Context-dependent Semantic Processing in the Human Brain: Evidence from Idiom Comprehension

Joost Rommers[1], Ton Dijkstra[2], and Marcel Bastiaansen[1,2]

Abstract

■ Language comprehension involves activating word meanings and integrating them with the sentence context. This study examined whether these routines are carried out even when they are theoretically unnecessary, namely, in the case of opaque idiomatic expressions, for which the literal word meanings are unrelated to the overall meaning of the expression. Predictable words in sentences were replaced by a semantically related or unrelated word. In literal sentences, this yielded previously established behavioral and electrophysiological signatures of semantic processing: semantic facilitation in lexical decision, a reduced N400 for semantically related relative to unrelated words, and a power increase in the gamma frequency band that was disrupted by semantic violations. However, the same manipulations in idioms yielded none of these effects. Instead, semantic violations elicited a late positivity in idioms. Moreover, gamma band power was lower in correct idioms than in correct literal sentences. It is argued that the brain's semantic expectancy and literal word meaning integration operations can, to some extent, be "switched off" when the context renders them unnecessary. Furthermore, the results lend support to models of idiom comprehension that involve unitary idiom representations. ■

INTRODUCTION

Most current neurocognitive models of sentence comprehension are compositional in nature, involving two main operations, which continuously interact and overlap in time: (1) incoming words' meanings are accessed and (2) these meanings are incrementally integrated (unified) with the preceding context to build up the message level meaning of the entire sentence (Hagoort, 2005; see also Werning, Hinzen, & Machery, 2011). Although this framework successfully accounts for a wide range of phenomena in the comprehension of literal sentences, little research has investigated whether these two brain operations can explain the processing of all possible linguistic input. For example, language use in everyday conversation contains many idioms (Sprenger, 2003; Jackendoff, 1995), a subclass of fixed expressions in which the meanings of the individual words are often completely unrelated to the meaning of the expression as a whole (e.g., *to spill the beans*, meaning *to let out a secret*).

What happens to accessing literal word meanings during the comprehension of idioms? And does the language system still engage in the process of unifying word meanings while processing idioms? Obviously, analyzing idioms compositionally would only lead to activation of the literal meaning (e.g., spilling actual beans) instead of the relevant figurative meaning of the idiom as a whole. Thus, the operations of activation and unification of literal word meanings are theoretically unnecessary for comprehending idioms. However, it is an empirical question whether or not these seemingly routine operations are actually less engaged during idiom comprehension than during the comprehension of literal sentences. Hence, idioms provide a test bed for investigating the roles of these operations during sentence comprehension. It is possible that whenever we hear or read a word, regardless of the context in which it occurs, we activate its meaning and attempt to unify it with the preceding context. On the other hand, the activation and unification of literal word meanings could be operations that are only carried out when they are really necessary, that is, when comprehending compositional linguistic input. This study aims to answer this question, employing both behavioral and electrophysiological methods.

Activation and Unification of Literal Word Meanings in Idiom Comprehension

The issue of literal word meaning activation has featured prominently in models of idiom comprehension. Some of these models pose no intrinsic limits on the activation of literal word meanings. For instance, the lexical representation hypothesis assumes that the entire idiom is retrieved as a whole, like any lexical item, but literal word meanings are activated because a literal analysis runs along in parallel with the retrieval of the figurative meaning (Swinney & Cutler, 1979). Hybrid models, on the other hand, represent idioms as units—lexical concept nodes

[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, [2]Radboud University Nijmegen, The Netherlands

Journal of Cognitive Neuroscience 25:5, pp. 762–776
doi:10.1162/jocn_a_00337

(Cutting & Bock, 1997) or superlemmas (Sprenger, Levelt, & Kempen, 2006)—that are connected to the representations of the single words they are composed of, thus allowing for activation to spread from the idiom to its literal word meanings.

Other models allow no literal word meaning activation or constrain it to certain situations. For example, the idiom list hypothesis predicts that literal word meanings should not be activated, because idioms are assumed to be stored as lexical items in a list separate from the rest of the lexicon (Bobrow & Bell, 1973). In the direct access hypothesis, literal word meanings are only computed when the idiomatic interpretation does not fit the context (Gibbs, 1980). According to the configuration hypothesis, incoming words are analyzed literally, but only until sufficient information has accumulated for the reader or listener to recognize the idiom as such (Cacciari & Tabossi, 1988). After this point (the "key" of the idiom) words are not analyzed literally anymore.

There is both evidence for and against literal word meaning activation in idiom comprehension. In support of the activation of literal word meanings, it has been shown that words in idioms can prime other words that are related to their literal meaning (e.g., *kick the bucket* primes *pail*; Sprenger et al., 2006; Hillert & Swinney, 2001; Colombo, 1993; Cacciari & Tabossi, 1988; Swinney, 1981). However, other studies have demonstrated a lack of literal word meaning activation in idioms. Peterson, Burgess, Dell, and Eberhard (2001) found that, when participants were asked to complete predictable sentence contexts with a spoken target word, there was a concreteness effect in literal sentences (shorter speech onset latencies for abstract than concrete target words) but not in idioms, suggesting that literal word meanings were not processed. Furthermore, in an fMRI study, somatotopically distributed activation of the motor system was observed for action verbs in isolation (e.g., *grab*/*kick*), in literal sentences (e.g., *The fruit cake was the last one so Claire grabbed it*), but not in idioms in sentences (e.g., *The job offer was a great chance so Claire grabbed it*; Raposo, Moss, Stamatakis, & Tyler, 2009). Motor cortex activation has only been found for action verbs in unpredictable idioms without much contextual support (e.g., *John kicked the bucket*; Boulenger, Hauk, & Pulvermüller, 2009). In summary, models and results on literal word meaning activation are mixed.

To our knowledge, all studies thus far have only looked at the activation of literal word meanings: no study has looked at the issue of semantic unification in idioms, and models of idiom comprehension have not explicitly addressed the issue. Semantic unification refers to the process of integrating word meanings by combining them into larger units (cf. Hagoort, 2005). It is conceivable that if idioms are partly represented as units and retrieved as such from memory, as in the hybrid models discussed previously (Sprenger et al., 2006; Cutting & Bock, 1997), the individual words need not be integrated, and therefore, the unification process should be less engaged in idioms compared with literal language.

One technique with which both word access processes and unification have been studied extensively is the recording of the brain's voltage fluctuations in the EEG. Before outlining this study, we turn to a discussion of the electrophysiological correlates of semantic processing.

Electrophysiological Signatures of Semantic Processing

In language comprehension studies, the EEG signal from multiple trials is most commonly averaged to form ERPs containing several components. The most well-known component is the N400 (Kutas & Hillyard, 1980), which is generally thought to index both word retrieval and integration processes (Kutas & Federmeier, 2011; Lau, Phillips, & Poeppel, 2008). In an experiment of particular relevance to this study, Federmeier and Kutas (1999) presented participants with contexts consisting of two sentences that were predictive of a specific word, such as "*They wanted to make the hotel look more like a tropical resort. So along the driveway they planted rows of…*." There were three conditions. The context ended with the expected word (e.g., *palms*), an anomalous word from the same semantic category as the expected word (e.g., *pines*), or an anomalous word from a different semantic category (e.g., *tulips*). Both anomalies elicited N400 effects relative to the expected word, but the N400 to the words from the same semantic category as the expected word was significantly smaller compared with the words from a different semantic category. In addition, the N400 to words from the expected semantic category was smaller in more highly constraining contexts compared with less constraining contexts, despite their lower rated plausibility in those sentences. From this pattern of results, Federmeier and Kutas concluded that context can lead to specific semantic expectations.

To our knowledge, few published ERP studies have investigated idiom comprehension, and their N400 results have been mixed. Laurent, Denhieres, Passerieux, Iakimova, and Hardy-Bayle (2006) observed smaller N400 amplitudes for the last word of an idiom when its idiomatic meaning was more salient than when it was less salient (as assessed in terms of conventionality, frequency, familiarity, or prototypicality). Proverbio, Crotti, Zani, and Adorni (2009) observed a larger N400 to words in idiomatic than literal phrases. Moreno, Federmeier, and Kutas (2002) observed that, relative to predictable words, "lexical switches" (synonyms) showed larger N400 amplitudes in both literal and idiomatic contexts, suggesting similar semantic processes (see also Liu, Li, Shu, Zhang, & Chen, 2010). Vespignani, Canal, Molinaro, Fonda, and Cacciari (2010) focused on predictions and manipulated words in idioms that were either at the key of the idiom or one position further downstream in the sentence. When the word at the key was replaced with a different word, this led to an N400 effect. One word further downstream in the sentence, a replacement

also led to an N400, but the control condition in which the entire idiom was left intact elicited a P300, suggesting the presence of a special type of predictability in idiom comprehension (cf. Molinaro & Carreiras, 2010).

