Brain and Language 209 (2020) 104837

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l

EEG signatures of elementary composition: Disentangling genuine composition and expectancy processes

Emilia Fló[a,b], Álvaro Cabana[a,c], Juan C. Valle-Lisboa[a,c,*]

[a] Center for Basic Research in Psychology, Facultad de Psicología, Universidad de la República, Tristán Narvaja 1674, Montevideo 11200, Uruguay
[b] Laboratorio de Neurociencias, Facultad de Ciencias, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay
[c] Instituto de Fundamentos y Métodos, Facultad de Psicología, Universidad de la República, Tristán Narvaja 1674, Montevideo 11200, Uruguay

ARTICLE INFO

Keywords:
Linguistic composition
EEG
Expectancy effects
Trial effects
Cluster based permutation

ABSTRACT

We adapted Bemis & Pylkkänen's (2011) paradigm to study elementary composition in Spanish using electroencephalography, to determine if EEG is sensitive enough to detect a composition-related activity and analyze whether the expectancy of participants to compose contributes to this signal. We found relevant activity at the expected channels and times, and a putative composition-related activity before the second word onset. Using threshold-free cluster permutation analysis and linear models we show a task-progression effect for the composition task that is not present for the list task. In a second experiment we evaluate two-word composition incorporating all conditions in a single task. In this case, we failed to find any significant composition-related activity suggesting that the activity measured with EEG may be in part carried by expectancy processes arising from the block design of the experiment, which can be prevented by using a non-blocked design and data-driven techniques to analyze the data.

1. Introduction

The ability to combine words in order to represent and convey new meanings is a fundamental operation in the comprehension and production of language (Martin & Baggio, 2020). One of the challenges in language research is to account for the functional neuroanatomical basis that underlie these processes (Friederici, 2017; Hagoort, Baggio, & Wlllems, 2009). More specifically, the challenge is to understand how the meaning of individual words is combined into complex meaning representations. Linguists recognize that different levels of knowledge are involved in the production or understanding of an utterance. In particular, syntax, semantics and world knowledge interact strongly when building the meaning of an expression (Hagoort, 2019). Research on the brain basis of language has attempted to disentangle these processes by designing tasks in which only one of these levels is varied (Friederici, Meyer, & Cramon, 2000; Humphries, Love, Swinney, & Hickok, 2005; Kutas & Hillyard, 1980; Mazoyer et al., 1993; Snijders et al., 2009). Nevertheless, on many occasions these approaches involve elaborate stimuli which elicit neural responses related to general performance. Additionally, it is questionable whether they manage to successfully isolate a specific linguistic processes. Furthermore, these paradigms do not usually shed light on the computations that occur at every step as words are combined to form a unified concept. According to Pylkkänen and collaborators, in order to determine the computational contributions of each type of processing, it is crucial to empirically characterize composition at its most basic level as a starting point from where to refine the study of syntactic and semantic computations (Poeppel, Emmorey, Hickok, & Pylkkänen, 2012; Pylkkänen, Brennan, & Bemis, 2011).

1.1. Elementary composition

Inspired by the classical sentences versus list of words tasks, Pylkkänen's research group (Bemis & Pylkkänen, 2011, 2013c, 2013a, 2013b; Pylkkänen, Bemis, & Blanco, 2014; Westerlund & Pylkkänen, 2014) designed a paradigm based on two-word modifier-noun phrases. The initial study of this series (Bemis & Pylkkänen, 2011) introduced a simple paradigm to evaluate composition at its minimum by restricting stimuli to pairs of adjectives and nouns, and comparing subject's brain response to a non-word - noun control condition. Furthermore, a list task was implemented in which noun-noun conditions and non-word - noun conditions were presented to subjects. The rationale behind the experimental design and data analysis was to find brain regions for which there was a difference in activity between the two-word

* Corresponding author at: Center for Basic Research in Psychology, Facultad de Psicología, Universidad de la República, Tristán Narvaja 1674, Montevideo 11200, Uruguay.
E-mail addresses: eflo@fcien.edu.uy (E. Fló), acabana@psico.edu.uy (Á. Cabana), juancvl@psico.edu.uy (J.C. Valle-Lisboa).

https://doi.org/10.1016/j.bandl.2020.104837
Received 19 December 2019; Received in revised form 27 June 2020; Accepted 3 July 2020

condition and the one-word condition in the composition task, and no (or a smaller) difference between conditions in the list task. By combining task and number of words in their design, they argue that any difference between conditions in the composition task that is not present in the list task cannot be due to an effect of word number. Applying an hypothesis-driven cluster permutation analysis on MEG recordings, they report that the left anterior temporal lobe (LATL) and the ventromedial prefrontal cortex (vmPFC) show an increased response only for nouns preceded by adjectives. Interestingly, these responses develop early: activity in LATL occurs 184–255 ms after noun-onset, and the vmPFC response was reported 331–480 ms post-stimulus. The authors interpret LATL and vmPFC activities as a reflection of the combinatorial processes elicited by the binding of adjective-noun stimuli. This paradigm has been adapted to study semantic and syntactic operations in MEG and fMRI studies (see Westerlund & Pylkkänen, 2014; Zaccarella & Friederici, 2015; Zaccarella, Meyer, Makuuchi, & Friederici, 2017; Ziegler & Pylkkänen, 2016; Zhang & Pylkkänen, 2015), and although other regions have shown activation for nouns in elementary combinatorial contexts (Bemis & Pylkkänen, 2011, 2013a; Pylkkänen et al., 2014), the most consistent brain area eliciting activation across studies is the LATL. This region's involvement in composition is supported by a series of fMRI studies on conceptual combination (Baron & Osherson, 2011;Baron, Thompson-Schill, Weber, & Osherson, 2010; Coutanche & Thompson-Schill, 2015). It has been postulated that the LATL is involved not in syntactic or semantic computations per se, but in the addition of conceptual features in order to construct complex conceptual representations (Patterson, Nestor, & Rogers, 2007; Poortman & Pylkkänen, 2016; Ralph, Jefferies, Patterson, & Rogers, 2016). In this line, the early LATL response obtained under the two-word paradigm in MEG experiments has shown to be sensitive to the characteristics of the concepts combined such that the specificity of both head and modifier modulate this signal (Pylkkänen, 2019, 2020; Westerlund, Kastner, Al Kaabi, & Pylkkänen, 2015; Zhang & Pylkkänen, 2015). In addition, this activity is elicited by the composition of complex numbers but not for two-word numerical phrases, (Blanco-Elorrieta & Pylkkänen, 2016; Prato & Pylkkänen, 2014), and by the addition of semantic features to an individual representation and not to multiple entities (Poortman & Pylkkänen, 2016). Therefore this basic composition-related activity seems to reflect LATL's involvement in conceptual composition.

1.2. Elementary composition measured by EEG

When neural activity occurs in a synchronous manner across a great number of neurons, the coherent field potentials and local magnetic fields produced become big enough to be picked up by EEG and MEG, respectively. Even though there is great overlap between the information provided by these techniques they have some important differences. Firstly, because electrical conductivity varies across the layers of tissue that separate the cortical sources and the scalp, electrical signals are reduced and distorted. This has lower impact on MEG signal as magnetic permeability is more consistent (Okada, Lähteenmäki, & Xu, 1999). Furthermore, whereas EEG can detect both tangential and radial dipoles, MEG is more sensitive to source orientation. MEG is not able to capture neuronal currents oriented radially as they do not generate a magnetic field outside the head (Malmivuo & Plonsey, 1995; Williamson & Kaufman, 1981), however sources meeting this criteria have been shown to correspond with relatively small regions of the cortex located at the crests of gyri (Hillebrand & Barnes, 2002). EEG source reconstruction requires numerous electrodes, knowing their positions with precision, having head shapes measurements and accurate estimates of tissue conductivities (Michel et al., 2004). As magnetic fields are not distorted by the different brain structures, it is typically considered that MEG is better suited to resolve source localization. Importantly, both techniques have millisecond resolution and therefore are useful when the temporal course of a neural process is of interest (Luck, 2005; Okada et al., 1999).

In spite of MEG and EEG similarities, few studies have attempted to obtain an EEG marker of elementary composition following the two-word, composition-list paradigm established in the MEG literature. To our knowledge, an adaptation to study syntactic composition was implemented in Segaert, Mazaheri, and Hagoort (2018) by measuring oscillatory changes in brain activity elicited by the syntactic binding of pairs of pseudowords, and only one study carried out a replication of Bemis & Pylkkänen initial study. Neufeld et al. (2016) performed a classical event related potential (ERP) analysis on EEG data obtained under this task. These authors selected a time window where the effect is expected according to the original study, and averaged the results over anterior and posterior electrodes. Their electrode grouping was motivated by trying to relate the combinatorial activity with classical N400 context effects. Although they found a broadly distributed negative composition-related activity in EEG recordings, they did not find a significant interaction between task (composition, list) and number of words (two-word, one-word), rendering their results inconclusive. Interestingly, their preprocessing approach enabled the authors to test for differences in the time window preceding the critical noun. A classical analysis on a visually selected time interval showed a difference between the composition task conditions before composition could have been achieved. They interpret this precombinatorial activity as reflecting the building of a syntactic structure to allocate the incoming noun.

The presence of a process taking place before the onset of the second word only in combinatorial contexts, points to an alternative interpretation of the experimental results. Given that the experiments in Neufeld et al. (2016) and Bemis and Pylkkänen (2011) were both based on a block design, it is possible that the structural properties of the stimuli in the different blocks influence to some extent their result. In a two-word trial during the composition task, the first word indicates with certainty to subjects that they would have to combine both items. Therefore, the processing of the first word was contingent on second word processing, and this was not the case during the two-word condition in the list task. Accordingly, it is possible that the activity identified as composition-related is conflated with anticipatory processes (c.f. Molinaro, Carreiras, & Duñabeitia (2012)).

1.3. Expectancy and task progression effects

A well-studied electrophysiological signature of expectancy-based processes is the Contingent Negative Variation (CNV), a slow negative event-related potential (ERP) associated to expectancy and prediction, which develops in a gradual manner during the realization of a task (Walter, Cooper, Aldrige, McCallum, & Winter, 1964). This complex expectancy wave is modulated by task demands (Jacobson & Gans, 1981; Rebert, McAdam, & Knott, 1967; Tecce, Savignano-Bowman, & Meinbresse, 1976), sensory processing (Gaillard, 1976; Loveless, 1975) and motor preparation (Brunia & Vingerhoets, 1981; Irwin, Knott, McAdam, & Rebert, 1966; Rohrbaugh & Gaillard, 1983). CNV is reported in the literature as a negative potential that increases as the contingency between two stimuli is learned (Cohen, 1969; Hillyard, 1969; Poon, Thompson, Williams, & Marsh, 1974; Proulx & Picton, 1980; Walter et al., 1964), with a post-learning behavior that is task-dependent (Donald, 1980). Besides CNV, task progression effects have also been reported for classical ERPs such as the P300 elicited during oddball paradigms, which shows a behavior consistent with habituation effects (Barry et al., 2019; Polich, 1989; Ravden & Polich, 1998). Likewise, N400 amplitude decreases during word repetition tasks (Bentin & McCarthy, 1994; Ströberg, Andersen, & Wiens, 2017).

This phenomenon challenges the ordinary analysis followed for ERPs, as averaging across trials neglects the possibility that the cognitive process of interest may vary as task unfolds. Furthermore, the prevalent averaging procedure reduces data to one data point per subject per condition, restraining the analysis to rigid ANOVAs. In contrast, modelling between-trial variation and using mixed regression

or non-parametric analyses allows the possibility to evaluate task progression effects (Brush, Ehmann, Hajcak, Selby, & Alderman, 2018; Vossen, Van Breukelen, Hermens, Van Os, & Lousberg, 2011; Volpert-Esmond, Merkle, Levsen, Ito, & Bartholow, 2018).

