What does memory retrieval *leave on the table*? Exploring Semantic Compositionality in Language Processing with MINERVA2 and sBERT

Anonymous ACL submission

Abstract

Despite being ubiquitous in natural language, collocations (e.g., *kick+habit*) incur a unique processing cost, compared to compositional phrases (*kick+door*) and idioms (*kick+bucket*). We confirm this processing cost with behavioural data as well as MINERVA2, a memory model, suggesting that collocations constitute a distinct linguistic category. While the model fails to fully capture the observed human processing patterns, we find that below a specific item frequency threshold, the model's retrieval failures align with human reaction times across conditions. This suggests an alternative processing mechanism that activates when memory retrieval fails, consistent with an analogical account of language processing.

1 Background

From *killing time* and *playing dead* to *running baths* and *making beds*, word combinations with semi-compositional meanings are ubiquitous in human language (Cowie, 1998). Often referred to as *collocations*, these idiosyncratic lexical elements comprise one word used in its literal sense and another in its figurative sense, constrained by an arbitrary restriction on substitution (Mel'čuk, 2003; Howarth, 1998). To illustrate, one can *raise questions* or *lift bans*, but neither #*lift questions* nor #*raise bans*. Collocations are syntactically well formed, but deviate from or violate the expected semantic representation (Culicover et al., 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*. Collocations are the largest subset of formulaic language (Barfield and Gyllstad, 2009a) with many being cross-linguistically attested (Yamashita, 2018). It is hardly surprising, then, that proper knowledge and use of such units provides fluency and idiomaticity to the language user (Pawley and Syder, 1983; Durrant and Schmitt, 2009). However, they pose an enormous hurdle to second-language learners and machines.

According to Howarth (1998), human language lies on a theoretical continuum of semantic compositionality—the degree to which the meaning of a phrase can be derived from the meaning of its constituent parts and their syntactic relations (Frege, 1892). Fully compositional combinations (e.g., *chase rabbits*, *chase thieves*, etc.) and fully non-compositional figurative idioms (e.g., *chase one's tail*, *chase rainbows*)[1] lie on extreme ends of the spectrum. Semi-compositional collocations (e.g., *chase dreams*, *chase money*, etc.) lie in between. The psychological validity of this continuum has been tested with the expectation that a decrease in compositionality is directly proportional to a decrease in processing time (Gyllstad and Wolter, 2016). However, empirical evidence shows that while collocations are processed slower and less accurately than fully compositional combinations (Gyllstad and Wolter, 2016; de Souza et al., 2024), fully opaque and non-compositional figurative idioms (e.g., *break the ice*) are processed faster and more accurately than compositional combinations (e.g, *break the cup*) (Carrol and Conklin, 2020; Tabossi et al., 2008). It appears that idioms are processed the fastest, followed by compositional units, and collocations are processed the slowest.

This disparity is also reflected in acquisition. Applied Linguistics research shows that second language (L2) learners—be they early sequential bilinguals (Nishikawa, 2019; Riches et al., 2022) or

[1] It is important to note that (Howarth, 1998) also specifies a fourth category called "pure idioms" (e.g., *blow the gaff*, *take a leak*, *shoot the breeze*). These do not possess well-specified literal meanings (see Mueller and Gibbs, 1987, for further reading) and comprise a very small subset of formulaic language occurring quite infrequently (Grant, 2005). Furthermore, most of the studies in this area focus on figurative idioms that have an additional literal reading (e.g., *kick the bucket*). Therefore, in order to constrain the scope of this paper, we limit our discussion to figurative idioms.

adults (Yamagata et al., 2023; Sonbul et al., 2024), even at high proficiency levels (Wolter and Gyllstad, 2013; Tsai, 2020)—have trouble acquiring and using collocations. In contrast, idioms are learned better and used more accurately than collocations (Fioravanti et al., 2021). Cast under the broader term of *conceptual metaphor* (Lakoff and Johnson, 1980), collocations are also found to be challenging for NLP systems (Liu et al., 2022; Zayed et al., 2018; Tayyar Madabushi et al., 2021) despite the fact that the last decade has seen immense progress (see Tong et al., 2021, for a review).

Although collocations are an important subset of human language, to the best of our knowledge there is no model of language processing that specifically accounts for collocations.

2 Accounting for Collocation Processing

It is generally agreed in the language processing literature that idioms are stored and retrieved from memory holistically (Carrol and Conklin, 2014; Noveck et al., 2023; Luo, 2019). There is less consensus on how compositional language is processed, and collocations are largely ignored. Drawing on ideas from the (in)famous Past Tense Debate in morphological processing, researchers in Applied Psycholinguistics have resorted to single- versus dual-route models to explain processing at the multi-word level (Wray, 2002).

Single-route and dual-route models. Assuming a domain-general hypothesis space, single-route models posit that all linguistic forms are stored in and retrieved from a single massive associative memory system[2] based on frequency of input and use (Bybee, 2012; Ambridge and Lieven, 2011). The more often a unit is encountered and/or used, the better it is entrenched in memory (Divjak, 2019; Langacker, 1987). Eventually, this leads to automatization—pure retrieval from memory[3] (Bybee, 2006) which makes processing fast and effortless. Positing such a homogenous mechanism makes for a parsimonious theoretical account of our language abilities, in particular, and our cognition in general. However, human memory is not only limited in capacity (Christiansen and Chater, 2008) but is also unstable (Kornell and Bjork, 2009). More importantly, behavioural evidence shows that collocations incur a processing cost versus compositional units even when frequency-matched (see de Souza et al., 2024). While memory undoubtedly plays an important role in language processing, it does not provide a satisfactory account for the processing cost of collocations which are frequently-occurring linguistic units (Barfield and Gyllstad, 2009b).

The dual-route model assumes a domain-specific hypothesis space, differentiating between words and rules (Pinker, 1991). Regular word forms are thought to be computed analytically (e.g., *walk* → *walk* + *ed*, *scratch* → *scratch* + *ed*) by way of rules, while irregular word forms (e.g., *run* → *ran*, *think* → *thought*) are processed via holistic storage and retrieval from memory (Pinker, 2013).

This theoretical distinction between computation and storage is a practical trade-off between two independent cognitive processes—procedural computation and declarative memory (Pinker and Ullman, 2002). More on-the-fly, rule-based computation means less storage. More storage means less computation. Positing such a heterogenous mechanism makes for a persuasive theoretical account of how human language can be infinitely compositional despite our limited cognitive capacities (O'Donnell et al., 2009; Galke et al., 2024). The dual-route explanation is used to account for formulaic language processing as a whole, i.e., it does not distinguish between the various subsets of multi-word units such as idioms, phrasal verbs, binomials, etc. (see Wray, 2002, 2008; Sidtis, 2020). All formulaic language is thought to be stored, while compositional language is computed on the fly. Memory retrieval is faster than analytic processing (Logan, 1997; Dasgupta and Gershman, 2021), therefore formulaic language is thought to be processed faster than non-formulaic language (Carrol and Conklin, 2014; Vilkaite and Schmitt, 2019). This is empirically consistent across a variety of tasks only in the case of fully non-compositional units like idioms (Noveck et al., 2023). However, dual-route hypotheses make a binary distinction between compositional and formulaic language which does not consider the effect of frequency on computation and retrieval. If collocations are frequent and retrieved from memory, the processing cost is unpredicted unlike the processing advantage for idioms.

Analogy. Neither the single-route nor the dual-route views satisfactorily explain collocational pro-

[2] Or that all forms are processed equally as in a connectionist network (see McClelland and Rumelhart, 1985).

[3] See Logan and Etherton (1994) for a domain-general cognitive account of automatization.

cessing, underscoring the need for a model which can account for a more fine-grained representation of semantic compositionality. One such plausible mechanism is analogical reasoning (Eddington, 2000; Ambridge, 2020). Like single-route models, this domain-general approach posits that all linguistic units are processed by a single mechanism (Skousen, 1990). However, in addition to memory retrieval, it posits on-the-fly analogy without resorting to any rule-based mechanisms. On receiving an input, a memory search is undertaken to find analogous exemplars previously experienced. The input is then evaluated based on the degree of similarity in order to find the most frequent category within the found set of most similar exemplars (Gentner and Namy, 2006). In principle, analogy is unbounded (Blevins and Blevins, 2009)—similarities can be detected at various levels ranging from perceptual (e.g., noticing the shared feature between red car and red flower) to relational (e.g., noticing that two objects from two different rows of objects share the middle position) and multiple mappings can be made (Smet and Fischer, 2017; Gentner and Markman, 1997). This property makes analogical reasoning a powerful processing mechanism as it supports flexible cross-domain generalization (Doumas et al., 2022). Proponents of this view see analogy as the core driver of human cognition (Hofstadter and Sander, 2013; Hofstadter, 1982).

