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 hypothesize that paraphrases may be a useful resource for understanding multiword expressions (MWEs). 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) to identify paraphrases in PPDB that might be interesting to MWE researchers. Although only a relatively small proportion of the many-to-one paraphrases in PPDB are classic MWEs, the resource contains millions of entries. We train a classifier to distinguish the interesting MWEs from other sorts of many-to-one paraphrases. We do a pilot study on parsing paraphrased sentences, to determine whether sentences with a MWE replaced by its paraphrase are parsed correctly more often than the original.

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. Baldwin 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, Li and Sporleder (2010) and Muzny and Zettlemoyer (2013) built a classifiers to identify 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

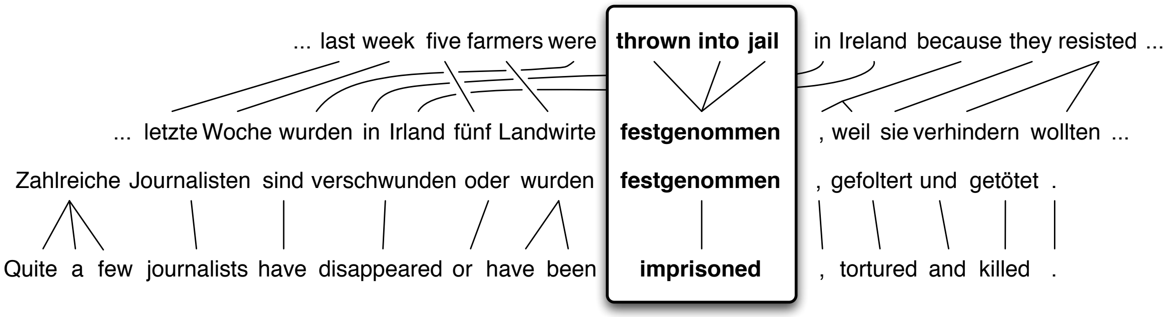


Figure 1: The paraphrase database contains one-to-many paraphrase where a single English word is paraphrased as a multiword phrase. The paraphrase of *imprisoned* as *thrown into jail* results because the shared foreign phrase *festgenommen* is sometimes translated using one word and sometimes as a phrase. We analyze PPDB's one-to-many paraphrases for their prevalence of more interesting MWEs.

2 The Paraphrase Database

In this paper, we analyze paraphrases within the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013), currently the largest available collection of paraphrases. Compared to other paraphrase resources such as the DIRT database (12 million rules) (Lin and Pantel, 2001) and the MSR paraphrase phrase table (13 million) (Dolan et al., 2004), PPDB contains over 150 million paraphrase rules. These rules were extracted automatically using Bannard and Callison-Burch (2005)'s "bilingual pivoting" method, in which two English phrases are assumed to be paraphrases if they both translate to the same foreign phrase. This is illustrated in Figure 1.

3 Related Work

Various definitions of MWEs have been used for NLP tasks. Sag et al. (2001) 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 their specific categories. We use an expanded set of categories to analyze the MWEs found in PPDB in section 4.

- Fixed expressions. The meaning of fixed expressions is non-compositional, and they do not not follow standard grammar rules. An example is the phrase "all of a sudden," which means "suddenly." The word sudden cannot be replaced with another adjective: "all of a happy" is nonsensical. Other examples of fixed expressions are "in short" and "kingdom come".

- Non-decomposable idioms. These are non-compositional expressions that do follow grammatical rules. For example, the grammatical structure of "kick the bucket," is valid (e.g., there exist other acceptable English phrases with the same structure, like "kick the stone,"), however the idioms meaning is not sum of its parts. Other examples in- clude "to take the bull by the horns" and "to beat swords into plowshares." (Nunberg, Sag, Wasow).***

- Compound nominals. These are noun combinations whose meaning is not automatically inferable from the individual nouns in the phrase. Contrast "orange juice", "hand lotion" and "newspaper column." Without word knowledge it is not possible to know whether the nouns relationship is *is made from*, *is intended for*, or *is located in*.

