Corpus-Based Investigation of the Markedness and Frequency of Japanese Passives in Contemporary Written Japanese

Tatsuya Aoyama
Georgetown University
ta571@georgetown.edu

Abstract

Japanese passives are traditionally considered to have two types: direct and indirect passives. However, more recent studies, such as Ishizuka (2012), suggest the two types can be unified under the same syntactic movement analysis. Utilizing the Balanced Corpus of Contemporary Written Japanese (BCCWJ; Maekawa, 2008; Maekawa et al., 2014), this study aims to investigate how likely different types of passives appear in the naturally occurring texts, especially in relation to markedness-based hierarchy called Noun Phrase Accessibility Hierarchy (NPAH; Keenan and Comrie, 1977), and to investigate if true indirect passives occur in contemporary written Japanese.

1 Introduction

1.1 Types of Japanese Passives

Japanese allows for a construction called indirect passive, which roughly means the use of passive voice with intransitive verbs (i.e., no internal argument), as exemplified by the difference between (1) and (2) (Ishizuka, 2012):

(1) a. Ken-ga keisatu-ni tukamae-rare-ta
Ken-SUB police-DAT catch-PASS-PST
Ken was caught by the police.

b. Keisatu-ga Ken-o tsukamae-ta
police-SUB Ken-ACC catch-PST
The police caught Ken.

(2) a. ?Naomi-ga Ken-ni oyog-are-ta
Naomi-SUB Ken-DAT swim-PASS-PST
Ken went and swam on Naomi.

b. Ken-ga *(Naomi-o) oyoi-da
Ken-SUB Naomi-ACC swim-PST
Ken swam Naomi.

In (1), the nominative NP *Ken-ga* in the passive sentence (1a) has a clear source in its active counterpart (1b), which is *Ken-o*. Therefore, (1a) can be called a passive construction with an *-o* marked accusative NP as its active source. On the other hand, in (2), the nominative NP *Naomi-ga* in the passive sentence (2a) has no source in its active counterpart, and its grammaticality is controversial. This type of passive construction is called indirect passive, whose nominative NP does not have a source in its active counterpart.

Traditionally, these two types of passive have been treated separately as direct and indirect passives (e.g., Hoshi, 1999). However, Ishizuka (2012) showed that the seemingly absent source of the nominative NP in an indirect passive is in fact present, and that it is marked as either one of the following three grammatical Cases in the active sentence: *-ni* marked dative Case, *-kara* marked ablative Case, or *-no* marked genitive Case, which corresponds to following examples (3), (4), and (5), respectively:

(3) a. Ken-ga Naomi-ni raburetaa-o
Ken-SUB Naomi-DAT raburetaa-ACC
watas-are-ta
give-PASS-PST
Ken was given a love letter by Naomi.

b. Naomi-ga **Ken-ni** raburetaa-o
Naomi-SUB **Ken-DAT** raburetaa-ACC
watas-ta
give-PST
Naomi gave a love letter ***to Ken****.*

(4) a. Ken-ga Naomi-ni niger-are-ta
Ken-SUB Naomi-DAT escape-PASS-PST
Ken was escaped from by Naomi.

b. Naomi-ga **Ken-kara** niger-ta
Naomi-SUB **Ken-ABL** escape-PST
Naomi escaped ***from ken****.*

(5) a. Ken-ga hahaoya-ni shin-are-ta
Ken-SUB mother-DAT die-PASS-PST
Ken was died on by his mother.

b. **Ken-no** hahaoya-ga shin-da
Ken-GEN mother-SUB die-PST
Ken's *mother died.*

In this analysis, a passive sentence whose matrix

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 403-407.
Amherst, Massachusetts, June 15-17, 2023

NP cannot be analyzed as any one of the abovementioned four Cases (accusative *-wo*, dative *-ni*, ablative *-kara*, and genitive *-no*), must be ungrammatical. Indeed, Ishizuka (2012) claims that (2a) is ungrammatical because the nominative NP *Naomi-ga* cannot exist in the active source (2b). In other words, changing the Case of *Naomi-o* to *-ni*, *-kara*, or *-no* does not make (2b) any less ungrammatical. This unified analysis (i.e., that all passives have active source) leaves us with a question of whether or not it is supported by empirical evidence. Although Ishizuka (2012) provides the result of grammaticality judgement tests (GJTs) to support her claim, artificial GJTs only measure the perception grammar and not the production grammar, which are arguably different from each other. Hence, this warrants an investigation of corpora of naturally occurring texts.

In hypothesizing the occurrences of various passives, it is unreasonable to assume that all passives discussed above occur equiprobably because so-called indirect passives seem to be more marked than direct passives, as is clear from the fact that indirect passives are not allowed in many other languages, including English (e.g., * Naomi was died by her mother). The next section will discuss the markedness of various passives, on which the predicted frequency order will be based.

