Language and Cognition (2025), 17, e55, 1–25
doi:10.1017/langcog.2025.10013

ARTICLE

‘Well, if you want to play with fire, make sure the flames are tamed’: investigating the integration of idiom literal completions across literal and creative contexts

Irene Pagliai

RTG 2636: Form-meaning mismatches, University of Göttingen, Göttingen, Germany
Email: irene.pagliai@uni-goettingen.de

(Received 24 January 2025; Revised 23 May 2025; Accepted 29 May 2025)

Abstract

This study investigates the integration of literal completions of idiomatic multiword expressions (MWEs) into two linguistic contexts: one promoting a literal interpretation and the other a figurative one, requiring reinterpretation to align with figurative bias. Sixteen Italian idioms were distributed in two groups by their Potential Idiomatic Ambiguity (PIA) score, an index of literal plausibility, decomposability and transparency. Using experimental dialogues, the study tested whether high-PIA idioms receive higher acceptability ratings across both contexts than low-PIA idioms. Eighty-four Italian-speaking participants rated idiom literal completions within literal and figurative contexts. Results show that literal completions of high-PIA idioms integrate better across contexts, while those of low-PIA idioms receive lower ratings and have longer combined reading and rating times. This supports hybrid models of idiom processing, emphasizing the role of idiomatic features and context in balancing figurative and compositional interpretations. This study also marks an initial effort to experimentally trace systematicity within idiomatic wordplay, challenging the idea that it lacks relevance for linguistic research while outlining limitations and directions for future work.

Keywords: acceptability ratings; ambiguity; creativity; crowdsourcing; experiment; idioms; idiom hybrid models; Italian; literal plausibility; wordplay

1. Introduction

Multiword expressions (MWEs) are a heterogeneous group of lexical units composed of at least two linguistic elements that co-occur more frequently than expected by chance (Farahmand & Nivre, 2015). Within this group, idioms are distinguished not

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

only by their formal and statistical idiosyncrasy but also by their non-compositional, figurative semantics (Cacciari & Tabossi, 2014; Monti & di Buono, 2019; Nunberg et al., 1994; Wagner, 2021). Consider, in this regard, example (1), an occurrence of the idiom *break the ice*:

(1) The preliminary chat which usually prefaces the interview proper is intended to *break the ice* and give both sides a chance to size each other up.
BNC

The MWE is typically associated with the figurative meaning 'overcome initial awkwardness', and the example illustrates the prototypical scenario in which the co-text surrounding an idiom only refers to its figurative meaning, with no allusion to the literal–compositional interpretation. Consider now the occurrences of *break the ice* in (2) and (3) (original in Italian, in which the equivalent MWE is *rompere il ghiaccio*):

(2) The **blizzard** had passed but the sky threatened more. We gingerly *broke the ice* in the **washing bowl**, washed, changed and joined the others in the small refectory below.
BNC

(3) 'Quanti anni…' gli dissi tanto per *rompere il ghiaccio*. Proprio di **ghiaccio** si trattava… Era più **freddo** di un **iceberg**!
('How many years…' I told him just to break the ice. It was ice indeed… He was colder than an iceberg!)
CORIS

Compared to the prototypical example, the co-texts now include references to the semantic domain of the MWE's literal constituents, albeit for two different reasons. In (2), *break the ice* is to be interpreted in its literal–compositional sense, whereas in (3), the speaker creatively draws on the idiom's literal constituents to enrich its figurative meaning.

The current study experimentally investigates the idea that idioms possessing inherent characteristics such as high literal plausibility, decomposability and transparency should be especially prone to appear in contexts where the literal aspect of the idiom is pivotal. These contexts include both strictly literal–compositional scenarios, such as (2), and creative contexts like (3), where the literal meaning of the idiom is extended to enrich its figurative interpretation.

Literal plausibility refers to the idiom compositionally denoting a meaning consistent with common world knowledge (Citron et al., 2016; Wagner, 2021). Decomposability describes the possibility of creating associative mappings between the idiom's figurative meaning and the literal constituents (Fadlon et al., 2013; Sailer, 2021). Finally, transparency refers to how clear the relationship between literal and figurative meanings is, making the latter easier to infer from the former when transparency is high (Moreno, 2005; Sailer, 2021).

To select idioms with varying degrees of the three variables described, a sample of Italian idioms is drawn from the Normed lexicon of English and Italian idioms. In this dataset, each idiom is assigned a Potential for Idiomatic Ambiguity (*PIA*) score, an index calculated as the sum of the average ratings for literal plausibility, decomposability and transparency based on the norming study detailed in Pagliai (2023a).

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

The index name reflects the degree of co-activation and resulting competition between figurative and literal interpretations unique to each idiom. Idioms with high literal plausibility exhibit an inherent semantic ambiguity, as they represent 'two different regions of meaning in the semantic space' (Wagner, 2021). As a result, these idioms can occur in both their figurative sense (example (1)) and their literal interpretation (example (2)). Moreover, when an idiom is also highly decomposable and transparent, the connections between the two 'regions of the semantic space' are stronger and more explicit. This may enable speakers to strategically leverage these links, navigating the literality–figurativity continuum with greater ease (Moreno, 2005). Such flexibility is argued to foster creative uses, as seen in example (3), where the coldness and thickness of ice metaphorically symbolize the challenge of overcoming social awkwardness.

Before thoroughly addressing the objectives of the present study in Section 1.3, we first outline in Section 1.1 experimental evidence concerning the three variables of interest for this work and their interaction with linguistic context. Subsequently, Section 1.2 provides an overview of research on creative idiomatic occurrences, emphasizing how an idiom's literality can be manipulated or functionally exploited in discourse.

1.1. Idiom features and linguistic context in hybrid models of idiom processing

The interplay between inherent idiom features and linguistic contexts has emerged as a key area of study in (psycho)linguistics, especially given the increasing evidence supporting hybrid models of idiom representation and processing. Models such as the Configuration Hypothesis (Cacciari & Tabossi, 1988), the Constraint-Based Model (Libben & Titone, 2008; Titone & Libben, 2014) and the Superlemma Theory (Sprenger et al., 2006) propose that idiom comprehension engages both compositional (bottom-up) and holistic (top-down) processing strategies. The necessity to postulate such hybrid models stems from the collection of apparently conflicting scientific evidence. On the one hand, idioms are processed faster than their literal equivalents+ ('idiom superiority effect'), supporting the direct retrieval of figurative meanings (Gibbs, 1980; McGlone et al., 1994). On the other hand, idioms also exhibit priming effects on semantically related literal words (Hillert & Swinney, 2001) and allow for the retrieval of the full idiom to be triggered by its individual components (Sprenger et al., 2006), highlighting therefore the role of compositional mechanisms.

Hybrid models resolve the tension between the apparently contradictory findings by postulating that the balance between the two routes of processing is dynamically adjusted according to a number of factors including the task to be performed (van Ginkel & Dijkstra, 2019), the cognitive characteristics of the individuals (Arnon & Lavidor, 2022; Tilmatine et al., 2021), the specific time points at which the stimulus is processed (Senaldi & Titone, 2024; Titone & Libben, 2014) and crucially for the present study, the interaction between linguistic contexts and inherent idiom features (Beck & Weber, 2020; Senaldi et al., 2022; Senaldi & Titone, 2024).

The degree of literal plausibility in idiomatic MWEs is central to the investigation of idiom processing. Idioms having literal meanings consistent with common world knowledge exhibit co-activation of figurative and literal interpretations, creating internal competition between the two. In contrast, literally implausible idioms are heavily biased toward figurative interpretation due to the inconsistency of their compositional meanings (Mancuso et al., 2019; Wagner, 2021).

In their study, Beck and Weber (2020) focused on how literal plausibility and context interact in idiom processing using a self-paced reading task. Their findings showed that literal plausibility modulates context effects on comprehension. Plausible idioms exhibited binary processing, with faster reading times in both figurative- and literal-biasing contexts due to simultaneous activation of both interpretations. In contrast, implausible idioms favored unitary processing, with faster reading times only in figurative-biasing contexts. They thus conclude that comprehension is facilitated only when context-driven expectations align with the idioms' intrinsic characteristics.

In a similar fashion, experimental studies have also demonstrated that an idiom's decomposability and transparency significantly influence the salience of its compositional-literal interpretation. The seminal eye-tracking study by Titone and Connine (1999) showed that highly decomposable idioms are processed faster in both figurative and literal contexts due to the relatedness between their figurative and literal meanings, which synergistically reinforces processing. More recently, the eye-tracking study of Titone et al. (2019) revealed that highly decomposable idioms generate pronounced competition between literal and figurative interpretations, making it especially difficult for participants to suppress literal constituents in favor of holistic figurative meanings. Similarly, van Ginkel and Dijkstra (2019)'s lexical decision task highlighted that transparency facilitates access to both figurative and literal meanings, with the authors noting that 'for native speakers, figurative meaning generally dominates, but literal components also receive partial activation, particularly when idioms are transparent'. Wagner (2021) finally corroborates this perspective, observing that opaque and non-decomposable idioms are more likely to be processed as single units, whereas highly decomposable and transparent idioms may engage both literal and figurative meanings simultaneously.