Besides ERPs, in recent years, an interest has grown in characterizations of the ongoing oscillatory activity in the EEG and MEG signal, mainly by examining power increases and decreases over time in different frequency bands, which reflect increasing and decreasing synchronization (Bastiaansen, Mazaheri, & Jensen, 2012; Bastiaansen & Hagoort, 2006). Semantic integration or unification operations have been found to induce power changes in the gamma frequency band (around 30–80 Hz). Although the functional significance of oscillatory changes during language processing is not as well understood as that of ERPs such as the N400, multiple studies now converge on the finding that normal semantic unification is accompanied by an increase in gamma band synchronization, which is disrupted upon encountering semantic unification difficulties, such as semantically anomalous words (Peña & Melloni, 2012; Urrutia, de Vega, & Bastiaansen, 2012; Wang, Zhu, & Bastiaansen, 2012; Hald, Bastiaansen, & Hagoort, 2006; Hagoort, Hald, Bastiaansen, & Petersson, 2004).

The Present Study

In this study, we examined the roles of semantic expectancy and semantic unification of literal word meanings in sentence comprehension. Specifically, we investigated whether these operations are carried out even in situations where they are theoretically unnecessary: namely, in the comprehension of (Dutch) idiomatic expressions (example: *tegen de lamp lopen*, lit. to walk against the lamp, meaning *to get caught*).

Participants were presented with two types of equally predictable sentence contexts: literal and idiomatic (see Table 1). Following Federmeier and Kutas (1999), in both contexts the critical word was (1) a correct and expected word (COR condition), (2) a word that was semantically related to the expected word (REL condition), or (3) a semantically unrelated word (UNREL condition). Both (2) and (3) were semantic violations.

Note that in this design the COR word was never actually presented in the REL and UNREL conditions but only predictable from the sentence context; thus, the contrast between the REL and UNREL conditions allowed for the investigation of the activation of literal word meanings based on the context alone. As such, this study builds on work concerning expectations for upcoming words (for a review, see Federmeier, 2007). Given our focus on semantics, we use the term "semantic expectancy."

Two experiments were carried out. In Experiment 1, participants read sentences and performed a lexical decision task on the critical words, a well-established measure of lexical access that is influenced by semantic variables (Meyer & Schvaneveldt, 1971). If literal word meanings are not activated by idiomatic contexts, we expected a semantic effect (i.e., shorter lexical decision latencies to words from the REL compared with the UNREL condition)

Table 1. Dutch Example Sentences and Their Literal Translation Equivalents in English

Condition	*Example*
Idiomatic Context	
COR	Na vele transacties liep de onvoorzichtige fraudeur uiteindelijk tegen de lamp gisteren.
	After many transactions the careless scammer eventually walked against the lamp yesterday.
REL	Na vele transacties liep de onvoorzichtige fraudeur uiteindelijk tegen de kaars gisteren.
	After many transactions the careless scammer eventually walked against the candle yesterday.
UNREL	Na vele transacties liep de onvoorzichtige fraudeur uiteindelijk tegen de vis gisteren.
	After many transactions the careless scammer eventually walked against the fish yesterday.
Literal Context	
COR	Na de lunch draaide de electricien het nieuwe peertje in de lamp gisteren.
	After lunch the electrician screwed the new light bulb into the lamp yesterday.
REL	Na de lunch draaide de electricien het nieuwe peertje in de kaars gisteren.
	After lunch the electrician screwed the new light bulb into the candle yesterday.
UNREL	Na de lunch draaide de electricien het nieuwe peertje in de vis gisteren.
	After lunch the electrician screwed the new light bulb into the fish yesterday.

The figurative meaning of the idiomatic context with the correct critical word (lamp) is *After many transactions the careless scammer eventually got caught yesterday*. Critical words are underlined.

in literal but not in idiomatic contexts. This finding would point to differences in word retrieval in idioms compared with literal sentences.

In Experiment 2, other participants read the same sentences while their EEG was recorded, but no task was involved other than to read for comprehension. In the ERPs to literal sentences, we expected to obtain a "graded" N400 pattern similar to what Federmeier and Kutas (1999) observed, with a reduction in N400 amplitude for the REL condition compared with the UNREL condition. For words in idioms, the same graded N400 pattern was predicted to occur only if semantic expectancy for literal word meanings extends to this type of context. If literal word meanings do not form part of expectancies in idiomatic contexts, the N400 response to the REL condition should not differ from that to the UNREL condition. Regarding power changes, in literal sentences, as in previous studies, increases in gamma band power were predicted to be disrupted by semantically anomalous words (the REL and UNREL conditions) compared with the COR condition. If the semantic unification load is higher for the UNREL than for the REL condition, there should also be a difference between these conditions in the gamma band. Importantly, we hypothesized that if semantic unification is less engaged in idioms than in literal sentences, any semantic effects should disappear in idioms. Finally, also in the correct sentences, more gamma power should be observed in literal sentences, where semantic unification continues, than in idioms.

EXPERIMENT 1

Methods

Participants

Twenty-four student volunteers from Radboud University Nijmegen (mean age = 22 years, range = 18–26 years, 17 women and 7 men) were paid for their participation in the experiment. All were right-handed native speakers of Dutch with normal or corrected-to-normal vision and no history of neurological disorders or language disorders. None of them took part in any of the pretests.

Materials and Design

The materials consisted of 90 sets of six sentences (see Table 1). In every set, the critical word was either the final noun of an idiom (for instance, "lamp" in the Dutch idiom "tegen de lamp lopen") or the same noun in a literal context. To avoid interference from sentence wrap-up effects, critical words were never placed in the final position of a sentence. The design included two within-subject factors: Idiomaticity, with two levels (Literal, Idiomatic) specifying the type of context, and Condition, with three levels (COR, REL, UNREL) specifying the type of critical word in the context. Across the item set, the same critical words were used in the Idiomatic sentences and their Literal counterparts. The materials were divided among six lists using a Latin square design such that no sentence context or critical word was repeated within a list. Each list consisted of 240 sentences: 90 experimental items (45 idioms, 45 literal sentences, with each condition [COR, REL, UNREL] being represented by 15 sentences), 45 idiomatic sentences with a pseudoword as the critical word, 45 literal sentences with a pseudoword as the critical word, and 60 fillers of which 30 had a pseudoword as the critical word. Thus, 50% of the trials contained a pseudoword. Each of the six lists was used for four participants, who each received a different randomization.

Pretests

The items were pretested on several dimensions (see below). We started out with 151 items. In the course of pretesting, items were discarded when they did not meet certain criteria, leaving a final set of 90 items for the experiments.

Idiom selection and ratings. We selected familiar and nontransparent (opaque) idioms, that is, well-known idioms of which the individual word meanings were unrelated to the overall meaning of the expression, based on a paper-and-pencil pretest. Using our own sense of transparency, 151 opaque idiomatic expressions (for instance, "tegen de lamp lopen," to walk against the lamp, *to get caught*) and 151 transparent idioms (for instance, "iemand voor de rechter slepen," to drag someone before the judge, *to sue someone*) were selected from a Dutch idiom dictionary (de Groot, 1999). The items were pseudorandomized. Idioms were presented in infinitival form (e.g., "tegen de lamp lopen," *to walk against the lamp*), with the critical word underlined for the transparency ratings. For the familiarity ratings, participants answered three multiple choice questions by circling one of three answers: (1) "How often do you come across this expression?" (often/sometimes/never), (2) "Do you know its meaning?" (yes/approximately/no), and (3) "Do you use it yourself?" (often/sometimes/never). In the analysis, answers were coded as 0 (*never/no*), 1 (*sometimes/approximately*), or 2 (*often/yes*) and added up for each idiom for each participant separately, yielding "idiom familiarity values" ranging from 0 (*unfamiliar*) to 6 (*highly familiar*). For the transparency ratings, participants were asked to rate each expression's transparency by judging "to what extent the underlined word has something to do with the figurative meaning of the expression" on a scale from 1 (*not transparent at all*) to 7 (*very transparent*). Twenty students (mean age = 21 years, range = 18–25 years, 14 women and 6 men) were paid for participation. Testing took approximately 1 hr. Only idioms for which at least 90% of the participants had a familiarity value of 2 or higher were included in the experiment, which led to the removal of 35 items. The average familiarity value of the 90 items used in the experiments (i.e.,

after removal of 61 items based on all pretests including this one and those discussed below) was 4.2 (*SD* = 0.6, range = 2.8–5.6).

Transparency ratings from participants who did not know an idiom (familiarity value below 2) were discarded before computing the average transparency per item. Four idioms were discarded because they were close to or on the transparent side of the scale (average transparency rating above 3.5). The final set of 90 items used in the experiments, with transparency ratings averaged over participants for each item, had a transparency rating of 2.0 (*SD* = 0.5, range = 1.1–3.5).

Word characteristics. Words for the REL condition were obtained in two steps. First, free association norms for Dutch words (van Loon-Vervoorn & van Bekkum, 1991; de Groot & de Bil, 1987; Lauteslager, Schaap, & Schievels, 1986) were consulted in order of recency of publication. Norms were available for 42 of the 90 COR words. Associated words were not used if they were grammatically inappropriate as a substitution for the COR word or semantically too appropriate in the sentence (i.e., they would not have yielded a semantic violation). On average, the third most frequent association was selected, with the average association frequency being 10%. The remaining items were selected and assessed using latent semantic analysis (LSA; Landauer & Dumais, 1997, available on-line at lsa.colorado.edu/), a semantic similarity measure that computes how often two words co-occur with the same set of other words. Following Chwilla and Kolk (2002), critical words were translated to English and submitted to the on-line "pairwise (term to term) comparison," once comparing the REL to the COR word and once comparing the UNREL to the COR word, using the General_Reading_up_to_1st_year_college topic space. The LSA values confirmed that words from the REL condition were on average more highly related (0.36, *SD* = 0.19) than the UNREL words (0.08, *SD* = 0.05), $t(89) = 13.68$, $p < .001$.

