# The Company They Keep: Extracting Japanese Neologisms Using Language Patterns

**James Breen**
University of Melbourne
Melbourne, Australia
jimbreen@gmail.com

**Timothy Baldwin**
University of Melbourne
Melbourne, Australia
tb@ldwin.net

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

## Abstract

We describe an investigation into the identification and extraction of unrecorded potential lexical items in Japanese text by detecting text passages containing selected language patterns typically associated with such items. We identified a set of suitable patterns, then tested them with two large collections of text drawn from the WWW and Twitter. Samples of the extracted items were evaluated, and it was demonstrated that the approach has considerable potential for identifying terms for later lexicographic analysis.

## 1 Introduction

As the coverage of lexicons (including wordnets) improves, deciding which words should be added next becomes an issue. New words are constantly being added to languages, and existing words are not always covered by current lexical resources.

This paper reports on an investigation as to whether it is possible to identify and extract neologisms (newly created words and expressions) from Japanese text based on the language patterns in which they occur. The genesis of the project is the observation that one often encounters in Japanese text terms which the writer thinks needs some explanation, either because they are new or uncommon. This may be signalled by following the term with phrases such as というのは (*to iu no wa* "as for that which is said ⟨term⟩") and とは (*to wa* "as for ⟨term⟩"), sometimes combined with the reading in parentheses, and then followed by an explanation. The phenomenon is well known to Japanese translators, who often will do a WWW search for "⟨term⟩ とは", etc. when encountering an unfamiliar term in order to identify cases where the term is being described, discussed or otherwise highlighted.

The investigation broadly breaks into two components:

a. the identification of the sorts of language patterns used to describe, discuss, highlight, etc. terms;
b. the extraction and evaluation of terms so targeted by those language patterns.

## 2 Prior Work

Research into the use of linguistic patterns in text to detect terms of interest has taken place in several contexts. In keyphrase extraction Hasan and Ng (2014) have produced a wide-ranging survey of the various techniques used in keyphrase extraction and their relative effectiveness, and Kim et al. (2013) evaluate the performance of a variety of supervised and unsupervised approaches. In term extraction, which is a major part of the broader field of terminology, usually in technical contexts (Kageura (2000)), Takeuchi et al. (2009) adapted the French ACABIT system, which detects morpho-syntactic sequences, to isolate terms in Japanese for later analysis. Le et al. (2013) used patterns of phrases to identify particular Japanese legal documents of interest. Mathieu (2013) successfully adapted a keyphrase extractor for use with Japanese, although its use was restricted to *kanji* sequences. The relationship between a text pattern and a term of interest is a form of **collocation**, i.e. lying between idiomatic expressions and free word combinations. In their survey of collocations in language processing, McKeown and Radev (2000) explore the role of the extraction of collocations in lexicography, although the focus is on the identification

of general terms rather than those which are highlighted as being of interest. Prior published research into the use of Japanese text patterns which target general terms of interest appears to be quite limited. Sato and Kaide (2010) employed a related technique for extracting English–Japanese name pairs by scanning texts for nearby occurrences of *Mr*, *Mrs*, etc. and the Japanese equivalents, e.g. さん (*san*).

## 3 Text Corpora

An essential element of the investigation is the availability of substantial quantities of Japanese text, preferably from a variety of sources. While there are number of Japanese corpora available for use in NLP work, most are actually quite small. In this study we used two text collections:

a. the Kyoto WWW Corpus. This is a collection of 500 million Japanese sentences collected from WWW pages in 2004. The main problem is that it is getting dated, and hence what may have been neologisms at the time of its collation may well be recorded and accepted now, or have totally faded from use.
b. Twitter text. We used a collection of 870 million Japanese text passages extracted from 2014 and 2015 Twitter data. This data provides the opportunity to see how the techniques under investigation perform with with contemporary and at times slangy text.

