Can Recognising Multiword Expressions Improve Shallow Parsing?

Ioannis Korkontzelos, Suresh Manandhar
Department of Computer Science
The University of York
Heslington, York, YO10 5NG, UK
{johnkork, suresh}@cs.york.ac.uk

Abstract

There is significant evidence in the literature that integrating knowledge about multiword expressions can improve shallow parsing accuracy. We present an experimental study to quantify this improvement, focusing on *compound nominals*, *proper names* and *adjective-noun constructions*. The evaluation set of multiword expressions is derived from *WordNet* and the textual data are downloaded from the web. We use a classification method to aid human annotation of output parses. This method allows us to conduct experiments on a large dataset of unannotated data. Experiments show that knowledge about multiword expressions leads to an increase of between 7.5% and 9.5% in accuracy of shallow parsing in sentences containing these multiword expressions.

1 Introduction

Multiword expressions are sequences of words that tend to co-occur more frequently than chance and are characterised by various levels of idiosyncracy (Baldwin et al., 2003; Baldwin, 2006). There is extended literature on various issues relevant to multiword expression; recognition, classification, lexicography, etc. (see Section 6). The vast majority of these publications identifies as motivation for multiword expression research its potential contribution to deep or shallow parsing. On the other side of this issue, the state-of-the-art parsing systems seem to ignore the fact that treating multiword expressions as syntactic units would potentially increase parser's accuracy.

In this paper, we present an experimental study attempting to estimate the contribution of integrating multiword expressions into shallow parsing. We focus on multiword expressions that consist of two successive tokens; in particular, *compound nominals* *proper names* and *adjective-noun constructions*. We also present a detailed classification method to aid human annotation during the procedure of deciding if a parse is correct or wrong. We present experimental results about the different classes of changes that occur in the parser output while unifying multiword expression components.

We conclude that treating known multiwords expressions as singletons leads to an increase of between 7.5% and 9.5% in accuracy of shallow parsing of sentences containing these multiword expressions. Increase percentages are higher for multiword expressions that consist of an adjective followed by a noun (12% to 15%); and even higher for non-compositional multiword expressions[1] that consist of an adjective and a noun (15.5% to 19.5%).

The rest of the paper is structured as follows: In Section 2 we present how multiword expressions can be annotated in text and used by a shallow parser. In Section 3 we present an overview of our experimental process. Section 4 explains how the set of target multiword expressions and textual corpora were created. In Section 5 we present and discuss the results of the experimental process. In Section 6 we present parts of the related literature. Section 7 concludes the paper and proposes some future work.

[1] Compositionality is defined as the degree to which the meaning of a multiword expression can be predicted by combining the meanings of its components (Nunberg et al., 1994).

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 636–644,
Los Angeles, California, June 2010.

2 Annotating Multiword expressions

In this paper, we present a study to inspect the extent to which knowledge of multiword expressions improves shallow parsing. Our approach focuses on English multiword expressions that appear as sequences in text. In particular, we focus on *compound nominals* (e.g. lemon tree), *proper names* (e.g. prince Albert) and *adjective-noun constructions* (e.g. red carpet).

Shallow or deep parsing should treat multiword expression as units that cannot be divided in any way. We replace the multiword expression tokens with a special made up token, i.e. the multiword expression constituents joined with an underscore. For example, we replace all occurrences of "lemon tree" with "lemon_tree".

We choose to replace the multiword expression words with a token that does not exist in the dictionary of the part of speech tagger. This is quite an important decision. Usually, a part of sheech tagger assigns to an unknown words the part of speech that best fits to it with respect to the parts of speech of the words around it and the training data. This is a desirable behaviour for our purposes.

The experimental results of our study quantify the difference between the shallow parser output of a big number of sentences after the replacement and the shallow parser output of the same sentences before the replacement. The comparison is done ignoring changes of parts of speech, assigned by the part of speech tagger.

3 Evaluation

The target of our experiment is to evaluate whether replacing the multiword expression tokens with a single token, unknown to the part of speech tagger, improves shallow parsing accuracy. The ideal way to perform this evaluation would be to use a corpus with manual annotation about parsing and multiword expressions. Given this corpus we would be able to measure the accuracy of a shallow (or deep) parser before and after replacing multiword expressions. However, to the best of our knowledge there is no corpus available to include this type of annotations in English.

