『日本語日常会話コーパス』の短単位解析：作業工程を中心に

著者	西川 賢哉, 渡邊 友香
雑誌名	言語資源活用ワークショップ発表論文集
巻	4
ページ	238-250
発行年	2019
URL	http://doi.org/10.15084/00002575

『日本語日常会話コーパス』の短単位解析：作業工程を中心に

西川 賢哉 (国立国語研究所 コーパス開発センター) *
渡邊 友香 (国立国語研究所 音声言語研究領域)

Morphological Analysis of the Corpus of Everyday Japanese Conversation

Ken'ya NISHIKAWA (National Institute for Japanese Language and Linguistics)
Yuka WATANABE (National Institute for Japanese Language and Linguistics)

要旨

国語研で構築中の『日本語日常会話コーパス』(CEJC) の短単位解析作業について報告する。CEJC における短単位情報は，アノテーションの一つであるにとどまらず，(i) 発音に関する情報を唯一持つ，(ii) 他のアノテーション（長単位・韻律）の初期値作成の際の入力となる，(iii) 転記誤りを発見する際の有力な手掛かりとなる，などの点で重要なアノテーションであり，高い精度が求められる。作業は次のように進める。まず，MeCab+UniDic で自動解析したのち，短単位付加情報の一つである「発音形」を，音を聴取しながら人手で修正する。これにより，発音形の精度向上を図る。さらに，修正された発音形を尊重しつつ再び形態素解析を行なうことにより，発音形以外の短単位情報（境界・付加情報）の精度向上をも図る（例：初出店「ショシュツ/テン」→「ハツ/シュッテン」）。その後，短単位解析結果を，形態論情報管理ツール「大納言」で検索・修正できるようにし，引き続き解析誤りを修正していく。修正が進んだ段階で，境界・付加情報に揺れがないかを系統的にチェックする（例：「ミリ/メートル」「ミリ＝メートル」）。

1. はじめに

国立国語研究所では，平成 28 年度から，日常場面で自然に生じるさまざまなタイプの会話を収録した『日本語日常会話コーパス』（Corpus of Everyday Japanese Conversation，以下 CEJC）の構築を進めている。CEJC の公開にあたっては，音声信号に加え，転記テキスト，単語（短単位・長単位），係り受け，談話行為，韻律情報といった各種アノテーション（研究用付加情報）が提供される予定である(1)。

本稿では，これらアノテーションのうち，短単位情報について，特に作業工程を中心に報告する。短単位は，言語の形態的側面に着目して規定した言語単位であり，用例検索での利用を主たる目的としている（概要については，小椋 (2014)，岡 (2019) を，詳細については小椋他 (2011) を参照されたい）。

* nishikawa[at]ninjal.ac.jp

(1) 転記テキストと短単位については，コーパス全体（約 200 時間）に対して，その他のアノテーションについては，「コア」と呼ばれる CEJC のサブセット（約 20 時間）に対して提供される予定である。

【短単位分割例】/国立/国語/研究/所/が/日本/語/コーパス/を/構築/し/た/（小椋 2014: 74）

CEJC における短単位情報は，数あるアノテーションの一つであるにとどまらず，(i) 発音に関する情報を唯一持つ，(ii) 他のアノテーション（長単位・韻律）の初期値作成の際の入力となる，(iii) 転記の誤りを発見する際の有力な手掛かりとなる，などの点で重要である。そのため，可能な限り精度を高められるよう（その一方で，作業を効率的に行なえるよう）留意し，作業工程を策定した。

2. 転記の仕様

短単位解析の作業工程を説明する前に，短単位解析の入力となる転記テキストの仕様について簡単に説明しておく（詳細は，臼田他 (2018) を参照）。転記テキスト例を図 1 に示す。

図 1 から明らかなように，発話内容は，原則として漢字仮名交じりで表記する。転記テキストには，発話内容とともに，開始時刻・終了時刻の情報も保持する(2)。これにより，ここから音声情報を簡単に参照できるようになる。話し言葉で生じる現象（語断片の出現，母音の延伸，発音の一時的なエラー等）は，表 1 に示すタグ（転記タグ）を用いて表現する。

