Simulating acceptability judgments using ARDJ data

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]National Institute of Communications and Information Technology [6]Nihon University, [7]Kyoto University, [8]Tsukuba University

1 Introduction

Acceptability judgment plays a crucial role in linguistic theorizing [5, 15, 17, 19]. 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 provided with high reliability. In our understanding, current theoretical linguistics is a strange mixture of best and worst practices [4]. In its best practice, it provides deep insights into human mind/brain. In its worst practice, research directions are under the strong influence of a few authorities, and discussions are plagued by confirmation biases [1, 2, 9]. This is why most "evidence" in current theoretical linguistics often falls close to anecdotes. This situation is reminiscent of the medical practices before "evidence-based medicine," aka EBM, was gaining ground. The essence of EBM is a fairly straightforward idea that all pieces of "evidence" need to be sorted out and hierarchically organized according to their effectiveness [6, 8].

As medicine needed EBM to alleviate the harmful effects of authorities, linguistics seems to follow a similar path: it would need "evidence-based linguistics," or EBL, to be renovated. In our view, *Acceptability Rating Data of Japanese (ARDJ)* is a potential contribution to a virtual movement for EBL in that it explores true nature of acceptability judgments using a sharable, large-scale, randomized survey, based on as little theoretically biased stimuli as possible. Its goal is to provide "proper evidence" in linguistics, though it is only the beginning, we admit.

ARDJ has completed two experiments. 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 for stimuli. Kuroda et al. [10] reported on this survey. The second experiment, called "Survey 2," was carried out in 2018. This was the main study, with the stimulus set expansed to 300 with some overlap with Survey 1. Samples of the used stimuli are presented in Table 1.

Survey 2 had two phases, Phases 1 and 2, and collected responses from 1,880 participants in total.[1)]. Phase 1 was a small scale paper-based survey, to which 201 participants (basically college or university students) contributed responses. This was comparable to the pilot study done in 2017. Phase 2 was a large scale web survey to which 1,679 participants contributed responses. They were significantly more varied in attributes and we would safely state responses obtained were randomized enough.

Kuroda et al. [11] analyzed the data at Phase 2 of Survey 2 and reported the results of Hierarchical Clustering and PCA applied to it, excluding data at Phase 1. See Appendix A for relevant information. The current paper serves as a supplement to the previous report in that it tries to directly simulate human's acceptability judgments rather than simply clustering obtained data. To this end, we conduct two analyses. In Analysis 1, we use Semi-supervised Local Fisher Discriminant Analysis (SELFA) [20] to see how clusters obtained are likely to be related to acceptability judgments. This analysis is exploratory in that it does not simulate acceptability judgments directly. In Analysis 2, we try a direct simulation of human categorical judgments using logistic regression [14] to yield a promising result. But it comes with a surprising suggestion that acceptability judgement is better characterized as a social decision rather than a personal/private one, against the popular view.

2 Partitioning responses by SELFA

In Kuroda et al. [11], 300 stimuli were analyzed using Hierarchical Clustering and PCA. See Figure 6 for relevant information. Results like this are undoubtedly useful, but you would argue they are not enough, because they do not tell us what acceptability *judgment* is. Explicit modeling of it is missing. To this end, we need to implement a function that maps, or "interprets," our data at hand to acceptability judgments, where $P = \langle p[0,1),\ p[1,2),\ p[2,3),\ p[3,\infty)\rangle$ are explanatory variables, and labels **A**(cceptables) and **UNA**(cceptables) (plus undecidables (**X**) if necessary) are objective variables.

2.1 Metrics to evaluate

Our data did not have "acceptable" and "unacceptable" labels as such, though explicit modeling requires them. Thus, we need to generate them, but no obvious way is known.

[1)]The response data we used for analysis is freely available at `https://kow-k.github.io/Acceptability-Rating-Data-of-Japanese/`, but you need to register to use it.

s.index	v.idex	pattern	author	edit 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。
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。
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。
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。

Table 1: Sample stimuli in gr0

We were forced to try out whatever metrics we could think of, ending up with the ones in (1).

