# 『日本語話し言葉コーパス』の概観

Version. 1.0

前川喜久雄　（国立国語研究所）

## 目 次



## １．『日本語話し言葉コーパス』

『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese:以下CSJと略称する)は、現代日本語の自発音声を種々の研究用付加情報とともに大量に格納したデータベースであり、語数にして約 750 万語、時間にして 660 時間の音声が含まれている。CSJ は、1)多数の話者による多少とも自発的な音声を対象としていること、2)豊富な研究用付加情報を提供していること、3)発話スタイルないし自発性に対する評価を与えていること、4)XML 文書化されたデータも公開していることなど、従来の音声データベースにはない多くの特長を有している。

CSJ は国立国語研究所と通信総合研究所によって開発された。開発資金は、東京工業大学の古井貞熙教授を総括責任者とする文科省科学技術振興調製費開放的融合研究制度課題 話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」(1999-2003)に拠った。古井教授にプロジェクト全般にわたるご指導をいただいたほか、ATR 音声言語コミュニケーション研究所、NTT コミュニケーション科学基礎研究所、京都大学等の研究者にも、多岐にわたるご協力をいただいた。

以下、本文書では CSJ の概要をできるだけ簡潔に説明する。本文書はあくまで概説であり、CSJ の利用に必要なすべての情報を網羅しているわけではない。CSJ に含まれる多様な研究用情報やツール類については、表 1 に示す解説文書が用意されているので、これらの文書を参照していただく必要がある。これらの文書は本文書と同じ場所（Disk 1 の/DOC ディレクトリ）に格納されている。

表１ 『日本語話し言葉コーパス』の解説文書

| 文書名 | ファイル名 |
|---|---|
| 『日本語話し言葉コーパス』の概観 | overview.pdf （本文書） |
| 音声収録作業の概要 | recording.pdf |
| 記録票データ・対話記録票データ・講演者属性データ・対話参加講演者の講演一覧の解説 | data_attribute.pdf |
| 転記テキストの仕様 | transcription.pdf |
| 文節の仕様について | bunsetsu.pdf |
| 『日本語話し言葉コーパス』の形態論情報の概要 | pos.pdf |
| 短単位・長単位データマニュアル | wdb.pdf |
| 短単位辞書マニュアル | suwdic.pdf |
| 『日本語話し言葉コーパス』の分節音ラベリング | segment.pdf |
| 『日本語話し言葉コーパス』のイントネーションラベリング | intonation.pdf |
| 印象評定データの概要 | impression.pdf |
| 『日本語話し言葉コーパス』における節単位認定 | clause.pdf |
| 『日本語話し言葉コーパス』における係り受け構造付与 | dependency.pdf |
| 『日本語話し言葉コーパス』における自由要約・重要文抽出データについて | summary data.pdf |
| 『日本語話し言葉コーパス』の談話境界情報について | discourse.pdf |
| 音声認識のための音響モデルと言語モデルの仕様 | asr.pdf |
| 『日本語話し言葉コーパス』XML 文書について | xml.pdf |

## ２．CSJ に収録された音声

CSJ はその名が示すように自発音声を対象としたデータベースである。しかし、ひとことに自発音声と言ってもその内容はきわめて多様であり、自発音声の全体像を一気に把握するようなデータベースを構築することは実際上不可能と考えられる。そこで、自発音声データベースを設計するにあたっては、対象とする音声の性格をあらかじめある程度限定しておく必要がある。

CSJ の設計では、CSJ を自発音声の自動音声認識の研究リソースとして活用することを想定して、以下の性格をもった音声を主要な対象とした。

1) まとまった内容をもつ（すなわち雑談ではない）
2) 全国共通語の（すなわち語彙と文法に方言的特徴のない）
3) 独話（モノローグ）音声

表 2はCSJ に含まれる音声を分類して、以下に説明するタイプ毎に、話者数、ファイル数 講演数)、総時間数を示している。表 2から学会講演と模擬講演という2種類の独話音声が時間にして CSJ の約 90%を占めていることがわかる。これらの音声タイプについては 2.1 以下で説明をくわえる。

表 2 音声のタイプと時間

| 音声のタイプ | タイプ | 話者数 異なり) | ファイル数 | 時間 |
|---|---|---|---|---|
| 学会講演 | 独話 | 819 | 987 | 274.4 |
| 模擬講演 | 独話 | ** 594 | 1,715 | 329.9 |
| その他 | 独話 | *** 16 | 19 | 24.1 |
| 学会講演インタビュー | 対話 | * (10) | 10 | 2.1 |
| 模擬講演インタビュー | 対話 | * (16) | 16 | 3.4 |
| 課題指向対話 | 対話 | * (16) | 16 | 3.1 |
| 自由対話 | 対話 | * (16) | 16 | 3.6 |
| 再朗読 | 朗読 | * (16) | 16 | 5.5 |
| 朗読 | 朗読 | * (248) | 507 | 15.5 |
| 計 | | 1,417 | 3,302 | 661.6 |

* ( )内は全員が学会講演話者もしくは模擬講演話者としてカウントされている
** 10 名は学会講演話者としてもカウントされている
*** 2名は学会講演話者としてもカウントされている

表 3は CSJ に含まれる形態論的単位数 語数)を音声のタイプごとに示している。CSJ では短単位、長単位という2種類の形態論的単位を用いているので (pos.pdf参照)、それぞれの数字を示した。言い誤りによって生じた語の断片などは除外されている。CSJ 全体での短単位総数は 7,525,125 であり、設計上の目標値 700 万語、3節参照)を大幅に超過している。

表の最終列は短単位数に占める長単位数の百分率である。この率は、学会講演で最低値をとり、模擬講演と対話で最大値をとっている。これは学会講演には専門語が多く用いられるために相対的に多くの複合語 複合辞)が含まれていることによると考えられる。この例が示唆するように、CSJ に格納されている音声は--上記 1-3)の制約にはしたがいつつも--多様性に富んでいる。以下、各タイプについて簡単に説明する。