1.4. The present study

The purpose of the present study is to investigate whether EEG is sensitive to detect a composition-related activity elicited by a Spanish adaptation of Bemis & Pylkkänen's paradigm, and to separate genuine composition from other non-specific expectancy-related responses. To this end, we carried out an adaptation of Bemis and Pylkkänen original experiment (Bemis & Pylkkänen, 2011) to Spanish while recording EEG activity (Experiment 1). Importantly, we used three methods to analyze the EEG data. In the first place, to be able to compare our results with the existent EEG replication (Neufeld et al., 2016), we followed their traditional ERP analysis. Following the original Bemis & Pylkkänen study, we performed an adapted version of their cluster permutation method on both ERP and time–frequency power representations. Finally, as expectancy processes change across trials and in order to avoid bias introduced by parameter selection, we implemented a threshold-free cluster permutation analysis that included trial number as a predictor variable to test for task progression effects on EEG activity. If the reported activity in Neufeld et al. (2016) is in fact related to composition, we would expect it to be also present during a task in which participants cannot anticipate if a given word will have to be used to perform composition. Conversely, if this activity is not elicited it would suggest the alternative hypothesis, that this brain activity is at least in part due to anticipatory processes generated by task demands. Therefore, to examine this possibility we introduce a novel task to evaluate two-word composition that avoids differential expectancy effects across conditions (Experiment 2).

2. Materials and methods

2.1. Experiment 1

2.1.1. Participants

Twenty-nine non-colorblind Uruguayan undergraduate students participated in this experiment (22 female, average age, 25.31 ± 0.56). All subjects were native Rioplatense Spanish speakers, right handed, and had normal or corrected vision. The experiments were approved by the ethics committee of the Facultad de Psicología Universidad de la República. All participants gave informed consent and were not awarded any economic or academic retribution, according to the national established guidelines (Decree Nž379/008).

2.1.2. Experimental design

The original paradigm to study elementary composition included two tasks: a composition and a list task. The composition task consisted of a one-word condition in which subjects were presented with a non-word followed by a word denoting a noun (*xkq boat*), and a two-word condition in which subjects saw two words, an adjective followed by a noun (*red boat*). After each condition subjects had to match the presented verbal stimulus to an image. In the list task, subjects were presented with a one-word condition exactly the same as in the composition task, and a two-word condition in which two consecutive nouns were displayed (*cup boat*). Subjects had to indicate whether a subsequent picture matched any preceding word. In contrast to English, in Spanish adjectives such as color, shape or evaluation tend to be used in a post-nominal way, and their pre-nominal uses are mainly restricted to literary resources as is the use of epithets (Bosque & Demonte, 1999). Taking this into account we adapted the composition task so that subjects were presented with *noun adjective* (NA) trials in the two-word condition and *non-word adjective* (XA) trials in the one-word condition. We also devised two versions of the list task. In the list of adjectives task, subjects saw either *adjective adjective* (AA) trials (two-word condition) or *non-word adjective* (XA) trials (one-word condition). In this task the critical second word was also an adjective (*noun adjective VS. adjective adjective*), matching the composition task. However, using a list of adjectives could be problematic as color adjectives are typically more abstract than regular nouns. Hence, subjects were presented with a list of nouns task in which they saw *noun noun* (NN) trials and *non-word noun* (XN) trials in the two-word and one-word conditions, respectively. During the composition task participants were instructed to answer whether the image matched the preceding words. In contrast, in the list tasks subjects had to answer whether the image matched any of the preceding words. Half of the participants started the experiment with the composition task, and half started with the list tasks. The list task order and yes/no response hand was counterbalanced across subjects. Subjects were encouraged to answer as quickly and accurately as possible. Each task consisted of 200 trials that were preceded by a 40 trial practice to ensure that participants understood the task and learned the response keys. During each trial participants saw a fixation cross for 1 s and all stimuli except for the images were presented for 300 ms and were followed by a 300 ms blank screen. The images presented at the end of each trial remained on screen until subjects pressed a key or after 3 s (see Fig. 1A). Following the end of each trial a sound was elicited to encourage participants to blink at that time. The inter-trial interval was randomly varied between 0.8–1.5 s. Stimuli presentation was coded in Psychopy (Peirce, 2007) and displayed on a CRT monitor with a 60 Hz refresh rate.

2.1.3. Stimuli construction

In order to generate the stimuli, 11 nouns - *bote, cable, cepillo, globo, gorro, lápiz, reloj, teléfono, tenedor, tren, zapato* (boat, cable, thoothbrush, balloon, hat, pencil, clock, telephone, fork, train, shoe) and 11 color adjectives - *amarillo, azul, blanco, celeste, marrón, naranja, negro, rojo, rosado, verde, violeta* - (yellow, blue, white, sky-blue, brown, orange, black, red, pink, green, purple) were selected. Nouns and adjectives were matched for frequency ($p = 0.059$), number of letters ($p = 0.51$), number of substitution neighbors ($p = 0.92$), number of phonemes ($p = 0.50$), number of syllables ($p = 0.58$), number of homophones ($p = 0.22$), and number of phonological neighbors ($p = 0.95$). It was not possible to compare words on familiarity, imageability and concreteness as there was no available information for 7 out of the 11 color adjectives. Non-words (*brnlqs, slgrl, grsd, vrpng, jlcrfsmt, cxgnff, drbcw, tphn, dpjzb, pkrdt, vqdfnsm*) were constructed using Wuggy software (Keuleers & Brysbaert, 2010), and did not differ in number of letters from the adjectives ($p = 0.40$) nor from the nouns ($p = 0.87$). p-values correspond to a two tails t-test and properties were taken from Duchon, Perea, Sebastián-Gallés, Martí, and Carreiras (2013). For the composition task a python script was created to automatically select 10 nouns and 10 adjectives, and to combine them in order to generate 100 trials. For the one-word condition the 11 non-words and the 11 adjectives were combined, 21 items were randomly selected and discarded. This was done independently for every subject. Half of the trials for each condition were incongruent. In the two-word condition a trial was incongruent if the noun (25 trials) or the adjective (25 trials) did not match the image. In the one-word condition a trial was incongruent if the adjective did not match the image (50 trials). A second python script was created to generate the stimuli for the list tasks. For each subject for each list task, permutations of the 11 words were obtained and 10 items were randomly selected and discarded. For the one-word condition the 11 non-words and the 11 words (nouns or adjectives) were combined, 21 items were randomly selected and discarded. For the one-word condition of both tasks, a trial was incongruent if the image did not match the word (50 trials). For the two-word condition a trial was incongruent if the image did not match any of the preceding words (50 trials).

We evaluated the corpus bigram frequency of all the noun-adjective pairs using the Spanish corpus of Google Ngram (Michel et al., 2011)

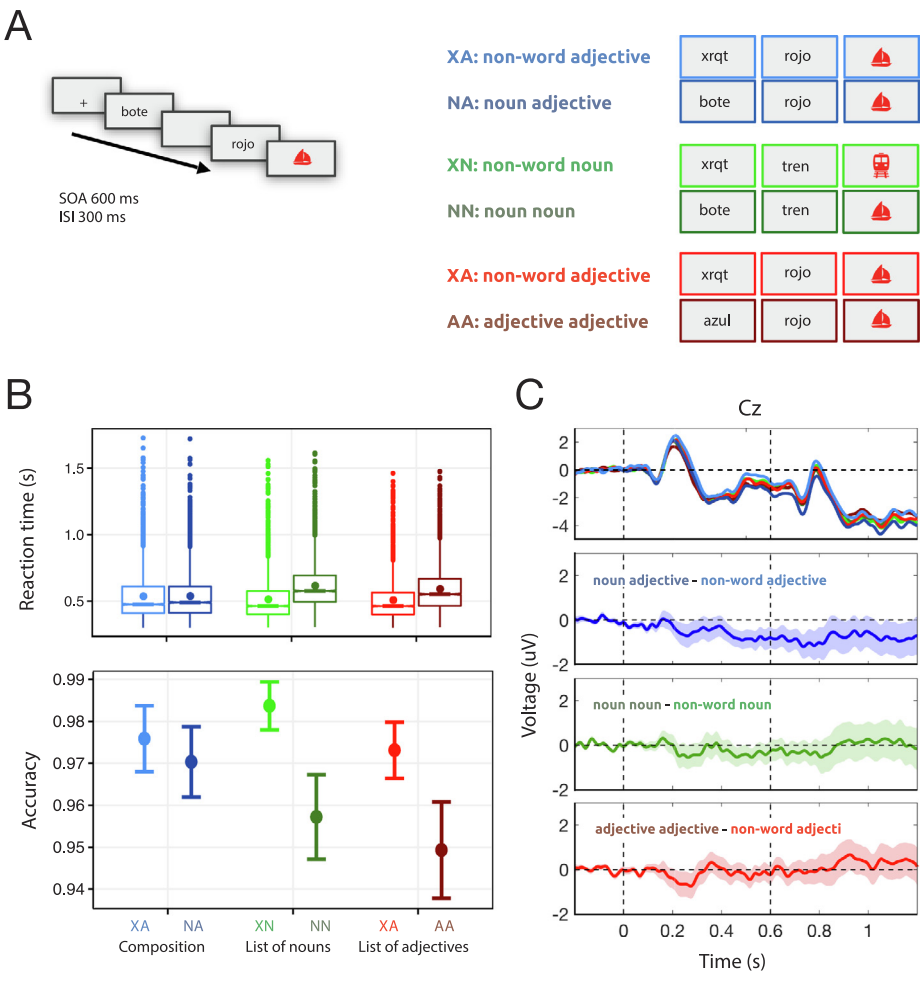


Fig. 1. Experiment 1. Spanish adaptation of Bemis & Pylkkänen's (Bemis & Pylkkänen, 2011) task. **A.** Subjects were presented with three tasks: a composition task and two list tasks with two-word and one-word conditions. **B.** Reaction times and accuracy results. Bars indicate 95% confidence interval. **C.** Top graph shows ERP grand average for each condition of each task superimposed for Cz. Bottom graphs show mean differences between conditions for each task with 95% confidence interval for Cz. XA: non-word adjective, NA: noun adjective, XN: non-word noun, NN: noun noun, and AA: adjective adjective. Vertical dashed lines indicate first and second. stimulus onset.

using the Phrasefinder API (https://phrasefinder.io/api). The Spanish corpus registers a total of 40.053.844 bigrams. Of the 121 possible noun-adjective pairs, 60 were not found in the corpus. The highest frequency was for the pair *lápiz rojo* (red pencil) with a frequency of 10579 instances (0.026%) and the lowest non-zero frequency was for the pair *gorro rosado* (pink cap) with a frequency of 49 (1.2×10^{-4} %). The overall mean relative bigram frequency was 8.9×10^{-4} %, 95% CI [3.9×10^{-4} %, 1.4×10^{-3} %]. None of the possible noun-noun pairs appear in the Google Ngrams Spanish corpus.

2.1.4. Behavioral data analysis

Response times were measured from image onset until subjects pressed the no or yes keys. For each subject and task, reaction times were analyzed by removing incorrect and missing responses as well as trials in which response times were over or under two standard deviations. A linear mixed model with task as a three level factor (*composition - list of nouns - list of adjectives*) and number of words as a two-level factor (*two-word - one-word*) as fixed effects, and subject, first and second word as random intercepts was fitted to log-transformed reaction times, using the *lme4* R package (Bates, Mächler, Bolker, & Walker, 2015; R Core Team, 2017). Accuracy data was analyzed using generalized linear mixed models using a logit link function with task as a three level factor (*composition - list of nouns - list of adjectives*), number of words as a two-level factor (*two-word - one-word*) and image congruency as a two-level factor (*congruent - incongruent*) as fixed effects, and random intercepts for each subject, first and second word. We used Wald χ^2 tests to evaluate factor main effects and their interactions, while Wald Z tests were employed in pairwise comparisons. Bonferroni adjustments were used to correct for multiple comparisons.

2.1.5. EEG recording and preprocessing

EEG signal was recorded using a Biosemi Active-Two system (Biosemi, B.V., Amsterdam, Netherlands). Sixty-four Ag-AgCl scalp electrodes where placed on a head cap following the location and label of the 10–20 system (Jasper, 1958). Ocular movements were monitored by 4 electrooculographic (EOG) electrodes (above, below the left eye, and on the outer canthi). The activity recorded was referenced online to the common mode sense (CMS; active electrode) and grounded to a passive electrode (Driven Right Leg, DRL), creating a feedback loop that drives the average potential of the participant to the AD-box reference potential. Data was digitized with a sample rate of 512 Hz with a fifth-order low-pass sinc filter with a −3 dB cutoff at 410 Hz. Data was preprocessed in MATLAB using fieldtrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011). Continuous data was two-pass filtered with a second-order high-pass Butterworth filter at 0.1 Hz and a fourth-order low-pass Butterworth filter at 30 Hz. Data was epoched 0.2 s prior to first word onset until 1.2 s (at image presentation). Epochs were baselined to activity 200 ms preceding first word onset. Noisy trials and channels were rejected following Junghöfer, Elbert, Tucker, and Rockstroh (2000). Trials in which participants responded incorrectly were discarded. The remaining trials for each condition were averaged in order to obtain one ERP per subject per condition.

