A Comprehensive Dictionary of Multiword Expressions

Kosho Shudo[1], **Akira Kurahone**[2], **and Toshifumi Tanabe**[1]
[1]Fukuoka University, Nanakuma, Jonan-ku, Fukuoka, 814-0180, JAPAN
{shudo,tanabe}@fukuoka-u.ac.jp
[2]TechTran Ltd., Ikebukuro, Naka-ku, Yokohama, 231-0834, JAPAN
kurahone@opentech.co.jp

Abstract

It has been widely recognized that one of the most difficult and intriguing problems in natural language processing (NLP) is how to cope with idiosyncratic multiword expressions. This paper presents an overview of the comprehensive dictionary (JDMWE) of Japanese multiword expressions. The JDMWE is characterized by a large notational, syntactic, and semantic diversity of contained expressions as well as a detailed description of their syntactic functions, structures, and flexibilities. The dictionary contains about 104,000 expressions, potentially 750,000 expressions. This paper shows that the JDMWE's validity can be supported by comparing the dictionary with a large-scale Japanese N-gram frequency dataset, namely the LDC2009T08, generated by Google Inc. (Kudo et al. 2009).

1 Introduction

Linguistically idiosyncratic multiword expressions occur in authentic sentences with an unexpectedly high frequency. Since (Sag et al. 2002), we have become aware that a proper solution of idiosyncratic multiword expressions (MWEs) is one of the most difficult and intriguing problems in NLP. In principle, the nature of the idiosyncrasy of MWEs is twofold: one is idiomaticity, i.e., non-compositionality of meaning; the other is the strong probabilistic affinity between component words. Many attempts have been made to extract these expressions from corpora, mainly using automated methods that exploit statistical means. However, to our knowledge, no reliable, extensive solution has yet been made available, presumably because of the difficulty of extracting correctly without any human insight. Recognizing the crucial importance of such expressions, one of the authors of the current paper began in the 1970s to construct a Japanese electronic dictionary with comprehensive inclusion of idioms, idiom-like expressions, and probabilistically idiosyncratic expressions for general use. In this paper, we begin with an overview of the JDMWE (Japanese Dictionary of Multi-Word Expressions). It has approximately 104,000 dictionary entries and covers potentially at least 750,000 expressions. The most important features of the JDMWE are:

1. A large notational, syntactic, and semantic diversity of contained expressions
2. A detailed description of syntactic function and structure for each entry expression
3. An indication of the syntactic flexibility of entry expressions (i.e., possibility of internal modification of constituent words) of entry expressions.

In section 2, we outline the main features of the present study, first presenting a brief summary of significant previous work on this topic. In section 3, we propose and describe the criteria for selecting MWEs and introduce a number of classes of multiword expressions. In section 4, we outline the format and contents of the JDMWE, discussing the information on notational variants, syntactic functions, syntactic structures, and the syntactic flexibility of MWEs. In section 5, we describe and explain the contextual conditions stipulated in the JDMWE. In section 6, we illustrate some important statistical properties of the JDMWE by comparing the dictionary with a large-scale Japanese N-gram frequency dataset, the LDC2009T08, generated by Google Inc. (Kudo et al. 2009). The paper ends with concluding remarks in section 7.

2 Related Work

Gross (1986) analyzed French compound adverbs and compound verbs. According to his estimate, the lexical stock of such words in French would be respectively 3.3 and 1.7 times greater than that of single-word adverbs and single-word verbs. Jackendoff (1997) notes that an English speaker's lexicon would contain as many MWEs as single words. Sag et al. (2002) pointed out that 41% of the entries of WordNet 1.7 (Fellbaum 1999) are multiword; and Uchiyama et al. (2003) reported that 44% of Japanese verbs are VV-type compounds. These and other similar observations underscore the great need for a well-designed, extensive MWE lexicon for practical natural language processing.

In the past, attempts have been made to produce an MWE dictionary. Examples include the following: Gross (1986) reported on a dictionary of French verbal MWEs with description of 22 syntactic structures; Kuiper et al. (2003) constructed a database of 13,000 English idioms tagged with syntactic structures; Villavicencio (2004) attempted to compile lexicons of English idioms and verb-particle constructions (VPCs) by augmenting existing single-word dictionaries with specific tables; Baptista et al. (2004) reported on a dictionary of 3,500 Portuguese verbal MWEs with ten syntactic structures; Fellbaum et al. (2006) reported corpus-based studies in developing German verb phrase idiom resources; and recently, Laporte et al. (2008) have reported on a dictionary of 6,800 French adverbial MWEs annotated with 15 syntactic structures.

Our JDMWE approach differs from these studies in that it can treat more comprehensive types of MWEs. Our system can handle almost all types of MWEs except compositional compounds, named entities, acronyms, blends, politeness expressions, and functional expressions; in contrast, the types of MWEs that most of the other studies can deal with are limited to verb-object idioms, VPCs, verbal MWEs, support-verb constructions (SVCs) and so forth.

