LINEデータベースの設計と属性情報付与の現状について

著者	宮嵜 由美
雑誌名	言語資源活用ワークショップ発表論文集
巻	3
ページ	176-184
発行年	2018
URL	http://doi.org/10.15084/00001651

LINE[1]データベースの設計と属性情報付与の現状について

宮嵜 由美（国立国語研究所音声言語研究領域）

Fundamental Planning of LINE Database and Participant's Information

Yumi Miyazaki (National Institute for Japanese Language and Linguistics)

要旨

本稿では，現在構築中の「LINE データベース」の設計と現状について，①データ収集方法，②データ提供者と参加者の属性，③研究用データベースとしての加工を，具体例とともに報告した。2016 年から収集を始めた本 LINE データベースへの協力者は，2018 年 6 月時点で延べ 183 名，約 35,800 行[2]のデータである。

１．はじめに

本稿では，2016 年 4 月から 2018 年 6 月現在まで，筆者が収集した LINE のデータのデータベース構築の基本設計を紹介する。LINE は，文字メッセージの送受信の他に，通話，テレビ電話などの機能も有しており，従来のコミュニケーションツールでの言語生活の多くの部分を兼務する。研究対象としても，再現性を確保したデータベース構築の為，また機械的な検索も可能となるデータベース構築の為にどのような手続きが必要であるか，提案する。

２．データ提供の依頼手順と収集方法

２.１　データ提供の依頼手順

LINE データベース構築にあたり，まず，筆者の担当する大学の講義受講生数人にデータ提供依頼を行った。そのうち，データ提供を許諾した数人から，さらに提供者の紹介を受ける方法でデータ提供者を募った。この方法により，「ある人物が形成するコミュニケーションネットワーク」と「言語生活」の一端をうかがい知ることができると考えた。

なお，本稿では，実際に筆者とコンタクトを取りデータ提供を行った，直接のデータ提供元となる協力者を「データ提供者」と呼び，提供されたデータに登場する参加者については，「参加者」と呼ぶ。「参加者」のデータ提供許諾と匿名加工の範囲については，データ提供者から参加者に個別に確認してもらった。

知り合いの知り合い同士が知り合い，という事もあり，同一人物同士がそれぞれの LINE コミュニティにおいて，例えば二者間の場合と，三者以上が同じ画面（以下，トークルーム）でやり取りに参加する LINE 内（以下，グループ LINE）ではどのように互いを待遇しているか，違いは生じているのか，いないのか，その様子を観察する事もできる。

例えば，図１に示したデータ提供者 A の例をみると，データ提供者 A が参加するグループ LINE の構成員の一人とは，別のトークルームでの LINE のやりとりを行っているというケースがみられる。さらには，収集したデータの一部には提供者 A の母親との LINE デ

---

[1] 「LINE」は株式会社 LINE の商標，または登録商標です。

[2] 本稿では改行をもって１行とし，データも改行毎に１セルに入力されている。

miyazakiyumi@gmail.com

ータを含むが，その母親は息子との LINE の他，母親がひとりの女性として参加している友人 A との LINE データ提供があるというケースが存在する。それぞれが，それぞれの場（トークルーム）内で振舞いが変わるか，変わらないのか，という視点からの分析も可能である。

