『日本語日常会話コーパス』における転記の基準と作成手法

著者	臼田 泰如, 川端 良子, 西川 賢哉, 石本 祐一, 小磯 花絵
雑誌名	国立国語研究所論集
号	15
ページ	177-193
発行年	2018-07
URL	http://doi.org/10.15084/00001602

国立国語研究所論集 (NINJAL Research Papers) 15: 177–193 (2018)
ISSN: 2186-134X print/2186-1358 online

『日本語日常会話コーパス』における転記の基準と作成手法

臼田泰如[a]　川端良子[b]　西川賢哉[c]　石本祐一[d]　小磯花絵[e]

[a] 国立国語研究所 研究系 音声言語研究領域 非常勤研究員
[b] 千葉大学大学院 博士課程／国立国語研究所 研究系 音声言語研究領域 非常勤研究員
[c] 国立国語研究所 コーパス開発センター 非常勤研究員
[d] 国立国語研究所 コーパス開発センター
[e] 国立国語研究所 研究系 音声言語研究領域

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』における転記の基準と作成手法について述べる。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や，聞き取りづらい，あるいは把握しづらい表現が頻出する。こうした会話データを多人数により均質に書き起こすには，転記のための基準を明確に定める必要がある。また，200 時間という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。本プロジェクトでは，実際の会話データを対象に転記を行いながら，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。本稿では，このようにして策定した転記基準と，作業を効率的に進めるために整備した方法について紹介する *。

キーワード：『日本語日常会話コーパス』，コーパス構築，転記基準，転記方法，タグ設計

1. はじめに

国立国語研究所では，平成 28 年度から機関拠点型基幹研究プロジェクト「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」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（小磯 2017）。

話し言葉の特性を多角的に分析するためには，さまざま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日本語の会話コーパスの多くは，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，協力者自身に日常のさまざまな場面，さまざまな相手との会話を収録してもらう。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（小磯他 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他（2017），収録方法については田中他（2017a，2017b）

* 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」（プロジェクトリーダー：小磯花絵）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謝意を表する。

参照)。このように，さまざま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には，収録した音声を転記したテキストが不可欠である。しかし，コーパスが対象とする 200 時間という規模を考えると，多くの人の手を経て転記テキストを作成する必要がある。そのため，コーパス内で均質な転記を作成するには，転記の基準を明確に策定する必要がある。また，転記作業は通常多くの時間を要するため，効率的に転記を行う方法を確立することが求められる。我々は，実際の会話データを対象に転記を行いながら，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進めてきた。こうした方法を記録し公開することは，将来のコーパス構築に対して有効な知見になるだろう。また転記テキストを用いて研究するには，どのような基準で転記したかの記録が必要であり，コーパスの利用者にとっても不可欠な情報と言える。そこで本稿では，転記基準と，作業を効率的に進めるために整備した転記作成手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り，今後転記作業が進むなかで，あるいは研究に利用する過程で，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

プロジェクトを開始するにあたり，準備研究として「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」（プロジェクトリーダー：小磯花絵，2014 年 7 月～ 2015 年 8 月）を実施した。ここ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。
- 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。

音声情報を書き起こす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。

CEJC の転記基準には，上記の既存のコーパスのものとは異なる点がある。例えば，CSJ の仕様と異なり，表記の統一（例：狐／きつね／キツネ）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が可能となるためである。また CSJ では，漢字仮名交じりのテキストに加え，仮名の範囲で発音を正確に記録したテキストも転記作業時に作成したが，後者の発音の正確な記録は転記作業時には行わないこととした。かわりに，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

転記対象とするのは，会話の参加者（以下，参加者）が会話中に発した言語音や，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，以下の条件を一つでも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。転記作業は映像分析ソフトウェア ELAN[2] や音声分析ソフトウェア 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である。CSJ では，同種の単位として節単位を導入したが（丸山他 2006），長い発話単位は統語的観点に立脚したこの節単位を基盤とし，そこに談話的・相互行為的な観点も加えて定義した単位である（丸山 2015）。なお転記単位は，上述の通り，発話単位の切れ目，あるいは発話単位中の知覚可能な休止，