Instead, there are two options: Firstly, we can use treebank data, where manual parsing annotation

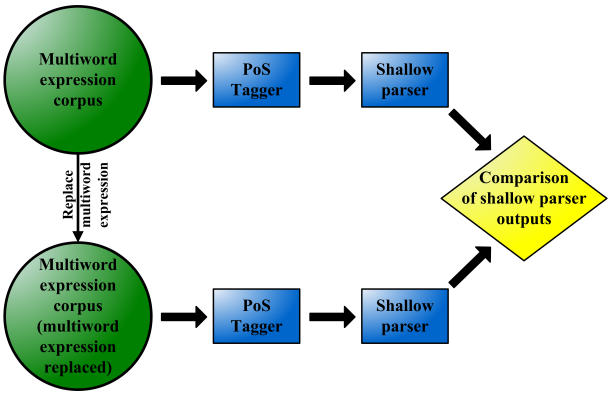

Figure 1: Evaluation process

is readily available, and manually annotate multiword expressions. The advantage of this approach is that results are directly comparable with other results of the literature, due to the use of benchmark data. Manual annotation of multiword expressions is a very time- and effort-consuming process due to the large size of most treebanks. Alternatively, multiword expression annotation could be done using a method of recognition. Annotating the multiword expressions that appear in *WordNet* could be a safe decision, in terms of correctness, however, *WordNet* is reported to have limited coverage of multiword expressions (Baldwin, 2006; Laporte and Voyatzi, 2008). *WordNet* covers only 9.1 % and 16.1 % of the datasets of Nicholson and Baldwin (2008) (484 noun compounds) and Kim and Baldwin (2008) (2169 noun compounds), respectively.

Secondly, we can use a set of multiword expressions as a starting point and then create corpora that contain instances of these multiword expressions. In succession, these sentences need to be manually annotated in terms of parsing, and this requires huge human effort. Alternatively, we can parse the corpora before and after replacing the multiword expression and then compare the parser output. This is the evaluation procedure that we chose to follow, and is shown in Figure 1.

The above procedure is only able to retrieve instances where the replacement of the multiword expression leads to a different parsing, a different allocation of tokens to phrases. It is not able to spot instances where the parser output remains unchanged after the replacement, no matter if they are correct. Since we are interested in measuring if replacing

Example A - Replacement causes no change	
Before:	*[NP they] [VP jumped] [PP over] [NP a bonfire] and [VP rolled] [NP a fire wheel] .*
After:	*[NP they] [VP jumped] [PP over] [NP a bonfire] and [VP rolled] [NP a fire_wheel] .*
Example B - Replacement corrects an error	
Before:	*[NP the blades] [VP ignited] and [NP he] [VP threw]* ***[NP the fire] wheel up*** *[PP into] [NP the air] .*
After:	*[NP the blades] [VP ignited] and [NP he] [VP threw]* ***[NP the fire_wheel] [PRT up]*** *[PP into] [NP the air] .*

Table 1: 2 shallow parsing examples. Multiword expression: *"fire wheel"*

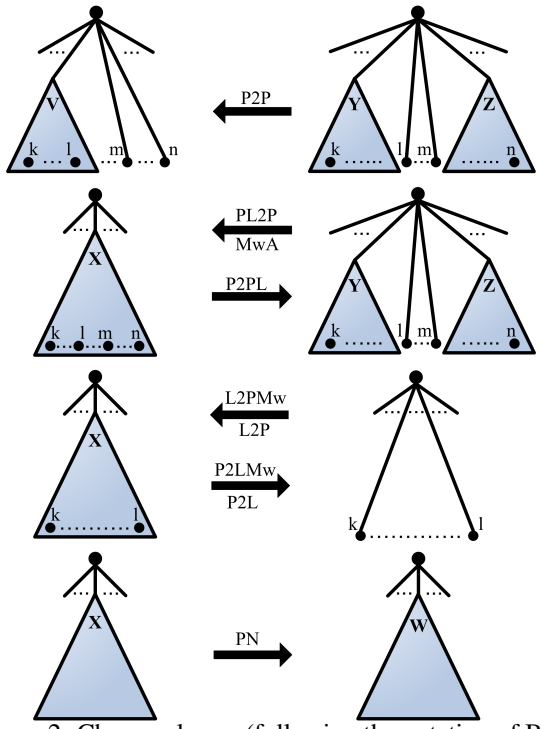

Figure 2: Change classes (following the notation of Bille (2005)). Triangles denote phrases and uppercase bold letters $V...Z$ denote phrase labels. Lowercase letters $k...n$ denote parsing leaves. For change classes *P2LMw* and *L2PMw*, X includes the multiword expression tokens. For change classes *P2L* and *L2P* it does not. For change class *MwA*, the multiword expression tokens are not assigned to the same phrase Y or Z.

multiword expressions with a single token improves parsing accuracy, we are not interested in instances that remain unchanged. We focus on instances that changed; either they were corrected or they were made wrong or they remain erroneous. For example, the shallow parser output for example A in Table 1 did not change after the replacement. Example B in Table 1 shows a sentence which was corrected after the replacement.

