Surprisal Dynamics for the Detection of Multi-Word Expressions in English

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

Abstract

This work examines the potential of surprisal slope as a feature for identifying multi-word expressions (MWEs) in English, leveraging token-level surprisal estimates from the GPT-2 language model. Evaluations on the DiMSUM and SemEval-2022 datasets reveal that surprisal slope provides moderate yet meaningful discriminative power with a trade-off between specificity and coverage: while high recall indicates that surprisal slope captures many true MWEs, the slightly lower precision reflects false positives, particularly for non-MWEs that follow formulaic patterns (e.g., adjective-noun or verb-pronoun structures). The method performs particularly well for conventionalized expressions, such as idiomatic bigrams in the SemEval-2022 corpus. Both idiomatic and literal usages of these bigrams exhibit negative slopes, with idiomatic instances generally showing a more pronounced decrease. Overall, surprisal slope offers a cognitively motivated and interpretable signal that complements existing MWE identification methods, particularly for conventionalized expressions.

1 Introduction

Regularity in language concerns not only structural aspects such as syntax and morphology, but also the patterned combination of words. Across languages, certain word combinations, referred to as multi-word expressions (MWEs), are recognized as conventional patterns associated with specific meanings or connotations. MWEs encompass a wide range of forms, from idioms that are structurally fixed and carry figurative meanings (e.g., *break the ice*), to compounds (e.g., *sea water*), which vary in their degree of compositionality, and phrasal verbs (e.g., *put up with*), which also range from compositional to idiomatic in meaning and are often lexically productive (cf. Avgustinova and Iomdin (2019)).

MWEs are widespread because they enhance language efficiency through highly predictable transitions between words. When highly conventionalized, MWEs can be retrieved holistically from the lexicon rather than processed incrementally, conferring a processing advantage over other word sequences (Siyanova-Chanturia et al., 2017). From a communicative perspective, MWEs thus provide a clear processing benefit for language users, serving as devices that ease the informational load within the signal (Conklin and Schmitt, 2012).

The identification and extraction of multi-word expressions (MWEs) constitute a significant area of research within natural language processing (NLP), focusing on the development of resources for use with machine learning, deep learning algorithms, and large language models (Ramisch et al., 2023). However, there is a lack of analysis from the perspective of interpretable NLP, particularly in identifying the specific features that characterize these linguistic units.

Rather than proposing a new extraction method for MWEs, this study aims to characterize them from a cognitive and information-theoretic standpoint. We focus on the predictability property of MWEs to examine whether variation in surprisal (measured through slope) reflects their potential to reduce cognitive load. Surprisal, as defined by Shannon (1948), is the negative logarithm of the probability of a word given its context, representing how unexpected or informative a word is in a sequence. Accordingly, if a sequence of tokens constitutes a MWE, we expect its surprisal slope to be negative, indicating increasing predictability of subsequent tokens and facilitating cognitive processing.

To test this hypothesis in the context of English, we use two publicly available MWE datasets: the DiMSUM corpus (Schneider et al., 2016) and the SemEval-2022 corpus for multilingual idiomaticity detection and sentence embedding (Tayyar Mad-

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

abushi et al., 2022). In-context surprisal was estimated using the GPT-2 small model (Radford et al., 2019), 124M parameters. For DiMSUM, variation in surprisal within MWEs was compared to that of randomly selected n-grams from the same dataset. For SemEval-2022, surprisal variation in idiomatic MWE instances was compared to their non-idiomatic counterparts. This approach represents a novel method for characterizing MWEs through surprisal dynamics.

The remainder of this paper is organized as follows. Section 2 reviews related work on the cognitive processing and predictability of MWE tokens, as well as studies on MWE characterization. Sections 3 and 4 present our methodology and results, respectively. In Section 5, we discuss the findings, followed by a summary and directions for future research in Section 6.

2 Related Work

A variety of studies have suggested that MWEs are easier to process than non-formulaic sequences of words. MWE frequency was found to be an important factor contributing to this processing advantage, with high-frequency MWEs being processed faster than lower-frequency control items. This effect was observed using different methods such as recognition times (Arnon and Snider, 2010), reaction times obtained in self-paced reading experiments (Tremblay et al., 2011; Conklin and Schmitt, 2008), eye tracking (Siyanova-Chanturia et al., 2011) as well as EEG recordings (Tremblay and Baayen, 2010). Moreover, MWEs are easier to retrieve from memory after processing, suggesting better pattern storage in memory (Tremblay et al., 2011; Tremblay and Baayen, 2010).

