Transcription Criteria and Guidelines for Processing “Corpus of Everyday Japanese Conversation”

『日本語日常会話コーパス』の転記基準と作業工程

川端 良子 (国立国語研究所 音声言語研究領域 / 千葉大学) *
臼田 泰如 (国立国語研究所 音声言語研究領域)
西川 賢哉 (国立国語研究所 コーパス開発センター)
徳永 弘子 (国立国語研究所 音声言語研究領域 / 東京電機大学)
小磯 花絵 (国立国語研究所 音声言語研究領域)

Transcription Criteria and Guidelines for Processing "Corpus of Everyday Japanese Conversation"

Yoshiko Kawabata (NINJAL / Chiba University)
Yasuyuki Usuda (NINJAL)
Ken'ya Nishikawa (NINJAL)
Hiroko Tokunaga (NINJAL / Tokyo Denki University)
Hanae Koiso (NINJAL)

要旨

本稿は，平成 28 年度から構築を進めている『日本語日常会話コーパス』の転記基準と転記作業工程を紹介する。本コーパスには，日常場面で自然に生じるさまざまなタイプの会話 200 時間がバランス良く収録される予定である。日常会話には，極めてくだけた表現も頻出する。こうしたデータを多人数で書き起こしをするためには，文字化をするための基準を明確に定める必要がある。また，大量の会話を限られた期間で書き起こすために，効率的に作業をするための工夫が必要になる。本発表では，これまでに収録された会話を転記しながら策定した転記基準と効率的に作業を行うために用いている方法を紹介する。

1. はじめに

国立国語研究所では，平成 28 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, 以下 CEJC と略称する) 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 (プロジェクトの詳細に関しては本ワークショップ予稿集所収の小磯 (2017) を参照)。

話し言葉の特性を多角的に分析するためには，様々な人々の多様な言語活動の実態を記録したデータが必要である。しかし，これまでに構築されたコーパスの多くは，会話場面や参加者の属性・関係などに偏りがある。本プロジェクトでは，各世代から均等に調査協力者 (以下，協力者) を募り，収録機材を渡して，調査者が立ち合わずに，協力者自身が日常のさまざまな

* kawabata@ninjal.ac.jp

場面の会話を収録する。こうして収録されたデータから，事前に行った会話行動調査 (小磯ほか 2016) を参考に幅広いレジスターをカバーするようにサンプルを選定しコーパスを構築する設計になっている (コーパスの設計については小磯ほか (2017)，収録方法については田中ほか (2017a)，田中ほか (2017b) 参照)。このようにして，様々な場面における現実の会話データが均衡的かつ大規模に収録される点が CEJC の最大の特徴となっている。

こうしたデータを対象に研究を行うためには，収録した音声を転記したテキストが不可欠である。コーパス内で均質な転記テキストを作成するためには，文字化の基準を策定し，基準に従って転記を行った後，実際に基準に従って転記がなされているかの二重，三重のチェックが欠かせない。そのため転記作業は通常多くの時間を要する。よって，適切かつ効率的に転記を行う方法を確立することが必要であり，またその方法の記録は将来のコーパス構築に対して有効な知見になると考えられる。我々は，実際のデータで転記を行いながら，効率的に作業をするための工程を検討し，ツールの開発や転記基準の改訂を行ってきた。本稿は試行錯誤の末に定まってきた転記基準と効率的に作業を行うために用いている方法について紹介する。

2. 転記基準

本節では，転記基準について説明する。ここで述べる基準はあくまで現時点での規定である。今後転記作業が進むなかで，規定が見直される場合があることに留意されたい。

2.1 基本方針

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：小磯花絵 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まじりで表記する。
- 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。
- 母音の延伸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によって失われる情報は非常に多い。可能な限り音声情報を転記するという方針も考えられるが，そのような方針では作業にかかる時間が増大するだけでなく，転記テキストの可読性が低下する。研究者ごとに会話中に生じる現象への興味は異なるため，必要以上の情報が含まれる転記テキストはかえって使いにくいものとなってしまう。そこで CEJC では，『日本語話し言葉コーパス (CSJ)』(国立国語研究所 2006)，および『千葉大学 3 人会話コーパス』(伝・榎本 2014) の転記基準や作業手順などを参考にして，読みやすさと作業効率を重視した転記仕様を策定することにした。