- Proper nouns. These are nouns that are names for unique entities including places (California, the Bronx), people (President Roosevelt), and events (the Industrial Revolution). This category is included in the taxonomy because names allow for some kinds of variation but not others. For example, *the San Francisco 49ers, the 49ers,* and *49ers* are all valid names for the sports team, but "the Bay Area 49ers" is not.

- Decomposable idioms. These are phrases whose meaning is compensational from the meaning of the individual words, but which still carry an idiomatic meaning. One example is the idiom "play with fire" (Wikipedia)***. Decomposable idioms can undergo syntactic changes to varying degrees. For example, the phrase "let the cat out of the bag" can be modified to "the cat was out of the bag," but also to "the cat was *really* out of the bag," adding an adverb.

- 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, or up the verb and the particle combined. Some examples are "wash out," "break down," and "follow up."

- Light verbs. The combination of light verbs (make, give, take) followed by a normalized verb are idiosyncratic in that it is difficult to predict which light verb will be used with the original verb. "To walk" can be paraphrased as "to take a walk," but not "to make a walk," whereas "to make a presentation" is fine and "to take a presentation" nonsensical.

Another piece of related work examined the possibility of using paraphrases to interpret MWEs. de Marneffe et al. (2009) examined the use of paraphrases to better align text-hypothesis pairs in the task of Recognizing Textual Entailment (Dagan et al., 2006). They examined paraphrase resources derived from Lin and Pantel (2001)'s DIRT algorithm and from Bannard and Callison-Burch (2005)'s bilingual pivoting method. They found that about one third of the MWEs in the RTE data set were contained in the paraphrase resources, but that the MWEs were sometimes mis-paraphrased, making it unclear whether the resources would help with RTE-alignment. Like de Marneffe et al. (2009), we do an analysis of a paraphrase resource derived from Bannard and Callison-Burch (2005)'s bilingual pivoting method, but at a larger scale.

4 Analysis of PPDB

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.

While categorizing the paraphrases we found that paraphrases sometimes did not fall under the classic MWE categories, but they were valid paraphrases because of a change in syntax between the two phrases, or because of some other known relationship between paraphrases (e.g., acronyms). Therefore, in addition to the MWE categories, we also included the following categories for syntactic changes from a word to its paraphrase:

- Change of tense followed by a paraphrase. Paraphrases in this category both rephrased the original word and changed the tense of the phrase. An example is the pair "say", "going to tell".

- Nominalizations. Paraphrases in this category had one of the phrases as a noun form of a verb. Two examples are "a monopoly" and "monopolization", and "operation" and "proper functioning."

- Infinitival to. Paraphrases in this category had the infinitival form a verb as the expansion, and either the same verb or one with similar meaning as the one-word paraphrase. One example is "answer," "to reply."

- Adverbial modifier. Paraphrases in this category had either the original word or a word with similar meaning, plus an adverbial modifier. Two examples are "interesting", "very interesting," and "teeny", "really little."

- One or more words the same as part of the original word. This category captures paraphrases where some of the paraphrase is the same as some of the original word, but it is not simply the original word with additional modifiers. This category is meant to capture paraphrases that are lexicographically related to each other even if the original word is not a substring of the paraphrase. Some examples of this category are “anytime”, “any point,” and “vice-president,” “the deputy president.”

- Determiner followed by a one-word paraphrase. Paraphrases in this category map one word to a phrase consisting of a determiner followed by a one-word paraphrase of the original word. Some examples are “photographs,” “the images,” and “mist,” “the smoke.”

- Determiner followed by the singular form. This category captures paraphrases where the original word is a plural, and the expansion is a phrase consisting of a determiner followed by the singular form of the original word. Some examples are “sounds,” “the sound” and “loans,” “a loan.”

- Change of tense. Paraphrases in this category have the same verb in both the original word and its expansion, but the tense is changed. Some examples are “changing,” “be changed,” and “attain”, “be attained.”

The following non-syntactic categories were motivated by the examples in the random sample:

- Acronyms. This category captures paraphrases where one or both of the entries in a paraphrase pair are acronyms. Sometimes the acronym is expanded (e.g., “hm,” “her majesty”), and sometimes the expansion contains a different acronym (e.g., “cmf,” “the fcm”).

- Hypernym-hyponym pairs. This category was included to capture paraphrases where either the original word was a hypernym of the paraphrase, or vice versa. There were no instances in the random sample.

