Development of Acceptability Rating Data for Japanese (ARDJ): An Initial Report*

Kow Kuroda[1], Hikaru Yokono[2], Keiga Abe[3], Tomoyuki Tsuchiya[4], Yoshihiko Asao[5], Yuichiro Kobayashi[6], Toshiyuki Kanamaru[7], and Takumi Tagawa[8]

[1]Kyorin University, [2]Fujitsu Laboratories Ltd., [3]Gifu Shotoku College, [4]Kyushu University, [5]NICT, [6]Nihon University, [7]Kyoto University, [8]Tsukuba University

1. Introduction: For proper description of acceptability

Many linguists rely on **acceptability**, or **grammaticality** in its special case, when they study language. It does not seem, however, that they fully understand what acceptability is and needs to be [2]. In fact, nobody has ever explored the entire acceptability space, i.e., the possibility (hyper)space of acceptabilities.

We decided to fill this gap in the case of Japanese linguistics. This is why we are building the **acceptability rating data of Japanese (ARDJ)**. We have recently finished a pilot study and obtained preliminary results. While Japanese is our targeted language, the methodology is easily applicable to other languages. This report gives some findings as well as the new methodology we adopted.

2. Design of Stimuli

2.1 A challenge

Our goal was to take the possibility space of acceptability and probe into it using both acceptable and unacceptable sentences. We need a large number of instances for both kinds, because the possibility space of acceptabilities is likely to be quite large, at least theoretically. We need to incorporate gradience between fully acceptable sentences and fully unacceptable sentences, including mildly deviant ones. How to achieve this aim?

The challenge we faced was the ironic fact that humans, even professional linguists, are really bad at systematic construction of **unacceptable** sentences. This is the case for two reasons: first, people can barely produce deviant sentences without a lot of training; second, even trained people, most of which are professional linguists, can only produce deviant sentences which are biased in many ways. It was thus obvious that we should not use human-generated sentences for probes. What we needed was a large set of unbiased, least theory-laden sentences for probes, which were systematically generated. After investigation, we decided to adopt a mutation-inspired approach and implement it in the way to be specified in §2.2.

*Contact person is the first author, who can be reached at kow.k@ks.kyorin-u.ac.jp.

2.2 Mutation-inspired generation

To meet the requirements above, we hit upon the idea of automatic generation of stimuli inspired by mutation on DNA, with which the possibility space is explored via "random walks." The basic idea is the following:

(1) Steps of randomized generation

Step 1. Construct sentences, called "originals," $O = \{o_1, o_2, \ldots, o_n\}$ with or without deviance.

Step 2. Introduce random mutations to instances of O. Let M denote the result.

Step 3. Mix O and M and use its subsets for acceptability rating tasks.

More details of Step 2 are given in the following:[1)]

(2) A) Random replacement of a lexical item under POS-identity (edit type: l(exical)); B) Random replacement of a postpositional case-marker (P) with another (edit type: p(ositional)); C) Random positional exchange of a given pair of NPs (or PPs) (edit type: s(wapping)).

There are many factors involving in acceptability. Actually, there are too many. It was not known how many factors we had to control. We ran the pilot study being reported here to estimate them. For this purpose, we constrained nominal and verbal mutations for random lexical mutation.

Mutation of (B) type requires similarity data, without which we will get too deviant candidates to explore the acceptability space effectively. We built one using word2vec method [3] implemented in `gensim` [5] package.

2.3 Construction of originals

The sentences we used in the experiment were constructed in the four steps. i) selection of verbs, ii) selection of templates to be lexicalized, iii) manual elaboration of the templates, and iv) mutation introduction and filtering. We give some relevant details of (i, ii) below.

Selection of verbs We selected 9 verbs in (3)[2)] based on frequency data available from NINJAL-LWP for *Bal-*

[1)]Pythons scripts were developed to perform three processes in (2). We plan to make them public.

[2)]The first author was planning to write a paper why these verbs were selected, but he didn't have enough time.

anced Corpus of Contemporary Written Japanese (BCCWJ) [3].

