# Segmentation and Translation of Japanese Multi-word Loanwords

**James Breen**
The University of Melbourne
jimbreen@gmail.com

**Timothy Baldwin**
The University of Melbourne
tb@ldwin.net

**Francis Bond**
Nanyang Technological University, Singapore
bond@ieee.org

## Abstract

The Japanese language has absorbed large numbers of loanwords from many languages, in particular English. As well as using single loanwords, compound nouns, multiword expressions (MWEs), etc. constructed from loanwords can be found in use in very large quantities. In this paper we describe a system which has been developed to segment Japanese loanword MWEs and construct likely English translations. The system, which leverages the availability of large bilingual dictionaries of loanwords and English *n*-gram corpora, achieves high levels of accuracy in discriminating between single loanwords and MWEs, and in segmenting MWEs. It also generates useful translations of MWEs, and has the potential to being a major aid to lexicographers in this area.

## 1 Introduction

The work described in this paper is part of a broader project to identify unrecorded lexemes, including neologisms, in Japanese corpora. Since such lexemes include the range of lexical units capable of inclusion in Japanese monolingual and bilingual dictionaries, it is important to be able to identify and extract a range of such units, including compound nouns, collocations and other multiword expressions (MWEs: Sag et al. (2002; Baldwin and Kim (2009)).

Unlike some languages, where there is official opposition to the incorporation of foreign words, Japanese has assimilated a large number of such words, to the extent that they constitute a sizeable proportion of the lexicon. For example, over 10% of the entries and sub-entries in the major Kenkyūsha New Japanese-English Dictionary (5th ed.) (Toshiro et al., 2003) are wholly or partly made up of loanwords. In addition there are several published dictionaries consisting solely of such loanwords. Estimates of the number of loanwords and particularly MWEs incorporating loanwords in Japanese range into the hundreds of thousands. While a considerable number of loanwords have been taken from Portuguese, Dutch, French, etc., the overwhelming majority are from English.

Loanwords are taken into Japanese by adapting the source language pronunciation to conform to the relatively restricted set of syllabic phonemes used in Japanese. Thus "blog" becomes *burogu*, and "elastic" becomes *erasutikku*. When written, the syllables of the loanword are transcribed in the *katakana* syllabic script (ブログ, エラスティック), which in modern Japanese is primarily used for this purpose. This use of a specific script means possible loanwords are generally readily identifiable in text and can be extracted without complex morphological analysis.

The focus of this study is on multiword loanwords. This is because there are now large collections of basic Japanese loanwords along with their translations, and it appears that many new loanwords are formed by adopting or assembling MWEs using known loanwords. As evidence of this, we can cite the numbers of *katakana* sequences in the the Google Japanese *n*-gram corpus (Kudo and Kazawa, 2007). Of the 2.6 million 1-grams in that cor-

James Breen, Timothy Baldwin and Francis Bond. 2012. Segmentation and Translation of Japanese Multi-word Loanwords. In *Proceedings of Australasian Language Technology Association Workshop*, pages 61–69.

pus, approximately 1.6 million are in *katakana* or other characters used in loanwords.[1] Inspection of those 1-grams indicates that once the words that are in available dictionaries are removed, the majority of the more common members are MWEs which had not been segmented during the generation of the corpus. Moreover the *n*-gram corpus also contains 2.6 million 2-grams and 900,000 3-grams written in *katakana*. Even after allowing for the multiple-counting between the 1, 2 and 3-grams, and the imperfections in the segmentation of the katakana sequences, it is clear that the vast numbers of multiword loanwords in use are a fruitful area for investigation with a view to extraction and translation.

In the work presented in this paper we describe a system which has been developed to segment Japanese loanword MWEs and construct likely English translations, with the ultimate aim of being part of a toolkit to aid the lexicographer. The system builds on the availability of large collections of translated loanwords and a large English *n*-gram corpus, and in testing is performing with high levels of precision and recall.