- Times. Paraphrases in this category were contained either times (e.g., “7:00,” “seven hours”) or dates (“2003/04,” “the 2003-04 fiscal year”) in either the original word or the paraphrase.

- Extra punctuation marks. This category captures paraphrases where either the original word or the expansion contains a punctuation mark (e.g., apostrophe or hyphen), which is what causes the word to be interpreted as a single word (e.g., “debt-servicing,” “debt servicing”) or the expansion to be interpreted as having multiple words (e.g., “what,” “somethin ’ ”).

- Numbers. Paraphrases in this category contain number in either the original word or its expansion. Some examples are “20,” “twenty of,” and “5,000,” “5 000.”

Finally, we defined categories to capture paraphrases that did not fall under the classic MWE, syntactic change, or non-syntactic categories.

- Unspecified expansions. These paraphrases do not fall under any of the other paraphrase categories defined until now.

- Bad paraphrases. These are paraphrases where the original word and the expansion do not seem to be related in any way, or where there is no context where the expansion could be replaced by the original word.

PPDB is released in a series of files. The different files are divided into one-to-one (synonyms), one-to-many, many-to-many, phrasal, and syntactic files. The lexical paraphrases are from one word to one word, the one-to-many are paraphrases between one word and a multi-word expression, many-to-many paraphrases are paraphrases between two multiword expressions, and syntactic paraphrases are paraphrases are those where the syntactic category is the same for both a phrase and its paraphrase. For this analysis we investigate the one-to-many paraphrase file as a source of potential paraphrases for multiword expressions.

The PPDB files are further subdivided by size, from S to XXXL (six sizes). Each paraphrase in the PPDB is scored according to how precise of a paraphrase it is likely to be. The smaller files contain better-scoring paraphrases, while the larger files contain incrementally more paraphrases, at the cost of precision. For this analysis we chose the L corpus as a compromise point among the various sizes for both coverage and quality of paraphrases.

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

Syntactic category	Freq in sample	Freq in PPDB-L
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
SQ	0	305
SBAR	0	165
WHNP	0	163
WHADVP	0	66
QP	0	39
LST	0	3
SBARQ	0	3
CONJP	0	1
PRT	0	1

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

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 Sag et al. (2001)l. 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.

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

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 those with the verb have, take, make, hold or give, followed by a noun phrase. The verb-particle constructions were any verbs followed by the particles down, up, on, out, over or upon. Negation instances had the word not either in the original or the expanded paraphrase. Comparatives had the word more, and superlatives had the word most.

To find instances of verb-particle constructions, we searched for 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 searched for the following regular expression:' up'_.

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 Table 3:

The results from these searches are summarized in the Tables 4 and 5. 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').

Among the 500 paraphrases in the sample, 37 fell into the category of interesting MWEs. Since the PPDB L one-to-many corpus contains 188,000 rules, based on our sample, we expect there to be around 14,000 relevant MWEs in the database.

5 Building a classifier for MWEs in PPDB

There are 188,000 rules in the PPDB L one-to-many file, which is too many to manually sift through and find relevant MWEs. Therefore, we built a classifier to automatically distinguish between trivial MWEs and paraphrases that are interesting to MWE researchers.

5.1 Experimental setup

We designed features tailored to find paraphrases in the following categories of MWE: verb-

