

An Analysis of Multiword Expressions in the Paraphrase Database

First Author

Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author

Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

We investigate whether paraphrases might be a useful resource for understanding multiword expressions (MWEs). In particular, we analyze the paraphrases in PPDB, the Paraphrase Database, where multiple words are re-written as a single word. By automatically mapping from multiword expressions onto single words, NLP systems could potentially process the re-written text more easily than the original text containing MWEs. We use the MWE categorization system as described in Sag et al. (2001). Although a small proportion of the many-to-one paraphrases in PPDB are classic MWEs, the resource contains millions of entries. We therefore train a classifier to distinguish interesting MWEs from other sorts of many-to-one paraphrases.

1 Introduction

Multiword expressions (MWEs) are phrases whose meanings are different than the literal interpretation of the words in the phrase. MWEs include verb-particle constructions, fixed expressions, compound nominals, and decomposable idioms, to name a few (Sag et al., 2001). MWEs are difficult to process both for non-native speakers of English, as well as for NLP systems. Studies using an eye-movement paradigm have found that non-native speakers of English required more time to retrieve figurative senses of phrases than literal ones, whereas native speakers retrieved the idiomatic meaning faster than the literal meaning (Siyanova-Chanturia and Martinez, 2014). These studies imply that L2 speakers of English may find it more difficult to understand MWEs than a similar phrase whose meaning was literal. Among

NLP systems, both parsers and information retrieval systems make errors on MWEs. As described by Villavicencio et al. (2007), Grammar et al. (2004) found that, among a random sample of 20,000 strings from the written portion of the British National Corpus (Burnard, 2000), using the English Resource Grammar (Copestake and Flickinger, 2000), MWEs caused 8% of all parse errors. When manually selected compound nominals were searched for as single terms it improved information retrieval results (Acosta et al., 2011).

Because MWEs are challenging for many NLP tasks, automatically identifying them could be useful for identifying and averting errors. Several research efforts have examined this topic. For example, Muzny and Zettlemoyer (2013) built a classifier to identify idioms, and Li and Sporleder (2010) use Gaussian Mixture Models to identify new idioms. In this paper, we use the Paraphrase Database (PPDB) as a resource to define a MWE lexicon, which could be incorporated into other NLP systems. In addition to being a potentially useful resource for *identifying* MWEs, it has the unique feature of potentially giving an *interpretation* of the MWEs by replacing them with a one word paraphrase.

In this paper we

- Analyze the PPDB for the prevalence of various types of MWEs
- Build a classifier that distinguishes interesting MWEs in the PPDB from more generic paraphrases
- Investigate whether parse quality can be improved by substituting MWEs with one word paraphrases

2 The Paraphrase Database

The Paraphrase Database (PPDB) is a database containing English paraphrases. PPDB was de-

veloped using alignment techniques from machine translation on bilingual parallel corpora, pivoting on a foreign language, to find English phrases that translate to the same foreign-language phrase Bannard and Callison-Burch (2005). The database takes into account syntactic information of both the English and foreign-language phrases: the entries in PPDB were found using SCFGs to come up with paraphrases that form constituents of the same syntactic category (Ganitkevitch et al., 2011; Ganitkevitch et al., 2013).

3 Related Work

Various definitions of MWEs have been used for NLP tasks. Sag et al. define a taxonomy of multiword expressions that is widely used in computational MWE research. They define broad categories for MWEs (fixed expressions, semi-fixed expressions, syntactically flexible expressions, institutionalized phrases), and more specific categories within each of these. The following is a brief outline of the categories defined by Sag et al. that we investigate in this paper:

- Fixed expressions. These are expressions that don't fit standard grammar rules and are not compositional.
- Non-decomposable idioms. These are expressions that follow grammatical rules, but whose meaning is not compositional.
- Compound nominal. Noun combinations whose meaning is different than the individual nouns in the phrase.
- Proper nouns. These are included in the taxonomy because names allow for some kinds of variation but not others (e.g., the San Francisco 49ers, the 49ers, 49ers, but not the Oakland 49ers).
- Decomposable idioms.
- Verb-particle constructions. These are verb phrases formed of a verb followed by a particle, where the meaning of the phrase is different than that of the verb alone.
- Light verbs. These are verb phrases formed of a light verb (make, give, do), followed by a noun phrase.