[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/

あるいは異なる音種で分割する。よって転記単位の方が発話単位よりも短く，後者が前者を内包する。転記の段階では，より短い単位で音声と同期をとりながら作業を進めるが，コーパスを公開する際は転記単位と発話単位の両方を提供する。2.4 節でも言及するように，転記作業の中で発話単位の境界に記号を付与することにより，転記単位から発話単位への自動変換が可能となる。

2.4 表記法

発話は原則として現代語の表記の習慣（現代仮名遣い）に従い漢字仮名交じりで表記する。使用する字種は，漢字，平仮名，片仮名を中心とする。「ＥＲ（イーアール）」などアルファベット読みをしている箇所は全角アルファベットで表記する。数字は漢数字を用いる。オノマトペは，「ぴゅーぴゅー」「ざっくざく」のように聞こえた通り長音記号（ー）と促音記号（っ）を含めて平仮名で表記する。感動詞は平仮名で表記する。詳細は後述する。

形態論情報を付与することで表記によらない柔軟な検索が可能となるため，転記の段階で表記の統一（例：狐／きつね／キツネ）は行わない。

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。可読性が著しく落ちる場合や，語や意味の同定に曖昧性が生じる場合などに半角のスペースを入れる。

2.5 感動詞の扱い

日常会話には，フィラーや感情表出系感動詞，応答系感動詞など，多様な感動詞が出現するが，表記に迷うことも多い。そこで CEJC ではこうした感動詞類の表記について次の通り定めた。

■フィラー　CEJC ではフィラーを，表 1 の形式であり，かつ場つなぎ機能を有するものに限定して認定する。表 1 において，( ) 内は省略可能であることを示す。これらの形式であっても場つなぎ機能を有していないものや，場つなぎ機能を有しているが表 1 の形式ではないものは，フィラーとはみなさない。

表 1　フィラーの形式一覧

基本表現	あ(ー)，い(ー)，う(ー)，え(ー)，お(ー)，ん(ー)， と(ー)*，っと(ー)*， あ(ー)(ん)(ー)の(ー)*，そ(ー)(ん)(ー)の(ー)*， う(ー)ん，う(ー)ん(ー)(っ)と(ー)*，ん(ー)(っ)と(ー)* あ(ー)(っ)と(ー)*，え(ー)(っ)と(ー)*	
組み合わせ	上記基本表現 +「ですね (:)」「っすね (:)」 上記*付きの表現 +「ね (:)」「さ (:)」	例) あのですね: 例) とね:，んーっとさ:

フィラーは平仮名で表記し，フィラーの前後にスペースを入れる。また，「あの」「その」に関しては，形態素解析においてフィラーなのか連体詞なのか区別することが難しいため，2.7.6 節で言及するタグ (F) を付与する。

■感情表出系感動詞 CEJC では，「えっ」や「あーあ」など，驚いた時や落胆した時などに発する表現を感情表出系感動詞と呼ぶ。感情表出系感動詞は語彙を定めず原則として聞こえた通り表記する。外国語由来の感動詞は片仮名で表記し，それ以外は平仮名で表記する。

```
イエーイ。
オッケー。
```

発話単位の境界に出現する感情表出系感動詞はその前後で発話単位を区切る。

```
あー。
名前出しちゃったよ。
```

■応答系感動詞・呼び掛け・掛け声 応答系感動詞（「はい」「ふーん」「いいや」など）や呼び掛け（「おい」「やあ」など），掛け声（「えい」「そら」など）については，基本となる語形が想定可能な場合には，曖昧に発音されている場合でもその基本となる語形で転記する。促音および長音が挿入される場合には，派生の語形として「っ」「ー」を用いて表記し，2.7.1 節で言及するタグ % やタグ : は用いない。基本となる語形が想定できない場合には聞こえた通りに転記する。感情表出系感動詞の場合と同様に，発話単位の境界に出現する場合はその前後で発話単位を区切る。

