



Knowledge is a river and education is like a stairway: An eye movement study on how L2 speakers process metaphors and similes

Henri Olkonieni^{1,2} , Raymond Bertram²  and Johanna K. Kaakinen^{2,3} 

¹Division of Psychology, Faculty of Education, University of Oulu, Oulu, Finland; ²Department of Psychology and Speech-Language Pathology, University of Turku, Turku, Finland and ³INVEST Research Flagship, University of Turku, Turku, Finland.

Research Article

Cite this article: Olkonieni H, Bertram R, Kaakinen JK (2022). Knowledge is a river and education is like a stairway: An eye movement study on how L2 speakers process metaphors and similes. *Bilingualism: Language and Cognition* **25**, 307–320. <https://doi.org/10.1017/S1366728921000869>

Received: 23 October 2020
Revised: 20 August 2021
Accepted: 22 August 2021
First published online: 8 November 2021

Keywords:

eye movements; metaphor; similes; reading; bilingualism

Address for correspondence:

Henri Olkonieni,
Email: henri.olkonieni@oulu.fi

Abstract

Very little is known about the processes underlying second language (L2) speakers' understanding of written metaphors and similes. Moreover, most of the theories on figurative language comprehension do not consider reader-related factors. In the study, we used eye-tracking to examine how native Finnish speakers ($N = 63$) read written English nominal metaphors ("education is a stairway") and similes ("education is *like* a stairway"). Identical topic-vehicle pairs were used in both conditions. After reading, participants evaluated familiarity of each pair. English proficiency was measured using the Bilingual-language Profile Questionnaire and the Lexical Test for Advanced Learners of English. The results showed that readers were more likely to regress within metaphors than within similes, indicating that processing metaphors requires more processing effort than processing similes. The familiarity of a metaphor and L2 English proficiency modulated this effect. The results are discussed in the light of current theories on figurative language processing.

1. Introduction

Metaphors are an essential and necessary part of human communication (Ortony, 1975). Metaphors refer to one thing by mentioning another as in 'knowledge is a river' and are used very frequently by native (L1) speakers. For example, already 20–50% of conversational content of kindergarten-aged children is metaphorical (Billow, 1981). However, recent research has shown that there are individual differences in processing metaphoric language (e.g., Olkonieni, Ranta & Kaakinen, 2016; Pierce, MacLaren & Chiappe, 2010), and one source of these differences is the level of language proficiency (e.g., Heredia & Cieślicka, 2016; Jankowiak, Rataj & Naskręcki, 2017; Vaid, López & Martínez, 2015). Language proficiency already widely diverges among natives, but is even more divergent among second language (L2) speakers. Yet over half of the world's population speaks more than one language (e.g., Bialystok, Craik & Luk, 2012), and many persons live and work in environments in which the non-native language needs to be used on a daily basis. Operating in daily life in one's non-native language may be challenging in general, but may come with specific costs in case of metaphor processing; for example, it has been shown that international students in UK often do not understand metaphors used in lectures leading to misunderstanding of important concepts (Littlemore, Chen, Koester & Barden, 2011). Thus, it is important to know how L2 speakers process figurative language, such as similes and metaphors, and at the same time to determine to what extent this processing depends on L2 proficiency. These are the objectives of the present article.

1.1 Processing of metaphors and similes

A metaphor typically takes the form *X is Y* (i.e., nominal or attributive metaphor; Cacciari & Glucksberg, 1994; Glucksberg, 1998): for example, *education is a stairway*. A metaphor expresses a relationship of its constituent parts called a topic (*education*) and a vehicle (*stairway*). In similes, in addition to the topic and the vehicle, the comparative word *like* is included (*X is like Y*): for example, *education is like a stairway*.

Traditionally it has been thought that when comprehending figurative expressions, such as metaphors and similes, the reader must first reject a non-fitting literal interpretation (i.e., education cannot be an actual stairway) and seek a non-literal interpretation that is appropriate to the text context (THE STANDARD PRAGMATIC VIEW, e.g., Grice, 1975). Moreover, it has been suggested that to comprehend metaphors a transformation into similes is needed, to match the features presented by the topic and vehicle (THE COMPARISON VIEW, e.g., see Glucksberg, 1998, 2003 for review). This view predicts that similes are processed faster than metaphors, because comprehending metaphors requires an additional transformation stage. Experimental

© The Author(s), 2021. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

evidence has lent support to these traditional theories. Studies have shown that metaphorical phrases take longer to process than their literal counterparts (e.g., Olkonemi et al., 2016; cf. Inhoff, Lima & Carroll, 1984) and that similes are faster to process than metaphors (e.g., Ashby, Roncero, de Almeida & Agauas, 2018; Gregory & Mergler, 1990; Shibata, Toyomura, Motoyama, Itoh, Kawabata & Abe, 2012). For example, Ashby et al. (2018) found that native English speakers spent longer processing metaphors than similes and this difference appears almost immediately during first-pass reading on the vehicle. The processing time results are supported by findings showing that people are more likely to agree that there are similarities between the topic and vehicle in similes than in metaphors (Gregory & Mergler, 1990; see also Aisenman, 1999; Zharikov & Gentner, 2002). Brain imaging studies have shown that processing of similes recruits different brain areas from processing of metaphors (e.g., Shibata et al., 2012), suggesting that comprehending metaphors is a more complex process than comprehending similes.

However, there is also evidence showing that processing of metaphors does not always take longer than processing of similes (see Glucksberg, 1998, 2003 for review). Moreover, phrases that are familiar or often used as metaphors are easier to process and comprehended than unfamiliar metaphorical phrases (e.g., Ashby et al., 2018; Blasko & Briehl, 1997; Columbus et al., 2015; Pexman, Ferretti & Katz, 2000) and there are individual differences in the ease of comprehending metaphors (e.g., Columbus et al., 2015; Heredia & Cieřlicka, 2016; Olkonemi et al., 2016). It has also been suggested that metaphor comprehension is mediated by aptness – the ease at which readers can apply the properties of a metaphor's vehicle to its topic (e.g., Jones & Estes, 2006). However, aptness falls out of the scope of this article, as here we focus on familiarity of metaphors and L2 proficiency.

In line with the findings reported above, recent theories propose that the familiarity of the phrase in its figurative meaning and individual differences between readers affect the comprehension of figurative statements. Several theoretical accounts actually suggest that the familiarity of the phrase as a metaphor is crucial in how easy it is to comprehend it (e.g., Bowdle & Gentner, 2005; Giora, 2003; Glucksberg, 2003), among them THE GRADED SALIENCE VIEW (Giora, 2003). The view states that the salience of a phrase, or familiarity, is crucial in how quickly its meaning is accessed. Typically, the literal meaning of a phrase is more salient and is accessed first. If the literal interpretation does not fit into the context, the alternative metaphoric meaning is searched for. This processing procedure would, thus, result in longer processing times for metaphoric than literal phrases. However, if the phrase is familiar and/or often used in its metaphorical meaning, the metaphorical meaning might be more salient and therefore be accessed first. In this case the time-consuming processes of first assessing and then rejecting the literal meaning can be skipped.

THE CAREER OF METAPHOR THEORY (Bowdle & Gentner, 2005) also emphasizes that phrases that are commonly used in everyday language in their metaphorical meaning (i.e., conventional metaphors) are processed differently from novel and unfamiliar metaphors, and similes. According to this view, similes are processed as direct comparisons to establish a relationship between the topic and the vehicle. As for metaphors, the assumption is that novel metaphors are processed similarly to the similes (i.e., as comparisons), whereas highly familiar conventional metaphors are comprehended by classifying the topic as a member of a category named by the vehicle. For example, in the metaphor *X is a real drama queen*, drama queen can refer to both its concrete

meaning 'a member of the royal family acting in a drama' and a more abstract categorical (or metaphoric) meaning 'somebody acting out'. As categorizing is assumed to be less taxing than a comparison, processing of conventional metaphors (e.g., 'drama queen' when referring to someone's dramatic behavior) can be as fast or even faster than processing of literal expressions (e.g., 'drama queen' when referring to a member of a royal family) or similes, which always require a comparison. These theoretical views are supported by experimental evidence showing that unfamiliar metaphors are slower to process than familiar metaphors (e.g., Ashby et al., 2018; Blasko & Briehl, 1997; Columbus et al., 2015; see also Raney, Campbell & Bovee, 2014). In a recent eye-tracking study by Ashby et al. (2018) participants read nominal metaphors and similes without text context. Ashby et al. showed that familiar metaphors and similes were read faster than unfamiliar ones. There was also an interaction between familiarity and figurative language type – unfamiliar similes were read faster than unfamiliar metaphors, but this reading time difference was smaller for familiar items.

