『日本語話し言葉コーパス』の設計と実装

『日本語話し言葉コーパス』の設計と実装

前川 喜久雄

独立行政法人国立国語研究所研究開発部門第2領域

〒115-8620 東京都北区西が丘3-9-14

E-mail: kikuo@kokken.go.jp

あらまし 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。コーパスの構築に関わる技術的な問題は本研究会の他の発表で詳しく紹介される予定であるので、本稿では以下の発表への導入を兼ねて、何故『日本語話し言葉コーパス』が必要とされるのか、そのためにどのような設計がおこなわれたのかに焦点をあてたデータベース全体の概説をおこなう。

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、設計

Design and Compilation of the *Corpus of Spontaneous Japanese*

Kikuo Maekawa

Department of Language Research, National Institute for Japanese Language

3-9-14, Nishiga'oka, Kita-ku, Tokyo 115-8620

E-mail: kikuo@kokken.go.jp

Abstract This paper introduces the *Corpus of Spontaneous Japanese*, a large-scale corpus of spontaneous Japanese. Since details of techniques used in the compilation will be reported in different papers of this workshop, I will rather concentrate on the overall description of the corpus, with special emphasis upon the basic aims and design issues of the corpus.

Keyword Corpus of Spontaneous Japanese, database, design

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。一方、話し言葉では録音された音声を文字に転記する手間が大変である。また、ただ単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、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もわからなくなる可能性がある。

さらに、言い誤りや言い淀みのような現象も転記が必要である。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。このようにして、話し言葉の本格的な研究は書き言葉に較べて立ち遅れてしまうのである。

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』[1]と『話しことばの文型』[2,3]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色をはなっており、現在でも引用されることが多い。しかし 1963 年を最後にこの種の調査研究は中断されてしまった。

本日紹介する『日本語話し言葉コーパス』は、この話し言葉調査の系譜に連なるデータベースである。その目標は、国語研究所における話し言葉研究の伝統を復活させると同時に、データベース自体を一般公開することによって情報処理も含めた現代日本語の話し言葉研究のインフラストラクチャを整備することにある。

『日本語話し言葉コーパス』(英名は Corpus of Spontaneous Japanese; 以下これを省略して CSJ と呼ぶ) は、国立国語研究所、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年度である[4]。

CSJには時間にして約660時間、語数では700万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。1999年以来継続されてきたCSJの開発作業は本2003年度で終了し、来年度には一般公開を予定している。

本稿の目的はCSJの紹介であるが、CSJに付加された研究用情報の詳細は、井佐原均氏と菊池英明氏の講演ならびにポスター発表の各論文に詳しいので、以下ではこれらの発表への導入としてCSJの設計と実装方法を概観する。またCSJを利用した音声認識・要約研究の成果については古井教授の講演と、南条浩輝氏ら、篠崎隆宏氏らによるポスター発表を参照していただきたい。

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。それらは以下のように大別することができる。

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[5]。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。音声信号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（現在はLondon-Lund Corpusの名で知られている）がその嚆矢となった[6]。

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、非常に付加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い。もちろん音声情報処理に利用することもできない。

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、1988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えはじめた当初からの方針である。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[7-9]。

2.2. データ量

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。

最初に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から700万語（短単位；3.4参照）と見積もった。これがCSJ全体のサイズである。この700万語に対しては、音声信号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。

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なデータ量を見積もることによって決定した。表1にCSJ全体とコアにおける研究用情報の相違をまとめた。

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、そこには様々なバリエーションがある。まず問題となるのが、独話と対話の別であるが、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が存在していないことを考えると、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、さらに多少ともまとまった内容の発話を対象に据えることにした。これは、やはり工学的応用として、誤りを含む音声認識結果を処理して簡潔なテキストにまとめる音声要約技術の研究をプロジェクトの目標のひとつに掲げていたことによる選択である。なお、一般に独話は対話よりも上記の性格に適うことが多いことは指摘するまでもないだろう。

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」という作業上の規定を採用し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収録の対象となっている。ただし、コアには韻律特徴のラベルを付与する関係上、韻律特徴が東京式の講演を選別して格納している。

表1：CSJが提供する研究用情報

CSJ全体
- 音声信号(サンプリング周波数16kHz, 量子化16bit)
- 転記テキスト（基本形と発音形。3.3節参照）
- 形態論情報（短単位と長単位の二種類。3.4節）
- 講演情報（講演のタイプ、模擬講演のテーマとタイトル等。3.1節）
- 話者情報（生年代、性別、出生地、両親の出生地、居住歴、学歴等）
- 印象評定値（単独評定）（3.2節）

コア全体　上記に加えて
- 分節音ラベル　（3.5節）
- 韻律ラベル　（3.5節）
- 節境界情報*　（3.6節）
- 重要文情報*　（3.7節）
- 印象評定値（集合評定）*（3.2節参照）

