Comprehensive Annotation of Multiword Expressions in a Social Web Corpus

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the Ninth International Conference on Language Resources and Evaluation

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Comprehensive Annotation of Multiword Expressions in a Social Web Corpus

Nathan Schneider†  Spencer Onuffer  Nora Kazour  Emily Danchik
Michael T. Mordowanec  Henrietta Conrad  Noah A. Smith
Carnegie Mellon University
Pittsburgh, PA 15213, USA
†nschneid@cs.cmu.edu

Abstract

Multiword expressions (MWEs) are quite frequent in languages such as English, but their diversity, the scarcity of individual MWEs, and contextual ambiguity have presented obstacles to corpus-based studies and NLP systems addressing them as a class. Here we advocate for a comprehensive annotation approach: proceeding sentence by sentence, our annotators manually group tokens into MWEs according to guidelines that cover a broad range of multiword phenomena. Under this scheme, we have fully annotated an English web corpus for multiword expressions, including those containing gaps.

Keywords: multiword expressions, corpus annotation, social media

1. Introduction

We present a 55,000-word corpus of English web text annotated for multiword expressions (MWEs) with the aim of full coverage. It uses a novel annotation scheme that emphasizes:

- heterogeneity—the annotated MWEs are not restricted by syntactic construction;
- shallow but gappy grouping—MWEs are simple groupings of tokens, which need not be contiguous in the sentence; and
- expression strength—the most idiomatic MWEs are distinguished from (and can belong to) weaker collocations.

We examine these characteristics in turn below. Details of the annotation process appear in §2, and an overview of the resulting corpus in §3. The annotations are available for download at http://www.ark.cs.cmu.edu/LexSem.

1.1. Heterogeneity

By “multiword expression,” we mean a group of tokens in a sentence that cohere more strongly than ordinary syntactic combinations: that is, they are idiosyncratic in form, function, or frequency. As fig. 2 shows, the intuitive category of MWEs or idioms cannot be limited to any syntactic construction or semantic domain. The sheer number of multiword types and the rate at which new MWEs enter the language make development of a truly comprehensive lexicon prohibitive. Therefore, we set out to build a corpus of MWEs without restricting ourselves to certain candidates based on any list or syntactic category. Rather, annotators are simply shown one sentence at a time and asked to mark all combinations that they believe are multiword expressions. Examples from our corpus appear in figs. [1] and [3].

Given that automatic detection of multiword expressions has shown promise for such diverse applications as machine translation (Carpuat and Diab 2010), keyphrase/index term extraction (Newman et al. 2012), and language acquisition research (Ellis et al. 2008), a common corpus with MWEs would be useful to develop and compare techniques that would cut across applications. To our knowledge, however, none of the corpus resources to encode multiword expressions have done so in a general fashion. For English, resources marking some kinds of lexical idioms include: lexicons such as WordNet (Fellbaum 1998), SAID (Kuiper et al. 2003), and WikiMwe (Hartmann et al. 2012); targeted lists (Baldwin 2005, 2008; Cook et al. 2008; Tu and Roth 2011, 2012); websites like Wiktionary and Phrases.net; and large-scale corpora such as SemCor (Miller et al. 1993), the French Treebank (Abellé et al. 2003), the Szeged-ParalellIFX corpus (Vincze 2012), and the Prague Czech-English Dependency Treebank (Cmejrek et al. 2005). But each of these prioritizes certain kinds of MWEs to the exclusion of others. Consequently, the computational literature on multiword expressions (reviewed in Baldwin and Kim 2010) has been fragmented, looking (for example) at subclasses of phrasal verbs or nominal compounds in isolation. With regard to the aforementioned corpora, annotations of multiword compounds in the French Treebank and light verb constructions in SzegedParalellIFX have been used as a testbed for statistical learning of sequence taggers (Constant and Sigogne 2011, 2012; Poét et al. 2012) and MWE-aware parsers (Green et al. 2011, 2012; Constant et al. 2012), while the SemCor-driven task of noun and verb supersense tagging (Ciaramita and Altun 2006; Paab and Reichertz 2009) involves the identification of some multiword expressions. We hope a resource with more comprehensive MWE annotations will lead to more general-purpose approaches to MWEs.

Figure 1: Two sentences from the corpus. Subscripts and text coloring indicate strong multiword groupings; superscripts and underlining indicate weak groupings. Boxes indicate gaps.

