統語・意味解析コーパスの開発と言語研究

Development of and Linguistic Research with a Parsed Corpus of Japanese

http://npcmj.ninjal.ac.jp/

“Grammatical principles for annotation and query”

Stephen Wright Horn (Adjunct Researcher, Theory and Typology Division, National Institute for Japanese Language and Linguistics)

「統語・意味解析コーパスの開発と言語研究」研究発表会
東北大学（2017/03/04）

Keyaki Treebank to NPCMJ

The goals of the Keyaki Treebank, in particular with respect to semantic expression, are the same as they always have been.

But allowing comparability with the grammars of other languages is a greater concern now. Thus, simplification and systematization based on generalizable principles is needed wherever it can be done without losing information peculiar to Japanese.

Ultimately, the corpus should be an instantiation of a coherent descriptive grammar of the Japanese language, allowing searches for a wide variety of grammatical phenomena.

Headedness

```
__ > /^NP¥b/
== !N|PP|Q|NUMCLP|PRO|WPRO|NPR|
CONJP|D|CP-
THT|NP|ADV|PRN|Q¥+N|ADVP|NML|NP
-RFL|-LRB-|-RRB-
|PU|WD|FS|INTJ|QUOT|SYM|LST|META|
NUM|FW|PNL|WADV|ACT|INTJP|Q-
1|ADVP-1|BIND|AX|CONJ|P|-
RRB|VB|CL|QP|ADJI
==!/^IP|^PP|¥*T¥*|^CP|^NP|^¥*|^N;¥{|P
RN;¥*/
```

Headedness

Unary branching between phrasal categories:

(NP (NUMCLP))

(NP (PP)

Xなど/だけの, etc.

(PP (NP (PP (NP X とY) と) (P を), etc.

「今のはなし」, etc. (N’ deletion)

Multiple candidates for head:

(NP (N) (PRN))

(NP (NUMCLP (NUM 百)(CL 人))(N 以上)), etc.

(NP (CONJP) (NP))

(PP (NP (PP Xから) (PP Yまで)) (P を)), etc.

Projection

Projecting PP: P, [CONJP $. PP], *ICH*

Projecting ADVP: ADV, WADV, ADJI, [ADJN $. AX], *ICH*

Projecting NUMCLP: [NUM $. CL], NUM, N

Projecting NP: N, Q, NUMCLP, PRO, WPRO, NPR, [CONJP $. NP], PRN, Q+N, NML, NP-RFL, FW, *nullpronouns*, *ICH*, *particles*, *

Part of speech (e.g., particles)

Particles that share a given phonological form can mark more than one function or role in the same group. For example, in the group of particles for core grammatical roles,に can mark SBJ, SBJ2, LGS, OB1, and OB2.

Particles that share a given phonological form can appear in more than one group. For example, と appears in the group of particles for core grammatical roles, the group of particles for peripheral grammatical roles, the group of CND disambiguated PP conjunctive particles, and the group of CP-THT particles.

A particle of a given group and function may have more than one phonological form. For example, the particle の that appears in the group of particles as clausal constituents has an alternative form ん.

Particles can share phonological form with elements of an entirely different class. For example, ぐらい in 「これぐらい」is a particle, but in 「このぐらい」 it is a noun. Particles の, に, と, で, にて, にして, and として all share forms with copulas, according to one possible analysis.

Nexus

= predication (subject - predicate relation)

A simple rule: Nexus relations hold under IP.

Some consequences: ADJI in nexus projects clauses; ADJI-く, [NP-PRD に], and [ADJN に] under control become small clauses before なる・する; resultative expressions become clauses, と, たる, に, なる, become AX after ADJN and NP-PRD, etc.

Nexus

Some problems:

Missing predicates:「僕も∅だ」,「太郎は面白く∅、花子はつまらなく思った」,

Missing subjects: 「僕はウナギだ」,「時間がなかったのだ」, etc.

[∅copula + P] or [infinitive copula + ∅P]?: 「馬鹿と思った」

Nexus

More problems:

- Naming constructions:「かちかち山と名付けたんじゃ」
- Depictive clauses: 「素足で土間に降りた」

Argument and adjunct

A simple principle: An argument of a predicate X is a constituent without which X cannot be fully interpreted. An adjunct is optional.

「山を登る」 vs. 「山を滑る」

「電車に乗る」 vs. 「電車に酔う」

「店を素通りする」 vs. 「店をぶらぶらする」

「駅に到着する」 vs. 「駅に行く」

「止めようと言った」 vs. 「止めようとブ
レーキを踏んだ。

Argument and adjunct

「AはBと同じだ」
「AはBに似ている」
「それは考えるに値する」

「宝くじが当たった」
「二三軒にあたってみた」
「それは失礼にあたる」

「トランプが大統領になった」
「息子を総理大臣にした」
「かちかち山と呼ぶ」
「通勤時間を無駄だと感じる」

Core grammatical roles

If a predication (nexus) has a subject argument and a predicate, and projects IP, then an IP always has a subject.

