Identification of Neologisms in Japanese Corpora using Synthesis

James Breen[1], Timothy Baldwin[2], Francis Bond[3]
[1]Monash University, Melbourne, Australia
[2]University of Melbourne, Melbourne, Australia
[3]Nanyang Technological University, Singapore

Abstract

We report on the investigation of a neologism detection approach involving the synthesis of possible Japanese words by mimicking Japanese morphological processes, followed by testing for the presence of candidate words in Japanese corpora. A 2-*kanji* compound generation and classification technique resulted in the detection of significant numbers of unrecorded terms.

Keywords: Japanese, term synthesis; *n*-gram corpora, neology

1. Introduction

This paper reports on part of a major study into the extraction of neologisms from Japanese corpora. In the study three main approaches were explored:

a) analysis of morpheme sequences in Japanese texts to determine the presence of potential new or unrecorded terms. The processes included processing the texts with a morphological analyzer to produce sequences of tagged morphemes, tagging of the morphemes with features derived from combinatory data derived from large lexicons and corpora, and processing the tagged morphemes with rule-based and machine-learning-based chunkers to assemble candidate words and expressions.
b) analysis of language patterns which are often used in Japanese in association with new and emerging terms. These patterns are usually associated with discussions or explanations of new terms. (Breen et al., 2018)
c) synthesis of possible Japanese terms by mimicking Japanese morphological processes, followed by testing for the presence of candidate terms in Japanese corpora.

In this paper we report on the third of these, based on the synthesis of possible Japanese words. (Another component, covering compound verbs, has been reported separately. (Breen and Baldwin, 2009))

A central issue when dealing with neologisms in Japanese is the nature of the orthography, with its use of multiple scripts, primarily *kanji* (Chinese characters), e.g. 猫, 犬, 鳥, 牛, etc., of which approximately 2,500 are in common use and which are used mainly for nouns and the roots of verbs, adjectives, etc.; and the *hiragana* and *katakana* syllabaries, each of 46 symbols plus diacritics. A major issue is the absence of any indication of the boundaries between the syntactic elements in texts. Automated text-processing in Japanese usually relies on morphological analysis software such as *MeCab* (Kudo, 2008) which employ large morpheme lexicons such as *UniDic* (Den et al., 2007), however unrecorded terms will (by definition) usually not be found in these lexicons.

2. Lexicographical Aspects

The study reported here has concentrated on the *identification* of neologisms with a view to possibly including them in a dictionary, and indeed many of those identified have been added to the online *JMdict* Japanese dictionary (Breen, 2004)(a major online Japanese dictionary). The process of assessing potential lexical items for such inclusion is central to lexicography, and along with the establishment of an accurate translation into the target language (English in this case) involves a range of processes and issues in determining their suitability. In our study we have targeted identifying neologisms which are suitable for inclusion in both *coding* and *decoding* dictionaries, an important issue as Japanese dictionaries compiled for native speakers do not typically contain terms such as 凹状 (*ōjō*– concavity) as it is seen as a prefix-plus-noun (concave shape), whereas the term should, and does, occur in dictionaries intended for non-native speakers.

The techniques used in this study focussed on terms that occur in corpora in significant enough quantities to warrant further investigation. In addition, several of the techniques identified 2-*kanji* terms which are typically nouns, and thus would be strong candidates for lexicalization. There remain questions such as whether identified compound nouns or multiword expressions have meanings which are idiomatic, novel or non-intuitive enough to warrant lexicalization. This is widely recognized as one of the major challenges in lexicography (Atkins and Rundell, 2008). Some have expressed the view that with the rise of electronic dictionaries, which do not have the size limitations of printed dictionaries, there is little harm in relaxing this evaluation and including larger numbers of non-idiomatic compounds, expressions, etc. In the case of Japanese there are additional issues such as multiple surface-forms, reading variations, the extensive use of pseudo-English constructions, etc. to take into account. A detailed analysis of the lexicographic issues associated with Japanese dictionary entries, including the criteria for lexicalization, is reported elsewhere (Breen, 2017).

