ARDJ の刺激文の読み時間の回帰木解析を使った要因分析
Regression Tree analysis of reaction-time data for ARDJ survey

黒田 航[1], 阿部 慶賀[2], 粟津 俊二[3], 寺井 あすか[4], 土屋 智行[5]
Kow Kuroda, Keiga Abe, Shunji Awazu, Asuka Terai, Tomoyuki Tsuchiya

[1]杏林大学, [2]和光大学, [3]実践女子大学, [4]公立はこだて未来大学, [5]九州大学
Kyorin Univ., Wako Univ., Jissen Woman's Univ., Future Univ. Hakodate, Kyushu Univ.
kow.k@ks.kyorin-u.ac.jp,k.abe@wako.ac.jp,awazu-shunji@jissen.ac.jp,aterai@fun.ac.jp,tsuchiya@flc.kyushu-u.ac.jp

概要

日本語容認度評定データ (ARDJ) 構築の第一期と第二期の調査では，容認度評定を得るために 466 種類の刺激文を使用した．その後，同一の文を刺激として使い，読み時間データを追加収集し，評定値データと対応づけた．黒田ら [14] はそのデータの回帰分析から．文の読み時間とその文の容認性判断との間に明確な相関が見出せない事を報告した．その後，決定木解析の一種 RPART を使ってこの結果を再評価した．本研究はその結果を報告する．読み時間と容認度判断との相関は，限定的かつ複雑である事が示唆された．

1 はじめに

1.1 ARDJ とは何か？

日本語容認度評定データ (Acceptability Rating Data for Japanese: ARDJ) [12] は，日本語の文の容認度評定値のデータベースであり，無償利用可能で，大規模かつ確証バイアスの影響が少ない．ARDJ はこれまでに第 1 期 [6], 第 2 期 [7] の成果を公開した[1)]．

ARDJ の大規模性には二つの意味がある．第一に，評定された刺激文の数が多い (第 1 期と第 2 期を合わせて異なり数が 466 であるが，一般に行なわれている言語学の研究で検討される刺激文の数はこれより一桁小さい)．第二に，評定者の数が多い事 (第二期では一文について 70 名以上の評定者から評定を得ているが，一般に行なわれている言語学の研究の評定者の数はこれより一桁小さい)．

ARDJ の脱確証バイアス性には二つの意味がある．第一に，刺激文の構築が特定の言語理論を確証，ないしは反証するために行なわれたものではない．刺激文の集合は理論的には可能性空間 (の部分空間) の無作為抽出の結果である (詳細は §2.2 を参照)．第二に，評定者を状況が許す限り無作為化している．評定者は性別，年齢，居住地の値の選択で明らかな片寄りが生じないように選ばれている．

更に ARDJ は部分的に社会調査として実施されている．それは調査の際に，単に日本語の文の容認度を評定して貰っているだけでなく，評定者の社会的属性 (年齢，性別，異国語学習歴，異国での生活歴，暮らした地域，教育歴，読書量，文系理系の程度など) を一緒に収集している．これにより，容認度に影響を与える要因の層別解析が可能になる．結果の一部を黒田ら [13] が報告している．

以上のように，ARDJ は刺激文を入念に設計し，評定者を入念に準備している．これは，ARDJ の目的が言語研究の方法論を，**証拠に基づく医療** (Evidence-based Medicine: EBM) [2, 4] を模範として再構築する事を意図しているからである．EBM は証拠の有効性を階層化し，その最上位に randomized controlled/contrastive trial (RCT) を置く．ARDJ は同じ地点を目指す．最終目的地にあるのは，確証バイアスや権威主義から可能な限り縁遠い**証拠に基づく言語学** (Evidence-based Linguistics: EBL) [10] の実現である．EBL の具体的動機は §4 で論じる．

1.2 本研究の目的

ARDJ は 466 種類の文に対する容認度評定値データを公開している．データの利用可能性の向上させると期待し，同一の刺激文の読み時間データを取得した．このデータの対応関係を，黒田ら [14] の続編として解析する事が本研究の目的である．

ただ本論文が主としているのは決定木解析結果の詳細な提示である．その結果として，結果の考察は十分ではない事は認めざるを得ない．そうしている理由は，次の通りであり，不本意なものである．1) 反応を取得した実験参加者に十分な代表性が期待できず，予備調査の域を出ていない; 2) そのための十分な紙面がない．

[1)] ARDJ データ公開用サイト https://kow-k.github.io/Acceptability-Rating-Data-of-Japanese/

表 1: RT 実験で使った刺激文 (gr0) の見本 [S (segmented) の部分を使用]

S.ID	RT.S.ID	V.ID	Pattern	Type	S (segmented)	#seg
s1-016	474	44	p3	o	高校生が/デートの場で/しらじらしさを/恋人に/感じた。	5
s1-062	270	338	p1	s	ころっと/相手に/大事な試合で/有望選手が/負けた。	5
s1-114	240	131	p3	o	刑事が/捜査で/手がかりを/手当たり次第に/探した。	5
s1-117	66	326	p1	n	客が/そのスーパーで/店員に/文句を/言うと/黙った。	6
s1-122	354	1197	p1	p	私が/遊園地で/インフルエンザに/家族に/感染した。	5
s1-144	114	326	p4	s	不安から/妊娠を/次女が/実家で/黙った。	5
s1-186	342	22	p3	v	船が/遠回りで/海路を/安全に/来た。	5

なお，本研究は探索型の研究であり，特定の仮説の確証を目的にしていない．そのため，本研究には結論と言えるものが伴っていない．それが欠点かどうかは，認知科学の研究に何を期待するかによって決まる．

2 データ収集

2.1 課題

調査の目的は RT の取得であるが，それを self-paced reading のパラダイムで実装した．実験参加者は，刺激文が事前に指定されている区切り (表 1 の S (segmented) で "/" で示されている単位) ごとに段階的に提示される．適当なキー (例えばスペースバー) を押すと，次の区切りが提示される．このように実験参加者は自分のペースで読み進め，全体の提示終了後に，文を (1) のいずれかに判定するように求められた:[2)]

(1) 1: 違和感がなく自然に理解できる文
2: 不自然で理解不能な文

この課題の結果，(rt1, rt2, …, rt5, RT, response) という数値が得られる．rt$i$ は i 番目の区画と $i+1$ 番目の区画の反応時間の差であり，次の区画に移動するまでの所要時間を意味する．RT は最後の区画を見た後に (1) の容認度判断を下すまでの時間で，response の値 (1 か 2) が判断の結果である[3)]．

