Cognition 251 (2024) 105875

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full Length Article

Hijacking limitations of working memory load to test for composition in language

Michael T. Ullman [a,*,1], Talat Bulut [b,c,1], Matthew Walenski [d,1]

[a] Brain and Language Laboratory, Department of Neuroscience, Georgetown University, Washington, DC, USA
[b] Neurobiology of Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands
[c] Department of Speech and Language Therapy, Istanbul Medipol University, Istanbul, Turkey
[d] Department of Communication Sciences and Disorders, East Carolina University, Greenville, NC, USA

ARTICLE INFO

Keywords:
Language
Composition
Working memory
Morphology
Regular
Irregular

ABSTRACT

Although language depends on storage and composition, just what is stored or (de)composed remains unclear. We leveraged working memory load limitations to test for composition, hypothesizing that decomposed forms should particularly tax working memory. We focused on a well-studied paradigm, English inflectional morphology. We predicted that (compositional) regulars should be harder to maintain in working memory than (non-compositional) irregulars, using a 3-back production task. Frequency, phonology, orthography, and other potentially confounding factors were controlled for. Compared to irregulars, regulars and their accompanying *−s/−ing*-affixed filler items yielded more errors. Underscoring the decomposition of only regulars, regulars yielded more bare-stem (e.g., *walk*) and stem affixation errors (*walks/walking*) than irregulars, whereas irregulars yielded more past-tense-form affixation errors (*broughts/tolded*). In line with previous evidence that regulars can be stored under certain conditions, the regular-irregular difference held specifically for phonologically consistent (not inconsistent) regulars, in particular for both low and high frequency consistent regulars in males, but only for low frequency consistent regulars in females. Sensitivity analyses suggested the findings were robust. The study further elucidates the computation of inflected forms, and introduces a simple diagnostic for linguistic composition.

1. Introduction

Language involves both storage and composition. This is not controversial at the extremes. There is broad agreement that idiosyncratic lexical knowledge must be stored in memory (e.g., a memorized link between the sound /dɑg/ and its referential meaning to the slobbery pets people keep), whereas novel sentences are clearly not memorized, but rather must be somehow composed out of smaller stored pieces (e.g., "*Triumphant pickles marched for saltier brine.*"). The status of many other forms in language, such as collocations, idioms, and morphologically derived and inflected words (e.g., *hackneyed*), is not as clear cut. Overall, the question of precisely what is memorized and what is assembled from smaller pieces has not been an easy question to answer. Here we explore the efficacy of a new method for testing the relative degree of composition of linguistic forms, and test it on a well-studied model organism in language, namely the distinction between regular and irregular inflected forms.

The representation of inflected morphological forms, such as past-tenses of English verbs, has been the source of contentious debate over the years, with various claims for regular and irregular verbs (for reviews and various perspectives, see Bybee, 1995; Lavric, Pizzagalli, Forstmeier, & Rippon, 2001; Pinker, 2006; Pinker & Ullman, 2002; Seidenberg & Plaut, 2014). According to some theories, such as the Words and Rules theory and the Declarative/Procedural model, regular inflected forms, that is, those that follow a default affixation rule, are generally (but not always; see Section 3.4) composed from their constituent parts when produced, and are broken down into these components when comprehended (Clahsen, 1999; Marslen-Wilson & Tyler, 1998; Pinker, 1999; Pinker & Ullman, 2002; Ullman, 2004, 2016). For example, on this view the production of English regular past-tense forms, whose stems take an *-ed* suffix (e.g., *walked*), typically involves composition of the stored stem (*walk*) with the inflectional past-tense

* Corresponding author at: Brain and Language Laboratory, Department of Neuroscience, Georgetown University, Box 571464, Washington, DC 20057-1464.
E-mail addresses: michael@georgetown.edu (M.T. Ullman), tbulut@medipol.edu.tr (T. Bulut), walenskim21@ecu.edu (M. Walenski).

[1] Equal contributions.

https://doi.org/10.1016/j.cognition.2024.105875
Received 22 June 2023; Received in revised form 2 July 2024; Accepted 3 July 2024
Available online 16 July 2024

affix (−*ed*). In contrast, some approaches have claimed that all regular inflected forms are represented in memory as whole words (Bybee, 1995; Manelis & Tharp, 1977), or, on a connectionist perspective, are learned and processed in associative memory (Joanisse & Seidenberg, 1999, 2005; McClelland & Patterson, 2002; Seidenberg & Plaut, 2014).

Irregular past-tense forms (e.g., *dig-dug*) are always posited to involve memorized representations, though the exact nature of these representations varies across models. Traditionally, it has been assumed that irregulars are stored as whole words in memory (Bybee, 1995; Kuczaj, 1977). On a connectionist view, irregulars are learned and processed in associative memory (Joanisse & Seidenberg, 1999, 2005; McClelland & Patterson, 2002; Seidenberg & Plaut, 2014). Similarly, the Words and Rules theory and the Declarative/Procedural model suggest that irregulars are represented in an associative memory system (Pinker, 1999; Pinker & Ullman, 2002; Ullman, 2016). Alternatively, according to Distributed Morphology, irregulars are partly compositional and partly memory-dependent (Fruchter, Stockall, & Marantz, 2013; Halle & Marantz, 1993; Stockall & Marantz, 2006). On this perspective, irregular verbs are linked in memory to specific affixes (e.g., *feel* + −*t*, *dig* + −*0*); inflecting an irregular verb involves combining the verb stem and the particular affix, as well as the application of any word-specific "stem-readjustment" rules (e.g., the vowel change from stem to past-tense form in *feel-felt* or *dig-dug*) (Halle & Marantz, 1993).

To summarize, depending on the theoretical construct, regular and/or irregular inflected forms may be compositional. Numerous empirical studies have examined the neurocognition of regular and irregular morphology, using a variety of psycholinguistic and cognitive neuroscience methods (for reviews, see Bybee, 1995; Clahsen, 1999; Lavric et al., 2001; Leminen, Smolka, Duñabeitia, & Pliatsikas, 2019; Marslen-Wilson & Tyler, 1998; McClelland & Patterson, 2002; Pinker, 1999; Pinker & Ullman, 2002; Seidenberg & Plaut, 2014; Ullman, 2004, 2016). Nevertheless, the conflicting claims discussed above have yet to be completely resolved. In particular, it has been difficult to test specifically for *composition.* Cognitive neuroscience methods such as functional neuroimaging (Bulut, 2022; Leminen et al., 2019) or lesion studies (Ullman, 2004) may show dissociations between forms that are posited to be compositional versus those that are not, but cannot easily provide direct evidence for or against composition. Similar arguments hold for behavioral measures. For example, although the presence of inflected-form frequency effects (Alegre & Gordon, 1999; Baayen, Dijkstra, & Schreuder, 1997; Prado & Ullman, 2009) has been used as a diagnostic of the storage of inflected forms, their absence can be difficult to interpret, and does not necessarily imply composition (Prado & Ullman, 2009). Similarly, larger inflected-form to stem priming effects for regulars (*walked* priming *walk*) than irregulars (*sang* priming *sing*), which have been interpreted as evidence for the composition of regulars (Marslen-Wilson & Tyler, 1998; Silva & Clahsen, 2008), might instead reflect other factors (Kielar, Joanisse, & Hare, 2008; Marslen-Wilson, 2007; Morgan-Short & Ullman, 2022; Rueckl, Mikolinski, Raveh, Miner, & Mars, 1997).

Here, we directly test for linguistic composition with a novel experimental method, focusing on inflected past-tense forms. Our method leverages working memory, which can be broadly defined as "a system of components that holds a limited amount of information temporarily in a heightened state of availability for use in ongoing processing" (Adams, Nguyen, & Cowan, 2018). This quoted definition is intended to be generic, and to apply across different working memory accounts. Indeed, a wide range of theoretical constructs of working memory have been proposed (Adams et al., 2018; Atkinson & Shiffrin, 1968; Baddeley, 2012; Baddeley, Hitch, & Allen, 2021; Cowan, 2001; Cowan, Morey, & Naveh-Benjamin, 2020; Engle & Kane, 2003; McElree, 2001, 2006; Oberauer, 2009, 2020; Öztekin, Davachi, & McElree, 2010; Wilhelm, Hildebrandt, & Oberauer, 2013). For instance, some accounts posit that working memory is subserved by components that are at least partly distinct from other neurocognitive systems, including long-term memory (Atkinson & Shiffrin, 1968; Baddeley, 2012; Baddeley et al., 2021), whereas others argue that the construct of working memory closely involves long-term memory (Crowder, 1993; Nairne, 1996, 2002; Neath & Surprenant, 2005), and, for example, can be largely explained by attentional processes focused on long-term representations (Cowan, 2001; Cowan et al., 2020; Engle & Kane, 2003; McElree, 2001, 2006). Overall, the various accounts of working memory invoke a variety of underlying mechanisms and concepts, including (somewhat overlapping) notions of encoding, maintenance, storage, capacity, load, rehearsal, attention, retrieval, interference, decay, processing, and manipulation (Adams et al., 2018; Baddeley et al., 2021; Cowan et al., 2020; Engle & Kane, 2003; Oberauer, 2020; Van Dyke & McElree, 2006). Regardless of the underlying mechanisms, the various accounts have attempted to explain the critical working memory characteristic of interest here, namely the limited amount of information that is temporarily available, which has been discussed in a variety of ways, including in terms of capacity limitations, attentional focus, and/or retrieval difficulties (Baddeley, 2012; Baddeley et al., 2021; Cowan, 2001; Cowan et al., 2020; McElree, 2001, 2006; Miller, 1956). Our experimental method is motivated by this limitation.

