

# Using a parsed corpus for linguistic research: A case study on the Coordinate Structure Constraint in Japanese

Yusuke Kubota      Ai Kubota

March 28, 2018

## Abstract

This paper presents a case study of using the NINJAL Parsed Corpus of Modern Japanese (NPCMJ) for theoretical linguistics research. NPCMJ is the first phrase structure-based treebank for Japanese that is specifically designed for application to linguistic (in addition to NLP) research. After discussing some basic methodological issues pertaining to the use of treebanks for theoretical linguistics research, we introduce our case study of the Coordinate Structure Constraint (CSC) in Japanese, showing that NPCMJ enables us to easily retrieve examples that support the key theoretical claim of Kubota and Lee (2015). The corpus-based study we conducted moreover revealed a previously unnoticed tendency that was highly relevant for further clarifying the nature of the CSC. We conclude the paper by briefly discussing some further methodological issues brought up by our case study pertaining to the relationship between linguistic research and corpus development.

## 1 Introduction

This paper presents a case study of applying the NINJAL Parsed Corpus of Modern Japanese (NPCMJ; <http://NPCMJ.ninjal.ac.jp/>) for theoretical linguistics research. Specifically, we explore the possibility of using NPCMJ for collecting attested data corroborating the claim made by Kubota and Lee (2015) (K&L) on the status of the Coordinate Structure Constraint (CSC) in Japanese. The results are positive. NPCMJ proved to be an effective tool for extracting sentences attesting the CSC violation patterns of the sort that crucially support K&L's argument. The treebank search moreover identified a tendency that was overlooked in previous studies (including K&L), but one that is arguably relevant for further clarifying the nature of the CSC. With these results, we hope to convince the reader that treebanks are highly effective tools for addressing questions that have direct theoretical relevance, and that they yield results that cannot be easily obtained by alternative methods (such as introspective judgments and unannotated or lightly annotated linguistic data). The present paper also discusses some possible challenges for using treebanks for linguistic research. This is not meant to discourage potential users of treebanks;

on the contrary, we believe that one can exploit the full power of treebanks only by having an accurate knowledge of what they are, and this is why it is important, on the part of users of treebanks, to understand at least some of the key issues involved in the construction of treebanks.<sup>1</sup>

The paper is structured as follows. We start with a discussion of some methodological and practical issues pertaining to the use of treebanks in linguistic research, so as to situate the present case study within a larger context. We then present our case study of the CSC. The case study reveals some further methodological issues pertaining to the relationship between treebank development and linguistic research. We conclude the paper by briefly commenting on these larger issues.

## 2 Some methodological issues

Treebanks are corpora with full syntactic annotation. This means that they are special type of corpora that come with unusually fine-grained linguistic/grammatical information as compared to ordinary large-scale corpora. For this reason, even if one is primarily interested in just using some existing treebank for one's research and not in actually developing such a corpus, it is still instructive to learn about the general challenges that treebank development faces.<sup>2</sup> Treebanks almost always come with detailed annotation manuals,<sup>3</sup> and these manuals are very instructive, as they tell us both the undergirding principles behind the development of the specific treebank in question and the specific ways in which various practical issues have been addressed.

In our own experience, the following issues stood out as potential bottlenecks in using NPCMJ as a tool for linguistic research:

- **Data size** – Treebanks are expensive to construct, and therefore tend to be small in size.
- **Annotation quality** – The distinction one is after may be encoded in the corpus, but the annotation may be unreliable/inconsistent.

---

<sup>1</sup>Previous studies employing treebanks for linguistic research are still surprisingly few in number despite the fact that treebanks for at least major European languages have been around for quite a long time by now. To get a feel for some representative studies, see Kübler and Zinsmeister (2015), which contains a concise summary of the studies by Bresnan et al. (2007) and Wasow et al. (2011). The proceedings of the conference 'Treebanks and Linguistic Theories' contain many papers exploiting the use of treebanks for addressing various issues in linguistic research.

<sup>2</sup>For readable and instructive discussions of using corpora (including treebanks) for linguistic research, see Meurers (2005), Meurers and Müller (2009) and Kübler and Zinsmeister (2015). Palmer and Xue (2013) contains a concise and useful discussion of treebank development from the viewpoint of corpus developers.

<sup>3</sup>In the case of NPCMJ, the annotation manual is available at [http://www.compling.jp/keyaki/manual\\_en/contents.html](http://www.compling.jp/keyaki/manual_en/contents.html).

- **Searchability** – If the phenomenon in question involves properties that are not purely structural, that information may not be encoded.

These issues essentially all pertain to the inherent nature of treebanks: by their very nature, treebanks embody the tension between ‘bottom-up’ descriptive archiving of linguistic data and ‘top-down’ linguistic theorizing guiding us what kinds of information to annotate.