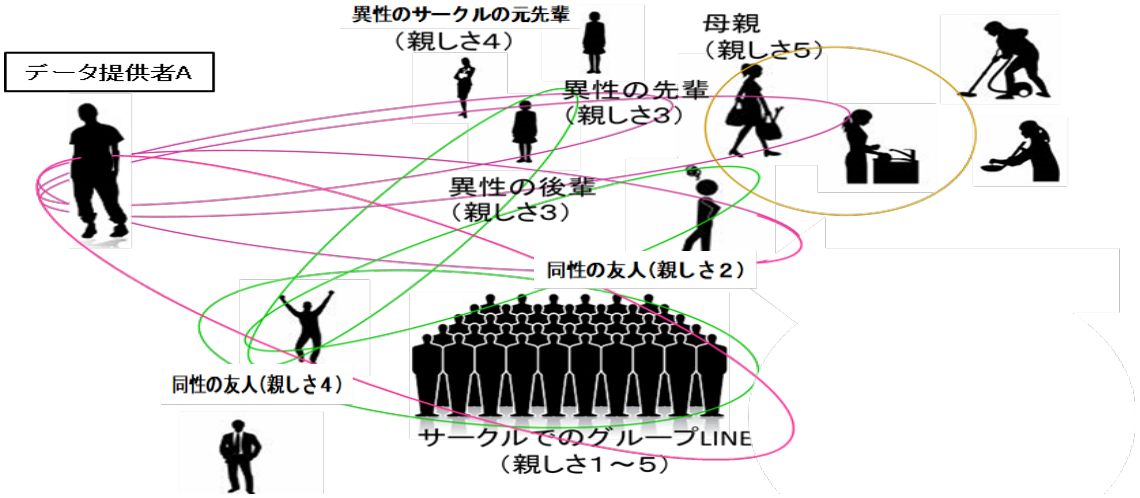


図１：LINE データ上のコミュニケーションネットワークの例（データ提供者 A の場合）

２.２　データ収集方法

LINE のコミュニケーションツールとしての特徴については三宅（2018）に詳しいが，今回データベースを整備するにあたり検討すべき点として，受信媒体の違いで文字化けが発生しないスタンプ[3]，画像，動画，通話機能などが文字メッセージとともに同じトークルーム内で連続性を持って送受信できるその様子を，検索可能な状態にどう再現するかという点にある。

収集したデータにも，スタンプのみの会話や，文字メッセージの送受信から突然音声通話に変わる，待ち合わせの場所は「位置情報」システムを使って教え合う，メッセージの送受信にどれだけの時間がかかったかといった既読・未読に関わる問題など，文字メッセージ以外の要素の再現すべき点は多々あり，それらも機械的な検索ができる限り可能となるよう考慮する必要がある。

これらの問題を勘案し，基本的には，LINE のデータテキスト化機能を使い，提供者任意の部分のみの①テキストデータを提供してもらうよう依頼した。さらに該当する部分の画面の②スクリーンショット画像を同時に提供するよう依頼した。

２.３　データ提供者の属性（2016 年 4 月～2018 年 6 月末時点）

【フェイスシート】

フェイスシートの質問項目は以下となっている。

Ａ．データ提供者自身に関する項目

１）名前・LINE 登録名・性別・年齢・職業・LINE 使用歴・出身地

Ｂ．LINE データ参加者に関する項目

１）名前・性別・年齢・職業・関係性・親しさ（データ提供者による記入）

[3] スタンプとは「LINE」で使用されるステッカーのような画像データを指す。

２）グループ LINE の場合はその「グループ名」と参加者全員の「LINE 登録名」

B, １）に示した親しさの判定については，データ提供者の視点から，５段階評価（とても親しい～全く親しくない）で評価されたものである。相手をどう認識し，待遇するかという問題であるが，この親しさの尺度以外にも，提供者と参加者の関係性の情報（サークルの同期や年上であっても職場の同僚など）も確認しており，この尺度はあくまで相手の待遇に関わる一つの要因として捉えていただきたい。

対面以外では携帯メールが主なコミュニケーションツールであった時代と同様（宮嵜：2004），親しい友人としか LINE はしない，という意見も聞かれた。しかし，実際にデータ収集を行ってみると，グループ LINE と呼ばれる複数人が同じトークルームに参加する場合などには，親しくないと判断される人物とのやり取りも少なくない（後述のグラフ１参照）。また，二者間の LINE でも，親しくない相手とのやり取りも行われており，今回のデータに含まれる。

