Cognition 269 (2026) 106387

Contents lists available at ScienceDirect

Cognition

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

Serial processing of two words becomes parallel when they combine to form a known compound word

Amritha Anupindi [1], Liana R. Eisler [1], Mariam Latif, Vassiki S. Chauhan, Alex L. White [*]

Department of Neuroscience & BehaviorBarnard College, Columbia University, 3009 Broadway, New York, NY 10011, United States of America

ARTICLE INFO

Keywords:
Word recognition
Visual attention
Divided attention
Language
Parallel processing

ABSTRACT

According to one model of reading, words are recognized one at a time with serial shifts of focused attention. This serial strategy would be required if, as some prior research suggests, there is a bottleneck in the brain that cannot process two words simultaneously. Consistent with this serial model, we first show that participants can judge the lexical status of only one of two unrelated words that are flashed briefly above and below the point of gaze fixation and then masked. We then investigate whether two words that together compose an existing compound word (e.g., bottle + neck) can nonetheless be processed in parallel. The results demonstrate that indeed, under the same conditions in which two unrelated words cannot be recognized simultaneously, accuracy for recognizing either or both of two words that form a compound exceeds the prediction of the serial model. This result complicates theories of a serial bottleneck in word recognition, especially in the context of natural reading. We propose a model that begins with parallel orthographic processing, initially serial lexical activation, and then interactive activations that can amplify the representations of two words that form a known compound.

1. Introduction

To understand how skilled readers read, we must determine how much information they can process at each glance at a page. Information intake is first constrained by limits inherent to the visual system: words are not legible in peripheral vision. Consequently, to read this page, you must process the text in small chunks by precisely focusing your attention and moving your eyes down each line (Rayner, 2009). Nonetheless, the "span" of legibility in the central visual field is often wide enough that it could allow multiple words to be read during a single gaze fixation (Legge et al., 2001; Veldre et al., 2023; Yeatman & White, 2021).

Does that mean that readers extract linguistic information from multiple words simultaneously? This question has been heavily debated in the study of natural reading (Snell & Grainger, 2019a, 2019b; White, Boynton, & Yeatman, 2019). Some models of reading allow attention to be distributed such that multiple words are processed in parallel during each fixation (Engbert et al., 2005; Snell et al., 2018). Other models assume that readers process one word at a time, via serial shifts of attention from word to word (Reichle et al., 2006).

To gain traction on this question, other lines of research have relied on controlled experiments to investigate how much information readers can extract from multiple words during a brief interval (e.g., Fallon & Pylkkänen, 2024; Snell & Grainger, 2019a, 2019b; White et al., 2025). One such series of experiments suggested that it is not possible to recognize two unrelated English words at exactly the same time (White et al., 2018, 2020). The study we report in this article is based on those experiments, which used the design illustrated in Fig. 1A. Two randomly selected words were flashed briefly on either side of the central fixation point and then replaced by post-masks. In a pre-test, the time between the words and the post-masks was set to each individual's *threshold* such that they had just barely enough time to recognize one word with focused attention. Then in each trial of the main experiment, there were two conditions that used the same stimulus timing: on *focused* attention trials, participants were pre-cued to attend covertly to one side in order to semantically categorize just one word. On *divided* attention trials, they were pre-cued to attend to both sides in order to categorize both words independently.

Accuracy in these conditions can be compared to quantitative models of processing capacity, which are plotted on an Attention Operating Characteristic (AOC; Sperling & Melchner, 1978), as illustrated in

* Corresponding author at: Department of Neuroscience & BehaviorBarnard College, Columbia University, 3009 Broadway, New York, NY 10011 United States of America.
E-mail address: alwhite@barnard.edu (A.L. White).
[1] co-first authors

https://doi.org/10.1016/j.cognition.2025.106387
Received 27 November 2024; Received in revised form 17 November 2025; Accepted 27 November 2025

Fig. 1B. The AOC plots accuracy for words above fixation against accuracy for words below fixation. The two focused-attention accuracies are pinned to their respective axes. Accuracy with divided attention forms a single point (open circle) in that 2-D space, alongside the predictions of three quantitative models of capacity limits (Bonnel & Prinzmetal, 1998; Scharff et al., 2011; Sperling & Melchner, 1978; White et al., 2018). Note that these models differ from the more complex models of attentional control during natural reading that were mentioned above. They are more general and provide relatively simple benchmarks for how well two stimuli can be processed at once. The three models are:

1. *Independent parallel model*: Two stimuli can be processed simultaneously just as well as single stimuli with focused attention. Thus, there is no cost to dividing attention. In the AOC, this model predicts that the divided attention point will fall at the intersection of the dashed lines. This model is also called an *unlimited-capacity* parallel model.
2. *Fixed-capacity parallel model*: The perceptual system can extract a fixed amount of information from the entirety of the stimulus display on each trial. In the focused attention condition, processing resources are devoted to just one stimulus. In the divided attention condition, resources are shared between two stimuli. Thus, there is a moderate accuracy deficit in the divided attention condition. This model traces the black curve in the AOC.
3. *All-or-none serial model*: Stimuli are processed one at a time. Because time is limited, participants can recognize only one stimulus in the divided attention condition, with equal accuracy as in the focused attention condition. When they try to continue to process the second stimulus, it has been replaced by the post-masks, which removed any sensory memory trace. The participant therefore must make a random guess about one of the two stimuli (hence, processing of each stimulus is "all or none"). This model traces out the diagonal line in the AOC. Where the accuracy point falls along that line depends on the proportion of trials in which the top stimulus gets processed.

In previous studies that used this paradigm with word recognition tasks, accuracy in the divided attention condition dropped so far that it supported the "all-or-none" serial model: participants could recognize one of the two words on each trial but had to make a random guess about the other (White et al., 2018, 2020). The post-masks were key to this result: they prevented participants from being able to process one word and then shift their attention, serially, to the other word.

Converging evidence for the serial model is provided by several other studies that used the same approach to limit the time available to process pairs of words, but with variations on the types of stimuli and task requirements (Brothers, 2022; Campbell et al., 2024; Johnson et al., 2022). These behavioral results were also supported by fMRI evidence that the "visual word form area" in the left ventral temporal cortex responds to only one word at a time (White, Palmer, et al., 2019).

In those prior studies, the words presented simultaneously were randomly selected and unrelated (for an exception with sentences, see Brothers, 2022). In natural text, however, neighboring words are systematically related to each other in several ways. A reader's prior knowledge of how words combine could facilitate parallel processing. Specifically, it is possible that two words can be processed in parallel if they are known by the reader to frequently co-occur or to compose a larger linguistic unit (Yoo & Joo, 2025).

We adapted the divided attention paradigm described above to test whether the serial model is violated when two words can combine to form another word. To do so, we introduced pairs of words that can form existing compound words when put together (e.g., stair + case). The compounds or combined words that we used are commonly written in English without spaces between the constituents and thus are distinguished from the much larger set of word pairs that *could* productively compose a new meaning (e.g., dragon + school). Our question is whether readers can process two words in parallel if they compose a compound word together, even under display conditions that prevent simultaneous recognition of two unrelated words that do not compose a compound.

On the one hand, the "serial bottleneck" in the reading circuitry could be so strict that relations between words (such as those that form a compound) are not detected until each has been identified serially. If words are always processed one at a time, accuracy in the divided-attention paradigm (Fig. 1) should not depend at all on how words are paired.

On the other hand, parallel processing could be facilitated when two words combine to form a known compound word. Theoretically, such an effect could be supported by several mechanisms. First, one constituent word could activate a single lexical unit dedicated to the compound word that feeds back to amplify both constituents at a sub-lexical level. Indeed, the "multi-constituent unit hypothesis" proposes that the linguistic units that are sequentially processed during reading might span multiple words that form familiar phrases or compounds (Zang, 2019). According to this proposal, frequent word combinations become "lexicalized" through experience, which allows their constituents to be processed in parallel. This hypothesis has some support from fixation patterns observed during reading (Cutter et al., 2014; Yu et al., 2016; Zang et al., 2024).

But even without a dedicated lexical unit for a compound, there could be facilitatory interactions between representations of the individual constituent words. More specifically, such facilitatory

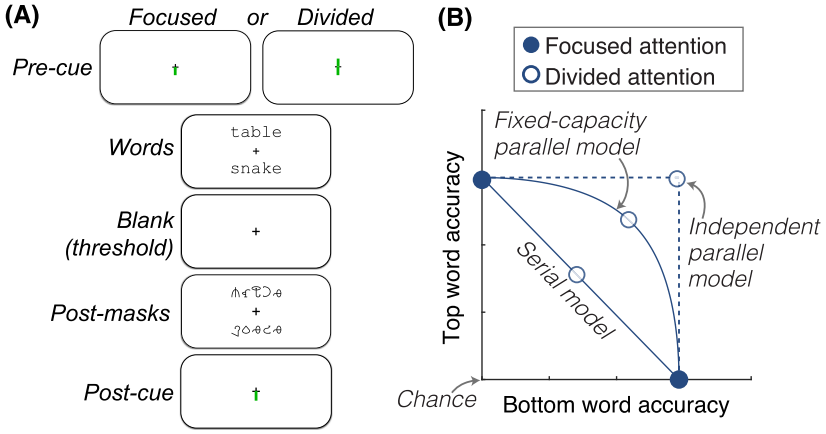

Fig. 1. **(A)** Diagram of the focused-vs-divided attention paradigm, as used by White et al. (2020) and in the present study. **(B)** A hypothetical attention operating characteristic (AOC) with the predictions of three models for accuracy in the divided attention condition.

connections between lexical representations could be driven by their *semantic similarity*, as occurs in the effects of priming (Meyer & Schvaneveldt, 1971). Alternatively, facilitatory connections could be driven by the statistics of co-occurrence of the words, or by prior knowledge about their compatibility as sub-units (or morphemes) of a larger lexical unit. For Korean words, accuracy in divided attention tasks (similar to Fig. 1) increases as a function of both co-occurrence statistics and the semantic relatedness between two words (Yoo & Joo, 2025). These alternative theoretical explanations are described more thoroughly in the General Discussion (see Fig. 6**).** Any of them could raise challenges for the simple serial model.

Our use of compound words relates to the more general study of morphologically complex words, of which they are a subset. Substantial research has investigated morphological decomposition, that is, a stage of processing in which the constituent morphemes (stems, in the case of compounds) are identified in the process of whole-word identification (Andrews, 1986; Fiorentino & Poeppel, 2007; Ji et al., 2011; Libben et al., 2020; Rastle et al., 2000; Sandra, 1990; Taft & Forster, 1976). The stimulus displays in the experiment reported here encourage decomposition because the two constituent words are spaced apart on opposite sides of fixation and are thus relatively easy to process separately. The distinct question we pursue is whether prior knowledge of how two words combine to form a compound facilitates simultaneous processing of both constituents. If that does occur, then the simple serial model proposed before needs to be revised or rejected.

In Experiment 1, participants first engaged in a divided-attention task like those described above, which we refer to as the "dual lexical decision task." Replicating prior results, divided-attention accuracy supported the serial model: participants could recognize only one word per trial. Then the same participants performed a second task with nearly identical stimulus sequences, except the task on each trial was to report whether the two words formed a compound word. If only one of the two words could be processed, then accuracy for this "compound word judgment" should be at chance. But as we report below, accuracy was substantially higher than chance.

In Experiment 2, we directly compared processing capacity for word pairs that either did or did not form compounds, using a novel task in which participants report one word by typing it in. Two words were flashed briefly and followed by post-masks, and participants were then prompted to type in the word they had seen at one post-cued location. Most word pairs were unrelated, but unbeknownst to the participants, some formed existing compound words (e.g., water + fall). Unlike in Experiment 1, participants in Experiment 2 were not searching for compound words and were motivated to process each of the two words independently on each trial. Nonetheless, task performance was affected by whether or not the two words happened to form a compound, defying the apparent one-word-at-a-time processing bottleneck.

These experiments leverage precise control of stimulus timing and measurements of word recognition accuracy to assess processing capacity, but they do not study natural reading. Rather than ask what readers do, we ask what readers *can* do. Both are important questions that must be linked, but the latter offers unique experimental levers to reveal the underlying mechanisms. Our experiments differ from natural reading in part because we positioned the words just above and below the point of gaze fixation. This is not the format in which words are typically read, nor is it a typical way to present the two constituents of a compound word. But we chose this arrangement for two reasons: first, it allows both words to be close to fixation and equally legible; second, this arrangement has provided strong evidence for serial processing of two unrelated words (White et al., 2020). Our experiments therefore provide a strict test of the hypothesis that the putative serial bottleneck in word recognition can be circumvented when two words compose an existing compound. Such a result would indicate that the efficiency with which written words can be processed depends on the linguistic context in which they are viewed.

2. Experiment 1

2.1. Methods

2.1.1. Participants

10 volunteers (6 female, 2 male, 2 non-binary, ages 18–23 years) with normal or corrected-to-normal visual acuity participated in Experiment 1 in exchange for fixed monetary payment ($20/h). Each participant gave informed consent in accordance with the Declaration of Helsinki and the Barnard College Institutional Review Board. All participants were naïve to the purposes of the experiment and reported learning English before the age of 5. On the composite TOWRE-II Test of Word Reading Efficiency (Torgesen et al., 1999), the mean score was 114 (SEM = 3.7). Nine of ten participants scored above the norm of 100. We used the TOWRE test to ensure that all our participants were proficient readers and to screen out dyslexia.

In the tradition of visual psychophysics, this study used a relatively small number of participants but many trials and multiple testing sessions per participant, with individually calibrated stimuli. The sample size was chosen ahead of data collection by a power analysis of a previous experiment that had a similar design (White et al., 2020, Experiment 2). We estimated the number of participants needed to distinguish between the fixed-capacity parallel and serial models, with 90 % power. In White et al. (2020), the mean difference between model predictions (at the points where they were closest to the empirical data) was 0.08. Empirically, the standard deviation of the difference in accuracy from the fixed-capacity parallel model's prediction was 0.06. Using that mean (0.08) and standard deviation (0.06), we simulated 10,000 paired *t*-tests for varying sample sizes, each representing an attempt to reject the fixed-capacity parallel model in favor of the serial model, as was done before. The minimum for 90 % power was 9 participants. We rounded up to 10 to be conservative and consistent with four previous experiments (two in White et al., 2018 and three in White et al., 2020), which all found similar effect sizes. Each participant completed two tasks: the dual lexical decision task, and then a "compound word judgment task."

2.2. Experiment 1a: dual lexical decision task with unrelated words

2.2.1. Equipment and stimuli

We used custom MATLAB software (MathWorks) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) to present stimuli on a ViewPixx 3D screen (VPixx Technologies) with a 120 Hz refresh rate and 1920 × 1080-pixel resolution. The background brightness was set to the screen's maximum (100 cd/m^2). The stimuli consisted of a black fixation mark and black letter strings. The fixation mark was made of a black cross 0.38 degrees of visual angle (°) wide, with a 0.1° white dot at its center, and a thin black ring around it (0.38° diameter). The letter strings were written in Courier New font. The height of the letter "x" was 0.6°.

The stimulus set was composed of 820 real English words and 785 pronounceable pseudowords.[2] Both categories were divided equally into strings of three, four, five, and six letters long. The real words came from all syntactic categories, ranging in lexical frequency from 2.4 to 6.3 Zipf (mean = 4.5, SD = 0.71). Zipf is a standardized measure of word frequency calculated through an adjustment of the log frequency per million words in the SUBTLEX-US corpus (Brysbaert & New, 2009; van Heuven et al., 2014). The pseudowords were generated from the MCWord database (Medler & Binder, 2005) to have constrained trigram statistics, which makes them pronounceable and similar to real words. **Tables S2 and S3** at the end of the *Supplementary Materials* list all the

[2] The slight imbalance in the numbers of items in the two categories (they differ by 4 %) could, theoretically, bias participants to report "real." However, as explained below, we compute accuracy with the unbiased estimator of area under the ROC curve.

words used in the experiment.

