



Iwan, Oliwia Anna & Wittenberg, Eva. 2026. Compositional parsing in adjective-noun phrases: The role of adjective semantics. *Glossa: a journal of general linguistics* 11(1). pp. 1–31. DOI: <https://doi.org/10.16995/glossa.23979>



## Compositional parsing in adjective-noun phrases: The role of adjective semantics

Oliwia Anna Iwan, Central European University, Vienna, Austria, [Iwan\\_Oliwia@phd.ceu.edu](mailto:Iwan_Oliwia@phd.ceu.edu)

Eva Wittenberg, Central European University, Vienna, Austria, [WittenbergE@ceu.edu](mailto:WittenbergE@ceu.edu)

---

A central question in language comprehension is how the mind combines the meanings of individual words. Minimal phrases such as adjective-noun pairs (*red cup*, *big boat*) provide a tractable model for studying how linguistic and perceptual features are integrated during comprehension. Previous research has often assumed a uniform compositional process, independent of adjective meaning. Across three experiments, we examined how adjective type (intersective vs. subsective) and word order influence visual verification performance. Experiment 1 used intersective colour adjectives and replicated the established facilitation for double-feature cues, showing faster responses to adjective–noun phrases than to single-word cues. Experiment 2, using subsective size adjectives, produced a similar but weaker pattern, with the advantage no longer significant once individual variability was modelled. Experiment 3 directly contrasted both adjective types within participants and revealed slower overall responses for subsective adjectives but no reliable interaction between adjective type and feature number. Taken together, these results indicate that the apparent “compositional advantage” in this paradigm depends on the semantic transparency of the features involved rather than on a categorical compositional mechanism. Intersective adjectives provide directly alignable perceptual information, yielding faster visual–linguistic integration, whereas subsective adjectives rely on context-dependent interpretation and show less consistent facilitation. Word order did not systematically affect comprehension in any experiment.

---

*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by the Open Library of Humanities. © 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 OPEN ACCESS



## 1 Introduction

A defining characteristic of human cognition is the capacity to generate an unbounded array of complex ideas through the combination of elements from a finite set of discrete building blocks (Humboldt 1876; Chomsky 2006). Consequently, compositional representations are widely recognized across diverse cognitive domains, from social and visual cognition to language. They are associated with efficient learning, enabling productive thought and generalization to novel contexts (Phillips & Wilson 2011; Kemp 2012; Chang et al. 2018). Understanding how compositional representations are formed and maintained in working memory is particularly crucial for capturing the dynamic nature of language processing, where we must rapidly combine individual word meanings into a complex whole that can be readily interpreted and understood. In this paper, we investigate how specific exemplars of intersective (e.g., *red*) and subsective (e.g., *big*) adjectives, together with order of presentation (e.g., *red boat* vs. *boat red*), influence the processing of adjective–noun combinations in visual matching tasks.

## 2 Compositionality in language

Compositionality, though evident across diverse cognitive domains (perhaps with domainspecific characteristics; Barsalou 2017), holds a particularly significant, and often debated, position in the study of language (Chomsky 1956; Jackendoff & Pinker 2005; Barker & Jacobson 2007). While the principle of compositionality, which states that the meaning of a complex expression is determined by the meanings of its parts and the rules governing their combination, provides a theoretical framework, its application to real-world language use faces numerous challenges (Chomsky 1956; Frege 1963; Partee 1995). For instance, phenomena like compounds, idioms, or semi-transparent constructions such as light verbs (e.g., Wittenberg 2016; Ziegler, Snedeker & Wittenberg 2018) and metaphors (e.g., Holyoak & Stamenković 2018) seem to challenge the straightforward application of compositional principles. The frequency of these phenomena underscores the continuing need for empirical investigation into the underlying mechanisms of compositionality in language. Nevertheless, the fact that language allows both transparent and non-transparent combinations highlights the importance of understanding when and how compositional mechanisms operate efficiently.

In the domain of simple modification, a growing body of work demonstrates that combining words into simple phrases can facilitate visual target matching. Early work by Potter & Faulconer (1979), for instance, used a picture-probe technique to examine how adjectival information influences the understanding of a noun. Participants heard sentences containing either adjective–noun phrases (e.g., *burning house*) or nouns alone, then saw a picture and indicated whether it matched the previous description. Responses were fastest when the picture and phrase corresponded: when cued with *burning house* participants responded fastest to an image of a burning house, whereas when cued with *house* they responded fastest to an image of a typical house.

The authors concluded that listeners rapidly integrate the meanings of adjectives and nouns to form a unified representation of the modified noun phrase. Comprehension involves accessing this integrated representation, rather than processing each word separately. The meaning of a noun is retrieved in conjunction with its modifying adjective, not independently, especially when the phrase describes a familiar concept. This argues against a model where noun meanings are retrieved in isolation and then combined with adjective meanings. Instead, the adjective influences retrieval of the noun's meaning from the outset.

Later studies (many designed primarily to identify the neural correlates of composition) used similar minimal-phrase tasks and yielded converging behavioral patterns. In experiments where participants indicated whether visual stimuli matched written cues (e.g., *red boat* vs. *xgf boat*) reaction times were reliably faster for two-word expressions (Bemis & Pylkkänen 2011; 2013a; 2013b). These effects demonstrate a consistent behavioral facilitation for matching minimal compositional phrases to visual targets compared to single words.

Subsequent behavioral work refined this phenomenon by identifying factors that constrain or modulate the compositional advantage. The facilitation observed for minimal phrases was found to diminish as phrase complexity increased, suggesting additional processing demands when relations among constituents become more intricate (Rabagliati, Doumas & Bemis 2017). The timing of feature presentation also affects the advantage: sequential presentation of features leads to faster responses than simultaneous presentation, although prolonged exposure to simultaneous features can ultimately produce a similar effect (Bocanegra, Poletiek & Zwaan 2022).

These behavioral patterns have motivated representational accounts that explain how multiple features are integrated during comprehension. According to these accounts, individuals form conjunctive (intersective) representations, treating multiple features as jointly necessary (such as combining *red* AND *square* into a single integrated concept) rather than maintaining them as fully independent throughout the matching process. Such representations create a precise mental target, allowing faster matching, with a congruent target, whereas disjunctive processing would require additional verification and thus slower responses (Rabagliati, Doumas & Bemis 2017). Evidence from visual working memory supports this view: Visual representations include bound conjunctions of features rather than separate attributes (Luck & Vogel 1997; Allen et al. 2006; Brockmole et al. 2008; Logie et al. 2009). Attending to a single feature tends to slow down responses, whereas linguistic cues specifying multiple bound features facilitate identification. The alignment between linguistic and visual representations therefore provides a plausible mechanism for the compositional advantage observed across these tasks.

Neuroimaging evidence provides converging support for rapid composition processes. Magnetoencephalography (MEG) studies have identified increased activation in the left anterior temporal lobe (LATL) approximately 200–250 ms after noun onset when participants process minimal adjective-noun phrases compared with unstructured or anomalous controls (Bemis & Pylkkänen 2011; Pylkkänen et al. 2014; Ziegler & Pylkkänen 2016). This “LATL-composition

effect” has been replicated across languages and stimulus types and is typically interpreted as a neural signature of early combinatorial integration. While these findings do not specify representational mechanisms, they converge with the behavioral results summarized above in suggesting that even minimal expressions engage rapid and efficient compositional processing.

### 3 Generalizability across types of composition

While, as we have seen before, adjective-noun phrases containing intersective adjectives are processed faster than individual words, this advantage is modulated by several factors, including representational complexity, processing dynamics, and presentation format. The question, therefore, remains whether this advantage extends beyond commonly studied kinds of composition such as those involving, for example, colour terms. In particular, while there is robust behavioral evidence for a compositional advantage with intersective colour adjectives (Bemis & Pylkkänen 2011; Bemis & Pylkkänen 2013a; Bemis & Pylkkänen 2013b; Rabagliati, Dumas & Bemis 2017; Bocanegra, Poletiek & Zwaan 2022), it remains unclear whether this advantage generalizes to other adjective classes, particularly those denoting different types of object features or dimensions.

Clarifying these processing asymmetries requires a closer examination of how adjective-noun composition is treated in formal semantics. Intersective adjectives are typically distinguished from a broader class of non-intersective adjectives, which includes subsective cases such as *big* or *small*, as well as privative and modal modifiers. The current study focuses on this subsective subclass, which, unlike intersective adjectives, cannot be interpreted in the same straightforward way (Partee 1995; Kennedy 2007; Toledo & Sassoon 2011). Intersective adjectives, such as *red* in *red boat*, contribute a property that composes directly with the noun’s meaning. Subsective adjectives, by contrast, require contextual interpretation; their meaning depends on properties that are evaluated relative to the noun’s extension or a relevant comparison class (Kennedy 2007; Solt 2009; Lassiter & Goodman 2017). While different classificatory schemes emphasize different aspects of adjectival meaning, they converge on the idea that some adjectives modify nouns independently, whereas others require contextual calibration.

In formal semantic theory, intersective adjectives are typically analyzed as predicates of type  $\langle e, t \rangle$ , which combine with nouns via predicate modification, yielding the intersection of their denotations (Heim & Kratzer, 1998). Subsective adjectives, in contrast, cannot be interpreted via predicate modification and are therefore classified as a subset of the broader non-intersective category. This category also encompasses adjectives such as *fake* or *alleged*, which are nonsubsective but share the key property of failing to combine via simple intersection, instead shifting or restricting the noun’s reference. These adjective classes can be distinguished using standard semantic diagnostics (Partee 1995; Kennedy 2007). For instance, intersective adjectives such as *red* pass the predicate conjunction test: From “That is a red boat,” one can validly infer both “That is red” and “That is a boat.” Subsective adjectives such as *big*, by contrast, fail this

test: From “That is a big boat,” one cannot infer “That is big,” since *bigness* is evaluated relative to the noun’s comparison class. Similarly, only intersective adjectives support set-intersection readings, whereas subjective adjectives require context-dependent threshold evaluation.

These diagnostics highlight that the two classes differ not only in their logical type but also in their inferential and interpretive behaviour. Subjective adjectives in particular introduce contextual sensitivity, often modeled via comparison-class variables or scalar thresholds. This prevents them from functioning as simple  $\langle e, t \rangle$  predicates and motivates analyses that treat them as higher-order functions (Kamp 2013), mapping predicates to predicates or introducing additional structure such as intensional or event-based arguments. These compositional differences, whether in semantic type, functional structure, or context dependence, reflect fundamental distinctions in how adjectives integrate with nouns.

