What can L1 speakers tell us about *killing hope*? A Novel Behavioral Measure for Identifying Collocations

Sydelle de Souza (sydelle.desouza@ed.ac.uk)
CDT in Natural Language Processing, School of Informatics, University of Edinburgh
Edinburgh, United Kingdom

Francis Mollica
Melbourne School of Psychological Sciences, University of Melbourne
Melbourne, Australia

Jennifer Culbertson
Centre for Language Evolution, University of Edinburgh
Edinburgh, United Kingdom

Abstract

Collocations, semi-productive lexical combinations with one figurative and one literal word, are said to be a “pain in the neck” for researchers and L2 learners. The present study aims: (i) to conceptually replicate the processing costs incurred by L1 speakers when processing collocations using a larger and more diverse set of items, (ii) to use literalness judgements to test whether L1 speakers are aware of the semi-transparent meaning of a collocation, and (iii) to test whether the presence of processing costs associated with collocations can be predicted from literalness judgements. If so, we propose that literalness judgements could be used as a diagnostic for reliably identifying collocations. We replicate the L1 processing costs with a larger stimulus set and demonstrate that speakers are aware of the semi-transparent meaning of the collocation. We further show that L1 speaker judgements about the literalness of a word combination can be used to predict its status as a collocation.

Keywords: semi-productive language; collocations; literalness judgements

Collocations

From *chasing dreams* and *drawing ire* to *heavy rain* and *catching fire*, semi-productive lexical patterns are ubiquitous in human language (Mel’čuk, 2003). Often referred to as *collocations*, these idiosyncratic lexical items are comprised of one word used in its literal sense and one other in its figurative sense, constrained by an arbitrary restriction on substitution (Howarth, 1998). To illustrate, one can *raise doubts* or *lift bans*, but not #*lift doubts* (to mean raise doubts) nor #*raise bans*. Collocations are syntactically well formed, but deviate from or violate the expected semantic representation (Culicover, Jackendoff, & Audring, 2017). For example, the verb *kill* prototypically requires an animate object, so one can *kill bugs* and *kill trees*, but not **kill books*. Yet one can *kill time*, *hope*, and *dreams*. Evidently, collocations are neither fully productive nor fully idiomatic.

Collocations constitute the largest subset of formulaic language (Barfield & Gyllstad, 2009), which together with other subsets such as idioms, binomials, and metaphors account for more than half of any given text, written or oral (Erman & Warren, 2000). A long history of research shows that proper knowledge and use of these lexical units is crucial to developing communicative competence, as they provide idiomaticity and fluency to the language user (Firth, 1957; Pawley & Syder, 1983; Nation, 2001; Durrant & Schmitt, 2010; Siyanova-Chanturia & Pellicer-Sanchez, 2018; Garner, 2022). It is hardly surprising then that they are considered to be crucial in various areas of linguistics, from language teaching and lexicography (Čermák, 2006) to natural language processing (NLP) applications such as human-computer interaction (Ford & Smith, 1982; Koulouri, Lauria, & Macredie, 2016) and machine translation (Dankers, Bruni, & Hupkes, 2022). While collocations have attracted a great deal of attention in these spheres, they have also gained notoriety. Due to their semi-transparent nature, collocations are considered to be a “pain in the neck” (Sag, Baldwin, Bond, Copestake, & Flickinger, 2002) for researchers and second language (L2) learners alike. However, collocations have largely been ignored in theoretical linguistics (Herbst, 2018; Wray, 2002) and perhaps as a result, in mainstream psycholinguistics as well. Little is known about how first language (L1) speakers acquire and process collocations, or if they are even a valid psychological construct, i.e., not merely a descriptive typology, in the first place. Therefore, the present study aims to investigate L1 speakers’ intuitions of the semi-transparent nature and meaning of a collocation.

The Identification Problem

For researchers, the traditional method for identifying collocations has relied on L1 speaker judgements, which can be tedious and expensive, especially with large corpora (Wahl & Gries, 2018). Furthermore, these judgements are subjective and often do not replicate across raters. For example, we all agree that one can *chase dreams*, but not everyone will agree that you can *hunt* them. This is problematic as it could leave collocations unidentified, or conversely, it could misidentify word combinations that have appeared together by chance as a collocation. The emergence of computational approaches for automatic collocation extraction have helped overcome some of these issues. Exploiting the properties that collocations are always syntagmatically related (Nesselhauf, 2003), well-formed (Culicover et al., 2017), and occur together more frequently than chance (Sinclair, 1991) might permit iden-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

tifying collocations using statistical measures such as mutual information scores to gauge the strength of associations (Wahl & Gries, 2018). However, purely statistical methods have their limitations. They often misidentify collocations, while simultaneously leaving low-frequency collocations unidentified (Seretan, 2018). Ultimately, collocations identified through these methods must be manually curated.

