ニュースのハード・ソフト分類における判断の齟齬とその要因分析

An Analysis of Factors Underlying Inter-Annotator Disagreement in Hard/Soft News Classification

杉本 麻衣 † 藤代 裕之 ‡ 松下 光範 ‡
Mai Sugimoto Hiroyuki Fujishiro Mitsunori Matsushita

† 関西大学 総合情報学部
Faculty of Informatics, Kansai University

‡ 法政大学 社会学部
Faculty of Social Sciences, Hosei University

Abstract: ニュース報道のハード・ソフト区分は，記事が読者に与える影響を解明する不可欠な指標である．しかし，この区分の分類基準は定義が曖昧であり，一貫性を保つことが困難である．そこで本研究は，判断の段階的な分解と根拠箇所の記録・可視化を可能にするアノテーション支援ツールを開発し，分類の齟齬が生じる原因を調査した．その結果，不一致の主因が，作業者の能力不足ではなく，専門家の暗黙知と言語化された定義とのズレや，記事の記述構造との不適合にあることを明らかにした．

1 はじめに

現代のデジタルニュース空間においては，記事の記述構造に大きな変化が生じている．かつての新聞報道は，最も重要な事実や結論を冒頭のリード文に配置する「逆三角形型」の構造が基本とされ，読者へ知識や情報を効率的に伝達することに重きが置かれていた [12]．しかし，インターネット以降のデジタルニュース空間においては，単なる客観的な事実の伝達にとどまらず，読者の関心や「共感」を強く惹きつける手法が求められるようになっている．より多くの読者を獲得し記事を拡散させるために，事象の背後にある文脈や特定の個人の感情・エピソードといった細部を描き出し，読者に追体験させるような「物語型」の手法の重要性が高まっている [10]．

このようなニュースの伝え方が読者に与える影響を分析するためのアプローチとして，メディア研究においては報道内容の「伝え方」の違いを「ハードニュース」と「ソフトニュース」に分類する枠組みが存在する [8]．しかし，従来の研究では，定義自体が曖昧であり，トピック（政治か芸能・スポーツか）による一次元的な分類に依存していた．そのため，読者に与える効果については一貫した結論が得られていなかった．この限界を克服するため，Reinemann ら [6] はニュース報道を単一の基準で二分するのではなく，「トピック」「フォーカス」「スタイル」という 3 つの次元で尺度化して分類する枠組みを提唱した．大森 [11] によるこの枠組みを用いた実証研究では，「フォーカスのソフト化」は，複雑な政治的問題が個人の生活にどう関わるのかを分かりやすく解説することで「政治の複雑さ」を解消し，政治関心や内的有効性感覚を高めるポジティブな効果を持つことが明らかになった．さらに，「スタイルのソフト化」は，過剰になると政治家等に対する外的有効性感覚を低下させ，政治不信を助長する副作用をもたらすことも示されている．

このように，ニュース報道の多次元的なハード・ソフト区分は，メディアの報じ方が読者の政治意識に与える影響を要因別に解明するための不可欠な指標として位置づけられる．しかし，この指標を用いた客観的な分析を行うためには，実際のニューステキストに分類枠組みを適用し，データとして分類する必要がある．テキストに対して分類枠組みを適用し，ラベルやカテゴリの情報を付与する作業は「アノテーション（あるいはコーディング）」と呼ばれる．Ziems ら [9] が指摘するように，この作業において特定の理論的構成概念をテキストに適用する際，分類基準の定義が曖昧になりやすく，客観的な判断の一貫性を保つことが困難になるという特有の難しさが存在する．また，Krippendorff[4] が述べるように，コーディングは単なる物理的なキーワードの測定ではなく，観測されたテキストを手がかりとして，背後にある

* E-mail: k570068@kansai-u.ac.jp
† E-mail: fujisiro@hosei.ac.jp
‡ E-mail: m_mat@kansai-u.ac.jp