The extent of this limitation, which varies according to item, subject, and other factors (e.g., word frequency, word length, phonological/orthographic similarity, neighborhood density, age) (Baddeley, 1966; Clarkson, Roodenrys, Miller, & Hulme, 2017; Conrad, 1964; Gathercole & Baddeley, 1993; Longoni, Richardson, & Aiello, 1993; McNab et al., 2015; Pliatsikas et al., 2019; Roodenrys, Hulme, Lethbridge, Hinton, & Nimmo, 2002; Saint-Aubin, Poirier, Yearsley, Robichaud, & Guitard, 2023), usually falls within a similar range across theoretical perspectives. Miller (1956) famously estimated that 7 plus or minus 2 items can be temporarily maintained. More recent estimates suggest that only 3 to 5 elements may generally be held together in working memory (Baddeley, 2012; Cowan, 2010), or even that only one item can be in focal attention at a time (McElree, 2001, 2006). Regardless of the limit, the crux of the idea is this: If maintaining (or retrieving)[2] compositional linguistic forms requires maintaining their constituent parts, then the greater the number of constituent parts, the greater the strain on working memory. *Thus, our technique hijacks working memory limitations to test for composition in language*. Of particular interest here, maintaining several compositional forms (e.g., regular inflected forms) in memory may be expected to tax an individual's working memory more than maintaining the same number of non-compositional forms (e.g., irregular inflected forms), leading to worse performance at the former than the latter, all else being equal.[3]

To examine this prediction for regular and irregular English past-tense forms, we created a production n-back task similar to previous digit or letter recall n-back tasks (Amon & Bertenthal, 2018; Dobbs & Rule, 1989; Monk, Jackson, Nielsen, Jefferies, & Olivier, 2011; Pelegrina et al., 2015), which are sensitive to working memory load limitations (Nikolin et al., 2021; Owen, McMillan, Laird, & Bullmore, 2005; Schmiedek, Lövdén, & Lindenberger, 2014; Wilhelm et al., 2013). In a *within-subjects design*, regular and matched irregular past-tense forms were presented one at a time on a computer screen in pseudo-

[2] For simplicity, we will henceforth use the term "maintaining" to refer to information available to working memory, but as explained below, we are agnostic as to whether information needed to perform the task is maintained in working memory or must be retrieved from long-term memory.

[3] This prediction appears to hold across different accounts of working memory, even though their underlying mechanisms differ. For example, whether items are maintained in a buffer, are the focus of attention, and/or are retrieved from long-term memory (Baddeley, 2012; Baddeley et al., 2021; Cowan, 2001; Cowan et al., 2020; McElree, 2001, 2006; Oberauer, 2009, 2020), the larger number of items associated with decomposed forms may be expected to generally result in a greater strain on working memory than matched items that are not decomposed. Nonetheless, we emphasize that different working memory accounts may explain this pattern at least somewhat differently; also see Discussion.

randomized orders, interspersed with other inflected forms and bare stems. Immediately after each word a fixation cross appeared, which cued participants to say out loud the third word back in the sequence (3-back recall). The task requires maintaining three words in working memory: the word just presented (1-back), the previous word (2-back), and the third word back (3-back, the target item). For example, in the sequence *walked*, *died*, *snoozes*, *flail*, …, the participant should say *walked* upon seeing *snoozes*, *died* after seeing *flail*, and so on.

Different outcomes are expected according to different morphological perspectives. If regular but not irregular past-tenses are generally compositional (Clahsen, 1999; Marslen-Wilson & Tyler, 1998; Pinker & Ullman, 2002; Ullman, 2004, 2016), as we predicted, performance at regular past-tense forms should generally be worse than performance at irregulars in the task. In addition, performance on the filler words that are maintained in working memory together with regular past-tenses should suffer more than those maintained together with irregular past-tenses, due to the extra burden that the regular stems and affixes put on working memory. In contrast, regular/irregular performance differences might not generally be expected if regular and irregular past-tenses are both compositional (Fruchter et al., 2013; Halle & Marantz, 1993; Stockall & Marantz, 2006) or if neither are (Bybee, 1995; Joanisse & Seidenberg, 1999; Seidenberg & Plaut, 2014), as long as phonological, lexical, and other factors are controlled for.

2. Methods

2.1. Participants

We tested 46 right-handed monolingual native English speakers (27 male, 19 female) with a mean age of 21.4 years (*SD* 5.2), and a mean of 14.5 years of education (*SD* 1.7). Participants had no uncorrected visual or auditory impairments and no known history of neurological, psychiatric, or neurodevelopmental disorders. The study was approved by the Georgetown University Institutional Review Board. Participants provided informed consent prior to testing.

2.2. Materials

The experimental items comprised 64 regular and 64 irregular past-tense forms (Supplementary Material, Table S1). The regular and irregular verbs were selected based on pairwise matching for 1) phonological structure (all verbs were monosyllabic and, to the extent possible, each regular/irregular pair had similar onsets and coda consonant clusters); and 2) surface frequency of the inflected forms (the combined past-tense and past-participle frequencies where these forms are identical). We matched on surface frequency rather than past-tense frequency because the past-tense items were presented in isolation in the task (i.e., not in sentence contexts), so these words could have been interpreted as past-participles. For the frequency measure, we used log10 of SUBTLEX-US contextual diversity, as this has been shown to predict lexical decision reaction times and accuracy more closely than various other frequency norms (Brysbaert & New, 2009). The contextual diversity measure reflects the number of contexts (number of films) in which a word occurs (Brysbaert & New, 2009). The log10 contextual diversity of the regular forms ($M = 2.656$) and irregular forms ($M = 2.707$) did not differ ($t(63) = 0.340$, $p = .735$). See Section 2.6 (Covariates) and Table 1 for additional item-specific characteristics and statistical comparisons of regular and irregular past-tense forms on these characteristics.

We attempted to avoid highly ‘inconsistent’ regular past-tense forms such as *glided*, whose stems are phonologically highly similar to (i.e., rhyme with) the stem of irregulars (e.g., *ride-rode*, *slide-slid, hide hid*), as evidence suggests that such phonologically inconsistent regular inflected forms are more likely to be stored than ‘consistent’ regular past-tenses, whose stems are not phonologically similar to the stems of irregulars (the storage of inconsistent regulars is thought to prevent their irregularization by analogy to similar irregulars, e.g., *glide-glode* or *glide-glid*) (Ullman, 1993, 2001; Walenski, Mostofsky, & Ullman, 2007). Along the same lines we avoided doublet verbs, which have both regular and irregular past-tense forms (e.g. *dive-dove/dived*), since previous evidence also suggests that doublet regulars (e.g., *dived*) are likely to be stored rather than decomposed (Ullman, 1993, 2001).

Table 1
Item-level covariates included in our main analyses of experimental past-tense items.

	Regulars	Irregulars	Difference
N	64	64	
Frequency (contextual diversity)	2.66 (0.74)	2.71 (0.74)	$t(63) = 0.340$, $p = .735$
Word-form frequency	2.78 (0.81)	2.87 (0.82)	$t(63) = 0.542$, $p = .590$
Length in phonemes	4.20 (0.67)	3.72 (0.68)	$t(63) = 3.677$, $p < .001$
Number of coda consonants	1.63 (0.58)	1.23 (0.56)	$t(63) = 3.772$, $p < .001$
Stem-past phonological changes	1.00 (0.00)	1.28 (0.49)	$t(63) = 4.621$, $p < .001$
Quintet phonological similarity (whole)	4.62 (0.39)	4.46 (0.51)	$t(63) = 1.674$, $p = .099$
Quintet phonological similarity (onset)	1.63 (0.28)	1.63 (0.28)	$t(63) = 0.039$, $p = .969$
Quintet phonological similarity (rhyme)	2.87 (0.42)	2.56 (0.40)	$t(63) = 3.823$, $p < .001$
Quintet orthographic similarity (whole)	5.30 (0.56)	5.26 (0.60)	$t(63) = 0.451$, $p = .654$
Phonological neighborhood density	1.45 (0.33)	1.28 (0.28)	$t(62) = 3.340$, $p = .001$
Orthographic neighborhood density	1.72 (0.19)	1.55 (0.29)	$t(62) = 4.621$, $p < .001$
Frequency of phonological neighbors	7.94 (0.65)	8.47 (0.79)	$t(62) = 4.435$, $p < .001$
Frequency of orthographic neighbors	6.83 (0.53)	8.11 (0.56)	$t(62) = 13.705$, $p < .001$

Note: Means (and SDs in parentheses) are provided for each covariate, along with the results from paired *t*-test for statistical comparisons between the 64 regular and matched 64 irregular verbs.

If only past-tense forms were presented, participants might resort to strategies in the production n-back task. For example, they might not maintain the full regular inflected form in working memory, but rather just maintain the stem and then strategically add −*ed* during production. To minimize such possibilities, and to examine the prediction (see above) about processing costs incurred on other words maintained together with inflected forms in working memory, we included an additional 128 items as fillers (Table S2). These included 32 verb stems, 32 −*s*-suffixed forms, and 32 −*ing*-suffixed forms, all from regular verbs, as well as 16 regular and 16 irregular past-tenses of doublet verbs (e.g., *dove, dived*). The verbs used for the filler items were not the same as any of the verbs used for the past-tense experimental items (e.g., since *hired* was used as a past-tense experimental item, we did not include *hires* or *hiring* as filler items).

2.3. Design

We created a pseudo-randomized sequence of 192 items, with 64 slots for the regular or irregular past-tense forms, and 128 slots for the filler verbs. The pseudo-random order followed several constraints. First, no more than two stem, −*s*-suffixed, or −*ing*-suffixed filler items were adjacent. Second, no more than 5 past-tense forms (past-tense experimental items, and regular and irregular doublet filler items) occurred in a row. Third, equal numbers of each item type (stem, −*s*-suffixed, −*ing*-suffixed, regular doublet and irregular doublet filler items, and past-tense items) were present in the first and second halves of the list.

This sequence formed the basis of two lists of 192 items: one for the regular past-tense forms, and the other for the irregular past-tense forms.

The regulars and their pairwise matched irregulars were inserted into the same positions in the lists. For example, the regular past-tense form *stepped* was in position 85 in the regular past-tense list, while its pairwise matched irregular *swept* was in position 85 in the irregular past-tense list. Filler items occurred in the same position in each list (e.g., *grieving* was in position 58 in both lists). Each list was divided into first and second halves, that is, 96 item sub-lists (A, B), to allow participants to take a break during the task. In order to preserve homogeneity among the test conditions for all items, two additional filler verbs were added at the beginning of each sub-list, so that all test items were maintained in working memory under the same conditions (the first item in the list has no other items in memory accompanying it, and the second has only one other). Two filler items were also added at the end of each sub-list, so that every test item would be produced (the last two items on the list are not produced because the task ends before they become the target items). These additional 8 filler items, which were either stem, −*ing*- or −*s*-suffixed forms, were distinct from all other verbs in the lists, and were not analyzed.

In addition, because we wanted to ensure that any differences that were found between the regular and irregular experimental items were not due to properties of the verbs themselves, but rather could be attributed specifically to their past-tenses, we also created regular and irregular verb stem lists. The stems on these lists were the same verbs used for the past-tense experimental items. Due to time constraints on the length of the testing session, only a subset of the stems of the past-tense items were included: 30 regular/irregular pairs (Table S2), matched pairwise on stem frequency (log10 of SUBTLEX-US contextual diversity; regulars: $M = 3.110$; irregulars: $M = 2.980$; $t(29) = 1.099$, $p = .281$), without additional filler items (except for 2 at the beginning and end of each list, to allow for each stem to be a target).

