(言語学者による) 容認度評定の認証システムを試作する構想
入念に設計された日本語文の容認度評定データベースに基づいて

黒田 航[1] 阿部 慶賀[2] 横野 光[3] 田川 拓海[4] 小林 雄一郎[5] 金丸 敏幸[6] 土屋 智行[7] 浅尾 仁彦[8]
[1] 杏林大学 [2] 岐阜聖徳学院大学 [3] 富士通研究所 [4] つくば大学 [5] 東洋大学 [6] 京都大学 [7] 九州大学 [8] NICT

概 要

本発表は研究発表とは違う．その目的は，第一著者を代表者として科学研究費の助成を受けた研究「言語研究者の容認度評定力の認証システムの試作：容認度評定データベースを基礎にして」の周知と，研究プロジェクトへの協力の呼びかけである．特に評定課題の刺激となる日本語文集合の選定で，プロジェクト外のいる研究者 —言語学者や心理学者— からの希望を受け付けたいと思っている．

1 はじめに

1.1 研究プロジェクトの目的

本研究の第一の目的は，項目反応理論に基づいた日本語文の十分な規模の容認度評定データベースの構築である．これは理論系 vs 実験系の別や基礎 vs 応用の区別を問わずに研究資源として有用であるため，構築が待望されているデータと思われるのだが，国内外問わず，今だに構築がない．このデータを，統制した日本語文集合への一般人の反応をサンプリングして構築する．

第二の目的は，当該データベースを土台にした，研究者向け容認度評定能力の認証システムの試作である．これが何のために必要，かつ有効かと言うと，想定システムの無償利用が可能になれば，個々の言語研究者が自分の容認度判断の信頼度を有能性のシグナルとして発信する事が可能となるからである．これは長期的に言語研究で使用される「証拠」の質を上げ，Evidence-based Linguistics (EBL) [10] の基礎になる事が期待できる．

1.2 研究の背景と位置づけ

言語研究者はしばしば，文の容認度を評定し，それに基づいて議論する．それは，自然な文 (=容認可能な文) と不自然な文 (容認不可能，あるいは困難な文) の区別を確定し，その差が生じる理由を明らかにする事が，言語学の目標の一つだからである．この判断を容認性判断，ないし容認度評定と呼ぶ．

容認性判断と容認度評定の違いは，前者が Yes/No の二値判断=カテゴリー判断なのに対し，後者は中間的段階を認める所にある（更に言うと，容認性判断と別に文法性判断もあるが，両者の区別はここでは取り上げない）．以下では，より一般的な概念である「容認度評定」の方で総称する．

例えば言語研究者は，次のような対比を用いて議論を進める:

(1) a. その獣は獲物を襲った．
b. その獣は獲物に襲いかかった．

(2) a. ?*その発作は患者を襲った．
b. *その発作は患者に襲いかかった．

*の付いた事例は容認が困難な事を，烙印のない事例は容認度に異常がない事を．?*が付いた事例は中間事例である事を，それぞれ表わす．

容認度評定を証拠に使う事は言語研究で一般的な方法であるが，明らかな難点が二つある．第一に，外部参照値がないため，個々の判断にどれ程の信頼が置けるのかが不明である．この種の判断が確証バイアス [1, 2, 3, 4, 6, 8] に左右される事が特に問題である．

第二の問題は，個々の評定者による評定値と「真」の容認度との関係のモデル化が単純過ぎる事である．これは私見では，第一の問題よりずっと深刻である．観測精度の問題を真剣に考慮するなら，任意の文について，容認度評定は事例ごとに独自の統計分布を示し，幾つかの分布クラスが存在すると考えるべき理由がある．だが，現在の言語学では，容認度評定の結果測定自体が行われていないので，その必要性すら自覚されてない．

容認度評定は長らく理論言語学の（ほぼ唯一の）証拠として使われて来たし，今だに完全な代替がない状態である．この理由で言語学の研究成果は玉石混交であり，分野外から見た信用は下がり続けているのだと思う．