The post-masks were strings of non-letter characters drawn in a false font, BACS-2 Serif, which has several visual features matched to Courier New (Vidal et al., 2017). Each post-mask was in fact a real word from the stimulus set, matched in length to the word that preceded it, but presented in the illegible false font.

2.2.2. Trial sequence

Fig. 2A illustrates an example trial. Each trial began with participants focusing on the fixation mark for a minimum of 200 ms, followed by a 500 ms pre-cue. On divided attention trials, two green pre-cue lines, 0.16° long, appeared superimposed on the upper and lower arms of the fixation cross. On focused attention trials, only one green pre-cue line indicated the side (above or below) to be post-cued. After a 500 ms blank interval containing only the fixation cross, two letter strings appeared for 33 ms, positioned above and below fixation and centered at 1.5° eccentricity, matching the stimulus positions used in the lexical decision task from Experiment 2 of White et al. (2020). Each letter string had an independent 50 % chance of being a real word or a pseudoword. The only constraints were that the letter strings on either side of fixation could not be identical, and neither string could have appeared in the previous trial.

The words were followed by an inter-stimulus interval (ISI), containing only the fixation mark. The ISI duration was set to each participant's threshold in the focused attention condition (mean ISI = 8 ms, SEM = 2.6 ms, range = [0 25]), as determined by a pre-test staircase (see below). After the ISI, two post-masks were presented for 250 ms, at the same locations and with the same number of letters as the preceding words. Following another 100 ms blank interval (not shown in Fig. 2A), a post-cue appeared to indicate which word should be judged. The post-cue was a green line like the focused pre-cue. 500 ms after the post-cue appearance, a 25 ms click sound was played, which prompted the participant to press a key to report the lexical category (real word or pseudoword) of the letter string on the side indicated by the post-cue. Keypresses before the click were not recorded. On focused attention trials, there was only one post-cue and one response (to the one word that was pre-cued). On divided attention trials, participants were asked to judge both words in a random order. After the first post-cue, click, and keypress response, the post-cue reversed to the other side. 300 ms later, a second click prompted the second response.

Participants pressed one of four keys (*a, s, d,* or *f*) with their left hand for words in the top location, or one of four keys (*m, <, >,* or *?*) with their right hand for words in the bottom location. For each hand, from left to right, the keys indicated "sure pseudoword," "guess pseudoword," "guess real word," and "sure real word."

Feedback about response accuracy (regardless of confidence level) was given with 100 ms auditory beeps (high pitch for correct, low pitch for incorrect). On focused attention trials, the feedback beep was played 350 ms after the participant's keypress. On divided attention trials, the two correct or incorrect feedback tones were played after both responses were recorded. After a 500 ms inter-trial interval (ITI), the next trial began.

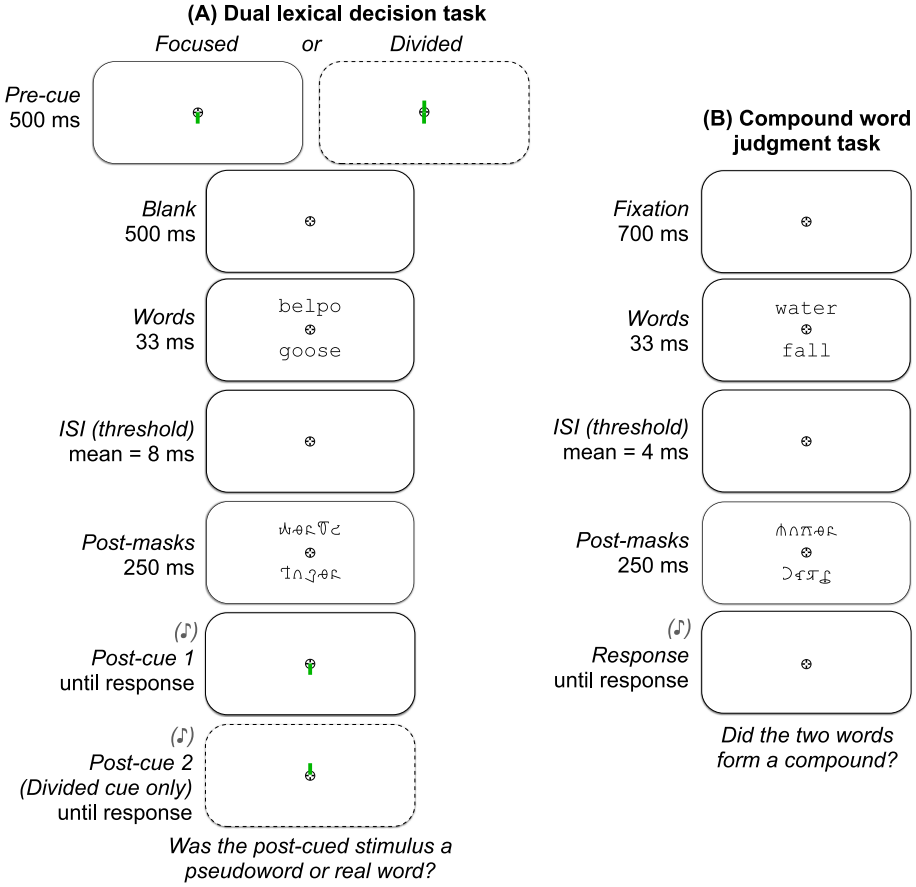

Fig. 2. Example trial sequences in Experiments 1a and 1b. **(A)** Dual lexical decision task (Experiment 1a). The example shows either a focused attention trial with the target on the bottom side, or a divided attention trial (stimulus frames with dashed outlines). The focused cues were equally likely to point to the top side or the bottom side. Not shown is a 100 ms blank between the post-masks and the post-cue. The musical notes (♪) indicate clicks that were played 500 ms after each post-cue onset, prompting the participant to respond. **(B)** Compound word judgment task (Experiment 1b). On a minority of trials (not represented here), the ISI was set to 400 ms. The musical note (♪) indicates a click that was played 300 ms after the post-masks offset, prompting the keypress response. In both tasks, feedback for each response's accuracy was delivered with a beep.

The cue condition (focused top, focused bottom, or divided) was blocked. Blocks of 20 trials each were run in sets of four: two divided attention and one of each focused attention condition (above and below), in a random order.

2.2.3. Pre-test staircase procedure to estimate ISI thresholds

The staircase was run in blocks of 20 trials, alternating between blocks with attention focused on the top word or bottom word. During each run, we adjusted the word-mask ISI in units of $\log_{10}$(seconds) from trial to trial following a weighted 1-up/1-down staircase procedure. The staircase was run with the Palamedes toolbox (Prins & Kingdom, 2009), and the step size down was one-third of the step size up, which makes the staircase converge on the 75 % correct threshold. Two staircases were randomly interleaved, and blocks continued until both staircases had reversed direction ten times. The threshold ISI was the mean value across all reversals.

2.2.4. Eye-tracking

Throughout the presentation of stimuli, we recorded the right eye's gaze position at 500 Hz with an Eyelink 1000+ video-based eye-tracker (SR Research). Fixation was established during the ITI at the start of each trial. The trial only advanced if the estimated gaze position was within 1.0° horizontally or 1.5° vertically from that fixation position. If a fixation break occurred between the pre-cue offset and post-mask offset, the trial was immediately terminated. The participant had to press a button to continue the next trial. Terminated trials were repeated at the end of the block, unless fewer than three trials remained. This applied to 10.7 % (SEM = 2.4 %) of trials on average in the lexical decision task.

To be sure that we excluded trials in which participants may have looked directly at a word, we also analyzed the eye traces offline. First, for each trial in a block, we computed the median gaze position (across measurement samples) in the 400 ms before the pre-cue onset (excluding intervals for blinks). Then we defined the "central gaze position" for the block as the across-trial median of those initial gaze positions. This analysis corrects for any error in the eye-tracker calibration by assuming that participants were fixating correctly in the interval before the pre-cue, when only the fixation mark was visible.

Then, for each trial, we analyzed gaze positions in the interval between the onset of the words and the offset of the post-masks. We defined an "offline fixation break" as a deviation that was more than 0.7° horizontally or 1.0° vertically from the central gaze position and that lasted more than 30 ms. In the analysis, we excluded all trials with offline fixation breaks.

2.3. Procedure

The dual lexical decision task required four to five 1-h sessions. In session 1, participants completed the TOWRE test of word reading efficiency, received instructions, practiced the task, and ran a staircase procedure that adjusted the duration of the ISI to their 75 % correct threshold in the focused attention condition. The main lexical decision trials began in session 2. The ISI was set to the staircase estimate and then adjusted as needed to maintain focused attention accuracy between 70 and 90 % correct. Any run of 4–12 blocks with an ISI that was too high or too low was discarded and re-run. This exclusion applied to four blocks for two participants, 8 blocks for one other, and 12 blocks for one other. The ISI did not differ between focused and divided attention conditions within each set of four blocks. Testing sessions continued until each participant had completed a total of 50 blocks (1000 trials).

2.4. Experiment 1b: Compound word judgment task

The same 10 participants then completed the following compound word judgment task. All methods were the same as in the dual lexical decision task (Experiment 1a) except as indicated below.

2.4.1. Stimuli

The stimulus set (see Supplementary **Table S4)** was constructed from 193 English "compound words" divided into their constituent words (e.g., waterfall → water + fall). These compound pairs were defined as whole words that typically appear in written form without a space between the two constituents. Seven of these pairs included prepositions (aftermath, afternoon, aftertaste, overflow, overweight, undercover and underweight). Also, two of the 193 combined words in the "compound pair" set were not technically compounds but rather derived nouns: friendship and knighthood. Trials with either of those two constituent pairs were excluded from all analyses. We did this because of two features of "knight + hood" and "friend + ship" that make them different from the rest of the compound pair set: (1) the second constituents are derivational suffixes, rather than free morphemes; and (2) the meanings of the second constituents ("hood" and "ship") when read separately are not related to the meanings of the combined words (knighthood and friendship). Therefore, trials with either of these two word pairs were excluded from the analysis in order to reduce the variability in the morphological structures in our "compound pair" stimulus set. This exclusion had minor effects on the statistics and caused no qualitative changes to any of the conclusions we draw. Nonetheless, it makes the results easier to interpret.

The combined compound words ranged in lexical frequency (Zipf) from 1.8 to 6.2 (mean = 3.08, SD = 0.70). 14 of the 193 compound words did not appear in the SUBTLEX-US Zipf database, but all were verified as being familiar compounds by five research assistants. The compound words varied in semantic transparency: how closely the meaning of the whole word relates to the meanings of its constituents. Our primary goal was to compile a large list of existing compound words, and thus we included some opaque compounds (words for which the meaning of the constituents is not clearly related to the meaning of the whole compound). More details on semantic transparency are reported in the "*Psycholinguistic variables*" section below.

The compound set was constructed from pairs of 293 unique constituent words, which came from multiple syntactic categories, ranging in length from three to six letters and ranging in lexical frequency from 2.8 to 6.3 Zipf (mean = 4.7, SD = 0.66). All of these constituent words also appeared in the dual lexical decision task. Each constituent word which appeared on average 2.6 times within the set of compound and scrambled pairs and 3.3 times in the experiment for each participant.

As described below, the constituent words were presented in "scrambled" pairs on half the trials. There were 193 unique scrambled pairs, each composed of the first constituent word from one compound and the second constituent from a different, randomly selected compound. By chance, five of those were listed as low-frequency unspaced compounds in the SUBTLEX-US corpus: cat + head, blue + stone, house + man, ship + man and snake + head (mean Zipf frequency = 2.1). Excluding trials with these word pairs from the analysis had a miniscule effect on accuracy (it increased by 0.02 %), so we include them in all the results reported below.

2.4.2. Trial sequence and procedure for the compound word judgment task

On each trial, two real words were presented simultaneously above and below fixation (Fig. 2B). The word above fixation was always the first constituent of a compound word (e.g., "stair" in staircase), and the word below fixation was always the second constituent of a compound word (e.g. "neck" in bottleneck). On half of all trials, the two constituent words were paired correctly to form a compound word together (e.g., "stair + case"). On the other half of trials, the two constituent words came from different compounds (e.g., "stair + neck"). A given pair of words appeared on average 1.3 times in the experiment for each participant. Each participant completed 500 trials total in this experiment (20 blocks of 25 trials each).

There were no pre- or post-cues. Each participant's word-to-mask ISI duration was matched to the last ISI used for that participant in the lexical decision task. The ISIs across participants ranged from 0 to 17 ms

(mean = 4 ms, SEM = 1.6 ms). On 33 % of all trials, the ISI was set to 400 ms instead of the threshold ISI. These provided the participants with some "easy" trials and allowed us to ensure that accuracy was high when participants were given more than enough time to process both words (see Supplementary Fig. S2).

300 ms after the offset of the post-masks, a beep prompted the participant's response. Responses before the beep were not recorded. Participants pressed one of four keys (*a, s, d,* or *f*), which indicated, in order from left to right, "sure not compound word," "guess not compound word," "guess compound word," and "sure compound word." Auditory feedback was given as in the lexical decision task. On average, 11 % (SEM = 2 %) of trials were terminated due to fixation breaks.

This compound word task required one 1-h session. Participants received instructions and practiced the task before completing 20 blocks (25 trials per block).

2.5. Analyses

2.5.1. Calculation of accuracy

In both Experiments 1a and 1b, we used the confidence ratings to compute accuracy in units of area under the receiver operating characteristic (ROC) curve, $\mathbf{A_g}$ (Pollack & Hsieh, 1969). (Note that the ROC is not to be confused with the attention operating characteristic, or AOC, which was described in the Introduction). A_g is a bias-free measure of accuracy that ranges from 0.5 (chance) to 1.0 (perfect). One can think of A_g as an unbiased estimate of proportion correct. More detail is provided in the **Supplementary Materials** and in White et al. (2018).

In these experiments, participants emphasized accuracy over speed and were forced to wait until a beep before responding. Thus, we do not analyze response times.

2.5.2. Bootstrapping and bayes factors

Throughout the results, we report bootstrapped 95 % confidence intervals (CIs) for average measurements. To compute these, we generated a distribution of 5000 resampled means, each calculated from a sample of ten values sampled with replacement from the original set of ten participants' means. The CI is the range from the 2.5th to the 97.5th percentile of the distribution of resampled means, with an "accelerated" bias correction (Efron, 1987).

Finally, we supplement our pairwise tests with Bayes Factors (BFs), which quantify strength of evidence. The BF is the ratio of the probability of the data under the alternate hypothesis (a distance is >0 or two conditions differ) relative to the probability of the data under the null hypothesis that there is no difference (Rouder et al., 2009). A BF of 10 indicates that the data are 10 times more likely under the alternate hypothesis than the null. Typically, BFs between 1 and 3 are regarded as weak evidence for the alternate hypothesis, BFs between 3 and 10 as substantial evidence, and BFs between 10 and 100 as strong evidence. Conversely, BFs between 1/3 and 1/10 are considered substantial evidence for the null hypothesis, etc. (Kass & Raftery, 1995). We computed BFs for pairwise *t*-test ANOVAs using the bayesFactor toolbox (Krekelberg, 2024).

