『日本語日常会話コーパス』に対する短単位情報付与：作業工程と評価

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	5
ページ	324-330
発行年	2020
URL	http://doi.org/10.15084/00003172

『日本語日常会話コーパス』に対する短単位情報付与：作業工程と評価

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)

Short Unit Word Annotation for the Corpus of Everyday Japanese Conversation: Procedures and Evaluation

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

要旨

『日本語日常会話コーパス』(CEJC) の短単位情報付与作業では，以下のような作業工程を踏んでいる：(i) 転記を MeCab（解析器）+UniDic（解析辞書）で自動解析，(ii) 音声を聴取しながら，付加情報の一つである「発音形」のみを人手修正，(iii) 人手修正された発音形を尊重しつつ再び自動解析，(iv) 短単位情報（境界情報，発音形以外の付加情報）を人手修正。この作業工程の妥当性を検証するため，人手修正済みデータを対象に，複数の版の現代話し言葉 UniDic（Ver2.2.0, 2.3.0, 3.0.1）で自動解析をしなおし，出力を比較した。その結果，どの版の UniDic を使っても，人手修正された発音形の情報を用いる方が，そうでない場合に比べ，短単位情報の精度向上を見込めることがわかった。特に，古い版の UniDic (Ver2.2.0) ではそれが顕著であった（境界＋品詞＋語彙素（F 値）：0.944 → 0.962）。一方で，最新版の UniDic (Ver3.0.1) では効果は限定的である（同：0.976 → 0.979）。

1. はじめに

国立国語研究所では，2016 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。公開時には，転記テキストの他，形態論情報（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が提供される予定である。

これらのアノテーションの一つである短単位情報は，CEJC 全体（約 200 時間）に対して提供される。短単位情報付与作業にあたっては，国語研で構築された他のコーパスと同様，機械による自動処理—形態素解析器 MeCab（工藤他 2004）および形態素解析用辞書 UniDic（岡 2019）を用いた形態素解析 (morphological analysis)—と，人手による処理—解析結果の確認および誤解析箇所の修正—とを併用する。ただし，CEJC の短単位情報付与作業では，限られた人的資源で高精度の解析を達成するため，独自の作業工程を設けることとした。本稿では，

* nishikawa[at]ninjal.ac.jp

fileID	speakerID	startTime	endTime	pause	text
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。

表 1　転記例

書字形	発音形	品詞	活用型	活用形	語彙素読み	語彙素
あー	アー	感動詞-一般			アア	ああ
雲取	クモトリ	名詞-普通名詞-一般			クモトリ	雲取
も	モ	助詞-係助詞			モ	も
イ	イ	言いよどみ				
一	ヒト	名詞-数詞			ヒト	一
組	クミ	名詞-普通名詞-助数詞可能			クミ	組
だけ	ダケ	助詞-副助詞			ダケ	だけ
外人	ガイジン	名詞-普通名詞-一般			ガイジン	外人
の	ノ	助詞-格助詞			ノ	の
ご	ゴ	接頭辞			ゴ	御
一行	イッコー	名詞-普通名詞-一般			イッコウ	一行
みたい	ミタイ	形状詞-助動詞語幹			ミタイ	みたい
の	ノ	助詞-準体助詞			ノ	の
帰る	カエル	動詞-一般	五段-ラ行	連体形-一般	カエル	返る
時	トキ	名詞-普通名詞-副詞可能			トキ	時
えー	エー	感動詞-一般			エエ	ええ

表 2　短単位解析例（ID 等は省略）

CEJC 短単位に関する作業工程を述べ，その工程の効果を検証する。

2. 作業工程

CEJC の短単位情報付与作業の工程について簡単に述べる(1)。形態素解析は，入力文字列（転記テキスト；表 1 参照(2)）に境界を与え（分割し），その境界で区切られた（分割された）単位に対して付加情報（発音形，品詞，語彙素，など）を付与する作業とみなすことができる（表 2 参照）(3)。

一般にこの作業は，(i) 形態素解析器と形態素解析辞書を用いて自動で形態素解析した後，(ii) 手動で境界・付加情報を修正する，という 2 ステップで実施される。それに対し，CEJC では，次に示すような若干複雑なステップを踏んでいる：

(1) CEJC の短単位情報付与作業の工程は，すでに西川・渡邊 (2019) で報告した。詳細はそちらを参照されたい。
(2) CEJC の転記テキストの仕様については，臼田他 (2018) を参照。
(3) 短単位の設計方針や，付加情報（品詞，語彙素，発音形，等）については，UniDic のサイトの「用語集」(`https://unidic.ninjal.ac.jp/glossary`)，小椋 (2014)，小椋他 (2011) などを参照。

1. ［自動］形態素解析
2. ［人手］発音形修正
3. ［自動］形態素解析（修正された発音形を考慮）
4. ［人手］発音形以外の短単位情報（境界・付加情報）修正

