2023 年 5 月 4 日

「加藤安彦ケータイメイルコーパス」

宮嵜由美（MIYAZAKI Yumi）

１．「加藤安彦ケータイメイル[1]コーパス」の構築と整備経緯

本コーパスは，故加藤安彦氏が 2004 年より 2010 年にかけて専修大学文学部日本語日本文学科加藤安彦ゼミナールにおいてデータ収集・構築を行っていたケータイメイルコーパスを，2016 年以降，同氏の遺志を引き継いだ整備者ら（田中ゆかり，三宅和子，宮嵜由美，林直樹）が一般公開を視野に入れ，整備したものである．

コーパス名はご遺族の意思を尊重し，『加藤安彦ケータイメイルコーパス』，（英名）『Kato Yasuhiko Mobile-mail Corpus』[2]とした．権利者，代表者，整備に関わる諸情報は 5.1 に記す．

２．データ概要

2.1　データ総数

収録されているデータは 2004 年より収集，構築が開始され（メイル送受信時期は 2001 年～2010 年），xlsx 形式によってまとめられたおおよそ 10 年間，延べ 271,598 通，462,874 行に及ぶデータである．

2.2　データ入力方針

データ入力は加藤（2005，2007）を基本方針とし，加藤ゼミ生による手入力（データ属性における「送受信主体者」）を原則として踏襲した．入力されたデータ総数等は以下となっている．

表１　年度ごと収録メイル数　（通）

年度	2004	2005	2006	2007	2008	2009	2010
メイル数	17,461	35,755	54,935	29,438	33,394	48,812	51,803

[1] 本コーパスでは、故加藤氏の用いた表記「ケータイメイル」を使用する．
[2] ケータイメイルの英名は Miyake(2007)を参考にした．

表２　送受信者の性別（延べ）　(通)

		送信者	
		男	女
受信者	男	12,681	27,000
	女	89,868	12,680

2.3 親密度の判定基準

3.1「送受信者間の関係」で記すように，本コーパスには「親密度」情報が付与されている．メイル送受信者間の親密度の判定は，以下加藤(2007)の基準をもとに，データ提供・入力者により行われている．

> 親密度は三段階に分け、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。
>
> 加藤（2007:10）

表３　男女別親密度（延べ）　(通)

	親密度１	親密度２	親密度３	合計
男	8,986	3,180	514	12,680
女	57,353	18,765	6,250	82,368

３．公開データ概要

3.1「第一次整備後」のデータ付与情報

公開を目指して整備を行った加藤安彦ケータイメイル研究会は，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を認識し，最終的には表４，表５の情報を付与し整備した．

情報に関わる数字，アルファベット，タグ記号類はすべて全角英数で統一した．表４で示した規則通りに入力されていなかったもの，未入力となっていたものは，基本的に「不明」を入力した．

表４　公開コーパスデータ付与情報

１通のメイルの構成目安	A列：年度（入力年度） B列：通し No S列：以下のいずれかを入力. 管理 ID（１通のメイル冒頭部） 題名[3] 本文
メイル本文	T列： ①題名が空白の場合には空白. ②元データに改行が挿入されていた場合は改行して入力.
Unicode のリンク[4]	U列以降： T列（本文列）内 Unicode に該当する画像を出現順に記載．クリックにより画像がポップアップされる． 本文処理と本列以降の処理の詳細は表５.
データ入力者情報 （データ提供・入力者）	C列：送受信主体者生年月日 データ提供・入力者（以下入力者[5]）は，「送信者」とも「受信者」ともなり得るため，「送受信主体者」と記す. D列：出身地 E列：出身地（国） 国内／国外 F列：性別 G列：送受信主体者携帯会社名： D : docomo, V : Vodafone/Softbank, E : au/TU-KA, P : PC, H : PHS, W : WILLCOM, 不特定多数の場合：DIV，メイリングリスト：ML
メイル送・受信者情報	H列：送信／受信 矢印の方向が左向きのものは，送受信主体者（入力者）が「受信者」であったメイルであり，「送信者」であった場合は右向きの矢印となっている.
メイル相手情報	I列：相手携帯会社名

[3] 題名に「Re:(Re>)」がついているもの，または「Re:(Re>)」の後の文が変わっていないときは，無題の時と同様とする.

