

## Neural and behavioural effects of typicality, denotation and composition in an adjective–noun combination task

Isabella Fritz & Giosuè Baggio

To cite this article: Isabella Fritz & Giosuè Baggio (2021): Neural and behavioural effects of typicality, denotation and composition in an adjective–noun combination task, Language, Cognition and Neuroscience, DOI: [10.1080/23273798.2021.2004176](https://doi.org/10.1080/23273798.2021.2004176)

To link to this article: <https://doi.org/10.1080/23273798.2021.2004176>



© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 22 Nov 2021.



[Submit your article to this journal](#)



Article views: 663



[View related articles](#)



[View Crossmark data](#)

# Neural and behavioural effects of typicality, denotation and composition in an adjective–noun combination task

Isabella Fritz<sup>a</sup> and Giosuè Baggio<sup>b</sup> 

<sup>a</sup>Language and Brain Lab, Faculty of Linguistics, Philology and Phonetics, University of Oxford, Oxford, UK; <sup>b</sup>Language Acquisition and Language Processing Lab, Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway

## ABSTRACT

Formal semantics states that the meanings of phrases are composed from the meanings of constituent parts and syntax. Little is known about how composition is neurally implemented. We studied ERP and behavioural responses to determiner-adjective-noun phrases. We assessed the effects of typicality and denotation, using intersective (typical: “A green turtle”, atypical: “An orange turtle”) or subsective adjectives (typical: “A slow turtle”, atypical: “A fast turtle”). After each phrase, participants responded to two questions (e.g., for “A fast turtle”: “Is it a common turtle?”; “Is it a fast animal?”). We contrasted these 4 semantic conditions, requiring composition, to 2 nonsemantic conditions, where the adjective was replaced with a pseudoword or a nonword. This contrast revealed a larger P600, if participants performed the task without instructions and feedback (experiment 1), or a larger sustained negativity, if they were nudged to pay attention to meaning by instructions and feedback (experiment 2). Typicality or denotation had an impact only on behavioural responses. We discuss implications for theories of language processing and compositional semantics.

## ARTICLE HISTORY

Received 19 January 2021  
Accepted 29 October 2021

## KEYWORDS



Semantics; composition; adjectives; ERP; P600


## 1. Introduction

Composing simple expressions, such as words or gestures, into more complex structures is a cornerstone of the human language capacity. Despite some progress in understanding the cortical mechanisms of speech and language processing, relatively little attention has been paid to linguistic composition at the phrasal level. Most psycho- and neurolinguistic studies have focused on sentence-level processing, where at least two different mechanisms are at play: bottom-up composition and top-down prediction. These two mechanisms can interact continuously in sentence or discourse processing and are therefore difficult to disentangle. Some studies have tried to isolate compositional operations by looking at minimal phrases (e.g. adjective–noun phrases), where the effects of predictive or anticipatory processing are minimised. MEG results suggest that processing phrases drives activity in the left anterior temporal lobe (L ATL) at ~200–250 msec after word onset (Bemis & Pykkänen, 2011, 2013; Blanco-Elorrieta & Pykkänen, 2016; Del Prato & Pykkänen, 2014; Zhang & Pykkänen, 2015; Ziegler & Pykkänen, 2016). However, it is

unclear whether this early LATL response is specifically a signature of syntax-driven semantic composition, as opposed to conceptual combination (Pykkänen, 2016). A recent ERP study (Fritz & Baggio, 2020) points to a later stage at which compositional operations might (also) occur, in the P600 frame (for a theory predicting a P600 “composition effect”, see Baggio, 2018, 2021), while other studies have implied that composition may also be reflected by N400 effects (Neufeld et al., 2016). Given these different experimental methods, dependent measures and results, we currently lack a firm empirical basis for models of syntax-driven meaning composition in the brain.

In addition, little is known about the brain correlates of different lexico-semantic variables within phrasal structures (e.g. [determiner [adjective noun]]) in which the predictability of the noun is minimised. Two well-investigated variables that affect language processing are *context-sensitivity* and *typicality*. Typicality and related semantic manipulations have been researched extensively in the field of sentence comprehension (Federmeier & Kutas, 1999; Molinaro et al., 2012).

**CONTACT** Giosuè Baggio  giosue.baggio@ntnu.no  Department of Language and Literature, Norwegian University of Science and Technology, NO-7491, Trondheim, Norway

 Supplemental data for this article can be accessed <https://doi.org/10.1080/23273798.2021.2004176>.

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Context-sensitivity has been studied in lexical and formal semantics, mainly in the adjectival modification literature (Solt, 2018) and in psycholinguistics (Lyu et al., 2019). One important distinction in the linguistic literature is between context-sensitive adjectives (e.g. *subsective* adjectives: “fast”, “slow” etc., whose contribution to the meaning of a phrase or sentence depends also on the modified noun) and relatively context-insensitive ones (e.g. *intersective* adjectives of colour, shape etc.). The interpretation of many *subsective* adjectives depends on the lexical semantics of the head noun and requires taking a relevant comparison class into consideration (details below). E.g. a fast turtle may be fast relative to other turtles, but it is still slow in comparison to other animals; it might therefore not count as a fast animal. In contrast, a green turtle is a green animal, irrespective of the comparison class one chooses to interpret the adjective. In two EEG experiments, we studied the neural bases of adjective–noun phrase composition using *intersective* (e.g. “green”) and *subsective* (e.g. “fast”) adjectives. Further, we examined the ERP effects of typicality in adjective–noun phrases, using typical and atypical *intersective* adjectives (e.g. “A green turtle” vs “An orange turtle”) and *subsective* adjectives (e.g. “A slow turtle” vs “A fast turtle”). We aimed to assess whether typicality and the *subsective* vs *intersective* distinction would affect the amplitude of language-related ERP components, such as the N400 (Federmeier & Kutas, 1999; Molinaro et al., 2012; Urbach & Kutas, 2010). An additional aim of the study was to replicate and to probe further the P600 “composition effect” we recently reported using the same paradigm, but a different set of stimuli and different adjective manipulations (Fritz & Baggio, 2020).

### 1.1. Neural correlates of phrasal composition

In a series of MEG studies, Pykkänen and colleagues investigated phrasal composition by comparing adjective–noun phrases (e.g. “red boat”) with nonword–noun combinations (“xkq boat”) and word lists (“cup, boat”) (for representative studies, see Bemis & Pykkänen, 2011, 2013; Blanco-Elorrieta & Pykkänen, 2016; Del Prato & Pykkänen, 2014; Westerlund et al., 2015; Zhang & Pykkänen, 2015; Ziegler & Pykkänen, 2016). Across multiple experiments, the left anterior temporal lobe (LATL) was found to be engaged in an early time frame, i.e. ~200–250 msec after the noun’s onset in phrases, but not in the other conditions.

Initially, this early effect was interpreted as a *syntactic effect*, in agreement with syntax-first models of language processing (Friederici, 2002, 2017), in which local syntactic operations precede semantic interpretation (Bemis &

Pykkänen, 2011, 2013). However, recent results suggest that the early LATL effect rather reflects combinatorial *conceptual* processing (for a review, see Westerlund & Pykkänen, 2017). Research using the same or similar paradigms and tasks has found activity for the composition conditions also in ventromedial prefrontal cortex (vmPFC), around 400 msec after the noun’s onset (Bemis & Pykkänen, 2011, 2013). Although this effect is not observed consistently across experiments, Pykkänen (2016) has proposed that, following early conceptual combination in the LATL, the vmPFC carries out compositional operations, together pointing to a multi-step model of semantic processing.

Seemingly in accord with the early MEG data are the results of an ERP study using a similar paradigm and stimuli as in Bemis and Pykkänen (2011). Neufeld et al. (2016) report an early ERP effect, at ~180–250 msec, when contrasting adjective–noun phrases with letter strings. This earlier negative-going response was followed by an ERP effect in a later time window, at ~300–400 msec, which the authors suggest could be interpreted as an N400 effect. This conclusion is compatible with the proposal that the N400 reflects the integration of lexical meaning in context (Cosentino et al., 2017; Hagoort et al., 2009).

However, although it is generally accepted that the N400 reflects lexico-semantic processes (Kutas & Federmeier, 2011), its functional interpretation remains to this day controversial. The controversy revolves around two functional models of the N400: semantic *retrieval* of a content word from memory (Brouwer et al., 2017; Delogu et al., 2019; Federmeier, 2007) and semantic *integration* of a word into the context (Calloway & Perfetti, 2017; Hagoort et al., 2004; Lau et al., 2016). Hybrid accounts, proposing that the N400 indexes both semantic access or retrieval and integration processes, have been discussed in the literature (Baggio, 2012; Baggio & Hagoort, 2011; Nieuwland et al., 2020 for ERP evidence). Importantly, in all these proposals, consistent with the experimental evidence, *the N400 reflects top-down processes, such as context-driven activation and integration of lexical meanings*, rather than bottom-up operations (syntax-driven composition). As for Neufeld et al.’s (2016) findings, the observed ERP effects are not fully compatible with the classic N400 onset and duration.

Another ERP component that has been linked to language processing is the P600. Although initially associated with grammatical processing (Hagoort, 2003), more recent studies have interpreted the P600 as an index of phrasal or sentential “integration” (Brouwer et al., 2017; Delogu et al., 2019), or of combining syntactic and semantic information (Bornkessel-

Schlesewsky & Schlesewsky, 2008; Kuperberg, 2007). Semantic composition indeed requires that lexical meanings and syntactic information are combined. The P600 is a possible candidate signature of semantic composition (Baggio, 2018, 2021; Fritz & Baggio, 2020).

## 1.2. Intersective and subsective adjectives

A prominent distinction, discussed in the linguistic literature about adjectival modification, is that between intersective and subsective adjectives. The relation between an intersective adjective and the modified noun may be seen as a *symmetric* one. For example, the sentence “Floyd is a Canadian surgeon” entails both that Floyd is Canadian and that he is a surgeon: Floyd is a member of the *intersection* of the two sets ( $[\text{Adj N}] = [\text{Adj}] \cap [\text{N}]$ ) (Morzycki, 2016). Noun phrases containing a subsective adjective do not give rise to the same entailments. In particular, the relationship here is *asymmetric*, in that the set denoted by the noun phrase is a subset of the denotation of the noun only and bears no fixed relation with the denotation of the adjective:  $[\text{Adj N}] \subseteq [\text{N}]$ . A specific type of subsectivity may be found in the semantics of *gradable adjectives* (e.g. “fast”, “slow”, “cold”, “warm”). Gradable modifiers are characterised by their vagueness and context-sensitivity. Kennedy (2012) has even argued that gradable adjectives, as such, do not directly denote properties. Rather, their denotational effects only come into play through composition. On this account, adjectives such as “slow” only denote a property once a threshold or standard of speed for a given object (e.g. a turtle) has been established. This requires reference to a relevant comparison class (Kamp, 1975; Kennedy, 2007): e.g. a “fast turtle” is fast relative to standards for turtles, not for animals in general.

Do the properties of gradable subsective adjectives result in on-line processing costs? In an MEG study, Ziegler and Pykkänen (2016) tested whether the context-sensitivity of subsective scalar adjectives (e.g. “fast”, “large”) influences the time-course of compositional operations in a phrasal context. Their study showed that the early effect of conceptual combination in the LATL at ~200 msec post-noun onset found in previous studies (e.g. Bemis & Pykkänen, 2011, 2013) was only present for intersective adjectives, but not for scalar ones. This result suggests that the early LATL effect only occurs for combinatorial operations on the relevant words in specific conditions. Due to the context-sensitive nature of the scalar adjectives, the combinatorial process may only be triggered when the noun has been processed. However, at later processing stages, following the noun, the intersective

adjectives did not trigger any additional effects, while processing of the scalar adjectives did elicit an effect in the LATL at around 400 msec. The authors tentatively suggest that combinatorial operations for Adj-N phrases containing a scalar adjective may be delayed until the meaning of the noun is “fully determined”. Interpreting functionally the patterns of later activation in the LATL is further complicated by the fact that, in addition to the adjective manipulation, Ziegler and Pykkänen (2016) also manipulated the noun’s specificity, as in “dog” vs “animal”. In a recent replication experiment in Dutch, Kochari et al. (2021) failed to observe the early combinatorial effect for intersective adjectives and the later effect for scalar adjectives. To our knowledge, there are no EEG studies on composition of intersective vs subsective adjectives in NPs. Research is needed to try to tease apart the effects of context sensitivity and composition in the time frames associated with the N400 and P600 in ERPs.

## 1.3. Typicality and adjectival modification

Studies investigating the neural correlates of semantic processing often employ paradigms using different linguistic manipulations. In many of these studies, the sentence or discourse context is kept constant, while the semantic fit of the target word is manipulated (see Kutas & Hillyard, 1980 for a seminal study). The N400 amplitude is proportional to the semantic fit of the eliciting word in the given context. For example, incongruous words elicit a larger N400 component relative to congruous sentence continuations (for reviews, see Lau et al., 2008; Kutas & Federmeier, 2011). Importantly, not only semantic violations, but also world and background-knowledge violations, are associated with increased processing costs and with increased amplitudes of the N400 component (Urbach & Kutas, 2010). Yet, from these studies one cannot conclude whether N400 effects stem from difficulties in composition or from (non-confirmed) predictions of the target-word. Studies investigating typicality more directly did not target composition effects. For example, Federmeier et al. (2010) presented category cues to participants followed by typical, atypical or incongruous exemplars of that category (e.g. “A kind of tree”; typical: “oak”, atypical: “ash”, incongruent: “tin”) and found that typicality modulated the N400 response, with low typicality items eliciting N400 responses whose amplitude falls between high typicality and incongruous target words. Although this study did produce an N400 effect of typicality, the setup and stimuli presentation are more consistent with an effect of prediction error, rather than composition. Using ERPs, Molinaro et al. (2012) aimed

to isolate composition and processes related to semantic pre-activation. They embedded within sentences adjective–noun combinations that could be semantically neutral (e.g. “Lonely monster”), low-typicality (“Lovely monster”), high-typicality (“Horrible monster”), or anomalous (“Geographic monster”). These adjective–noun combinations were equally (un)predictable across conditions. As in previous work, they found an N400 effect when comparing the neutral and anomalous conditions. However, there was no N400 when comparing the neutral condition to the high-typicality or the low-typicality conditions. This contrasts with earlier behavioural works that found off-line differences of typicality (Lucas, 2001; Smith et al., 1988). Rather, the two different typicality conditions both elicited a late frontal positivity (after 500 msec) compared to the neutral cases, which the authors related to increased processing demands when embedding an adjective–noun combination into its sentence context. Thus, from this study, it is unclear whether those ERP effects were driven by adjective–noun composition as such or instead by integration of the target noun/NP into the sentence context. Also, the authors only compared the typical and atypical conditions to the baseline (neutral) condition and not to each other.

