Type grammar meets Japanese particles

Kumi Cardinal

Keio University, Japan
`cardinal@sfc.keio.ac.jp`

Abstract. This paper presents a computational analysis within the framework of a type grammar for the treatment of Japanese particles. In Japanese, particles express a number of functional relations; they follow a word to indicate its relationship to other words in a sentence, and/or give that word a particular meaning. We explain our parsing technique and discuss about various constructions using case particles and focus particles. We show how troublesome phenomena such as scrambling and omission of case particles are treated.

1 Introduction

As the need of software modules performing natural language processing tasks is growing, in depth grammatical analyses of sentences must be properly carried out. Grammatical analyses based on theoretically sound grammar formalisms are thus essential.

Treatment of case particles constitute an essential part of a grammar for the Japanese language, where the word order is relatively flexible. The role of case particles is functionally determined within a sentence: they indicate that the accompanying noun functions as subject, object, etc. But because case components are often scrambled or omitted and because case particles disappear when case components are accompanied by the topic marker *wa* or other special particles, it makes it difficult to syntactically analyze Japanese sentences.

Various studies in the literature discuss about the Japanese argument case marking and the treatment of Japanese focus particles. Here, we explore the treatment of Japanese particles within the Lambek style pregroup grammar.

The application of pregroups in natural language processing provides a rigorous formulation of the grammar of a given language. Pregroup calculations are very simple from a computational point of view. Furthermore, in analyzing a sentence, we go from left to right and imitate the way a human hearer might proceed: recognizing the type of each word as it is received and rapidly calculating the type of the string of words up to that point.

The reader might be curious to see a comparison of our grammar formalism with other existing formalisms such as HPSG. Indeed, it would be interesting to write our proof-theoretic analysis in terms of the model-theoretic HPSG framework. We could perhaps follow the HPSG analysis of Japanese presented by Siegel in [13], where particles are analyzed as heads of their phrases and the relation between case particle and nominal phrase is a head-complement relation. To account for the omission and scrambling of verbal arguments, Siegel introduces the attributes SAT, which denotes whether a verbal argument is already saturated, optional or adjacent, and VAL, which contains the agreement information for the verbal argument. Siegel also presents a Japanese head-complement schema which accounts for optional and scramblable arguments as well as for obligatory and adjacent arguments. Due to limited space, however, page-filling representations in the HPSG framework will not be further discussed.

2   The calculus of Pregroup

The concept of pregroup has been developed as an algebraic tool to recognize grammatically well-formed sentences in natural languages [8–11]. Pregroups are a simplification of the Lambek calculus [7]. In [6], Kiślak compares the strenght of the Lambek calculus and the calculus of pregroup, and shows that syntactic analyses can be translated from one framework to the other one by means of basic translation. Furthermore, Buszkowski formally proved that grammars based on free pregroups are context-free [1].

We formally introduce the notion of pregroup [8].

Definition 1. *A pregroup is a partially ordered monoid in which each element a has a left adjoint a^l and a right adjoint a^r such that $a^l a \rightarrow 1 \rightarrow aa^l$ and $aa^r \rightarrow 1 \rightarrow a^r a$.*

Here the arrow is used to denote the order[1]. Consequences of the definition of *pregroup* are the following identities:

$$1^l = 1, \quad a^{rl} = a, \quad (ab)^l = b^l a^l, \quad aa^l a = a, \quad a^l aa^l = a^l;$$

$$1^r = 1, \quad a^{lr} = a, \quad (ab)^r = b^r a^r, \quad aa^r a = a, \quad a^r aa^r = a^r;$$

and the following implication:

$$\text{if } a \rightarrow b \text{ then } b^l \rightarrow a^l \text{ and } b^r \rightarrow a^r.$$

In linguistic applications, we work with the pregroup *freely generated* by a partially ordered set of *basic* types. From the basic types, we construct *simple types*: if a is a simple type, then so are a^l and a^r. Thus, if a is a basic type, then

$$\cdots, a^{ll}, a^l, a, a^r, a^{rr}, \cdots$$

are simple types. The compound types are strings of simple types. The only computations required are *contractions*, $a^l a \rightarrow 1, aa^r \rightarrow 1$; and *expansions*, $1 \rightarrow aa^l, 1 \rightarrow a^r a$, where a is a simple type. Expansions are not needed for the purpose of sentence verification, but only contractions combined with some rewriting induced by the partial order.

