モノローグを対象とした自発音声コーパス －その設計について－

前川喜久雄，籠宮隆之，小磯花絵，菊池英明，小椋秀樹

（国立国語研究所）

{kikuo/koiso/kagomiya/kikuchi/ogura}@kokken.go.jp

1. はじめに

国立国語研究所，郵政省通信総合研究所，東京工業大学の３機関は，科学技術振興調整費開放的融合研究制度の研究資金を得て，自発性の高い話し言葉処理技術の開発に取り組んでいる．プロジェクトの正式名称は「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』，総括責任者は東京工業大学の古井貞熙教授であり，研究期間は 1999 年から 2003 年までの５年間である．

本稿では上記プロジェクトの一環として，国立国語研究所が中心となって構築を進めている日本語共通語モノローグの自発音声コーパス（データベース）について，その設計と構築作業の現状を報告する．

2. コーパスの目的と内容

大規模な研究用コーパスの開発には，多大の人的資源と研究資源の投入が必要とされる．したがって，コーパスの設計にあたっては，その利用目的が明確化されていなければならない．本稿で紹介する「日本語話し言葉コーパス Corpus of Spontaneous Japanese: CSJ」にはふたつの利用目的が想定されている．

第一に CSJ は自発音声(spontaneous speech)の音声認識研究用のコーパスである．現在の統計的手法に基づく音声認識技術では，言語モデル（語の連鎖確率のモデル，bigram や trigram）や音響モデル（音素のコンテキストによる変形のモデル）が必要とされる．しかし現在までに開発されてきている言語モデルは，新聞記事などの書き言葉データから構築されているために，自然に発話された話し言葉にはそのまま適用することができない．また音響モデルも，新聞記事の読み上げなどの朗読音声と以下に述べる自発音声とでは，大幅に相違していることが予想される．

CSJ の第二の利用目的は，話し言葉そのものの言語学的研究である．音声学，音韻論，形態論など言語学の諸領域をはじめ，談話分析や言語心理学，さらには話し言葉を対象とした自然言語処理研究や計量語彙論などでも利用可能なコーパスを目指している．

以上ふたつの目的をもつコーパスには以下の諸情報が格納されていることが望まれる．

1) 音声自体
2) 音声の書き起こしテキスト（タグ付）
3) 書き起こしテキストの形態論情報（語境界と品詞）
4) 分節音や韻律のラベル情報
5) 話者や発話の状況に関する情報

1)，2)と 5)はどのような研究にも必須である．2)の書き起こし作業は自動化することはまったく考えられず，人手によって実施するしか方法がない．3)は研究目的によっては不要となる場合もあるだろう．要求される情報の精度や粒度も研究目的によって相違してくると予想される．また近年では漢字仮名まじりテキストの自動形態素解析技術が進歩したので，ある程度までは，情報付与の自動化が期待できる．4)はもっぱら音声研究に必要な情報である．分節音ラベリングの場合，既存技術の応用で一定範囲の自動情報付与が期待できる．一方，韻律ラベリングの自動化には，はるかに大きな困難が予想される．要するに 4)に関する情報付与のコストは 3)に比べるとかなり高い．つまり専門知識をもった専門家を長時間拘束する必要がある．

3. コーパスの設計

3.1 コーパスのサイズ

ところで，上記ふたつの研究目的がコーパスに要求する特性は，必ずしも一致していない．音声認識研究のためには，何よりもまずコーパスのサイズが大きいことが要請される．言語モ

デルの悉皆性を高めるためである．言語モデルは最低限書き起こしテキストと語境界情報があれば構築することができるので，書き起こしテキストへの情報付与はかなりの程度まで計算機処理にまかせることができそうである．一方，言語研究では，コーパスの量よりも質が問われる．例えば数十時間程度の量であっても，正確な情報付与がおこわれているコーパスの方が，精度に問題のある数百時間のコーパスよりもはるかに研究に貢献すると考えられる．

