A Comprehensive Japanese MWE Lexicon : JMWEL

Masahito Takahashi[1], Toshifumi Tanabe[2], Jack Halpern[3] and Kosho Shudo[4]

[1]Kurume Institute of Technology, JAPAN

[2]Fukuoka University, JAPAN

[3]CJKI Co. , JAPAN

[4]Fukuoka University, emeritus, JAPAN

[1]taka@kurume-it.ac.jp

[2]tanabe@fukuoka-u.ac.jp

[3]jack@cjki.org

[4]viggo_ksf@jcom.home.ne.jp

Abstract

JMWEL (Japanese MWE Lexicon) is a comprehensive lexicon of Japanese Multiword Expressions (MWEs) with a rich set of grammatical attributes fine-tuned for phrase-based processing of a wide range of Japanese

documents. It has about 160,000 MWE lemmas covering almost every kind of linguistically idiosyncratic but commonly used Japanese phrases, e.g., idioms, quasi-idioms, collocations, quasi-collocations, clichés, quasi-clichés, institutionalized phrases, proverbs and old sayings, excepting technical terms in specialized fields or named entities. JMWEL consists of sixteen sub-lexicons reflecting their distinctive features. The comprehensiveness of the collected MWEs and the detailed morpho-syntactic information given to each MWE which may include internal modifiers are notable features of JMWEL. In this paper, we introduce the newest version of JMWEL.

Keywords: Multiword Expression (MWE), Phrase-based NLP, Phrase-Based Machine Translation (PBMT), Phrase-Based Statistical Machine Translation (PBSMT), Neural Machine Translation (NMT), Lexical bundles, Formulaic language, Construction grammar, Phraseology

1. Introduction

Neural machine translation (NMT) based on deep learning technology is

rapidly expanding its scope of use. However, even state-of-the-art NMT has some difficulty dealing with the semantic non-compositionality of multiword expressions (MWEs) such as idioms or clichés. Since the beginning of this century, the importance of these idiosyncratic expressions has received particular attention. These expressions have now been widely studied in various correlated notions such as Multiword Expression (MWE) (Sag et al., 2002), Formulaic Language (Corrigan et al., 2009), Phraseology (Cowie, 1998), Lexical Bundles (Biber et al., 1999), and Construction Grammar (Fillmore et al., 1988). Many researchers have pointed out that these expressions are used with considerable variety and frequency in everyday language.[1] However, the automatic extraction of MWEs from corpora has not been sufficiently successful and the overall picture of the MWE phenomena is still unclear.

The authors of this paper have developed a Japanese MWE Lexicon (JMWEL) (Tanabe et al., 2014). JMWEL, which now lemmatizes about 160,000 MWEs, covers the majority of MWEs appearing in ordinary

[1] Sag et al. (2002) pointed out that about 41% of the entries in WordNet 1.7 are MWEs.

Japanese written articles of newspapers, novels, journals, etc. The research on JMWEL started in the late 1960s with the intention to partly incorporate the human language understanding process, i.e., the "word-chunking" mechanism into Natural Language Processing (NLP).

JMWEL is now in the mature stage of its development. In this paper, we introduce the current status of JMWEL, primarily focusing on its organization and encoding schema.

In Chapter 2, we describe the procedure used to collect MWEs for JMWEL. Chapter 3 describes the organization of JMWEL, i.e., what sub-lexicons comprise JMWEL. Chapter 4 describes the information encoded in JMWEL. We briefly discuss some applications of JMWEL in Chapter 5. Chapter 6 is a short survey of the related work and we give some concluding remarks in Chapter 7.

2. MWE lemma

For the development of JMWEL, MWEs were extracted from a wide range of documents such as newspaper, magazine, novels, essays, encyclopedias

and dictionaries, mainly through the editor's introspection based on two criteria. i.e., non-compositionality and probabilistic idiosyncrasy.

2.1 Non-compositionality

If the word sequence $w_1w_2...w_n$ has a coherent syntactic, semantic or pragmatic function and $w_1w_2...w_{i-1}x_iw_{i+1}...w_n$, which replaces some word w_i ($1 \leq i \leq n$) in $w_1w_2...w_n$ with its synonym or quasi-synonym x_i, ceases to make sense, has a meaning that is entirely different from the original, or is unnatural, then we approximate that the sequence $w_1w_2...w_n$ is a non-compositional MWE.[2] This judgment is essentially made by introspection.

