『日本語話し言葉コーパス』の概要

『日本語科学』15(2004年4月) 111-133

『日本語話し言葉コーパス』の概要

前川 喜久雄
(国立国語研究所)

キーワード

『日本語話し言葉コーパス』, 自発音声, データベース, XML

要 旨

現代日本語の大規模な自発音声データベースである『日本語話し言葉コーパス』を紹介する。まず話し言葉研究におけるデータベースの必要性を指摘したのち,『日本語話し言葉コーパス』公開版の仕様を紹介する。締めくくりとして, 日本語のコーパス言語学について簡単な展望を述べる。

1. はじめに

書き言葉と話し言葉の研究を比較すると, 話し言葉の研究には何かと制約が多い。書き言葉のテキストは, 電子的手段で作成されたものであれば, ほぼそのまま研究の一次資料として利用できる。さらに, テキストを語に分割して品詞情報を付与することも, 現在ではかなり高い精度で自動実行することができるようになっている。

これに比べて話し言葉の研究では, 録音した音声を文字に転記する作業がまず大変な手間を要求する。しかも—この点が重要なのだが—, ただ単に音声を文字に転記しただけではイントネーションやポーズなどの韻律的特徴が脱落してしまうので, 理想的にはこれらの情報まで含めた転記が必要になる。そうしないと, 或る発話が断定なのか質問なのか, 発話のどこに強調が置かれているかといった基本的な情報が分明でなくなることがありうる。韻律情報(正確にはパラ言語情報)は, 話し言葉と書き言葉の本質的な相違点にかかわる情報である(前川2000)。

さらに, 言い誤りや言い淀みのような現象の転記も必要である。これらの現象は会議録などの書き起こしでは省略されるのが普通であるが, 言語心理学的な研究のためには, こうした非流暢性の要素が重要であることがわかっている。そのため転記テキストは一層複雑化し, 作成コストが増大する。話し言葉研究用データのコスト高は, 話し言葉の研究が書き言葉にくらべて著しく立ち遅れていることの最大の理由のひとつであろう。

国立国語研究所は1948年の創立以来多くの調査研究を実施してきているが, やはり, その大部分は書き言葉を対象とした調査であった。そのなかで『談話語の実態』(国語研1955)と『話しことばの文型』(国語研1960, 1963)の報告書にまとめられた調査は, 話し言葉を正面きってとりあげた研究として異彩をはなっており, 現在でも引用されることが少なくない。しかし『話しことばの文型』以降は, 話し言葉そのものの特色を解明するための研究は国立国語研究所の公式な研究課題から姿を消してしまうことになった[1]。本稿で紹介しようとする『日本語話し言葉コ

ーパス』は，この話し言葉調査の系譜を現在に蘇らせる試みである。

『日本語話し言葉コーパス』（以下ではその英語正式名称であるCorpus of Spontaneous Japaneseを略してCSJと呼ぶ）は国立国語研究所，情報通信研究機構（旧通信総合研究所），東京工業大学の三者が共同開発した現代日本語の話し言葉研究用データベースであり，プロジェクトの総括責任者は東京工業大学の古井貞熙教授である。開発費用の多くは科学技術振興調整費開放的融合研究制度補助金に拠った。研究課題名は「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」，研究期間は1999-2003年度であった（古井他2000）。

CSJには時間にして約660時間，語数にして750万語以上の話し言葉が格納されている。上述した『談話語の実態』で分析された録音資料が約9時間分であることと較べれば，CSJの大きさを理解していただけるだろう。CSJは日本語の音声データベースとして最大であるだけでなく，世界の主要音声データベースと比較しても遜色がない。研究用に付加された情報の多様性と精度の高さにおいては，むしろ諸外国のデータベースを凌駕している。データベースの価値が，そのデータ量と付加情報の多様性の積で決まるとすれば，明らかに現時点における世界最高の音声言語データベースである。CSJは，2004年3月をもって予定通りに開発を終了し，近く一般公開する予定である。次節以下ではCSJ公開版の仕様を紹介する。

2．設計

2.1．基本方針

CSJのような音声言語データベースはこれまでにも世界各地で構築されてきている。それらは二種に大別できる。ひとつは，1980年代から世界中で盛んに構築され始めた音声情報処理用のデータベースである。これは，大量の学習データを用いて音声の自動認識や合成を行なおうとする工学的研究に用いられたもので，その内容は，単語と文章を多数の話者が読み上げたものが中心である[2]。この種の音声は朗読音声（read speech）と呼ばれている。

朗読音声の話者は職業的な朗読者（ナレーターやアナウンサー）であることが多く，当然ながら，誤りのない理想化された音声になっている。音声の他に提供されるのは，朗読用テキストとその音素表記程度であり，韻律情報が提供されることは稀である。

もうひとつは音声学や言語学のために構築されたデータベースである。英国で1959年に開始されたSurvey of English Usage（SEU）のデータがその嚆矢となった（現在はLondon-Lund Corpusの名で知られている。Svartvik and Quirk(1980)参照）。SEUは書き言葉と話し言葉の双方を対象とした調査であり，全体の半分，約50万語分が話し言葉データにあてられていた。そのうち76%が独話音声，24%が対話音声である。話し言葉データの大半は，一般話者による，練習無しの自発音声（spontaneous speech）であり，さらに韻律特徴や言い淀み等の情報も付与されているので，非常に利用価値の高いデータであるのだが，残念なことに肝心の音声そのものは提供されていない。そのため，ユーザーは転記テキストに埋め込まれた複雑な音声記号群から音声を想像しなければならない。もちろん音声情報処理に利用することもできない。

もうひとつ，BNC（British National Corpus）の例を挙げておこう。BNCは英語の辞書学や

コーパス言語学で広く活用されているデータベースであり，１億語のうち1000万語を話し言葉に充てているが，やはり音声は公開されておらず，音声記号によるアノテーションも与えられていない（Aston and Burnard 1998)[3]。

我々は上に述べた二種類の音声言語データベースそれぞれの特長をCSJで同時に実現しようと考えた。これは，1998年にATR音声翻訳通信研究所（当時）の山本誠一氏の肝煎で我々が科学技術振興調整費への応募を考慮しはじめた当初からの方針であった。具体的には，対象を自発音声とし，自発音声の音声認識技術を開発するために必要なデータ量を確保しながら，一方で音声・言語研究のための付加情報も豊富に提供しようというよくばった設計方針である(Maekawa et al. 2000；前川他2000；前川2001)。

2.2. CSJの構造

一定の研究コストの制約内で上記の設計方針を実現するためには，それなりの工夫が要る。我々はデータベースに一種の階層構造を導入して付加情報に濃淡をつけるという方策を採用した。

図１にCSJの階層構造と，研究用付加情報の濃淡を示した。最初にCSJ全体のサイズを700万語（短単位；4.2参照）と見積もった[4]。これは音声認識研究に最低限必要なデータ量を朗読音声の認識研究での知見に依拠して推定したものである。この700万語に対しては，音声の他に精密な転記テキストと形態論情報（つまりテキストを語に区切って品詞をつけた情報）を提供する。これらは音声認識研究を実施するために最低限必要な情報である。また，講演音声が聴き手にどのような印象を与えたかを主観的に評定したデータ（印象評定データ）と，転記テキストにおける節（clause）境界の情報（節単位情報）も提供し，さらに話し手に関する種々の属性情報(性別，年齢など）も提供する。

