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 MWE *types*, 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 corpus 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*.[1] 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 (Abeillé et al., 2003), the Szeged-ParalellFX corpus (Vincze, 2012), and the Prague Czech-English Dependency Treebank (Čmejrek 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 SzegedParalellFX have been used as a testbed for statistical learning of sequence taggers (Constant and Sigogne, 2011; Constant et al., 2012; Vincze et al., 2013) 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; Paaß and Reichartz, 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.

(1) My wife had taken$_1$ her '07$_2$ Ford$_2$ Fusion$_2$ in$_1$ for a routine oil$_3$ change$_3$.

(2) he was willing to budge$_1$ a$_2$ little$_2$ on$_1$ the price which means^4 a$^4_3$ lot$^4_3$ to^4 me^4 .

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/connective:** *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* <oneself> *up to* <one>*'s full height*
13. **fixed phrase:** *easy as pie*, *scared to death*, *go to hell in a handbasket*, *bring home the bacon*, *leave of absence*, *sense of humor*
14. **phatic:** *You're welcome. Me neither!*
15. **proverb:** *Beggars can't be choosers. The early bird gets the worm. To each his own. One man's* $\langle thing_1 \rangle$ *is another man's* $\langle thing_2 \rangle$.

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

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. 4) or syntactic parses.

1.3. 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,[2] 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 bucks the usual grammatical conventions of the language, the expression is almost certainly an MWE. It often helps to test how fixed/fossilized the expression is, by substituting words with synonyms/antonyms, adding or removing modifiers, or rearranging the syntax. Another strategy is to search large corpora for the expression to see if it is much more frequent than alternatives. In practice, it is not uncommon for annotators to disagree even after considering these factors, and to compromise by marking something as a weak MWE.

For purposes of annotation, the only constraints on MWE groupings are: (a) a group must consist of two or more tokens; (b) all tokens in a group must belong to the same sentence; (c) a given token may belong to at most one strong group and at most one weak group; and (d) if a token belongs to both a strong group and a weak group, all other tokens in the strong group must belong to the same weak group.

2. Annotation

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-

[2] But note that *close shave* and *near miss* are other idioms using the same "proximity to danger" metaphor.

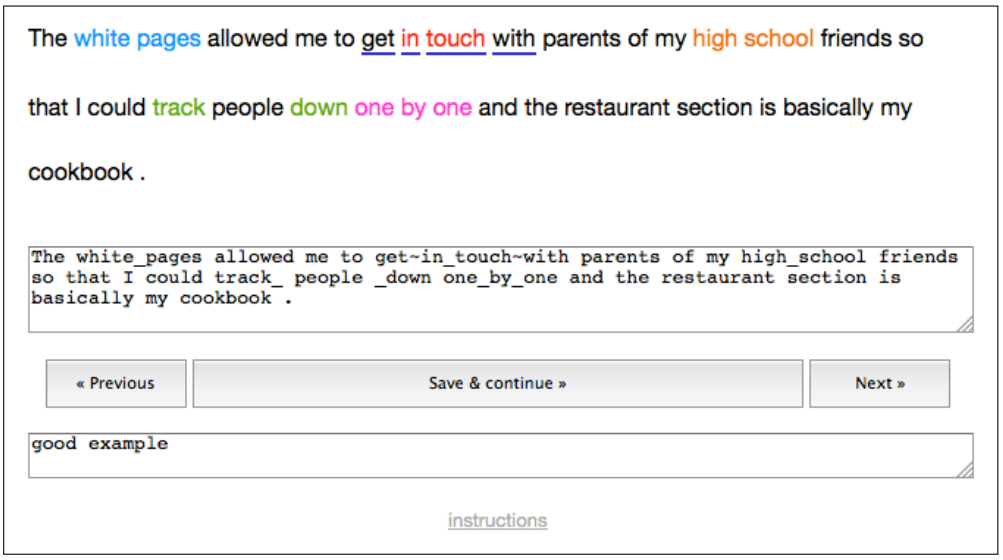

Figure 3: MWE annotation interface. The user joins together tokens in the textbox, and the groupings are reflected in the color-coded sentence above. (Invalid markup results in an error message.) A second textbox is for saving an optional note about the sentence. The web application also provides capabilities to see other annotations for the current sentence and to browse the list of sentences in the corpus (not shown).

tic manipulations and substitution of synonyms/antonyms, 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***.[3]
- 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.

