Resolving figurative expressions during reading

Carrol, Gareth; Littlemore, Jeannette

DOI:
10.1080/0163853X.2020.1729041

Document Version
Peer reviewed version

Citation for published version (Harvard):
Carrol, G & Littlemore, J 2020, 'Resolving figurative expressions during reading: the role of familiarity, transparency and context', *Discourse Processes*, vol. 57, no. 7, pp. 609-626. https://doi.org/10.1080/0163853X.2020.1729041

Abstract

Native speakers understand familiar idioms (e.g. *over the moon*) and conventional metaphors (e.g. describing time as a doctor) quickly and easily. In two eye-tracking studies we considered how native speakers are able to make sense of fundamentally *unfamiliar* figurative expressions. In Experiment 1, compared to literal paraphrases of the same meaning, known idioms had a clear advantage, unknown idioms showed a significant disadvantage, and conventional metaphors showed no difference between figurative and literal versions. In Experiment 2, readers saw known and unknown idioms (or paraphrases) in contexts that either supported the intended meaning, or were neutral. Strength of context had minimal effect on reading patterns for either idiom type, and had no effect when readers were asked to subsequently identify the meaning. Context may be helpful in terms of sense selection but not when new senses need to be generated, when aspects such as transparency become more important.

Keywords: Idioms, metaphor, figurative language, context, eye-tracking

Introduction

Understanding figurative meaning requires the language user to employ a complex set of skills, such as semantic analysis (consideration of properties such as transparency or decomposability) and making inferences based on the broader context of use (Levorato & Cacciari, 1995, 1999). However, a sometimes overlooked aspect is simply the knowledge of conventionalized expressions that speakers can call on when required, and which builds up during a lifetime of linguistic and cultural experience (Sprenger, la Roi & van Rij, 2019). The importance of prior knowledge to how figurative expressions are understood is reflected in the range of experimental studies that have investigated this. Idioms are often processed more quickly than matched novel phrases (Carrol and Conklin, 2014, 2017; Carrol, Conklin and Gyllstad, 2016; McGlone, Glucksberg, and Cacciari, 1994; Siyanova-Chanturia, Conklin and Schmitt, 2011; Swinney and Cutler, 1979), and this is primarily a property of their being 'known' rather than their figurativeness per se (Carrol and Conklin, 2019; Tabossi, Fanari and Wolf, 2009). Whilst factors such as transparency and decomposability are important in classifying idioms (e.g. Cacciari and Glucksberg, 1991; Gibbs, 1986; Gibbs, Nayak and Cutting, 1989) and sometimes show effects in terms of processing (Caillies and Butcher, 2007; Titone and Libben, 2014), familiarity seems to be the overarching predictor of how idioms are processed and understood (Libben and Titone, 2008; Schweigert, 1986, 1991).

These phrase internal effects have been shown for other forms of figurative language as well, where conventionality as well as individual familiarity is found to be an important factor. Blasko and Connine (1993) and Blasko and Briihl (1997) demonstrated a clear effect of familiarity in how metaphors are understood (see also Bambini, Canal, Resta & Grimaldi, M., 2019; Columbus et al., 2015; Mashal and Faust, 2009). Bowdle and Gentner (2005) presented their Career of Metaphor hypothesis, whereby only novel metaphors are understood through a process of comparison and analysis, but for more established, conventionalised metaphors, a

(less effortful) process of categorisation is employed. (See also Cardillo et al., 2012, for complementary neuroimaging evidence whereby activation of brain areas varies as a function of how conventional a metaphor is). Alongside metaphors, Frisson and Pickering (1999, 2007) found similar effects whereby conventional metonymies were processed more quickly than less conventional uses. Giora (2003) argued that conventional ironies are easier to process than non-conventional uses, in line with the broader claims of her Graded Salience Hypothesis.

As in language processing more generally, context is also important in how figurative language is understood. Here, we consider context to be largely independent of the semantic aspects discussed above, although there may be situations where these interact (e.g. a strong context may make an unknown phrase easier to understand, which may make it seem more transparent). A supportive context helps to distinguish figurative and literal readings of idioms (Beck & Weber, 2018; Cieślicka, 2011; Colombo, 1993; Holsinger, 2013; Titone & Connine, 1999), and facilitates their comprehension more generally (Fanari, Cacciari & Tabossi, 2010; Schweigert & Moates, 1988). A relevant context can also be helpful for other types of figurative phrase. Inhoff, Lima and Carroll (1984) found that literal interpretations of noun phrases were read more quickly than metaphorical uses when they appeared in single sentences, but when a prior context was provided, differences between literal and metaphorical uses disappeared (see also Gildea & Glucksberg, 1983; Bambini, Bertini, Schaeken, Stella & Di Russo, 2016). Frisson and Pickering (2007) found that a prior sentence introducing an otherwise unfamiliar metonymy (e.g. establishing 'Needham' as an author, prior to using this as a producer-for-product metonymy, e.g. 'read a lot of Needham') was enough to remove any disruption to reading times experienced when the same referent was used without a preceding context. However, a relatively under-explored question is what role context plays when language users are faced with the task of interpreting less familiar

phrases. Guessing from context is often identified as a key strategy for non-native speakers in understanding idioms (e.g. Liontas, 2002), although Boers, Eyckmans and Stengers (2007) suggested that learners may be significantly less successful here than is sometimes assumed. Zuo (2008) and Xie (2017) both found that use of context was a major factor in how Chinese learners attempted to identify previously unknown English idioms. In support of this, Wray, Bell and Jones (2016) found that both native speakers and learners used context as the main strategy when asked to guess the meaning of unknown phrases.

In attempting to establish what role context plays in the comprehension of different types of phrases, one useful source of comparison is the literature on how language users identify the meaning of individual words they have not heard before. Whilst not exactly the same challenge (an unknown idiom is likely to include at least some words that are already known, to act as a starting point), encountering a previously unseen word may require some of the same skills in terms of utilising the surrounding context to infer a likely meaning. In this situation, readers have been found to spend longer on unknown words, and also spend more time reading the following context, especially when this is informative (Chaffin, Morris & Seely, 2001; Joseph et al., 2015; Lowell & Morris, 2014; Williams & Morris, 2004). More time spent reading unknown words also leads to better performance on subsequent testing for both native and non-native speakers (Dolgunsöz, 2015; Godfroid, Boers & Housen, 2013; Pellicer-Sánchez, 2016) As far as single words can be equated to longer phrases, when encountering unknown items, readers do seem to make use of "context-based strategies" (Pellicer- Sánchez, 2016, p.101) with some level of success.

In this paper we therefore set out to explore the role of context, and how this affects the real-time processing of both familiar and unfamiliar figurative phrases using eye-tracking to record reading patterns. We also aim to investigate how factors such as transparency affect processing, and how these variables interact.

Experiment 1

The aim of Experiment 1 was to compare the reading of highly familiar idioms, previously unseen idioms and conventional metaphors. These three types of phrase vary in that they are familiar in both form and meaning (idioms), familiar in meaning but not specific fixed form (conventional metaphors), and not familiar in either form or meaning (unknown idioms). We decided not to also include a set of novel metaphors on the grounds that these are effectively equivalent to unknown idioms: both are unfamiliar in form and meaning, hence relative transparency should be an important factor in either case.

Materials and methods

Materials were taken from Carrol, Littlemore and Dowens et al. (2018; see this paper for a full description of the selection of materials) and consisted of 20 L1 (English) idioms, 40 L2 (translated) idioms, and 20 metaphors. English idioms were chosen from previously published studies and were of the form X-det-Y, e.g. *over the moon*. Translated idioms were chosen from published studies that contained normative data (including English translations) for Bulgarian, German and Chinese phrases, and none had a congruent equivalent in English, e.g. *have hair on your teeth* (German, meaning "be easily annoyed"). These idioms were more variable than the English items in syntactic structure and length. Metaphors were of the form "A is B", e.g. *time is a doctor*, and were taken from previously published norming data (Katz et al., 1988). Ratings for familiarity, transparency and decomposability were collected from 31 native speakers during the development of materials in Carrol et al. (2018). Note that we consider transparency (how easily the meaning can be guessed before it is known) and decomposability (how much the figurative and literal meanings map onto one another) as separate variables here, and address this in more detail in the General Discussion.

For each phrase we created a literal paraphrase matched with the original for length in characters. For the metaphors, this often involved restating some aspect of the original but without the figurative element, e.g. *a smile is a knife* became *smiles can be nasty*. Balancing the individual frequency of all individual words would have placed too much of a restriction in terms of creating plausible paraphrases, so we accounted for component word frequency by obtaining British National Corpus (BNC) frequency counts for each content word in each phrase and calculating an aggregate frequency for the phrase or paraphrase (expressed on the Zipf scale, for ease of comparison).[1] We included these values in our analysis as a way of controlling for differences in frequency between conditions. Table 1 summarises key characteristics for all items (a list of phrases and paraphrases is provided in the appendix).