この量と質との相克関係を解消するために，CSJ の設計では，コーパス全体とその一部とでは付加する情報の内容を変更することにした．具体的には，音声認識研究の経験的予測に従って，コーパス全体のサイズを 700 万語と決め（時間に換算すれば，700 から 800 時間に達するサイズであるが，話し言葉の音声認識のためには最低限必要なサイズである），その全体に対して，書き起こしテキストと形態論情報を提供する．一方，コーパスの一部，50 万語程度に対しては，情報付与のコストの増加を容認して，人手を多用した高精度の情報付与を実施することにした．この高精度部分をコア(core)と呼ぶことにする．コーパス全体とコアの関係を表 1 にまとめる．上記 4)の音声研究用情報はコアに対してのみ付与される．

表 1 コーパス全体とコアの関係

	サイズ	格納される情報
コーパス全体	700 万語 (700～800 時間)	音声 書き起こしテキスト 語境界＋品詞情報 話者＋発話情報
コア	50 万語 (50～60 時間)	(上記に加えて) 分節音ラベル 韻律ラベル

3.2 発話の選択基準

上に述べたように CSJ は自発音声のコーパスである．自発音声とは発話内容があらかじめ定められていない発話のことである．反対に発話内容が定められた発話のことは朗読音声と呼ぶ．音声認識においても，音声研究一般においても，自発音声が最終的な研究対象であることは，議論をまたないであろう．しかし，朗読音声と比較した場合，自発音声にはきわめて多種多様の発話が含まれる．そのすべてを網羅することは実際上不可能であるから，コーパスの設計においては，特定の発話に焦点をあわせることが必要になる．

CSJ では，まず発話スタイルをモノローグ（独話）に限定することにした．ダイアローグ（対話）が興味深い研究対象であることは論をまたないが，信頼のおける対話音声のデータベースがすでにある程度は存在していること（例えば市川，堀内，土屋，2000 参照），一方自発的モノローグに関してはまとまったデータが存在していないこと，さらには，音声認識研究の現下のターゲットがモノローグであること，の三点を考慮して，モノローグを選択した．

次に CSJ は日本語の「全国共通語」を収録の対象としている．共通語を明確に規定することは困難であるが，我々は文字に書き起した際に分節音，語彙および文法が東京語に酷似しており，方言的特徴が認められない発話という作業上の定義を採用している．この定義は音声の韻律特徴，つまりアクセントやイントネーションには言及していないので，韻律特徴が非東京的な発話もコーパスに格納されることになる．

最後に発話の自発性の問題に立ちもどると，発話の自発性には濃淡の差があり，完全に自発的な発話からほとんど朗読音声に近いものまでが連続的に分布している．以下で触れるように，CSJ には学会発表音声が多量に格納されているのだが，文科系の学会では発表用の原稿を用意して，それを朗読する発表者が少なくない．当初，この種の講演音声はコーパスに格納しない予定であったが，実際に収録してみると，この種の講演においてもフィラーや言い淀みなどをまったく含まない音声は皆無であり，また種々の理由（講演原稿のミスや時間不足）によって，部分的には自発性の高い音声が観察されることがわかった．そのため，現在は原稿を朗読している学会発表音声も収録の対象にくわえている．

3.3 発話の収録源

CSJ に格納される音声は表 2 に示す三つのソースから収録されている．学会講演とは，様様な専門学会の全国大会などでおこなわれる研究発表をライブ録音したものである．日本音声学会全国大会でも昨年から収録作業を行わせていただいているが，その他に，日本音響学会，言

語処理学会，人工知能学会，国語学会，社会言語科学会，行動計量学会，その他の学会の協力を仰いで収録作業を続けている．

学会では短期的に大量の音声を収録できるが，その反面，発表者が 20～30 代に偏る，自然科学系の学会では男性に偏る，などの問題がある．また，学会という場の性格上，学会ごとに使用語彙に大きな変動があり，発話スタイルも相当に高いことが多い．

