HOLISTIC PROCESSING OF REGULAR FOUR-WORD SEQUENCES

Holistic Processing of Regular Four-word Sequences: A Behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall.

Antoine Tremblay and Harald Baayen

University of Alberta

Author Note

Antoine Tremblay and Harald Baayen, Department of Linguistics, University of Alberta.

This paper is supported by a Major Collaborative Research Initiative Grant (number 412-2001-1009) and a Doctoral Fellowship (number 752-2006-1315) from the Social Sciences and Humanities Research Council of Canada (SSHRC). We wish to thank Dr. Gary Libben, Dr. Ruth Ann Atchley, Dr. Patrick Bolger, Dr. Jeremy Caplan, Dr. Kathryn Conklin, and several attendees from the 6th *International Conference on the Mental Lexicon*, held in Banff, 7-10 October 2008, the *Canadian Linguistic Association* conference, held at the University of British Columbia, 31 May-2 June 2008, as well as from the *Formulaic Language Research Network Third International Postgraduate Conference*, held in Nottingham, UK, 19-20 June 2008.

Correspondence concerning this article should be addressed to Antoine Tremblay, Department of Linguistics, University of Alberta, 4-32 Assiniboia Hall, Edmonton Alberta, T6G 2E7, Canada. E-mail: antoinet@ualberta.ca; Web site: www.ualberta.ca/~antoinet

Introduction

It is generally accepted that we store representations of words in a mental dictionary, which we call the lexicon. However, what exactly is stored in the mental lexicon remains an open question. For example, do we store the word *dog* as well as its plural form *dogs*, or do we only store *dog* and have a rule (NOUN + *-s* = plural) to compute the plural form. A similar question arises regarding the storage vs. computation of multi-word units, wherein a single meaning is attached to a string of words. The canonical examples are phrasal verbs (*give up*), compounds (*jailbird*), and idioms (*kick the bucket*). By their very nature, these items offer us an opportunity to understand the interplay between storage and computation. Corpus-based research has shown that the tendency for words to occur together in discourse extends far beyond the canonical (e.g., Biber, Johansson, Leech, Conrad, and Finegan, 1999; Bod, Scha, and Sima'an, 2003). In fact, other sequences of words, such as *in the middle of*, pattern together with such frequency that it may be enough to treat them as single units in their own right (Biber et al., 1999). There is a good psycholinguistic basis for proposing that the mind stores and processes these multi-word units as wholes (e.g., Bod, 2001; Schmitt and Underwood, 2004; Underwood, Schmitt, and Galpin, 2004; Jiang and Nekrasova, 2007; Conklin and Schmitt, 2008; Tremblay, Derwing, Libben, and Westbury, 2008). The main reason may be the structure of the mind itself, which stores a vast number of information units in long-term memory, but is only able to process about 4-7 of them online, in working memory (Miller, 1956). In effect, the mind might make use of a relatively unlimited resource (long-term memory) to compensate for relatively limited one (working-memory) by storing a number of frequently needed/used multi-word units as wholes. Such units could be easily retrieved and used as wholes without the need to compose them on-line through word selection and grammatical sequencing. Such an ability would place less

demand on cognitive resources because the multi-word units would be "ready to go" and require little or no additional processing.

In the realm of psycholinguistics, research on questions of storage and computation has for the most part disregarded sentences on grounds that they are necessarily derived via general rules from individual words (Chomsky, 1988). That is, the meaning of a sentence such as *I play soccer* can be derived from the individual words that compose it and is therefore not stored in the lexicon. Such approaches to language are further supported by the observation that storing every possible utterance one has ever heard and/or seen is clearly infeasible. There is mounting evidence, however, suggesting that the repertoire of sentences native speakers commonly use is more restricted and repetitive than was previously thought (e.g., Biber et al., 1999). As a result, the notion that we store regular and irregular utterances becomes more credible. This has led researchers such as Goldberg (1995) and Bod, Scha, and Sima'an (2003) to propose models of language where more or less abstract "patterns" or "constructions" of variable lengths and degrees of idiosyncrasy emerge from the accumulation of stored instances (e.g., *to pull X's leg*, *It is X that Y*, *Subject – Verb – Object*). When confronted with the need to produce a novel sentence, for example, one would choose the appropriate construction and fill out its open slots with the appropriate material (potentially other constructions). Recent findings suggest that, in addition to full sentences, regular sentence fragments are also stored in the mental lexicon. For instance, Biber et al.'s (1999) study of the *British National Corpus* found that frequent regular multi-word strings such as *I think that* and *I don't know* are more likely to be repeated as wholes (e.g., *I think that I think that DNA is a very good example, because erm, it presumably, it was initially a piece of jury search*) and that pauses frequently occur at their boundaries (e.g., *I mean they fought valiantly for peace but I, I think that erm <pause> the maternity bill I think is what*

everybody admits that we shall always go down as being noted for). Such a hypothesis implies that the mental lexicon keeps track of how many times it has experienced not only words, but also regular sentences and sentence fragments. To put it in terms of Hebb's law of neural plasticity, one could say that words used together wire together.

Supporting evidence for this idea is provided by a handful of recent psycholinguistic studies that report reduced processing loads for regular high frequency multi-word sequences (e.g., *I said to her*) relative to regular low frequency items (e.g., *I was to her*) in L1 and L2 speakers of English (e.g., Bod, 2001; Jiang and Nekrasova, 2007; Tremblay et al., 2008). In a recent study, Tremblay et al. (2008) found that highly frequent four- and five-word sequences referred to as lexical bundles (>= 10 and 5 occurrences per million respectively) provide on-line processing advantages over comparable, low-frequency sequences (< 10 and 5 per million respectively). The impression arising from such findings is that of a sharp lexical vs. non-lexical bundle dichotomy. It is conceivable, however, that these categories are epiphenomenal to the factorial design Tremblay et al. (2008) used in their study. Would the same distinction have emerged had they considered sequences ranging from very low to very high frequencies? Moreover, is it reasonable to believe that non-lexical bundles with a frequency of 1 per million behave (exactly) like those with frequency of 9 per million? Are the latter strings radically different from lexical bundles with a frequency of 10 or 11 per million? What about lexical bundles with a frequency of 20, 50, or 100? In order to investigate these issues, we conducted an immediate free recall task where the stimuli consisted of 432 regular four-word sequences with whole-string frequencies ranging roughly from 0.01 to 100 per million.

Immediate Free Recall

In immediate free recall tasks, participants are asked to recall without delay items from a previously studied list in any order. In such tasks, single word frequency was revealed to be a paradoxical predictor. When comparing pure lists of high-frequency words (e.g., *letter*, *money*, *people*) to pure lists of low-frequency items (e.g., *dike*, *strong*, *key*), recall is usually better for high-frequency items (e.g., DeLosh and McDaniel, 1996; Merritt, DeLosh, and McDaniel, 2006). Surprisingly however, in lists consisting of mixed high- and low-frequency words, the advantage is robustly given to low-frequency items (e.g., DeLosh and McDaniel, 1996; Merrit et al., 2006; Tse and Altarriba, 2007). In line with classical theories of information processing (e.g., Johnston and Heinz, 1978), DeLosh and McDaniel (1996) argue that this effect is attributable to the fact that a greater amount of attentional resources is allocated to the processing and interpretation of salient low-frequency items than trivial high-frequency items, which allows for suppression and inattention.

In light of this and given the mixed-frequency stimulus list used in this experiment, we expect that items associated with lower frequencies and lower frequency-related measures such as probability of occurrence (e.g., LogitABCD; see Table 1 below) will be correctly recalled more often than strings associated with higher frequencies and higher frequency-related variables.