In the rest of this section, we discuss these issues briefly. First, large-scale treebanks require great amount of time and extensive human resource (both in quality and quantity) to construct. This means that treebanks with reliable quality tend to be small in size as compared to other types of corpora. For example, even including unreleased data (for which final checking is pending), NPCMJ contains only about 650,000 words. This is less than 1% of the Balanced Corpus of Contemporary Written Japanese ([http://pj.ninjal.ac.jp/corpus\\_center/bccwj/en/](http://pj.ninjal.ac.jp/corpus_center/bccwj/en/)), which contains 1 billion words and which is widely used in linguistic research. Comparable treebanks for other languages are only twice or three times larger than NPCMJ.<sup>4</sup> Data size may become a real issue depending on the type of question one is interested in finding an answer for. In fact, this can pose a real difficulty in theoretical research, given that in many cases theoretical linguists are interested in rare phenomena. There is no easy solution for this dilemma, but one promising approach is to increase the data size by including machine-annotated data without manual correction. For a case study exploiting this possibility, see Hinrichs et al. (2015), which convincingly shows that this approach is particularly effective for identifying occurrences of rare phenomena. The challenge that such an approach faces is of course quality control. For some discussion about how one might go about estimating error rate with respect to a particular grammatical feature one is interested in within a large-scale machine-annotated corpus, see Bloem (2016).

The second challenge also pertains to the fact that treebanks are expensive to construct. Treebanks annotate full syntactic structure for each sentence (and some treebanks, including NPCMJ, encode even more fine-grained information such as grammatical relations). Carrying out such detailed linguistic analysis on real-world data consistently would be a great challenge even for experienced syntacticians. See Palmer and Xue (2013) for some discussion on annotation evaluation for treebanks. They report that the Chinese Treebank has achieved inter-annotator agreement of over 95% accuracy (in terms of the Parseval F score, which takes into consideration the bracketing, but not the category labels) after adjudication of inter-annotator discrepancies. This sounds encouraging, but one should keep in mind that there is no guarantee that quality control of approximately the same level is enforced in all existing treebanks. Annotation quality may then turn out to pose an issue for linguistic application. If the key distinction one is looking for is annotated incorrectly/unreliably, the search result will inevitably contain false hits (negatively affecting precision) and one will inevitably miss some positive instances (negatively affecting recall).

---

<sup>4</sup>For example, the Penn Chinese Treebank contains 1.5 million words, and the Tübingen Treebank of Written German contains about 1.8 million words.

In many cases, false positives may not be a serious issue in treebank-based research, since if the corpus itself is not huge, the results for a moderately complex (thus linguistically interesting) search will typically return a small enough number of hits allowing for exhaustive manual inspection. But missing positive instances remains an annoying issue.

Finally, even though treebanks encode rich linguistic information, in actual linguistic research, it often turns out that structural information alone is not enough to formulate a linguistically interesting search. In particular, even in syntactic research, important and interesting theoretical questions often involve multiple levels of linguistic information. A typical situation of this is when one is interested in the distribution of a particular class of words in particular syntactic environments.

Let us provide a concrete example demonstrating this last point. In our initial exploration of using NPCMJ for theoretical research, we identified one possible topic which we decided not to pursue further for the time being for practical reasons. The linguistic issue we were interested in was the distribution of evaluative adverbs in the scope of sentential operators—interrogatives, imperatives, modals and negation. It has been reported in the literature (Sawada, 1978) that evaluative adverbs in Japanese don't appear in the scope of such operators. Examples involving interrogative and negation are given in (1).

- (1) a. \*Orokanimo John-wa odot-ta no?  
       stupidly John-TOP dance-PAST Q  
       ‘Stupidly, did John dance?’  
    b. Orokanimo John-wa odotte-i-nai.  
       stupidly John-TOP dance-be-NEG-PAST  
       ‘Stupidly, John hasn't danced.’ (ADV > NEG, \*NEG > ADV)

However, a recent study by Kubota (2015, 23) suggests that the facts are somewhat more complex. In this work, Kubota notes that at least for the interrogative sentences like (1a), the acceptability of the allegedly ill-formed examples can be improved significantly by embedding them in a certain kind of discourse context.

We believe that a need to search for this type of correlation (or lack thereof) between a particular syntactic configuration and a particular lexical class is very typical in linguistic research, since many interesting and important issues in theoretical linguistics pertain directly to such correlations (e.g., NPI licensing, the licensing and interpretation of *wh*-indefinites in Japanese by the interrogative *ka* and the ‘universal’ *mo* particles). In the case of the evaluative adverb distribution, the lack of appropriate lexical resource unfortunately made it difficult to formulate an effective search query. Adverbs are not subclassified in the NPCMJ corpus, and that practically meant that all we could do was either to run a very coarse-grained search retrieving all occurrences of all adverbs in the relevant syntactic configurations or to prepare a list of evaluative adverbs manually in advance. While the first strategy has an advantage in terms of recall (in the sense that we are not going to miss any true hits), it is impractical in terms of precision (the search results will contain

too many false hits). The second strategy is often the best compromise in situations like this. It will work particularly well when the lexical class in question is a closed or near-closed class (such as focus particles). Unfortunately, for the case of evaluative adverbs, this alternative strategy did not seem promising either. The class of evaluative adverbs is not a closed class. This makes it difficult to come up with a reasonably complete list of such items manually. Thus, conducting research on this topic will be realistic only when a reasonably large-scale lexical resource with a fine-grained subclassification of different parts of speech (in this case adverbs) is available.