転記テキストの 1 行は，一つの転記単位を構成する。転記単位とは，以下のいずれかの条件を満たす箇所で区切られた単位である。

(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」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。発話単位の末尾にはタグとし

fileID	speakerID	startTime	endTime	pause	text
T004_003	IC01	23.733	24.403	2.716	いいよ いいよ (D # #)。
T004_003	IC02	23.851	24.172	2.09	うん。
T004_003	IC02	26.262	26.947	1	でかいんだよ。
T004_003	IC01	27.119	27.302	1.798	うん。
T004_003	IC02	27.947	28.506	0.002	だから。
T004_003	IC02	28.508	28.99	23.849	あれが。
T004_003	IC01	29.1	29.59	0.097	そうだね。
T004_003	IC01	29.687	30.603	14.262	(W デシ|出し) にくいんだ。
T004_003	IC03	38.577	39.252	1.109	あー。
T004_003	IC03	40.361	41.516	0.196	雲取も:。
T004_003	IC03	41.712	41.968	0.736	(D イ)
T004_003	IC03	42.704	44.705	0.541	一組だけ外人のご一行みたいの
T004_003	IC01	44.865	45.601	0.899	えー。
T004_003	IC03	45.246	45.935	2.32	帰る時。

図 1　転記テキスト例（タブ区切り）

(2) 書き起こし作業は，映像解析ソフトウェア ELAN (`http://tla.mpi.nl/tools/tla-tools/elan/`) や音声分析ソフトウェア Praat (`http://www.praat.org/`) を用いて行なっているため，時刻情報は自動的に記録される。

表 1　転記テキストに使用されるタグの一覧（臼田他 (2018: 182) 表 2 をもとに最新の仕様に更新）

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 な感じ)
(X)	語が不明な箇所	(X リョウゴ) アタック，(X ＃＃) の

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 短単位内の知覚可能な休止により転記単位が分割される場合	狭＋い，六＋本木
␣	（スペース）形態統語的単位の境界（任意）	十 二十 三十，川 渡り

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
(I)	「感動詞-一般」のうち，「感動詞-フィラー」と誤解析されやすいもの（発話単位の内部に出現するもの）	(I あー)，(I うー)

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，(R ＊＊＊)
@	転記単位に対するコメント	スパ@車の愛称

て「。」が付与される。

転記単位は，(3) 発話単位境界の他，(1) 知覚可能な休止や (2) 異なる音種の境界でも分割されるため，転記単位は必ず発話単位と同じか，それより短い単位となる。図 1 に挙げた例では，ほとんどの転記単位が単独で一つの発話単位を構成するが，末尾 4 行の IC03 による発話（下から 2 行目にある IC01 による発話「えー。」は除く）は，

(D イ) ［休止］一組だけ外人のご一行みたいの［休止］帰る時。

のように，三つの転記単位が一つの発話単位を構成する。

3. 短単位解析の工程

前節に述べた転記テキストをもとに，短単位解析を施す。作業工程は大きく以下の 6 つに分けられる：

1. 形態素解析 (1)：初期値作成
2. 発音形修正
3. 形態素解析 (2)：発音形を考慮した解析
4. 解析結果の形態論 DB へのインポート
5. 大納言での人手修正
6. 系統的チェック・修正

以下，各作業について説明する。

3.1 形態素解析 (1)：初期値作成

形態素解析器 MeCab（工藤他 2004）と形態素解析用辞書 UniDic（岡 2019）を用いて，機械的に形態素解析を施し[(3)]，初期値を作成する。解析にあたり，以下の処理（前処理・後処理）を行なう。

