# A New Comprehensive Annotation of Multiword Expressions

Jiaying Li Katie Krajovic

COSI 140: Natural Language Annotation for Machine Learning - Final Report

## Abstract

This paper presents a 15,722 word corpus of English text taken from the reviews section of the English Web Treebank. The corpus is annotated for Multiword Expressions (MWEs), under a novel annotation scheme in which each MWE is classified as Fixed, Semi-Fixed, or Flexible. A model trained on this corpus shows promising preliminary results for the task of lexical segmentation. At this point there does not appear to be a significant difference in performance when using the data from the classification of the MWEs as compared to a binary IOB labeling scheme. More data and further revision of the model will be necessary to determine the viability of this specific annotation scheme for the generation of a corpus to support lexical segmentation.

## 1 Introduction

A Multiword Expression can be broadly defined as a lexical item that can be decomposed into multiple lexemes and displays one or more of lexical, syntactic, semantic, pragmatic or statistical idiomaticity (Baldwin and Kim 2010). MWEs have traditionally been a challenging problem in NLP. Sag et al. present four specific problems: overgeneration, idiomaticity, flexibility, and lexical proliferation, associated with MWEs (Sag et al. 2002).

If all MWEs were simply treated as individual tokens, NLP systems could display a problem of overgeneralization in which expressions like *telephone booth* or *telephone box* could be generated, but additional expressions like *telephone cabinet* or *telephone closet* could be generated erroneously (Sag et al. 2002). A system like this would also likely struggle to predict the meaning of an idiomatic expression like *the cat's out of the bag*, in which the meaning of the expression is entirely distinct from the meaning of the component parts. If instead, MWEs were treated simply as words that contain whitespace, the flexibility of many MWEs would likely be lost. A parser might be able to correctly assign two interpretations to an expression like *look up the tower*, but be unable to disambiguate the similar but unambiguous expression *look the tower up* (Sag et al. 2002). Systems like these can rely heavily on hard coded lexicons of idioms, or multi-word verbs, typically do not generalize well and can be difficult to maintain, requiring manual updates as new expressions enter the language.

For these reasons, an automatic identification of MWEs could be useful in many applications across the NLP domain. Schneider et al. noted the potential usefulness of a tool like this, but also the lack of comprehensive corpora identifying MWEs (Schneider et al. 2014). They presented a new corpus, STREUSLE, annotated for all kinds of MWEs, and substantially higher performance from a model trained on this corpus for MWE identification than a model using traditional lexicon lookup.

The present study aims to build on this work, using a modified annotation scheme involving further classification of MWEs. The following sections of this paper will go into greater detail on the motivation of this new annotation, describe the annotation process, present the annotation scheme and corpus, and some preliminary results of a model trained on this data.

## 2 Background

### 2.1 STREUSLE Corpus

Schneider et al. present a 55,000 word corpus of informal English text, taken from the reviews section of the English Web Treebank (Schneider et al. 2014). Prior to this corpus, there had been no comprehensive resource for MWEs. Most of the tools that existed were fixed lexicons of idioms, or idiomatic expressions. All annotation that had been done on MWEs typically focused on only one subgroup, for example named entities. The hope was that a general corpus annotated for all kinds of MWEs would provide a better resource for general MWE related tasks.

In the corpus, each MWE is represented as a grouping of tokens. These groupings can be discontiguous due to syntactic constructions, or flexibility within the MWE itself. Each MWE was additionally classified as either strong or weak, with expressions with more opaque meanings classified as strong, and more transparent meaning classified as weak. For example, the meaning of *close call,* particularly the use of *call*, is not easily interpretable from the meaning of the component parts. An expression like *narrow escape*, while not completely transparent in meaning, is more easily interpretable from the components. In these two examples *close call* would be classified as a strong MWE, while *narrow escape* would be classified as a weak MWE (Schneider et al. 2014).

Results from a model trained on the STREUSLE corpus for the task of lexical segmentation, the splitting of a text into lexical units (in which case MWEs should be grouped together), showed a substantial increase in performance when compared to models using prior techniques like lexicon lookup. They also observed a drop in performance when they trained the model on data without the strength distinction, suggesting that the classification of MWEs as strong or weak was aiding the lexical segmentation.

It is mentioned by Schneider et al., that during annotation, deciding whether or not a given expression was a MWE was often difficult, even after discussion between multiple annotators. It is also noted that in these cases of disagreement, the compromise was frequently to classify something as a weak MWE (Schneider et al. 2014). This, along with the previous discussion, seem to suggest that the distinction between strong and weak MWEs is not, and perhaps can not, be completely concretely defined, but the annotated distinction still helped to identify MWEs.

The present study aims to build on this idea of MWE classification, but with a different set of more concretely defined categories. We hope to see whether more standardization in the classification of MWEs could further improve results in a lexical segmentation task.

### 2.2 A Novel Approach

We propose a novel annotation scheme in which MWEs are classified into one of three categories: Fixed, Semi-Fixed, and Flexible. These categories are based on the syntactic distribution of the MWE, and should be distinguishable regardless of the relative strength of the expression. Definitions and examples are provided in the following section.