Even though the theoretical views introduced above provide accounts of how certain textual features impact comprehension of metaphors and similes, none of them provides a view that could accommodate individual differences in the processing and comprehending of metaphors and similes. One theory that can be used to explain how individual differences impact comprehension is THE PREDICTIVE CODING THEORY (Fabry, 2021; Friston, Kilner & Harrison, 2006). The theory is based on the assumption that all biological systems minimize the amount of free energy, which is the amount of work that can be extracted from the system (FREE ENERGY PRINCIPLE, e.g., Friston et al., 2006). To minimize free energy, systems aim to minimize prediction error, which is the key element of this theory from the perspective of figurative language processing (Fabry, 2021). Perception, cognition, and emotion are then realized by the continuous minimization of prediction error. The predictive coding theory has been already used to explain lower level language processes such as word processing and word learning (e.g., Lupyan & Clark, 2015; Ylinen, Bosseler, Junttila & Huotilainen, 2017), but also higher-level language processes such as irony comprehension (Fabry, 2021). Upon encountering figurative expressions, in most of the situations literal language use is expected. Consequently, use of a figurative expression (e.g., *education is a stairway*) is not expected and it results in prediction error, which requires then a corrective process to form a more suitable interpretation. The theory assumes that the more familiar a metaphor is, the more predictable its figurative interpretation is, which would result in reduced prediction error. Moreover, it predicts that similes are faster to process than metaphors, as the comparison in similes is made explicit and, thus, reduces prediction error. Individual differences in the ease and efficiency of resolving metaphorical meaning of phrases could also be associated with differences in prediction error. For example, high language proficiency should lead to smaller prediction error when encountering metaphors, as a proficient reader has probably encountered them more often than a less proficient reader and has therefore its figurative meaning available as a potential interpretation for the phrase.

1.2. Bilingualism and figurative language use

Figurative language use is notably difficult for L2 speakers. As a non-native speaker one may misinterpret figurative language expressions such as metaphors or idioms due to a too literal

interpretation. In their article on processing of metaphoric reference by Spanish–English bilinguals, Heredia and Cieřlicka (2016, p. 1) cite a professor saying “Back then in the late 1980s, I used to write poetry, now I only write sleeping pills” to which a bilingual student responded with “You mean prescription drugs”. The literal interpretation prevented the student from grasping the intended meaning, which is that the professor stated that his nowadays writings are not all that interesting anymore. This anecdotal evidence is supported by experimental studies. For instance, Cieřlicka, Heredia and Garcıa (2017) found that Spanish–English bilinguals dominant in Spanish only activated the literal meaning in an implicit lexical decision task for expressions that could be interpreted idiomatically as well (e.g., ‘kick the bucket’). More typical is that figurative language is understood – in comparison to L1 speakers – with a delay. Siyanova-Chanturia, Conklin and Schmitt (2011) found that L2 speakers processed idioms at a similar speed as novel phrases and slower than literal phrases, whereas L1 speakers processed idioms faster than novel phrases and equally fast as literal ones. To explain these findings, it has been suggested that natives bypass literal interpretations of words but instead map words or word sequences directly onto their idiomatic meaning (Boers & Lindstromberg, 2012; cf. Bowdle & Gentner, 2005). In contrast, and in line with the cross-modal priming results of Cieřlicka (2006), L2 speakers would – at least initially – activate literal meanings of composite words in an idiomatic expression. Given that activation of literal meanings will lead the reader/listener astray (as for instance in a phrase like: “at the end of the day, what really matters is ...”), there will be a need for reinterpretation, which in turn will lead to longer processing times. However, despite the processing time differences of idiomatic expression between monolinguals and bilinguals, bilinguals are also sensitive to the familiarity of the idiomatic expression (Titone, Columbus, Whitford, Mercier & Libben, 2015).

L2 language proficiency influences the ease of processing figurative language in L2. Heredia and Cieřlicka (2016) examined how language dominance affects reading times of metaphoric vs literal sentences in a text context. They found that metaphors were overall slower to read than their literal counterparts, but language dominance modulated the reading times. Participants who read in their non-dominant language showed a larger reading time difference between metaphoric and literal sentences than balanced bilinguals or readers reading in their dominant language. Similar findings have been reported in other figurative language domains. Cieřlicka, Heredia and Olivares (2014) found in an eye movement study that idiom processing is less efficient for non-dominant in comparison to dominant bilinguals. Also, in the realm of sarcasm, differences – as a function of L2 proficiency – have been found. For instance, Tiv, Rouillard, Vingron, Wiebe and Titone (2019) found that greater global second language proficiency was linked to greater use of sarcasm in daily life.

As far as we are aware, there is no study that directly pits the processing of similes against metaphors within L2 speakers. The one bilingual study that considers both means of figurative language use is the study of Harris and Mosier (1999) on bilingual memory. Yet, processing of similes and metaphors by L2 speakers would be interesting to investigate, as there may be an intricate interplay between the use of similes vs metaphors (education is like a stairway vs. education is a stairway), level of proficiency, and familiarity of the figurative expression. Moreover, the time course of processing the topic and vehicle is of interest, as both are linked to each other through comparison and the resolution of this comparison may again be modulated by familiarity and

L2 proficiency. The current study addresses these issues by means of an eye movement experiment.

1.3. Overview of the present study

In the present study, we used the same materials as Ashby et al. (2018) to investigate how familiarity of the phrases as metaphors and English L2 proficiency influence processing of metaphors and similes using eye-tracking. By using eye-tracking methodology, it is possible to tap into moment-to-moment cognitive processing during reading (Rayner, 1998, 2009). That is, eye-tracking offers an excellent method to assess the detailed time-course of processing written language, which is especially important when assessing linguistically long-distance dependencies as is the case in metaphors and similes. The metaphors and similes used in this study were also rated for familiarity by L2 speakers.

Based on the literature, we expect three main effects. First, we expect an effect of familiarity: the more familiar an expression is, the faster it will be processed, both in case of metaphors and similes. Second, a main effect of proficiency is to be expected, with more proficient L2 readers showing faster processing times. Third, by virtue of explicitness of the comparison, we also expect that similes will be processed faster than metaphors. In terms of interactions, we hypothesize that less proficient L2 readers benefit more from explicitness of comparison than proficient L2 readers. Hence, we predict a larger processing time difference between similes and metaphors for less proficient in comparison to more proficient L2 readers. However, this effect may be cancelled out for very unfamiliar similes and metaphors. These hypotheses are in line with the predictive coding theory described above (Fabry, 2021; Friston et al., 2006).

2. Method

2.1. Participants

Sixty-three University of Turku students (59 women), between ages of 19–44 ($M_{age} = 25$, $SD_{age} = 6$), participated in the study to fulfill a course requirement. All were native speakers of Finnish and had normal or corrected-to-normal vision.

The study was conducted in accordance with the Declaration of Helsinki, and each participant gave their written consent. The Ethics Committee for Human Sciences at the University of Turku approved the study.

2.2. Apparatus

Eye movements were recorded monocularly using EyeLink 1000 (SR Research Ltd., Ontario, Canada) at 1000 Hz sampling frequency. The stimuli were presented on a 24” BenQ XL2411-monitor with refresh rate of 100 Hz and a resolution of 1920 × 1080 pixels. Participants were seated 70 cm from the screen, and a chin-and-forehead rest was used to stabilize the head.

2.3. Materials

Each participant read a total of 124 sentences on a computer screen (font: Courier New, font size: 14) while their eye movements were recorded. Sixty-two of the sentences were metaphor and simile pairs of the form “X is Y” or “X is like Y”, and there were 62 filler items. Items in the simile and metaphor conditions were identical, except that the word *like* appeared in the simile

condition, an example pair is presented in Table 1. All items were the same as in the Ashby et al. (2018) study, and they were of moderate aptness ($M = 6.3$ on a scale from 1: not all apt to 10: very apt, see Ashby et al., 2018 for more details).

Each participant saw half of the topic–vehicle pairs in the simile condition and half in the metaphor condition. The target phrase was followed by an explanation of the metaphor or simile (i.e., “they are big and strong.”), so that the vehicle would not appear in a sentence-final position. After reading, participants answered a yes/no comprehension question for 25% of the sentences. The answer accuracy to comprehension questions was high ($M = 91.22\%$, $SD = 28.31\%$).

FAMILIARITY RATINGS. After completing the reading task, each participant evaluated how familiar the topic-vehicle pairs were to him/her as metaphors (“How often have you encountered the following metaphors”). Topic-vehicle pairs were presented to participants on a computer screen one at a time in the form “X is Y” (e.g., “wrestlers are gorillas”), and the familiarity of each was evaluated on a scale from 1 (not at all familiar) to 7 (very familiar). Mean familiarity score of metaphors was 3.02 ($SD = 2.10$).

