Title

The Role of Working Memory for Syntactic Formulation in Language Production

Journal

Journal of Experimental Psychology Learning Memory and Cognition, 45(10)

ISSN

0278-7393

Authors

Ivanova, Iva
Ferreira, Victor S

Publication Date

2019-10-01

DOI

10.1037/xlm0000672

Peer reviewed

RUNNING HEAD: Working memory in syntactic formulation

The role of working memory for syntactic formulation in language production

Iva Ivanova* and Victor S. Ferreira

Department of Psychology, University of California, San Diego, 9500 Gilman Dr, La Jolla, CA, 92093, USA

*Corresponding author, now at
500 W. University Ave
Department of Psychology
University of Texas at El Paso
El Paso (TX), 79962
USA
Tel.: +1 915-747-7365
Fax.: +1 915-747-6553
Email: imivanova@utep.edu

Abstract

Four picture-description experiments investigated if syntactic formulation in language production can proceed with only minimal working memory involvement. Experiments 1-3 compared the initiation latencies, utterance durations and errors for syntactically simpler picture descriptions (adjective-noun phrases, e.g., *the red book*) to those of more complex descriptions (relative clauses, e.g., *the book that is red*). In Experiment 4, the syntactically more complex descriptions were also lexically more complex (e.g., *the book and the car* vs. *the book*). Simpler and more complex descriptions were produced under verbal memory load consisting of two or four unrelated nouns, or under no load. Across experiments, load actually made production more efficient (as manifested in shorter latencies, shorter durations or both), and sped up the durations of relative clauses more than those of adjective-noun phrases. The only evidence for disproportional *disruption* of more complex descriptions by load was a greater increase of production errors for these descriptions than for simpler descriptions under load in Experiments 2 and 4. We thus conclude that syntactic formulation in production (for certain constructions or in certain situations) can proceed with minimal working memory involvement.

Keywords: verbal load; dual task; picture naming; speech planning; internal monitoring

Introduction

We listen to podcasts while driving and scan emails while talking on the phone, thinking we are doing both perfectly. But multitasking is detrimental to performance (e.g., Redelmeier & Tibshirani, 1997), as concurrent activities impose demands on our limited cognitive resources. Speaking, an activity we often perform concurrently with others, is no exception: While talking and doing something else at the same time, we might become less able to remember what we want to say, produce shorter or simpler utterances, more clichés, become more disfluent, make more mistakes or speak more slowly (Allport, Antonis, & Reynolds, 1972; Kemper, Herman, & Lian, 2003; Norman & Bobrow, 1975; Power, 1985; Shaffer, 1975; Yngve, 1973). Such evidence suggests that language production is a limited-capacity system and requires processing resources, among them working memory. But do all component processes of language production always require working memory resources? Specifically, what is the role of working memory for syntactic formulation in language production, and can syntactic formulation proceed with minimal working memory involvement? We address these questions here by asking participants to produce phrase-length picture descriptions with or without verbal memory load.

To produce an utterance, speakers conceptualize their intended message, select appropriate lexical items and build syntactic structure, activate the corresponding phonological and metrical structure, and finally engage in articulation (Levelt, 1989; Bock & Levelt, 1994). The process of building syntactic structure (syntactic formulation) involves assigning grammatical functions (e.g., subject, object) to the selected lexical items, computing the relationships between grammatical constituents and their linear order, and assigning the appropriate morphology (Bock, 1990; Bock & Levelt, 1994; Chang, Dell, & Bock, 2006). Further, computing the relationship between grammatical constituents involves construction of hierarchical representations. This is necessary because the structural relationships between sentence constituents, at least sometimes, are not reflected in their linear order in an utterance.

Here we contrast two mutually-exclusive accounts of the role of working memory in syntactic formulation. On a *memory-heavy account*, syntactic formulation requires working memory (a limited-capacity cognitive mechanism responsible for short-term storage and manipulation of information with the aim of performing a task: Baddeley, 1986; 1995). Such involvement could be graded depending on the nature and complexity of the computations, but even relatively simple syntactic computations in production involve some detectable amount of working memory.

Working memory could be necessary for the construction of hierarchical structures, by retaining already constructed constituents or chunks of structure until the addition of subsequent ones, or by retaining simultaneously activated alternative structures until committing to a single one (Myachykov, Scheepers, Garrod, Thompson, & Fedorova, 2013). Working memory could also be necessary for computing the linear order of syntactic constituents from previously constructed hierarchical representations, insofar as such representations do not always map directly onto constituents' eventual linear order. For example, in *The person behind Carla gave me a surprised look*, *person* may need to be kept in working memory until the verb of which it is a subject, *gave*, is formulated for production.

Keeping constituents in working memory might also be necessary for pre-articulation monitoring. For example, the perceptual loop monitoring theory (Levelt, 1989; see Postma, 2000, for a review of monitoring theories) postulates that already prepared utterance components are held in a phonological-loop buffer, played back through inner speech and monitored via the comprehension system. Other theories postulate that monitoring is performed by separately checking the output of each production stage (Laver, 1980), by the detection of conflict between intended and produced utterances (Nozari, Dell, & Schwartz, 2011), or discrepancies between predictions based on forward models of utterances and actual utterances (Pickering & Garrod, 2014). Independently of how monitoring is performed, planned utterance components (including their syntactic constituent information) might need to be held in working memory until they are checked for errors.

Working memory also seems necessary if syntactic formulation proceeds by retrieval of stored syntactic frames from long-term memory instead of computing them online. This might happen for syntactically formulaic utterances (e.g., a response such as *I am doing well* or *I am doing great* to the question *How are you today?*). Some grammatical theories are consistent with the possibility that structures are retrieved as a whole (e.g., Construction Grammar, Goldberg, 1995; 2006; Two-stage Competition model, consisting of a structure-selection stage and a structure-planning stage, Segaert, Menenti, Weber, & Hagoort, 2011; Segaert, Wheeldon, & Hagoort, 2016). If so, abstract syntactic frames, once retrieved, would *have to* be maintained active in working memory until retrieval of all the lexical items necessary to fill these frames.

Yet another possibility is the Bayesian framework of Fragment Grammars (O'Donnell, 2015), in which computations of structures are stored in memory as a function of how likely they are to be reused in the future. The more frequent the estimated need for reuse of a computation, the more likely it is to be stored in memory, to maximize efficiency and avoid expending resources to perform a frequent computation from scratch. When reusing stored computations, working memory would be required if whole computational sequences are retrieved at the same time, to maintain them active until their turn to be performed.

Some support for the memory-heavy account of syntactic formulation comes from word order choice in spoken production. For example, heavy noun-phrase shifts occur in sentences such as *The chair brought to our attention the pressing need to remodel the available space*, and refer to speakers' tendency to produce complex noun phrases (e.g., *the* *pressing need to remodel the available space*) in a non-canonical sentence-final position. In a corpus analysis, Wasow (1997) found higher rates (47%) of heavy noun-phrase shifts in (transparent) collocations (e.g., *brought to our attention*), which are likely planned together, than in non-collocations (15%). This suggests that that speakers tend to avoid having to keep part of the collocation in memory until after production of the longer noun phrase. Similarly, Bader (2017) showed that participants with shorter working memory spans (as measured by a reading span task) were more likely to produce extrapositions (memory-saving structures in that heavy constituents appear to the right of their canonical position). Both studies suggest that speakers' working memory demands affect syntactic formulation in production, and specifically the linear ordering of hierarchically planned syntactic constituents.

Support for the involvement of working memory in syntactic formulation also comes from studies examining how concurrent working memory load and working memory capacity influence subject-verb agreement errors in written and spoken production. Fayol, Largy, and Lemaire (1994) found that participants made more agreement errors in French written recall when they kept in memory a list of five monosyllabic words or counted series of 6-10 clicks than when they did not have additional memory load (see also Hupet, Fayol, & Schelstraete, 1998). Hartsuiker and Barkhuysen (2006) found that participants with low working memory spans (as measured by the Daneman & Green's (1986) speaking span test) made more agreement errors in Dutch spoken fragment completion under a three-word memory load than low-span participants under no load (while high-span participants showed no difference). These results might suggest that working memory is required for a controlled monitoring process checking the results of the automatically completed agreement against grammatical knowledge and correcting any errors (Fayol et al., 1994), as well as a process of syntactic integration, in which the number specifications of the head noun phrase (e.g., *the key* in *the key of the cabinets is*) and the local noun phrase (e.g., *the cabinets*) are reconciled by a competition process (Hartsuiker & Barkhuysen, 2006). Working memory involvement is also evidenced by patterns of attraction errors (e.g., *the key of the cabinets are*) in healthy and brain-injured speakers. Such errors index similarity-based interference between head and local nouns and hence suggest that these nouns are held in working memory until the production of verbs that agree with them (Badecker & Kuminiak, 2007; Slevc & Martin, 2016).

However, there is an alternative theoretical view about the involvement of working memory in syntactic formulation in language production. Bock (1982) pointed out that the language production system is a limited-capacity system which requires a constant need to balance the locus of deployment of the limited available cognitive resources. For example, content planning is generally effortful, and can be more effortful in certain situations (e.g., explanations require more resources than descriptions, Goldman-Eisler, 1968; Levin, Silverman, & Ford, 1967; see also Deese, 1978, 1980). For production not to suffer, the system needs to minimize resources in other processing components. Bock (1982) proposed that grammatical encoding (including lexical selection and syntactic formulation) might mediate such interplay between automatic and controlled processing. She hypothesized that grammatical encoding *can* proceed by automatic execution of a certain alternative from a set of syntactic options as a way to free up processing resources for other production components that require them in a given situation. That is, the involvement of working memory in syntactic formulation can be minimal (even though it does not have to be so if resources are available). In this proposal, referential (i.e., related to conceptual processing) and phonetic representations are maintained in working memory, but the products of intermediate processing components (such as syntactic formulation) are not and are instead only governed by the language production system itself. Such products are in principle accessible to working memory, but language production routinely proceeds without accessing them.

Following the spirit of Bock's (1982) proposal, we derive a *memory-light account* of syntactic formulation. In this account, syntactic formulation (for certain structures or in certain situations) *can* proceed with only minimal working memory involvement. This might be possible when event structure can map directly onto a syntactic structure, and hierarchically-structured constituents can map directly onto their eventual linear order. For such utterances, once utterance semantics are determined, a syntactic formulation plan would be executed by the language production system relatively automatically; constituents would not need to be held in working memory. In such situations, internal monitoring (assumed to require working memory, e.g., Fayol et al., 1994) might also be minimized. Although there is ample evidence for internal monitoring (review in Postma, 2000), it might not be performed thoroughly in some situations. For example, when speakers are under time pressure to initiate their utterances, they are more likely to fail to take into account common ground, presumably because of a monitoring failure (Horton & Keysar, 1996). Also, studies eliciting speech errors (e.g., the spoonerism of "barn door") include a production deadline to ensure a higher number of such errors, which are assumed to reflect monitoring failures (Baars, Motley & Mackay, 1975; Hartsuiker, Corley, & Martensen, 2005; although note that these studies do not specifically address the monitoring of syntactic structure). Retrieval from long-term memory of stored computations of syntactic structure as in Fragment Grammars (O'Donnell, 2015) would also be compatible with the memory-light account if computations are retrieved and executed one at a time, a process that could be implemented by the production system with only minimal working memory involvement.

Note that the memory-light account does not deny working memory involvement in some syntactic-formulation situations – for example, when the eventual linear order does not map directly onto hierarchical constituent structure (i.e., in long-distance dependencies), or when several structural options need to be maintained active. Instead, this account simply *allows* for syntactic formulation to proceed without substantial working memory involvement. It is further possible that memory-light syntactic formulation might occur as a system mechanism to free up processing resources for other (resource-demanding) production processes (as proposed by Bock, 1982) – and as such is situation-specific and not construction-specific – but we do not test this possibility here.

Results from sentence-generation studies are consistent with the memory-light account. Power (1985) asked participants to generate spoken sentences from two semantically related (e.g., street - road) or unrelated nouns (e.g., editor - basket), either under no memory load, or keeping in memory three or six digits. Most relevant for present purposes, sentences produced under load were marginally shorter than those produced under no load, but did not differ in number of clauses or judged complexity. If working memory played a role in syntactic formulation, sentences generated under load would have been less structurally complex. Counterintuitively, sentences' initiation latencies *decreased* under both 3- and 6-digit load, likely driven by a strategy to complete the sentence generation before forgetting the digits. Such speeding up under load, however, came at the cost of conceptual planning: As digit load increased, the generated sentences were more stereotyped and conveyed less information; thus, Power (1985) concluded that working memory demands affected conceptual processing in sentence generation. This conclusion is consistent with the memory-light account in implying that syntactic formulation in production can proceed relatively independently from working memory. (See also Kellogg, 2004, for similar findings in written sentence generation, but note that he placed working memory demands at the level of lexical selection.)

Further, across three experiments, Bock and Cutting (1992) found only a weak relationship between participants' agreement errors and working memory capacity (speaking span, from Daneman & Green, 1986).

In sum, there is evidence for both the memory-heavy and memory-light accounts of working memory involvement in syntactic formulation. However, because of the differences in utterance structures and task contexts across studies (e.g., the possibility for similarity-based interference in some studies), the role of working memory for syntactic formulation in production remains unclear. Indeed, Fyndanis, Arcara, Christidou, and Caplan (in press) showed that subject-verb agreement errors without the possibility for similarity-based interference are not observed more for individuals with lower working-memory capacity such as aphasic speakers (but see Kok, van Doorn, & Kolk, 2007, for evidence that agreement in a similar population and task is affected by processing demands).

Of note, versions of the memory-heavy and memory-light accounts have been discussed in sentence comprehension. On the one hand, Just and Carpenter (1992) proposed that language comprehension is a limited-capacity, resource-demanding process which relies on working memory resources shared with all verbal tasks. Sentences could require more or less working memory capacity depending on their complexity. On the other hand, Caplan and Waters (1999) proposed that basic comprehension processes (*interpretative processes*, which include syntactic processing, lexical and meaning processing, integration with discourse) rely on a highly specialized working memory component that enables them to function relatively automatically (in contrast to post-interpretative processes such as inference drawing or reanalysis, which share resources with other language tasks). Note, however, that the memory-heavy and memory-light accounts presented here do not differ in the domain-generality versus domain-specificity of working memory (of which they are agnostic). In other words, the memory-light account does not assume that syntactic formulation in production engages a domain-specific working-memory system. Instead, in this account, the language production system is a dynamic system that has the capacity to execute portions of formulation in a relatively implicit and automatic fashion (see General Discussion for a discussion of automaticity), without expending considerable cognitive resources (or dedicating considerable attentional focus; see e.g. Engle, 2002) to maintain and manipulate information.

Effects of working memory load on advance planning

The memory-heavy and memory-light accounts should be distinguishable even before speech onset. Speakers plan at least portions of utterances prior to speech onset (e.g., Lindsley, 1975, Smith & Wheeldon, 1999), and effects of working memory seem to be present at this stage. Much evidence suggests that the typical advance planning scope is the first phrase (Martin, Crowther, Knight, Tamborello, & Yang, 2010; Smith & Wheeldon, 1999; Wheeldon, Ohlson, Ashby, & Gator, 2013), even when it is not the head phrase (as in head-final languages such as Japanese: Allum & Wheeldon, 2007; 2009; but see Brown-Schmidt & Konopka, 2008; Griffin, 2001; 2003, for evidence inconsistent with a phrasal planning scope).

Martin, Yan, and Schnur (2014) studied the effects of verbal memory load on advance planning. In response to three-picture displays, participants produced sentences which either began with a simple noun phrase (*The drum is above/below the package and the squirrel*) or with a complex noun phrase (*The drum and the package are above/below the squirrel*). Participants performed this task under spatial memory load (the position of two dots on a grid), verbal load (two words, which could further specifically tap into semantic or phonological processing), or no load. Participants initiated sentences beginning with complex noun phrases more slowly than sentences beginning with simple ones (replicating previous findings, e.g., Smith & Wheeldon, 1999), but sentences beginning with complex noun phrases were not disproportionally affected by load. The authors thus concluded that the well-attested noun-phrase complexity effects on planning did not stem from either a semantic or a phonological source, and, because of this, were not disproportionally affected by their semantic and phonological load. Instead, the authors proposed that noun-phrase complexity effects were syntactic in nature, because their participants made more syntactic (but not lexical) errors under load on sentences beginning with complex noun phrases than those beginning with simple ones.