## 3 Multiword Expressions

### 3.1 Definition

A Multiword Expression is a single lexical unit, composed of multiple tokens delimited by whitespace, that displays some form of idiomaticity.

### 3.2 Idiomaticity

Idiomaticity can be lexical, syntactic, semantic, pragmatic, or statistical and generally refers to some aspect of the meaning of the expression being uninterpretable from the meanings or characteristics of the individual lexemes that comprise it (Baldwin and Kim 2010). Understanding idiomaticity and the various types is crucial in determining whether a given expression is a MWE.

**Lexical.** Lexical idiomaticity occurs when one or more of the lexemes in the MWE are not of the English lexicon. If the lexeme has no English meaning on its own, then the meaning of the expression cannot be derived from the meaning of the lexemes. Some examples include: *mutatis mutandis, ad hoc, summa cum laude, c'est la vie.*

**Syntactic.** Syntactic idiomaticity occurs when the part of speech of one or more of the lexemes in the MWE do not correspond with the syntactic category of the expression. For example, *by and large* is an adverbial phrase consisting of a Preposition, Conjunction, and Adjective (Baldwin and Kim 2010).

**Semantic.** Semantic idiomaticity occurs when the meaning of an expression does not correspond to the meaning of its component parts. For example, the expression, *neither here nor there,* means roughly 'of no importance', and has nothing to do with locations near or far. The meaning may not be entirely opaque, but it is also not readily interpretable from the comprising lexemes. This is arguably the most common form of idiomaticity, and often even if a MWE displays another form, it will display this form as well. There are many examples, spanning various syntactic categories, ranging from *kick the bucket* to *around the clock.*

**Pragmatic.** Pragmatic idiomaticity is evident in expressions that are associated with a specific context, and often have a specialized meaning in that context, that they do not maintain elsewhere. For example the expression *you're welcome* is frequently used to acknowledge some expression of thanks. In this default context, the meaning is very different than in a context like *you're welcome to stay with me while you're in Boston.*

**Statistical.** Statistical idiomaticity simply refers to expressions composed of lexemes that occur together more than would be expected based simply on the distribution of the individual lexemes themselves. These are frequently referred to as collocations. Examples include: *narrow escape* and *quick bite.*

### 3.3 Classification

MWEs can be further broken down into three classifications based on their syntactic distribution: Fixed, Semi-Fixed, and Flexible.

**Fixed.** A Fixed MWE is one that does not undergo morphosyntactic variation or internal modification. A Fixed MWE can never be made of discontiguous tokens, has completely fixed word order, and cannot be inflected. Examples typically include proper names, MWEs that display lexical idiomaticity, fully fixed idioms, and some prepositional expressions: *The White House, Mr. Rogers, at all, on foot, beggars can't be choosers, c'est la vie.*

**Semi-Fixed.** A Semi-Fixed expression is one that can undergo some morphosyntactic variation, like inflection. A Semi-Fixed MWE can never be composed of discontiguous tokens, and maintains a fixed word order. Examples include nominal expressions, like noun noun compounds, modal constructions, and verbs with prepositions: *family tree, have to, want to, look for, work with, depend on.*

**Flexible.** Flexible expressions can undergo syntactic variation, they may have changing word order, can be composed of discontiguous tokens, and may take many different arguments. Examples include verbs with particles, quantity phrases, and light verbs: *pick* the kids *up, pick up* the kids, *a* large *number of*, *a* small *number of, a number* *of, take* a *walk*, *make* the most horrible *mistake*.

# 4 Annotation

## 4.1 The Annotation Task

The data used in this annotation task were 1,000 sentences taken from the STREUSLE corpus. Only the raw text was used.

Annotation was completed over three weeks by three annotators. The annotators were two master’s students and an undergraduate, each with minimally one course in Syntactic Theory completed. They were compensated for their annotation with course credit.

Each annotator received 680 short sentences to annotate, resulting in a final data set of two completed annotations for each of 1000 sentences. The 680 sentences were broken up into files each containing 20 sentences. These files could be used directly in the annotation tool and sent back when the annotation was completed.

A Slack channel was set up during the annotation period to facilitate communication between the annotators and the guideline writers. Annotators were encouraged to ask questions about any cases in which they were unsure.

**Guidelines.** The annotation guidelines were a PDF document shared with the annotators. They provided detailed descriptions of MWEs, as well as instructions and suggested steps for identifying and classifying each MWE. Several pages of example expressions and tricky cases were also provided. They were also shared in the form of a google doc, in case any subtle changes needed to be made as a result of cases identified during annotation. Ultimately this was not the case, and no updates were made to the guidelines.

Additional task-specific instructions included: a MWE can never span multiple sentences, each sentence must be evaluated as a completely new context; punctuation like “’ (apostrophe) in a possessive or ‘.’ in an abbreviation, can be part of a MWE, but punctuation signaling the end of a sentence should not be included; spelling and grammatical errors should be ignored, and annotators should take their best guess at the intended meaning; determiners should not be included unless they are part of a fixed expression; discontiguous MWEs should only contain one gap.