観測できない文脈やニュアンスを読み解く，高度な推論を伴う解釈的なプロセスである．本研究で扱うニュースの「フォーカス」や「スタイル」の判定においても，記事の焦点が社会と個人のどちらに向いているか，あるいは表現にどの程度感情や個人的見解が含まれているかといった，文脈依存的な推論が必要となる．そのため，前後の文脈や作業者の主観によって必然的に解釈の揺らぎが生じる．その結果，作業者が持つ一般的な語彙の理解と，専門家が想定する厳密な定義との間にギャップが生じやすく，その解釈が作業者の主観に委ねられることで判断基準の齟齬が発生しやすい性質を持っている．さらに，「物語型」の記事構造は，客観的な事実と主観的な感情や個人のエピソードが複雑に入り混じるため，多次元的な分類基準を適用する際の判断を一層困難にしていると考えられる．

そこで本研究では，従来の「逆三角形」構造の記事と，現代的な「物語（ナラティブ）」構造の記事の双方に対してアノテーションを実施し，評価の一貫性を保つことが難しい原因を比較・調査した．その検証にあたり，複雑な分類作業を小さな判断ステップに順次分解し，「記事のどの記述を根拠としてその判断を下したのか」という思考プロセスを記録・可視化するアノテーション支援ツールを開発した．このツールを用いて分類作業の実験を行うことで，最終的なラベルのみのアノテーションではわからない「判断のどの段階で」「記事のどのような表現によって」判断の齟齬が生じたのかを詳細に特定することができる．これにより，分類の一貫性を阻害する要因が，作業者の単なる能力不足によるものなのか，専門家が持つ暗黙知と言語化された定義とのズレにあるのか，あるいは現代のネット記事が持つ特有の記述構造との不適合にあるのかを多角的に調査し，より信頼性の高い分類枠組みを構築するための知見を提示する．

2 関連研究

複数の作業者がアノテーションタスクを行う際の一致性は，質的研究や内容分析における関心課題の一つである．Lombard ら [5] は，アノテーションタスクにおいて作業者間で生じる不一致の要因として「基準の曖昧さ」「コードの定義の不十分さ」「作業者の訓練不足」を挙げている．また，Guest ら [3] は，作業仕様書の継続的更新や訓練手続きの明示化が一致率向上に寄与することを示す一方で，作業者にかかる人的負担の大きさも指摘している．

こうした手作業の制約を克服するための支援技術の必要性の高まりに伴い，アノテーションタスクを大規模言語モデル（Large Language Model; LLM）を用いて行うことも検討されている [7]．一部のタスクでは良好な結果が報告されているが，専門的なアノテーション付与タスクを LLM で代替することには課題が残る．Ziems ら [9] は，25 の代表的なタスクを用いて LLM の性能を評価し，その限界を指摘している．これらのタスクは日常的な用語に専門的，あるいは非標準的な定義を適用する必要があり，LLM が事前学習で獲得した一般的な意味論とは異なる非慣習的な言語理解が求められるためである．特定の理論的枠組みに基づいた文脈依存的なニュアンスを正確に捉えることは依然として計算機には難しく，信頼性の高いデータセットを構築するには人間の解釈に基づくアノテーション付与が不可欠である．

また，アノテーションにおける作業者間の齟齬は，排除すべきノイズとして従来は扱われてきたが，Aroyo ら [1] は，その齟齬をタスクの曖昧さや作業者の多様な視点を反映した重要な手がかりであると指摘している．同様に，Cabitza ら [2] は機械学習用のデータセット構築に関わるアノテーション作業において，齟齬を排除するのではなく，判断の多様性を保持し，複数の視点を正解データの構築プロセスに統合するプロセスを提案している．このアプローチを採ることは，予測能力の向上だけでなく，モデルの解釈可能性や公平性向上にも貢献する可能性があるとしている．

3 タスクの定義と判断指標

本研究では，Reinemann らの枠組み [6] をもとに大森が翻訳・整理した指標 [11] を，アノテーションタスクを実施する際の判断基準とした．以下に，大森が提示する各側面の判断指標を示す．