Other work has described a division of MWEs extracted from paraphrases: one-to-many, many-to-one, decomposable and non-decomposable (de Marneffe et al.,). They also investigated using paraphrases to automatically extract multiword expressions. (de Marneffe et al.,) consider dependency-based paraphrases and parallel corpora-based paraphrases as sources for MWEs, finding that 34% of the MWEs in the MSR manual word alignments of the RTE 2006 corpus were present in these resources (Brockett, 2007).

4 Experimental Design

The Paraphrases Database (PPDB) contains English paraphrases. We have characterized a subset of the paraphrases found in the PPDB, according to categories of multi-word expressions (MWEs), syntactic changes in the expansion from a word to its paraphrase, and what parts of speech appear in the corpus. We also looked at how many of the paraphrases in the PPDB appear to be spurious.

The categories of MWEs we looked at were light verbs, verb-particle constructions, negation, and superlatives. We also included the categories for MWEs described in Sag et al: fixed expressions, non-decomposable idioms, compound nominals, proper names, and decomposable idioms.

In addition to MWE categories, we also included categories for syntactic changes from a word to its paraphrase: change of tense followed by a paraphrase, nominalizations, infinitival to, adverbial modifier, one or more words the same as part of the original word, determiner followed by a one-word paraphrase, determiner followed by the plural form, and change of tense. Finally, we included acronyms, hypernym-hyponym pairs, times, extra punctuation marks, and numbers as categories, as well as unspecified expansions and bad paraphrases.

5 Results

Of a random sample of 500 paraphrases from the L one-to-many paraphrase file, the most common types of paraphrase were expansions using the same morphological form (117 instances, or 23.4%), determiner followed by a one-word paraphrase (86 instances, or 17.2%), and paraphrases that did not fall into a particular category (62 instances, or 12.4%). Of the sample, 43 were bad paraphrases (8.6%). The full list of categories and the number of instances in each are in the table

below.

Of the 500 paraphrases in the random sample, 37 (7.4%) fell into the MWE categories defined by Baldwin et al. Of these MWE paraphrases, the most common were verb-particle constructions (14 instances, or 37.8%) , followed by proper names (7, or 18.9%). There were no instances of compound nominals in the sample.

The full list of categories with the number of instances of paraphrases in each, as well as illustrative examples for each, are in the table below (Table 1). MWEs, as defined by Sag et al. (2001), are marked in bold.