Lau et al. (2016) contrasted a “multiple generators” theory of the N400 (Baggio & Hagoort, 2011) with models that account for the N400 in terms of either semantic access (Lau et al., 2008) or semantic composition. They manipulated the predictability and the congruency of the noun in adjective–noun phrases, e.g. in “runny nose” vs “dainty nose” (predictability) and “yellow bag” vs “innocent bag” (congruency). The phrases “runny nose” and “dainty nose” can result in congruent meanings, but “nose” is more predictable in the former than in the latter. Instead, “bag” is unpredictable in both “yellow bag” and “innocent bag”, but only in the former case does it result in a congruent meaning. Access-based accounts predict a larger N400 in “dainty nose” than in “runny nose”, and composition-based accounts predict a larger N400 in “innocent bag” than in “yellow bag”. Neither of those models predicts an N400 effect for both “dainty nose” and “innocent bag”. Hybrid accounts of the N400 (Baggio & Hagoort, 2011) can accommodate N400 effects of *both* predictability (top-down) and congruency (bottom-up; Baggio, 2018). Indeed, Lau et al. (2016) found a larger N400 predictability effect for “dainty nose” vs “runny nose” and a smaller N400 for “innocent bag” vs “yellow bag”. The N400 effect of congruency showed a different topographical distribution compared to the predictability effect. That is suggestive of contributions from multiple cortical sources (for additional EEG evidence along the same lines, see

Nieuwland et al., 2020). These studies do not show that the N400 is a neural correlate of preactivation or composition, but they do point to multiple functional factors or concurrent processes potentially affecting the N400’s amplitude. The reason why these studies cannot directly bear on composition and its ERP correlates is that all of them compare conditions that require or involve composition, and none of them uses a baseline or control condition in which composition is not engaged. This is not the case for Pykkänen’s lab studies, which did include a low-level no-composition condition. As noted above, however, EEG/ERP data are needed to integrate LATL and vmPFC results from MEG studies with the electrophysiology of language at large, including research on the N400 and P600. This also applies to interactions between composition and other linguistic variables.

#### 1.4. The present study

Initially, the present study had two aims. The first was to further assess the hypothesis that the P600 is one candidate ERP signature of composition, by comparing semantic conditions to nonsemantic conditions. The former were noun phrases of the form [Det [Adj N]], where all three elements are actual words in Norwegian (Bokmål). In the nonsemantic conditions, the adjectives were replaced by either nonwords (phonotactically illegal consonant strings) or pseudowords (phonotactically legal strings; see Methods; Table 1). We assume that the semantic conditions involve semantic composition and the nonsemantic conditions do not: because the nonword and pseudoword carry no meaning, there is nothing to be composed with the meaning of the noun, i.e. no applicable modifier. This follows closely the design of previous MEG experiments by Pykkänen’s group. Additionally, the distinction between the nonword and pseudoword conditions has been used previously in an ERP study (Neufeld et al., 2016) and in an MEG study (Kochari et al., 2021). We slightly depart from those studies, however, in that our stimuli (semantic and nonsemantic) also include a determiner and are therefore complete, syntactically licensed noun phrases (NPs). The reason is theoretical. In linguistics (formal semantics), composition is often assumed to be a syntax-driven process: it only applies given some syntactic representation of the phrase or sentence. As a result, it is possible that composition *qua* cognitive/neural process is only triggered for syntactically licensed structures. By contrasting the semantic conditions, which require composition, to the nonsemantic conditions, we tried to isolate a candidate ERP correlate of composition in complete [Det [Adj N]] phrases, in which the



**Table 1.** Experimental design and examples of stimulus phrases.

| Det | Adj          | N ↓ERP           | Composition           | Denotation   | Typicality | Label |
|-----|--------------|------------------|-----------------------|--------------|------------|-------|
| en  | xkqh         | skilpadde turtle | Syn– Sem–             | ∅            | ∅          | NW    |
| a   | [nonword]    |                  |                       |              |            |       |
| en  | tæff         | skilpadde turtle | Syn+ Sem–             | ∅            | ∅          | PW    |
| a   | [pseudoword] |                  |                       |              |            |       |
| en  | grønn        | skilpadde turtle | Syn+ Sem+             | Intersective | Typical    | IST   |
| a   | green        |                  | [Adj N] = [Adj] ∩ [N] |              |            |       |
| en  | oransje      | skilpadde turtle | Syn+ Sem+             | Intersective | Atypical   | ISA   |
| a   | orange       |                  | [Adj N] = [Adj] ∩ [N] |              |            |       |
| en  | langsom      | skilpadde turtle | Syn+ Sem+             | Subsective   | Typical    | SST   |
| a   | slow         |                  | [Adj N] ⊆ [N]         |              |            |       |
| en  | rask         | skilpadde turtle | Syn+ Sem+             | Subsective   | Atypical   | SSA   |
| a   | fast         |                  | [Adj N] ⊆ [N]         |              |            |       |

effects of contextual prediction are minimised. In our study, pseudowords were effectively pseudo-adjectives and expressed morphological features (a grammatical gender suffix) that always agreed with both the Det and the N. The [Det [pseudo-Adj N]] condition may then be expected to engage syntactic composition, but not semantic composition (see Kochari et al., 2021). This implies that any candidate neural signature of semantic composition, revealed in ERPs in the semantic/nonsemantic contrast at the N, must also hold up in the semantic/pseudoword contrast. Based on theory (Baggio, 2018, 2021), on earlier studies on syntax-semantic integration (see above), and on our own findings (Fritz & Baggio, 2020), we expect that signature to be a modulation of the P600 component in the present paradigm (see Olstad et al., 2020 for a different paradigm).

The paradigm adopted here is exactly as described by Fritz and Baggio (2020), but the stimuli differ. The second aim of this study was to disentangle brain responses underlying different types of nominal modification, by manipulating lexical-semantic features of the adjectives. We are interested in *how* and *when* lexical-semantic properties of adjectives are taken into account during the online processing of a phrase's head noun. In Fritz and Baggio (2020), we studied the real-time effects of *intensionality* (modal vs temporal adjectives) and *denotation* (privative vs nonprivative adjectives). In the current study, we investigated the ERP effects of *denotation* (adjective type: intersective vs subsective) and *typicality* (typical vs atypical combinations) in the semantic conditions. A few studies reported that typicality affects the N400 amplitude (Federmeier et al., 2010; Urbach & Kutas, 2010), but they could not decide whether this typicality effect stems from prediction errors or composition. Moreover, in the Molinaro et al. (2012) study, where prediction was kept constant across typical or atypical adjective–noun combinations, no N400 effect was observed, but a P600 was obtained when comparing a neutral baseline condition to the atypical condition. Based on earlier research, it is not clear which ERP components, if any,

are modulated by (a) typicality, (b) denotation, and (c) composition *in phrasal contexts*, when prediction is minimised or absent. This is the problem we set out to address in this study. Predictions for typicality or denotation effects are difficult to formulate on theoretical grounds. The distinction between intersective and subsective adjectives concerns denotational aspects of meaning—formal properties of the real-world structures that can verify (or satisfy) an NP. These properties are often modelled in terms of sets and relations between them (see above). In the ERP literature, referential processing has been associated with modulations of post-N400 components, e.g. sustained negativities (SAN, Baggio et al., 2008; Nref, Van Berkum et al., 1999; Van Berkum et al., 2003). This is the most likely candidate ERP response, also in the present study (Fritz & Baggio, 2020).

We conducted two subsequent experiments, collecting EEG and behavioural data: accuracies and response times to questions targeting the typicality and denotation manipulations (see Table 2). Task-free, naturalistic reading or listening paradigms are increasingly used in the neuroscience of language. But given our aims, a task was necessary and had to be designed so as to (a) ensure that participants are indeed composing the meanings of the NPs and (b) assess how they interpret Adj-N phrases, e.g. whether they are sensitive to the distinctions between intersective vs subsective adjectives and between typical vs atypical combinations. The task should then include *two questions*, to check that participants would judge “a green turtle” (intersective typical) as a common turtle (typicality question) and as a green animal (denotation question). We used a *superordinate category* in denotation questions (Table 2). “An orange turtle” (intersective atypical) should be judged as a non-common turtle and an orange animal. “A slow turtle” (subsective typical) should be judged as a common turtle and a slow animal. “A fast turtle” (subsective atypical) should be judged as a non-common turtle, but not as a fast animal: a turtle that is fast relative to other turtles may not be fast relative to other animals.

**Table 2.** Experimental task and examples of questions and expected answers.

| Label | Det | Adj          | N ↓ERP    | Typicality question<br>[Expected answer] | Denotation question<br>[Expected answer] |
|-------|-----|--------------|-----------|--|--|
| NW    | en  | xkqh         | skilpadde |  | Is it an animal?                         |
|       | a   | [nonword]    | turtle    |  | [Yes]                                    |
| PW    | en  | tæff         | skilpadde |  | Is it an animal?                         |
|       | a   | [pseudoword] | turtle    |  | [Yes]                                    |
| IST   | en  | grønn        | skilpadde | Is it a common turtle?                   | Is it a green animal?                    |
|       | a   | green        | turtle    | [Yes]                                    | [Yes]                                    |
| ISA   | en  | oransje      | skilpadde | Is it a common turtle?                   | Is it an orange animal?                  |
|       | an  | orange       | turtle    | [No]                                     | [Yes]                                    |
| SST   | en  | langsom      | skilpadde | Is it a common turtle?                   | Is it a slow animal?                     |
|       | a   | slow         | turtle    | [Yes]                                    | [Yes]                                    |
| SSA   | en  | rask         | skilpadde | Is it a common turtle?                   | Is it a fast animal?                     |
|       | a   | fast         | turtle    | [No]                                     | [No]                                     |

In the first experiment, we did not give feedback and explicit instructions to participants on how they should respond. However, our results showed that participants tend to interpret intersective and subsective adjectives similarly, for example judging “A fast turtle” to be a fast animal. We therefore decided to conduct a second experiment, in which participants were provided with instructions and trial-to-trial correct/incorrect feedback to highlight the distinction between intersective and subsective adjectives. The difference in instructions and in-task feedback between the two experiments was intended to achieve a third emerging aim of the study, namely to encourage participants to pay attention to the meaning of phrases, which could also increase our chances to detect meaningful behavioural and ERP effects of typicality, denotation, and composition.

## 2. Methods

In what follows, we describe jointly the methods used in experiments 1 and 2 (abbreviated as E1 and E2). Differences in experimental design between E1 and E2 will be noted. Unless stated otherwise, the information provided below applies to both experiments. For both E1 and E2, we describe (in Methods) and report (in Results) all measures, conditions, and data and participant exclusions. Determination of sample sizes is addressed in section 2.2. Data and scripts are available at DataverseNO (<https://doi.org/10.18710/K849XH>).

### 2.1. Stimuli

We constructed 176 Norwegian Bokmål phrases with the syntactic form [Det [Adj N]]. The stimuli formed 44 quadruplets with the same Det and N, but varying the Adj preceding the N. The length of the Ns ranged between 3 and 10 letters (E1 & E2:  $M=6.2$ ,  $SD=2.1$ ). Ns were drawn from different categories, e.g. animals, transportation, food, or drinks. All Adjs were either intersective

or subsective. We manipulated the Adj-N combinations in the typicality dimension: adjectives were either *intersective* ( $[Adj\ N] = [Adj] \cap [N]$ , either typical or atypical Adj-N phrases) or *subsective* ( $[Adj\ N] \subseteq [N]$ , either typical or atypical Adj-N phrases). Apart from these 4 semantic conditions, we included a nonword condition (an unpronounceable consonant string) and a pseudoword condition (a phonotactically licensed, pronounceable string, but not a real Adj in Norwegian). Each nonword and pseudoword was used just once in the stimulus set. They were matched in length with the 4 Adjs, i.e. for each real adjective, 1 nonword and 1 pseudoword were created with the same length (examples in Table 1).

Across all 6 experimental conditions, within each item, the exact same noun was used (e.g. “turtle”, in Tables 1 and 2), resulting in exactly matched stimuli. For the *subsective* condition, 11 pairs of gradable antonyms (e.g. fast/slow; big/small) were created. We then combined those Adjs with Ns to form typical Adj-N combinations (e.g. “a slow turtle”) and atypical but non-anomalous combinations (e.g. “a fast turtle”). Importantly, Ns were selected so that the scalar meaning of the Adj becomes apparent: e.g. if all turtles are slow animals, then even a fast turtle is a slow animal. Each of the subsective adjectives was repeated 4 times, twice in each typicality condition. For the *intersective* conditions, we used 20 Adjs in E1 and 19 Adjs in E2, denoting either colour or shape attributes. Substance Adjs as used in previous studies (e.g. wooden, glass, plastic) were not used, because such Adj-N combinations are relatively infrequent in Norwegian (Schumacher, 2013). For some intersective Adjs, it was impossible to create Adj pairs that can be used with different Ns and are matched across the typicality conditions. Rather, we matched the category from which Adjs were drawn, across typicality conditions within the same item (e.g. colour or shape). Moreover, each Adjective was used at least 3 times, and no more than 6 times in E1 and 7 times in E2, and at least once in each of the two typicality conditions (see Supplementary Materials).

To assess differences in adjective–noun co-occurrences across the four semantic conditions, we extracted Adj-N frequencies from NoWaC (Norwegian Web as Corpus, containing ~700 million words; Guevara, 2010). Co-occurrences of the adjective–noun combinations, based on lemma forms, were generally very low, or even 0 (in 97 items). As expected, frequencies were higher in the typical conditions than in the atypical ones, but in any case well below 1 occurrence per million (typical intersective: 0.05; typical subsective: 0.02). Corpus data are given in the Supplementary Materials. In NoWaC, individual lemma frequencies per million entries are: intersective Adjs, E1:  $M=26.42$ ,  $SD=31.6$ , and E2:  $M=27.45$ ,  $SD=32.1$ ; subsective Adjs, E1 and E2:  $M=228.36$ ,  $SD=418$ ; Ns, E1 =  $33.96$ ,  $SD=95.4$ , and E2 =  $19.49$ ,  $SD=27.7$ .