2.5.3. Psycholinguistic variables

We explored the effects of several psycholinguistic variables on task performance, using linear mixed effect models fit to single-trial responses. Some variables are properties of single words: orthographic length (the number of letters); the number of morphemes; lexical frequency on the Zipf scale (Brysbaert & New, 2009; van Heuven et al., 2014); orthographic neighborhood size (the number of words that can be created by changing a single letter in the target word); and concreteness ratings (Brysbaert et al., 2014). For the concreteness measure, each word has a rating from 1 to 5, where 1 means "abstract" (something that cannot be experienced through the senses) and 5 means "concrete" (something that can be experienced through the senses). In lexical decision tasks, more concrete words tend to be judged more quickly and accurately than abstract words (e.g., Binder et al., 2005; James, 1975; Kroll & Merves, 1986).

The remaining linguistic variables are properties of word *pairs*. First, to assess the *semantic similarity* between the words in each pair, we calculated the cosine similarity between their GloVe distributional semantic vectors (Pennington et al., 2014). These vectors, calculated from machine learning analyses of large text corpora, represent the position of each word in a 300-dimensional semantic space. Words with higher cosine similarity (which ranges from −1 to 1) more often occur in similar contexts. For example, two pairs with the least semantic similarity in the compound word judgment experiment were "day + washer" and "lime + book", while the two most similar pairs were "web + site" and "butter + milk."

We also estimated the frequencies of each word pair as a spaced combination (or two-word phrase), by extracting data from Google Books Ngrams, using English books published between 1980 and 2022 (Michel et al., 2011; https://books.google.com/ngrams/info). From this we extracted the frequency of each two-word combination, F_{pair}, as well as the frequencies of each constituent word (F_t and F_b for the top and bottom word, respectively). F_{pair} is the frequency of the pair of words written with a space between them. These statistics were available for 100 % of the compound pairs and 52 % of the scrambled pairs (because 48 % of the scrambled pairs never appeared side-by-side in the corpus).

We then calculated the *surprisal* (*S*) of each word given the other it was paired with (Onnis et al., 2022). The "forward" *surprisal* S_b of the second (bottom) word given the first word is: $S_b = -\log2(F_{pair} / F_t)$. The "backward" surprisal S_t of the top word is: $S_t = -log2(F_{pair} / F_b)$. Given that S_b and S_t were correlated ($r = 0.73$), we used the average of them as our measure of surprisal, *S*. Again, note that F_{pair} is based on the frequency of the pair written with a space between the two words, and we calculate it for both the 'compound' and 'non-compound' pairs used in the experiments. The compound pairs with the lowest surprisal were "web site" (S = 2.8) and "tool kit" (6.16), because they are sometimes written with a space between constituents, and those with the highest surprisal were "after math" (14.4) and "keep sake" (15.5), because they are rarely written with a space. The scrambled pairs with the lowest surprisal were "blue stone" (10.3) and "after wave" (10.4) and those with the highest surprisal were "pick ring" (20.1) and "book cup" (20.6).

Finally, the last variable we analyzed was relevant only to the compound pairs: *semantic transparency*. We had four native English speakers rate each compound word as either 1 for fully opaque (the meaning of the compound is not related from the meanings of its constituents), 2 for intermediate, and 3 for fully transparent (the meaning of the compound is clearly related to both its constituents' meanings). The raters all agreed on 20 of the 193 compounds that were fully opaque (e.g., "haywire") while all agreed that 114 were fully transparent (e.g. "bedsheet"). The rest were intermediate.

3. Results

3.1. Experiment 1a: Accuracy in the dual lexical decision task supports the all-or-none serial model

Lexical decision accuracy, reported in units of area under the ROC curve (A_g), was on average 0.85 (SEM = 0.01) in the focused cue condition. It was considerably lower in the divided cue condition, by a mean difference of 0.18 (SEM = 0.05, t(9) = 10.79, $p < 10^{-5}$; BF = 8826. To compare this large deficit to the predictions of three different models, we plot our data on attention operating characteristics (AOCs). The models are explained in the Introduction.

To test each model statistically, we constructed AOCs for each participant (see Supplementary **Fig. S1**) and then computed the Euclidean distances between the divided attention point and the nearest points on the serial model's prediction line, the fixed capacity parallel model's prediction curve, and from the independent parallel model's predicted point. Distances for accuracy below the predictions were

assigned negative values. Table 1 lists statistics on the means of these three distances.

As shown in Fig. 3**A,** mean accuracy for the dual lexical decision task was best predicted by the all-or-none serial model, which traces out the diagonal line between the two focused-attention accuracy points. Divided attention accuracy was significantly below the fixed-capacity model's prediction (the solid blue curve; see statistics in Table 1). All but one participant performed worse than the prediction of the fixed-capacity parallel model (**Fig. S1**).

Thus, the AOC analysis allows us to reject both parallel processing models and supports the serial model: participants could fully process only one of the two words (or pseudowords) presented on each trial. These results match what has been reported before in a similar study with unrelated word pairs (White et al., 2020).

The AOC in Fig. 3A also illustrates that in the divided cue condition, accuracy tended to be higher for the top word (mean = 0.73, SEM = 0.03) than for the bottom word (mean = 0.61, SEM = 0.01). That difference was statistically significant (t(9) = 3.98, $p = 0.003$; BF = 15.6). However, the asymmetry does not seem to be due to an inherent difference in how legible the words were at the two positions. That is because in the focused attention condition, the asymmetry was not observed, with accuracy being slightly better for the *bottom* side (mean = 0.87, SEM = 0.01) than the top (mean = 0.84, SEM = 0.01). The difference was not statistically significant, however (t(9) = 1.50, $p = 0.17$; BF = 0.74). To summarize, in the divided attention condition, when it was effectively impossible to judge both words, participants were biased to process the top word, even though that one was not easier to perceive on its own. This pattern is consistent with a previous dual lexical decision task with a similar design (White et al., 2020). The bias towards the top may be learned from the standard order of reading pages and lists from top to bottom.

3.1.1. *A stimulus processing tradeoff also supports the serial model for the dual lexical decision task*

The all-or-none serial processing model further assumes that there is a trial-by-trial tradeoff between the two words, because participants can process only one per trial. This predicts that accuracy for each side (top word or bottom word) should be lower on trials when the response to the other side was *correct* than on trials when the response to the other side was incorrect (Braun & Julesz, 1998; Lee et al., 1999; White et al., 2018, 2020). For example, if the response to the top word is correct, the top word was probably processed and therefore the response to the bottom word is less likely to be correct.

We tested that prediction by separating all responses on divided attention trials into two sets: (1) the response to the other side on the same trial was correct, and (2) the response to the other side on the same trial was incorrect. Within each set, we computed accuracy (A_g). These data are plotted in Fig. 3**B.** The horizontal bars are the means and each participant's two data points are connected by a thin gray line. Accuracy was significantly worse when the response to the other side was incorrect than correct (mean difference = 0.053, SEM = 0.012; t(9) = 4.06, $p = 0.003$; BF = 17.3). This negative correlation between accuracies is rare for dual-task performance, and, like the AOC analysis, rejects the two parallel models described above and supports the serial model.

3.2. *Experiment 1b: Accuracy in the compound word task nonetheless exceeds the serial model's prediction*

In the compound word judgment task (Experiment 1b), participants again viewed pairs of words flashed at the same positions and masked after the same brief interval. The task was to report whether the two words together formed an existing compound (like stair + case or grand + father), which was true on a random 50 % of the trials. On the remaining trials, the same individual words were mismatched to not form compounds (e.g., stair + father).

We devised this novel task to test a simple model prediction. A strict version of the serial model would assume that participants are able to identify only one of the two letter strings per trial, just like they were in the dual lexical decision task with unrelated word pairs. This predicts that accuracy in the compound word judgment task should be at chance ($A_g = 0.5$, or 50 % correct), because identifying any one word provides no information as to whether the other word formed a compound with it or not.

However, accuracy on trials with the threshold-level ISI greatly exceeded the chance level (mean $A_g = 0.767$, SEM = 0.037, 95 % bootstrapped CI of the difference from chance= [0.19 0.33]; t(9) = 6.83, $p < 0.0001$; BF = 360). This rejects the strict all-or-none serial model's prediction. (See Supplementary **Fig. S2** for an analysis of accuracy in units of *d'*).

Fig. 3C further demonstrates how the serial model fails to account for compound word accuracy at the level of individual participants. Each participant's accuracy in the compound word judgment task (y-axis) is plotted as a function of their distance from the serial model in the divided-attention lexical decision task (x-axis). These distance metrics are negative when lexical decision accuracy is lower than the serial model predicted.

All participants but one performed significantly above chance in the compound word task. That can be seen in Fig. 3C, where their vertical 95 % CIs exclude 0. That is true even when their lexical decision performance was near or worse than the serial model's prediction. (Only one participant, represented by the rightmost dot in Fig. 3C, significantly outperformed the serial model in the dual lexical decision task). This result should not have occurred if the participants could recognize only one half of each compound pair, as would be predicted by their dual lexical decision performance.

We may also make a prediction based on the independent parallel model. This model assumes that both letter strings are processed simultaneously, with the same accuracy as when attention is focused on one word at a time. We model the probability of correctly determining that the two words form a compound as the joint probability of recognizing both words independently.

To estimate the probabilities of correctly identifying the top and bottom words separately, we use data from the focused attention conditions of the dual lexical decision task (Experiment 1a). In doing so we also take into account properties of each individual word in both experiments. Specifically, for each participant, we fit two logistic mixed-effects models to predict their accuracy for real words in Experiment 1a: one for trials with attention focused on the top word, and one for trials with attention focused on the bottom word. The length, lexical frequency, orthographic neighborhood size, and concreteness of each word were predictors.

Then we use those fitted models to predict each participant's

Table 1
Lexical decision performance in Experiment 1a: Mean distances of the divided attention accuracy point from the nearest points on the predictions of the three capacity limit models, calculated for each individual participant. Statistics were computed by two-tailed *t*-tests, with 95 % confidence intervals and Bayes Factors.

Model	Mean distance	SEM	$\|t\|$	p	95 % Confidence Interval	Bayes Factor
Independent parallel	−0.289	0.018	15.4	10^{-7}	[−0.327–0.260]	1×10^5
Fixed-capacity parallel	−0.105	0.021	4.80	0.001	[−0.138–0.051]	42
All-or-none serial	−0.012	0.020	0.54	0.601	[−0.044 0.036]	0.35

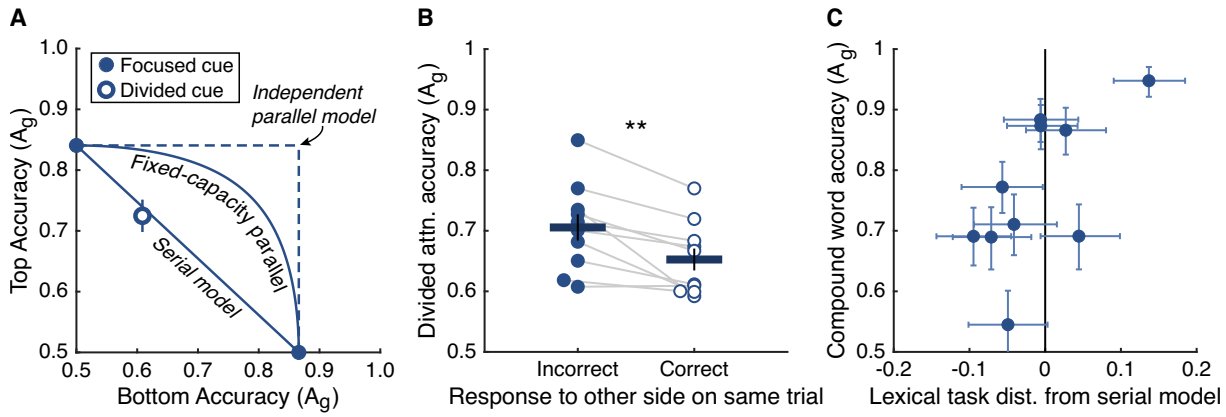

Fig. 3. Results of Experiment 1. **(A)** Mean attention operating characteristic ($N = 10$) for the dual lexical decision task. Solid points pinned to the axes represent focused attention accuracy (in units of area under the ROC curve, A_g). The open point represents divided attention accuracy. Error bars are ±1 SEM. Divided attention accuracy is closest to the all-or-none serial model's prediction. Individual participant AOCs are plotted in the Supplementary Material. **(B)** Stimulus processing tradeoff effect on accuracy (A_g) in the divided attention condition of the dual lexical decision task. Horizontal bars show across-participant means. Each dot is an individual participant's data, with each participant's two points connected by a thin gray line. The asterisk indicates $p < 0.01$. **(C)** Individual participants' accuracy (A_g) in the compound word judgment task plotted against their result from the dual lexical decision task; specifically, the x-axis is the distance of each participant's divided attention point to the all-or-none serial model (see Table 1). Both horizontal and vertical error bars represent 95 % bootstrapped confidence intervals.

probability of correctly identifying each word (top and bottom, separately) in the compound word judgment task, using the four properties (listed above) of each word as predictors. The independent parallel model prediction is that the probability of correctly judging whether both words form a compound is the product of the two probabilities for the top and bottom word on that trial. Integrating across trials, we calculate a predicted percent correct for the compound word judgment task for each participant.

On average, the predicted compound word accuracy was 0.751 (SEM = 0.025), which is statistically indistinguishable from the measured mean proportion correct of 0.747 (SEM = 0.036; t(9) = 0.07, $p = 0.94$; BF = 0.31). Thus, compound word judgment accuracy is not significantly lower than the prediction based on independent parallel processing.

One concern is that participants saw the words more than once during the experiment. Although each individual word appeared just as often in compound pairs as in scrambled pairs, learning compound words might have boosted accuracy. To test that, we analyzed the subset of trials when each word pair appeared for the first time (see Supplementary **Fig. S2**). Accuracy was only 0.025 lower than in the main analysis and still significantly greater than 0.5 (mean = 0.742, SEM = 0.041, t(9) = 5.97, $p = 0.0002$; BF = 152). Therefore, the above-chance accuracy is not due to learning of compound pairs within the experiment.

Another possible concern relates to the diversity of words in the stimulus set. First, two of the constituent words were themselves multimorphemic (e.g., writer, washer). Compounds containing more than two morphemes might be more difficult to detect, and non-compound pairs containing more than two morphemes might be more difficult to reject. Second, seven of the combined words in the "compound pair" contained prepositions (e.g., overflow, undercover). Many of those could be described as a stem modified by a prepositional particle and might be processed differently than compounds composed of two noun stems. Third, 20 of the 193 compound pairs were rated as being semantically opaque (e.g., deadline, haywire), and might be represented differently from transparent compounds. In a subsidiary analysis we excluded all those words, leaving in the 92 % of trials with monomorphemic constituent words and transparent compounds made of two nouns. (Note that the two derived nouns were already excluded from the primary analyses reported above). The results changed only slightly (mean accuracy = 0.77, SEM = 0.038; comparison to chance: t(9) = 6.78, $p < 10^{-4}$; BF = 343). Even when excluding those trials and including only the 1st appearance of each word, mean accuracy was 0.744 (SEM = 0.042), significantly above chance (t(9) = 5.74, $p = 0.0003$; BF = 119). That analysis included 62 % of the data but still over 200 trials/participant.

