1 Application Detail

1. The type of presentation (poster, short paper, long paper or panel)

   POSTER

2. A title

   Design of Serial Comparison Model for the Diachronic Corpus Study of Japanese

3. A list of keywords (up to five)

   Corpus Linguistics, Diachronic Language Study, Comparative Study,

4. The name, status and affiliation of the presenter (s)

   - Hilofumi Yamamoto, Tokyo Institute of Technology / University of California, San Diego
   - Makiro Tanaka, National Institute of Japanese Language and Linguistics
   - Yasu-Hiro Kondo, Aoyama Gakuin University / National Institute of Japanese Language and Linguistics

5. A contact email address

   - yamagen@ryu.titech.ac.jp
   - mtanaka@ninjal.ac.jp
   - yhkondo@cl.aoyama.ac.jp

6. A postal address

   - TokyoTech: 2-12-1 O-okayama, Meguro-ku, Tokyo, 152-8550, Japan
   - NINJAL: 10-2 Midori-cho, Tachikawa City, Tokyo, 190-8561, Japan
   - AOYAMA-GU: 4-4-25, Shibuya, Shibuya-ku, Tokyo, 150-8366, Japan

2 A biography up to 100 words

Hilofumi Yamamoto is an associate professor at Tokyo Institute of Technology. He earned a Ph. D. in Linguistcs at the Australian National University. He is currently working on linguistic change and complexity.

Makiro Tanaka is an associate professor at the National Institute for Japanese Language and Linguistics. He is examining the history of Japanese vocabulary in the Heian period in terms of the theory of language layers.

Yasu-Hiro Kondo is a professor at Aoyama Gakuin University. He graduated from the University of Tokyo where he earned a Ph. D. in Literature. His contributions to Japanese syntax studies brought him the Dr. Kindaiichi Prize in 2000.

3 Abstract

The project "the design and development of a diachronic corpus of Japanese" began in 2009 at the Department of Corpus Study, the National Institute of Japanese Language and Linguistics (NINJAL) as a collaborative research project by linguists and scholars of literature from NINJAL and the University of Oxford. Its focus is on collecting representative Japanese literary works and classical documents from the tenth century to the nineteenth century.

We are currently working on the development of a prototype version of the diachronic Japanese corpus: i.e. we are focusing on the selection of materials, digitizing texts, adding alternative texts (different orthography) to original texts, compiling a basic thesaurus that differentiates between different spellings, and word segmentation.

This paper addresses the discussion of the basic concepts encountered during our work on the project: synchronic and diachronic analysis, which led us to the design of a serial comparison model which allows us to examine language change between documents or literary works with respect to time.

Saussure (1983) defines diachronic change as a language feature transforming according to time, and synchronic as a language feature at a certain moment in time. In either event it is necessary to compare any two types of texts to examine differences in terms of diachronic or synchronic change. What we examine is not differences in content but those in language use. Therefore we must extract, in a clear way, only the differences in language use between the target texts.

Many studies have tried to examine differences in language use. However, in almost every case, examples or clues of language change were collected not by computers but by time-consuming manual work. The methods they employed are thus impractical for the collection of a large number of examples or clues, as is often the case in corpus studies. Furthermore, there is a risk that utilizing such methods could lead researchers into only collecting sentences conforming with their hypotheses. Therefore, it is necessary to use computers and corpora for conducting more exact and unbiased research than was ever achieved or possible before. This, however, requires us to prepare a framework that can clearly differentiate ways of language use and contents of texts.

To this end, we developed a "serial comparison model" which allows us to examine evidence of language change by searching in corpora, and classify vocabulary items into differences in content or differences in language use.

The finding of this study is that the serial comparison model allows us to 1) discern language change between any two points in time, and 2) classify vocabulary items into five categories: unchanged, similar but not used in the new text, similar but not used in the old text, newly used, and abandoned. When comparing two texts, the model extracts differences in language change from differences in content. (466 words)