Analogy allows speakers to make abstract generalizations from known patterns to novel structures, providing a flexible account for a variety of phenomena ranging from narrative building (Fish et al., 2024) to language change (Smet and Fischer, 2017). However, analogy is effortful (Gick and Holyoak, 1980; Noveck et al., 2023), and it is unlikely that a language user is using on-the-fly analogical processes every single time an input is encountered. Furthermore, an adequate model of analogy must be constrained enough to explain why speakers generalize over certain relations and not others and it is hard to predict which analogies will actually be drawn (Albright, 2009; Dunbar).

3 The Present Study

Based on the review above, it would be uncontroversial to say that memory is critical to all forms of language processing (see also Divjak, 2019; Divjak et al., 2022; Corballis, 2019). It encapsulates single-route processes, is an integral component of dual-route models, and is the first step in analogy (Gentner and Colhoun, 2010). Therefore, as a first step towards a model that accounts for collocational processing, we test the extent to which simple memory retrieval is sufficient to reproduce human processing trends.

We begin by confirming the trends, which we surmise from the literature. To the best of our knowledge, there exists no behavioural study which has investigated the processing of compositional units, collocations and idioms in the same task. We do so by extending de Souza et al. (2024) and test L1 English speakers on an acceptability judgement task (AJT) using stimuli from all three conditions. We analyse reaction times (RTs) and accuracy.

Next, we simulate memory retrieval under two empirically ascertained factors that affect collocational processing: frequency (Wolter and Gyllstad, 2013) and semantics (Gyllstad and Wolter, 2016; Fioravanti et al., 2021). We implement a well-established frequency-based model of memory—MINERVA2 (Hintzman, 1984), and adopt distributional semantic representations (Landauer and Dumais, 1997; Mikolov et al., 2013). We modify MINERVA to simulate RTs and load its memory with contextualized vector representations from Sentence-BERT (Reimers and Gurevych, 2019), according to the frequency of the stimuli in the corpus. We explore successful and failed retrievals to assess their influence on the processing signatures of different item conditions. Our research questions are as follows: **(1)** Is there a statistically significant difference in processing speed and accuracy between compositional items, collocations and idioms? **(2)** Does MINERVA replicate the processing trends observed in humans?

In keeping with the literature, we expect idioms to be processed fastest, followed by compositional items, with collocations being the slowest. We expect MINERVA to also show differences between conditions. However, based on the review above, we do not expect it to match human processing signatures. Instead, we expect to see only frequency effects. Given that all items would be familiar to an L1 English speaker and present in MINERVA's memory, we expect no differences in accuracy.

4 Behavioural Experiment

4.1 Methodology

Stimuli de Souza et al. (2024) introduced a stimulus set consisting of 100 Verb-Noun collocations (e.g., *spill secrets*) and 100 compositional Verb-

Noun combinations containing the same verb as the collocation (e.g., *spill water*). We attempted to augment this stimulus set with a matching figurative idiom (e.g., *spill the beans*) for each verb with the help of the 'word sketch' function in The Sketch Engine's enTenTen21 corpus (Kilgarriff et al., 2024). However, we were only able to identify idioms for 82 verbs in the dataset resulting in a final dataset of 246 target items (1 collocation, one composition, and one idiom for each of the 82 verbs). 82 baseline items, nonsense Verb-Noun combinations (*fry knob*), were created to use as distractors in the experiment. The dataset was divided into 3 folds of 82 items wherein no two items had the same verb. As expected, there are statistically significant differences between the mean frequencies of all three constructions with idioms being the most frequent, followed by collocations and compositional items being the least frequent group (see Appendix C for more details). We account for this discrepancy by including frequency as a covariate in our statistical models.

Participants & Task A total of 186 L1 English speakers ($F = 112$; $M = 71$; $NB = 3$) were recruited using Prolific. They were remunerated £1.50 for their participation[4]. The mean age of the sample was 38.6 years ($SD = 10.81$). They 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). They were asked to respond as quickly and accurately as possible, by pressing the 'y' key for yes or the 'n' key for no. During testing, each participant saw 164 items: 82 target items and 82 distractors. Items were presented in an individualized random order. 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.

Data Pre-processing Data pre-processing was carried out using R version 4.4.1 "Race for Your Life" (R Core Team, 2024). Due to an error in data collection, data of four participants were replaced. We also remove all incorrect trials for reaction time analyses ($2,752$; s.f. Appendix C.1. We then eliminated responses below 450 ms and responses over 3.5 standard deviation from the grand mean including time-outs. These outliers accounted for 1.484% of the total data ($n = 30,504$ including distractors).

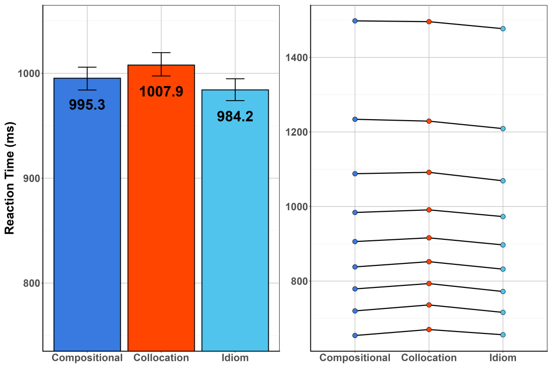

Figure 1: **Left:** mean reaction times (ms) by condition. Error bars indicate bootstrapped confidence intervals. **Right:** decile plot of reaction times by condition. Note the differences in the y-axes.

In terms of accuracy, all participants scored above 50%. However, we found 4 items with a mean accuracy of less than 50%. We eliminated those items along with other items that comprised the same verbs from our analyses. We do not analyse distractors ($15,252$). All reaction time (RT) analyses are conducted on this final dataset ($n = 13,369$).

4.2 Statistical Modelling

We first specified a maximal model as "justified by the design" (Barr et al., 2013). The main dependent variable was the reaction times (RTs) from the acceptability judgement task while the main predictor variable was Condition (Compositional, Collocation, Idiom; treatment coded, with idiom as the reference level). Phrasal Frequency (scaled) was included as a covariate. The maximal converging random effect structure included intercepts for Participant and Verb. The analysis model in R syntax specified using the 'lme4' (Bates et al., 2015) package is as follows:

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

4.3 Results

Figure 1 shows the mean reaction times (RTs) by condition, as well as a breakdown by decile. Collocations have the slowest responses with a mean of 1007.87 ms ($SD = 370.84$ ms) compared to compositional items (995.32 ms, $SD = 375.76$ ms) and idioms (984.20 ms, $SD = 365.39$ ms).

Our statistical results showed a small, significant difference in RTs between compositional items and idioms ($\beta = 4.69; SE = 2.240; p = 0.037$), suggesting that compositional units were

[4] The study received ethics approval.

processed slower than idioms. A larger difference was found between collocations and idioms ($\beta = 13.80; SE = 1.760; p < 0.001$), replicating the processing costs predicted by the literature. Unsurprisingly, Phrasal Frequency also has a significant effect on RTs ($\beta = -18.50; SE = 1.640; p < 0.001$), corresponding to a 18.5 ms decrease in RT for every 1 standard deviation increase in phrasal frequency. In terms of accuracy, we found no significant difference between idioms and compositional items, but we do see a marginal difference ($p = 0.04$) between idioms and collocations. See Appendix C.1 for detailed results.

5 Modelling Memory Retrieval with MINERVA

As a first step toward elucidating the cognitive mechanisms underlying the processing trend that humans display across the compositionality continuum, we investigate the extent to which we can account for the trend with memory retrieval alone. MINERVA is an instance-based model of episodic memory that has been successfully applied to many cognitive phenomena from frequency judgements (Hintzman, 1988) to false memory (Arndt and Hirshman, 1998). It has also been used to model artificial grammar learning (Jamieson and Mewhort, 2009) and, recently, to metaphor recognition (Nick Reid and Jamieson, 2023).

MINERVA's core assumptions are: (i) every item encountered leaves a memory trace, represented as a distributed set of features, and (ii) similar items have similar traces. Similarities between present and past encounters drive item-specific and parallel memory retrieval. As a global memory model, it encapsulates both episodic and semantic memory which communicate with each other. On encountering a stimulus, the episodic memory sends a probe to the semantic memory to retrieve traces from past encounters. The familiarity of the probe is then calculated as the sum of the values of a similarity measure between the probe and each stored trace.

MINERVA is instantiated in a linear algebra system. The MINERVA memory $\mathbf{M}$ is an $n \times d$ matrix, each row of which contains a d-dimensional memory trace vector. When cued for retrieval with a probe $p \in \mathbb{R}^d$, MINERVA retrieves the representation of the probe iff the probe's familiarity f is greater than a threshold $K \in [0, 1)$. Familiarity is calculated by taking the cosine similarity s of the probe to all instances stored in memory, scaling s to reflect activation (weighting) of memory items a over elapsed time τ, and linearly combining instances in memory to compute a memory echo e. The familiarity score at timestep τ is the cosine similarity of the echo to the probe, following this system of equations:

$$s = \text{sim}(p, \mathbf{M}) \tag{1}$$

$$a_\tau = s^\tau \text{sign}(s) \tag{2}$$

$$e_\tau = a_\tau \mathbf{M} \tag{3}$$

$$f_\tau = \text{sim}(e_\tau, p) \tag{4}$$

Modelling AJT Responses with Taus (τ) The free parameter τ is used to accentuate differences in similarity values (Hintzman, 1988; Nick Reid and Jamieson, 2023). By raising the value of τ, higher-similarity memory traces will elicit exponentially more activation, allowing those traces to play a larger role in the overall activation profile versus pooling a potentially large number of low-similarity items.

