Cognitive Signatures of Multi-Word Expressions: Reading-Time and Surprisal

Diego Alves and Sergei Bagdasarov and Elke Teich
Saarland University
Saarbrücken, Germany
diego.alves@uni-saarland.de, sergeiba@lst.uni-saarland.de,
elke.teich@uni-saarland.de

Abstract

This study investigates whether eye-tracking measures predict if a word is the final token of a multi-word expression (MWE), focusing on two understudied MWE types: fixed expressions (e.g., *due to*) and phrasal verbs (e.g., *turn out*). Using mixed-effects logistic regression, we compared tokens in MWE contexts with the same tokens in non-MWE contexts. Results reveal a clear difference in processing. For fixed expressions, reading-time measures significantly predict MWEhood. In contrast, phrasal verbs show no consistent predictive effects. Additionally, we compared the reading-time models to models that included GPT-2 surprisal as a predictor. While surprisal does predict MWEhood, it fails to capture the distinction between types. These findings highlight the need to consider MWE typology in models of formulaic language processing.

1 Introduction

Across languages, certain word combinations, known as multi-word expressions (MWEs), are conventional patterns associated with specific meanings or connotations. MWEs take diverse forms, ranging from structurally fixed idioms with figurative meanings (e.g., *break the ice*), to compounds (e.g., *sea water*), which vary in compositionality, and phrasal verbs (e.g., *carry out*), which can be either compositional or idiomatic and are often lexically productive (Avgustinova and Iomdin, 2019).

MWEs are ubiquitous because they enhance language efficiency through predictable transitions between words. Highly conventionalised MWEs can be retrieved holistically from the lexicon rather than incrementally processed, providing a processing advantage over novel sequences (Siyanova-Chanturia et al., 2017). From a communicative perspective, MWEs reduce cognitive load for language users, serving as devices that streamline processing and facilitate comprehension (Conklin and Schmitt, 2012). Many studies have demonstrated the processing advantages of MWEs using eye-tracking and event-related potentials (ERP). These studies show that MWEs are generally read and processed more efficiently than novel sequences, with facilitation influenced by factors such as frequency, predictability, familiarity, and type-specific properties (e.g., Siyanova (2010); Carrol and Conklin (2020); Kessler et al. (2021)).

As shown by Carrol and Conklin (2020), different types of MWEs exhibit different cognitive processing patterns. In the present study, we focus on two types: fixed expressions (e.g., *due to*, *out of*) and phrasal verbs (e.g., *turn out*, *rush in*). These small lexical units have been understudied in research on MWE processing. Our analysis focuses on the final token of each sequence because MWEs are characterized by highly predictable transitions between constituent tokens. This predictability advantage is expected to manifest most clearly at the final token, which is processed more rapidly when it completes an MWE than when the same token appears in a non-MWE context. To do this, we compare tokens appearing in MWEs with the same tokens when they occur in non-MWE contexts. Additionally, we compare the results obtained using reading-time predictors with models based on surprisal estimates from a large language model, to examine whether the surprisal behaviour of the final token also differs according to MWE type.

2 Related Work

Eye-tracking studies have long shown that gaze patterns are sensitive to linguistic and contextual factors, including lexical frequency, verb complexity, and ambiguity (Rayner (1975); Rayner and Duffy (1986); Rayner et al. (2012)), providing a foundation for understanding real-time processing of multi-word expressions (MWEs) and formulaic language.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 48–53

Frequency strongly influences MWE processing. Siyanova (2010) found that high-frequency MWEs are processed more efficiently by native speakers, whereas non-native speakers benefit mainly from very high-frequency items. Conklin and Schmitt (2012) review evidence that MWEs are generally read faster than novel sequences, with speed modulated by frequency, predictability, and transparency. Pellicer-Sánchez and Perez (2024) similarly highlight frequency, familiarity, predictability, and decomposability as robust predictors of processing ease, especially for L1 readers.

Different MWE types exhibit distinct patterns. Carrol and Conklin (2020) reported a general processing advantage for idioms, binomials, and collocations, with type-specific effects: idioms were sensitive to frequency, familiarity, and decomposability; binomials to predictability and semantic association; collocations to mutual information. Kessler et al. (2021) extended this to spoken idioms, showing listeners fixate predicted completions and early semantic associates, with ERP data indicating facilitated processing for correct completions.

Late gaze measures, including regressions and re-reading, reliably distinguish MWEs from novel sequences. Rohanian et al. (2017) showed that combining gaze features with part-of-speech and frequency enables computational models to predict MWEs, consistent with findings that early gaze measures are less informative (Siyanova-Chanturia, 2013).

