Probabilistic Approach to the Iterative NP Construction in Japanese

Sumiyo Nishiguchi
Department of Management, Tokyo University of Science

nishiguchi@rs.tus.ac.jp

1 Iterative NP Construction in Japanese

In Japanese, noun phrases (NPs) can be repeated if they are "glued," or interceded with the possessive/genitive marker *no* "'s/of".

(1) a. NP_1-no NP_2

b. NP_1-no NP_2-no NP_3

c. NP_1-no NP_2-no NP_3-no NP_4

d. NP_1-no NP_2-no,...,NP_n

The "$(\text{NP-no})^n$ NP" construction is similar to the "$(\text{NP's})^n$ NP" construction in English. (2a) is structurally and semantically ambiguous in two ways as in (2b) and (2c).

(2) a. Mary's dog's scarf

b. $[_{DP}[_{DP}$ Mary's dog$][_{D'}$ $[_{D}$'s$][_{NP}$ scarf$]]]$
"A scarf of Mary's dog"

c. $[_{DP}$ $[_{DP}$ Mary$][_{D'}$ $[_{D}$'s$][_{DP}$ $[_{DP}$ dog$]$ $[_{D'}$ $[_{D}$'s$][_{NP}$ scarf$]]]]]$
"Mary's scarf which is the kind that dogs usually wear"

1.1 Structural Ambiguity

The multiple sequences of NPs produce syntactic and semantic ambiguities. NP_n may modify either the following NP_{n+1} or the NP_{n+2} which follows after.

(3) a. NP_1-no NP_2-no NP_3

b. heya-no akari-no suicchi
room-GEN light-GEN switch
"the room light switch

[[POSS P NP]POSS] NP]

c. nyushi-no sansu-no mondai
entrance exam-GEN math-GEN question
"a math problem for the entrance exam"

[POSS P [POSS P [NP]]

(4) a. NP_1-no NP_2-no NP_3-no NP_4

b. Aren-no oku-no eiga-no seisaku
Allen-GEN many-GEN movie-GEN production
"production of many Allen's movies"

[[POSS P [POSS P NP]]POSS NP]

c. niwa-no ajisai-no murasakiiro-no tsubomi
garden-GEN hydrangea-GEN purple-GEN bud
"the purple buds of hydrangeas in the garden"

[[POSS P NP]POSS [POSS P NP]]

(5) a. NP_1-no NP_2-no NP_3-no NP_4-no NP_5

b. jidosha-no shiyo-no honkyo-no bunpu-no jotai
car-GEN use-GEN base location-GEN distribution-GEN condition
"the situation with distribution of base locations for car use"

[POSS P [POSS P [POSS P [POSS P NP]]]]

Therefore, the syntactic and semantic modification have two patterns. NP_n may modify either the following NP_{n+1} or the ones after. The modification relation can be represented as below, when R is a modification relation and m and n are natural numbers:

(6) a. $R(NP_n,NP_{n+1})$

b. $R(NP_m,NP_{n>m+1})$

1.2 Syntactic Trees

The two patterns of modification are syntactically represented in Figure 1 and 2 respectively. The trees consider NPs to be DPs (determiner phrases), following the DP Hypothesis [1]. [1] When all "NP_n-no" or

[1] While [2] considers the *'s* morpheme in English to be a clitic and the D head as a zero determiner with possessive feature $\emptyset_{[poss]}$, *no* in Japanese does not appear to be a clitic.

POSS Ps modify the following NP_{n+1}, the preceding DPs are in the specifier position of the following DP in (7) whose tree is Figure 1. On the other hand in (8), the tree of which is Figure 2, the multiple possessive phrases DP-adjoin to the previous possessive DPs. The possessive phrases are considered to be full DPs with null NP, which is deleted and phonologically unpronounced.

(7) Jon-no heya-no denki-no suicchi
Jon-GEN room-GEN light-GEN switch
"The switch of the light in Jon's room"

R(Jon, room) ∧ R(room, light) ∧ R(light, switch)

(8) nyushi-no sugaku-no bekutoru-no mondai
entrance exam-GEN math-GEN vector-GEN question
"a math question on vectors for the entrance exam"

R(entrance exam, question) ∧ R(math, question) ∧ R(vector, question)

2 Data Analysis

I have manually classified the 128 instances of *NP_1-no NP_2-no NP_3* construction "NP_1-GEN NP_2-GEN NP" in Chunagon[2] by searching the sequence *NP_1-no NP_2-no NP_3*.

Out of 128 instances, there were 116 instances that NP_1 modified NP_2 whereas 12 instances were such that NP_1 modified NP_3 and also NP_2 modified NP_3.

modification type	instances	
R(NP_n, NP_{n+1})	115	89.8%
R(NP_n, NP_{n+2})	13	10.2%
total	128	

Figure 3: Distribution of Two Types of Modification

Among R(NP_n, NP_{n+2}), there were several cases of "NP_1-no NP_2-no NP_3-no NP_4" construction, in which the NP occurred more than three times.

modification type	n	instances		total
R(NP_n, NP_{n+1})	3	111	96%	115
	4	3	2.6%	
	5	1	0.8%	
R(NP_n, NP_{n+2})	3	5	38%	13
	4	8	61%	
	5	0	0%	

Figure 4: Modification Type and the Number of NPs

[2]The search engine for the *Balanced Corpus of Contemporary Written Japanese*, https://chunagon.ninjal.ac.jp/

3 Probabilities

3.1 Disjoint events

Clearly, R(NP_n, NP_{n+1}) and R(NP_n, NP_{n+2}) are disjoint events, that is, the distribution of the two events is complementary. Therefore, the sum of their probabilities is 1.

