NINJAL-LWP for BCCWJ: A Lexical Profiling Based Browsing System

Prashant Pardeshi (National Institute for Japanese Language and Linguistics)
Shiro Akasegawa (Lago Institute of Language)

The National Institute for Japanese Language and Linguistics (NINJAL) and the Lago Institute of Language (LIL) have jointly developed an online corpus browsing system called NINJAL-LWP for BCCWJ (NLB). Its site (http://ninjal-lwp-bccwj.ninjal.ac.jp) went public in June, 2012 and has been accessed by researchers and teachers of Japanese language all over the world. In this presentation, we will report: (A) the salient features and functions of NLB and (B) two practical applications of NLB in the domain of linguistic analysis and compilation of a Japanese basic verb usage handbook.

(A) **The salient features and functions of NLB**: One of the main characteristics of this system is that it introduces the lexical profiling methodology, which was first proposed by K. Church and P. Hanks back in 1989, although they didn't use the expression "lexical profiling." Text processing for lexical profiling varies from one language to another. NLB, the first lexical profiling system for Japanese, deals with the complexity of the Japanese writing system. Japanese is usually written in three types of characters: hiragana, katakana and kanji. This means a word could be written in at least three ways. The word '人', which means *a person*, can be written as 'ひと' in hiragana or 'ヒト' in katakana with different connotations. In the case of compound verbs, things are more complicated due to the fact that some verbs have two or more kanji candidates with slightly different meanings. The compound verb '取り入れる', which means *to introduce,* can also be written as '採り入れる', resulting in more than eight ways of orthographic rendering. To handle the rich variation found in Japanese orthographic forms, NLB incorporates the idea of a representative orthographic form for the lexical unit, which is analogous to a headword in a dictionary.This is an indispensable feature for Japanese dictionary compilation.

(B) **Practical applications of NLB**: We will report two cases of practical application of NLB. The first one is related to the research of idiomatic expressions in Japanese involving the basic verb EAT (*kuu* or *kurau*) in expressions such as *katasukashi o kurau*, *homuran o kurau*, *panchi o kurau*, *koosuauto o kurau*, *awa o kuu, michikusa o kuu*, etc. Japanese has more than 150 such expressions and the repertoire of such EAT-expression is an open set allowing addition of new members. We will demonstrate how NLB proves to be effective in the collection of such expressions as well as in determining the degree of indiomaticity of the EAT-expressions. Further, we will also demonstrate how synchronic and diachronic research can be done using corpus data. The second application of NLB is related to the field of lexicography or dictionary compilation. In our talk we will report the progress of the collaborative research project entitled "Compilation of Japanese Basic Verb Usage Handbook for JFL Learners" currently being carried out at NINJAL.