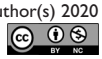
Article

Language *and* Speech

Language *and* Speech
2021, Vol. 64(3) 594–624



DOI: 10.1177/0023830920943625
journals.sagepub.com/home/las


Activation of Literal Word Meanings in Idioms: Evidence from Eye-tracking and ERP Experiments

Ruth Kessler, Andrea Weber and Claudia K. Friedrich
University of Tübingen, Germany

Abstract

How the language processing system handles formulaic language such as idioms is a matter of debate. We investigated the activation of constituent meanings by means of predictive processing in an eye-tracking experiment and in two ERP experiments (auditory and visual). In the eye-tracking experiment, German-speaking participants listened to idioms in which the final word was excised (*Hannes let the cat out of the . . .*). Well before the offset of these idiom fragments, participants fixated on the correct idiom completion (*bag*) more often than on unrelated distractors (*stomach*). Moreover, there was an early fixation bias towards semantic associates (*basket*) of the correct completion, which ended shortly after the offset of the fragment. In the ERP experiments, sentences (spoken or written) either contained complete idioms, or the final word of the idiom was replaced with a semantic associate or with an unrelated word. Across both modalities, ERPs reflected facilitated processing of correct completions across several regions of interest (ROIs) and time windows. Facilitation of semantic associates was only reliably evident in early components for auditory idiom processing. The ERP findings for spoken idioms compliment the eye-tracking data by pointing to early decompositional processing of idioms. It seems that in spoken idiom processing, holistic representations do not solely determine lexical processing.

Keywords

Idioms, eye-tracking, ERP, online processing

Introduction

There is an open debate in psycholinguistic research on whether and how formulaic sequences or multi-word expressions, as for example in collocations (*black coffee*), phrasal verbs (*dig into something*), or idioms (*kick the bucket*), are stored in the mental lexicon (for a review, see Conklin & Schmitt, 2012). In some accounts, the linguistic system is assumed to store formulaic sequences as

Corresponding author:
Ruth Kessler, Developmental Psychology, University of Tübingen, Schleichstraße 4, Tübingen, D-72076, Germany.
Email: ruth.kessler@uni-tuebingen.de

larger units and to process them holistically (e.g., Jackendoff, 2002; Swinney & Cutler, 1979; Wray, 2005). According to this account, formulaic sequences have their own lexical entry comparable to "long words." More recently, other accounts emphasize the internal syntactic and semantic structure of these multi-word expressions (e.g., Kyriacou, Conklin, & Thompson, 2020; Mancuso et al., 2020; Marantz, 2005; Snider & Arnon, 2012; Sprenger, Levelt, & Kempen, 2006; Tremblay & Baayen, 2010). While parsers are indeed sensitive to phrase frequencies, they access representations of all individual constituents in a phrase simultaneously (Arnon & Christiansen, 2017). According to these accounts, single constituents within multi-word units can be accessed separately.

In order to capture the hybrid nature of multi-word sequences, accounts of idiom processing have been proposed in which the structural properties of an idiom are preserved, while its meaning and form are also stored holistically. For example, the Configuration Hypothesis by Cacciari and Tabossi (1988) assumes that idioms are processed like novel, literal language, but only until the parser recognizes a phrase as an idiom. After this "idiom key," the parser directly retrieves the idiom configuration and associated meaning from the mental lexicon. According to a multidetermined view of idiom processing, factors such as familiarity or literal plausibility in addition to predictability determine the time point of recognition (Libben & Titone, 2008; Titone et al., 2019). Thus it is not surprising that in highly predictable idioms the recognition of the phrase can occur prior to the final word (Cacciari & Corradini, 2015). The Superlemma Hypothesis for speech production states that single word meanings within idiomatic expressions are necessarily activated (Sprenger et al., 2006). According to this view, idioms are accessible as both individual words (*simple lemmas*) and lexical units (*superlemmas*). In the present study, we investigated the neurocognitive reality of the representation of formulaic language in the mental lexicon by tracing the temporal dynamics of online activation of idiom constituent meaning.

The assumption of holistic processing is typically backed up by empirical evidence of greater processing ease for formulaic than for comparable non-formulaic language (e.g., Conklin & Schmitt, 2008; Gibbs, 1980; Siyanova-Chanturia, Conklin, & Schmitt, 2011; Strandburg et al., 1993; Swinney & Cutler, 1979; Tabossi, Fanari, & Wolf, 2009; Tremblay et al., 2011; Underwood, Schmitt, & Galpin, 2004). Several studies have found, for example, that participants read fixed multi-word expressions faster than novel phrases (e.g., Conklin & Schmitt, 2008; Tremblay et al., 2011), and that they fixate on words in idioms less extensively than on words in control sentences (Siyanova-Chanturia et al., 2011; Underwood et al., 2004). Based on holistic processing accounts, it has been argued that formulaic sequences are retrieved faster from the semantic memory than novel controls because there is no need for the parser to access single word meanings.

However, processing advantages might not originate exclusively from a purely holistic representation of formulaic phrases. They might also emerge from phrase frequency, predictive mechanisms for frequently co-occurring constituents, or phrase familiarity (e.g., Carrol & Conklin, 2020). For example, Canal et al. (2010) propose that predictive mechanisms within idioms are based on the knowledge of their specific lexical form in the mental lexicon and these predictive mechanisms might differ qualitatively from predictions within literal, non-formulaic expressions. Arguably, more direct evidence for holistic representations would show that the linguistic system does not process single words and their meanings separately, but receives multi-word expressions unanalyzed from the mental lexicon (see the discussion in, e.g., Siyanova-Chanturia, 2015).

Idioms are well suited for investigating holistic processing versus decomposition into single constituents of formulaic expressions. In many idioms, the figurative meaning cannot be inferred from the compositional meaning of the constituent words. For example, the figurative meaning of *to let the cat out of the bag* (to reveal a secret unintentionally) is not derived from the meaning of the single noun constituents (*bag* and *cat*) or from their combination with the verb (*to let*). Therefore, evidence that such multi-word idioms are obligatorily decomposed into their single

constituents would strongly speak against a model assuming solely holistic processing of idioms not allowing access to single words. One way to test whether single constituents within idioms are processed individually is to measure the activation of semantic associates (*basket*) of these constituents (*bag*). Because, in general, activation of a word in the mental lexicon will spread to semantically related words (Collins & Loftus, 1975), activation of semantic associates within idioms would indicate that the parser processes individual constituents.

Following the approach of spreading semantic activation, priming and word production studies have indeed shown that parsers have single word meanings available quickly during idiom processing (e.g., Beck & Weber, 2016; Smolka, Rabanus, & Rösler, 2007; Sprenger et al., 2006; van Ginkel & Dijkstra, 2019). In these studies, participants typically first read idioms (Rabanus et al., 2008; Smolka et al., 2007; van Ginkel & Dijkstra, 2019) or listened to idioms (Beck & Weber, 2016), such as *to pull someone's leg* (meaning "to spoof someone"), and subsequently performed a lexical decision task on immediately following written target words. Across these studies, participants responded faster to targets that were semantically related to the literal meaning of a constituent word (e.g., *walk*) compared to unrelated targets. In two experiments conducted by Sprenger and colleagues (2006, Experiment 2 and Experiment 3), participants read idiom fragments (e.g., *Jan liep tegen de [lamp]*, literally translated: *Jan walked against the [lamp]*, meaning "to get caught" in Dutch) and were asked to complete the idiom by speaking aloud the final, missing noun (e.g., *lamp*). Both experiments tested whether participants have semantic associates of idiom-final words (e.g., *candle*) available while they prepare their responses. In Experiment 2, participants received a spoken prime while they prepared their response. Semantic associates facilitated participants' responses compared to unrelated primes. In Experiment 3, participants were prompted to produce the idiom-final word when a question mark appeared on the screen. However, when another word appeared on the screen instead of the question mark, they had to switch the task and produce that word. In this production task, participants responded faster to semantic associates of the idiom-final constituent compared to unrelated probes.

Evidence for spreading semantic activation originating from single idiom constituents was also found in an eye-tracking study by Holsinger (2013). Participants listened to idiomatic phrases (*hit the hay*) while they saw four printed words on the screen, including an associate of a constituent word (*barn*). Fixations showed that participants considered the semantic associate more often than they considered unrelated distractors. Together, priming and eye-tracking results are in line with accounts assuming that the parser has idiom internal structures available (e.g., Marantz, 2005; Snider & Arnon, 2012; Sprenger et al., 2006; Tremblay & Baayen, 2010).

In contrast to priming and eye-tracking work, data from an event-related potentials (ERP) study found no apparent involvement of single word meanings during idiom processing (Rommers, Dijkstra, & Bastiaansen, 2013). In this experiment participants read highly predictable Dutch idioms (e.g., literally translated *to walk against the lamp*). In a related condition, a semantic associate replaced the idiom's final noun (*candle*), and in an unrelated condition, an unrelated word replaced the final noun (*fish*). Semantic associates of idiom-final nouns did not elicit different ERPs than completely unrelated words did (see Experiment 2, for further discussion of the specific ERP effects elicited in this study). Rommers and colleagues argued that participants did not form semantic predictions of idiom-final constituents. Results rather indicated holistic processing of idioms, as would be suggested by representational accounts viewing idioms as "large words" (Jackendoff, 2002) or "lexical items" (Swinney & Cutler, 1979), which are processed as a whole.

Design-related differences (such as modality, paradigm, and idiom characteristics) in previous studies might account for the mixed results regarding the processing of idiom constituents. For example, the modality in which idioms were presented differed between experiments and this comes with different amounts of linguistic information available to participants at any given point

in time. While connected spoken language makes single words only sequentially available (as they are unfolding over time), written language makes complete words or phrases available at once. Using spoken idioms combined with written probes, Beck and Weber (2016) and Holsinger (2013) found semantic activation of single idiom constituents. Other studies presented idioms and probes visually, either phrase-wise (Sprenger et al., 2006, Experiment 3) or word-by-word (Rabanus et al., 2008; Rommers et al., 2013; Smolka et al., 2007). The experiments by Sprenger et al. (2006, Experiments 2 and 3) using phrase-wise presentation, where the whole idiom fragment was available at once, revealed semantic activation of the idiom constituent. Results were mixed for experiments using word-by-word presentation in a rapid serial sequence (Rabanus et al., 2008; Rommers et al., 2013; Smolka et al., 2007). Clearly, the time course of word recognition and semantic activation might differ depending on the amount of linguistic information available at a certain point in time (e.g., Anderson & Holcomb, 1995; Van Petten et al., 1999) and this might play a role in processing differences found across different studies.

Different experimental paradigms could also relate to different results. In most studies that support decomposition of idioms (Beck & Weber, 2016; Holsinger, 2013; Rabanus et al., 2008; Smolka et al., 2007), activation of semantically related words might have resulted from bottom-up spreading activation, due to the critical idiom constituent being actually presented. For example, the eye-tracking study by Holsinger (2013) reported biased eye movements towards semantic associates (*barn*) shortly after the participants heard the critical idiom constituent (*hay*) as part of the idiom. Similarly, the critical constituent was part of the primes in priming studies showing semantic activation (Rabanus et al., 2008; Smolka et al., 2007; van Ginkel & Dijkstra, 2019). In these studies, the critical idiom constituent might have briefly activated semantic associates in a bottom-up fashion without the idiom representation being involved. In contrast, participants were not presented with the critical idiom constituent (*lamp*) in the ERP study by Rommers et al. (2013), which did not find evidence for activation of semantic associates (*candle*). According to the authors of the latter study, the prediction of the correct idiom-final word might not be sufficient to activate single word meanings within idioms and, thus, no processing benefit for semantically related words was found. However, while critical idiom constituents were also not presented in the production study by Sprenger et al. (2006, Experiments 2 and 3), these authors did find that facilitation of semantically related words was induced merely by planning to produce the idiom-final constituent.

