Dropping Beans or Spilling Secrets: How Idiomatic Context Bias Affects Prediction

Manon Hendriks*, Wendy van Ginkel*, Ton Dijkstra, and Vitória Piai

Abstract

■ Idioms can have both a literal interpretation and a figurative interpretation (e.g., to "kick the bucket"). Which interpretation should be activated can be disambiguated by a preceding context (e.g., "The old man was sick. He kicked the bucket."). We investigated whether the idiomatic and literal uses of idioms have different predictive properties when the idiom has been biased toward a literal or figurative sentence interpretation. EEG was recorded as participants performed a lexical decision task on idiom-final words in biased idioms and literal (compositional) sentences. Targets in idioms were identified faster in both figuratively and literally used idioms than in compositional sentences. Time–frequency analysis of a prestimulus interval revealed relatively more alpha–beta power decreases in literally than figuratively used idiomatic sequences and compositional sentences. We argue that lexico-semantic retrieval plays a larger role in literally than figuratively biased idioms, as retrieval of the word meaning is less relevant in the latter and the word form has to be matched to a template. The results are interpreted in terms of context integration and word retrieval and have implications for models of language processing and predictive processing in general. ■

INTRODUCTION

Predicting upcoming events in perception and in social interactions plays an important role in establishing fluent and appropriate behavior in the world (see Huettig, 2015, for a review of studies). This insight motivates the ongoing research interest in the phenomenon of prediction in the perception and action domains of cognitive neuroscience (e.g., Clark, 2013; Sebanz, Knoblich, & Prinz, 2003). Remarkably, studies on predictive performance have been more pessimistic with respect to language processing, because, as has been argued, the number of upcoming words that could be predicted in sentences is staggering (e.g., Jackendoff, 2007).

Although recent psycholinguistic studies have collected evidence in favor of predictive processing, there has been a fierce debate as to its validity and consequences (e.g., Lau, Holcomb, & Kuperberg, 2013; DeLong, Urbach, & Kutas, 2005) and even how "prediction" should be defined. For the purposes of the current article, we define prediction as the (pre)activation of conceptual or word information stored in long-term memory before the appearance of that information in the linguistic input stream. This will likely involve accumulation of evidence over time until enough evidence is collected that a certain conceptual item or word form will follow, when preactivation occurs.

Evidence for predictive processing in language is sometimes inconclusive. For instance, in an ERP study manipulating the expectancy of upcoming words, DeLong et al. (2005) provided evidence in favor of probabilistic preactivation of word forms in sentence context. Unfortunately, in a large replication study across nine labortories, Nieuwland et al. (2018) were unable to replicate the prediction effect, weakening "the view that listeners routinely pre-activate the phonological forms of predictable words." However, Yan, Kuperberg, and Jaeger (2017) argue that a reanalysis of Nieuwland et al. using a surprisal measure rather than probabilities is in favor of anticipatory "semantic" predictions. Martin, Branzi, and Bar (2018) hold that prediction in language comprehension, measured by ERPs and the N400 ERP component in particular, might actually be the result of production processes. Without taking such a strong theoretical position, we can at least describe such predictions as involving retrieval of lexical–semantic information from memory before the start of the upcoming event (Jafarpour, Piai, Lin, & Knight, 2017; Piai, Roelofs, Rommers, Dahlslätt, & Maris, 2015).

On the whole, these studies suggest that predictive processing in sentence comprehension does indeed occur. However, they also indicate that predictive processing is not an all-or-none process of a single kind. In fact, prediction in sentence comprehension could entail the availability or preactivation of word form information, meaning information, or both, before a predicted word is actually presented in the sentence. Moreover, prediction might not always be equally strong; it might be strongest in case there is only one, highly expected sentence continuation. In many other cases, information activated on the basis of

Radboud University, Nijmegen, The Netherlands
*The first two authors contributed equally to this work.

https://doi.org/10.1162/jocn_a_01798

the preceding sentence context might arrive too late or be too weak to directly affect activation and retrieval of the upcoming word. We would refer to the later combination of information sources of sentence and target word as integration rather than prediction.

This analysis further suggests that finding evidence for predictive processing may also be dependent on the experimental paradigm and stimulus materials used. In this study, we argue that comparing the processing of idioms versus literal sentences is an optimally suited alternative for investigating this issue. Idioms consist of relatively fixed and therefore quite predictable sequences of words that can be placed in rather natural discourse contexts. Consider, for instance, the following sentence pair: "The farmer was old. He kicked the bucket" and "The farmer was angry. He kicked the bucket." Readers may predict the word "bucket" in either sentence pair rather than just integrate the word after it is recognized, depending on probabilistic continuation. However, if predictive processes differ for idiomatic expressions when they are biased toward a figurative or literal sentence interpretation, their "fingerprint" in the associated brain waves might well be different. In this article, we consider this issue using the EEG, in particular oscillatory activity and ERPs, and behavioral measures. Before zooming in on our own study, we will briefly summarize recent key studies on idiomatic processing.

Rommers, Dijkstra, and Bastiaansen (2013) investigated to what extent upcoming constituent word meanings are activated during idiom comprehension. The meanings of individual words in an idiom are theoretically unnecessary to comprehend the figurative meaning of the expression. In fact, they might even be detrimental to processing, for instance, in opaque idioms where the figurative meaning of the idiom as a whole is not easily or directly derived from its constituent words. In two experiments combining behavioral and electrophysiological measures, participants' brain activity was measured in response to words completing an idiomatic or literal sentence. Target words were either the correct, expected completion of the sentence; a semantic associate of the expected word; or an unrelated word. In both RT and ERP results, a graded pattern emerged for the literal sentences: The correct word showed most semantic expectancy and fastest responses, followed by the semantic associate and then the unrelated target. However, for the idiom sentences, there was no graded pattern: The correct target was processed fastest and showed most semantic activation, but there was no difference between the semantic associate and the unrelated target. These findings indicate that, at least in opaque idioms in a biasing context, literal word processing is suppressed by the presence of an idiomatic expression. In other words, there is a lack of semantic expectancy in idiomatic contexts, and prediction might instead be oriented toward word form. Because Rommers et al. measured the participants' brain responses in time intervals simultaneous with target word presentation, their effects can be interpreted as evidence that expectations affect EEG signals and RTs, but it is unwarranted to conclude that prediction already accrued before the target appeared. In other words, on the basis of their results, no difference can be made between predictive and integrative processes (note this was not the intention of Rommers et al. anyway). In this study, we therefore performed time–frequency analyses on a time window before the target word appeared. Any effects arising in this interval could be ascribed to prediction rather than integration.

Canal, Pesciarelli, Vespigniani, Molinaro, and Cacciari (2016) considered how one and the same idiomatic phrase was interpreted either figuratively or literally in contexts that bias either meaning. Matched control sentences were created where the idiom-final word was presented in isolation in a literal sentence. Using EEG, Canal et al. found no N400 differences between literal and idiomatic meanings at the last word in an idiomatic expression, but they did observe amplitude differences in the post-N400 positivity (PNP). The PNP has been associated with sentence reanalysis mechanisms and is thought to be modulated by prediction accuracy and context plausibility of the upcoming word string (Brothers, Swaab, & Traxler, 2015). Words near the end of idioms embedded in a figurative context were found to elicit a larger PNP than the same words in the same idioms embedded in a literal context (i.e., when used literally). Interpreting these findings in terms of sentence reanalysis, the authors concluded that idiom processing may be more cognitively demanding, especially when idioms can also be used as literal word strings. It must be noted that all idioms used by Canal et al. were highly plausible in their literal sense. This may have hindered their figurative interpretation, especially if they were highly transparent (van Ginkel & Dijkstra, 2019). However, the transparency of their idioms (and its interaction with literal plausibility) was not controlled or tested for. Thus, it remains unclear whether their results generalize to different types of idioms.

Furthermore, Canal et al. examined oscillations, showing differences in power specifically in the middle gamma frequency band (50–70 Hz) between idiomatic and literal uses of idioms. In literal context sentences, an increase in gamma power was observed that may be reflective of successful sentence processing or of a match between a predicted word and the characteristics of the incoming word (Lewis & Bastiaansen, 2015; Monsalve, Pérez, & Molinaro, 2014; Penolazzi, Angrilli, & Job, 2009). This increase of power in the gamma band was absent in the idiomatic context compared to the literal context, suggesting that processing of the idiom string in the idiomatic context occurs at a lower level than that of the same string used in a literal context. This finding is also in line with Rommers et al. (2013), who reported semantic unification to be less engaged in idiom processing than in literal sentence processing.