In the present study we operationalize this distinction by contrasting colour adjectives (e.g., *red, green*), which exemplify intersective modification, with size adjectives (e.g., *big, small*) which instantiate subjective, context-dependent modification. This pairing provides an illustrative instantiation of the broader intersective-subjective contrast, enabling an empirical comparison of how these adjective types pattern in task performance. Importantly, colour and size adjectives differ not only in intersectivity but also in other semantic dimensions (including gradability, context sensitivity and perceptual feature type) which are likely to jointly shape their processing profiles (Kennedy & McNally 2005; Solt 2015).

Because intersective and subjective adjectives differ in how they semantically combine with nouns, this distinction also predicts differences in how adjective-noun combinations are processed in real time. The present study investigates whether these theoretically grounded differences surface in real-time comprehension, using visual matching tasks to test how colour and size adjectives combine with nouns during language processing.

Magnetoencephalography studies have revealed that combinatory LATL effects arise selectively for context-insensitive intersective adjectives (e.g., *dead*), but not for scalar, context-dependent ones (e.g., *large*, Ziegler & Pyllkkänen 2016). These findings have been interpreted as supporting a two-stage model of semantic composition, in which early, rapid integration is characteristic of intersective adjectives, while subjective ones undergo delayed, contextsensitive interpretation. These results align with the hypothesis that intersective adjectives afford more efficient composition during language processing. Complementary evidence from eye-tracking-studies further shows that the integration of adjectival meaning is modulated by contextual demands and adjective class. In visual-world-paradigms, colour adjectives permit rapid referent identification, whereas gradable adjectives (both relative and absolute) yield later or more variable facilitation patterns. When contrastive visual contexts are available, facilitation emerges earlier for relative adjectives but only later for absolute ones. This asymmetry reflects the fact that absolute adjectives rely on fixed scalar standards and require threshold-based evaluation, which introduces additional processing demands (Aparicio, Xiang & Kennedy 2015).

Developmental findings reveal a comparable hierarchy: intersective adjectives such as colour terms are processed more efficiently and accurately than relative or absolute adjectives, with children gradually approaching the adult pattern of earlier adjective integration (Redolfi & Melloni 2025). While compositional advantages have been consistently observed for minimal phrases, these effects have been tested almost exclusively with intersective adjectives describing perceptual features such as colour. It remains an open question whether similar facilitation occurs for context-dependent adjectives such as size terms. The current study directly compares these classes, assessing how adjective type and word order affect behavioral performance in a visual matching paradigm.

## 4 Current studies

In what follows, we present three experiments using a visual matching task to investigate how the semantic type of adjectives, specifically the distinction between intersective and subjective modification, affects compositional processing. Building on previous work by Bocanegra et al. (2022), we ask whether the compositional advantage (faster responses to two-word cues compared to single-word cues) arises uniformly across adjective types, or whether it reflects deeper semantic differences in how adjective-noun combinations are interpreted.

Experiment 1 replicates Bocanegra et al.'s (2022) Experiment 2b using intersective adjectives, specifically colour terms. Participants are shown linguistic cues consisting of either one feature (e.g., *red* or *square*) or two features (e.g., *red square*). In dual-feature trials, we further manipulate word order by comparing an adjective-first sequence (*red square*) to a noun-first sequence (*square red*). We predict that intersective adjective-noun combinations will yield a compositional advantage, replicating prior results. We also examine whether word order influences performance in this minimal compositional context. If modifiers require a referent to be fully integrated, then noun-first order might support more efficient composition. Alternatively, if processing is shaped by structural markedness or linear parsing expectations, adjective-first order may produce faster responses.

Experiment 2 extends this design to subjective adjectives such as *big* or *small*, which do not contribute independent intersective properties but rather depend on the noun for interpretation. This experiment tests whether a compositional advantage can arise when adjective-noun meaning is computed through context-sensitive, function-argument composition. If the previously observed advantage is tied specifically to predicate conjunction, then we should not observe facilitation for subjective phrases. However, if other compositional mechanisms can support efficient integration, a processing benefit may still emerge. We also explore word order effects. Because subjective adjectives depend more strongly on the noun for interpretation, presenting the noun first could potentially support more efficient composition. On the other hand, if a noun-first sequence introduces additional processing cost, we may see slower responses in that condition.

Experiment 3 directly compares intersective and subjective adjectives within a single task. Participants encounter both adjective types in blocked trials, allowing for a direct test of whether the compositional advantage varies systematically with semantic type. This experiment does not manipulate word order and focuses solely on the presence or absence of the advantage across adjective categories. If the efficiency of composition is modulated by semantic type, we expect a stronger or more reliable advantage for intersective than for subjective modification. Together, these experiments investigate whether the observed compositional advantage is a general feature of adjective-noun composition or whether it is modulated by differences in semantic type that are central to formal theories of modification. The results speak to the interface between lexical semantics and real-time language comprehension.

## 5 Experiment 1: Colour adjectives (replication of Bocanegra et al. 2022)

Experiment 1 examined the speed of visual verification following linguistic descriptions, replicating the design of Bocanegra et al., (2022). Participants were presented with images of varying shapes and colours and tasked with rapidly identifying targets matching either singlefeature (e.g., *red*) or double-feature (e.g., *red square*) descriptions. This design investigated whether the presence of two features facilitates visual search, reflecting efficient integration of word meanings during comprehension.

### 5.1 Participants

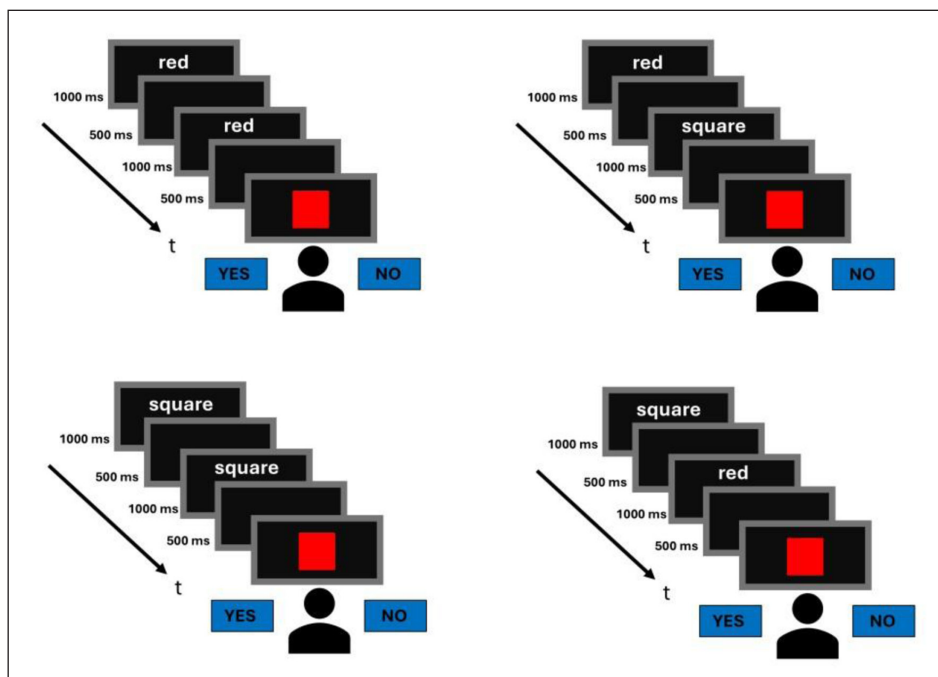
Fifty-nine proficient English speakers with normal or corrected-to-normal vision were recruited through [university name redacted] Research Participation System and were compensated 5 Euros for their participation. To obtain the pre-registered sample of 48 complete data sets, additional participants were tested to replace cases with missing data. Participants were on average 25 years old.

### 5.2 Materials and procedures

The study was conducted in a laboratory setting. Stimuli were presented and responses recorded using PsychoPy (version 2023.2.3). Closely following Bocanegra et al.'s design, the experiment used the same adjective and noun sets (*red, green; diamond, square*) and trial structure. Each trial began with a 500 ms fixation cross, followed by the linguistic stimulus. In double-feature trials, participants saw an adjective and noun presented sequentially for 1 s each, separated by a 2 s blank screen (e.g., *red diamond* or *diamond red*). In single-feature trials, a single word was presented twice (1 s each, separated by a 2 s blank screen) to match the duration of doublefeature trials. Adjective-noun combinations were presented sequentially for one second per word, separated by a two second blank screen. In single-feature trials, the word was presented twice (for one second each presentation, separated by a two second blank screen) to maintain equal trial durations across all trial types.

After the linguistic stimulus a 2 s blank screen preceded the presentation of a single icon, which remained onscreen until the participant responded. Participants pressed the “A” key to indicate a match and the “L” key to indicate a mismatch between the linguistic and visual stimulus. Mismatch trials were designed such that the presented icon always mismatched the linguistic stimuli on all relevant semantic dimensions. For example, if the linguistic stimulus was *red diamond*, the icon would be a green square. If the linguistic stimulus was *red*, the icon would be either a *green square* or a *green diamond*, balanced across trials. Similarly, if the linguistic stimulus was *diamond*, the icon would be either a *red square* or a *green square*, balanced across trials.

The experiment comprised 128 trials (32 trials  $\times$  4 blocks). The four blocks corresponded to the four conditions (adjective–noun order, noun–adjective order, adjective-only, noun-only); blocks were presented once per participant in randomized order, and trials were randomized within blocks. Each experimental session consisted of an equal number of match and mismatch trials (50% each). On match trials, the visual display corresponded to the linguistic cue (e.g., red square  $\rightarrow$  red square), whereas on mismatch trials, the display and cue did not correspond (e.g., red square  $\rightarrow$  green diamond, see Figure 1). The preregistration is available at <https://doi.org/10.17605/OSF.IO/M4ZPT>.



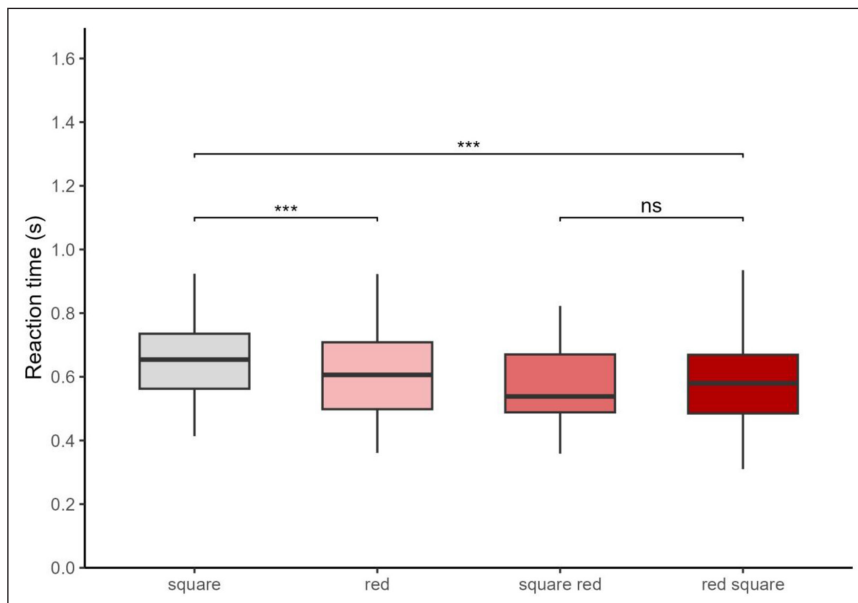
**Figure 1:** Illustration of the task design in Experiment 1. The top-left panel shows an example of a single-feature trial with an adjective cue (*red*) and the bottom-left panel shows a noun cue trial (*square*). The top-right panel depicts a double-feature trial with an adjective-first order (*red square*), and the bottom-right panel shows a doublefeature trial with a noun-first order (*square red*).

