1—Q—31 共通日本語話し言葉コーパスの設計*

○前川 喜久雄，△小磯花絵，籠宮隆之（国立国語研究所）
古井貞熙（東京工業大学），△井佐原 均（通信総合研究所）

1 はじめに

現在の音声科学や音声言語情報処理の研究は，程度の差はあっても基本的に朗読音声(read speech)を対象としている．これを拡張ないし再構築して自発音声(spontaneous speech)の本質を把握し，その処理技術を開発することは，21世紀初頭における音声言語研究の最重要課題のひとつと言ってよいだろう．

ところで，自発的な話し言葉（以下では単に話し言葉と呼ぶ）の研究を行なう者がまず直面する困難はデータの収録である．自発音声は，その定義上，発話内容を収録者側であらかじめ指定して統制することが許されない．内容を話者まかせにする以上，いきおい多量のデータが必要となる．さらに話し言葉は，発話の状況に応じてきわめて多様なスタイルで実現されると予想されることもまた，大規模データの必要性を高める要因である．

我々は平成11年度科学技術振興調整費開放的融合研究制度研究課題「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」の一環として，日本語を対象とした大規模な話し言葉コーパスの作成を開始した．研究期間は5年である．本発表では，この研究で作成し公開する予定のコーパスの基本設計と作業の進行状況を報告する．

2 コーパスの設計

2.1 コーパスの規模

コーパスの設計にあたっては，最初にコーパスの利用目的を明確に設定しなければならない．特に話し言葉のように本質的に多様性に富む現象については，この設定が一層肝要であると考えられる．

我々は，1）音声認識，特に言語モデルの構築に最低限必要な量を有し，同時に 2）音声・言語（処理）の研究にも利用可能な精度と付加情報を有する話し言葉コーパスを構築したい．しかし，これらふたつの目的は二律背反の関係にある．言語モデルのためにはコーパスは大きいほどよいが，コーパスが巨大化するほど，その全体に対して精度の高い情報付与を行なうコストが増大するからである．

我々は表 1 に示すようにコーパスの総量に対しては 1)の目標を優先して適用し，コーパスの一部（中核部）に対しては 2)の目標を満たす情報付与を実施することによって，上記の二律背反を回避しようとしている．

また，本コーパスは独話（モノローグ）を主たる対象としており，一部を除いて対話を収録する予定はない．現在の音声認識用言語モデルが基本的に独話を対象としていること，真に自発的な対話音声はしばしば転記が不可能なほどに朗読音声とは異なっていることが，この決定の理由である．

2.2 言語変種

日本語にも様々な変種があるが，本コーパスはそのうち，全国共通語を対象とする．現代の日本では高齢者を除外すれば，地域言語（方言）だけを話す話者はほとんど存在しない．日常的には方言で生活している話者も，発話の状況がおおやけの場になると（特に不特定の聴衆が存在する状況では），方言色を大きく減じた変種，すなわち全国共通語にスイッチした発話をおこなうのが普通である．

全国共通語の文法と音韻は東京語を近似していることが多い．しかし，韻律面，特にアクセントは必ずしもそうでないことがあり，特に近畿地方や無アクセント方言地域出身の話者は例外となることが多い．

本コーパスでは，韻律面の特徴が東京語とは大幅に相違している発話であっても，文法と分節音の音韻が東京方言を近似していればコーパスに格納し，反対に文法や音韻がしばしば東京語のそれから逸脱する発話は除外することにしている．

また次節に述べるように，本コーパスでは多少ともおおやけの状況下での発話を収録しており，聴衆も不特定であることが多いので，全国共通語の使用が期待され，実際にもそうなっていると認められる．

表 1 目標とするコーパスの量と付加情報

	データ量	付加情報
全体	700万形態素 (約 1000 時間)	音声(16KHz,16bit) 転記テキスト 形態素解析情報
中核部	そのうち 50万形態素 (約 70 時間)	(上記に加えて) 分節音ラベル情報 韻律ラベル情報

2.3 収録源

次に，実際にどこでどのような話し言葉を収録するかが問題となる．現在のところ，各種学会・研究会における研究発表の音声と模擬的な講演会の音声を中心に収録を進めている．対象とする学会はできるだけ幅広い学術分野を対象としたいところであるが，現時点では協力の得やすい音声・言語関係の諸学会が中心となっている．

模擬的な講演会とは，一般の話者が自由に選択した日常的な話題について簡単なメモを頼りに10分から20分程度まとまった内容のスピーチをおこなった音声である．コーパスのために企画した講演であるので模擬的と呼んでいる．

放送番組の音声収録も望まれるので，複数の放送局と交渉を進めている．コーパスの公開を前提としているために話者と機関の著作権をクリアする必要があり，現時点で交渉が終了した機関はない．

2.4 自発性の高低

本コーパスに収録される発話の自発性は決して均一ではない．完全に自発的と言える発話は少なく，学会講演の中には前準備された原稿を朗読している音声も一部含まれている．しかし，この種の講演も，言い淀みやフィラーの分布に関しては，職業的話者

Design of the Spontaneous Speech Corpus of Common Japanese. Kikuo Maekawa, Hanae Koiso, Takayuki Kagomiya (Natl. Lang. Res. Inst.), Sadaoki Furui (Tokyo Inst. Tech.), Hitoshi Isahara (Com. Res. Lab.). Contact: kikuo@kokken.go.jp

の朗読とは歴然と相違しているので，コーパスに格納している．また，収録作業の一環として，すべての発話の自発性を主観的に評定したデータを作成しており，コーパスと同時に公開する予定である．