2.4. Language proficiency measures

THE LEXICAL TEST FOR ADVANCED LEARNERS OF ENGLISH (LexTale; Lemhöfer & Broersma, 2011) was used to assess vocabulary knowledge of the participants. In the task participants were shown 60 letter strings. Most of them were existing English words ($n = 40$, e.g., ‘bewitch’) and some of them non-words ($n = 20$, e.g., ‘kermshaw’). They were asked to indicate for each item whether it was an existing English word or not, by pressing either the *yes* or *no* key from the keyboard. The order of items was fixed, such that no more than five words or nonwords appeared in a row. Several possible methods to score the task exists (Lemhöfer & Broersma, 2011). In the current study, the test was scored by summing all the correct answers together. This scoring scheme was selected over the corrected percentage score suggested by Lemhöfer and Broersma (2011) as the sum score correlated more strongly with the other language proficiency assessments. The score can vary between 0 and 60 points. For descriptive statistics, see Table 2.

THE FINNISH–ENGLISH BILINGUAL LANGUAGE PROFILE questionnaire (adapted from the English–French version of Birdsong, Gertken & Amengual, 2012) is a questionnaire to assess language dominance and English and Finnish language proficiency. The English language proficiency section used in the present experiment consists of four questions (i.e., How well do you *speak/understand/read/write* English?), which participants rated on a scale from 0 (= not well) to 6 (= very well). The score can vary from 0 to 24 points, higher scores indicating higher self-evaluated English proficiency. Descriptive statistics of the scores are presented in Table 2.

A revised version of the **AUTHOR RECOGNITION TEST** (Acheson, Wells & MacDonald, 2008) was used to evaluate print exposure to literature. The test is a paper-and-pencil test which consists of a list of 65 real authors (e.g., Ernest Hemingway) and 65 foils (e.g., John Landau). Each correctly marked author is awarded with one point, and for each incorrectly indicated author one point is reduced. The maximum score is 65 points, with higher scores indicating better knowledge about who is an author and who is not. The ART score was eventually left out from the composite measure reported below as it correlated weakly with both the LexTale and English proficiency scores (see Table 2), and recent studies indicate that ART may not be as informative in

Table 1. An example of simile and metaphor pair used in the experiment and their respective analysis regions.

| Type | | Vehicle | Spillover | Explanation |
|-----------------|-------------------------------|----------|-----------|--------------------------|
| Metaphor | Brian says wrestlers are | gorillas | because | they are big and strong. |
| Simile | Brian says wrestlers are like | gorillas | because | they are big and strong. |

assessing exposure to text for L2 speakers as it is for L1 speakers (e.g., McCarron & Kuperman, 2021). Descriptive statistics of the tasks are presented in Table 2.

THE FINAL COMPOSITE SCORE OF ENGLISH PROFICIENCY used in the main analyses was calculated by taking the mean of the standardized values (z-score) of the LexTale vocabulary scores and the self-rated English Proficiency scores. The values were standardized to give both measures equal weight in the analysis. The use of self-ratings as such is motivated by the fact that they are often used to assess proficiency level of L2 speakers and correlate with more objective measures (Marian, Blumenfeld & Kaushanskaya, 2007). However, they nevertheless may over- or underestimate L2 speakers’ proficiency level and are therefore not to be used as a unique index of language proficiency (Gollan, Weissberger, Runnqvist, Montoya & Cera, 2012). The vocabulary test, on the other hand, is an objective measure which taps into vocabulary knowledge, which in turn strongly correlates with other language proficiency dimensions like L2 speech fluency (Hilton, 2008), L2 reading comprehension (Jeon & Yamashita, 2014) and L2 listening comprehension (Zhang & Zhang, 2020). Among different kinds of vocabulary tests LexTale has become increasingly popular which is underlined by the development of variants of LexTale in multiple languages (Mandarin Chinese, Dutch, English, Finnish, French, German, Italian, and Spanish; see Salmela, Lehtonen, Garusi & Bertram, 2021, for a more detailed discussion). The current composite measure includes therefore adequate and common proxies of language proficiency and is in line with the recommendation of De Bruin, Carreiras and Duñabeitia (2017) to combine objective measures of language proficiency with subjective ones.

2.5. Procedure

Participants were tested individually. They were naïve to the purpose of the experiment: upon arrival, they were only informed that the experiment was about reading in English. The specific purpose of the task was explained to the participants only after the experiment. Each participant signed an informed consent form before experimentation. Subsequently, the eye-tracker was set up and each participant was calibrated using a three-point calibration screen. Participants were instructed to read each sentence at their own pace and were told to press the spacebar on the keyboard when they wanted to move from one sentence to another. Participants answered a yes/no comprehension question after 25% of the sentences, by pressing designated buttons on a keyboard. After the participant had answered the question, the next sentence was presented.

The reading task was followed by familiarity rating of each metaphor, LexTale, ART, and the Bilingual Language Profile

Table 2. Descriptive statistics of the language proficiency measures and correlations between the measures

| Measure | <i>M</i> | <i>SD</i> | <i>Min</i> | <i>Max</i> | <i>r</i> [95% CI] | |
|---------------------|----------|-----------|------------|------------|-------------------|----------------|
| | | | | | ART | LexTale |
| ART | 9.67 | 7.73 | 1 | 40 | | |
| LexTale | 46.18 | 7.17 | 35 | 60 | .36 [.35, .37] | |
| English Proficiency | 17.27 | 3.61 | 6 | 24 | .002 [−.01, .01] | .46 [.45, .47] |

Note. ART = Author Recognition Task, LexTale = Lexical Test for Advanced Learners of English; English Proficiency = Bilingual-language Profile Questionnaire, English Language Proficiency Score

questionnaire. Experimental sessions lasted for about 45–60 minutes.

3. Results

Fixations shorter than 50 ms were either merged with a nearby fixation (if the distance between the fixations was $< 1^\circ$) or removed from the data. Three different reading time measures were computed for the VEHICLE, SPILLOVER REGION, and EXPLANATION REGION from the eye movement data: gaze duration, go-past time, and regression out proportion. GAZE DURATION is the duration of all fixations on the target region until the eyes fixate away from it either to the left or right. GO-PAST TIME is the total duration of all fixations that occurred from the first fixation on a target region until the target region was exited in a progressive manner. REGRESSION OUT PROPORTION is the likelihood of reader returning to already read text regions from the target region. The measures reported are the same as in Ashby et al. (2018). Using these measures in combination with the analysis of different regions allows for direct comparison with the results of that study alongside a detailed assessment of the time course of simile and metaphor processing. Gaze duration on the vehicle informs about initial processing, whereas go-past time and proportion of regressions inform about how well the vehicle can be integrated with the topic (the subject of the simile/metaphor). Effects in the spillover region inform to what extent metaphor processing lingers on and effects in the explanation region reflect how well the initially activated interpretation of the simile/metaphor can be integrated with the explicit explanation.

The reading time measures were skewed and consequently transformed. The best fitting transformation was selected to normalize the measures; all the reading time measures were logarithmically transformed, except gaze duration on explanation region was square-root transformed. Observed means and standard deviations of the eye movement measures are presented in Table 3.

3.1. Statistical analyses

Data were analyzed with linear mixed-effects models (LMM) using the *lme4* package (Bates, Mächler, Bolker & Walker, 2015) in the R statistical software (Version 4.0.1; R Core Team, 2020). Separate models were built for each eye movement measure for each text region. It has been recommended that only minimal data filtering would be conducted when analyzing data with LMMs, applied with model criticism (Baayen & Milin, 2010). We compared models using non-filtered reading times to filtered models, in which reading times > 2.5 SD were excluded, and R^2 values were either better for non-filtered models or virtually

equal to filtered models, favoring the use of non-filtered values. Consequently, no filtering for the data was applied.

Text type (simile vs. metaphor) was fitted to each model as a deviation coded fixed effect variable. English proficiency and familiarity rating of the topic-vehicle pairs were continuous variables and were added to the models as centered fixed effects variables. Correlation between English proficiency and familiarity rating was low, $r = .05$, 95% CI = [.03, .06]. Participants and items were entered to the models as random intercepts (Baayen, Davidson & Bates, 2008). The maximal random structure was fitted to the model (Barr, Levy, Scheepers & Tily, 2013). If the model failed to converge with the full random structure, the random structure of the model was trimmed top-down, starting with correlations between factors (see Brauer & Curtin, 2018). Two models failed to converge with only random intercepts for participants and items fitted in the model (i.e., models on regression out proportion from spillover region and go-past time on explanation region); consequently, non-significant interaction terms between the fixed effects were removed from the model, starting with smallest t or z value.