The two sub-lists (A and B) of both the regular and irregular past-tense lists, and the regular and irregular verb stem lists, were presented in two distinct orders, counterbalanced across participants (roughly half of the male and half of the female participants saw each sequence). Sequence 1: Regular A, Irregular B, Irregular A, Regular B, Regular Stems, Irregular Stems. Sequence 2: Irregular A, Regular B, Regular A, Irregular B, Irregular Stems, Regular Stems. Because every past-tense sub-list was presented to each participant, every participant saw every past-tense item once, but every filler item twice, once in the first half of the sequence of sub-lists, and once in the second half (the same filler items were used in the Regular and Irregular A sub-lists, and in the Regular and Irregular B sub-lists). The repetition of filler items was addressed analytically, by including a covariate representing list half (first half, second half) of the experimental items (see Section 2.6).

2.4. Procedure

The verb forms were presented visually, one word at a time, with the *E*-Prime stimulus presentation program, on the center of a PC laptop computer screen. Each word was displayed for 1500 ms, followed immediately by a fixation cross for 2000 ms, followed in turn by a blank screen for one second. Presentation of the fixation cross initiated a software timer which was terminated by the oral response. Participants were instructed to respond each time the fixation cross appeared on the screen by saying out loud the third word back in the sequence. The experiment began with a practice sequence of ten nouns to become familiar with the task and to make sure the participants understood the instructions. This practice session was repeated until participants were comfortable with the task (usually after one or two sessions). They were given a break for a few minutes between each sub-list, and before the stem list. The experiment was digitally audio-recorded.

2.5. Response coding

A native-English speaking experimenter transcribed all responses during testing. These transcribed responses were checked against the recording by a different native-English speaker. A third native-English speaker reconciled the few cases where the two transcriptions did not match. Accuracy of the first response constituted the dependent variable. Incorrect responses included no response (participant did not say anything), or any first response that did not match the 3-back item in its entirety. Responses that began with a false start (e.g., *st…started*) or had somewhat reduced audibility (in whole or in part), but were clearly identifiable, were considered correct. These accounted for less than 0.2% of the data.

Response times (RTs) were analyzed as a separate dependent variable. However, because participants tend to respond in a rhythm in this task, we expected no regular/irregular differences in RTs. Response time analyses were performed over only the correct responses, using the criteria described above. Additionally, those responses for which the experimenter noted that the microphone had been triggered by a noise other than the subjects' first response were excluded from RT analysis (but were included in accuracy and error rate analyses). These comprised 20 responses for the past-tense experimental items, 65 for the filler items, and 3 from the verb stems. Due to experimenter error, RT data was not recorded for four subjects, requiring them to be excluded from RT analyses.

2.6. Covariates

We controlled for a number of potentially confounding variables by including them as covariates in our main past-tense analyses. Table 1 provides descriptive statistics of all item-level covariates for regular and irregular past-tense forms, as well as statistical comparisons for these covariates between regulars and irregulars.

First of all, we included inflected-form surface frequency as a covariate, as indexed by log10 of SUBTLEX-US contextual diversity (Brysbaert & New, 2009), to account for word frequency effects in working memory (Gregg, Freedman, & Smith, 1989; Hulme et al., 1997; Roodenrys et al., 2002). Note that even when variables do not show significant differences between conditions, including them as covariates can reduce the error term, and thus lead to more accurate results. Thus, we included frequency as a covariate even though the regular and irregular past-tense forms did not differ significantly on this measure (Table 1). Table 1 also includes log10 of word-form frequency (Brysbaert & New, 2009), simply the log of the number of times each word appears in the corpus (SUBTLEX-US), since one of our robustness analyses includes this frequency measure in lieu of contextual diversity.

Next, we addressed potential phonological differences between regular and irregular forms. Even though all inflected forms were monosyllabic, and regular and irregular forms were matched pairwise to the extent possible on phonological structure, we included three phonological covariates to minimize possible phonological confounds. First, we included the number of phonemes in the past-tense forms as a covariate to control for phonological length; indeed, the number of phonemes differed between the regular and irregular past-tenses (Table 1). Second, we included a covariate to account for the complexity of consonant clusters in the coda of past-tense forms, since regular past-tense forms tend to have more consonants in their codas than do irregulars, due to the presence of the −*ed* suffix (e.g., *walked*, *helped* vs. *dug*, *thought*). Indeed, the regular past-tense forms had a greater number of coda consonants than the irregular past-tense forms (Table 1). Third, we also quantified the number of stem-past phonological changes for all past-tense items in the study. Stem-past vowel substitutions, consonant substitutions, consonant additions, and consonant deletions were each considered to be a single change (Marcus et al., 1992). For example, *feed-fed*, *dig-dug*, *sing-sang*, *send-sent*, *dwell-dwelt*, and all regulars have one change, *keep-kept*, *feel-felt*, *teach-taught* and *buy-bought* have two changes, while *lose-lost* and *think-thought* each have three. As expected, the irregular verbs had significantly more stem-past changes than the regular verbs (Table 1).

We also included four covariates to account for phonological and orthographic similarity among items that were held together in working memory in our task, as such similarity has been shown to affect performance in working memory tasks (Baddeley, 1966; Conrad, 1964; Gathercole & Baddeley, 1993; Longoni et al., 1993). These similarity measures were calculated based on Levenshtein distance (Levenshtein, 1966), which is the minimum number of single-character edits (insertion, deletion, or substitution operations) required to turn one word into the other. The similarity covariates were computed on the basis of the similarity of the current item (1-back) and the four prior items (2-back, 3-back, 4-back, 5-back). Although our 3-back task required participants to say the third word back in the sequence, we also took into account the two items preceding the target word, since they may lead to proactive interference (Cowan, 2001; Engle, 2002; Neath & Surprenant, 2005; Öztekin & McElree, 2007; Wickens, Born, & Allen, 1963), and in keeping with the upper boundary of recent estimates of working memory limitations; i.e., 3 to 5 items (Cowan, 2010). For example, in the sequence *reeling, gleams, fanned, peeving, sneezes,* where the current item (1-back) is *sneezes* and the target is *fanned*, for the target we computed similarity among not only the target and the following two items, but also the two items preceding it (that is, over all five items: *reeling, gleams, fanned, peeving, sneezes*). Thus, similarity in each case was calculated over five items, hereafter referred to as quintet similarity. Although the phonological similarity of list items has long been found to lead to interference in immediate serial recall (Baddeley, 1966; Conrad, 1964; Longoni et al., 1993), it has been shown that rhyming words can actually facilitate list retrieval (Chow, Macnamara, & Conway, 2016; Gupta, Lipinski, & Aktunc, 2005). In order to consider these two potentially opposing effects, we calculated quintet phonological similarity in three different ways: based on whole words (e.g., /glimz/ and /f{nd/, which represent SAMPA phonetic transcription (Wells, 1997) for *gleams* and *fanned*), based on syllable onsets (e.g., /gl/ and /f/), and based on syllable rhymes (e.g., /imz/ and /{nd/). For rhymes, quintet phonological similarity was calculated on the basis of the stressed syllable; e.g., between /{nd/ and /iv/ for *fanned* and *peeving* (note that although all experimental and filler items had monosyllabic stems, and all experimental past-tense items were monosyllabic, many filler items were bisyllabic in their affixed forms, such as *peeving* or *sneezes*). Finally, because words are presented orthographically in the task, quintet orthographic similarity was also calculated, in this case based on the whole forms (since syllabic structure is phonological in nature, and the potentially opposing effects of syllabic subparts were already addressed by including the different phonological similarity covariates). We used Excel lambda code (https://gist.github.com/ncalm/715a0507805ff1df95cde2a04a9709be) to calculate Levenshtein distances for each pair of a quintet (10 pairs in total), and then took the mean of these to obtain the final quintet similarity measure (separately for the four quintet similarity measures). As shown in Table 1, quintet phonological similarity for rhymes was significantly greater for regulars than irregulars, while none of the other quintet similarity measures differed significantly between the two verb types.

We moreover considered phonological and orthographic neighborhood density, which have both been shown to affect performance in working memory tasks (Clarkson et al., 2017; Roodenrys et al., 2002; Saint-Aubin et al., 2023). We included as covariates the following four neighborhood measures from the *E*-Lexicon Project (Balota et al., 2007; Yarkoni, Balota, & Yap, 2008): the mean of the closest 20 Levenshtein-distance neighbors for both phonology and orthography (phonological and orthographic neighborhood density; see Table 1), and the mean log frequencies of these neighbors for both phonology and orthography (frequency of phonological neighbors and of orthographic neighbors; Table 1). The regular and irregular past-tenses differed significantly on all four neighborhood measures (Table 1). Note that for one regular (*spied*) and one irregular past-tense form (*wrung*), no neighborhood measures could be included as their values were not included in the E-Lexicon database.

In addition to these item-specific covariates, we included two task-related covariates. First, we included item order (coded from the very start of the n-back task), to account for autocorrelation between responses and for changes in performance over the course of the task (Baayen & Milin, 2010). Second, the repetition of the filler items between the first and second half of the experimental items might make the second half of the task easier, reducing overall working memory load (which could potentially change the predicted differences between regular and irregular past-tenses). Therefore, we also included list half (1 vs. 2) as a covariate (only for the experimental items, since the stem items were not presented with fillers).

2.7. Data analysis

The statistical analyses were conducted in R version 4.1.2 (R Core Team, 2013) using the lme4 package version 1.1.27.1 (Bates, Mächler, Bolker, & Walker, 2015). Response times were analyzed using linear mixed-effects models, and accuracy data were analyzed using generalized linear mixed-effects models with a binomial link function. Response times faster than 2 ms were removed to reduce skewness in the data (8.3% of responses; these extremely short response times likely occurred because the participants were instructed to wait to respond until a fixation cross appeared, which was always 1500 ms after the word appeared, allowing time for them to prepare their responses), followed by removal of outliers 3 standard deviations below or above the overall mean. Response times were then natural-log transformed, prior to analysis.