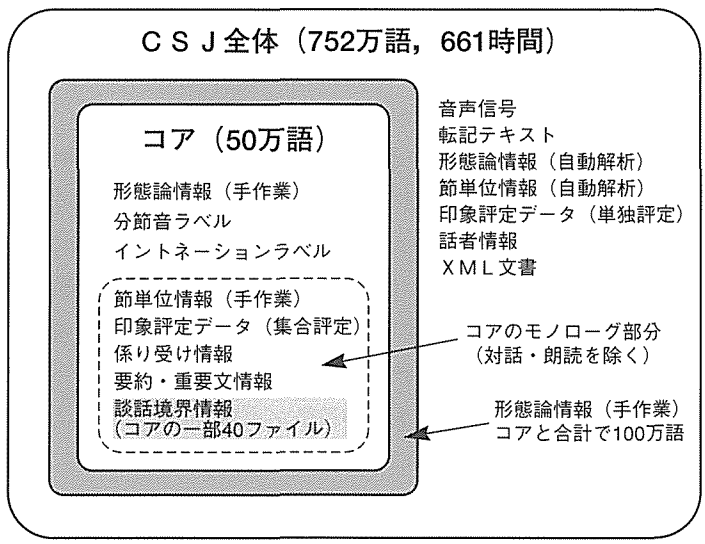


図１　CSJにおける研究用情報の付与方式

一方，CSJの一部，約50万語に限っては上よりもはるかに豊富な研究用情報を提供することにした。我々はこの50万語をデータベースの中核部分という意味で「コア」(Core) と呼びならわしている。50万語というサイズは，研究コストから逆算して処理可能な最大データ量を見積もって決定した。コアだけに付与される研究用付加情報は以下のものである。

— 分節音ラベル　(4.5.参照)
— イントネーションラベル　(4.5.参照)
— 印象評定データ（集合評定）　(2.5.2.および4.4.参照)
— 係り受け構造情報　(4.6.1.参照)
— 要約・重要文情報　(4.6.2.参照)
— 談話境界情報　(4.6.3.参照)

形態論情報と節単位情報はCSJの全体に提供される情報であるが，コアに対しては手作業で綿密な分析を実施しており，コア以外に対するものよりも情報の精度が高い。コアの形態論情報は，コア以外のデータを自動解析するための学習データとして利用されている。

最後に図1中の網掛けを施した部分は，コア以外であるが，手作業による形態論情報が付与されている部分である。手作業による高精度形態論情報は，コアを含めて約100万語に対して付与されている。

2.3. 対象とする音声

CSJの対象は自発音声である。しかしひとくちに自発音声といっても実際には多種多様である。まず問題となるのが，独話（モノローグ）と対話（ダイアローグ）の別であるが，CSJでは独話を中心に据えることにした。その理由は，現在の音声認識研究が基本的に独話を対象としているからである。言語研究者のなかには対話にしか興味がないという人もいるようだが，実は日本語の場合，対話のデータベースは少量であっても或る程度整備されているのに対して，自発的な独話のデータベースは存在していない。このことを考えると，言語研究の観点からも独話データは価値が高いと考えられる。

次に自発性には高低さまざまな段階がある。CSJでは親しい間柄での雑談のように極端に自発性の高い発話は対象とせず，従来研究されてきた朗読音声よりは自発性が高いが，音声だけを聞いても内容が十分に理解できる，まとまった内容をもつ発話を対象に据えることにした。これもやはり工学的応用を考えての選択である（自発性の問題については2.5参照)。

またCSJではいわゆる標準語を対象とすることにした。標準語という概念を正確に規定することは難しいが，我々は「高校卒業程度の教育をうけた現代人が多少とも公的な場面で用いる日本語で，分節音の音韻特徴および語彙・文法上の特徴が東京方言に酷似したもの」というやや大雑把な規定によってデータを選別することにした。

この規定は，韻律特徴については何も言及していないので，アクセントが明らかに東京方言とは異なる発話もCSJには収録されている。ただしコアには韻律特徴のラベルを付与する関係上，韻律特徴が東京式と判断された音声だけを格納している。

2.4. 音声の種類と量

2.4.1. 学会講演と模擬講演

2.3節に述べた方針に適う音声として学会講演と模擬講演を収録することにした。学会講演は，人文，理工，社会の各領域にまたがる様々な学会での研究発表を実況録音した音声である。学会講演は内容が論理的であると期待できるから，上述の音声認識・要約技術が最初に対象としてとりあげるべき種類の音声である。

各学会から承諾をいただいた後に講演者に連絡をとり，データベースが公開されることを承知のうえで承諾書を提出してくださった講演者の口頭発表を収録した。1999年から2001年にかけて収録した学会講演は延べ987件に達している（後掲の表2参照）。

しかし，学会講演の話者には言語学的に見て強い偏りがある。どの学会でも講演者には大学院生が多いため年齢が20代半ばから30代前半に集中しており，理工系学会では大半が男性である。また専門領域ごとに使用語彙の著しい偏りがあることも想像に難くない。さらに学会講演は一般にスタイルの高い発話が多く，少数ではあるが原稿を朗読しているに近い講演もある。

これらの問題を解決するために企画されたのが模擬講演である。人材派遣会社に依頼して年代(20代から60代まで）と性別に偏りのない話者を派遣してもらい，当方で指定した一般的テーマにそった10分程度のスピーチを各人に3種類語ってもらった（ただし最初期に収録した一部のデータに関しては話者のバランスがとれておらず，テーマも指定していない)。表1に指定したテーマのリストを示す。人材派遣会社にはできるだけ首都圏出身の話者を選択するよう依頼したが，この要望は必ずしも叶えられていない（3節参照)。

話者には収録の二日ほど前にテーマを連絡した。話者は収録までに各テーマについて具体的なスピーチを考え，その概要を簡単なアウトラインにまとめてタイトルをつける。例えばテーマ1

表1　模擬講演のテーマ

0	（指定なし）
1	人生を振り返って嬉しかった・楽しかった出来事
2	人生を振り返って悲しかった・つらかった出来事
3	住んでいる町や地域について
4	よく知っていること，興味・関心のあることの客観的説明
5	人生を振り返って印象に残っていること
6	過去数年の間にマスコミで扱われたニュース
7	無人島に持っていくもの3つ
8	～のやり方，作り方*
9	～の歴史*
10	自分にとっていちばん大事なもの・人
11	21世紀に残したいもの・残したくないもの

* ～は話者が選択する

であれば「大学に合格したこと」，テーマ２であれば「母の死」などである。講演用の朗読原稿を準備することは禁止した[5]。模擬講演の話者からもデータ公開の承諾書を頂戴している。

