【Research Note】

Morphological analysis of sentences in the Japanese language textbooks for sixth graders in Japanese elementary schools

Taeko Ogawa[1] and Chikako Fujita[2]
(1: Tokai Gakuin University, 2: Nanzan University)

Abstract

This study aims to exhaustively extract words included in sentences in elementary school Japanese language textbooks and to investigate the characteristics of words that are used as learning materials in Japanese language education. We conducted morphological analysis to extract words from sentences in a sixth-grade textbook. A total of 5,002 words were extracted from all the printed material in the textbook. Words were counted for each part of speech for these words, and word token frequency was reported. Problems related to morphological analysis of sentences in Japanese language textbooks are discussed.

Keywords: morphological analysis, Japanese language textbook, corpus linguistics

Introduction

In recent years, the importance of corpus linguistics has been increasing in many areas related to language research and language education (Halliday, Teubert, Yallop, & Čermáková, 2004; Ishikawa, 2012; Maekawa, 2013; Maekawa, Yamazaki, Ogiso, Maruyama, Ogura, Koshino, Koiso, Yamaguchi, Tanaka, & Den, 2014; Meyer, 2002). Corpus linguistics is a research field that observes various characteristics of languages mainly from an empirical point of view and investigates and analyzes them based on a corpus which is a collection of texts of written (or spoken) language presented in electronic form. Corpus analysis has revealed language characteristics that could not be found in conventional experimental methods and investigations that use a relatively small number of language materials. In this study, which focuses on the word reading process, we use elementary school Japanese language textbooks in Japan as a corpus. This was done to clarify the organization of a mental lexicon that develops as children acquire knowledge such as the orthographic, phonological, and semantic information of words.

When looking for various kinds of information such as part-of-speech (POS) and phonology of words, it is necessary to extract individual words from the sentences included in the corpus. The technique of extracting the words is called “morphological analysis”; a text written in Japanese is used as a text input, and a sentence is divided into words by using a morphological analyzer with a pre-built system dictionary.

“Morpheme” in morphological analysis does not mean morpheme as defined in linguistics, but generally refers to individual words. Since words are separated using spaces in alphabetical languages such as English, it is relatively easy to specify individual words that constitute a sentence. However, the boundary between words is unclear in Japanese sentences, and in addition, three types of scripts (Kanji, Hiragana, and Katakana) are used, so the orthographic variation is large (Joyce, Hodošček, & Nishina, 2012); it is thus difficult to define a word. In this regard, in the process of building “the Balanced Corpus of Contemporary Written Japanese” (BCCWJ, hereafter) and “UniDic,” a

morphological analyzer dictionary developed by the National Institute for Japanese Language and Linguistics (NINJAL hereafter), two linguistic units have been defined: short-unit word (SUW) and long-unit word (LUW) (Maekawa et al., 2014; Ogura, Koiso, Fujiike, Miyauchi, Konishi, & Hara, 2011).

BCCWJ is a corpus built to understand the overall picture of the modern Japanese written language and is currently the only balanced corpus available for Japanese. The corpus contains approximately 100 million words, and the words cover genres such as books, magazines, newspapers, white papers, internet blogs, internet bulletin board, school textbooks, laws, and others. Samples are extracted randomly from each genre (Maekawa et al., 2014). SUW and LUW, as adopted by BCCWJ, were assigned morphology information such as POS and other important information for reading. The SUWs are morphological units used for collection purposes and are defined as the smallest units with meaning in modern languages. The LUWs are syntactical units and are specified by dividing each phrase into content words and function words according to specific LUW rules (Ogura et al., 2011). In other words, the SUWs are a unit representing the level in the entry of a traditional dictionary; basically, morphemes. LUWs, however, generally represent the level of compound words such as compound nouns and compound verbs, are close to the phrase level (Joyce et al., 2012; Maekawa et al., 2014).

