『日本語シナリオコーパス』構築の試行とその調査例

長谷川　守寿

1．目的

本稿の目的は、『日本語シナリオコーパス』構築を念頭に、作成方針を仮設し入力を行う過程で、どのような問題が発生するかを明らかにし、その対処策を検討し、今後の規模拡大の際に役立てることにある。『日本語シナリオコーパス』を構築する目的は後述するが、これまで作られてこなかったシナリオを対象とした時、作成の過程でどのような一括処理に適さない形式が見られ、どのような対処が望ましいかを検討し、シナリオコーパスに適した形式を提案する。

さらに、構築中の『日本語シナリオコーパス』を対象とし、他の現存するコーパスと比較する調査を行い、その特徴を量的な観点から検討する。

2．先行研究

大規模コーパスに関しては、設計から構築まで詳細な報告がある。日本語に限定した場合、書き言葉コーパスの代表格とも言える『現代日本語話し言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese、以後BCCWJ）の設計や構築については多くの報告書が国立国語研究所のウェブサイトで公開されており、そのエッセンスに相当するものとして山崎編（2014）がある。同様に話し言葉コーパスの構築では、小磯編（2015）や小磯他（2023）などが挙げられる。日本語教育に関連したコーパスでは、『多言語母語の日本語学習者横断コーパス』の書き起こしとタグ付けの概要について迫田他（2016）に詳細な記述がある。

これらは多種多様な情報を含むデータをコーパス化していく際には必須と思われるが、単一のデータをコーパス化する際にも同様の仕様書を準備する必要があるか、疑問が残るところである。

また類似するコーパスとして、脚本コーパスが挙げられる。日本脚本アーカイブズ推進コンソーシアムにより、過去のテレビドラマやラジオの脚本を収集し、アーカイブする活動が行われている。そして、それらを元に脚本家個人の作品をまとめたものとして『鎌田敏夫脚本コーパス』『市川森一脚本コーパス』があり、これらを用いた研究に松下・丸山（2018）、松下（2019）、松下・丸山（2019）が挙げられる。しかし、これらのコーパスは公開されておらず、どのような方針の元、どのような形式で作られているのかは、明らかになっていない。

このように大規模なコーパスに関しては、構築に関する報告は多々ある。しかしシナリオや脚本などを対象とした小規模なコーパスの構築に関しては、詳細な報告は行われていない。本研究

で対象とするシナリオは著作権処理が済んでいないため、シナリオ自体を公開することはできず、結果として研究者の個人使用に限定された小規模なコーパスとなる。調査者が自らの研究のためにコーパスを作成する場合、最小限の負担で調査に使用できる形にするには、どのような問題が見られ、どのような対処が考えられるのか、構築を進めながら検討していくこととする。

3．方法

3．1．シナリオコーパスの必要性について

本研究の背景には、二つの出来事が関わっている。まずは『日本語日常会話コーパス』（Corpus of Everyday Japanese Conversation、以後CEJC）が、2021年度末に公開されたことである。CEJCは日本語における日常的な会話の自然なパターンを収集・分析するために設計されたデータセットである。このコーパスは、日常生活における多様な状況やトピックを網羅し、実際の会話を反映した音声とその書き起こしを含んでいるため、自然言語処理や音声認識の研究において貴重なリソースとなっている。また、日本語教育の素材と考えた場合、CEJCは文法の誤り・言いよどみ・言い間違い・フィラーなどを含み、正真正銘の話し言葉の日本語であると言える。

もう一つは、2021年６月に公表された「日本語教育の参照枠」（文化審議会国語分科会 2021）である（以後「参照枠」と略す。また出典のない引用は全て「参照枠」からである）。「参照枠」とは、「日本語の習得段階に応じて、求められる日本語の内容及び方法を明らかにし、外国人等が適切な日本語教育を継続的に受けられるようにするための、日本語教育に関わる全ての人が参照できる、日本語学習、教授、評価のための枠組み」（p.9）である。「参照枠」では、日本語能力の熟達度[(1)]について、全体的な尺度として六つのレベルで示し、聞くこと、読むこと、話すこと（やり取り）、話すこと（発表）、書くことの五つの言語活動別に、熟達度を示した上で、言語能力記述文を提示している。

そこで、「参照枠」の聞くことで記述されているような能力に引き上げる活動のために、語彙表の作成、表現文型の抽出、さらには授業で使用するタスクの作成を考えてみる。調査対象としてCEJCを選択した場合、多様な場面が含まれているとは言え、比較的平穏に会話が進んでいる場面のみで、多少ストレスを感じるような状況を含む、より多くの場面での会話は含まれていないという問題が考えられる。「参照枠」で示された言語能力記述文を達成するためのタスクの素材とし、「どんな種類の話し言葉も実質的に容易に理解できる」（p.24）レベルに達せられる教材を作成するには、CEJCでは含まれている場面が不足していると考える。

そこで筆者が取り上げるのはシナリオである。シナリオを取り上げる理由として、映画は従来から日本語教育の分野で、特に日本文化を学ばせる目的で使われてきたことにある。一つの映画を選んで、それを日本語教育学的に論じている論考・研究には、吉村（2010）や中山（2012）等いくつか見られる。

もう一つの理由は、「参照枠」で映画が取り上げられていることである。五つの言語活動の聞くことの中に記載されている言語能力記述文（pp.24-26）をレベルの下から挙げると以下のよ