2.6 口語表現の扱い

「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，一時的な発音の怠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 | わかん) 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。

それに対し，「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 よりもこうしたくだけた表現が格段に多く見られる。現在，プロジェクトで形態素解析を担当するグループと連携し，口語表現の対象やその基準を定め，形態素解析用辞書 UniDic の拡張を行っているところである。

2.7 タグの設計

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, CSJ や『千葉大学 3 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 2 に示す。

表 2　転記テキストに使用されるタグの一覧

1）非語彙的な発音の変化や言いよどみに関わるもの

タグ	概要	使用例
:	非語彙的な母音の引き延ばし	すご:い，デー:タ
%	非語彙的な音の詰まり	す%ごい，解%析
(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)
(D)	語の言いさし	(D コ)明日から

2）韻律・パラ言語的情報に関わるもの

?	疑問上昇調	行きます?，コップ?
(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)	泣きながら発話している，あるいは単独の泣き	(C)，(C なにが)
(S)	歌いながら発話している，あるいは歌詞を伴わない歌	(S)，(S ふるさと)，(S ヘイヘイホー)
<>	発音に類する行為のうち会話の流れに関わるもの	<舌打ち>，<咳>，<口笛>

3）聞き取り等の判断の信頼性に関わるもの

(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＃＃＃)
(X)	語が不明な箇所（最終的には(U)に統合）	(X リョウゴ)アタック，(X ルトラ)のさ

4）転記テキストの可読性や内容理解の補助に関わるもの

(K)	タグ等のために漢字表記できず可読性が落ちる箇所	(K シ:ツ|質)問，(K リ% ツ|律)
(M)	音や言葉が言及対象とされており内容が把握しづらい箇所	(M すごい)を(M すっごい)と発音する
(O)	一般的に理解が難しい外国語・方言が用いられる箇所	(O ボッソワー)，(O ＃＃＃)

5）発話単位・転記単位に関わるもの

。	発話単位末	食べます。，やったけど。，うん。
+	1語内の知覚可能な休止により転記単位が分割される場合	す+ごい，神田+川

6）形態素解析のための作業上のもの

(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)
(G)	解析が困難な口語表現，口語表現かどうか迷う表現	(G 嫌|や:)，(G ちょっと|ちょっ)
(F)	「あの」「その」類がフィラーとして使用された場合	(F その)なんというんですか

7）その他，個人情報保護やコメントに関わるもの

(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
@	転記単位に対するコメント	スパ@車の愛称

タグは大きく次の七つに分けられる。以降に各タグについて説明する。

1）非語彙的な発音の変化や言いよどみに関わるタグ
2）韻律・パラ言語的情報に関わるタグ
3）聞き取り等の判断の信頼性に関わるタグ
4）転記テキストの可読性や内容理解の補助に関わるタグ
5）発話単位・転記単位に関わるタグ
6）形態素解析のための作業上のタグ
7）その他，個人情報保護やコメントに関わるタグ

なお，タグは発話単位末を示す句点「。」および発話に類する行為を示すタグ＜＞以外はいずれも半角である。

2.7.1 非語彙的な発音の変化や言いよどみに関わるタグ

■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（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
```

■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。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の場合。言いかけた語が推測できる場合は，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。言いかけた語が推測できない場合はタグ（D）を単独で用いる。

```
知らない(W (D ヒ)|人)知らない人に　←「人」と言いかけて「ひ」で中断したと判断
えー(D ダ)例えば左　← 言いかけた語が推測できない場合
```

- 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の場合。