Instead of manually annotating the sentences whose parser output changed after the replacement as corrected or not, we identify a number of change classes under which we classify all these sentences. In the following section, we present the change classes. For each we thoroughly discuss whether its form guarantees that its sentences are wrongly parsed before the change and correctly parsed after the change. In this case, the sentences of the corresponding class should be counted as false positives. We also discuss the opposite; if the form of each change class guarantees that its sentences are correctly parsed before the change and wrongly parsed after the change. In this case, the sentences of the corresponding class should be counted as true negatives. For this discussion we hypothesize that among the possible output shallow parses for a given sentence the correct one has (a) the smallest number phrases, and (b) the smallest number of tokens not assigned to any phrase.

3.1 Shallow parsing change classes

In this section, we present a classification of cases where the shallow parser output of the sentence is different from the parser output of the same sentence after replacing the multiword expression with a single token. The secondary focus of this discussion is to estimate whether the specific form of each change class can lead to a safe conclusion about if the parser output of the sentence under discussion: (a) was wrong before the replacement and was then corrected, (b) was correct before the replacement and was then made wrong, or (c) was wrong before the replacement and remained wrong. For this discussion, we refer to words that are not assigned to any phase in the shallow parser output as "leaves".

Hypothesis: We base our analysis on the hypothesis that among the possible output shallow parses for a given sentence the correct one has (a) the smallest number phrases, and (b) the smallest number of leaves. The theoretical intuitions behind the hypothesis are: (a) parse trees with just leaves are partial parse trees and hence should not be preferred over complete parse trees. (b) when mistaken parse

trees are generally larger (with more phrases). We checked the hypothesis by manually annotating 80 randomly chosen instances; 10 for each change class that is counted as correct or wrong (see Table 2). 74 instances validated the hypothesis (92.5%).

Table 2 shows one example for each change class. Figure 2 presents the classes as transformations between trees, following the notation of Bille (2005).

Change class *P2LMw* (***P***hrase to ***L***eaves including the ***M***ulti***w***ord expression) Before replacing the multiword expression sequence with a single token, the multiword expression is assigned to some phrase, possibly together with other words. After the replacement, the components of that phrase are not assigned to any phrase, but instead as leaves.

Change class *P2L* (***P***hrase to ***L***eaves excluding the multiword expression) Similarly to change class *P2LMw*, before the replacement, some successive tokens excluding the multiword expression itself are assigned to some phrase. After the replacement, the components of that phrase appear as leaves.

Change class *L2PMw* (***L***eaves to ***P***hrase including the ***M***ulti***w***ord expression) The changes covered by this class are the opposite changes of change class *P2LMw*. Before the replacing the multiword expression sequence with a single token, the multiword expression sequence is not assign to any phrase possibly among other words. After the replacement, the multiword expression is assigned to a phrase.

Change class *L2P* (***L***eaves to ***P***hrase excluding the multiword expression) Similarly to change class *L2PMw*, before the replacement, one or more successive tokens excluding the multiword expression itself appear as leaves. After the replacement, these tokens are assigned to a phrase.

Change class *PL2P* (***P***hrases or ***L***eaves to ***P***hrase) After the replacement, the tokens of more than one phrases or leaves are assigned to a single phrase.

Change class *P2PL* (***P***hrase to ***P***hrases or ***L***eaves) In contrast to change class *PL2P*, after the replacement, the tokens of one phrase either are assigned to more than one phrases or appear as leaves.

Change class *PN* (***P***hrase label ***N***ame) After replacing the multiword expression sequence with a single token, one phrase appears with a different phrase label, although it retains exactly the same component tokens.

Change class *PoS* (***P***art ***o***f ***S***peech) After replacing the multiword expression sequence with a single token, one or more tokens appears with a different part of speech. This class of changes comes from the part of speech tagger, and are out of the scope of this study. Thus, in the results section we show a size estimate of this class, and then we present results about change classes, ignoring change class *PoS*.

Change class P2P (***P***hrases to less ***P***hrases) After replacing the multiword expression sequence with a single token, the component tokens of more than one successive phrases α are assigned to a different set of successive phrases β. However, it is always the case that phrases α are less than phrases β ($|\alpha| < |\beta|$).

Change class MwA (***M***ulti***w***ord expression ***A***llocation) Before replacing the multiword expression sequence, the multiword expression constituents are assigned to different phrases.

The instances of change classes where the parser output after the replacement has more parsing leaves or phrases than before are counted towards sentences that were parsed wrongly after the replacement. For these classes, change classes *P2LMw*, *P2L* and *P2PL*, most probably the parser output after the replacement is wrong.