Predictability of single MWE components was also argued to play an important role in processing. It has beed suggested that the smooth transitions from one MWE component to the other help to activate pre-fabricated mental templates, resulting in a reduced cognitive load. For instance, Siyanova-Chanturia et al. (2017) found an increased P300 and a reduced N400 effect at the last word in those binomial expressions with a more predictable second conjunct, suggesting a more efficient processing. The cognitive advantage persists even in modified phrases that still allow the prediction of the final word (Chantavarin et al., 2022).

While predictability is indeed often a function of frequency, this is not the case for relatively low-frequency MWEs like proper noun MWEs or rare collocations. To account for different properties of MWEs, Gries (2022) and Youssef (2024) proposed a multi-dimensional and highly information-theoretical approach that can identify and describe MWEs of a wide range of types.

Another common measure for MWE detection is mutual information (Church and Hanks, 1990), a statistical metric that quantifies the strength of association between two words based on how frequently they co-occur relative to chance. This measure, or its variations, forms the basis of several statistical approaches to MWE or formulaic language extraction (e.g., Zhang et al. (2009); Simpson-Vlach and Ellis (2010)).

A more straightforward way to address MWE predictability is surprisal – an information-theoretic measure that quantifies the (un)expectedness of a word in a given context.[1] Onnis and Huettig (2021) used the surprisal of the last word in a pre-selected list of four-word expressions to distinguish between MWEs and non-formulaic sequences. They found that surprisal is a better predictor of formulaicity than frequency. Based on these results, MWE internal predictability operationalized with surprisal should be an effective instrument to identify MWEs in naturally occurring texts.

3 Methodology

3.1 Data

As previously mentioned, to analyze the impact of surprisal variation on the identification of MWEs, we used two datasets: the DiMSUM corpus and the SemEval 2022 Shared Task dataset on multilingual idiomaticity detection.

The DiMSUM corpus (Schneider et al., 2016), developed for the SemEval 2016 Shared Task, comprises three datasets: the STREUSLE 2.1 corpus of web reviews (Schneider and Smith, 2015), along with the Ritter (Ritter et al., 2011) and Lowlands Twitter datasets (Johannsen et al., 2014). It is annotated for the majority of major multi-word expression (MWE) categories, including nominal (e.g., *business book* and *Lady Gaga*), verbal (e.g., *get lost* and *check out*), adverbial (e.g., *so far* and *at all*), and functional MWEs (e.g., *due to*), although it does not provide specific category labels. The annotation follows the BIO tagging scheme. The training set contains 4,799 sentences, while the

[1] For a more detailed account on surprisal, see Section 3.

test set includes 1,000 sentences. Overall, the corpus features 5,044 MWE occurrences, comprising 4,541 continuous and 503 discontinuous expressions. Among these, there are 3,304 unique continuous MWEs and 416 unique discontinuous MWEs. We used the entire dataset (train and test) and discontinuous MWEs were not treated differently, as we assume that the predictability effect persists even when words intervene between the components.

The SemEval 2022 Shared Task dataset on multilingual idiomaticity detection (Tayyar Madabushi et al., 2022) focuses on the binary classification of MWEs based on their usage context. The English training set consists of 3,327 sentences containing 163 distinct MWEs (one MWE per sentence), all of which are fixed two-word expressions containing at least one noun (continuous bigrams). Each instance is annotated with a binary label: 0 for idiomatic usage and 1 for non-idiomatic usage. In this dataset, MWEs occur in both senses; for example, heavy cross appears 12 times (9 idiomatic, 3 literal). For this analysis, 2,755 sentences were retained due to tokenization issues arising from the LLM's subword segmentation when converting back to words, likely caused by encoding. Since the dataset remained sufficiently large, we proceeded with the sentences that did not present any issues.

3.2 Surprisal Estimation

To estimate token-level surprisal in the two selected datasets (DiMSUM and SemEval-2022), we employed the GPT-2 language model using the surprisal[2] Python library. In this context, surprisal refers to the information-theoretic measure of how unexpected a word is given its preceding context, as originally proposed by Shannon (1948). Formally, the surprisal of a word w_i ($S(w_i)$) given a context $w_1, w_2, \ldots, w_{i-1}$ is defined as:

$$\mathrm{S}(w_i) = -\log_2 P(w_i \mid w_1, w_2, \ldots, w_{i-1}) \quad (1)$$

Higher surprisal values indicate lower predictability and typically correlate with increased cognitive processing effort (Demberg and Keller, 2009). For implementation, we used the AutoHuggingFaceModel class from the surprisal library to load the pre-trained GPT-2 model[3]. Each sentence was processed to compute token-level surprisal scores, which were subsequently written to an output file for further analysis. Surprisal was calculated at the subword token level (as defined by the GPT-2 tokenizer), and word-level surprisal values were obtained by summing the surprisal of all subwords composing each word.