The exact degrees of freedom are difficult to determine for the t -statistics estimated by LMMs, leading to problems in determining exact p -values (Baayen et al., 2008). Consequently, degrees of freedom or p -values are not reported; statistical significance at the .05 level is indicated by values of the $|t|$ or $|z| > 1.96$. For the sake of brevity, only significant effects are reported in the text. The final models are reported in the Appendix A Tables A1–A9. Data and analysis code are available at <https://osf.io/jmekh/>.

3.2. Vehicle region

The model for GAZE DURATION ON THE VEHICLE revealed a main effect of readers' English proficiency, $\beta = -0.11$, 95% CI [−0.18, −0.05], $t = -3.42$. The result indicated that as English proficiency of the reader increased the gaze duration on vehicle decreased. The model for GO-PAST TIME ON VEHICLE showed two main effects. First, there was a main effect of sentence type, indicating that go-past time on similes was shorter than on metaphors, $\beta = 0.12$, 95% CI [0.08, 0.16], $t = 6.15$. Second, there was a main effect of English proficiency, indicating that as English proficiency of the reader increased, the go-past time on vehicle decreased, $\beta = -0.13$, 95% CI [−0.20, −0.05], $t = -3.38$. Last, the model for REGRESSION OUT PROPORTION FROM THE VEHICLE showed main effects of sentence type and familiarity. First, readers were less likely to regress from the simile than from the metaphor, $\beta = 0.74$, 95% CI [0.54, 0.93], $z = 7.45$. Second, as familiarity of the metaphor increased, the likelihood of regressing from the vehicle decreased, $\beta = -0.14$, 95% CI [−0.25, −0.03], $z = -2.49$.

Table 3. Means and standard deviations of the eye movement measures

| Region | Measure | Simile | | Metaphor | |
|-------------|-----------------------------|----------|-----------|----------|-----------|
| | | <i>M</i> | <i>SD</i> | <i>M</i> | <i>SD</i> |
| Vehicle | Gaze duration | 346 | 212 | 359 | 222 |
| | Go-past time | 406 | 324 | 470 | 361 |
| | Regression out (proportion) | 0.10 | 0.30 | 0.18 | 0.38 |
| Spillover | Gaze duration | 268 | 111 | 265 | 106 |
| | Go-past time | 321 | 251 | 347 | 272 |
| | Regression out (proportion) | 0.08 | 0.27 | 0.13 | 0.33 |
| Explanation | Gaze duration | 1415 | 694 | 1370 | 704 |
| | Go-past time | 2077 | 1441 | 2065 | 1440 |
| | Regression out (proportion) | 0.51 | 0.50 | 0.50 | 0.50 |

3.3. Spillover region

The model for GAZE DURATION ON THE SPILLOVER REGION revealed a main effect of familiarity, $\beta = -0.02$, 95% CI [-0.03, -0.002], $t = -2.22$. The result indicates that as familiarity of a metaphor increased the gaze duration on spillover region decreased.

The model for GO-PAST TIME ON THE SPILLOVER REGION revealed main effects of sentence type, familiarity, and English proficiency. First, similes were faster to process than metaphors, $\beta = 0.06$, 95% CI [0.02, 0.09], $t = 3.42$. Second, when the familiarity of a metaphor increased, the go-past time on spillover region got faster, $\beta = -0.03$, 95% CI [-0.05, -0.01], $t = -3.43$. Similarly, go-past reading time on spillover region decreased, when the English proficiency of reader increased, $\beta = -0.07$, 95% CI [-0.13, -0.02], $t = -2.54$. There was no indication of an interaction between sentence type and familiarity or English proficiency. However, there was a three-way interaction between sentence type, familiarity, and English proficiency, $\beta = -0.05$, 95% CI [-0.08, -0.01], $t = -2.86$ (see Figure 1). The result indicated that for readers with relatively low English proficiency, go-past reading time of similes on the spillover region (i.e., because) was faster than go-past reading time of metaphors as the familiarity increased. In contrast, for readers with relatively high English proficiency, go-past time of similes on the spillover region was slower than go-past time of metaphors, when the familiarity was low, but the difference between similes and metaphors decreased as the familiarity increased.

Last, the model for REGRESSION OUT PROPORTION FROM THE SPILLOVER REGION showed main effects of sentence type, familiarity, and English proficiency. First, readers were more likely to regress from the spillover region for metaphors than similes, $\beta = 0.59$, 95% CI [0.37, 0.81], $z = 5.16$. Second, the likelihood of regressing from the spillover region decreased as the familiarity of the metaphor increased, $\beta = -0.18$, 95% CI [-0.32, -0.04], $z = -2.47$. Similarly, the likelihood of regressing from the spillover region decreased as the English proficiency of the reader increased, $\beta = -0.27$, 95% CI [-0.53, -0.01], $z = -2.02$.

3.4. Explanation region

The model for GAZE DURATION ON THE EXPLANATION REGION revealed main effects of sentence type and English proficiency. The result indicated that gaze duration on the explanation region was shorter

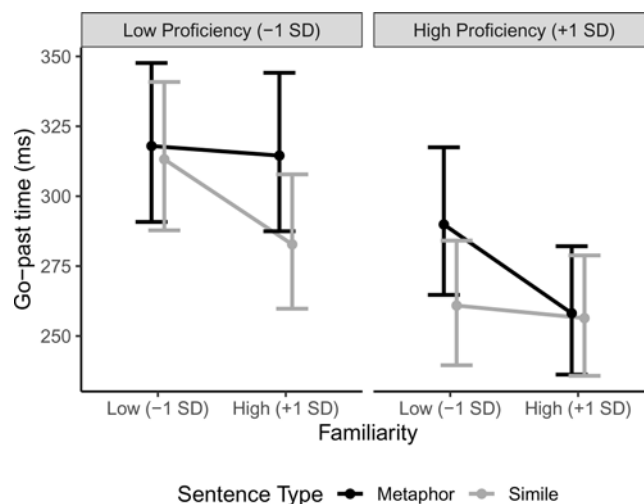


Fig. 1. Model estimates for the go-past time on the spillover region. For illustration purposes, English Proficiency and Familiarity are divided into High and Low (± 1 SD), and the model means and confidence intervals are back-transformed from log values. Error bars represent 95% CI.

for metaphors than similes, $\beta = -0.75$, 95% CI [-1.18, -0.33], $t = -3.46$. In addition, gaze duration decreased as English proficiency of the reader increased, $\beta = -1.77$, 95% CI [-3.03, -0.52], $t = -2.77$. Moreover, the model showed an interaction between sentence type and familiarity, $\beta = 0.44$, 95% CI [0.01, 0.87], $t = 1.99$; indicating that readers showed longer gaze durations on explanation regions that followed similes than metaphors when the familiarity was relatively low, but the effect wore off as the familiarity increased (see Figure 2). This effect was confirmed by follow-up comparisons, in which sentence type was fitted to the same model (see Table A7), but with both similes and metaphors fitted separately as baseline condition (i.e., treatment coding). The model fitted with similes as baseline showed an effect of familiarity, indicating that gaze durations on the explanation region decreased as familiarity increased, $\beta = -0.40$, 95% CI [-0.76, -0.04], $t = -2.61$. In contrast, the model with metaphors set as baseline showed no effect of familiarity, $\beta = 0.04$, 95% CI [-0.31, 0.39], $t = 0.22$.

The model for GO-PAST TIME ON THE EXPLANATION REGION revealed main effects of familiarity and English proficiency. The result

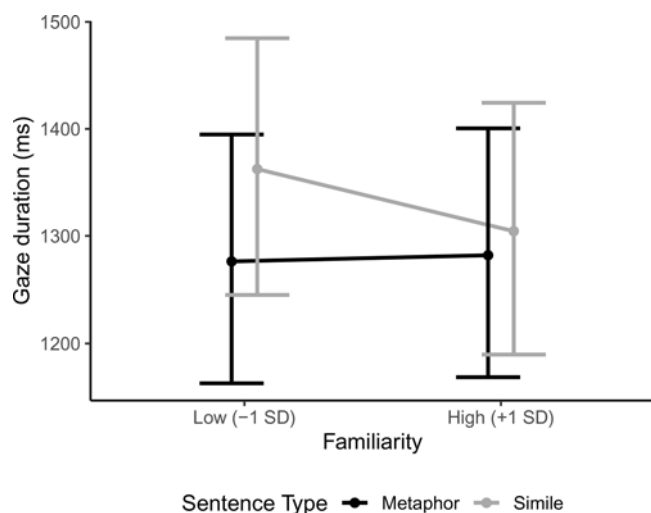


Fig. 2. Model estimates for the gaze duration on the explanation region. For illustration purposes, Familiarity is divided into High and Low (± 1 SD), and the model means and confidence intervals are back-transformed from square-root values. Error bars represent 95% CI.

indicated that go-past time on the explanation region decreased as the familiarity of the metaphor increased, $\beta = -0.03$, 95% CI [-0.05, -0.02], $t = -3.96$. Similarly, go-past time got faster as the English proficiency of the reader increased, $\beta = -0.14$, 95% CI [-0.25, -0.03], $t = -2.45$. Last, the model for REGRESSION OUT PROPORTION FROM THE EXPLANATION REGION revealed a main effect of familiarity, $\beta = -0.16$, 95% CI [-0.25, -0.07], $z = -3.58$, indicating that readers' likelihood of regressing decreased as the familiarity of a metaphor increased.

