

Expanding the Lexicon: the Search for Abbreviations

James BREEN

Monash University

Clayton 3800, Australia

jwb@csse.monash.edu.au

Abstract

This paper describes a small experimental project to determine whether it is possible to discover hitherto unrecorded abbreviations in Japanese by mimicking the natural-language abbreviation process using a large set of source words, then using the WWW as a corpus of Japanese texts to determine whether the synthesized abbreviations exist. The process of formation of the abbreviations is described, along with the WWW search and the resulting analysis of text material. While only a small number of cases have been investigated completely, and the project is only at a proof-of-concept stage, a number of "new" abbreviations have been detected.

1. Introduction

Any person learning Japanese as a foreign language quickly discovers that a large number of abbreviations are in regular use. Many loan-words are truncated, and even the governing party in Japan, the Liberal-Democratic Party (自民党: *jimintou*) has a title that is abbreviated from the rarely-used 自由民主党.

A problem for the learner is that these abbreviations are poorly lexicalized. Some only appear in newspaper headlines, and thus are considered unimportant by lexicographers. They rarely appear in learners' dictionaries, where space is at a premium, and are often overlooked in the major Japanese-English dictionaries, as these are produced primarily for the domestic Japanese market and few native Japanese speakers would wish to look up such an abbreviation in English.

In this paper, a project to determine whether it would be possible to expand the representation of such abbreviations using natural-language processing techniques is described. The project involves the synthesis of possible abbreviations, then the search of a text corpus to determine whether the synthesized abbreviation is being used.

2. Abbreviation Formation in Japanese

One of the major processes in Japanese word formation is *compounding*, i.e. the combination of two or more words to form a new word. This process, which is common to many languages, may involve independent words, or may involve morphemes which are not normally independent words. Compounds in Japanese may involve native Japanese components, e.g. 近道 (*chikamichi*: *shortcut*), Sino-Japanese components, e.g. 殺人 (*satsujin*: *murder, manslaughter*), or hybrids of native, Sino-Japanese and loan-word components, e.g. 台所 (*daidokoro*: *kitchen*) and 石油ショック (*sekiyuushokku*: *oil shock*). (Tsujimura, 1996)

Compounding extends to combining two existing compounds to form a new compound which is typically a noun, noun-phrase or multi-word expression consisting of four or more kanji, e.g. 為替 (*kawase*: *exchange, money order*) and 相場 (*souba*: *market price*) combine to form 為替相場 (exchange rate). There are many such extended compounds in use in Japanese, and a large number of them appear in dictionaries as independent entries. As confirmed by inspection of various lexicons, the majority of such extended compounds are made up of four kanji, typically formed from two two-kanji compounds, although there is also a large number of longer compounds.

Another word formation process is that of *abbreviation* (sometimes also called *clipping*), in which words are shortened. In many cases, particularly for long loan-words, the word is simply truncated, e.g. プラットホーム (*purattohoomu*: *platform*) becomes just ホーム, and スーパーマーケット (*suupaamaaketto*: *supermarket*) becomes just スーパー. For long

compounds, the process typically involves selecting the first two morae of each constituent compound (Tsjumura, 1996). For example, 学生割引 (*gakuseiwaribi: student discount*), is usually abbreviated to 学割, with the latter occurring four times more often in Japanese WWW documents. In some ways this process is analogous to the creation and use of acronyms in languages which use alphabets.

While this form of abbreviation is by no means unique to Japanese (e.g. in the 1960s the Ministry of Technology in the UK was often referred to as "MinTech"), it appears that the process is more strongly embedded and more commonly employed than in many other languages.

