Insights from a large scale web survey for Acceptability Rating Data for Japanese (ARDJ) project*

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

Acceptability plays a crucial role in linguistic theorizing. But it is far from fully understood how "ordinary" people react to sentences with varied degrees of deviance or anomaly, simply because there is no such data. ARDJ is a project that aims at exploring true nature of acceptability judgment/rating based on a large-scale survey with least theoretically biased stimuli.

ARDJ has done two experiments for this purpose. The first one, called "survey 1," was carried out in 2017. It was intended to be a pilot study with only a limited variety of responders (roughly 200 college students only) on 200 sentences. The results we obtained are reported in [3].

The second experiment, called "survey 2," was carried in 2018. It was the main study of the ARDJ project, with the expansion of stimulus set to 300. It has two phases: phrase 1 is a small scale experiment in which roughly 300 college students participated, which is comparable to the pilot study done in 2017. Phase 2 is a large scale web survey in which responses were obtained from over 1,600 participants with significantly more variations in their attributes. This paper reports on phase 2 of ARDJ survey 2.

2. ARDJ survey 2, Phrase 2

2.1 Stimuli construction

Table 1: Mutation types and ratios in survey 2 stimuli

edit.type	count	ratio
o(riginal)	36	0.12
s(wapping)	70	0.23
p(ostposition)	58	0.19
v(erb)	65	0.22
n(ominal)	71	0.24
sum	300	1.00

Effective exploration into the possibility space of acceptability requires careful manipulation of stimuli. It is evident, however, that humans are not capable of producing a large number of potentially and actually deviant sentences systematically, and without theoretical biases. This necessitates automatic method. The procedure we followed was the semi-automatic method described in [3]. In summary, candidates for deviant sentences were constructed by randomly replacing lexical items on either nominal, verbal or positional sites, or by randomly swapping of any pair of phrases in them. Mutation types and their ratios in survey 2 are given in Table 1.

Survey 2 exploited the 27 Japanese verbs in Table 2.

Table 2: Verbs used in survey2

v.index	v.form	count	ratio
v18	聞いた	23	0.077
v22	行った	1	0.003
v25	入れた	16	0.053
v26	話しかけた	1	0.003
v40	教えた	3	0.010
v44	感じた	10	0.033
v111	伝えた	17	0.057
v116	答えた	1	0.003
v131	探した	1	0.003
v145	聞こえた	17	0.057
v155	繰り返した	9	0.030
v183	届いた	8	0.027
v210	遊んだ	17	0.057
v326	黙った	2	0.007
v338	負けた	1	0.003
v345	助けた	17	0.057
v377	表れた	1	0.003
v447	つないだ	8	0.027
v450	載った	18	0.060
v470	襲った	29	0.097
v713	間違った	9	0.030
v807	直した	23	0.077
v829	助かった	9	0.030
v831	届けた	17	0.057
v958	習った	29	0.097
v1147	知り合った	12	0.040
v1197	感染した	1	0.003
sum		300	1.000

To start with, candidates were selected semi-randomly from the verb list of NINJAL-LWP for BCCWJ[1)]. They were manual coded with 6 attributes, F1: [effect is physical], F2: [effect is mental], F3: [effect is social], F4: [event is interactive], F5: [event is interactional], F6: [effect is intended (if subject is agent)]. Kuroda [5] gives a detailed report of the data and analysis. Then, we obtained the Formal Concept Analysis [1] presented in Figure 1, where verbs with red frames correspond to the candidates for the supplement to the 9 verbs used in survey 1.

Candidate sentences were constructed by manually fill-

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

1)http://nlb.ninjal.ac.jp/search/

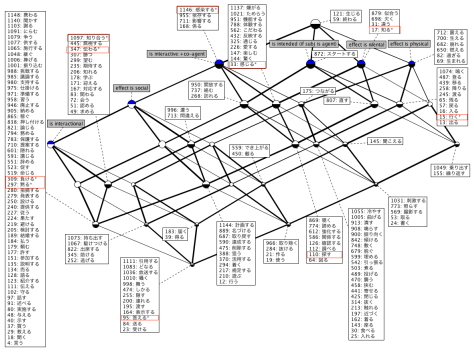

Figure 1: FCA of sampled verbs

ing in the four lexical gaps in the constructions in (1).

(1) P1: ___ が ___ で ___ に ___ と V した.
P2: ___ が ___ で ___ に ___ を V した.
P3: ___ が ___ で ___ を ___ に V した.
P4: ___ が ___ で ___ から ___ を V した.
P5: ___ が ___ で ___ と ___ を V した.

