a single prefabricated item in her or his lexicon.” (Peters, 1983: p. 2)
“a unit of clause length or longer whose grammatical form and lexical content is wholly or largely fixed” (Pawley & Syder, 1983: p. 191)

Multiword sequences are also relevant to research in education where they have been documented to pose challenges for second-language (L2) learners (e.g., Wray, 1999). For example, as discussed further in Arnon and Christiansen (2017), L2 learners tend to produce fewer multiword sequences than native speakers and are more likely to produce word combinations that sound odd to native speakers (e.g., *cause happiness, put more attention to*). Even researchers in literacy development might benefit from further insights into multiword sequences, as text-based chunking has been shown to facilitate reading (e.g., Carver, 1970; Rasinski, 1989). Moreover, as discussed by Wray (2017), the study of multiword sequences can provide insight into social-communicative behavior of potential interest to sociologists and sociolinguists alike. In the remainder of this brief introductory article, we delve further into the specific contributions to cognitive science reflected by the articles in this topic.

## 2. The relevance of multiword sequences for cognitive science

The importance of multiword sequences to the study of language has been recognized for more than a century (see Wray, 2002, for a review), going back to de Saussure (1916), who noted that they provide a “shortcut” that applies to “a whole cluster of signs, which then becomes a simple unit” (p. 177). Two-thirds of a century later, multiword sequences were recognized as crucial not only for language acquisition (e.g., Peters, 1983) but also for native-language fluency (e.g., Pawley & Syder, 1983). However, the general impact on mainstream cognitive science research has until recently remained relatively minimal. Here, we highlight several ways in which the research presented in this topic addresses key issues of relevance to the broader cognitive science community.

### 2.1. Implications for linguistics

Research on multiword sequences has already had considerable impact on theories of language, from functional and cognitive-linguistic approaches to language (e.g., Goldberg, 2006; Langacker, 1987) to generative grammar (e.g., Culicover & Jackendoff, 2005; Jackendoff, 1997). Culicover et al. (2017) and Wray (2017) provide an extensive sampling of the different types of multiword sequences and their various functions from two different theoretical perspectives. In both cases, the prevalence of multiword sequences is suggested to call for a reappraisal of the standard words-and-rules perspective.

From a generative viewpoint, Culicover et al. (2017) note that although multiword constructions cannot generally be accounted for by simple compositional processes, they still follow standard grammatical constraints. They argue, though, that multiword sequences pose problems for linguistic accounts building on the traditional notions of procedural rules used to derive the syntactic structure of a sentence (e.g., rewrite rules such as,  $S \rightarrow NP VP$ , indicating that a sentence can be rewritten as a noun-phrase followed by a verb-phrase). Instead, Culicover et al. suggest that rules function as schemas and templates that can motivate or support fragments of well-formed expressions. Multiword

sequences are viewed as having the same status as morphologically complex words, stored in the lexicon. In this sense, multiword constructions are viewed as being parasitic on the productivity of grammar, relying on the productive use of compositionally interpreted expressions. This account challenges notions of rule-based grammar formalisms often used in cognitive science, from mainstream generative grammar (e.g., Chomsky, 1995) to computational linguistics (e.g., Clark, Fox, & Lappin, 2010).

Whereas Culicover and colleagues seek to understand the impact of multiword sequences on the notion of grammar, Wray (2017) aims to understand their communicative function using insights from cognitive linguistics and socio-pragmatics; that is, how their use relates to everyday social practices. She highlights the importance of considering the communicative impact of a speech event; that is, how successful it is in achieving the speakers' goal (see Rendall, Owren, & Ryan, 2009; for a discussion of animal communication in a similar light). She surveys a number different functions that formulaic sequences may have (e.g., to maintain fluency) and which may help alleviate various cognitive pressures (e.g., the pressure from dealing with language in the here-and-now; Christiansen & Chater, 2016). A key function of multiword sequences may be to maintain general fluency (Pawley & Syder, 1983)—especially when the speaker is under pressure. For example, people with dementia may rely more on formulaic sequences to help overcome general processing problems, whereas an L2 learner may not have acquired a sufficient number of them to improve fluency. Thus, Wray's article suggests that to fully understand the broad extent of multiword sequences, cognitive scientists need to move beyond purely formal approaches and incorporate communicative function into their accounts.