3.3 Surprisal Slope

After estimating surprisal values for each token in the selected datasets, we quantified the variation of surprisal within multi-word expressions (MWEs) by calculating the slope of surprisal values across the MWE tokens. Given that MWEs typically exhibit predictable transitions between tokens, our hypothesis is that surprisal values would tend to decrease within the MWE, resulting in a negative slope. While some transitions may present an increase in surprisal, the overall slope should indicate a facilitation in processing.

To measure this, a linear regression (first-degree polynomial fit) was applied to the surprisal sequence $S(w_1), S(w_2), \ldots, S(w_n)$ for each MWE of length n. The analysis was implemented in Python using the `numpy` library for numerical computations.

3.4 Surprisal Slope as a Discriminative MWE Feature

To analyze whether surprisal slope serves as a discriminative feature of MWEs, we focused on the data provided by the DiMSUM corpus. Specifically, we calculated the surprisal slope values for all 5,044 MWE occurrences, including both continuous and discontinuous types. Additionally, we randomly extracted 5,044 non-MWE n-grams from the corpus, matching the length distribution of MWEs as shown in Figure 1, and computed their surprisal slopes as well.

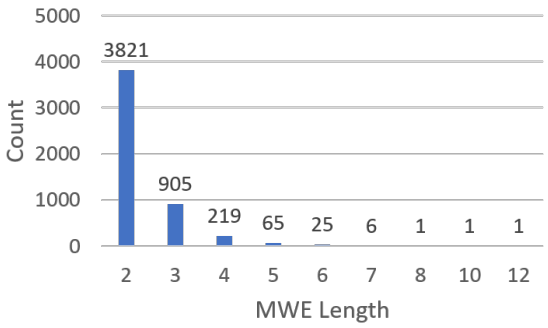

Figure 1: Count distribution of MWEs by length in the DiMSUM corpus, categorized by the number of tokens per MWE (e.g., bigrams = 2 tokens, trigrams = 3 tokens, etc.).

[2] https://pypi.org/project/surprisal/
[3] openai-community/gpt2

To investigate whether surprisal slope serves as a significant predictor for identifying MWEs, we conducted a logistic Generalized Additive Model (GAM) analysis. We prepared a combined dataset containing slope values extracted from both MWE and randomly selected non-MWE token sequences, labeled accordingly. Using the Python pygam library[4], we fitted a GAM with slope as the sole predictor, incorporating a smooth term to capture potential nonlinear effects. The statistical significance of slope was assessed via p-value. A dependence plot was generated to visualize the relationship between surprisal slope and the probability of a sequence being an MWE.

3.5 Idiomatic Usage and Surprisal Variation

The second analysis aimed to examine whether surprisal slope differs between common bigrams used idiomatically versus literally. Using the SemEval 2022 Shared Task dataset on multilingual idiomaticity detection corpus, which contains annotated examples of idiomatic and literal usage, we performed logistic regression with surprisal slope as the predictor and usage label (idiomatic = 0, literal = 1) as the outcome. The analysis was implemented in Python with the statsmodels library[5]. The model was fitted to assess whether surprisal delta significantly distinguishes between the two usage types. A box plot was also generated to visualize the distribution of surprisal delta across idiomatic and literal usages of the bigrams.

4 Results

4.1 Surprisal-Based Differentiation of MWEs and Non-MWEs

As outlined in Subsection 3.4, the objective is to assess whether surprisal slope constitutes a significant factor in distinguishing MWEs from randomly selected sequences of words.

Thus, we compared the surprisal slope values of the 5,044 MWE occurrences in the DiMSUM corpus with an equal number of randomly extracted non-MWE sequences, matched for n-gram length distribution, from the same corpus.

Table 1 presents the number of occurrences from the MWE and non-MWE lists that present a negative slope.