Each stimulus group contains 30 sentences, of which s281 and s282 were carried over from survey 1, and were used as common ground. 280 sentences were randomly sampled from the candidates to fill in the other 28 sentences. Table 3 presents samples from gr0.

2.2 Rating task

Questionnaire was implemented by Google Forms. Each group consists of 11 questions for personal and/or social attributes specified in (3), followed by 30 questions for acceptability rating. On rating the stimuli, participants, whom we call "responders," are asked to choose one of the four choices on a scale in (2).[2)]

(2) **0.** natural and easy to understand [違和感がなく自然に理解できる文]; **1.** more or less deviant but comprehensible [違和感を感じるが理解可能な文]; **2.** deviant and difficult to understand [違和感を感じて理解困難な文]; **3.** quite unnatural and incomprehensible [不自然な理解不能な文]

Prefixed indices roughly encode the degrees of deviance.

Note that the questionnaire does not directly ask if such and such sentence are acceptable (or not). We avoided it for two reasons. First, the very notion of acceptability is not as simple as ordinary people can fully appreciate. Second, we aimed at factoring out of acceptability, by explicitly trying to dissociate semantic anomaly from grammatical (morphological and/or syntactic) anomaly.

2.3 Social attributes collected

ARDJ is primarily a project of data collection targeting acceptability ratings. Aware of the social nature of grammar, we found it necessary to embed our data collection within the framework of social survey. Thus, we collected the 11 social attributes specified in (3). Layered analysis using those social attributes is planned, but not carried out yet.

(3) **Q1.** [Age] How old are you now?; **Q2.** [Gender] Which gender is yours? (Male/Female/Unsure); **Q3.** [Native language] Is Japanese your mother tongue? (Yes/No/I don't know); **Q4.** [Lived places] Which area have you lived in the past? Answer as many prefectures as necessary); **Q5.** [Experience of living abroad] Have you lived for more than one year in an area where people don't speak Japanese?); **Q6.** [Number of known languages] How many languages have you ever learned? Answer the number of them irrespective of their duration; **Q7.** [Length of foreign language learning] How long did you spend learning foreign languages? Answer in number of years); **Q8.** [Daily contact with foreigners] Do you have regular contacts with those who speak foreign languages?); **Q9.** How many books do you read in a month, roughly?; **Q10.** How many years did you spend learning after elementary school? Answer in the number of years; **Q11.** Which intellectual orientation describes you better? (Science-oriented, relatively science-oriented, neutral, relatively humanity-oriented, humanity-oriented)

Some facts. The ratios of male and female responders are nearly equal. The ages of responders range from "under 13 yo" to "over 70 yo," as illustrated in Figure 2. Weak over- and undersampling can be pointed out, but overall effect is not bad. The rank distribution of lived places (in terms of prefecture) is as illustrated in Figure 3. It is safe to say that places are well diversified, as far as we take Zipf's law into account.

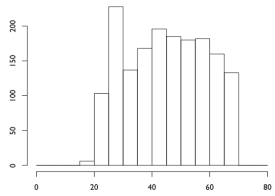

Figure 2: Age distribution over 5-year ranges

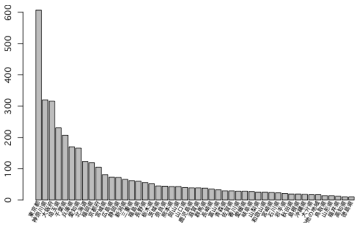

Figure 3: Rank distribution of lived places

2.4 Outlier removal: data cleaning

Outliers in survey are not rare because responders always have incentive to play "satisficing" [2]. This means that

[2)] One of the responders pointed out that "違和感を感じる" was a wrong expression, and claimed that this error debunked the authenticity of the survey.

Table 3: Sample stimuli in gr0

s.index	v.idex	pattern	author	type	gr	ver	gr.index	sentence
s10	v25	P4	3	n	gr0	A	1	担当者が携帯で出張もさから電話を入れた。
s30	v831	P3	1	n	gr0	A	3	伝書鳩が戦地で進攻を司令官に届けた。
s50	v831	P3	1	v	gr0	A	5	伝書鳩が戦地で戦況を司令官に送り届けた。
s70	v345	P1	3	s	gr0	A	7	続編で宿敵がピンチに主人公と助けた。
s100	v470	P4	2	o	gr0	A	10	暴漢が鋭利な刃物で背後から人を襲った。
s140	v958	P5	3	v	gr0	A	14	弟が家で妹と料理を習わせた。
s170	v145	P3	1	v	gr0	A	17	ランナーが路上で悲鳴を夕暮れ時に聞き取れた。
s210	v345	P1	3	n	gr0	A	21	宿敵が続編で苦境に主人公と助けた。
s250	v958	P1	1	s	gr0	A	25	医学生が解剖実習で看護師と医師に習った。
s281.0	v1147	P1	1	p	gr0	A	29	夫が職場で真夜中に妻へ知り合った。
s282.0	v44	P4	1	n	gr0	A	30	学生が合格発表の場で足下から幸福を感じた。

survey data always comes with outliers, at least potentially. Analysis goes awry when their effect is ignorable.