Molinaro, Monsalve, and Lizarazu (2016) compared the processing of words at the end of multiword units (e.g., "on the other hand") with processing of the same words at the end of literal, yet highly semantically constrained, sequences. Measures based on the prestimulus interval of the sentence (e.g., before actual presentation of the word completing the multiword-unit or literal sentence) revealed relatively more anterior beta-band power decreases in the multiword-unit condition (i.e., figurative condition) than in the literal condition. The authors interpreted this finding as possibly reflecting engagement of a more detailed preparation process in the multiword-unit condition as compared to the compositional condition. In other words, prediction of the upcoming word is stronger when the word completes a multiword unit rather than a literal unit. The authors argue this is because, for the word in the literal condition, the prediction is more likely to be made for a "semantic field": An array of possible targets is preactivated, although cloze probability is matched. For the multiword unit, prediction is more deterministic in that the unit can be retrieved from memory as a whole with its associated continuation contained in the figurative unit.

Interestingly, the beta-band power decrease in the multiword condition was followed by a larger alpha-band power increase, which was absent in the compositional condition. Previous research has suggested that posterior alpha power increases reflect active functional inhibition of task-irrelevant information (e.g., Jensen, Gips, Bergmann, & Bonnefond, 2014). Following this proposal, the presence of alpha power increases in the multiword-unit condition could reflect the inhibition of processing the target words in this condition as compared to the compositional condition. The strength of the prediction of the final word in the figurative multiword-unit condition cancels the need for detailed processing or encoding of the target word, whereas in the compositional condition, such detailed encoding is still engaged. In other words, encoding of the word in the figurative condition may be shallow in comparison to the compositional condition.

In summary, it may be suggested that deriving figurative and literal interpretations of an idiomatic expression involves different processes that are sensitive to sentence context. In a biasing context, such differences are reflected in ERP (e.g., N400) and oscillatory effects (e.g., alpha–beta band power decrease) in the figurative and literal conditions. In this study, we therefore examined both ERPs and oscillations for evidence of biasing context on ease and speed of figurative versus literal interpretation of idioms.

This Study

To assess if prediction has different characteristics for idioms and literally interpreted sentences in different biasing discourse contexts, we used the following research paradigm. Participants were presented with Dutch sentence pairs like "The farmer was old. He kicked the bucket" and "The farmer was angry. He kicked the bucket." The second sentence in these pairs consisted of a phrase that could be interpreted either figuratively or literally. The first sentence biased either the figurative or literal interpretation of the second sentence. In half of the sentences, the last word of the second sentence was replaced by a pseudoword. Participants had to decide as quickly and accurately as possible if the last item of the second sentence was an existing word or a pseudoword (lexical decision). In a control condition, we substituted some words of the idiomatic sequence by other words that were matched in length and frequency, so the sequence was not formulaic anymore. This sequence was then preceded by an appropriate context sentence, as in the sentence pair "The child was playing. He kicked the marble." Hypotheses were formulated to test if predictions differ for figurative and literally interpreted sentences in a biasing context. Our study focused on the time–frequency domain of the EEG in the various conditions. If predictions take place during idiomatic processing depending on the biasing context, this might become visible in the brain waves of the participants in the time window before the target actually appears. In terms of time–frequency analysis, we will consider predictions for idiomatic effects in both the gamma frequency band (50–70 Hz) and the alpha–beta frequency bands (8–30 Hz). With respect to the gamma frequency band, we hypothesized a relative power decrease in the figurative compared to the literal context condition in line with previous research. Gamma frequency band power is thought to reflect semantic unification processes (Rommers et al., 2013). Note that the conclusions from Rommers et al. are based on measurements from stimulus onset onward, whereas we are measuring a prestimulus interval. However, semantic unification may still be less involved in the figurative sentences if prediction of the final word is based on template matching of the form rather than word semantics: The idiom is already recognized as such and retrieved from memory before presentation of the final word, so semantic unification processes may already be less involved in the prestimulus interval (Canal et al., 2016).

However, as discussed above, the alpha and beta bands are equally of interest when considering idiom processing. Beta-band activity has been associated with prediction mechanisms in multiple areas of human cognition and action, such as the motor and visual domains and, critically, the language domain (e.g., Weiss & Mueller, 2012; Jenkinson & Brown, 2011; Engel & Fries, 2010). In particular, when comparing sentences that bias a final target word (e.g., "The farmer milked the …," target: cow) with neutral sentences (e.g., "The child drew a …"), relative alpha and beta power decreases have been consistently found in the pretarget interval (Piai, Rommers, & Knight, 2018; Rommers, Dickson, Norton, Wlotko, & Federmeier, 2017; Piai, Roelofs, Rommers, Dahlsätt, et al., 2015; Piai, Roelofs, & Maris, 2014). Following an alternative (and not mutually exclusive) interpretation, a relative power decrease in the alpha and beta bands reflects the retrieval of complex conceptual representations (e.g., during prediction; Hanslmayr, Staresina, & Bowman, 2016; Hanslmayr, Staudigl, & Fellner, 2012; for lexical–semantic retrieval in the language domain, see Piai et al., 2018; Piai, Roelofs, Rommers, & Maris, 2015).

In summary, if there is a lack of semantic expectancy in the figurative context condition, participants may only be processing the word form, that is, template matching the word. Prediction will then be limited to the word form, as semantic information would not be retrieved in the prestimulus interval. Therefore, we hypothesized less power decrease in the alpha–beta frequency band in the figurative than in the literal context condition.

Next, we analyzed the N400 effects in the ERPs for evidence that sentence context influences the processing of the correct target word. Available evidence suggests there is an inverse relationship between N400 amplitude and cloze probability (Lau et al., 2013; DeLong et al., 2005). We define cloze probability, or the degree to which the upcoming item is expected in offline questionnaires, as a form of word predictability: Words that are highly expected may show predictive processing before occurring in the input stream, when preactivation of the word may (partially) occur before it appears. In our study, we therefore expect smaller N400 effects to arise for correct targets in the figurative and literal conditions than in the control condition, because cloze probability is high in the first two conditions but lower in the control condition. In line with previous literature (reviewed in Kutas, Van Petten, & Kluender, 2006), we expect to find a larger N400 amplitude for pseudoword than for word targets in all context conditions. Note that the correct word should inherently be more expected than any pseudoword.

Finally, as a check on the sensitivity of the paradigm we applied, we tested the behavioral data for the presence of basic effects of condition and lexical status. First, we expect shorter RTs to target words in the figurative and literal context conditions relative to the control sentence, because in both conditions, the word continuation should be strongly expected on the basis of a high cloze probability. In contrast, no RT differences were expected for correct target words in the figurative and literal context conditions, precisely because target words in the two conditions are high in cloze probability.

METHODS

Participants

Twenty-four students from Radboud University Nijmegen and the HAN University of Applied Sciences (mean age = 23.25 years, 18 women) gave informed consent and received course credit or monetary reward 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 or language disorders. Two participants were excluded from behavioral analyses because of poor performance in the task, therefore reducing the sample size to 22 participants.

Materials and Design

Experimental materials consisted of 55 Dutch idioms that could be used in both a figurative sentence and a literally biasing sentence. Idiom selection was based on a database of Dutch idioms developed by the Idiomatic Second Language Acquisition Group of Radboud University Nijmegen (Hubers, van Ginkel, Cucchiarini, Strik, & Dijkstra, 2018). Idioms were selected for pretesting if they (1) contained no Dutch–English cognates; (2) consisted of at least three words; (3) ended in a noun, adjective, or preposition when put into a sentence; and (4) could easily be interpreted both figuratively and literally. This left us with a set of 203 idioms. These idioms were provided with figurative and literal biasing context sentences and extensively pretested (see Appendix A). Rating studies led to the final selection of 55 idioms that were highly familiar and frequent, where the target sentence logically followed the context sentence and where both sentences had a natural feel. Furthermore, the control sentences were matched to the literal target sentences in terms of link and naturalness ratings (for link: p = .511, for naturalness: p = .292; see Table 1 for values).

Participants performed a Dutch lexical decision task (LDT) on the last word of each target sentence/idiom, which required them to indicate as fast as possible whether this word was an existing Dutch word or a pseudoword. Each of the 55 idioms was presented twice to each participant: in a literally biasing context and in a figuratively biasing context. Furthermore, matched control sentences were created that were fully literal. Presentation of targets was counterbalanced across participants; for example, if a participant saw the existing target word in the figurative context condition, they would see a pseudoword in the literal context condition. In total, the figurative and literal context sentences combined with the matched control sentences made for an even number of 166 trials per participant.

