『日本語日常会話コーパス』の転記基準と特徴について

臼田 泰如 † 川端 良子 †,‡ 徳永 弘子 †,§ 西川 賢哉 † 小磯 花絵 †

† 国立国語研究所　‡ 千葉大学　§ 東京電機大学

1 はじめに

国立国語研究所では，2016 年度から「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」プロジェクトを進めている．このプロジェクトでは，さまざまなタイプの日常会話をバランス良く収録した大規模なコーパス『日本語日常会話コーパス (Corpus of Everyday Japanese Conversation, CEJC)』を構築し，そのコーパスの分析を通して日常会話を含む話し言葉の特性を多角的に解明することを目指している．本稿では，CEJC の転記基準，および現段階でのコーパスの特徴について報告する．なお詳細な転記基準や作業の流れについては川端ほか (2017) を参照されたい．

2 転記基準

国立国語研究所共同研究プロジェクト「均衡性を考慮した大規模日本語会話コーパス構築に向けた基盤整理」(リーダー：小磯 2014 年 7 月 ～ 2015 年 8 月) での検討内容をベースに以下を転記の基本方針とした。

- 発話内容はテキストで表現できる範囲で転記し，原則として漢字仮名交じりで表記する。
- 転記テキストと音声情報の同期をとることで，転記テキストから音声情報を容易に参照できるようにする。
- フィラーや発音エラーなどの会話で生じる現象は転記する対象を定め，各種タグを用いて表現する。
- 転記テキストに対して自動形態素解析を実施し，語彙素・語形・発音形等の情報を付与する。

音声情報を文字化することで多くの情報が失われるが，失われる情報をタグで補いつつ慣習的な表記で書き起こすことで，可読性や作業効率を担保しつつ音声データの持つ情報を可能な限り盛り込むという方針をとっている．

2.1 転記対象

転記対象となるのは，参与者によって発話された言語音，言語音とは独立に生じる笑い・泣き・歌，および会話の流れに深く関わるその他の発音に類する行為（会話上意味があると考えられる舌打ちなど）である．また参与者からは会話の収録・公開に関する同意書を取得するが，原則として同意書のない参与者の発話は書き起こさない．ただし飲食店での収録などにおいて，店員が注文をとるなど，当たり障りがないと考えられるものについてはその限りではない．

2.2 転記の単位

ELAN[*1]や Praat[*2]などを用いて，音声と映像を参照しながら人手で転記テキストを作成する．その際，下記の条件に基づいて転記を区切り，音声にアライメントする．

1. 知覚可能な休止がある場合
2. 異なる音種（言語音・言語音を伴わない笑い・泣き・歌）が続く場合
3. 発話単位の切れ目がある場合

ここで発話単位とは，JDRI (2014) における「長い発話単位」を指す．話し手と聞き手が行為や情報を交換する際の基本単位に相当し，統語的・談話的・相互行為的なひとまとまりに対応する．

2.3 表記法

発話内容は原則として，現代仮名遣いに従って，漢字仮名交じりで表記する．使用する字種は漢字，平仮名，片仮名を中心とし，必要に応じてローマ字での表記も可とする．数字は漢数字を用いる．読点は付与しないが，後述するように発話単位の境界を示すタグとして「。」を使用する．

転記テキストは転記作業の段階では原則的に表記の統一（例：狐／きつね／キツネ）は行わない．形態素解

[*1] http://tla.mpi.nl/tools/tla-tools/elan/
[*2] http://www.fon.hum.uva.nl/praat/