In survey, there are predictably two typical ways of playing satisfice, and fortunately, the two kinds of satisficing can be easily detected by statistical assessment of data. In survey, case 1) a responder always returns the same value to all questions; or case 2) a responder chooses values fully or nearly randomly. The first case can be detected if someone's responses do not have enough standard deviation. The second case can be detected if someone's responses shows too much randomness, which can be measured using such measures as Mahalanobis distance.

Table 4: Distribution of outliers (M-dist discard rate = 0.05)

ID	#sd	#M-dist	#shared	#unified	#effectives
gr0	6	8	1	13	153
gr1	5	9	0	14	160
gr2	8	8	0	16	153
gr3	5	8	0	13	155
gr4	7	8	0	15	151
gr5	10	8	0	18	145
gr6	3	9	0	12	166
gr7	5	8	0	13	155
gr8	7	8	0	15	147
gr9	4	8	0	12	152
sum	60	82	1	141	1,538

We used double filters to remove the outliers in the data at hand: responses in each group are discarded if either not $0.5 < \mathrm{sd} < 1.6$ or out of the 95% range[3)] of Mahalanobis distance. Table 4 gives a brief description of the statistics.

2.5 Overview of the obtained data

We have 300 sentences to examine and it is clearly unrealistic to give all the plots here. So, we present only 24 cases nearly randomly selected, given in Figure 4.

The selection here is nearly random, because, first, the 10 groups were randomly constructed and second, the presentation order of the 30 stimuli was randomized. What is not random about this selection is inclusion of s281 and s282, which are common to all stimulus sets.[4)]

Inspection of the 300 responses, sampled in Figure 4, leads to the following insights: i) the same stimuli produce

[3)]This discard rate was determined through trial and error. It is possible to increase the number of common outliers, but discard rate needs to go up to 0.3 if we want to each group has at least one common between sd-based and Mahalanobis-based detections.

[4)]s281.i and s281.i encode the stimuli in group i.

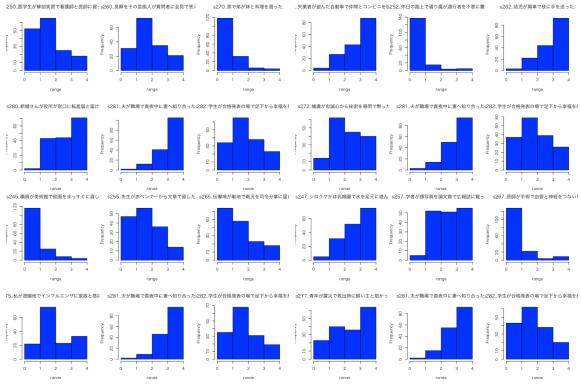

Figure 4: Selected response patterns (last 6 stimuli for gr0, gr2, g5, gr7)

roughly the same responses, as are the case with s281 and s282, which are common to all groups; ii) despite this, responses are rather varied: different sentences are likely to cause different responses; iii) still, responses are far form random. These findings, if combined, suggest a hypothesis that there seem to be only a limited number of classes of responses, thereby predicting a typology of response patterns.

3. Analysis of responses

The suggested existence of response typology naturally begs the question, "How to recognize response classes, and associate each to other?"

Hierarchical clustering is a popular method for grouping data. Principal Component Analysis (PCA) is a popular method for revealing a simple geometry in the data. An `R` package `FactoMineR` [4] provides the combination of the two. We decided to use it to reveal the hidden structures in the data.

3.1 Standardizing responses

We needed one intermediate step, however. Note that gr0, gr1, ..., gr9 are different data sets, and they cannot be directly compared. Comparison of them requires standardization. It was carried out in the following way: sentences in each set were transformed into binned density array, $p[0,1), p[1,2), p[2,3), p[3,4)$, over four rating ranges $r[0,1), r[1,2), r[2,3), r[3,4)$. All sentences in our dataset correspond to quadruples $(p[0,1), p[1,2), p[2,3), p[3,4))$.

These are commensurable even if the sets of responders are different for groups.

