Corpus-based Extraction of Japanese Compound Verbs

James Breen and **Timothy Baldwin**
Department of Computer Science & Software Engineering
University of Melbourne
Australia
jimbreen@gmail.com, tb@ldwin.net

Abstract

We describe two methods for Japanese compound verb (JCV) extraction, based on synthesis and pattern matching over the Google Japanese n-gram corpus. We devise a number of filters to boost the precision of the corpus-based method, and evaluate the two methods based on a sample of JCVs occurring in varying frequency bands. We also investigate the distribution of JCV token frequency, and the type frequency of their components.

1 Introduction

This paper describes work conducted in a project to extract Japanese compound verbs (JCVs) from corpora and corpus-based resources. Compound verbs in Japanese have attracted considerable attention in Japanese linguistics as they are a highly productive and flexible element of the language (Shibatani, 1990; Baldwin and Bond, 2002; Tsujimura, 2006). Apart from some manually-prepared verb lists they have received relatively little attention in corpus linguistics.

The reasons for collecting and studying Japanese compound verbs include:

a. the development of reliable methods for extraction of the verbs from corpora;

b. investigation of the distribution of the verbs and their constituents;

c. investigation of the coverage of the verbs in the major lexicons

In particular, it is hoped that by isolating JCVs which are in use, but are not currently recorded or lexicalized, and eventually by developing and verifying Japanese meanings and English translational equivalents for these verbs, the lexicon of JCVs can be expanded.

At the current stage of the project, two methods for extracting compound verbs have been developed and applied over a major Japanese corpus. The result has been the identification of a large number of potential JCVs, which we show to have a high level of precision, relative to a sample of JCVs across varying frequency bands. We also investigate the distribution of the frequency of the verbs and their components.

2 Overview of Japanese Compound Verbs

The compound verb in Japanese (複合動詞 *fukugôdôshi*, hereafter JCV) is a concatenation of two or more verbs which function as a single multiword verb. There are several classes of JCV, however in this work we concentrate of the largest and most common class in which the first verb is in the continuative form (also known as the *masu*-stem because it forms the base for the polite spoken *-masu* group of inflections) (Uchiyama et al., 2005; Kubota, 1992).[1] In common with most studies of JCVs, we concentrate on verbs where both components are native Japanese verbs, not loanwords or Sino-Japanese words. This exclusion is because these latter verbs are much less

[1] In some cases this undergoes phonetic variation, e.g. the gemination in 引っ越す *hikkosu* as an alternative to the regular 引き越す *hikikosu*.

common and have a different morphology.

As a JCV consists of two adjacent verb components, we will refer to these components as the V1 and V2. A typical JCV is 行き過ぎる *ikisugiru* “to go too far”, where the V1 is the continuative form of 行く *iku* “to go” and the V2 is 過ぎる *sugiru* “to be excessive; to be too much”. 過ぎる is a particularly productive V2. A less productive V2 is 汚す *yogosu* “to make dirty”, found in the JCV 食べ汚す *tabeyogosu* “to eat messily”.

JCVs play a role in Japanese which is analogous to several different structures in other languages. English equivalents include compound verbs (e.g. *to start to eat*), verb-plus-gerund (e.g. *to start swimming*) and verb particle constructions (e.g. *to kick up (ball, fuss), to pull down*).

The JCV is a highly productive form, with some particular V1s and V2s being strongly represented, however there is no real restriction on a verb being used within a JCV, subject to issues such as aspect and valency (Kubota, 1992), and the result being meaningful. Some popular references list many hundreds of JCVs (Tagashira and Hoff, 1986) and major dictionaries typically include several thousand as entries, however it is generally recognized that many more JCVs are in use than are lexicalized. The incompleteness of the lexicalization of JCVs arises not only from their productivity, but from the fact that their meaning is often obvious to a Japanese speaker, and hence dictionary editors usually concentrate on JCVs which are polysemous, or have idiosyncratic meanings. Extension of the coverage of recorded and translated JCVs would be of assistance in areas such as language learning, and in lexicons used by morphological analysis and machine translation systems.