Finally, experiments differed in terms of idiom characteristics such as predictability. Depending on the amount of given linguistic constraints, individual idioms can be recognized prior to their last constituent (Libben & Titone, 2008). Earlier versus later activation of the idiomatic form might result in higher versus lower predictability of the idiom-final word (Canal et al., 2010). According to the Configuration Hypothesis (Cacciari & Tabossi, 1988), predictability might affect the activation of literal constituent meanings. Since in highly predictable idioms the idiom key should be well before the final constituent, literal activation of the latter would be less likely. Nevertheless, Rabanus et al. (2008), Rommers et al. (2013), Smolka et al. (2007), and Sprenger et al. (2006) measured lexical activation of highly predicted, idiom-final constituents and came to different conclusions. Taken together, different idioms used across different studies render comparisons of results obtained with different paradigms and presentation modalities difficult.

In the present study, we targeted the previously obtained inconsistencies regarding literal meaning activation of single idiom constituents. To this end, (a) we varied presentation modality by presenting idioms and probes cross-modally (Experiment 1), auditorily (Experiment 2), and visually (Experiment 3), (b) we focused on top-down prediction mechanisms, for example by not presenting the critical constituent in the input in order to discourage pure bottom-up spreading of semantic information (Experiments 2 and 3), and (c) we kept the idiom characteristics constant by using the same highly predictable idioms across experiments. Furthermore, we employed different

implicit methods by relying on eye-tracking (Experiment 1) and ERPs (Experiments 2 and 3) measures. Implicit online measures might be more sensitive in detecting spreading semantic activation (Heil, Rolke, & Pecchinenda, 2004).

2 Experiment 1

In Experiment 1, we addressed the question of semantic activation of idiom constituent meanings through predictive processing by conducting an eye-tracking study. We exploited the tendency of gaze behavior (e.g., time course and amount of fixations) to be biased towards implicit linguistic aspects of displayed words (for a review, see Huettig, Rommers, & Meyer, 2011). Fixation biases include semantic associates of target words as reflected, for example, in more fixations towards the printed word *shark* while the word *turtle* is mentioned (Huettig & McQueen, 2011). These results imply that eye movements are a powerful tool to investigate bottom-up spreading semantic activation exerted by spoken input.

In the eye-tracking study on idiom processing by Holsinger (2013), participants' eye movements were attracted by semantic associates of idiom constituents while they listened to the idiom containing the respective constituent. For example, while listening to *hay* in *hit the hay*, participants fixated the printed word *BARN* more often than unrelated control words. That is, the design of this former study does not allow disentangling rapid bottom-up semantic spread exerted by the presentation of the single word and decomposition of the idiom during processing. In order to study the latter, we have to rely on a paradigm that does not present the critical idiom constituent in the input.

In order to avoid presentation of the critical idiom constituent, we exploited predictive processing in online comprehension. Numerous eye-tracking studies have shown that participants use sentence contexts to predict upcoming words and their semantic properties (Altmann & Kamide, 1999, 2007; Kamide, Altmann, & Haywood, 2003). For instance, when participants listened to a sentence such as *the boy will eat the cake*, they fixated on the picture of a cake in a visual scene at the offset of the verb *eat* (Altmann & Kamide, 1999). That is, eye fixations reflect predictions built during online processing before the critical word can exert bottom-up semantic spread. Therefore, prediction of semantic features for idiom constituents that are not part of the input can indicate decomposable memory traces for idioms.

In order to investigate prediction within idioms, we measured predictive fixations to displayed words before the full idiom has been heard and processed. In Experiment 1, we used highly predictive German idiomatic phrases. Participants listened to incomplete idioms, missing the final critical word, without any biasing context (e.g., *Hannes ließ die Katze aus dem . . .*, "Hannes let the cat out of the . . ."). Visual displays included four printed words: the correct idiomatic completion (*SACK*, "BAG"), a semantic associate of the correct completion (*KORB*, "BASKET"), and two unrelated distractors, with a semantic relation to each other (*ARM*, "ARM" and *BAUCH*, "STOMACH"). Participants had to choose which of the displayed words was the correct completion of the idiomatic phrase. In order to fixate the correct item, participants had to anticipate the complete idiom. This should result in fixations of the correct completions. Fixation biases to correct completion will be informative about idiom recognition. If semantic associates of single idiom constituents are available for predictive processing, this would be indicated in fixations to semantic associates of the correct completion as soon as the idiom is recognized. This would support word-by-word predictions based on decomposable memory traces for idioms. If semantic associates of single idiom constituents do not attract more fixations than unrelated distractors, this would speak for holistic idiom representations, not allowing a word-by-word analysis.

Table 1. German Example Sentence for Types 1–4 with English Equivalent.

(a) Person	(b) Sentence body	(c) Target words			
		(1) Correct	(2) Related	(3) Unrelated 1	(4) Unrelated 2
Hannes	ließ die Katze aus dem	Sack	Korb	Bauch	Arm
Hannes	*let the cat out of the*	*bag*	*basket*	*stomach*	*arm*

2.1 Methods

2.1.1 Participants. Thirty-one adults (mean age = 20.97, range = 18–30, 22 female, 9 male) participated in the experiment. Participants were recruited at the University of Tübingen and received subject credits as compensation. Prior to the experiment, participants gave written informed consent. All participants were native, monolingual speakers of German. Participants had no hearing impairments and normal or corrected-to-normal vision. Experiment 1 was approved by the Ethical Committee for Psychological Research at the University of Tübingen (reference number: 2016/1027/22).

2.1.2 Stimuli and design. We selected 20 well-known German idioms (see Appendix).[1] The idioms were embedded in sentences with a comparable structure (see Table 1): (a) a person carrying out the action of the sentence, (b) a sentence body that originated from a German idiom, and (c) the final target word of the idiom (which was not presented auditorily in Experiment 1). All idiomatic sentences were spoken in their complete form by a native speaker of German and digitally recorded. For Experiment 1, we removed the final target word from the recording. Participants heard each idiomatic sentence fragment once, while seeing four visual words on a computer screen. The four words represented one of these four types: (1) Correct Completion: correct completion of the idiomatic phrase, (2) Related Distractor: semantic associate of the correct completion, and (3&4) Distractors Unrelated 1 and Unrelated 2: semantically unrelated to the correct completion. Unrelated 1 and Unrelated 2 words were matched word pairs from Correct Completions and Related Distractors used with other sentence fragments in the experiment (avoiding phonological overlap). All words on the screen had the same grammatical gender fitting the preceding sentence context.

We ensured semantic relatedness between correct and related words by comparing pairwise semantic spaces using the R package LSAfun (Günther, Dudschig, & Kaup, 2015) and testing these similarity values with a Wilcoxon signed rank test. On average, semantic similarity between correct-related word pairs was significantly higher than between correct-unrelated1 ($Z = 189$, $p < .001$) and correct-unrelated2 ($Z = 185$, $p = .002$) word pairs. Semantic similarity between correct-unrelated1 and correct-unrelated2 word pairs did not differ ($Z = 75$, $p = .28$). Furthermore, semantic similarity between correct-related word pairs was significantly higher than between related-unrelated1 ($Z = 180$, $p = .004$) and related-unrelated2 ($Z = 182$, $p = .003$) word pairs.

Displayed words were presented in white font (Arial, font size 28) on a gray background. The position of the displayed words was counterbalanced across items and participants. The order of the trials was randomized.

2.1.3 Procedure. Participants completed the experiment in a single session. For the experimental task, participants received both written and oral instructions. Prior to the experimental task, each participant received a 5-point grid for calibration and a practice block consisting of five trials.

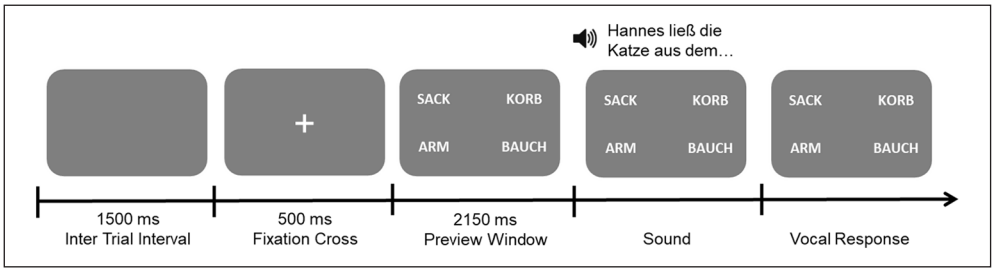


Figure 1. Example with times indicating the duration of the respective displays.

An exemplary trial scheme is displayed in Figure 1. Each trial began with a 1500 ms inter-trial interval followed by a 500 ms presentation of a fixation cross. Then the visual display of the set of four words appeared on the screen and remained until the end of the trial. The presentation of the audio stimuli started after a total of 2150 ms and was presented via headphones. After they heard the auditory stimuli, the task of the participants was to decide for each item which of the visually presented words was the best completion for the idiom by saying their choice out loud.[2] The experimenter noted the participants' responses. Participants were instructed to press a button after their oral response in order to continue on to the next trial.

We recorded fixations using a portable Tobii eye-tracker with a sampling size of 60 Hz. In total, the eye-tracking experiment took around 20 minutes including instructions, calibration and the experimental task, which took around 10 minutes.

2.2 Results

For the analysis, we divided the screen into four areas of interest. The analysis time window was aligned to the offset of each audio stimulus (offset = 0 ms). For the statistical analysis, we only included items responded to correctly, that is, in which the participants completed the sentence aloud with the correct final word of the idiom (these were 98.87% of all trials). Figure 2 (Panel A) shows fixations proportions towards correct, related, and aggregated unrelated words as fixation proportions from 800 ms before to 1000 ms after the offset of the spoken stimuli. Running *t*-tests comparing fixations towards correct completions and unrelated distractors at succeeding measurement points (every 16.67 ms) showed that participants' fixations were biased towards the correct idiomatic completion 464 ms prior to the offset of the audio stimuli ($p < .01$). This can be interpreted as the recognition point of the idiom. To compare the amount and time course of fixations towards related and unrelated distractors, we conducted a growth curve analysis (GCA) with orthogonal polynomials (Mirman, Dixon, & Magnuson, 2008). As the starting point of the GCA time window, we chose the start of observable anticipation (464 ms prior to the offset) for a duration of 1200 ms.

Fixation proportions were modeled with third-order orthogonal polynomials, because visual inspection of the time course bent at two points. To test the effect of *Distractor Type* (related vs. unrelated), we compared models using the –2LL deviance statistic. Including the effect of *Distractor Type* significantly improved the model fit ($\chi^2 = 42.98, p < .0001$). Estimated parameter terms of distractor type are summarized in Table 2. The intercept term reflects the average magnitude of the curve. Thus, the significant effect on the intercept term indicates that participants fixated more on related than on unrelated distractors across the complete time window. The linear term is comparable to the overall slope of the curve. In this case, the significant effect on the linear

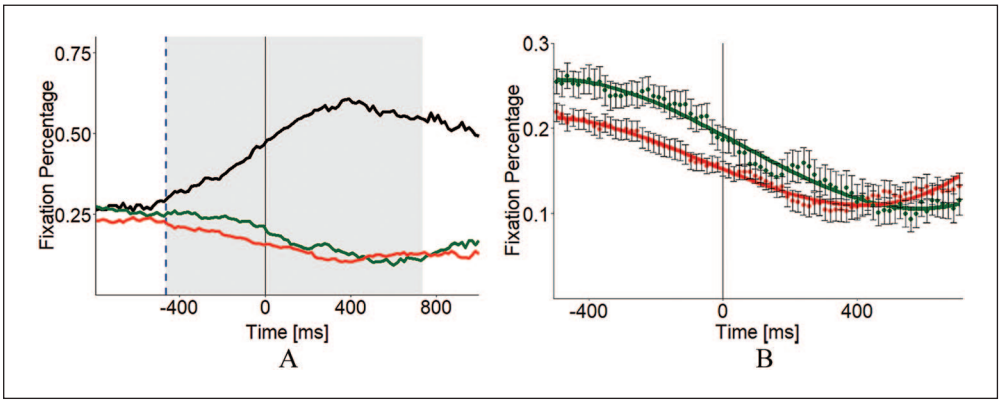


Figure 2. Panel (A) Fixation percentage for correct completions (black), related distractors (green) and mean of unrelated distractors (red); black vertical line = offset of spoken stimuli (0 ms); blue vertical, dashed line = start of the anticipation (-464 ms); gray background = time window for GCA. Panel (B) Fixation percentage for semantically related and unrelated distractors (points = mean; error bars = standard error) with fit of the growth curve model (line).