２．４　データ提供者・参加者の属性：属性の流動性

2018 年６月末時点で整備が進んでいるデータ提供者・参加者の年齢と性別は表１の通りである。前述の通り，起点としたデータ提供者が友人同士という場合もあり，LINE コミュニティの形成にも重複がみられる。つまり，同一人物が複数の LINE に参加している場合があり，表１は延べ人数を示す。年齢は提供された LINE データ送受信時のものである。

さらに，データ提供者・参加者の職業と延べ人数を表２に示す。職業についても，LINE 送受信時のものである。同一人物が，それぞれの LINE データ送信時に別の職業に移行したケースがあり，表１の合計とは一致しない。

グラフ１に示したのは，データ提供者からみた参加者との親しさである。データ提供者の判断による協力者との親しさであり，データ提供者が異なると，同一人物であっても，データ提供者の評価によっては親しさが異なる場合もある。

表１ データ提供者・参加者の年齢と性別（人）

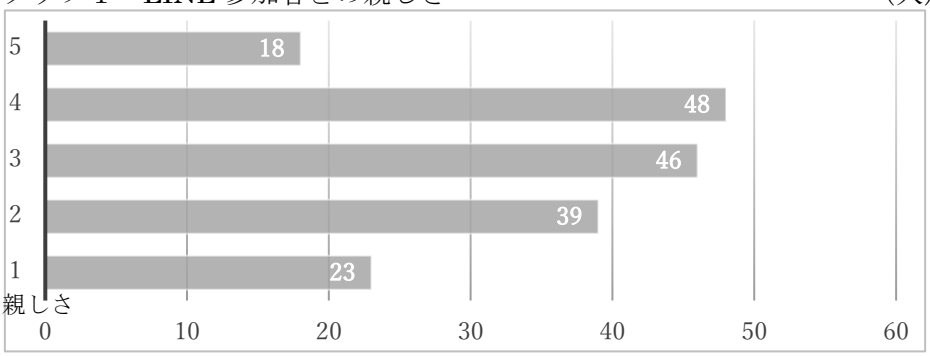
	男性	女性	合計
19 歳	6	4	10
20 歳	28	32	60
21 歳	30	20	50
22 歳	20	8	28
23 歳	5	4	9
24 歳	1	1	2
25 歳	1	0	1
54 歳	0	1	1
55 歳	0	5	5
56 歳	0	1	1
57 歳	0	5	5
58 歳	0	10	10
59 歳	0	1	1

データ送受信時の年齢は，筆者が大学生へのデータ提供依頼開始時に起点としたことから，現時点では大学生学部生に該当する年齢と，その親の世代の，大きく分けて２つの年齢層が存在するデータとなった。

表２ データ提供者・参加者の職業 （人）

学生	137
会社員	16
主婦／パート	9
アルバイト	8
専業主婦	6
介護士	3
教員	2
主婦／歯科衛生士	1
会社役員	1
教会牧師	1

グラフ１ LINE 参加者との親しさ （人）



３．研究用データベースとしての加工について

本調査では，LINE に備わっている「トーク履歴の送信」機能を使用し，文字列情報を忠実に再現できる収集方法を取った[4]。そこから研究用として匿名性やデータベースとしての検索性を考慮し，以下の加工方針に沿った xlsx 形式のデータベースを作成した。

また，スタンプや画像等の参照の為，本文と対応するトークルーム画像に対し，個人や場所等を特定できる部分を匿名加工した画面がポップアップされるようにした。

図２に匿名加工後のデータベースの例を示し，以下，入力規則について述べる。

３．１ データに付与される属性情報

＜データ属性＞

当該のトークルームの参加者の構成を示す。

① 二者間 LINE，10 代・20 代女性同士の例：ＦＦ １２００３

[4] 提供されたデータの一部には，画像データのみの提供もあり，手作業でのテキスト化を行ったデータも部分的に含んでいる。

② グループ LINE，50 代女性同士の例：G　F F　５５　００１
（実際のデータに空白は含まない）

左から，提供者が男性の場合には「M」女性には「F」を付与。冒頭に G が付く場合はグループ LINE である事を示し，次に提供者の性別・協力者の性別（グループ LINE の場合は参加者の性別構成）を示す。