An example of a polysemous JCV is 引き抜く *hikinuku*, from 引く *hiku* “to draw; to pull”, and 抜く *nuku* “to extract”. It means both “to uproot” and “to pull out”, and less obviously “to head-hunt” and “to lure away”.

3 General Approach and Resources

3.1 Approaches

A fundamental problem with searching Japanese corpora for unrecorded words is that Japanese text does not usually have spaces or any other marking between words. Thus the identification of words in text necessitates the use of a morphological analysis process to separate the words, and all such processes currently rely on extensive lexicons. The absence of a word in the lexicon usually results in the analysis software defaulting to producing a sequence of untagged morphemes until it can resynchronize.

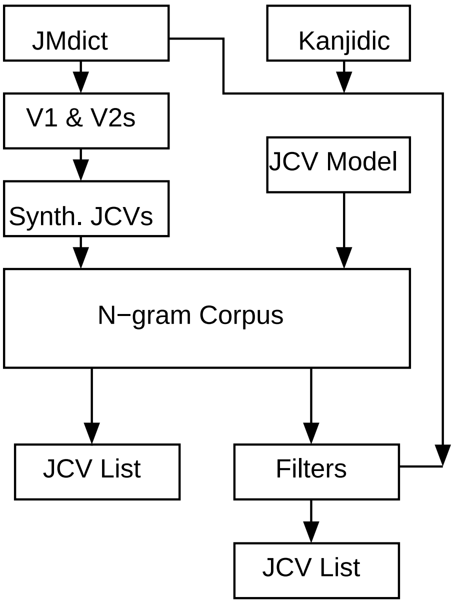

Figure 1: An outline of the two proposed approaches for JCV generation/extraction

An approach that has had some previous success is to synthesize possible words by mimicking Japanese morphological processes, and then testing, e.g. using a WWW search engine, to determine whether the word is in use (Breen, 2004a). A variant of this approach has been applied in the current project. A second approach in which the Google n-gram corpus (described below) was scanned using a filter designed to detect the character patterns consistent with JCVs. These approaches are described in detail below. Figure 1 shows a diagram of the two approaches.

3.2 Resources Used

The project uses several lexical resources to assist with the identification and extraction of JCVs. The JMdict Japanese-English dictionary database (Breen, 2004b) and the associated Kanjidic database (Breen, 2009) were used to establish sets of possible V1 and V2 components, and a combined lexicon was constructed from the JMdict file and the following:

- the Kôjien Japanese dictionary (Ootsuka, 1998)
- the Daijirin Japanese dictionary (Matsumura, 1995)
- the GoiTaikei lexicon (Ikehara et al., 1999)
- the Japanese Linguistics Database (JLD) (Halpern, 2008)

The Kôjien and Daijirin are major Japanese dictionaries with extensive coverage of the Japanese lexicon. The GoiTaikei and JLD commercial lexicons are primarily used for Japanese NLP projects and research. The combined lexicon contains over 650,000 surface forms of Japanese words.[2]

The main corpus used in this project has been the Google Japanese n-grams (Kudo and Kazawa, 2007). The set of n-grams in the corpus was compiled by extracting text from a complete crawl of Japanese WWW pages for the month of July, 2007, and analyzing that text using the MeCab morphological analysis system (Kudo, 2008). n-gram sequences of up to 7-grams are recorded in the corpus if they were identified in 20 or more text segments. As systems such as MeCab are observed to break JCVs into up to three morphemes depending on the inflection of the verb, only the 1-gram, 2-gram and 3-gram section of the corpus were used in this study, and then only those n-grams which began with a *kanji* character.

A similar Google n-gram corpus has been used successfully in the extraction of verb-particle constrictions in English (Kummerfeld and Curran, 2008).

