New words in Japanese and the design of *UniDic* electronic dictionary

Teruaki Oka

Corpus Development Center, National Institute for Japanese Language and Linguistics

teruaki-oka@ninja.ac.jp

Abstract

The National Institute for Japanese Language and Linguistics (NINJAL) is involved in developing Japanese language corpora, including the Balanced Corpus of Contemporary Written Japanese, Corpus of Spontaneous Japanese, Corpus of Historical Japanese, and NINJAL Web Japanese Corpus. In the development processes we often encounter new words that are formed by using different character types (e.g., Hiragana, Katakana, Kanji) and their heterographs, with their combinations, even for writing a single word (e.g., *big*: おおきい, 大きい, オオキイ, ぉぉきぃ, 大キィ), which could be 'literal' (e.g., *as it was expected*: 矢張り), 'somewhat colloquial' (やっぱり), 'colloquial' (やっぱし), 'abbreviated' (やぱ), and so on. Thus, new words can appear as orthographic variants (おおきい vs. 大キィ), form variants (矢張り vs. やぱ) and new lemmas (such as エモい*emotional*), and be classified at these three levels (orthographic, form, lemma).

We apply a design policy called "hierarchical definition of word indexes" to register new words in *UniDic*, our electronic Japanese word dictionary for annotating plain texts with morphological information. Using the hierarchical definition of word indexes, a single lemma (e.g., 矢張り) has its various word forms written in Katakana characters (e.g., 矢張り←ヤハリ, ヤッパリ, ヤッパシ, ヤパ) as its children, with each form having its orthographic variants as its children (e.g., ヤハリ←矢張り, やはり, ヤハリ). *UniDic* contains about 200 thousand lemmas

and one million of their form and orthographic variants with rich morphological information (e.g., part of speech, lemmatized form, pronunciation, accent). To annotate morphological information in plain unsegmented texts, we select optimal records for character strings in the texts from UniDicDB, a word database system. The records and their morphological information are manually registered to UniDicDB when new words are detected during the annotation phase. We also employ UniDicExplorer, an annotator-friendly user interface capable of searching and registering words. Another feature is UniDicMA, a dictionary software for the morphological analyzer, which is derived from UniDicDB and can attach the hierarchical structure of *UniDic* to each word in an input plain unsegmented text automatically (https://unidic.ninjal.ac.jp/). Only UniDicMA is open to the public, whereas all other UniDics are not accessible outside NINJAL.

In this paper, we will discuss what is a 'new word' in Japanese, our hierarchical definition of word indexes, and how to register new words in UniDicDB using UniDicExplorer.

Keywords

electronic dictionary, Japanese, corpus, annotation, database system, morphological analyzer, neologisms