The UniDic morphological analyzer dictionary (Den, Ogiso, Ogura, Yamada, Minematsu, Uchimoto, & Koiso, 2007) was developed in accordance with the development of BCCWJ as a dictionary for dividing Japanese text into words and giving morphological information. MeCab is a morphological analyzer developed by Kudo, Yamamoto, and Matsumoto (2004), and UniDic-Mecab is software that can use UniDic as a system dictionary. There are four main features of UniDic. (1) It is designed using SUW, which is a uniform unit without orthographic variation. (2) It has three hierarchical structures: *lemma*, *word form*, and *written form*. It also can give the same heading regardless of the orthographic variation or changes of word form. The top level of this hierarchical structure is a lemma. The next layer distinguishes the differences in word form, and the lowest layer is a written form for distinguishing orthographic differences. Figure 1 shows an example of the hierarchical structure of UniDic (Ogura et al., 2011). By setting up such a hierarchical structure, we can obtain information on the type of change of a word form, and how much the orthographic variation fluctuates. (3) Useful morphological information for language studies such as POS can be given. (4) It can give information on accent and sound change, and it can be used for research on speech processing. It was reported that the analytical accuracy of morphological analysis by UniDic-Mecab is extremely high as high as 97% or more in research on newspaper articles and literary works (Ogiso, 2014).

語彙素 /goiso/ 'lemma'	語形 /gokei/ 'word forms'	書字形/shojikei/ 'orthographic forms'
矢張り(adverb) 'also; absolutely'	ヤハリ	やはり
		矢張り
	ヤッパリ	やっぱり
		矢っ張り
	ヤッパ	やっぱ

Figure 1. An example of three basic levels of UniDic cited from Ogura et al. (2011).

Our study focuses on the development of language expected from pupils across the six years. We aim to develop materials to examine the development of the mental lexicon for written language based on textbooks used in Japanese language classrooms. Specifically, we cover sixth grade textbooks; this is the highest grade in elementary school. Kyōiku kanji (“education kanji”), which is learned in elementary school, has 1,006 kanji characters, and the school year's structure has been determined according to curriculum guidelines (“courses of study”) for elementary schools (the Ministry of Education, Culture, Sports, Science and Technology: MEXT, 1989). For textbooks for lower grades, if a word that includes an unlearned kanji script is part of a text, hiragana scripts are usually assigned instead of kanji.

For example, the second character of the word “文章” (pronounced /bun-shou/, meaning ‘sentence’) is assigned to third graders. Since the second character “章” is not used in the second-grade textbook, it is written using a kanji-kana mixed form, such as “文しょう.” This means that for lower grades, there are more kanji-kana mixed forms and characters constituting words usually written in kanji are instead written in hiragana. This suggests that it is highly possible to encounter words that are not registered in the morphological analysis dictionary, and therefore it is estimated that the accuracy of morphological analysis decreases.

Fujita, Taira, Kobayashi, and Tanaka (2014) point out that accuracy of morpheme analysis declines in picture books that contain a lot of words in the hiragana script compared to newspapers that contain many words written in the kanji script. Similar problems are expected to occur in lower grade textbooks as well. Therefore, in this study, we first picked up a sixth-grade textbook that uses all the kanji learned in elementary school, and then conducted a morphological analysis using the UniDic-Mecab morphological analyzer. Based on the results of this analysis, we can also consider morpheme analysis for lower grade textbooks containing a lot of words written using mixed kanji and kana script.

Method

Materials

This study looks at the Japanese language textbook “*Kokugo Roku Souzou*” (2016 version) published by the Mitsumura Tosho Publishing Co., Ltd., which is the most widely adopted textbook in elementary schools in Japan. This textbook consists of a total of 284 pages. It contains various themes such as stories, essays, explanatory texts, poetry, and kanji learning, and consists of nine learning units.

Procedure

A textbook corpus was created excluding the following from all characters, numbers and symbols printed in the textbook.

(1) Front cover, inner cover, back cover, and spine.

(2) The page numbers printed on the lower part of each page.