Many attempts have been made to extract MWEs automatically using statistical corpus-based methods. For example, Pantel et al. (2001) sought to extract Chinese compounds using mutual information and the log-likelihood measure. Fazly et al. (2006) attempted to extract English verb-object type idioms by recognizing their structural fixedness in terms of mutual information and relative entropy. Bannard (2007) tried to extract English syntactically fixed verb-noun combinations using pointwise mutual information, and so on.

In spite of these and many similar efforts, it is still difficult to adequately extract MWEs from corpora using a statistical approach, because regarding the types of multiword expressions, realistically speaking, the corpus-wide distribution can be far from exhaustive. Paradoxically, to compile an MWE lexicon we need a reliable standard MWE lexicon, as it is impossible to evaluate the automatic extraction by recall rate without such a reference. The conventional idiom dictionaries published for human readers have been occasionally used for the evaluation of automatic extraction methods in some past studies. However, no conventional Japanese dictionary of idioms would suffice for an MWE lexicon for the practical NLP because they lack entries related to the diverse MWE objects we frequently encounter in common textual materials, such as quasi-idioms, quasi-clichés, metaphoric fixed or partly fixed expressions. In addition, they provide no systematic information on the notational variants, syntactic functions, or syntactic structures of the entry expressions. The JDMWE is intended to circumvent these problems.

In past Japanese MWE studies, Shudo et al. (1980) compiled a lexicon of 3,500 functional multiword expressions and used the lexicon for a morphological analysis of Japanese. Koyama et al. (1998) made a seven-point increase in the precision rate of kana-to-kanji conversion for a commercial Japanese word processor by using a prototype of the JDMWE with 65,000 MWEs. Baldwin et al. (2003) discussed the treatment of Japanese MWEs in the framework of Sag et al. (2002). Shudo et al. (2004) pointed out the importance of the auxiliary-verbal MWEs and their non-propositional meanings (i.e., modality in a generalized sense). Hashimoto et al. (2009) studied a disambiguation method of semantically ambiguous idioms using 146 basic idioms.

3 MWEs Selected for the JDMWE

The human deliberate judgment is indispensable for the correct, extensive extraction of MWEs. In

view of this, we have manually extracted multiword expressions that have definite syntactic, semantic, or communicative functions and are linguistically idiosyncratic from a variety of publications, such as newspaper articles, journals, magazines, novels, and dictionaries. In principle, the idiosyncrasy of MWEs is twofold: first, the semantic non-compositionality (i.e., idiomaticity); second, the strong probabilistic affinity between component words. Here we have treated them differently.

The number of words included in a MWE ranges from two to eighteen. The length distribution is shown in Figure 1.

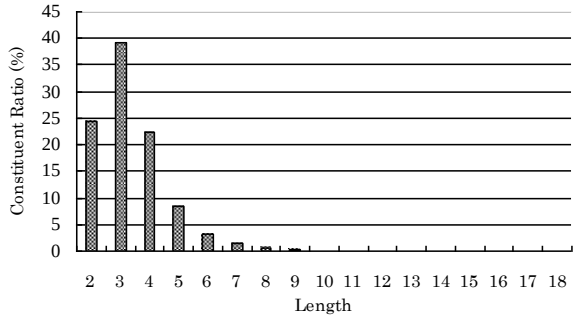

Figure 1: Length distribution of MWEs

type	example
Idiom: Semantically Non-Compositional Expression	赤−の−他人 *aka-no-tanin* (lit. red stranger) “complete stranger”
Morphologically or Syntactically Non-Compositional Expression, Cranberry-Type Expression	と−は−いえ *to-ha-ie* “however”
SVC: Support-Verb Construction	批判−を−加える *hihan-wo-kuwaeru* (lit. add criticism) “criticize”
Compound Noun; Compound Verb; Compound Adjective; Compound Adjective-Verb	打ち−拉がれる *uti-hisigareru* (lit. be hit and smashed) “become depressed”
Four-Character-Idiom	支離−滅裂 *siri-meturetu* “incoherence”
Metaphorical Expression	命−の−限り *inoti-no-kagiri* (lit. limit of life) “at the risk of life”
Quasi-Idiom	辞書−を−引く *jisho-wo-hiku* (lit. pull dictionary) “look up in a dictionary”

Table 1: Non-Compositional Expressions

3.1 Non-Compositional MWEs

In our approach, we use non-substitutability criterion to define a word string as an MWE, the logic being that an MWE expression is usually fixed in its form and the substitution of one of its constituent words would yield a meaningless expression or an expression with a meaning that is completely different from that of the original MWE expression. Formally, a word string $w_1w_2\cdots w_i\cdots w_n$ ($2\leq n\leq 18$) is an MWE if it has a definite syntactic, semantic, or communicative function of its own, and if $w_1w_2 \cdots w_i' \cdots w_n$ is either meaningless or has a meaning completely different from that of $w_1w_2\cdots w_i\cdots w_n$ for some i, where w_i' is any synonym or synonymous phrase of w_i. For example, 赤(w_1)-の-他人 *aka-no-tanin* (lit. “red stranger”) is selected because it has a definite nominal meaning of “complete stranger” and neither 真紅(w_1')-の-他人 *sinku-no-tanin* nor レッド(w_1')-の-他人 *reddo-no-tanin* means “complete stranger”. The evaluation of semantic relevance of MWEs was carried out by human judges entirely. It is just too difficult to judge the semantic relevance automatically and correctly. Table 1 shows a number of MWEs of this type.[1]