Table 1. Mean (and Standard Deviation) of Cloze Probability (0–1) and Link and Naturalness Ratings on a Scale from 1 (*Very Low*) to 7 (*Very High*) for Figuratively Used Idioms (idiom-FIG), Literally Used Idioms (idiom-LIT), and Literal Compositional Sentences (lit-CON)

Context	*Cloze Probability*	*Link*	*Naturalness*
idiom-FIG	0.83 (0.20)	5.67 (0.56)	4.80 (0.68)
idiom-LIT	0.74 (0.30)	4.92 (0.64)	3.90 (0.83)
lit-CON	0.26 (0.32)	5.00 (0.78)	4.06 (0.87)

In half of the trials (83 trials), the target word was replaced by a nonexisting word. This pseudoword could be either similar or dissimilar to the original existing word. The pseudowords did not exist in the English or Dutch language but were created by substituting roughly a third of the letters of the existing word for similar pseudowords and two thirds of the letters in dissimilar pseudowords.

In the 166 experimental trials, there were 25 idiomatic context sentences with existing words as targets, 25 literal context sentences with words as targets, 33 control sentences with existing words as targets, 30 idiomatic context sentences with pseudowords as targets (half similar, half dissimilar), 30 literal context sentences with pseudowords as targets (half similar, half dissimilar), and 23 control sentences with pseudowords as targets (half similar, half dissimilar). In total, 50% of the items in each experimental list were words and 50% were pseudowords. To illustrate the conditions in the experiment, Table 2 provides the example of the Dutch idiom *op je tenen lopen* (English translation: "to walk on your toes"), which means "wanting to achieve more than you can handle."

Procedure

All participants were tested individually in a soundproof, electrically shielded room. The experiment was programmed in Psychopy (Peirce et al., 2011) and presented on a computer screen. RTs were recorded via a dedicated button box developed by the Donders Centre for Cognition (BitsiBox).

Participants received Dutch written instructions before giving their informed consent. Participants were presented with printed sentences in rapid serial visual presentation, preceded by context sentences. Their task was to decide as fast as possible whether or not the last word of the target sentence (presented in yellow on a black background) was an existing Dutch word by pressing one of two designated buttons on the button box with their left hand (see Piai et al., 2015). Half of the participants

Table 2. Overview of the Different Manipulations Used in This Experiment with Literal English Translations

Condition	*Example*		*Target Word*
Existing words			
idiom-FIG	Wendy heeft het ontzettend druk.	Ze loopt op haar	tenen.
	Wendy is very busy.	*She is walking on her*	*toes.*
idiom-LIT	Wendy wil graag groter lijken dan ze is.	Ze loopt op haar	tenen.
	Wendy wants to look taller than she is.	*She is walking on her*	*toes.*
lit-CON	Mia is eigenaresse van een café.	Ze werkt in haar	kroeg.
	Mia is the owner of a café.	*She is working in her*	*bar.*
Similar pseudowords			
idiom-FIG	Wendy heeft het ontzettend druk.	Ze loopt op haar	teben.
	Wendy is very busy.	*She is walking on her*	*boes.*
idiom-LIT	Wendy wil graag groter lijken dan ze is.	Ze loopt op haar	teben.
	Wendy wants to look taller than she is.	*She is walking on her*	*boes.*
lit-CON	Mia is eigenaresse van een café.	Ze werkt in haar	kroog.
	Mia is the owner of a café.	*She is working in her*	*bor.*
Dissimilar pseudowords			
idiom-FIG	Wendy heeft het ontzettend druk.	Ze loopt op haar	paven
	Wendy is very busy.	*She is walking on her*	*waas.*
idiom-LIT	Wendy wil graag groter lijken dan ze is.	Ze loopt op haar	paven
	Wendy wants to look taller than she is.	*She is walking on her*	*waas.*
lit-CON	Mia is eigenaresse van een café.	Ze werkt in haar	spoog.
	Mia is the owner of a café.	*She is working in her*	*tir.*

The meaning of the idiom "to walk on your toes" is "wanting to achieve more than you can handle." Examples are given for figuratively used idioms (idiom-FIG), literally used idioms (idiom-LIT), and literal compositional sentences (lit-CON).

responded to words by pressing the left button (with their left middle finger); and the other half, by pressing the right button (with their left index finger). This was reversed for the pseudowords.

The experiment started with a practice block of 12 trials. Each trial began with a fixation cross (+) lasting for 750 msec. Then, the context sentence appeared in full on the screen. All words were presented in a white color on a black background in Arial font (font size: 28.5). After participants read the entire sentence, they pressed a random button of their choice to continue to the target sentence. The target sentence appeared on the screen word by word. Each word remained on the screen for 350 msec, alternated with a blank screen of 300 msec between each word. The final word was presented in a different color (yellow), was 1.5 times bigger than the other words (font size: 42.5), and was indicated with a dot. These were cues to indicate that the participant should perform the LDT on this word.

The experiment was divided into three blocks with breaks between blocks, each consisting of a mix of the idiom context conditions and control sentences. For a particular idiom, each context was positioned in a different block. Participants processed a pseudorandomized list of items for which they never had to press the same button more than three times in a row. Furthermore, the order of blocks was randomized across participants. After each block, participants were allowed to take a break for as long as they wanted. The session as a whole, including capping for EEG, took approximately 120 min. Afterward, participants were presented with a multiple-choice questionnaire regarding the meaning of the idioms presented in the LDT to ensure participants were familiar with the respective idioms.

ERP Data Recording and Preprocessing

The EEG signals were recorded from 64 Ag–AgCl active electrodes, of which 62 were mounted in a cap (ActiCAP 64Ch; Brain Products), and referenced online to the left mastoid. Two separate electrodes were placed on the left and right mastoids. The ground electrode was placed at the AFz location. Four passive electrodes were placed above and beneath the left eye, and at each outer canthus, to measure eye blinks and horizontal eye movements, respectively. The ground electrode for the passive electrodes was placed on the tip of the nose. Electrode impedance was kept below 15 kΩ. Participants were asked to blink only during the presentation of the context sentence, to keep the number of eye blinks to a minimum in the time frame of interest.

Before EEG analyses were conducted, the data were rereferenced offline to the average of the left and right mastoids. The continuous EEG signal was segmented into epochs of 2250 msec, lasting from 950 msec before the onset of the target word until 1300 msec after word onset. The linear trend was removed from the data per trial. Before statistical analysis, all trials were excluded where participants were unfamiliar with the meaning of the idiom (as assessed through a multiple-choice test). In addition, all trials with incorrect responses on the LDT were removed from both the EEG and behavioral analyses.

Preprocessing of the data was performed with the Fieldtrip software package, an open-source MATLAB toolbox for neuropsysiological data analysis (Oostenveld, Fries, Maris, & Schoffelen, 2011). First, an independent component analysis was performed to identify and remove components related to eye blinks and muscle activity. Afterward, bad channels were identified, and the signal was replaced with the interpolated activity from the surrounding channels. Finally, trial outliers were removed after visual inspection. Approximately 4% of the trials were rejected on this basis, and the number of rejected trials was comparable across conditions ($F = 0.06, p = .94$).

RESULTS

Time–Frequency Analysis

Time-resolved spectra were computed using a Hanning taper of length equals to 3 cycles of each frequency being estimated (2–70 Hz). The taper was advanced in 1-Hz frequency steps and 10-msec time steps. Power estimates were averaged across trials for each context condition (figurative, literal, control) separately for each participant.

We used cluster-based permutation tests (Maris & Oostenveld, 2007) to assess the differences between conditions in a way that naturally takes care of the multiple comparisons problem by identifying clusters of significant differences between conditions in the time, space, and frequency dimensions. The statistical tests were performed for the alpha–beta range (8–30 Hz; Piai et al., 2018; Piai, Roelofs, Rommers, Dahlslätt, et al., 2015; Piai, Roelofs, & Maris, 2014; Rommers et al., 2013) and gamma range (50–70 Hz) separately. All available channels were entered in the statistical analyses, but given that the hypotheses were specific to the pretarget stimulus interval, the time window analyzed was −300 to 0 msec, or the blank screen between presentation of the penultimate word and final word of the sentence. First, an F test was performed to compare across the three context conditions (i.e., control, literal, and figurative). If the F test was significant, showing sensitivity to the experimental manipulation as a whole, paired-samples t tests were conducted to compare the three levels of the context condition in a pairwise manner. Monte Carlo p values were calculated on the basis of 1000 random permutations.