- 形態素解析前処理
  - 要素を持たないタグ (L), (C), (S), ＜＞に，ダミーの要素◇を与える（開始・終了時刻をそれに帰属させるため）
  - 転記単位から発話単位をくみ上げる（発話単位を形態素解析の単位とするため）。
  - 転記タグは外して解析器に渡す。タグ (D) が付与された要素（語断片），タグ (W), (K), (Y), (G) の左項，タグ (U) の第 2 候補以降は解析器に渡さない。
    - ただし「。」は解析器に渡す（解析精度の向上を期待できるため；後処理で転記タグとする）。
  - 転記単位に対するコメント（@以降）は削除する。
  - 短単位を範囲に付与されるタグ (W), (D), (U), (X), (M), (O), (G), (F), (R) については，その情報を利用し，タグ付与範囲の開始・終了位置で必ず短単位が分割されるようにする。

(3) 慣用に従い「形態素解析」という用語を用いる。この用語については，小木曽 (2014: 89)，岡 (2019: 10–12) を参照。

  - 転記単位境界（ただしタグ + が使用されている箇所を除く）およびスペースの位置で必ず短単位が分割されるようにする。
- 形態素解析後処理
  - 解析器に渡さなかったもの（タグ記号，タグ (D) が付与された要素など）を復元する。
  - 「。」を転記タグとする。
  - タグ (D) の範囲の品詞は「言いよどみ」，(X) の範囲の品詞は「形態論情報付与対象外」，(R ＊) の範囲の品詞は「伏せ字」，(S ＊) の範囲の品詞は「歌」とする(4)。
  - 「(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，すでに定義されているものについて，適切な品詞を与える。

このように，CEJC 独自の転記タグの情報を活用することで，短単位解析精度の向上を図る。

3.2 発音形修正

ここでは，前節の処理により作成された短単位解析初期値データのうち，付加情報の一つである「発音形」をチェック・修正する。

3.2.1 発音形修正（自動）

まず，前もって作成しておいた発音形書き換えリスト（図 2 参照）に基づき，機械的に発音形を書き換える(5)。このリストには，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（発音形［誤］）が，書字形および正しい発音形（発音形［正］）とともに記されている（このとき短単位境界を “/” で指定する）。このリストの「書字形」と「発音形［誤］」の両方に合致するものが形態素解析初期値にあれば，その箇所の発音形を「発音形［正］」のように機械的に書き換える。図 3 に発音形の書き換え例を示す（「お父（さん）」「お母（さん）」）。この処理で書き換えられたレコードの品詞は「処理保留」となる。「処理保留」は作業上の品詞で，のちの作業で適切な品詞に書き換えられることになる。

なお，ここで修正されるのは発音形のみであり，他の情報（発音形以外の付加情報・境界）の誤りは修正されない。そのため，短単位境界に関する誤りがあったとしても，切り出された「短単位」を尊重して発音形書き換えリストに記述する。例えば，「初出店」（ハツ/シュッテン）が「初出/店」（ショシュツ/テン）のように解析された場合，発音形書き換えリストの「発音形［正］」フィールドには（「ハツ/シュッテン」ではなく）「ハツシュッ/テン」と記す（図 2 の最終行を参照）。

(4) タグ (R), (S) の要素が＊でなければ，通常の品詞を付与する。例えば，「(R 佐藤)」ならば，「名詞-固有名詞-人名-姓」という品詞を付与する。

(5) UniDic は，CEJC 短単位修正作業中も更新が続けられており，発音形書き換えリストにある発音形誤りのうち，かなりのものは，最新の UniDic（執筆時のバージョンは 2.3.0）では正しく解析されるようになっている（「お父（さん）」「お兄（さん）」など）。

書字形	発音形［誤］	発音形［正］
お/母	オ/ハハ	オ/カー
お/経	オ/ヘ	オ/キョー
お/香	オ/カ	オ/コー
お/終い	オ/ジマイ	オ/シマイ
お/大事	オ/オーゴト	オ/ダイジ
お/父	オ/チチ	オ/トー
お/兄	オ/アニ	オ/ニー
お/姉	オ/アネ	オ/ネー
お/人形	オ/ヒトガタ	オ/ニンギョー
：	：	：
初出/店	ショシュツ/テン	ハツシュッ/テン