In contrast, the instances of change classes where a sequence of tokens is assigned to a phrase, or many phrases are merged are counted towards sentences that were parsed wrongly before the replacement and correctly after the replacement. These changes, that are described by classes *L2PMw*, *L2P*, *PL2P* and *P2P*, most probably describe improvements in shallow parsing. The instances of change class *MwA* are counted as correct after the replacement because by definition all tokens of a multiword expression are expected to be assigned to the same phrase.

The instances of change class *PN* can be either correct or wrong after the replacement. For this reason, we present our results as ranges (see Table 4). The minimum value is computed when the instances of class *PN* are counted as wrong after the replacement. In contrast, the maximum value is computed when the instances of this class are counted as correct after the replacement.

3.2 Shallow parsing complex change classes

During the inspection of instances where the shallow parser output before the replacement is dif-

P2LMw	B	***[NP the(DT) action(NN) officer(NN)]*** *[NP logistic(JJ) course(NN)] [VP is(VBZ) designed(VBN)] [VP to(TO) educate(VB)] and(CC) [VP train(VB)] [NP military(JJ) personnel(NNS)] ...*	✗
	A	***the(DT) action_officer(NN)*** *[NP logistic(JJ) course(NN)] [VP is(VBZ) designed(VBN)] [VP to(TO) educate(VB)] and(CC) [VP train(VB)] [NP military(JJ) personnel(NNS)] ...*	
P2L	B	*... [NP the(DT) action(NN) officer(NN)]* ***[PP in(IN)]*** *[NP armenia(NN)] [VP signed(VBN)] ...*	✗
	A	*... [NP the(DT) action_officer(NN)]* ***in(IN)*** *[NP armenia(NN)] [VP signed(VBN)] ...*	
L2PMw	B	*"(")* ***affirmative(JJ) action(NN) officer(NN)*** *"(") [NP aao(NN)] [VP refers(VBZ)] [PP to(TO)] [NP the(DT) regional(JJ) affirmative(JJ) action(NN) officer(NN)] or(CC) [NP director(NN)] ...*	✓
	A	*"(")* ***[NP affirmative(JJ) action_officer(NN)]*** *"(") [NP aao(NN)] [VP refers(VBZ)] [PP to(TO)] [NP the(DT) regional(JJ) affirmative(JJ) action_officer(NN)] or(CC) [NP director(NN)] ...*	
L2P	B	*[NP the(DT) action(NN) officer(NN)]* ***usually(RB)*** *[VP delivers(VBZ)] ...*	✓
	A	*[NP the(DT) action_officer(NN)]* ***[ADVP usually(RB)]*** *[VP delivers(VBZ)] ...*	
PL2P	B	*... [VP to(TO) immediately(RB) report(VB)] [NP the(DT) incident(NN)] [PP to(TO)]* ***[NP the(DT) equal(JJ) opportunity(NN)] and(CC) [NP affirmative(JJ) action(NN) officer(NN)]*** *.(.)*	✓
	A	*... [VP to(TO) immediately(RB) report(VB)] [NP the(DT) incident(NN)] [PP to(TO)]* ***[NP the(DT) equal(JJ) opportunity(NN) and(CC) affirmative(JJ) action_officer(NN)]*** *.(.)*	
P2PL	B	*... [NP action(NN) officer(NN)]* ***[VP shall(MD) prepare(VB) and(CC) transmit(VB)]*** *...*	✗
	A	*... [NP action_officer(NN)]* ***[VP shall(MD) prepare(VB)] and(CC) [VP transmit(VB)]*** *...*	
PN	B	*... [NP an(DT) action(NN) officer(NN)]* ***[SBAR for(IN)]*** *[NP communications(NNS)] ...*	*?*
	A	*... [NP an(DT) action_officer(NN)]* ***[PP for(IN)]*** *[NP communications(NNS)] ...*	
PoS	B	*... [NP security(NN) officer(NN)] or(CC) "(") [NP* ***youth(JJ)*** *action(NN) officer(NN)] .(.) "(")*	*?*
	A	*... [NP security(NN) officer(NN)] or(CC) "(") [NP* ***youth(NN)*** *action_officer(NN)] .(.) "(")*	
P2P	B	*... ,(,) [PP as(IN)]* ***[NP a(DT) past(JJ) action(NN) officer(NN)] and(CC) command(NN) and(CC) control(NN) and(CC)*** *[NP intelligence(NN) communications(NNS) inspector(NN)] ...*	✓
	A	*... ,(,) [PP as(IN)]* ***[NP a(DT) past(JJ) action_officer(NN) and(CC) command(NN) and(CC) (control(NN)] and(CC)*** *[NP intelligence(NN) communications(NNS) inspector(NN)] ...*	
MwA	B	***the(DT) campus(NN) affirmative(JJ) action(NN) [NP officer(NN)]*** *[VP serves(VBZ)] ...*	✓
	A	***[NP the(DT) campus(NN) affirmative(JJ) action_officer(NN)]*** *[VP serves(VBZ)]...*	