4. Discussion

The purpose of the present eye movement experiment was to investigate L2 reading of similes vs. metaphors while at the same time exploring the role of metaphor familiarity and L2 English proficiency.

The results showed that similes were read faster than their metaphoric counterparts. This replicates the previous findings (e.g., Ashby et al., 2018), and is in line with theories proposing similes to be faster to process than novel metaphors (e.g., Bowdle & Gentner, 2005). One possible explanation for this finding is that the word "like" makes the figurative expression more predictable by making the comparison explicit (e.g., Fabry, 2021). In contrast, for nominal metaphors, the X is Y form doesn't explicitly signal figurative meaning, making metaphors harder to predict (i.e., increasing the amount of prediction error) and resulting in longer reading times. Explicit signaling of the relationship between topic and vehicle – as in similes – may play an even more important role in the present experiment, where metaphors and similes were presented in single sentences without preceding context. Some previous studies have shown that the amount of previous supporting contextual information affects the ease of processing of metaphors (Inhoff et al., 1984), arguably because supporting context reduces the amount of prediction error. Although, metaphors and similes sometimes exist without context: for example, in social media, literature, and advertisements, they more typically appear within a larger context. More studies are needed to explore the role of contextual

support in processing differences between similes and metaphors in detail, and to further explore theoretical assumptions why similes are faster to process than metaphors.

Second, the results showed that the familiarity of the topic-vehicle pair as metaphor sped up reading times of both metaphors and similes. The result replicates previous findings (e.g., Ashby et al., 2018), and supports theoretical accounts, such as the graded salience view (e.g., Giora, 2003) and the career of metaphor theory (Bowdle & Gentner, 2005), emphasizing the role of familiarity in processing figurative language. Findings on the role of familiarity in the processing of figurative expressions have been contradictory though (see Olkonien & Kaakinen, 2021, for a review of findings related to irony). One reason for this is that some researchers use the terms conventionality and familiarity interchangeably, whereas others consider these two variables to be distinct (Bowdle & Gentner, 2005; Gentner & Wolff, 1997; Jones & Estes, 2006; Roncero & de Almeida, 2015); moreover, different authors have different definitions for these variables. Another possible reason is that ratings on the familiarity of expressions are typically provided by different participants than those in the actual reading experiment. From here it follows that the familiarity scores are not directly linked to how familiar the phrases are experienced by the individual readers. In the present experiment, this practice was remedied by using the same participants to rate the familiarity of the topic-vehicle pairs as the ones that read the sentences including metaphoric and similes. Having established this, the items used were on average in the mid-range of the familiarity scale, and can be considered as ranging from novel to moderate familiar. Thus, the findings from the current study cannot be generalized to the processing of highly familiar (i.e., lexicalized metaphors; see e.g., Geiger & Ward, 1999; Giora, 2003) or conventionalized metaphors. Theoretical accounts would actually predict equally fast or even faster reading times for such metaphors in comparison to similes (e.g., Bowdle & Gentner, 2005; Giora, 2003). These hypotheses should be tested in the bilingual context in future studies.

Third, higher L2 English proficiency was related to overall faster reading of the English metaphors and similes. These results are in line with more general findings showing that L2 reading proficiency has an impact on the speed of processing metaphors; for instance, reading metaphors is slower in one's non-dominant than in one's dominant language (Heredia & Cieslicka, 2016). On the other hand, our results contrast with the findings of Ashby et al. (2018) within L1 speakers. That is, they did not find an impact of reading proficiency on processing metaphors and similes, despite using the same text materials as in the current study. We come back to this apparent discrepancy later in the Discussion.

As we hypothesized, we observed a delicate interplay between sentence type, phrase familiarity, and L2 English proficiency in different measures in the spillover and explanation region. That is, the effect of sentence type was modulated by interactions with familiarity in the explanation region, and with familiarity and English L2 proficiency in the spillover region. The latter interaction indicated that metaphor processing of readers with higher L2 proficiency is affected by metaphor familiarity, whereas this is not the case for lower L2 proficiency readers; for them it seems that metaphor processing is difficult anyway, no matter whether metaphors are familiar or not. This contrasts with simile processing, since here readers with lower L2 proficiency are affected by familiarity, whereas higher L2 proficiency readers are not; for them it seems that simile processing is easy anyway,

no matter whether they are unfamiliar or not. When directly comparing simile and metaphor processing, we observed that higher L2 proficiency readers showed similar go-past time on the spill-over region for familiar metaphors and similes, but longer go-past time for unfamiliar metaphors compared to unfamiliar similes. For lower proficiency L2 readers we observed similar go-past time for unfamiliar metaphors and similes but shorter go-past time for familiar similes compared to unfamiliar ones. The result suggests that all readers are able to minimize prediction error based on familiarity, but that L2 language proficiency dictates whether this is on a level of lower (similes) or higher linguistic complexity (metaphors). In other words, it could be argued that readers with higher L2 proficiency are able to use higher order information (e.g., familiarity of metaphor) to minimize prediction error in metaphor interpretation processes.

The interaction between familiarity and sentence type in the explanation region is driven by longer gaze durations for unfamiliar than familiar similes across proficiency level. This implies both low-proficiency and high-proficiency readers had not processed unfamiliar similes deeply enough the first time around. For higher proficiency L2 readers this suggests that they initially rely on the explicitness provided by “like” to interpret unfamiliar phrases as direct comparison, but that they need to elaborate on their interpretation while reading the explanation region, suggesting that the initial interpretation may have been inaccurate. For lower L2 proficiency readers, familiarity seems to benefit the processing of simile phrases throughout all regions. This suggests that making the comparison explicit makes the processing of the intended meaning less taxing for them, allowing them to use familiarity information during the interpretation of similes. In general, the findings are in line with the predictive coding theory (e.g., Fabry, 2021; Friston et al., 2006), which holds that factors such as familiarity, sentence complexity, and readers’ proficiency affect the ability to minimize the prediction error.

Finally, as our study has used exactly the same materials and procedure as Ashby et al. (2018), a direct comparison between L1 and L2 processing of metaphors and similes can be made. While doing so, it should be noted that our L2 speakers were all university undergraduate students using English on a daily basis in their studies. Consequently, they should be comfortable with reading in English and may represent a specific group of L2 readers due to which the results cannot be automatically generalized to all L2 readers. Yet the average scores and the range of the vocabulary test and the self-ratings indicated that many if not most L2 speakers were not on native-like level in English and this was reflected in the reading measures as well. For instance, L2 readers had about 2.5 times longer average go-past time than L1 readers in the explanation region (about around 800 ms vs. almost 2100 ms) and made about 50% regressions from this region against about 5% regressions by the L1 readers. This implies that L2 readers need much more reinspection of the figurative expressions than L1 speakers do. In a similar vein, both L1 and L2 readers show faster processing times for similes compared to metaphors in several measures, but L2 readers did not show these effects quite as early (i.e., not in gaze duration on the vehicle). The effect is very pronounced for L2 speakers in go-past time though. This indicates that while L1 readers more readily dedicate time to metaphor resolution on the vehicle, the L2 reader would need more reinforcement from earlier parts in the text for this, most likely from the topic. Both observations are in line with what we noted in the Introduction, that figurative language processing is notoriously difficult for L2 speakers and a

source for confusion (Littlemore et al., 2011). This has been found to be related to an initially more literal interpretation of the figurative expression. Idioms are for instance processed by L2 speakers at a similar speed as novel phrases and slower than literal phrases, whereas L1 speakers process idioms faster than novel phrases and equally fast as literal ones (Sivanova-Chanturia et al., 2011). Given that activation of literal meanings will lead L2 readers more often astray than L1 readers, there will be more need for reinterpretation, which in turn will lead to longer processing and reinspection times. Most likely, the initial activation of the literal meaning also delays processing of the similes and especially metaphors in our study, certainly when they are not so familiar and/or when L2 speakers are not so proficient.