## 2.2. Implications for language acquisition

Multiword sequences have played an important role in theories of language acquisition for quite some time (e.g., Peters, 1983). Theakston and Lieven (2017) provide a broad overview suggesting that much of child language can be described in terms of reuse of multiword strings. They highlight the importance of learning abstractions over these strings, resulting in slot-and-frame type schemas (e.g., generalizing *Where's mommy?* to *Where's PERSON?*). They note how children's reliance on multiword sequences show up in their error patterns, such as when children make more *me-for-I* errors (e.g., *Me do it*) when their caregivers produce a relatively high number of sentences in which *me* appears pre-verbally (e.g., *Let me do it*). Theakston and Lieven's review does not just focus on English but also considers other languages as well. They conclude that despite the relative paucity of work on languages with complex morphology and flexible word order, the available evidence still assigns a key role to multiunit strings, highlighting the importance of studying such units and understanding their function.

Similarly, Ellis and Ogden (2017) also approach multiword strings from within a general usage-based approach, but rather than discussing their general role in acquisition, they delve into the detailed patterns of verb argument constructions in English, such as Verb *about* Noun (e.g., *dream about summer*). They first demonstrate that the verbs used

in these constructions have a Zipfian distribution,<sup>2</sup> indicating that a small number of high-frequency verbs accounts for most of the use of these constructions. Ellis and Ogden then report results from corpus analyses showing that child use of these constructions closely matched the frequency with which they occurred in the speech to them from the adults around them. Further psycholinguistic evidence from processing experiments demonstrates that people are highly sensitive to the distributional patterns of the specific verb argument constructions.

Together, the articles by Theakston and Lieven (2017) as well as Ellis and Ogden (2017), along with the contributions by Arnon and Christiansen (2017) as well as McCauley and Christiansen (2017) underscore the fundamental importance of multiword sequences for understanding first-language (L1) acquisition. Given the broad interest in language acquisition within the cognitive science community, these four articles should provide much food for thought, while also challenging several common assumptions about the nature of linguistic representations and the primacy of words in the process of language acquisition.

### *2.3. Implications for second-language learning*

Although multiword sequences appear to facilitate both L1 acquisition as well as adult L1 usage, the picture is more complex for L2 learning (see also Wray, 2002, 2017). Arnon and Christiansen (2017) explore the role of multiword sequences in explaining L1-L2 differences in learning. They suggest that part of the difficulties that L2 learners face may be attributed to their problems with picking up on and learning from multiword sequences. As L2 learners already know that words exist, they may tend to focus on individual words rather than multiword combinations. This may prevent them from learning certain grammatical patterns, such as grammatical gender, as suggested by a recent artificial language learning study (Arnon & Ramscar, 2012). Arnon and Christiansen draw on psycholinguistic, developmental, and computational findings to support the prediction that L2 learners use multiword units in different ways from L1 learners. By emphasizing the role of multiword units as building blocks for language acquisition and use, this perspective promises to offer new insights into the challenges faced by L2 learners.

McCauley and Christiansen (2017) provide a direct test of the hypothesis that L2 learners may use multiword sequences differently than L1 speakers. They employ a computational model—the Chunk-based Learner (CBL; McCauley & Christiansen, 2011)—inspired by the real-time processing constraints on language (Christiansen & Chater, 2016). The model learns incrementally to chunk incoming words into multiword building blocks, and it is able to model both aspects of comprehension and production when exposed to corpora of child-directed speech. McCauley and Christiansen use CBL to assess the “chunkedness” of speech produced by children, native adult speakers, and adult L2 learners. The results indicate that the speech of L2 learners appears to involve less reuse of multiword chunks compared to both children and native adult speakers. Additional simulations indicated that when L2 learners do use multiword sequences, these tend to rely more on simple overall frequency of the multiword unit, rather than the statistical relationship between the

component words as is characteristic of L1 learners (see also, Ellis, Simpson-Vlach, & Maynard, 2008).

