『日本語歴史コーパス 明治・大正編VI落語 SP 盤』
音声データアラインメントの概要

2023 年 1 月 10 日
松崎安子・西川賢哉・石本祐一・中村壮範・小木曽智信

『日本語歴史コーパス　明治・大正編VI落語 SP 盤』（短単位データ 1.0）は、2022 年 3 月に公開された『日本語歴史コーパス　明治・大正編VI落語 SP 盤』（短単位データ 0.9）のテキストデータに対し、落語家の音声データを対応付けアップデートしたものである。本文書では音声データアラインメントの工程と、中納言検索結果を通しての音声データ試聴の方法を解説する。

なお、以下『日本語歴史コーパス明治・大正編VI落語 SP 盤』は CHJ「明治・大正編VI落語 SP 盤」と略し表記する。

１．音声アラインメントの工程

１．１　音声データにおけるノイズ除去

CHJ「明治・大正編VI落語 SP 盤」の音源は岡田コレクション「学術研究用デジタル音源集」（日外アソシエーツ株式会社）を用いている。ただし、音源の音声データには口演者の声以外の音が混入しており、そのままでは聞きづらいものとなっていることから、口演者の声の音質を維持しながらその他の雑音成分を減らすため、Wiener フィルタによる雑音抑圧と調波成分の再生成を組み合わせた手法(Plapous et al. 2006）による雑音抑圧処理を行った。なお、この雑音抑圧処理手法の提案者の一人により MATLAB 言語による実装コードが無償で公開されている。

１．２　Praat を使用した音声アラインメント

前節のように用意した音声データと CHJ「明治・大正編VI落語 SP 盤」の書き起こしテキストとの対応づけにあたっては音声分析用フリーソフトウェア Praat（Version6.2.12 2022 年 4 月 17 日リリース版)を使用した。

Praat 使用下では、音声データに対し既存のテキストを転記単位で割り付けることで TextGrid（Praat アノテーション形式）を作成した（図 1、図 2 参照)。作成にあたって、国立国語研究所が構築した日本語日常会話コーパス(CEJC)における転記テキストについて規定している臼田泰如ほか(2018)の基準を参考とし、落語家の発声の切れ目について次のような場合を転記単位の境界とした。

【転記単位の境界】

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

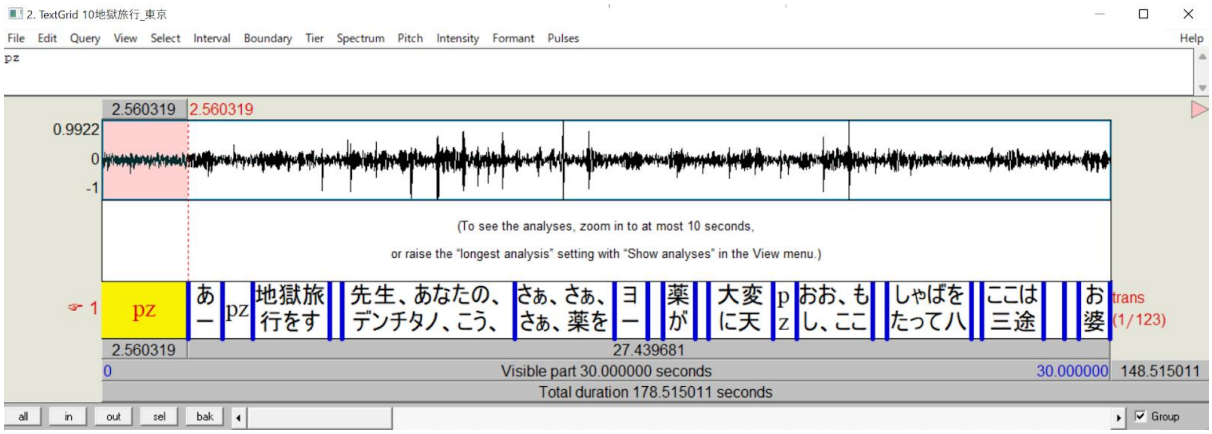

図１　Praat で TextGrid を作成

```
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0
xmax = 178.51501133786850
tiers? <exists>
size = 1
item []:
    item [1]:
        class = "IntervalTier"
        name = "trans"
        xmin = 0
        xmax = 178.51501133786850
        intervals: size = 123
        intervals [1]:
            xmin = 0
            xmax = 2.5603191509176035
            text = "pz"
        intervals [2]:
            xmin = 2.5603191509176035
            xmax = 3.5755255765126255
            text = "あー、お前は、"
```

