イノベーティブな言語使用は集団的アイデンティティの指標になりうるか？

伊藤 和浩 [1] 矢田 竣太郎 [1,2] 若宮 翔子 [1] 荒牧 英治 [1]
[1] 奈良先端科学技術大学院大学 [2] 筑波大学
{kazuhiro-it,wakamiya,aramaki}@is.naist.jp yada@slis.tsukuba.ac.jp

概要

所属集団との結びつきについての感覚を表す集団的アイデンティティは，人々の認知や行動に影響を与えることが知られている．近年，オンラインコミュニティの集団的アイデンティティを推定する重要性が高まっている．既存の辞書を使った指標では，辞書メンテナンスのコストや，低頻度語が考慮できないなどの課題がある．そのため本稿では，語のイノベーティブな使用（Linguistic Innovation）の頻度が，集団的アイデンティティの指標になるという仮説を検証する．データは YouTube 上の動画へのコメントを用いた．結果は，対象とした 4 カテゴリのうち 3 つのみで仮説を支持し，指標としての機能にはコミュニティの性質による条件が示唆された．

1 はじめに

人は個人の独自性を示す個人的アイデンティティとは別に，所属集団との結びつきを表す**集団的アイデンティティ**を持つ [20]．集団的アイデンティティは，人々の認知や行動に影響を与えることがよく知られている [18, 22, 23]．近年，ソーシャルメディアの発展が顕著になるにつれ，オンラインコミュニティにおける集団的アイデンティティを調査する重要性が高まっている．

これを受け，大規模なテキストデータから集団的アイデンティティを辞書ベースで推定する手法が提案されている [2]．具体的には，Linguistic Inquiry and Word Count (LIWC) [15] を用いて次の 2 つの心理的概念をスコア化した．

親和（Affiliation） 集団内の他者を自己と同一視する感覚を表すもの．コミュニケーションを表す言葉（例：‘愛’，‘コミュニティ’，‘SNS’，‘われわれ’，など）が該当する．

認知プロセス（Cognitive processing） 所属集団に対する疑義を表すもの．認識や推論を表す言葉（例：‘なぜ’，‘らしい’，‘だけど’，‘つまり’，など）が該当する．

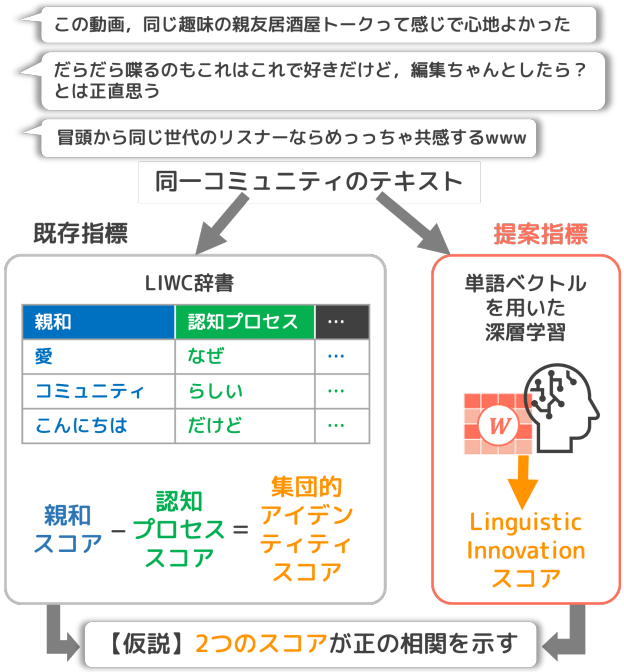

図 1: 本研究の概要．LIWC から集団的アイデンティティを推定した既存研究 [2] より汎用性の高い指標として，Linguistic Innovation スコアを提案する．

集団的アイデンティティは，親和スコアと正の相関を，認知プロセススコアと負の相関を示し，これらは社会心理学における知見とも矛盾しない [8, 20]．しかし，LIWC ベースの指標は，辞書のメンテナンスコストや，辞書に含まれない新語や低頻度語は考慮できないこと，また他言語対応が難しいことが課題である．そのため辞書などの高コストな言語リソースを必要としない指標が理想的である．