The argument of a 1-place predicate is NP-SBJ.

A non-subject nominal argument of a 2-place predicate is NP-OB1.

A 2-place predicate may have NP-SBJ and NP-SBJ2.

For 3-place predicates, if one of the non-subject arguments is marked with を, that argument is assigned NP-OB1.

What is the difference between NP-OB1 and NP-OB2? There are no other clear criteria in the manual.

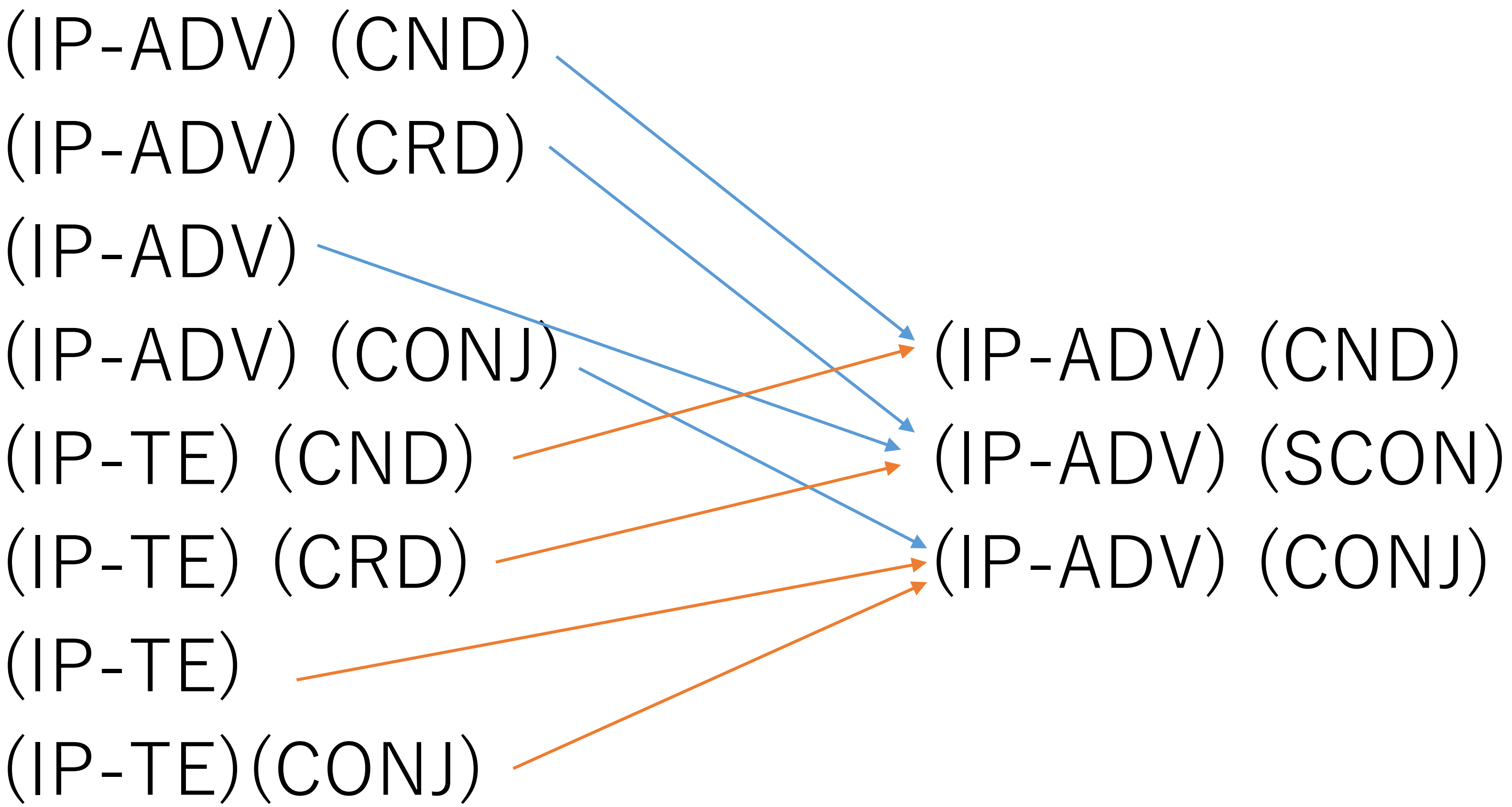
There is no consistently applied mechanism to mark non-nominal arguments as arguments (as distinct from adjuncts).

Changes in practice

Some of the ideas presented above might be incorporated into the annotation in a principled way. The improvements that might be made aren’t limited to these. But each change needs to be tested for clarity of implementation.

In the spirit of making the annotation task simpler, making nomenclature more intuitive, and improving the semantic parse output, the following changes will be implemented.