Here we lay out our statistical approach for the main accuracy and RT analyses on the experimental past-tense items; see just below for additional analyses. The fixed-effects structure of the initial models for our main analyses included verb type (regular, irregular) as the predictor of interest, and the twelve item-level (all covariates in Table 1 other than word-form frequency) and two task-related covariates explained above. Verb type was sum-coded as −0.5 and 0.5 for regulars and irregulars, respectively. All continuous predictors were scaled and centered.

We performed backwards elimination to identify the best-fitting models, using the drop1() function in the lme4 package, to successively eliminate covariates that did not improve model fit ($p \geq .100$) (Reifegerste, Meyer, Zwitserlood, & Ullman, 2021). Participants and items were included as random effects. The maximal random-effects structure was used (Barr, Levy, Scheepers, & Tily, 2013), with verb type entered as a by-subject random slope. Any convergence and/or singular fit errors were resolved using an optimizer (bobyqa) (Powell, 2009).

All additional analyses used a similar approach. The main analyses on the filler items and on the verb stems differed only in the covariates that were included in initial models; see Results Section 3.1. We also performed several exploratory analyses on the experimental past-tense items; Section 3.4. For simplicity and comparability, these exploratory analyses used the same model structure as the final model of the main analysis on the experimental past-tense items; that is, the covariates included in the exploratory analyses were the same as those included in the final model of that main analysis. We also carried out two sets of robustness analyses for all analyses, that is, for all main analyses and all exploratory analyses. The robustness analyses followed a similar approach as the main analyses and exploratory analyses on which they were based; see Sections 3.2 and 3.5.

All significant interactions were followed up with planned comparisons based on the estimated marginal means from the relevant models, using the emmeans package (Lenth, 2021). For error analyses, *t*-tests on the differences in error rates (out of all responses) between regular and irregular verbs were conducted both by subject (t_1) and by item (t_2).

3. Results

3.1. Main analyses

Regular past-tenses were produced significantly less accurately than irregular past-tenses (regular: $M = 64\%$; irregular: $M = 72\%$; $b = 0.3019$, $z = 3.175$, $p = .001$; see Fig. 1A and Table S3). (Here and below, key results are presented in the main text, with the full model output tables presented in Supplementary Material.) No differences were observed between the regular and irregular past-tense items in response times (untransformed RTs; regular: $M = 525$ ms; irregular: $M = 525$ ms; $b = 0.0059$, $t = 0.222$, $p = .825$; Table S4). As mentioned above (Section 2.5), this response time pattern was expected as a consequence of the rhythmic response pattern created by this task. Therefore, no further response time analyses were conducted (other than robustness analyses for this analysis on regular/irregular past-tense RTs). In addition, we predicted that if regulars but not irregulars are decomposed, items held in working memory *together* with regular past-tenses should be more adversely affected than those held together with irregulars, since working memory should be under a greater load while regular past-tenses are being maintained. Thus, filler items should yield more errors when held together with regular than irregular past-tense forms. Indeed, even though the filler items themselves were *identical* in the regular and irregular past-tense lists, and occurred in the *same positions* in the two lists, significantly more errors were observed on these forms when they occurred in the regular past-tense list than in the irregular past-tense list (accuracy on regular list: $M = 62\%$; on irregular list: $M = 65\%$; $b = 0.1812$, $z = 3.266$, $p = .001$; Fig. 1B and Table S5).

It might be argued that the irregular verbs were easier to maintain in working memory than the regular verbs due to semantic differences or some other inherent property of the verbs themselves, rather than due to the compositionality of the regular inflected forms. We anticipated this issue by giving the same participants the same production n-back task, but with only the bare stems of the experimental regular and irregular past-tense items; see Section 2.3. On the stem lists (which did not contain fillers), participants did not show the same pattern of worse performance on regulars than irregulars. Instead, they showed significantly higher accuracy on regulars than irregulars (regular: $M = 90\%$; irregular: $M = 87\%$; $b = -0.6660$, $z = -2.624$, $p = .009$; Fig. 1C and Table S6). The reasons for the higher performance on regular than irregular stems is unclear, and may warrant further investigation. Importantly, the results suggest that the irregular inflected forms were not easier than regular inflected forms for reasons unrelated to inflection.

3.2. Main analyses: robustness analyses

We performed two sets of robustness (i.e., sensitivity) analyses to examine the reliability of the patterns obtained in the main analyses. Crucially, the robustness analyses yielded the same pattern of results as the main analyses.

First, given that correlations between predictors can considerably alter regression estimates, and even cause sign reversals (Friedman & Wall, 2005; Veríssimo, Verhaeghen, Goldman, Weinstein, & Ullman, 2021; Wurm & Fisicaro, 2014), we inspected whether the observed regular/irregular verb type effects would remain without adjustment for any covariates. In other words, we refitted the original models without including any covariates. As in the main analyses, regular past-tenses were produced significantly less accurately than irregular past-tenses ($b = 0.4023$, $z = 4.158$, $p < .001$; Table S7), while no differences were observed between the two in response times ($b = 0.0097$, $t = 0.374$, $p = .710$; Table S8). Likewise, significantly more errors were observed on the fillers when they occurred in the regular past-tense list than in the irregular past-tense list ($b = 0.1765$, $z = 2.997$, $p = .003$; Table S9). Finally, as in the main analysis on the stems, significantly higher accuracy was observed on the regular than irregular verb stems when no covariates were included in the model ($b = -0.6002$, $z = -2.351$, $p = .019$; Table S10).

The second set of robustness analyses pertained to the use of a different frequency measure. Although contextual diversity, which was used in the main analyses, has advantages over other frequency measures (Brysbaert & New, 2009), it is arguably a less direct frequency

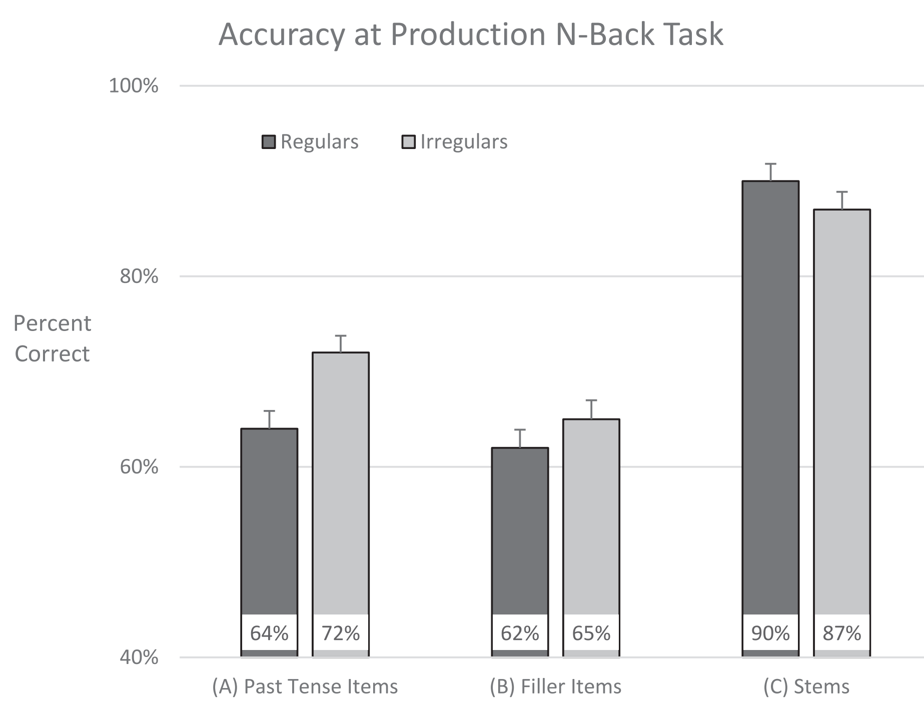

Fig. 1. Percent correct accuracy (unadjusted - that is, raw - means) on the production n-back task, for (A) the experimental regular and irregular past-tense items and (B) the filler items from the regular and irregular past-tense lists. (C) displays performance on the task for stems, from the separately-tested regular and irregular stem lists. Error bars reflect standard errors (based on participant means).

metric, and seems to be less commonly used, than word-form frequency, which is simply the number of times each word appears in the corpus. Therefore, the main analyses were repeated by replacing log10 of SUBTLEX-US contextual diversity with log10 of SUBTLEX-US word-form frequency in the initial models, followed by backwards elimination. The resulting final models yielded the same results as the main analyses. Regular past-tenses were produced significantly less accurately than irregular past-tenses ($b = 0.2667$, $z = 2.783$, $p = .005$; Table S11), with no regular/irregular difference in response times ($b = 0.0059$, $t = 0.222$, $p = .825$; Table S12). As in the main stem analysis, significantly higher accuracy was observed on regular than irregular stems ($b = -0.6834$, $z = -2.680$, $p = .007$; Table S13). Note that this alternate-frequency metric robustness analysis was not run for the fillers, since frequency is not a covariate in that analysis.

3.3. Error rates

Differences between regular and irregular past-tense forms in the error rates of specific types of incorrect responses (Table 2) further support the decomposition of regulars but not irregulars. First, if regulars but not irregulars are decomposed into stem and affix, there should be more bare-stem errors for regular than irregular past-tense forms (*walk* in lieu of *walked*, versus *dig* in lieu of *dug*). Indeed, there were significantly more (15 times as many) bare stem errors on the regular than irregular items (6.0% vs. 0.4% of responses; see Table 2); see next paragraph for an alternative account of this pattern.

Second, if regular but not irregular past-tense forms are maintained as stems and affixes in working memory alongside −*ing* and −*s* suffixed filler items, and then are recomposed when the response is produced, bare-stem affixation errors should be more likely for regulars than irregulars (e.g., *walking* or *walks* for *walked* should be more likely than *digging* or *digs* for *dug*). Indeed, there were significantly more (26 times as many) −*ing* or −*s* suffixed forms for regulars than for irregulars (5.2% vs. 0.2% of responses).

Third, if affixes are decomposed from their stems, then irregular past-tenses may also be occasionally affixed (*dugged*, *dugs* or *dugging*), while additional affixes should be much rarer for regulars (*walkeded*, *walkeds*, *walkeding*). Indeed, the rate of affixed irregular past-tenses was significantly higher than the rate of affixed regular past-tenses (0.6% vs. 0% of responses).