Following Nick Reid and Jamieson (2023), we depart from prior work wherein τ is kept constant for a particular experiment and model reaction times by dynamically increasing τ for a particular probe p until a desired threshold of familiarity $K \in [0, 1)$ is reached. At this point, we take the final value of τ as a proxy for the time required to recognize p from memory, i.e, a proxy for reaction time (RT). We set a time-out at $\tau = 300$ after which the next probe is presented.

In human acceptability judgements, reaction times serve as a proxy for processing difficulty. We implicitly model acceptability judgements in MINERVA as a function of whether the familiarity threshold K is reached within the allowable time window. If the familiarity score surpasses K before the time-out, i.e., successful recognition, we treat this as a "yes". Conversely, if familiarity remains below the threshold when $\tau = 300$, we treat the failure to retrieve as a "no" response.

5.1 Motivations & Assumptions

Collocational processing is known to be driven by two factors: *semantic transparency* and *frequency* (see Gyllstad and Wolter, 2016; Fioravanti et al., 2021). Our model captures semantic transparency by means of distributional semantics, i.e, vector embeddings, while frequency is captured by means of phrasal frequency in an Internet-wide text corpus. We demonstrate the effect of both factors in our ablations in Appendix E.

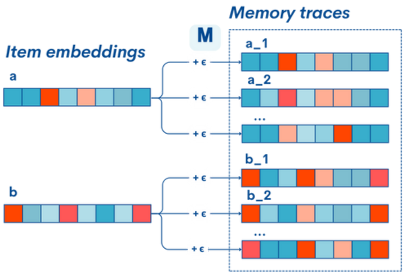

Figure 2: Illustration of how embeddings are noised and loaded into MINERVA's memory matrix M. Colors depict values within a vector. Note that the noise vectors ϵ are independently sampled for each memory trace.

Semantics of Memory Traces Using distributed vector representations as memory traces for MINERVA is well-established in the literature (Chubala and Jamieson, 2013; Jamieson et al., 2018; Nick Reid and Jamieson, 2023). Given that the figurative idioms (e.g., *spill the beans*) also have a compositional reading, we need a contextualized, fine-grained vector representation to capture the semantics of each word combination. Therefore, we rely on Sentence-BERT (sBERT) which provides semantically meaningful vector embeddings for sentences (Reimers and Gurevych, 2019). To derive the vector embedding for each of the 246 target stimuli, we follow Vulić et al. (2020). First, we collect a set of 100 sentences of the word combination[5] from the enTenTen21 corpus, in which the noun occurs as the direct object of the verb. We feed each sentence to sBERT obtaining a set of contextualized word embeddings representing each word in the sentence (we perform mean pooling over sub-words). Given that the higher layers of BERT architectures are the most sensitive to lexical semantics (Reif et al., 2019), we take our embeddings from the last hidden layer of the model. From each of the 100 sentences, we extract the embeddings corresponding to the verb and the noun and average across them separately, resulting in the mean contextualized representation of the verb when paired with the noun, and of the noun when paired with the verb. Finally, we concatenate the mean embedding for the verb with the mean embedding for the noun to form the vector representation of our stimulus[6].

Memory Frequencies & Forgetting In accordance with the instance theory, MINERVA's retrieval time is inversely proportional to the number of memory traces that strongly respond to a particular probe (Nick Reid and Jamieson, 2023). Therefore, we populate MINERVA's memory matrix using $10,000$ items sampled proportionally to their phrasal frequency. Following prior work, we simulate forgetting by adding zero-centered Gaussian noise to each memory trace vector such that each dimension of each trace has an independent probability $F \in [0, 1)$ of being corrupted with noise. The more frequent a particular item, the more traces it will have in memory, averaging out the noise and making high-frequency items easier to retrieve.

5.2 Simulations

To explore the extent to which simple memory retrieval is sufficient to reproduce processing trends for each condition, we load the memory matrix as described above (see Figure 2) and then test MINERVA's recognition capabilities using a noiseless vector embedding of the target stimulus as the probe. To simulate N different participants who are exposed to different samplings of items from the same environmental distributions, as well as different patterns of forgetting, we run each simulation $N = 300$ times with different random seeds, re-sampling and re-noising the memory matrix each time. We perform a thorough hyperparameter sweep of activation threshold K and forgetting probability F to ensure robustness. We discuss results for hyperparameter values $K = 0.99$ and $F = 0.8$, although our results are robust across many hyperparameter combinations (see Figure 9).

We use the same statistical model described in Section 4 to analyse the effect of semantics and frequency on retrieval (i.e., Tau).

5.3 Results

The results of our computational experiment are shown in Figure 3. As MINERVA was not presented with any baseline items and as all items were in MINERVA's memory, it should have succeeded at recognizing all items. Thus, we first considered only successful retrievals. Despite being provided with meaningful embeddings and frequencies, the model failed to capture human processing trends. Collocations were retrieved faster than idioms ($\beta = -0.41; SE = 0.004; p < 0.001$) while

[5] Distractor items were not included in the simulations as they are nonsense combinations, have no context sentences and would have very low frequency in MINERVA's memory.

[6] We use concatenation instead of mean pooling as our stimuli are all Verb + Direct Object and concatenation preserves word order and therefore, syntactic role information. However, see Section E.3.

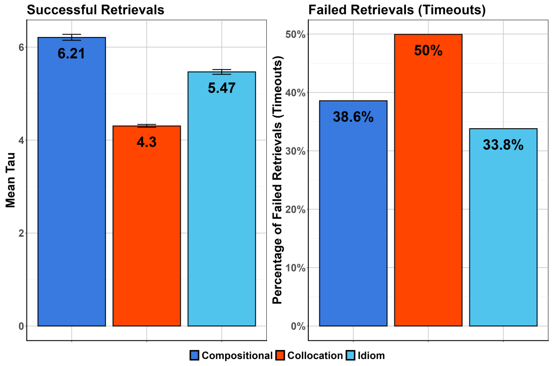

Figure 3: **Left:** mean Tau (τ) by condition for successful retrievals in MINERVA. The y-axis represents mean Tau, the model's output which acts as a proxy for reaction times. Error bars indicate bootstrapped confidence intervals. **Right:** percentage of failed retrievals, i.e., timeouts, per condition. Note that while the pattern of Taus on *successful retrievals* is different from the pattern of human RTs, the pattern of *timeouts* per condition matches the pattern of human RTs.

compositional items were retrieved slower than idioms ($\beta = 0.62; SE = 0.004; p < 0.001$). See Appendix D for more details.

Given the surprising results, we visually inspected the failures to retrieve, i.e., timeouts (see 3, right panel). MINERVA timed out on 50% of the retrievals for collocations, followed by compositional items (38.6%), with idioms timing out the least (33.8%). The pattern of retrieval failures in MINERVA appears to qualitatively capture the trend in human RTs across the three conditions.

Additionally, we found that MINERVA always succeeds at retrieving items above a high frequency threshold (Figure 4, black line). We find a similar frequency boundary in humans (Figure 4, green line), which lies very close to the MINERVA threshold. On items above this threshold (16 compositional, 18 collocations, 17 idioms), participants did not show a significant difference in RT by condition, while still showing a significant effect of frequency.

6 Discussion

Our behavioural results show that idioms are processed fastest with compositional items coming a close second (although the significance was marginal) and that collocations are processed by far the slowest. This effect occurs despite collocations and compositional items being very close in frequency (with the balance in favour of collocations). The result is mirrored in our computational findings. These stark differences in processing speed for collocations indicate that they should be treated as a separate class of linguistic items, apart from the broad umbrella of formulaic language.

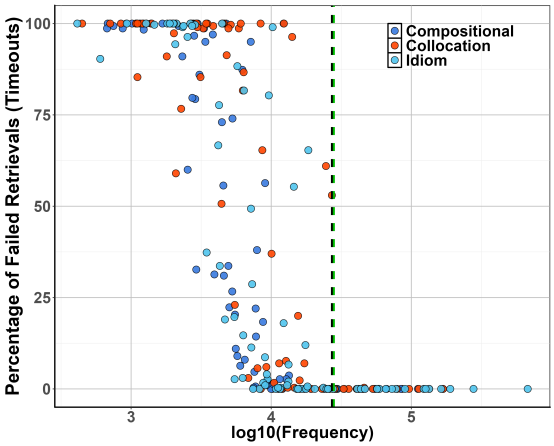

Figure 4: Percentage of failed retrievals (i.e., timeouts) in MINERVA per stimulus item, as a function of the frequency of the item. The x-axis is displayed in log scale. The black line indicates the frequency threshold ($f = 27123$) above which MINERVA times out less than 1% of the time. The green line ($f = 28000$) indicates the frequency threshold above which condition stops being a significant predictor of human RTs.