## 2 Prior Work

There has not been a large amount of work published on the automatic and semi-automatic extraction and translation of Japanese loanwords. Much that has been reported has been in areas such as back-transliteration (Matsuo et al., 1996; Knight and Graehl, 1998; Bilac and Tanaka, 2004), or on extraction from parallel bilingual corpora (Brill et al., 2001). More recently work has been carried out exploring combinations of dictionaries and corpora (Nakazawa et al., 2005), although this lead does not seem to have been followed further.

Both Bilac and Tanaka (2004) and Nakazawa et al. (2005) address the issue of segmentation of MWEs. This is discussed in 3.1 below.

[1] In addition to *katakana,* loanwords use the ー *(chōoN)* character for indicating lengthened vowels, and on rare occasions the ヽ and ヾ syllable repetition characters.

## 3 Role and Nature of Katakana Words in Japanese

As mentioned above, loan words in Japanese are currently written in the *katakana* script. This is an orthographical convention that has been applied relatively strictly since the late 1940s, when major script reforms were carried out. Prior to then loanwords were also written using the *hiragana* syllabary and on occasions *kanji* (Chinese characters).

The *katakana* script is not used exclusively for loanwords. Other usage includes:

a. transcription of foreign person and place names and other named entities. Many Japanese companies use names which are transcribed in *katakana*. Chinese (and Korean) place names and person names, although they are usually available in *kanji* are often written in *katakana* transliterations;
b. the scientific names of plants, animals, etc.
c. onomatopoeic words and expressions, although these are often also written in *hiragana;*
d. occasionally for emphasis and in some contexts for slang words, in a similar fashion to the use of italics in English.

The proportion of *katakana* words that were not loanwords was measured by Brill et al. (2001) at about 13%. (The impact and handling of these is discussed briefly at the end of Section 4.)

When considering the extraction of Japanese loan words from text, there are a number of issues which need to be addressed.

### 3.1 Segmentation

As mentioned above, many loanwords appear in the form of MWEs, and their correct analysis and handling often requires separation into their composite words. In Japanese there is a convention that loanword MWEs have a "middle-dot" punctuation character ( ・ ) inserted between the components, however while this convention is usually followed in dictionaries, it is rarely applied elsewhere. Web search engines typically ignore this character when indexing, and a search for a very common MWE: トマトソース

*tomatosōsu* "tomato sauce", reveals that it almost always appears as an undifferentiated string. Moreover, the situation is confused by the common use of the ・ character to separate items in lists, in a manner similar to a semi-colon in English. In practical terms, systems dealing with loanwords MWEs must be prepared to do their own segmentation.

One approach to segmentation is to utilize a Japanese morphological analysis system. These have traditionally been weak in the area of segmentation of loanwords, and tend to default to treating long *katakana* strings as 1-grams. In testing a list of loanwords and MWEs using the ChaSen system (Matsumoto et al., 2003), Bilac and Tanaka (2004) report a precision and recall of approximately 0.65 on the segmentation, with a tendency to undersegment being the main problem. Nakazawa et al. (2005) report a similar tendency with the JUMAN morphological analyzer (Kurohashi and Nagao, 1998). The problem was most likely due to the relatively poor representation of loanwords in the morpheme lexicons used by these systems. For example the IPADIC lexicon (Asahara and Matsumoto, 2003) used at that time only had about 20,000 words in *katakana*, and many of those were proper nouns.

In this study, we use the MeCab morphological analyzer (Kudo et al., 2004) with the recently-developed *UniDic* lexicon (Den et al., 2007), as discussed below.

As they were largely dealing with non-lexicalized words, Bilac and Tanaka (2004) used a dynamic programming model trained on a relatively small (13,000) list of *katakana* words, and reported a high precision in their segmentation. Nakazawa et al. (2005) used a larger lexicon in combination with the JUMAN analyzer and reported a similar high precision.

### 3.2 Non-English Words

A number of loanwords are taken from languages other than English. The JMdict dictionary (Breen, 2004) has approximately 44,000 loanwords, of which 4% are marked as coming from other languages. Inspection of a sample of the 22,000 entries in the Gakken *A Dictionary of Katakana Words* (Kabasawa and Satō, 2003) indicates a similar proportion. (In both dictionaries loanwords from languages other than English are marked with their source language.) This relatively small number is known to cause some problems with generating translations through transliterations based on English, but the overall impact is not very significant.