本研究では，心理言語学の知見に基づき，語彙のイノベーティブな使用（**Linguistic Innovation**）の頻度から，**集団的アイデンティティ**の強さが推定できるという仮説を検証する．Linguistic Innovation は，

心理言語学の分野において，円滑なコミュニケーションのためにコミュニティ独自の文脈に応じて生み出されるもの [4] とされている．例えば，オンラインのシューティングゲームのプレイヤーの間では，対戦相手から隠れ続ける振る舞いを「芋」「芋る」とネガティブなニュアンスを込めて呼ぶ．このような，コミュニティの文脈が独自であるほど発生しやすい Linguistic Innovation は，メンバーシップが強いほど頻度が高いと推論できる．このことから，Linguistic Innovation が集団的アイデンティティと正の相関を示し，よって代替指標として機能するという仮説を立てた（図 1）．

仮説を検証するための実験として，YouTube 上の動画へのコメントに対して LIWC ベースの集団的アイデンティティスコアと Linguistic Innovation の発生頻度を表すスコアを算出し，相関係数を調査した．

2 関連研究

Linguistic Innovation については，言語学や社会学の領域を中心に研究されてきた．Andersen ら [1] は，Linguistic Innovation を語義の変化 (change) ではなく語義が創出される瞬間と定義している．語義の変化が時期間の語義の対応関係を指す一方，Linguistic Innovaiton はより限定的に，新たな語彙または語義が出現する事象を指す．本稿における Linguistic Innovation は，Andersen らの定義に依拠する．

Tredici ら [6] は，Linguistic Innovation の定着に関して，ソーシャルネットワークにおけるメンバー間の紐帯の強さ [12] の影響を検証した．Reddit のデータでの検証により，他コミュニティとの弱い紐帯が多いメンバーは Linguistic Innovation を新たに発生させ，コミュニティ内の強い紐帯が多いメンバーは生まれた語の普及に貢献することを明らかにした．

自然言語処理分野では，Linguistic Innovation の検出と似たタスクが盛んに行われてきた．例えば，人による語義の揺らぎの定量化 [14] や，語義の拡張のモデル化による語義曖昧性解消タスクの性能向上が試みられている [24]．

集団間での語義の違いという観点では，経時的な意味変化を捉える研究が広く行われてきた [11, 19]．Nagata らは，比較するテキスト間の埋め込み空間の対応付けを必要としない計算コストの低い手法を提案している [13]．本研究では，Nagata らの手法を用い，コミュニティ間の語義の違いを検出することで Linguistic Innovation スコアを算出する．

表 1: YouTube データセットの統計量

	Game	Politics	Sports	Variety
コミュニティ数	100	52	47	49
各コメント数	20,000	15,000	10,000	40,000
各平均トークン数	13.9	15.8	19.1	19.3
平均動画数	484.3	411.5	421.9	289.7

3 データ

コミュニティと解釈できる構造が存在し，ユーザが頻繁にテキストで発信するプラットフォームとして YouTube 上の動画への投稿コメントを使用し，各チャンネルをコミュニティと定義した．話題や規模，期間を大まかに層別するため，対象コミュニティは 4 つのカテゴリ (Game，Politics，Sports，Variety) にそれぞれ関連し [1]，2024 年 10 月時点で登録者数が閾値を超える日本語コンテンツのみとした．登録者数の閾値はカテゴリごとに，Game は 50 万人，Politics は 20 万人，Sports は 40 万人，Variety は 150 万人と設定した．

データ収集には YouTube Data API v3 [2] を用い，コミュニティ間でコメント数が均一になるようにランダムサンプリングを行った．統計量を表 1 に示す．なお，コメントの分かち書きには MeCab [10] を用い，辞書は NEologd [21] を用いた[3]．

4 手法

4.1 集団的アイデンティティスコア

コミュニティごとの親和スコア，認知プロセススコア，集団的アイデンティティスコアを，日本語版の LIWC として開発された J-LIWC [9] を用いて，それぞれ算出する．