3.3 Synthesis of Compound Verbs

In this approach, a set of JCVs were synthesized as follows:

a. The JMdict lexicon was examined and JCVs identified. Including alternative surface forms, some 2,900 JCVs in which *kanji* were used in both the V1 and v2 were extracted. These were divided into the V1 and V2 components, yielding approximately 700 V1s and 600 V2s.

b. Using the V1 and V2 components, 420,000 synthetic JCVs were created via all combinations of the V1s and V2s. For each verb two forms were generated: the form in which the V2 used *kanji* as the root of the verb, and the form in which the V2 was entirely in the *hiragana* script. Both these forms are freely used in Japanese, for example 抱き付く *dakitsuku* "to cling to; to embrace" can equally well be written 抱きつく, and in fact the latter is more commonly used.

   For each of these, as well as the plain non-past tense (which is considered to be the reference form of Japanese verbs and is used for dictionary headwords) two inflections were generated: the continuative *te*-form and the plain past tense. These three are the most commonly used inflections in written Japanese, and it was considered appropriate to focus on them in order to detect whether words were in use.

Each synthetic JCV was initially checked against the combined lexicon, resulting in a total of 6,094 matches.

Each synthetic JCV was then checked against the Google n-gram corpus. As there were three inflections of two written forms of 420,000 JCVs, a total of 2,520,000 words were tested. The sections of the n-gram files which began with a *kanji* were preprocessed to recombine each 2-gram and 3-gram into a single character string, then sorted, resulting in a file of 270M unigrams in a file of 5.8GB. This facilitated processing in a single pass against the sorted verb file, thus enabling a rapid comparison and collation of results.

Initially, approximately 26,000 of the synthesized JCVs were matched in one or more of their

[2] Japanese is written using a combination of Chinese characters (*kanji*) and two syllabaries: *hiragana* and *katakana*. It has considerable flexibility as to whether words are written in *kanji*, one of the syllabaries, or a mixture. Also, alternative *kanji* are often used. For example, the JCV 詰め合わせる *tsume-awaseru* "to pack an assortment of goods, etc." can also be written: 詰め合せる, 詰合わせる, 詰合せる, 詰めあわせる or つめあわせる, and 並び変える *narabikaeru* "to put things in order" can also be written 並変える, 並び替える, 並替える, etc..

inflections. On inspection, the JCV form in which the V2 was in *hiragana* did not contribute significantly to the matches, and as this form has an increased chance of homophones which cannot be resolved without textual context, it was removed from the analysis. Also removed were a number of JCVs which were effectively alternative conjugations (passive, potential, etc.). This reduced the matched JCVs to 22,692.

Of the 6,094 JCVs which were found in the combined lexicon, 4,779 matched n-grams in the corpus, i.e. 1,315 which were found in the combined lexicon were not in the corpus. On inspection it was noted that many of the 1,315 were archaic and literary words.

The distribution of the counts of occurrences in the corpus is sharply asymptotic, with a small number having very high counts and declining to a long tail with over 15,000 having counts below 500.

3.4 Direct n-gram Search

In addition to the synthesis approach, an alternative approach was devised in which the n-gram corpus was scanned for character strings which conformed to the structural pattern of JCVs. From the examination of known JCVs, it can be determined that the common structural pattern is:

- a V1 consisting of one or two *kanji* followed by one to three *hiragana*;
- a V2 consisting of one or two *kanji* followed by one to four *hiragana*.

As there are many other valid text fragments which also conform to this pattern, e.g. noun/particle/verb, noun/particle/adjective, etc. filters were applied as follows:

a. the V1 component was limited to the *masu*-stems of known or potential verbs. To do this, a list of verb *masu*-stems was created and merged from:

   i. all the verbs in the JMdict dictionary;
   ii. all the V1 components used in the synthesis methods;
   iii. all the kanji in the Kanjidic database which had the potential to form a verb (this information is detailed in the database.)