表 3 形態論的単位数

| 音声のタイプ | 短単位数 | 長単位数 | %長単位 |
|---|---|---|---|
| 学会講演 | 3,279,364 | 2,654,823 | 81.0 |
| 模擬講演 | 3,605,729 | 3,115,302 | 86.4 |
| その他 | 282,728 | 239,989 | 84.9 |
| 朗読と再朗読 | 207,478 | 172,216 | 83.0 |
| 対話 | 149,826 | 131,544 | 87.8 |
| 全体 | 7,525,125 | 6,313,874 | 83.9 |

### ２．１　学会講演と模擬講演

学会講演(Academic Presentation Speech: APS)は、理工学、人文、社会の 3領域におよぶ種々の学会における研究発表のライブ録音である。講演時間は 10分から25分程度が大半であるが、1時間前後に及ぶ特別講演の類も少数含まれている。学会講演の多くをしめる理工学系の学会では、男性の大学院生であることが多いので、学会講演の話者は、年齢と性別の偏りがある。発話スタイルは概してあらたまり度が高い (4節参

照)。

模擬講演(Simulated Public Speaking: SPS)は、できるだけ年齢と性別のバランスをとった一般話者による、日常的話題についての講演である。話者の大部分は人材派遣会社からの派遣であり、あらかじめ指定された3種の一般的テーマ(例えば「人生で一番嬉しかったこと」「人生で一番悲しかったこと」「私の住んでいる街」等)に基づいて、具体的な講演内容を決めてタイトルをつけ、1講演10〜15分程度のスピーチをおこなった。発話スタイルは概して学会講演よりもくだけたものとなっている(4節参照)。

## 2．2　その他

学会講演にも模擬講演にも該当しない独話音声は、「その他」に分類されている。具体的には、以下のものである。

- 研究機関が一般聴衆を対象に企画した連続講演会の講演音声。対象は歴史や民俗学など。
- 国語研究所が一般聴衆むけに開催した講演会の講演音声、および国語研究所員を聴衆とした識者による講演。後者は1講演のみである。
- 専門学校における日本語教師養成関係の講義音声

その他に分類した音声は、独話である点では学会講演や模擬講演と同一であるが、話者と聴き手の関係が専門家と一般聴衆の関係である点において、学会講演とも模擬講演とも異なっている。

## 2．3　朗読

CSJ の大部分を占める学会講演と模擬講演は自発的な独話(モノローグ)音声である。しかし、独話音声だけでは、自発音声の限定された一面しか検討することができない。また、自発音声の特徴を把握するためには、朗読された音声との比較も必要である。そこで、独話との対照用に、対話と朗読の音声も収録した。

朗読には、「朗読音声」と「再朗読音声」の二種類がある。「朗読音声」は模擬講演話者の一部が、書き言葉のテキストを朗読したものである。テキストとしては、野本陽代著「宇宙の果てにせまる」(岩波新書新赤版570,1988)および中村桂子著「あなたのなかのDNA」(ハヤカワノンフィクション文庫 176,1994),p.9-10)の一節を利用させていただいた。以下にその一部を示す。このうち「あなたのなかのDNA」は、会話体の書き言葉であり、その話し手としては女性が想定されているため、男性話者による朗読では、性差にかかわる終助詞を一箇所修正して朗読に供している(朗読テキストの詳細については recording.pdf 参照)。

私たちの住んでいる宇宙は、いったいどこまで広がっているのだろうか。いつ生まれて、いつまで存在するのだろうか。そこには、どんな天体があって、それらは将来どうなるのだろうか。

(「宇宙の果てにせまる」)

となりの家のかな子ちゃんは女子大生。好奇心にあふれた、元気な女の子です。彼女がある日、聞きました。

―「先生、DNAって何ですか？」

ん？

―「DNAって、ほら、遺伝子なんでしょ。」

(「あなたのなかのDNA」)

上記2種類の他、さらに3種類のテキストを用いた朗読音声を収録しておりCSJ にも格納しているが、これらを朗読した話者はそれぞれ4名のみである(recording.pdf 参照)。

## ２．４　再朗読

再朗読音声とは、学会講演ないし模擬講演として収録された音声の転記テキストを同一の話者が朗読した音声である。フィラーや言い直しも朗読の対象としている。再朗読の話者は、学会講演話者から選ばれた 10 名と、模擬講演話者から選ばれた 6名の合計 16名である。これらの話者は次に述べる対話音声の話者でもある。詳しくは 2.6 参照。

## ２．５　対話

対話音声には「学会講演インタビュー」「模擬講演インタビュー」「課題指向対話」「自由対話」の 4種類がある。2種類のインタビューは、上記 16名による学会講演ないし模擬講演（10 名は両方、6名は模擬講演のみ）に関してインタビュワーが様々な質問を発し、講演者がこれに答える形式の対話である。予想されるように、発話の大半は、質問に対する回答によって占められている。

インタビュワーは 20代と30代の女性各 1名である。インタビュワーは、学会講演インタビューに関しては事前に予稿集論文に目を通したうえで、また、インタビュー対象の模擬講演については、その講演の収録現場で講演を聴取したうえで、インタビューに臨んでいる。

課題指向対話では、インタビューとの対比のため、参加者 2名（上記インタビューと同一ペア）の発話量が等しくなりやすい課題を選定した。具体的には、実在の芸能人に講演を依頼した場合の謝礼（ギャラ）の額を想像し、その多寡の順に、芸能人 9 ないし10 名をソートするタスク（ギャラ・タスク）を考案した。対話開始時点で各話者に手渡されている人名リストは、わざと一致しないように作成してあるので、謝礼額の推定に先立って（あるいは同時に）、推定対象となる芸能人の完全なリストを作成するための対話も必要とされる。

最後に自由対話では、話題の制約なしに、10 分程度、自由に対話をおこなってもらっている。以上、4種類の対話音声は、同一の話者ペア（講演者とインタビュアー）によって発話されている。

## ２．６　再朗読および対話の収録についての補遺

先に触れたように、再朗読および対話の話者は、10名が学会講演の経験者から、6名が模擬講演の経験者から選ばれている。このうち学会講演経験者の大部分は CSJ の関係者かその知人である。

CSJ に格納された各音声タイプ間には自発性の程度差が存在していると考えられるが（4節参照）、これら16名の話者については、自発性が最も低いと考えられる再朗読音声から、自発性が最も高いと考えられる対話音声まで、自発性の幅広い領域にまたがる音声の比較が可能である。