Paraphrase Category	Number of instances in sample	Examples
determiner + one-word paraphrase	86	photographs, the images; duties, the responsibilities
expansion	62	WHADVP suffocate, need air; cumbersome, time consuming
expansion, same morphological form	47	OP fun, sounds like fun; signs, signs and signals
change of tense + paraphrase	43	EST initiated, has undertaken; say, going to tell
inaccurate/bad paraphrase	43	SBARQ iron a par; also, do i
implicit/most common modifier	41	CONJP Baghdad, iraqi capital baghdad; PR1 enrichment, uranium enrichment
implicit type	29	training, training course; customs, customs offices
adverbial modifier	27	interesting, very interesting; more, even more
acronym	21	gpa, the global programme of action; cras, in the entire PPDB E file, sorted by frequency in PPDB L
one or more words	17	anytime, any point; enslavement, slave labour
the same as part of original word	14	In both samples, the most common part of speech is NP, followed by VP.
verb-particle	9	gloves, the glove; militaries, the military
determiner + plural	7	In addition to categorizing a random sample of paraphrases, we searched for instances of light
proper noun	7	markov, mr markov; karadzic, radovan karadzic
change of tense	7	changing, be changed; attain, be attained
fixed expression	6	verbs, verb-particle constructions, negation, comparatives, and superlatives. The light verbs were
superlative	5	applied, put into effect; plenty, a whole host
copula	5	notably, most particularly; best-known, most famous
decomposable idiom	5	those with the verb have, take, make, hold or give, qualify, are eligible; reason, been right
light verb	4	followed by a noun phrase. The verb-particle constructions were any verbs followed by the particles
number	4	issued, made available; place, make way
infinitival to	4	20, twenty of; 5,000, 5 000.
nominalization	3	track, to follow; answer, to reply
negation	3	the expanded paraphrase. Comparatives had the operation, proper functioning
time variation	2	word more, and superlatives had the word most.
punctuation	2	unused, not utilized; non-parties, not parties
non-decomposable idiom	2	To find instances of verb-particle constructions, we used the Linux command 'grep' and the regular
alternate spelling	2	debt-servicing, debt servicing; what, something
compound nominal	2	expression 'particle' to find phrases where the potential particle was not the first word, and to ensure
	1	al-najaf, al nagaf
	0	that it was not a substring of another word (e.g., 'onto' instead of 'on'). For example, to find

Table 1: Number of instances from each category in random sample of 500.

The distribution of the syntactic categories from this random sample is depicted in Table 2.

Syntactic category	Frequency in sample	Frequency in PP
NP	254	101563
VP	109	39380
X	39	17005
ADJP	27	14513
ADVP	16	5451
S	9	4103
INTJ	3	1730
TOP	4	1476
NX	2	1113
PP	1	532
FRAG	3	388
Examples	0	305
SQ	0	165
SBAR	0	163
WHNP	0	86
WHADVP	0	39
OP	0	3
EST	0	3
SBARQ	0	1
CONJP	0	1
PR1	0	

Table 2: Distribution of syntactic categories in random sample of 500 words from PPDB L, and in the entire PPDB E file, sorted by frequency in PPDB L.

In both samples, the most common part of speech is NP, followed by VP.

In addition to categorizing a random sample of paraphrases, we searched for instances of light verbs, verb-particle constructions, negation, comparatives, and superlatives. The light verbs were notably, most particularly; best-known, most famous those with the verb have, take, make, hold or give, qualify, are eligible; reason, been right followed by a noun phrase. The verb-particle constructions were any verbs followed by the particles issued, made available; place, make way 20, twenty of; 5,000, 5 000. Negation instances had the word not either in the original or the expanded paraphrase. Comparatives had the operation, proper functioning word more, and superlatives had the word most. unused, not utilized; non-parties, not parties To find instances of verb-particle constructions, we used the Linux command 'grep' and the regular expression 'particle' to find phrases where the potential particle was not the first word, and to ensure that it was not a substring of another word (e.g., 'onto' instead of 'on'). For example, to find potential instances of verb-particle constructions with the particle 'up', we ran the following command: grep ' up' ppdb-1.0-l-o2m

We then determined manually whether the phrases found were verb-particle constructions. The potential particle was sometimes a preposition or adverb, in which case it did not fit into this MWE category.

We used similar commands to find instances of the other paraphrase categories. A list of the commands used is in the following table:

Paraphrase category	Search command
verb-particle construction	grep 'up \b' ppdb.filename
light verbs	grep '\b make ' ppdb.filename
negation	grep '\b not ' ppdb.filename
comparatives	grep '\b more ' ppdb.filename
superlatives	grep '\b most ' ppdb.filename

Table 3: Linux commands to search for paraphrases of different types.

The results from these searches are summarized in the tables below. We considered only light verb phrases with no extra terms (e.g., 'have a word', but not 'have a word with you' or 'can i have a word').

Light verb	Number of light verb phrases found by grep	Number of light verb phrases
make	46	35
have	105	45
give	11	4
take	48	33
hold	3	1

Table 4: Light verbs: number of potential light verb phrases found by grep, the number of these that turned out to be light-verb phrases, and some examples.