Table 2: Examples for change classes. Multiword expression: *"action officer"*. Parts of speech appear within parentheses. "B" stands for "before" and "A" for "after" (multiword expression replacement). ✓or ✗ denote change classes that count positively or negatively towards improving shallow parsing. *?* denotes classes that are treated specially.

ferent from the shallow parser output after the replacement, we came across a number of instances that were classified in more than one class of the previous subsection. In other words, two or more classes of change happened. For example, in a number of instances, before the replacement, the multiword expression constituents are assigned to different phrases (change class *MwA*). After the replacement, the tokens of more than one phrases are assigned to a single phrase (change class *PL2P*). These instances consist new complex change classes and are named as the sum of names of the participating classes. The instances of the example above consist the complex change class *PL2P+MwA*.

4 Target multiword expressions and corpora collection

We created our set of target multiword expressions using *WordNet 3.0* (Miller, 1995). Out of its $52,217$ multiword expressions we randomly chose 120. Keeping the ones that consist of two tokens resulted in the 118 expressions of Table 3. Manually inspecting these multiword expressions proved that they are all *compound nominals*, *proper names* or *adjective-noun constructions*. Each multiword expression was manually tagged as compositional or non-compositional, following the procedure described in Korkontzelos and Manandhar (2009). Table 3 shows the chosen multiword expressions together with information about their compositionality and the parts of speech of their components.

Compositional Multiword expressions (Noun - Noun sequences)				
action officer (3119)	bile duct (21649)	cartridge brass (479)	field mushroom (789)	fire wheel (480)
key word (3131)	king snake (2002)	labor camp (3275)	life form (5301)	oyster bed (1728)
pack rat (3443)	palm reading (4428)	paper chase (1115)	paper gold (1297)	paper tiger (1694)
picture palace (2231)	pill pusher (924)	pine knot (1026)	potato bean (265)	powder monkey (1438)
prison guard (4801)	rat race (2556)	road agent (1281)	sea lion (9113)	spin doctor (1267)
tea table (62)	telephone service (9771)	upland cotton (3235)	vegetable sponge (806)	winter sweet (460)
Non-Compositional Multiword expressions (Noun - Noun sequences)				
agony aunt (751)	air conditioner (24202)	band aid (773)	beach towel (1937)	car battery (3726)
checker board (1280)	corn whiskey (1862)	corner kick (2882)	cream sauce (1569)	fire brigade (5005)
fish finger (1423)	flight simulator (5955)	honey cake (843)	jazz band (6845)	jet plane (1466)
laser beam (16716)	lemon tree (3805)	lip service (3388)	love letter (3265)	luggage van (964)
memory device (4230)	monkey puzzle (1780)	motor pool (3184)	power cord (5553)	prince Albert (2019)
sausage pizza (598)	savoy cabbage (1320)	surface fire (2607)	torrey tree (10)	touch screen (9654)
water snake (2649)	water tank (5158)	wood aster (456)		
Compositional Multiword expressions (Adjective - Noun sequences)				
basic color (2453)	cardiac muscle (6472)	closed chain (1422)	common iguana (668)	cubic meter (4746)
eastern pipistrel (128)	graphic designer (8228)	hard candy (2357)	ill health (2055)	kinetic theory (2934)
male parent (1729)	medical report (3178)	musical harmony (1109)	mythical monster (770)	red fox (10587)
relational adjective (279)	parking brake (7199)	petit juror (991)	taxonomic category (1277)	thick skin (1338)
toxic waste (7220)	universal donor (1454)	parenthesis-free notation (113)		
Non-Compositional Multiword expressions (Adjective - Noun sequences)				
black maria (930)	dead end (5256)	dutch oven (4582)	golden trumpet (607)	green light (5960)
high jump (4455)	holding pattern (3622)	joint chiefs (2865)	living rock (985)	magnetic head (2457)
missing link (5314)	personal equation (873)	personal magnetism (2869)	petit four (1506)	pink lady (1707)
pink shower (351)	poor devil (1594)	public eye (3231)	quick time (2323)	red devil (2043)
red dwarf (6526)	red tape (2024)	round window (1380)	silent butler (332)	small beer (2302)
small voice (4313)	stocking stuffer (7486)	sweet bay (1367)	teddy boy (2413)	think tank (4586)

Table 3: 118 multiword expressions randomly chosen from *WordNet*. The size of the respective corpus in sentences appears within parentheses.

