

「NINJAL-LINE コーパス（仮）」の設計と公開

落合 哉人（国立国語研究所）[†]

Design and Public Availability of the NINJAL-LINE Corpus (Provisional)

Kanato Ochiai (National Institute for Japanese Language and Linguistics)

要旨

首都圏及び関西圏の若年層日本語話者を中心に LINE のスクリーンショットのデータを募集し、文字化・個人情報のマスキング処理を施したコーパスの構築を進めている。本コーパスの特徴は、やりとりに関わるすべての参加者からあらかじめ研究利用及び公開の同意を得る点にある。本稿では、本コーパスに含まれるデータの概要を示した上で、収集方法と整備方法について説明した。また、現時点で収集できたデータについてその一部を例示するとともに、今後の公開のあり方を論じた。

1. はじめに

本稿は、現在構築中の「NINJAL-LINE コーパス（仮）」（以下、本コーパスと呼ぶ。）について設計方針を説明するとともに、現時点で文字化が完了した一部のデータに関して周知を図るものである。アプリケーションソフトウェア「LINE」は、2011年に登場して以来、多くの日本語話者の日常的な言語生活において欠かせない連絡手段として定着してきた。しかし、LINE の主要な機能であるテキストチャット（以下、これを LINE と呼ぶ。）のデータは秘匿性の高さから自動収集することが難しく、これまで言語使用の実態が正確に把握されてきたとは言い難い。

そこで、筆者の取り組みでは、昨年度（2024 年度）より若年層を中心的な対象として LINE のスクリーンショットの収集を進めており、文字化・形態素解析・メタ情報の付与を施した上で（サブ）コーパスとすることを計画している。同様の大規模なデータ収集に関しては、筆者によるものも含めてこれまでも先行する取り組みがあるが（加納他，2017；宮寄，2018；三宅，2019；落合，2019，2025b；楊・倉田，2023），本コーパスの新たな点は、やりとりに関わるすべての参加者から、あらかじめ研究利用・公開の許諾を得ている点にある。

以下、収集中のデータの概要を述べた上で（2 節）、スクリーンショットを文字化する際の方針を説明する（3 節）。また、すでに収集したデータの一部を示すとともに、研究者間での共有のあり方に関して論じる（4 節）。

2. データの収集

2.1 対象とするデータ

まず、現在収集を進めているデータについて説明する。本コーパスには、300 発信程度の連続したやりとりを1談話として¹、大きく以下3種類の談話を収めることを予定している。

[†] katakanato@ninjal.ac.jp

¹ ただし、2.2 で述べる通り、実際にはスクリーンショットの枚数（7 枚）を基準としてデータの募集を行なっているため、送り手のコミュニケーションスタイルに起因して 1 発信が非常に長いものに関してはその限りでない。

- (a)すべての参加者が首都圏出身（東京都・千葉県・埼玉県・神奈川県・茨城県南部）かつ、10代～20代の、友人同士を中心とする親しい間柄同士のやりとり（計1万発信／年）
- (b)すべての参加者が首都圏出身（東京都・千葉県・埼玉県・神奈川県・茨城県南部）かつ、30代の、友人同士を中心とする親しい間柄同士のやりとり（計1万発信／年）
- (c)すべての参加者が関西圏出身（大阪府・京都府・兵庫県・奈良県・和歌山県・滋賀県・三重県）かつ、10代～20代の、友人同士を中心とする親しい間柄同士のやりとり（計1万発信／年）

これらのデータを対象とする理由として、収集が比較的容易であることと、地域差・年代差を観察できること、の2点が挙げられる。このうち、収集の難易度については、若年層ほど日常的なコミュニケーション系メディアの平均利用時間が長く（総務省情報通信政策研究所, 2025）、スマートフォンにおいて多くのデータを保持している可能性が高いことを指す。また、地域差・年代差については、方言利用（田中, 2014; 三宅, 2018）や、携帯メールの利用経験に基づく言語使用の違いといった問題について実態把握の余地があることを考慮している。