Table 1. Mean (SD in brackets) ratings for familiarity, transparency and decomposability (all out of 7), length (characters) and aggregate frequency for phrases and literal paraphrases

	Fam	Trans	Decomp	Length		Freq (Zipf)	
				Phrase	Paraphrase	Phrase	Paraphrase
L1 idioms	6.1 (0.78)	4.3 (0.56)	4.9 (0.61)	14.6 (2.2)	15.5 (2.4)	5.0 (0.5)	5.6 (0.7)
L2 idioms	1.8 (0.52)	2.7 (0.61)	3.5 (1.04)	25.0 (6.3)	24.8 (6.1)	5.2 (0.5)	5.8 (0.5)
Metaphors	3.2 (1.03)	4.5 (0.64)	6.0 (0.42)	21.6 (4.0)	23.4 (3.6)	6.5 (0.1)	6.0 (0.5)

We created a two-sentence context for each phrase, consisting of an introductory sentence followed by a sentence in which the phrase was embedded. These were constructed so that a

[1] The Zipf scale (Van Heuven et al. 2014) is a logarithmic scale to express relative frequency, taking into account the size of the corpus. A value of 1 represents 1 occurrence per 100 million words; 2 represents 10 occurrences per 100 million words; 3 represents 100 occurrences; and so on.

phrase and its paraphrase could be included in the same context with no other alterations. The first sentence (mean length = 55 characters, SD = 1.5) was designed to be relatively neutral, rather than providing any clues as to the meaning of the phrase. The second sentence (mean = 63, SD = 1.6 for figurative phrases; mean = 64, SD = 2.8 for paraphrases) had a short preamble to avoid the target phrase appearing at the start of a line, then the phrase or paraphrase appeared, then the sentence completed in a way that made sense within the context. Examples of experimental items (figurative phrase / literal paraphrase in bold) are:

Over the moon (L1 idiom)

The baby arrived last week and everyone is doing well.

They're all **over the moon / extremely happy** as they've wanted a family for a long time.

Have hair on your teeth (L2 idiom)

My father has always been a little bit short-tempered.

He **has hair on his teeth / is very easily annoyed** so just be careful what you say to him.

Time is a doctor (metaphor)

It seems bad now but it may look better in the future.

I think **time is a doctor / time improves things** so often you just need to wait things out.

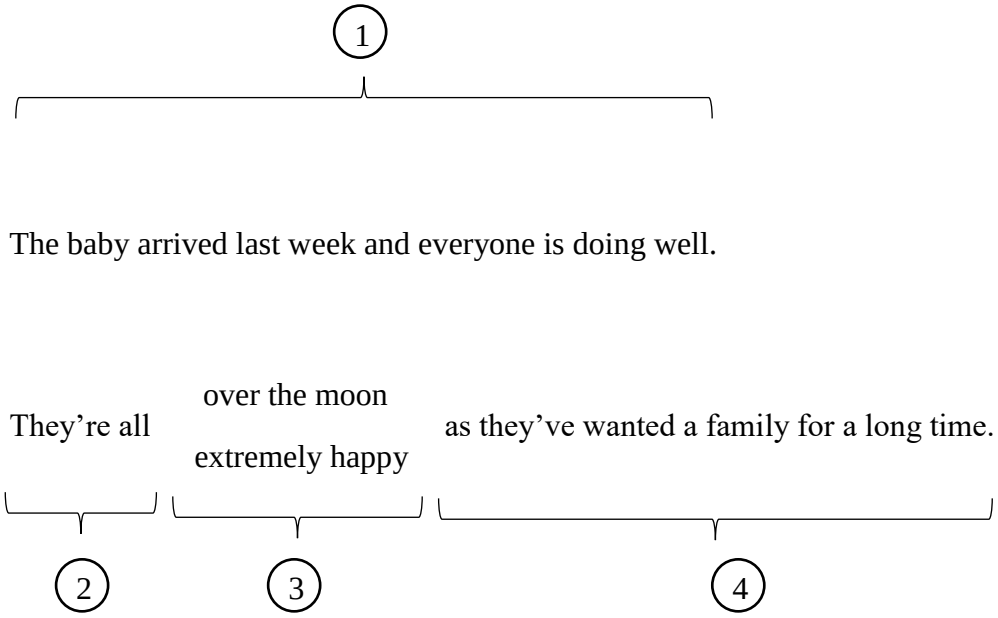
Items were counterbalanced over two lists so that each person saw equal amounts of figurative phrases and literal paraphrases for each phrase type. Lists were balanced for familiarity, transparency and decomposability. Participants saw 80 experimental items and 40 filler sentences of the same form (two connected sentences). For all 120 items the two sentences appeared on screen at the same time, one above the other, with double line spacing in between in Courier New, 18 point text. For each trial, participants were asked to read the sentences as naturally as possible and press the spacebar once they had done so. Eight practice items (same form as the fillers in the main experiment) were presented to begin, then all 120 items were presented in random order, with a break midway through. After each item, participants saw either a "Ready?" prompt or were asked a simple yes/no comprehension question to encourage attention throughout. Each participant was positioned at a comfortable height approximately 60cm from a 1280 x 1024 display. A chin rest was used to stabilise their head position, and eye-movements were recorded using an Eyelink 1000+ desk-mounted eye-tracker from SR Research Ltd at a sample rate of 500Hz. For all participants the left eye was tracked. Prior to starting, the camera was calibrated and validated using a nine-point calibration grid to ensure tracking accuracy. Additional calibrations were conducted as required during the study.

Results

Data from 51 participants (native English speakers, university students, age range = 18-22) was analysed. Data were manually checked and vertically aligned where required, and trials where track loss or timeout (no response 10 seconds after the start of the trial) had occurred were removed, leading to the loss of 1.8% of the data. Data were cleaned according to the default four-stage process within Eyelink DataViewer and very long (over 1000ms) or very short (under 80ms) fixations were removed. The eye tracking data was divided into four regions of interest (ROIs): the first sentence (1), then for the second sentence, the pre-phrase

region (2), the phrase region (3), and the post-phrase region (4). Figure 1 demonstrates the different regions for analysis. Any region that received no fixation during first pass reading was discounted from further analysis.

Figure 1. Example of experimental item for the L1 idiom *over the moon.* Sections are divided into Regions of Interest (ROIs) for subsequent analysis: 1 (context sentence), 2 (pre-phrase), 3 (phrase) and 4 (post-phrase).



Early and late eye-tracking measures were selected for analysis. Broadly, early measures relate to processes of recognition / initial lexical access and retrieval, while late measures reflect processes such as integration into the broader context (Staub & Rayner, 2007). Since our ROIs were all several words long, we used first pass reading time (first pass RT: an early measure of the sum total of all fixations within an ROI before the gaze exits to the left or right) and total reading time (total RT: a late measure defined as the sum of all fixations made

in a region throughout the trial, including re-reading) as our main fixation duration measures. We also analysed number of regressions into an ROI (for regions 1, 2 and 3) and number of regressions out (for region 4). For region 3 (the phrase itself) we included regression path duration (also known as go past time), which measures time spent in a region before gaze exits to the right, including any regressions made before gaze moves forward out of the ROI.

We constructed separate linear mixed-effects models for each region and for each measure using the lme4 package (version 1.1-21; Bates et al., 2015) in R (version 3.5.3; R Core team, 2019) and R Studio (version 1.1.463). Duration measures were log-transformed to reduce skewing. For measures involving count data (number of regressions into or out of a particular region) we used a poisson distribution in a generalised linear model. All models included the interaction of the treatment-coded variable phrase type (L1 idiom = baseline, vs. L2 idiom and metaphor) and condition, which was sum-coded. Differences between levels of factors were calculated using the difflsmeans function in the lmerTest package (version 3.1-0; Kuznetsova, Brockhoff & Christensen, 2017). Trial order, list and length of the region (centred and scaled to help with model fitting) were included as covariates in all models.[2] We included random intercepts for subject and item, by-subject random slopes for the effect of condition and type and by-item random slopes for the effect of condition, where model comparison showed that these were warranted. A summary of results is provided in table 2.

[2] For the phrase region we added the aggregate frequency score as a covariate to control for differences between figurative and literal phrases. Since the three phrase types varied considerably in length, we also included length of the phrase as an interaction with phrase type and condition, where model comparison showed that this was an improvement over inclusion as a fixed effect.

Table 2. Mean reading times in milliseconds (SD in brackets) for each phrase type (L1 idioms, L2 idioms, metaphors) for figurative (Fig) vs. literal paraphrase (Lit) conditions.