1.2. Idiomatic creativity

Building on Langlotz (2006), we understand idiomatic creativity as the diverse contextual adaptation of conventional idiomatic expressions to meet novel communicative purposes. Aligned with the multidetermined and dynamic view of idioms proposed by hybrid models, recent research into idiomatic creativity challenges the traditional view of idioms as fixed and non-extendable lexical units (Fellbaum, 2019). Indeed, the investigation of idiomatic flexibility has grown to encompass multiple levels of linguistic analysis, including morpho-syntactic, lexical and semantic–pragmatic dimensions (Bargmann et al., 2021; Carrol & Segaert, 2024; Fellbaum, 2019; Kyriacou et al., 2019; Langlotz, 2006; Mancuso et al., 2019). This growing body of research highlights the systematic potential for varying and enriching idiomatic MWEs, further underscoring their dynamic and adaptable nature.

For instance, earlier studies based on the Idiom Decomposition Hypothesis (Gibbs & Nayak, 1989) posited a positive correlation between an idiom's decomposability and its morphosyntactic flexibility, suggesting that non-decomposable idioms are especially frozen and resistant to significant syntactic changes like passivization (Langlotz, 2006). More recent data-driven research has significantly revised this assumption: corpus-based and experimental studies consistently demonstrate that idioms can generally undergo passivization while retaining their figurative meaning (Fellbaum, 2019; Kyriacou et al., 2019; Maher, 2013; Mancuso et al., 2019).

The analysis of lexical variation in idioms has also revealed a richer and more systematic creativity than previously recognized. Fellbaum (2019) observes that

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

lexical substitutions in idiomatic MWEs follow principles of phonological similarity, contextual relevance and paradigmatic semantic relations. She further argues that this systematicity extends to creative cases where 'both literal and figurative readings of idioms are accessed', such as zeugma involving the 'conjunction of canonical idiom components with non-idiomatic components within a single idiom' (p. 762). Such adjustments often produce additional pragmatic effects like irony, and are thus frequently regarded as unsystematic wordplay of limited relevance to linguistic analysis (Langlotz, 2006; Mel'čuk, 2014). However, Fellbaum's corpus-based analysis challenges this view, arguing that while these adjustments are indeed humorous and context-dependent, they are far from arbitrary: 'because such word play is highly systematic, it cannot be dismissed as a linguistic epiphenomenon and must instead be included in the study of idioms' (Fellbaum, 2019, p. 764).

These references to idiomatic wordplay prompt a closer examination of the semantic—pragmatic creativity central to this study. In preparing the items for this experiment, contexts similar to (3) are replicated by first preceding the idiom with a figurative-biasing context and then following it with one involving 'literal-scene manipulation' (Langlotz, 2006). This shift reflects an extension of the idiom's figurative meaning through manipulation of its literal components. Langlotz (2006, p. 207) describes this type of idiomatic creativity as the strategic exploitation of an idiom's literal components to adapt its figurative meaning to a given context. More specifically, since prior recognition of the idiomatic string and its figurative meaning is crucial for achieving the creative effect, the author defines this type of manipulation as 'parasitic elaboration' (p. 202), arguing that this nonsystematic idiomatic wordplay leverages the literal scene to generate alternative interpretations while preserving the idiomatic meaning.

Additionally, we propose that another category of idiomatic manipulation aligns closely with the type of creativity relevant for this study: 'conjunction modification' (Ernst, 1981). Following Bargmann et al. (2021, p. 249), conjunction modification refers to a type of idiom manipulation in which a modifier applies to the literal meaning of a noun within the idiom, without disrupting its figurative meaning. Rather, this process semantically introduces an additional, independent proposition alongside the idiom's figurative interpretation. One example provided by the authors is 'he kicked the golden bucket' (p. 255), where the idiomatic interpretation ('he died') is combined with a literal one ('the bucket was gold'), which is then reinterpreted figuratively as 'he was rich'. In the present experiment, in its creative variation, the idiom is first preceded by a context that biases its figurative interpretation and then is followed by one that uses its literal components to extend the figurative meaning. Accordingly, the creative contexts created in this study can be seen as akin to cases of conjunction modification already split into two propositions.

In conclusion, this study examines whether the systematicity of idiomatic creativity found in other linguistic domains extends to the semantic–pragmatic domain, particularly in cases labeled as wordplay. To this end, we echo Bargmann et al. (2021, p. 253), who claim that 'even if conjunction modification were to fall within "word play" (however we define it), it would still involve language and thus should be analysable'.

1.3. The present research

The overarching aim of this study is to experimentally investigate 'the complex interaction and integration of context, literal scene, and idiomatic meaning in idiom

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

processing' (Langlotz, 2006, p. 245). More specifically, the main research question addresses whether idiom literal completions are integrated in both literal (example (2)) and figurative (example (3)) contexts, depending on the PIA of the idioms.

Providing an answer to this question addresses several central aspects in idiom research. First, it contributes to further testing hybrid models of idiom representation and processing. High-PIA idioms, characterized by high literal plausibility, decomposability and transparency, are expected to engage compositional processing more readily in contexts that emphasize their literal meaning, thereby facilitating smoother contextual integration. In contrast, the opposite traits of low-PIA idioms should favor holistic, figurative processing, which arguably makes them more difficult to integrate into contexts that highlight the literal meaning of the MWE.

Secondly, this study seeks to address a gap in linguistic research on creative uses of idiomatic language. To the best of our knowledge, the type of semantic–pragmatic idiomatic creativity explored here has been considered exclusively in qualitative theoretical or corpus-based studies, without being subjected to experimental investigation.[1] This is motivated by the difficulty of constructing linguistic contexts that are both sufficiently extensive and naturalistic while remaining controllable enough to test this form of idiomatic creativity. This methodological challenge is tackled through the fictional dialogues developed for this study. Such dialogues provide linguistic contexts that feature idiom literal-scene manipulation, while balancing naturalness, extensiveness and experimental controllability.

Finally, by comparing creative instances of idiomatic language with idiom literal uses, we seek to assess the systematicity (or lack thereof) of literal-scene manipulation instances. Based on previous research (see Section 1.1), we confidently expect that literal completions of high-PIA idioms will show greater integration ease into fully compositional-literal contexts than those of low-PIA idioms. However, when it comes to creative-figurative contexts, the study takes a more exploratory approach: do the intrinsic characteristics of idioms influence the ease of integrating idiom literal completions into contexts requiring their figurative reinterpretation, or do they not? We believe that a positive answer would indicate the emergence of systematicity rooted in the intrinsic characteristics of idioms, even in creative idiomatic usage at the semantic–pragmatic level. This would further corroborate what has already been suggested by Bargmann et al. (2021, p. 277), namely that 'the processes involved are far from unsystematic and should not be dismissed as mere linguistically inexplicable creative wordplay'.

2. Methods

2.1. Materials

A selection of 16 Italian idioms was drawn from the Normed lexicon of English and Italian idioms (Pagliai, 2023b). Although modest in size, this item set was limited to

[1] An attentive reviewer drew our attention to the recent experimental study by Carrol and Segaert (2024), which investigates idiomatic creativity via lexical manipulation. Their study focuses on idiom variants created by replacing the canonical final noun of an idiom with a semantically related noun, maintaining the underlying figurative meaning. By contrast, the present study explores a distinct type of idiomatic creativity: here, the idiom remains formally intact, and creativity is realized at the semantic–pragmatic level, since participants are asked to reinterpret literal completions of idioms in light of a figurative bias, generating meaning through context-driven resemantization rather than lexical change.

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

preserve methodological rigor. The stimuli in the present study required careful balancing of ecological validity and experimental control (see below). While a larger sample could have increased statistical power, it would have introduced greater variability, thereby reducing the precision and reliability of the experiment. The smaller item set also reflects the study's partly exploratory nature, which future work may build upon and refine (see Section 4).

The 16 idioms were chosen from the three most frequent syntactic categories in the dataset: V-NP, V-NP-PP and V-PP. Selection was guided by the continuum of the PIA index, whose distribution was split into 16 percentiles, and one idiom was chosen from each. This approach guarantees that the selected idioms reflect the entire PIA spectrum. Additionally, the idioms have high mean ratings for familiarity ($M = 4.37$, $SD = 0.32$) and meaningfulness ($M = 4.58$, $SD = 0.31$). Tables 1 and 2 report high- and low-PIA idioms, respectively (individual values for *Literal Plausibility*, *Decomposability* and *Transparency* are omitted for brevity. Details are available in the publicly accessible dataset).

To construct the experimental items, the idioms were embedded in a dialogical context, represented by two lines exchanged between speakers A and B. Simulated dialogue was chosen based on the idea that linguistic creativity tends to emerge more frequently in informal, noninstitutionalized and symmetrical social contexts (Carter, 2015; Carter & McCarthy, 2004). In this sense, simulated dialogue mimics everyday speech, where language is more flexible and conducive to the pragmatic effects typically associated with linguistic creativity.