図２　テキストエディタに TextGrid を展開

１．３　テキストデータの文字位置と音声データの時刻情報の関連付け

次に、テキストデータの文字位置と音声データの時刻情報とを関連付けを行った。音声データの時刻情報と関連付けるためのテキストの文字位置については、データベースから出力した短単位データを使用した。短単位データには短単位境界とその開始･終了文字位置の情報が付与されており（表 1)、前節の Praat での作業で作成した TextGrid と短単位データを照合させることで転記単位での開始･終了の時刻と文字位置の対応表を作成した。

さらに、この対応表に基づいて、形態論情報テーブルに、当該短単位が属する転記単位の開始時刻・終了時刻を付加した（表 3)。

表１　データベースから出力した短単位データ（一部）

file	start	end	boundary	orthToken
10地獄旅行_東京	10	30	B	あー
10地獄旅行_東京	30	40	I	、
10地獄旅行_東京	40	60	I	お前
10地獄旅行_東京	60	70	I	は
10地獄旅行_東京	70	80	I	、
10地獄旅行_東京	80	100	I	地獄
10地獄旅行_東京	100	120	I	旅行
10地獄旅行_東京	120	130	I	を
10地獄旅行_東京	130	150	I	する
10地獄旅行_東京	150	160	I	薬
10地獄旅行_東京	160	170	I	を
10地獄旅行_東京	170	190	I	好む
10地獄旅行_東京	190	200	I	か
10地獄旅行_東京	200	210	I	い
10地獄旅行_東京	210	220	I	。

表２　開始・終了の時刻と転記単位の文字位置の対応表（一部）

file名	文字 開始位置	文字 終了位置	音声 開始時刻	音声 終了時刻	転記（書字形）	音声ファイル名
10地獄旅行_東京	10	80	2.56	3.576	あー、お前は、	10地獄旅行_東京
10地獄旅行_東京	80	220	4.478	6.786	地獄旅行をする薬を好むかい。	10地獄旅行_東京
10地獄旅行_東京	220	640	7.135	12.269	先生、あなたの、デンテタノ、こう、みやげで評判のお話をするたあを伺いとうがすなぁ。	10地獄旅行_東京
10地獄旅行_東京	640	830	12.269	15.238	さぁ、さぁ、さぁ、薬をお上がんなさい。	10地獄旅行_東京
10地獄旅行_東京	830	860	15.238	16.225	ヨー。	10地獄旅行_東京
10地獄旅行_東京	860	920	16.687	17.643	薬が早いね、	10地獄旅行_東京
10地獄旅行_東京	920	1100	18.022	19.958	大変に天気は暗くなってきちまったね。	10地獄旅行_東京
10地獄旅行_東京	1100	1300	20.644	22.909	おお、もし、ここは日本ではありませんよ。	10地獄旅行_東京
10地獄旅行_東京	1300	1500	23.34	25.85	しゃばをたって八幡行きと踏み込んでくる。	10地獄旅行_東京
10地獄旅行_東京	1500	1610	26.169	27.982	ここは三途川の入口だ。	10地獄旅行_東京