The two articles by Arnon and Christiansen (2017) and McCauley and Christiansen (2017) combine to highlight the importance of using theoretical insights about the role of multiword units in L1 acquisition to understand the challenge of L2 learning. This novel perspective should therefore be of interest to cognitive scientists working on L2 learning, as well as to researchers in education more generally (see also Wray, 2017).

#### 2.4. *Implications for computational approaches to language*

A key issue in cognitive science is the nature of the representations and mechanisms that subserve language and cognition. Although the exact details may differ, most of the papers in this topic agree that multiword sequences in one way or another are represented in our brains. However, Geeraert et al. (2017) take a different stance. Focusing on idioms—often considered the canonical stored multiword unit—they report rating data suggesting that people are willing (to some degree) to accept non-canonical variations in idioms. The results were then simulated using a naïve discriminative learning model that maps letter pairs onto semantic features (corresponding to content words + the determiners *a* and *the*). Geeraert et al. conclude from their successful simulations that a direct representation of multiword sequences is not required.

In contrast, Arnon and Christiansen (2017) propose that representations of multiword units emerge as consequence of two separate processes, involving either (a) undersegmentation where children fail to separate out multiple words from one another, or (b) chunking whereby frequently co-occurring word sequences are grouped together. The CBL model by McCauley and Christiansen (2017) captures the latter chunking process (see Monaghan & Christiansen, 2010; for a model of undersegmentation). A possible advantage of this model is that it can use distributional information acquired in the service of comprehension to assist in the production of language (while capturing the comprehension-production asymmetry observed in language acquisition; Chater, McCauley, & Christiansen, 2016). Although the naïve discriminative learning model used by Geeraert et al. (2017) might be able to model aspects of comprehension, it seems unlikely that the same model would be able to generalize what it has thus learned to produce grammatically correct sentences.

### 3. Looking ahead

These are exciting times for the study of language. Recent years have seen the development of new theoretical perspectives and empirical paradigms that allow us to take a new look at old questions. These developments are driven, among other things, by the existence of large corpora, new experimental techniques, and more accessible statistical and computational tools, as well as by the growing number of collaborative interactions between computer scientists, linguists, and psychologists. This special topic brings

together various hypotheses and data about the role of multiword sequences in language. In doing so, the papers provide multiple fresh perspectives on the age-old debate about the representations and mechanisms responsible for language productivity. Moving away from the classic words-and-rules approach, the findings provide strong evidence for the prevalence of multiword units in language; their impact on child language learning; and their possible role in explaining L1-L2 differences. Thus, these findings join the growing literature on the importance of larger patterns in language more generally—in machine learning, corpus linguistics, the study of formulaic language, and many more—pointing to new avenues of research within the cognitive science of language.

## Acknowledgment

This work was partially supported by BSF grant number 2011107.

## Notes

1. Contributions to recent special issues of computational linguistics journals (e.g., Ramisch, Villavicencio, & Kordoni, 2013; Villavicencio, Bond, Korhonen, & McCarthy, 2005) have underscored the immense scale of the challenge for computer science.
2. If words from a sufficiently large corpus are ranked in a table according to their frequency, then Zipf's (1935) law states that the frequency of a word is inversely proportional to its rank in the table. This means that the most frequent word will occur about twice as often as the second most frequent word, which in turn occurs twice as often as the fourth most frequent word, and so on. We note here that multiword sequences, more generally, provide a better fit with a Zipfian distribution than do single words (Williams et al., 2015). This underscores the importance of multiword building blocks for normal language use.

## References

- Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, 122, 292–305.
- Carver, R. P. (1970). Analysis of “chunked” test items as measures of reading and listening comprehension. *Journal of Educational Measurement*, 7, 141–150.
- Chater, N., McCauley, S. M., & Christiansen, M. H. (2016). Language as skill: Intertwining comprehension and production. *Journal of Memory and Language*, 89, 244–254.
- Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.
- Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral & Brain Sciences*, 39, e62.

