Vol.2012-NL-209 No.7
2012/11/22

英語の複単語表現辞書の構築と品詞タグ付けへの応用

重藤 優太郎[1,a)] 東 藍[1,b)] 近藤 修平[1,c)] 北裏 龍太[1,d)] 坂口 慶祐[1,e)] 光瀬 智哉[1,f)]
久本 空海[1,g)] 吉本 暁文[1,h)] Frances Yung[1,i)] 松本 裕治[1,j)]

概要：これまで，英語における品詞タグ付けは様々な手法が提案されており，それらの手法を用いることで高い精度を得ることが示されてきた．しかし，それらの手法は単語のみに対して品詞タグ付けを行うものであり，連語などの複数の構成要素からなる複単語表現に対して品詞タグ付けを行うことができなかった．そこで本稿では，複単語表現辞書を構築し，可変長 CRF を用いて複単語表現を考慮した品詞タグ付けを行い，品詞タグ付けの精度と複単語表現の認識精度の評価を行った．その結果，複単語表現辞書を用いることで複単語表現を考慮しないタグ付けを行うよりも高い精度が得られた．

Construction of English Multiword Dictionary and its Application to POS Tagging

Yutaro Shigeto[1,a)] Ai Azuma[1,b)] Shuhei Kondo[1,c)] Ryuta Kitaura[1,d)] Keisuke Sakaguchi[1,e)]
Tomoya Kose[1,f)] Sorami Hisamoto[1,g)] Akifumi Yoshimoto[1,h)] Frances Yung[1,i)]
Yuji Matsumoto[1,j)]

Abstract: Many previous studies have proposed various methods for English Part-Of-Speech (POS) tagging, and these methods have been shown to give high accuracy. However POS tagging in these previous methods are only on each word, and they cannot tag on multiword expressions consisting of multiple lexical items, such as collocations. In this research we constructed a multiword expression dictionary, and we used variable length CRF to conduct POS tagging which takes into consideration the multiword expressions. We, then, evaluated the tagging accuracy and the multiword recognition accuracy. In result, the tagger with the multiword dictionary achieved higher accuracy compared with the tagger which does not consider multiword expressions.

1. はじめに

タグ付きコーパスを利用した統計的学習手法の進展により，品詞タグ付けや統語解析など基盤的な言語解析技術は大きな進歩を遂げているが，統語的に特殊な現象によって困難な問題が残されている．その要因の一つに複単語表現 (Multiword Expression, MWE) の取り扱いの問題がある．MWE は，単語境界 (英語の場合はスペース) を越える表現で特異な解釈 (“idiosyncratic interpretation that cross word boundaries”[1]) をもつ表現と定義される．MWE は，単語と統語の間にある言語要素と考えられ，固定した表現からある種の統語的な自由度を持つもの，また，コロケーションのような意味的な単語の共起まで様々な表現をさし得る．Sag ら [1] は，MWE を大きく次のように分類している．

- Lexicalized phrases
  - fixed expressions: 固定した表現であり，常に同じ語順と形をもつ
  - semi-fixed expressions: 語順の自由度はないが，複数

[1] 奈良先端科学技術大学院大学 情報科学研究科
Nara Institution Science of Thecnology
[a)] yutaro-s@is.naist.jp
[b)] ai-a@is.naist.jp
[c)] shuhei-k@is.naist.jp
[d)] ryuta-k@is.naist.jp
[e)] keisuke-sa@is.naist.jp
[f)] tomoya-kos@is.naist.jp
[g)] sorami-h@is.naist.jp
[h)] akifumi-y@is.naist.jp
[i)] pikyufrances-y@is.naist.jp
[j)] matsu@is.naist.jp

形や活用などの変化形，冠詞の選択などの自由度をもつ
  - syntactically flexible expressions: 修飾語や目的語などの挿入，受け身形などの統語的変形などを許す
- Institutionalized phrases
  - 統語的，意味的には構成的な表現，いわゆるコロケーションなど

