Vocabulary Learning and Instruction
Volume 8, Issue 2, October 2019
http://vli-journal.org

A Procedure for Determining Japanese Loanword Status for English Words

David Allen
Ochanomizu University
allen.david@ocha.ac.jp
https://doi.org/10.7820/vli.v08.2.Allen

Abstract

Japanese loanwords are mainly derived from English. These loanwords provide a considerable first-language (L1) resource that may assist in second-language (L2) vocabulary learning and instruction. However, given the huge number of loanwords, it is often difficult to determine whether an English word has a loanword equivalent and whether the loanword is likely to be widely known among the Japanese. This article demonstrates an efficient method of answering these two questions. The method employs corpus frequency data from the *Balanced Corpus of Contemporary Written Japanese,* from which the existence and frequency of loanwords in Japanese can be determined. Following the guidelines presented herein, researchers will be able to use data from the corpus themselves to check cognate frequency, thereby determining the cognate status of items used in research.

Keywords: Japanese loanwords; Japanese-English cognates; loanword frequency; cognate frequency; different-script cognates

1. Background

There are thousands of Japanese loanwords and most are derived from English (e.g., テーブル /teeburu/ "table"). In fact, a recent study showed that of the most common 10,000 headwords in English according to the British National Corpus – Corpus of Contemporary American English corpus wordlists (Nation, 2012), half have a loanword equivalent in Japanese, and a quarter have one that is relatively high-frequency (Allen, 2018b). This loanword resource has been referred to as a "built-in lexicon" that may support English learning in Japan (Daulton, 2008). However, while there has been growing interest in the role of loanwords in language learning, teaching, and assessment in the Japanese context, many important questions remain unanswered (Allen, 2019). Therefore, with the aim of supporting and promoting research into the impact of loanwords, this paper describes an efficient method of identifying which English words have Japanese loanword equivalents, and which of these Japanese loanwords are likely to be widely known to Japanese speakers.

Japanese loanwords from English are often referred to as *Japanese–English cognates* because they are perceived to share form (phonology) and meaning (semantics) across the two languages. As with cognates in other languages, such

Please cite this article as: Allen, D. (2019). A Procedure for Determining Japanese Loanword Status for English Words. *Vocabulary Learning and Instruction, 8* (2), 1–7. https://doi.org/10.7820/vli.v08.2.Allen

as Dutch and English, Japanese–English cognates are processed more quickly and accurately than noncognates when reading words in the L2 and when producing L2 words from picture stimuli (Allen & Conklin, 2013; Hoshino & Kroll, 2008; Miwa, Dijkstra, Bolger, & Baayen, 2014). This beneficial effect is referred to as the "cognate effect." Furthermore, classroom-based research indicates that Japanese–English cognates may support learning (e.g., Daulton, 2008). In language assessment, too, Japanese–English cognates have been shown to be more accurately recognized and comprehended than noncognates (e.g., Allen, 2018a; Rogers, Webb, & Nakata, 2014; Stubbe & Yokomitsu, 2012).

Japanese–English cognates can also, however, create difficulties for learners in some cases (e.g., Masson, 2013). These difficulties may arise from differences in pronunciation, meaning, and use of the words in the two languages. If such differences are not recognized, learners may erroneously apply L1 cognate knowledge during L2 use, resulting in potential comprehension problems.

To help understand the benefits and drawbacks associated with Japanese–English cognates, there is a clear need for further research. As a matter of course, such research will need to address the following two important questions: *Which English words have loanword equivalents in Japanese? Which of these Japanese loanwords are likely to be known to Japanese speakers?* These questions can be addressed using one or more of the following sources: loanword dictionaries, informants, and corpora.

Loanword dictionaries may be used as an authoritative source to determine whether a loanword exists in Japanese. However, there is considerable variation across such dictionaries in terms of the loanwords included (e.g., Daulton, 2008), which casts doubt on their reliability. Moreover, dictionaries traditionally do not provide information on whether the loanword is likely to be widely known among Japanese speakers, meaning they cannot be used to answer the second question above.

Informants, such as students or teachers, may be asked both whether and how well they know a particular loanword. However, this is a subjective measure, which may not accurately predict whether others will know the loanword, especially if only a small number of informants (e.g., two or three) are consulted. Recruiting a larger number of informants to do a word familiarity rating task would provide a more reliable measure; however, this is more time-consuming to perform.