模擬講演の総数は1715件である。初期に収録した一部を除けば，すべて国立国語研究所内の音声スタジオで収録した。模擬講演を収録する目的のひとつは，学会講演よりも低いスタイルの発話を収録することにあったから，可能なかぎりリラックスした状態で講演してもらうために工夫をこらした（2.5参照）。後述する印象評定値および収録されたデータの予備的分析結果をみると，統計的には模擬講演のスタイルが学会講演よりも低下していることがわかる（前川2001；Maekawa et al. 2003）。このスタイル差はCSJを社会言語学的な研究に利用しようとする研究者に利便をもたらすと考えられる。

2.4.2. 対話など

表２にCSJに収録された音声の内訳を示す。CSJの約90％（605時間）は学会講演と模擬講演であるが，それ以外に約55時間の音声が収録されている。そのうち「その他」に分類されているのは，一般向けに開催された博物館の連続講演会，専門学校における日本語学の講義，国立国語研究所が開催した一般向け講演会などの独話である。いずれも学術的な講演であるが，話し手が専門家，聴き手が一般人という点で，学会講演とは相違している。

「学会講演インタビュー」から「自由対話」までは合計約12時間分の対話音声である。また「朗読」（新書から抜粋した自然科学に関するテキスト２種類を模擬講演話者が朗読したもの）と「再朗読」（収録済の自発音声の転記テキストを同一話者が朗読した音声）も合計約21時間分が収

表２　CSJに格納された音声の種類とその内訳

音声の種類	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

録されている。これらは，CSJの中心を占める学会講演ないし模擬講演と比較対照してCSJに格納された独話の性質を評価するために収録したものである。16名分と量は限られているが，同じ話者による学会講演（10名のみ），模擬講演，4種類の対話，再朗読データも提供されているので，独話と対話の違いなど，音声の種類による影響を同一の話者グループにおいて比較できる。12時間程度とはいえ，対話音声も従来の水準からすれば少なからぬ量が収録されているので，目的によっては独話と切り離して単独で分析することも不可能でない。

2.5. 音声の自発性

話し言葉の多様性を考える際に重要な観点となるのが音声の自発性（spontaneity）の問題である。音声研究では，音声を「朗読音声」と「自発音声」に二分することが多い（2.1参照）。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[6]。データベースに存在する自発性の程度差は，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或る程度客観的に評価できることが望ましい。CSJで採用した方策を以下に説明する。

2.5.1. 自発性の序例

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（記号'<'はその左側よりも右側の方が自発性が高いことを示し，'<<'はその差が顕著であることを示す）。

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る。また，学会講演と模擬講演は量的にCSJの大部分を占める音声でもある。そのために，両者間の差異をきわだたせる対策をとることにした。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等の対策である（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（前川2001）。

2.5.2. 印象評定

上に示した序列は，しかしながら，絶対的なものではない。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。印象評定には，このようにして収集したコーパスのほぼ全体に対するデータ（単独評定データ）の他に，コアの独話だけを対象としてより詳細な評定をおこなったデータ（集合評定データ）がある。印象評定データは4.4で紹介する。

3．話者の分布

話し言葉の多様性の一部は，性別，出生地，居住歴，学歴，講演経験の有無など，話者の社会

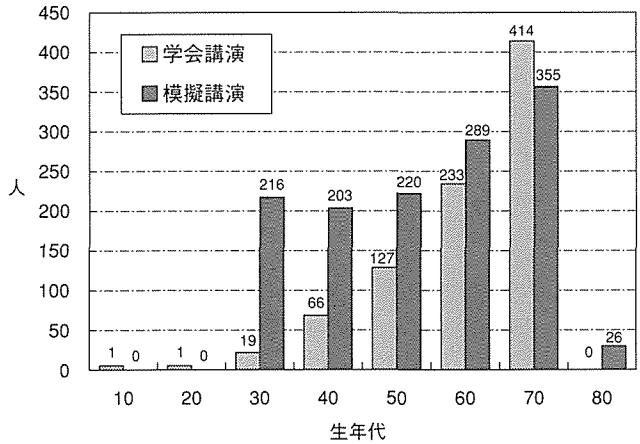


図2　学会講演話者と模擬講演話者の生年による分布(延べ)

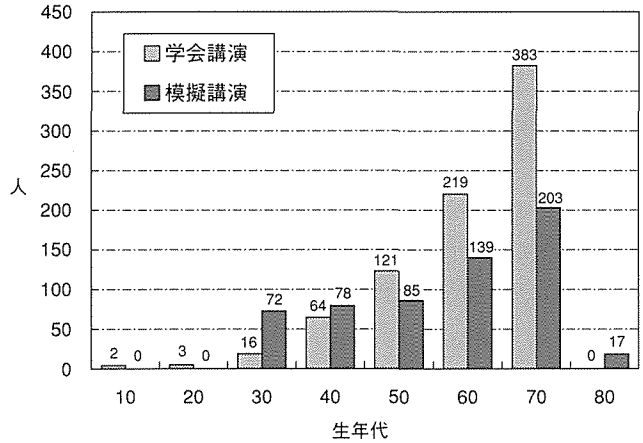


図3　学会講演話者と模擬講演話者の生年による分布(異なり)

的属性に起因している。そのため，話し言葉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。ここでは，CSJ公開版を対象として，最も代表的な話者属性である話者の生年代，性別，出生地の分布を概観する。

まず，図2，3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2，3ではこれを10年ごとに区分しなおして集計した。図2は生年代ごとの延べ話者数，図3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2においても図3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。

次に，表3，4に話者の性別と音声の種類によるクロス集計を示す。表3が延べ話者数，表4が異なり話者数である。表4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種類にまたがってデータを提供している場合に重複してカウントしているためであり（ひとつの種類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。

先にも述べたように，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学会において著しい（ちなみに図2，3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

表3　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

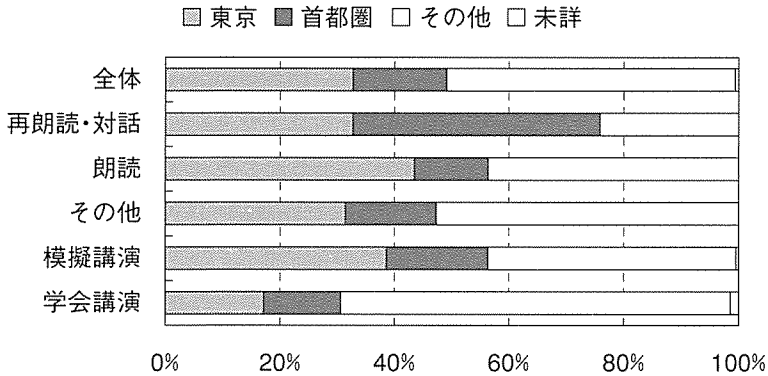
表4　話者の性別の分布(異なり)