It is notable that language proficiency did not impact metaphor effects in Ashby et al., but clearly modulated metaphor processing in this study. One reason for this may be that the participants in the Ashby et al. study were a homogeneous group of highly proficient native speakers sampled from the university population. Ashby et al. note that further research is needed to examine whether reading ability affects how children process metaphors, as it can be expected that reading proficiency is more variable among them. The lower level of proficiency in combination with a larger range in proficiency among the L2 speakers in this study probably allowed for the main effects of proficiency and the modulated metaphor effects to occur.

5. Conclusions

The results of this study showed that metaphors and similes are processed in different ways, and that this processing is modulated by a text-related property such as phrase familiarity and by a reader-related property such as language proficiency. The results lent support to the predictive coding theory (Fabry, 2021; Friston et al., 2006), as this theory holds that readers’ abilities (e.g., perceptual, cognitive, emotional, and linguistic) are in continuous interaction with text-related factors in order to minimize prediction error. Our study shows that, in cases where language proficiency is at the lower end, more complex figurative language processing remains challenging, even when the expressions become more familiar.

Acknowledgements. This research was supported by grants from the Ella and Georg Ehrnrooth Foundation and the Åbo Akademi University Foundation awarded to Henri Olkonemi. Portions of the data have been presented at the 61st Annual Meeting of the Psychonomic Society. We would like to thank Dr. Jane Ashby for providing the experimental sentences, research assistant Satu Savo for aiding in the data collection, and Angus MacGyver from the Phoenix Foundation for the inspiration.

References

- Acheson DJ, Wells JB and MacDonald MC (2008) New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods* 40, 278–289. <https://doi.org/10.3758/BRM.40.1.278>
- Aisenman RA (1999) Structure-mapping and the simile-metaphor preference. *Metaphor and Symbol* 14, 45–51. https://doi.org/10.1207/s15327868ms1401_5
- Ashby J, Roncero C, de Almeida RG and Agauas SJ (2018) The early processing of metaphors and similes: Evidence from eye movements. *Quarterly Journal of Experimental Psychology* 71, 161–168. <https://doi.org/10.1080/17470218.2016.1278456>
- Baayen RH, Davidson DJ and Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59, 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>

- Baayen RH and Milin P (2010) Analyzing reaction times. *International Journal of Psychological Research* 3, 2011–2084. <https://doi.org/10.21500/20112084.807>
- Barr DJ, Levy R, Scheepers C and Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68, 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates D, Mächler M, Bolker B and Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bialystok E, Craik FIM and Luk G (2012) Bilingualism: Consequences for mind and brain. *Trends in Cognitive Science* 16, 240–250. <https://doi.org/10.1016/j.tics.2012.03.001>
- Billow RM (1981) Observing spontaneous metaphor in children. *Journal of Experimental Child Psychology* 31, 430–445. [https://doi.org/10.1016/0022-0965\(81\)90028-X](https://doi.org/10.1016/0022-0965(81)90028-X)
- Birdsong D, Gertken LM and Amengual M (2012) *Bilingual Language Profile: An easy-to-use instrument to assess bilingualism*. Retrieved from <https://sites.la.utexas.edu/bilingual/>
- Blasko DG and Briihl DS (1997) Reading and recall of metaphorical sentences: Effects of familiarity and context. *Metaphor and Symbol* 12, 261–285. https://doi.org/10.1207/s15327868ms1204_4
- Boers F and Lindstromberg S (2012) Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics* 32, 83–110. <https://doi.org/10.1017/S0267190512000050>
- Bowdle BF and Gentner D (2005) The career of metaphor. *Psychological Review* 112, 193–216. <https://doi.org/10.1037/0033-295X.112.1.193>
- Brauer M and Curtin JJ (2018) Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods* 23, 389–411. <https://doi.org/10.1037/met0000159>
- Cacciari C and Glucksberg S (1994) Understanding figurative language. In M.A. Gernsbacher (eds), *Handbook of psycholinguistics*. San Diego: Academic Press, 447–477.
- Cieślicka A (2006) Literal salience in on-line processing of idiomatic expressions by second language learners. *Second Language Research* 22, 115–144. <https://doi.org/10.1191/0267658306sr263oa>
- Cieślicka AB, Heredia RR and Olivares M (2014) It's all in the eyes: How language dominance, salience, and context affect eye movements during idiomatic language processing. In M Pawlak and L Aronin (eds), *Essential topics in applied linguistics and multilingualism: Studies in honor of David Singleton*. New York: Springer, pp. 21–42. https://doi.org/10.1007/978-3-319-01414-2_2
- Cieślicka AB, Heredia RR and García T (2017) Task effects in bilingual idiom comprehension. *Poznań Studies in Contemporary Linguistics* 53, 95–117. <https://doi.org/10.1515/psic-2017-0005>
- Columbus G, Sheikh NA, Côté-Lecaldare M, Häuer K, Baum SR and Titone D (2015) Individual differences in executive control relate to metaphor processing: An eye movement study of sentence reading. *Frontiers in Human Neuroscience* 8, 1057. <https://doi.org/10.3389/fnhum.2014.01057>
- De Bruin A, Carreiras M and Duñabeitia JA (2017) The BEST Dataset of Language Proficiency. *Frontiers in Psychology* 8, 522. <https://doi.org/10.3389/fpsyg.2017.00522>
- Fabry RE (2021) Getting it: A predictive processing approach to irony comprehension. *Synthese* 198, 6455–6489. <https://doi.org/10.1007/s11229-019-02470-9>
- Friston K, Kilner J and Harrison L (2006) A free energy principle for the brain. *Journal of Physiology – Paris* 100, 70–87. <https://doi.org/10.1016/j.jphysparis.2006.10.001>
- Geiger O and Ward LW (1999) Metaphors and the mental lexicon. *Brain and Language* 68, 192–198. <https://doi.org/10.1006/brln.1999.2080>
- Gentner D and Wolff P (1997) Alignment in the processing of metaphor. *Journal of Memory and Language* 37, 331–355. <https://doi.org/10.1006/jmla.1997.2527>
- Giora R (2003) *On our mind: Saliency, context, and figurative language*. New York: Oxford University Press
- Glucksberg S (1998) Understanding metaphors. *Current Directions in Psychological Science* 7, 39–43. <https://doi.org/10.1111/1467-8721.ep13175582>
- Glucksberg S (2003) The psycholinguistics of metaphor. *Trends in Cognitive Sciences* 7, 92–96. [https://doi.org/10.1016/S1364-6613\(02\)00040-2](https://doi.org/10.1016/S1364-6613(02)00040-2)
- Gollan TH, Weissberger GH, Runnqvist E, Montoya RI and Cera CM (2012) Self-ratings of spoken language dominance: A multi-lingual naming test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. *Bilingualism* 15, 594–615. <https://doi.org/10.1017/S1366728911000332>
- Gregory ME and Mergler NL (1990) Metaphor comprehension: In search of literal truth, possible sense, and metaphoricality. *Metaphor and Symbolic Activity* 5, 121–173. https://doi.org/10.1207/s15327868ms0503_2
- Grice H (1975) Logic and conversation. In P Cole and J Morgan (eds), *Syntax and semantics 3: Speech acts*. New York: Academic Press, pp. 41–58.
- Harris RJ and Mosier NJ (1999) Memory for metaphors and similes in discourse. *Discourse Processes* 28, 257–270. <https://doi.org/10.1080/01638539909545084>
- Heredia RR and Cieślicka AB (2016) Metaphorical reference: An eye movement analysis of Spanish-English and English-Spanish bilingual readers. *Frontiers in Psychology* 7, 439. <https://doi.org/10.3389/fpsyg.2016.00439>
- Hilton H (2008) The link between vocabulary knowledge and spoken L2 fluency. *The Language Learning Journal* 36, 153–166. <https://doi.org/10.1080/09571730802389983>
- Inhoff AW, Lima SD and Carroll PJ (1984) Contextual effects on metaphor comprehension in reading. *Memory & Cognition* 12, 558–567. <https://doi.org/10.3758/BF03213344>
- Jankowiak K, Rataj K and Naskręcki R (2017) To electrify bilingualism: Electrophysiological insight into bilingual metaphor comprehension. *PLoS One* 12, e0175578. <https://doi.org/10.1371/journal.pone.0175578>
- Jeon EH and Yamashita J (2014) L2 reading comprehension and its correlates: A meta-analysis. *Language Learning* 64, 160–212. <https://doi.org/10.1111/lang.12034>
- Jones LL and Estes Z (2006) Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language* 55, 18–32. <https://doi.org/10.1016/j.jml.2006.02.004>
- Lemhöfer K and Broersma M (2011) Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods* 44, 325–343. <https://doi.org/10.3758/s13428-011-0146-0>
- Littlemore J, Chen PT, Koester A and Barnden J (2011) Difficulties in metaphors comprehension faced by international students whose first language is not English. *Applied Linguistics* 32, 408–429. <https://doi.org/10.1093/applin/amr009>
- Lupyan G and Clark A (2015) Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science* 24, 279–284. <https://doi.org/10.1177/0963721415570732>
- Marian V, Blumenfeld HK and Kaushanskaya M (2007) The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research* 50, 940–967. [https://doi.org/10.1044/1092-4388\(2007\)067](https://doi.org/10.1044/1092-4388(2007)067)
- McCarron SP and Kuperman V (2021) Is the author recognition test a useful metric for native and non-native English speakers? An item response theory analysis. *Behavior Research Methods* 47, 1095–1109. <https://doi.org/10.3758/s13428-021-01556-y>
- Olkoniemi H and Kaakinen JK (2021) Processing of irony in text: A systematic review of eye tracking studies. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 75, 99–106. <https://doi.org/10.1037/cep0000216>
- Olkoniemi H, Ranta H and Kaakinen JK (2016) Individual differences in the processing of written sarcasm and metaphor: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42, 433–450. <https://doi.org/10.1037/xlm0000176>
- Ortony A (1975) Why metaphors are necessary and not just nice. *Educational Theory* 25, 45–53. <https://doi.org/10.1111/j.1741-5446.1975.tb00666.x>
- Pexman PM, Ferretti TR and Katz AN (2000) Discourse factors that influence online reading of metaphor and irony. *Discourse Processes* 29, 201–222. https://doi.org/10.1207/S15326950dp2903_2
- Pierce RS, MacLaren R and Chiappe DL (2010) The role of working memory in the metaphor interference effect. *Psychonomic Bulletin & Review* 17, 400–404. <https://doi.org/10.3758/PBR.17.3.400>
- Raney GE, Campbell SJ and Bovee JC (2014) Using eye movements to evaluate the cognitive processes involved in text comprehension. *Journal of Visualized Experiments* 83, e50780. <https://doi.org/10.3791/50780>