**Annotation Interface.** MAE was used as the annotation interface. A DTD file containing the annotation schema, which could be loaded directly into MAE, was provided to the annotators. Their data was provided to them in 34 XML files which could be opened in MAE and updated with tags as the annotation was completed. Annotators were able to easily select a span, contiguous or discontiguous, and tag it with the classification corresponding to the type of MWE selected. MAE was also used during adjudication.

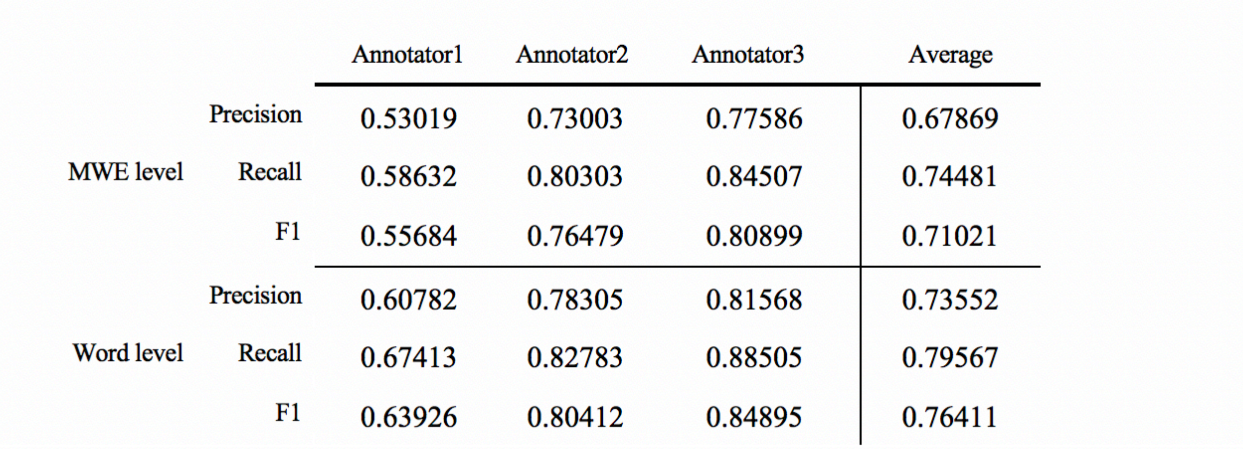
**Adjudication.** Two rounds of adjudication were performed, with one guideline writer performing an initial pass, and generating the first draft of the gold standard from the two completed annotations of any given file, and the other guideline writer performing a second pass comparing the two annotations as well as the gold standard draft. Any points of disagreement were discussed, and agreement was reached relatively easily. After the second pass was completed, the gold standard was considered finalized.

## 4.2 Inter-annotator Agreement

Several metrics were used to compute Inter-annotator agreement (IAA). In order to measure agreement on the identification of MWEs, precision, recall, and F1 score were calculated for each of the three annotators against the gold standard. In order to measure agreement on classification of MWEs, Cohen’s Kappa was calculated.

### 4.2.1 Identification Agreement

The first calculations (MWE level) for precision, recall, and F1 considered only



|  |  | Annotator1 | Annotator2 | Annotator3 | Average |
|---|---|---|---|---|---|
| MWE level | Precision | 0.53019 | 0.73003 | 0.77586 | 0.67869 |
|  | Recall | 0.58632 | 0.80303 | 0.84507 | 0.74481 |
|  | F1 | 0.55684 | 0.76479 | 0.80899 | 0.71021 |
| Word level | Precision | 0.60782 | 0.78305 | 0.81568 | 0.73552 |
|  | Recall | 0.67413 | 0.82783 | 0.88505 | 0.79567 |
|  | F1 | 0.63926 | 0.80412 | 0.84895 | 0.76411 |

**Table 1:** Agreement metrics for MWE identification between each of three annotators and the gold standard. One set of scores given for exact matches of full MWEs, the other for matches at the word level.

MWEs with identical spans. If an annotator's MWE differed from one identified in the gold standard by even one character, they received no credit for the identification of this MWE. A second set of calculations (word level) were then completed, in which annotators were awarded credit for any word that they identified as part of a MWE, that was also identified as part of a MWE in the gold standard. The results of both of these sets of calculations are shown in Table 1.