3.2.1. *Interim discussion of Experiment 1 (dual lexical decision and compound judgment tasks)*

For the dual lexical decision task with pairs of unrelated words (Experiment 1a), we reject the independent parallel and fixed-capacity parallel models in favor of a serial model (one word processed per trial). The serial model accounts for both the cost of dividing attention and the negative stimulus processing tradeoff in accuracy, replicating prior experiments with unrelated word pairs (White et al., 2020). Nonetheless, the serial model cannot account for above-chance accuracy in the compound word judgment task (Experiment 1b), in which participants viewed very similar stimulus sequences as in the lexical decision task. This suggests that when two words combine to form a known compound word, they can pass through any "bottleneck" in the word recognition circuitry together.

One interpretation of this result is still consistent with a serial bottleneck that arises at the stage of lexical access, when a set of input letters is identified with a single word stored in memory. To explain the data, this interpretation must allow that the set of input letters processed together could be divided into two strings with space between them (as in the experiments here). Thus, the compound word result is compatible with the serial model, if we assume the serial processing does not apply to single letter strings, but rather to single lexical items that can span multiple letter strings (each of which functions like a morpheme in the compound; Zang, 2019).

To dig deeper into the underlying mechanisms, we conducted exploratory analyses of how accuracy in the compound word judgment was affected by six psycholinguistic properties of the words on each trial. These properties were entered as predictors into logistic linear mixed effect (LME) models of single-trial accuracy (with random slopes and intercepts by participant). Note that these are exploratory analyses that must be interpreted with caution, especially given covariation between the predictor variables. These variables were: (1) the total words' orthographic length (the sum of the number of letters in both words); (2) lexical frequency (Zipf), either for the compound word or the mean of the two constituent words; (3) concreteness ratings, either of the compound word or the mean of the two constituent words (Brysbaert et al., 2014); (4) Semantic similarity between the two constituent words in

each pair, calculated as cosine similarity of the two GloVe vectors. (5) Mean surprisal, averaged across surprisal of the top word and of the bottom word. This last metric was calculated using Google NGrams statistics as described above in the Methods section. (6) Semantic transparency ratings of how closely the meaning of each compound word is related to the meanings of the two constituents.

We fit three models: one for all trials, one for compound pair trials only (including the seven words that contain prepositions), and one for non-compound (scrambled pair) trials. The outputs of the model fits are shown in Table 2**.** In the first model fit to all trials, the combined word length and mean constituent concreteness had significant positive effects on accuracy. Semantic similarity, transparency, and surprisal were not included in this first analysis, because they might predict opposite effects for accuracy on compound (target present) and scrambled (target absent) trials.

Second, in the analysis of compound word trials, we initially included all six predictors. Semantic transparency ratings had a significant positive effect (mean estimate = 0.45, $p = 0.002$). Indeed, in a simple comparison we found for compound pairs, hit rates were significantly lower on the 10 % of trials with fully opaque pairs than all other trials (means = 0.61 vs 0.53; t(9) = 3.57, $p = 0.006$; BF = 9.4). That would be consistent with prior reports that opaque compounds are more difficult to process when presented in a spaced format that encourages decomposition (Frisson et al., 2008). However, the LME with transparency as a predictor did *not* fit better than a simpler model without it (AIC = 6861 vs 6831, respectively; log likelihood = −3395 vs −3389). The results of this simpler model, with five predictors, is shown in the middle section of Table 2**.** The total orthographic length of the compound and the compound's concreteness had significantly positive effects on the probability of a correct report of seeing a compound. The semantic similarity (cosine distance) of the constituents and their surprisal relative to each other did not have significant effects, nor did the compound's lexical frequency.

The same results arose when excluding semantic similarity as a predictor from the start. Also, in a subsidiary analysis in which we excluded all multimorphemic constituent words, compounds containing prepositions, and opaque compounds, we still found null effects of the semantic similarity and the surprisal between the constituent words of compound pairs. Thus, variability in the compound pair set was not obscuring those effects.

To summarize this analysis of accuracy for detecting compound words: semantic similarity and surprisal did not significantly affect hit rates. There was mixed evidence when it comes to semantic transparency of the compounds: hit rates were lower overall for the most opaque compounds, but the effect of transparency does not add any predictive value over the other variables. However, the compounds were easier to detect when they contained more letters and referred to more concrete objects. The concreteness effect is not surprising given past analyses of lexical decision and naming performance (e.g., Binder et al., 2005; Forster & Chambers, 1973; James, 1975; Kroll & Merves, 1986). The positive effect of orthographic length is contrary to the typical (small) negative effect on a variety of visual word recognition measures (Barton et al., 2014). In this compound word task, longer words might provide less uncertainty for distinguishing compound from non-compound pairs.

The third analysis in Table 2 is just for non-compound trials (and therefore did not include compound transparency as a predictor). We found a weak and not statistically significant trend for accuracy to decrease with semantic similarity ($p = 0.075$). A simpler model *without* semantic similarity as a predictor provided a better fit to the data than the model with semantic similarity (AIC = 4645 vs. 4687; log likelihood = −2302 vs −2316). Surprisal, in contrast, did have a significant and positive effect on accuracy (see bottom of Table 2). The model with surprisal as a predictor fit significantly better than a simpler model without it (AIC = 4687 vs 8759, likelihood ratio test $\chi^2(7) = 4086$, $p < 0.001$). The positive effect of surprisal on non-compound trials means that participants were more likely to make a false alarm (incorrectly report seeing a compound word) when the two words happened to co-occur more often in natural text (with lower surprisal).

The evidence in these secondary, exploratory analyses must be interpreted cautiously. We do not draw strong conclusions from the null effect of semantic similarity or transparency. The significant effect of surprisal for non-compounds suggests that participants are sensitive to the statistical regularities of word pairings, although that effect was absent for trials with compounds.

To recap, the primary result in Experiment 1 was that participants could detect when two words form an existing compound, even under the same presentation conditions when lexical decisions for randomly paired words supported the serial processing model. That suggests that some degree of parallel or holistic processing of multiple constituent words can occur.

Importantly, the compound word judgment task (Experiment 1b) explicitly required participants to look for compound words. The participants had to attend to both constituent words on each trial and judge whether they formed a single compound together. Does parallel processing of two constituents depend on such voluntary effort to process them as a single compound? To find out, we conducted a second experiment in which compound words appeared without the participant's knowledge and were not relevant to their task.

Table 2
The results of fitting generalized linear mixed effect models to single-trial accuracy in the compound word judgment task (Experiment 1b). Significant predictors ($p < 0.05$) are bolded. "Frequency" is lexical frequency in Zipf units.

Predictors of accuracy across all trials

Factor	Mean estimate	*t*(3302)	*p*
total length	0.12	2.80	0.005
mean constituent frequency	0.13	1.29	0.197
mean constituent concreteness	0.28	3.11	0.002

Predictors of accuracy for compound word trials

Factor	Mean estimate	*t*(1507)	*p*
total length	0.13	2.10	0.036
compound frequency	0.16	1.70	0.090
compound concreteness	0.21	2.58	0.010
semantic similarity	−0.28	−0.50	0.614
surprisal	−0.06	−1.50	0.134

Predictors of accuracy for non-compound trials

Factor	Mean estimate	*t*(871)	*p*
total length	0.05	0.40	0.686
mean constituent frequency	0.12	0.44	0.663
mean constituent concreteness	−0.07	−0.32	0.750
semantic similarity	−2.88	−1.78	0.075
surprisal	0.15	2.46	0.014

4. Experiment 2

The goal of Experiment 2 is to investigate whether two words that happen to form a compound can be processed in parallel, under conditions when compound words are rare and unexpected, and when participants attempt to process the two words independently.

One approach to accomplish that goal would be to use a dual lexical decision task as in Experiment 1a, adding in some pairs of real words that together form a known compound. There is a drawback to that approach, however: two real words occur on only 25 % of trials. So, if we wished to make both compound pairs and scrambled compound pairs appear in rare subsets of those trials, we would end up with very few trials per key condition. We therefore devised a novel typed report task that is similar to Experiment 1a, but the task on each trial is simply to

type in the word on one post-cued side. This design has several advantages: (1) All the stimuli are real words so we can collect more trials per condition, while inserting compounds on a small minority of trials. (2) Chance accuracy is much lower than in the lexical decision task. In the lexical decision task, chance accuracy was 50 % because there were only two possible responses (word or non-word). In this typed report task, each word was chosen from a set of 918, so chance accuracy is at most 0.1 % (but effectively lower than that because the participant did not know the set of 918 possible words). This increases the distance between serial and parallel models and thus our statistical power. (3) We can learn more about *what* participants perceived (or guessed) by analyzing the words they typed in. For instance, some incorrectly reported words might share letters with the target words, and thus provide evidence of parallel orthographic processing (although it did not, as shown below). (4) We can examine whether participants were *biased to guess* words that could form compounds. That would have been impossible to determine with the lexical decision task.

It is important to note that this typed report task may not measure the same type of lexical processing as the lexical decision task (used in Experiment 1a). In the lexical decision task, participants must judge the whole word form to distinguish it from pseudowords with matched trigram statistics. In the typed report task, the participant may perceive only partial orthographic information (i.e., they may identify only some of the letters in the word) and then correctly guess the whole word. So, the first important question about this new typed report task is whether the serial model would continue to predict accuracy for unrelated word pairs. We investigate this by analyzing accuracy as a binary variable (whole word correct or not) and as a continuous measure of the proportion of letters correctly reported.

In more detail: participants in Experiment 2 were presented with two real words on each trial and their task was to report one post-cued word by *typing* it into a text box (see Fig. 4). In separate blocks, a pre-cue instructed participants either to focus attention on the top side or bottom side (because they would only be asked to report the word on that side), or to divide attention between both sides (because they could be asked to report either one of the two words). On 70 % of all trials, the two words presented together were unrelated (either randomly selected pairs or scrambled compounds, as explained below). Participants were always motivated to process the two words independently as best they could. On a random 15 % of trials, the two words formed an existing compound word – for example, "cow" above fixation and "boy" below. Participants were not informed of the presence of compound words and almost never noticed them, as demonstrated by debriefing interviews after the experiment.

The primary control condition that provides the baseline against which we compare accuracy used "scrambled compound" pairs: the halves of two different compound words were mismatched (just as in Experiment 1b). Thus, the same constituent words at the same positions were used in both the "compound" and "scrambled" conditions. Any differences in accuracy across conditions cannot be attributed to the properties of the constituent words, but only to how they combine.

We assumed that when our English-reading participants are presented with a compound pair, they would (perhaps implicitly) consider the word at the top position to be 'first', like the left half of an unspaced compound word. But to test that, we also included 15 % of trials with the same compound pairs in the "*reversed*" order, that is, with the second constituent on top. The inclusion of this reversed compound condition also reduces the expectation that the top word might be the first half of a compound word; if participants had that expectation, they might adopt a strategy to attend only to the top word and then guess a bottom word that could form a compound with it. We also analyze the specific words reported to rule out this biased guessing strategy.

4.1. Methods

All methods were the same as in Experiment 1 except as described below.

4.1.1. Participants

13 new volunteers (all female, ages 18–22) were recruited in the same manner and from the same population as in Experiment 1, and all provided informed consent. None had participated in Experiment 1. The sample size was determined based on a power analysis of an independent pilot experiment with a similar design and 12 participants. In that experiment we estimated the difference in accuracy between compound word pairs and randomly selected word pairs: mean = 23.5 %, SD = 7.8

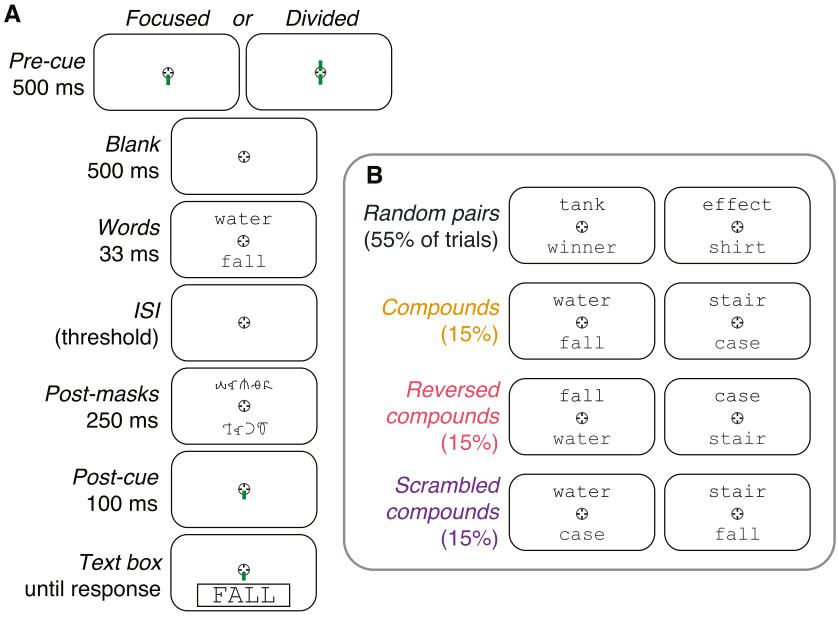

Fig. 4. Design of Experiment 2. (**A**) Example trial sequence, which could be a focused attention or divided attention trial, depending on the pre-cue. The mean threshold-level inter-stimulus interval (ISI) was 12 ms (range 0 to 33 ms). The participant's task was to type in the word on the post-cued side (in this case, the bottom side). Not shown is a 100 ms blank interval between the post-masks and the post-cue. Central gaze fixation was enforced. (**B**) Examples of the four different word pair types, along with the percentage of trials in which they appeared.

%. That comparison was not perfect because the individual constituent words were not identical. That is why we undertook the second experiment presented here. For the power analysis, we made the rough prediction that in this improved experiment, the mean effect would be 1/3 as large with the same SD. We then simulated 10,000 paired *t*-tests of an effect of that size and variability and found that 13 participants are sufficient to reject the null hypothesis with 90 % power.

On the composite TOWRE-II Test, the participants' mean score was 112 (SEM = 2.4), with all but one of the participants scoring above the norm of 100.

4.1.2. Stimuli

The stimulus set (which is reproduced in full in the *Supplementary Materials*, **Tables S4-S6***)* consisted of 918 English nouns from the MCWord database, ranging in length from three to six letters and in lexical frequency from 2.0 to 6.7 Zipf (mean = 4.4). 242 words were selected such that they could form 121 "compound pairs" when combined (e.g., "water" and "fall"). 66.4 % of the compound words also appeared in Experiment 1. As in Experiment 1, this set included both semantically transparent and opaque compound words. Our raters agreed that 10 of the 121 were fully opaque (e.g., hopscotch, milestone, deadline), 70 were fully transparent, and the rest were intermediate. Fourteen of the constituent words were themselves multimorphemic (e.g., affixed nouns like "hopper" in "grasshopper"). Four of the second constituent words were plural. One of the compound pairs was "knight+hood," which does not technically form a compound but is a derived noun. Trials with "knight" or "hood" as either constituent were excluded from the analyses. This applied to 0.5 % of the 8356 trials in the data set. The exclusion caused minor changes to the statistics and did not qualitatively change any of the conclusions we draw. As in Experiment 1, excluding this derived word nonetheless makes the 'compound pair' set more consistent and the results therefore easier to interpret.

The remaining 676 individual nouns that did not form compound word pairs were used for the "random pair" condition. The distributions of length, concreteness, orthographic neighborhood, and lexical frequency in the "compound" word set were roughly matched to the "random" set (see the *Supplementary Materials*).