Due to the three methodological issues mentioned above, using treebanks for linguistic research isn't always straightforward. But this does not mean that such an attempt is futile. On the contrary, we believe that as long as the user of the corpus is aware of the potential limitations and pitfalls, currently existing treebanks do already give us invaluable resource for gaining insight into questions that are directly relevant for theoretical linguistic research. In the rest of the present paper, we aim to demonstrate this point through a concrete case study.

### 3 Case study: A corpus-based study of the Coordinate Structure Constraint in Japanese

In this section, we report on a case study we have conducted on the so-called Coordinate Structure Constraint (CSC) in Japanese using the NPCMJ corpus. Specifically, we were interested in the question of whether corpus data can be used to either validate or invalidate the claim made by Kubota and Lee (2015) that the CSC should be viewed as a pragmatic principle rather than a syntactic constraint. To state the conclusion first, the results were positive, in that we were able to find occurrences of sentences in the corpus that crucially support K&L's key claim; we were moreover able to identify a tendency that was overlooked in the previous literature (including K&L), but one which is highly relevant for the overall conclusions of K&L's study. To make the present paper self-contained, we start by reviewing the relevant theoretical literature, both on the CSC and on the relevant facts we focused on in Japanese. However, due to space-limitations, the background discussion below is kept to the minimum. Readers interested in the details are referred to Kubota and Lee (2015) and references cited therein.

#### 3.1 The CSC and its status in the grammar

The CSC is a famous constraint in the syntactic literature first identified by Ross (1967) as one of the 'island constraints'. It prohibits extraction out of a single conjunct in a coordinate structure:

(2) **Coordinate Structure Constraint**

In a coordinate structure, no conjunct may be moved nor may any element contained

in a conjunct be moved out of that conjunct.

(Ross, 1967, 89)

The first part of the constraint prohibiting extraction of the whole conjunct is called the ‘conjunct constraint’ and the second part prohibiting extraction of an element out of a conjunct (i.e., the underlined part) is called the ‘element constraint’. In the rest of this paper, we focus on the element constraint, and for this reason, when we just say ‘the CSC’ in what follows, we mean just the element constraint, not the entirety of the CSC including the conjunct constraint.

The CSC is supposed to account for contrasts like the following, and has standardly been taken to be a syntactic constraint:

- (3) a. \*This is the book that he [[bought \_\_ ] and [read the newspaper]].  
b. This is the book that he [[bought \_\_ ] and [didn’t read \_\_ ]].

However, Ross himself was the first to note counterexamples to the CSC. There are well-known examples of extraction out of a single conjunct, such as the following, which are perfectly acceptable and which remain problematic for any formulation of the CSC as a syntactic constraint along the lines of (2).

- (4) a. Here’s the whiskey which I [[went to the store] and [bought \_\_ ]]. (Ross, 1967)  
b. That’s the stuff that the guys in the Caucasus [[drink \_\_ ] and [live to be a hundred]]. (Schmerling, 1972)

These and similar data have led some researchers to propose an alternative analysis of the patterns of acceptability displayed in the data in (3) and (4) based on pragmatic, rather than syntactic factors. This line of analysis starts with Lakoff (1986). Kehler (2002) elaborates this approach further and embeds it in a general theory of discourse coherence that has robust empirical applications outside of the CSC facts (such as the licensing conditions of VP ellipsis and Gapping). We summarize Kehler account of the CSC in what follows since it forms the basis of K&L’s work on Japanese and Korean.

Kehler’s account of the CSC builds on a general theory of discourse relations in terms of the following three-way classification:

- (5) **Resemblance**  
a. Mary is a linguist and Sue is a psychologist.  
b. Mary voted for Clinton, but Bill voted for Trump.
- (6) **Cause-Effect**  
a. George is a politician, and therefore he’s dishonest.  
b. George is a politician, but he’s honest.
- (7) **Contiguity**

- a. Larry went into a restaurant. The baked salmon sounded good and he ordered it.
- b. George picked up the speech. He began to read.

Among the three, the Resemblance relation is unique in that this discourse relation largely depends on the form of the sentence, most typically the predicate argument structure. For example, in (5b), the two clauses share the same predicate and the two arguments of the predicate in one clause have counterparts in the other clause occupying the same argument position (subject and object).

The two other discourse relations, on the other hand, are supported by the semantic, rather than the structural relations between their components. The Cause-Effect relation holds between two clauses if the events or situations described by them are related via some kind of causal relation in the broader sense. This includes cases such as (6b), which signals the *lack* of an expected causal or inferential relation.

Finally, the Contiguity relation relies most heavily on extragrammatical factors among the three. It holds between event descriptions that are most typically sequentially ordered, and which, taken as a whole, form a coherent narrative. The following CSC violation example from Lakoff (1986) is a typical example of the Contiguity relation:

- (8) This is the kind of brandy that you can sip after dinner, watch TV for a while, sip some more of, work a bit, finish off, go to bed, and still feel fine in the morning.

Here, the conjoined clauses form a coherent discourse which, as a whole, characterizes the quality of the brandy in question. The factors that support the Contiguity relation seem to be varied, including world knowledge, common sense, social/cultural conventions and patterns of human cognition.