[3] 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 {*threw*, *party*}, {*surprise*, *party*}, and {*birthday*, *party*} all would have been reasonable; but, as they share a token in common, the compromise decision was to annotate {*birthday*, *party*} as a strong MWE and {*threw*, {*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%.[4] 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*? *listen to*? *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 (Bies et al., 2012), 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.

	constituent tokens				
	2	**3**	**4**	**≥5**	**total**
strong	2,257	595	126	46	3,024
weak	269	121	44	25	459
	2,526	716	170	71	3,483

(a) MWE instances by number of constituent word tokens

number of gaps		
0	**1**	**2**
2,626	394	4
322	135	2
2,948	529	6

(b) MWEs by number of gaps

gap length		
1	**2**	**≥3**
259	98	45
93	38	8
352	136	53

(c) Gaps by length (in tokens)

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). ◆ *Gaps:* 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).

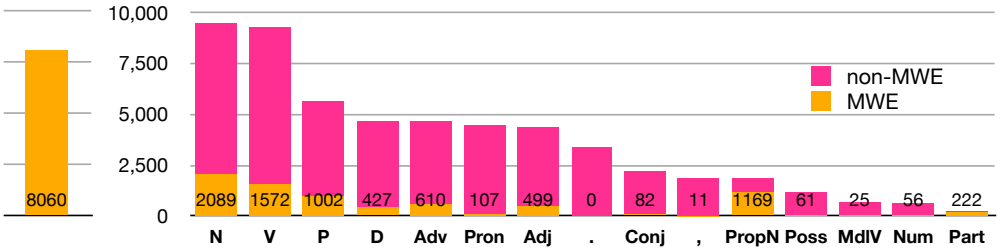

Figure 4: Distribution of tokens in the corpus by gold POS grouping and whether or not they belong to an MWE. Overall, 8,060 tokens are within an MWE; this not much less than the total number of common nouns (left). The rarest POS categories are not shown; of these, the only ones with large proportions of MWE tokens are hyphens (79/110) and incomplete words (28/31).

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).[5]

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

- The 3,483 MWEs include 15% of all tokens in the corpus. As a point of reference, 17% of all tokens are common nouns.
- 57% of sentences (72% of sentences over 10 words long) and 88% of documents contain at least one MWE.
- 87% of the MWEs are strong/13% are weak.
- 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.[6])
- 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–common 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.[7] 82% of these types occur only once; just 183 occur three or more times. The most frequent are *highly recommend(ed)*, *customer service*, *a lot*, *work with*, and *thank you*. The longest are 8 lemmas long, e.g. *do n't get catch up in the hype* and *do n't judge a book by its cover*.

[5] See, e.g., Nunberg et al. (1994, p. 493: "idioms are typically associated with relatively informal or colloquial registers and with popular speech and oral culture"), Moon (1998, p. 267: "[fixed expressions/idioms] can be seen as part of a discourse of familiarity. . . [they can] increase solidarity between the speaker/writer and hearer/reader"), and Simpson and Mendis (2003, p. 434: "possible communicative effects [of idioms] include exaggeration, informality, and rhetorical flair").

[6] They are: *offers*^{1} *a decent bang*$^{1}_{2}$ *for*$^{1}_{2}$ *the buck*$^{1}_{2}$; *take*$_{3}$ *this as*$_{3}$ *far*$_{3}$ *as*$_{3}$ *we can*$_{3}$; *passed*$^{5}_{4}$ *away*$^{5}_{4}$ *silently in*^{5} *his sleep*^{5}; *asked*$_{6}$ *Pomper for*$_{6}$ *my money back*$_{6}$; *putting*$_{7}$ *me at*$_{7}$ *my ease*$_{7}$; *tells*$_{8}$ *me BS to*$_{8}$ *my face*$_{8}$

[7] Our operational definition of *MWE type* combines a strong or weak designation with an ordered sequence of lemmas, using the WordNet API in NLTK (Bird et al., 2009) for lemmatization.