うになり、C2になると「どんな種類の話し言葉も実質的に容易に理解できる」レベルとしている（下線は筆者、以下同）。

- B1 【テレビや映画を見ること】映像と人の行動が話の大筋を伝え、はっきりとした簡潔な言葉で話されていれば、かなりの映画が理解できる。
- B2 【テレビや映画を見ること】共通語による言葉遣いのドキュメンタリー、生のインタビュー、トークショー、演劇、大部分の映画を理解できる。
- C1 【テレビや映画を見ること】相当数の俗語や慣用表現のある映画が理解できる。
- C2 【包括的な聴解】熟達した日本語話者にかなり速いスピードで話されても、生であれ、放送であれ、どんな種類の話し言葉も実質的に容易に理解できる。

これらに対応するためには少なくとも、「はっきりとした簡潔な言葉で話されてい」る映画にはどのようなものが該当し、「相当数の俗語や慣用表現のある映画」にはどのようなものが挙げられるのか情報がなければ、現場の日本語教師は何を取り上げたらいいのか困ってしまう。さらに、映画に含まれる「相当数の俗語や慣用表現」を探るにはコーパスが必要となるが、管見の限り、シナリオとして入手可能だったのは日本語教育支援システム研究会（CASTEL/J）が以前に配布した『男はつらいよ』だけであり、客観的なデータを得るには、シナリオを集めたデータの必要性が挙げられる。

また映画の台本であるシナリオには、日常会話だけではない場面での会話も含んでおり、教材の元とするには、適当だと考えられる。これらを元に、調査・教材化することで、記述文を考慮した目標やタスクを設計し、教育に使用することができるのではないかと思われる。

このように聞くことに対応したタスクを作成するためにも、シナリオを収集した広義でのコーパスが必要となり、本稿では作成していく過程で必要となる仕様を考えていく。

3．2．方針

本研究で作成するコーパス名は、『日本語シナリオコーパス』とする（本論文中ではシナリオコーパスと呼ぶ）。作成の大方針として「最小限の負担で構築する」ことを目指し、手段としては「なるべくそのまま入力する」こととする。そして、ファイル形式は筆者にとって一番簡単なテキストファイルで、エンコードの種類はShift-JISとし、一つの作品は一つのテキストファイルにまとめることとする。

入力にShift-JISを採用する理由は、研究開始時に手元にあった数冊の『年鑑代表シナリオ集』の表記の形式を見る限り、山口（2014）で述べられている表記ほど複雑ではないと思われるため、Shift-JISで十分であると考えられたからである。

次に入力の範囲である。図１の線で囲んだ部分（『顔』、00年）[(2)]が入力しない部分であり、それ以外は全て入力することとする。具体的にはヘッダーにあるタイトル、フッターにあるページ番号、脚本家の顔写真以外は入力する。よって現在のところ、脚本家の略歴、スタッフ、キャス

トも入力することとする。

またシナリオ内では、図２（『独立少年合唱団』、00年）のように線で囲まれた掲示物を示す部分が見られ、この場合は線の中の文字だけを入力する。またアラビア数字・アルファベットは、２バイト文字で入力する。

次に「そのまま入力」しない部分について述べる。そのまま入力しない部分については、今後行われる調査を考え、修正しておいた方がよいものは適宜修正していく方針とする。例えば、誤字や脱字は語彙調査の際に支障となるので、前後から判断して正しい文字を入力する。他にそのまま入力しない例は形式に揺れがある場合である。調査の際にはプログラミング言語を使用して、正規表現を用いた処理で該当する箇所を抽出する。そのためなるべく一括処理が可能なように、形式等に揺れがある場合は統一を行う。

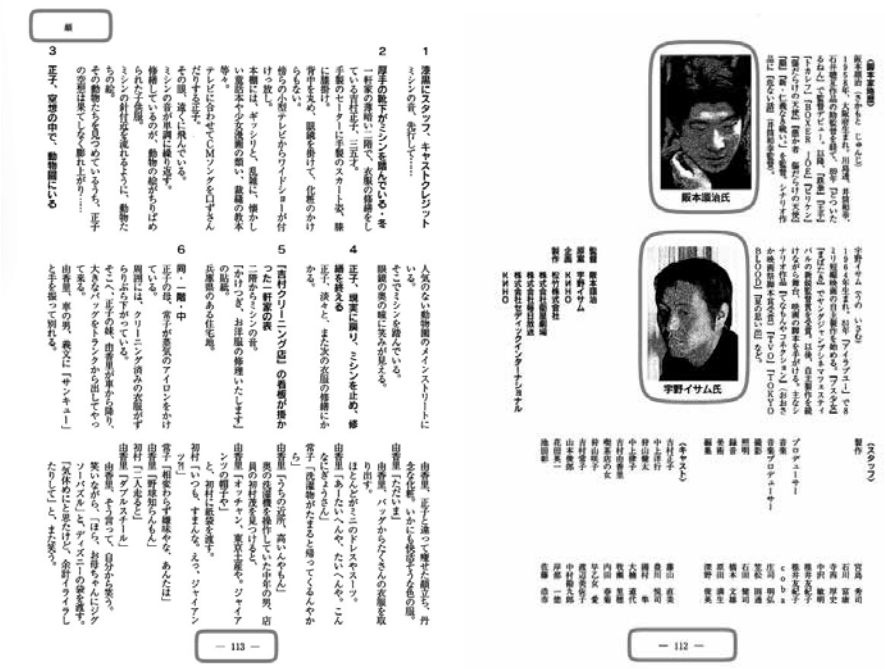

図１．入力しない箇所の例

次にタグについてである。出現が予想される誤字脱字の場合でも、シナリオコーパス作成の目的が本文研究ではないため、タグで記録を残すことはしない。また作成開始時の『年鑑代表シナリオ集』には、ルビは見当たらないため、ルビを表すタグは不要と考える。ただし、疑問が出た場合に参照できるように、元のファイルはpdfファイルで残すこととする。