   A total of 6,023 actual or potential verb stems were identified and used to filter the potential JCVs.

b. the inflecting part of the V2 was limited to the *hiragana* strings associated with valid verbs in the plain non-past, plain past and *te*-form inflections. A list of 208 such inflections was compiled and used as a filter.

A scan of the n-gram corpus for unigrams which conformed to the structural model and passed the filters yielded just on 135,000 potential JCVs. As these included many inflected forms, a considerable amount of post-processing was carried out to reduce them to a consolidated set of reference (plain non-past) forms. Some of the processes involved included:

a. matching the inflected and reference forms and combining the counts;

b. detecting and removing additional inflections such as the potential and passive forms, which share some of the inflection patterns of the reference form;

c. detecting and removing adjectives. In Japanese, adjectives inflect in a manner similar to verbs, and a number had been collected in the scan.

From this a reduced list of approximately 80,000 potential JCVs was produced. When tested against the combined lexicon, 6,203 matched with one or more of the dictionaries, an increase of 1,424 over the synthesis method. It is clear that this approach has an improved recall, i.e. the number of actual JCVs identified as a proportion to the number in existence, relative to the published lexicons, but possibly at the price of a reduced precision, i.e the proportion of actual JCVs among the potential JCVs. As with the synthesized JCVs, the distribution of the counts is asymptotic with a long tail.

4 Analysis of the Potential Compound Verbs

A detailed comparison of the potential JCVs compiled in the two approaches revealed that all the synthesized JCVs which had matched unigrams

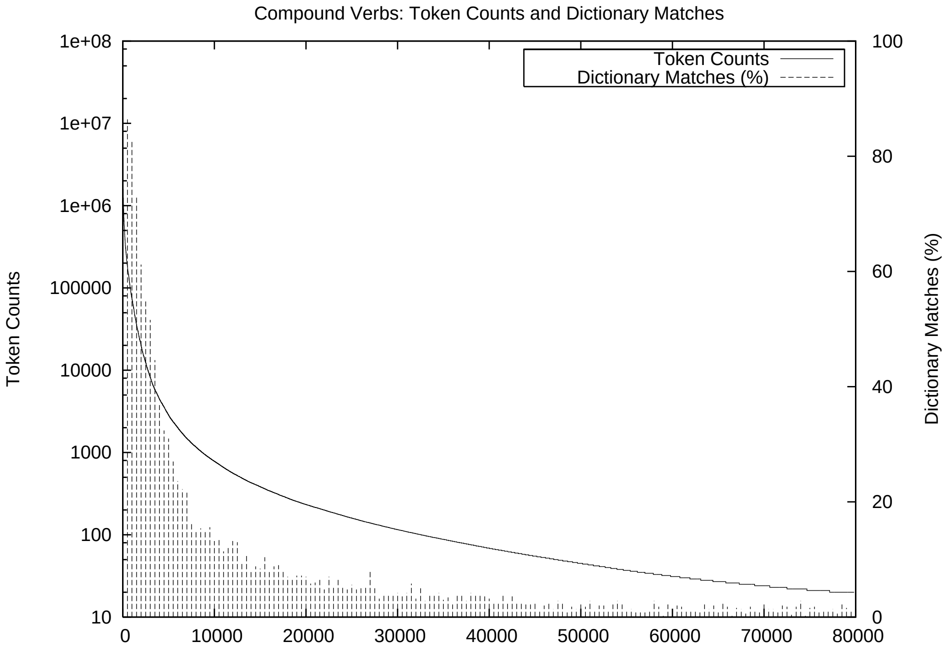

Figure 2: Analysis of the token counts and dictionary matches of JCVs, ranking in decreasing order of token frequency

in the corpus has also been collected in the search, and moreover the n-gram counts were almost always identical. This meant that a combined set could be used for further analysis, with tagging as to whether a JCV had been detected by both methods, or by the search alone.