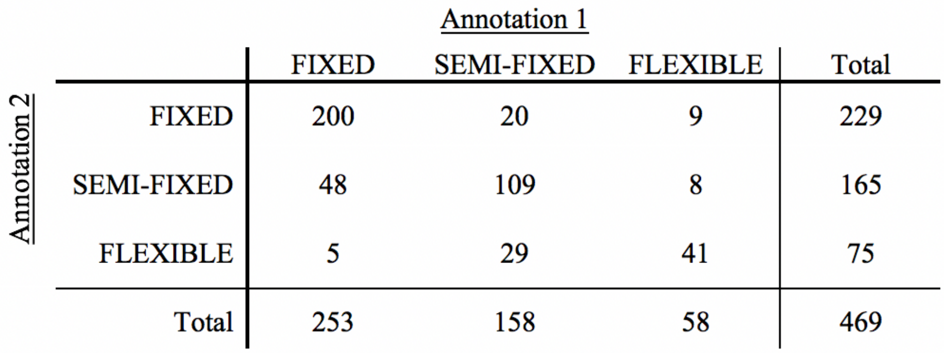
Word level agreement is expectedly higher than agreement at the full MWE level. This is because calculating agreement at the word level effectively gives annotators partial credit. There was much consideration of whether or not these partial credit scores were warranted for this specific task. Since the task is focused on identifying expressions that function as single lexical units, it is important that annotators are identifying only, and the entirety of, the full expressions. Looking only for perfect matches though, completely ignores cases in which the expression identified by the annotator and the expression in the gold standard differ by a single determiner, for example, or even a single space. Enough cases like this were noted during adjudication that it seemed worthwhile to present metrics for both levels of agreement.

For two of the annotators, almost all scores were in the 75-90% range, with one exception being the 0.73003 precision score for Annotator2 at the MWE level. The third annotator produced scores between 50-60% at the MWE level, and between 60-70% at the word level. In all cases, the recall was higher than the precision, which suggests that annotators were consistently annotating more expressions than were in the gold standard. This seems consistent with the nature of the task, which in some cases can come down to an annotator's own intuition about how idiomatic a given expression is.

During adjudication there was discussion about tricky cases and disagreements in judgements. This was not something the annotators had the opportunity to do, as annotation was completed independently. MWEs were only added to the gold standard when both adjudicators agreed that they should be. This effectively narrowed the MWEs in the gold standard as they had to pass both adjudicators' intuitions about idiomaticity. Because of this difference in procedure between creating the gold standard and completing one round of annotation, the higher recall is unsurprising.

### 4.2.2 Classification Agreement

Cohen's Kappa was calculated to represent the agreement between classifications of MWEs. Since in order for there to be agreement, an MWE must be classified by more than one annotator, the only data used in this calculation were those MWEs identified by both annotators who looked at any given file. The data were therefore a subset of the MWEs in the gold standard, as



| | | Annotation 1 | | | |
|---|---|---|---|---|---|
| | | FIXED | SEMI-FIXED | FLEXIBLE | Total |
| Annotation 2 | FIXED | 200 | 20 | 9 | 229 |
| | SEMI-FIXED | 48 | 109 | 8 | 165 |
| | FLEXIBLE | 5 | 29 | 41 | 75 |
| | Total | 253 | 158 | 58 | 469 |

**Table 2:** Agreement matrix for the classification of MWEs from which Cohen's Kappa was calculated.

some MWEs in the gold standard were identified by only one annotator.

The resulting Kappa score was 0.575917. This number alone does not seem to suggest particularly high agreement. It can be seen from the agreement matrix given in Table 2, that there are 3 noticeably high instances of disagreement: Ann1:Semi, Ann2:Fixed; Ann1:Fixed, Ann2:Semi; Ann1:Semi, Ann2:Flexible; with counts 20, 48, and 29 respectively.

Throughout adjudication there were two very consistent errors in classification observed. Both of these errors were present in the batch of annotation represented in this table in Annotation 1.

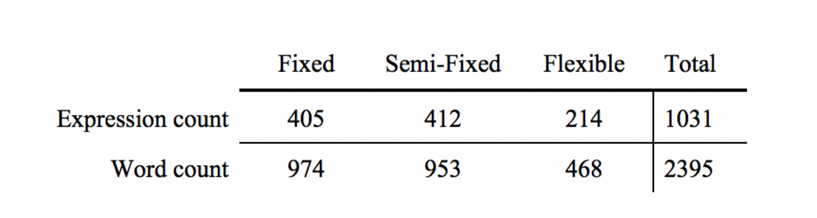
The first error, demonstrated consistently by one annotator, was to classify Noun-Noun compounds, which almost always fall into the Semi-Fixed category, as Fixed. This classification was made consistently by this annotator, for almost every instance of NN compound identified. It appeared to be a misinterpretation of the guidelines regarding the classification of NN compounds, rather than a random or case-by-case classification error.

The second error was also demonstrated consistently and by a single annotator, whose classifications here are included in Annotation 1. According to the annotation guidelines, any MWE that contains a gap is automatically classified as Flexible. This annotator seemed to have overlooked this aspect of the guidelines for classification, and instead used some other tests for deciding on a MWE type. He did occasionally classify MWEs with gaps as flexible, but frequently classified them as Semi-Fixed.

The effects of these consistent misclassifications can be seen in the values 48 and 29 in Table2. Since these mistakes represent a consistently applied misunderstanding, rather than legitimate confusion or disagreement between annotators, it is interesting to estimate what agreement might have looked like if these two simple misinterpretations had been clarified and corrected.

