

# アノテーション作業ログの提示が 仕様書の改善過程に及ぼす影響

杉本 麻衣<sup>1,a)</sup> 松下 光範<sup>1,b)</sup>

**概要：**本研究の目的は、複数の作業員で行うアノテーションにおいて、仕様書の改善を支援するインタフェースの実現である。作業員間で判断が分かると、設計者はこれを手がかりに仕様書を改善するが、得られるのは票数や一致度などの集計結果だけで、どの文でなぜ分かれたかは分からない。前報で筆者らは、作業員がどの文・語句をどう判断したかを記録するインタフェースを開発し、仕様書を改訂してその効果を一致率で評価したが、一致率は結果しか測らず、設計者が作業ログを見て推察をどう変えるかは明らかにならない。本稿では設計者1名に、集計結果のみの条件と作業ログを見る条件を与え、ハード／ソフトニュース分類の単一事例で原因分析を比較した。作業ログを示すと、不一致の原因を作業員の判断記録に紐づけて論じる記述が現れ、標準解の見直しにまで及んだ。ただしこの照合は、分布の全体に対して選択的でありうる。作業ログは原因分析を正解へ導く装置ではなく、推察を実際の判断と照合可能にする材料として働くこと示唆される。

## 1. はじめに

アノテーションは、複数の作業員が仕様書（内容分析におけるコードブックにあたる）に従って対象にラベルを付与する工程であり、社会科学の内容分析でも機械学習用のデータ構築でも広く行われている [1]。ここでは同じ対象に異なるラベルが付与される不一致が頻繁に生じる。この不一致を手がかりに仕様書のどこをどう直すかを判断するのは設計者であり、その判断に基づいて仕様書は繰り返し改訂される [2]。このとき設計者は、不一致という結果からその原因を推察する作業を行っている。

設計者が原因を推察するときの手がかりは、ある文章に対して何人の作業員が共通のラベルを付与したか、という集計結果である。集計結果からは、どのラベル対で判断が分かれたかまでは分かるが、どの文で、なぜ判断が分かれたかを特定することは難しい [3]。そのため設計者は、結果と原因の対応関係を自ら推察して補うことになるが、その推察を作業員の判断と照合するには、作業員への事後の聞き取りのような別の手続きを要する。

この対応関係を補うには、設計者の推察を、作業員が判断を下した時点の事実と照合する手段が必要である。筆者らはこれまでに、不一致が仕様書の不備によるものか解釈の多様性によるものかを判別するには判断の根拠の記録が

必要であるという考えの下、作業員がどの文にどのラベルを付けたかを記録するインタフェースを開発してきた [4]。本稿では、このインタフェースが作業員の操作を記録したものを作業ログと呼ぶ。作業ログには、作業員がどの文にどのラベルを付与したか、どの語句を判断の根拠として選択したか、どこで文を分割したかが含まれる。設計者は作業ログを参照することで、自らの推察を作業員の実際の判断と照合できる。ただし作業ログが提供するのは判別のための材料であり、不備として仕様書を修正するか多様性として残すかは設計者が判断する。前報 [4] では、このインタフェースを用いて仕様書を改訂し、改訂前後の作業員間一致率によってその効果を検証したが、設計者が作業ログを参照しながらどのように仕様書を改訂するのかという判断過程は明らかにしていない。

そこで本稿では、記録された作業ログを設計者に提示するダッシュボードを設計し、設計者が不一致の原因を推察し仕様書の改訂方針を決めるまでの過程が、集計結果のみを提示した場合と比べてどのように変化するかを観察する。題材は、ニュースの報道スタイル进行分类する枠組みである Reinemann ら [5] のハード／ソフトニュース分類である。読売新聞政治セクションの5記事に情報学部および同大学の学生13名がアノテーションを行い、仕様書設計者1名が、記事全体に対する判定のみを提示する条件と作業ログを提示する条件の両方で不一致の原因を分析する。観察するのは、集計結果のみから立てられた推察が作業ログの

<sup>1</sup> 関西大学

<sup>a)</sup> k570068@kansai-u.ac.jp

<sup>b)</sup> m\_mat@kansai-u.ac.jp

提示によってどのように変化した（あるいは変化しなかった）かであり、その変化が正しい方向であるかは問わない。

## 2. 関連研究

本章では2つの観点から関連研究を整理する。2.1節では、不一致の原因を特定しようとする試みと、判断の過程が記録として残る設計を概観し、本研究が対象とする問題を示す。2.2節では、設計者の推察が何から引き出されるかを述べ、その支援に関する研究を整理する。

### 2.1 不一致の原因を事後に特定する試み

本節ではまず、不一致をどう位置づけるかについての2つの立場を確認する。

Lombard ら [6] は、一致度の測定と報告を内容分析の中核の手続きと位置づけ、一致度が低い場合には操作的定義やカテゴリを見直すべきだとした。これに対し Aroyo と Welty [7] は、不一致をノイズとみなす立場を批判し、不一致はむしろ解釈の多様性を反映するシグナルであると主張した。彼らはまた、指示を精密にすれば不一致は解消されるが、作業者に自身が妥当と考えていない選択を強いることで一致度が取り繕われ、曖昧な事例がもつシグナルは失われるとも述べている。本研究が対象とするのは、個々の不一致が生じた後の段階で、それが仕様書の不備か解釈の多様性かをどう判別するか、という問題である。

