NLP モデルの性能の再現可能な測定に向けて：再現性の時間軸モデルと日本の NLP 研究の再現性の簡易的調査

竹下昌志 [1]　ジェプカ・ラファウ [2]　荒木健治 [2]
[1] 北海道大学大学院情報科学院
[2] 北海道大学大学院情報科学研究院
{takeshita.masashi,rzepka,araki}@ist.hokudai.ac.jp

概要

再現性は研究の信頼性を評価する上で重要な指標であるが，「再現性」とその関連用語の定義は混迷している．そこで本稿では，時間軸で再現性の諸側面を分類する概念モデルを提案する．これによって既存の定義と今後提案される定義を包括して再現性を理解することができる．次に，再現性を改善するための方法を検討する．最後に，近年の日本の NLP 研究の再現性を簡易的に調査する．その結果，特にコードが公開されておらず，追試する者にとって難しい状況であることが明らかとなった．

1　はじめに

再現性は科学にとって重要である[1)]．「再現性」とは，広義には，論文で記述されていることを繰り返すことができるという研究およびその研究結果の性質である．実験結果が再現できない場合はその実験結果に対する疑義が生じ，再現できればその実験結果の信頼性が向上する．また追試が行われていない場合，その実験結果の信頼性は不明である．本大会のワークショップ「NLP における再現性」[2)]の概要で述べられているように，「実験の再現性は健全な議論のために不可欠の条件である」．

しかし，「再現性」やその関連用語の定義は混迷しており，対立する定義も存在する（**表 1** 及び**付録 A**）．「再現性」の定義が定まらなければ，人々が「再現性」で意味することが異なるために議論が整理されず，研究の再現性の評価が困難になる．

そこで本稿では「再現性」の様々な定義を捉える包括的な概念モデルである**再現性の時間軸モデル**を提案する．この概念モデルでは，再現性を何らかの測定の再現性であるとし，測定前，測定結果，測定後の再現性をそれぞれ区別する．また本稿では，この再現性モデルの下で二つのことを議論する．第一に，再現性をどのように向上させるかを検討する（**3 節**）．第二に，日本の NLP 研究で再現性がどれほど確保されているかを簡易的に調査する（**4 節**）．

2　再現性の時間軸モデル

再現性とその他の関連用語に関しては様々な定義がなされているが，定義は一致していない[3)]．そこで本稿では，再現性の定義について各提案を包括的に整理するために，測定前，測定結果，測定後の再現性にそれぞれ分類する再現性の時間軸モデルを提案する．既存の定義と，私たちの提案する再現性の分類との対応を**表 1** に示す．

本節ではまず再現性の時間軸モデルの利点を説明し，次にこの概念モデルの詳細を説明する．

2.1　再現性の時間軸モデルの利点

既存の再現性の定義の仕方は，「再現性」に一つの側面だけを認めるか，実験に関する様々な側面に関して再現性をそれぞれ定義するかのいずれかである（**表 1** 及び**付録 A**）．しかし，測定には他にも様々な事物が関わるため，このように再現性を定義することは有限ではなく，望ましくないと考える．

本稿で提案する再現性の時間軸モデルでは，既存の定義と今後提案される定義を時間軸上に分類することで包括的に捉えられるため，再現性の理解や分類に有用であると考える．また再現性の諸側面を時間という概念的に同じ側面で分類することができるため，分類の仕方として適切であると考える．

1）　近年は科学における再現性に対する懸念が高まっており，「再現性の危機」とも呼ばれている．これについての哲学的認識論も含めた概要については Fidler と Wilcox [1] を参照．
2）　https://sites.google.com/view/reproducible-nlp-ws
3）　再現性や反復性の定義をめぐる議論を本稿で網羅的に扱うことはできない．他分野での定義を含めた用語法を概観するものとして Plesser [6] を参照．