The log-transformed lemma frequency per million and word length of the critical words were extracted from the Dutch CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1993). The conditions did not differ significantly in word frequency, $F(2, 267) = 1.42$, $p = .244$, which was 1.5 on average (*SD* = 0.5), but there was a difference in word length, $F(2, 267) = 5.78$, $p < .01$, with words from the REL condition being longer compared with both the COR words, $p < .01$, and the UNREL words, $p < .05$. However, the effect size was only half a character (REL: 5 characters, vs. UNREL and COR: 4.5 characters). Because the same critical words were used in literal and idiomatic contexts, any effects of idiomaticity cannot be attributed to lexical characteristics.

Cloze probability. For the purpose of this study, we wanted to vary the context categorically (literal vs. idiomatic) while controlling for the cloze probability of the COR words; therefore, we performed a paper-and-pencil cloze probability test. All 302 sentence contexts up to the critical word were included (this was at a stage before removal of items based on familiarity and transparency) and divided into two randomized lists. Idiomatic and literal contexts that were predictive of the same word did not appear on the same list. Thirty participants (mean age = 21 years, range = 18–30 years, 20 women and 10 men; one additional participant was excluded from analysis for having grown up outside the Netherlands) were instructed to complete each sentence fragment with the first word or words that came to mind, keeping it short but grammatical without trying to be original. Fifteen participants were randomly assigned to each list, with eight of them receiving the items in reversed order compared with the other seven. Testing took approximately half an hour. A subset of eight literal items was rewritten and clozed separately by 15 other students (mean age = 22 years, range = 19–24 years, eight women and seven men). For the 90 items used in the experiments (i.e., after selection based on all pretests including the plausibility test described below), critical words in literal and idiomatic sentences had similar cloze probabilities (.82, *SD* = .20 and .85, *SD* = .20, respectively), $t(89) = 1.262$, $p = .210$. Cloze probability for the two types of violations was always zero.

Plausibility ratings. To test whether one of the two types of violation was an unlikely but plausible ending and the word could therefore be integrated into its sentence context, an independent group of 30 participants (mean age = 22 years, range = 18–32 years, 20 women and 10 men) was asked to rate the plausibility of each sentence on a scale from 1 (*highly implausible*) to 7 (*highly plausible*). Where Dutch grammar allowed for it, sentence fragments after the critical words were removed. A set of 666 sentences (111 sets of six sentences each, consisting of the 96 sets of sentences that had met the familiarity, transparency, and cloze probability criteria and some additional items that had not yet been discarded) were divided among three lists using a Latin square design to avoid repetition of critical words or sentence contexts. Each of the three lists was completed by 10 participants, of whom five received the items in reversed order. Testing took approximately 1 hr.

Items were excluded if the REL or UNREL condition received an average plausibility rating above 4 (i.e., was on the plausible side of the scale), which was the case for 6 of the 96 items. The ratings for the final 90 sets of six items that were used in the EEG experiment (see Table 2) were submitted to a by-participants 2 × 3 repeated-measures ANOVA with two within-subject factors: Idiomaticity (Literal, Idiomatic) and Condition (COR, REL, UNREL). All p and mean squared error (*MSE*) values are Greenhouse–Geisser corrected where necessary, but the original degrees of freedom are reported. Plausibility ratings were higher in Literal contexts (3.51, *SE* = 0.10) than in Idiomatic contexts (3.29, *SE* = 0.09), $F(1, 29) = 20.955$, $MSE = .106$, $p < .001$. There was a main effect of Condition, $F(2, 58) =$

Table 2. Mean Plausibility Ratings for the Sentences on a Scale from 1 (*Highly Implausible*) to 7 (*Highly Plausible*)

	Context	
Condition	*Literal*	*Idiomatic*
COR	6.2 (0.6)	6.3 (0.4)
REL	2.4 (0.8)	1.9 (0.8)
UNREL	1.9 (0.7)	1.7 (0.7)

Standard deviations are indicated between parentheses.

799.20, $MSE = .747$, $p < .001$, and the interaction between Idiomaticity and Condition was also significant, $F(2, 58) = 30.12$, $MSE = .067$, $p < .001$. Simple effects tests revealed that all three conditions differed from one another in both contexts (all ps $< .01$). COR words were marginally more plausible in Idiomatic contexts than in Literal contexts, $F(1, 29) = 3.999$, $MSE = .162$, $p = .055$, whereas the REL and UNREL conditions were more implausible in Idiomatic than in Literal contexts, $F(1, 29) = 47.980$, $MSE = .200$, $p < .001$ and $F(1, 29) = 17.689$, $MSE = .103$, $p < .001$, respectively. As described in the Discussion, the plausibility ratings were not directly related to the effects observed in the experiments.

Procedure

Participants were tested individually in a single session in a soundproof room. The presentation of the visual stimuli and the recording of the RTs were controlled by a computer program and a dedicated button box developed by the Donders Centre for Cognition. The participants sat at a table with the computer monitor at a 60-cm distance. They received a Dutch-written instruction, repeated orally, which informed them that they would see a series of sequentially presented words. One of these words was presented in red color (similar to Schwartz & Kroll, 2006), and they were asked to decide whether this was a Dutch word or not by pressing one of two buttons on the button box in front of them. The participants were told to react as quickly as possible without making too many errors. The session began with a short practice block consisting of 10 sentences to allow participants to familiarize themselves with the task. Each trial began with a fixation cross (+) that remained on the screen for a duration of 1000 msec. A sentence was then presented one word at a time in the center of the screen, in black letters of the font Tahoma, size 30, on a white background. Each word was presented for 300 msec, with a 300-msec blank screen following each word. The experiment was divided into 12 short blocks of 20 sentences each (each block lasting approximately 4 min). After each block, participants were allowed to take a break for as long as they wanted. In total, the session lasted nearly an hour.

Results

Incorrect responses and RTs shorter than 300 msec or longer than 1500 msec (3.6%) were removed from the experimental trials. The resulting RTs in all word conditions are presented in Table 3. The nonword conditions resulted in a mean RT of 700 msec ($SD = 187$) and a mean accuracy rate of .95. To test for effects of Context and Condition, two-factor repeated-measures ANOVAs were performed on the mean RTs with participant as a random variable. No ANOVAs with items as a random variable were conducted in accordance with the EEG data and because the selected items were matched on word frequency (Raaijmakers, Schrijnemakers, & Gremmen, 1999).

Idiomatic conditions were responded to faster than Literal conditions, resulting in a main effect Idiomaticity, $F(1, 23) = 9.41$, $MSE = 1380.29$, $p < .01$. Note that this converges with Swinney and Cutler (1979), who explained this effect in terms of quickly accessed unitary idiom representations. Correct conditions were responded to faster than incorrect conditions, resulting in a main effect of Condition, $F(2, 22) = 41.28$, $MSE = 3003.35$, $p < .001$. The interaction of Idiomaticity and Condition did not reach significance, $F(2, 22) = 1.48$, $MSE = 1718.28$, $p = .24$.

Next, we considered the planned pairwise comparisons between the levels of each condition for Literal and Idiomatic contexts separately. For the Literal context, significant differences were found between all conditions: COR yielded faster RTs than REL, $p < .001$, than UNREL, $p < .001$, and responses in the REL condition were significantly faster than those in the UNREL condition, $p < .034$. For the Idiomatic context, participants responded faster in the COR than the REL and UNREL conditions, both ps $< .001$. However, the REL and UNREL conditions did not differ significantly, $p = .243$, suggesting that literal word meanings were not activated. An additional ANOVA on log-transformed RTs yielded the same pattern of results. Given the very low error rates no error ANOVAs are reported.

EXPERIMENT 2

Methods

Participants

Twenty-four student volunteers from Radboud University Nijmegen and the HAN University of Applied Sciences

Table 3. Mean RTs (msec) in Experiment 1

	Context	
Condition	*Literal*	*Idiomatic*
COR	594 (98)	565 (80)
REL	669 (108)	661 (105)
UNREL	698 (93)	677 (96)

Standard deviations are indicated between parentheses.

(mean age = 22 years, range = 18–30 years, 21 women and 3 men) gave informed consent and were paid for their participation in the EEG experiment. All were right-handed, native speakers of Dutch with normal or corrected-to-normal vision and no history of neurological disorders or language disorders. None of them took part in Experiment 1 or in any of the pretests. Data from two additional participants were discarded because of excessive blinking or high amplitude alpha activity and tiredness.

Materials and Design

The same materials and design as in Experiment 1 were used, except that, as no lexical decisions were required, no pseudowords occurred in Experiment 2. The six lists from Experiment 1 were merged to three lists, removing all pseudowords, such that on every list each of the six conditions was represented by 30 items. Each sentence within a pair (e.g., the literal and the idiomatic "lamp" sentence) was used to represent a different condition within a list (e.g., if the idiomatic sentence contained "lamp," the literal sentence would contain "fish" or "candle"), such that no sentence contexts or critical words were repeated. In every list 50% of the critical words was anomalous.