## 4 Initial Exploration

### 4.1 Pattern Frequencies

Initially we explored whether the text patterns typically associated with the discussion of particular terms occur in sufficient quantities to make them useful search keys by examining their frequencies in the Google Japanese *n*-gram Corpus (Kudo and Kazawa, 2007) (see Table 1).

The high-scoring とは is really a common form of topic marker without any particular association with new or unusual terms, and almost certainly would produce very noisy results if used as a search pattern. On the other hand というのは, という言葉, という意味 and の意味は are typically associated with particular terms and are probably worth further investigation.

| Term | Frequency |
|---|---|
| とは *to wa* "as for" | 169,756,339 |
| というのは/と言うのは *to iu no wa* "as for the said" | 19,134,679/1,207,555 |
| という言葉/ということば *to iu kotoba* "said term" | 5,360,613/167,095 |
| という意味/といういみ *to iu imi* "said term's meaning" | 4,544,800/10,364 |
| という意味は *to iu imi wa* "as for the said term's meaning" | 51,726 |
| の意味は/のいみは *no imi wa* "as for the meaning of" | 1,979,108/1,169 |

Table 1: Google *n*-gram Corpus Frequencies of Text Patterns

### 4.2 Testing Contexts of Known New Terms

We also investigated the sorts of contexts in which known new terms are being used to see if any useful additional patterns could be identified. As an initial exploration 5 terms were chosen from recent additions to the JMdict database (Breen, 2004) which had been noted as popular new words/expressions. The 5 terms were:

- マタハラ *matahara* abbreviation meaning "workplace discrimination against pregnant women";
- こじらせ女子 *kojirase joshi* "girl who has low self-esteem";
- ナマポ *namapo* slang for "welfare recipient"
- 美魔女 *bimajo* "middle-aged woman who looks very young for her age"
- 隠れメタボ *kakure metabo* abbreviation meaning "normal weight obesity"

10 sentences for each term were extracted using a WWW search. While this is clearly a small number of samples, it emerged that there were relatively few of the という/とは/etc. sorts of patterns used; only four occurred a total of seven times in the 50 sentences, and quite a number of the terms being tested occurred encapsulated by some form of parentheses, either "..." (5 occurrences), 「...」

| Term | Frequency |
|---|---|
| 造語 *zōgo* “neologism, coinage” | 232,837 |
| 新語 *shingo* “neologism, new word” | 152,785 |
| 現代用語 *gendai yōgo* “neologism, recent word” | 62,705 |
| 新造語 *shinzōgo* “neologism, new coinage” | 3,978 |
| 言語新作 *gengo shinsaku* “neologism (esp. medical)” | 220 |
| 造語症 *zōgoshō* “neologism (esp. medical)” | <20 |
| ネオロジズム *neorojizumu* “neologism” | <20 |
| ネオレジズム *neorejizumu* “neologism” | <20 |

Table 2: Google *n*-gram Frequencies for Words Meaning Neologism

| Term | Frequency |
|---|---|
| という造語/と言う造語 (*to iu zōgo*) | 10042/491 |
| という新語/と言う新語 (*to iu shingo*) | 3140/117 |
| という現代用語/と言う現代用語 (*to iu gendaiyōgo*) | 50/<20 |

Table 3: Google *n*-gram Frequencies for Extended Neologism Patterns

(10 occurrences) or『...』(1 occurrence).[1]

### 4.3 Explicit Neologism Labelling

We then investigated the use of terms in Japanese which can mean neologism, some of which are given in Table 2, along with their relative frequencies from the Google *n*-grams. As the first three account for almost all the usage, these were investigated further for their use in combination with the という and と言う (“as said”) patterns (Table 3).

As the frequencies for という造語 and という新語 looked promising, a sample of 10 sentences for each was identified via a Google WWW search. These sample sentences indicate the approach seems to have considerable promise. Quite a few relatively new terms, such as ブロマンス *buromansu* “bromance”, were in the samples. It is also interesting to note that all the terms referenced by the patterns were encapsulated in some forms of parentheses.