To assess the *semantic similarity* between each pair of words, we calculated the cosine similarity between their GloVe distributional semantic vectors (Pennington et al., 2014), as in Experiment 1. The two pairs with the smallest semantic similarity were "number + twig" and "wave + ware", while the two most similar pairs were "girl + woman" and "knee + injury." The means of this measure for each pair type are in Table 3, along with summary statistics for frequency and concreteness.

As in Experiment 1, we also analyzed the "surprisal" for each pair. These statistics were available for 100 % of the compound pairs and 33 % of the other pair types. The two pairs with the lowest surprisal (averaging 'backward' and 'forward' surprisal for the top and bottom word, respectively), were "rib cage" (a compound pair, with mean surprisal 3.5) and "orange county" (a random pair, mean surprisal 5.1). Another low-surprisal random pair was "wild turkey" (mean surprisal 6.6) The two pairs with the highest surprisal were "hotel deal" and "member cent" (mean surprisal 22).

4.1.3. Trial sequence and procedure

The trial sequence, illustrated in Fig. 4A**,** was the same as the dual lexical decision task of Experiment 1a except as indicated here. At the end of each trial, a post-cue (green line) pointed to either the top or bottom side, followed 100 ms later with a black-outlined text box that was centered 2.1° from fixation on the same side as the post-cue (which remained visible). The post-cued side matched the pre-cue on focused attention trials and was equally likely to be the top or bottom on divided attention trials. The response was prompted by the appearance of the text box, with no accompanying beep (unlike Experiment 1). The participant then had unlimited time to *type in* the one word they had seen on the post-cued side. The letters they typed appeared in capital letters within the text box, and they could delete letters as necessary, then press the return key once they were satisfied. Unlike in Experiment 1, they only ever had to judge one of the two words, even on divided attention trials. Feedback for correct or incorrect responses was delivered with beeps as in Experiment 1.

As shown in Fig. 4B, there were four types of word pairs: (1) *Random pairs*: two randomly selected nouns from the set that do not form compounds. (2) *Compound pairs*: the two words formed a compound word in the correct order. For example: "sun + flower" and "grand + father". The first constituent word (e.g., sun in sunflower) appeared above fixation, while the second constituent (e.g., flower) appeared below fixation. (3) *Reversed compound*: two words could form a compound but were presented in reverse order, with the second constituent on top and the first constituent on bottom. For example: "flower + sun" and "father + grand. (4) *Scrambled compound*: the two words presented together were mismatched constituents from two different compound word pairs. The top word was the first constituent from one compound pair, and the bottom word was the second constituent from another. For example: "sun + father" and "grand + flower".

The "random pair" condition was presented in 55 % of trials, while the other three conditions were presented in 15 % of trials each (randomly intermixed). The participants received no explicit information about these conditions, were not aware that some word pairs would form compound words, and were only instructed to report the *single* post-cued word for each trial (even on divided attention trials).

Importantly, the scrambled compound condition is the baseline against which we compare the reversed and compound conditions, as they all used the same set of constituent words and occurred just as frequently.

The study required 2–3 one-hour sessions per participant. Following the instructions in the initial session, the participant practiced the task with slow stimulus presentation. Then we used a staircase procedure to estimate participant's threshold for the stimulus-mask ISI, as in Experiment 1a, using focused cue trials only. The staircase converged on the 67 % correct threshold (which was lower than the threshold in Experiment 1 because chance accuracy is lower in this typed report task). Then each participant completed 640 trials of the main experiment (in 32 blocks of 20 trials). The cue condition (focused top, focused bottom, or

Table 3
Psycholinguistic properties of word pairs of each type in Experiment 2. Each column lists the mean (and standard deviation in parentheses). "Constituent freq" is the mean Zipf lexical frequency of the two constituents in each pair. "Compound freq" is the lexical frequency of the compound formed by concatenating the two constituents (e.g., staircase). (n/a*) indicates that 0 of the random pairs appeared unspaced in the corpus. (n/a^) applies to the reversed compound pairs, for which only 2 of the pairs appeared in the corpus datasets: townhome and slipcover. "Semantic similarity" is the cosine similarity of the GloVe semantic embedding vectors of the constituent words. "Compound concreteness" is the concreteness rating from Brysbaert et al. (2014). "Surprisal" is calculated from Google Ngram frequencies of the spaced word pair and the constituent words. Surprisal could be calculated for 100 % of compound pairs, 28 % of random pairs, 78 % of reversed pairs, and 48 % of scrambled pairs. The remaining pairs did not appear in the corpus because they are very unlikely combinations.

Pair Type	Constituent freq (Zipf)	Compound freq (Zipf)	Compound concreteness	Semantic similarity	Surprisal
Random	4.33 (0.48)	(n/a*)	(n/a*)	0.09 (0.10)	16.0 (2.5)
Compounds	4.58 (0.52)	2.98 (0.67)	4.36 (0.77)	0.26 (0.11)	10.1 (1.9)
Reversed	4.58 (0.52)	(n/a^)	(n/a^)	0.26 (0.11)	14.7 (2.5)
Scrambled	4.58 (0.48)	(n/a*)	(n/a*)	0.12 (0.11)	15.8 (2.3)

divided) was blocked, and the word pair types were randomly intermixed.

Throughout the experiment we adjusted the word-to-mask ISI to keep each participant's focused-attention accuracy between 65 % and 85 % correct. Any run of 4 to 12 blocks with an ISI that was too high or too low was discarded and re-run. The mean number of excluded blocks per participant was 4 (ranging from 0 to 16). The mean ISI on included trials was 12 ms (SEM = 2.5 ms, range across participants 0 to 33 ms).

4.1.4. Analysis

Our primary method of analyzing accuracy in this typed report task is to score a response as "correct" if the typed letter string exactly matches the post-cued word. This is a strict measure of how well the words were perceived, and it is the basis of the primarily analyses presented below. Specifically, *p(correct)* is the proportion of trials with an exactly correct response. To construct the AOCs from these data, we must specify the chance level of accuracy (accuracy achieved if the participant makes totally random guesses). This chance level forms the origin of the AOCs. To estimate the chance level, we assume that a participant with 0 information about the target word would type a word drawn randomly from the set of all 1794 words that includes the stimulus set and all words entered by all participants. Thus, the chance level is effectively 0: 1/1794 = 0.00057.

For these AOCs, we can compute the all-or-none serial model's prediction (diagonal line between the two focused cue accuracy points) and the independent parallel model's prediction (focused cue accuracy = divided cue accuracy). However, we lack the fixed-capacity parallel model for this task, as it has only been developed for 2-alternative forced-choice tasks like that used in Experiment 1. Therefore, we simply computed the minimum distance between each participant's divided-attention accuracy point and their own serial model's prediction.

An alternate method to compute accuracy is the "edit distance" between the participant's typed response and the target word that was presented. This metric could be sensitive to partial information about the words that the participant perceived. However, as briefly described in the discussion section below and reported at length in the *Supplementary Materials*, the analyses of the edit distance led to the same conclusions as the analyses of p(correct).

5. Results

5.1. Divided-attention accuracy is highest when the two words forms a known compound

Fig. 5A plots the mean p(correct) in the focused and divided cue conditions, separated by word pair type. To analyze these data, we fit a linear mixed effects model to each participant's proportion correct, with fixed effects of cue and word pair type, and with random slopes and intercepts by participant. First, overall accuracy in the divided attention condition was significantly lower than in the focused attention condition ($F(1,96) = 415$, $p < 10^{-35}$). Mean accuracy on focused cue trials was 0.76 (SEM = 0.01), while accuracy on divided cue trials was 0.37 (SEM = 0.02). Accuracy also varied across word pair types ($F(3,96) = 17.4$, $p < 10^{-8}$), and word pair type interacted with cue condition: $F(3,96) = 16.5$, $p < 10^{-8}$).

To interpret that interaction, within each cue condition we computed pairwise comparisons between each word pair type and the "scrambled compound" condition, which serves as our baseline. After correcting for false discovery rate (q = 0.05), only one significant difference emerged:

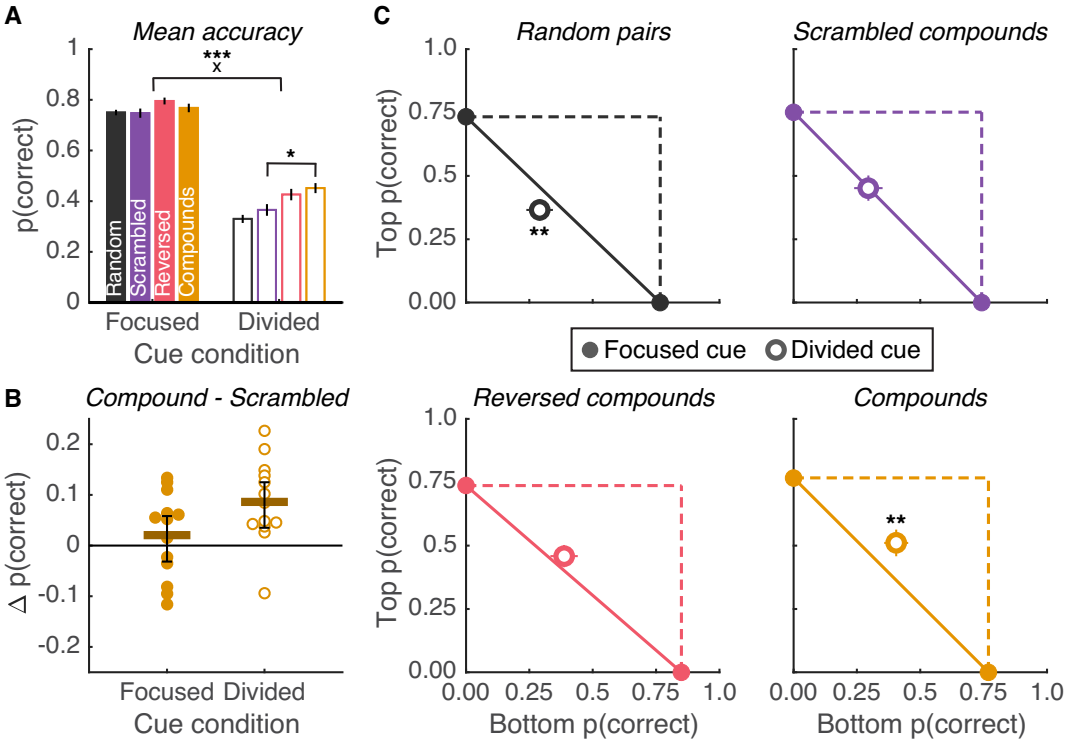

Fig. 5. *Results of Experiment 2.* **(A)** Mean accuracy: the proportions of trials with correctly reported words in each condition. Error bars are ±1 SEM. The central asterisks and "x" indicate the significant main effect of cue condition and the significant interaction between cue condition and word pair type. Within each cue condition, we compared all word pair types to the "scrambled compounds" baseline condition. The lower asterisk indicates the only significant difference: between compound pairs and scrambled compounds in the divided attention condition. **(B)** The *differences* in accuracy between trials with compound word pairs and scrambled compound pairs, separately for each cue condition. Each dot is an individual participant, and the horizontal lines are the means. Error bars in black indicate 95 % *confidence intervals*. **(C)** Attention Operating Characteristics (AOCs) for each word pair type, constructed from mean accuracy levels. Error bars are ±1 SEM. Asterisks indicate that the minimum distance between the divided-attention accuracy point, and the serial model's prediction line is significantly different from 0 (FDR-corrected for the 4 comparisons).

under divided attention, accuracy in the correctly ordered "compound" condition was greater than the "scrambled" compound condition (mean difference = 0.09, SEM = 0.02; t(12) = 3.75, FDR-corrected *p*-value = 0.019, BF = 17; 95 % CI = [0.04 0.13]). Fig. 5B plots that difference estimated from individual participants' accuracy levels, for both focused and divided cue conditions. In the divided cue condition, all but one participant had higher accuracy on compound word trials than scrambled word trials.

This finding demonstrates that when participants divided attention between both words, they were sensitive to the association between two words that formed a compound word together. That association improved the participants' ability to report either of the constituent words.

In the divided attention condition, mean accuracy for words in reversed compound pairs was intermediate between the scrambled and correctly ordered compound pairs. Reversed compound accuracy did not differ significantly from the compound pair condition (mean difference = −0.026, SEM = 0.020, t(12) = 1.24, FDR-corrected $p = 0.34$; BF = 0.52) or from the scrambled compound condition (mean difference = 0.061, SEM = 0.029, t(12) = 1.97, FDR-corrected $p = 0.17$; BF = 1.2).

The next step is to test the serial model with AOCs for each word pair type (Fig. 4C**)**. For each word pair type, we constructed AOCs for each participant, computed the minimum distance between the divided-attention accuracy point and the serial model prediction, and then conducted a one-sample *t*-test on the 13 participants' distance measures. Those statistics are reported in Table 4. For the random pair condition, accuracy was significantly *worse* than the serial model predicts, highlighting the great difficulty in processing two unrelated words at the same time. For the scrambled compound and reverse compound conditions, accuracy was not distinguishable from the serial model's prediction. But for the compound word pairs, accuracy significantly exceeded the serial model's prediction, by about 10 percentage points in accuracy (BF = 13.7). Thus, under the exact same conditions in which participants could only report one of two words presented in an unrelated pair, their accuracy improved when the two words happened to form a known compound together. The AOC suggests that some parallel processing of the two words is possible in that condition, although with limited capacity.

As in Experiment 1a, we examined differences in accuracy across the top and bottom sides, collapsing across word pair types. On divided cue trials, there was again a trend for higher accuracy at the top (mean = 0.41, SEM = 0.04) than at the bottom (mean = 0.32, SEM = 0.05), but it was not reliable across participants, and the paired t-test was not significant (t(12) = 1.00, p = 0.34; BF = 0.42). Also, as in Experiment 1a, this trend reversed in the focused cue trials (top mean = 0.74, SEM = 0.01; bottom mean = 0.78, SEM = 0.02) but was also not significant (t(12) = 1.50, $p = 0.16$; BF = 0.69).

5.1.1. Ruling out awareness of compound words, biased guessing, and word repetitions

Participants were generally not aware that compound words were presented. Immediately after completing their last block of trials, each participant was asked, "Did you notice any relationship or pattern between the words?" All but two of the participants said no. Of these two participants, one noticed varying levels of difficulty among certain word pairs but remained skeptical that there was an underlying reason for it. The other reported noticing a single compound word (grass + hopper).

Nonetheless, there are two concerns to address: the first is biased guessing. The guessing hypothesis is that participants perceived only one word on the "compound" word pair trials, and when prompted to report the other side, they guessed a word that *would* form a compound with the word that they did perceive. This could explain the higher performance for compound word pairs, even if only one of the two words was recognized per trial. To test this hypothesis, we analyzed *incorrect* responses on scrambled compound trials in the divided attention condition. We counted how often participants reported a word that *did* form a compound with the word on the other side of the screen (which was not post-cued). 9 of 13 participants never did that. The remaining participants guessed a compound on only 1 or 2 of these error trials. In the whole set of 619 error responses on scrambled-compound divided-attention trials, there were a total of 7 responses that did form a compound. Therefore, the relatively high accuracy for compound word pairs cannot be attributed to biased guessing.