For example, 赤-の-他人 *aka-no-tanin* (lit. “red stranger”), which when properly translated means “complete stranger”, is judged to be a non-compositional MWE because the expression 深紅-の-他人 *sinku-no-tanin* (lit. “crimson stranger”) produced by replacing 赤-の *aka-no* to its synonym 深紅-の *sinku-no* makes no sense.

Some examples of non-compositional MWEs are shown in Table 1.

[2] The maximum value of n in the MWEs recorded in JMWEL is 18.

Table 1: Examples of non-compositional MWEs

Types	Examples
Expressions with semantic non-compositionality	赤-の-他人 *aka-no-tanin* (lit. “red stranger”) “complete stranger”), 顔-を-売る *kao-wo-uru* (lit. “to sell one’s face”) “to make oneself known to the public”
Expressions with inadequate or unclear morpho-syntactic composition	と-は-いえ *to-ha-ie* “however”, ありがとう *arigato-u* “Thank you”
Some support verb constructions	批判-を-加える *hihan-wo-kuwaeru* (lit. “to add criticism”) “to criticize”, 計画-を-立てる *keikaku-wo-tateru* (lit. “to stand a plan”) “to make a plan”
Some compound words	打ち-拉がれる *uti-hishigareru* (lit. “to be hit and smashed”) “to become depressed”, 袋-叩き *fukuro-dataki* (lit, ”bagging and beating

	on”) “beating on”
Four-kanji-character-idioms	一生-懸命 *issyou-kenmei* (lit. “all through life and earnestly”) “with all one’s might”, 一心-不乱 *isshin-furan* (lit, “single-mindedly and composedly”) “being single minded”
Idiomatic metaphorical expressions	命-の-限り *inochi-no-kagiri* (lit. “limitation of one’s life”) “as long as one lives”, 血-の-雨-が-降る *chi-no-ame-ga-furu* (lit. “to be showered with blood” “There is bloodshed”

2.2 Probabilistically idiosyncratic MWEs

Expressions that have strong probabilistic affinities among component words are selected as probabilistically idiosyncratic MWEs.

If the word sequence $w_1w_2...w_n$ has a coherent syntactic, semantic, and pragmatic function and the forward-transition-probability $p_f(w_i|w_1...w_{i-1})$ for

some word w_i ($2 \leq i \leq n$) in $w_1w_2...w_n$ or the backward-transition-probability $p_b(w_j|w_{j+1}...w_n)$ for some word w_j ($1 \leq j \leq n-1$) in $w_1w_2...w_n$ is relatively high, $w_1w_2...w_n$ is judged to be a probabilistically idiosyncratic MWE.

For example, we consider ぐっすり-眠る *gussuri-nemuru* "to have a good sleep" to be an MWE because it is clear that p_f (眠る|ぐっすり) is very high. Although this criterion was judged by introspection, the validity of the judgment was statistically confirmed using a twenty-billion-sentence web-corpus (Tanabe et al., 2014) (Kudo & Kazawa, 2009).

Some examples of probabilistically idiosyncratic MWEs are shown in Table 2.

Table 2: Examples of probabilistically idiosyncratic MWEs

Types	Examples
Expressions with particularly strong affinity among component words	風前-の-灯 *fuuzen-no-tomosibi* (lit. "light in front of the wind") "a candle flickering in the wind", 手-を-拱く *te-wo-komaneku* "to fold one's arms"

Aphorisms, proverbs, or clichés	急が-ば-回れ *isogaba-maware* (lit. “Make a detour when in a hurry”) “more haste, less speed”, 初心-忘る-可から-ず *shoshin-wasuru-bekara-zu* “Don't forget your first resolution”
Collocations with onomatopoeic or mimetic words	キャンキャン-鳴く *kyankyan-naku* “to yelp”, ポッカリ-と-空く *pokkari-to-aku* “to be empty” “to gape”
Expressions with strong affinity among component words	肩-の-荷-を-下ろす *kata-no-ni-wo-orosu* (lit. “to take a load off one's shoulders” “take a load off one's mind”, メリハリ-の-利いた *merihari-no-kiita* “explicit”
Fixed phrases to describe particular concepts (Institutionalized phrases)	女流-作家 *jyoryuu-sakka* “a woman writer”, 機械-翻訳 *kikai-honyaku* “machine translation”

2.3 Distribution of MWEs in JMWEL

We examined about 2,000 MWEs randomly extracted from JMWEL and found that about 38% of the MWEs have non-compositionality, about 92% have strong probabilistic affinity among component words and about 30% have both of these properties. For example, 油-を-売る *abura-wo-uru* (lit. “to sell oil”) “to waste time in idle” is a semantic non-compositional MWE, ぐっすり-眠る *gussuri-nemuru* “sleep soundly” is a MWE with strong probabilistic affinity between component words, and 肩-の-荷-を-下ろす *kata-no-ni-wo-orosu* (lit. “to take a load off one's shoulders”) “to take a load off one's mind” is a MWE which has both properties.