Our simulation results suggest that simple memory retrieval, as implemented in a frequency based model of memory, is insufficient to fully explain human processing trends across idioms, collocations, and compositional items. MINERVA was fastest to retrieve collocations, followed by idioms, and finally compositional items were the slowest. Furthermore, MINERVA exhibited many more incorrect responses, i.e., unsuccessful memory retrievals. However, unlike the pattern of Taus for successful retrievals, retrieval failures do appear to capture the key asymmetries in human processing. Once again this effect is especially noticeable for collocations. These findings indicate that additional cognitive mechanisms may be required to fully account for human behavioural patterns.

Notably, both asymmetries—retrieval failures in MINERVA and reaction times in humans—only occur below a certain frequency threshold. We suggest that items above this threshold are sufficiently frequent so as to be holistically retrieved across conditions. Below this threshold, retrieval starts to fail. Given that MINERVA does not have any processing mechanism beyond memory retrieval, it simply times out on these items. We conjecture that at this point, humans invoke other processing mechanisms to facilitate interpreting of the stimulus and incur a cost in reaction time. This is consistent with usage based theories of language, which posit that frequent encounters with a linguistic units lead to entrenchment (Langacker, 1987).

The fact that collocations incur such a processing cost compared to frequency-matched idioms and compositional items show that single-route accounts provide an incomplete picture. They further demonstrate that dual-route accounts, which posit a binary class of formulaic versus compositional language, are also insufficient to account for the processing of this large and frequent subset of language. We posit that that the analogical account of language processing may provide a more complete explanation of these findings, and that further work should explore this proposal.

As discussed above, memory retrieval is the first step in analogical processing. Hence, processing a sufficiently frequent item via analogy will simply resort to memory retrieval

. Such a mechanism would be invariant to the semantic compositionality of the item in question, as we have seen in humans. Below this frequency threshold, however, proper analogical machinery comes into play.

In compositional items, both the verb and the noun play a prototypical role. Thus, even though the language user may not recall this exact verb-noun pairing from memory, it is relatively easy to map the verb and noun to similar instances of the same, due to the high semantic overlap between compositional uses of the verb and the noun. In collocations, however, the verb is not used in its prototypical sense. Resolving the meaning of the verb requires a much "farther" mapping, which may involve increased search over possible abstractions of the verb or extensive structure-mapping. Engaging such machinery inevitably incurs a processing cost with respect to compositional items (Gentner and Namy, 2006), as reflected in RTs.

Despite the fact that idioms also time out, we hypothesize that they are still processed via memory retrieval even below the frequency threshold. At first, this appears contradictory. However, note that the idioms in our dataset are also highly frequent in the corpus. We speculate that when encountering a novel idiom, the learner implicitly has a choice to either memorize the idiom, or discard it entirely. By choosing idioms that are highly frequent in the corpus, our dataset is conditioned upon learners already having memorized the idioms. The lack of an analytic mechanism makes it difficult to interpret novel idioms. Furthermore, the fact that every learned idiom incurs a fixed cost in remembering it may account for the relative sparsity of idioms in language, i.e., lower type frequency, and their high frequency of occurrence, i.e., higher token frequency (Grant, 2005) .

The retrieval failures for idioms seem to stem from a limitation of our dataset—the fact that we only consider figurative idioms which have a compositional reading. We were unable to ascertain the relative frequency of idiomatic versus literal readings in the context sentences of every idiom in our stimuli set which we use to generate embeddings. It is also unknown to what precise extent sBERT can accurately represent idiomatic meanings, nor whether our human participants who interpreted idiomatic stimuli in a figurative sense. Combined, these factors suggest that the semantics of our set of idioms are somewhat akin to our set of compositional items, and some of the processing trends which pertain to compositional items are inadvertently present in the trend of responses to idioms. In line with the holistic retrieval hypothesis, we surmise that idioms for which the literal reading is much less frequent than the idiomatic one (e.g., *kick the bucket*) will tend to be processed faster and with fewer timeouts than more ambiguous ones (e.g., *hold the key*). Future work will attempt to investigate this prediction and further augment our understanding of idiomatic processing by including pure idioms, i.e., those without a literal reading, in the dataset, and employing other behavioural tasks which involve presentation of items within context (e.g., self-paced reading).

One intriguing implication of our computational experiment may be of interest to the NLP community. Specifically, MINERVA's retrieval mechanism bears similarities to retrieval-augmented generation (RAG) approaches (Lewis et al., 2020, i.a.). Our findings suggest that sBERT embeddings of semi-compositional language are particularly prone to failures in retrieval. Given the prevalence of collocations in language, this may significantly impair language understanding and generation in RAG models.

Overall, the present study demonstrates that for both humans and machines, collocations are a bigger "pain in the neck" (Sag et al., 2002) than other subsets of the semantic compositionality continuum, and that memory retrieval does *leave something on the table*.

References

Adam Albright. 2009. Modeling analogy as probabilistic grammar. In James P. Blevins and Juliette Blevins, editors, *Analogy in Grammar: Form and Acquisition*, pages 185–213. Oxford University Press.

Ben Ambridge. 2020. Against stored abstractions: A radical exemplar model of language acquisition. 40(5-6):509–559.

Ben Ambridge and E. Lieven. 2011. *Child language acquisition: Contrasting theoretical approaches*. Cambridge University Press, Cambridge.

Jason Arndt and Elliot Hirshman. 1998. True and False Recognition in MINERVA2: Explanations from a Global Matching Perspective. *Journal of Memory and Language*, 39(3):371–391.

Andy Barfield and Henrik Gyllstad. 2009a. Introduction: Researching L2 Collocational Knowledge. In Henrik Gyllstad and Andy Barfield, editors, *Researching Collocations in Another Language: Multiple Interpretations*, pages 1–18. Palgrave Macmillan UK, London.

Andy Barfield and Henrik Gyllstad, editors. 2009b. *Researching Collocations in Another Language*. Palgrave Macmillan UK, London.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278. Publisher: Elsevier Inc.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

James P. Blevins and Juliette Blevins. 2009. Introduction: Analogy in grammar. In James P. Blevins and Juliette Blevins, editors, *Analogy in Grammar: Form and Acquisition*, page 0. Oxford University Press.

Joan Bybee. 2006. From usage to grammar: The mind's response to repetition. *Language*, 82(4).

Joan Bybee. 2012. A usage-based perspective on language. In *Language, Usage and Cognition*, pages 1–13. Cambridge University Press, Cambridge.

Gareth Carrol and Kathy Conklin. 2014. Getting your wires crossed: Evidence for fast processing of L1 idioms in an L2. *Bilingualism: Language and Cognition*, 17(4):784–797. Publisher: Cambridge University Press.

Gareth Carrol and Kathy Conklin. 2020. Is All Formulaic Language Created Equal? Unpacking the Processing Advantage for Different Types of Formulaic Sequences. *Language and Speech*, 63(1):95–122.

Morten H. Christiansen and Nick Chater. 2008. Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489–509.

Chrissy M. Chubala and Randall K. Jamieson. 2013. Recoding and representation in artificial grammar learning. *Behavior Research Methods*, 45(2):470–479.

Michael C. Corballis. 2019. Language, Memory, and Mental Time Travel: An Evolutionary Perspective. *Frontiers in Human Neuroscience*, 13. Publisher: Frontiers.

Anthony Paul Cowie. 1998. *Phraseology : theory, analysis, and applications*. Clarendon Press, Oxford. Series Title: Oxford studies in lexicography and lexicology.

Peter W. Culicover, Ray Jackendoff, and Jenny Audring. 2017. Multiword Constructions in the Grammar. *Topics in Cognitive Science*, 9(3):552–568.

Ishita Dasgupta and Samuel J. Gershman. 2021. Memory as a Computational Resource. *Trends in Cognitive Sciences*, 25(3):240–251.

Sydelle de Souza, Francis Mollica, and Jennifer Culbertson. 2024. What can L1 speakers tell us about killing hope? A Novel Behavioral Measure for Identifying Collocations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0).

Dagmar Divjak. 2019. *Frequency in Language: Memory, Attention and Learning*. Cambridge University Press, Cambridge.

Dagmar Divjak, Petar Milin, Srdan Medimorec, and Maciej Borowski. 2022. Behavioral Signatures of Memory Resources for Language: Looking beyond the Lexicon/Grammar Divide. *Cognitive Science*, 46(11):e13206.

Leonidas A. A. Doumas, Guillermo Puebla, Andrea E. Martin, and John E. Hummel. 2022. A theory of relation learning and cross-domain generalization. *Psychological Review*, 129(5):999–1041.

Kevin Dunbar. The Analogical Paradox: Why Analogy Is so Easy in Naturalistic Settings, Yet so Difficult in the Psychological Laboratory.

Philip Durrant and Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *IRAL - International Review of Applied Linguistics in Language Teaching*, 47(2):157–177.

David Eddington. 2000. Analogy and the dual-route model of morphology. *Lingua*, 110(4):281–298.