After each trial, participants were asked two questions in all four semantic conditions. The questions were chosen to address the typicality and denotation (subsective vs intersective) manipulations: “Is it a common [N]?” (typicality); “Is it a [Adj] [superordinate category N]?” (denotation). The denotation question provided an appropriate comparison class for the N. The comparison class was a superordinate category, such as animal, building, food etc. The categories were chosen so that only world knowledge, and no specialised knowledge, was needed to answer the questions. Nonword and pseudoword trials were followed by a single question with the following form: “Is it a [superordinate category N]?” (Table 2). Note that it would not be possible to include typicality and denotation questions also for nonsemantic trials, because of the missing Adj. For semantic trials, 50% of the questions required a “yes” answer and 50% required a “no” answer (randomised and counterbalanced over trials and blocks). The resulting 264 trials were shown in 6 blocks. Each block contained all 44 nouns just once, whilst adjectives were drawn equally from the 6 conditions. Trials in each block were randomised so that 2 or more items from the same condition were never shown after each other. The order of the 6 blocks was randomised, resulting in 6 experiment versions.

## 2.2. Participants

Twenty-six native Norwegian speakers (19 women; mean age: 23.3 years; age range 19–33 years) participated in E1 and were included in the final analyses. Thirty native Norwegian speakers (21 women; mean age: 22.9 years; age range 19–43 years) participated in E2 and were included in the final analyses. No participants were excluded based on their responses to the two task questions in either experiment. Sample sizes

were determined based on our previous study (Fritz & Baggio, 2020;  $N=23$ ), which produced robust effects using the same paradigm with Adj-N phrases, the same number of conditions (6) and trials per condition, a similar 2-question task, the same nonsemantic conditions, and a similar 2×2 design for the semantic conditions. All participants in E1-E2 were right handed, had corrected-to-normal or normal vision, and had no history of neurological or psychiatric disorders. The study was approved by the Norwegian Center of Research Data (NSD; projects 60081 and 719026).

## 2.3. Procedure

Participants were seated in a dimly lit, sound-attenuated booth, approximately 90 cm away from an LCD monitor. Phrases were delivered visually using Presentation (Neurobehavioral Systems, Inc.), in lowercase letters, with a white 30-point size Arial font against a dark grey background. Each trial started with a white fixation cross shown on screen for 500 msec, followed by word-by-word presentation of a [Det [Adj N]] phrase. Each word was displayed for 400 msec and followed by a 400 msec inter-word interval (blank screen). After the NP, a white fixation cross was shown for 500 msec, before the first question appeared. Compared to other studies with similar trial presentations, we extended the latencies of the word and of the inter-word intervals to 400 msec. This allowed us to detect possible late ERP effects. The fixation cross appeared 800 msec after the onset of the noun. To answer the two task questions, participants were instructed to use the key “F” on a standard QWERTY keyboard for “Yes” and “J” for “No”, or vice versa: the pairings of the F/J keys to Yes/No responses were counterbalanced across participants. The experiment continued as soon as the participant provided an answer or, if no response was given, after a time limit of 4 sec. In the nonword and pseudoword conditions, only one question was asked, after which a new trial started. In the semantic conditions, the second question followed the first, after a fixation cross in between. The order of the typicality and denotation questions was randomised over trials.

In E1, participants were asked to read silently and carefully each phrase. They were told to answer each question quickly and accurately and that there would be a limited amount of time to answer. However, no instructions were given on *how* to answer the questions. In E2, instead, we highlighted the intersective vs subsective distinction and the atypical vs typical distinction by training participants on how to answer each question appropriately, prior to the EEG session: we showed participants 2 examples of each of the 4 semantic



conditions, including the 2 questions and their correct answers; for each answer, we provided a short explanation as to why that was the expected, appropriate response: e.g. for the denotation question in the subsective atypical condition (e.g. “a fast turtle”), we glossed the example as follows: ““Is it a fast animal?” — No, because fast turtles are not fast compared to most other animals” (see Supplementary Materials). During training, participants could ask questions. The experiment began with a practice block with 2 trials from each condition. Importantly for the purposes of E2, both in the practice block and in the actual experimental blocks, for each trial we provided immediate feedback on whether the participant answered correctly. Between experimental blocks (44 trials), participants could take a break and continue with the experiment when they were ready. E1 took about 45 min on average to complete, including breaks; E2 took approximately 55 min.

## 2.4. Data acquisition

The EEG was recorded from 31 active electrodes (Fp1, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, TP10, P7, P3, Pz, P4, P8, PO9, O1, Oz, O2, PO10), using the actiCAP system by Brain Products, GmbH. The implicit reference channel and the TP10 channel were placed outside the elastic cap on the left and right mastoid respectively. All EEG channels were then re-referenced offline to the averaged signal from the mastoids. EEG data were sampled at 1000 Hz resolution with a 1000 Hz high cutoff filter and a 10 sec time constant. Impedance was kept below 5 kOhm across all channels.

## 2.5. Data analysis

### 2.5.1. Behavioural responses

Accuracy and response time data were analyzed by fitting generalised linear mixed-effects models (GLMMs) in R by using the *lmer* and *glmer* functions of the *lme4* package (Bates et al., 2015). To obtain *p*-values when fitting GLMs with *lmer*, the Satterthwaite approximation was used, as implemented in the *lmerTest* package (Kuznetsova et al., 2015). Tables 6–9 present the results of the models that best fit the accuracy and RT data. Typicality (atypical vs typical) and denotation (intersective vs subsective) were fixed factors (treatment coded). As baselines, we used *atypical* for the typicality variable and *intersective* for the denotation variable. For the random effects structure, we treated Subject and Item as random effects. For each model, we started with the maximal random effects structure

(Barr et al., 2013). Because none of the maximal models converged, we simplified the random effect structure by first dropping covariance between random slopes and random intercepts for item and then for subject, typicality, and denotation. If the model still did not converge, we then also removed the slopes contributing the least to the variance explained, until convergence was achieved. Only correct responses were included in RT models. We further excluded all trials exceeding a 2.5 standard deviation from the mean in each condition per participant and all responses faster than 200 msec. This resulted in the exclusion of 3.01% of all correct trials in E1 and 2.46% of all correct trials in E2. Responses from the two questions were analyzed separately. Only answers to the task questions in the 2 semantic conditions were analyzed, but we report RT means and accuracies also for the two nonsemantic conditions (Table 3).<sup>1</sup>

### 2.5.2. Event-related potentials

EEG data were analyzed using the FieldTrip toolbox (Oostenveld et al., 2011). Epochs were extracted from 200 msec before word onset (the onset of the N or the prenominal stimulus in each phrase, i.e. a nonword, pseudoword, or real adjective). A 1600 msec post-stimulus interval was used to analyse ERPs time locked to the prenominal stimuli, spanning the full Adj-N epoch (excluding the Det and the response interval following the fixation cross). An 800 msec post-stimulus interval was used to analyse ERPs time locked to the N. In all cases, the 200 msec pre-stimulus data were used for baseline correction. Artifacts in the epoched data were detected and rejected using two FieldTrip functions: (1) trials where amplitude values exceeded a threshold of  $\pm 150 \mu\text{V}$  relative to baseline were discarded; (2) trials that contained eye blinks or movements were rejected by means of thresholding *z*-transformed values of the preprocessed EEG data from channels Fp1 and Fp2, in the 1-15 Hz band. The data were filtered with a digital low-pass filter at 30 Hz. Based on these criteria, on average 41.3 trials per condition per participant were retained for ERP data analysis in E1

**Table 3.** Descriptive statistics of response accuracy (0-1) and response time (msec) data for the nonsemantic conditions in experiment 1 (N=26) and experiment 2 (N=30).

| Condition (experiment 1) | Response accuracies |      | Response times |        |
|--------------------------|---------------------|------|----------------|--------|
|                          | Mean                | SD   | Mean           | SD     |
| Nonword (NW)             | 0.67                | 0.43 | 1118.91        | 326.12 |
| Pseudoword (PW)          | 0.68                | 0.39 | 1080.81        | 256.91 |
| Condition (experiment 2) | Mean                | SD   | Mean           | SD     |
| Nonword (NW)             | 0.86                | 0.26 | 1514.64        | 270.98 |
| Pseudoword (PW)          | 0.88                | 0.25 | 1497.89        | 250.85 |

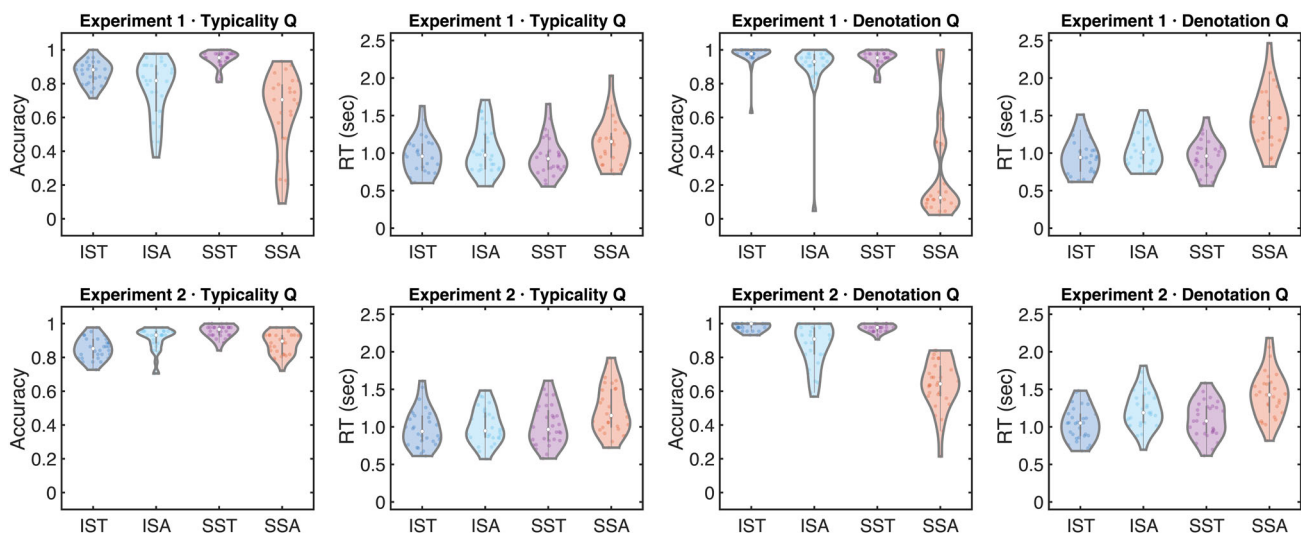
and 40.3 in E2. Next, artifact-free epochs were averaged per participant in each condition to obtain condition-specific ERPs. In our statistical analyses, we used the cluster-based permutation approach of Maris and Oostenveld (2007) with the standard  $\alpha=0.05$  for both the sample- and cluster-level (Monte Carlo)  $p$ -values. To identify the channels and time points at which two conditions differed, ERPs were compared via a  $t$  test in each sample: (electrode, time point) pair. Samples in which the  $t$  values exceeded the 95th quantile of a  $T$  distribution were used to derive spatiotemporally connected clusters of electrode-time point pairs (samples). Our clustering criteria assumed a minimum of 2 adjacent time points and 2 adjacent channels. The algorithm searched for clusters in the entire epochs,  $[-200, 1600]$  or  $[-200, 800]$  msec, where 0 msec is the onset of the critical word. Cluster-level  $t$  values were computed as the sum of sample-level  $t$  values, while cluster-level  $p$  values were estimated using Monte Carlo simulations: participant-specific ERP averages in each of two experimental conditions were collected in one set, which was then randomly partitioned into two equally sized subsets; a dependent-samples  $t$  test was used to compare the means of the subsets; this was repeated 1000 times;  $p$  values were then estimated as the proportion of random partitions ( $x/1000$ ) resulting in a larger  $t$  statistic than the observed one. The output is: a (possibly empty) set of spatio-temporal clusters where the conditions, considered pairwise, differed; the sum of  $t$  statistics in each given cluster ( $T_{\text{sum}}$ ); Monte Carlo  $p$ -value estimates; and cluster size ( $S$ ) in number of samples (Tables 10 and 11). This method addresses the multiple comparisons problem by (a) requiring samples to cluster together spatially and temporally, instantiating a

neurophysiologically plausible constraint on MEEG effects, and (b) replacing sample-level with cluster-level statistics for inferential purposes (Maris & Oostenveld, 2007). We further used GLMs to test whether ERP amplitudes in each condition (in the intervals and channels corresponding to statistical clusters; see Results) predicted accuracies or response times in those conditions. We fitted GLMs in R by using the same functions and model structures as above, but now including ERP amplitudes in each condition as predictors, for each question and in each experiment separately.

### 3. Results

#### 3.1. Behavioural responses

In experiment 1, we found high accuracies across conditions, with the exception of *atypical subsective* trials (SSA; Figure 1; Table 4), where accuracy was down to 0.63 in the typicality question and to 0.29 in the denotation question. This suggests that participants interpret a substantial fraction of SSA trials in ways inconsistent with the patterns expected according to linguistic theory. In particular, in the typicality question, many trials (i.e.  $\sim 37\%$ ) were interpreted as typical (e.g. a “fast turtle” is a common turtle); in the denotation question, a large majority of trials (i.e.  $\sim 71\%$ ) was interpreted as non-subsective or intersective (e.g. a “fast turtle” is a fast animal). It took longer to produce correct responses in SSA trials than in the other semantic conditions (Figure 1; Table 4), especially in the denotation questions. This indicates that the task is most demanding with subsective atypical (SSA) trials. These observations are confirmed by statistical analyses, which revealed



**Figure 1.** Violin plots and box plots of response accuracies (0-1) and response times (RT) for correct responses to the typicality and denotation questions in experiments 1 and 2. See Tables 4 and 5 for descriptive statistics and Tables 6–9 for inferential statistics.

clear interaction effects of typicality and denotation, for both questions, on both accuracy and RT data (Tables 6–9). Further, we found effects of typicality and denotation on response accuracies and response times, for both typicality and denotation questions (Tables 6–9), indicating that differences between conditions are not driven only by lower accuracies and longer RTs for SSA trials.