For the composition task the number of rejected electrodes and trials was 1.52 ± 1.16 and 23.19 ± 6.97, respectively. A t-test showed no difference in the number of trials rejected between conditions ($p = 0.16$). For the list of adjectives task the average of discarded electrodes was 2.17 ± 1.97 and the number of trials rejected was 21, 76 ± 6.80, there was no difference in the number of trials rejected between conditions ($p = 0.12$). For the list of nouns task the average of

electrodes rejected corresponded to 1.55 ± 1.22 and the average for trials was 23.30 ± 7.22. No difference between conditions for the number of rejected trials was found ($p = 0.96$). A one-way ANOVA was conducted to compare the effect of task on number of channels rejected, no difference was found ($F(2, 78) = 1.86$, $p = 0.16$). Accordingly, a second one-way ANOVA showed no difference between task and number of discarded trials ($F(2, 78) = 0.008$, $p = 0.99$).

Subjects' ERP signal-to-noise ratio was evaluated following Parks, Gannon, Long, and Young (2016). A lower bound of signal-to-noise ratio confidence interval (SNRLB) of 3.0 dB was used as threshold to ensure signal quality. For the composition task the SNRLB ranged from 3.54 to 21.87 dB (mean = 11.25, median = 10.26 dB, SD = 5.23 dB). The SNRLB for the list of adjectives task ranged from 0.15 to 16.10 dB (mean = 8.31 dB, median = 8.34 dB, SD = 3.41 dB). Subjects 13 and 19 failed to meet the SNRLB criterion. Finally for the list of nouns task the SNRLB was between 4.37 and 22.14 dB (mean = 11.76 dB, median = 10.46 dB, SD = 3.66 dB). In order to compare the data between tasks we excluded from the analysis the two participants that did not met the critical value in the list of adjectives task.

2.1.6. Cluster permutation analysis on ERPs

We used a cluster permutation analysis to evaluate all electrodes and epoch data points from −0.2 s to 1.2 s (until image presentation). We used two-tailed t-tests to contrast the difference between the composition conditions (NA - XA) with the difference between the list of nouns conditions (NN - XN). Independently, we used the same approach to contrast differences between the composition (NA - XA) conditions with the difference between the list of adjectives conditions (AA - XA). This analysis aims at finding differences between the composition task conditions that are not present between the list task conditions. Additionally, as we decided to use two controls, conditions in the list of nouns and in the list of adjectives were also contrasted to test for possible differences.

In each analysis, a t-test was performed on every sample and t values were clustered depending on if they exceeded a dependent samples t-test threshold of $p < 0.05$ (two-tailed). t values for each data point within each cluster were summed in order to obtain a value per cluster. The maximum negative and positive cluster statistic values were kept. This was done for 5000 permutations of the data resulting in a null hypothesis distribution of the statistic, against which we tested the real data. We considered the critical α level here to be 0.025. For all the analyses using this method on ERPs we set to three the minimum number of electrode neighbors that had to be significant for a given time point in order to be part of a cluster. All cluster permutation analyses were conducted on MATLAB using Fieldtrip toolbox (Oostenveld et al., 2011) and neighbors were defined following Biosemi 64 neighbors template.

2.1.7. Threshold free cluster permutation analysis

In order to explore task progression effects on ERPs we performed threshold-free cluster enhancement (TFCE) (Mensen & Khatami, 2013; Smith & Nichols, 2009) analyses on t-values derived from linear models fitted to epoched EEG data. TFCE is a non-parametric cluster randomization method inspired by random field theory that uses an enhanced statistic designed to boost weaker but broadly distributed signals. In contrast to the cluster permutation method implemented in Fieldtrip (Oostenveld et al., 2011), this method does not have threshold as a free parameter, reducing the researcher degrees of freedom (Wicherts et al., 2016). We fitted two linear models to each time point and electrode using data from both tasks combined (as in the previous analyses). A simple model contained parameters for trial number, task (composition vs. list), number of words (two-word vs. one-word), their interaction (task × number of words), and the interaction between task and trial number. This last parameter was introduced to test for different effects of task progression across tasks. A second model (the "complete" model) also included the interaction between number of words and trial number, and the triple interaction (task × number of words × trial number). Also, to further dissect the relation between task progression and experimental condition, we fitted follow-up models with trial number, number of words and their interaction as parameters for each task separately. For each time and electrode, t-values were obtained for each parameter estimate. Following Mensen and Khatami (2013), positive and negative t-values were separated and t_{tfce} statistics were computed. The maximum (in absolute value) t_{tfce} statistic was used in cluster randomization tests with 5000 permutations to obtain p-values for each parameter of each model.

2.1.8. Voltage correlation to reaction times

In order to quantify the association between voltage and participants' reaction times we used the rmcorr R package to obtain repeated-measures correlation coefficients (Bakdash & Marusich, 2018). This analysis takes into account the fact that observations within participants are not independent and has greater statistical power as no averaging across individuals is performed, similar to linear mixed model approaches. We selected the electrode-time point pair for which the t_{tfce} statistic for the trial parameter was maximal and averaged voltage across a 100 ms time window centered on that point. Using this data, we obtained correlation coefficients between trial-to-trial voltage values and reaction times for the composition and lists tasks.

2.2. Experiment 2

2.2.1. Participants

Thirty-nine non-colorblind Uruguayan undergraduate students participated in this experiment (23 female, average age, 23.13 ± 3.65). All subjects were native Rioplatense Spanish speakers, right handed, and had normal or corrected vision.

2.2.2. Experimental design

In this experiment we tried to maintain the design similar to Experiment 1 while trying to reduce possible confounding of expectancy in composition effects. During this task participants were presented with four different conditions: *noun adjective: NA*, *non-word adjective: XA*, *noun noun: NN* and *non-word noun: XN*. In all conditions, each word pair was followed by an image. Subjects were instructed to answer if the image matched the preceding verbal material and were encouraged to answer as quickly and accurately as possible. Hence, there were two-word conditions (NA and NN) and one-word conditions (XA and XN), and composition was required only for the NA condition. A representation of the trials is shown in Fig. 5B. The main differences with the previous experiment were that subjects saw all types of trials in the same block, and the image presented in NN trials had two elements. The task consisted of 400 trials that were preceded by a 40 trial practice. In each trial, participants saw a fixation cross for 1 s and all stimuli except for the images were presented for 300 ms followed by a 300 ms blank screen. Every subject was presented with 100 trials of each condition, and half of the trials of every condition were incongruent to the image. For the XN and XA conditions, a trial was incongruent if the image did not match the noun or adjective respectively (50 trials for each condition). For the NA condition a trial was incongruent if the image did not match the noun (25 trials) or did not match the adjective (25 trials). Finally, for the NN condition a trial was incongruent if the image did not match the first noun (25 trials) or the second noun (25 trials).

In order to obtain results comparable to Experiment 1, we conceptually organized the aforementioned four conditions as two tasks with two conditions each. The composition task comprised of a *two-word* (NA) and a *one-word* (XA), condition, and a *list of nouns* task also comprised of a *two-word* (NN) and a *one-word* (XN) condition.

2.2.3. Stimuli construction

The same pool of words and non-words used in the Experiment 1

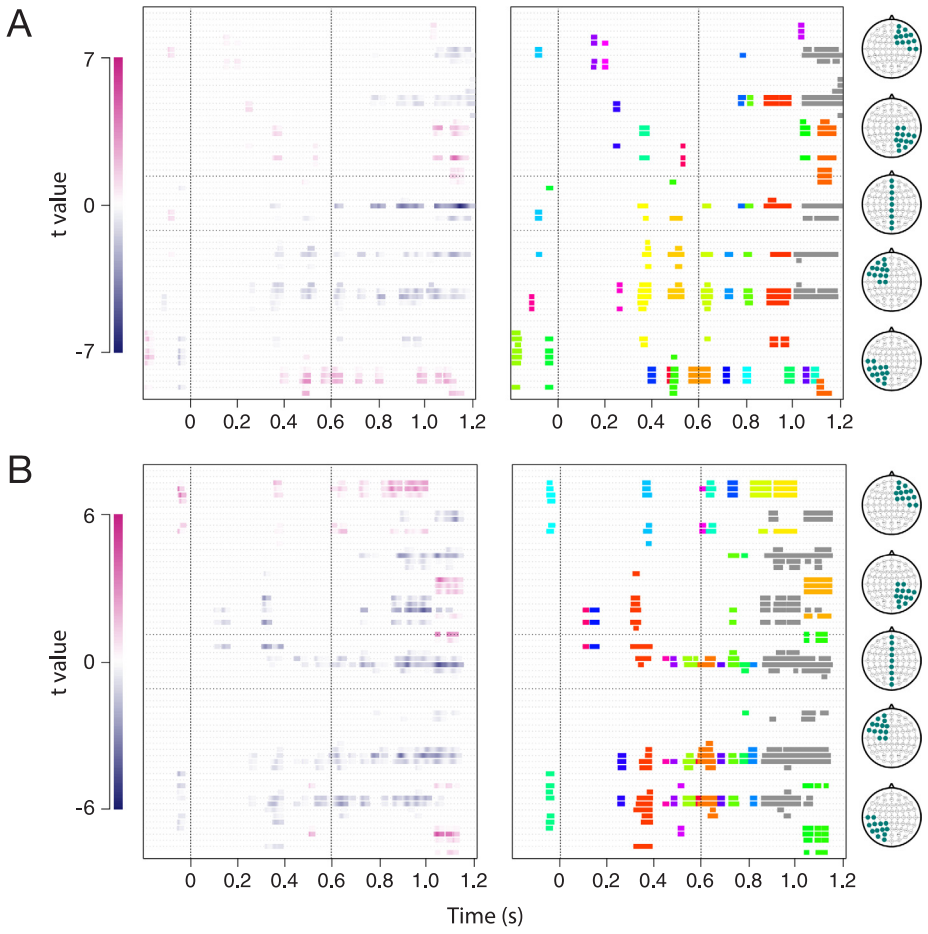


Fig. 2. Cluster permutation analysis results for the interaction between task and number of words. **A**. Composition vs. list of adjectives. Left: t-values for the clusters obtained. Right: Electrode–time points clusters. Color code represents data points that participate in a given cluster. Only the gray cluster is statistically significant ($p = 0.007$, t = 1.01–1.20 s). **B**. Composition vs. list of nouns. Left: t-values for the clusters obtained. Right: Electrode–time clusters. Only the gray cluster is statistically significant ($p = 0.006$, t = 0.86–1.15 s).

was employed to create the stimuli. A python script was coded to combine words and non-words for each subject in order to create 400 trials. For the XN, XA and the NA conditions, a combination of the 11 elements of each pool was made, resulting in 121 pairings with the desired structure. Subsequently, 21 items were randomly discarded. For the NN condition, permutations of the nouns were generated and 10 items were discarded, resulting in 100 pairings. Both the NN and the NA pairs have the same bigram frequency properties as the ones used in Experiment 1.

2.2.4. Behavioral data analysis

Response times were measured from image onset until subjects pressed the no or yes keys. For each subject, reaction times were analyzed by removing incorrect and missing responses as well as trials in which response times were over or under two standard deviations. A linear mixed model with task as a two level factor (*composition - list of nouns*) and number of words as a two-level factor (*two-word - one-word*) as fixed effects, and subject, first and second word as random intercepts was fitted to log-transformed reaction times, using the `lme4` R package (Bates et al., 2015; R Core Team, 2017). Accuracy data was analyzed using generalized linear mixed models using a logit link function with task as a two-level factor (*composition - list of nouns*), number of words as a two-level factor (*two-words - one-word*) and image congruency as a two-level factor (*congruent - incongruent*) as fixed effects, and random intercepts for each subject and second word. We used Wald χ^2 tests to evaluate main effects and their interactions, while Wald Z tests were employed in pairwise comparisons. Bonferroni adjustments were used to correct for multiple comparisons.