コアの一部　上記すべてにくわえて
- 談話境界情報　（3.7節参照）
- 係り受け情報　（3.7節参照）

*対話と再朗読の音声は対象外

3. 実装

3.1. 音声収録

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の話者は延べ1000名を越している。

しかし、学会講演の話者には強い偏りがあることに注意する必要がある。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。さらに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。つまり、学会講演を現代日本語の代表とみなすには問題があると考えられる。

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。これらの偏りはCSJを用いて社会言語学的な研究を実施しようとする場合に好ましくない影響をおよぼす。

こ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首都圏出身の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三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2に指定したテーマのリストを示す。

表2：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

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。模擬講演の話者数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 1700 件以上に達する。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。収録に先立って収録スタッフと一定時間雑談する、講演中には収録スタッフができるだけ相槌を返すなどの工夫である。これらの工夫の効果を測定することはできないが、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、模擬講演のスタイルは、明らかに学会講演よりも低下していることがわかる[9]。学会講演と模擬講演のスタイル差については前川らのポスターを参照していただきたい。

表３に CSJ に収録された音声の内訳を示す。CSJ の 95%は学会講演と模擬講演であるが、それ以外に約 32 時間の音声が収録されており、うち約 12 時間は種々の対話音声である。また朗読音声（新書から抜粋した自然科学に関するテキスト二種類を模擬講演話者が朗読したもの）と再朗読音声（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もの）も合計 20 時間収録されている。

これらは、独話音声と比較対照して 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、同じ話者が学会講演、模擬講演、４種類の対話、再朗読をおこなったデータも提供されているので、発話状況の差が音声に及ぼす影響を同一話者において幅広く比較することもできる。

CSJ の 5%に過ぎないとはいえ、これらの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、独話と切り離して分析することもできるだろう。

表３：CSJ に格納された音声の内訳

音声の種類	話者数	講演数	独話／対話の別	自発／朗読の別	時間数
学会講演	838	1007	独話	自発	299.5
模擬講演	580	1699	独話	自発	324.1
朗読音声	*(244)	491	独話	朗読	14.1
インタビュー話者による模擬講演	*(16)	16	独話	自発	3.4
学会講演に関するインタビュー	*(10)	10	対話	自発	2.1
模擬講演に関するインタビュー	*(16)	16	対話	自発	3.4
課題指向対話	*(16)	16	対話	自発	3.1
自由対話	*(16)	16	対話	自発	3.6
再朗読	*(16)	16	独話	朗読	5.5
				総時間数	658.8

*( )内の話者は独話話者の一部

3.2. 印象評定値

CSJ には種々様々な自発音声が収録される。それらが聴き手に与える印象もまた一様でない。印象評定値とは、講演音声が聴き手に与える印象を主観的に評定したデータである。印象評定値には二種類がある。ひとつは音声収録の現場で収録スタッフ１名が調査票に記入したデータ[10]、もうひとつは収録が終了した後に、コアの独話を 20 名の評定者が評定したデータである[11]。これらをそれぞれ単独評定データ、集合評定データと呼ぶことにする。

単独評定データは時間の制約から簡単な方法で記録した。ひとつは評定シートに記入された 31 種の評価語（たどたどしい、流暢な、単調な、表情豊かな、等）のうち該当すると思われるものにマルをつける方法、もうひとつは「講演の自発性」「発話スピード」「発話スタイル」「発音の明瞭さ」等を五段階尺度で評定する方法である。

一方、集合評定データでは実験心理学的に厳密な手続きに従った尺度構成をおこなった。これについては籠宮隆之氏らのポスター発表を参照。

印象評定値はスピーチの巧拙など、独話の印象がどのように形成されるかを客観的に検討するために作成したデータであるが、その他に発話スタイルの指標として利用することも想定している。先に模擬講演の発話スタイルが学会講演よりも低いと述べたが、これは統計的な事実であって個々の講演のスタイルを保障するものではない。実際、非常にくだけた学会講演もあれば堅苦しい模擬講演もある。印象評定値のうちスタ

イルに関係する部分を利用すれば、個々の講演をスタイルに関して順序づけることが可能になる。このような情報は言語変異現象の分析などにおいては非常に有益である[24,25]。

3.3. 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。書き起こし作業については小磯花絵氏らのポスターに詳しいが、ここでは、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ことを強調しておきたい。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。

CSJ の転記テキストには、漢字仮名まじりで表記された基本形と片仮名だけで表記された発音形の二種類が提供される。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[12]。

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。

転記テキストには上記のほかにも多くのタグ記号が挿入されている。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(?)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。

なお CSJ の転記テキストは長めのポーズ（原則として 0.2 秒以上）位置で転記基本単位に分割されている。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ので、これによって転記基本単位ごとの発話速度を計算することができる。このように転記テキストだけを用いて実施できる研究も少なくない。