CSJ の仕様と異なり，表記の統一（例：狐／きつね／キツネ）や発音を記録したテキストの作成は行わない。UniDic(1)に基づく形態素解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柔軟な検索を可能とする。また，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。このように，形態素解析や自作の自動変換プログラムを用いることを前提にし，作業時の転記テキストは変換時に曖昧性が残らない範囲で簡略化して効率化を図る。

2.2 転記対象

転記対象となるのは，会話の参加者 (以下，参加者) が会話中に発した言語音，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。基本的には同意書が得られた話者の発話を転記し，同意書のない話者の発話は原則書き起こさない。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，当り障りがないと考えられる発話に関してはその限りではない。

2.3 転記単位

転記テキストは音声との同期をとるために，以下の条件を満たす発話ごとに転記テキストを区切っている。ここで区切られた転記テキストを「転記単位」と呼ぶ。転記作業は ELAN(2)や Praat(3)などを用い，映像と音声を参照しながら人手で行い，転記単位ごとに音声にアライメントをする。

1. 知覚可能な休止がある場合
2. 異なる音種（言語音・単独の笑い・泣き・歌・その他）が続く場合
3. 発話単位の切れ目がある場合

3 の「発話単位」は，Japanese Discourse Research Initiative によって策定された「長い発話単位」(JDRI 2017) に準拠する。長い発話単位とは，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・相互行為的な一まとまりに対応する単位とされる。

2.4 表記法

発話内容は原則として，現代語の表記の習慣 (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。発話単位の境界を示すタグとして句点「。」を使用するが，読点は用いない。

■発音の扱い　語によっては表記と発音に差異があるが，それが一般的に受け入れられているものがある。たとえば，綴り字において母音が続する言葉「しいたけ (siitake)」，「通院 (tsuuiN)」，「講座 (kouza)」などは発音が長音化して「シータケ」，「ツーイン」，「コーザ」と

(1) 国立国語研究所が規定した「短単位」に対して形態論情報を付与する電子化辞書。http://pj.ninjal.ac.jp/corpus_center/unidic/
(2) http://tla.mpi.nl/tools/tla-tools/elan/
(3) http://www.praat.org/

発音される場合が多い。このように，発音が表記から予測可能である語については，一律に標準的な表記 (「しいたけ」「通院」「講座」) を用いる。

こうした予測可能な発音ではなく，強調のために「すごい」を「スゴーイ」と母音を引き延したり，「スッゴイ」のように子音を引き延して発音する場合は，後述する「:」や「%」のタグ (表 1 参照) を用いて「すご:い」「す % ごい」と表記する。

「わからない」と言おうとして「ワアンナイ」と言ってしまうような，一時的な発音のなまけやエラーについては後述するタグ (W) を用いて，「(W ワアン|分から) ない」のように，実際の発音に加え，本来言おうとした表現を補足する。一方，「こりゃすげえ (これはすごい)」のような，(1) 音の転訛を伴い，(2) くだけた場面で（意図的に）使用される表現で，(3) 一個人に限らず幅広く観察されるものという条件を満たす表現は，発音の一時的なエラーとはみなさず，口語表現としてそのままの語形を表記する。CSJ の構築の際にも，口語表現を積極的に認定したが，CEJC が対象とする日常会話では，講演を中心とする CSJ よりもこうした表現が多く見られることから，形態素解析担当班と連携して，積極的に登録する口語表現を定め，形態素解析用辞書 UniDic の拡張を行う。

感動詞類 (フィラーを含む) やオノマトペについては，基本的には語彙を定めず，聞こえた通りに表記する(4)。

フィラーの例

あー， あのー，んー，えーっとー

オノマトペの例

びゅーびゅー風が吹く， ぼろっぼろの鞄

発音が全く聞き取れなかった部分は，予想される発話の長さ（モーラ数）に応じてシャープ (＃) をタグと組み合わせて記す。

2.5 タグの設計

転記には，発音エラーや非語彙的な音（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。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。本節では，一部のタグについて簡単に説明する。なお，タグは発話単位末を示す句点「。」以外はいずれも半角である。