特に取り組みの初年度（2024年度）は、上記のうち、(a)のデータについて31談話（1対1のやりとりが20談話／3人以上によるやりとりが11談話）計10,811発信を収集した。また、2年目（2025年度）以降は、直近1年以内になされた(a)(b)(c)すべてのデータについて、1年あたり各1万発信（計3万発信）を目安として募集を進めている。したがって、2026年度までの3年間を一区切りとすると、データサイズは少なくとも7万発信となり²、1発信10語として約70万語が分析可能となる見込みである。

なお、文字を基調とするCMC (Computer-Mediated Communication)の特徴として、音声を基調とする会話と比べても、ひとつの話題が終結するスピードが早く、相対的に話題の移り変わりが激しいことが挙げられる（Herring, 1999）。そのため、少なくとも協力者自身の判断では統制が困難であると思われることから、やりとりの内容はデータを募集する時点で特に条件を設けなかった。同様に、参加者の関係性については、募集時点で「友人同士（親しい間柄であれば、先輩後輩の関係や上司部下の関係等でもかまわない）」と指定しているが、一部、家族（姉妹）間のデータについても提供があった。

2.2 収集の手続き

データの収集においては、2.1で挙げた(a)(b)(c)の属性に合致する日本語話者を広く募り、自らのLINEアカウントにおいて提供できそうなデータがあるかどうか探してもらった³。その際、1談話の分量は、やりとりの始発部または区切りのよい箇所からLINEの「スキャン」機能（図1）で範囲を指定した際に保存できる最大の長さの画像7枚分を目安として提示した。このような形でデータの分量を指定するメリットとして、（実際には例外もあるが、）

² データサイズの目安として、LINEに関する先行研究のうち、これまで最も大規模にデータの収集を行なっていると思われる「M-ZAK LINE データベース」では、約38,500行（発信数は不明）が収められている（宮寄, 2018）。また、公開型コーパスとしてはLINE同様、受け手が特定・少人数のソフトウェア・サービス（落合, 2025a）を対象としたデータ収集の取り組みのうち、日本語の携帯メールを収める「加藤安彦ケータイメールコーパス」（田中他, 2023）において約10万発信のメールが、英語及び中国語のSMS (Short Message Service) を収める NUC SMS Corpus (Chen & Kan, 2013)において約8万発信のメッセージが、それぞれ収められている。したがって、本コーパスは、LINEを対象とする取り組みとしては最も大規模であり、国内外の類似環境を対象とする公開型コーパスとしても比較的大規模であると言える。

³ 個人を特定できる情報を含めデータの収集・管理・分析に関しては、事前に国立国語研究所研究倫理委員会の研究倫理審査を受審しており、承認の判定を受けている。

1 談話あたり最低でも 250 発信程度が確保できることや、あらかじめ、提供されるデータが一連のやりとりである保証が得られること等が挙げられる。

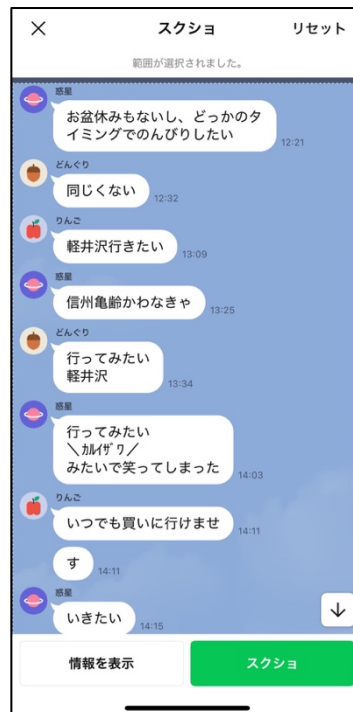


図 1 「スクショ」機能

データ提供の意思がある者（以下、協力者と呼ぶ。）には、まず、本コーパス構築に際して作成した LINE Works アカウントに協力申し出の連絡を行なってもらった。その上で、当該の連絡に対する返信として本コーパスの概要を説明したほか、提供を行うデータに関わる他の参加者（以下、単に参加者と呼ぶ。）にデータ提供の内諾を得ることと、参加者からも LINE Works アカウントに連絡をとるよう依頼することの 2 点を依頼した。