	L1 idioms			L2 idioms			Metaphor	
	Fig	Lit		Fig	Lit		Fig	Lit
Context sentence								
First pass RT	1669	1706	*	1714	1710		1759	1746
	(587)	(560)		(588)	(574)		(628)	(593)
Total RT	1843	1900	*	1917	1893		1924	1902
	(663)	(681)		(681)	(700)		(704)	(653)
Regs in	0.40	0.43		0.39	0.40		0.34	0.37
	(0.55)	(0.57)		(0.54)	(0.55)		(0.49)	(0.53)
Pre-phrase region								
First pass RT	344	312	**	346	342		285	278
	(178)	(147)		(199)	(196)		(149)	(134)
Total RT	401	380		458	408	**	321	314
	(211)	(206)		(306)	(263)		(176)	(172)
Regs in	0.25	0.28		0.41	0.27	***	0.39	0.36
	(0.50)	(0.52)		(0.60)	(0.51)		(0.55)	(0.53)
Phrase region								
First pass RT	367	407		701	578	***	523	539
	(173)	(244)		(432)	(341)		(334)	(339)
Total RT	443	536	**	1037	762	***	791	801
	(199)	(330)		(527)	(387)		(392)	(366)
Reg path	482	593	**	969	785	***	843	836
	(271)	(492)		(538)	(466)		(511)	(483)
Regs in	0.08	0.10		0.29	0.13	***	0.17	019
	(0.27)	(0.31)		(0.53)	(0.37)		(0.40)	(0.45)
Post-phrase region								
First pass RT	1074	1089		865	865		1007	1003
	(484)	(472)		(510)	(482)		(476)	(527)
Total RT	1167	1183		1066	956	***	1109	1152
	(480)	(444)		(526)	(486)		(490)	(562)
Regs out	0.44	0.49		0.66	0.48	***	0.45	0.49
	(0.57)	(0.58)		(0.71)	(0.61)		(0.57)	(0.63)

Significant differences between figurative and literal conditions (* p <.05; ** p < .01; *** p < .001) are based on values extracted using the difflsmeans function in the lmerTest package (version 3.1-0; Kuznetsova, Brockhoff & Christensen, 2017) in R.

Effects were minimal for the context sentence, with no consistent differences for any phrase type. For the pre-phrase region there were significant effects of condition for L2 idioms for

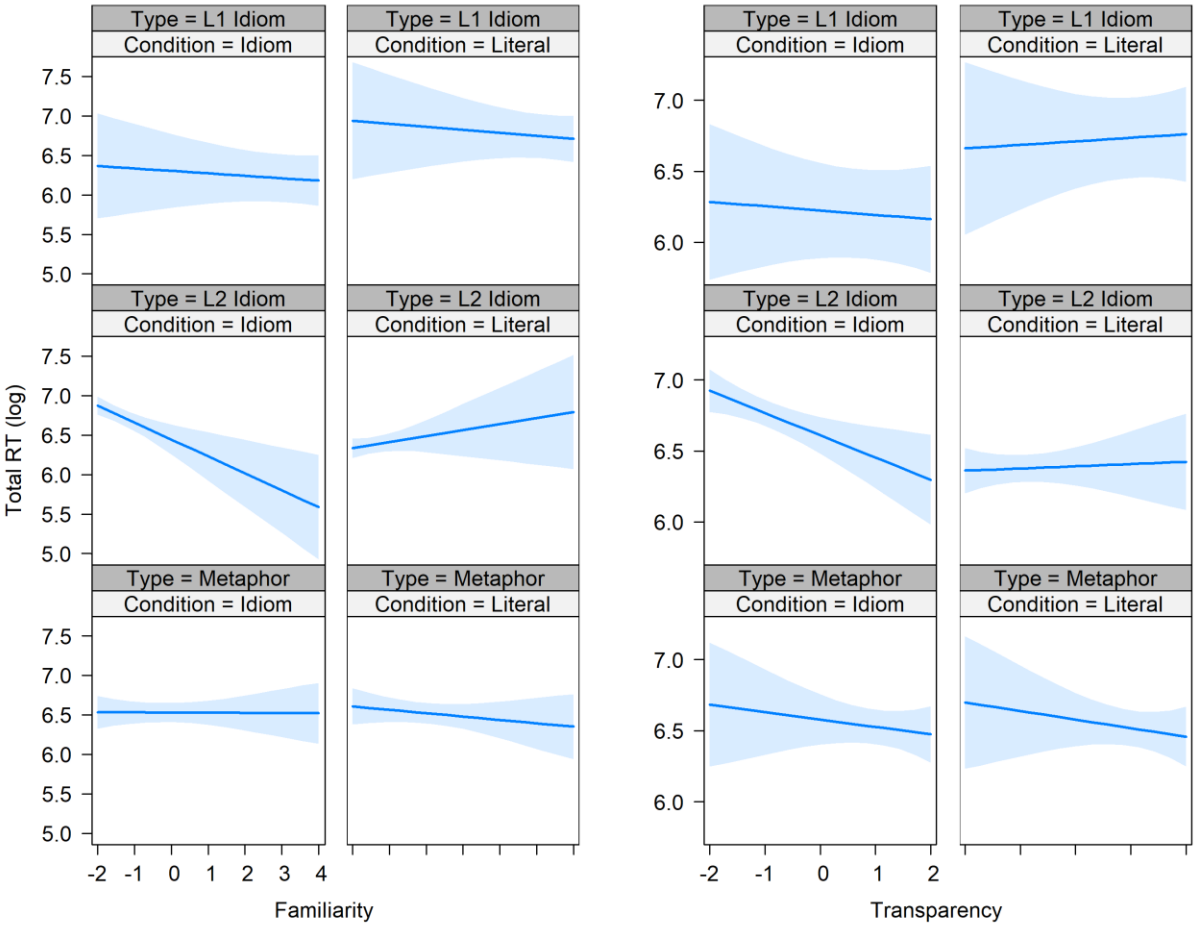
total RT (β = 0.09, *t* = 3.07, *p* = .003) and regressions in (β = -0.42, *z* = -5.35, *p* < .001). For the phrase region there were significantly shorter total RTs (β = -0.54, *t* = 2.70, *p* = .008) and regression paths (β = -0.65, *t* = 3.39, *p* = .001) for L1 idioms, and for L2 idioms significantly longer first pass RTs (β = 0.19, *t* = 4.17, *p* < .001), total RTs (β = 0.37, *t* = 7.01, *p* < .001), regression paths (β = 0.26, *t* = 5.37, *p* < .001) and more regressions in (β = 0.48, *z* = -8.22, *p* < .001). In the post-phrase region there were significant effects of condition for L2 idioms for total RT (β = 0.11, *t* = 4.46, *p* < .001) and regressions out (β = -0.31, *z* = -5.13, *p* < .001).

We next considered the effects of familiarity, transparency and decomposability (all centred) and compared whether inclusion of these made an improvement as a fixed effect, then as an interaction with condition, for each region. For the context sentence, familiarity made an improvement to the model for total RT as a fixed effect ($\chi 2$ (1) = 4.09, *p* = .043), whereby greater familiarity led to shorter RTs across all phrase types and conditions. For the pre-phrase region, the model for number of regressions was improved by the addition of interactions with transparency ($\chi 2$ (6) = 18.20, *p* = .006), familiarity ($\chi 2$ (6) = 22.19, *p* = .001) and decomposability ($\chi 2$ (6) = 14.07, *p* = .029), where higher ratings led to fewer regressions into the region for figurative phrases for L1 and L2 idioms in the figurative condition.

For the phrase region, the model for first pass RT was marginally improved by the addition of a fixed effect of familiarity ($\chi 2$ (1) = 3.34, *p* = .068), whereby more familiar phrases (of all types) were read more quickly. For total RT, there was a marginal improvement by adding an interaction with transparency ($\chi 2$ (6) = 11.38, *p* = .077) and a significant improvement by adding an interaction with familiarity ($\chi 2$ (6) = 19.03, *p* = .004). In both cases, higher ratings had a facilitative effect for L2 idioms in the figurative condition only, but no effect for other phrase types (Figure 2). Regression path duration was improved by the addition of transparency as a main effect ($\chi 2$ (1) = 4.29, *p* = .038; more transparent phrases had shorter regression paths across the board) and familiarity as an interaction ($\chi 2$ (6) = 18.73, *p* = .005;

more familiar phrases had shorter regression paths for L2 idioms in the figurative condition only). Number of regressions in was improved by interactions with each of transparency ($\chi 2$ (6) = 18.21, p = .006), familiarity ($\chi 2$ (6) = 15.66, p = .016) and decomposability ($\chi 2$ (6) = 13.27, p = .039). Higher transparency and familiarity showed a trend toward fewer regressions in for all phrase types in the figurative condition, while decomposability showed an effect (fewer regressions in the figurative but not literal condition) for L1 idioms only.