2.2.5. EEG recording and preprocessing

We followed the same recording and preprocessing steps implemented in the first experiment. The number of rejected electrodes and trials was 4.44 ± 2.21 and 55.75 ± 15.28, respectively. A one-way repeated-measures ANOVA was conducted to test for differences in the number of trials across conditions, no difference was found $F(3, 124) = 0.74$, $p = 0.53$. Subjects' ERP SNRLB ranged from 1.74 to 20.31 dB (mean = 10.99, median = 11.52 dB, SD = 4.78 dB). Subjects 3, 12, 14, 15, 26 and 29 failed to meet the SNRLB criterion and were excluded from the analysis.

2.2.6. EEG data analysis

A cluster permutation analysis and a repeated-measures correlation analysis between trial-to-trial voltage values and reaction times were carried out following the procedures described for Experiment 1.

2.3. Further analyses

In order to test for specific hypotheses and compare to published results we performed a classical ERP analysis of Experiment 1. Furthermore both experiments were submitted to time–frequency analysis coupled to a cluster permutation test. The details are described in the Supplementary Materials section.

3. Results

3.1. Experiment 1: Adaptation to Spanish of the original study

3.1.1. Behavioral results

We found a significant main effect of task ($\chi^2(2) = 54.67$, $p < 0.001$) as well as an effect of number of words ($\chi^2(1) = 729.02$, $p < 0.001$). More interestingly, an interaction between task and number of words was also significant ($\chi^2(2) = 471.35$, $p < 0.001$). Pairwise comparisons showed no difference in reaction times between the two-word condition and the one-word condition for the composition task ($Z = -0.81$, $p = 0.42$). Contrarily, reaction times were smaller for the one-word condition for the list of nouns ($Z = -27.37$, $p < 0.001$) and for the list of adjectives ($Z = -22.89$, $p < 0.001$), compared to the two-word conditions.

Participant's accuracy was in line with the reaction time results. There was a main effect of task ($\chi^2(2) = 11.03$, $p = 0.004$) and number of words ($\chi^2(1) = 33.66$, $p < 0.001$), as well as a congruency effect ($\chi^2(1) = 7.63$, $p = 0.006$) which was not further explored. The analysis shows a significant interaction between task and number of words ($\chi^2(2) = 11.44$, $p = 0.003$). Pairwise comparisons revealed no difference between conditions in accuracy for the composition task ($Z = 1.23$, $p = 0.22$) and a significant difference between conditions in the list of nouns ($Z = 5.48$, $p < 0.001$), as well as for the list of adjectives task ($Z = 4.26$, $p < 0.001$) (Fig. 1B).

3.1.2. Cluster permutation analysis results

The permutation cluster analysis to test the interaction between the list of adjectives and composition tasks and the number of words yielded a significant negative cluster comprised of 16 electrodes with a central-frontal distribution ($p = 0.007$, t = 1.01–1.20s) (Fig. 2A; see Fig. 1 in Supplementary Material for th topographical distribution of the cluster). A post hoc analysis was carried out for each task taking only the electrodes and data points that participate in the interaction cluster. No cluster was obtained for the list of adjectives. We found a significant negative voltage cluster for the composition task ($p = 5.0 \times 10^{-4}$, t = 1.01–1.20 s).

The same analysis comparing the list of nouns and composition tasks showed similar results. A significant negative cluster was obtained for the interaction between task and number of words ($p = 0.006$, t = 0.86–1.15 s) composed of 26 electrodes with a central distribution (Fig. 2B; see Fig. 2 in Supplementary Material for the topographical distribution of the cluster). The post hoc analysis showed no difference between conditions for the list of nouns task, and two significant negative voltage clusters for the composition task ($p = 8.0 \times 10^{-4}$, t = 0.95–1.15 s and $p = 6.6 \times 10^{-3}$, t = 0.86–094 s).

We also carried out the same analysis contrasting both list tasks. No significant clusters were obtained.

3.1.3. TFCE cluster permutation analyses and linear models

For the sake of brevity we present results for the composition task and the list of nouns task, since results for both list tasks follow the same pattern.

Threshold-free cluster enhancements statistics were obtained for two linear models fitted to the EEG data, the simple and the complete model (See Table 1 in supplementary material).

For the simple model, we found significant effects of task (FCz, t = 0.423 s, $t_{tfce} = -2.68 \times 10^3$, $p = 0.027$), number of words (PO7, t = 0.503 s, $t_{tfce} = 8.55 \times 10^3$, $p < 0.001$), trial number (Oz, t = 0.415 s, $t_{tfce} = 4.45 \times 10^3$, $p < 0.001$) (Fig. 3) and a significant interaction between task and number of words (Cz, t = 1.05 s, $t_{tfce} = 2.75 \times 10^3$, $p = 0.027$) (Fig. 3A). Importantly, we found a significant effect for the interaction between task and trial number (F2, t = 0.460 s, $t_{tfce} = -2.46 \times 10^3$, $p = 0.043$).

The analysis of the complete model with all the interactions showed the following results. A significant task effect (F1, t = 0.417 s, $t_{tfce} = -1.82 \times 10^3$, $p = 0.032$), a significant effect of number of words (FC3, t = 0.398 s, $t_{tfce} = -2.84 \times 10^3$, $p = 0.024$), a significant trial number effect (F3, t = 0.404 s, $t_{tfce} = -2.82 \times 10^3$, $p = 0.026$) (Fig. 3B, Fig. 4A), no significant interaction between task and number of words (C6, t = 0.374 s, $t_{tfce} = -1.44 \times 10^3$, $p = 0.49$) (Fig. 3B), no significant interaction between task, number of words and trial number (Iz, t = 0.915, $t_{tfce} = -1.26 \times 10^3$, $p = 0.67$), no significant interaction between task and trial number (F1, t = 0.398s, $t_{tfce} = 1.41 \times 10^3$, $p = 0.52$) and no significant interaction between number of words and trial number (Iz, t = 0.899 s, $t_{tfce} = 1.72 \times 10^3$, $p = 0.30$).

For each task a model was fitted to compare the one-word and two-word conditions. For the composition task an effect of number of words (FC1, t = 0.439 s, $t_{tfce} = -2.52 \times 10^3$, $p = 0.020$), an effect of trial (F3, t = 0.402 s, $t_{tfce} = -2.68 \times 10^3$, $p = 0.013$), and no effect for an interaction between number of words and trial (Iz, t = 0.888 s, $t_{tfce} = 1.50 \times 10^3$, $p = 0.31$) was found.

For the list of nouns task an effect of number of words (POz, t = 0.499 s, $t_{tfce} = 2.23 \times 10^3$, $p = 0.020$), no effect of trial (C6, t = 0.723 s, $t_{tfce} = 1.58 \times 10^3$, $p = 0.20$), and no effect for an interaction between number of words and trial (F5, t = 0.061 s, $t_{tfce} = -1.26 \times 10^3$, $p = 0.49$) was obtained.

3.1.4. Correlation between voltage and reaction times

For the composition task, lower reaction times correlated with more negative potentials (F3, rrm (1436) = 0.055, 95% CI [0.004, 0.107], $p = 0.036$). This correlation was not found for the list of nouns (F3, rrm (1428) = −0.015, 95% CI [-0.067, 0.037], $p = 0.56$) (Fig. 4 B).

3.1.5. Classical ERP and time–frequency analysis

In the Supplementary Materials we present the results for the classical ERP and the time–frequency analyses. As in Neufeld et al. (2016) we found no interaction between task and number of words, although post hoc analyses yielded a difference between conditions for the composition task and no difference between conditions for the list tasks. No significant differences were found for the composition task in the time points evaluated before second word onset. Finally, we found no significant clusters for the interaction between task and number of words in any of the studied power bands (gamma, alpha and beta bands).

3.2. Experiment 2

3.2.1. Behavioral results

For reaction times, a main effect of task was found ($\chi^2(1) = 458.36$, $p < 0.001$) as well as an effect of number of words ($\chi^2(1) = 546.51$, $p < 0.001$). Moreover, the interaction between task and number of words was significant ($\chi^2(1) = 1284.94$, $p < 0.001$). Pairwise comparisons were tested between number of words for each group. The two-word and one-word conditions in the list of nouns task were different ($Z = -37.04$, $p < 0.001$). Furthermore, there was a significant difference between conditions for the composition group as well ($Z = -4.10$, $p < 0.001$) (Fig. 5B).

In regard to accuracy, there was a main effect of task ($\chi^2(1) = 11.78$, $p < 0.001$) and number of words ($\chi^2(1) = 47.32$, $p < 0.001$), as well as a congruency effect ($\chi^2(1) = 42.88$, $p < 0.001$). The analysis showed a significant interaction between task and number of words ($\chi^2(1) = 15.86$, $p < 0.001$). Pairwise comparisons revealed no difference between conditions for the composition task ($Z = 1.91$, $p = 0.057$) and a significant difference between conditions in the list of nouns ($Z = 8.36$, $p < 0.001$) (Fig. 5B).

3.2.2. Cluster permutation analysis results

To evaluate an interaction between number of words (two-word – one-word) and task (composition – list of nouns) we carried out a cluster permutation analysis on the EEG data averaged for each subject.

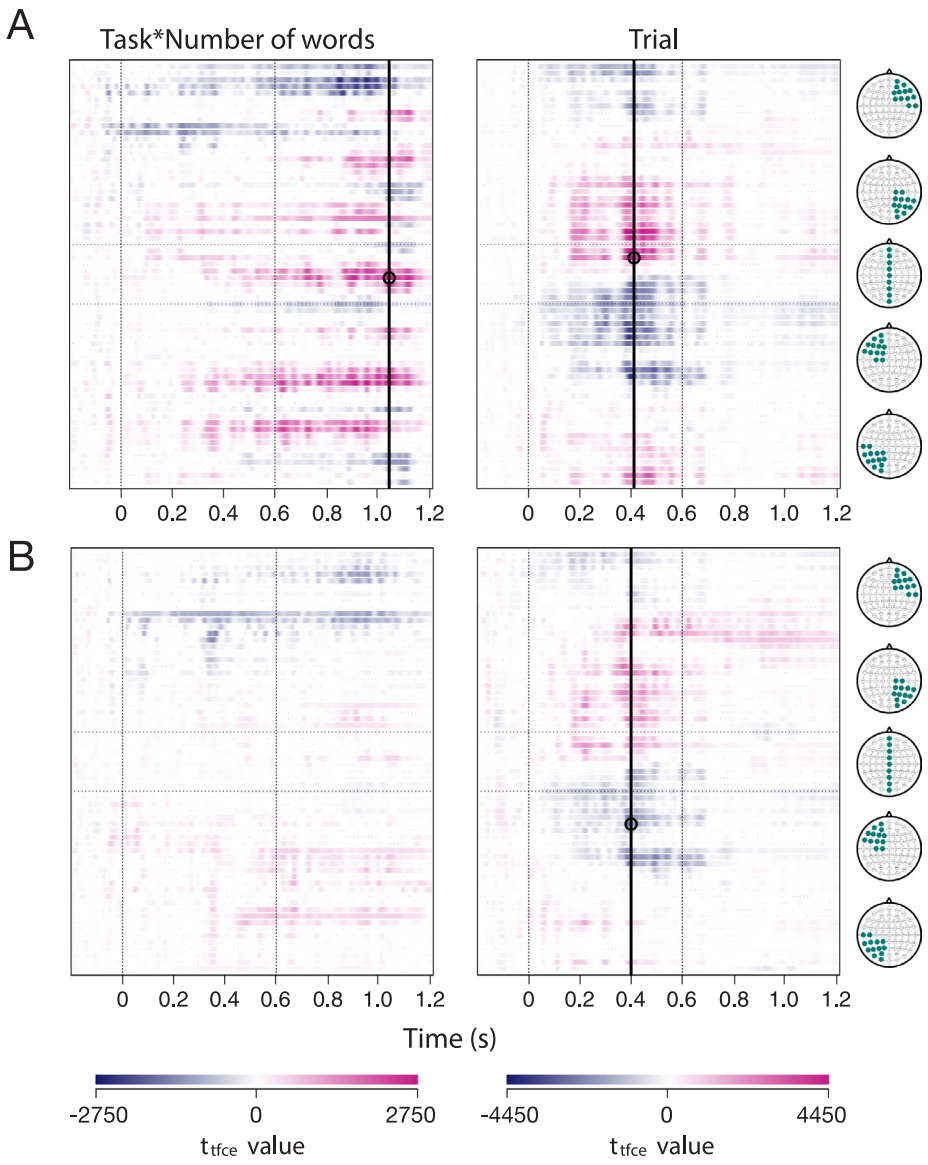
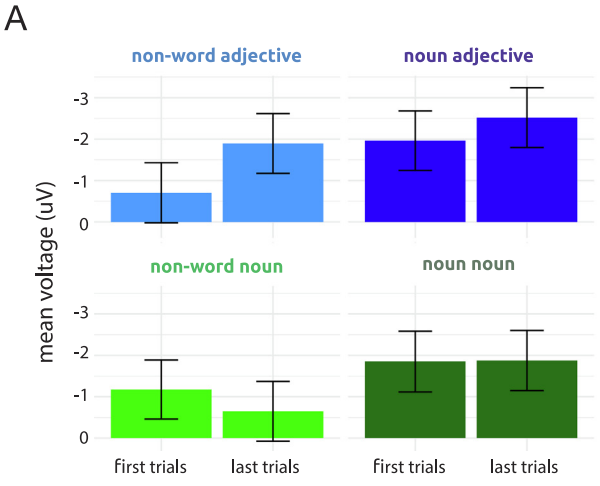