この判別のため、不一致の原因を特定しようとする研究がある。Abercrombie ら [8] は同一作業者が時間を隔てて与えるラベルの一致度を併せて測ることで、主観性に由来する不一致と曖昧さに由来する不一致を切り分けられるとし、Sandri ら [9] は不一致の理由を4カテゴリに整理する分類体系を提案した。ただしいずれの手法でも、原因を判定するのは結果を事後に確認する人間である。Abercrombie ら自身が、事後の原因分類は分類者間で一致せず、付与された原因ラベルからは実際の一致度のパターンを説明できなかったと報告している。また Schaeckermann ら [10] は、作業者どうしの議論によって判断が収束するか否かで不一致を分類したが、収束しなかった不一致が仕様書の不備によるものかどうかは判定していない。Wacholder ら [3] はこの限界に対して不一致の位置をテキスト上で特定する手法を提案したが、これも付与済みのラベル範囲を事後に集約するものである。

一方、判断の過程が記録として残る設計は存在する。Krippendorff [1] は、記事全体のような大きな記録単位を直接記述するのではなく、文のように容易に合意できる小さな記録単位を記録し、そこから分析手続きを通じて大きな記録単位を記述することを推奨している。この二段階の手続きをとるかぎり、どの文にどの判断が下されたかは記録として残るが、この記録を設計者に提示する仕組みは一般的でなく、不一致を診断する段で参照されるとは限ら

ない。本研究は、設計者が集計結果だけを手がかりとする場面にこの記録を提示する点で既存研究と異なる。

なお、仕様書は一度作成すれば完成するものではない。Pustejovsky と Stubbs [2] は、アノテーション開発の初期段階において仕様書の修正と試験的なアノテーションを反復する手続きを MAMA サイクル (Model-Annotate-Model-Annotate) と呼んだ。本研究が観察の対象とするのは、このサイクルのうち、設計者が不一致を見て修正の方向を決める局面である。

### 2.2 設計者の推察の由来とその支援

設計者は仕様書を作成した当事者だが、作業者が個々の判断を下す場には立ち会っておらず、手元にあるのは集計結果だけである。このとき設計者は、自分が見ていない判断の原因をその結果だけから説明することになる。Nisbett と Wilson [11] は、判断の過程についての報告が、本人の内省であっても結果の記述だけを読んだ第三者の説明であっても精度が変わらないことを示した。いずれの報告も、事前にもつ因果についての想定から引き出されるためである。この知見に従えば、設計者の推察もまた、判断時の事実ではなく自身の想定を反映している可能性がある。

この問題に対し Schmer-Galunder ら [12] は、作業者を協働的・反復的に関与させ、議論と内省を通じて不一致の理由を作業者本人から引き出す手法を報告している。本研究は同じ問題に対し、作業者に理由を語らせるのではなく、判断時の操作を作業ログとして記録し、設計者に可視化する方法をとる。

設計者による分析の支援については、Pirolli と Card [13] のセンスメイキングモデルを援用する。これは分析者が大量の情報から結論を導く過程を、情報の探索・絞り込みを行う foraging loop と、証拠に適合する概念モデルを構築する sensemaking loop の2つに定式化し、各ループについて技術による介入が効きうる点 (leverage point) を挙げたものである。本研究で提示される作業ログは固定されており、新たな仮説に応じて追加のデータを収集することはないため、foraging loop に属する点は扱わない。本研究が参照するのは sensemaking loop に属する3点、すなわち証拠と仮説への注意の幅、代替仮説の生成、確証バイアスである。

## 3. 提案手法

本研究のツールは、作業者がアノテーションを行うツール (前報 [4] のものをを用いる) と、記録された作業ログを仕様書設計者に提示するダッシュボードからなる。後者が本研究の設計であり、前報では作業ログの分析を著者らが直接行っていたため、設計者に作業ログを提示する手段はなかった。

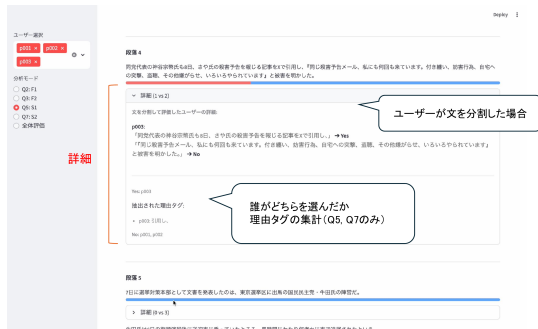


図 1 設計者に提示するダッシュボード。サイドバーで表示対象の作業者と指標を選び、本体に記事本文と文ごとの判断分布・作業者の内訳・根拠語句・文の分割操作を重ねて表示する。

### 3.1 作業側ツールと記録される操作

作業側側のツールは、仕様書の判断手続きを工程ごとに分割して順次提示し、各工程における操作を作業ログとして記録する\*1。画面は記事本文の表示領域と設問への回答領域からなる。F1 と F2 については、作業者は該当する文を青または赤でハイライトする\*2。F1 では青が「個人」、赤が「社会」、F2 では青が「エピソード」、赤が「テーマ」に対応する。S1 については、各文が該当するか否かを判断し（本稿ではこの判断を yes / no と表記する）、該当する場合はその根拠にあたる語句を本文から抽出する。S2 については文単位の該当判断を求めず、記事全体に対して該当する語句を抽出する工程のみを置いた。したがって S2 の抽出語句は、記録上どの文に対するものかの識別子をもたない。ハイライトと語句抽出はいずれも任意であり、付与しないまま次工程へ進める場合がある。文の分割はツール側が行うが、区切りが不適切な場合には作業者が動的に分割できる。