MWE の取り扱いは近年注目を集め，2003 年の ACL 併設のワークショップ以来，MWE に関するワークショップ (Workshop on Multiword Expression) が例年のように開催されている[*1]．しかし，多くの研究は，MWE の自動抽出や文中での自動識別，あるいは，句動詞 (phrasal verbs) や複合名詞 (compound nouns) など特定の MWE の解釈や翻訳などの応用研究を対象としている．

我々のグループでは，英語の MWE 辞書の構築とそれを用いた言語解析器の構築を目指している．本稿では，その第一歩として英語の fixed MWE 辞書の構築と品詞タグ付けシステムに関する現状の報告を行なう．具体的には，次の 3 点について報告する．

(1) Wiktionary を利用した英語の機能表現に相当する MWE 辞書の構築
(2) Penn Treebank に出現する MWE のアノテーション
(3) MWE 辞書を用いた品詞タグ付けシステムの構築と性能評価

以下，次節で MWE に関する関連研究の概要を述べる．3 節では，Wiktionary からの MWE の抽出について述べ，4 節で，Penn Treebank に出現する MWE の用法の分類のための我々が取った手続きについて説明する．5 節では，MWE アノテーションが施された Penn Treebank を用いた品詞タグ付け実験について報告し，6 節でまとめと今後の活動予定について述べる．

2. 関連研究

MWE 研究の多様性に比べて，MWE に関する言語資源の構築は意外に進んでいない．例えば，フランス語の副詞 MWE 辞書の構築 [2][3]，オランダ語の MWE 辞書構築 [4] などの研究があるが，英語の網羅的かつ自由に入手可能な MWE 辞書は構築されていない．日本語では，複単語機能語表現辞書 [5] や網羅的な MWE 辞書 [6] の構築が行われている．本稿では，Wiktionary の英語辞書[*2]中の空白を含む見出し語のうち，副詞，前置詞，接続詞，限定詞，代名詞，前置詞句に分類される固定的な表現を抽出することによって，英語の機能語相当の MWE 辞書構築を行った．一方，MWE のアノテーションを施したコーパスも多くはなく，フランス語，スウェーデン語など限られた言語のコーパスしか存在しない．British National Corpus には一部 MWE のアノテーションが施されているが，網羅的ではない．我々は，Wiktionary から取り出した複単語機能表現の Penn Treebank における出現を抽出し，それぞれが MWE 用法か字義通りの (リテラルな) 用法かを区別し，MWE 用法にはその表現全体の品詞情報のタグ付けを行った．

MWE 辞書を利用した言語解析に関する研究は，種々行われている．Nivre ら [7] は，MWE の認識が統語解析，特に依存構造解析にどのような効果をもつかを調査した．彼らは，スウェーデン語のコーパスにおいて，MWE を認識することができれば，依存構造解析の解析精度の向上が見られることを確認した．Korkontzelos ら [8] は，複合名詞や固有名詞など複単語名詞をアノテートすることで基本句チャンキングの解析性能が大きく向上できることを示した．他に，MWE の識別が語義曖昧性解消に有効に働くとの報告 [9] などがある．

文中の MWE 認識については，単語の共起度や頻度に基づいて未知の表現の抽出が試みられることが多いが，大規模な MWE 辞書や MWE タグ付きコーパスを用いた言語解析による MWE 認識の研究も行われている．Constant ら [11] は，CRF を用い，BIO タグ付けに基いて MWE の認識実験を行なっている．Green ら [12] は，MWE を木構造としてタグ付けしたコーパスを構築し，Tree Substitution Grammar を利用して MWE を認識する手法を提案した．前者を MWE 認識の前処理とした句構造解析と，後者の手法に対して reranking を適用した手法の比較が Constant ら [10] によって行われている．

前節で述べたように，MWE には固定的な表現から統語的な自由度をもつものまで多様性があるが，MWE を考慮した言語的な解析では，主として固定的な表現が対象とされることが多い．上で参照した Nivre ら [7] の実験で彼らが対象にした MWE も主として固定的な表現であり，次のような品詞に分類されるものだった．