表2 発話の収録源

収録源	収録目標時間
学会講演	300 時間
模擬講演	300 時間
その他(一般講演会や大学の講義等)	100～200 時間

これらの問題点を解消するために考案されたのが模擬講演である．模擬講演とは CSJ に対するデータ提供を目的として雇用された一般人が行う，比較的短め（10～15 分）のスピーチである．実際には人材派遣会社を利用して，20 代から 60 代までの男性女性を均等に派遣してもらっている．模擬講演においては努めて一般的な話題がと語彙が選択されることを期待して，50 名を 1 クールとして，おおまかなテーマを指定している．例えば第 1，第 2 クールの各 50 人に対して指定したテーマは以下の通りであった．

第 1 クール（記憶に残った感情的出来事と感情的に中立の出来事）
テーマ 1 ：過去を振り返って楽しかったこと・嬉しかったことについて述べる
テーマ 2 ：過去を振り返って悲しかったこと・つらかったことについて述べる
テーマ 3 ：自分の住んでいる町，または自分の住んでいた町について述べる

第 2 クール（説明的話題を重視）
テーマ 1 ：過去を振り返って印象に残っていることについて述べる
テーマ 2 ：過去数年の間にマスコミを賑わした話題をとりあげ，事実を説明した後に自分の意見を述べる
テーマ 3 ：自分の良く知っていることについて客観的に説明する

収録は国語研究所内の録音スタジオで実施し，1 テーマについて 10～15 分のスピーチを 3 本，1 日のうちに録音している．

その他のカテゴリに属する発話は雑多である．現在，種々の学校や放送機関とソース提供に関する交渉をおこなっている最中である．

3.4 書き起こしテキスト

書き起こしテキストは，（現存する話し言葉書き起こしコーパスのように）それ自体を研究素材とすることも可能であるが，CSJ においては，むしろ音声信号に対する言語情報（およびパラ言語情報の一部）の検索用インデックスとしての役割を担わされている．CSJ では，以下の方法でテキストと音声の時間同期を確保している．まず，収録された音声を計算機に入力し，200ms 以上のポーズ（学会講演などでは室内のノイズがかなり混入しているので，必ずしも物理的な無音区間とは限らない）を検索し，このポーズに挟まれた音声区間を発話単位として同定する．ただし，言語学的な文末形式（用言の終止形や終助詞）の直後に 50ms 以上 200ms 以下のポーズが存在する場合は，そこも発話単位末と認定する．発話単位には，その開始時間と終了時間が付与される．

音声は発話単位ごとに書き起こされるが，その際，「基本形」と「発音形」という二種類のフォーマットで書き起こしが実施される．基本形は日本語の正書法に近い漢字仮名まじりテキストであるが，検索用情報としての利用を考慮して，表記のユレを可能な限り低減させるために，漢字と仮名の使い分け，漢字列への仮名のおくり方，カタカナ語の表記法などについて厳重な制約を設けている．また，話し言葉を特徴づける「ワタシャ」，「ソースリャ」，「ツーコトワ」などの融合形や「ヤルンダ」のような音便形は，そのまま基本形に表記されている．

一方，発音形は片仮名のみで表記されており，収録された音声の発音が仮名表記の範囲で可能な限り精密に記録されている．助詞の「は」「を」や連母音の発音はもちろん，話し言葉に特有の発音上の転訛も可能な限り記録している．例えば「国語研」が「コッゴケン」，「六義園」

が「リッギエン」と発音されている場合，基本形には漢字列が，発音形には上記の片仮名が記入されることになる（下記も参照）．

書き起こしテキストには，() や＜＞に囲まれて種々のタグが挿入される．(D)は，「(D あたら)最新の研究で」のように，何かを言いかけて，別の表現で言い換えた場合の，言いかけの部分に付与される．「学校 大学の」のように，単語が言い換えられているものは対象とせず，単語より短い語の断片，および助詞，助動詞の類の機能語が言い換えられている場合のみ対象とする．(?)は，音の聞き取りや表記に自信のない場合に用いる．(? スズキ,ススキ)のように複数の候補を指定することもできる．アルファベットや算用数字は，(A ダット;ＤＡＴ)や(A 千九百九十五;１９９５)のように A タグを利用し，漢字仮名に併記する形で記述する．