The second concern is the repetition of words. Each constituent word was presented to each participant on average 2.3 times (including practice, staircase, and main experiment blocks). Each word was the post-cued target 1.7 times on average. Upon seeing a word for the second time during the experiment, the participant may process it faster, which could in theory be especially beneficial on compound word trials. Overall, accuracy in the divided attention condition was slightly higher when a word was repeated than when it appeared for the first time (mean benefit = 0.04, SEM = 0.02). To test whether the benefit for repeated words specifically benefited words of any pair type, we fit a linear mixed effect model to single-trial accuracies in the divided attention condition, with fixed effects for word repetition, word pair type, and the interaction of those two effects, and random slopes and intercepts by participant. The increase in p(correct) caused by word repetitions was nearly significant (F(1,4162) = 3.86, $p = 0.050$), and it did not interact with word pair type (F(3,4162) = 0.29, $p = 0.83$). Thus, the small benefit to overall accuracy when words repeated did not specifically benefit compound word pairs and cannot explain the main results reported above.

Lastly, we excluded the 16 % of trials that contained multimorphemic constituent words (e.g., washer, payers). This was in addition to the exclusion, already applied to the analyses above, of trials containing "knight" or "hood." The results were unchanged by these exclusions. The mean difference in accuracy between compound and scrambled pairs on divided attention trials was 0.08, SEM = 0.024 (t(12) = 3.41, $p = 0.006$).

5.1.2. Consistent results in an analysis of the partial reports of the letters within words

The proportion-correct data analyzed above provide a strict test for how well participants could precisely report whole words. But might participants have perceived *partial* information about two words at once? If so, when they made an error, they may have reported a word that shares some letters with the target word. We also addressed this possibility by computing the "edit distance" between each target word and the word participants typed in, then normalized by the length of each word. This alternate measure of accuracy, which we call p(letters correct), can be considered the mean proportion of *letters* correctly reported.

Table 4
Tests of the serial model in Experiment 2. For each word pair type, this table lists statistics on the distances between the divided attention accuracy point (in units of proportion correct) and the nearest point on the serial model's prediction line. BF = Bayes Factor; CI = confidence interval.

Word pair type	Mean	SEM	*t*(12)	FDR-corrected *p*	BF	95 % CI
Random pairs	−0.073	0.020	−3.51	0.009	11.57	[−0.110–0.032]
Compound	0.099	0.026	3.62	0.009	13.74	[0.048 0.149]
Reversed	0.052	0.033	1.51	0.208	0.70	[−0.015 0.112]
Scrambled	−0.008	0.044	−0.18	0.856	0.28	[−0.085 0.091]

The results of this analysis are fully reported in the Supplementary Materials and are largely consistent with the analyses of whole-word accuracy reported above. See **Supplementary Fig. S4 and Table S1** for details**.** For compound word pairs, accuracy was clearly higher than the serial model's prediction (mean distance = 0.14, SEM = 0.02, t(12) = 5.61, FDR-corrected $p = 0.0005$, BF = 258). Interestingly, for reversed compound pairs, accuracy was also significantly above the serial model, but by a smaller amount (mean distance = 0.09, SEM = 0.03, t(12) = 2.66, FDR-corrected $p = 0.041$, BF = 3.2). Thus, there is some evidence that the association between the two halves of a compound influenced performance even when they were in the wrong order. Overall, the analysis of p(letters correct) demonstrates that the 'serial bottleneck' for randomly selected word pairs also applies to partial or sub-lexical processing (Campbell et al., 2024). It also confirms the significant benefit for target words in the compound pair condition.

5.2. Interim Discussion of Experiment 2

Experiment 2 confirms that when participants tried to recognize both of two *unrelated* words presented above and below fixation, they could report only one and made random guesses when asked about the other. Nonetheless, when the two words happened to form an existing compound word (as occurred unexpectedly on a minority of trials), accuracy rose modestly but significantly above the serial model's prediction. We propose that there is a processing benefit specifically when two words combine to form a single known compound.

To further test this proposition, we next conducted exploratory analyses of psycholinguistic properties of the constituent words and the word pairs. Our goal is to determine whether some properties could explain the overall benefit for compound word pairs observed above. First, it must be emphasized that any properties of the individual constituent words cannot explain accuracy differences between the compound, reversed, and scrambled pair conditions reported above, because those three conditions used the same words, just in different pairings and orders.

But trials also differed in properties of the word *pairs*. We investigated two properties that depend on the relations between the words and their co-occurrence statistics. The first property is the semantic *similarity* between the constituent words, which we estimated as the cosine similarity between their GloVe semantic embedding vectors (see Methods above). As shown in Table 3, the compound pairs did have greater mean cosine similarity than the scrambled pairs or random pairs. The semantic association between the compound constituents might make them easier to recognize quickly than two words that are far in semantic space. Perhaps two semantically similar words "prime" each other (Meyer & Schvaneveldt, 1971).

The second property was the *surprisal* of the target word given its pairing, which is proportional to inverse of the probability that the target word would be written next to the other in natural text. Accuracy for recognizing one word might be higher when it is paired with a word that it is paired with more often in written language. This could arise if the learned statistics of common two-word combinations facilitates processing of each constituent.

We entered those two properties of each word pair, semantic similarity and surprisal, as predictors in a logistic mixed effects model of single-trial accuracy, along with five control variables: the total number of letters in both words; the total number of morphemes in both words; the target word's lexical frequency; the target word's orthographic neighborhood size; and the target word's concreteness. Those last three variables are of the one constituent word that the participant was post-cued to type in on each trial, in contrast to the other variables that are defined by the combination of both words in the pair. Table 5 shows the results of this analysis, across all divided attention trials for all word pair types.

All three properties of the individual target word significantly predicted accuracy: positive effects of increasing lexical frequency and concreteness, and a negative effect of orthographic neighborhood size. The frequency and concreteness effects are consistent with past research (e.g., Forster & Chambers, 1973; James, 1975; Schwanenflugel & Stowe, 1989). Orthographic neighborhood size is known have a range of effects in different tasks (e.g., Andrews, 1997; Carreiras et al., 1997). In this typing task with limited perceptual input, words with more orthographic neighbors are easier to confuse with their neighbors. For example, imagine that the target was a 5-letter word and the participant is unsure about the identity of the third letter. If the target was a word like "water" that has many orthographic neighbors, the participant might incorrectly report "wader", "wafer", "wager", or "waver." If the target is a word like "wagon" with no orthographic neighbors, even without knowing the third letter the participant is quite likely to correctly report the only possible word.

Table 5
Fixed effect statistics from a generalized linear mixed model fit to divided attention trials of Experiment 2 (including all word pair types). Significant predictors are bolded. "Lexical freq" is lexical frequency on the Zipf scale. "Orthographic N" is the number of orthographic neighbors.

Factor	Mean estimate	$t(2048)$	p
target lexical freq.	0.159	2.09	0.036
target orthographic N	−0.017	−2.01	0.044
target concreteness	0.224	2.78	0.006
total num. Letters	−0.056	−1.26	0.208
total num. Morphemes	0.069	0.46	0.643
semantic similarity	0.016	0.03	0.974
surprisal	−0.038	−2.27	0.024

In terms of the properties of the word *pair*, the total number of letters and morphemes did not significantly affect accuracy, nor did the semantic similarity of the two constituent words did not predict accuracy. Surprisal did have a significant effect, however. The model with surprisal fit significantly better than a simpler model that did not have it as a predictor (AIC = 8902 vs 17,957; likelihood ratio test $\chi^2(8) = 9071$, $p < 0.001$). Thus, participants were relatively less likely to correctly report target words that were more surprising (i.e., less likely) given the other word they were paired with, according to their frequencies in English books.

The significant effect of surprisal on accuracy across all pair types could be due to the differences between the compound and non-compound pairs. As noted in Table 3, the compound pairs tended to have lower surprisal than the other pair types. This is because many of them can be written either with or without a space between the constituents (e.g. "tool kit", "jelly bean") and appear in the corpus as common two-word combinations with low surprisal. But if there is a general effect of surprisal, beyond this difference in the word pair types, then it should also arise within the scrambled and random pair trials, for which surprisal also varied across a wide range. We tested that by fitting fit the same linear mixed effect model (as reported in Table 5) to divided attention trials with only scrambled compound or random word pairs. However, there was no significant effect of surprisal (mean slope estimate = 0.016, t(927) = 0.49, $p = 0.62$). This null result must be interpreted cautiously, as this exploratory analysis may be underpowered. It does, however, temper our confidence in the explanation that words within compound pairs were easy to report *only* because of their relatively low surprisal.

The analysis in Table 5 also reported a null effect of semantic similarity, but semantic similarity was negatively correlated with surprisal across word pairs ($r = -0.38$, $p = 10^{-70}$). Given that collinearity, we also separately analyzed the effect of semantic similarity. Of particular interest is whether it would predict accuracy on *non-compound* trials, which could suggest a sensitivity to semantic similarity even for word pairs that are not commonly written side-by-side. We fit a linear mixed effect model to accuracy on divided attention trials with scrambled and random word pairs, with the individual target word properties (frequency, concreteness, orthographic neighborhood size, all of which

were significant). The effect of cosine semantic similarity was not significant (mean estimate = 0.29, t(2895) = 0.65, $p = 0.51$). A simpler model *without* semantic similarity had a better (lower) AIC score (3680 vs 3688) and could not be rejected by a likelihood ratio test ($\chi^2(6) = 3.40$, $p = 0.76$).

We conclude that semantic similarity does not, on its own, make two words significantly easier to process in parallel. Rather, the benefit observed for correctly ordered compound pairs is likely due to their combination into a single known lexical unit.

Lastly, we analyzed the effects of four properties of the combination of the two words only on trials with compound pair trials. These properties were: (1) the semantic *transparency* of each compound word, as established by ratings of how related the compound's meaning is to its constituents' meanings; (2) the concreteness of each compound word (Brysbaert et al., 2014); (3) the lexical frequency of each compound word; (4) the surprisal of the word pair, as defined above. We fit a linear mixed-effects model to predict accuracy using those four predictors on divided attention trials with compound pairs only (including all words in the compound pair set). None of those predictors had a significant effect (all $t(547) < 1.35$, $p > 0.18$). That was true even when we analyzed the effect of each predictor separately (with three control variables: the target word's frequency, orthographic neighborhood size, and concreteness).

We are cautious about drawing strong conclusions from these exploratory null effects. Nonetheless, they suggest that the overall accuracy benefit for words in compound pairs arises because the two words combine as morphemes into a single known word. The higher-level properties of that compound word do not seem to have strong effects. That informs the explanations provided in the General Discussion that are based on orthographic and lexical activations without appeal to semantic variables.

The *reversed compound* condition is also relevant to this discussion. As shown in Fig. 5A, when attention was divided, accuracy for reversed pairs was intermediate between the 'correctly' ordered compound pairs and the scrambled pairs. Although those differences were not statistically significant, it is worth noting that the compound pairs and the reversed compound pairs had the same semantic similarity (by definition). In terms of surprisal, the reversed pairs were on average intermediate between the scrambled pairs and the correctly ordered compound pairs, because surprisal takes order into account (considering the top word to be first). Thus, although semantic similarity may have some small effect, the data altogether are most consistent with an explanation based on prior experience with letter strings in a particular order. The slight benefit for reversed compound pairs (relative to scrambled pairs) could also be due to the fact that neither arrangement with words above/below fixation is the standard presentation for these unspaced compound words. We return to this issue in the General Discussion.

Another unexpected result is that divided-attention accuracy for the "random pairs" was significantly *below* the serial model's prediction (upper left of Fig. 5C). This suggests that participants could recognize only one word per trial, but they were slightly more likely to make an error in reporting it than in the focused-attention trials. This additional loss of accuracy is not easy to explain, especially since it did not occur for the *scrambled* compound pairs, which were also unrelated to each other. But it is the scrambled compound condition that provides the best control to compare against the compound word condition. The words in each scrambled pair were also unrelated to each other (like in the "random pairs" condition) and came from the same word set as those in the compound condition. In the scrambled condition, mean accuracy fell directly on top of the serial model's prediction.

To summarize the key theoretical point: the existence of a difference in accuracy between the compound and scrambled pairs proves that the simple serial model does not strictly hold. It is not the case that only one word was ever processed at all. However, accuracy for compounds was far from the *independent* parallel prediction (see the bottom right of Fig. 5C). This could be because the facilitation by compound pairs is relatively weak, provided by a limited-capacity form of parallel processing. Alternatively, the fact that accuracy for compound words was intermediate between model predictions could reflect averaging over a mixture of trials with different outcomes: many trials in which only 1 word was processed at all (consistent with the serial model) and a smaller number of trials in which both constituent words were processed in parallel with no deficit, fully benefitting from the holistic recognition of the compound. Also, the relatively small benefit for compound words in this experiment, compared to Experiment 1b's compound word judgment task, could be partly explained by the lower probability of word pairs forming a compound in Experiment 2 (15 % vs 50 %).

6. General discussion

6.1. Summary

In two experiments, we found that when two words combine to form an existing compound word, they can, at least to some degree, pass together through the theorized "serial bottleneck" in the word recognition system (White, Boynton, & Yeatman, 2019).

The dual lexical decision task in Experiment 1a supported the serial model: two *unrelated* words could not be fully processed (to the point of lexical access) simultaneously. This is consistent with several previous studies (White et al., 2018, 2020; White, Palmer, et al., 2019). However, the same participants, in the same time-limited conditions, were able to perceive when the two words flashed together formed an existing compound word (Experiment 1b). That result is evidence for parallel or holistic processing of the two halves of a compound word, even when presented on opposite sides of fixation.

The second experiment demonstrated that such parallel processing occurs even when participants do not attempt to combine the two letter strings into one word. In Experiment 2, we presented compound word pairs on a random minority of trials during a task that motivated participants to process two words independently. Compound words were rarely noticed by the participants. The serial model again accounted for the low accuracy when participants divided attention between two unrelated words, but it could not account for the rise in accuracy when the two words formed a compound. Thus, while the data for unrelated word pairs suggests that there is a "serial bottleneck" in word recognition, compound words seem to defy its constraints.

It is remarkable that participants were sensitive to the presence of the compound pairs when the interval between the words' onset and the post-masks was so short: on average, it was only 37 ms in Experiment 1b and 45 ms in Experiment 2. That limited input time may constrain the depth of processing that could play a role in generating effects. In the masked priming literature, it has been suggested when a word is available for such a short duration, there can be orthographic and morphological processing but only weak semantic activation, if any (Rastle et al., 2000). That is one reason why, in the explanation proposed in the next section, we invoke only orthographic and lexical representations.

None of the three capacity limit models tested in the attention operating characteristics can account for the entire pattern of results. The all-or-none serial model could account for accuracy in some conditions but not others. Thus, a more sophisticated model is needed. We now outline two possibilities.