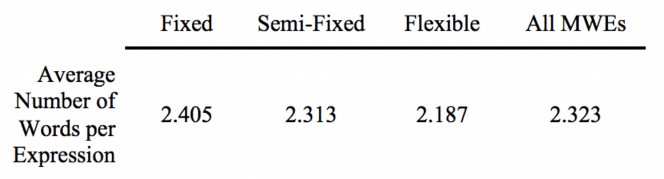
If these two values were instead replaced with the average (10) of the rest of the cells representing disagreement in this matrix, as an attempt of estimating what the disagreement might have looked like if those consistent mistakes had not occurred, the Kappa score improves drastically to 0.74355.

This may actually be a more informative metric, and an updated table could help to understand the shortcomings of the current guidelines. For example, there was not observed any single consistent mistake that could have led to the third "high" value in the agreement matrix, where Ann1 classified things as Semi-Fixed and Ann2 classified them as Fixed. The distinction between Fixed and Semi-Fixed MWEs therefore, seems to be the most challenging to make, and efforts could be focused on refining the guidelines regarding this distinction.



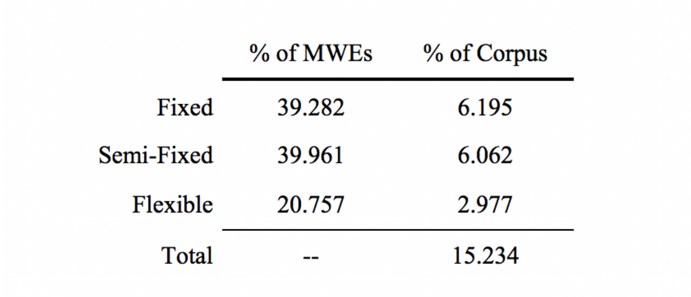
| | Fixed | Semi-Fixed | Flexible | Total |
|---|---|---|---|---|
| Expression count | 405 | 412 | 214 | 1031 |
| Word count | 974 | 953 | 468 | 2395 |

**Table 3:** The number of expressions of each category of MWE found in this corpus, as well as the number of words in the expressions in each category.



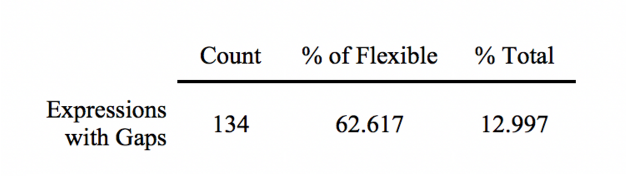
| | Fixed | Semi-Fixed | Flexible | All MWEs |
|---|---|---|---|---|
| Average Number of Words per Expression | 2.405 | 2.313 | 2.187 | 2.323 |

**Table 4:** Average number of words in each type of MWE.



| | % of MWEs | % of Corpus |
|---|---|---|
| Fixed | 39.282 | 6.195 |
| Semi-Fixed | 39.961 | 6.062 |
| Flexible | 20.757 | 2.977 |
| Total | -- | 15.234 |

**Table 5:** The distribution of MWEs across the three types, as well as the percent of words in the corpus that are part of each of the three types, and MWEs in general.



| | Count | % of Flexible | % Total |
|---|---|---|---|
| Expressions with Gaps | 134 | 62.617 | 12.997 |

**Table 6:** Distribution of the gappy MWEs in the corpus.

## 5 Corpus

The presented corpus is composed of 15,722 words, 1,000 sentences, of informal English text taken from the reviews sections of the English Web Treebank (Bies et al. 2012). It is comprehensively annotated for Multiword Expressions, with each expression tagged as Fixed, Semi-Fixed, or Flexible.

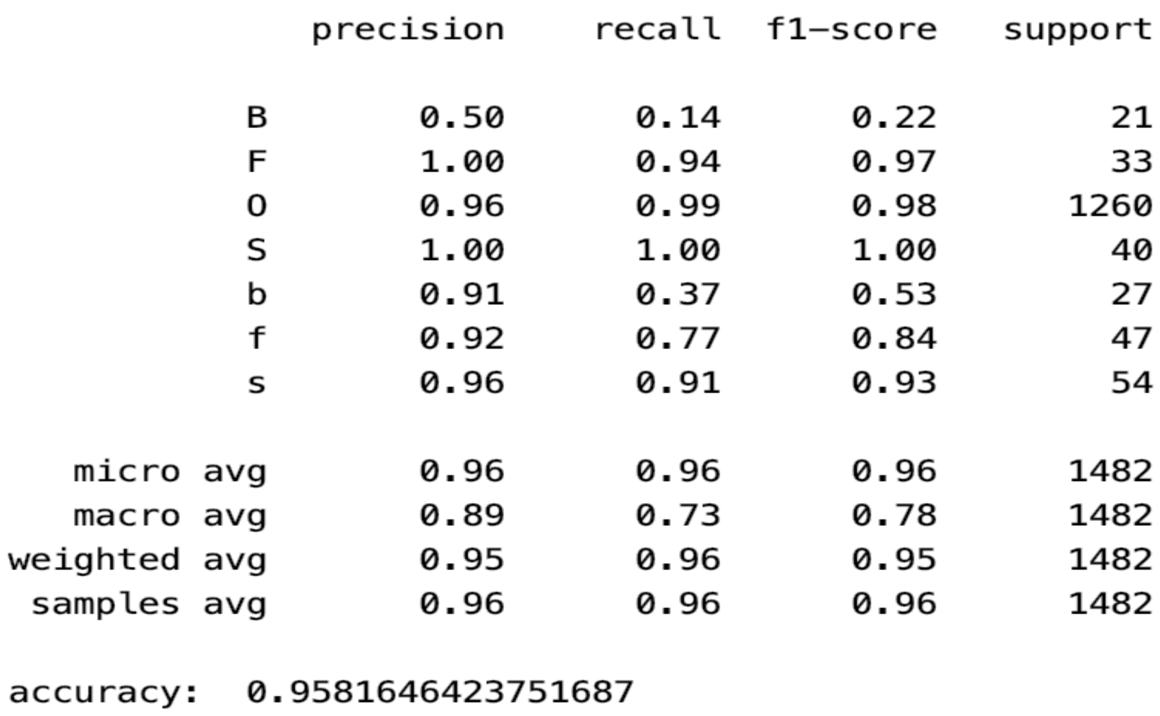
The 1,000 sentences are a random sample taken from the STREUSLE corpus, which is an annotation of the entire reviews section of the EWT. No sentences in the corpus are specifically selected as containing MWEs, they are simply regular English sentences written in informal online reviews. Therefore, there are obviously sentences that contain MWEs in the corpus, but it is not the case that every sentence is guaranteed to contain a MWE, or that MWEs are any more common in this corpus than they would be in any random text.