Figure 2 shows the distribution of the n-gram counts and also the proportion of the potential JCVs which were found in a lexicon.[3]

A key issue is the extent to which these methods have revealed actual JCVs as opposed to character sequences which simply share symbolic characteristics with JCVs. To examine this aspect fully would require an evaluation of each JCV candidate in context, e.g. as it is used in WWW texts, to determine its status. In order to estimate the effectiveness of the JCV extraction approaches, samples of 50 potential JCVs were selected at random from each of three bands based on token frequency

a. High: JCVs with over 5,000 counts in the n-gram corpus (3,795 JCVs)

b. Medium: JCVs with 1,000 to 4,999 counts (4,886 JCVs)

c. Low: JCVs with 20 to 999 counts (71,138 JCVs)

The sample JCV candidates were classified as to whether they were in a lexicon or not, and if not, whether they were actually verbs. For the latter analysis, each potential JCV was manually checked against WWW pages via a search engine to verify whether it was being used as a verb. (At some later stage it may be possible to employ deeper linguistic analysis to carry out this process automatically.)

The summary of this classification is in Table 1. The figures in parentheses are numbers of JCVs in each category resulting from the search approach alone.

The JCV candidates which were classified as “other”, i.e. not verbs, fell into several categories. The most common were inflected adjectives which had not been detected in filtering, adverbs such as 再び, V1s such as 見て which probably should have been filtered out (see below), and apparent typographical or grammatical errors associated with other verbs. Some were other constructs such as noun/verb without the usual intervening particle.

[3] For the purposes of depicting this, JCVs were examined in batches of 500, and the percentage which matched were plotted.

	High	Med	Low
In lexicon	27 (3)	12 (2)	1 (0)
Not in lexicon	23 (14)	38 (20)	49 (35)
verb	7 (1)	26 (8)	27 (14)
other	16 (13)	12 (12)	22 (21)

Table 1: Analysis of sample JCV candidates over the three frequency distribution bands, in terms of their occurrence in the lexicon; for JCV candidates not in the lexicon, we additionally break down the counts into verb and non-verb candidates

	High	Med	Low	Total
JMdict	1,788 (0.47)	453 (0.09)	671 (0.01)	2,912
Kôjien	1,420 (0.37)	401 (0.08)	976 (0.01)	2,797
Daijirin	1,626 (0.43)	491 (0.10)	970 (0.01)	3,087
GoiTaikei	1,375 (0.36)	377 (0.08)	661 (0.01)	2,413
JLD	2,172 (0.57)	932 (0.19)	2,023 (0.03)	5,126

Table 2: Occurrence of potential JCVs in the different dictionaries across the three frequency bands, in terms of the raw type count and proportion of overall types (in parentheses)

Of considerable interest is the relative performance of the JCVs identified by the synthesis approach; in each of the three bands almost all of these JCVs were valid verbs.

In terms of the precision of the different approaches, the full set of potential JCVs achieved precisions of 0.68, 0.76 and 0.56 respectively across the three selected bands, however within this the synthesis approach achieved precisions of 0.91, 1.00 and 0.93.

The comparative recall is difficult to measure as there is no gold standard for the number of JCVs in use or able to be used. Certainly the direct search approach achieved a greater recall but at the price of a lower precision.

As reported above, 6,203 of the potential JCVs matched with one or more of the dictionaries which make up the combined lexicon. It was noticed that while the high-ranking JCVs tended to match all the dictionaries, lower-ranking JCVs tended to match more sparsely, with often only one or two dictionaries matching. The specific dictionary matches were extracted for each of the high, medium and low bands. These are shown in Table 2. The figures in parentheses are the proportions of the dictionary matches against the total dictionary matches for the band.

	Combined n-gram	Synthetic n-gram	Lexicon	First 10,000
V1	2,601	680	1,294	1,290
V2	8,883	591	1,314	1,597

Table 3: V1 and V2 frequencies for the two proposed methods and in the lexicon

5 Productivity Measures of V1 and V2 Components

The productivity of the JCV is well known, as is the frequency with which some V1 and V2 components appear. It is useful, having established a reasonably large collection of JCVs, to use this to analyze the frequency of usage of the components.