The structure of the dialogues has been kept as consistent as possible across all items. In the first line, speaker A introduces a situation with 'you know…' and sets up

Table 1. High-PIA idioms

Item	Idiom	Translation	Meaning	PIA
1	prendere per mano	take by hand	guide	12.66
2	giocare a carte scoperte	play with uncovered cards	act without hiding	12.03
3	gettare la maschera	throw the mask	show one's intentions	11.31
4	rimettersi in piedi	get back on one's feet	feel good again	10.90
5	rompere il ghiaccio	break the ice	overcome initial awkwardness	10.41
6	parlare al muro	talk to the wall	waste words	9.89
7	forzare la mano	force the hand	force	9.38
8	prendere il toro per le corna	take the bull by the horns	tackle a situation head-on	9.00

Table 2. Low-PIA idioms

Item	Idiom	Translation	Meaning	PIA
9	alzare il gomito	raise the elbow	drink too much alcohol	8.73
10	prendere per il naso	take by the nose	mock	8.31
11	morire dietro	die behind	desire intensely	7.92
12	sputare il rospo	spit the toad	confess	7.63
13	toccare corde sensibili	touch sensitive strings	deal with delicate topics	7.28
14	fare acqua da tutte le parti	make water from all sides	fail to function	6.93
15	costare un occhio della testa	cost an eye of the head	be very expensive	6.24
16	fare l'avvocato del diavolo	do the devil's lawyer	object for the sake of it	5.35

Table 3. Examples of experimental items: high-and low-PIA idioms

Idiom	PIA	Condition	A	B
prendere per mano (take by hand)	High	FIG	You know, at school I have to follow a child with learning disabilities. I have to guide her in learning basic math skills.	Well, if you have to take her by hand, make sure you have a firm grip.
		LIT	You know, next week we are going with the whole family to Tokyo. I'm terrified of those huge street crossings, especially for my 4-year-old daughter.	
fare acqua da tutte le parti (make water from all sides)	Low	FIG	You know, I thought I had a good idea for the doctoral project. But at the moment, I do not think it's working.	Well, if it has to make water from all sides, make sure you have large buckets.
		LIT	You know, yesterday a pipe burst in our bathroom and it's continuing to leak. The floor is a mess, it's completely wet!	

one of two biasing conditions: figurative (FIG) or literal (LIT). Speaker B's reply is always biased toward the literal meaning of the idiomatic expression, following the pattern 'well, if you want/have to [MWE], then make sure [reference to the MWE's literal components]'. Refer to Table 3, which provides (English-translated) examples of two experimental items, one with a high-PIA idiom and one with a low-PIA idiom.

In addition to the 16 critical items, 28 filler items were developed in two categories. Fourteen items address syntactic attachment ambiguity, where a sentence's structure allows for multiple interpretations (Hindle & Rooth, 1991). These fillers have three conditions: preferred (PREF), where B's response confirms the most probable syntactic parsing with the PP adverbial interpretation ('A: You know, yesterday I was playing a fighting video game. My character had a bat, and the opponent had a spear'. 'B: Well, then your character hit the opponent with the bat'.); dispreferred (DISP), a garden-path sentence (Slattery et al., 2013) with the PP attributive reading requiring syntactic reanalysis (A's line is the same as in PREF. 'B: Well, then your character hit the opponent with the spear'.); ambiguous (AMB), where two syntactic structures are equally plausible ('A: You know, yesterday I was playing a fighting video game. My character had a bat, and the opponent also had a bat'. 'B: Well, then your character hit the opponent with the bat'.).

The remaining 14 filler items focus on the subjunctive–indicative alternation in Italian (Digesto, 2022; Zucchini, 2023), and have two conditions: correct (RIGHT), where the subordinate clause employs the subjunctive mood, following normative Italian usage ('A: You know, my husband is spending far too much money on furnishing the house. I am desperate!' 'B: Well, if you think your husband is (*sia* in Italian) a spendthrift, make sure the joint account card is (*sia* in Italian) no longer in his possession!'); and incorrect (WRONG), where the indicative mood replaces the subjunctive (A's line remains the same as in RIGHT. 'B: Well, if you think your husband is (*è* in Italian) a spendthrift, make sure the joint account card is (*è* in Italian) no longer in his possession!').

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

These engaging filler items serve a dual purpose. First, they divert participants' attention in a meaningful way. Second, they act as a tool for monitoring the overall performance of the participants during the experiment, as ratings for the syntactic ambiguity items are anticipated to be lower for DISP than for AMB and PREF, with AMB potentially rated lower than PREF. Similarly, ratings for the subjunctive items are anticipated to be lower for WRONG than for RIGHT.

Finally, following a Latin square design, all 102 experimental items prepared were counterbalanced across six experimental lists, each containing 44 items.

2.2. Participants

Eighty-four participants were recruited through the crowdsourcing platform Prolific (www.prolific.com). Recruitment was restricted to native Italian speakers residing in Italy, with no language-related disorders or literacy difficulties. Additional filters ensured participants had completed at least 50 prior submissions on Prolific with a minimum approval rate of 90%.

Of the 84 participants, 52 were male, 31 female and 1 identified as other. Seventy-three participants were right-handed, 10 left-handed and 1 ambidextrous. The average age was 35.4 years (*SD* = 11.1, range = 21–63).

2.3. Experimental procedure

Before taking part, all participants were informed via Prolific of the experiment's essential details. The requirements included being of legal age, having Italian as their native language and completing the task on a computer (excluding tablets and smartphones). They were also instructed not to interrupt the experiment, if possible. Participants were also briefed on the nature of the task, the different sections of the experiment and the demographic information they would be asked to provide. As for duration and remuneration, they read that the experiment would take approximately 15 minutes (based on pre-test data), with a compensation of £9 per hour, categorized by Prolific as 'good'. Accordingly, participants received £2.25 for their contribution. This preliminary briefing allowed participants to make an informed decision about whether or not to proceed. Participants then clicked a link to begin the experiment, which was created using PennController for Internet-Based Experiments (PCIbex, Zehr & Schwarz, 2018), a free, JavaScript-based platform for designing, hosting and managing online behavioral experiments.

Participants first read the informed consent and could only proceed after agreeing to it. They then provided demographic information (see Section 2.2). Next, the experiment instructions were displayed, explaining that participants would read fictional dialogues between two speakers, A and B. Initially, only A's part appeared on the screen; after reading it, participants pressed the space bar to reveal B's response. This division allowed the tracking of reading times for each section of the dialogue separately.

The task was explained as follows: participants were asked to 'evaluate the acceptability of B's answer based on the context provided by A' using a Likert scale from 1 (completely unacceptable) to 7 (completely acceptable), which appeared alongside B's response. Importantly, the time recorded for B contexts reflects both the reading of B's response and the subsequent act of providing a rating. This was an intentional design choice, aimed at capturing not only the time required for reading but also any additional

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

cognitive effort involved in the evaluative process, such as hesitation or reinterpretation. For ease of reference, we occasionally refer to the time measures for A and B contexts as 'reading times' throughout the manuscript. However, readers should be aware that B times reflect a composite of reading and rating times.

Once participants rated B's response, they could either press the space bar again or click a button to proceed to the next dialogue. The instructions themselves included two examples (one with a high rating and one with a low rating), but participants were also given the opportunity to practice with two additional example items before the actual trial began.

Participants were assigned to one of the six experimental lists, which are balanced by participant count (14 per list). The item presentation order was automatically randomized at each session. Upon completing the experimental trial, participants were redirected to Prolific via a link to confirm task completion, so as to receive compensation.

The experiment is openly accessible at this demonstration link. Additionally, by clicking 'Click here to edit a copy in the PCIbex Farm' in the ocher bar at the top left, readers can access the PCIbex experiment script, along with all related materials, including the csv file containing all experimental items (in Italian) distributed in the Latin square design.

2.4. Statistical analysis

For clarity in statistical analysis, variable names are written in italics (*Condition*, *PIA*, etc.) and, when dealing with categorical variables, levels are written in capital letters (FIG, LIT, HIGH, LOW, etc.).

All statistical analyses were performed in the R Statistical Software (v4.3.2, R Core Team, 2023). Cumulative link mixed models (clmms, library `ordinal`, Christensen, 2023, estimates expressed in probits) were used to model ordinal data (that is ratings), while continuous data (that is time data) were modeled using linear mixed-effects models (lmms, libraries `lme4` and `lmerTest` Bates et al., 2015; Kuznetsova et al., 2017).

Statistical results for the filler items are presented in Section 3.1. For both filler categories (syntactic ambiguity and modal alternation), the analysis explored whether ratings are predicted by the categorical variable *Condition*, split into the levels AMB, PREF and DISP for the syntactic item clmm, and into WRONG and RIGHT for the modal alternation one. In both models, *Condition* was encoded using treatment contrast, with AMB as the reference level for syntactic items and WRONG for modal items. Each model includes a maximal random effects structure (Barr et al., 2013), with by-participant and by-item random intercepts and slopes for *Condition.*

The results on idiomatic item ratings are presented in Section 3.2. Pertinent variables in this case are *Condition*, here divided into FIG and LIT and *PIA*. Unless otherwise specified, *PIA* is treated as a binary, categorical variable split into LOW and HIGH.[2] After inspection of data distribution and summary descriptive statistics, a

[2] A reviewer rightly questioned the choice of treating the variable *PIA* as categorical, given its underlying continuous nature as an index. To address this point, an ad hoc Appendix (see the 'Supplementary Materials' of the manuscript) has been prepared, where we explain that the binary coding of *PIA* was preferred due to the limited number of idioms, which likely provides insufficient coverage of the PIA continuum and limits the potential for fine-grained, robust analyses using a continuous predictor. The Appendix furthermore includes a comparison between two clmms, one with categorical *PIA* and one with continuous *PIA*, showing that the continuous version does not improve model fit.

clmm is run aiming to investigate the effect of *Condition*, *PIA* and their interaction on the ratings obtained. Both variables were coded using sum contrasts. The random effects structure is specified to be maximal, with random intercepts and slopes for the *Condition***PIA* interaction by participant, and a random intercept and slope for *Condition* by item.

The correlations presented in Section 3.2.1 explore the relationships between each variable comprising the PIA index and participant ratings across both conditions. Specifically, the associations of the numerical variables *Literal Plausibility*, *Decomposability* and *Transparency* with experimental ratings were investigated using the nonparametric Spearman's rank correlation coefficient.