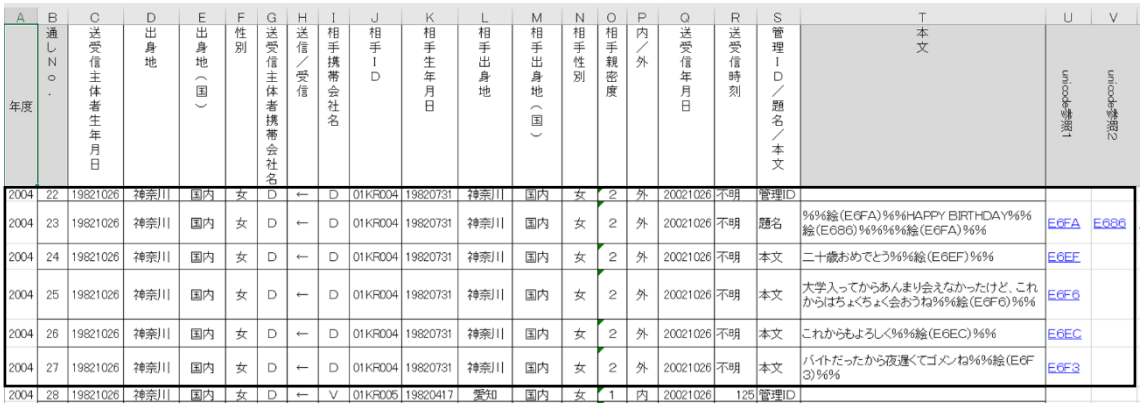
[4] Unicode のデータ内での再現方法は宮嵜(2018)を参考にした．Unicode の挿入・整備，ポップアップ機能の付与は本コーパス整備において最も労力を要した作業であった．元データ１行目情報名のゆれの統一など，訂正の詳細を表５に記載した.

[5] 同ゼミメンバーとのやり取りの場合は，データが重複しないよう受信メイルのみを入力.

	Ｊ列：相手 ID[6] K列：相手生年月日 Ｌ列：相手出身地 M列：相手出身地（国） 　　国内／国外 N列：相手性別 ①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，不完全な生年月日の記入や男女の別が未入力の場合はそのままとした． ②相手が「DIV（不特定多数）」の場合，K～N列，O列は空白となっている．
送受信者間の関係	O列：相手親密度 入力者の判断により１から３まで付与．その基準と内訳は本文中の 2.3 に記載． Ｐ列：内／外 メイル送受信相手とは専修大学内における関係か，学外における関係か．
送受信年月日・時刻	Q列：送受信年月日．８桁で記載． R列：時間と分を分ける「:」は省き 24 時間表示の 4 桁で入力されていた．セルは「時刻表示」設定になっているものの，値の意が不明であったものは，整備にあたり「不明」を付与した．

[6] 相手 ID で「7.54E+03」と表示されるものは，実際には「7540」と入力されている．検索時には注意されたい．

表４の情報を付与したデータの具体例を図１に示す.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
年度	通しNo.	送受信主体者生年月日	出身地	出身地（国）	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	相手出身地（国）	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
2004	22	19821026	神奈川	国内	女	D	←	D	01KR004	19820731	神奈川	国内	女	2	外	20021026	不明	管理ID			
2004	23	19821026	神奈川	国内	女	D	←	D	01KR004	19820731	神奈川	国内	女	2	外	20021026	不明	題名	%%絵(E6FA)%%HAPPY BIRTHDAY%%絵(E686)%%%%絵(E6FA)%%	E6FA	E686
2004	24	19821026	神奈川	国内	女	D	←	D	01KR004	19820731	神奈川	国内	女	2	外	20021026	不明	本文	二十歳おめでとう%%絵(E6EF)%%	E6EF	
2004	25	19821026	神奈川	国内	女	D	←	D	01KR004	19820731	神奈川	国内	女	2	外	20021026	不明	本文	大学入ってからあんまり会えなかったけど、これからはちょくちょく会おうね%%絵(E6F6)%%	E6F6	
2004	26	19821026	神奈川	国内	女	D	←	D	01KR004	19820731	神奈川	国内	女	2	外	20021026	不明	本文	これからもよろしく%%絵(E6EC)%%	E6EC	
2004	27	19821026	神奈川	国内	女	D	←	D	01KR004	19820731	神奈川	国内	女	2	外	20021026	不明	本文	バイトだったから夜遅くてゴメンね%%絵(E6F3)%%	E6F3	
2004	28	19821026	神奈川	国内	女	D	←	V	01KR005	19820417	愛知	国内	女	1	内	20021026	125	管理ID			

図１　『加藤安彦ケータイメイルコーパス.xlsx』データ例

【具体例：図１，太枠で囲った部分参照】

＜１通のメイル構成目安＞

図１の四角で囲った部分に示す通り，１通のメイルであることを示す３つの情報［管理ID］［題名行[7]］［本文行］がＳ列に付与されている。図１の四角で囲った１通のメイルの具体的な送受信情報は以下の通りとなる。

2002 年 10 月 26 日　送受信時間不明

送受信者主体 1982 年 10 月 26 日生まれ（女性：神奈川県出身：docomo 使用）が，1982 年 7 月 31 日生まれ（女性：神奈川県出身：docomo 使用）に送ったメイルであることを示す．二人の関係性は親しさ 2 と入力者に判断された学外の友人である．

データは一文一行体裁を取っている．一文末は「句点･絵文字･顔文字･記号・改行」が挿入されていた箇所と判断し入力されている[8].