話者 16 名の講演者 ID と音声ファイル名の対応表を本文書末に付録 1 として掲載する。この表中の模擬講演（模擬」）は、学会講演経験者のみならず、模擬講演経験者 6 名についても新規に収録したものであることに注意してほしい。これによって模擬講演のテーマは統一が保たれている。

## ２．７　話者の分布

自発音声の多様性の一部は、性別、出生地、居住歴、学歴、講演経験の有無など、話者の社会的属性に起因していると考えられる。そのため、自発音声の研究では話者の属性への配慮が欠かせない。CSJ では、話者のプライバシーを侵害しないと判断された範囲で話者の属性情報を公開している（属性情報の詳細については recording.pdf および data_attribute.pdf 参照）。ここでは、最も代表的な属性として、話者の生年代と性別と出生地の分布を概観する。

まず、図 1, 2に学会講演と模擬講演における話者の生年代の分布を示した。CSJ のデータでは、話者の生

年を西暦で5年刻みに区分して公開しているが、図1, 2ではこれを10年ごとに区分しなおして集計した。

図1は生年代ごとの延べ話者数、図2は同じく異なり話者数の分布を示している。延べと異なりの区別が必要となるのは、模擬講演だけでなく、学会講演においても同一話者の音声が複数回収録されていることがあるからである。これを重複してカウントしたのが延べ話者数、何回講演しても1名としてカウントしたのが異なり話者数である。

図1においても図2においても、学会講演話者数は生年代が下がるにつれ単調に増加している。一方、模擬講演話者は、学会講演に較べれば相対的にバランスのとれた分布を示している。なお、学会講演話者のうち9名については生年が不明であるために集計から除外している。

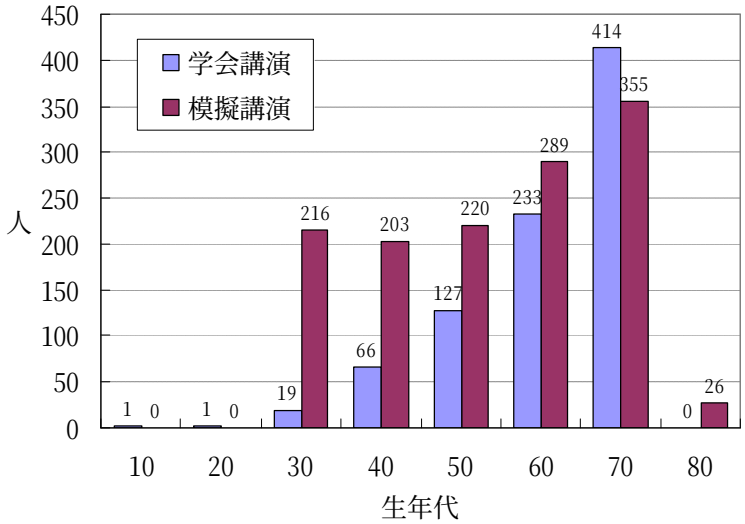



図1 学会講演と模擬講演話者の生年による分布（延べ）

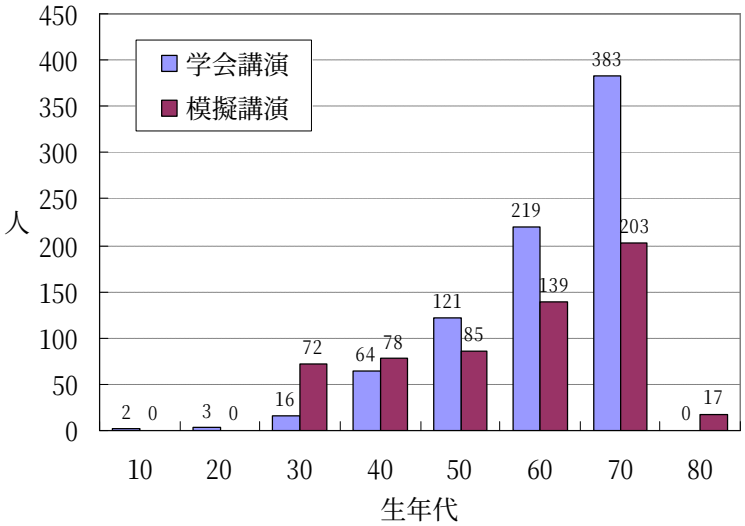



図2 学会講演と模擬講演話者の生年による分布（異なり）

次に、表4、5に話者の性別と音声タイプのクロス集計を示す。表4が延べ話者数、表5が異なり話者数である。表5では学会講演から対話までの合計が「全体」欄の数字と一致していない。これは同一話者が複数の音

声タイプの講演を行っている場合に重複してカウントしているためであり（同一タイプ内部での重複はカウントしていない）、再朗読と対話の話者を学会講演ないし模擬講演の話者から選択していることと「その他」の話者のうち2名が学会講演話者でもあることが、その原因である。

学会講演話者の大多数は男性である。これは学会発表の多くが大学院生によっておこなわれており、その大部分が男性であることによる。この傾向は特に理工系大学院において著しい（ちなみに図1, 2の学会講演において70年代生まれの話者数が突出しているのも大学院生の多さによる）。一方、模擬講演以下では、男女がほぼ均等に分布している。

表4 話者の性別の分布（延べ）

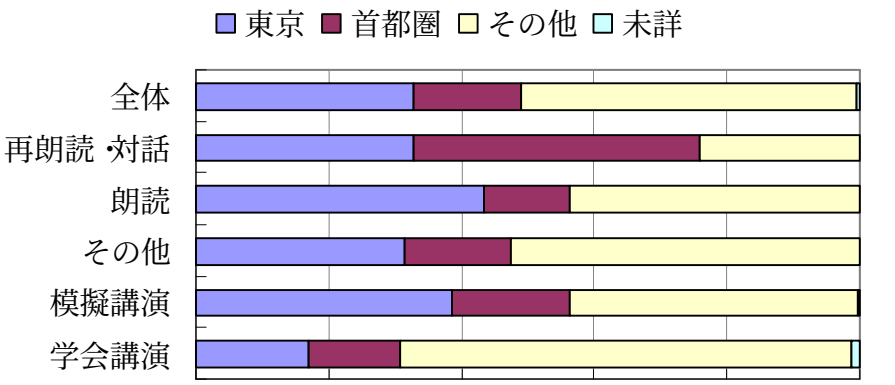
| 性別 | 学会講演 | 模擬講演 | その他 | 朗読 | 再朗読 | 対話 | 全体 |
|---|---|---|---|---|---|---|---|
| 女 | 173 | 910 | 9 | 252 | 8 | 29 | 1381 |
| 男 | 814 | 805 | 10 | 255 | 8 | 29 | 1921 |
| 計 | 987 | 1715 | 19 | 507 | 16 | 58 | 3302 |

