LEXICAL ANALYSIS OF IDIOMS USING FRAMENET AND LEXICAL SETS

Irena Srdanović Erjavec, Kikuko Nishina 仁科喜久子
Tokyo Institute of Technology

Abstract: This paper discusses some new insights about multi-word units that can be revealed through corpus-based lexical analysis. First, the corpus frequencies of two Japanese idioms, *ki ni suru* and *ki ni kakeru,* are examined and compared with frequencies and dictionary descriptions of single-word units of similar meanings. Then, the theory of Frame Semantics and its contribution to lexical analysis are introduced, with special emphasis on the information for multi-word units presented in the FrameNet database. Finally, words that function as objects of the idioms are classified into different lexical sets, which helps to identify the senses and usage of the idioms.

Keywords: Multiword units, idioms, FrameNet, lexical sets, corpus-based lexical analysis, Japanese, dictionaries

INTRODUCTION

Consulting learners' and standard dictionaries of the Japanese language often leaves one almost empty-handed regarding the meaning or usage of even frequently used idioms. And, it is even more difficult if a dictionary user wants to grasp the differences between idioms bearing similar meanings. Corpora and various methods of meaning analysis, which are becoming standards in the dictionary creation process nowadays, can also provide us with much more precise and complete data on the meanings and usages of multi-word units. In this paper, we define multi-word units as groups of words that occur together more often frequently that would be expected by chance. Idioms refer to multi-word units whose meaning cannot be completely understood from the meanings of their component parts.

IDIOMS IN DICTIONARIES AND CORPORA

To investigate the dictionary definitions of *ki ni suru* and *ki ni kakeru* idioms, we consulted some of the best-known Japanese monolingual and Japanese-English bilingual dictionaries, *Koujien*, *Daijirin*, *Kenkyusha* and *Sanseido*. The idioms can be found under the common headword *ki* (気) in long and not always easily searchable lists of numerous idioms (also proverbs) containing the word *ki*. The lists are sorted either by gojuonjun (pronunciation) or by senses and then by gojuonjun – the later one shows to be more time consuming. The definitions provided in the dictionaries are not descriptive, but rather offer near-synonyms as an explanation of the meaning. This might be helpful to understand the basic meaning of the idioms, but it is not sufficient to aid the user to grasp the various senses and usage of each idiom, as well as the difference between the idioms, which is what learners of Japanese usually need. Koujien and Sanseido do not offer an explanation for the *ki ni kakeru* idiom, and Kenkyusha only provides a link to the *ki ni suru* entry. Examples are only presented in Sanseido (for *ki ni suru*).

The corpus-based analysis indicates that some idioms are so frequent that they deserve much more space in dictionary explanations, but they are allotted less space than their less frequent one-word unit friends. In order to compare the frequencies of *ki ni suru* and *ki ni kakeru* idioms with some semantically related words (emotion verbs) (Teramura, 1982: 152-3), we employed two Japanese corpora: 1) Aozora bunko corpus (8 million tokens) and Chakoshi, a system for corpus data exploration, available at the web site of the University of Nagoya; 2) JapWaC, a large web corpus (400 million tokens) searchable through the Japanese Sketch Engine (Erjavec et al, 2007).

The frequency of *ki ni suru* is quite high: in the corpus of literary texts it is generally as frequent as the verbs *kinchou suru*, *hotto suru*, *urotaeru*, *shitsubou suru*, and it is more frequent than *kitai suru*, *koi suru*, *ai suru*, *natsukashimu*. In the web corpus, it is even more frequent – as frequent as *shimpai suru*, *okoru*, *anshin suru*, and more frequent than *kinchou suru*, *hotto suru*, *kirau*, *konomou* and many other verbs. In contrast, *ki ni kakeru* is much less frequent, but it is more frequent than *koi suru, ai suru* and *natsukashimu* in the Aozora bunko corpus and more frequent than *urotaeru* and *natsukashimu* in the web corpus. Although the listed verbs are of similar frequencies, or even less frequent than *ki ni suru* and *ki ni kakeru*, the dictionaries provide much more space for descriptions of their meaning and usage - sometimes five to ten times more than for the idioms. Comparing the two corpora also provides interesting results about their differences in lexica.