(9) p(R(NP_n, NP_{n+1}) + p(R(NP_n, NP_{n+2})= 1

modification type	instances	probability
R(NP_n, NP_{n+1})	115	0.898
R(NP_n, NP_{n+2})	13	0.102
total	128	1

Figure 5: Disjoint Events

3.2 Joint Probability

The joint probability of the number of NP sequences and the type of modification can be calculated by multiplying the probabilities of each. According to Figure 6, the highest probably pattern is "NP_1-no NP_2-no NP_3" whose NP_1 modifies NP_2 and NP_2 predicates NP_3.

(10) p(n, R) = p(n) × p(R)

p(n, R)	instances	probability
p(3, R(NP_n, NP_{n+1}))	111	0.867
p(4, R(NP_n, NP_{n+1}))	3	0.023
p(5, R(NP_n, NP_{n+1}))	1	0.007
p(3, R(NP_n, NP_{n+2}))	5	0.039
p(4, R(NP_n, NP_{n+2}))	8	0.062
p(5, R(NP_n, NP_{n+1}))	0	0
total	128	1

Figure 6: Joint Probabilities

n	instances	probability
3	116	0.906
4	11	0.086
5	1	0.008
total	128	1

Figure 7: The Number of NPs and Probabalities

3.3 Conditional Probability

If n = 3, the conditional probability of R(NP_n, NP_{n+1}) is:

(11) $p(R(NP_n, NP_{n+1}|n=3)$
$= \frac{p(n=3, R(NP_n, NP_{n+1}))}{p(n=3)}$
$= \frac{0.867}{0.906}$
$= 0.956$

Figure 8 presents conditional probabilities for the modification types given the number of NPs.

p(n, R)	instances	probability
$p(R(NP_n, NP_{n+1}) \mid 3)$	111	0.956
$p(R(NP_n, NP_{n+1}) \mid 4)$	3	0.267
$p(R(NP_n, NP_{n+1}) \mid 5)$	1	0.875
$p(R(NP_n, NP_{n+2}) \mid 3)$	5	0.039
$p(R(NP_n, NP_{n+2}) \mid 4)$	8	0.043
$p(R(NP_n, NP_{n+1}) \mid 5)$	0	0
total	128	1

Figure 8: Conditional Probability

4 Summary

This paper analyzed structural and semantic ambiguities of the multiple NP sequences interceded with the possessive/genitive marker in Japanese. The structural ambiguities were represented syntactically. The corpus data drew the joint and conditional probabilities.

Acknowledgement

I thank the Workshop on Bayesian Natural Language Semantics and Pragmatics at ESSLLI2013 which gave me an opportunity to share the basic ideas of this paper during the discussion period.

References

[1] Steven Paul Abney. *The English Noun Phrase in its Sentential Aspect*. PhD thesis, MIT, 1987.

[2] Chris Barker. *Possessive Descriptions*. CSLI Publications, Stanford, 1995.

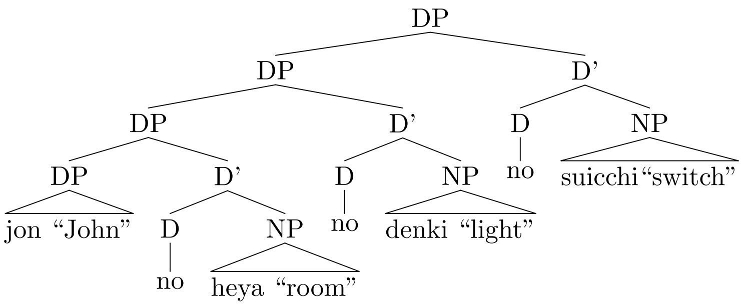

Figure 1: (7)

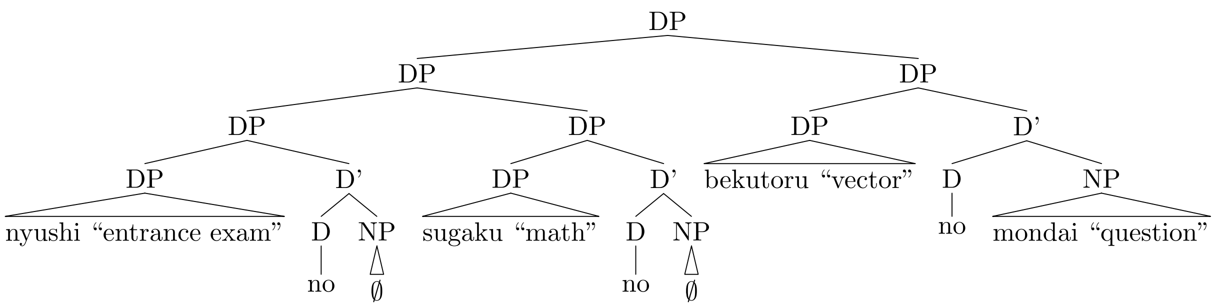

Figure 2: (8)