In experiment 2, we found a different pattern of responses overall—one that matches more closely the predictions of linguistic theory. In the typicality question, accuracy improved, in particular in SSA trials (Figure 1; Table 5), now up to 0.89 (between-groups Wilcoxon test:  $W=81$ ,  $p<0.001$ ). The effect of typicality, observed in experiment 1 (more errors in atypical trials), disappears in experiment 2, as a result of better accuracies in ISA and SSA (Table 6). The interaction between typicality and denotation persists in experiment 2: SSA is still the condition with most errors in typicality questions, in both experiments. We found a similar pattern in RT data in typicality questions: the typicality effect (slower responses in atypical trials) in experiment 1 is clearly reduced in experiment 2. However, in SSA trials, responses were *slower* in experiment 2 ( $W=900$ ,  $p<0.001$ ). We found again an interaction of typicality and denotation (Table 7): SSA is the condition in which producing a correct response takes the longest, in both experiments. In brief, the performance gain in the typicality question in experiment 2 comes at the cost of slower responses.

In the denotation question, too, we found a clear accuracy improvement in experiment 2, in particular in SSA trials (Figure 1; Table 5), where it now reached 0.64 ( $W=124.5$ ,  $p<0.001$ ). Yet, the interaction of typicality and denotation remained significant, reflecting the fact that most errors are still made in the SSA condition (Figure 1; Table 8): a sizeable portion of SSA trials was interpreted as non-subjective (e.g. a “fast turtle” is a

**Table 4.** Descriptive statistics of response accuracy (0-1) and response time (msec) data for the semantic conditions in experiment 1 (N=26).

| Condition (Typicality question) | Response accuracies |      | Response times |        |
|---------------------------------|---------------------|------|----------------|--------|
|                                 | Mean                | SD   | Mean           | SD     |
| Intersective Typical (IST)      | 0.86                | 0.07 | 961.61         | 250.82 |
| Intersective Atypical (ISA)     | 0.77                | 0.17 | 1033.36        | 307.09 |
| Subjective Typical (SST)        | 0.95                | 0.05 | 950.16         | 265.28 |
| Subjective Atypical (SSA)       | 0.63                | 0.24 | 1154.47        | 306.31 |
| Condition (Denotation question) | Mean                | SD   | Mean           | SD     |
| Intersective Typical (IST)      | 0.97                | 0.07 | 951.57         | 233.96 |
| Intersective Atypical (ISA)     | 0.89                | 0.18 | 1043.64        | 236.41 |
| Subjective Typical (SST)        | 0.97                | 0.05 | 958.55         | 220.40 |
| Subjective Atypical (SSA)       | 0.29                | 0.31 | 1468.53        | 406.07 |

Typicality question: “Is it a common [Noun]?” Denotation question: “Is it a [Adj] [superordinate category Noun]?”

**Table 5.** Descriptive statistics of response accuracy (0-1) and response time (msec) data for the semantic conditions in experiment 2 (N=30).

| Condition (Typicality question) | Response accuracies |      | Response times |        |
|---------------------------------|---------------------|------|----------------|--------|
|                                 | Mean                | SD   | Mean           | SD     |
| Intersective Typical (IST)      | 0.85                | 0.06 | 983.22         | 251.77 |
| Intersective Atypical (ISA)     | 0.91                | 0.07 | 1010.28        | 234.76 |
| Subjective Typical (SST)        | 0.95                | 0.04 | 1021.56        | 268.86 |
| Subjective Atypical (SSA)       | 0.89                | 0.07 | 1231.57        | 315.27 |
| Condition (Denotation question) | Mean                | SD   | Mean           | SD     |
| Intersective Typical (IST)      | 0.98                | 0.02 | 1044.97        | 222.59 |
| Intersective Atypical (ISA)     | 0.87                | 0.13 | 1231.79        | 254.53 |
| Subjective Typical (SST)        | 0.98                | 0.02 | 1101.14        | 244.13 |
| Subjective Atypical (SSA)       | 0.64                | 0.15 | 1431.83        | 316.69 |

Typicality question: “Is it a common [Noun]?” Denotation question: “Is it a [Adj] [superordinate category Noun]?”

**Table 6.** Summary of effects of the full mixed logistic model for response accuracy data for the typicality question in experiment 1 (N=26) and experiment 2 (N=30).

| Predictor (experiment 1)       | $\beta$ | SE    | z      | p      |
|--------------------------------|---------|-------|--------|--------|
| Intercept                      | 1.633   | 0.269 | 6.068  | <.0001 |
| Typicality                     | 0.933   | 0.418 | 2.236  | .025   |
| Denotation                     | -0.831  | 0.286 | -2.906 | .004   |
| Denotation $\times$ Typicality | 2.332   | 0.308 | 7.571  | <.0001 |
| Predictor (experiment 2)       | $\beta$ | SE    | z      | p      |
| Intercept                      | -2.443  | 0.149 | 16.405 | <.0001 |
| Typicality                     | 0.220   | 0.239 | 0.921  | .357   |
| Denotation                     | 0.226   | 0.129 | 1.753  | .080   |
| Denotation $\times$ Typicality | -1.644  | 0.203 | -8.083 | <.0001 |

**Table 7.** Summary of effects of the full mixed logistic model for response time data for the typicality question in experiment 1 (N=26) and experiment 2 (N=30).

| Predictor (experiment 1)       | $\beta$  | SE     | t       | p      |
|--------------------------------|----------|--------|---------|--------|
| Intercept                      | 3.959    | 0.0225 | 175.818 | <.0001 |
| Typicality                     | -0.02456 | 0.0112 | -2.190  | .0311  |
| Denotation                     | 0.04457  | 0.0093 | 4.769   | <.0001 |
| Denotation $\times$ Typicality | -0.0481  | 0.0124 | -3.883  | <.0001 |
| Predictor (experiment 2)       | $\beta$  | SE     | t       | p      |
| Intercept                      | 3.958    | 0.019  | 203.10  | <.0001 |
| Typicality                     | -0.015   | 0.001  | -1.55   | .127   |
| Denotation                     | 0.085    | 0.011  | 7.64    | <.0001 |
| Denotation $\times$ Typicality | -0.068   | 0.012  | -5.89   | <.0001 |

**Table 8.** Summary of effects of the full mixed logistic model for response accuracy data for the denotation question in experiment 1 (N=26) and experiment 2 (N=30).

| Predictor (experiment 1)       | $\beta$ | SE     | z       | p      |
|--------------------------------|---------|--------|---------|--------|
| Intercept                      | 2.3768  | 0.1902 | 12.497  | <.0001 |
| Typicality                     | 1.3744  | 0.1991 | 6.903   | <.0001 |
| Denotation                     | -3.8547 | 0.1408 | -27.387 | <.0001 |
| Denotation $\times$ Typicality | 3.9030  | 0.2819 | 13.846  | <.0001 |
| Predictor (experiment 2)       | $\beta$ | SE     | z       | p      |
| Intercept                      | -2.609  | 0.307  | -8.503  | <.0001 |
| Typicality                     | -2.476  | 0.240  | 10.334  | <.0001 |
| Denotation                     | 1.871   | 0.343  | 5.450   | <.0001 |
| Denotation $\times$ Typicality | -1.184  | 0.317  | -3.74   | .0002  |

**Table 9.** Summary of effects of the full mixed logistic model for response time data for the denotation question in experiment 1 ( $N=26$ ) and experiment 2 ( $N=30$ ).

| Predictor (experiment 1)       | $\beta$ | SE      | $t$     | $p$    |
|--------------------------------|---------|---------|---------|--------|
| Intercept                      | 3.978   | 0.0193  | 206.689 | <.0001 |
| Typicality                     | -0.0354 | 0.0066  | -5.378  | <.0001 |
| Denotation                     | 0.1172  | 0.01410 | 8.310   | <.0001 |
| Denotation $\times$ Typicality | -0.1144 | 0.01294 | -8.843  | <.0001 |
| Predictor (experiment 2)       | $\beta$ | SE      | $t$     | $p$    |
| Intercept                      | 4.055   | 0.017   | 243.42  | <.0001 |
| Typicality                     | -0.065  | 0.012   | -5.641  | <.0001 |
| Denotation                     | 0.066   | 0.010   | 6.318   | <.0001 |
| Denotation $\times$ Typicality | -0.053  | 0.012   | 4.444   | <.0001 |

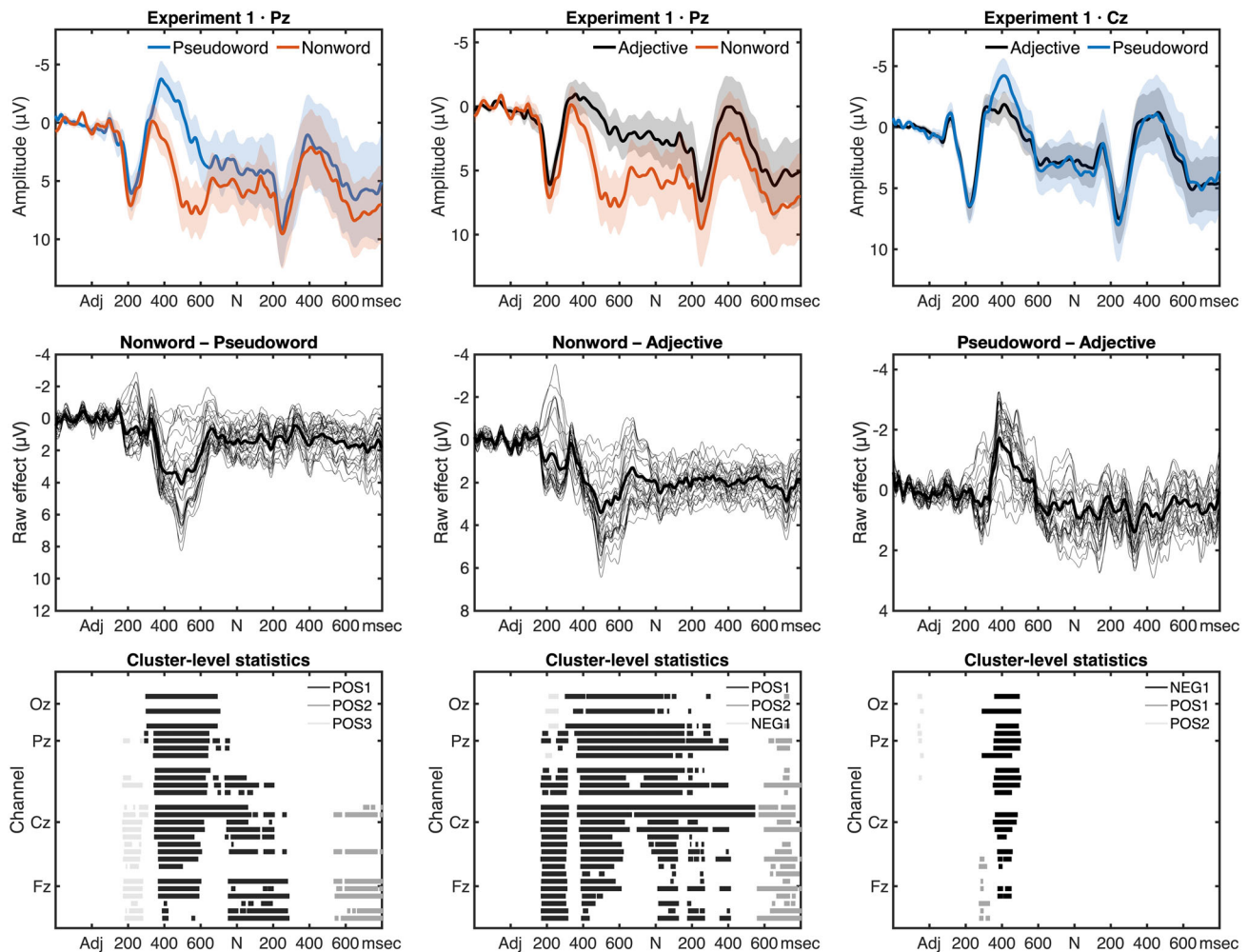
fast animal). Errors were in the same direction as in experiment 1, in spite of training, feedback, and clear performance gains. This parallels what we found for the typicality question. However, for the denotation questions, accuracy improvements were not accompanied by longer response times, which were similar to experiment

1 (Tables 4 and 5;  $W=359$ ,  $p=0.8102$ ). SSA is the slowest condition, as is reflected by the typicality by denotation interaction on RT data (Table 9). Performance gains were also observed between experiments 1 and 2 for the baseline questions, but this resulted once again in longer response times (Table 3), as for the typicality question. These results show that our task manipulation in experiment 2 had a clear effect on performance, across all conditions, with a particular impact on subjective atypical phrases.