Table 2. Parameter Estimates for the Model including Distractor Type (Related vs. Unrelated).

Term	Estimate	Standard Error	*t*	*p* <
Intercept	−0.0080	0.0031	−2.6224	**.012**
Linear	0.0655	0.0298	2.1950	**.029**
Quadratic	−0.0249	0.0049	−5.0924	**.001**
Cubic	−0.0066	0.0049	−1.3428	.180

term implies variation across time with larger differences between the distractor types at the beginning of the time window. The quadratic term reflects symmetric inflection of the curve around the center meaning. Thus, the curve of the related distractor is shallower than the curve of the unrelated distractor, and towards the end of the time window the proportions of looks to related and unrelated distractors converge. The cubic term reflects inflections of the curve at the ends of the analysis time window. We found no significant effect on this term.

2.3 Discussion

The predictive eye movements recorded in an eye-tracking paradigm in Experiment 1 were in line with previous priming and eye-tracking research (Beck & Weber, 2016; Holsinger, 2013; Smolka et al., 2007; Sprenger et al., 2006) in showing that single word meanings are available in online processing of idioms. In Experiment 1, participants looked more often at distractor words that were related to the idiom's final word than at unrelated distractor words. Moreover, this fixation bias emerged anticipatorily, meaning well before the point in time at which the idiom's final word would have become evident in the speech signal. In fact, participants started to anticipate — that is, look at — correct idiomatic completions around 460 ms prior to the offset of the phrase fragment. Because programming of saccades after onset of the critical word typically takes around 200

ms (Saslow, 1967), we can assume that recognition of the idiom occurred even before 460 ms. Simultaneously with the increase in fixations on correct idiomatic completions, the fixation bias for semantic associations emerged. The fixation bias towards semantic associates diminished over time and ended 400 ms after the offset of the phrase fragment. In sum, our eye-tracking data suggest rapid prediction of upcoming idiom completions revealing that listeners represented these ordered strings in their mental lexicon. In addition, predictive eye movements to semantic associates of idiom completions demonstrate that listeners not only pre-activate and predict words within idioms in a holistic fashion, but they also appear to pre-activate single constituents together with their respective meanings.

For some of the used idiom fragments, the related distractor provided a literally plausible interpretation which might have compromised the fixation towards this distractor. In a post-hoc visual inspection, we plotted fixation data for items that allow a literal interpretation of the related completion (*Sarah band sich einen Klotz ans Knie*., literally translated: *Sarah tied herself a chunk to her knee*.) and implausible, related completions (*Hannah schlug sich die Zeit um die Augen*., literally translated: *Hannah hit herself the time around the eyes.*) separately. We did not observe decreased semantic activation for literally implausible, related completions supporting the interpretation of pre-activation of idiom constituents together with their semantic features. This complements the results from a visual world experiment using literal, novel phrases that show anticipatory fixations towards predicted words and semantic competitors, although the latter were implausible completions of the phrase (Ito & Husband, 2017).

The relatively long preview window that we implemented in the present experiment might have biased participants towards predictive processing. For example, Ferreira, Foucart, and Engelhardt (2013), suggested that preview time is associated with the strength of expectations participants form. Accordingly, longer preview of words or objects on the display is associated with stronger expectations participants form with regard to which word on the display is likely to be referred to. In this case, we would expect participants of Experiment 1 to build up stronger expectations for the correct idiom completion as part of the conventionalized phrase, and therefore weaken any tendency to look at related words. As a result of these stronger expectations, we might have overestimated the timing of the anticipation onset.

Another aspect of the eye-tracking design in Experiment 1 potentially limits its straightforward interpretation: the visually presented probes might have induced spreading semantic activation in a bottom-up fashion. Although we did not present a spoken version of the idiom-final constituent, a written version of it was available on the visual display, simultaneously with a written version of its semantic associate. Thus, fast fixations towards the correct idiomatic completion might have induced fast visual word processing and spreading semantic activation, which might have rapidly biased fixations towards the semantic associate. However, similar onsets of the fixation biases towards correct completions, on the one hand, and semantic associates, on the other hand, somewhat restrict an interpretation in terms of spreading activation exerted by the visual versions of the correct completions, because this mechanism might need some extra processing time (i.e., visual word recognition of the correct completion, spreading activation, and elicitation of eye movements towards its semantic associate). Nevertheless, similar to results of other studies (Beck & Weber, 2016; Holsinger, 2013; Smolka et al., 2007), the present eye-tracking data might overestimate decomposition because an instance of the critical constituent was visually included in each trial. In Experiment 2 and Experiment 3, we attempted to further rule out this alternative interpretation by avoiding any presentation of the critical idiom constituent for which we attempt to measure prediction effects in ERP experiments.

3 Experiment 2

In the following two experiments, we exploited semantic expectancy in spoken (Experiment 2) and written (Experiment 3) idioms in an ERP paradigm comparable to that of Rommers et al. (2013). As in the former study, we focused on N400 effects. Typically, reduction of the N400 ERP component is related to facilitated semantic processing, including semantic expectancy mechanisms (e.g., Federmeier & Kutas, 1999; Kutas & Federmeier, 2011; Laszlo & Federmeier, 2009). The N400 is a centro-posterior negative-going ERP component peaking around 400 ms after word onset. In N400 experiments, semantic expectations are usually determined via the cloze probability of a critical word within a given context. This measure reflects how often participants complete a phrase or sentence with a specific word. The N400 amplitude inversely correlates with this index: the higher the cloze probability of a word, the smaller the N400 amplitude it elicits (Kutas & Hillyard, 1984). Respective predictive mechanisms are so strong that even the processing of an unexpected word (with a low cloze probability) might reduce N400 amplitude if it shares semantic features with the expected stimulus (e.g., Federmeier & Kutas, 1999; Federmeier et al., 2002).

Evidence for the sensitivity of the N400 to the prediction of semantic features originally came from Federmeier and Kutas (1999), who presented participants with written versions of highly predictive sentences, such as "*They wanted to make the hotel look more like a tropical resort. So along the driveway, they planted rows of . . .*" Sentences ended with either a highly expected word (*palms*), an unexpected word from the same semantic category (*pines*) or an unexpected word from a different semantic category (*tulips*). In this experiment, the N400 amplitude for unexpected words from both categories clearly differed from the N400 amplitude for expected words. Moreover, N400 amplitudes were graded: words from the same semantic category as the expected word elicited a significantly smaller N400 amplitude than words from a different semantic category. Therefore, the N400 effect shows that semantic features of expected words are co-activated during online comprehension and words sharing these features benefit from predictive processing.

In the context of written idioms, Rommers et al. (2013) did not replicate the N400 prediction effect for semantic associates of final words. Participants read Dutch idioms embedded in figuratively biasing contexts (*After many transactions the careless scammer eventually walked against the lamp yesterday*.) in which the final word of the embedded idiom was either correct (*lamp*), not expected but from the same semantic category as the correct completion (*candle*), or not expected and from a different semantic category (*fish*). An N400 reduction for correct idiom-final words was found. This effect emerged with the typical topography (posterior) and within the typical time window of the N400 (300–400 ms). Yet semantic associates of correct idiom completions did not elicit an N400 reduction. That is, ERPs did not indicate facilitated processing of semantic associates of single idiom constituents. In addition to the N400 effect, Rommers et al. (2013) found a reduced late positivity ranging between 500 and 800 ms for correct idiom completions compared to related and unrelated completions. Again, the related and the unrelated condition did not differ. Rommers and colleagues interpreted this positivity as an instance of the P600 component reflecting a violation of the idiom representation as a linguistic unit.

In Experiment 2 and Experiment 3, we adopted the semantic expectancy ERP paradigm by Rommers et al. (2013) to preclude possible bottom-up spreading semantic activation (as in the eye-tracking paradigm in Experiment 1). In Experiment 2, we examined spoken versions of idioms in a unimodal design in order to be able to relate the results to previous cross-modal designs with spoken idioms that found activation of semantic associates of idiom constituents (eye-tracking paradigm in Experiment 1, Beck & Weber, 2016; Holsinger, 2013). In the literature, results for the semantic N400 effect in sentences is fairly comparable for visual and auditory processing (Connolly et al., 1992; Federmeier et al., 2002; Hagoort & Brown, 2000). This includes semantic expectancy

effects (Federmeier et al., 2002). Only the onset of N400 might differ, in that it starts earlier for auditory than for visual processing. It is still a matter of debate whether this early onset is functionally different from the N400 or not (Connolly & Phillips, 1994; Diaz & Swaab, 2007; Nieuwland, 2019; Van Den Brink, Brown, & Hagoort, 2001).

We again presented German idioms in short sentences without further context, including the ones we used in our eye-tracking study (Experiment 1) as well as additional items. Participants listened to highly predictive idiomatic phrase onsets (e.g., *Hannes ließ die Katze aus dem . . .*, "Hannes let the cat out of the . . ."). Phrase onsets were completed either (1) with the expected and correct final idiom word (*Sack*, "bag"), (2) with an unexpected but semantically related completion (*Korb*, "basket"), or (3&4) with an unexpected and semantically unrelated completion (*Arm*, "arm"; *Bauch*, "stomach"). If processing is solely holistic, the words in related and unrelated conditions should show comparable ERP amplitudes, as was shown by Rommers et al. (2013). Such a finding would suggest that fixations towards semantic associates of correct completions in Experiment 1 were merely an epiphenomenon of bottom-up spreading activation exerted by the visual probe being presented together with the correct completion within the same display. If literal meanings of expected words are accessed, the processing of semantically related words should benefit more from this expectation when compared to unrelated words. This would yield graded ERP amplitudes for related and unrelated completions.

3.1 Methods

3.1.1 Participants. Forty-two healthy participants volunteered for Experiment 2. None of the participants had taken part in Experiment 1. We excluded the data of one bilingual participant and of one participant for whom we had technical problems with the ERP recording. Participants whose data were included in the analysis ($N = 40$, mean age = 22.9 years, range = 18–32, 20 female and 20 male) were right-handed as assessed by the Edinburgh Handedness Questionnaire (Oldfield, 1971), monolingual native speakers of German, and had no history of a neurological, psychiatric, or hearing disorder. As compensation, subjects were paid for the experiment or provided with subject credits. Experiment 2 was approved by the Ethical Committee of the German Psychological Society (reference number: RK 112015).

3.1.2 Stimuli. In order to arrive at a sufficient number of trials for an ERP study, we extended the experimental materials from Experiment 1 from 20 to 40 phrases using the same criteria of familiarity and predictability (see Appendix). Linguistic stimuli resulted from the combination of the sentence body with the four sentence final target words in four conditions with a combination logic following that of Experiment 1 (see Table 1). The conditions were the following: (1) Correct Condition: the target word was the correct completion of the idiomatic phrase, (2) Related Condition: the target word was semantically related to the correct completion, and (3&4) Conditions Unrelated 1 and Unrelated 2: the target word was semantically unrelated to the correct completion. Unrelated 1 and Unrelated 2 words were matched word pairs from Correct and Related Conditions used with other sentence bodies in the experiment (no phonological or semantic overlap). Each sentence body was repeated four times, once in all four conditions. This resulted in 160 different combinations of sentence bodies and target words. The same native speaker of German as in Experiment 1 spoke all linguistic stimuli. The linguistic stimuli that were repeated across conditions (sentence body and final words) were realized as the same recordings.