数字の一桁目は提供者，二桁目は参加者の，それぞれ提供された LINE を送受信した時期の年齢の最小値と最大値を示す。下三桁はデータベース全体におけるトークルームの通番を示す。

よって，①の場合は，10 代女性と 20 代女性間のトーク履歴であることを示し，②の場合は，50 代女性間のグループラインのトーク履歴であることを示す。

図２ LINE データベース加工例

	データ属性	通番	管理ID	性別	年齢	職業	関係	親しさ	送信日	送信時間	吹き出し内行数	本文
2	FF12003	1	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:37		<Lf003>ですお久しぶりです
3	FF12003	2	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:37		ブロックしないでね(ノω`)
4	FF12003	3	Rf001	2	19	学生	Lf003の友人	4	2016/09/03(土)	8:41		<Lf003>ーーーーーー
5	FF12003	4	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:41	1	<人名>から勝手にもらいました～
6	FF12003	5	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:41	2	連絡もせずに消えててごめんね。
7	FF12003	6	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:41	3	わたしのLINEはあんまり回さないでもらえると嬉しいです(*_*)
8	FF12003	7	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:42		やっぱ早いね既読
9	FF12003	8	Rf001	2	19	学生	Lf003の友人	4	2016/09/03(土)	8:42		うん、わかった、大丈夫だよ！
10	FF12003	9	Rf001	2	19	学生	Lf003の友人	4	2016/09/03(土)	8:42		ちょーど目覚まし止めた(笑)
11	FF12003	10	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:42		おはよう[絵文字]
12	FF12003	11	Rf001	2	19	学生	Lf003の友人	4	2016/09/03(土)	8:42		[スタンプ]<ギュー･･>
13	FF12003	12	Rf001	2	19	学生	Lf003の友人	4	2016/09/03(土)	8:42		おはよう
14	FF12003	13	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:42		本当ごめんね
15	FF12003	14	Rf001	2	19	学生	Lf003の友人	4	2016/09/03(土)	8:43		大丈夫だよ(笑)
16	FF12003	15	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:43		<人名>から<Rf001>と<人名>が心配してるって聞いて、申し訳なくなった
17	FF12003	16	Lf003	2	19	浪人生	Rf001の友人	4	2016/09/03(土)	8:44		LINEは最近始めた！

＜通番＞
同トークルームにおいて，時系列で並べた場合の本文の通し番号を示す。

＜管理 ID＞
本文の送信者の管理番号を示す。
① 提供者 A の例：Rf001　　② 参加者 B の例：Lf002

①の例はそれぞれ，左から，トークルーム右側[5]「R」の吹き出し，女性，提供者通番を示す。
②の例はそれぞれ，左から，トークルーム左側「L」の吹き出し，女性，参加者通番を示す。

＜性別[6]＞
男性には１，女性には２を付与。

[5] 図２を参照されたい。
[6] 性別，年齢，職業については，公開を希望しない場合もあり，その場合は X が付与される。

＜年齢＞
本文送受信時の年齢を示す。（年齢の情報はあるが，生年月日の情報がない場合は年度で判断）

＜職業＞
本文送受信時の職業を示す。ただし，表２に示したように筆者がコーディングした。

＜関係＞
データ提供者が判断した参加者との関係を示す。
① 提供者 C（Rf004）の例：（Lf003 の）サークルの後輩
② 参加者 D（Lf003）の例：（Rf004 の）サークルの先輩

＜親しさ＞
データ提供者が判断した参加者との親しさを示す。親しさの判定については，データ提供者の視点から，５段階評価（とても親しい～全く親しくない）で評価されたものである。尺度の捉え方については，前述の【フェイスシート】に詳しい。