Figure 2. Effects plots showing the relative effects of transparency (left) and familiarity (right) on total RTs for the whole phrase. Only L2 idioms (middle row) in the figurative condition showed significant effects (transparency, t = -3.36, p = .001; familiarity, t = -3.82, p < .001). Shading shows 95% confidence intervals. Plots produced using the Effects package (version 4.1-3, Fox and Weisberg, 2019) in R.



In the post-phrase region, each of transparency ($\chi 2$ (1) = 4.39, p = .036) and familiarity ($\chi 2$ (1) = 4.32, p = .038) improved the model for first pass RT as fixed effects (higher ratings led to shorter reading times for all phrase types and conditions). Each of transparency ($\chi 2$ (6) = 20.58, p = .002), familiarity ($\chi 2$ (6) = 17.15, p = .009) and decomposability ($\chi 2$ (6) = 13.11, p = .041) improved the model for total RT as interactions, and in each case showed a trend

toward shorter time overall spent reading the rest of the sentence following a figurative phrase. Each of transparency ($\chi 2$ (6) = 15.30, p = .018), marginally familiarity ($\chi 2$ (6) = 11.36, p = .078) and decomposability ($\chi 2$ (6) = 13.40 p = .037) improved the model for regressions out, with trends toward fewer regressions for more transparent, familiar and decomposable phrases of all types.

Discussion

L1 idioms were read more quickly than literal paraphrases, L2 idioms were read more slowly than literal versions, and metaphors showed no difference compared to their literal equivalents. For L1 idioms, this advantage mirrors the widespread effects reported in the literature and were apparent in total RT and regression path duration (with a trend for first pass RT). This suggests that not only are idioms recognised quickly as examples of formulaic sequences, but also that their overall meanings are activated and integrated into the surrounding context more easily than their literal paraphrases. Since this advantage for well-known idioms is consistent with studies that have used a range of different tasks in the literature (e.g. meaningfulness judgments, as well as reading times), we assume that this reflects an ability to quickly and easily understand each phrase, rather than demonstrating broader reading strategies that facilitate overall understanding. Factors like transparency and decomposability had limited effects for L1 idioms, hence for highly familiar phrases that can be simply ‘retrieved’, aspects of their semantic make-up have little influence during on-line processing. Relative familiarity also had little effect, which is unsurprising as the items were all chosen to be highly familiar in the first place.

In contrast, L2 idioms showed clear disruption across all measures: readers experienced difficulty immediately as they were reading them, returned to the preceding context to attempt to resolve this, and subsequently spent longer reading both the phrase and the

following sentence. As we might expect, when phrases are fundamentally unknown, readers need longer to (attempt to) work out the intended meaning, relative to a literal phrase expressing an equivalent meaning. Importantly, each of transparency and familiarity had an effect on reading times (principally total RT, but also regression path duration) for L2 idioms. For transparency, this means that readers can more easily parse the intended meaning when this is easier to work out, compared to a phrase that is more opaque. Whilst this seems logical, the fact that it exerts an influence in real-time processing is noteworthy. Familiarity might be better seen here as *perceived* familiarity (since all L2 idioms were fundamentally unknown), hence when a phrase expresses a familiar-sounding idea the meaning can be more easily retrieved, even if the specific idiom itself has not been heard before.

For metaphors, no consistent effects were observed for any measure in any region. Since the metaphors were highly conventional, this supports Bowdle and Gentner's (2005) model (as well as other studies discussed in the introduction showing clear effects of familiarity and conventionality), where it is assumed that conventional metaphors are recognised as examples of an established category rather than understood through a (more effortful) process of active comparison between topic and vehicle. Novel metaphors might be expected to behave in a similar way to more transparent L2 idioms: since neither form nor meaning is familiar, relative transparency would be the primary driver of how much difficulty is encountered in working out the meaning in real time. Importantly for the metaphors used in this study, although the meaning may be conventionalised, the form itself is not fixed, hence we see no advantage here in the same way as for idioms.[3]

[3] Here and elsewhere, 'fixed form' refers to the lexical composition for any given item. Hence idioms are 'fixed', in the sense that they are a specific combination of words that cannot be substituted or changed. Although the metaphors we chose all had the same general schema/structure 'A is B', their lexical composition is not fixed or formulaic in the same way as idioms.

Finally, evidence that readers were actively using the surrounding context was limited to L2 idioms, where both the pre-phrase and post-phrase regions had longer total RTs for figurative phrases. Of note, there was no indication at all that readers returned to the initial context sentence, even for unknown phrases. Given this, we next considered whether providing a more informative context may reduce the disruption seen for unknown idioms.

Experiment 2

In Experiment 2 we varied the strength of the context that was provided in the initial sentence. We also added a multiple-choice test following the eye-tracking experiment, to assess how well participants were able to identify the meaning of idioms that they had seen during the reading task. We predicted that a more informative context may have negligible benefit for already known phrases (L1 idioms), but may reduce the disruption seen for unfamiliar phrases, and make subsequently identifying the meaning easier.

Materials and methods

We used the same items as in Experiment 1, but discounted metaphors from further investigation because of the lack of any significant effects. For each of the 20 English idioms and 40 L2 idioms we then adapted the contexts in order to provide two versions: one where the initial sentence was neutral (low context) and one where the initial sentence was biased to provide an indication of the meaning of the upcoming sentence (high context). We adapted the second sentence so that it was congruent with both versions of the context sentence. The second sentence consisted of a minimum of four words in the pre-phrase region, then the phrase or literal paraphrase, then an ending to the sentence that was consistent with the meaning in both low and high contexts. In order to ensure that the contexts were sufficiently biasing, we placed all literal versions of the items into an online survey to collect ratings of the strength of the context. We presented one version of the context sentence

(counterbalanced over two lists, so that people rated only the low or high version for any item) followed by the sentence containing the literal paraphrase and asked raters (n = 60, all native speakers of British or American English) to decide on a five-point scale how strongly the first sentence predicted the second. Data was collected via Amazon Mechanical Turk and participants were paid a small fee for completing the survey. Following this initial rating we made some adjustments to items where the contexts were not judged as high/low as intended and redistributed the survey to a new set of raters (n = 48). In this final version of materials, low context sentences were rated 2.8/5 (SD = 0.29), and high context sentences were rated 3.8/5 (SD = 0.32); a paired samples *t*-test confirmed that these were significantly different: $t(59) = 20.99$, $p < .001$. Examples of a high and low context sentence are:

Over the moon (L1 idiom)

High: Did John and Mary tell you their good news last week?

Low: Did you get chance to speak to John and Mary last week?

They seem to be **over the moon / extremely happy** as they've wanted this for a long time.

Have hair on your teeth (L2 idiom)

High: My father has always been a little bit short-tempered.

Low: I don't think you've ever really got to know my father.

We know he's always **had hair on his teeth / been easily annoyed** so watch what you say.

The items were arranged over four counterbalanced lists so that each participant saw 10 L1 idioms (5 x low context; 5 x high context) and 10 literal paraphrases (5 x low context; 5 x high context), and 20 L2 idioms (10 x low context; 10 x high context) and 20 literal paraphrases (10 x low context; 10 x high context). No participant saw the same figurative and literal version of any one item, or saw the same item in more than one context condition. We added in the 40 filler items from Experiment 1 to make a total of 100 experimental sentences. Procedures for the experiment were the same as in Experiment 1, but following the main reading task, participants undertook a short unrelated task lasting around 10 minutes, then were presented with all 60 idioms (20 L1, 40 L2) and asked to identify what they thought was the correct figurative meaning from a set of four options. These options represented plausible meanings for each phrase and were developed as part of the procedure in Carrol et al. (2018), where native speakers identified 89% of L1 idioms correctly and 38% of L2 idioms correctly. Because of the counterbalancing of items, some of the idioms seen during this follow-up test would have been seen during the reading task, while others would have been seen in their paraphrase form (i.e. depending on what version of materials was assigned, a participant might see *over the moon* in the reading task and then be asked the meaning in the follow up, but might see *be easily annoyed* in the reading task and then be asked the meaning of *have hair on your teeth* in the follow up). This allowed us to assess how much a prior encounter with a phrase in any kind of context helped with subsequent identification of the meaning, compared to phrases that had not been encountered at all during the reading task.

Results

Data were analysed for 64 participants drawn from the same population as in Experiment 1. The same data checking, cleaning and removal procedures were applied, leading to the removal of 3.6% of the data. The ROIs were again defined as the initial context sentence (1), then in the second sentence the pre-phrase (2), phrase (3) and post-phrase (4) regions, and the

same eye-tracking measures were analysed. Each model included the three-way interaction of the sum-coded factors phrase type (L1 idiom vs. L2 idiom), condition (figurative vs. literal) and context (high vs. low), along with the length of the region, trial order and list as covariates, random intercepts for subject and item, by-subject random slopes for the effects of phrase type, condition and context type, and by-item random slopes for the effects of condition and context type, where model comparison showed that these were warranted. A summary of results is provided in table 3.