### 3.2. Event-related potentials

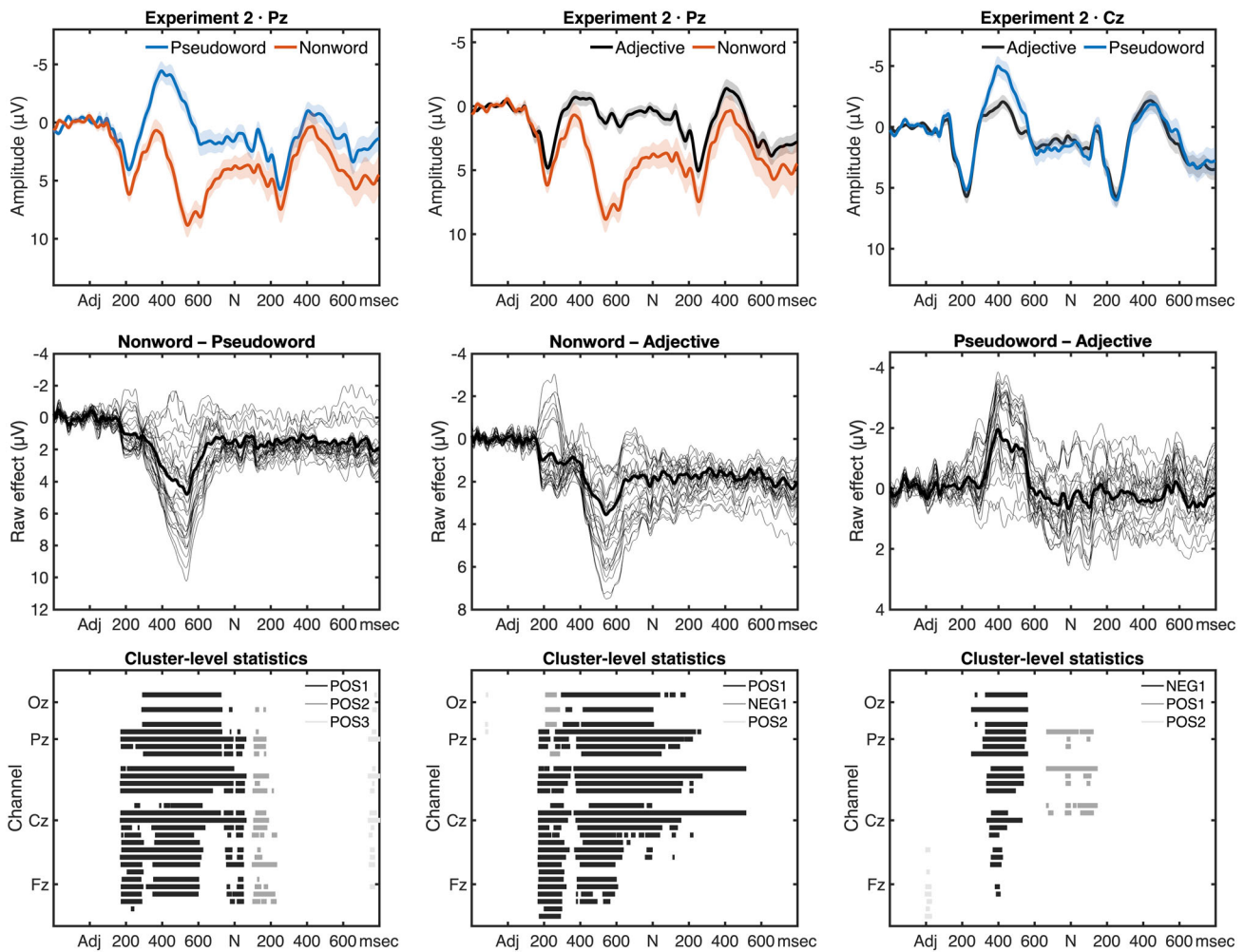
#### 3.2.1. Long epochs: stimulus type effects in the adj-N interval

Pairwise contrasts between pseudowords, nonwords, and real adjectives showed clear ERP effects, in line with earlier results (Barber et al., 2013; Kim & Lai, 2012;



**Figure 2.** Grand-average ERP waveforms and raw effects of pairwise comparisons between nonwords, pseudowords, and adjectives and results of cluster-based permutation statistics from experiment 1 ( $N=26$ ). The middle-row plots show raw effects across channels in each pairwise contrast: each line is a difference wave from one channel; the thick black line is the mean. The bottom-row plots display statistical clusters from Table 10: a tick mark indicates that the ERP difference between conditions is significant (at sample-level  $\alpha=0.05$ ) in a given sample, (electrode, time point) pair. Adjective onset is at 0 msec.





**Figure 3.** Grand-average ERP waveforms and raw effects of pairwise comparisons between nonwords, pseudowords, and adjectives and results of cluster-based permutation statistics from experiment 2 ( $N=30$ ). The middle-row plots show raw effects across channels in each pairwise contrast: each line is a difference wave from one channel; the thick black line is the mean. The bottom-row plots display statistical clusters from Table 10: a tick mark indicates that the ERP difference between conditions is significant (at sample-level  $\alpha=0.05$ ) in a given sample, (electrode, time point) pair. Adjective onset is at 0 msec.

Kounios & Holcomb, 1994; Meade et al., 2019; Ziegler et al., 1997). Nonwords elicited larger positivities compared to both pseudowords and real adjectives (Figures 2 and 3). In both experiments, we found a large positive cluster (POS1; Table 10), with a broad distribution over time and channels. The polarity, latency, and scalp topography of this effect are compatible with a P300-type of response (Ziegler et al., 1997): raw effects (Figures 2 and 3, middle rows) appear at  $\sim 300$  msec post stimulus onset (nonword, pseudoword, adjective) and wane around 700 msec. Statistical effects persist into the noun interval (Figures 2 and 3, bottom rows). In both datasets, we found an N400-type of response to pseudowords relative to adjectives (Figures 2 and 3), manifested as a single largest negative cluster (NEG1; Table 10) between 300 and 500 msec after stimulus onset. Raw effects show a characteristic peak at

400 msec (Figures 2 and 3, middle rows), and statistical effects indicate that the N400 here has a broad topographical distribution, but does not extend to the noun window (Figures 2 and 3, bottom rows). Typicality and denotation (the intersective/subsective distinction) were expected to have an impact on real-time neural processing only when the noun is encountered. Still, we assessed whether ERP signals already at the adjective differ along those dimensions. We found that they did not. There were no clusters of differential activity for either subsective vs intersective adjectives, or for atypical vs typical adjectives (i.e. for adjectives that, if combined with a noun, result in atypical or typical combinations), in either experiment (Table 10). In summary, ERP effects in the Adj-N interval are primarily driven by stimulus type: nonwords elicited P300-type effects relative to both pseudowords and real adjectives,



**Table 10.** Summary of cluster-level permutation statistics for ERP data time locked to the adjective and spanning the Adj-N interval. For each cluster, we report the sum of sample-level statistics ( $T_{sum}$ ), the cluster-level permutation-based Monte Carlo  $p$ -value, and cluster size in number of samples ( $S$ ).

| Experiment 1 (Fig. 2)     | Cluster I ( $T_{sum}; p; S$ ) | Cluster II ( $T_{sum}; p; S$ ) | Cluster III ( $T_{sum}; p; S$ ) |
|---------------------------|-------------------------------|--------------------------------|---------------------------------|
| Nonword – Pseudoword      | POS1: 43394.75; <0.001; 9878  | POS2: 4744.39; 0.052; 1609     | POS3: 4110.04; 0.061; 1403      |
| Nonword – Adjective       | POS1: 53090.3; <0.001; 16304  | POS2: 6633.41; 0.027; 2475     | NEG1: -547.67; 0.321; 192       |
| Pseudoword – Adjective    | NEG1: -7320.39; 0.007; 2204   | POS1: 440.89; 0.355; 178       | POS2: 220.7; 0.562; 92          |
| Atypical – Typical        | POS1: 824.9; 0.176; 348       | POS2: 464.21; 0.296; 185       | POS3: 319.82; 0.392; 118        |
| Subsective – Intersective | NEG1: -261.28; 0.518; 96      | NEG2: -261.24; 0.518; 105      | NEG3: -249.93; 0.533; 102       |
| Experiment 2 (Fig. 3)     | Cluster I ( $T_{sum}; p; S$ ) | Cluster II ( $T_{sum}; p; S$ ) | Cluster III ( $T_{sum}; p; S$ ) |
| Nonword – Pseudoword      | POS1: 54688.58; 0.002; 11910  | POS2: 2574.1; 0.086; 1078      | POS3: 979.13; 203; 419          |
| Nonword – Adjective       | POS1: 52317.65; 0.004; 14111  | NEG1: -703.46; 0.25; 238       | POS2: 60.94; 0.865; 27          |
| Pseudoword – Adjective    | NEG1: -10574.82; 0.031; 3114  | POS1: 2410.89; 0.115; 935      | POS2: 408.2; 0.439; 163         |
| Atypical – Typical        | POS1: 494; 0.278; 206         | POS2: 203.24; 0.529; 91        | POS3: 129.73; 0.654; 56         |
| Subsective – Intersective | NEG1: -179.76; 0.645; 71      | NEG2: -146.41; 0.693; 59       | POS2: 58.67; 0.867; 27          |

**Table 11.** Summary of cluster-level permutation statistics for ERP data time locked to the noun. For each cluster, we report the sum of sample-level statistics ( $T_{sum}$ ), the cluster-level permutation-based Monte Carlo  $p$ -value, and cluster size in number of samples ( $S$ ).

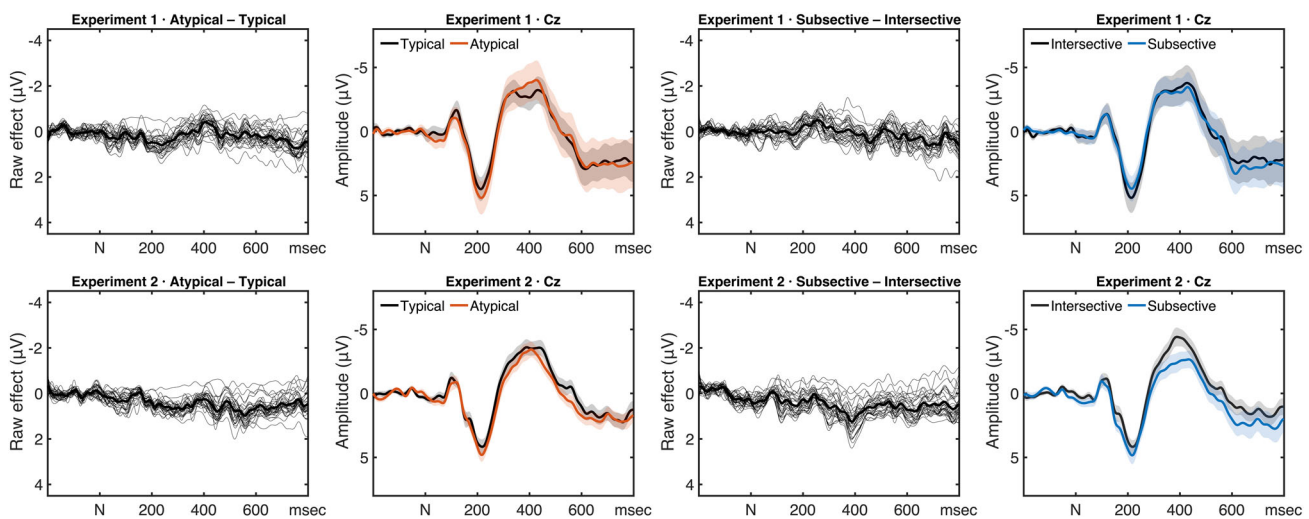
| Experiment 1 (Figs. 4, 5) | Cluster I ( $T_{sum}; p; S$ ) | Cluster II ( $T_{sum}; p; S$ ) | Cluster III ( $T_{sum}; p; S$ ) |
|---------------------------|-------------------------------|--------------------------------|---------------------------------|
| Atypical – Typical        | POS1: 48.16; 0.851; 22        | POS2: 46.55; 0.857; 21         | POS3: 45.64; 0.859; 20          |
| Subsective – Intersective | NEG1: -735.44; 0.239; 292     | POS1: 364.27; 0.461; 148       | NEG2: -142.61; 0.713; 59        |
| Semantic – Nonsemantic    | POS1: 7064.02; 0.014; 2474    | POS2: 3216.33; 0.05; 1091      | POS3: 1950.05; 0.081; 674       |
| Semantic – Pseudoword     | POS1: 15487.32; 0.003; 5378   | POS2: 710.84; 0.24; 239        | NEG1: -221.29; 0.574; 87        |
| Experiment 2 (Figs. 4, 6) | Cluster I ( $T_{sum}; p; S$ ) | Cluster II ( $T_{sum}; p; S$ ) | Cluster III ( $T_{sum}; p; S$ ) |
| Atypical – Typical        | POS1: 898.08; 0.185; 358      | POS2: 829.21; 0.206; 331       | POS3: 398.68; 0.368; 167        |
| Subsective – Intersective | POS1: 844.29; 0.165; 330      | POS2: 505.53; 0.264; 223       | POS3: 413.09; 0.298; 155        |
| Semantic – Nonsemantic    | NEG1: -8045.86; 0.02; 2720    | POS1: 1742.58; 0.068; 721      | POS2: 940.8; 0.139; 370         |
| Semantic – Pseudoword     | NEG1: -2085.62; 0.068; 771    | NEG2: -1063.85; 0.142; 413     | POS1: 950.69; 0.176; 356        |

while pseudowords evoked N400-type effects relative to real adjectives, similarly in the two experiments.

### 3.2.2. Short epochs: semantic effects in the N interval

In the noun interval, we assessed semantic effects across the four semantic conditions (IST, ISA, SST, SSA; effects of

typicality and denotation and pairwise contrasts) and between the semantic and nonsemantic (NW, PW) conditions. We could not detect effects of typicality or denotation in either experiment (Table 11): ERP waves at the noun were similar in atypical and typical trials, and in subsective and intersective trials (Figure 4). There is a discrepancy between behaviour and ERP data:

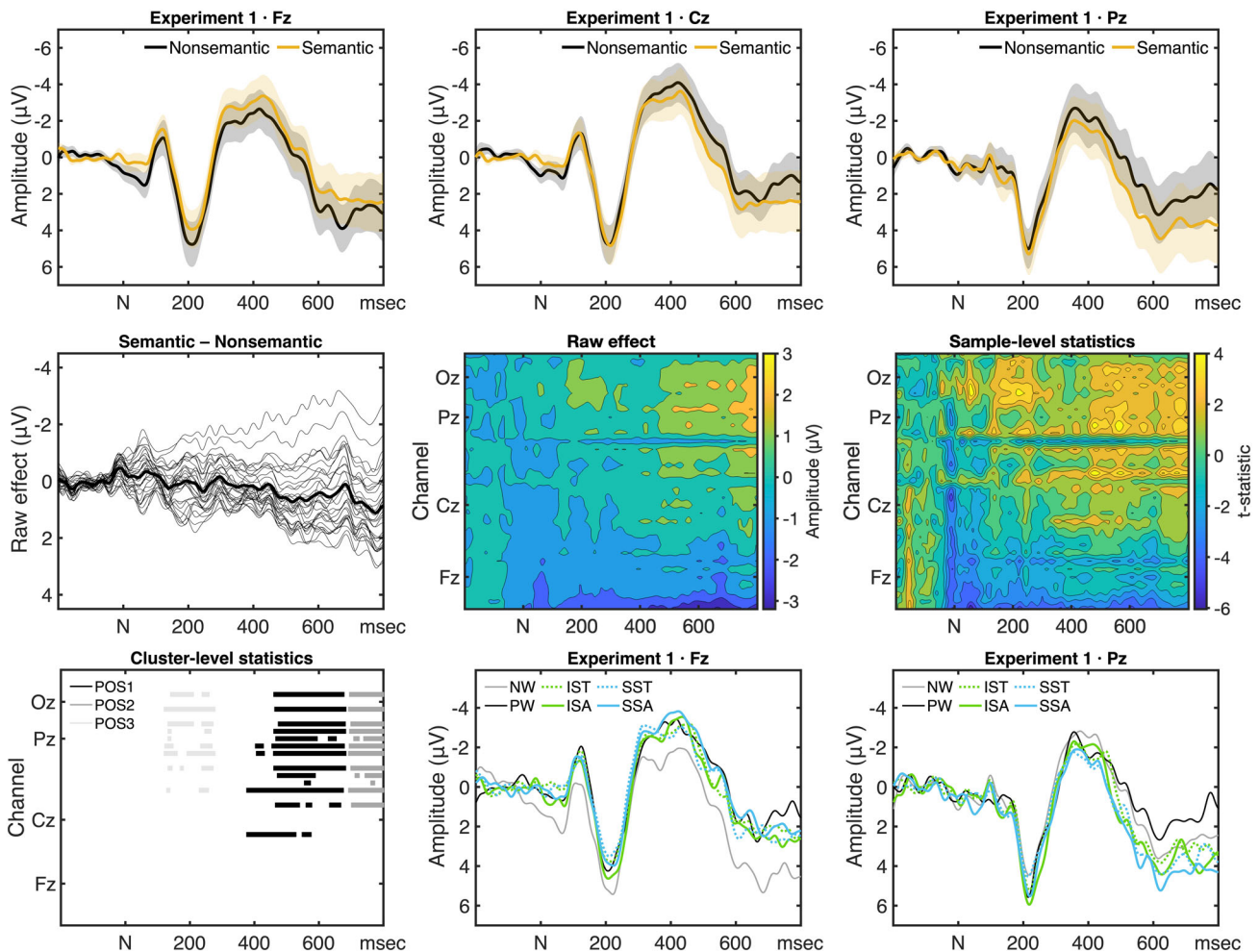


**Figure 4.** Grand-average ERP waveforms and raw effects of pairwise comparisons between atypical vs typical trials and subsective vs intersective trials from experiments 1 and 2. No effects (clusters) in either case were found (Table 11). Noun onset is at 0 msec.