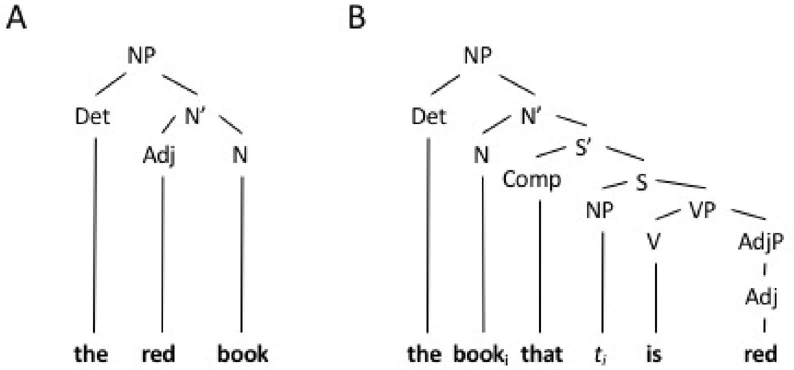
However, there is evidence that extrinsic memory load of the type used by Martin et al. (2014) does not influence advance planning scope (Wagner, Jescheniak, & Schriefers, 2010). These authors showed that planning scope can be narrowed when processing resources are scarce, but only by cognitive load directly implicated in utterance planning (such as an object-size decision which then determines the mention of a size adjective), and not by external digits or adjectives. On the other hand, external load which shares features with elements in the production task (e.g., load consisting of nouns instead of adjectives or digits) *might* lead to a reduction in planning scope.

The present study

We explored the role of working memory for syntactic formulation in language production, aiming to distinguish between the memory-heavy and memory-light accounts. We compared working memory involvement in the production of relative clauses (e.g., *the book that is red*) and of adjective-noun phrases (e.g., *the red book*). Both phrases contain the same content words (e.g., *book* and *red*), but, crucially, relative clauses are more complex than adjective-noun phrases: they contain more nodes in a tree structure (Ferreira, 1991; see Figure 1) and feature morphological agreement, clausal embedding, and the establishment of a dependency between the subject position of the embedded clause and the nominal head of the relative clause. At the same time, both phrases are relatively simple and even the more complex relative clauses, produced in isolation, do not involve dependencies over a long distance. For example, in such phrases there is only one surface constituent (the relative pronoun *that*) separating the head noun (*book*) from the gap that it licenses and the verb (*is*) agreeing with it in number.

Our complexity manipulation is based on the assumption that the greater linguistic complexity of relative clauses relative to adjective-noun phrases leads to a greater psychological complexity because the former requires performing a greater number of computations (agreement, dependency formation). But we note that, in our view, there is no one-to-one relationship between psychological complexity and production speed. This is because production speed might depend on a host of factors (e.g., structure frequency, number of alternatives), and the production system might feature mechanisms allowing it to produce more complex utterances more quickly in certain situations.

Figure 1. Syntactic tree representations for adjective-noun phrases (A) and relative clauses (B).



The properties of adjective-noun phrases and relative clauses allow us to put to the test the memory-heavy and memory-light accounts. Under the memory-heavy account, the production of both adjective-noun phrases and relative clauses would require working memory resources, but more so for relative clauses, because of their greater structural complexity described above. For example, the verb phrase, if planned together with the head noun, might need to be kept in memory during production of the relative pronoun, and the adjective phrase (*red*), if planned together with the head noun (*book*), kept in memory during production of both relative pronoun and verb (cf. Lee, Brown-Schmidt, & Watson, 2013). Under the memory-heavy account, production of adjective-noun phrases would demand less working memory insofar as they consist of a single phrase, do not contain a verb, and the adjective and noun (even if they are planned together: Schriefers & Teruel, 1999) are immediately adjacent.

Conversely, under the memory-light account, neither adjective-noun phrases nor relative clauses involve complexity that would uncontroversially require working memory resources (such as dependencies that arise from a longer-distance anaphors or extrapositions). In this account, the involvement of working memory in the syntactic formulation of both types of phrase would be minimal.

We aimed to distinguish between the memory-heavy and memory-light accounts in four picture-naming experiments, manipulating picture descriptions' syntactic complexity and concurrent verbal working memory load. The basic set-up in all experiments required participants to either memorize unrelated nouns (load trials) or just view a row of *x*s (no load trials), then describe a picture with either an adjective-noun phrase or a relative clause (except in Experiment 4, which involved simple and complex noun phrases), and then report the recall words on load trials. Load trials involved two nouns in Experiments 1, 3 and 4, and four nouns in Experiment 2. The design of all experiments was 2 x 2, with the factors load (no load, load) and syntactic complexity (simpler description, more complex description).

We opted for verbal memory load because it would be more likely to interact with aspects of linguistic processing. Further, we used a type of verbal load (several unrelated nouns) that is commonly used in studies addressing questions similar to ours and has produced the predicted effects (e.g., three nouns in Hartsuiker & Barkhuisen, 2006, for syntactic formulation in production; two nouns in Slevc, 2011, for accessibility effects on syntactic formulation in production; one or three nouns in Fedorenko, Gibson, & Rohde, 2006, for syntactic processing in comprehension). Further, the kind of load we used involves maintenance of information in working memory, which should interfere with the necessary maintenance of any information that (by hypothesis) would be manipulated during syntactic formulation. In other words, we expected memory load requiring maintenance of information to interfere with processes involving both maintenance and manipulation of information as studied here.

We assumed that structural processing in production incurs costs that are separable from both conceptual and lexical processing (Smith & Wheeldon, 2001). Our main dependent measures were initiation latencies (to capture advance planning), utterance durations (assuming that duration lengthening would at least in part reflect planning difficulties, e.g., Ferreira & Swets, 2002; Lee et al., 2013) and production errors (given that increased syntactic complexity leads to greater error rates, Scontras et al., 2014).

Our main prediction for all measures was that a disproportionate slowing and disruption of more complex relative to simpler descriptions under load (statistically, an interaction between load and complexity) would support the memory-heavy account, while an equivalent slowing and disruption of more complex and simpler descriptions under load (statistically, no interaction between load and complexity) would support the memory-light account. (Note that such a pattern could also emerge with secondary tasks not involving working memory, but because of a different kind of interference with linguistic processing than the one assumed here.)

Further, we predicted that concurrent memory load would cause greater slowing and reduced accuracy than no load (statistically, a main effect of load). This prediction is based on (1) the standard assumption in dual-task experiments that concurrently performing two tasks disrupts performance (including slowing it down; non-linguistic performance: e.g., de Fockert, Rees, Frith, & Lavie, 2001; Lavie, Hirst, de Fockert, & Viding, 2004; Stins, Vosse, Boomsma, & de Geus, 2004; linguistic performance: e.g., Ferreira & Pashler, 2002); (2) a conception of the language and working memory systems as limited capacity systems (Baddeley, 1995; Bock, 1982); and (3) numerous studies showing that external working memory load including the type used here slows down linguistic processes (in production: Belke, 2008; Martin et al., 2014; in comprehension: Baddeley & Hitch, 1974; Fedorenko et al., 2006; Fedorenko, Gibson, & Rohde, 2007) and causes more production errors (Fayol et al., 1994; Hartsuiker and Barkhuysen., 2006).

In the context of this study, we considered that a phrasal planning scope would span the whole relative clause, insofar as what would technically be the first phrase contained a single noun (cf. Smith & Wheeldon, 1999). If so, syntactically more complex utterances would be initiated or uttered more slowly and less accurately than syntactically simpler utterances (statistically, a main effect of complexity). Since our target utterances were not matched for length, we performed additional analyses on length-corrected utterance durations, reported after Experiment 4; also note that more complex utterances take longer to produce even when they are length-matched: Scontras et al. (2014). [1]

[1] We also manipulated head noun frequency, to see if the outcome of our load manipulation was influenced by lexical accessibility (Slevc, 2011). It was not: Frequency did not interact with the variables of interest in any of the four experiments we report. Frequency effects were overall weak, possibly globally reduced by interference between the load words and picture-description head nouns.

Experiment 1

In this experiment, participants named pictures with adjective-noun phrases (simple descriptions, e.g., "*the red book*") and relative clauses (complex descriptions, e.g., "*the book that is red*") while keeping in working memory a two-word verbal load (e.g., *collection* and *mountain*).

Method

Participants. Forty-eight undergraduates at the University of California, San Diego (UCSD) participated for course credit. All were native speakers of English. Four further participants were excluded: two because they were non-native speakers of English, one because the data were lost, and one because of consistently producing utterances which were different from the target ones.

Materials. The experimental materials consisted of 48 black-and-white pictures, six color patches and 48 memory word pairs. The pictures were selected from the International Picture Naming Project (IPNP) database (Szekely, Jacobsen, D'Amico, Devescovi, Andonova, et al., 2004); see Table 1 for picture name characteristics. Additionally, half of the pictures had high-frequency names, and half had low-frequency names (frequency difference: $p < .001$; age of acquisition difference: $p < .001$; remaining factors: all *p*s > .2). Twelve additional pictures served as fillers, and another twelve pictures as practice items.

The six color patches were red, green, blue, orange, pink and purple. Each picture was paired with a unique color (e.g., the picture of a book occurred together with the red color patch for all subjects), which produced 48 experimental, 12 filler and 12 practice picture-color combinations. In the experiment proper, each color occurred eight times, and each picture only once.

Table 1

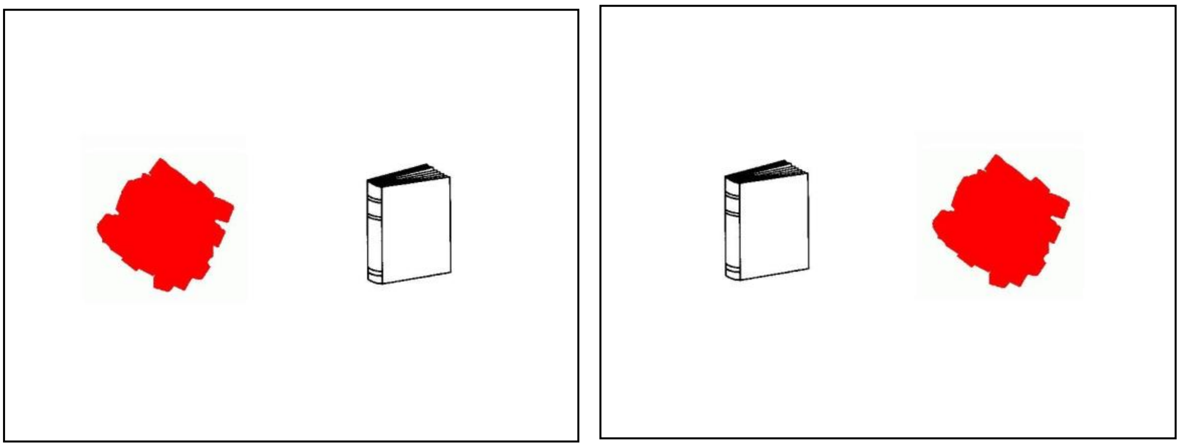
Characteristics of the pictures and picture names used in all experiments

	M	SD	Min	Max
Frequency per million, average	95	132	1.3	557.1
High-frequency names	175	148	51.7	557.1
Low-frequency names	14	9	1.3	39.1
Imageability (100 – 700 scale)	614	21	573	661
Age of acquisition (years)	1.71	.94	1	3
Phonological neighborhood (# of neighbors)	17	8	1	33
Name agreement (# of alternative names)	1.8	1.2	1	5
Length (phonemes)	3.4	.8	1	5
Length (syllables)	1.2	.4	1	2
Visual complexity of pictures	15791	8982	6551	52543

Note: The frequency values were obtained from the SUBTLEX-US corpus for American English (51 million words; Brysbaert & New, 2009; http://expsy.ugent.be/subtlexus/). The imageability, phonological neighborhood and length values were obtained from N-watch (Davis, 2005). The age of acquisition, name agreement and visual complexity values (the latter reflecting image size of the digitalized picture files) were obtained from the IPNP database.

For each experimental picture-color combination, two images were created. In one, the respective color patch was on the left-hand side and the black-and-white line drawing was on the right-hand side (Figure 2, left panel); such images were intended to elicit noun-phrase descriptions such as *the red book*. In the other image, the positions of color patch and picture were reversed (Figure 2, right panel). These images were intended to elicit relative clause descriptions such as *the book that is red*.

Figure 2. Examples of an experimental picture-color combination. Left panel: designed to elicit the noun-phrase description *the red book*. Right panel: designed to elicit the relative-clause description *the book that is red*.



The 48 experimental memory word pairs were formed by combining 96 medium-frequency words (M = 36 per million words, SD = 48.4, values obtained from SUBTLEX-US, Brysbaert & New, 2009) into semantically-unrelated pairs (e.g., *collection mountain*). Pairs were four syllables long on average (range: 3-6), with individual words one to four syllables long. Six additional pairs were used during practice, and six further pairs in filler items. Each color-picture combination was paired with a unique pair of memory words for all subjects (i.e., memory word pairs were not rotated across items). Experimental items thus consisted of a memory word pair and a color-picture display.

Four versions of each item were created by crossing the factors *load* (no load, load) and *description syntactic complexity* (simpler syntax, more complex syntax). The four item versions were used to create four lists such that each item occurred only once per list. Each list contained an equal number of items (12) in each experimental condition. Each subject only saw one list. The load and no load trials were interleaved and presented in random order, with the restriction that no more than three trials of the same type occur in a row.

The four lists were divided into two experimental blocks (such that item order within a block was the same across the four lists). The picture names in the two blocks were matched on all characteristics (all *p*s < .33) and, as much as possible, on semantic category. The two blocks contained an equal number of trials from each experimental condition, an equal number of trials with each color, and an equal number of high- and low-frequency words. Block order was counterbalanced for each of the four lists, thus creating a total of eight lists. Each block began with six filler items which were not analyzed.

Procedure. Participants were tested individually on iMac computers with 21-inch monitors. They read written instructions that stepped them through the procedure, which the experimenter then repeated verbally. A trial had the following components. First, an asterisk was displayed for 500 ms. Then, a row of *x*s (xxxxxxx xxxxxxx) or two memory-load words (e.g., *collection mountain*) were displayed for 5500 ms. To standardize how participants went about memorizing the two words, they were instructed to silently repeat the word pairs five times and lightly tap on the desk with a pen after each repetition, to signal to the experimenter that they were actually repeating the words. Participants then saw a picture-color combination like the one depicted in Figure 1, and were instructed to describe it with a single phrase as quickly and as accurately as they could. They were to say something like "the red couch" if the color patch preceded the drawing (from left to right), or something like "the couch that is red" if the color patch followed the drawing. We asked participants to produce the full instead of abbreviated version of the relative pronoun and verb to discourage the retrieval of *that's* as a single element. Also, pilot participants were instructed to produce "that's" but spontaneously switched to "that is", indicating that "that is" was the preferred alternative.

The color-picture displays remained on the screen until participants had produced a response. Then, the experimenter pressed the space bar and the word "RECALL" appeared on the screen, signaling to participants to say aloud the two words they were asked to memorize (or say "don't remember" if they could not recall any of the two words). If the trial had started with rows of *x*s instead of two words, the experimenter moved on to the next trial.

The experiment began with twelve practice trials, which included an equal number of the experimental conditions and were presented to participants in increasing complexity. If participants changed or omitted a part of the intended descriptions (e.g., omitting the definite article or saying "that's" instead of "that is"), they were corrected. After the practice, participants were familiarized with the names of the pictures they were to see throughout the experiment (48 experimental + 12 filler). To that aim, participants named each of the black-and-white pictures (presented in alphabetical order of the names they were supposed to elicit) and were corrected if they produced a different-from-intended name.

The experiment was presented with DMDX (Forster & Forster, 2003). Participants wore a headset microphone and the program recorded their spoken responses. After the experimental session, participants completed a questionnaire verifying they were native English speakers and probing for their performance strategies and beliefs about the purposes of the experiment.

The study procedures conformed to U. S. federal guidelines for the protection of human subjects and were approved by the UCSD Institutional Review Board. Informed consent was obtained from all participants prior to testing and after the procedures of the study had been fully explained.

Data analysis. *Description latencies, description durations and production errors*. In this and subsequent experiments, we analyzed initiation latencies, utterance durations and rate of production errors for the picture descriptions. Initiation latencies and utterance durations were extracted from the digital recordings generated by DMDX with CheckVocal. Production errors included incorrect objects or colors (e.g., saying "rat" instead of "mouse" or saying "pink" instead of "purple"), incorrect structures (e.g., saying "the red book" instead of “the book that is red” or vice versa, as well as saying “the book *that’s* red instead of “that is red”), disfluencies (e.g., saying “the r-red book” or “the red ba-book”) and self-corrections (saying “the green.. I mean red book”). Only trials on which both load-words were recalled correctly were kept for analyses (but analyses including the trials on which at least one load-word was recalled produced an identical pattern of results). Trials which involved a production error, an experimenter or recording error or were outliers (3 standard deviations above or below the mean) were excluded from the latency and duration analyses.

