一般公開を視野に入れた「携帯メイルコーパス」整備の試み

ー加藤安彦氏の遺志を受けてー

宮嵜由美(明治大学)　林直樹(日本大学) 田中ゆかり(日本大学) 三宅和子(東洋大学)

1. はじめに

本発表は，故加藤安彦氏が 2004 年よりデータを収集し，構築を行っていた「Senshu-Univ. Keitai-mail Corpus」（以下，SKC と仮称）を，同氏の遺志により引き継いだ発表者らが，一般公開[1]を視野に入れ 2016 年より再整備を行った経過報告である．

加藤氏は国語研究所在任中よりコーパスの重要性を提唱し，SKC は「打ちことば」コーパスの実現を目指したものとして位置づけられる．具体的には，2004 年より構築が開始され，xlsx 形式によってまとめられたおおよそ 10 年間（メイル[2]送受信時期は 2002 年～10 年），延べ 262598 通，約 463,000 行に及ぶデータである．

SKC の基本設計は加藤（2004，2007）に詳しい．本発表では，今後広く研究者に公開するため検索性と再現性を主眼に再整備を行った際に改善した点を示し，研究の一例として，頻出語彙が親疎によってどのように異なるかを調査した結果を報告する．

2. 「Senshu-Univ. Keitai-mail Corpus (SKC)」の意義

現在日本では，「話しことば」では『日本語話し言葉コーパス』『日本語日常会話コーパス』や『多言語母語の日本語学習者横断コーパス』など，「書きことば」では，現代日本語書き言葉均衡コーパス，『日本語歴史コーパス』等々，共時的にも通時的にもコーパスの整備が進んでいる．しかし，2000 年以降，日本語社会における主要なコミュニケーション媒体の一つとなった「打ち言葉」を記録し，コーパス化されているものは存在していない．「打ちことば」は，この 20 年間の日本の言語生活の大部分を占めるようになってきたが，すでにスマートフォンにシフトしてしまった現在，携帯メイルの記録は言語資源としても非常に貴重なものである．

SKC が整備・一般公開されれば，近年年齢を問わず利用が謙虚になってきた LINE によるコミュニケーションを収集したコーパス（例えば宮嵜 2018）などとの比較も可能になる．

収録されているデータは，実際に送受信された自然データである点でも，日本語学や社会言語学以外の他分野で利用が可能であり高い資料性をもつと位置づけられる．

3. SKC データ入力概要

3.1 SKC データ入力方針　-全体を通して-

発表者らが加藤氏から引き継いだ原 SKC は，2004 年度から 2010 年度にかけて，専修大学文学部加藤安彦ゼミナールにおいて収集された，データ入力者（データ属性では「送受信主体者」）とデータ提供者間において公開承諾を得た自然データが収録されている．

入力は，加藤（2004，2007）を元に，ゼミ生による手入力であった．また，経年により新たに追加されたメイル機能に付与する記号の検討と統一など，多様な背景を持つ研究者への一般公開に向け再度整備を行う必要があった．再整備後，現在以下の情報が付与されている．