1.2 Noun Phrase Accessibility Hierarchy

In determining which passives are expected to occur more in a corpus, we focus on the similarity between Japanese passive construction and relative clause construction. Consider the following examples:

(6) a. Otoko-no hahaoya-ga shin-da
Man-GEN mother-SUB die-PST
A man's mother died.

b. Otoko-ha hahaoya-ni shin-are-ta
Man-TOP mother-DAT die-PASS-PST
A man was died by his mother.

c. Hahaoya-ga shin-da otoko
Mother-SUB die-PST man
A man whose mother died.

The passivization from (6a) to (6b) and the relativization from (6a) to (6c) has an interesting property in common: the grammatical Case of the source NP is masked after the alteration. More concretely, the *-no* marked genitive NP *otoko* (man) is no longer *-no* marked in either (6b) or (6c), obscuring the source. Ishizuka (2012) cogently puts this point:

> The absence of original Case under movement is in fact a general property of Japanese; it is also operative in relative constructions (...). This property obscures the source of the nominative NP in the passive and has often misled linguists to conclude that the indirect passive is gapless (p. 8).

Based on this common property, a framework used for relativization could be extended to passivization.

Keenan and Comrie (1977) proposed Noun Phrase Accessibility Hierarchy (NPAH), a hierarchy of relative clause construction based on cross-linguistic markedness, which is reproduced below:

SU > DO > IO > OBL > GEN > OCOMP

SU, DO, IO, OBL, GEN, and OCOMP stands for subject, direct object, indirect object, oblique, genitive, and object of comparison, respectively. As has been discussed in the previous section, the source of nominative NP in passives is either DO (*-o* marked direct passive), IO (*-ni* marked indirect passive), OBL (*-kara* marked indirect passive), and GEN (*-no* marked indirect passive). Although true indirect passives with no source NP do not exist purportedly, if it does at all, its frequency ranking should be after OCOMP in the above hierarchy. Table 1 summarizes the parts of the hierarchy relevant to this study:

Rank	NPAH	Case	Examples	
1	DO	*-o*	*nagur-*	(to punch)
	DO	*-o*	*homer-*	(to praise)
2	IO	*-ni*	*watas-*	(to pass)
3	OBL	*-kara*	*niger-*	(to escape)
4	GEN	*-no*	*shin-*	(to die)
5	∅	∅	*nak-*	(to cry)
	∅	∅	*ner-*	(to sleep)
	∅	∅	*aruk-*	(to walk)

Table 1: Expected Frequency Ranking of Japanese Passive Types based on NPAH.

1.3 Research Questions and Hypotheses

In light of all these, the following research questions and hypotheses were formulated:

1. What is the relationship between the grammatical Case of a source NP and the frequency of its passivization?

(a) It is hypothesized that, the relative frequency of passivization for each verb follows NPAH proposed by Keenan and Comrie (1977).

2. Are there any truly indirect passive constructions, whose matrix NPs do not have active sources?

(a) It is hypothesized that, based on the unified analysis by Ishizuka (2012), no true indirect passives exist.

2 Methodology

This section introduces the corpus used in this study, data extraction, and data annotation.

2.1 Corpus

In this study, Balanced Corpus of Contemporary Written Japanese (BCCWJ; Maekawa, 2008; Maekawa et al., 2014) was used. BCCWJ is a richly annotated 104.3 million words corpus, freely available to anyone upon registration. It contains written Japanese from various genres, including books, magazines, newspapers, blogs, online bulletin boards, textbooks, and laws.

2.2 Data Extraction

Because no semantic annotation is included in the corpus, relevant matches were searched based on the lemma of the verbs and suffixes. In Japanese, depending on the verb's conjugation type, one of the two passivization suffixes, *-reru* or *-rareru* is attached to the verbs with imperfective form. For example, to search for all occurrences of passivized verb *naguru* (to punch), lemma was specified as *-reru*, preceded by a verb with lemma *naguru*. For verbs of a different conjugation type, such as *neru* (to sleep), lemma was specified as *-rareru*, preceded by a verb with lemma *neru*.

2.3 Data Annotation

Because Japanese morphology is highly polysemous, the abovementioned passivization suffixes, *-reru* and *-rareru*, have several different meanings, such as honorific, potential, and passive. For example, an imperfective form of the verb *taber-* (to eat) followed by the passivization suffix *-rarer*, which combines to *taberareru*, can mean (1) to be able to eat, (2) (honorific) to eat, and (3) to be eaten, and these meanings can only be disambiguated by looking at the contexts. Therefore, once all the matching occurrences are exported to a spreadsheet, each occurrence was manually checked by a Japanese native speaker (i.e., the author) for the semantic disambiguation.