このように本ツールが求めるのは、根拠の記述ではなく、根拠にあたる箇所の指定である。作業者に事後に理由を語らせる手法 [12] では理由が作業者自身の言葉で与えられるのに対し、本ツールが記録し設計者に提示するのは、ハイライト付与、文単位の該当判断、根拠語句の抽出、文の分割、記事全体判定の選択であり、いずれも解釈の言語化を要求しない。根拠が本文への参照として残るため、設計者は作業者の判断を本文上の位置として参照できる。ただし、同じ理由から抽出した場合でも語句の範囲が一致しないことはありうるため、この記録は根拠の同一性を保証するものではない。

### 3.2 ダッシュボードの構成

ダッシュボードは、標準解を含む全作業者の作業ログを読み込み、文単位の判断分布の記事本文に重ねて表示する（図 1）。サイドバーでは、表示する指標と、表示対象とする作業者を選択する（作業者は任意の部分集合に絞り込め

る）。指標を選択した場合、本体は記事本文を段落ごとに区切り、各文について最大で次の 4 点を表示する。(1) 2 つのラベルが選ばれた比率を示す 2 色の帯、(2) 各ラベルを選んだ作業者の一覧、(3) 根拠語句の抽出を伴う指標については作業者が抽出した語句、(4) 当該文を分割して評価した作業者の操作である。帯の比により収束と不一致を一覧で判別でき、抽出語句によって作業者が本文のどこを指したかを比較できる。ただし表示される内容は指標によって異なる。3.1 節で述べたとおり S2 には文単位の該当判断がないため、S2 では (1) の帯と (2) の作業者の一覧が表示されない。また S2 の抽出語句は文の識別子をもたないため、ダッシュボードは語句と本文との文字列一致によって表示位置を決めている。記事全体評価を選択した場合は、主題設問および各指標の記事全体設問について作業者ごとの回答を一覧表として表示する。ただし文単位の表示と記事全体の表示は別のモードであり、両者を同一画面に並べて対応づける表示はもたない。

### 3.3 分析支援としての位置づけ

2.2 節で述べた sensemaking loop の 3 つの leverage point に照らして、本ツールの設計を整理する。証拠と仮説への注意の幅については、作業者ごと・文ごとに分散して記録された判断を、記事本文に重ねた分布として一覧可能な形に再構成する。代替仮説の生成については、どの文で判断が分かれたか、根拠とされた語句が一致しているか、文の区切り方が分かれていないかを表示する。確証バイアスについては能動的な対応をもたない。表示可能な分布を全件表示するため、設計者の推察を支持する文も支持しない文も画面上に存在するが、本ツールは両者を区別して表示せず、どこを検討するかは設計者に委ねられる。この設計が実際に何をもたらしたかは 6.1 節で述べる。

## 4. 実験

### 4.1 題材と設問構成

本研究は、報道スタイルの分類枠組みである Reinemann ら [5] のハード／ソフトニュースの多次元的な操作化を判断基準として用いる（翻訳と整理は大森 [14] に依拠した）。用いるのは 4 つの下位次元、すなわち「個人-社会」関連度 (F1)、「エピソード-テーマ」フレーム度 (F2)、「個人-非個人」的なりポート度 (S1)、「感情-非感情」的なりポート度 (S2) である。F1 は記事の結論が個人と社会のどちらに関わるか、F2 は主題を特定の事例と広い問題のどちらとして描くか、S1 は記者個人の見解が含まれるか、S2 は感情や戦いに関わる表現が用いられるかを問う。政治的に関連するニュースは常によりハードだとする Reinemann ら自身の主張に従えば、政治セクションから採られた対象記事は定義上ほぼ必然的にハードとなるため、トピック次元は除外した [14]。

\*1 画面構成の詳細は前報 [4] を参照されたい。

\*2 F1-S2 は本研究が用いる 4 指標であり、内容は 4.1 節で述べる。

表 1 アノテーションタスクの設問構成

| No. | 指標 | 単位 | 設問内容                 |
|-----|----|----|----------------------|
| Q1  | 前提 | 記事 | 主題を第 1 段落から抜き出す      |
| Q2  | F1 | 文  | 該当文を「個人」「社会」に色分け     |
| Q3  | F1 | 記事 | 記事全体としての F1 判定       |
| Q4  | F2 | 文  | 該当文を「エピソード」「テーマ」に色分け |
| Q5  | F2 | 記事 | 記事全体としての F2 判定       |
| Q6  | S1 | 文  | 個人の見解が含まれるか          |
| Q7  | S1 | 記事 | 記事全体としての S1 判定       |
| Q8  | S2 | 記事 | 感情・戦いに関わる語句を抽出       |
| Q9  | S2 | 記事 | 記事全体としての S2 判定       |

本タスクの特徴は、記事全体を判定単位とする Reine-mann らの概念を、文単位の判断を経て記事全体の判定に至る手続きとして構成した点にある。作業者は指標ごとに文単位の判断または語句の抽出を行ったうえで記事全体を判定する(表 1)。これにより、記事全体の判定という同一の結果に対して、そこへ至る文単位の判断の分布を観測できる。ただし S2 のみは文単位の該当判断をもたず、記事全体に対する語句抽出にとどまる。なお、本稿でいう仕様書とは、作業側ツールが画面上に提示する設問文と判断基準・判断例の総体を指す。