FRAMENET AND IDIOMS

As the procedure and results of the lexical analysis are discussed in more detail in Srdanovic Erjavec (2006), this paper focuses more on how the FrameNet project (http://www.icsi.berkeley.edu/~framenet/) tackles multi-word units.

The project aims to create an online lexical resource based on Frame Semantics theory and corpus data, originally for the English language. It describes the combinatorial properties of words in a sentence, annotating their semantic roles and their syntactic forms and functions. A lexical unit (a given sense of a word) is provided with a frame it belongs to, including a description of the frame and its elements (semantic roles), with a definition, annotated corpus examples and valences information. Multi word units can be also assigned to semantic frames like one-word units, but, depending on their type, the information provided varies in the database (Ruppenhofer et al, 2002).

One category represents multi-word lexical units that are marked as such and assigned semantic frames. These are, for example, lexicalized noun compounds (*calendar year*), verb-particle collocations, with an obligatory particle (*back out*), and various types of idioms (*a pain in the neck*). Such multi-word units are placed into frames based on their meaning – the meaning of all the constituents of a unit, and not on their single meanings. The unit *calendar year* belongs to the *Calendric_unit* frame, the *back out* unit is assigned to the *Going_back_on_a_commitment* frame, and *a pain in the neck* is part of the *Difficulty* frame. Another type are multi-word units not presented as one lexical unit including all of the constituents, but their parts represent core Frame Elements of the unit or their syntactical realizations (valence roles). These are, for example, verb-particle collocations, where verbs can also be free (*object to*, *prevent from,* etc.), noun collocations that include modifiers (*navy captain*), collocations with transparent nouns (*swarm of bees*), and event nouns in collocation with support verbs (*make complaint*).

Ki ni suru and *ki ni kakeru* belong to the first category and we handle them as one-word units. We assign different Semantic Frames and Elements to the idioms, and, accordingly, divide them into different senses, as suggested by Atkins et al (2003) – if a word belongs to different frames it can be divided into different senses. Based on English FrameNet resources, which are under development and incomplete for the emotional domain, and *ki ni suru* and *ki ni kakeru* examples from the Aozora Bunko corpus, we assign five different frames/senses to *ki ni suru* and two different frames/senses to *ki ni kakeru*. More information about the frame descriptions and the senses in Japanese is provided in Srdanović Erjavec (2006).

Ki ni suru: *[Semantic_frame (Element1 and Element2)]*

1. to worry (feel uneasy) about sb/sth Emotion_active (Experiencer and Topic)
   Sadayo no koto wo ki ni shite iru.
2. to be bothered by sth / to mind sth. Emotion_directed (Experiencer and Stimulus)
   Kyuukutsuna uwagi no kata wo ki ni shinagara ...
3. to have one's thoughts occupied with
   / to be on one's mind constantly Cognition (Cognizer and Topic)
   Kono bou no koto bakari ki ni shite ita.
4. to care for / to pay undo attention to sth (often in negation) Interest (Cognizer and Content)
   Kimono nado wa amari ki ni shinai.
5. touch or rub constantly Reaction_to_stimulus (Agent and Stimulus_affected)
   Tebukuro no yabure wo ki ni shinagara chiisai kodomo...

Ki ni kakeru:

1. to worry (feel uneasy) about sb/sth,
   to bother one's mind Emotion_active (Experiencer and Topic)
   Hahaoya wa kono kuse wo ki ni kakete inakatta.
   Doko e itta mono de arou to akekure Kumiko fujin ga ki ni kakete iru uchi ni...
2. ?Interest (Cognizer and Content) (a larger corpus is needed to confirm this sense)

LEXICAL SETS AND BEYOND

We also classify the words that function as objects of the idioms into various lexical groups. As suggested by Hanks (1997), different lexical sets of neighboring words that function as subjects or as objects should reveal different senses of the word in question. The subjects of *ki ni suru* and *ki ni kakeru* are always human, but the objects are more heterogeneous and, therefore, we divide them into 11 groups for *ki ni suru* and 8 groups for *ki ni kakeru* (for example, *human: activities or state of affairs in connection with people*; *other people's talk/opinions*; *body parts*; *parts of clothes*, etc.) (Srdanović Erjavec, 2006).

