Mathematical Linguistics, Vol.31 No.1 (June 2017) pp.1-12. 〔Times New Roman 9pt, left-aligned〕

Paper B

A quantitative approach to addressee-honorific markers
Identification of crucial independent variables and prototypes

Akitaka Yamada
Georgetown University

Abstract: This paper quantitatively examines the alternation between the two negative addressee-honorific constructions in contemporary Japanese; *-mas-en* (the prescriptive form) and *nai desu* (the new form). Based on the interpretation of the best statistical model, this paper makes two novel observations. First, presence of the epistemic suffix *-yoo* is the strongest in effect size, favoring the new form. Second, although previous studies hypothesized that stative predicates (*e.g.*, *wakar-* 'understand' and *deki-* 'can') are prototypes for the new form (Noda 2004; Kawaguchi 2014), they do not show a systematic, strong preference when important fixed effects are all controlled. Instead, this paper identifies the prototypes for each construction by looking at the estimated value of the random effect, revealing that *-tai* 'want' is a prototype of the *nai desu* form and *nega-e-* 'can wish' is a prototype of the *-mas-en* form.

Keywords: corpus linguistics, cognitive linguistics, usage-based linguistics, generalized linear mixed effect models, addressee-honorific constructions, prototypes, schema structures

1. Introduction

Since the influential study of Rosch (1975), importance of prototypes has been acknowledged in linguistics. But identifying prototypical examples is not an easy task in practice. Researcher's intuition, though giving a good approximation in some cases, cannot escape from their (unknown) biases, which should make other researchers wonder how prototypical or generalizable the alleged 'prototypical' examples are. In pursuit of representative examples, researchers collect sentences from corpora and such corpus-aided/driven studies have been appreciated within the tradition of usage-based communities (Bybee 2010; Taylor 2012). However, in many cases, a corpus is treated as an unstructured storage of real data and researchers sometimes arbitrarily cite examples to justify their theory without a serious consideration on how prototypical the cited examples are. For better usage-based studies, we need to discuss the degree of prototypicality of the cited example as objectively as we can.

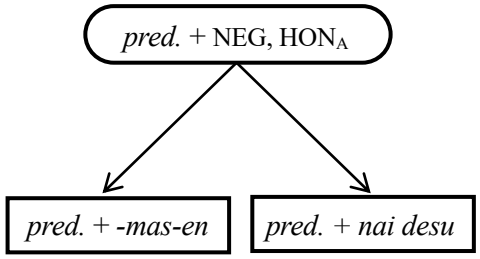
This study demonstrates how statistical models help us identify prototypical examples by examining the variation of negative addressee-honorific constructions in contemporary Japanese. For example, observe the two sentences in (1) and (2). Japanese is equipped with two grammatical patterns when the addressee-honorification and the negation cooccur within the same sentence. In the paradigm of cognitive linguistics, they are analyzed

as two instances of the same construction scheme; that is, the scheme [*akai* NEG, HON_A] 'red NEG, HON_A' gets realized either as [*akaku ari-mas-en*] or as [*akaku nai desu*]. The relation between the instances and the scheme is graphically represented as in (3) (Langacker 2008, among others); here, a comma is used to refer to a scheme whose linear precedence is left underspecified. Every time the speaker wants to instantiate the scheme, s/he has to pick up one of the two forms.

(1) *Akaku ari-**mas-en**.*
red be-HON_A-NEG
'(it is) not red; polite.' [Prescriptive form]

(2) *Akaku **nai** **desu**.*
red NEG HON_A
'(it is) not red; polite.' [New form]

(3)



Previous studies have reported several factors relevant for the selection of the two variants; *e.g.*, the part-of-speech of the main predicate (the *pos*-effect) and the presence of sentence-final particles (the *sfp*-effect) (Tanomura 1994; Fukushima and Uehara 2003; Noda 2004; Ozaki 2004; Kawaguchi 2010, 2014; Ochiai 2012; Banno 2012). In reexamining such influential factors, this study reconsiders prototypical examples of each variant. It has been said that the prescriptive form is preferred if the predicate that precedes [NEG, HON_A] is a verbal phrase. But, not all the verbs are as strongly tied to one form. Our goal is to quantitatively identify prototypical examples that exhibit a strong association with one of the construction schemes even after general factors such as the *pos*/*sfp*-effects are controlled.

In Section 2, this study summarizes the findings of previous studies. After making some preliminary remarks on the research design (Section 3), this paper builds and compares different statistical models in order to discuss the difference in effect size among these factors and quantitatively identify the prototypes of each construction (Section 4 and 5). The discussion is wrapped up in Section 6 with some remarks on future studies.

2. Previous studies

Previous studies have revealed different linguistic and/or sociolinguistic factors that affect the choice between the two variants. First, *the register* or *the formality* has been considered as an important factor; when the situation is very formal, the prescriptive form is preferred. The difference between the written document and the spoken register has also been discussed (Tanomura 1994; Noda 2004; Ochiai 2012; Banno 2012). Especially, Ochiai (2012) and Banno (2012) quantitatively examined the effect of register, taking advantage of the annotation in BCCWJ. But they did not take into account all the registers annotated in this corpus; instead, they arbitrarily selected a subset of the possible registers. It is, thus, necessary for us to examine wider registers to see how generalizable their findings are. Second, *the part-of-speech of the preceding element* is also considered as an indispensable factor. Previous studies have agreed that adjectives and nouns are used more easily with the new variant while verbs tend to take the prescriptive form (Tanomura 1994; Noda 2004; Kawaguchi 2010, 2014). Third,

	formality	prec.cat	sfp	eventuality	speech act
-mas-en	formal	V	w/o sfp	action V	command/invitation
nai desu	informal	A&N	with sfp	stative V (*wakar-/deki-*)	others

Table 1 Factors discussed in previous studies.

presence of a sentence-final particle (SFP) makes the new variant more likely to be produced (Tanomura 1994; Noda 2004; Ozaki 2004; Kawaguchi 2010, 2014). Tanomura (1994) observes that this effect is only salient for nouns and adjectives. Noda (2004), however, reports that the sentence-final particle effect is also observed with verbs. Fourth, some researchers have proposed that *the eventuality of the verb* affects the variable selection (Noda 2004; Kawaguchi 2014). They argue that stative verbs are used more easily in the new variant. For example, *wakar-* 'understand' and *deki-* 'can' have been treated as clear 'prototypes' of the *nai desu* form. However, not all the researchers agree with this conclusion (Ochiai 2012). Since previous studies have not built any statistical models, it is difficult for us to discuss spurious effects among the factors. The reason why *wakar-* seems to go along with the *nai desu* form is perhaps due to other correlated factors; for example, it may be because this verb tends to appear with SFP and, due to this correlation, *wakar-* superficially selects the new variant to a great degree. Of course, it may also be the case that *wakar-* intrinsically gravitates to the new variant. But we cannot clinch the argument until we estimate the effect of each individual word and the other possible factors at the same time. Fifth, *the speech act* of the sentence is discussed as an important factor. For example, sentences with a speech act of invitation do not fit with the new variant (Tanomura 1994; Ozaki 2004). Ozaki (2004) also observes that *nai desu* forms are reluctant to appear when making a command or an invitation. Table 1 summarizes the findings in previous studies.