LIWC により推定する集団的アイデンティティスコア [2] は，次のように定義される：

$$\begin{aligned} &\text{集団的アイデンティティスコア} \\ &= z(Score_C^{Aff}) - z(Score_C^{Cog}), \end{aligned} \tag{1}$$

$$Score_C^{Aff}, Score_C^{Cog} = \frac{1}{N_C} \sum_{s \in S_C} \frac{LIWC(s)}{WC(s)} \tag{2}$$

ここで，$Score_C^{Aff}$，$Score_C^{Cog}$ はそれぞれコミュニティごとの親和スコアと認知プロセススコア，$z(\cdot)$ は標準化（カテゴリ内で平均 0，分散 1 になるよう

1） カテゴリの判定は https://yutura.net/ 内の，ユーザ投票によるタグ付けを参考にした．
2） https://developers.google.com/youtube/v3?hl=ja
3） 使用したチャンネル ID のリストは Github にアップした https://github.com/sociocom/Linguistic-Innovation

変換），S_C はコミュニティ C のコメント集合，N_C はコミュニティ C のコメント数，$LIWC(\cdot)$ はテキストを分かち書きし，J-LIWC の心理カテゴリ（親和・認知プロセス）内の語数をカウントする関数，$WC(\cdot)$ は文字数をカウントする関数を示す．なお，前処理としてストップワードは削除した[4]．

4.2 Linguistic Innovation スコア

Linguistic Innovation の発生頻度（LI スコア）をコミュニティごとに算出する．Nagata ら [13] に倣い，語ごとにコミュニティ内・コミュニティ外のコメント集合での埋め込み表現のノルムの平均値を比較する．なお，語の埋め込み表現は日本語版 Wikipedia で事前学習を行ったモデル[5]を用いて獲得し，ノルムを比較する語の頻度の閾値は 10 回とした．Nagata らによれば，語埋め込み表現のノルムは語義の集中度と解釈でき，ノルム差は語義の集中度の違いの大きさと解釈できる．コミュニティ内での語義の集中度がコミュニティ外よりも低い語は，コミュニティ内で意味が集中しており，独自に使用されているとみなし，ノルム差を LI スコアに使用する．埋め込み空間同士のノルム差 $c(S,T)$ は次の式で定義される：

$$c(S,T) = \log \frac{l_T(1-l_S^2)}{l_S(1-l_T^2)}. \quad (3)$$

本研究の文脈では，T は対象コミュニティのコメント集合，S は対象コミュニティを除いたコメント集合の全語の埋め込み表現を表す．l はコメント集合内の特定の語のノルムを表す．この定義に従い，ノルムの差が閾値以上となった語数を LI スコアとし，このスコアをコミュニティごとに算出した．なお，各語のノルムは、構成するサブワードのノルムの平均値として算出した。また，ストップワード，数字のみの語，日本語が含まれない語はスコア算出の対象から除外した．対象コミュニティを除いたコメント集合には，他コミュニティのテキストから，データ数の 10 倍の件数をサンプリングした．

ノルム差の閾値について述べる．語義の集中度の閾値が大きいほど，メンバー自身が語の用法の独自性を認識している可能性が高い語が抽出できる一方で，対象語が減少するため結果の信頼性は低下するトレードオフが存在する．我々は予備実験の結果，どのコミュニティでも 1 語以上が算出対象となる限りで大きい閾値として 0.4 を採用した．

4) Slouthlib プロジェクトが提供するリストを使用．
http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt

5) https://huggingface.co/tohoku-nlp/bert-base-japanese-whole-word-masking

表 2: 各カテゴリにおける LI スコアと集団的アイデンティティスコアのピアソン相関係数．相関係数の右肩の*は有意な値であることを示す（p<0.01）．

カテゴリ	相関係数	親和スコア (平均)	認知プロセススコア (平均)
Game	.400*	0.286	0.783
Politics	.032	0.369	2.589
Sports	.461*	0.227	1.086
Variety	.536*	0.328	0.869