(1) a. *Condition 0*: If $p[0,2) > 0.5$ then **A**; else **UNA**.

b. *Condition 1*: If $p[0,2) > \{ 0.6, 0.72, 0.85, \dots \}$ then **A**; else **UNA**.

c. *Condition 2*: If $p[0,2) > p[1,3)$ AND $p[1,3) > p[2,\infty)$ then **A**; else **UNA**.

d. *Condition 3*: If $p[0,2) > 0.5$ then **A**; if $p[2,\infty) > 0.5$ then **UNA**; else **X**.

e. *Condition 4*: If $p[0,2) > 0.5$ then **A**; if $p[1,3) > 0.5$ then **UNA**; else **X**.

f. *Condition 5*: If MAX($p[0,2)$, $p[1,3)$, $p[2,\infty)$) = $p[0,2)$ then **A**; if MAX($p[0,2)$, $p[1,3)$, $p[2,\infty)$) = $p[2,\infty)$ then **UNA**; else **X**.

g. *Condition 6*: If MAX($p[0,1)$, $p[1,2)$, $p[2,3)$, $p[3,\infty)$) = $p[0,1)$ then **A**; if MAX($p[0,1)$, $p[1,2)$, $p[2,3)$, $p[3,\infty)$) = $p[3,\infty)$ then **UNA**; else **X**.

h. *Condition 6r*: If MAX($p[0,1)$, $p[1,2)$, $p[2,3)$, $p[3,\infty)$) = $p[0,1)$ then **A**; if MAX($p[0,1)$, $p[1,2)$, $p[2,3)$, $p[3,\infty)$) = $p[3,\infty)$ then **UNA**; if $p[1,2) > p[2,3)$ then **X**; else **Y**.

where $p[i,j)$ means density over ranges from i next to j.

(1)a,b,d,e are simple dichotomies with particular ad-hoc thresholds (e.g., 0.5, 0.6, ...). In contrast, (1)c, d–h are not simple dichotomies, involving **X** (and **Y**) for buffering, and more importantly they are distribution-aware ones in that MAX(...) is used for decision.

The metrics in (1) are evaluated using Semi-supervised Local Fisher Discriminant Analysis" (SELFA) [20][2] to evaluate the models in (1).[3] We selected this for two reasons: 1) it is resistant to outliers and avoids over-fitting, and 2) it allowed us to seek the best mix of supervised and unsupervised classifications.

To generate simulated "(in)correct judgments," we assigned **A**, **X**, and **UNA** to Clusters 1, 2 and 3 in Figure 6.

[2] An `R` package `ldfa` [21] was chosen for this purpose.

[3] SELFA, like LFDA, has a parameter r to specify the dimensionality of reduced space. $r = 3$ gave us the most reasonable results.

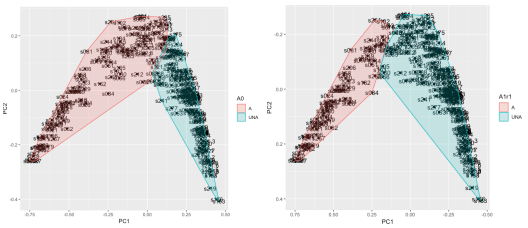

Figure 1: SELFAs (Conditions 0 and 1) of 300 stimuli

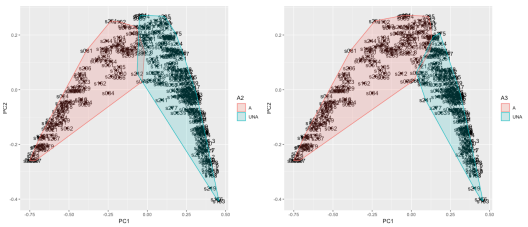

Figure 2: SELFA (Conditions 2 and 3) of 300 stimuli

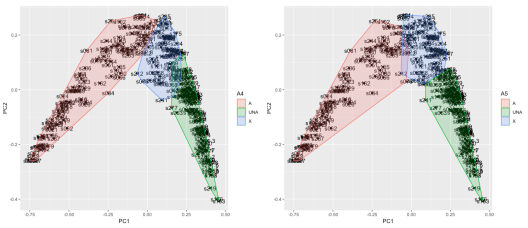

Figure 3: SELFA (Conditions 4 and 5) of 300 stimuli

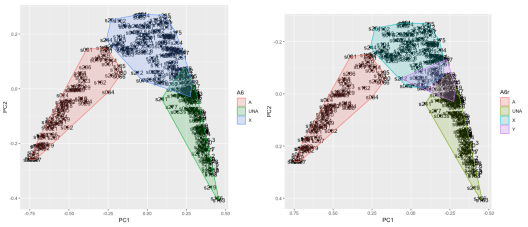

Figure 4: SELFAs (Conditions 6 and 6r) of 300 stimuli

Condition	C0	C1	C2	C3	C4	C4r1	C4r2	C4r3	C5	C6r1	C6
correct count	182	179	184	184	237	245	280	282	256	255	263
match rate	0.607	0.597	0.613	0.613	0.790	0.817	0.933	**0.940**	0.853	0.850	0.877

Table 2: correct counts and match rates for Conditions

Figure 1 shows the SELFAs for Conditions 0 and versions of Condition 1. Figure 2 shows the results for Conditions 2 and 3. Figure 3 shows the results for Conditions 4 and 5. Figure 4 shows the results for Conditions 6 and 6r.