3. Resources

As the experimentation with synthesized terms takes the form of create-and-test, a key requirement is access to appropriate large-scale corpora to test for the presence and usage patterns of the terms.

Corpora such as the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014) which contain relatively small quantities of specially selected documents, or collections of newspaper articles (e.g. the Kyoto Corpus), are unlikely to include new or ephemeral terms in quantities that would enable automated detection and evaluation. The largest collection of accessible text is the WWW, and several studies have shown it is broadly representative in areas such as term frequency (Keller and Lapata 2003.) It is appropriate, therefore, to concentrate on WWW-based text collection for analyzing synthesized terms. (An alternative source of contemporary text is Twitter. A large collection of Twitter-derived text was used in another part of the neologism detection study.)

The main accessible WWW-derived Japanese corpus for this type of testing is the Google *n*-gram Corpus (Kudo and Kazawa, 2007), based on approximately 20 billion text segments extracted from WWW pages, and is provided in the form of sets of 1-grams to 7-grams with counts of the numbers of occurrences. The "grams" in this case are morphemes as identified by morphological analysis software working with large morpheme lexicons (mentioned above.) As the Google corpus only reports *n*-grams which occur 20 or more times a second *n*-gram corpus was assembled using the smaller Kyoto University WWW Corpus, containing about 500 million text segments.

A reference lexicon of known Japanese lexical items is also essential, both for filtering candidate synthesized compounds, and for establishing training data for the evaluation processes. A lexicon of over 1.5 million surface-form/reading combinations was assembled from entries in the following dictionaries: the Kōjien Japanese-Japanese dictionary (広辞苑) with 229,000 surface-form/reading combinations; the Daijirin Japanese-Japanese dictionary (大辞林) with 215,000 combinations; the JMdict (Japanese-Multilingual dictionary) with 240,000 combinations; the JMnedict (Japanese-Multilingual named-entity dictionary) with 744,000 combinations; the Kenkyūsha New Japanese-English Dictionary with 128,000 combinations; the NTT Goi Taikei Japanese Lexicon (語彙大系) with 296,000 combinations ; the Japanese Lexical Database (JLD) from the CJK Dictionary Institute with 302,000 combinations; and miscellaneous subject-specific glossaries (e.g. legal, biomedical, engineering, etc.) with 329,000 combinations.

Also the most recent UniDic morpheme lexicon (756,000 entries) (Den et al., 2007) is a useful additional resource as it has a very comprehensive coverage of known bound morphemes in modern Japanese.

4. Investigation Targets

Three types of synthesized term formation were investigated:

a) **Abbreviation/Clipping**. This is a very common and productive process in Japanese, wherein the (usually) leading character of each of the components of a composite are taken to form an abbreviated compound (Tsujimura, 2006, p. 153). An example of this is 学割 *gakuwari* "student discount" from the full compound 学生割引 *gakuseiwaribiki*. The terms produced by this process are typically nouns or adjectival nouns, and hence if valid would be clear candidates for lexicalization.

a) **Affixation**. The addition of prefixes and suffixes, often written with a single *kanji*, is a very common morphological process in Japanese (Tsujimura, 2006). Vance (1991) describes 63 single-*kanji* affixes commonly employed. The process is very productive and the resulting terms are not usually lexicalized unless they have an idiomatic meaning or unusual reading. Most of the terms arising from this process are nouns, e.g. those arising from the affixation of 化 (*ka*: -ization). Some others, such as 的 (*teki:* -like, -ical, -ish, etc.) form adjectives.

b) **Compounding**. As in many languages, the formation of terms by combining two or more words or morphemes is very common. The components can be independent words, as in 秋空 *akizora* "autumn sky" where both 秋 *aki* and 空 *sora* can be used independently, or bound morphemes as in 警告 *keikoku* "warning" where neither component can be used alone. (See Tanaka (2002) and Baldwin and Tanaka (2004) for earlier work in this area.) In this study two types of synthesized compounds were investigated:
   i. 2-*kanji* compounds, as in the examples above;
   ii. composites formed by aggregating known 2-*kanji* compounds, for example combining 警告 (above) with 射撃 *shageki* "firing, shooting" forms a composite 警告射撃 *keikoku shageki* meaning "warning shot". This process is very common and the resulting terms often have a clear sum-of-the-parts meaning, which results in them often not being lexicalized. (警告射撃 is in several major bilingual dictionaries, but is not in most Japanese dictionaries.)

