

『日本語歴史コーパス 明治・大正編VI落語 SP 盤』 テキストの凡例と「中納言」表示項目について

2023 年 1 月 10 日 服部紀子 松崎安子

1. はじめに

『日本語歴史コーパス 明治・大正時代編VI落語 SP 盤』(Ver.0.9) は、金澤裕之・矢島正浩編 (2019)『SP 盤落語レコードがひらく近代日本語研究』(笠間書院) で扱われた東京 76 作品、大阪 51 作品を収録対象としている。

本コーパスの最大の特徴は、『日本語歴史コーパス』の中で初めて音声データを取り扱った点にある。また、従来、書き言葉を扱ってきた「明治・大正編」の中で初めて話し言葉を 主対象とした点も新しい。「江戸時代編」においては、「Ⅰ洒落本」、「Ⅱ人情本」、「Ⅲ近松浄 瑠璃」と近世の話し言葉研究には欠かせない言語資料を収録して公開しているが、SP 盤に 録音された落語は、近世から連続する話し言葉の資料として価値を持つものである。

本コーパスの本コーパスの電子化テキストは、金澤裕之・矢島正浩編 (2019) で文字化されたデータを基礎としつつ、可能な範囲でノイズを取り除いた音声データをもとに修正を加えて作成したものである。さらに『日本語歴史コーパス』(以下、CHJ) の他の時代のコーパスと整合的な形態論情報を付与するため、適宜校訂を加えている

校訂を経たコーパスは、検索対象サンプルとして原則 1 演目を単位として分割され¹、各サンプルは個別の ID と書誌的情報(落語家・録音年・地域等に関する情報)を有する。

この文書では、コーパス本文の成り立ちと、検索アプリケーション「中納言」における検索結果の表示項目について、例示しながらその概要を示す。

なお、研究上必要と思われる情報を、できるだけ音声資料に即して記述するよう努めたが、不十分・不適切な箇所が残存する可能性もある。適宜、原資料の情報を確認されることを推奨する。

2 テキストの凡例

本コーパスでは、本文に対して、CHJ の他の時代や『現代日本語書き言葉均衡コーパス』(BCCWJ) 等の他のコーパスと整合的な形態論情報を提供することを目的として、形態素解析辞書 UniDic に基づいた短単位情報を付与している。そのため、UniDic による形態素解析に適した解析用本文を用意する必要があり、金澤裕之・矢島正浩編 (2019) で文字化されたデータを基礎としつつ、適宜修正を加えた。

本文の表記は、全体的に漢字ひらがな交じり文をベースとしながら、長音や促音表記や漢字への変換なども可能な限り現代の一般的な表記法や用字となるようにした²。例えば、語彙素「言う」の終止形および連体形は、音声によればその出現発音形は「ユー」である。だが、コーパスの本文としてのテキストでは「ゆう」とせず、

¹ 作品名において「大学・手と足の喧嘩」「近江八景・小噺」のように中黒点のあるものは、二つの演目の語りが連続し、明確に分割が困難であったため 1 サンプルにまとめて収録している。

² 本コーパスの本文テキストに使用した文字の範囲は、JIS X 0213 (JIS の文字コード規格) の文字集合 (JIS 漢字の第 4 水準までを含む) に準拠している。

現代仮名遣いの規則に従い「いう」の表記を採った³。このように形態素解析や検索に適した形式に改変したところがある。

上記のような事情から、「中納言」検索画面においては「検索用文字列」、「原文文字列」、「原文 KWIC」の表示に違いはない。

サンプル ID	開始位置	連番	コア	前文脈	キー	後文脈	語彙素読み	語彙素	語形	品詞	活用型	活用形	原文文字列	振り仮名
60Rむら 1903_01036	1850	1210	1	最初から相談を致しますと、大家が録談の事で断られるといかん	言う	ので、あきぢいさんの久八という、多助の友達を呼びました。#さあさあ	イウ	言う	ユウ	動詞-一般	五段-ワ行	連体形-一般	言う	
				初めから相談を致しますと、大家が録談の事で断られるといかん	言う	ので、あきぢいさんの久八という、多助の友達を呼びました。#さあさあ、こっちずーっと、梅屋さんお入り。								