Power in the alpha–beta frequency range was sensitive to the manipulation of sentence context (F test, Monte Carlo: $p = .02$; see Figure 1). Pairwise comparisons showed less alpha–beta power in the literally used idiom condition (idiom-LIT) than in the compositional control (lit-CON) condition (Monte Carlo $p = .01$) in the

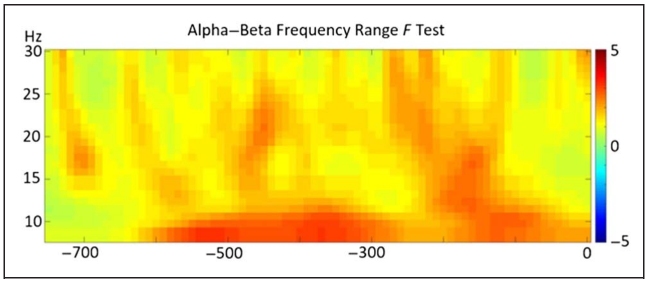

Figure 1. Visualization of the *F* test in the alpha–beta frequency range (8–30 Hz).

prestimulus interval examined. The comparison between the figuratively used idiom condition (idiom-FIG) and the lit-CON condition did not yield a significant effect (Monte Carlo $p = .27$), despite vast differences in cloze probability scores between these conditions. Comparing literally and figuratively used idioms, we found more alpha–beta power decreases for the idiom-LIT condition as compared to the idiom-FIG condition (Monte Carlo $p = .04$). For a visualization of the pairwise comparisons, see Figure 2.

When examining gamma power, the Monte Carlo *F* test did not show a significant effect ($p = .13$) in the hypothesized frequency range between 50 and 70 Hz. Therefore, individual contrasts were not examined. For visualization of the *F* test in a 50- to 70-Hz range, see Figure 3. To exclude the possibility that gamma power was found in lower gamma frequency ranges, we conducted an *F* test in the 30- to 50-Hz range as well. This Monte Carlo *F* test did not show a significant effect either ($p = .11$).

ERP Analysis

The single-trial epochs were averaged per condition and participant. No baseline correction or filtering was further applied. As the inclusion of a baseline correction is quite common and the validity of excluding such a correction may be subject of debate, we ran additional analyses on the ERPs after baseline correcting the signal using the segment of −950 to 0 msec. Appendix Figure A1 presents the comparison between the ERPs with and without baseline correction, which were virtually identical to each other. As expected, the inferential statistics was virtually identical between the ERPs with and without baseline correction, as also reported in the Appendix A. The hypotheses regarding the ERPs were tested using cluster-based permutation tests (Maris & Oostenveld, 2007). All available channels were entered in the statistical analyses, but given that the

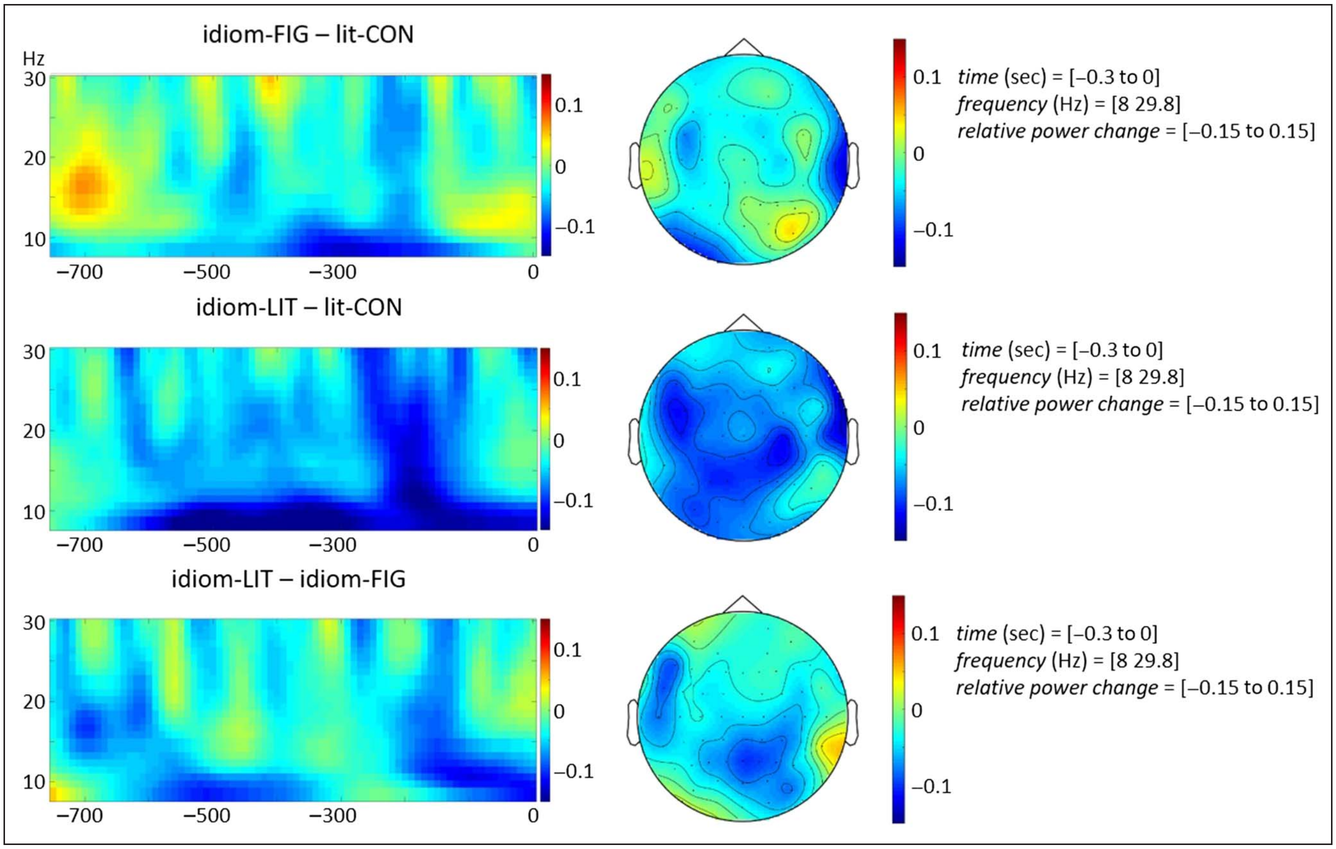

Figure 2. Comparison between time–frequency representations of the power changes for figuratively biased idioms (idiom-FIG), literally biased idioms (idiom-LIT), and literal control condition (lit-CON), with their associated topographies. The time–frequency representations are shown for the average over all channels associated with the significant cluster. For the nonsignificant contrast, all channels were used for the average. The topographies show the distribution of the differences across the scalp and indicate frequencies between 8 and 30 Hz and in a time range from −0.30 to 0 sec before target onset for the idiom-FIG – lit-CON, idiom-LIT – lit-CON, and idiom-LIT – idiom-FIG contrasts.

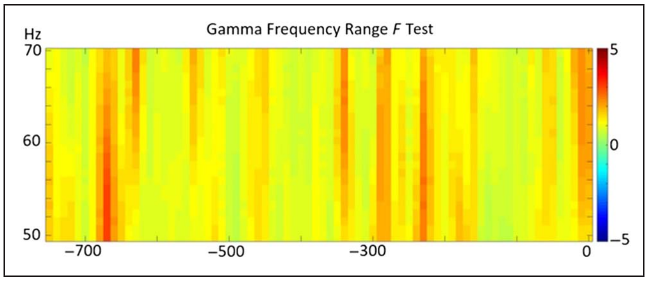

Figure 3. Visualization of the (nonsignificant) F test in the gamma frequency range (50–70 Hz).

hypotheses were specific to the N400 component, the time window analyzed was 250–650 msec. First, an F test was performed on the real words to compare across the three context conditions (i.e., control, literal, and figurative). Paired-samples t tests were then used to compare the levels of the context condition pairwise. To examine the lexical status effect, the ERPs were averaged across the three context conditions for words and pseudowords separately, and paired-samples t tests were used to compare them.

Post hoc, an ERP analysis was conducted over the time point of −950 to 0 msec (presentation of the target word), to ensure that no relevant ERP differences were caused by aspects of the sentence (such as determiners, possessives, etc.) before presentation of the target. The F test indicated no significant differences across the three conditions in this period (Monte Carlo $p = .574$). Although the F test was not significant, we ran paired tests as an additional check. None of these tests came back significant (all Monte Carlo $ps > .613$). Therefore, we are confident that no aspects of the sentences before the target word have confounded the ERP analyses of the critical target word summarized below.