次に、すべての参加者と連絡がとれ次第、LINE Works のアンケート機能を用いて、協力者から提供を受け付けるデータの研究利用・公開に同意するかどうかを回答してもらった。また、この時点で協力者に「スクショ」機能を用いてアイコン等の情報を隠したスクリーンショットを送ってもらい、やはり研究利用・公開に同意しているかどうか回答を依頼した⁴。

最後に、すべての協力者・参加者の同意が確認でき次第、データの分量に応じて調整した金額の謝金を協力者・参加者に支払った。

なお、LINE Works のアンケート機能を用いて同意を取得する際には、同時に「年齢」「出身地」「性別（男性、女性、その他）」「LINE の利用歴（例：5 年）」「LINE の主な利用手段（例：スマートフォン、PC、タブレット）」「スマートフォンの機種（例：iPhone、Android）」に関しても質問を行なった⁵。

⁴ LINE のトーク履歴は、バックアップ機能を利用することでテキストファイルに保存することもできる。ただし、すべての履歴が保存されてしまう都合上、協力者にデータを加工する負担が発生すること等の事情に鑑みて本コーパスではこの方法を採用しないこととした。

⁵ 研究利用・公開に関する同意の対象には、それらの情報も含まれることを事前に案内している。なお、協力者・参加者には同意の有無だけでなく、「この範囲であれば同意する」という条件も申告できるようにした。

3. データの文字化

収集したスクリーンショットについては、個人情報の特定につながり得る人名や地域名、施設名、画像・動画等にマスキングを施したほか、時系列で発信に関する情報の文字化を行った⁶。具体的には、1 談話ごとに 1 フォルダを設けた上で、その直下にマスキング済みの画像ファイルと以下のフォーマットで構成される CSV ファイルを収めた。

フォルダ名、送信日、送信時刻、通番、送信者、リプライ、本文、リアクション、元画像ファイル名、備考

CSV ファイルのフォーマットのうち、「フォルダ名」はその談話の名称⁷を指し、「送信日」と「送信時刻」は文字通り各発信の送信日時を指す⁸。また、「通番」は行番号を、「送信者」は「スクショ」機能で個人情報を隠した際、代わりに表示される名称（例：惑星、どんぐり、りんご）を、「リプライ」はリプライ機能（図 2）が用いられた際に対応する先行発信の行番号を、それぞれ記す⁹。同様に「本文」は談話内の各発信の内容であり、「リアクション」は各発信に付けることのできるリアクション（図 3）の種類を表したものである。さらに「元画像ファイル名」には対応するマスキング済み画像ファイルの名称を記載するほか、「備考」には画像が不鮮明である箇所をはじめ、特に「本文」の内容に関する留意点を記す。

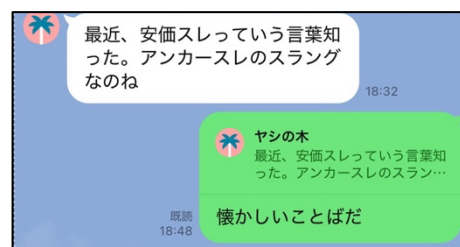


図 2 リプライ機能

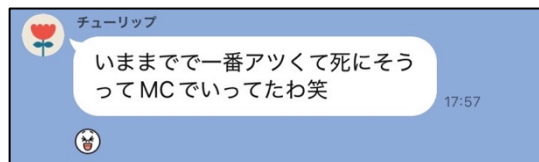


図 3 リアクション

「本文」を書き起こす際は基本的に「送信」ボタンを一度押すことで送信できる吹き出しやスタンプ、画像、動画、その他の PDF ファイル等を 1 発信と捉えた。また、1 発信ごとに 1 行を設けた上で、先行するデータ収集の試みである宮寄(2018)に準じた以下の方針に基づいて文字化を行なった¹⁰。

⁶ 2024 年度に収集したスクリーンショットのマスキング処理及び文字化は、東京反訳株式会社に作業を委託した。

⁷ 2024 年度に収集したデータについては便宜的に参加者が 1 対 1 のやりとりを 001 以降の番号で、3 人以上のやりとりを 00A 以降のアルファベットで表すこととした。