だが，事態が改善される兆しは微かである．例外的な研究として [11, 12, 13] がある位である．

1.3 到達目標

この状況は明らかに泥沼であり，それから脱するには少なくとも次の A が必要であり，B が実現できれば更に望ましい:

A. 第一に，統制された刺激文で構成される文集合に対する非専門家の容認度評定値の分布をデータベース化し，容認度評定という課題の反応分布を明らかにし，結果を研究者の共有資源とする．

B. 第二に，上の容認度評定値データベースを (部分的に) 参照し，個々の研究者の判断の信頼度を保証する認証制度を実現する．

容認度評定データと認証システムが利用可能となり，言語研究者が判断の信頼性を自己開示する事が可能になれば，言語学の研究成果の信頼度が長期的に引き上げられる．それが期待できるのは，学会の権威への挑戦が簡単になるからである．

2 研究方法と計画

2.1 方法

容認度評定の認証は次の 3 つの手順 (正確には 4 手順) で実現できる．

(3) 認証システムの構築手順

Step 1. 刺激文集合 E の設計と構築

Step 2. E に対する一般被験者の反応のデータベース化 (結果を D)

Step 3i. D に基づく認証クラスの特定

Step 3ii. 容認度評定力の認証システムの開発

2.2 計画

今後の実行計画は次の通りである:

(4) 計画

H28 年: Step 1 を実施し，可能であれば Step 2 の予備実験も実施

H29 年: Step 2 (と Step 3i) を実施

H30 年: (Step 3i と) Step 3ii の実施

具体的手順を以下で順に説明する．

2.3 Step 1.

なるべく大きな規模の，統制の取れた文の集合 E を用意する．ただし，

(5) a. E の事例数は少なくとも数百の規模．可能ならば千台に乗せたい．

b. E は逸脱のない (=容認度の高い) 文だけでなく，逸脱のある (=容認度の低い) 文を十分に含んでいる必要がある．

c. E はサンプルとは言え，それなりのカバーを持ち，選定基準が学派の利害から独立している必要がある．

E の構築で制御すべき要因の一つが逸脱の度合いである．これは不自然な文を作る作業を前提としており，これは言語学を学んだ研究者の協力を得て実現する．

もう一つの条件は，E の代表性の確保のために日本語の基本構文をなるべく多くカバーする事である．この際，基本構文に強く結びついた動詞を選定する事が必要になる．これは大規模言語データを日常的に扱っている研究者の協力を得て実現する．

ここで宣伝

刺激文集合 E の選定でプロジェクトに参加していない方々 (特に言語学者や心理学者) を対象に希望調査を行う予定です．Web 調査を考えていますので，関与を希望する方は第一著者に連絡を下さい．

2.4 Step 2.

言語学を専門としない一般被験者 R に一定の指示 I を与え，E のそれぞれの文 e_i の容認度を評定して貰う．e_i に平均 m_i と分散 v_i が割り当てられる．この結果 D は，(e_i, m_i, v_i) という三組である．ただし，

(6) a. R は (少なくとも) 数百人の規模であるべき．

b. I の例は ⟨ 4. 違和感なし — 3. 言っている事はわかるが，軽く違和感を感じる — 2. それなりに違和感を感じるが，言いたい事は理解できる — 1. 強く違和感を感じるが理解できない程ではない — 0. 意味不明 ⟩ の 5 件法など．

