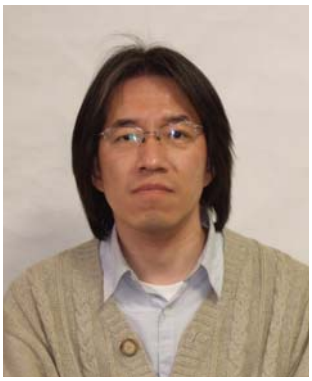
Humanities and Social Sciences

Challenges in Japanese Corpus Linguistics



Professor, Faculty of Letters, Division of Behavioral Science

Yasuharu Den

Background of Research

What kinds of words do we normally encounter, and how often? What are the differences in writing style between a “hard” document like a government white paper and a “soft” document like a blog? Finding answers to these sorts of questions will identify the real nature of our linguistic activity, and will also shed light on how the mind understands words and language.

This line of research requires a body of linguistic data (a *corpus*) that encompasses a variety of genres and literary styles. The study of the Japanese language using such large corpora is called *Japanese corpus linguistics*, and is a field in its infancy. As a driving project in this field, as Grant-in-Aid for Scientific Research for Priority Areas begun in 2006, a huge research group, led by the National Institute for Japanese Language and Linguistics, has worked on the development of a contemporary written Japanese corpus containing more than 100 million words. As part of this project I put together an “electronic dictionary” team, which has played a central role in the creation of this corpus.

Achievements of Research

As an illustration of our research, consider the sentence “さくらの木が1本ありました” (“*sakura-no ki-ga ippon ari-mashi-ta*”: “There was one cherry tree.”). What exactly are the individual words that make up this Japanese sentence? Since, unlike English, Japanese does not use spaces to separate individual words from each other, we cannot answer this question at a single glance. In order to study this, or any other, Japanese text, we will need to divide it into separate words in order to assign information such as parts of speech (See the chart below).

The technology that automatizes the performance of this task using a computer is called *morphological analysis*, and has been studied now for more than 30 years. Morphological analysis requires a dictionary that contains the words that can appear in any text and that also defines information such as what part of speech a word is. Our electronic dictionary was developed for such purpose. With the existing electronic dictionaries it was not possible to perform certain tasks, such as collating orthographic variants (for example, “桜”, “さくら”, and “サクラ” for the word “*sakura*” (“cherry”), written in the three Japanese writing systems of kanji, hiragana, and katakana), distinguishing between heteronyms (such as the pair of Japanese kanji characters “生物”, which can be read as either “*seibutsu*” or “*namamono*”), and changing the pronunciation of words as required from context (such as the pair of the Japanese numeral for “one” and the numeral classifier for a long tree-shaped object, “一本” , from the verbatim pronunciation “*ichihon*” to the standard pronunciation “*ippon*”). We have tackled these issues by employing a novel approach when designing our dictionary. The electronic dictionary that we have

developed as a free public resource contains more than 200,000 vocabulary entries, and has an accuracy, in automatic morphological analysis of texts in various genres, of more than 98.5% (with fewer than three words in 200 being incorrect). In addition to being used in the development of the Grant-in-Aid for Scientific Research for Priority Areas corpus and in Japanese corpus linguistics research such as vocabulary research and the analysis of genre characteristics, this dictionary is also finding commercial applications through free licensing agreements entered into with several companies including Apple Inc. in the U.S.

Prospective developments

The development of spoken Japanese corpora has been less explored compared to that of written Japanese corpora. Particularly for the study of conversations, which form the core of our daily linguistic activities, it becomes essential to gather and analyze real data on an enormous scale, since spoken conversations inherently have less occasion of being kept a record than written words do, and because they also often involve spontaneous uses of language such as disfluency and back channels. With the aim of expanding the scope and range of Japanese corpus linguistics, alongside our current Grant-in-Aid for Scientific Research for Priority Areas work, we have begun to investigate the development of a Japanese conversation corpus as a Grant-in-Aid for Scientific Research B project.

* * * * * * *

さくらの木が1本ありました (“*sakura-no ki-ga ippon ari-mashi-ta*”: “There was one cherry tree.”)

Orthography	Pronunciation	Dictionary heading	Part of speech	Conjugation type	Conjugation form
さくら	sakura	さくら【桜】(cherry)	Common noun		
の	no	の (GEN)	Case particle		
木	ki	き【木】(tree)	Common noun		
が	ga	が (NOM)	Case particle		
1	ip	いち【一】(one)	Numeral		
本	pon	ほん【本】(NC)	Nominal classifier		
あり	ari	ある【有る】(be)	Verb	Vowel-inflection	Adverbial
まし	mashi	ます (POL)	Auxiliary verb	Aux-masu	Adverbial
た	ta	た (PAST)	Auxiliary verb	Aux-ta	Predicative