Finally, results on reading times are presented in Section 3.3. Time data were first normalized by dividing reading times in ms by the number of words in each section of the dialogical stimuli (A, B and whole item). Due to the positive skewness of the time data (see Section 3.3), outliers were removed using the interquartile range (IQR) method: data points lying below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR were considered outliers and replaced with NA values, ensuring they did not influence further analyses.

Descriptive statistics on time data include average times for each section (A, B, whole item) of the short dialogues, along with a focused analysis of idiomatic items. Specifically, this examines whether the combined reading and rating times for context B vary as a function of *Condition* and *PIA*. To this end, a lmm was employed with both variables as predictors. They were coded using treatment contrasts, with LIT and HIGH serving as the reference levels for *Condition* and *PIA*, respectively. The model includes a maximal random effects structure, comprising by-participant random intercepts and slopes for *Condition* and *PIA*, as well as by-item random intercepts and slopes for *Condition.*

3. Results

3.1. Filler items

The results for the filler items are positive, aligning well with expectations. For the syntactic ambiguity items, average ratings for the DISP condition (M = 3.12, SD = 1.89) are lower than those for both the AMB (M = 5.05, SD = 1.74) and PREF (M = 5.11, SD = 1.78) conditions. This finding is further supported by the clmm run, which reveals a significant difference in ratings between DISP and AMB conditions ($b = -1.498$, $SE = 0.213$, $z = -7.020$, 95% $CI = [-1.916, -1.079]$, $p < .001$). However, ratings for the PREF condition do not significantly differ from those for the AMB condition ($b = 0.076$, $SE = 0.169$, $z = 0.450$, 95% $CI = [-0.256, 0.408]$, $p = .653$).

For the subjunctive–indicative alternation items, average ratings for the WRONG condition (M = 5.03, SD = 1.86) are lower than those for the RIGHT condition (M = 6.26, SD = 1.24). Again, the significance of the difference was confirmed by the clmm, which shows that ratings for the RIGHT condition are significantly higher than those for the WRONG condition ($b = 1.164$, $SE = 0.173$, $z = 6.741$, 95% $CI = [0.826, 1.502]$, $p < .001$).

3.2. Idiomatic items

Figure 1 shows the sorted distributions of raw ratings across the FIG and LIT conditions.

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

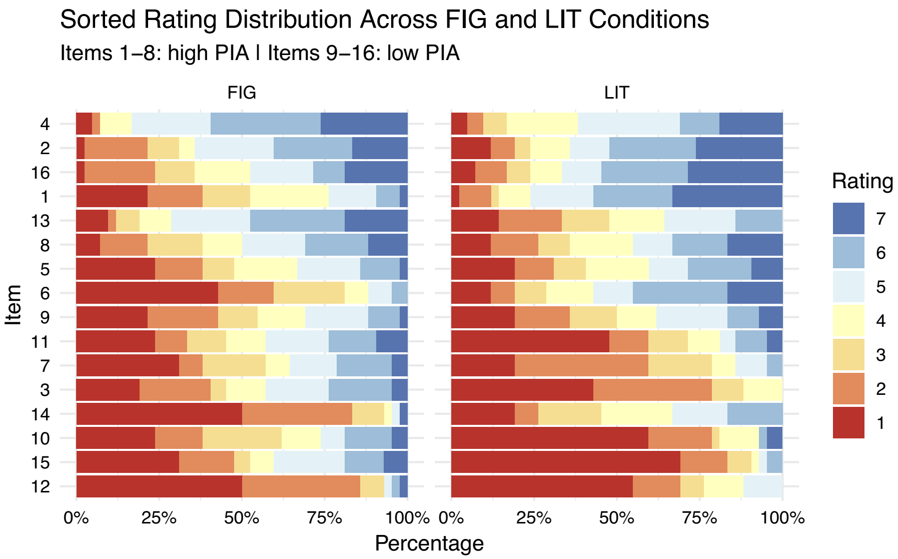

Figure 1. Sorted rating distributions for idiomatic items in the figurative (FIG) and literal (LIT) conditions. Items are arranged from bottom to top in ascending order based on increasingly higher mean ratings on the 1 to 7 Likert scale.

Overall, visual inspection of the distributions shows that, in both conditions, most high-PIA items (1–8) received higher ratings, as indicated by their placement near the top of the plot. Conversely, most low-PIA items (9–16) appear near the bottom, reflecting their generally lower ratings.

Some items, however, deviate from this pattern in both conditions. Items 3 (*gettare la maschera*) and 7 (*forzare la mano*) received mostly low ratings, despite being in the high PIA category. At the same time, items 13 (*toccare corde sensibili*) and 16 (*fare l'avvocato del diavolo*) received mostly high ratings, although they belong to the low PIA category.

A closer analysis also reveals items that behave differently across the two conditions, contradicting expectations in only one. In the LIT condition, item 14 (*fare acqua da tutte le parti*) shows a relatively balanced distribution of ratings across the scale levels. In the FIG condition, the situation is more complex: items 1 (*prendere per mano*) and 6 (*parlare al muro*) received lower ratings than expected. Lastly, item 11 (*morire dietro*) also stands out, having received a notable number of high ratings.

The observations made from the raw distributions are visually identifiable in Figures 2 and 3, which show mean ratings per idiomatic item by *PIA* (both continuous and categorical) for the FIG and LIT conditions, respectively.

The figures also clearly indicate a positive relationship between mean ratings of idiomatic items and the PIA index across both conditions, with a notably stronger relationship in the LIT condition than in the FIG condition. In this regard, Table 4 shows that, while ratings do not seem to diverge significantly by *Condition*, the switch from low to high PIA items leads to an overall increase in average ratings, with a greater increase observed for LIT than for FIG (*SD* in brackets).

In light of the descriptive statistics results, a clmm was fitted to further explore the data (see details in Section 2.4). Model results are reported in Table 5.

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

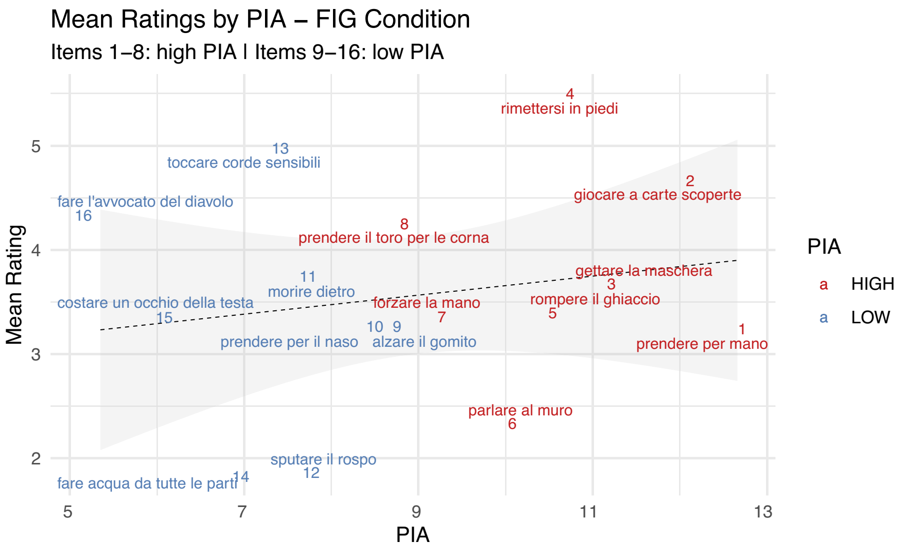

Figure 2. Mean item ratings by *PIA* in the figurative (FIG) condition, with *PIA* shown continuously (x-axis) and categorically (HIGH versus LOW by color). The regression line illustrates the positive relation between *PIA* and mean ratings.

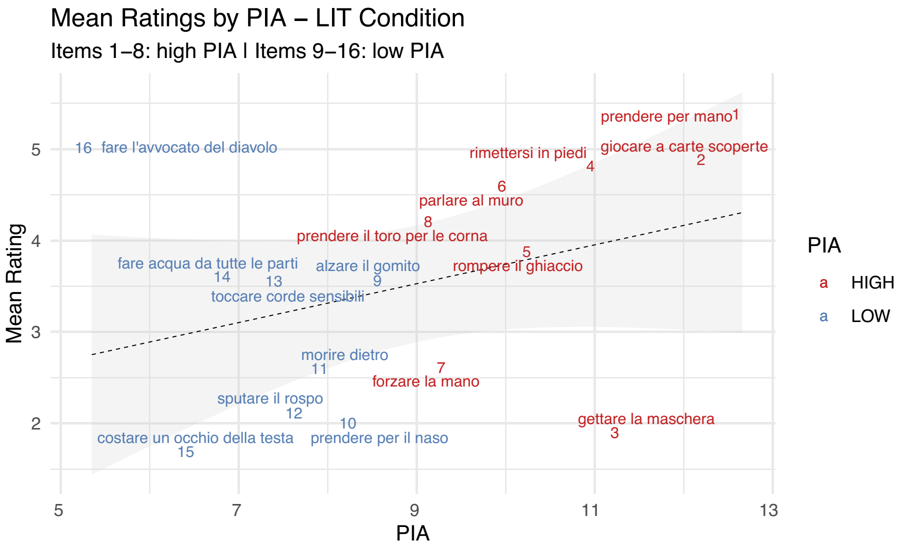

Figure 3. Mean item ratings by *PIA* in the literal (LIT) condition, with *PIA* shown continuously (x-axis) and categorically (HIGH and LOW by color). The regression line illustrates the positive relation between *PIA* and mean ratings.