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，****インタビュワーを含めると471名

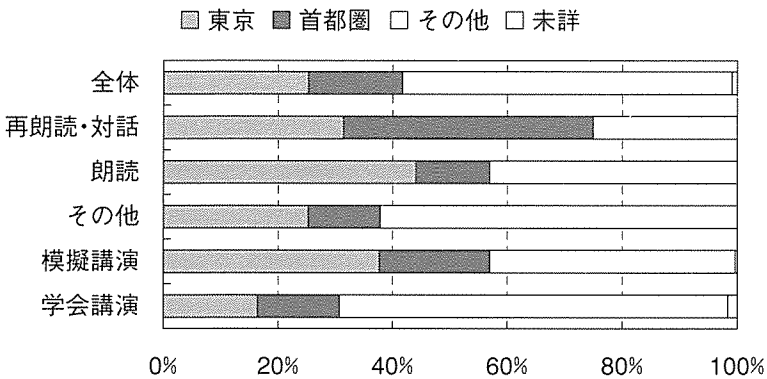
大学院生の多さによる)。一方，模擬講演以下では，男女がほぼ均等に分布している。

最後に，図4，5に話者の出生地による分布を示す。図4が延べ話者数，図5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。社会言語学などの研究においてはさらに詳しい履歴が必要



	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4　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5　話者の出生地の分布(異なり)

とされるであろうが，その種の情報もプライバシーを侵害しない範囲で公開されている。

4．研究用付加情報（アノテーション）

本節ではCSJに付与された研究用付加情報について概観する。紙幅の関係で細部には触れることができないので，詳細な情報はCSJに同梱されるマニュアル類を参照していただきたい。

4.1. 転記テキスト

収録された音声は，そのままでは検索することができないので，これを文字に書き起こした転記テキストを作成する必要がある。この作業の精度によってデータベースの価値が決まると言ってよい重要な作業である。音声認識に用いる言語モデルの精度もこの作業に強く依存する。

話し言葉を転記しようとする際に必ず遭遇する重要な問題は，転記の単位をどう定めるかという問題である。CSJでは文法的ないし統語的な基準は採用せず，長めのポーズ（原則として0.2秒以上）位置で音声を転記基本単位に分割している。各転記基本単位には開始時刻と終了時刻の情報が提供されている。ポーズという物理的な指標によって転記の単位を定めたのは，CSJのように大量のデータを扱う場合，その全体に一貫して適用可能な「文末」の言語学的特徴を客観的に規定することが，実際上不可能であると判断されたからである。そのためCSJの転記基本単位は統語的な文末と一致しているとは限らない。

CSJの転記テキストには，発話を漢字仮名まじりで表記した基本形と片仮名だけで表記した発音形の2種類がある。基本形は主として情報検索のための利用を想定しているので表記にゆれを生じさせないことを徹底して追及した（小磯他2001）。

一方，発音形の役割は，基本形の漢字の読みを確定させると同時に，発音上の変異を正確に表記することにある。「私」が「ワタクシ」か「アタクシ」か，「本当」が「ホントー」か「ホント」か，「前川」が「マエカワ」か「マエカー」か，「国語研」が「コクゴケン」か「コッゴケン」か等々が，人間の耳で聞き分けられ仮名文字で表現できる範囲で，可能なかぎり正確に表記されている。発音形は，日本語話し言葉の音声変異について貴重な情報を提供する。また近年の音声認識技術で重要性を増している発音辞書の構築にも利用できる（堤他2004；河原2004）。

転記テキストには多くのタグが挿入されている。代表的なタグに「エー」「アノー」等の言い淀みを表す（F），言いさしによって断片化された語を示す（D），聞き取りが困難な箇所を示す（?），発音の転訛ないし不正確な発音を示す（W），非語彙的な母音の延長（「あれが」が「アーレガ」と発音されるようなケース）を示す＜H＞などがある。タグの多くは当該文字列を囲む形で転記テキスト中に挿入されている（タグの詳細は小磯他(2001)およびCSJに同梱されているマニュアル—Disk1/DOCのtranscription.pdf—参照[7]）。図6にCSJの転記テキストの例を示した。

```
0087 00187.217-00193.684 L:
(F あーの)                    &(F アーノ)
オーストラリアに              & オーストラリアニ
行くと                        &(? イク)ト
大概                          & タイガイ
(F あの)                      &(F アノ)
ビーチの                      & ビーチノ
あるところに                  & アル<H>トコロニ
滞在したりとか                & タイザイシタリトカ
住んだりっていう              & スンダリッテユー
経験が                        & ケーケンガ
あるんですが                  & アルンデスガ
0088 00194.417-00194.918 L:
で                            &(? デ)
(F えー)                      &(F エー)
0089 00195.255-00195.979 L:
(F ま)                        &(F マ)
どうしても                    & ドーシテモ
こう                          & コー
0090 00196.284-00197.702 L:
ビーチに                      &(W ピーチ;ビーチ)ニ
行くと                        & イクト
嬉しいので                    & ウレシーノデ<H>
(以下略)
```

図6　転記テキストの例

模擬講演の例。「0087」等の数字で始まる行は，転記基本単位の通し番号，開始時刻，終了時刻を示している。転記基本単位中は文節で改行されており，&で区切られた左側が基本形，右側が発音形である。図中で用いられているタグについては本文参照。

4.2. 形態論情報

形態論情報とは発話を語に分解して品詞分類を施した情報である。その際，当然ながら，語をどう規定するかによって結果が異なってくる。この問題はあらゆる言語に存在するが，日本語のように造語法上の自由度が高い言語では殊に重要である。理論上は，漢字のひとつひとつが単位となってしまうような短い単位から，いわゆる臨時一語（例えば「国立国語研究所外部評価委員会報告書」）が一単位となるような長い単位までを考えることができる。

CSJでは，国語辞典の見出し語に該当するような短めの単位と，それよりも長めの単位との2種類を採用して二重の形態論情報を提供している。これらをそれぞれ短単位，長単位と呼ぶ。例えば「これからディズニーワールドについてお話しいたします」というテキストは，短単位では「これ|から|ディズニー|ワールド|に|つい|て|お|話し|いたし|ます」と11単位に，長単位では「これ|から|ディズニーワールド|について|お話しいたし|ます」と6単位に分解される。

短単位および長単位の設計については，CSJに同梱されるマニュアル（小椋他2004）に詳し

いが，２種類の形態論情報を同時に提供することによって，日本語の造語法についての貴重な知見を得ることができる。また，語と韻律特徴との関係を吟味する研究のためにも，二重の形態論情報は有益であると思われる。

CSJ の形態論分析では，まず，コアの全体を含む短単位で100万語相当のテキストを国語研究所の研究員が手作業で分析した。このデータは情報通信研究機構に渡されて，形態素自動解析ソフトウェアの学習用データとして利用された。CSJ のうち上記100万語を除外した残り650万短単位は，このソフトウェアによって自動解析されたものである[8]。自動解析結果には国立国語研究所において可能な限りの手修正をくわえている。

手作業による短単位形態論情報の精度はランダムサンプリングによって約99.9%と推定されている。これを1000語にひとつも誤りがあると考える方もあるかもしれないが，実際に話し言葉のデータを分析してみると，語境界や品詞を一意に決定しがたいケースが1000語にひとつ程度は出現するので，この数字は人知の限界であると考えている。自動形態素解析の精度は手作業に較べると若干低下することは避けられず，おそらく98%前後である。そのためコアを含む100万短単位とそれ以外とでは形態論情報の精度が相違している。