5. Analysis Approach

At the heart of this task is the development of techniques which will detect if a sequence of characters is being used as a term in written Japanese. Synthesized terms which were not already recorded in a reference lexicon were considered for evaluation if they occurred more than 100 times in the corpus (that number being a reasonable level of "commonness" to warrant recording in a dictionary). The approach which is being evaluated is based on the assumption that if a compound is actually being used, it will occur in textual syntactic contexts which are typical for such terms. Japanese uses a large number of particles which are usually written in the *hiragana* syllabary, and it is the encapsulation by these particles that we use as the basis of the analysis. For example, if a term is being used as a noun, it could be expected that it is observed in text followed by the が *ga* subject-marking particle, or the は *wa* topic-marking particle, and will often occur in encapsulations such as <が,を> where を *wo* marks the object of a clause or sentence.

For testing purposes combinations of the following common pre/postpended particles were selected:

pre: は (*wa* – topic), が (*ga* - subject), に (*ni* – location, indirect object, etc.), の (*no* – possessive, modifier, etc.), な (*na* - adjectival), て (*te* - continuative), や (*ya* - conjunction)
post: を (*wo* - object), が (*ga*), に (*ni*), の (*no*), な (*na*), や (*ya*)

The initial tests were carried out using all combinations apart from repeated particles (e.g. が ... が), giving 37 combinations. The n-gram counts for 10 common encapsulations are given in Table 1 for three compounds: the common noun 男子 (*danshi*) - youth, young man, which occurs 5 million times in the n-grams, the less common noun: 人魚 (*ningyo*) - mermaid, merman (607,772 times), and the rather less common 間者 (*manja*) - spy (24,543 times).

Table 1. Examples of n-gram counts for particle encapsulations

Encapsulation	男子	人魚	間者
はXを	2631	616	100
はXが	18242	1002	47
はXに	9117	1150	26
はXの	24367	4290	54
はXな	1135	247	0
はXや	164	103	0

がXを	5676	285	33
がXに	4340	1015	35
がXの	6363	1703	35
がXな	423	75	0

As can be seen from the counts associated with these known nouns, there are clear patterns associated with the encapsulations which have the potential to drive the analysis of synthesized terms to detect actual nouns among them.

Two types of evaluation were carried out on the sets of counts:

a) a machine-learning approach in which models were trained on the features associated with known terms and non-terms;
b) a heuristic approach using rules based on the numbers of encapsulations.

To evaluate whether a machine-learning approach to classifying synthesized compounds can be used in this case we tested the application of a support vector machine (SVM) to the problem. SVMs have been demonstrated to be a very effective approach to this form of classification problem (Joachims 1999; Steinwart and Christmann 2008). The SVM package used is LIBSVM (Chang and Lin 2011).

To train the SVM a set of 400 terms which met the 100-occurrence threshold was selected and the n-gram counts of their encapsulations extracted. On examination it emerged that terms with fewer than 5 non-zero encapsulations were generally not valid nouns, so this was used as the basis for dividing the training data into valid/non-valid sections. During this initial evaluation it was also determined that there were some useful refinements possible with the training model:

a) seven of the 37 initial encapsulation patterns, e.g. <て,な> and <や,な>, did not contribute significantly to the evaluation, and hence could be removed from the basic model
b) the addition of the counts with six postpended particles (は, を, が, に, の, な) improved the performance (this formed an extended model.)

The models were tested against a set of 840 synthesized terms which had been hand-checked for their validity. The basic model classified 71 of these, of which 68 were in the reference lexicon and the other 3 were valid, albeit rare terms. The extended model classified 83, of which 80 were in the reference lexicon and the other 3 were also valid but rare terms. This indicated a very high level of precision with the technique. The establishment of the level of recall of the technique, i.e. the proportion of valid terms classified, is a more of a problem as a valid but rarely-occurring term is unlikely to be classified by the models. Further testing with randomly-selected sets of valid terms indicated an overall recall of around 70%.