Discourse relations are inherently pragmatic notions, and, for this reason, identifying the discourse relation that holds of a particular sequence of sentences in real text is not always clear-cut (in particular, drawing a boundary between the Cause-Effect and Contiguity is often difficult). But one relatively simple diagnostic one can employ for this purpose is the insertion of adverbial expressions that explicitly signal discourse relations, such as *on the other hand*, *by contrast* (Resemblance), *therefore*, *and so*, *because* (Cause-Effect), and *then*, *after that* (Contiguity). If the use of such explicit discourse relation markers is felicitous and if they don't change the overall meaning of the whole discourse, then this disambiguation strategy helps clarify the underlying discourse relations that are often implicit in real text.

Building on his theory of discourse relations, Kehler reformulates the CSC as a pragmatic condition along the following lines:

- (9) **Pragmatic Reformulation of CSC (Element Constraint)**  
When the discourse relation between the two conjuncts is Resemblance, extraction needs to take place from all conjuncts.

The key idea behind this proposal should be intuitively clear. Extraction involves positing a link between the head noun and a gap position inside the clause that hosts extraction. If such a link is established only in one of the two conjuncts, it will break the structural symmetry requirement imposed by the Resemblance relation on the two conjuncts. Thus, if the Resemblance relation holds, only ‘across-the-board’ extraction is allowed. By contrast, Cause-Effect and Contiguity relations do not impose any symmetry requirement on the two conjuncts to begin with. Thus, in these case, extraction from a single conjunct does not give rise to unacceptability.

### 3.2 CSC patterns in Japanese

We now turn to the CSC patterns in Japanese. To state the conclusion first, the overall patterns are essentially the same as in English. However, in Japanese both the constructions corresponding to English coordination and the constructions corresponding to English extraction display properties that are quite different from the respective constructions in English. We first review the basic properties of the relevant constructions, and then discuss the CSC patterns that they exhibit.

In Japanese, the *-te* form and the ren’yoo (or the ‘continuative’) form are the natural candidates for coordination constructions at the clausal level, given the range of meaning relations they can express, which largely overlaps that of English coordination.<sup>5,6</sup> As shown in (10), in these constructions, the first clause is marked by a non-finite *-te* or ren’yoo form, and the second clause carries the tense morphology.

- (10) a. [Taroo-ga **utai/utat-te**], [Hanako-ga odot-ta].  
 Taro-NOM sing/sing-TE Hanako-NOM dance-PAST  
 ‘Taro sang, and Hanako danced.’
- b. Taroo-wa [mise-ni **iki/it-te**] [hon-o kat-ta].  
 Taro-TOP store-DAT go/go-TE book-ACC buy-PAST  
 ‘Taro went to the store and bought a book.’

The clearest morphosyntactic evidence for coordinationhood is whether the conjuncts can independently serve as an expression of the given category. The *-te* and ren’yoo forms

---

<sup>5</sup>Coordination is a somewhat tricky notion to define/characterize (see, e.g., Haspelmath (2007) for an overview the relevant typological issues). For the purpose of the present paper, we follow Kubota and Lee (2015) in taking syntactic and morphological properties as the primary criteria, since the hypothesis under consideration is the validity of the CSC as a *syntactic* constraint. Typologically, the Japanese *-te* form and the ren’yoo form seem to fit most comfortably in the category of ‘cosubordination’, which shares certain properties with both coordination and subordination (cf. Hasegawa (1996); Velupillai (2012)).

<sup>6</sup>We exclude other types of connectives in the ensuing discussion such as *-tari* and *-si*. These forms express more specific types of semantic relations between the clauses they combine than the *-te* form and the ren’yoo form, and for this reason, don’t display the full patterns of interactions with discourse relations exhibited by the *-te* form and the ren’yoo form. We believe that the overall distributional patterns of these other connectives are entirely consistent with K&L’s proposal.

do not carry their own tense and thus cannot stand alone as an independent sentence, as shown in (11):

- (11) \*Taroo-ga uta-o utai/utat-te.  
 Taro-NOM song-ACC sing/sing-TE  
 intended: ‘Taro sang a song.’

Given this, it seems reasonable to assume that *-te*-marked and *ren’yoo* clauses are morpho-syntactically subordinate clauses, at least without any clear evidence to the contrary.

Turning now to the counterpart of English extraction, we again see that the syntactic properties of the relevant constructions are quite different. ‘Relative clauses’ are formed in Japanese by placing a finite clause in front of the modified noun as in (12):

- (12) [Taroo-ga \_\_ yon-da] hon  
 Taro-NOM read-PAST book  
 ‘the book that Taro read’

Two properties sharply distinguish the Japanese NPCMJ corpus from English relative clauses, as noted by a number of previous authors (cf., e.g., Kuno (1973), Teramura (1975–1978), and Matsumoto (1997)). First, as exemplified in (13a), the Japanese NPCMJ corpus does not obey typical island constraints (except for the CSC). Second, it has the so-called ‘gapless’ variants, such as (13b), in which the relationship between the modifying clause and the head noun is mediated by a pragmatic inference partly supported by world knowledge.

- (13) a. [[\_\_ ki-te i-ru] yoohuku-ga kitanai] hito  
 wear-TE PROG-NPST clothes-NOM dirty.NPST person  
 ‘the person such that the clothes that s/he is wearing are dirty’  
 b. [atama-ga yoku-naru] hon  
 mind-NOM good-become book  
 ‘a book such that one becomes smart by reading it’

Given these properties, and given the fact that Japanese is a pro-drop language, Matsumoto (1997) concludes that the null hypothesis about noun-modifying constructions in Japanese is that the semantic relation between the head noun and the modifying clause is underspecified in the syntax and is mediated pragmatically. In this paper, we follow Matsumoto in taking this view on NPCMJ corpus in Japanese.