⁸ ただし、送信日についてはスクリーンショットに表示されておらず、不明なものもある（その場合は空欄とした）。

⁹ ただし、吹き出しが右から出る場合、送信者の名称は一律に「自分」と記載することとした。同様に、リプライ機能において対応する先行発信が、提供されたスクリーンショットの範囲に含まれていない場合は[不明]と記載した。

¹⁰ 例外として、音声通話やボイスメッセージ、「送信取り消し」、アルバムの作成及び更新も 1 発

- ①人物や場所、施設名称等で個人の特定ができる可能性が高いものはそれぞれ全角<>で囲い、フォルダ内で現れた順に<人名+全角大文字アルファベット>、<地名+全角大文字アルファベット>、<施設名+全角大文字アルファベット>とする（敬称等は<>に含めない）。

例：<人名 B>さんって<施設名 A>で働いてたよね？

- ②送信者の位置情報が示されている場合は、半角[]を用いて[位置情報]と記入する。

- ③文字、スタンプ、デコ文字、画像、動画等

- ・絵文字については、半角[]で括り、[絵文字]と入力する。
- ・顔文字については、半角[]で括り、[顔文字]と入力する。
- ・スタンプは、半角[]で括り、[スタンプ]と入力する。
- ・スタンプに文字列が付与されている場合は、]の前に当該の文字列を全角（ ）で括り文字列を入力する。

例：[スタンプ（おはよう）]

- ・デコ文字で記入された範囲は、開始部[文字スタンプ S]と終了部[文字スタンプ E]で括る。その際、ブラケットは半角，アルファベットは全角とする。

例：[文字スタンプ S]おめでとう[文字スタンプ E]

- ・画像データは全角<>で括り、一律に<写真>と入力する。
- ・アルバムは全角<>で括り、<アルバム>と入力する。
- ・動画は全角<>で括り、<動画>と入力する。
- ・通話が行われた場合は半角[]を用いて[通話]と記入する。]の後、半角スペースを挟んで通話時間を記入する（不在着信の場合は「不在着信」と記入する）。

例：[通話] 0:54

例：[通話] 不在着信

- ・グループ通話が行われた場合は半角[]を用いて[グループ通話]と記入する。通話時間は記載しない。
- ・ボイスメッセージは半角[]を用いて[ボイスメッセージ]と記入する。]の後、半角スペースを挟んで時間を記入する。

例：[ボイスメッセージ] 09:49

- ・PDF ファイルは全角<>で括り、<pdf>と入力する。その他の拡張子のファイルも<拡張子>の形で記載する。拡張子は半角で記入する。

例：<zip>

同様に、リアクションについては図 4 に示す 6 種類について、左から順に「いいね・了解」「うれしい・大好き」「笑顔・面白い」「感謝・お願い」「悲しい・涙」「驚き・すごい」と名称を付け、付与された発信の「リアクション」列にいずれかを記載した（付与されていない場合は空欄）。その際、1 発信に複数のリアクションが付いている場合は、その数だけ記載を行なった。

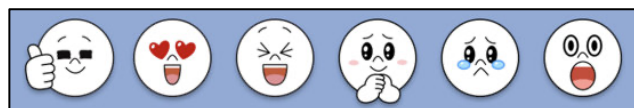


図 4 リアクション

例：3 つのリアクションが付いている場合

信相当とみなした。なお、「送信取り消し」の場合は日時が表示されないため、「送信日」と「送信時刻」は空欄とした。

「いいね・了解」「いいね・了解」「感謝・お願い」

ただし、LINE のリアクション機能では、2025 年 5 月以降のバージョンにおいて、図 4 の 6 種類に限らずすべての絵文字が利用可能となった。そのため、2025 年度以降に収集するデータに関しては、直接絵文字を用いてリアクションを記載することを検討している。

4. 整備済みのデータとその公開のあり方

4.1 整備済みデータの一例

上記の手続きを経て収集・整備されたデータの一部を以下に示す。図 5 はマスキング済みのスクリーンショットであり、表 1 は CSV ファイルのフォーマットに沿ってスクリーンショットを文字化したものである（ただし、「フォルダ名」は割愛する）。なお、図 5 においては 4.2 で述べる通り、スタンプ及び背景の一部をさらに白抜きで隠している。



図 5 マスキング済みのスクリーンショット (00D)

