Multinomial Mixed-Effects Models and Linguistic Variation

Competitions among Japanese Subject-Honorific Constructions

Akitaka Yamada[1]

[1]Osaka University

1 Introduction

When examining linguistic variation, it is important to detect the factors affecting selection tendencies. For example, selection tendencies may change per verb (a linguistic factor) or per register (a social factor), while some are attributed to sentence-level factors, and other idiosyncrasies are ascribed to no clear factors at all. Thus, when only personal introspection is used, proofs concerning how far each of these many factors contributes to a selection become less convincing and objective.

To separate lexical idiosyncrasies from general tendencies, this research promotes a corpus study employing an underused statistical model: the Multinomial Mixed-Effects Model (MMEM) (Leshvina, 2016). Applying this model to variations among Japanese subject-honorifics, this study shows how scrutinizing random effects reveals lexical idiosyncrasies and contributes to the discussion in theoretical linguistics.

2 Background

Japanese has several subject-honorifics, as illustrated in (1), and how they differ in use is of great concern.

(1) a. *sensei-ga* ***go***- *tootyaku-****ni nat****-ta.*
teacher-NOM HON-arriving-ni become-PST
b. *sensei-ga tootyaku-****nasat****-ta.*
teacher-NOM arrive-hons-PST
c. *sensei-ga* ***go***- *tootyaku-****nasat****-ta.*

teacher-NOM HON-arrive-hons-PST
‘The teacher arrived.’

Previous studies have revealed that (HYP 1) a verb of one mora cannot be used with the honorific prefix, and (HYP 2) a verb of Chinese origin favors the *nasar*-construction when compared to the *go…ni nar*-construction. Our goal is to examine to what extent these well-accepted conclusions are supported by the real corpus data.

3 Data, Model, and Estimation

The data set is a sample from the BCCWJ. Among the 20,864 hits, we concentrate on verbs used at least more than 24 times, because it is not easy to detect the distributional profile of a verb with a lower frequency. Based on these 11,223 cases, the study builds the following MMEM model by developing the model proposed in Yamada (2019a):

(2)

$$y_{i(jk)} \sim \mathrm{Categorial}\left(\vec{\pi}_{i(jk)}\right)$$

$$\vec{\pi}_{i(jk)} = \mathrm{inv\_logit}\left(\vec{\eta}_{i(jk)}\right)$$

$$\vec{\eta}_{i(jk)} = \begin{pmatrix} 0 \\ \\ \gamma_{00}^{(2)} + \gamma_{01}^{(2)} x_{ORIGIN,j} + \gamma_{02}^{(2)} x_{MORA,j} + \gamma_{03}^{(2)} x_{SUPPL,j} \\ +\beta_{01}^{(2)} x_{AUX,i(j)} + \beta_{02}^{(2)} x_{IMP_s,i(j)} + \beta_{03}^{(2)} x_{IMP_w,i(j)} + u_j^{(2)} + v_k^{(2)} \\ \\ \gamma_{00}^{(3)} + \gamma_{01}^{(3)} x_{ORIGIN,j} + \gamma_{02}^{(3)} x_{MORA,j} + \gamma_{03}^{(3)} x_{SUPPL,j} \\ +\beta_{01}^{(3)} x_{AUX,i(j)} + \beta_{02}^{(3)} x_{IMP_s,i(j)} + \beta_{03}^{(3)} x_{IMP_w,i(j)} + u_j^{(3)} + v_k^{(3)} \end{pmatrix}$$

The outcome variable is the choice of subject-honorific constructions in (1); the *go…ni nar* construction is set as the baseline category; $\vec{\gamma}^{(2)}$ and $\vec{\beta}^{(2)}$ represent the construction in (1b), and $\vec{\gamma}^{(3)}$ and $\vec{\beta}^{(3)}$ represent the construction in (1c). The considered predictors are as follows:

(3) Group-level fixed effects:

a. SUPPL: 1 when the main verb has a suppletive form; 0 otherwise

b. MORA: 1 when the main verb is formed by one mora; 0 otherwise

c. ORIGIN: 1 when the main verb is a word of Chinese origin; 0 otherwise

(4) Population-level fixed effects:

a. IMP_s: 1 when the sentence is a strong imperative; 0 otherwise

b. IMP_w: 1 when the sentence is a weak imperative; 0 otherwise

c. AUX: 1 when the subject-honorific marker appears on an auxiliary; 0 otherwise

(5) Group-level random effects:

a. LEXEME: a grouping variable indicating the predicate to which the subject-honorific marking is attached

b. REG: a grouping variable indicating the register from which the sentence is taken.