➢メイルの送受信時の情報

A 列：（入力）年度　B 列：通し No.（本文行の通し番号．年度毎）　O 列：送受信年月日　P 列：送受信時刻

➢メイル送受信主体者（以下，データ入力者）の情報

C 列：送受信主体者生年月日　D 列：（データ提供者の）出身地　E 列：性別　F 列：送受信主体者携帯会社名　G 列送信／

[1] 同氏の遺志により「言語資源協会」での公開を予定している．

[2]本研究では故加藤氏の用いた表記「携帯メイル」を使用する．

受信（データ提供者が送信者であったか，受信者であったか）

➢データ受信者の情報

H 列：相手携帯会社名　I 列：相手 ID　J 列：相手生年月日　K 列：相手出身地　L 列：相手性別

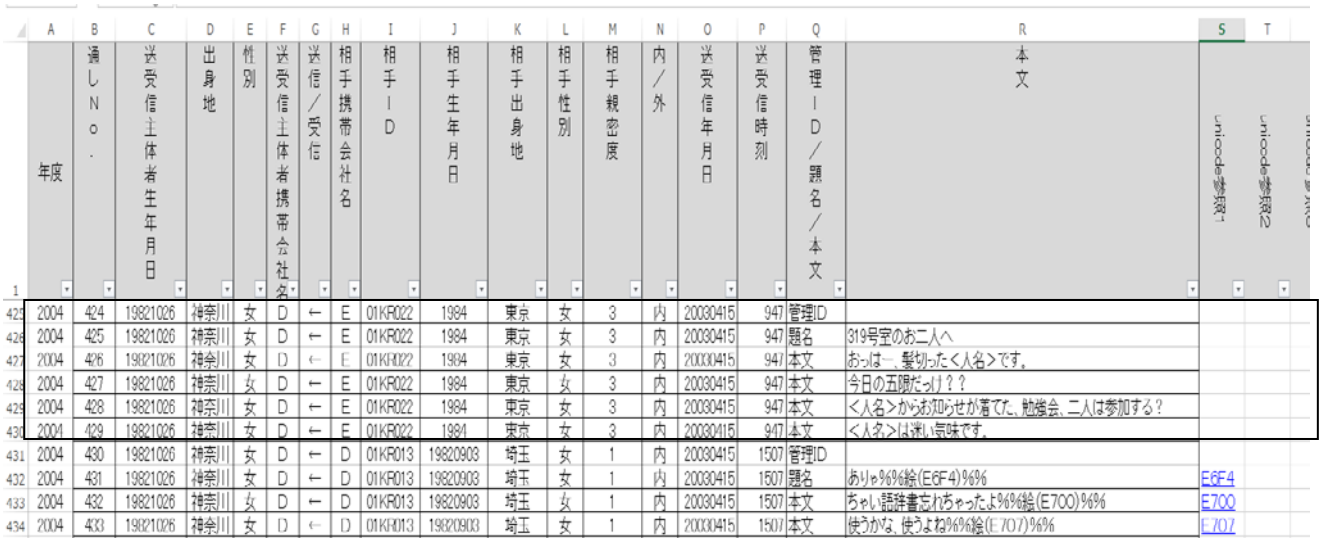
M 列：相手親密度（提供者による判断）　N 列：内／外（相手が同大学内の関係であるかどうか）

➢メイル本文に関する情報

Q 列：管理 ID／題名／本文　R 列：本文　S 列以降：絵文字 Unicode 参照行

個人情報はデータ提供者による公開を前提とした任意の部分であるため，例えば，生年月日などは，生年のみの入力もみられる．同様に男女の別が入力されていないケースもみられる．

１通のメイルの構成例として，下記図１の四角で囲んだ「通し No.」424 から 429 を挙げる．メイル開始行には Q 列に「管理 ID」が入力されており，次に題名行，本文行が続く．本文は，元データに改行が挿入されていた場合は行を変えて入力される．従って，１通のメイルは「管理 ID」行をはじめに，「通し No.」を目安に管理される．以上の入力規則によって整備された年度毎メイルデータ数を表１に示す．