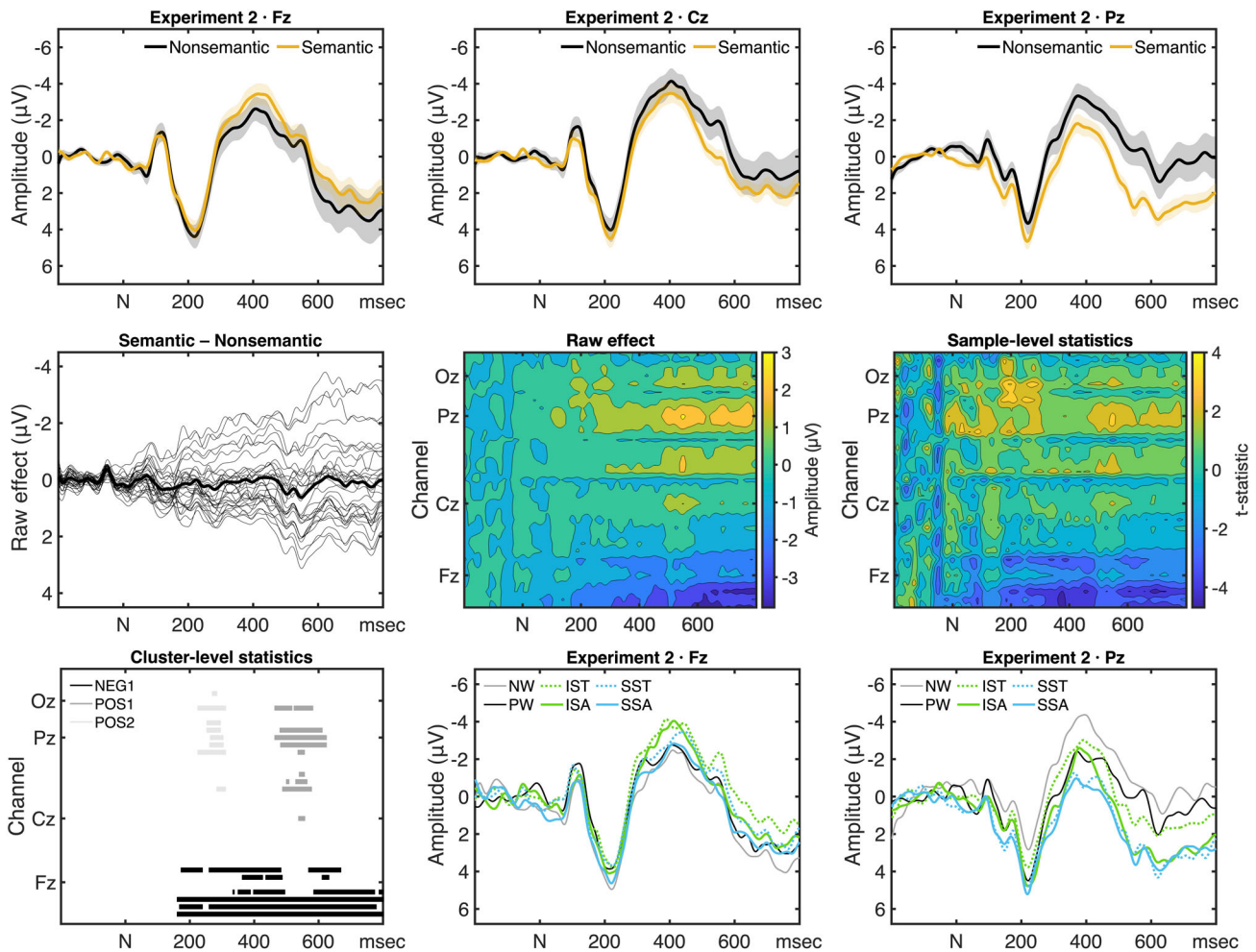
accuracies and response times showed effects of typicality or denotation, or both, in both questions, whereas those effects are absent in ERP data. Our GLM analyses also showed that ERP amplitudes in each of the four semantic conditions did not predict accuracies or response times in those conditions, in the typicality or denotation questions, in both experiments (all effects,  $p > 0.1$ ). We did not find ERP differences between SSA and the other three semantic conditions, which largely patterned together through the N epoch (Figures 5 and 6; all cluster-level  $p$ -values  $> 0.05$ ). Further, comparing the four semantic conditions in pairs (ISA vs IST, SSA vs SST; SST vs IST; SSA vs ISA) failed to reveal any effects (all cluster-level  $p$ -values  $> 0.05$ ). This is the main apparent discrepancy between behavioural data and ERP data: the interaction of typicality and denotation, driven by SSA in accuracies and RTs, in both questions and both experiments, did not correspond to any

detectable ERP differences between SSA and the other conditions.

We found both similarities and differences across the two experiments at the critical noun in semantic trials compared to nonsemantic trials. In experiment 1, semantic trials resulted in more positive amplitude values than nonsemantic trials, particularly in a post-N400 time frame. Specifically, we identified three distinct positive clusters of activity: a highest-ranked cluster (POS1) over centro-parietal electrodes between 350 and 700 msec; a second cluster (POS2) with the same spatial distribution as the first one and adjacent to it temporally; and a third activity cluster (POS3) between 100 and 300 msec (Figure 5). Only cluster POS1 had an associated Monte Carlo  $p$ -value below the 0.05 alpha threshold (Table 11). Cluster POS1 was also identified in the contrast between semantic and pseudoword trials (Table 11). The distributions of the raw effect and



**Figure 5.** Grand-average ERP waveforms and raw effects for the contrast between semantic and nonsemantic (NW, PW) conditions and results of cluster-based permutation statistics from experiment 1 ( $N=26$ ). The middle-row plots display difference waves across channels and contour maps of the ERP effect and of sample-level  $t$ -statistics over time and channels. The bottom-row plots show the statistical clusters in the semantic vs nonsemantic contrast from Table 11 and ERP waves for all conditions from two channels. Noun onset is at 0 msec.



**Figure 6.** Grand-average ERP waveforms and raw effects for the contrast between semantic and nonsemantic (NW, PW) conditions and results of cluster-based permutation statistics from experiment 2 ( $N=30$ ). The middle-row plots display difference waves across channels and contour maps of the ERP effect and of sample-level  $t$ -statistics over time and channels. The bottom-row plots show the statistical clusters in the semantic vs nonsemantic contrast from Table 11 and ERP waves for all conditions from two channels. Noun onset is at 0 msec.

of sample-level statistics over time and channels and the spatio-temporal properties of POS1 are compatible with the characteristics of a P600 effect (Figure 5). We take up issues of classification and functional interpretation in Discussion.

In experiment 2, we found a similar positive-going ERP shift in semantic trials compared to nonsemantic trials, especially late in the N epoch and over posterior channels: this is visible in ERP waveforms and in the temporal evolution of the raw ERP effect and of sample-level statistics (Figure 6). We found two positive clusters, similar to experiment 1, around 5–600 msec and around 2–300 msec (POS1, POS2; Figure 6), but these were the second- and third- ranked clusters overall, and their associated  $p$ -values were above the 0.05  $\alpha$  (Table 11). The highest ranked cluster here was a *negative* one (NEG1; Table 11; Figure 6), with an anterior topographical distribution and a sustained temporal profile,

starting around 200 msec and lasting through the epoch. NEG1 was also found in the comparison between semantic trials and pseudoword trials. In the dataset from experiment 1, we could not find, among clusters with rank lower than 3, negative clusters with the same spatiotemporal characteristics. The spatial and temporal features of the late positivity (POS1) are consistent with a P600 effect, while those of the observed slow negativity (NEG1) match the profile of sustained anterior negativities (SAN or Nref) reported in other language processing experiments.

#### 4. Discussion

In the present study, we aimed at investigating the neural correlates of semantic processing in minimal phrases where a noun (the critical word) is combined with a determiner and an adjective, e.g. “a green

turtle". We compared four semantic conditions, which are assumed to involve composition, [Det [Adj N]], with two nonsemantic conditions, which do not undergo semantic composition, [Det nonword N] and [Det [pseudo-Adj N]]. This contrast revealed a larger P600, if participants performed the task without explicit instructions or trial-to-trial feedback (experiment 1), or a larger sustained anterior negativity (SAN), if they were made to pay attention to meaning by instructions and in-task feedback (experiment 2). Our study does not provide conclusive evidence that either the P600 or the SAN are ERP *correlates* (in a narrow, technical sense) of semantic composition. Yet, the fact that these ERP components and effects have been associated to linguistic processes in the literature suggests that, here too, they may reflect linguistic or cognitive processes involved in the real-time construction of phrasal meaning. Importantly, we view this as a hypothesis to be further assessed, not as a conclusion following from our results. This issue is discussed in section 4.1.

Our study comprised a manipulation of two well-studied variables in the (psycho)linguistic literature: *typicality* (e.g. "an orange turtle"/"a fast turtle") and *denotation*, using subsective adjectives (e.g. "a slow turtle"/"a fast turtle"). A 2-question task shared by both experiments was devised to test whether the distinction between intersective and subsective adjectives is apparent to speakers of Norwegian (denotation question) and whether they distinguish between typical and atypical phrasal combinations (typicality question). In neither of the two experiments did we find differences in ERP waveforms when comparing intersective vs subsective trials, typical vs atypical trials, or the 4 semantic conditions in pairs: there were no on-line ERP effects of typicality and denotation, and there were no interactions between the two. However, we found off-line behavioural differences in both experiments. Notably, in experiment 1, a large portion of all trials in the subsective atypical (SSA) condition received an intersective reading, as indicated by low accuracies in both the typicality and denotation questions. Although instructions and feedback were given in experiment 2, the intersective reading prevailed in about one third of trials for the denotation question. Still, performance improved in experiment 2: responses matched more closely the expected patterns based on linguistic theory. Apparent discrepancies between the behavioural and ERP data are further discussed in section 4.2.

ERP effects at the prenominal stimuli were very similar in the two experiments. We found a larger P300 for nonwords compared to pseudowords and real adjectives, and a larger N400 for pseudowords

relative to adjectives. A P300 effect in response to nonwords, compared to real words and pseudowords, has been previously reported by Ziegler et al. (1997). In their experiment, all three stimulus types had the same occurrence probability, and the P300 had a similar latency and distribution to the P300 effect we describe here. In our study, different instructions and trial-to-trial feedback had no effect at the prenominal stimuli, but they had different effects on on-line ERPs at the noun (P600 vs SAN in the semantic vs nonsemantic contrast) and on behaviour (performance gains, especially in the SSA condition).

#### **4.1. Effects of phrasal meaning composition: positive and negative components**

Replicating the results of Fritz and Baggio (2020), we observed both early and late positivities in the comparison between semantic vs nonsemantic trials. The late positivity (P600) is the strongest ERP effect at the noun in experiment 1. It is also found in experiment 2, where it is, however, accompanied by a statistically stronger SAN effect. The waveforms time-locked to the prenominal adjective, letter string, or pseudoword indicate that these effects are not due to spill-over from the preceding stimulus: the sign of effects in the prenominal interval (positive for nonwords vs pseudowords or real adjectives; negative for pseudowords vs real adjectives) is not consistent with the sign of the effects in the noun interval. We are inclined to exclude that the ERP effects observed at the noun are driven primarily by differences in task demands across semantic and nonsemantic trials. If that was the case, we should find ERP effects already at the prenominal stimulus, when it becomes clear to participants what task questions they will get: pseudowords and nonwords should then pattern together (one question) against real adjectives (two questions) in the ERP data. But that was not the case in either experiment. This is, however, an empirical question that deserves to be addressed in a separate experiment. Indeed, one possibility is that in a naturalistic or task-free setting, the effects we found at the noun will be greatly reduced or will disappear altogether.

Nouns following adjectives elicited a more positive P200 effect relative to nouns following pseudowords or nonwords. This P200 response was found in both experiments. In neither experiment did the P200 reach a cluster-level significance threshold. This effect is unlikely to correspond to the early conceptual combination effects in the LATL reported by previous MEG studies (e.g. see Bemis & Pykkänen, 2011, 2013; Blanco-Elorrieta & Pykkänen, 2016; Del Prato & Pykkänen, 2014). Establishing systematic correspondences between EEG and



MEG effects is difficult for several methodological, physiological, and neurophysiological reasons to do with the nature of the signals involved. The temporal profiles of the P200 and of MEG effects from LATL are different: the onset time is comparable, but the duration of our effect is longer, i.e.  $\sim 150$  msec, vs  $\sim 50$ – $70$  msec of the LATL effects of Bemis and Pykkänen's (2011, 2013). As a P200, the effect observed here is likely to be a manifestation of modulations of attention between semantic and nonsemantic trials. Participants may (re)allocate attention differently when processing the noun, depending on whether it is preceded by an adjective (normal or heightened attention) vs by a nonword or a pseudoword (decreased attention as compared to real adjectives). Research suggests that ERP effects in the 150–300 msec range, such as the P200, reflect dynamic, task- or stimulus-dependent (re)allocation of attentional resources (for a review, see Crowley & Colrain, 2004). P200 effects were weak in our study, and our experiments were not designed to manipulate attention independently of semantic load. Further research is needed to draw firmer conclusions on the possible role of dynamic attention reallocation in semantic processing and semantic composition tasks, as well as on the distinctness and independence of the P200 and P600 effects as reported here.