### 5.3 Data analysis and results

Accuracy rates for match<sup>1</sup> trials were overall high, with 95.8% and 95.1% correct responses in the double-feature and single-feature trials, respectively. Reaction times faster than 100ms and slower than 2000 ms were excluded and only the mean reaction times from correct trials were analysed (< 1% of all valid trials).

Following Bocanegra et al. (2022), a repeated-measures analysis of variance (ANOVA) was conducted to examine the effects of feature number (single vs. double) and headedness (adjective-first [noun-headed] vs. noun-first [colour-headed] word order) on reaction times.<sup>2</sup> All analyses were conducted in R (R Core Team 2023). ANOVA models were fitted using the afex package (Singmann et al. 2021), which reports generalized eta squared ( $\eta_g^2$ ) as a measure of effect size.<sup>3</sup>

The analysis revealed a significant main effect of feature number,  $F(1, 47) = 15.89, p < .001, \eta_g^2 = .036$ , with slower responses for single-feature trials ( $M = 610$  ms,  $SD = 191$  ms) compared to double-feature trials ( $M = 565$  ms,  $SD = 183$  ms). A significant main effect of headedness also emerged,  $F(1, 47) = 9.93, p = .003, \eta_g^2 = .011$ , with faster responses to colour-headed cues ( $M = 572$  ms,  $SD = 183$  ms) than to noun-headed cues ( $M = 602$  ms,  $SD = 183$  ms, see **Figure 2**).



**Figure 2:** Mean reaction times in Experiment 1 for double-feature trials (adjective-first vs. noun-first order) and single-feature trials (colour vs. noun cues). Asterisks indicate statistically significant effects based on the overall ANOVA and post hoc pairwise comparisons.

<sup>1</sup> A detailed analysis of the correct mismatch trials is provided in the *Supplementary Materials*; results closely paralleled those observed for the match trials.

<sup>2</sup> Some data sets were automatically excluded during ANOVA due to missing values.

<sup>3</sup> Effect sizes are reported as generalized eta squared  $\eta_g^2$ , as recommended for repeated-measures designs (Olejnik & Algina 2003; Bakeman 2005).

The interaction between feature number and headedness was significant,  $F(1, 47) = 4.78, p = .034, \eta_g^2 = .005$ , indicating that the effect of feature number on reaction time also depended on cue headedness. In particular, the cost of single-feature trials was greater when the cue was noun-headed ( $M = 627$  ms,  $SD = 191$  ms) compared to colour-headed ( $M = 592$  ms,  $SD = 190$  ms), indicating that the effect of feature number varied slightly as a function of cue headedness. While all effects reached statistical significance, the effect of feature number ( $\eta_g^2 = .036$ ) was notably the strongest, suggesting that the primary influence on reaction times was the number of features rather than the headedness itself.

To assess the robustness of these effects while accounting for random variation across participants and items we analyzed log-transformed reaction times, using a linear mixed-effects model (LMM) fitted in R (R Core Team, 2023) with the *lme4* package (Bates et al. 2015). The model included fixed effects of feature number (single vs. double), headedness (noun-headed vs. colour-headed), and their interaction, with random slopes for both predictors by participant and random intercepts for items.

The model replicated the ANOVA findings, revealing a significant main effect of feature number,  $t(45) = 3.72, p < .001$  and headedness ( $t(41) = 3.38, p = .002$ ), as well as a significant interaction between the two factors ( $t(2207) = 2.84, p = .005$ ).

Post hoc pairwise comparisons were conducted using the Kenward–Roger method for degrees of freedom estimation, with Tukey adjustments applied to control the family-wise error rate (via the *emmeans* package; Lenth 2023).

Participants responded significantly slower in the single-feature noun-headed condition (e.g., *square*,  $M = 667$  ms) compared to the single-feature colour-headed condition (e.g., *red*,  $M = 626$  ms),  $t(2378) = 4.40, p < .001$ . Reaction times for single-feature noun-headed trials were also significantly slower than for both the double-feature colour-headed condition (e.g., *red square*,  $M = 595$  ms),  $t(2379) = 7.74, p < .001$ , and the double-feature noun-headed condition (e.g., *square red*,  $M = 607$  ms),  $t(2374) = 6.59, p < .001$ .

In contrast, responses to *red* ( $M = 626$  ms) were significantly slower than to *square red* ( $M = 552$  ms),  $t(2379) = 3.46, p = .003$ , but did not differ significantly from *red square* ( $M = 577$  ms),  $t(2378) = 2.20, p = .123$ . Together, these comparisons show that the apparent advantage for double-feature cues was primarily driven by slower responses to nouns presented alone; for adjectives, combining them with a second feature only led to significantly faster responses in adjective-first (e.g., *red square*), but not in adjective second phrases (e.g., *square red*).

## 5.4 Discussion of experiment 1

This study investigated how feature number (single vs. double) and word order (adjective-first vs. noun-first) influence reaction times in a visual feature-matching task. Replicating prior work

(Bocanegra et al. 2022; Bemis & Pylkkänen 2011; 2013a; 2013b; Rabagliati et al. 2017), responses were faster in double-feature than in single-feature trials, indicating a facilitation effect due to redundancy rather than increased syntactic complexity.

Both the repeated-measures ANOVA and linear mixed-effects analyses revealed a robust main effect of feature number, with faster responses when both features were available. Although greater linguistic complexity might intuitively predict slower responses, this result aligns with evidence that redundant or congruent cues can facilitate visual processing and decision making (Rubio-Fernández 2016). A double-feature cue provides a more precise and format-congruent template for the two-dimensional visual target, allowing the linguistic representation to map more efficiently onto the perceptual input. In contrast, a single-feature cue under-specifies the visual object and leaves an irrelevant dimension to be filtered, increasing decision time. The interaction between feature number and headedness reached statistical significance but appears to reflect lexical category differences (adjective vs. noun) rather than a genuine order effect. In single-feature trials, headedness corresponds to lexical category (adjective vs. noun), whereas in double-feature trials it marks word order. Post hoc comparisons confirmed that word order did not reliably affect performance when both features were present. The small interaction likely reflects slower access to shape nouns relative to colour adjectives rather than differences in compositional processing.

The absence of a word-order effect is theoretically informative. Despite English grammar favouring the adjective-noun sequence, participants responded equally quickly to noun-first phrases. This pattern suggests that, under rapid matching conditions, participants relied primarily on feature-based integration of visual information rather than full syntactic composition. Both words independently activated relevant conceptual features, and their linear order had little influence on the efficiency of visual comparison. Because English adjectives lack agreement morphology, both sequences are locally interpretable and can serve as effective retrieval cues. Future work could test this further in morphologically richer languages or with tasks requiring deeper syntactic processing, where word order constraints might exert stronger effects.

The asymmetry between lexical categories (specifically, slower responses to nouns than to adjectives) likely reflects lexical-semantic and perceptual factors. Colour terms are less semantically complex, more visually diagnostic, and more rapidly accessed than nouns denoting shapes or objects (Nicholson & Humphrey 2004; Werning 2010). This interpretation also accounts for the absence of a difference between single-colour and double-feature cues: the inclusion of a salient colour cue in either case already maximizes efficiency.

In sum, the present findings highlight that in rapid visual matching, performance depends primarily on feature-based congruency and lexical salience. The observed “composition effect” is best understood as a behavioural consequence of redundant and perceptually diagnostic cues, rather than evidence of explicit syntactic composition.

## 6 Experiment 2: Size adjectives

The results of Experiment 1 indicated that redundant, converging linguistic cues facilitate efficient visual matching. Experiment 2 replicated this paradigm using size adjectives (*big, small*) instead of colour terms to test whether the facilitation observed in Experiment 1 generalizes across feature dimensions. If the effect reflects a general benefit of redundant feature specification, similar performance patterns should emerge for size-based descriptions. We used a similar verification paradigm as in Experiment 1. Participants judged whether presented icons corresponded to single- or double-feature descriptions. Crucially, this experiment presented only subsective adjectives (e.g., *big, small*), enabling investigation into whether facilitated processing of double-feature descriptions is a universal phenomenon in adjective-noun- phrase processing or restricted to specific adjective types.

We use a limited space (four icons) and simplified size distribution (two size adjectives) to create a processing environment similar to that of intersective adjectives. Thus, the integrated representations formed during composition should yield a processing advantage regardless of adjective type.

### 6.1 Participants

Fifty native English speakers were recruited via Prolific ([www.prolific.co](http://www.prolific.co)) and compensated at a rate of £9.00 per hour. The experiment took an average of 22 minutes to complete. To obtain the pre-registered sample of 48 complete data sets, additional participants were tested to replace cases with missing reaction time data. No demographic information (e.g., age, gender) was collected, as participants were recruited anonymously via Prolific.<sup>4</sup>