The latencies, durations and production errors were analyzed with linear or logistic mixed-effects regression (LMER) modeling (Baayen, 2008; Jaeger, 2008) in R. The fixed predictors in these models were load condition (no load, coded as -0.5, load, coded as 0.5), description syntactic complexity (simple, coded as -0.5, complex, coded as 0.5), and their interaction. All predictors were centered around the mean. If the complexity x load interaction was significant, we also ran simple effects models to shed light on the interaction. These models only contained the complexity predictor and the interaction term as predictors; the load predictor was removed (Jaeger, 2013). All models had the maximal random effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013), unless otherwise specified. If a model with the full random effects structure did not converge, we simplified the full model by step-wise removal from the full model of the random effect which accounted for the least variance, and we report the results of the first model that converged.

Recall performance. Additionally, we compared the recall performance after simple and complex descriptions. For this purpose, we ran a logistic mixed-effects regression model on the data from the load condition only (excluding the trials involving production errors), with the fixed predictor syntactic complexity (simple, coded as -0.5, complex, coded as 0.5). We counted as correct the trials on which both memory load words were recalled correctly in either order (coded as 0), and as incorrect, the trials on which only one word was recalled, or no words were recalled (coded as 1).

Results

Data exclusions. The experiment contained 2304 trials. There were 68 incorrect recall trials (2.95% of all trials, or 5.90% of load trials), excluded from the analyses of latencies, durations and errors. Production errors were made on 113 of the remaining trials, and these were excluded from the analyses of response latencies and durations. Twenty further trials were discarded because of experimenter error. In the latency data, another 25 trials were removed as outliers. Of the total of 45 unanalyzed latency trials, 23 involved simple descriptions and 22 complex descriptions. In the duration data, another 66 trials were discarded because the endings of utterances were cut off from the recordings, and 15 trials were removed as outliers. Of the total of 101 unanalyzed durations trials, 47 involved simple descriptions and 54 complex descriptions. In total, 226 trials (9.81%) were discarded from the latency data, and 282 trials (12.24%), from the duration data.

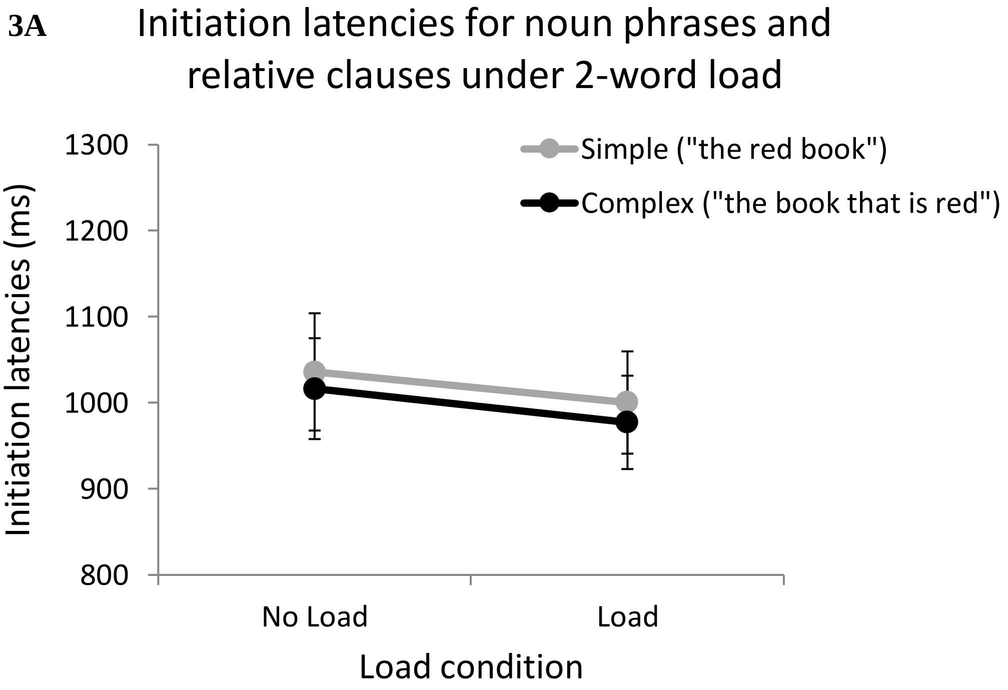
Analyses of description latencies, description durations, and production errors. Descriptive statistics are presented in Figure 3, and inferential statistics are summarized in Table 2. The analyses of initiation latencies (Panel 3A) revealed that participants did not take any longer to initiate complex (relative clause) descriptions than simple (noun phrase) descriptions (complexity was not a significant predictor). Furthermore, they were *faster* to initiate both types of description under load than under no load (load was a significant predictor). Importantly for our purposes here, load did not differentially affect simple and complex descriptions (the complexity x load interaction was not significant).

The analyses of utterance durations (Panel 3B) revealed that participants took longer to utter complex than simple descriptions (complexity was a significant predictor); this is to be expected since complex descriptions contained more words than simple descriptions. More relevantly, participants took no longer to utter their descriptions under load than under no load (load was not a significant predictor). However, participants took less time to utter complex descriptions under load than under no load, but a similar amount of time to utter simple descriptions under load and no load (the complexity x load interaction was significant).

The analyses of production-error rates (Panel 3C) revealed no significant effects.

Recall performance. Recall accuracy was high (94.25%). Participants correctly recalled a similar number of memory load words after producing simple (504, or 94.56% of the valid load trials in this condition) and complex descriptions (513, or 93.96%). In this analysis, syntactic complexity was not a significant predictor [*Estimate* = .31, *SE* = .44, *z* = .70, *p* = .49].

3A



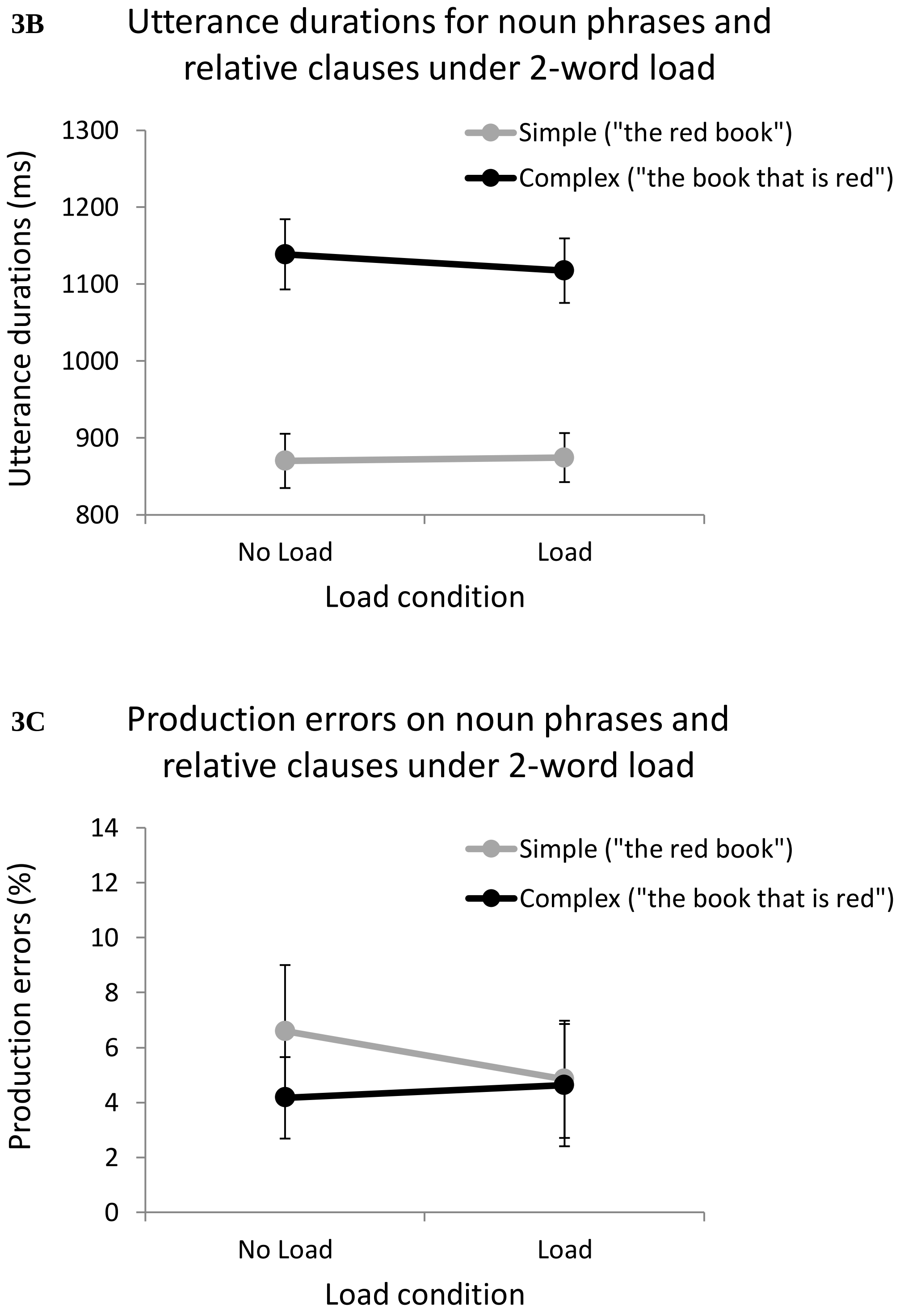


Figure 3. Initiation latencies, utterance durations and production errors in Experiment 1. Panel 3A: initiation latencies; Panel 3B: utterance durations; Panel 3C: production errors. Error bars represent 95% confidence intervals.

Table 2

LMER analyses of initiation latencies, utterance durations and production errors in Experiment 1

Model	Predictors	Estimate	SE	*t or z*	*p*
Initiation latencies					
	Complexity	-21.02	14.80	-1.42	.16
	Load	-36.34	10.41	-3.49	.001
	Complexity * Load	-3.14	20.15	-.16	.88
Utterance durations					
Main model	Complexity	256.20	10.50	24.40	< .001
	Load	-9.66	7.82	-1.24	.22
	Complexity * Load	-27.62	12.02	-2.28	.02
Simple effects of Load	Simple descriptions	-4.17	10.58	-.40	.70
	Complex descriptions	23.45	9.09	2.58	.01
Production errors[a]					
	Complexity	-.29	.24	-1.19	.24
	Load	-.35	.24	-1.43	.15
	Complexity * Load	.43	.44	.98	.33

Note: Shaded rows highlight significant effects.

[a] This model had the following random effects structure:

err ~ load * complexity + (0 + complexity + load:complexity | subj) + (1 | item).

Discussion

At the outset, we predicted that participants would initiate and utter their descriptions more slowly and less accurately under load than under no load, independently of the role of working memory for syntactic formulation in production. Initiation latencies in Experiment 1 showed the opposite pattern: Participants initiated their descriptions more quickly under load than under no load. This result replicates the finding of Power (1985), as well as some results by Antje Meyer and colleagues (personal communication), and suggests that participants might start their utterances faster under load to avoid forgetting the load words (although in our study trial length was controlled by the experimenter). Still, we find it noteworthy that participants had the capacity at all – at least in the current situation – to speed up language production while still maintaining accuracy when processing resources were taxed. This result evidences speakers' flexibility to manage task goals and sub-goals of sentence production in ways that can be optimized to the specific task or context at hand.

Also contrary to our predictions, we found that participants took a similar amount of time to initiate adjective-noun phrases and relative clauses. This result might reflect full planning of adjective-noun phrases (in accordance with Schriefers & Teruel, 1999), but only partial planning of relative clauses, prior to speech onset. Although Smith and Wheeldon (1999) provided evidence for a phrasal scope of advance planning, they showed less thorough advance processing for the second noun in relative clauses than for the second noun in complex noun phrases. In our experiment, the color adjectives in relative clauses might have been similarly planned less thoroughly than these adjectives in adjective-noun phrases.

The design of our description-eliciting displays raises alternative interpretations of the results. First, the entities on our visual displays always appeared, from left to right, in the same order as they were supposed to be produced in linearized descriptions. This might have encouraged a radically incremental processing strategy, such that participants planned only the first content word, or even only the definite article, prior to initiating speech. Some aspects of the results are, however, inconsistent with such a possibility. Adjectives should have been easier to plan because they were repeated multiple times, while each noun occurred only once during the experiment. If only the first content word was planned in both adjective-noun phrases and relative clauses, description latencies for adjective-noun phrases should have been shorter than latencies for relative clauses, inconsistent with what we found. Also, the relatively long initiation latencies in this experiment (~1000 ms) undermine the possibility that only the definite determiner was planned prior to speech onset. (Analyses of determiner durations in Experiment 2 further undermine this possibility.)

Second, different visual displays elicited adjective-noun phrases and relative clauses. But, to produce our finding, the display eliciting adjective-noun phrases should have been harder to process, thus slowing down the initiation latencies for these utterances and making them similar to the ones of relative clauses. However, the displays eliciting adjective-noun phrases should be, if anything, easier to process, because the more salient entity on these displays (the color patch) is also the one to be mentioned first.

Of most relevance here, participants sped up to a similar extent initiating simpler and more complex descriptions under load, indicating that load did not disproportionally affect the production of relative clauses relative to that of adjective-noun phrases. However, participants took less time to *utter* relative clauses (but not adjective-noun phrases) with load than with no load – in other words, load affected the durations of more complex but not of simpler descriptions. We are hesitant, however, to take this as support for the memory-heavy account. This is because the taxing of resources in a limited-capacity system should disrupt performance, not facilitate it, but we did not find any disruption for either initiation latencies, description durations or production errors in Experiment 1. Instead, to avoid forgetting the two load words, participants might have sped up uttering specifically relative-clause descriptions because they are longer and thus provide more opportunity to speed up articulation. (We report analyses of length-corrected durations after Experiment 4.) The un-predicted speed-up of production under load, outside of the study's initial goals, seems interesting in itself, but at this point it is unclear whether it was a function of the specific features of Experiment 1. Experiments 2-4 contain manipulations that test for this possibility, and we return to the topic of production flexibility in the General Discussion.

Meanwhile, it is fair to note that our load manipulation (keeping two words in memory) might not have been taxing enough to produce effects on syntactic formulation during the production task. This is because participants sped up their descriptions instead of slowing them down under load, and recalled correctly a high percentage of load words (94.25%). To address the possibility that our task was not taxing enough, we increased the memory load in Experiment 2 from two to four words. Such a load should be comparable to the agreement-disrupting 5-word load in Fayol et al. (1994) and 7-word load in Hupet et al. (1998), because their words were monosyllabic and ours were mostly multisyllabic.

Experiment 2

In this experiment, participants named pictures with adjective-noun phrases (simple descriptions, e.g., "*the red book*") and relative clauses (complex descriptions, e.g., "*the book that is red*") while keeping in working memory a four-word verbal load (e.g., *collection mountain octopus razor*).

Method

Participants. Forty-eight undergraduates from the same population as Experiment 1 participated for course credit. All were native speakers of English. Seven further participants were excluded: five because of consistently producing utterances that were different from the ones they were instructed to produce, and two because of experimenter error.

Materials and procedure. These were the same as in Experiment 1, except that the memory load was increased from two to four words. To this aim, the memory word pairs from Experiment 1 were kept the same, and two new words with similar characteristics were added to each pair to form quadruplets (e.g., "collection mountain octopus razor").

Data analysis. In addition to the analyses performed for Experiment 1, we conducted three further analyses to aid the interpretation of our findings. First, we analyzed the durations of the utterance-beginning determiners (*the* in e.g. the *car that is red* or the *red car*), to probe for effects of complexity and load undetected by the latencies and durations analyses. Second, we analyzed recall performance as a continuous variable as a function of the length of the preceding picture description while controlling for syntactic complexity, to see if recall performance in this experiment supported the memory-heavy account or was due to the distance between encoding and recall. Third, we analyzed latencies and durations on only partial-recall trials (on which participants recalled no, one or two words correctly out of four) to see if a possibly heavier memory burden on these trials would influence the pattern of results. For the reader's processing ease, the statistical models used in these additional analyses are described together with their results.

Results

Data exclusions. The experiment contained 2304 trials. There were 588 incorrect recall trials (trials on which fewer than all four words were recalled correctly; these were 25.52% of all trials, or 51.04% of load trials); they were excluded from the analyses reported here. Analyses keeping trials on which at least one word was recalled, leading to the exclusion of only 33 trials (2.86% of load trials) produced an identical pattern of results. Production errors were made on 86 of the remaining trials and were excluded from the analyses of description latencies and durations. Fifteen further trials were discarded because of experimenter error (13) or failure of a participant to follow the instructions (2). Another 12 trials were removed as outliers from the latency data, and another 10 from the duration data. Of the 27 unanalyzed-latency trials, 12 involved simple descriptions and 15 complex descriptions; of the 25 unanalyzed-durations trials, 11 involved simple descriptions and 14 complex descriptions. In total, 701 trials (30.43%) were discarded from the latency data, and 699 trials (30.34%), from the duration data.