The Selection Problem

For second language (L2) learners, collocations are known to be notoriously difficult to acquire and use, even for those reaching high language proficiency levels (Wolter & Gyllstad, 2013; Fioravanti, Senaldi, Lenci, & Siyanova-Chanturia, 2021). While L1 children seem to learn collocations with apparent ease, research suggests that even early sequential bilinguals have trouble converging on the "native-like" use of collocations as adults (Nishikawa, 2019). Part of the difficulty may stem from the fact that collocations often comprise high-frequency words, which causes learners to overestimate their knowledge of the semi-transparent meaning involved, leading to errors in comprehension (Laufer, 2011; Martinez & Murphy, 2011). In production, L2 speakers tend to allow substitutions in collocations that to an L1 speaker sound odd and erroneous (Fioravanti et al., 2021; Cowie & Howarth, 1996). A widely cited study by Pawley and Syder (1983) claims that L1 speakers do not make lexical choices based on word-level syntax or semantics in a way that two synonyms could be substituted in a given combination. To illustrate, *heavy* and *weighty* are both adjectives with similar meanings. However, L1 speakers will produce the combination *heavy smoker* but not *weighty smoker*. Evidently, these seemingly arbitrary lexical choices made by L1 speakers—that Pawley and Syder (1983) term "native-like selection"—could hold the key to solving the identification problem.

This difficulty faced by L2 speakers is reflected in behavioral (Wolter & Yamashita, 2015, 2018) and electrophysiological (Pulido & Dussias, 2019) data which show that L2 learners process collocations slower and less accurately than productive language. Interestingly, L1 speakers tested as a control group in these studies also incur processing costs for collocations over productive combinations, at least in terms of reaction times (RTs) (Gyllstad & Wolter, 2016; Souza & Chalmers, 2022). However, these studies are underpowered, especially in terms of items, as they have to be curated within the constraints of the L2 speakers' first language.

The Present Study

Based on the review of the literature and the gaps identified therein, the present study aims: (i) to conceptually replicate the costs incurred by L1 speakers when processing collocations using a larger and more diverse set of items, (ii) to test whether L1 speakers are aware of the semi-transparent meaning of a collocation, and (iii) to test whether the presence of processing costs associated with collocations can be predicted from literalness judgements. If so, we propose that literalness judgements could be used as a diagnostic for reliably identifying collocations. In the following we present behavioral experiments and statistical modelling aimed at addressing these issues.

Experiment

To answer the questions posed above, this study tests L1 English speakers in two behavioral tasks—a timed acceptability judgement task and a novel literalness judgement task. In the acceptability judgement task, we ask speakers to judge whether a word combination is acceptable to them in English. In the novel literalness judgement task, we ask them to judge whether the verb in a given word combination is being used literally.

We first conceptually replicate L1 results from L2 collocational processing studies. Specifically, we test how well RTs in the acceptability judgement task can be predicted by our expert "gold standard" judgements of whether or not a given word combination is a collocation. In other words, we test whether the processing cost associated with collocations (as compared to productive combinations) that has been reported in previous literature is indeed a robust behavioural signature of collocations. We then investigate L1 speakers' intuitions of the semi-transparent meanings via literalness judgements and compare them to our gold standard. We look at whether speakers' own literalness judgement task responses can equally well predict this processing signature. Specifically, we compare a model that predicts acceptability judgement task RTs using our "gold standard" judgements to a model that predicts them using L1 speakers' literalness judgements. This comparison allows us to test whether literalness judgements can be used as a diagnostic for identifying collocations.

Materials

First, we generated a preliminary list of 158 Verb-Noun collocations based on previous literature, L1-speaker intuitions (of the first author) and the *Collins COBUILD Advanced Learner's Dictionary* (Collins, 2018). We then computed each collocation's phrasal frequency and association scores in The Sketch Engine's enTenTen21 corpus, a massive (61.6M tokens; 52.3M words), dynamic web corpus containing texts from various genres and from all varieties of English. The corpus was queried using the verb as the node and restricting the collocate to the direct object position. Light verb constructions, i.e., collocations involving "neutral" verbs such as *make* and *take* (e.g., *make a decision*, *take a walk*) were not included as their meanings can be gleaned from the noun itself (Culicover et al., 2017).