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	年度	通しNo.	送受信主体者生年月日	出身地	性別	送受信主体者携帯会社名	送信／受信	相手携帯会社名	相手ID	相手生年月日	相手出身地	相手性別	相手親密度	内／外	送受信年月日	送受信時刻	管理ID／題名／本文	本文	unicode参照1	unicode参照2
425	2004	424	19821026	神奈川	女	D	←	E	01KR022	1984	東京	女	3	内	20080415	947	管理ID			
426	2004	425	19821026	神奈川	女	D	←	E	01KR022	1984	東京	女	3	内	20080415	947	題名	319号室のお二人へ		
427	2004	426	19821026	神奈川	女	D	←	E	01KR022	1984	東京	女	3	内	20080415	947	本文	おっは一、髪切った<人名>です。		
428	2004	427	19821026	神奈川	女	D	←	E	01KR022	1984	東京	女	3	内	20080415	947	本文	今日の五限だっけ？？		
429	2004	428	19821026	神奈川	女	D	←	E	01KR022	1984	東京	女	3	内	20080415	947	本文	<人名>からお知らせが着てた、勉強会、二人は参加する？		
430	2004	429	19821026	神奈川	女	D	←	E	01KR022	1984	東京	女	3	内	20080415	947	本文	<人名>は迷い気味です。		
431	2004	430	19821026	神奈川	女	D	←	D	01KR013	19820903	埼玉	女	1	内	20080415	1507	管理ID			
432	2004	431	19821026	神奈川	女	D	←	D	01KR013	19820903	埼玉	女	1	内	20080415	1507	題名	ありゃ%%絵(E6F4)%%	E6F4	
433	2004	432	19821026	神奈川	女	D	←	D	01KR013	19820903	埼玉	女	1	内	20080415	1507	本文	ちゃい語辞書忘れちゃったよ%%絵(E700)%%	E700	
434	2004	433	19821026	神奈川	女	D	←	D	01KR013	19820903	埼玉	女	1	内	20080415	1507	本文	使うかな、使うよね%%絵(E707)%%	E707	

図１　SKC データ属性付与状況

表１ 各年度メイル数　　(通)

年度	2004	2005	2006	2007	2008	2009	2010
メイル数	17461	35755	54935	20438	33394	48812	51803

3.2 SKC データ入力方針　-再整備点-

今回の再整備において変更を行った点は，大きく以下の４点である．

①各年度 A～R 列の名称統一と，本文行に含まれる絵文字・顔文字を機械的に検索する際に使用される記号の統一

②顔文字以外の文字記号「形[3]」を含む記号の削除　「%%形（orz）%%」→再整備後「%%（orz）%%」

③本文には絵文字画像に対応する Unicode が入力されている．再整備時にはこれらを「Unicode 参照」として Q 列以降に出現順に抽出．xlsx 画面と同期した画像とのリンクを貼った．

画像リンクの具体例として表２に 2004 年 12 月 24 日 10 時 57 分，1982 年生まれの女性（愛知県出身）に対し 1942 年生まれの女性．親密度２）から送られたメイルを示す．Unicode 参照欄にはき出された Unicode には画像が個別にリンクされており，検索画面上にポップアップされる．

表２ 1942 年生まれ女性（福島県出身）のメイル送信履歴

題名	おはようございます%%絵（Ｅ０３３）%%	Ｅ０３３

[3] 「形」について，加藤(2007)pp.6-7 によれば，「意味を持たされた記号の組み合わせに対しては，すべて「%%(」と「)%%」でくくり，「形」の文字を付与しておく.」とあり，さらに詳細な定義がなされているが，付与基準に対し入力者による揺れが顕著であるため，再整備時に「形」は削除した．

本文	昨夜は２時に帰宅して起きたらこんな時間でした%%顔（f^_^;）%%	
本文	今＜＜人名＞さん＞からのプレゼント開けました%%絵（Ｅ１１２）%%	Ｅ１１２
本文	何やら箱も小物入れかティッシュケースに活用できるみたいだし，タオルも私好みな色柄でしたよ～☆	
本文	有難うございましたっ%%絵（Ｅ０５６）%%	Ｅ０５６

Unicode の入力も手入力によるものであるため，「おやすみ%%絵（9）%%」のように，Unicode 自体の判別が難しいものがある．このような場合，参照列に「未処理」を入力，現在対応する画像が確認できない場合は「現在不明」画像がポップアップされるよう変更を行った．