図1 「中納言」での本文表示欄と「原文文字列」「原文 KWIC」表示欄

ただし、音声資料の文字化について不十分な箇所が残存する可能性もある。利用に際しては、必要に応じて書誌情報に記す音声情報を確認されたい。

3. テキストの範囲とサンプル

本コーパスに収録したテキストは、落語家が語った演目が SP 盤に録音された音声そのまま全て書き起こしたものである。それらの各演目を、本コーパスでは 1 サンプルとし、それぞれに書誌的・言語的な情報を付与することとした。

個々のサンプルは、サンプル ID という個別に認識される ID を持つ、サンプル ID は 15 桁からなり、その構成は以下の表 1 のとおりである。

表 1 本コーパスにおけるサンプル ID の構成

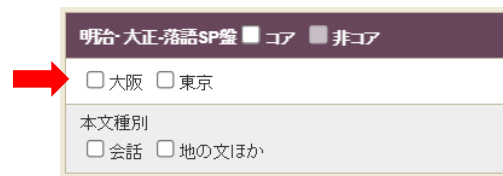
1-2桁目		3桁目		4-5桁目		6-9桁目	10桁目	11-12桁目		13-15桁目	
時代 通し番号		ジャンル		作品ID		録音年 (西暦)	区切り 記号	地域		通し番号	
60	明治	R	落語	nn	落語家略名	yyyy	-	01	東京 大阪	地域内通番	
70	大正							02			

4. 「中納言」における検索対象の選択方法と表示項目

4. 1 検索対象の選択方法

本コーパスには東京と大阪の 2 地域の落語の演目が収録されている。検索する際には「中納言」の「検索対象の選択」画面（図 2）において「東京」「大阪」のいずれかを選択することができる。

³ 「現代仮名遣い 本文 第 2（表記の慣習による特例）」（昭和 61 年 7 月 1 日内閣訓令、内閣告示）を参照。



明治・大正落語SP盤 ■ コア ■ 非コア

☐ 大阪 ☐ 東京

本文種別

☐ 会話 ☐ 地の文ほか

図2 検索対象の選択画面

4. 2 表示項目

本コーパスは、言語研究を目的とした利用に資する、さまざまな情報を有している。これらの情報は、コーパス検索アプリケーション「中納言」において参照可能な形で提供される。

表2として「中納言」の検索画面で参照できる付加情報のうち、初期設定で表示される主な項目を挙げ、「明治・大正時代編VI落語 SP 盤」で特に注意が必要な項目については説明を付す。空欄となっている項目については他のコーパスと共通する内容となるため、「中納言」検索インターフェイス⁴の解説を参照いただきたい。

表2 「中納言」検索画面で参照可能な付加情報

情報種別	項目名	内容
コーパス情報	サンプル ID	検索対象が含まれるサンプルの ID (→ 3 節)。
	多重化種別	本コーパスでは、一箇所につき複数の形態論情報を持たせる例はないため、常に空欄となる。
形態論情報	語形	検索対象の語形。語形はカタカナ表記され、語彙素で統合される。 例1：音声データに基づく語形「アスビ」 →語彙素「遊び」(名詞-普通名詞-一般)に統合。 例2：音声データに基づく語形「ナハレル」 →語彙素「為される」(動詞-非自立可能、上一段-ラ行)に統合。 例3：音声データに基づく語形「アジョ」 →語彙素「なんじょう」(副詞)に統合。
	品詞	特殊な品詞は以下のように表示される。 「言いよどみ」…非語彙的な発音。本来発音される短単位の一部の音などが、短単位とはならないレベルで発音されたもの。 例：「エー、いつでも私が、<u>お</u>、起きて開けてあげますからなんてエのは、」 (60R 小 さ 1911_01029,2090) 例：あなたの倅、佐兵衛の倅の佐太郎が、ただ今、<u>帰</u>…。 (60R 小 さ 1911_01029,6440) 「外国語」…英単語のローマ字表記部分。 例：英国の格言で、<u>Marry in haste, and repent at leisure</u> という言葉がございます。 (60R ブラ 1903_01076,70)