(4) ただし，応答系感動詞 (「はい」，「うん」等) は，ある程度語彙化されているため，明らかに発音のエラーと思われる場合 (例えば，「はい」と言おうとして「アイ」と言ってしまう場合) についてはタグ (W) を用いて表記する。例の場合，(W アイ|はい) とする。

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 シ:ツ|質) 問，(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	(M 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 これじゃないのか)
(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 へイへイホー)
(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	食べます。，やったけど。，うん。
<>	発音に類する行為	<舌打ち>，<咳>，<口笛>
@	転記単位に対するコメント	スパ@車の愛称
(X)	語が不明な箇所	(X リョウゴ) アタック，(X ルトラ) のさ
(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

■タグ :　語彙的には母音の引き延ばしが含まれないにもかかわらず，強調や言い淀みなどのために一時的に母音が引き延ばされた箇所に「:」(コロン) を付与する。

> 冷た:い視線で
> す:ごい腹立ったな:っていう

■タグ %　強調や言い淀みなどのために，一時的に音が詰まった箇所に「%」(パーセント記号) を付与する。

> き % ついね
> なん % かね:

■タグ ?　「?」は上昇調の句末に付与し，発話が聞き手への質問や確認などであることをを示す。上昇の音調であっても，質問や確認など聞き手への働きかけでないもの (例えば強調など) は付与対象外とする。

■タグ (D)　以下のケースで生じる「語の断片」にタグ (D) を付与する。語の断片は片仮名で表記する。なお，ここで「語」とは「短単位」(小椋 2014) を指す。

- 言いかけて語の途中で発話をやめた場合の中断した語。 言いかけた語が推測できる場合は，後述のタグ (W) と合せて用いる。推測できない場合はタグ (D) を単独で用いる。

> えー (D ダ) 例えば
> っていうだけじゃ (D ワカ) だめだよ。

- 語を言いかけたと言うよりは，発声上の問題で生じたと考えられる断片的な音声。

> その (D ン) 問題は

■タグ (W) 言い誤りや発音の怠けなどによって，一時的に非標準的な発音が生じた場合，(W 実際の発音|意図された語) の形で表記する。 実際の発音は片仮名で表記する。

> (W ワアン|わかん) ない ←「わかん (ない)」を「わあん (ない)」と発音
> (W ジュブン|自分) 一人でできるよ。 ←「じぶん」を「じゅぶん」と発音

語の断片のうち何を言いかけたか分る場合はタグ (W) を使用して言いかけた語を補い，言いさしであることを タグ (D) で示す。

> 知らない (W (D ヒ)|人) 知らない人に ←「人」と言いかけて「ひ」で中断

■タグ (M) 「あ という文字は め と非常によく似ている」のように，音や言葉自体を言及の対象としているような発話 (メタ的引用) のうち， そのままでは可読性が著しく低くなる場合や，タグ: % (W) などを用いて表記すると意図が通じなくなる場合は， その範囲にタグ (M) を付与して可読性を高める。

> (M 僕が) の (M が) は格助詞で (M 行って) の (M て) は接続助詞
> (M すごい) を (M すっごーい) のように促音を入れ強調して話す

■タグ (T) いわゆるささやき声など通常の会話時よりも明らかに小さな声で発話している箇所に付与する。声の大きさに関しては，通常の会話より音量が大きい場合と小さい場合がある。小さい場合のみタグを付与する理由は，声が小さい場合は，聞き手への働きかけではなく，いわゆる「独り言」である可能性があるからである。ただし，転記作業では独り言であるかどうかの判断を行わず，音量の小ささのみからタグの付与を判断する。

■タグ (L)(C)(S) 言語音以外として「笑い」と「泣き」および「歌」を転記対象とし，それぞれ以下のタグを付与する。

- 笑い: タグ (L)
- 泣き: タグ (C)
- 歌: タグ (S)

笑いながら，泣きながら，歌いながら発話している場合，その範囲に上記タグを付与する。非言語音が単独で出現する場合，あるいは歌詞を伴わない (聞きとれない) 歌の場合には，それぞれ (L),(C),(S) を単独で記す。

■タグ (O)　外国語など，現代標準日本語の語彙，文法体系とは異なる体系の言語のうち，日本語の日常会話では一般的に用いられない表現の箇所に付与する。発音は可能な範囲で聞きとり，片仮名で表記する。

> (O チャッチャッカンマンミヤネ)。@韓国語「待って ごめんね」か?