The predictability of a final MWE element can also be formalized with surprisal, the negative log probability of an event (Shannon (1948)), with higher surprisal leading to longer fixations. Onnis and Huettig (2021) applied this to MWEs, showing that frequent and predictable sequences are easier to integrate, whether stored as chunks or composed. Moreover, Alves et al. (2025) show that the negative surprisal slope over token sequences is a strong predictor of MWEhood.

In this study, we focus on two under-studied MWE types, using regression models to examine whether reading-time measures predict MWEhood. We also compare these effects with surprisal estimates from a large language model, which have been shown to predict reading times (Wilcox et al., 2023).

3 Methodology

3.1 Data

We used two eye-tracking corpora: UCL (Frank et al., 2013) and Provo (Luke and Christianson, 2018).

The UCL dataset includes self-paced reading times and eye-tracking data from 361 English sentences drawn from three novels. The participants were native speakers and first-year psychology students (104 self-paced readers and 42 eye-tracking participants; mostly native speakers). Reading-time measures include word-by-word response times, first-fixation, first-pass gaze duration, and total fixation.

The Provo Corpus contains eye-tracking data from 84 native English-speaking adults reading 55 short passages (134 sentences, 2,745 words) from news, fiction, and popular-science texts. Measures include fixation durations, number of fixations, skipping, regressions, and cloze-based predictability norms. Unlike isolated sentence corpora, Provo captures more naturalistic, continuous reading, making it particularly suitable for studies of predictive processing.

Sentences from both corpora were automatically annotated using the Universal Dependencies framework with Stanza (Qi et al., 2020) and the combined English model. Fixed expressions and phrasal verbs were identified from tokens labeled as `fixed` and `compound:prt`, respectively, and assigned a value of 1 (MWE), while tokens with the same surface form but different labels were assigned 0 (non-MWE).

3.2 Reading-time Measures and Surprisal

In this study, we focus on three widely used reading-time measures (Rayner, 1998). First fixation duration refers to the duration of the initial fixation on a word during first-pass reading. Gaze duration is the sum of all first-pass fixations on a word, while total fixation duration represents the total time spent fixating on a word, including regressions.

First fixation duration reflects early lexical access, gaze duration captures lexical and syntactic processing during initial reading, and total fixation duration indexes later comprehension stages such as reanalysis and integration difficulties (Rayner, 1998).

For the comparison of reading-time measures with surprisal, we estimated the surprisal of each

word using the smallest GPT-2 model[1] (Radford et al., 2019). We use GPT-2 because prior work has shown that surprisal estimates from larger transformer-based language models often provide a poorer fit to human reading times than smaller models, likely because increased capacity leads to representations that diverge from human incremental processing (Oh and Schuler, 2023). Surprisal values were extracted using the `surprisal`[2] Python library. Word-level surprisal was computed by summing the surprisal values of the constituent subword tokens.

3.3 Regression Models

We performed logistic mixed-effects regression analyses in R, using lme4 library (Bates et al., 2015), to examine whether reading-time (RT) measures, tested one at a time, predict the likelihood that a word is part of a multi-word expression (MWEhood). Our analysis focusses specifically on the final tokens that occur either in fixed expressions or in phrasal verbs, comparing their behaviour when they appear in MWE contexts versus non-MWE contexts. For each token, we fitted a logistic mixed-effects model with predictors including the current word's RT, word length, their interaction, spillover RTs and word lengths of the two preceding words, and random intercepts for participants (Equation 1).

$$\begin{aligned} \text{MWEhood}_{ij} \sim\ & \text{RT}_{ij} \times \text{WordLength}_{ij} \\ + \text{RT}_{i,j-1} + & \text{RT}_{i,j-2} + \text{WordLength}_{i,j-1} \\ + \text{WordLength}_{i,j-2} + & (1 \mid \text{Subject}_i) \end{aligned} \quad (1)$$

The same type of regression was conducted in a second step, replacing the reading-time measure with surprisal estimates derived from a GPT-2 language model for the occurrences of the MWEs in the Brown corpus (Francis, 1965).

Finally, to complement our analysis, we calculated the pointwise mutual information (PMI) for each fixed expression and phrasal verb identified in the corpora[3]. The idea is to test whether PMI values can account for the differences observed between the reading-time models and the surprisal models.

[1] https://huggingface.co/openai-community/gpt2

[2] https://pypi.org/project/surprisal/

[3] All MWEs extracted from the corpora for this study were bigrams.

4 Results

4.1 MWE Identification

From the parsed sentences of both corpora, we extracted several MWEs. In the case of fixed expressions, six were identified in the UCL corpus; however, for four of these, the final token did not appear in a non-MWE context (e.g., *at least*, *in order*). Consequently, only *instead of* and *out of* were considered, with the preposition *of* as the analyzed token. In the Provo corpus, nine fixed expressions were identified, but only *due to* and *out of* had final tokens that also occurred in non-MWE contexts.