表 1　既存の再現性関連の用語の定義と，再現性の時間軸モデルとの対応表．各項目の説明は**付録 A** を参照．

再現性の時間軸モデル（提案）	測定前の再現性	測定結果の再現性	測定後の再現性
Goodman ら [2] の定義	方法の再現性	結果の再現性（＝複製性）	推論の再現性
Cohen ら [3] の定義	複製性（＝反復性）	知見の再現性，値の再現性	結論の再現性
Belz ら [4]・Belz [5] の定義		再現性，複製性（＝反復性）	

2.2　測定前の再現性

測定前の再現性は，測定を再現するために必要な手続きや機器などを繰り返すことができるか否かを意味する．Belz [5] は測定に関わる条件を対象条件，測定法条件，測定手続条件に分類している．対象条件は測定機器に関する条件であり，測定に用いたコードやモデルなどが含まれる[4]．測定法条件には評価指標とその実装者が含まれる．測定手続条件には用いたテストデータや実行環境などが含まれる．これらの条件を繰り返せるということが測定前を再現できることを意味する．

2.3　測定結果の再現性

測定結果の再現性は，測定結果を繰り返すことができるか否かを意味する．「結果」には，測定によって得られた値やその値を統計的に処理した値（例：平均値）を含み，また Cohen ら [3] に合わせて値の比較も含める．そのため，信頼区間の推定結果や統計検定の結果も「結果」に含まれる．

測定結果の再現性に，測定前の再現性は必要でも十分でもない．必要でない理由は，例えば高価な GPU やコンパイル・実行環境を再現することは困難だが，それにもかかわらず同一の測定結果（正解率など）を得ることは可能である．十分でない理由は，深層学習モデルの場合，完全に測定前の再現性を満たさない限り，重みの初期値などのランダム性のために測定結果が厳密に再現されないことがありうるからである．

2.4　測定後の再現性

測定後の再現性とは，測定結果を用いた推論とその結論を繰り返すことができるか否かを意味する．

別の測定結果から同じ結論を導く推論も，同じ測定結果から別の結論を導く推論も可能である [2]．前者の例として，BERT [8] は様々なタスクで成功を収めていることから，別々の測定結果から「BERT は優れた性能をもつ」という同じ結論を導く推論は可能である．また後者の例として，ある研究者が自身の提案手法を英語の分類タスクで実験し，良い分類精度が得られた後に「提案手法は優れた分類性能をもつ」と推論したとする．しかし別の研究者は他の言語では良い精度が得られない（一般化できない）可能性を考慮し，この推論に同意しないとする．この場合，同じ測定結果から異なる結論・推論になったため，測定後の再現は失敗している．

3　測定の再現性の改善に向かって

本節では以上の概念整理の下で，各再現性を改善させるための方法を検討する．

3.1　測定前・結果・後の再現性の改善

すべての再現性の改善に共通することは，実験を繰り返すために必要な事柄を明示することである．ACL Rolling Review（ARR）は実験設定の詳細の明示化のために，付録のページ数を論文のページ数制限に含めないとし，また副次資料の提出も認めている[5]．また ARR は責任ある研究チェックリスト [9] を提示しており，一部は再現性のチェックリストとしての役割が意図されている．こうしたチェックリストを満たすことは重要である[6]．

3.2　測定前の再現性の改善

測定前の再現性の改善には，**2.2 節**で紹介した Belz [5] の三条件を満たすことが重要である．特に，コードとデータの公開は測定前の再現性の改善に重要であり，「必須の前提条件」 [12] である．

また，実験で用いたランダムシード値（RS 値）も明示されるのがよいと考える．ARR や NeurIPS のチェックリスト [13] には，RS 値をいくつ用いたのかや，RS 値によるランダム性を明示することが含まれているが，RS 値自体の明示化には何も述べてない．測定前の再現性を厳密に満たすことを目指す

4）　Gundersen と Kjensmo [7] は AI 手法と AI プログラムを区別している．AI 手法は抽象的なアルゴリズムやアイデアであり，AI プログラムはそのアルゴリズムをハードウェア上で実行できるものにしたものである．各条件を満たす上でこうした細かな区別は重要である．