2.2 刺激

ARDJ はこれまでに調査 1 (survey 1) と調査 2 (survey 2) を実施し，2 回の調査で延べ 466 種類の刺激文の評定値を得ている．その内訳は次の通り: 調査 1 では 200 種類の文を刺激に使った．調査 2 では 300 種類の文を刺激に使ったが，調査 1 から 12 種類を再利用した．そのため二つの調査で使った刺激文は 466 (=188 + 280 − 2) 種類[4)]．

表 1 に RT 取得に使われた刺激の見本を示す．刺激文の分節数は 5 か 6 である．大半を 5 の場合が占めるが少数ながら 6 個の刺激がある[5)]．表 1 の変数の簡単な説明は次の通り:[6)]

(2) a. S.ID は刺激文の ID (s1- は第 1 期のみで使われた刺激，s2- は第 2 期で使われた刺激を意味する).
b. RT.S.ID は S.ID とは別に今回の実験でデータの無作為化のために利用した ID.
c. V.ID は刺激文を作成する際に種文に使われた動詞の ID (NINJAL-LWP for BCCWJ[7)] の動詞の頻度順位に弱く対応).
d. Pattern は動詞の値に拠らずに事前に決めた 5 種類の文の雛型
e. Type は §2.3 で説明する刺激文の編集の型.
f. S (segmented) は刺激文 (分割箇所は "/" で補助的に示している).
g. #seg は分割数

2.3 刺激文の作成手順

刺激文は 65 種類の原文 (originals) に (「変異」と呼ぶ) 無作為な編集を適用して自動生成された．刺激文の作成で調査者の意図を反映させていないのは意図的な決定に拠るもので，RCT の設計理念に基づいている．

原文の作成は次の手順で行った: Step 1) 5 つの文パターン P_1–P_5 の選定: P_1: __ が __ で __ に __ と V した; P_2: __ が __ で __ に __ を V した; P_3: __ が __ で __ を __ に V した; P_4: __ が __ で __ から __ を V した; P_5: __ が __ で __ と __ を V した (ただし __ は名詞が実現する変項，V は動詞が実現する変項)．この際，i) 曖昧性を除外するため [__ は] は用いない，ii) [__ が __ で…V した] は

2) 調査 1, 調査 2 の容認度評定は [0: 違和感がなく自然に理解できる文; 1: 違和感を感じるが自然に理解できる文; 2: 違和感を感じ理解が困難な文; 3: 不自然で理解不能な文] の 4 件法だったが，カテゴリー判断を模するように，両端の二値を使った．

3) それぞれの実験で使われた数値は異なるが，ここでは {1,2} に統一した．また，回帰解析では 1 → 0，2 → 1 の数値変換を施し，逸脱の程度を [0,1] の値に正規化してある．

4) s2u データで得られる異なり数が 468 でなく 466 なのは，s1-010=s2-010=s2-281 と s1-127=s2-127=s2-282 が別の文として扱われているため．

5) これは意図した事ではなくて刺激作成の際の不注意による．

6) 詳細は [6, 7] を参照されたい．

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

固定すると決めた．[…V した] の他に，[…V する]，[…V している]，[…V していた] を候補に加え，時制と時相の影響は見たかったが，事例当たりに十分な反応数が確保できないのは明らかだったので，断念した[8]．

Step 2) 動詞 V の選定 (NINJAL の LWP for BCCWJ[9]の動詞検索の頻度情報を参考にして，高頻度域と中頻度域から候補をほぼランダムに採取し，動詞を選定．ARDJ の第 1 期 (Survey 1) [6] の原文作成では [行く, 知る, 教える, 感じる, 答える, 探す, 黙る, 負ける, 伝わる, 知り合う, 感染する] の 11 動詞を選定，第 2 期 (Survey 2) [7] の原文作成では (第 1 期から転用された事例を除くと) [襲う, 入れる, 伝える, 聞こえる, 繰り返す, 遊ぶ, 助ける, つなぐ, 載る, 間違える, 直す, 届ける, 習う] の 13 動詞を選んだ．第 2 期の動詞選定は単純な頻度基盤ではなく，事前に Formal Concept Analysis (FCA) を使った動詞分類で異なりが多くなるように候補を絞り込んだ (詳細は [11] を参照)．

Step 3) $P_1, P_2, \ldots, P_5$ の V (P_5 は第 2 期で候補に追加) を step 2 で選定した動詞で埋め，そのように定義された文の雛型 (e.g., “__ が __ で __ を __ に探した”, “__ が __ で __ に __ を入れた”, “__ が __ で __ に __ を届いた”) のうち，明白な逸脱を示すもの (e.g., “__ が __ で __ に __ を届いた”) を除き，残った雛型のすべての空所 (i.e., __) を人間が語彙的に実現する[10]．

Step 4) このように人手で生成された候補から適当な数の事例をほぼランダムに選ぶ．こうして第 1 期では 33 個の，第 2 期では 33 個の原文が選ばれた．

変異は i) 動詞の置換 (mutated verb)，ii) 名詞の置換 (mutated nominal)[11]，iii) 格助詞の置換 (mutated positional)，iv) 句の入れ替え (phrase swapping) である (詳細は [6, 7] を参照)．変異の分布は表 2 の通り[12]．

表 2: 変異の割合

code	type of mutation	count	ratio
o	original [no mutation]	65	0.139
v	mutated verb	90	0.193
n	mutated nominal	108	0.232
p	mutated postpositionl	95	0.204
s	swaped phrases	108	0.232
	sum	466	1.00

このような無作為生成には 2 つの意味がある．まず，確証バイアスの抑制が期待できる．これは，評定者の属性が可能な限り無作為化されている事と合わせて，確証バイアスの抑制の度合いを高める．次に，無作為化には逸脱例と非逸脱例の違いを滑らかにし，2 クラスの境界 (超) 平面の認識を容易にする効果が期待できる．

2.4 反応

合わせた 466 種類の刺激文をランダムに 6 つのグループ (gr0, gr1, …, gr5) に分割した (1 グループに含まれるのは約 80 文)．実験の際には，特定の刺激文のグループに反応の取得が片寄らないように，ランダム化をした．

反応の取得は，函館[13]，東京[14]，岐阜[15]の 3 ヶ所で大

---

[8] 理論的言語学をやっている人には滅多に理解されない事だが，観察の範囲と信頼度とはトレードオフの関係にある．信頼できないデータを広い範囲で集めても，得られる結果は単なるガラクタである．従来の言語学はガラクタの上に成立している疑似科学の域を出なていという懸念が ARDJ の出発点にある．この事は指摘しておく必要があるだろう．