Inspection of the lexicalization of this two-kanji class of abbreviations reveals:

a. many such abbreviations do not appear in dictionaries. For example, the common 工博 abbreviation of 工学博士 (Doctor of Engineering) does not appear in any major dictionaries. While this may be due to many such words being neologisms, and they may appear in dictionaries at a later stage if they persist, their absence, according to advice received from a lexicographer at Kenkyusha, is due to them being recognizable by Japanese native speakers as being abbreviations.

b. many abbreviations are handled differently in domestic Japanese dictionaries and Japanese-foreign-language dictionaries. For example, both the Koujien and Daijirin Japanese dictionaries have entries for 学生割引, and for 学割 only note that it is an abbreviation of 学生割引, whereas major Japanese-English dictionaries such as Kenkyusha's New Japanese-English Dictionary and Sanseido's Grand Concise Japanese-English Dictionary both have entries for 学割 and only mention 学生割引 in the compound list within the 学生 (*gakusei: student*) entry. This pattern can be observed for most abbreviations that appear in these dictionaries.

3. Abbreviations - the Search Process

Given that a large number of four-kanji compounds have already been lexicalized, e.g. the JMdict file (Breen, 2004) has over 8,000 four-kanji compounds recorded, it is possible to use that set of compounds as the basis for an automated search of a Japanese corpus to determine if hitherto unrecorded abbreviations are in use.

The process employed in this project is:

- a. construct a set of possible two-kanji abbreviations from the four-kanji compounds in the JMdict file;
- b. remove from that set the character pairs which:
 - i. were already recorded in a dictionary. Often these were an already-recorded abbreviation, but in some cases were an independently-formed two-kanji compound;
 - ii. are used as a person or place name. While this does not preclude the additional use of the pair as an abbreviation, it is likely to result in a number of false matches with the corpus;
 - iii. contain a kanji numeric character, or contain a kanji commonly used as a single character prefix or suffix (e.g. 非, 的, etc.) as these do not typically appear in this form of abbreviation.

Approximately 30% of candidates were removed at this stage, of which the majority were identifiable as person or place names.

- c. test the remaining possible abbreviations against Japanese pages in the WWW, both for frequency of occurrence and for syntactic contexts where they may operate as an independent word.

The Japanese pages in the WWW were used as a corpus in this investigation for several reasons:

a. it is a very large collection of text, with over 300 million pages indexed by common search engines. Japanese is the second most common language used in WWW pages, after English (Breen and Tokita, 2004);

b. it is freely available and is amenable to effective searches using search engines. Comparable large corpora, such as newspaper archives, were not available without a prior commercial arrangement;

c. prior studies have indicated a high level of correlation between the WWW and large corpora in such areas as word frequencies (Keller and Lapata, 2003).

The examination of the WWW was made using the Google search engine via the API (Application Program Interface), which provides for programmed searches. The API interface enables a number of filters to be set, and in this case the text language was limited to Japanese, i.e. only pages which have been classified by Google to be in Japanese. The language classification appears to be quite conservative, however it was considered important to exclude pages containing Chinese or Korean as the Google database holds pages in Unicode coding and thus false matches are possible for kanji search keys.

A further restriction was to ensure that the pair of characters were adjacent in the text. For poorly lexicalized terms the Google indices appear to handle kanji as separate tokens, and as a default the search may return a match on non-adjacent kanji. By specifying a key in quotation marks it is possible to restrict the match to a sequence of kanji, however the match will still occur if the kanji are separated by space or punctuation characters, necessitating a finer analysis of the search results.

The examination of possible abbreviations proceeded in two stages:

a. an initial analysis was made to determine the "hits", i.e. the number of WWW pages which contained each possible abbreviation. These were then sorted in descending order of the number of hits;

b. a second detailed analysis was made in which "snippets" of text were retrieved from Google for detailed examination.

The ordering of the candidates according to frequency of hits was done in order to concentrate on the more commonly occurring sequences which would, if valid, be worth including in a lexicon. If the process was successful for these candidates, it could be repeated for less-frequently occurring candidates. As Google ranks pages according to a weighting system based on, among other things, the number of hyperlinks pointing to a page, it is reasonable to expect that valid uses of an abbreviation would be discernible in the higher-ranked pages. Approximately 23% of the remaining candidates received over 1,000 hits in the Google search.