6. Analysis of Synthesized Abbreviations

As mentioned earlier, abbreviations in Japanese are typically formed by taking the leading character of the morpheme components of the full term, as in the case of 学割/学生割引. There are known exceptions, for example the common term for "high school" is 高校 *kōkō,* which is formed from the first character of 高等 *kōtō* (high grade) and the second character of 学校 *gakkō* (school). A brief test of the validity of this (common) assumption was carried out by examining the sources of a sample of 75 known abbreviations. 78% followed the leading morpheme character pattern, with the remainder either coming from non-initial characters (as with 高校) or from multi-morpheme source terms where some morphemes did not contribute to the abbreviation. It was considered that the 78% validated the approach sufficiently to warrant continuing the approach.

Using the leading-character method, 33,000 possible abbreviations were generated from known multi-morpheme terms. Of these, 7,900 were not already in the reference lexicon and met the threshold of 100 occurrences in the n-gram corpus. The SVM classifier identified 162 for further consideration. Inspection of these determined:

- 15 were abbreviations of the source term;

- 33 were valid terms, but not identifiable as abbreviations of the source term, For example 一話 had been generated from 一夕話 *issekiwa* (short story) but is actually a different term meaning "episode (of a series, serial, etc.)";
- the remainder were mostly on-the-fly compound collocations such as 五年 *go'nen* "five years", 三万 *sanman* "thirty thousand", 主論 *shuron* "principal theory", etc. Constructions using productive affixes such as 的 *teki* "-like" and 化 *ka* "-ification" were quite common too.

It must be concluded that to find 15 abbreviations after processing 33,000 source compounds certainly indicates it is not a highly productive technique. With the level of precision resulting from this test being only 9%, one must conclude there are probably better ways of searching for new unrecorded terms.

7. Analysis of Affixation

As mentioned above, affixation, often through the use of single-*kanji* prefixes and suffixes, is a very common and productive morphological process in Japanese. For example the noun 衒学 *gengaku* "pedantry" can take suffixes such as the personalizing suffix 者 *sha* to form 衒学者 *gengakusha* "pedant" or the adjective-forming suffix 的 to form 衒学的 *gengakuteki* "pedantic".

There is the potential to synthesize extended compounds using these affixes and test for their presence in Japanese texts. However initial testing demonstrated that the productiveness of this morphological process is so high, and in general the meanings of the extended compounds are so predictable, that as expected the resulting terms are usually lightly lexicalized. The reasons for considering the identification and possible lexicalization of such terms include:

a) idiomatic meanings which are not readily apparent from the components;
b) covering targets for English-Japanese dictionary entries;
c) special meanings in certain contexts, e.g. 拮抗 *kikkō "*rivalry, antagonism" where 拮抗的 *kikkōteki* is used in medical contexts for "antagonistic" and can be found in formations such as 拮抗薬 *kikkōyaku* "antagonist drug".

The challenge is to determine whether proceeding further with this particular investigation could bring up anything useful. Certainly a straight extraction process is virtually guaranteed to identify large numbers of unlexicalized formations, however identifying those worthy of lexicalization, e.g. those with idiomatic or special meanings, is largely outside the scope of this study.

8. Analysis of Compound Generation

The development of the techniques described earlier allows for the rapid testing of large numbers of synthesized compounds against the n-gram Corpus, and are achieving high levels of precision and generally satisfactory levels of recall. While there has been some success with targeted compound synthesis in the form of potential abbreviation formation, there is also the possibility of simply generating *kanji* combinations using commonly occurring *kanji*, and using the classification system as a filter to detect combinations which are actually being used as terms.

While the full number of *kanji* in use in written Japanese is very large, only about 2,500 are in common use, leading to over 6 million possible 2-*kanji* compounds. While this is a relatively large number, it is a quite manageable processing task to:

- create synthetic compounds simply by combining *kanji*;
- filter out known compounds;
- classify the remaining compounds using the techniques described above.