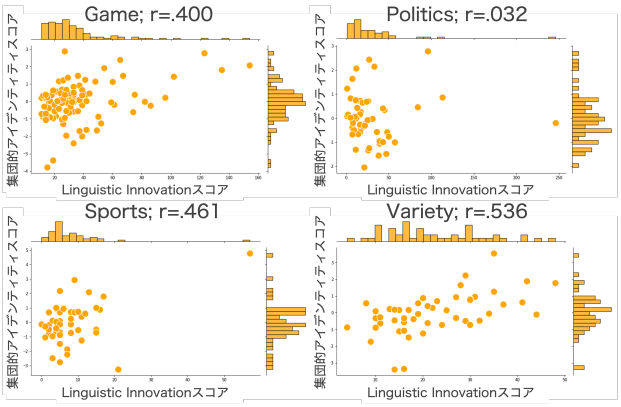

図 2: LI スコアと集団的アイデンティティスコアの散布図．

4.3 仮説の検証

集団的アイデンティティスコアと LI スコアとのピアソン相関係数により評価する．社会心理学の分野では，指標間のピアソン相関係数が 0.4 以上であれば，一方の指標がもう一方の指標の代替指標と解釈できるとされている [17]．そのため本研究でも，相関係数 0.4 以上を評価の基準とする．

5 結果

Game，Sports，Variety のカテゴリは相関係数 0.4 の基準を満たす一方，Politics はおよそ無相関となった（表 2，図 2）．我々の仮説は多くのコミュニティで成り立つことを示した一方で，コミュニティの性質に関する制約が示唆された．

6 考察および今後の課題

6.1 Linguistic Innovation の例

Linguistic Innovation が起こった語の例を示す (表 3 の (1)(2))．(1) の「圧倒的」の例では，「圧倒的 [動画

投稿者名] 大好き」および，その派生語の流行が示されている．(2) の「レモン」は，あるアニメ作品のセリフをそのまま含めたコメントが多い．

逆に，本稿で Linguistic Innovation として捉えてはいない，ノルム差が負の値であった語の例を示す (表 3 の (3)(4))．(3) の「エキゾチック」は，全コメントでゲーム内の固有名詞として使用されていた．(4) の「バチバチ」の例では，語自体は頻出するものの，人物の関係性を表したり，動作を表したりと，文脈が多様であることがうかがえる．

得られた知見として，Linguistic Innovation は一つの語ではなく，複単語表現（例：「はじける」＋「レモン」）として現れることが多い．そのため，bi-gram や tri-gram でのトークン化や，係受けを考慮することで，より正確に Linguistic Innovation を捉えられるだろう．加えて，ノルム差が負になる語は，一貫した独自の語義が創造されているのではなく，多様な文脈で発話される語であることが示唆された．

6.2 仮説を支持しないカテゴリの解釈

Politics カテゴリでは，親和スコアと比較して認知プロセススコアが非常に高い．コメントには特定の組織や人物を批判する内容が多く，「べき」「必要」といった LIWC の認知プロセスカテゴリに含まれる語が頻出する．LIWC を利用した既存手法では，認知プロセスは所属集団への疑義を表すとされていた．しかし，Politics カテゴリの結果は，批判的な発言と集団的アイデンティティの正の相関を示唆し，Common-enemy theory [5] に沿う結果と解釈できる．この対立から，集団的アイデンティティ推定時に認知プロセススコアを除算，あるいは加算するかを，認知プロセスに関わる語が集団の内外いずれに向けられているかに基づき判断する必要があるだろう．

一方で，LI スコアによる集団的アイデンティティ推定指標に条件があるという解釈もできる．そのため，認知プロセススコアが高いコミュニティを対象に，LI スコアと質問紙調査 [7] との相関を調査することも今後の課題である．加えて，今回は LI スコアと LIWC ベース指標との相関を見たが，両者の違いに焦点を当てることも有意義だろう．

6.3 LI スコアの算出方法

今回はコミュニティ外よりもコミュニティ内で語義の集中度が閾値以上に高いものを Linguistic Innovation と定義したが，表 3(3)(4) のように語義の集中度が低い，つまり多様な文脈で使われる語を含める解釈も考えられる．加えて，複単語表現 [16] は語義の創出に寄与していることが自然言語処理の分野でも示唆されていることや [25]，語義の創出が発生しやすい語の意味論的・形態論的な規則が存在すること [3, p37-40; p42-43] が議論されてきており，Linguistic Innovation の検出に言語学的な観点を組み込む有用性が示唆されている．

表 3: Linguistic Innovation が起こった語を含むテキスト例 (1)(2) と，ノルムがより小さい負の値になった語を含む発生コミュニティのテキスト例 (3)(4).

