Language and Cognition (2025), 17, e72, 1–30
doi:10.1017/langcog.2025.10029

ARTICLE

High ceilings and ingenuine allies: tapping into the idiom meaning knowledge of first and second language speakers of English

David O'Reilly[1], Alexander Onysko[2], Carina Rasse[2], Lisa Papitsch[2], Herbert Colston[3] and Iris van der Horst[2]

[1]Centre for Advanced Studies in Language and Education (CASLE), University of York, York, UK; [2]Department of English, University of Klagenfurt, Klagenfurt, Austria and [3]Department of Linguistics, University of Alberta, Edmonton, AB, Canada
Corresponding author: David O'Reilly; Email: david.oreilly@york.ac.uk

(Received 01 November 2024; Revised 15 August 2025; Accepted 19 August 2025)

Abstract
Idioms are undoubtedly important for second language (L2) learners, who encounter them in instructed learning, textbooks/resources and in out-of-class language use. While research on first language (L1) and L2 idiom comprehension shows how well L1/L2 speakers understand various idioms and the role of different predictors, important questions remain about how knowledge varies with more difficult task types and stimuli, how well L1 'norms' serve L2 learners, how subjective and objective predictors of idiom knowledge interact and how L2 learner inferencing works in learning idioms. To address these issues, university-age L1 and L2 English (L1 German) participants provided meaning descriptions and familiarity ratings for 100 challenging idioms from learner resources, and each idiom was assigned an OpenAI-generated transparency rating, corpus-based frequency and to one of six cross-language overlap (CLO) types. Descriptive statistics showed lower and more varied idiom meaning knowledge than might be expected, especially for the L1ers, who were some way off ceiling level. Mixed-effects regression revealed familiarity and transparency as positive L1 and L2 knowledge predictors, but groups differed in sensitivity to idiom frequency, which only mattered for the L1ers and CLO, which (as expected) only mattered for the L2ers, who mistook false friends as genuine allies.

Keywords: idioms; knowledge; L1/L2; meaning descriptions; meaning knowledge; metaphors; transfer

1. Introduction

Idioms such as *over the hill* and *to be/stay on the ball* are conventionalised phrases conveying figurative meaning overall or in at least one element (Aydin, 2019; Carrol et al., 2018; Hubers et al., 2020). For second language (L2) learners, idiom mastery increases perceived fluency (Beck & Weber, 2016; Park & Chon, 2019), reduces cognitive load/processing (Beck & Weber, 2016; Conklin & Schmitt, 2008; Pawley & Syder, 1983; Siyanova-Chanturia, 2013) and enriches engagement with discourse communities. Consequently, idiom competence is widely recognised in language proficiency frameworks underpinning high-stakes testing (e.g., The Common European Framework of Reference for Languages [CEFR]). While first language (L1) idiom competence is generally well-developed by early adolescence, maturing alongside broader inferencing and semantic analysis skills (Carrol et al., 2018), L2ers are often unaware of idioms (Kim, 2016), do not understand them or misunderstand them (Littlemore, 2001).

To further this line of research, the current study provides new quantitative and qualitative evidence on L1 and L2 English idiom meaning recall/inferencing,[1] focusing on underexplored interactions between four key predictors of idiom comprehension: subjective familiarity and objective transparency, frequency and cross-language overlap (CLO). The study has two main theoretical contributions. First, findings related to the objective transparency predictor add to our understanding on the parameters around the particular importance of this aspect for L2 (compared with L1) speakers, who, as the *Literal Salience Hypothesis* (Cieślicka, 2006, 2015) suggests, should be more prone to a literal, word-by-word analysis of idioms (see *throw the baby out with the bathwater* example in Section 2.2). Second, findings related to the CLO predictor expand Soto-Sierra and Ferreira’s (2024) recent application of the *Parasitic Strategy* (Hall, 2002) to L2 idiom learning, showing how, in our case, both participant groups exploit pre-existing lexical and conceptual knowledge when inferencing unfamiliar idioms, the issues they encounter and the apparent entrenchment of mislearned idiom meanings.

As part of a larger educational project to develop conceptual metaphor-related visual aids for L1 German and L1 Russian L2 English learners, the findings also contribute pre-intervention, baseline data on how well L1/L2ers know 100 challenging metaphorical idioms from learner resources. To motivate the research questions, we first contextualise idioms within broader vocabulary knowledge, delineate recall from other knowledge types and examine research on L2 idiom inferencing and factors influencing comprehension.

2. Background

2.1. Idiom knowledge(s) in the L1 and L2

Idiom knowledge sits within broader vocabulary (or lexical) knowledge, described, variably, as receptive and productive mastery of multiple word knowledge components (González-Fernández & Schmitt, 2020) and different knowledge continua (e.g., partial knowledge such as being able to pronounce a written word, to more precise knowledge such as being able to identify common associate words), which language users have variable real-time control over (Henriksen, 1999; Laufer & Goldstein, 2004).

In recent studies, González-Fernández and Schmitt (2020) and González-Fernández (2022) found that mixed-proficiency Spanish, and in the later study Chinese,

[1] Generally, we use ‘recall’ as a catch-all term for the ability to produce an idiom’s meaning when presented with its form. For clarity, in Section 4.3, we indicate the point at which we differentiate the terms ‘recall’ and ‘inferencing’ to denote knowledge of familiar and unfamiliar idioms, respectively.

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

learners of English displayed recognition mastery for all eight components tested before recall mastery of any component, and a generally consistent acquisition order. The easiest component was Form–Meaning link meaning recognition via multiple choice, for example, in ‘it is the best season’ recognising that the underlined word form has the meaning *time*, not *animal*, *appearance* or *place*, when these are presented as different options. The most difficult was Multiple-Meanings recall, requiring participants to use either the L1 or L2 to supply three meanings per target word (synonyms, translation and/or descriptions were accepted), when given word classes and hints. For example, for the stimulus ‘season’, hints included ‘(Noun = year)_____, (Verb = cooking)_____, (Noun = animals in season)_____.’ These findings are important because they extend the previously established recognition-before-recall developmental trajectory from form-meaning link knowledge (Laufer & Goldstein, 2004) to at least three other components of Nation’s (2013) word knowledge framework: (1) collocations – words that commonly occur together (e.g., for recall, supplying ‘promise’ when given ‘fulfill his p_____’); (2) derivatives – different version of a word according to its class (e.g., for recall, providing ‘seasonally’ when given the stimulus ‘season’ and ‘Adverb – In this country, the temperature variations occur _____’) and (3) multiple meanings (see above, this paragraph). While idiom knowledge might be expected to show the same trajectory, it did not receive focus in these studies.

In the empirical L1/L2 idiom knowledge literature, several themes emerge that help contextualise the current study’s measures, although neat comparisons are complicated by varying designs.

First, L1ers do not perform at as high a ceiling level as with other figurative multi-word units (e.g., phrasal verbs, O’Reilly, 2017). For example, in a multiple-choice meaning recognition test involving 110 idioms selected from a norming database of 393, Hubers et al.’s (2020) L1 Dutch speakers averaged 88% (standard deviation [*SD*] = 18, hereafter in parenthesis) compared with 62% (26) for moderate-to-high proficiency L1 German learners of Dutch. Carrol et al.’s (2018) L1 English group averaged 88% (31) multiple-choice meaning recognition of 22 high-familiarity English idioms, compared with 73% (44) for a mixed L1 ESL (English as a Second Language) group, and 47% (50) for an L1 Chinese EFL (English as a Foreign Language) group. Idiom recall appears to be even harder than recognition (as implied in González-Fernández & Schmitt, 2020 and in González-Fernández, 2022). In Guo and Xiang (2023), mixed proficiency L1 Chinese EFL learners averaged 55% (15) meaning recall for 120 low-familiarity L1 idioms, compared with 50% (19) and 31–36% (15–41) for high- and low-familiarity L2 idioms. More recently, in a study with 18–80-year-old British English speakers, Carrol (2023) found the equivalent of 90.5% (23) idiom familiarity, highest for older participants, who also had fewer unknown idioms or mismatches between perceived and actual knowledge. While educational level played some role, age and vocabulary knowledge combined were superior predictors of idiom knowledge measured in this way.

Second, both L1 and L2 speakers show considerable *variation* in idiom knowledge. Considering that under a normal distribution, 68.2% of scores lie one standard deviation above and below the mean, L1 standard deviations of 31 (Carrol et al., 2018) and even 18 (Hubers et al., 2020) seem large compared with other figurative multi-word units (cf. O’Reilly, 2017). Confusingly, while L2 dispersion is probably typically higher (e.g., 43, linked to an upper-intermediate L1 Spanish EFL learner mean of 76% in Soto-Sierra & Ferreira, 2024), standard deviations equivalent to ≤17%

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

(i.e., less than the L1ers in Carrol et al., 2018 and Hubers et al., 2020) were found with meaning recognition and form recall measures from EFL learners in Japanese (Vasiljevic, 2016) and Dutch (Boers et al., 2007) contexts, discrepancies which likely arise from differing study designs and stimuli, suggesting more within-study comparisons are needed.

Third, idiom knowledge studies have focused more on *how well*, than *how*, L1/L2ers interpret idiom meaning, seldom combining quantitative and qualitative approaches. In the current study, we incorporate these two elements, focusing on meaning recall, a highly informative knowledge type that yields qualitatively richer data on learner comprehensions than meaning recognition (cf. O'Reilly & Marsden, 2021; O'Reilly & Yan, 2025).

2.2. Knowledge and inferencing

When inferencing unfamiliar idiom meanings, L2ers seem strongly attracted to contextual information when available, which could include context from surrounding discourse and/or the scenario in which the idiom is encountered (who said it, the location and purpose of the interaction etc.). For example, Park and Chon's (2019) L1 Korean adolescent EFL learners most frequently reported guessing meanings from discourse context (see also Aljabri, 2013), then using background knowledge, literal translation and least frequently, their L1. Only the first two strategies were significant, positive meaning recognition predictors for these learners; the latter two were generally detrimental to inferencing across hierarchical regression models with predictors entered one at a time. However, this finding may be partly due to the study design since idiom stimuli were selected to be unfamiliar to participants, contain high frequency individual constituent words and balance out prepositional idioms (e.g., 'by the way') and verbal idioms (e.g., 'look after'), the most common types encountered by Korean middle school learners encounter in their pedagogic materials (Park & Chon, 2019, p. 224). As such, item variation in terms of high versus low amounts of background knowledge required for deciphering, high versus low figurative-literal correspondence and the degree of L1–L2 equivalence was not directly considered (see Section 2.3).

At the theoretical level, the *Parasitic Strategy* (also referred to as the *Parasitic Hypothesis*, *Model*, *Vocabulary, Lexicon* and *View*)[2] attempts to provide an account of how known words affect words being learned, claiming 'that on initial exposure to a word, learners automatically exploit existing lexical material in the L1 or L2 in order to establish an initial memory representation' (Hall, 2002, p. 69). Rooted in connectionist models of the lexicon (in the SLA context, see e.g., Ellis, 1998; MacWhinney, 1997), the *Parasitic Strategy* suggests that as L2 proficiency increases, stronger and more direct L2 form-meaning links and conceptual representations then develop (Soto-Sierra & Ferreira, 2024). In setting out the model, Hall (2002, pp. 73–74) focuses on: 'true cognates' – lexical items common to the L1 and L2 despite minor

[2] In the current study, we typically refer to this phenomenon as the *Parasitic Strategy* to emphasise the strategic purpose served by connecting new lexical material to established mental representations. In our view, the parasite metaphor discussed in the literature offers an evocative and memorable heuristic for conceptualising the linguistic and cognitive mechanisms at work. However, to state the obvious, it is not (and should not be) understood as intended to frame language users themselves as parasites.

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

orthographic and/or phonological differences, emerging from shared linguistic history and/or borrowing (e.g., *rose* [English] and *rosa* [‘rose’ in Spanish]); and two kinds of ‘false cognates’: ‘true false cognates’ – historically unconnected but coincidentally similar lexical items (e.g., *tuna*, in English a fish, in Spanish a prickly pear) and ‘indirect cognates’ – partially semantically overlapping lexical items with divergent (and therefore deceptive) meanings (e.g., *library* [English] vs *librería* [‘bookshop’ in Spanish]).

