Vol.2026-HCI-216 No.18
2026/1/14

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

杉本 麻衣[1,a)]　松下 光範[1,b)]　藤代 裕之[2,c)]

概要：感情意図ラベル付与タスクのように，仕様書に基づいて複数の作業者がアノテーションを行う場合，仕様書の曖昧さにより作業者間で判断基準に齟齬が生じることがある．このような場合には，齟齬が発生する曖昧な箇所を特定し，仕様書を継続的に改善していく必要がある．しかし，仕様書のどの手続きに，どのような不具合が生じているかを把握するには，最終的なアノテーション結果の突合だけでは不十分である．そこで本研究では，アノテーションタスクを段階的に実施することを前提としたアノテーション支援ツールを提案する．提案システムでは，仕様書の判断手続きを工程ごとに分割し，それを参照しながら段階的に作業を進める．さらに，各段階で判断の根拠を記録することで，アノテーション間の齟齬が生じた際に，その原因を容易に特定できる．

1. はじめに

社会学をはじめとする人文科学の研究領域では，新聞記事や政策文書，識者の論考など，多様なテキスト資料を対象としてその記述内容を分析し，その背後にある構造や意味を明らかにしようとしてきた．アノテーション（コーディング）は，こうした分析を行うために，テキストにラベルやカテゴリの情報を付与する作業である．しかし，アノテーションタスクは，テキストから，前後の文脈や背景等を踏まえて「意味」を読み取る解釈的な行為であり，作業者の主観によって意味解釈のゆらぎが生じ得る [4]．特に，特定の学問分野の理論的構成概念に基づくアノテーションは，日常的な用語に対して専門的で非標準的な定義を適用する作業であり，その解釈も作業者の主観に依ることから，判断基準の齟齬が生じやすいという性質を持つ [9]．

こうした主観性に起因する齟齬を最小化するために，詳細なアノテーション仕様書を作成し，判断基準の標準化が試みられてきた．しかし，Aroyo ら [1] は，仕様書の詳細化によって齟齬を無理に解消しようとすることは，解釈の曖昧さといった重要な手がかりを排除してしまう可能性があると指摘している．

この手がかりを適切に扱うためには，その解釈の齟齬が「多様な解釈が許容されるべき曖昧さ」によるものなのか，「仕様書の曖昧さ」によるものなのかを判別する必要がある．従来の一致率による評定方法では，作業者間の結果に生じた齟齬は確認できても，仕様書のどの手続きに，どのような不具合が生じているかを特定することは困難である．齟齬の原因を解明し，仕様書に必要な修正を反映するには，結果だけでなく，そこに至る判断手続きを詳細に記録・分析できる仕組みが必要である．

こうした齟齬の原因特定と仕様書の修正は，単発の対応ではなく，継続的なサイクルとして実施していく必要がある．Pustejovsky ら [6] は，アノテーション開発の初期段階において，仕様書修正と試験的なアノテーションを反復する MAMA サイクル（Model-Annotate-Model-Annotate）を提唱している．これは，実際のデータに対する作業者の判断と，その結果に基づく定義の修正を繰り返すことで，仕様書の不備や曖昧さを解消するプロセスである．このことから，アノテーション仕様書は一度作成すれば変わらない静的なものではなく，齟齬を通じて更新され続ける動的なものとして捉え直す必要がある．

こうした背景の下，本研究では作業者間の齟齬を仕様書改善の手がかりとして記録・分析できるアノテーション支援ツールを提案する．提案ツールでは，仕様書の判断基準を工程ごとに分割し．作業者がそれに沿って段階的に判断を行えるようにする．さらに，各段階で判断の根拠を記録することで，仕様書のどの手続きにどのような不具合が生じているかを特定することを可能にする．

[1] 関西大学
[2] 法政大学
[a)] k570068@kansai-u.ac.jp
[b)] m_mat@kansai-u.ac.jp
[c)] fujisiro@hosei.ac.jp

2. 関連研究

2.1 アノテーションタスクにおける作業者間の齟齬

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の不一致が生じる要因として「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している．しかし，アノテーションコード体系の複雑化やテキスト量の増大によって人的負担が大きくなることが指摘されており，支援技術の必要性が高まっている．

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 [8]．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，意見の多様性を保持し，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．