2.5 音声の収録

音声はヘッドセット型の接話型指向性マイクロホンを話者に装着してもらい，48KHz, 16bit でDATに収録している．同時に話者の上半身映像と講演に利用されたOHP等の画像をデジタルビデオに記録している．これは後述する転記作業のための参考資料であり，コーパスとしては公開しない．

2.6 音声の転記

音声認識用言語モデルの構築のためにも，種々の言語(処理)学的研究のためにも，収録した音声をテキストに転記する必要がある．その際，以下のような問題を検討しなければならない．

a) 発話単位の認定：話し言葉では書きことばの文に該当する単位を認定できないことが多い．発話をどのような基準で単位分割するか．

b) 表記法：正書法を用いるか，実際の音声に忠実な表記を用いるか．

c) 非流暢性(disfluencies)：フィラー，言い淀み等，話し言葉に固有の音声の表記法．

d) 種々の雑音源：呼吸音，リップノイズ，笑い，デモ音，室内外の雑音等の示しかた．

a)については，客観性と確実性を重視して，ポーズによって発話単位を認定することにした．200msを閾値として，それより長いポーズがあれば発話単位として分割するのが原則である．ただし，50ms 以上(200ms以下)のポーズについては，その直前に言語学的な文末形式，つまり述語の終止形や終助詞などが存在していれば，発話単位として分割している．また言語学的な文末形式が全くポーズをともなわずに出現している場合は，発話の分割はおこなわないが，転記テキスト中にその所在を示すタグを埋め込むことにしている．

b)については，ほぼ正書法に準じた表記法(基本形)と発音よりの表記法(発音形)との二種類の転記テキストを作成して対処している．発音形は，正書法に規定された例外的用法(助詞の「わ，を，へ」や連母音の長音化など)や，漢字仮名混じり正書法表記に読みが複数存在する場合(「今日は」がコンニチワかキョーワか，「研究所」がケンキューショか～ジョか)や，正書法では想定されていないが，実際には存在する発音のユレ(「国語研」がコッゴケン，「六義園」がリッギエン，「縁側」かエンガーと発音される)などの場合に，その発音を確定させる役割を果たす．ただし「アンタ」「タベッカラ(食べるから)」など，話し言葉で頻発する形態は，そのまま基本形として表記する予定である．

c)については，転記テキスト中に各種タグを埋め込んでフィラー，言い淀み等の位置を示している．

d)のうち大きな雑音源については発話単位の認定作業のなかで雑音を同定し，その時間情報を提供し，それ以外は可能な範囲で転記テキスト中に所在を示している．

以上については，図1に示した転記データの実例を参照されたい．

2.7 形態素情報

転記テキストには形態素解析を施し，語境界情報と品詞情報とを提供する．中核部分に対しては手作業で高精度の分析を施す．コーパス全体については，中核部分を学習データとして自動形態素解析を施す予定である．

ちなみに日本語における語の認定基準はコーパスの利用目的によって相当異なってくる．本コーパスでは「国立国語研究所」を「国立/国語/研究/所」と細かく分割する単位と全体を一語とする単位，長短二種類の単位について解析結果を提供する方針で検討をすすめている [1].

2.8 中核部への情報付加

中核部に対しては，分節音ラベリングと韻律ラベリングを実施する予定である．分節音ラベルは音素よりも音声に近い (subphonemic)水準のラベルセットを構築して利用する予定である．

韻律ラベリングにはトーンモデルに基づくラベリング体系を採用する予定である．ただし，現在提案されている JToBI[2]をそのまま話し言葉に適用できるか否かについては慎重な検討が必要である．相当程度の拡張・改変が必要となるかもしれない．

3 今後の予定

本コーパスは日本語の自発音声を対象とする初めての大規模コーパスであり，音声・言語研究の広い領域での利用を想定している．中核部分は平成14年度に，コーパス全体は平成16年度に構築し終える予定であるが，完成以前にも学術的利用を目的としたモニター公開を想定している．

```
0181 04:02:173-04:03:589 L:
(F え)          &    (F エ)
例えば          &    (W ツトエバ;タトエバ)
後ろに          &    ウシロニ
付いた          &    ツイタ
ラベル          &    ラベル
0182 04:03:942-04:04:127 L:
(D を)          &    (D オ)
0183 04:04:602-04:04:867 L:
(F えー)        &    (F エー)
0184 04:07:078-04:09:694 L:
の              &    ノ
情報も          &    ジョーホーモ
見ながら        &    ミナガラ
```

図1 転記データの一例 1行目は発話IDと発話単位の開始時間と終了時間である．2 行目から 6 行目がこの発話単位の転記テキストであり，基本型と発音形が&で区切られて1行の左右に配されている．可読性を高めるために概略文節に該当する単位毎に改行がおこなわれている．以下同様に全部で四つの発話単位が転記されている．

2 行目などの(F)はフィラーのタグ．3 行目の(W)は修正のない言い誤りで，セミコロンの後ろが正しいと想定される語形．8 行目などの (D)は自立語ないし分節よりも小さな単位での言い誤りである．(D)は話者自身によって言い直されることが多い．本例では 12 行目が 8 行目の言い直しである．

参考文献

[1] 国立国語研究所(1983).『高校教科書の語彙調査』(国立国語研究所報告 76).

[2] J. Venditti (1995). "Japanese ToBI labelling guidelines." http://ling.ohio-state.edu/Phonetics/J_ToBI/