To avoid the problem of complete separation, we assume weakly informative priors $N(0, 3)$ for fixed effect parameters, and Half-t distribution for the variance of random effects Half-$t_4(0, .5)$.

The Hamiltonian Markov Chain Monte Carlo algorithm is adopted to estimate the parameters of the Bayesian MMEM. Stan is employed on R, allowing allows us the No U-Turn Sampler. After confirming the convergence in estimation (with the R-hats of the estimated parameters being < 1.01), the posterior distribution is interpreted for each parameter.

4 Results

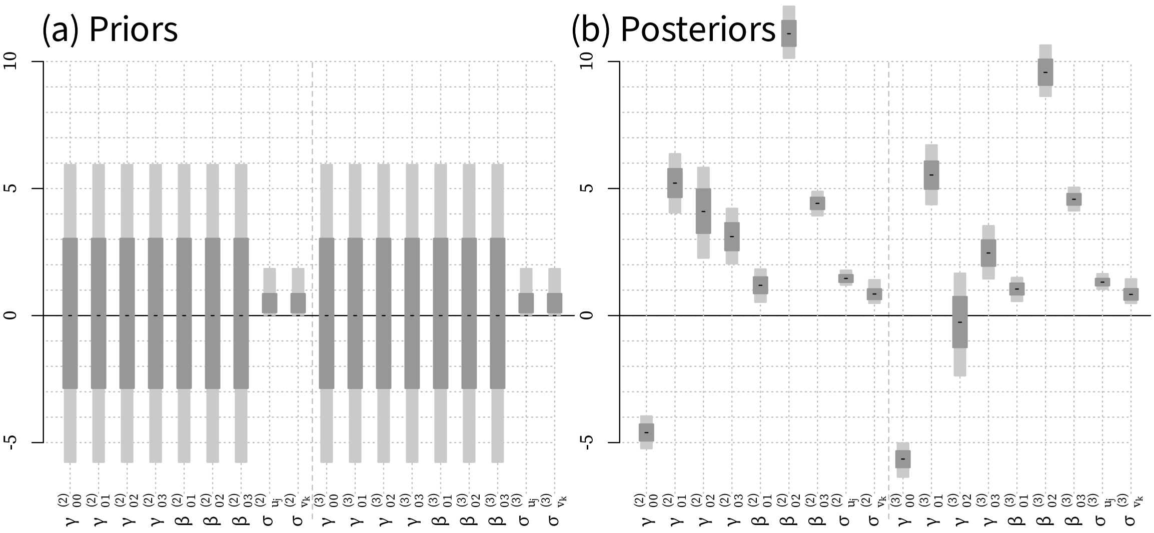
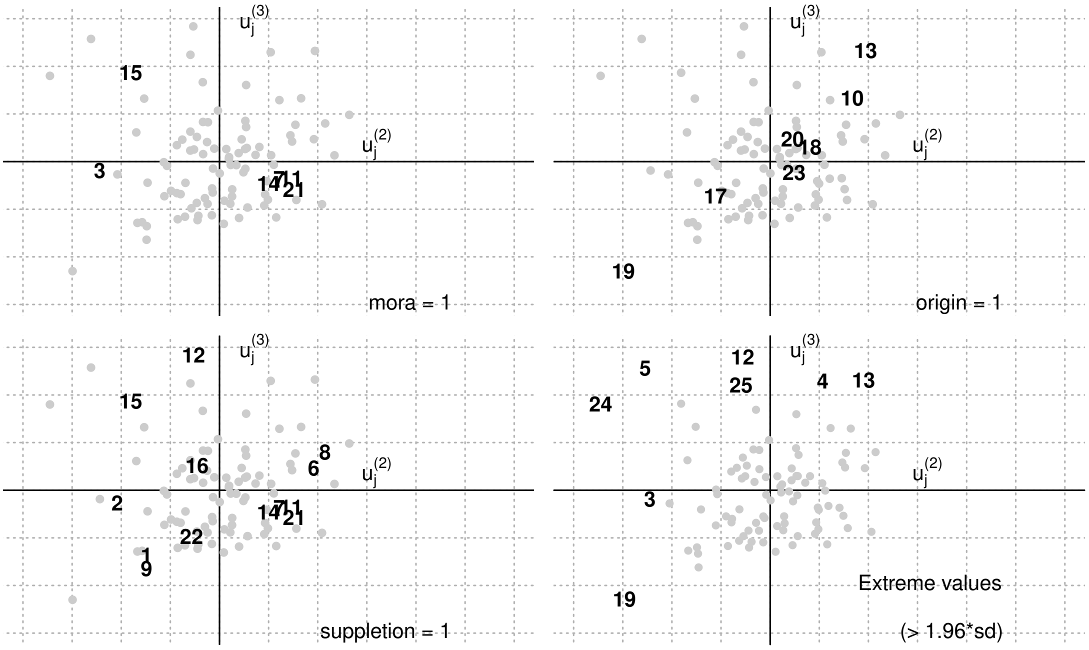


Fig. 1.