As shown in Table 1, a substantial proportion of MWEs exhibit negative surprisal slopes, compared to a considerably smaller percentage of non-MWE sequences. This difference suggests that MWEs tend to show a more consistent decrease in surprisal across their tokens, aligning with the hypothesis that MWEs involve more predictable transitions between words. While surprisal reduction is expected during any incremental processing of language (Hale, 2001), the stronger and more systematic decrease observed in MWEs points to their higher degree of contextual predictability and entrenched usage.

	Negative Slope	%
MWE	3998	79.3
non-MWE	2591	51.4

Table 1: Count of Sequences with Negative Surprisal Slopes in MWE and Matched Non-MWE Sets.

In terms of precision, recall, and F1-measure:

- **Precision**: 60.6%
- **Recall**: 79.3%
- **F1-measure**: 68.7%

These results indicate that the surprisal slope is a reasonably strong discriminative feature for identifying MWEs. The relatively high recall (79.3%) suggests that the method is effective at capturing a large proportion of actual MWEs, while the lower precision (60.6%) indicates that some non-MWEs are also classified as MWEs based only on this feature.

This is further supported by the GAM analysis, which identifies surprisal slope as a significant predictor of MWE status. Figure 2 illustrates the effect of slope on the probability of a sequence being classified as an MWE, showing how this probability varies across different slope values.

As illustrated in Figure 2, a negative surprisal slope increases the likelihood that a sequence is classified as an MWE. While the predicted probability approaches 1 for strongly negative slopes, this region includes relatively few instances. The most densely populated region with negative slopes is between –20 and 0, where the predicted probability remains around 0.8. This probability then drops sharply to approximately 0.2 as the slope becomes increasingly positive, especially in the 0 to 10 range, which contains the highest density of positive-slope instances. This pattern supports the hypothesis that MWEs tend to exhibit more predictable token transitions, reflected in their steeper

[4] https://pypi.org/project/pygam/
[5] https://pypi.org/project/statsmodels/

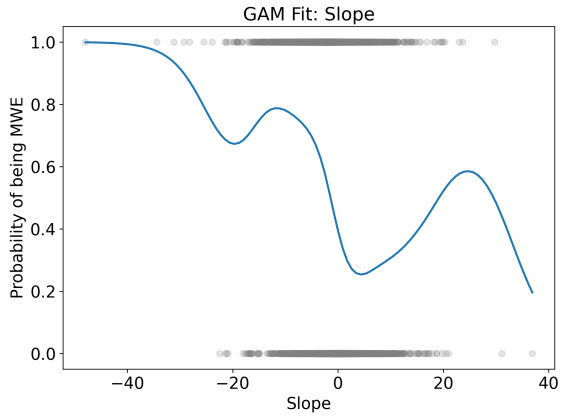

Figure 2: GAM Analysis of Surprisal Slope Impact on MWE Identification Probability.

negative surprisal slopes. We also observe a secondary peak in predicted probability (around 0.6) near a slope value of 20, suggesting that some sequences labelled as MWEs in the DiMSUM corpus do not exhibit the expected decrease in surprisal across tokens. These cases may reflect atypical or less compositional MWEs where predictability is not strongly reflected in surprisal patterns.

The GAM model achieved a pseudo R-squared of 0.0911, indicating a modest but meaningful proportion of deviance explained. Importantly, the effect of surprisal slope was found to be statistically significant ($p < 0.001$). While this confirms that slope contributes to distinguishing MWEs from non-MWEs, the relatively low explained variance suggests that other linguistic or contextual factors also play a role in determining MWE status.

One factor that may influence this result is that different parts of speech exhibit varying surprisal tendencies, with content words generally showing higher surprisal values than function words. Figure 3 illustrates the mean surprisal and standard deviation for each part-of-speech (PoS), calculated over the entire DiMSUM corpus annotated with surprisal and PoS tags (following the Universal Dependencies guidelines, De Marneffe et al. (2021)).

Thus, certain PoS combinations tend to exhibit negative surprisal slopes more frequently (e.g., VERB–ADP, PRON–AUX, etc.).

To account for PoS-related differences in surprisal, we applied two types of normalization to the surprisal values before calculating slopes. First, we normalized surprisal by dividing each token's value by the average surprisal of its PoS category, which adjusts for baseline differences across PoS types and centers the analysis on relative surprisal within each category rather than on absolute values. Second, we applied min–max normalization per PoS, rescaling each token's surprisal to the [0,1] range based on the minimum and maximum values observed for that PoS. This enables direct comparison of surprisal patterns across PoS categories with different distributional ranges.