表5 話者の性別の分布（異なり）

| 性別 | 学会講演 | 模擬講演 | その他 | 朗読 | 再朗読および対話 | 全体 |
|---|---|---|---|---|---|---|
| 女 | 138 | *331 | 6 | (122) | (8) | ****470 |
| 男 | 681 | **263 | ***10 | (124) | (8) | 947 |
| 計 | 819 | 594 | 16 | (246) | (16) | 1417 |

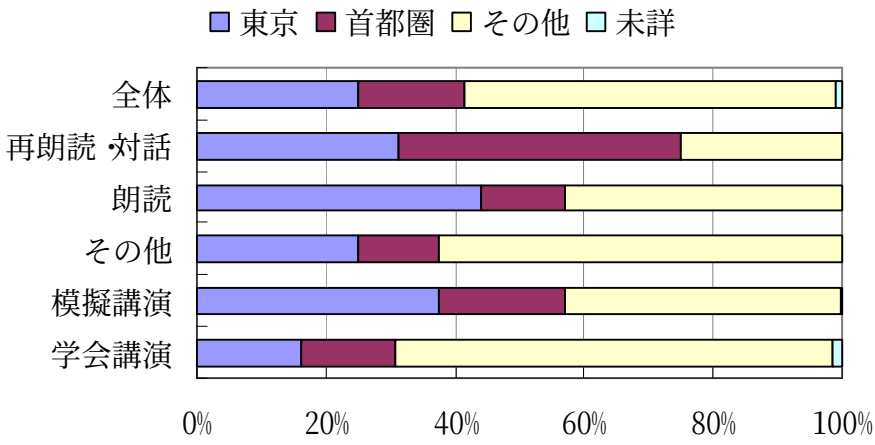
( )内の数字は学会講演もしくは模擬講演と重複、 * 5名が学会講演と重複、 ** 5名が学会講演と重複

*** 2名が学会講演と重複、****インタビュワーを加えると471名




| | 学会講演 | 模擬講演 | その他 | 朗読 | 再朗読・対話 | 全体 |
|---|---|---|---|---|---|---|
| 未詳 | 12 | 3 | 0 | 0 | 0 | 15 |
| その他 | 672 | 747 | 10 | 221 | 14 | 1664 |
| 首都圏 | 135 | 305 | 3 | 66 | 25 | 534 |
| 東京 | 168 | 660 | 6 | 220 | 19 | 1073 |

図3 話者の出生地の分布（延べ）




| | 学会講演 | 模擬講演 | その他 | 朗読 | 再朗読・対話 | 全体 |
|---|---|---|---|---|---|---|
| 未詳 | 12 | 1 | 0 | 0 | 0 | 13 |
| その他 | 555 | 255 | 10 | 106 | 4 | 816 |
| 首都圏 | 120 | 115 | 2 | 32 | 7 | 233 |
| 東京 | 132 | 223 | 4 | 108 | 5 | 355 |

図４ 話者の出生地の分布（異なり）

図３、４に話者の出生地による分布を示す。図３が延べ話者数、図４が異なり話者数である。いずれの図においても「東京」「首都圏」「それ以外」に分類し百分率で示している。「首都圏」とは千葉、埼玉、神奈川の３県をさす。なお、ここで、出生地とは文字通り話者が生まれた土地であって生育地ではない。社会言語学的な研究などにおいてさらに詳しい履歴が必要とされる場合は、CSJ の講演者属性データ(Disk1 /DATA の speaker_data.dat)に転居歴のデータが含まれているので、それを利用できる(data_attribute.pdf 参照)。

## ２．８　ファイルの命名

CSJ のファイル名(本体部分)には、以下の規則に則って音声タイプとその内訳、さらに話者の性別が反映されている（recording.pdf も参照）。

- ファイル名先頭のアルファベット1文字は音声のタイプを示す
  - A:学会講演
  - S:模擬講演
  - M:その他
  - D:対話音声
  - R:朗読（含再朗読）音声
- 続く2 桁の数字は、音声タイプ毎に、その内訳を示す
  - A では学会の別
  - S では講演テーマの別
  - D では対話の種類の別（インタビュー2 種、課題指向、自由）
  - R では朗読テキストの別
- 続くアルファベット1文字は、話者の性別を示す
  - F:女性
  - M:男性
- 残る4桁の数字は各音声タイプ内での識別番号

## ３．研究用付加情報（アノテーション）

CSJ には豊富な研究用情報（アノテーション）が付加されている。ただし、アノテーションは全体に対して斉一的に実施されていないことに注意が必要である。CSJ のうち、「コア」と呼ばれる約 50 万語分については、特に多くの情報を集中的に付与した。図 5はコアとそれ以外における情報付与の異同の概念図、表 6はコアにおけるファイルの分布を音声タイプと話者の性別に関する内訳である。