この段階で重要なのは，幅広い範囲の評定者と最適反応を導く I の設計である．これは心理実験と効果測定の専門家の協力を得て実現する．

2.5 Step 3

2.5.1 Step 3i.

評定者 P の容認度判定の信頼度を，次の指標で評価する (ただし，本手順の b が Step 2 の一部として実行可能な場合には，そうする)．

(7) a. D から適当な部分集合を選定し，これを T とする (T の要素数は数十)．なお，T は D の公開部分から除外する．

b. P に R に与えたのと同じ指示 I を与え，T のそれぞれの要素 t_i について，容認度を評価して貰い，かつ反応の分散の大きさを予想して貰う．

c. $x_i = [(t_i$ の平均値$) - (P$ による t_i の評定値$)]$ と $y_i = [(t_i$ の分散のクラス$) - (P$ の予想した t_i の分散のクラス$)]$ を評価する．

段階 b で使う分散評価の基準は，{ L: 大きい (=評定値が人によって割れる); S. 小さい (=評定値が人によって割れない); M. 両者の中間ぐらい } の 3 値分類を想定する．

段階 c の評定者 P の容認度評定の信頼度の評価法は次の通り: D のうちの k 個が T として選ばれて，P に刺激として与えられるとして，それぞれの刺激 t_i について，

$$X = \Sigma_{i=1}^{k} x_i^2, \quad Y = \Sigma_{i=1}^{k} y_i^2$$

を計算する．この時，P による評価の R による評価に対するズレの指標 $Z^2 = (X + Y)/k$ の値が小さいほど，P による容認度評定はそれだけ (R に対し高い代表性をもっていると判断できるので) 信頼できると考える．

なお Y の個々の要素については，分散クラスの予想値と真値とが

(8) a. (L, L), (M, M), (S, S) のように一致した場合は 0,

b. (L, M), (M, S) のように 1 度の食い違いを生じた場合は δ,

c. (L, S) という 2 度の食い違いを生じた場合は 2δ

の罰則が発生すると想定する (L, M, S の区分の基準や δ の値は事後的に求める).

2.5.2 Step 3ii.

ズレの分布 Z を予備調査で確定すれば，その後は任意の人が行った容認度判定の信頼度を，例えば事後的に定める α, β を用いる事で，次の 5 クラスのどれかに分類する事が可能である:

(9) a. S 級 ($+2\alpha$ 以上),

b. A 級 ($+1\alpha$ と $+2\alpha$ の間),

c. B 級 (-1β と $+1\beta$ の間),

d. C 級 (-2β と -1β の間),

e. D 級 (-2β 以下)

正規分布なら，$\alpha = \beta = \sigma$

とは言え，Z の分布は不明であり，言語研究者に対象とする予備調査の結果に基づいて，分布の性質を確認する必要がある．

以上の手順で容認度評定の認証システムのエンジンが得られる．本格的な運営は将来の課題とする．

3 議論

3.1 本研究の挑戦性

自然言語文の十分な規模の容認度評定データベースは，構築が待望されている言語研究の重要な基礎データである．その理由は，この種のデータが理論系 vs 実験系の別や基礎 vs 応用の区別を問わずに言語研究のリファレンスとなるからである．しかし，そのようなデータベースは国内外問わず，今だに存在していない．それを世界に先駆けて実現しようとする本研究には少なくとも次の二つの斬新性がある．

第一の斬新性は，容認度評定データベースの構築のための方法論を提示する所にある．具体的に言うと，容認度評定を，刺激に対する個々人の反応パターンの統計分布と見なし，項目反応理論 (Item Response Theory) [14, 15, 16] を応用して記述する．その背景として，容認度分布に質的に幾つかの異なるクラスを見出す事を想定している．これは従来の言語観には皆無だった視点であり，この想定の正しさが確証されれば，それだけでインパクトの強い成果が得られる．

第二の斬新性は，当該のデータベースを，専門の研究者が行う容認度評定の信頼性の認証評価システムの構築という応用を見込んだ形で構築する所にある．認証評価の対象となった人の容認度評定の信頼性は，評定の代表性の保証という形で与えられる．このような評定の質的保証がないために，言語学者による専門研究の成果は玉石混交である．言語研究者が本認証システムを使って自己情報の開示 (シグナリング) を始めれば，その効果の蓄積の結果として，玉を石から見分ける事が将来は可能となるはずである．