After normalization, the surprisal slopes were recalculated following the same method described earlier. Table 2 presents the precision, recall, and F1 scores for these normalization approaches compared to the baseline model without surprisal normalization.

	Precision	Recall	F1
Baseline	**60.58**	79.26	68.67
Avg. srp. norm.	59.19	77.93	67.28
Min-max norm.	60.01	**82.81**	**69.59**

Table 2: Precision, recall, and F1 scores for surprisal slope-based MWE classification using different surprisal normalization strategies. The baseline corresponds to scores obtained without any normalization. For each metric, the highest value is shown in **bold**.

Of the tested normalization methods, only the min–max surprisal normalization yielded a modest improvement in recall and F1 score, though it resulted in a slight reduction in precision. Our results are relatively higher compared to the models presented in the SemEval-2016 Shared Task (Schneider et al., 2016), where the best-performing systems achieved F1 scores ranging from 54.8 to 61.09 depending on the source text (reviews, tweets, TED talks). They are also higher than those reported by Williams (2016), whose model achieved F1 scores between 0.48 and 0.62.

To better understand the limitations of using surprisal slope for MWE identification, Table 3 presents the top five PoS patterns associated with false negatives (i.e., MWEs exhibiting positive slope) and false positives (i.e., non-MWEs exhibiting negative slope).

In terms of false negatives, we observe that specific combinations of proper nouns and nouns often do not exhibit a negative surprisal slope. However, the DiMSUM corpus also includes other sequences with the same PoS patterns that do present negative slopes (e.g., *Lady Gaga*, *Justin Bieber*, *guest editor*, *birthday card*). This suggests that surprisal slope is more effective at identifying combinations that are more conventionalized and thus more frequent

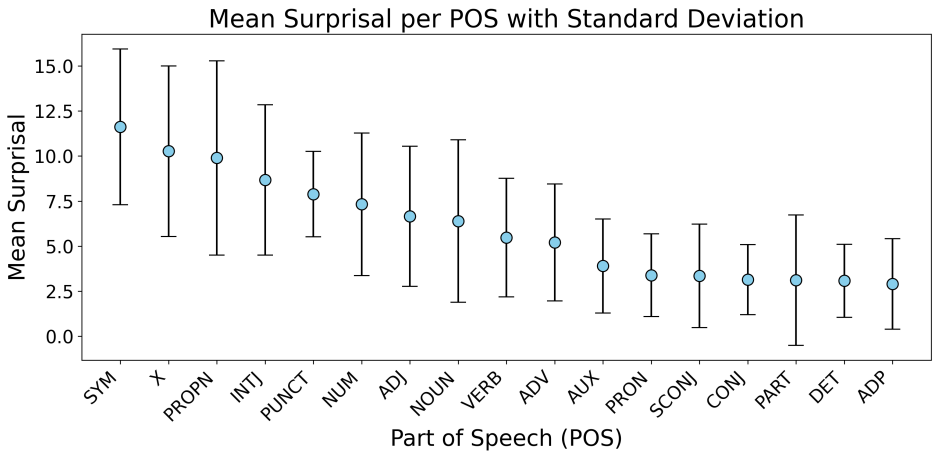

Figure 3: Mean Surprisal and Standard Deviation by Part-of-Speech in the DiMSUM Corpus.

MWEs with Positive Surprisal Slopes (False Negatives)		**Non-MWEs with Negative Surprisal Slopes (False Positives)**	
PoS Pattern	**Examples**	**PoS Pattern**	**Examples**
PROPN PROPN	*Jill Konrath, Kevin Wakeford, Nissan Terrano*	ADJ NOUN	*last week, amazing atmosphere, specific type*
NOUN NOUN	*Today stats, business agility, Job Search*	VERB PRON	*enjoy it, do it, offer you*
VERB NOUN	*taking rest, take chances, have experience*	NOUN ADP	*patient of, people at, lot of*
DET NOUN	*a couple, a pleasure, a lot*	VERB DET	*order a, use this, giving any*
ADP NOUN	*on time, in fact, in detail*	PRON VERB	*it was, we were, who wants*

Table 3: Top 5 PoS patterns among (left) MWEs with positive surprisal slopes (false negatives) and (right) non-MWEs with negative surprisal slopes (false positives).

in the training data of the language model used to estimate surprisal values.