In support of the *Parasitic Strategy*, Hall (2002) reports an experiment involving single-word pseudocognates (English nonwords designed to resemble real Spanish words), which L1 Spanish L2 English learners perceived as more familiar than other nonwords designed with no overlap. More recently, Soto-Sierra and Ferreira (2024) applied the *Parasitic Strategy* to idiom learning, interpreting their L2 learners’ reliance on literal idiom meanings as most consistent with this theoretical view and also with the *Literal Salience Hypothesis* (Cieślicka, 2006), which posits that L2 learners conduct a literal, word-by-word analysis of new idioms by default, compared with L1 speakers, for whom figurative idiom meanings are more familiar and salient (Carrol et al., 2018).[3] However, except for a brief paragraph in Hall’s (2002, p. 81) article discussing how three pseudocognates were mistaken for other, real, L2 words (e.g., **gan* mistaken for the real English word ‘gun’, shown in the responses *pistola* [‘pistol’] and *arma* [‘weapon’]), the language forms and concepts that L2 learners latch on to in their *Parasitic Strategy* were not considered in these studies.

When contextual information is absent, as in the current study, L2ers more frequently rely on semantic analysis than form/meaning retrieval, homing in on phrase constituents (Carrol et al., 2018; Wray et al., 2016). For example, when confronted with an unknown idiom *throw the baby out with the bathwater* (meaning ‘discard something valuable along with an undesirable thing’), an L2 learner may arrive at the correct meaning by analysing the individual words and drawing an analogy (a baby represents something desirable to be kept, used bathwater represents something undesirable to be discarded, throwing out one with the other represents a mistake arising from a failure to separate the desirable and undesirable etc.). Research by Skoufaki (2008), involving advanced L2 English learners, found increased idiom transparency (clarity of overall meaning from constituent words) related to a decreased range of different meaning interpretations, suggesting intuitions are shaped by both constituents and overall familiarity. Once transparency is accounted for, L1/L2 speakers have comparable inferencing skills, higher L1 vocabulary and cultural knowledge offering some advantage (Carrol et al., 2018). In addition, when L1ers believe an idiom has a particular meaning, whether correctly or incorrectly, they map its meaning elements onto the idiom constituents, increasing perceived transparency and precluding openness to other meanings (Keysar & Bly, 1995, 1999).

Other research suggests that learners can be trained, to some extent, in associative thinking (Littlemore, 2008) for successful inferencing. For instance, Boers et al. (2007, p. 58) showed that providing etymological information (e.g., ‘jump the gun [act too

[3] Soto-Sierra and Ferreira’s (2024) findings also lend partial support to a further model, the *Dual Idiom Representation* (DIR) model (Abel, 2003), which is not discussed here because we did not operationalise decomposability as distinct from transparency (see Section 2.3) and because the DIR’s claim that non-decomposable idioms (e.g., *kick the bucket*, with a meaning indecipherable from individual words) require a single, idiom entry in the mental lexicon cannot be true for new idiom encounters (for an L2-specific refinement of this model, see the *Heuristic* approach discussed in Liu, 2008).

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

soon]' originating from a false start in athletics, before hearing the starting pistol) before (rather than after) revealing idiom meaning improved Dutch university English majors' recall and that guessing using etymological information and sentence examples was preferable to just the latter or idioms in isolation. Similarly, Wang et al. (2024) observed that better initial stage guessing, from idioms presented first without and then with etymological information, predicted better recall for Chinese university English majors 1 week later and that errors lingered despite feedback. Pedagogically, these findings suggest benefits for developing inferencing strategies over reliance on feedback.

2.3. Factors influencing idiom comprehension

In addition to knowledge type and discourse context, numerous other factors shape idiom comprehension. Here, we unpack the role of familiarity, transparency, frequency and CLO (for the current study's operationalisations, see Sections 3.2 and 3.3) and other factors as they relate to these variables (for more detailed overviews, see Beck & Weber, 2016; Hubers et al., 2020; Kim, 2016; Soto-Sierra & Ferreira, 2024).

Idiom familiarity can denote perceived encounters with spoken or written form, regardless of meaning knowledge (Bonin et al., 2013; Soto-Sierra & Ferreira, 2024) and/or perceived meaning knowledge. The two (form/meaning) are sometimes considered different familiarity types, or the latter is viewed as a separate construct, meaningfulness (Beck & Weber, 2016; Bulkes & Tanner, 2016). Alternatively, Hubers et al. (2020, p. 2) describe self-perceived frequency of encounter as 'idiom frequency' and familiarity with an expression's meaning as 'idiom familiarity,' while others (e.g., Aljabri, 2013; Nippold & Taylor, 2002) view idiom familiarity as objective frequency of occurrence in the language.

Idiom transparency refers to how clear an idiom's overall figurative meaning is from its parts, with participants either also given the correct meaning to work with (e.g., Hubers et al., 2020) or not (e.g., Soto-Sierra & Ferreira, 2024). Since idioms appear more transparent once learned (Carrol et al., 2018; Keysar & Bly, 1995, 1999), perceived transparency in the subjective sense, which differs from person to person, needs distinguishing from true conceptual transparency in the more objective sense, i.e., as an idiom-inherent property. Transparency has also been described as metaphorical motivation (Cieślicka, 2015), figurative meaning (Skoufaki, 2008) and semantic value (Steinel et al., 2007) and used interchangeably with analysability and decomposability (Carrol et al., 2018) although in other studies there is an attempt to tease these variables apart (e.g., Bulkes & Tanner, 2016).

Idiom frequency, in the objective sense, means commonality within language, typically established using a corpus. Here, estimates can be derived from: basic phrasal frequencies of an idiom's linguistic variations (e.g., *take/took/taking* [etc.] *him/her/someone* [etc.] *for a long/short/wild* [etc.] ride); idiom constituent noun and verb frequencies (e.g., Libben & Titone, 2008; Soto-Sierra & Ferreira, 2024) sometimes combined with morphosyntactic pattern matching to try and delineate figurative usage (e.g., Hubers et al., 2020) or, as the sum of constituent frequencies divided by their number, an approach dating back to Kucera and Francis (1967) used in Bonin et al. (2013) and Cronk et al. (1993). Each has its advantages and disadvantages. The first maximises valid hits, but capturing creative modifications and rejecting literal meanings is labour-intensive. The second and third offer efficient

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

searching but risk imprecise estimates that are constituent (rather than phrase) focused.

CLO denotes how closely idiom forms and meanings correspond between languages, either objectively, where idioms are pre-assigned to different categories (as in the current study), or subjectively, where learners' self-perceived CLO is measured. In recent studies (e.g., Hubers et al., 2020; Soto-Sierra & Ferreira, 2024), Titone et al.'s (2015) rating system has informed four CLO types: (1) no L1 equivalent; (2) L1 equivalent exists with completely different content words; (3) L1 equivalent exists with some content words in common; (4) L1 equivalent exists with word-to-word equivalence. While CLO is essentially an L2-specific variable (Soto-Sierra & Ferreira, 2024), it is often also modelled as an L1 idiom knowledge predictor, where it is not expected to show any effect, in turn more reliably establishing its effect for L2ers (e.g., Carrol et al., 2016, 2018).

Undoubtedly, familiarity, frequency, transparency and CLO are important for idiom comprehension. In a study on eight predictors of L1 Spanish EFL learners' idiom meaning recognition via multiple choice, Soto-Sierra and Ferreira (2024) argue that familiarity generally matters for comprehension, and transparency when idioms are less familiar, such as theirs were. Similarly, with L1 German learners of Dutch, Hubers et al. (2020) found L2-rated transparency was a key meaning recognition predictor when idioms were unfamiliar. However, while Hubers et al. (2020) operationalised L2 participants' own intuitions about transparency and other variables (e.g., familiarity, imageability), in Soto-Sierra and Ferreira (2024) the superiority of transparency as a knowledge predictor may partly have arisen because only this variable was operationalised using L2 participant intuitions, others (e.g., familiarity, meaningfulness, literal plausibility) were constructed from published L1 norms.

Similarly, in Carrol et al. (2018) idiom meaning recognition was also operationalised via four-option multiple choice but familiarity (not transparency) was the key knowledge predictor for L1ers, and to a lesser extent, mixed L1 and L1 Chinese L2 English groups. In line with earlier research (Keysar & Bly, 1995, 1999), the study also showed a positive familiarity–transparency relationship, suggesting the importance of testing for interactions. While interactions between the various subjective measures mentioned above have been modelled, there is scope for further enquiry regarding subjective and objective measures.

In the studies cited so far, objective frequency seems to matter somewhat for L1ers and little for L2ers (Hubers et al., 2020; Soto-Sierra & Ferreira, 2024), as usage-based models predict (Carrol et al., 2018; Tomasello, 2003). Even for L1ers though, it can be a highly variable knowledge predictor due to the low frequency of many idioms, whether perceived as familiar or unfamiliar.

Across the literature, CLO plays a positive role, with increased L1–L2 equivalence predicting higher offline meaning recognition (Charteris-Black, 2002; Hubers et al., 2020; Kainulainen, 2006; Soto-Sierra & Ferreira, 2024), meaning recall (Deignan et al., 1997) and form recall (e.g., Charteris-Black, 2002; Laufer, 2000) and smoother online processing (e.g., Carrol & Conklin, 2014, 2017; Carrol et al., 2016; cf. Beck & Weber, 2016). Hubers et al.'s (2020) CLO effects are noteworthy, since L2ers confidently drew on their L1 for exact equivalents but not for partial or non-equivalents. In Soto-Sierra and Ferreira (2024), partial—/non-equivalents were not compared in this way, so whether they corroborate or contradict this interpretation is unclear. Moreover, qualitative inspections of response data were

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

not presented in these studies but could shed valuable light on L2 approaches to meaning recognition/recall.

2.4. Research questions (RQs)

While L1 idiom knowledge is often tacitly assumed as a reliable benchmark for L2ers, L1ers seem to have less mastery and higher variation than might be expected, even for the easiest knowledge type (meaning recognition) with highly familiar idioms. Further evidence is needed on more difficult L1 and L2 knowledge types (e.g., meaning recall) and norms for more difficult idiom sets, especially if used in pedagogy.

Second, the need to consider the unique and interactive effects of different idiom knowledge predictors has been increasingly recognised (Carrol et al., 2018; Soto-Sierra & Ferreira, 2024). While recent studies have modelled familiarity interacting with transparency (Carrol et al., 2018; Hubers et al., 2020; Soto-Sierra & Ferreira, 2024) and imageability (Hubers et al., 2020) and objective idiom frequency with decomposability (Soto-Sierra & Ferreira, 2024), interactions between familiarity and objective frequency and involving CLO remain unexplored (cf. Carrol et al., 2018). Furthermore, more data is needed on objective, learner-independent, transparency to complement what is known about subjective, perceived transparency and its interactions.

Third, exploring further CLO types that L2ers encounter (e.g., different conceptual metaphor correspondences, false friends), beyond the basic four from recent research, would be useful for helping further inform theoretical perspectives on how pre-existing lexical and conceptual knowledge shapes language users’ understandings of unfamiliar idiomatic language and for practice. Here, combined quantitative and qualitative methods can offer a better comprehensive understanding of knowledge, inferencing and for L2 learners, the role of the L1.

We address these issues via three research questions (RQs):

1. How do L1 and L2 English speakers compare in their recall/inferencing of the meanings of 100 challenging idioms found in learner resources?
2. To what extent is idiom meaning recall/inferencing by L1 and L2 English speakers explained by subjective idiom familiarity and objective transparency, frequency and CLO?
3. For L2 English speakers, which themes characterise their interpretations of idioms of different CLO types?

Given the number of available participants and considerations of model sensitivity, we took a parsimonious approach and selected familiarity as a single subjective measure for quantitative investigation alongside objective transparency, frequency and CLO, leaving questions about the plethora of other subjective measures and possible interactions for further research (see Section 6).

3. Methods

All materials, data, analyses and further methodological information are available on the study’s Open Science Framework (OSF) page: https://osf.io/b69pf/.