日本語とは異る体系の言語であっても，日常会話で一般的に使用される，あるいは理解できる表現にはタグ (O) は付与しない。

> ハロー ジャクソンとかいったら
> イエーイ
> アイムジャパニーズって言ってあげ (L れば良かっ (U た))。

■タグ (U)　聞き取りや語の判断に自信がない場合は，その範囲にタグ (U) を付与する。複数の候補がある場合は，候補を 「/」(スラッシュ) で区切り，可能性の高い順に列挙する。形態論情報はここで最も可能性が高いとされた語を解析の対象とする。

> (U 底/そこ) に付いている草や泥を取り除き
> 相手も何かきらいだ (U っていうんで)

■タグ (R)　個人情報保護などの観点から問題となる箇所については，その範囲にタグ (R) を付し，データを公開する際に仮名化・伏字化するなどの処理を施す。具体的には次のようなものが対象となる。

- 参加者を含む一般人の名前（愛称を含む，ただし著名人の名前は対象外）
- 参加者を含む一般人の所属する組織名（学校や職場の名称）など。
- 参加者を含む一般人の自宅や所属組織の住所など。
- 誹謗中傷や差別語のうち，特に問題になると判断されたもの。
- 会話者が非公開を希望した箇所。

■タグ (X)　身近な人達同士の会話では，そのコミュニティでのみ通用する語や略語が用いられることがあり，転記作業者が語を特定できないことがある。発話された表現が辞書に登録されていない場合，もしくは辞書に登録されていたとしても，その語の使用は文脈から考えて不自然である場合にタグ (X) 付与する。

> 九十 (X プチポ)。←なんらかの単位と推測できるが，そのような語が存在するか不明
> あのー (X ルトラ) のさ あのー ←文脈からブランド名か店名と推測できるが，不確定

協力者への聞きとり等によって語が判明した場合はこのタグは除かれる。最終的に語が判明しない場合は未知語とされて，タグ (U) が付与される。

■タグ (F)(A) 音の引き伸ばしや音の詰まりのある発話は，タグ「:」と「ー」(長音)，タグ「%」と「っ」(促音) のどちらで表記されているかによって，語が判定できる場合がある。例えば，「あの:」と表記された場合は，連体詞の「あの」と解析できる。「あのー」であればフィラーと判断できる。しかし，長音や促音を伴わない「あの」と「その」は，形態素解析において連体詞とフィラーを区別をすることが難しく，しかも会話中に頻出する。そのため，この 2 つの語形に対してのみ語を区別するためのタグを付与することにした。「その」がフィラーとして使用された場合は タグ (F) を，「あの」がフィラーではなく連体詞として使用された場合はタグ (A) を付与する。「その」は連体詞，「あの」はフィラーで用いられることが全般的に多かったため，作業効率の観点から，出現頻度の少ない使われ方をした場合に限定してタグを付与することにした。形態素解析後はこれらのタグを削除する。

2.6 転記テキストの例

図 1 に作業用転記テキストの例を示す。これは，ELAN で転記したものをタブ区切りテキストに変換したものである。1 行が 1 つの転記単位であり，発話の開始時間と終了時間が割り当てられている。句点「。」は発話単位の境界を示している。テキストには必要に応じて各種タグが付与されている。