### 3.3 Pseudo-English Constructions

A number of *katakana* MWEs are constructions of two or more English words forming a term which does not occur in English. An example is バージョンアップ *bājon'appu* "version up", meaning upgrading software, etc. These constructions are known in Japanese as 和製英語 *wasei eigo* "Japanese-made English". Inspection of the JMdict and Gakken dictionaries indicate they make up approximately 2% of *katakana* terms, and while a nuisance are not considered to be a significant problem.

### 3.4 Orthographical Variants

Written Japanese has a relatively high incidence of multiple surface forms of words, and this particularly applies to loan words. Many result from different interpretations of the pronunciation of the source language term, e.g. the word for "diamond" is both ダイヤモンド *daiyamondo* and ダイアモンド *daiamondo*, with the two occurring in approximately equal proportions. (The JMdict dictionary records 10 variants for the word "vibraphone", and 9 each for "whiskey" and "vodka".) In some cases two different words have been formed from the one source word, e.g. the English word "truck" was borrowed twice to form トラック *torakku* meaning "truck, lorry" and トロッコ *torokku* meaning "trolley, rail car". Having reasonably complete coverage of alternative surface forms is important in the present project.

## 4 Approach to Segmentation and MWE Translation

As our goal is the extraction and translation of loanword MWEs, we need to address the twin tasks of segmentation of the MWEs into their constituent source-language components, and generation of appropriate transla-

tions for the MWEs as a whole. While the back-transliteration approaches in previous studies have been quite successful, and have an important role in handling single-word loanwords, we decided to experiment with an alternative approach which builds on the large lexicon and *n*-gram corpus resources which are now available. This approach, which we have labelled "CLST" (Corpus-based Loanword Segmentation and Translation) builds upon a direction suggested in Nakazawa et al. (2005) in that it uses a large English *n*-gram corpus both to validate alternative segmentations and select candidate translations.

The three key resources used in CLST are:

a. a dictionary of *katakana* words which has been assembled from:
   i. the entries with *katakana* headwords or readings in the JMdict dictionary;
   ii. the entries with *katakana* headwords in the Kenkyūsha New Japanese-English Dictionary;
   iii. the *katakana* entries in the Eijiro dictionary database;[2]
   iv. the *katakana* entries in a number of technical glossaries covering biomedical topics, engineering, finance, law, etc.
   v. the named-entities in *katakana* from the JMnedict named-entity database.[3]

   This dictionary, which contains both base words and MWEs, includes short English translations which, where appropriate, have been split into identifiable senses. It contains a total of 270,000 entries.

b. a collection of 160,000 *katakana* words drawn from the headwords of the dictionary above. It has been formed by splitting the known MWEs into their components where this can be carried out reliably;

c. the Google English *n*-gram corpus[4]. This contains 1-grams to 5-grams collected from the Web in 2006, along with frequency counts. In the present project we use a subset of the corpus consisting only of case-folded alphabetic tokens.

| |
|---|
| ソーシャル・ブックマーク・サービス |
| ソーシャル・ブックマーク・サー・ビス |
| ソーシャル・ブック・マーク・サービス |
| ソーシャル・ブック・マーク・サー・ビス |
| ソー・シャル・ブックマーク・サービス |
| ソー・シャル・ブックマーク・サー・ビス |
| ソー・シャル・ブック・マーク・サービス |
| ソー・シャル・ブック・マーク・サー・ビス |

Table 1: Segmentation Example

The process of segmenting an MWE and deriving a translation is as follows:

a. using the *katakana* words in (b) above, generate all possible segmentations of the MWE. A recursive algorithm is used for this. Table 1 shows the segments derived for the MWE ソーシャルブックマークサービス *sōsharubukkumākusābisu* "social bookmark service".