An additional possible confound is that regular past-tense forms usually contain their stems both phonologically and orthographically (e.g., *walked* contains *walk*), whereas this is not the case for irregulars (e.g., *dug* does not contain *dig*). Participants might therefore be more likely to process and maintain only the stem in the case of regulars than in the case of irregulars. For example, participants might stop reading when they reach the end of a real word, either by skipping over graphemes at the end of the word, or by not fully encoding the word phonologically. Indeed, as we have seen, stem errors (e.g., *walk* for *walked*) were more common for regular than irregular past-tense forms. However, if past-tense forms are being processed up to but not beyond the point where a phonological (real) word exists, this may also affect irregular past-tense forms that contain a phonological word (e.g., *told* contains *toll*), leading to containment errors on irregulars such as *toll*. Fifteen of the irregular items phonologically contained other words: told (toll); crept (crepe); lost (loss); sold (sole, soul); held (hell); dealt (dell); sought (saw); caught (caw); sank (sang); built (bill); felt (fell); stole (stow); wrote (row, roe); rode (row, roe); thought (thaw). Contrary to a phonological containment account, the rate of phonological containment errors for these irregulars was very low (0.72% of responses; which was not significantly greater than zero: by subjects $t_0(45) = 1.53$, $p = .134$, by items $t_0(14) = 2.09$, $p = .056$). Moreover, no containment errors were found for regulars (e.g., *chain* for *changed*) – other than bare stems, which were produced at a relatively high rate. For further evidence against a containment account, see the exploratory analyses below.

Table 2
Percentages of different types of errors for regular and irregular past-tense experimental items.

Errors	Examples or description	Regular	Irregular	Regular vs. irregular paired *t*-tests
Bare stem	walk, dig	6.0%	0.4%	$t_1(45) = 5.26$, $p < .001$; $t_2(63) = 7.93$, $p < .001$
Affixed stem	walking, digs	5.2%	0.2%	$t_1(45) = 8.90$, $p < .001$; $t_2(63) = 6.86$, $p < .001$
Affixed past-tense	walkeded, dugs, spenting	0.0%	0.6%	$t_1(45) = 3.44$, $p < .005$; $t_2(63) = 2.47$, $p = .020$
Other	Includes non-words, no responses, and inflected and uninflected forms of other words	24.8%	26.8%	$t_1(45) = 1.65$, $p = .110$; $t_2(63) = 1.34$, $p = .190$

Note. Percentages (unadjusted means) are computed out of all responses (not out of only incorrect responses).

3.4. Exploratory analyses

Here we consider additional explanations and factors that the stimuli were not initially designed to test. We emphasize that these analyses are exploratory, and leave further quantitative evaluation of these findings to carefully designed replication efforts.

First, as mentioned in the Introduction, it has been suggested that all past-tense forms are computed compositionally (Fruchter et al., 2013; Halle & Marantz, 1993; Stockall & Marantz, 2006). On this Distributed Morphology view, it is not only regulars that are affixed. Additionally, irregular past-tenses take either a null affix (e.g., in *dig* ➔ *dug* + −*0*) or an overt −*t* or −*d* affix (e.g., in *feel* ➔ *felt* + −*t*, *flee* ➔ *fle* + −*d*), over and above any phonological changes to the stem. If all affixed forms are predicted to be decomposed and thus tax working memory, one might expect no regular/irregular difference in our task, contrary to the observed pattern over all items. Nevertheless, it might be argued that, unlike overt affixes, null affixes should not tax working memory, and thus the inclusion of the 42 null affixed irregulars among the 64 irregular past-tense items in this study could have overshadowed poorer performance on the 22 overtly affixed items. We therefore split the items according to whether irregular past-tenses are or are not expected to have an overt affix, with each subset of irregular verbs compared against their paired (matched) regulars. The verb type (irregular vs. matched regular) by affix type (null vs. overt) interaction was not significant ($b = 0.0979$, $z = 0.598$, $p = .550$; Table S14), whereas the main effect of verb type remained significant, as in our main analysis ($b = 0.3089$, $z = 3.155$, $p = .002$; Table S14). Together with our main results, this argues against the view that irregulars as well as regulars are composed.

Second, it has been suggested by the Declarative/Procedural model and the Word and Rules approach it grew out of that although regulars *can* be assembled and disassembled by rule-governed processes, they can *also* be stored in memory, with the likelihood of their storage depending on a range of item, subject, and other factors (Johari, Walenski, Reifegerste, Ashrafi, & Ullman, 2019; Pinker, 1999; Pinker & Ullman, 2002; Ullman, 2004). One such factor is 'consistency'. As mentioned in the Section 2.2, inconsistent regular past-tense forms such as *glided*, whose stems are phonologically similar to (i.e., rhyme with) the stems of irregulars, are more likely to be stored than consistent regular past-tenses, whose stems are not phonologically similar to the stems of irregulars (Ullman, 1993, 2001; Walenski et al., 2007). The regular items in this study were selected to avoid high levels of phonological similarity with the stems of irregulars, that is, to avoid as much as possible regulars that rhyme with the stems of irregulars. Nevertheless, we could not avoid all such regulars: 25 of the 64 regular verbs had stems that rhymed with at least one irregular stem (number of irregular rhymes of the 25 verbs: $M = 2.32$; $SD = 1.11$). In contrast, the remaining 39 regulars had stems that did not rhyme with the stems of any irregulars.

Thus, in the next exploratory analysis, we examined regular/irregular accuracy differences for these 25 inconsistent 'rhyming' regulars, as compared to the 39 consistent 'non-rhyming' regulars. The analysis revealed a significant inconsistent/consistent by regular/irregular interaction ($b = -0.3436$, $z = -2.192$, $p = .028$; Table S15). Planned comparisons showed that the participants performed significantly worse on the consistent regulars ($M = 61\%$) than their matched irregulars ($M = 72\%$; $b = -0.4570$, $z = -4.030$, $p < .001$; Table S16), whereas no such difference was found between the inconsistent regulars ($M = 69\%$) and their matched irregulars ($M = 71\%$; $b = -0.1130$, $z = -0.872$, $p = .383$; Table S16). This pattern is in line with prior evidence that inconsistent regulars are more likely to be stored than consistent regulars (Ullman, 1993, 2001; Walenski et al., 2007), and further suggests that even a very small number of irregular rhymes can be sufficient to lead to the storage of regular inflected forms. The finding also argues against the possibility that worse performance on only consistent regulars than irregulars here, as well as the pattern of lower accuracy on regulars than irregulars in the main analyses, are due to phonological or orthographic containment errors (i.e., that regulars are worse than irregulars only because they phonologically or orthographically contain their stems; Section 3.3).

Third, two other factors have also been suggested to affect the storage vs. composition of regular past-tenses: their surface-form frequencies, with a greater likelihood of storage for higher frequency regulars (Alegre & Gordon, 1999; Baayen et al., 1997; Prado & Ullman, 2009) and the subject-level variable of sex, with females more likely than males to store regulars, likely due to a (estrogen-modulated) female advantage at declarative memory (Dye, Walenski, Prado, Mostofsky, & Ullman, 2013; Johari et al., 2019; Prado & Ullman, 2009; Reifegerste et al., 2020; Ullman, 2004, 2016; Ullman, Miranda, & Travers, 2008). Moreover, these factors appear to interact, such that high-frequency past-tenses of consistent regulars tend to be stored by women but not men, while low-frequency consistent regular past-tenses are not stored by either group (Prado & Ullman, 2009). We therefore examined the 39 consistent regulars for differences in composition as a function of frequency and sex. To emphasize the low/high frequency difference, we split the 39 verbs into low and high frequency terciles ($n = 13$ items per tercile) based on the SUBTLEX-US contextual diversity of their inflected forms.

Accordingly, in the third and final exploratory analysis, we examined regular/irregular accuracy differences in these low and high frequency regular forms, together with their matched irregulars, for males and females (Table 3). The three-way interaction between verb type (consistent regular vs. matched irregular), sex (male vs. female), and frequency (low vs. high) was significant ($b = -0.7607$, $z = -2.041$, $p = .041$; Table S17). To examine this interaction, we releveled the dataset separately for the low and high frequency items; that is, we changed the reference level of frequency by recoding the frequency levels (low and high), first with low frequency as 0 and high frequency as 1, in order to examine verb type by sex effects for low frequency items, and then vice versa, to examine effects in high frequency items (Uygun & Clahsen, 2020; Veríssimo et al., 2021). The two-way interaction between verb type and sex was significant for the high frequency verbs ($b = -0.5995$, $z = -1.991$, $p = .046$; Table S18), but not for the low frequency verbs ($b = 0.1612$, $z = 0.541$, $p = .588$, Table S18). Planned comparisons revealed that for low-frequency past tenses both males and females showed worse performance on regulars than irregulars (males: $b = -0.6070$, $z = -2.466$, $p = .014$; females: $b = -0.7680$, $z = -2.826$, $p = .005$, S19). In contrast, for high-frequency past tenses, males showed worse performance on regulars than irregulars ($b = -0.8960$, $z = -3.632$, $p < .001$), while females did not ($b = -0.2960$, $z = -1.106$, $p = .269$, S19). This pattern, which is consistent with prior evidence suggesting the same pattern, using a different methodological approach (Prado & Ullman, 2009), lends further support to the view espoused by the Declarative/Procedural model that the storage vs. composition of consistent regulars is modulated both by their surface inflected-form frequency and the subject-level factor of sex (Prado & Ullman, 2009; Ullman, 2004, 2016; Ullman et al., 2008). Moreover, along the lines discussed just above for the findings from consistent vs. inconsistent regulars, the results further argue against the possibility that the regular/irregular accuracy differences found in this study are due to containment errors.

Table 3
Mean (unadjusted) accuracy, with standard errors, for low and high frequency consistent regular past-tenses and their matched irregulars, for male and female participants.

Sex	Frequency	Regulars	Irregulars	Difference
Male ($n = 26$)	Low	60% (3.5%)	73% (3.8%)	13%
	High	62% (3.3%)	77% (2.9%)	15%
Female ($n = 19$)	Low	54% (4.3%)	71% (3.5%)	17%
	High	64% (3.3%)	68% (3.8%)	4%

3.5. Exploratory analyses: robustness analyses

As with the main analyses, we also carried out two sets of robustness analyses for each exploratory analysis, that is, without covariates, and with word-form frequency instead of contextual diversity. As in the robustness analyses with word-form frequency for the main analyses, here we replaced log10 of SUBTLEX-US contextual diversity with log10 of SUBTLEX-US word-form frequency in the initial models, followed by backwards elimination. In all cases, the robustness analyses yielded the same pattern of results as the exploratory analyses reported above.