(1) Multiword names: 人名，地名など
(2) Numerical expressions: 数値，数式など
(3) Compound function words: 複合機能語
  (a) Adverbs: 副詞
  (b) Prepositions: 前置詞
  (c) Subordinating conjunctions: 従属接続詞
  (d) Determiners: 限定詞
  (e) Pronouns: 代名詞

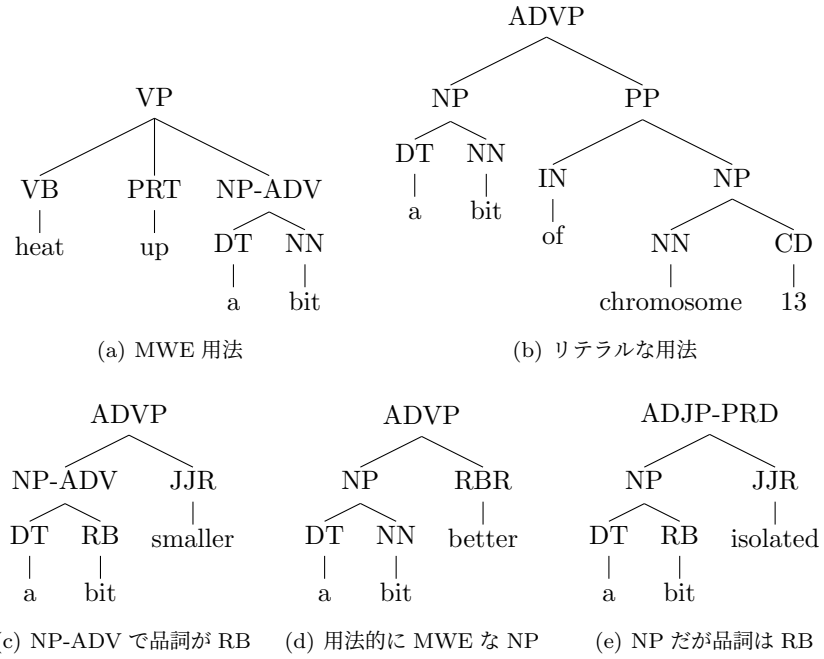
これらのうち，人名，地名などの複合名詞や数値などは構造的な多様性や統語的な機能は限定されている．一方，複合機能語は，構造的にも統語的にも役割の多様性が見られ，これらの識別が品詞タグ付けや統語解析の精度に与える影響が大きいと考えられる．

本稿では，英語版 Wiktionary から上記の複合機能語に相当する品詞に分類されるものを抽出し，これらを解析することができる品詞タグ付けシステムを (MWE の認識

[*1] http://multiword.sourceforge.net/
[*2] https://en.wiktionary.org/wiki/Wiktionary:Main_Page

表 1　Wiktionary および Penn Treebank 中の MWE 数
Table 1　The Number of MWE Entries in Wiktionary and the Penn Treebank.

	Adverb	Conjunction	Determiner	Prepositional Phrase	Preposition	Pronoun
Wiktionary	1501	49	15	165	110	83
Penn Treebank	468	35	9	66	77	18



(a) MWE 用法　(b) リテラルな用法

(c) NP-ADV で品詞が RB　(d) 用法的に MWE な NP　(e) NP だが品詞は RB

図 1　「a bit」が持つ様々な構造の例
Fig. 1　Examples of phrase structures annotated to "a bit"