```
その(D ン)問題は
```

- 言い淀みや発声上の問題による音声が転記困難な場合，＃（全角シャープ）を用いて表記する。

```
(D ＃)これすごいさ いい写真だけどさ:
```

2.7.2 韻律・パラ言語的情報に関わるタグ

■タグ？　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示す。上昇の音調であっても，強調など聞き手への働きかけでないものは付与対象外とする。

■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)　「笑い」「泣き」「歌」に関連するタグとして以下を用いる。

- 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（聞き取れない）歌の場合には，それぞれ (L),(C),(S) を単独で記す。

```
てめっちゃ断っ(L たんですけど)　← 笑いながら発話
(L)　← 発話を伴わない単独の笑い
```

■タグ＜＞　言語音・笑い・泣き・歌以外で，転記対象者が行った発音に類する行為のうち，特に会話の流れに関わるものを＜＞タグ（全角）の内部に記載する。タグの前後で転記単位を分割する。行為の名称は別途定めるリストを参照し，同種の行為は同じラベルとなるよう統一する。なお，転記対象とする行為は，原則として発声器官を使用してなされる音であり，かつ，それがやりとりに関連する場合（会話の相手への働きかけや応答としてなされていることが理解できるなど）に限定する。

2.7.3 聞き取り等の判断の信頼性に関わるタグ

■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。形態論情報などの各種アノテーションは，ここで最も可能性が高いとされた語を対象に付与する。聞き取りが著しく困難で全く語が特定できない場合は，考えられる発話の長さ（モーラ数）に応じて＃（全角のシャープ）を記す。

```
(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)
全部まとめて(U ＃＃＃＃＃)
```

■タグ (X)　身近な人達同士の会話では，発音の明瞭さにかかわらず，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がこれに相当する。

```
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
(F あの:)(X ルトラ)のさ (F あの:)　← 文脈からブランド名か店名と推測できるが，不確定
```

タグ (X) を付けた箇所は，転記テキスト作成後に行う協力者へのヒアリングで語を確認し，語が判明した場合はこのタグを削除する。語が判明しない場合，形態論情報において品詞付与対象外のものとして扱い，転記テキストはタグ (U) に統一する。よって公開する転記テキストに本タグは含まれない。

2.7.4 転記テキストの可読性や内容理解の補助に関わるタグ

■タグ (K)　漢字やアルファベットで表記することが自然な語であっても，各種転記用のタグを付与したり，知覚可能な休止に伴い転記単位を分割した結果，仮名で表記せざるを得ず，かなり不自然な表記になり，発話内容を把握しづらくなることがある。そのような場合，次のように (K 実際の発音 | 漢字表記) で表記する。実際の発音は片仮名で表記する。タグは漢字 1 字を範囲とする。

```
五(K マ:ン|万)
なんかそういう気(K ヅ%カ|疲)れがない
```

■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，タグ：やタグ % などは使用せず聞こえた通りに表記する。

(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)のように強調して話す

■タグ (O)　外国語や一部の方言など，現代標準日本語の語彙，文法体系とは大きく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取り，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))

2.7.5 発話単位・転記単位に関わるタグ

■タグ。　2.3 節で言及した発話単位の切れ目には「。」（句点）を付与する。それ以外では句読点は使用しない。可読性が著しく落ちる場合，語や意味の同定に曖昧性が生じる場合，形態素解析の精度を下げる可能性がある場合に，半角のスペースを入れる。

棒でなんかこう ね お参りするのがあったんだよね。
違うお医者さん かかることにして。

■タグ +　語の内部に休止が入ることで転記単位が分割される場合，タグ + を使用し，転記単位をまたいで語が続いていることを示す。ここで 1 語とは「一つの纏まった意味をもつ表現」という程度のゆるい纏まりを意味し，短単位に限定しない。

じゃ百四だから変わ +　←「変わる」のカワとルの間に休止があり転記単位が分割
るだと(U 思う)

2.7.6 形態素解析のための作業上のタグ

ここでは，自動形態素解析におけるエラーを回避し精度を向上させることを意図して導入したタグについて概説する。

■タグ (Y)　漢字やアルファベット表記した語が一般的な読みと異なり，その表記から実際の発音が特定できない恐れがある場合や，複数の読みがある場合，その範囲に (Y 読み | 標準的表記) を付与し，読みを補足する。読みは片仮名で表記する。