The research interest of the present work is focused on the *PIA* variable, whose effect is estimated at $b = 0.552$ ($SE = 0.291$, $z = 1.897$, 95% $CI = [-0.018, 1.122]$). The p-value of 0.058 indicates that the effect of *PIA* approaches conventional thresholds for statistical significance ($p < 0.05$), pointing to a trend where high-PIA idioms may

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

Table 4. Mean ratings (*SD* in brackets) by condition and PIA category

		PIA	
		LOW	HIGH
Condition	FIG	3.33 (2.04)	3.8 (2)
	LIT	3.02 (2)	4.04 (2.07)

Table 5. Fixed effects results for the cumulative link mixed model (clmm) exploring the interaction between *Condition* and *PIA*

Fixed effect	*b*	*SE*	*z*	*CI* 2.5%	*CI* 97.5%	*p*
Condition (LIT)	−0.032	0.213	−0.149	−0.450	0.386	0.881
PIA (HIGH)	0.552	0.291	1.897	−0.018	1.122	0.058
Condition:PIA	0.458	0.416	1.101	−0.357	1.273	0.271

be associated with higher ratings across both FIG and LIT conditions. The 95% confidence interval narrowly includes zero, with a lower bound of −0.018. This suggests that the observed positive trend in the effect of *PIA* is likely real but requires further confirmation with larger sample sizes or additional data.

As predictable from descriptive data, *Condition* shows no significant effect ($b = -0.032$, $SE = 0.213$, $z = -0.149$, 95% $CI = [-0.160, 0.421]$, $p = 0.38$). The positive interaction between *Condition* and *PIA* ($b = 0.458$, $SE = 0.416$, $z = -1.101$, 95% $CI = [-0.357, 1.273]$) is also nonsignificant ($p = 0.271$).

The positive effect of *PIA* is finally illustrated in Figure 4 in which individual participant behavior is compared with the grand mean ratings by *PIA* and *Condition*. Although ratings increase in both conditions from low to high PIA idioms, the greater variability in participant responses within the FIG condition results in a less pronounced overall increase compared to the LIT condition.

3.2.1. Role of literal plausibility, decomposability and transparency within PIA

Given that *PIA* is a composite index comprising *Literal Plausibility*, *Decomposability* and *Transparency*, the relation between each individual variable and the experimental ratings can be explored. This was done by calculating the Spearman correlations reported in Table 6. In the FIG condition, only *Literal Plausibility* shows a significant, weak positive correlation with ratings ($\rho = 0.11$, $p = 0.004$). Neither *Decomposability* ($\rho = -0.03$, $p = 0.397$) nor *Transparency* ($\rho = 0.06$, $p = 0.142$) seems to be significantly correlated with the ratings.

In the LIT condition, *Literal Plausibility* again shows a significant weak positive correlation with ratings (= 0.25, $p < 0.001$). While *Decomposability* remains uncorrelated with ratings (= −0.02, $p = 0.676$), *Transparency* now shows a significant, weak positive correlation (= 0.24, $p < 0.001$).

3.3. Time data

The average experiment duration was 13.81 minutes ($SD = 7.08$, min = 6.53, max = 43.29, skewness = 2), aligning well with the timeframe communicated to the participants.

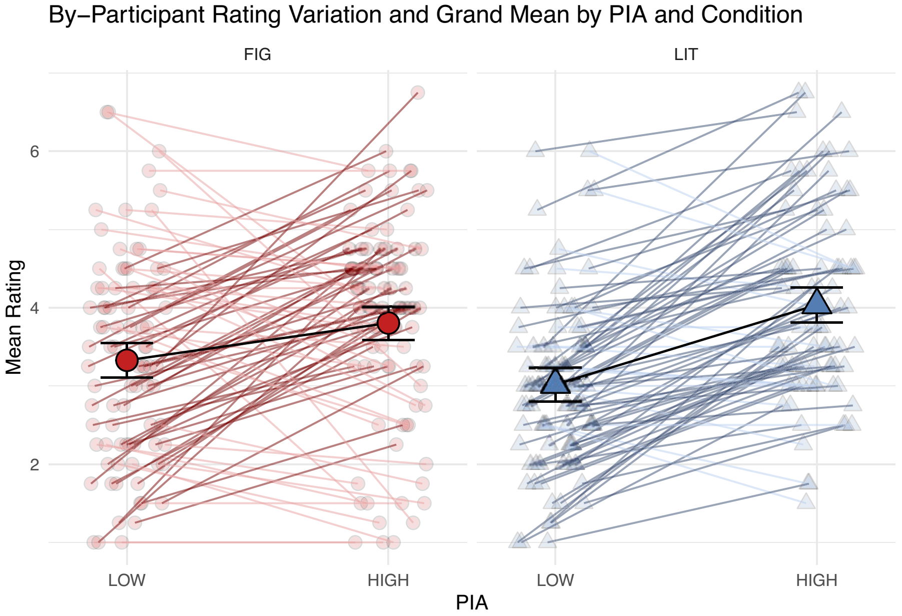

Figure 4. Overall grand mean ratings with 95% CIs by *PIA* in FIG and LIT conditions, shown alongside individual participant behaviors.

Table 6. Spearman correlations between single PIA variables and ratings in both FIG and LIT conditions

	Ratings - FIG	Ratings - LIT
Lit. Plausibility	0.11*	0.25*
Decomposability	−0.03	−0.02
Transparency	0.06	0.24*

Note: *indicates p-value <0.05.

Following outlier removal, the average reading times for each section of the experimental stimuli were calculated. The mean reading time for A contexts is 185.15 ms per word (*SD* = 87.79), while the mean reading and evaluation time for B contexts in relation to A is 436.33 ms per word (*SD* = 190.62). The overall mean time per item is 644.93 ms per word (*SD* = 242.2).

Let us now turn our attention to the idiomatic items. Table 7 shows the average reading times for each stimulus section in relation to the two experimental conditions and the two *PIA* groups in each.

Reading times across the conditions show minimal differences, with only slightly higher values for the LIT condition. However, a pattern appears in section B when comparing high and low-PIA items: combined reading and rating times are consistently longer for low-PIA items than for high-PIA items. Specifically, in the FIG condition, combined reading and rating times are 393.92 ms/word for high-PIA and 433.74 ms/word for low-PIA items, while in the LIT condition, they are 414.31 ms/word for high-PIA and 438.24 ms/word for low-PIA items. This trend results in higher overall mean times per stimulus for low-PIA idiomatic items (FIG = 640.02 ms/word,

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

Table 7. Mean reading times (ms/word) by *Condition* and *PIA* for each section of the idiomatic items (*SD* in brackets)

	FIG		LIT	
Section	High PIA	Low PIA	High PIA	Low PIA
A	184.20 (90.93)	174.06 (80.90)	180.92 (87.39)	180.12 (90.12)
B	393.92 (158.56)	433.74 (170.57)	414.31 (160.75)	438.24 (172.46)
Stimulus	601.93 (211.72)	640.02 (231.23)	620.14 (207.01)	636.74 (219.25)

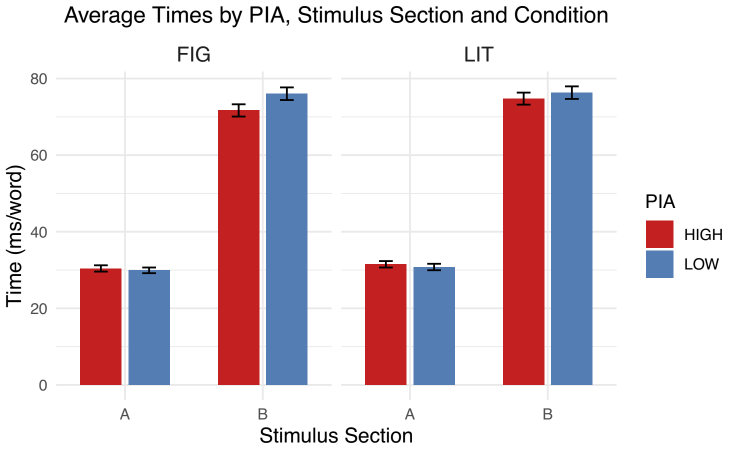

Figure 5. Bar plots with error bars (Standard Error) showing average reading times (ms/word) for the FIG and LIT conditions, grouped by *PIA* across stimulus sections A and B (note: B includes both reading and evaluation times).

LIT = 636.74 ms/word) compared to high-PIA idiomatic items (FIG = 601.93 ms/word, LIT = 620.14 ms/word). Figure 5 provides a visual representation of the described patterns, displaying average times by stimulus section (part A and B), *PIA* and *Condition*.

Based on the patterns identified via descriptive statistics, combined reading and rating times for contexts B were analyzed using a lmm (see Section 2.4). Table 8 presents the model output for the fixed effects.

Descriptively, Table 7 presents slightly lower times in the FIG condition compared to the LIT condition. However, Table 8 shows that this difference is not statistically reliable, and no significant effect of *Condition* was found ($b = -13.610$, $SE = 11.937$, $df = 19.703$, $t = -1.140$, 95% CI [−37.520, 10.405], $p = 0.268$).