Fig. 3. Threshold-free cluster enhancement results on t-values derived from linear models fitted to EEG data from experiment 1. **A**. Simple model t_{tfce} statistics for the interaction between task and number of words (left) and for trial (right). **B**. Complete model t_{tfce} statistics for the interaction between task and number of words (left) and for trial (right). Line and circle indicate time-electrode maximum statistic values that reached significance in permutation tests.



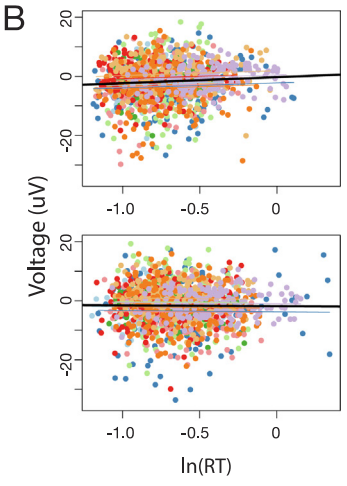


Fig. 4. Effect of task progression on the evoked response and reaction times for the composition and list of nouns tasks. Voltage values correspond to the average in a window centered at the maximum value for the trial statistic (F3, 352–452 ms). **A**. Voltage values for the first and last 33 trials of each condition for each task were averaged. There is an increase in negative voltage as the composition task unfolds. Composition task conditions are shown in shades of blue, list of nouns conditions in shades of green. Error bars represent 95% confidence intervals. **B**. Correlations between voltage and reaction times for the composition task (top) and the list of nouns task (bottom).

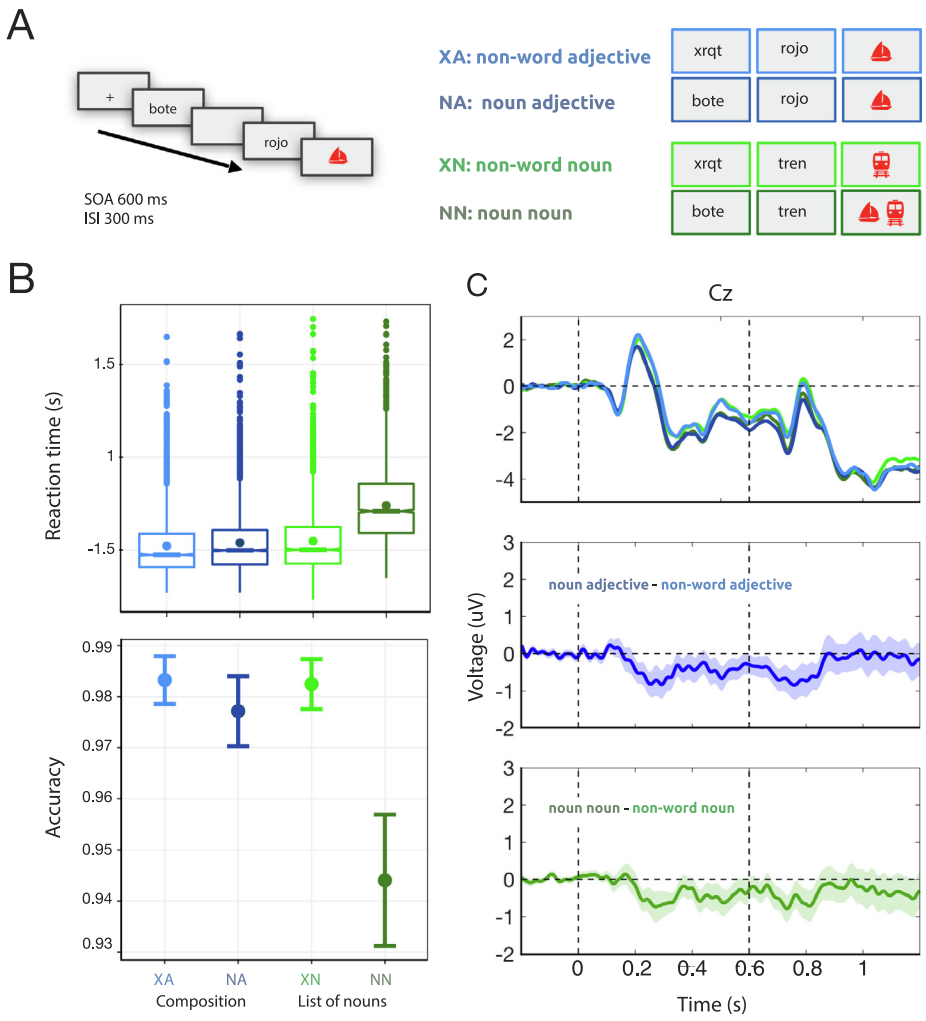


Fig. 5. Experiment 2. Composition without expectancy. **A**. Participants were presented with two two-word conditions: noun adjective (NA) and noun noun (NN) conditions, and two one-word conditions: non-word adjective (XA) and non-word noun (XN) conditions. **B**. Reaction times and accuracy results. Bars indicate 95% confidence interval. **C**. Top graph shows ERP grand averages for each condition at Cz. Bottom graphs show mean differences between conditions with 95% confidence intervals. Vertical dashed lines indicate first and second stimulus onset.

The cluster permutation analysis yielded 9 positive and 8 negative clusters; however, none of them reached significance ($p > 0.64$). To further explore this result, we proceeded to do a post hoc analysis for the difference between each pair of conditions. The comparison between the composition task conditions yielded: three significant positive voltage clusters (1) from 0.617 to 0.836 s, $p < 0.001$, (2) from 0.336 to 0.397 s, $p = 0.006$ and (3) from 0.217 to 0.279 s, $p = 0.007$; and two significant negative voltage clusters: (1) from 0.195 to 0.434 s, $p < 0.001$, (2) from 0.621 to 0.859 s, $p < 0.001$. The analysis for the list of nouns conditions gave a comparable result: two significant positive voltage clusters (1) from 0.314 to 0.533 s, $p < 0.001$and (2) from 0.555 to 0.756, $p < 0.001$. Furthermore, three significant negative voltage clusters were obtained: (1) from 0.199 to 0.543 s, $p < 0.001$, (2) from 0.625 to 0.758 s, $p = 0.002$ and (3) from 0.762 to 0.843 s, $p = 0.008$ (see Fig. 3 in Supplementary Material).

3.2.3. TFCE cluster permutation analyses and linear models

In order to keep analyses similar to Experiment 1, we fitted two linear models to each time-electrode point in epoched EEG data, and subjected each parameter to the TFCE procedure (see Table 1 in supplementary material). The simple model yielded no significant effect of task (FCz, t = 0.872 s, $t_{tfce} = -1.37 \times 10^3$, $p = 0.48$), a significant effect of number of words (PO8, t = 0.267 s, $t_{tfce} = 9.07 \times 10^3$, $p < 0.001$), a significant trial number effect (FCz, t = 0.964 s, $t_{tfce} = -4.89 \times 10^3$, $p < 0.001$) (Fig. 6A), no significant interaction between task and number of words (FC6, t = 0.527 s, $t_{tfce} = -955.2$, $p = 0.87$) (Fig. 6A) and no significant effect for the interaction between task and trial (PO4, t = 0.950 s, $t_{tfce} = -1.42 \times 10^3$, $p = 0.49$).

The complete model with all the trial interactions showed the following results. We found no significant task effect (P7, t = 0.911 s, $t_{tfce} = 2.21 \times 10^3$, $p = 0.33$), a significant effect of number of words (P8, t = 0.269 s, $t_{tfce} = 1.70 \times 10^4$, $p = 0.001$), a significant trial number effect (FCz, t = 1.08 s, $t_{tfce} = -7.19 \times 10^3$, $p < 0.001$) (Fig. 6B), no significant interaction between task and number of words (F7, t = 0.950 s, $t_{tfce} = 1.52 \times 10^3$, $p = 0.30$) (Fig. 6B), no significant interaction between task, number of words and trial number (CP4, t = 0.572, $t_{tfce} = -1.24 \times 10^3$, $p = 0.59$), no interaction between task and trial number (P7, t = 0.007s, $t_{tfice} = 1.28 \times 10^3$, $p = 0.62$) and no effect between number of words and trial number (PO8, t = 0.972 s, $t_{tfce} = -1.59 \times 10^3$, $p = 0.25$).

3.2.4. Correlation between voltage and reaction times

We found a significant positive correlation between the average voltage for FCz between 950–1050 ms (t_{tfce} maximum statistic was obtained at around 1000 ms on that electrode) and reaction times on a trial-to-trial basis. Lower reaction times correlated with more negative potentials (FCz, rrm (10376) = 0.069, 95% CI [0.050, 0.088],