Focus of attention, probability and frequency are known to modulate a number of ERP components, among others the P1, N1, and P2 deflections. The P1 is the earliest visual event-related potential known to vary with spatial attention, state of arousal, lexical frequency, and probability. It arises at occipital scalp sites 60-90 msec and peaks 100-150 msec post-stimulus

(e.g., Luck, 2005, and references cited therein; Penolazzi, Hauk, and Pulvermüller, 2007, and references cited therein). The early portion of the P1 (peak latency 98-110 msec) is believed to have extrastriate generators (in the middle of the occipital gyrus) that possibly include areas V2 and V4 of the visual cortex, whereas the later portion (peak latency 136-146 msec) arises from the ventral extrastriate cortex of the fusiform gyrus (Hillyard, Teder-Sälejärvi, and Münte, 1998, Russo, Martinez, Sereno, Pitzalis, and Hillyard, 2002; Luck, 2005).

The N1 is composed of at least three subcomponents, one which peaks at frontal and central sites ~ 100-150 msec after stimulus onset (N1a), and two later ones at posterior and occipital scalp sites with a peak latency ~ 150-200 msec (N1b). The N1 and particularly the anterior N1, believed to originate from centro-parietal sources (Di Russo et al., 2002), is known to be sensitive to spatial attention (Luck, 2005 and references cited therein) as well as lexical frequency and probability of occurrence (e.g., Penolazzi et al., 2007, and references cited therein). The P2 typically onsets 150 to 220 msec after stimulus presentation at frontal and central scalp sites. It is known to be modulated by the amount of attention directed at features of an event as well as stimulus probability, expectancy, and frequency (e.g., Luck, 2005; Dambacher, Kliegl, Hofmann, and Jacobs, 2006; Wlotko and Federmeier, 2007).

Against the backdrop of the word-frequency effect, we expect that lower frequency sequences will elicit larger P1, N1, and P2 deflections. Furthermore, we anticipate these early components to be followed by a slow wave at frontal sites known as the slow anterior negativity, which onsets ~ 250 msec poststimulus, peaks ~ 400 msec, and lasts until ~ 500 msec. This wave is thought to reflect short-term memory processes (e.g., Kluender and Kutas, 1993). Given that lower-frequency sequences are expected to attract more attentional resources than higher-frequency items and therefore be recalled more readily, the amount of resources devoted to

short-term memory processes indexed by slow anterior negativity amplitudes is expected to decrease as whole-string frequency increases.

Participants

Eleven female students from the University of Alberta were paid for their participation in the experiment. (Mean age = 23.4; SD = 1.6; Min = 22; Max = 27). All were native speakers of English. The Research Ethics Board approved the study. Participants gave informed consent after the nature of the study was explained to them. They were asked to fill out the Edinburgh Inventory handedness questionnaire (Oldfield, 1971). The questionnaire was presented on a PC using E-Prime (a stimulus presentation software). Ten were right-handed (Mean handedness score = 79.5/100; SD = 15.8; Min = 54.5/100; Max = 100/100) and one was left-handed (handedness score = -47.4/100). We also assessed participants' reading span and working memory capacity (henceforth WMC) using an adaptation of the Salthouse and Babcock (1991) test (Mean WMC score = 73.3/100; SD = 10.4; Min = 53.6/100; Max = 87.5/100). The WMC test items were presented on a PC using E-Prime.

Materials

The stimuli list consisted of 432 four-word sequences taken from the *British National Corpus*. Frequencies, obtained from the *Variations in English Words and Phrases* search engine, ranged from 0.03 to 105 occurrences per million.

Experimental Design and Procedure

Participants first completed a practice block, which consisted of six trials. In each trial, six three-word sequences were presented in a random order (for a total of 36 practice items). At the end of

each trial, participants were asked to recall as many sequences as possible. The experimental portion consisted of 72 blocks. Each block was divided into 18 trials, where, in each trial, six four-word sequences were randomly presented. A trial looked like the following: Participants first saw the word "Ready …" for 2,500 msec (font: Courier New; size: 18; position: Center), then a fixation cross "+", which was uniformly presented for 250 to 1,000 msec (font: Times New Roman; size: 16; position: Center), then a blank screen for 1,500 msec, followed by the first of six four-word sequences presented all at once for 1,500 msec (font: Times New Roman; size: 14; position: Center), followed by a fixation cross (as previously described) and the second of six sequences (as previously detailed), and so on until six four-word sequences were shown. At the end of each trial, participants were prompted to type in as many sequences as they could recall. Participants had three two-minute breaks. Sequences subtended on average ~ 5º x 0.4º visual angle; the longest four-word string (*becoming increasingly clear that*) subtended ~ 8º x 0.4º visual angle.

Behavioural Analysis and Results

While examining the data, we realized that one item was a three-word sequence and another one appeared twice in the list; they were thus removed leaving us with 430 items. The remaining data were analyzed using linear mixed-effects regression (LMER; Baayen, 2007; Baayen, Davidson, and Bates, 2008). Our main interest here was to determine whether the number of times a sequence would be correctly recalled varied as a function of whole-string frequency/probability. Responses were coded as "correctly recalled" or "incorrectly recalled". In order to be correctly recalled, the sequence had to be recalled exactly. That is, if the target sequence was *in the middle of*, any response other than *in the middle of* was considered to be incorrect such as for instance *in the middle*, *in the middle and*, *in the middle of a*, or *at the middle of*. We did accept, however,

minor misspelling such as *in the mdle of* or *n the midle of*. Given that whole-string frequency and probability correlate with a number of variables such as for instance a sequence's length, the frequencies of the words that compose it, as well as sequence-internal bigram and trigram frequencies and probabilities, we considered in addition to whole-string frequency and probability a number of variables (fixed effects), which are listed and briefly described in Table 1.

[Insert Table 1 about here]

This would ensure that other potential sources of variation in recall would be controlled for and confirm that a significant whole-string frequency/probability effect, if it were found, would be independent of confounded variables.

Subjects and items were entered in the model as random effects. The most parsimonious and generalizable model consisted of WMC, Position, FreqABC, FreqBCD, PhraseABCD*FreqC, PhraseABCD*FreqD, and PhraseABCD*LogitABCD. Collinearity between model variables was acceptable, that is, there was no significant overlap in predictive power between model variables. Results of the linear mixed-effects regression are summarized in Table 2.

[Insert Table 2 about here]

Figure 1 illustrates the effects of each predictor on probability of recall. Note that the modulation of each variable is *independent* of other model predictors and *additive*. That is, the probability of

recall of an item in this particular case is equal to the sum of the effects of WMC, Position, FreqABC, FreqBCD, PhraseABCD*FreqC, PhraseABCD*FreqD, and PhraseABCD*LogitABCD. Given space constraints, we will only discuss results regarding the PhraseABCD and LogitABCD variables, which are the two variables of main interest.

[insert Figure 1 about here]

Previous studies uncovered a positive correlation between number of words recalled and the amount of linguistic structure existing between them (e.g., Miller and Selfridge, 1950; Tulving and Patkau, 1962). It was thus expected that, in general, phrasal four-word sequences such as *in the United States* would be recalled more readily than non-phrasal strings such as *by the end of*. We believe this is due to the fact that phrases instantiate (relatively) complete concepts compared to non-phrases.

The finding that higher whole-string probability (LogitABCD) facilitate recall is contrary to expectations. Indeed, it was predicted that lower frequency/probability sequences would have been more readily recalled, as was found elsewhere for words in mixed-frequency lists (e.g., DeLosh and McDaniel, 1996; Merrit et al., 2006; Tse and Altarriba, 2007). If more salient items are more easily recalled, then saliency, in the case of regular multi-word sequences, appears to be related to lexical activation rather than to novelty: Lower activation thresholds and/or higher levels of activation relate to higher multi-word string saliency, which in turn is associated with higher probability of recall. While token frequency provides an indication of an item's salience relative to all other items in a language, whole-string probability offers an indication of its salience relative to its "family". The following will clarify this notion.

Let us first restate the equation used to calculate the LogitABCD value of a four-word sequence.