[1] Japanese orthography uses a variety of symbols for text encapsulation, with the「」pair commonly used where inverted commas are used in English. Other symbols used for this include: ◇, 《》, <>, 〔〕, 【】 and 〖〗

### 4.4 Parenthesized Kana

It has been observed that explanations of terms in Japanese are often accompanied by the reading of the term in parentheses.

To evaluate whether parenthesized readings are present in association with the sorts of language patterns under consideration, and if so whether they are in sufficient quantities to include them in the text analysis, a scan was made of the Kyoto Corpus to extract all sentences containing the patterns described above (という言葉, という造語, etc.). Approximately 2.4 million sentences were extracted, and these were analyzed to determine if they contained parenthesized strings of kana. Only 116 text lines contained “(*kana*)” patterns, and of these there was only one passage containing the “term (reading)” pattern, which indicated that this pattern was not common enough to make it worth a lot of attention.

### 4.5 Expansion of Linguistic Patterns

Discussions were held with several native speakers of Japanese in order to identify possible patterns which may be used with new terms. From this a number of additional patterns were identified. Some also typically followed the term in question, e.g. xx という言葉を聞き *to iu kotoba wo kiki* “hearing the said word xx” and xx という不思議な *to iu fushigi na* “the said xx is strange/curious”.

In addition, a set of phrases which would precede a target word was identified, e.g. この頃よく聞く xx *kono goro yoku kiku*, 近頃よく聞く xx *chikagoro yoku kiku*, and 最近はやりの xx *saikin yoku kiku*, all of which mean “the often heard recently xx”.

This resulted in an overall set of 37 text patterns, some of which have alternative surface forms, e.g. このごろ and この頃 (*kono goro*).

### 4.6 Initial Evaluation of the Language Patterns

The 37 text patterns were tested against the Kyoto WWW Corpus. For each pattern a sample of 20 sentences was examined in detail, with each sentence being classified into one of three groups: sentences which did not directly discuss any identifiable word or term (1); sentences which focussed on a word or term which is already established in one or more lexicons (2); and sentences which focussed on a word or term which is not in an accessible lexicon, and which warrants further investigation (3).

It was clear that some of the text patterns were quite effective in identifying text passages which focus on words or terms of interest, and in some cases the precision appeared to be quite high; in three of the sets of samples (という造語, という新語, という新しい言葉) all of the passages had such a focus, and in another five (という言葉を聞き, という言葉を耳に, という言葉が話題に, という言葉がはやって, という流行語) 85% or more had that focus.

Around half of the sampled passages (349) were classified into Groups 2 and 3, and these were about evenly split between those where the target term was in parentheses (177) and those where it was not (172).

Overall the numbers of sentences extracted with the selected patterns only made up a very small proportion of the sentences in the Corpus. Of the approximately 500 million sentences the high precision patterns only extracted 2,600 sentences. When combined with lower precision patterns the numbers extracted came to about 280,000 (about 0.06%), and it was observed that most of these were from one pattern (という言葉).

## 5 Detailed Investigation

From the original set of 37 patterns, a set of 18 were chosen for further experimentation. The selection process was to choose those patterns which had resulted in the higher proportion of Group 2/3 being detected in the sampling.

Excluded from the original set were three of the more commonly occurring patterns: と言うのは/というのは, というと and といういみ/という意味. Although between them they accounted for about 80% of the of the sentence selections, they performed comparatively poorly in being associated with possibly useful terms. Of the chosen patterns ということば/という言葉 accounted for over 90% of the remaining extracted lines, and 最近はやりの/最近流行の/最近流行りの accounted for a further ∼7%. Thus the overwhelming majority of remaining extractions come from two patterns. They are among the middle-ranking performers according to the sampling, and certainly cannot be ignored. While there are other patterns which performed considerably better in the sampling in terms of precision, the number of actual extractions associated with them is much lower.