表５は CSJ に含まれる長短単位数を音声の種類ごとに示している。言い誤りによって生じた語の断片と言い淀みは除外されている。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。山口(2003)は学会講演と模擬講演とでは品詞の分布に組織的な差が生じていることを報告している。

表５　形態論的単位数

音声の種類	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

4.3. 節単位情報

自発的な独話音声では，形態論的に典型的な文末特徴が生じることなく発話が連綿と続いてゆくことがある。「みんなで相談したんですけど，賛成しようということになって，私は反対だったんだけど，それでもみんなは賛成なんで，一応賛成しようということになったんだけど，やっぱり私は…」というような発話である。

書き言葉を基準にしてこの種の発話を分析すると大変な長文が生じてしまう。しかし，話し言葉として見た場合，「節」(clause) が情報処理上の単位として機能している可能性が高い。上例に読点を挿入した箇所である。節境界の情報は，以下に述べる談話境界情報や係り受け情報を作成する際の単位の切り出しに利用することができるし，それ以外にも多くの利用が可能であると考えられる。

CSJ には，転記テキストを解析して節境界の位置を検出し分類した情報が提供される。この情報付与作業には，ATR 音声言語コミュニケーション研究所で開発された節境界解析プログラム CBAP（丸山他2003）を CSJ 用に改良して利用した。また，既に述べたようにコアに関しては自動解析結果を人手で修正した高精度情報を提供している。節単位情報の一例を下に示す。ゴチック部分が付加情報，それ以外は転記テキストである。

私は旅行が大好きで／**並列節デ**／＋今までもあちこち行きましたけれども／**並列節ケレドモ**／；**主題の共有**

この例にはふたつの節が含まれており，いずれも並列節に分類されている。最後の付加情報は，係助詞「は」でマークされた主題（「私」）がふたつの節で共有されていることを記述している（高梨他2003)。

4.4. 印象評定データ

既に述べたように，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，その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。例えば発話の自発性に関する項目としては「自発性」と「発話スタイル」があり，いずれも 5 段階で評定されている。前者は「講演に際してどれだけ原稿を読みあげているか」（原稿への依存度が低いほど自発性の評定値が高い)，後者は発話が「どれだけ改まっているか」（改まり度が高いほどスタイルの評定値が高い）を評定している。上記以外の 5 段階評定項目には「難関な専門用語の多少」「発話スピード」「発音の明瞭さ」「方言の多少」がある。

5 段階評定項目以外に，講演の印象にあてはまる形容語句をリスト中から自由に選択する形式の評定も実施した。リストには以下の語句が含まれており，評定者は複数の語彙を選択できる：「たどたどしい，流暢な，単調な，表情ゆたかな，自信のある，自信の無い，優しい，落ち着いた，落ち着きのない，いらいらした，緊張した，リラックスした，大きい声，小さい声，かすれた声，裏返った声，こもった声，重厚な，軽薄な，若々しい，年寄じみた，元気のある，元気のない，聞き取りやすい，聞き取りにくい，生意気な，尊大な，鼻にかかった，高い，低い，きっぱりした」。

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。また，

上に示した形容語句のリストも慎重に検討して作成したものではなかった。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，全データの収録を終えた後に，コアに含まれる独話音声を対象として作成した。1講演の冒頭，中程，終盤からそれぞれ1分程度の連続した音声サンプルを抽出し，それを20名の評定者が独立に評定した。CSJには，そのうち，評定の再現性が高いことが保証されている10名分のデータが格納されている。また，単独評定の評定が心理学的な厳密性に欠けるのに対して，集合評定では実験心理学的に厳密な手順で構成した5種類の評定尺度を用いている点も特徴である（籠宮他2003）。データベースの全体にわたる評価が必要な場合は単独評定データを，対象とする講演は少なくとも信頼性の高い評価値を利用したい場合は集合評定データを，それぞれ利用することができる。

印象評定データは，社会言語学などの研究において必要とされる発話スタイルの外的指標となるほかに，講演が聴き手に与える印象そのものの研究に利用できる。籠宮他(2003)は，集合評定データで把握された講演の巧拙の印象と発話速度との関係を分析して，両者の関係が線形でないことを報告している。

4.5. 分節音情報とイントネーション情報

我々は多くの場合，ただ音声を聞くだけで朗読音声と自発音声を区別することができる。つまり両者間には何らかの音声学上ないし言語学上の差が存在していると考えられる。また印象評定で「単調な」と評定される音声と「表情豊かな」と評定される音声の間にも当然何らかの音声学な差異があるものと予想される。

こうした差異を客観的に検討するためには，音声自体の検討が必要になる。そのために，CSJではコアに含まれる音声に対して分節音（子音や母音）のラベルとイントネーション（声の高さの時間変化）のラベルを提供した。これらは話し言葉の本質に最も直接的にかかわる情報と言ってよい。特にイントネーションについては自発音声の多様性が顕著に表れることが予想されたので，従来のラベリング手法であるJ_ToBI（Venditti 1997）を大幅に拡張したX-JToBI（前川他2001；Maekawa et al. 2002）を新たに考案して作業に臨んだ。朗読音声に分節音や韻律のラベルを付与することは，従来から行なわれてきており，また自発音声のラベリングも試験的には世界各地で試みられてきている（Nat. Inst. Jap. Lang. 2004参照）。しかし44時間（50万短単位）というまとまった量の自発音声をラベリングしたのは世界で初めての試みである。

図7にX-JToBIによるラベリングの例を示す。図上部に音声信号と音声基本周波数（Fo）曲線が表示されており，その時間軸に同期させて各種ラベルがそれぞれ別の窓に表示されている。上から順に「分節音層」（子音や母音のラベル），「単語層」（短単位の音素表記），「トーン層」（イントネーションの構成要素としての音韻論的tone），「Break index（BI）層」（発話の韻律境界の深さを示す指標），「プロミネンス層」（トーン層の解釈を補助するための情報）が表示されている。プロミネンス層の右端に表示されているラベル“FR”はいわゆる「浮き上がり調」（川上1963）の上昇イントネーションを示している。イントネーションラベルの仕様については