Then each verb in the preliminary list was checked to ensure that all initial collocations were moderately to strongly associated by referencing their logDice scores. We opted to use logDice scores instead of mutual information as they are not affected by corpus size and have the added benefit of being easy to interpret. Mutual information is strongly affected by frequency and corpus size, wherein low-frequency words tend to have a higher mutual information (indicating stronger association), which could be misleading (Rychlý, 2008) and the larger the corpus, the more skewed the mutual information will be. Dice calculates association without accounting for corpus size. It is expressed as:

$$D = \frac{2\, f_{\mathrm{xy}}}{f_{\mathrm{x}} + f_{\mathrm{y}}}, \tag{1}$$

where f_x and f_y are the number of occurrences of words x and y in the corpus respectively, and f_{xy} is the number of co-occurrences of x and y. However, the values of the Dice score (D) are usually very small numbers. Therefore, Rychlý (2008) proposes the logDice, which is easier to interpret and is expressed as:

$$logDice = 14 + log_2\, D. \tag{2}$$

LogDice expresses the strength of the association on an easy-to-interpret scale with a theoretical maximum value of 14 which indicates that the two words always occur together in a given corpus, while a score of zero means that they never occur together. A negative score indicates dissociation—the words are likely to be unrelated. In essence, the closer the value is to 14, the stronger the association.

We eliminated collocations with a logDice score lower than five. This resulted in a list of 100 base collocations with a mean phrasal frequency of 20470.27 (SD = 37504.29) and a mean logDice score of 7.28 (SD = 1.26). For every collocation, we created a corresponding productive combination which shared the verb but differed in the noun. This noun was chosen by identifying an alternative which resulted in the minimum difference in raw corpus frequency between the collocation and the productive combination. The final item set comprised 200 target items—100 Verb-Noun collocations (e.g., *chase dream, freeze account*) and 100 productive Verb-Noun combinations containing the same verb as the collocation (e.g., *chase rabbit, freeze vegetable*). We also constructed 40 nonsense items (e.g., *roast bells, stay music*) with unique verbs and nouns for use as fillers to balance the acceptability judgement task.

Experimental Procedure

As previously mentioned, the study comprised two tasks, an acceptability judgement task (AJT) and a literalness judgement task (LJT). Each participant performed both tasks. Figure 1 depicts the experimental procedure to which the participants were subjected. The median completion time was 6.4 minutes. A total of 230 L1 English speakers (F=117; M=113; Non-binary=0) with a mean age of 42.03 years (SD = 13.41) were recruited using Prolific. They were remunerated £1.50 for their participation. The study was certified according to theUniversity of Edinburgh's School of Philosophy, Psychology and Language Sciences Research Ethics Process (RT number: 339-2122/4).

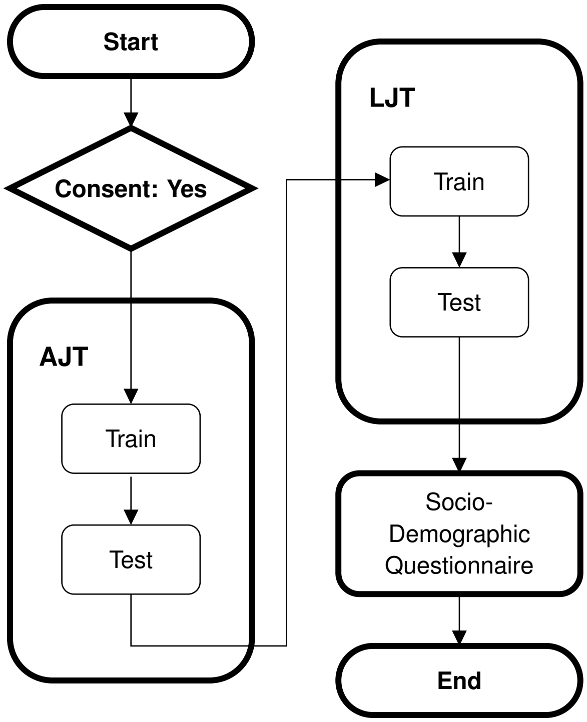

Figure 1: Experimental Procedure Flow Diagram

Acceptability Judgements In this task, participants were asked to judge whether or not the word combination presented to them sounded acceptable, i.e., would they as L1 English speakers use this word combination in their everyday speech, by pressing the 'y' key for yes or the 'n' key for no. They were asked to respond as quickly and accurately as possible. During testing, each participant saw 10 collocations, 10 productive combinations, and 20 nonsense combinations. Items were presented in an individualized random order with the constraint that no participant saw a collocation and a productive item with the same verb. A fixation cross with an inter-stimulus interval of 350 ms was presented between trials. Trials timed out at 8,000 ms if no decision was taken. See figure 2 for a visualization of how the stimuli were presented to the participants in the AJT.

Literalness Judgements The procedure for this task was similar to the previous one. The participants were asked to judge whether or not the action expressed by the verb in the word combination was *really happening*, by pressing the 'y' key for yes or the 'n' key for no. The participants judged the same 20 target items that were presented to them in the acceptability judgement task. Trials were not set to time out and participants were made aware that this was not a speeded judgement task.