Irene Fioravanti, Marco Silvio Giuseppe Senaldi, Alessandro Lenci, and Anna Siyanova-Chanturia. 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.

Robert D Fish, Gail E Austen, Jacob W Bentley, Martin Dallimer, Jessica C Fisher, Katherine N Irvine, Phoebe R Bentley, Maximilian Nawrath, and Zoe G Davies. 2024. Language matters for biodiversity. *BioScience*, 74(5):333–339.

Gottlob Frege. 1892. *Über Sinn und Bedeutung*, 1. auflage edition. Zeitschrift für Philosophie und philosophische Kritik, Neue Folge. Pfeffer, Leipzig.

Lukas Galke, Yoav Ram, and Limor Raviv. 2024. Deep neural networks and humans both benefit from compositional language structure. *Nature Communications*, 15(1):10816. Publisher: Nature Publishing Group.

Dedre Gentner and Julie Colhoun. 2010. Analogical Processes in Human Thinking and Learning. In Britt Glatzeder, Vinod Goel, and Albrecht Müller, editors, *Towards a Theory of Thinking: Building Blocks for a Conceptual Framework*, pages 35–48. Springer, Berlin, Heidelberg.

Dedre Gentner and Arthur B. Markman. 1997. Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45–56. Place: US Publisher: American Psychological Association.

Dedre Gentner and Laura L. Namy. 2006. Analogical Processes in Language Learning. *Current Directions in Psychological Science*, 15(6):297–301. Publisher: SAGE Publications Inc.

Mary L Gick and Keith J Holyoak. 1980. Analogical problem solving. *Cognitive Psychology*, 12(3):306–355.

Lynn E. Grant. 2005. Frequency of 'core idioms' in the British National Corpus (BNC). *International Journal of Corpus Linguistics*, 10(4):429–451. Publisher: John Benjamins.

Henrik Gyllstad and Brent Wolter. 2016. Collocational Processing in Light of the Phraseological Continuum Model: Does Semantic Transparency Matter? *Language Learning*, 66(2):296–323.

Douglas L. Hintzman. 1984. MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, 16(2):96–101.

Douglas L. Hintzman. 1988. Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, 95(4):528.

Douglas Hofstadter and Emmanuel Sander. 2013. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Surfaces and essences: Analogy as the fuel and fire of thinking. Basic Books, New York, NY, US. Pages: xiv, 578.

Douglas R. Hofstadter. 1982. Analogies and Metaphors to Explain Godel's Theorem. *The Two-Year College Mathematics Journal*, 13(2):98–114. Publisher: Mathematical Association of America.

Peter Howarth. 1998. Phraseology and second language proficiency. *Applied Linguistics*, 19(1):24–44.

Randall K. Jamieson, Johnathan E. Avery, Brendan T. Johns, and Michael N. Jones. 2018. An Instance Theory of Semantic Memory. *Computational Brain & Behavior*, 1(2):119–136.

Randall K. Jamieson and D. J.K. Mewhort. 2009. Applying an exemplar model to the artificial-grammar task: Inferring grammaticality from similarity. *Quarterly Journal of Experimental Psychology*, 62(3):550–575.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2024. The Sketch Engine.

Nate Kornell and Robert A. Bjork. 2009. A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, 138(4):449–468.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Ronald W Langacker. 1987. *Foundations of cognitive grammar*. Stanford University Press, Stanford, Calif.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the Ability of Language Models to Interpret Figurative Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Gordon D. Logan. 1997. Automaticity and Reading: Perspectives from the Instance Theory of Automatization. *Reading & Writing Quarterly*, 13(2):123–146.

Gordon D Logan and Joseph L Etherton. 1994. What Is Learned During Automatization? The Role of Attention in Constructing an Instance. *Journal of Experimental Psychology: Learning Memory, and Cognition*, 20(5):1022–1050.

Han Luo. 2019. Idiomaticity and Semantic Extension. In Han Luo, editor, *Particle Verbs in English: A Cognitive Linguistic Perspective*, pages 127–145. Springer, Singapore.

James L McClelland and David E Rumelhart. 1985. On learning the past tense of English verbs. In J. L. McClelland and D. E. and the PDP Research Group Rumelhart, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 2. Bradford Books/MIT Press, Cambridge, MA.

Igor Mel'čuk. 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'.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Rachel A. G. Mueller and Raymond W. Gibbs. 1987. Processing idioms with multiple meanings. *Journal of Psycholinguistic Research*, 16(1):63–81.

J. Nick Reid and Randall K. Jamieson. 2023. True and false recognition in MINERVA 2: Extension to sentences and metaphors. *Journal of Memory and Language*, 129:104397. Publisher: Elsevier Inc.

Tomomi Nishikawa. 2019. 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).

Ira A. Noveck, Nicholas Griffen, and Diana Mazzarella. 2023. Taking stock of an idiom's background assumptions: an alternative relevance theoretic account. *Frontiers in Psychology*, 14.

Timothy J O'Donnell, Noah D Goodman, and Joshua B Tenenbaum. 2009. Fragment Grammars : Exploring Computation and Reuse in Language Fragment Grammars. *Computer Science and Artificial Intelligence Laboratory Technical Report*.

Andrew Pawley and Frances Hodgets Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and Communication*, pages 191–226. ISBN: 9781317869634.

Steven Pinker. 1991. Rules of Language. In *Science*, volume 253, pages 530–535. American Association for the Advancement of Science. Issue: 5019.

Steven Pinker. 2013. *Learnability and Cognition, New Edition: The Acquisition of Argument Structure*. MIT Press, Cambridge, MA.

Steven Pinker and Michael Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Sciences*, 6(11):456–463.

R Core Team. 2024. R: a language and environment for statistical computing. manual, R Foundation for Statistical Computing, Vienna, Austria.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and Measuring the Geometry of BERT. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint*. ArXiv:1908.10084 [cs].

Nick Riches, Carolyn Letts, Hadeel Awad, Rachel Ramsey, and Ewa Dąbrowska. 2022. Collocational knowledge in children: a comparison of English-speaking monolingual children, and children acquiring English as an Additional Language. *Journal of Child Language*, 49(5):1008–1023.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 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.

Diana Sidtis. 2020. Familiar Phrases in Language Competence. In Alexander Haselow and Gunther Kaltenböck, editors, *Grammar and Cognition : Dualistic Models of Language Structure and Language Processing*, pages 38–67. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Royal Skousen. 1990. *Analogical Modeling of Language*. Springer Netherlands, Dordrecht.

Hendrik De Smet and Olga Fischer. 2017. The Role of Analogy in Language Change: Supporting Constructions. In Marianne Hundt, Sandra Mollin, and Simone E. Pfenninger, editors, *The Changing English Language: Psycholinguistic Perspectives*, Studies in English Language, pages 240–268. Cambridge University Press, Cambridge.

Suhad Sonbul, Dina Abdel Salam El-Dakhs, and Rezan Alharbi. 2024. Rendering natural collocations in a translation task: The effect of direction, congruency, semantic transparency, and proficiency. *International Journal of Applied Linguistics*, 34(1):117–133. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijal.12482.

Patrizia Tabossi, Rachele Fanari, and Kinou Wolf. 2008. Processing idiomatic expressions: Effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(2):313–327.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and Methods for the Exploration of Idiomaticity in Pre-Trained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Stroudsburg, PA, USA. Association for Computational Linguistics. ArXiv: 2109.04413.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online. Association for Computational Linguistics.

Mei-Hsing Tsai. 2020. Teaching L2 collocations through concept-based instruction: The effect of L2 proficiency and congruency. *International Journal of Applied Linguistics*, 30(3):553–575. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijal.12311.

Laura Vilkaite and Norbert Schmitt. 2019. Reading collocations in an L2: Do collocation processing benefits extend to non-adjacent collocations? *Applied Linguistics*, 40(2):329–354.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing Pretrained Language Models for Lexical Semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Brent Wolter and Henrik Gyllstad. 2013. Frequency of Input and L2 Collocational Processing: A Comparison of Congruent and Incongruent Collocations. *Studies in Second Language Acquisition*, 35:451–482.

Alison Wray. 2002. *Formulaic Language and the Lexicon*, volume 80. Cambridge University Press, Cambridge. Publication Title: Language ISSN: 1535-0665.

Alison Wray. 2008. *Formulaic language: pushing the boundaries*. Oxford University Press, Oxford. Series Title: Oxford Applied Linguistics.

Satoshi Yamagata, Tatsuya Nakata, and James Rogers. 2023. Effects of distributed practice on the acquisition of verb-noun collocations. *Studies in Second Language Acquisition*, 45(2):291–317.

Junko Yamashita. 2018. Possibility of semantic involvement in the L1-L2 congruency effect in the processing of L2 collocations. *Journal of Second Language Studies*, 1(1):60–78.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2018. Phrase-Level Metaphor Identification Using Distributed Representations of Word Meaning. In *Proceedings of the Workshop on Figurative Language Processing*, pages 81–90, New Orleans, Louisiana. Association for Computational Linguistics.