表 1　転記テキストに使用されるタグの一覧

タグ	概要	使用例
:	非語彙的な母音の引き伸ばし	すご：い，デー：タ
%	非語彙的な音の詰まり	す % ごい，解 % 析
?	疑問上昇調	行きます?，コップ?
(D)	語の言いさし	(D コ) 明日から
(W)	言い誤り・発音の怠け等の一時的な発音エラー	(W コエ｜これ)，(W ギーツ｜技術)
(K)	タグ付与等のために漢字表記ができない箇所	(K ア:マ｜甘) い，(K リ % ツ｜律)
(M)	音や言葉自体が言及の対象とされている発話	すごい を (M すっごい) と発音する
(T)	小さい声で発話している箇所	(T それで)，これ (T じゃないのか)
(L)	笑いが生じている箇所	(L)，これ (L なんですけど)
(C)	泣きが生じている箇所	(C)，(C なにが)
(S)	歌が生じている箇 所	(S)，(S ふるさと)，(S ヘイヘイホー)
(O)	一般的でない外国語/方言が用いられる箇所	(O ボッソワー)，(O ＃＃＃)
(U)	聞き取りや語の判断に自信がない箇所	(U 外国/外交)，(U な＃＃＃)
(R)	個人情報などに関わる仮名・伏字処理候補	(R 国語研究所) の (R 佐藤) さん
。	発話単位末	うん。，やったけど。，食べます。
<>	発音に類するその他の行為	<舌打ち>，<咳>，<口笛>
@	転記単位に対するコメント	スパ@車の愛称
(Y)	漢字表記の一般的な読みと発音が異なる箇所	(Y ゼツ｜舌)，(Y ギョク｜玉)
(F)	「その」がフィラーとして使用された場合	(F その) 研究所への行き方については
(A)	「あの」が連体詞として使用された場合	(A あの) 人が
(X)	語彙不明な箇所	(X リョウゴ) アタック，(X ルトラ) のさ

析によって形態論情報を付与することで，転記テキスト自体に表記の揺れがあっても検索は可能である．また，転記テキストは漢字仮名交じりで表記するため読みは一意に同定できない場合があるが，自動解析の結果得られる発音情報を人手でチェック・修正することにより，形態論情報から正確な発音の情報を得ることができるようにする．

2.4　タグの設計

転記には，発音エラーや非語彙的な音韻（延伸，促音挿入），語の言いさしなどを体系的に示すため，『千葉大学三人会話コーパス』の転記の仕様を参考に定めたタグを使用する．

タグの一覧を表 1 に示す (川端ほか, 2017)．非語彙的な発音の変化（:，%，W）やパラ言語的情報（L，C，S，T）を記述するものや，表記に関わるもの（K，M），個人情報など仮名化や伏字化などの後処理に関わるもの（R）のほか，転記テキストを対象に行われる自動形態素解析におけるエラーをあらかじめ回避するためのもの（Y，F，A，X）などがある．形態素解析用のタグは作業上のものであり，解析後に転記テキストから削除する予定である．

3　転記からみた日常会話の傾向

前節で述べた基準に従って，2016 年 12 月 28 日までに 25.6 時間のデータに対して一次作業以上の転記を終えている（詳細は表 2 参照）．その転記テキストを対象に，川端ほか (2017) に示した手続きに従って UniDic+MeCab で自動解析した．本節では，その自動解析済みデータを用いて，話し言葉を収めた他のコーパスとの比較を通じて CEJC の特徴を記述する．比較対象とするのは『日本語話し言葉コーパス（CSJ)』から学会講演，模擬講演，対話と，『名大会話コーパス（名大 C)』である．なお CSJ 模擬講演は一般話者による日常的話題についての講演，CSJ 対話は学会講演および模擬講演に関して演者になされたインタビュー，インタビューと同一ペアによる課題指向対話および自由対話からなる．また名大 C は約 100 時間の雑談を収録したコーパスである．

CSJ および名大 C の分析には，CEJC と同様に UniDic 体系による解析が施され，オンラインコーパス

検索アプリケーション「中納言」で公開されているバージョンのデータを用いた．これにより同一の基準に基づく短単位の比較が可能になった．

3.1 時間あたりの語数

表 2 に各コーパスの合計時間，合計語数 (短単位数)，時間あたりの語数を示す．

表 2　語数及び時間

	学会講演	対話模擬	自由対話	名大 C	CEJC
合計時間	275.03	330.60	12.45	100	25.62
合計語数	3,321,313	3,639,689	150,720	1,137,800	272,760
語数/時	12,076	11,009	12,103	11,378	10,647

CEJC 以外のコーパスは，参与者が講演や会話をするためにその場に参与して行った発話を収録したものである．従って一定時間以上誰も発話しないという状況は生じにくい．これに対して CEJC は日常的な状況で自発的に生じた会話を収録している．従って CEJC においては会話のない時間や，会話が必須ではない活動に集中する状況が時として生じる．CEJC が他のコーパスに比べて時間あたりの語数がやや少ないのはそうした前提を反映している可能性がある．

3.2 語種・品詞

図 1 は出現した語彙の語種および品詞の比率 (%) を示したものである．ここでは紙面に限りがあるため，小磯ほか (2009) などを参考に，特にレジスター間の特徴記述に有益であると考えられる項目に絞って掲載した．