トピック: ニュース項目の政治的関連性

1. 2 つ以上の政治的アクターが登場するか
2. 立法・行政・司法といった意思決定機関が登場するか
3. ニュースの取り上げる主題・問題に対し実現された政策的決定や措置プログラムに言及するか
4. ニュースの取り上げる主題・問題に対し実現された政策的決定や措置プログラムに関係する個人やグループが登場するか

フォーカス: ニュース項目の焦点に着目した分類

1. **「個人ー社会」との関連度 (F1)**：ニュースの内容の帰結が，個人の生活などのミクロな範囲に関連するものか，社会全体の問題に関連するものか

2. **「エピソードーテーマ」フレーム度 (F2)**：ニュース内容が，特定の個人のエピソードに着目するものか，より広範な問題のテーマに着目するものか

スタイル: ニュース報道で用いられる様式

1. **「個人ー非個人」的なリポート度 (S1)**：フォーカス面の「個人ー社会」関連度とは異なり，ニュースにリポーターや解説者，あるいはゲストといった個人の見解が含まれているか
2. **「感情ー非感情」的なリポート度 (S2)**：従来の戦略型フレーム報道やソフトニュースの研究でたびたび注目されてきた，戦いに関連するような言葉や表現を用いているか

本研究では，解釈の揺らぎが生じやすい「フォーカス」と「スタイル」の 2 次元（4 項目）に焦点を当ててアノテーションタスクを設計する．

4 実験

本稿では，判断の齟齬が生じる箇所を詳細に記録可能なアノテーション支援ツールを用いてアノテーション付与の過程を記録し分析する．

4.1 アノテーション支援ツール

本実験で利用するアノテーション支援ツールは，アノテーションタスクの遂行に当たり，作業者の認知プロセスを支援しつつ，作業者間の判断の齟齬が生じる箇所を詳細に把握することを目的としたツールである [13]．一般的なアノテーションでは，作業者の最終ラベルのみが記録されるため，どのような齟齬が生じたのかは把握できても，それがどのような判断や手続きによって生じたのかを特定することが難しい．この課題に対し，本ツールは，作業者が判断の根拠となったテキスト箇所の選択や理由記述を通じて，齟齬の原因と発生箇所の特定を紐づけて捉えられるようにしている．

図 1 に提案ツールのインタフェースを示す．このツールは，記事閲覧エリア（図 1 上部）と作業・判断エリア（図 1 下部）の 2 つから構成されている．記事閲覧エリアには，対象テキストが表示される．作業者はテキスト内の根拠箇所をクリックすることで，ハイライト（青/赤）を入れることができる．また，テキストの区切りが不適切な場合は，動的に文を分割することが可能である．

作業・判断エリアには，現在の作業工程における設問と操作パネルが表示される．複雑な仕様書を一度に提示するのではなく，「主題の特定」→「根拠のハイライト」

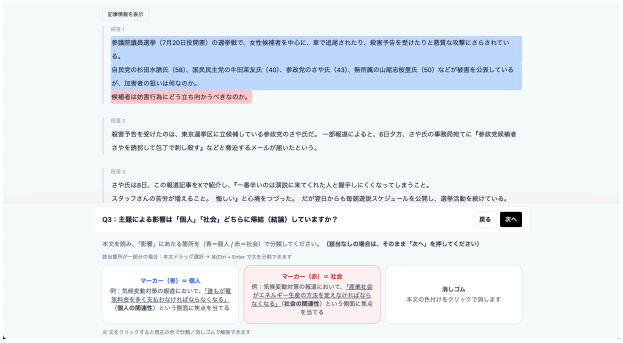

図 1: 提案ツールのインタフェース

→「全体判定」と段階的にタスクを提示することで，作業者の認知的負荷を低減することを企図している．

本ツールでは，これら各ステップでの操作（選択肢，ハイライト箇所，記述内容）および滞在時間をすべてプロセスログとして記録する．このログ機能により，最終ラベルと判断根拠がデータとして保存される．同時に，蓄積されたログは仕様書改善のサイクルを回すための客観的な分析資源として機能し，作業者の迷いや解釈の齟齬を定量的に特定することを可能にする．