Initial tests were carried out using 10,000 compounds generated from combinations of the 100 most common *kanji*. Of these approximately 2,700 were in the in the reference lexicon, and of the remainder 97 were classified using the SVM model described earlier. Inspection of the 97 classified compounds revealed that many involved numerics (二円 *ni'nen* "two yen, circles, etc.", 十五 *jūgo* "fifteen", etc.) and several involved very common compositional suffixes such as 氏 *shi* "surname", e.g. 東氏 *higashishi* "Mr Higashi". This resulted in a form of noise that tended to obscure some valid compounds in the set, e.g. 化調 *kachō* which is

an unrecorded abbreviation of 化学調味料 *kagakuchōmiryō* "chemical seasoning". Sufficient valid results were detected that it was decided to investigate more thoroughly using less common kanji and excluding the major single-*kanji* affixes, which it was hoped would not produce as many obvious compositional compounds.

8.1. 2-*kanji* Compound Generation and Testing

In carrying out large-scale testing of synthesized 2-*kanji* compounds, the SVM classification system described earlier was supplemented in several ways:

a) three additional sets of training data were created of 10,000, 40,000 and 90,000 compounds respectively (small, medium and large models). Compounds which did not appear in the n-gram corpora or did not have a minimum number of encapsulation features were removed, and the remainder were checked against the reference lexicon to establish a valid/not-valid classification.
b) both the Google and Kyoto models were used in parallel to test whether the 20 n-gram cut-off in the Google corpus had a significant impact.
c) a heuristic approach based on the number of encapsulation features was also tested.

Two testing sets each of 40,000 compounds were synthesized, and then filtered against the criteria of having n-gram counts and meeting a minimum number of counted encapsulations. For example, in the first set 202 compounds were counted in the Google corpus and had counts for 5 or more features. Of these 202, 135 were in the reference lexicon. As expected somewhat larger numbers were counted in the Kyoto corpus as the cut-off is lower.

The filtered compounds were then classified by the various SVM models. For example of the 202 compounds mentioned above 89 of the 135 in-lexicon compounds were classified along with 20 of the 67 not-in-lexicon compounds. The overall assessment of these classification tests was:

a) the overall precision, using presence in the lexicon as the criterion, was around 75% across the models;
b) the recall level, i.e. the proportion of in-lexicon terms identified, varied across the models and was highest with the Google/5+ features model at around 65%.
c) virtually identical results were achieved by using the feature count heuristic.

It is inappropriate, however, to focus too much on the effectiveness of the models in recovering known terms as it loses sight of the actual goal, which is to identify terms which are not in references. To evaluate that properly, we needed to look in detail at the compounds which have been classified but did not occur in the reference lexicon. A total of 38 unlexicalized compounds from the first test set which had either been classified by the Google model or had 7 or more features were examined in detail, using reference material and the WWW pages in which they were used. The compounds turned out to consist of 11 nouns, 23 named-entities and 4 non-terms (apparently resulting from mis-parses in the n-gram generation). The *kanji* ranges used in the generation happened to include *kanji* such as 霊 (spirit), 魔 (demon), 姫 (princess), 鬼 (ogre), 龍 (dragon), 雷 (lightning), etc. thus virtually guaranteeing the generation of a number of manga, anime, game, etc. character names. Since the investigation is not attempting to make value judgements about the compounds, but needs to assess them as to whether they are valid terms in use which may be incorporated in some lexicon (even one of manga characters), it has to be concluded that 34 of the 38 are valid, and hence the overall precision achieved was 89.5%. The second set of 40,000 compounds was also evaluated, with a similar outcome. It was observed that valid terms identified by the process tended to have larger numbers of encapsulation features, indicating that a heuristic approach based on this is probably more effective that using machine learning.

While the overall yield of valid unlexicalized compounds detected during the testing is, predictably, not particularly high, the level of precision, approaching 90%, indicates that this technique has considerable potential for being used on a large scale to identify potential neologisms which can be passed onto detailed analysis.

Some examples of valid unlexicalized 2-*kanji* compounds detected during the investigation are in Table 2.