6.1.1. Explanation based on an interactive activation model

The first explanation is based on *interactive activation* models of word recognition (McClelland & Rumelhart, 1981). This type of model assumes that written words (the input) activate "letter units," which in turn activate "lexical units" or representations for single words. Such a model is diagrammed in Fig. 6, where there are two sets of letter units, one for the top word's position and one for the bottom word's position. The activation strength of each unit is graded. Each lexical unit corresponds to a string of letters in the orthographic lexicon but does not

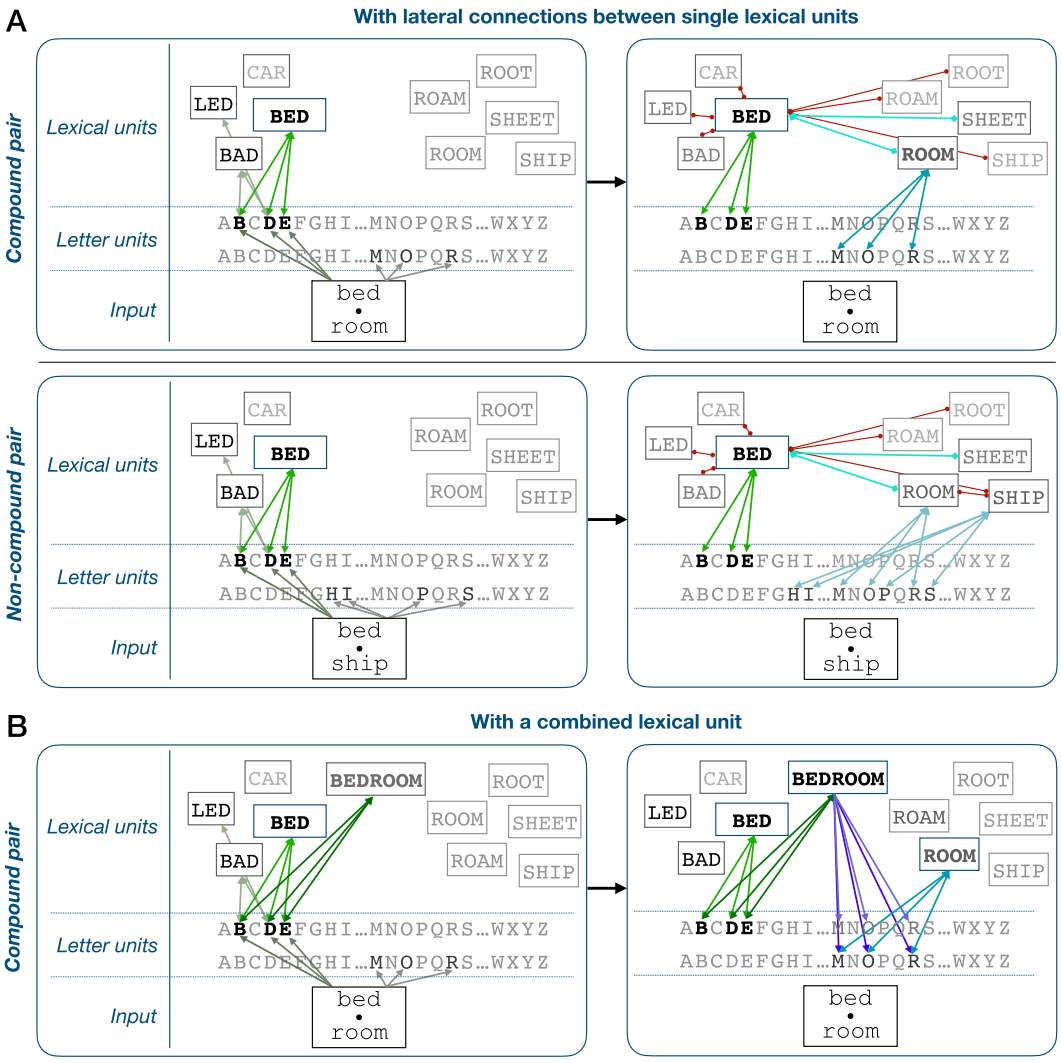

Fig. 6. Sketch of how two different versions of an interactive activation model would respond to a compound pair of words (bed + room) or a non-compound pair (bed + ship). For each pair of words there are two panels that represent two points in time, from left to right. Stronger activations in letter and lexical units are represented by darker and thicker font. **(A)** A version of the model in which there are excitatory lateral connections between constituent lexical units that can combine into a known compound. Excitatory connections between lexical units are represented in teal with square endpoints; inhibitory connections are represented in dark red with round endpoints. **(B)** A version of the model with a dedicated lexical unit for the combined compound word (BEDROOM). Feedback from that lexical unit is key to activating both constituent lexemes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

specify semantic meaning on its own. (For example, there is only one lexical unit for the string "ship", but that string could, in different contexts, refer to a type of boat or to the act of transporting goods). The activation of each lexical unit scales with the overlap between the letters that its word contains and the letter units that are active. Importantly, each active lexical unit then *feeds back* to the letter level, further activating the units for the letters contained in that word. Processing in this model is "cascaded," meaning that activation continuously flows from one level to the next, and back, rather than each unit needing to first reach a threshold (McClelland, 1979). Such models allow multiple iterations of activation across levels of representation before any individual words are definitively identified (Wen et al., 2019).

Beyond that generic architecture, the model's processing capacity is shaped by four key features:

(1) An initial *parallel* stage of orthographic activations across both words, which may have limited capacity, meaning that letter units are activated less strongly when attention is divided than focused;
(2) A subsequent *serial* stage of lexical activation by only one input letter string at a time;
(3) Lateral connections between lexical units, which can be either excitatory or inhibitory. We assume that most are inhibitory, except for excitatory connections between pairs of lexical units that are morphologically compatible. Two lexical units are compatible if they are known to combine side-by-side as constituents or morphemes of a larger unit. (Other forms of "compatibility", in terms of syntax or semantics for example, are theoretically possible but not necessary here).
(4) Interactive activation: feedback from all activated lexical units to their corresponding letter units, followed by forward activation again to the matching lexical units.

Fig. 6A shows how such a model would respond to a compound pair of words like "bed room," and to a non-compound pair like "bed ship." The initial parallel stage allows some activation of all the letter units for the visible letters in both words. Because of limited capacity, one word activates its letter units more strongly ("bed" in this case). Then there is

a *serial stage* of lexical activation: only one word's letter units can, in a feedforward manner, activate the lexical units that have those letters. In Fig. 6A, the letters in "bed" activate the lexical unit for BED and others with overlapping letters to a lesser degree, including BAD and LED. The letters in "room" do not activate any lexical units. In the next step (the top right panel in Fig. 6), the unit for BED inhibits other lexical units except for those that it is compatible with, namely "ROOM" and "SHEET" (because bedroom and bedsheet are known compounds). The teal lines indicate excitatory lateral connections, and the red lines indicate inhibitory lateral connections. Then, the activated lexical units feed back to activate their corresponding letter units, which include the R, O and M from "ROOM". This feedback activation therefore potentiates the weak parallel bottom-up activation from the second input word "room" and can then iteratively strengthen the activation of the lexical unit for ROOM. Thus, despite an initial serial stage of lexical activation, the lexical units for both constituent words are activated.

The situation is different for a non-compound pair like "bed + ship," as shown in the second row of Fig. 6A. The initial stages are the same, with parallel activation of all letter units and then lexical activation only by one constituent's letters ("bed"). Again, BED laterally activates compatible lexical units such are ROOM and SHEET, which feed back to their input letter units. However, there are no bottom-up signals to amplify for those letters, and thus there is a smattering of weak activations across the letters in "ship" and "room," which could weakly activate corresponding lexical units. But no one lexical unit is much more active than the others, and most of them mutually inhibit each other. The correct lexical unit SHIP is no more likely to reach a threshold for report than any others, and thus the initial serial bottleneck in lexical activation prevents both words from being identified.

How would this model account for performance in the experiments reported above? First, Experiment 1a's dual lexical decision task with randomly paired words: this corresponds to the situation in the bottom row of Fig. 6A. The initial serial stage of lexical activation means that only one of the two words can activate the correct lexical unit, and rarely is a compatible lexical unit activated that could feed back to amplify the correct letter units for the other word. In fact, the inhibitory connections between most lexical units can help explain why accuracy is so low for non-compound pairs (across both experiments). The initial strong activation of one unit shuts down any weak activation of the other, incompatible word's lexical unit.

Second, Experiment 1b's compound word judgment task: when the two words *do* form a compound word, one lexical unit is activated strongly first, which then laterally activates its partner's lexical unit, which then initiates a round of interactive activation with the letters in the second constituent. In order for the participant to decide that the correct answer is "compound," there must be an additional stage when both lexical units activate a single representation, or the participant reads out that there were two lexical units active rather than one, and judges that they are compatible. The positive effect of the compound words' concreteness that we observed in this experiment could emerge from this late stage of processing. When the two words are a scrambled pair like "bed ship", only one lexical unit is strongly activated, which at a later stage participants can judge is not a compound. The model can therefore explain the high accuracy in the compound word judgment task.

Third, Experiment 2's divided-attention task: when the input is a scrambled pair, only one of the two words' lexical units is strongly active. If the participant is post-cued to report the other word, they must pick from a large set of weakly active word units and are very unlikely to report the correct one. Again, that is because parallel orthographic activation is followed by a *serial* stage of lexical activation. When the input is a compound pair, if the participant is asked to report the word that is activated first, they are very likely to be correct. If they are asked to report the other word, they have a moderate chance of reporting it correctly because of the excitatory lateral connections between lexical units and the interactive activation between lexical and letter units. The probability correct for that second word is less than for the first word (which benefits from the initial bottom-up lexical activation), but it is greater than when the words do not form a compound. This model can therefore explain why divided-attention accuracy for compound words was moderately above the serial model's prediction in Experiment 2.

Could this model explain accuracy in the *reversed* compound condition of Experiment 2? As described so far, the model would predict equivalent accuracy for targets in reversed and correctly ordered compound pairs, because any one constituent lexical unit would activate its partner, regardless of where on the screen either input word was located. Empirically, the data are somewhat ambiguous. On the one hand, proportion correct in the reversed compound condition did not significantly exceed the serial model (see Table 2). On the other hand, accuracy with divided attention was only slightly and not statistically lower in the reversed compound condition than in the correctly ordered compound condition. And when measured in units of p(letters correct), accuracy in the reversed compound condition *did* significantly exceed the serial model (See Supplementary Fig. S4). Thus, the data might match the model's basic prediction. In order to explain the trend for worse accuracy in the reversed than correctly-ordered compound condition, two additional assumptions are needed: (1) The excitatory lateral connections between lexical units are directional and stronger from the 1st constituents to the 2nd constituents of a compound than in the reverse direction; (2) Participants are biased, at least slightly, to process the top word first when dividing attention. Thus, for reversed compound pairs, the benefit of interactive activation would arise primarily on the relatively uncommon trials when the bottom word, which is the first constituent, gets processed first.

There is another way to implement interactive activation of both lexical units: via a single lexical unit for the whole compound words. For this idea we give credit to Joshua Snell. This variant of the model is diagrammed in Fig. 6B, showing its response to the compound pair "bed + room." This model does not assume lateral activations between lexical units. Rather, the letter units from one input word *partially* activate the compound's lexical unit, which then feeds back to amplify the letter units that were weakly activated by the second constituent word. Those then iteratively activate the lexical unit for that second word. Thus, indirectly, both lexical units are active.

This variant of the model could, in theory, make similar predictions as the first variant. It is consistent with the "multi-constituent unit" (MCU) hypothesis, which has been previously advanced to explain fixation patterns in the presence of *spaced* English compounds (Cutter et al., 2014) and parafoveal processing during Chinese reading (Zang, 2019; Zang et al., 2024). As Zang (2019) said, "If words that are processed in parallel comprise an MCU, and MCUs are represented and stored in the lexicon as a single lexical entry, then in fact any demonstration of parallelism over the constituent elements of a MCU need not violate serial processing assumptions." Thus, this model may provide some resolution to the parallel vs. serial debate.

The two variants of the interactive activation model in Fig. 6 differ in how prior knowledge of compound words is implemented: either with lateral connections between constituents, or with dedicated lexical units for the whole words. Our data cannot rule out one in favor of the other. It is possible that both variants are valid. For instance, whole-word lexical units might exist for opaque compounds but not for transparent compounds, because in the latter case the meaning can be derived from the constituents.

These are just sketches of an interactive activation model that provides, in theory, an explanation for our results. A full model could have more detailed mechanics, such as: a specific timecourse of interactive activations (which is relevant for understanding the role of the backwards masking); intermediate units for common sublexical letter combinations (e.g., bigrams or morphemes); and specific strengths of inhibitory or excitatory connections between units (e.g. Snell et al., 2018). The lexical units for constituent words could be considered to lie at a lower 'level' than a unit for the compound word, alongside

representational units for bound morphemes (such as "ness" in darkness; Rastle et al., 2004). In that case, the model could also include feedforward and feedback connections directly between the constituents and the compound's lexical unit, as well as laterally between constituent units. Different strengths of connections between constituent and compound units could account for any effects of surprisal (co-occurrence familiarity) and semantic transparency. However, the core idea works even without that additional hierarchical structure.

Overall, this model includes a mix of parallel processing at one stage, serial processing at another, and then "interactive activation" that allows word pairs that are part of an existing compound to be processed more effectively despite the serial stage. This type of "cascaded parallel processing" blurs the distinction between strictly serial processing of one word at a time and independent parallel processing of two words at once.

It is also important that the semantic relation between the two constituents is not explicitly represented in either variant of the model. That choice is justified by the absence of effects of semantic similarity in our data, and weak or absent effects of semantic transparency. It is also consistent with the proposal (supported by priming data) that morphological decomposition can occur even for semantically opaque words (Rastle et al., 2004). Representational links between the constituent parts of a word (and between the constituents and the whole word) need not be based in semantics. Thus, based on the model proposed above, we do not propose that any pair of words that *could* form a compound, on the basis of their meaning, would benefit from this kind of processing. For example, a reader could productively infer the meaning of the pair "key + box", but it would not benefit from cascaded parallel processing as much as "key + chain," which is already known to the reader.

6.1.2. Explanation based on unconscious parallel lexical activation followed by serial conscious access

There is at least one alternative model to consider. This model assumes *unconscious* parallel lexical access, followed by serial activation in consciousness (Snell & Grainger, 2019a). In more detail: there can be parallel lexical activation of *any* pair of words, even if they are unrelated. However, the participant is not aware of those activations, meaning that the participant cannot report both lexical representations in our divided-attention tasks. What does reach the level of conscious access is a single higher-level "linguistic unit," which could be a semantic or conceptual representation of a word (assuming that semantic representations can be activated in such brief presentations; see below).

Thus, if two *unrelated* words are presented, both are processed to the lexical level but only one of them can activate a representation that reaches awareness, which is what the participant uses to perform the task. If two words that form a compound are presented, both are processed to the lexical level and then a single unit for the compound is activated, which reaches awareness. With that compound in mind, a participant in our Experiment 1 could easily report that they saw a compound pair. A participant in our Experiment 2 could report either constituent word once they have become aware of the unified compound word.

Thus, this model differs from the interactive activation model in Fig. 6 in terms of when a serial bottleneck arises. The interactive activation model has a serial bottleneck at the first stage of lexical access. In this alternate model, the serial bottleneck arises at a higher level, after parallel activations in the orthographic lexicon. The late bottleneck constrains how many representations can reach consciousness and be reported. Compound words effectively *compress* two representations into one, allowing some information about both constituent words to be reportable.

There are some reasons to prefer the first model (interactive activation between lexical units and constituent letter units) over the second (unconscious parallel lexical activation followed by serial conscious access). The first model can explain most of our data with the minimum of processing stages and the types of feedforward and feedback processing that it assumes have been supported by other experiments. The second model is straightforward if we assume that the "higher-level unit" that reaches consciousness is a *semantic* representation. But priming experiments suggest that words do not strongly activate semantic representations when masked after such brief presentations (Rastle et al., 2000). Nonetheless, to rule out the second model, more investigation is necessary. Masked priming could in fact be used as a tool to test whether two unrelated words are processed in parallel to the morphological or lexical level, even if the observer is not aware of their meanings.