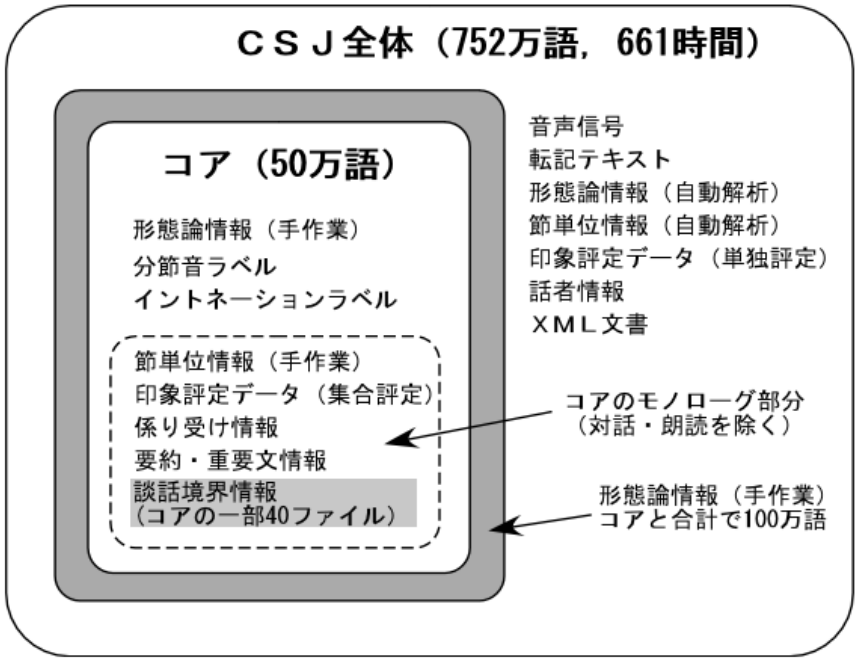



図５ CSJ のアノテーションにおける階層構造

表６ コアの内訳

| 音声のタイプ | 女性話者 | 男性話者 | 計 |
|---|---|---|---|
| 学会講演 | 24 | 46 | 70 |
| 模擬講演 | 54 | 53 | 107 |
| 対話 | 9 | 9 | 18 |
| 朗読 | 3 | 3 | 6 |
| 計 | 90 | 111 | 201 |

CSJ に、このような階層構造を導入した理由は以下の通りである。「話し言葉工学」プロジェクトの目標は、自然な話し言葉（自発音声）の音声言語処理技術のための基盤技術の開拓であった。工学領域における研究としては、1)自発音声の自動認識、2)自発音声認識結果の自動要約、3)話し言葉の自動形態素解析、等を想定しており、これらに加えて、4)自発音声の韻律特徴、5)自発音声における音声変異、6)自発音声の統語的、談話的構造と韻律特徴の関係、7)講演音声が聴き手に与える印象、等の言語学的ないし音声学的研究における利用も、念頭においた。

CSJ の設計にあたって問題となったのが、音声認識研究（上記の 1と2)とそれ以外の関係である。統計的学習に基礎をおく音声認識研究においてはデータの量が重視されるのに対して、自然言語解析や言語研究においては、量よりもむしろ多様かつ精密なアノテーション情報を提供することが大切と考えられた。そこで、

コアを設定し、そこにアノテーションを集中することにした。

まず、音声認識研究に最低限必要なデータ量を 700 万語（短単位）と推定し、これをデータベース全体のサイズの目標値とした。この700万語分（実際のCSJでは表3に示したように約752万語）に対しては、音声信号(16bit, 16kHz)の他に、精密な転記テキスト、形態論情報（品詞情報）、節単位情報を提供して、音声認識研究での利便性を確保した。一方、コアに対しては、上記の情報に加えて以下の付加情報を追加することによって、言語研究等における利便性を追求した。1)分節音ラベル、2)イントネーションラベル、3)節単位情報、4)係り受け情報、5)要約・重要文情報。さらに、コアに含まれる模擬講演の一部、40 ファイルに対しては、6)談話境界ラベルも付与した。これらの付加情報の仕様については、表 1 に示した解説文書に詳しい。

以上のほかにも、コアとコア以外の関係について指摘しておくべきことがある。

— コアのサイズは、5年間という研究実施期間において実現可能な上限として推定した。また、コアに含まれる音声は、東京ないし首都圏で出生した話者のものとした。これはイントネーションラベリング方式が東京方言のアクセント体系に依存しているからである。

— 形態論情報もコアに関連した異同がある。CSJ のうちコアを含む約 100 万短単位分の形態論情報は人手で実施されており、精度が高い（ランダムサンプリングによる推定では 99.9%）。一方、残る 650 万短単位分はコンピュータにより自動解析されたものを、時間が許す範囲で人手修正したものであり、その精度は 97～98%と推定される。また、人手による解析結果と自動解析結果とでは、用言の活用型と活用形の分類が一部相違しているので注意が必要である（wdb.pdf参照）。

— 節単位情報にも上と類似の相違がある。コアの節単位情報は自動解析結果を人手で修正しているが、コア以外に関しては自動解析結果をそのまま提供している。

— 節単位情報（手作業）、係り受け情報、要約・重要文情報は、コアに含まれる独話と下に説明する「テストセット」だけを対象としている。

— 印象評定（集合評定）データは、コアの独話（学会講演ないし模擬講演）だけを対象としている。

— 談話境界情報は、コアの独話のうち 40 講演（学会講演ないし模擬講演）にだけ付与されている。

— CSJ を音声認識研究で利用する過程で、認識システムの性能を客観的に評価するための「テストセット」30 ファイルを選定して利用した。このテストセットにはコア以外のファイルが 22 ファイル含まれている。この 22ファイルには、節単位情報（手作業）、係り受け情報、要約・重要文情報が付与されている（談話境界情報は付与されていない）。テストセットに含まれるファイルのリストを本文書末に付録 2として掲載した。

このように、CSJには大きくは、コア、コアを含む 100 万語、それ以外の三階層があり、またコア内部にも付加情報の濃淡がある。CSJに格納された3302個の音声ファイルのそれぞれにどのような研究用情報が与えられているかは、Disk1 の/DATA ディレクトリにある correspondence_data.dat というデータファイルに記載されている。

## ４．音声の自発性

CSJは「自発」音声データベースである。しかし、音声ないし言語の自発性という概念を明確化することは、実は簡単でない。自発性について用いられる説明のひとつに「発話の時点において、あらかじめ発話の形式が決定されていない発話」というものがある。CSJに収められた音声は、朗読および再朗読音声を除外すれば、総じてこの規定に適う。しかし、2779個におよぶ講演ないし対話音声を比較すると、そこには自発性の程度差が存在していることが明らかである（さらに、CSJの場合、朗読・再朗読音声にもかなりの数のフィラーや語断片が含まれている。これらいわゆる非流暢性が自発音声の特徴であるとすれば、朗読音声にもまた自発性の程度差が存在することになる）。

データベースに存在する自発性の差は、それを擾乱要因ととらえるにせよ、あるいは積極的に利用するにせよ、それを何らかの方法である程度客観的に評価できることが望ましい。CSJで採用したな方策を以下に説明する。