図 2　発音形書き換えリスト（抜粋）

修正前				修正後		
書字形	発音形	品詞		書字形	発音形	品詞
えっとー	エットー	感動詞		えっとー	エットー	感動詞
お	オ	接頭辞		お	オ	接頭辞
父	**チチ**	名詞	→	父	**トー**	**処理保留**
さん	サン	接尾辞		さん	サン	接尾辞
と	ト	助詞		と	ト	助詞
お	オ	接頭辞		お	オ	接頭辞
母	**ハハ**	名詞	→	母	**カー**	**処理保留**
さん	サン	接尾辞		さん	サン	接尾辞
です	デス	助動詞		です	デス	助動詞

図 3　発音形自動書き換え例：「（お）父」と「（お）母」の発音形がリストに従い書き換えられる

3.2.2　発音形修正（手動）

次に，手動で発音形をチェックし，必要な箇所を修正する。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（例：研究会って，×「ケンキュー /アッ/テ」，○「ケンキュー/カイ/ッテ」）によるものである。前者に関して，特に数詞には注意が必要である（十「ジュー」「ジッ」「ジュッ」など，バリエーションが豊富であるため）。

この作業は，専用の発音形修正ツールを用いて行なう。これは Excel VBA で実装している。使用例を図 4 に示す。このツールから短単位解析結果を Excel スプレッドシートに読み込むと，発音形修正モードになる。このモードでは，シートへの直接入力は禁止され，専用ツールからの発音形修正のみ許される（ソート等はこのツールから可能）。読み込み後，チェック対象の短単位（主に漢字表記を含む短単位）を Excel スプレッドシート上で抽出し，個々の短単位について，音を聴取しながら，発音形の誤りがないかをチェックする(6)。誤りを発見した

(6) Excel スプレッドシート上から Praat を起動し，当該箇所を表示するという機能を持つツールを別途用意しており，音声の聴取はそのツールを用いて行なう。