#### 4.2 対象記事と作業者

対象記事は、読売新聞オンラインの政治セクションに掲載された 5 記事である。2026 年 5 月 12 日 13 時 04 分時点の掲載記事を新しい順に走査し、文単位の判断分布を観察するのに必要な分量として 4 段落以上を満たす上位 5 件を採った。得られた記事は 514-806 字、4-7 段落である。設計者への提示順は収集順とは独立に無作為に決定しており、以降ではその提示順に従って記事 1-記事 5 と呼ぶ。内容は順に、大阪都構想を巡る党内対立、「国力研究会」の発足、チームみらい結成 1 年の会見、皇位継承に関する党見解案、ふるさと納税の手数料問題である。

本タスクの判断基準は前報 [4] の実験 2 と同一である。ただし前報の実装では F1 の記事全体設問が欠けていたため、4 指標すべてが記事全体設問を備えるよう F1 の記事全体設問(表 1 の Q3)を追加した。設問番号は前報と対応せず、対象記事および作業者も新たに集めたものである。

作業者は情報学部および同大学院の学生 13 名であり、各作業者は 5 記事のうち 3 記事を担当した。各記事の担当は後述の標準解を含めて 8 名または 9 名であり、担当人数と提示位置は釣り合っていない。

これとは別に、仕様書設計者自身が全 5 記事に対して標準解を作成している。標準解は作業者 ID p000 として作業ログに記録され、ダッシュボードが表示する票数にも 1 票として含まれる。したがって設計者は、分布の中に自らの判断を含む形で他の作業者の判断を見ることになる。

#### 4.3 分析実験と分析の手続き

分析実験の対象者は、本タスクの仕様書設計者 1 名であり、新聞社での記者経験を経てニュースを専門とする研究者である。不一致を観察して仕様書を改訂する立場にある者として選定した。対象者は前報 [4] の共著者でもあるが、本稿では分析に関与していない。前報では作業ログの分析を著者らが議論しながら行っており、設計者単独の推論ではなかったが、本稿では対象者が単独で分析し、著者はデータの収集とダッシュボードへの投入のみを担当した。

両条件に共通して、記事原文、設問、作業者の背景情報(学年と普段のニュース接触)を参照可能とし、標準解の作成に用いた作業者向けツールを開いて読み返すことも認めた。そのうえで、条件 A では記事全体設問に対する作業者の最終回答の集計表を提示した。この集計表には設計者自身の標準解も 1 行として含まれており、記事全体の水準では条件 A でも自他の判断を対比できる。一方、どの文にどの判断が下されたかは提示せず、自らの文単位のアノテーション結果も参照しないよう指示した。条件 B では、ダッシュボードを通じて文単位の判断分布、抽出された根拠語句、個別作業者の判断履歴を提示した。

両条件において対象者は記事ごとに、(1) 解釈が分かれた原因の分析、(2) その根拠、(3) 仕様書のどの記述に不具合があるか(あるいは保持すべき正当な解釈の多様性か)、(4) 修正案、(5) 対処方針への確信度(5 段階)とその理由を記述した。手続きは条件 A → 条件 B の固定順である。同一の記事集合を両条件で用いるため、条件 B を先に実施すると続く条件 A に条件 B で得た知識が混入することによる。条件 B の実施時に、対象者自身の条件 A での回答は提示していない。

分析は、記事×指標を単位として条件 A と条件 B の記述を突き合わせる形で行った。分類体系はあらかじめ設けなかった。対応づけは、対象者が本文中で名指しした指標によった。そのうえで、何がどちらへ変化したか(あるいは変化しなかったか)を両条件の原文を根拠として記述した<sup>\*3</sup>。さらに各変化について、それが条件 B でしか観測できない情報(文単位の票分布、抽出語句、文の分割操作、標準解との文単位の対比)に紐づくかどうかを判定した。条件 B でのみ提示される情報を参照した記述は、条件 A の段階では書きえなかったものである。ただしこの判定は記述の性質についてのものであり、当該情報の提示が変化を引き起こしたことまでを示すものではない。この点については 6.2 節で改めて述べる。

<sup>\*3</sup> ダッシュボードは文単位の表示モードを「Q2: F1」「Q3: F2」「Q5: S1」「Q7: S2」と表記する一方、記事全体評価のモードでは表 1 の番号を用いるため、同一画面内で Q7 が S2 の文単位表示と S1 の記事全体設問の双方を指す番号の混同があった。本稿では引用時に設問番号を指標名に置き換え、対象者が自らの標準解を指す固有名は作業者 ID p000 に置き換え □ を付す。

表 2 記事×指標ごとに観察された変化

|      | F1  | F2 | S1   | S2   |
|------|-----|----|------|------|
| 記事 1 | 具体化 | 消失 | 対応不能 | 対応不能 |
| 記事 2 | 変更  | —  | 具体化  | 新出   |
| 記事 3 | 具体化 | 新出 | 具体化  | —    |
| 記事 4 | 消失  | 新出 | 新出   | —    |
| 記事 5 | 具体化 | 維持 | 対応不能 | 対応不能 |

## 5. 結果

### 5.1 対照の概観

記事×指標を単位として条件 A と条件 B の記述を対照した概観を表 2 に示す。表中の語は、設計者の推察（以下、その内容を仮説と呼ぶ）がどう変化したかを言い表すための語彙であり、変化したことと変化しなかったことは等価な観察である。「具体化」は仮説が条件 B で下位の水準の根拠を得た場合、「変更」は原因の帰属先が変わった場合、「新出」は条件 A に前身をもたない仮説が現れた場合、「消失」は条件 A で論じた指標が条件 B で論じられなかった場合、「維持」は同一の仮説が保たれた場合、「対応不能」は対照ができなかった場合、「—」は両条件とも論じられなかった場合である。