Both tasks began with a short training set of six trials with feedback. Participants were informed that there were no right or wrong answers and that we were only interested in their judgements. Therefore, for feedback during training we opted for a smiley face emoji for answers that matched our gold standard and a confused face emoji for answers that did not. The decision to include this feedback was made on the basis of suggestions from participants in a pilot experiment. See

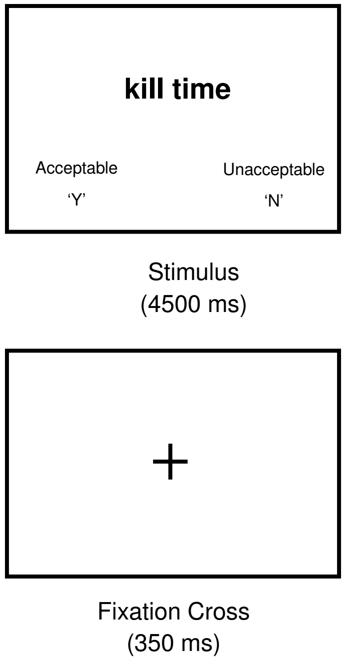

Figure 2: Stimulus Presentation for the Acceptability Judgement Task

figure 3 for a visualization of how the stimuli were presented to the participants in the LJT.

Data Analysis

Data pre-processing Our initial data set comprised 4,600 observations. We eliminated responses with reaction times (RTs) slower than 450 ms (0.1%), outliers greater than 3.5 standard deviations from the mean (1.61%), and incorrect acceptability judgement responses (4.58%). The resulting data set comprised 4,266 observations. All statistical models were run on this data.

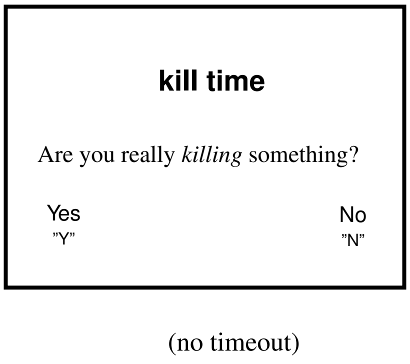

Figure 3: Stimulus Presentation for the Literalness Judgement Task

Statistical Models We specified linear mixed-effects models using the 'lme4' (Bates, Mächler, Bolker, & Walker, 2015) package in R version 4.3.2 (R Core Team, 2023). We first specified a maximal model as "justified by the design" (Barr, Levy, Scheepers, & Tily, 2013). The main dependent variable was the reaction times (RTs) from the acceptability judgement task. The main predictor variables were Condition (Collocation or Productive), i.e., our "gold standard", Phrasal Frequency (logged and scaled) and the interaction between the two. The logDice scores (scaled) were included as a covariate. Full crossed random effects for `(Condition || ID)` and `(Condition|| Verb)` and random intercepts `(1 | Verb)` and `(1 | Participant)` were specified. The maximal model is expressed as:

Maximal Model:

```
RT ∼ Condition * Phrasal Frequency + Score +
    (Condition || ID) + (Condition || Verb)
```

Due to convergence and singular fit issues, the maximal model was simplified by step-by-step elimination of the random effects structure. The final model was expressed as:

Gold Standard Model:

```
RT ∼ Condition * Phrasal Frequency + Score + (1
              | ID) + (1 | Verb)
```

The second set of models was the same as the simplified first set, except that instead of our "gold standard", the literalness judgements (Yes or No) were included as the main predictor variable. The final model was stated as:

Human Judgements Model:

```
RT ∼ Literalness Judgements * Phrasal Frequency
         + Score + (1 | ID) + (1 | Verb)
```

Results

Global Results

The mean RT and mean accuracy for each condition in the acceptability judgement task was first calculated (see Figure 4). The mean RT was 1051.89 ms (SD = 330.28 ms) for productive combinations and 1091.14 ms (SD = 337.58) for collocations. The mean accuracy for productive items was 95.6% and for collocations was 93.2%. Figure 5 depicts the mean agreement between the literalness judgements and the gold standard. 91.5% of the literalness judgements for productive combinations agree with our gold standard, while only 81.5% of literalness judgements match our gold standard for collocations.

