2S－5

Japanese Idiom Frequencies in Literary and Newspaper Corpora - A Statistical Approach -

Danny MINN[†], SANO Hiroshi[‡]
Tokyo University of Foreign Studies
Graduate School of Area and Cultural Studies[†], Faculty of Foreign Studies[‡]

1 Introduction

For non-native learners of Japanese, much time and effort is required to learn the proper usage of Japanese idioms. The main reason for the difficulty is that two or more words form an idiom to represent one meaning that may not relate to the individual words directly. Depending on the type of word (a basic word, a specialized word, etc.), the frequency with which the word is used varies. For example, the frequency distribution of words used in spoken expressions and written expressions differs. Similarly, the frequency of idioms should differ according to the type of writing or speech.

In order to learn idioms efficiently, Japanese learners should prioritize and learn the most frequently used idioms. This paper will discuss how two corpora (literary and newspaper) were used to examine the frequency of Japanese idioms. From this research, we were able to find the idioms that appeared in both genres frequently and not so frequently, and their distributions. The next step will be to use other types of corpora, and to organize the findings for Japanese learners.

2 Research Goals

For non-native learners of Japanese, the study of most Japanese idiomatic expressions is reserved for intermediate to advanced levels. As is the case with all second language acquisition, fundamentals such as grammar, pronunciation, and vocabulary take precedence over idiom learning. However, after these fundamentals are mastered, language learners inevitably come across idiomatic expressions. There is a wealth of idioms used in the Japanese language, and there are many books written on the subject. The books surveyed here have anywhere from 100 to 2,000 idioms described (Akiyama, Garrison, Maynard, Miyaji, Sasaki). With the aid of these reference books, one can find the meaning of a particular idiom, its English equivalent, some etymological information, or even some example sentences. Some of these books even have illustrations for comedic effect or as memory devices. These references do not, however, include much information about how frequent a particular idiom is used.

Even with these references, one would find it difficult to determine an idiom's relative frequency of use without the advice of a native speaker or rigorous research. In lieu of those options, using corpora and computer programs to process and search the data, statistical analysis of idioms is possible. Of course, native speakers and especially teachers of Japanese have an innate sense of which idioms are used frequently. However, it would be difficult for them to produce concrete evidence of that innate sense. By using the results of corpus analysis, both teachers and learners of Japanese can recognize which idioms are the best to learn according to one's language ability level or field of work or study.

In this paper, we will discuss three points:

1. Idiom frequencies in two corpora were examined, producing results that were genre-specific.
2. Non-native speakers of Japanese do not know many of the most frequently used idioms, even after several years of study. This may indicate a gap in textbooks or reference materials.
3. Based on frequency, an idiom database would be desirable and helpful to Japanese learners and teachers.

3 Creating a "Top" List of Frequency Data

In order to find out how well non-native Japanese speakers know Japanese idioms, a questionnaire was to be distributed. However, as there are far too many idioms that exist, it was necessary to find a method to pick the most frequently used idioms. Rather than inundating volunteers with a long list of idioms, or just picking a shorter list of idioms at random, it was deemed preferable to pick the most frequently used idioms.

The first task at hand was to enter a list of idioms into a computer. The preliminary list was taken from "The Complete Japanese Expression Guide" (Sasaki 1993). The next task was to search for this list of about 300 idioms in corpus data, namely the "Shinchoubunko No 100 Satsu, CD-ROM" and the "Mainichi Shinbun, 2000, CD-ROM." The Shinchoubunko data consisted of 100 literary works published by the Shinchoubunko publishing company. The corpus used contained approximately 10.5 million characters. The Mainichi Shinbun, 2000 data consisted of one year's worth of newspaper articles published by the Mainichi Shinbun newspaper. This corpus contained approximately 65 million characters. Through the use of a computer program written in Perl, it was fairly simple to find and count all of the occurrences of the idioms in the preliminary list.

It became clear that there were just too many types of idioms, and that it would be helpful to just concentrate on one type of idiom. At this stage, 127 idioms from the previous list that fall into the category of "noun + particle + verb" (NPV) were selected. Again, using the previous Perl programs, the frequency of the 127 idioms in the Shinchoubunko and Mainichi data was determined. With this data, it was easy to see which idioms from our list

were used most frequently, at least in the two corpora (Tables 1 and 2).