Constructing a pregroup grammar for a language consists of assigning one or more types to each word in the dictionary, and then verifying the grammaticality and sentencehood of a given string of words by a calculation on the corresponding types.

3   Analyzing Japanese grammar

We will study the pregroup freely generated by a partially ordered set of basic types for some fragments of the Japanese language[2]. To begin with, there are a number of basic types such as the following:

---

[1] Lambek originally used the '$\leqslant$' symbol to denote the order in the pregroup but since the terminology is borrowed from category theory, he later adodpted the arrow for the partial order [11].

[2] The analysis presented is based on parts of my Master's thesis [2].

π = pronoun;

$\bar{n}$ = proper name;

n = noun;

s = statement when the tense is irrelevant;

$\bar{s}$ = topicalized sentence;

s_i = statement,

      with $i = 1$ for the non perfective tense;

           $i = 2$ for the perfective tense;

c_1 = nominative complement;

c_2 = genitive complement;

c_3 = dative complement;

c_4 = accusative complement;

c_5 = locative complement.

We also postulate:

$$s_i \rightarrow s;$$
$$\bar{s} \rightarrow s;$$
$$n \rightarrow \bar{n} \rightarrow \pi.$$

To account for the free word order, we assign the type $(c_4^r, c_1^r)s_i$ to a transitive verb, and the type $(c_1^r)s_i$ to an intransitive verb. What occurs between the parentheses is optional. Furthermore, the order of the elements in the parentheses can be random.

3.1 Case particles

In (1b), the topic marker *wa* replaces the nominative case particle *ga*; *wa* is assigned the type $\pi^r c_1$, which is the type for the particle *ga*. In the example sentences given in (1), we use the partial order $n \rightarrow \pi$ to get the simplification of the type of the accusative complement.

However, we will prefer the alternative analysis in which we assign the new type $\pi^r \bar{s} s^l$ to the topic marker *wa*, as in (1c), such that the resulting sentence is of type $\bar{s}$, that is, a topicalized sentence. One of the motivation for the choice of the type $\pi^r \bar{s} s^l$ is that we can differentiate topicalized sentences from sentences; other reasons will be given in a subsequent section.

(1)   a. Watasi   ga   ringo   o   taberu.

       π $(\pi^r c_1)$ n $(\pi^r c_4)$ $(c_4^r c_1^r s_1) \rightarrow s_1$

       I   NOM apple ACC eat

     I eat an apple.

  b. Watasi   wa   ringo   o   taberu.

       π $(\pi^r c_1)$ n $(\pi^r c_4)$ $(c_4^r c_1^r s_1) \rightarrow s_1$

       I   TOP apple ACC eat

     I eat an apple.

  c. Watasi   wa   ringo   o   taberu.

       π $(\pi^r \bar{s} s^l)$ n $(\pi^r c_4)$ $(c_4^r s_1) \rightarrow \bar{s}$

        I   TOP apple ACC eat
    I eat an apple.

The sentence *Watasi ga ringo o taberu* 'I eat an apple' has several variants, all meaning the same. In (2a), the word order is changed; in (2b), the object is missing; in (2c), the subject is missing; and in (2d), both the subject and the object are missing.

The word-order flexibility and the omission of complements phenomena are tackled by assigning different types to the verb. For example, in (2a), the verb *taberu* is assigned the type $c_1^r c_4^r s_1$; in (2b), *taberu* is assigned the type $c_1^r s_1$ while in (2c), it is assigned the type $c_4^r s_1$; and finally, *taberu* is assigned the simple type s_1 in (2d).

(2) a. Ringo o watasi ga taberu.
          n $(\pi^r c_4)$ π $(\pi^r c_1)$ $(c_1^r c_4^r s_1) \rightarrow s_1$
        apple ACC   I   NOM eat
        I eat an apple.
    b. Watasi ga taberu.
          π $(\pi^r c_1)$ $(c_1^r s_1) \rightarrow s_1$
        I   NOM eat.
        I eat (an apple).
    c. Ringo o taberu.
          n $(\pi^r c_4)$ $(c_4^r s_1) \rightarrow s_1$
        apple ACC eat
        (I) eat an apple.
    d. Taberu.
          s_1
        eat
        (I) eat (an apple).