(1) $LogitABCD = log(FreqABCD/(FreqABC* – FreqABCD)+1))$

That is, LogitABCD is equal to the frequency of the whole string divided by the sum of the frequencies of every four-word sequence that share the first three words minus the frequency of that string. By way of example, let us consider the sequence *in the middle of*, which has a token frequency of 28.46 occurrences per million in the *British National Corpus*. There are 243 other sequences that begin with the words *in the middle*, which we refer to as a sequence's "family". Some examples are given in (2), where whole-string frequencies and their respective rankings relative to other members of the family appear in parentheses (*in the middle of* is the most frequent sequence and thus ranked 1).

(2) a. *in the Middle East* (frequency = 4.99; rank = 2)

b. *in the Middle Ages* (frequency = 2.2; rank = 4)

c. *in the middle and* (frequency = 1.23; rank = 6)

d. *in the middle to* (frequency = 0.2; rank = 9)

e. *in the middle are* (frequency = 0.12; rank = 17)

If we only consider the part of the equation that provides the actual ratio between the frequency of *in the middle of* and the summed frequencies of all other sequences of its family, that is, *FreqABCD/(FreqABC* – FreqABCD)+1)*, we obtain the value 28.46/((47.27 – 28.46)+1) =

28.46/19.81 = 1.44, which points to the fact that *in the middle of* stands out from other sequences in the family (in fact, it is the most salient one). Indeed, a ratio greater than 1 indicates that a sequence is more frequent than (most) other members, whereas a ratio smaller than 1 means that it is less frequent. Compare *in the middle of* to *in the middle and*, which has a ratio of 1.23/((47.27 – 1.23)+1) = 0.03 or with *in the middle portion*, which has a ratio of 0.01/((47.27 - 0.01)+1) = 0.0002.

To summarize, the fact that sequence-internal trigrams and single words modulate recall in addition to whole-string probability of occurrence suggests that four-word sequences are both stored as wholes *and* as parts. These results are exactly in line with usage-based accounts of grammar (e.g., Goldberg, 1995; Bod et al., 2003) according to which regular multi-word sequences as wholes leave memory traces in the brain (whatever the definition of the term "memory trace"). The behavioral results, however, are silent as to what type of memory trace might be left behind. Are they procedural or declarative memory traces? In other words, are four-word sequences put together on-line or retrieved as parts and wholes? Because of its high temporal resolution (to the millisecond), electroencephalography is the perfect tool to distinguish between fast computation and holistic retrieval. If it turns out that whole-string probability affects early ERP components such as the P1, the N1, and the P2 deflections, one could argue for holistic retrieval. Indeed, it is believed that words are accessed within 200 msec of presentation (e.g., Sereno, Rayner, and Posner, 1998) irrespective of whether they appear in or out of the context of a sentence. It would thus be impossible to retrieve four words, let alone perform the necessary computations to integrate them, within 200-250 msec.

EEG Recordings and Processing

Electroencephalogram (EEG) recordings were made with Ag/AgCl active electrodes from 32 locations according to the international 10/20 system (www.biosemi.com/headcap.htm) at the midline (Fz, Cz, Pz, Oz) and left and right hemisphere (Fp1, Fp2, AF3, AF4, F3, F4, F7, F8, FC1, FC2, FC5, FC6, C3, C4, T7, T8, CP1, CP2, CP5, CP6, P3, P4, P7, P8, PO3, PO4, O1, O2). Electrodes were mounted on a nylon cap. Additional electrodes were placed at the left and right mastoids, which served as off-line re-reference. Eye movements were monitored by electrodes placed above and below the left eye and at the outer canthi of both eyes, which were bipolarized off-line to yield vertical (VEOG) and horizontal (HEOG) electrooculograms. Analogue signals were sampled at 8,192 Hz using a BioSemi (Amsterdam, Netherlands) Active II digital 24 bits amplification system with an active input range of –262 mV to +262 mV per bit and were band-pass filtered between 0.01 and 100 Hz. The digitized EEG was initially processed off-line using Analyzer version 1.05; it was downsampled to 128 Hz, DC detrended 100 msecs before stimulus markers (Henninghausen, Heil, and Rosler, 1993), band-pass filtered from 0.01 to 32 Hz using an inverse discrete wavelets transform (14 levels), and corrected for eye movements and eye blinks using vertical and horizontal EOGs (Gratton, Coles, and Donchin, 1983). The processed signal was then segmented into epochs of 3,000 msec (1,500 before stimulus onset and 1,500 msec after). Each epoch was baseline corrected on the 1,500 msec segment immediately preceding stimulus onset using the baseline correction option of the inverse discrete Haar-Dauberchies 2 wavelet transform in Analyzer. This was done in order to obtain brain activity measures for each item that, as much as possible, would be uncontaminated by activity from previously presented segments. The data were then exported for further processing and analysis

in R version 2.7.2. Data points exceeding ±100 μV at any channel were excluded from the analysis. We further inspected the data by drawing voltage density and quantile-quantile plots for each channel of each subject; channels showing a significant departure from the normal distribution and failing to reach a peak voltage density of 0.035 were removed.[1] Overall, we discarded 10.4% of our data (2,752,128 over 26,419,200 data points).

Electrophysiological Analysis and Results

We used the generalized additive modeling approach (henceforth GAM) to analyze the ERP data (see Wood, 2006; see Baayen, Hendrix, and Tremblay, 2008, for an application of GAM to ERP data analysis). In essence, GAM determines a linear and/or non-linear equation that strikes a balance between overfitting and overgeneralizing a set of data through a process called penalized iteratively re-weighted least squares (see Wood, 2006, for details). The main advantages of using the GAM method for ERP analysis over the traditional ERP averaging method are as follows: (i) the possibility to fully appreciate the effects of graded variables, such as frequency; (ii) the potential to identify non-linear effects; (iii) the ability to estimate longitudinal effects in the data; and (iv) the power to determine a predictive model of brain activity. In short, this data analysis technique affords the opportunity both to conceive of and investigate research questions previously unthought of or dismissed as untestable.

Though we could have included the sole left-handed participant in the ERP analysis, we restricted our analysis to the ten right-handed participants in order to reduce variability between subjects and increase statistical power. Indeed, it is well known that the structure of the brain in right-handed people differs from that of left-handed people. It is thus very probable that brain potentials elicited from right- and left-handed participants would vary quite significantly in terms of voltage distribution across the scalp.

Given that our main interest in this study is to determine whether whole-string frequency and/or probability affects the retrieval and processing of regular multi-word sequences, and that we do not know exactly what stimuli elicited the event-related potentials recorded to incorrectly recalled sequences, we decided to restrict the ERP analysis to correctly recalled sequences only (i.e., 32.3% of our processed data, which represents 7,635,149 over 23,667,072 data points). By doing so, the ERP component of the study becomes one of lexical access/processing and ceases being one of memory. We relegate the comparison of ERPs to both types of responses to future work.

In the ERP analysis, we used only those variables that reached significance in the behavioral analysis. Baseline corrected epochs were segmented into seven 250 msec windows overlapping 50 msec at edges (mostly because models are not as robust at the edges). For each time window we averaged over subjects. Using GAM, we also removed main time trends and variability due to individual items, Time*FreqC*PhraseABCD, Time*FreqD*PhraseABCD, Time*FreqABC, and Time*FreqBCD. We subsequently assessed, again using GAM, whether the remaining voltage variability was modulated by Time*LogitABCD*PhraseABCD. Main voltage trends (in mcirovolts) for the first time windows are shown in Figure 2.

[insert Figure 2 about here]

We will not be concerned with the other six time windows here given space limitations and given that our analysis focuses on very early ERP components. Each panel represents an electrode: Fp1 and Fp2 are at the top (i.e., the front of the head) and O1, Oz, and O2 at the bottom (i.e., the back of the head). The *x*- and *y*-axes represent time in milliseconds and (baseline-corrected)

microvolts respectively; positive is plotted up. Red dots are scalp voltages averaged over items; blue lines correspond to fitted curves for time obtained from the GAM analysis.