### 6.2 Materials and procedure

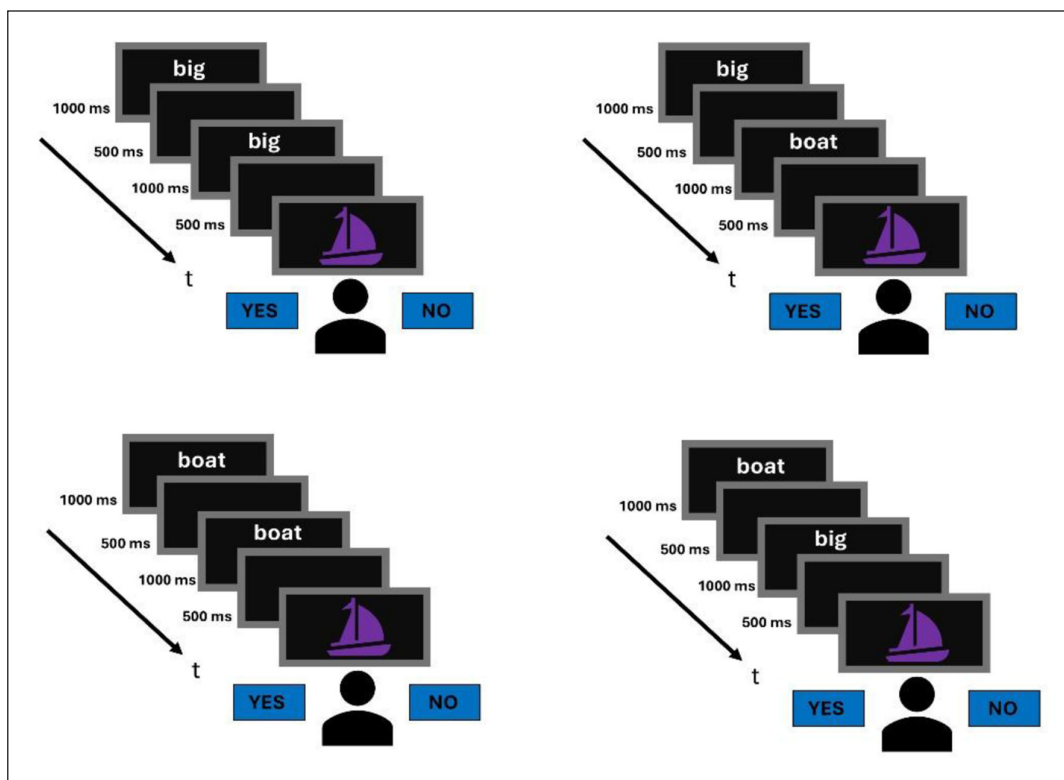
To maintain consistency across experiments, Experiment 2 adapted the replication procedure, utilizing new adjectives (*big, small*), nouns (*car, boat*), and corresponding images. Data collection took place online using Pavlovia (version 2024.1.4).

Simple purple line drawings of these objects in two sizes served as icons. The small car icon was approximately 1.5 cm wide ( $\approx 130$  px), and the large car icon was approximately 10 cm wide ( $\approx 370$  px) when displayed on a  $1920 \times 1080$  pixel screen. The boat icons were similarly scaled.

---

<sup>4</sup> Given the challenges of recruiting native English speakers locally, Experiments 2 and 3 were conducted online. This change in recruitment strategy aimed to mitigate potential confounds arising from non-native language processing in a study investigating subtle semantic effects.

As before, the experiment consisted of 128 trials (32 trials per block  $\times$  4 blocks), equally divided among the four trial types: adjective-first order, noun-first order, adjective-only, and noun-only. The four blocks corresponded to these conditions and were presented once per participant in a randomized order. For double-feature trials (adjective-noun phrases), word order was fixed within each block (adjective-noun vs. noun-adjective). Single-feature trials were presented in dedicated adjective-only and noun-only blocks. Within each block, trials were randomized. Stimuli were presented and responses recorded using PsychoPy (version 2023.2.3) and Pavlovia (platform version 2024.1.4). Each experimental session consisted of an equal number of match and mismatch trials (50 % each). On match trials, the visual display corresponded to the linguistic cue (e.g., *big boat*  $\rightarrow$  big boat, see **Figure 3**), whereas on mismatch trials, the display and cue did not correspond (e.g., *big boat*  $\rightarrow$  small car). The preregistration is available at <https://doi.org/10.17605/OSF.IO/DQUGW>.



**Figure 3:** Illustration of the task design in Experiment 2. The top-left panel shows an example of a single-feature trial with an adjective cue (*big*) and the bottom-left panel shows a single-feature trial with a noun cue (*boat*). The top-right panel depicts a double-feature trial with an adjective-first order (*big boat*), and the bottom-right panel shows a double-feature trial with a noun-first order (*boat big*).

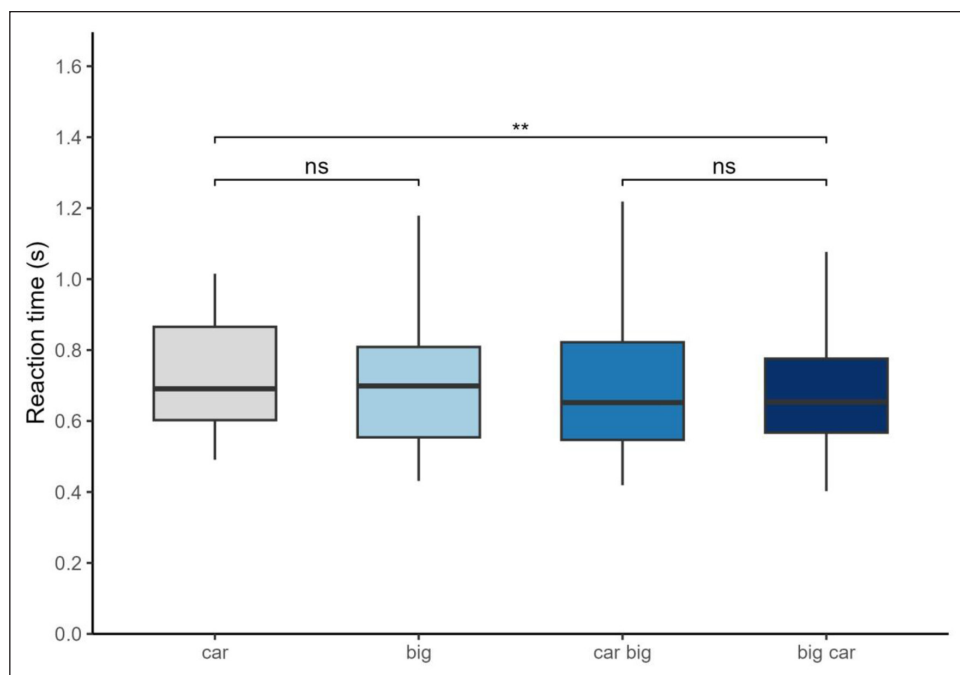
### 6.3 Data analysis and results

Data were analysed using the same procedures and packages as in Experiment 1. Accuracy was high overall. On match<sup>5</sup> trials participants responded correctly on 92.0% of double-feature trials and 89.9% of single-feature trials (24.1% of correct match trials were excluded due to RT trimming < 100 ms or > 2000 ms).

The ANOVA revealed a significant main effect of feature number,  $F(1, 47) = 8.48, p = .005, \eta_g^2 = .014$ , with faster responses in the double-feature condition ( $M = 666$  ms,  $SD = 257$  ms) than in the single-feature condition ( $M = 698$  ms,  $SD = 260$  ms).

No significant main effect of headedness was observed,  $F(1, 47) = 0.46, p = .549, \eta_g^2 < .001$ , indicating no reliable difference in response times between size-headed cues ( $M = 678$  ms,  $SD = 263$  ms) and noun-headed cues ( $M = 692$  ms,  $SD = 243$  ms).

The interaction between feature number and headedness was not significant,  $F(1, 47) = 1.99, p = .17, \eta_g^2 < .001$  (see **Figure 2**). Across conditions, mean RTs were 677 ms for single-size trials, 717 ms for single-noun trials, 666 ms for double-size trials, and 666 ms for double-noun trials, indicating that responses were generally slower for single cues and especially for single nouns (see **Figure 4**).



**Figure 4:** Mean reaction times in Experiment 2 for double-feature trials (adjective-first e.g., *big boat* vs. nounfirst order e.g., *boat big*) and single-feature trials (size e.g., *big* vs. noun cues e.g., *boat*). Asterisks indicate statistically significant effects based on the overall ANOVA.

<sup>5</sup> A detailed analysis of the correct mismatch trials is provided in the *Supplementary Materials*; results paralleled those observed for the match trials.

To verify the robustness of these effects, a linear mixed-effects model was fitted with by-participant random slopes for both predictors and a by-item random intercept.

This model did not yield any reliable fixed effects (all  $p > .20$ ), indicating that once individual variability was explicitly modelled, no contrasts reached statistical significance.

Estimated marginal means (back-transformed from log RTs) ranged from 679 ms (single-size) to 663 ms (double-noun). Pairwise comparisons confirmed the absence of significant differences between any conditions (all Tukey-adjusted  $p > .49$ ). Together, the ANOVA and mixed-effects analyses converge on a numerically consistent but attenuated double-feature advantage. Participants were, on average, faster when both features were present, yet this effect was small and variable across individuals. The lack of a significant interaction or word-type difference suggests that the advantage was not modulated by lexical category (adjective vs. noun) and does not reflect sensitivity to word order.

## 6.4 Discussion of experiment 2

The results of Experiment 2 replicated the visual-matching paradigm of Experiment 1 using size adjectives (*big*, *small*) paired with nouns denoting concrete objects (*boat*, *car*). The repeated-measures ANOVA showed a significant main effect of feature condition, with faster responses for double-feature than single-feature cues. This pattern numerically replicated the facilitation observed in Experiment 1, but the linear mixed-effects model, which incorporated by-participant random slopes, did not show any reliable fixed effects. Thus, once individual variability was fully modelled, the advantage for double-feature cues no longer reached statistical significance.

The attenuation of the effect is theoretically meaningful. Subjective adjectives such as *big* and *small* express relational properties that depend on comparison classes (a *big car* differs from a *big boat*). Unlike intersective colour terms, which map directly onto perceptual dimensions, size adjectives require an internal standard of comparison before evaluation can occur. For subjective adjectives such as *big* and *small*, efficient use of the noun information requires establishing a contextually appropriate comparison standard (e.g., *big for a car* vs. *big for a boat*). This evaluative process cannot be completed until the noun concept is fully accessed, which likely offsets any potential facilitation from having both words present. In other words, while the noun provides essential interpretive information, it also introduces a processing bottleneck: integrating size adjectives with their comparison class is slower and more variable than the direct perceptual mapping available for intersective (colour) adjectives.

As a result, their meanings cannot be mapped directly onto the visual input but must be interpreted relative to context, making the linguistic-visual correspondence inherently noisier. This attenuation is best understood as a consequence of semantic indeterminacy, not of syntactic composition. Whereas colour adjectives denote more fixed, perceptually grounded features,

size adjectives are context-sensitive and therefore provide less stable and less redundant cues for matching. Participants could not simply align each word with an independent visual feature, since the notion of “bigness” or “smallness” is graded and depends on object category. Consequently, double-feature cues offered less additive benefit, because the size term did not contribute a discrete, perceptually verifiable dimension in the way colour did in Experiment 1. Importantly, word order again had no measurable effect, even though subsective adjectives were hypothesised to depend more on grammatical adjective-noun structure for interpretation. If participants had engaged in full syntactic composition to derive phrase-level meaning (e.g., *big boat*), the adjective-first order should have conferred an advantage over *boat big*.