### ４．１　自然の序列

CSJに格納された種々の音声は、その種類によって、かなりの程度まで自発性の程度が組織的に異なっており、全般的な傾向としては、音声種別間で以下のような序列を想定してよいものと考えられる。このような自然の序列は、話し言葉の研究にとって積極的に利用すべき性質の情報である。

**自発性低** ◀――――――――――――――――――――▶ **自発性高**

再朗読ないし朗読　＜＜　学会講演　＜　模擬講演　＜＜　インタビュー　＜　課題指向ないし自由対話

ただし、このうち学会講演と模擬講演との間の差異については、他のカテゴリ間の差よりも小さい可能性があり、また、量的にCSJの大部分を占める音声でもあるために、両者間の差異をきわだたせる対策をとった。具体的には、模擬講演話者ができるだけリラックスした状態で録音に臨めるよう、収録に先立って収録スタッフと雑談を交わす時間を設ける、収録中も収録スタッフは積極的にうなずき等の反応をかえすといった対策である（学会講演話者には働きかけようがないので、何も対策を施していない）。いずれも素朴な対策であるが、一定の効果を発揮したことは、データの解析によって確認できる。

### ４．２　印象評定

自然の序列は、しかし、範疇的なものではない。特に学会講演と模擬講演の間では、前節に述べた収録上の対策にも関わらず、序列の逆転が生じていることが少なくないと思われる。そこで、個々の講演についても、その自発性をある程度客観的に評価する手段があるとよい。そのような評価の一助として、CSJのデータ収録作業では、原則としてすべての講演音声に対して音声収録記録票を作成し、その一部を音声が聴き手に及ぼす印象の主観評定に充てた。これを印象評定データと呼ぶ。印象評定項目の一部は自発性の程度差に関する外的指標として利用することができる。

印象評定には単独評定データと集合評定データの2種類がある。単独評定データは、ほぼすべての講演音声を、収録の現場において収録スタッフ中の1名が評価したデータである。自発性に関する項目としては、五段階で評定された「自発性」と「発話スタイル」がある（発話の自発性に関するもの以外にも種々の項目を評定している）。

ただし、単独評定データには、ひとつの講演に評定者が1名だけであること、データベース全体では多数

の評定者が参加していること、講演の前半を聴いた時点で評定を行っているため講演中のどの部分が特に印象形成に影響したかが明らかでない場合がありうること等の問題がある。これらの問題を解決するために作成したのが集合評定による印象評定データである。集合評定データは複数の評定者による評定結果であり、評定方法も実験心理学的により厳密な手順を踏んだものとなっている。単独評定データについて上で指摘した問題はほぼ解決されているが、対象となっているのはコアの独話だけである。印象評定データの詳細については data_attribute.pdf（単独評定データ）および impression.pdf（集合評定データ）参照。

### ４．３　音声収録記録票とアンケート

印象評定データの他にも、自発性の評価に利用できるデータがある。ひとつは、音声収録スタッフが音声収録記録票に記入したコメントである。スタッフは、話者が原稿を読み上げていることが明らかな場合には、その旨のコメントを記録票の備考欄に記している。もうひとつ、音声収録に先立って話者に実施したアンケートも自発性の評価に関連した項目を含んでいる。これらの情報は Disk1 の /DOC ディレクトリ中の talk_data.dat というデータファイルに記録されている(data_attribute.pdf 参照)。

## ５．XML 文書

CSJ に付与された多種多様な研究用情報を一括して表現するために、XML 文書化をおこなった。XML 化は元来、各種データの整合性をチェックするために実施したものであるが、XML 文書は、複雑な情報検索の用途にも向いているので『日本語話し言葉コーパス』の一部として公開することにした。XML 文書の仕様については xml.pdf 参照。

## ６．言語モデルと音響モデル

CSJ には京都大学と東京工業大学で開発された音声認識用音響モデルと言語モデルが Disk1 の /LM（言語モデル）、/AM（音響モデル）ディレクトリに同梱されている。これらについては asr.pdf 参照。

## ７．話者の個人情報

本文書に述べたように CSJ には 1417 名（インタビュワーを加えると 1418 名）にのぼる話者の音声が記録されている。自発音声の研究では話者の個人的属性に関する知識が必要となることが多いと考えられるので、Disk1 の/DATA ディレクトリにおかれた speaker_data.dat というデータファイルに話者の個人情報を記録して公開している(data_attribute.pdf 参照)。

一方、話者から音声の収録とデータ公開の許諾を得るにあたっては、話者の個人情報を不必要に開示しないことを条件としていた。これらの相反する制約を両立させるために、以下の方針に則ることとした。

1) 話者の氏名は開示しない。学会講演を収録した学会名も開示しない
2) 講演中の話者ないし関係者の個人情報に関係する言語情報（自己紹介等）は消去する
3) 話者の年齢は 5 歳刻みで開示する
4) 音声信号に含まれる個人性情報はそのまま開示する（音声は加工しない）

収録した音声中に話者やその関係者の身元を特定できる情報が含まれていた場合には、該当部分の転記テキストを伏せ字とし、あわせて伏せ字を含む転記基本単位全体に対応する音声信号をホワイトノイズで置換する処理をとった。また、ごく稀にではあるが、いわゆる差別語に属すると判断される表現が用いられている

場合も、同様の処理を施している(transcription.pdf のタグ(R)についての記述参照)。

しかし、以上の処理にもかかわらず、CSJ の話者を特定することは不可能ではない。知人の音声であれば、それを聞くことによって個人を同定することができるし、学会講演は、その内容から、学会名を特定することが不可能でない。CSJ の利用許諾に関する覚書に規定されているように、CSJ の利用者がデータを解析することによって知りえた個人情報を開示することは固く禁止されている。

## 8．予備的解析結果の所在

CSJ の構築過程では、想定したとおりのデータが採れているかどうかを確認するために、折に触れて予備的解析を実施した。模擬講演は本当に学会講演よりも発話スタイルがくだけているのか、自発音声と朗読音声では発話速度がどの程度異なっているのか、印象評定データと種々の言語変異現象との間には相関が認められるか、等々の検証である。予備的解析結果の一部は、国立国語研究所のホームページで公開しているので参照していただきたい。

http://www2.kokken.go.jp/~csj/public/index.html　（英文）
http://www2.kokken.go.jp/~csj/public/index_j.html　（和文）

また以下に CSJ の設計に関する文献と予備的解析結果に関する文献のうち比較的に入手が容易なものを挙げておく。


## 謝辞

**『日本語話し言葉コーパス』に音声をご提供いただいた話者の皆様、および、学会講演音声の収録を許可していただいた諸学会に心より感謝いたします。**


## 文献


Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, Hitoshi Isahara. “Spontaneous speech corpus of Japanese.” *Proceedings of the Second International Conference of Language Resource and Evaluation*, 2, pp.947-952, 2000.