4.2 実験準備

アノテーションを付与する対象の記事は，2025 年 7 月 14 日に Yahoo!ニュースに掲載されていた記事の中から，情報の効率的な伝達を目的とした従来の「逆三角形」構造である「ブラジルのペルアスー洞窟国立公園，ユネスコの世界自然遺産に登録される」[*1]（以下，ユネスコ記事）と，「物語」構造である「参政党・さや氏や山尾志桜里氏には殺害予告，国民民主党・牛田氏は車で追尾され… 女性候補への攻撃が相次ぐ理由」[*2]（以下，選挙妨害記事）の 2 記事を用いた．

本実験におけるアノテーションの判断基準は，第 3 章で述べた大森の指標を採用したが，これらは抽象的なアノテーション指針であり，ツール上の具体的な操作手順へ落とし込む必要があった．そこで，新聞記者経験のあるニュース研究者（本稿第 3 著者）の監修の下，ニュース記事 85 件を用いたアノテーションを実施して判断基準の具体化と合意形成を行い，本実験で使用する 2 記事に対する標準解を策定した．具体化された手順を表 1 に示す設問構成として本ツールに実装した．実験参加者は，画面上に順次表示される設問に回答することでアノテーションを進める．

[*1] https://news.yahoo.co.jp/articles/7bdc90774000336a5b4a86aa66fcd65c1ca02e60 (2025/7/14 確認).
[*2] https://news.yahoo.co.jp/articles/8fc4366e34ae12f3611a5eb9acde2116655d20c3 (2025/7/14 確認).

表 1: 大森の基準 [11] に基づき設計された設問

No.	対象側面	設問・タスク内容
Q1	前提	記者が記事で伝えたい「主題」を記述する
Q2	前提	記事が「速報・天気予報」であるか判定する
Q3	**F1**	主題による影響が「個人 (青)」「社会 (赤)」どちらに帰結するか抽出し，全体の傾向を判定する
Q4	前提	記事は「社会的な問題」を扱ったものか判定する
Q5	**F2**	社会的問題の語られ方を「エピソード (青)」「テーマ (赤)」で抽出する
Q6	**F2**	抽出箇所に基づき，記事全体のフレーム (エピソード型/テーマ型) を 3 段階で判定する
Q7	**S1**	リポーター等の「個人の見解」が含まれる文と，その根拠語句を抽出する
Q8	**S1**	抽出箇所に基づき，記事全体のスタイル (個人的/非個人的) を 3 段階で判定する
Q9	**S2**	感情を刺激する表現や，戦いに関連する語句を抽出する
Q10	**S2**	抽出語句に基づき，記事全体のスタイル (感情的/非感情的) を 3 段階で判定する

4.3 評価方法

本実験では，異なる記述構造を持つニュース記事間で，分類枠組みの適用可能性を比較・評価するため，以下のデータ処理および判定を行った．

1. **評価単位**：意味的に完結した一文を最小の評価単位と規定した．これに基づき，システム上で不自然に分割または結合されて提示された箇所については，本来の文構造に合わせて再構成（結合・分割）を行い，評価単位としての整合性を担保した．
2. **ラベル付与の判断基準**：再定義された文に対して，作業者が分割操作等により複数の箇所でラベルを付与していた場合，いずれか一箇所でも付与されていれば，その文に対してラベルが付与されたものとした．

評価指標の選定においては，ニュース記事の特性を考慮した．ニュース記事の 1 文には，「個人の帰結/エピソード」と「社会の帰結/テーマ」の双方が含まれる可能性があるため，これらをどちらが支配的かという排他的な基準で評価することは，情報の欠落を招く恐れがある．そこで本研究では，各カテゴリ（F1 における個人/社会，F2 におけるエピソード/テーマ）の独立性を担保するため，各々について選択されたか否かを個別に評価する 2 値分類のアプローチを採用した．これに基づき，分析の目的と粒度に応じて以下の指標を用いた．

単純一致率（文単位）
作業者間または標準解との間で，ラベルを付与するか否かの判断が一致した割合（κ 係数算出における観測一致率と同義）を算出する．本研究では，文単位での単純一致率をプロセスログと照合することで，判定のパターンや迷いやすい文を抽出する質的分析の手がかりとしても利用した．