1 Disciplines such as phraseology and language acquisition have dozens of other terms for various notions of MWEs: among these are fixed expression, formulaic sequence, fossilized idiom, phraseological unit, and prefabricated pattern (Moon 1998, Wray 2000).
1. **MW named entities:** Chancellor of the Exchequer Gordon Brown
2. **MW compounds:** red tape, motion picture, daddy longlegs, Bayes net, hot air balloon, skinny dip, trash talk
3. **conventionally SW compounds:** snapdragon, overlook (v. or n.), blackjack, shootout, sunscreen, somewhere
4. **verb-particle:** pick up, dry out, take over, cut short
5. **verb-preposition:** refer to, depend on, look for, prevent from
6. **verb-noun-(preposition):** pay attention to, go bananas, lose it, break a leg, make the most of
7. **support verb:** make decisions, take breaks, take pictures, have fun, perform surgery
8. **other phrasal verb:** put up with, miss out (on), get rid of, look forward to, run amok, cry foul, add insult to injury, make off with
9. **PP modifier:** above board, beyond the pale, under the weather, at all, from time to time, in the nick of time
10. **coordinated phrase:** cut and dry, more or less, up and leave
11. **conjunction/connector:** as well as, let alone, in spite of, on the face of it, on its face
12. **semi-fixed VP:** smack <one>‘s lips, pick up where <one> left off, go over <thing> with a fine-tooth(ed) comb, take <one>‘s time, draw <one>s up to <one>‘s full height
13. **fixed phrase:** easy as pie, narrowly escape, narrow escape, close call, therefore would treat <one> closely calling <one> near call, nor can the danger be described as *closely calling or *calling close. We would therefore treat close call as a strong MWE. On the other hand, the expression narrow escape is somewhat more transparent and flexible—one can narrowly escape/avoid an undesirable eventuality, and the alternative formulation close escape is acceptable, though less conventional—so it would therefore qualify as a weak MWE.

1. **Shallow token groupings**

Concretely, we represent each MWE as a grouping of tokens within a sentence. The tokens need not be contiguous: **gappy** (discontinuous) uses of an expression may arise due to internal arguments, internal modifiers, and constructions such as passives. For example, sentence \[1\] in fig. \[1\] contains a gappy instance of the verb-particle construction take in. It also contains two contiguous MWEs, the named entity 07 Ford Fusion and the noun-noun compound oil change. Syntactic annotations are not used or given as part of the MWE annotation, though MWEs can be syntactically categorized with part-of-speech tags (as in table \[2\] and fig. \[3\] or syntactic parses.

1. **Strength**

Qualitatively, the strength of association between words can vary on a continuum of lexicality, ranging from fully transparent collocations to completely opaque idioms (Hermann et al., 2012). In the interest of simplicity, we operationalize this distinction with two kinds of multiword groupings: **strong** and **weak**. For example, the expression close call describes a situation in which something bad nearly happened but was averted (He was late and nearly missed the performance—it was a close call). This semantics is not readily predictable from the expression: the motivation for call in this expression is opaque; and moreover, *near call and *far call are not acceptable variants nor can the danger be described as *closely calling or *calling close. We therefore would treat close call as a strong MWE. On the other hand, the expression narrow escape is somewhat more transparent and flexible—one can narrowly escape/avoid an undesirable eventuality, and the alternative formulation close escape is acceptable, though less conventional—so it would therefore qualify as a weak MWE.

While there are no perfect criteria for judging MWE-hood, several heuristics tend to be useful when a phrase’s status is in doubt. The strongest cues are semantic opacity and morphosyntactic idiosyncrasy: if a word has a function unique to a particular expression, or an expression buck

---

2 But note that close shave and near miss are other idioms using the same “proximity to danger” metaphor.

---

1. **Figure 2:** Some of the classes of idioms in English. The examples included here contain multiple lexicalized words—with the exception of those in (3), if the conventionally single-word spelling is used.

### 1.2. Shallow token groupings

Over the course of 5 months, we fully annotated the 55,000-word REVIEWS section of the English Web Treebank (Bies et al., 2012). Annotators were the first six authors of this paper. All are native speakers of English, and five hold undergraduate degrees in linguistics.

The annotation took three forms: (a) **individual** annotation (a single annotator working on their own); (b) **joint** annotation (collaborative work by two annotators who had already worked on the sentence independently); and (c) **consensus** annotation (by negotiation among three or more annotators, with discussion focused on refining the guidelines). Initially, consensus annotation sessions were held semi-weekly; the rate of these sessions decreased as agreement improved.