図 1 転記テキストの例

発話者	開始時間	終了時間	テキスト
IC01	2502.617	2503.920	(U この前) 飲み会どこで飲んだの。
IC03	2504.661	2505.651	えっと 赤坂。
IC04	2507.718	2508.495	赤坂の
IC03	2508.791	2509.744	(L)
IC04	2509.287	2510.202	料亭。
IC03	2510.912	2511.480	(L いやいや)。
IC01	2511.432	2512.185	違う違う。
IC01	2512.749	2513.451	居酒屋。
IC03	2513.641	2514.236	(W イサカヤ|居酒屋)。
IC03	2515.464	2516.201	(X フタヘルモ)。
IC03	2516.999	2519.648	同期の (D ヒ)(D フ) 同期と二人で飲んだぐらいで。
IC05	2519.670	2521.713	芸能人もいっぱい歩いてるんじゃないですか。
IC05	2521.713	2522.074	外。
IC03	2522.237	2522.865	(W ンナ|そんな) 見る余裕。
IC03	2522.869	2526.534	もう 仕事終わったら家帰ることしか頭に (L ないです)。
IC05	2523.585	2524.039	ね:。
IC03	2526.541	2527.636	(L)
IC01	2530.214	2531.759	前ＴＢＳの地下で:
IC01	2532.456	2533.398	(R 仮名処理) さん ジュリー見た。

3. 転記作業工程

本節では，転記の作成工程について説明する。おおまかな流れを図 2 に示す。作業は大きく 5 つの工程に分けられる。以降に転記の 5 つの工程で行う作業について説明する。

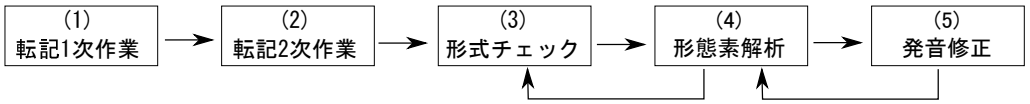

図 2　転記作業工程

■（1）転記 1 次作業　人手で会話音声の文字化と転記単位ごとの音声へのアライメント作業を行う。この作業は二つの方法で行う。一つは，上述の転記基準について知識を有する作業者が，ELAN を用いて文字化・タグ付け・音声へのアライメント作業を同時に行う方法である。もう一つは，いわゆる素起こしのレベルで文字化を外注した上で，転記基準について知識を有する作業者が Praat を用いて音声へのアライメント・タグ付け作業を行う方法である。後者は電話会話や 2 名の（比較的重複の少ない簡単な）会話を対象に行う。いずれの方法においても，次項の転記 2 次作業では，ELAN で映像音声を参照しながら修正作業をする。後者の方法を導入したのは，調査協力者へのフォローアップインタビューを行う時に文字化されたテキストが必要であり，ELAN での方法だけでは間に合わないためである。しかし実際に導入してみると，作業効率がかなり良いことから，電話や簡単な 2 名の会話についてはこの方法も積極的に採用している。現在は，文字化テキストを自動でアライメントした上で人手で修正する方法も検討中である。

■（2）転記 2 次作業　訓練を受けた作業者が，1 次作業で作成された転記テキストを対象に，ELAN 上で映像音声を参照しながら，文字化された内容や付与されたタグなどを確認・修正する。転記 1 次作業ではスピードを重視し，転記基準を完全に満たしたテキストの作成を作業者に求めていない。例えば，正しい転記テキストを作成するには，短単位の知識が必要だが，自信がない場合に詳しく調べる必要はないものとしている。発話単位の認定も，形式的・形態的に特定できる簡単なもののみの認定に留めている。また，基準ではタグはすべて半角，発音は片仮名で表記するが，作業者の入力しやすい文字（全角/平仮名）の使用も認めている。このうち，次の工程で自動変換可能なものを除き，訓練を受けた 2 次作業者が修正を行う。

■（3）形式チェック　転記テキストの形式的なチェックとして，以下の作業を行う。

- 文字種 (半角/全角，平仮名/片仮名) や典型的な転記エラーの自動修正
- タグの種類やタグの入れ子関係などの自動チェック・人手修正
- タグの範囲（短単位を範囲として付与するタグなど）の自動チェック・人手修正
- 発話単位の自動チェック・人手修正（「ケレドモ」節など形態的特徴に基づく自動チェック，発話単位長や発話単位中の無音時間などを参照したチェックなど）

修正作業は，ELAN，Praat，Excel 等のソフトウェアを用いて行う。それぞれで用いるファイル形式 (XML，TextGrid，タブ区切テキスト) を相互変換するスクリプトを整備しており，各作業ごとに最も効率の良い環境で作業できるようにしている。