A total of 1,031 MWEs are annotated in the corpus, with roughly 40% being Fixed, 40% Semi-Fixed, and 20% Flexible. Roughly 62% of the Flexible expressions contained gaps. Just over 15% of all words in the corpus are part of a MWE. This is roughly the same percentage demonstrated in the STREUSLE corpus, which is an excellent indication that both of these corpora are in fact annotated comprehensively for MWEs. The average length of a MWE is 2.323 words. Additional details and statistics are presented in Tables 3-6.

## 6 Model

### 6.1 CRF model

Multiword expression (MWE) identifications can be viewed as a sequence labeling task, which involves assigning categories to each word. We have two sets of labeling schemes, one specifying MWEs into fixed, semi-fixed and flexible, the other without specifying the categories of MWE and only discriminating between MWE and non-MWE. For the first labeling scheme, we used seven labels including “F”, “f”, “S”, “s”, “B”, “b”, “0". In more detail, considering that we have three categories of MWEs, fixed, semi-fixed and flexible, “F” is the first token in a fixed MWE, “f” is any subsequent token in a fixed MWE, “S” is the first token in a semi-fixed MWE, “s” is any subsequent token in a semi-fixed MWE, “B” is the first token in a flexible MWE, “b” is any subsequent token in a flexible MWE, and “0” represents a token



```
              precision    recall  f1-score   support

           B       0.50      0.14      0.22        21
           F       1.00      0.94      0.97        33
           O       0.96      0.99      0.98      1260
           S       1.00      1.00      1.00        40
           b       0.91      0.37      0.53        27
           f       0.92      0.77      0.84        47
           s       0.96      0.91      0.93        54

   micro avg       0.96      0.96      0.96      1482
   macro avg       0.89      0.73      0.78      1482
weighted avg       0.95      0.96      0.95      1482
 samples avg       0.96      0.96      0.96      1482

accuracy:  0.9581646423751687
```

**Table 7:** evaluation report on the first labeling scheme on fold 9

that is not part of an MWE. For example, the sentence ”Best Store in Boothbay Harbor” can be labeled as Best/0 Store/0 In/0 Boothbay/F Harbor/f, where “BoothBay Harbor” is a fixed MWE so these two words are labeled as “F” and “f” respectively, while the other words are labeled as “0” for they are not parts of any MWEs. For the second more general scheme, we used three labels including “M” for the first token of the MWE, “m” for the subsequent token of the MWE and “0” for the non-MWE token.

For this sequence labeling task, we used the conditional random field(CRF) machine learning model. CRF is a discriminative model and it’s based on the conditional probability distribution. One advantage of the CRF model is that it doesn’t rely on the independence assumption that the labels are independent of each other and therefore avoids label bias.