[9] https://nlb.ninjal.ac.jp/search/

[10] この作業は本論文の第 1, 第 2, 第 5 著者の他にもう 1 名 (浅尾仁彦 (NICT)) を加えた計 4 名が行った．参考までに述べておくと，この作業は相当に大変であった．P_1–P_5 を元にした雛型は 4 つ名詞変項を持つ．そのうち 3 つまでの語彙的実現は困難ではないが，最後の 4 つ目の実現は言語学を専攻した者にとっても簡単ではない．

[11] nominal は形容動詞を含む．

[12] 査読者の一人から「表 1・表 2 を見たところ，刺激文の分類が言語学の観点から十分でない．表 1 を見ると，短文構造・複文構造・二つの文が順に出てきて主語が共有されている例が混在している．さらに，動詞のアスペクトがそもそも文法的に正しくないものも含まれている．…一方で表 2 の分類コードを見ると，語順や格助詞など，非常に表面的な特徴でしか文を捉えておらず，文構造の根本的な部分が捉えられていない．そのように雑多で言語学の観点からは未分類のデータの読み時間と容認度評定の相関を見ていくことにどのような意義があるのか，明示することが望まれる」というコメントを頂いた．本調査で使った刺激文が「[そ] の分類が言語学の観点から十分でな」く「表面的な特徴でしか文を捉えておらず，文構造の根本的な部分が捉えられていない」のは，意図したものである．その意図は EBM で RCT による確証バイアスの抑制と相同である．具体的に言うと，EBL の観点では，複雑な文構造が存在し，それに理論的な説明が必要であるという入来の言語学の想定が観察妥当性を阻害する確証バイアスと見なされる．それを回避するためにランダム生成した刺激文 (= 研究者の調査意図/確証バイアスを反映しない刺激文) を使っている．それが現段階で十分な変異の大きさを持っていないのは事実であるが，調査規模を考えるとこれ以上の変異は望めない．検討範囲が狭過ぎるという批判は，観察の範囲と得られた結果の信頼度との間のトレードオフを理解していない空論である．

同じ査読者から「(本研究は) EBM と EBL の統合を目指しているとあるが，医療における診察は基本的には個別の事例に個々に対応していくので，どの枠組みをとるにせよ，データから一般的特性を抽出して分類・分析する言語学とは根本的に相入れないように思われる」というコメントを頂戴している．この点については，私たちは査読者にまったく同意できない．まず「医療における診察は基本的には個別の事例に個々に対応していく」というのは医療行為が偶発的に被っている望ましくない制約であり，医療の本質条件ではない．医療は確かに処置の再現性の欠如，対照条件設定の困難により，大きく制限されている．実際，EBM は RCT を利用する事でその不本意な制約を緩和する方法論として提唱され，導入された (関連する議論は [10] を参照されたい)．更に，「データから一般的特性を抽出して分類・分析する言語学」という規定は，データ収集にバイアスがかかった状況で言語学に何が起きているかを方法論的に自覚した上での規定とは思われない．理由は §4 で簡単に論じる．

[13] データは MatLab Psychotoolbox で取得．

[14] データは SuperLab 4.5 で Self-Paced Reading の設定で取得．

[15] データは PsychoPy3 (Windows 10) で取得．

学生を対象に行った．函館では 10 名から (1 名当たり 90 文への反応で) 合計 900 反応を，東京では 15 名から (1 名当たり 80–83 文への反応で) 合計 1,223 反応を，岐阜では 10 名から (1 名当たり 76–79 文への反応で) 合計 778 反応を得た．こうして合計 2,901 反応を収集した．

2.5 前処理: はずれ値除去

生データにはエラーが含まれる．はずれ値はその副産物の一種である．検討した手法は 2 つある: rt1, rt2, ..., rt5, RT ごとの i) 標準偏差 (SD) を評価し，逸脱反応を除去; ii) Mahalanobis 距離を評価し，逸脱反応を除去．結果として i) の手法で SD < 3 の基準ではずれ値を除外した (詳細は [14] を参照されたい)[16]．これにより，2,697 個 (約 93%) の反応が有効と判断され，解析の対象となった．

本調査は予備調査として実際された．そのため，被験者の属性の無作為化はできていない (そもそも，十分な数の反応を集められていない)．更に，得られたデータの rt*i*, RT の分布を見ると，データ収集がすべての場所で同じ条件で実施されたと言いがたい[17]．この意味で，本調査で得たデータは代表性を持つとは言いがたく，その解析結果は (示唆的だとは言え) 一般性に限界がある．

3 RPART を使った階層的回帰分析

ARDJ で容認度評定値を得た 466 種類の刺激文のすべてで読み時間を取得したが，設定の不備から rt5 を持つ刺激 (31 種類) と持たない刺激 (435 種類の) とが生まれた．これらを同様に扱わず，分離した．以下の解析では rt5 を持たない 435 種類の刺激のみを対象としている．

3.1 背景

黒田ら [14] では，以上の手順で取得した RT データの回帰分析[18]の結果を示した．検討された変数は次の 20 種類である:

(3) a. resp (容認可能性の判断値: 0, 1), rt1, rt2, rt3, rt4, RT (分節ごとの反応時間と最後の分節を読んでから容認度判断までの所要時間)
b. seg1,size, seg2.size, seg3.size, seg4.size, seg5.size (刺激文の分節中の文字数)
c. r01, r12, r23, r3x (刺激文の評定値の区間 [0, 1), [1, 2), [2, 3), [3, ∞) の確率濃度)
d. edit:o, edit:p, edit:s, edit:v, edit:n (刺激文を生成するのに使った編集のタイプ)

r01, r12, r23, r3x の 4 種類は Kuroda et al. [7] で取得したデータ，resp, rt1, rt2, rt3, rt4, RT の 6 種類はその後に RT 実験で取得したデータ，残りは刺激文に固有の属性である．

用語法に関する注意: (3) の a, b を RT (reaction time) データと呼び，(3) の c, d を AR (acceptability rating) データと呼ぶ．これらすべてを一緒にしたものを RT x AR の混合データと呼ぶ．

この解析で得られた結果の概要を図 1 の表に示す．行が説明変数に対応し，列が独立変数に対応している．これから言えるのは，次の事である:

(4) a. resp の他変数による回帰では，切片, r01, r12 の影響が $p < .001$ 水準で，edit:{p,v}, seg1.size の影響が $p < .05$ 水準で，rt4 の影響が $p < .1$ 水準で有意
b. rt1 の他変数による回帰では，切片, rt2 の影響が $p < .001$ 水準で，rt3 の影響が $p < .01$ 水準で，edit:p の影響が $p < .05$ 水準で，edit:{o,s,v}, seg1.size の影響が $p < .1$ 水準で有意
c. rt2 の他変数による回帰では，rt1, rt3, edit:p の影響が $p < .001$ 水準で，rt4, edit:v の影響が $p < .01$ 水準，edit:{o,s} の影響が $p < .05$ 水準で有意 (切片の影響は非有意)
d. rt3 の他変数による回帰では，切片, rt2, rt4, seg3.size の影響が $p < .001$ 水準で，rt1 の影響が $p < .01$ 水準で，r01 の影響が $p < .05$ 水準で，edit:p の影響が $p < .1$ 水準で有意
e. rt4 の他変数による回帰では，rt3, seg4.size の影響が $p < .001$ 水準で，rt2 の影響が $p < .01$ 水準で，edit:{p,v}, seg1.size の影響が $p < .05$ 水準で，resp の影響が $p < .1$ 水準で有意 (切片の影響は非有意)
f. RT の他変数による回帰では，切片, seg4.size の影響が $p < .001$ 水準で有意，edit:p の影響が $p < .01$ 水準で有意
g. r01 の他変数による回帰では，切片, r12, r23, r3x の影響が $p < .001$ 水準で，seg4.size の影響が $p < .01$ 水準で，edit:s の影響が $p < .05$ 水準で，edit:o の影響が $p < .1$ 水準で有意
h. r12 の他変数による回帰では，切片, r01, r23, r3x の影響が $p < .001$ 水準で，resp, seg4.size の影響が $p < .01$ 水準で有意
i. r23 の他変数による回帰では，r01, r12, r3x の影響が $p < .001$ 水準で，切片, resp, seg4.size の影響が $p < .01$ 水準で，edit:s の影響が $p < .05$ 水

[16] その一方で，SD による濾過では，rt1, rt2, rt3 の極端な値の濾過が十分でない可能性がある．
[17] 例えば，函館で実施された実験で明らかに rt4 が長く，RT が短か目である．
[18] R package `stats` (v3.6.3) の glm (gaussian) で実行．

var	切片	resp	rt1	rt2	rt3	rt4	RT	edit:o	edit:p	edit:s	edit:v	r01	r12	r23	r3x	seg1.size	seg2.size	seg3.size	seg4.size	seg5.size
resp	***	–				.			*		*	***	***			*				
rt1	***		–	***	**			.	*	.	.					.				
rt2			***	–	***	**		*	***	*	**									
rt3	***		**	***	–	***			.			*						***		
rt4		.		**	***	–			*		*					*			***	
RT	***						–		**											
edit:o								–												
edit:p									–											
edit:s										–										
edit:v											–									
r01	***							.		*		–	***	***	***				**	
r12	***	**								*		***	–	***	***				**	
r23	**	**					.	.		*		***	***	–	***				**	
r3x	***	***						.		*		***	***	***	–				**	

図 1: RT× AR の混合回帰の有意性 [‘***’: $p < 0.001$; ‘**’: $p < 0.01$; ‘*’: $p < 0.05$; ‘.’ : $p < 0.1$]

準で，RT, edit:o の影響が $p < .1$ 水準で有意

j. r3x の他変数による回帰では，切片, resp, r01, r12, r23 の影響が $p < .001$ 水準で，seg4.size の影響が $p < .01$ 水準で，edit:s の影響が $p < .05$ 水準で，edit:o の影響が $p < .1$ 水準で有意

k. edit.x の他変数による回帰で有意な影響は認められない

重要な点は，rt1, rt2, rt3, rt4 の間の相互の影響は (当然のように計測値の間に相関があるため) 有意に認められるが，それらへの r01, r12, r23, r3x からの影響，つまり容認度評定値からの影響が認められなかった点にある．rt3 への r01 の影響が $p < .05$ 水準で認められたが，これは例外的な挙動である．

得られた結果が示唆しているのは，刺激文の容認度と反応速度との間に強い相関は認められず，おそらくほとんど因果性もないという可能性である．これは，逸脱のある文の処理には余分な時間がかかるという想定 [9] に合わない．この結果をどう解釈すべきか？ 本研究は予備調査の域を出ていないので，確定的に何かを言うのは難しい．それでも，この結果がデータ収集の片寄りから生じた偶発的なものなのか，そうではなく真の結果なのかを本調査で確認する必要があるのは，明らかである．

更に言うと，このような調査の必要性が意味しているのは，従来の言語学研究が一貫して方法論的基盤の確認を疎かにして来た事，その原因は特定の言語理論に由来する確証バイアスだったという事であり，証拠に基づく言語学が必要とされる理由は，そのような方法論的基盤の再確認である．動機は §4 で簡単に説明する．

3.2 決定木解析

回帰分析は，個々の変数の独立性を仮定し，それぞれが相対的にどの程度重要かを評価するが，独立変数の間の依存関係については積極的に語らない．複数の独立変数が単一の説明変数と相関している時，独立変数間にも相関がありそうだと予想できる程度である．私たちが知りたいのは，逆に独立変数の間の依存関係，特に影響の階層性である．決定木解析 (decision tree analysis) は，それを知りたい時に有用な解析手法である．この目的のために本研究では決定木の一種である RPART (Recursive Partitioning and Regression Trees) 解析 [1] を実行した[19]．説明変数は rt1, rt2, rt3, rt4, RT, resp, r01, r12, r23, r3x の 10 個に限定している．木の分岐を制御する complexity parameter (cp) の値として 0.01, 0.02, 0.03, 0.05 を試した．結果的に 0.02 を最適な値として選んだ (0.01 は分岐過剰，0.03 は分岐不足を生むと評価した)．

RPART を RT x AR の混合データに適用した結果を以下に示すが，概要を先に述べるとこうである．

(5) a. resp に rt4 との領域間相関が
b. rt2 に (rt1, rt3, RT との領域内相関に加えて) edit, r12 との領域間相関が
c. RT に (rt3, rt4 との領域内相関に加えて) edit, r01, r23 との領域間相関が

それぞれ認められたが，他の場合では領域間相関は認められなかった．

(6) a. resp と rt4 との相関はデータの性質を考えると驚くべきものではないが，b. rt2 と edit, r12 との領域間相関，c. RT と edit, r01, r23 との領域間相関の存在の確認は自明な結果ではない．