Figure 3 depicts main effects of Time*LogitABCD on scalp voltages and Figure 4 Time:LogitABCD:PhraseABCD (phrase) interactions. As in Figure 2, each panel represents an electrode: Fp1 and Fp2 are at the top of the plots (the front of the head) and O1, Oz, and O2 are at the bottom (the back of the head). The name of the electrodes appears at the very top of the panel. The *x-axis* represents time (msec); the first vertical dashed line represents the 50 msec time point and the following two broken lines the 100 and 200 msec time points. The *y-axis* represents LogitABCD values (log probability of occurrence) from very small at the bottom of the panel (≈ -6) to very high at the top of the panel (≈ 4). The *z-axis* encodes scalp voltages in microvolts, which are represented by both little red contour lines and colors. Voltage values are indicated in red on the contour lines and are also given via color-coding: The hotter the color (yellow), the more positive the voltage and similarly, the colder the color (blue), the more negative the voltage; various shades of green represent voltages around 0. These voltage maps are very similar to topographic maps, where the height of a mountain or the depth a valley is indicated by values appearing on the lines that form them and their steepness by the amount of space separating those lines (the closer the lines, the steeper the incline/decline). At the bottom of each panel *p*-values are provided; significant effects are marked by an asterisk (Bonferroni corrected significance threshold = 0.05 / 32 electrodes / 7 time windows = 0.00022).

[insert Figure 3 about here]

[insert Figure 4 about here]

A significant Time*LogitABCD main effect (after Bonferroni correction) was found at electrode FC1 (0-250 msec time window; $p = 0.00003$), and significant Time*LogitABCD*PhraseABCD (phrase) interactions at electrodes P3 and P7 (0-250 msec time window; $p = 0.00005$ and 0.00007 respectively). We believe the effects found at these sites are real given that other electrodes in their vicinity also recorded the same electrical pattern (though they did not reach significance). We did not find any significant LogitABCD modulations on either the P2 or the slow anterior negativity. We discuss the electrophysiological results in the following section.

Early fronto-central negativity (N1a)

A significant Time*LogitABCD main effect was found in the 0-250 msec time window at electrode FC1. In order to interpret this effect, it is necessary to consider it in the context of its associated Time smooth (i.e., the blue line in Figure 2, electrode FC1). Indeed, Figure 3 merely illustrates the manner in which LogitABCD (for phrases and non-phrases alike) modulates the electroencephalogram (EEG) in this time window, that is, that the N1 component is more positive for lower probability sequences and more negative for higher probability ones. To observe the actual N1, it is necessary to add the Time*LogitABCD curve to the Time curve (hence the term "additive" in "generalized additive model"). This is shown in Figure 5. Note that the bottom panel of Figure 5 is merely intended to give an approximate representation of what the actual EEG at this time window looks like.

[insert Figure 5 about here]

In Figures 5, it can be observed that the N1 component increases in amplitude as the probability of occurrence of a regular four-word sequence increases. Given that stimulus characteristics such as length are known to affect N1 amplitudes, it is possible that the modulations we observe in our data are attributable to Length rather than to LogitABCD. Note that we did not include length from start of the ERP analysis because this variable did not reach significance in the behavioral analysis (recall that we only considered those variables that significantly accounted for variability in the behavioral data, which did not include Length).[2] We thus examined whether the LogitABCD effect would survive the addition of Length to the model. We first took out from the total variance that portion explained by individual items, Time, Time*FreqC*PhraseABCD, Time*FreqD*PhraseABCD, Time*FreqABC, Time*FreqBCD, *and* Time*Length*PhraseABCD and then fitted another model on the remaining variance where Time*LogitABCD*PhraseABCD was entered as the only predictor. The correlation existing between Length and LogitABCD is very small (r = -0.1). It was thus probable that the LogitABCD effect would remain even after removing the variability due to Length. Neither the Time*Length main effect nor the Time:Length:PhraseABCD (phrase) interaction reached significance (p = 0.9072 and 0.0328 respectively). As expected, the Time*LogitABCD main effect survived the addition of Length to the model (p = 0.00001; $\alpha_{Bonferroni}$ = 0.00022).

As mentioned in the introduction, the anterior N1 is thought to originate from centro-parietal sources including areas in and adjacent to the intraparietal sulcus (Di Russo et al., 2002). Mevorach, Shalev, Allen, and Humphreys (in press) report higher activation levels along the left intraparietal sulcus (IPS) for low relative to high saliency target stimuli. If, according to classic theories of information processing (e.g., Johnston and Heinz, 1978), novel stimuli are more salient (in this case lower probability sequences), it is conceivable that less salient higher

probability sequences elicited higher activation along the left IPS and concomitantly higher N1 amplitudes at anterior scalp sites, which ultimately lead to better recall. This hypothesis fits well with the unexpected outcome that higher rather than lower probability four-word strings were more readily recalled. Given the present findings, one could argue that memory traces associated with at least some aspects of regular four-word sequences are present in the centro-parietal pathway.

Early parietal positivity (P1)

We now turn to the Time:LogitABCD:PhraseABCD (phrase) interaction found at electrodes P3 and P7 in the 0-250 time window. Figure 6 depicts how P1 amplitudes vary as a function of time and the probability of occurrence of *phrasal* four-word sequences (electrode P7 shown).

[insert Figure 6 about here]

Stimulus characteristics such as length are also known to affect P1 deflection. We thus assessed whether the addition of Length to the model would remove the Time:LogitABCD:PhraseABCD (phrase) effect. Neither the Time*Length main effect (P3: $p = 0.00085$; P7: $p = 0.961$) nor the Time:Length:PhraseABCD (phrase) interaction (P3: $p = 0.01293$; P7: $p = 0.755$) reached significance. The Time:LogitABCD:PhraseABCD (phrase) interaction is robust to the addition of Length to the model at both electrodes (P3: $p = 0.00016$; P7: $p = 0.00018$; $\alpha = 0.00022$).
As mentioned in the introduction, the early P1 (peak ~ 98-110 msec) is believed to originate from the dorsal extra-striate cortex of the middle occipital gyrus and the late P1 (peak ~136-146 msec) from the ventral extra-striate cortex of the fusiform gyrus (Di Russo et al., 2002). Functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) studies

have reported activation of this complex during word, object, and face presentations, which diminished with repeated presentations (Rossion, Schiltz, Robaye, Pirenne, and Crommelinck, 2001, and references cited therein). These observations are generally attributed to "better (or faster) performance at processing these stimuli, thus indicating the neural correlates of perceptual priming or implicit memory processing … In other words, these deactivations reflect a facilitation in neural computations when the same information is processed again" (Rossion et al., 2001, p. 1027). Given these findings, it is conceivable that the P1 amplitudes observed in the present study, which decrease as the probability of phrasal four-word sequences increase, reflect the level of entrenchment of at least some aspects of these items in the occipito-temporal pathway.

Conclusion

We investigated the processing of regular four-word sequences from both a behavioral and an electrophysiological perspective. The fact that whole-string probability as well as sequence-internal word and trigram frequency affected recall suggests that multi-word strings are stored both as parts and wholes. Furthermore, frequency/probability was found to modulate recall and event-related potentials in a continuous rather than categorical manner, thus indicating that lexical bundles and non-lexical bundles are best viewed as two extremes of a "whole-string frequency/probability" continuum.

It was unclear form the behavioral results whether the whole-string probability effect reflected fast computation or holistic retrieval. The electrophysiological results provided evidence to the effect that four-word sequences are retrieved in a holistic manner (whatever the definition of the term holistic) rather than computed on-line via rule-like processes. Indeed, the

fact that whole-string probability modulated P1 and N1 amplitudes ~ 110-150 msec after stimulus onset strongly advocates for this deduction. If the earliest frequency/probability effect on event-related potentials to single word processing is reported to be ~ 110 msec (e.g., Penolazzi et al., 2007), it is most unlikely that *four words* can be accessed, let alone stringed together, within this time frame.