Table 3. Summary of reading measures (means, with SD in brackets) for L1 idioms vs. L2 idioms in low vs. high contexts for figurative vs. literal conditions. Duration measures are in milliseconds.

	L1 idioms					L2 idioms				
	High			Low		High			Low	
	Figurative	Literal		Figurative	Literal	Figurative	Literal		Figurative	Literal	
Context sentence											
First pass RT	1851 (682)	1880 (709)		1822 (686)	1787 (732)	1910 (693)	1898 (703)		1810 (674)	1837 (691)	
Total RT	2010 (744)	2031 (767)		1955 (721)	1989 (813)	2064 (734)	2061 (768)		1994 (750)	2001 (744)	
Regressions in	0.29 (0.49)	0.34 (0.53)		0.29 (0.49)	0.36 (0.53)	0.30 (0.49)	0.32 (0.51)		0.36 (0.54)	0.33 (0.52)	
Pre-phrase region											
First pass RT	590 (295)	551 (286)	*	581 (303)	536 (256)	570 (312)	565 (307)		581 (312)	569 (297)	
Total RT	660 (328)	661 (327)		660 (330)	641 (292)	726 (397)	661 (361)	**	758 (391)	683 (354)	***
Regressions in	0.21 (0.44)	0.31 (0.56)	*	0.21 (0.42)	0.29 (0.53)	0.43 (0.63)	0.27 (0.54)	***	0.43 (0.64)	0.30 (0.54)	***
Phrase region											
First pass RT	382 (174)	420 (235)	*	408 (189)	436 (243)	*	730 (429)	591 (345)	***	736 (436)	602 (335)	***
Total RT	455 (223)	528 (302)	*	479 (209)	543 (276)	*	1110 (535)	786 (416)	***	1134 (536)	821 (395)	***
Regression path	483 (364)	561 (412)	**	510 (360)	555 (349)	*	1033 (574)	768 (467)	***	1064 (585)	812 (466)	***
Regressions in	0.11 (0.31)	0.12 (0.37)		0.09 (0.31)	0.11 (0.33)	0.35 (0.57)	0.22 (0.45)	***	0.33 (0.56)	0.24 (0.48)	***
Post-phrase region											
First pass RT	897 (483)	933 (492)		857 (410)	855 (412)	724 (431)	739 (456)		762 (482)	712 (434)	
Total RT	989 (489)	1041 (493)		964 (451)	995 (459)	928 (488)	863 (460)	**	952 (509)	851 (481)	***
Regressions out	0.40 (0.57)	0.46 (0.66)		0.39 (0.54)	0.50 (0.61)	0.66 (0.72)	0.53 (0.65)	**	0.68 (0.72)	0.55 (0.66)	**
Meaning correct (%)	91 (28)	91 (29)		93 (26)	89 (31)	*	50 (50)	40 (49)	***	56 (50)	43 (49)	***

Significant differences between figurative and literal conditions (* p <.05; ** p < .01; *** p < .001) are based on values extracted using the difflsmeans function in the lmerTest package (version 3.1-0; Kuznetsova, Brockhoff & Christensen, 2017) in R.

For the context sentence there was a main effect of context for first pass RT (β = 0.04, *t* = 2.23, *p* = .030), whereby high context sentences took slightly longer to read than low context sentences. This may reflect that fact that high context sentences contained more informative content, while low context sentences were neutral. There were no effects of phrase type or condition, and no interactions. For the pre-phrase region there was a main effect of condition (β = 0.02, *t* = 2.64, *p* = .008) whereby readers spent longer in the pre-phrase region during first pass RT for figurative phrases. For total RT there was an interaction between phrase type and condition (β = 0.02, *t* = -2.64, *p* = .011) whereby L2 idioms had longer RTs in the figurative condition but L1 idioms did not. For number of regressions in, there was an interaction of phrase type and condition (β = -0.02, *z* = -2.64, *p* = .011), whereby fewer regressions were made to the preceding context for L1 idioms in the figurative condition (compared to literal), but more were made for L2 idioms.

For the phrase itself we included aggregate frequency as a covariate, and included the interaction of Length and Condition as in Experiment 1. There was a significant interaction of phrase type and condition for first pass RT (β = -0.16 *t* = -4.10, *p* < .001), total RT (β = -0.22, *t* = -4.79, *p* < .001) and regression path duration (β = -0.19, *t* = -5.15, *p* < .001), and a main effect of phrase type (β = -0.44, *z* = -2.52, *p* = .012) for number of regressions in. This meant that L1 idioms were in general read more quickly than literal paraphrases, while the opposite was true for L2 idioms. There was also an overall effect of context for total RT (β = -0.02, *t* = -3.58, *p* < .001) and regression path duration (β = -0.02, *t* = -2.84, *p* = .005). Inspection of between-group differences suggested that this was driven by figurative uses of L1 idioms, which had shorter total RTs (β = -0.08, *t* = -2.54, *p* = .011) and marginally shorter regression paths (β = -0.07, *t* = -1.93, *p* = .054) in high contexts compared to low contexts; and also by literal uses of L2 idioms, which had shorter total RTs (β = -0.05, *t* = -2.06, *p* = .039) and shorter regression paths (β = -0.06, *t* = -2.64, *p* = .008) in high compared to low contexts.

For the post-phrase regions total RT showed an interaction of condition and type (β = -0.04, *t* = -3.69, *p* < .001) whereby more time was spent overall on L2 idioms in the figurative condition, relative to literal. For regressions out of this region there was an interaction of phrase type and condition (β = -0.11, *z* = -4.26, *p* < .001) whereby L2 idioms showed more regressions in the figurative condition than the literal.

We next considered the effects of transparency, familiarity and decomposability. For both the context sentence and the pre-phrase region, none of transparency, familiarity or decomposability improved the model for any measure. For the phrase region, the models for total RT ($\chi 2$ (8) = 23.87, *p* = .002) and number of regressions in ($\chi 2$ (8) = 50.55, *p* < .001) were improved by the addition of an interaction with familiarity. Here, higher familiarity was broadly facilitative for both L1 and L2 idioms in figurative conditions, with shorter reading times and fewer regressions for more familiar phrases, but there was no difference between high and low contexts. For regressions in, the model was also improved by the addition of an interaction with transparency ($\chi 2$ (8) = 34.63, *p* < .001) and marginally decomposability ($\chi 2$ (8) = 13.70, *p* = .090), whereby higher ratings for L2 idioms in both high and low contexts led to fewer regressions to the phrase in the figurative but not literal condition. For the post-phrase, first pass RT was improved by the addition of fixed effects for transparency ($\chi 2$ (1) = 4.47, *p* = .034) and decomposability ($\chi 2$ (1) = 4.42, *p* = .036), whereby more transparent and more decomposable phrases of both types meant that less time was spent reading the rest of the sentence on first encounter. Total RT was improved by the addition of interactions with each of familiarity ($\chi 2$ (8) = 27.78, *p* < .001) and transparency ($\chi 2$ (8) = 21.32, *p* = .006). Here, higher ratings led to shorter overall RTs for L2 idioms in the figurative condition, and shorter RTs for figurative L1 idioms in the low context condition only. For regressions out of the region the model was improved by interactions with familiarity ($\chi 2$ (8) = 27.69, *p* < .001), transparency ($\chi 2$ (8) = 26.34, *p* < .001) and marginally decomposability ($\chi 2$ (8) = 14.42, *p* =

.071). Here, familiarity was broadly facilitative (higher ratings = fewer regressions) for both phrase types in the figurative condition, and transparency and decomposability were facilitative for figurative L2 idioms and figurative L1 idioms in low contexts only.

We finally considered performance on the post-test asking participants to identify the meaning of the idioms included in the study. The percentage of correctly identified idiom meanings was significantly higher overall for L1 idioms (mean = 91%) compared to L2 (mean = 47%), with main effects of phrase type ($\beta = 1.55$, $z = 8.66$, $p < .001$) and condition ($\beta = 0.25$, $z = 4.18$, $p < .001$), but no effect of context. Transparency made a marginal improvement as a fixed effect ($\chi2$ (1) = 3.21, $p = .073$) with more transparent phrases being better identified for both phrase types. Familiarity made an improvement as an interaction ($\chi2$ (6) = 15.71, $p = .047$). Here, familiarity was more beneficial in identifying meaning for L2 idioms, with no effects of condition or context. Decomposability improved the model as a fixed effect ($\chi2$ (1) = 21.13, $p < .001$), and for both phrase types seen in both contexts, higher decomposability led to better identification of meaning. We additionally considered whether the time spent reading each phrase was a predictor of successful meaning identification, i.e. whether people who spent longer reading an idiom subsequently performed better in terms of correctly guessing the meaning in the post-test. We analysed the phrase region only, but found no effects of any of first pass RT, total RT or regression path duration on subsequent success in the post-test for either L1 or L2 idioms.