表３　形態論情報と転記単位の開始・終了時刻の対応表（一部）

	subCorpusName	sampleID	start	end	boundary	orthToken	pronToken	reading	lemma	pos	soundFileName	startTime	endTime
1	明治・大正-落語SP盤	60R圓遊1903_01010	10	30	B	あー	アー	アー	あー	感動詞-フィラー	10地獄旅行_東京	2.560	3.576
2	明治・大正-落語SP盤	60R圓遊1903_01010	30	40	I	、			、	補助記号-読点	10地獄旅行_東京	2.560	3.576
3	明治・大正-落語SP盤	60R圓遊1903_01010	40	60	I	お前	オマエ	オマエ	御前	代名詞	10地獄旅行_東京	2.560	3.576
4	明治・大正-落語SP盤	60R圓遊1903_01010	60	70	I	は	ワ	ハ	は	助詞-係助詞	10地獄旅行_東京	2.560	3.576
5	明治・大正-落語SP盤	60R圓遊1903_01010	70	80	I	、			、	補助記号-読点	10地獄旅行_東京	2.560	3.576
6	明治・大正-落語SP盤	60R圓遊1903_01010	80	100	I	地獄	ジゴク	ジゴク	地獄	名詞-普通名詞-一般	10地獄旅行_東京	4.478	6.786
7	明治・大正-落語SP盤	60R圓遊1903_01010	100	120	I	旅行	リョコー	リョコウ	旅行	名詞-普通名詞-サ変可能	10地獄旅行_東京	4.478	6.786
8	明治・大正-落語SP盤	60R圓遊1903_01010	120	130	I	を	オ	ヲ	を	助詞-格助詞	10地獄旅行_東京	4.478	6.786
9	明治・大正-落語SP盤	60R圓遊1903_01010	130	150	I	する	スル	スル	為る	動詞-非自立可能	10地獄旅行_東京	4.478	6.786
10	明治・大正-落語SP盤	60R圓遊1903_01010	150	160	I	薬	クスリ	クスリ	薬	名詞-普通名詞-一般	10地獄旅行_東京	4.478	6.786
11	明治・大正-落語SP盤	60R圓遊1903_01010	160	170	I	を	オ	ヲ	を	助詞-格助詞	10地獄旅行_東京	4.478	6.786
12	明治・大正-落語SP盤	60R圓遊1903_01010	170	190	I	好む	コノム	コノム	好む	動詞-一般	10地獄旅行_東京	4.478	6.786
13	明治・大正-落語SP盤	60R圓遊1903_01010	190	200	I	か	カ	カ	か	助詞-終助詞	10地獄旅行_東京	4.478	6.786
14	明治・大正-落語SP盤	60R圓遊1903_01010	200	210	I	い	イ	イ	い	助詞-終助詞	10地獄旅行_東京	4.478	6.786
15	明治・大正-落語SP盤	60R圓遊1903_01010	210	220	I	。			。	補助記号-句点	10地獄旅行_東京	4.478	6.786

２．中納言検索画面からの音声視聴

前節までのように CHJ のテキストと対応付けられた音声情報は、中納言の検索結果画面において直接試聴することができる。 図 3 には中納言での検索画面の一例をあげている。このうち、左端のサンプル ID の欄には用例のサンプル ID とともに音声再生用ボタン ▶ が表示されている。これをクリックすると、図 4 のように「(サンプル ID の表示) 再生中」のポップアップが表示され、検索したキーが属する転記単位を含んだ前後 10 秒ほどの音声が再生される。

図 3　中納言検索結果画面から音声を試聴する（イメージ図）

図 4　音声再生中のポップアップ（イメージ図）

付記

本コーパスは、国立国語研究所共同研究プロジェクト「開かれた共同構築環境による通時コー パスの拡張」の研究成果によるものである。

関連資料

SP 盤貴重音源　岡田コレクション「学術研究用デジタル音源集」日外アソシエーツ株式会社

関連 URL

コーパス検索アプリケーション「中納言」　https://chunagon.ninjal.ac.jp/

国立国語研究所『日本語歴史コーパス明治・大正編VI落語 SP 盤』

https://clrd.ninjal.ac.jp/chj/meiji_taisho.html#rakugo

Praat 公式サイト https://www.fon.hum.uva.nl/praat/

Wiener filter for Noise Reduction and speech enhancement - MATLAB Central

https://www.mathworks.com/matlabcentral/fileexchange/24462

参考文献

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
http://doi.org/10.15084/00001602

服部紀子・松崎安子・小木曽智信（2022）「『日本語歴史コーパス 明治・大正編 VI 落語 SP 盤』の公開」『日本語学会 2022 年度春季大会予稿集』pp.199-204

松崎安子・西川賢哉・石本祐一・中村壮範・小木曽智信（2022）「『日本語歴史コーパス明治・大正編VI落語 SP 盤』音声アラインメントの公開」pp.115-118

Cyril Plapous, Claude Marro, and Pascal Scalart (2006). “Improved Signal-to-Noise Ratio Estimation for Speech Enhancement.” IEEE Transactions on Audio, Speech and Language Processing, 14:6, pp. 2098–2108.