を可能にするため可変長の単語を取り扱うように拡張した)CRF を用いて構築し，解析性能の評価を行った．以後，この拡張した CRF を可変長 CRF と呼ぶ．

3. Wiktionary からの MWE の抽出

英語の MWE 辞書を構築するために，2012 年 7 月 14 日時点の英語版 Wiktionary から見出し語に空白を含む記事を抽出した．今回は副詞，接続詞，限定詞，前置詞，前置詞句，代名詞のいずれかのカテゴリーに該当する MWE を対象とした．名詞は固有表現抽出等の既存の解析手法を流用することが容易なため，また，動詞は semi-fixed あるいは syntactically flexible expression に相当するものが多く fixed expression とは異なる解析手法が必要なため，それぞれ対象外とした．形容詞も fixed expression ではないものが多数を占めることに加え，形容詞句として be 動詞等の補語となることはあっても名詞を前置修飾することがない．そのため，通常の形容詞とは性質の異なるものが多く，原則として今回は対象外とした．ただし，形容詞を含む複数のカテゴリーを持つ MWE で，副詞などの今回抽出対象としたカテゴリーにも該当するものは対象に含めた．結果として得られた見出し語の数を表 1 に示す．

4. Penn Treebank に出現する MWE の用法の分類

Wiktionary から抽出した MWE を元に，Penn Treebank に出現する，MWE である可能性がある単語列を対象としたアノテーションを行った．Penn Treebank 中に一致する単語列が出現する MWE の数を表 1 に示す．MWE かリテラルな用法であるかの判断基準として，MWE の場合はリテラルな用法とは異なる木構造や品詞列になっていること，同一の用法では木構造が統一されていることが期待されていた．しかし，実際にそのようになっている例も多かったが，（図 1(a)，1(b)），同一の用法で木構造や品詞のアノテーションに揺れがある例（図 1(c)，1(d)，1(e)）や，異なる用法で木構造が共通している例（図 1(b)，1(d)）も存在した．このため，木構造と品詞列をもとに MWE の出現例を分類した上で，人手による確認と修正を行った．また，MWE であるかリテラルな用法であるかにかかわらず，同一の用法内で Penn Treebank の品詞タグ付けに表記揺れがある場合，その用法内で多数を占める品詞列に統一した．このアノテーションは MWE になり得る単語列のみに対して行っており，MWE になり得ない単語列に対しては行っていない．

5. 実験および考察

5.1 実験設定

本節では構築した MWE 辞書と Penn Treebank を用いて行った品詞タグ付けの実験について述べる．この実験では，Penn Treebank のセクション 0-18 をトレーニングデータ，セクション 22-24 をテストデータとし品詞タグ付けを行った．これは，英語の品詞タグ付けにおける一般的な設定である．ここで，ピリオドなどの空白の無い単語の境界（e.g., I have a pen .）はすでにわかっているものとする．このデータセットに対して既存の品詞タガーでは 97%を超える精度 (Per token accuracy) を得られることが報告されている [13].

実験には以下の辞書，および，それぞれに対応する Penn Treebank を用いた．

(a) Penn Treebank に出現した単語から構成される辞書
(b) MWE 辞書を構築する上で，アノテーションし直した Penn Treebank から構成される辞書
(c) MWE を BIO タグを用いて表現した辞書
(d) 本稿で構築した MWE 辞書

各辞書の見出し語は素性として大分類の品詞（CPOS），小分類の品詞（PPOS），表層（Surface form）の情報を持っている．それに加え，MWE 辞書は見出し語に MWE を含み，MWE の品詞と MWE の各構成要素となる単語の品詞の情報を持っている．以後，MWE の構成要素となる単語のことを構成要素と呼ぶ．例えば，「a number of」の場合，これが持つ MWE の品詞は DT であり，各構成要素の品詞は DT，NN，IN である．これらの辞書は Penn Treebank に出現する単語に基づいて構築されているため，テストデータに未知語は存在しない．各辞書における，MWE の表現方法の例として副詞的用法（RB）である「about to」を表 2 に示す．表 2 の（c）BIO POS の「-B」は MWE の構成要素の先頭（Beginning）の品詞に対して付与され，「-I」は MWE の構成要素の 2 番目以降（Inside）の品詞に対して付与されたものである．（d）MWE の RB は MWE である「about to」の品詞を表す．解析器には MWE を取り扱うことのできる可変長 CRF を用いており，Penn Treebank とこれらの辞書で品詞タグ付けを行う．MWE を含んだトレリスの例を図 2 に示す.