前川喜久雄・籠宮隆之・小磯花絵・小椋秀樹・菊池英明「日本語話し言葉コーパスの設計」『音声研究』, 4 (2), pp.51-61, 2000.

古井貞熙・前川喜久雄・井佐原均「科学技術振興調整費開放的融合研究制度：大規模コーパスに基づく話し言葉工学』の構築」『日本音響学会誌』, 56 (11), pp.752-755, 2000.

前川喜久雄「スピーチのデータベース－『日本語話し言葉コーパス』について－」『日本語学』20 (6), pp.12-27, 2001.

前川喜久雄「話し言葉における長母音の短呼－『日本語話し言葉コーパス』を用いた音声変異の分析－」『国語学会 2002 年度春季大会要旨集』, pp.43-50, 2002.

籠宮隆之・小磯花絵・小椋秀樹・山口昌也・菊池英明・間淵洋子・土屋菜穂子・斎藤美紀・西川賢哉・前川喜久雄「大規模自発音声コーパス『日本語話し言葉コーパス』の仕様と作成」『国語学会 2002 年度春季大会要旨集』, pp. 225-232, 2002.

小磯花絵・斎藤美紀・間淵洋子・前川喜久雄「話し言葉における助詞の撥音化現象の実態－『日本語話し言葉コーパス』を用いて－」『第 10 回社会言語科学会研究大会予稿集』, pp. 215-220, 2002.

前川喜久雄「『日本語話し言葉コーパス』を用いた言語変異研究」『音声研究』, 6 (3), pp.48-59, 2002.
小磯花絵　『コーパスによる音声談話の研究」『日本語学』22(4 月臨時増刊号), pp.200-209,2003.
Kikuo Maekawa, Hanae Koiso, Hideaki Kikuchi, and Kiyoko Yoneyama. “Use of a large-scale spontaneous speech corpus in the study of linguistic variation.” *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, pp. 643-646, 2003.

## 付録１　再朗読および対話における話者とファイル名の対応

| 出自 | ID | 再朗読 | 元講演 | 課題指向 | 自由 | INT 模擬 | INT 学会 | 模擬 |
|---|---|---|---|---|---|---|---|---|
| 学会 | 8 | R00M0132 | A01M0074 | D02M0016 | D03M0048 | D01M0005 | D04M0041 | S05M1506 |
| 学会 | 19 | R00F0028 | A05F0043 | D02F0015 | D03F0058 | D01F0002 | D04F0022 | S05F1041 |
| 学会 | 373 | R00M0415 | A11M0469 | D02M0024 | D03M0038 | D01M0020 | D04M0021 | S05M1623 |
| 学会 | 423 | R00M0187 | A01M0056 | D02M0028 | D03M0017 | D01M0019 | D04M0056 | S05M0613 |
| 学会 | 463 | R00F0178 | A01F0122 | D02F0031 | D03F0008 | D01F0030 | D04F0044 | S05F0463 |
| 学会 | 471 | R00M0134 | A11M0369 | D02M0051 | D03M0004 | D01M0009 | D04M0010 | S05M1666 |
| 学会 | 514 | R00F0407 | A06F0128 | D02F0025 | D03F0006 | D01F0023 | D04F0050 | S05F1600 |
| 学会 | 674 | R00F0149 | A01F0861 | D02F0032 | D03F0001 | D01F0003 | D04F0011 | S05F1385 |
| 学会 | 685 | R00M0036 | A11M0846 | D02M0039 | D03M0037 | D01M0047 | D04M0052 | S05M1236 |
| 学会 | 1185 | R00F0304 | A11F0703 | D02F0054 | D03F0034 | D01F0049 | D04F0029 | S05F0391 |
| 模擬 | 68 | R00F0061 | S02F0852 | D02F0033 | D03F0036 | D01F0046 | - | S05F0378 |
| 模擬 | 251 | R00M0286 | S01M0998 | D02M0035 | D03M0053 | D01M0043 | - | S05M0417 |
| 模擬 | 790 | R00M0071 | S02M0444 | D02M0014 | D03M0013 | D01M0012 | - | S05M1406 |
| 模擬 | 808 | R00F0306 | S01F0664 | D02F0027 | D03F0045 | D01F0055 | - | S05F0899 |
| 模擬 | 910 | R00M0019 | S01M1401 | D02M0026 | D03M0007 | D01M0042 | - | S05M0956 |
| 模擬 | 1125 | R00F0287 | S03F0737 | D02F0018 | D03F0040 | D01F0057 | - | S05F1664 |