Procedure

Participants were tested individually in a single session in a soundproof, electrically shielded room. They were seated in a comfortable chair at a distance of approximately 60 cm from a computer screen and instructed to read the sentences for comprehension while avoiding blinks and movements. The instructions were given in written form and then repeated orally. The session began with a short practice block consisting of 10 sentences. Each trial began with a fixation cross (+) that remained on the screen for a duration of 1000 msec to orient the participant toward the center of the screen. A sentence was then presented one word at a time in the center of the screen, in black letters of the font Tahoma, size 30, on a white background. Each word was presented for 300 msec, with a 300-msec blank screen following each word. After every sentence, three asterisks (***) appeared for a duration of 3000 msec. During this period, participants were free to blink. The experiment was divided into 12 short blocks of 20 sentences each, each block lasting approximately 4 min. After each block, participants were allowed to take a break for as long as they wanted. The session included half an hour of electrode application and instruction and approximately 1 hr of reading sentences.

EEG Recording

EEG was recorded from 61 active Ag/AgCl electrodes, of which 59 were mounted in a cap (actiCap), referenced to the left mastoid. Two separate electrodes were placed at the left and right mastoids. Blinks were monitored through an electrode on the infraorbital ridge below the left eye. Horizontal eye movements were monitored through two electrodes in the cap (LEOG and REOG), placed approximately at each outer canthus. The ground electrode was placed on the forehead. Electrode impedance was kept below 10 kΩ. EEG and EOG recordings were amplified through BrainAmp DC amplifiers with a bandpass filter of 0.016–100 Hz, digitized on-line with a sampling frequency of 500 Hz, and stored for off-line analysis.

ERP Analysis

For the ERPs, the data were rereferenced off-line to the average of the left and right mastoids. Bipolar vertical EOG was computed as the difference between the electrode on the infraorbital ridge of the left eye and the electrode right above this eye in the cap. Bipolar horizontal EOG was computed as the difference between the LEOG and REOG electrodes. A bandpass filter of 0.1–30 Hz was applied. The continuous EEG was then segmented into epochs of 1350 msec, lasting from 150 msec before word onset until 1200 msec after word onset. An average baseline of 150 msec before stimulus onset was subtracted. Trials were screened for artifacts semiautomatically, and those contaminated by blinks, eye movements, muscle activity, and so forth, were removed. Approximately 6% of the trials was lost because of such artifacts, and the number of rejected trials was comparable across conditions, $F < 1$. Average ERPs were then computed across trials for each type of critical word (COR, REL, UNREL) in each type of context (Idiomatic, Literal).

To assess the significance of observed N400 effects, mean voltage measures were taken. On the basis of possible component overlap with a late positivity in our data (see Results) and following other work on nonliteral language indicating short-lived (300–400 msec) N400 effects (de Grauwe, Swain, Holcomb, Ditman, & Kuperberg, 2010), we used an Early N400 (300–400 msec) and a Late N400 (400–500 msec) time window. Late positivities (500–800 msec) were of interest as well. For each time window, mean voltage measures were averaged over quadrants of nine electrodes each, which divided the data into anterior, posterior, left, and right parts of the scalp (see Figure 1). The data were then subjected to 2 × 2 × 2 × 3 repeated-measures ANOVAs with the factors Hemisphere (Left, Right), Anteriority (Anterior, Posterior), Idiomaticity (Literal, Idiomatic), and Condition (COR, REL, UNREL).

Time–Frequency Analysis of Power

Time–frequency (TF) analysis was performed with the Fieldtrip software package, an open-source Matlab toolbox for neurophysiological data analysis (Oostenveld, Fries, Maris, & Schoffelen, 2011). We used a multitaper approach, as described by Mitra and Pesaran (1999) for computing TF representations, ranging from 20 to 100 Hz. The power

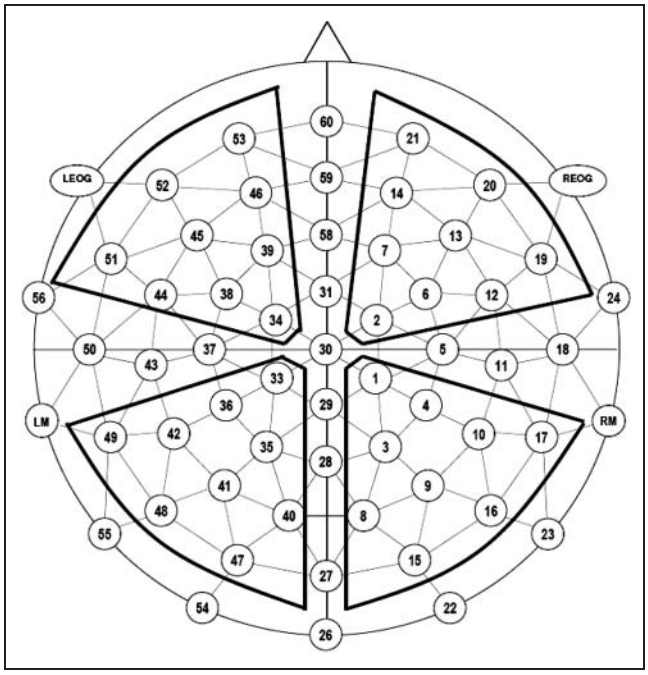

Figure 1. Electrode layout. Thick lines indicate quadrants used for mean amplitude analyses on the ERPs. Left mastoid (LM) and right mastoid (RM) electrodes were placed outside the cap onto the mastoids.

changes were computed with 400-msec time-smoothing and a 5-Hz frequency-smoothing window in 2.5-Hz frequency steps and 10-msec time steps. Power estimates thus obtained were then averaged across trials separately for each condition, for each subject. The resulting subject-averaged power changes in the poststimulus interval were expressed as a relative change from the baseline interval (from −0.5 to −0.15 sec).

The significance of the difference between conditions was evaluated by means of a cluster-based random permutation test (Maris & Oostenveld, 2007) that naturally takes care of the multiple comparisons problem by identifying clusters of significant differences between conditions in the time, space, and frequency dimensions. Note that this procedure only allows for pairwise comparisons. We therefore computed separate contrasts between COR and UNREL, COR and REL, and REL and UNREL, for each context separately. In addition, we computed the contrast between COR in the literal context and COR in the idiomatic context.

Results

ERPs

As can be seen from Figure 2, in Literal contexts critical words from the REL and the UNREL condition elicited negativities relative to words in the COR condition, lasting from approximately 300 to 500 msec and exhibiting a centroparietal maximum (N400s). N400s were short-lived in the Idiomatic contexts (Figure 3), where violations also led to a late positivity. In the following, main effects and interactions are reported only when they involve the factors Idiomaticity and/or Condition. Nonsignificant effects are reported only when the lack of significance is relevant for the interpretation. All p and *MSE* values are Greenhouse–Geisser corrected where necessary, but the original degrees of freedom are reported.

In the early N400 time window (300–400 msec), as supported by the statistics in Table 4, critical words in Idiomatic contexts had a more positive voltage (1.10 μV) compared with critical words in Literal contexts (−0.57 μV). Although not the focus of this study, note that it is possible that this effect reflects a P300 for idioms rather than an N400 reduction, as discussed in Vespignani et al. (2010). The effect was larger over the right hemisphere (1.86 μV difference) compared with the left hemisphere (1.47 μV difference). There were differences between conditions (COR: 1.62 μV, REL: −0.06 μV, UNREL: −0.77 μV). To test

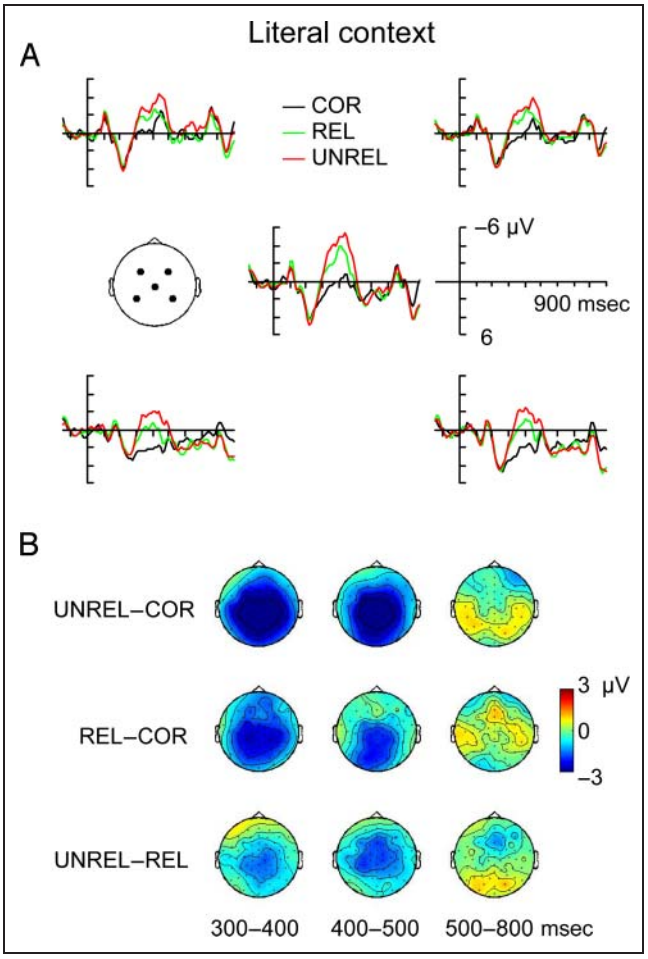

Figure 2. Grand-averaged (n = 24) ERPs in Literal contexts. (A) Waveforms (negative plotted up, critical word [e.g., "lamp"] presented at 0 msec, next word [e.g., "gisteren"] at 600 msec) for five representative electrodes, of which the locations are indicated on a head map (left). The four corner electrodes are taken from each of the quadrants used in the statistical analyses. (B) Scalp topographies of the difference between each of the conditions.