Note that the boundary between Clusters 1 and 2 in Figure 6 in Appendix A does not fit the **A/X/UNA**boundaries in Figures 1 and 2, suggesting that simple threshold-based partitioning is inadequate. The results for Conditions 2 and 3 look better but not drastically. The results for Conditions 4 and 5 in Figure 3 look good, and so do the results for Conditions 6 and 6r in Figure 4. The problem here is that it is hard to select the best result if there is one.

2.2 Simple evaluations of SELFA results

Results in Figures 1–4 are informative, but they cannot be quantitatively assessed. To make our assessment quantitative, we tentatively measured the "correctness" of each metric in (1) by measuring the match rates between simulated "judgments" and cluster assignment interpretations $1 \rightarrow$ **A**, $2 \rightarrow$ **X**, and $3 \rightarrow$ **UNA**.

Table 2 shows its results. It says that in terms of simple match rate, Condition C4r3 performed the best, and Condition C4r2 the next best. Conditions 4rN are variants of Condition 4 in (1e) with threshold values other than 0.5 determined by manual tweaking. Specifically,

(2) a. Condition 4r2: If $p[0,2) > 0.8$ then **A**; if $p[1,3) > 0.60$ then **UNA**; else **X**.

b. Condition 4r3: If $p[0,2) > 0.8$ then **A**; if $p[1,3) > 0.65$ then **UNA**; else **X**.

These results are far from definitive, however. Note that correctness thus defined is artificial, if not arbitrary. While we do not believe that this setting is seriously unrealistic, but we are not certain how much truth it can capture. Definitely, a better modeling is in need.

3 Logistic regression

3.1 Analysis 2 and its results

Our strong motivation in the current research is to see if we can construct a reasonable model for acceptability judgment as as a categorical decision. SELFA successfully link clustering results to the models in (1) but does not give us exactly what we need. We need to take another route.

For this purpose, logistic regression [14] was applied to the discrimination models listed in (1). This was done for two purposes. First, we had the impression that Condition 6(r) gave a better fit qualitatively and wanted to see whether our intuition was correct or not. Second, we wanted to see if parameter-free models could work, because C4r2, and C4r3, high performers, require manual parameterization for thresholds.

Logistic regression was performed using `glm` package for `R` with the formula as follows:[4)]

(3) decision $\sim p[0,1) + p[1,2) + p[2,3) + p[3,\infty)$
where decision = 1.0 if label is **A** (or **X**); otherwise, decision = 0.0.

Roughly, the formula in (3) checks if probability p of **A** (with or without **X**) against probability $(1-p)$ of **UNA** is predictable from the following estimate:

$$\ln \tfrac{p}{1-p} = c + w_1\, p[0,1) + w_2\, p[1,2) + w_3\, p[2,3) + w_4\, p[3,\infty)$$

where c, w_1, …, w_4 are given in Table 3. The left-hand side is the log odds of p against $(1-p)$ and the right-hand side is a linear combination of $p[0,1)$, $p[1,2)$, $p[2,3)$ and $p[3,\infty)$ with appropriate weights and an intercept c.

Two settings were tried out for comparison. In one setting, only **A** was set to 1.0, and both **X** and **UNA** were into 0.0. In another, **A** and **X** were converted into 1.0 and only **UNA** was into 0.0, on interpreting **X** as part of **A**.

In the first setting, none of the 11 conditions, C0, C1, …, C6, C6r1, reached convergence. In the second, only C6 and C6r1 reached convergence[5)] with the estimation presented in Table 3, and the others all failed. This indicates that **X** needs to be treated as **A** for convergence. But admittedly this contradicts with the relative closeness of Clusters 2 and 3. We will return to this in §3.2.