A Limitations

Our approach relies on contextual embeddings to capture semantic information. However, these embeddings do not always differentiate clearly between compositional and idiomatic readings. Given that our idiomatic stimuli also have a productive reading, the same embedding may be used for both literal and figurative interpretations. Similarly, we cannot ensure that our task is eliciting an idiomatic reading in humans as human listeners disambiguate based on context.

The current dataset was not built from scratch with frequency-matching criteria for idioms. Frequency is a well-established predictor of language processing and an ideal dataset would equate or carefully control the frequency distributions of idioms relative to other word types.

Our study exclusively examined verb–noun (VN) collocations. While these are a critical class of multiword expressions, little is known about other collocational structures (e.g., adjective–noun, phrasal verbs, etc.) which are also prevalent in natural language and may be processed differently. Extending our investigation to these additional types will be important for assessing the generalizability of our findings across the broader spectrum of semi-compositional linguistic units.

MINERVA2 provides a parsimonious framework for modelling memory retrieval, yet it inherently simplifies many aspects of human cognitive processing. The model does not integrate attentional mechanisms or dynamic contextual cues beyond the static embeddings provided, and it does not account for developmental changes in memory and language processing. These simplifications may limit the model's ability to capture the full complexity of human language processing, particularly in cases where retrieval failures (time-outs) interact with other cognitive processes.Our simulations relied on specific hyperparameter settings (e.g., activation threshold K=0.99 and forgetting probability F=0.8) that were chosen based on qualitative assessments. Although results were robust across a range of parameter values, the possibility remains that different parameterizations could yield different patterns.

B Dataset Statistics

Table 1: Descriptive statistics of phrasal frequency by condition

Condition	Mean	SD	N
Compositional	19374.47	30671.53	78
Collocation	21528.21	30971.42	78
Idiom	36784.68	87468.40	78

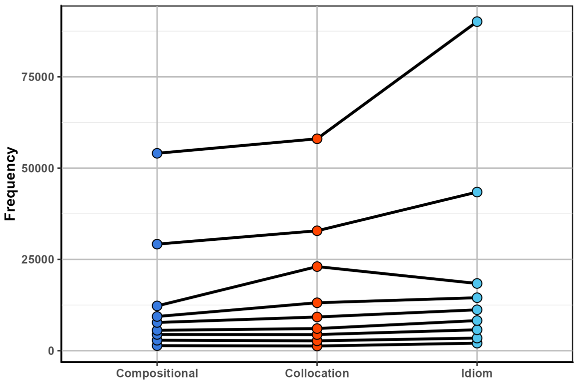

Figure 5: Item frequencies across conditions, by decile

C Human Data

C.1 Reaction times & Accuracy

Table 2: Descriptive statistics of human reaction times (ms) by condition

Condition	Mean	SD	N
Idiom	984.20	365.39	4462
Compositional	995.32	375.76	4423
Collocation	1007.87	370.84	4484

Table 3: Descriptive statistics of human accuracy by condition

Condition	Mean	SD	N
Idiom	0.93	0.25	4785
Compositional	0.92	0.27	4791
Collocation	0.94	0.24	4772

Table 4: Number of incorrect trials by condition

Condition	n
Compositional	400
Collocation	464
Idiom	433
Baseline	1455

Table 5: GLMM results of accuracy in humans

	Dependent variable:
	Accuracy
Compositional	−0.076
	(0.085)
Collocation	0.196^{**}
	(0.089)
Frequency	1.020^{***}
	(0.187)
Constant	3.510^{***}
	(0.148)
N	14,348
Note:	**p<0.05; ***p<0.01

D GLMM Results for Main Simulation

Table 6: GLM results for human AJT reaction times, compared to Tau, a proxy for reaction times, simulated in MINERVA. MINERVA is run with $K = 0.99$, $F = 0.8$

	Dependent variable:	
	RT	Tau
	Human	MINERVA
Compositional	4.690**	0.624***
	(2.240)	(0.004)
Collocation	13.800***	−0.410***
	(1.760)	(0.004)
Frequency	−18.500***	−0.541***
	(1.640)	(0.004)
Constant	1,047.0***	5.900***
	(2.140)	(0.004)
N	13,369	43,708
Note:	*p<0.1; **p<0.05; ***p<0.01	

E Ablations

E.1 Semantics-only

In the semantics-only ablation wherein the model was loaded with all instances being equally frequent, we visually observed that idioms were retrieved slower than compositional items when timeouts were not included. The proportions of timeouts are similar to those of the main experiment, with collocations timing out much more frequently. In the semantics-only condition, however, compositional items time out less frequently than idioms. This result is not surprising given the marginal significance of the difference between human RTs for idioms and compositional items. However, investigating the cause of this discrepancy is an interesting avenue for future work.

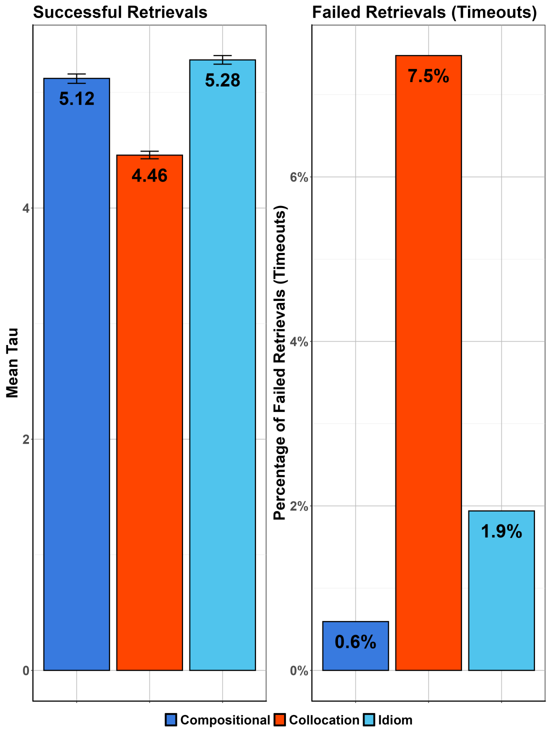

Figure 6: **Left:** mean Tau (τ) by condition for successful retrievals in Ablation 1, wherein frequency information was eliminated. The y-axis represents mean Tau, the model's output which acts as a proxy for reaction times. Error bars indicate bootstrapped confidence intervals. **Right:** percentage of failed retrievals, i.e., timeouts, per condition in Ablation 1.

E.2 Frequency-only

In the frequency-only ablation, the model was loaded with embeddings comprised entirely of Gaussian noise. However, each noise-item was sampled according to correct frequency information. For successful retrievals, we visually observed that idioms and collocations were retrieved equally quickly, whereas compositional items were retrieved slower. Given that frequency drives MINERVA's retrieval mechanism, the pattern of timeouts for Ablation 2 are not surprising. Idioms which are the most frequent subset time out the least, followed by collocations which are slightly more frequent than compositional units which, in turn, time out the most.

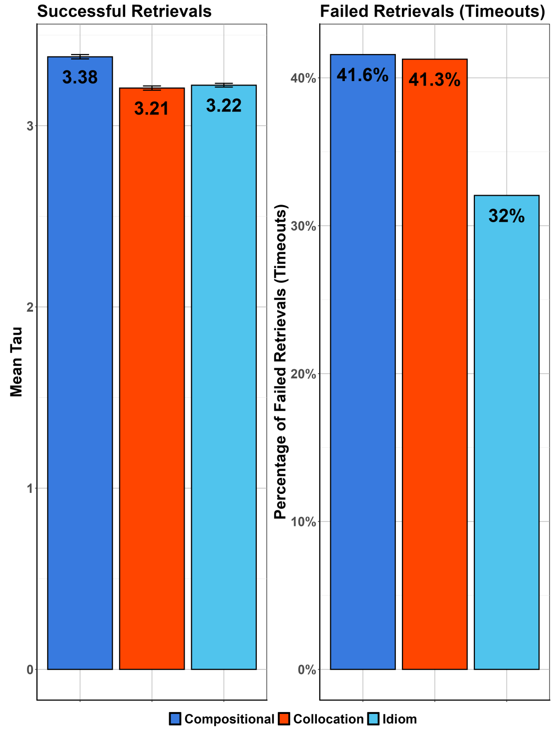

Figure 7: **Left:** mean Tau (τ) by condition for successful retrievals in Ablation 2, wherein semantic information was eliminated while leaving the correct item frequency distribution. The y-axis represents mean Tau, the model's output which acts as a proxy for reaction times. Error bars indicate bootstrapped confidence intervals. **Right:** percentage of failed retrievals, i.e., timeouts, per condition in Ablation 2.

E.3 Averaging vs Concatenating sBERT Embeddings

In this ablation, we investigate the impact which concatenating verb and noun embeddings has on our modelling results. Instead of concatenating verb and noun embeddings, we perform mean-pooling across them, the same as we do for subword tokens. As shown in Figure 8, the trends exhibited by the model in the $K = 0.99, F = 0.8$ hyperparameter configuration are largely the same as those reported in the main text.