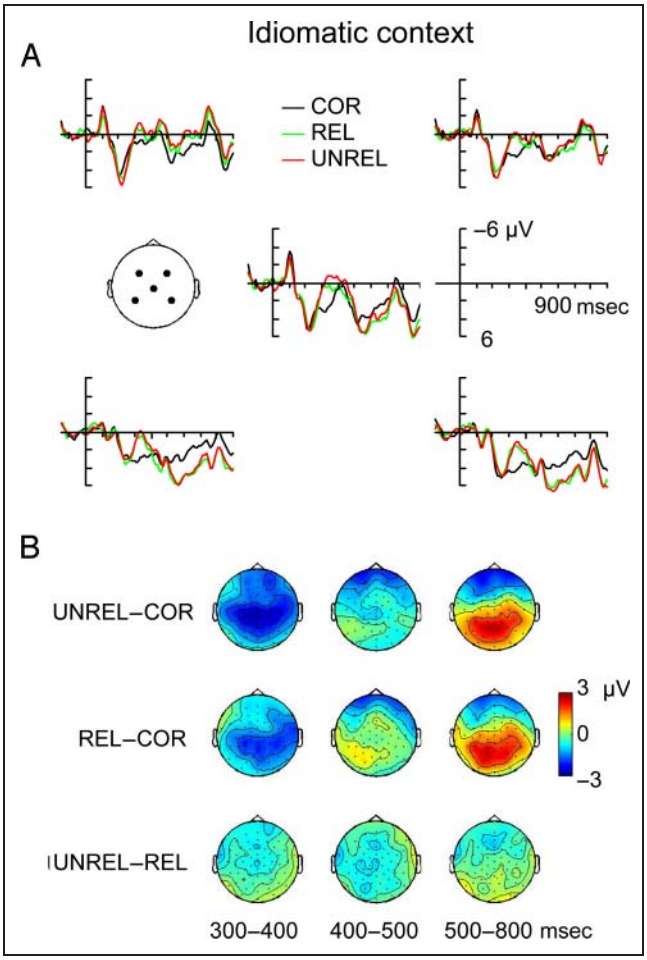

Figure 3. Grand-averaged ERPs in Idiomatic contexts. (A) Waveforms (negative plotted up, critical word presented at 0 msec) for the same electrodes as in Figure 2. (B) Scalp topographies of the difference between each of the conditions.

for semantic effects in Literal and Idiomatic contexts, planned comparisons between the conditions in the two posterior quadrants were carried out for each context, based on the known N400 scalp distribution (e.g., Kutas & Hillyard, 1980). In Literal contexts, all conditions differed significantly from one another: UNREL (−1.02 μV) was more negative than COR (2.46 μV), $F(1, 23) = 56.24$, $MSE = 5.14$, $p < .001$, REL (0.21 μV) was more negative than COR, $F(1, 23) = 31.97$, $MSE = 3.78$, $p < .001$, and importantly, UNREL was more negative than REL, $F(1, 23) = 9.64$, $MSE = 3.76$, $p < .01$. In Idiomatic contexts, not all conditions differed significantly from one another: Both UNREL (1.40 μV) and REL (1.78 μV) were more negative than COR (3.50 μV), $F(1, 23) = 14.54$, $MSE = 7.26$, $p < .01$ and $F(1, 23) = 17.20$, $MSE = 4.15$, $p < .001$, respectively, but REL and UNREL did not differ significantly, $F(1, 23) = .78$, $MSE = 4.28$, $p = .387$. Even when a more liberal analysis was performed, using only the data from the central electrode with the maximum raw effect (electrode 30; 1.01 μV difference), a planned comparison again did not indicate a significant difference between the REL and the UNREL conditions in Idiomatic contexts, $F(1, 23) = 2.04$, $MSE = 12.15$, $p = .167$, thus confirming the quadrant analysis.

In the late N400 time window (400–500 msec), as supported by the statistics in Table 5, the mean voltage to critical words was more positive in idioms (1.49 μV) than in literal contexts (−1.13 μV). This effect was larger over the right hemisphere (2.91 μV difference) than the left hemisphere (2.34 μV difference). There were differences between the three conditions (COR: .96 μV, REL: .27 μV, UNREL: −.69 μV). The Anteriority × Idiomaticity × Condition interaction (Table 5) was because of an Idiomaticity × Condition interaction being not significant at Anterior sites, $F(2, 46) = 1.10$, $MSE = 3.37$, $p = .342$,

Table 4. Omnibus ANOVA on ERPs in the Early N400 Time Window (300–400 msec)

Effect	*df*	*F*	*MSE*	*p*
Idiomaticity	(1,23)	19.94	19.96	<.001
Condition	(2,46)	36.73	9.47	<.001
Hemisphere × Idiomaticity	(1,23)	4.88	1.11	.037
Anteriority × Idiomaticity	(1,23)	.01	2.34	.926
Hemisphere × Anteriority × Idiomaticity	(1,23)	.12	.26	.734
Hemisphere × Condition	(2,46)	.38	1.91	.636
Anteriority × Condition	(2,46)	2.24	4.23	.128
Hemisphere × Anteriority × Condition	(2,46)	1.13	.23	.324
Idiomaticity × Condition	(2,46)	1.35	10.25	.268
Hemisphere × Idiomaticity × Condition	(2,46)	.77	.72	.465
Anteriority × Idiomaticity × Condition	(2,46)	1.60	1.68	.216
Hemisphere × Anteriority × Idiomaticity × Condition	(2,46)	.04	.16	.954

Table 5. Omnibus ANOVA on ERPs in the Late N400 Time Window (400–500 msec)

Effect	*df*	*F*	*MSE*	*p*
Idiomaticity	(1,23)	69.10	14.35	<.001
Condition	(2,46)	9.50	14.71	<.001
Hemisphere × Idiomaticity	(1,23)	10.07	1.20	.004
Anteriority × Idiomaticity	(1,23)	3.66	3.81	.069
Hemisphere × Anteriority × Idiomaticity	(1,23)	.58	.15	.453
Hemisphere × Condition	(2,46)	.06	1.15	.928
Anteriority × Condition	(2,46)	.23	3.33	.784
Hemisphere × Anteriority × Condition	(2,46)	1.02	.22	.358
Idiomaticity × Condition	(2,46)	4.06	10.73	.024
Hemisphere × Idiomaticity × Condition	(2,46)	.14	1.13	.868
Anteriority × Idiomaticity × Condition	(2,46)	7.20	1.62	.002
Hemisphere × Anteriority × Idiomaticity × Condition	(2,46)	2.26	.20	.121

but significant at Posterior sites, $F(2, 46) = 8.39$, $MSE = 2.82$, $p < .01$. Here, simple effects tests indicated an effect of Condition in Literal contexts, $F(2, 46) = 17.58$, $MSE = 3.37$, $p < .001$, but not in Idiomatic contexts, $F(2, 46) = .89$, $MSE = 3.60$, $p = .414$. Planned comparisons within Literal contexts revealed a graded N400 effect with UNREL (−1.42 μV) more negative than COR (1.70 μV), $F(1, 23) = 32.02$, $MSE = 7.27$, $p < .001$, REL (−.08 μV) more negative than COR, $F(1, 23) = 12.56$, $MSE = 6.07$, $p < .01$, and importantly UNREL more negative than REL, $F(1, 23) = 6.39$, $MSE = 6.67$, $p < .05$. In idioms, there were no differences between the conditions (all ps > .20). In summary, in Literal contexts the graded N400 effect from the Early N400 time window continued into the Late N400 time window, whereas in idioms there was no graded effect.