The text snippets supplied by the Google API typically contain about 70 characters of Japanese text surrounding the target word. The text in the snippets was stripped of residual HTML tags, then examined to determine:

a. if the target word was present. As explained above, on occasions Google's indexing system will indicate a match even if the kanji in the key are separated by other characters. Also, Google will return a page for which the key is in the text associated with the URL of a link to that page. This accounts for a large proportion of cases where the text is not actually present. The text of each snippet was scanned and the snippet rejected if the candidate compound was not present as an adjacent pair of characters. In some cases this leads to false rejections, as the snippets returned from converted document formats such as Microsoft Word and Adobe PDF occasionally have spaces between characters, however it is difficult to detect and remedy these cases automatically;

b. if the target word was found to be present, the text was examined to determine if the characters were likely to form an independent word. A three-level classification was used:

- i. if the candidate was preceded by and followed by either kana or punctuation, parentheses, etc. it was classified as a strong indication;
- ii. if it was adjacent to more kanji, but preceded by or followed by either punctuation or kana, it was classified as a moderate indication;
- iii. if it was both preceded and followed by more kanji, it was classified as a weak indication. This is because there is the possibility that the two kanji detected are from two distinct adjacent compounds.

In order to assess this classification, the analysis was applied to a small set of recognized abbreviations: 拡販, 学割, 郵貯 and 労組, and to a set of common Japanese compounds: 先生, 学校, 政府 and 工場. Table 1 shows the results from the 110 highest-ranking pages for those words. The column marked "Confidence" is the ratio of pages classified as either Strong or Moderate to the total number of pages containing the candidate, and may be seen as a crude measure of the precision of the technique.

Word	Strong	Moderate	Weak	Not Present	Confidence
拡販	62	20	9	19	0.90
学割	56	28	2	24	0.98
郵貯	84	5	0	21	1.00
労組	32	35	13	30	0.84
先生	36	21	0	53	1.00
学校	31	25	4	50	0.93
政府	4	36	12	58	0.77
工場	24	41	2	43	0.97

Table 1: WWW Search and Analysis Results: Common Abbreviations and Words.

From this it is reasonable to conclude that a strong representation in the Strong and Moderate classifications may be grounds for concluding that word exists and is in use.

To test this assumption, the analysis was carried out on a selection of possible abbreviations. Table 2 shows the results from candidates which had resulted in Google reporting several thousand matched pages, and Table 3 shows the results for candidates for which about two hundred matches were reported. The compound from which the abbreviation candidate was formed is also shown.)

Word	Strong	Moderate	Weak	Not Present	Confidence
工技 (工業技術)	6	15	73	16	0.22
再利 (再生利用)	0	73	27	10	0.73
国補 (国家補償)	7	61	24	18	0.74
国計 (国土計画)	31	34	18	27	0.78
国展 (国際展開)	39	35	22	14	0.77
工化 (工業化学)	51	17	3	39	0.96
最賃 (最低賃金)	35	50	7	18	0.92
国都 (国際都市)	28	37	32	13	0.67
国関 (国際関係)	12	11	7	80	0.77
県病 (県立病院)	10	51	34	15	0.64
高建 (高層建築)	36	42	6	26	0.93
財相 (財産相続)	1	93	5	11	0.95

印電 (印刷電信)	1	21	62	26	0.26
合皮 (合成皮革)	77	21	2	10	0.98

Table 2: WWW Search and Analysis Results: High-rank Candidate Abbreviations.

Word	Strong	Moderate	Weak	Not Present	Confidence
作指 (作況指数)	3	9	59	39	0.17
最限 (最小限度)	93	7	0	10	1.00
再制 (再販制度)	3	94	3	11	0.97
債保 (債務保証)	0	2	68	40	0.03

Table 3: WWW Search and Analysis Results: Low-rank Candidate Abbreviations.

These results do not lend themselves to straightforward interpretation.