3.4. 形態論情報

形態論情報とは既に述べたように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、語をどう規定するかによって結果が異なってくることは当然である。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り、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二種類を採用して二重の形態論情報を提供している[13]。これらをそれぞれ短単位、長単位と呼ぶ。一例を示せば「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。

これらの単位の設計については小椋秀樹氏らのポスター発表に詳しいが、二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、二重の分析が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。このデータは通信総合研究所に渡されて、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 600 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（若干の手修正も加えている）。自動解析の詳細は井佐原氏の講演に詳しい。

ちなみに手作業による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ので、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に差がある。CSJ の活用にあたって注意が必要となろう。

3.5. 分節音情報と韻律情報

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転記テキストの分析だけでは不十分であり、音声信号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を対象として分節音（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）のラベルを提供している。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。

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。しかし50万語（約44時間）というまとまった量の自発音声のラベリングは世界で初めての試みである。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(J_ToBI[14])を大幅に拡張した X-JToBI [15-18]を新たに考案して作業に臨んだのであるが、作業の進展にともなって当初予期していなかった韻律現象も多い。自発音声の多様性を改めて認識させられた。

分節音や韻律特徴に関する予備的分析は、あまり進展していないが、一部のデータについてアクセント句末に生じるイントネーションを比較したところ、学会講演と模擬講演とで用いられるイントネーションのタイプに顕著な差異が生じていた。今後、多くの発見が可能であろうと期待している。CSJ の音声ラベリングについては菊池氏らのポスターに詳しい。

3.6. 節境界情報

独話においては、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。このような節境界の情報は、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思われる。

CSJ のコアには、転記テキストを解析して節境界の位置を検出した情報が提供される。この情報は、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 による解析結果をもとに、通信総合研究所で人手修正されたものである[19]。コア以外についても、自動検出結果を提供する方向で検討を進めている。詳細は高梨克也氏ら、丸山岳彦氏らのポスター参照。

3.7. 係り受け構造情報・重要文・談話境界情報

係り受け構造情報は、前節で紹介した節を領域として、その内部での文節間の修飾関係を示した情報であり、発話の統語構造に関係する。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広く利用価値の認められる情報である。係り受け情報付与作業は通信総合研究所で実施されており、コアの一部に対して提供される予定である（内元清貴氏らのポスター参照）。

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。例えば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、転記テキスト中の単位を選択する。その際、選択の単位としては上述の節を利用する。なお、上記の方法によって抽出した重要文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、これも公開する予定である。

重要文は、音声認識に基づく自動要約結果を人手による重要文抽出結果と比較して、その精度を評価するために利用するが、その他に、自然言語処理の研究でも利用でき、また、人間による要約作業そのものの研究資料にもなると思われる。重要文もコアに対して提供される情報である。

談話境界情報は、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[20]。いわゆる談話研究に書くことのできない情報であるが、独話への情報付与はかなり難しく、コアの一部に対してだけ提供する予定である。詳細は竹内和広氏らのポスター参照。

3.8. XML 表現

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、話し言葉に関する新事実が数多くもたらされると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。

例えば、アクセント句末に位置する終助詞のイントネーション形状が、アクセント句が有核であるか、節の末尾に位置しているかによってどのように変動するかを検討したいとしよう。この場合、少なくとも、節境界の有無、アクセント核の有無（韻律ラベルのうち単語層と呼ばれる層に属する情報）、短単位の品詞、そしてイントネーションの形状を表すラベル（韻律ラベルのうちトーン層と呼ばれる層に属する情報）を統合的に検索しなければならない。

このような検索を可能にするひとつの方法は、種々の情報を階層化して表現することである。現在、我々は CSJ の研究用情報を XML と呼ばれるマークアップ言語によって階層的に表現することを試みている。

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）、困難をともなう作業であるのだが、データの階層化は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必要不可欠であると考えている。この問題については菊池氏が講演で触れる予定である。

4. CSJ の公開

以上『日本語話し言葉コーパス』の設計と実装を概観した。CSJ の構築作業は現在最終段階にあり、現在は来春の公開をめざした作業が続いている。データの総量はまだ最終的に確定していないが、DVD-ROM で 10 枚以上になる予定である。無償とはゆかないが、できるだけ多くの人に利用していただける頒価で提供したい。

CSJ に関する情報は、サンプル音声や予備的分析の結果も含めて、国語研究所のホームページに記載している(http://www2.kokken.go.jp/~csj/public/index.html)。公開に関する情報もホームページ等で順次お知らせする予定である。

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。この情報も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。なお CSJ の公開版には簡潔なユーザーズマニュアル類を同梱する。

5. 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、社会言語学研究[21-25]、心理学研究[26-28]、談話研究[29,30]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 年の二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 350 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。