In the 500–800 msec time window (Table 6), critical words had a more positive voltage in idioms (2.14 μV) than in Literal contexts (.56 μV). The difference was larger at Posterior sites (2.18 μV difference) than Anterior sites (.98 μV difference). The four-way interaction (Table 6) was further clarified by testing for the effects of Hemisphere, Idiomaticity and Condition at the two levels of Anteriority. At Anterior sites, ERPs to critical words were more positive in idioms (.61 μV) than literal sentences (−.37 μV), $F(1, 23) = 7.495$, $MSE = 9.168$, $p < .05$. This

Table 6. Omnibus ANOVA on ERPs in the P600 Time Window (500–800 msec)

Effect	*df*	*F*	*MSE*	*p*
Idiomaticity	(1,23)	20.67	17.37	<.001
Condition	(2,46)	1.15	12.02	.326
Hemisphere × Idiomaticity	(1,23)	3.50	1.13	.074
Anteriority × Idiomaticity	(1,23)	23.85	2.19	<.001
Hemisphere × Anteriority × Idiomaticity	(1,23)	.14	.12	.715
Hemisphere × Condition	(2,46)	.56	1.11	.552
Anteriority × Condition	(2,46)	12.49	2.80	<.001
Hemisphere × Anteriority × Condition	(2,46)	3.15	.29	.069
Idiomaticity × Condition	(2,46)	.70	8.78	.499
Hemisphere × Idiomaticity × Condition	(2,46)	.55	.78	.572
Anteriority × Idiomaticity × Condition	(2,46)	9.22	1.69	.001
Hemisphere × Anteriority × Idiomaticity × Condition	(2,46)	3.54	.19	.042

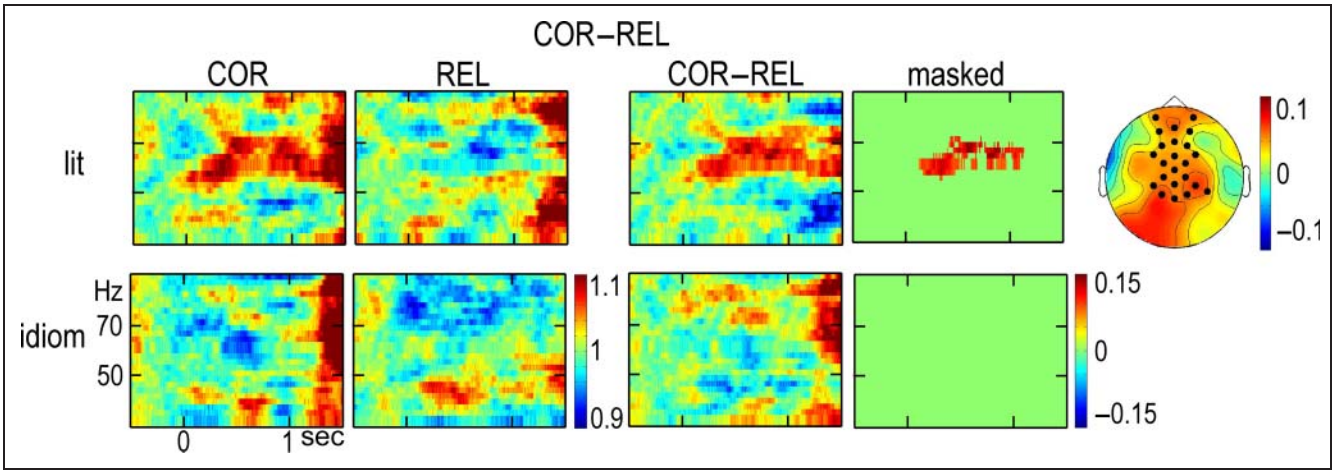

Figure 4. Comparison between the TF representations of the power changes in the COR and REL conditions in Literal (lit) and Idiomatic (idiom) contexts. A representative channel (4, see Figure 1) is shown, with blue representing power decreases and red representing power increases. The "masked" panels show the TF points with significant differences between the conditions. The scalp map shows the distribution of the significant differences across the scalp in a time range from 0.2 to 1 sec after CW onset and a frequency range of 55–70 Hz. Black dots indicate the electrodes that participate in the significant cluster.

was also the case at Posterior sites (Idiomatic: 3.67 μV, Literal: 1.49 μV), $F(1, 23) = 32.969$, $MSE = 10.389$, $p < .001$. There were differences between conditions at Posterior sites (COR: 1.85 μV, REL: 2.88 μV, UNREL: 3.02 μV), $F(2, 46) = 6.391$, $MSE = 6.468$, $p < .01$. An Idiomaticity × Condition interaction, $F(2, 46) = 3.900$, $MSE = 5.234$, $p < .05$, indicated an effect of Condition for idioms, $F(2, 46) = 8.320$, $MSE = 3.608$, $p < .01$, but not literal sentences, $F(2, 46) = .559$, $MSE = 2.235$, $p = .575$. Posterior contrasts in idioms indicated positivities for both the REL (4.30 μV) and the UNREL condition (4.30 μV) relative to the COR condition (2.42 μV), $F(1, 23) = 12.649$, $MSE = 6.691$, $p < .01$ and $F(1, 23) = 6.54$, $MSE = 8.63$, $p < .05$, respectively, and no difference between UNREL and REL, $F(1, 23) < .001$, $MSE = 6.691$, $p = .996$. In summary, there was a significant posterior positivity for violations in idioms, regardless of semantic relatedness.

TF Analysis

Power in the gamma frequency band was sensitive to the manipulations. In the COR–REL contrast, significantly larger gamma power was observed for COR than for REL in literal contexts ($p < .005$), but not in idiomatic contexts (see Figure 4). This gamma power increase was most prominent between 60 and 70 Hz and was sustained throughout the entire time interval that entered the analysis. Similarly, in the COR–UNREL contrast, larger gamma power was observed for COR than for UNREL in the literal context only ($p = .032$), in a roughly similar time and frequency range and with similar scalp topography as in the COR–REL contrast (see Figure 5). In the REL–UNREL contrast, more gamma power was observed for REL than for UNREL, again only in literal contexts, although this effect was only marginally significant ($p = .052$), and was observed in a higher frequency band (around 80 Hz) than in the other two

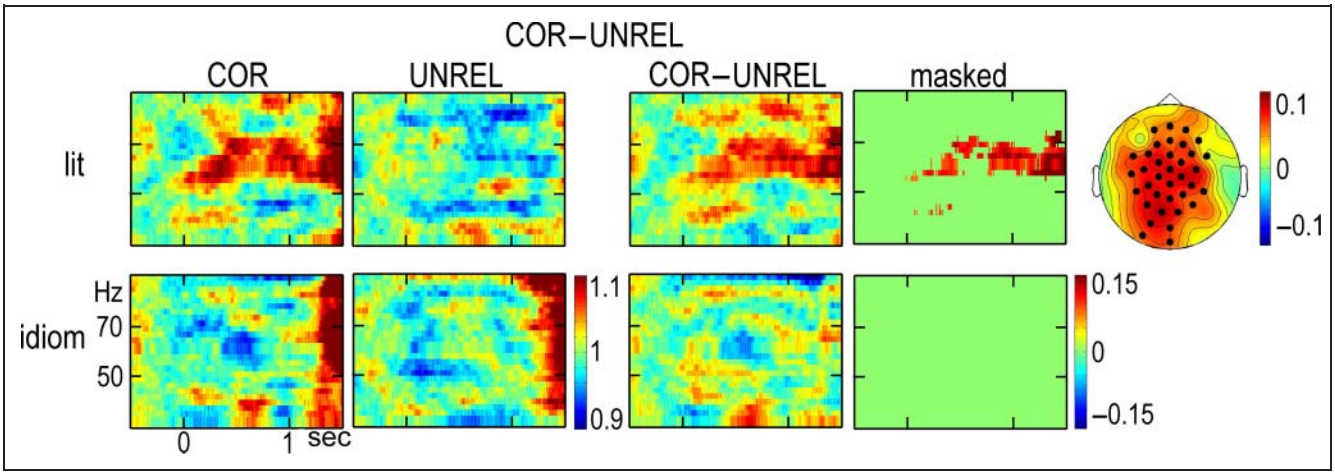

Figure 5. Comparison between the TF representations of the power changes in the COR and UNREL conditions in Literal (lit) and Idiomatic (idiom) contexts. See legend of Figure 4 for details.

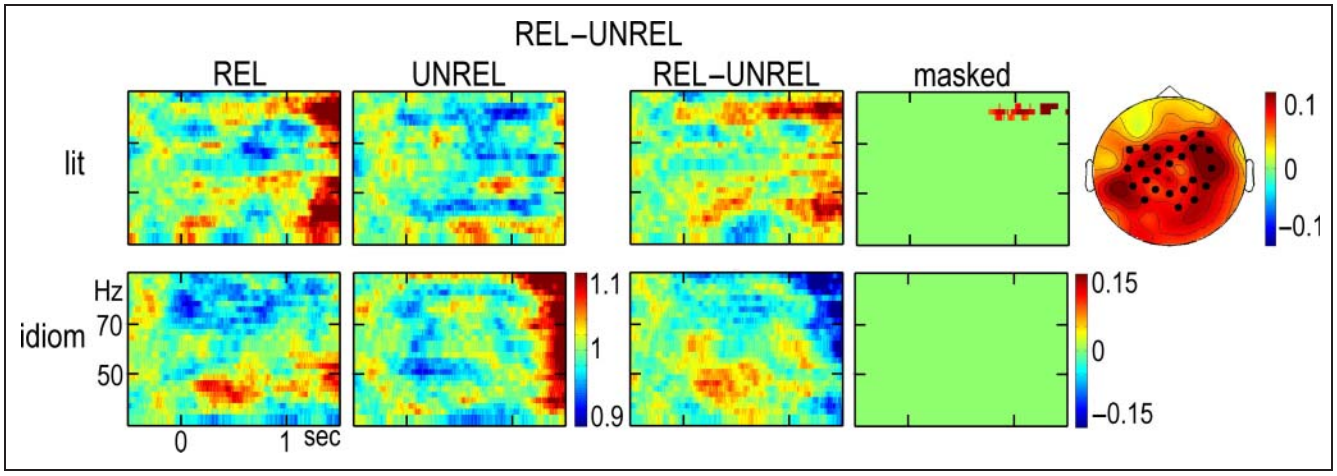

Figure 6. Comparison between the TF representations of the power changes in the REL and UNREL conditions in Literal (lit) and Idiomatic (idiom) contexts. The scalp map shows the distribution of the significant differences across the scalp in a time range from 0.4 to 1.5 sec after CW onset and a frequency range of 80–90 Hz. See Figure 4 for details.

contrasts (see Figure 6). Because the random permutation approach does not in itself provide measures of variability, participants' invidual power changes are shown in Figure 7. Finally, in the direct contrast between the two COR conditions in the different contexts, larger gamma power was observed for the COR condition in the literal context relative to the COR condition in the idiomatic context ($p = .045$; see Figure 8). An additional analysis was performed on a lower frequency range (2–30 Hz), but this yielded no significant differences between any of the conditions.