3.2 PCA and hierarchical clustering

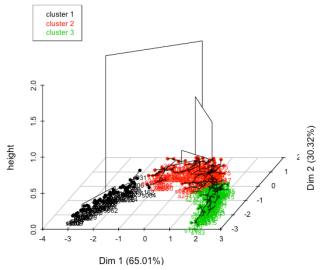

Figure 5: HCxPCA of combined responses in gr0-gr9

Building on standardized responses, a multivariate analysis was conducted where PCA was combined with hierarchical clustering, resulting in visualization in Figure 5. In this, we recognize three major classes of stimuli: clusters 1 (in black, of acceptable stimuli), cluster 2 (in red, of undecidable stimuli) and cluster 3 (in green, of unacceptable stimuli). Clusters 2 and 3 form a larger cluster.

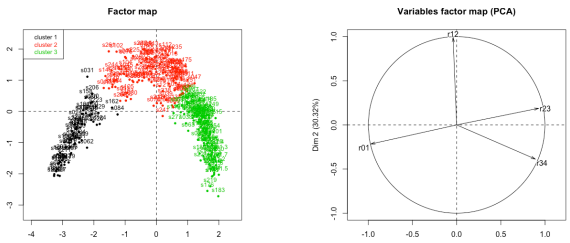

Figure 6: Factor Map/PCA of combined responses

The first dimension (Dim 1) of PCA roughly corresponds to the polar opposition of $r[0,1)$ and $r[2,3)$. The second dimension (Dim 2) is mildly encoded by $r[1,2)$, and weakly by $r[3,4)$.

The interpretation of Dim 1 is straightforward. It encodes the degree of deviance from right to left, or of acceptability, from left to right. In contrast, the interpretation of Dim 2 is not as simple as Dim 1. A few likely interpretations come to mind, but the most convincing one would be that Dim 2 encodes semantic and/or syntactic complexity that often blurs the judgment. At least, this seems to provide a natural account for the parabolic shape of the area of plotted points.

We should point out here that response structures local to gr0, ..., gr9 are similar to the one in Figure 6, thereby predicting that the geometry is an invariant self-similar image.

4. Discussion

We have seen that 1) unsupervised grouping (i.e., hierarchical clustering) of responses resulted in three clusters rather than two clusters; 2) dimension reduction of responses revealed that acceptability is encoded by Dimension 1, with correlation with the polar opposition between

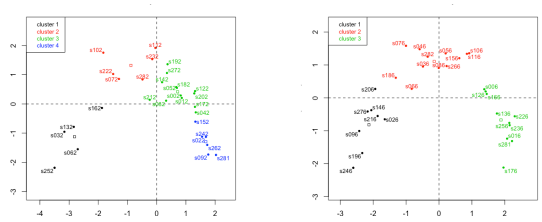

Figure 7: PCA/Factor Map of responses in gr2 and g6

$r[0,1)$ and $r[2,3)$, but Dimension 2 cannot be ignored. The first point confirms that acceptability cannot be properly modeled as a categorical judgment. The second point, coupled with the first point, strongly suggests that acceptability is not a monolithic notion and is likely to be affected by a number of factors.

The analysis of phrase 1 of survey 2, which is not described in this paper, gave us different profiles in terms of PCA/Factor Map. The shape of data points is not parabolic. The two results from phases 1 and 2 need to be reconciled.

5. Conclusion

Analysis of responses from a large scale survey, through its unsupervised clustering and dimension reduction, revealed some true properties of acceptability judgment. Our results are robust, based on the unbiased nature of stimuli, scale of survey and controlled varieties of responders.

Let us close with future plans. First and foremost, we make the obtained data public. Second, we try out layered analyses. Does the result involve gender difference? We suspect so. Does the result involve location/dialect difference? We suspect so. They are definitely worth trying.

Acknowledgement

This research was supported by JSPS through Grant-in-Aid (16K13223).

References

[1] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, Berlin, 1999.

[2] Jon A. Krosnick. Response strategies for coping wiht the cognitive demands of attitude ameasure in surveys. *Applied Cognitive Psychology*, 5(3):213–236, 1991.

[3] Kow Kuroda, Hikaru Yokono, Keiga Abe, Tomoyuki Tsuchiya, Yoshihiko Asao, Yuichiro Kobayashi, Toshiyuki Kanamaru, and Takumi Tagawa. Development of Acceptability Rating Data of Japanese (ARDJ): An initial report. In *Proc. of the 24th Annual Meeting of the Association for NLP*, pp. 65–68, 2018.

[4] Sebastien Le, Julie Josse, and Francois Husson. FactoMineR: An R package for multivariate analysis. *J. of Statistical Software*, 25(1):1–18, 2008.

[5] 黒田 航. 意味の社会性を意識した動詞の分類とその理論的含意. In **認知科学会第** *35* **回大会発表論文集**, pp. 65–68, 2018.