None of these studies has proposed any concrete statistical models, resulting in the following problems. First, it is difficult to identify the difference in effect size among the factors. Maybe, grammatical effects are larger than sociolinguistic effects or vice versa. The lack of estimated effect sizes prevents us from discussing issues like this. Second, as mentioned above, the risk of having a spurious factor has not been considered at all. Apparent effects of some factors may be attributed to some correlated predictors. Finally, the arguments in the previous studies rely on the descriptive statistics, which makes it difficult for us to make a quantitative inference on the population from which the data is assumed to be extracted.

3. Data

3.1 Corpus

This study uses the Balanced Corpus of Contemporary Written Japanese (BCCWJ) as its primary source of data, for the following three reasons. First, it is released for free, so interested readers can easily replicate the process or propose a better model if they want to criticize this current study. Second, this is the largest annotated corpus for contemporary Japanese. When we build complicated models, the accuracy of the estimation may not be guaranteed with a small data set. Third, this corpus allows us to compare different genres,

making it easier for us to examine sociolinguistic effects.

The data was accessed through Chunagon, the web interface for BCCWJ, on 07/29/2018. With the following formulae, 176,899 examples are extracted from the corpus. We will restrict ourselves to a subset of these examples as discussed shortly below.

(4) a. *key*: (*lexeme* = "無い" OR *lexeme* = "ない") AND *the following context*: *lexeme* = "です" ON 1 WORDS FROM key [31,016 examples]

b. *key*: (*lexeme* = "無い" OR *lexeme* = "ない") AND *the following context*: *lexeme* = "た" ON 1 WORDS FROM *key* AND *the following context*: *lexeme* = "です" ON 2 WORDS FROM *key* [1,530 examples]

c. *key*: *lexeme* = "ます" AND *the following context*: *lexeme* = "ず" ON 1 WORDS FROM *key* [144,353 examples]

3.2 Variables examined in this study

A. *variant* (the outcome variable). This is the outcome variable, which can take two values; *i.e.*, the baseline prescriptive form (*-mas-en*; = 0) and the new variant (*-nai desu*; = 1).

B. *tns* (the presence of the past tense suffix). This predictor variable is categorical and takes two levels, *i.e.*, *the present form* (*the baseline form*; = 0) and *the past tense* (= 1).

C. *yoo* (the epistemic modal suffix). This binary variable indicates presence or absence of *the epistemic modal morpheme -yoo*, as in (5). If it is present, 1; 0 otherwise.

(5) a. [[*aruki-mas-en*] ***des-yoo***]. b. [[*aruka-nai*] ***des-yoo***].

walk-HON$_A$-NEG COP. HON$_A$-EPI walk-NEG COP.HON$_A$-EPI

'It is likely that (he) will not walk.' 'It is likely that (he) will not walk.'

D. *hon* (honorifics). This variable indicates whether one of the following expressions in (6) is used within the same sentence (Kikuchi 1997). Though Japanese has many other honorific markers, the honorific elements in (6) are all and the only expressions we have under the restrictions we will discuss in Section 3.3. If an honorific expression is present, 1; 0 otherwise.

(6) a. [Subject honorifics] *irassyar-* 'go, come,' *kudasar-* 'give,'

b. [Object honorifics] *itadak-(e-)* '(can) receive,' *mair-* 'come, go,' *moosiage-* 'say,'

c. [Teicho-go] *zonzur-* 'think, know,' *zonzi-* 'be knowing,' *itas-* 'do,' *or-* 'be,' *moos-* 'say'

E. *sfp* (sentence-final particle). This variable indicates whether the variant is followed by a sentence final particle (*e.g.*, *-ne* and *-yo*). The conjunctive particle (*i.e.*, *setuzoku zyosi*) is also treated as a different level, resulting in three categorical levels (0: none; 1: sentence-final particles; 2: conjunctive particles).

F. *int* (the interrogative particle). This predictor indicates whether the sentence is used with an interrogative particle *-ka* or not. If it is present, 1; 0 otherwise.

G. *reg* (the register). This variable refers to the register from which the sentence is retrieved. In BCCWJ, the following 12 registers are identified; *magazines*, *Yahoo blogs*, *Yahoo news papers*, *Chiebukuros*, *the Diet proceedings*, *poems*, *published books*, *library books*, *white papers*, *best-sellers*, *textbooks* and *PR magazines*. This study treats them as a random variable, because the register does not form a closed class.

H. *prec.word* (the preceding word) and *prec.cat (*the part-of-speech of the prec.word). It is

likely that some preceding words have positive or negative effects on the variable selection. To examine such effects, this study has two variables, *prec.word* (the preceding word) and *prec.cat* (the category/part-of-speech of the preceding word). For example, in (7), the variant *-mas-en* is preceded by the verb *aruki*. So, the *prec.word* is *aruk-* 'to walk' and the *prec.cat* is a 'verb.' In addition to verbs, *na*-adjectives, *i*-adjectives, nouns and auxiliaries are taken into account.[1,2] The *prec.word* is treated as a random effect, because *prec.words* form an open class, while the *prec.cat* is treated as a fixed effect because parts-of-speech form a closed class.

(7) a. [$_{prec.w}$ ***aruki***] ***-mas-en***. b. [$_{prec.w}$ ***aruka***] ***-nai-desu***. verbs
walk -HON$_A$-NEG walk -NEG-HON$_A$
'(I) do not walk.' '(I) do not walk.'

(8) a. [$_{prec.w}$ ***raku***]*-de-wa* *ari-**mas-en***. b. [$_{prec.w}$ ***raku***]*-de-wa* ***nai-desu***. *na*-adj.
easy-COP-FOC be-HON$_A$-NEG easy-COP-FOC NEG-HON$_A$
'(It) is not easy.' '(It) is not easy.'

(9) a. [$_{prec.w}$ ***uresiku***] *ari-**mas-en***. b. [$_{prec.w}$ ***uresiku***] ***nai-desu***. *i*-adj.
happy be-HON$_A$-NEG happy NEG-HON$_A$
'(I) am not happy.' '(I) am not happy.'