The absence of such an effect suggests that even subsective adjectives did not trigger syntactic composition under these task conditions. Instead, participants relied on direct, feature-based mappings between linguistic and visual representations, with both words processed in parallel as partially independent semantic cues. The additional variability observed for subsective adjectives thus reflects greater ambiguity in conceptual mapping, not the engagement of syntactic mechanisms. Taken together, the results show that the double-feature advantage generalises only weakly to subsective adjectives. Whereas intersective adjectives support rapid, perceptually grounded feature matching, subsective adjectives engage more variable, contextdependent semantics. This pattern thus reinforces the view that the “composition effect” in this paradigm arises from redundant feature overlap and perceptual congruency.

## 7 Experiment 3: Colour and size adjectives within participants

Experiment 3 combined the designs of the previous studies to directly compare intersective and subsective adjectives within the same participants. Using the same image-verification task, participants encountered trials containing both adjective types, allowing a within-subject test of whether the redundancy-based facilitation observed earlier differed by semantic class. Word-order manipulations were omitted to focus specifically on the impact of adjective type on performance in single- and double-feature cues.

### 7.1 Participants

Ninety-nine native English speakers were recruited via Prolific ([www.prolific.co](http://www.prolific.co)) and compensated at a rate of £10.80 per hour. The experiment took an average of approximately 21 minutes to complete. To obtain the pre-registered sample of 98 complete data sets, one additional participant was added to replace a case with missing reaction time data. After exclusions based on pre-registered criteria, 96 complete data sets were retained for analysis. No demographic information (e.g., age, gender) was collected, as participants were recruited anonymously via Prolific.

## 7.2 Materials and procedure

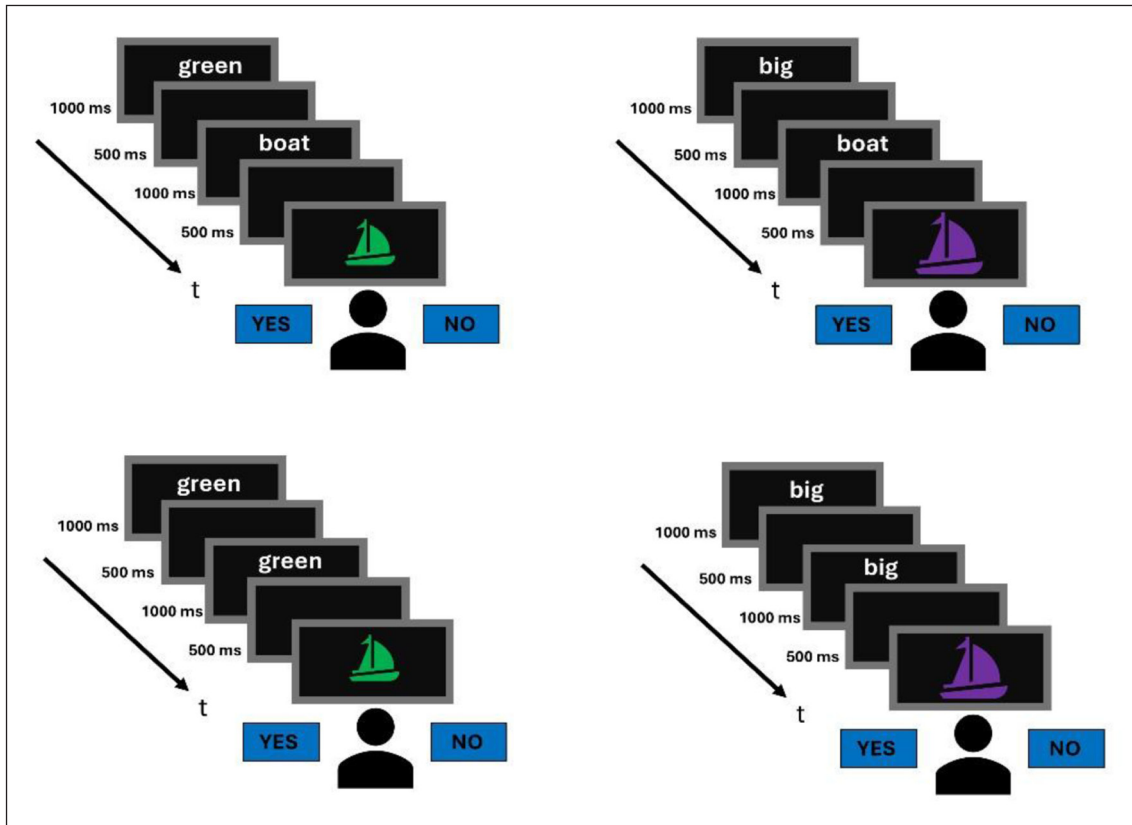
Experiment 3 largely mirrored the design and procedure of Experiments 1 and 2, closely following Bocanegra et al., (2022) but with key modifications. Most importantly, adjective type (colour vs. size) was manipulated within-subjects, allowing for direct comparison of processing differences while controlling for individual variability. Stimuli were drawn from Experiment 2 and the replication study, with *orange* replacing *red* in the intersective condition. Intersective trials used *orange* and *green* while subsective trials used *big* and *small*, both paired with the same nouns (*boat* and *car*). This design enabled a direct assessment of the compositional advantage for each adjective type within the same participants.

The experiment included blocks of adjective-noun phrases and blocks of single-feature trials presenting only the adjective (one block for each adjective type). A 2000 ms blank screen separated the linguistic stimuli and the presentation of a coloured line drawing of the object. Participants judged whether the icon matched or mismatched the preceding linguistic stimuli by pressing ‘A’ for match or ‘L’ for mismatch. Two trial types, intersective and subsective, employed different visual icons. Intersective trials used green or orange cars and boats presented in a medium size (approximately 12 cm or 400 px wide on a 1920 px-wide screen). Subsective trials used purple versions of the same objects in two sizes: small (approximately 4 cm or 133 px wide) and large (36 cm or 1200 px wide). These icons consisted of simple line drawings of the respective object categories. Prior to the main experiment, participants completed four practice trials.

The experiment consisted of 128 trials (32 trials per block  $\times$  4 blocks), equally divided among the four trial types: intersective adjective-noun phrase, subsective adjective-noun phrase, intersective-only and subsective-only. The four blocks corresponded to these conditions and were presented once per participant in a randomized order. Within each block, trials were randomized. Stimuli were presented and responses recorded using PsychoPy (version 2023.2.3) and Pavlovla (platform version 2024.1.4). Each experimental session consisted of an equal number of match and mismatch trials (50% each). On match trials, the visual display corresponded to the linguistic cue (e.g., *big boat*  $\rightarrow$  big boat), whereas on mismatch trials, the display and cue did not correspond (e.g., *big boat*  $\rightarrow$  small car, see **Figure 5**).

In Experiment 3, single-feature trials were restricted to adjectives rather than including nouns. This choice was motivated by two considerations. First, Experiment 1 already established that object-noun cues elicit slower responses than adjectives, likely reflecting differences in conceptual complexity and perceptual diagnosticity. Including additional single-noun trials would therefore not have provided new theoretical leverage. Second, limiting single-feature cues to adjectives allowed us to maintain a balanced number of trials across conditions and to keep total session length within feasible limits, given that Experiment 3 introduced both intersective and subsective adjective types within participants. This ensured that the experiment targeted

the intended lexical–semantic contrast without imposing excessive task duration or fatigue. The preregistration is available at <https://doi.org/10.17605/OSF.IO/J8K92>.



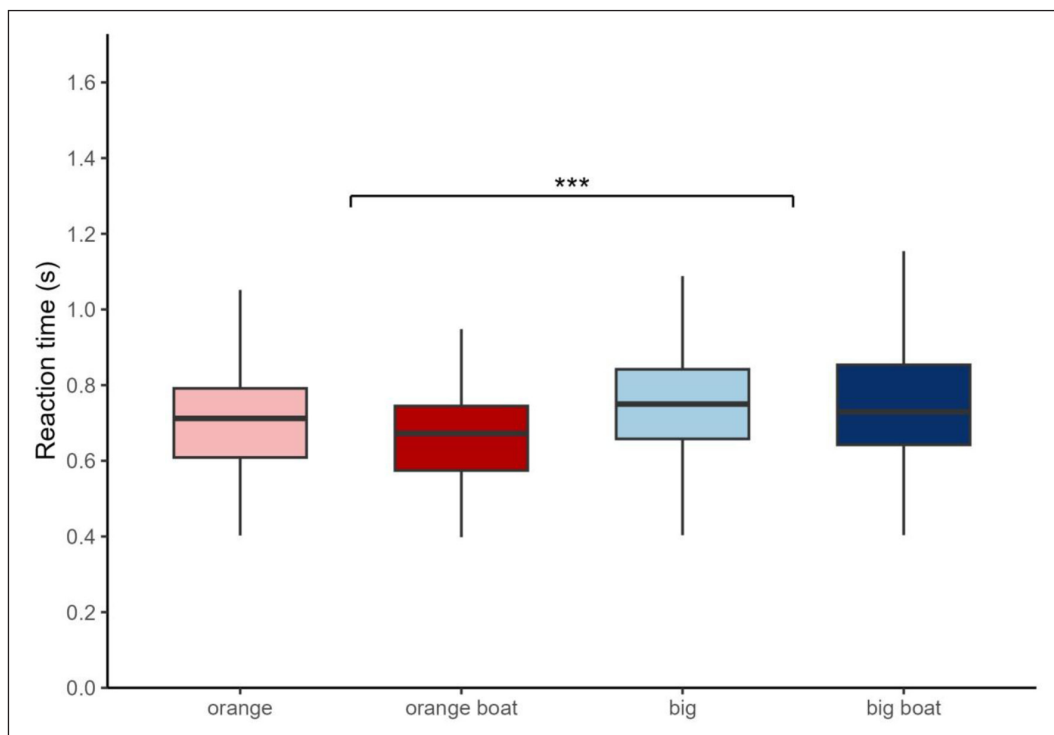
**Figure 5:** Illustration of the task design in Experiment 3. The top-left panel shows an example of a double-feature trial with a colour adjective (e.g., *green boat*), and the top-right panel shows the double-feature trial featuring a size adjective (e.g., *big boat*). The bottom-left panel depicts a single-feature trial with a colour adjective (e.g., *green*), and the bottom-right panel shows a single-feature trial with a size adjective (e.g., *big*).

### 7.3 Data analysis and results

Data were analysed using the same procedures and packages as in Experiments 1 and 2. Accuracy in match<sup>6</sup> trials was high across all conditions (91–94%) and did not differ systematically by feature condition or adjective type, indicating that response-time patterns were not confounded by speed–accuracy trade-offs (21.3% of correct match trials were excluded due to RT trimming <100 ms or >2000 ms). Analyses of correct match trials revealed a consistent main effect of

<sup>6</sup> A detailed analysis of the correct mismatch trials is provided in the *Supplementary Materials*; results closely paralleled those observed for the match trials.

adjective type. A repeated-measures ANOVA with factors feature condition (single vs. double) and adjective type (colour vs. size) showed significantly slower responses for size than for colour adjectives,  $F(1, 95) = 23.97, p < .001, \eta_p^2 = .024$ . The main effect of feature condition was not significant,  $F(1, 95) = 3.07, p = .083$ , and there was no interaction,  $F(1, 95) = 0.28, p = .60$ . Mean reaction times (based on participant averages) were 691 ms and 661 ms for single- and double-feature intersective trials, and 745 ms and 728 ms for single- and double-feature subjective trials, respectively (see **Figure 6**).



