容認度判定の実態調査の報告
その正体は不均一な反応からなる，バイアスのかかった心理評定である *

黒田 航† 仲村 哲明‡ 河原 大輔‡,+

†杏林大学 医学部 ‡京都大学 情報学研究科 +科学技術振興機構 さきがけ

kow.k@ks.kyorin-u.ac.jp, tnakamura@nlp.ist.i.kyoto-u.ac.jp, dk@i.kyoto-u.ac.jp

1. はじめに

言語学は何らかの文が (文法的に) 適格であるかとか容認可能であるかを問題にする．前者は文法性判断 (grammaticality judgments)，後者は容認性判断，あるいは容認度判定 (acceptability judgments) と呼ばれる．本研究は，容認度判定の一般化として容認度評定 (acceptability rating) を考え，その実態を探りつつ，理論言語学が無自覚に想定している反応の均一性の仮定 (homogeneous response hypothesis) を放棄した容認度評定のモデル化を提示する．予備的に実データとのつきあわせの方法も考察する．

2. 容認度評定という心理学課題のモデル化

表現 e の容認度評定とは，評定者 r が刺激としての e に評定値 $a(e)$ を与える作業である．その通常のモデル化では評定者の反応の個体差が捨象されている．だが，一般に異なる評定者 r と r' が同一の表現 e に与えた容認度評定の結果 $a(e,r)$, $a(e,r')$ が同一になる保証はどこにもないし，実際に異なっている．これはクラウドソーシングのような多人数の行動データに基づいたデータ処理では (平均への回帰が強く働かない限り) 問題となり得る．データ駆動型の研究を志向するならば，反応の不均一性を効果として取りこめるモデルが望ましい．

2.1 $a(e,r) \neq a(e,r')$ となる理由

異なる評定者が同一表現に異なる容認度を与えるのはなぜなのか? 原因として考えられるのは次の二つ:

(1) 表現 e を固定しても，評定者 r と評定者 r' とは，

a. e を文脈 $C = (c_1, c_2, \ldots, c_m)$ に対応づける評価関数 $a_{i,j} = a(e_i, c_j)$ (=容認度評定関数) が異なる．

b. 同一の評価関数を使っていても，C の異なる部分集合 D, D' を使っている．

方法論的には (1a) は (1b) よりも強い条件であり，他の条件が同じであれば，(1b) による説明を優先すべきである．従って，次のように仮定することは合理的だろう:

(2) 評定者は一般に，e の容認度評定で C の部分空間 $D \subset C$ しか考慮していない．

この問題の解決には十分に強力な C の指定法/表示法の確立が必要であるが，現状では望みがたい．その代わり，以下では一定の誤差の範囲で (1a) が妥当である場合のモデル化を示す．

2.2 $a_{i,j}$ への r の変異の取り込み

r と r' が C 全体を使って容認度評定をしていると想定した上で考慮すべきことは，評価関数それ自体の変異である．この場合には，評価関数の変異をモデル化すれば十分である．結果が表 1 にある $a_{i,j,k}$ への拡張である．

表 1: $A_{i,j}$ と $R \times E$ の対応 ($R = (r_1, \ldots, r_N)$, $E = (e_1, \ldots, e_n)$)

	e_1	$\cdots$	e_j	$\cdots$	e_n
r_1	$A_{1,1}$	$\cdots$	$A_{1,j}$	$\cdots$	$A_{1,n}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
r_{i-1}	$A_{i-1,1}$	$\cdots$	$A_{i-1,j}$	$\cdots$	$A_{i-1,n}$
r_i	$A_{i,1}$	$\cdots$	$A_{i,j}$	$\cdots$	$A_{i,n}$
r_{i+1}	$A_{i+1,1}$	$\cdots$	$A_{i+1,j}$	$\cdots$	$A_{i+1,n}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
r_N	$A_{N,1}$	$\cdots$	$A_{N,j}$	$\cdots$	$A_{N,n}$

