Collecting figurative expressions using indicators and a semantic tagged Japanese corpus

Rei Kikuchi[†], Sachi Kato[‡], Masayuki Asahara[‡]
[†]Chuo University, [‡]National Institute for Japanese Language and Linguistics
a12.7cyh@g.chuo-u.ac.jp

Keywords: simile, figurative expression, figurative indicator, semantic tagged Japanese corpus

We constructed a large-scale figurative expression database using indicators and a semantic tagged Japanese corpus. Corpus surveys using indicators such as co-occurrence and patterns are useful for example collection. Collecting examples of similes with elements of figurative indicators in Japanese figurative expressions from a corpus appears to be easy. However, collecting figurative expressions with indicators requires human judgement to determine whether the collected samples are really figurative expressions. For example, if candidate expressions containing the possible figurative indicator “like” or “as” are collected, it is necessary to judge the candidates to extract the figurative ones, as this word has multiple meanings such as estimation, illustration, appearance, and similarity. Furthermore, for indicators such as “conscious”, which has multiple possible meanings including “experience,” “feel,” “perceive,” and “sense,” it is difficult to conduct an exhaustive search of examples containing synonymous expressions. In other words, using indicators to collect figurative expressions from a corpus involves collecting examples containing elements that could act as figurative indicators, including synonymous expressions, followed by extracting the actual figurative language from these potential candidates. Given this difficulty in collecting large volumes of figurative expression data, determining the extent to which a figurative expression is used, as well as the kinds of figurative expressions used in different genres, has been impossible.

In this study, we aimed to collect figurative expression data from large-scale corpora by first collecting simile expressions that contained elements that could act as figurative indicators, and used human judgement to extract the figurative expression from these example groups. The corpus data used was the core data of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (approximately 1 million words) and the semantic tagged BCCWJ (approximately 350,000 words from newspapers, magazines, books, and other publications (Kato et al., 2018)). The 441 indicators for figurative elements defined by Nakamura (1977) and the semantic tags assigned to them were used as cue phrases. When judging whether an expression was figurative, the preceding and succeeding 100 words of the target phrases that could act as indicators (cue phrases) were shown as context, and the portion with the figurative expression was extracted. The source and target domains for the expressions were simultaneously annotated, along with the type information.

Through this research, we collected 97,118 examples of potential figurative expression candidates, with 923 determined to be figurative. In other words, less than 1% (0.95%) of the examples was figurative. Through this study, we demonstrated the difficulty of collecting figurative expressions. We also added information on the topic, vehicle, source domain, target domain, and type to the collected examples. Furthermore, by using a semantic classification, we expanded the cue phrases that could contain indicator elements and were able to collect figurative expressions synonymous with these indicators. Through this large-scale collection of figurative expressions and the addition of defining information, BCCWJ enables us to study the current state of figurative expression usage in Japanese contemporary written works.

References

Kato, Sachi, Masayuki Asahara, Makoto Yamazaki. (2018). Annotation of `Word List by Semantic Principles' Labels for the Balanced Corpus of Contemporary Written Japanese. PACLIC-32.

Nakamura, Akira. (1977). *Theory and Category of Metaphor Expression.* Shuei Shuppan. (in Japanese).