3.3 rt1 の RPART

図 2 に rt1 値の他変数による階層回帰の結果を示す．rt1 と有意味な相関が認められたのは rt2 のみである．近傍にある rt 同士は相関し，この結果は自然なものである．ただ，rt1 値の決定木解析は単純であり，RT データ内で

[19] R package の rpart (v4.1) を利用した．

閉じており，AR データとの相互作用を示していない．

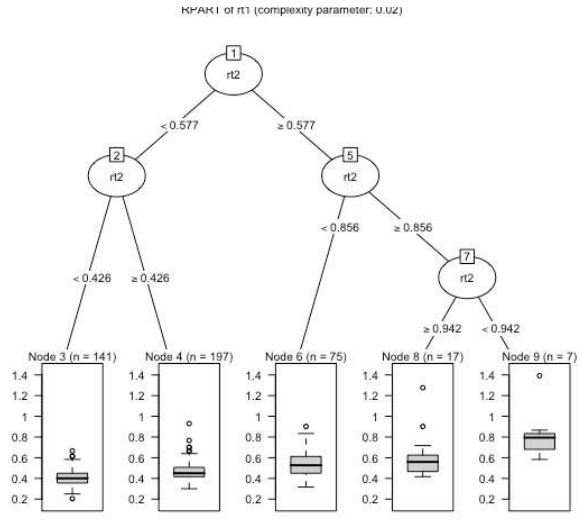

図 2: rt1 の RPART

3.4 rt2 の RPART

図 3 に rt2 値の他変数による階層回帰の結果を示す．この結果は rt2 には rt1, rt3, RT, r12, edit と rt2 の間に複雑な相関がある事を示している．この決定木は rt1, rt3, rt4 のそれとは質的に異なっており，意外な結果である．

rt2 にもっとも強く相関しているのは rt3 で，それに rt1 が次ぐ事がわかる．近傍にある rt 同士は相関するので，この結果は自然なものである．しかし，先の分岐に領域を跨ぐ edit:o,p,s,v との相関，r12 との相関が示されている．RT との相関は，rt3, rt4 との隔たりを考えると意外であるが，これは edit:o,p,s,v や r12 との相関の副作用だと考えれば，それほど不思議な結果ではない．

rt2 の決定木で重要なのは，それが RT データ内で閉じておらず，AR データとの相互作用の存在を示している点にある．その相互作用が rt1, rt3, rt4 には認められず，rt2 に認められた事に意味があるのかは今後検討したい．

3.5 rt3 の RPART

図 4 に rt3 値の他変数による階層回帰の結果を示す．rt3 には rt2 と seg2.size, seg3.size と有意味な相関がある．近傍にある rt 同士は相関するので，rt2 からの影響は自然なものである．ただ rt4 との相関はない．rtN のすべてに segN.size との有意な相関がある訳ではなく，このタイプの相関が認められるのは，ここにある rt3 と次に検討する rt4 だけである．rt3 値の決定木解析は単純で RT データ内で閉じ，AR データとの相互作用を示していない．

3.6 rt4 の RPART

図 5 に rt4 値の他変数による階層回帰の結果を示す．

rt4 と有意味な相関を持っているのは rt2, rt3, seg4.size である．近傍にある rt 同士は相関するので，rt3, rt4 との相関は自然なものである．seg4.size が rt4 と相関するのも自然である．rt4 値の決定木解析は単純で RT データ内で閉じており，AR データとの相互作用を示していない．

3.7 RT の RPART

図 6 に RT 値の他変数による階層回帰の結果を示す．RT と有意味な相関を持っているのは，edit, rt3, r01, rt4, r23 である．この決定木で重要なのは，それが RT データ内で閉じておらず，AR データとの相互作用の存在を示している点にある．その相互作用が認められたのは，先述の rt2 と RT のみである．rt2 が AR データに言及しているのは不思議だが，RT の決定木がそうなっているのは，変数の特性から理解でき，言語学の想定と合致している．

ただ，生じている相互作用は複雑である．まず，RT の値は edit の値で {o, p} 型と {n, s, v} 型に大別される．{o, p} 型は反応が短い．o は源文で，原則として逸脱していない文であり，反応が早い事は予測できる．格助詞のランダムな置換を伴っている p が o と同じぐらい早く判断されている．これは格助詞の用法の逸脱は一目瞭然なので，時間がかかっていない事を示唆してる．それに対し，n, s, v の逸脱はすぐに処置できるものではないという事を示唆しているように思われる．

その次に，RT の値は rt3 の値が 0.367 より小さい場合と大きい場合に大別される．小さい場合は，反応時間は短か目である．rt3 の値が 0.367 より大きい場合，容認度評定の確率分布濃度 r01, r23 と相関する．

3.8 r01 の RPART

図 7 に r01 値の他変数による階層回帰の結果を示す．

r01 は，r12, r23, r3x と同じく確率密度で [0,1] 間にあり，全体に占める割合を意味する．

r01 と有意な相関を持っているのは r23, r12, r3x である．決定木は複雑な階層性をなしているが，RT データや編集タイプや分節の大きさは影響していない．

3.9 r12 の RPART

図 8 に r12 値の他変数による階層回帰の結果を示す．この決定木解析は，後述の r23, rt2 の決定木と同じくノードが 7 個あり，複雑さが最大である．だが，その複雑さは AR データ内で閉じており，RT データとの相互作用を示していない．

3.10 r23 の RPART

図 9 に r23 値の他変数による階層回帰の結果を示す．この決定木解析は，前述の r12，後述の rt2 の決定木と同じくノードが 7 個あり，複雑さが最大である．だが，その複雑さは AR データ内で閉じており，RT データとの相互作用を示していない．