このように，(i) 短単位情報を，発音形とそれ以外の情報（境界，品詞，語彙素）に分離し，前者を最初の人手修正の対象とすること，(ii) 形態素解析を 2 回実施し，2 回目の解析では，人手修正された発音形を考慮した解析を行なうこと，の 2 点が CEJC 短単位解析の工程の特徴である。

発音形の修正に特化した工程を設ける狙いは次の二つである。

A. **発音形それ自体の精度向上**

話し言葉のコーパスにおいては，発音に関する情報は重要であり，高い精度が求められる。CEJC での発音に関する情報は，部分的には転記でも提供されるものの，一般的には短単位情報の一つである発音形で提供される(4)。そのため，発音形に関する独立の工程を設け，網羅的にチェック・修正することとした。この工程では，「お/父/さん」（チチ），「お/母/さん」（ハハ）のような単純な解析誤り(5)や，「日本」（ニホン，ニッポン），「研究/所」（ショ，ジョ）のような，候補が複数想定される発音(6)に関する誤りを，音声を聴取しつつ正しく修正する。

B. **発音形を活用した短単位情報全般の精度向上**

発音形が決まればその他の情報（境界，品詞，語彙素）が決まる場合がある。例えば，「初出店」「女将軍」「発表会って」は，以下（矢印左側）のように誤解析されることがあるが，

初出/店（ショシュツ/テン）	→	初/出店（ハツ/シュッテン）
女将/軍（オカミ/グン）	→	女/将軍（オンナ/ショーグン）
発表/会っ/て（ハッピョー/アッ/テ）	→	発表/会/って（ハッピョー/カイ/ッテ）

発音形の情報が与えられれば，矢印右側に示すように，正しく解析することができる（ここでは境界の情報しか示していないが，品詞・語彙素の情報も正しく解析される(7)）。このように，人手修正された発音形を使って短単位情報全般の精度の向上を図る。

---

(4) 同じ話し言葉のコーパスである『日本語話し言葉コーパス』(CSJ) では，発音に関する情報は，短単位情報とは独立の，転記テキストにおいて提供されている。

(5) ここに挙げた「お/父/さん」（チチ→トー），「お/母/さん」（ハハ→カー），また，「言語/道断」（ゲンゴ→ゴンゴ）など，直前または直後に来る要素によって発音形が一意に決定できるものについては，あらかじめリストを作成しておき，そのリストに基づき，人手による作業の前に機械的に発音形を書き換えることにしている。

(6) この種の曖昧性が発生しうる「読み」情報については，CEJC 転記テキストにおいて，タグ (Y) を用いて，(Y ニッポン | 日本)，(Y ショ | 所) のように表現される。ただし，タグ (Y) は転記において必ずしも網羅的に付与されているわけではない。また，形態素解析結果には，「お/父/さん」（チチ）のような，タグ (Y) の対象とならないような単純な解析誤りも含まれる。したがって，短単位レベルでの発音形のチェックを省略することはできない。

(7) CEJC に出現したものではないが，業界で有名（？）な例の一つである「外国人参政権」についても，発音形の情報があれば，正しい短単位情報（境界・品詞・語彙素）を得ることができる。

外国/人参/政権（ガイコク/ニンジン/セーケン） → 外国/人/参政/権（ガイコク/ジン/サンセー/ケン）

上記 B の目的のため，発音形を人手修正した後，形態素解析を再度実施する。再解析の対象は，発音形が修正された短単位を含む発話単位[8]とする。再解析の際には，形態素解析器 MeCab のオプション-N を用いて，複数の解析結果（N-best 解）を出力し，修正した発音形に合致する最初の解析結果を選ぶ。これにより，上に挙げた「初出店」「女将軍」「発表会って」は正しく解析される。

3. 精度と評価

前節で，人手修正された発音形を用いて形態素解析を実施することで，短単位情報全体の精度を向上させる手法について述べた。この手法の妥当性を検証するため，人手修正済みの短単位データのサブセットを正解データとみなしたうえで，対応する転記テキストを解析しなおし，解析精度を求めた。結果を表 3 に示す。

詳細は以下の通りである：

- 正解データとして，CEJC 第 2 期内部公開用データ（137 会話，約 62 万短単位）を用いた。これらは学習用データとしては使用されておらず，UniDic にとっては未知のデータである。
- 解析に用いた辞書は，以下の 3 種類の UniDic である（括弧内は公開年月日）[9]：
  - 現代話し言葉 UniDic Ver2.2.0 （2017 年 9 月 5 日）
  - 現代話し言葉 UniDic Ver2.3.0 （2018 年 4 月 10 日）
  - 現代話し言葉 UniDic Ver3.0.1 （2019 年 12 月 17 日）