Paraphrase Category	**Number of instances in sample**	**Examples**
determiner + one-word paraphrase	86	photographs, the images; duties, the responsibilities
expansion	62	suffocate, need air; cumbersome, time consuming
expansion, same morphological form	47	fun, sounds like fun; signs, signs and signals
change of tense + paraphrase	43	initiated, has undertaken; say, going to tell
inaccurate/bad paraphrase	43	iron, a par; also, do i
implicit/most common modifier	41	baghdad, iraqi capital baghdad; 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, credit-rating agencies
one or more words the same as part of original word	17	anytime, any point; enslavement, slave labour
verb-particle	14	torched, burnt down; done, carried out
determiner + plural	9	gloves, the glove; militaries, the military
proper noun	7	markov, mr markov; karadzic, radovan karadzic
change of tense	7	changing, be changed; attain, be attained
fixed expression	6	applied, put into effect; plenty, a whole host
superlative	5	notably, most particularly; best-known, most famous
copula	5	qualify, are eligible; reason, been right
decomposable idiom	4	nuts, out of your mind; sleeping, get a good night's sleep
light verb	4	issued, made available; place, make way
number	4	20, twenty of; 5,000, 5 000
infinitival to	3	track, to follow; answer, to reply
nominalization	3	operation, proper functioning
negation	2	unused, not utilized; non-parties, not parties
time variation	2	7:00, seven hours; 2003/04, the 2003-04 fiscal year
punctuation	2	debt-servicing, debt servicing; what, somethin '
non-decomposable idiom	2	furious, as mad as hell; entails, brings with it
alternate spelling	1	al-najaf, al nagaf
compound nominal	0	

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

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 methodology used in the classifier could also be applied more generally for finding relevant expressions in PPDB.

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 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 Table 6.

5.2 Results

We used an SVM classifier (libsvm) with these features on our sample of 500 words from the PPDB L corpus. With ten-fold cross-validation, the accuracy of the classifier was 76.4%. By comparison, a baseline classifier that returns the majority class has an accuracy of 69.2%.

6 Using paraphrased MWEs for an NLP task

6.1 Experimental design

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), as well as one paraphrase from each of the other categories 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.

6.2 Results

Of the 33 paraphrases sampled from MWE categories, 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 2 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 2 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 3 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 4 shows the trees for a case where both sentences are parsed incorrectly.

Of the 18 other paraphrases sampled, all were parsed correctly in their original form, while 15 were parsed correctly using the one-word paraphrase.

Bibliography

References

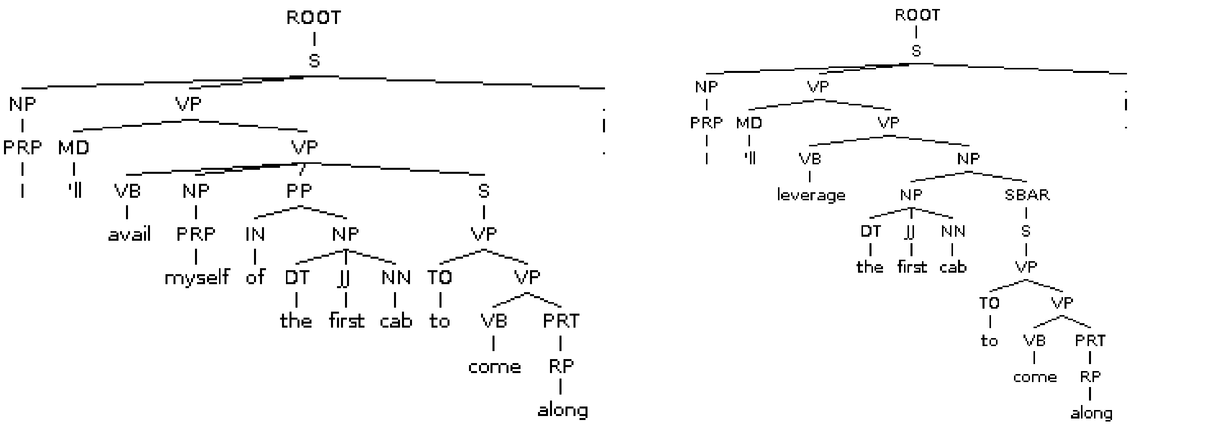
Otavio Acosta, Aline Villavicencio, and Viviane Moreira. 2011. Identification and treatment of multiword expressions applied to information retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 101–109, Portland, Oregon, USA, June.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

Timothy Baldwin, John Beavers, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Beauty and the beast: What running a broad-coverage precision grammar over the bnc taught us about the grammar – and the corpus.