3.1. Participants

L1 English participants were 94 Canadian university students (68 female, 25 male, one 'prefer not to say') aged around 21 (*M* = 20.86, *SD* = 5.35) studying various individual or combined subjects. All spoke L1 English, with 12 reporting additional L1s (three Cantonese; two Bengali, French and Punjabi; one Hindi, Korean and Tamil). At least some knowledge of one or more of 22 L2s was reported. L1ers completed the idioms task in January 2023. Data from one further participant, whose responses showed evidence of dictionary use, were not retained.

L2 English participants were 88 advanced proficiency Austrian university students (66 female, 19 male, one diverse, two 'prefer not to say') aged around 24 (*M* = 23.51, *SD* = 4.96). All studied English, and overall, many additional subjects. All were L1 German speakers, with 8 additional L1s (Arabic, Bosnian, English, French, Greek, Romanian, Serbian, Serbo-Croatian) and 13 additional second languages reported. L2ers completed the idioms task between November 2022 and January 2023. Data from a further 11, non-L1 German participants were not retained.

3.2. Data collection instrument

The idioms task comprised 100 idioms presented using six lists (three lists each with two random presentation orders, A and B) to balance item coverage and participant fatigue. Participants were each assigned to one list (Table 1).

The task was administered online via Google Forms (https://www.google.co.uk/forms/about/) for all participants except L1ers on Linguistics programmes with a research participant requirement, who completed tasks via Sona Systems (https://www.sona-systems.com). L1ers worked in their own time and L2ers during class. In each list, participants were provided with the bare idiom form and asked:

- Q1. Do you know this idiom? (Yes/No)
- Q2. If you know the idiom, please write down its meaning. If you do not know the idiom, try to guess what it might mean.

3.3. Stimuli, outcome measure, and predictors

3.3.1. Stimuli

Idioms were selected from a pool of close to 600 from the following sources: Cambridge Idioms Dictionary (2006), Collins Cobuild Idioms Dictionary (2020),

Table 1. Assignment of 100 idioms and 182 participants to three paired lists

Idiom lists (idioms within list)[a]	*k* idioms in list (total 100)	*n* participants assigned to list L1 English (total 94)	 L2 English (total 88)[b]
1A (1–33)	33	14	15
1B (1–33)	33	15	15
2A (34–66)	33	16	15
2B (34–66)	33	16	15
3A (67–100)	34	17	14
3B (67–100)	34	16	16

[a] Idioms 1–100, A lists presented idioms in ascending order, B lists in a different, random order.
[b] One participant (A10) misunderstood their assignment and completed lists 1A, 2A and 3A. Since there were no indications of invalid responses, with A10's permission, we retained all their data.

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

Idioms Organiser (Wright 1999), Oxford Idioms Dictionary for Learners of English (Parkinson & Francis 2006) and Oxford Dictionary of Idioms (Ayto 2020). Selection was geared towards: (a) identifying idioms likely to be difficult for advanced L2ers and (b) with English idiomatic meaning based on conceptual metaphor(s). After conceptual analysis of the idiom pool, the research team at the University of Klagenfurt established a core list of 100 idioms to allow the testing of varying levels of English-German lexical/metaphorical equivalence (see CLO predictor below). We presented idioms as bare forms rather than reformulating using a carrier phrase and aimed not to introduce any subtle clues that may have helped or hindered meaning recall/inferencing.

3.3.2. Idiom meaning recall (outcome)

L1 idiom meaning explanations (see Section 3.2) were scored by all six authors and L2 explanations by the four authors at the University of Klagenfurt. Scorers worked through the responses idiom-by-idiom and independently, but could see one another's emerging scoring in the data file, assigning either 2 (correct), 1 (partially correct) or 0 (incorrect). At regular intervals, scoring teams discussed discrepancies, which, to reflect the 'fuzziness' and challenges of pinpointing certain meanings, were permitted to stand with rationale. Mean and percentage scores, used in the main analyses, were calculated from all available ratings. Interrater reliability analyses of cleaned data (see below) showed high levels of scorer agreement (L1: 84% 'almost perfect', Fleiss Kappa = .87; L2: 97% 'almost perfect', Fleiss Kappa = .95, Landis & Koch, 1977).

3.3.3. Idiom familiarity (predictor)

Given the idiom familiarity question phrasing (see Section 3.2), we take this measure to reflect subjective, self-perceived meaning knowledge, and not simply encounters with the form regardless of knowledge, i.e., a malleable, speaker-specific variable (e.g., as in Carrol et al., 2018) rather than a fixed, norm-indexed property (e.g., as in Soto-Sierra & Ferreira, 2024). A dichotomous response format was employed in favour of a wider (e.g., 7-point) scale because it minimised participant burden/fatigue, promoting more sustained engagement. The intraclass correlation coefficients (see OSF document 'RQ1_Descriptive statistics') showed that the reliability of familiarity ratings was high for both the L1 participants (.95) and L2 participants (.87).

3.3.4. Idiom transparency (predictor)

Since relevant norming data were unavailable for all 100 idioms, we operationalised this predictor by asking ChatGPT-4o, a generative artificial intelligence chatbot based on the GPT-4o large language model (LLM) (OpenAI, 2025), to rate each idiom's 'transparency' from 1 to 5 (very unclear to very clear) using Hubers et al.'s (2020, p. 6) definition 'how clear the meaning of this expression is based on the individual words in the expression.' ChatGPT-4o was not informed about the current study and its variables, and it is important to note that we operationalised and modelled the transparency predictor during the article revision stage in response to reviewer suggestions to provide a more systematic investigation here (originally, we had considered transparency only in a qualitative manner as part of the RQ3 analysis). A 'large' strength correlation (Cohen, 1988) was observed between ratings obtained from two independent interrogations conducted several weeks apart ($r = .70$ [.59, .79], $r_s = 0.64$ [.51, .74)]), in line with recent research showing the test retest

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

stability of Open-AI, in particular GPT-4o and other larger models, in rating figurative language properties (Mangiaterra et al., 2025).

3.3.5. Idiom frequency (predictor)

To estimate each idiom's usage frequency, we used English Web corpus 2020 (enTenTen20) of 36 billion words with genre annotation and topic classification (Jakubíček et al., 2013) in Sketch Engine (Kilgarriff et al., 2014). To maximise variant coverage, we searched relevant lemmas, morphemes and possible word spacings (e.g., *walk* * on air* yielding *walk/ed/s/ing*[etc.] *home/high*[etc.] *on air*) until total hits fell below 100 or idiomatic uses below 10. For each idiom variant, idiomatic usage was estimated by multiplying total hits by the proportion of idiomatic usages manually counted in the first 200 hits (or less, for totals <200). Final variant estimates were then summed to obtain a single idiomatic usage frequency per idiom.

3.3.6. CLO (predictor)

The 100 English idioms fit the following CLO types (*k* idioms):

1. lexically and metaphorically similar (15)
2. lexically different, metaphorically similar (31)
3. neither lexically nor metaphorically similar (but a German idiomatic expression denoting this meaning exists) (34)
4. no idiomatic equivalent for the English idiom exists in German (10)
5. lexically different and German idiomatic equivalent is not based on metaphor (5)
6. false friend (5)

For example, the English idiom *walking on air* was indexed as follows:

- Meaning (English) = 'to feel extremely excited or happy'
- Conceptual metaphor (English) = BEING HAPPY IS BEING UP IN THE AIR
- German idiomatic equivalent = *auf Wolken schweben* ['floating on clouds', lit. on clouds floating]"
- Conceptual metaphor (German) = BEING HAPPY IS BEING UP IN THE AIR
- CLO type 2: lexically different, metaphorically similar.

In contrast to previous studies involving meaning recognition (Hubers et al., 2020), processing (Carrol et al., 2018; Titone et al., 2015) and expected acquisition difficulties (Charteris-Black, 2002), this CLO classification distinguished lexical, metaphorical and semantic aspects. Among the more fine-grained classifications so far, Hubers et al.'s (2020) reflects formal L1 to L2 equivalence, while Charteris-Black's (2002) focuses on the nexus of linguistic form and conceptual basis. Our six CLO types encompass Charteris-Black's lexical/conceptual metaphor overlap, as well as cases of non-idiomatic, non-metaphoric (e.g., metonymy-based) and false friend equivalence. At the phrase level, false friend idioms are generally akin to (true) false cognates (Hall, 2002) in the sense that English and German expressions have arisen from common human experiences (e.g., literally climbing a hill), rather than via common linguistic routes, yielding formally similar idioms with semantic aspects that serve different metaphorical comparisons. For example, in the English idiom

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

over the hill (meaning 'becoming old/obsolete/tired'), the fatigue of climbing a hill is relevant, whereas in the German idiom *über den Berg* (meaning 'overcoming the worst' [lit. over the hill/mountain]) the accomplishment is relevant.

Given the available exemplars in the 100 idioms, it was preferable to use a general 'lexically and metaphorically similar' category for 'word-for-word equivalence' and 'some common content word equivalence' categories from previous studies and a 'lexically different, metaphorically similar' category for the 'L2 equivalent with completely different content words' category from Hubers et al. (2020) and Soto-Sierra and Ferreira (2024). Idioms within the CLO type 1 'lexically and metaphorically similar' category generally align with Hall's (2002) notion of true cognates at both the phrase level (e.g., *gird your loins* and *seine Lenden gürten* as independent literal translations of a biblical source texts) and for certain constituents (e.g., *gird* and *gürten* originating from Proto-Germanic *gurdijaną*), but not all constituents (English *loins* from Old French *loigne* and Latin *lumbus*, German *Lenden* from Old High German *lenti*).

3.3.7. Other measures

To better understand stimuli generalisability/specificity, we asked ChatGPT-4o (OpenAI, 2025) to classify each idiom's regional/dialect characteristics, register, and discourse domain. Since L2ers often rely on constituents when inferencing without context, we also obtained enTenTen20 frequencies of English idiom content words.

3.3.8. Data cleaning

From an initial 95 L1ers with 3,168 responses, one (33 responses) was removed for dictionary use. From an initial 99 L2ers with 3,366 responses, 11 (366 responses) were removed as non-German L1. Although 12 idiom familiarity question responses were missing, all available meaning interpretations for these participants were retained. For the L1ers and L2ers, respectively, these removals resulted in 1% and 11% data reductions, leaving 3135 (from 3168) and 2988 (from 3366) responses.

3.4. Data analysis

Data were analysed using R programming language (R Core Team, 2025) with several packages and Microsoft Excel.

To address RQ1, percentage means and standard deviations of L1/L2 idiom meaning recall were calculated and grouped by familiarity (Yes/No) and CLO type (1–6). Means, standard deviations and ranges are also reported for the AI-generated transparency ratings and idiom frequency estimates. In Section 3.3, we summarised the most relevant reliability information for the knowledge, familiarity and transparency variables, which were found to be generally comparable with those observed in recent research (e.g., Hubers et al., 2020; Mangiaterra et al., 2025) and the wider L2 field (Plonsky & Derrick, 2016). For more detailed reporting and discussion of instrument and rater reliability in the current study, we refer readers to the study's OSF page.

To address RQ2, we ran linear mixed-effects regression models using the lmer function in the 'lme4' R package (Bates et al., 2019). To maximise interpretability and

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

sensitivity, separate models were built for L1 and L2 groups rather than entering group as a predictor. With idiom meaning recall percentage score as the continuous outcome variable, initial additive models were run with all four predictors: familiarity (categorical, two levels, deviation-coded: no = −1, yes = 1); transparency (1–5, very unclear to very clear) and frequency (continuous), both standardised to aid interpretability; and CLO (categorical, six levels, deviation-coded, switching the non-compared level to obtain all estimates). Final, optimal-fitting models were identified as the combination of added or interactive predictors with the lowest Akaike Information Criterion (AIC) value, a criterion balancing likelihood, complexity, and interpretability, yielding a generalisable model more likely to predict novel data and less susceptible to overfitting than one positing more parameters (Winter, 2019). Optimal random participant and item effects were identified using the 'buildmer' R package (Voeten, 2022).