このような方針を設定した場合、どのような問題が生ずるか観察し、対応策を検討する。

３．３．対象とするデータについて

『年鑑代表シナリオ集』は、出版委員会による選考を受けたもので、毎年刊行されており、質的にも保証されたものであると考える。毎年９本から10本の作品が掲載されており、今回は掲載されている全てのシナリオを対象とする。筆者が順次購入し作業しているため、本稿の対象は22年分のデータ（2000年、2001年、2004年、2006年～2024年）、201作品である[(3)]。

小木曽・山崎（2024）によれば、平成30（2018）年に行われた著作権法改正の「柔軟な権利制限規定」によって、コーパス構築に関わる権利制限規定が設けられ（文化庁著作権課 2019）、「収集した資料の機械可読テキスト化等は原著者の許諾なく行うことが可能」（小木曽・山崎 2024、p.85）となり、さらに「構築したコーパスの利用・分析等も許諾なく行えるようになった」（同上）とのことであり、本研究も適法だと考えられる。

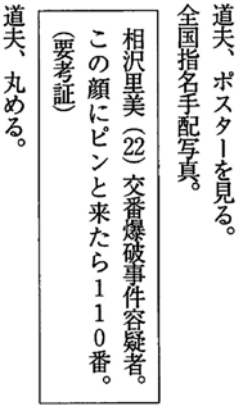

図２．線で囲まれた例

３．４．構築手順

紙面をpdfファイル化し、OCRソフト（「読み取り革命Ver.15」）で文字化する。その後、OCRソフト上でのOCR結果と画像ファイル（pdfファイルを変換したもの）の照合、印刷した結果と原本の照合の都合２回行う（なお、結果として文字化の誤りは残る可能性がある）。

４．結果

結果に入る前に、シナリオの各部分の呼び方について述べる。シナリオにおいて、シーンの場所・時間・天候などを示す部分（（１）の１行目）を「柱」、登場人物の動作や情景を描写する部分（（１）の２行目）を「ト書き」、登場人物が話す言葉を示す部分（（１）の３行目）を「セリフ」と呼ぶ。また、シナリオの作者を本稿では脚本家と呼ぶこととする。なお、引用の際、文字列の場合は「文字列」、記号の場合は“記号”と表記することとする。

（１）　１２　『吉村クリーニング店』・表・夕刻
　　由香里が初村に「オッチャン、またなー」と声を掛けて出てくる。
　　常子「正子に謝らんで行くんか。あんな、嘘言うて」　（『顔』、00年）

以後、全体で観察された、方針に沿わない表記の問題（４．１、４．２）を挙げ、次にそれぞれの部分で観察された問題（４．３、４．４）を挙げ、最後に全体の結果（４．５、４．６）を述べる。

4．1．文字コード

方針の通りShift-JISで入力を行ったが、入力できない文字がいくつか見られた。まず図３のハングルであるが、『福田村事件』（23年）で１箇所見られた。OCRソフトでも文字化はできず、セリフ内の日本語訳を入力している。

次に（２）の「尾骶骨」の「骶」である。これはShift-JISでは入力できないため、現在「尾てい骨」と漢字と平仮名の交ぜ書きで書かれている。語彙調査を行う際、このままでもWeb茶まめ（chamame.ninjal.ac.jpで現代語辞書使用）で正しく形態素解析できるため、本研究の目的が表記の研究ではないことを鑑みれば、このままで進めて問題ないと思われる。

（２）　大石「だから今でも、尾骶骨というのが残っているんです。～（中略）」　（『窓ぎわのトットちゃん』、23年）

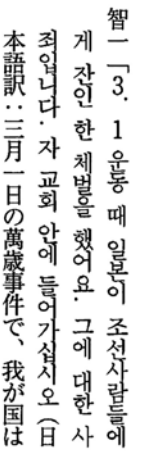
智一「3．1 운동 때 일본이 조선사람들에게 잔인한 체벌을 했어요. 그에 대한 사죄입니다. 자 교회 안에 들어가십시오（日本語訳：三月一日の萬歳事件で、我が国は

図３．ハングルの例

なお、映画化の情報も入力しているが、（３）のようにキリル文字が使われた会社名は、Shift-JISでは入力できなかった。映画情報のデータベースを作成しているわけではないため、どこまでデータ化するかも含めて、さらに検討が必要になる。

また、ルビの出現は当初予想していなかった。しかし非常に少数ではあるが、（４）のような例が見られたため、ルビが振られた漢字は、青空文庫に倣い、「罪人《つみびと》」のような形式で入力することとする。その場合、語に区切り、計量する際には、漢字とルビを統合する処理が求められるが、正確な処理を行うにはタグが必須となる。

（３）　企画：ＫＩＨＯ　（『顔』、00年）
（４）　所長「やめろよ、みんな罪人（つみびと）なんだから」　（『独立少年合唱団』、00年）

4．2．誤字・脱字

誤字と考えられるものには、元々のテキストを入力する際に仮名漢字変換を誤ったもの（５）（６）と、OCRの認識誤りがそのまま残ってしまったと考えられるもの（７）（８）が観察された。これらは正しいと思われる文字で入力する。