(10) a. [$_{prec.w}$ ***hako***]*-de-wa* *ari-**mas-en***. b. [$_{prec.w}$ ***hako***]*-de-wa* ***nai-desu***. noun
box-COP-FOC be-HON$_A$-NEG box-COP-FOC NEG-HON$_A$
'(It) is not a box.' '(It) is not a box.'

(11) a. [$_{prec.w}$ ***mi-rare***] ***-mas-en***. b. [$_{prec.w}$ ***mi-rare***] ***-nai-desu***. aux.
see-can -HON$_A$-NEG see-can NEG-HON$_A$
'(I) cannot see it.' '(I) cannot see it.'

Example. Assume that the sentence has the scheme as shown in (12)a. This sentence can get realized either as (12)b or as (12)c. Here, the values for *tns* and *int* are set to 1, while *yoo* and *hon* take the value of 0. *Prec.cat* is "adj" and *prec.word* is "*uresii*." Our task is to predict the choice of these constructions based on the surrounding expressions and contextual information.

(12) a. *uresiku* [ ]*-ta* [ ]*-ka?* b. *uresiku* [*ari-mas-en desi*]*-ta-ka?*
happy -PST -Q happy be-HON$_A$-NEG HON$_A$ -PST-Q

c. *uresiku* [*nak at*]*-ta* [*desu*]*-ka?*
happy NEG be -PST HON$_A$-Q

3.3 Restrictions: this study examines a subset of the extracted examples.

A. Restriction on the *prec.cat*. First, miscellaneous *prec.cats* are discarded, such as *empty spaces* and *unknown words*, because many of these categories result from the annotation schema only specific to BCCWJ. Second, there exist marked constructions that would deserve our attention but are not considered as a canonical negation. For instance, mirative/exclamative

[1] **Nouns:** Pronouns are subsumed under *nouns*. Nominal suffixes, such as *-ya* 'shop' in *bideo-ya* 'video shop' is also counted as a noun.

[2] ***zonzur-* 'think.HON':** BCCWJ inconsistently annotates *zonzi* as a verb or a noun. This study manually re-annotated and unified the examples as a verb.

	i -adjectives	*na* -adjectives	auxiliaries	verbs	nouns	total
-mas-en (= 0)	1,026	360	6,727	90,637	4,619	103,369
nai-desu (=1)	936	90	1,299	6,630	1,081	10,036
total	1,962	450	8,026	97,267	5,700	113,405

Table 2 Raw frequencies for the preceding part-of-speech.

constructions such as *hasiru-de-wa ari-mas-en-ka* run-COP-FOC be-HON$_A$-NEG-Q and *hasiru-de-wa nai-desu-ka* run-COP-FOC NEG-HON$_A$-Q '(unexpectedly) (he) ran!' are different from the sentences in (7), in that the running event is not the target of the negation. Admittedly, the examination of such constructions would be indispensable for the complete understanding of variable selection but, due to its complexity, this paper focuses only on those parts-of-speech introduced above, leaving the exhaustive research to future studies. With this restriction, this study extracted 125,715 examples out of the original 176,899 tokens. When we build a model in Section 4, the *i*-adjective is set to the default case.

B. Restriction with respect to the frequency. At this moment, we have 3,993 different *prec.words* but 2,088 *prec.words* appear only once (*i.e.*, hapax legomena; *cf.*, Zipf's law). This study excludes *prec.words* with a low frequency for fear that the ratio between the prescriptive form and the new form may be affected too much by the random errors in the corpus. For example, if a word X is observed only once in the corpus and takes the form of *nai desu*, we treat it as an expression that is *always* used with the new variant. This conclusion might be true, but such an apparent strong preference may also be due to the small sample size. This study, thus, examines those that appear more than or equal to 30 times (≥ 30) in this corpus. Examples are, then, boiled down to 114,098 examples and we have 247 different *prec.words*.[3]

C. Restriction with respect to *the prec.word.* Although the accuracy of annotation in BCCWJ is quite reliable, there are few erroneous cases, most of which are due to the difficulty in the morphological segmentation process. Especially, the following *prec.words* in (13) and (14) result in many non-negligible errors. Besides, there are some cases where BCCWJ inconsistently annotate *-nai* as a part of a single word or as a word in its own right (= (15)). These instances are taken away as well as a dialectal expression in (16). In this study, we will examine the remaining 113,405 tokens; that is, we will discuss 243 *prec.words*. Table 2 summarizes the *prec.cats* of these 113,405 examples.

(13) **Auxililary *-ri* (archaic)** [61 tokens]: many errors stem from the mis-segmentation of *aari-mas-en*, which is a variant of *ari-mas-en* with the first vowel lengthened and *wakari-mas-en* when the *wakari* is written as "解かり".

(14) **Auxililary *-da*** [206 tokens]: when the verb *de-* 'to come out' is used, it is sometimes wrongly parsed as the copula *-da*; *e.g.*, *boonasu denaidesyoo* is structurally ambiguous between the reading of *a bonus won't come out (be given)* and *(it is) not the bonus*.

[3] **Arbitrariness:** the choice of 30 as our threshold is arbitrary. The larger the threshold value is, the more reliable data we can gain. But, at the same time, the type frequency of *prec.words* gets smaller. That is, there is a trade-off relation between the reliability and the type frequency. I consider 247 types are enough for our purpose of examining the variation in *prec.words* and adopt this threshold.

Mathematical Linguistics, Vol.31 No.1 (June 2017) pp.1-12. 〔Times New Roman 9pt, left-aligned〕

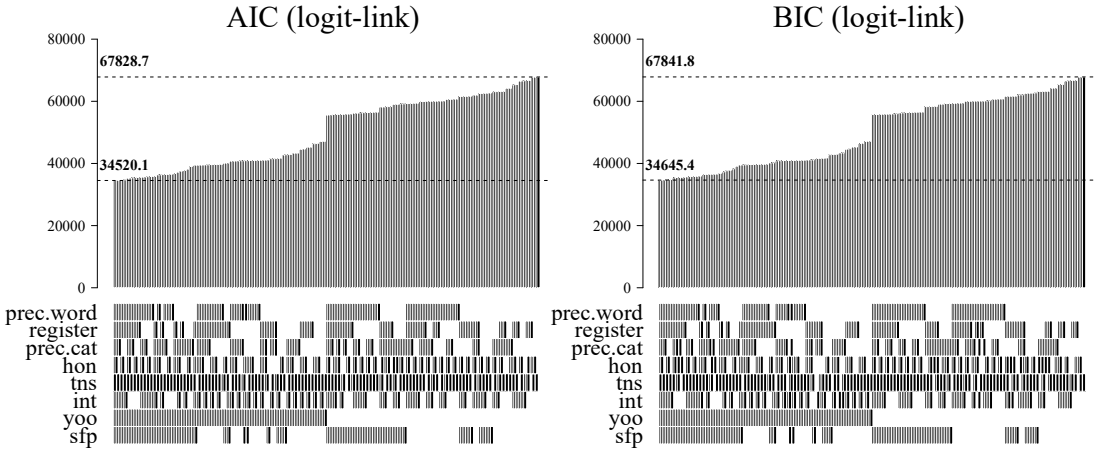


Figure 1 AIC and BIC for the logit-link models. Each plot shows the AIC/BIC values for the 256 models. They are plotted in increasing order *wrt* these information criteria.