Though consensus annotations are only available for 1/5 of the sentences, every sentence was at least reviewed independently and jointly. The annotation software recorded the full version history of each sentence; during some phases of annotation this was exposed so that analyses from different annotators could be compared.

The judgment of whether an expression should qualify as an MWE relied largely on the annotator’s intuitions about its semantic coherence, idiosyncrasy, and entrenchment in the language. As noted in §1.3, the decision can be informed by heuristics. Judgments about the acceptability of syntac-
The guidelines document describes general issues and considerations along with informal web searches, were often used to investigate the fixedness of candidate MWEs; a more systematic use of corpus statistics (along the lines of [Wulff, 2008] might be adopted in the future to make the decision more rigorous.

**Annotation guidelines.** Annotation conventions were recorded on an ongoing basis as the annotation progressed. The guidelines document describes general issues and considerations (e.g., inflectional morphology; the spans of named entities; date/time/address/value expressions; overlapping expressions), then briefly discusses about 40 categories of constructions such as comparatives (as X as Y), age descriptions (N years old), complex prepositions (out of, in front of), discourse connectives (to start off with), and support verb constructions (make a decision, perform surgery).

Some further instructions to annotators include:

- **Groups should include only the lexically fixed parts of an expression (modulo inflectional morphology); this generally excludes determiners and pronouns:** made the mistake, pride themselves on.
- **Multiword proper names count as MWEs.**
- **Misspelled or unconventionally spelled tokens are interpreted according to the intended word if clear.**
- **Overtokenized words (spelled as two tokens, but conventionally one word) are joined as multiwords.** Clitics separated by the tokenization in the corpus—negative n’t, possessive ’s, etc.—are joined if functioning as a fixed part of a multiword (e.g., T’s Cafe), but not if used productively.
- **Some constructions require a possessive or reflexive argument (see semi-fixed VP examples in fig. 2). The possessive or reflexive marking is included in the MWE only if available as a separate token; possessive and reflexive pronouns are excluded because they contain the argument and the inflection in a single token. This is a limitation of the tokenization scheme used in the corpus.**

In some cases idiosyncratic constructions were rejected because they did not contain more than one lexicalized element: e.g., the construction have + <evaluative adjective> + <unit of time> (have an excellent day, had a bad week, etc.).

**Overlap.** A handful of cases of apparent MWE overlap emerged during the course of our annotation: e.g., for threw a surprise birthday party, the groups {throw, party}, {surprise, party}, and {birthday, party} would all have been reasonable; but, as they share a token in common, the compromise decision was to annotate {birthday, party} as a strong MWE and {throw, {birthday, party}} as a weak MWE.

**Annotation interface.** A custom web interface, fig. 3, was used for this annotation task. Given each pretokenized sentence, annotators added underscores (_) to join together strong multiwords and tildes (~) for weak MWEs. During joint annotation, the original annotations were displayed, and conflicts were automatically detected.

**Inter-annotator agreement.** Blind inter-annotator agreement figures show that, although there is some subjectivity to MWE judgments, annotators can be largely consistent. E.g., for one measurement over a sample of 200 sentences, the average inter-annotator $F_1$ over all 10 pairings of 5 annotators was 65%.

When those annotators were divided into two pairs and asked to negotiate an analysis with their partner, however, the agreement between the two pairs was 77%, thanks to reductions in oversights as well as the elimination of eccentric annotations.

**Difficult cases.** Prepositions were challenging throughout; it was particularly difficult to identify prepositional verbs (speak with? look for?). We believe a more systematic treatment of preposition semantics is necessary. Nominal compounds (pumpkin spice latte) and alleged support verbs (especially with get: get busy? get a flat?) were frequently controversial as well.

3. The Corpus

The MWE corpus consists of the full REVIEWS subsection of the English Web Treebank, comprising 55,579 words in 3,812 sentences. Each of the 723 documents is a user review of a service such as a restaurant,

4 Our measure of inter-annotator agreement is the precision/recall–based MUC criterion ([Vilain et al., 1995]. Originally developed for coreference resolution, it gives us a way to award partial credit for partial agreement on an expression.
dentist, or auto repair shop. As the Web Treebank does not provide metadata for reviews, one of our annotators coded all the documents for topic and sentiment. The distribution is shown in Table 3. The writing style of these reviews is informal, so we would expect a lot of colloquial idioms, perhaps for dramatic effect (especially given the strong opinions expressed in many reviews).