Table 2. Examples of synthesized 2-*kanji* compounds

Compound	**Meaning**
桃豆 *momomame*	bean-based sweet rolled in peach powder
低弦 *teigen*	low-pitched string instruments (abbr)
周堤 *shūtei*	circular bank, e.g. round *Jōmon*-period grave sites
整膚 *seifu*	alternative therapy involving pinching and pulling the skin
移弦 *igen*	string-crossing (violin, etc. technique)
紙筒 *kamidzutsu*	paper tube (var. of 紙管 *shikan*)
製縫 *seihō*	suffix to business names meaning clothes-making
賞暦 *shōreki*	awards, list of years prizes were won
丸刀 *gantō*/*marutō*	gouge with U-shaped blade
士紳 *shishin*	archaic term for ranking official
帰寮 *kiryō*	returning to one's accommodation, etc.
春苗 *shunbyō*	spring seedlings (also a girl's name)
母珠 *moshu*	large bead(s) in a Buddhist rosary
白粒 *shirochibu*	type of ceramic used in Kitaniware, etc.
紙函 *kamibako*	paper box
親鴨 *oyagamo*	can mean "parent duck", but mostly used in 親鴨会, a society of former IBM staff

8.2. 4-*kanji* Compound Generation and Testing

The potential numbers of 4-*kanji* compounds which can be generated from known 2-*kanji* compounds is very large, for example if we take the approximately 50,000 2-*kanji* compounds in the JMdict data as a basis for generation we will get several billion candidates. The total number of recorded 4-*kanji* compounds is, however, in the 100-200,000 range so it can be expected that the proportion of generated compounds which turn out to be valid will be very small.

For the testing of synthesized 4-*kanji* compounds a similar approach to the 2-*kanji* compounds has been followed. In this case the generation process was combined with the initial filtering, and compounds were only passed to the classification stage if they reached a threshold count in the Google n-grams (initially 1,000, later reduced to 100), and recorded a minimum number of 5 or more encapsulation features. In an initial test of 1,000,000 compounds, 310 met the 1,000 count threshold and of these 181 had had the required number of features. 80 of these were classified by the SVM model, of which 16 were in the reference lexicon. A selection of the classified compounds is in Table 3. Reducing the count threshold to 100 increased the candidate compounds by 29, however no more were classified and only two more of the unclassified compounds were in the lexicon (also in the table.)

T**able 3.** Examples of classified 4-*kanji* compounds

Compound	**Count**	**Features**	**Lexicon**	**Meaning**
一軒一軒 *ikken'ikken*	48601	22	Y	house-to-house
一言一句 *ichigon'ikku*	29723	15	Y	word-by-word
一向一揆 *ikkōikki*	30691	27	Y	*Jodo Shinshu* Buddhist uprising
一個一個 *ikkoikko*	72951	24	N	one-by-one
一曲一曲 *ikkyokuikkyoku*	49027	23	N	piece-by-piece (in music)

一言一行 *ichigen'ikkō*	533	9	Y	every word and act
一齣一齣 *hitokomahitokoma*	258	8	Y	frame by frame

The fact that the 2-*kanji* compounds used for generation contained a set beginning with the *kanji* 一 i*chi, hitotsu* "one" was somewhat unfortunate; all but 7 of the classified compounds began with 一. It appears there are many non-idiomatic expressions of the 一 X 一 X and 一 X 一 Y variety with meanings such as "X occurring one after another" or "one X per Y" which are not necessarily lexicalized. As this distorted the results estimation of the precision in this sample was not attempted

A further two sets of 1,000,000 compounds were tested of which 115 met the criteria for classification. The numbers classified and whether they were in the lexicon is in Table 4.

Table 4. Evaluation outcome for synthesized 4-*kanji* compounds

	In Lexicon	Not In Lexicon
Classified	9	17
Not Classified	8	81

Examination of the classified out-of-lexicon compounds determined that several are real terms:

英国王室 *eikokuōshitsu* - "British royal family"
液晶演出 *eishōenshutsu* - name of a video/*pachinko* game
欧州遠征 *ōshūensei* - "European campaign" (esp. with sporting teams)
横浜駅前 *yokohamaekimae* - a typical address, in this case at the front of Yokohama Station
下限価格 *kagenkakaku* - "price floor"
化粧下地 *keshōshitaji* - "foundation base"

Most of the in-lexicon but unclassified compounds were rather obscure, e.g. 寡占価格 *kasenkakaku* "oligopolistic price" has a low count and barely made the list to be tested.