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

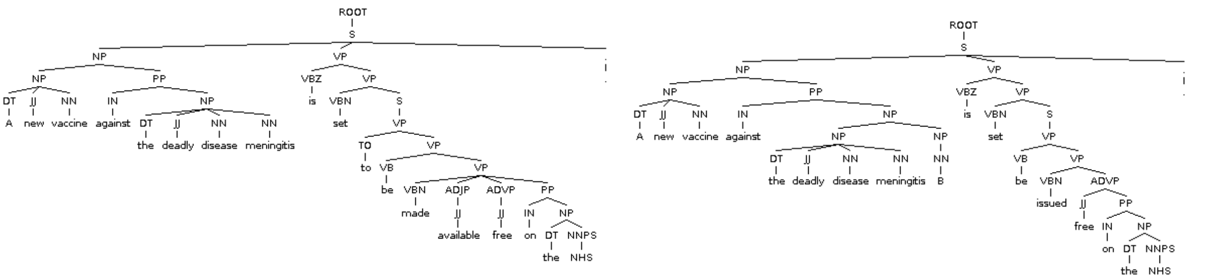
Ann Copestake and Dan Flickinger. 2000. An open-source grammar development environment



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

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

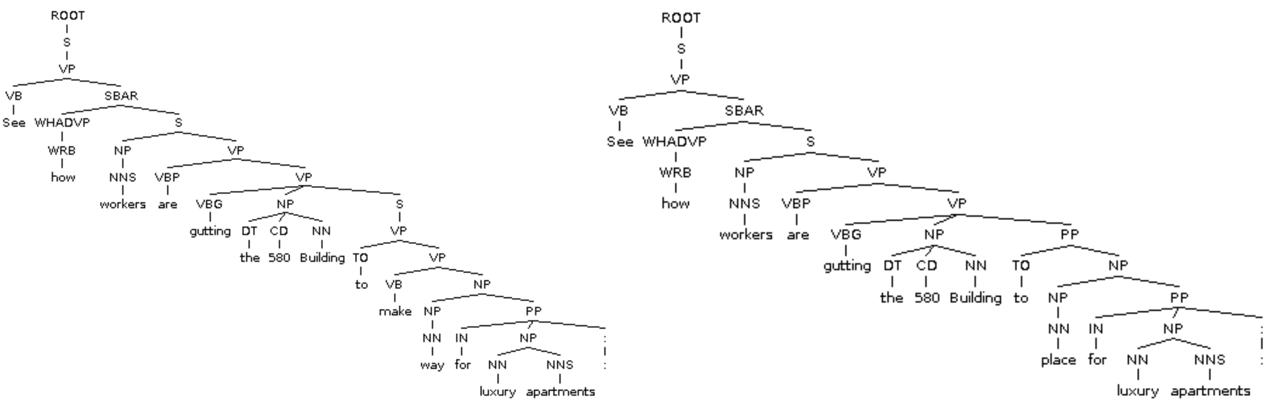
Figure 2: Parse trees for a sentence pair where substituting the single-word paraphrase for the MWE (in figure (b)) yields a better parse than the original sentence (in figure (a)).



(a) Original sentence: A new vaccine against the deadly disease meningitis is set to be made available free on the NHS.

(b) Sentence with substitution: A new vaccine against the deadly disease meningitis is set to be issued free on the NHS.

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 (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 place for luxury apartments.

Figure 4: 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)).

and broad-coverage English grammar using HPSG. In *Proceedings of the second international conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190. Springer.

Marie-Catherine de Marneffe, Sebastian Pado, and Christopher D. Manning. 2009. Multi-word expressions in textual inference: Much ado about nothing? In *Proceedings of the 2009 Workshop on Applied Textual Inference*, pages 1–9, Suntec, Singapore, August. Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350. 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.

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.

Dekang Lin and Patrick Pantel. 2001. DIRT – Discovery of Inference Rules from Text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328. ACM.

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.

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.

Light verb	**Number of light verb phrases found by grep**	**Number of light verb phrases**	**Matching verbs**
make	46	35	make a call, make a decision, make a suggestion
have	105	45	have a ball, have a conversation, have a word
give	11	4	give a damn, give a reply
take	48	33	take a break, take a look, take a walk
hold	3	1	hold a debate

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.

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
out	383	383
over	65	67

Table 5: Verb-particle constructions.

prepositions	'about', 'around', 'back', 'down', 'in', 'off', 'on', 'out', 'over', 'up'
light verbs	'give', 'have', 'hold', 'make', 'take'

Table 6: Words used as features for classifier.