私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長期的な課題として追求したいと考えている。例えば CSJ を含む現代日本語の大規模なデータベースを解析して日本語学習者用のコロケーション辞書を開発するなどすれば、斯界への貢献は絶大なものがあるだろうと想像する。

さて、本稿を終えるにあたり、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を指摘しておきたい。

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遷過程を組織的かつ正確に記録しておくことが必要になる。

ここで指摘しておきたいことは、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である。我々が江戸時代やそれ以前の文書に文化財としての価値を認めるように、今日の日本語は百年後二百年後の日本人にとってきわめて貴重な文化財となるに違いない。まして CSJ のように音声を伴った記録であれば、その価値は倍増するであろう。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付加価値が存することは、もっと広く認識されるべきだと思う。

最後に、国民レベルで現代日本語について議論するためにも日本語の現状を正確に把握しておかねばならいことは当然である。近年、日本語の正しさ、美しさについての議論や漢字の使用制限の改修に関する議論が盛んであるが、この種の議論を真に有益なものとするためには、研ぎ澄まされた言語感覚に基づく推察の基礎として、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。

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。

文　献

[1] 国立国語研究所,『談話語の実態』,秀英出版, 1955.
[2] 国立国語研究所,『話し言葉の文型(1)』,秀英出版, 1960.
[3] 国立国語研究所,『話し言葉の文型(2)』,秀英出版, 1963.
[4] 古井 他「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, 2001.
[5] http://www.ciair.coe.nagoya-u.ac.jp/jpn/db/dbciair/speech_corpus.htm に日本語の音声データベースが概観されている。
[6] Svartvik, J. & R. Quirk. *A Corpus of English Conversation,* LiberLäromedel, Lund, 1980.
[7] Maekawa, K. et al., "Spontaneous speech corpus of Japanese," *Proceedings of the Second International Conference of Language Resources and Evaluation (LREC2000),* Athens, 2, pp.947-952, 2000.
[8] 前川 他「日本語話し言葉コーパスの設計」音声

研究, 4:2, pp.51-61, 2000.

[9] 前川「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.

[10] 籠宮 他「自発音声コーパスにおける印象評定とその要因」日本音響学会 2001 年秋季研究発表会講演論文集, 1, pp.381-382, 2001.

[11] 籠宮 他「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, 2003.

[12] 小磯 他「『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, 2001.

[13] 小椋「話し言葉コーパスの単位認定基準について」話し言葉の科学と工学ワークショップ講演予稿集, pp.21-28, 2001.

[14] Venditti, J. "Japanese ToBI Labeling Guidelines." *OSU Working Papers in Linguistics*, 50, pp.127-162, 1997(http://www.ling.ohio-state.edu/phonetics/J_ToBI/).

[15] 菊池 他「自発音声に対する J_ToBI ラベリングの問題点検討」日本音響学会 2001 年春季研究発表会講演論文集 1, pp.383-384, 2001.

[16] 前川 他「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[NLC2001-71, SP2001-106], pp.25-30, 2001.

[17] Maekawa, K. et al., X-JToBI: An extended J_ToBI for spontaneous speech, *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP2002)*, Denver, 3, pp.1545-1548, 2002.

[18] Kikuchi, H. et al., "Evaluation of the effectiveness of 'X-JToBI': A new prosodic labeling scheme for spontaneous Japanese speech," *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, pp. 579-582, 2003.

[19] Takanashi, K., et al., "Identification of 'sentence' in spontaneous Japanese –Detection and modification of caluse boundaries," *Proceedings of ISCA&IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, Tokyo, pp. 183-186, 2003.

[20] Takeuchi, K., et al., "Committee-based discourse purpose assignment: Discourse structure annotations of spontaneous Japanese monologue," *Proceedings of ISCA&IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, Tokyo, pp. 199-202, 2003.

[21] 前川「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.

[22] 斎藤 他「「ギジツ」と「ギジュツ」:『日本語話し言葉コーパス』に基づく直音化現象の分析」第 10 回社会言語科学会研究大会予稿集, pp.209-214, 2002.

[23] 小磯 他「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.

[24] 前川「『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.

[25] Maekawa, K., et al. "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, 2003.

[26] 槙・前川「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, 2001.

[27] 山住 他「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, 1, pp.371-372, 2003.

[28] 天野 他「言語心理学の新展開：大規模データベースの構築と利用」日本心理学会第 67 回大会発表論文集,S34, 2003.

[29] Yoneyama, K. et al., "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, 2003.

[30] 小磯 他「『日本語話し言葉コーパス』を用いた談話構造と韻律との関係に関する一考察」人工知能学会言語・音声理解と対話処理研究会 (SIG-SLUD-A203), 139-144, 2003.

追記：CSJ を用いた工学的研究成果は下記の論文集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.