The composition contrast between the semantic and nonsemantic trials resulted in a larger P600 component in phrases where the noun followed a real adjective compared to phrases where it followed a nonword or a pseudoword. This P600 effect is as described by Fritz and Baggio (2020). It was especially evident in experiment 1 and detected in experiment 2, too. One reason for us to include pseudowords (i.e. pseudo-adjectives) is that they resemble real adjectives phonotactically and morphologically: in our stimuli, in each phrase pseudowords agreed with the noun in both gender and number. Pseudowords may therefore be assumed to involve a form of syntactic composition, but not semantic composition (see also Kochari et al., 2021; Neufeld et al., 2016). If that is correct, then the observed P600 could be taken to reflect *semantic* composition: the P600 was also found in a contrast with pseudowords only, which suggests it requires that both the adjective and noun have meaning. A more prudent stance would be to argue that, if pseudowords and nonwords block syntactic *and* semantic composition, the P600 would reflect syntactic *and* semantic composition. In either case, our data do not conclusively link the P600 to syntactic or semantic composition (see below for further discussion), but they are consistent with the hypothesis that syntax-driven meaning composition, along with other processes, can modulate the P600's

amplitude (Baggio, 2018; 2021) and with several proposals relating the P600 to computation at the syntax-semantics interface (Bornkessel-Schlesewsky & Schlewsky, 2008; Kuperberg, 2007). Also, P600-like effects have been described in studies on pragmatic phenomena, such as scalar implicature (Spychalska et al., 2016), metonymy (Schumacher, 2013), and inference (Burkhardt, 2006). These P600 effects have been attributed to the processing costs of reanalysis or updating of the current discourse model (Brouwer et al., 2017; Delogu et al., 2019). Likewise, our results are compatible with the view that the P600 indexes “semantic integration” (Brouwer et al., 2017). There are important differences between these proposals that our data cannot address, however. Our results do not support the idea that meaning composition in phrases is reflected by the N400 (Neufeld et al., 2016). Moreover, the time course of the P600 seems to conflict with the proposal, based on MEG findings (Bemis & Pykkänen, 2011, 2013), that compositional processing occurs at around 400 msec from noun onset (in the vmPFC). That different methods—EEG vs MEG—may reveal different neural signatures of composition, in brain space and time, is a real possibility (but see Fló et al., 2020). Future work will need to integrate these different data types in new ways to resolve apparent inconsistencies.

An alternative account of the observed P600 is that it is, in fact, an instance of P300, which would, in this study, reflect task contingencies. We are open to the suggestion that the P600 does not reflect composition or, more generally, syntactic or semantic processes, but rather decision making or other processes driven by task demands. Dedicated experiments may be designed to pit the composition and decision making accounts against each other. However, prior to that, a more explicit decision-making account should be produced that meets three requirements: (1) it must not be ad hoc or post hoc, but should follow from theory, as does the hypothesis that the P600 reflects composition or syntax/semantics interface processes (see above); (2) it must not be based on assumptions of continuity or connectedness of the observed P200 and P600, which should be demonstrated with an experimental design that, by manipulating the stimuli or the task, allows these two ERP effects to vary independently; needless to say, this line of research would also benefit the development and further testing of the composition account; (3) it has to be distinguished from the hypothesis that the P600 is part of the “P300 family”; this is a long-running and recently revived discussion (Coulson, 1998; Osterhout, 1999; Sassenhagen et al., 2014; Sassenhagen & Bornkessel-Schlesewsky, 2015; Sassenhagen &



Fiebach, 2019), which is still consistent with the existence of different “members of the family”, both phenomenologically and neurophysiologically, as the P300 we found for nonwords and the P600 observed in the semantic conditions in our experiment.

The idea that the P600 is associated with composition or processes at the syntax/semantics interface should be combined with specific hypotheses on how other ERP components may be modulated by processes that feed into composition, that depend on it, or that impact the way it unfolds in real time. For example, it may be expected that on-line effects of typicality and denotation should be manifested as changes in the amplitude of the P600. However, in previous work (Fritz & Baggio, 2020), we found effects of intensionality (temporal vs modal adjectives) and denotation (privative vs nonprivative adjectives) in [Det [Adj N]] phrases on N400 and post-N400 effects at the noun, respectively, while a larger P600 was observed for all semantic vs nonsemantic trials, as in the current study. In formal semantics, composition is seen as a discrete operation that either applies or not does apply to given input (Martin & Baggio, 2019). Different adjective types may determine the structures that are composed or those that result from composition, as well as their inferential properties (e.g. entailment), but will not affect the composition operation as such. On the other hand, complex meanings may be constructed in multiple ways by the brain, depending on properties of the input and context and on available resources (Baggio, 2018; 2021). Our results suggest that sustained anterior negativities (SAN) may also be elicited during composition, when participants have to pay special attention to the meaning of the input in order to perform a task as instructed and receive (positive) feedback at each trial. This result is consistent with work linking SAN effects to higher-order semantic processes (e.g. Baggio et al., 2008; Van Berkum et al., 1999; Van Berkum et al., 2003) and with theories positing a dynamic balance between syntax-driven composition, as hypothetically indexed by the P600, and context-driven interpretive processes, as reflected by N400 and SAN effects (Baggio, 2018, 2021; Michalon & Baggio, 2019).

#### **4.2. On discrepancies between on-line and off-line effects**

Our study design, in particular the distinction between intersective vs subsective adjectives, is based on linguistic theory. By taking formal analyses as a starting point, one can avoid ad hoc notions or distinctions of meaning and generate experimental results that are relevant to confirming or redressing linguistic theory.

Experimental results on adjectives have been used mainly to strengthen the intuition-based “data” in which formal theories are grounded and secondarily to refine or further develop those theories. This is reflected in the task- and dependent measures-choices of experimental studies building on formal semantic theories, which utilise mainly off-line behavioural tasks, such as verification or acceptability judgment tasks, inferencing, and interpretation tasks (Solt, 2018). These studies may help us obtain a clearer picture of what meanings may be derived from phrases or sentences, but they tell us little on how different types of expressions (e.g. adjectives) are processed on-line.

An important observation here is that classical, truth conditional semantic theories specify, for any given expression in the language, what it can mean when the information provided by lexical or constituent meanings and syntax is fully exploited, as required by the principle of compositionality (Baggio et al., 2012). Gradable subsective adjectives exemplify well this point: their meaning is assumed to be such that the entities denoted by a noun phrase (e.g. “a fast turtle”) are not necessarily a subset of the entities denoted by the adjective (i.e. fast Xs, where X is any relevant comparison class, e.g. all animals). It is possible that speakers of a language can assign such “maximal” meanings to all or most phrases or sentences, but it is unclear whether (in what conditions and to what extent) such capacity can play out on-line during processing. Semantic theories that do not follow the truth conditional tradition have assumed that meaning representations can be “minimal” or “underspecified” (for discussion, see Baggio, 2018). This proposal goes along well with psycholinguistic results showing that on-line language processing does not normally result in automatic extraction of the kind of “maximal” meanings posited by traditional theories. Our ERP results fit this narrative. They show that the intersective/subsective distinction, and even the typical/atypical distinction, as tested in the present study, may not be reflected by ERP signals. Behavioural results from experiment 1 show that off-line responses may not follow the formal analyses either, as we found for many atypical subsective trials, where participants classified a fast turtle as a fast animal. When participants paid attention to the subsective vs intersective distinction, their accuracy rates increased, especially for typicality questions. It may be suggested that there are no discrepancies between on-line and off-line effects of denotation: readers may simply not be sensitive to the intersective vs subsective distinction. Note that the typical subsective (SST) condition is not relevant for distinguishing between the two alternative readings, and that the behavioural denotation effect is largely driven

by atypical subsective (SSA) trials. So, the main discrepancy between behavioural and ERP data would concern the latter condition.

How can these observations be accommodated in light of MEG reports of early composition effects in NPs featuring an intersective but not a subsective adjective (Ziegler & Pykkänen, 2016)? Different methodologies may well detect different neural signals that contribute to semantic processing. Moreover, Ziegler and Pykkänen (2016) included a manipulation of the noun's specificity (e.g. "dog" vs "animal"). This makes it difficult to compare their findings of an early and late composition effects to our own results. Finally, Kochari et al. (2021), using the same paradigm in Dutch, did not find the scalar effect of Ziegler and Pykkänen (2016).

The discrepancy between ERP data and behavioural responses in our dataset, particularly in relation to subsective atypical phrases, may also be interpreted in light of the "good enough" approach to language comprehension, put forward by Ferreira et al. (2002). Their proposal is based on the assumption that language processing can (and, in fact, typically does) result in partial, or underspecified, semantic representations. Yet, these meanings are often "good enough" for the purposes of comprehension, given task demands, time, processing load, and cognitive limitations (for a review and synthesis, see Karimi & Ferreira, 2016). However, in addition to this hypothesis, and to formal semantic theories that emphasise minimality and underspecification (see above), we envisage a possible explanatory role for the distinction between on-line and off-line representations. Although meanings computed on-line may be underspecified or "good enough", off-line representations need not be. The latter may "flesh out" elements that remain implicit in the initial on-line representation: that may be the case especially for those semantic elements that are not strictly necessary for understanding the "gist" of the message (see Baggio, 2018; Baggio et al., 2016a, 2016b; Johnson-Laird, 1983; Stenning & van Lambalgen, 2008); the distinctions between typical and atypical phrases and between intersective and subsective adjectives might be of that kind. This may also explain the observed discrepancies between ERP and behavioural effects. What is computed on-line is a representation where the meanings of the adjective and noun are effectively composed together, but where relations of typicality and denotation, among other aspects of meaning, are not explicitly fleshed out—that only happens off-line, after meaning composition, under pressure from the task and given goals; only then do the effects of typicality and denotation become apparent and result in processing costs, such as errors and longer response times. Not all aspects of

meaning as specified by semantic theories leave traces on on-line neural signals, although most (or even all) such aspects are in principle recoverable by competent speakers of a language on the basis of the representations computed on-line. Fritz and Baggio (2020) for example found that privativity as encoded in an adjective (e.g. "fake", "former") is reflected in ERPs for modal privative adjectives, not for temporal privative adjectives. ERPs time-locked to the noun in the phrase "A fake president" elicited a larger negativity in a post N400 frame relative to "A real president". No difference in ERPs was found when comparing "A current president" to "A former president", suggesting that specific aspects of denotation (privativity) or intensionality (modal vs temporal) are used on-line. The behavioural results by Fritz and Baggio (2020) are comparable to those from the current study: null results in the ERPs (e.g. for the non-privative temporal "current" vs the privative temporal "former") were accompanied by off-line differences (main effects of intensionality and denotation).

A large body of work described on-line ERP effects of manipulations that are close enough to the present typicality manipulation to warrant the expectation of an N400 effect in our experiments (Kutas & Federmeier, 2011). However, it is important to note that earlier work that has reproduced reliable N400 typicality effects *always used sentences or discourses*. The current study therefore cannot be seen as a failed replication of a well-known (N400) effect, because there has not been enough research on typicality manipulations at the phrase level. The study which is closest to our own in terms of linguistic stimuli, by Lau et al. (2016), did not manipulate typicality within the adjective–noun combination: it manipulated congruity, resulting in semantic anomaly (e.g. "Innocent bag" vs "Yellow bag"). As shown by Federmeier et al. (2010), low typicality items trigger N400 components whose amplitudes fall between high typicality and incongruous target words. Typicality manipulations as in our study may indeed be less marked than semantic anomalies or violations (Lau et al., 2008; Kutas & Federmeier, 2011). More semantically deviant stimuli than atypical adjective–noun phrases in our stimuli can elicit N400 effects also in minimal phrases (Bekemeier et al., 2019). Further, N400 effects of semantic manipulations tend to be larger toward the end of a sentence, and early on in sentences mostly frequency effects emerge (Van Petten & Kutas, 1990): semantic constraints need to be built up by several successive words before they can modulate N400 amplitudes. This would explain why variables that modulate N400 effects in sentences may not produce comparable effects in phrases. As a final

point, we note that even the stronger congruity manipulation at the phrasal level, as used by Neufeld et al. (2016), only elicited a weak effect of congruity in two studies. In one study, they just found an interaction between congruity and hemisphere in a by-quadrant analysis. No other contrasts involving congruity yielded significant effects. Also, although some studies found off-line behavioural differences of typicality (Lucas, 2001; Smith et al., 1988), Molinaro et al. (2012) found no N400 effects of typicality in phrases embedded in sentences (Segaert et al., 2019). Understanding the conditions in which the N400 component is modulated in phrases, as opposed to sentences or discourses, is an important area of future research. Likewise, the P600 and SAN effects we report here, and that we hypothetically associate with meaning construction processes, should be replicated and further tested in alternative versions of our paradigm and in other paradigms, with a focus on matching more closely composition and non-composition trials, in terms of both stimuli and task.

## Notes

1. Regarding accuracies in the nonsemantic conditions, participants consistently answered the questions with either “Yes” or “No” and kept to one response throughout the session. As the task in the nonsemantic conditions is not relevant to our research questions, we refrain from interpreting these results here.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This research is funded by Norwegian Research Council FRIPRO grant 251219.

## Author contributions

IF and GB designed the experiment. IF constructed the stimuli and collected the data. IF and GB analyzed the data, interpreted the data, and wrote the paper.

## ORCID

Giosuè Baggio  <http://orcid.org/0000-0001-5086-0365>

## References

Baggio, G. (2012). Selective alignment of brain responses by task demands during semantic processing. *Neuropsychologia*, 50(5), 655–665. <https://doi.org/10.1016/j.neuropsychologia.2012.01.002>

Baggio, G. (2018). *Meaning in the brain*. MIT Press.