### 5.1 Text Scanning and Target Term Extraction

With over a billion lines of text to examine for the presence of the language patterns a reasonably fast searching technique is desirable. The possibility of training a machine learning model was considered, however since we are dealing with a constrained set of patterns a direct pattern-matching approach is clearly more appropriate. Also the nature of the patterns lends itself to a fast character-by-character search using a search tree. The patterns being used begin with only four different characters: こ, と, 近 and 最, and initially each character in a line of text only has to be compared with them to determine whether more of the tree is to be searched. Similarly at each level of the tree only a few characters typically need to be tested.

The 500 million lines in the Kyoto Corpus had 280,574 matches with these patterns, and the 870 million tweets had 130,310 matches. The hit rate for these patterns in Twitter is thus only about 30% that of the WWW text, which is probably indicative of both the brevity of many tweets, and possibly a very different text style for longer tweets.

From the extracted lines of text, it was necessary to isolate the target terms associated with the patterns. The approach taken was:

a. divide the patterns into those where the target usually precedes the pattern (these always begin with という), and those where the target usually follows (the rest).
b. detect and extract text which occurred in some form of parentheses before or after the pattern. The extraction was re-

stricted to parenthesized terms beginning 3 or fewer characters before or after the pattern. This margin was to allow for the occasional punctuation characters and words such as など *nado* "et cetera". Also it was clear that there were occasionally quite long strings of parenthesized text, typically quotations, which were not going to be considered valid lexical items, so the extraction was restricted to strings of up to 10 characters.

c. where there are no parenthesized target strings associated with the text patterns, it is necessary to attempt to extract target terms from the text preceding or following the patterns. Inspection of a number of passages indicated that most likely candidates were made up of combinations such as noun–noun, prefix–noun, noun–suffix, adverb–noun, adjective–noun, etc. and that a reasonable heuristic would be to collect morphemes until one which typically lies on the boundary of an expression, such as a particle or a verb, was encountered.

   To implement this approach, the text following or preceding the pattern was passed through the *MeCab* morphological analyzer (Kudo et al., 2004; Kudo, 2008)[2] operating with the Unidic morpheme lexicon (Den et al., 2007), and adjacent morphemes which met a limited set of part-of-speech (POS) attributes were aggregated

For each text collection the target term extraction as described above was run, the extracted terms were filtered against a large reference lexicon (as the aim of the investigation is to determine whether the method is extracting new or unrecorded terms), and the remaining unlexicalized extractions were sorted and aggregated to determine how often they occur. This is to enable evaluation of the hypothesis that more frequently-occurring terms are more likely to be potential lexical items. The numbers of target terms extracted from the text collections is shown in Table 4.

Some general observations that can be made about these extractions are:

a. the extractions comprise a very small proportion of the text in the two collections. The passages extracted from the WWW Corpus represent only 0.056% of the text and the ones from the Twitter collection only 0.015%.

b. the ということば/という言葉 pattern is relatively much more common in the Kyoto Corpus (0.054%) than in the Twitter collection (0.013%). The 最近流行りの/etc. pattern is also more common in the Kyoto Corpus, but not to such a degree.

c. the target terms are clearly less likely to be parenthesized in Twitter text, and also the target terms associated with という... patterns are more likely to be parenthesized than the others where the target follows the pattern.

## 6 Evaluation of Extracted Target Terms

The extracted terms were then categorized according to the usefulness of the term as a lexical item. This involved examining the term both in the context of the text passage(s) in which it was detected, in other text passages such as those discovered from WWW searches, and in reference material such as glossaries which were not part of the reference lexicon. From this categorization codes were assigned to the terms as follows: (A) in the reference dictionary in different surface form, e.g. partially or fully in kana instead of kanji; (B) an inflected or variant form of existing entry; (C) definitely of interest as it has the potential to be a valid lexical item; (D) other, e.g. a phrase not of particular interest; (E) corrupted text.