Summary statistics of the MWEs in the corpus are given in Table 1. Among the highlights:

• 15% of MWEs contain at least one gap, and 35% of gaps contain more than one token. 1.5% of tokens fall within a gap; 0.1% of tokens belong to an MWE nested within a gap (like ‘07 Ford Fusion and a little in fig. 1).

• 65% of the gaps are one word long; another 25% are two words long.

These figures demonstrate (i) that MWEs are quite frequent in the web reviews genre, and (ii) that annotators took advantage of the flexibility of the scheme to encode gappy expressions and a strength distinction.

Figure 4 shows the distribution of intra-MWE and extra-MWE words by part of speech. The MWE words are syntactically diverse: common nouns, verbs, proper nouns, prepositions, adverbs, adjectives, determiners, and particles account for most of them. Nearly all particles and nearly two thirds of proper nouns were marked as part of an MWE.

Categorizing MWEs by their coarse POS tag sequence, we find only 8 of these patterns that occur more than 100 times: common noun–proper noun, proper noun–proper noun, verb-preposition, verb-particle, verb-noun, adjective-noun, and verb-adverb. But there is a very long tail—460 patterns in total. For the interested reader, Table 2 shows the most frequent patterns, with examples of each. Many patterns are attested with and without gaps; a handful occur more frequently with gaps than without. About 78% of gaps are immediately preceded by a verb. There are 2,378 MWE types; 82% of these types occur only once; just 183 occur three or more times. The most frequent are:

• don’t get caught up in the hype
• don’t judge a book by its cover
• do n’t get catch up in the hype and do n’t judge a book by its cover.

Table 1: Annotated corpus statistics over 723 documents (3,812 sentences). 57% of sentences (72% of sentences over 10 words long) and 88% of documents contain at least one MWE. 8,060/55,579=15% of tokens belong to an MWE; in total, there are 3,024 strong and 459 weak MWE instances. 82 weak MWEs (18%) contain a strong MWE as a constituent (e.g., means a lot to me in fig. 1 and get in touch with in fig. 3).

 spokesmen (59 tokens). They are:

• 15% of MWEs contain at least one gap. (Only 6 contain two gaps.)
• 73% of the MWEs consist of two tokens; another 21% consist of three tokens.
• 16% of the MWEs are strong and contain a gold-tagged proper noun—most of these are proper names.
• 73% of the MWEs consist of two tokens; another 21% consist of three tokens.
• 15% of the MWEs contain at least one gap. (Only 6 contain two gaps.)

5 We can increase solidarity between the speaker/writer and hearer/reader), and Simpson and Mendis (2003) p. 434: “possible communicative effects [of idioms] include exaggeration, informal, and rhetorical flair.”

6 They are: offers’ a decent bang for the buck; take3 this as you far as we can; passed away silently in his sleep; asked me BS to bring my face back; putting me at ease; tells me
Table 2: All POS sequences occurring in at least 10 MWEs in the corpus (49 patterns). Contiguous and gappy MWE instances are counted separately. POS groupings are abbreviated with a single character (N for common nouns, ` for proper nouns, T for particles, etc.). Strong MWEs are joined with _ and weak MWEs with ~; weak MWE examples are italicized. MWE types occurring at least 10 times are bolded.

4. Conclusion

We have described a process for shallow annotation of heterogeneous multword expressions in running text. The annotation guidelines and our annotations for the English Web Treebank can be downloaded at: [http://www.ark.cs.cmu.edu/LexSem](http://www.ark.cs.cmu.edu/LexSem). Licensing restrictions prevent us from publishing the full text of every sentence, so we provide annotations in terms of token offsets in the original corpus. Tokens within the span of an MWE are retained.

An MWE identification system trained on our corpus is presented in Schneider et al. (2014). Other ongoing and future work includes extending the annotation scheme to new datasets; developing semi-automatic mechanisms to detect or discourage inconsistencies across sentences; and integrating complementary forms of semantic annotation of the
MWEs (such as WordNet synsets). These improvements will facilitate NLP tools in more accurately and informatively analyzing lexical semantics for the benefit of downstream applications.

Acknowledgments
This research was supported in part by NSF CAREER grant IIS-1054319, Google through the Reading is Believing project at CMU, and DARPA grant FA8750-12-2-0342 funded under the DEFT program. We are grateful to Kevin Knight, Martha Palmer, Chris Dyer, Lori Levin, Ed Hovy, Tim Baldwin, members of the NLP group at Berkeley and the Noah’s ARK group at CMU, and anonymous reviewers for valuable feedback.

References


Philadelphia, PA.