(3) 22: *iku* (行く) [go]; 26: *shiru* (知る) [know]; 44: *kanjiru* (感じる) [feel]; 116: *kotae-ru* (答える) [answer]; 326: *damaru* (黙る) [be(come) silent]; 338: *makeru* (負ける) [lose]; 377: *tutawaru* (伝わる) [carry, propagate, get through]; 1147: *shiri+au* (知り+合う: VV-compound) [know each other]; 1197: *kansen+suru* (感染+する: NV-compound) [acquire, contract, catch, develop (a disease)] [Note: ID's correspond to the frequencies in BCCWJ data.]

Construction of original/seed sentences The 9 verbs above were inserted into the V slots of the four templates in (4) and then the slots for nominals, indicated by "__", were lexically realized by humans (i.e., the first, third, fourth, and fifth authors):

(4) P1: __-ga __-de __-ni __-to V-(shi)ta [ex. s111: *Douryoo-ga shitumon-de aite-ni ina-to kotae-ta*];
P2: __-ga __-de __-ni __-wo V-(shi)ta [ex. s151: *Kazokudure-ga shiohigari-de umi-ni kai-wo sagashi-ta*];
P3: __-ga __-de __-wo __-ni V-(shi)ta [ex. s197: *Kanojo-ga tegami-de shinjitu-wo fui-ni shit-ta*];
P4: __-ga __-de __-kara __-wo V-(shi)ta [ex. s71: *Horyo-ga jinmon-de chuuseishin-kara himitu-wo damat-ta*]

Note: *-ga* is the nominative marker, *-de* the instrumental/locative marker, *-ni* the goal/recipient marker, and *-shita* the perfective marker of verb.

2.4 Characteristics of stimuli

We then introduced random mutations and selected 167 cases so that we have 200 sentences in total. The result consists of 33 originals and 167 mutations. The size of stimuli, 200, was determined based on the size of participants available for experiment. The derived sentences underwent edit of either n-, v-, p-, or s-type. Their numbers and proportions are the following: o-type: 0.165% (33/200) n-type: 0.245% (49/200); v-type: 0.145% (29/200); p-type: 0.180% (36/200); s-type: 0.265% (53/200).

3. Experiment

3.1 Arrangement of stimuli

The 200 stimuli were randomly separated into 10 groups, gr0, gr1, . . . , gr9, each of which consists of 20 stimuli. Each group had four versions, A, B, C and D, in which the orders of the 20 sentences were randomized differently.

3.2 Task: Acceptability rating

For each of the 20 sentences, participants were asked to select one of the four values in (5):

(5) **1.** [felt no deviance]; **2.** [felt a slight deviance but found little difficulty in comprehension]; **3.** [felt a noticeable deviance and had difficulty in comprehension]; **4.** [found erroneous and incomprehensible].

We avoided asking simply if each sentence was acceptable or not, because we had already known that the setting did not have enough descriptive power. In addition, we avoided using the central/neutral value so that this simulates forced choice task.

3.3 Raters

3.3.1 Collection of attributes

Desirably, a survey for acceptability rating/judgment is conducted as a social survey. This is partly because we simply do not know exactly what factors contribute to acceptability to what extent, and partly because personal developmental history is sure to affect acceptability. [1] Under this consideration, we additionally collected 10 attributes in (6):

(6) **Attributes collected:** **1**) age [number]; **2**) sex [f/m/other]; **3**) native place [number encoding prefecture]; **4**) if Japanese is the rater's mother tongue [Yes/No]; **5**) if the rather has lived in other country or countries for more than one year [Yes/No]; **6**) the number of foreign languages the rater has learned [number]; **7**) the total length of foreign language learning [number]; **8**) if the rater has a frequent contact with non-Japanese speaking foreigners [Yes/No/Don't know]; **9**) the number of books read in a month [number]; **10**) years of education [number].

3.3.2 Participants

We collected responses at three different places, Tokyo[4], Gifu and Fukuoka with the aim of reducing the location bias as much as possible. We had 93 participants in Tokyo, 109 participants in Gifu, and 49 participants in Fukuoka, and 251 participants in total. Later, 35 of them were excluded under a statistical measure.[5] So, 216 is the number of effective participants.

In the end, each of the 200 stimuli were rated by 21.3 raters on average (max 29, min 15).

4. Analysis and Results

4.1 Data encoding

For each stimulus, we calculated a response probability distribution on four rating value ranges, i.e., [1,0], [2,1), [3,2) and [4,3), and treated them as representational vectors.

[3] Available at `http://nlb.ninjal.ac.jp`

[4] This experiment was ran by Shunji Awazu (Toyo University), who kindly helped us in this survey.