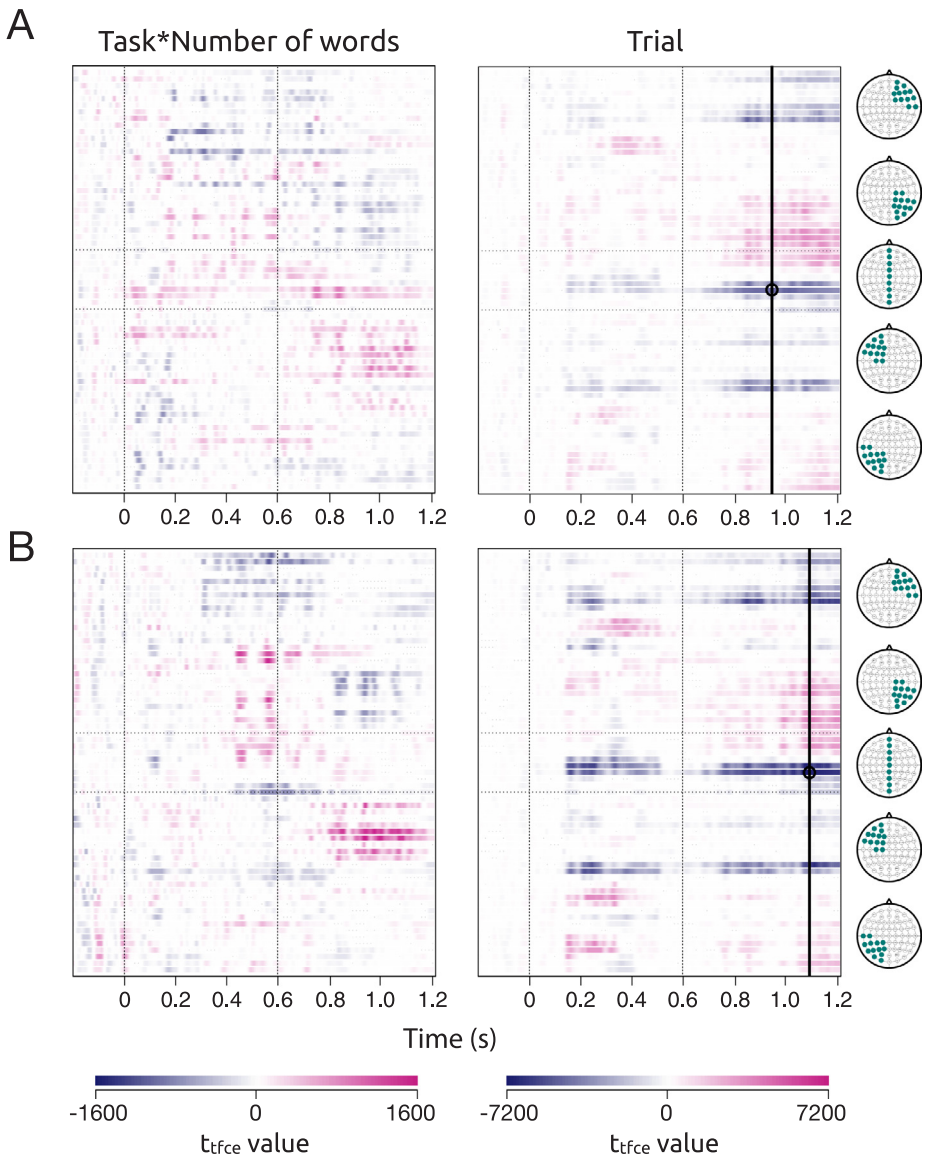


Fig. 6. Threshold-free cluster enhancement results on t-values derived from linear models fitted to EEG data from experiment 2. **A**. Simple model t_{tfce} statistics for the interaction between task and number of words (left) and for trial (right). **B**. Complete model t_{tfce} statistics for the interaction between task and number of words (left) and for trial (right). Line and circle indicate time-electrode maximum statistic values that reached significance in permutation tests.

$p < 0.001$) across conditions.

3.2.5. *Time–frequency analysis*

In the Supplementary Materials we present the results of the time–frequency analysis. We found no significant clusters for the interaction between task and number of words for the gamma, alpha and beta bands.

4. Discussion

In this work we have adapted Bemis & Pylkkänen's (Bemis & Pylkkänen, 2011) experimental paradigm in order to study elementary language composition in Spanish using EEG. As the original experiment was done in MEG and with English stimuli, our adaptation required many changes. In contrast to English, canonical adjectivation of a noun in Spanish is done post-nominally; this imposed the need to add a second control task to keep the critical word type consistent in both the composition and the list task. Furthermore, we adapted the analysis to EEG data, and used two different cluster finding techniques. Despite these important modifications, we found comparable results to those of the original study. The binding of two items into a single concept yielded a processing advantage, as shown by faster reaction times and a lower error rate for the two-word condition of the composition task compared to conditions in which composition was not possible. Our cluster permutation analysis comparing brain responses to the composition task and the list of nouns task, showed a significant interaction between task and number of words 260–550 ms after second word onset. This effect was driven by a difference between two-word and one-word conditions in the composition task. A similar result in a later time window (410–600 ms) was found when comparing the composition and the list of adjectives tasks. Although consisting of different word classes (nouns and adjectives), the clusters obtained in both comparisons showed a similar temporal and topographical distribution. Furthermore, no difference was found between the list tasks, suggesting that both tasks were equally appropriate as controls. This results were not accompanied by differences in the non-phase-locked activity as no modulation of power was found for gamma, alpha and beta bands (see Supplementary Material for analyses and results).

To the best of our knowledge this is the first study to show an EEG signal temporally consistent with the MEG composition-related activity

using an unbiased rigorous cluster permutation analysis and Spanish stimuli. Although a previous study using EEG was published (Neufeld et al., 2016), the reported composition-related activity is not supported by the interaction analysis between task and number of words, leaving open the possibility that their result is indexing the presence of two words over one word, and not a genuine composition activity. This inconclusive result may be a consequence of the electrode partition choice and the time window targeted.

Interestingly, Neufeld et al. (2016) reported a difference between the composition task conditions before second word onset. The authors interpret this pre-combinatorial activity as a syntactic building process, arguing that the adjective-noun syntactic structure is initialized before second word onset to allocate the expected noun. Even though carrying the same classical ERP analysis on our data did not yield the exact same results (see Supplementary Material), our cluster permutation analysis output (Fig. 2A and Fig. 2B) shows that the same electrodes that comprise the significant clusters seem to also take part in smaller clusters before second word onset (this is clear for anterior and posterior left hemisphere electrodes). Although Neufeld and collaborators' interpretation is reasonable, we propose an alternative explanation. For all tasks and conditions, after seeing the first word participants knew with certainty what type of word would follow. Nevertheless, it was only the case for the composition task that subjects had to manipulate the first word in relation to the second word. In this way, participants' processing of the first word was conditional to the second word, and a contingency between the first and second stimuli had to be established only during the composition task. This pre-composition activity could then be more related to general expectancy rather to a specific syntactic mechanism. Given that the experimental design imposes the need to establish a relation between the stimuli, CNV is a good candidate explanation for our results. In particular, this component's amplitude is sensitive to attentional demands (Jacobson & Gans, 1981; Low, Coats, Rettig, & McSherry, 1967; Rebert et al., 1967; Simons, ÖHman, & Lang, 1979; Tecce & Scheff, 1969) and task progression (Walter et al., 1964), and it has been consistently shown that CNV increases as the contingency between two stimuli is learned (Cohen, 1969; Hillyard, 1969; Proulx & Picton, 1980). Consistent with CNV behavior, a TFCE permutation analysis allowed us to evidence the presence of a task progression effect specific to the composition task, i.e. amplitude increased on a trial-to-trial basis (Fig. 4A). The maximum effect of trial number occurred before second word onset and was independent of condition. Moreover, a more exhaustive model including all interactions between predictor variables rendered the crucial interaction between number of words and task non-significant. This results suggest that the composition-related activity obtained in our EEG experiment is contaminated by an anticipatory process. In agreeement with the electrophysiological response described above, we found a correlation between the voltage amplitude for each trial at the location of maximum effect of trial number (as indicated by the TFCE analysis) and the participants's response times (Fig. 4B), such that response times are lower as voltage values get more negative, suggesting its involvement in response preparation. Importantly, this correlation was significant only for the composition task. These results suggest that the composition task shows electrophysiological and behavioral patterns compatible with non-stationary and learning effects which are not present for the list task, indicating that the tasks differ not only in the composition requirement. Hence, the crucial manipulation designed to identify a neural signature of composition may not allow to separate an elementary composition activity from an expectancy-based response in EEG. In order to test this hypothesis we designed an experiment to engage participants in combining two words, but ensuing that expectancy processes affect both tasks similarly. For this, conditions were grouped in a single block such that each two-word trial started with a noun, and an adjective or another non-composable noun could follow. In this manner, participant's expectancy would be equally affecting both tasks, as on any given trial a noun gave no indication of whether the subsequent word would be a composable item or not. Behavioral data indicates that subjects were unifying both elements when a noun was followed by an adjective, as reaction times were similar to the one-word condition and lower than the two nouns condition. However, a cluster permutation analysis on ERP data yielded no interaction between task and number of words. Similarly, no differences in the frequency domain were found (see Supplementary Material). The comparison of each two-word condition to its one-word control shows very similar temporal and topographical distributions (see Fig. 3 in Supplementary Material). After the initial noun is presented, a negative activity starting before second word onset develops and extends to the time-window where a composition effect would be expected, whether the second word allows composition or not. This suggests that the activity we detected in the first experiment may not reflect basic composition, as no interaction effect is obtained when controlling for expectancy. Moreover, a TFCE permutation analysis employing the same models used to evaluate the first experiment showed no significant interaction between task and number of words. Interestingly, a trial number effect was identified after second word onset, which could reflect an anticipatory response to the image to which subjects were required to answer. In this line, negative voltage was correlated with lower reaction times. Although the maximum trial number effect takes place after the second word, following the same behavior than for the composition task a comparable trial number effect can be noticed 200 ms after first word onset (see Fig. 6B). It is important to point out that contrary to experiments in which a typical CNV is observed, Bemis & Pylkkänen's paradigm consists of two linguistic stimuli followed by an image, and therefore two expectancy processes would be elicited: one between the first and second word, and another between second word onset and image presentation. For all tasks and conditions, after second word presentation an anticipatory activity was probably elicited as subjects had to maintain the verbal material available in memory and prepare to give a motor response. We argue that this process would be equal for all conditions. However, a critical difference across tasks is the contingency between stimuli in the composition task, manifested as a task progression effect. In order to determine if our results are compatible with a CNV interpretation it would be relevant to implement the same task with a larger number of trials. If a learning effect accounts for our results, the increase in amplitude of the CNV should reach a plateau once the contingency is fully established. Another interesting alternative would be to increase the inter stimulus intervals (ISIs) between first word, second word and image presentation. This would allow the negative potentials to develop in time, enabling a comparison with the two-component response described for the CNV (Loveless & Sanford, 1974; Weerts & Lang, 1973). In our experiment, the short time that separates the three stimuli probably produces a superposition of anticipatory waves, preventing a proper characterization of these components. Is this anticipatory response in part specific to language processing? According to predictive coding models (Clark, 2013; Farmer, Brown, & Tanenhaus, 2013; Kutas & Federmeier, 2011) the expectancy-related activity could be part of language processing; in particular, it could be elicited by a composition-related process. It is clear though that the activity we observed in our first experiment (only in the composition task) does not correlate with linguistic composition since it appears to be present for both tasks in experiment 2. In this case both conditions with nouns as first word create a similar expectancy as the continuation is not known. Crucially, after the second word the situation is disambiguated, but we do not measure any difference in activity between conditions NA and NN that could be attributable to composition.

It has been shown that a negative potential is elicited when a delay is introduced before sentence-final words or when a specific linguistic stimulus can be predicted (Besson, Faita, Czternasty, & Kutas, 1997; Kaan & Carlisle, 2014; León-Cabrera, Rodríguez-Fornells, & Morís, 2017). Evidence against a simple syntactic interpretation of this anticipatory response (Neufeld et al., 2016) comes from Bentin (1987). In this work, lexical stimuli separated by long ISIs were used to study

neural responses to semantic expectancy. On some trials the first stimulus was a word, and subjects were asked to respond whether the next word was an antonym of the first. In contrast, on trials starting with a non-word, subjects had to perform a lexical decision task on the second stimulus. The elicited complex response (characterized by a sustained negative activity) had a larger amplitude in the antonym task that required the semantic content of the first word to be held in memory. This supports the possibility that the pre-combinatorial activity addressed in Neufeld et al.'s and in our experiments may reflect an anticipatory mechanism related to linguistic processing that is not specific to syntax or semantic composition.

A point that could be raised is whether the similar electrophysiological response for the *noun adjective* and *noun noun* conditions in our second experiment reflects subjects' attempt to compose both nouns. Although noun-noun compounds are widely used in Spanish, compounding is not overly productive. There are hierarchical compounds, where the head-noun can be in first or last position (*hombre arańa* "spiderman", *hidroterapia* "hydrotherapy") and concatenative compounds which according to the final representation can be considered coextensional (*cantante-bailarńc* "Singer-dancer"), additive (*espacio-tiempo* "space–time") or intersective (*centro-derecha* "center-right") (Moyna, 2011). Nevertheless, none of the possible noun pairs used as stimuli correspond to established lexical compounds and we verified that they are not used in natural Spanish constructions. Although we cannot be certain that composition was not attempted by subjects, it would be a bad strategy to perform the task. Participants would benefit from maintaining the verbal material as independent units in order to check each word against the images presented. Moreover, open noun-noun composition would lead to ambiguous results (is a *shoe fork* a type of fork or a type of shoe?). Even though reaction times are higher and accuracy is lower for the noun-noun condition, participants still show a very good performance. This is not surprising, as during reading or hearing a sentence there is no certainty about which word will follow a specific stimuli and people can still correctly compose meanings. Therefore, the uncertainty introduced by our task is in this sense ecological. It could be argued that our results are related to the linguistic differences between Spanish and English, specifically in relation to the order of adjectival modification, such that presentation of the head-noun in the first position elicits a linguistic process that would be absent when presenting an adjective. Nevertheless, there is evidence from Arabic that the LATL shows an increased response to noun phrases resulting from postnominal modification (Westerlund et al., 2015) which is no different from that observed using English stimuli. Moreover, this activity is elicited for English color-object noun phrases presented in reverse order (object-color) when the task demanded subjects to match both words to a single image representation (Bemis & Pylkkänen, 2013c). Notice that our first experiment reproduces all these features. Besides, the initial suggestion of a pre-composition activity for nouns preceded by adjectives was described for EEG using English noun-phrases in their canonical order, so there is at least one published report in English showing this pre-composition expectancy. It might also be questioned whether the difference we observe might be due to the lexical properties of the stimuli used. The original experiment consisted of monosyllabic stimuli whereas we use multisyllabic words. Nevertheless they are moderately frequent and easy to read (well within the window where reading time is insensitive to word length (Rayner, 1998)). In addition, following the English stimuli used in the previous studies the vast majority of words in our study are mono-morphemic, with the word "telephone" as the only exception, and hence we do not think this difference could explain the results.