評価は品詞の精度と MWE の認識精度で行った．まず，上記 (a)，(b)，(c)，(d) の辞書を用いた場合の品詞タグ付けの精度の評価を示す．この品詞タグ付けにおいて，(d) の辞書を用いた実験では，MWE と認識された語については各構成要素の品詞に分解し，分解した各構成要素の品詞で評価を行う．

次に，辞書 (c)，(d) を用いて MWE の認識精度の評価を示す．これら以外に，ベースラインとして，MWE と一致する単語列全てに対して，Penn Treebank で出現頻度が高い用法（MWE 用法かリテラルな用法）を割り当てる場合の精度も測定した．なお，このベースラインにおいて複数の MWE と一致する単語列に対しては，長さが長い用法を優先して用法を決定した．このベースラインの認識例として「as well as」が入力された場合を考える．認識され得る MWE は「as well」と「as well as」の 2 通りある．ベースラインは長さが長い MWE を認識することにしたので，2 単語から構成される「as well」ではなく，3 単語から構成される「as well as」を MWE として認識する．

表 2 各辞書における MWE の表現方法（about to）
Table 2 Representations of MWE in each dictionary（"about to"）

Dicrionary	Word/POS
(a) Original POS	about/RB to/TO
(b) Revised POS	about/IN to/TO
(c) BIO POS	about/RB-B to/RB-I
(d) MWE	about to/RB

品詞タグ付けの精度と MWE の認識精度の評価尺度は以下の式で評価する．

$$Precision = \frac{\#\ of\ correct\ tokens}{\#\ of\ tokens\ in\ output} \quad (1)$$

$$Recall = \frac{\#\ of\ correct\ tokens}{\#\ of\ tokens\ in\ corpus} \quad (2)$$

$$F-measure = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \quad (3)$$

この際，MWE を考慮しない解析の場合，英単語の境界は空白で明示的に示されているため，品詞タグ付けの精度は $Precision = Recall = F-measure$（Per token accuracy）となる．

5.2 素性

この小節では，可変長 CRF に利用した素性について述べる．利用した素性を表 3 に示す．表 3 の Bigram 素性に記してある Head POS と Tail POS は，MWE の構成要素の先頭と末尾にある構成要素の品詞を表現している．例えば「a lot of/DT」（a/DT lot/NN of/IN）の場合，Head POS は「a/DT」の DT であり，Tail POS は「of/IN」の IN を表す.

5.3 実験結果および考察

ここでは，品詞タグ付けの精度と MWE の認識精度の実験結果について述べる．まず，品詞タグ付けの精度を表 4 に示す．表 4 より，(a) Original POS よりも (b) Revised POS の方が 0.02%精度が良いことがわかる．これは，元々存在していた Penn Treebank の品詞の誤りを訂正したことで，システムのタグ付けした品詞が正解と一致することに

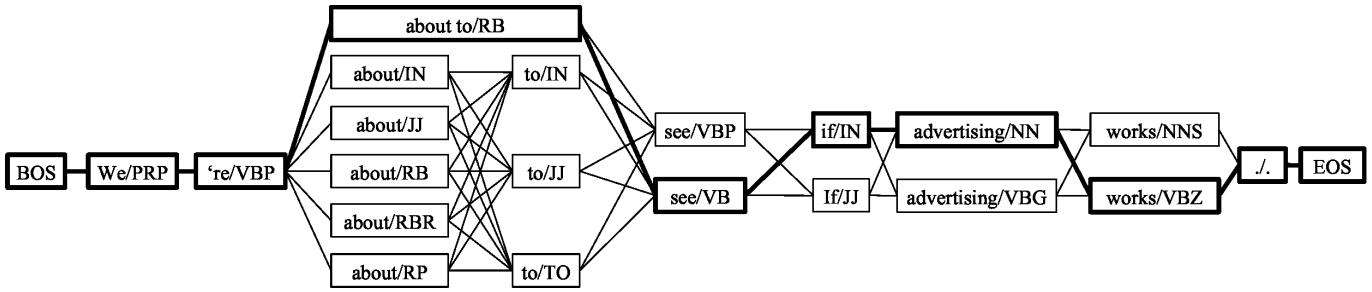


図 2 MWE（about to/RB）を含むトレリスの例（正解は太枠）
Fig. 2 Example of lattice containing MWE (”about to/RB”)（correct path is marked with bold boxes.）

