Constructing and Using the NINJAL Parsed Corpus of Modern Japanese

Stephen Wright HORN (Adjunct Researcher, National Institute for Japanese Language and Linguistics)

The NINJAL Parsed Corpus of Modern Japanese under development at the National Institute for Japanese Language and Linguistics is a treebank of Japanese sentences parsed according to syntactic principles in the Penn Historical Treebank format.

This presentation describes the production pipeline that generates the bracketed tree structures, and the syntactic principles that govern the process.

Text is passed through a series of morphological parsers which segment it and annotate it with information.

Annotation is rewritten into terminal node labels.

Structure is built up from segment-label pairs to help correct the node label assignments by reference to context.

Corrected string-node pairings are then passed to a statistical parser.

The resulting trees are then hand-annotated.

An initial version of the corpus has been made publicly available, with easy-to-use interfaces.

The presentation includes a demonstration of 1) a hand-annotation tool, and 2) the on-line user interfaces.