Finally, corpora can be used to answer both questions. Not only can frequency data be collected quickly and straightforwardly, they are also highly correlated with rated familiarity of the same words (e.g., $r = 0.75$, Allen, 2018b). In other words, the more frequent the loanword is in Japanese, the more likely it is to be rated as known to most Japanese speakers. Likewise, if a loanword is very low-frequency or has zero-frequency (i.e., does not occur in the corpus), it is likely to be unknown to many Japanese speakers. Moreover, Allen (2018b) showed that loanwords that have a frequency of occurrence of at least one per million words in a corpus of Japanese tended to be known to the majority of undergraduates who rated their familiarity of the loanwords. Thus, this "one per million" threshold can be used as a general guide to determining loanwords that exist *and* are likely to be known.

Having argued that Japanese loanword frequency data is the optimum method of determining the cognate status of English words, I will present a method that allows researchers to access these data for use in their studies. In the following sections, I will first introduce the corpus that can be used to identify loanwords and the likelihood that they will be known. I will then illustrate the procedure of checking for loanwords by referring to a sample wordlist. After describing the process, I will briefly discuss the limitations of the resultant data so that they may be used judiciously.

2. Methods

2.1. The Japanese corpus

The corpus used to identify loanwords is the *Balanced Corpus of Contemporary Written Japanese* (BCCWJ; Maekawa et al., 2014; National Institute for Japanese Language and Linguistics, 2013). The BCCWJ contains 104.3 million words and is currently the most representative source of frequency data for words in the Japanese language (for more information, see https://pj.ninjal.ac.jp/corpus_center/bccwj/en/). A major advantage of this corpus for the present purposes is that all Japanese loanwords have been annotated with a "subLemma," which shows the English word from which it is derived. Therefore, when a search for an English word (e.g., *table*) is performed, the corresponding Japanese loanword (i.e., テーブル) can be identified along with its frequency data. This process can be automated using Excel so that a list of English words can be cross-referenced quickly and easily.

2.2. Identifying cognates

To identify loanwords, a wordlist (i.e., a list of English words for which we want loanword information) and the BCCWJ corpus wordlist are required, both open in separate spreadsheets in Microsoft Excel. The VLOOKUP formula in Excel can be used to extract information from the BCCWJ spreadsheet into our wordlist spreadsheet. The VLOOKUP formula is explained on many "how to" pages available on the Internet. A particularly helpful and relevant explanation is provided at http://crr.ugent.be/subtlex-nl/SUBTLEX_Excel.pdf. The authors explain basically the same process as below for obtaining English word frequency data using a different corpus wordlist. Here, I will explain a simple procedure that is specific to the BCCWJ data. This procedure is performed on a Mac operating system, so there may be minor variations when using other operating systems.

1. Visit the BCCWJ website page https://pj.ninjal.ac.jp/corpus_center/bccwj/en/freq-list.html and download the wordlist from the corpus (Short Unit Word list data: BCCWJ_frequencylist_suw_ver1_0.zip). The file is in .tsv format. Open the file, select all the data, copy it, and paste it into an Excel spreadsheet. Save this spreadsheet and keep it open.
2. To simplify and speed up the process, it is a good idea to select only the most relevant columns from the BCCWJ data and arrange them. Select the

following columns by clicking on them and holding command+c: lemma (the katakana loanword), subLemma (the English cognate), wType, frequency, and pmw (word frequency per million words). Copy these to a new spreadsheet and save it. Then, rearrange the columns such that subLemma is in Column A and the other columns are adjacent in Columns B to D. You can then sort the data by "wType," select all the rows that do *not* have have 外 (/gai/ "foreign") in the wType column and delete them. This will give you only the loanwords.

