Problems in TEI:P5 Encoding on Colloquial Japanese Documents of the Early Modern Period

The National Institute for Japanese Language and Linguistics (NINJAL) is conducting morphological analysis on Japanese classics.

Digitization has been done thus far on the literature of several ages and various text corpora are published. However, each element (tag) of the text corpora is marked up under NINJAL's Document Type Definition, which is basically neither unified nor standardized. Under this circumstance causes problem with structural analysis and numerical analyses between several corpora. Thus it is necessary to design and mark up a unified definition from a higher level in order to conduct analyses concurrently. In this study, we examine the possibilities to convert documents of classical Japanese, an old block book from Sharebon's "Keisei-kai futasuji-no-michi" (published in 1798) as a model case, with TEI-compliant XML and discuss its issues.