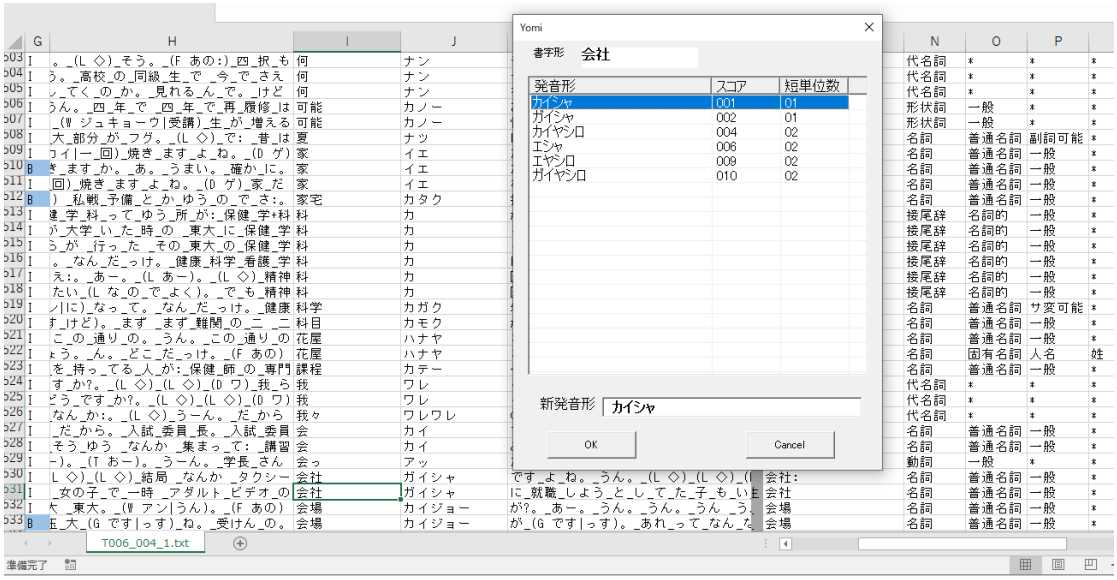

図 4　発音形修正ツール（スプレッドシート J 列が発音形；専用ツールで「会社」の発音形を修正中）

ら，修正するためのウインドウを起動する（図 4 の Yomi ウインドウ）。すると，そこに当該書字形に対する発音形候補が提示されるので，その中から正しいものを選ぶ。提示される候補は，MeCab と UniDic の出力結果に基づく。MeCab のオプション-N（N-Best 解の出力）を使って当該の書字形に対する複数の形態素解析結果をその場で取得し，さらにそこから可能な発音形を求めている。Yomi ウインドウに提示される候補の中に正しいものがない場合，「新発音形」のフィールド（図 4 のウインドウ下部）に手入力することで，正しい発音形を与える。このように，専用のツールを用いることで，作業の効率化を図っている。

ここまでの作業で，明らかな発音形誤りが発見された場合，発音形書き換えリスト（前述）にその誤りを追加し，次回から正しい発音形を与えられるようにしておく。

発音形自動書き換えのケースと同様，このツールを使って発音形を修正すると，品詞欄には自動的に「処理保留」というラベルが付与される。また，手動による発音形修正においても，発音形以外の誤りは修正しない。短単位境界に関する誤りがあったとしても，切り出された「短単位」を尊重して発音形を与える。例えば，「研究会って」が「研究/会っ/て」（ケンキュー/アッ/テ）のように解析された場合，切り出された「会っ」に対して正しい発音形「カイッ」を付与する。「オトツイ」と発音された「一昨日」が「一/昨日」（イッ/サクジツ）のように解析された場合には(7)，ひとまず「オ/トツイ」あるいは「オト/ツイ」といった発音形を与えておく（空の発音形は許容していないので，強制的に分割し，発音形に何らかの値を与えておく）。

CEJC のような話し言葉コーパスにおいては，発音に関する情報は非常に重要であり，高い精度が求められる。しかし，話はそれにとどまらない。作業工程全体から見た場合，この段階で発音形修正作業を行なうことには，次の二つの作業工程上の利点がある：

(7)「一昨日」（オトツイ）は説明のために用いている例であり，CEJC の転記テキストでは「オトツイ」「オトトイ」などは平仮名で表記される。

1. 作業者には必ずしも短単位に関する深い知識は要求されない（境界誤りや品詞誤りなどはここでは無視するため，作業者には音を正しく聞き取り，それを表記する能力さえあればよい[8]）。そのため，作業者を確保しやすい。
2. 正しい発音形をこの段階で与えておくことで，その情報を利用し，短単位情報（境界，および発音形以外の付加情報）の解析精度向上を図ることができる。

2 について，次の節で説明する。

3.3 形態素解析 (2)：発音形を考慮した解析

発音形修正の後に，再び形態素解析を行なう。入力は，発音形を修正した短単位解析データで，そのうち，品詞が「処理保留」となっているレコードを含む発話単位を再解析の対象とする。3.1 節で述べた形態素解析の前処理・後処理はここでも行われる。再解析にあたっては，MeCab のオプション-N（N-Best 解の出力）を使って複数の形態素解析結果を出力し，そのうち，発話単位で見た時に，入力の発音形に合致する最初の候補を選び，それで元の解析結果を上書きする。こうすることで，品詞「処理保留」のレコードが，適切な品詞が付与されたレコードで置き換えられることになる。また，短単位境界も正しいものになることが期待される。