Many of the candidates in Table 2 with reasonably high confidence measures turn out to be valid words, but not all are abbreviations. For example, 国補, 合皮, 最賃, 高建, 国関 and 県病 are abbreviations of the original four-kanji compound, however 国展 and 国都 are words formed independently and 国計 and 工化 are abbreviations of other words.

While it is tempting to dismiss candidates such as 工技, which has a low confidence measure, inspection of the WWW pages that contain it reveals that it used as an abbreviation of 工業技術 in such things as the titles of prefectural industrial research centres, e.g. the 工技ネット新潟 in Niigata. Similarly 印電 resulted in one page where it clearly was used as an abbreviation of 印刷電信, but in all others the matches resulted from the juxtaposition of 印 (seal) and 電話番号 (telephone number) on forms.

In the cases of 財相, 最限 and 再制 in Table 3 the confidence measure was very high, although the number of hits was low. The result was skewed either to the strong or moderate classifications. On inspection the reasons for this become apparent:

- the matches for 再利 were all occurrences of 再利用 (reuse; recycling). This candidate should perhaps have been filtered on the grounds that it is a partial match on an existing word;
- 財相 matched on occurrences of 経財相, which is a contraction of 経済財政担当相 (Minister for Economics and Finance);

c. 最限, which recorded a very high confidence level, is something of a mystery. On investigation, it was apparent that the total number of unique occurrences was quite low, as many of the pages containing hits were plainly copied from each other. The author discussed the word on a translators' mailing list, which resulted in the following suggestions:

- i. it is an input-method entry error for 際限 (which has the same pronunciation: *saigen*);
 - ii. it is an abbreviation of the phrase 最も限度に近い;
 - iii. it is indeed an abbreviation of either or both of 最小限 and 最大限.
- d. The consensus was that it means "limit", as does 際限.

4. Discussion

Although only a relatively small number of cases has been examined in depth, it appears that provided the confidence measure is above about 0.60 and there is a reasonable representation of strongly classified hits, there is a good chance that a "new" word has been identified. At present this has only been tested for candidates with total hits in the thousands.

The situation with candidates with relatively low numbers of hits is far less clear. If the actual of target pages is small, there is a risk of the results being influenced by input errors, spelling mistakes, etc. Also the impact of having pages with related material, as was the case with 最限, becomes greater.

It is appropriate to question at this stage whether the line of investigation taken in this project is worthwhile. Of the original 8,000 candidates derived from four-kanji compounds fewer than 1,500 meet the criteria of not already being in a lexicon and achieve a suitably large number of page hits. Of these, it is unlikely that more than 50% will result in a "new" word being lexicalized. As the validation usually requires reading several WWW pages to determine the meaning and context of the word, the overall process can be quite time-consuming.

If the purpose of the process is simply to expand the lexicon, there are probably easier and less time-consuming ways to do this, such as calling for donations of material from native speakers. However as a method of detecting unrecorded abbreviations, it appears that the technique is worth applying to completion.

5. Conclusion

This project has demonstrated that it is possible to identify numbers of Japanese abbreviations by synthesizing candidate abbreviations from longer compound words, then testing for their presence in WWW pages. A semi-automated process was developed which identified which candidates had a high likelihood of being either a valid abbreviation or a hitherto unrecorded neologism.

References

Breen, J.W. *JMdict: a Japanese-Multilingual Dictionary*, COLING-2004 Multilingual Linguistic Resources Workshop, Geneva, August 2004 Also:
<http://www.csse.monash.edu.au/~jwb/jmdictart.html>

Breen, J.W. and Tokita, A, *The WWW in Japan: a threat to cultural identity, or a domesticated system?*, First International Conference on Cultures and Technologies in Asia, Mumbai, India, Feb 2004

Keller, F and Lapata, M., *Using the Web to Obtain Frequencies for Unseen Bigrams*, Computational Linguistics, Vol. 29, No. 3, September 2003, MIT Press.

Tsujimura, N. *An Introduction to Japanese Linguistics*, Blackwell, 1996.