κ 係数 （記事単位）
記事内の全単位に対する判断列（ベクトル）を入力とし信頼性係数 κ を算出する．なお，複数属性（青・赤）を統合した全体の評価においては，各色の判定ベクトルを結合した長さ $2N$ のベクトルを対象として κ を算出した．

4.4 実験手続き

アノテーションタスクにおいて生じる判断の齟齬が，単なる「作業手順や定義の曖昧さ」に起因するものか，あるいは「現代のニュース記事が持つ構造的複雑さと分類枠組みの本質的な不適合」に起因するものかを分離して検証するため，2 段階の実験を設計した．まず 1 段階目の実験（実験 1）では，初期状態の仕様書を用いてアノテーションを実施し，提案ツールのプロセスログから齟齬が生じた箇所を特定する．続いて 2 段階目の実験（実験 2）では，実験 1 で明らかになった定義の曖昧さや認知的なバイアスを排除するよう仕様書を修正し，再度アノテーションを実施する．なお，両実験ともに被験者は大学生（実験 1: 9 名, 実験 2: 7 名）とした．各実験において，参加者はツールの事前説明を受けた後，対象記事に対してアノテーションを行うよう指示された．

4.5 実験 1：結果

記事全体での評価を問う設問 (Q2, Q4, Q6, Q8, Q10) における一致率を表 2 に示す．Q2（速報性）や Q4（社会性）において，選挙妨害記事では概ね高い一致を示したが，ユネスコ記事では判断が割れた．Q8/Q10 においては，作業者間一致率は高い一方で，標準解との一致率がほぼゼロであるという乖離が見られた．

表 2: 実験 1：記事全体に対する設問の回答一致率

No.	記事	対 標準解 (正答率)	作業者間 (一致率)
Q2	選挙妨害記事	1.000	1.000
	ユネスコ記事	0.556	0.444
Q4	選挙妨害記事	0.889	0.778
	ユネスコ記事	0.556	0.444
Q6	選挙妨害記事	0.500	0.429
	ユネスコ記事	0.200	0.600
Q8	選挙妨害記事	0.556	0.361
	ユネスコ記事	0.000	0.778
Q10	選挙妨害記事	0.333	0.278
	ユネスコ記事	0.111	0.778

ログおよびアンケート分析に基づき，以下の設計上の課題を特定し，実験 2 へ向けた修正を行った．

- **主題特定の制約強化（Q1）**
  Q1 では「記者が記事で伝えたい主題を説明する」という自由記述形式を採用していたが，抽象度や記述の粒度に作業者間のばらつきが大きく，主題特定そのものが一致率の低下要因となっていた．そこで実験 2 では，ニュース記事に一般的に採用される逆三角形構造[*3]を踏まえ，「第一段落（リード文）から主題を抜き出す」ことを明示的な制約としてインタフェース上に実装した．この制約により，主題特定が作業者の読解力や要約方略に依存しにくい条件を整えることを意図した．さらに，第 1 段落に記事の主題が十分に含まれていない場合には，作業者の誤りとはみなさず，「該当なし」として Q1 の回答を省略できるよう設計した．これにより，本ツールはアノテーション仕様書や作業者の判断だけでなく，ニュース記事自体の構造的な良し悪しを評価する補助的な指標としても機能することが期待される．
- **フィルタリング設問の撤廃（Q2, Q4）**
  「速報性」や「社会性」の判定は，比較対象に依存する相対的なものであり，Q2, Q4 のような二値分類フィルタとして実装するには不適切であることが判明した．曖昧な基準によるフィルタリング設問で後続の分析データが欠損することを防ぐため，実験 2 ではこれらの事前フィルタを廃止し，全データを主要な分析フローへ回す方針とした．