Figure 4A shows the ERPs time-locked to the onset of the target words (only existing words) for each context condition. Testing for an ERP effect in the latency range from 250 to 650 msec poststimulus, the cluster-based permutation test revealed a significant difference on the real target words across the three context conditions idiom-FIG, idiom-LIT, and lit-CON (F test, Monte Carlo $p < .001$). In the pairwise tests, a difference was observed between the idiom-LIT and lit-CON conditions and between the idiom-FIG and lit-CON conditions (both Monte Carlo $ps = .002$), but there was no statistically significant difference between the idiom-FIG and idiom-LIT conditions (Monte Carlo $p = 1$). In this latency range, the difference between the idiom-LIT and idiom-FIG conditions relative to the lit-CON condition was most pronounced over centro-posterior channels (averaged activity over 250–490 msec), as shown in the topographical maps in Figure 4A.

Figure 4B shows the ERPs time-locked to the onset of the target stimulus for real words and pseudowords, collapsed over context type. Testing for an ERP effect in the latency range from 250 to 650 msec poststimulus, the cluster-based permutation test revealed a significant positive cluster (i.e., larger amplitude for real words vs. pseudowords; Monte Carlo $p = .002$) and a significant negative cluster (i.e., larger amplitude for pseudowords vs. real words; Monte Carlo $p = .040$). The positive cluster was most pronounced in the time window of 250–590 msec; and the negative cluster, in the time window of 460–650 msec. The difference between the real words and the pseudowords was most pronounced over central channels, as shown in the topographical map in Figure 4B, for the activity averaged over 250–590 msec.

Behavioral Results

Behavioral analyses were conducted only on trials with correct responses to the LDT. If a participant incorrectly answered the multiple-choice question on the meaning of an idiom, that idiom's trials were removed from analysis for that participant as he or she was unfamiliar with the idiom's meaning. In total, 5% of all trials were rejected following these criteria. The number of rejected trials was comparable across conditions ($F = 1.961$, $p = .141$), with 1083 trials in the idiom-FIG condition, 1074 trials in the idiom-LIT condition, and 1050 trials in the CON condition. Two participants were then excluded for having average RTs more than 2.5 *SD*s from the mean of all participants. Three idioms were then excluded from analysis because they led to a large number of errors (18%–21%: "een held op sokken zijn," "iemand blij maken met een dode mus," and "de mussen vallen van het dak"). Finally, individual trials above 2.5 *SD*s of the mean per participant were

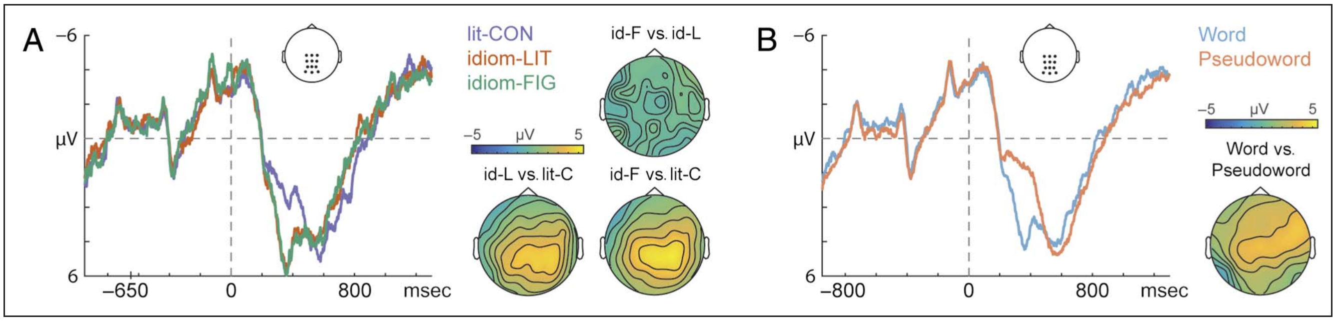

Figure 4. ERPs for the literally biased idioms (idiom-LIT), figuratively biased idioms (idiom-FIG), and literal control sentences (lit-CON). On the left, A depicts the ERPs for the contrasts between experimental conditions. The −650-msec mark corresponds to the onset of the penultimate word of the sentence. On the right, B depicts the ERP comparison between existing words and pseudowords across all experimental conditions.

Table 3. Mean RTs in Milliseconds and Error Rate per Context Condition (Standard Deviation between Parentheses)

	Words		*Pseudowords*	
Context	*RT*	*Error Rate*	*RT*	*Error Rate*
idiom-FIG	536 (161)	.04 (.20)	634 (163)	.05 (.22)
idiom-LIT	549 (155)	.04 (.18)	645 (143)	.05 (.22)
lit-CON	642 (143)	.07 (.26)	692 (136)	.06 (.20)

removed for each participant. The number of rejected trials was comparable between context conditions ($F = 0.046, p = .955$). In total, 19.02% of the data was removed before analysis, of which 5.5% was because of the exclusion of the three idioms with higher error rates.

Linear mixed effects model regression analyses were conducted in Rstudio (*lmerTest* package in R, Version 3.4.1; R Project for Statistical Computing, Vienna, Austira). Mean RTs and error rates are summarized in Table 3 for both existing words and pseudowords. For the behavioral analysis, we included responses to existing words only as

Table 4. Results for the Releveled Linear Mixed Effects Regression Analysis Subdivided for the Three Levels of the Model for Simple Effects

	Estimate	*SE*	*df*	*t Value*	*p Value*
Overall context effects					
idiom-FIG vs. lit-CON	.07312	.02453	1088	2.981	.003
idiom-LIT vs. lit-CON	.09130	.02221	1103	4.111	.000
idiom-FIG vs. idiom-LIT	−.01818	.02003	1452	−0.908	.364
Simple effects in the lit-CON context					
Cloze probability	−.1441	.03911	7614	−3.684	.000
Word Frequency	−.008071	.02377	3569	−0.340	.734
Cloze × Word Frequency	.03484	.04546	6761	0.766	.444
Simple effects in the idiom-FIG context					
Cloze probability	−.3484	.05652	9609	−6.164	.000
Word Frequency	−.02333	.02044	3660	−1.142	.254
Cloze × Word Frequency	−.04362	.06482	7876	−0.673	.501
Simple effects in the idiom-LIT context					
Cloze probability	−.2390	.04201	8407	−5.690	.000
Word Frequency	.01583	.01586	2349	0.998	.319
Cloze × Word Frequency	.03349	.04283	7207	0.782	.435
Context × Cloze Probability					
(idiom-FIG)lit-CON × Cloze	.2043	.06828	9369	2.992	.003
(idiom-LIT)lit-CON × Cloze	.09495	.05810	7314	1.634	.103
(idiom-FIG)idiom-LIT × Cloze	.1094	.06880	1119	1.590	.112
Context × Word Frequency					
(idiom-FIG)lit-CON × Word Frequency	.01526	.02903	1080	0.526	.599
(idiom-LIT)lit-CON × Word Frequency	−.02390	.02664	9540	−0.897	.370
(idiom-FIG)idiom-LIT × Word Frequency	.03916	.02308	1410	1.697	.090

Interaction effects for context conditions with cloze probability or word frequency are listed under the respective headers.

we were interested in how the correct target word was processed across different context conditions. Furthermore, we were interested in the effect of cloze probability, and the cloze probability of pseudowords is inherently zero. Model selection began with a theoretically maximal model including predictors at the level of the target and the idiom. At the word level, these were word frequency, word length, and cloze probability. At the idiom level, several predictors were taken into account for the idiom-bearing sentences (Hubers et al., 2018): usage rates, subjective frequency measures, imageability, transparency, and familiarity scores. Insignificant interaction terms and predictors were removed from the model in an iterative manner, with each model tested against its predecessor in an ANOVA and the most explanatory model being selected to proceed with. The final model took the log-transformed RTs as the dependent variable and included a random slope for participant over trial number to take into account trial order effects as well as a random slope for item at the level of the idiom. Fixed effects consisted of a three-way interaction between Condition (idiom-FIG, idiom-LIT, lit-CON), Cloze probability (centered), and Word frequency as well as a fixed effect for Trial order.

All p values reported are given by the lmerTest statistics package. Comparisons between conditions were examined by releveling the Context conditions factor to change the condition on the intercept and to allow for comparisons within the same linear mixed effects regression model. The results of this analysis are summarized in Table 4 for all the levels of the model. As there were no significant interaction effects with Word frequency between any of the context conditions (all ps $> .2$), these comparisons are not listed in the table for brevity.

Targets in the FIG context condition were identified as words faster than targets in the lit-CON condition ($p = .003$), but there was no difference between the idiom-FIG and idiom-LIT conditions. RTs to targets in the LIT condition were faster than those in the lit-CON condition ($p < .001$).