具体例を使って説明する。図 5 は，「これが初出店です」（コレガハツシュッテンデス）という発話の形態素再解析結果を示している。入力に「処理保留」という品詞があるため，形態素再解析の対象となる。そこで，書字形から発話単位「これが初出店です」をくみ上げ，それを解析器の入力とし，MeCab+UniDic で複数の解析結果を出力する。そうしてから，出力された各候補について上から発話単位レベルで（短単位境界を無視して）入力と候補とで発音形とのマッチングを取る。第 1 候補は入力の発話単位発音形とのマッチングが取れないので，次の候補に進む。第 2 候補を見ると，入力の発話単位発音形と完全に一致するので，第 2 候補を選び，元の解析結果と置き換える。入力の発話単位発音形と一致するものが得られたので，第 3 候補以降は参照せず，この発話単位についての処理を終了させる。採用された第 2 候補を見ると，正しく境界が認定されており（「初出/店」→「初/出店」），それぞれの付加情報も正しく与えられている（処理保留/接尾辞→名詞/名詞）ことがわかる。このように，発音形を制約として用いることで，解析精度の向上を図っているわけである。

どの候補の発音形も入力の発音形と一致しない場合，書き換えは行なわず，入力をそのまま出力とする。品詞が「処理保留」となっている短単位が残るが，これはのちほど手動で修正する（それでも発音形の情報は正しく保持される）。

---

[8] 発音形表記の規約として，
(i) 「ヲ」「ヂ」「ヅ」は用いず，代わりにそれぞれ「オ」「ジ」「ズ」を用いる。
(ii) 助詞の「は」「へ」はそれぞれ（発音の通り）「ワ」「エ」と表記する。
(iii) 長音で発音されるものは（仮名遣いにかかわらず）「ー」で表記する
（例：高校，×「コウコウ」，○「コーコー」）。
などがあり，このうち (i)(ii) は発音形修正ツールによって入力の誤りを抑制できるが，(iii) は，作業者が手動で発音形を入力した場合，チェックが若干難しい。この種の誤りは別途チェックすることにしている。

	書字形	発音形	品詞				発話単位発音形
入力	これ	コレ	代名詞	*	*	*	コレガハツシュッテンデス
	が	ガ	助詞	格助詞	*	*	
	初出	ハツシュッ	処理保留	*	*	*	
	店	テン	接尾辞	名詞的	一般	*	
	です	デス	助動詞	*	*	*	
第 1 候補	これ	コレ	代名詞	*	*	*	コレガショシュツテンデス
	が	ガ	助詞	格助詞	*	*	
	初出	ショシュツ	名詞	普通名詞	サ変可能	*	
	店	テン	接尾辞	名詞的	一般	*	
	です	デス	助動詞	*	*	*	
第 2 候補	これ	コレ	代名詞	*	*	*	コレガハツシュッテンデス
	が	ガ	助詞	格助詞	*	*	
	初	ハツ	名詞	普通名詞	一般	*	
	出店	シュッテン	名詞	普通名詞	サ変可能	*	
	です	デス	助動詞	*	*	*	
第 3 候補	これ	コレ	代名詞	*	*	*	コレガショシュッテンデス
	が	ガ	助詞	格助詞	*	*	
	初	ショ	接頭辞	*	*	*	
	出店	シュッテン	名詞	普通名詞	サ変可能	*	
	です	デス	助動詞	*	*	*	

図 5　「これが初出店です」の形態素再解析結果（第 4 候補以降は省略）

3.4　解析結果の形態論 DB へのインポート

ここまでの処理で得られた短単位解析データをさらに加工したうえでいくつかのファイルに分割し，それらを形態論データベース (database; DB) にインポートする。ここで作成するテーブルは次の通り：

- 短単位テーブル：短単位情報
- 文字テーブル　：文字と文字位置の情報
- タグテーブル　：転記タグの種別，位置の情報
- 一般テーブル　：発話単位の開始・終了位置，発話単位の開始・終了時刻の情報

これらのテーブルを活用することにより，転記タグでの検索，当該短単位が属する発話単位の開始・終了時刻の取得（その情報を用いた音声の再生）などが可能となる。

3.5　大納言での人手修正

形態論情報管理ツール「大納言」（小木曽・中村 2014）を用い，データベースに格納された短単位情報を人手で修正していく。原則として，1 次作業では一つの会話を参加者ごとに，発話順にチェックしていく。品詞「処理保留」となっている短単位は（発音形以外は）誤りなので，適切な付加情報を付与する。この時，発音形を不用意に書き換えないよう注意する（特に分割・結合時）。現行の大納言には，一部の話し言葉系コーパスの検索結果から Praat を起動