Regarding phrasal verbs, seventy were extracted from the UCL corpus. The ones retained for our analysis included three with the particle *on* (*knock on*, *go on*, and *caught on*), six with *in* (e.g., *fill in*, *step in*), and seventeen with *out* (e.g., *let out*, *knock out*). In the Provo corpus, fifteen phrasal verbs were extracted, of which one had the particle *in* (*rush in*) and five had *out* (*turned out*, *help out*, *dig out*, *built out*, *looked out*).

4.2 Reading Time as MWEhood Predictor

Table 1 shows the significance and AIC values of first fixation, gaze, and total fixation for the final tokens of fixed expressions and phrasal verbs in the UCL and Provo corpora, in predicting whether a token is part of an MWE. Stars indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, ns = not significant), and numbers in the adjacent column represent the corresponding AIC values of the regression models.

It is can be observed that reading-time measures are statistically significant predictors of MWEhood for fixed expressions in both corpora. The coefficients indicate, consistent with previous work, that tokens are read faster when they form part of an MWE. However, in contrast to earlier findings (cf. Siyanova-Chanturia (2013)), we also find significant and consistent effects for first-fixation duration, suggesting that MWE processing advantages can emerge at earlier stages of lexical access than previously reported.

On the other hand, for phrasal verbs, we observed no significant effects (with the exception of the particle *out* in the Provo corpus). Although Kissane et al. (2024) reported that phrasal-verb particles tend to be read more rapidly than verb–preposition bundles, our results align with the findings of Yaneva et al. (2017), who showed that

Corpus	Token	MWE Type	First Fix.	AIC	Gaze	AIC	Total Fix.	AIC
Provo	to	Fixed Expression	**	294	*	294	***	1959
	of	Fixed Expression	*	107	ns	113	ns	491
UCL	of	Fixed Expression	**	522	**	525	ns	531
Provo	in	Phrasal Verb	ns	57	ns	57	ns	508
	out	Phrasal Verb	ns	230	*	224	ns	1045
UCL	on	Phrasal Verb	ns	513	ns	493	ns	504
	in	Phrasal Verb	ns	655	ns	658	ns	661
	out	Phrasal Verb	ns	806	ns	805	ns	805

Table 1: Significance and AIC values of reading-time measures for fixed expressions and phrasal verbs across corpora. Stars indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, ns = not significant).

the final word in verb–particle combinations does not differ in processing between MWEs and control phrases for either native or non-native speakers. This is likely because readers often extract sufficient information about particles before directly fixating on them, resulting in high skipping rates.

In terms of AIC, models trained on the Provo data show better predictive accuracy compared to UCL models for first fixation and gaze duration. However, when total fixation duration is used as a predictor, the Provo-based models are less predictive of MWEhood, suggesting that the status of the token as part of an MWE has a stronger influence during earlier stages of cognitive processing.

When replacing the reading-time measures in equation 1 with surprisal estimates from GPT-2 (also for the previous tokens), we observe a significant effect (*** $p < 0.001$) for all tokens except *in*. This suggests that surprisal does not distinguish between MWE types in a way that reflects the cognitive patterns observed in the eye-tracking data. Additionally, the AIC values for models using surprisal as a predictor are relatively high, over 3,000 for fixed expressions and over 1,200 for phrasal verbs, indicating lower predictive accuracy compared to models using reading-time measures as predictors.

The differences observed in the reading-time models may be due to structural differences between fixed expressions and phrasal verbs. While the former function as grammatical units, the latter behave as lexical items, which entails differences in their overall cognitive processing. Moreover, reading time reflects multiple stages of processing such as lexical access, syntactic integration, and comprehension, whereas surprisal is more limited, capturing only how predictable a token is given the preceding context. Table 2 presents the mean PMI values for the fixed expressions and phrasal-verb particles included in our analysis.

Token	MWE Type	Mean PMI
to	fixed	2.53
of	fixed	3.81
on	PV	3.42
in	PV	1.65
out	PV	3.94

Table 2: Mean PMI values for fixed expressions and phrasal verbs.

Analysing the PMI values of the fixed expressions and phrasal verbs shows that fixed expressions and verb–particle combinations with *out* and *on* generally show the highest PMI scores, although some variability is evident (e.g., the low PMI of *not to*, 0.81). In contrast, phrasal verbs with *in* show the lowest PMI values, which may help explain the lack of significant effects when using surprisal as a predictor. Overall, these results indicate that PMI alone cannot account for the differences observed in the cognitive processing of fixed expressions and phrasal verbs.

5 Conclusion and Future Work

This study examined whether eye-tracking measures predict whether a word is the final token of a multi-word expression (MWE), focusing on fixed expressions (e.g., *due to*) and phrasal verbs (e.g., *go out*). Logistic mixed-effects regression analyses were used to compare reading-time measures for tokens appearing in MWEs versus the same tokens in non-MWE contexts.