DISCUSSION

This study investigated whether the basic sentence processing operations of word meaning retrieval and semantic integration are routinely carried out, even when they are unnecessary. This was done by examining whether behavioral and electrophysiological manifestations of semantic processing typically observed for the comprehension of compositional literal sentences extend to the case of opaque idioms where the individual word meanings are unrelated to the figurative and ultimately relevant meaning. Semantic manipulations in predictable literal sentences yielded previously established behavioral and electrophysiological signatures of semantic processing: semantic facilitation in lexical decision and semantic effects on N400 amplitude and on power in the gamma frequency band. However, the same manipulations in idioms yielded none of these effects. Instead, semantic violations elicited a late positivity in idioms. Moreover, gamma band power was lower in idioms than in literal sentences. The results are relevant for idiom comprehension as well as for neurocognitive accounts of semantic processing in general.

Context-dependent Semantic Processing

The finding of significant semantic facilitation (faster responses in the REL than in the UNREL condition) in lexical decision in literal sentence contexts, but not in idioms, suggests that idiomatic contexts did not lead to activation of literal word meanings. Similarly, a significant reduction in N400 amplitude for the REL condition relative to the UNREL condition was observed in literal sentence contexts (replicating Federmeier & Kutas, 1999), but not in idioms. In idioms, both types of violation (REL and UNREL) elicited a short-lived N400 of which the amplitudes were statistically indistinguishable. Assuming that the N400 reflects both word retrieval and integration (Baggio & Hagoort, 2011; Kutas & Federmeier, 2011), the results indicate that either literal word meaning activation or semantic unification or both are less engaged during idiom comprehension than in literal sentences. The N400 findings cannot be explained by integration alone, because our plausibility ratings do not match with the ERPs: Violations were more implausible in idioms than in literal sentences but led to smaller N400 amplitudes, and the REL and UNREL conditions in idioms differed in plausibility but not in N400 amplitude. Therefore, we interpret this result as a lack of semantic expectancy in idiom comprehension.

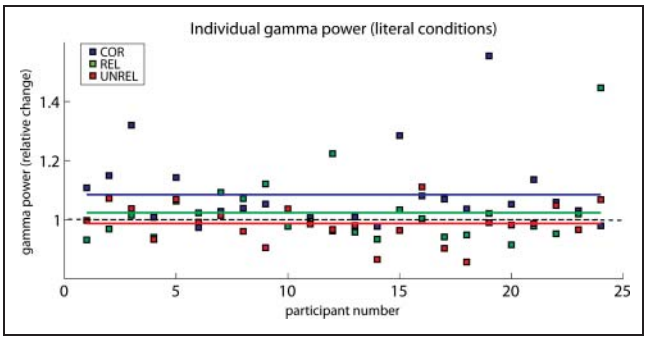

Figure 7. Individual gamma band power changes in Literal contexts. Colored squares indicate individual participant's average relative power change in the relevant TF windows described in Figures 4–6. The dashed horizontal line indicates a relative change of 1 (no change). The solid horizontal lines indicate the condition means.

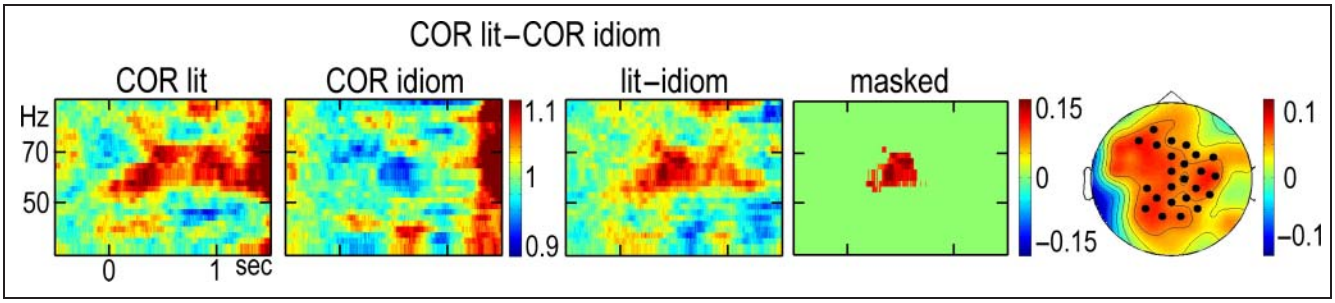

Figure 8. Comparison between the TF representations of the power changes in the COR conditions in Literal (lit) and Idiomatic (idiom) contexts. The scalp map shows the distribution of the significant differences across the scalp in a time range from 0.3 to 0.7 sec after CW onset and a frequency range of 60–80 Hz. For other details, see Figure 4.

To provide further supporting evidence, TF analyses were carried out. Literal contexts revealed a gamma increase in the COR condition relative to the REL and UNREL conditions, roughly between 50 and 70 Hz. This is consistent with previous studies suggesting that the process of semantic unification reflected by the gamma band increase is disrupted upon encountering semantic anomalies (Hald et al., 2006; Hagoort et al., 2004). A marginally significant semantic effect (REL vs. UNREL) on gamma power in literal contexts, though somewhat higher in frequency (between 70 and 80 Hz), further suggests that relationships between expected and presented words may increase the engagement of unification operations. Unlike previous studies (e.g., Hald et al., 2006), we did not observe a theta band power increase following semantic violations. Crucially, none of the gamma band effects were observed in idiomatic contexts, again suggesting that unification operations were less engaged in idiom comprehension than in the comprehension of literal sentences. A comparison between the correct words (COR condition) across contexts also revealed larger gamma power in literal contexts than in idiomatic contexts. This particular result shows the difference in gamma band activity without using semantic violations and further explains the absence of other gamma effects in idioms: Semantic integration operations reflected in gamma band synchronization were less engaged during idiom comprehension in the first place, even when the sentence continued normally (the COR condition), and therefore, this process could also not be disrupted by the semantic violations.

In summary, the behavioral data, ERP data, and TF data all converge onto the same view on semantic processing in literal language and idioms. The (pre)activation and integration of word meanings are dependent on sentence context. When reading idioms, these operations are clearly less engaged.

The results have implications for neurocognitive accounts of language processing such as the memory, unification, and control model (Hagoort, 2005). At present, this model seems to need further specification to deal with idiom comprehension. We suggest two possible extensions. One possibility is that the memory, unification, and control model implements context-dependent semantic processing, allowing for semantic unification to be "switched off" when it is not helpful. A second possibility is to describe the units that enter into the unification process differently. At present, these building blocks are almost always called words. It would be better to speak of lexical items instead and specify that these lexical items can be idioms too.

Literal Word Meaning Activation in Idioms as a Bottom–up Process

The lack of evidence for semantic expectancy of literal word meanings in idioms may appear inconsistent with previous results showing that words in idioms can prime semantically related words (Sprenger et al., 2006; Hillert & Swinney, 2001; Colombo, 1993; Cacciari & Tabossi, 1988; Swinney, 1981). This discrepancy is probably because of the fact that these studies usually presented participants with the word in question (e.g., *beans* in *spill the beans*), such that the word could not only be (pre)activated by the idiomatic context but was also processed from the input in a bottom–up fashion. In contrast, we substituted words with semantically related or unrelated words, such that in the anomalous conditions the word in question was never actually presented, which likely emphasizes top–down effects such as anticipation. This explanation is supported by the observation that, in the previously discussed study by Peterson et al. (2001), the final words of the idioms were also not presented but had to be produced by the participants themselves. In line with our results, this study did not report evidence for literal word meaning activation. This suggests that literal word meaning activation in idioms stems mainly from the bottom–up processing of words, that is, from the words actually present in the text or speech stream. Top–down contextual expectations are not enough to yield measureable literal word meaning activation during idiom comprehension.

The results also differ from Liu et al. (2010), who replaced characters in Chinese idioms with, among others, synonyms and unrelated words. They observed a smaller N400 for synonyms than unrelated words, consistent with semantic expectancy even in idioms. However, Liu et al.

(2010) were investigating different issues and chose idioms that come across as more transparent than ours ("hiding a dagger behind one's smiles"), such that processing literal word meanings could have been helpful. Our study used highly opaque idioms as a strict test of whether literal word meanings are always processed.