6.1.3. Relation to other claims of parallel processing

Our findings may reconcile the apparent serial result for unrelated word pairs (White et al., 2020) and phenomena such as the "sentence superiority effect" that have been used to argue for parallel processing of multiple words. In the sentence superiority effect, accuracy for identifying single words is higher when they appear in the context of a briefly flashed four-word sentence than in a string of words that do not form a grammatical sentence (Snell & Grainger, 2017). It is important to note that this occurs even when the words in the grammatical sentences are not predictable. It does not seem to depend on combinations of words in known high-probability combinations, which is the key to our results for existing compound words. The sentence superiority effect may instead arise from rapid extraction of syntactic phrase structure (Fallon & Pylkkänen, 2024), rather than parallel representation of combinations of particular words or semantic meaning. Another argument is that the sentence superiority effect is due to biased guessing (Staub et al., 2025). In contrast, the results consistent with parallel processing that we reported in the present study were dependent on words appearing in known combinations, and could not be due to biased guessing.

On the other side of the debate, Brothers (2022) argued for serial processing of words even when they were in meaningful combinations. They adapted the divided-attention paradigm (like used here) but for four-word sentences. Participants had to judge the *grammaticality* of two-word pairs on either the left or right side of fixation. The large drop in accuracy with divided attention supported the serial model, even with words embedded in sentence context. Thus, it may not be the case that all types of syntactic or grammatical structures support parallel processing of multiple words, or that grammaticality judgments have a different processing capacity than lexical or semantic judgments.

Lastly, a recent study with Korean words used a similar divided-attention design as our Experiment 2, but with a semantic categorization task and words positioned to the left and right of fixation (Yoo & Joo, 2025). The two words presented on each trial were either randomly paired, or formed a compound, or were semantically related. Accuracy was lower for random pairs than the other two conditions, but in *all three* conditions accuracy exceeded the serial model's prediction. For the compound and semantically related word pairs, accuracy was closest to the fixed-capacity parallel model's prediction. On the one hand, the benefit for compounds matches what we found here with English readers. On the other hand, we did not find an influence of semantic similarity in our exploratory analyses. Moreover, a key facet of our results is that accuracy for randomly paired words matched the serial model. That places greater constraints on how a benefit for compound pairs might have emerged. In Yoo & Joo's study, there seemed to be some degree of parallel processing even for unrelated words, and thus more opportunity for relations between the words to have an effect. As those authors discussed, there may also be differences between the English and Korean languages that account for the differing results; or, as we address next, an influence of visual field positions.

6.2. Limitations & future directions

This study is the first to measure how well two English words are processed in parallel with divided attention while introducing linguistic

relations between the words. In this case, both words sometimes combined into a known unspaced compound. There are several limitations that may be addressed in the future. The first concerns the applicability of these results to natural reading. We assessed processing capacity with strict experimental control, to determine the limits of what skilled readers can perceive. These findings then constrain models of reading (White, Boynton, & Yeatman, 2019).

However, the positions of the words above and below fixation clearly differed from how English words are placed during reading. As explained in the Introduction, we chose those positions because prior studies found strong evidence for serial processing of two unrelated words positioned there, and because it allows for all the letters to be relatively close to fixation to maximize legibility of all of them. A priori, this choice also minimized the chance that we would find any influence of compound word pairs, as readers only have experience reading those words arranged horizontally (and usually without a space between them). The effect we did find in these 'unnatural' conditions is likely to be even stronger when the words are positioned more like in natural reading. It would also be valuable to place one word directly in the fovea and the other (or others) to the side in the parafovea, more like during natural English reading.

Second, there are open questions about different types of morphologically complex words. For instance, most of our "compound pairs" formed transparent compound words. But some compounds were opaque, to varying degrees. In exploratory analyses we did not find clear evidence that performance in our tasks differed across transparent vs. opaque compound pairs, or between compounds and derived or inflected multi-morphemic nouns. Further investigation may reveal how the impact of those differing representational structures.

Third, it will be important to explore other types of word combinations, beyond compounding. Which types of multi-word combinations are coded in the lexicon as compatible combinations such that they can be activated in parallel? It has been proposed that "multi-constituent units" could include common compounds, phrases, or idioms (Zang, 2019). The experiments reported above were not designed to explore other types of common word combinations, but we did conduct exploratory analyses of the "*surprisal*" of the words in each pair, based on their statistics in a large corpus of books. These analyses provide mixed results: in the compound word judgment task of Experiment 1, low surprisal predicted incorrect responses to non-compound pairs. In other words, when looking for compound pairs, participants were sensitive to the likelihood of a two-word combination even if it did not form compounds. But in Experiment 2, there were no such effects of the surprisal in non-compound pairs. We conclude therefore that if participants ever did detect a familiar combination between those non-compound words, they did so rarely or that had a very weak effect. In contrast, accuracy did significantly exceed the serial model when the words combined into an already existing, unspaced compound word (one that is in the dictionary).

Given these mixed results, we suggest an important avenue for future research would be to conduct experiments like these but inserting word pairs that form common phrases with various types of morpho-syntactic structures (e.g., "yes please", "sunny day"). It is possible that some types of high-frequency phrases also have stored representations that can facilitate parallel processing – implemented either with dedicated "lexical units" or excitatory connections between the consistent units. The protocol in our Experiment 2 provides a way to determine which types of word combinations are represented in such an efficient way. It will be important to account for the diversity of syntactic structures in such multi-word combinations when modeling how parallel processing emerges in more naturalistic reading conditions.

7. Conclusion

The debate about parallel vs. serial processing in reading often focuses on single words as discrete units, defined as letter strings separated by spaces. The results reported here suggest a shift towards evaluating units of text that may encompass multiple letter strings (Cutter et al., 2014; Zang, 2019). Specifically, the data demonstrate that two words may be processed in parallel if they compose a known compound word, even under the conditions when two unrelated words are processed serially. Thus, letter strings can be processed more efficiently when they combine into units the reader has learned before. In other words, processing capacity limits are alleviated by higher-order representations. That emergent parallel processing may be a key component of reading skill.

We suggest two potential explanations: 1) Parallel orthographic processing is followed by serial lexical activation by one word at a time. But learned connections between the "lexical units" belonging to a known compound can amplify the other constituent word's representation as well. (2) Multiple words can be processed to the lexical level in parallel, but conscious perception is limited to one higher-level linguistic unit at a time. The initial parallel lexical stage is only revealed when the two words combine to activate a single representation that can be reported. Further investigation is required to test these hypotheses and their implications for natural reading.

CRediT authorship contribution statement

Amritha Anupindi: Writing – review & editing, Writing – original draft, Investigation, Data curation, Conceptualization. **Liana R. Eisler:** Writing – review & editing, Writing – original draft, Investigation, Data curation, Conceptualization. **Mariam Latif:** Writing – review & editing, Formal analysis. **Vassiki S. Chauhan:** Writing – review & editing, Formal analysis. **Alex L. White:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Acknowledgements

This project was supported by funding from the National Eye Institute (grant R00 EY-029366). We are grateful to Liina Pylkkänen for advice at early stages of this project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2025.106387.

Data availability

All raw data and analysis code are available via the Open Science Framework: https://osf.io/r5cfh/.

References

Andrews, S. (1986). Morphological influences on lexical access: Lexical or nonlexical effects? *Journal of Memory and Language, 25*, 726–740.
Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review, 4*, 439–461.
Barton, J. J. S., Hanif, H. M., Eklinder Björnström, L., & Hills, C. (2014). The word-length effect in reading: A review. *Cognitive Neuropsychology, 31*(5–6), 378–412.
Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience, 17*, 905–917.
Bonnel, A.-M., & Prinzmetal, W. (1998). Dividing attention between the color and the shape of objects. *Perception & Psychophysics, 60*, 113–124.
Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 443–446.
Braun, J., & Julesz, B. (1998). Withdrawing attention at little or no cost: Detection and discrimination tasks. *Perception & Psychophysics, 60*, 1–23.
Brothers, T. (2022). Capacity limits in sentence comprehension: Evidence from dual-task judgements and event-related potentials. *Cognition, 225*, Article 105153.
Brysbaert, M., & New, B. (2009). Moving beyond Kucera 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*, 977–990.
Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904–911.
Campbell, M., Oppenheimer, N., & White, A. L. (2024). Severe processing capacity limits for sub-lexical features of letter strings. *Attention, Perception, & Psychophysics, 86*, 643–652.
Carreiras, M., Perea, M., & Grainger, J. (1997). Effects of orthographic neighborhood in visual word recognition: Cross-task comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 857–871.
Cutter, M. G., Drieghe, D., & Liversedge, S. P. (2014). Preview benefit in english spaced compounds. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 40*, 1778–1786.
Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*, 171–185.
Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112*(4), 777–813.
Fallon, J., & Pylkkänen, L. (2024). Language at a glance: How our brains grasp linguistic structure from parallel visual input. *Science Advances, 10*, Article eadr9951.
Fiorentino, R., & Poeppel, D. (2007). Compound words and structure in the lexicon. *Language & Cognitive Processes, 22*, 953–1000.
Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*, 627–635.
Frisson, S., Niswander-Klement, E., & Pollatsek, A. (2008). The role of semantic transparency in the processing of English compound words. *British Journal of Psychology, 99*, 87–107.
van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*, 1176–1190.
James, C. T. (1975). The role of semantic information in lexical decisions. *Journal of Experimental Psychology: Human Perception and Performance, 104*(2), 130–136.
Ji, H., Gagné, C. L., & Spalding, T. L. (2011). Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque English compounds. *Journal of Memory and Language, 65*, 406–430.
Johnson, M. L., Palmer, J., Moore, C. M., & Boynton, G. M. (2022). Evidence from partially valid cueing that words are processed serially. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-022-02230-w
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.
Krekelberg, B. (2024). *Matlab toolbox for Bayes factor analysis*. https://doi.org/10.5281/zenodo.13744717
Kroll, J. F., & Merves, J. S. (1986). Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(1), 92–107.
Lee, D. K., Koch, C., & Braun, J. (1999). Attentional capacity is undifferentiated: Concurrent discrimination of form, color, and motion. *Perception & Psychophysics, 61*, 1241–1255.
Legge, G. E., Mansfield, J. S., & Chung, S. T. L. (2001). Psychophysics of reading XX. Linking letter recognition to reading speed in central and peripheral vision. *Vision Research, 41*, 725–743.
Libben, G., Gagné, C. L., & Dressler, W. (2020). The representation and processing of compound words. In V. Pirelli, I. Plag, & W. U. Dressler (Eds.), *Word knowledge and word usage* (pp. 336–352). De Gruyter.
McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86*, 287–330.
McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*, 375–407.
Medler, D. A., & Binder, J. R. (2005). MCWord: An on-line orthographic database of the English language. Retrieved from http://www.neuro.mcw.edu/mcword/.
Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227–234.
Michel, J. B., Kui Shen, Y., Presser Aiden, A., Veres, A., Gray, M. K., Pickett, J. P., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*, 176–182.
Onnis, L., Lim, A., Cheung, S., & Huettig, F. (2022). Is the mind inherently predicting? Exploring forward and backward looking in language processing. *Cognitive Science, 46*. https://doi.org/10.1111/cogs.13201
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.
Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).
Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under the ROC-curve and of d'e. *Psychological Bulletin, 71*, 161–173.
Prins, N., & Kingdom, F. A. A. (2009). *Palamedes: Matlab routines for analyzing psychophysical data*.
Rastle, K., Davis, M. H., Marslen-Wilson, W. D., & Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language & Cognitive Processes, 15*, 507–537.
Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin and Review, 11*, 1090–1098.
Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506.
Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research, 7*, 4–22.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.
Sandra, D. (1990). On the representation and processing of compound words: Automatic access to constituent morphemes does not occur. *The Quarterly Journal of Experimental Psychology Section A, 42*, 529–567.
Scharff, A., Palmer, J., & Moore, C. M. (2011). Extending the simultaneous-sequential paradigm to measure perceptual capacity for features and words. *Journal of Experimental Psychology: Human Perception and Performance, 37*, 813–833.
Schwanenflugel, P. J., & Stowe, R. W. (1989). Context availability and the processing of abstract and concrete words in sentences. *Reading Research Quarterly, 24*, 114–126.
Snell, J., & Grainger, J. (2017). The sentence superiority effect revisited. *Cognition, 168*, 217–221.
Snell, J., & Grainger, J. (2019a). Consciousness is not key in the serial-versus-parallel debate. *Trends in Cognitive Sciences, 23*, 814–815. Elsevier ltd.
Snell, J., & Grainger, J. (2019b). Readers are parallel processors. *Trends in Cognitive Sciences, 23*, 537–546.
Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. *Psychological Review, 125*, 969–984.
Sperling, G., & Melchner, M. J. (1978). The attention operating characteristic: Examples from visual search. *Science, 202*, 315–318.
Staub, A., Deutsch, E., Greene, J., & Hammond, J. (2025). The 'sentence superiority effect' is due to guessing. *Cognition, 263*(June), Article 106202. https://doi.org/10.1016/j.cognition.2025.106202
Taft, M., & Forster, K. I. (1976). Lexical storage and retrieval of polysyllabic words. *Journal of Verbal Learning and Verbal Behavior, 15*, 607–620.
Torgesen, J., Rashotte, C., & Wagner, R. (1999). *TOWRE-2: Test of word Reading efficiency, 2nd Ed*.
Veldre, A., Reichle, E. D., Yu, L., & Andrews, S. (2023). Lexical processing across the visual field. *Journal of Experimental Psychology: Human Perception and Performance, 49*, 649–671.
Vidal, C., Content, A., & Chetail, F. (2017). BACS: The Brussels artificial character sets for studies in cognitive psychology and neuroscience. *Behavior Research Methods, 49*, 2093–2112.
Wen, Y., Snell, J., & Grainger, J. (2019). Parallel, cascaded, interactive processing of words during sentence reading. *Cognition, 189*, 221–226.
White, A. L., Boynton, G. M., & Yeatman, J. D. (2019). You can't recognize two words simultaneously. *Trends in Cognitive Sciences, 23*, 812–814.
White, A. L., Palmer, J., & Boynton, G. M. (2018). Evidence of serial processing in visual word recognition. *Psychological Science, 29*, 1062–1071.
White, A. L., Palmer, J., & Boynton, G. M. (2020). Visual word recognition: Evidence for a serial bottleneck in lexical access. *Attention, Perception, & Psychophysics, 82*, 2000–2017.
White, A. L., Palmer, J., Boynton, G. M., & Yeatman, J. D. (2019). Parallel spatial channels converge at a bottleneck in anterior word-selective cortex. *Proceedings of the National Academy of Sciences, 116*, 10087–10096.
White, A. L., Palmer, J., Sanders, G., Hossain, J., & Zabinsky, Z. B. (2025). Negative effects of redundant targets. *Journal of Experimental Psychology: Human Perception & Performance*. https://doi.org/10.31234/osf.io/vwgz4_v2. In press.
Yeatman, J. D., & White, A. L. (2021). Reading: The confluence of vision and language. *Annual Review of Vision Science, 7*, 487–517.
Yoo, S., & Joo, S. J. (2025). Korean Hangul is more robust to a serial bottleneck: Co-occurring and semantically related Korean words can be processed in parallel. *Journal of Experimental Psychology: General, 154*, 1878–1887.
Yu, L., Cutter, M. G., Yan, G., Bai, X., Fu, Y., Drieghe, D., & Liversedge, S. P. (2016). Word n + 2 preview effects in three-character Chinese idioms and phrases. *Language, Cognition and Neuroscience, 31*, 1130–1149.
Zang, C. (2019). New perspectives on serialism and parallelism in oculomotor control during reading: The multi-constituent unit hypothesis. *Vision, 3*. https://doi.org/10.3390/vision3040050
Zang, C., Wang, S., Bai, X., Yan, G., & Liversedge, S. P. (2024). Parafoveal processing of Chinese four-character idioms and phrases in reading: Evidence for multi-constituent unit hypothesis. *Journal of Memory and Language, 136*, Article 104508.