（５）　【誤】喉の弾機から顎関接の、首回りの柔軟等の発声方法。
　　【正】喉の暖機から顎関接の、首回りの柔軟等の発声方法。　（『独立少年合唱団』、00年）
（６）　【誤】やむをえず公務失効妨害で手配した
　　【正】やむをえず公務執行妨害で手配した　（『太陽の傷』、06年）
（７）　【誤】隣の居開にはベビー用品が置かれている。
　　【正】隣の居間にはベビー用品が置かれている。　（『宮本から君へ』、19年）

（8）【誤】三田ハツエの姿と長女の娘は、ハツエの遺影を～
【正】三田ハツエの次女と長女の娘は、ハツエの遺影を～　（『理由』、04年）

また、脱字と思われるもの（9）、誤入力ではあるが上記の分類に合わないもの(10)等も見られた。これらも修正した形で入力する。

（9）【誤】微かに笑を浮かべる。
【正】微かに笑みを浮かべる。　（『ほかげ』、23年）
(10)【誤】女「坊や、（裁縫箱を取り出し）こっち来なさない」
【正】女「坊や、（裁縫箱を取り出し）こっち来なさいな」　（『ほかげ』、23年）

4．3．ト書きの問題

ト書きの部分は、ほとんどの場合、（9）(11)のように行が句読点で終わる形式になっているが、(12) のように句読点がない形式のト書きも見られた（なお、ト書きは行頭２文字が全角空白文字という形式であるため、ト書きと分かるように本節以降はそのままの形で引用する）。

また、ト書きに書かれる内容も多様で、(13)のようにト書きの中にセリフが書かれているものや、画面に映される字幕を記したもの(14)、張り紙の内容(15)、効果音（Sound Effect、演出のために加えられる音）(16)、歌詞の一部(17)など、演出に関する様々な内容がト書きの部分に書かれている。今後ト書きだけを抽出して調査を行う場合も考えられるが、例えば語数調査ではセリフを外して処理するなど、目的にあわせて適宜処理するのがよいと思われる。

(11)　矢尾板、ナイフの刃で机をコツコツ叩きながら、　（『スリ』、00年）
(12)　翁、解せぬままに壁から笊を取り、それを持ってついて行きながら――
（『かぐや姫の物語』、13年）
(13)　蔦枝と里子、振り向いて、「おはようございます」。　（『この国の空』、15年）
(14)　T「６ヶ月後」　（『Ｗｉｎｎｙ』、23年）
(15)　岩田たち、井戸を封鎖し、張り紙を貼る。
――『飲ムナ　毒ニチュウイ』　（『福田村事件』、23年）
(16)　（ＳＥ）りぃぃん……　（『鬼太郎誕生　ゲゲゲの秘密』、23年）
(17)　まわれ　めぐれ　めぐれよ　はるかなとき　（『かぐや姫の物語』、13年）

4．4．セリフの問題

セリフの部分には様々な問題が見られたので、記号の問題（4．4．1）、形式の問題（4．4．2）、内容の問題（4．4．3）、使用言語の問題（4．4．4）に分けて記述する。

4．4．1．記号の問題

セリフの中には、図4（右が『顔』(00年)、左が『ナビィの恋』(00年）より）に示すように、1文字（2バイト文字）としての“‼”や“⁉”が出現する。これを1文字として見た場合、入力するには、“‼”（UnicodeでU+203C)、“⁉”（UnicodeでU+2049）のように、UTF-8での入力が必要になる。しかしこれらは語ではないため、語数の比較の際には影響しない。よって語数を対象とした調査ならば、Shift-JISで2文字として入力して問題ないと思われる。

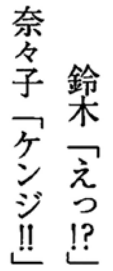

図4．Shift-JISで入力できない文字の例

しかし、それ以外の記号の例として、(18)のように“♡”が使用されるシナリオがあった（現在「＜ハートマーク＞」と入力している)。“‼”や“⁉”も含め、シナリオから話者の感情を読み取るなどのタスクを考えた場合、無視できない記号であるため、今後の出現の状況によっては文字コードの検討も必要となる。

(18)　麗子「気はやさしくて、力持ちよ〜、ねぇ、ダーリン♡♡」　（『ナビィの恋』、00年）

4．4．2．形式の問題

セリフは、多くの場合「人名「話す言葉」」のような形式で書かれているが、非常に希に句点“。”を最後に持つ形式(19)や、セリフの最後に「J」のようにそのセリフが話される言語を示すアルファベットが添えられている場合(20)がある。「J」は日本語を表すため、入力がなくても問題ないが、その他の言語については他に様々な形式があるため後述する。

(19)　翔子「内緒で乗って来たからね、仕方無いんです」。　（『理由』、04年）
(20)　ラック「あなた日本帰るでしょ？　わたし仕事いく」J　（『バンコクナイツ』、17年）

(19)のような形式については、統一した表記を目指すことから“。”を削除した形で入力する。(20)については、言語名の表示のバリエーションの一つとしてそのまま入力し、セリフ抽出などの一括処理の際には例外として扱う。

さらに、『理由』には(21)のように人名が“――”で表記されている部分がある。前後の文脈があれば理解できるが、検索して検討する場合などは、誰のセリフか分からないことになる。そこで、この場合、“――”を人名に置き換えて入力を行うことが必要となる。