3.2 Focus particles

Japanese case particles are frequently omitted when the topic marker *wa* or a focus particle, such as *made, bakari, sae*, is added to a noun phrase. Moreover, when a sentence has a particular syntactic construction, a case particle can mark a different case than it usually does.

Various functional relations are expressed by particles in Japanese. For instance, particles such as *bakari, dake, nomi* specify focus in sentences. Focus particles bear different syntactic functions depending on where they appear in the sentence, so a Japanese parsing system needs to be able to correctly treat these particles.

In (3a), the focus particle *mo* replaces the accusative case particle *o* while in (3b), *mo* replaces the nominative case particle *ga*. The particle *mo* is therefore assigned the type $\pi^r c_4$ in (3a) and $\pi^r c_1$ in (3b) respectively.

(3) a. Watasi ga ringo mo taberu.
          π $(\pi^r c_1)$ n $(\pi^r c_4)$ $(c_4^r c_1^r s_1) \rightarrow s_1$
        I   NOM apple also eat
      I eat an apple, too.

b. Watasi mo ringo o taberu.
   π $(\pi^r c_1)$ n $(\pi^r c_4)$ $(c_4^r c_1^r s_1) \to s_1$

   I   also apple ACC eat

I, too, eat an apple.

In (4), the particle *dake* follows the proper noun *Taroo*. Assigning the type $\pi^r \pi$ to the focus particle *dake* and then using the partial order $\bar{n} \to \pi$, we can analyze the nominative complement *Taroo dake ga* and thus, (4) can be successfully parsed.

(4) Taroo dake ga ringo o tabeta.
    $\bar{n}$ $(\pi^r \pi)$ $(\pi^r c_1)$ n $(\pi^r c_4)$ $(c_4^r c_1^r s_2) \to s_2$

Taro   only NOM apple ACC eat-PAST

Only Taro ate an apple.

In (5), the particle *dake* follows the clause *Taroo wa taberu*. In this case, the type $\pi^r \pi$ for *dake* is inappropriate. We need to introduce a different type which will correspond to the particle *dake* occurring after a verb. We thus assign the type $s^r s$ to *dake*.

(5) Taroo wa taberu dake da.
    $\bar{n}$ $(\pi^r \bar{s} s^l)$ s $(s^r s)$ $(s^r s_1) \to \bar{s}$

Taro TOP eat   only be

Taro only eats.

In (6b), the focus particle *bakari* follows the gerund *tabete*. We therefore assign the type ss^l to *bakari*.

(6) a. Taroo wa ringo o tabete iru.
     $\bar{n}$ $(\pi^r \bar{s} s^l)$ n $(\pi^r c_4)$ $(c_4^r s s^l)$ $s_1 \to \bar{s}$

Taro   TOP   apple ACC eat-GERUND is

Taro is eating an apple.

   b. Taroo wa ringo o tabete bakari iru.
      $\bar{n}$ $(\pi^r \bar{s} s^l)$ n $(\pi^r c_4)$ $(c_4^r s s^l)$ (ss^l) $s_1 \to \bar{s}$

Taro   TOP   apple ACC eat-GERUND only is

Taro has been eating just apples.

In (7b), *bakari* appears in a different context; it follows the locative case particle *de*. We therefore assign the new type $c_5^r c_5$ to *bakari*.

(7) a. Taroo wa gakko de hon o yomu.
     $\bar{n}$ $(\pi^r \bar{s} s^l)$ n $(\bar{n}^r c_5)$ n $(\pi^r c_4)$ $(c_4^r c_5^r s_1) \to \bar{s}$

Taro TOP school LOC book ACC read

Taro reads a book in school.

   b. Taroo wa gakko de bakari hon o yomu.
      $\bar{n}$ $(\pi^r \bar{s} s^l)$ n $(\bar{n}^r c_5)$ $(c_5^r c_5)$ n $(\pi^r c_4)$ $(c_4^r c_5^r s_1) \to \bar{s}$

Taro TOP school LOC just book ACC read

Taro reads a book just in school.

3.3   Relative clauses

In (8a), the verb *kaita*, the past tense of *kaku*, is assigned the type $c_1^r s_2$, since the object and other complements are omitted. The resulting type of the sentence *gakusei ga kaita* is then s_2.