記事 1 と記事 5 については、S1（個人見解）と S2（感情・戦い）の両指標、すなわち表 2 の 4 セルを対照から除外した。両記事の条件 B の記述は各記事 1 件のみだが、いずれも S1 を論じながら、参照データが記事 1 では S2 と確定でき、記事 5 ではどちらとも確定できないためである（設計者が挙げた記事 1 の「勝利した」「おかしい」は S2 の抽出にのみ現れるが、記事 5 の「問題視されている」「指摘された」は両指標の抽出に現れる）。この交錯は条件 B に固有ではない。条件 A の記事 1 でも設計者は仕様書評価の欄で「一文のなかに一箇所でも個人的な表現や感情的な表現があれば」と両指標を連結し、作業者の側でも記事 1 の「冷ややかな見方が目立つ」は S1 の抽出語句に 6 名、S2 に 5 名が登録している。なお条件 B の交錯は、S2 が文単位の該当判断をもたないという工程設計（3.1 節）と、4.3 節の脚注に述べた番号の混同のもとで生じている。本研究はこの 2 記事を、仕様書の不備と解釈の多様性のいずれとも判別しがたい対象として、対応不能のまま開示する。ただしこれらのセルは仕様書の不備の候補としても有力であり、対照から外したことが、変化を報告した残りのセルに偏りを生みうる。

### 5.2 聞き取り要求の消失と標準解の再考

条件 A と条件 B で記述の有無が分かれたのは、作業者への事後聞き取りを求める記述である。条件 A において設計者は、作業者本人に理解を尋ねることを 6 回求めた。内訳は記事 2 で 1 回、記事 3 で 1 回、記事 5 で 4 回である。記

事 2 では自身の標準解が学生の回答と食い違ったことについて「むしろ学生側の理解を聞いてみたほうが良い気がする」と述べており、記事 5 ではこの要求が集中する。条件 B では、この種の要求は 5 記事を通じて一度も現れない<sup>\*4</sup>。

条件 B にのみ現れたもう 1 つの観察が標準解の再考である。条件 A でも記事全体の水準では自他のズレが見えており、設計者は記事 2 について「ほぼ学生側が一致しているのに対して [p000] のだけが異なっている」と書いている。しかしそこでの反応は「学生側の理解を聞いてみたほうが良い」という外部への照会であった。条件 B では、同型のズレに対して自らの標準解を見直す記述が現れる。記事 3 の S1 について設計者は、段落 1 の「課題となる」で判断が分かれていることを指摘し、『『課題となる』は p000 は抽出していないが抽出しても良い』と述べている。作業ログでは「課題となる」は yes 5 名・no 4 名に分かれており、標準解 p000 はこれを抽出していない。設計者はこの分布を参照したうえで、標準解の側を見直している。

ただし条件 A の記事 5 にも「[p000] は発表を主題としたが、間違いとも言い切れない気がする」という自己相対化の記述はあり、条件 B にのみ現れたのは文および抽出語句の単位で標準解を見直す記述である。またこの見直しは自らが少数側にあることを示す表示のもとで生じており、判断の再考と多数側への同調とを本データから判別することはできない。

### 5.3 根拠として引用された情報の紐づけ

条件 A の記述が具体性を欠いていたわけではない。記事 2 では S1 について『『狙いがある』『みられる』... という推定表現をとれば、0 になると思ったが、2 になったのは意外であった』と記事中の語句を引用している。しかしこれらの語句は、設計者が記事を読み返して拾ったものであり、作業者が実際にそこを見ていたかは確かめられていない。また条件 A の根拠欄は、5 記事を通じて「回答が 3 つにバラけているため」といった分布の形についての記述に帰着している。

条件 B では、同じ種類の語句が作業者の判断記録に紐づいた形で引用される。記事 2 の S1 について設計者は『『狙いがある』の場合は 4 対 4、それが『狙いがあるとみられる』になると 7 対 1 となっている。みられるのような言葉があれば個人と明確に分かるようだ』と述べる。作業ログでは、段落 1 の「... 足元を固める狙いがある。」は yes 4 名・no 4 名、段落 4 の「... 囲い込む狙いがあるとみられ

<sup>\*4</sup> 設計者は条件 B の中で F1 の修正案（Q2 に「記事中の個人や組織の問題ではない」旨を明記する案）を記事 1・記事 2・記事 3 と 3 度繰り返しており、条件 A でも記事 5 の 1 記事内で聞き取り要求を 4 度反復している。すなわち、この設計者は同一の内容を反復して書く書き手である。ただし、条件 B で書かなかったことを直ちに要求の消失と読むことはできない。条件 A ですでに要求を書いたため反復しなかった可能性、および同一記事を 2 周目に見たことによる可能性がある。

る。」は yes 7 名・no 1 名であり、票数は設計者の述べたとおりである。ただし比較された 2 文は段落も主語も異なり、この対比のみから「みられる」を分岐の要因として同定できるわけではない。本稿が観察するのは推論の当否ではなく、根拠が語句の水準に求められたことであり、語句ごとの票の分かれ方は、記事全体単位の票数しかもたない条件 A では参照できない。同様に、根拠が語句の記録に紐づく例は記事 1 の F1、記事 3 の S1、記事 4 の S1 にも見られる。

また、記事 2 の F1 では、原因の帰属先が条件 A と条件 B とで変わっている。条件 A では「自民党内の記事をどう捉えるかというニュースへの理解度が出た」と作業者の理解度に帰属されていたのに対し、条件 B では「記事に個人の動向を書いたものを『個人』と判断し、政府の場合は『社会』となっている」と、文の書かれ方に帰属されている。ただし両者は排他的ではなく、条件 B で理解度に言及しなかったことがその仮説の放棄を意味するとは限らない。