3. Open your wordlist (the words for which you want loanword information) in a separate Excel spreadsheet. For this illustration, there should be a list of individual words in the first column (column A) of the spreadsheet.
4. Click on the empty cell in Column B which is next to the first word in your list.
5. Go to "Formulas," "Lookup and reference," and select VLOOKUP. The formula builder will appear on the right of the spreadsheet.
6. Click on the first dialogue box "Lookup_value". Then, select the first word in your list (Row 1, Column A).
7. Click on the second dialogue box for "Table_array". Then, on your modified BCCWJ data spreadsheet, click and hold on Column A at the top to highlight the whole column and continue to hold down the mouse button while dragging rightward. This will allow you to select all four columns. You will see that the information has now been entered in the dialogue box "Table_Array".
8. Click on the dialogue box for "Col_Index". Input the column number for the data that you wish to extract: I want to extract the katakana loanword data first, so I enter the numeral 2 because the data for "lemma" is in Column B, which is the second column.
9. Click on the next dialogue box "Range_lookup". Enter the numeral 0. This means you will only extract data for words that exactly match the words in your list.
10. Click "done". If there is a loanword in the BCCWJ corresponding to the first word in your list, you should see the loanword in katakana. If the word was not identified as having a loanword in the BCCWJ, you will see "#NA".
11. Click on the cell with the katakana loanword data for your first word (Column B, Row 1). Click again on the bottom right corner of the cell and drag down. This will copy the VLOOKUP formula to all the cells in this column, giving you the katakana loanword data for your full wordlist.
12. Repeat the above process from step 4 to step 11 to add additional data, including "frequency" and "pmw". You can use Columns C and D in the same spreadsheet. After this, you will see the English word along with the Japanese loanword, the loanword frequency and the loanword frequency per million words.

3. Results and Discussion

Following the above process, I collected the cognate data for 20 words in approximately 10 min. Using the "sort" function, I ordered the words by "frequency" so that I could see which had loanwords and how frequent they are (Table 1).

Table 1. Sample word list and cognate data

Word	Lemma	Frequency	pmw
Alley	#NA	#NA	#NA
Beard	#NA	#NA	#NA
Burden	#NA	#NA	#NA
Cattle	#NA	#NA	#NA
Cellar	#NA	#NA	#NA
Dirt	ダート	5	0.05
Choir	クワイア	31	0.30
Poem	ポエム	71	0.68
Alarm	アラーム	126	1.20
Price	プライス	219	2.09
Mind	マインド	351	3.36
Rifle	ライフル	352	3.36
Bucket	バケツ	620	5.93
Money	マネー	855	8.17
Desk	デスク	1,033	9.87
Circle	サークル	1,434	13.71
Story	ストーリー	1,909	18.25
Tomato	トマト	2,570	24.57
Drama	ドラマ	4,861	46.47
Hotel	ホテル	10,503	100.40

The data show that 15 out of the 20 words have a loanword in the corpus, whereas five do not. The loanwords identified range in frequency from five occurrences to 10,503 occurrences. Twelve of the loanwords occur at a frequency above 1.00 per million words and are thus probably known to most Japanese speakers, whereas the remaining eight words are less likely to be known as loanwords, according to Allen (2018b).

In addition to the basic procedure above, there are a number of further considerations regarding derivatives, homonyms, and spelling, which researchers should be aware of.

3.1. Derivatives

In the procedure outlined, an *exact match* search was used. This means each word in the list (e.g., *hotel*) was cross-referenced with the BCCWJ list to see if it appeared as a "subLemma" in English. Any derivatives of the target word, such as *hotels* or *hotelier*, would not be identified because they do not exactly match the input word (*hotel*). If frequency data for word families is required, the individual frequencies of each derivative must be identified and summed.

3.2. Homonyms

Some Japanese loanwords have multiple distinct meanings and are therefore homonyms (e.g., ライト /raito/ can mean "not heavy" or "not dark"). In most cases,

the creators of the BCCWJ have differentiated frequencies for each specific meaning of the loanwords. For example, the frequency of the meanings of ライト in Japanese were as follows: *light* (光 /hikari/ "not dark") = 1,002 and *light* (軽い /karui/ "not heavy") = 1,530. When searching for cognate frequencies using the exact match method, the result for *light* is #NA, suggesting it does not occur in the corpus. This is because in the BCCWJ list, there is "light (光)" and "light (軽い)," neither of which exactly correspond to *light* because of the inclusion of the Japanese definition in the cell. To resolve this issue, the search function (command+F) can be used to locate each entry of the target word, then the frequencies can be summed to provide the loanword frequency that includes both meanings in Japanese (e.g., *light* = 2,532 occurrences). In general, when a search results in #NA, it is good practice to manually search for the English words to double-check that they do not occur in the corpus. For words such as *light*, this would reveal the existence of homonyms.

3.3. Spelling

Because the BCCWJ uses primarily US spellings, the researcher should be careful of spelling differences in the two lists, as alternative spellings will give null results (e.g., *grey*/*gray*).