判断の代表性の保証は，判断の正確さに自分が代表する集団の反応の分散の予想を組み込む事で高精度化される．共有されている刺激（この場合，自然言語文）に対する反応の平均値を予測する課題 (John Maynard-Keynes の言う「美人コンテスト」) は典型的であるが，当刺激に対する反応の分散を予想させ，その精度を評価に組み入れるのは類を見ないモデル化であり，一定の効果が期待できる．

なお，システムの本格的な運用は試作成功後の課題とする．

3.2 補足的注意

本研究の第一段階の成果は，二点目の斬新性が真でなくても，それだけで容認度評定の基礎理論の刷新に貢献する．ヒトの認知活動の複雑さを考えると，同一の刺激に対し一様な反応が得られ，それが真の反応であるという，従来の言語学の理想化は素朴過ぎる．だが，そのモデルの改訂は今ま

で誰も試みていない．構築された容認度評定データをサンプルとする事で，ヒトの容認度評定の緻密なモデル化が将来的に可能となると期待できる．この点については§4.1 で再説する．

容認度データベース構築では日本語が対象となる，それには三つの理由がある．i) 第一に，異国語 (例えば英語) の言語資源の構築は非母語者に手に負える程に簡単なものではない．ii) 第二に，仮に構築が可能だとしても，日本国内での有用性が限定される．iii) 第三に，構築されたものが英語のデータでないとしても，このようなデータベースの構築が可能である事実を立証するだけでインパクトは十分に大きく，方法論自体は言語に依存しない透明性の高いものなので，他言語の研究でも本研究の成果を追従する事が確実に起こると期待できる．

一方，日本語データの構築が済んでしまえば異言語への移植は難しい事ではない．特に期待できるのは，日本語教育への展開性である．

4 最後に

4.1 射程の限界

最後に幾つか，現時点で取り上げる予定にない事を明示しておきたい．

本研究の射程は二つの意味で限られている．第一に，本研究は萌芽的ものであり，一気に完成度の高い成果を上げる事は目指していない．方法論の確立が最優先事項であり，応用的有用性の追求はそれに準じる．

第二に現実的な面として，したいと思っている事のすべてをやり遂げるのに (研究者の頭数や力量を含めて) 十分な研究資源が備わっている訳ではない．端的に言えば，期間内，予算内で達成できる事が限られている．

以上の制限の下で，将来的に達成すべきであるが射程外にあるのは次である:

(10) a. 容認度評定に文脈が与える影響の測定

b. 得られた結果の社会言語学的考察

まず (10a) について言うと，ベースライン反応がわかっていないと文脈の影響は調べようがない．本研究の目標は，将来的に文脈効果の研究を可能とするためのベースライン反応の記述である．

一般的に容認度を考えるためには，表現 E に内在する容認度，文脈効果 C，評定者効果 R の三組で構成される容認度空間 $E \times C \times R$ を考える必要がある．このモデル化で重要な事は，C と R の分離である (R は評定者の個性 $\approx$ 癖と見なして良い)．R を C から分離するモデル化は，ほとんどの言語の意味構築の理論で明示的に行われていない事であるが，容認度評定が心理反応の一種である事を考えれば，モデルの精緻化のために明らかに導入すべき仮定の一つである．

E，C，R のそれぞれは，理想的には他の二つの要因を固定した状態で測定すべき抽象的対象である．今回の調査に関して言うと，E と C を固定して R を記述する事が主な目的である．これが確立したら，その後に，E と R を固定して，C を記述する事が可能になる．期待を込めて言うと，これは実験意味論/語用論 (experimental semantics/pragmatics) [7, 5] の研究精度を向上させるだろう．