```
(Y エンザン|遠山)の中へ
一二三(Y ヨン|四)五
```

3 節の「発音修正」で述べるように，自動形態素解析で得られた「発音形」を人手でチェック・修正して発音を特定する工程があるため，転記の段階で厳密に発音の情報を補うことはしない。

■タグ (G)　口語表現に関連して，次の二つの事例に対してタグ (G) を付与する。

1. 口語表現であるもののうち，形態素解析が困難な一部の表現
2. 口語表現かどうか判断に迷う表現

該当する箇所は (G 標準的表現 | 口語表現（候補）) のように，左側は標準的な表現を記し，右側には出現した口語表現（上記 2 の場合は口語表現候補）を表記する。

1 については，これまでの形態素解析の精度や出現頻度を考慮し，本タグを付与する対象を定めている。表 3 に記す表現がこれに該当する。

表 3　タグ(G)を付与する口語表現

口語表現	意味	例
や	形状詞「嫌」 感動詞「いや(否)」 感動詞「いや」	そんなの(G 嫌|や)。 (G いや|や)。そうじゃなくて (G いや|や)。それ程のことでも
ま	副詞「まあ」 感動詞「まー」	(G まあ|ま)いいか。 あら(G まー|ま)驚いた。
そ	感動詞相当(応答)「そう」	(G そう|そ)(G そう|そ)。そうゆうこと。
ちょ,ちょっ	副詞「ちょっと」	(G ちょっと|ちょ)待てよ。

2 については，口語表現として扱うか否かが現時点で確定していない表現に付与する。2.6 節で言及したように，CEJC が対象とする日常会話にはくだけた表現が頻出するため，講演を中心に整備した CSJ の基準では十分に対応することができない。そこで，ある程度データを蓄積し，日常の言葉として定着した口語表現とみなすか，一時的な発音の怠けとしてタグ (W) の扱いとするかを判断するために，口語表現の可能性のある箇所に本タグを付与する。現在，その事例を見ながら，口語表現のリスト・基準を作成しているところである。

■タグ (F)　「あの」「その」類は，連体詞の場合とフィラーの場合が存在するが，音を聞かないと判断できないことも多く，また会話に頻出する。そのため，音を聞きながら作業する転記の段階で，連体詞かフィラーかを区別することとした。これらの表現がフィラーとして使用された場合にタグ (F) を付与する。本タグが付与されていない「あの」「その」類は連体詞として解析される。