We conducted rating studies to determine some characteristics of the materials essential for ERP research. In a cloze probability task, 17 participants read the 40 sentence bodies and filled in the word that they considered to be the most likely completion. The mean cloze probability of the correct idiom-final word was 93.82% ($SD = 9.69$).

Furthermore, we controlled for the semantic relatedness between critical words by means of a second rating study. Fifteen participants received lists of word pairings of the target words and judged their relatedness on a scale from 1 to 7. The association strength between words presented as critical words in the Correct Condition (i.e., between the correct idiom completion) and words presented in the Related Condition (see Table 1) was significantly higher than the association strength of critical words presented in the Correct Condition and both Unrelated Conditions (Wilcoxon signed rank test: Unrelated 1 $Z = 120$, $p < .001$; Unrelated 2 $Z = 120$, $p < .001$). The association strength between critical words presented in the Correct Conditions and those presented in both Unrelated Conditions did not differ ($Z = 78$, $p = .32$).

3.1.3 Procedure. Participants completed the experimental task in a single session. After signing an informed consent form, participants sat in a comfortable chair facing a computer screen in a dimly lit room. During the experimental task, they were instructed to sit still and avoid eye movements including blinking. Later, participants took part in a calibration task at the beginning and the end of the session. In this task, eye movements were systematically evoked for offline ocular correction. Before the experimental task, participants received both written and oral instructions. The participants received a practice block consisting of eight trials to ensure that they were familiar with the procedure and the task.

For each experimental trial, a sentence was presented auditorily via loudspeakers on both sides of the computer screen. During the presentation of the sentences, participants viewed a fixation cross at the center of the screen. After the auditory presentation, the task of the participants was to decide for each sentence whether it was a correct idiomatic phrase or not by pressing buttons with the index fingers of the right or the left hand.[3] The side for yes- and no-buttons was counterbalanced across participants. The response type was a delayed response; 1200 ms after onset of the target stimulus a question mark appeared at the center of the screen to signal the start of the response window for the participants. If they responded before the start of the response window, participants were given feedback (*too fast*). The interval between succeeding trials was 1500 ms.

The experiment consisted of eight blocks of 20 trials, 160 trials in total, with five trials in each condition in each block. The order of trials was pseudorandomized in such a way that the same sentence body or target word never occurred in the same block. After each block, participants had the opportunity to take a self-timed break. The order of blocks was randomized using the Latin Square method. In total, the EEG experiment took around 1.5 hours including electrode application, instruction, calibration and the experimental task; the experimental task itself took around 15–20 minutes.

3.1.4 Electrophysiological recordings. Electrophysiological brain potentials were recorded with 46 active electrodes (Ag/AgCl) mounted in an elastic cap (Easycap GmbH, Herrsching, Germany) according to the 10–20 system (see Figure 3), online referenced to the nose. The ground electrode was positioned at the location of the AF3. In order to record eye movements, we attached two ocular electrodes below both eyes. The raw data were sampled at 500 Hz (bandpass filter 0.01–100 Hz, BrainAmpStandard, Brain Products, Gilching, Germany).

3.1.5 EEG analysis. For the ERP analysis, the raw data were re-referenced offline to the average reference and filtered with a 0.3 Hz Low-Cut-Off filter. Using surrogate MultipleSource EyeCorrection (MSEC) by Berg and Scherg (1994), we removed horizontal and vertical eye movements as well as blinks from the continuous EEG signal. The EEG data were segmented into trials in epochs from 100 ms before and 1000 ms after the stimulus onset with a 100 ms pre-stimulus baseline subtraction. We excluded trials contaminated with artifacts and in which participants

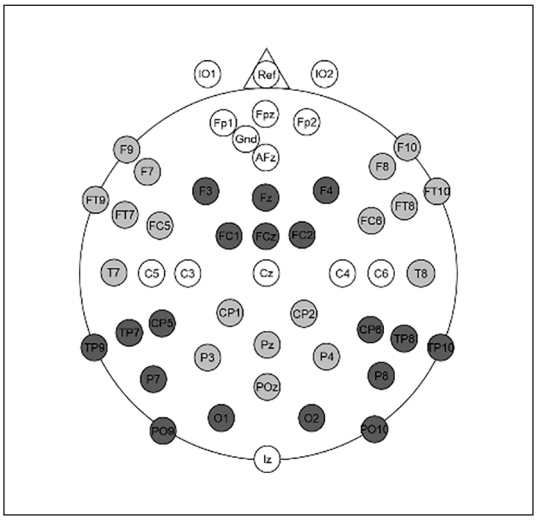


Figure 3. Electrode configuration used in the experiment. Anterior-Left, Anterior-Right, and Posterior-Central ROIs are highlighted in light gray. Anterior-Central, Posterior-Left, and Posterior-Right ROIs are highlighted in dark gray.

responded before the onset of the response time window -1200 ms after stimulus onset. Further, we only included individual items that participants responded to correctly in the Correct Condition in the analysis in all conditions, because we assumed that when participants recognized the idiom correctly in the Correct Condition (94.31%), they had established memory traces of the correct idiom form. These inclusion criteria resulted in the following percentage of trials per condition: Correct: 79.4%; Related: 81.3%; Unrelated 1: 78%; Unrelated 2: 78.7%. For further analyses, we aggregated the conditions Unrelated 1 and Unrelated 2 into one condition Unrelated by averaging the mean voltages of the two conditions for each participant. Following this process, the final three conditions discussed in the analyses were: Correct, Related, and Unrelated.

Based on visual inspection of ERP results, we chose six regions of interest (ROIs), covering lateral and midline anterior and posterior sites (see Figure 3). Both lateral anterior ROIs included six electrode positions over both temporal cortices (left: F9, F7, FT9, FT7, FC5, T7; right: F10, F8, FT10, FT8, FC6, T8). The anterior midline ROI covered six fronto-central electrodes (F3, Fz, F4, FC1, FCz, FC2). Both lateral posterior ROIs included six temporo-parietal electrode positions (left: TP9, TP7, CP5, P7, PO9, O1; right: CP6, TP8, TP10, P8, PO10, O2). The posterior midline ROI covered six centro-parietal electrode positions (CP1, CP2, P3, Pz, P4, POz).

For statistical analysis, we conducted a 3 x 3 x 2 repeated-measures ANOVA (RM-ANOVA) with the within-participant factors Condition (Correct, Related, Unrelated), Hemisphere (Left, Central, Right), and Region (Anterior, Posterior). First, we conducted RM-ANOVAs for each 100 ms time window. We identified three relevant time windows, which showed three-way interactions for Condition, Region, and Hemisphere (see Table 3): 100–200 ms, 300–500 ms, and 700–1000 ms. Both later time windows approximately align with the effects obtained in Rommers et al. (2013), with the 300–500 ms time window reflecting an N400 effect, and the 700–1000 ms time window reflecting a late positivity. The early time window does not find a parallel in previous ERP work on idiom processing. We label it as "pre-N400" throughout the results section. For further analysis, we aggregated amplitudes across these time windows.

Table 3. RM-ANOVAs. C–Condition, R–Region, H–Hemisphere. * for significant main effects and interactions.

	100–200 ms	200–300 ms	300–400 ms	400–500 ms	500–600 ms	600–700 ms	700–800 ms	800–900 ms	900–1000 ms
C	*								
CxR			*	*		*	*	*	*
CxH	*	*	*	*	*	*		*	*
CxRxH	*		*	*			*	*	*

3.2 Results

Figure 4 (Panel A) depicts Grand-Average ERPs aggregated over ROIs. Visual inspection of grand-averaged ERPs justified the selected time windows. As shown in the difference topographies (Figure 4, Panel B), the effect is most prominent over posterior sites. Moreover, a late positivity was observable over posterior sites.

RM-ANOVAs revealed significant three-way interactions for 100–200 ms, $F(4, 156) = 3.04$, $p = .03$, 300–500 ms, $F(4, 156) = 12.62$, $p < .0001$, and 700–1000 ms, $F(4, 156) = 4.27$, $p = .004$. All reported p-values are Greenhouse-Geisser or Bonferroni (for post-hoc t-tests) corrected.

3.2.1 100–200 ms time window (pre-N400). Post-hoc analyses of the three way interaction revealed a significant Condition effect for the Anterior-Left, Anterior-Central, Anterior-Right, and Posterior-Central ROIs, all $F(2, 78) \geqslant 7.18$, $p \leqslant .002$. Over the Anterior-Left sites only, all three conditions differed from each other: Correct vs. Related, $t1(39) = -2.93$, $p < .018$, Correct vs. Unrelated, $t2(39) = -6.21$, $p < .001$, and Related vs. Unrelated, $t3(39) = -3.09$, $p < .018$. Over the remaining three sites, we found differences between the Correct Condition vs. the Related Condition, all $t2(39) \geqslant |3.71|$, all $p \leqslant .002$, and for the Correct Condition vs. the Unrelated Condition, all $t3(39) \leqslant |1.04|$, all $p \geqslant .91$. In sum, we found parallel effects of semantic activation and no semantic activation.

3.2.2 300–500 ms time window (N400). For the 300–500 ms time window, a Condition effect was only evident over Posterior-Central sites, $F(2, 78) = 38.17$, $p < .0001$. Bonferroni-corrected post-hoc tests revealed significant differences between all three conditions: Correct vs. Related, $t1(39) = 5.91$, $p < .001$, Correct vs. Unrelated, $t2(39) = 7.43$, $p < .001$, and Related vs. Unrelated, $t3(39) = 2.71$, $p < .03$. Across Posterior-Central electrodes, amplitudes for the Unrelated Condition were more negative than those for the Related Condition, and amplitudes for the Correct Condition were most positive. Together, we found graded condition effects for a Posterior-Central electrode cluster typically associated with the N400.

3.2.3 700–1000 ms time window (late positivity). For the 700–1000 ms time window, we report those ROIs where a condition effect was significant, $F(2, 78) > 8.12$, $p \leqslant .002$. Post-hoc tests for these regions revealed differences of Related and Unrelated Conditions with Correct Conditions, but not between Related vs. Unrelated. Over Left-Anterior sites, amplitudes for the Correct Condition were more positive than for the Related Condition, $t1(39) = 3.94$, $p < .001$, and the Unrelated Condition, $t2(39) = 3.35$, $p < .006$, but amplitudes for the Related Condition and the Unrelated Condition did not differ significantly, $t3(39) = -1.7$, $p = .294$. Similarly, over Right-Anterior sites, amplitudes for the Correct Condition were more positive than for the Related