(15) **verb *tumar-*** [79 tokens]: the sequence of *tumar-* followed by *-nai* is ambiguous. One reading is the negation of a clogging (*tumar-*) event. The other reading is to treat *tumar-anai* as an unanalyzable adjective meaning 'boring.' Since it is not easy to automatically disambiguate these readings, this study excludes all these instances.

(16) **Auxiliary *-ya*** [60 tokens]: this auxiliary is used in the western dialect. For simplicity's sake, this paper only examines those that can be found in the standard Tokyo Japanese.

4. Statistical modeling

4.1 Model comparison

We build a set of linear models that assume that every outcome variable is independently generated by identical Bernoulli distributions. Since we are considering 8 variables, there are $2^8 = 256$ possible models we can compare; for simplicity's sake, we do not consider any complicated models, for example, with interaction terms, non-linear effects and/or different link functions, leaving such models to future studies. As a criterion for the comparison, this study calculates AIC and BIC for every model; the results are illustrated in Figure 1. In these two plots, the height of each bar represents the information criterion of a model and the barcode images below show which predictors are included in the model. From these results, we conclude that, irrespective of the choice of AIC/BIC, the best model (= the leftmost model) is the one with all the variables, which is formally stated in (17).

(17) Best model (among the models considered in this study)

$$y_{ijk} \sim Bernoulli\left(\pi_{ijk}\right),\ i \in \{1, \ldots, n_{jk}\},\ j \in \{1, \ldots, 243\},\ k \in \{1, \ldots, 12\}$$

$$\begin{aligned}\mathrm{logit}(\pi_{ijk}) = \ & \beta_{0jk} + \beta_{\mathrm{sfp}} x_{\mathrm{sfp}\,ijk} + \beta_{\mathrm{conj}} x_{\mathrm{conj}\,ijk} + \beta_{\mathrm{yoo}} x_{\mathrm{yoo}\,ijk} + \beta_{\mathrm{int}} x_{\mathrm{int}\,ijk} \\ & + \beta_{\mathrm{tns}} x_{\mathrm{tns}\,ijk} + \beta_{\mathrm{hon}} x_{\mathrm{hon}\,ijk} + \beta_{\mathrm{na.adj}} x_{\mathrm{na.adj}\,ijk} \\ & + \beta_{\mathrm{noun}} x_{\mathrm{noun}\,ijk} + \beta_{\mathrm{vrb}} x_{\mathrm{vrb}\,ijk} + \beta_{\mathrm{aux}} x_{\mathrm{aux}\,ijk}\end{aligned}$$

$$\beta_{0jk} = \gamma_{00} + u_{0j} + w_{0k},\ u_{0j} \sim N(0, \sigma_0^2),\ w_{0k} \sim N(0, \tau_0^2)$$

In (17), y_{ijk} is the outcome variable (either 0, the *mas-en* form, or 1, the *nai desu* form). π_{ijk} is the probability of the *i*-th observation of *prec.word j* from register *k*. We use *logit* for the link

		Estimate	Std. Err.	z value	Pr(>|z|)
(intercept)	$\hat{\gamma}_{00}$	-1.48	0.339	-4.4	1.31E-05
sfp (sfp)	$\hat{\beta}_{\text{sfp}}$	2.73	0.039	70.1	0
sfp (conj)	$\hat{\beta}_{\text{conj}}$	1.12	0.043	26.0	2.74E-149
yoo	$\hat{\beta}_{\text{yoo}}$	7.41	0.097	76.1	0
int	$\hat{\beta}_{\text{int}}$	-1.55	0.060	-25.7	1.17E-145
tns	$\hat{\beta}_{\text{tns}}$	0.28	0.056	4.9	1.00E-06
hon	$\hat{\beta}_{\text{hon}}$	-3.93	0.417	-9.4	4.56E-21
prec.cat					
na-adjective	$\hat{\beta}_{\text{na.adj}}$	-1.97	0.381	-5.2	2.15E-07
noun	$\hat{\beta}_{\text{noun}}$	-1.49	0.257	-5.8	7.02E-09
verb	$\hat{\beta}_{\text{vrb}}$	-3.11	0.203	-15.3	7.19E-53
aux	$\hat{\beta}_{\text{aux}}$	-2.76	0.359	-7.7	1.61E-14

Table 3 Results for the fixed effects.

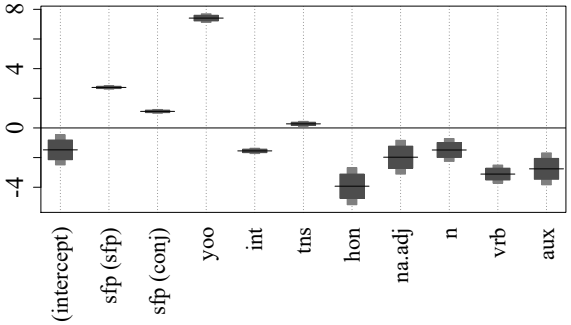


Figure 2 Confidence intervals. For each variable, ±1.96*Std Error (thicker region) and ±3*Std Error are illustrated (thinner region).

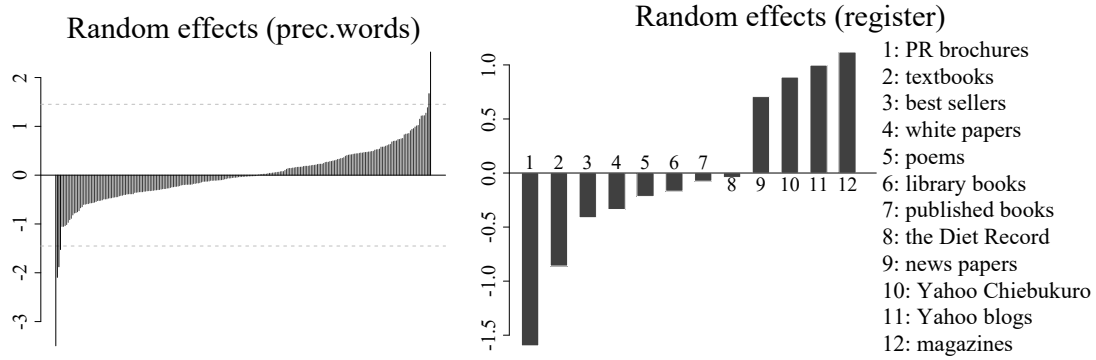


Figure 3 Results of the random effects. (Left) the estimated $\boldsymbol{u_{oj}}$'s (j = {1, …, 243}) in the increasing order $\widehat{\boldsymbol{\sigma_0^2}} = \mathbf{0.54}$; (Right) the estimated $\boldsymbol{w_{0k}}$'s (k = {1, …, 12}); $\widehat{\boldsymbol{\tau_0^2}} = \mathbf{0.76}$.

function. The intercept β_{0jk} constitutes γ_{00} and two random effects u_{0j} (the effect of the *j*-th *prec.word*) and w_{0k} (the effect of the *k*-th *register*) and it is assumed that they both come from the Normal distribution with mean 0 and variance σ_0^2 and τ_0^2. The other betas are the coefficients for the predictors.