For the final models we report estimates and corresponding 95% CIs, standard errors, degrees of freedom, Wald *t* values, *p* values and standardised effect sizes (Cohen's *d*) for mixed-effects models, i.e., the fixed-effect estimate divided by the square root of the sum of the random intercept, slope and residual variances, assuming deviation coding of fixed-effects (Brysbaert & Stevens, 2018; Judd et al., 2017; Westfall et al., 2014), interpreted as small (0.2), medium (0.5) and large (0.8) (Cohen, 1988). Importantly, this method is more powerful and stricter than traditional approaches that generate larger estimates by first averaging participants/items per condition (F1/F2), thus dismissing their inherent variance (Brysbaert & Stevens, 2018; for a recent study showing how this can impact substantive findings in L2 vocabulary research, see Nicklin et al., 2025). Marginal and conditional R^2 showing variance proportions explained by fixed and combined fixed and random effects were computed using the 'performance' R package (Lüdecke et al., 2021) and interpreted as small (.18), medium (.32) or large (.51) (Plonsky & Ghanbar, 2018).

To analyse the other measures, we modelled regionality (categorical, four levels, deviation-coded: general, British English, American English, British and American English), register (1–6; informal, informal-to-neutral, neutral-to-informal, neutral, neutral-to-formal, formal-to-neutral) and constituent frequencies (tokens per million, more easily reported the typically large hits) as additional L1 and L2 predictors. Discourse domain information was rich and varied and so not operationalised in regression analyses but where relevant, we report counts of general occupations and usage.

To address RQ3, we first assigned and plotted percentage knowledge and familiarity for the 100 idioms, then thematically analysed L2 responses for different CLO types. While this analysis focuses on L2 responses, corresponding L1 patterns are reported where relevant.

4. Results

4.1. RQ1: L1 and L2 English speakers' idiom meaning recall

The L1ers perceived only 47% (1463/3135) of the idioms as familiar and the L2ers only 26% (776/2988). Figure 1 shows actual meaning recall scores as means plus and minus half standard deviations (for easier visualisation, given large dispersion), overall and by familiarity and CLO levels. Overall knowledge was low (i.e., <50%) except for L1ers responding to familiar idioms (*M* = 72.16, *SD* = 38.08). Both groups

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

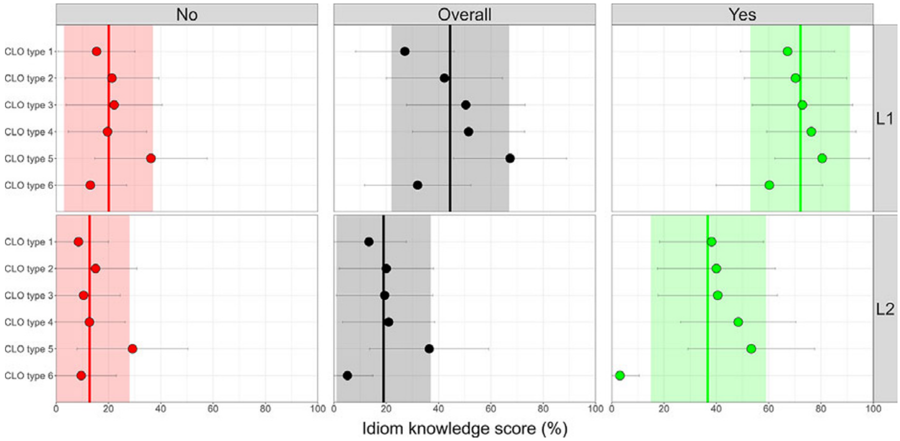

Figure 1. Point plot showing idiom averages (means) and spread (+/− half standard deviations) by group (vertical pane), familiarity (horizontal pane), and CLO type (see Section 3.3). Vertical lines and shaded areas show averages and spread by group and familiarity, points and horizontal when CLO type is also regarded.

had higher scores for familiar (than unfamiliar) idioms (L1: $M = 72.16$, $SD = 38.08$ vs $M = 20.09$, $SD = 34.29$; L2: $M = 36.66$, $SD = 44.22$ vs $M = 12.78$, $SD = 29.56$). For CLO, L2 scores were highest for type 5 (lexically different, non-metaphor equivalent) ($M = 36.42$, $SD = 45.52$) and lowest for type 6 (false friend) when disregarding familiarity ($M = 5.17$, $SD = 19.49$), and if regarding it, when type 6 (false friend) idioms were familiar ($M = 3.16$, $SD = 14.72$).

Figure 1 also shows L1 scores by CLO for comparative purposes, which were modelled in the RQ2 analyses (and expected to be unimportant). Descriptive statistics for the other predictors (not shown in Figure 1) show that idioms were skewed towards mid-to-higher transparency (range 1–5, $M = 3.72$, $SD = 0.91$) and varied widely in frequency (range = 6–33947.98, $M = 3388.70$, $SD = 4673.28$).

Idioms were mostly general/non-dialect specific (= 69%, British English = 14%, American English = 10%, British and American English = 7%). Their register was mostly informal-to-neutral (1–6 scale [informal to formal-to-neutral], $M = 2.06$, $SD = 1.32$) and constituent content word frequencies were much higher than for idioms as phrases (rate per million, $M = 276.62$, $SD = 1087.13$). Kruskal–Wallis tests indicated that none of these variables significantly differed by CLO type (regionality: $\chi^2(15, 100) = 8.12$, $p = .92$; register: $\chi^2(25, 100) = 27.96$, $p = .31$; constituents: $\chi^2(5, 100) = 7.61$, $p = .18$). Multiple discourse domains were represented throughout the stimuli with no discernible imbalance between CLO types (e.g., even the smallest categories, CLO types 5 and 6, both contained idioms spanning formal, professional, social and more personal, casual, informal contexts).

4.2. RQ2: Idiom meaning recall predicted by idiom familiarity, transparency, frequency and cross-language overlap

A comparison of AIC values across the different possible models showed that L1 idiom meaning recall was best predicted by a model containing an interaction between familiarity and transparency and an added frequency term, but without

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

CLO. L2 idiom meaning recall was best predicted by a model containing an interaction between familiarity and CLO and an added transparency term, but without frequency.

Figure 2 shows the estimates and 95% confidence intervals for all predictors in the final L1 and L2 models.

In the final L1 English model, the combined fixed and random effects explained a medium amount of total variance (conditional $R^2 = 0.502$) while the fixed effects explained a small amount (marginal $R^2 = 0.321$). Idiom familiarity had a medium, positive effect, i.e., the meanings of familiar idioms were better recalled (estimate = 20.784 [18.664, 22.903], *SE* = 1.081, *df* = 106.116, *t* = 19.219, *p* < .001***, *d* = 0.578); transparency had a very small effect, i.e., more transparent idioms were better recalled (estimate = 6.704 [3.331, 10.077], *SE* = 1.721, *df* = 113.16, *t* = 3.896, *p* < .001***, *d* = 0.187) and the familiarity-transparency interaction had a very small effect, i.e., familiarity and transparency increase together in predicting better idiom recall (estimate = 2.507 [0.375, 4.639], *SE* = 1.088, *df* = 117.17, *t* = 2.305, *p* = .023*, *d* = 0.07). Alongside these predictors, idiom frequency made only a very small, non-significant contribution (estimate = 1.554 [−1.879, 4.987], *SE* = 1.752, *df* = 106.59, *t* = 0.887, *p* = .377, *d* = .043), although a model with this predictor was superior to one without it.

In the final L2 English model, the fixed and random effects explained a medium amount of total variance (conditional $R^2 = 0.394$) while the fixed effects explained a small amount (marginal $R^2 = 0.125$). Controlling for other effects, meanings were significantly better recalled for familiar idioms, with a small, positive effect (estimate = 9.477[6.523, 12.431], *SE* = 1.507, *df* = 111.786, *t* = 6.288, *p* < .001***, *d* = 0.272); for more transparent idioms, with a very small positive effect

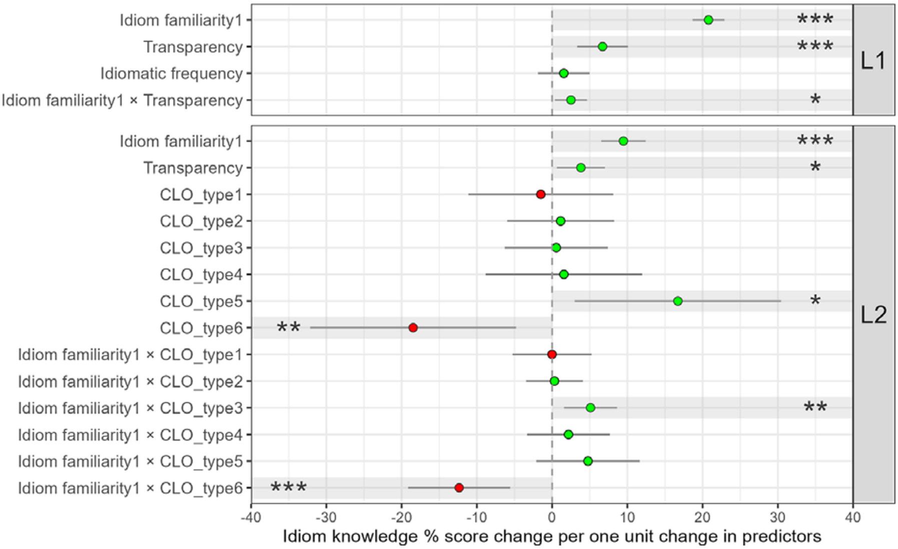

Figure 2. Dot-and-whisker plot showing final L1 (top) and L2 (bottom) model fixed effects estimates (dots) and 95% confidence intervals (whiskers): outcome (x-axis) = idiom meaning recall percentage score (L1 and L2 models); predictors (y-axis) = deviation-coded idiom familiarity (−1 = no, 1 = yes, L1 and L2 models), standardised transparency (1–5) and frequency (L1 model only), and deviation-coded CLO type (see Section 3.3; L2 model only); asterisks and shading show significance at .05*, .01**, and .001*** levels.

(estimate = 3.827 [0.628, 7.026], *SE* = 1.6322, *df* = 85.545, *t* = 2.345, *p* = .021*, *d* = 0.110) and for CLO type 5 idioms (lexically different, non-metaphor equivalent), with a small to medium, positive effect (estimate = 16.714 [2.998, 30.429], *SE* = 6.998, *df* = 87.769, *t* = 2.388, *p* = .019*, *d* = 0.479) while CLO type 6 idioms (false friends) were significantly worse recalled, with a medium, negative effect (estimate = −18.470 [−32.151, −4.788], *SE* = 6.9804, *df* = 87.075, *t* = −2.646, *p* = .010**, *d* = −0.530).

The strongest L2 interaction was between familiarity and German CLO type 6, which had a small, negative effect, i.e., false friend idioms perceived as familiar had worse meaning recall (estimate = −12.353 [−19.133, −5.574], *SE* = 3.459, *df* = 64.829, *t* = −3.571, *p* = <.001***, *d* = −0.354), suggesting over-confidence in familiarity based on a mistaken assumption of equivalence. A very small, but significant, interaction was also observed between familiarity and German CLO type 3, i.e., when perceived as familiar, the meanings of lexically and metaphorically different idioms where a German idiom exists were better recalled (estimate = 5.105 [1.583, 8.628], *SE* = 1.797, *df* = 71.169, *t* = 2.840, *p* = .006**, *d* = 0.146). Surprisingly, CLO type 1 (lexically and metaphorically similar) idioms were neither significantly better recalled (as might be expected) nor worse recalled, but knowledge seems low (see Figure 1).

Finally, neither the L1 nor L2 models were improved by entering regional/dialect characteristics, register or idiom constituents as additive or interactive predictors.

4.3. RQ3: Themes characterising L2 recall/inferencing of idioms of different CLO types.

Figure 3 plots L2ers' idiom meaning recall and familiarity percentages, while Figure 4 shows the size of difference (i.e., matches/mismatches) between them, where positive values imply underestimated knowledge (meaning recall > familiarity) and negative values imply overestimated knowledge (familiarity > meaning recall).

Generally, the L2ers were more prone to overestimation, occurring with 62% of the 100 idioms (*M* [*SD*] difference = −18.95 [18.42]) than underestimation, occurring with 36% of the 100 idioms (*M* [*SD*] difference = 13.31 [11.80]) or no meaning recall-familiarity difference observed for the remaining 2% of idioms.