	Estim.	Std. err.	z-val.	Pr(>|z|)	signif.
Interc c	-82.7	25.2	-3.29	0.0010	**
w_1	94.9	31.3	3.04	0.0024	**
w_2	120.3	36.5	3.30	0.0010	***
w_3	135.5	42.2	3.21	0.0013	**
w_4	NA	NA	NA	NA	

Table 3: Coefficients for C6 and C6r1 (significance codes: 0 '***'; 0.001 '**'; 0.01 '*')

In this regression, w_4 turned out to be ineffective. In other words, the probability of **A** can be predicted from three values $p[0,1)$, $p[1,2)$ and $p[2,3)$ only.

[4)] Used link function is binomial.

[5)] Incidentally, the fittings of C6 and C6r1 returned the same result.

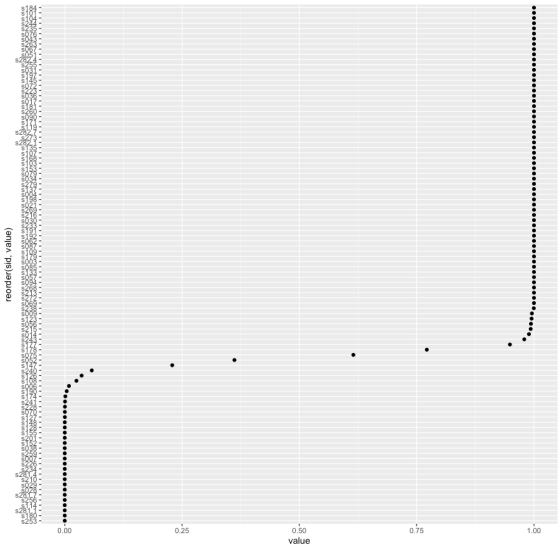

Figure 5: Predicated acceptabilities by Condition 6(r1)

To give readers a good grasp of the result, Figure 5 plots the predicted values (between 0.0 and 1.0) for sampled 100 sentences.[6)]

Quite interestingly, only C6 and C6r1 reached convergence and all other Conditions, including C4r2 and C4r3, high performers in simple metric, failed to converge. This result is noteworthy, because it suggests that effectiveness in terms of simpler measures can be faulty.

3.2 Discussion for Analysis 2

The status of Cluster 2 (in red, of **X** stimuli) in Figure 6 in Appendix A is troublesome. Are they part of **A** or **UNA**? This is a choice of theoretical importance that makes a difference. In Analysis 2, we compared two cases where **X** is included into **UNA** and alternatively it is included into **A**. The result revealed that the second treatment was necessary for convergence. Though it is far from definitive, it suggests that the best way to treat **X** is to include it into **A**.

But this conclusion cannot be debate-free, because in terms of closeness of clusters, Clusters 2 and 3 form a larger cluster. So, the result is fairly puzzling, making the obtained solution look somewhat opportunistic. We admit that we can offer no good explanation for this.

Whatever status Cluster 2 has, though, it would not be a serious problem as far as we consider the preconditions for convergence of logistic regression. Under logistic regression, only Conditions 6 and 6r1 converged and all other conditions failed. Given this result is not accidental, it poses interesting theoretical implications for the question of what mental process acceptability judgement demands.

Note that Conditions 6(r1) uses MAX(...) function. Because **X** is included in **A**, Condition 6(r) is equivalent to:

(4) If MAX($p[0,1)$, $p[1,2)$, $p[2,3)$, $p[3,\infty)$) = $p[3,\infty)$ then **UNA**; else **A**.

This evaluation metric is "collective," in the sense of "collective intelligence" [13, 18], or at least "population-aware," in that it is density-based. But how can a judgment be population-aware? What is crucial is that each rater should "know" how others respond, or at least they should be able to relativize their own ratings to those by other raters, most likely performing a kind of mental simulation of judgements by others. Why is this so? Because, otherwise, higher function like MAX(...) would not be unnecessary.

If this conclusion is correct, it suggests that acceptability judgment is not only a fairly complex decision, but also a "socialized" decision, we would like to claim. This conclusion is debatable but intriguing enough from the perspective of language acquisition in social context [22] and perspective of cultural inheritance [3, 7, 16, 23].

Another important implication is that acceptabilities, as something measurable, are not an intrinsic property of stimuli, i.e., sentences. Rather, they are a property "distributed" over a population of speakers. This likely to debunk the methodological basis of a certain brand of linguistics that do not see acceptabilities in this way.