5）　https://aclrollingreview.org/cfp

6）　他分野での実践を参考に，事前登録制度を実施することも再現性の改善につながると考える [10]．NLP における事前登録制度に関しては van-Miltenburg ら [11] を参照．

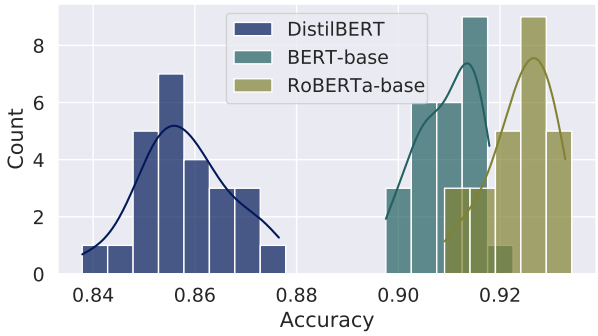

図 1　3.4 節の実験について，livedoor ニュースコーパス上で日本語のマスク言語モデルを 25 個の RS 値（0 〜 24）で fine-tuning した結果のヒストグラムを示す．縦軸はその階級値に属する個数で，横軸は 9 クラス分類の正解率を表す．実験の詳細を**付録 B** に示す．

表 2　3.4 節の実験での各モデルの正解率の平均値と標準偏差．

	BERT	RoBERTa	DistilBERT
平均	0.9096	0.9241	0.8581
標準偏差	0.005775	0.006091	0.009026

のであれば，RS 値の明示化は重要である．

3.3　測定結果の再現性の改善

測定結果の再現性の改善のためには，測定によって得られた結果の明示化が重要である．一部の論文では統計検定が行われているが，どの実験のどの値に関して何の統計検定を行ったのかが明示されていないことがある [14]．これを明示することはチェリーピッキングの疑義を晴らす上で重要である．

また，ハイパーパラメータやモデル構造の探索結果を仮説検証型の研究として報告するのは HARKing [15] につながり，問題がある [16, 10]．HARKing とは “Hypothesizing After the Results are Know”（結果を知った後に仮説を作る）の略称で，結果を得た後にその結果に合うように仮説を形成することを意味する．これが問題であることは，テストデータを用いた学習モデルでのテストデータの予測結果を，あたかもテストデータを学習に使ってないかのように報告することと類似していることから理解できる．Gencouglu ら [16] は複数の院生を用いた手作業によるハイパーパラメータやモデルの探索を院生降下法（Grad Student Descent）と称し，これによってなぜそのモデルやハイパーパラメータが優れているのかが分からなくなると論じている．

測定結果の再現性の観点からは，こうした実践は「知見の再現性」 [3] を低下させる．「知見の再現性」

表 3　測定前の再現性の調査結果．

		NLP2021（計 25 本）	『自然言語処理』（計 17 本）
データ公開	完全公開	10 (40%)	11 (約 65%)
	部分公開	4 (16%)	2 (約 12%)
	非公開	11 (44%)	4 (約 24%)
コード公開	完全公開	0	0
	部分公開	0	1 (約 6%)
	非公開	25 (100%)	16 (約 94%)

には，測定によって得られた値の比較を伴い，統計検定の結果などが含まれる．例えば，p 値が 5%の有意水準を下回っていたとしても，20 回に 1 回はそのような結果が期待される．そのため，院生降下法などによる探索的研究で複数回の試行を行った場合に p 値が偶然有意水準を下回ることがありえる．その結果を元に論文の「ストーリー」を組み立てて仮説検証型の研究として報告するのは HARKing にあたり，問題がある．

3.4　測定後の再現性の改善

測定後の再現性を改善するためには，測定によって得られた結果の限界に注意すべきである．ARR のチェックリストの最初の項目は「A1. あなたの研究の限界を述べたか？」であり，強い仮定の明示，自分の主張の範囲の反省などが含まれる．