C に対応づけられた e_i の容認度評定の列を $A_i = (a_{i,1}, \ldots, a_{i,j}, \ldots, a_{i,m})$ (m は C の要素の数) とし，評定者 $r_k \in R$ の異なりの次元 R を追加すると，$a_{i,j,k}$ を構成できる．これによって評定者の違いによる表現 E の容認度分布の変異を考えることができる．

容認度は一般的には，n 個の表現列 E, m 個の文脈 C, N 個の評定者列 R を次元とする 3 次元空間中の (確率) 分布 $a_{i,j,k}$ となる．概略図は図 1 である．表 1 が表わしているのは，文脈 C の異なりの効果を A に圧縮した状態である．

*本発表は第一著者による Japan Cognitive Linguistics Association (JCLA), 2016 のワークショップ "「見えない」言語をどう「見る」か: 言語知識へ至る方法論に関する考察と議論" での口頭発表「心理学的により現実的な容認度判定のモデルを求めて」の補足である．

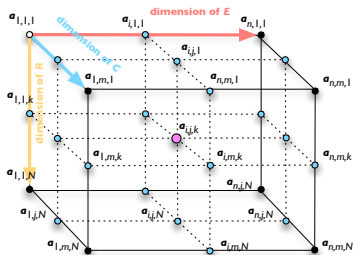

図 1: Tensor of $E \times C \times R$

3. 実データ解析による検証

以上の一般論を具体例に適応した事例分析を一つ紹介する．ただし，この実験は以上のモデル化の妥当性を確認するために設計実行されたものではないので，結果の解釈に制限がある．

3.1 データ

[4] は “Yahoo!クラウドソーシング” を利用して大規模な容認度評定データの収集を実行した．合計 254 名の評定者に逸脱表現を含んだ 1945 種類の文 S (= E) を刺激として提示し，それぞれを容認可能か不可能かの 2 件法で判定させた．図 2 に判定作業の見本を示す．提示刺激文は格フレーム辞書 [3] を基にして自動作成された．

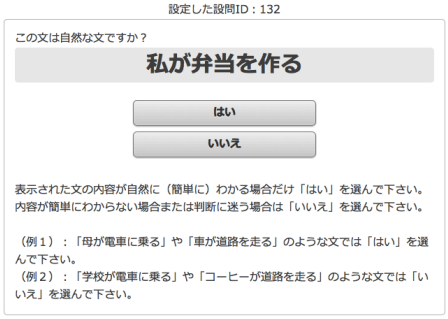

図 2: E の提示の例

S は評価用の D = {d1, d2, …, d10} と実験用の T = {s1, s2, …, s1935} からなる．全評定者が D を評定しているが，T は全員が評定していない．平均して 10 例ほどの刺激を共有しているのみ．

3.2 分析法

設定から明らかな事だが，分析対象のデータは欠損値が非常に多く，通常のクラスタリングは適用できない．対策として欠損値ありクラスタリング (clustering with missing values) を利用した．統計解析統合環境 R [2] 用の `cluster` [1] パッケージで実装されているものを利用した．以下に示すのは，欠損値ありのクラスタリングを，E (図では S と表記する事がある) と R の両方に適用した結果である

3.3 刺激文 E の階層クラスタリング

3.3.1 全事例のクラスタリング

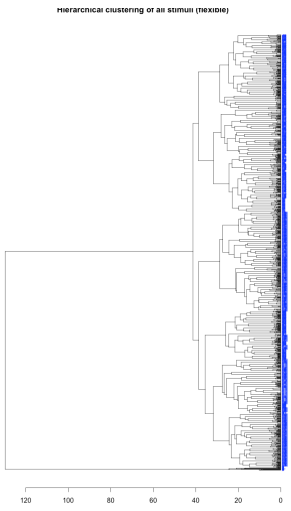

図 3: E の階層クラスタリングの例 (N = 1945)

図 3 に S の全 1945 事例のクラスタリング結果を示した．細部は読み取れないが，D が完全に分離しているのが確認できる (詳細は後のサンプリングで明確になる)．