4.2 Estimation

The parameters of this generalized linear mixed effect model are estimated by maximum likelihood estimation (Adaptive Gauss-Hermite Quadrature), based on the *lme4* package for R; this method is chosen because of the speed and the ease of convergence (in comparing different models, these criteria are practically important). The estimated results for the fixed effects are summarized in Table 3, which are visually represented in Figure 2. The results for the random intercepts are shown in Figure 3. As for the *prec.words*, for the limited space, this paper refrains from showing the results of all 243 estimated values. Instead, the *prec.words* with the largest values (Table 3 and Table 5) and the smallest values (Table 4) are reported. The "No." refers to the ranking of the estimated *prec.words* in increasing order. For example, the leftmost bar in the left panel in Figure 3 refers to *nega-e-* (-3.50), which is No. 1, and the rightmost bar is No. 243 *-tai* (2.52).

No.	prec.word	u_{0j}	No.	prec.word	u_{0j}	No.	prec.word	u_{0j}
1	*negae-* 'can wish'	-3.50	5	*sadaka-* 'clear'	-1.05	9	*ser-* 'CAUS'	-0.97
2	*sum-* 'finish'	-2.09	6	*mezurasii* 'rare'	-1.05	10	*kire-* 'can cut'	-0.91
3	*sukunai-* 'few'	-1.88	7	*hito* 'person'	-1.04	11	*reigai* 'exception'	-0.89
4	*kudasar-* 'give me (hon)'	-1.53	8	*kuwasii* 'detailed'	-1.01	12	*byooki* 'disease'	-0.82

Table 3 Random effects (*prec.words* with large negative values).

No.	prec.word	u_{0j}	No.	prec.word	u_{0j}	No.	prec.word	u_{0j}
126	*mous-* 'say (hon)'	-0.02	130	*nobi-* 'grow'	-0.01	134	*miatar-* 'can be found'	0.01
127	*mousiage-* 'say (hon)'	-0.02	131	*suw-* 'smoke'	0.00	135	*ire-* 'insert'	0.02
128	*maniaw-* 'be enough'	-0.02	132	*nuke-* 'come off'	0.01	136	*her-* 'decease'	0.02
129	*itadak-* 'receive (hon)'	-0.01	133	*umare-* 'be born'	0.01	137	*ike-* 'can go'	0.02

Table 4 Random effects (*prec.words* with small values).

No.	prec.word	u_{0j}	No.	prec.word	u_{0j}	No.	prec.word	u_{0j}
232	NP-*san* 'Mr./Ms. NP'	0.95	236	*ir-* 'be necessary'	1.15	240	*hosii-* 'want'	1.27
233	*yom-* 'read'	0.99	237	*yoi* 'good'	1.21	241	*tar-* 'be enough'	1.38
234	*aki-* 'get bored'	1.01	238	*kanzi* 'feeling'	1.22	242	*irassyar-* 'go (hon)'	1.67
235	*uresii* 'happy'	1.02	239	*-kire-* 'do perfectly'	1.22	243	*-tai* 'want'	2.52

Table 5 Random effects (*prec.words* with large positive values).

4.3 Interpretation

First, the intercept is negative ($\hat{\gamma}_{00} = -1.48$), suggesting that, in the baseline case --- when the addressee-honorific marker is used with an *i*-adjective in the present tense and sentence-final particles but honorific expressions, conjunctive particles, *-yoo* and *-ka* are all absent (*e.g.*, (1) and (2)) --- the prescriptive form is preferred.

Second, corroborating the results in the previous studies, the sentence-final particle ($\hat{\beta}_{\mathrm{sfp}}$) and the conjunctive particle ($\hat{\beta}_{\mathrm{conj}}$) show a positive value, suggesting that they favor the new variant. The variable with the largest positive effect size is the presence of the suffix *-yoo* ($\hat{\beta}_{\mathrm{yoo}} = 7.41$), which is not reported in the previous studies. The past tense (*tns*) has a very small effect size ($\hat{\beta}_{\mathrm{tns}} = 0.28$). With a large sample size, the confidence interval gets smaller, as seen in Figure 2. So, despite its apparent significance (under the common threshold of $\alpha = 0.05/0.001$), they do not seem as important as the other variables in practice. The *ka* ($\hat{\beta}_{\mathrm{int}} = -1.55$) and the honorific form ($\hat{\beta}_{\mathrm{hon}} = -3.93$) have a negative effect.

Third, the fixed effects of the four parts-of-speech are all negative, suggesting that, compared to *i*-adjectives, they are more inclined to take the prescriptive form. In other words, the *i*-adjective is the most liberal to the new variant (*cf.*, in terms of the effect size, *verbs* and *auxiliaries* show larger values, favoring the prescriptive form, as has been claimed in the previous studies). A possible account is the influence of the non-negative sentence (Kawaguchi 2014: 85). The corresponding affirmative sentence for (18)a is not (18)b but (18)c. That is, *i*-adjectives are split into the *mas*-based addressee-honorification system (when it is negative) and the *des*-based system (when it is affirmative). By adopting the *des*-form for the negative sentence, *i*-adjectives can level the morpho-syntactic paradigm.[4]

[4] **The sentence on (18)b:** This hypothesis would become more persuasive if we could explain why *i*-adjectives had been reluctant to take *-mas* in the affirmative sentence. But, in order to provide an analysis, we need a theory on the distribution of the *do/be*-support element or the

(18) a. Negative form
Uresiku *ari-mas-en.*
happy be-HON$_A$-NEG
'I am not happy.'

b. Affirmative form (I)
*?*Uresiku* *ari-mas-u.*
shappy be-HON$_A$-PRS
'I am happy (intended).'

c. Affirmative form (II)
Uresii *desu.*
happy HON$_A$
'I am happy.'