Model Results

Results for the Gold Standard Model showed a significant difference in RTs between Conditions (treatment coded, with productive combination as the baseline; $\beta = 41.511; SE = 5.960; p < 0.001$), suggesting that collocations are processed significantly slower than productive items. This replicates the processing cost observed in L1 control groups in previous L2 studies, but with a larger and more diverse set of stimuli. Unsurprisingly, Phrasal Frequency also has a significant effect on RTs ($\beta = -71.849; SE = 8.206; p < 0.001$), corresponding to a 71.849 ms decrease in RT for every 1 standard

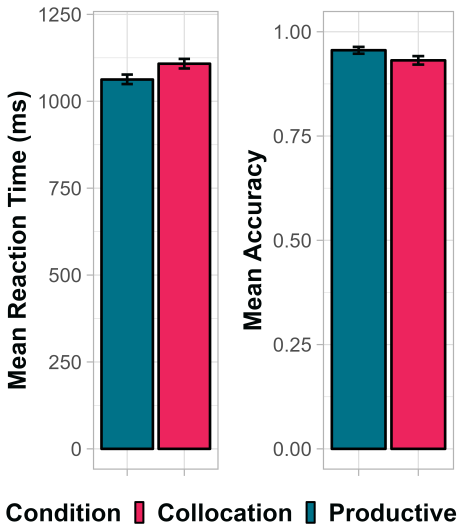

Figure 4: Mean Reaction Times and Accuracy in the Acceptability Judgement Task for Productive Combinations and Collocations. Error bars indicate bootstrapped confidence intervals.

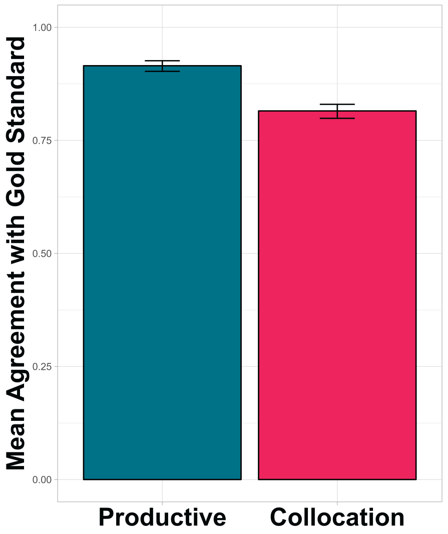

Figure 5: Mean Agreement in the Literalness Judgement Task for Productive Combinations and Collocations. Error bars indicate bootstrapped confidence intervals.

deviation increase in phrasal frequency. A statistically significant interaction between condition and phrasal frequency was also found, suggesting that condition had more of an impact on lower frequency combinations ($\beta = -18.484; SE = 6.206; p = 0.003$). No significant effect for logDice score was detected ($\beta = -0.297; SE = 6.349; p = 0.963$).

Importantly, results for the Human Judgements Model were very similar. We found statistically significant differences in RTs between Literalness Judgements (treatment coded, with judgement of the verb as literal as the baseline; $\beta = 35.205; SE = 6.174; p < 0.001$), suggesting that participants were 35.205 ms slower to respond to figurative verbs than literal ones. Similarly, an increase in phrasal frequency led to faster RTs ($\beta = -68.598; SE = 8.067; p < 0.001$) and a statistically significant interaction between the literalness judgements and phrasal frequency was also found, suggesting that literalness judgement had more of an effect for lower frequency items ($\beta = -17.8384; SE = 6.347; p = 0.005$). Once again, there was no significant effect for logDice score ($\beta = -0.476; SE = 6.314; p = 0.939$).

Models were compared using Akaike Information Criterion (AIC). The AIC score for the Gold Standard Model was 60517 and for the Human Judgements Model was 60533, suggesting a small difference in favour of the Gold Standard.

Discussion

The present study set out: (i) to conceptually replicate the processing trends of L1 speakers reported in L2 collocational processing studies, (ii) to investigate whether L1 speakers are aware of the semi-transparent meaning of a collocation, and (iii) if so, to determine whether literalness judgements could be used as a reliable method to identify collocations. Results from our experiments show that L1 speakers do indeed incur a processing cost for collocations over productive language, they are aware of the figurative meaning of the verb and literalness judgements can be used as a method to identify collocations. This has several implications, both theoretical and practical.

Idioms which are fully opaque (e.g., *kick bucket*, *break leg*) are processed faster and more accurately than productive language in both the L1 and the L2, a phenomenon often referred to as the *idiom superiority effect* (Noveck, Griffen, & Mazzarella, 2023). Studies investigating compositionality have shown that familiarity, which comes from frequent exposure, plays an important role in determining processing advantages for idioms in comparison to matched novel phrases (Tabossi, Fanari, & Wolf, 2009). This finding is in line with usage-based models such as those put forth by N. Ellis, Simpson-Vlach, and Maynard (2008), which posit that frequent exposure leads to chunking, i.e., holistic storage, retrieval, and access in short-term and long-term memory. Furthermore, there is evidence that formulaic units (in general) are encountered and used more frequently by L1 speakers and are therefore processed faster than productive language (N. Ellis et al., 2008; N. C. Ellis, 2008; Arnon & Snider, 2010; Carrol & Conklin, 2020; Wahl & Gries, 2018). Our results suggest that collocations do not enjoy this superiority effect as other subsets of formulaic language do, and as such, should not be classified under this umbrella term as is currently done.