3. Organization of JMWEL

MWEs as a whole are so diverse that the comprehensive realization of an MWE lexicon depends on how we treat properly classified subsets of the phenomena. JMWEL has been compiled and maintained as a collection of eleven sub-lexicons created on the basis of the expression’s grammatical functions and five sub-lexicons created cross-sectionally to the grammatical sub-lexicons based on topics. The former eleven sub-lexicons comprise the

basic JMWEL, i.e., every MWE is included in some of them.

The following is a brief explanation of sub-lexicons and the recorded expressions.

3.1 Sub-lexicons compiled based on grammatical functions

MWEs are classified by how they syntactically function in context, i.e., as the non-terminal category in the Phrase-Structure-Grammar (PSG)-framework.

(1) Sub-lexicon of nominal MWEs

This contains about 28,600 MWEs each of which is a nominal phrase or a noun-predicate-sentence.

(Example) 真っ赤な-嘘 *makkana-uso* (lit. “a crimson lie”) “complete lie”

(2) Sub-lexicon of verbal MWEs (class1)

This contains about 37,000 verbal MWEs each of which is a verb phrase of the form: NOUN - α - VERB, where α is a case-marking-particle が *ga*, を *wo* or に *ni* which mark the subjective-, objective-, and dative-cases,

respectively.[3]

(Example) 脛-を-齧る *sune-wo-kajiru* (lit. “to bite ….’s shin”) “to depend on the financial support of …”

(3) Sub-lexicon of verbal MWEs (class2)

This contains about 37,100 MWEs each of which is a verb phrase of the form: NOUN - α - VERB, where α is a case-marking-particle other than が *ga*, を *wo*, and に *ni*, or of the form: NOUN $_1$ -α_1 - NOUN $_2$ - α_2 - ….. - NOUN $_n$ - α_n - VERB, where α_i (i≦n) means a case-marking-particle and n≧2.

(Example) 目-に-物-を-見せる *me-ni-mono-wo-miseru* (lit. “to show things to …’s eyes”) “to teach … a good lesson”

(4) Sub-lexicon of verbal MWEs (class3)

This contains about 4,200 non-compositional compound verbs.

(Example) 空-とぼける *sora-tobokeru* (lit. “to be blurred with the sky”) “to pretend not to know”

(5) Sub-lexicon of adjectival MWEs

[3] This is the simplest and fundamental form of Japanese sentence.

This contains about 5,800 MWEs each of which is an adjective phrase.

(Example) 腰-が-低い *kosi-ga-hikui* (lit. “to have a low waist”) “to be quite modest”

(6) Sub-lexicon of adjective-verbal MWEs

This contains about 2,800 MWEs each of which is an adjective-verb phrase.

(Example) 傍若-無人 *boujaku-bujinn* (lit. “boujaku-bujin”) “unmannered, haughty behavior”

(7) Sub-lexicon of adverbial MWEs

This contains about 17,600 expressions each of which is an adverbial phrase.

(Example) かみ砕い-て *kamikudai-te* (lit. “by chewing”) “in detail and clearly”

(8) Sub-lexicon of adnominal MWEs

This contains about 17,100 MWEs each of which is an adnominal phrase.

(Example) 地-に-足-の-着い-た *chi-ni-ashi-no-tui-ta* (lit. “with feet on the ground”) “firm and steady”

(9) Sub-lexicon of discourse-marking MWEs

This contains about 1,900 discourse-marking or sentence-adverbial MWEs,

each of which is used at the beginning of a sentence to make the dialogue or discourse smooth.

(Example)予め-断っ-て-おく-けど *arakajime-kotowat-te-oku-kedo* (lit. “Though I notice you in advance”) “I want you to know in advance”

(10) Sub-lexicon of postpositional function-MWEs

This contains about 2,700 functional MWEs, each of which acts as a case-marking-, adverbial-, or a connective-particle. This kind of MWE, indicates a semantic relationship between words.