Finally, as for the register, as expected, formal media (*e.g.*, PR brochures and textbooks) show high preference for the *-mas-en* form, while casual registers (*e.g.*, the Internet blogs/sleds and magazines) are more generous to the new variants. The details of the other random effect, *prec.words*, will be discussed in Section 5.2 and 5.3.

5. Discussion

5.1 Epistemic *-yoo*

By examining the effect sizes of the predictors, we can conclude that the impact of having the suffix *-yoo* is the largest, though the importance of this predictor has not been discussed by any of the previous studies. This conclusion motivates us to ask a deeper question. Why is this morpheme so fond of the new variant? Though this paper does not have an ultimate answer, it seems important to look at the fact that this epistemic morpheme in contemporary Japanese has a requirement that it be linearly adjacent to a copula element. Observe that, even in the plain form, one cannot omit the copula element *dar* (< *de ar-* 'ASSERT be') in front of *-(y)oo*, as shown in (19).

(19) Epistemic *-yoo* in the plain form

a. *Karera-wa baa-ni iku *(dar)-oo.* *affirmative*
they-TOP bar-to go COP-EPI
'It is likely that they will go the bar.'

b. *Karera-wa baa-ni ika-nai *(dar)-oo.* *negative*
they-TOP bar-to go-NEG COP-EPI
'It is likely that they will not go the bar.'

Of all the two addressee-honorific markers, only *des* has the copula function in addition to encoding the addressee-honorific meaning. Since *-yoo* has this morpho-syntactic requirement, the new form is selected despite the general tendency.[5]

---

copula construction in Japanese, which is beyond the scope of this paper.

[5] **Proposal-making *-yoo* and speech acts:** The morpheme *-yoo* has another distinct function to trigger the speech act of making a proposal. Unlike the epistemic *-yoo* in (5)b, proposal-making *-yoo* is incompatible with *desu*. As Ozaki (2004: 35) points out, the *(nai) desu* form seems strongly related with the speech act of making a statement. Other illocutionary forces are not as frequently used with the new variant. One can understand that this *-yoo* dislikes the new form because the sentence in (i)a is used to make a proposal to the addressee, *i.e.*, another speech act different from making an assertion. The fact that it cannot be negated as in (ii)(1)b supports the conclusion that it does not make a statement.

(i) Volitional *-yoo* in the polite form

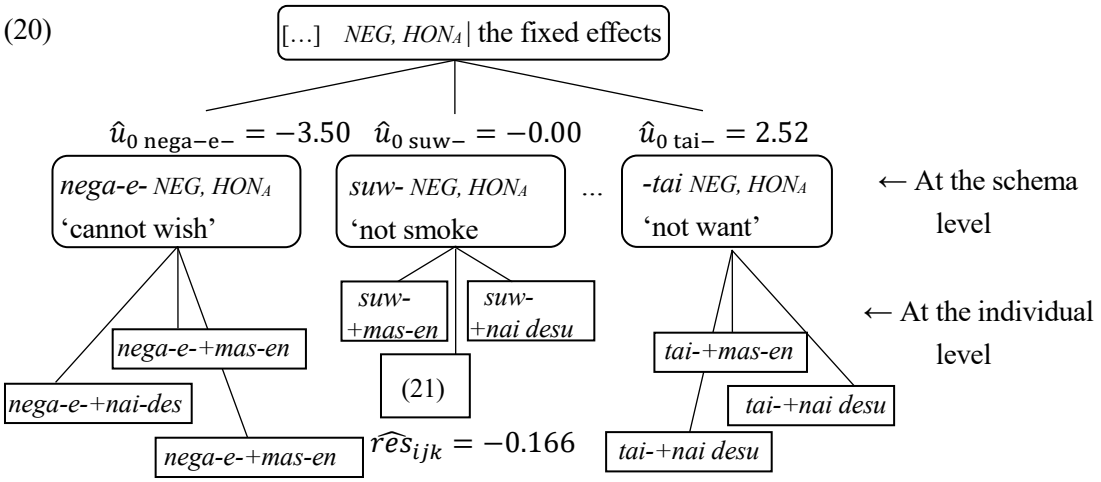
a. *Issyoni baa-ni {iki-mas-yoo/ * iku des-yoo}.* *affirmative*
together bar-to go-HON$_A$-VOL/ go HON$_A$-VOL

Mathematical Linguistics, Vol.31 No.1 (June 2017) pp.1-12. 〔Times New Roman 9pt, left-aligned〕

5.2 Conditional prototypes

In construction grammars, it is assumed that our knowledge of language is represented as a schema structure made up of taxonomies of constructional schemas with different levels of abstraction (Langacker 2008). For example, the schema structure in (20) depicts how NEG HON$_A$ construction forms a network. The top node in the tree illustrates the abstract superschema that represents the negation and the addressee-honorific meaning, given the fixed effects.

(20)



When the bracketed slot is replaced by a specific *prec.word*, the constructional schema gets elaborated. For example, "*suw-+ NEG, HON$_A$*" is an elaboration of the top node. But, this node is still an abstract schema for its instances. For instance, the sentence in (21) is regarded as an elaboration of this schema, in which *NEG, HON$_A$* gets realized as *-mas-en* (in (20), the round node is used for the conceptual, latent schema and the boxed node for observed instances).

(21) *demo, ie-no naka-de-wa sui-**mas-en**.* ($\hat{\pi}_{ijk} = 0.03$; $\widehat{res}_{ijk} = -0.166$)
but house-GEN inside-at-FOC smoke-HON$_A$-NEG [register: *Yahoo blog*]
'But I do not smoke inside (the building) of my house.'

We can interpret the statistical model in (17) as quantifying this schema structure by assigning a weight to each node. First, from Table 4, we know *suw-* has the smallest estimated value for the random intercept in magnitude ($\hat{u}_{0\ \mathrm{suw}-}$is almost 0). This means that the variable selection of *suw-* is highly predictable when the fixed effects are controlled. On the other hand, "*nega-e-*+NEG HON$_A$" is assigned the largest negative value ($\hat{u}_{0\ \mathrm{nega}-\mathrm{e}-}$ = -3.5), suggesting

---

'Let us go to the bar together.'

b. *Issyoni baa-ni {*iki-mas-en des-yoo/* ika-nai des-yoo}. negative*
together bar-to go-HON$_A$-NEG HON$_A$-EPI go-NEG HON$_A$-EPI
'Let us not go to the bar together (intended).'