At the group level, L1ers and L2ers used similar numbers of words in their responses (L1: 1–43, *Mdn* = 5, *IQR* = 4; L2: 1–37, *Mdn* = 4, *IQR* = 4), and linguistic and metaphorical themes were rich and varied. In the remainder of this section, we report on those themes underlying the significant CLO effects (first the interactions for types 6 and 3, then type 5), followed by the surprising non-significant CLO type 1 effect. Due to space constraints, we omit focus on CLO types 2 and 4, which unsurprisingly did not show any effect. We would also like to make clear that in the remainder of the article, 'recall' and 'inferencing' are used differently to refer to our participants' knowledge of familiar or unfamiliar idioms respectively.

4.3.1. Negative CLO type 6-familiarity interaction

As Figures 3 and 4 show, overestimation was strongest for CLO type 6 idioms (false friends; *M* [*SD*] difference = −63.50 [30.74]), occurring for each idiom within this category, suggesting a strong, negative L1–L2 transfer effect (the non-effect for CLO type 1 idioms is discussed below).

Qualitative evidence of false friend meanings can be found, in varying amounts, in responses to all CLO type 6 idioms. For example, when perceived as familiar,

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

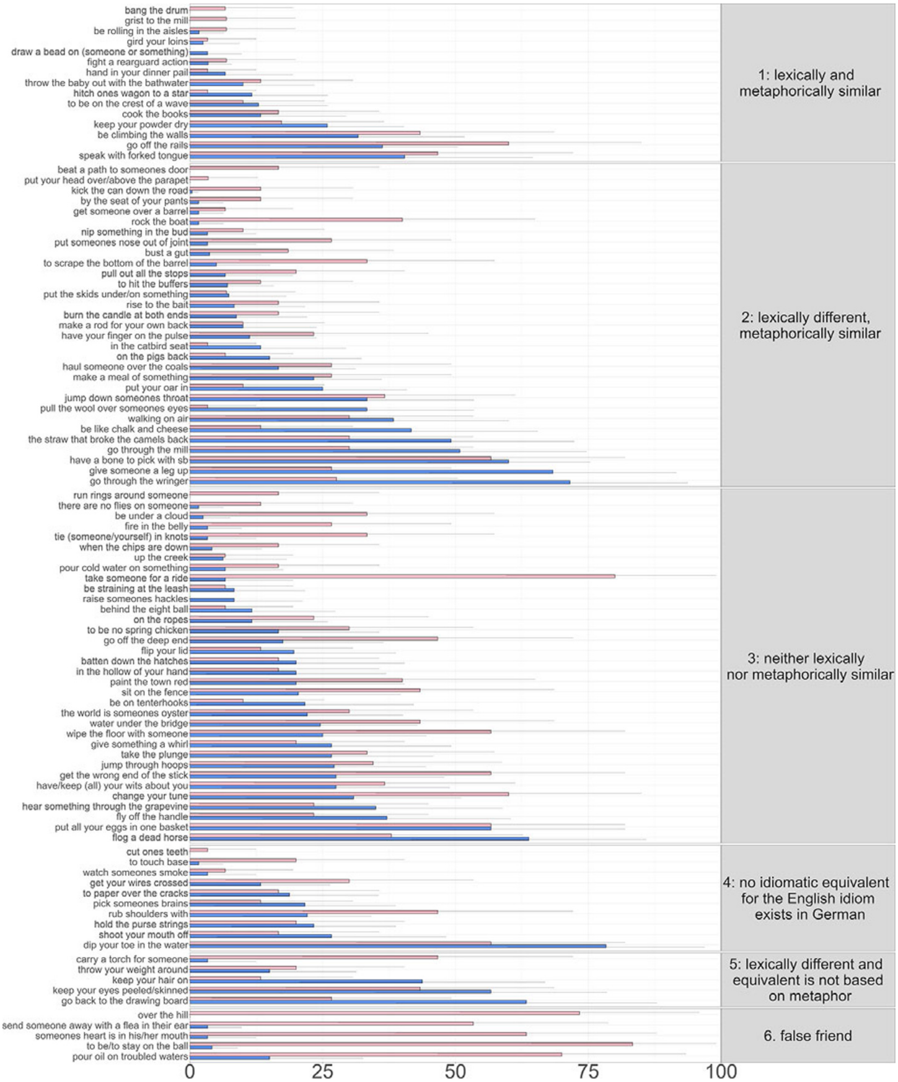

Figure 3. Percentages for L2 English idiom meaning recall (blue) and familiarity (pink) shown by bars and half SDs shown by faint blank lines, idioms grouped by CLO type (1–6, see Section 3.3).

responses to *over the hill* (‘old/obsolete’ [Eng.] vs ‘over the worst’ [Ger.]) mention overcoming general/non-specified “difficulty (A37, A46, A61, A48, A49, A58, A47),” recovery from physical illness such as “cancer (A44),” and stabilising following “an accident and then an operation (A54)” and “[illness/injury] (A55, A57, A60).” While the L1ers mostly correctly recalled this idiom as meaning “past your prime (C34, C42, C47)” or “old (C37, C44, C46, C61),” even some of their recall (C59, C40, C36, C50) and inferencing (C35, C39, C41, C49, C51, C56) mentions overcoming or escaping hardship. The theme of literal distance is common in the L2 data, as in “gone/far away (A40, A41, A53, A59, A59, A63, A64, A65, A67),” while both L1ers and L2ers also

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

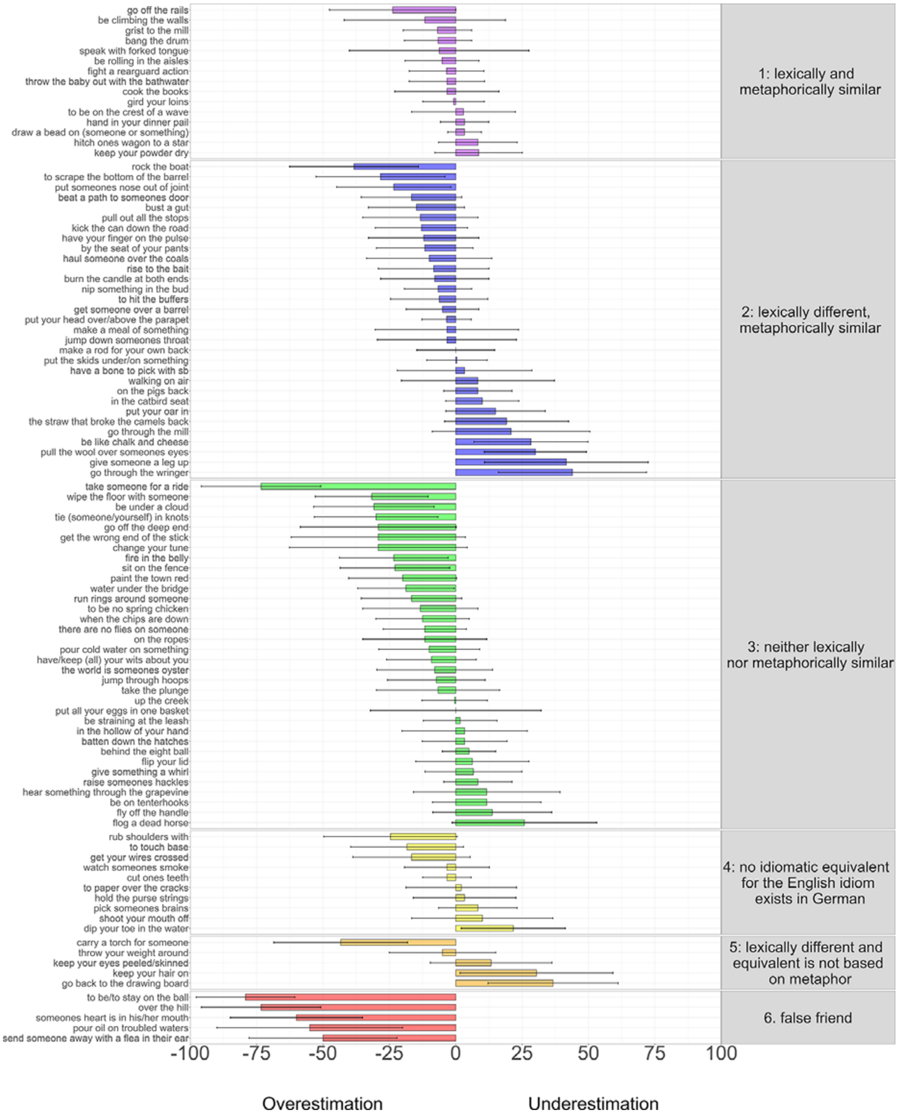

Figure 4. Bars showing differences between L2 English recall and familiarity percentage means and difference SDs shown by faint blank lines, per idiom grouped by German CLO type (1–6, see Section 3.3). Overestimated knowledge = bars < 0 (i.e., familiarity > meaning recall); underestimated knowledge = bars > 0 (i.e., meaning recall > familiarity).

mentioned the opposite, closeness: recalling “when something is not too far away (A39)” and “something is nearby (C53)” and inferencing “close but out of sight (C58),” which resembles the meaning of similar, but more positive expressions *on the horizon*, *around the corner* and *just over the hill*, underpinned by a different conceptualisation of EXPERIENCING IS SEEING.

Similarly, with *to be/stay on the ball* (‘quick to react’ [Eng.] vs ‘persistent’ [Ger.]), when perceived as familiar, L2ers more commonly invoked the false friend meaning via "keep[ing] doing (A68, A75, A76, A81, A82, A87, A91, A92)," "[continuing] (A72, A85, A88, A89, A90, A95, A96)," "to pursue (A83, A97)," "[being] motivated (A74)" and "finish[ing] (A69, A79)," compared with L1ers, who more typically referred to being "on track/task/time/top of things (C65, C67, C68, C72, C73, C76, C81, C89, C90)," "[attentive] (C63, C87)," "aware (C70, C77)," "alert (C94, C95, C91)," "focused (C71, C78, C84, C88)" and in "control (C82, C83)." Similar patterns can be found in other type 6 idioms: *pour oil on troubled waters* ('calm an argument' [Eng.] vs 'aggravating a bad situation' prompted by the close German equivalent *Öl ins Feuer gießen* ['pour oil in the fire']), *send someone away with a flea in their ear* ('angrily dismiss' [Eng.] vs 'put an idea in someone's head' [Ger.] which corresponds to the meaning of the German idiom *jemanden einen Floh ins Ohr setzen*) and *someone's heart is in his/her mouth* ('excited/worried/frightened' vs 'open/vulnerable' [Ger.]).

4.3.2. Positive CLO type 3-familiarity interaction

The positive interaction between familiarity and German CLO type 3 idioms (lexically and metaphorically different, German idiom exists) is difficult to observe in Figures 3 and 4, but these visuals do show how close matches between familiarity and knowledge span idioms at the lower and upper ends. At the lower end, the idiom, *up the creek* was generally unfamiliar and poorly inferred (knowledge: $M = 6.25$, $SD = 23.84$; familiarity: $M = 6.67$, $SD = 25.37$). When unfamiliar, incorrect guesses predominantly focus on the alternative meanings of: *up*, including "Onto the mountain (A76)," "high up (A81, A82)," "somewhere up? (A90)" and various other responses (A80, A94, A97, A98, A99); "on a high horse (A92)," which may be literal or metaphorical (i.e., arrogance) and "right there (A85, A91)." Unsurprisingly, L2ers did not focus on the low-frequency constituent *creek* (only 32 hits per million words in enTenTen corpus), but L1ers did, as in "upstream (C69, C85)," and potentially "go with the flow (C74)" and "towards the source of the problem (C94)." Like the L2ers, L1ers mentioned "[literal higher] (C87)" but more commonly inferenced "[closeness] (C66, C72, C75, C82, C86, C89)."