(Example) を-良い-こと-に *wo-yoi-koto-ni* (lit. “making ….. good thing”) “for the reason that ….”

(11) Sub-lexicon of post-predicative function-MWEs

This contains about 5,200 functional MWEs, each of which acts as an auxiliary-verb or a sentence-ending-particle. This kind of MWE attaches the modality, polarity, tense, aspect, or pragmatic meaning in a broad sense to a predicate.

(Example) て-くださる-と-有難かっ-た-ん-です-が *te-kudasaru-to-arigatakat-ta-n-desu-ga* (lit. “If you ….., it was thankful for me”) “I hoped

that you……but unfortunately you didn't do it"

3.2 Sub-lexicons organized by topics.

JMWEL contains five sub-lexicons organized by topics. They are cross-sectional to sub-lexicons organized by grammatical function explained in 3.1.

(1) Sub-lexicon of standard idioms

This covers about 4,900 idioms, which are found in commercial Japanese idiom dictionaries for human daily use.

(example) 手-を-切る *te-wo-kiru* (lit. "to cut …'s hand") "to cut a connection with….."

(2) Sub-lexicon of aphorisms, proverbs, old-sayings, clichés, etc.

This covers about 4,100 expressions each of which is classified as an aphorism, proverb, old-saying or other cliché.

(example) 背-に-腹-は-代え-られ-ない *se-ni-hara-ha-kae-rare-nai* (lit. "back can not be replaced by stomach") "to have no choice but …."

(3) Sub-lexicon of four-kanji-character-idioms

This covers about 3,200 four-kanji-character-idioms.

(Example) 四面-楚歌 *simen-soka* (“to be surrounded by So’s songs on four sides”)[4] “to be surrounded by enemies on all sides”

(4) Sub-lexicon of collocations with onomatopoeia or mimetic words

This contains about 41,900 collocations that feature about 2,000 onomatopoeia or mimetic-words. The Japanese language is characterized by an abundance of onomatopoeia or mimetic words, most of which have strong probabilistic affinities with particular predicates.

(Example) バリバリ-働く *BariBari-hataraku* (lit. “to work BariBari”) “to work quite energetically”

(5) Sub-lexicon of syntactically incomplete idioms

This covers about 470 expressions each of which is not a complete phrase in the sense of PSG-framework, i.e., it has some constituents missing but is sometimes used as a syntactical unit.

(Example) 病-は-気-から *yamai-ha-ki-kara* (lit. “illness …. from bad mental state”) “illness comes from bad mental state”

[4] “So” is a name of an ancient Chinese dynasty.

4. Data Entries in JMWEL

Every sub-lexicon of JMWEL is created in tabular form using Microsoft Excel and saved as a xlsx file. Each line of the file corresponds to an MWE and each column of the file to its attribute information. In view of the diversity of information contained in the actual JMWEL, we explain here only the major information encoded in JMWEL.

A sample image of the data entries is illustrated in Fig.1, whose example MWE is described in 4.8.

Column A type label	Column B lemma	Column C segmentation into a mopheme string	Column D notational variant	Column E syntactic function	Column F morpho-syntactic structure
verba (class1)	**めにあう** *meniau*	**め-に-あう** *me-ni-au*	**目-に-会う** *me-ni-au*	**VP**	**[[*Nni]*V30]**

Column G forward-context-condition	Column H backward-context-condition	Column I final predicate	Column J necessity of inflection	Column K interpretation
<adnom modifier>	**…**	**会う** *au*	**…**	**…**

Fig. 1 Sample image of data entries in JMWEL

4.1 Type Label

A type label of MWE is noted in Column A, e.g., “verbal (class1)”, in the case of the verbal (classs1) sub-lexicon.

4.2 Lemma

Column B contains the lemma of the MWE given as a sequence of hiragana (Japanese phonetic symbols) without pauses (spaces). The inflectional words at the end of verbal or adjectival MWEs are recorded in the plain-form in JMWEL.

4.3 Constituent Morphemes

The lemma is segmented into a sequence of morphemes in Column C. Every constituent morpheme is written in hiragana and the delimiter between morphemes is a hyphen “-”. A morpheme in Japanese is one of a word, a prefix, a suffix, or a word-element.[5]

4.4 Orthographic Variants

[5] A word-element is a bound morpheme which is noted by a single kanji and less productive than a prefix or suffix in word forming.