表 3 素性テンプレート
Table 3 Feature templates

Unigram features
Surface form
PPOS, Surface form
CPOS, Surface form
Bigram features（left context features/right context features）
Surface form / PPOS, Surface form
PPOS, Surface form / Surface form
Tail POS, Surface form / Head POS, Surface form
Surface form / Head POS
Tail POS / Head POS
Tail POS / Surface form

表 4 単語の品詞精度
Table 4 Per token accuracy.

Dictionary	Accuracy
(a) Original POS	97.54
(b) Revised POS	97.56
(c) BIO POS	97.32
(d) Splitting MWE	97.62

表 5 MWE の認識精度
Table 5 Recognition accuracy of MWE.

Dictionary	Precision	Recall	F-measure
Baseline	78.79	80.26	79.51
(c) BIO	92.81	90.90	90.18
(d) MWE	95.75	97.16	96.45

表 6 MWE の認識エラー数
Table 6 Recognition error of MWE.

Error types	# of errors
False recognition	33
Unrecognizable	19
POS errors	15
Misrecognition	2

より精度が向上したと考えられる．さらに，(b) Revised POS よりも (d) Splitting MWE の方が 0.06%精度が高くなっている．これより，各単語ごとに直接品詞タグ付けを行うよりも，MWE を考慮した品詞タグ付けを行い，その後，MWE と認識された語の各構成要素の品詞に分解し，分解した各構成要素の品詞を用いる方が良い結果が得られることがわかった．

次に MWE の認識精度を表 5 に示す．表 5 では，(d) MWE の $F-measure$ が baseline と比べて約 17%，(c) BIO と比較して約 6%精度が向上していることがわかる．これより，baseline や既存の辞書よりも今回構築した辞書を用いる方が高い認識精度を得ることがわかった．

MWE の認識のエラーには大きく分けて以下の 4 種類のエラーが存在した．

- システムが誤って MWE と認識してしまったもの
- システムが MWE と認識できなかったもの
- システムが MWE と認識はしたが，誤った品詞タグ付けを行ったもの
- システムが異なる MWE と認識してしまったもの

表 6 に MWE の各認識エラーの数を示している．

誤って MWE と認識したものに「a bit/RB」などがある．この「a bit/RB」の例を図 3 に示す．これらの MWE は MWE とリテラルな用法の 2 通りの解釈が可能であり，リテラルな用法を MWE 用法と誤って認識したと考えられる．全体で 33 回の誤りが存在したが，MWE の種類 (type) 数で見ると 23 種類であった．中でも 18 種類の MWE はただ 1 回のみ間違ったものであり，この認識結果はまれなものであったと考えられる．実際に，「a bit/RB」は図 3 の文でのみ間違っていた．しかし，残りの 5 種類の MWE は，複数回間違いを繰り返していた．

次に，システムが MWE と認識できなかったものに「in black and white/RB」などがある．これらの MWE が認識できなかった理由は 2 通りあり，まず 1 つはトレーニングデータに出現していない，もしくは出現頻度が少ないことが原因であると考えられる．これには「after all/RB」などが該当する．（図 3 では「after all/RB」も認識できていない．）認識できなかった大半の MWE がこの理由だと考えられるが，これらとは対象的に出現頻度の多い「so far」なども，まれではあるが認識できていないことがあった．

gold: ···· who/WP after all/RB is/VBZ really/RB a/DT bit/JJ player/NN on/IN the/DT stage/NN ····

system: ···· who/WP *after/IN *all/DT is/VBZ really/RB *a bit/RB player/NN on/IN the/DT stage/NN ····

図 3 「after all/RB」と「a/DT bit/JJ」の例
Fig. 3 Example of "after all/RB" and "a/DT bit/JJ".