深層学習モデルの実験には RS 値に依存したランダム性があるため [17, 18]，単一の RS 値のみを用いた実験によって既存手法の精度を超えたとしても，それが偶然そうなった可能性を考慮すべきである[8]．RS 値による精度の変動の一例として，livedoor ニュースコーパス[9]を用いたニュース記事分類タスクの実験結果を**図 1** 及び**表 2** に示す．実験に用いたモデルは日本語コーパスで事前学習された BERT[10]，RoBERTa [20][11]，DistilBERT [21][12] である．実験の詳細を**付録 B** に記載する．また，用いたコードを以下の脚注の URL で公開している[13]．結果として，例えば，BERT の正解率と RoBERTa

7）　このうち 1 本の論文では複数の RS 値を使用したかどうかの記述が見つからなかったが，複数回の試行を行っていた．
8）　深層学習のランダム性を考慮した適切な実験方法とその結果の報告に関する議論として Agarwal ら [19] を参照．
9）　https://www.rondhuit.com/download.html#ldcc
10）　https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking
11）　https://huggingface.co/rinna/japanese-roberta-base
12）　https://github.com/BandaiNamcoResearchInc/DistilBERT-base-jp/blob/main/docs/GUIDE.md
13）　https://github.com/Language-Media-Lab/reproducibility-random-seed/

表 4　関連する論文の測定後の再現度の調査結果．NLP2021 の調査対象 25 本中残りの 6 本の論文のうち 3 本はルールベースの手法であり，うち 1 本では交差検証を行っていた．その他の 3 本はデータ分析を行っていた．

	複数の RS 値の使用	交差検証	エラーバーの表示	統計検定の実行
NLP2021 （計 19 本）	3 (約 16%)[7)]	0	2 (約 11%)	0
『自然言語処理』（計 17 本）	3 (約 18%)	1 (約 6%)	2 (約 12%)	7 (約 41%)

の正解率をそれぞれ一つずつ組み合わせた場合に BERT の正解率が RoBERTa の正解率を超えた割合は 4.64% であった．この結果より，二つのモデルを比較して一方が他方より優れていると言うためには，複数の RS 値を用いた実験を行うことが重要であることが示唆される．また，ここでは統計検定の結果や効果量の報告が有用である[14)]．例えば，RoBERTa での実験の平均値と BERT での実験の平均値が等しいという帰無仮説の下でのスチューデントの両側 t 検定での p 値は 3.83×10^{-11} であった．このことから，RoBERTa の livedoor ニュースコーパスでの平均的な分類性能は BERT より優れていると推論することは合理的だと考える．

4　日本の NLP 研究における再現度

本節では日本の NLP 研究の再現性を簡易的に調査する．調査対象は，言語処理学会第 27 回年次大会（NLP2021）で発表された論文からランダムに選ばれた 25 本[15)]，会誌『自然言語処理』28 巻 3，4 号に掲載された一般論文 17 本である．

4.1　測定前の再現度

測定前の再現性を確保するためにはコードやデータの公開が重要である．しかし，少なくともコードはあまり公開されてない．Wieling ら [12] による調査によれば，コード公開とデータ公開のどちらも改善されつつあるが，2016 年の ACL で発表された論文に関して，データが共有されたのは全体の 86.3%，コードが共有されたのは全体の 59.3% であった[16)]．

本稿で調査したコード・データの公開状況の結果を**表 3** に示す．データ公開については半分以上の論文が少なくとも部分的に公開しているか公開されているデータを用いていた．プライバシーの問題があるため，データは公開可能な範囲で公開されるのがよいと考える．一方コード公開に関しては，『自然言語処理』の一つの論文以外はすべて非公開だった．本稿の調査ではメール等で問い合わせてないため，もしメール等で問い合わせればこの割合は増えるかもしれないが，論文自体に公開情報を載せることがよいと考える．

4.2　測定結果の再現度

Belz ら [4] による様々な追試のメタ分析によれば，同一条件下での実験（つまり測定前の再現性が確保されている実験）で元論文と厳密に同じスコアが得られた実験は全体の 14.03% であり，1% 以上スコアが異なる追試結果は全体の約 6 割であった．