#### 5.4 規則と文単位ハイライト分布の照合

本節では、条件 B で設計者が述べた規則を作業ログの分布全体と照合する。対象となる F1 の仮説は条件 A の記事順のなかで発展したものであり、記事 1 の「1 は個人を吉村の話と誤解した可能性があるのではないか」から、記事 4 の修正案の「個人というのは記事に個人が出ているということではなく、記事の結論が個人（読者など）に帰結するか」という定式化に到達している。登場人物から帰結への捉え直しは条件 B の提示以前に生じており、条件 B が加えたのは、この仮説を文単位の要素に結びつけることであった。

記事 1 の条件 B で設計者は、仮説の根拠として「判断が分かれているところは吉村氏、馬場氏と名前がでており、まとまっているところは与党といった組織になっている」を挙げる。この規則を F1 の文単位ハイライト分布の全体と照合した結果を表 3 に示す。3.1 節で述べたとおりハイライトの付与は任意であるため、文ごとの合計票数は一様でない。

規則を支持する箇所は存在する。段落 4 の 3 文は人名を含み、青 2: 赤 2, 青 1: 赤 1, 青 2: 赤 1 といずれも拮抗し、段落 5 の与党に言及した文は青 0: 赤 4 で社会側に収束している。一方で規則に反する箇所も同じ分布のなかにある。段落 1 の吉村代表に言及した文は人名を含みながら青 0: 赤 4 で収束し、段落 6 の「市の廃止の賛否を府民が判断するのはおかしい」は人名を含まないにもかかわらず青 2: 赤 2 と判断が分かれ、段落 7 の吉村氏に言及した文は青 4: 赤 1 で個人側に偏っている。観察されたのは、設計者が作業ログによって仮説を確証したのではなく、仮説を作業ログの特定の要素に結びつけたことである。

#### 5.5 「大部分」に関する仮説の消失

条件 A において設計者は、「大部分」という語の閾値が不明瞭であるという仮説を記事 1 で提出し、その修正案の文面にまで書き、記事 3 ではこの仮説を反復した。「大部分」は、4 指標のうち F2・S1・S2 の記事全体設問の選択肢に用いられている語であり、文単位の判断の記事全体の判定にまとめる際に、どの程度を占めれば一方の極とみなすかを表す。

しかし、この仮説は、条件 B において 5 記事を通じて一度も再登場しない。記事 4 でも論じられる指標そのものが入れ替わっており、設計者は条件 A の記事 4 で F1 の定式化に到達し修正案まで書いているのに対し (5.4 節)、条件 B の記事 4 で論じられるのは F2 と S1 である。

消失の理由は単一事例では判別できないが、記述の不在を観察単位として扱う限界のもとで、少なくとも次の 4 つが考えられる。(1) 作業ログが分岐点を特定の語句として示したため、閾値という説明が不要になった。(2) ツールが提示した語句単位の情報に注意が向き、検証されるべき仮説が放置された。(3) 本ツールは文単位の判断分布と記事全体の回答を別々のモードで表示するため、この仮説が問題にしていた手続きが画面上に並んで現れなかった。(4) 条件 A の記事 1 ですでに修正案の文面まで書いたため、条件 B で再び論じる必要がなかった。第 1 の説明では作業ログが有効に、第 2 の説明では有害に働いたことになり、第 3・第 4 の説明ではそのいずれでもない。同一の観察が有効・有害・中立のいずれとも読めるため、本研究はどれかを選ばず 4 つを併記して開示する。

なお確信度は条件 A で 3, 3, 4, 4, 4, 条件 B で 5, 5, 4, 3, 3 であるが、差分が記事の提示順に対して単調に減少しており、順序による説明を排除する特徴をもたないため、主たる観察としては扱わない。

### 6. 考察

#### 6.1 センスメイキングの介入点との対応

3.3 節の 3 点に照らして、観察された変化を整理する。証拠と仮説への注意の幅については、条件 A で根拠が記事全体の票の散らばりであったのに対し、条件 B では特定の文・語句の票の分布が根拠となった (5.3 節)。条件 A でも語句への着目はあったが、それを作業者ごと・文ごとの分布として一覧しうるのは条件 B である。ただし、条件 B でしか提示されない情報に紐づく記述が条件 A に現れないのは定義上のことである。必然でないのは、設計者がその情報を根拠として実際に採用したことのほうであり、条件 B においても分布の形についての記述を続けることはできた。代替仮説の生成については、条件 A に前身をもたない仮説が現れたほか (表 2 の「新出」のセル。ただし対応不能とした 4 セルを除く)、記事 2 の F1 で不一致の原因の帰属先が作業者から文へ移った (5.3 節)。これは既存の仮説

