**Hodošček — Visualizing**

<0 images>

**Visualizing Japanese Language Change during the Past Century**

Hodošček, B.

This study introduces an online system for the visualization and analysis of over a century (1874–2008) of Japanese language change. A comprehensive account of register variation in contemporary Japanese has recently become possible with the public availability of the Balanced Corpus of Contemporary Written Japanese (BCCWJ), a 100-million-word corpus that contains a wide variety of written Japanese collected and curated by the National Institute for Japanese Language and Linguistics. Increasingly, too, public releases of new corpora that record various genres of modern (Meiji-era) Japanese writing have paved the way to enabling more comprehensive diachronic analysis (analysis of language development and evolution through time) of Japanese. Still, especially compared to recent efforts in English, which include large book digitalization projects such as the Google Books corpus (Michel et al., 2011), as well as more curated historical corpora such as the Corpus of Historical American English (COHA) or the register-balanced Corpus of Contemporary American English (COCA) (Davies, 2010; 2011), the available resources and research tools for investigating diachronic language change as well as register variation in Japanese lack in two respects: balanced representation of registers throughout time, and unified and sophisticated search interfaces.

We combine the use of corpus metadata and annotations with textual features to model language change through time and between different registers from the following six corpora:

• The Balanced Corpus of Contemporary Written Japanese (c. 1975–2008).

• The Sun corpus (c. 1895–1925).

• The Meiroku Zasshi corpus (c. 1874–1875).

• The Kindai Josei Zasshi corpus (c. 1894–1925).

• The Kokumin no Tomo corpus (c. 1887–1888).

• A subset of the Aozora Bunko (c. 1890s–).

All text is first converted into a unified structured format that includes structural information (paragraphs, headings, titles, lists, etc.) as well as other information (spoken text, quotations, etc.), where available, from the different textual or XML encodings of the corpora. Next, we process sentences into morpheme tokens using the morphological analyzer MeCab and, depending on the time period, the modern or contemporary version of the UniDic morphological dictionary. A unique property of both variants of UniDic is their organization of word tokens under lemma that cover the many orthographic variants observed in Japanese writing. Taking the basic lemma, word orthography, and POS triplets as a base, we construct co-occurrence networks between all words occurring in the same sentence or paragraph. This co-occurrence network is constructed so that we are able to generate sub-networks that match some metadata query, such as year and NDC code, which can then be used to compare with other sub-networks. The query and visualization interface thus provides a timeline for choosing specific time-related subsets from the corpora, as well as a visual way of selecting from other metadata, including categorical (gender, media type, etc.) and hierarchical (NDC, topic, etc.) information, which allows the user to further constrain the scope of investigation into language change to some register within a chosen time period or to instead focus on the differences between registers by comparing between two or more different registers within a set time period.

**References**

**Davies, M.** (2010). The Corpus of Historical American English: 400 Million Words, 1810–2009. http://corpus.byu.edu/coha/ (accessed 1 November 2014).

**Davies, M.** (2011). Google Books (American English) Corpus (155 Billion Words, 1810–2009). http://googlebooks.byu.edu/ (accessed 1 November 2014).

**Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., . . . Orwant, J., et al.** (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science,* **331**(6014): 176–82.