日本の NLP 研究における測定結果の再現性を調査するには実際に追試が行われる必要があるが，私たちの知る限り，公開で行われたことはない．国際的には，人工知能系の学会で再現実験チャレンジがここ数年行われている [23, 24]．今後日本でも行われることを期待する．

4.3　測定後の再現度

測定後の再現性を直接評価することは困難であるが，複数の RS 値の使用や統計検定の実施を調べることで間接的かつ定量的に評価できる．Dror ら [14] による調査によれば，2017 年の ACL と TACL で，統計検定を行うべき研究で適切な検定の実施を確認できたのは，ACL では 180 本中 36 本（20%）で，TACL では 33 本中 18 本（約 54.5%）だった．

本稿での調査結果を**表 4** に示す．査読付きの『自然言語処理』では半分以上の論文がランダム性を考慮した結果の報告を行っているが，NLP2021 では半分に満たなかった．今後はランダム性を考慮した実験を行い，結果が適切に報告されるのが望ましい．

5　まとめ

本稿では，既存の再現性の定義や説明を包括して整理できる再現性の時間軸モデルを提案した．またこのモデルの下で，各時間軸での再現性を改善する方法について議論した．最後に，日本の NLP 研究の再現性の簡易的調査を行った．本稿の議論は再現性の改善への小さな一歩であり，また今後，再現性に関する議論が活発に行われることを期待する．

14)　（不）適切な統計検定の実施に関しては阿部 [22, 第 9 章] や Dror ら [14] を参照．
15)　調査対象とした論文の発表番号を**付録 C** に示す．
16)　これは，論文に共有のためのリンクが含まれなかった場合には著者にメールで問い合わせた上での結果である．

謝辞

日本の NLP 研究の再現性を調査するにあたって手伝っていただいた同僚の吉井瑞貴さんに感謝します.

参考文献

[1] Fiona Fidler and John Wilcox. Reproducibility of Scientific Results. In Edward N. Zalta, editor, **The Stanford Encyclopedia of Philosophy**. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

[2] Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. What does research reproducibility mean? **Science Translational Medicine**, Vol. 8, No. 341, pp. 341ps12–341ps12, 2016.

[3] K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. Three dimensions of reproducibility in natural language processing. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[4] Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. A systematic review of reproducibility research in natural language processing. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 381–393, Online, April 2021. Association for Computational Linguistics.

[5] Anya Belz. Quantifying reproducibility in NLP and ML. **arXiv preprint arXiv:2109.01211**, 2021.

[6] Hans E. Plesser. Reproducibility vs. replicability: A brief history of a confused terminology. **Frontiers in Neuroinformatics**, Vol. 11, p. 76, 2018.

[7] Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 32, No. 1, Apr. 2018.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] ACL Rolling Review. Arr responsible nlp research checklist, 2021, https://aclrollingreview.org/responsibleNLPresearch/.

[10] Samuel J Bell and Onno P Kampman. Perspectives on machine learning from psychology's reproducibility crisis. **arXiv preprint arXiv:2104.08878**, 2021.

[11] Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. Preregistering NLP research. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 613–623, Online, June 2021. Association for Computational Linguistics.

[12] Martijn Wieling, Josine Rawee, and Gertjan van Noord. Squib: Reproducibility in computational linguistics: Are we willing to share? **Computational Linguistics**, Vol. 44, No. 4, pp. 641–649, December 2018.

[13] NeurIPS 2021. Neurips paper checklist, 2021, https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist.

[14] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[15] Norbert L. Kerr. Harking: Hypothesizing after the results are known. **Personality and Social Psychology Review**, Vol. 2, No. 3, pp. 196–217, 1998. PMID: 15647155.

[16] Oguzhan Gencoglu, Mark van Gils, Esin Guldogan, Chamin Morikawa, Mehmet Süzen, Mathias Gruber, Jussi Leinonen, and Heikki Huttunen. Hark side of deep learning–from grad student descent to automated machine learning. **arXiv preprint arXiv:1904.07633**, 2019.

[17] Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 338–348, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[18] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. **arXiv preprint arXiv:2002.06305**, 2020.