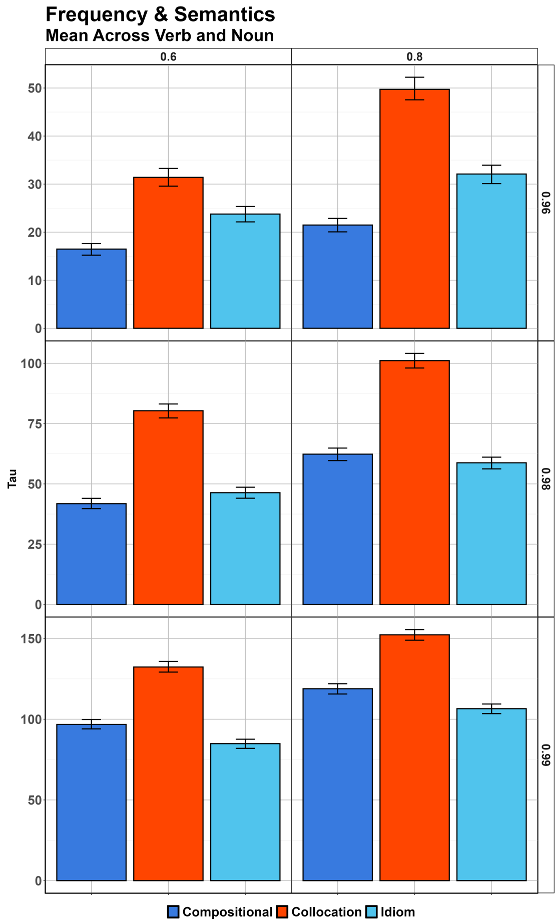

Figure 8: Reduced hyperparameter sweep showing the effects of mean-pooling the verb and noun embeddings before loading them into MINERVA, instead of concatenating them. Note that the hyperparemeter combination reported in the main text is $K = 0.99, F = 0.8$.

F Hyperparameter Sweeps for Simulation Experiments

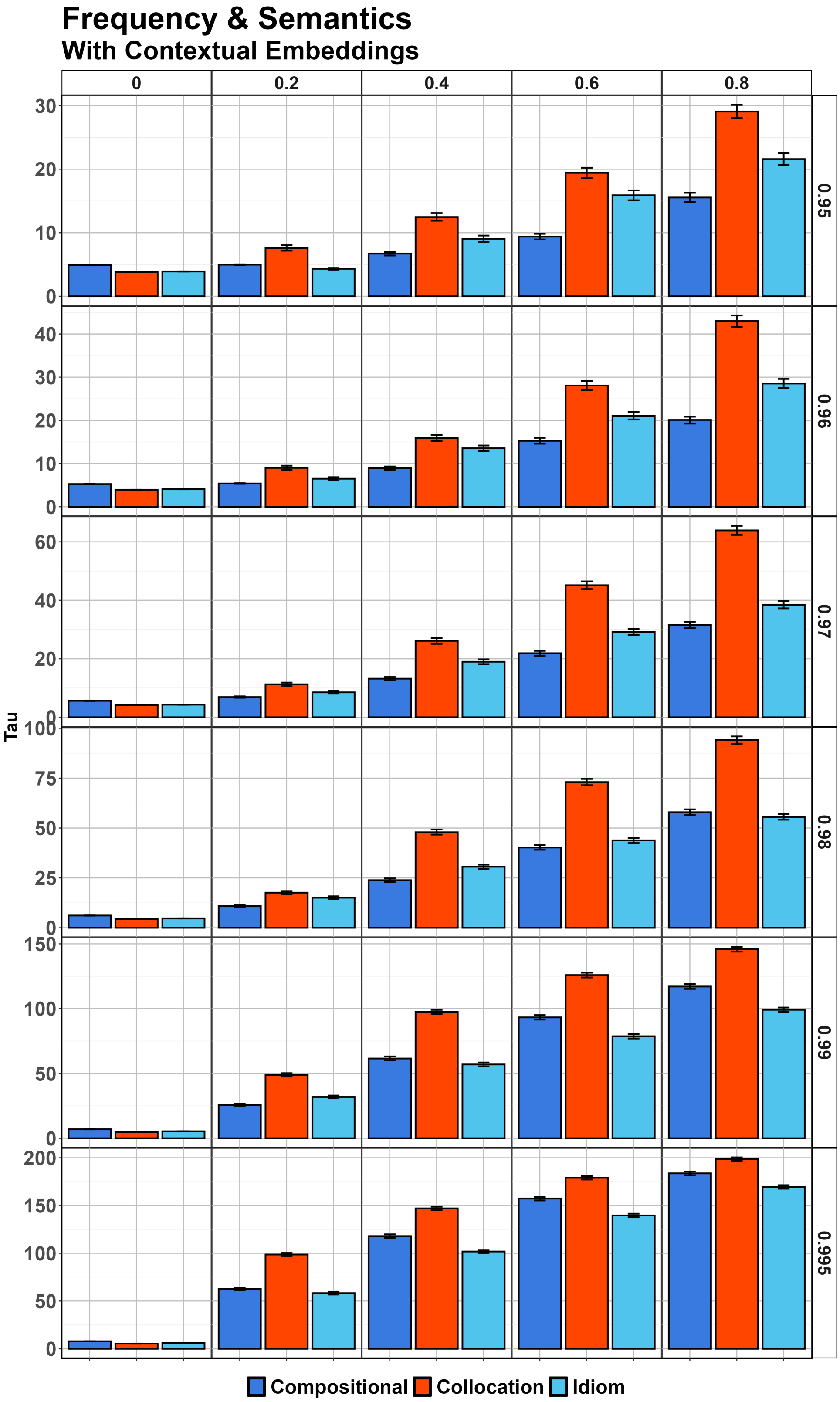

Figure 9: Results of the hyperparameter sweep for all values of activation threshold K and forgetting probability F for our main experiment. Error bars indicate bootstrapped confidence intervals. Note the difference in scales on the y-axis.

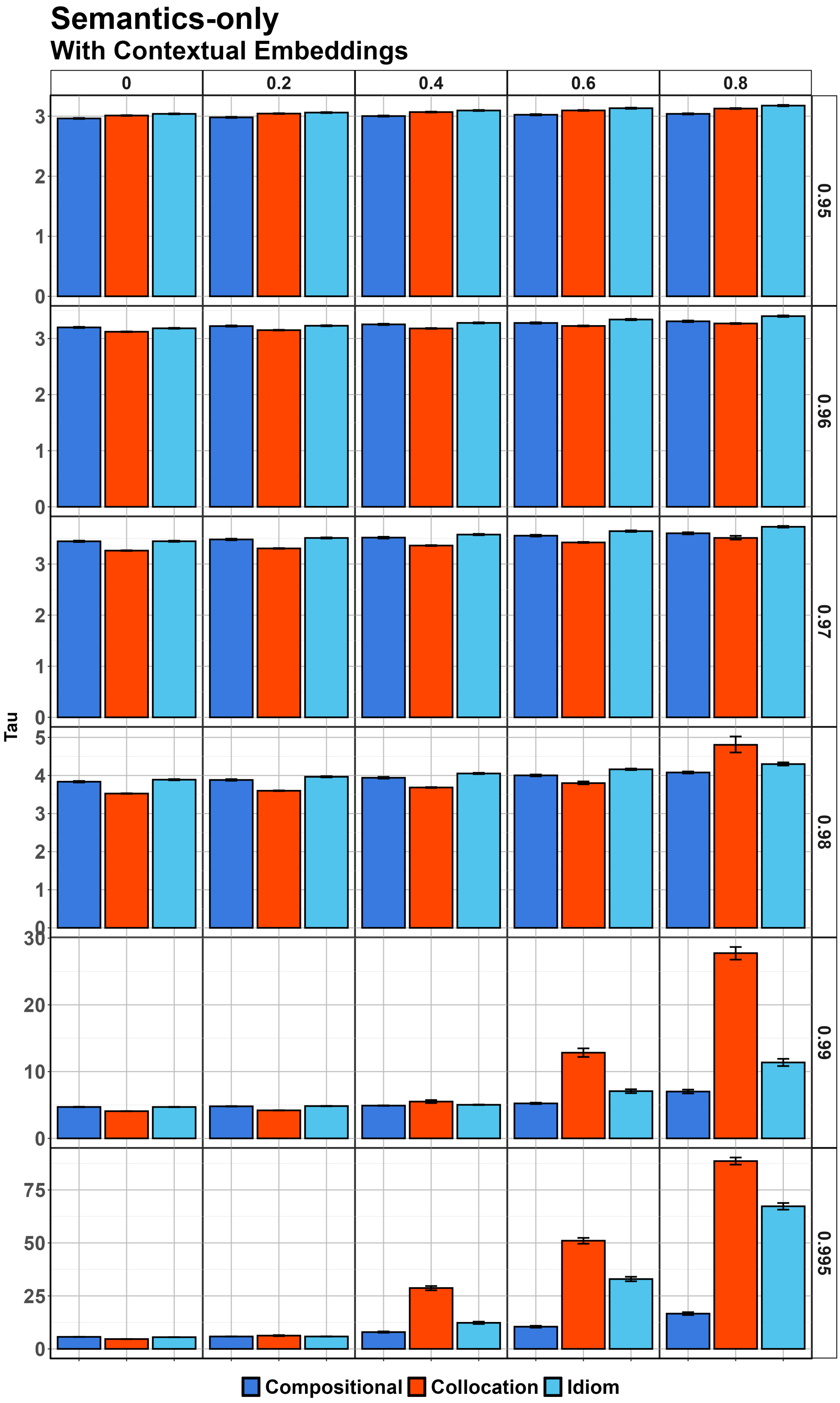

Figure 10: Results of the hyperparameter sweep for all values of activation threshold K and forgetting probability F for Simulation 2: Semantics-only wherein the matrix was loaded with all items having equal frequency. Error bars indicate bootstrapped confidence intervals. Note the difference in scales on the y-axis.