■（4）形態素解析　上記 (3) の形式チェックを徹底するため，この段階で形態素解析を行う。形態素解析は，形態素解析器 MeCab(工藤ほか 2004) と形態素解析用辞書 UniDic を用いる。入力は発話単位とする。解析にあたっては以下の処理を行う。

- タグが付与されたテキストはそのまま解析できないため，タグを外して解析器に渡す。その際，タグ (D) が付与された言いよどみ要素，タグ (W) の左項（発音のなまけやエラーを含む実際の発音），タグ (U) の第 2 候補以降は解析器に渡さない。
- 短単位を範囲に付与されるタグについては，その情報を利用し，タグ付与範囲の開始・終了位置で必ず単語が分割されるようにする。
- 「(F その)」の品詞を「感動詞-フィラー」，「(A あの)」の品詞を「連体詞」にする。
- 解析器には渡さなかった要素（転記単位の開始・終了時刻の情報などを含む）を解析結果に埋め込み，転記テキストに記された情報を保持する。これにより，転記テキストが再生成できるようにする。なお，タグ (D) の範囲の品詞は「言いよどみ」とする。

以上の処理の結果を用いて，再び形式チェックを行う。

■（5）発音修正　この工程では，工程 (4) にて自動で付与された「発音形」を人手でチェック・修正する。修正対象となるのは，発音が一意に同定できない語 (例：一日「イチニチ/ツイタチ」，日本「ニホン/ニッポン」) や解析誤りによるものである。明らかな誤りや必ずしも誤りとは言えないが低頻度と思われる発音形を機械的に置換した上で，音を聴取しながら発音形の修正を行う。後者の作業は，発音形修正ツールを用いて効率化を図る。

修正した発音情報を参照することで，単位境界・付加情報も正しく解析されることがあるため，発音形修正の終了後，修正した発音に基づき，再び形態素解析を行う。

4. おわりに

本稿では，現在構築中の日本語日常会話コーパス (CEJC) の転記基準と作業工程について紹介した。現在のコーパス構築状況については小磯ほか (2017)，コーパスの特徴については，臼田ほか (2017) を参照されたい。CEJC の公開は，2021 年度末を予定している。また，2018 年度，50 時間のデータをモニター公開する予定である。

謝　辞

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。コーパスの収録にご協力・ご参加くださった皆様に感謝します。

文　献

小磯花絵 (2017).「『日常会話コーパス』プロジェクトーコーパスに基づく話し言葉の多角的研

究－」　言語資源活用ワークショップ 2016 発表論文集.
小磯花絵・土屋智行・渡部涼子・横森大輔・相澤正夫・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.
田中弥生・柏野和佳子・角田ゆかり・伝康晴・小磯花絵 (2017a).「『日本語日常会話コーパス』構築における会話収録方法と進捗状況」　言語資源活用ワークショップ 2016 発表論文集.
田中弥生・柏野和佳子・角田ゆかり・伝康晴・小磯花絵 (2017b).「『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.
国立国語研究所 (2006).『国立国語研究所報告 124 日本語話し言葉コーパスの構築法』　国立国語研究所.
伝康晴・榎本美香 (2014).「『千葉大学 3 人会話コーパス』使用説明書 Release 1」　. http://research.nii.ac.jp/src/files/Chiba3Party_manual.pdf
JDRI (2017).「『発話単位ラベリングマニュアル』version 2.1」　. http://www.jdri.org/open-data/ から入手可能
小椋秀樹 (2014).『書き言葉コーパス ―設計と構築―』, 第 4 章　pp. 68–86. 講座　日本語コーパス 2　朝倉書店.
工藤拓・山本薫・松本裕治 (2004).「Conditional Random Fields を用いた日本語形態素解析」情報処理学会研究報告自然言語処理（NL）, 2004:47, pp. 89–96.
臼田泰如・川端良子・徳永弘子・西川賢哉・小磯花絵 (2017).「『日本語日常会話コーパス』の転記基準と特徴について」　言語処理学会年次大会発表論文集.

関連 URL

『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』プロジェクトのウェブサイト
http://pj.ninjal.ac.jp/conversation/