**Figure 6:** Mean reaction times in Experiment 3 for double-feature trials featuring size adjectives (e.g., *big boat*) colour adjectives (e.g., *orange boat*) and single-feature trials (size e.g., *big* vs. colour e.g., *orange*). Asterisks indicate statistically significant effects based on the overall ANOVA.

A linear mixed-effects model including random slopes for both within-subject factors replicated the ANOVA pattern. Responses were reliably slower for size than colour adjectives ( $\beta = -0.038 \pm 0.009, t(19.9) = -4.11, p = .001$ ), while neither the effect of feature condition ( $\beta = 0.013, p = .18$ ) nor the interaction ( $\beta = 0.006, p = .46$ ) reached significance. These converging results indicate that participants responded more slowly overall to size than to colour adjectives, but the advantage of double- over single-feature cues did not differ reliably between adjective types.

Given the absence of an interaction, exploratory follow-up models were fit within each adjective type to illustrate, rather than test, potential differences in magnitude. For colour adjectives, responses tended to be faster for double- than single-feature cues ( $\beta = 0.022 \pm 0.008$ ,  $t(78) = 2.67$ ,  $p = .009$ ), corresponding to an advantage of roughly 35 ms ( $\approx 4\text{--}5\%$ ). For size adjectives, no difference was observed ( $\beta = 0.005 \pm 0.016$ ,  $t(7.5) = 0.31$ ,  $p = .77$ ). This pattern should be interpreted cautiously, as the between-type contrast was not statistically reliable.

### 7.4 Discussion of experiment 3

Experiment 3 directly compared intersective (colour) and subsective (size) adjectives within the same participants, allowing a within-subject assessment of whether the redundancy-based facilitation observed in the previous experiments generalizes across adjective types. Overall, participants responded more slowly to subsective adjectives than to intersective ones, but there was no reliable interaction between adjective type and feature number. This indicates that both adjective types supported broadly similar patterns of visual–linguistic matching, with only modest numerical differences in facilitation that did not reach statistical significance. The consistent main effect of adjective type likely reflects differences in semantic specificity rather than differences in compositional processing. Intersective adjectives such as *green* denote stable, perceptually grounded properties that can be directly mapped onto visual features, facilitating rapid comparison between linguistic and visual representations. Subsective adjectives such as *big*, in contrast, describe relational or context-dependent properties whose interpretation depends on the object category (e.g., *big for a car* vs. *big for a boat*). Because these meanings are evaluated relative to an internal comparison standard, their mapping to visual input is less direct, resulting in slower and more variable responses overall.

The absence of a reliable interaction supports the interpretation that redundancy effects in this paradigm arise from the availability of directly alignable perceptual information rather than from syntactic or compositional integration. When both adjectives and nouns refer to perceptually separable dimensions, as in colour–shape combinations, redundant cues yield modest facilitation. When one of the cues is semantically context-dependent, as with size adjectives, this facilitation is attenuated, not because compositional mechanisms fail to apply, but because the relevant features are less perceptually determinate.

Together with Experiments 1 and 2, these findings indicate that apparent “combination” effects in this task are best understood as reflecting perceptual alignment and cue redundancy, rather than obligatory syntactic composition. The efficiency of linguistic–visual matching thus depends on the perceptual transparency and diagnosticity of the features being described, rather than on their grammatical configuration.

## 8 General discussion

This study investigated how different types of adjectives influence the processing of adjective–noun combinations in real time. Across three experiments, we examined whether the semantic properties of adjectives, specifically the distinction between intersective and subsective types, modulate the efficiency of integrating linguistic and visual information during comprehension. Across experiments, a consistent but graded pattern emerged.

In Experiment 1, which used intersective (colour) adjectives, participants were faster to verify visual targets following adjective–noun phrases than after single-word cues. This replicated the previously observed redundancy-based facilitation effect. In Experiment 2, which tested subsective (size) adjectives, a similar but weaker trend was observed, and the facilitation no longer reached significance once random variation was accounted for in a mixed-effects model. In Experiment 3, which directly contrasted both adjective types within the same participants, responses to subsective adjectives were generally slower than to intersective ones, but no reliable interaction between feature number and adjective type was found. Taken together, these results suggest that the efficiency of visual–linguistic matching varies with the semantic transparency of the features involved, rather than reflecting a categorical difference in linguistic compositional processes between adjective types.

Intersective adjectives, such as *green*, denote stable, perceptually grounded properties that can be mapped directly onto visual features. Subsective adjectives, such as *big*, describe relational or context-dependent properties that require a comparison class and therefore depend more heavily on contextual information. This difference likely underlies the overall slowdown observed for subsective adjectives, but it does not constitute evidence for qualitatively distinct compositional mechanisms.

The present findings therefore do not demonstrate a processing dissociation between intersective and subsective adjectives. Instead, they suggest that redundancy and perceptual alignment, rather than differences in linguistic compositional processes, determine the speed and efficiency of performance in this task. When cues provide multiple directly alignable features, as in colour–shape combinations, responses are faster. When one feature is semantically less determinate, as in size–object combinations, this facilitation is attenuated, not because compositional mechanisms fail to apply, but because the relevant features are less perceptually determinate. This interpretation is consistent with recent proposals that composition-like effects in such paradigms may reflect feature-based mappings between linguistic cues and visual dimensions, rather than syntactic assembly alone (e.g., Bemis & Pykkänen 2011; Flick et al. 2022). Participants appear to rely on feature-based mappings between words and visual dimensions, processing both components in parallel without necessarily computing a fully *linguistically* compositional meaning.

Accordingly, the absence of word order effects across all three experiments indicates that participants did not engage in syntactic composition under these task conditions, likely because both cues were sufficient to identify the visual referent.

Across all three experiments, exploratory mismatch trial analyses<sup>7</sup> showed uniformly high accuracy (typically above 90%) and closely mirrored the reaction-time patterns observed for match trials. Participants reliably rejected nonmatching displays, with minimal variability across conditions. In Experiment 1, mismatch responses were faster and more accurate for double-feature cues and for adjective-headed (colour-based) phrases, reflecting redundancy-driven facilitation similar to that observed in match trials. In Experiments 2 and 3, which included size adjectives, accuracy again remained near ceiling, and the same small advantage for double cues appeared only with intersective (colour) adjectives. Subjective (size) adjectives showed no such effect. Taken together, the mismatch data confirm that the facilitation associated with redundant or perceptually diagnostic features is not limited to affirmative (match) responses but extends to rejection trials as well. These converging patterns strengthen the interpretation that the observed effects reflect general efficiency in feature-based mapping rather than speed–accuracy trade-offs or response biases.

Although the current results align with formal semantic distinctions between adjective types, they do not imply that these distinctions have direct and categorical processing consequences. Instead, they point to a more graded relationship between semantic determinacy and processing efficiency: adjectives that denote concrete, perceptually grounded properties support faster visual matching than those that encode relational or context-sensitive ones. This graded pattern may reconcile prior theoretical and neurocognitive findings by suggesting that differences between adjective types emerge only under conditions requiring detailed compositional analysis.

It is worth noting that in the size-adjective conditions, the objects were depicted in a uniform purple hue that was not mentioned in the linguistic cue (for example, *big boat*). While this visual–linguistic asymmetry could have introduced a minor perceptual mismatch, all objects necessarily possessed some colour, and the hue was constant across conditions. Future work could directly test whether such incidental colour incongruities contribute to the slower responses observed for subjective adjectives. The consistent processing advantage for intersective adjectives likely arises from their direct perceptual grounding: their meaning can be mapped onto visual features without additional contextual computation. In contrast, subjective adjectives require evaluating the noun relative to a comparison class, introducing variability and slowing identification even when both words are available.

---

<sup>7</sup> See section *Supplementary Materials* for details.

## Supplementary Materials

### Experiment 1

#### Supplement A: Trial and Learning Effects

To evaluate potential practice or learning effects introduced by blocking, we re-analyzed the preregistered 48-participant dataset using three linear mixed-effects models (LMMs) on logtransformed reaction times (RTs). Each model included by-participant random intercepts and a random slope for the relevant trial index (*lme4*; Bates et al. 2015).

We defined three trial indices capturing learning at different levels of granularity (global, within-block, and within-condition), as described in Experiment 1.

### Results

Participants showed reliable speed-up over time:

- **Global model:**  $\beta = -0.050$  ( $SE = 0.0105$ ),  $t(\approx 35.1) = -4.79$ ,  $p < .001$ .  
Interactions with feature number and headedness were **not** significant (all  $p > .15$ ) **except** a small three-way interaction,  $\beta = 0.0167$  ( $SE = 0.0073$ ),  $t(\approx 1288) = 2.29$ ,  $p = .022$ .
- **Within-block model:**  $\beta = -0.0249$  ( $SE = 0.0061$ ),  $t(\approx 42.1) = -4.09$ ,  $p < .001$ . The interactions were **not** reliably different from zero ( $|t| \leq 1.69$ ,  $p \geq .09$ ).
- **Within-condition model:**  $\beta = -0.0399$  ( $SE = 0.0073$ ),  $t(\approx 42.4) = -5.45$ ,  $p < .001$ . Interactions with feature number, headedness, and their two-way were **not** significant (all  $p \geq .12$ ).

Taken together, these analyses show a general practice effect (faster responses later in the session) that is similar across blocks and conditions. Thus, the blocking scheme did not introduce condition-specific learning that could explain the main results.