Additionally, many of the major false negatives involve common collocations (e.g., VERB–NOUN pairs) and sequences where a function word precedes a content word. On the other hand, regarding false positives, we observe the opposite tendency: random associations of a content word followed by a function word often exhibit negative surprisal slopes (e.g., VERB–PRON and NOUN–ADP). Additionally, some adjective–noun combinations also appear as false positives, likely due to their frequency in language use, which may suggest they function as potential MWEs, although not annotated as such in DiMSUM.

4.2 Surprisal-Based Differentiation of Idiomaticity

As detailed in Subsection 3.5, this analysis focuses on comparing the surprisal slopes of a specific type of MWE: bigrams containing at least one noun that can occur either in a literal or an idiomatic sense.

For this analysis, we use the English data from the SemEval-2022 corpus for multilingual idiomaticity detection, which contains 2,755 sentences featuring 163 bigrams. Each instance is labelled as 0 for idiomatic usage and 1 for literal usage.

In this specific case, since the MWEs consist of only two tokens, the surprisal slope is equivalent to the surprisal delta, which corresponds to the surprisal of the second token minus that of the first. Only the sentence where the bigram occurs was considered for the surprisal estimation.

Table 4 presents the surprisal delta distribution,

showing the number of positive and negative deltas for both literal (1) and idiomatic (0) occurrences. The accuracy is also reported, assuming that the negative deltas indicate the MWE nature of the bigram.

Thus, regarding this specific set of bigrams, we notice that the vast majority present a negative delta, which is higher than the results obtained for DiMSUM, a more general dataset. We identified in Table 3 that NOUN-NOUN combinations in DiMSUM were among the MWE classes with the highest number of false negatives. However, the high accuracy observed with the SemEval 2022 bigrams can be attributed to their more conventionalized usage, which likely enhances the predictability of the second unit given the first.

Additionally, idiomatic usage appears to positively influence the prediction of the second token, making such bigrams easier to process. Consequently, the accuracy for bigrams used idiomatically is higher than that for those used literally.

We also attempted to include the preceding sentence from the corpus to assess whether additional context would increase the occurrence of negative deltas; however, no improvement was observed.

Regarding the logistic regression, Figure 4 illustrates the differences in surprisal delta between idiomatic and literal usages.

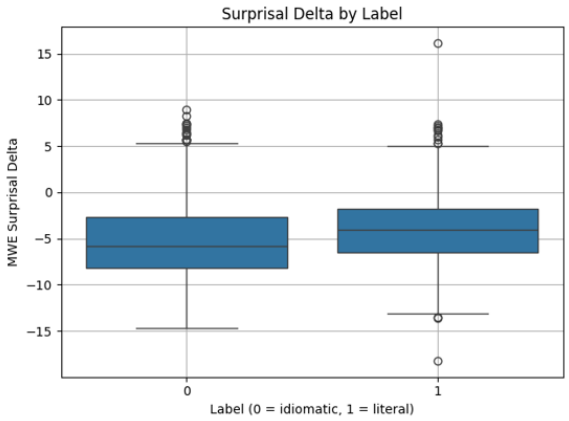

Figure 4: Surprisal Delta Distribution by Usage Label (0 = Idiomatic, 1 = Literal).

As anticipated from the previous analysis of the number of negative deltas per label, the boxplot primarily indicates a negative delta overall, with a notable tendency for the idiomatic usage to be more negative. Furthermore, using surprisal delta as a predictor in the logistic regression model reveals a significant effect (p-value < 0.001).

One advantage of this corpus is that each bigram type appears in a varying number of sentences. Therefore, we also examined, for each type, how many instances exhibit positive versus negative surprisal deltas. Among the 163 types in the dataset, only 7 show more positive deltas; the vast majority (95.9%) have a higher number of occurrences with negative surprisal delta. These findings suggest that surprisal is a more effective factor for determining the MWE status of a token sequence when evaluated across multiple utterances containing the MWE candidate.

Table 5 presents the bigrams for which the number of occurrences with positive surprisal delta exceeds that of negative ones.

We observe a clear imbalance between idiomatic and literal usage across compounds. For example, *heavy cross* and *white spirit* show a strong bias toward idiomatic occurrences, while *day shift* appears more frequently in literal contexts. Additionally, some bigrams such as *spinning jenny* occur only once or twice in the dataset, indicating limited data. This sparsity may influence surprisal estimation, potentially contributing to the observed positive deltas.