For each multiword expression we created a different corpus. Each consists of webtext snippets of length 15 to 200 tokens in which the multiword expression appears. Snippets were collected following Korkontzelos and Manandhar (2009). Given a multiword expression, a set of queries is created: All synonyms of the multiword expression extracted from WordNet are collected[2]. The multiword expression is paired with each synonym to create a set of queries. For each query, snippets are collected by parsing the web-pages returned by *Yahoo!*. The union of all snippets produces the multiword expression corpus.

In Table 3, the number of collected corpus sentences for each multiword expression are shown within parentheses. *GENIA* tagger (Tsuruoka et al., 2005) was used as part of speech tagger. *SNoW-based Shallow Parser* (Munoz et al., 1999) was used for shallow parsing.

[2]e.g. for "red carpet", corpora are collected for "red carpet" and "carpet". The synonyms of "red carpet" are "rug", "carpet" and "carpeting".

5 Experimental results and discussion

The corpora collecting procedure of Section 4 resulted in a corpus of $376,007$ sentences, each one containing one or more multiword expressions. In $85,527$ sentences (22.75%), the shallow parser output before the replacement is different than the shallow parser output after the replacement. 7.20% of these change instances are due to one or more parts of speech changes, and are classified to change class *PoS*. In other words, in 7.20% of cases where there is a difference between the shallow parses before and after replacing the multiword expression tokens there is one or more tokens that were assigned a different part of speech. However, excluding parts of speech from the comparison, there is no other difference between the two parses.

The focus of this study is to quantify the effect of unifying multiword expressions in shallow parsing. Part of speech tagging is a component of our approach and parts of speech are not necessarily parts of the parser output. For this reason, we chose to ignore part of speech changes, the changes of class *PoS*. Below, we discuss results for all other classes.

Multiword expressions			Shallow Parsing improvement	
class	*PS*	sentences	min.	max.
On average	-	376,007	7.47%	9.49%
Comp.	N N	93,166	5.54%	7.19%
Non-Comp.	N N	127,875	3.66%	4.44%
Comp.	J N	68,707	7.34%	9.21%
Non-Comp.	J N	86,259	15.32%	19.67%
-	N N	221,041	4.45%	5.60%
	J N	154,966	11.78%	15.03%
Comp.	-	161,873	6.30%	8.05%
Non-Comp.	-	214,134	8.36%	10.57%

Table 4: Summary of results. *PS*: parts of speech, Comp: compositional, N: noun, J: adjective, min.: minimum, max.: maximum.

Table 4 shows a summary of our results. The first two columns describe classes of multiword expression with respect to compositionality and the parts of speech of the component words. The first line accounts for the average of all multiword expressions, the second one for compositional multiword expressions made of nouns, etc. The third column shows the number of corpus sentences of each class.

For each one of the classes of Table 4, the fourth and fifth columns show the minimum and maximum improvement in shallow parsing, respectively, caused by unifying multiword expression tokens. Let $\|X\|$ be the function that returns the number of instances assigned to change class X. With respect to the discussion of Subsection 3.1 about how the instances of each class should be counted towards the final results, the minimum and maximum improvements in shallow parsing are:

$$\begin{aligned} min \;=\; & -\|P2LMw\|-\|P2L\|+\|L2PMw\|+\|L2P\|+ \\ & +\|PL2P\|-\|P2PL\|+\|PL2P{+}MwA\|+ \\ & +\|P2P\|+\|P2P{+}MwA\|-\|PN\| \qquad (1) \\ max \;=\; & -\|P2LMw\|-\|P2L\|+\|L2PMw\|+\|L2P\|+ \\ & +\|PL2P\|-\|P2PL\|+\|PL2P{+}MwA\|+ \\ & +\|P2P\|+\|P2P{+}MwA\|+\|PN\| \qquad (2) \end{aligned}$$

On average of all multiword expressions, unifying multiword expression tokens contributes from 7.47% to 9.49% in shallow parsing accuracy. It should be noted that this improvement is reported on sentences which contain at least one known multiword expression. To project this improvement on any general text, one needs to know the percentage of sentences that contain known multiword expressions. Then the projected improvement can be computed by multiplying these two percentages.

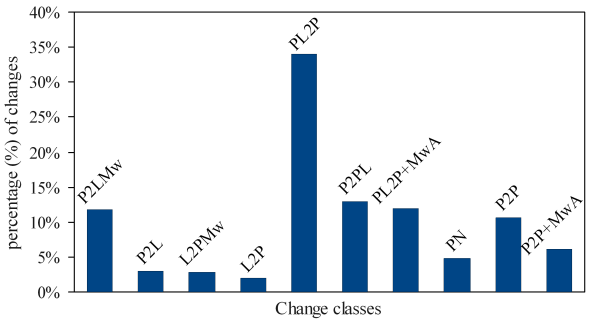

Figure 3: Average change percentages per change class.