[19] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. **Advances in Neural Information Processing Systems**, Vol. 34, , 2021.

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. **Computing Research Repository**, Vol. arXiv:1907.11692, , 2019.

[21] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. **arXiv preprint arXiv:1910.01108**, 2019.

[22] 阿部真人. データ分析に必須の知識・考え方　統計学入門：仮説検定から統計モデリングまで重要トピックを完全網羅. ソシム, 2021.

[23] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, Hugo Larochelle. Improving reproducibility in machine learning research: a report from the neurips 2019 reproducibility program. **Journal of Machine Learning Research**, Vol. 22, , 2021.

[24] Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In **Proceedings of the 14th International Conference on Natural Language Generation**, pp. 249–258, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics.

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **International Conference on Learning Representations**, 2019.

A　表 1 の各項目の説明

Goodman ら [2] は「方法の再現性」,「結果の再現性」,「推論の再現性」という 3 つの再現性を提案している.「方法の再現性」とは，同じ結果を得るための実験方法や手続き自体を繰り返すことができるという性質である.「結果の再現性」は可能な限り同じ手続きで別の研究チームが同じ結果を得られるという性質であり，Goodman らによれば，これは従来の複製性（Replicability）に対応する.「推論の再現性」は結果の分析や考察から同じ結論を導けるという性質である.

Cohen ら [3] は「結論の再現性」「知見の再現性」「値の再現性」という 3 種類の再現性を提案し，別の概念として複製性または反復性（Repeatability）を提案している．ここで複製性は「方法の再現性」に対応し，「知見の再現性」は「結果の再現性」に対応するが，「発見の性質」は値の比較に基づくとされる.「値の再現性」は測定結果の値自体を繰り返して確認できるかという性質であり，結論の再現性は，論文の「結論」の項目の記述であるとしている.

Belz ら [4] および Belz [5] は，再現性と複製性（または反復性）の両方を測定結果の性質として定義している．Belz らは，複製性を実験時刻以外が同じ条件の下での測定結果（の分布）であると定義し，再現性を異なる条件下での測定結果（の分布）であると定義している.

B　livedoor ニュースコーパス上での実験詳細

livedoor ニュースコーパスのデータセットの統計情報を**表 5** に記載する．本実験では本データセットを 6 : 2 : 2 の割合で学習セット（4,420 件），検証セット（1,473 件），テストセット（1,474 件）に分割した．各データにはタイトルと本文が含まれているが，本実験では本文のみを用いた.

本実験で使用したハイパーパラメータを**表 6** に示す．使用した三つのモデルでハイパーパラメータは共通である．また，本実験では早期終了を用いており，検証セットでの損失が 3 回連続で更新されなかった場合に，その時点までで検証セットで最も損失が小さいモデルを用いてテストセットで評価した.

表 5　livedoor ニュースコーパスの各カテゴリ内のデータ数.

カテゴリ名	データ数
独女通信	870
IT ライフハック	870
家電チャンネル	864
livedoor HOMME	511
MOVIE ENTER	870
Peachy	842
エスマックス	870
Sports Watch	900
トピックニュース	770
合計	7367

表 6　実験に用いたハイパーパラメータ．本実験では三つのモデルすべてで共通のハイパーパラメータを使用した．AdamW の学習率以外のハイパーパラメータは Pytorch のデフォルト設定を用いた[17]．

ハイパーパラメータ	設定値
epoch 数	最大 30
batch size	32
入力の系列長	256 トークン
RS 値	0〜24
最適化手法	AdamW [25]
学習率	2×10^{-5}
β_1	0.9
β_2	0.999
ϵ	1×10^{-8}
weight decay	0.01

C　再現性調査の対象とした NLP2021 の論文の発表番号一覧

A6-1, A8-4, B7-2, B7-3, C4-2, C8-4, D1-4, D2-2, D5-3, D6-3, D8-1, E9-3, P1-3, P1-6, P3-4, P3-12, P3-17, P4-6, P4-21, P5-14, P6-15, P6-20, P7-14, P8-1, P9-14

17）https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html