Table 8. Fixed effects results of the linear mixed model exploring the impact of *Condition* and *PIA* on context B combined reading and rating times

Fixed effect	*b*	*SE*	*df*	*t*	*CI* 2.5%	*CI* 97.5%	*p*
Intercept	418.62	13.709	42.386	30.536	391.584	445.505	< 0.001
Condition (FIG)	−13.610	11.937	19.703	−1.140	−37.520	10.405	0.268
PIA (LOW)	31.274	12.650	16.209	2.472	5.465	56.999	0.025

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

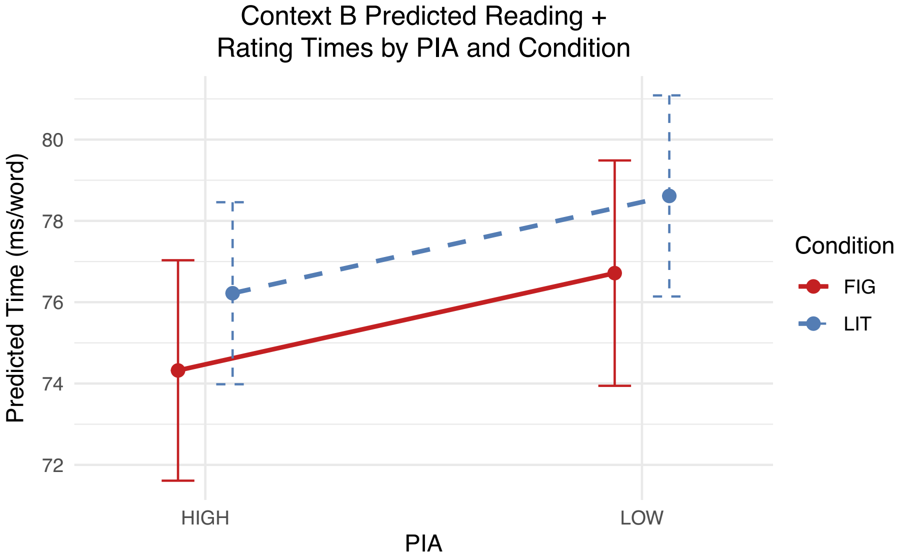

Figure 6. Times (ms/word) by *PIA* and *Condition* predicted by the linear mixed-effects model for reading and assessing context B of the experimental stimuli. Error bars indicate Standard Errors (*SE*).

In contrast, the effect of *PIA* (LOW) is positive and significant, with an estimate of 31.274 (*SE* = 12.650, *df* = 16.209, *t* = 2.472, 95% *CI* [5.465, 56.999], *p* = 0.025), suggesting that combined reading and rating times are significantly higher for low-PIA items compared to high-PIA items. Overall, the model output highlights that while times do not change significantly across the two experimental conditions, they vary significantly by *PIA* level, with stimuli including low-PIA idioms associated with slower times than stimuli including high-PIA idioms. Model predictions are shown in Figure 6.

4. Discussion

The results for filler items indicate that crowdsourced participants understood the task and responded as expected to the properties of the stimuli. For syntactic ambiguity, the dispreferred garden-path sentences received lower ratings compared to preferred and ambiguous conditions, aligning with the reanalysis model of syntactic ambiguity (van Gompel et al., 2005). For the modal alternation, the indicative in contexts prescribing the subjunctive was rated lower, though its relatively high scores likely reflect its growing use in neo-standard Italian (Ballarè, 2021; Digesto, 2022).

With regard to the critical items with idiomatic MWEs, results show that acceptability ratings for literal completions increase in both the entirely literal and figurative–creative conditions as we move from low- to high-PIA MWEs. Literal completions are thus better integrated following high-PIA idioms compared to low-PIA idioms. While this positive trend is observable in both conditions, it is more pronounced in the entirely literal condition than in the figurative–creative one. On top of this, time analysis revealed significantly longer durations in section B of the

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

dialogue – where participants assigned ratings – for stimuli containing low-PIA idioms across both conditions. Time results must, of course, be contextualized within the limitations of this offline experimental setup, particularly the lack of strict control over variables like word length and frequency. However, stimuli were designed to be as homogeneous as possible across items and conditions, and times were normalized by word count for each stimulus section. As such, the combined findings of lower ratings and longer combined reading and rating times for low-PIA idioms suggest higher integration costs for literal completions across both biasing contexts.

These results support the hybrid model of idiom representation and processing, which highlights the dual nature of idioms as both formulaic and compositional, depending on the context and the idiom's properties (Libben & Titone, 2008; Senaldi & Titone, 2024; Sprenger et al., 2006). In this study, idioms with high ambiguity potential facilitate the integration of post-idiom literal completions, both when this is to be interpreted entirely compositionally and when it is to be figuratively reinterpreted in a creative manner. In line with Beck and Weber (2020), this occurs because such idioms favor binary processing, where holistic and compositional analyses coexist. In contrast, low-ambiguity idioms are biased toward a unitary, figurative interpretation.

Building again on the findings of Beck and Weber (2020), integration is most efficient when context-driven expectations align with those arising from the idiom's internal characteristics. In this study, this alignment occurs in the literal condition with high-PIA idioms, where a context favoring literal interpretation complements the dual nature of these idioms. Table 4 confirms this, showing that the LIT/high-PIA condition has the highest mean rating (4.04/7) among the four scenarios.

While the LIT/high-PIA condition shows the highest mean rating, it is worth noting that this is not an exceptionally high score, and two interrelated factors may account for this: the statistical idiosyncrasy of MWEs and the meaning dominance or semantic salience specific to each idiom. As noted in the introduction of the current contribution, MWEs are notable for the high statistical attraction between their constituent parts (Farahmand & Nivre, 2015). The co-occurrence of those precise lexemes makes MWEs specialized in conveying specific meanings, typically non-compositional ones. The semantic specialization of an idiom toward a specific meaning is referred to as its semantic salience or meaning dominance (Giora, 1997, 2003; Milburn & Warren, 2019). This concept captures the relative prominence of one meaning of an MWE compared to its alternative interpretation. When there is a significant imbalance in the frequency with which the meanings are used or recognized, the MWE becomes more salient for one meaning, often the figurative one. In addition, semantic dominance is further accentuated when an idiom is very familiar (Laurent et al., 2006).

In this regard, the findings of Milburn and Warren (2019) offer a particularly relevant insight. Their eye-tracking study revealed an interaction effect whereby, under conditions of high figurative dominance, increased semantic relatedness between the figurative and literal meanings of an idiom slowed processing, both when the preceding context was biased toward a literal interpretation and when it was biased toward a figurative one. This effect suggests that strong entrenchment of the figurative meaning may hinder the accessibility of the alternative literal meaning, including when this is plausible and the two meanings are closely related. Applied to the present data, a high-PIA idiom like *gettare la maschera*, which likely exhibits strong figurative dominance, may exemplify this phenomenon. In the literal

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

condition, it stands out as an outlier, having failed to receive any ratings above 4 despite its high literal plausibility, decomposability and transparency (see idiom No. 3 in Figure 1). This result could plausibly be attributed to interference from its dominant figurative interpretation, which hinders literal access even in the presence of a context that strongly biases toward a literal reading.

In summary, the LIT/high-PIA scenario likely unfolded as follows: (a) reading of context A biased toward a literal interpretation; (b) initial encounter in context B with a highly familiar MWE and with a potential figurative semantic dominance; (c) continued literal bias in context B; (d) the idiom's inherent high ambiguity enabled participants to access its literal meaning and integrate it into the context. The cognitive cost of this integration likely varied depending on the idiom's degree of semantic dominance.

A similar situation occurred in the LIT/low-PIA condition, with the added challenge that these idioms were not inherently ambiguous, being distinctly biased toward a figurative interpretation. This internal bias heightened the cognitive cost of integration in step (d). As a result, this scenario yielded the lowest ratings among all four conditions (3.02, Table 4), despite the congruence of the context, which was literal in both parts A and B. This finding highlights the critical role of idiomatic features in determining the ease of contextual integration (Beck & Weber, 2020).

Noting that the extreme values for the idiom PIA groups occur in the literal condition implies that the shift from low-PIA to high-PIA idioms is more pronounced in this condition compared to the figurative–creative one (as illustrated in Figures 2 and 3). This outcome does not come as a surprise, given the distinctive nature of the dialogues in the figurative condition, where the two contexts (A and B) exhibit incongruent biases – figurative in A and literal in B. Nonetheless, the positive impact on ratings in the transition from idioms with low to high ambiguity potential is observable. We interpret this as an indication of an emergent systematicity even in the case of playful, creative idiomatic instances at the semantic–pragmatic level. In other words, the inherent features of the idioms influenced the ease with which participants could reinterpret the literal completions in light of the pre-established figurative bias. Considering that this is, to the best of our knowledge, the first experimental investigation of idiomatic parasitic elaboration, the outcome of the present study can be regarded as an encouraging step forward. Naturally, it must be contextualized in light of the relatively small item set. Still, as noted in Section 2.1, the study is partly exploratory in nature, and a central aim of exploratory research is precisely to identify promising patterns that merit further investigation.

Looking at Figure 1, a good example in this respect is idiom No. 4, *rimettersi in piedi* ('get back on one's feet'), whose dialogue is here reported in English: 'A: You know, after the bankruptcy I had a real hard time. But now I think I'm ready to start a new business. B: Well, if you want to get back on your feet, first make sure your legs are stable'. Reading this dialogue, participants were arguably able to reinterpret the stability of the legs as a metaphorical representation of the security and preparedness that the imaginary speaker A requires to successfully launch a new business following a financial crisis.