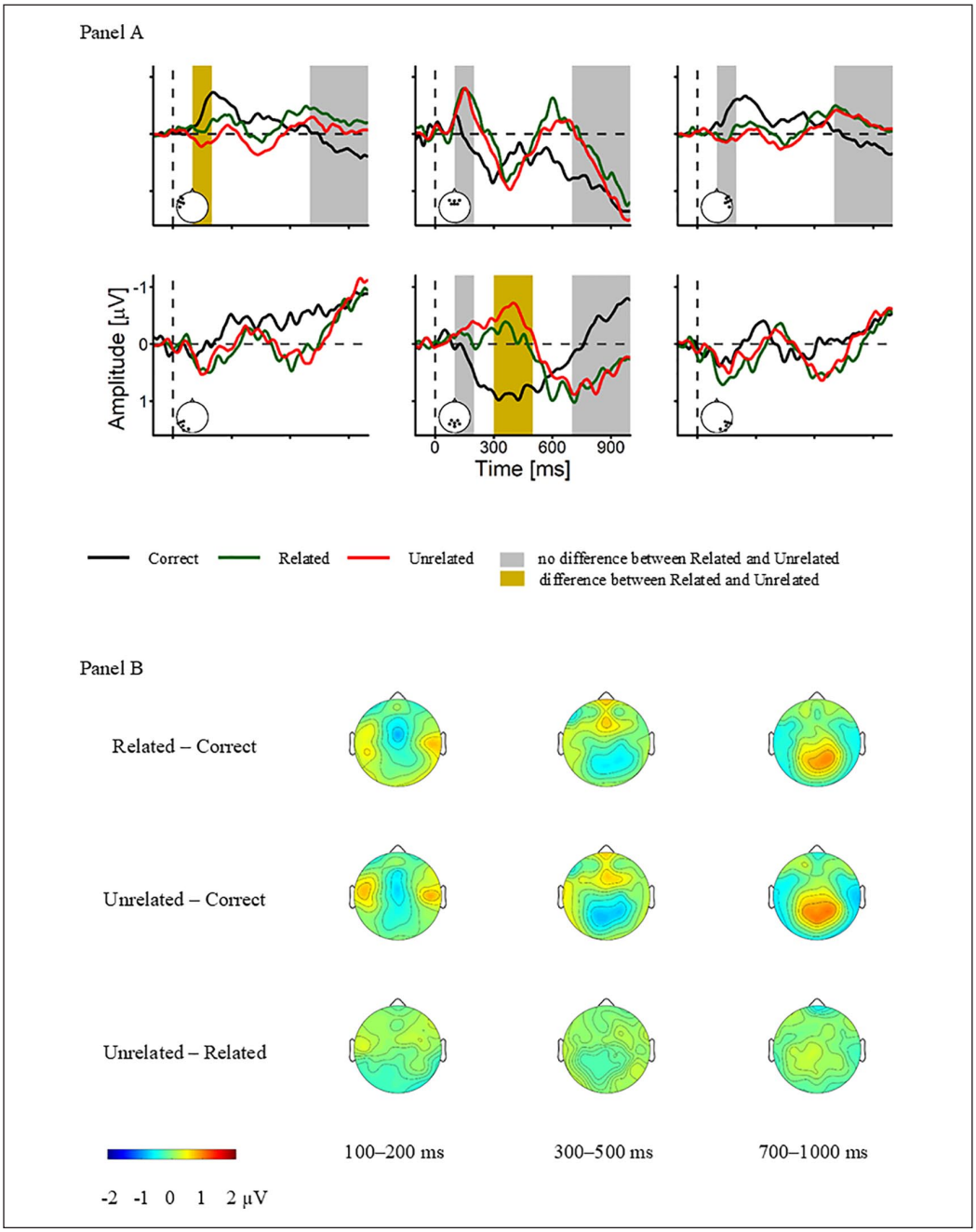


Figure 4. Grand-Averaged ERPs (A) ERP-waveforms for the ROIs Anterior-Left, Anterior-Central, Anterior-Right, Posterior-Left, Posterior-Central, and Posterior-Right. (B) Difference topographies for the time windows 100–200 ms, 300–500 ms, and 700–1000 ms.

Condition, $t1(39) = 2.87$, $p < .020$, and the Unrelated Condition, $t2(39) = 3.58$, $p < .003$, while amplitudes for the Related Completion and the Unrelated Condition did not differ significantly,

t3(39) = 0.05, $p = 1$. Over Posterior-Central sites, amplitudes for the Correct Condition were more negative than for the Related Condition, *t1*(39) = -5.55, $p < .001$, and the Unrelated Condition, *t2*(39) = -5.89, $p < .001$, but amplitudes for the Related Condition and Unrelated Condition did not differ significantly, *t3*(39) = -0.28, $p = 1$. In sum, late ERPs show that related and unrelated violations of the idiom yield comparable amplitudes of a late positivity with posterior distribution (and reversed amplitudes over anterior regions).

3.3 Discussion

Using a semantic expectancy ERP paradigm in Experiment 2, we investigated processing mechanisms in highly predictive spoken idiomatic phrases. In contrast to Experiment 1, the critical idiom constituent itself did not appear in trials in which we probed the activation of semantic associates of this idiom constituent. This way, we aimed to rule out potential bottom-up spread from sensory input, which could have biased results in the visual world eye-tracking design exploited in Experiment 1.

Across ERP amplitudes, there was a clear effect of expectancy of the correct idiom: both related and unrelated violations showed significantly higher ERP amplitudes than correct completions. This indicates that correct completions of an idiom were highly expected and easier to access than both related and unrelated substitutes. Because idioms were presented without biasing context, this broadly supports the notion that predictability within idioms mainly stems from the knowledge of the idiom form (Vespignani et al., 2010).

Using spoken idioms, N400 amplitudes reflected semantic expectancy within violation trials. That is, we not only obtained N400 reductions for correct completions, but also for semantic associates of correct completions. Since N400 reductions are interpreted in terms of facilitated semantic processing, including semantic expectancy mechanisms (for a review, see Kutas & Federmeier, 2011), it seems that the anticipation of the correct completion activated semantic associates, for which semantic processing was facilitated. In this sense, the N400 effect observed here is compatible with the eye-tracking data from Experiment 1. It appears that single constituents and their individual meanings are available when these are predicted. These results in the auditory modality do not replicate those obtained for visually presented idioms obtained by Rommers et al. (2013), and challenge the conclusions drawn by these authors, who concluded that the top-down prediction of idiom completions does not lead to beneficial processing of substitutes that are semantically related to idiom constituents.

The ERPs obtained in Experiment 2 mainly reflect an N400 effect followed by a late positivity. Recently, it has been discussed whether during the processing of idioms or other formulaic sequences the N400 is preceded by a P300 effect (Molinaro & Carreiras, 2010; Siyanova-Chanturia et al., 2017; Vespignani et al., 2010). The authors of those studies found an enhanced P300 amplitude for correct and expected idiomatic forms compared to violations of those forms. They concluded that the P300 reflects a template matching process. Although we cannot rule out that the present N400 effect might also include an instance of the P300, we hypothesize that an activation of semantic information as found in Experiment 2 would only be detectable in the N400 component. We therefore conclude that the graded ERP effect between 300 and 500 ms in Experiment 2 are indeed an instance of the semantic N400 effect.

A late positivity between 700 and 1000 ms across posterior sites was independent of semantic relatedness, that is, it did not show amplitude differences between related and unrelated violations. This effect converges with findings by Rommers et al. (2013), who interpreted this late effect as a violation of the idiom as a lexical item. More recently, the late positivity following the semantic N400 (post-N400 positivity, PNP) in prediction paradigms has been interpreted as revision of a

predicted sentence representation (Brothers, Swaab, & Traxler, 2015; Kuperberg & Wlotko, 2020) irrespective of the semantic relations between presented sentence-final words (Thornhill & Van Petten, 2012). Thus, in idiom processing the late positivity might also reflect that listeners revise the activated representation of the idiom string when hearing related or unrelated violations. Together with the N400 effect suggesting decomposition, the late positivity effect could be interpreted as evidence for a dual representation of idioms in the mental lexicon as both individual words and chunked items (Sprenger et al., 2006).

Similarly, early ERP effects obtained in the present study suggest that decomposition is not the only strategy followed by the parser. In contrast to the results from the written presentation of idioms by Rommers et al. (2013), we obtained ERP effects preceding the N400 in our study with spoken materials. This is comparable with previous findings (for a review, see Nieuwland, 2019). Already early on (between 100 and 200 ms), we see evidence for parallel processing. Across anterior-left electrodes, amplitudes for related conditions significantly differed from amplitudes for unrelated conditions. Across central and anterior-right sites, amplitudes for related and unrelated conditions did not differ. These early ERP effects might relate to parallel pre-activation of lexical representations (e.g., Friedrich & Kotz, 2007). If so, the present ERP results dissociate two types of lexical idiom representations: a form and a meaning representation of the single constituents. The former is indicated by the mid to right-anterior ERP deflections, while the latter is indicated by the left-lateralized ERP deflection. Thus, within familiar and highly predictable idioms, final constituents including their semantic properties can be pre-activated before they are fully processed (Smolka & Eulitz, 2020).

In general, the results of Experiment 2 corroborate other studies presenting idioms auditorily (Beck & Weber, 2016; Holsinger, 2013) by showing that listeners activate idiom constituents and have semantic associates of these constituents available. Possibly, the pre-N400 ERP effects and the graded N400 effect that we found might be due to modality-related differences compared to the study by Rommers et al. (2013). In contrast to the written and serial presentation (word-by-word) in that study, we presented idioms and violated idioms auditorily in Experiment 2. Semantic information might be accessible earlier in spoken language processing compared to written language processing. For example, preceding information speeds up spoken word identification even before enough acoustic information has accumulated (Van Petten et al., 1999). Therefore, we conducted a third experiment in which we used the same task and material as in Experiment 2, but presented them in the written modality.

4 Experiment 3

In Experiment 3, we conducted a semantic expectancy ERP experiment using the same material as in Experiment 2, but with written instead of spoken idioms. While experiments on spoken idiom processing clearly point to decomposition within idioms (Experiments 1 and 2; Beck & Weber, 2016; Holsinger, 2013), the evidence from word-by-word presentations of idioms is mixed (Rabanus et al., 2008; Rommers et al., 2013; Smolka et al., 2007). Therefore, we aimed to address the question of processing differences across modalities in Experiment 3 by using the same idiomatic expressions and violations of these forms as in Experiment 2. If there are any prediction effects inherent to the idioms we used, we should not replicate results by Rommers et al. (2013).

4.1 Methods

4.1.1 Participants. Thirty adults participated in Experiment 3, of whom we had to exclude data of five participants, due to incorrect instructions (3), a psychiatric disorder (1), and insufficient eye movement correction (1). This resulted in a sample of 25 participants for statistical analysis (N = 25, mean age = 21.4 years, range = 18–27, 18 female and 7 male). Participants were recruited at

the University of Tübingen and received subject credits or payment as compensation. All participants included in the analysis were native, monolingual speakers of German, right-handed as assessed by the Edinburgh Handedness Questionnaire (Oldfield, 1971), and had no history of a neurological, psychiatric, or hearing disorder and normal or corrected-to-normal vision. None of the participants took part in Experiments 1 or 2. Prior to the experiment, participants gave written informed consent.

4.1.2 Stimuli. In Experiment 3, we used the same stimuli as in Experiment 2, but these were presented visually at the center of a computer screen.

4.1.3 Procedure. The procedure was the same as in Experiment 2 except for the presentation modality of the stimuli. We used the same timing of presentation as in the EEG study by Rommers et al. (2013). Each trial started with a fixation cross (+) for 1500 ms. Sentences were presented word-by-word with 300 ms presentation duration of a word and 300 ms blank screen. At 900 ms after the presentation of the sentence-final word, a question mark (?) appeared on the screen, resulting in a 1200 ms delayed response after the onset of the target word. When the question mark appeared, participants were asked to decide whether the presented sentence was a correct idiom or not. They gave their answers via button press. The response hand was counterbalanced across participants.

4.1.4 Electrophysiological recordings. Same as in Experiment 2.

4.1.5 EEG analysis. As in Experiment 2, we included items that did not contain artifacts and to which participants responded correctly in the Correct Condition (92.8%) and after the onset of the response time window (1200 ms after stimulus onset). This resulted in the following percentage of included trials per condition for the analysis: Correct: 69.5%, Related: 72.3%, Unrelated 1: 72.4%, Unrelated 2: 71.5%. Compared to Experiment 2, the number of trials was lower, because the EEG recordings were more artifactual.

Conducting RM-ANOVAs for 100 ms time window steps, we identified two relevant time windows, which showed three-way interactions for Condition, Region, and Hemisphere (see Table 4): 300–400 ms, and 500–700 ms. The first time window aligns with the early N400 time window found in Rommers et al. (2013). The later time window partly aligns with the time window for the late positivity in Rommers et al. (2013, 500–800 ms). For further analysis, we aggregated amplitudes across these time windows.

4.2 Results

Figure 5 (Panel A) depicts Grand-Average ERPs aggregated over ROIs. As shown in the difference topographies (Figure 5, Panel B), the N400 effect is most prominent over posterior sites. RM-ANOVAs revealed significant three-way interactions for 300–400 ms, $F(4, 96) = 5.31$, $p = .002$, and 500–700 ms, $F(4, 96) = 5.82$, $p < .001$. All reported p-values are Greenhouse-Geisser or Bonferroni (for post-hoc t-tests) corrected.