Analyses of description latencies, description durations, and production errors. Descriptive statistics are presented in Figure 4, and inferential statistics are summarized in

Table 3. The analyses of initiation latencies (Panel 4A) revealed no significant effects. Participants took a similar amount of time to initiate complex and simple descriptions, as well as descriptions under load and under no load, and load did not differentially affect simple and complex descriptions.

The analyses of utterance durations (Panel 4B) revealed that participants took longer to utter complex descriptions than simple descriptions (complexity was a significant predictor). Further, participants uttered their descriptions more quickly under load than under no load (load was a significant predictor). Unlike Experiment 1, load did not differentially affect the duration of complex and simple descriptions (the complexity x load interaction was not significant).

The analyses of production-error rates (Panel 4C) revealed that participants made a similar number of errors on simple and on complex descriptions, as well as with or without memory load (complexity and load were not significant predictors). However, load differentially affected simple and complex descriptions: Load increased the production errors for complex, but not for simple descriptions (in this analysis, the complexity x load interaction was significant, and the simple effect of load was significant for complex, but not for simple descriptions; see Table 3).

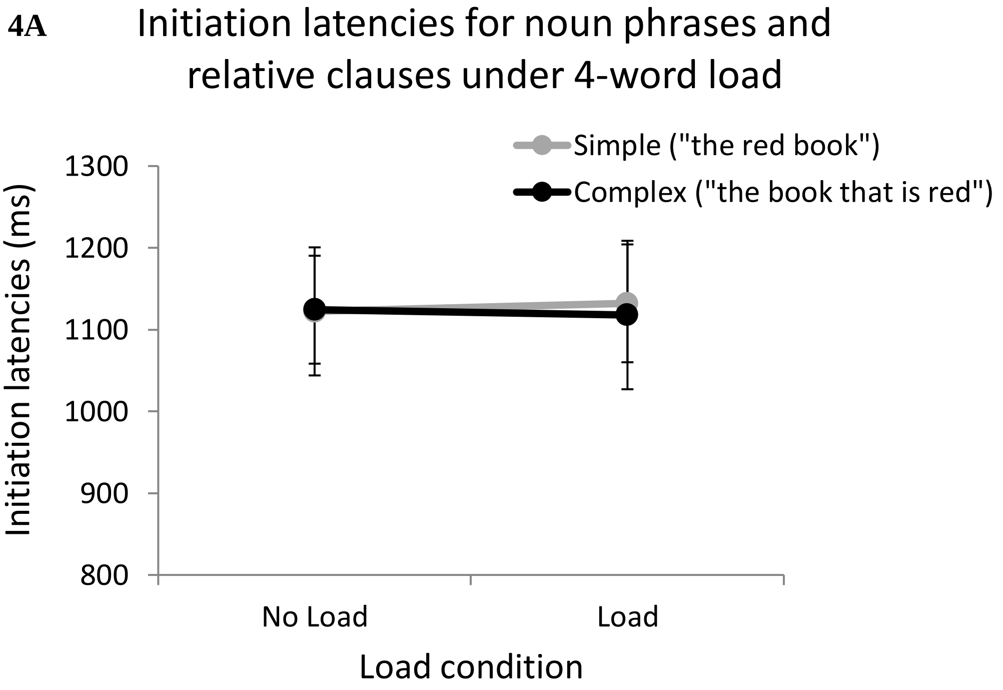
Determiner durations. The absence of effects of complexity and load on initiation latencies in this experiment (and the speed-up of latencies under load in Experiment 1) might stem from radically incremental processing, such that participants initiated their descriptions as quickly as possible but then lengthened the duration of their descriptions' first word (always the determiner *the*). To address this possibility, we extracted and analyzed the durations of the phrase-initiating determiners. The fixed predictors in the model were load, complexity and their interaction (same as for the main analyses). We found that participants took longer to utter determiners in simple than in complex descriptions (complexity was a significant predictor [*Estimate* = -16.13, *SE* = 5.44, *z* = -2.97, *p* = .005]), but the effect of load and the complexity x load interaction were not significant [both *p*s > .5]. This pattern undermines the possibility of radically incremental processing and hints instead at strategic processing: Durations were sped up under load to minimize the separation between encoding and recall of the load words, and this happened more for the longer relative clauses than for the shorter adjective-noun phrases.

Recall performance. Recall performance was low (49.81%), and significantly lower than in Experiment 1 [*Estimate* = .23, *SE* = .02, *z* = 11.86, *p* < .001], confirming that the load task in this experiment was sufficiently taxing. Participants correctly recalled fewer load words after producing complex descriptions (236, or 45.56% of the valid load trials in this condition) than after producing simple descriptions (284, or 53.99% of the valid load trials; syntactic complexity was a significant predictor [*Estimate* = .45, *SE* = .18, *z* = 2.55, *p* = .01]).

The worse recall after producing more complex than after producing simpler descriptions might indicate a trade-off in working memory demands between the production and recall tasks, and thus support the memory-heavy account. However, our more complex descriptions were also longer, and we suspected that recall was worse after these descriptions primarily because of the longer separation between encoding and recall than for simpler descriptions. If so, description length should predict recall performance after controlling for syntactic complexity. To check for this possibility, we regressed length in phonemes for the descriptions in Experiment 2 onto syntactic complexity (numerically coded as -0.5 and 0.5). We then used the residuals of this model – a measure of the length of speakers’ utterances with complexity regressed out – as a fixed predictor in a model analyzing recall performance as a continuous dependent variable (0, 1, 2, 3 or 4 words correctly recalled). The effect of this residual (complexity-corrected) length was marginally significant [*Estimate* = -.10, *SE* = .05, *z* = -2.00, *p* = .051]. This result suggests that longer descriptions had adverse effects on recall even after correcting for syntactic complexity, and thus undermines the possibility that recall performance in this experiment was specifically affected by syntactic complexity.

To see if recall success affected performance, we analyzed latencies and durations for only those trials on which participants recalled none, one or two words (out of four). Differently from the main analyses, load slowed latencies [main effect of load: *Estimate* = 111.92, *SE* = 37.29, *z* = 3.00, *p* = .005] and left durations unaffected [*Estimate* = -7.12, *SE* = 15.53, *z* = -.46, *p* = .65]. Still, there was no complexity x load interaction for either measure [both *p*s > .59]. These effects may suggest that working memory was most taxed on no-recall or partial-recall trials (resulting in adverse effects of load) but it may also be that these trials involved processes in some ways different from the ones we assume operate on correct-recall trials. In any case, most important was that an interaction between complexity and load was absent even in the presence (for latencies) of the predicted adverse effect of load.

4A



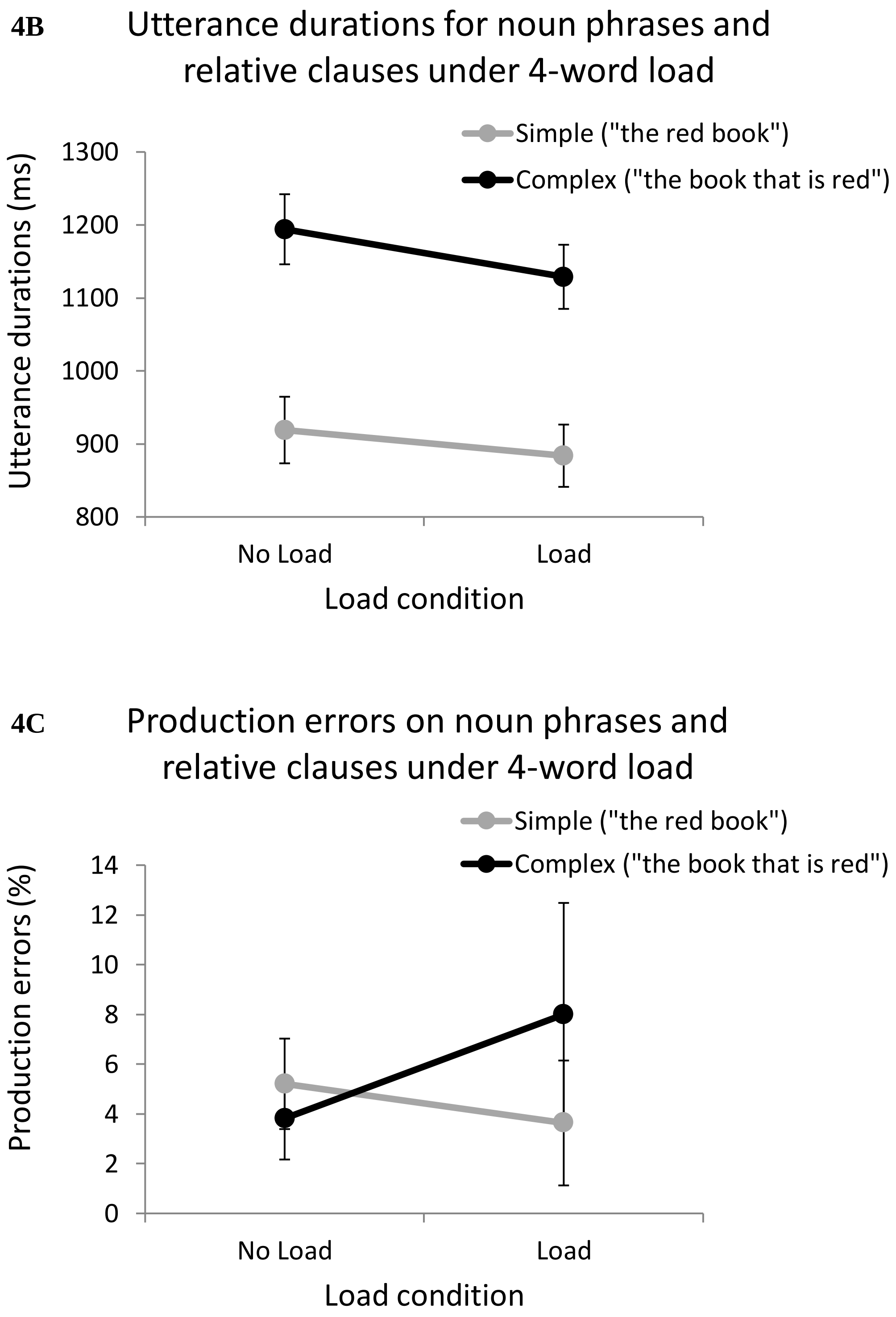


Figure 4. Initiation latencies, utterance durations and production errors in Experiment 2. Panel 4A: initiation latencies; Panel 4B: utterance durations; Panel 4C: production errors. Error bars represent 95% confidence intervals.

Table 3

LMER analyses of initiation latencies, utterance durations and production errors in Experiment 2

Model	Predictors	Estimate	SE	*t or z*	*p*
Initiation latencies					
	Complexity	-10.01	20.35	-.49	.63
	Load	9.97	22.32	.45	.66
	Complexity * Load	-40.21	30.71	-1.31	.20
Utterance durations[a]					
	Complexity	269.89	11.84	22.80	< .001
	Load	-44.10	11.09	-3.98	< .001
	Complexity * Load	-20.84	16.15	1.29	.20
Production errors					
Main model[b]	Complexity	.16	.24	.68	.50
	Load	-.02	.39	-.05	.96
	Complexity * Load	1.51	.50	3.02	.003
Simple effects of Load[c]	Simple descriptions	.43	.38	1.13	.26
	Complex descriptions	-1.07	.32	-3.32	< .001

Note: Shaded rows highlight significant effects.

[a] This model did not have a by-item random slope for load.

[b] This model had the following random-effects structure:

err ~ load * complexity + (1 + load | subj) + (1 | item).

[c] This model had only by-subject and by-item random intercepts.

Discussion

The aim of Experiment 2 was to increase the amount of memory load participants kept during the production task, to make sure load maintenance detracted enough resources from it to observe an influence of load on this task. The poor recall in this experiment (49.81%), as well as the longer production latencies (around 1100 ms, relative to around 1000 ms in Experiment 1) suggest that we succeeded in this intention.

This increased difficulty affected description latencies relative to Experiment 1: they remained unchanged under load (instead of decreasing), suggesting that the increased load did have an effect on utterance initiation. This might have been the case because encoding the four load words failed to complete until after picture onset, or that the load words interfered with the picture descriptions, making it impossible to complete the (minimal necessary) planning prior to speech initiation. In both cases, a comparison between Experiment 1 and Experiment 2 (see below) suggests that a heavier load did not leave production unaffected. Despite this, it seems that the resource capacity of the production system was still not depleted, in that description latencies did not *increase* under load. This result thus further evidences the flexibility of the production system.

Production flexibility was further supported by an additional analysis of description-initiating determiner durations, which were shorter for more complex than for simpler descriptions (presumably to further speed up uttering the longer relative clauses and avoid forgetting the load words) but were unaffected by load. This analysis undermines the possibility that participants initiated speech after planning production of only the determiners (i.e., radically incremental processing, see e.g. Brown-Schmidt & Konopka, 2008, 2015), and the predicted complexity and load effects leaked into the production of the determiners.

As in Experiment 1, we found that participants took a similar amount of time to initiate relative clauses and adjective-noun phrases, again suggesting that relative clauses (unlike adjective-noun phrases) were not fully planned before speech onset.

In this experiment, participants uttered both simpler and more complex descriptions more quickly under load than under no load (while they sped up uttering only more complex utterances under load in Experiment 1), again, presumably in an attempt to avoid forgetting the load words. In other words, participants could still optimize their production to the task demands even under the considerably increased load.

Of most relevance here, neither initiation latencies nor utterance durations for more complex descriptions were disproportionally affected by load relative to simpler descriptions. This was so even in the presence of the predicted adverse effects of load, shown by latency analyses on partial-recall trials. These results suggest little or no involvement of working memory in syntactic formulation in our experiments, in support of the memory-light account.

However, participants made more production errors on relative-clause descriptions with load than with no load, but a similar number of production errors on adjective-noun descriptions with load and no load. This first piece of evidence so far for the memory-heavy account might suggest a role of working memory in pre-articulation monitoring (cf. accounts in Fayol et al., 1994; Hartsuiker & Barkhuysen, 2006; Hupet et al., 1998). This is because the presumed role of the internal monitor is to avoid production errors; an increase of such errors under load might indicate that the monitor's optimal functioning was compromised when working memory resources were taxed. To investigate what type of monitoring might have been affected (of syntactic or lexical content), we report analyses of different error types after Experiment 4, and we return to this point in the General Discussion.

In Experiment 2, participants recalled fewer load words after producing relative clauses than after producing adjective-noun phrases. This result might evidence complexity effects, in that recall was disrupted to a greater extent after producing the more complex relative to simpler utterances. It thus seems consistent with the memory-heavy account in that performance on the production task disrupted performance on the (here very demanding) recall task, suggesting shared resources. However, additional analyses suggested that this effect was instead likely attributable to the greater length of relative clauses, leading to longer separation between encoding and recall.

In Experiment 2, we found that heavier load did not slow down either description initiation or duration, and the presence of load did not interact with description complexity for either of these measures. Before drawing conclusions, we aimed to address another possible explanation for these results, related to a conceptualization of working memory as a set of several different cognitive mechanisms instead of as a unitary construct. Shipstead, Lindsey, Marshall, and Engle (2014) discuss three such mechanisms: primary memory, secondary memory, and attention control, and present evidence that each of these mechanisms contributes to explaining individual differences in what is collectively known as working memory capacity (see also Badecker & Kuminiak, 2007; Cowan, 1999; McElree, 2001, for similar distinctions). In such a conceptualization of working memory, primary memory would reflect a limited capacity storage of 3-5 items (corresponding to the size of an individual's focus of attention, e.g., Cowan, Elliott, Saults, Morey, Mattox, Hismjatullina et al., 2005) whose function is to protect the stored units of information from proactive interference and allow the formation of new connections between them (Cowan, 2001; Oberauer, Süß, Wilhem, & Sander, 2007). Secondary memory would correspond to a portion of long-term memory where relevant information too large for primary memory is displaced and, at least for some individuals, remains less prone to interference and more searchable with appropriate cues (Unsworth & Engle, 2007; see also Wixted & Rohrer, 1994). Attention control is the mechanism for selecting goal-relevant information and avoiding distractions in the presence of conflicting information or prepotent responses (Engle, 2002).

Adopting such a multi-mechanism conceptualization of working memory raises the possibility that processing of the memory load words and picture-description syntactic structure might not have overlapped in a single working memory component in our experiments. This is because, in both Experiments 1 and 2 (but especially in Experiment 1 where the load was smaller), load-encoding time (5500 ms) might have been long enough for encoding and displacement to secondary memory. As a result, the load in our experiments might not have taxed the attended portion of working memory at exactly the same time as it was needed for production. To address this possibility, load encoding time was reduced to 1500 ms in Experiment 3 (which was otherwise identical to Experiment 1). In doing so, we assumed that load encoding processes were more likely to at least partially overlap with utterance planning and production.