	語	ノルム差	テキスト
(1)	圧倒的	1.47	圧倒的 [動画投稿者名] 大好き
			圧倒的 [動画投稿者名] 感謝
			あ、 [チャンネル名] 新しい動画上げてるーんー何コメントしようかなー、、「圧倒的 [動画投稿者名] 大好き」
(2)	レモン	0.70	はじけるレモンの香り！で察した
			はじけるレモンの香り！←この世代やったわ w 懐かしいな
			はじけるレモンの香り！キュアレモネード懐かしいいいい
(3)	エキゾチック	-0.22	エキゾチック強すぎる
			今回のエキゾチックくらいの武器があった方が楽しい！
			エキゾチック武器はランダムエキゾチックで手に入ります
(4)	バチバチ	-0.20	可愛いのにバチバチのやり合いで笑った w
			[動画投稿者名] くんぜんぜん喋らずにバチバチにキルしてくの気持ちいい
			ピアスバチバチな [動画投稿者名] をこのチャンネルで見るのは珍しいかも

7 おわりに

集団的アイデンティティについて，Linguistic Innovation の発生頻度からの推定を試みた．YouTube 上の 4 つのカテゴリ（Game, Politics, Sports, Variety）における動画へのコメントを使用した．結果は集団的アイデンティティスコアと Linguistic Innovation スコアは，一部を除いたカテゴリで正の相関を示し，指標としての有用性を示したが，コミュニティの性質に関する条件も示唆された．今後は，集団的アイデンティティの質問紙を使った検証や言語学的な知見を活用することで，指標の頑健性を向上させ，コミュニティの安全性や創造性への貢献を目指す．

謝辞

本研究は, JST, 未来社会創造事業, JPMJMI21J2 および「戦略的イノベーション創造プログラム（SIP）」「統合型ヘルスケアシステムの構築」JPJ012425 の支援を受けたものである.

参考文献

[1]Henning Andersen. Understanding linguistic innovations. **Language change: Contributions to the study of its causes**, pages 5–27, 1989.

[2]Ashwini Ashokkumar and James W Pennebaker. Tracking group identity through natural language within groups. **PNAS Nexus**, 1(2):pgac022, 06 2022.

[3]E.M. Bender, A. Lascarides, and G. Hirst. **Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics**. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2019.

[4]Herbert H. Clark. **Using Language**. “Using” Linguistic Books. Cambridge University Press, 1996.

[5]Kris De Jaegher. Common-enemy effects: Multidisciplinary antecedents and economic perspectives. **Journal of Economic Surveys**, 35(1):3–33, October 2020.

[6]Marco Del Tredici and Raquel Fernández. The road to success: Assessing the fate of linguistic innovations in online communities. In **Proceedings of the 27th International Conference on Computational Linguistics**, pages 1591–1603, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[7]Ángel Gómez, Matthew L. Brooks, Michael D. Buhrmester, Alexandra . Vázquez, Jolanda Jetten, and William B. Swann. On the nature of identity fusion: insights into the construct and a new measure. **Journal of personality and social psychology**, 100 5:918–33, 2011.

[8]Michael A. Hogg. **Uncertainty–Identity Theory**, page 69–126. Elsevier, 2007.

[9]Tasuku Igarashi, Shimpei Okuda, and Kazutoshi Sasahara. Development of the Japanese version of the linguistic inquiry and word count dictionary 2015. **Frontiers in Psychology**, 13, March 2022.

[10]Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pages 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[11]Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In **Proceedings of the 27th International Conference on Computational Linguistics**, pages 1384–1397, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[12]Lesley Milroy. **Language and social networks**. Basil Blackwell, 1980.

[13]Ryo Nagata, Hiroya Takamura, Naoki Otani, and Yoshifumi Kawasaki. Variance matters: Detecting semantic differences without corpus/word alignment. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pages 15609–15622, Singapore, December 2023. Association for Computational Linguistics.