Discussion

Results from Experiment 2 largely mirror Experiment 1, where L1 idioms were in general read more quickly than literal paraphrases, and L2 idioms were in general read more slowly, with more looks back to the immediate pre-phrase region, and more time spent reading the remainder of the sentence. As in Experiment 1, familiarity, transparency and decomposability

had facilitative effects primarily for L2 idioms, with higher ratings reducing reading times and regressions. There was limited evidence that context played any role. There was some indication that a high context was beneficial to the reading of L1 idioms (but not literal paraphrases). L2 idioms in the literal condition were also somewhat helped by an informative context, but notably figurative uses were not; crucially, this suggests that the lack of an effect was not simply the result of the contexts not being strong enough to bias the meanings in the way that we intended. Transparency and decomposability seemed to play a role for L1 idioms only in low contexts, but had limited effects when a high context was provided. Readers did not look back more often to the initial context sentence, despite this being the location of the helpful additional information, for either phrase type or in either condition.

The ability to correctly guess the meaning of items was similarly unaffected by context. L1 idioms were all at or above 90% accurate, suggesting that as familiar phrases they were easily identified by participants, regardless of the context and regardless of whether they had actually been seen in the reading task. For L2 idioms, however, whilst strength of context made no difference, condition had a clear effect, with better identification of the correct meaning for phrases that had already been encountered during the reading task compared to when a literal paraphrase had been seen instead. While strength of context may therefore have limited benefit in a one-time exposure (in terms of working out a precise meaning), the fact that an unknown idiom had been seen in any kind of context was beneficial for readers in narrowing down the range of possible meanings. In other words, usage of a phrase in any sentence immediately limits and constrains what it could mean, e.g. encountering *have hair on your teeth* in the context of interacting with someone's father tells us that it is likely to describe some aspect of his personality. This process of honing potential meanings may be one of the contributing mechanisms that help us to work out the actual definition of an unknown word or phrase. Notably, longer time spent considering L2 idioms during the

reading task did not lead to any greater success in identifying the meaning, contrary to results discussed in the introduction (Dolgunsöz, 2015; Godfroid, Boers & Housen, 2013; Pellicer-Sánchez, 2016), where more time spent on unknown words in general leads to better performance in subsequent vocabulary tests. This may be a reflection of the greater difficulty posed by figurative phrases (as opposed to single words), where more time spent reading the phrase may instead reflect the process of trying and failing to assign a logical meaning on first encounter. In contrast, a single (literal) word may be much more straightforward and/or constrained, hence easier for readers to work out.

General Discussion

In two experiments we compared reading patterns for figurative phrases, relative to literal paraphrases expressing an equivalent meaning. The established 'idiom superiority effect' for well-known phrases was observed in both, with little indication that semantic factors (transparency and decomposability) or context had a strong effect on how the phrases were recognised or understood. Factors such as literal plausibility and dominance of figurative vs. literal meanings do interfere with idiom processing (e.g. Findlay & Carrol, 2019; Milburn & Warren, 2019; Titone & Libben, 2014) and the component words of idioms do contribute to how they are understood and processed (e.g. Hamblin & Gibbs, 1999; Smolka, Rabanus & Rösler, 2007), but at least in our study, these seem to have little effect on how highly familiar phrases are read.

In contrast, we observed clear disruption across all measures for fundamentally unknown phrases. Readers spent more time on first encounter with the phrase, were more likely to immediately regress to the preceding context, and subsequently spent more time reading and re-reading the remainder of the sentence. The pattern mirrors that of non-native speakers for English idioms where figurative expressions often show longer reading times than literal ones

(e.g. Carrol & Conklin, 2017; Siyanova-Chanturia et al., 2011). Cieślicka (2006) described this in terms of 'literal salience', whereby a literal, word-by-word reading remains the default approach for language learners, even when they have some knowledge of the figurative meaning of that idiom. The fact that we found the same pattern of results for unknown phrases may indicate that readers simply process incoming material in an incremental manner unless there is a reason not to (i.e. unless a known configuration is recognised and 'retrieved'). Approaches such as the Graded Salience Hypothesis (Giora, 2003) suggest that the most 'salient' meaning (figurative or literal) is the one that is most readily processed, hence for phrases where no figurative meaning is known, a literal reading would by necessity be the most salient; if this fails to deliver a logical interpretation then readers must derive a plausible alternative (see also the hierarchy of utterance interpretation outlined by Kecskés, 2006). Facilitative effects of transparency (in both experiments) suggest that this later, meaning generation phase is easier when a phrase is more amenable to semantic analysis than when it is not.

In Experiment 2 we saw no evidence that context helps readers when faced with unknown figurative expressions, which was unexpected. There was some limited evidence that high context helped the processing of L1 idioms (in line with studies such as Fanari, Cacciari & Tabossi, 2010; Schweigert & Moates, 1988), but there was no indication that this helped to work out the meaning for unknown phrases, either in terms of online reading patterns or the ability to subsequently identify the correct meaning in a post-test. Along with semantic analysis, inferring from context is one of the pillars of the global elaboration model (Levorato & Cacciari, 1995, 1999), as well as other approaches such as constraint-satisfaction accounts (e.g. Titone & Connine, 1999; Libben & Titone, 2008 as they relate to idioms; MacDonald & Seidenberg, 2006 for language processing more generally). These models suggest that all possible information is used to reach an appropriate interpretation, although they do also

stress that this process is grounded in probabilistic information based on prior experience. Similarly, Frisson and Pickering (2001) proposed an under-specification model for how the language processor makes sense of figurative uses of words and phrases, but point out that this can only really account for senses that are fundamentally known in the first place. One possibility is therefore that context matters much less for the process of sense generation (coming up with an appropriate interpretation for a previously unseen phrase) than for the process of sense selection (deciding between possible meanings for an already known phrase). When language users are required to do the former (work out previously unknown meanings), semantic analysis is the predominant factor. Effects here may also be highly task-dependent: those studies that have found guessing from context to be a significant strategy did so using a think-aloud task (Wray, Bell & Jones, 2016; Zuo, 2008) or by asking participants to actively guess the meaning from the context provided (Xie, 2017), which may encourage this strategy in a way that is not reflected in natural reading. This is potentially very important, since such tasks are not generally part of 'normal' language use. The situations when language users are most likely to encounter figurative expressions – everyday spoken or written discourse – may therefore be precisely those where contextual support has the least benefit.

The overall finding regarding context seems to be that any kind of context is helpful when it comes to subsequently identifying the meaning of known phrases. In the post-test in Experiment 2, those idioms that had been seen in the figurative condition were better identified, suggesting that even this one-time exposure was enough to at least begin to constrain the meaning. There is some evidence that the salience of idioms adds to their learnability (e.g. Reuterskiöld & Van Lancker Sidtis, 2012), which may also help in terms of subsequent identification, but as with the literature on the learning of words from incidental reading (e.g. Pellicer-Sánchez, 2016), it may be that several repetitions are required before

any real benefit can be observed. More important are the semantic factors identified above, which can help readers to process text more easily in real time. While familiarity remains the most obvious predictor (and obviates the need for active analysis), transparency and to a lesser extent decomposability made clear contributions. We make an important distinction here between transparency and decomposability, and our results support a view where transparency is more likely than decomposability to predict processing difficulty for unknown idioms. In other words, encountering a phrase for the first time requires the language user to actively generate possible meanings, and a variety of factors contribute to how successfully this can be achieved (such as vocabulary knowledge, underlying metaphors, knowledge of the source domain and cultural knowledge). Once a meaning is revealed (i.e. once we are told what an idiom *actually* means), the relationship between the literal and figurative meaning of the phrase (its decomposability) can be evaluated, but this is more of a post-hoc process. Notably, decomposability showed its clearest effects in the identification of meaning in Experiment 2, where participants could actively compare literal and (possible) figurative meanings in order to judge the most likely.

Finally, while we have addressed the role of context in a fairly narrow, linguistic sense here, it is clear that in real life, context extends beyond simply what is said (or written), and incorporates a range of information that jointly help to shape meaning (e.g. Halliday, 1991; Kecskés, 2006). Our study therefore has methodological implications for how we might investigate these questions further, and raises the question of how to balance the need for naturalistic contexts with controlled experimental materials, to best reflect the experience of language users ‘in the wild’ when they encounter idioms in a range of discourse environments.