Note that we will not be able to fully discard the possibility that maintenance of information in the attended portion of working memory is fully serial across tasks; in our case, that the information necessary to encode the load words and to plan and produce an utterance cannot be kept in the attended portion of working memory at the same time. However, we think that such a possibility goes against most working memory dual task research, which has found effects of keeping information in working memory on subsequent processing in another task (most relevantly here, Belke, 2008; Hartsuiker and Barkhuysen, 2006; Martin et al., 2014; Wagner et al, 2010).

Experiment 3

In this experiment, participants named pictures with noun phrases (simple descriptions, e.g., "*the red book*") and relative clauses (complex descriptions, e.g., "*the book that is red*") while keeping in working memory a two-word verbal load (e.g., *collection mountain*). To encode the load words, participants were given only 1500 ms (and not 5500 ms as in Experiments 1 and 2).

Method

Participants. Forty-eight undergraduates from the same population as Experiment 1 participated for course credit. All were native speakers of English.

Materials, procedure and data analysis. These were the same as in Experiment 1, with the following exceptions. The asterisk beginning each trial was presented for 400 ms, the two memory words were then presented for 600 ms, followed by a mask (&!=#%+?#&!=#%+?) for 500 ms (to avoid encoding the words' visual characteristics), and a blank screen followed for 400 ms. Participants were warned that the two words would disappear from the screen quickly and were asked to pay attention. In this experiment, they were given no instructions on how to go about memorizing the words.

As in Experiment 2, we also analyzed latencies and durations on imperfect recall trials (none or one out of two words).

Results

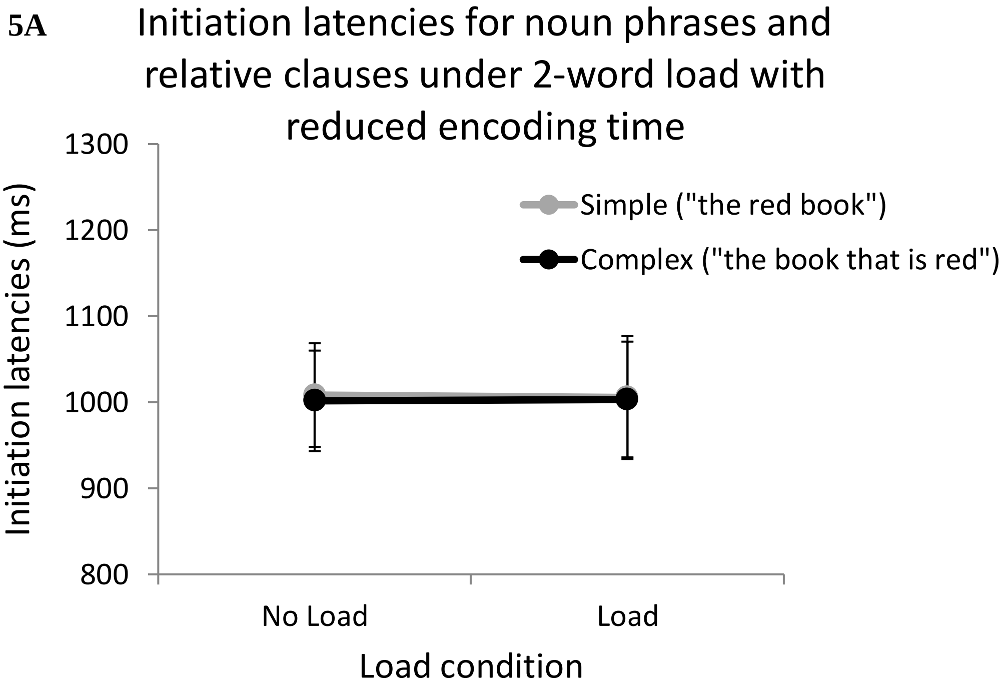
Data exclusions. The experiment contained 2304 trials. There were 146 incorrect recall trials (trials on which fewer than all two words were recalled (correctly); 6.34% of all trials, or 12.67% of load trials). These trials were excluded from the analyses reported here; analyses keeping trials on which at least one word was recalled, leading to the exclusion of 61 trials (5.30% of load trials), produced an identical pattern of results. Production errors were made on 103 of the remaining trials, and were excluded from the analyses of response latencies and durations. Eleven further trials were discarded because of experimenter error. From the latency data, another 30 trials were removed because of a coding error, and 63 as outliers. Of the 104 unanalyzed-latency trials, 53 involved simple descriptions and 51 complex descriptions. From the duration data, another 26 trials were removed because of a coding error, and 62 as outliers. Of the 99 unanalyzed-durations trials, 65 involved simple descriptions and 34 complex descriptions. In total, 353 trials (15.32%) were discarded from the latency data, and 348 trials (15.10%), from the duration data.

Analyses of description latencies, description durations, and production errors. Descriptive statistics are presented in Figure 5, and inferential statistics are summarized in Table 4. As in Experiment 2, the analyses of initiation latencies (Panel 5A) revealed no significant effects. Participants took a similar amount of time to initiate complex and simple descriptions, as well as descriptions under load and under no load, and load did not differentially affect simple and complex descriptions.

Also as in Experiment 2, the analyses of utterance durations (Panel 5B) revealed that participants took longer to utter complex descriptions than simple descriptions (complexity was a significant predictor), uttered their descriptions more quickly under load than under no load (load was a significant predictor), and load did not differentially affect the duration of complex and simple descriptions (the complexity x load interaction was not significant).

The analyses of production-error rates (Panel 5C) revealed that participants made more errors on no load than on load trials (load was a significant predictor). No other predictors were significant.

5A



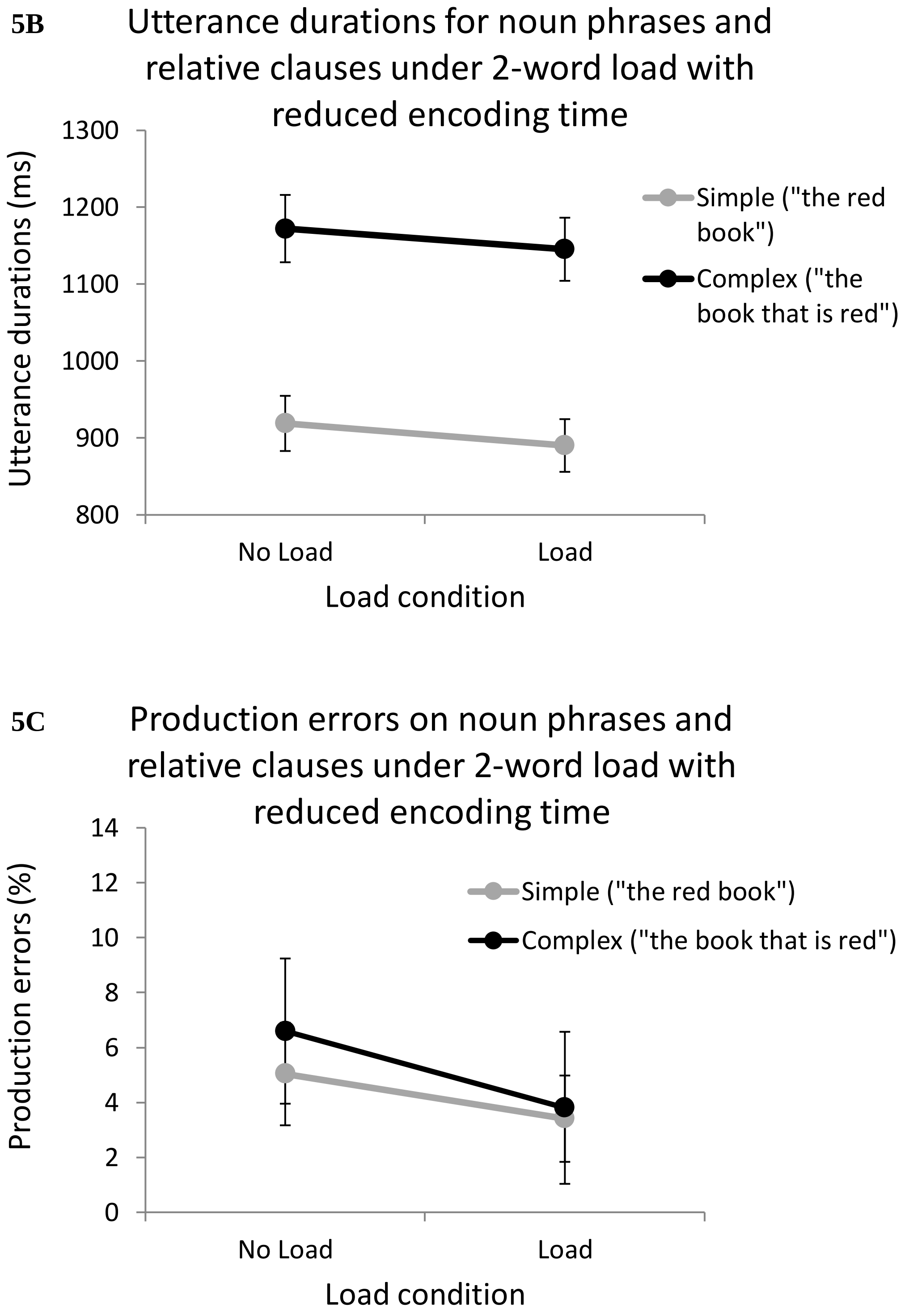


Figure 5. Initiation latencies, utterance durations and production errors in Experiment 3. Panel 5A: initiation latencies; Panel 5B: utterance durations; Panel 5C: production errors. Error bars represent 95% confidence intervals.

Recall performance. Recall performance (88.57%) was significantly lower than in Experiment 1 [*Estimate* = .03, *SE* = .01, *z* = 3.79, *p* < .001]. Participants correctly recalled a similar number of memory load words after producing simple (474, or 86.97% of the valid load trials in this condition) and complex descriptions (487, or 90.19% of the valid load trials), that is, syntactic complexity was not a significant predictor [*Estimate* = .23, *SE* = .31, *z* = .75, *p* = .45].

Analyses of latencies and durations for imperfect-recall trials (on which participants recalled none or one word, out of two) showed that, differently from the main analyses, load slowed down both latencies, as in Experiment 2 [main effect of load: *Estimate* = 94.65, *SE* = 41.80, *z* = 2.26, *p* = .03] and durations, unlike Experiment 2 [*Estimate* = 63.34, *SE* = 30.39, *z* = 2.09, *p* = .04]. Most importantly, and as in Experiment 2, there was no complexity x load interaction for either measure [both *p*s > .60].

Table 4

LMER analyses of initiation latencies, utterance durations and production errors in Experiment 3

Model	Predictors	Estimate	SE	*t or z*	*p*
Initiation latencies					
	Complexity	-1.11	17.98	-.06	.95
	Load	7.36	18.91	.39	.70
	Complexity * Load	7.53	25.31	.30	.77
Utterance durations					
	Complexity	259.78	11.29	23.02	< .001
	Load	-23.88	9.53	-2.51	.02
	Complexity * Load	2.40	14.28	.17	.87
Production errors[a]					
	Complexity	.18	.22	.81	.42
	Load	-.50	.22	-2.30	.02
	Complexity * Load	-.22	.44	-.50	.62

Note: Shaded rows highlight significant effects.

[a] This model had only by-subject and by-item random intercepts.

Discussion

In Experiment 3, we reduced the load encoding time, to ensure that load encoding and maintenance taxed the same resources as those necessary for syntactic formulation during production. Despite this manipulation, results were very similar to those of Experiment 2: Participants initiated simpler and more complex descriptions with similar speed under load and no load, and description durations decreased under load. Thus, the reduction of encoding time for the two load words produced a similar effect on production to increasing the load to four words in Experiment 2 (although recall accuracy was overall much higher in this experiment (88.57%) than in Experiment 2 (49.81%)).

In Experiment 3, participants made fewer production errors overall with load than with no load. Of most relevance here, neither initiation latencies nor utterance durations nor production errors for more complex descriptions were disproportionally affected by load relative to simpler descriptions. These findings provide further support for the memory-light account.

Experiment 4 was the strongest test of the memory-heavy versus memory-light accounts. The (accountable) absence of detectable complexity effects in the latencies and durations in Experiments 1-3 (although not in production errors) might have obscured any effects of load. Experiment 4 introduced lexical complexity in addition to syntactic complexity, because it has consistently produced complexity effects in numerous prior studies (e.g., Smith & Wheeldon, 1999; Martin et al., 2010; 2014), likely due to the need to retrieve two lexical items instead of one. If the latency and duration effects we found in Experiments 1-3 (supporting the memory-light account) were due to the absence of detectable complexity effects, the presence of such effects (driven by any kind of complexity) should lead to also detecting the predicted adverse effects of load. Specifically, in Experiment 4 production should slow down, rather than speed up, under load (as in Martin et al., 2014), and load should disproportionally affect the production of more complex relative to simpler utterances. Since load did not disproportionally disrupt initiation latencies for sentences beginning with complex relative to simple noun phrases in Martin et al. (2004), we would expect such effects in Experiment 4 to manifest more strongly in the description durations. Conversely, a replication of the load effects from our Experiments 1-3 would strongly support the memory-light account.

Experiment 4 thus involved the production of phrases which were both lexically and syntactically more complex (e.g., *"the book and the ball"*) than the simpler phrases (e.g., *"the car"*). Complex noun phrases are lexically more complex because of the need to retrieve two lexical items instead of one, and syntactically more complex because of the need to plan more structure and compute a plural referent instead of a simpler structure and a singular referent.

Experiment 4

In this experiment, participants named pictures with simple noun phrases (e.g., "*the book*") and complex noun phrases (e.g., "*the book and the car*") while keeping in working memory a two-word verbal load (e.g., *collection mountain*). Load encoding time was 5500 ms, as in Experiments 1 and 2.

Method

Participants. Forty-eight undergraduates from the same population as Experiment 1 participated for course credit. All were native speakers of English. One further participant was excluded because of consistently producing utterances which were different from the target ones.

Materials, procedure and data analysis. These were the same as in Experiment 1, except that there were no color patches. Instead, on simple-description trials, participants saw a single picture, positioned at the center of the screen, which they were instructed to name with a definite noun phrase (e.g., *the book*); on complex-description trials, they saw two pictures next to each other, which they were instructed to name with a definite coordinated noun phrase (e.g., *the book and the car*). The 24 additional pictures were selected from the IPNP database and had similar characteristics to the original 48 pictures (half had high-frequency names, and half had low-frequency names, $p < .001$). In the complex-syntax condition, the pictures in a pair had names of the same frequency type, forming either high-frequency or low-frequency pairs. Before the beginning of the experiment proper, participants were familiarized with all the pictures to appear during the experiment.

Results

Data exclusions. The experiment contained 2304 trials. There were 58 incorrect recall trials (trials on which fewer than all two words were recalled (correctly); 2.52% of all trials, or 5.03% of load trials). These trials were excluded from the analyses reported here; analyses keeping trials on which at least one word was recalled, leading to the exclusion of 33 trials (1.74% of load trials) produced an identical pattern of results. Production errors were made on 100 of the remaining trials, and these were excluded from the analyses of response latencies and durations. Seventeen further trials were discarded because of experimenter error. Another 26 trials were removed as outliers from the latency data, and another 5 from the duration data. Of the 43 unanalyzed-latency trials, 15 involved simple descriptions and 28 complex descriptions; of the 22 unanalyzed-durations trials, 7 involved simple descriptions and 15 complex descriptions. In total, 201 trials (8.72%) were discarded from the latency data, and 180 trials (7.81%) from the duration data.