After manually identifying categories, we developed a classifier to automatically find paraphrases that are interesting to MWE researchers. We built the classifier to find the following categories of MWE: verb-particle construction; expansion, same morphological form; implicit/most common modifier; implicit type; light verb; fixed expression; non-decomposable idiom; proper noun; decomposable idiom; light verb.

The features we used were whether a preposition appeared in the expansion of the paraphrase (feature value either 1 or 0), whether the original word appeared in the expansion, whether a

Particle	Number of verb-particle constructions found	Number of verb-particle phrases found by grep
up	439	439
about	10	115
around	44	47
back	145	156
down	190	195
in	32	207
off	134	134
on	109	196
over	383	383
under	65	67

Table 5: Verb-particle constructions.

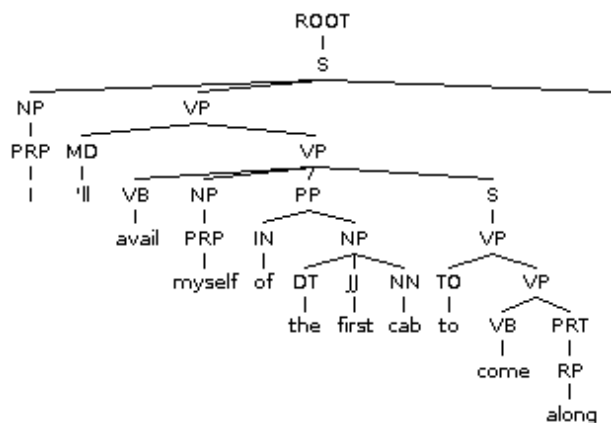
light verb appeared in the expansion, and all of the scores from the PPDB (every number that appeared in the entry following an equal sign).

The prepositions and light verbs used for the features are listed in the table below.

prepositions	'about', 'around', 'back', 'down', 'in', 'off', 'on'
light verbs	'give', 'have', 'h'

Matching Words
used as features for classifier.
We used an SVM classifier (libsvm) with these features on a sample of 500 words. The accuracy of the classifier was 76.4%. For comparison, a baseline classifier that returns the majority class has an accuracy of 6.2%.

To determine whether a lexicon of paraphrases extracted from the PPDB would be useful for the task of parsing, we took a sample of 33 paraphrases that fall into the categories described by Sag et al. (2001) (verb particle, fixed expression, non-decomposable idiom, compound nominal, proper name, decomposable idiom, and light verb) from the random sample of 500 words. We then did a Google search to find a sentence containing the expansion, selecting sentences where the part of speech matched that of the paraphrase, by manual inspection. We restricted the sentences to those less than 50 words long. We then created an equivalent sentence by replacing the expansion with the one-word paraphrase. We used the online demo of the Berkeley parser (Petrov et al., 2006) to parse these sentence pairs, and manually evaluated the resulting parse trees for each sentence of the pair.



(a) Original sentence: I'll avail myself of the first cab to come along.

Figure 1: Parse trees for a sentence pair where substituting the single-word paraphrase for the MWE (in figure (b)) yields a better parse than the original sentence (in figure (a)).