(W)は，先に述べた「リッギエン」のように，発音のなまけや転訛，言い間違いなどが生じた場合に用いる．(W リッギエン;リクギエン)のように，セミコロンの左に実際に発音された音を書き表わすと同時に，セミコロンの右には丁寧に発音された場合に生じる(と予想される)音を併記する．

(笑)，(泣)，(咳)，(あくび)は，これらの非言語行動と発話が同時もしくは入り混じりながら進行している区間に付与され，(L)は小さな声で発話されている区間に付与される．

この他にも種々のタグが用意されているが，省略に従う．

3.5 形態論情報

CSJ には先述した語境界と品詞情報の他に，代表形（見出し語形），代表表記（見出し語形に対する漢字等の表記），活用の種類，活用形などの形態論情報が格納される．

語境界は，「国立国語研究所」を全体でひとつとする長い単位（「長単位」）と，「国立」「国語」「研究」「所」と四つに分割する短い単位（「短単位」）の二種類を採用することになっている．そのため，代表形，代表表記，品詞などは，長短それぞれの単位に対して付与することになる．

短単位と長単位を併用する理由は，根本的には日本語では「語」の規定が明快でないことに由来するのだが，それと同時に，CSJ を利用した研究において，これら二種類の単位がともに必要とされると予想されるからである．一般的傾向として，音声認識の言語モデルでは長めの「語」が利用される一方，自然言語処理や言語学の世界では，短めの「語」が愛好される傾向がある．

代表形と代表表記は，簡単に言えば辞書における見出し語とその表記である．つまり，先に述べたとおり，CSJ の書き起こしテキストの基本形の表記は，必ずしも慣習的な正書法と一致しないし，活用する語の場合，基本形や発音形を分割して得られる形態だけでなく，活用形を抽象した見出し語形が存在する方が，コーパスの検索の利便性が向上すると考えられる．

品詞と活用形の分類については，検討を進めている最中であるが，いわゆる学校文法における品詞体系と大きくは相違しない品詞セットを利用する予定である．また，品詞セットには階層性をもたせて，大分類（名詞，動詞，etc.）から中分類（サ変名詞，動詞活用タイプ）を経て，さらに小分類へと進む形での情報付与を行う予定である．

3.6 分節音ラベリング

分節音ラベルは，発話を構成する分節音と音声の時間同期をとるための情報である．概略音素に該当するラベルセットを用意しているが，母音の無声化，イ段母音の前での子音の口蓋化，破裂音・破擦音における閉鎖区間などについては，音声学レベルのラベルも導入している．

スペクトログラムや音声波形に対して分節音ラベリングを実施した体験のある方にはただちに理解してもらえることであるが，分節音ラベリングには多大の労力が要求される．コアに含まれる 50 時間以上の音声をすべて人手でラベリングすることは非現実的であるので，ラベルの初期値を計算機によって自動的に付与することを試みている．

具体的には，まず書き起こしテキストの発音形から音声に含まれる音素特徴の系列を生成する．その際，母音の無声化や母音の脱落など，経験的に予想される音声変化については，複数の系列を生成しておく．

次に，上で生成された複数の音素系列候補の割当(alignment)を連続音声認識技術を利用して実行する．具体的には，あらかじめ大量の音声データを学習して音素毎のモデルを構築し，時

間分割された入力音声信号の音響的尤度を与えられた音素系列の各音素毎に計算する．その結果，区間を通して最も尤度の高い系列がラベリング結果として選択される．

次に自動ラベリングによって得られた音素ラベルを分節音ラベルに変換する．これは，自動ラベリングで用いる音素ラベル体系と最終的に付与する分節音ラベルが必ずしも同一でないために必要となる作業である（例えばカ行子音の母音/i/の前での口蓋化は音素列としては表現されないが，CSJ では非口蓋化子音とは別のラベルをあたえる）．