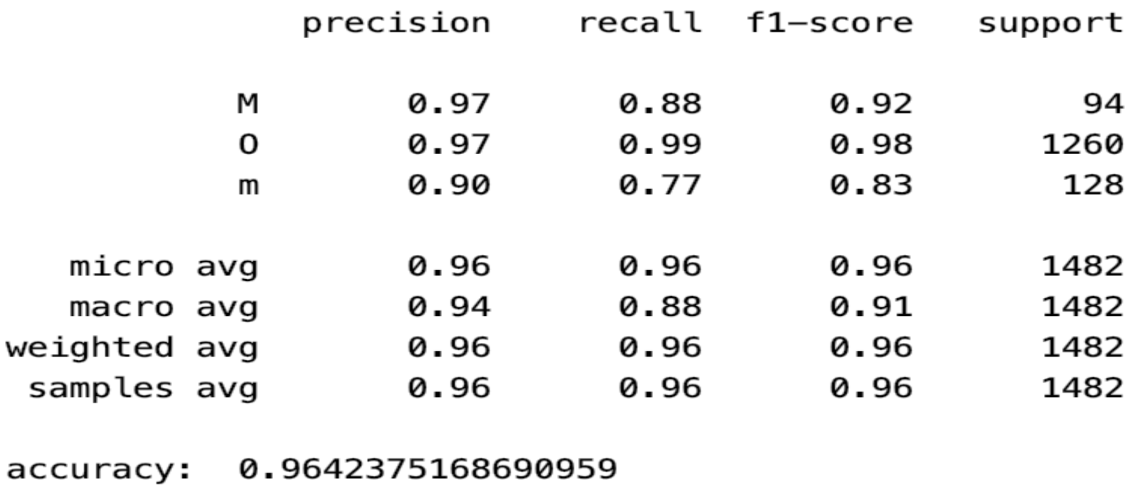
For the feature extraction, we extracted information from a window of four words including the current word “word”, two previous words “word-1” and “word-2”, two previous tags “label-1” and “label-2”, and the next word “word+1”. For instance, in a sentence “The/0 worst/0 Burger/F King/f restaurant/0 !!!/0”, for the current word “Burger”, its feature set is {word=Burger, word-1=worst, word-2=The, word+1=King, label-1=0, label-2=0}.

In our modeling, we divided 1,000 sentences to a training set of size 800, a development set of size 100 and a test set of size 100. And we also performed a cross-validation of ten folds to compare the performance of two labeling schemes. We used the default L-BFGS training algorithm with Elastic Net (L1+L2) regularization and training iterations set to 50 for a better result.

### 6.2 Evaluation report

With the first labeling scheme specifying the categories of MWEs as fixed, semi-fixed and flexible in model training, we obtained an accuracy of 0.95. Details can be seen in Table 7.

After examining the performances of seven labels, we noticed that the classifier identifying “B” and “b” performs poorly compared with identifying other labels. Remember that “B” and “b” represent flexible MWEs, indicating that the model



```
              precision    recall  f1-score   support

           M       0.97      0.88      0.92        94
           O       0.97      0.99      0.98      1260
           m       0.90      0.77      0.83       128

   micro avg       0.96      0.96      0.96      1482
   macro avg       0.94      0.88      0.91      1482
weighted avg       0.96      0.96      0.96      1482
 samples avg       0.96      0.96      0.96      1482

accuracy:  0.9642375168690959
```

**Table 8:** evaluation report on the second labeling scheme on fold 9

rarely identifies the flexible MWEs correctly. So we concluded that the poor performance of MWE identification is mainly due to the lack of accuracy in identifying flexible MWEs.

One of the explanations could be that quite a few flexible MWEs are gappy expressions and our feature sets only extract the information of a window of four words, which made it hard for the model to catch the full range of flexible MWEs. For instance, in the sentence "Because they cut_FLEXIBLE1 me good deals_FLEXIBLE1 if I paid in_FIXED1 cash_FIXED1.", the flexible MWE is "cut…deals" and the fixed MWE is "in cash"; for the fixed MWE "in cash", the feature set of either word would catch the information about the other, while for the flexible MWE "cut…deals", the feature set of either word could not catch the information about the other, as a result it might be omitted as a flexible MWE in model training.

The model with specific categorization into fixed, semi-fixed, and flexible ones helped us find out the poor performance of MWE identification relies on the identification of flexible MWE but we still identified different categories of MWEs as "_MWE" for our final segmentation purposes.

With the second labeling scheme, MWE and non-MWE, without specifying the categories further, the results are as seen in Table 8.

The accuracy of MWE and non-MWE Identification is 0.96, a bit higher than the model with specific categorization. To compare whether two labeling schemes make a difference in the identification of MWE, we performed a cross-validation of ten folds for two labeling schemes, the results are seen in Table 9, at the top of the next page.

We can tell from the statistics that there are no huge differences between categorizing MWE into fixed, semi-fixed or flexible and just discriminating between MWE and non-MWE. And the reason could be that our corpus contains a rather small data with 1,000 sentences so that there is no way to tell whether one labeling scheme is better than the other.

## 7 Conclusion

The preliminary results of a model trained on the minimal data presented in this paper are inconclusive as to whether the presented annotation scheme could produce a useful resource for the task of lexical segmentation. Comparison to the STREUSLE corpus suggests that both of these resources are indeed annotated comprehensively for MWEs. There could be value in the expansion or continuation of this project, as

|  | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Four categories (7 labels)** | 0.96337 | 0.96905 | 0.96197 | 0.95898 | 0.9714 | 0.96683 | 0.95455 | 0.95769 | 0.95816 | 0.9646 |
| **Two categories (3 labels)** | 0.96337 | 0.96715 | 0.9601 | 0.96552 | 0.97081 | 0.96441 | 0.96287 | 0.96968 | 0.96424 | 0.95809 |

**Table 9:** accuracy comparison for two labeling schemes on the 10 folds of cross validation, and red mark indicates higher value

additional data and feature engineering could hopefully break what is currently an inconclusive equivalence in performance between a binary labeling scheme and the classification scheme proposed in this paper. Below are listed the things that would be changed if the work progresses.