まず，比較した 5 つのコーパスのなかで，CEJC が全体的にどのような傾向をもつかについてみていきたい．図 1 から見て取れるように，CEJC は語種構成や品詞の出現率において名大 C と類似した特徴をもっている．和語率は学会講演において最も低く，模擬講演，会話を収録したコーパス（CSJ 対話コーパス，名大 C，CEJC）の順に高くなる．いずれのコーパスも和語と漢語が大部分を占めるため，漢語率 (%) はおおむね「100 – 和語率 (%)」に相当し，和語と逆の傾向を示す．また名詞率は学会講演において高く，模擬講演，会話の順に低くなる．形容詞率はそのおおむね逆に学会講演から会話へ増加の傾向を示す．図は省略したが，副詞も形容詞と同様の分布を示した．格助詞については学会講演，模擬講演，会話の順に減少し，終助詞は学会講演，模擬講演，会話の順に増加している．

小磯ほか (2009) では，話し言葉，特に専門性や改まり度の低い場面における話し言葉について，和語率が高く名詞率が低くなる傾向について論じている．CEJC をここでの他のコーパス，特に学会講演・模擬講演と比較してみても，おおむね和語の使用が多く，名詞の割合が低いという傾向が見て取れる．日常的な会話には改まり度の低い状況における会話が多いことを考えると (小磯ほか, 2016)，日常会話の様相を反映している点であると考えられる．

逆に形容詞率については，厳密な表現や事実の伝達を指向する場面より，ものごとの様相や印象を表出することが優先される場面において多用されることが予想される．名大 C と CEJC においては他のコーパスより形容詞が比較的多用されており，これについても日常会話における参与者の指向を反映した分布であるといえるだろう．

また格助詞と終助詞については，くだけた場面においては格助詞の脱落が起きやすく，会話場面においては対人関係的機能をもつ終助詞の使用が増加することが予想される．この点についても概ね予測に沿う分布となった．

ついで，類似した傾向をもつ名大 C と CEJC について比較してみたい．語種については，和語率が名大 C に比べて CEJC ではやや低く，外来語率がやや高いことがみてとれる．名大 C が雑談を収録したコーパスであるのに対し，CEJC は日常の中の多様な場面における会話を収録したコーパスである．収録された会話には雑談が全体の 70% 程度と多い一方で，用談・相談が 25%，仕事や学業に関する会議・会合・レッスンなどが 5% ほど含まれている (小磯ほか, 2017; 田中ほか, 2017)．このように CEJC には，雑談だけでなく専門性の高い活動に関連する会話も一定量含まれている．そうした会話においては雑談より漢語や外来語の使用が増加するだろう．本研究の集計結果はそのような CEJC の収録データの多様性を反映していると考えられる．