Irrespective of what turns out to be the explanation of our results, it is doubtful that the activity we found in our first experiment is entirely a composition-based activity. Certainly, our work shows that some results previously published could be confounded with expectancy-based processes. This has been addressed clearly for Neufeld et al. (2016); however this issue is also present in other reports using visual (Bemis & Pylkkänen, 2011, 2013c, 2013a) and auditory (Bemis & Pylkkänen, 2013a) stimuli. In these cases, the authors do not analyze or plot the activity elicited before second word onset, and they do not discuss the potential issues that their experimental design entails. Furthermore, in studies in which expectancy would equally affect noun-adjective constructions (Westerlund & Pylkkänen, 2014; Westerlund et al., 2015; Zhang & Pylkkänen, 2015; Ziegler & Pylkkänen, 2016), conditions are only compared to each other or contrasted with a non word-word condition, disregarding the crucial list control condition. Importantly, these possible objections cannot be applied to all the studies produced on this subject. In particular, the compositional interpretation is supported by production studies. These authors show an increase in LATL's activity when subjects had to describe pictures with adjective-noun constructions in comparison to pictures described by enlisting two concepts (Blanco-Elorrieta, Kastner, Emmorey, & Pylkkänen, 2018; Prato & Pylkkänen, 2014; Pylkkänen et al., 2014). Although experiments were designed such that conditions were arranged in separate blocks there were no expectancy differences across stimuli. Thus our results do not question the involvement of LATL in linguistic composition, but show that EEG measures of brain responses elicited by this paradigm do not provide reliable evidence for an elementary composition marker, and suggest that some experimental designs used in MEG may be subject to the same limitation. On this line, we cannot rule out that the expectancy-related activity showed in this work is more apparent in EEG. This technique could be more sensitive to capture this anticipatory neural activity than MEG thus hindering the detection of a signal specific to composition. Additionally, it is possible that the hypothesis driven analyses carried out on specific regions of interest in the MEG studies manage to isolate composition from other processes. In summary, we successfully adapted Bemis and Pylkkänen (2011) minimal composition paradigm to EEG and a language like Spanish that uses post-nominal adjectivation. We found an increased negativity for nouns followed by adjectives in a time window consistent with the composition-related activity described in the MEG literature. We have also shown the relevance of applying data-driven analyses that take into consideration task development effects, and adapted appropriate methods to do so. Finally, we introduced a non-blocked variant of the experiment to separate the contributions of composition and general expectancy effects to the measured signals. We suggest that the composition-related activity measured with EEG may be at least in part carried by expectancy-related processes arising from the block design of the experiment. Whether it is possible to find an unequivocal electrophysiological marker of elementary composition is a question that remains open and should be addressed in future work.

Acknowledgements

EF was supported by the Agencia Nacional de Investigación e Innovación, POS_NAC_2015_1_109472, and by CAP-UDELAR. AC and JCVL received support from SNI-ANII, CSIC-UDELAR and PEDECIBA.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.bandl.2020.104837.

References

Bakdash, J. Z. & Marusich, L. R. (2018). *rmcorr: Repeated Measures Correlation*. https://cran.r-project.org/package=rmcorr.

Baron, S. G., & Osherson, D. (2011). Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage, 55*, 1847–1852. https://doi.org/10.1016/j.neuroimage.2011.01.066.

Baron, S. G., Thompson-Schill, S. L., Weber, M., & Osherson, D. (2010). An early stage of conceptual combination: Superimposition of constituent concepts in left anterolateral temporal lobe. *Cognitive Neuroscience, 1*, 44–51. https://doi.org/10.1080/

17588920903548751.
Barry, R. J., Steiner, G. Z., De Blasio, F. M., Fogarty, J. S., Karamacoska, D., & MacDonald, B. (2019). Components in the P300: Don't forget the Novelty P3! *Psychophysiology* (pp. 1–15). doi:10.1111/psyp.13371.
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 251–264. https://doi.org/10.18637/jss.v067.i01 http://www.jstatsoft.org/v67/i01/.
Bemis, D. K., & Pylkkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience, 31*, 2801–2814. https://doi.org/10.1523/JNEUROSCI.5003-10.2011 http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.5003-10.2011.
Bemis, D. K., & Pylkkänen, L. (2013a). Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral Cortex, 23*, 1859–1873. https://doi.org/10.1093/cercor/bhs170.
Bemis, D. K., & Pylkkänen, L. (2013b). Combination across domains: An MEG investigation into the relationship between mathematical, pictorial, and linguistic processing. *Frontiers in Psychology, 3*, 1–20. https://doi.org/10.3389/fpsyg.2012.00583.
Bemis, D. K., & Pylkkänen, L. (2013c). Flexible composition: MEG evidence for the deployment of basic combinatorial linguistic mechanisms in response to task demands. *PLoS ONE, 8*. https://doi.org/10.1371/journal.pone.0073949.
Bentin, S. (1987). Event-related potentials, semantic processes, and expectancy factors in word recognition. *Brain and Language, 31*, 308–327. https://doi.org/10.1016/0093-934X(87)90077-0.
Bentin, S., & McCarthy, G. (1994). The effects of immediate stimulus repetition on reaction time and event-related potentials in tasks of different complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 130–149. https://doi.org/10.1037/0278-7393.20.1.130 http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.20.1.130.
Besson, M., Faita, F., Czternasty, C., & Kutas, M. (1997). What's in a pause: Event-related potential analysis of temporal disruptions in written and spoken sentences. *Biological Psychology, 46*, 3–23. https://doi.org/10.1016/S0301-0511(96)05215-5.
Blanco-Elorrieta, E., Kastner, I., Emmorey, K., & Pylkkänen, L. (2018). Shared neural correlates for building phrases in signed and spoken language. *Scientific Reports, 8*, 1–10. https://doi.org/10.1038/s41598-018-23915-0.
Blanco-Elorrieta, E., & Pylkkänen, 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.
Bosque, I. & Demonte, V. (1999). *Gramática Descriptiva de la Lengua Española. Number v. 1 in Nebrija y Bello*. Librería Tirant lo Blanch. https://books.google.fr/books?id=atcp_KuI4ysC.
Brunia, C., & Vingerhoets, A. (1981). Opposite hemisphere differences in movement related potentials preceding foot and finger flexions. *Biological Psychology, 13*, 261–269. https://doi.org/10.1016/0301-0511(81)90041-7 http://linkinghub.elsevier.com/retrieve/pii/0301051181900417.
Brush, C. J., Ehmann, P. J., Hajcak, G., Selby, E. A., & Alderman, B. L. (2018). Using multilevel modeling to examine blunted neural responses to reward in major depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*, 1032–1039. https://doi.org/10.1016/j.bpsc.2018.04.003.
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–204. https://doi.org/10.1017/S0140525X12000477.
Cohen, J. (1969). Very slow brain potentials relating to expectancy: The CNV. *Average evoked potentials: Methods, results, and evaluations* (pp. 143–198). Washington: US National Aeronautics and Space Administration. https://doi.org/10.1037/13016-004 https://ntrs.nasa.gov/search.jsp?R=19700007575 http://content.apa.org/books/13016-004.
Coutanche, M. N., & Thompson-Schill, S. L. (2015). Creating concepts from converging features in human cortex. *Cerebral Cortex, 25*, 2584–2593. https://doi.org/10.1093/cercor/bhu057.
Donald, M. W. (1980). Memory, Learning and Event-Related Potentials. In H.H. Kornhubek & L. Deecke (Eds.), *Progress in brain research. Progress in brain research* (Vol. 54, pp. 615–627). Elsevier. doi:10.1016/S0079-6123(08)61681-7. http://linkinghub.elsevier.com/retrieve/pii/S0079612308616817 https://linkinghub.elsevier.com/retrieve/pii/S0079612308616817 http://www.sciencedirect.com/science/article/pii/S0079612308616817.
Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods, 45*, 1246–1258. https://doi.org/10.3758/s13428-013-0326-1.
Farmer, T. A., Brown, M., & Tanenhaus, M. K. (2013). Prediction, explanation, and the role of generative models in language processing. *Behavioral and Brain Sciences, 36*, 211–212. https://doi.org/10.1017/S0140525X12002312 https://www.cambridge.org/core/product/identifier/S0140525X12002312/type/journal_article.
Friederici, A. D. (2017). *Language in our brain: the origins of a uniquely human capacity*. https://mitpress.mit.edu/books/language-our-brain.
Friederici, A. D., Meyer, M., & Cramon, D. Y. V. (2000). Auditory language comprehension: An event-related fMRI study on the processing of syntactic and lexical information. *Brain and Language, 300*, 289–300. https://doi.org/10.1006/brln.2000.2313.
Gaillard, A. (1976). Effects of warning-signal modality on the contingent negative variation (CNV). *Biological Psychology, 4*, 139–153. https://doi.org/10.1016/0301-0511(76)90013-2 http://linkinghub.elsevier.com/retrieve/pii/0301051176900132.
Hagoort, P. (2019). The neurobiology of language beyond single-word processing. *Science, 366*, 55–58. https://doi.org/10.1126/science.aax0289 http://www.sciencemag.org/lookup/doi/10.1126/science.aax0289.
Hagoort, P., Baggio, G., & Wlllems, R. M. (2009). Semantic unification. *The cognitive neurosciences* (pp. 819–835). (4th ed.). Cambridge, MA, US: Massachusetts Institute of Technology.
Hillebrand, A., & Barnes, G. R. (2002). A quantitative assessment of the sensitivity of whole-head MEG to activity in the adult human cortex. *NeuroImage, 16*, 638–650. https://doi.org/10.1006/nimg.2002.1102.
Hillyard, S. A. (1969). Relationships between the contingent negative variation (CNV) and reaction time. *Physiology and Behavior, 4*, 351–357. https://doi.org/10.1016/0031-9384(69)90188-7.
Humphries, C., Love, T., Swinney, D., & Hickok, G. (2005). Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. *Human Brain Mapping, 138*, 128–138. https://doi.org/10.1002/hbm.20148.
Irwin, D. A., Knott, J. R., McAdam, D. W., & Rebert, C. S. (1966). Motivational determinants of the contingent negative variation. *Electroencephalography and Clinical Neurophysiology, 21*, 538–543. https://doi.org/10.1016/0013-4694(66)90172-6.
Jacobson, G. P., & Gans, D. P. (1981). The contingent negative variation as an indicator of speech discrimination difficulty. *Journal of Speech Language and Hearing Research, 24*, 345. https://doi.org/10.1044/jshr.2403.345 http://jslhr.pubs.asha.org/article.aspx?doi=10.1044/jshr.2403.345.
Jasper, H. H. (1958). *The tenn twenty electrode system of the international federation*. doi:10.1016/0013-4694(58)90053-1.
Junghöfer, M., Elbert, T., Tucker, D. M., & Rockstroh, B. (2000). Statistical control of artifacts in dense array EEG/MEG studies. *Psychophysiology, 37*, 523–532. https://doi.org/10.1017/S0048577200980624.
Kaan, E., & Carlisle, E. (2014). ERP indices of stimulus prediction in letter sequences. *Brain Sciences, 4*, 509–531. https://doi.org/10.3390/brainsci4040509.
Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*, 627–633. https://doi.org/10.3758/BRM.42.3.627.
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*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123 http://www.annualreviews.org/doi/10.1146/annurev.psych.093008.131123.
Kutas, M., & Hillyard, S. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*, 203–205. https://doi.org/10.1126/science.7350657 http://www.sciencemag.org/cgi/doi/10.1126/science.7350657.
León-Cabrera, P., Rodríguez-Fornells, A., & Morís, J. (2017). Electrophysiological correlates of semantic anticipation during speech comprehension. *Neuropsychologia, 99*, 326–334. https://doi.org/10.1016/j.neuropsychologia.2017.02.026.
Loveless, N. E. (1975). The effect of warning interval on signal detection and event-related slow potentials of the brain. *Perception & Psychophysics, 17*, 565–570. https://doi.org/10.3758/BF03203970.
Loveless, N. E., & Sanford, A. J. (1974). Slow potential correlates of preparatory set. *Biological Psychology, 1*, 303–314. https://doi.org/10.1016/0301-0511(74)90005-2.
Low, M. D., Coats, A. C., Rettig, G. M., & McSherry, J. W. (1967). Anxiety, attentiveness-alertness: A phenomenological study of the CNV. *Neuropsychologia, 5*, 379–384. https://doi.org/10.1016/0028-3932(67)90009-7 http://www.rocksbackpages.com/Library/Article/eels-the-freak-shall-inherit-the-earth http://linkinghub.elsevier.com/retrieve/pii/0028393267900097.
Luck, S.J.S.J. (2005). *An introduction to the event-related potential technique*. doi:10.1118/1.4736938. arXiv:9780262621960.
Malmivuo, J., & Plonsey, R. (1995). *Bioelectromagnetism principles and applications of bioelectric and biomagnetic fields*, *Vol. 15*Oxford University Presshttps://doi.org/10.1093/acprof:oso/9780195058239.001.0001.
Martin, A. E., & Baggio, G. (2020). Modelling meaning composition from formalism to mechanism. *Philosophical Transactions of the Royal Society B: Biological Sciences, 375*, 20190298. https://doi.org/10.1098/rstb.2019.0298 https://royalsocietypublishing.org/doi/10.1098/rstb.2019.0298.
Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., & Salamon, G. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience, 5*, 467–479. https://doi.org/10.1162/jocn.1993.5.4.467.
Mensen, A., & Khatami, R. (2013). Advanced EEG analysis using threshold-free cluster-enhancement and non-parametric statistics. *NeuroImage, 67*, 111–118. https://doi.org/10.1016/j.neuroimage.2012.10.027.
Michel, J. B., Kui Shen, Y., Presser Aiden, A., Veres, A., Gray, M. K., Pickett, J. P., ... Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*, 176–182. https://doi.org/10.1126/science.1199644.
Michel, C. M., Murray, M. M., Lantz, G., Gonzalez, S., Spinelli, L., & Grave de Peralta, R. (2004). EEG source imaging. *Clinical Neurophysiology, 115*, 2195–2222. https://doi.org/10.1016/j.clinph.2004.06.001 https://linkinghub.elsevier.com/retrieve/pii/S1388245704002135.
Molinaro, N., Carreiras, M., & Duñabeitia, J. A. (2012). Semantic combinatorial processing of non-anomalous expressions. *NeuroImage, 59*, 3488–3501. https://doi.org/10.1016/j.neuroimage.2011.11.009.
Moyna, M. I. (2011). *Compound words in Spanish: Theory and history*.
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*, e0158446. https://doi.org/10.1371/journal.pone.0158446 https://dx.plos.org/10.1371/journal.pone.0158446.
Okada, Y., Lähteenmäki, A., & Xu, C. (1999). Comparison of MEG and EEG on the basis of somatic evoked responses elicited by stimulation of the snout in the juvenile swine. *Clinical Neurophysiology, 110*, 214–229. https://doi.org/10.1016/S0013-4694(98)00111-4.
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. doi:10.1155/2011/156869. arXiv:156869.
Parks, N. A., Gannon, M. A., Long, S. M., & Young, M. E. (2016). Bootstrap signal-to-noise