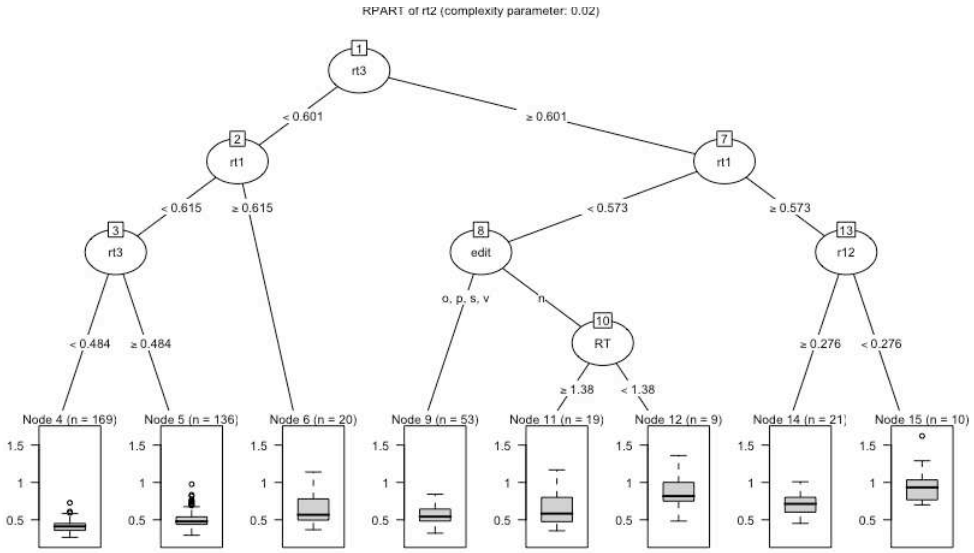

図 3: rt2 の RPART

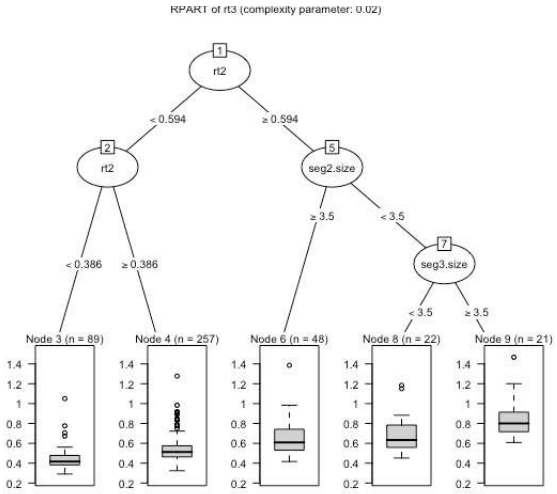

図 4: rt3 の RPART

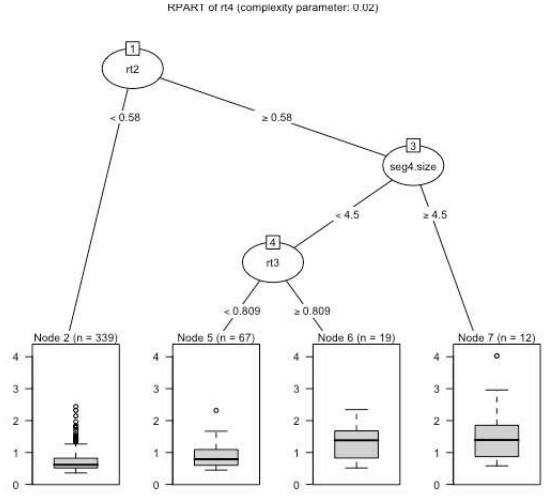

図 5: rt4 の RPART

3.11 r3x の RPART

図 10 に r3x 値の他変数による階層回帰の結果を示す．r3x 値の決定木解析は単純であり，AR データ内で閉じており，RT データとの相互作用を示していない．

3.12 resp の RPART

図 11 に resp 値の他変数による階層回帰の結果を示す．resp = 0 が刺激文が容認可能である事を，resp = 1 が容認不可能である事を表わす (resp は標準化された逸脱度を表わす)．

resp と有意味に相関しているのは r01, rt4, r12 である．r01 がもっとも強く resp と相関しており．それに rt4 と r12 との相関が次ぐ．r01 の値は逸脱度 resp の値と逆相関する．そのため，図 11 の決定木は，r01 が十分に大きくない (= 容認度が高くない) 場合に，r12 が resp と相関を持つという関係を表している．r12 が小さいならば，r01 と r12 の和が小さい = 容認度が低い事を意味している．

図 11 の決定木は，r01 がある程度大きくても，その後の処理で rt4 が大きい場合には容認度が下がる，どんでん返しの存在をエンコードしている．この結果は RT データと AR データの相互作用の存在を明らかにしている．その上，rt4 と resp との相関が有意で，かつ resp と RT との相関が有意でない事は，理論的に興味深い．

3.13 RPART 解析の結果のまとめ

§3.3–§3.12 の RPART の結果から次の事が窺える:

(7) a. AR データの変数 (容認度判断の構成変数) は RT データの変数から予測できない．容認度を処理時間で説明する事はできない．

b. RT データの変数のうち，rt2 と RT (と resp) は AR データの変数から予測できるが，残りはでき

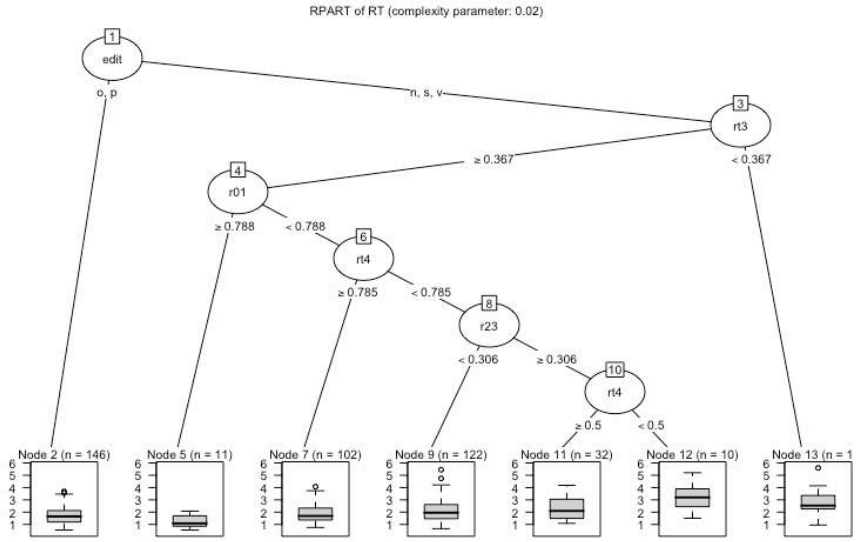

図 6: RT の RPART

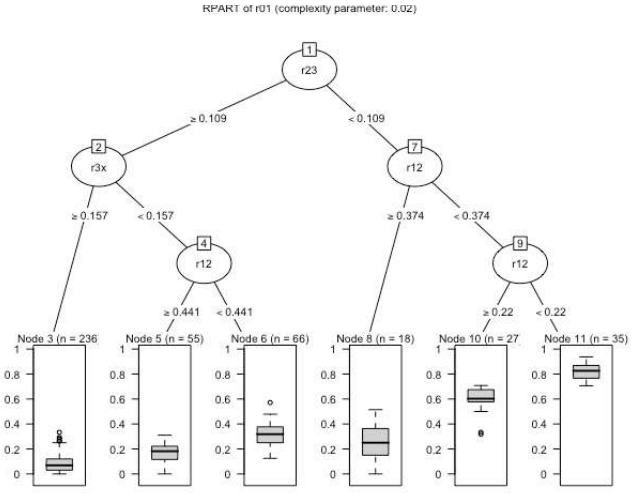

図 7: r01 の RPART

ない．具体的には rt1, rt3, rt4 を容認度から説明する事はできない．

AR データと RT データの変数の依存関係は非対称である．これは当然と言えば当然であるが，バイアスが少ない大規模データで実証的に示された事はないと思われる．

4 議論: なぜ EBL なのか？ [20)]