Owing to previous research that focused on the identification of P1 and N1 generators (e.g., Di Russo et al., 2002), we are in a position to put forth the hypothesis that at least some aspects of non-phrasal and phrasal four-word sequences leave memory traces in the centro-parietal pathway and that *phrasal* multi-word strings leave additional ones in the occipito-parietal pathway. These results are exactly in line with usage-based accounts of grammar (e.g., Goldberg, 1995; Bod et al., 2003, Bybee and McClelland, 2005, McClelland and Bybee, 2007).

References

Azizian, A., and Polich, J.(2007). Evidence for attentional gradient in the serial position memory curve from event-related potentials. *Journal of Cognitive Neuroscience, 19*, 2071-2081.

Baayen, R.H. (2007). *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Baayen, R.H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. To appear in *Journal of Memory and Language, special issue on Emerging Data Analysis Techniques*.

Baayen, R.H., Hendrix, P., & Tremblay, A. (2008). Generalized additive modeling: An application to event-related brain potential data from word naming and free recall tasks. Manuscript preparation.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Grammar of Spoken and Written English*. Harlow: Longman.

Bod, R. (2001). Sentence Memory: Storage vs. Computation of Frequent Sentences. Abstract retrieved on October 29, 2006, from http://staff.science.uva.nl/~rens/cuny2001.pdf.

Bod, R. Scha, R., & Sima'an, K. (2003). *Data-oriented Parsing*. Stanford, CA: Studies in Computational Linguistics.

Bower, G.H. (2000). A brief history of memory research. In E. Tulving & F.I.M Craik (Eds.), *The Oxford Handbook of Memory* (pp. 3-32). Oxford: Oxford University Press.

Bybee, J. & McClelland, J.L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review, 22*, 381-410.

Chomsky, N. (1988). *Language and Problems of Knowledge*. Cambridge, MA: MIT Press.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than Nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*, 72-89.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A.M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research, 1084*, 89-103.

DeLosh, E.L., & McDaniel, M.A. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1136-1146.

Di Russo, F., Martinez, A., Sereno, M.I., Pitzalis, S., & Hillyard, S.A. (2002). Cortical sources of the early components of the visual evoked potential. *Human Brain Mapping, 15*, 95-111.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: The University of Chicago Press.

Gratton, G., Coles, M.G.H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55*, 468-484.

Hennighausen, E., Heil, M., & Rosler, F. (1993). A correction method for DC drift artifacts. *Electroencephalography and Clinical Neurophysiology, 86*, 199-204.

Hillyard, S.A., Teder-Sälejärvi, W.A., & Münte, T.F. (1998). Temporal dynamics of early perceptual processing. *Current Opinion in Neurobiology, 8*, 202-210.

Jiang, N., & Nekrasova, T.M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, *91*, 433-445.

Johnston, W.A., & Heinz, S.P. (1978). Flexibility and capacity demands of attention. *Journal of Experimental Psychology, 107*, 420-435.

Kluender, R., & Kutas, M. (1993). Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience*, *5*, 196–214.

Luck, S.J. (2005). *An Introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT Press.

McClelland, J.L. & Bybee, J. (2007). Gradience of gradience: A reply to Jackendoff. *The Linguistic Review, 24*, 437-455.

Merovach, C., Shalev, L., Allen, H.A., & Humphreys, G.W. (in press). The left intraparietal sulcus modulates the selection of low salient stimuli. *Journal of Cognitive Neuroscience*.

Merritt, P.S., DeLosh, E.L., & McDaniel, M.A. (2006). Effects of word frequency on individual-item and serial order retention: Tests of the order-encoding view. *Memory and Cognition, 34*, 1615-1627.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, *63*, 81-97.

Miller, G.A., & Selfridge, J.A. (1950). Verbal context and the recall of meaningful material. *The American Journal of Psychology*, *63*, 176-85.

Oldfield, R.C. (1971). The assessment and analysis of handedness: The Edinburgh Inventory. *Neuropsychologica*, 9, 97-113.

Penolazzi, B., Hauk, O., & Pulvermüller, F. (2007). Early semantic context integration and lexical access as revealed by event-related brain potentials. *Biological Psychology, 74*, 373-388.

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org.

Rossion, B., Schiltz, C., Robaye, L., Pirenne, D., & Crommelinck, M. (2001). How does the brain discriminate familiar and unfamiliar faces?: A PET study of face categorical perception. *Journal of Cognitive Neuroscience, 13*, 1019-1034.

Salthouse, T.A., & Babcock, R.L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology, 27*, 763-776.

Schmitt, N., & Underwood, G. (2004). Exploring the processing of formulaic sequences through a self-paced reading task. In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 171-89). Amsterdam and Philadelphia: John Benjamins.

Sereno, S.C., Rayner, K., & Posner, M.I. (1998). Establishing the time-line of word recognition: evidence from eye movements and event-related potentials. *NeuroReport, 9*, 2195-2200.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2008). Processing Advantages of Lexical Bundles: Evidence form Self-paced Reading Experiments, Word and Sentence Recall tasks, and Off-Line Semantic Ratings. Manuscript submitted for publication.

Tse, C.-S., & Altarriba, J. (2007). Testing the associative-link hypothesis in immediate serial recall: Evidence from word frequency and word imageability effects. *Memory, 15*, 675-690.

Tulving, E., & Patkau, J.E. (1962). Concurrent effects of contextual constraint and word frequency on immediate recall and learning of verbal material. *Canadian Journal of Psychology*, *16*, 83-95.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 153-172). Amsterdam and Philadelphia: John Benjamins.

Wlotko, E.W., & Federmeier, K.D. (2007). Finding the right word: Hemispheric asymmetries in the use of sentence context information. *Neuropsychologia, 45*, 3001-3014.

Wood, S.N. (2006). *Generalized Additive Models. An Introduction with R*. Boca Raton, FL: Chapmans and Hall/CRC.

Footnotes

[1] A peak voltage density under 0.035 means that voltages are too widely distributed around the channel's peak density (usually ~ 0 μV). In other words, the channel is too noisy. This threshold value was obtained by visually comparing EEG epochs and their voltage density plots.

[2] Given that recalling an item and encoding it in short-term memory are two different cognitive processes (with some possible overlap), strictly speaking the ERP data should have been

modeled independently; variables that did not significantly account for variability in the behavioral data might have done so in the ERP data.

Tables

Table 1

Variables Taken into Consideration in the Statistical Analysis

Variable	Description
WMC	Reading span and working memory capacity score.
EI	Handedness score.
Trial	The block in which an item was presented (out of 72 blocks).
Position	Position of an item within a trial (either 1, 2, 3, 4, 5, or 6).
Length	Length of the whole sequence in number of letters.
PhraseABCD	Whether the whole sequence was a phrase (e.g.,), or a non-phrase (e.g., *I think it's the*). Phrases are sequences that can stand alone such as *in the United States*, *they don't have to*, *she was going to*, and *he shook his head*, while non-phrases such as *but there is no*, *the result of a*, and *I don't think it's* cannot.
WordTypeABCD	Patterns of content (Con) and non-content (N) words. For example, *in the middle of* has the structure NNConN.
FreqA, FreqB, FreqC, FreqD	Frequency of the first, second, third, and forth word of the

	sequence. Considering the sequence *in the middle of*, FreqA = frequency of *in*, FreqB = frequency of *the*, FreqC = frequency of *middle*, and FreqD = frequency of *of*.
FreqAB, FreqBC, FreqCD	Frequency of the sequence formed by the first and second word (FreqAB), the second and third word (FreqBC), and the third and forth word (FreqCD).
FreqABC, FreqBCD	Frequency of the sequence formed by the first, second, and third word (FreqABC) and second, third, and fourth word (FreqBCD) of a sequence.
FreqABCD	Frequency of the whole sequence (e.g., *in the middle of*).
LogitAB, LogitBC, LogitCD	The (log) probability of obtaining word B, C, or D given word A, B, or C respectively. For example, *LogitAB = log(FreqAB/((FreqA* – FreqAB)+1))*.
LogitABC, LogitBCD	The (log) probability of obtaining word C or D given the sequence AB or BC, respectively.
LogitABCD	The (log) probability of obtaining word D given the sequence ABC.