4 おわりに

本発表では CEJC の転記の仕様について報告するとともに，これまでに作成した転記テキストと自動形

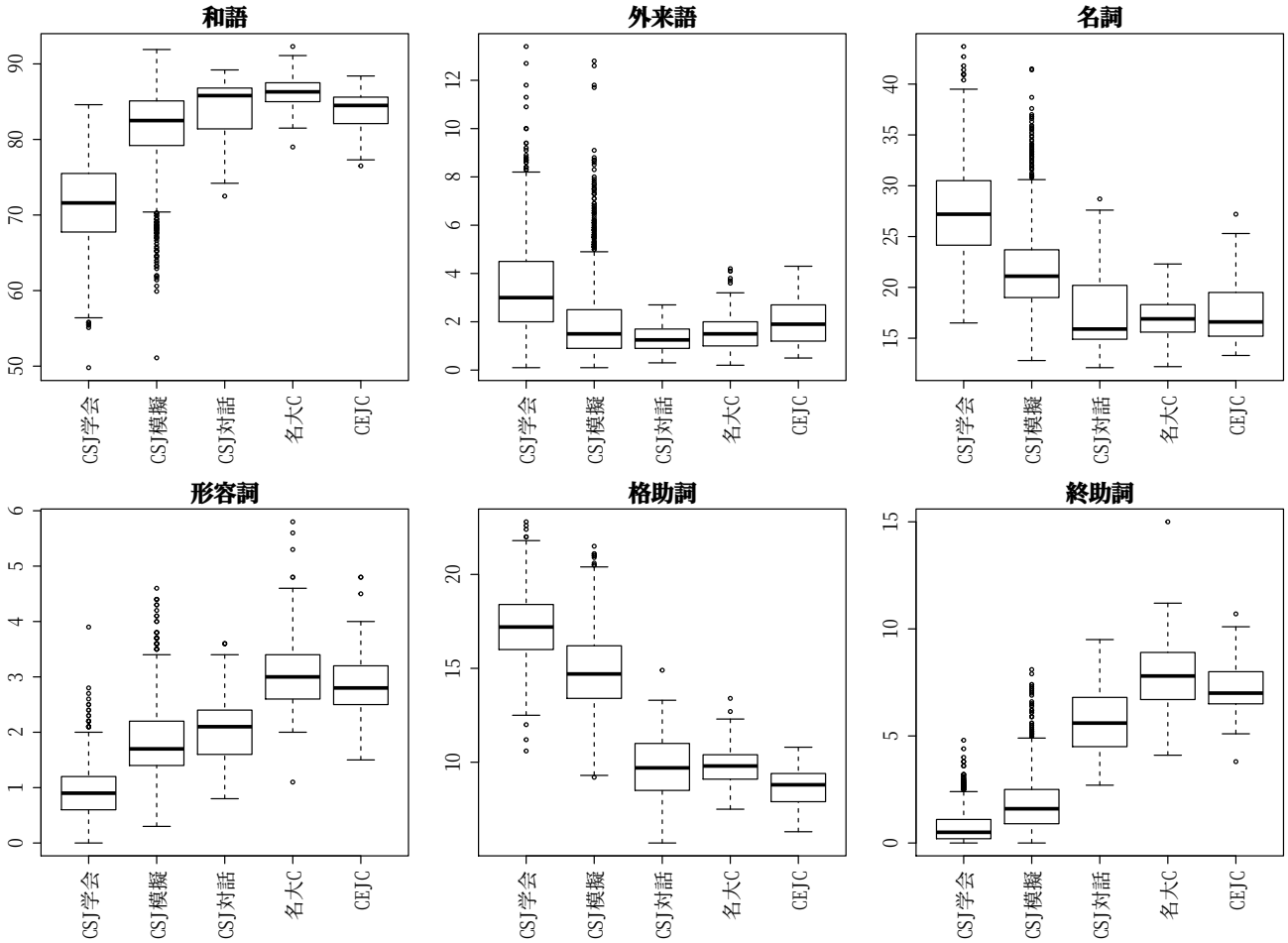

図 1　語種・品詞の比較

態素解析の結果を用い，類似のコーパスとの比較を通じて CEJC の特徴を明らかにした．大量の音声・映像データから質の揃った文字データを得るため，また高精度の形態論情報を付与する上で転記の仕様は重要な役割を持つ．また CEJC は，従来からある対話コーパスと近い傾向を示しつつも，より均衡的な「日常会話」の切り取りを達成しつつあるといえるだろう．

謝辞　収録にご協力いただいた皆さまに感謝申し上げる．本研究は国立国語研究所の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」の研究成果を報告したものである。

参考文献

JDRI, Japanese Discourse Research Initiative (2014)『発話単位ラベリングマニュアル version2.1』. http://www.jdri.org/resources/manuals/uu-doc-2.0.pdf

川端良子・臼田泰如・西川賢哉・徳永弘子・小磯花絵 (2017)「『日常会話コーパス』の転記基準と作業工程」　『言語資源活用ワークショップ 2016 発表論文集』　1 巻.

小磯花絵・小木曽智信・小椋秀樹・宮内佐夜香 (2009)「コーパスに基づく多様なジャンルの文体比較――短単位情報に着目して――」　『言語処理学会年次大会発表論文集』 15 巻　pp. 594–597.

小磯花絵・土屋智行・渡部涼子・横森大輔・相沢正夫・伝康晴 (2016)「均衡会話コーパス設計のための一日の会話行動に関する基礎調査」　『国立国語研究所論集』, 10, pp. 85–106.

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉 (2017)「『日本語日常会話コーパス』の構築」　『言語処理学会年次大会発表論文集』 23 巻.

田中弥生・柏野和佳子・角田ゆかり・伝康晴・小磯花絵 (2017)「『日本語日常会話コーパス』構築における会話収録方法」『言語処理学会年次大会発表論文集』　23 巻.