また，本コーパスではデータの保存性を重視し，絵文字は Unicode によって入力されている[9]．今回の整備で本文内に Unicode を持つ絵文字がある場合は，U列以降にその出現順にはきだした．Unicode をクリックすることで，当該の絵文字がポップアップ画面で表示される．

[7] 元データに入力がなかった場合は空白.

[8]「絵文字･顔文字･記号が文中の語として捉えられる場合はそれを一文の終了部とせず，改行はしない」「ただし，一行の文章があまりにも長すぎる場合は，読点もしくは意味の区切れで改行としてよい」「句点と見なされる空白はそれを終端記号としてそこで改行とする」「上記以外の空白(行頭，読点代わりなど)はそのまま入力して，終端記号とはしない」加藤(2007:5，一部用語改訂).

[9] 送受信者の携帯キャリアの違いによって文字化けが生じた場合には受信状況のまま「=」「・」マークが付与されている．誤字脱字についても忠実に入力されている.

４．固有名詞，記号類，タグ処理一覧

入力作業時のゆれの統一や，加藤（2007）以降，新規に登場した機能などに付与された情報の統一などを行う必要があった．整備後の詳細は表５の通りである

①タグは全て全角英数で入力，②タグ以外の記号は送受信時のまま全角，半角の区別はデータ入力時のままとした．

表５　固有名詞，記号類，タグ処理一覧

	加藤(2007)	本コーパス内表記（変更後）
記号類	①!，？，♪，☆，★などのパソコンで変換，入力可能なものについてはそのまま入力． ②「↑みたいに～」や「下記参照↓」などの矢印は絵記号タグで括らない．	
固有名詞	固有名詞は，プライバシーに関わるようなものはすべて抽象名詞に置き換えて括弧「＜＞」タグで括る． 【具体例】＜大学名＞，＜人名＞，＜＜人名＞ちゃん＞， ＜携帯番号＞，＜アドレス＞，＜住所＞	
【例外】	①商品名，有名人，文意に関わって，意味が通らなくなるようであれば，大学に関係する部署名，組織名，地名などはそのまま入力． ②その他，プライバシーに関することは，個々人の判断に任せ，本人が公表しても構わないと判断したものはそのまま入力．	
顔文字	顔文字は「%% 顔(」と「) %%」タグが付与されている． 【具体例】%%顔（(ToT)）%% %%顔（(p_q)゛）%%	
絵文字	絵文字は絵文字 Unicode 参考サイト[10]を元に Unicode 番号を記入．「%% 絵(」と「) %%」タグが付与されている． 【具体例】 「うんっ%%絵（(１M)）%%%%絵（Ｅ６	①U 列以降，出現順に Unicode を抽出. 当該の絵文字画像をリンクさせた． ②Unicode が入力されているが調べると不明だったものは，「現在不明」画像にリ

[10] http://trialgoods.com/emoji/　2023 年 3 月現在は閉鎖

	ＥＤ）%%」	ンクさせた. ③入力ミスによる Unicode 不明の場合，本文はそのまま，U 列以降に「未処理」を付与した（左記下線部）.
デコレーションメイル	2009，2010 年度データに出現. 【具体例】 %%デコメ（((おやすみ*´３｀）ノ)）%% %%デコメ（((アンパンマン)(えいえいおー)）%%	入力規則が恣意的であったため，以下の処理を行った. ①本文行では「%%デコメ（ ）%%」タグで括り，丸括弧内は原文のままとした. ②U 列以降では一括して「デコメ」を付与した.
フレーム	2010 年度データに出現. 携帯電話にイラスト画面が送られ，自由にメッセージが挿入できる機能. 【具体例】 %%フレーム(ＨＡＰＰＹ　ＢＩＲＴＨＤＡＹ to you）%%	入力規則が恣意的であったため，以下の処理を行った. ①本文行では「%%フレーム（ ）%%」タグで括り，括弧内は原文のままとした. ②U 列以降では一括して「フレーム」を付与した.
パラ言語的表現	「笑」や「泣」といった，文字本来の意味によるパラ言語的な表現とそれに付随する部分. ①丸括弧「（ ）」で括られた場合と②括られない場合が存在するが，パラ言語的表現として機能していたかは入力者の判断による. 顔文字同様に「%% （ ）%%」タグによって括られる. 【具体例】 ①本文内当該表現が括弧で括られていた場合　%%（(笑)）%% 　　%%（wwwww）%% ②本文内当該表現が括弧で括られていない場合　%%（笑）%%	
形	顔文字ではないが，絵文字･顔文字･記号に準	入力時の規則性が恣意的で