第 3 章で述べた各側面（F1, F2, S1, S2）について，本文中の根拠箇所の抽出結果に基づき，記事全体としての傾向がどれだけ一致していたかを検証する．実験 1 における標準解との一致率，および，作業者間一致率を表 3, 表 4 に示す．全体として一致率は低調であり，特に選挙妨害記事の F1（社会）は 0.093，ユネスコの F2（全体）は 0.040 と，統計的に偶然レベルの一致に留まる項目が散見された．特定された要因に基づき，実験 2 に向けた仕様の修正を行った．

表 3: ユネスコ記事における一致率（実験 1・2 比較）

		実験 1		実験 2	
対象側面	評価ラベル	対 標準解	作業者間	対 標準解	作業者間
F1	個人	0.778	0.580	0.857	0.714
	社会	0.085	0.203	0.301	0.191
	全体	0.058	0.199	0.311	0.197
F2	エピソード	0.139	0.282	-0.247	0.253
	テーマ	0.139	0.171	-0.247	0.253
	全体	0.040	0.126	-0.584	0.290
S1	該当あり	0.111	0.469	0.429	0.352
S2	該当あり	0.066	0.582	-0.007	0.714

表 4: 選挙妨害記事における一致率（実験 1・2 比較）

		実験 1		実験 2	
対象側面	評価ラベル	対 標準解	作業者間	対 標準解	作業者間
F1	個人	0.111	0.016	0.143	0.460
	社会	0.093	0.055	0.011	0.487
	全体	0.050	0.027	-0.003	0.465
F2	エピソード	0.591	0.493	0.045	0.156
	テーマ	0.591	0.493	0.045	0.156
	全体	0.602	0.504	0.173	0.363
S1	該当あり	0.313	0.411	0.209	0.337
S2	該当あり	0.099	0.102	-0.003	0.292

- **F1：用語の認知的な齟齬**
  選挙妨害記事において，記事冒頭の問いかけを社会への帰結として誤ってマーキングする傾向がログから確認された．これは「帰結」という専門用語が，作業者に直感的に理解されていないことを示唆している．
- **F2：定義の境界における迷い**
  「テーマ」の定義に含まれる「専門家の解説」という記述に引きずられ，個人的なエピソードであっても専門家が登場するだけで「テーマ」と分類してしまう誤りが多発した．
- **S2: 対象とスタイルの混同**
  悲惨な事件（殺人予告など）の事実記述に対し，記者の表現自体は中立であるにも関わらず「感情的」と判定するケースが見られた．これは「出来事の性質」と「記事のスタイル」の切り分けが仕様書上で不明確であったことに起因する．