Table 4 shows that the increase in shallow parsing accuracy is lower for expressions that consist of nouns than for those that consist of an adjective and a noun. Moreover, the improvement is higher for non-compositional expressions than compositional ones. This is expected, due to the idiosyncratic nature of non-compositional multiword expressions. The highest improvement, 15.32% to 19.67%, occurs for non-compositional multiword expressions that consist of an adjective followed by a noun.

Figure 3 shows the percentage of each class over the sum of sentences whose parse before unifying multiword expression tokens is different for the parse after the replacement. The most common change class is *PL2P*. It contains sentences in the shallow parser output of which many phrases or leaves were all assigned to a single phrase. 34.03% of the changes are classified in this class. The least common classes are change classes *P2L*, *L2PMw* and *L2P*. Each of these accounts for less than 3% of the overall changes.

6 Related Work

There have been proposed several ways to classify multiword expressions according to various properties such as compositionality and institutionalisation[3] (Moon, 1998; Sag et al., 2002; Baldwin, 2006). There is a large variety of methods in the literature that address recognising multiword expressions or some subcategory. McCarthy (2006) divides multiword expression detect-

[3]Institutionalisation is the degree that a multiword expression is accepted as lexical item through consistent use over time.

ing methods into statistical (e.g. pointwise mutual information (*PMI*)), translation-based, dictionary-based, substitution-based, and distributional. Statistical methods score multiword expression candidates based on co-occurrence counts (Manning and Schutze, 1999; Dunning, 1993; Lin, 1999; Frantzi et al., 2000). Translation-based methods usually take advantage of alignment to discover potential multiword expressions (Venkatapathy and Joshi, 2005).

Other methods use dictionaries to reveal semantic relationships between the components of potential multiword expressions and their context (Baldwin et al., 2003; Hashimoto et al., 2006). Substitution-based methods decide for multiword expressions by substituting their components with other similar words and measuring their frequency of occurrence (Lin, 1999; Fazly and Stevenson, 2006). These techniques can be enriched with selectional preference information (Van de Cruys and Moirón, 2007; Katz and Giesbrecht, 2006). Fazly and Stevenson (2007) propose measures for *institutionalisation*, *syntactic fixedness* and *compositionality* based on the selectional preferences of verbs. There are several studies relevant to detecting compositionality of noun-noun, verb-particle and light verb constructions and verb-noun pairs (e.g. Katz and Giesbrecht (2006)).

To the best of our knowledge there are no approaches integrating multiword expression knowledge in deep or shallow parsing. However, there are several attempts to integrate other forms of lexical semantics into parsing. Bikel (2000) merged the Brown portion of the Penn Treebank with SemCor, and used it to evaluate a generative bilexical model for joint word sense disambiguation and parsing. Similarly, Agirre et al. (Agirre et al., 2008) integrated semantic information in the form of semantic classes and observed significant improvement in parsing and PP attachment tasks. Xiong et al. (2005) integrated first-sense and hypernym features in a generative parse model applied to the Chinese Penn Treebank and achieved significant improvement over their baseline model. Fujita et al. (2007) extended this work by implementing a discriminative parse selection model, incorporating word sense information and achieved great improvements as well. Examples of integrating selectional preference information into parsing are Dowding et al. (1994) and Hektoen (1997).

7 Conlusion and future work

In this paper, we presented an experimental study attempting to estimate the contribution of unifying multiword expression components into shallow parsing. The evaluation is done based on 118 multiword expressions extracted from *WordNet 3.0*. They consist of two successive components and are in particular, *compound nominals*, *proper names* or *adjective-noun constructions*.

Instead of using pre-annotated text, we collected sentences that contain the above multiword expressions from the web. We applied shallow parsing before and after unifying multiword expression tokens and compared the outputs. We presented a detailed classification of changes in the shallow parser output to aid human annotation during the procedure of deciding if a parser output is correct or wrong.

We presented experimental results about change classes and about the overall improvement of unifying multiword expression tokens with respect to compositionality and the parts of speech of their components. We conclude that unifying the tokens of known multiwords expressions leads to an increase of between 7.5% and 9.5% in accuracy of shallow parsing of sentences that contain these multiword expressions. Increase percentages are higher on *adjective-noun constructions* (12% to 15%); and even higher on non-compositional *adjective-noun constructions* (15.5% to 19.5%).

Future work will focus in conducting similar experiments for multiword expressions longer than two words. One would expect that due to their size, a wrong interpretation of their structure would affect the shallow parser output more than it does for multiword expressions consisting of two words. Thus, unifying multiword expressions longer than two words would potentially contribute more to shallow parsing accuracy.