The Japanese language uses ideographic characters (kanji) and phonetic characters (hiragana and katakana) which have the same phonetic values. In addition, there are many same-sound/similar-meaning/different-kanji words in Japanese. Furthermore, a certain hiragana in the stem of some inflectional word, called “okurigana”, can be omitted in the sentence when a kanji is used for the word. Thus, MWEs can often have diverse orthographic variants. We have tried to present notational variants as much as possible to make JMWEL applicable to a wide range of Japanese texts. Variants are encoded using the closure-free regular-expression in Column D.

For example, 行(な/ε)う *okonau* “to act” indicates the possibility of 行なう and 行う and (馬鹿/莫迦/バカ)-に-する *baka-ni-suru* “to look down on” indicates 馬鹿-に-する, 莫迦-に-する, and バカ-に-する.[6]

4.5 Syntactic Function

The syntactic function of an MWE, i.e., how the MWE syntactically works in the context, is presented in Column E by a non-terminal symbol in PSG-

[6] “ε” is a null symbol.

framework, e.g., “AdnP” which means an adnominal phrase in case of adnominal sub-lexicon.

4.6 Morpho-syntactic Structure

The morpho-syntactic structure of an MWE is given in Column F based on the dependency grammar formalism for Japanese. The dependency tree structure is encoded in a parenthesized linear representation that we name a “Dependency Marker (DM)”, which is a formula defined recursively as follows:

(Base1) : A Part-of-Speech (POS)-code given in an upper-case English letter (possibly along with digits to indicate the inflectional variant) for a content-word, a prefix, a suffix, or a word-element is a DM whose head is itself.

(Base2) : A spelling of a function-word given as a string of lower-case English letters is a DM whose head is itself.

(Recursion Step 1) : If α and β are DMs, then [αβ] is a DM whose head is the

head of β. [αβ] means that the word denoted by the head of α modifies (depends on) the word denoted by the head of β.

This definition reflects the head-final, nested-structure principle of Japanese syntax.[7] We apply this formalization to the non-dependency concatenation of two words, for simplicity.

For example, a DM of a verbal MWE 顔-を-揃える *kao-wo-soroeru*, (lit. “to align faces"), “to get together” will be [[Nwo]V30], because, firstly, 顔 and を constitute an adverbial postpositional phrase 顔-を and, secondly, 顔-を depends on 揃える. Here, three component words 顔 *kao* “face”, を *wo* and 揃える *soroeru* “to align” are assigned a POS-code N which means a noun, a spelling of an objective-case-marking-particle ‘wo’, and a POS-code V30 which means a plain-form of a verb, respectively. The structure is illustrated in Fig. 2.

[7] The parallel structure in MWE is consistently incorporated in the DM formalism, but we omit their discussion in this paper.

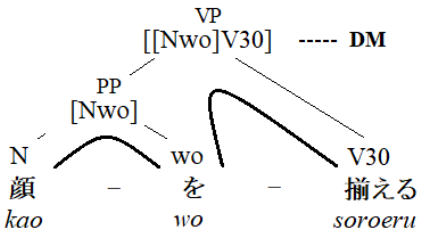

Fig.2 Dependency structure of a verbal MWE 顔-を-揃える

“to get together” given by its DM [[Nwo]V30]

4.7 Internal Modification

The definition of DM is significantly extended for actual JMWEL by adding the following clauses:

(Recursion Step 2) : If α is a DM, then *α is a DM whose head is the head (governing content-word) of α. *α means that the word denoted by the head of α may have an additional modifier located in the position of *.

Based on this definition, the example verbal-MWE 顔-を-揃える is actually given a DM [[*Nwo]*V30] in Column F. Here, *N and *V30 in the DM mean that a noun 顔 *kao* “face” (N) and a verb 揃える *soroeru* “align” (V30) might be modified by a left-hand adjacent adnominal-phrase such as

元気-な *genki-na* “cheery” and by a left-hand adjacent adverbial phrase such as 皆-が *mina-ga* “everyone”, respectively. This yields a naturally extended discontinuous MWE 元気-な-顔-を-皆-が-揃える *genki-na-kao-wo-mina-ga-soroeru* “everyone gets together cheerily”. Fig.3 illustrates the structure of this MWE.