表 1 文字化したデータ

送信日	送信時刻	通番	送信者	リブライ	本文	リアクション	元画像ファイル名	備考
	2:39	1	鉛筆		9/12 ってみなさん空いてますか？	「いいね・了解」	53201800...	
	2:40	2	鉛筆		<人名 A> とのご飯会その日どう？	「いいね・了解」	53201800...	
	7:34	3	鉛筆		19 もどうみんな？	「いいね・了解」	53201800...	
	7:34	4	鉛筆		ちゃんと教えてくれよな		53201800...	
	7:35	5	自分		[スタンプ（既読スルーしてはいけない）]		53201800...	
	8:53	6	梨		空いてます！		53201800...	
	10:25	7	月		12 日は空いてないけど 19 日はあいてる！		53201800...	
	10:42	8	チューリップ	7	おなじく！		53201800...	
	11:54	9	はち		19 の夜ならギリ行けそうです！		53201800...	
	12:09	10	自分		[スタンプ（ありがとうございます）]		53201800...	
	13:40	11	鉛筆		お！じゃあ 19 にしよっか！		53201800...	
	13:47	12	月		りょーかい！		53201800...	
	13:47	13	月		連絡取り合ってくれてありがとう！		53201800...	
	13:52	14	鉛筆		全然大丈夫だよ！		53201800...	
	19:52	15	鉛筆		<施設名 A> 祭行く人ー？		53201800...	
	19:52	16	月		行きたいけど行く人いない人		53201800...	
	19:54	17	自分		21 なら行ける人		53201800...	
	19:54	18	月		うわ 22 なら行ける人		53201800...	
	21:21	19	鉛筆		他は！？		53201800...	
	21:23	20	自分		午前中だったら 22 確実に行ける！		53201800...	
	22:31	21	チューリップ		行きたい！ちなどっちも暇です		53201800...	
	22:31	22	月		21 日午前なら行ける		53201800...	
	22:41	23	梨		どっちも行けない[顔文字]		53201800...	
	22:42	24	梨		けど 22 なら<施設名 A> 祭にいるから会うことはできる		53201800...	
9 月 5 日	18:05	25	チューリップ		結局どっちに行く？22 日？		53201800...	
9 月 5 日	18:36	26	月		土曜日にでも話す？		53201800...	
		27	月		[月がメッセージの送信を取り消しました]		53201800...	
9 月 5 日	18:39	28	月		<写真>		53201800...	
9 月 5 日	18:39	29	月		この時間で取ってくれた？飛行機		53201800...	
9 月 5 日	19:06	30	自分	26	確かに！そうしょ！		53201800...	

4.2 公開のあり方と課題

2024 年度のデータ募集では、協力者・参加者あわせてのべ 84 名（ことなり 82 名）からやりとりの研究利用・公開の同意が取得できた。その結果得られた約 1 万発信（2.1 で前述）は、すでにスクリーンショットのマスキング処理及び文字化が完了しており、公募型の国立国語研究所共同利用型共同研究(B)（随時受付）に申請することで機関外の研究者が利用することが可能である。また、本コーパスは今後、形態素解析及びメタ情報の付与を施した上

で随時、「現代日本語書き言葉均衡コーパス 2」(BCCWJ2 (山崎他, 2025)) のサブコーパスとして収録することを計画している。したがってコーパス検索アプリケーション「中納言」での利用も数年以内に可能となる見込みがある。近年, BCCWJ(2), 「日本語日常会話コーパス」, 「加藤安彦ケータイメールコーパス」といった諸コーパスの整備・公開によって, 直近四半世紀における書きことば・話しことば・CMC のことば (あるいは「打ちことば」(田中, 2014)) を対象とした実証的な日本語研究が展開されつつあるが, 本コーパスの構築を通して, そのような流れが一層進むことが期待される。

一方で, 本コーパスのデータを一般公開する上での課題として, 特にスクリーンショットに含まれるスタンプやデコ文字, 一部の背景については, 製作者・著作権者から一切の研究利用の同意を得ていないことが挙げられる。それらはいずれも CMC に独自の要素である点で使用実態の把握が欠かせないため, 今後さらなるマスキング処理及びコード化等を検討する余地がある。