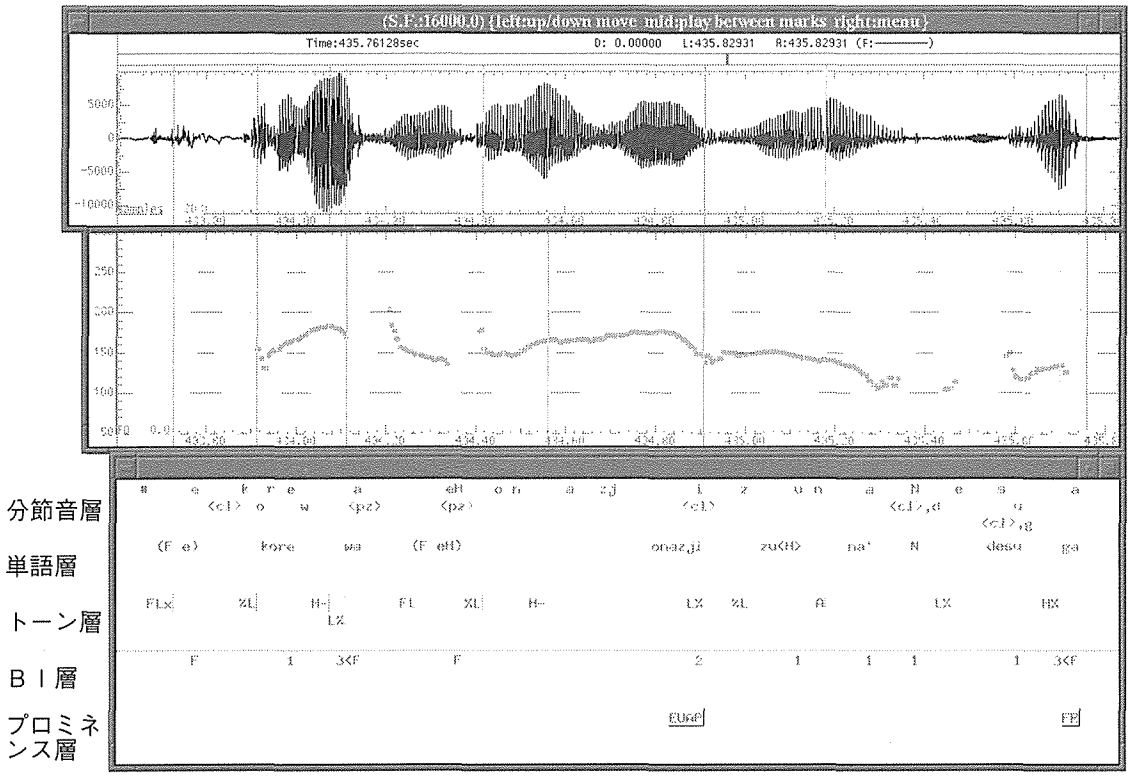


図 7　CSJ の分節音ラベル・イントネーションラベルの例
（発話は「え，これは，えー，同じ図なんですが」）

CSJ に同梱されているマニュアル—Disk1/DOC の intonation.pdf—ないし Maekawa et al. (2002)，前川(2004)参照[9]。

4.6. その他の付加情報

紙幅の関係でその他の研究用付加情報にはごく簡単に言及するにとどめる。これらの情報付与作業は上記の節単位情報とともに情報通信研究機構で実施された。

4.6.1. 係り受け構造情報

係り受け構造情報は，4.3で紹介した節単位を領域として，その内部での文節間の修飾関係を示した情報である。話し言葉の文法研究だけでなく，統語構造とイントネーションの関係の研究などにも利用価値の認められる情報である。係り受け構造情報はコアに対してだけ提供される。

CSJ の係り受け構造分析の仕様は，新聞の書き言葉を対象とした「京大コーパス」の仕様（黒橋・長尾1997）を話し言葉用に拡張した仕様で実施されている。拡張仕様は，話し言葉に頻出する，言いさし・言い直し・倒置・係りのねじれ等の現象に関するものである。以下に言い直しと倒置に関する例を示す（内元他2003)。ゴチックが注目する要素である。倒置では，例外と

して右から左への係り受けが許容されている。

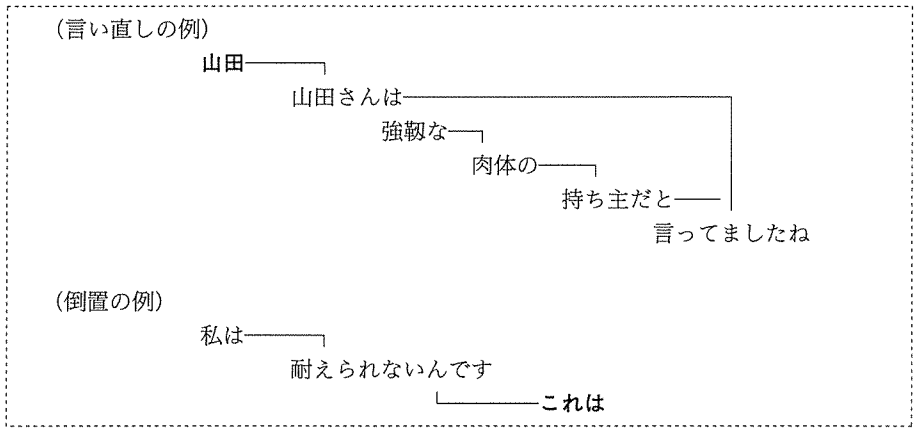


図8　係り受け構造情報の例

4.6.2. 要約・重要文情報

重要文とは，講演を要約する目的で抽出された転記テキスト中の重要部分のことである。テキストの要約は自然言語処理の重要な研究対象である。また話し言葉の自動音声認識研究でも，認識結果をそのまま出力するのではなく，言語情報の伝達には無駄な部分を省略して要約したテキストを出力することが多い。こうした研究のためには，人間が与えられた転記テキストをどのように要約するかの情報が必要である。

CSJ の要約・重要文情報作成作業では，作業者に50％と10％の 2 種類の基準で要約を作成させた。50％の要約率を指定された作業者は，与えられた転記テキストの分量がちょうど半分になるように転記テキストを取捨選択する。取捨選択の単位としては4.3で説明した「節」を利用している（野畑他2004)。

また，上記の手法とは別に，転記テキストを自由に要約した自由要約データも作成しており，これも CSJ の一部として公開する。要約・重要文情報はコアに対して提供される情報である。

4.6.3. 談話構造情報

談話構造情報は，談話（例えばひとつの学会講演や模擬講演）内部における話題の階層構造を示す情報である。談話構造の表示方法には様々な流儀があるが，CSJ では Grosz と Sidner が提唱した「意図」に基づく談話構造理論に依拠した分析をおこなっている（Grosz and Sidner 1986；竹内他2003)。簡単に言えば，話し手が或る発話をおこなった際に保有していた意図（何故そのような発話をおこなったか）を推測し，それによって談話を分割し，分割された単位間の階層構造を決定する作業である。図 9 に談話構造タグの実例を示す。WHY の後にゴチックで示されているのが認定された「意図」である。本例では最初の意図に属する談話区分が更にふたつの意図に下位区分されており，そのうち前半が更にふたつに下位区分されている。分析されてい

る講演は音声学に関するものである。分析対象のテキストは転記テキストの基本形を変形したものであり，タグ（M）はメタ言語的表現を囲っている。

談話構造情報は，いわゆる談話研究に欠くことのできない情報であり，自然言語処理にも重要な情報であるが，自発性の高い独話への情報付与はかなり難しいので，コアの一部に対してだけ作業を実施した。