3.2 Probabilistically Idiosyncratic MWEs

An MWE must form a linguistic unit of its own. This and the following transition probability condition constitute another criterion that we adopt to define what an MWE is. Formally, a word string $w_1w_2\cdots w_i\cdots w_n$ ($2\leq n\leq 18$) is an MWE if it has a definite syntactic, semantic, or communicative function of its own, and if its forward or backward transition probability $\boldsymbol{p}_f(w_{i+1}|w_1\cdots w_i)$ or $\boldsymbol{p}_b(w_i|w_{i+1}\cdots w_n)$, respectively is judged to be in the relatively high range for some i. With this definition, for example, 手-を-拱く *te-wo-komaneku* “fold arms” is selected as an MWE because it is a well-formed verb phrase and $\boldsymbol{p}_b$(手|を-拱く) is judged empirically to be very high. No general probabilistic threshold value can be fixed a priori because the value is expression-dependent. Although the probabilistic judgment was performed, for each expression in turn, on the basis of the developer’s empirical language model, the resulting dataset is consistent with this criterion on

[1] These classes are not necessarily disjoint.

the whole as shown in section 6.1. Table 2 lists some MWEs of this type.[2]

type	example
Cliché, Stereotyped, Hackneyed, or Set Expression	風前−の−灯 *fuuzen-no-tomosibi* (lit. light in front of the wind) "candle flickering in the wind"
Proverb, Old-Saying	急が−ば−回れ *isoga-ba-maware* (lit. make a detour when in a hurry) "more haste, less speed"
Onomatopoeic or Mimetic Expression	ノロノロ−と−歩く *noronoro-to-aruku* (lit. slouchingly walk) "walk slowly"
Quasi-Cliché, Institutionalized Phrase	肩−の−荷−を−下ろす *kata-no-ni-wo-orosu* (lit. lower lord from the shoulder) "take a big load off one's mind"

Table 2: Probabilistically Idiosyncratic Expressions

With entries like these, an NLP system can use the JDMWE as a reliable reference while effectively disambiguating the structures in the syntactic analysis process.

Of the MWEs in the JDMWE, approximately 38% and 92% of them were judged to meet criterion 3.1 and criterion 3.2, respectively. These are illustrated in Figure 2.

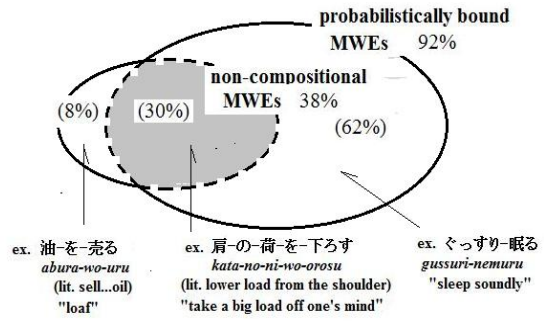

Figure 2: Approximate constituent ratio of non-compositional MWEs and probabilistically bound MWEs

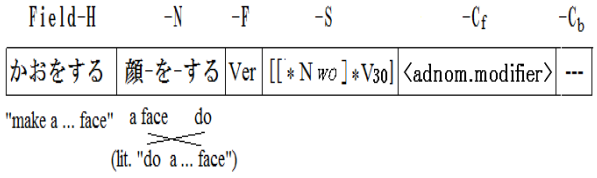

Figure 3: Example JDMWE entry

[2] These classes are not necessarily disjoint.

4 Contents of the JDMWE

The JDMWE has approximately 104,000 entries, one for each MWE, composed of six fields, namely, Field-H, -N, -F, -S, -C_f, and -C_b. The dictionary entry form of an MWE is stated in Field-H in the form of a non-segmented hira-kana (phonetic character) string. An example is given in Figure 3.

4.1 Notational Information (Field-N)

Japanese has three notational options: hira-kana, kata-kana, and kanji. The two kanas are phonological syllabaries. Kanji are originally Chinese idiographic characters. As we have many kanji characters that are both homophonic and synonymous, sentences can contain kanji replaceable by others. In addition, the inflectional suffix of some verbs can be absent in some contexts. The JDMWE has flexible conventions to cope with these characteristics. It uses brackets to indicate an optional word (or a series of interchangeable words marked off by the slash "/") in the Field-N description. Therefore, the entry whose Field-H (the first field) is きのいいやつ *ki-no-ii-yatu* (lit. "a guy who has a good spirit") "good-natured guy", can have (き/気)-の-(い/良/好/善)い-(やつ/奴/ヤツ) in its Field-N. The dash "-" is used as a word boundary indicator. This example can stand for twenty-four combinatorial variants, i.e., きのいいやつ,…, 気の良い奴,…, 気の善いヤツ.