Label changes: clause linkage



(IP-ADV) (CND)

(IP-ADV) (CRD)

(IP-ADV)

(IP-ADV) (CONJ)

(IP-TE) (CND)

(IP-TE) (CRD)

(IP-TE)

(IP-TE)(CONJ)

(IP-ADV) (CND)

(IP-ADV) (SCON)

(IP-ADV) (CONJ)

More label changes

IP-INF	IP-SMC
CARD	NUM
NUMCL	CL

VB2 should be marker VB when primary verb (yet to be rolled out in the corpus)

Structural changes

PNL (prenominal) is a new part of speech for 連体詞

PNLP is a new phrasal category for difficult cases of 名詞修飾

ADJI with a subject argument always projects an IP, even in contexts of 名詞修飾

Q projects an NP (QP is not a category)

NUMCLP projects an NP

Structural changes: complex quotations

Complex quotations used to be treated a constituent coordinations:

ただ若い読者はそこまで深く考えないので、「何？雑誌なんてお金出して買うの？もったいな〜い」となる。

```
(CP-THT
        (IP-MAT
            (IP-MAT)
            (CONJP (CP-QUE))
            (CONJP (CP-QUE))))
```

Structural changes: complex quotations

Hereon, sentences forming complex quotations are directly under a new category:

ただ若い読者はそこまで深く考えないので、「何？雑誌なんてお金出して買うの？もったいな〜い」となる。

```
(CP-THT
              (multi-sentence
                     (IP-MAT)
                     (CP-QUE))
                     (CP-QUE)))
```

Structural changes: binding information

Prenominal quantifying expressions are marked ‘;*’ to indicate that they quantify the N that they are complement to:

```
(NP
        (PP;*
                (NP
                        (NUMCLP
                                (NUM 3)
                                (CL 匹)))
                (P の)
        (N 子豚))
```

Structural changes: binding information

Prenominal expressions that don’t quantify aren’t marked ‘;*’

```
(NP
    (PP
        (NP
            (PP
                (NP
                    (NUMCLP
                    (NUM 二人))
            (P だけ)))
        (P の)
    (N 秘密))
```

Structural changes: binding information

Appositive quantifying expressions are marked ‘;*’ to show they quantify their preceding sisters.

```
(PP
    (NP
        (NML
            (N 私たち))
        (NP;*
            (NUMCLP
                (NUM 二人))))
    (P が))
```

Structural changes: binding information

SENSE information has changed.
Previously this came as a separate SENSE node, e.g.:

```
( (IP-MAT (PP (NP (PRO 私))
              (P は))
          (NP-SBJ *)
          (PP (NP (N 空))
              (P を))
          (NP-OB1 *を*)
          (VB 仰ぎ見)
          (SENSE *仰ぎ見る.01*)
          (AXD た)
          (PU ．))
  (ID 1_vv-lexicon_20150226;1-1-1;VV;仰ぎ見る.01;MJ))
```

Structural changes: binding information

Such SENSE information is now included with curly braces at the POS level:

```
( (IP-MAT (PP (NP (PRO 私))
              (P は))
          (NP-SBJ *)
          (PP (NP (N 空))
              (P を))
          (NP-OB1 *を*)
          (VB;{仰ぎ見る.01} 仰ぎ見)
          (AXD た)
          (PU .   ))
  (ID 1_vv-lexicon_20150226;1-1-1;VV;仰ぎ見る.01;MJ))
```

Structural changes: binding information

BIND information has changed. Previously this came as a separate BIND node, e.g.:

```
( (IP-MAT (PP (IP-ADV (NP-SBJ *speaker*)
                      (PP (NP (PP (NP (BIND *MIRROR*)
                                      (N 鏡))
                                  (P の))
                              (N 数))
                          (P を))
                      (NP-OB1 *を*)
                      (VB 勘定)
                      (VB0 し))
              (P たら))
          (SCON *)
          (NP-SBJ (BIND *MIRROR*)
                  (ZERO *pro*))
          (NP (NUMCLP (NUM 六)
                      (CL つ)))
          (BIND *SBJ*)
          (VB あっ)
          (AXD た)
          (PU 。))
  (ID 457_Natsume;MJ))
```

Structural changes: binding information

Such BIND information is now included with curly braces (corresponding to what was BIND information placed UNDER an argument projection) or star notation (corresponding to adjacent BIND information that was sister to a immediately preceding floating quantifier):

```
( (IP-MAT (PP (IP-ADV (NP-SBJ *speaker*)
              (PP (NP (PP (NP;{MIRROR} (N 鏡))
                     (P の))
                   (N 数))
                (P を))
              (NP-OB1 *を*)
              (VB 勘定)
              (VB0 し))
         (P たら))
      (SCON *)
      (NP-SBJ;{MIRROR} *pro*)
      (NP;*SBJ* (NUMCLP (NUM 六)
                  (CL つ)))
      (VB あっ)
      (AXD た)
      (PU 。))
 (ID 457_Natsume;MJ))
```