Also recorded was whether the occurrences of the terms were parenthesized or not, and which pattern(s) generated the extraction. (This was done for the "C" terms.)

### 6.1 WWW Corpus

Of the 234,733 terms extracted from this Corpus, 68,644 were not in the reference lexicon. Of these 52,277 were terms that occurred only once, and the remainder occurred multiple times (the maximum was 55 times).

A detailed analysis of 120 terms was carried out as follows: the most common 50 terms (13–55 occurrences), a sample of 20 terms which occurred 5 times each, and a sample of 50 terms which occurred once each. The categorization of the terms is shown in Table 5.

[2] http://taku910.github.io/mecab/

| Source | Total lines | Extractions (Paren.) | Extractions (Non-paren.) | None extracted |
|---|---|---|---|---|
| WWW Corpus (all patterns) | 280574 | 124371 | 110362 | 45841 |
| Twitter (all patterns) | 130310 | 37083 | 71995 | 21232 |
| WWW Corpus (という言葉) | 270553 | 122727 | 103111 | 44715 |
| Twitter (という言葉) | 119871 | 36074 | 64254 | 19543 |
| WWW Corpus (最近流行りの) | 6711 | 573 | 5653 | 485 |
| Twitter (最近流行りの) | 7635 | 314 | 6530 | 791 |
| WWW Corpus (the rest) | 3310 | 1071 | 1598 | 641 |
| Twitter (the rest) | 2805 | 696 | 1211 | 898 |

Table 4: Target Term Extraction Counts

| Category | Top 50 | 5 Times (20) | Once (50) |
|---|---|---|---|
| A | 15 | 2 | 0 |
| B | 6 | 6 | 1 |
| C | 18 | 10 | 3 |
| D | 8 | 2 | 46 |
| E | 3 | 0 | 0 |

Table 5: Categorizations of Extracted Text — WWW Corpus

Some examples of the extractions are:
(A) がんばれ *ganbare*: *kana* form of 頑張れ "go for it!"
(A) ガイジン *gaijin*: *katakana* form of 外人 "foreigner"
(B) 愛している *aishiteiru*: from the verb 愛する and meaning "to be in love"
(B) 感動した *kandōshite* — past tense of 感動する "to be moved"
(C) ゲーム性 *gēmusei* "quality of a video game; game rating"
(C) 共創 *kyōsō* "growing together; joint development"
(D) シンプルイズベスト *shinpuru izu besuto* ("Simple Is Best": pop song name)

The relatively high proportion of "C" terms in the multiply-occurring sets (36–50%) is interesting. It might seem intuitively obvious that more commonly used or discussed terms would be more likely to be potential lexical items, but it could well not have been the case. More sampling of the 2, 3 and 4 batches may be appropriate, but it seems clear that multiple occurrences of a term, at least among the terms extracted here, is a signal of its likelihood to be of interest.

### 6.2 Significance of Multiple Occurrences

It was noted that the three singly-occurring C extractions in Table 5 all had reasonably high counts of occurrences in the *n*-gram Corpus (258–473). That raises the question of whether the number of Corpus occurrences is linked or correlated to the usefulness of extracted terms. To test this a sample of 10 of the singly-occurring "D" terms was checked to determine the number of occurrences in the Corpus. 6 of these occurred fewer than 10 times and the others occurred 39, 52, 62 and 1,561 times respectively. Also checked were the Corpus counts of the 8 "D" terms in the "top 50" set. While they varied, they were noticeably lower than the "C" counts. This seems to indi-

cate support for a (quite reasonable) hypothesis that low overall occurrence counts are related to the usefulness of extracted terms.

As a further test of this hypothesis, a set of 2,000 of the singly-extracted terms was chosen and their overall counts in the Corpus established. About 160 of these (8%) each occurred 400 or more times. Examination of a sample of 20 of these more commonly occurring terms resulted in the following category counts: B: 1, C: 14, D: 6.