- Rayner K** (1998) Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* **124**, 372–422. <https://doi.org/10.1037/0033-2909.124.3.372>
- Rayner K** (2009) The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology* **62**, 1457–1506. <https://doi.org/10.1080/17470210902816461>
- R Core Team** (2020) R: A language and environment for statistical computing. Retrieved from <http://www.R-project.org>.
- Roncero C and de Almeida RG** (2015) Semantic properties, aptness, familiarity, conventionality, and interpretive diversity scores for 84 metaphors and similes. *Behavior Research Methods* **47**, 800–812. <https://doi.org/10.3758/s13428-014-0502-y>
- Salmela R, Lehtonen M, Garusi S and Bertram R** (2021) Lexize: A test to quickly assess vocabulary knowledge in Finnish. *Scandinavian Journal of Psychology*. Advance online publication. <https://doi.org/10.1111/sjop.12768>
- Shibata M, Toyomura A, Motoyama H, Itoh H, Kawabata Y and Abe J-I** (2012) Does simile comprehension differ from metaphor comprehension? A functional MRI study. *Brain and Language* **121**, 254–260. <https://doi.org/10.1016/j.bandl.2012.03.006>
- Siyanova-Chanturia A, Conklin K and Schmitt N** (2011) Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research* **27**, 251–272. <https://doi.org/10.1177/0267658310382068>
- Titone D, Columbus G, Whitford V, Mercier J and Libben M** (2015) Contrasting bilingual and monolingual idiom processing. In RR Heredia and A B. Cieślicka (eds), *Bilingual figurative language processing*. Cambridge University Press, pp. 171–207 <https://doi.org/10.1017/CBO9781139342100.011>
- Tiv M, Rouillard V, Vingron N, Wiebe S and Titone D** (2019) Global second language proficiency predicts self-perceptions of general sarcasm use among bilingual adults. *Journal of Language and Social Psychology* **38**, 459–478. <https://doi.org/10.1177/0261927X19865764>
- Vaid J, López BG and Martínez FE** (2015) Linking the figurative to the creative: Bilinguals' comprehension of metaphors, jokes, and remote associates. In RR Heredia and AB Cieślicka (eds), *Bilingual figurative language processing*. Cambridge University Press: New York, pp. 53–86 <https://doi.org/10.1017/CBO9781139342100.006>
- Ylinen S, Bosseler A, Junntila K and Huotilainen M** (2017) Predictive coding accelerates word recognition and learning in the early stages of language development. *Developmental Science* **20**, e12472. <https://doi.org/10.1111/desc.12472>
- Zhang S and Zhang X** (2020) The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*. Advance online publication. <https://doi.org/10.1177/1362168820913998>
- Zharikov S and Gentner D** (2002) Why do metaphors seem deeper than similes? In WD Gray and CD Schunn (eds), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*. Fairfax, VA: George Mason University, pp. 976–981.

Appendix A

Final models for each eye movement measure

Table A1. Final model for gaze duration on vehicle

| Random Effects | <i>n</i> | Variance | <i>SD</i> |
|---|------------------|---------------|---------------|
| Participant (Intercept) | 63 | 0.05 | 0.21 |
| Item (Metaphor) | 62 | 0.04 | 0.20 |
| Item (Simile) | | 0.03 | 0.19 |
| Item (English Proficiency) | | 0.003 | 0.05 |
| Residual | | 0.17 | 0.41 |
| Fixed Effects | <i>Estimates</i> | <i>95% CI</i> | <i>t</i> |
| Intercept | 5.72 | 5.64 – 5.79 | 155.20 |
| Sentence Type (Simile vs. Metaphor) | 0.02 | –0.01 – 0.05 | 1.36 |
| Familiarity | –0.003 | –0.02 – 0.01 | –0.29 |
| English Proficiency | –0.11 | –0.18 – –0.05 | –3.42 |
| Sentence Type (Simile vs. Metaphor) × Familiarity | 0.01 | –0.02 – 0.04 | 0.60 |
| Sentence Type (Simile vs. Metaphor) × English Proficiency | 0.001 | –0.03 – 0.03 | 0.04 |
| Familiarity × English Proficiency | 0.004 | –0.01 – 0.02 | 0.44 |
| Sentence Type (Simile vs. Metaphor) × Familiarity × English Proficiency | 0.01 | –0.02 – 0.04 | 0.62 |

Note. *t* values > |1.96| are bolded.

Table A2. Final model for go-past time on vehicle

| Random Effects | <i>n</i> | Variance | <i>SD</i> |
|---|------------------|---------------|---------------|
| Participant | 63 | 0.06 | 0.24 |
| Item | 62 | 0.04 | 0.21 |
| Item (Simile vs Metaphor) | | 0.01 | 0.09 |
| Item (English Proficiency) | | 0.004 | 0.06 |
| Residual | | 0.22 | 0.47 |
| Fixed Effects | <i>Estimates</i> | <i>95% CI</i> | <i>t</i> |
| Intercept | 5.88 | 5.80 – 5.96 | 145.02 |
| Sentence Type (Simile vs. Metaphor) | 0.12 | 0.08 – 0.16 | 6.15 |
| Familiarity | –0.02 | –0.04 – 0.001 | –1.84 |
| English Proficiency | –0.13 | –0.20 – –0.05 | –3.38 |
| Sentence Type (Simile vs. Metaphor) × Familiarity | 0.003 | –0.03 – 0.04 | 0.19 |
| Sentence Type (Simile vs. Metaphor) × English Proficiency | –0.01 | –0.04 – 0.03 | –0.46 |
| Familiarity × English Proficiency | –0.01 | –0.03 – 0.01 | –1.40 |
| Sentence Type (Simile vs. Metaphor) × Familiarity × English Proficiency | –0.02 | –0.05 – 0.02 | –0.84 |

Note. *t* values > |1.96| are bolded.