表 3 記事 1 における F1 の文単位ハイライト分布 (青=個人, 赤=社会, \*=人名を含む文)

| 段落 | 文 (冒頭)         | 青 | 赤 | 段落 | 文 (冒頭)         | 青 | 赤 |
|----|----------------|---|---|----|----------------|---|---|
| 1  | 日本維新の会が, 看板…   | 0 | 1 | 4  | 馬場伸幸前代表は 4 月…* | 1 | 1 |
| 1  | 来年 4 月の知事任期…*  | 0 | 4 | 4  | 馬場氏は, 都構想は…*   | 2 | 1 |
| 2  | 吉村氏は 8 日, 大阪…* | 1 | 1 | 5  | 与党が 3 月にまとめた…  | 0 | 4 |
| 2  | その上で, 任期中に…    | 0 | 2 | 5  | それでも維新内には…     | 1 | 4 |
| 3  | 維新は結党当初から…     | 0 | 2 | 6  | 住民投票の手法を…      | 1 | 3 |
| 3  | だが, 過去 2 回の…   | 1 | 2 | 6  | 吉村氏は, 副首都の…*   | 2 | 2 |
| 3  | 悲願実現にこだわる…*    | 1 | 2 | 6  | 一方, 党内では「市の…   | 2 | 2 |
| 4  | もっとも国会議員団…*    | 2 | 2 | 7  | 党勢は伸び悩んで…*     | 4 | 1 |

の精緻化ではなく, 帰属先の移動として読める。

確証バイアスについては, 互いに逆を向く 2 つの観察が得られた。1 つは照合が可能になったことである。条件 A において設計者は自らの推察を検証する手段を手元にもたず, 照合を作業への聞き取りという外部の手続きに委ねようとしていた。条件 B ではその照合が設計者自身の手で行われ, 対象は他者の判断にとどまらず自らの標準解にまで及んでいる (5.2 節)。もう 1 つは, 根拠として挙げられた文が規則を支持する側に限られていたことである (5.4 節)。支持しない文も同じ画面上に同じ形式で表示されており, 記述に現れなかったことが検討されなかったことを意味するとは限らない。設計者自身も, 分析終了後の事後アンケートにおいて, 不一致箇所の把握しやすさを 5 段階の 5 と評価する一方, 「検討する部分が多く, どこを見る必要があるのかを迷った」と述べている。

## 6.2 情報を強調して提示することのリスク

前節では, 作業ログの提示のもとで観察された変化を介入点に沿って整理した。しかし, 判断時の情報を目立つ形で示すことは, 設計者の推察を実際の分岐とは異なる方向へ導きうるという別の可能性も伴う。

Nisbett と Wilson[11] は, 人が判断の原因を報告するとき, 過程の記憶をたどるのではなく, その場で思いつく候補のうちもっともらしいものを選んでいと述べ, 実際には影響していない候補の利用可能性が高められた場合には, それが誤って原因として報告されることも示している。本ツールは判断時の情報を再び利用可能な形で提示して候補を増やす。利用可能性の回復と誤帰属は切り離せるものではなく, 条件 B で観察された変化の一部がツールの強調した情報への誤帰属である可能性を本研究は排除しない。なお, 本研究が観察の単位とする設計者の記述もまた事後に構成された説明であり, この留保は観察された変化だけでなく, 照合が行われたという記述そのものにも及ぶ。

## 6.3 照合の手段としての作業ログ

本研究が示したのは, 作業ログを提示した条件で, 設計者が推察を組み立てる過程が集計結果のみの条件とは異なっていたという観察であり, 仕様書の修正方向の判別を

正しく導くことを示したのではない。集計のみの条件では, 設計者は他の作業者の文単位の判断を参照できず, 作業への聞き取りを 3 記事で計 6 回求めた。これは集計結果だけでは推察を照合しえないことを, 設計者自身の行動が示したものと読める。作業ログの条件では, 設計者自身が分布を参照して照合する記述が現れ, 既存の仮説は語句の水準の要素に結びつけられた。聞き取り要求の消失そのものは条件 B に帰属できるとは限らないが, Schmer-Galunder ら [12] が聞き取りによって扱おうとした同じ問題に, 作業ログが別の手段を与える可能性を示す。ただし作業ログが代替しうるのは照合の手続きであって, 作業者が語る理由そのものではない。

本研究のツールが提供するのは判別のための材料であり, 不備として仕様書を直すか解釈の多様性として残すかは設計者が判断する。しかし, その材料が判別そのものを動かすことは, 5.2 節で報告した標準解の見直しに現れている。設計者は自らが抽出していなかった語を, 他の作業者の分布を見たうえで抽出してもよいと述べた。これが判断の再考が多数側への同調かは判別できないが, いずれであれ, 何を正解とするかの検討に他者の分布が働いたことは共通する。ただし発言は「抽出しても良い」とどまり, 標準解そのものは変更されていない。

## 7. 本研究の射程と今後の展開

本研究は単一事例の質的観察であり, 報告した変化は分析者の解釈に基づく示唆であって, 効果の存在やその一般化を主張するものではない。

同一の記事集合を両条件で用いる本デザインでは順序をカウンターバランスできない。条件 B の記述には, 語句ごとの票や標準解との対比のように条件 A では物理的に参照できない情報によるものがある一方, 修正方針の変化や 5.2 節の聞き取り要求の消失のように, 条件 A でも参照可能であった情報についてのものもある。いずれについても, 条件 B の提示が変化をもたらしたのか同一記事を 2 周目に見たことによる再考がもたらしたのかは切り分けられない (6.2 節)。さらに, 条件 A で仮説と修正案を言語化させたこと自体が条件 B の記述に一貫性圧力を与えた可能性, および標準解を分布に含めて提示した設計が 5.2 節の

見直しに影響した可能性も、本デザインでは切り分けられない。記事の提示順は無作為に決定したが、対象者が1名である以上1つの順序しか残らず、実際、設計者は先行する記事の分析を後続の記事に持ち越している。