We also observe that in the majority of cases, the bigram consists of an adjective-noun (ADJ-NOUN) structure, with only two exceptions (i.e., *day shift* and *fashion plate*), which are compounds (NOUN-NOUN). This suggests that surprisal deltas may be particularly effective for identifying MWEs of the compound type.

5 Discussion

The results presented in Section 4 indicate that surprisal slope (or delta for bigrams) is indeed a strong predictor of MWEs, consistent with the findings of Onnis and Huettig (2021), who used a different methodology. Additionally, we have identified that idiomatic usages of MWEs tend to exhibit more negative deltas; however, literal usages also show a decrease in surprisal between the first and second units.

The GAM analysis using DiMSUM data indicates that, although surprisal slope is a significant predictor of MWEs, it alone is not sufficient for their precise identification. As previously explained, using surprisal slope tends to favor the identification of MWEs composed of specific parts of speech, typically a sequence of a content word followed by a function word as showed in Table 3.

Label	Positive Delta	Negative Delta	Accuracy
0	170	1254	88.1
1	231	1100	82.6
All	401	2354	85.4

Table 4: Distribution of Surprisal Deltas and Classification Accuracy in SemEval 2022 English MWEs (Label 0 = Idiomatic, Label 1 = Literal).

Bigram	0 +Δ	0 -Δ	1 +Δ	1 -Δ	Total +Δ	Total -Δ
heavy cross	7	2	1	2	8	4
big cheese	15	4	3	3	18	7
day shift	0	0	13	10	13	10
big wig	10	1	3	0	13	1
spinning jenny	1	0	0	0	1	0
white spirit	17	1	2	0	19	1
fashion plate	18	2	0	0	18	2

Table 5: Bigrams for which positive surprisal deltas outnumber negative ones, broken down by usage type (0 = idiomatic, 1 = literal).

However, the results using the SemEval 2022 corpus show that negative slopes are characteristic of conventionalized bigrams.

The PoS patterns presented in Table 3 also indicate that, although all compounds and combinations of proper nouns are labeled as MWEs in the DiMSUM data, not all of them meet the criteria for MWEhood if defined by a decrease in surprisal (i.e., facilitation in cognitive processing). Only conventionalized combinations of nouns or proper nouns exhibit this surprisal decrease.

Thus, our results indicate that surprisal variation can be used as a complementary methodology when using automatic methods for identification of MWEs (e.g., Klyueva et al. (2017) and Gries (2022)), especially if the analysis focuses on cognitive processing aspects of the usage of MWEs in specific registers (e.g., Alves et al. (2024)).

6 Conclusion and Future Work

This study explored the viability of surprisal slope as a feature for identifying multi-word expressions (MWEs) in English, using token-level surprisal patterns derived from the GPT-2 language model. Our evaluation on the DiMSUM and SemEval-2022 datasets demonstrated that surprisal slope provides moderate but meaningful discriminative power, with precision (60.6%) and recall (79.3%) indicating a trade-off between specificity and coverage. While the high recall suggests that surprisal slope effectively captures a majority of true MWEs, the lower precision highlights its tendency to misclassify some non-MWE sequences, particularly those with part-of-speech patterns resembling formulaic structures (e.g., adjective-noun or verb-pronoun combinations).

Notably, the method yielded better results for conventionalized MWEs, such as the bigrams in the SemEval-2022 corpus (used either in the idiomatic or literal meaning), where negative surprisal slopes (reflecting increased predictability) led to 85.4% accuracy in identifying bigrams as MWEs. This underscores the feature's strength for MWEs with strong contextual entrenchment. However, its performance varied across MWE categories, with adjective-noun bigrams less reliably exhibiting the expected slope patterns. Moreover, both idiomatic and literal usages showed negative slopes, with idiomatic instances tending to be more strongly negative.

These findings suggest that surprisal slope complements, but does not fully replace, existing MWE identification methods. Nevertheless, it offers a cognitively grounded perspective on formulaicity.

As future work, we intend to extend our analyses to additional MWE datasets, particularly those with well-defined annotation guidelines, such as PARSEME (Savary et al., 2017), and to include datasets in other languages. Furthermore, we aim to explore alternative ways of quantifying surprisal variation within MWE units, such as measuring surprisal fluctuation or distributional shifts across contexts. It may also be worth applying the whitespace correction technique proposed by Oh and Schuler

(2024) and Pimentel and Meister (2024) to ensure that estimated surprisals form a proper probability distribution. Relatedly, it would be interesting to investigate potential parallels between intra-word surprisal dynamics (within multi-token words) and inter-word dynamics (within MWEs). These directions may help capture subtler patterns of predictability and improve the robustness of surprisal-based MWE identification.