If fully expanded with this information, the JDMWE's total number of MWEs can exceed 750,000.

4.2 Functional Information (Field-F)

Linguistic functions of MWEs can be simply classified by means of codes, as shown in Tables 3 and 4. Field-F is filled with one of those codes which corresponds to a root node label in the syntactic tree representation of a MWE.

code	function	size	example
Cdis	Discourse-Connective	1,000	言い−換えれ−ば *ii-kaere-ba* (lit. if (I) paraphrase) "in other words"
Adv	Adverbial	6,000	不思議−と *fusigi-to* "strangely enough"
Pren	Prenominal-Adjectival	13,700	確−たる *kaku-taru* "definite"

Nom	Nominal	12,000	灰汁−の−強さ *aku-no-tuyosa* (lit. strong taste of lye) “strong harshness”
Nd	Nominal/ Dynamic	4,700	一目−惚れ *hitome-bore* “love at first sight”
Nk	Nominal/State-describing	5,400	二−枚−舌 *ni-mai-jita* “being double-tongued”
Ver	Verbal	49,000	油−を−売る *abura-wo-uru* (lit. sell oil) “idle away”
Adj	Adjectival	4,600	眼−に−入れ−ても−痛く−ない *me-ni-ire-temo-itaku-nai* (lit. have no pain even if put into eyes) “an apple in ones eye”
K	Adjective-Verbal	3,500	経験−豊か *keiken-yutaka* “abundant in experience”
Ono	Onomatopoeic or Mimetic Expression	1,300	スラスラ−と *surasura-to* “smoothly”, “easily”, “fluently”

Table 3: Syntactic Functions and Examples

code	function	size	example
_P	Proverb, Old-Saying	2,300	百聞−は−一見−に−如か−ず *hyakubun-ha-ikken-ni-sika-zu* (lit. hearing about something a hundred times is not as good as seeing it once) “a picture is worth a thousand words”
_Self	Soliloquy, Monologue	200	困っ−た−なあ *komat-ta-naa* “Oh boy, we’re in trouble!”
_Call	Call, Yell	150	済み−ませ−ん−が *sumi-mase-n-ga* “Excuse me.”
_Grt	Greeting	200	いらっしゃい−ませ *irasshai-mase* “Welcome!”
_Res	Response	350	どう−いたし−まし−て *dou-itasi-masi-te* “You’re welcome.”

Table 4: Communicative Functions and Examples

4.3 Structural Information (Field-S)

4.3.1 Dependency Structure

The dependency structure of an MWE is given in Field-S by a phrase marker bracketing the modifier-head pairs, using POS symbols for conceptual words.[3] For example, an idiom 真っ赤-な - 嘘 *makka-na-uso* (lit. “crimson lie”) “downright lie” is given a marker [[K_{00} *na*] N]. This description represents the structure shown in Figure 4, where K_{00} and N are POS symbols denoting an adjective-verb stem and a noun, respectively.

[3] The intra-sentential dependency relation in Japanese is unilateral, i.e., the left modifier depends on the right head.

The JDMWE contains 49,000 verbal entries, making this the largest functional class in the JDMWE. For these verbal entries, more than 90 patterns are actually used as structural descriptors in Field-S. This fact can indicate the broadness of the structural spectrum of Japanese verbal MWEs. Some examples are shown in Table 5.

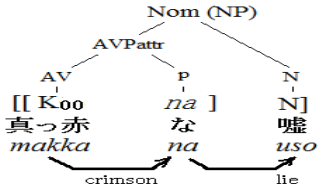

Figure 4: Example of dependency structure given in Field-S

example of structural pattern of verbal MWE	example of MWE
[[N *wo*] V_{30}]	異−を−唱える *i-wo-tonaeru* (lit. chant the difference) “raise an objection”
[[N *ga*] V_{30}]	撚り-が-戻る *yori-ga-modoru* (lit. the twist comes undone) “get reconciled”
[[N *ni*] V_{30}]	手−に−入れる *te-ni-ireru* (lit. put...into hands) “get”, “obtain”
[[[[N *no*] N] *ga*] V_{30}]	化け−の−皮−が−剥げる *bake-no-kawa-ga-hageru* (lit. peel off disguise) “expose the true colors”
[[[[N *no*] N] *ni*] V_{30}]	玉−の−輿−に−乗る *tama-no-kosi-ni-noru* (lit. ride on a palanquin for the nobility) “marry into wealth”
[[N *de*][[N *wo*] V_{30}]]	顎−で−人−を−使う *ago-de-hito-wo-tukau* (lit. use person by a chin) “order a person around”
[[N *ni*][[N *ga*] V_{30}]]	尻−に−火−が−付く *siri-ni-higa-tuku* (lit. buttocks catch fire) “get in great haste”
[[V_{23} *te*] V_{30}]	切っ−て−落とす *ki-te-otosu* (lit. cut and drop) “cut off”
[[V_{23} *ba*] V_{30}]	打て−ば−響く *ute-ba-hibiku* (lit. reverberate if hit) “respond quickly”
[[[[N *ni*] V_{23}] *te*] V_{30}]	束−に−なっ−て−掛かる *taba-ni-nat-te-kakaru* (lit. attack someone by becoming a bunch) “attack all at once”
[Adv [[N *ga*] V_{30}]]	どっと−疲れ−が−出る *dotto-tukare-ga-deru* (lit. fatigue bursts out) “being suddenly overcome with fatigue”