Note. The capital letters A, B, C, and D refer to words in the first, second, third, and forth position of a four-word sequence (e.g., *in the middle of* where A = *in*, B = *the*, C = *middle*, and D = *of*). The asterisk * is a wildcard representing any single word; if A = *in* then A* could stand for *in the*, *in a*, *in your*, etc. Con stands for "content word" (e.g., *middle*), and N for "non-content word" (e.g., *the*).

Table 2

Linear Mixed-effects Regression Results

Random Effects			
Groups	Name	Variance	*SD*
Subject	(Intercept)	0.0547	0.2339
Item	(Intercept)	0.1663	0.4078
Fixed Effects			
	Estimate	*SE*	*z* value
Intercept	-2.1450	0.6814	-3.1**
WMC	2.6230	0.7737	3.4***
1st restricted cubic spline for Position	-0.2574	0.0555	-4.6***
2nd restricted cubic spline for Position	0.6856	0.0608	11.3***
PhraseABCD (phrases)	0.4798	0.6453	0.7
FreqC	-0.1025	0.0250	-4.1***
FreqD	-0.0189	0.0365	-0.5
FreqABC	0.1137	0.0333	3.4***
FreqBCD	-0.0833	0.0386	-2.2*
LogitABCD	0.1074	0.0394	2.7**
PhraseABCD(phrases) by FreqC	0.1850	0.0538	3.4***

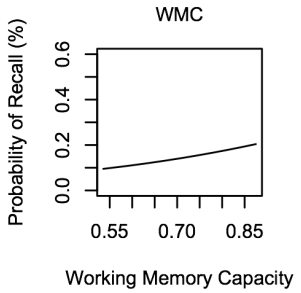
PhraseABCD(phrases) by FreqD	-0.1338	0.0612	-2.2*
PhraseABCD (phrases) by LogitABCD	0.2128	0.0698	3.1**

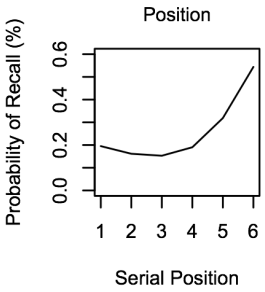
Note. Estimates and standard errors correspond to log probability of recall (i.e., *logit(P) = log(P/(1-P))*). Probabilities (%) are obtained from the following equation: *P = exp(logit(P))/(1+exp(logit(P)))*. Restricted cubic splines (rcs) with three knots were used for Position, indicating that the effect is non-linear. 4,730 observations, where one observation is equal to one four-word sequence correctly recalled or not by one participant. Collinearity index between model variables is 12.6, which is acceptable (15 is considered to be too high).

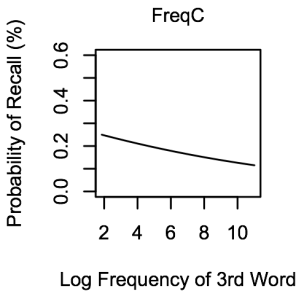
* = $p < .05$. ** = $p < .01$. *** = $p < .001$.

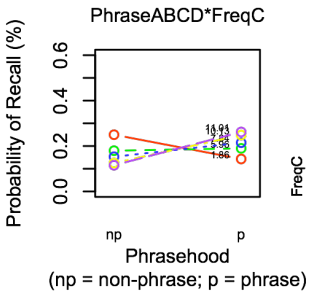
Figures

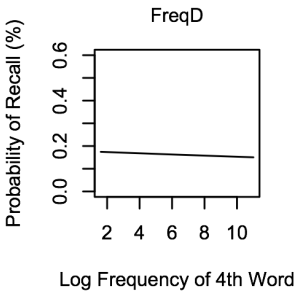
Figure 1. Results of the linear mixed-effects regression analysis. Each panel shows the effect sizes of significant variables on probability of recall. From top left to bottom right: WMC, Position, FreqC, FreqD, FreqABC, FreqBCD, PhraseABCD by FreqC, PhraseABCD by FreqD, and PhraseABCD by LogitABCD. In the fourth, sixth, and tenth panels, each line represents the 1st, 2nd, 3rd, 4th, and 5th quantiles of the FreqC, FreqD, and LogitABCD distributions (i.e., 1.86, 5.96, 7.84, 10.13, and 11.01 for FreqC; 1.61, 7.74, 9.97, 10.27, and 11.01 for FreqD; -9.36, -1.90, -0.85, 0.28, and 3.44 for LogitABCD).

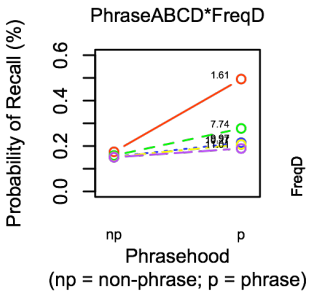

WMC
Probability of Recall (%)
0.0 0.2 0.4 0.6
0.55 0.70 0.85
Working Memory Capacity

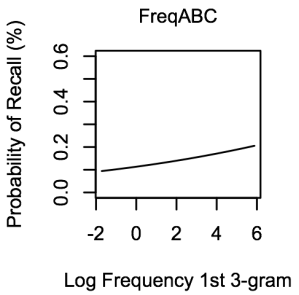

Position
Probability of Recall (%)
0.0 0.2 0.4 0.6
1 2 3 4 5 6
Serial Position

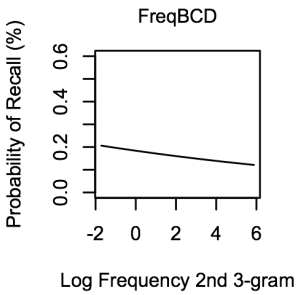

FreqC
Probability of Recall (%)
0.0 0.2 0.4 0.6
2 4 6 8 10
Log Frequency of 3rd Word

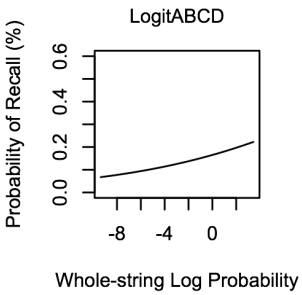

PhraseABCD*FreqC
Probability of Recall (%)
0.0 0.2 0.4 0.6
np
p
Phrasehood
(np = non-phrase; p = phrase)
FreqC


FreqD
Probability of Recall (%)
0.0 0.2 0.4 0.6
2 4 6 8 10
Log Frequency of 4th Word


PhraseABCD*FreqD
Probability of Recall (%)
0.0 0.2 0.4 0.6
1.61
7.74
np
p
Phrasehood
(np = non-phrase; p = phrase)
FreqD


FreqABC
Probability of Recall (%)
0.0 0.2 0.4 0.6
-2 0 2 4 6
Log Frequency 1st 3-gram


FreqBCD
Probability of Recall (%)
0.0 0.2 0.4 0.6
-2 0 2 4 6
Log Frequency 2nd 3-gram


LogitABCD
Probability of Recall (%)
0.0 0.2 0.4 0.6
-8 -4 0
Whole-string Log Probability

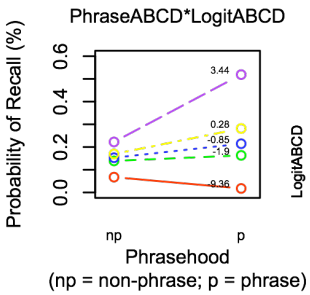

PhraseABCD*LogitABCD
Probability of Recall (%)
0.0 0.2 0.4 0.6
3.44
0.28
-0.85
-1.9
-9.36
np
p
Phrasehood
(np = non-phrase; p = phrase)
LogitABCD

Figure 2. 0 – 250 msec time window.