(3) “*Rōmaji* (Roman alphabet) table” (p.264 of textbook) and “Table of kanji learned in six years” (p.265-284) included in the appendix.

(4) The characters in the pictures of the cover of the books which are introduced in the textbook.

(5) Arabic numerals referencing pages and rows (e.g., the Arabic numerals 9 and 8 from the text “line 9 on page 8”).

(6) Arabic numerals surrounded by circles that are only used in Japanese sentences (Figure 2 shows examples of such numerals to be excluded). Other Arabic numerals were not excluded.

(7) Furigana is given in kana script to explain the reading of kanji. In the textbooks, there are many kanji words, have not yet been learned, to which furigana are assigned. When performing morphological analysis, the existence of these furigana in the texts reduces accuracy, so they were excluded from this study.

(8) Symbols of geometric patterns (e.g., square and inverted triangle) were excluded using the search replacement function of MS word (see Figure 2). Other symbols (e.g., diamond) and symbols such as parentheses were also excluded from the analysis.

names	symbols
Arabic numerals in circles	①, ②, ③, ④, ⑤, ⑥
Symbols	▼, →, ←, □, ○, △, ×
Special characters representing repetition	〳

Figure 2. Examples of symbols excluded from morphological analysis.

(9) Special characters used to represent character repetition (e.g., “〳”). For example, the second “ari” of the word “ari-ari” is written using this special character. We replaced the special character used for the second “ari” with the kana character corresponding to that orthography. (This special character was used in the old haiku [p.199] [see Figure 2]).

(10) Novelist Motojiro Kajii (p.254 of textbook) and Dr. Noguchi Hideyo's mother's letter (p.255 in the textbook) were deleted because they were represented using classical terms, not modern languages. The morphological analyzer used in the present study does not correspond to classical words. However, the modern translation of these two letters was included in the study.

(11) A list of the characters included in the table "Origin of Hiragana" and "Origin of Katakana" printed in the section "Characters used in Japan" (p.170) was excluded.

After the 11 criteria above were removed, the remainder was saved in an electronic file as a text document that contained 97,624 characters.

The morphological analyzer system

"Cha-mame ver.20.," which was developed by Dr. Ogiso, is morphological analysis software that was used in conjunction with UniDic-Mecab ver. 2.1.2. When analyzing numbers, the numbers shown in half-width characters were converted to full-width characters. In addition, we did not use "NumTrans," which is a function for specifying digit.

Results and Discussion

As a result of the morphological analysis on the text file composed of 97,624 characters that were extracted using the procedure described above; the total number of rows, that is, the number of morphemes, came to 64,784 words. However, since it contained blank lines (3,700 lines) and supplementary symbols (9,020 lines) such as parentheses, those 12,720 were deleted and the remaining 52,064 words were extracted. For these 52,064 words, the following three points were analyzed.

First, there were 32 unknown words that were not registered in UniDic. There were 5 types of these. (1) 13 words were proper nouns/names of people (e.g., a girl's name "マジャミン" '*Majamin*'), place names (e.g., "モリーオ" '*Moriio*'). (2) Three words were dictionary entries that have no description in that script (e.g., "ザトウクジラ" '*humpback whale*'). (3) Twelve words were furigana characters for some kanji characters at the end of each unit in the textbook (e.g., "翌" is pronounced "ヨク" /yoku/). (4) Three words were not registered in the UniDic (e.g., "Genji," a title of classical literature in the Heian period of Japan). (5) One additional word was an onomatopoeia that is not registered in the UniDic (e.g., "ゲロロッ," pronounced /geroro/, representing a sound of frogs croaking). For these unknown words, it was necessary to reassign POS by hand.

