# Construction of the "Corpus of Historical Japanese: Meiji–Taishō Series I – Magazines"

**Toshinobu Ogiso**
togiso@ninjal.ac.jp
National Institute for Japanese Language and Linguistics, Japan

**Asuko Kondo**
kondo@ninjal.ac.jp
National Institute for Japanese Language and Linguistics, Japan

**Yoko Mabuchi**
mabuchi@meiji.ac.jp
Meiji University, Japan

**Noriko Hattori**
nhattori@ninjal.ac.jp
National Institute for Japanese Language and Linguistics, Japan

In this talk, we wish to discuss the construction of the corpus "Meiji-Taishō Series I - Magazines," (hereinafter called CHJ-Magazines) which was released as a part of the Corpus of Historical Japanese (CHJ). This corpus contains magazines published in Japan in the Meiji and Taishō periods (1868–1911) representative of Modern Japanese language, with the total number of words used in the text reaching a size of some 14,000,000 items. This corpus is the first large-scale corpus with morphological information of Modern Japanese and will contribute to research into that period of the Japanese language.

At the National Institute for Japanese Language and Linguistics, a joint research project entitled "Construction of a Diachronic Corpus and New Developments in Research of the Japanese Language" was begun in 2016. CHJ Magazines is one part of this project, using a number of representative magazines for each of several periods, with the selected published material spread out over set time intervals. The result is a large-scale corpus that makes it possible to examine the state of the written Japanese language in the Meiji and Taishō periods, as well as to examine how the language underwent changes in those periods (Table 1).

| | 1870's | 1880's | 1890's | 1900's | 1910's | 1920's | *Total* |
|---|---|---|---|---|---|---|---|
| "Meiroku Zasshi" | 180 | | | | | | *180* |
| "Kokumin no Tomo" | | 1,000 | | | | | *1,000* |
| "Taiyo" | | | 2,280 | 4,300 | 2,050 | 2,300 | *10,930* |
| "Iogaku Zasshi" | | | 680 | | | | *1,890* |
| "Iogaku Sekai" | | | | 590 | | | |
| "Fujin Kurabu" | | | | | | 620 | |
| *Total* | *180* | *1,000* | *2,960* | *4,890* | *2,050* | *2,920* | *14,000* |

Table 1. Corpus Size for Each Year of Publication (Units: Thousand Words)

One characteristic of this corpus is that morphological (word) information is annotated to each text. In order to annotate highly accurate morphological information to the corpus, it was necessary to develop and utilize a dictionary that could enable automatic morphological analysis for the colloquial and literary styles that was mixed in the written Japanese of the aforementioned periods. We customized UniDic, a dictionary for morphological analysis of Japanese, which can lemmatize variations of orthography and word forms. We also used MeCab, a state-of-the-art Japanese morphological analyzer with this customized UniDic.

While the accuracy of automatic analysis is high at approximately 93%, it is extremely difficult to append morphological information uniformly without any mistakes in such a large amount of text. To address this difficulty, it was necessary to establish a "core" data set, for which we could guarantee the attachment of highly accurate morphological information (with accuracy higher than 99%) through the utilization of automatic computer analysis together with manual (human) correction. For core data, 500,000 words were selected, taking into consideration the balance of publication year, style and genre of each of the articles sampled. Core data is suitable for Japanese research which requires highly accurate morphological information. This data was also used as training data for the dictionary for morphological analysis described above. The automatic analysis results of the "non-core" data set, meanwhile, received a certain degree of manual correction which, while not being exhaustive, strikes a balance between quality and volume.

This corpus is made publicly accessible by way of an online search application called "Chūnagon." With "Chūnagon," it is possible to carry out searches that specify complex combinations of different morphological information (lemma, part-of-speech,

conjugation type, lexical strata, etc.). The search results also display information on the source (magazine title, article title, year of publishing), author (name, sex, year of birth), data type (core, non-core), text type (conversation, quotation, other), text style (literary, colloquial, mixture, verse), etc. The author information includes a link to catalogues in the National Diet Library. By making use of this convenient service, it is possible to check how to read the name of the author of the article in question, to check when he or she was born and died, and, at the same time, to see if he or she made use of multiple other names. The "Chūnagon" search results also provide links that allow the viewing of images of corresponding pages in the original magazines.

## Bibliography


**Center, National Institute for Japanese Language and Linguistics** (2014 - 2016) *Corpus of Historical Japanese*. http://pj.ninjal.ac.jp/corpus_center/chj/overview-en.html

**Ogiso, T.** (2014), "Dictionary and Morphological Analysis in the Historical Corpus." *Nihongogaku*, 432, pp. 83-95. Meiji Shoin. (In Japanese)

**Mabuchi, Y., and Asuko, K.** (2016), "Explanatory notes of symbols in the text, Display categories for Chunagon: Corpus of Historical Japanese, Meiji – Taisho Series I – Magazines (Short Unit Ver. 0.9)". http://pj.ninjal.ac.jp/corpus_center/chj/doc/abstract-meiji-taisho-2016.pdf (In Japanese)