b. for each possible segmentation of an MWE, assemble one or more possible glosses as follows:
   i. take each element in the segmented MWE, extract the first gloss in the dictionary and assemble a composite potential translation by simply concatenating the glosses. Where there are multiple senses, extract the first gloss from each and assemble all possible combinations. (The first gloss is being used as lexicographers typically place the most relevant and succinct translation first, and this has been observed to be often the most useful when building composite glosses.) As examples, for ソーシャル・ブックマーク・サービス the element サービス has two senses "service" and "goods or services without charge", so the possible glosses were "social bookmark service" and "social bookmark goods or services without charge". For ソーシャル・ブック・マーク・サービス the element マーク has senses of "mark", "paying attention",

[2] http://www.eijiro.jp/e/index.htm

[3] http://www.csse.monash.edu.au/~jwb/enamdict_doc.html

[4] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

"markup" and "Mach", so the potential glosses were "social book mark service", "social book markup service", "social book Mach service", etc. A total of 48 potential translations were assembled for this MWE.

ii. where the senses are tagged as being affixes, also create combinations where the gloss is attached to the preceding or following gloss as appropriate.

iii. if the entire MWE is in the dictionary, extract its gloss as well.

It may seem unusual that a single sense is being sought for an MWE with polysemous elements. This comes about because in Japanese polysemous loanwords are almost always due to them being derived from multiple source words. For example ランプ *ranpu* has three senses reflecting that it results from the borrowing of three distinct English words: "lamp", "ramp" and "rump". On the other hand, MWEs containing ランプ, such as ハロゲンランプ *harogenranpu* "halogen lamp" or オンランプ *onranpu* "on-ramp" almost invariably are associated with one sense or another.

c. attempt to match the potential translations with the English *n*-grams, and where a match does exist, extract the frequency data. For the example above, only "social bookmark service", which resulted from the ソーシャル・ブックマーク・サービス segmentation, was matched successfully;

d. where match(es) result, choose the one with the highest frequency as both the most likely segmentation of the MWE and the candidate translation.

The approach described above assumes that the term being analyzed is a MWE, when in fact it may well be a single word. In the case of as-yet unrecorded words we would expect that either no segmentation is accepted or that any possible segmentations have relatively low frequencies associated with the potential translations, and hence can be flagged for closer inspection. As some of the testing described below involves deciding whether a term is or is not a MWE, we have enabled the system to handle single terms as well by checking the unsegmented term against the dictionary and extracting *n*-gram frequency counts for the glosses. This enables the detection and rejection of possible spurious segmentations. As an example of this, the word ボールト *bōruto* "vault" occurs in one of the test files described in the following section. A possible segmentation (ボー・ルト) was generated with potential translations of "bow root" and "baud root". The first of these occurs in the English 2-grams with a frequency of 63, however "vault" itself has a very high frequency in the 1-grams so the segmentation would be rejected.

As pointed out above, a number of *katakana* words are not loanwords. For the most part these would not be handled by the CLST segmentation/translation process as they would not be reduced to a set of known segments, and would be typically reported as failures. The transliteration approaches in earlier studies also have problems with these words. Some of the non-loanwords, such as scientific names of plants, animals, etc. or words written in *katakana* for emphasis, can be detected and filtered prior to attempted processing simply by comparing the *katakana* form with the equivalent *hiragana* form found in dictionaries. Some of the occurrences of Chinese and Japanese names in text can be detected at extraction time, as such names are often written in forms such as "...金鍾泌(キムジョンピル)..."[5].

## 5 Evaluation

Evaluation of the CLST system was carried out in two stages: testing the segmentation using data used in previous studies to ensure it was discriminating between single loanwords and MWEs, and testing against a collection of MWEs to evaluate the quality of the translations proposed.