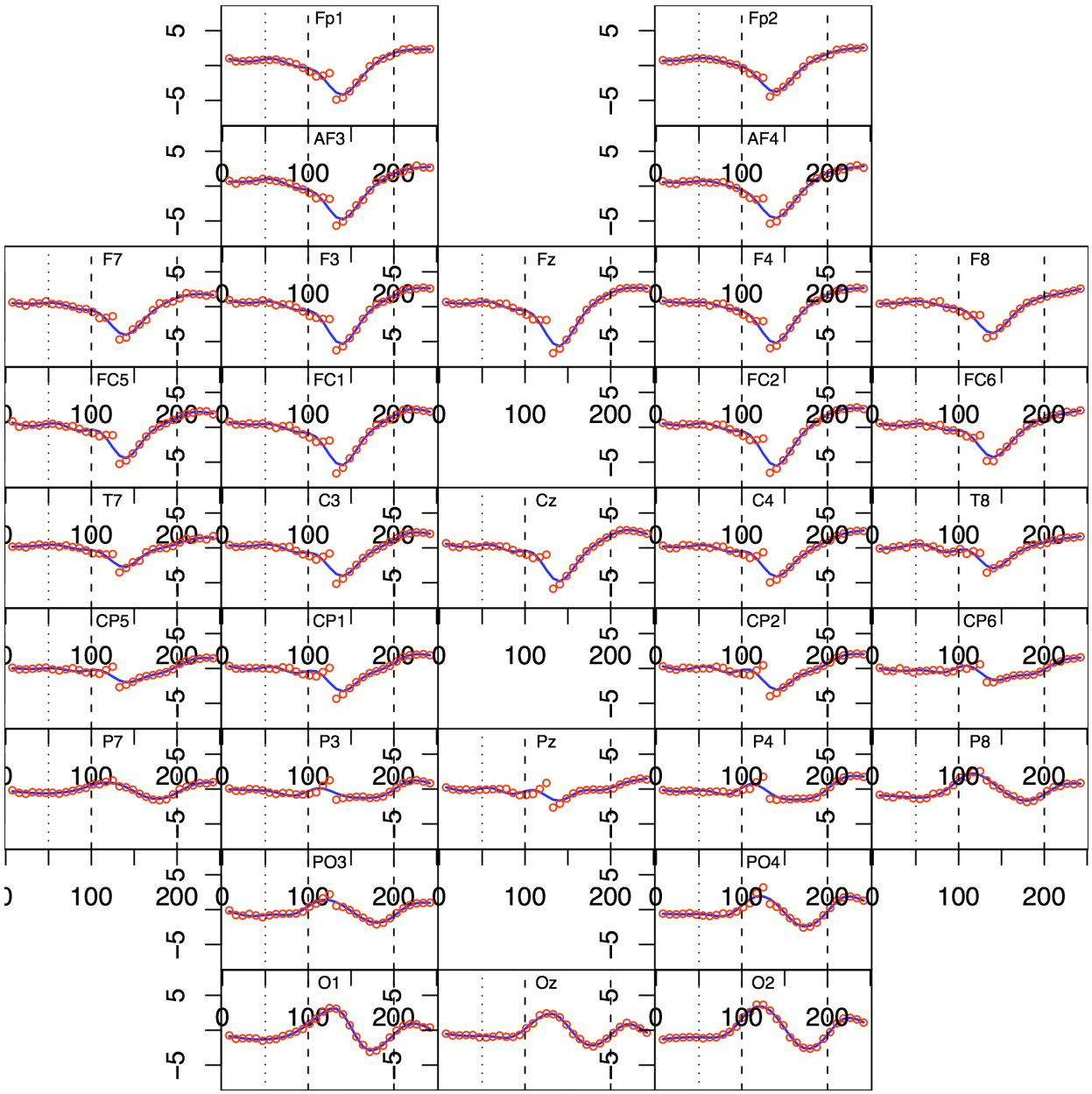


Figure 3. Time*LogitABCD main effect in the 0 – 250 msec time window.

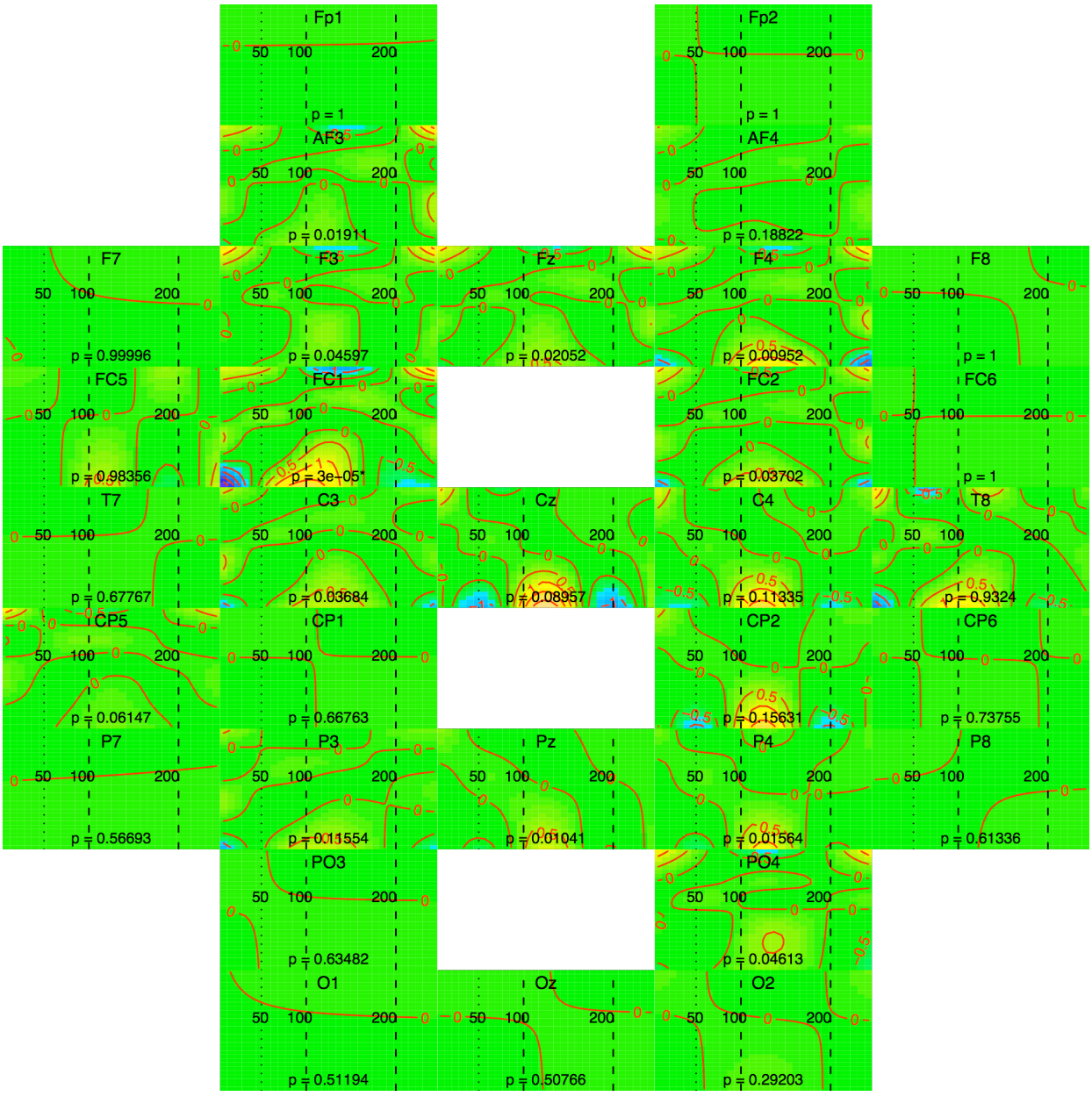


Figure 4. Time:LogitABCD:PhraseABCD (phrase) interaction in the 0 – 250 msec time window.