References

Bambini, V., Bertini, C., Schaeken, W., Stella, A. and Di Russo, F. (2016). Disentangling Metaphor from Context: An ERP Study. *Frontiers in Psychology, 7*, 59. DOI: 10.3389/fpsyg.2016.00559

Bambini, V., Canal, P., Resta, D. and Grimaldi, M. (2019) Time Course and Neurophysiological Underpinnings of Metaphor in Literary Context. *Discourse Processes, 56*(1), 77-97. DOI: 10.1080/0163853X.2017.1401876

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. DOI: 10.18637/jss.v067.i01

Beck. S. and Weber, A. (2018). Context Matters, Figuratively, for L2 Readers: Evidence from Self-paced Reading. *Proceedings of the 11th International Conference on the Mental Lexicon 2018*, Edmonton, AB, Canada.

Blasko, D., and Briihl, D. (1997). Reading and Recall of Metaphorical Sentences: Effects of Familiarity and Context. *Metaphor and Symbol, 12*(4), 261-285. DOI: 10.1207/s15327868ms1204_4

Blasko, D., and Connine, C. (1993). Effects of Familiarity and Aptness on Metaphor Processing. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*(2), 295-308. DOI: 10.1037/0278-7393.19.2.295

Boers, F., Eyckmans, J., and Stengers, H. (2007). Presenting figurative idioms with a touch of etymology: more than mere mnemonics? *Language Teaching Research, 11*(1), 43-62. DOI: 10.1177/1362168806072460

Bowdle, B. F., and Gentner, D. (2005). The career of metaphor. *Psychological Review, 112*(1), 193-216. DOI: 10.1037/0033-295X.112.1.193

Cacciari, C., & Glucksberg, S. (1991). Understanding idiomatic expressions: The contribution of word meanings. In G. Simpson (Ed.), *Understanding Word and Sentence* (pp. 217-240). North Holland: Elsevier.

Caillies, S., & Butcher, K. (2007). Processing of Idiomatic Expressions: Evidence for a New Hybrid View. *Metaphor and Symbol, 22*(1), 79-108.

Cardillo, E., Watson, C., Schmidt, G., Kranjec, A. and Chaterjee, A. (2012). From novel to familiar: Tuning the brain for metaphors. *Neuroimage, 59*(4), 3212–3221. DOI: 10.1016/j.neuroimage.2011.11.079.

Carrol, G. and Conklin, K. (2014). Getting your wires crossed: Evidence for fast processing of L1 idioms in an L2. *Bilingualism: Language and Cognition, 17*(4), 784-797. DOI: 10.1017/S1366728913000795

Carrol, G., and Conklin, K. (2017). Cross language priming extends to formulaic units: Evidence from eye-tracking suggests that this idea “has legs”. *Bilingualism: Language and Cognition, 20*(2), 299-317. DOI: 10.1017/S1366728915000103

Carrol, G., and Conklin, K (2019). Is All Formulaic Language Created Equal? Unpacking the Processing Advantage for Different Types of Formulaic Sequences. *Language and Speech*, first published online Jan 29th 2019. DOI: 10.1177/0023830918823230

Carrol, G., Conklin, K., and Gyllstad, H. (2016). Found in Translation: The influence of L1 on the processing of idioms in L2. *Studies in Second Language Acquisition, 38*(3), 403-443. DOI: 10.1017/s0272263115000492

Carrol, G., Littlemore, J., and Dowens, M. G. (2018). Of false friends and familiar foes: Comparing native and non-native understanding of figurative phrases. *Lingua*, *204*, 21-44. DOI: 10.1016/j.lingua.2017.11.001

Chaffin, R., Morris, R. and Seely, R. (2001). Learning New Word Meanings from Context: A Study of Eye Movements. *Journal of Experimental Psychology Learning Memory and Cognition 27*(1), 225-235.

Cieślicka, A. B. (2006). Literal salience in on-line processing of idiomatic expressions by second language learners. Second Language Research, 22(2), 115-144. DOI: 10.1191/0267658306sr263oa

Cieślicka, A. B. (2011). Suppression of literal meanings in L2 idiom processing: Does context help? *Studies in Second Language Learning and Teaching, 1*(1), 13-36.

Colombo, L. (1993) The comprehension of ambiguous idioms in context. In C. Cacciari C and P. Tabossi (Eds.), *Idioms: Processing, structure, and interpretation.* Hillsdale, NJ: Erlbaum, 163–200.

Columbus, G., Sheikh, N. A., Cote-Lecaldare, M., Hauser, K., Baum, S. R., and Titone, D. (2015). Individual differences in executive control relate to metaphor processing: an eye movement study of sentence reading. *Frontiers in Human Neuroscience, 8*, 1057. DOI: 10.3389/fnhum.2014.01057

Dolgunsöz, E. (2015). Measuring Attention in Second Language Reading Using Eye-tracking: The Case of the Noticing Hypothesis. *Journal of Eye Movement Research*

8(5),4, 1-18.

Fanari, R., Cacciari, C., and Tabossi, P. (2010). The role of idiom length and context in spoken idiom comprehension. *European Journal of Cognitive Psychology, 22*(30), 321-334.

Findlay, H. and Carrol, G. (2019). Contributions of semantic richness to the processing of idioms. *The Mental Lexicon, 13*(3), 311-332. DOI: 10.1075/ml.18014.fin

Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression, 3rd edition*. Sage, Thousand Oaks CA. http://tinyurl.com/carbook.

Frisson, S., and Pickering, M. J. (1999). The Processing of Metonymy: Evidence From Eye Movements. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*(6), 1366-1383. DOI: 10.1037//0278-7393.25.6.1366

Frisson, S., and Pickering, M. J. (2001). Obtaining a Figurative Interpretation of a Word: Support for Underspecification. Metaphor and Symbol, 16(3-4), 149-171. DOI: 10.1080/10926488.2001.9678893

Frisson, S., and Pickering, M. J. (2007). The processing of familiar and novel senses of a word: Why reading Dickens is easy but reading Needham can be hard. *Language and Cognitive Processes, 22*(4), 595-613. DOI: 10.1080/01690960601017013

Gibbs, R. (1986). Skating on thin ice: Literal meaning and understanding idioms in conversation, *Discourse Processes, 9*(1), 17-30. DOI: 10.1080/01638538609544629

Gibbs, R. W., Nayak, N. P., and Cutting, C. (1989). How to Kick the Bucket and Not Decompose: Analyzability and Idiom Processing. *Journal of Memory and Language, 28*, 576-593.

Gildea, P. and Glucksberg, S. (1983). On understanding metaphor: the role of context. *Journal of Verbal Learning and Verbal Behavior, 22*(5), 577-590.

Giora, R. (2003). On our mind: Salience, context and figurative language. Oxford, England: Oxford University Press.

Godfroid, A., Boers, F., and Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye tracking. *Studies in Second Language Acquisition, 35,* 483–517.

Halliday, M. A. K. (2007 [1991]). The Notion of "Context" in Language Education. In J. Webster (Ed.), *Language and Education* (Volume 9 in the Collected Works of MA. A. K. Halliday, pp. 269-290). London/New York: Continuum.

Hamblin, J., and Gibbs, R. (1999). Why You Can't Kick the Bucket as You Slowly Die: Verbs in Idiom Comprehension. *Journal of Psycholinguistic Research, 28*(1), 25-39. DOI: 10.1023/A:1023235403250

Holsinger, E. (2013). Representing Idioms: Syntactic and Contextual Effects on Idiom Processing. *Language and Speech, 56*(3), 373-394. DOI: 10.1177/0023830913484899

Inhoff, A., Lima, S., and Carroll, P. (1984). Contextual effects on metaphor comprehension in reading. *Memory and Cognition, 12*(6), 558-567. DOI: 10.3758/BF03213344

Joseph, H., Wonnacott, E., Forbes, P and Nation, K. (2014). Becoming a written word: Eye movements reveal order of acquisition effects following incidental exposure to new words during silent reading. *Cognition, 133*(1), 238-248.

Katz, A. N., Paivio, A., Marschark, M., and Clark, J. M. (1988). Norms for 204 Literary and 260 Nonliterary Metaphors on 10 Psychological Dimensions. *Metaphor and Symbolic Activity, 3*(4), 191-214. DOI: 10.1207/s15327868ms0304_1

Kecskés, I. (2006). On my mind: Thoughts about salience, context and figurative language from a second language perspective. *Second Language Research, 22*, 2, 219-237. DOI: 10.1191/0267658306sr266ra

Kuznetsova, A., Brockhoff, P.B., and Christensen, R.H.B. (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software, 82*(13), 1–26. DOI: 10.18637/jss.v082.i13

Levorato, M. C., and Cacciari, C. (1995). The Effects of Different Tasks on the Comprehension and Production of Idioms in Children. *Journal of Experimental Child Psychology, 60*, 261-283.