The Representation of Idioms in the Brain

Similar to Liu et al. (2010) and Moreno et al. (2002), the N400 to semantic violations was followed by a late positivity (P600) in idioms but not in literal contexts. The P600 has previously been reported as a response to violations of agreement (Hagoort, Brown, & Groothusen, 1993) and of orthography (Münte, Heinze, Matzke, Wieringa, & Johannes, 1998) and different interpretations of the component exist in the literature (Kos, Vosse, van den Brink, & Hagoort, 2010; Bornkessel-Schlesewsky & Schlesewsky, 2008; Kuperberg, 2007; Kim & Osterhout, 2005; Hoeks, Stowe, & Doedens, 2004; Kolk, Chwilla, van Herten, & Oor, 2003). We explain this finding by assuming similar representations for words and idioms, consistent with configurations, lexical concept nodes, or superlemmas (Sprenger et al., 2006; Cutting & Bock, 1997; Cacciari & Tabossi, 1988). Because at some level of representation, both words and idioms are lexical items, both violations of agreement/orthography and word substitutions in idioms can be considered form violations within lexical items, thus eliciting the same ERP response. Unitary representations of idioms are further supported by the TF results, which suggested a lack of semantic unification in idioms, because if idioms are stored as a whole and retrieved as such from long-term memory, the individual word meanings need not be unified.

Note that a lack of semantic unification does not preclude that unification continues at other levels: The words within an idiom must be combined syntactically to recognize the idiom, and the recognized idiom as a whole must be integrated with its context semantically and syntactically.

Conclusions

To conclude, although it is clear that an important part of language comprehension consists of activating and combining word meanings, it appears that the extent to which these operations are carried out can vary across contexts. When reading predictable and opaque idiomatic expressions, for which literal word meanings are irrelevant, the processing of literal word meanings can to some extent be "switched off."

Acknowledgments

We thank Lonneke Bücken and Kevin Lam for running Experiment 1, the research assistants of the Neurobiology of Language Department for their assistance with running Experiment 2, Lilla Magyari for help with some of the figures, and Francesco Vespignani and three anonymous reviewers for comments on a previous version of this article.

Reprint requests should be sent to Joost Rommers, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands, or via e-mail: joost.rommers@mpi.nl.

REFERENCES

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1993). *The CELEX lexical database [CD-ROM]*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes, 26,* 1338–1367.

Bastiaansen, M. C. M., & Hagoort, P. (2006). Oscillatory neuronal dynamics during language comprehension. In C. Neuper & W. Klimesch (Eds.), *Event-related dynamics of brain oscillations* (pp. 179–196). Amsterdam: Elsevier.

Bastiaansen, M. C. M., Mazaheri, A., & Jensen, O. (2012). Beyond ERPs: Oscillatory neuronal dynamics. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components* (pp. 31–50). New York: Oxford University Press.

Bobrow, S. A., & Bell, S. M. (1973). On catching on to idiomatic expressions. *Memory & Cognition, 1,* 343–346.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews, 59,* 55–73.

Boulenger, V., Hauk, O., & Pulvermüller, F. (2009). Grasping ideas with the motor system: Semantic somatotopy in idiom comprehension. *Cerebral Cortex, 19,* 1905–1914.

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language, 27,* 668–683.

Chwilla, D. J., & Kolk, H. H. J. (2002). Three-step priming in lexical decision. *Memory & Cognition, 30,* 217–225.

Colombo, L. (1993). The comprehension of ambiguous idioms in context. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure and interpretation* (pp. 163–189). Hillsdale, NJ: Erlbaum.

Cutting, J. C., & Bock, K. (1997). That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory & Cognition, 25,* 57–71.

de Grauwe, S., Swain, A., Holcomb, P. J., Ditman, T., & Kuperberg, G. R. (2010). Electrophysiological insights into the processing of nominal metaphors. *Neuropsychologia, 48,* 1965–1984.

de Groot, A. M. B., & de Bil, J. M. (1987). *Nederlandse woordassociatienormen met reactietijden*. Lisse: Swets & Zeitlinger.

de Groot, H. (1999). *Van Dale Idioomwoordenboek: Verklaring en herkomst van uitdrukkingen en gezegden*. Amsterdam: The Reader's Digest NV.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44,* 491–505.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41,* 469–495.

Gibbs, R. W., Jr. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition, 8,* 149–156.

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences, 9,* 416–423.

Hagoort, P., Brown, C. M., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes, 8,* 439–483.

Hagoort, P., Hald, L. A., Bastiaansen, M. C. M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304,* 438–441.

Hald, L. A., Bastiaansen, M. C. M., & Hagoort, P. (2006). EEG theta and gamma responses to semantic violations in online sentence processing. *Brain and Language, 96,* 90–105.
Hillert, D., & Swinney, D. A. (2001). The processing of fixed expressions during sentence comprehension. In A. Cienki (Ed.), *Conceptual structure, discourse, and language* (pp. 107–121). Stanford: CSLI.
Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research, 19,* 59–73.
Jackendoff, R. (1995). The boundaries of the lexicon. In M. Everaert, E.-J. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 133–165). Hillsdale, NJ: Erlbaum.
Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language, 52,* 205–225.
Kolk, H. H. J., Chwilla, D. J., van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language, 85,* 1–36.
Kos, M., Vosse, T. G., van den Brink, D., & Hagoort, P. (2010). About edible restaurants: Conflicts between syntax and semantics as revealed by ERPs. *Frontiers in Psychology, 1,* E222.
Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146,* 23–49.
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.
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207,* 203–205.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.
Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience, 9,* 920–933.
Laurent, J. P., Denhieres, G., Passerieux, C., Iakimova, G., & Hardy-Bayle, M. C. (2006). On understanding idiomatic language: The salience hypothesis assessed by ERPs. *Brain Research, 1068,* 151–160.
Lauteslager, M., Schaap, T., & Schievels, D. (1986). *Schriftelijke woordassociatienormen voor 549 zelfstandige naamwoorden*. Lisse: Swets & Zeitlinger.
Liu, Y., Li, P., Shu, H., Zhang, Q., & Chen, L. (2010). Structure and meaning in Chinese: An ERP study of idioms. *Journal of Neurolinguistics, 23,* 615–630.
Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164,* 177–190.
Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90,* 227–234.
Mitra, P. P., & Pesaran, B. (1999). Analysis of dynamic brain imaging data. *Biophysical Journal, 76,* 691–708.
Molinaro, N., & Carreiras, M. (2010). Electrophysiological evidence of interaction between contextual expectation and semantic integration during the processing of collocations. *Biological Psychology, 83,* 176–190.
Moreno, E. M., Federmeier, K. D., & Kutas, M. (2002). Switching languages, switching palabras (words): An electrophysiological study of code switching. *Brain and Language, 80,* 188–207.
Münte, T. F., Heinze, H.-J., Matzke, M., Wieringa, B. M., & Johannes, S. (1998). Brain potentials and syntactic violations revisited: No evidence for specificity of the syntactic positive shift. *Neuropsychologia, 36,* 217–226.
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,* 156869.
Peña, M., & Melloni, L. (2012). Brain oscillations during spoken sentence processing. *Journal of Cognitive Neuroscience, 24,* 1149–1164.
Peterson, R. R., Burgess, C., Dell, G. S., & Eberhard, K. M. (2001). Dissociation between syntactic and semantic processing during idiom comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27,* 1223–1237.
Proverbio, A. M., Crotti, N., Zani, A., & Adorni, R. (2009). The role of left and right hemispheres in the comprehension of idiomatic language: An electrical neuroimaging study. *BMC Neuroscience, 10,* 116.
Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41,* 416–426.
Raposo, A., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2009). Modulation of motor and premotor cortices by actions, action words and action sentences. *Neuropsychologia, 47,* 388–396.
Schwartz, A. I., & Kroll, J. F. (2006). Bilingual lexical activation in sentence context. *Journal of Memory and Language, 55,* 197–212.
Sprenger, S. A. (2003). *Fixed expressions and the production of idioms*. PhD thesis, MPI Series in Psycholinguistics, 21, University of Nijmegen.
Sprenger, S. A., Levelt, W. J. M., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language, 54,* 161–184.
Swinney, D. A. (1981). Lexical processing during sentence comprehension: Effect of higher order constraints and implications for representation. In T. Meyers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 201–209). Amsterdam: North-Holland (Advances in Psychology Series).
Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior, 18,* 523–534.
Urrutia, M., de Vega, M., & Bastiaansen, M. (2012). Understanding counterfactuals in discourse modulates ERP and oscillatory gamma rhythms in the EEG. *Brain Research, 1455,* 40–55.
van Loon-Vervoorn, W. A., & van Bekkum, I. J. (1991). *Woordassociatielexicon*. Amsterdam: Swets & Zeitlinger.
Vespignani, F., Canal, P., Molinaro, N., Fonda, S., & Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *Journal of Cognitive Neuroscience, 2,* 1682–1700.
Wang, L., Zhu, Z., & Bastiaansen, M. (2012). Integration or predictability? A further specification of the functional role of gamma oscillations in language comprehension. *Frontiers in Psychology, 3,* article 187.
Werning, M., Hinzen, W., & Machery, M. (Eds.) (2011). *The Oxford handbook of compositionality*. Oxford, UK: Oxford University Press.