する機能が備わっているので（西川 2018），音声を聴取したいときにはその機能を活用する。辞書 (UniDic) に未登録の語が出現したときは，その語の仮登録を行なう。

作業者には，月末に作業報告書を提出していただき，扱いに困る箇所，現行のマニュアルが想定していない箇所などを報告していただいている。それを踏まえ，月に 1 回，所内の関係者が集まり，対応を協議する会合を開いている。

3.6 系統的チェック・修正

1 次作業が一定量終了した段階で，全体を対象に系統的なチェックを行なう。チェック項目は数多くあるので，ここでは代表的なものだけ紹介する。

- 転記タグと短単位情報との整合性
  - タグ付き書字形に (D) が含まれているのに，品詞が言いよどみではない
  - 品詞が言いよどみなのに，タグ付き書字形に (D) が含まれていない
  - そのスコープ内では短単位分割されないはずのタグ
- 発話単位と短単位付加情報との整合性
  - 文末に出現した連体形（終止形の可能性）
  - 文頭以外の位置に出現した接続詞（接続詞以外の可能性）
  - 直後に名詞を伴う活用語終止形（連体形の可能性）
- 短単位付加情報一般
  - 品詞が空
  - 形状詞の直後に格助詞「に」「で」（助動詞「だ」の活用形の可能性）
  - 直後に助動詞「れる」「せる」を伴わない「さ」（「する」未然形）
  - 直後に特定の助動詞を伴わない語幹
  - 直後に特定の助動詞を伴わない未然形
  - 直後に特定の助詞・助動詞を伴わない「連用形-促音便」
- 短単位境界チェック
  - コーパスに含まれるすべての短単位書字形，および隣接する二つの短単位書字形列について，(i) 分割されているもの／結合しているもので揺れている，(ii) 分割位置が揺れている，ものを抽出し，チェック・修正

【例】

前文脈	書字形	後文脈
リメートル＼答えは一	ミリ/メートル	ん百センチメートルか
リですか＼まず百十七	ミリメートル	って書いてごらん＼答
商店街あるじゃん＼で	美園/町	に向かって右サイドな
かり売ってるお店がさ	美園町	にあるじゃん＼知らな

- 助詞・助動詞・接尾辞の連接チェック
  - すべての助詞・助動詞・接尾辞について，その直前の短単位の活用形（空の場合もある）に揺れがあるものを抽出し，チェック・修正

【例】

頻度	活用形	品詞	語彙素読み	語彙素
81		助詞-格助詞	ヨリ	より
16	終止形-一般	助詞-格助詞	ヨリ	より
2	終止形-撥音便	助詞-格助詞	ヨリ	より
35	連体形-一般	助詞-格助詞	ヨリ	より

これらの項目の抽出は，大納言で SQL（問い合わせ言語）を使うか，あるいはデータベースからエクスポートしたテキストファイルに対してスクリプト言語を使って行なう。なお，抽出されたものが必ずしも誤りだとは限らない点には注意が必要である。また，抽出されたものが，短単位情報ではなく，転記テキストの誤りの可能性もある（転記タグに関する誤り，聞き取りに関する誤り，単純な誤り，など)。その場合には，転記テキストの担当者に報告し，そちらの修正を待つ。

4. 転記テキスト修正に伴う短単位 DB 書字形更新および短単位付加情報再付与

転記テキストもアノテーションの一つである以上（小磯 2014: 49–50)，誤りは避けられない。転記テキストの修正は，短単位 DB にインポートした後も継続して行なわれているので（実際のところ，短単位情報を付与する中で，あるいは短単位情報を用いることで発見される転記誤りが数多くある)，最新の転記テキストと形態論 DB との間には齟齬が生じてしまう。この種の齟齬を解消するため，ある段階で，形態論 DB 書字形の更新作業を行なう。