- Church, K. (2013). How many multiword expressions do people know? *ACM Transactions on Speech and Language Processing (TSLP)*, 10(2), 4.
- Clark, A., Fox, C., & Lappin, S. (Eds.) (2010). *The handbook of computational linguistics and natural language processing*. Oxford, England: Blackwell-Wiley.
- Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, 32, 45–61.
- Culicover, P. W., & Jackendoff, R. (2005). *Simpler syntax*. New York: Oxford University Press.
- DeCock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learning English on computer* (pp. 67–79). London: Addison, Wesley, Longman.
- Durrant, P. (2013). Formulaicity in an agglutinating language: The case of Turkish. *Corpus Linguistics and Linguistic Theory*, 9, 1–38.
- Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second- language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 41, 375–396.
- Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. New York: Oxford University Press.
- Jackendoff, R. (1997). *The architecture of the language faculty*. Cambridge, MA: MIT Press.
- Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*. Stanford, CA: Stanford University Press.
- Liberman, M. (1989). *How many words do we know?* Invited talk delivered at the 27th Annual Meeting of the Association for Computational Linguistics. Vancouver, BC.
- McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCINO model. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1619–1624). Austin, TX: Cognitive Science Society.
- Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, 37, 545–564.
- Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning*, 48, 323–363.
- Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards, & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). New York: Longman.
- Peters, A. M. (1983). *The units of language acquisition*. Cambridge, UK: Cambridge University Press.
- Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.
- Ramisch, C., Villavicencio, A., & Kordoni, V. (2013). Introduction to the special issue on multiword expressions: From theory to practice and use. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(2), 3.
- Rasinski, T. V. (1989). Adult readers' sensitivity to phrase boundaries in texts. *Journal of Experimental Education*, 58, 29–40.
- Rendall, D., Owren, M. J., & Ryan, M. J. (2009). What do animal signals mean? *Animal Behaviour*, 78, 233–240.
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. Gelbukh (Ed.), *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 1–15). Berlin: Springer.
- de Saussure, F. (1916). *Course in general linguistics*. New York: McGraw-Hill.
- Villavicencio, A., Bond, F., Korhonen, A., & McCarthy, D. (2005). Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language*, 19, 365–377.
- von Humboldt, W. (1836/1999). *On language: On the diversity of human language construction and its influence on the mental development of the human species*. Cambridge, UK: Cambridge University Press. (Original work published 1836)

- Williams, J. R., Lessard, P. R., Desu, S., Clark, E. M., Bagrow, J. P., Danforth, C. M., & Dodds, P. S. (2015). Zipf's law holds for phrases, not words. *Scientific Reports*, 5, 12209.
- Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching*, 32, 213–231.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.
- Zipf, G. K. (1935). *The psycho-biology of language*. New York: Houghton, Mifflin.

## Articles in this topic

- Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining L1-L2 differences. *Topics in Cognitive Science*, 9(3), 621–636.
- Culicover, P. W., Jackendoff, R., & Audring, J. (2017). Multiword constructions in the grammar. *Topics in Cognitive Science*, 9(3), 552–568.
- Ellis, N. C., & Ogden, D. C. (2017). Thinking about multiword constructions: Usage-based approaches to acquisition and processing. *Topics in Cognitive Science*, 9(3), 604–620.
- Geeraert, K., Newman, J., & Baayen, R. H. (2017). Idiom variation: Experimental data and a blueprint of a computational model. *Topics in Cognitive Science*, 9(3), 653–669.
- McCauley, S. M., & Christiansen, M. H. (2017). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science*, 9(3), 637–652.
- Theakston, A., & Lieven, E. (2017). Multiunit sequences in first language acquisition. *Topics in Cognitive Science*, 9(3), 588–603.
- Wray, A. (2017). Formulaic sequences as a regulatory mechanism for cognitive perturbations during the achievement of social goals. *Topics in Cognitive Science*, 9(3), 569–587.