[5] The exclusion procedure was rather complicated to describe succinctly. Suffice it to say that if responses had either too little or too much standard deviations in the respective groups, they were excluded as illegitimate.

4.2 Response patterns (of a random sample)

Plotting all responses does not give an intelligible result. Instead, Fig. 1 gives a graph of a random sample of 20.

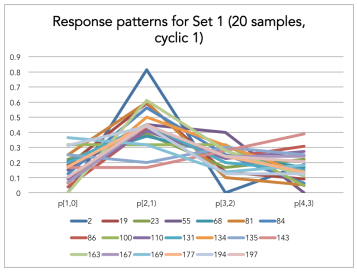

Figure 1: Patterns of 20 random samples (cycle=1)

4.3 Clustering responses by edit type

To begin with, we clustered responses by edit types. It turned out that this classification did not give us interesting results: no significant discrepancies were detected.

For space limitation, we show only results of o-type and p-type. Fig. 2 gives a graph of responses to the o-type stimuli. Fig. 3 gives a graph of responses to the p-type stimuli.

4.4 Clustering responses using k-means

We then tried k-means method. Fig. 4 gives the result of k-means clustering ($k = 6$) of all the 200 stimuli.[6] Membership under hard clustering is the following: C6.1: 55 instances (maximum); C6.2: 41 instances; C6.3: 20 instances; C6.4: 34 instances; C6.5: 7 instances (minimum); C6.6: 43 instances.

4.4.1 Properties of C6.1, C6.2, . . . , C6.6

How do the six cluster differ? The properties of each cluster can be stated in terms of partial ordering as those in (7):

(7) C6.1: [2,1] > **[1,0]** > [3,2], [4,3] (mild deviance 1)

C6.2: [2,1] > **[1,0]**, [3,2], [4,3] (mild deviance 2)

C6.3: **[1,0]**, [2,1] > [3,2] > [4,3] (slight deviance)

C6.4: [4,3], [2,1] > [3,2] > **[1,0]** (strong deviance)

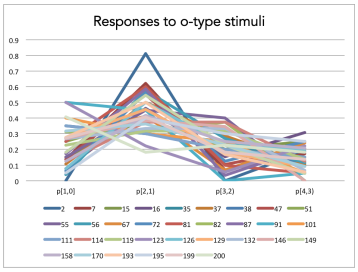

Figure 2: Responses to o-type stimuli

[6] We tried other cases where $k = 5$ and $k = 7$ and concluded that $k = 6$ gave the best result.

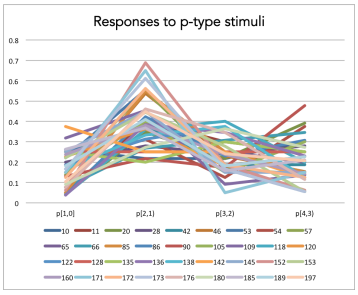

Figure 3: Responses to p-type stimuli

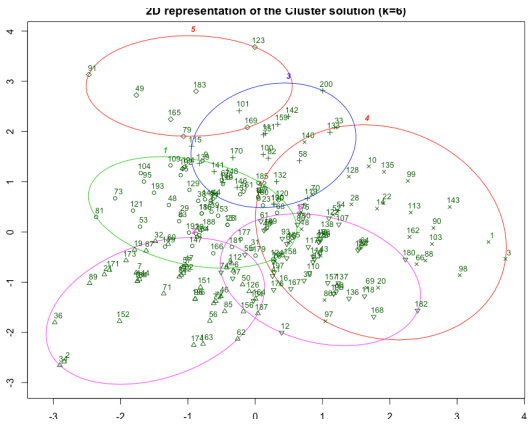

Figure 4: k-means clustering ($k = 6$)

C6.5: **[1,0]** > [2,1] > [4,3] > [3,2] (no deviance)

C6.6: [2,1], [3,2] > **[1,0]**, [4,3] (mild deviance 3)

This suggests the following: 1) C6.5 collects stimuli with no (detectable) deviance; 2) There are two distinct directions of deviance, encoded by C6.2 and C6.4, respectively; and 3) C6.3 is a mixture of C6.5 and C6.4; C6.1 a mixture of C6.5 and C6.2; and C6.6 a mixture of C6.2 and C6.4;

4.5 Classfication by clusters

For space limitation, we show only graphs of the responses in C6.2, C6.4 and C6.5. Fig. 5 gives a graph of the responses in C6.2. Fig. 6 gives a graph of the responses in C6.4. Fig. 7 gives a graph of the responses in C6.5.