For the purpose of this analysis, the V1s and V2s were extracted, counted and ranked from:

- the full collection of possible JCVs collected from the n-gram corpus;
- the synthesized JCVs which had matches in the n-gram corpus;
- the JCVs extracted from the combined lexicon;
- the highest-ranked 10,000 JCVs in the full collection (to see if there is a bias in component use according to the how common the JCV is).

In addition, V2 rankings collected by Kubota (1992) from her own corpus analysis and from an earlier published collection (Nomura and Ishii, 1987) were added for comparison.

The number of V1s and V2s which were extracted are shown in Table 3. As can be seen, over 80% of the V2s in the combined file are only found in the relatively low-frequency JCVs.

While there was some correlation of the frequency rankings of the V1s and V2s, there were also some notable differences. This can be seen in Tables 4 and 5, which show the 20 most common components as they were found in the combined JCV list, with their comparative rankings in the other lists.

With regard to the V1s in Table 4, it will be noted that there are some which do not appear in

V1	Combined n-gram	Synth. n-gram	Lexicon	First 10,000
通り	1	2	190	1
再び	2	1	642	3
思い	3	3	11	2
見て	4	4	520	4
限り	5	–	857	9
同じ	6	27	–	51
入り	7	13	52	29
余り	8	–	726	18
使い	9	11	45	13
出て	10	7	823	37
取り	11	5	1	5
入れ	12	19	49	24
行き	13	16	17	22
買い	14	24	24	11
持ち	15	22	25	19
作り	16	15	68	30
考え	17	17	113	25
言い	18	9	4	10
及び	19	–	1242	132
感じ	20	30	428	41

Table 4: V1 rankings for the two proposed methods, the lexicon and the most frequent 10,000 JCVs

V2	Combined n-gram	Synth. n-gram	Lexicon	First 10,000	Kubota
始める	1	1	6	1	1
続ける	2	2	3	2	8
過ぎる	3	4	4	5	9
出す	4	3	2	3	3
合う	5	5	5	6	4
得る	6	13	152	11	2
行く	7	10	59	12	–
初める	8	8	106	43	–
終わる	9	7	74	7	13
切る	10	11	9	10	5
難く	11	–	–	28	–
込む	12	6	1	4	7
多く	13	–	–	56	–
直す	14	9	10	9	11
見る	15	14	109	27	–
忘れる	16	15	18	16	39
上げる	17	12	8	8	21
歩く	18	19	50	25	34
来る	19	22	291	31	–
頑張る	20	–	–	68	–

Table 5: V2 rankings for the two proposed methods, the lexicon and the most frequent 10,000 JCVs

all lists, or have very different rankings. These almost all relate to differing interpretations at to what comprises a JCV. For example:

a. the V1s 限り *kagiri* “restricted”, 余り *amari* “remain” and 及び *oyobi* “and”, while deriving from verbs, are almost invariably use as conjunctions or adverbials in modern Japanese, and hence would not normally be part of a JCV. They should be added to the filter rules.

b. 再び *futatabi* “again” is more commonly regarded as an adverb, and should also probably be excluded.

c. the two “te-form” V1s (見て *mite* and 出て *dete*) could be classed as either part of JCVs, or as the common (V$1_{\mathrm{te-form}}$,V2) sequence which has the sense of simultaneous occurrence or activity. They are often excluded from JCV classifications.

It will be noted that few of the high-ranking V1s in the lexicon appear in Table 4. On inspection it proved that they mostly lie in the 20–40 range in the n-gram lists. Given that dictionary compilers concentrate on JCVs which are polysemous or have idiosynchratic meaning, a lack of frequency alignment with corpus-based lists is to be expected.