### 5.1 Segmentation

The initial tests of CLST were of the segmentation function and the identification of single words/MWEs. We were fortunate to be

[5]Kim Jong-Pil, a former South Korean politician.

| Method | Set | Recall | Precision | F |
|---|---|---|---|---|
| CLST | EDR | 98.67 | 100.00 | 99.33 |
| MeCab | EDR | 92.67 | 97.20 | 94.88 |
| CLST | NTCIR-2 | 94.87 | 100.00 | 97.37 |
| MeCab | NTCIR-2 | 95.52 | 92.75 | 89.37 |

Table 2: Results from Segmentation Tests

able to use the same data used by Bilac and Tanaka (2004), which consisted of 150 out-of-lexicon *katakana* terms from the EDR corpus (EDR, 1995) and 78 from the NTCIR-2 test collection (Kando et al., 2001). The terms were hand-marked as to whether they were single words or MWEs. Unfortunately we detected some problems with this marking, for example シェークスピア *shēkusupia* "Shakespeare" had been segmented (shake + spear) whereas ホールバーニング *hōrubāningu* "hole burning" had been left as a single word. We considered it inappropriate to use this data without amending these terms. As a consequence of this we are not able to make a direct comparison with the results reported in Bilac and Tanaka (2004). Using the corrected data we analyzed the two datasets and report the results in Table 2. We include the results from analyzing the data using *MeCab/UniDic* as well for comparison. The precision and recall achieved was higher than that reported in Bilac and Tanaka (2004). As in Bilac and Tanaka (2004), we calculate the scores as follows: $N$ is the number of terms in the set, $c$ is the number of terms correctly segmented or identified as 1-grams, $e$ is the number of terms incorrectly segmented or identified, and $n = c + e$. Recall is calculated as $\frac{c}{N}$, precision as $\frac{c}{n}$, and the F-measure as $\frac{2 \times precision \times recall}{precision + recall}$.

As can be seen our CLST approach has achieved a high degree of accuracy in identifying 1-grams and segmenting the MWEs. Although it was not part of the test, it also proposed the correct translations for almost all the MWEs. The less-than-perfect recall is entirely due to the few cases where either no segmentation was proposed, or where the proposed segmentation could not be validated with the English $n$-grams.

The performance of *MeCab/UniDic* is interesting, as it also has achieved a high level of accuracy. This is despite the *UniDic* lexicon only having approximately 55,000 *katakana* words, and the fact that it is operating outside the textual context for which it has been trained. Its main shortcoming is that it tends to over-segment, which is a contrast to the performance of *ChaSen/IPADIC* reported in Bilac and Tanaka (2004) where under-segmentation was the problem.

### 5.2 Translation

The second set of tests of CLST was directed at developing translations for MWEs. The initial translation tests were carried out on two sets of data, each containing 100 MWEs. The sets of data were obtained as follows:

a. the 100 highest-frequency MWEs were selected from the Google Japanese 2-grams. The list of potential MWEs had to be manually edited as the 2-grams contain a large number of over-segmented words, e.g. アイコン *aikon* "icon" was split: アイコ+ン, and オークション *ōkushon* "auction" was split オーク+ション;
b. the *katakana* sequences were extracted from a large collection of articles from 1999 in the *Mainichi Shimbun* (a Japanese daily newspaper), and the 100 highest-frequency MWEs extracted.

After the data sets were processed by CLST the results were examined to determine if the segmentations had been carried out correctly, and to assess the quality of the proposed translations. The translations were graded into three groups: (1) acceptable as a dictionary gloss, (2) understandable, but in need of improvement, and (3) wrong or inadequate. An example of a translation graded as 2 is マイナスイオン *mainasuion* "minus ion", where "negative ion" would be better, and one graded as 3 is フリーマーケット *furīmāketto* "free market", where the correct translation is "flea market". For the most part the translations receiving a grading of 2 were the same as would have been produced by a back-transliteration system, and in many cases they were the *wasei eigo* constructions described above.

Some example segmentations, possible translations and gradings are in Table 3.

| MWE | Segmentation | Possible Translation | Frequency | Grade |
|---|---|---|---|---|
| ログインヘルプ | ログイン・ヘルプ | login help | 541097 | 1 |
| ログインヘルプ | ログ・イン・ヘルプ | log in help | 169972 | - |
| キーワードランキング | キーワード・ランキング | keyword ranking | 39818 | 1 |
| キーワードランキング | キー・ワード・ランキング | key word ranking | 74 | - |
| キャリアアップ | キャリア・アップ | career up | 13043 | 2 |
| キャリアアップ | キャリア・アップ | carrier up | 2552 | - |
| キャリアアップ | キャリア・アップ | career close up | 195 | - |
| キャリアアップ | キャリア・アップ | career being over | 188 | - |
| キャリアアップ | キャリア・アップ | carrier increasing | 54 | - |