The crux of the issue with collocations is understanding the "native-like selection" and the arbitrary restrictions on substitution that makes collocations difficult to predict *a priori*. The issue of understanding conventionalized structure in potentially arbitrary word-meaning mappings is reminiscent of the homonomy-polysemy continuum. Homonymous words are useful because they keep the inventory low and they are easily disambiguated across contexts (Piantadosi, Tily, & Gibson, 2012). Polysemous words can obey fixed indexical/relational/metonymical rules (e.g., food and animal) or they can have historic relations that to present day speakers are non-apparent (Port for Portugal; for dessert wine; for the docks). Understanding this relationship with regards to collocations might be useful for shedding light on these similar restrictions at the single word level. Furthermore, polysemy is useful as it enables languages to compress multiple concepts into individual word-forms thereby allowing for a compact lexicon in the face of limited cognitive resources such as memory (Xu, Malt, & Srinivasan, 2017). In collocations, a word is reused by mapping an existing sense from its literal domain to a figurative domain, based on structural similarities. For example, in the collocation *freeze accounts*, the verb *freeze* shares the sense of being able to revert to its original state with a productive use such as *freeze water*. Understanding how these mappings occur could further help uncover the underlying patterns in the "native-like selection" that we see in collocations.

From a methodological standpoint, the upshot of our work is that we now have a possible tool for testing novel semi-productive collocations. While we understand that selectional restrictions are somewhat arbitrary, at the same time they appear to cluster (e.g., abstract concepts). We can now try and understand novel constructions by looking at literalness judgement tasks and processing costs in acceptability judgements. Practically, our work shows that we can rely on L1 speaker judgements for large-scale identification of collocations. This could be used as tool to build extensive lists of collocations that can be used for more robust item sets in experiments, in textbooks for L2 learners, for computerised language teaching, for training NLP systems, etc.

Finally, our results from the literalness judgement task suggest that L1 speakers are indeed aware of the semi-transparent meaning of the collocations, whereas research shows that L2 speakers are unaware of the same (Martinez & Murphy, 2011; Laufer, 2011; Laufer & Waldman, 2011; Nesselhauf, 2004). We recommend that these semi-transparent lexical items should be explicitly taught to L2 learners.

Limitations & Future Work

Only Verb+Noun collocations were used in this study, future work should investigate other syntactic categories of collocations such as Noun+Noun, Adverb+Verb, and Adverb+Adjective. Furthermore, the stimuli were presented out of context. Many collocations are context-dependent and therefore should be tested using stimuli that in context. A possible confound in the experimental design, is the feedback during training. It could be that our participants were biased by our feedback. We plan to re-run this experiment with no feedback and compare the results to those of the present study. It must be noted that several items saw low agreement ratings with our gold standard. Although we do not perform a qualitative study of these results in the present paper, our future work will consider how judgements vary based on individual verbs and their verb classes (Levin, 1993).

Finally, our immediate next steps are to extend this study by testing synonyms of the figurative word in the collocation to further investigate the "native-like" selection procedure by comparing whether a novel collocation like *hunt dreams* would show a similar processing cost and literalness judgements as an existing one like *chase dreams*.

Conclusion

In the present study, we conceptually replicated the processing trends of L1 speakers reported in L2 collocational processing studies by means of an acceptability judgement task. Our results confirmed that L1 speakers do indeed incur a processing cost for collocations over productive language. We further investigated whether L1 speakers are aware of the semi-transparent meaning of a collocation using a novel literalness judgement task. We found that in contrast with L2 speakers, L1 speakers are indeed aware of the figurative meaning of the verb in the collocation and that there may be credence given to collocations as a valid psychological construct in the mind of a speaker. Finally, we attempted to determine whether literalness judgements could be used as a reliable method to identify collocations. Results from our experiments show that the literalness judgements provided by L1 speakers show broad agreement with our expert gold standard judgements, and therefore can be used to identify collocations.

Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP\S022481 \1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

References

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. (Publisher: Elsevier Inc.) doi: 10.1016/j.jml.2009.09.005

Barfield, A., & Gyllstad, H. (2009). Introduction: Researching L2 Collocational Knowledge. In H. Gyllstad & A. Barfield (Eds.), *Researching Collocations in Another Language: Multiple Interpretations* (pp. 1–18). London: Palgrave Macmillan UK.

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

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

Carrol, G., & Conklin, K. (2020, March). Is All Formulaic Language Created Equal? Unpacking the Processing Advantage for Different Types of Formulaic Sequences. *Language and Speech*, *63*(1), 95–122. doi: 10.1177/0023830918823230

Collins. (2018). *COBUILD Advanced Learner's Dictionary*. Glasgow, Scotland.