Furthermore, the evaluation results presented in this paper could be strengthened by adding manual multiword expression annotation to some treebank. This would provide a way to avoid the change class analysis presented in Subsection 3.1 and compute statistics more accurately. Finally, the results of this paper suggest that implementing a parser able to recognise multiword expressions would be very helpful towards high accuracy parsing.

References

E. Agirre, T. Baldwin, and D. Martinez. 2008. Improving parsing and PP attachment performance with sense information. In *Proceedings of ACL*, pages 317–325, USA. ACL.

T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows. 2003. An empirical model of multiword expression decomposability. In *proceedings of the ACL workshop on MWEs*, pages 89–96, USA. ACL.

T. Baldwin. 2006. Compositionality and multiword expressions: Six of one, half a dozen of the other? In *proceedings of the ACL workshop on MWEs*, Australia. ACL.

D. Bikel. 2000. A statistical model for parsing and word-sense disambiguation. In *proceedings of the 2000 Joint SIGDAT conference: EMNLP/VLC*, pages 155–163, USA. ACL.

P. Bille. 2005. A survey on tree edit distance and related problems. *Theoretical Computer Science*, 337(1-3):217–239.

J. Dowding, R. Moore, F. Andryt, and D. Moran. 1994. Interleaving syntax and semantics in an efficient bottom-up parser. In *proceedings of ACL*, pages 110–116, USA. ACL.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

A. Fazly and S. Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of EACL*, pages 337–344, Italy.

A. Fazly and S. Stevenson. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *proceedings of the ACL workshop on MWEs*, pages 9–16, Czech Republic. ACL.

K. Frantzi, S. Ananiadou, and H. Mima. 2000. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130.

S. Fujita, F. Bond, S. Oepen, and T. Tanaka. 2007. Exploiting semantic information for hpsg parse selection. In *proceedings of DeepLP*, pages 25–32, USA. ACL.

C. Hashimoto, S. Sato, and T. Utsuro. 2006. Detecting japanese idioms with a linguistically rich dictionary. *Language Resources and Evaluation*, 40(3):243–252.

E. Hektoen. 1997. Probabilistic parse selection based on semantic cooccurrences. In *proceedings of IWPT*, pages 113–122, USA.

G. Katz and E. Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *proceedings of the ACL workshop on MWEs*, pages 12–19, Australia. ACL.

S. Kim and T. Baldwin. 2008. Standardised evaluation of english noun compound interpretation. In *proceedings of the LREC workshop on MWEs*, pages 39–42, Morocco.

I. Korkontzelos and S. Manandhar. 2009. Detecting compositionality in multi-word expressions. In *proceedings of ACL-IJCNLP*, Singapore.

E. Laporte and S. Voyatzi. 2008. An Electronic Dictionary of French Multiword Adverbs. In *proceedings of the LREC workshop on MWEs*, pages 31–34, Marocco.

D. Lin. 1999. Automatic identification of non-compositional phrases. In *proceedings of ACL*, pages 317–324, USA. ACL.

C. Manning and H. Schutze, 1999. *Foundations of Statistical NLP, Collocations*, chapter 5. MIT Press.

D. McCarthy. 2006. Automatic methods to detect the compositionality of MWEs. presentation slides. url: www.sunum.org/myfiles/B2/McCarthyCollocIdioms06.ppt last accessed: 28/11/2009.

G. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

R. Moon. 1998. *Fixed Expressions and Idioms in English. A Corpus-based Approach.* Oxford: Clarendon Press.

M. Munoz, V. Punyakanok, D. Roth, and D. Zimak. 1999. A learning approach to shallow parsing. In *proceedings of EMNLP/VLC*, pages 168–178, USA.

J. Nicholson and T. Baldwin. 2008. Interpreting compound nominalisations. In *proceedings of the LREC workshop on MWEs*, pages 43–45, Morocco.

G. Nunberg, T. Wasow, and I. Sag. 1994. Idioms. *Language*, 70(3):491–539.

I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *proceedings of CICLing*, pages 1–15, Mexico.

Y. Tsuruoka, Y. Tateishi, J. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *Advances in Informatics*, pages 382–392.

T. Van de Cruys and B. Moirón. 2007. Semantics-based multiword expression extraction. In *proceedings of the ACL workshop on MWEs*, pages 25–32, Czech Republic. ACL.

S. Venkatapathy and A. Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *proceedings of HLT*, pages 899–906, USA. ACL.

D. Xiong, S. Li, Q. Liu, S. Lin, and Y. Qian. 2005. Parsing the penn chinese treebank with semantic knowledge. In *proceedings of IJCNLP*, pages 70–81, Korea.