4 Conclusion

Two models for acceptability judgment as a categorical decision were considered in this paper. One is an indirect modeling of it using SELFA. Another was a direct modeling of it using logistic regression. The second modeling turned out to be successful, suggesting that it is possible to simulate human's acceptability judgments, at least in a rudimentary form. Moreover, and theoretically quite interestingly, successful simulation requires population-aware metrics rather than simple threshold-based ones. This is a bald conjecture, but is supported by the analyses presented in this study.

But this conclusion also begs questions: 1) given acceptability judgement is a part of our "collective intelligence," how do individuals internalize it? We admit that this question is open to further research and theorizing.

Acknowledgements

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

We are grateful for Shunji Awazu (Jissen Women's University), Asuka Terai (Future University of Hakodate) and Minoru Yamaizumi (Osaka University), who kindly helped our execution of Phase 1 of ARDJ Survey 2.

[6)] Sentences were sampled because a full plot of 300 predicated values was very likely be unreadable.

References

[1] Jonathan Baron. *Thinking and Deciding*. Cambridge University Press, 2000 [1988, 1994].

[2] H. W. Bierhoff and R. Klein. Expectations, confirmation bias, and suggestibility. In V. A. Gheorghiu, et al., editors, *Suggestion and Suggestibility*, pp. 337–346. New York: Springer, 1989.

[3] Robert Boyd and Peter J. Richerson. *Culture and the Evolutionary Process*. University of Chicago Press, 1988.

[4] Holly P. Branigan and Martin Pickering. An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, 40, 2017.

[5] Wayne Cowart. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Sage Publishing, Inc., 1997.

[6] Evidence-based Medicine Group. Evidence-based medicine: A new approach to teaching the practice of medicine. *Journal of the American Medical Association*, 268(17):2420–5, 1992 (Nov 4).

[7] Joseph Henrich. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton University Press, 2017. [邦訳: 文化がヒトを進化させた: 人類の繁栄と<文化-遺伝子革命>. 白揚社, 2019.].

[8] David Isaacs and Dominic Fitzgerald. Seven alternatives to evidence based medicine. *The British Medical Journal*, 319(7225):1618, 1999. EBM.

[9] Joshua Klayman. Varieties of confirmation bias. *Psychology of Learning and Motivation*, 32:384–418, 1995.

[10] 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.

[11] Kow Kuroda, Hikaru Yokono, Keiga Abe, Tomoyuki Tsuchiya, Yoshihiko Asao, Yuichiro Kobayashi, Toshiyuki Kanamaru, and Takumi Tagawa. Insights from a large scale web survey for Acceptability Rating Data for Japanese (ARDJ) project. In *Proc. of the 25th Annual Meeting for the Association of NLP*, pp. 253–256, 2019.

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

[13] P. Levy. *Collective Intelligence*. Basic Books, 1997.

[14] Scott Menard. *Applied Logistic Regression Analysis*. Sage Publications, 2nd edition, 2001.

[15] Gary Dean Prideaux, Bruce L. Derwing, and William J. Baker. *Experimental Linguistics: Integration of Theories and Applications*. John Benjamins, 1979.

[16] Peter J. Richerson and Robert Boyd. *Not By Genes Alone: How Culture Transformed Human Evolution*. University of Chicago Press, 2008.

[17] Carson T. Schütze. *The Empirical Base of Linguistics: Grammaticallity Judgments and Linguistic Methodlogy*. Language Science Press/University of Chicago Press, 2016 [1996].

[18] Steven Sloman and Philip Fernbach. *The Knowledge Illusion: The Myth of Individual Thought and the Power of Collective Wisdom*. Picador, 2017.

[19] Jon Sprouse. Three open questions in experimental syntax. *Linguistics Vanguard*, 1(1):89–100, 2015.

[20] Masaru Sugiyama, Tsuyoshi Ide, Shin'ichi Nakajima, and Jun Sese. Semi-supervised local Fisher discriminant analysis for dimensionality reduction. *Machine Learning*, 78(35), 2010.

[21] Yuan Tang and Wenxuan Li. LFDA: An `R` package for local fisher discriminant analysis and visualization, 2016.

[22] Michal Tomasello. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Harvard University Press, revised edition, 2005.

[23] 田村 光平. 文化進化の数理. 森北出版, 2020.

A Clustering and PCA of sentences

A.1 Collection of ratings/responses

On acceptability rating, participants were asked to choose one of the four choices in (5).

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

Prefixes 0, 1, 2 and 3 are added to indicate the degrees of deviance, though they need not be on a single scale.