従来の言語学の研究成果は玉石混淆であり文字通りに受け取る事はできない．1) 精度と被覆率のトレードオフ関係を気にしていない，2) 記述性能と予測性能のトレードオフ関係を気にしていないの 2 つの難点があるばかりでなく，次の意味で証拠の扱い方が素朴過ぎるからである．

(8) a. すべての証拠が同程度に信頼できる訳ではない事を知らず，どの証拠がどれぐらい信用できるかに関する経験的見積もりの蓄積がない．
b. その結果，信用できる証拠と信用できない証拠の区別がついていない．

(9) それと相関して立証が確証バイアス (confirmation bias) [3, 5, 8] に汚染されている．

確証バイアスとは，人は証拠集めで無意識かつ体系的に，自説に好都合な事例に注目し，不都合な事例を無視する傾向の事である．

言語学と同じような確証バイアスに悩まされながら，その抑制に成功した分野がある．それは証拠に基づく医療 (Evidence-based medicine: EBM) [2, 4] である．表 3 は EBM が認定する証拠を信頼度の順に並べたものである．

表 3: EBM が想定する証拠のレベルの分類

Level	内容と例
1a	無作為化のある比較治験 (RCT) のメタ分析
1b	少なくとも 1 つの無作為化のある比較治験 (RCT)
2a	無作為化なしの同時対照群を伴うコホート研究
2b	無作為化なしの過去の対照群を伴うコホート研究
3	症例対照研究
4	前後比較や対照群を伴わない研究
5	症例報告やケースシリーズ
6	専門家個人の意見 (専門家委員会報告を含む)

証拠の強さは [Grade A: 言い切れる強い根拠あり; Grade B: 言い切れる根拠あり; Grade C: 言い切れる根拠なし] の 3 段階に大別される．Grade A には少なくとも 1 つの Level 1 (=1a か 1b) の研究が，Grade B には少なくとも 1 つの Level 2 (=2a か 2b) の研究が求められる．