While there is generally good agreement between the rankings of the lists of potential V2s in Table 5, some attract similar comments:

a. two of the potential V2s (難く *gataku* and 多く *ôku*) are clearly derived from adjectives, and should be added to the filter rules.

b. V2s which rank lower in the lexicon list (得る *eru* “to attain”, 初める *hajimeru* “to start”, etc.) are usually parts of semantically regular JCVs, and hence are less likely to be in a dictionary.

c. the appearance of 頑張る *gaNbaru* “to persist; to insist on; to stand firm” is of interest. It is not usually regarded as a JCV component, yet from its occurrence in the n-gram lists, it is clearly being used as such.

The foregoing comments are confined to the V1 and V2 components appearing among the 20 highest ranking counts in the combined list; similar comments can be made about a number of lower-ranking components. There is scope for considerable analysis of the ranking lists of V1 and V2 components.

6 Conclusions and Future Work

The work so far in the project has demonstrated that large numbers of JCVs are in regular use and can be detected through the application of NLP techniques to corpora. The two detection techniques which have been developed and tested have been demonstrated to have good levels of precision, especially in the case of the JCV synthesis method.

A substantial list of JCVs which are not recorded in commonly-used dictionaries has been identified for further study. In addition data on the frequency of usage of JCVs and their V1 and V2 components has been collected and can be made available for other Japanese NLP projects.

Future work in the project will include the development and testing of the meanings of unrecorded JCVs. The approach developed in earlier work (Uchiyama et al., 2005), which employs rule-based and statistical methods based on extensive classification of the V1 and V2 components, will be followed.

7 Acknowledgment

The assistance of Mr Jack Halpern, who has kindly provided a copy of his Japanese Linguistic Database for use in the project, is gratefully acknowledged.

References

Timothy Baldwin and Francis Bond. 2002. Multiword expressions: Some problems for Japanese NLP. In *Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing (Japan)*, pages 379–382, Keihanna, Japan.

James Breen. 2004a. Expanding the Lexicon: the Search for Abbreviations. In *Proceedings of the Papillon Workshop*, Grenoble, France.

James Breen. 2004b. JMdict: a Japanese-Multilingual Dictionary. In *Proceedings of the Multilingual Lin-*

guistics Resources Workshop, Geneva, Switzerland. COLING2004.

James Breen. 2009. Kanjidic/Kanjd212 Project. `http://www.csse.monash.edu.au/~jwb/kanjidic.html`.

Jack Halpern. 2008. Japanese Lexical Database. `http://www.cjk.org/cjk/samples/japword.htm`.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi, editors. 1999. *GoiTaikei - A Japanese Lexicon*. Iwanami Shoten.

Mariko Kubota. 1992. *Japanese Compound Verbs*. Ph.D. thesis, Monash University.

Taku Kudo and Hideto Kazawa. 2007. Japanese Web N-gram Corpus version 1. `http://www.ldc.upenn.edu/Catalog/docs/LDC2009T08/`.

Taku Kudo. 2008. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. `http://mecab.sourceforge.net/`.

Jonathan Kummerfeld and James Curran. 2008. Classification of verb particle constructions with the google web1t corpus. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, pages 55–63, Hobart, Australia, December.

Akira Matsumura, editor. 1995. *Daijirin*. Sanseido, 2nd edition.

Masaaki Nomura and Masahiko Ishii. 1987. **複合動詞資料集** *(Fukugôdôshi Shiryôshû - Compound Verb Collection)*. National Language Research Institute, Tokyo.

Nobukazu Ootsuka, editor. 1998. *Kôjien*. Iwanami Shoten, 5th edition.

Masayoshi Shibatani. 1990. *The Languages of Japan*. Cambridge University Press, Cambridge, UK.

Yoshiko Tagashira and Jean Hoff. 1986. *Handbook of Japanese Compound Verbs*. Hokuseido Press (Tokyo).

Natsuko Tsujimura. 2006. *An Introduction to Japanese Linguistics*. Blackwell, Oxford, UK, 2nd edition.

Kiyoko Uchiyama, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese Compound Verbs. *Computer Speech and Language*.