④「デコレーションメール」機能や「フレーム」機能の処理（例 1，2 下線部参照）

例 1：「残念%%デコメ（(φ)(リラックマ)(落ち込む)）%%＜人名＞も行かないって%%絵（ＥＡＣ０）%%」

例 2：「%%フレーム(ＨＡＰＰＹ ＢＩＲＴＨＤＡＹ to you) %%)」

デコレーションメールは 2009 年度，2010 年度のデータにみられたが，統一された入力規則がなかったため，「%%デコメ () %%」で囲み，括弧内に既に入力された情報をそのまま挿入する方針を取った．ただし画像は特定できないため，Unicode 参照欄にはその出現順と出現頻度のみ「デコメ」として抽出した．

2010 年度に 2 件のみ見られたフレーム機能も同様の処理を行った．

3.3 メイル送受信者の概要

１通の目安となる「管理 ID」に入力されている SKC 収録のメイルの送受信者の概要を説明する．

データ入力者の生年は，1982 年から 1989 年である．メイル相手の生年には戦前の 1932 年，1933 年，1942 年生まれがおり，戦後は 1947 年から 1950 年代，60 年代，80 年代，そして 1996 年生まれのデータとなっている．

送受信者間における男女の別を表 3 に示す．また，表 4 には「性別」とデータ入力者によって判断された「相手親密度[4]」の別を示す．それぞれ，属性情報の付与されていないデータは除いた．

表 3　送受信者間性別　（通）

		送信者	
		男	女
受信者	男	12681	27000
	女	89868	12680

表 4 男女別送受信者との親密度　（通）

	親密度１	親密度２	親密度３	合計
男	8986	3180	514	12680
女	57353	18765	6250	82368

4.　親密度別高頻度語（上位語）

本稿では，親密度 1 と 3 と判断されたメイルデータの中から空白行を除き，ランダムサンプリングにより各年度 300 行を抽出した．それぞれの高頻度語[5]を表 5 に示す．算出した高頻度語では，語彙的形態素（名詞・動詞・形容詞）のみを取り上げたが，形態素解析ソフトによっては名詞と感動詞との分別が揺れるものがあるため，本稿では感動詞も含めた．また，携帯メイルにおける「打ちことば」の文体は，いわゆる「書きことば」とは違い，表記の揺れが激しい．そこで，表 5 では語彙素読みを挙げ，必要に応じ括弧内に相当する漢字表記を付与した．

親密度による大きな違いは見られず，親密度 1 と 3 ともに最頻出語は「人名」である．ただし，親密度 3 の場合には，相手[6]を呼び捨てにせず，「サン（7 位），チャン（16 位）」が付与されることが多い．その他，親密度 1 に頻出し親密度 3 には頻出しない語として，否定を表す「ナイ」や，今回欄外となったが 22 位に「バイト（29 件）」23 位に「ダイジョウブ（28 件）」などが挙げられる．「フィラー」は，親密度 1（「エ」が最多で 31 件出現／延べ 77 件），親密度 3（「ナンカ」が最多で

[4] 加藤（2007）p.10 によれば，親密度は「深刻な内容の話をすることのできる相手，家族や親友と呼べる関係の友人を「親密度 1」，「親密度 1」に比べると親しさは薄れるが，比較的行動を共にすることの多い仲のよい友人などを「親密度 2」とし，顔を知っていて，ことばも交わす普通の友人を「親密度 3」としてある．これは必ずしも厳密な分類ではないが，例えば顔文字を使用する頻度を分析すると，親密度による傾向が見てとれるということから有意味な情報であるといってよい．」との基準により，データ入力者によって判断されたものである．

[5] 表 5 における高頻出語は，絵文字・顔文字・パラ言語表現等に関連する記号類は除いた．具体的には数字，絵文字の一部，「（笑）」，これら表現を囲むデータ整備上の記号類である．

[6] 話題の人物が，送受信者間であるか，第三者であるかは今回分析の対象としていない．

10 件出現／延べ 36 件）の約 2 倍の使用がみられた.