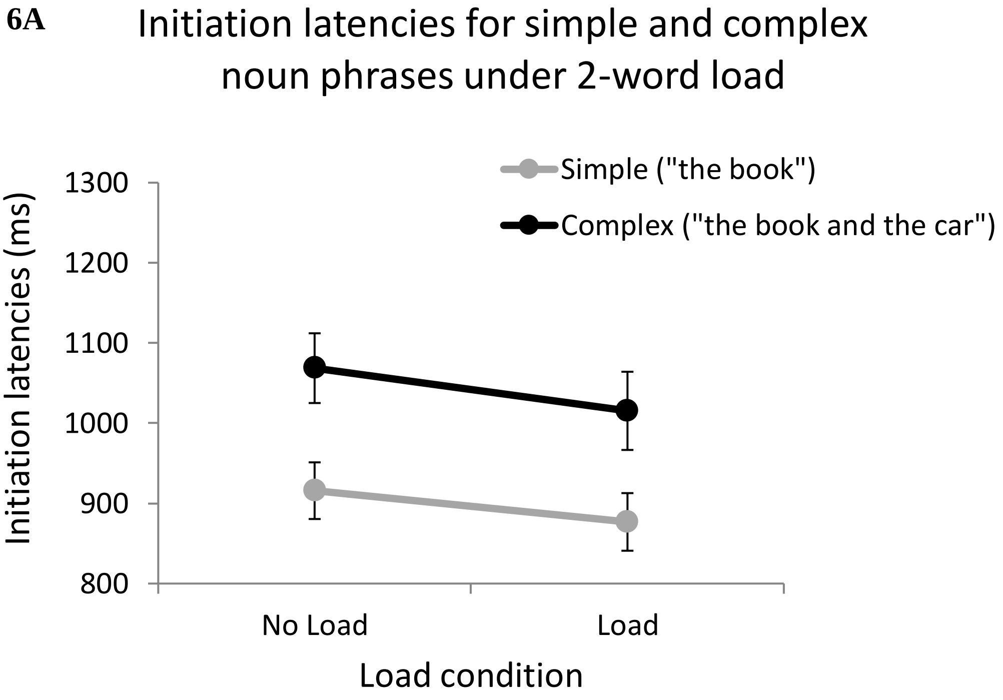
Analyses of description latencies, description durations, and production errors. Descriptive statistics are presented in Figure 6, and inferential statistics are summarized in Table 5. The results of all four experiments are presented in Table 6, for easy comparison. The analyses of initiation latencies (Panel 6A) revealed that participants took longer to initiate complex than simple descriptions (complexity was a significant predictor). Also, participants were faster to initiate both types of description under load than under no load (load was a significant predictor). However, load did not differentially affect simple and complex descriptions (the complexity x load interaction was not significant).

The analyses of description durations (Panel 6B) revealed that participants took longer to utter complex descriptions than simple descriptions (complexity was a significant predictor). There were no other significant effects: Participants took no longer to utter their descriptions under load than under no load (load was not a significant predictor), and load did not differentially affect simple and complex descriptions (the complexity x load interaction was not significant). Analyses with durations regressed for length produced an identical pattern of results (see the Additional Analyses section following this experiment for an explanation of how these analyses were performed).

The analyses of production-error rates (plotted in Panel 6C) suggested that participants made a similar number of errors on simple as on complex descriptions, but more errors with load than with no load (load was a significant predictor; but see note under Table 5). There was also an indication that load differentially affected complex and simple descriptions: there were more errors on complex descriptions under load than under no load, but a similar number of errors on simple descriptions under load and no load (the complexity x load interaction in the main model was not significant, but the simple effect of load was significant for complex descriptions and not for simple descriptions).

Recall performance. Recall performance was very high overall (95.70%) and did not differ from Experiment 1 [*Estimate* = -.004, *SE* = .01, *z* = -.57, *p* = .57]. Participants correctly recalled fewer load words after producing complex descriptions (486, or 93.64% of the valid load trials in this condition) than after producing simple descriptions (538, or 97.64%): In this analysis, syntactic complexity was a significant predictor [*Estimate* = 1.40, *SE* = .70, *z* = 2.00, *p* = .05]; this model did not have a by-item random intercept. Here, analyses on how complexity-corrected description length influenced recall were not performed because of the high recall performance and few recall options (three total: 0, 1 and 2 words correctly recalled; cf. Experiment 2).

6A



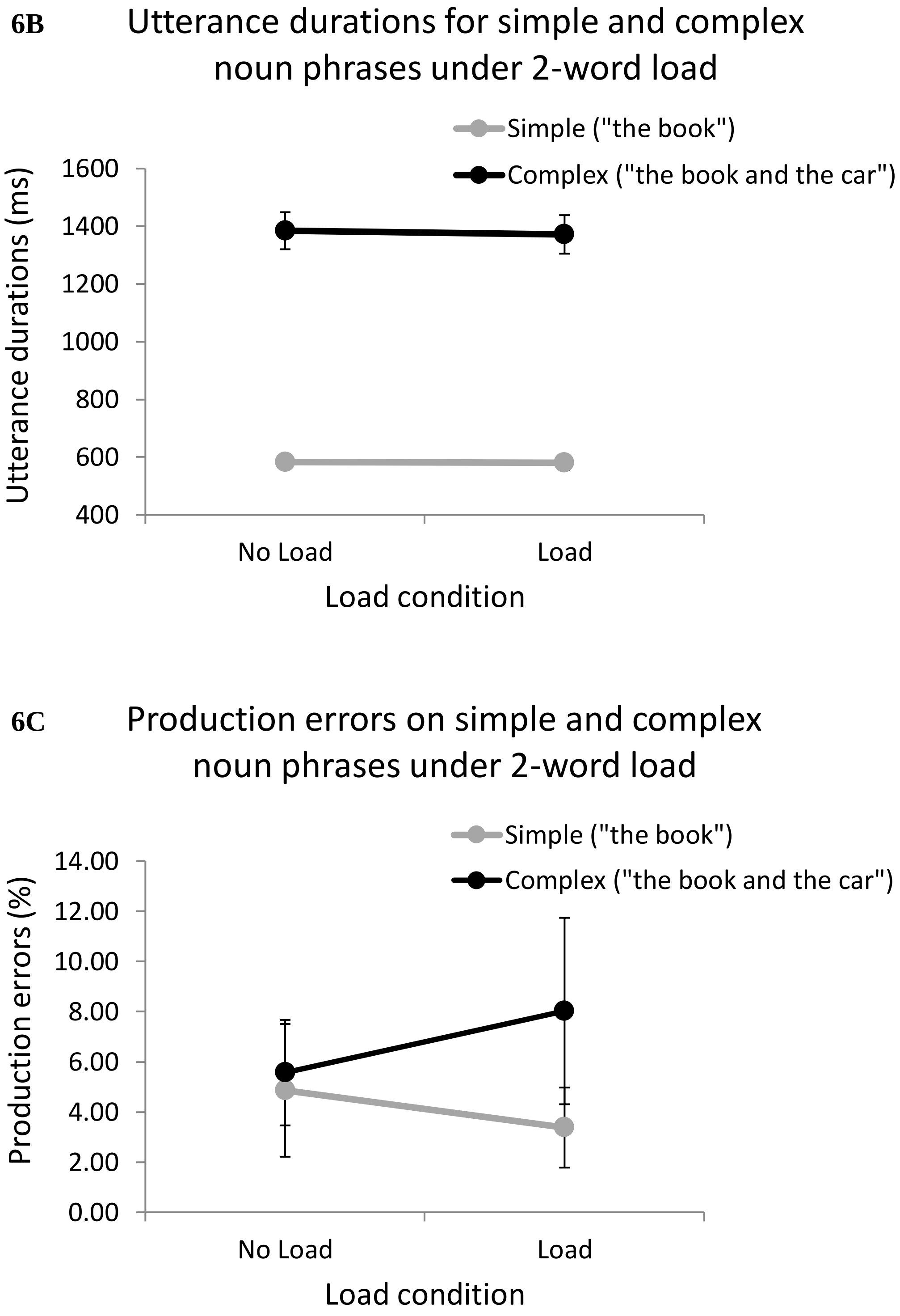


Figure 6. Initiation latencies, utterance durations and production errors in Experiment 4. Panel 6A: initiation latencies; Panel 6B: utterance durations (note that this panel is on a different scale than all other figures in this manuscript); Panel 6C: production errors. Error bars represent 95% confidence intervals.

Table 5

LMER analyses of initiation latencies, utterance durations and production errors in Experiment 4

Model	Predictors	Estimate	SE	*z*	*p*
Initiation latencies					
	Complexity	142.11	12.81	11.09	< .001
	Load	-45.40	12.61	-3.60	< .001
	Complexity * Load	-14.83	21.99	-.67	.50
Utterance durations					
	Complexity	793.93	28.33	28.03	< .001
	Load	-10.66	7.96	-1.34	.19
	Complexity * Load	-12.61	16.27	-.76	.44
Production errors					
Main model[a,b]	Complexity	.64	.41	1.55	.12
	Load	.56	.28	2.04	.04
	Complexity * Load	.11	.47	.22	.82
Simple effects of Load[c]	Simple descriptions	-.50	.39	-1.29	.20
	Complex descriptions	.65	.28	-2.30	.02

Note: Shaded rows highlight significant effects.

[a] This model had the following random-effects structure:

err ~ load * complexity + (1 + complexity | subj) + (0 + load | item).

[b] In the main model analyzing error rate, the load predictor was significant, but the complexity x load interaction was not. However, a visual inspection of Figure 6C suggests that load differentially affected the errors made on simple and complex descriptions. We suspected that the statistical model might have attributed to the load predictor some variance accounted for by the interaction predictor (the correlation between the two was r_2 = .29), and we also analyzed the simple effects of load.

[c] This model had only by-subject and by-item random intercepts and a by-subject random slope for complexity.

Table 6

Initiation latencies, utterance durations and production errors in all experiments

		Latencies (ms.)			Durations (ms.)			Errors (%)		
		Simple	Compl.	*Diff.*	Simple	Compl.	*Diff.*	Simple	Compl.	*Diff.*
Exp. 1	No load	1036	1016	*-20*	870	1139	*269*	6.60	4.17	*-2.43*
	Load	1000	977	*-23*	875	1118	*243*	4.84	4.63	*-.21*
	Diff.	*-36*	*-39*		*5*	*-21*		*-1.76*	*.46*	
Exp. 2	No load	1122	1124	*2*	919	1194	*275*	5.21	3.82	*-1.39*
	Load	1132	1118	*-14*	884	1129	*245*	3.64	8.00	*4.36*
	Diff.	*10*	*-7*		*-35*	*-65*		*-1.57*	*4.18*	
Exp. 3	No load	1008	1002	*-6*	919	1172	*253*	5.03	6.60	*1.57*
	Load	1006	1003	*-3*	890	1145	*255*	3.41	3.80	*.40*
	Diff.	*-2*	*1*		*-29*	*-27*		*-1.63*	*-2.79*	
Exp. 4	No load	916	1069	*153*	583	1385	*802*	4.86	5.57	*.71*
	Load	877	1015	*139*	581	1372	*791*	3.38	8.02	*4.65*
	Diff.	*-39*	*-54*		*-2*	*-3*		*-1.48*	*2.46*	

Discussion

In replication of prior studies, participants initiated complex noun phrases more slowly than simple noun phrases (Smith & Wheeldon, 1999; Martin et al., 2010; 2014), likely because of the additional time needed to retrieve two lexical items instead of one before speech onset. We acknowledge that our design conflated phrase complexity with visual complexity (one versus two pictures for simple versus complex noun phrases). However, our complexity effect (153 ms with no load) was comparable to the complexity effect of *The dog and the kite move up* relative to *The dog moves up* in experiments without such a confound (195 ms and 164 ms respectively in Experiments 2 and 3, Smith & Wheeldon, 1999).

Still, in replication of Experiment 1, participants initiated their descriptions more quickly under load than under no load. This result differs from the study of Martin et al. (2014), who found that latencies were longer, not shorter, under load. The reason for this discrepancy could be that longer utterances such as the sentences in Martin et al. require more advance planning than shorter utterances such as the single phrases we used (see Holmes, 1988, for related evidence). If so, there might be a limit to the flexibility of utterance planning under load. (But note that Power's (1985) participants, who also sped up utterance initiation under load, also produced sentences instead of phrases.)

Unlike Experiments 1-3, there were no effects of load on the description durations. It might be that some utterance elements give more room for faster planning and articulation, and that more such elements are present in adjective-noun phrases and relative clauses than in complex noun phrases. Specifically, color adjectives in Experiments 1-3 (present in both adjective-noun phrases and relative clauses) might be planned and uttered more quickly than nouns because they repeated eight times throughout each experiment, while nouns appeared only once. In contrast, the complex noun phrases in Experiment 4 contained only nouns but not adjectives. Further, other elements that might enable speeding up articulation are longer in relative clauses (the relative pronoun *that* and verb *is*) than in complex noun phrases (the conjunction *and*).

In Experiment 4 participants made more production errors on complex phrases under load than under no load, but a similar number of production errors on simple phrases under load and no load. This result replicates the production-error pattern in Experiment 2 and indicates a possible role of working memory in pre-articulation monitoring processes (Fayol et al., 1994; Hartsuiker & Barkhuysen, 2006; Hupet et al., 1998).

Contrary to our predictions, utterance durations for complex noun phrases were not disproportionally disrupted by load in this experiment (and neither were initiation latencies, cf. Martin et al., 2014). These patterns provide strong support for the memory-light account. They also suggest that memory load of the type tested here (i.e., involving active maintenance but no manipulation of information) does not seem to affect the process of lexical selection in language production.

In Experiment 4, participants recalled fewer load words after producing complex than after producing simple noun phrases. As in Experiment 2, this effect might be due to the greater length of complex noun phrases relative to simple ones, but also to interference between the two nouns in the picture description and the two load nouns (although they were unrelated in meaning).

Additional Analyses

Can lack of power explain the lack of evidence for working memory involvement in syntactic formulation?

At the outset, we predicted that the memory-heavy account would be supported by disproportional slowing and disruption under load of relative clauses relative to adjective-noun phrases, or a statistical interaction between load and complexity. Across three measures in each of three experiments, we found such an interaction in only two instances. In Experiment 1, the durations for complex descriptions decreased under load while the durations for simple descriptions remained unchanged. In Experiment 2, production errors on complex descriptions increased under load, while production errors on simple descriptions remained unchanged.

While the direction of the effect of load on durations is inconsistent with the memory-heavy account, we probed further into the presence of interactions between load and complexity. To increase power, we pooled the data from Experiments 1-3 (N = 144). Experiment 4 was not included in the pooled analyses because of methodological differences: It involved the production of utterance types (simple and complex noun phrases without adjectives) that were different from those produced in each of Experiments 1-3 (adjective-noun phrases and relative clauses). For the pooled data, we again analyzed the latencies, durations and production errors. For each measure, we ran two models. First, we compared

Experiment 1 with Experiments 2 and 3 together. The fixed predictors in this model were load (no load, load), complexity (simple, complex), experiment (Experiment 1, coded as -0.5, Experiments 2 and 3, each coded as 0.25), and their interactions. The second model compared Experiment 2 with Experiment 3. It had as fixed predictors load, complexity, experiment (Experiment 2, coded as -0.5, Experiment 3, coded as 0.5), and their interactions.

Initiation latencies. The comparison of Experiment 1 with Experiments 2 and 3 confirmed the observation that participants initiated their descriptions more quickly under load than under no load in Experiment 1, but with similar speed under load and no load in Experiments 2 and 3 [load x experiment interaction: *Estimate* = 58.57, *SE* = 27.03, *t* = 2.17, *p* = .03]. The comparison of Experiment 2 with Experiment 3 showed that participants initiated their descriptions more quickly overall in Experiment 3 than in Experiment 2 [main effect of experiment: *Estimate* = -113.43, *SE* = 42.82, *t* = -2.65, *p* = .009]. No other effects or interactions were significant in either analysis.

The disproportional effect of load on the latencies of more complex relative to simpler utterances could be so small that even the pooled data from three experiments was not enough to detect it. To assess our confidence in the absence of a load x complexity interaction, we computed a Bayes Factor for the latency data with the package BayesFactor in R. We ran linear models (lmBF) with and without the interaction term (both, with random intercepts for subjects and items[2]). The Bayes factor in favor of the full model was .05 +/- 4.16% (which amounts to a Bayes factor of 20.17 in favor of the null model). Bayes factors of more than 3 are considered to strongly support the theoretical prediction, while Bayes

[2] Note that the models we used for significance testing had larger random effects structures, but we consider it unlikely that the unequivocal Bayes factor value we obtained crucially depended on the details of the models' random effects structures.

factors of less than .33 are considered to strongly support the null hypothesis; a Bayes factor of .05 thus strongly favored the null hypothesis for our latency data.

Utterance durations. The comparison of Experiment 1 with Experiments 2 and 3 showed main effects of load and complexity, similarly to the analyses of individual experiments. Most relevantly, participants seemed to speed up uttering complex descriptions more than simple descriptions under load [marginal load x complexity interaction: *Estimate* = -15.44, *SE* = 8.23, *z* = -1.88, *p* = .06], and speed up uttering their descriptions under load in Experiments 2 and 3, but not in Experiment 1 [marginal load x experiment interaction: *Estimate* = -28.54, *SE* = 15.02, *z* = -1.90, *p* = .06]. The comparison of Experiment 2 with Experiment 3 also showed that participants sped up uttering complex descriptions more than simple descriptions under load [marginal load x complexity interaction: *Estimate* = -16.29, *SE* = 8.65, *z* = -1.88, *p* = .06].

It is possible that the durations of relative clauses appeared to speed up under load more than those of adjective-noun phrases simply because they are longer, thus giving more opportunities for speeding up articulation than adjective-noun phrases. To control for effects of length on description durations, we conducted analyses on length-corrected durations. To compute those, we regressed the raw durations from Experiments 1-3 onto utterance length in phonemes (here, a more sensitive measure of length than number of syllables, given that 42 of our picture names and four of our color names were monosyllabic). This linear mixed-effects regression model had length as a fixed predictor, a by-subject random intercept, and a by-subject random slope for length (following Fine, Jaeger, Farmer, & Qian, 2013). The residuals of this model were then analyzed with the two joint-experiment models described at the beginning of this section.