POS pattern	MWEs *contig.*	*gappy*	most frequent types (lowercased lemmas) and their counts
N_N	331	1	**customer service: 31** oil change: 9 wait staff: 5 garage door: 4
^_^	325	1	santa fe: 4 dr. shady: 4
V_P	217	44	**work with: 27** **deal with: 16** **look for: 12** **have to: 12** ask for: 8
V_T	149	42	**pick up: 15** **check out: 10** show up: 9 end up: 6 give up: 5
V_N	31	107	take time: 7 give chance: 5 waste time: 5 have experience: 5
A_N	133	3	front desk: 6 top notch: 6 last minute: 5
V_R	103	30	**come in: 12** come out: 8 take in: 7 stop in: 6 call back: 5
D_N	83	1	**a lot: 30** **a bit: 13** a couple: 9
P_N	67	8	**on time: 10** in town: 9 in fact: 7
R_R	72	1	**at least: 10** at best: 7 as well: 6 of course: 5 at all: 5
V_D_N	46	21	**take the time: 11** do a job: 8
V~N	7	56	*do job: 9* *waste time: 4*
^_^_^	63		home delivery service: 3 lake forest tots: 3
R~V	49		***highly recommend: 43*** *well spend: 1* *pleasantly surprise: 1*
P_D_N	33	6	over the phone: 4 on the side: 3 at this point: 2 on a budget: 2
A_P	39		pleased with: 7 happy with: 6 interested in: 5
P_P	39		**out of: 10** due to: 9 because of: 7
V_O	38		**thank you: 26** get it: 2 trust me: 2
V_V	8	30	get do: 8 let know: 5 have do: 4
N~N	34	1	*channel guide: 2* *drug seeker: 2* *room key: 1* *bus route: 1*
A~N	31		*hidden gem: 3* *great job: 2* *physical address: 2* *many thanks: 2* *great guy: 1*
V_N_P	16	15	**take care of: 14** have problem with: 5
N_V	18	10	mind blow: 2 test drive: 2 home make: 2
^_$	28		bj s: 2 fraiser 's: 2 ham s: 2 alan 's: 2 max 's: 2
D_A	28		**a few: 13** **a little: 11**
R_P	25	1	all over: 3 even though: 3 instead of: 2 even if: 2
V_A	19	6	make sure: 7 get busy: 3 get healthy: 2 play dumb: 1
V_P_N	14	6	go to school: 2 put at ease: 2 be in hands: 2 keep in mind: 1
#_N	20		5 star: 9 2 star: 2 800 number: 1 one bit: 1 ten star: 1 360 restraunt: 1
N_A	18		year old: 9 month old: 3 years old: 2 cost effective: 1 lightning fast: 1
V~R	11	6	*stay away: 3* *go in: 2* *bring back: 2* *recommend highly: 2* *work hard: 1*
N_P_N	14	2	chest of drawers: 2 man of word: 1 bang for buck: 1 sister in law: 1
N~V	6	10	*job do: 2* *work do: 2* *picture take: 1* *care receive: 1* *operation run: 1*
R_V	15	1	well do: 4 never mind: 2 better believe: 1 well know: 1
N_R	15		night out: 3 hands down: 3 thanks again: 3
N_-_N	14		a / c: 2 jiu - jitsu: 2
P~D~N	14		*in the world: 3* *around the corner: 2* *for some reason: 2*
V_R_P	12	1	look forward to: 3 talk down to: 2 have yet to: 1 be there for: 1
A_A	13		west indian: 3 old fashioned: 1 up front: 1 spot on: 1 tip top: 1 dead on: 1
V_T_P	11	2	watch out for: 2 make up for: 2 put up with: 2 turn over to: 1
P_P_N	10	2	out of business: 3 out of town: 2 out of date: 1
N_P	12		nothing but: 2 increase in: 1 damage to: 1
P_N_P	11		in front of: 3 on top of: 2 in need of: 1 in spite of: 1 in search of: 1
A_N_N	11		criminal defense lawyer: 2 purple hull pea: 1 social security numbers: 1
N_N_N	11		search engine optimization: 2 kung pao chicken: 1
N_&_N	10		spay and neuter: 2 give and take: 1 bar and grill: 1 hit or miss: 1
G_A	10		over priced: 4 over cooked: 1 miss informed: 1 out standing: 1
^_^_^_^	10		bexar county tax office: 1 anna maria jose mudo: 1
P_R	10		by far: 8 if ever: 1 of late: 1

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 multiword 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.[8]

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

[8] 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.

Topical category	# docs
Food/restaurant	207
Retail	115
Home services	74
Automotive	73
Medical/dental	52
Entertainment/recreation	45
Travel	44
Health/beauty	30
Pet	16
Other	65
Unsure	2

Perceived sentiment	# docs
++ strongly positive	310
+ positive	214
− negative	88
−− strongly negative	111

Table 3: Distribution of review topics and sentiment as coded by one of the annotators.

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 ARK's group at CMU, and anonymous reviewers for valuable feedback.

References

Abeillé, Anne, Clément, Lionel, and Toussenel, François (2003). Building a treebank for French. In Abeillé, Anne and Ide, Nancy, editors, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 165–187. Springer Netherlands.

Baldwin, Timothy (2005). Looking for prepositional verbs in corpus data. In *Proc. of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 115–126. Colchester, UK.

Baldwin, Timothy (2008). A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In *Proc. of MWE*, pages 1–2. Marrakech, Morocco.

Baldwin, Timothy and Kim, Su Nam (2010). Multiword expressions. In Indurkhya, Nitin and Damerau, Fred J., editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

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

Bird, Steven, Klein, Ewan, and Loper, Edward (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., Sebastopol, CA.

Carpuat, Marine and Diab, Mona (2010). Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proc. of NAACL-HLT*, pages 242–245. Los Angeles, California.