Thus, we have independent reasons to believe that both the coordination-like constructions (i.e., the *-te* and *ren’yoo* forms) and the noun-modifying construction have properties that distinguish them from their counterparts in English, at least in respects that are crucial for the applicability of the CSC as a syntactic constraint. Given this, it is remarkable that Japanese exhibits basically the same patterns of regular CSC behavior and its ‘apparent’ exceptions as in English in these syntactic constructions. The relevant data are as follows:

- (14) a. \*Taroo-ga [zassi-o kat-te/kai] [\_\_ yon-da] hon  
 Taro-NOM magazine-ACC buy-TE/buy read-PAST book  
 intended: ‘the book such that Taro bought the magazine and read it’
- b. Taroo-ga [\_\_ kat-te/kai] [\_\_ yon-da] hon  
 Taro-NOM buy-TE/buy read-PAST book  
 ‘the book that Taro bought and read’ (Resemblance)
- (15) a. [\_\_ syutuensi-te/syutuensi] [kookaisi-ta] sakuhin  
 appear-TE/appear regret-PAST piece  
 ‘the piece (movie) that he appeared in and regretted’ (Cause-Effect)
- b. [Kinokuniya-ni it-te/iki] [\_\_ kat-ta] hon  
 Kinokuniya-LOC go-TE/go buy-PAST book  
 ‘the book that I bought when I went to Kinokuniya’ (Contiguity)

The above patterns fall out straightforwardly by taking the CSC as a pragmatic constraint along the lines of (9). Regardless of the syntactic properties of the constructions involved, linking the head noun to a missing argument position breaks the structural symmetry imposed on the two clauses by the Resemblance relation. But if the discourse relation between the two clauses is either Cause-Effect or Contiguity, no such symmetry requirement exists for the two clauses. Thus, the patterns exhibited by (14) and (15) are predicted without any extra stipulation. By contrast, the CSC pattern in (14) and (15) raise a number of serious issues to the syntactic approach to the CSC, as discussed in detail by K&L.<sup>7</sup>

### 3.3 Results of the corpus search

At the time of writing K&L, no treebank for Japanese or Korean was available as a convenient tool for searching for grammatical structures in a form easy to use for working linguists.<sup>8</sup> K&L do nevertheless offer attested data instantiating CSC violation struc-

<sup>7</sup>One might wonder whether an independently motivated syntactic (or syntactico-semantic) distinction among different subtypes of the *-te* form or ren’yoo form could be linked to the CSC patterns observed here. We believe that such an approach would be difficult to maintain. For example, Uchimaru (2006) distinguishes between different subtypes of the *-te* form based on tests such as NPI licensing and focus particle interpretation, but on her classification, *-te* forms expressing all the three discourse relations exemplified in (14) and (15) fall under the same ‘TP coordination’ class. Similarly, in relation to the influential three-level classification of clausal hierarchies in Japanese, Minami (1974) posits that the *-te* form and ren’yoo form can belong to all of the three classes A–C. However, since noun-modifying constructions by definition belong to Class A/B and since Class A (which according to Minami is basically limited to clauses functioning as manner adverbs) is too small to host most of the examples of the form in (14) and (15), we are forced to conclude that all of our examples belong to Minami’s Class B. But then, this classification cannot be used to distinguish between cases like (14) and cases like (15).

<sup>8</sup>The dependency-based Kyoto Corpus (<http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KyotoUniversityTextCorpus>) was available back then, but it was released and used primarily as

turally identical to (15). The strategy they took was to formulate an approximate query in Google string search which in one instance took the following form:

(16)           \* si|site \* sita \* wa  
Gloss: do/do.TE do.PAST TOP

The intention here is that the arbitrary strings before the two occurrences of light verb *suru* (*si|site* (*-te/ren'yoo* form) and *sita* (past tense)) would correspond to the first and second clauses and the third arbitrary string (followed by the topic marker *wa*) would correspond to the modified head noun. Since no further constraints are imposed, this query returns many false hits. But it was sufficient (but only barely sufficient) for the purpose of finding examples to cite in the paper.

K&L's compromise shows that Google string search is actually useful (perhaps already an obvious point for working syntacticians). The size of the corpus is certainly a big advantage. But it also shows the disadvantages of this approach. First of all, having to wade through many irrelevant search results is inconvenient. Second, in order to identify sequences of two clauses, K&L had to restrict the search to target only the light verb *suru*. For the case at hand, this compromise was good enough, but in the general case, there is no guarantee that restricting the search arbitrarily for the sake of searchability would not have a negative effect on either the quality or quantity of the search results.

The development of NPCMJ at NINJAL starting from 2016 gave us an opportunity to revisit this issue. NPCMJ comes with an explicit encoding of both 'gapped' and 'gapless' noun-modifying constructions;<sup>9</sup> it distinguishes different grammatical relations; and it enables one to search for the *-te* form and the *ren'yoo* form separately. In short, it has all the distinctions we need to search for examples like (15).