At the higher end, for *put all your eggs in one basket* (L2 knowledge and familiarity: $M = 56.67$, $SD = 50.40$), when perceived as familiar, L2 interpretations such as "[to put everything on] one card (A4, A35)" and "to go all in (A5, A30)" were successful, depicting the general theme of GAMBLING. Similarly, L1ers mentioned "[gambling/risky investment] (C1, C3, C30, C8, C15, C29)," general reliance and confidence in "one thing (C11, C12, C13, C20, C21, C22, C27)," "one person (C28)," "one path (C19)" and so forth Neither group invoked egg or basket imagery, but rather, incorrect inferences centred around *put all* and *one*: for L2ers: "get one's life together (A1)," "To collect your ideas or thoughts (A7)," "Use all your available tools (A8)" and "Giving everything to achieve something (A18)""; for L1ers: "keep it together (C7)" and "put all your ideas down (C25)."

4.3.3. Positive CLO tyrpe 5 effect controlling for familiarity

Thematically, the elevated knowledge for this type is particularly attributable to three idioms (*keep your hair on*, *keep your eyes peeled/skinned*, and *go back to the drawing board*), which had a comparatively large influence given the number of items for this type. In these three idioms, L2ers picked up on basic conceptualisations that underlie the idiomatic meanings. For example, in *go back to the drawing board*, their responses focus on the metaphor of REDOING IS RETURNING TO THE START, as in evident in the mention of "[re]start[ing] (A2, A5, A14, A19, A23, A24, A26, A27, A30, A31, A33, A34)" and "back (A4, A5, A15, A16, A17, A22, A23, A25, A26, A35)".

Similarly, in *keep your hair on*, the L2ers invoked being "calm (A39, A42, A43, A44, A46, A47, A48, A58)…and collected (A42)" and "keep[ing] your shirt on; to calm down (A48)." These associations can be related to metaphoric conceptualisations of calmness as a non-disruptive bodily state, which is also evident in the answer keep[ing] a cool head (A67)," expressing the metaphor of ANGER IS A HOT FLUID RISING IN A CONTAINER. The L1ers used more alternative metaphorical phrases, as in "freak[ing] out (C34, C41)" and "stay[ing] level-headed (C50)," with similarly rare (but present) evidence of temperature imagery, with "…chill (C60)," and "dont lose your cool (C48)."

Given the non-interaction with familiarity for this type, underestimated knowledge was offset by overestimated knowledge for the other two items, especially *carry a torch for someone* ('to be in love with someone') which, shows incorrect recall themes linked to irrelevant constituent meanings, as in "help (A40, A43, A45, A56, A59, A61, A63)" and "[lighting the way] (A36, A61)," which also appeared in the L1 responses "help (C37, C39)" and "[lighting the way] (C39, C49, C50, C56)," alongside unique L2 and L1 themes, respectively, "[defence/protection] (A46)" and "[continuing someone's] legacy (C32, C47, C59)."

4.3.4. Non-effects for CLO type 1 idioms

Surprisingly, no effect was found for CLO type 1 idioms (lexically and metaphorically similar), which were not significantly better or worse known than other types, either when controlling for or in interaction with familiarity. At both the higher and lower ends, similar L1 and L2 response themes were observed.

For the easiest CLO type 1 idiom *speak with forked tongue* ('to deceive'), which has a very close German equivalent (*mit gespaltener Zunge reden* [lit. with split tongue talking]), correct L2 recall/inferencing was characterised by "[lying] A39, A10, A42, A44, A45, A46, A47, A54, A58, A64) "[being] two-faced (A47)", "speak in lies (like a snake) (A54)" etc., patterns which were remarkably similar for L1ers, who very often mentioned "[lying] (C34, C37, C38, C40, C41, C42, C43, C44, C46, C47, C52, C53, C54, C56, C58, C59, C61)" and who also once mentioned "speak deceitfully, like a snake (C56)". Incorrect L2 responses focused on speaking "[unclearly] (A38, A40, A56, A67, A43, A36)," "[incompletely/with restraint] (A61, A63, A48, A60)," "[meanly/without restraint] (A66, A41, A53, A49)" and "[of secrets or plot twists] (A55, A59)." Again, L1 similarly mentioned "[difficult expression/indecision] (C31, C33)," "[obfuscation/unclear/odd speech] (C35, C36, C57, C62)," "[meanness/directness] (C45, C60)," "[with restraint/caution] (C 32, C39, C50)," alongside their own unique theme "[nonnative language use] (C51)."

For *bang the drum* ('promoting someone/something'), the joint hardest CLO type 1 idiom with a German equivalent involving a different action (*die (Werbe-)Trommel für etwas rühren* [lit. The (advertising-)drum for something stir]), all L2 responses (mostly inferences) focused on "[loudness] (A6, A9, A15, A17, A20, A26, A32, A34)," "[progress/completion] (A11, A23, A33, A31)," "[craziness/unrestraint] (A4, A14, A21, A22)," "[initiation/control] (A16, A19, A25, A30)," "[risk] (A8, A24, A35)," "[excelling/achieving] (A27, A29)," "[attention seeking/shock] (A1, A3, A5)" and "[overcomplication] (A2)." L1ers showed similar themes of "[loudness, both neutral

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

and negative] (C1, C7, C28, C30, C9, C10, C13, C27)," "[progress/striving] (C17, C24) [and steadily-paced work] (C26)," "[excitement] (C14)," "[initiation/control] (C8, C15)" and "[troublemaking] (C20)," and their own unique themes of "[pre-existing support] (C4)," "[warning] (C21)" and "[alertness] (C23)."

5. Discussion

5.1. RQ1 discussion

RQ1 asked how well L1/L2 English speakers could recall/inference the meanings of 100 challenging idioms from learner resources. In many studies, there is a tacit assumption that L1ers should perform at (near) ceiling level, whereas L2ers are more varied (Carrol et al., 2018). The first set of results add important nuances to L1, and indeed L2, norms in this respect. We observed that knowledge, as measured, could only be considered high (but certainly not ceiling) for L1ers responding to familiar idioms, in all other cases it was low (< 50%). However, even high-end L1 scores were lower and more varied than L1 meaning recognition levels with familiar idioms in previous studies, where idioms were thought to be so well known that meaning was retrieved without much inferencing (Carrol et al., 2018; Hubers et al., 2020).

Similarly, our L1ers' inferencing of unfamiliar idioms was even lower than with Guo and Xiang's (2023) recall test. While these differences might be partly attributable to decontextualised idiom presentation and test format generally (far fewer cues than in real-world interaction), even with the provision of discourse context in Guo and Xiang (2023), meaning recall remained challenging (see Section 2.1).

Taken together, these findings confirm that the 100 idioms used in the current study represent a particularly difficult pool extracted from the various pedagogical resources. While the general pattern of L1 knowledge mirrors previous research (familiar idioms better known than unfamiliar ones), it is noteworthy that our L1ers were mostly *unable* (~20%, on average) to infer idioms perceived as unfamiliar. This chimes in with Carrol et al.'s (2018) argument that when idioms are unfamiliar, L1–L2 knowledge is more on par, perhaps with marginal L1 advantage from cultural familiarity. The non-effects for regionality and register do not speak to this issue per se since these predictors reflect variation in the idioms, not in participants' relevant cultural knowledge. Rather, the non-effects indicate that neither L1 nor L2 knowledge was sensitive to variations in these generalised usage patterns, as operationalised.

The stronger ceiling effects in research on other figurative multiword units such as phrasal verbs (O'Reilly, 2017) are probably explained by comparatively higher frequency and entrenchment, and less semantic information. For example, compared with the phrasal verb *break up*, the idiom *send someone away with a flea in their ear* ('angrily tell someone to go away') has more processing elements, some with possibly misleading connotations (e.g., *flea* = parasite/disease/jumping/itchiness). For L2ers, on the other hand, phrasal verbs are usually nontransparent lexical units to be learned by heart, especially when absent or different in the L1.

Unlike many L2-focused studies (Aljabri, 2013; Aydin, 2019; Boers et al., 2007; Park & Chon, 2019; Soto-Sierra & Ferreira, 2024; Vasijelvic, 2016), our data also offer L1–L2 comparisons for the same idiom set (as in Carrol et al., 2018; Hubers et al., 2020). Our L2ers had more markedly low and variable meaning recall/inferencing than with meaning recognition measures from other EFL contexts (Hubers et al., 2020; Soto-Sierra & Ferreira, 2024), which is intuitive, given the established

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

recognition-to-recall trajectory (González-Fernández, 2022; González-Fernández & Schmitt, 2020). In the context of the larger educational project (see Section 1), L2ers' unfamiliarity and difficulty with the 100 idioms affirms their suitability for pedagogical interventions (see the resource *59 Advanced English Idioms: Illustrated and Exemplified with Learning Games*, available at https://www.aau.at/wp-content/uploads/2025/04/59_Advanced_English_idioms_ue-komprimiert.pdf).

5.2. RQ2 and RQ3 discussion

RQ2 asked about the extent to which L1 and L2 idiom meaning recall/inferencing predicted by subjective familiarity, and objective frequency, transparency and CLO, while RQ3 asked about the themes characterising L2ers' interpretations of idioms of different CLO types. The quantitative results (RQ2) showed that L1 knowledge was positively predicted by familiarity, and to a lesser extent, transparency, the familiarity-transparency interaction and frequency. L2 idiom recall/inferencing, on the other hand, was positively predicted by familiarity and its interaction with CLO type 3 (lexically and metaphorically different, German idiom exists), transparency and CLO type 5 (lexically different, non-metaphor equivalent) and negatively predicted by CLO type 6 (false friends) and its interaction with familiarity. The qualitative results (RQ3) corroborated L2ers' misguided intuitions based on false-friend meanings and provided insight into their relevant intuitions for lexically and metaphorically different idioms and points of L1–L2 comparison.

5.2.1. Familiarity, transparency, frequency

The clear familiarity effect in this and other studies affirms that L1ers and L2ers have a good grasp of their idiom knowledge, whether measured via meaning recognition or recall format. For both, subjective familiarity overshadows objective, corpus-based frequency, providing a more meaningful index of true idiom encounters (Bonin et al., 2013). The fact that objective frequency still mattered, to some extent, for L1ers, is explained by their overwhelmingly high and less variable language exposure (Carrol et al., 2018; Cronk et al., 1993; Cronk & Schweigert, 1992; Libben & Titone, 2008; Tomasello, 2003) compared with L2ers, for whom frequency has also been a non-effect for meaning recognition (Hubers et al., 2020; Soto-Sierra & Ferreira, 2024).

The L1 familiarity-transparency interaction suggests that nonfrequency factors predominate in making certain idioms seem familiar and their meanings memorable. Even with limited exposure, idioms are typically salient and often encountered with various contextual, connotational, affective, and auditory cues that can aid rapid uptake. For instance, the young and adolescent female L1ers in Reuterskiöld and Van Lancker Sidtis (2013) had better recognition and comprehension of low frequency idioms introduced once in a conversation compared with nonidiomatic expressions and other idioms and nonidiomatic expressions not introduced.

For L2ers, the familiarity and transparency effects bore both similarities and differences with previous studies (e.g., familiarity better predicted knowledge in Carrol et al., 2018 but not in Hubers et al., 2020 or Soto-Sierra & Ferreira, 2024). Certainly, the positive transparency effect for L2 learners aligns with theoretical perspectives emphasising the primacy of literal meanings and word-by-word inferencing for L2 learners (Cieślicka, 2006, 2015), which in Soto-Sierra and Ferreira's (2024) view, is also in line with L2 reliance on L1 lexical knowledge. The fact that transparency was a weaker predictor for our L2 (than L1) participants, and not the strongest predictor overall (cf. Soto-Sierra & Ferreira, 2024), is probably best explained by the difficulty of many of the idioms, pushing L1 participants to employ semantic analysis where they lacked phrase-level mental representations, which they did with comparatively more success. The finding may also be partly attributable to the operationalisation of transparency as an AI-generated variable, which given the GPT-4o large language model's account of its training data and the variety of sources used (see the study's OSF page), suggests that ratings are likely more representative of patterns of L1 usage and norms than commensurate with transparency in the minds of our L2 learners. Extrapolating from Hubers et al. (2020), whose L2ers had a comparable lack of familiarity to ours (22.8% vs 26%), we would expect that in testing with these 100 idioms, L2 participant-ratings, had we taken them, would supersede human-generated L1 norms as knowledge predictors (as shown in Hubers et al., 2020), or indeed AI-generated ratings, as we used, although the comparative efficacy of the latter two is less clear (see Section 6).