Baggio, G. (2021). Compositionality in a parallel architecture for language processing. *Cognitive Science*, 45(5), e12949. <https://doi.org/10.1111/cogs.12949>

Baggio, G., Cherubini, P., Pischredda, D., Blumenthal, A., Haynes, J. D., & Reverberi, C. (2016b). Multiple neural representations of elementary logical connectives. *NeuroImage*, 135, 300–310. <https://doi.org/10.1016/j.neuroimage.2016.04.061>

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, 26(9), 1338–1367. <https://doi.org/10.1080/01690965.2010.542671>

Baggio, G., Stenning, K., & van Lambalgen, M. (2016a). Semantics and cognition. In M. Aloni, & P. Dekker (Eds.), *The Cambridge Handbook of formal semantics* (pp. 756–774). Cambridge University Press.

Baggio, G., Van Lambalgen, M., & Hagoort, P. (2008). Computing and recomputing discourse models: An ERP study. *Journal of Memory and Language*, 59(1), 36–53. <https://doi.org/10.1016/j.jml.2008.02.005>

Baggio, G., Van Lambalgen, M., & Hagoort, P. (2012). The processing consequences of compositionality. In M. E. Werning, W. E. Hinzen, & E. E. Machery (Eds.), *The Oxford handbook of compositionality* (pp. 655–672). Oxford University Press.

Barber, H. A., Otten, L. J., Kousta, S.-T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, 125(1), 47–53. <https://doi.org/10.1016/j.bandl.2013.01.005>

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 48. <https://doi.org/10.18637/jss.v067.i01>

Bekemeier, N., Brenner, D., Klepp, A., Biermann-Ruben, K., & Indefrey, P. (2019). Electrophysiological correlates of concept type shifts. *PLoS One*, 14(3), e0212624. <https://doi.org/10.1371/journal.pone.0212624>

Bemis, D. K., & Pykkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of Neuroscience*, 31(8), 2801. <https://doi.org/10.1523/JNEUROSCI.5003-10.2011>

Bemis, D. K., & Pykkänen, L. (2013). 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>

Blanco-Elorrieta, E., & Pykkänen, L. (2016). Composition of complex numbers: Delineating the computational role of the left anterior temporal lobe. *NeuroImage*, 124, 194–203. <https://doi.org/10.1016/j.neuroimage.2015.08.049>

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on “semantic P600” effects in language comprehension. *Brain Research Reviews*, 59(1), 55–73. <https://doi.org/10.1016/j.brainresrev.2008.05.003>

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A Neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, 41(S6), 1318–1352. <https://doi.org/10.1111/cogs.12461>

- Burkhardt, P. (2006). Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related brain potentials. *Brain and Language*, 98(2), 159–168. <https://doi.org/10.1016/j.bandl.2006.04.005>
- Calloway, R. C., & Perfetti, C. A. (2017). Integrative and predictive processes in text reading: The N400 across a sentence boundary. *Language, Cognition and Neuroscience*, 32(8), 1001–1016. <https://doi.org/10.1080/23273798.2017.1279340>
- Cosentino, E., Baggio, G., Kontinen, J., & Werning, M. (2017). The time-course of sentence meaning composition. N400 effects of the interaction between context-induced and lexically stored affordances. *Frontiers in Psychology*, 8, 813. <https://doi.org/10.3389/fpsyg.2017.00813>
- Coulson, S. (1998). ERPs and domain specificity: Beating a straw horse. *Language and Cognitive Processes*, 13(6), 653–672. <https://doi.org/10.1080/016909698386410>
- Crowley, K. E., & Colrain, I. M. (2004). A review of the evidence for P2 being an independent component process: Age, sleep and modality. *Clinical Neurophysiology*, 115(4), 732–744. <https://doi.org/10.1016/j.clinph.2003.11.021>
- Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, 135, 103569. <https://doi.org/10.1016/j.bandc.2019.05.007>
- Del Prato, P., & Pyllkänen, L. (2014). MEG evidence for conceptual combination but not numeral quantification in the left anterior temporal lobe during language production. *Frontiers in Psychology*, 5, 524. <https://doi.org/10.3389/fpsyg.2014.00524>
- Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, 44(4), 491–505. <https://doi.org/10.1111/j.1469-8986.2007.00531.x>
- Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4), 469–495. <https://doi.org/10.1006/jmla.1999.2660>
- Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language*, 115(3), 149–161. <https://doi.org/10.1016/j.bandl.2010.07.006>
- Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, 11(1), 11–15.
- Fló, E., Cabana, Á., & Valle-Lisboa, J. C. (2020). EEG signatures of elementary composition: Disentangling genuine composition and expectancy processes. *Brain and Language*, 209, 104837. <https://doi.org/10.1016/j.bandl.2020.104837>
- Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Science*, 6(2), 78–84. [https://doi.org/10.1016/S1364-6613\(00\)01839-8](https://doi.org/10.1016/S1364-6613(00)01839-8)
- Friederici, A. D. (2017). *Language in our brain: The origins of a uniquely human capacity*. MIT Press.
- Fritz, I., & Baggio, G. (2020). Meaning composition in minimal phrasal contexts: Distinct ERP effects of intensionality and denotation. *Language, Cognition and Neuroscience*, 35(10), 1295–1313. <https://doi.org/10.1080/23273798.2020.1749678>
- Guevara, E. R. (2010). NoWaC: a large web-based corpus for Norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, Association for Computational Linguistics, 1–7.
- Hagoort, P. (2003). How the brain solves the binding problem for language: A neurocomputational model of syntactic processing. *Neuroimage*, 20, 18–S29. <https://doi.org/10.1016/j.neuroimage.2003.09.013>
- Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In M. Gazzaniga (Ed.), *The Cognitive neurosciences, 4th ed.* (pp. 819–836). MIT press.
- Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, 304(5669), 438. <https://doi.org/10.1126/science.1095455>
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.
- Kamp, H. (1975). Two theories of adjectives. In E. Keenan (Ed.), *Formal semantics of natural language* (pp. 123–155). Cambridge University Press. Reprinted in 1984.
- Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *The Quarterly Journal of Experimental Psychology*, 69(5), 1013–1040. <https://doi.org/10.1080/17470218.2015.1053951>
- Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable predicates. *Linguistics and Philosophy*, 30(1), 1–45. <https://doi.org/10.1007/s10988-006-9008-0>
- Kennedy, C. (2012). Adjectives. In G. Russell, & D. Graff Fara (Eds.), *Routledge companion to philosophy of language*. Routledge (pp. 328–341). Routledge.
- Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from ERPs. *Journal of Cognitive Neuroscience*, 24(5), 1104–1112. [https://doi.org/10.1162/jocn\\_a\\_00148](https://doi.org/10.1162/jocn_a_00148)
- Kochari, A. R., Lewis, A. G., Schoffelen, J. M., & Schriefers, H. (2021). Semantic and syntactic composition of minimal adjective-noun phrases in Dutch: An MEG study. *Neuropsychologia*, 155, 107754. <https://doi.org/10.1016/j.neuropsychologia.2021.107754>
- Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4), 804–823. <https://doi.org/10.1037/0278-7393.20.4.804>
- Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146, 23–49. <https://doi.org/10.1016/j.brainres.2006.12.063>
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, 62(1), 621–647. <https://doi.org/10.1146/annurev.psych.093008.131123>
- Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427), 203–205. <https://doi.org/10.1126/science.7350657>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmerTest'. *R Package Version*, 2.
- Lau, E. F., Namyst, A., Fogel, A., & Delgado, T. (2016). A direct comparison of N400 effects of predictability and incongruity in adjective-noun combination. *Collabra*, 2(1), 13. <https://doi.org/10.1525/collabra.40>



- Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, 9(12), 920–933. <https://doi.org/10.1038/nrn2532>
- Lucas, M. (2001). Essential and perceptual attributes of words in reflective and on-line processing. *Journal of Psycholinguistic Research*, 30(6), 605–625. <https://doi.org/10.1023/A:1014283106728>
- Lyu, B., Choi, H. S., Marslen-Wilson, W. D., Clarke, A., Randall, B., & Tyler, L. K. (2019). Neural dynamics of semantic composition. *Proceedings of the National Academy of Sciences*, 116(42), 21318–21327. <https://doi.org/10.1073/pnas.1903402116>
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG and MEG data. *Journal of Neuroscience Methods*, 164(1), 177–190. <https://doi.org/10.1016/j.jneumeth.2007.03.024>
- Martin, A. E., & Baggio, G. (2019). Modeling meaning composition from formalism to mechanisms. *Philosophical Transactions of the Royal Society B*, 375(1791), 20190298. <https://doi.org/10.1098/rstb.2019.0298>
- Meade, G., Grainger, J., & Holcomb, P. J. (2019). Task modulates ERP effects of orthographic neighborhood for pseudowords but not words. *Neuropsychologia*, 129, 385–396. <https://doi.org/10.1016/j.neuropsychologia.2019.02.014>
- Michalon, O., & Baggio, G. (2019). Meaning-driven syntactic predictions in a parallel processing architecture: Theory and algorithmic modeling of ERP effects. *Neuropsychologia*, 131, 171–183. <https://doi.org/10.1016/j.neuropsychologia.2019.05.009>
- Molinaro, N., Carreiras, M., & Duñabeitia, J. A. (2012). Semantic combinatorial processing of non-anomalous expressions. *Neuroimage*, 59(4), 3488–3501. <https://doi.org/10.1016/j.neuroimage.2011.11.009>
- Morzycki, M. (2016). *Modification*. Cambridge University Press.
- Neufeld, C., Kramer, S. E., Lapinskaya, N., Heffner, C. C., Malko, A., & Lau, E. F. (2016). The electrophysiology of basic phrase building. *PLoS ONE*, 11(10), e0158446. <https://doi.org/10.1371/journal.pone.0158446>
- Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., & Husband, M. E. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 375(1791), 20180522. <https://doi.org/10.1098/rstb.2018.0522>
- Olstad, A. M. H., Fritz, I., & Baggio, G. (2020). Composition decomposed: Distinct neural mechanisms support processing of nouns in modification and predication contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(11), 2193–2206. <https://doi.org/10.1037/xlm0000943>
- Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). Fieldtrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011. <https://doi.org/10.1155/2011/156869>
- Osterhout, L. (1999). A superficial resemblance does not necessarily mean you are part of the family: Counterarguments to Coulson, King and Kutas (1998) in the P600/SPS-P300 debate. *Language and Cognitive Processes*, 14(1), 1–14. <https://doi.org/10.1080/016909699386356>
- Pylkkänen, L. (2016). Composition of complex meaning: Interdisciplinary perspectives on the left anterior temporal lobe. In G. Hickok, & S. L. Small (Eds.), *Neurobiology of language* (pp. 621–631). Academic Press.
- Sassenhagen, J., & Bornkessel-Schlesewsky, I. (2015). The P600 as a correlate of ventral attention network reorientation. *Cortex*, 66, A3–A20. <https://doi.org/10.1016/j.cortex.2014.12.019>
- Sassenhagen, J., & Fiebach, C. J. (2019). Finding the P3 in the P600: Decoding shared neural mechanisms of responses to syntactic violations and oddball targets. *NeuroImage*, 200, 425–436. <https://doi.org/10.1016/j.neuroimage.2019.06.048>
- Sassenhagen, J., Schlewsky, M., & Bornkessel-Schlesewsky, I. (2014). The P600-as-P3 hypothesis revisited: Single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain and Language*, 137, 29–39. <https://doi.org/10.1016/j.bandl.2014.07.010>
- Schumacher, P. B. (2013). When combinatorial processing results in reconceptualization: Toward a new approach of compositionality. *Frontiers in Psychology*, 4, 677. <https://doi.org/10.3389/fpsyg.2013.00677>
- Segaert, K., Markiewicz, R., & Mazaheri, A. (2019). *Effects of semantic binding and plausibility on ERPs and oscillatory power in the theta, alpha and beta band*. 11th meeting of the Society for the Neurobiology of Language (SNL). Helsinki, poster presentation. August 20–22, 2019
- Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science*, 12(4), 485–527. [https://doi.org/10.1207/s15516709cog1204\\_1](https://doi.org/10.1207/s15516709cog1204_1)
- Solt, S. (2018). Adjective meaning and scales. In C Cummins & N Katsos (Eds.), *The Oxford Handbook of experimental semantics and pragmatics* (pp. 263–282). Oxford University Press.
- Spychalska, M., Kontinen, J., & Werning, M. (2016). Investigating scalar implicatures in a truth-value judgement task: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, 31(6), 817–840. <https://doi.org/10.1080/23273798.2016.1161806>
- Stenning, K., & Van Lambalgen, M. (2008). *Human reasoning and cognitive science*. MIT Press.
- Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, 63(2), 158–179. <https://doi.org/10.1016/j.jml.2010.03.008>
- Van Berkum, J. J., Brown, C. M., & Hagoort, P. (1999). Early referential context effects in sentence processing: Evidence from event-related brain potentials. *Journal of Memory and Language*, 41(2), 147–182. <https://doi.org/10.1006/jmla.1999.2641>
- Van Berkum, J. J., Brown, C. M., Hagoort, P., & Zwitserlood, P. (2003). Event-related brain potentials reflect discourse-referential ambiguity in spoken language comprehension. *Psychophysiology*, 40(2), 235–248. <https://doi.org/10.1111/1469-8986.00025>
- Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition*, 18(4), 380–393. <https://doi.org/10.3758/BF03197127>
- Westerlund, M., Kastner, I., Al Kaabi, M., & Pylkkänen, L. (2015). The LATL as locus of composition: MEG evidence from English and arabic. *Brain and Language*, 141, 124–134. <https://doi.org/10.1016/j.bandl.2014.12.003>



- Westerlund, M., & Pykkänen, L. (2017). How does the left anterior temporal lobe contribute to conceptual combination? Interdisciplinary perspectives. In J. A. Hampton, & Y. Winter (Eds.), *Compositionality and concepts in Linguistics and psychology* (pp. 269–290). Springer International Publishing.
- Zhang, L., & Pykkänen, L. (2015). The interplay of composition and concept specificity in the left anterior temporal lobe: An MEG study. *NeuroImage*, 111, 228–240. <https://doi.org/10.1016/j.neuroimage.2015.02.028>
- Ziegler, J. C., Besson, M., Jacobs, A. M., Nazir, T. A., & Carr, T. H. (1997). Word, pseudoword, and nonword processing: A multitask comparison using event-related brain potentials. *Journal of Cognitive Neuroscience*, 9(6), 758–775. <https://doi.org/10.1162/jocn.1997.9.6.758>
- Ziegler, J., & Pykkänen, L. (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>