＜送信日＞
本文の送信日を示す。

＜送信時間＞
本文の送信時間を示す。

＜吹き出し内行数＞
LINE 本文は，「吹き出し」画像内に提示され送られる。どの文字列，絵記号までをひとつの吹き出し内に収めるかは，送信者の任意による。
ひとつの吹き出し内に複数の改行がある場合，改行毎にセルを分け，出現順に番号を付与。改行が存在しない場合は何も入力しない。

３．２本文セル内の加工規則

＜本文セル＞
以下，付与する記号類はすべて全角とする。
① １吹き出し内，１行を１セルに入力。
② １吹き出し内に複数の改行による文字列が入力されている場合は，改行毎に１行下に入力。
③ １吹き出し内に複数の改行による文字列が入力されている場合は，「吹き出し内行数」列に改行毎に通し番号を付与。改行がない場合は何も記入しない。
④ 登場した人物や場所，施設名称については，それぞれ全角＜＞で囲い，＜人名＞，＜地名＞，＜施設名＞とする。
⑤ 元データ画像に「▷」マークが付与され，送信者の位置情報が示されている場合は，［位置情報］[7]と記入。

[7] 図２を参照されたい。

⑥ [8]日時が不明の場合は，文頭に▲を付与し　▲不明瞭　と記載。

⑦ 本文が不明瞭の場合は，文頭に▲を付与し　▲不明瞭　と記載。
一部推測できる場合は，文頭に▲を付与し，全角「」で括り本文を入力する。
例：▲「おお！！」

絵文字

① 絵文字については，半角[]で括り，[絵文字]と入力。
例：　きいてないよー  ！！　→　きいてないよー[絵文字]！！

顔文字

① 顔文字（記号の組み合わせによるもの）については，できる限りそのまま入力する。

② 再現が難しい場合は，半角[]で括り，[顔文字保留]と入力。

スタンプ

① スタンプは，半角[]で括り，[スタンプ]と入力。１スタンプ毎に１セルに入力。

② スタンプに文字列が付与されている場合は，[スタンプ]の後に文字列のみ，全角＜＞で括り文字列を入力する。　例：[スタンプ]＜さすけねー＞

③ 動くスタンプと確認できた場合は，[スタンプ]の前に，全角【】で括った【動】を入力する。記号は全角。
例：【動】[スタンプ]
例：【動】[スタンプ]＜おはようございます！＞

④ 「  」マークが付与された音の出るスタンプの場合は，[スタンプ]の前に，全角【】で括った【音】を入力する。記号は全角。どのような音が出ているかわかる場合は，[スタンプ]記号の後に＜＞で括り文字列を入力する。音声が確認できない場合は，＜＞内に＜音声不明＞と入力する。
例：【音】[スタンプ]＜さすけねー＞
例：【音】[スタンプ]＜音声不明＞