First, we performed both robustness analyses on the exploratory analysis examining the effect of affix type (null vs. overt) (Fruchter et al., 2013; Halle & Marantz, 1993; Stockall & Marantz, 2006). In both robustness analyses, the verb type (irregulars vs. matched regulars) by affix type (null vs. overt) interaction was not significant (without covariates: $b = 0.0773$, $z = 0.434$, $p = .665$; Table S20; with word-form frequency: $b = 0.1624$, $z = 1.002$, $p = .316$; Table S21), while the main effect of verb type remained significant (without covariates: $b = 0.4151$, $z = 4.134$, $p < .001$; with word-form frequency: $b = 0.2654$, $z = 2.720$, $p = .007$).

Second, we ran the two robustness analyses on the exploratory analysis of regular inflected-form consistency (consistent vs. inconsistent). As in the exploratory analysis above, both robustness analyses revealed a significant interaction between verb type (regulars vs. matched irregulars) and consistency (consistent vs. inconsistent) (without covariates: $b = -0.4386$, $z = -2.607$, $p = .009$; Table S22; with word-form frequency: $b = -0.3292$, $z = -2.062$, $p = .039$, Table S23). Planned comparisons for these interactions revealed that the participants performed significantly worse on the consistent regulars than their matched irregulars (without covariates: $b = -0.5690$, $z = -4.983$, $p < .001$; Table S24; with word-form frequency: $b = -0.4310$, $z = -3.656$, $p < .001$; Table S25), while no such difference was found between the inconsistent regulars and their matched irregulars (without covariates: $b = -0.1300$, $z = -0.930$, $p = .352$; Table S24; with word-form frequency: $b = -0.1020$, $z = -0.782$, $p = .435$; Table S25).

Third, the two robustness analyses were performed on the exploratory analysis examining effects of sex and frequency. The robustness analyses yielded the same patterns of three-way interactions between sex (male vs. female), verb type (regular vs. matched irregular) and tercile (high vs. low) as the exploratory analysis above (without covariates: $b = -0.7422$, $z = -2.001$, $p = .045$; Table S26; with word-form frequency, which is also used to define the tercile split in this analysis: $b = -0.7169$, $z = -1.923$, $p = .054$, with this interaction approaching significance; Table S27). As before, to examine these interactions, we releveled the dataset separately for each of the two frequency terciles. As for the exploratory analysis above, in both robustness analyses the two-way interaction between verb type and sex was significant for the high frequency tercile (without covariates: $b = -0.6102$, $z = -2.019$, $p = .043$; Table S28; with word-form frequency: $b = -0.6642$, $z = -2.164$, $p = .030$; Table S29), but not for the low frequency tercile (without covariates: $b = 0.1532$, $z = 0.513$, $p = .608$; Table S28; with word-form frequency: $b = 0.0527$, $z = 0.174$, $p = .862$; Table S29).

As above, planned comparisons revealed that for low-frequency past tenses both males and females showed worse performance on regulars than irregulars in both the model without covariates (males: $b = -0.7880$, $z = -3.411$, $p = .001$; females: $b = -0.8450$, $z = -3.247$, $p = .001$; Table S30), and in the model with word-form frequency (males: $b = -0.7770$, $z = -3.173$, $p = .002$; females: $b = -0.8300$, $z = -3.045$, $p = .002$; Table S31). In contrast, for high-frequency past tenses, males showed worse performance on regulars than irregulars (without covariates: $b = -0.8830$, $z = -3.738$, $p < .001$; Table S30; with word-form frequency: $b = -0.9480$, $z = -3.608$, $p < .001$; Table S31), while females did not (without covariates: $b = -0.1980$, $z = -0.757$, $p = .449$; Table S30; with word-form frequency: $b = -0.2830$, $z = -0.994$, $p = .320$; Table S31).

4. Discussion

In summary, participants were given a 3-back production task that required them to maintain either regular or irregular past-tense experimental items in working memory, together with intermixed filler items (stem, –*s*-affixed, and –*ing*-affixed forms, and doublet regular/irregular past tenses). The participants made more errors on regular than irregular past-tenses. They also made more errors on the filler items maintained in working memory together with regulars than on the same filler items maintained together with irregulars. The regular past-tense items elicited more bare-stem errors (e.g., *walk*) and more stem affixation errors (e.g., *walking*, *walks*) than irregular past-tense items. In contrast, the irregulars yielded more errors involving the addition of affixes to already-inflected past-tense forms (e.g., *dugs*, *tolded*). In a control 3-back task requiring the participants to maintain in working memory the stems from the same regular and irregular verbs, no such regular disadvantage was found. Sensitivity analyses showed that the patterns were robust. The results from the experimental past-tense items are particularly striking in light of previous evidence indicating worse accuracy at irregular than regular past-tense forms in simple production tasks in healthy individuals, in English (Prado & Ullman, 2009; Ullman, 1993; Ullman et al., 1997) as well as in other languages (Nemeth et al., 2015; Walenski, Sosta, Cappa, & Ullman, 2009).

A number of confounding factors and explanations do not appear to account for the observed pattern. The examination of the verb stems, the consideration of potentially confounding covariates, as well as further analyses, suggest that the results are unlikely to be due to (1) intrinsic lexical, semantic, or syntactic properties of the verbs themselves, independent of their past-tense inflections; (2) inflected-form frequency differences between the regular and irregular items; (3) phonological length differences between the regular and irregular past-tense forms; (4) regular/irregular differences in past-tense consonant cluster complexity; (5) regular/irregular differences in the phonological faithfulness of stem-past mappings; (6) differences in phonological and orthographic similarity among items within the regular and irregular lists; (7) phonological and orthographic neighborhood density and frequency of neighbors; (8) effects due to item order or to the repetition of the filler items; (9) regular/irregular differences in phonological or orthographic containment errors; or (10) subject-level factors such as age or education, since regulars and irregulars are compared within subjects. Further, the findings do not appear to be explained by a theoretical perspective that posits that both regular and irregular forms are affixed (Fruchter et al., 2013; Halle & Marantz, 1993; Stockall & Marantz, 2006), or by the view that regular inflected forms are stored and retrieved as whole forms, on a retrieval-based working memory account (McElree, 2001, 2006).[4]

The results are consistent with our predictions, and appear to support the view that regular inflected forms are generally (de)composed, whereas irregular inflected forms are stored (Clahsen, 1999; Marslen-Wilson & Tyler, 1998; Pinker & Ullman, 2002; Ullman, 2004, 2016). To further investigate composition in English past-tense inflection, and to further test the ability of our methodological approach to distinguish composition and storage, we also performed a set of exploratory analyses to examine whether the regular disadvantage at the task might in fact not be found in item- or subject-level conditions where regulars are likely to be stored (Johari et al., 2019; Pinker, 1999; Pinker & Ullman, 2002; Ullman, 2004). Indeed, in line with previous findings using other methods, consistent but not inconsistent regulars yielded more errors than their matched irregulars; further, for consistent regulars the regular/irregular difference was found for low frequency verbs for both sexes, but for high frequency verbs only for males. Overall, the pattern of findings supports the perspective that regular inflected forms are generally (de)composed, while irregulars are stored, but that regulars can in fact be stored as a function of various item, subject, and other factors (Ullman, 2004, 2016). The results are consistent with, and indeed were largely predicted by, the perspective espoused by the Words and Rules theory and Declarative/Procedural model. Although to our knowledge the findings are not predicted by other models, such as connectionist accounts or Distributed Morphology, and do not seem to be broadly consistent with such accounts, we emphasize that we are not claiming that these accounts cannot explain the results. Future studies examining this issue seem warranted.

In addition to elucidating the storage vs. composition of inflectional morphology, in particular for English past tense, the study crucially demonstrates the utility of our novel empirical approach, which leverages working memory load limitations to test for compositionality in language. We emphasize that our approach is designed to test for compositionality rather than particular working memory accounts, and indeed the findings appear to be consistent with various working memory accounts, whether they invoke capacity limitations, attentional focus, and/or retrieval difficulties to explain load limitations (Baddeley, 2012; Baddeley et al., 2021; Cowan, 2001; Cowan et al., 2020; Crowder, 1993; McElree, 2001, 2006; Nairne, 1996, 2002; Neath & Surprenant, 2005). The findings appear to be consistent with the results of previous studies that examined the impact of working memory on inflectional or derivational complex forms in other languages (e.g., as compared to simple forms), primarily in children (Cohen-Mimran, Adwan-Mansour, & Sapir, 2013; Marton, Schwartz, Farkas, & Katsnelson, 2006; Service & Maury, 2015; Service & Tujulin, 2002). Our approach is simple and affordable (e.g., it does not require neuroimaging), increasing and

[4] It might be argued that the worse performance at regulars than irregulars could be explained if in the task the regular past-tense and affixed filler forms are stored and retrieved as whole forms, under a retrieval-based working memory account (McElree, 2001, 2006). On this view, such errors might arise from blends across different inflected forms retrieved from memory (e.g., in the sequence *walked*, *kicking*, *plays*, one might utter *walking* via a blend error between retrieved forms *walked* and *kicking*). Indeed, such affixed stem errors were more common for regulars than irregulars (Table 2). However, this error pattern is also expected if regulars are decomposed; see Section 3.3 in Results. Further, it is not clear how the storage and retrieval of whole affixed forms could easily explain the higher prevalence of affixed stem for irregulars (e.g., *spented*, *spents* or *spenting*) than regulars (e.g., *planneded*, *planneds*, *planneding*), nor the finding in our exploratory analyses that the disadvantage for regulars occurred specifically for consistent (not inconsistent) regulars, and for males for lower and higher frequency forms, but for females only for lower frequency forms. Thus overall, the evidence does not seem to suggest that regular inflected forms are stored and retrieved in this task, on any working memory account – even while the findings appear to be compatible with different working memory accounts, including a retrieval-based account, if regulars and other inflected forms are decomposed.

broadening its utility. Although here the use of the technique was restricted to regular and irregular English past-tense inflection, we believe that it is likely to be more widely applicable in the investigation of composition in language, including for syntax (e.g., in the examination of the compositionality of collocations, idioms, and other fixed phrases) as well as for derivational morphology and various other aspects of inflectional morphology. For example, the technique could be used to investigate inflectional systems that may be ambiguous with respect to composition, such as Italian verbs in the 2nd or 3rd conjugation with irregular stems but regular affixes (Say & Clahsen, 2002). The method should also be useful when employed parametrically. That is, complex linguistic forms that are hypothesized to be composed from a larger number of constituent parts (e.g., in agglutinative languages such as Turkish) should yield a larger number of errors with this method. Given debates about rule-governed composition vs. storage in other domains, such as reading or math (Coltheart, Curtis, Atkins, & Haller, 1993; Dehaene, 2011; Dehaene & Cohen, 1997; Evans & Ullman, 2016; Pritchard, Coltheart, Palethorpe, & Castles, 2012), the technique may also be useful in the study of cognition beyond language.