Finally, a number of important limitations must be borne in mind when using frequency data to predict human language knowledge and behavior. Firstly, the BCCWJ, which is collected from newspapers, novels, and other text genres, will be most representative of lexical knowledge of Japanese speakers who read some or all of these kinds of texts. Secondly, some loanword frequencies may be exaggerated or understated due to the make-up of the corpus and therefore may less accurately predict to what extent loanwords are widely known. For example, クワイア /kuwaia/ "choir", which is very low frequency in the corpus, may in fact be well known to many Japanese speakers, while マインド /maindo/ "mind" may not, even though it occurs at a relatively high frequency. When using any corpus data, researchers must be cognizant of these limitations.

4. Conclusions

In this short methodology paper, I have described a simple procedure using the BCCWJ corpus to identify which English words have loanwords and whether these loanwords are likely to be known in Japanese. Using the BCCWJ data will allow researchers to control and predict the potential influence of cognates in vocabulary learning and assessment. Using corpus data is not only more efficient than referring to intuitions of informants and/or dictionary entries but will also improve the generalizability and replicability of research findings.

While this paper has focused specifically on Japanese–English cognates, the issue of identifying cognates/loanwords for research and educational purposes is not restricted to these languages. However, to my knowledge, there are no corpora in other languages that have been annotated for cognate/loanword status in the way that the BCCWJ has, which makes their identification infinitely more demanding. Nevertheless, given the importance of both word frequency and cross-linguistic influence in language learning, such annotated corpora would surely be well received by researchers in the language sciences.

Acknowledgements

I am very grateful to Koji Miwa for initially pointing me in the direction of the BCCWJ frequency wordlists back in 2016. I also wish to thank two anonymous reviewers for their helpful comments on an earlier version of this manuscript.

References

Allen, D. (2018a). Cognate frequency and assessment of second language lexical knowledge. *International Journal of Bilingualism*. Published Online: 2018-06-22. doi:10.1177/1367006918781063

Allen, D. (2018b). The prevalence and frequency of Japanese-English cognates: Recommendations for future research in applied linguistics. *International Review of Applied Linguistics in Language Teaching*. Published Online: 2018-02-22. doi:10.1515/iral-2017-0028

Allen, D. (2019). An Overview and Synthesis of Research on Japanese-English Loanwords. *Vocabulary Learning and Instruction*, *8*(2), 8–25. doi:10.7820/vli.v08.2.allen2

Allen, D., & Conklin, K. (2013). Cross-linguistic similarity and task demands for Japanese – English bilingual processing. *PLoS One, 8*(8), e72631. doi:10.1371/journal.pone.0072631

Daulton, F.E. (2008). *Japan's built-in lexicon of English-based loanwords*. Clevendon: Multilingual Matters.

Hoshino, N., & Kroll, J. F. (2008). Cognate effects in picture naming: Does cross-linguistic activation survive a change of script? *Cognition, 106*, 501–511.

Maekawa, K., M. Yamazaki, T., Ogiso, T., Maruyama, H., Ogura, W., Kashino, H., … Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources & Evaluation, 48,* 345–371.

Masson, M. E. (2013). How L1 loanwords can create a false sense of familiarity with L2 vocabulary meaning and usage. *Vocabulary Learning and Instruction, 2*(1), 8–14.

Miwa, K., Dijkstra, T., Bolger, P., & Baayen, H. (2014). Reading English with Japanese in mind: Effects of frequency, phonology, and meaning in different-script bilinguals. *Bilingualism: Language and Cognition, 17*(3)*,* 445–463.

Nation, I. S. P. (2012). *The BNC/COCA word family lists*. Unpublished paper. Retrieved from www.victoria.ac.nz/lals/about/staff/paul-nation

National Institute for Japanese Language and Linguistics. (2013). *BCCWJ word list*. Retrieved from https://pj.ninjal.ac.jp/corpus_center/bccwj/en/freq-list.html

Rogers, J., Webb, S., & Nakata, T. (2014). Do the cognacy characteristics of loanwords make them more easily learned than noncognates? *Language Teaching Research, 19(1),* 9–27.

Stubbe, R., & Yokomitsu, H. (2012). English loanwords in Japanese and the JACET 8000. *Vocabulary Education and Research Bulletin, 1,* 10–11.