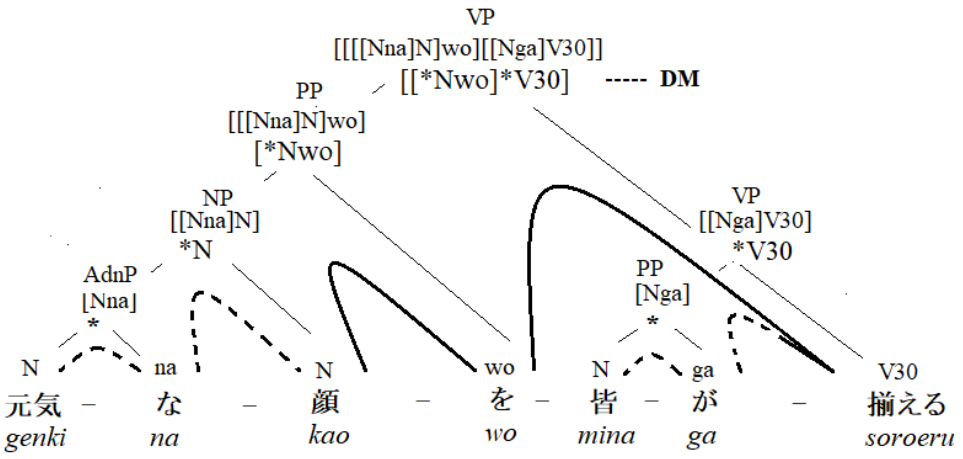

Fig.3 Dependency structure of a discontinuous MWE 元気-な-顔-を-皆-が-揃える “Everyone gets together cheerily.” which has additional modifiers in asterisk’s positions in its DM [[*Nwo]*V30]

With this indication of the possibility of additional internal modification, JMWEL potentially covers numerous generalized MWEs which might be rare or unseen in the training corpus.

In addition, the syntactic flexibility/rigidity of MWE becomes scalable in the sense that the fewer asterisks the DM includes, the more syntactically rigid the MWE is. For example, while the DM of a syntactically fixed MWE 手-に-汗-を-握る *te-ni-ase-wo-nigiru* (lit. “to grasp tears in hands”) “to be in breathless suspense” is [[Nni][[Nwo]V30]] that contains no asterisk, the DM for a free-word-combination 箱-に-物-を-入れる *hako-ni-mono-wo-ireru* “put a thing into a box” will be [[*Nni]*[[*Nwo]*V30]] which indicates every component content-word could have its modifiers in the positions of the asterisk.

The information in Column F might contribute to establishing an elaborate “syntactic typology” of Japanese MWEs.

4.8 Forward Context Condition

Some MWEs require a specific modifier in the left-hand side position in the sentence. This kind of requirement is specified in Column G. For example, a verbal MWE 目-に-会う *me-ni-au*, (lit. “to meet eyes”) “to have some negative experience” requires an adnominal modifier such as 悲しい *kanasii* “sad” in the left-hand adjacent position to make a correct verbal phrase 悲しい-目-に-会う *kanasii-me-ni-au* “to have a sad experience”. The requirement is specified by a category code such as <adnom. modifier>. There are about thirty kinds of forward context requirements in JMWEL.

4.9 Backward Context Condition

Some MWEs require their specific heads in the right-hand side position in the sentence. This kind of requirement is specified in Column H. For example, adverbial MWE 何-一つ *nani-hitotu* (lit. “a single thing”) “(not)…at all” requires its head phrase which includes a negation auxiliary-verb (or a negation adjective) ない *nai* “not” in the right-hand side position to yield such a correct phrase as 何-一つ-苦痛-を-与え-ない *nani-hitotu-kutuu-wo-atae-nai* “do not give pain at all”. This is stipulated by a code

<negation> in Column H. There are about three hundred backward requirement codes in JMWEL. Some of the long-distance-dependency phenomena in Japanese are dealt with these conditions.

4.10 Inflection

In order to identify MWEs in texts or corpora, it is necessary to create in advance all inflected-forms of inflectional MWEs in JMWEL because the lemma of inflectional MWE is given only in plain-form in JMWEL. For the user's ease of this inflected-form creation, the inflectional word in the final position of each verbal or adjectival MWE is re-cited in Column I.

MWEs with an exceptionally high degree of fixedness such as sayings, proverbs, clichés, and some archaic expressions are rarely used in their inflected-forms. This type of MWE is accompanied with a statement "inflection-not-required" in column J. This field can be regarded as a sign for the "maximum fixedness" of the expression.