5.2.2. CLO types 3, 5, and 6

What mattered most for L2 idiom meaning recall/inferencing was CLO. The advantage of increased CLO is well documented in studies measuring offline knowledge types (Charteris-Black, 2002; Deignan et al., 1997; Kainulainen, 2006; Laufer, 2000) and online processing (e.g., Carrol & Conklin, 2014, 2017; Carrol et al., 2016). Our study contributes new evidence here, showing a positive familiarity interaction, which unexpectedly was not with lexically and metaphorically similar idioms, and a negative effect for false friends.

The false friend results are more easily interpreted since the mechanism is clearer; L2ers perceive an equivalence, albeit in error, and draw on the L1 idiom to explain L2 idiom meaning. However, even though the L2 learners largely declared these idioms as familiar, the reality of this familiarity needs careful consideration. One possibility is that the L2 learners had in fact never previously encountered (or noticed encountering) these English idiom forms, but the force of the perceived L1 equivalent, a false friend, was so strong that it caused the English idiom to be erroneously recognised as familiar in the moment. As previous studies have shown, even nonexistent L2 words/phrases can seem familiar if resembling those in the L1 (e.g., Carrol et al., 2018; Hall, 2002). Another possibility is that this process had already happened in the past, and rather, the current study's data provide evidence of a more established, form-meaning connection, i.e., 'the fossilization of false cognates' (Hall, 2002, p. 82).

Thematically, L1 and L2 responses for this type were clearly different, but not entirely. The fact that some L1 participants actually recalled/inferenced the German false friend meaning of *over the hill* ('over the worst') was highly revealing, not for any L1–L2 cross-linguistic reason, but because it shows that even they can be led astray by salient competing metaphors, including those that also happen to underpin false friend meanings for L2ers. (As an aside, to the extent that people tend to remember idioms most relevant to them, it is also not surprising that an idiom to denote feeling old and obsolete is not well-ingrained with young adults, see also Carrol, 2023). The caveats here, though, are that the current study did not operationalise the learner-internal reality of the apparent metaphorical mappings (e.g., their online processing), and that it remains unclear whether or how, in their day-to-day speaking/writing,

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

participants would produce such idioms to convey these misunderstood meanings (e.g., *over the hill* to mean 'overcoming/closeness/distance' etc., see Section 4.3). To the extent that erroneous form-meaning links inhabit their productive idiom vocabulary, the question is then about the effect of any feedback (if any) from interlocutors noticing the issue.

From a theoretical perspective, although Hall initially implied an L2-only scope for the *Parasitic Strategy*, a model 'formulated to account for early stages of vocabulary development in second language learners' (2002, p. 69), his subsequent discussion and examples (see Section 2.2) lead to the eventual argument that 'the Parasitic Strategy is essentially promiscuous with regard to the language source of potential form associates accessed, i.e., that any form in L1 or L2 (or L3) can influence the processing of any other form in L1 or L2 (or L3) with which it overlaps' (2002, p. 81–82). Our findings show how this promiscuity also manifests as L1–L1 connections, as some L1 participants reach for alternative, less relevant meanings, for them salient, either during inferencing or recall, suggesting possible fossilisation (also see Onysko, 2016 on meaning associations to novel compounds by monolinguals and multilinguals).

While the familiarity-CLO type 3 interaction is more difficult to interpret, the response themes, which were remarkably similar for both groups, suggest incorrect inferencing characterised by barking up the wrong tree with irrelevant/unhelpful constituent meanings and correct recall/inferencing characterised by core concepts shared in the L1 and L2, rather than the specifics of more immediate conceptual metaphors and metonymies at the phrase and constituent levels (Hall, 2002; Soto-Sierra & Ferreira, 2024). This latter phenomenon, offset by overestimation, probably also explained the positive CLO type 5 effect. For CLO types 5 and 6 though, despite small-to-medium size effects (higher if based on participant or item averages, cf. Brysbaert & Stevens, 2018; Nicklin et al., 2025) and 95% confidence intervals not containing zero, it is important to emphasise that further substantiation with larger idiom samples is required.

5.2.3. CLO type 1 (non-significant effect)

The finding that CLO type 1 idioms (lexically and metaphorically similar) did not stand out as easier or more difficult was surprising given previous research (see Section 2.3). This finding does not seem to be accounted for by low constituent frequencies within certain CLO type 1 idioms (e.g., less than one hit per million for *rearguard* and *grist*), since constituent frequency did not contribute to the optimal L2 model, and comparable infrequency occurs within other CLO types (e.g., *wringer* [type 2] and *batten* [type 3]). Rather, we suggest two possible reasons, pending further research.

First, given our L1ers' surprisingly low and variable idiom knowledge, by analogy, the L2ers may also have rather fragile knowledge of (many of) the corresponding idioms in German, their L1, and a very limited basis for successful L1–L2 idiom transfer despite the availability of phrase-level connections that might have been exploited (Hall, 2002). Although we did not measure the L2ers' knowledge of the German equivalents, to the extent that frequency plays some L1 role, it is worth noting that even before filtering out nonidiomatic usages, total phrasal hits (an overestimate of idiomatic frequency) of CLO type 1 German idioms within the German Web Corpus (deTenTen23) are lower ($M = 1524.33$, $SD = 2155.31$) than

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

English idiomatic usage frequencies overall ($M = 3388.70$, $SD = 4673.28$) and for CLO type 1 idioms specifically ($M = 2008.26$, $SD = 1906.42$). Perhaps this hindered the potential for the recognition and exploitation of L1–L2 links.

Second and possibly related, inspection of the lexical and metaphorical correspondences between English idioms and German counterparts shows that CLO type 1 idioms were typically more closely aligned with Hubers et al.'s (2020) 'some common content words' category, which had no effect, along with 'completely different content words' and 'no equivalent' (i.e., only their 'word-for-word equivalence' category predicted better meaning recognition). Thus, for a strong facilitating effect to occur, lexical and metaphorical correspondence between L1–L2 idioms might need to be identical (or very close) rather than merely similar. However, this pattern is not clearly seen in our data. Although several better-known CLO type 1 idioms have very close L1–L2 lexical/metaphorical similarity (*speak with forked tongue*, *be climbing the walls*, and *keep your powder dry*), others (*throw the baby out with the bathwater*, *gird your loins*) had very low knowledge scores, and even the best-known type 1 idiom (*speak with a forked tongue*) was superseded by 12 idioms from CLO types 2, 3, 4 and 5.

Finally, our models compared each CLO type to the grand mean of all types rather than to full equivalents (Hubers et al., 2020) or nonequivalents (Soto-Sierra & Ferreira, 2024). If we had used CLO type 1 (lexically and metaphorically similar) as the reference level, only familiarity, transparency and the negative familiarity-CLO type 6 interaction remain as significant, giving further confidence that lexically and metaphorically similar idioms were not better or worse known relative to the stimuli as a whole.

6. Limitations and future research

First, while we sought to contextualise results alongside previous studies, more systematic and direct, within-participants investigations are needed into how the vocabulary knowledge component hierarchy applies to idioms, as distinct from the collocates and multiple meanings featured in González-Fernández and Schmitt's (2020) test battery, and the variety of moderating factors.

Second, to maximise parsimony and model sensitivity, we selected four key predictors for quantitative investigation. Where greater sample and item sizes are possible (Brysbaert, 2019), interactions between other subjective/objective predictors such as plausibility, predictability, imageability, saliency, valence, arousal etc. (see Beck & Weber, 2016) could be further explored through unidirectional and/or more complex models (e.g., SEM-based, as in González-Fernández/Schmitt's work). Particularly intriguing are the role of perceived (rather than pre-categorised/objective) CLO, and the efficacy of participant versus L1/L2-normed versus AI-rated/coded variables, in the context of growing interest in human and AI in Applied Linguistics (e.g., Lamprianou, 2025). Given the size, diversity and temporal relevance of the AI information base, this information source could provide several advantages over published norms. In the current study, ChatGPT-4o produced stable ratings over different sessions (test–retest reliability) and a transparency variable that uniquely explained (albeit small) amounts of L1 and L2 idiom knowledge (predictive validity). While the use of OpenAI in the current study should be understood as exploratory and in need of further investigation, its performance compares well with emerging

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

research by Mangiaterra et al. (2025), who found that GPT-4o (in particular, compared with other LLMs) was highly effective at producing metaphor familiarity and comprehensibility ratings closely aligned with human-generated ratings in eight published datasets (convergent validity), similarly predictive of behavioural and neurolinguistic effects (predictive validity), and stable across separate interrogation sessions (test–retest reliability).

Third and related, although our variables seemed sufficiently reliable for present purposes (see OSF page), more evidence is needed to understand the precision/burden trade-off linked to different scale sizes (Rodriguez, 2005) and on reported strategy use. While we provided some qualitative L1–L2 comparisons, further idiom knowledge studies could also expand this aspect.

Finally, although the homogeneity in our L1 and L2 groups meant that data did not speak to the effects of age and educational level, we echo Carrol's (2023) suggestion for more research on the dynamic nature of idiom knowledge throughout L1, L2 and multilingual lifespans, and on different language input sources (e.g., reading, media/entertainment, occupation).

7. Conclusion

The current study provides new insights into how well L1 and L2 speakers of English can recall/infer the meanings of challenging, pedagogically relevant idioms underpinned by conceptual metaphor. The finding that the L1ers were even further off ceiling-level performance than in previous research suggests that gold-standard notions of L1 norms cannot be taken for granted for a given idiom set, but need to be empirically established, especially when knowledge is operationalised as recall/inferencing, rather than recognition, of form-meaning links. The modelling revealed nuances in how familiarity, transparency, frequency and CLO explain knowledge, individually and/or in interaction; overall, familiarity mattered for both L1 and L2 groups, and objective frequency for L1ers only, given their richer English language experience. While the *Literal Salience Hypothesis* would suggest a prominent role of transparency for L2 learners (Cieślicka, 2006, 2015), the significant (but comparatively lesser) role of this factor for both groups in our study and its stronger effect for L1 participants, suggests that idiom difficulty was such that even the L1ers were forced to rely on literal, word-by-word analysis in many cases. As an objective, AI-generated predictor, the transparency effects also suggest that this variable, as operationalised, is somewhat more reflective of L1 (than L2) norms and usage, and although predictive of idiom knowledge for both groups, is probably less informative as a knowledge predictor than participants' subjective intuitions (Hubers et al., 2020; Soto-Sierra & Ferreira, 2024).

Building on a recent application of the *Parasitic Strategy* (Hall, 2002) to L2 idiom knowledge research (Soto-Sierra & Ferreira, 2024), the response themes revealed the deceptiveness and tenacity of highly salient pre-existing form-meaning representations as ingenuine allies (CLO type 6, false friend idioms), and that L2 learners are not always able to draw on crosslinguistic connections that could, if known and recognised, help them (CLO type 1, lexically and metaphorically similar idioms). Regarding the L1 participants, the study provides evidence of the *Parasitic Strategy* operating across forms and meanings within the L1, a promiscuity that was suspected, but not directly observed in its original formulation.

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

Data availability statement. All materials, data, analyses and further methodological information are available on the study's Open Science Framework (OSF) page: https://osf.io/b69pf/.

Acknowledgements. The authors would like to thank the anonymous reviewers for numerous helpful comments on the manuscript and Giulia Bencini for help with the initial planning and setup of the study.

Funding statement. This research was funded in whole or in part by the Austrian Science Fund (FWF; grant https://doi.org/10.55776/I5280). For open-access purposes, the author has applied a CC BY public copyright licence to any author-accepted manuscript version arising from this submission.

Competing interests. The authors declare none.

References

Abel, B. (2003). English idioms in the first language and second language lexicon: A dual representation approach. *Second Language Research*, 19, 329–358.

Aljabri, S. S. (2013). EFL students' judgments of English idiom familiarity and transparency. *Journal of Language Teaching and Research*, 4(4), 662–669. https://doi.org/10.4304/jltr.4.4.662-669.