A speculative observation can be made regarding the different average ratings in the literal and figurative conditions. It is important to note that statistical analysis reported no significant difference between the two conditions. Consequently, the following observations are specific to this experiment and should not be generalized as of now. Nevertheless, given the partly exploratory nature of this study, we believe it is important to highlight preliminary insights that future research may confirm or refute with greater precision.

Table 4 shows that the average rating for the FIG/low-PIA scenario is higher than for LIT/low-PIA and is accompanied by a shorter average combined reading and rating time (see Table 7). This may suggest that, in the figurative condition, the contextual incongruence within the dialogues including low-PIA idioms may act as a quasi-advantage compared to the contextual congruence observed in the literal condition with the same idioms. That is to say, in the LIT/low-PIA condition, the intrinsic features of the idiom may hinder smooth integration into a context that predominantly favors a literal interpretation. In contrast, in the FIG/low-PIA condition, at least half of the context aligns with the idiom's intrinsic tendency to prompt direct memory retrieval of its figurative meaning. Paradoxically, this contextual inconsistency might have facilitated participants' evaluations. However, as we move from low-PIA to high-PIA idioms, the contextual incongruence in the FIG condition could no longer offer a relative advantage. In the LIT/high-PIA condition, the congruent context aligns with the expectations likely generated by the inherent characteristics of high-PIA idioms. Consequently, the pattern of average ratings reverses for high-PIA idioms, with the literal condition receiving higher ratings than the figurative one.

To verify the impact of the individual variables within the PIA index on the ratings, correlations were tested between literal plausibility, decomposability, transparency and ratings in both literal and figurative conditions. In the literal condition, ratings showed a significant positive correlation with literal plausibility, thus corroborating again the findings in Beck and Weber (2020). Interestingly, transparency also correlated positively with ratings, aligning with van Ginkel and Dijkstra (2019), who found that higher transparency between literal and figurative meanings supports compositional retrieval. Decomposability, however, showed no correlation with ratings.

In the figurative–creative condition, only literal plausibility shows a significant positive correlation with the ratings. This finding positions literal plausibility as the key variable across both context types, confirming its prominent role in studies on idiomatic MWEs (Beck & Weber, 2020; Mancuso et al., 2019; Titone & Libben, 2014). By contrast, neither decomposability nor transparency shows significant associations with the ratings in the figurative–creative condition. Before drawing conclusions on the variables themselves, a methodological remark should be made. The statistical analysis of the dataset from which the idioms were drawn showed that literal plausibility has a bimodal distribution, in contrast to decomposability and transparency, which instead have a uniform and a more normal distribution, respectively (Pagliai, 2024). Selecting only 16 idioms, along with the distributional differences among the variables, led to more extreme values for literal plausibility (M = 3.17, range = 1.17–5.00) compared to decomposability (M = 2.93, range = 1.90–3.76) and transparency (M = 2.90, range = 2.07–3.90). This likely reduces the potential statistical impact of decomposability and transparency, making them less prominent compared to the heightened role of literal plausibility. These observations further underscore the importance of expanding the idiom set in future experimental research on idiomatic wordplay, both to balance variable distributions and to build on the patterns identified in this exploratory study.

Concerning decomposability not being significant in either condition, we can make a further observation regarding the high familiarity and meaningfulness of all the idioms included in the present study. The experiments conducted by Libben and Titone (2008) revealed that the effects of decomposability on processing are attenuated when idioms are highly familiar. Therefore, one way to further investigate whether decomposability truly plays no role would be to expand the experiment to include idioms with varying degrees of familiarity and meaningfulness. By incorporating less familiar and less meaningful idioms, it would be possible to test whether decomposability exerts a more pronounced effect under such conditions, as suggested by prior research.

Drawing from the discussion thus far, improvements for future research can be outlined. First, an important variable may be missing from the current framework: the previously mentioned semantic salience or meaning dominance. This variable captures a critical dimension of idiom semantics and could potentially offer greater explanatory power than decomposability or transparency. Including it in future studies would complement the PIA variables and provide deeper insights into the mechanisms underlying the integration of idiom literal completions in biasing contexts. Second, expanding the idiomatic experimental items is essential to achieving more generalizable results (see Section 2.1). While this poses challenges, given the difficulty of crafting sufficiently uniform yet adequately extended and naturalistic experimental contexts (Wagner, 2021), a step in this direction has been taken with the dialogue items developed in this study, which can serve as a foundation for refinement and future research.

5. Conclusions

The rating experiment reported here examined how idiom literal completions are integrated into two distinct biasing contexts: one promoting a fully compositional interpretation and the other a figurative one, which requires creative reinterpretation of the literal completion to align with the figurative bias. Specifically, the study examined whether the ease of integrating literal completions is modulated by the inherent characteristics of idioms, categorized into two groups based on their Potential Idiomatic Ambiguity (*PIA*) scores, calculated as the average of literal plausibility, decomposability and transparency for each idiom. Building on prior research, high-PIA idioms involve competition between figurative and literal meanings, while low-PIA idioms are biased toward figurative interpretation. Therefore, we hypothesized that literal completions of high-PIA idioms would integrate more effectively, especially in the literal–compositional condition. For the figurative–creative condition, the study adopted an exploratory approach to assess whether and to what extent idiom PIA would exert an influence.

Overall, results align well with the expectations outlined in Section 1.3, showing that in both experimental conditions, high-PIA idioms received higher acceptability ratings for literal completions compared to low-PIA idioms. This provides further experimental evidence supporting hybrid models of idiom representation and processing, which emphasize the critical role of context and idiom features in adaptively balancing compositional and holistic comprehension strategies.

Even if the positive trend on ratings is less pronounced in the figurative–creative condition – reflecting its greater difficulty of comprehension — we find the results for this condition encouraging. To the best of our knowledge, this is the first study providing preliminary evidence of systematicity within instances of idiomatic

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

wordplay, particularly in literal-scene manipulation (or parasitic elaboration, Langlotz, 2006). This work, therefore, aligns with the perspectives in Fellbaum (2015, 2019) and Bargmann et al. (2021), advocating for idiomatic wordplay to be viewed as more than a linguistic epiphenomenon. Indeed, linguistic history has already shown that once peripheral phenomena, like MWEs, can in fact move to the forefront of (psycho)linguistic inquiry.

In pursuing this goal, it is crucial to recognize how challenging it is to identify patterns of systematicity within creative uses of idiomatic language, given all the sources of diversity and variability involved. First, idioms are varied and distinct, a characteristic that underpins the present work. Second, Gibbs (1995) already observed that speakers do not consistently analyze or interpret idiom semantics in the same way, leading to inter-individual variation in the mappings between literal and figurative meanings. Moreover, idiomatic wordplay, as a form of creative language use, inherently fosters innovation and diversity (Langlotz, 2006). Adding to this whole diversity are the individual cognitive characteristics of speakers, which recent research has shown to influence how idioms are produced and understood (Arnon & Lavidor, 2022; Geeraert et al., 2020; Tilmatine et al., 2021).

Given this multilayered diversity, it also becomes evident that generalizations cannot rely solely on corpus-based qualitative analyses, as idiomatic wordplay is infrequent (Langlotz, 2006), often based on a limited number of idioms and shaped by the subjective interpretations and cognitive schemas of individual speakers. Nonetheless, theoretical and corpus-based qualitative studies provide a crucial foundation from which new experiments can be designed to generate sufficient data for progressively uncovering systematicity. In other words, accepting the challenge of identifying systematicity in creative idiomatic uses requires researchers to adopt approaches that are both methodical and innovative. We believe this research marks a step in that direction and invites (psycho)linguists to further embrace this challenge by designing innovative experiments to systematically explore the ever-fascinating creativity of language.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/langcog.2025.10013.

Data availability statement. The experiment, its script and all experimental items can be accessed via this demonstration link. The experiment results, as well as the R script for statistical analysis replication, are available in the publicly accessible repository of the University of Göttingen (Pagliai, 2025).

Funding statement. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK2636.

Competing interests. The authors declare none.

References

Arnon, T., & Lavidor, M. (2022). Cognitive control in processing ambiguous idioms: Evidence from a self-paced reading study. *Journal of Psycholinguistic Research*, 52(1), 261–281. https://doi.org/10.1007/s10936-022-09861-z.

Ballarè, S. (2021). L'italiano neo-standard oggi: Stato dell'arte. *Italiano LinguaDue*, V. 12 N. 2 (2020). https://riviste.unimi.it/index.php/promoitals/article/view/15013; https://doi.org/10.13130/2037-3597/15013

Bargmann, S., Gehrke, B., & Richter, F. (2021). Modification of literal meanings in semantically non-decomposable idioms. In B. Crysmann & M. Sailer (Eds.), *One-to-many relations in morphology, syntax, and semantics* (pp. 245–279). Language Science Press. https://doi.org/10.5281/zenodo.4729808.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Beck, S. D., & Weber, A. (2020). Context and literality in idiom processing: Evidence from self-paced reading. *Journal of Psycholinguistic Research*, 49(5), 837–863. https://doi.org/10.1007/s10936-020-09719-2.

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language*, 27(6), 668–683. https://doi.org/10.1016/0749-596x(88)90014-9.

Cacciari, C., & Tabossi, P. (2014). *Idioms: Processing, structure, and interpretation*. Psychology Press. https://doi.org/10.4324/9781315807133.

Carrol, G., & Segaert, K. (2024). As easy as cake or a piece of pie? Processing idiom variation and the contribution of individual cognitive differences. *Memory & Cognition*, 52(2), 334–351. https://doi.org/10.3758/s13421-023-01463-x.