Table 5: Examples of structural types of verbal MWEs (N: noun, V_{23}: verb (adverbial form), V_{30}: verb (end form), Adv: adverb, *wo*, *ga*, *ni*, *no*, *de*, *te*, and *ba*: particle)

4.3.2 Coordinate Structure

Approximately 2,500 MWEs in the JDMWE contain internal coordinate structures. This information is described in Field-S by bracketing with "<" and ">", and the coordinated parts by "(" and ")". The coordinative phrase specification usually requires that the conjuncts must be parallel with respect to the syntactic function of the constituents appearing in the bracketed description. For example, an expression 後-は-野-と-なれ-山-と-なれ *ato-ha-no-to-nare-yama-to-nare* (lit. "the rest might become either a field or a mountain") "what will be, will be", has an internal coordinate structure. Thus, its Field-S is [[N *ha*]<([[N *to*] V_{60}])([[N *to*] V_{60}])>]. This description represents the structure shown in Figure 5, where V_{60} denotes an imperative form of the verb.

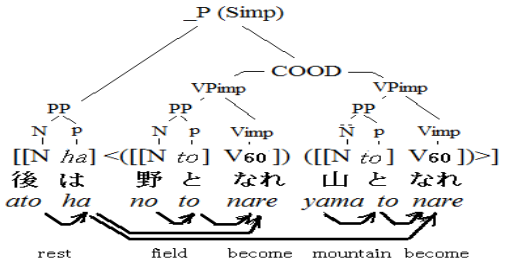

Figure 5: Example of the coordinate structure shown by "<" and ">" in Field-S

4.3.3 Non-phrasal Structure

Approximately 250 MWEs in the JDMWE are syntactically ill-formed in the sense of context-free grammar but still form a syntactic unit on their own. For example, 揺り籠-から-墓場-まで *yurikago-kara-hakaba-made* "from the cradle to the grave" is an adjunct of two postpositional phrases but is often used as a state-describing noun as in 揺り籠-から-墓場-まで-の-保証 *yurikago-kara-hakaba-made-no-hoshou* (lit. security of from cradle to grave) "security from the cradle to the grave". Thus Field-F and Field-S have a functional code N_k and a description [[N *kara*][N *made*] $]], respectively. The symbol "$" denotes a null constituent occupying the position of the governor on which this MWE depends. This structure is shown in Figure 6.

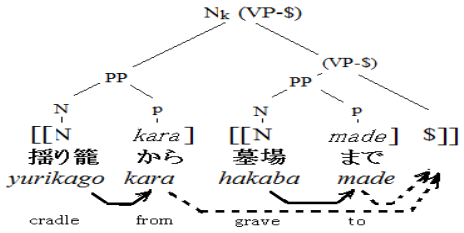

Figure 6: Example of a non-phrasal expression with a null constituent marked with "$" in Field-S

The total number of structural types specified in Field-S is nearly 6,000. This indicates that Japanese MWEs present a wide structural variety.

4.3.4 Internal Modifiability

Some MWEs are not fixed-length word strings, but allow the occurrence of phrasal modifiers internally. In our system, this aspect is captured by prefixing a modifiable element of the structural description stated in the Field-S with an asterisk "*". An adverbial MWE 上-に-述べ-た-様-に *ue-ni-nobe-ta-you-ni* "as I explained above" is one such MWE and thus has a description [[[[[N *ni*] *V_{23}] *ta*] N] *ni*] in Field-S, meaning that the third element V_{23} is a verb that can be modified internally by adverb phrases. Since the asterisk designates such optional phrasal modification, our system allows a derivative expression like 理由 -を-上-に-詳しく-述べ-た-様-に *riyuu-wo-ue-ni-kuwasiku-nobe-ta-you-ni* "as I explained in detail the reason above", which contains two additional, internal modifiers. The structure is shown in Figure 7.[4]

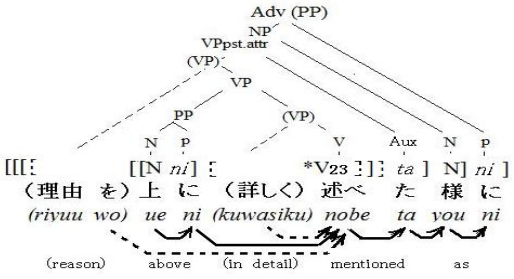

Figure 7: Example of internal modifiability marked by "*" in Field-S

[4] The positions to be taken by an internal modifier can be easily decided by the structural description given in Field-S along with the nest structure requirement.