[14]Daisuke Oba, Naoki Yoshinaga, Shoetsu Sato, Satoshi Akasaki, and Masashi Toyoda. Modeling personal biases in language use by inducing personalized word embeddings. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pages 2102–2108, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15]James W. Pennebaker, Martha E. Francis, and Roger J. Booth. **Linguistic Inquiry and Word Count**. Lawerence Erlbaum Associates, Mahwah, NJ, 2001.

[16]Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbukh, editor, **Computational Linguistics and Intelligent Text Processing**, pages 1–15, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[17]David L. Streiner, Geoffrey R. Norman, and John Cairney. **Health Measurement Scales: A practical guide to their development and use**. Oxford University Press, 11 2014.

[18]William B. Swann and Michael D. Buhrmester. Identity fusion. **Current Directions in Psychological Science**, 24(1):52–57, February 2015.

[19]Nina Tahmasebi, Lars Borin, and Adam Jatowt. Survey of computational approaches to lexical semantic change detection. **Computational approaches to semantic change**, 6(1), 2021.

[20]Henri Tajfel. An integrative theory of intergroup conflict. **The social psychology of intergroup relations/Brooks/Cole**, 1979.

[21]Taiichi Hashimoto Toshinori Sato and Manabu Okumura. Implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval (in Japanese). In **Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing**, pages NLP2017–B6–1. The Association for Natural Language Processing, 2017.

[22]Jay J. Van Bavel and Andrea Pereira. The partisan brain: An identity-based model of political belief. **Trends in Cognitive Sciences**, 22(3):213–224, 2018.

[23]Jacquelien van Stekelenburg. **Collective Identity**. Wiley, January 2013.

[24]Lei Yu and Yang Xu. Word sense extension. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pages 3281–3294, Toronto, Canada, July 2023. Association for Computational Linguistics.

[25]大友. 和幸 and 橋本 敬. コーパス分析によって明らかにする新語によるコミュニケーションを可能にする複合名詞の語形成. In **言語処理学会 第 29 回年次大会 発表論文集**. 言語処理学会, 2023.

表 4: J-LIWC から算出したスコアが高い（左）または低い（右）コメント例.

高いスコアの例			低いスコアの例		
カテゴリ	スコア	テキスト	カテゴリ	スコア	テキスト
親和	5	昔から、ゲームやってて楽しむ [動画投稿者] くんが好きです!けど、これからの [動画投稿者] くんも大好きです!	親和	0	この動画の編集 tiktok で似たようなの見たようなぁ、気のせいかな?
親和	4	ゲームで遊んでるはずなのにゲームに操作されてる?という気持ちになりました。仕掛けも会話もストーリーも結末も全てが最高でした。この最高のゲームを [動画投稿者] さんの実況で見られた事感謝します。	親和	0	[動画投稿者] 家の猫ちゃんだいぶご長寿だよねかなり前から飼ってるって聞いてる気がする
認知プロセス	8	大阪住んで 0 年経ったけど全部わかった。関西弁の違いわかんないけど、周りみんな使ってるから、京都弁じゃないのかも	認知プロセス	0	明けましておめでとうございます!お正月から嬉しすぎます!!!!今年もよろしくお願いします
認知プロセス	7	耐久?配信お疲れ様でした!初めから最後まで寝ずに見れたの初めてかも知れない...、よく考えたらゲームプレイするのに時間忘れてずっとプレイするのは良くある話だけど、誰かのプレイを時間を忘れて見ているって結構凄い事では!?	認知プロセス	0	配信お疲れ様でした!!だんだんできること増えてくるね!それぞれわちゃわちゃしながらも仲良し組が協力するの見ていてほんと楽しい!次回も楽しみだよー!!

A　親和スコア・認知プロセススコアが高い・低いテキスト例

親和スコアおよび認知プロセススコアが高いテキスト，低いテキストの例を示す 4．親和スコアが高いテキストには視聴チャンネルの動画投稿者に呼びかけながら褒める内容が多く，低いテキストには感情をあまり直接的に表出せず，投稿者に呼びかけるというよりはひとりごとのような内容が多い．認知プロセススコアが高いテキストは自身の推論（感覚ではなく）を述べているものが多く，低いテキストは推論（理由や逆接，「考えた」「思った」など）を表す語を使わずに感想を述べるものが多い．