4.2.1 300–400 ms time window (N400). Across Left sites, we found a main effect for Condition, $F(2, 48) = 21.21$, $p < .001$, which was due to significant amplitude differences between Correct vs. Related, $t1(24) = -3.60$, $p = .004$, Correct vs. Unrelated, $t2(24) = -6.11$, $p < .001$, and Related vs. Unrelated, $t3(24) = -3.01$, $p = .018$. Across Central sites, we also found a main effect for Condition, $F(2, 48) = 9.93$, $p < .001$. Post-hoc t-tests revealed amplitude differences between Correct vs. Related, $t1(24) = 4.91$, $p < .001$, and Correct vs. Unrelated, $t2(24) = 6.02$, $p < .001$, to be significant. Amplitudes for the Related and Unrelated Conditions did not differ, $t3(24) = 1.08$, $p = .869$.

Table 4. RM-ANOVAs. C–Condition, R–Region, H–Hemisphere. * for significant main effects and interactions.

	100–200 ms	200–300 ms	300–400 ms	400–500 ms	500–600 ms	600–700 ms	700–800 ms	800–900 ms	900–1000 ms
C		*	*	*	*	*			*
CxR				*	*	*	*		
CxH	*	*	*						*
CxRxH			*		*	*			

Across Right sites, we found an interaction of Condition with Region, $F(2, 48) = 3.64$, $p = .04$. For Right-Anterior electrodes there was no effect of Condition, $F(2, 48) = 2.51$, $p = .11$, but for Right-Posterior electrodes the main effect for Condition was significant, $F(2, 48) = 13.24$, $p < .001$. Across the latter region, amplitudes between Correct vs. Related, $t1(24) = -3.66$, $p = .004$, and Correct vs. Unrelated, $t2(24) = -4.75$, $p < .001$, differed significantly. There was no amplitude difference between Related vs. Unrelated, $t3(24) = -0.29$, $p = 1$. Altogether, in the typical N400 time window and region we did not find evidence for a graded pattern of semantic expectancy.

4.2.2 500–700 ms time window (late positivity). For the 500–700 ms time window, we found a main effect of Condition across Left sites, $F(2, 48) = 5.41$, $p = .008$, and an interaction of Region and Condition across Central sites, $F(2, 48) = 6.00$, $p = .006$. Over Left electrodes, we found no amplitude differences between Correct vs. Related, $t1(24) = -0.70$, $p = 1$, significant amplitude differences between Correct vs. Unrelated, $t2(24) = -3.02$, $p = .018$, and marginally significant amplitude differences between Related vs. Unrelated, $t3(24) = -2.53$, $p = .055$. Across Central electrodes, there was only a Condition effect for Central-Anterior electrodes, $F(2, 48) = 4.92$, $p = .02$, with a significant amplitude difference between Correct vs. Unrelated, $t2(24) = 2.65$, $p = .042$, and no amplitude differences between Correct vs. Related, $t1(24) = 1.92$, $p = .199$, and Related vs. Unrelated, $t3(24) = 1.44$, $p = .489$. In sum, for the late effect the amplitude differences show a mixed pattern.

4.3 Discussion

In Experiment 3, we again conducted a semantic expectancy ERP experiment to investigate top-down spreading semantic activation within idioms. In contrast to Experiment 2, where we used the same material in a unimodal auditory paradigm, we presented idioms as written stimuli on the screen to further explore potential modality-related differences in processing. Therefore, the design was directly comparable to that of Rommers et al. (2013), who did not find activation of semantic associates of final constituents within written versions of idioms.

For written idioms, we found a clear expectancy effect on the N400: amplitudes for related and unrelated completions differed significantly from amplitudes for correct completions. This parallels findings for spoken idioms in Experiment 2. However, in contrast to Experiment 2, we did not find amplitude differences between related and unrelated targets in the typical semantic N400 region. Thus, we did not find an effect of semantic expectancy here. Based on the results by Rommers et al. (2013) and the present Experiment 2, these results might indicate that for online prediction of semantic features within idioms, the modality in which the idioms are presented might indeed play a role.

Nevertheless, we found differences between related and unrelated targets over left-hemispheric electrode leads in the N400 time window. Since there was no evidence of an N400

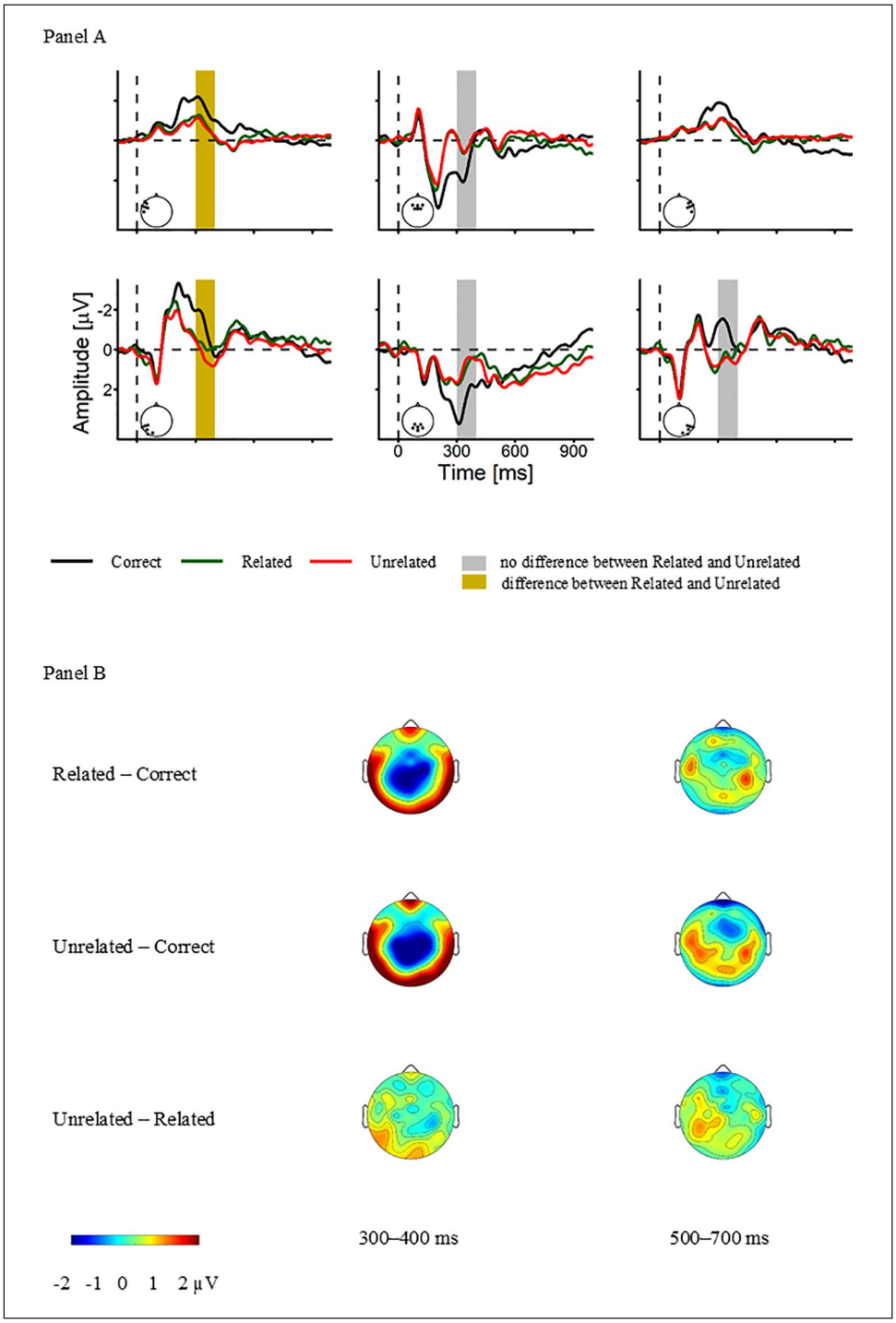


Figure 5. Grand-Averaged ERPs (A) ERP-waveforms for the ROIs Anterior-Left, Anterior-Central, Anterior-Right, Posterior-Left, Posterior-Central, and Posterior-Right. (B) Difference topographies for the time windows 300–400 ms and 500–700 ms.

effect localized in this region either in Experiment 2, in the experiment by Rommers et al. (2013), or in the literature on the semantic N400 effect in non-idiomatic language an interpretation of this effect is difficult.

During a later time window, we did not replicate effects of a late positivity found in Experiment 2 and by Rommers et al. (2013). This effect was previously interpreted as indexing a violation of the holistic idiom representation. In Experiment 3, we only found consistent differences between correct and unrelated words. Amplitude differences between related and unrelated completions were mixed. Rommers et al. (2013) interpreted the late positive ERP effect as reflecting the difficulty to revise a predicted idiomatic multi-word representation. However, as we did not replicate such a late positive ERP effect with written idioms (Experiment 3), we are not confident about an interpretation at this point.

Like Rommers et al. (2013), we did not find a pre-N400 component for written idioms in an early time window. This suggests that the early component found in Experiment 2 was indeed specific to processing in the auditory modality (Connolly & Phillips, 1994; Connolly, Phillips, & Forbes, 1995).

Different ERP effects obtained for spoken idioms in Experiment 2 and written idioms in Experiment 3 challenge an alternative interpretation of activation effects for semantic associates in our study. Even though some semantically related completions of the idioms we have presented might have a literally plausible interpretation, the results of Experiment 3 reveal that it is unlikely that the N400 is modulated by literal plausibility. If this were the case, we would also see an effect of semantic activation for written idioms in Experiment 3, because we used the same material for both modalities. Instead semantic activation was only observable for the processing of spoken idioms in Experiment 2. Moreover, Rommers et al. (2013) did not find N400 amplitude differences between related and unrelated completions although related completions were rated as more plausible than unrelated completions. For these reasons, we argue that literal plausibility does not account for the effects of semantic co-activation on the N400 component found in Experiment 2. Furthermore, research on plausibility and predictability in literal language suggests that rather than the N400 component, a post-N400 positivity is affected by the plausibility of the interpretation (DeLong, Quante, & Kutas, 2014; Quante, Bölte, & Zwitserlood, 2018). In the present study, we found no amplitude reduction of the late positivity for related completions indicating effects of plausibility. Furthermore, in the idiom literature amplitude differences in the N400 were not associated with semantic integration processes (Canal et al., 2017). Altogether, we hypothesize that the reduction of the N400 amplitude for the spoken idioms that we obtained in Experiment 2 results from a short-lived semantic activation of the final constituent.

5 General discussion

In the present study, we aimed to shed light on previous contradictory evidence on the extent to which idioms are processed holistically or decomposed into single items. Indirect evidence for holistic processing stems from studies showing faster processing for idioms compared to novel phrases (e.g., Conklin & Schmitt, 2008; Swinney & Cutler, 1979; Tabossi et al., 2009). Here, we tested idiom processing more directly by measuring their possible decomposition by means of semantic activation of individual idiom components (see e.g., Siyanova-Chanturia, 2015). Previous research demonstrated that semantic features of idiom constituents are available at least for priming processes in reading and listening (Beck & Weber, 2016; Holsinger, 2013; Smolka et al., 2007). However, when focusing on prediction mechanisms in reading, evidence for decomposition in idiom processing was lacking (Rommers et al., 2013). To rule out design

and modality-related differences, we measured the level of semantic expectancy during online processing of highly predictive, spoken idioms in an eye-tracking paradigm with written words (Experiment 1) as well as in a semantic expectancy ERP paradigm with spoken (Experiment 2) and written (Experiment 3) idioms.