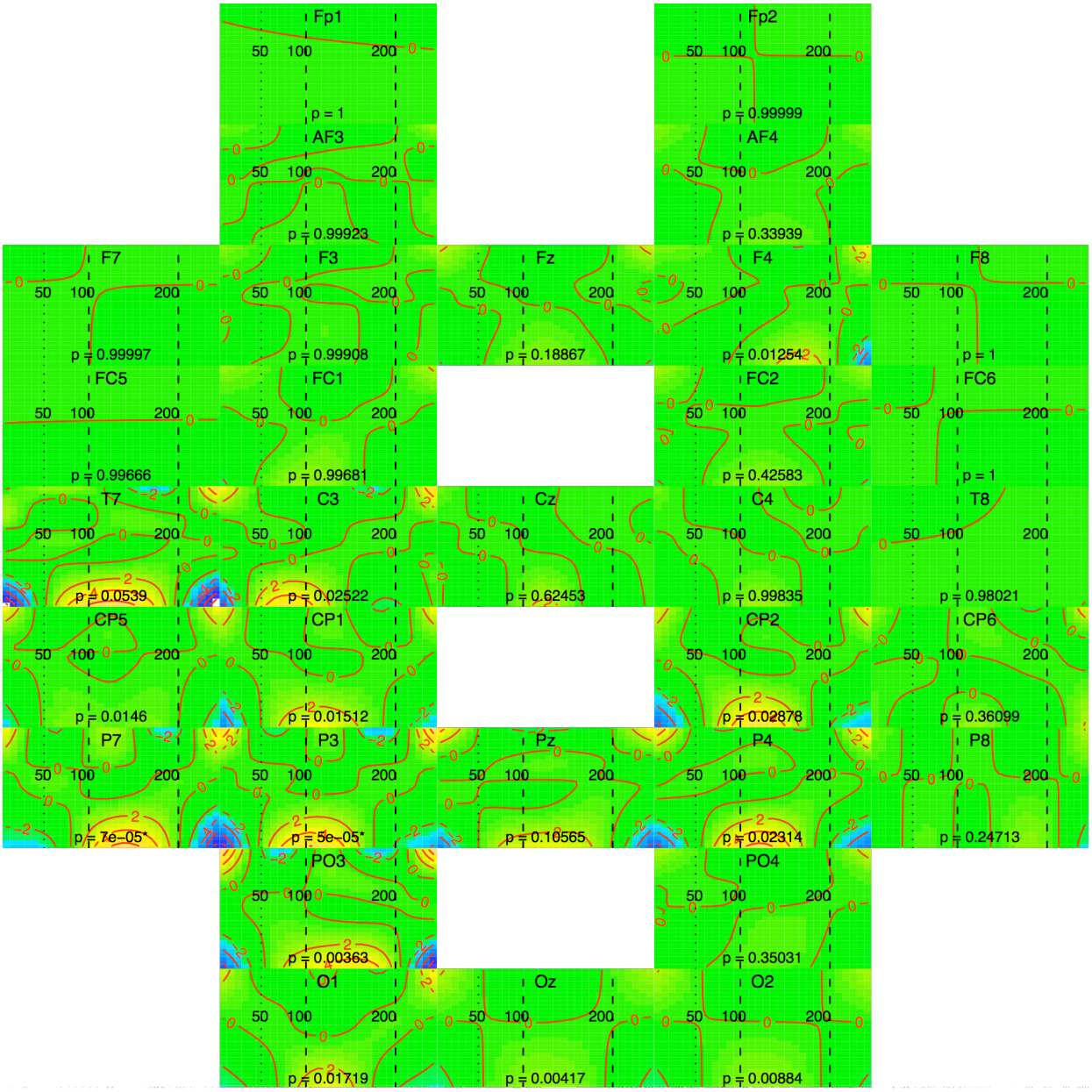


Figure 5. Adding the Time*LogitABCD smooth to the Time smooth at electrode FC1. Top panel: Time smooth (red points: voltages averaged over items; blue line: fitted Time smooth); the *x-axis* is time in msec and the *y-axis* is voltage in μV (positive is plotted up). Middle panel: Time*LogitABCD smooth; the *x-axis* is time in msec; the *y-axis* is LogitABCD; the *z-axis* (contours and colors) is voltage in μV (red colors are positive and blue colors are negative). Bottom panel: The sum of the top two panels.

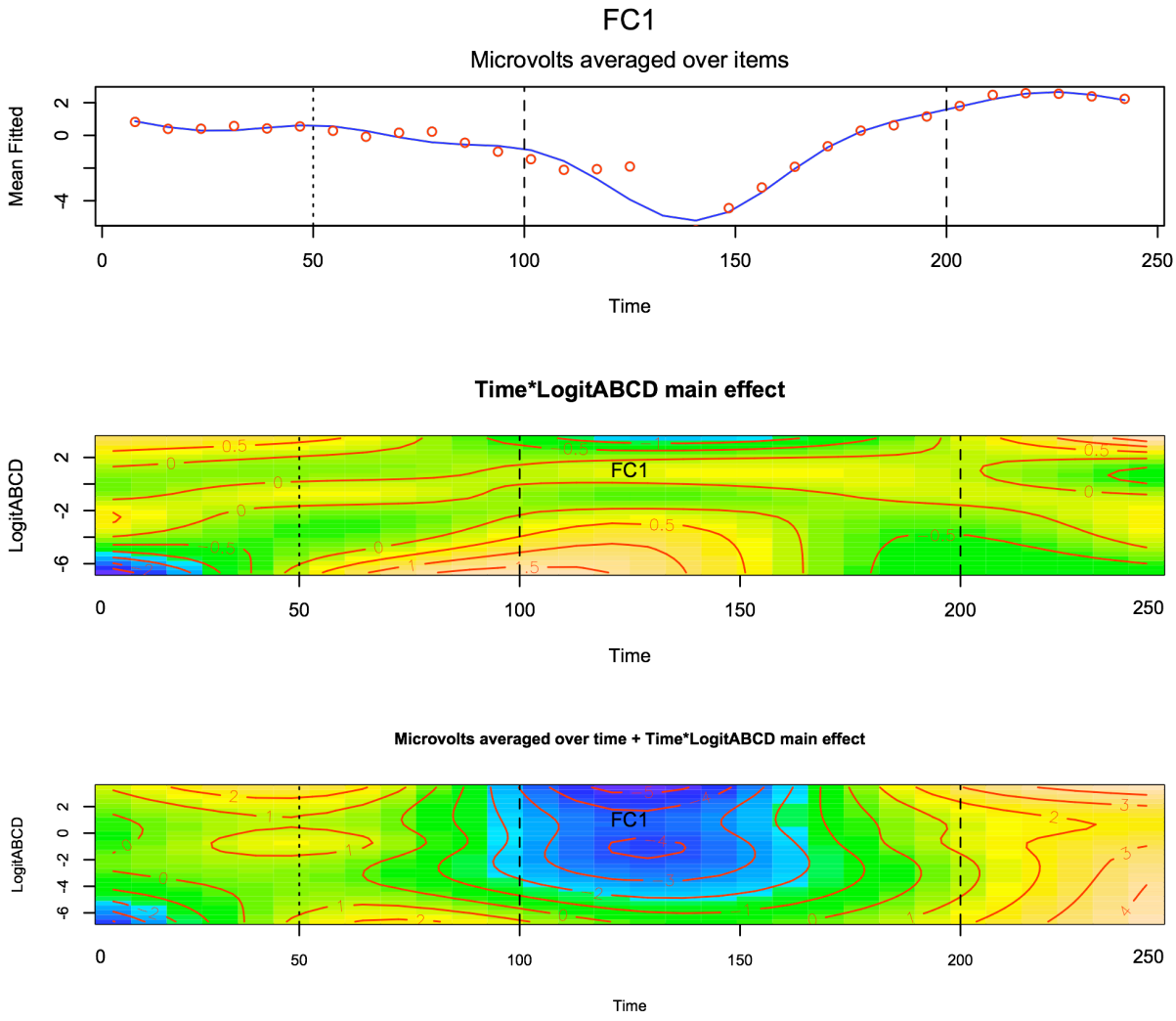


Figure 6. Adding the Time:LogitABCD:PhraseABCD (phrase) smooth to the Time smooth at electrode P7.