Cowie, A. P., & Howarth, P. (1996, March). Phraseology - a Select Bibliography. *International Journal of Lexicography*, *9*(1), 38–51. doi: 10.1093/ijl/9.1.38

Culicover, P. W., Jackendoff, R., & Audring, J. (2017, July). Multiword Constructions in the Grammar. *Topics in Cognitive Science*, *9*(3), 552–568. doi: 10.1111/tops.12255

Dankers, V., Bruni, E., & Hupkes, D. (2022). The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 4154–4175). Stroudsburg, PA, USA: Association for Computational Linguistics. (arXiv: 2108.05885) doi: 10.18653/v1/2022.acl-long.286

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, *26*(2), 163–188. doi: 10.1177/0267658309349431

Ellis, N., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 61–78. doi: 10.1515/CLLT.2009.003

Ellis, N. C. (2008, February). Phraseology: The periphery and the heart of language. In F. Meunier & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching* (pp. 1–13). John Benjamins Publishing Company. Retrieved 2024-01-30, from `https://benjamins.com/catalog/z.138.02ell` doi: 10.1075/z.138.02ell

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*(1), 29–62. (ISBN: 0165-4888/00/0020-0029)

Fioravanti, I., Senaldi, M. S. G., Lenci, A., & Siyanova-Chanturia, A. (2021). Lexical fixedness and compositionality in L1 speakers' and L2 learners' intuitions about word combinations: Evidence from Italian. *Second Language Research*, *37*(2), 291–322. doi: 10.1177/0267658320941560

Firth, J. (1957). A Synopsis of Linguistic Theory. In *Studies in Linguistic Analysis* (pp. 1–31). Oxford: Blackwell.

Ford, W. R., & Smith, R. N. (1982). Collocational Grammar as a Model for Human-Computer Interaction. In *Coling 1982 Abstracts: Proceedings of the Ninth International Conference on Computational Linguistics Abstracts.* Retrieved 2024-01-30, from `https://aclanthology.org/C82-2023`

Garner, J. (2022, September). The cross-sectional development of verb–noun collocations as constructions in L2 writing. *International Review of Applied Linguistics in Language Teaching*, *60*(3), 909–935. doi: 10.1515/iral-2019-0169

Gyllstad, H., & Wolter, B. (2016, April). Collocational Processing in Light of the Phraseological Continuum Model: Does Semantic Transparency Matter? *Language Learning*, *66*(2), 296–323. doi: 10.1111/lang.12143

Herbst, T. (2018). Is Language a Collostructicon? A Proposal for Looking at Collocations, Valency, Argument Structure and Other Constructions. In P. Cantos-Gómez & M. Almela-Sánchez (Eds.), *Lexical Collocation Analysis: Advances and Applications* (pp. 1–22). Cham: Springer International Publishing. Retrieved 2024-01-29, from `https://doi.org/10.1007/978-3-319-92582-0`$_1$ doi: 10.1007/978-3-319-92582-0$_1$

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1), 24–44. doi: 10.1093/applin/19.1.24

Koulouri, T., Lauria, S., & Macredie, R. D. (2016, January). Do (and Say) as I Say: Linguistic Adaptation in Human–Computer Dialogs. *Human–Computer Interaction*, *31*(1), 59–95. Retrieved 2024-01-30, from `https://doi.org/10.1080/07370024.2014.934180` (Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07370024.2014.934180) doi: 10.1080/07370024.2014.934180

Laufer, B. (2011). The contribution of dictionary use to the production and retention of collocations in a second language. *International Journal of Lexicography*, *24*(1), 29–49. doi: 10.1093/ijl/ecq039

Laufer, B., & Waldman, T. (2011, June). Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. *Language Learning*, *61*(2), 647–672. (Publisher: John Wiley & Sons, Ltd) doi: 10.1111/j.1467-9922.2010.00621.x

Levin, B. (1993). *English Verb Classes and Alternations: A preliminary investigation.* Chicago: University of Chicago Press.

Martinez, R., & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly*, *45*(2), 267–290. doi: 10.5054/tq.2011.247708

Mel'čuk, I. (2003). Collocations: définition, rôle et utilité. *Travaux et recherches en linguistique appliquée. Série E, Lexicologie et lexicographie.*(1), 23–31. (Num Pages: 9 Place: Amsterdam Publisher: Editions 'De Werelt')

Nation, P. (2001). The goals of vocabulary learning. In *Learning Vocabulary in Another Language* (pp. 6–25). Cambridge: Cambridge University Press. (ISSN: 08894906) doi: 10.1016/s0889-4906(02)00014-5

Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*, *24*(2), 223–242. (Place: Basel)

Nesselhauf, N. (2004). *Collocations in a Learner Corpus*. Philadelphia: John Benjamins Publishing Company.