We used the internal version of the NPCMJ corpus as of August 4, 2017, containing 59,557 sentences.<sup>10</sup> The Stanford *tregex* (<https://nlp.stanford.edu/software/tregex.shtml>) search tool was used for formulating queries and retrieving data. We formulated eight queries by specifying different values for the following three binary parameters:

- the form of the first clause (*-te* vs. *ren'yoo*)
- grammatical relation of the gap site (subject vs. object)<sup>11</sup>

---

resource for NLP research.

<sup>9</sup>Although empty positions in noun-modifying constructions are tagged as 'traces' in NPCMJ, this is not meant to embody any theoretical claim. More generally, annotations in treebanks are just notations that are meant to help the user to retrieve useful information as readily as possible. We return to the relationship between linguistic theory and corpus development in section 4.

<sup>10</sup>This is mostly identical to Keyaki Treebank version 1 (<http://www.compling.jp/keyaki/>), which contains about 40,000 sentences. The internal data we used contained constructed examples from a thesaurus and textbooks, which is why the size of the corpus is apparently 1.5 times larger (in terms of the number of sentences), but we excluded all constructed data from our search results manually.

<sup>11</sup>We ignored indirect object gaps since they were few in number. Gaps in oblique positions (e.g. temporal/locative PPs) were often inconsistently tagged, and for that reason we set them aside as well.



All of the 21 true hits were cases of either Contiguity or Cause-Effect, and no clear instance of Resemblance was found. Thus, although the size of the corpus is limited, we can conclude that the pragmatic reformulation of the CSC in (9) is confirmed by the NPCMJ data.

The rest of the hits are classified into three types: ambiguous examples, false hits, and annotation errors. Of the three, annotation errors are simple. Most of the examples have the following form string-wise:

(21) ... **V**(-te) ... **V**-ta N ... **V**

This is structurally ambiguous in terms of the attachment site of the first clause, which can either be a constituent of the matrix clause or part of the noun-modifying clause. We manually excluded cases that were arguably annotation errors due to misanalysis from true hits.

Ambiguous cases are cases in which the clause that is annotated as ‘gapless’ could alternatively be analyzed as having a gap position, either as some oblique dependent (e.g. temporal/locative PPs) or as a genitive modifier of some overt argument. Although these decisions may reflect some linguistic insight (either due to the human annotator(s) or the annotation manual), it did not seem prudent to take a blind faith in the annotation decisions on a subtle analytic question like this when using corpus data to support a particular theoretical claim. For this reason, we have decided to exclude all these cases from true hits.<sup>12</sup>

Finally, false hits mainly consisted of two types of cases. First, since the queries only specified the presence of a gap in either the subject or the object position in one clause and the absence of a gap in the same position in the other clause, they returned examples in which there were gaps in both clauses, one in the subject position and the other in the object position. The other false hits consisted of cases of the *-te* form or the ren’yoo form that can be analyzed as complex postpositions. An example of this latter type is given in (22).

(22) [kagaku-teki-**ni mi-te**] [\_\_ kooka-ga utagawasii] tiryoo-kooi  
 scientifically see-TE effect-NOM dubious treatment  
 ‘treatment whose effects are dubious from a scientific perspective’  
 10\_newswire-closed\_BCCWJ\_25\_PN1c\_00006;JP

These cases are distinguished from the truly clausal *-te* form and ren’yoo form in that the verb has lost its original sense and argument structure, and the combination of the postposition and the verb functions as a unit. The defective clausal status of these examples can be tested, for example, by whether all the arguments of the verb are clear from the context (for example, in (22), the subject of *mi-te* ‘see’ is not clear). But here too, the

<sup>12</sup>Annotation inconsistency was (unsurprisingly) frequent in such cases. This itself is an interesting phenomenon. Studying inconsistencies of this sort carefully may give us some insight into the correspondence between linguistic theory (which inevitably involves a high level of abstraction) and actual linguistic data.

boundary between full-fledged clauses and complex postpositions is not always clear-cut. We followed the same principle as above of excluding all suspicious cases from true hits.

In addition to confirming the reformulated CSC in (9), our corpus search has revealed two notable facts. First, there are cases of CSC violation both with the *-te* form and the ren'yoo form. Though few in number (only two out of 126 hits), the fact that the corpus search yielded instances of the ren'yoo form (such as (19) above) was significant since this directly counterexemplifies a claim made in the literature by Tokashiki (1989) that the ren'yoo form, unlike the *-te* form, was true conjunction and that it would therefore obey the CSC. Our corpus search result shows that the incompatibility of the ren'yoo form with the CSC violation pattern is, if anything, only a tendency.

The second notable fact was that most of the attested examples of CSC violation involved 'extraction' from the second clause rather than the first clause. We had not expected this skew in the distribution of data before conducting the search—no previous literature reports on any such tendency, so, this was an unexpected result. However, as suggested to us by David Oshima (p.c.), this fact may provide yet another piece of evidence that the CSC facts in the Japanese noun-modifying construction are governed by non-syntactic factors. Because of the head-final word order of Japanese, the head noun appears after the modifying clause in the noun-modifying construction. But then, given the relative distance between the head noun and the 'gap', it would not be particularly surprising if examples with the 'gap' in the second clause would be easier to process (and hence occur more frequently in the corpus) than examples with the 'gap' in the first clause.