形態論 DB 書字形の更新を行なうために，まず形態論 DB 短単位テーブルの書字形と，最新の転記テキストとの差分を取り，それをもとに書字形更新のためのテーブルを作成する。このテーブルを使って，DB 上で短単位テーブルの書字形を書き換える。

書字形が書き換えられると，そのレコードの品詞は「転記修正」となり，他の付加情報は空となる。更新終了後，作業者は「転記修正」を検索し，そのレコードに適切な付加情報を与える。発音形の情報も失われてしまうので，作業は必要に応じて音を聴取しながら行なう。

なお，書字形を持つ短単位テーブルだけでなく，それ以外のテーブル（文字テーブル，タグテーブル，一般テーブル；3.4 説参照）も更新が必要である。これらについては，現状のテーブルとの差分を取ることはせず，最新の転記テキストから新規に作成し，現状のテーブルをそれらで置き換えることで対応する。

5. おわりに

『日本語日常会話コーパス』(CEJC) の短単位解析について，作業工程を中心に報告した。作業工程の策定にあたっては，作業の効率化と高精度の解析の両方を可能にするものを目指した。このことがどこまで実現されているかは今後の検討が必要だが，少なくとも，どのみち必要な発音形のチェック・修正を早めに行ない（3.2 節)，修正した発音形をもとに短単位情報の更新を行なう（3.3 節)，という点では，作業の効率化と高精度の解析の両方をそれなりに達成できていると思われる。

ここで述べた作業工程を経て作成された CEJC 短単位データをもとに，長単位情報と韻律情報の初期値が作成され，現在人手による修正が進められている。この種のデータが利用可能に

なれば，さらに別の，効率的な短単位のチェックが可能になり，精度の向上につながることが期待される。そのための方法を現在検討中である。

最後に，本稿で述べた作業工程は，同じく現在構築中の『昭和話し言葉コーパス』（丸山2019）にも適用され，同コーパスの短単位解析作業が進行中であることを付言しておく。

謝　辞

本研究は国立国語研究所共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」および「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」の成果です。

文　献

Boersma, Paul and Weenink, David (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.40, retrieved 11 May 2018 from http://www.praat.org/

小磯花絵 (2014)「話し言葉の書き起こし」小磯花絵（編）『話し言葉コーパス：設計と構築』（講座日本語コーパス 3）. 朝倉書店, pp. 33–53.

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, pp. 89–96.

丸山岳彦 (2019)「『通時音声コーパス』の可能性と問題点：『昭和話し言葉コーパス』の構築と分析」『言語資源活用ワークショップ発表論文集』

西川賢哉 (2018)「『日本語日常会話コーパス』構築における Praat の利用」『言語資源活用ワークショップ発表論文集』3, pp. 142–147.（http://doi.org/10.15084/00001647 よりダウンロード可能）

小木曽智信 (2014)「形態素解析」山崎誠（編）『書き言葉コーパス：設計と構築』（講座日本語コーパス 2）. 朝倉書店, pp. 89–115.

小木曽智信・中村壮範 (2014)「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」『自然言語処理』21 巻 2 号, pp. 301–332.

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』（講座日本語コーパス 2）. 朝倉書店, pp. 68–88.

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕 (2011)「『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（下）」特定領域研究「日本語コーパス」平成 22 年度研究成果報告書（JC-D-10-05-02）（http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf よりダウンロード可能）

岡照晃 (2019)「言語研究のための電子化辞書」伝康晴・荻野綱男（編）『コーパスと辞書』（講座日本語コーパス 7）. 朝倉書店, pp. 1–28.

臼田泰如・川端良子・西川賢哉・石本祐一・小磯花絵 (2018)「『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」『国立国語研究所論集』15, pp. 177–193.（http://doi.org/10.15084/00001602 よりダウンロード可能）

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	https://www2.ninjal.ac.jp/conversation/
UniDic	https://unidic.ninjal.ac.jp/
MeCab	https://taku910.github.io/mecab/
ELAN	http://tla.mpi.nl/tools/tla-tools/elan/
Praat	http://www.praat.org/