(21)　　すると、思いのほか鋭い声で、翔子が言った。
――「お巡りさん、行きなよ（とハンカチを返す)」。
石川「あん？」。　（『理由』、04年）

また、(21)の「(とハンカチを返す)」や(22)のように、セリフの中にト書きに相当する部分が

入っているものがある。ト書きに相当する部分“（）”を削除し、括弧“「」”内のセリフだけを抜き出す一括処理が可能なため、このまま入力することとする。

(22)　恭一「ハナちゃん（と微笑みかけるが）、姪っ子かぁ。全っ然実感湧かない」
（『ある男』、22年）

4．4．3．内容の問題

一つの内容に対し多様な形式が用いられるのとは逆に、いくつかの異なる内容を一つの形式で記述しているものがあった。例えば、セリフには(21)のように会話を示すものがほとんどだが、(23)(24)(25)のように、モノローグやナレーションを示す「M・N・声」の表記と共に、独話がセリフ同様に記されている。独話は会話中の一部である(26)(27)とは使用されている文型・語彙が異なるため、調査に使用することを念頭に置いた場合、別に記述する必要がある。

(23)　内海（M）「塾へ行くまでの１時間30分の暇つぶしが始まる」　（『セトウツミ』、16年）
(24)　N　「この不思議な赤ん坊は、その日から、竹取りの翁と媼の手で大切に育てられることになりました」　（『かぐや姫の物語』、13年）
(25)　一雄の声「一九八〇年……山口百恵が引退して、松田聖子がデビューした……」
（『お父さんのバックドロップ』、04年）
(26)　内海「お前はええなぁ、大学行けへんねやろ？」　（『セトウツミ』、16年）
(27)　一雄「僕のお父さんがプロレスラーなの、ぜったい秘密だからね……」
（『お父さんのバックドロップ』、04年）

そこで、シナリオコーパスでは「人名『』」の形式で、独話を表すこととする。これは、入力の際に本文を読んで判断する必要がある。例えば、モノローグやナレーションだけではなく、手紙を読んでいる場面(28)や、歌を歌っている場面(29)、留守番電話の再生音(30)なども、“『』”を用いることとする。ただし、その場面では声だけであるが、次の場面ではその声の人と他の人との会話が始まるようなセリフも見られ、会話か独話かの判定は実際の前後から判断する必要がある。

(28)　みゆき『（読んで）バンド・コンテストご応募ありがとうございました。残念ながらお送りいただいたデモテープは、第二次審査にて落選となり……』　（『東京ゴミ女』、00年）
(29)　三上『♪モンキーマージック！　モンキーマージック！　モンキーマージック！　モンキーマージック！』　（『すばらしき世界』、21年）
(30)　携帯電話の声『メッセージを消しました。次のメッセージ……』
（『雪に願うこと』、06年）

4．4．4．使用言語の問題

シナリオには、吃音、方言、手話、外国語など、様々な言語とその形式が含まれる。(31)は吃音の例である。これらに対し現在のところタグは用いていないが、正確な語彙調査を行う場合は手作業での修正が望ましいので、どのようなタグを用いるか検討が必要である。

(31)　道夫「(深呼吸して）や柳田みみちおです」　（『独立少年合唱団』、00年）

また、方言がシナリオの一部に含まれている例は多々あり、含まれる方言は、津軽方言(32)、京都方言(33)、広島方言(34)など様々である。これらについてもそのまま入力しているが、「参照枠」に関連するので「６．考察」で再度言及する。

(32)　ハツヱ「しばらく弾いでねえはんで、糸道も消えでまってろ。なして弾がねえのさ」
（『いとみち』、21年）
(33)　おばさん「そりゃもうえらい酔い方どっせ、トシコバーのトシコはんが、連れてきてくれはったんどす。どないしやはりましてん。ナニがあったんどすか」　（『三文役者』、00年）
(34)　五十子「ちいと野暮用があってこっちに来たんじゃがのう、呉原で一番の美人がおる店に連れてけえて言うたらこの店じゃったんよ。(以下略)」　（『孤狼の血』、18年）

さらにシナリオの中には、(35)のように、セリフの一部に日本語以外の語が使われる場合もあるが、セリフ全体が日本語以外で話されるものもある。例えば、“（ ）”内で実際に用いられる言語を示し、その後に意味に相当する日本語を示している形式(36)や、セリフでは実際に用いられる言葉を示し“()”内でその該当する日本語を示す形式(37)、“「」”の後ろに実際に用いられる言語の略号（この場合、Tはタイ語）を示す形式(38)がある。さらに“「」”内には実際のセリフだけで意味の記述がない（この場合他の人のセリフで説明がある）形式(39)や、ト書きの中に言語を示す形式(40)など、様々な形式が見られた。

(35)　リエ「オッパ！　痛みはない？　大丈夫？」　（『かぞくのくに』、12年）
(36)　教師「(朝鮮語）おい！　どこ行くんだ！」　（『ＧＯ』、01年）
(37)　英姫「アイゴ、イノム（この野郎）！」　（『血と骨』、04年）
(38)　運転手「シーロムのどこ？」T　（『バンコクナイツ』、17年）
(39)　フェン「チャン・チュー・フェン・カ。クゥン・ラ・カ？」　（『サウダーヂ』、11年）
(40)　　しかし杉原は軽く身をかわす。(以下すべて朝鮮語)　（『ＧＯ』、01年）