Roughly speaking, 30,000 MWEs in the JDMWE have no asterisk in their Field-S. Our rigid examination reveals that internal modification is not allowed for them.

5 Contextual Condition (Field-C_f , C_b)

Approximately 6,700 MWEs need to be classified differently because they require particular forward contexts, i.e., they require co-occurrence of a particular syntactic phrase in the context that immediately precedes them. For example, 顔-を-する *kao-wo-suru* (lit. "do face") which is a support-verb construction, cannot occur without an immediately preceding adnominal modifier, e.g., the adjective 悲しい *kanasii* "sad", yielding 悲しい-顔-を-する *kanasii-kao-wo-suru* (lit. "do sad face") "make a sad face". This adnominal modifier co-occurrence requirement is stipulated in Field-C_f by a code <adnom. modifier>. There are about 30 of these forward contextual requirements. Similarly, backward contextual requirements, of which there are about 70, are stated in Field-C_b. Approximately 300 MWEs require particular backward contexts.

6 Statistical Properties

Without a rule system of semantic composition, it is difficult to evaluate the validity of the JDMWE concerning idiomaticity. However, we can confirm that 3,600 Japanese standard idioms that Sato (2007) listed from five Japanese idiom dictionaries published for human readers are included in the JDMWE as a proper subset. In addition, the JDMWE contains the information about their syntactic functions, structures, and flexibilities.

6.1 Comparison with Web N-gram Frequency Data

We examined the statistical properties of the JDMWE using the Japanese Web N-gram, version 1: LDC2009T08, which is a word N-gram ($1 \leq N \leq 7$) frequency dataset generated from 2×10^{10} sentences in a Japanese Web corpus, supplied by Google Inc. (Kudo et al. 2009). We will refer to this (or the Web corpus examined) subsequently as GND. We will refer to trigram $w_1w_2w_3$ as an NpV-trigram only when w_1 and w_3 are restricted to a noun and a verb (end form), respectively, and w_2 is one of the following case-particles: accusative を *wo*, subjective が *ga*, or dative に *ni*.[5] We write the number of occurrences of an expression x, counted in the GND, as C(x).

First, we obtain from the GND sets G, T, D, B, and R_i's defined below, using a Japanese word dictionary IPADIC (Asahara et al. 2003):

G={$w_1w_2w_3$| $w_1w_2w_3 \in$ GND, $w_1w_2w_3$ is an NpV-trigram.}
T={$w_1w_2w_3$| $w_1w_2w_3 \in$ JDMWE, $w_1w_2w_3$ is an NpV-trigram.}
D={w_1w_2|$\exists w_3$, $w_1w_2w_3 \in G$}
B={w_1w_2|$\exists w_3$, $w_1w_2w_3 \in T$}
R_i={$w_1w_2w_3$| $w_1w_2w_3 \in T$, C($w_1w_2w_3$) is the i-th largest among C(w_1w_2v)'s for all $w_1w_2v \in G$}.

We then found the following data:
・$|B|$=10,548
・$|D|$=110,822
・$|R_1|$=4,983, $|R_2|$=1,495, $|R_3|$=786, $|R_4|$=433, …
From these, we realize, for example, that 47.2% =($|R_1|/|B|$)×100 of trigrams in T have verbs that occur most frequently in the GND, succeeding the individual bigrams. An example of such a trigram is アクション-を-起こす *akushon-wo-okosu* (lit. "raise action") "take action". Similarly, 14.0%= ($|R_2|/|B|$)×100 have the second most frequent verbs, 7.5% have the third most frequent verbs, and so on. Figure 8(a) illustrates the results. From this, we can assume that the higher probability $\boldsymbol{p}_f(w_3|w_1w_2)$ a trigram $w_1w_2w_3$ has, the more likely w_3 is chosen for each w_1w_2 in the JDMWE. This is consistent with what we wrote in section 3.2. Figure 8(b) is the accumulative substitute of Figure 8(a). Extrapolating Figure 8(b) suggests that 10% of NpV-trigrams in the JDMWE do not occur in the GND. This implies that the size, i.e., 2×10^{10} sentences of the Web corpus used by the GND is not sufficiently large to allow MWE extraction.[6]

[5] The NpV-trigrams represent the typical forms of shortest Japanese sentences, corresponding roughly to subject-verb, verb-object/direct, and verb-object/indirect constructions in English.
[6] Otherwise, the frequency cut-off point of 20 adopted in GND is too high.