2.2 アノテーションタスクの効率化支援

人間によるアノテーションの認知的負荷を低減し，タスク効率と一貫性を向上させるためのアプローチとして，ツールのインタフェース設計によるタスクの単純化が挙げられる．Prodigy[*1]は，自然言語処理から画像・音声まで幅広いアノテーションを対象とする拡張性の高いアノテーションツールであり，モデルによる事前予測に対して作業者が応答する形でアノテーションを進める人間参加型の設計思想を採っている．これは，複雑なアノテーションタスクを小さな意思決定単位に分解し，必要に応じて Yes/No 判断や少数の候補選択によって効率的にラベル付けを行わせることで，判断の迷いを軽減し，速度と一貫性の向上を図っている．しかし，Prodigy が重視するのは「効率的に正解ラベルを収集する」ことであり，作業過程で生じる迷いや意見の分岐点は排除される設計となっている．一方，本研究は，判断の迷いそのものを排除すべきノイズではなく，「どこで・どのように意見が分岐したか」を理解するための重要な情報として扱う．段階的なタスク構造も，意思決定を単純化するためではなく，意見の分岐点を抽出・可視化し，仕様改善や合意形成のための議論材料として活用することを目的とする．

[*1] https://prodi.gy/ （最終アクセス：2025 年 11 月 25 日）．

3. 対象とする課題

本研究では，アノテーションタスクの実践例として，ニュース報道を「ハードニュース」と「ソフトニュース」に分類する枠組みを扱う．これはニュース研究において広く参照される区分であり，報道内容の「伝え方」の違いを捉える概念として位置づけられている．この区分の分類基準は，コミュニケーション研究領域において長年議論されてきたが，その定義や測定方法は研究者によって異なり，概念的な曖昧さが指摘されてきた．従来の研究は，ニュースをトピック（政治・経済か，芸能・スポーツか）のみで分類する一次元的なアプローチに固執する傾向があったが，Reinemann ら [7] は，トピック・フォーカス・スタイルという 3 つの側面を組み合わせた多次元的なアプローチを提案している．この枠組みは，ニュースを単一の基準で二分するのではなく，各側面の強度の組み合わせによって，ニュース記事をハードからソフトへの連続的な尺度の中に位置づけることを企図している．

本研究が対象とするのは，この Reinemann らが提案した多次元的な分類枠組みである．この枠組みを，アノテーションタスクの実践例として対象とした理由は，日常的な用語に対して特定の理論的構成概念に基づく専門的な定義を適用し，文脈依存的な意味を読み取るという高度で非慣習的な判断が求められることから，その解釈が作業者の主観に委ねられやすく，判断基準の齟齬が生じやすいという性質を持つためである．本研究では，Reinemann らの枠組みを大森が翻訳整理した指標 [10] 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–社会」との関連度 **(F1)**：そのニュースの内容の帰結が，個人の生活等ミクロな範囲に関連するものであるのか，それとも社会全体の問題に関連するものであるのか

(2)「エピソードーテーマ」フレーム度 **(F2)**：ニュースストーリー自体が，特定の個人のエピソードに着目するものか，それともより広範に問題のテーマに着目するものであるのか

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，個人の見解が含まれているか，それとも事実に基づいた記述か

(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注目されてきた，戦いに関連するような言葉や感情的な表現を用いているか

4. アノテーション支援ツールの設計と実装

4.1 デザイン指針

アノテーションタスクの実施において，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特定可能にするため，以下の 3 点を設計指針として定めた．

(1) **判断手続きの段階的提示により，作業者の認知負荷を低減すること**

アノテーション仕様書を長文のまま一括で提示すると，作業者に高い認知的負荷がかかり，解釈や判断の齟齬が生じやすくなる．そこで本ツールでは，複雑な仕様書を一度に参照させるのではなく，判断に必要な手続きを分岐構造として再構成し，インタフェース上ではこれに沿って設問を順次提示することで，作業者の認知的負荷を低減させる設計とした．

(2) **最終ラベルだけでなく，その根拠となる箇所の記録を可能にすること**

従来の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，どの判断手続きで，どんな齟齬が生じたのかを特定することが難しい．この課題に対し，各工程において，作業者が判断の根拠となったテキスト箇所を選択，または理由を記述することで，齟齬の原因と発生箇所の特定を可能にする仕組みとした．

(3) **プロセスログを用いて齟齬の発生箇所を特定し，仕様書の改善を支援すること**

解釈の難しい社会学的なアノテーションでは，初期段階で完全な仕様書を作成することは難しく，運用過程で生じた齟齬に基づいて仕様を更新する仕組みが必要である．そこで，収集されたプロセスログを分析することで，齟齬が発生した箇所を特定できる．特定の箇所で多くの作業者が迷ったり，判断が割れたりしている場合，その設問に対応する仕様書の記述が曖昧であるか，不十分であることを示唆する．この手がかりをもとに，仕様書の該当箇所を修正・追記することで，継続的な改善サイクルを効率的に回せるよう支援する設計とした．