Noun + Particle + Verb Idiom	Number of Occurrences
1. kidou ni noru, 軌道に乗る*	395
2. te wo dasu, 手を出す*	194
3. te wo utsu, 手を打つ*	169
4. te wo nuku, 手を抜く	108
5. toukaku wo arawasu, 頭角を現す	100
6. akaji ni naru, 赤字になる	90
7. urame ni deru, 裏目に出る	89
8. te ga todoku, 手が届く	85
9. mizu wo sasu, 水を差す	75
10. kimo ni mejiru, 肝に銘じる*	62

Table 1: Top 10 List of NPV Idioms in "Mainichi Shinbun, 2000"

Noun + Particle + Verb Idiom	Number of Occurrences
1. te wo dasu, 手を出す*	97
2. te wo utsu, 手を打つ*	47
3. haji wo kaku, 恥をかく*	40
4. aizuchi wo utsu, 相槌を打つ	32
5. uchouten ni naru, 有頂天になる	30
6. touge wo kosu, 峠を越す*	27
7. kubi ni naru, 首になる	24
8. toppyousi mo nai, 突拍子もない	22
9. hana ni kakeru, 鼻にかける	13
10. kidou ni noru, 軌道に乗る*	12

Table 2: Top 10 List of NPV Idioms in "Shinchoubunko No 100 Satsu"

* Appear in both the Top 20 of Mainichi data and the Top 10 of Shinchoubunko data

4 A Survey of Non-native Speakers of Japanese

From the results, it was possible to construct a questionnaire that contained a "Top 20" list, which consisted of the top 19 idioms appearing in the Mainichi Shinbun and one from the Shinchoubunko data (The top three in the Shinchoubunko data also appear in the top 20 of the Mainichi data, so No. 4 from the Shinchoubunko data was added to our surveyed list). The questionnaire asked whether or not the subject knew each idiom. It also asked whether or not they often used or saw each idiom, and where they often encountered each idiom. The preliminary results of a survey of non-native Japanese speakers showed that they only knew about half of the "Top 20" list, even though they had been studying Japanese for an average of over five years. At a minimum, we can interpret from the frequency data that 19 of the 20 idioms appear frequently in newspapers and the remaining idiom appears frequently in literary works. We cannot assume that they all appear in everyday conversation, but it is fair to say that adult, native Japanese speakers know all of the idioms almost without exception. We can also interpret from the data that non-native students are probably not taught many of these idioms in the classroom.

5 Genre-Specific Frequency Data

Not only can we find the most frequent idioms, we can determine where they usually appear. For example, in the case of "kidou ni noru," (meaning, to get on track; literally, to ride an orbit) the most frequent in our preliminary list, there were 395 instances in the Mainichi data. 34% of the instances appeared in the economics section, 14% in the international news section, and 9% on the front page. We can reasonably deduce from our data, that this idiom would be useful to learn if one desires to read Japanese newspapers, and more specifically, economics related news. The idiom "toukaku wo arawasu," (meaning, to be outstanding; literally, to show one's antlers) the fifth most frequent in our list, appeared in the Mainichi data 100 times. 32% of the instances appeared in the sports section, 14% in the general affairs section, and 10% on Page 2. This idiom also would be useful to learn for those who want to read Japanese newspapers, as it appears often in the sports and general affairs sections. This kind of analysis can easily be applied to the rest of the idioms.

6 Conclusion

The results of this preliminary analysis of Japanese idioms proved to be promising. Using the two corpora, we were able to determine the frequency distributions of 127 NPV idioms. As is the case with individual words, the frequency of idioms also depends on genre. The next step will be to increase the number and types of idioms to search and the size and types of corpora. This is a daunting task as there are thousands of idioms, and hence, there will be an enormous amount of matched data to sift through. The results will, however, give us more complete frequency data. Based on this data, we hope to create a database of Japanese idioms that will aid learners of Japanese.

References

Akiyama, Nobuo and Carol Akiyama. 2001 Japanese and English Idioms. New York: Barron's Educational Series, Inc., 1996.

Garrison, Jeffrey G. "Body" Language. Tokyo: Kodansha International, 1990

Garrison, Jeff and Kayoko Kimiya. Communicating with Ki, The "Spirit" in Japanese Idioms. Tokyo: Kodansha International, 1994

Maynard, Michael L. and Senko K. Maynard. 101 Japanese Idioms. Chicago: Passport Books, NTC Publishing, 1996.

Miyaji, Yutaka. Kanyouku No Imi To Youhou. Tokyo: Meiji Shoin, 1982.

Sasaki, Mizue. The Complete Japanese Expression Guide. Tokyo: Charles E. Tuttle Co., 1993.