表 5　親密度別高頻度語（上位 20 語）

順位	親密度1		
	語彙素読み(漢字表記)	品詞	語数
1	ジンメイ(人名)	名詞	189
2	シ	動詞	166
3	テ	動詞	83
3	テル	動詞	83
5	イイ	形容詞	77
6	ノ	名詞	71
7	キ(着／来)	動詞	64
8	キョウ(今日)	名詞	63
9	ワタシ	名詞	57
10	イッ(言／行)	動詞	55
11	イ(居る)	動詞	54
12	アシタ(明日)	名詞	52
12	ジ(時)	名詞	52
14	スル	動詞	49
15	コト(事)	名詞	48
16	ニチ(日)	名詞	45
17	アル	動詞	43
18	ゴメン	感動詞	39
19	ナッ	動詞	38
20	イマ(今)	名詞	37
20	ナイ	形容詞	37

順位	親密度3		
	語彙素読み(漢字表記)	品詞	語数
1	ジンメイ(人名)	名詞	211
2	シ	動詞	209
3	テル	動詞	92
4	ニチ(日)	名詞	74
5	キョウ(今日)	名詞	67
6	テ	動詞	64
7	イイ	形容詞	61
7	サン	名詞	61
9	イ(居る)	動詞	60
10	アシタ(明日)	名詞	56
11	ワタシ	名詞	54
12	コト(事)	名詞	50
13	ナッ	動詞	49
14	キ(着／来)	動詞	48
15	ジ(時)	名詞	45
16	スル	動詞	40
16	イマ(今)	名詞	40
16	チャン	名詞	40
19	イク(行く)	動詞	39
19	イッ(言／行)	動詞	38
20	クダサイ	動詞	38

4. おわりに

2018 年, 2019 年に筆頭執筆者が担当する私立女子大学（東京都内）のある授業時の「親しい友人への依頼」という課題に対し, 約 700 名の半数近くが選択したコミュニケーション媒体は「直接会う」であった. 携帯メイルの使用が一般化しはじめて 20 年を経た現在でもなお, 対面時に得られる非言語の要素を排したツールでのコミュニケーションに困難を感じている者は少なくない. (携帯) メイルや LINE など,「打ち言葉」によるコミュニケーションは, その多くの場合, 対面や通話と連続的に位置し, それぞれの特性と不足を補い合って我々の言語生活に存在していると考えられる.

なお, 形態素解析にあたっては, 前述の通り「打ちことば」の表記のゆれの問題がある. 例えば,「ありがとう」を伝える際の特殊拍の記述の有無,「ありがとーう」「ありがと～う」「ありがとっ」「ありがとん」「サンキュ」, 絵文字との組合せ「🙇 がとう」, 敬体の場合は「あざっす」など, その処理と位置づけについては「話しことば」同様今後の課題となる.

5. 謝辞

本研究は JSPS 科研費 16K02714, JSPS 科研費 18H00680, 平成 28 年度・29 年度日本大学文理学部人文科学研究所総合研究費, 平成 30 年度・31 年度日本大学学術研究助成金社会実装研究の助成を受けたものである。

参考文献

加藤安彦 編(2004)「加藤ゼミ携帯メイル入力マニュアル」『加藤ゼミ研究報告集１ - 携帯メイルによるコミュニケーション研究 - 2004 年度』pp. 7-10 私家版

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 (2019)「モバイル・メディアにおける配慮-LINE 依頼談話の特徴-」山岡政紀編『日本語配慮表現の原理と諸相』第 11 章 pp. 163-180

宮嵜由美(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 2018 発表論文集』3 巻 pp. 176-184 言語資源 WS2018　国立国語研究所 学術情報リポジトリ(http://doi.org/10.15084/00001651)

田中ゆかり(2019)「5-2 携帯メイルコーパス（限定公開版）に基づく絵文字の出現状況」荻野綱男他「コーパス言語学の学際的研究」『研究紀要』第 97 号 pp. 284-289 日本大学文理学部人文科学研究所