**出自：**　**講演者が学会講演経験者か模擬講演経験者かの別**

ID：　**講演者 ID**

**再朗読：**　**再朗読のファイル名**

**元講演：**　**再朗読の対象としたオリジナルの学会講演ないし模擬講演ファイル名**

**課題指向：**　**課題指向対話のファイル名**

**自由：**　**自由対話のファイル名**

INT **模擬：**　**模擬講演（＝模擬）に関するインタビューのファイル名**

INT **学会：**　**学会講演（＝元講演）に関するインタビューのファイル名。出自＝学会の話者のみ。**

**模擬：**　**新規に収録した模擬講演のファイル名（これについてインタビューを実施）**

## 付録２　音声認識研究用テストセットの構成

| ファイル名 | コア | 形態論情報 | セット |
|---|---|---|---|
| A01M0110 | コア | 手作業 | 1 |
| A01M0137 | コア | 手作業 | 1 |
| A01M0097 | コア | 手作業 | 1 |
| A01M0056 | コア | 手作業 | 2 |
| A03F0072 | コア | 手作業 | 2 |
| S00M0112 | コア | 手作業 | 3 |
| S00F0066 | コア | 手作業 | 3 |
| S00M0213 | コア | 手作業 | 3 |
| A04M0123 | 非コア | 自動解析 | 1 |
| A04M0121 | 非コア | 自動解析 | 1 |
| A04M0051 | 非コア | 自動解析 | 1 |
| A03M0156 | 非コア | 自動解析 | 1 |
| A03M0112 | 非コア | 手作業 | 1 |
| A03M0106 | 非コア | 手作業 | 1 |
| A05M0011 | 非コア | 手作業 | 1 |
| A02M0012 | 非コア | 自動解析 | 2 |
| A03M0016 | 非コア | 自動解析 | 2 |
| A06M0064 | 非コア | 自動解析 | 2 |
| A06F0135 | 非コア | 自動解析 | 2 |
| A01F0034 | 非コア | 手作業 | 2 |
| A01F0063 | 非コア | 手作業 | 2 |
| A01F0001 | 非コア | 手作業 | 2 |
| A01M0141 | 非コア | 手作業 | 2 |
| S00F0019 | 非コア | 自動解析 | 3 |
| S00M0079 | 非コア | 自動解析 | 3 |
| S01F0105 | 非コア | 自動解析 | 3 |
| S00F0152 | 非コア | 手作業 | 3 |
| S00M0070 | 非コア | 手作業 | 3 |
| S00M0008 | 非コア | 手作業 | 3 |
| S00F0148 | 非コア | 手作業 | 3 |

**コア：　　ファイルがコアに含まれるかどうか**
**形態論情報：形態論情報が手作業（高精度）か自動解析か（図５参照）**
**セット：　テストセットの下位区分（詳細は** asr.pdf **参照）**

# 付録３　主要なファイルの配置とファイル名拡張子

CSJ を格納した DVD-ROM セットに、どのような方法でファイルが記録されているかを簡単に説明する。詳しくは各ディスクのルートディレクトリに置かれた README ファイルを参照すること。

１）Disk1

Disk1 は以下のディレクトリ構造をもち、各種解説文書、データブラウジング用ツール、さらに個々の講演から独立したデータ類を格納している。

| | |
|---|---|
| /DOC | 本文書の表 1 に記載した解説文書類 |
| /TOOL/XSL | XML 関係のスクリプト類 |
| /MNFC | 音声・テキストブラウザー |
| /XML_BROWSER | XML 検索ツール |
| /AM | 音響モデル |
| /LM | 言語モデル |
| /DATA | 講演者属性データ、記録票データ、印象評定（集合評定）データなど、CSJ 全体にかかわるデータ各種 |
| /TRN-EUC | 転記ファイル(EUC)* |
| /TRN-SJIS | 転記ファイル(シフト JIS)* |

*個々の講演のディレクトリとは別にすべての転記ファイルがここにも格納されている。

| | |
|---|---|
| /SUWDIC | 短単位辞書 |
| /LDB | 長単位形式形態論データ** |
| /SDB | 短単位長単位混合形式データ** |

**個々の講演のディレクトリとは別にすべての形態論情報データがここにも格納されている。

２）Disk2

Disk2 には XML 文書をまとめて格納している。個々の XML 文書は Disk3 以降の各講演のディレクトリにも格納されていることに注意。

３）Disk3 以降

Disk3 以降には、個々の講演に関わるデータ類を格納している。ルートディレクトリの直下に講演 ID(“ A01M0001 ”,“ S03F0099 ”等)を名称とするディレクトリが多数並んでいる。各ディレクトリには、すべての講演に対して提供されるファイル群が格納されている。講演 ID として“ A01M0001 ”を例にとると以下のとおりである。

| | |
|---|---|
| /A01M0001/A01M0001.wav | （音声ファイル）* |

*独話はモノラル、対話は２チャンネルインターリーブ

| | |
|---|---|
| /A01M0001/A01M0001-L.wav | （対話音声の左チャンネルを分離したモノラル音声） |
| /A01M0001/A01M0001-R.wav | （対話音声の右チャンネルを分離したモノラル音声**） |

**上記２種類のファイルは対話音声にだけ提供される

| | |
|---|---|
| /A01M0001/A01M0001.trn | （転記ファイル；シフト JIS） |
| /A01M0001/A01M0001.xml | （XML 文書；UTF-8） |
| /A01M0001/A01M0001.ldb | （長単位形式形態論データ） |
| /A01M0001/A01M0001.sdb | （短単位長単位混合形式形態論データ） |

ある講演がコアに含まれているか、要約・重要文抽出の対象となっていれば、講演 ID に対応するディレクトリの下に３個のサブディレクトリが作成されている。再び“A01M0001”を講演 ID とすると以下のとおりである。

/A01M0001/PLABEL　　（分節音およびイントネーション(X-JToBI)のラベル）
/A01M0001/SUMMARY/50PER（自由要約データ、要約率 50%）
/A01M0001/SUMMARY/10PER（自由要約データ、要約率 10%）

Disk3 以降の収録内容は以下のように分類されている。

Disk3　コアに含まれる学会講演・対話・再朗読
Disk4　コアに含まれる模擬講演
Disk5　コア以外の女性による学会講演および女性による「その他（M01～M03）」
Disk6　コア以外の男性による学会講演（A01）
Disk7　コア以外の男性による学会講演（A02～A03）
Disk8　コア以外の男性による学会講演（A04～A06）
Disk9　コア以外の男性による学会講演（A07～A10）
Disk10 コア以外の男性による学会講演（A11～A13）および男性による「その他（M01～M03）」
Disk11 コア以外の女性による模擬講演（S00～S02）
Disk12 コア以外の女性による模擬講演（S03～S06）
Disk13 コア以外の女性による模擬講演（S07～S11）
Disk14 コア以外の男性による模擬講演（S00～S03）
Disk15 コア以外の男性による模擬講演（S04～S07）
Disk16 コア以外の男性による模擬講演（S08～S11）
Disk17 コア以外の対話・朗読・再朗読

Disk1 の/DATA ディレクトリ中の talk_data.dat ファイルの「収録ディスク」フィールドには、講演 ID と収録ディスクの対応が記入されている。これによって、ある特定の講演がどのディスクに収録されているかを検索することができる (data_attribute.pdf 参照)。

以上。