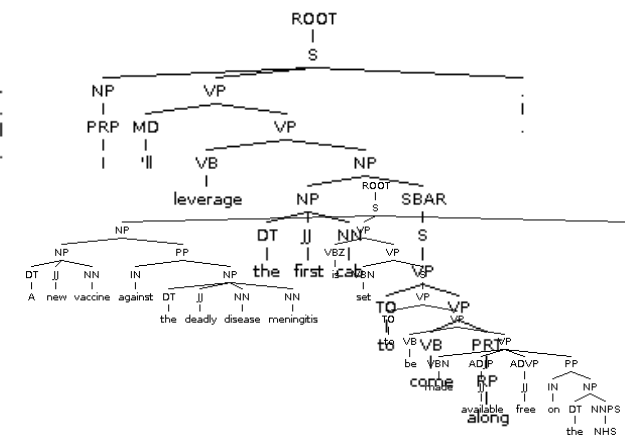
Of the 33 paraphrases sampled, 30 were correctly parsed using the one-word paraphrase, and 29 were correctly parsed using the expansion. 2 were parsed correctly for the one-word paraphrase but not for the expansion, 1 was parsed correctly for the expansion but not the one-word paraphrase, and 1 was parsed incorrectly for both. Figure 1 shows the trees for a case where the sentence containing the one-word paraphrase ("leverage") is parsed correctly, but the sentence containing the expansion ("avail myself of") is not. Figure 1 shows the trees for a case where the sentence containing the expansion ("leverage") is parsed correctly, but the sentence containing the one-word paraphrase ("avail myself of") is not. Figure 2 shows the trees for a case where the sentence containing the one-word paraphrase ("issued") is parsed correctly, but the sentence containing the expansion ("made available") is not. Figure 3 shows the trees for a case where both sentences are parsed incorrectly.

Bibliography

References

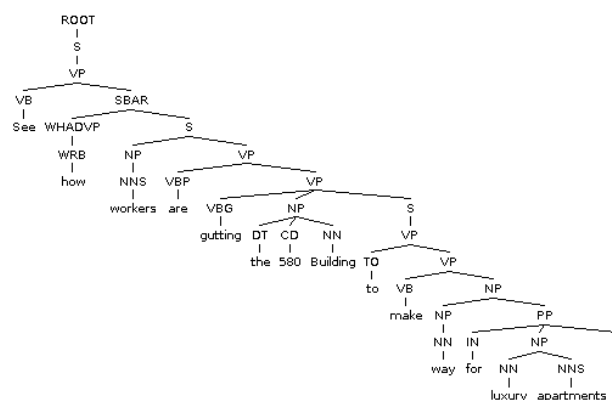
Octavio Acosta, Aline Villavicencio, and Viviane Moreira. 2011. Identification and treatment of multi-word expressions applied to information retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 101–109, Portland, Oregon, USA, June. Association for Computational Linguistics.

Colin Bannard and Chris Callison-Burch. 2005. Para-



(b) Sentence with substitution: I'll leverage the first cab to come along.

Figure 2: Parse trees for a sentence pair where the original sentence (in figure (a)) yields a better parse than substituting the single-word paraphrase for the MWE (in figure (b)).



(a) Original sentence: See how workers are gutting the 580 Building to make way for luxury apartments.

(b) Sentence with substitution: See how workers are gutting the 580 Building to make way for luxury apartments.

Figure 3: Parse trees for a sentence pair where the original sentence (in figure (a)) yields a better parse than substituting the single-word paraphrase for the MWE (in figure (a)).

- phrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- Lou Burnard. 2000. User reference guide for the british national corpus, technical report. Oxford University Computing Services.
- Ann Copestake and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the second international conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece.
- Marie-Catherine de Marneffe, Sebastian Pado, and Christopher D. Manning.
- Juri Ganitkevitch, Chris Callison-Burch, Courtney Napoles, and Benjamin Van Durme. 2011. Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1168–1179, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.
- Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.
- Broad-Coverage Precision Grammar, Timothy Baldwin, John Beavers, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Beauty and the beast: What running a.
- Linlin Li and Caroline Sporleder. 2010. Using gaussian mixture models to detect figurative language in context. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 297–300, Los Angeles, California, June. Association for Computational Linguistics.
- Grace Muzny and Luke Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, Washington, USA, October. Association for Computational Linguistics.
- Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia, July. Association for Computational Linguistics.
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for nlp. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15.
- Anna Siyanova-Chanturia and Ron Martinez. 2014. The idiom principle revisited. In *Applied Linguistics*, January.
- Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1034–1043.