Besides the disambiguation of the suffixes, the presence and absence of a *-ni* phrase was annotated for each occurrence of passive constructions. This was a rather straightforward task for intransitive and transitive verbs; however, for ditransitive verbs, the first task was to decide which one of the two objects is passivized, and the second task was to decide if corresponding postpositional phrases were present or not. For example, an active sentence with a ditransitive verb *watas-* (to pass), typically takes the following construction (7a), and passivizes into either (7b) or (7c):

(7) a. Naomi-ga hon-o Ken-ni
Naomi-SUB book-ACC Ken-DAT
watas-ta
pass-PST
Naomi passed the book to Ken.

b. Hon-ga Naomi(-kara|-ni) Ken(-he|-ni)
book-SUB Naomi-ABL|DAT Ken-DAT
watas-rer-ta
pass-PASS-PST
The book was passed (from|by) Naomi to Ken.

c. Ken-ga Naomi(-kara|-ni)
Ken-SUB Naomi-DAT
hon-wo watas-rer-ta
Ken-ABL|DAT pass-PST
Ken was passed a book (from|by) Naomi.

Because only the passive sentences whose matrix NP originated from the *-ni* marked dative Case NP in the active sentence qualify for IO in NPAH, only the passivization in (7c) was counted.

On a related note, *-no* marked NPs (GEN category in Table 1) have to be annotated with a special care, making sure that a sensible relationship can be naturally established between the two NPs. For example, the grammaticality of (5b) is licensed by the fact that the two NPs (i.e., Ken and mother) have a natural *-no* markable relationship (i.e., mother-son relationship). Hence, for GEN category, only the passive sentences that satisfy this criterion were counted. Conversely, for ∅, only the passive sentences where no natural *-no* markable relationship can be established between the two NPs were counted, since they would otherwise belong to GEN category.

3 Results

3.1 Research Question 1

The summary of the absolute and relative frequencies of passivization and the presence of -ni phrase for each verb lemma is shown in Table 2.

lemma	all	pass.	%	-ni	%
nagur-	1780	656	36.9	146	22.3
homer-	2381	560	23.5	182	32.5
watas-	5041	212	4.2	60	28.3
niger-	5930	208	3.5	81	38.9
shin-	16081	101	0.6	85	84.2
nak-	7071	69	1.0	13	18.8
ner-	10679	5	0.05	0	0
aruk-	18130	6	0.03	0	0

Table 2: Absolute and Relative Frequencies of Passivization and -ni Phrase for each Verb Lemma. The columns correspond to all occurrences of the lemma (all), frequency of passivization (pass.), pass. / all (%), passivization with an accompanying -ni phrase (-ni), and -ni / pass. (%), respectively.

The first % column represents the proportion of passive sentences of all occurrences of the corresponding verb lemma in the row. The second % column represents the proportion of passive sentences with -ni phrase out of all the passive sentences of the corresponding verb lemma in the row. As shown above, the relative frequency of passivization differs substantially among verb lemmas. To test if these differences are significant, a series of pairwise proportion tests were conducted. To correct for the multiple testing, Bonferroni correction, one of the more conservative correction methods, was employed. Table 3 summarizes the results.

		1	2	3	4	5	6	7
1	***nagur-***	-	-	-	-	-	-	-
2	***homer-***	*	-	-	-	-	-	-
3	***watas-***	*	*	-	-	-	-	-
4	***niger-***	*	*	*	-	-	-	-
5	***shin-***	*	*	*	*	-	-	-
6	***nak-***	*	*	*	*	.16	-	-
7	***ner-***	*	*	*	*	*	*	-
8	***aruk-***	*	*	*	*	*	*	1.00

Table 3: Pairwise Proportion Tests. Note: * indicates *p* < .0001

The above table of *p* values show that the relative frequencies of passivization are significantly different between all possible pairs of verbs after Bonferroni correction, except for the differences between verbs *nak-* (to cry) and *shin-* (to die), and between *aruk-* (to walk) and *ner-* (to sleep). To determine what these differences mean in terms of NPAH, verbs are grouped into NPAH categories in Figure 1.

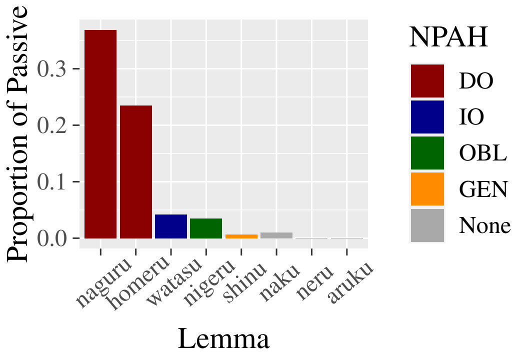

Figure 1: Relative Frequencies of Passivization for each Verb Lemma.