4.11 Interpretation

Some idioms, proverbs, sayings, and clichés are annotated interpretations in Column K for the convenience of JMWEL users.

5. Applications

There is an advantage in using the semantically coherent MWEs as processing units in NLP. Suppose that three MWEs 手-に-付か-ず *te-ni-tsuka-zu* "as…is unable to get down to", 散歩-に-出る *sanpo-ni-deru* "to go for a walk" and こと-に-する *koto-ni-suru* "to decide to….." are given in JMWEL, their syntactic functions are noted in Column E as AdvP (adverbial phrase), VP (verb phrase) and Aux (auxiliary-verb equivalent), respectively, and their syntactic structures DMs are specified in Column F as [[Nni][V12zu]], [[Nni]V30] and [[Nni]suru], respectively. Also suppose that a postpositional phrase marked with a case-marking-particle が *ga* like 仕事-が *sigoto-ga* is required by a forward context condition in Column G of 手-に-付か-ず *te-ni-tsuka-zu*. Then, the input sentence 彼-は-仕事-が-手-に-付か-ず-散歩-に-出る-こと-に-し-た *kare-ha-shigoto-ga-te-ni-tsuka-zu-sanpo-ni-deru-koto-ni-shi-ta* "he couldn't work anymore and decided to go

for a walk” could be parsed as shown in Figure 4, using regular syntax-rules, for example. Roughly, the fifteen-word input sentence is processed as if it were an eight-word sentence by chunking each MWE in this example. That is, MWE chunking possibly reduces the number of useless syntactic substructures in the parsing process and, at the same time, provides semantically plausible output.

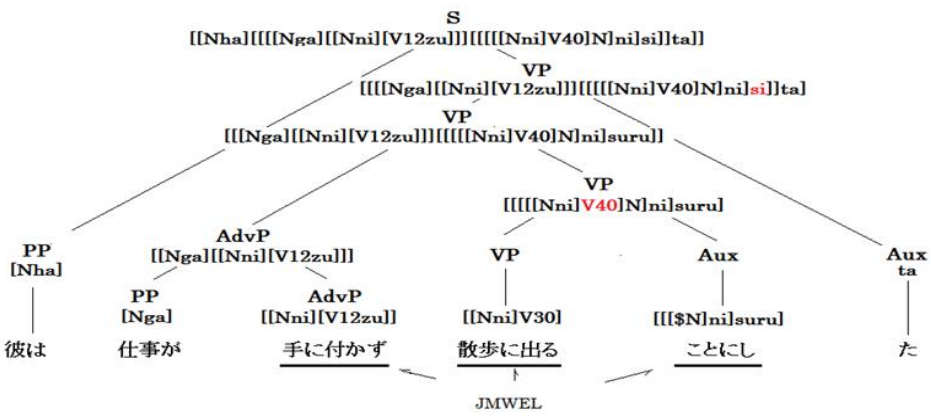

Fig.4 Sample image of the rule-based MWE-oriented parsing

If the above three MWEs are accompanied by frames for English production, such as “as SUB is unable to get down to doing SUB's N”, “go out for a walk” and “decide to”, respectively, then a semantically plausible Japanese-to-

English translation will be performed as shown in Figure 5, in a manner parallel to the analysis shown in Figure 4. Thus, it will be fruitful in principle to apply JMWEL's data to PBMT (Phrase-Based Machine Translation), or PBSMT (Phrase-Based Statistical Machine Translation).