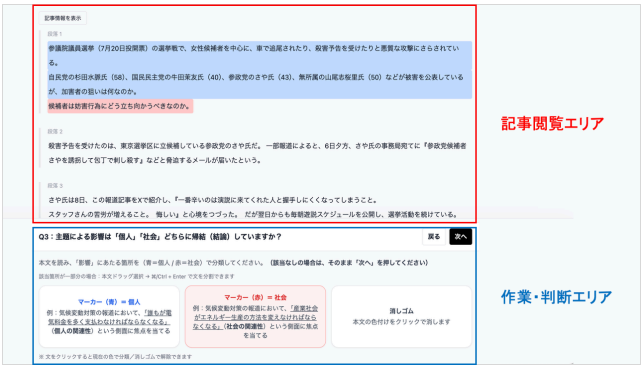

図 1 提案ツールのインタフェース

4.2 ユーザインタフェースと機能

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，画面を上下 2 つのエリアから構成されており，上記の課題 (1)〜(3) を解決するため，以下の機能を備えている．

(1) **記事閲覧エリア（図 1 上部）**：画面上部に配置され，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．

(2) **作業・判断エリア（図 1 下部）**：画面下部に固定され，現在の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」→「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させる設計とした．

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．

5. 実験

提案したツールを用いることで，アノテーションタスクにおける齟齬の発生箇所と原因を特定し，その知見を仕様

改善へと還元を支援できるかについての検証実験を行った．

5.1 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「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」（以下，ユネスコ記事）および「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」（以下，選挙妨害記事記事）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．

5.2 評価方法

本実験では，評価の厳密性と再現性を担保するため，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，それぞれ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．このアプローチに基づき，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を算出する．本研究では，この単純一致率を，後述する κ 係数における観測一致率に対応する指標として用いるとともに，実装上は文単位での，作業者ごとの一致率を算出し，プロセスログと照合することで，誤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，偶然による一致を補正した信頼性係数 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．

5.3 実験手続き

アノテーションタスクにおける齟齬の発生箇所と原因を特定し，その知見を仕様改善への還元を支援できるかを検証するため，2 段階の実験を計画した．まず 1 段階目の実験（実験 1）では，初期状態の仕様書と判断フローを用いてアノテーションを実施し，提案ツールによって記録されたプロセスログを分析することで，齟齬がどの文・どの判断工程で生じたかを抽出した．続いて 2 段階目の実験（実験 2）では，実験 1 で明らかになった課題に基づき仕様書および判断フローを改修し，第 5.2 節で定義した 2 値の κ 係数（記事単位）および単純一致率（文単位）を用いて一致率の変化を評価した．

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．

5.4 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．

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[*2]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

[*2] http://www.at-s.com/blogs/nie/study/howto.html （最終アクセス：2025 年 12 月 12 日）．

表 1　大森の基準 [10] 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

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	Senkyo	1.000	1.000
	Unesco	0.556	0.444
Q4	Senkyo	0.889	0.778
	Unesco	0.556	0.444
Q6	Senkyo	0.500	0.429
	Unesco	0.200	0.600
Q8	Senkyo	0.556	0.361
	Unesco	0.000	0.778
Q10	Senkyo	0.333	0.278
	Unesco	0.111	0.778

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

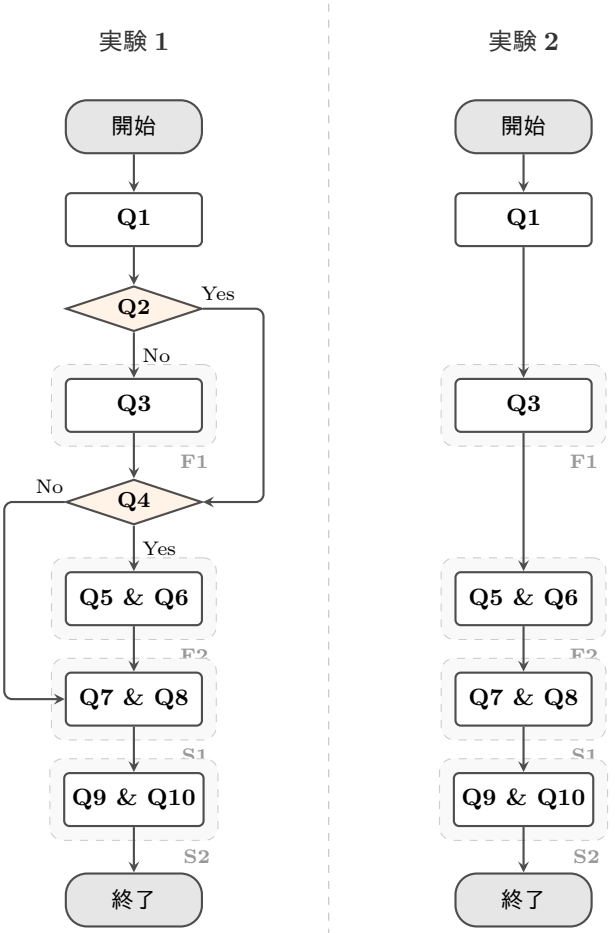

図 2　実験フローの比較

「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．

5.5　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，全体としても改善が見られた．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車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．