(ii) Volitional *-yoo* in the polite form

a. *Issyoni baa-ni ik-oo.*
together bar-to go-PM
'Let us go to the bar together.'

b. **Issyoni baa-ni ik-anak-oo*
together bar-to go-NEG-PM
'Let us not go to the bar together (intended).'

that this verb shows the strongest preference for the prescriptive form. When we take into account all the fixed effects under the formula in (17), it is predicted that this verb could take more *nai desu* forms. If we use the term PROTOTYPE to refer to such constructions that have an idiosyncratic propensity that favors one variant over the other to a substantial degree, "*nega-e-*+*mas-en*" is a prototype for the prescriptive form among the instances of "[…]-+NEG, HON_A" schema.[6] Similarly, Table 5 shows that "*tai-*+NEG, HON_A" is one of the most prototypical schemas for the new form. Under the model, we predict that *-tai* could be used in the prescriptive form much more frequently given the fixed effects but, in reality, it is much more strongly associated with the new form than we have predicted.

Second, with the residual, we can measure the discrepancy between the expected outcome and the real observation. For example, "*suw-*+ NEG, HON_A" gets realized as *sui-mas-en* in (21). The estimated $\hat{\pi}_{ijk}$ for this sentence is 0.03, which means, given the fixed effects, we expect this scheme to be realized as the prescriptive form. In this case, this prediction is borne out and *-mas-en* form is indeed selected. Pearson residual quantifies our surprise ($\widehat{res}_{ijk}$= -0.166), measuring the difference between our expectation and the real outcome.

Of course, there are a few cases where the predicted form and the real form do not match. For example, while the scheme "*nega-e-*+NEG HON_A" has many examples with *-mas-en*, *e.g.*, (22), we still have some instances where it gets realized as *nai desu*, *e.g.*, (23).

(22) *Dekitara, o-hikitori nega-e-**mas-en**?* ($\hat{\pi}_{ijk} = 2.6\text{e}-04$; $\widehat{res}_{ijk} = -0.02$)
if possible HON-leave wish-can-HON_A-NEG [register: *library book*]
'If possible, can I wish that you would leave (this place)?'

(23) […] *o-yurusi nega-e-**nai** **des**-yoo-ka* ($\hat{\pi}_{ijk} = 0.58$; $\widehat{res}_{ijk} = 0.84$)
HON-forgive wish-can-NEG HON_A-EPI-Q [register: *published book*]
'[…], can I wish that you would forgive (me)?'

Similarly, the examples with the least/smallest Pearson residual are given in (24) and (25).

(24) *monbukagakusyoo-wa tebanasi-**taku-nai** **des**-yoo-ne.* ($\hat{\pi}_{ijk} = 1.0$; $\widehat{res}_{ijk} = 0.008$)
MEXT-TOP give up-want-NEG COP.HON_A-EPI-SFP [register: *magazine*]
'It is likely that the Ministry of Education, Culture, Sports Science and Technology (MEXT) does not want to give (it) up.'

(25) *zyosei-tosite wasure-**taku** ari-**mas-en**-ne.* ($\hat{\pi}_{ijk} = 0.89$; $\widehat{res}_{ijk} = -2.911$)
woman-as forget-want be-HON_A-NEG-SFP [register: *magazine*]
'(we) do not want to forget (it) as a woman, right?'

Even though they are all real examples from the corpus, they are different in prototypicality. The sentences in (22) and (24) have a smaller residual value. This means that

[6] **The verb *nega-e-*:** Detailed examinations of *nega-e-* is beyond the scope of this short paper but, presumably, its strong preference for the prescriptive form comes from the speech act or the performativity in the sense of Austin (1982). This verb is typically used as an explicit performative predicate in the interrogative clause and the sentence is used to ask for permission, rather than seeking for new information. It can be the case that the speaker avoids selecting the non-prescriptive form, because the use of non-prescriptive form may give a bad impression to the addressee, which may prevent the addressee from offering a help that the speaker wants.

No.	prec.word	u_{0j}	No.	prec.word	u_{0j}	No.	prec.word	u_{0j}
1	*negae* 'can wish'	-3.50	120	*dase-* 'can give'	-0.03	180	*ie-* 'can say'	0.29
7	*kire-* 'can cut'	-0.91	121	*tukae-* 'can use'	-0.03	181	*kae-* 'can buy'	0.30
17	*sire-* 'can know'	-0.67	125	*mire-* 'can see'	-0.02	182	*ute-* 'can hit'	0.31
19	*-re-* 'can'	-0.60	137	*ike-* 'can go'	0.02	187	*tukame-* 'can grasp'	0.36
32	*susume-* 'can go'	-0.49	139	*iikire-* 'can assert'	0.04	191	*tukure-* 'can make'	0.42
40	*nore-* 'can ride'	-0.44	148	*nome-* 'can drink'	0.07	204	*kike-* 'can hear'	0.48
68	*nozome-* 'can wish'	-0.30	150	*omoe-* 'can think'	0.12	208	*ae-* 'can meet'	0.52
70	*nemure* 'can sleep'	-0.28	155	*hanase-* 'can talk'	0.15	210	*itadake-* 'can receive (hon)'	0.53
75	*hanase-* 'can separate'	-0.25	157	*erabe-* 'can select'	0.16			
85	*tore-* 'can get'	-0.19	159	*kikoe-* 'can hear'	0.16	218	*kake-* 'can write'	0.68
88	*-rare-* 'can'	-0.19	167	*tabere-* 'can eat'	0.20	220	*mote-* 'can hold'	0.70
92	*minogase-* 'can miss'	-0.16	169	*yuruse-* 'can allow'	0.21	221	*ure-* 'can sell'	0.70
105	*hazuse-* 'can remove'	-0.10	173	*mie-* 'can see'	0.23	229	*yome-* 'can read'	0.85
112	*deki-* 'can do'	-0.06	176	*iname-* 'can deny'	0.27	239	*-kire-* 'can finish'	1.22
118	*kikitore-* 'can hear'	-0.04	177	*tore-* 'can take'	0.27			

Table 6 Potential verbs.

the schemes are elaborated as we (= the model) expect. But the sentence in (23) and (25) have a larger residual value, showing that, given the schema, these elaborations are less expected, *i.e.*, less prototypical than the other sentences.

Quantified schema structures, thus, enable us to select the very sentence we need. First, if we want to cite an example of a prototypical instance from a prototypical *prec.word*, it is reasonable to use the sentence in (22), not (21) or (23). Second, if we are interested in how Japanese is changing, then the sentence in (23) may be quite informative, because this illustrates a case where even the most conservative construction takes the new variant.

The notion of *prototype* is a relativized concept. In terms of the *precc.words* (and the fixed effects in the model), *nega-e-*+NEG HON_{A} is a CONDITIONAL PROTOTYPE for *-mas-en* at the schema level. Sentences with a small residual are a conditional prototype at the individual level. Models with the random intercepts, thus, make us quantitatively analyze the hierarchical schema structure in the least subjective way.