Levorato, M. C., and Cacciari, C. (1999). Idiom Comprehension in Children: Are the Effects of Semantic Analysability and Context Separable? *European Journal of Cognitive Psychology, 11*(1), 51-66. DOI: 10.1080/713752299

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

Liontas, J. (2002). Context and idiom understanding in second languages. In S. Foster-Cohen, T. Ruthenberg, & M.-L. Poschen (Eds.), *EUROSLA Yearbook: Annual conference of the European second language association* (Vol. 2, pp. 155-185). Amsterdam/Philadelphia: John Benjamin Publishing Company.

Lowell, R. and Morris, R. (2014). Word length effects on novel words: Evidence from eye movements. *Attention, Perception, & Psychophysics, 76*(1), 179-189.

Mashal, N. and Faust, M. (2009).Conventionalization of novel metaphors: a shift in hemispheric asymmetry. *Laterality*, *14*, 573–589. DOI: 10.1080/13576500902734645

MacDonald, M. C. and Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. In M. J. Traxler, & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (2nd edition, pp. 581-611). London: Elsevier.

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

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

Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition *from* and *while* reading: An eye-tracking study. *Studies in Second Language Acquisition, 38*, (1)97-130. DOI: 10.1017/S0272263115000224

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Reuterskiöld, C., and Van Lancker Sidtis, D. (2012). Retention of idioms following one-time exposure. *Child Language Teaching and Therapy, 29*(2), 219-231. DOI: 10.1177/0265659012456859

Schweigert, W. (1986). The comprehension of familiar and less familiar idioms. *Journal of Psycholinguistic Research, 15*, 33-45.

Schweigert, W. (1991). The Muddy Waters of Idiom Comprehension. *Journal of Psycholinguistic Research, 20*(4), 305-314.

Schweigert, W., and Moates, D. (1988). Familiar Idiom Comprehension. Journal of Psycholinguistic Research, 17(4), 281-296.

Siyanova-Chanturia, A., Conklin, K., and 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. DOI: 10.1177/0267658310382068

Smolka, E., Rabanus, S., and Rösler, F. (2007). Processing Verbs in German Idioms: Evidence Against the Configuration Hypothesis. *Metaphor and Symbol, 22*(3), 213-231. DOI: 10.1080/10926480701357638

Sprenger, S., la Roi, A. and van Rij, J. (2019). The Development of Idiom Knowledge Across the Lifespan. *Frontiers in Communication, 4*, 29.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 327-342). Oxford: Oxford University Press.

Swinney, D., and Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behaviour, 18*, 523-534.

Tabossi, P., Fanari, R., and Wolf, K. (2009). Why are idioms recognized fast? *Memory and Cognition, 37*(4), 529-540. DOI: 10.3758/MC.37.4.529

Titone, D., and Connine, C. (1999). On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics, 31*, 1655-1674.

Titone, D., and 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. DOI: 10.1075/ml.9.3.05tit

Van Heuven, W.J.B., Mandera, P., Keuleers, E., and Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*, 1176-1190. DOI: 10.1080/17470218.2013.850521

Williams, R. and Morris, R. (2004) Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology, 16*(1-2), 312-339, DOI:

10.1080/09541440340000196

Wray, A., Bell, H., and Jones, K. (2016). How native and non-native speakers of English interpret unfamiliar formulaic sequences. *European Journal of English Studies, 20*(1), 47-63. DOI: 10.1080/13825577.2015.1136163

Zuo, H. (2008). 'Comprehension of Unfamiliar English Idioms by Chinese EFL Learners in Reading.' *Journal of the Chinese English Language Association 31*(3), 9–29.

Xie, H. (2017). Investigating Chinese EFL Learners' Comprehension of English Idioms. *Journal of Language Teaching and Research, 8*(2), 329-336.

Appendix

Stimuli (figurative expressions / literal paraphrases) used in both experiments (Experiment 1: L1 idioms, L2 idioms and metaphors; Experiment 2: L1 idioms and L2 idioms)

Figurative expression	Literal paraphrase	Type
Bite my tongue	Not speak freely	L1 Idiom
Changed his tune	Altered his opinion	L1 Idiom
Dropped the ball	Made a bad error	L1 Idiom
Eat his words	Say he was wrong	L1 Idiom
Find his feet	Get used to it	L1 Idiom
Hit the roof	Get furious	L1 Idiom
Jumped the gun	Started too soon	L1 Idiom
Let off steam	Release tension	L1 Idiom
Over the moon	Extremely happy	L1 Idiom
Playing with fire	Taking big risks	L1 Idiom
Popped the question	Proposed to her	L1 Idiom
Pulling their leg	Playing a trick	L1 Idiom
Rings a bell	Sounds familiar	L1 Idiom
Smell a rat	Am suspicious	L1 Idiom
Spilled the beans	Let out the secret	L1 Idiom
Stole the show	Was the best one	L1 Idiom
Tightened my belt	Reduced my spending	L1 Idiom
Twist her arm	Persuade her	L1 Idiom
Under the weather	Feeling a bit unwell	L1 Idiom
Walking on air	Really happy	L1 Idiom
A duck drank my mind	I feel really silly	L2 Idiom
A horse doesn't stop its hooves	Things carry on without stopping	L2 Idiom
Add oil and vinegar	Exaggerate stories	L2 Idiom
Beat the grass and scare the snake	Alert them by acting too hastily	L2 Idiom
Bite into the sour apple	Do some unpleasant things	L2 Idiom
Bought it for an apple and an egg	Got it for a really cheap price	L2 Idiom
Brought sticks to put out a fire	Made the whole situation worse	L2 Idiom
Calls a deer a horse	Misrepresents things	L2 Idiom
Chase the wind and grasping at shadows	Make accusations without any real proof	L2 Idiom
Chicken feathers and garlic skins	Lots of small and trivial things	L2 Idiom
Cold coffee	News I know	L2 Idiom
Covering his ears to steal a bell	Not being very honest with himself	L2 Idiom
Different mouths but one sound	Everyone saying the same thing	L2 Idiom
Draw a snake and add feet	Ruin it by being too fussy	L2 Idiom
Earns a pig's money	Earns a lot of money	L2 Idiom
Eyes bright like torches	Alert and really focused	L2 Idiom
Gave him a basket	Wasn't interested	L2 Idiom

Gives the word to everyone	Invites everyone to speak	L2 Idiom
Has hair on his teeth	Is very easily annoyed	L2 Idiom
He is naked water	He is unqualified	L2 Idiom
He lives five for four	He lives quite recklessly	L2 Idiom
He's missed his first seven years	He's been really badly raised	L2 Idiom
I stepped on the lion's tail	I angered someone dangerous	L2 Idiom
I'm gathering my hammers	I'm leaving this place	L2 Idiom
It came out salty	I'm paying too much	L2 Idiom
Leave the church in the village	Not exaggerate about everything	L2 Idiom
Neither three nor four	Not really trustworthy	L2 Idiom
One gun and a horse	Without any company	L2 Idiom
Pour them clear wine	Show the true situation	L2 Idiom
Seven hands and eight feet	Too many people involved	L2 Idiom
She can spoon the soup out	She can solve her own problems	L2 Idiom
She's a big stick	She's important	L2 Idiom
Stepped into the grease bowl	Really embarrassed himself	L2 Idiom
Talking into the blue	Talking quite aimlessly	L2 Idiom
Three long and two short	Unexpected disasters	L2 Idiom
Trick the sky to cross the sea	Use extremely devious methods	L2 Idiom
White clouds change into grey dogs	Life changes in unpredictable ways	L2 Idiom
Wine and meat friends	Friends when it suits	L2 Idiom
Without shirt or shoes	Very scruffy and untidy	L2 Idiom
Won't share the same sky	Really hate each other	L2 Idiom
A blackmailer is a leech	Blackmailer saps your life	Metaphor
A degree is a doorway	Education is opportunity	Metaphor
A friend is a ray of sunshine	A friend makes things brighter	Metaphor
A good lover is a teddy bear	A partner is a great comfort	Metaphor
A mind is a sponge	Minds take lots in	Metaphor
A museum is a history book	You can learn from museums	Metaphor
A rumour is a plague	Rumours spread quickly	Metaphor
A smile is a knife	A smile can be nasty	Metaphor
Alcohol is a crutch	They drink for support	Metaphor
An accountant is a juggler	Accountants balance things	Metaphor
Anger is a storm	Anger is unstable	Metaphor
Books are treasure chests	Books contain precious things	Metaphor
Danger is a spice	Risk can be exciting	Metaphor
Discipline is a fertilizer	Self-control helps you grow	Metaphor
Education is a lantern	Learning helps you to see	Metaphor
Hard work is a ladder	Hard work is beneficial	Metaphor
History is a mirror	The past shows things	Metaphor
Humour is a medicine	Humour improves things	Metaphor
Money is a lubricant	Money gets things done	Metaphor
Time is a doctor	Time improves things	Metaphor