Table 3: Sample Segmentations and Translations

| Data Set | Failed Segmentations | Translation Grades 1 | 2 | 3 | Precision | Recall | F |
|---|---|---|---|---|---|---|---|
| Google | 9 | 66 | 24 | 1 | 98.90 | 90.00 | 94.24 |
| Mainichi (Set 1) | 3 | 77 | 19 | 1 | 98.97 | 96.00 | 97.46 |
| Mainichi (Set 2) | 1 | 83 | 16 | 0 | 100.00 | 99.00 | 99.50 |

Table 4: Results from Translation Tests

The assessments of the segmentation and the gradings of the translations are given in Table 4. The precision, recall and F measures have been calculated on the basis that a grade of 2 or better for a translation is a satisfactory outcome.

A brief analysis was conducted on samples of 25 MWEs from each test set to ascertain whether they were already in dictionaries, or the degree to which they were suitable for inclusion in a dictionary. The dictionaries used for this evaluation were the commercial Kenkyusha Online Dictionary Service[6] which has eighteen Japanese, Japanese-English and English-Japanese dictionaries in its search tool, and the free WWWJDIC online dictionary[7], which has the JMdict and JMnedict dictionaries, as well as numerous glossaries.

Of the 50 MWEs sampled:

a. 34 (68%) were in dictionaries;
b. 11 (22%) were considered suitable for inclusion in a dictionary. In some cases the generated translation was not considered appropriate without some modification, i.e. it had been categorized as "2";
c. 3 (6%) were proper names (e.g. hotels, software packages);
d. 2 (4%) were not considered suitable for inclusion in a dictionary as they were simple collocations such as メニューエリア *menyūeria* "menu area".

As the tests described above were carried out on sets of frequently-occurring MWEs, it was considered appropriate that some further testing be carried out on less common loanword MWEs. Therefore an additional set of 100 lower-frequency MWEs which did not occur in the dictionaries mentioned above were extracted from the *Mainichi Shimbun* articles and were processed by the CLST system. Of these 100 MWEs:

a. 1 was not successfully segmented;
b. 83 of the derived translations were classified as "1" and 16 as "2";
c. 8 were proper names.

The suitability of these MWEs for possible inclusion in a bilingual dictionary was also evaluated. In fact the overwhelming majority of the MWEs were relatively straightforward collocations, e.g. マラソンランナー *marasonrannā* "marathon runner" and ロックコンサート *rokkukonsāto* "rock concert", and were deemed to be not really appropriate as dictionary entries. 5 terms were assessed as being dictionary candidates.

[6] http://kod.kenkyusha.co.jp/service/

[7] http://www.edrdg.org/cgi-bin/wwwjdic/wwwjdic?1C

Several of these, e.g. ゴールドプラン *gōrudopuran* "gold plan" and エーススストライカー *ēsusutoraikā* "ace striker" were category 2 translations, and their possible inclusion in a dictionary would largely be because their meanings are not readily apparent from the component words, and an expanded gloss would be required.

Some points which emerge from the analysis of the results of the tests described above are:

a. to some extent, the Google *n*-gram test data had a bias towards the types of constructions favoured by Japanese webpage designers, e.g. ショッピングトップ *shoppingutoppu* "shopping top", which possibly inflated the proportion of translations being scored with a 2;
b. some of the problems leading to a failure to segment the MWEs were due to the way the English *n*-gram files were constructed. Words with apostrophes were split, so that "men's" was recorded as a bigram: "men+'s". This situation is not currently handled in CLST, which led to some of the segmentation failures, e.g. with メンズアイテム *menzuaitemu* "men's item";

## 6 Conclusion and Future Work

In this paper we have described the CLST (Corpus-based Loanword Segmentation and Translation) system which has been developed to segment Japanese loanword MWEs and construct likely English translations. The system, which leverages the availability of large bilingual dictionaries of loanwords and English *n*-gram corpora, is achieving high levels of accuracy in discriminating between single loanwords and MWEs, and in segmenting MWEs. It is also generating useful translations of MWEs, and has the potential to being a major aide both to lexicography in this area, and to translating.