In all three context conditions, RTs to targets were faster if their cloze probability was higher (all ps $< .001$). There was an interaction effect of cloze probability and context condition that revealed that the effect of cloze probability differed between the idiom-FIG and lit-CON contexts ($p < .01$). There was no difference between the idiom-FIG and idiom-LIT conditions or between the idiom-LIT and lit-CON conditions. This effect showed that the facilitation of RTs because of higher cloze probability in the idiom-FIG condition was significantly larger than that in the lit-CON condition.

Error Analysis

Table 3 reports means and standard deviations for the error analysis. A binary logistic regression run on correctness of judgments did not yield differences in error rates between any of the experimental conditions overall. There was also no difference in accuracy for pseudowords and existing words in the idiom-FIG and idiom-LIT contexts, but in the lit-CON context, pseudowords were rejected slightly more reliably than existing words were accepted as words (estimate $= -.7769$, $SE = .3081$, $Z = -2.521$, $p = .012$).

DISCUSSION

In this study, we hypothesized that prediction processes for idiomatic and literal sentences differ from each other. Prediction was considered to be the (pre)activation of conceptual or word information stored in long-term memory before the appearance of that information in the linguistic input stream. We tested our hypothesis by examining whether behavioral and electrophysiological manifestations of prediction would differ between figuratively and literally biased idioms and literal (compositional) sentences. We examined the EEG signal in terms of both ERPs and time–frequency modulations as these two measures are known to capture different aspects of brain activity and have been shown to dissociate under certain circumstances (see Piai, Roelofs, Jensen, Schoffelen, & Bonnefond, 2014; Laaksonen, Kujala, Hultén, Liljeström, & Salmelin, 2012; Davidson & Indefrey, 2007).

In the time–frequency domain, we examined idiomatic effects in both the alpha–beta (8–30 Hz) and gamma (30–70 Hz) frequency bands in the pretarget interval. With respect to the alpha–beta band, an analysis of the EEG data indicated differences in predictive processes between conditions. More power decreases were found in the literally used idiom condition than in the figuratively used idiom and control conditions. There was no difference in power between the figuratively used idioms and the control sentences. Under the hypothesis that power decrease in the alpha and beta bands is sensitive to prediction, our finding suggests that prediction played a bigger role during the literally interpreted idioms in the interval immediately preceding the target word. In particular, assuming that alpha–beta power decrease reflects lexical–semantic retrieval (Piai, Roelofs, Rommers, Dahlslätt, et al., 2015; Piai, Roelofs, Rommers, & Maris, 2015), there might be less semantic and/or lexical activation in the interval immediately preceding the target word during the figurative use of idiom sentences compared to the literal use of these sentences, as well as compared with weakly biasing contexts. We hypothesize that these differences may arise because a literally biased idiom is processed in two ways: both literally and figuratively, as the phrase may still be recognized as an idiom. Therefore, more predictive processes are at play here than in figuratively biased idioms or fully literal sentences. In contrast, fully literal sentences are only processed one way (literally), and figuratively biased idioms are primarily being processed in a figurative way (however, some superficial literal word processing may still be necessary). In terms of lexical–semantic retrieval, this means that a literally biased idiom affords more information to be retrieved than in the case in the other two conditions. Future research should examine these potential differences in underlying processing. As of yet, a mechanistic theory linking alpha–beta power decreases to lexical–semantic retrieval is lacking (for developments in this direction, refer to Piai & Zheng, 2019; Meyer, 2018). We note that the current study used strongly biasing contexts to bias the interpretation of a sentence that has a potential idiomatic interpretation as either a literal or figurative sentence. We controlled for subjective frequency of these idioms (how often participants in a large survey reported encountering the idiom themselves). In the current study, we did not consider objective frequency such as frequency counts of the idioms in corpora, as these counts differ vastly between the different types of corpora examined (e.g., newspaper corpora, spoken language corpora, corpora of Internet text, provided largely differing measures). Especially when considering idiomatic sequences presented without a strongly biasing context, how often the sentence is used figuratively and literally in corpora can potentially be a relevant factor to consider, along with how often the sequences are used in their figurative and literal forms in these corpora.

Next, we analyzed gamma band frequency effects between 50–70 and 30–50 Hz. On the basis of previous research (e.g., Canal et al., 2016; Rommers et al., 2013), we hypothesized a power decrease in the figuratively compared to the literally used idioms, reflecting increased semantic unification in literal versus figurative language. Whereas Rommers et al. reported this pattern for literal versus figurative sentences in a time window after the onset of a stimulus, Canal et al. showed similar effects for literally used idioms versus figuratively used idioms in a prestimulus interval, suggesting that semantic unification is less involved in figuratively used idioms before presentation of the idiom-final word. However, in our study, we found no effects in the gamma frequency bands (50–70 and 30–50 Hz). The extent to which gamma band effects are informative about processing of idiomatic expressions should be examined in future studies.

With respect to the EEG, we also analyzed if sentence context influences the processing of the correct target word in terms of the N400. Earlier studies suggest there is an inverse relationship between N400 amplitude and cloze probability (Lau et al., 2013; DeLong et al., 2005). Because cloze probability in our study was high in both the figurative and literal conditions and was lower in the control condition (see above), we expected smaller N400 effects for correct targets in the figurative and literal conditions than in the control condition. Indeed, we observed amplitude differences for these comparisons. We further hypothesized that an N400 difference would arise between the figurative and literal conditions if encoding of the word in the figurative condition would be relatively shallow in comparison to the literal condition (in line with Molinaro et al., 2016). However, there was no significant ERP difference between the two conditions. It is possible that any effect was too subtle in light of differences in the cloze probability of individual sentences, but we cannot be sure as we found no difference. We also found amplitude differences between pseudowords and real-word targets in all context conditions, in line with previous literature (reviewed in Kutas et al., 2006). We note that our ERP results were measured on the last word of the sentence, which may induce wrap-up effects as the sentence is fully processed. However, the time–frequency effects observed in the alpha–beta band were measured on a pretarget interval and showed differences between the conditions. In addition, our N400 results followed the expected pattern given the N400 literature, indicating that any wrap-up effects were not confounded in our study.

Behavioral analyses confirmed the sensitivity of our paradigm, revealing basic effects of condition and lexical status. As expected, in all conditions, RTs were shorter for words than for pseudowords. Finally, we observed shorter RTs when cloze probability was high. RTs were shorter to target words in both the figurative and literal idiom context conditions than in the control condition, likely because cloze probability was lower in the latter condition. This facilitation effect is in line with previous studies in which faster responses were found to formulaic sequences compared to compositional sequences (Siyanova-Chanturia, Conklin, & Schmitt, 2011; Conklin & Schmitt, 2008). There were no significant RT differences between target words in the figurative and literal context conditions.

In summary, the results of the behavioral and ERP analyses attest to the sensitivity of our manipulations, providing a solid ground for interpreting the time–frequency effects in the time window just before the target appears. Measures time-locked to target word onset (RTs and ERPs) follow the pattern of cloze probability, where targets in conditions with a higher cloze probability are processed faster and show more semantic expectancy than targets in the lower cloze probability control condition. Crucially, time–frequency results measured before target-word onset revealed a different pattern, discordant with cloze probability, suggesting that predictions differ as a function of the type of sentence context. In particular, we found evidence for stronger lexical–semantic retrieval in the interval preceding target onset in the literally biased idiom condition than in a weakly biasing literal sentence context or a strongly biasing figurative context. We note that the trials used in the analysis for the behavioral and ERP results were not always the same: Because of our stringent selection criteria on the quality of the trials coupled with a relatively low number of trials per condition, removing trials based on the criteria of both types of analyses left too few trials for adequate power.

Interestingly, in our study, the figurative interpretation of the idiomatic expressions did not differ from the control condition in terms of alpha–beta oscillations. One explanation for this finding is that the idiom might already be recognized earlier, necessitating only a very superficial processing of the (word form of the) target word. This interpretation needs to be tested in future research by considering the temporal aspects of activation in figuratively and literally interpreted sentences in more detail. Experimental manipulations could try to slow down or speed up the relative availability of information on the figurative or literal interpretation of the idioms. For instance, a stronger context manipulation leading to higher cloze probabilities might affect the temporal availability of form and meaning information.

Conclusion

We examined predictive and integrative processing of words contained within idioms that were biased toward a literal or figurative interpretation by a preceding context sentence. Measures after presentation of the idiom-final word (ERP and RT), when this word was available for integration, showed patterns following cloze probability. Measures before presentation of the idiom-final word deviated from cloze probability patterns. Differences in alpha–beta band power showed evidence for stronger semantic word retrieval in the literally used idioms compared to the figuratively used idioms and fully literal (compositional) sentences. In contrast, no such differences were found between figuratively biased idioms and compositional sentences, despite vast differences in cloze probability. We interpret these findings as reflecting the type of prediction (preactivation of lexical information) that is made in figurative versus literal language. A word completing a literally used idiom is subject to semantic retrieval before it is presented, but the same word completing a figuratively used idiom may only be subject to a process of "template matching" where the word form is matched to the expected word form once the word is encountered.