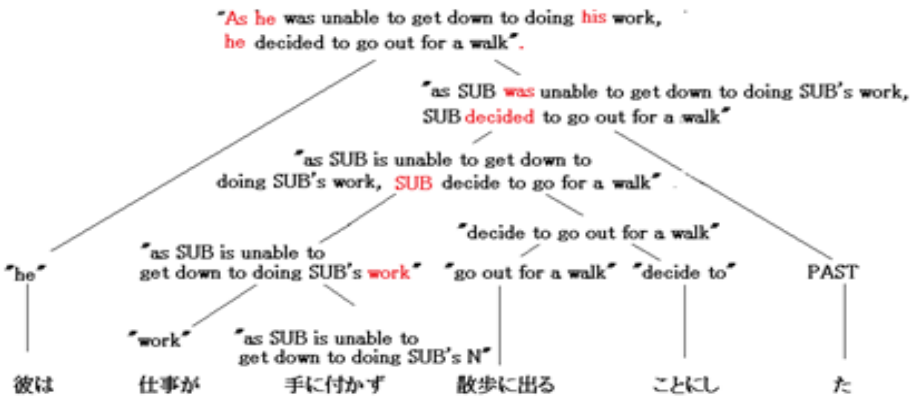

Fig.5 Sample image of the rule-based MWE-oriented Japanese-to-English MT

Although recent NMTs have made rapid progress specifically in dealing with frequent expressions, such as function-MWEs or some frequent clichés in Japanese, MWEs as a whole have not yet been fully addressed. In particular, there are many problems with notational variety of expressions and

ambiguous MWEs that have both compositional and non-compositional meanings. The manual annotation of large corpora with ambiguous MWEs, distinguishing their non-compositional usage from compositional usage is necessary for further progress of NMT, i.e. for more effective machine learning. JMWEL will contribute to this annotation on the one hand, and to the discovery of unseen MWEs in the corpus using the latent syntactic variants framed by the information on the internal modification given in JMWEL, on the other.

6. Related Work

Since Sag et al. (2002) pointed out that the treatment of idiosyncratic MWEs is a key problem in NLP, many attempts for the automatic MWE-identification or MWE-discovery in texts or corpora have been made, applying various rule-based or corpus-based stochastic methods to distinctive type of MWEs, i.e., the named entity, verb-particle construction, light-verb-construction, compositional verbal MWE, non-compositional verbal MWE, discontinuous MWE, seen- or unseen-MWE in corpora. However, it remains

uncertain which frameworks proposed in these studies are sufficiently effective for the variety of the MWE phenomena.

Recently, there have been active discussions on how to handle MWEs in machine translation. Ramisch (2018) pointed out that it is crucial to identify MWEs having non-compositionality in the source text before translating them. He also noted that the discontinuity of MWEs is a problem for both identification and translation, and structural methods based on trees and graphs will give solutions for the problem. Savary et al. (2019) highlighted the necessity of the syntactic lexicon connected to the MWE-identification or MWE-discovery task, stressing the problem of discontinuous MWEs and MWEs unseen in the training corpus. These indications will be crucial for the forthcoming advanced MWE processing in NLP.

7. Concluding Remarks

JMWEL is a handcrafted, NLP-oriented syntactic lexicon of Japanese MWEs. The notable features of JMWEL are summarized as follows:

(1) A large variety of MWE categories are covered.

(2) A large orthographic variety of each MWE is covered.

(3) The detailed morpho-syntactic structure of each MWE is encoded.

(4) A large syntactical variety of each MWE is potentially covered by indicating the possibility of internal modification for each MWE.

Acknowledgements

We would like to thank the late professor Toshikiko Kurihara who inspired our current research in the late 1960s. We are also grateful to everyone who assisted in the development of JMWEL and to Richard Lee for his advice on matters of English style.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (Eds.) (1999). Longman Grammar of Spoken and Written English. Harlow: Pearson Education Limited.

Corrigan, R., et al. (Eds.) (2009). Formulaic Language, vol.1 distribution and historical change. John Benjamins Publishing Company.

Cowie, A. P. (Ed.) (1998). Phraseology: Theory, Analysis, and Applications. (Oxford Studies in Lexicography and Lexicology). Clarendon Press Oxford.

Fillmore, C., Kay, P., & O'Connor, M. C. (1988). Regularity and Idiomatical Grammatical Construction: The Case of Let Alone. Language 64 (pp.501-538).

Kudo, T., & Kazawa, H. (2009). Japanese Web N-gram Version 1. Linguistic Data Consortium 2009. T08. Philadelphia.

Ramish, C. (2017). Putting the Horses Before the Cart: Identifying Multiword Expressions Before Translation. The EUROPHRAS 2017 (MUMTTT 2017). Proceedings (pp.69-84). Invited Talk. London.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: a pain in the neck for NLP. The 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002). Proceedings (pp.1-15). Mexico City.

Savary, A., Cordeiro, S. R., & Ramisch, C. (2019). Without lexicons, multiword expression identification will never fly: A position statement.

The EUROPHRAS 2019. Proceedings (pp.79-91). Malaga.

Tanabe, T, Takahashi, M., & Shudo, K. (2014). A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing. Computer Speech and Language. 28-6 (pp.1317-1339). Elsevier.