#### Supplement B: Mismatch Data

Accuracy in mismatch trials was uniformly high across all conditions (95-98%; *single-noun*: 95.2%; *single-adjective*: 97.1%; *double adjective-first*: 97.7%; *double noun-first*: 98.3%). A generalized linear mixed-effects model with random intercepts for participants and items showed higher accuracy for double-feature cues than for single-feature cues ( $\beta = -0.34$ ,  $SE = 0.13$ ,  $z = -2.54$ ,  $p = .011$ ). The main effect of headedness did not reach significance ( $\beta = -0.23$ ,  $SE = 0.13$ ,  $z = -1.71$ ,  $p = .087$ ), although the numerical pattern suggested slightly higher accuracy for colour-headed than for shape-headed cues. There was no interaction ( $p = .69$ ). For correct rejections, reaction times (RTs) were analyzed with a linear mixed-effects model on log-transformed RTs including random intercepts for participants and items. Responses were faster for double- than

single-feature cues ( $\beta = 0.039$ ,  $SE = 0.006$ ,  $t(7.02) = 6.13$ ,  $p < .001$ ) and for adjective-headed (*single-colour* or *noun-first*) than for noun-headed (*single-noun* or *adjective-first*) cues ( $\beta = 0.021$ ,  $SE = 0.006$ ,  $t(7.02) = 3.31$ ,  $p = .013$ ), with no interaction ( $p = .60$ ). For correct rejections, reaction times (RTs) were analyzed with a linear mixed-effects model on log-transformed RTs including random intercepts for participants and items. Responses were faster for double- than single-feature cues ( $\beta = 0.039$ ,  $SE = 0.006$ ,  $t(7.02) = 6.13$ ,  $p < .001$ ) and for adjective-headed (*single-colour* or *noun-first*) than for noun-headed (*single-noun* or *adjective-first*) cues ( $\beta = 0.021$ ,  $SE = 0.006$ ,  $t(7.02) = 3.31$ ,  $p = .013$ ), with no interaction ( $p = .60$ ). Consistent with these effects, back-transformed estimated means showed faster responses for colour than for noun cues in single-feature trials (625 vs. 657 ms) and for adjective-first compared to noun-first sequences in double-feature trials (582 vs. 603 ms).

### Supplement C: Accuracy Analysis Match Data

To complement the RT analyses, we examined accuracy patterns on *match* trials in Experiment 1. Accuracy was very high overall. In the full sample, mean accuracy ranged from 95.0 % to 96.2 % across conditions. A binomial GLMM predicting trial-level accuracy from featurecondition (single vs. double) and headedness (noun vs. colour), with random intercepts for participant and item, revealed no significant effects or interactions (all  $|z| < 1$ ,  $p > .38$ ).

Estimated marginal probabilities on the response scale confirmed near-ceiling performance (*single-noun*: 0.959 ( $SE = 0.009$ ), *double adjective-first*: 0.964 ( $SE = 0.008$ ), *single-colour*: 0.963 ( $SE = 0.008$ ), *double adjective-first*: 0.970 ( $SE = 0.007$ ). The model showed a boundary (singular) fit due to negligible item variance, consistent with uniformly high accuracy.

## Experiment 2

### Supplement D: Learning Effects

To evaluate potential practice or learning effects introduced by blocking, we re-analysed the preregistered 48-participant dataset with three linear mixed-effects models (LMMs) on logtransformed RTs. Each model included by-participant random intercepts and a random slope for the corresponding trial index (lme4; Bates et al. 2015). Three trial indices were constructed to capture learning at different levels of granularity: a global index counting trials across the whole session, a within-block index (for the four observed blocks: *single-size*, *single-noun*, *double adjective-first*, *double noun-first*), and a within-condition index collapsing across single vs. double trials.

## Results

No reliable evidence of systematic practice or learning effects was found.

- **Global model:**  $\beta = -0.0029$  ( $SE = 0.0103$ ),  $t(\approx 41) = -0.29$ ,  $p = .78$ . A small interaction between trial and feature condition was significant ( $\beta = 0.0184$ ,  $SE = 0.0070$ ,  $t(\approx 1882) = 2.63$ ,  $p = .009$ ), indicating a slightly flatter slope for double- than single-feature trials. The main effects reflected the expected baseline RT differences between conditions, but there was no evidence that these patterns changed systematically over trials.
- **Within-block model:**  $\beta = 0.0079$  ( $SE = 0.0056$ ),  $t(\approx 526) = 1.42$ ,  $p = .16$ . Interactions with block label were not significant (all  $p \geq .28$ ). Mean RTs differed across blocks (single > double; noun > size), but these differences remained constant over trials.
- **Within-condition model:**  $\beta = -0.0024$  ( $SE = 0.0099$ ),  $t(\approx 33.1) = -0.24$ ,  $p = .81$ . A small trial  $\times$  feature-condition interaction ( $\beta = 0.0111$ ,  $SE = 0.0054$ ,  $t(\approx 2125) = 2.06$ ,  $p = .04$ ) and a marginal trial  $\times$  word-type interaction ( $p = .058$ ) suggested weak, inconsistent trends. The three-way interaction was not significant ( $p = .81$ ).

Taken together, these analyses show no systematic RT reductions over time and no evidence of condition-specific learning. Participants' response speeds remained stable across the session, indicating that the observed processing asymmetries reflect persistent representational differences rather than cumulative practice effects.

### Supplement E: Mismatch Data

Accuracy in mismatch trials was uniformly high across all conditions, indicating that participants reliably rejected nonmatching displays.

For correct rejections, reaction times (RTs) were analyzed using a linear mixed-effects model on log-transformed RTs with random intercepts for participants and items, and by-participant slopes for both within-subject factors.

There was no reliable main effect of feature condition ( $\beta = 0.012$ ,  $SE = 0.007$ ,  $t(\approx 13) = 1.67$ ,  $p = .12$ ). Responses were faster overall for colour than for size adjectives ( $\beta = -0.037$ ,  $SE = 0.008$ ,  $t(\approx 17) = -4.87$ ,  $p < .001$ ). The interaction between feature condition and adjective type was small and only marginally significant ( $\beta = 0.014$ ,  $SE = 0.006$ ,  $t(\approx 7) = 2.27$ ,  $p = .056$ ), suggesting that any double-feature-related facilitation may depend on adjective type.

Back-transformed estimated means (geometric means) mirrored this pattern: for colour adjectives, double cues were faster than single cues (663 vs. 697 ms), whereas size adjectives showed no difference (733 vs. 731 ms). In sum, mismatch performance remained near ceiling, and double features yielded only a modest, adjective-specific RT advantage

### Supplement F: Accuracy Analysis Match Data

Accuracy in match trials was high across all conditions (87–92%; single-size: 87.4%; singlenoun: 92.4%; double-size: 91.9%; double-noun: 92.1%). A generalized linear mixed-effects model with

random intercepts for participants and items showed no significant effects of feature condition, word type, or their interaction ( $\beta = -0.14$ ,  $SE = 0.14$ ,  $z = -1.00$ ,  $p = .32$ ;  $\beta = -0.17$ ,  $SE = 0.14$ ,  $z = -1.21$ ,  $p = .23$ ; interaction  $p = .26$ ).

Estimated accuracy probabilities confirmed that responses were comparably accurate across all cue types (*single-size*: 0.87; *single-noun*: 0.92; *double-size*: 0.92; *double-noun*: 0.92). Thus, while size cues tended to produce slightly lower accuracy numerically than noun cues, this difference did not reach significance, and accuracy was uniformly high overall.

## Experiment 3

### Supplement G: Learning Effects

To assess potential practice or learning effects in Experiment 3, we analyzed log-transformed reaction times (RTs) from correct match trials (0.1–2 s) using three linear mixed-effects models (LMMs) with by-participant random intercepts and random slopes for the relevant trial index. Each model examined RT change over the course of the experiment at a different level of aggregation. Specifically, the *global model* assessed overall change across the full session; the *within-block model* captured possible practice effects within each of the four block types (*singlecolour*, *single-size*, *double-colour*, *double-size*); and the *within-condition model* tested learning within feature conditions (single vs. double) while retaining adjective type (colour vs. size) as a second within-subject factor. In all models, coefficients associated with trial indices indicate learning slopes, whereas cross-sectional condition effects serve only as reference baselines.

## Results

No evidence for systematic practice or fatigue effects that could account for the main experimental outcomes

- **Global model:** Across the full session, there was no evidence of general practice-related speed-up ( $\beta = 0.0073$ ,  $SE = 0.0075$ ,  $t = 0.97$ ,  $p = .34$ ). Interactions of trial position with feature number or adjective type were not significant (all  $p \geq .12$ ), indicating that learning slopes did not differ across conditions. The only reliable effects were constant RT differences between conditions (responses were faster in double-feature than single-feature trials and slower for size than colour adjectives) reflecting stable baseline differences rather than learning.
- **Within-block model:** When examining trial order within each block, there was again no reliable indication of within-block practice ( $\beta = 0.0029$ ,  $SE = 0.0052$ ,  $t = 0.57$ ,  $p = .57$ ). Several main effects of block label mirrored the overall RT differences between conditions (e.g., faster responses for *double-colour* than *double-size* trials), but their interactions with

the trial index were small, inconsistent, and did not form a systematic pattern across blocks.

- **Within-condition model:** A similar picture emerged when learning was modeled within each feature condition. The overall trend toward faster RTs over trials was weak and non-significant ( $\beta = -0.0104$ ,  $SE = 0.0059$ ,  $t = -1.76$ ,  $p = .08$ ). However, a small but significant three-way interaction between trial order, feature condition, and adjective type ( $\beta = -0.0161$ ,  $SE = 0.0049$ ,  $t = -3.28$ ,  $p = .001$ ) indicated that the minimal learning observed was modulated by both factors. Inspection of fitted trends suggested that responses to colour adjectives showed a slightly greater improvement over time than those to size adjectives, particularly in the double-feature condition. No other trial-related effects reached significance.

Across all analyses, reaction times did not generally decrease over the session or within blocks, and only a minor interaction indicated slightly greater improvement for colour adjectives. Thus, the observed condition differences in Experiment 3 are stable and cannot be attributed to differential learning across time.

### Supplement H: Mismatch Data

Accuracy in mismatch trials was uniformly high across all conditions, indicating that participants reliably rejected nonmatching displays.

Reaction times (RTs) for correct rejections were analyzed using a linear mixed-effects model on log-transformed RTs, with random intercepts for participants and items and by-participant slopes for both within-subject factors.

There was no reliable main effect of feature condition ( $\beta = .012$ ,  $SE = .007$ ,  $t(\approx 13) = 1.67$ ,  $p = .12$ ). Responses were faster overall for colour than size adjectives ( $\beta = -.036$ ,  $SE = .008$ ,  $t(\approx 17) = -4.73$ ,  $p < .001$ ). The interaction between feature condition and adjective type was small and only marginally significant ( $\beta = .014$ ,  $SE = .006$ ,  $t(\approx 7) = 2.29$ ,  $p = .055$ ), suggesting that any redundancy-related facilitation depended on adjective type.

Back-transformed means showed that colour adjectives elicited faster responses for double- than single-feature cues ( $\approx 665$  ms vs. 690 ms), whereas size adjectives showed virtually identical latencies across cue types ( $\approx 730$  ms for both). In sum, mismatch performance remained near ceiling, and redundancy yielded only a modest, adjective-specific RT advantage.

### Supplement I: Accuracy Analysis Match Data

Accuracy in match trials was uniformly high across all conditions (92–96%), indicating that participants responded reliably when the cue and display matched.

