A case study on the Coordinate Structure Constraint in Japanese

Yusuke Kubota and Ai Kubota
University of Tsukuba and ex-NINJAL

In designing treebanks, one of the central questions is how much and what kind of information to annotate. This of course greatly depends on the purpose(s) of the final product, but in a large- scale project that runs for a long time, it is often difficult to have a clear answer to that question in advance. Thus, in practice, corpus development and (at least the initial, exploratory phase of) the use of the corpus to reveal linguistic generalizations run in tandem. In this paper, we sketch one aspect of this situation in the development of the NINJAL Parsed Corpus of Modern Japanese (NPCMJ), and discuss both the potential benefit of guiding the annotation by (what one takes to be) linguistic insight and challenges that one faces when taking this approach. Our case study focuses on the question of whether the parsed corpus of Modern Japanese currently under development can be used to shed light on a theoretical debate about the status of the Coordinate Structure Constraint (CSC) in Japanese.

In Japanese, semantic relations typically expressed by coordination in English are expressed by the so-called *-te* form and the renyookei form, both of which are morpho-syntactically subordination rather than coordination constructions (this can be seen most clearly from the fact that both involve non-finite verb forms):

(1) a. John-ga utai/utat-te, Mary-ga odot-ta.
John-NOM sing/sing-TE Mary-NOM dance-PAST
'John sang and Mary danced.'
b. John-wa mise-ni iki/it-te, hon-o kat-ta.
John-TOP store-DAT go/go-TE book-ACC buy-PAST
'John went to the store and bought the book.'
c. John-wa sono sakana-o tabe/tabe-te, byooki-ni nat-ta.
John-TOP that fish-ACC eat/eat-TE illness-DAT become-PAST
'John ate the fish and got sick.'

Despite this, both the *-te* form and the renyookei form initially appear to exhibit the same patterns of CSC as English coordination:

(2) a. *This is the book that John bought __ and Mary bought the magazine.
b. * Kore-ga John-ga kai/kat-te Mary-ga zassi-o kat-ta hon-da.
this-NOM John-NOM buy/buy-TE Mary-NOM magazine-ACC buy-PAST book-COP

intended: ‘This is the book such that John bought it and Mary bought the magazine.’

It is well-known that in English, well-formed examples that violate the CSC can be found when the semantic relation between the two clauses is not ‘parallel’ (Ross 1967; Schmerling 1972):

(3) a. This is the book that John went to the store and bought___ .
b. This is the stuff that the guys in the Caucasus drink___and live to be a hundred.

Kubota and Lee (2015) observe that basically the same patterns of ‘CSC violation’ can be found in Japanese and Korean and conclude, on the basis of this cross-linguistic observation, that the CSC should be viewed as a semantic/pragmatic principle rather than a syntactic constraint: the relevant facts receive the most straightforward explanation by assuming that extraction out of a single clause is disallowed when the two clauses stand in a semantically parallel relation, but is possible when the two clauses stand in non-parallel relations.

In this paper, we first demonstrate that the type of examples that support Kubota and Lee’s (2015) claim can be searched and found easily in a parsed corpus like NPCMJ. This clearly demonstrates the utility of a parsed corpus in theoretical linguistics research. In particular, a parsed corpus is especially useful for finding examples that instantiate specific patterns that can be used to refute claims made in the literature. To give just one example, in the literature on the CSC in Japanese, Tokashiki (1989) has claimed that only the -te form (and not the renyookei) allows for the type of CSC violation analogous to the English examples in (2) (similar claims have been made for related constructions in Korean by Cho (2005) and Yoon (1997)), based on examples such as the following:

(4) Kore-ga Taroo-ga oki-te/?*oki arat-ta kutu-da.
this-NOM Taro-NOM wake.up-te/wake.up wash-PAST shoes-COP
‘These are the shoes that Taro woke up and washed.’

We show that well-formed examples refuting Tokashiki’s claim are attested in NPCMJ.

Based on this simple demonstration of the use of a parsed corpus in linguistic research, we then turn to a subtler, and somewhat more complex relationship between corpus development and linguistic research. The current annotation scheme of NPCMJ (Butler et al. 2017) distinguishes between ‘coordinate coordination’ and ‘subordinate coordination’ via two tags CONJ and SCON, which is in part based on requirements from the semantic calculation system that assigns predicate calculus representations for parsed sentences (Butler 2015) and in part guided by the general idea that the notion of ‘co-subordination’ (Hasegawa 1996) is relevant in linguistic analysis and classification.

Although the CONJ/SCON distinction currently employed in the NPCMJ annotation scheme and the notion of parallel vs. non-parallel semantic relations in the sense of Kubota and Lee (2015) are motivated by different kinds of considerations and hence should not be conflated, it is true that at an intuitive level, the two are expected to broadly converge in practice. We propose to test this hypothesis by taking a subset of NPCMJ and examining the degree to which the two criteria converge. We discuss what conclusions can be drawn about the relationship between linguistic research and corpus development depending on the results of this correlation analysis.

References

Butler, Alastair. 2015. *Linguistic Expressions and Semantic Processing*. Dordrecht: Springer.

Butler, Alastair, Stephen Wright Horn, Kei Yoshimoto, Iku Nagasaki, and Ai Kubota. 2017. The Keyaki Treebank manual. Available at http://www.compling.jp/keyaki/contents.html.

Cho, Sae-Youn. 2005. Non-tensed VP coordination in Korean: Structure and meaning. *Language and Information* 9(1):35–49.

Hasegawa, Yoko. 1996. *A Study of Japanese Clause Linkage: The Connective TE in Japanese*. Stanford, California: CSLI Publications.

Kubota, Yusuke and Jungmee Lee. 2015. The Coordinate Structure Constraint as a discourse-oriented principle: Further evidence from Japanese and Korean. *Language* 91(3):642–675.

Ross, John Robert. 1967. Constraints on Variables in Syntax. Ph.D. thesis, MIT.

Schmerling, Susan. 1972. Apparent counterexamples to the Coordinate Structure Constraint: A canonical conspiracy. *Studies in Linguistic Sciences* 2(1):91–104.

Tokashiki, Kyoko. 1989. On Japanese Coordinate Structures: An Investigation of Structural Differences between the -Te Form and the -I Form. Master’s thesis, The Ohio State University.

Yoon, James Hye Suk. 1997. Coordination (a)symmetries. vol. 7, 3–30. Seoul: Hanshin Publishing Company.