以上の手順によって自動付与されたラベルは必ずしも完全ではない．そこで最後に，音声波形やスペクトログラムなどを参照しながら人手でラベルの種類や分節音境界を修正する作業が必要となる．自発音声には，これまでの音声研究の予想を越える現象も出現するので，効率と一貫性を考慮しつつ，あらゆるケースに対応するラベリングスキームを構築してゆくことが必要である．

3.7 韻律ラベリング

韻律ラベリングは，音声のパラ言語情報の伝達に重要な役割を果たすイントネーションの情報を記号化する作業である．記号化には種々の手法が提案されているが，CSJ では多くの言語に応用されている ToBI(Tones and Break Indices)システム，具体的には日本語用の J_ToBI システムを採用する．

ただし，ToBI は基本的に言語情報の表現のみを目的として設計された韻律ラベリング手法である．そのため，CSJ に必然的に含まれる種々のパラ言語情報を担う音声現象（例えばフィラーやポーズ）の記号化には問題が生じることがある．以下にその代表例をふたつ指摘する．

まず，フィラーとポーズであるが，ToBI ではフィラーはラベリングの対象から除外されていると考えてよい．少なくとも日本語においては，フィラーは平坦なピッチで実現されることが多く，一般のアクセント句の韻律特徴（句頭にピッチ上昇を有するが，その実現形態は句頭音節の重さによって変異する，また，最大 1 個までのアクセント核を有する等）とは大きく異なっているのだが，それを現行の J_ToBI で表現することはできない．

また，言い淀み等によって，アクセント句もしくは語の内部にポーズが生じた場合，それをどう表現するかも規定されていない．自発音声を観察すると，この種のポーズは，韻律句を新たに生成する場合と（この場合は J_ToBI で対応可能），韻律句の継続に影響をおよぼさない場合（対応が必要）とが共に存在するようである．

次に，ToBI で boundary pitch movement と呼ぶ韻律句末のピッチ変化現象をとりあげる．よく知られているように，日本語では発話末尾で聞き手への情報要求やその他の語用論的意味を表出するために，ピッチが局所的に上昇・下降することがある．ToBI ではこれを boundary tone と呼ばれるトーンの分布として記述するのだが，現行の boundary tone には最低以下のふたつの問題がある．

第一に boundary tone のインベントリを現在以上に拡張する必要がある．現在の ToBI では単純な H%，L% の他には，念押しや強い疑念の表現で利用される L%H%連鎖が認められているだけであるが，これらの他に例えば H から L に下降する boundary tone(H%L%)などが必要である．

また，一層大きな問題として，boundary tone の生起位置に関する問題がある．現在の J_ToBI では boundary tone は BI3，つまり intermediate phrase の末尾にだけ生起すると想定されている．しかし，CSJ に格納される予定の自発音声のピッチ曲線を眺めていると，図１に示すように，boundary pitch movement が連続的に生じる発話で，複数の boundary tone を越えて，downstep が継続していると思われる例が散見する．これが問題となるのは，ＢＩ３ という韻律境界そのものが downstep の生起領域として定義されているからである．また図１ではピッチの自然下降現象である declination もまた boundary tone を越えて継続しているようにも見える．J_ToBI の韻律境界認定規約には declination に関係する記述は含まれていないが，J_ToBI の理論的基礎をなす日本語のイントネーション理論である Pierrehumbert & Beckamn の Japanese Tone Structure に徴する限り，これは一種の矛盾である．

さらに現在の BI3 認定基準がポーズの存在を重視していること（つまり intermediate phrase 境界にはしばしばポーズが生じるという規定）が，自発音声の記述で問題を引き起こすことは，本節冒頭の記述から明らかである．