If this reasoning is on the right track, we predict that, other things being equal, the opposite tendency should be found in English relative clauses (where the constituent order is the opposite). It is also expected that in other types of displacement constructions in Japanese such as scrambling and topicalization, in which the displaced element appears on the left of the host clause, again, the opposite tendency should be found. We leave it for future research to test these predictions.

One point we would like to emphasize before moving on is that the overwhelming tendency for the gap to appear in the second clause would most likely have been difficult to notice in alternative methods. In particular, Google searches can identify hits from a huge body of text, but it is impossible to estimate precisely the relative difference in frequency, especially when the search results themselves contain a large amount of false hits (which actually was the case when K&L conducted Google search) and when, for that reason, it is impossible to look through all the hits manually. Only by examining the number of hits of each syntactic pattern from the same fixed corpus, as we have done above, did it become fully evident that a frequency difference does exist between extraction from the first clause and extraction from the second clause.

## 4 Some further methodological issues

We now turn to some further methodological issues that the present case study brings up, pertaining to the relationship between linguistic theory and corpus development. The ideal relationship between the two is one in which they inform each other. This is easier said than done. At least in our own experience (of being involved in the NPCMJ project as theoretical linguists), the distance between the thinking behind theoretical linguistics and that behind corpus development is often a source of confusion. We ourselves perhaps had an unfair advantage in this respect, since being part of the development team gave us an opportunity to familiarize ourselves with the annotation policy and practice. This was immediately useful, for example, in making educated guesses about places in which annotation inconsistency was likely. But the majority of treebank users are not themselves developers of the resources they use, and in such cases, figuring out the exact relationship between the annotation reflected in the corpus and one's own theoretical persuasion can be fairly challenging. The most important source of information is of course the annotation manual. In consulting the annotation manual, it is essential to keep in mind the overall design policy of the corpus, which is sometimes explicitly stated, but sometimes only implicit.

In the case of the NPCMJ corpus, we found it useful to distinguish two major factors that guide the current annotation policy, in order to make better sense of the particular treatments of specific cases. One of these is the obvious and sound principle, explicitly noted in the annotation manual, that the annotation aims to achieve descriptive adequacy and maximum convenience in the retrieval of information. The other, less obvious factor is the fact that the NPCMJ corpus (or, more precisely, the Keyaki treebank that it is based on) was originally designed to serve as a learning model for the syntactic component of an integrated system of syntactic/semantic parser (Butler, 2015).

The policy to distinguish ‘gapped’ and ‘gapless’ noun-modifying clauses can be seen as reflecting the first annotation policy, since this distinction roughly corresponds to the well-known distinction between the ‘inner’ and ‘outer’ relations (‘uchi-no kankei’ and ‘soto-no kankei’) due to Teramura’s (1975–1978) influential work in descriptive grammar. We want to emphasize here that this annotation policy served our purpose of identifying ‘CSC violation’ noun-modifying clauses *despite* the fact that, on the face of it, it was directly at odds with our own theoretical position (where, following Matsumoto (1997), we take all noun-modifying clauses to be gapless). As theoretical linguists, we are often (mis)led to take the annotations reflected in the corpus to embody some theoretical claim, but when one is using the corpus as a *tool*, this is not the right way to make sense of corpus annotation.

This, however, raises the vexing issue of whether a completely theory-free treebank development is possible (a moment’s reflection should tell one that such a thing is impossible). And here, the second, ‘hidden’ guiding principle of NPCMJ becomes relevant. Again, taking the distinction between ‘gapped’ and ‘gapless’ relative clauses as an example,

the policy to explicitly distinguish these two types of noun-modifying constructions finds support not only from convenience for search but also (at least partly) from the need to use the syntactic annotation as a basis for building semantic representations. If we take this other consideration seriously, there is a sense in which this annotation policy (at least implicitly) reflects a particular theoretical/analytic stance regarding the architecture of the syntax/semantics/pragmatics interface. To the extent that this is a theoretical claim, it contradicts the theoretical position of K&L (and of the authors of the present paper). This is so because it is a claim that the distinction between ‘gapped’ and ‘gapless’ noun-modifying clauses needs to be encoded at the level of syntax in order to obtain adequate semantic representations.

This discussion ultimately leads to a quite complex and fundamental issue: can this sort of tension (which one often encounters in treebanking) be safely regarded as a gap that inevitably exists between theory and implementation, or is it the case that freely admitting the existence of such a gap is, after all, an evasion in the name of theoretical idealization? This is of course a version of the frequently observed gaps between theory and practice found in many domains of inquiry (including computational linguistics), and in general, there is no simple answer to this question. The relationship between linguistic theory and treebank development is thus quite nuanced, and we think that it ultimately relates to fundamental issues in theory development in linguistics in the sense alluded to above. We find it stimulating that the practice of developing a treebank inherently gives rise to issues that potentially demand a fundamental rethinking of the key premises in theoretical linguistics.