The apparent success of an approach based on a combination of large corpora and relatively simple heuristics is consistent with the conclusions reached in a number of earlier investigations (Banko and Brill, 2001; Lapata and Keller, 2004).

Although the CLST system is performing at a high level, there are a number of areas where refinement and experimentation on possible enhancements can be carried out. They include:

a. instead of using the "first-gloss" heuristic, experiment with using all available glosses. This would be at the price of increased processing time, but may improve the performance of the segmentation and translation;
b. align the searching of the *n*-gram corpus to cater for the manner in which words with apostrophes, etc. are segmented. At present this is not handled correctly;
c. tune the presentation of the glosses in the dictionaries so that they will match better with the contents of the *n*-gram corpus. At present the dictionary used is simply a concatenation of several sources, and does not take into account such things as the *n*-gram corpus having hyphenated words segmented;
d. extend the system by incorporating a back-transliteration module such as that reported in Bilac and Tanaka (2004). This would cater for single loanwords and thus provide more complete coverage.

## References


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

Timothy Baldwin and Su Nam Kim. 2009. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*. CRC Press, Boca Raton, USA, 2nd edition.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the ACL and 10th Conference of the EACL (ACL-EACL 2001)*, Toulouse, France.

Slaven Bilac and Hozumi Tanaka. 2004. A Hybrid Back-transliteration System for Japanese. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Geneva, Switzerland.

James Breen. 2004. JMdict: a Japanese-Multilingual Dictionary. In *Proceedings of the COLING-2004 Workshop on Multilingual Resources*, pages 65–72, Geneva, Switzerland.

Eric Brill, Gary Kacmarcik, and Chris Brockett. 2001. Automatically Harvesting Katakana-

English Term Pairs from Search Engine Query Logs. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, November 27-30, 2001, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan*, pages 393–399.

Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Minematsu, Kiyotaka Uchimoto, and Hanae Koiso. 2007. The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics (in Japanese). *Japanese Linguistics*, 22:101–123.

EDR, 1995. *EDR Electronic Dictionary Technical Guide*. Japan Electronic Dictionary Research Institute, Ltd. (in Japanese).

Yōichi Kabasawa and Morio Satō, editors. 2003. *A Dictionary of Katakana Words*. Gakken.

Noriko Kando, Kazuko Kuriyama, and Masaharu Yoshioka. 2001. Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop. In *Proceedings of the Second NTCIR Workshop*, Jeju, Korea.

Kevin Knight and Jonathan Graehl. 1998. Machine Transliteration. *Comput. Linguist.*, 24(4):599–612, December.

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

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 230–237, Barcelona, Spain.

Sadao Kurohashi and Makoto Nagao. 1998. *Nihongo keitai-kaiseki sisutemu JUMAN* [Japanese morphological analysis system JUMAN] version 3.5. Technical report, Kyoto University. (in Japanese).

Mirella Lapata and Frank Keller. 2004. The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proceedings of the Human Langauge Techinology Conference and Conference on Empirical Methods in National Language Processing (HLT/NAACL-2004)*, pages 121–128, Boston, USA.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara. 2003. *Japanese Morphological Analysis System ChaSen Version 2.3.3 Manual*. Technical report, NAIST.

Yoshihiro Matsuo, Mamiko Hatayama, and Satoru Ikehara. 1996. Translation of 'katakana' words using an English dictionary and grammar (in Japanese). In *Proceedings of the Information Processing Society of Japan*, volume 53, pages 65–66.

Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2005. Automatic Acquisition of Basic Katakana Lexicon from a Given Corpus. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 682–693, Jeju, Korea.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Watanabe Toshiro, Edmund Skrzypczak, and Paul Snowdon (eds). 2003. *Kenkyûsha New Japanese-English Dictionary, 5th Edition*. Kenkyûsha.