Ciaramita, Massimiliano and Altun, Yasemin (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, pages 594–602. Sydney, Australia.

Constant, Matthieu and Sigogne, Anthony (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proc. of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 49–56. Portland, Oregon, USA.

Constant, Matthieu, Sigogne, Anthony, and Watrin, Patrick (2012). Discriminative strategies to integrate multiword expression recognition and parsing. In *Proc. of ACL*, pages 204–212. Jeju Island, Korea.

Cook, Paul, Fazly, Afsaneh, and Stevenson, Suzanne (2008). The VNC-Tokens dataset. In *Proc. of MWE*, pages 19–22. Marrakech, Morocco.

Ellis, Nick C., Simpson-Vlach, Rita, and Maynard, Carson (2008). Formulaic language in native and second language speakers: psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 42(3):375–396.

Fellbaum, Christiane, editor (1998). *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.

Green, Spence, de Marneffe, Marie-Catherine, Bauer, John, and Manning, Christopher D. (2011). Multiword expression identification with tree substitution grammars: a parsing tour de force with French. In *Proc. of EMNLP*, pages 725–735. Edinburgh, Scotland, UK.

Green, Spence, de Marneffe, Marie-Catherine, and Manning, Christopher D. (2012). Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227.

Hartmann, Silvana, György Szarvas, and Iryna Gurevych (2012). Mining multiword terms from Wikipedia. In Pazienza, Maria Teresa and Stellato, Armando, editors, *Semi-Automatic Ontology Development*. IGI Global.

Hermann, Karl Moritz, Blunsom, Phil, and Pulman, Stephen (2012). An unsupervised ranking model for noun-noun compositionality. In *Proc. of *SEM*, pages 132–141. Montréal, Canada.

Kuiper, Koenraad, McCann, Heather, Quinn, Heidi, Aitchison, Therese, and van der Veer, Kees (2003). SAID. Technical Report LDC2003T10, Linguistic Data Consortium,

Philadelphia, PA.

Miller, George A., Leacock, Claudia, Tengi, Randee, and Bunker, Ross T. (1993). A semantic concordance. In *Proc. of HLT*, pages 303–308. Plainsboro, NJ, USA.

Moon, Rosamund (1998). *Fixed expressions and idioms in English: a corpus-based approach*. Oxford Studies in Lexicography and Lexicology. Clarendon Press, Oxford, UK.

Newman, David, Koilada, Nagendra, Lau, Jey Han, and Baldwin, Timothy (2012). Bayesian text segmentation for index term identification and keyphrase extraction. In *Proc. of COLING 2012*, pages 2077–2092. Mumbai, India.

Nunberg, Geoffrey, Sag, Ivan A., and Wasow, Thomas (1994). Idioms. *Language*, 70(3):491–538.

Paaß, Gerhard and Reichartz, Frank (2009). Exploiting semantic constraints for estimating supersenses with CRFs. In *Proc. of the Ninth SIAM International Conference on Data Mining*, pages 485–496. Sparks, Nevada.

Schneider, Nathan, Danchik, Emily, Dyer, Chris, and Smith, Noah A. (2014). Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*. To appear.

Simpson, Rita and Mendis, Dushyanthi (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly*, 37(3):419–441.

Tu, Yuancheng and Roth, Dan (2011). Learning English light verb constructions: contextual or statistical. In *Proc. of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 31–39. Portland, Oregon, USA.

Tu, Yuancheng and Roth, Dan (2012). Sorting out the most confusing English phrasal verbs. In *Proc. of *SEM*, pages 65–69. Montréal, Canada.

Čmejrek, Martin, Cuřín, Jan, Hajič, Jan, and Havelka, Jiří (2005). Prague Czech-English Dependency Treebank: resource for structure-based MT. In *Proc. of EAMT*, pages 73–78. Budapest, Hungary.

Vilain, Marc, Burger, John, Aberdeen, John, Connolly, Dennis, and Hirschman, Lynette (1995). A model-theoretic coreference scoring scheme. In *Proc. of MUC-6*, pages 45–52. Columbia, Maryland.

Vincze, Veronika (2012). Light verb constructions in the SzegedParalellFX English-Hungarian parallel corpus. In Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Doğan, Mehmet Uğur, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, and Piperidis, Stelios, editors, *Proc. of LREC*. Istanbul, Turkey.

Vincze, Veronika, Nagy T., István, and Zsibrita, János (2013). Learning to detect English and Hungarian light verb constructions. *ACM Transactions on Speech and Language Processing*, 10(2):6:1–6:25.

Wray, Alison (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, 21(4):463–489.

Wulff, Stefanie (2008). *Rethinking idiomaticity: a usage-based approach*. Research in Corpus and Discourse. Continuum International Publishing Group, London.