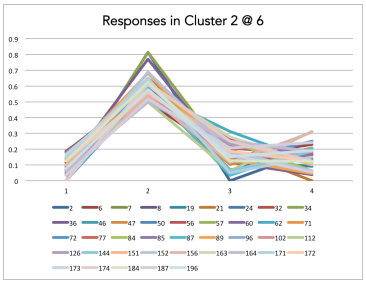

Figure 5: Responses to stimuli in C6.2

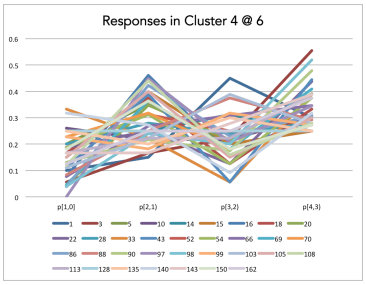

Figure 6: Responses to stimuli in C6.4

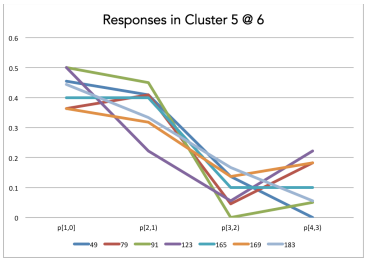

Figure 7: Responses to stimuli in C6.5

4.6 Cross-comparison of clusters

We then compared clusters in terms of proportion of edit types. Fig. 8 gives the proportions of responses to o-, n-, v-, p- and s-type stimuli in each clusters. Different clusters have significantly different proportions of edit types. If this was not an accident, we can conclude the following: 1) v-type mutations drastically lowers acceptability; 2) n-type mutations change acceptability in both directions, and the effects counterbalance; 3) s-type mutations change acceptability noticeably, and sometimes raise it for unknown reasons; and 4) p-type mutations lower acceptability, but not drastically.

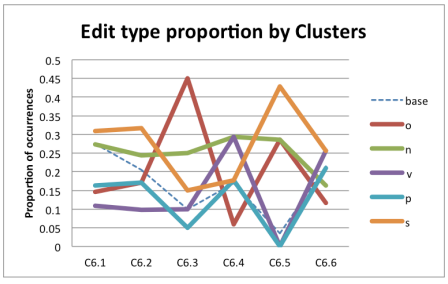

Figure 8: Type proportions in 6 clusters

5. Discussion

What we did was a pilot study on a small scale. We only ran superficial analyses such as k-means clustering which do not tell much. But some results are worth mentioning.

5.1 Findings

First, simple classification by mutation type revealed no differentiation in response. We admit that this was surprising: we expected otherwise. Second, we confirmed that deviances were brought about in two distinct dimensions, though we are not certain what they really are. Further investigation awaits this finding. Third, and perhaps most interestingly, cluster analysis suggested differentiated effects of mutation types into clusters. Six clusters identified through k-means method ($k = 6$) comprise show different proportions of edit types. While this runs counter to the first finding, this would have simply told us that simple across-the-board classification was not informative enough.

5.2 Future work

Comparison among response classes by originals revealed that different originals receive different response patterns. We are currently trying to group the response patterns. In addition, we will also do the following in future: 1) classification of raters, 2) research into interaction between personal attributes and response patterns. And surely, we will run full version of experiments where more variations are added.

6. Concluding Remarks

We have not completed a full analysis and any of our conclusions are inevitably tentative. We believe, however, that our project saw a good start, so that we can expect a full survey will bring on more interesting findings, including suprising and even counterintuitive ones.

Acknowledgements

The current research was supported by JSPS Grant 16K13223.

References

[1] E. Dąbrowska. The LAD goes to school: A cautionary tale for nativists. *Linguistics*, 35(4):755–766, 1997.

[2] E. Dąbrowska. Naive vs. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, 27:1–23, 2010.

[3] T. Mikolov, K. Chen, G. Corrado, and D. Jeffrey. Efficient estimation of word representations in vector space. 2013. `arXiv:1301.3781`.

[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[5] R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta, May 2010. ELRA.