これらの分析に基づき，実験 2 では仕様書の用語変更や，注釈の追加を行った．

[*3] http://www.at-s.com/blogs/nie/study/howto.html（2025/12/12 確認）．

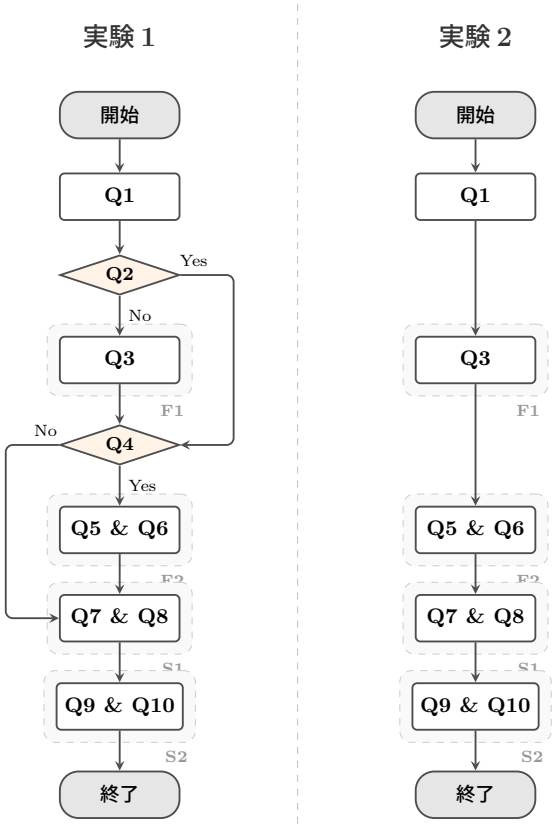

図 2: 実験フローの比較

4.6 実験 2：結果

実験 1 で特定された課題に対し，仕様書およびツール上の注釈を修正するとともに，判断フローを図 2（右）のように修正して実施した実験 2 の結果を比較する．

表 3 および表 4 に示した通り，一部の項目で一致率の向上が確認された一方，F2 や S2 では悪化も確認された．

- **作業者間の一致率の向上**：
  改善が確認された項目として，実験 1 で偶然レベル（$\kappa = 0.027$）と判定された選挙妨害記事の F1（全体）がある．実験 2 ではこれが $\kappa = 0.465$ まで改善が見られた．これは，実験 1 の分析で明らかになった「問い・事実」と「帰結」の混同に対し，「帰結」を「結論」と言い換えたことが，作業者の解釈を統一する上で有効性が示唆された．また，ユネスコの S2（感情）においても，作業者間一致率は 0.582 から 0.714 へと向上した．「出来事の悲惨さと，表現の感情的スタイルを区別する」という注釈を追加したことで，作業者間で判断基準が共有されやすくなったと考えられる．
- **標準解との一致率の向上**：
  ユネスコの F1 においては，「個人」ラベルの正答率が 0.778 から 0.857 へ，「社会」ラベルが 0.085 から 0.301 へと向上し，改善が見られた．一方で，作業者間の一致率は向上したものの，標準解との一致率が必ずしも連動して向上しないケース（例：選挙妨害記事の F1 やユネスコの F2 など）も確認された．これは，仕様書の改善によって「作業者集団の中での解釈」は収束したものの，その解釈が標準解作成者の意図とは異なる方向で収束した可能性を示唆している．

5 考察

実験結果が示すように，アノテーションの不一致は単なるランダムな誤りではなく，仕様書の定義，記事構造，作業者の解釈といった複数の要因が絡んだ構造的問題として発生している．本節では，得られた知見を「記事構造による差異」「標準解との乖離」の 2 点から考察する．

5.1 記事構造による差異

Q1 において「第一段落（リード文）から主題を抽出する」という制約を導入したことで，主題抽出の判断基準が記事構造に明示的に紐づけられ，作業者の読解力や要約能力の差異による影響を大幅に低減できた．実験 2 ではユネスコ記事において全ての作業者がほぼ同一の主題を抽出し，実験 1 で顕著であった主題記述の長さや抽象度に関する記述ゆれが解消された．逆三角形構造のニュース記事では主題が第一段落に集約されるため，この操作は本来，作業者間の一致を高める効果を持つ．しかし，選挙妨害記事においては作業者の回答が三つの小グループに分かれていた．自由記述ではなく第一段落からの抜き出しという操作に統一されているにもかかわらずこの分裂が生じたことは，作業者の読解力や要約能力の差異ではなく，リード文自体が複数の主題候補を含む構造になっていることを示唆する．

客観的な事実伝達を基本とする「逆三角形」構造のユネスコ記事では，スタイル面の作業者間一致率が比較的高かった．対照的に，「物語」構造を持つ選挙妨害記事においては，同項目の一致率が 0.102 に留まった．悲惨な出来事を扱う記事では，事実の描写そのものが強い感情的な文脈を帯びて読者に提示される．そのため，作業者が「客観的な事実の描写」と「記者の感情的な執筆スタイル」を明確に分離できず，結果としてアノテーションの判断が大きく分散したと考えられる．

5.2 作業者間の一致と標準解との乖離

実験 2 の結果において，仕様書の改善により作業者間の一致率は向上したものの，標準解との一致率は必ずしも向上しない，あるいは低下する事例が確認された．作業者同士の解釈は統一されたが，専門家の意図とは異なる方向へ収束したという現象は，アノテーションの品質管理において重要な示唆を与える．