A generalized linear mixed-effects model (GLMM) with random intercepts for participants and items revealed significantly higher accuracy for double- than single-feature cues ( $\beta = -0.12$ ,  $SE = 0.05$ ,  $z = -2.33$ ,  $p = .020$ ) and for size compared to colour adjectives ( $\beta = 0.25$ ,  $SE = 0.05$ ,  $z = 4.84$ ,  $p < .001$ ). The interaction between feature condition and adjective type was not significant ( $p = .21$ ).

### **Supplement J: Global Analysis of Exp. 1 and 2**

To assess the consistency of redundancy and adjective-type effects across experiments, reaction times (RTs) from Experiments 1 and 2 were combined and analyzed using a linear mixed-effects model on log-transformed RTs. The model included fixed effects of feature condition (single vs. double), adjective type (colour vs. size), and their interaction, with by-participant random slopes for feature condition and by-item random intercepts.

Responses were faster overall for double- than single-feature cues ( $\beta = 0.032$ ,  $SE = 0.011$ ,  $t(\approx 29) = 2.79$ ,  $p = .009$ ), and for colour compared to size adjectives ( $\beta = -0.068$ ,  $SE = 0.024$ ,  $t(\approx 115) = -2.79$ ,  $p = .006$ ). The interaction between feature condition and adjective type was not significant ( $p = .71$ ).

These results replicate the main double-feature and adjective-type effects observed in the individual experiments, indicating that both the facilitation from multiple cues and the speed advantage for colour adjectives generalize across experiments.

### **Data availability**

Data and analysis scripts are available in the Open Science Framework repository, which can be accessed using the following link: <https://osf.io/3nqa9>.

### **Ethics and consent**

This study was approved by, and carried out in accordance with the recommendations of, the Central European University Psychological Research Ethics Board. All participants gave written informed consent in accordance with the Declaration of Helsinki.

### **Acknowledgments**

We thank the editor and two anonymous reviewers for their thoughtful and constructive comments and Attila Balla for assistance with data collection.

### **Competing interests**

The authors have no competing interests to declare.

---

## References

- Allen, Richard & Alan, Baddeley & Hitch, Graham. 2006. Is the binding of visual features in working memory resource-demanding? *Journal of Experimental Psychology: General* 135. 298–313. <https://doi.org/10.1037/0096-3445.135.2.298>
- Aparicio, H. & Xiang, M., & Kennedy, C. (2015). Processing gradable adjectives in context: A visual world study. <https://doi.org/10.3765/salt.v25i0.3128>
- Bakeman, Roger. 2005. Recommended effect size statistics for repeated measures designs. *Behavior Research Methods* 37(3). 379–384. <https://doi.org/10.3758/BF03192707>
- Barker, Chris & Jacobson, Pauline. 2007. *Direct compositionality* (Oxford Linguistics 14). Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780199204373.001.0001>
- Barsalou, Lawrence W. 2017. Cognitively plausible theories of concept composition. In Hampton, James A. & Winter, Yoad (eds.), *Compositionality and Concepts in Linguistics and Psychology* (Language, Cognition, and Mind), vol. 3, 9–30. Cham: Springer International Publishing. [https://doi.org/10.1007/978-3-319-45977-6\\_2](https://doi.org/10.1007/978-3-319-45977-6_2)
- Bates, Douglas & Mächler, Martin & Bolker, Ben & Walker, Steve. 2015. Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software* 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Bemis, Douglas K. & Pylkkänen, Liina. 2011. Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience*. 31(8). 2801–2814 <https://doi.org/10.1523/JNEUROSCI.5003-10.2011>
- Bemis, Douglas K. & Pylkkänen, Liina. 2013a. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral Cortex* 23(8). 1859–1873. <https://doi.org/10.1093/cercor/bhs170>
- Bemis, Douglas K. & Pylkkänen, Liina. 2013b. Flexible composition: MEG evidence for the deployment of basic combinatorial linguistic mechanisms in response to task demands. *PLoS ONE* 8(9). e73949. <https://doi.org/10.1371/journal.pone.0073949>
- Bocanegra, Bruno R. & Poletiek, Fenna H. & Zwaan, Rolf A. 2022. Language concatenates perceptual features into representations during comprehension. *Journal of Memory and Language* 127. 104355. <https://doi.org/10.1016/j.jml.2022.104355>
- Brockmole, James & Parra, Mario & Sala, Sergio Della & Logie, Robert. 2008. Do binding deficits account for age-related decline in working memory? *Psychonomic Bulletin & Review* 15. 543–547. <https://doi.org/10.3758/PBR.15.3.543>
- Chang, Michael & Gupta, Abhishek & Levine, S. & Griffiths, T. 2018. Automatically Composing representation transformations as a means for generalization. *ArXiv*. <https://www.semanticscholar.org/paper/1766648967f6206a944a4bd18bbbd92a74c164bd> (13 May, 2025).
- Chomsky, N. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 2(3). 113–124. <https://doi.org/10.1109/TIT.1956.1056813>
- Chomsky, Noam. 2006. *Language and mind*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511791222>

- Frege, Gottlob. 1963. Compound thoughts. *Mind* 72(285). 1–17. <https://doi.org/10.1093/mind/LXXII.285.1>
- Heim, Irene & Kratzer, Angelika. 1998. *Semantics in generative grammar*. 1st edn. Blackwell
- Holyoak, Keith J. & Stamenković, Dušan. 2018. Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin* 144(6). 641–671. <https://doi.org/10.1037/bul0000145>
- Humboldt, Wilhelm Freiherr von. 1876. *Ueber die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwicklung des Menschengeschlechts*. S. Calvary.
- Jackendoff, Ray & Pinker, Steven. 2005. The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition* 97(2). 211–225. <https://doi.org/10.1016/j.cognition.2005.04.006>
- Kemp, Charles. 2012. Exploring the conceptual universe. *Psychological Review* 119(4). 685–722. <https://doi.org/10.1037/a0029347>
- Kamp, Hans. 2013. Two theories about adjectives. In von Heusinger, Klaus & ter Meulen, Alice G. B. (eds.), *Meaning and the dynamics of interpretation: Selected papers of Hans Kamp* (Current Research in the Semantics/Pragmatics Interface volume 29). Leiden ; Boston: Brill. <https://doi.org/10.1163/9789004252882>
- Kennedy, Christopher. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30(1). 1–45. <https://doi.org/10.1007/s10988-006-9008-0>
- Lenth, Russell V. 2023. emmeans: Estimated marginal means, aka least-squares means (R package version 1.8.5). *Vienna: R Foundation for Statistical Computing*. <https://CRAN.Rproject.org/package=emmeans>
- Logie, Robert H. & Brockmole, James R. & Vandenbroucke, Annelinde R. E. 2009. Bound feature combinations in visual short-term memory are fragile but influence long-term learning. *Visual Cognition* 17(1–2). 160–179. <https://doi.org/10.1080/13506280802228411>
- Luck, Steven J. & Vogel, Edward K. 1997. The capacity of visual working memory for features and conjunctions. *Nature* 390(6657). 279–281. <https://doi.org/10.1038/36846>
- Nicholson, Karen G. & Humphrey, G. Keith. 2004. The effect of colour congruency on shape discriminations of novel objects. *Perception* 33(3). 339–353. <https://doi.org/10.1068/p5136>
- Olejnik, Stephen & Algina, James. 2003. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods* 8(4). 434–447. <https://doi.org/10.1037/1082-989X.8.4.434>
- Partee, Barbara. 1995. Lexical semantics and compositionality. In *An invitation to cognitive science: Language* 1. 311–360. 2nd edn. Cambridge, MA: MIT Press. <https://www.cs.brandeis.edu/~jamesp/classes/cs216-2009/readings2009/ParteeSemanticsAndCompositionality.pdf> (13 May, 2025).

- Phillips, Steven & Wilson, William H. 2011. Categorical compositionality II: Universal constructions and a general theory of (quasi-)systematicity in human cognition. *PLoS Computational Biology* 7(8). e1002102. <https://doi.org/10.1371/journal.pcbi.1002102>
- Potter, Mary C. & Faulconer, Barbara A. 1979. Understanding noun phrases. *Journal of Verbal Learning and Verbal Behavior* 18(5). 509–521. [https://doi.org/10.1016/S0022-5371\(79\)90274-3](https://doi.org/10.1016/S0022-5371(79)90274-3)
- Pylkkänen, Liina & Bemis, Douglas K. & Elorrieta, Estibaliz Blanco. 2014. Building phrases in language production: An MEG study of simple composition. *Cognition* 133(2). 371–384. <https://doi.org/10.1016/j.cognition.2014.07.001>
- R Core Team. 2023. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org>
- Rabagliati, Hugh & Dumas, Leonidas A. A. & Bemis, Douglas K. 2017. Representing composed meanings through temporal binding. *Cognition* 162. 61–72. <https://doi.org/10.1016/j.cognition.2017.01.013>
- Redolfi, M. & Melloni, C. 2025. Processing adjectives in development: Evidence from eyetracking. *Journal of Child Language* 52(2). 270–293. <https://doi.org/10.1017/S0305000923000703>
- Rubio-Fernández, Paula. 2016. How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology* 7. <https://doi.org/10.3389/fpsyg.2016.00153>
- Singmann, Henrik & Bolker, Ben & Westfall, Jacob & Aust, Frederik & Ben-Shachar, Mattan S. 2021. *afex: Analysis of factorial experiments (R package version 1.0-1)*. CRAN. <https://CRAN.Rproject.org/package=afex>
- Solt, S. 2009. *The semantics of adjectives of quantity* (Order No. 3349494). Available from ProQuest Dissertations & Theses Global: The Humanities and Social Sciences Collection. (304859514). Retrieved from <https://www.proquest.com/dissertations-theses/semantics-adjectives-quantity/docview/304859514/se-2>
- Solt, Stephanie. 2015. Vagueness and Imprecision: Empirical Foundations. *Annual Review Linguistics* 1. 107–127. <https://doi.org/10.1146/annurev-linguist-030514-125150>
- Werning, Markus. 2010. Complex first? On the evolutionary and developmental priority of semantically thick words. *Philosophy of Science* 77(5). 1096–1108. <https://doi.org/10.1086/656826>
- Wittenberg, Eva. 2016. *With light verb constructions from syntax to concepts*. Universitätsverlag Potsdam.
- Ziegler, Jayden & Pylkkänen, Liina. 2016. Scalar adjectives and the temporal unfolding of semantic composition: An MEG investigation. *Neuropsychologia* 89. 161–171. <https://doi.org/10.1016/j.neuropsychologia.2016.06.010>
- Ziegler, Jayden & Snedeker, Jesse & Wittenberg, Eva. 2018. Event structures drive semantic structural priming, Not thematic roles: Evidence from idioms and light verbs. *Cognitive Science* 42(8). 2918–2949. <https://doi.org/10.1111/cogs.12687>