WHY? 実験の結果の説明
で結果ですが
WHY? 4つの「あ」を混合した結果の説明
まずこれを見てこれはお手許にある図と同じでございます
これは何を表わしているかと言いますとその（M ささだが）という発話に含まれる四つの（M あ）それを十回繰り返したもの全てですね
のこっちの左側がホルマント周波数の分布 右側がT3中舌面のコイルの位置であります
で軸を変換いたしましていわゆる母音四角形のように読めるように表示しております
WHY? ホルマント周波数との相関の説明
でホルマント周波数F2を見ていただきますとS
Sというのはサスピションで疑いですが疑いの場合はF2が高い
それからAがアドミレーションで感心なんですがその場合は低いという関係がはっきり見て取れます
WHY? 調音運動との相関の説明
そして同じように今度は調音運動の方を見ますと
Sにおいては T3X つまり前後方向の値が小さいということは前寄り
それからAにおいては T3X が大きいということはより後ろ寄りという関係が見て取れます
NDに関しましては中立および落胆に関してはその中間に分布するという結果が出ております
WHY? それぞれの「あ」の個別の結果
で今のは四つのモーラ（M ささだが）の全ての（M あ）を
プールした結果でありますがそれぞれ個々のモーラに分離いたしますとこういう結果が出ます

図9 談話構造情報の例（竹内他2003より引用）

4.7. XML文書

以上の説明からわかるようにCSJには豊富な研究用情報が含まれている。これらの情報を相互参照することによって，話し言葉研究に新たな展開が期待されるのであるが，研究用情報が豊富になればなるほど，それらを統合して検索することが困難になってくる。

この問題を回避するためには，種々の情報を階層化して統合的に表現すればよい。近年普及しはじめたXMLは，この目的によく適ったマークアップ言語である。CSJに含まれる研究用付加情報は，単独のファイルとして提供されるとともに単一のXML文書に統合された形でも提供される（付加情報のなかにはXML文書としてのみ提供されるものもある）。

話し言葉のデータでは，階層構造に破綻が生じることが稀ではないので（例えば節の内部に200ms以上のポーズが生じると，文法的には単一の節がふたつの転記基本単位に分割されてしまう），階層化は簡単ではない。しかし，情報検索のためだけでなく巨大なデータベースを論理的に一貫した方法で管理してゆくためにもデータの階層化は必要不可欠である。CSJのXML文書化については，菊池他(2004)，塚原他(2004)，Maekawa et al.(2004)参照。

5．CSJの公開

以上，本稿では『日本語話し言葉コーパス』公開版の仕様を概観した。5年間にわたったCSJの開発は2004年3月に完了し，近日中の一般公開を予定している。国立国語研究所のホームページでは，サンプル音声や予備的分析の結果も含めて，既にCSJの情報を提供してきているが（http://www.kokken.go.jp），一般公開に関する情報もホームページで提供する予定である[10]。

またCSJの構築過程で蓄積してきた各種作業マニュアルは現在700ページ以上に達している。これらのマニュアルは日本のコーパス言語学にとって貴重な財産であるので，国語研究所の報告書その他の形で順次公開してゆく予定である。またCSJ公開版には270ページ程度の解説文書類を同梱する。

6．今後の展望

我々は過去5年間にわたってCSJの構築に全力を注いできた。今後はCSJを言語研究や音声情報処理研究のみならず幅広い研究領域で有効活用してゆくことが重要な課題になる。これまでに実施した予備的解析では，社会言語学（前川2002a，2002b；Maekawa et al. 2003），心理学（槙・前川2001；山住・籠宮・前川2003），音声談話研究（Yoneyama, et al. 2003；小磯2003）などにおける有効性を示してきた。しかし，これが利用可能な領域のすべてではあるまい。2001年と2002年の2回にわたって実施したCSJのモニター公開に対しては，合計で300件を超える試用申込みをいただいたが，希望者の専門は，音声情報処理，自然言語処理，言語学，日本語教育学，心理学，社会学などの領域に広がっていた。これらの領域でCSJが幅広く活用されてゆくことを期待している。

筆者個人としてはいわゆるコーパス言語学的な専門的言語研究とならんで，辞書編纂など応用面での可能性も中長期的な課題として追求したいと考えている。例えば，中期的な課題として発音辞典への応用が考えられる。現在刊行されている日本語の発音・アクセント辞典類では，発音のゆれの存在は記述されていてもその社会言語学的実態は報告されていない。この問題は，CSJ転記テキストの発音形やコアの分節音・イントネーションラベルを解析することによって，或る程度まで解決することができるだろう。長期的課題としては，書き言葉データを含む現代日本語の総合的なデータベースを解析して，話し言葉と書き言葉の双方におよぶ日本語のコロケーション辞書を開発することなどが考えられる。

ここで指摘しておく必要があるのは，今後どのような目的で利用されるにせよ，CSJのよう

な言語データベースの構築作業は一回実施すればそれで完了してしまう性質のものではないことである。言語には堅固な構造が備わっていると同時に，時代とともに変化してゆく側面がある。これは話し言葉も書き言葉も同様であり，音声や言語に関わる情報処理技術はその影響を免れることができない。そのため，一定の時間間隔で日本語の変化を組織的かつ正確に記録しておくことが必要になる。

本稿を終えるにあたり，このようにして構築されるデータベースには情報処理技術上の価値だけでなく，広く国民の財産としての価値が認められることを強調しておきたい。我々が江戸時代や平安時代の文書に文化財としての価値を認めるように，今日の日本語は未来の日本人にとって貴重な文化財となることは間違いない。CSJ のように多量の付加情報を伴った記録であれば，その価値は倍増する。言語データベースの構築には未来の文化財を創成するという大きな付加価値が存しているのである[11]。

謝辞：『日本語話し言葉コーパス』に音声を提供してくださった話者の方々ならびに関係諸学会に心より感謝いたします。

注

1　科研費による研究として野元(1980)，石井(1983)もあった。

2　http://www.ciair.coe.nagoya-u.ac.jp/jpn/db/dbciair/speech_corpus.htm に日本語に関する音声データベースの概観が掲載されている。

3　BNC の spoken part のために録音された音声は大英図書館(British Library)で聴取することができる。

4　実際の CSJ には図 1 に示されているように752万短単位が格納されている。これは形態論的な解析が終了するまでは語数を正確に推定することができないため，音声を多めに収録した結果である。

5　ただし実際には原稿を朗読したに等しい話者も若干名含まれている。

6　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。

7　transcription.pdf に記述された内容が最新の仕様であり，他の文献とは異同がある。

8　厳密に言えば，手作業と自動解析の品詞体系は細部（用言の活用形の下位区分）において異なっている。これは自動解析に必要とされる言語素性を解析時に追加したために生じた相違である。

9　intonation.pdf に記述された仕様が最終版であり，他の文献の内容とは異同がある。

10　本稿の刊行と前後して一般公開を開始する予定である。

11　いわゆる日本語の乱れについて有益な議論をおこなうためにも，いま眼前に広がっている日本語の多様性を的確に把握しておかねばならない。大規模な言語データベースの構築と解析は，この目的を達するための，ほとんど唯一の有効手段である。

文　献

石井久雄(1983).『形態結合における音融合』(昭和58年度科研費奨励研究(A)報告書).