Figure 1 illustrates the priors (left panel) and the corresponding posterior distributions (right panel); the light and dark gray regions represent the 95% and the 66% credible intervals, respectively. In general, the posterior means of the fixed parameters do not substantially differ between $\vec{\gamma}^{(2)}$ and $\vec{\gamma}^{(3)}$, or between $\vec{\beta}^{(2)}$ and $\vec{\beta}^{(3)}$, except for $\gamma_{02}^{(2)}$ and $\gamma_{02}^{(3)}$, in agreement with the view of (HYP 1). The $\gamma_{01}^{(2)}$ and $\gamma_{01}^{(3)}$ are also in line with (HYP 2). The IMP$_s$ shows the largest effect size; (1a) rarely takes the strong imperative form, while (1b/c) have no such restrictions (Yamada, 2019b). The positive value for AUX means that the construction in (1a) is not as easily used as an auxiliary as those in (1b/c) (Yamada, 2019a).



1: *age* 'raise'
2: *ar* 'be'
3: *de* 'come out'
4: *hanas* 'talk'
5: *ide* 'come out'
6: *ik* 'go'
7: *ir* 'be'
8: *iw* 'say'
9: *kaw* 'buy'
10: *kekkon* 'marry'
11: *kur* 'come'
12: *kure* 'give'
13: *kurou* 'have difficulty'
14: *mi* 'see'
15: *ne* 'sleep'
16: *nom* 'drink'
17: *riyou* 'utilize'
18: *sinpai* 'be worried'
19: *siteki* 'point out'
20: *soudan* 'consult'
21: *sur* 'do'
22: *tabe* 'eat'
23: *touben* 'state'
24: *yasum* 'take a rest'
25: *yos* 'avoid'

Fig. 2.

While the results of the fixed effects essentially coincide with the findings of previous literature, MMEMs also allow us to examine idiosyncrasies not attributed to the aforementioned structural/general tendencies. Observe the scatterplots in Figure 2, which show how each lexeme has its own lexical idiosyncrasy. The following findings are worth attention.

- Despite the general tendency, some Shino-verbs do not favor the construction in (1a) (*kuroo* and *kekkon*).
- Despite the general tendency, the verb *de-* can appear in the construction in (1a).

5 Implications for Future Study

These findings show that the generalizations in (HYP 1) and (HYP 2) are not absolute rules, and can be overwritten by lexical idiosyncrasies. Although due to space limitations, this research refrains from providing an analysis, it will be important for future study to examine why these “outliers” exist. The fixed effects findings also require further investigation. For example, the interaction with an imperative is a legitimate concern, which would be better discussed from pragmatic perspectives (Yamada, 2019b). The issue of auxiliary status can also be approached via the grammaticalization theory. Whatever theoretical framework is adopted, analysis must be based on empirical facts. This topic cannot be approached without data analysis, such as that completed in this study, which provides descriptive desiderata and, thus, serves as a necessary departure point for theoretical investigations.

Acknowledgements

This work was supported by JSPS Grant-in-Aid for Young Scientists (Start-up) #20K21957, and the NINJAL collaborative research project “Cross-linguistic Studies of Japanese Prosody and Grammar.”

References

Levshina, N. (2016) “When variables align: a Bayesian multinomial mixed-effects model of English permissive constructions.” *Cognitive Linguistics*, 27(2): 235-268.

Yamada, A. (2019a). “Competing subject-honorific constructions with predicates of Yamato origin [Wago kigen no doosi to kyoogoosuru sonkeigo koobun].” *Proceeding of the 168th Meeting of the Mathematical Linguistic Society of Japan*, pp. 66-71.

Yamada, A. (2019b). “An OT-driven dynamic pragmatics: high-applicatives, subject-honorific markers and imperatives in Japanese.” *Proceedings of Logic and Engineering of Natural Language Semantics 16*, pp. 172-185.