文字フォント（LINE 特有の文字スタンプ）

①  　のような，LINE 特有の絵文字フォントの場合，開始部[文字スタンプＳ]と終了部[文字スタンプＥ]で括る。ブラケットは半角，アルファベットともに全角。

② 文字スタンプが１吹き出し内の文中の一部にある場合
例：でさー，[文字スタンプＳ]ありえない[文字スタンプＥ]わけ！

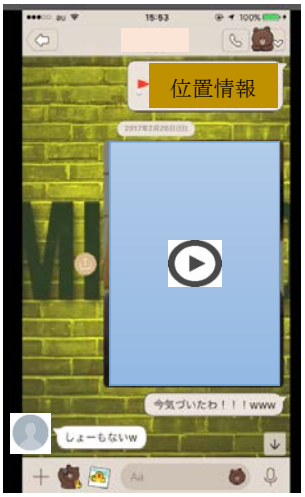
写真

写真やスクリーンショットの画像の場合，半角 [] で括り，[写真]と入力する。写真毎に１セルに入力。

8 ＜本文セル＞⑤, ⑥, 顔文字②については，画像ファイルのみ提供されたデータに付与したものである。

動画

写真に ▶ が付与されている場合は動画である。その場合半角［］で括り，［動画］と入力。動画毎に１セルに入力。



Rm025	20:26		おけー
Rm025	21:28		着いた
Lm026	21:35		どこいるの？
Lm026	21:36		☎ 不在着信
Rm025	21:50		☎ 通話時間 0:54
Rm025	21:50	1	[位置情報]
Rm025		2	
Rm025	18:26		[動画]
Rm025	18:26		今気づいたわ！！！www
Lm026	20:51		しょーもないw
Rm025	15:51		今夜地元で＜人名＞さんと飲むんだけど来る？？
Lm026	16:14		いくわ
Lm026	16:14		何時から？
Rm025	16:15		まだ決めてないー
Rm025	16:15		前から話してた、＜地名＞のホルモンの店

図３　動画，通話，不在着信，位置情報通知の処理例

通話

① 通話が行われた場合は「☎」マークを挿入する。半角の空白を挿入後，同セル内に通話時間を記入。

② 不在着信の場合は，「☎」マークを挿入し，半角の空白を挿入後，不在着信と記入。

３．３文字列以外の情報参照のためのポップアップ機能

図４　ポップアップ画像の加工例

トークルーム画像データは，図２，３，４に示すように個人情報に匿名化の処理を施し，ポップアップ画像で参照できるようにした。

検索機能を使用する際は xlsx 画面の本文列を参照し，スタンプや絵文字等，文字列以外の情報については該当部分の画像番号をクリックすることで，ポップアップ画面が別途立ち上がり参照できる（図２）。

その他画面左脇に表示される参加者のアイコン，写真や動画画面，人名等個人が特定できる情報，位置情報なども加工した。人名等は，既に管理 ID が付与されている人物の場合はその番号が付与され，それ以外の人物には＜人名＞が付与される。

４．おわりに

本稿では，2018 年 6 月末時点で収集したデータと，研究用データベース加工の際に付与した属性情報やデータにおける個人情報の加工の方針とともに，データの概要を紹介した。

本稿報告の通り，現時点で整備が行われているのは，主に 20 代の大学生を中心とした LINE コミュニティのデータであるが，現在 30 代，40 代の提供者とそのコミュニティにおける LINE のデータ収集を行っている。本発表で得た助言を元に引き続きデータ整備を行う。

謝　辞

本研究は国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（代表：小磯花絵），JSPS 科研費 16K02714（代表：宮嵜由美）の助成を受けたものである。

文　献

岡本能里子(2016)「雑談のビジュアルコミュニケーションーLINE チャットの分析を通して」村田和代・井出里咲子『雑談の美学』pp.213-236,ひつじ書房

加藤安彦(2007)「ケータイメイルにおける顔文字と記号の出現頻度とその関係ーケータイメイルコーパスの紹介とともにー」『専修国文』81 巻，pp.1-17，専修大学日本語日本文学会

三宅和子(2018)「LINE の中の「方言」ー場と関係性を熟成する言語資源ー」小林隆編『コミュニケーションの方言学』第 14 章 pp.319-337, ひつじ書房

宮嵜由美(2004)『場面における言語行動のストラテジーの考察 ー携帯メールを中心にー』東京都立大学人文科学研究科国文学専攻修士論文

宮嵜由美(2015)「LINE を用いた依頼場面における送受信者の言語行動ー表現の担う機能と構造に着目して」西尾純二他編『言語メディアと日本語生活の研究』pp.5-20,大阪府立大学人間社会学部／大学院人間社会学研究科

宮嵜由美(2017) 「LINE を使用した依頼：2 者間・3 者間での受け手のフォローと共話性」第 21 回「ひと・ことばフォーラムーSNS の教育・研究の可能性についてー」ひと・ことば研究会

水谷信子(2001)「あいづちとポーズの心理学」『言語』第 30 巻第 7 号 pp.47-51,大修館書店