Carter, R. (2015). *Language and creativity: The art of common talk* (1st ed.). Routledge. https://doi.org/10.4324/9781315658971.

Carter, R., & McCarthy, M. (2004, 03). Talking, creating: Interactional language, creativity, and context. *Applied Linguistics*, 25(1), 62–88. https://doi.org/10.1093/applin/25.1.62

Christensen, R. H. B. (2023). Ordinal—Regression models for ordinal data [computer software manual]. https://CRAN.R-project.org/package=ordinal (R package version 2023.12–4.1)

Citron, F. M., Cacciari, C., Kucharski, M., Beck, L., Conrad, M., & Jacobs, A. M. (2016). When emotions are expressed figuratively: Psycholinguistic and affective norms of 619 idioms for german (panig). *Behavior Research Methods*, 48, 91–111. https://doi.org/10.3758/s13428-015-0581-4.

Digesto, S. (2022). Lexicalization and social meaning of the italian subjunctive. *Cadernos de Linguística*, 2(3), e609. https://doi.org/10.25189/2675-4916.2021.v2.n3.id609.

Ernst, T. (1981). Grist for the linguistic mill: Idioms and 'extra'adjectives. *Journal of Linguistic Research*, 1(3), 51–68.

Fadlon, J., Horvath, J., Siloni, T., & Wexler, K. (2013). The acquisition of idioms: Stages and theoretical implications. In *Poster presentation, generative approaches to language acquisition*. The University of Oldenburg.

Farahmand, M., & Nivre, J. (2015). Modeling the statistical idiosyncrasy of multiword expressions. In *Proceedings of the 11th workshop on multiword expressions*. Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-0905

Fellbaum, C. (2015). Is there a grammar of idioms. In *8th Brussels conference on generative linguistics, Brussels* (pp. 4–5).

Fellbaum, C. (2019). How flexible are idioms? A corpus-based study. *Linguistics*, 57(4), 735–767.

Geeraert, K., Newman, J., & Baayen, R. H. (2020). Variation within idiomatic variation: Exploring the differences between speakers and idioms. *East European Journal of Psycholinguistics*, 7(2). https://doi.org/10.29038/eejpl.2020.7.2.gee.

Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition*, 8(2), 149–156. https://doi.org/10.3758/bf03213418.

Gibbs, R. W. (1995). Idiomaticity and human cognition. In M. Everaert, E.-J. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 97–116). Lawrence Erlbaum Associates, Inc.

Gibbs, R. W., & Nayak, N. P. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology*, 21(1), 100–138.

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *cogl*, 8(3), 183–206. https://doi.org/10.1515/cogl.1997.8.3.183.

Giora, R. (2003). *On our mindsalience, context, and figurative language*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195136166.001.0001.

Hillert, D., & Swinney, D. (2001). The processing of fixed expressions during sentence comprehension. In *Conceptual and discourse factors in linguistic structure*, 107–122.

Hindle, D., & Rooth, M. (1991). Structural ambiguity and lexical relations. In *Proceedings of the 29th annual meeting on association for computational linguistics* (pp. 229–236). Association for Computational Linguistics. https://doi.org/10.3115/981344.981374

https://doi.org/10.1017/langcog.2025.10013 Published online by Cambridge University Press

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Kyriacou, M., Conklin, K., & Thompson, D. (2019). Passivizability of idioms: Has the wrong tree been barked up? *Language and Speech*, 63(2), 404–435. https://doi.org/10.1177/0023830919847691.

Langlotz, A. (2006). *Idiomatic creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English*. John Benjamins Publishing Company. https://doi.org/10.1075/hcp.17.

Laurent, J.-P., Denhières, G., Passerieux, C., Iakimova, G., & Hardy-Baylé, M.-C. (2006). On understanding idiomatic language: The salience hypothesis assessed by erps. *Brain Research*, 1068(1), 151–160. https://doi.org/10.1016/j.brainres.2005.10.076.

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

Maher, Z. (2013). *Opening a can of worms: Idiom flexibility, decomposability, and the mental lexicon*. Yale University MA thesis.

Mancuso, A., Elia, A., Laudanna, A., & Vietri, S. (2019). The role of syntactic variability and literal interpretation plausibility in idiom comprehension. *Journal of Psycholinguistic Research*, 49(1), 99–124. https://doi.org/10.1007/s10936-019-09673-8.

McGlone, M. S., Glucksberg, S., & Cacciari, C. (1994). Semantic productivity and idiom comprehension. *Discourse Processes*, 17(2), 167–190. https://doi.org/10.1080/01638539409544865.

Mel'čuk, I. (2014). Phrasemes in language and phraseology in linguistics. In *Idioms* (pp. 167–232). Psychology Press.

Milburn, E., & Warren, T. (2019). Idioms show effects of meaning relatedness and dominance similar to those seen for ambiguouswords. *Psychonomic Bulletin & Review*, 26(2), 591–598. https://doi.org/10.3758/s13423-019-01589-7.

Monti, J., & di Buono, M. P. (2019). PARSEME-it: An italian corpus annotated with verbal multiword expressions. *IJCoL*, 5(2), 61–93. https://doi.org/10.4000/ijcol.483.

Moreno, R. E. V. (2005). *Idioms, transparency and pragmatic inference* (tech. Rep.). *UCL Working Papers in Linguistics*, 17, 389–425.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, 70(3), 491–538. https://doi.org/10.1353/lan.1994.0007.

Pagliai, I. (2023a). Bridging the gap: Creation of a lexicon of 150 pairs of english and italian idioms including normed variables for the exploration of idiomatic ambiguity. *Journal of Open Humanities Data*. https://doi.org/10.5334/johd.123.

Pagliai, I. (2023b). Normed lexicon of English and Italian idioms. https://doi.org/10.25625/EPSWDY.

Pagliai, I. (2024). Some cakes have icing, others have a cherry: Does it make a difference? A cross-linguistic norming study on 150 pairs of english and italian idioms. *Submitted to the John Benjamins book series Figurative Thought and Language.*

Pagliai, I. (2025). *Italian idioms in literal and creative contexts: Experimental results on the contextual integration of idiom literal completions*. GRO.data. https://doi.org/10.25625/ODOEST.

R Core Team. (2023). *R: A language and environment for statistical computing [computer software manual]*. Vienna, Austria.

Sailer, M. (2021). *Idioms*. Language Science Press. https://doi.org/10.5281/zenodo.5599850.

Senaldi, M. S. G., & Titone, D. (2024). Idiom meaning selection following a prior context: Eye movement evidence of l1 direct retrieval and l2 compositional assembly. *Discourse Processes*, 61(1–2), 21–43. https://doi.org/10.1080/0163853x.2024.2311637.

Senaldi, M. S. G., Wei, J., Gullifer, J. W., & Titone, D. (2022). Scratching your tête over language-switched idioms: Evidence from eye-movement measures of reading. *Memory & Cognition*, 50(6), 1230–1256. https://doi.org/10.3758/s13421-022-01334-x.

Slattery, T. J., Sturt, P., Christianson, K., Yoshida, M., & Ferreira, F. (2013). Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language*, 69(2), 104–120. https://doi.org/10.1016/j.jml.2013.04.001.

Sprenger, S., Levelt, W., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, 54(2), 161–184. https://doi.org/10.1016/j.jml.2005.11.001.

Tilmatine, M., Hubers, F., & Hintz, F. (2021). Exploring individual differences in recognizing idiomatic expressions in context. *Journal of Cognition*, 4(1), 37. https://doi.org/10.5334/joc.183.

Titone, D., & Connine, C. (1999). On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics*, 31(12), 1655–1674. https://doi.org/10.1016/s0378-2166(99)00008-9

Titone, D., & Libben, M. (2014). Time-dependent effects of decomposability, familiarity and literal plausibility on idiom meaning activation: A cross-modal priming investigation. *The Mental Lexicon*, 9(3), 473–496. https://doi.org/10.1075/ml.9.3.05tit.

Titone, D., Lovseth, K., Kasparian, K., & Tiv, M. (2019). Are figurative interpretations of idioms directly retrieved, compositionally built, or both? Evidence from eye movement measures of reading. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 73(4), 216–230. https://doi.org/10.1037/cep0000175.

van Ginkel, W., & Dijkstra, T. (2019). The tug of war between an idiom's figurative and literal meanings: Evidence fromnative and bilingual speakers. *Bilingualism: Language and Cognition*, 23(1), 131–147. https://doi.org/10.1017/s1366728918001219.

van Gompel, R. P., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, 52(2), 284–307. https://doi.org/10.1016/j.jml.2004.11.003.

Wagner, W. (2021). *Idioms and ambiguity in context*. De Gruyter. https://doi.org/10.1515/9783110685459.

Zehr, J., & Schwarz, F. (2018). *Penncontroller for internet based experiments (IBEX)*. OSF Preprints. https://doi.org/10.17605/OSF.IO/MD832

Zucchini, E. (2023). *L'italiano neostandard nella lingua a scuola: Il Caso dell'alternanza fra indicativo e congiuntivo*. (Doctoral dissertation, Alma Mater Studiorum Bologna). https://doi.org/10.48676/unibo/amsdottorato/10505

Cite this article: Pagliai, I. (2025). 'Well, if you want to play with fire, make sure the flames are tamed': investigating the integration of idiom literal completions across literal and creative contexts, *Language and Cognition*, 17, e55, 1–25. https://doi.org/10.1017/langcog.2025.10013