Aydin, B. (2019). Cognitive processing of second language idiom comprehension: A comparative study. *Journal of Language and Linguistic Studies*, 15(1), 307–325. https://doi.org/10.17263/jlls.547750.

Bates, D., Mächler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., & Fox, J. (2019). Lme4: Linear mixed-effects models using 'Eigen' and S4 (Version 1.1–21) [R package]. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/package=lme4

Beck, S. D., & Weber, A. (2016). Bilingual and monolingual idiom processing is cut from the same cloth: The role of the L1 in literal and figurative meaning activation. *Frontiers in Psychology*, 7, 1–16. https://doi.org/10.3389/fpsyg.2016.01350.

Boers, F., Eyckmans, J., & Stengers, H. (2007). Presenting figurative idioms with a touch of etymology: More than mere mnemonics? *Language Teaching Research*, 11(1), 43–62. https://doi.org/10.1177/1362168806072460.

Bonin, P., Méot, A., & Bugaiska, A. (2013). Norms and comprehension times for 305 French idiomatic expressions. *Behavior Research Methods*, 4(45), 1259–1271. https://doi.org/10.3758/s13428-013-0331-4.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, 2(1), 16. https://doi.org/10.5334/joc.72

Brysbaert, M. & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, 1(1) 9, 1–20, https://doi.org/10.5334/joc.10

Bulkes, N. Z., & Tanner, D. (2016). Going to town": Large-scale norming and statistical analysis of 870 American English idioms. *Behavior Research Methods*, 2(49), 772–783. https://doi.org/10.3758/s13428-016-0747-8.

Carrol, G. (2023). Old dogs and new tricks: Assessing idiom knowledge amongst native speakers of different ages. *Journal of Psycholinguistic Research*, 52, 2287–2302. https://doi.org/10.1007/s10936-023-09996-7

Carrol, G., & 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. https://doi.org/10.1017/S1366728915000103

Carrol, G., & Conklin, K. (2014). Getting your wires crossed: Evidence for fast processing of L1 idioms in an L2. *Bilingualism Language and Cognition*, 17, 784–797. https://doi.org/10.1017/s1366728913000795.

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

Carrol, G., Conklin, K., & Gyllstad, H. (2016). Found in translation. *Studies in Second Language Acquisition*, 38(03), 403–443. https://doi.org/10.1017/s0272263115000492.

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

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

Charteris-Black, J. (2002). Second language figurative proficiency: A comparative study of Malay and English. *Applied Linguistics*, 23, 104–133.

Cieślicka, A. (2006). Literal salience in on-line processing of idiomatic expressions by second language learners. *Second Language Research*, 22(2), 115–144.

Cieślicka, A. (2015). Idiom acquisition and processing by second/foreign language learners. In R. R. Heredia & A. B. Cieślicka (Eds.), *Bilingual figurative language processing* (pp. 208–244). Cambridge University Press. https://doi.org/10.1017/CBO9781139342100.012.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, 29(1), 72–89. https://doi.org/10.1093/applin/amm022.

Cronk, B. C., Lima, S. D., & Schweigert, W. A. (1993). Idioms in sentences: Effects of frequency, literalness, and familiarity. *Journal of Psycholinguistic Research*, 22(1), 59–82. https://doi.org/10.1007/BF01068157.

Cronk, B. C., & Schweigert, W. A. (1992). The comprehension of idioms: The effects of familiarity, literalness, and usage. *Applied PsychoLinguistics*, 13(2), 131–146. https://doi.org/10.1017/S0142716400005531.

Deignan, A., Gabry's, D., & Solska, A. (1997). Teaching English metaphors using cross-linguistic awareness-raising activities. *ELT Journal*, 51(4), 352–360. https://doi.org/10.1093/elt/51.4.352

Ellis, N. C. (1998). Emergentism, connectionism and language learning. *Language Learning*, 48(4), 631–664.

González-Fernández, B. (2022). Conceptualizing L2 vocabulary knowledge: An empirical examination of the dimensionality of word knowledge. *Studies in Second Language Acquisition*, 44(4), 1124–1154. https://doi.org/10.1017/S0272263121000930.

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, 41(4), 481–505. https://doi.org/10.1093/applin/amy057.

Guo, Y., & Xiang, M. (2023). Investigating the factors influencing the comprehension of idiom variation in a second language. *English Language Teaching*, 16(2), 43–52. https://doi.org/10.5539/elt.v16n2p43.

Hall, C. J. (2002). The automatic cognate form assumption: Evidence for the parasitic model of vocabulary development. *International Review of Applied Linguistics in Language Teaching*, 40(2), 69–87. https://doi.org/10.1515/iral.2002.008.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, 21(2), 303–317. https://doi:10.1017/S0272263199002089

Hubers, F., Cucchiarini, C., & Strik, H. (2020). Second language learner intuitions of idiom properties: What do they tell us about L2 idiom knowledge and acquisition? *Lingua*, 246, 102940. https://doi.org/10.1016/j.lingua.2020.102940.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen corpus family. In *Proceedings of the 7th international corpus linguistics conference* (pp. 125–127). Masaryk University.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, 68(1), 601–625.

Kainulainen, T. (2006). *Understanding idioms: A comparison of Finnish third grade students of national senior secondary school and IB diploma programme*. University of Jyväskylä.

Keysar, B., & Bly, B. (1995). Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans? *Journal of Memory and Language*, 34, 89–109.

Keysar, B., & Bly, B. (1999). Swimming against the current: Do idioms reflect conceptual structure? *Journal of Pragmatics*, 31, 1559–1578.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The sketch engine: Ten years on. *Lexicography*, 1, 7–36. http://www.sketchengine.eu

Kim, C. (2016). L2 learners' recognition of unfamiliar idioms composed of familiar words. *Language Awareness*, 25(1–2), 89–109. https://doi.org/10.1080/09658416.2015.1122025.

Kucera, H., & Francis, W. (1967). *A computational analysis of present-day American English*. Brown University Press.

Lamprianou, I. (2025). Network analysis for the investigation of rater effects in language assessment: A comparison of ChatGPT vs human raters. *Research Methods in Applied Linguistics*, 4(2). https://doi.org/10.1016/j.rmal.2025.100205.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. https://doi.org/10.2307/2529310.

Laufer, B. (2000). Avoidance of idioms in a second language: The effect of L1-L2 degree of similarity. *Studia Linguistica*, 54(2), 186–196. https://doi.org/10.1111/1467-9582.00059.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x.

Libben, M. R., & Titone, D. A. (2008). The multidetermined nature of idiom processing. *Memory & Cognition*, 36(6), 1103–1121. https://doi.org/10.3758/MC.36.6.1103.

Littlemore, J. (2001). The use of metaphor in university lectures and the problems that it causes for overseas students. *Teaching in Higher Education*, 6(3), 333–349. https://doi.org/10.1080/13562510120061205.

Littlemore, J. (2008). The relationship between associative thinking, analogical reasoning, image formation and metaphoric extension strategies. In M. S. Zanotto, L. Cameron, & M. C. Cavalcanti (Eds.), *Confronting metaphor in use: An applied linguistic approach* (pp. 199–222). John Benjamins.

Liu, D. (2008). *Idioms: Description, comprehension, acquisition, and pedagogy*. Routledge.

Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). Performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60), 3139.

MacWhinney, B. (1997). Second language acquisition and the competition model. In M. B. d. G. Annette & J. F. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives* (pp. 113–142). Lawrence Erlbaum Associates.

Mangiaterra, V., Al-Azary, H., Barattieri di San Pietro, C. & Bambini, V. (2025). GPT as a rater: Systematic evaluation of machine-generated norms for English and Italian metaphors. In *17th researching and applying metaphor (RaAM) conference on metaphor, technology, and communication*. Lawrence Technical University.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nicklin, C., McLean, S., & Vitta, J. P. (2025). Contrasting fixed- and mixed-effects modeling in vocabulary research: Reanalyzing Laufer (2024) and McLean et al. (2020). *Language Learning*, 1–38. https://doi.org/10.1111/lang.12715.

Nippold, M. A., & Taylor, C. L. (2002). Judgments of idiom familiarity and transparency: A comparison of children and adolescents. *Journal of Speech, Language, and Hearing Research*, 45(2), 384–391. https://doi.org/10.1044/1092-4388(2002/030.

O'Reilly, D. (2017). *An investigation into metaphoric competence in the L2: A linguistic approach*. University of York.

O'Reilly, D., & Marsden, E. (2021). Eliciting and measuring L2 metaphoric competence: Three decades on from low (1988). *Applied Linguistics* 42(1), 24–59. https://doi.org/10.1093/applin/amz066

O'Reilly, D., & Yan, L. (2025). Playing with second language metaphor: An exploration with advanced Chinese learners of English. *Applied Linguistics*, 46(1), 53–74. https://doi.org/10.1093/applin/amad067.

Onysko, A. (2016). Enhanced creativity in bilinguals? Evidence from meaning interpretations of novel compounds. *International Journal of Bilingualism*, 20(3), 315–334. https://doi.org/10.1177/1367006914566081.

OpenAI. (2025). ChatGPT (GPT-4o) [Large language model]. https://chat.openai.com/

Park, J., & Chon, Y. V. (2019). EFL learners' knowledge of high-frequency words in the comprehension of idioms: A boost or a burden? *RELC Journal*, 50(2), 219–234. https://doi.org/10.1177/0033688217748024.

Parkinson, D., & Francis, B. (Eds.). (2006). *Oxford idioms dictionary for learners of English* (2nd rev. ed.). Oxford University Press.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–225). Longman.

Plonsky, L., & Derrick, D. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, 100(2), 538–553. https://doi.org/10.1111/modl.12335.

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, 102(4), 713–731.

R Core Team (2025). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/

Reuterskiöld, C., & Van Lancker Sidtis, D. (2013). Retention of idioms following one-time exposure. *Child Language Teaching and Therapy*, 29(2), 219–231. https://doi.org/10.1177/0265659012456859

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3–13.

https://doi.org/10.1017/langcog.2025.10029 Published online by Cambridge University Press

Siyanova-Chanturia, A. (2013). Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon*, 8, 245–268. https://doi.org/10.1075/ml.8.2.06siy.

Skoufaki, S. (2008). Investigating the source of idiom transparency intuitions. *Metaphor Symbol*, 24(1), 20–41.

Soto-Sierra, V., & Ferreira, R. A. (2024). The influence of cross-language similarity and transparency on idiom knowledge in non-immersed L2 speakers. *System*, 122, 1–11. https://doi.org/10.1016/j.system.2024.103287.

Steinel, M. P., Hulstijn, J. H., & Steinel, W. (2007). Second language idiom learning in a paired-associate paradigm: Effects of direction of learning, direction of testing, idiom imageability, and idiom transparency. *Studies in Second Language Acquisition*, 29(3), 449–484. https://doi:10.1017/S0272263107070271

Titone, D., Columbus, G., Whitford, V., Mercier, J., & Libben, M. (2015). Contrasting bilingual and monolingual idiom processing. In R. R. Heredia & A. B. Cieślicka (Eds.), *Bilingual figurative language processing* (pp. 171–207). Cambridge University Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Vasiljevic, Z. (2016). Effects of etymology and pictorial support on the retention and recall of L2 idioms. *Electronic Journal of Foreign Language Teaching*, 12(1), 35–55.

Voeten, C. C. (2022). buildmer: Stepwise elimination and term reordering for mixed-effects regression (Version 2.3) [R package]. https://CRAN.R-project.org/package=buildmer

Wang, X., Boers, F., & Warren, P. (2024). Prompting language learners to guess the meaning of idioms: Do wrong guesses linger? *Language Awareness*, 33(1), 94–116. https://doi.org/10.1080/09658416.2022.2153859.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5), 2020–2045. https://doi.org/10.1037/xge0000014.

Winter, B. (2019). *Statistics for linguists: An introduction using R (1st ed.)*. Routledge. https://doi.org/10.4324/9781315165547

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

Cite this article: O'Reilly, D., Onysko, A., Rasse, C., Papitsch, L., Colston, H., & van der Horst, I. (2025). High ceilings and ingenuine allies: tapping into the idiom meaning knowledge of first and second language speakers of English, *Language and Cognition*, 17, e72, 1–30. https://doi.org/10.1017/langcog.2025.10029