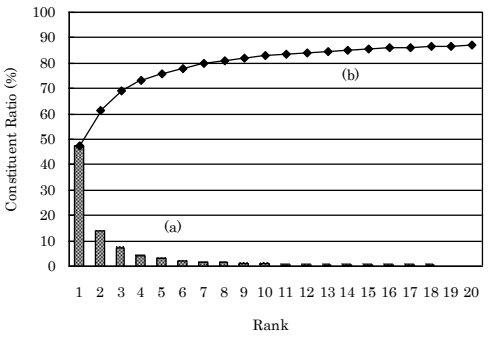

Figure 8 (a): Constituent ratio ($|R_i|/|B|$)×100 for rank i of probability $\boldsymbol{p}_f(w_3|w_1w_2)$; (b): Accumulative variant of (a) for rank i of probability $\boldsymbol{p}_f(w_3|w_1w_2)$

Second, we calculate the (normalized) entropy $H_f(w_3|w_1w_2)$ for each $w_1w_2 \in D$ defined below, where the probability $\boldsymbol{p}_f(w_3|w_1w_2)$ is estimated by $C(w_1w_2w_3)/C(w_1w_2)$. This provides a measure of the flatness of the $\boldsymbol{p}_f(w_3|w_1w_2)$ distribution canceling out the influence of the number N of verb types w_3's.

$$H_f(w_3|w_1w_2) = -\Big(\sum_{w_3} \boldsymbol{p}_f(w_3|w_1w_2) \log \boldsymbol{p}_f(w_3|w_1w_2)\Big) / \log N$$

After arranging 110,822 bigrams in *D* in ascending order of $H_f(w_3|w_1w_2)$, we divided them into 20 intervals $A_1, A_2, \ldots, A_{20}$ each with an equal number of bigrams (5,542). We then examined how many bigrams in *B* were included in each interval. Figures 9(a) and (b) plot the resulting constituent ratio of the bigrams in *B* and the mean value of $H_f(w_3|w_1w_2)$'s in each interval, respectively. We found, for example, that 1,262 out of 5,542 bigrams are in *B* for the first interval, i.e., the constituent ratio is 22.8%=(1,262/5,542) × 100. Similarly, we obtain 22.5%=(1,248/5,542) × 100 for the second interval, 20.5%=(1,136/5,542)×100 for the third, and so on. From this, we realize the macroscopic tendency that the larger the entropy $H_f(w_3|w_1w_2)$, or equivalently the perplexity of the succeeding verb w_3, a bigram w_1w_2 has, the less likely it is adopted as a prefix of a trigram in *T*.

Taking the results in Figure 8 and Figure 9 together, we can presume that not only frequently but also exclusively occurring verbs would be the preferred choice in *T*.

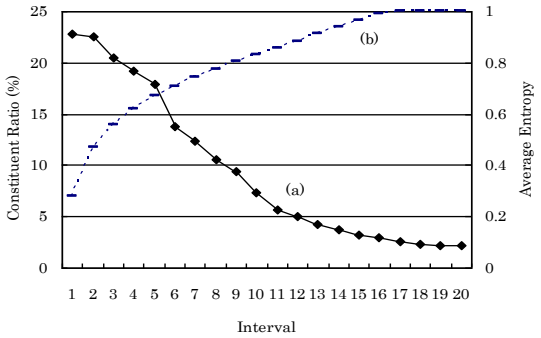

Figure 9 (a): Constituent ratio of the bigrams in *B* among bigrams in *D* in interval k ($1 \le k \le 20$); (b): Mean value of entropies $H_f(w_3|w_1w_2)$'s in the interval k ($1 \le k \le 20$)

This suggests the general feasibility of the JDMWE, for its relative compactness, in effectively disambiguating the syntactic structures of input word strings.

The above investigations were carried out on the forward conditional probabilities for restricted types of MWEs. However, the results imply a general validity of the JDMWE since the same criteria for selection were applied to all kinds of multiword expressions.

6.2 Occurrences in Newspapers

We examined 2,500 randomly selected sentences in Nikkei newspaper articles (published in 2009) to determine how many MWE tokens of the JDMWE occur in them. We found that in 100 sentences an average of 74 tokens of our MWEs were used. This suggests a large lexical coverage of the JDMWE.

7 Concluding Remarks

The JDMWE is a slotted tree bank for idiosyncratic multiword expressions, annotated with detailed notational, syntactic information.

The idea underlying the JDMWE is that the volume and meticulousness of the lexical resource crucially affects the outcome of the rule-oriented, large-scale NLP. In view of this, the JDMWE was designed to encompass the wide range of linguistic objects related to Japanese MWEs, by placing importance on the recall rate in the selection of the

candidate expressions.[7] The statistical properties clarified in this paper imply the general feasibility of the JDMWE at least in the probabilistic respect.

Possible fields of application of the JDMWE include, for example:

- Phrase-based machine translation
- Phrase-based speech recognition
- Phrase-based kana-to-kanji conversion
- Search engine for Japanese corpus
- Paraphrasing system
- Japanese dialoguer
- Japanese language education system

Another aspect of the JDMWE is that it would provide linguists with lexicological data. For example, the usage of Japanese onomatopoeic adverbs, which are mostly bound probabilistically to specific verbs or adjectives, is extensively catalogued in the JDMWE.