ここで，R は際限なく変異するもの，つまり評定者の数だけ異なりがあるものでなく，幾つかの基本戦略 (strategies) に分類できるものと想定している．これは未検証の想定であるが，予備的な調査 [9] から示唆されている可能である．本研究の目標の一つはこの想定の妥当性の検証であり，容認度評定の認証システムを実装は，その検証の肯定的結果の応用となる．

4.2 社会調査との統合の必要性

上述の R を構成する要因が，単に認知的なものであるだけでなく，社会言語学なものである事が想像できる．例えば，容認度評定の戦略が，学齢や学習履歴や学習態度と相関を持つ事は容易に想像しうる．だが，容認度評定の戦略が具体的にど

いう形で評定者の社会的属性と結びついているかは，しっかりした社会調査を実施しなければ確定しようがない．これは将来的に実現すべき目標の一つであるが，調べるべき社会的属性の選定を含め，現時点では準備が整っていない．この方向性は本研究が一定の成果を収めた後に着手したい．

参考文献

[1] Jonathan Baron. An analysis of confirmation bias. In *Paper presented at the 22nd Annual Meeting of the Psychonomic Society, 6–8 November, 1981, Philadelphia*, 1981.

[2] H. W. Bierhoff and R. Klein. Expectations, confirmation bias, and suggesttibility. In V. A. Gheorghiu, P. Netter, H. J. Eysenck, and R. Rosenthal, editors, *Suggestion and Suggestibility*, pp. 337–346. New York: Springer, 1989.

[3] B. Evans. *Bias in Human Reasoning: Causes and Consequences*. Psychology Press, 1990. [refers to the confirmation bias?].

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

[5] Jörg Meibauer and Markus Steinbach, editors. *Experimental Pragmatics/Semantics*. Linguistics Today 175. John Benjamins, Amsterdam/Philadelphia, 2011.

[6] Ian I. Mitroff. Scientists and confirmation bias. In Ryan D. Tweeney, MIcheal E. Doherty, and Clifford R. Mynatt, editors, *On Scientific Thinking*, pp. 170–175. Columbia University Press, 1981.

[7] Ira A. Noveck and Dan Sperber, editors. *Experimental Pragmatics*. Palgrave Macmillan, 2005.

[8] T. ギロビッチ. 人間この信じやすきもの: 迷信・誤信はどうして生まれるか. 新曜社, 1993. [原典: Thomas Gilovich (1993). *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*, Free Press].

[9] 黒田 航. 言語表現の容認度とは何か? また何であるべきか?— 言語学者であるはずなのに，容認度判断が何であるかに自信をもって答えられない (大半の) 人々への手引き. 2011. URL: http://clsl.hi.h.kyoto-u.ac.jp/~kkuroda/papers/on-acceptability.pdf

[10] 黒田 航. 証拠に基づく医療 (ebm) との比較を通じて理論言語学の方法論を見直す. In 第 *16* 回日本認知言語学会発表予稿集, 2015. URL: http://clsl.hi.h.kyoto-u.ac.jp/~kkuroda/papers/kuroda-jcla2015proc.pdf.

[11] 斎藤 幹樹. 下位構文スキーマが容認性判断に与える影響の統計的評価. In 日本認知科学会第 *32* 回大会発表論文集, pp. 1–8, 2015.

[12] 斎藤 幹樹. 容認性判断に関わる認知的要因: 認知文法的観点からの分析. 未刊行修士論文, 京都大学, 2015.

[13] 藤田 元. 多重主格構文の容認性判断について: 容認性の一般理論に向けて. In 日本認知科学会第 *32* 回大会発表論文集, pp. 598–603, 2015.

[14] 豊田 秀樹, 他. 項目反応理論 *[*入門編*]*. 朝倉書房, 第 2 版 edition, 2012.

[15] 豊田 秀樹, 他. 項目反応理論 *[*中級編*]*. 朝倉書房, 第 2 版 edition, 2013.

[16] 野口 裕之, 他. 組織・心理テスティングの科学. 白桃書房, 2015.