APPENDIX A

This appendix contains the pretests and piloting of stimulus materials.

Pretest 1: Rating Study

Ratings on subjective frequency ("how often have you seen or heard this expression," scale: 1–5, with 1 being the least often) and familiarity ("how familiar are you with the meaning of this expression," scale: 1–5, with 1 being the least familiar) were available for 30 of the selected idioms. For another 55 of our selected idioms, 30 participants filled out a rating study consisting of one of two randomly distributed lists. Idioms with ratings of at least 3.5 of 5 on both frequency and familiarity were selected for the next phase. This left us with 83 idioms for the next validation round.

Pretest 2: Sentence Selection and Validation

For each of the 83 idioms, we created a literal sentence and an idiomatic context sentence. This sentence was followed by a target sentence containing the idiom (see Table 2 in the main text). The target sentence was kept identical for both conditions. Furthermore, each idiom was converted into a control target sentence. The structure of the literal target sentence was kept, but some words (matched to the original words in length and frequency as presented in the Subtlex-NL database [Keuleers, Brysbaert, & New, 2010]) were replaced to produce an unrelated but plausible, literal alternative. Furthermore, we created a control context sentence to precede each control target sentence. Next, all context and target sentence pairs were pseudorandomly distributed across three different lists, and participants viewed one of three lists. Each idiom appeared only once in each list.

The items were then rated by 66 independent participants in an online survey. First, we estimated the cloze probability of each item as an operationalization of predictability. Participants processed the context and target sentences with the final word of the target sentence omitted. They then completed the sentence with the first word that they could think of, without trying to be original. In a second task, participants provided two judgments on a 7-point Likert scale: the semantic link between the context sentence and the target sentence ("How well does the target sentence relate to the context sentence?") and the naturalness of the items ("Would you ever use or encounter this sentence in your daily life?"). Items were selected only if they received a rating of at least 3.8 on the semantic link between context and target sentence, as we wanted the target sentence to be a logical continuation of the context. Furthermore, as the control items were based on the literal condition, we made sure the ratings on the link and naturalness did not differ significantly between the literal and control items in a dependent t test (for link: p = .511, for naturalness: p = .292). The cloze probability was allowed to differ between conditions (p < .01), as this is an inherent feature of our selected context conditions. For this same reason, ratings between figurative and literal sentences were allowed to be different, as long as they surpassed the minimal threshold rating (for link and naturalness: p < .01, cloze probability: p = .037). For the means of the dimensions tested in the pretests, see Table A1.

Table A1. Mean of Cloze Probability (0–1) and Link and Naturalness Ratings on a Scale from 1 (*Very Low*) to 7 (*Very High*)

Context	*Cloze Probability*	*Link*	*Naturalness*
idiom-FIG	0.83 (0.20)	5.67 (0.56)	4.80 (0.68)
idiom-LIT	0.74 (0.30)	4.92 (0.64)	3.90 (0.83)
lit-CON	0.26 (0.32)	5.00 (0.78)	4.06 (0.87)

Standard deviation is listed in parentheses.

Pilot

The remaining 56 idioms (168 items) were subjected to a pilot study in which 25 participants took part. On the basis of the results, we decided to discard the idiom "een klein hartje hebben" (literal translation: "to have a small heart") as too few people gave the correct meaning of this idiom. Thus, we conducted the actual experiment with 55 idioms.

APPENDIX B: ERP ANALYSIS WITH BASELINE CORRECTION

Testing for an ERP effect in the latency range from 250 to 650 msec poststimulus, the cluster-based permutation test revealed a significant difference on the real target words across the three context conditions idiom-FIG, idiom-LIT, and lit-CON (*F* test Monte Carlo $p < .010$). In the pairwise tests, a difference was observed between the idiom-LIT and lit-CON conditions, and between the idiom-FIG and the lit-CON conditions (both Monte Carlo $ps = .002$), but no statistically significant difference between the idiom-FIG and idiom-LIT conditions (Monte Carlo $p = 1$).

For real words and pseudowords, collapsed over context type, testing for an ERP effect in the latency range from 250 to 650 msec poststimulus, the cluster-based permutation test revealed a significant positive cluster (i.e., larger amplitude for real words vs pseudowords, Monte Carlo $p = .002$).

Acknowledgments

This research was funded as part of the research program Free Competition in the Humanities with Project Number 23000349 NWO ISLA FdL, financed by the Netherlands Organisation for Scientific Research (NWO) and approved by the Ethical Committee of the Faculty of Social Sciences at Radboud University Nijmegen (ECG2012-2711-059). The authors gratefully acknowledge the open access funding provided by the Donders Centre for Cognition of the Donders Institute for Brain, Cognition and Behaviour, Radboud University

Reprint requests should be sent to Wendy van Ginkel, Kapittelweg 29, Room B.00.026, 6525 EN Nijmegen, Radboud University, Nijmegen, The Netherlands, or via e-mail: W.vanGinkel@donders.ru.nl.

Funding Information

Research funded as part of the research program Free Competition in the Humanities with project number 23000349, financed by the Netherlands Organisation for Scientific Research (NWO).

Diversity in Citation Practices

A retrospective analysis of the citations in every article published in this journal from 2010 to 2020 has revealed a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated

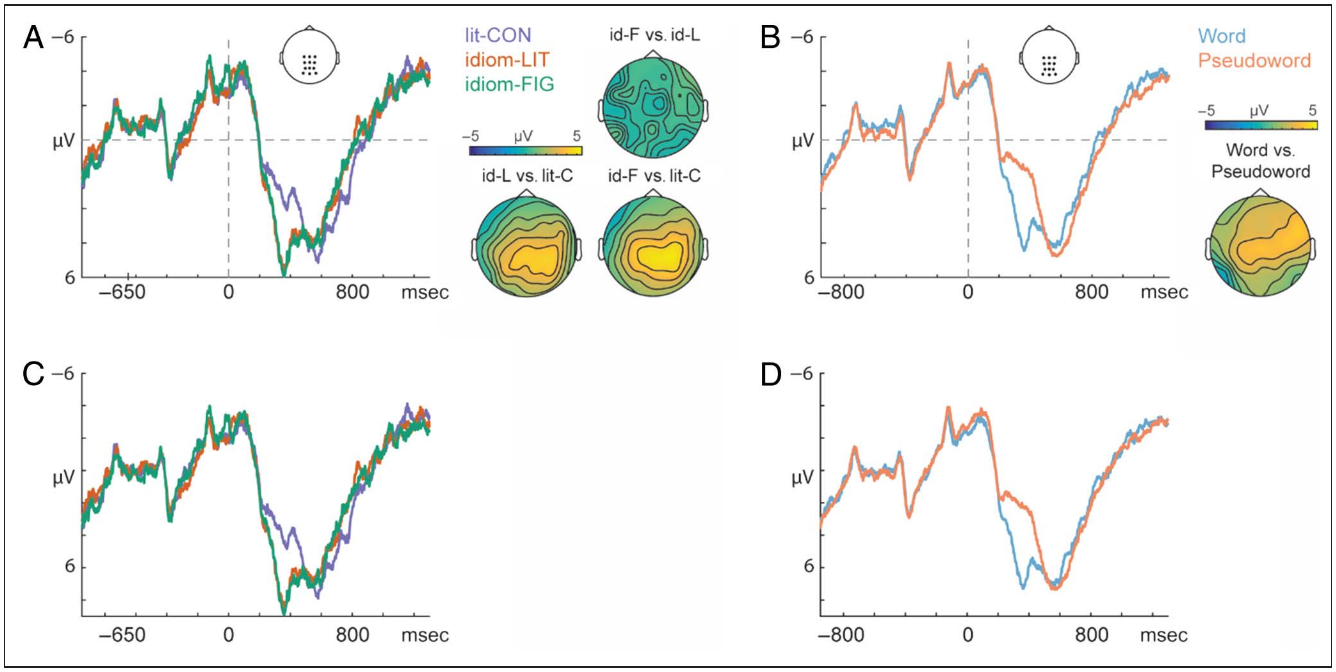

Figure A1. Demonstration of the EEG data from the original analysis without baseline correction (A and B) and with baseline corrections applied (C and D).

gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .408, W(oman)/M = .335, M/W = .108, and W/W = .149, the comparable proportions for the articles that these authorship teams cited were M/M = .579, W/M = .243, M/W = .102, and W/W = .076 (Fulvio et al., *JoCN*, 33:1, pp. 3–7). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

REFERENCES

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition*, *136*, 135–149. https://doi.org/10.1016/j.cognition.2014.10.017, PubMed: 25497522

Canal, P., Pesciarelli, F., Vespigniani, F., Molinaro, N., & Cacciari, C. (2016). Basic composition and enriched integration in idiom processing: An EEG study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 928–943. https://doi.org/10.1037/xlm0000351, PubMed: 28068127

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–204. https://doi.org/10.1017/S0140525X12000477

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. https://doi.org/10.1093/applin/amm022

Davidson, D. J., & Indefrey, P. (2007). An inverse relation between event-related and time–frequency violation responses in sentence processing. *Brain Research*, *1158*, 81–92. https://doi.org/10.1016/j.brainres.2007.04.082, PubMed: 17560965

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*, 1117. https://doi.org/10.1038/nn1504, PubMed: 16007080

Engel, A. K., & Fries, P. (2010). Beta-band oscillations—Signalling the status quo? *Current Opinion in Neurobiology*, *20*, 156–165. https://doi.org/10.1016/j.conb.2010.02.015, PubMed: 20359884

Hanslmayr, S., Staresina, B. P., & Bowman, H. (2016). Oscillations and episodic memory: Addressing the synchronization/desynchronization conundrum. *Trends in Neurosciences*, *39*, 16–25. https://doi.org/10.1016/j.tins.2015.11.004, PubMed: 26763659

Hanslmayr, S., Staudigl, T., & Fellner, M. C. (2012). Oscillatory power decreases and long-term memory: The information via desynchronization hypothesis. *Frontiers in Human Neuroscience*, *6*, 74. https://doi.org/10.3389/fnhum.2012.00074, PubMed: 22514527

Hubers, F., van Ginkel, W., Cucchiarini, C., Strik, H., & Dijkstra, T. (2018). Normative data on Dutch idiomatic expressions: Native speakers. DANS [Dataset]. https://doi.org/10.17026/dans-zjx-hnsk

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014, PubMed: 25708148

Jackendoff, R. (2007). A parallel architecture perspective on language processing. *Brain Research*, *1146*, 2–22. https://doi.org/10.1016/j.brainres.2006.08.111, PubMed: 17045978

Jafarpour, A., Piai, V., Lin, J. J., & Knight, R. T. (2017). Human hippocampal pre-activation predicts behavior. *Scientific Reports*, *7*, 5959. https://doi.org/10.1038/s41598-017-06477-5, PubMed: 28729738

Jenkinson, N., & Brown, P. (2011). New insights into the relationship between dopamine, beta oscillations and motor function. *Trends in Neurosciences*, *34*, 611–618. https://doi.org/10.1016/j.tins.2011.09.003, PubMed: 22018805

Jensen, O., Gips, B., Bergmann, T. O., & Bonnefond, M. (2014). Temporal coding organized by coupled alpha and gamma oscillations prioritize visual processing. *Trends in Neurosciences*, *37*, 357–369. https://doi.org/10.1016/j.tins.2014.04.001, PubMed: 24836381

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*, 643–650. https://doi.org/10.3758/BRM.42.3.643, PubMed: 20805586

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics electrified II (1994–2005). In M. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 659–724). Academic Press. https://doi.org/10.1016/B978-012369374-7/50018-3

Laaksonen, H., Kujala, J., Hultén, A., Liljeström, M., & Salmelin, R. (2012). MEG evoked responses and rhythmic activity provide spatiotemporally complementary measures of neural activity in language production. *Neuroimage*, *60*, 29–36. https://doi.org/10.1016/j.neuroimage.2011.11.087, PubMed: 22173296

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience*, *25*, 484–502. https://doi.org/10.1162/jocn_a_00328, PubMed: 23163410

Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, *68*, 155–168. https://doi.org/10.1016/j.cortex.2015.02.014, PubMed: 25840879

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024, PubMed: 17517438

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is production: The missing link between language production and comprehension. *Scientific Reports*, *8*, 1079. https://doi.org/10.1038/s41598-018-19499-4, PubMed: 29348611

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms. *European Journal of Neuroscience*, *48*, 2609–2621. https://doi.org/10.1111/ejn.13748, PubMed: 29055058

Molinaro, N., Monsalve, I. F., & Lizarazu, M. (2016). Is there a common oscillatory brain mechanism for producing and predicting language? *Language, Cognition and Neuroscience*, *31*, 145–158. https://doi.org/10.1080/23273798.2015.1077978

Monsalve, I. F., Pérez, A., & Molinaro, N. (2014). Item parameters dissociate between expectation formats: A regression analysis of time–frequency decomposed EEG data. *Frontiers in Psychology*, *5*, 847. https://doi.org/10.3389/fpsyg.2014.00847, PubMed: 25161630

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468, PubMed: 29631695

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. https://doi.org/10.1155/2011/156869, PubMed: 21253357

Peirce, J., Gray, J., Halchenko, Y., Britton, D., Rokem, A., & Strangman, G. (2011). PsychoPy—A psychology software in Python. https://buildmedia.readthedocs.org/media/pdf/psychopy-hoechenberger/latest/psychopyhoechenberger.pdf

Penolazzi, B., Angrilli, A., & Job, R. (2009). Gamma EEG activity induced by semantic violation during sentence reading. *Neuroscience Letters*, *465*, 74–78. https://doi.org/10.1016/j.neulet.2009.08.065, PubMed: 19723559

Piai, V., Roelofs, A., Jensen, O., Schoffelen, J. M., & Bonnefond, M. (2014). Distinct patterns of brain activity characterise lexical activation and competition in spoken word production. *PLoS One*, *9*, e88674. https://doi.org/10.1371/journal.pone.0088674, PubMed: 24558410

Piai, V., Roelofs, A., & Maris, E. (2014). Oscillatory brain responses in spoken word production reflect lexical frequency and sentential constraint. *Neuropsychologia*, *53*, 146–156. https://doi.org/10.1016/j.neuropsychologia.2013.11.014, PubMed: 24291513

Piai, V., Roelofs, A., Rommers, J., Dahlslätt, K., & Maris, E. (2015). Withholding planned speech is reflected in synchronized beta-band oscillations. *Frontiers in Human Neuroscience*, *9*, 549. https://doi.org/10.3389/fnhum.2015.00549, PubMed: 26528164

Piai, V., Roelofs, A., Rommers, J., & Maris, E. (2015). Beta oscillations reflect memory and motor aspects of spoken word production. *Human Brain Mapping*, *36*, 2767–2780. https://doi.org/10.1002/hbm.22806, PubMed: 25872756

Piai, V., Rommers, J., & Knight, R. T. (2018). Lesion evidence for a critical role of left posterior but not frontal areas in alpha–beta power decreases during context-driven word production. *European Journal of Neuroscience*, *48*, 2622–2629. https://doi.org/10.1111/ejn.13695

Piai, V., & Zheng, X. (2019). Speaking waves: Neuronal oscillations in language production. In K. D. Federmeier (Ed.), *Psychology of learning and motivation* (Vol. 71, pp. 265–302). Academic Press. https://doi.org/10.1016/bs.plm.2019.07.002

R Development Core Team. (2009). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rommers, J., Dickson, D. S., Norton, J. J., Wlotko, E. W., & Federmeier, K. D. (2017). Alpha and theta band dynamics related to sentential constraint and word expectancy. *Language, Cognition and Neuroscience*, *32*, 576–589. https://doi.org/10.1080/23273798.2016.1183799, PubMed: 28761896

Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience*, *25*, 762–776. https://doi.org/10.1162/jocn_a_00337, PubMed: 23249356

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: Just like one's own? *Cognition*, *88*, B11–B21. https://doi.org/10.1016/S0010-0277(03)00043-X, PubMed: 12804818

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, *27*, 251–272. https://doi.org/10.1177/0267658310382068

van Ginkel, W., & Dijkstra, T. (2019). The tug of war between an idiom's figurative and literal meanings: Evidence from native and bilingual speakers. *Bilingualism: Language and Cognition*, *23*, 131–147. https://doi.org/10.1017/S1366728918001219

Weiss, S., & Mueller, H. M. (2012). "Too many betas do not spoil the broth": The role of beta brain oscillations in language processing. *Frontiers in Psychology*, *3*, 201. https://doi.org/10.3389/fpsyg.2012.00201, PubMed: 22737138

Yan, S., Kuperberg, G. R., & Jaeger, T. F. (2017). Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and DeLong et al. (2005). *bioRxiv*, 143750.