```
(F その:)なんてゆうの
```

2.7.7 その他，個人情報保護やコメントに関わるタグ

■タグ（R）　データを公開する際に仮名化・伏字化する箇所については，その範囲にタグ（R）を付し，あとで仮名化・伏字化の処理を施す。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）
- 個人識別符号
- 参加者を含む一般人の所属する組織（学校や職場など）の名称や住所・電話番号等
- 誹謗中傷や差別語のうち，特に問題になると判断されるもの
- 上記以外で会話者が非公開を希望する箇所

■タグ @　当該の転記単位に関するコメントは，各転記単位の末尾に @ コメント内容の形式で記述する。

2.8 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 行が一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。

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。

図 1　転記テキストの例

3. 転記作成手法

本節では，転記の作成手法について説明する。おおまかな流れを図 2 に示す。

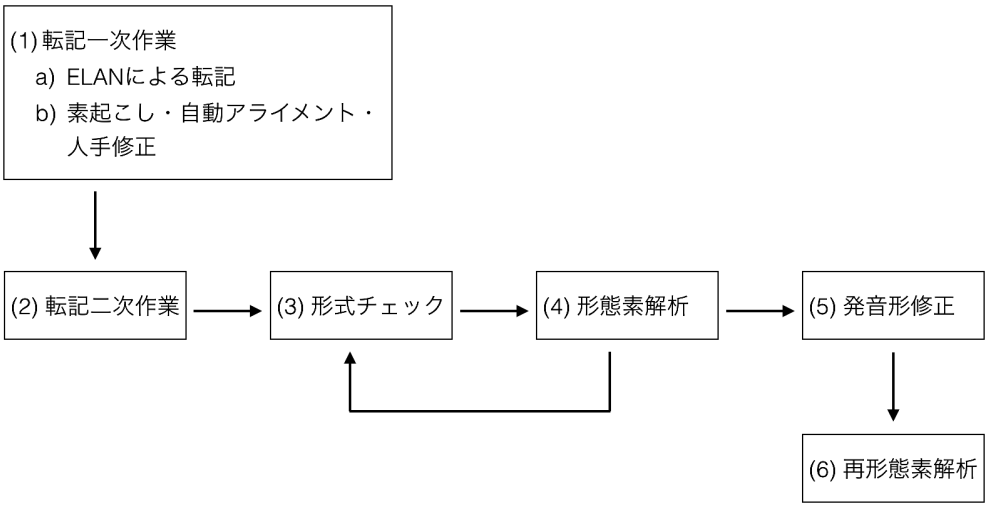

図 2　転記作業工程

作業は大きく五つの工程に分けられる。以降に転記の五つの工程で行う作業について説明する。

(1) 転記一次作業　人手で会話音声の書き起こし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，自動アライメントのシステムによって音声に割り付け，転記基準について知識を有する作業者が Praat を用いてタグ付け作業などの人手修正を行う方法である。後者の方法を導入した理由は次の通りである。

音声から文字化，さらにアライメントまでの作業を音声情報処理によりすべて自動化することができれば大幅な作業効率の向上が望める。しかし現在の音声認識技術では，さまざまな環境音が重畳したり自発発話の非流暢化が生じやすい CEJC の音声に対しては認識精度に難があり，転記テキスト付与の全自動化は実用的な段階に至っていない。もっとも現時点でも，あらかじめテキストが存在すれば，テキストと音声の自動アライメントを高い精度で行うことが可能な段階にある。そこで，2 名の会話を対象に自動アライメントを試行した結果，90% 以上の発話に対しその開始位置を推定することができることが分かった。また，開始位置の誤差は人手で付与した場合と比較しておよそ 300 ms 程度となっている。このように，自動アライメントの精度が実用可能な水準にあり，この工程を導入することによって作業効率がかなり良くなることから，会話者数の少ない（主として 2 名の）会話についてはこの方法を積極的に採用することとした。自動ア

ライメントについては，音声認識を用いた自動字幕作成システム（秋田他 2015，河原他 2016）の技術を応用している。このシステムは，音声ファイルや映像ファイルを入力とし，音声認識による書き起こしを時刻情報付きで出力して字幕として提示する目的で構築されたものだが，あらかじめ入力されたテキストに対して音声を同期させる機能も有している。CEJC では上述のように，素起こしテキストと音声に対しこのシステムを適用する。ただし，現在のシステムには以下に挙げる問題もあり，自動アライメント結果に対し必ず人手で修正するようにしている。

- フィラーや小声の発話等で，開始時刻が推定できない場合がある
- 発話の終了時刻が正確には推定できない

自動アライメント性能の詳細は石本（2017）を参照されたい。

(2) 転記二次作業　訓練を受けた作業者が，一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。転記一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，訓練を受けた二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，タブ区切りテキスト）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（工藤他 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（発音の怠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。
- 転記単位境界（ただしタグ + が使用されている個所を除く）およびスペースの位置で必ず単語が分割されるようにする。
- 「(F その)」「(F あの)」の品詞を「感動詞 - フィラー」にし，タグのない「あの」「その」の品詞を「連体詞」にする。
- タグ (G) が付与された語のうち, 表3に挙げられているものに対して, 適切な品詞を与える。
- 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（3）の形式チェックを行う。

(5) 発音形修正　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，修正した発音形に基づき，再び形態素解析を行う。

このように，転記テキストの作成においては自動アライメント技術も部分的に導入しつつ，転記テキストのエラー検出・修正については UniDic と MeCab を用いた自動形態素解析結果を活用しながら，転記作業を効率化すると同時にその精度を一定以上に保つようにしている。

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）における転記の基準と作成手法について紹介した。現在のコーパス構築状況については小磯他（2017），コーパスの特徴については，臼田他（2017），川端他（2017）を参照されたい。CEJC の公開は，平成 33 年度末を予定している。また平成 30 年度に 50 時間のデータをモニター公開する予定である。

参照文献

秋田祐哉・三村正人・河原達也（2015）「音声認識を用いた講義・講演の字幕作成・編集システム」『情報処理学会研究報告』2015-SLP-108(2): 1–6.