これらは，上述したようにMWEとリテラルな用法の2通りの解釈が可能なMWEであり，システムがリテラルな用法であると認識したためである．

システムがMWEとして認識したが，誤った品詞タグ付けを行ったものに「how much」や「as much」などがある．これは，システムの品詞タグ付けの間違いであり，MWE以外の単語の場合と同様に，品詞の推定にはさらに多くの情報や工夫が必要だと考えられる．

また，異なるMWEとして認識したものは次の2通りであった．

- 「quite a few/RB」を「quite/RB a few/PRP」と間違って認識
- 「the hell out of/RB」を「the hell/RB out of/IN」と間違って認識

これらのMWEもシステムが認識できなかったMWEと同様に「quite a few/RB」と「the hell out of/RB」がトレーニングデータに出現していないことが原因であると考えられる．

6. おわりに

本稿ではMWE辞書を構築し，品詞タグ付けの精度とMWEの認識精度の評価を行った．実験の結果，品詞タグ付け，MWEの認識精度において，既存の辞書やこれまで提案された手法を用いるよりも高い精度を得ることを示した．ただし，1，2節で述べた通り本稿のMWE辞書には形容詞や名詞，動詞などは含んでおらず，また，「a (special) kind of」のように修飾語を内部に含み得るMWEを含んでいない．

今後の課題として，本稿で構築したMWE辞書に含んでいないMWE辞書を構築することや，MWEを考慮した依存構造解析や構文解析などを行うことが考えられる．

参考文献

[1] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. "Multiword Expressions: A Pain in the Neck for NLP," *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics 2002*, pp.1-5, 2002.

[2] Eric Laporte and Stavroula Voyatzi. "An Electronic Dictionary of French Multiword Adverbs," *Language Resources and Evaluation Conference. Workshop Towards a Shared Task for Multiword Expressions*, 2008.

[3] Eric Laporte, Takuya Nakamura, and Stavroula Voyatzi. "A French Corpus Annotated for Multiword Nouns," *Language Resources and Evaluation Conference. Workshop Towards a Shared Task for Multiword Expressions*, 2008.

[4] Nicole Gregoire. "Design and Implementation of a Lexicon of Dutch Multiword Expressions," *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, 2007.

[5] 松吉俊, 佐藤理史, 宇津呂武仁, "日本語機能表現辞書の編纂," **自然言語処理**, 14(5), pp.123–146, 2007.

[6] Kosho Shudo, Akira Kurahone, and Toshifumi Tanabe. "A Comprehensive Dictionary of Multiword Expressions," *49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[7] Joakim Nivre and Jens Nilsson. "Multiword Units in Syntactic Parsing," *Workshop on Methodologies and Evaluation of Multiword Units in Real-world Applications*, LREC-2004, 2004.

[8] Ioannis Korkontzelos and Suresh Manandhar. "Can Recognising Multiword Expressions Improve Shallow Parsing?" *11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

[9] Mark Alan Finlayson and Nidhi Kulkarni. "Detecting Multi-Word Expressions improves Word Sense Disambiguation," *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, ACL2011, 2011

[10] Matthieu Constant, Anthony Sigogne, and Patrick Watrin. "Discriminative Strategies to Integrate Multiword Expression Recognition and Parsing,' *50th Annual Meeting of the Association for Computational Linguistics*, 2012.

[11] Matthieu Constant and Anthony Sigogne. "MWU-Aware Part-of-Speech Tagging with a CRF Model and Lexical Resources," *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, ACL2011, 2011

[12] Spence Green, Marie-Catherine de Marneffe, John Bauer and Christopher D. Manning. "Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French," *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011.

[13] Christopher D. Manning. "Part-of-speech tagging from 97% to 100% is it time for some linguistics?," *Computational Linguistics and Intelligent Text Processing*, pp. 171-189, 2011.