3.3.2 無作為抽出標本+D のクラスタリング

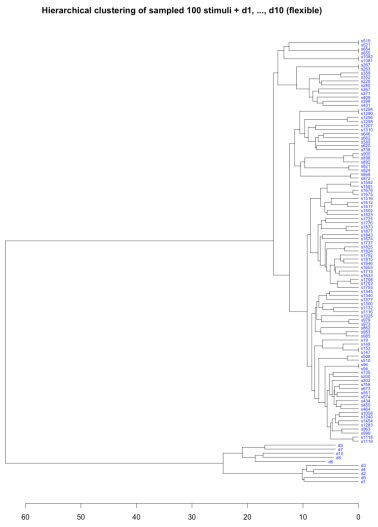

図 4: 標本 100 例+D の階層クラスタリングの例

E(= S) から 100 例を無作為抽出し，それに D を加えた S の部分集合のクラスタリング結果を図 4 に示す．図 3 の形状は良く保存されている．また，D が {d1, …, d5} (完全に容認可能と期待される刺激の集合) と {d6, …, d10} (完全に容認不可能と期待される刺激の集合) とに分離されているのがわかる．

3.3.3 D を除いた無作為抽出標本のクラスタリング

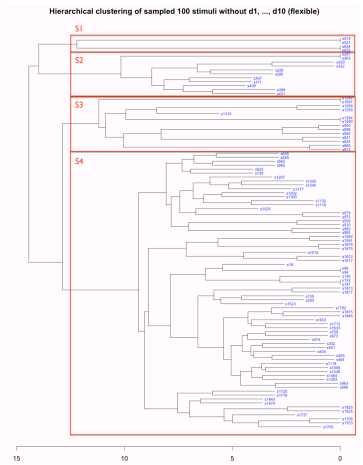

図 5: E の標本 100 例の階層クラスタリングの例

D を除いた無作為抽出した 100 例のみのクラスタリング結果を図 5 に示した．図 4 と比較すると，評価用=違反検知用の D と他の「通常」の刺激に対する反応は (クラスターとして分離しているという意味で) 明らかに質が違う．これは言語学の反応の均一性の想定が D のような理想的な刺激にしか成立しない事を意味する．

図 5 には S1, S2, S3, S4 の四つのクラスターを認める事ができる．それより粗い粒度なら，{S1, S2}, {S3, S4} の二つに大別できる (グラフでは E を S と記号化している)．D を除いた刺激に階層クラスターが生じている理由の説明が必要である．

3.3.4 逸脱の種類とクラスターの対応づけ

S1, ..., S4 は次の仕方で逸脱の種類に対応していると事後分析で判明した:

(3) a. 外心構造 S4 ⇒ S3 ⇒ S2 ⇒ S1 が成立している (中心の S4 が逸脱のない，普通の刺激).

b. (i) S3 では必須要素が欠ける事で逸脱が生じている，(ii) S3 では選択制限が満足されない事で逸脱が生じている，(iii) S3 では不要な要素が現われている事で逸脱が生じている．

(3b) の根拠となる具体例を幾つか示す [1].

(4) a. S1 の例: 1) 選手が監督を有るにやる; 2) プロが試合をやるにやる; 3) 数々が両方を楽しめる; 4) 引きが展開を楽しめる

b. S2 の例: 1) 情報が改札を画面に出る; 2) 記事が出口を上に出る; 3) 人がモーツァルトを聞く; 4) 人が期間を入れる; 5) 人が伸びるを待つ; 6) 行動が域を裏目に出る; 7) 人がシリーズを含む; 8) 人がカードを含む; 9) 人がフォームを含む; 10) 人が効果を含む; 11) 人が円を掛ける

c. S3 の例: 1) 金が頭をところに使える; 2) 金が頭を使える; 3) 道路が下を真っ直ぐに走る; 4) 私がトップを主義に走る; 5) 激震が世界を業界に走る, 6) 風が海岸を走りに走る; 7) ランナーがメートルをトイレに走る

d. S4 の例: 1) 海水浴がサーフィンを楽しめる' 2) 人が有るを学ぶ; 3) コントラストが効果を楽しめる; 4) 人が臨席を賜る; 5) 人が数式を用いる; 6) 人が計算を用いる; 7) 社員が仕事をやる; 8) 人が役をやる

[1] 一部に欠損値の影響で帰属クラスターが適切でない例がある事は指摘しておきたい.