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.（2018 年 6 月 12 日確認）

石本祐一（2017）「コーパス構築における発話アライメントの現状」『言語資源活用ワークショップ 2016 発表論文集』30–37.

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/.（2018 年 6 月 12 日確認）

川端良子･臼田泰如･西川賢哉･徳永弘子･小磯花絵 (2017)「『日本語日常会話コーパス』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』296–306.
河原達也・秋田祐哉・広瀬洋子（2016）「自動音声認識を用いた放送大学のオンライン授業に対する字幕付与」『情報処理学会研究報告』2016-AAC-2(5): 1–4.
小磯花絵（2017）「『日常会話コーパス』プロジェクト―コーパスに基づく話し言葉の多角的研究を目指して―」『言語資源活用ワークショップ 2016 発表論文集』114–119.
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: 775–778.
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: 85–106.
国立国語研究所（2006）『日本語話し言葉コーパスの構築法』（国立国語研究所報告 124）東京：独立行政法人国立国語研究所.
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）』47: 89–96.
丸山岳彦（2015）「発話の単位」小磯花絵（編）『話し言葉コーパス―設計と構築―』54–80. 東京：朝倉書店.
丸山岳彦・高梨克也・内元清貴（2006）「節単位情報」国立国語研究所 (2006), 255–322.
小椋秀樹（2014）「形態論情報」山崎誠（編）『書き言葉コーパス―設計と構築―』68–86. 東京：朝倉書店.
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』収録の進捗状況」『言語資源活用ワークショップ 2016 発表論文集』248–257.
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』23: 481–484.
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』23: 174–177.

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/（2018 年 6 月 12 日確認）

Criteria and Composition Method of Transcription for the Corpus of Everyday Japanese Conversation

USUDA Yasuyuki[a] KAWABATA Yoshiko[b] NISHIKAWA Ken'ya[c]
ISHIMOTO Yuichi[d] KOISO Hanae[e]

[a]Adjunct Researcher, Spoken Language Division, Research Department, NINJAL
[b]Doctoral Student, Chiba University / Adjunct Researcher, Spoken Language Division, Research Department, NINJAL
[c]Adjunct Researcher, Center for Corpus Development, NINJAL
[d]Center for Corpus Development, NINJAL
[e]Spoken Language Division, Research Department, NINJAL

Abstract
This paper describes the criteria and composition method of transcription for the Corpus of Everyday Japanese Conversation, which has been in construction since 2016 and will contain 200 hours of various types of conversations in a balanced distribution. As some expressions are extremely informal, hard to hear, or hard to understand, it is necessary to establish clear criteria for transcription to ensure homogeneous transcription quality from a large number of staff. Methods are also required to transcribe no less than 200 hours of conversations efficiently and in a timely manner. As part of this project, procedures for efficient transcription have been considered, and the development of tools and the revision of criteria of transcription have been conducted. This paper presents said transcription criteria and methods.

Key words: Corpus of Everyday Japanese Conversation, corpus construction, transcription criteria, transcription method, tag design