6.　考察

本研究で提案したアノテーション支援ツールは，従来の最終ラベルのみを対象とした分析では困難であった「齟齬の発生箇所と原因の分離」の可能性を示した．実験結果が示すように，アノテーションにおける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み合った構造的な問題として発生している．本節では，得られた知見をプロセスログの役割，標準解との乖離，仕様書運用の再定位およびツール使用感という 4 点から議論する．

6.1　プロセスログによる曖昧さの発見

従来のアノテーション研究では，不一致の原因の特定は困難であり，「どこで」「なぜ」齟齬が生じたのかはブラックボックスとなっていた．最終ラベルから仮説を立てること自体は可能であるが，その仮説をどの文のどの判断過程に紐づく問題として位置づけることは容易ではない．提案ツールは，Q1〜Q10 の段階化された判断手続きと根拠ハイライトの記録を通じて，不一致の発生箇所を文単位で特定することの可能性を示した．特に，F1 における「帰結/問い・事実」の混同，S2 における「対象とスタイル」の概念の混同といった例は，最終ラベルや自由記述だけからでは，推測にとどまっていたものを，具体的な文と判断工程に対応づけてピンポイントに特定できた点は，本ツールによるプロセスログ可視化の効果だといえる．この知見は，提案ツールが単なる作業効率化支援のインタフェースではなく，仕様書のデバッグと改善を支援する設計環境として機能し得ることを示唆する．

6.2　作業者間の一致と標準解との乖離

実験 2 では，仕様書を改善した結果として作業者間の一致率は向上したが，標準解との一致率は必ずしも向上しなかった．この結果は，改善後の仕様書が作業者にとっては明確な指示として機能した一方で，標準解作成者の暗黙知とは異なる方向に解釈を安定化させていた可能性を示す．つまり，「高一致＝正確」とは限らず，一致率だけでは評価

できない齟齬が存在することを明らかにした．この点においても，プロセスログによる判断分岐の分析が，表面的な一致率指標だけでは捉えられない問題の発見に寄与した．

6.3 仕様書の動的運用とツールの役割

本研究の知見は，アノテーション仕様書を，作業者との相互作用を通じて更新され続ける動的な設計対象として捉え直す必要性を示唆している．本ツールは，初期段階での齟齬を明示化し，それを元に仕様書を継続的に改善していくための基盤として機能する．

また，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．以上より、本ツールはアノテーション仕様書の運用改善だけでなく，記事の構造的問題の検出にも寄与する可能性がある．

6.4 参加者によるツール使用感の評価

事後アンケート（n=16）の「ツールは使いやすかったか」という設問に対して，「非常にそう思う」（3 名），「ややそう思う」（10 名），「どちらともいえない」（2 名），「あまりそう思わない」（1 名）という回答が得られた．この結果は，本ツールが一定の受容性を持つことを示している．一方で，「判断が難しかった箇所とその理由」に関する自由記述では，以下のようなコメントが寄せられた：「読者に戦いや感情を訴えるような言葉がどのようなものか想像がしづらかったため，判断が難しかった」「個人か社会かの判断が難しかった」「テーマとエピソードの分類で迷った」しかし，これらは曖昧さの種類を列挙するに留まり，判断のどのステップで齟齬が生じたのか，原因が仕様書なのか文章構造なのかを特定するには至らなかった．このことは事後アンケート形式の自己報告のみでは，曖昧さの原因の精緻な特定には限界があることを示している．本ツールは，この限界を補完し，不一致の原因を具体的な判断手続きとして抽出可能にするための手段として有効であるといえる．

7. おわりに

本研究では，ニュース記事に対するアノテーションタスクにおいて，判断手続きを段階的に記録するツールを提案した．Q1〜Q10 の判断系列と根拠のログにより，最終ラベルのみでは把握できなかった齟齬の発生箇所を特定可能であることを示した．特に，仕様書改訂の初期段階においても，齟齬の原因を局所化しピンポイントで修正可能であるという示唆が得られた．本ツールは，仕様書改訂の後期フェーズにおける微細な齟齬の発見により大きく寄与すると期待される．

8. 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 (online), 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. The 37th AAAI Conference on Artificial Intelligence*, Vol. 37, No. 6, pp. 6859–6867 (online), 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* (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 (online), DOI: 10.1111/j.1468-2958.2002.tb00826.x (2006).

[6] Pustejovsky, J. and Stubbs, A.: *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*, " O'Reilly Media, Inc." (2012).

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

[8] 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, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 930–957 (online), DOI: 10.18653/v1/2024.emnlp-main.54 (2024).

[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 (online), DOI: 10.1162/coli_a_00502 (2024).

[10] 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).