This is a very different outcome to that shown by the randomly selected singly-extracted terms, and it seems likely that a high extraction count and/or a high overall Corpus count are good indicators that an extracted term has has a chance of being a term of interest. The overall Corpus count of a term may not be a particularly useful metric as it would be difficult to obtain in a general harvesting process. They are only available with the Kyoto WWW Corpus because an $n$-gram corpus and associated utility software are available. However a useful corpus count could well be taken from a different comprehensive corpus such as the Google $n$-gram Corpus.

### 6.3 Twitter Data

A similar analysis was carried out on the text of 2014/15 Twitter data. Some additional analysis was carried out on two aspects of this data: where the text passages were identified as "re-tweets" these were aggregated and a separate investigation made of the term to see if occurrence within a re-tweet was any different to other target terms in terms of usefulness; and since the Twitter text was associated with specific dates, an analysis was made to determine if identified terms were clustered and if so whether this was associated with greater usefulness.

### 6.4 Re-tweets

The fact that Twitter text contains "re-tweets", i.e. messages repeated by Twitter users to their followers, raises a number of issues in terms of the analysis of the text. On the one hand the re-tweeting can seriously distort any analysis which attempts to use frequency information with regard to such things as extracted terms (Lu et al., 2014). On the other hand the fact that a passage is being relayed by Twitter users may in itself be useful in the analysis of the passage.

The actual identification of re-tweets has proved to be a significant problem as we observed that often the users make minor amendments before sending the message as though it were new; often such relays of modified tweets outnumbered the formal re-tweets.

### 6.5 Analysis of Re-tweets

The terms extracted from re-tweets were aggregated and ranked according to the numbers of times the tweet was repeated in order to see if greater repetition was associated with the usefulness of the extracted term. Samples of terms from the over 100 repetitions, 10 to 99 repetitions and 5 repetitions groups were selected and examined. From this examination it was concluded that the occurrence of extracted terms in re-tweets was not a strong indication of usefulness.

A similar investigation was made of a sample of multiply-occurring terms that were not in re-tweets, and as with the investigation of the extractions from the WWW Corpus discussed above, it does appear that the number of times a term is extracted is correlated with the likelihood it is of interest.

As with the WWW Corpus terms, a sample of singly-occurring terms was checked against an $n$-gram corpus, in this case the Google $n$-gram Corpus. A selection of 2,000 singly-extracted candidate terms was matched against the Corpus and a sample of 20 of the higher-ranking terms was evaluated. The results were 7 terms ranked as A or B, 5 as C and 8 as D. While this is only a small sample, it does seem to indicate that a high count in an $n$-gram Corpus indicates a greater likelihood that a term is of interest.

### 6.6 Classification of Names

In contrast to the terms identified in the WWW Corpus, a significant proportion of the terms extracted from Twitter text were names, e.g. *anime* characters, Pokemon characters, singers, etc. In hindsight there probably should have been a category for them, as they have been treated as "D" (not of interest). The fact they are being collected is an indication of the efficacy of the approach.

### 6.7 Issue of Parenthesized Terms

As previously described, the method for extracting possible terms involves either collecting a string of text in parentheses associated with the pattern, or collecting a string of morphemes with restricted POSs associated with the pattern. It is worth examining the relative outcomes of these two approaches to determine if there is a qualitative difference.

Of the approximately 27,000 potential terms extracted from the Twitter text, 12,650 were parenthesized and 14,348 were not parenthesized. Samples were selected from the two groups of terms and examined in detail. From this it was determined that there is no clear domination of one approach over the other.