Nishikawa, T. (2019, June). Non-nativelike outcome of naturalistic child L2 acquisition of Japanese: The case of noun–verb collocations. *International Review of Applied Linguistics in Language Teaching*(Lenneberg 1967). doi: 10.1515/iral-2018-0292

Noveck, I. A., Griffen, N., & Mazzarella, D. (2023). Taking stock of an idiom's background assumptions: an alternative relevance theoretic account. *Frontiers in Psychology*, *14*.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and Communication*, 191–226. (ISBN: 9781317869634) doi: 10.4324/9781315836027-12

Piantadosi, S. T., Tily, H., & Gibson, E. (2012, March). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. Retrieved 2024-02-02, from `https://www.sciencedirect.com/science/article/pii/S0010027711002496` doi: 10.1016/j.cognition.2011.10.004

Pulido, M. F., & Dussias, P. E. (2019, May). The Neural Correlates of Conflict Detection and Resolution During Multiword Lexical Selection: Evidence from Bilinguals and Monolinguals. *Brain Sciences*, *9*(5), 110. doi: 10.3390/brainsci9050110

R Core Team. (2023). *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `http://www.r-project.org` (Publication Title: R Foundation for Statistical Computing)

Rychlý, P. (2008). A lexicographer-friendly association score. In *RASLAN 2008 - Recent Advances in Slavonic Natural Language Processing: 2nd Workshop on Recent Advances in Slavonic Natural Language Processing, Proceedings* (pp. 6–9).

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2276*, 1–15. (ISBN: 3540432191) doi: 10.1007/3-540-45715-1$_1$

Seretan, V. (2018). Bridging Collocational and Syntactic Analysis. In P. Cantos-Gómez & M. Almela-Sánchez (Eds.), *Lexical Collocation Analysis: Advances and Applications* (pp. 23–38). Cham: Springer International Publishing. Retrieved 2024-01-29, from `https://doi.org/10.1007/978-3-319-92582-0`$_2$ doi: 10.1007/978-3-319-92582-0$_2$

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia, A., & Pellicer-Sanchez, A. (2018). *Understanding Formulaic Language*. London: Routledge. (Series Title: Second language acquisition research series) doi: 10.4324/9781315206615

Souza, S., & Chalmers, H. (2022). Processing Collocations in the L2: When semantic transparency & congruency collide. In *Proceedings of the 58th Annual Meeting of the Chicago Linguistics Society.* [In preparation].

Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast? *Memory & Cognition*, *37*(4), 529–540. Retrieved from `https://doi.org/10.3758/MC.37.4.529` doi: 10.3758/MC.37.4.529

Wahl, A., & Gries, S. T. (2018). Multi-word Expressions: A Novel Computational Approach to Their Bottom-Up Statistical Extraction. In P. Cantos-Gómez & M. Almela-Sánchez (Eds.), *Lexical Collocation Analysis: Advances and Applications* (pp. 85–109). Cham: Springer International Publishing. Retrieved 2024-01-29, from `https://doi.org/10.1007/978-3-319-92582-0`$_5$ doi: 10.1007/978-3-319-92582-0$_5$

Wolter, B., & Gyllstad, H. (2013). Frequency of Input and L2 Collocational Processing: A Comparison of Congruent and Incongruent Collocations. *Studies in Second Language Acquisition*, *35*, 451–482. doi: 10.1017/S0272263113000107

Wolter, B., & Yamashita, J. (2015). Processing collocations in a second language: A case of first language activation? *Applied Psycholinguistics*, *36*, 1193–1221. doi: 10.1017/S0142716414000113

Wolter, B., & Yamashita, J. (2018). Word Frequency, Collocational Frequency, L1 Congruency, and Proficiency in L2 Collocational Processing: What accounts for L2 performance? *Studies in Second Language Acquisition*, *40*(2), 395–416. (ISBN: 0272263117) doi: 10.1017/S0272263117000237

Wray, A. (2002). *Formulaic Language and the Lexicon* (Vol. 80). Cambridge: Cambridge University Press. (Publication Title: Language ISSN: 1535-0665) doi: 10.1353/lan.2004.0209

Xu, Y., Malt, B. C., & Srinivasan, M. (2017). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive Psychology*, *96*, 41–53. Retrieved 2024-01-30, from `https://www.sciencedirect.com/science/article/pii/S001002` doi: 10.1016/j.cogpsych.2017.05.005

Čermák, F. (2006). Collocations, Collocability and Dictionary. In *Proceedings of XII EURALEX International Congress* (pp. 929–937).