The results reveal a clear processing distinction between these MWE types. For fixed expressions, reading times, including early measures such as first-fixation duration, significantly predicted

MWEhood. In contrast, phrasal verbs showed no consistent reading-time differences. Additionally, while surprisal estimates from GPT-2 generally predicted MWEhood, they did not capture this type-specific distinction, and PMI values also failed to account for the observed processing differences.

These findings highlight that MWE type matters: fixed expressions, which function as grammatical units, and phrasal verbs, which behave as lexical items, engage distinct cognitive mechanisms despite both being formulaic.

Although the present study focuses specific classes of English MWEs, the proposed approach is not inherently language-specific. It could be extended to other languages by leveraging tokens labelled as fixed in the Universal Dependencies (UD) framework, which capture a wide range of multiword expressions cross-linguistically. Moreover, while phrasal verbs are characteristic of English, the same methodology could be applied to other MWE types, such as light verb constructions, which are prominent in many languages.

Future work should extend this investigation to additional eye-tracking corpora and other types of MWEs not included in the present study.

Limitations

The findings of this study should be considered in light of its limitations. First, the analysis relies on data from only two eye-tracking corpora (UCL and Provo), which constrains the number and variety of multi-word expressions (MWEs) available for examination. Consequently, many fixed expressions and phrasal verbs were excluded because their final tokens did not appear in comparable non-MWE contexts, reducing statistical power and generalisability. Second, the findings are specific to two MWE types (fixed expressions and phrasal verbs); other important categories were not tested with the same regression approach. Consequently, idiomaticity, transparency, and semantic compositionality are not examined in this paper. Ideally, future eye-tracking experiments would include compounds in both compositional and non-compositional contexts, enabling direct comparison of reading-time measures.

Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Diego Alves, Sergei Bagdasarov, and Elke Teich. 2025. Surprisal dynamics for the detection of multi-word expressions in english. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1185–1194.

Tania Avgustinova and Leonid Iomdin. 2019. Towards a typology of microsyntactic constructions. In *International Conference on Computational and Corpus-Based Phraseology*, pages 15–30. Springer.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Gabor Grothendieck, Peter Green, and Maintainer Ben Bolker. 2015. Package 'lme4'. *convergence*, 12(1):2.

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.

Kathy Conklin and Norbert Schmitt. 2012. The processing of formulaic language. *Annual review of applied linguistics*, 32:45–61.

W Nelson Francis. 1965. A standard corpus of edited present-day american english. *College English*, 26(4):267–273.

Stefan L Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior research methods*, 45(4):1182–1190.

Ruth Kessler, Andrea Weber, and Claudia K Friedrich. 2021. Activation of literal word meanings in idioms: Evidence from eye-tracking and erp experiments. *Language and Speech*, 64(3):594–624.

Hassane Kissane, Konstantin Tziridis, Achim Schilling, Patrick Krauss, and Thomas Herbst. 2024. Cognitive dynamics of verb-particle constructions: An eye-tracking study. *bioRxiv*, pages 2024–12.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50(2):826–833.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Luca Onnis and Falk Huettig. 2021. Can prediction and retrodiction explain whether frequent multi-word phrases are accessed 'precompiled' from memory or compositionally constructed on the fly? *Brain Research*, 1772:147674.

Ana Pellicer-Sánchez and Maribel Montero Perez. 2024. Eye-tracking in vocabulary research: Introduction to the special issue. *Research methods in applied linguistics*, 3(1):100095.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Keith Rayner. 1975. The perceptual span and peripheral cues in reading. *Cognitive psychology*, 7(1):65–81.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Keith Rayner and Susan A Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. *Psychology of reading*. Psychology Press.

Omid Rohanian, Shiva Taslimipoor, Victoria Yaneva, and Le An Ha. 2017. Using gaze data to predict multiword expressions.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Anna Siyanova. 2010. On-line processing of multi-word sequences in a first and second language: Evidence from eye-tracking and erp. Technical report, University of Nottingham.

Anna Siyanova-Chanturia. 2013. Eye-tracking and erps in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon*, 8(2):245–268.

Anna Siyanova-Chanturia, Kathy Conklin, Sendy Caffarra, Edith Kaan, and Walter JB van Heuven. 2017. Representation and processing of multi-word expressions in the brain. *Brain and language*, 175:111–122.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Victoria Yaneva, Shiva Taslimipoor, Omid Rohanian, and Le An Ha. 2017. Cognitive processing of multiword expressions in native and non-native speakers of english: Evidence from gaze data. In *International conference on computational and corpus-based phraseology*, pages 363–379. Springer.