### 6.8 Burstiness

As the Twitter texts have dates in their metadata it was possible to examine whether multiple occurrences were in bursts, and whether this might be associated with greater or lesser relevance. A sample of ten non-re-tweet multiply-occurring extractions ranging from 16 to 48 occurrences was examined. Of the 10, 3 were clustered into a relatively short period, e.g. a few days, and the other 7 were spread over the whole period of the data. From this is does not appear that clustered multiple occurrences of candidate terms have any particular advantages. The clustering may indicate a degree of topicality of a term, although it may lead to focus on an ephemeral term, when a greater spread of usage over time may indicate more general usage.

## 7 Precision and Recall

The establishment of precision and recall metrics in this area poses an interesting challenge. In terms of precision the testing reported above indicates that some patterns, e.g. という造語/という新語, are likely to result in fairly high levels, however if they result in a relatively small number of lexical items being collected it is of limited use in lexicon building. Casting a wider net and being prepared to sift results is probably a better course.

In terms of measuring recall the typical approach would be to identify how many terms-of-interest there are in a corpus, and test how often they are identified by the extraction method. To probe this issue the 10 candidate terms examined in Section 6.8 above were tested to see how often they occurred in the text, both in and out of the extraction patterns.

In 8 of the 10 terms over half of the occurrences in the Twitter text had been identified, and in three cases over 95% were identified. The proportions identified in the WWW Corpus were noticeably lower.

## 8 Discussion and Conclusions

From the investigations described above, a number of conclusions can be drawn and observations made about the techniques being investigated. Among them are:

a. it is clear that the technique is quite effective in highlighting terms suitable for further investigation, as it identifies candidates that are often very worthy of detailed examination and subsequent lexicalization.
b. it is interesting and not a little frustrating that after all the early work in identifying useful text patterns for identifying possible terms, the outcome has been so totally dominated by two patterns, to the extent that the others may as well be ignored. Several of the other text patterns have demonstrably better precision, but their recall of useful terms is so low as to make them of little use in a practical harvesting exercise. (That is no reason, of course, to exclude them as they add little overhead to process and at the margin can improve the outcome.)
c. the technique can clearly be enhanced by association with an $n$-gram corpus with frequency counts. A term, particularly one which has not been extracted often, is much more likely to be a useful candidate if it has a high $n$-gram count.
d. at present we have no real indication of the recall of the techniques being investigated. Objective analysis of recall would be a major task and best left for further work.
e. one could envisage this technique being attached to something like a Twitter feed, and passing extracted candidate terms through a frequency and $n$-gram analysis, and ultimately on to lexicographers for analysis.

## References


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

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.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273. Association for Computational Linguistics, Baltimore, Maryland. URL `http://www.aclweb.org/anthology/P14-1119`.

Kyo Kageura. 2000. *The Dynamics of Terminology: a descriptive theory of term formations and terminological growth*. John Benjamins.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2013. Automatic keyphrase extraction from scientific articles. *Language Resources and Evaluation*, 47(3):723–742.

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

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.

Tho Thi Ngoc Le, Minh Le Nguyen, and Akira Shimazu. 2013. Unsupervised keyword extraction for Japanese legal documents. In *26th International Conference on Legal Knowledge and Information Systems, Bologna, Italy*.

Yao Lu, Peng Zhang, Yanan Cao, Yue Hu, and Li Guo. 2014. On the frequency distribution of retweets. In *2nd International Conference on Information Technology and Quantitative Management, ITQM 2014*.

Jérôme Mathieu. 2013. Adaptation of a key phrase extractor for Japanese text. In *Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI*.

Kathleen McKeown and Dragomir Radev. 2000. Collocations. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing*, chapter 21. Marcel Dekker.

Satoshi Sato and Sayoko Kaide. 2010. A person-name filter for automatic compilation of bilingual person-name lexicons. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.

Koichi Takeuchi, Kyo Kageura, Teruo Koyama, Béatrice Daille, and Laurent Romary. 2009. Pattern based term extraction using ACABIT system. *CoRR*, abs/0907.2452. URL `http://arxiv.org/abs/0907.2452`.