Across all three experiments, we found evidence that participants built up an expectation of the idiom-final word. They fixated the correct idiom completion well before the idiom fragments presented in Experiment 1 ended, and they showed reduced N400 amplitudes for correct idiom completions compared to unrelated words in Experiments 2 and 3. Based on this, we conclude that idioms and their conventionalized forms can be recognized and activated before their offset (Libben & Titone, 2008; Smolka & Eulitz, 2020; Vespignani et al., 2010). Together, these findings are evidence for multi-word representation of idioms. It appears that the mental lexicon stores information about the co-occurrence of specific words making up an individual idiom. Activation of respective multi-word representations triggers expectation of individual words that are part of these multi-word expressions. Here, we do not preclude a certain flexibility of these multi-word representations, but propose rather a strong coherence between the words of which they are composed (Cacciari & Tabossi, 1988; Geeraert, Baayen, & Newman, 2017; Kyriacou et al., 2020; Mancuso et al., 2020).

Across eye-tracking and ERP methods with spoken idioms, we found evidence for early, short-lived semantic activation of individual idiom constituents. As soon as participants fixated correct idiom completions, they also fixated respective semantic associates (Experiment 1). In the ERPs, semantic associates of correct idiom completions elicited effects in the same early time window in which correct completions elicited effects (Experiment 2). Since we found anticipation of correct idiom completions in the fixation data, we conclude that early effects for semantic associates of idiom completions in the ERPs indeed relate to pre-activation of idiom constituents. Based on knowledge of conventionalized idiom forms, parsers seem to pre-activate a multi-word representation before the respective idiom is completely available in the auditory signal and this pre-activation includes single word representations that spread semantic activation within the network. It appears that even though literal constituent meanings typically do not contribute to the understanding of the idiomatic meaning, their processing is still automatically carried out. This conclusion is comparable to the notion that semantic processing cannot be switched off, as for example Connolly, Stewart, and Phillips (1990) showed for spoken language processing. We speculate that this is similar to a Stroop-like effect (Stroop, 1935) where the literal meaning of the word is not informative, but is nevertheless activated (cf. Glucksberg, 1993; McGlone, Glucksberg, & Cacciari, 1994).

It appears that semantic activation of constituent words within spoken idioms rapidly declines over time, as proposed for automatic spreading activation within the semantic network (e.g., Neely, O'Connor, & Calabrese, 2010). Neither fixation data nor ERPs gave evidence for long-lasting semantic activation of idiom constituents. Across Experiments 1 and 2, there was no longer a processing benefit for semantic associates compared to unrelated words after respective initial effects. Within spoken idioms, the present effect is comparable to that obtained by Sprenger et al. (2006), where semantic activation appeared to be strongest during early processing stages.

Here, we tentatively speculate that a rapid decay of semantic activation of constituent associates accounts for the presently and previously found mixed results for spoken and written idioms. Across paradigms using spoken idioms (Experiments 1 and 2), we consistently found evidence for activation of semantic associates of idiom-final words. Using written versions of the same idioms as in Experiments 1 and 2, we did not find effects of semantic activation in Experiment 3 and this replicates results that Rommers et al. (2013) obtained for written idioms (word-by-word presentation). If automatic semantic activation of the idiom constituent decays

rapidly, the time between idiom recognition and measurement of the semantic activation is crucial for observing respective effects. In general, it takes more time to present an idiom visually word-by-word (e.g., Experiment 3 of the present study or Rommers et al., 2013) than it takes to present a spoken version of the same idiom (e.g., Experiments 1 and 2). According to this timing difference, short-lived semantic activation might be still measurable at final constituents of spoken idioms, while it might have decayed already before the measurement in word-by-word reading (Rommers et al., 2013; Experiment 3 of the present study).

For priming experiments, where semantic spread presumably occurs in a bottom-up fashion, activation of semantic associates of idiom constituents was found for both modalities (Beck & Weber, 2016; Rabanus et al., 2008; Smolka et al., 2007). Since in those experiments the idiom constituent itself was always presented in the input, the recognition of the idiom and resulting preactivation of its constituents is not the only source of spreading semantic activation. This led us to conclude that there is an interplay of the processing mechanism (top-down vs. bottom-up) and the modality-related rate of presentation. In addition, the results imply that even the top-down prediction of idiom-final words is sufficient to activate single word meanings, but this is only measurable in the auditory modality in the present experiment. More research is needed to dissociate differences in these processes directly and to validate this explanation.

To account for individual idiom knowledge, we conducted an overt idiom recognition task in all experiments. In Experiment 1, participants had to choose the correct idiom completion among four alternatives. In Experiments 2 and 3, participants had to indicate whether the spoken or written strings were idioms. By performing these tasks, the participants might have been biased to activate canonical idiom forms only. However, if the participants would only have compared the incoming input with the activated idiom form, we should not have obtained a semantic activation of single word meanings in Experiments 1 and 2. In any case, general effects of the task cannot explain the differences between the results regarding activation of associates of idiom constituents of Experiments 2 and 3. Using the same task in both experiments, we show modality-related differences in online processing of idiomatic expressions.

The present results challenge the assumption that idioms are solely unanalyzed "long words" (Jackendoff, 2002). In general, our results support hybrid models such as the Superlemma Hypothesis (Sprenger et al., 2006), in which idioms are represented as both multi-word representations (*superlemmas*) and simple lemmas of single constituents on a lexical level. The hybrid nature of idioms may allow the linguistic system to rely on single constituent and multi-word representations in parallel (Arnon & Christiansen, 2017; Tremblay & Baayen, 2010). We hypothesize that the meanings of simple lemmas within idioms are available for only a short time after their activation.

Acknowledgements

We would like to thank Anne Bauch, Sara Beck, Stacie Boswell, Birte Herter, Babette Jakobi, Sören Koch, Tobias Kopp, Matteo Marks, Anne Rau, Ulrike Schild, and Charlotte Veil. We also warmly thank all participants. Furthermore, we would like to thank two anonymous reviewers for their helpful comments on a previous version of the manuscript.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 75650358 – SFB 833. The founding source had no involvement in the study.

ORCID iD

Ruth Kessler https://orcid.org/0000-0001-6443-6856

Notes

1. Since we planned to test children with the same material and paradigm in the future, we only selected highly familiar short idioms that German children would already be expected to know.
2. Since we planned to test children with the same material and paradigm in the future, we had to adapt the paradigm. Therefore, we chose this long preview window of the four printed words so that there was enough time to read all four words before the onset of the auditory stimuli. Since the specific idiom knowledge of children is very different, we only wanted to include idioms that are known to the individual children. Therefore, we chose an overt task where participants had to find the correct idiomatic completion.
3. As in Experiment 1, we chose an overt idiom recognition task because we wanted to conduct the same experiment with children. In order to account for differing idiom knowledge between children, we wanted to include only idioms that participants recognized correctly. A similar idiom recognition task was used in Qualls et al. (2003).

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518.

Anderson, J. E., & Holcomb, P. J. (1995). Auditory and visual semantic priming using different stimulus onset asynchronies: An event-related brain potential study. *Psychophysiology*, *32*(2), 177–190.

Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining L1–L2 differences. *Topics in Cognitive Science*, *9*(3), 621–636.

Beck, S. D., & Weber, A. (2016). Bilingual and monolingual idiom processing is cut from the same cloth: The role of the L1 in literal and figurative meaning activation. *Frontiers in Psychology*, *7*, 1305.

Berg, P., & Scherg, M. (1994). A multiple source approach to the correction of eye artifacts. *Electroencephalography and Clinical Neurophysiology*, *90*(3), 229–241.

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.

Cacciari, C., & Corradini, P. (2015). Literal analysis and idiom retrieval in ambiguous idioms processing: A reading-time study. *Journal of Cognitive Psychology*, *27*(7), 797–811.

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language*, *27*(6), 668–683.

Canal, P., Pesciarelli, F., Vespignani, F., Molinaro, N., & Cacciari, C. (2017). Basic composition and enriched integration in idiom processing: An EEG study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(6), 928–943.

Canal, P., Vespignani, F., Molinaro, N., & Cacciari, C. (2010). Anticipatory mechanisms in idiom comprehension: Psycholinguistic and electrophysiological evidence. In M. Balconi (Ed.), *Neuropsychology of Communication* (pp. 131–144). Springer.

Carrol, G., & Conklin, K. (2020). Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech*, *63*(1), 95–122.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*(1), 72–89.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61.

Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience*, *6*(3), 256–266.

Connolly, J. F., Phillips, N. A., & Forbes, K. A. (1995). The effects of phonological and semantic features of sentence-ending words on visual event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, *94*(4), 276–287.

Connolly, J. F., Phillips, N. A., Stewart, S. H., & Brake, W. G. (1992). Event-related potential sensitivity to acoustic and semantic properties of terminal words in sentences. *Brain and Language*, *43*(1), 1–18.

Connolly, J. F., Stewart, S. H., & Phillips, N. A. (1990). The effects of processing requirements on neurophysiological responses to spoken sentences. *Brain and Language*, *39*(2), 302–318.

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162.

Diaz, M. T., & Swaab, T. Y. (2007). Electrophysiological differentiation of phonological and semantic integration in word and sentence contexts. *Brain Research*, *1146*, 85–100.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495.

Federmeier, K. D., McLennan, D. B., De Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*, *39*(2), 133–146.

Ferreira, F., Foucart, A., & Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language*, *69*(3), 165–182.

Friedrich, C. K., & Kotz, S. A. (2007). ERP evidence of form and meaning coding during online speech recognition. *Journal of Cognitive Neuroscience*, *19*(4), 594–604.

Geeraert, K., Baayen, R. H., & Newman, J. (2017). Idiom variation: Experimental data and a blueprint of a computational model. *Topics in Cognitive Science*, *9*(3), 653–669.

Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition*, *8*(2), 149–156.

Glucksberg, S. (1993). Idiom meanings and allusional content. In C. T. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, Structure, and Interpretation* (pp. 3–26). Erlbaum.

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun—An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, *47*(4), 930–944.

Hagoort, P., & Brown, C. M. (2000). ERP effects of listening to speech compared to reading: The P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia*, *38*(11), 1531–1549.

Heil, M., Rolke, B., & Pecchinenda, A. (2004). Automatic semantic activation is no myth: Semantic context effects on the N400 in the letter-search task in the absence of response time effects. *Psychological Science*, *15*(12), 852–857.

Holsinger, E. (2013). Representing idioms: Syntactic and contextual effects on idiom processing. *Language and Speech*, *56*(3), 373–394.

Huettig, F., & McQueen, J. M. (2011). The nature of the visual environment induces implicit biases during language-mediated visual search. *Memory & Cognition*, *39*(6), 1068–1084.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171.

Ito, A., & Husband, E. M. (2017). How robust are effects of semantic and phonological prediction during language comprehension? *A visual world eye-tracking study*. https://doi.org/10.13140/RG.2.2.33577.49765

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156.

Kuperberg, G., & Wlotko, E. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35.

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. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Kyriacou, M., Conklin, K., & Thompson, D. (2020). Passivizability of idioms: Has the wrong tree been barked up? *Language and Speech*, *63*(29), 404–435.

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, *61*(3), 326–338.

Libben, M. R., & Titone, D. A. (2008). The multidetermined nature of idiom processing. *Memory and Cognition*, *36*(6), 1103–1121.

Mancuso, A., Elia, A., Laudanna, A., & Vietri, S. (2020). The role of syntactic variability and literal interpretation plausibility in idiom comprehension. *Journal of Psycholinguistic Research*, *49*(1), 99–124.

Marantz, A. (2005). Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review*, *22*(2–4), 429–445.

McGlone, M. S., Glucksberg, S., & Cacciari, C. (1994). Semantic productivity and idiom comprehension. *Discourse Processes*, *17*(2), 167–190.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494.

Molinaro, N., & Carreiras, M. (2010). Electrophysiological evidence of interaction between contextual expectation and semantic integration during the processing of collocations. *Biological Psychology*, *83*(3), 176–190.