現在はそのまま入力しているが、映画の場合、字幕などで出ていれば理解できるであろうし、字幕も含め、より正確に映画で使われる日本語を抽出する場合は、日本語でその発話の意味を記した部分だけにするなど、変更が必要となる。例えば、今後は字幕として拾われるであろう部分

が分かる(36)のように統一して、その形式とは異なる(37)については「(朝鮮語) この野郎！」のような形式が望ましいと思われる。ただし、このような対応をとった場合でも、(39)については対応できず、これら日本語以外の会話はどうするのか、さらに検討が必要となる。

4．5．構築結果

現在までの構築結果を挙げる。文字数は秀丸エディタver9.17を用い、語数はWeb茶まめ（使用辞書は「現代語話し言葉」、解析前の数値処理なし）を用いた。

セリフ・モノローグの部分については、人名、括弧“「」”、セリフ内の“（）”とその中に入っている文字列を除いた部分を対象とする。例えば、(41)の場合は、「わかったよ行ってきます」の部分のみ、語数・文字数を調べる。

その結果、全体で約436万語（記号を含む）・約660万字であり、会話に関連するセリフの部分は、約168万語（記号を含む）・約262万字であった。モノローグ部分は、約３万３千語（記号を含む）・約５万１千字であった。概数にしたのは、前述の通り、シナリオのセリフには様々な方言が含まれており、また他の言語で話されるセリフもそのまま対象としているため、正しく形態素解析が出来ていない例が多く見られるからである。例えば、前出の(32)の「弾がねえのさ」は、「弾（ヒ）がねえのさ」（丸括弧内は語彙素読み）となるべきところが、「弾（タマ）がねえのさ」という結果になる。

(41)　英雄「(釈然としないが) わかったよ（と受け取り）行ってきます」
（『あいときぼうのまち』、14年）

4．6．全般的な問題

調査開始時の『年鑑代表シナリオ集』には含まれていなかったため、掲載されている全てのシナリオを入力したが、ピンク映画のシナリオやそれに類する場面を含むシナリオが見られ、当然であるが語彙調査や教材化の際には注意が必要となる。

また、シナリオの全てがそのまま映像化されているわけではない。(42)のように決定稿というだけであって、撮影時に使用されたシナリオとは異なる場合があり、映像と合わせて利用する際には注意が必要である。

(42)　掲載シナリオは「決定稿」なので、完成した映画とは若干異なります。
（『毎日母さん』、11年）

5．量的調査の例

現段階のデータを用いて、特定の語と記号の使用について他のコーパスとの比較も含めた調査を行い、傾向を探り、その結果から作成中であるシナリオコーパスの特徴について考察する。

5．1．出現形の偏りについて

OCR結果の確認の際に、「ホント」には「ポンド」「ホンド」など誤認識が多く見られ、またセリフに「ホント」という表記が多いことに気付いた。そこで副詞「本当」を取り上げる。方法としてセリフの部分も実際の発話ではなく書かれたものであるという点から、書き言葉としての表記を見るためにBCCWJとの比較を行い、また実際に話される（た）という共通点があることから、CEJCと発音形出現形の比較を行う。

検索方法であるが、BCCWJは全てのジャンルを対象に「語彙素」で検索し、「書字形出現形」別に集計した。シナリオコーパスは、秀丸エディタで正規表現を使用し文字列検索を行った。出現頻度が低い「ホンット」「本っ当」、シナリオコーパスに出現しない「ほんたう」「ほんっっと」「ほーんと」を「その他」にまとめて集計したのが表1である（「シナリオ」はシナリオコーパスを示す。表2も同様）。CEJCも「語彙素」で検索し、「発音形出現形」別に集計した。その結果が表2である。

表1において出現数の多い順上位5位までの出現形とその頻度についてカイ自乗検定を行った結果、「本当」の出現形の頻度に関してBCCWJとシナリオコーパスの間では有意な偏りが見られた（$\chi^2(4)=882.259$、$p<.01$）。

表2から、実際に話された「ホントー」（書字形出現形では「本当」）と、「ホント」（書字形出現形では「ほんと」）は、CEJCとシナリオコーパスの間では出現傾向が逆であることが分かり、カイ自乗検定を行うと出現に有意な偏りが見られる（$\chi^2(1)=1822.157$、$p<.01$）。

表1．語彙素「本当」の書字形出現形の頻度（BCCWJでの頻度の多い順）

	BCCWJ	シナリオ
本当	24918	1141
ほんと	4048	561
ほんとう	3556	50
ホント	2353	318
ほんっと	57	28
それ以外	142	12
合計	35074	2110

表2．語彙素「本当」の発音形出現形の頻度（CEJCでの頻度の多い順）

	CEJC	シナリオ
ホントー	4705	879
ホント	545	1191
ホンット	0	32
ホンットー	0	8
合計	5250	2110

この結果から、シナリオコーパスは書き言葉としての表記とも、話し言葉としての発音とも分布が異なる独自のものとして位置付けられるであろう。なお、シナリオには同一作品中の一つのセリフの中にも(43)のように、表記の揺れが見られ、後続する形式による脚本家の使い分けが存在する可能性があり、さらに詳細な調査が必要となる。