  これらは実際に CEJC 短単位解析作業で使用されたものである（最新版 Ver3.0.1 は現在も使用されている）。新しい版は，以前の版よりも登録語数が多い（すなわち、語数は統制されていない）。また，どの版にも登録されていない語（未知語）が正解データには含まれる。
- 形態素解析は，発音形を考慮しない解析（発音なし）と，発音形を考慮した解析（発音あり）の 2 回実施した。
- 評価にあたっては，小木曽 (2014) に従い，「境界」「品詞」「語彙素」の 3 段階の評価基準を設けた：
  - 境界：短単位境界が正しいかどうか
  - 品詞：境界に加えて，短単位の品詞・活用型・活用形を正しく選択できたか
  - 語彙素：境界と品詞に加えて，語彙素・語彙素読みが正しく行われたか

  品詞の評価は境界が正しいことを前提としており，語彙素の評価は境界と品詞が正しいことを前提としている。そのため，必ずこの順に厳しい評価となる（小木曽 2014:103）。

表 3 を見ると，どの版の UniDic を使っても，人手修正された発音形の情報を用いる方が，そうでない場合に比べ，短単位情報の精度向上を見込めることがわかる。特に，古い版の UniDic (Ver2.2.0) ではそれが顕著である（境界＋品詞＋語彙素（F 値）：0.944 → 0.962）。このこと

[8] 「発話単位」については，臼田他 (2018) を参照。
[9] 執筆時点で，これらのバージョンの UniDic はすべて UniDic のサイトからダウンロード可能である。

		Ver2.2.0	Ver2.3.0	Ver3.0.1
発音		0.974	0.979	0.981
境界	発音なし	0.984	0.991	0.993
	発音あり	0.986	0.992	0.993
品詞	発音なし	0.958	0.971	0.978
	発音あり	0.963	0.972	0.979
語彙素	発音なし	0.944	0.957	0.976
	発音あり	0.962	0.971	0.979

表 3　解析精度（F 値）

は，前節で述べた手法がそれなりに効果的であることを示すと考えられる。

一方で，最新版の UniDic (Ver3.0.1) では効果は限定的である（同：0.976 → 0.979）。これは，UniDic (Ver3.0.1) の解析精度が，以前の版に比べ格段に向上したからだと推測できる。このことは，発音形の情報を使用するか否かにかかわらず，より新しい版の UniDic の方が，それ以前の版の UniDic より高い解析精度を示すことからも裏付けられる。

4. 誤解析

人手修正した発音形を与えたとしても，また，高精度の解析辞書を用いたとしても，解析にはなお誤りが残る。また，当然のことながら，もともと発音形は正しいのに他の短単位情報が誤るケースもある。今後の修正作業の参考のため，誤解析の内訳を集計した。

表 4 に発音形書き換え箇所の短単位誤解析数を，表 5 に発音形書き換え箇所以外の短単位誤解析数を示す。

集計の詳細は以下の通り：

- 前節での評価と平行的に，「境界」「品詞」「語彙素」の 3 段階の評価基準を設けた：
  - 境界：短単位境界が誤り
  - 品詞：境界は正しいが，短単位の品詞・活用型・活用形が誤り
  - 語彙素：境界・品詞は正しいが，語彙素・語彙素読みが誤り
- 「未解析」は，形態素解析器が「未知語」として出力したもの，あるいは，人手で修正した発音形を考慮すると解析できなかったもの，のいずれかである

	Ver2.2.0	Ver2.3.0	Ver3.0.1
未解析	648	584	431
誤解析：境界	373	345	239
誤解析：品詞	229	163	145
誤解析：語彙素	2	4	2
合計	1262	1096	817

表 4　発音形書き換え箇所の短単位誤解析

	Ver2.2.0	Ver2.3.0	Ver3.0.1
未解析	742	605	385
誤解析：境界	4690	2581	2215
誤解析：品詞	13259	11269	8016
誤解析：語彙素	261	433	171
合計	18952	14888	10787

表 5　発音形書き換え箇所以外の短単位誤解析

誤解析の傾向を一般化するのは難しいが，いずれの場合も，未知語（未登録語）が誤解析を引き起こす一因となっていることが多いように思われる。詳しい分析は今後の課題としたい。

5. おわりに

本稿では,『日本語日常会話コーパス』(CEJC) の短単位解析作業の工程について述べ，その工程の妥当性を検証した。

現在までに，284 万短単位（公開予定の CEJC データの約 8 割）が，2 節で述べた工程を経て，形態論情報管理システムに登録されており，管理システム上で短単位情報の修正を進めているところである。今後は，修正作業を継続するとともに，4 節で述べたような誤解析をより詳しく分析し，修正の手掛かりとすることで，高精度の短単位データを提供できるよう努めたい。

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

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.

西川賢哉・渡邊友香 (2019) 「『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」『言語資源活用ワークショップ発表論文集』4, pp. 238-250.（http://doi.org/10.15084/00002575 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/