Committee on Japanese Materials Meeting, Thursday March 26, Chicago.

Michiko Ito, Chair of CJM, welcomed participants and introduced the committee members and their activities, such as officially launching JpnLibLiaison listserv and updating CJM website. She also briefly reported on the activities of the Committee and the Subcommittee on Japanese Rare Books and the CTP/CJM Joint Working Group on the Japanese Romanization Table. She then introduced the speakers, Dr. Toshinobu Ogiso, Department of Corpus Studies, National Institute for Japanese Language and Linguistics (NINJAL) and Dr. Hoyt Long, Associate Professor of Japanese Literature, East Asian Languages and Civilizations, University of Chicago.

Japanese Corpora by NINJAL, by Dr. Toshinobu Ogiso, Department of Corpus Studies, NINJAL

NINJAL was founded in 1948 as the National Language Research Institute (NLRI) and reorganized as an "Inter-University Research Institute" in 2009. Dr. Ogiso introduced some of the databases created by NINJAL.

1) Balanced Corpus of Contemporary Written Japanese (BCCWJ)

Launched in 2009, this corpus includes 100 million words drawn from a wide range of source texts, all of them copyright cleared. It is morphologically annotated with word-segmentation and POS-tagging provided by a new original machine-readable dictionary, UniDic, and the morphological analyzer MeCab. The accuracy of morphological analysis is as high as 98%.

There are three methods of accessing these corpora:

- Shonagon: Online free access without registration
- Chunagon: Online free access with registration
- Himawari: DVD release (charged service)

2) Corpus of modern magazines

Consists of the following components:

- Taiyō corpus (DVD package). Texts of 3,409 articles in 60 issues of the magazine *Taiyō* 太陽, published over the period of 1895-1925. Approximately 9 million words.
- Corpus of modern women's magazines (free download).
- Meiroku Zasshi Corpus (free download). Full-text and morphologically annotated corpus of *Meiroku zasshi* 明六雑誌, Japan's first modern magazine, published in 1874 and 1875. Approximately 180,000 words.
- Kokumin-no-Tomo Corpus『國民之友』(free download).

3) Corpus of Historical Japanese

Currently under development. It contains material covering the 8th to 20th centuries and is divided into two series:

- Heian period series: 14 texts including Genji monogatari, Makura no soshi.
- Muromachi period series : Kyōgen

Dr. Ogiso gave two examples of how the corpora could be used for research purposes.
Case 1: Orthography of shōgai
It is sometimes said that the word *shōgai* (difficulty, handicap) was written as 障碍 before 1946, but was subsequently written as 障害 following the introduction of *Tōyō kanji* in that year.
Analysis of texts in the Taiyo and BCCWJ corpora revealed that the characters 障害 were also widely used in the Meiji period.

Case 2: Usage of *fujin* 婦人 and *josei* 女性 (woman)
Dr. Ogiso provided results of a number of searches using the Taiyo and BCCWJ corpora to test the argument that *fujin* was more widely used in the Meiji and Taisho periods than *josei*. For example, a search of the Taiyo Corpus revealed:

1895: *josei* = 19 hits, fujin =241 hits
1925: *josei* = 299 hits, fujin =350 hits
In modern texts *josei* is used in 90% of cases.

Dr. Ogiso explained that the corpora could be used across a range of disciplines

1) <u>Linguistics</u> – increasing trend to corpus-based analysis.
2) <u>Lexicography</u> – dictionaries can be built based on corpora. E.g. Collins Cobuild Dictionary of English and *A Frequency Dictionary of Japanese: Core Vocabulary for Learners* (Routledge, 2013), the world's first corpus-based frequency dictionary of Japanese.
3) <u>Language education</u> – corpora are being used for teaching materials, dictionaries, and e-learning tools.
4) <u>Language policy</u> - BCCWJ was one of the reference materials used for the 2010 revision of *Jōyō kanji*.
5) <u>Study of Classical literature</u> – the Corpus of Historical Japanese is a useful resource for the study of Japanese classical literature.
6) <u>Information technology</u> – e.g. the Japanese Input Method program of iOS (iPhone, iPad) uses statistical information and dictionary entries from BCCWJ

Dr. Ogiso concluded his presentation with the view that free access corpora will remove barriers to the study of Japanese language and culture, especially for foreign scholars, and urged the audience to make the existence of NINJAL 's corpora more widely known to scholars and students.

Digital Humanities in Japanese Studies, by Dr. Hoyt Long, University of Chicago
Dr. Long began by posing the question, "What are digital humanists doing?" He summarized their activities as:

1) Findings way to archive and better curate digital collections – building image collections and text collections, creating databases with linked metadata, and developing new presentation platforms such as Omeka and Neatline.
2) Developing methods for small/large-scale data analysis – e.g. text mining, network analysis, GIS image analysis.

3) Visualizing humanities data to facilitate data exploration – e.g. tools for search and pattern detection such as Voyant and Palladio.
4) In Japanese studies much effort is currently devoted to 1) but gradually shifting to 2) and 3).

Dr. Long then described how he believed librarians could aid digital scholarship. They should:

1) Learn about and help disseminate tools – DH Centers and Labs, DH conferences (especially the Japanese Association for *Digital Humanities,* or JADH), LibGuides, online tutorials, and digital collections.
2) Work to enhance access to existing collections and make them available for data analysis – e.g Aozora Bunko, NINJAL corpus, HathiTrust, e-books, and electronic databases.
3) Create workflow for building local digital collections and transforming print resources into dynamic archives – e.g. by scanning, OCR, data hosting, online access, copyright issues – think of digital collections as resources for viewing and content analysis.

Dr. Long demonstrated an example of transforming print data into a database. He purchased data from Nichigai Associates *General index of modern Japanese poetry: 1920-1944* 現代詩 1920-1944—モダニズム詩誌作品要覧, which contains over 100,000 poems by 4,000 poets taken from 166 journals, and converted it into an SQL database. Using this dataset, Dr. Long began a research project focused on translated content within Japanese modernist journals (results of this project to appear as "Fog and Steel: Mapping Communities of Literary Translation in an Information Age" (Journal of Japanese Studies, Summer 2015).

The data could also be used in a variety of ways, such as basic statistical analysis of works translated into Japanese by year to establish larger historical context and general trends, finer-grained statistical analysis using richer categories of metadata such as region of origin of the translated works, and complex network analysis to explore patterns of relation within the data (using Excel and Gephi).

Dr. Long described various kinds of analysis that could be carried out with literary text data. He gave examples of how it was possible to obtain basic linguistics measures using corpora of digitized texts and to make simple comparisons of lexical difference using digitized texts and online tutorials. He also demonstrated how he had assembled a corpus of English-language Haiku.

Dr. Long concluded by considering future prospects for the study of Japanese literature. In the short term, researchers would use existing collections like Aozora, which contains over 12,000 works, only 1,000 of which are novels, and would need to build smaller sub-corpora of pre-processed texts.

In the long term, they should their build own mini-collections geared to individual research interests or large-scale projects. This would involve scanning, OCR, and pre-processing texts to make them available for analysis of the types he had described.