Outlier responders were filtered out using standard deviation ($0.6 < \mathrm{sd} < 1.5$) and Mahalanobis distance (< 0.95). See Kuroda et al [11] for relevant details.

A.2 Standardizing responses

s.id	$r[0,1)$	$r[1,2)$	$r[2,3)$	$r[3,\infty)$	sum
s029	0	31	43	79	153
s099	42	57	36	18	153
s136	0	33	42	91	166
s180	5	25	34	89	153
s231	119	27	10	4	160
s281.4	3	18	55	75	151

Table 4: Frequency table by ranges (6 samples)

s.id	$p[0,1)$	$p[1,2)$	$p[2,3)$	$p[3,\infty)$	sum
s029	0.000	0.203	0.281	0.516	1.00
s099	0.275	0.373	0.235	0.118	1.00
s136	0.000	0.199	0.253	0.548	1.00
s180	0.033	0.163	0.222	0.582	1.00
s231	0.744	0.169	0.063	0.025	1.00
s281.4	0.020	0.119	0.364	0.497	1.00

Table 5: Density table by ranges (6 samples)

Note that gr0, ..., gr9 are different data sets, and cannot be directly compared. Comparison of them requires standardization. All groups were collapsed and responses were counted for each of the four rating ranges $r[0,1)$, $r[1,2)$, $r[2,3)$, and $r[3,\infty)$.[7] Table 4 shows 10 samples of this process. These raw counts were then converted into proportions to item-wise sums. This gave us density array, $P = \langle p[0,1),\ p[1,2),\ p[2,3),\ p[3,\infty)\rangle$, as exemplified in Table 5. The arrays of ranged response probabilities in this format are to be called "response potentials." They are commensurable among groups with different sets of responders, and were used as encodings of the stimuli in the following multivariate analyses.[8]

[7] Note that allowed response values were 0, 1, 2, and 3. These numbers are reinterpreted as ranges $r[0,1)$, $r[1,2)$, $r[2,3)$, and $r[3,\infty)$, respectively, where $r[i,j)$ means a sum of response counts between i and j with i included and j excluded.

[8] Another route to follow is data imputation in which missing values are imputed. We made a few attempts but it turned out that our data have too many unobserved data points to be handled successfully.

A.3 Hierarchical clustering and PCA

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` [12] was used to PCA and Hierarhical Clustering in combination.

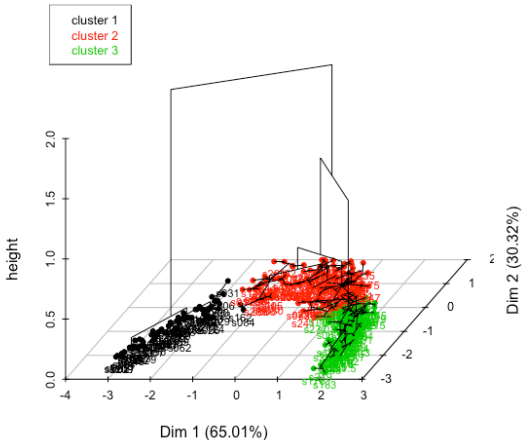

Figure 6: HCxPCA of combined responses for gr0–gr9

Building on standardized responses, a multivariate analysis was conducted where PCA was combined with hierarchical clustering, resulting in visualization in Figure 6. In this, we recognize three major classes of stimuli: clusters 1 (in black, of supposedly "acceptable" stimuli), cluster 2 (in red, of undecidable stimuli) and cluster 3 (in green, of supposedly "unacceptable" stimuli). Clusters 2 and 3 form a larger cluster, contrasting with cluster 1.

A.4 Interpreting PCs

	PC1	PC2	PC3	PC4
Variance	2.60	1.21	0.18	0.01
% of var.	64.94	30.23	4.60	0.22
Cumulative % of var.	64.94	95.17	99.78	100.00

Table 6: Effects of principal components

Table 6 gives the contribution of the three factors identified by PCA. PC 1 roughly corresponds to the polar opposition of $r[0,1)$ and $r[2,3)$. PC 2 is mildly encoded by $r[1,2)$, and weakly by $r[3,\infty)$.

The interpretation of PC1 is straightforward. It encodes the degree of deviance (read from right to left), or of acceptability (read from left to right). In contrast, the interpretation of PC2 is not as simple as PC1. A few likely interpretations come to mind, but the most convincing one would be that PC2 encodes semantic and/or syntactic complexity that often blurs the judgment.