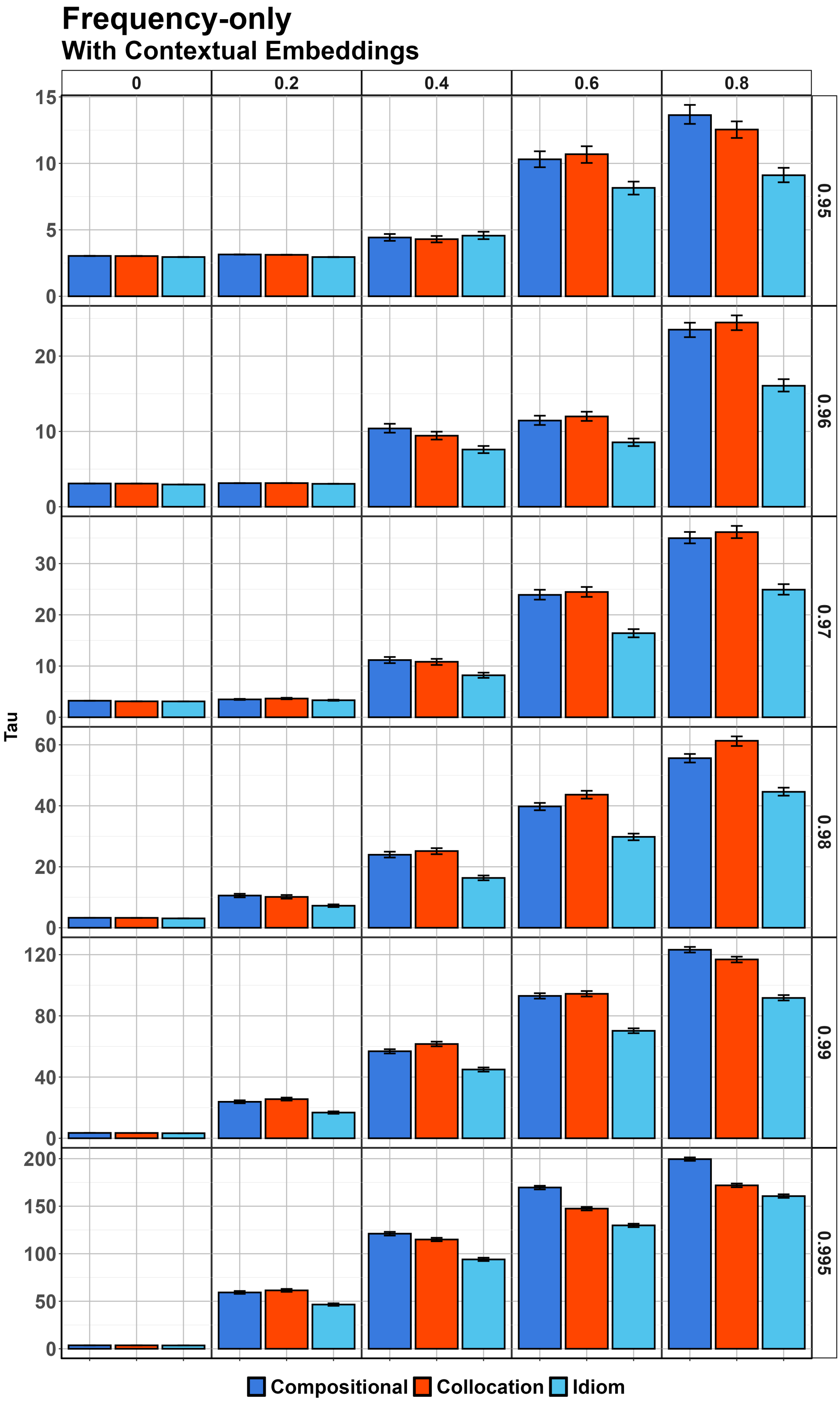

Figure 11: Results of the hyperparameter sweep for all values of activation threshold K and forgetting probability F for Simulation 2: Semantics-only wherein the matrix was loaded with noised embeddings but with the correct frequency. Error bars indicate bootstrapped confidence intervals. Note the difference in scales on the y-axis.

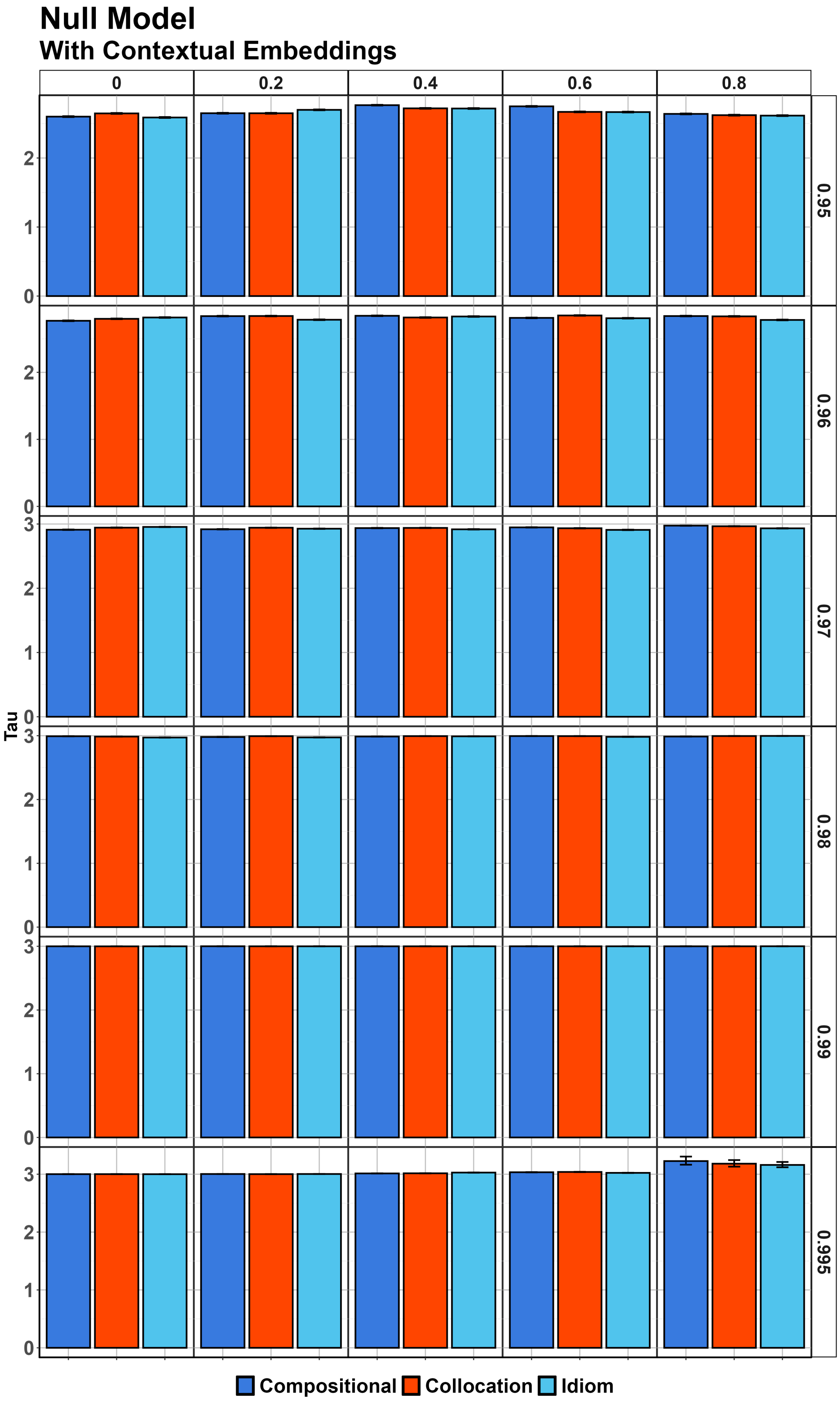

Figure 12: Results of the hyperparameter sweep for all values of activation threshold K and forgetting probability F for the Null Model wherein all the items in the matrix were loaded with noised embeddings and equal frequency. Error bars indicate bootstrapped confidence intervals. Note the difference in scales on the y-axis.