5.3 Eventuality

Some previous studies have claimed that *wakar-* 'understand' and *deki-* 'can' are the prototypes of the new variant and have hypothesized that stative predicates (*e.g.*, potential verbs) are tied to the new variant (Noda 2004: 234-235; Kawaguchi 2014: 152), while Ochiai (2012) argues against this view saying that eventuality of the verb does not play a role. By looking at the estimated values for the random intercepts, we can see which view is the more reasonable conclusion to adopt.

First, the random intercepts of these words do not show a large positive value; $u_{0\text{wakar}-} = 0.05$ and $u_{0\text{deki}-} = -0.06$. Compared to other *prec.words* (*e.g.*, those in Table 5), we cannot conclude that they are extremely leaned toward the *nai desu* form, when the fixed effects are all controlled. Second, the estimated random intercepts for potential verbs are extracted and shown in Table 6. If they prefer the new variant, then they should systematically take a (large) positive value. However, such a tendency is not easily inferred from this table.

The statistical model, thus, supports Ochiai's (2012) view that the eventuality of *prep.word* is not crucial in variable selection.

6. Conclusion

Based on the best model in terms of information criteria, this paper has made two novel claims. First, the epistemic suffix *-yoo* is the strongest in effect size, favoring the new form. Second, although some previous studies hypothesized that stative predicates, *e.g.*, *wakar-* 'understand' and *deki-* 'can', are prototypes for the new construction (Noda 2004; Kawaguchi 2014), they do not show a strong preference for the new form when important fixed effects are controlled. Instead, this paper identifies prototypes by looking at the estimated value of random effects, pointing out that *-tai* is a candidate for a prototype of the *nai desu* form and *nega-e-* is a prototype of the prescriptive form.

Of course, any statistical study cannot be complete by itself. We can improve the model by incorporating variables not taken into account in the previous study or by making a model with a more elaborated structure; *cf.*, Box's (1979) famous remark that "[a]ll models are wrong but some are useful." The fact that all the variables remain wrt. AIC/BIC suggests that none of the variables is dominant, so it is likely that we may find another important factor in future studies. For example, some have argued that the speech act of the utterance plays a pivotal role (Ozaki 2004; Kawaguchi 2010, 2014). Though, for simplicity's sake, this paper only examines the clause type distinction between *ka* and non-*ka* sentences, one may build a better model by considering fined-grained distinction in speech acts; since BCCWJ does not provide the information on the speech act, we would manually annotate the data. In addition, the model may also get ameliorated by having interaction terms, non-linear effects and/or changing the link function. It is important to note, however, that, whatever better model we may build in future studies, the core proposal of this paper remains effective; once we have identified the best model, we can discuss which *prec.word* +NEG HON_{A} is the biggest in random intercept and, by examining the residuals, we can detect which instance is the most prototypical/marginal, which enhances our understanding of the variable selection.

Reference

Austin, J. (1962). *How to do things with words.* New York: Oxford University Press.

Banno, E. (2012). Kōpasu o tukatta zyutugo hitei kei "masen" to "naidesu" no siyō zittai tyōsa. [A corpus-based study on the negative predicate forms *"masen"* and *"nai desu"*]. *Journal of international student education* 17. 133-140.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In Launer, R. L.,Wilkinson, G. N. (eds.), *Robustness in Statistics*. New York: Academic Press, 201-236.

Bybee, J. (2010). *Language, usage and cognition*. New York: Cambridge University Press.

Fukushima, E. & Uehara, S. (2003). Nihongo teineitai hiteizi nikeisiki ni kansuru tūziteki kenkyū: tekisuto bunseki ni yoru kēsu sutadī [A study on two negative polite forms in Japanese]. *Journal of the Graduate School of International Cultural Studies* 11. 79-89.

Kawaguchi, R. (2010). "Masen" kei kara "naidesu" kei e no sihuto ni kakawaru yōin ni tuite: dōsi hiteikei no gengo henka toiu kanten kara [Factors related to the shift from the *-masen* to the *-nai desu* form: in terms of language change of the negative polite verb form]. *Nihongo Kyoiku* 144. 121-132.

Kawaguchi R. (2014). *Teineitai hiteikei no bariēsyon ni kansuru kenkyū [A study on the variation in polite negative forms]*. Tokyo: Kuroshio Publishers.

Langacker, R. (2008). *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

Masuoka, T. (2007). *Nihongo modaritii tankyuu [Investigation on Japanese modalities]*. Tokyo: Kuroshio Publishers.

Miyagawa, S. (2012). Agreements that occur mainly in main clauses. In Aelbrecht, L., Haegeman, L., Nye, R., (eds.), *Main clause phenomena: new horizons*. 79-112. Amsterdam: John Benjamins Publishing Company.

Miyagawa, S. (2017). *Agreement beyond phi*. Cambridge, MA: MIT Press.

Noda, H. (2004). Hitei teineikei "masen" to "naidesu" no siyō ni kakawaru yōin: yōrei tyōsa to zyakunensō ankēto tyōsa ni motozuite [Factors relating to the use of -masen and -nai desu as polite negative form --- based on a usage investigation and on a questionnaire investigation of young generation speakers]. *Mathematical Linguistics* 24 (5). 228-244.

Ochiai, T. (2012). Kakikotoba ni arawareru "masen" to "naidesu" ["masen" and "naidesu" in the written language]. *Kokubun Mejiro* 51.14-22.

Ozaki, N. (2004). Hitei no teineikei "naidesu" to "masen" ni tuite [On the polite negative forms "naidesu" and "naidesu"]. *Okayama Daigaku Gengogaku Ronso* 11. 29-42.

Rizzi, L. (1997). The fine structure of the left periphery. In Haegeman, L. (ed.), *Elements of Grammar*. 281-339. Dordrecht: Kluwer.

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of experimental pscychology: general* 104. 192-233.

Tanaka, A. (2008). "Masu" kara "desu" e: teineitai no henyō [From "masu" to "desu": a change in polite forms]. *Kidaigo Kenkyû* 14. 326-341.

Tanomura, T. (1994). Teineitai no zyutugo hiteikei no sentaku ni kansuru keiryō teki tyōsa: "-masen" to "-nai desu" [A corpus-based study on the polite negative predicate forms in Japanese]. *Journal of Osaka University of Foreign Studies* 11.51-66.

Taylor, J. R. (2012). *The mental corpus: how language is represented in the mind*. Oxford: Oxford University Press.

Yamada, A. (2018). *Historical developments/variations of Japanese addressee-honorific markers and economy principles.* Speed Presentation at the workshop of Cross-linguistic Variation in the Left Periphery at the Syntax-Discourse Interface. SNU International Conference on Linguistics, Seoul National University 2018.