5. おわりに

本稿では, 本コーパスに含まれる予定のデータに関して概要を説明した上で, 収集方法と整備方法について示した。また, データの一部を例示したほか, 今後の公開のあり方及び, 課題について述べた。本コーパス構築に関する現時点での進捗は未だ初年度のデータ収集と文字化が完了したに過ぎず, 今後本稿で述べた諸手続きも少なからず変更が生じる可能性がある。引き続きデータを提供可能な協力者を募るとともに, 折を見て進捗の報告を行うこととしたい。また同時に, すでに整備できたデータを対象として言語使用の分析も進めていきたい。

謝 辞

本稿は JSPS 科研費 25K16275 による成果の一部を報告するものです。LINE のデータ提供に協力してくださった皆様及び, 関西圏でのデータ募集にご協力くださっている同志社大学大学院文学研究科の谷口悠氏に深く感謝を申し上げます。

なお, 本コーパスに収録するデータは引き続き募集を行なっています。特に 2.1 の(a)(b)(c) に合致するデータの収集にご協力いただける方は筆者のメールアドレスまたは以下の QR コードからアクセスできる LINE Works アカウントにご一報いただけますと幸いです。



<https://works.do/R/ti/p/ochiai.thatright@works-26932>

また, 現時点で収集・整備したデータの利用をご希望の方もまずは筆者のメールアドレスにご連絡いただけますと嬉しく存じます。

文 献

- 落合哉人 (2019). 「LINE テキストチャットにおける分析単位の規定と接続表現の使用傾向」『筑波日本語研究』23, pp.83-112.
- 落合哉人 (2025a). 「「打ちことば」の特質—分類と実態の解明—」『日本語学』44:2, pp.14-25.
- 落合哉人 (2025b). 『「打ちことばの研究—モバイルメディアコミュニケーションから再考する日本語—」 ひつじ書房.

- 総務省情報通信政策研究所(2025)「令和 6 年度情報通信メディアの利用時間と情報行動に関する調査報告書」https://www.soumu.go.jp/main_content/001017241.pdf (2025 年 8 月 17 日確認).
- 田中ゆかり (2014).「ヴァーチャル方言の 3 用法ー「打ちことば」を例としてー」石黒圭・橋本行洋編『話し言葉と書き言葉の接点』pp.37-55. ひつじ書房.
- 田中ゆかり・三宅和子・宮寄由美・林直樹(2023)「加藤安彦ケータイメールコーパスー親密度・絵文字 Unicode 情報が付与された 2000 年代 10 年間のデーター」『計量国語学』34(2), pp.111-121.
- 三宅和子 (2018).「LINE の中の方言ー場と関係性を醸成する言語資源ー」小林隆編『コミュニケーションの方言学』pp.319-337. ひつじ書房.
- 三宅和子 (2019).「モバイル・メディアにおける配慮ーLINE の依頼談話の特徴ー」山岡政紀編『日本語配慮表現の原理と諸相』pp. 163-180. くろしお出版
- 宮寄由美 (2018).「LINE データベースの設計と属性情報付与の現状について」『言語資源活用ワークショップ発表論文集』3, pp.176-184.
- 山崎誠・高橋雄太・小木曾智信 (2025).「「現代日本語書き言葉均衡コーパス」の拡張ーBCCWJ2 の構築ー」『言語処理学会第 31 回年次大会発表論文集』pp.414-417.
- 楊虹・倉田芳弥 (2023).「LINE チャットの会話における感動詞の分析ー日本語母語場面と日韓接触場面の比較を通してー」『語用論研究』24, pp.79-88.
- Chen, Tao & Kan, Min-Yen (2013) Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus. *Language Resources and Evaluation*, 47:2, pp.299-355.
- Herring, Susan C. (1999) Interactional Coherence in Cmc. *Journal of Computer-Mediated Communication*, 4:4. <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1083-6101.1999.tb00106.x> (2025 年 8 月 16 日確認).

関連 URL

- | | |
|-------------------------|---|
| コーパス検索アプリケーション『中納言』 | https://chunagon.ninjal.ac.jp/ |
| BCCWJ2 -現代日本語書き言葉均衡コーパス | https://www2.ninjal.ac.jp/BCCWJ2/ |