The study has a number of limitations, which in turn suggest additional lines of research. Due to testing time constraints, only a subset of the verb stems was used (Section 2.3). Moreover, since the stem forms were presented in isolation, and some of them are identical to forms of other word categories (e.g., *plan* is also a noun), it is possible that participants did not interpret all forms as verbs. Future studies could address this issue. Further, the findings here were based on a production n-back task. Although n-back tasks are well-studied and widely used, other working memory paradigms (e.g., recognition n-back tasks, or simple or complex span tasks) should also be employed to test our hypothesis that (de)composition particularly taxes working memory, especially since different working memory tasks may tap into somewhat different neurocognitive processes (Miller, Price, Okun, Montijo, & Bowers, 2009; Redick & Lindsey, 2013).

In conclusion, this study elucidates the storage vs. composition of inflectional morphology, and demonstrates the utility of a simple new methodological approach to test for composition in language. We believe that the technique may prove more broadly useful in the investigation of composition in language and perhaps other aspects of cognition.

CRediT authorship contribution statement

Michael T. Ullman: Funding acquisition, Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Supervision. **Talat Bulut:** Formal analysis, Writing – review & editing. **Matthew Walenski:** Investigation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare no competing interests.

Data availability

All the data reported in the present study, as well as accompanying documentation, are available for download via the Open Science Framework (https://osf.io/vfpgs/). The R code used for running the analyses reported in the paper can be made available on request.

Acknowledgements

This work was supported NIH R21HD087088 and NSF BCS1439290 and BCS1940980 to MTU, and TÜBİTAK (the Scientific and Technological Research Council of Turkey) grant 121S573 to TB. We thank Elizabeth Prado, Karsten Steinhauer, Kaori Ozawa, and João Veríssimo for help on various aspects of the study.

Appendix A. Supplementary data

Supplementary material to this article can be found online at https://doi.org/10.1016/j.cognition.2024.105875.

References

Adams, E. J., Nguyen, A. T., & Cowan, N. (2018). Theories of working memory: Differences in definition, degree of modularity, role of attention, and purpose. *Language, Speech, and Hearing Services in Schools, 49*(3), 340–355. https://doi.org/10.1044/2018_LSHSS-17-0114

Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language, 40*(1). https://doi.org/10.1006/jmla.1998.2607

Amon, M. J., & Bertenthal, B. I. (2018). Auditory versus visual stimulus effects on cognitive performance during the N-back task. In *Proceedings of the 40th annual meeting of the cognitive science society, CogSci 2018*.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *, 2. Psychology of learning and motivation - Advances in research and theory* (pp. 89–195). https://doi.org/10.1016/S0079-7421(08)60422-3

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language, 37*(1). https://doi.org/10.1006/jmla.1997.2509

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12–28. https://doi.org/10.21500/20112084.807

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A., Hitch, G., & Allen, R. (2021). A multicomponent model of working memory. In *Working memory* (pp. 10–43). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0002.

Baddeley, A. D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology, 18*(4), 362–365. https://doi.org/10.1080/14640746608400055

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The english lexicon project. *Behavior Research Methods*. https://doi.org/10.3758/BF03193014

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*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

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. https://doi.org/10.3758/BRM.41.4.977

Bulut, T. (2022). Neural correlates of morphological processing: An activation likelihood estimation meta-analysis. *Cortex, 151*, 49–69. https://doi.org/10.1016/j.cortex.2022.02.010

Bybee, J. (1995). Regular morphology and the lexicon. *Language & Cognitive Processes, 10*(5), 425–455. https://doi.org/10.1080/01690969508407111

Chow, M., Macnamara, B. N., & Conway, A. R. A. (2016). Phonological similarity in working memory span tasks. *Memory & Cognition, 44*(6), 937–949. https://doi.org/10.3758/s13421-016-0609-8

Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X99002228

Clarkson, L., Roodenrys, S., Miller, L. M., & Hulme, C. (2017). The phonological neighbourhood effect on short-term memory for order. *Memory, 25*(3). https://doi.org/10.1080/09658211.2016.1179330

Cohen-Mimran, R., Adwan-Mansour, J., & Sapir, S. (2013). The effect of morphological complexity on verbal working memory: Results from Arabic speaking children. *Journal of Psycholinguistic Research, 42*(3), 239–253. https://doi.org/10.1007/s10936-012-9200-z

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100*(4), 589–608. https://doi.org/10.1037/0033-295X.100.4.589

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology, 55*(1), 75–84. https://doi.org/10.1111/j.2044-8295.1964.tb00899.x

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Cowan, N., Morey, C. C., & Naveh-Benjamin, M. (2020). An embedded-processes approach to working memory. In *Working memory* (pp. 44–84). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0003.

Crowder, R. G. (1993). Short-term memory: Where do we stand? *Memory & Cognition, 21*(2), 142–145. https://doi.org/10.3758/BF03202725

Dehaene, S. (2011). *The number sense: How the mind creates mathematics, revised and updated edition. The number sense how the mind creates mathematics rev and updated ed*.

Dehaene, S., & Cohen, L. (1997). Cerebral pathways for calculation: Double dissociation between rote verbal and quantitative knowledge of arithmetic. *Cortex, 33*(2). https://doi.org/10.1016/S0010-9452(08)70002-9

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging, 4*(4). https://doi.org/10.1037/0882-7974.4.4.500

Dye, C. D., Walenski, M., Prado, E. L., Mostofsky, S., & Ullman, M. T. (2013). Children's computation of complex linguistic forms: A study of frequency and imageability effects. *PLoS One, 8*(9). https://doi.org/10.1371/journal.pone.0074683

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Engle, R. W., & Kane, M. J. (2003). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In *, 44. Psychology of learning and motivation - Advances in research and theory* (pp. 145–199). https://doi.org/10.1016/S0079-7421(03)44005-X

Evans, T. M., & Ullman, M. T. (2016). An extension of the procedural deficit hypothesis from developmental language disorders to mathematical disability. *Frontiers in Psychology, 7*(SEP), 1–9. https://doi.org/10.3389/fpsyg.2016.01318

Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *American Statistician, 59*(2). https://doi.org/10.1198/000313005X41337

Fruchter, J., Stockall, L., & Marantz, A. (2013). MEG masked priming evidence for form-based decomposition of irregular verbs. *Frontiers in Human Neuroscience, 7*(NOV), 1–16. https://doi.org/10.3389/fnhum.2013.00798

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Gregg, V. H., Freedman, C. M., & Smith, D. K. (1989). Word frequency, articulatory suppression and memory span. *British Journal of Psychology, 80*(3). https://doi.org/10.1111/j.2044-8295.1989.tb02326.x

Gupta, P., Lipinski, J., & Aktunc, E. (2005). Reexamining the phonological similarity effect in immediate serial recall: The roles of type of similarity, category cuing, and item recall. *Memory & Cognition, 33*(6), 1001–1016. https://doi.org/10.3758/BF03193208

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In *The view from building 20: Essays in honor of Sylvain Bromberger* (hereafter DM).

Hulme, C., Roodenrys, S., Brown, G. D. A., Schweickert, R., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 23*(5). https://doi.org/10.1037/0278-7393.23.5.1217

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences, 96*(13), 7592–7597. https://doi.org/10.1073/pnas.96.13.7592

Joanisse, M. F., & Seidenberg, M. S. (2005). Imaging the past: Neural activation in frontal and temporal regions during regular and irregular past-tense processing. *Cognitive, Affective, & Behavioral Neuroscience, 5*(3), 282–296. https://doi.org/10.3758/CABN.5.3.282

Johari, K., Walenski, M., Reifegerste, J., Ashrafi, F., & Ullman, M. T. (2019). Sex, dopamine, and hypokinesia: A study of inflectional morphology in Parkinson's disease. *Neuropsychology, 33*(4), 508–522. https://doi.org/10.1037/neu0000533

Kielar, A., Joanisse, M. F., & Hare, M. L. (2008). Priming English past tense verbs: Rules or statistics? *Journal of Memory and Language, 58*(2). https://doi.org/10.1016/j.jml.2007.10.002

Kuczaj, S. A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior, 16*(5). https://doi.org/10.1016/S0022-5371(77)80021-2

Lavric, A., Pizzagalli, D., Forstmeier, S., & Rippon, G. (2001). Mapping dissociations in verb morphology. *Trends in Cognitive Sciences*. https://doi.org/10.1016/S1364-6613(00)01703-4

Leminen, A., Smolka, E., Duñabeitia, J. A., & Pliatsikas, C. (2019). Morphological processing in the brain: The good (inflection), the bad (derivation) and the ugly (compounding). *Cortex, 116*, 4–44. https://doi.org/10.1016/j.cortex.2018.08.016

Lenth, R. V.. emmeans: Estimated marginal means, aka least-squares means. R package version 1.6.2-1. https://CRAN.R-project.org/package=emmeans.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8).

Longoni, A. M., Richardson, J. T. E., & Aiello, A. (1993). Articulatory rehearsal and phonological storage in working memory. *Memory & Cognition, 21*(1). https://doi.org/10.3758/BF03211160

Manelis, L., & Tharp, D. A. (1977). The processing of affixed words. *Memory & Cognition, 5*(6), 690–695. https://doi.org/10.3758/BF03197417

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development, 57*(4). https://doi.org/10.2307/1166115

Marslen-Wilson, W., & Tyler, L. K. (1998). Rules, representations, and the English past tense. *Trends in Cognitive Sciences, 2*(11), 428–435. https://doi.org/10.1016/S1364-6613(98)01239-X

Marslen-Wilson, W. D. (2007). Morphological processes in language comprehension. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (1st ed., pp. 495–524). Oxford University Press.