In (8b), the same verb form *kaita* appears in the context of a relative clause. The phrase *gakusei ga kaita* modifies the noun *ronbun* 'article', therefore, it must be of a different type than s_2. This leads us to the introduction of the new type $c_1^r n n^l$ for the verb *kaita* occurring in a relative clause.

(8) a. Gakusei   ga   kaita.
  $\quad$ n $\quad$ $(\pi^r c_1)$ $\;$ $(c_1^r s_2) \rightarrow s_2$
  $\quad$ student $\;$ NOM $\;$ wrote
  $\quad$ The student wrote (an article).

  b. [gakusei   ga   kaita]   ronbun
  $\quad$ n $\qquad$ $(\pi^r c_1)$ $\qquad\quad$ n
  $\quad$ student $\;$ NOM $\;$ wrote $\;$ article
  $\quad$ The article that the student wrote.

The type assignment of relative clause verbs is summarized in the following metarule.

Metarule 1 *Any verb of type $(c_j^r, c_k^r)s_i$ also has type $(c_j^r)nn^l$, where $i = 1, 2$ and $j, k = 1, \ldots, 5$.*

If we apply the above metarule to the verb *kaita* of the sentence (9a), we obtain the two relative clauses in (9b) and (9c).

(9) a. gakusei   ga   ronbun   o   kaita.
  $\quad$ n $\quad$ $(\pi^r c_1)$ $\quad$ n $\quad$ $(n^r c_4)$ $\;$ $c_4^r c_1^r s_1 \rightarrow s_1$
  $\quad$ student NOM article ACC wrote
  $\quad$ A student wrote an article.

  b. gakusei   ga   kaita   ronbun
  $\quad$ n $\quad$ $(\pi^r c_1)$ $\,$ $(c_1^r n n^l)$ $\quad$ n $\;$ $\rightarrow n$
  $\quad$ student NOM wrote $\;$ article
  $\quad$ The article that the student wrote

  c. ronbun   o   kaita   gakusei
  $\quad$ n $\quad$ $(\pi^r c_4)$ $\,$ $(c_4^r n n^l)$ $\quad$ n $\;$ $\rightarrow n$
  $\quad$ article $\;$ ACC wrote student
  $\quad$ The student who wrote the article

In the following two example sentences, the topic marker *wa* replaces the nominative case particle *ga*. In (10a), *sensei* is the subject of *yonda*. In (10b), *sensei* is the subject of *yonda* and the subject of *kaita* is omitted. That is, *sensei* modifies *yonda* rather than *kaita*. This grammatical phenomenon can be generalized by the following rule: *wa* cannot cause a topicalization and an omission of the nominative case particle *ga* in a relative clause [12].

(10) a. sensei   wa   ronbun   o   yonda.
  $\quad$ n $\quad$ $(\pi^r \bar{s} s^l)$ $\quad$ n $\quad$ $(\pi^r c_4)$ $\;$ $(c_4^r s_2) \rightarrow \bar{s}$
  $\quad$ teacher $\;$ TOP $\;$ article $\;$ ACC $\;$ read-PAST
  $\quad$ The teacher read an article.

b. sensei   wa   kaita   ronbun   o   yonda.

$$n \quad (\pi^r \bar{s} s^l) \quad (nn^l) \qquad n \quad (\pi^r c_4) \, (c_4^r s_2) \to \bar{s}$$

teacher TOP write-PAST article ACC read-PAST

The teacher read the article that (someone) wrote.

However, if *wa* is assigned the type $\pi^r c_1$ as first suggested, the sentence *sensei wa kaita ronbun o yonda* is not correctly parsed, as shown in (11). Under this analysis, *sensei* wrongly becomes the subject of *kaita*. In order to avoid this problem, the topic marker *wa* must be assigned the type $\pi^r s s^l$.

(11) sensei   wa   kaita   ronbun   o   yonda.

$$* \quad n \quad (\pi^r c_1) \, (c_1^r n n^l) \quad n \quad (\pi^r c_4) \, (c_4^r s_2) \to s_2$$

teacher TOP write-PAST article ACC read-PAST

The teacher read the article that (someone) wrote.

3.4   Ga-no conversion

The genitive case particle *no* usually transforms a noun into a possessive, as in *gakusei no ronbun* 'the student's article', but it can also be used as a subject marker in a relative clause. In fact, when a relative clause includes a noun phrase marked with the nominative case particle *ga*, the case can be replaced by the genitive case particle *no* without difference in meaning (see [3]).