(43) 女「え、これ。働いたの？　盗んだんじゃなくて…？　本当？…すごいじゃん。ほんとうにちゃんとした仕事なんだね」（『ほかげ』、23年）

5．2．「……」の頻出

シナリオコーパスの特徴として(44)の石川・翔子のセリフに見られるような「……」形式の頻出が挙げられる。

(44) 信子「週刊誌で見た人が、……うちにいるの。新聞にも載ってた人だよ」
　　下顎をがくがくさせながら、信子はそう言い、ゆるゆると顔を上げた。
　石川「……」
　翔子「……」
　　声も無く、信子を見詰める、二人。（『理由』、04年）

無言を表す「……」をBCCWJの全ジャンルを対象に文字列検索した場合、15件[(4)]だけであった。14件はYahoo!ブログで、(45)のようにオリジナルの小説を書いたものかと思われ、1件はYahoo!知恵袋で、曲名の一部である(46)。なお、発話のみを採取している『日本語話し言葉コーパス』やCEJCでは見られない例である。

(45) 「(略) 川村の心臓部付近に刺し傷が二つありました」「……」「おそらく刺創（しそう）の具合から見て、(略)」（OY13_00240, 32410）
(46) 「(略) 2．サウスポー（ピンク・レディー）3．渚の「……」（うしろゆびさされ組）4．白い色は恋人の色（ベッツィ＆クリス）(略)」（OC01_02636,2890）

そこでさらに無言を表す補助記号“…”について、コーパス全体に含まれる比率をBCCWJとシナリオコーパスで比較する。

まずBCCWJであるが、「『現代日本語話し言葉均衡コーパス』語彙表ver1.1解説」の短単位の語数（104,612,418）には、品詞に「空白」「補助記号」「記号」の文字列を含むものが除かれている。そのため、BCCWJ表記表（Version1.1）のBCCWJ_WritingFormTable.xlsxより「空白」「補助記号」「記号」の記号数（19,236,006）を加えて、全体の語数（123,848,424）とし、“…”の出現数（143,261）は上述のBCCWJ表記表から求めた。その結果、記号も含む全体の語数に対する“…”の比率は約0.12%（143,261／123,848,424）であった。

これに対しシナリオコーパスでは、全体の語数（記号を含む）は延べ語数4,364,621語、“…”の頻度が86,556回であり、約1.98%（86,556/4,364,621）と16倍近い違いが見られた。

このようにシナリオには“…”が多いという特徴が見られ、どのような時に日本語母語話者が黙るのか、学習者にとって必要な情報となるであろうし、このような特徴も活かした会話の教材化も一つの試みと言える。

6．考察

前述の通りシナリオには、方言を含むものが多々見られた。「参照枠」では方言と共通語を区別して示しており、以下のように述べている。

> 「共通語とは、国内において異なる地域社会に属する人や未知の人などとの意思疎通に必要となる全国に共通する言葉であり、全国共通語とも呼ばれる。一方、日本語には地方の伝統文化や地域社会の豊かな人間関係を担う言葉としての方言がある。こうした日本語の多様性を尊重した上で、ここでは相手や場面に応じて共通語と方言を使い分けることも社会言語能力の一側面として必要な能力と捉えている」（p.21脚注）。

さらに「社会言語能力」として以下のレベルで言及しており（p.61）、「聞き慣れないなまり」でも大枠は理解でき「適切に応じ」「仲介することができる」ことを目指すタスクも必要となる。

- C1 （略）特に聞き慣れないなまりの場合、時々細部を確認する必要があるかもしれない。俗語や慣用句がかなり使われている映画の筋を追うことができる。（略）
- C2 （略）熟達した日本語話者が言語を使用する際の実質的に全ての社会言語的、および社会文化的な意味を十分に理解し、適切に応じることができる。社会文化的、及び社会言語的な違いを考慮しながら、日本語話者と自分自身の生活地域の言語の話者との間を、効果的に仲介することができる。

日本語教師個人が様々な方言を操ることは難しいが、シナリオコーパスに含まれる方言から、方言を含む映画を示し、学習者に方言に触れる機会を映画が提供し、大枠の理解を求めるようなタスクが可能にならないか、検討する意義はあると思う。また本稿では、「参照枠」の聞くことを取り上げたが、シナリオのセリフに含まれる慣用表現、口語体表現の抽出は、以下に挙げる話すこと（やりとり）のC2（p.23）にも有効であると思われる。

- 〈話すこと〉C2　慣用表現、口語体表現をよく知っていて、いかなる会話や議論でも努力しないで加わることができる。

これらの用途に応えるためにも、「参照枠」では方言と共通語を区別して示しているように、シナリオコーパスでも、それらを区別した正確な語数調査などを行う必要があり、方言の部分を特定する新たなタグを用いた表記が求められる。

7．まとめと今後の課題

本稿では、最小限の負担でシナリオを入力することを考え、全体で約436万語、シナリオ部分では、168万語のコーパスが作られた。またその過程で、調査を念頭に置いた場合、形式の多様性により、どのような問題が生じるか考察してきた。その結果、セリフ・ト書き・柱の抽出自体は正規表現で可能であるが、セリフに含まれるト書きや、ト書きに含まれるセリフなど、注意する点も明らかになった。さらに、ごく少数ではあるが、Shift-JISでは入力できない漢字や、シナリオを理解する上では重要となるような記号が見られ、文字コードを変える必要性やルビに対応するタグの必要性も見えてきた。最後に、副詞「本当」と記号“…”を対象に、他のコーパスとの比較を行い、『日本語シナリオコーパス』の特徴を検討した。

