Treebank meets descriptive grammar: The NPCMJ Explorer

Introductoin This paper reports on the development of a novice-friendly interface called 'NPCMJ Explorer' for the NINJAL Parsed Corpus of Modern Japanese (NPCMJ), a phrase structure-based treebank for Japanese currently (2016–2022) being developed at NINJAL (containing 20k (at present) and 50k sentences (by 2022)). In developing this interface, we encountered several issues which seemed to have larger implications not limited to either Japanese or treebanks, and which seemed to merit discussion in a wider context of corpus development and linguistic research.

Background NPCMJ is a successor of the Keyaki Treebank (Butler et al. 2017), whose primary purpose was to construct a resource that would serve as a model for a parser that outputs explicit predicate logic-based semantic representations. In addition to this design feature inherited from its precursor, NPCMJ is guided by its another major goal: developing a treebank that can be used as a useful research tool for ordinary linguists (i.e. not just NLP/computational linguistics researchers).

Motivation and specification The NPCMJ Explorer was developed precisely for this latter purpose of wider dissemination. It enables users without special knowledge of treebanks to quickly search for major grammatical phenomena in Japanese through a web browser-based GUI interface. Moreover, users can familiarize themselves with the annotation scheme by examining search expressions and search results. The search expressions were prepared by a specialist of Japanese linguistics familiar with the annotation scheme. Items were taken from the set of grammatical phenomena listed as section headers of Masuoka and Takubo (1992), a reputable descriptive grammar book on Japanese.

Result The design feature of NPCMJ had both advantages and disadvantages in NPCMJ Explorer development. Of the 136 grammatical phenomena listed in Masuoka and Takubo (1992), we were able to write adequate search expressions for 74 (i.e. about half). Phenomena for which the NPCMJ annotation scheme proved advantageous include the distinctions between 'direct' and 'indirect' passives and 'gapped' and 'gapless' relative clauses. Many of these were cases involving syntactically similar structures with distinct truth-conditional meanings. By contrast, phenomena involving finer-grained meaning differences (going beyond truth-conditional differences) turned out to pose major difficulties. These include the distinction between 'thematic' and 'contrastive' *wa* and phenomena involving focus-sensitive particles (*toritate-shi* in descriptive grammar). For these latter cases, we have either dealt with them by approximating a search with a manually prepared list of lexical items (*toritate-shi* were dealt with this way), or otherwise have given up on writing a search query.

Discussion One of the major challenges that face corpora with annotation of fine-grained information (such as treebanks) is the fact that their annotation policies tend to be complex, being guided by multiple (not necessarily mutually compatible) underlying principles. We have learned from our experience of developing the NPCMJ Explorer that NPCMJ was no exception. There is an implicit incongruence (in terms of what should be prioritized) between the design feature inherited from its precursor Keyaki Treebank and the goal to make the treebank a maximally useful resource for general linguists. The gap that consequently exists between what ordinary linguists take to be 'grammar' (represented by Takubo and Masuoka (1992)) and what is explicitly annotated in (or is otherwise easily retrievable from) the current form of NPCMJ was unexpectedly large.

Concluding remarks We believe that the difficulty we faced was a version of dilemma that is universal in corpus development and corpus use, and that, for that reason, sharing our experience with the wider community would be useful. The development of the NPCMJ Explorer taught us many things which we would otherwise have never learned. One definitely positive result is that we now know more precisely what to tell to linguists when we advertise NPCMJ as resource for doing linguistic research, by sharing our experience of what turned out to be 'easy' and 'difficult' cases in formulating queries. Another benefit is that we are now much clearer than before about what kinds of linguistic distinctions (of potential interest to linguists) are still lacking in the NPCMJ annotation (and why), which is essential information for the future development of the corpus. Finally, we think that the very fact that the gap between treebank and grammar was unexpectedly large is itself highly illuminating. This means that even a relatively fine-grained treebank which gives sufficient information to capture (at least the core aspects of) truth conditional meanings still glosses over many properties of language that linguists have identified to be important. This reminds us of the fact that language is an extremely intricate system, a fact which we of course are all aware of but which we perhaps tend to not see quite clearly unless we are really forced to do so (like the way we were).

NPCMJ Webpage: `http://npcmj.ninjal.ac.jp/`

References

Butler, Alastair, Kei Yoshimoto, Shota Hiyama, Stephen Wright Horn, Iku Nagasaki, and Ai Kubota. 2017. The Keyaki Treebank Parsed Corpus. http://www.compling.jp/Keyaki/.

Masuoka, Takashi and Yukinori Takubo. 1992. *Kiso Nihongo Bunpoo: Kaitee-Ban (Basic Japanese Grammar: Revised Edition)*. Tokyo: Kurosio Publishers.