Marton, K., Schwartz, R. G., Farkas, L., & Katsnelson, V. (2006). Effect of sentence length and complexity on working memory performance in Hungarian children with specific language impairment (SLI): A cross-linguistic comparison. *International Journal of Language & Communication Disorders, 41*(6), 653–673. https://doi.org/10.1080/13682820500420418

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences, 6*(11), 465–472. https://doi.org/10.1016/S1364-6613(02)01993-9

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(3), 817–835. https://doi.org/10.1037/0278-7393.27.3.817

McElree, B. (2006). Accessing recent events. In *, 46. Psychology of learning and motivation - Advances in research and theory* (pp. 155–200). https://doi.org/10.1016/S0079-7421(06)46005-9

McNab, F., Zeidman, P., Rutledge, R. B., Smittenaar, P., Brown, H. R., Adams, R. A., & Dolan, R. J. (2015). Age-related changes in working memory and the ability to ignore distraction. *Proceedings of the National Academy of Sciences, 112*(20), 6515–6518. https://doi.org/10.1073/pnas.1504162112

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2). https://doi.org/10.1037/h0043158

Miller, K. M., Price, C. C., Okun, M. S., Montijo, H., & Bowers, D. (2009). Is the N-back task a valid neuropsychological measure for assessing working memory? *Archives of Clinical Neuropsychology, 24*(7). https://doi.org/10.1093/arclin/acp063

Monk, A. F., Jackson, D., Nielsen, D., Jefferies, E., & Olivier, P. (2011). N-backer: An auditory n-back task with automatic scoring of spoken responses. *Behavior Research Methods, 43*(3). https://doi.org/10.3758/s13428-011-0074-z

Morgan-Short, K., & Ullman, M. T. (2022). Declarative and procedural memory in second language learning. In *The Routledge handbook of second language acquisition and psycholinguistics*. https://doi.org/10.4324/9781003018872-30

Nairne, J. S. (1996). Short-term/working memory. In *Memory* (pp. 101–126). Elsevier. https://doi.org/10.1016/B978-012102570-0/50006-9.

Nairne, J. S. (2002). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology, 53*(1), 53–81. https://doi.org/10.1146/annurev.psych.53.100901.135131

Neath, I., & Surprenant, A. M. (2005). Mechanisms of memory. In *Handbook of cognition* (pp. 222–239). United Kingdom: SAGE Publications Ltd., 1 Oliver's Yard, 55 City Road, London EC1Y 1SP https://doi.org/10.4135/9781848608177.n9.

Nemeth, D., Janacsek, K., Turi, Z., Lukacs, A., Peckham, D., Szanka, S., … Ullman, M. T. (2015). The production of nominal and verbal inflection in an agglutinative language: Evidence from hungarian. *PLoS One, 10*(3). https://doi.org/10.1371/journal.pone.0119003

Nikolin, S., Tan, Y. Y., Schwaab, A., Moffa, A., Loo, C. K., & Martin, D. (2021). An investigation of working memory deficits in depression using the n-back task: A systematic review and meta-analysis. *Journal of Affective Disorders*. https://doi.org/10.1016/j.jad.2021.01.084

Oberauer, K. (2009). Design for a working memory. In *, 51. Psychology of learning and motivation - Advances in research and theory* (pp. 45–100). https://doi.org/10.1016/S0079-7421(09)51002-X

Oberauer, K. (2020). Towards a theory of working memory. In *Working memory*. https://doi.org/10.1093/oso/9780198842286.003.0005

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. In *, 25. Human brain mapping*. https://doi.org/10.1002/hbm.20131

Öztekin, I., Davachi, L., & McElree, B. (2010). Are representations in working memory distinct from representations in long-term memory? Neural evidence in support of a single store. *Psychological Science, 21*(8), 1123–1133. https://doi.org/10.1177/0956797610376651

Öztekin, I., & McElree, B. (2007). Proactive interference slows recognition by eliminating fast assessments of familiarity. *Journal of Memory and Language, 57*(1), 126–149. https://doi.org/10.1016/j.jml.2006.08.011

Pelegrina, S., Lechuga, M. T., García-Madruga, J. A., Elosúa, M. R., Macizo, P., Carreiras, M., … Bajo, M. T. (2015). Normative data on the n-back task for children and young adolescents. *Frontiers in Psychology, 6*(OCT). https://doi.org/10.3389/fpsyg.2015.01544

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

Pinker, S. (2006). Whatever happened to the past tense debate? In E. Baković, J. Ito, & J. J. McCarthy (Eds.), *Wondering at the natural fecundity of things: Essays in honor of Alan Prince*. Santa Cruz, CA: Linguistics Research Center, University of California, Santa Cruz.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6*(11), 456–463. https://doi.org/10.1016/S1364-6613(02)01990-3

Pliatsikas, C., Veríssimo, J., Babcock, L., Pullman, M. Y., Glei, D. A., Weinstein, M., … Ullman, M. T. (2019). Working memory in older adults declines with age, but is modulated by sex and education. *Quarterly Journal of Experimental Psychology, 72*(6), 1308–1327. https://doi.org/10.1177/1747021818791994

Powell, M. J. D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives*. Cambridge NA Report NA2009/06. Cambridge: University of Cambridge.

Prado, E. L., & Ullman, M. T. (2009). Can imageability help us draw the line between storage and composition? *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*(4). https://doi.org/10.1037/a0015286

Pritchard, S. C., Coltheart, M., Palethorpe, S., & Castles, A. (2012). Nonword reading: Comparing dual-route cascaded and connectionist dual-process models with human data. *Journal of Experimental Psychology: Human Perception and Performance, 38*(5). https://doi.org/10.1037/a0026703

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Redick, T. S., & Lindsey, D. R. B. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic Bulletin and Review, 20*(6). https://doi.org/10.3758/s13423-013-0453-9

Reifegerste, J., Estabrooke, I. V., Russell, L. E., Veríssimo, J., Johari, K., Wilmarth, B., … Ullman, M. T. (2020). Can sex influence the neurocognition of language? Evidence from Parkinson's disease. *Neuropsychologia, 148*. https://doi.org/10.1016/j.neuropsychologia.2020.107633

Reifegerste, J., Meyer, A. S., Zwitserlood, P., & Ullman, M. T. (2021). Aging affects steaks more than knives: Evidence that the processing of words related to motor skills is relatively spared in aging. *Brain and Language, 218*(March), Article 104941. https://doi.org/10.1016/j.bandl.2021.104941

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6). https://doi.org/10.1037//0278-7393.28.6.1019

Rueckl, J. G., Mikolinski, M., Raveh, M., Miner, C. S., & Mars, F. (1997). Morphological priming, fragment completion, and connectionist networks. *Journal of Memory and Language, 36*(3). https://doi.org/10.1006/jmla.1996.2489

Saint-Aubin, J., Poirier, M., Yearsley, J. M., Robichaud, J.-M., & Guitard, D. (2023). Modeling verbal short-term memory: A walk around the neighborhood. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*(2), 198–215. https://doi.org/10.1037/xlm0001226

Say, T., & Clahsen, H. (2002). *Words, rules and stems in the Italian mental lexicon*. https://doi.org/10.1007/978-94-010-0355-1_4

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2014). A task is a task is a task: Putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in Psychology, 5*(DEC). https://doi.org/10.3389/fpsyg.2014.01475

Seidenberg, M. S., & Plaut, D. C. (2014). Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science, 38*(6), 1190–1228. https://doi.org/10.1111/cogs.12147

Service, E., & Maury, S. (2015). Differential recall of derived and inflected word forms in working memory: Examining the role of morphological information in simple and complex working memory tasks. *Frontiers in Human Neuroscience, 8*(JAN), 1–16. https://doi.org/10.3389/fnhum.2014.01064

Service, E., & Tujulin, A.-M. (2002). Recall of morphologically complex forms is affected by memory task but not dyslexia. *Brain and Language, 81*(1–3), 42–54. https://doi.org/10.1006/brln.2001.2505

Silva, R., & Clahsen, H. (2008). Morphologically complex words in L1 and L2 processing: Evidence from masked priming experiments in English. *Bilingualism, 11*(2). https://doi.org/10.1017/S1366728908003404

Stockall, L., & Marantz, A. (2006). A single route, full decomposition model of morphological complexity. *The Mental Lexicon, 1*(1), 85–123. https://doi.org/10.1075/ml.1.1.07sto

Ullman, M. T. (1993). *The computation of inflectional morphology. Cambridge, MA*.

Ullman, M. T. (2001). The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research, 30*(1). https://doi.org/10.1023/A:1005204207369

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition, 92*(1–2), 231–270. https://doi.org/10.1016/j.cognition.2003.10.008

Ullman, M. T. (2016). The declarative/procedural model: A neurobiological model of language learning, knowledge, and use. In *Neurobiology of language* (pp. 953–968). Elsevier. https://doi.org/10.1016/B978-0-12-407794-2.00076-6.

Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., Growdon, J. H., Koroshetz, W. J., & Pinker, S. (1997). A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience, 9*(2), 266–276. https://doi.org/10.1162/jocn.1997.9.2.266

Ullman, M. T., Miranda, R. A., & Travers, M. L. (2008). Sex differences in the neurocognition of language. In *Sex differences in the brain: From genes to behavior*. Oxford University Press.

Uygun, S., & Clahsen, H. (2020). Morphological processing in heritage speakers: A masked priming study on the Turkish aorist. *Bilingualism*. https://doi.org/10.1017/S1366728920000577

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language, 55*(2), 157–166. https://doi.org/10.1016/j.jml.2006.03.007

Veríssimo, J., Verhaeghen, P., Goldman, N., Weinstein, M., & Ullman, M. T. (2021). Evidence that ageing yields improvements as well as declines across attention and executive functions. *Nature Human Behaviour, 6*(1), 97–110. https://doi.org/10.1038/s41562-021-01169-7

Walenski, M., Mostofsky, S. H., & Ullman, M. T. (2007). Speeded processing of grammar and tool knowledge in Tourette's syndrome. *Neuropsychologia, 45*(11), 2447–2460. https://doi.org/10.1016/j.neuropsychologia.2007.04.001

Walenski, M., Sosta, K., Cappa, S., & Ullman, M. T. (2009). Deficits on irregular verbal morphology in Italian-speaking Alzheimer's disease patients. *Neuropsychologia, 47*(5). https://doi.org/10.1016/j.neuropsychologia.2008.12.038

Wells, J. C. (1997). SAMPA computer readable phonetic alphabet. In D. Gibbon, R. Moore, & R. Winski (Eds.), *Handbook of standards and resources for spoken language systems*. Berlin and New York: Mouton de Gruyter. https://doi.org/10.1515/9783110804508-008.

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 2*(5–6), 440–445. https://doi.org/10.1016/S0022-5371(63)80045-6

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology, 4*(JUL). https://doi.org/10.3389/fpsyg.2013.00433

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language, 72*(1). https://doi.org/10.1016/j.jml.2013.12.003

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971