Relative clauses may be accounted for by the following metarule:

Metarule 2 *Any verb of type $c_1^r n n^l$ has also type $c_2^r n n^l$.*

The relative clause *gakusei ga kaita* in (12a) contains the nominative case particle *ga*, therefore, we can replace the particle *ga* by the genitive case particle *no* and get the alternate relative clause *gakusei no kaita*, which is illustrated in the example (12b).

The type of the verb *kaita* in the phrase (12a) is $c_1^r n n^l$. If we then apply the above metarule, we get the alternate phrase (12b), where *kaita* has type $c_2^r n n^l$.

(12) a. gakusei   ga   kaita   ronbun

$$n \quad (\pi^r c_1) \, (c_1^r n n^l) \quad n \to n$$

student NOM wrote article

The article that the student wrote

   b. gakusei   no   kaita   ronbun

$$n \quad (\pi^r c_2) \, (c_2^r n n^l) \quad n \to n$$

student GEN wrote article

The article that the student wrote

4   Conclusion

In this paper, we proposed a parsing method based on pregroup grammar that properly treats Japanese case particles. Despite the introduction of several different types for one single word, we could successfully analyze and cover the numerous constructions with case particles and focus particles. It would certainly be even more efficient to combine our method with the use of a case frame dictionary [4].

Aknowledgments

References

1. Wojciech Buszkowski. Lambek grammars based on pregroups. In P. de Groote, G. Morrill, and C. Retoré, editors, *Logical Aspects of Computational Linguistics*, pages 95–109. Springer LNAI 2099, Berlin: Springer Verlag, 2001.
2. Kumi Cardinal. An algebraic study of Japanese grammar. Master's thesis, McGill University, Montréal, 2002.
3. S. Harada. Ga-no conversion and idiolectal variations in Japanese. *Gengo Kenkyuu*, 60:25–38, 1976.
4. Daisuke Kawahara, Nobuhiro Kaji, and Sadao Kurohashi. Japanese case structure analysis by unsupervised construction of a case frame dictionary. In *Proceedings of the 18th conference on Computational linguistics*, pages 432–438, Morristown, NJ, USA, 2000. Association for Computational Linguistics.
5. Sue A. Kawashima. *A dictionary of Japanese particles*. Kodansha International, Tokyo, 1999.
6. Aleksandra Kiślak. Pregroups versus English and Polish grammar. In V.M.Abrusci and C. Casadio, editors, *New Perspectives in Logic and Formal Linguistics, Proceedings of the Vth Roma Workshop*, pages 129–154, 2002.
7. Joachim Lambek. The mathematics of sentence structure. *American mathematical monthly*, 65:154–170, 1958.
8. Joachim Lambek. Type grammar revisited. In A. Lecompte, F. Lamarche, and G. Perier, editors, *Logical Aspects of Computational Linguistics*, pages 1–27. Springer LNAI 1582, 1999.
9. Joachim Lambek. Pregroups: A new algebraic approach to sentence structure. In C. Martín-Vide and G. Păun, editors, *Recent topics in mathematical and computational linguistics*. Editura Academei Române, Bucharest, 2000.
10. Joachim Lambek. Type grammars as pregroups. *Grammars*, 4:21–39, 2001.
11. Joachim Lambek. What are pregroups? In C. Casadio, P.J. Scott, and R.A.G. Seely, editors, *Language and Grammar: Studies in Mathematical Linguistics and Natural Language*, pages 129–136. CSLI Publications, Stanford University, 2005.
12. Hiroshi Masuichi, Tomoko Ohkuma, Hiroki Yoshimura, and Yasunari Harada. Japanese parser on the basis of the lexical-functional grammar formalism and its evaluation. In *Proceedings of the 17th Asia Pacific Conference (PACLIC-17)*, pages 298–309, 2003.
13. Melanie Siegel. HPSG analysis of Japanese. In Wolfgang Wahlster, editor, *Verbmobil:Foundations of Speech-to-Speech Translation*, pages 264–279. Springer, Berlin, Germany, 2000.
14. Natsuko Tsujimura. *An Introduction to Japanese Linguistics*. Blackwell Publishers, 1996.