[Kare] ga [kimono nado] wa amari ki ni shinai.
[Cognizer] [Content: *clothes_general*] (4.) *to care for*, Cognition frame)

A specific lexical group is usually combined with a specific meaning and the usage of the idioms, so the classification provides new information about the idioms' usage. For example, the fourth meaning combines with objects classified under the lexical group of *clothes/appearance_general*; it is usually in negation and describes a person's attitude. However, the meaning and usage are also conditioned by the wider context, in other words, by the situations in which the idioms occur, and we therefore further classify the examples into different situational types. These can yield new insights into the differences in the uses of the two idioms. For example, the relationships between participants in *ki ni suru* examples are closer than is the case for *ki ni kakeru*. Also, for the sense *to worry about sb.*, *ki ni suru* is used for more serious problems than *ki ni kakeru* (death, illness vs. how is somebody, what he is doing).

CONCLUSION

This conducted corpus-based analysis of idioms first of all reveals that dictionaries underestimate the importance of idioms. They should provide more space for the description of idioms' meanings and usage; generally as much space as given to equally frequent one-word units. Also idiom descriptions should be more easily accessible for dictionary users. Assigning different Semantic Frames to idioms, based on English FrameNet, can help to divide idioms into different senses. The classification of the idioms' relevant neighbor words into different lexical sets reveals some information about the usages of the idioms and also helps to identify their senses. Only by consulting the context and by classifying examples into situational types can we obtain clearer information about how these two synonymous idioms actually differ. As future work, the analysis can be repeated using other data and thus to make comparisons among various types of corpora – for example, the recently available JapWaC, or a balanced Japanese corpus, when one becomes available.

REFERENCES

Atkins, S., Charles J. F., Christopher R. J.(2003). *Lexicographic Relevance: Selecting Information from Corpus Evidence. International Journal of Lexicography 16/3*. Oxford: Oxford University Press. 251–280.

Collick, M., Kazuo, H., Munekazu, T., ed. (1995). *Kenkyusha's New College Japanese-English Dictionary. 4.ed.* Tokyo: Kenkyusha.

Erjavec, T., Kilgarriff, A., Srdanovic Erjavec, I. (2007). *A large public-access Japanese corpus and its query tool. Inaugural Workshop on Computational Japanese Studies*. .

Fillmore, C. J., Johnson, C. R., Petruck, M. R. L.(2003). *Background to FrameNet. International Journal of Lexicography 16/3*. Oxford: Oxford University Press. 235–250.

Hanks, P.(1997). *Lexical sets: relevance and probability. Translation and Meaning. Part 4*. Maastricht: School of Translation and Interpreting.

Konishi, T., Inoue, N., ed.(1996). *Nyuu senchurii waei jiten. 2.ed*. Tokyo: Sanseido.

Matsumura, A., ed.(1995). *Daijirin*. 2.ed. Tokyo: Sanseido.

Ruppenhofer, J., Baker, C.F., Fillmore C.J. (2002). *Collocational Information in the FrameNet Database. Proceedings of the Tenth Euralex International Congress Vol. I.* Denmark: EURALEX. 359-369.

Shinmura, I., ed. (1998). *Koujien daigohan CD han*. Tokyo: Iwanami Shoten.

Srdanović Erjavec, I. (2006). *Ko-pasu jishogaku ni okeru imi bunseki: "ki ni suru" to "ki ni kakeru" no kanyouku wo rei to shite. Dai19kai Nihongo kyouiku renraku kaigi ronbunshuu*. (in print)

Teramura, H. (1982). *Nihongo no shinakusu to imi I*. Tokyo: Kuroshio shuppan.