confidence intervals: An objective method for subject exclusion and quality control in ERP studies. *Frontiers in Human Neuroscience, 10*, 1–15. https://doi.org/10.3389/fnhum.2016.00050 http://journal.frontiersin.org/Article/10.3389/fnhum.2016.00050/abstract.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8*, 976–987. https://doi.org/10.1038/nrn2277.

Peirce, J. W. (2007). PsychoPy-psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017.

Poeppel, D., Emmorey, K., Hickok, G., & Pylkkänen, L. (2012). Towards a new neurobiology of language. *Journal of Neuroscience, 32*, 14125–14131. https://doi.org/10.1523/JNEUROSCI.3244-12.2012 http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3244-12.2012.

Polich, J. (1989). Habituation of P300 from auditory stimuli. *Psychobiology, 17*, 19–28. https://doi.org/10.3758/BF03337813.

Poon, L. W., Thompson, L. W., Williams, R. B., & Marsh, G. R. (1974). Changes of Anteroposterior distribution of CNV and late positive component as a function of information processing demands. *Psychophysiology, 11*, 660–673. https://doi.org/10.1111/j.1469-8986.1974.tb01135.x http://doi.wiley.com/10.1111/j.1469-8986.1974.tb01135.x.

Poortman, E. B., & Pylkkänen, L. (2016). Adjective conjunction as a window into the LATL's contribution to conceptual combination. *Brain and Language*. https://doi.org/10.1016/j.bandl.2016.07.006.

Prato, P. D., & Pylkkänen, L. (2014). MEG evidence for conceptual combination but not numeral quantification in the left anterior temporal lobe during language production. *Frontiers in Psychology, 5*, 1–11. https://doi.org/10.3389/fpsyg.2014.00524.

Proulx, G. B., & Picton, T. W. (1980). The CNV during cognitive learning and extinction. *Progress in Brain Research, 54*, 309–313. https://doi.org/10.1016/S0079-6123(08)61640-4.

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science, 366*, 62–66. https://doi.org/10.1126/science.aax0050 http://www.sciencemag.org/lookup/doi/10.1126/science.aax0050.

Pylkkänen, L. (2020). Neural basis of basic composition: What we have learned from the red-boat studies and their extensions. *Philosophical Transactions of the Royal Society B: Biological Sciences, 375*. https://doi.org/10.1098/rstb.2019.0299.

Pylkkänen, L., Bemis, D. K., & Blanco, E. (2014). Building phrases in language production: An MEG study of simple composition. *Cognition, 133*, 371–384. https://doi.org/10.1016/j.cognition.2014.07.001.

Pylkkänen, L., Brennan, J. R., & Bemis, D. K. (2011). Grounding the cognitive neuroscience of semantics in linguistic theory. *Language and Cognitive Processes, 26*, 1317–1337. https://doi.org/10.1080/01690965.2010.527490 http://www.tandfonline.com/doi/abs/10.1080/01690965.2010.527490.

Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2016). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience, 18*, 42–55. https://doi.org/10.1038/nrn.2016.150.

Ravden, D., & Polich, J. (1998). Habituation of P300 from visual stimuli. *International Journal of Psychophysiology, 30*, 359–365. https://doi.org/10.1016/S0167-8760(98)00039-7.

Rayner, K. (1998). Eye movements in reading and information processing: 20 Years of research. *Psychological Bulletin, 124*, 372–422. https://doi.org/10.1037/0033-2909.124.3.372.

R Core Team (2017). *R: A language and environment for statistical computing*. Austria: R Foundation for Statistical Computing Vienna https://www.r-project.org/.

Rebert, C. S., McAdam, D. W., & Knott, J. R. (1967). *Slow potential change in human brain related to level of motivation*. doi:10.1037/h0024146.

Rohrbaugh, J. W. & Gaillard, A. W. (1983). 13 Sensory and motor aspects of the contingent negative variation. In *Tutorials in ERP research: Endogenous components* (pp. 269–310). doi:10.1016/S0166-4115(08)62044-0. http://linkinghub.elsevier.com/retrieve/pii/S0166411508620440 https://linkinghub.elsevier.com/retrieve/pii/S0166411508620440.

Segaert, K., Mazaheri, A., & Hagoort, P. (2018). Binding language: Structuring sentences through precisely timed oscillatory mechanisms. *European Journal of Neuroscience, 1*, 1–12. https://doi.org/10.1111/ejn.13816.

Simons, R. F., ÖHman, A., & Lang, P. J. (1979). Anticipation and response set: Cortical, cardiac, and electrodermal correlates. *Psychophysiology, 16*, 222–233. https://doi.org/10.1111/j.1469-8986.1979.tb02982.x.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage, 44*, 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061.

Snijders, T. M., Vosse, T., Kempen, G., Berkum, J. A. V., Petersson, K. M., & Hagoort, P. (2009). Retrieval and unification of syntactic structure in sentence comprehension: An fMRI study using word-category ambiguity. *Cerebral Cortex, 19*, 1493–1503. https://doi.org/10.1093/cercor/bhn187.

Ströberg, K., Andersen, L. M., & Wiens, S. (2017). Electrocortical N400 effects of semantic satiation. *Frontiers in Psychology, 8*, 1–14. https://doi.org/10.3389/fpsyg.2017.02117.

Tecce, J. J., Savignano-Bowman, J., & Meinbresse, D. (1976). Contingent negative variation and the distraction-arousal hypothesis. *Electroencephalography and Clinical Neurophysiology, 41*, 277–286. https://doi.org/10.1016/0013-4694(76)90120-6.

Tecce, J. J., & Scheff, N. M. (1969). Attention reduction and suppressed direct-current potentials in the human brain. *Science, 164*, 331–333. https://doi.org/10.1126/science.164.3877.331 http://www.sciencemag.org/cgi/doi/10.1126/science.164.3877.331.

Volpert-Esmond, H. I., Merkle, E. C., Levsen, M. P., Ito, T. A., & Bartholow, B. D. (2018). Using trial-level data and multilevel modeling to investigate within-task change in event-related potentials. *Psychophysiology, 55*, 1–12. https://doi.org/10.1111/psyp.13044.

Vossen, H., Van Breukelen, G., Hermens, H., Van Os, J., & Lousberg, R. (2011). More potential in statistical analyses of event-related potentials: A mixed regression approach. *International Journal of Methods in Psychiatric Research, 16*. https://doi.org/10.1002/mpr.348 arXiv:1106.4512.

Walter, W. G., Cooper, R., Aldrige, V. J., McCallum, W. C., & Winter, A. L. (1964). Contingent negative variation: An electric sign of sensori-motor association and expectancy in the human brain. *Nature, 203*, 380. https://doi.org/10.1038/203380a0.

Weerts, T. C., & Lang, P. J. (1973). The effects of eye fixation and stimulus and response location on the Contingent Negative Variation (CNV). *Biological Psychology, 1*, 1–19. https://doi.org/10.1016/0301-0511(73)90010-0.

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 https://linkinghub.elsevier.com/retrieve/pii/S0093934X14001990.

Westerlund, M., & Pylkkänen, L. (2014). The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia, 57*, 59–70. https://doi.org/10.1016/j.neuropsychologia.2014.03.001.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid P-hacking. *Frontiers in Psychology, 7*, 1–12. https://doi.org/10.3389/fpsyg.2016.01832.

Williamson, S., & Kaufman, L. (1981). Biomagnetism. *Journal of Magnetism and Magnetic Materials, 22*, 129–201. https://doi.org/10.1016/0304-8853(81)90078-0 https://linkinghub.elsevier.com/retrieve/pii/0304885381900780.

Zaccarella, E., Meyer, L., Makuuchi, M., & Friederici, A. D. (2017). Building by syntax: The neural basis of minimal linguistic structures. *Cerebral cortex* (New York, N.Y.: 1991) 27, 411–421. doi:10.1093/cercor/bhv234.

Zaccarella, E., & Friederici, A. D. (2015). Merge in the human brain: A sub-region based functional investigation in the left pars opercularis. *Frontiers in Psychology, 6*, 1–9. https://doi.org/10.3389/fpsyg.2015.01818.

Zhang, L., & Pylkkänen, L. (2015). NeuroImage The interplay of composition and concept speci fi city 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., & Pylkkänen, L. (2016). Neuropsychologia 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.