These analyses confirmed that participants sped up complex descriptions more than simpler descriptions under load, independently of utterance length [interaction between load and complexity in the model comparing Experiment 1 with Experiments 2 and 3: *Estimate* = -15.93, *SE* = 7.88, *t* = -2.02, *p* = .05; interaction between load and complexity in the model comparing Experiment 2 with Experiment 3: *Estimate* = -15.81, *SE* = 8.16, *t* = -1.94, *p* = .06]. The other significant effects in these analyses were the same as in the analyses of raw description durations, except the main effects of length were no longer significant.

So far, the joint analyses of Experiments 1-3 provided little evidence that load disproportionally disrupted more complex relative to simpler descriptions in our experiments. For initiation latencies, there was no such evidence. Durations were differentially impacted by load, but, again, such that more complex descriptions were uttered more quickly, rather than more slowly, under load. This effect remained even after controlling for description length, possibly because participants might have shortened repeated and predictable words (such as *that is* in relative clauses) more than they were able to do so with nouns or adjectives; as such, this effect was not accounted for by controlling for length in phonemes. In any case, we do not consider this result as evidence for shared resources in a limited capacity system, because there was no *disruption* of production caused by load.

Production errors. The comparison of Experiment 1 with Experiments 2 and 3 showed that participants made significantly fewer errors (a difference of 1.39%) on complex than on simple descriptions in Experiment 1, but more errors (a difference of 0.94%) on complex than on simple descriptions in Experiments 2 and 3 [complexity x experiment interaction: *Estimate* = 0.73, *SE* = .34, *z* = 2.18, *p* = .03]. Also, for complex descriptions, participants made 0.33% more errors in the load than in the no load condition, but, for simple descriptions, they made 1.59% fewer errors in the load than in the no load condition [load x complexity interaction: *Estimate* = 0.48, *SE* = .25, *z* = 1.94, *p* = .052; simple effect of load for simple descriptions: *Estimate* = 0.35, *SE* = .18, *z* = 1.97, *p* = .05; the simple effect of load for complex descriptions was not significant, *p* > .6]. (These models had by-subject random intercepts only.) For the comparison of Experiment 2 with Experiment 3, none of the models converged, even with the simplest possible random-effects structure. We thus do not report any results from these models.

For production errors, the counterintuitive result that fewer errors were made on complex than on simple descriptions in Experiment 1 was not replicated in Experiments 2 and 3, thus providing some support for our assumption that relative clauses were psychologically more complex than adjective-noun phrases. But production errors did not show a disruption under load in the combined analyses – there was a reduction of errors under load for simple descriptions, but no increase of errors under load for complex descriptions.

To shed more light on the error patterns in our experiments, we coded and analyzed the type of errors committed in Experiments 1-3. In Martin et al. (2014), such analyses showed complexity effects but only for syntactic and not for lexical errors: More syntactic errors were made on sentences beginning with complex noun phrases than on sentences beginning with simple noun phrases, while lexical errors showed no such effect. Such a finding in our experiments would attest to a greater psychological complexity of relative clauses than of adjective-noun phrases. Further, analyses of error types could shed light on whether load affected syntactic or lexical processing in any way. These analyses would also provide information about whether our participants computed syntactic structure or retrieved pre-assembled syntactic frames from long-term memory (although note that the absence of effects of load on latency and duration data seem inconsistent with retrieval of pre-assembled frames, as explained in the Introduction). If some syntactic computations were performed, syntactic but not lexical errors should be affected by complexity, insofar as adjective-noun phrases and relative clauses differ in syntactic but not lexical complexity. Conversely, if syntactic frames were retrieved as pre-assembled chunks, syntactic and lexical errors should show similar patterns, because both of them would be retrieved from long-term memory in a similar way.

Following Martin et al. (2014), we coded as lexical errors utterances containing the wrong noun or wrong color adjective (e.g., “the blue rat” or “the purple mouse” instead of “the blue mouse”), such utterances that were interrupted and self-corrected (e.g., “the pur- I mean blue mouse” or “the blue rat-mouse”), and utterances containing disfluencies such as “uh”s and “um”s. We coded as syntactic errors utterances missing the initial definite determiner “the” (e.g., “red ball”), utterances missing the relative pronoun “that” (e.g., “the ball is red”), and utterances with the wrong structure (e.g., “the red ball” when “the ball that is red” was required), including such utterances that were self-corrected (e.g., “the red.. uh, the ball that is red”; although note that producing the wrong structure in our task as well as in Martin et al., 2014, might not have been syntactic in nature).

We performed two separate analyses of lexical errors and syntactic errors. In the analyses of lexical errors, we coded lexical errors as 1 and the rest of the trials as 0; in the analyses of syntactic errors, we coded syntactic errors as 1 and the rest of the trials as 0. The fixed predictors in these models were load, complexity and their interaction.

The analysis of lexical errors revealed (counterintuitively) that more such errors (118, 3.41%) were made with no load than with load (65, 2.47%) [main effect of load: *Estimate* = -.33, *SE* =.16, *z* = -2.08, *p* = .04; this model had only by-subject and by-item random intercepts]. No other effects were significant. In contrast, the analyses of syntactic errors showed that more such errors were made on relative clauses (69, 2.27%) than on adjective-noun phrases (46, 1.51%) [main effect of complexity: *Estimate* =.46, *SE* =.20, *z* = 2.30, *p* = .02; this model had by-subject and by-item random intercepts and a by-subject random slope for the interaction term]. Note that if this result were uniquely due to the fact that relative clauses have more words and thus provide more opportunities for making an error, errors on each of the words would be equally probable. If so, also more lexical errors should have been made on relative clauses than on adjective-noun phrases, different from what we found.

In sum, across Experiments 1-3, more syntactic (but not lexical) errors were made on sentences beginning with complex noun phrases than on sentences beginning with simple noun phrases, consistent with our assumption of a greater psychological complexity of relative clauses relative to adjective-noun phrases. However, across the three experiments, load affected (in the opposite direction) only lexical errors but not syntactic errors, in support of the memory-light account of syntactic formulation in production. Importantly, these results also provide some evidence that our participants did not retrieve pre-assembled syntactic frames from long-term memory. This is because syntactic and lexical errors showed different patterns, and only syntactic errors showed an effect of syntactic complexity. If syntactic frames were retrieved as pre-assembled chunks, effects should have been similar for the two types of errors, because, in such a case, syntactic and lexical information would rely on similar retrieval processes.

Can routinization or strategic planning explain the lack of evidence for working memory involvement in syntactic formulation?

To address this possibility, we included trial order (ranging between 1 and 48 and centered around the mean) and its interactions as fixed predictors (in addition to complexity, load and their interactions) in an LMER model analyzing the pooled latency and duration data from Experiments 1-3 (to increase power). If routinization or the prioritization of recall over picture description accounted for the lack of evidence for working memory involvement in syntactic formulation, we should see a different pattern over the first few trials (before participants adapted to the task and adopted production routines) than the one we report in our main analyses.

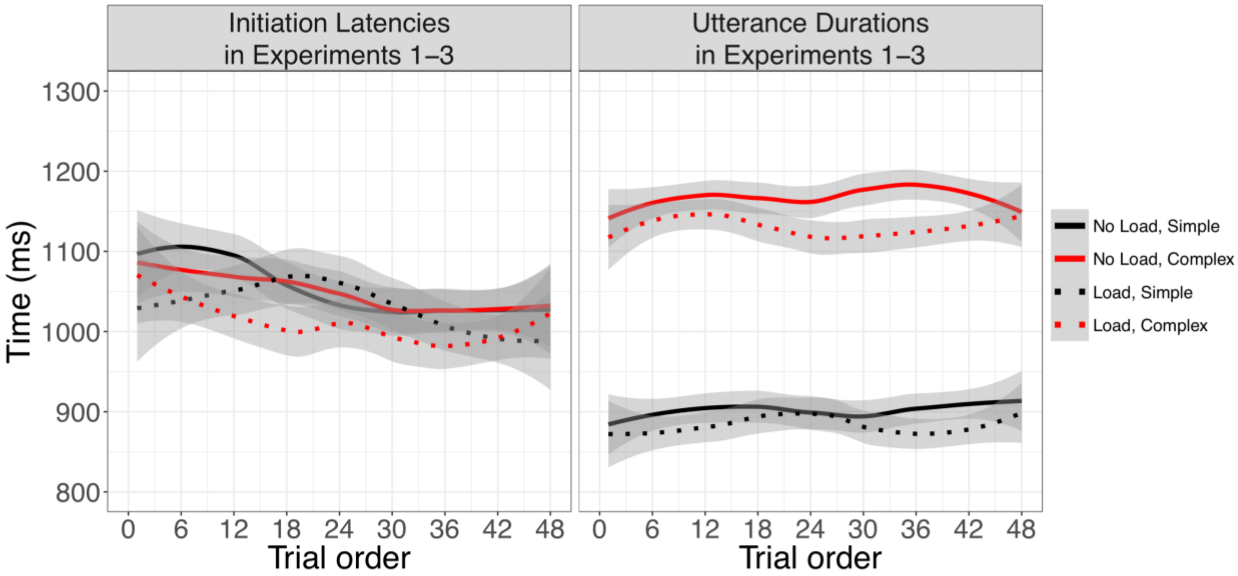


Figure 7. Initiation latencies and utterance durations as a function of trial order in the pooled data of Experiments 1-3. For visualization, trial-level latencies and durations are LOESS-smoothed. Values were computed from the relevant raw data and do not represent F1 means. Ribbons represent 95% confidence intervals.

The latencies and durations as a function of trial order are plotted in Figure 7. An examination of this figure suggests no effects of load at the beginning of the experiment for either latencies or durations, with a speed-up of the durations of more complex (but not simpler) descriptions under load towards the second part of the experiment. The analyses confirmed these observations. The latency analysis showed no significant effects, except a main effect of trial order indicating a gradual speed-up of latencies throughout the experiment [*Estimate* = -1.90, *SE* =.22, *z* = -8.46, *p* < .001]. The durations analysis showed a marginal three-way interaction [*Estimate* = -1.06, *SE* =.58, *z* = -1.84, *p* = .07], hinting at the duration speed-up for complex descriptions under load in the second part of the experiment.[3] These results strongly suggest that routinization or strategies cannot account for our results.

General Discussion

We investigated the involvement of working memory for syntactic formulation in language production. We contrasted a memory-heavy account, under which syntactic formulation requires detectable (even if potentially variable) working memory involvement in all situations or constructions, with a memory-light account, under which it is possible for syntactic formulation to proceed with only minimal working memory involvement in some situations or constructions. To distinguish between these accounts, we conducted four experiments manipulating the complexity of picture descriptions and the presence of verbal load.

Analyses of initiation latencies in all experiments support the memory-light account. In Experiments 1 and 4, participants initiated their descriptions faster under load than under no load, and in Experiments 2 and 3, latencies were unaffected by heavier load or shorter encoding time. Importantly, in neither case were more complex descriptions disproportionally affected by load relative to simpler descriptions.

Analyses of raw and length-corrected description durations across the first three experiments also support the memory-light account. These analyses showed that relative

[3] These models were specified with the following formulas:

Latencies: rt ~ load * complexity * trialOrder + (1 + load + complexity + load:complexity | subj) + (1 + load + complexity + load:complexity | item)

Durations: dur ~ load * complexity * trialOrder + (1 + load + complexity + load:complexity + load:complexity:trialOrder | subj) + (1 + load + complexity + load:complexity | item) clauses were affected by concurrent load more than adjective-noun phrases (and in Experiment 4 durations were unaffected by load). But the direction of this effect was inconsistent with the memory-heavy account: Load disproportionally *sped up* relative-clause descriptions instead of slowing them down. To assume that such a disproportionate speed-up evidences working memory involvement is inconsistent with our original predictions for the memory-heavy account and indeed would violate the basic assumptions behind theories of working memory (Badecker, 1995; Jacobs, 1987; James, 1890). Crucially, under the assumptions that working memory has limited capacity and language production is resource-demanding, evidence for resource demands would be provided by disruption, not facilitation. We consider below the locus of the speed-up in the language production system.

In contrast, production errors in two of the four experiments seem to support the memory-heavy account. In Experiments 2 and 4, errors increased under load for more complex descriptions but remained unchanged for simpler descriptions. In contrast to the latency and duration data, these results suggest greater working memory involvement in the production of more complex than of simpler descriptions. Assuming that internal monitoring processes are ultimately responsible for error detection, an increase of errors under load might suggest a role of working memory for pre-articulation monitoring. Such a conclusion is consistent with the conclusions reached in prior studies of the role of working memory in syntactic formulation in production (Fayol et al., 1994; Hartsuiker & Barkhuysen, 2006; Hupet et al., 1998), as well as with studies suggesting that monitoring processes are compromised under production pressure (Baars, Motley & Mackay, 1975; Horton & Keysar, 1996). Such a conclusion also provides an explanation for the speed-up of initiation latencies and utterance durations under load. A disproportionate increase of production errors for more complex than for simpler utterances under load, in the presence of a latency and duration speed-up, might suggest that monitoring processes were compromised for the sake of greater production speed.

However, separate analyses of lexical and syntactic errors across Experiments 1-3 showed complexity effects only for syntactic errors, and load effects (in the opposite direction) only for lexical errors. These analyses attest to the greater psychological complexity of relative clauses relative to adjective-noun phrases (although see below our discussion of complexity), but undermine the involvement of working memory in syntactic formulation.

Taken together, the results of this study provide stronger support for the memory-light account than for the memory-heavy account. By contrasting two structures which differ in structural complexity but do not involve long-distance dependencies or extrapositions, the contribution of the present study is in demonstrating that it is in principle possible for syntactic formulation in production to proceed with minimal working memory involvement.

This conclusion opens several possibilities (left for future research) about how working memory resources are recruited in real-life production. It is possible that working-memory use is structure-independent but instead depends on resource availability in different circumstances – it might be used even for the type of structures tested here when resources are ample but not used when resources are scarce. It is also possible that working memory is consistently not recruited for specific types of (simpler) structures, freeing up recourses for other aspects of production such as conceptual formulation (Bock, 1982), other aspects of syntactic formulation within the same (longer) utterance such as long-distance dependencies or agreement (Fayol et al., 2004), or other aspects of cognition such as attending to unexpected events.

One might argue that we would have found the predicted effects of working memory load (and, potentially, support for the memory-heavy account) with a larger complexity difference between the simple and complex descriptions. However, we think that a larger complexity difference would have made the two structures different in too many other ways and hence less straightforwardly comparable. For example, a more complex structure might have required more computation of agreement, possibly long-distance dependencies, and certainly more lexical items. We therefore chose two structures which differ in complexity yet are as comparable as possible in other respects.

Also, because of the relative simplicity of both structures and the nature of the task, syntactic formulation in our experiments could have involved retrieving preassembled syntactic frames from long-term memory instead of performing syntactic computations from scratch. However, this possibility was undermined by the different patterns of syntactic and lexical errors across Experiments 1-3. Further, for reasons explained above, such a scenario would predict greater effects of concurrent load for more complex relative to simpler utterances, inconsistent with our findings. Lastly, syntactic formulation in natural-language production would at least occasionally involve retrieval of preassembled syntactic frames, as in formulaic utterances; thus, such retrieval in our experiments is not inconsistent with our conclusions. Also note that paradigms involving much more continuous repetition than ours have evidenced the computation of abstract structure in other domains. For example, production times were longer for structurally dissimilar (e.g., KIL – KILP.NER and KILP – KIL.PER) than structurally similar syllables (e.g, KIL – KIL.PER and KILP – KILP.NER) during continuous (4 sec) repetition of these syllables, suggesting that speakers compute abstract syllable structure separable from phonemic content (Sevald, Dell, & Cole, 1995; see also Sevald & Dell, 1994).

Our findings have implications for the flexibility of the language production system in general. We found that participants initiated (similarly to Power, 1985) or uttered their descriptions faster under load (while, in Experiments 1 and 3, also maintaining accuracy). This pattern suggests that the production system has the *capacity* (albeit presumably up to a point) to modulate (and, if necessary, speed up) the execution of different production processes in response to task demands or context. But what is the mechanism behind such flexibility? One possibility during the planning stage is incremental planning: Ferreira and Swets (2002) showed that speakers produced utterances incrementally only with a response deadline, but not without. Note, however, that such incremental planning could not have been too radical (i.e., limited to the definite determiner in our descriptions) because we did not find any effects of load on determiner durations in Experiment 2. Further, during the articulation stage, repeated utterance elements (such as the relative pronoun *that* and verb *is* in relative clauses, and color adjectives in both noun phrases and relative clauses) might provide room for speeding up both planning and articulation. (Although note that planning and articulation of the repeated conjunction *and* in complex noun phrases did not produce any decrease in description durations in Experiment 4.) Yet another possibility is that production flexibility can be achieved by a reduction in internal monitoring, as discussed above. Indeed, it seems that participants in all experiments prioritized the recall task over the production task, which accounts for their increased production speed. But this pattern of performance still shows that the production system, at least for our relatively simple target utterances, can function not only relatively normally but also faster when priority is given to another activity (as could be driving or operating tools in real life).

