# Instructions for COLING-2014 Proceedings

**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 Penn Paraphrases Database (PPDB) as a resource to define a MWE lexicon, which could be incorporated into other NLP systems.

The Penn Paraphrases 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 Penn 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.

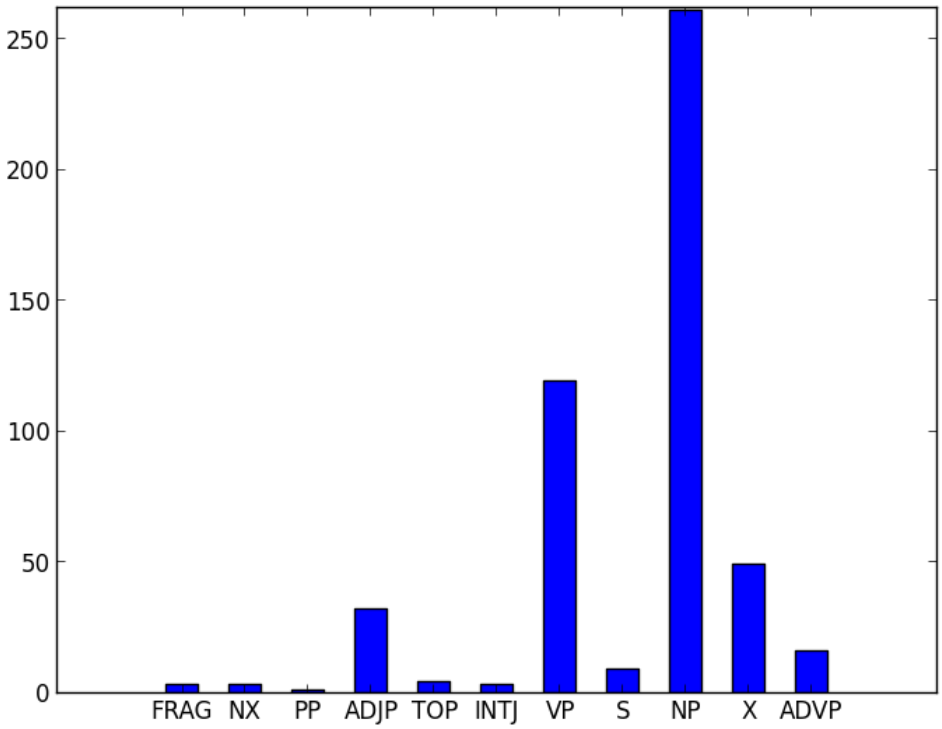
Of the 500 paraphrases in the random sample, 37 (7.4%) fell into the MWE categories defined by Baldwin et al. Of these MWE paraphrases, the most common were verb-particle constructions (14 instances, or 37.8%) , followed by proper names (7, or 18.9%). There were no instances of compound nominals in the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

The full list of categories and the number of instances of paraphrases in each are in the table below (Table 1). MWEs, as defined by Baldwin et al., are marked in bold.

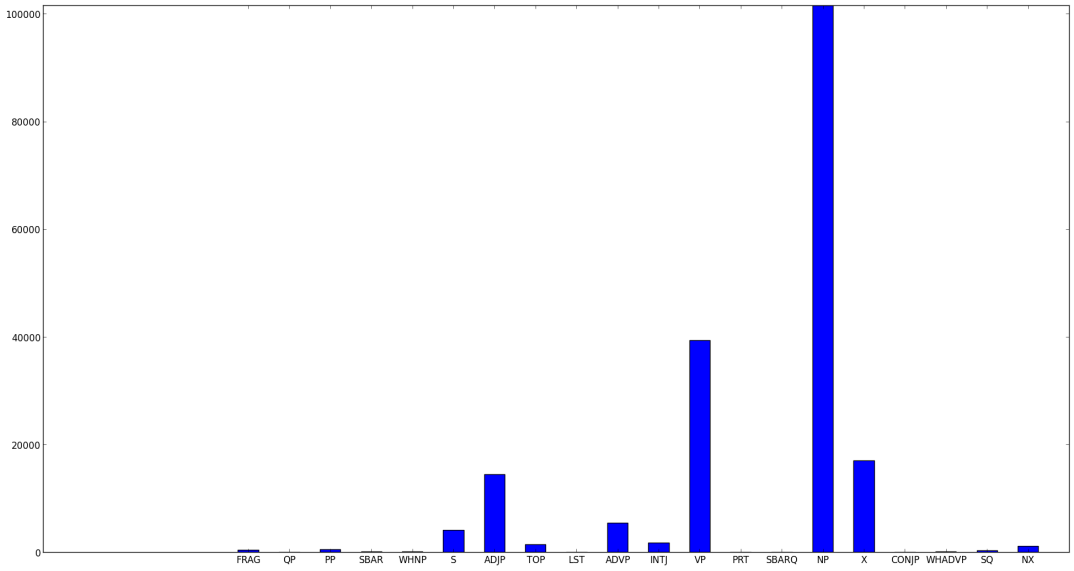
| **Paraphrase Category** | **Number of instances in sample** |
|---|---|
| expansion, same morphological form | 117 |
| determiner + one-word paraphrase | 86 |
| expansion | 62 |
| change of tense + paraphrase | 43 |
| inaccurate/bad paraphrase | 43 |
| adverbial modifier | 27 |
| acronym | 21 |
| one or more words the same as part of original word | 17 |
| **verb-particle** | 14 |
| determiner + plural | 9 |
| **proper noun** | 7 |
| change of tense | 7 |
| **fixed expression** | 6 |
| superlative | 5 |
| copula | 5 |
| **decomposable idiom** | 4 |
| **light verb** | 4 |
| number | 4 |
| infinitival to | 3 |
| nominalization | 3 |
| negation | 2 |
| time variation | 2 |
| punctuation | 2 |
| **non-decomposable idiom** | 2 |
| alternate spelling | 1 |
| **compound nominal** | 0 |

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

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:




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

Below (Table 1) are illustrative examples of all categories of paraphrases. Multiword expressions (as defined in Baldwin et al.) are marked in bold.

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

| MWE Category | Example |
|---|---|
| **verb-particle** | torched, burnt down |
| **fixed expression** | applied, put into effect |
| **non-decomposable idiom** | furious, as mad as hell |
| **proper noun** | markov, mr markov |
| **decomposable idiom** | nuts, out of your mind |
| **light verb** | issued, made available |
| superlative | notably, most particularly |
| negation | unused, not utilized |
| expansion | suffocate, need air |
| acronym | gpa, the global programme of action |
| change of tense + paraphrase | initiated, has undertaken |
| infinitival to | track, to follow |
| adverbial modifier | interesting, very interesting |
| one or ore words the same as part of original word | anytime, any point |
| determiner + one-word paraphrase | entry, the recordal |
| negation | unused, not utilized |
| copula | qualify, are eligible |
| expansion, same morphological form | revenue, tax revenue |
| nominalization | operation, proper functioning |
| inaccurate/bad paraphrase | peninsula, al jazeera |

Table 2: Examples from each category.

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

The results from these searches are summarized in the tables below.

## 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

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

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

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

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

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