3.4 反応 R の階層クラスタリング

3.4.1 前処理

撹乱要因を除去するため，次のデータ洗浄を行った: 1) 平均容認度が 0.15 以上で 0.86 以下の評定者のみに限定した．2) 評定した事例数が 80 以上 (0.15, 0.86, 80 のような閾値は事後的に選択).

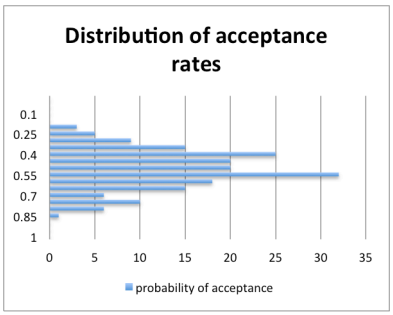

図 6: 選別後の平均容認度の分布 (n = 185)

以上の選別作業後の平均容認度の分布は図 6 に示す通りである．見て取れる通り，分布は正規分布に近い．これは意外な結果であり，次の重要な含意がある:

(5) 平均容認度の分布が正規分布で近似できるとすれば，課題としての容認度判定は実質的にランダムに近い.

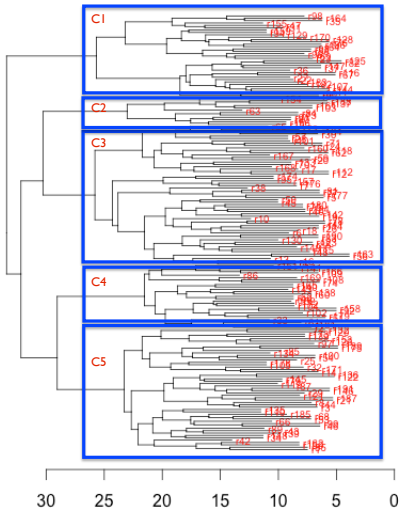

図 7: R の階層クラスタリングの例 (n = 185)

3.4.2 *R* の階層クラスタリングの結果と解釈

図 7 に示すのは，得られた評定値を属性として階層クラスタリングを実行した結果である．

図 7 には適度な粒度で見ると 5 つのクラスター C_1–C_5 が，粒度を粗くすると 3 つのクラスター C_1, $\{C_2, C_3\}$, $\{C_4, C_5\}$ が認められる．これは興味深い結果である．だが，この階層化は何を表わしているのか？

目標は次である: これらの反応クラスターの成立条件を明らかにできれば，容認度評定に生じるバイアスの類型化が実現される [2]．その恩恵として，少なくとも理論的にはデータの事後補正が可能であり，生データをバイアスなしの理想データに近づくように加工できる．

3.5 *R* のクラスターの実質

図 7 の解釈の際に重要な点を一つ明記しておく: 同一クラスターに属している評定者は，評定した文集合の重なり度が高いからそうなっている訳ではない．これは確認済みである．では，何が階層クラスターの要因か？

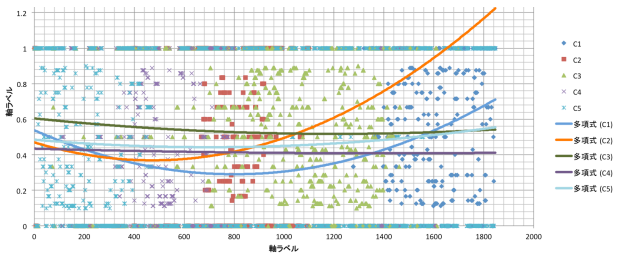

図 8: 2 次多項式 fitting (*x* 軸は ID，*y* は平均容認度)