Another indication of this flexibility is the fact that description latencies for adjective-noun phrases and relative clauses were approximately 1000 ms on no load trials in Experiments 1 and 3, but approximately 1100 ms on such trials in Experiment 2. This again suggests that production processes are modulated in response to global task demands (in this case, slowed down, to ensure normal production under the overall more demanding task in Experiment 2), possibly by allowing more time for individual production processes. (Note, however, that the longer latencies in Experiment 2 could also stem from lexical interference carried over from the four words to be remembered on preceding load trials.)

The results we report here bear on the relationship between linguistic complexity, psychological complexity, and measures of production difficulty. Specifically, it is often assumed that psychological complexity translates directly into production difficulty, and such difficulty is measured with structure choices in production (more difficult structures are produced less often), and utterance latencies and durations (more difficult structures are produced more slowly). However, structure choices in production seem determined by a host of factors instead of or beyond linguistic complexity – such as availability of alternative structures, head-noun animacy, interference from semantically-similar nouns, case marking, word-order flexibility and frequency (e.g., Gennari, Mirković, & MacDonald, 2012; see MacDonald, Montag, & Gennari, 2016, for arguments against dependency distance itself leading to production difficulty). In view of such evidence, our finding that structural complexity did not slow down production is not unexpected (although note that relative clauses are typically produced less frequently than adjective-noun phrases in experimental contexts similar to ours but involving a structure choice: e.g., Santesteban, Pickering, & McLean, 2010). While we do not claim that relative clauses are equally easy as adjective-noun phrases (because we defined complexity in terms of necessary number of computations, and hence assumed that it translated into psychological complexity and thus difficulty), we caution that neither of the latter must directly translate into production slowing, for example because of the flexibility of the production system discussed above. In other words, production difficulty might amount to performing more mental operations and recruiting more cognitive resources, but not be evident in behavioral measures assumed to reflect it. (Also note that we did find some psychological complexity effects in error analyses.) Taken together, our results and others suggest that there is not a one-to-one correspondence between psychological complexity and *measures* of production difficulty, and caution is necessary when defining their relationship.

Finally, what do our results imply for the automaticity of syntactic formulation in production? Bock (1982) discussed features of automatic and non-automatic processing, and seemed to equate lack of working memory involvement with automaticity of processing. A useful discussion regarding the automaticity in language processing was recently provided by Hartsuiker and Moors (in press). These authors point out that automaticity is traditionally defined by a number of features (e.g., lack of intentionality or conscious awareness, efficiency, uncontrollability) but is regarded as an all-or-none phenomenon: A process is either automatic or non-automatic. Instead, Hartsuiker and Moors (following Moors, 2016) suggest that different activities may be characterized by only some of these features, and only to some degree. They conclude (p. 18) that "it may be more fruitful to consider automaticity features as a subset of many mutually compensatory factors that jointly influence whether and how a particular process will be carried out." We agree with this reasoning, and we do not assume that we have found one language production process which might be fully automatic; for example, it might be that syntactic formulation in production tasks similar to ours would be affected by increased attentional demands, or a different operationalization of memory load. What we show here is that syntactic formulation in language production *can* proceed without detectable temporary maintenance of syntactic constituents in working memory before they reach the next processing stage.

Acknowledgements

This research was supported by an R01from NICHD (HD051030), awarded to Victor S. Ferreira. Heartfelt thanks go to Coy Shaffstall, Alexandra Schuman, Emily Barker, Hanako Shimizu and William Kim for data collection, to Coy Shaffstall, Emily Barker and Cristina Lloyd for time-stamp extraction and coding, to Aziz Atiya for coding recall performance in Experiment 3, and to Adrian Staub and Daniel Kleinman for help with the Bayes factor calculations and the trial-level plots, respectively. The authors declare no conflict of interest.

References

Alario, F. X., Costa, A., & Caramazza, A. (2002). Frequency effects in noun phrase production: Implications for models of lexical access. *Language and Cognitive Processes, 17*(3), 299-319.

Allport, D. A., Antonis, B., & Reynolds, P. (1972). On the division of attention: A disproof of the single channel hypothesis. *The Quarterly Journal of Experimental Psychology*, *24*(2), 225-235.

Allum, P. H., & Wheeldon, L. R. (2007). Planning scope in spoken sentence production: The role of grammatical units. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 791-810.

Allum, P. H., & Wheeldon, L. (2009). Scope of lexical access in spoken sentence production: Implications for the conceptual–syntactic interface. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1240-1255.

Baars, B. J., Motley, M. T., & MacKay, D. G. (1975). Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior*, *14*(4), 382-391.

Baayen, R.H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Baddeley, A. D. (1986). *Working memory.* New York: Oxford University Press.

Baddeley, A. D. (1995). Working Memory. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (pp. 755-764). Campidge, Mass.: The MIT Press.

Baddeley, A. D., & Hitch, G. J. (1974). *The Psychology of Learning and Motivation* . New York: Academic Press.

Bader, M. (2017). Working memory involvement in grammatical encoding. Poster presented at the 30th CUNY Conference on Human Sentence Processing, Cambridge, MA.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255-278.

Belke, E. (2008). Effects of working memory load on lexical-semantic encoding in language production. *Psychonomic Bulletin & Review, 15*(2), 357-363.

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review, 89*(1), 1-47.

Bock, J. K. (1990). Structure in language: Creating form in talk. *American Psychologist, 45,* 1221-1236.

Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language, 31*(1), 99-127.

Bock, K., & Levelt, W. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.) *Handbook of Psycholinguistics* (pp. 945-984). San Diego: Academic Press.

Brown-Schmidt, S., & Konopka, A. E. (2008). Little houses and casas pequeñas: Message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition, 109*(2), 274-280.

Brown-Schmidt, S., & Konopka, A. E. (2015). Processes of incremental message planning during conversation. *Psychonomic Bulletin & Review, 22*(3), 833-843.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences, 22*(1), 77-94.

Chang, F., Dell, G. S., & Bock, J. K. (2006). Becoming syntactic. *Psychological Review, 113,* 234-272.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62-101). Cambridge: Cambridge University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87-185.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., et al. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51,* 42-100.

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language, 25*(1), 1-18.

Davis, C.J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods, 37,* 65-70.

Deese, J. (1978). Thought into speech: Linguistic rules and psychological limitations in processing information determine how we put our ideas into words. *American Scientist, 66*(3), 314-321.

Deese, J. (1980). Pauses, prosody, and the demands of production in language. In W. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honor of Frieda Goldman-Eisler.* The Hague: Mouton.

de Fockert, J. W., Rees, G., Frith, C. D., & Lavie, N. (2001). The role of working memory in visual selective attention. *Science, 291*(5509), 1803-1806.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*, 19-23.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Fayol, M., Largy, P., & Lemaire, P. (1994). Cognitive overload and orthographic errors: When cognitive overload enhances subject–verb agreement errors. A study in French written language. *The Quarterly Journal of Experimental Psychology*, *47*(2), 437-464.

Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language, 54*(4), 541-553.

Fedorenko, E., Gibson, E., & Rohde, D. (2007). The nature of working memory in linguistic, arithmetic and spatial integration processes. *Journal of Memory and Language, 56*(2), 246-269.

Ferreira, V. S., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1187-1199.

Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*(1), 57-84.

Fyndanis, V., Arcara, G., Christidou, P., & Caplan, D. (in press). Morphosyntactic Production and Verbal Working Memory: Evidence from Greek Aphasia and Healthy Aging. *Journal of Speech, Language, and Hearing Research*.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS one, 8*(10), e77661.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35,* 116-124.

Gennari, S. P., Mirković, J., & MacDonald, M. C. (2012). Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive Psychology*, *65*(2), 141-176.

MacDonald, M. C., Montag, J. L., & Gennari, S. P. (2016). Are there really syntactic complexity effects in sentence production? A reply to Scontras et al. (2015). *Cognitive Science*, *40*(2), 513-518.

Goldberg, A.E. (1995). *Constructions: A construction grammar approach to argument structure.* Chicago: Chicago University Press.

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

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech.* New York: Academic Press.

Gregg, V. H., Freedman, C. M., & Smith, D. K. (1989). Word frequency, articulatory suppression and memory span. *British Journal of Psychology*, *80*(3), 363-374.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition, 82,* B1-B14.

Griffin, Z. M. (2003). A reversed word length effect in coordinating the preparation and articulation of words in speaking. *Psychonomic Bulletin & Review*, *10*(3), 603-609.

Hartsuiker, R. J., & Barkhuysen, P. N. (2006). Language production and working memory: The case of subject-verb agreement. *Language and Cognitive Processes*, *21*(1-3), 181-204.

Hartsuiker, R. J., Corley, M., & Martensen, H. (2005). The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related beply to Baars et al.(1975). *Journal of Memory and Language*, *52*(1), 58-70.

Hartsuiker, R. J., & Moors, A. (in press). On the Automaticity of Language Processing. In H.-J. Schmid (Ed.), *Entrenchment, memory and automaticity. The psychology of linguistic knowledge and language learning* (pp. 201-223). Berlin: De Gruyter Mouton.

Holmes, V. M. (1988). Hesitations and sentence planning. *Cognition, 3,* 323-361.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91-117.

Hupet, M., Fayol, M., & Schelstraete, M. A. (1998). Effects of semantic variables on the subject-verb agreement processes in writing. *British Journal of Psychology, 89*(1), 59-75.

Jacobs, J. (1887). Experiments on "prehension." *Mind, 12,* 75-79.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434-446.

Jaeger, T. F. (2013, Feb 14). Re: queries regarding significance of interaction term, simple effects, and false convergence [Electronic mailing list message]. Retrieved from https://mailman.ucsd.edu/pipermail/ling-r-lang-l/2013-February/000461.html

James, W. (1890). *Principles of psychology.* New York: Holt.

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824-843.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99(1),* 122-149.

Kellogg, R. T. (2004). Working Memory Components in Written Sentence Generation. *The American Journal of Psychology, 117 (3),* 341-361.

Kemper, S., Herman, R. E., & Lian, C. H. (2003). The costs of doing two things at once for young and older adults: Talking while walking, finger tapping, and ignoring speech of noise. *Psychology and Aging*, *18*(2), 181-192.

Kok, P., van Doorn, A., & Kolk, H. (2007). Inflection and computational load in agrammatic speech. *Brain and Language*, *102*, 273-283.

Laver, J. D. M. (1980). Monitoring systems in the neurolinguistic control of speech production. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand.* New York: Academic Press.

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General, 133,* 339-354.

Lee, E. K., Brown-Schmidt, S., & Watson, D. G. (2013). Ways of looking ahead: Hierarchical planning in language production. *Cognition*, *129*(3), 544-562.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Levin, H., Silverman, I., & Ford, B. L. (1967). Hesitations in children's speech during explanation and description. *Journal of Verbal Learning and Verbal Behavior*, *6*(4), 560-564.

Lindsley, J. R. (1975). Producing simple utterances: How far ahead do we plan? *Cognitive Psychology*, *7*, 1-19.

Martin, R. C., Crowther, J. E., Knight, M., Tamborello, F. P., & Yang, C. L. (2010). Planning in sentence production: Evidence for the phrase as a default planning scope. *Cognition*, *116*(2), 177-192.

Martin, R. C., Yan, H., & Schnur, T. T. (2014). Working memory and planning during sentence production. *Acta Psychologica*, *152*, 120-132.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *27*(3), 817-835.

Meyer, A. S. (1996). Lexical access in phrase and sentence production: Results from picture–word interference experiments. *Journal of Memory and Language*, *35*(4), 477-496.

Moors, A. (2016). Automaticity: Componential, causal, and mechanistic explanations. *Annual Review of Psychology*, *67*, 263-287.

Myachykov, A., Scheepers, C., Garrod, S., Thompson, D., & Fedorova, O. (2013). Syntactic flexibility and competition in sentence production: The case of English and Russian. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1601-1619.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, *7*(1), 44-64.

Nozari, N., Dell, G. S., & Schwartz, M. F. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology*, *63*(1), 1-33.

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. Cambridge, MA: MIT Press.

Oberauer, K., Sü., H.-M., Wilhem, O., & Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In A. R. A. Conway, C. Jarrold, M. H. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 49–75). New York: Oxford University Press.

Oldfield, R.C., &Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology, 17*, 273-281.

Pickering, M. J., & Garrod, S. (2014). Self-, other-, and joint monitoring using forward models. *Frontiers in Human Neuroscience, 8.*

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition*, *77*(2), 97-132.

Power, M. J. (1985). Sentence production and working memory. *The Quarterly Journal of Experimental Psychology*, *37*(3), 367-385.

Rosenthal, R., & Rosnow, R. L. (1991). Essentials of behavioral research: Methods and data analysis (2nd ed.). New York, NY: McGraw-Hill Series in Psychology.

Santesteban, M., Pickering, M. J., & McLean, J. F. (2010). Lexical and phonological effects on syntactic processing: Evidence from syntactic priming. *Journal of Memory and Language*, *63*(3), 347-366.

Segaert, K., Menenti, L., Weber, K., & Hagoort, P. (2011). A paradox of syntactic priming: Why response tendencies show priming for passives, and response latencies show priming for actives. *PLoS One, 6(10)*, e24209.

Segaert, K., Wheeldon, L. & Hagoort, P. (2016) Unifying structural priming effects on syntactic choices and timing of sentence generation. *Journal of Memory and Language, 91*, 59-80.

Shaffer, L. H. (1975). Multiple attention in continuous verbal tasks. *Attention and performance V*, 157-167.

Redelmeier, D. A., & Tibshirani, R. J. (1997). Association between cellular-telephone calls and motor vehicle collisions. *New England Journal of Medicine*, *336*(7), 453-458.

Sevald, C. A., & Dell, G. S. (1994). The sequential cuing effect in speech production. *Cognition*, *53*(2), 91-127.

Sevald, C. A., Dell, G. S., & Cole, J. S. (1995). Syllable structure in speech production: Are syllables chunks or schemas? *Journal of Memory and Language*, *34*(6), 807-820.

Schriefers, H., & Teruel, E. (1999). The production of noun phrases: A cross-linguistic comparison of French and German. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 637-642). Mahwah, NJ: Erlbaum.

Scontras, G., Badecker, W., Shank, L., Lim, E., & Fedorenko, E. (2014). Syntactic complexity effects in sentence production. *Cognitive Science, 39*(3), 559-583.

Shah, P., & Miyake, A. (1999). Models of working memory: An Introduction. In A. Miyake and P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 1-27). Cambridge: Cambridge University Press.

Shipstead, Z., Lindsey, D. R., Marshall, R. L., & Engle, R. W. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language*, *72,* 116-141.

Slevc, L. R. (2011). Saying what's on your mind: working memory effects on sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1503-1514.

Slevc, L. R., & Martin, R. C. (2016). Syntactic agreement attraction reflects working memory processes. *Journal of Cognitive Psychology*, *28*(7), 773-790.

Smith, M., & Wheeldon, L. (1999). High level processing scope in spoken sentence production. *Cognition*, *73*(3), 205-246.

Smith, M., & Wheeldon, L. (2001). Syntactic priming in spoken sentence production – an online study. *Cognition*, *78*(2), 123-164.

Stins, J. F., Vosse, S., Boomsma, D. I., & De Geus, E. J. (2004). On the role of working memory in response interference. *Perceptual and Motor Skills*, *99*(3), 947-958.

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., Lu, C.C., Pechmann, T., Pléh, C., Wicha, N., & Federmeier, K. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language, 51*(2), 247-250.

Unsworth, N., & Engle, R. W. (2007b). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114,* 104-132.

Wagner, V., Jescheniak, J. D., & Schriefers, H. (2010). On the flexibility of grammatical advance planning during sentence production: Effects of cognitive load on multiple lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 423-440.

Wasow, T. (1997). End-weight from the speaker's perspective. *Journal of Psycholinguistic Research*, *26*(3), 347-361.

Wheeldon, L., Ohlson, N., Ashby, A., & Gator, S. (2013). Lexical availability and grammatical encoding scope during spoken sentence production. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1653-1673.

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of empirical literature. *Psychonomic Bulletin & Review, 1,* 89-106.

Wright, C. E. (1979). Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory & Cognition*, *7*(6), 411-419.

Yngve, V. H. (1973). I forget what I was going to say. In *Papers from the Ninth Regional Meeting, Chicago Linguistic Society* (pp. 688-699). Chicago, IL: University of Chicago Press.