In Figure 1, the relative frequency ranking follows the NPAH hierarchy almost perfectly. That is to say, the following relationship was empirically observed: DO > IO > OBL > GEN > None. The only deviation is the unexpectedly frequent passivization of verb *nak-*. Although this will be further discussed in the next section, given that the differences between *shin-* (GEN) and *ner-* (None) and between *shin-* (GEN) and *aruk-* (None) were statistically significant (see Table 3), it is reasonable to say that the inequality GEN > None was maintained in general, with the verb *nak-* deviating from the rest of None category.

Similarly, the significant difference within the DO category (i.e., *nagur-* and *homer-*) was unexpected; however, it could be due to the adversative connotation that commonly accompanies passivization, which is clearly more compatible with *nagur-* (to punch) than with *homer-* (to praise).

3.2 Research Question 2

Especially noteworthy in Table 2 is the occurrences of passivized verbs of None category; namely, passives whose matrix NP does not have a source in their active counterpart. The equally infrequent passivization of the verbs *neru* and *aruku* (p = 1), imply that they indeed belong to the same category (which we call None). Although they are both extremely unlikely, the fact they both occur in naturally occurring texts must be accounted for. Also, the proportion of *-ni* phrase in passive sentences for None type verb, presented in Figure 2, indicates its distinctive property.

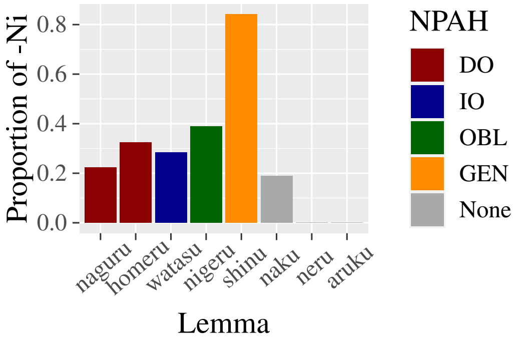

Figure 2: Relative Frequencies of -ni Phrase in Passive Sentences for each Verb Lemma.

For DO, IO, and OBL type passives, the proportion of sentences with *-ni* phrase ranges from about 20% to 40%. This optional nature of *-ni* phrase does not seem to hold for GEN type passives, as the proportion exceeds 80%. This is presumably due to the fact that omitting a *-ni* phrase in GEN type passive results in a floating genitive marked NP in its active counterpart:

(8) a. Ken-ga (hahaoya-ni) shin-are-ta
Ken-SUB mother-DAT die-PASS-PST
Ken was died by his mother.

b. Ken-no (hahaoya-ga) shin-da
Ken-GEN mother-SUB die-PST
Ken's mother died.

Omitting the parenthetical *-ni* phrase in the passive sentence (8a) is equivalent to omitting the parenthetical nominative *-ga* phrase in the active counterpart (8b), which results in the clearly ungrammatical floating genitive *-no* marked phrase. This is one plausible explanation of the strong preference to attach a *-ni* phrase in GEN type passive sentences.

Given this analysis, the low proportion of sentences with *-ni* phrase for None type passives seems to suggest that they constitute a unique category different from GEN type. The None type passive is indeed indirect in that they have no active source, as shown below in the actual example taken from the corpus (ID: LBl7_00047):

(9) a. Taako-ni-ha mou-ichi-do
Taako-DAT-TOPIC another-one-time
nak-are-ta-koto-ga aru.
cry-PASS-PST-event-SUB exist.
There is another occasion where I was cried by Taako.

In this indirect passive, the omitted subject (i.e., I) does not have a source in its active counterpart; in other words, there is no possible Case (neither *-wo*, *-ni*, *-kara*, nor *-no*) that the omitted subject (i.e., I) can be marked with, such that its active sentence can be grammatically formed.

4 Conclusion

This paper has shown that the relative frequency of passivization follows the NPAH proposed by Keenan and Comrie (1977), and that although extremely infrequent, true indirect passives do occur in naturally occurring written Japanese, as opposed to the claim by Ishizuka (2012).

However, given that the sample size is rather small in this study, and that only 8 verbs were considered, the results have to be interpreted with caution. For example, the observed differences could have been due to the unique semantic properties of each verb, rather than the larger syntactic categories verbs were coerced into (e.g., IO). To tackle this problem, future studies could annotate a wider range of verbs, and compare the within-class differences and the between-class differences to test if NPAH overrides the differences among individual verbs.

References

Hiroto Hoshi. 1999. Passives. In Natsuko Tsujimura, editor, *The handbook of Japanese linguistics*, pages 191–235. Blackwell.

Tomoko Ishizuka. 2012. *The passive in Japanese: A cartographic minimalist approach*, volume 192. John Benjamins Publishing.

Edward L Keenan and Bernard Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic inquiry*, 8(1):63–99.

Kikuo Maekawa. 2008. Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the 6th Workshop on Asian Language Resources*.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written japanese. *Language resources and evaluation*, 48:345–371.