以上のような問題点はあるものの，ToBI は利用実績を誇るすぐれたラベリング手法である．

今後は，ToBI の開発者とも連絡をとりながら，J_ToBI システムに必要な変更を加えてゆく予定である．なお，ToBI ラベリングの自動化はきわめて望ましいのであるが，自発音声の場合，その実現には多くの問題が山積していると予想される．可能な範囲で対応してゆきたい．

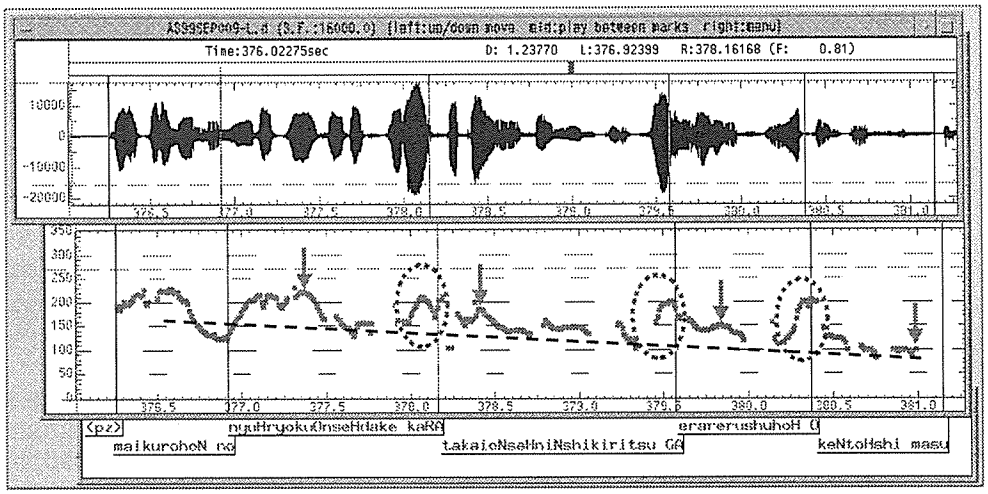

図 1 ToBI で記述が困難な boundary pitch movement．上段は音声波形，下段はピッチ曲線．「マイクロホンの入力音声だけから高い音声認識率が得られる手法を検討しました」の下線部に BPM が生じている．楕円はピッチ曲線中の BPM 該当部分を，矢印はアクセント核の位置と高さを，点線は declination の傾向を示している．

4. CSJの公開

英語研究の世界では，London - Lund コーパスを嚆矢とする一連の話し言葉のコーパスが存在している．しかし，それらの仕様を検討すると，音声を種々の方法で転記したテキストがコーパスの本体となっており，音声そのものを提供するコーパスは稀である．また音声が提供される場合も，ニュースなどの放送音声が大半で自発音声のコーパスと言えない場合がある（例えば英国の Spoken English Corpus）．近年米国では CallHome コーパスのように，きわめて自発性の高い音声コーパスも提供されるようになってきているが，このコーパスの音声は自発性が高すぎて，音声の科学的研究には役にたっても音声認識等への応用は困難なようである．

本稿で紹介した CSJ は自発モノローグの音声コーパスとしては世界に類を見ない規模と品質を達成することを目指しており，完成時には，情報処理学，言語学，さらには心理言語学などの諸領域で広く利用されることを期待している．

CSJ は現在までに約 170 時間の音声を収録し，100 時間程度を書き起こしたに過ぎない．前途遼遠ではあるが，2004 年 3 月のプロジェクト終了時を期して公開し，研究目的の利用には原則無償で公開する予定である．また，2000 年度末以降，各年度末にも一定量（100 時間程度）のデータをモニター公開する予定である（ただし内容は音声と書き起こしテキストに限定する）．詳細は近くインターネット等で広報する予定である（http://www.crl.go.jp/pub/orc-speech/ にプロジェクトのホームページが開設されている）．

謝辞： CSJ のデータ収録にご協力いただいた話者の方々に心より御礼申しあげます．また CSJ の設計には古井貞熙教授以下，融合研究プロジェクトメンバーの貢献が多大であったことを記して感謝します．