	ずるケータイメイル表現として「形」という分類を設け,「%%形（）%%」タグで括られる. 【具体例】「;」(汗を意味するセミコロン),「やった↑↑」(気分の上昇, 下降),「orz」「(挫折，自己嫌悪，絶望感)など	あるため, 整備の際,「形」のみを削除. T 列では「%%（ ）%%」タグのみを付与した. 【具体例】 「全然%%（orz）%%」

５．その他

5.1 権利者について

所有者（権利者）: 加藤宇温

研究会名：加藤安彦ケータイメイルコーパス研究会

研究会員：加藤宇温（代表者），田中ゆかり，三宅和子，宮嵜由美，林直樹

データ提供者からの権利譲渡について : 所有者が「権利譲渡（ゼミ生より加藤安彦氏が承諾書を受理）がなされている」と判断・保証する.

研究会 Email アドレス: kymobilemailcorpus@gmail.com

5.2　データ整備について

データの整備作業は,研究会員が整備方針を示した上で以下の業者に委託した.

・社会福祉法人　東京コロニー　トーコロ情報処理センター

・パーソルテクノロジースタッフ株式会社

６．　文献・成果

6.1 整備に関わる参考文献

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－ケータイメイルコーパスの紹介とともに－」『専修国文』81 巻，pp.1-17 専修大学日本語日本文学会

Miyake, Kazuko (2007) How Young Japanese Express Their Emotions Visually in Mobile Phone Messages: A Sociolinguistic Analysis, Japanese Studies 27:1, pp.53-72

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)

6.2 β版（整備中のコーパス）を利用した文献

田中ゆかり（2019）「コーパス言語学の学際的研究　5「携帯メイルコーパス」の公開を目

指したデータ整備についての経過報告」『日本大学文理学部　人文科学研究所　研究紀要』97, pp.284-289.
宮嵜由美, 林直樹, 田中ゆかり, 三宅和子 (2020)「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 ―加藤安彦氏の遺志を受けて― 」『社会言語科学会第 44 回大会発表論文集』pp.262-265.

6.3 本コーパス（整備完成版）を利用した文献

田中ゆかり・林直樹(2021) 「「打ちことば」コミュニケーションにおける絵文字使用－「加藤安彦ケータイメイルコーパス」を用いた分析－」『語文』170, pp.124(1)-111(14).
三宅和子(2022)「2000 年代のケータイメイルの実態を捉える－「加藤安彦ケータイメイルコーパス」を利用した研究の可能性」『日本文学文化』21, pp.56(1)-43(14).
宮嵜由美・林直樹・田中ゆかり・三宅和子(2022)「『加藤安彦携帯メイルコーパス ver.1.0』の整備と研究の可能性」『計量国語学会 第 66 回大会予稿集』pp. 18-23.
宮嵜由美(2023)「「句点」の文末使用の経年変化－加藤安彦ケータイメイルコーパスの整備を通して－」『専修国文』112, pp.1-15.

付記

本研究は JSPS 科学研究費補助金：基盤研究(C) 16K02714(宮嵜由美代表 2016), 挑戦的萌芽研究 15K12898(三宅和子代表 2017), 基盤研究(B)18H00680(三宅和子分担 2018), 基盤研究（C）18K00623（田中ゆかり代表 2019), 日本大学文理学部人文科学研究所総合研究費(田中ゆかり分担 2017), 日本大学学術研究助成金社会実装研究費(田中ゆかり分担 2018)の助成を受けたものである.

以上

付録

Windows10 以上
Office2016 以上

ダウンロード時の注意点

本データのダウンロードに，以下圧縮フォルダから作業用フォルダへの移行時に時間を要する場合があることにご留意ください.

圧縮フォルダ解凍後

圧縮 zip フォルダを解凍後そのフォルダ内で作業を続けず，必ず別に作業用フォルダを作成し，ファイルを移動した上で作業を行ってください．そうしないと Unicode のリンクがうまく作動しない可能性があります.

エクセルファイル移行の際には年度毎にひとつずつ移行作業を行うと時間が短縮されます．移行が進まない場合は，一度すべてのプログラムを閉じ，PC を再起動後，圧縮フォルダに再度アクセスし，移行作業を行ってみてください.

絵文字をリンクさせるために

本データの容量と xlsx のハイパーリンク数の上限により，全年度一括データでは絵文字リンクが起動しません．絵文字リンクを起動させるためには各年度に分割したデータをご利用ください.

➢絵文字リンクなしのデータ：一括データ
➢絵文字リンクありのデータ：各年度データ

Unicode 再現ソフト

Windows フォト（フォトビューアー）