筆者は『日本語シナリオコーパス』を2000年以降の『年鑑代表シナリオ集』で構成することを計画しており、今後も追加・整備を続け、さらにルビやモノローグなど、本稿で挙げたタグが正しく入力されているか全体の再確認も行う予定である。

また調査の面では、語彙表を作成し、シナリオでよく使われている慣用句や口頭表現を明らかにし、B2に該当する共通語による言葉遣いが大部分を占める映画やC1に該当する「相当数の俗語や慣用表現」のある映画には、どのようなものが挙げられるかなど、示していきたい。

ただし、2000年以降の『年鑑代表シナリオ集』で充分なのかは検討の余地がある。例えば、慣用句などを収集した場合、出現状況によっては、さらなるデータの拡張が必要となる。その場合、二つの方向性が考えられる。一つは、90年代の『年鑑代表シナリオ集』、80年代の『年鑑代表シナリオ集』と時代を遡る方向であるが、入手が困難だという問題と、日本語教育での使用を考えるならば、なるべく現在の話し言葉に近いことが望ましいなど問題がある。もう一つは、『月刊シナリオ』という月刊誌に掲載されたシナリオが選考されて『年鑑代表シナリオ集』に掲載される経緯があるため、『月刊シナリオ』まで対象を広げることである。後者が有力だと思われるが、その際には対象が増えることにより本稿では見られなかった表記の多様性が存在する可能性もあるため、構築を続ける中で、適宜タグを決めていく必要も出てくる。

最後に今後の研究の可能性を挙げる。現在、データには性別や場面に関する情報が欠けている。どのような場面が含まれているかは必須の情報だと思われるが、シナリオ自体にはこれらの情報は付与されていない。そこで、生成AIによる情報付与の自動化が挙げられる。生成AIと人が行う性別や場面の判断はどのように異なるのか、その理由は何が考えられるのかなど、様々な可能性が広がると思われる。

注

（1）日本語の習熟度は、基礎段階（A1）から熟達段階（C2）までの六つのレベルであり、Aは基礎的段階の言語使用者のレベル、Bは自立した言語使用者のレベル、Cは熟達した言語使用者のレベルである。

（2）括弧内の『＊＊』は作品名、その後の年数は2000年以降何年の『年鑑代表シナリオ集』に収録されているかを示し、00年は2000年を示す。以後、用例の出典はこの形式とする。

（3）ただし『'03年鑑代表シナリオ集』にページを分けて別々に掲載されている『ソロモンの偽証　前篇・事件』（pp.61-101）と『ソロモンの偽証　後篇・裁判』）（pp.103-143）を一つの作品と考えた場合は、200作品である。

（4）なお、"…"が一つの「…」は1,843件であり、三つの「……」は２件であった。そのため入力の際に制限があった可能性もある。

参考文献

小木曽智信・山崎誠（2024）「現代語書き言葉コーパスと著作権処理―BCCWJ2の構築に向けて―」『日本語学会2024年度秋期大会予稿集』、日本語学会、pp.85-88.

小磯花絵編（2015）『講座日本語コーパス　3．話し言葉コーパス―設計と構築』、朝倉書店.

小磯花絵・天谷晴香・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・渡邊友香（2023）「『日本語日常会話コーパス』設計と特徴」『国立国語研究所論集』24、pp.153-168.

迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子（2016）「多言語母語の日本語学習者横断コーパス」『国語研プロジェクトレビュー』Vol.6No.3、pp.93-110.

中山英治（2012）「日本語教育における映画の一般的な教材価値と社会参画を支援できる教材価値―『男はつらいよ』を資料として―」『早稲田日本語教育実践研究』巻１、早稲田大学日本語教育研究センター、pp.119-137.

文化審議会国語分科会（2021）『日本語教育の参照枠　報告』https://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/hokoku/pdf/93736901_01.pdf（2025年12月26日最終閲覧）

文化庁著作権課（2019）「デジタル化・ネットワーク化の進展に対応した柔軟な権利制限規定に関する基本的な考え方」https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30_hokaisei/pdf/r1406693_17.pdf（2025年12月26日最終閲覧）

松下晶子・丸山岳彦（2018）「脚本テキストに基づくコーパス文体論の可能性―テレビドラマ脚本に注目して―」『言語資源活用ワークショップ2018発表論文集」、pp.257-266.

松下晶子（2019）「テレビドラマ脚本で用いられる終助詞について―年代差・性差・表記差の分析―」『コーパスと文体論のインタフェース2018発表論文集』、pp.1-14.

松下晶子・丸山岳彦（2019）「複数の脚本コーパスに現れた終助詞の比較分析」『言語資源活用ワークショップ2019発表論文集』、国立国語研究所、pp.19-29.

山口昌也（2014）「第３章　文書構造の電子化」『講座日本語コーパス　2．書き言葉コーパス―

設計と構築―』、朝倉書店、pp.45-67.
山崎誠編（2014）『講座日本語コーパス　2．書き言葉コーパス―設計と構築―』、朝倉書店.
吉村弓子（2010）「映画を用いた日本語教育」『北海道言語文化研究』No.8、北海道言語文化研究会、pp.3-12.

参照データ

「『現代日本語話し言葉均衡コーパス』語彙表ver1.1解説」https://doi.org/10.15084/00003282
BCCWJ_WritingFormTable.xlsx（https://clrd.ninjal.ac.jp/bccwj/bcc-chu.html、2025年12月26日最終閲覧）