One observation from this investigation of 4-*kanji* compounds is that identifying candidates on the basis of n-gram counts and feature numbers almost always out-performed the SVM machine-learning approach. Moreover, it usually identified a compound which was actually being used in text. This is not that surprising when one considers the process of forming these sorts of extended compounds in Japanese - one has considerable freedom and flexibility. It is not unreasonable to conclude that almost of the 4-*kanji* compounds passing the (heuristic-driven) filters are actual compounds being used. Quite possibly this part of the study needs to be viewed from the position of discovering 4-*kanji* compounds that are in common enough use to be considered for lexicalization. That consideration needs to take into account issues such as:

a) whether the compound has an idiomatic meaning (e.g. 援助交際 *enjokōsai*, where the component compounds mean "assistance" and "company", but the full compound is a common euphemism for teenage prostitution.)
b) whether it is a useful inclusion for reasons such as for English-to-Japanese reverse searches.

A point to note is that many of the accepted terms have meanings which are readily apparent from the components, and hence are unlikely to be included in a general dictionary unless the usual criteria were relaxed. It was noted that compounds formed from components which were polysemous were more likely to have non-intuitive meanings, which indicates a possibly fruitful area of future study.

9. Conclusion

In this paper we describe the investigation of a neologism detection approach involving the synthesis of possible Japanese terms by mimicking Japanese morphological processes, followed by testing for the presence of candidate terms in Japanese corpora in appropriate syntactic contexts. Synthesized terms which passed this evaluation were assessed by regular lexicographic processes to determine their suitability for lexicalization.

Of the techniques tested: abbreviation, affixation and compounding, the latter showed particular promise, with the 2-*kanji* compound generation and classification resulting in significant numbers of unrecorded terms deemed suitable for inclusion in dictionaries.

The investigation used both a SVM-based machine-learning technique and heuristics to identify neologisms, and concluded that the simpler heuristic approach performed marginally better. It would be interesting as future work to test the use of a neural network approach, as these have proved very successful in language-processing applications. Given that the techniques used in this investigation achieved relatively high levels of precision, any improvement would probably be small.

References

B.T. Sue Atkins and Michael Rundell. 2008. T*he Oxford Guide to Practical Lexicography.* Oxford University Press, Oxford, UK.

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of compound nominals: Getting it right. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing.*

James Breen. 2004. JMdict: a Japanese-Multilingual dictionary. In *Proceedings of COLING-2004 Workshop on Multilingual Resources.*(ANON)

James Breen and Timothy Baldwin. 2009. Corpus-based Extraction of Japanese Compound Verbs. In *Proceedings of the Australasian Language Technology Workshop (ALTW 2009).*(ANON)

James Breen. 2017. Extraction of Neologisms from Japanese Corpora *(PhD Thesis, The University of Melbourne)* Chapter: "Neologisms: Lexicographic Issues and Terminology" . https://minerva-access.unimelb.edu.au/handle/11343/211675(ANON)

James Breen, Timothy Baldwin, and Francis Bond. 2018. The Company They Keep: Extracting Japanese Neologisms Using Language Patterns. In *Proceedings of the Global Wordnet Conference*.(ANON)

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

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.

Thorsten Joachims, 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods — Support Vector Learning*.

Frenk Keller, and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. Computational Linguistics 29.459–484.

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/.

Kikuo Maekawa, 2008. Balanced Corpus of Contemporary Written Japanese. In Proceedings of the 6th Workshop on Asian Language Resources, Hyderabad, India.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In Proceedings of LREC 2010.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. Language Resources and Evaluation 48.345–371.

Ingo Steinwart, and Andreas Christmann. 2008. Support Vector Machines. Springer Publishing Company, Incorporated, 1st edition.

Takaaki Tanaka. 2002. Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *Proceedings of Coling 2002*.

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

Timothy J Vance. 1991. *Instant Vocabulary through Prefixes and Suffixes*. Kodansha International, Tokyo.