The first version of the JDMWE will be released after proofreading.[8] If possible, we would like to add further information to each MWE on morphological variants, passivization, relativization, decomposability, paraphrasing, and semantic disambiguation for future versions.

Acknowledgments

We would like to thank the late Professors Toshihiko Kurihara and Sho Yoshida, who inspired our current research in the 1970s. Similar thanks go to Makoto Nagao. We are also grateful to everyone who assisted in the development of the JDMWE. Further special thanks go to Akira Shimazu, Takano Ogino, and Kenji Yoshimura for their encouragement and useful discussions, to those who worked on the LDC2009T08 and IPADIC, to the three anonymous reviewers for their valuable comments and advice, and to Stephan Howe for advice on matters of English style in the current paper.

References

Asahara, M. and Matsumoto, Y. 2003. IPADIC version 2.7.0 User's Manual (in Japanese). NAIST, Information Science Division.

Baldwin, T. and Bond, F. 2003. Multiword Expressions: Some Problems for Japanese NLP. Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing (Japan): 379–382.

Bannard, C. 2007. A Measure of Syntactic Flexibility for Automatically Identifying Multiword Expressions in Corpora. Proceedings of A Broader Perspective on Multiword Expressions, Workshop at the ACL 2007 Conference: 1–8.

Baptista, J., Correia, A., and Fernandes, G. 2004. Frozen Sentences of Portuguese: Formal Descriptions for NLP. Proceedings of ACL 2004 Workshop on Multiword Expressions: Integrating Processing: 72–79.

Fazly, A. and Stevenson, S. 2006. Automatically Constructing a Lexicon of Verb Phrase Idiomatic Combinations. Proceedings of the 11th Conference of the European Chapter of the ACL: 337–344.

Fellbaum, C. (ed.) 1999. WordNet. An Electronic Lexical Database, Cambridge, MA: MIT Press.

Fellbaum, C., Geyken, A., Herold, A., Koerner, F., and Neumann, G. 2006. Corpus-Based Studies of German Idioms and Light Verbs. International Journal of Lexicography, Vol. 19, No. 4: 349-360.

Gross, M. 1986. Lexicon-Grammar. The Representation of Compound Words. Proceedings of the 11th International Conference on Computational Linguistics, COLING86:1–6.

Hashimoto, C. and Kawahara, D. 2009. Compilation of an Idiom Example Database for Supervised Idiom Identification. Language Resource and Evaluation Vol. 43, No. 4 : 355-384.

Jackendoff, R. 1997. The Architecture of Language Faculty. Cambridge, MA: MIT Press.

Koyama, Y., Yasutake, M., Yoshimura, K., and Shudo, K. 1998. Large Scale Collocation Data and Their Application to Japanese Word Processor Technology. Proceedings of the 17th International Conference on Computational Linguistics, COLING98: 694–698.

Kudo, T. and Kazawa, H. 2009. Japanese Web N-gram Version 1. Linguistic Data Consortium, Philadelphia.

Kuiper, K., McCan, H., Quinn, H., Aitchison, T., and Van der Veer, K. 2003. SAID: A Syntactically Anno tated Idiom Dataset. Linguistic Data Consortium 2003T10.

Laporte, É. and Voyatzi, S. 2008. An Electronic Dictionary of French Multiword Adverbs. Proceedings of the LREC Workshop towards a Shared Task for Multiword Expressions (MWE 2008): 31–34.

[7] The time required to compile this dictionary is estimated at 24,000 working hours.

[8] A portion of the JDMWE is available at http://jefi.info/.

Pantel, P. and Lin, D. 2001. A Statistical Corpus-Based Term Extractor. Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, Springer-Verlag: 36–46.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. 2002. Multiword Expressions: A Pain in the Neck for NLP. Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics, CICLING2002: 1–15.

Sato, S. 2007. Compilation of a Comparative List of Basic Japanese Idioms from Five Sources (in Japanese). IPSJ SIG Notes 178: 1-6.

Shudo, K., Narahara, T., and Yoshida, S. 1980. Morphological Aspect of Japanese Language Processing. Proceedings of the 8th International Conference on Computational Linguistics, COLING80: 1–8.

Shudo, K., Tanabe, T., Takahashi, M., and Yoshimura, K. 2004. MWEs as Non-Propositional Content Indicators. Proceedings of ACL 2004 Workshop on Multiword Expressions: Integrating Processing: 31–39.

Uchiyama, K. and Ishizaki, S. 2003. A Disambiguation of Compound Verbs. Proceedings of ACL 2003. Workshop on Multiword Expressions: Analysis, Acquisition and Treatment: 81–88.

Villavicencio, A. 2004. Lexical Encoding of MWEs. Proceedings of ACL 2004 Workshop on Multiword Expressions: Integrating Processing: 80–87.