条件Aは現行の実践をそのまま再現したものではない。設計者は標準解の作成を通じて自らの文単位の判断を作成しており、条件Aではその参照を教示によって禁じている。また対象者はツールの目的と期待される結果を知る立場にあり、デマンド特性の影響も順序効果と同様に切り分けられない。加えて、本稿の分析手続き自体が、単一の読み手がコードブックなしに事後に結果を読んで変化の種類を定めるという、本研究が問題としている状況と同型であり、この点で本稿の記述もまた独立の確認を経ていない。

なお、5.4節の反例の判定は、「名前がでており」を人名の有無と読む著者の操作化に依存し、ハイライトが任意であるため、表3の票のない文が非該当か未判断かも区別できない。また、ある文を反例と判定するには、その票分布を「判断が分かれた」と見做すか「まとまった」と見做すかを定める必要があるが、本稿はその基準を定めていない。そのため反例の同定は網羅的でない。設計者の発話が記事1の特定箇所を指したものであった可能性も残る。さらに、本ツールが記録するのは箇所の指定であって理由ではないため、記録から不備と多様性の判別に直接には至らない(5.1節の対応不能セルはその現れである)。

以上を踏まえ、今後の展開として、記事集合を2群に分けて条件間で入れ替えること、複数の対象者を立て記事の提示順を対象者ごとに振ること、および変化の読み取りについて別の読み手による独立の確認を行うことに取り組む。ツール側では、文単位の判断分布と記事全体の回答を対応づけて提示すること、指標間で提示される情報の粒度と設問番号の表記を揃えること、および検討すべき箇所の絞り込みを支援することが、改良の方向として挙げられる。

## 8. おわりに

本稿では、作業者の判断過程を記録した作業ログを設計者に提示するダッシュボードを設計し、集計結果のみを提示する条件との対照によって設計者の分析過程を観察した。

本稿の達成点は次の3点である。第1に、作業者に理由を語らせることなく判断時の操作を設計者に提示し、推察を作業者の実際の判断と照合可能にする手段を示した。第2に、その提示のもとで推察がどう変化したかを両条件の原文に紐づけて記述し、根拠が分布の形から特定の語句へ移ったこと、聞き取りの要求が現れなくなったこと、標準解自体の見直しが生じたことを観察した。第3に、同じ提示が仮説の選択的な照合を伴うことを、設計者の述べた規則と分布全体との照合によって示した。これらは、作業ログが原因分析を正解へ導く装置ではなく、設計者の推察を実際の判断と照合可能にする材料として働くことを

示唆する。

今後は、7章に述べた制約を解いた条件下で観察を重ね、この照合が仕様書の改訂にどう反映されるかを検討する。

**謝辞** 本研究はJST RISTEX(課題番号JPMJRS23L2)の支援を受けた。

## 参考文献

- [1] Krippendorff, K.: *Content Analysis: An Introduction to Its Methodology*, Sage Publications, 2nd edition (2004).
- [2] Pustejovsky, J. and Stubbs, A.: *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*, O'Reilly Media (2012).
- [3] Wacholder, N., Muresan, S., Ghosh, D. and Aakhus, M.: Annotating multiparty discourse: Challenges for agreement metrics, *Proc. 8th Linguistic Annotation Workshop*, pp. 120–128 (2014).
- [4] 杉本麻衣, 松下光範, 藤代裕之: 曖昧さを含む仕様書の改善を目的としたアノテーション支援ツールの検討, 情報処理学会研究報告, Vol. 2025-HCI-216, No. 18, pp. 1–7 (2026).
- [5] Reinemann, C., Stanyer, J., Scherr, S. and Legnante, G.: Hard and soft news: A review of concepts, operationalizations and key findings, *Journalism*, Vol. 13, No. 2, pp. 221–239 (2012).
- [6] Lombard, M., Snyder-Duch, J. and Bracken, C. C.: Content analysis in mass communication: Assessment and reporting of intercoder reliability, *Human Communication Research*, Vol. 28, No. 4, pp. 587–604 (2002).
- [7] Aroyo, L. and Welty, C.: Truth is a lie: Crowd truth and the seven myths of human annotation, *AI Magazine*, Vol. 36, No. 1, pp. 15–24 (2015).
- [8] Abercrombie, G., Dinkar, T., Cercas Curry, A., Rieser, V. and Hovy, D.: Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement, *Proc. 4th Workshop on Perspectivist Approaches to NLP*, pp. 63–74 (2025).
- [9] Sandri, M., Leonardelli, E., Tonelli, S. and Jezek, E.: Why don't you do it right? Analysing annotators' disagreement in subjective tasks, *Proc. 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2428–2441 (2023).
- [10] Schaekermann, M., Goh, J., Larson, K. and Law, E.: Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work, *Proc. ACM on Human-Computer Interaction*, Vol. 2, No. CSCW, pp. 154:1–154:19 (2018).
- [11] Nisbett, R. E. and Wilson, T. D.: Telling more than we can know: Verbal reports on mental processes, *Psychological Review*, Vol. 84, No. 3, pp. 231–259 (1977).
- [12] Schmer-Galunder, S., Wheelock, R., Jalan, Z., Chvasta, A., Friedman, S. and Saltz, E.: Annotator in the loop: A case study of in-depth rater engagement to create a prosocial benchmark dataset, *Proc. AAAI ACM Conference on AI, Ethics, and Society*, Vol. 7, No. 1, pp. 1319–1328 (2024).
- [13] Pirolli, P. and Card, S.: The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis, *Proc. International Conference on Intelligence Analysis*, Vol. 5, No. 1, pp. 2–4 (2005).
- [14] 大森翔子: メディア変革期の政治コミュニケーション: ネット時代は何を変えるのか, 勁草書房 (2023).