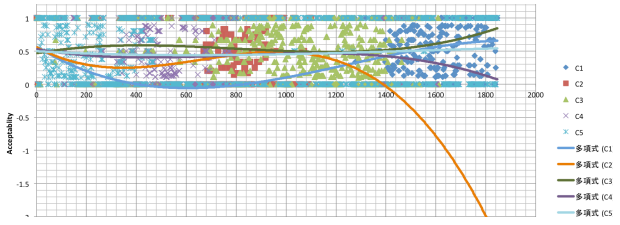

図 9: 3 次多項式 fitting (*x* 軸は ID，*y* は平均容認度)

図 8 と図 9 の解析結果から推測して，(6) か (7) のどちらかが成立していると推測された:

(6) 可能性 1: ID の変化につれて，刺激の質が変わった．

(7) 可能性 2: ID の変化は提示の順序に対応する．

a. “Yahoo! クラウドソーシング” は作業者に ID の順に刺激を提示した．

b. 図 7 のクラスターは作業者の “癖” = 反応タイプに対応する．

[2] ただし欠損値ありの階層クラスタリングでは，クラスター間の構造は高精度に再現されるが，分類対象の正確な位置は不正確になる．

(6) は多項式近似の解釈と整合しにくい．解析後に “Yahoo!クラウドソーシグ” にデータの提示法を確認したところ，(7) の推測の通りだった (従って，提示法の逆行エンジニアリングに成功した)[3]．

2 次近似を元にした対応の解釈は次の通り:

(8)
a. 一定の評定を続けるタイプ: C_3 (緑色): 一定して高目（か妥当な）の評定値を与え続ける; C_4 (紫色): 一定して低目（か妥当な）評定値を与え続ける

b. 一定の評定を続けないタイプ: C_1 (青色), C_2 (橙色): 比較的高目に評価を始めるが，途中で評定が低くなり，最後に高目に戻る．ただし，C_2 は全体として評定

c. 両者の中間 C_5 (空色): 一定に近い評価を与え続けるが，後半に評価が緩く高目に変化する

3 次近似を元にした対応の解釈は次の通り:

(9) C_4 (紫色) と C_5 (空色) は一貫した評定を続けるが，C_4 は最後に評定が低くなる．C_1 (青色) と C_2 (橙色) と C_3 (緑色) は一貫した評定を続けない．C_1 は中庸な評価から始めるが途中で何も容認しなくなり，また元に戻る．C_2 は中庸な評価から始めるが途中で容認度が低くなり，一旦元に戻るが，最後にすべてを容認しなくなる．C_3 は中庸な評価から始めるが，途中で甘くなり，また元に戻る．

4. 結論

本研究の結果から言えるのは，容認性判断と容認度評定は別物であり，容認度評定は作業者の内部モデルによってバイアスされているという事である．これは言語学が無自覚に想定している均一反応の仮定を反証する．

その一方，これらのバイアスは十分なデータがあればうまく分類でき，かつ事後補正も可能である．予想外の成果として，クラウドソーシング利用の落とし穴も発見できた．

参考文献

[1] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2015. R package version: 2.0.4.

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

[3] 河原 大輔, 黒橋 禎夫. 高性能計算環境を用いた Web からの大規模格フレーム構築. 情報処理学会研究報告, Vol. 2006-NL-171, pp. 67–73, 2006.

[4] 仲村 哲明, 河原 大輔. 集合知を用いた事態参与者の特徴変化に関する知識の獲得. 言語処理学会第 22 回年次大会発表論文集, pp. 901–904, 2016.

[3] これが NLP 業界にとって意味する事の一つは，クラウドソーシングを利用する時，無作為化の必要がある時，依頼者側でそれをしてデータを渡す必要があるという事である．