An Analysis of Multiword Expressions in PPDB: 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

In this paper we present a feasibility study for rewriting multiword expressions as single words, which NLP systems could potentially process more easily than the original phrases. Here we investigate PPDB: The Paraphrase Database to get a mapping from multiword expressions onto single words, using the MWE categorization system as described in Baldwin, et al.

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, 2002).

MWEs pose difficulties 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), Baldwin et al. (2004) found that, among a random sample of 20,000 strings from the written portion of the British National Corpus (BNC: Burnard, 2000), using the English Resource Grammar (ERG: Copestake and Flickinger, 2000), MWEs caused 8

Based on this information, it seems that identifying MWEs could be useful for NLP tasks. 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.

The Paraphrase Database (PPDB) is a database containing English paraphrases. PPDB was developed 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 (Ganitkevich et al, 2011; Ganitkevich et al, 2013).

2 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 Tim Baldwin's categories for MWEs: 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.

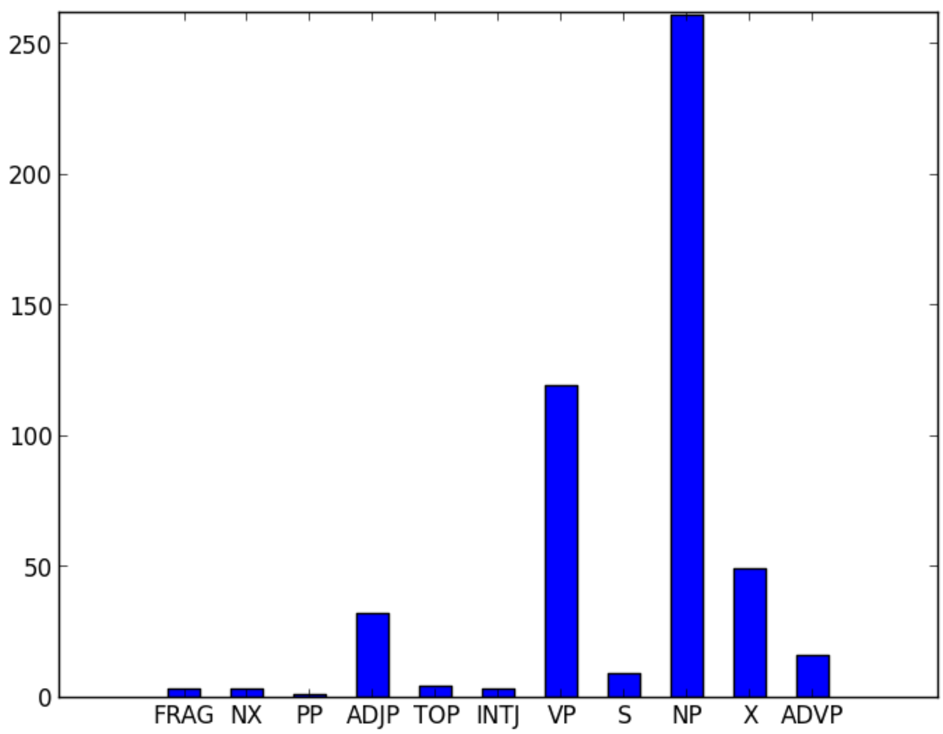
3 Results

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 Baldwin et al., are marked in bold.

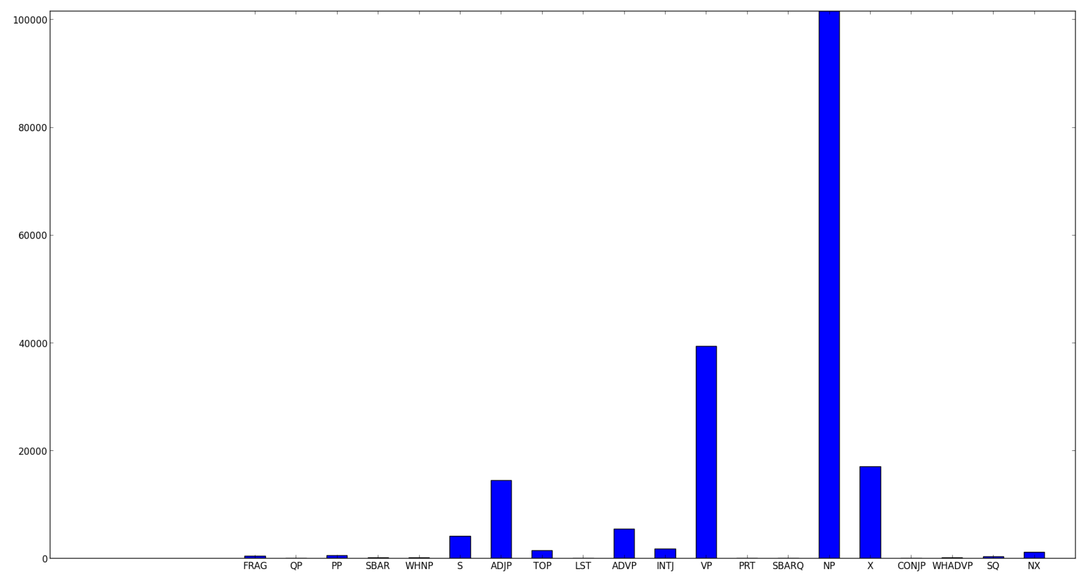
The distribution of the parts of speech from this random sample is depicted in the histogram below.



The distribution of all of the parts of the speech from the L one-to-many paraphrase file is depicted in the following histogram:

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.



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

Table 3: Light verbs.

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 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.

We used an SVM classifier (libsvm) with these features on our sample of 500 words. With ten-fold cross-validation, the accuracy of the classifier was 75.4

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 4: Verb-particle constructions.

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

Table 5: Words used as features for classifier.

4 Analysis

Acknowledgements

The acknowledgements should go immediately before the references. Do not number the acknowledgements section. Do not include this section when submitting your paper for review.

References

....

Yutaro Shigeto, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kose, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, and Yuji Matsumoto. 2013 Construction of English MWE Dictionary andits Application to POS Tagging *Proceedings of the 9th Workshop on Multiword Expressions*,139–144.

Matthieu Constant, Anthony Sigogne, and Patrick Watrin. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. *ACL '12 Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, 204-212.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.