Second, when we counted the number of different words at the written form level, we found 6,266 words. Some words have different forms for the same entry, such as cases where the conjugation forms of verbs are different or where a word was written in different script in spite of being the same word. For example, the verb "行く," meaning '*go*' differs in the written form and there are 17 patterns of conjugation forms (for example, "行っ," "行き," "ゆく," "行か," and "行け.") When counting word printed frequencies (token frequencies), traditionally each lemma is regarded as an independent SUW and is counted as a separate entry word. For this reason, words having different written forms were listed in the frequency norm as different headwords, based on the word frequency norms by the National Language Research Institute (1970). However, even if the conjugation forms are different, the word is the same word. Therefore, in this study, we also counted the token frequency of words at the lemma level as follows.

When we counted the number of different words at the lemma level, it came to 5,002 words. The most frequent word was the case marking particle "の" (2,774 words), followed by "を" (2,194 words). Because these are functional words, they do not themselves hold meaning. Therefore, in this study, we created a table that extracts 20 words in descending order of frequency of occurrence, limited to nouns, verbs, and adjectives as content words commonly used in language research that use words as materials. Table 1 lists these 20 words in descending order of token frequencies for those three types of POS.

Looking at the top 4 words of common nouns, the most frequent "事" (meaning *things*) shows up 515 followed by "自分" (meaning *self*) at 204 times, and "言葉" (meaning *words*) at 174 times. The word "事" is

Table 1. A list of the top 20 word frequencies in basic content words for 3 types of part of speech

	Adjectives			Verbs			Nouns		
	Japanese word	English translation	Word fequency	Japanese word	English translation	Word fequency	Japanese word	English translation	Word fequency
1	楽しい	pleasant	32	言う	say	271	事	thing	515
2	大きい	large	32	書く	write	193	自分	self	204
3	新しい	new	28	考える	think	190	言葉	word	174
4	長い	long	27	読む	read	167	考え	idea	131
5	多い	many	21	思う	suppose	115	人	person	129
6	美しい	beautiful	17	つく[1)]	about	97	本	book	109
7	深い	deep	16	使う	use	97	父	father	96
8	早い	early	16	感ずる	feel	91	絵	painting	76
9	白い	white	16	作る	make	87	漢字	Kanji	76
10	悪い	bad	14	伝える	convey	87	人物	person	74
11	黒い	black	14	聞く	listen	82	山伏	mountain priest	68
12	詳しい	detailed	12	分かる	turn out	60	世界	world	66
13	強い	strong	11	表わす	represent	59	筆者	writer	64
14	高い	high	11	生きる	live	58	文章	writing	64
15	旨い	delicious	10	因る	cause	55	物語	story	61
16	少ない	few	10	持つ	have	55	次	next	57
17	青い	blue	10	纏める	to put together	52	宇宙	space	53
18	面白い	amusing	10	話す	speak	52	文	sencence	52
19	難しい	difficult	9	捕らえる	capture	47	森	forest	50
20	美味しい	delicious	9	取る	take	46	気	life energy	47

Note. [1)] This represents a combination of conjunctive form "て" attached to the stem form of the conjunctive particle of the verb "つく."

often used as the infinitive “to do,” “to think,” “to feel,” and so on. Next, “自分” is frequently used, but not because it is part of the texts of a story or explanations, but because it is a section of learning in the textbook that emphasizes “listening, speaking, writing and reading” based on stories and explanatory texts read by children in many cases. It is often used as pointing to the children themselves who are learners of "their ideas" and "their opinions". Likewise, “言葉” and “考え” (meaning *idea*) is also a word frequently used in the learning section. Also in the results of verbs, the words “言う” (meaning *say*), “書く” (meaning *write*), “考える” (meaning *think*), and “読む” (meaning *read*) are words that are also frequently used in the learning section. On the other hand, in the results of adjectives, there is no tendency to bias to the learning section, and words frequently used in various units and sentences are located in the higher rank.

In this article, only the top 20 nouns, verbs, and adjectives are shown as a table and the 21st and below are omitted (we plan to release them on the Internet in the future.). In addition, we omitted the different written word forms from this article. In the future, comparative studies on the analysis of word counts and word frequencies in lemmas and word forms are necessary. Furthermore, it is necessary to carry out additional morphological analysis for fifth grade and under textbooks, and to clarify the influence of the change of learning words from hiragana to kanji scripts over time from the first grade to the sixth grade.