Limitations

There are a few limitations to consider in the present work. First, our analyses were based on only two MWE datasets, which, while representative, may not capture the full diversity of MWE types and usage contexts. Incorporating additional resources could improve the generalisability of our findings.

Second, our investigation was limited to English. Since MWEs are known to manifest differently across languages, especially in morphologically rich or syntactically flexible languages, it remains an open question whether the surprisal-based patterns observed here hold cross-linguistically.

Third, our regression models focused primarily on surprisal-based features. Future work should consider integrating complementary features such as lexical association measures (e.g., PMI, t-score) and dispersion statistics, which may offer additional insights into the structural and contextual properties of MWEs.

Finally, surprisal variability within MWE units was measured solely using the surprisal slope. Exploring alternative measures of surprisal variation, such as fluctuations or distributional properties, may capture more nuanced aspects of MWE predictability.

Acknowledgments

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

References

Diego Alves, Stefan Fischer, Stefania Degaetano-Ortlieb, and Elke Teich. 2024. Multi-word expressions in english scientific writing. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 67–76.

Inbal Arnon and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1):67–82.

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

Sirikarn Chantavarin, Emily Morgan, and Fernanda Ferreira. 2022. Robust processing advantage for binomial phrases with variant conjunctions. *Cognitive Science*, 46(9):1–43.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Kathy Conklin and Norbert Schmitt. 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, 29(1):72–89.

Kathy Conklin and Norbert Schmitt. 2012. The processing of formulaic language. *Annual Review of Applied Linguistics*, 32:45–61.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Vera Demberg and Frank Keller. 2009. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.

Stefan Th. Gries. 2022. Multi-word units (and tokenization more generally): A multi-dimensional and largely information-theoretic approach. *Lexis*, (19). Online since 26 March 2022.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Anders Johannsen, Dirk Hovy, Héctor Martínez Alonso, Barbara Plank, and Anders Søgaard. 2014. More or less supervised supersense tagging of twitter. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, pages 1–11.

Natalia Klyueva, Antoine Doucet, and Milan Straka. 2017. Neural networks for multi-word expression detection. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 60–65. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. *arXiv preprint arXiv:2406.10851*.

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.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. *arXiv preprint arXiv:2406.14561*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Carlos Ramisch, Abigail Walsh, Thomas Blanchard, and Shiva Taslimipoor. 2023. A survey of MWE identification experiments: The devil is in the details. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 106–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Alan Ritter, Sam Clark, Oren Etzioni, and 1 others. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534.

Agata Savary, Carlos Ramisch, Silvio Ricardo Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemi Zadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and 1 others. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *The 13th Workshop on Multiword Expression at EACL*, pages 31–47.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. Semeval-2016 task˜ 10: Detecting minimal semantic units and their meanings (dimsum). In *10th International Workshop on Semantic Evaluation*, pages 546–559. Association for Computational Linguistics.

Nathan Schneider and Noah A Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *A Corpus and Model Integrating Multiword Expressions and Supersenses*, pages 1537–1547.

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

Rita Simpson-Vlach and Nick C Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4):487–512.

A. Siyanova-Chanturia, K. Conklin, and W. J. B. van Heuven. 2011. Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3):776–784.

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.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Antoine Tremblay and R. Harald Baayen. 2010. Holistic processing of regular four-word sequences: A behavioral and erp study of the effects of structure, frequency, and probability on immediate free recall. In David Wood, editor, *Perspectives on Formulaic Language: Acquisition and Communication*, pages 151–173. Continuum International, London, UK.

Antoine Tremblay, Bruce Derwing, Gary Libben, and Chris Westbury. 2011. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, 61:569–613.

Jake Ryland Williams. 2016. Boundary-based mwe segmentation with text partitioning. *arXiv preprint arXiv:1608.02025*.

Chiraz Ben Youssef. 2024. *mMERGE: A Corpus Driven Multiword Expressions Discovery Algorithm*. Ph.D. thesis, University of California, Santa Barbara. ProQuest ID: BenYoussef_ucsb_0035D_16720; Merritt ID: ark:/13030/m5457324.

Wen Zhang, Taketoshi Yoshida, Xijin Tang, and Tu-Bao Ho. 2009. Improving effectiveness of mutual information for substantival multiword expression extraction. *Expert Systems with Applications*, 36(8):10919–10930.