内元清貴・丸山岳彦・高梨克也・井佐原均(2003).「『日本語話し言葉コーパス』における係り受け構造付与」『平成15年度国立国語研究所公開研究発表会予稿集』pp.35-36.

小椋秀樹・山口昌也・西川賢哉・石塚京子・木村睦子(2004).「『日本語話し言葉コーパス』の形態論情報の概要」(『日本語話し言葉コーパス』公開版添付文書).

籠宮隆之・山住賢司・槙洋一・前川喜久雄(2003).「講演音声に対する評定尺度の作成」『第17回日本音声学会全国大会予稿集』pp.135-140.

川上蓁(1963).「文末などの上昇調」『国語研究』16, pp.25-46.

河原達也(2004).「『日本語話し言葉コーパス』を用いた音声認識の進展」『第3回話し言葉の科学と工学ワークショップ講演予稿集』pp.61-66.

菊池英明・塚原渉・前川喜久雄(2004).「XMLを利用した『日本語話し言葉コーパス』(CSJ)の整合性検証」『第3回話し言葉の科学と工学ワークショップ講演予稿集』pp.27-32.

黒橋禎夫・長尾眞(1997).「京都大学テキストコーパスプロジェクト」『言語処理学界第3回年次大会発表論文集』pp.115-118.

小磯花絵(2003).「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.

小磯花絵・土屋菜穂子・間淵洋子・斉藤美紀・籠宮隆之・菊地英明・前川喜久雄(2001).「『日本語話し言葉コーパス』における書き起こしの方法とその基準について」『日本語科学』, 9, pp.43-58.

小磯花絵・斎藤美紀・間淵洋子・前川喜久雄(2002).「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.

国立国語研究所(1955).『談話語の実態』秀英出版.

国立国語研究所(1960).『話し言葉の文型(1)』秀英出版.

国立国語研究所(1963).『話し言葉の文型(2)』秀英出版

高梨克也・内元清貴・丸山岳彦・井佐原均(2003).「『日本語話し言葉コーパス』における節境界認定」『平成15年度国立国語研究所公開研究発表会予稿集』pp.45-46.

竹内和広・森本郁代・高梨克也・小磯花絵・井佐原均(2003).「『日本語話し言葉コーパス』における談話構造タグの仕様」『平成15年度国立国語研究所公開研究発表会予稿集』pp.37-38.

塚原渉・菊池英明・前川喜久雄(2004).「『日本語話し言葉コーパス』のXML検索環」『第3回話し言葉の科学と工学ワークショップ講演予稿集』pp.33-38.

堤怜介・加藤正治・小阪哲夫・好田正紀(2004).「発音変形依存と教師なし適応による講演音声認識の性能改善」『第3回話し言葉の科学と工学ワークショップ講演予稿集』pp.93-98.

野畑周・高梨克也・内元清貴・井佐原均(2004).「『日本語話し言葉コーパス』における要約データの作成」『第3回話し言葉の科学と工学ワークショップ講演予稿集』pp.99-104.

野元菊雄(1980).『日本人の知識層における話しことばの実態』(科研費特定研究「日本語教育のための言語能力の測定」報告書).

古井貞熙・前川喜久雄・井佐原均(2000).「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.

前川喜久雄(2000).「パラ言語的情報」『別冊国文学「現代日本語必携」』53, pp.172-175.

前川喜久雄(2001).「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.

前川喜久雄(2002a).「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.

前川喜久雄(2002b).「『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.

前川喜久雄(2004).「『日本語話し言葉コーパス』の韻律アノテーション」『韻律に着目した音声言語情報処理の高度化2003年度研究成果報告書』pp.1-4 (東京大学新領域創成科学研究科).

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

前川喜久雄・菊池英明・五十嵐陽介(2001).「X-JToBI:自発音声の韻律ラベリングスキーム」『電子情報通信学会技術報告』(SP2001-106, NLC2001-71), pp.25-30.

槙洋一・前川喜久雄(2001).「自伝的な出来事の想起に関する世代差」『日本認知科学会第18回大会発表論文集』, pp.96-97.

丸山岳彦・柏岡秀紀・熊野正・田中英輝(2003).「節境界自動検出ルールの作成と評価」『言語処理学界第9回年次大会発表論文集』pp.517-520.

山口昌也(2003).「『日本語話し言葉コーパス』における品詞分布の分析」『平成15年度国立国語研究所公開研究発表会予稿集』pp.45-46.

山住賢司・籠宮隆之・前川喜久雄(2003).「講演音声の特徴を捉える評価尺度の構築」『日本音響学会2003年秋季研究発表会講演論文集』pp.371-372.

Aston, G. and L. Burnard (1998). *The BNC Handbook: Exploring the British National Corpus with Sara.* Edinburgh Univ ersity Press.

Grosz, B. and C. Sidner (1986). "Attention, intention, and the structure of discourse," *Computational Linguistics*, 12 (3), pp.175-204.

Maekawa, K., H. Koiso, S. Furui, and H. Isahara (2000). "Spontaneous speech corpus of Japanese," *Proceedings of the Second International Conference of Language Resources and Evaluation (LREC2000)*, Athens, 2, pp.947-952.

Maekawa, K., H. Kikuchi, Y. Igarashi, and J. Venditti (2002). "X-JToBI: An extended J_ToBI for spontaneous speech," *Proceedings of the 7th International Conference on Spoken Language Processing* (ICSLP2002). Denver, pp.1545-1548.

Maekawa, K., H. Koiso, H. Kikuchi, and K. Yoneyama (2003). "Use of a large-scale spontaneous speech corpus in the study of linguistic variation," *Proceedings of the 15th International Congress of Phonetic Sciences* (ICPhS2003), Barcelona, pp.643-646.

Maekawa, K., H. Kikuchi, and W. Tsukahara (2004). "Corpus of Spontaneous Japanese: Design, Annotation, and XML Representation," *Proceedings of the International Symposium on Large-scale Knowledge Resources* (LKR2004), pp.19-24 (Tokyo Institute of Technology 21st Century COE Program).

National Institute for Japanese Language (2004). *Spontaneous Speech: Data and Analysis (Proceedings of the 1st session of the 10th international symposium).*

Svartvik, J. and R. Quirk (1980). *A Corpus of English Conversation.* LiberLäochromedel, Lund.

Venditti, J. (1997). "Japanese ToBI Labeling Guidelines." *OSU Working Papers in Linguistics,*

50, pp.127-162, (http://www.ling.ohio-state.edu/phonetics/J_ToBI/).
Yoneyama, K. J. Fon, and H. Koiso (2003). "Durational and prosodic patterning at discourse boundaries in Japanese spontaneous monologs," *Proceedings of the 15th International Congress of Phonetic Sciences* (ICPhS2003). Barcelona, pp.2637-2640.

追記：CSJ を用いた工学的研究成果は下記の文献に多数収録されている。*Proceedings of ISCA& IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003),* Tokyo, 2003.

前川　喜久雄
国立国語研究所　研究開発部門第二領域
115-8620　東京都北区西が丘3—9—14
kikuo@kokken.go.jp