[20)] 本節の記述は [10] の簡略版である．

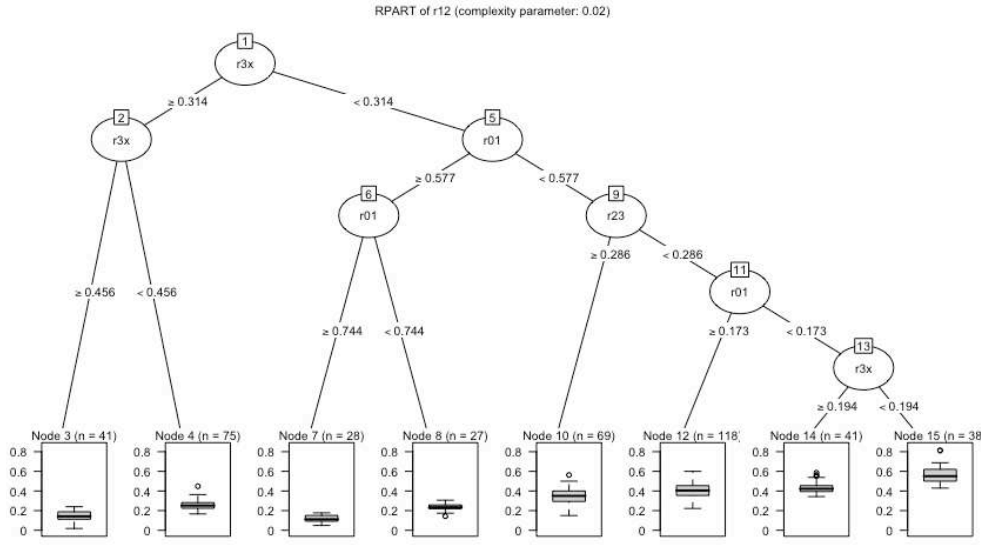

図 8: r12 の RPART

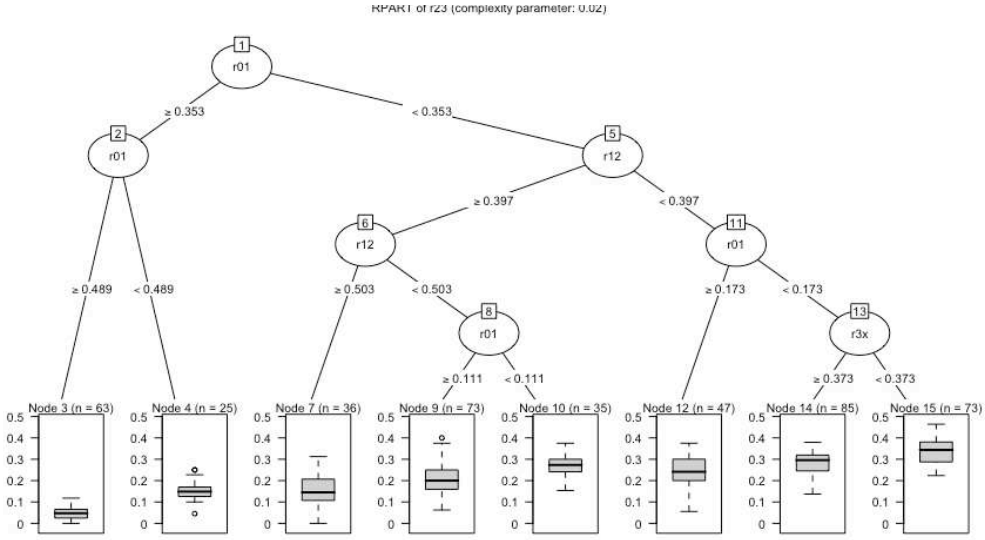

図 9: r23 の RPART

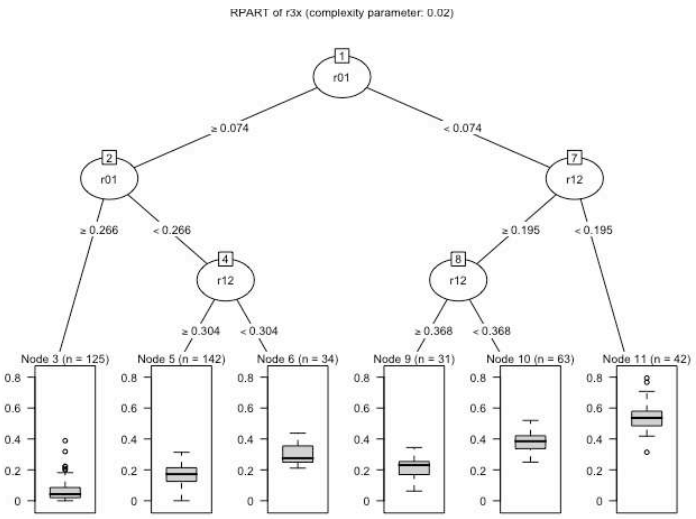

図 10: r3x の RPART

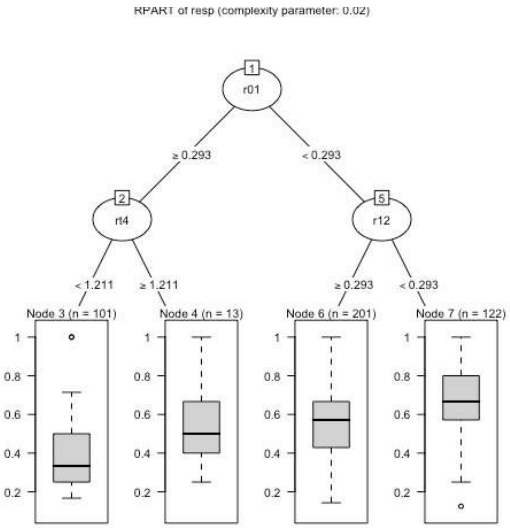

図 11: resp の RPART

EBM は臨床的意思決定 (代表例が診断) の支援を目的にしている．ここで科学的立証と医学/臨床的診断との間にアナロジーが成立すると考えてみよう．具体的に言うと，特定の説明を受入れるという科学的意思決定が，特定の診断を下すという医学的意思決定とアナロジー上の対応物だと考える．言語学の説明について少なくとも次が言えるはずである:

(10) 表 3 の EBM の信頼性の水準で言うと
   a. 言語学の立証の大半は，Level 5, 6 であり，意欲的な立証でも Level 4 止まり．
   b. 特に第三者の検証を得ていない独自判断 (e.g., 自作例の容認度判断) は (高名な研究者に拠るものであろうと)，Level 6 であり，言語学に権威主義が蔓延しているのは一目瞭然.

ARDJ の目標は言語研究で Level 1 の証拠を提供する事である．

5 終わりに

日本語の容認度評定データ (ARDJ) は大規模かつ確証バイアスの少なく，無償利用可能な日本語の文の容認度評定値のデータベースである．それは**証拠に基づく言語学** (Evidence-based Linguistics) [10] を将来的に実現するために必須の参照データだと筆者らは考える．本論文が報告しているのは予備調査の結果だとは言え，その可能性が垣間見れたと思う．

とは言え，本研究は探索型の研究であり，得られた結果から結論と言えるものを引き出すのは難しい．強いて結論らしきものを挙げるとするならば，容認性判断と処理負荷 (読み時間) との関係は言語学が想定しているより複雑かも知れないと言う可能性の提示だろう．得られた解析からは少なくとも，容認度評定と反応時間には単純な相関がなく，かつ意外な相関がある事が示されている．

読み時間と容認度評定値との対応づけは，行動データとの様々な対応づけの一つに過ぎない．眼球運動データや脳活動データとの対応づけが考えられる．ARDJ はそのような形で今度も拡充させて行く価値のある研究資源であると著者たちは信じる．

Acknowledgments

第 38 回認知科学会の 2 名の匿名査読者から有益な示唆を頂戴した．一方は理解に基づき，もう一方は無理解に基づくものであったが，それらはどちらも本論文の完成に当って有益だった．それらの貢献に感謝する．

RPART は RStudio (1.1.x) 上の R (version 3.6.3) で，glm 回帰は RStudio (1.1.x) 上の R (version 3.5.3) で実行した．

参考文献

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Routledge, 1983.

[2] G. 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).

[3] Th. Gilovich. *How We Know What Isn't So*. Free Press., 1993. [(翻訳) 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.].

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

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

[6] K. Kuroda, H. Yokono, K. Abe, T. Tsuchiya, Y. Asao, Y. Kobayashi, T. Kanamaru, and T. 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.

[7] K. Kuroda, H. Yokono, K. Abe, T. Tsuchiya, Y. Asao, Y. Kobayashi, T. Kanamaru, and T. 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.

[8] S. A. Vyse. *Believing in Magic: The Psychology of Superstition*. Oxford University Press, reprinted edition, 2000. [翻訳: 人はなぜ迷信を信じるのか: 思い込みの心理学. 朝日新聞社, 1999.].

[9] B. B. Wulfeck. A reaction-time study of grammaticality judgments in children. *Journal of Speech, Language, and Hearing Research*, 36(6):1208–1215, 1993.

[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論文集**, pp. 580–585, 2016.

[11] 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. 602–611, 2018.

[12] 黒田 航, 阿部 慶賀, 横野 光, 田川 拓海, 小林 雄一郎, 金丸 敏幸, 土屋 智行, and 浅尾 仁彦. (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. In **日本認知科学会第 *33* 回大会発表論文集**, pp. 557–562. 日本認知科学会, 2016.

[13] 黒田 航, 阿部 慶賀, 横野 光, 土屋 智行, 小林 雄一郎, 金丸 敏幸, 浅尾 仁彦, and 田川 拓海. 容認度評定に影響する要因の定量的評価: 日本語容認度評定データ (ARDJ) から得られた知見. In **日本認知科学会第 *36* 回大会発表論文集**, pp. 727–736, 2019.

[14] 黒田 航, 阿部 慶賀, 粟津 俊二, 寺井 あすか, and 土屋 智行. ARDJ を基にした容認度評定値と反応時間の多変量解析と混合回帰分析. In **認知科学会第 *37* 回大会発表論文集**, pp. 919–928, 2020.