### 7.1 Things to Improve Upon

**Guideline Improvements.** The first, and greatest change that needs to be made to our guidelines is to further emphasize that expressions are not to be marked as MWEs unless they are made of multiple tokens/contain white space. All three annotators frequently marked acronyms or hyphenated words as MWEs. While these tokens do, usually, represent the meaning of more than just a single lexeme, they are already grouped together as a single lexical unit, due to their lack of whitespace. Since the end goal of this annotation is to perform lexical segmentation, these sorts of phrases do not need to be annotated, as they are already a single token. It is mentioned in the guidelines, during the definition of MWEs, that an expression is only a MWE if it contains whitespace, but clearly this was not emphasized enough, or re-emphasized in the correct places. The guidelines should be updated to make this clearer.

**Annotation Procedure Improvements.** Probably the single greatest improvement that could have been made to the annotation process in this study would have been to have each annotator complete annotation of a small sample of files (maybe five), and then to assess and provide feedback on the annotation of these initial files before the rest of the annotation was completed. This would not have solved every problem or inconsistency, but it would very likely have been able to eliminate the consistent errors made in classification due to misinterpretation of the guidelines. More specifically, this could have eliminated the errors discussed previously, which heavily impacted the Kappa score. Luckily these consistent errors were easy to identify, and could be corrected during adjudication, so they did not have a large impact on the resulting data. They did however, impact the Kappa score to such a degree that it was difficult to interpret the strengths and weaknesses of our guidelines. An initial round of annotation with feedback could have resulted in more representative metrics, and greater ability to improve guidelines.

During the generation of the STREUSLE corpus, three different types of annotation occurred: individual, joint, and consensus (Scnheider et al. 2014). The present work, largely due to the timeframe, made use of only individual annotation. We feel there could have been large benefits to making use of joint and consensus annotation as well.

It has already been noted in the analysis of the precision and recall for MWE identification, that having two annotators look at and discuss the same text narrows and refines the MWEs that are selected. This was observed in the generation of the gold standard, and would likely be observed in an early round of annotation as well.

A further benefit of this approach would be enabling the annotators to help each other learn the guidelines. The guidelines are long, and detailed, and could be overwhelming, particularly at the beginning of annotation. It is likely that different annotators may more easily internalize specific parts of the guidelines than others. Allowing two annotators to work together in these early stages would pool their knowledge and facilitate more complete, and probably faster, learning of the guidelines. This would also likely help to eliminate consistent errors due to misinterpretation of the guidelines. This joint annotation, with two annotators, would likely happen relatively early on in the annotation process, possibly immediately after the initial five files were received and evaluated.

Another step that could be useful around the same time would be group or consensus annotation. The main purpose of this would be refining the guidelines, rather than directly helping the annotators. This should come at a time when all annotators have a fairly solid understanding of the guidelines and any consistent errors or misinterpretations have been identified and corrected. In this way, during group annotation, the guideline writers would be able to see where the tricky cases were, and where the guidelines were perhaps unclear or needed to be revised. This is the point in the process where hopefully, any confusion that caused the final divergent value in the agreement matrix for classification of MWEs (in which Annotator1 marked it as Semi-Fixed and Annotator 2 marked it as Fixed) , could be identified and eliminated. The large disagreement in this case suggests that many of the tricky classifications were happening between these two categories, and consensus annotation could help to identify what specifically caused the confusion.

**Model Improvements.** Though we can not tell which labeling scheme, one specifying MWE into fixed, semi-fixed, and flexible, the other only discriminating between MWE and non-MWE, produces better accuracy, we learned from the first labeling scheme that the flexible MWEs are a rather tricky case to identify among all the categories. And the gappy flexible MWEs mainly account for the failure of the identification of flexible MWEs because our feature sets only capture a window of four words and we couldn't include more words as the vectors would get sparser that way. As statistics suggests, we have 10% of gappy flexible MWEs, and this 10% is our first priority to take care of if we are going to improve the performance of our model. But from the feature sets we already had, there is no way to modify the number of words or labels to get a better result for a distinct reason. So we might need to take a fresh look at our feature sets and import new features, which are not limited to word spans. One idea could be including dependency relations between words, in which we can link discontinuous words together in our feature sets.

## 8 References

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. *Handbook of natural language processing*, *2*, 267-292

Bies, Ann, Mott, Justin, Warner, Colin, and Kulick, Seth (2012). English Web Treebank. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, PA

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002, February). Multiword expressions: A pain in the neck for NLP. In *International conference on intelligent text processing and computational linguistics* (pp. 1-15). Springer, Berlin, Heidelberg.

Schneider, N., Danchik, E., Dyer, C., & Smith, N. A. (2014). Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*, *2*, 193-206.

Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M. T., Conrad, H., & A Smith, N. (2014). Comprehensive annotation of multiword expressions in a social web corpus.