Neely, J. H., O'Connor, P. A., & Calabrese, G. (2010). Fast trial pacing in a lexical decision task reveals a decay of automatic semantic activation. *Acta Psychologica*, *133*(2), 127–136.

Nieuwland, M. S. (2019). Do "early" brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience and Biobehavioral Reviews*, *96*, 367–400.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Qualls, C. D., Treaster, B., Blood, G. W., & Hammer, C. S. (2003). Lexicalization of idioms in urban fifth graders: A reaction time study. *Journal of Communication Disorders*, *36*(4), 245–261.

Quante, L., Bölte, J., & Zwitserlood, P. (2018). Dissociating predictability, plausibility and possibility of sentence continuations in reading: Evidence from late-positivity ERPs. *PeerJ*, *6*, e5717.

Rabanus, S., Smolka, E., Streb, J., & Rösler, F. (2008). Die mentale Verarbeitung von Verben in idiomatischen Konstruktionen. *Zeitschrift für Germanistische Linguistik*, *36*(1), 27–47.

Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience*, *25*(5), 762–776.

Saslow, M. G. (1967). Effects of components of displacement-step stimuli upon latency for saccadic eye movement. *Journal of the Optical Society of America*, *57*(8), 1024–1029.

Siyanova-Chanturia, A. (2015). On the "holistic" nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, *11*(2), 285–301.

Siyanova-Chanturia, A., Conklin, K., Caffarra, S., Kaan, E., & van Heuven, W. J. B. (2017). Representation and processing of multi-word expressions in the brain. *Brain and Language*, *175*, 111–122.

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*(2), 251–272.

Smolka, E., & Eulitz, C. (2020). Can you reach for the planets or grasp at the stars? Modified noun, verb, or preposition constituents in idiom processing. In S. Schulte im Walde & E. Smolka (Eds.), *The Role of*

Constituents in Multiword Expressions: An Interdisciplinary, Cross-lingual Perspective (pp. 179–204). Language Science Press.

Smolka, E., Rabanus, S., & Rösler, F. (2007). Processing verbs in German idioms: Evidence against the Configuration Hypothesis. *Metaphor and Symbol*, *22*(3), 213–231.

Snider, N., & Arnon, I. (2012). A unified lexicon and grammar? Compositional and non-compositional phrases in the lexicon. In S. Gries & D. Divjak (Eds.), *Frequency Effects in Language* (pp. 127–163). Mouton de Gruyter.

Sprenger, S., Levelt, W., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, *54*(2), 161–184.

Strandburg, R., Marsh, J., Brown, W., Asarnow, R., Guthrie, D., & Higa, J. (1993). Event-related potentials in high-functioning adult autistics: Linguistic and nonlinguistic visual information processing tasks. *Neuropsychologia*, *31*(5), 412–434.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(6), 643–662.

Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, *18*(5), 523–534.

Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast? *Memory & Cognition*, *37*(4), 529–540.

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, *83*(3), 382–392.

Titone, D. A., Lovseth, K., Kasparian, K., & Tiv, M. (2019). Are figurative interpretations of idioms directly retrieved, compositionally built, or both? *Canadian Journal of Experimental Psychology*, *73*(4), 216–230.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on Formulaic Language: Acquisition and Communication* (pp. 151–173). Continuum.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*(2), 569–613.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 155–172). John Benjamins.

Van Den Brink, D., Brown, C. M., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Journal of Cognitive Neuroscience*, *13*(7), 967–985.

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*(1), 131–147.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 394–417.

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., & Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *Journal of Cognitive Neuroscience*, *22*(8), 1682–1700.

Wray, A. (2005). *Formulaic Language and the Lexicon*. Cambridge University Press.

Appendix Materials Experiment 1, Experiment 2 and Experiment 3.

Item	Experiment 1	Idiom Body	Correct Completion	Related Word	Familiarity (N=20), Scale from 1 (not familiar) to 7 (highly familiar)		Relation of final word to figurative meanings (N=25), Scale from 1 (not related) to 7 (highly related)		Cloze Probability for correct idiomatic completion (N = 17)
					Mean	Standard Error	Mean	Standard Error	Percentage
1	x	Julia rutschte das Herz in die *Julia slid the heart in the*	Hose *pants*	Jacke *coat*	6.05	0.31	3	0.34	100%
2	x	Lena setzte alle Hebel in *Lena put all levers in*	Bewegung *move*	Sprünge *jumps*	6.35	0.21	4.92	0.33	88%
3	x	Marie stand Gabriel Rede und *Marie stood Gabriel speech and*	Antwort *answer*	Frage *question*	5.65	0.36	5.72	0.32	100%
4	x	Hannah schlug sich die Zeit um die *Hannah hit herself the time around the*	Ohren *ears*	Augen *eyes*	5.35	0.44	2.17	0.37	100%
5		Sarah band sich einen Klotz ans *Sarah tied herself a chunk to her*	Bein *leg*	Knie *knee*	5.25	0.32	3.4	0.34	100%
6		Paula malte den Teufel an die *Paula painted the devil on the*	Wand *wall*	Tür *door*	6.2	0.26	2.44	0.34	100%
7	x	Sofia brachte die Aufgaben unter Dach und *Sofia brought the duties under roof and*	Fach *shelf*	Schrank *cupboard*	5.6	0.37	3.08	0.36	100%
8	x	Annika war Balsam für die *Annika was balm for the*	Seele *mind*	Gefühle *emotions*	5.95	0.20	4.96	0.35	94%
9	x	Amelie hatte einen Frosch im *Amelie had a frog in the*	Hals *throat*	Rücken *back*	6	0.29	5.88	0.29	100%
10		Emma packte den Stier bei den *Emma grabbed the bull by the*	Hörnern *horns*	Zähnen *teeth*	4.35	0.41	3.58	0.37	88%

Appendix. (Continued)

Item	Experiment 1	Idiom Body	Correct Completion	Related Word	Familiarity (N=20), Scale from 1 (not familiar) to 7 (highly familiar)		Relation of final word to figurative meanings (N=25), Scale from 1 (not related) to 7 (highly related)		Cloze Probability for correct idiomatic completion (N = 17)
					Mean	Standard Error	Mean	Standard Error	Percentage
11		Laura streute Salz in die *Laura sprinkled salt into the*	Wunde *wound*	Narbe *scar*	6.45	0.18	5.36	0.33	94%
12		Johanna hatte Tomaten auf den *Johanna had tomatos on her*	Augen *eyes*	Ohren *ears*	5.1	0.35	4.92	0.34	88%
13	x	Isabell hatte Schmetterlinge im *Isabell had butterflies in the*	Bauch *stomach*	Arm *arm*	6.5	0.24	4.96	0.34	100%
14	x	Jasmin lebte wie die Made im *Jasmin lived like the grub in the*	Speck *bacon*	Käse *cheese*	4.4	0.41	3.12	0.36	94%
15		Lisa packte das Übel an der *Lisa grabbed the evil at the*	Wurzel *root*	Blüte *blossom*	4	0.36	4.12	0.34	65%
16	x	Melina ließ die Kirche im *Melina left the church in the*	Dorf *village*	Feld *field*	5.15	0.32	2.36	0.29	94%
17		Eva war das schwächste Glied der *Eva was the weakest link if the*	Kette *chain*	Linie *line*	5.75	0.28	3.8	0.36	76%
18	x	Helena fiel ein Stein vom *Helena fell a stone from the*	Herzen *heart*	Magen *stomach*	6.8	0.12	4.96	0.31	100%
19		Tabea hatte mehr Glück als *Tabea had more luck than*	Verstand *mind*	Geist *spirit*	6.15	0.28	5.76	0.31	94%
20	x	Nora hatte ihr Herz am rechten *Nora had her heart on the right*	Fleck *spot*	Platz *place*	6.35	0.20	2.92	0.35	100% 0%
21		Lukas war am Ende seines *Lukas was at the end of his*	Latein *Latin*	Spanisch *Spanish*	5.25	0.37	3.28	0.43	71%

Item	Experiment 1	Idiom Body	Correct Completion	Related Word	Familiarity (N=20), Scale from 1 (not familiar) to 7 (highly familiar)		Relation of final word to figurative meanings (N=25), Scale from 1 (not related) to 7 (highly related)		Cloze Probability for correct idiomatic completion (N = 17)
					Mean	Standard Error	Mean	Standard Error	Percentage
22		Leon nahm eine Mütze voll	Schlaf	Traum	4.15	0.35	6.16	0.32	65%
		Leon took a hat full	*sleep*	*dream*					
23		Max schlug zwei Fliegen mit einer	Klappe	Kiste	6.35	0.17	3.4	0.37	88%
		Max hit two flies with one	*flap*	*box*					
24	x	Nico erblickte das Licht der	Welt	Venus	5.75	0.32	5.68	0.32	88%
		Nico beheld the light of the	*world*	*venus*					
25		Simon hatte bei Nina einen Stein im	Brett	Nagel	4.2	0.47	2	0.22	94%
		Simon had with Nina a stone in the	*board*	*nail*					
26	x	Robin legte für Natalie die Hand ins	Feuer	Holz	6.45	0.15	3	0.32	100%
26		*Robin put for Natalie the hand in the*	*fire*	*wood*					
27	x	Linus verlor den Boden unter den	Füßen	Händen	6.35	0.21	3.84	0.38	100%
		Linus lost the floor under the	*feet*	*hands*					
28		Lars fand in jeder Suppe ein	Haar	Kinn	4.85	0.40	2.96	0.34	100%
		Lars found in every soup a	*hair*	*chin*					
29	x	Jannis fiel die Decke auf den	Kopf	Bart	6.45	0.15	3.6	0.37	100%
		Jannis fell the ceiling on the	*head*	*beard*					
30		Daniel brachte das Fass zum	Überlaufen	Austrocknen	6.55	0.14	4.88	0.41	100%
		Daniel brought the barrel to	*overflow*	*dry-out*					
31		Florian brachte Marius an den Rand der	Verzweiflung	Angst	6.4	0.22	6.12	0.23	82%
		Florian brought Marius to the edge of the	*desperation*	*fear*					

Appendix. (Continued)

Item	Experiment 1	Idiom Body	Correct Completion	Related Word	Familiarity (N=20), Scale from 1 (not familiar) to 7 (highly familiar)		Relation of final word to figurative meanings (N=25), Scale from 1 (not related) to 7 (highly related)		Cloze Probability for correct idiomatic completion (N = 17)
					Mean	Standard Error	Mean	Standard Error	Percentage
32	x	Julian hielt den Kopf über *Julian held the head above*	Wasser *water*	Regen *rain*	5.1	0.27	3.42	0.37	100%
33		Erik fiel mit der Tür ins *Erik fell with the door in the*	Haus *house*	Zelt *tent*	5.95	0.32	2.92	0.32	100%
34	x	Moritz begab sich in die Höhle des *Moritz repaired himself to the hole of the*	Löwen *lion*	Hasen *bunny*	5.45	0.37	3.44	0.4	100%
35		Fabian stellte Emils Geduld auf die *Fabiel put Emils patience*	Probe *test*	Übung *exercise*	6.1	0.23	5.16	0.33	100%
36	x	Timo brachte den Stein ins *Timo brought the stone in the*	Rollen *rolling*	Kugeln *rolling* (syn.)	5.2	0.32	4.4	0.33	100%
37	x	Hannes ließ die Katze aus dem *Hannes let the cat out of the*	Sack *bag*	Korb *basket*	5.75	0.31	2.76	0.31	94%
38		Antons Entscheidung stand auf Messers *Antons decision stood on knifes'*	Schneide *blade*	Klinge *blade* (syn.)	5.1	0.43	3.8	0.39	100%
39		Dennis packte die Gelegenheit beim *Dennis grabbed the occasion at the*	Schopfe *tuft*	Scheitel *parting*	4.9	0.35	2.92	0.34	100%
40		Emil war das fünfte Rad am *Emil was the fifth wheel on the*	Wagen *car*	Zug *train*	6.5	0.15	2.92	0.41	94%