## 5 Conclusion

In this paper, we have presented a case study of using the NPCMJ corpus for theoretical linguistics research. In the first part of the paper, we discussed some methodological issues pertaining to the use of treebanks for theoretical research in order to situate the present case study in a larger context. Although the value of treebanks as resource for research is well recognized in the NLP community, treebanks are still novel research tools in the theoretical linguistics community, and their potentials in this latter context don’t yet seem to be fully appreciated. We hope that the preliminary discussion of the methodological points we have offered here will provide a starting point for further examining these methodological issues and for innovative uses of treebanks for approaching theoretical important questions in linguistic research.

Our own case study has taught us many things. First and foremost, we were encouraged by the fact that, when certain conditions are met, treebanks enable us to directly search for linguistic phenomena that have immediate theoretical relevance. Second, the treebank search has revealed a previously overlooked tendency that can potentially shed further light on the essential characteristics of the phenomenon under investigation. Finally, the complex

and often subtle relationship between treebank development and theoretical linguistics research is thought-provoking in many respects—for pencil-and-paper linguists like the two of us, it gave us an opportunity to pause and reflect on the strengths and limitations of doing linguistic research in the usual familiar way.

## References

- Bloem, Jelke. 2016. Evaluating automatically annotated treebanks for linguistic research. In *Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora*, 8–14.
- Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Krämer, and J. Zwarts, eds., *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.
- Butler, Alastair. 2015. *Linguistic Expressions and Semantic Processing: A Practical Approach*. Springer.
- Hasegawa, Yoko. 1996. *A Study of Japanese Clause Linkage: The Connective TE in Japanese*. Stanford, California: CSLI Publications.
- Haspelmath, Martin. 2007. Coordination. In T. Shopen, ed., *Language typology and syntactic description*, vol. 2, 1–51. Oxford: OUP, 2nd edn.
- Hinrichs, Erhard, Daniël de Kok, and Çağrı Çöltekin. 2015. Treebank data and query tools for rare syntactic constructions. In *Proceedings of Treebanks and Linguistic Theory 14*, 106–118.
- Kehler, Andrew. 2002. *Coherence, Reference and the Theory of Grammar*. Stanford, California: CSLI Publications.
- Kübler, Sandra and Heike Zinsmeister. 2015. *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury.
- Kubota, Ai. 2015. *Adverbs of Evaluation in Japanese: A Conditional Account*. Ph.D. thesis, Michigan State University.
- 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.
- Kuno, Susumu. 1973. *The Structure of the Japanese Language*. Cambridge, Mass.: MIT Press.

- Lakoff, George. 1986. Frame semantic control of the Coordinate Structure Constraint. In A. M. Farley, P. T. Farley, and K.-E. McCullough, eds., *CLS 22 Part 2: Papers from the Parasession on Pragmatics and Grammatical Theory*, 152–167. Chicago: Chicago Linguistic Society.
- Matsumoto, Yoshiko. 1997. *Noun-Modifying Constructions in Japanese: A Frame Semantics Approach*. Amsterdam: John Benjamins.
- Meurers, W. Detmar. 2005. On the use of electronic corpora for theoretical linguistics: Case studies from the syntax of German. *Lingua* 115(11):1619–1639.
- Meurers, Walt Detmar and Stefan Müller. 2009. Corpora and syntax. In A. Lüdeling and M. Kytö, eds., *Corpus Linguistics. An International Handbook*, chap. 42, 920–933. Berlin: Walter de Gruyter.
- Minami, Fujio. 1974. *Gendai Nihongo-no Koozoo (The structure of the Modern Japanese Language)*. Tokyo: Taishukan.
- Palmer, Martha and Nanwen Xue. 2013. Linguistic annotation. In *The Handbook of Computational Linguistics and Natural Language Processing*, 238–270. Wiley-Blackwell.
- Ross, John Robert. 1967. *Constraints on Variables in Syntax*. Ph.D. thesis, MIT. Reproduced by the Indiana University Linguistics Club.
- Sawada, Harumi. 1978. Nichi-eego bunfukushi-rui no taishoo gengogaku-teki kenkyuu (A contrastive study of Japanese and English sentence adverbials: From the viewpoint of speech act theory). *Gengo Kenkyu* 74:1–36.
- Schmerling, Susan. 1972. Apparent counterexamples to the Coordinate Structure Constraint: A canonical conspiracy. *Studies in Linguistic Sciences* 2(1):91–104.
- Teramura, Hideo. 1975–1978. Rentaishuushoku no shintakusu to imi (The syntax and semantics of adnominal modification). Originally published in *Nihongo Nihonbunka* and reprinted in *Teramura Hideo Ronbunshuu I: Nihonbunpoo-hen [Collected works of Hideo Teramura I: Works on Japanese grammar]* (Kurosio Publishers 1992). English translation available at <http://adnominal-modification.ninjal.ac.jp>.
- 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.
- Uchimaru, Yukako. 2006. Dooshi-no te-kei-o tomonau setsu-no toogo-koozoo-ni tsuite (On the syntactic structure of clauses with the -te form). *Nihongo-no Kenkyuu (Study on the Japanese Language)* 1–15.

Velupillai, Viveka. 2012. *An Introduction to Linguistic Typology*. John Benjamins.

Wasow, Thomas, T. Florian Jaeger, and David Orr. 2011. Lexical variation in relativizer frequency. In H. J. Simon and H. Wiese, eds., *Expecting the Unexpected: Exceptions in Grammar*, 175–195. De Gruyter.