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

Hilofumi Yamamoto Tokyo Institute of Technology

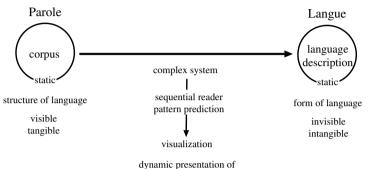
Tokyo Institute of Technology

National Institute for Japanese Language and Linguistics, Japan

Development of Diachronic Corpus

Project by the National Institute for Japanese Language and Linguistics, Japan, NINJAL: 2009–13, 4 year project.

Main purpose: Study of Japanese language (sub) purpose: Study of Japanese (classic) literature



diachronic language change

Figure 1: Corpus and Description, Langue and Parole: The nature of language is dynamic and always changing while the phenomena of language might be static. We should consider the dynamic change of language as a component comprised of various elements. The feature of language we usually observe is a complex system and tangled with wide-ranging elements.

Contents of Diachronic Corpus

- 1. The Tale of the Bamboo-Cutter
- (ca. 890; Taketori monogatari; 12,583 tokens) 2. Tales of Ise
- (ca. 901; Ise monogatari; 15,900 tokens)
- 3. Tales of Yamato
- (ca. 950; Yamato monogatari; 26,733 tokens) 4. The Tosa Diary
- (ca. 935; Tosa nikki; 8,113 tokens)
- 5. The Pillow Book
- (ca. 996; Makura no sōshi; 79,861 tokens) 6. Tale of Genji
- (ca. 1100; Genji monogatari; 510,711 tokens)

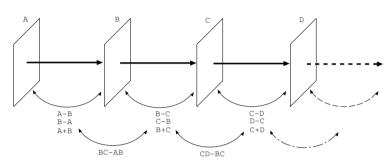


Figure 2: Extraction of delta from each synchronic layer: A, B, C and D are arbitrarily-assigned synchronic layers on the time axis. Examination of linguistic transitions is achieved through the comparison of lexical items in each layer with those in other layers, and the discovery of common principles appearing in the delta of data extracted from both systems as well.

A case study: use of SAKURA(cherry blossoms) in Mt. Yoshino \rightarrow Kokinshū (ca. 905) vs Shinkokinshū (1205)

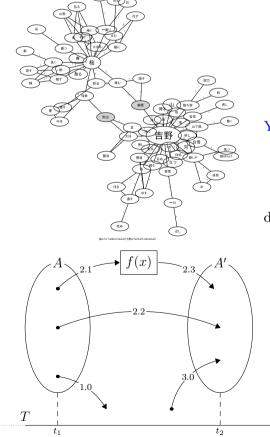


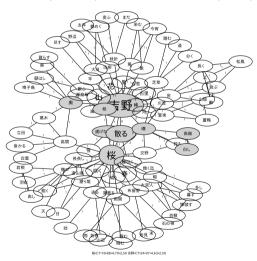
Figure 3: Serial comparison model; differential model of transitional linguistic elements of target texts; A is a set of elements that occurred at Time t_1 ; A' is a set of elements that occurred at Time t_2 ; T is the time axis; f(x) is a function for converting an element x of A into that of A'.



Sakura (桜) and Yoshino (吉野), a place name in Nara prefecture

← Kokinshū (ca. 905) Shinkokinshū (1205) →

during 300 years differences.



<u>Future Task</u>

- To define linguistic units suitable for each era
- To develop a dictionary for machine analysis
- \rightarrow it allows us syntagmatic and paradigmatic analyses

Conclusion

- Addressed basic concepts and framework of diachronic corpus
- Illustrated the serial comparison model for historical analysis
- \rightarrow Lexical differences between any two groups of texts.