Acknowledgments

This research was supported by Grants-in-Aid for Scientific Research form the Japan Society for the Promotion of Science (16K04434).

References

Den, Y., Oghiso, T., Ogura, H., Yamada, A., Minematsu, N., Uchimoto, K., & Koiso, H. (2007). *Koopus nihongogaku no tame no gengo shigen: Keitaisokaisekiyo denshijisho no kaihatsu to ouyo* [The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics]. *Nihongo Kagaku* [Japanese Linguistics], *22*, 101-122.

Fujita, S., Taira, H., Kobayashi, T., & Tanaka, T. (2014). Japanese morphological analysis of picture books. *Journal of Natural Language Processing*, *21*, 516-539.

Halliday, M. A. K., Teubert, W., Yallop, C., & Čermáková, A. (2004). *Lexicology and Corpus Linguistics: An Introduction.* London: Continuum.

Ishikawa, S. (2012). *A basic guide to corpus linguistic*. Tokyo: Hitsuji Shobo.

Joyce, T., Hodošček, B., & Nishina, K. (2012). Orthographic representation and variation within the Japanese writing system. *Written Language and Literacy*, *15*, 254-278.

Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying conditional random field to Japanese morphological analysis. *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp.230-237.

Maekawa, K. (2013). *Koopasu no sonzai igi.* [Significance of corpus existence]. In K. Maekawa (Ed.), *Kouza Nihongo Koopasu: Koopasu Nyuumon.* [Series of Japanese corpus, Vol.1, Introduction to corpus], (pp.1-31). Tokyo: Asakura Shoten.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Koshino, W., Koiso, H., Yamaguchi, M., Tanaka, M., & Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, *48*, 345-371. doi: 10.1007/s10579-013-9261-0

Meyer, C. F. (2002). *English corpus linguistics: an introduction*. Cambridge: Cambridge University Press.

Mitsumura Tosho (2016). *Kokugo Roku Souzou* [Japanese language textbook for sixth-grade elementary school]. Tokyo: Mitsumura Tosho Publishing.

National Language Research Institute (1970). *Denshikeisanki niyoru shinbun no goi-chosa* [Studies on the vocabulary of modern newspapers (Volume 1), General descriptions and vocabulary frequency tables]. Tokyo: Shuei Shuppan.

Ogiso, N. (2014). *Keitaisokaiseki.* [Morpholocical analysis]. In M. Yamazaki (Ed.), *Kouza Nihongo Koopasu: Kakikotoba Koopasu: Sekkei to Koutsiku.* [Series of Japanese corpus, Vol.2, Written language corpus: Design and construction]. (pp.89-115). Tokyo: Asakura Shoten.

Ogura, H., Koiso, H., Fujiike, Y., Miyauchi, S., Konishi, H., & Hara, Y. (2011). *"Balanced Corpus of Contemporary Written Japanese" Morphological information regulations collection 4th edition*. National Institute for Japanese Language and Linguistics report, LR-CCG-10-05-01, 02.

小学校六年生の国語教科書を対象とした形態素解析

小河妙子・藤田知加子
東海学院大学・南山大学

要　約

本研究の目的は，小学校国語教科書に掲載されている文章に含まれる単語を網羅的に抽出し，本邦の国語教育において教材とされている単語の特徴を調査することにある。教科書に掲載されている文章から単語を抽出するために，形態素解析を実施した。その結果，小学校六年生の教科書に掲載されているすべての印刷された文字から，5,002 語の語彙素が抽出された。これらの単語を対象として品詞ごとに語数を数え，単語出現頻度を報告した。国語教科書に掲載されている文章を対象とした形態素解析に関する問題点が論じられた。

キーワード：形態素解析，国語教科書，コーパス言語学