この乖離が生じた主な要因として，標準解作成者が有する「暗黙知」が仕様書に十分に言語化されていなかった点が挙げられる．専門家が判断の際に用いた文脈依存的なニュアンスや判断基準が明文化されていなかったため，仕様書の記述を忠実に参照した作業者との間に認識の齟齬が生じたと考えられる．加えて，標準解そのものが，専門家の無意識のバイアスや感覚的な判断によって，本来定められた定義から逸脱していた可能性もある．

従来のアノテーション手法では，最終的なラベルのみを比較するため，この乖離が「作業者の理解不足」によるものなのか，「標準解の妥当性の欠如」によるものなのかを判別することは困難であった．しかし，本研究で提案したツールは，判断の根拠となるテキスト箇所を記録している．これにより，標準解作成者は，作業者が「なぜその選択をしたのか」という思考プロセスを追跡することが可能となる．

もし，作業者が仕様書の記述を忠実に守った結果として標準解と異なる判断を下しているのであれば，修正すべきは「作業者の判断」ではなく，「仕様書の記述」あるいは「標準解そのもの」であると判定できる．すなわち，本ツールは単に作業者を正解に導くだけでなく，標準解作成者自身が「自分の判断を見直すべきか，仕様書を修正すべきか」を客観的に判定するためのデバッグツールとしても機能する．これは，専門家の意図を仕様書やツール上でどう明確に伝えるかという設計プロセスにおいて，不可欠な機能であると言える．

6 おわりに

本研究では，ニュース報道の多次元的なハード・ソフト区分において分類基準の一貫性を保つことが困難であるという課題に対し，判断の段階的な分解と根拠箇所の記録・可視化を可能にするアノテーション支援ツールを開発し，分類の齟齬が生じる原因を調査した．

ツールから得られたプロセスログと検証実験の分析により，アノテーションにおける判断の不一致の主因が，作業者の単なる能力不足によるものではないことが確認された．専門家が持つ暗黙知と言語化されたマニュアル定義との間に生じるズレや，現代のネット記事が持つ特有の記述構造（物語性）との不適合が，評価の一貫性を阻害する構造的な要因であることを明らかにした．

特に，読者の共感を引き出すために客観的な社会課題の背後に個人の心情やエピソードを意図的に混在させる現代のネット記事（物語構造）では，客観的な事実描写と感情的な表現が融合していると考えられる．

謝辞

本研究は JST RISTEX (課題番号 JPMJRS23L2）の支援を受けた．また，本研究の実施にあたり森野穣氏から示唆を受けた．記して謝意を表す．

参考文献

[1] 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, DOI: 10.1609/aimag.v36i1.2564 (2015).

[2] Cabitza, F., Campagner, A. and Basile, V.: Toward a Perspectivist Turn in Ground Truthing for Predictive Computing, *Proc. AAAI Conf. on Artificial Intelligence*, Vol. 37, No. 6, pp. 6859–6867, DOI: 10.1609/aaai.v37i6.25840 (2023).

[3] Guest, G., MacQueen, K. M. and Namey, E. E.: *Applied Thematic Analysis*, Sage publications (2011).

[4] Krippendorff, K.: *Content Analysis: An Introduction to Its Methodology*, Sage Publications (2004).

[5] 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, DOI: 10.1111/j.1468-2958.2002.tb00826.x (2006).

[6] 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, DOI: 10.1177/1464884911427803 (2012).

[7] Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L. and Liu, H.: Large Language Models for Data Annotation and Synthesis: A Survey, *Proc. 2024 Conf. on Empirical Methods in Natural Language Process-*

ing, pp. 930–957, DOI: 10.18653/v1/2024.emnlp-main.54 (2024).

[8] Tuchman, G.: Making news by doing work: Routinizing the unexpected, *American Journal of Sociology*, Vol. 79, No. 1, pp. 110–131 (1973).

[9] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z. and Yang, D.: Can Large Language Models Transform Computational Social Science?, *Computational Linguistics*, Vol. 50, No. 1, pp. 90–138, DOI: 10.1162/coli_a_00502 (2024).

[10] 石戸諭: ニュースの未来, 光文社 (2021).

[11] 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).

[12] 斉藤友彦: 新聞記者がネット記事をバズらせるために考えたこと, 集英社 (2025).

[13] 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).