Table A3. Final model for regression out proportion from vehicle

| Random Effects | <i>n</i> | Variance | <i>SD</i> |
|---|------------------|---------------|---------------|
| Participant (Intercept) | 63 | 0.37 | 0.61 |
| Item (Intercept) | 62 | 0.05 | 0.23 |
| Item (English Proficiency) | | 0.03 | 0.17 |
| Fixed Effects | <i>Estimates</i> | <i>95% CI</i> | <i>z</i> |
| Intercept | -2.04 | -2.24 – -1.85 | -20.51 |
| Sentence Type (Simile vs. Metaphor) | 0.74 | 0.54 – 0.93 | 7.45 |
| Familiarity | -0.14 | -0.25 – -0.03 | -2.49 |
| English Proficiency | -0.06 | -0.27 – 0.16 | -0.50 |
| Sentence Type (Simile vs. Metaphor) × Familiarity | -0.02 | -0.22 – 0.18 | -0.18 |
| Sentence Type (Simile vs. Metaphor) × English Proficiency | -0.16 | -0.38 – 0.06 | -1.40 |
| Familiarity × English Proficiency | -0.08 | -0.21 – 0.04 | -1.37 |
| Sentence Type (Simile vs. Metaphor) × Familiarity × English Proficiency | -0.06 | -0.29 – 0.17 | -0.51 |

Note. *z* values > |1.96| are bolded.

Table A4. Final model for gaze duration on spillover region

| Random Effects | <i>n</i> | Variance | <i>SD</i> |
|---|------------------|----------------|---------------|
| Participant (Intercept) | 63 | 0.02 | 0.15 |
| Item (Intercept) | 62 | 0.001 | 0.03 |
| Item (Familiarity) | | 0.001 | 0.03 |
| Item (English Proficiency) | | 0.001 | 0.03 |
| Residual | | 0.11 | 0.32 |
| Fixed Effects | <i>Estimates</i> | <i>95% CI</i> | <i>t</i> |
| Intercept | 5.51 | 5.47 – 5.55 | 273.87 |
| Sentence Type (Simile vs. Metaphor) | -0.01 | -0.03 – 0.01 | -1.00 |
| Familiarity | -0.02 | -0.03 – -0.002 | -2.22 |
| English Proficiency | -0.04 | -0.08 – 0.01 | -1.54 |
| Sentence Type (Simile vs. Metaphor) × Familiarity | -0.01 | -0.03 – 0.01 | -0.59 |
| Sentence Type (Simile vs. Metaphor) × English Proficiency | -0.004 | -0.03 – 0.02 | -0.30 |
| Familiarity × English Proficiency | -0.01 | -0.02 – 0.01 | -1.01 |
| Sentence Type (Simile vs. Metaphor) × Familiarity × English Proficiency | -0.02 | -0.04 – 0.005 | -1.58 |

Note. *t* values > |1.96| are bolded.

Table A5. Final model for go-past time on spillover region

| Random Effects | <i>n</i> | Variance | <i>SD</i> |
|---|------------------|---------------|---------------|
| Participant | 63 | 0.03 | 0.18 |
| Item | 62 | 0.01 | 0.11 |
| Item (Simile vs Metaphor) | | 0.01 | 0.08 |
| Residual | | 0.19 | 0.44 |
| Fixed Effects | <i>Estimates</i> | <i>95% CI</i> | <i>t</i> |
| Intercept | 5.66 | 5.60 – 5.71 | 203.96 |
| Sentence Type (Simile vs. Metaphor) | 0.06 | 0.02 – 0.09 | 3.42 |
| Familiarity | –0.03 | –0.05 – –0.01 | –3.43 |
| English Proficiency | –0.07 | –0.13 – –0.02 | –2.54 |
| Sentence Type (Simile vs. Metaphor) × Familiarity | –0.001 | –0.03 – 0.03 | –0.09 |
| Sentence Type (Simile vs. Metaphor) × English Proficiency | –0.002 | –0.03 – 0.03 | –0.13 |
| Familiarity × English Proficiency | –0.003 | –0.02 – 0.01 | –0.30 |
| Sentence Type (Simile vs. Metaphor) × Familiarity × English Proficiency | –0.05 | –0.08 – –0.01 | –2.86 |

Note. *t* values > |1.96| are bolded.

Table A6. Final model for regression out proportion from spillover region

| Random Effects | <i>n</i> | Variance | <i>SD</i> |
|---|------------------|---------------|---------------|
| Participant (Intercept) | 63 | 0.58 | 0.76 |
| Item (Intercept) | 62 | 0.58 | 0.76 |
| Fixed Effects | <i>Estimates</i> | <i>95% CI</i> | <i>z</i> |
| Intercept | –2.65 | –2.96 – –2.34 | –16.93 |
| Sentence Type (Simile vs. Metaphor) | 0.59 | 0.37 – 0.81 | 5.16 |
| Familiarity | –0.18 | –0.32 – –0.04 | –2.47 |
| English Proficiency | –0.27 | –0.53 – –0.01 | –2.02 |
| Sentence Type (Simile vs. Metaphor) × Familiarity | –0.02 | –0.25–0.22 | –0.15 |
| Sentence Type (Simile vs. Metaphor) × English Proficiency | 0.06 | –0.19–0.31 | 0.65 |

Note. *z* values > |1.96| are bolded.

Table A7. Final model for gaze duration on explanation region

| Random Effects | <i>n</i> | Variance | <i>SD</i> |
|---|------------------|---------------|--------------|
| Participant (Intercept) | 63 | 18.04 | 4.25 |
| Item (Intercept) | 62 | 21.15 | 4.60 |
| Residual | | 45.74 | 6.76 |
| Fixed Effects | <i>Estimates</i> | <i>95% CI</i> | <i>t</i> |
| Intercept | 36.14 | 34.57 – 37.71 | 45.20 |
| Sentence Type (Simile vs. Metaphor) | –0.75 | –1.18 – –0.33 | –3.46 |
| Familiarity | –0.18 | –0.46 – 0.11 | –1.23 |
| English Proficiency | –1.77 | –3.03 – –0.52 | –2.77 |
| Sentence Type (Simile vs. Metaphor) × Familiarity | 0.44 | 0.01 – 0.87 | 1.99 |
| Sentence Type (Simile vs. Metaphor) × English Proficiency | –0.13 | –0.63 – 0.37 | –0.53 |
| Familiarity × English Proficiency | 0.01 | –0.25 – 0.27 | 0.09 |
| Sentence Type (Simile vs. Metaphor) × Familiarity × English Proficiency | 0.14 | –0.36 – 0.64 | 0.56 |

Note. *t* values > |1.96| are bolded.

Table A8. Final model for go-past time on explanation region

| Random Effects | <i>n</i> | Variance | <i>SD</i> |
|---|------------------|---------------|---------------|
| Participant (Intercept) | 63 | 0.14 | 0.38 |
| Item (Intercept) | 62 | 0.06 | 0.25 |
| Residual | | 0.15 | 0.39 |
| Fixed Effects | <i>Estimates</i> | <i>95% CI</i> | <i>t</i> |
| Intercept | 7.45 | 7.33 – 7.56 | 128.63 |
| Sentence Type (Simile vs. Metaphor) | –0.01 | –0.03 – 0.02 | –0.46 |
| Familiarity | –0.03 | –0.05 – –0.02 | –3.96 |
| English Proficiency | –0.14 | –0.25 – –0.03 | –2.45 |
| Sentence Type (Simile vs. Metaphor) × Familiarity | 0.02 | –0.01 – 0.04 | 1.56 |
| Sentence Type (Simile vs. Metaphor) × English Proficiency | 0.01 | –0.02 – 0.04 | 0.60 |
| Familiarity × English Proficiency | 0.01 | –0.005–0.03 | 1.34 |
| Sentence Type (Simile vs. Metaphor) × Familiarity × English Proficiency | –0.01 | –0.04–0.02 | –0.50 |

Note. *t* values > |1.96| are bolded.

Table A9. Final model for regression out proportion from explanation region

| Random Effects | <i>n</i> | Variance | <i>SD</i> |
|---|------------------|---------------|--------------|
| Participant (Intercept) | 63 | 2.12 | 1.46 |
| Item (Intercept) | 62 | 0.10 | 0.31 |
| Fixed Effects | <i>Estimates</i> | <i>95% CI</i> | <i>z</i> |
| Intercept | 0.01 | –0.36 – 0.39 | 0.07 |
| Sentence Type (Simile vs. Metaphor) | –0.02 | –0.16 – 0.13 | –0.21 |
| Familiarity | –0.16 | –0.25 – –0.07 | –3.58 |
| English Proficiency | –0.18 | –0.62 – 0.25 | –0.83 |
| Sentence Type (Simile vs. Metaphor) × Familiarity | –0.07 | –0.22 – 0.08 | –0.97 |
| Sentence Type (Simile vs. Metaphor) × English Proficiency | 0.07 | –0.10 – 0.25 | 0.85 |
| Familiarity × English Proficiency | 0.04 | –0.04 – 0.13 | 0.97 |
| Sentence Type (Simile vs. Metaphor) × Familiarity × English Proficiency | 0.01 | –0.16 – 0.18 | 0.10 |

Note. *z* values > |1.96| are bolded.