⁴ 「中納言」検索インターフェイスの URL は <https://clrd.ninjal.ac.jp/chj/chj-interface.html>

		「経文」…経文や願文の一部で、連続的に漢字の字音で発音していたもの。 例：<u>我昔所造</u>、<u>諸悪業皆由</u>、<u>無始貪瞋癡</u> (60R 枝雀 1909_02026,6420) 「解釈不明」…語の解釈が困難なため、解釈を保留した文字列。 「読取不可」…音声の聞き取りが困難で、それが二字以上にわたるような場合は、テキストとして具体的な表記が不可能であることを表す表示「<u>ㄥ</u>」をあて、その箇所を「読取不可」とした。 例：日本は少し、少しさかい、<u>ㄥ</u>そうかって言ったら嘘はない。 (60R 文三 1903_02018,2120) これらの特殊な品詞が付与された短単位は、形態論情報に関する項目のうち、「前文脈」「キー」「後文脈」「原文 KWIC」「品詞」「原文文字列」「振り仮名」以外は空欄となっている。
	活用形	検索対象活用語の活用形。文法的に特定の活用形が期待される箇所（単語同士の接続関係や文末等）で、それとは異なる形態が用いられている場合は、語の形態に即して活用形を割り当てる。 例：以後ご別懇に<u>願</u>い…、(五段-ワア行/連用形-一般) (70R 文雀 1924_02038_25570)
	原文文字列	検索対象と同じ文字列を示す（→2 節）。
	振り仮名	キーの文字列について、その読みが一般とは異なるような場合にひらがなで示す。
本文情報	本文種別	検索対象が落語中の登場人物によるセリフは「会話」と示す。それ以外は落語家の語り（地語り）で本文種別は空欄となる。
	話者	検索対象の本文種別が「会話」である場合、その話者名を示す。具体的な人名のほか、話中における立場（例：大家、住人）や役割（例：訪問者、接待者）、単に性別（例：男（甲）、男（乙））を示す場合もある。
	文体	検索対象を含む文の文体を示す。本コーパスでは常に「口語」と表示する。
作品情報	ジャンル	検索対象を含むサンプルのジャンル。すべて「落語」とする。
	作品名	検索対象を含むサンプル全体の演目名。
	成立年	検索対象を含む演目が収録された年。
	巻名等	空欄とする。
	部	検索対象を含む演目を語った落語家の所属地域を示す。本コーパスでは「東京」か「大阪」が表示される。
作者情報	作者	検索対象が含まれるサンプルの演目を語った落語家名。世襲の区別（代または世）を丸括弧に収めて表示する。落語家が「国立国会図書館典拠データ検索・提供サービス（Web NDL Authorities）」に収録さ

		れている場合は、そのウェブページへのリンクを付与する。
	生年	検索対象が含まれるサンプルの演目を語った落語家の生年。
底本情報	底本	検索対象が含まれる演目と SP レコードの音盤番号。
	ページ番号	空欄とする。
	出版社	検索対象を含むサンプルの演目を録音、発売したレーベル名。
その他	外部リンク	提供していない。

参考文献

岡田則夫編（2016）『落語研究資料解題 明治～平成 速記本・SP レコードデータ付き』（日外アソシエーツ）

金澤裕之・矢島正浩編（2019）『SP 盤落語レコードがひらく近代日本語研究』（笠間書院）

文化庁「国語政策」 https://www.bunka.go.jp/seisaku/kokugo_nihongo/kokugo_shisaku/index.html（2022 年 3 月 22 日確認）