『日本語話し言葉コーパス』について ―開発の経緯と普及の現状―*

○前川喜久雄（独立行政法人国立国語研究所研究開発部門）

1　CSJ とは

『日本語話し言葉コーパス』（Corpus of Spontaneous Japanese,以下 CSJ）は現代日本語の自発音声コーパスである。多少とも自発性を有する音声を対象としている点に最大の特徴があるが、規模の大きさとアノテーションの豊富さにおいても、従来の音声データベースとは一線を画している。2004 年の公開後、広い領域で利用されつつある。

2　開発の経緯

2.1 研究体制

CSJ の開発費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。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」。東京工業大学の古井貞熙教授を総括責任者に迎えた CRL（通信総合研究所、現在は情報通信研究機構）と国立国語研究所の共同研究である。CRL では井佐原均氏が、国語研では私がそれぞれサブリーダーとして研究グループを統括した。研究期間は 1999-2003 年度の５年間であった。

CSJ の開発は専ら国語研が担当したが自動形態素解析とアノテーションの一部（談話ラベリング、節境界ラベリング、重要分抽出）は CRL で実施した。

プロジェクトには京都大学、ATR、NTT などの研究者にも参加していただいた。また ATR 音声翻訳通信研究所（当時）の山本誠一氏には研究の立ち上げ時に様々な助力を頂戴した。

2.2 設計

CSJ の設計にあたっては、音声認識研究での利用を最重要の応用領域として念頭においたが、副次的には言語研究での利便性も勘案した。

設計において最初に決断したのはデータサイズである。プロジェクトの申請書には音声認識研究に最低限必要なデータ量を 1000 万語と見積もったが、予算の制約で実際には 700 万語を目標値とすることになった。

次に収録する音声のタイプを決定した。やはり音声認識への応用を重視してモノローグ、それもまとまった（要約を作成するに足る）内容をもった知的なモノローグを主要な対象とすることにした。表１に CSJ に格納された音声の種類と分量を示す。

Table 1 Speech Type in CSJ

音声のタイプ	話者数	時間
学会講演 APS	819	274.4
模擬講演 SPS	594	329.9
一般講演会 PL	(16)	24.1
インタビュー Dialogue	(10)	2.1
課題指向対話 Dialogue	(16)	3.4
自由対話 Dialogue	(16)	3.1
朗読 Reading	(248)	15.5
再朗読 Reading	(16)	5.5
計	1417	661.6

( )内の話者数には重複が含まれる

音声に対するアノテーションとしては、精密な転記テキストと、その形態素情報をコーパス全体に対して提供することとした（実際にはその他に節境界情報と印象評定情報もコーパス全体に提供した）。しかし、人文系での言語研究を考えると、これだけではいかにも物足りない。そこでデータの一部、約 50 万語について、様々なアノテーションを追加することとした。この部分を CSJ の「コア」と呼んでいる。

表 2 はコアに追加で施されたアノテーションの一覧である。これらに詳しく触れている余裕はないが、我国ないし世界初の試みとなったものが多い。また形態素解析では長短２種類の単位で提供するなどの新しい試みもある。興味のある方は文献[1]ないし CSJ のマニ

*Corpus of Spontaneous Japanese: Its Construction and Dissemination by Kikuo MAEKAWA (Dept. Language Research, Nat'l Inst. for Japanese Language).

ュアル（国語研のホームページで公開している）を参照していただきたい。

Table 2 Annotation of the CSJ-Core

アノテーション	備考
分節音ラベル	X-JToBI 仕様
イントネーションラベル	
節単位情報（手解析・高精度）	コア以外は自動解析
印象評定データ（集合評定）	コア以外は単独評定
係り受け情報	---
要約・重要文情報	---
談話境界情報	コアの一部
形態素解析（手解析・高精度）	コア以外は自動解析

3 普及の現状

CSJ は一部を構築中にもモニター公開したが、研究基幹終了後の 2004 年 6 月にコーパス全体の一般公開を開始した。DVD18 枚組の巨大なコーパスとなった。また 2005 年春には XML のエラーを修正した追加 DVD をユーザーに無償送付した。

公開後、本稿執筆時点までの約 1 年半に 270 セットを出荷した。契約形態と利用者の所属が文科系か理工系かでクロス集計すると表 3 のようになる。契約形態のうち無償はプロジェクト関係者への贈呈分である。

Table 3 Dissemination of CSJ

所属 / 契約形態	文科系	理工系
大学・独法等	75	57
企業	0	25
学生	38	0
個人	37	3
無償	27	8
計	177	93

CSJ を利用した研究成果がどの程度出版されているかも分析した。1999 年以降 2005 年 10 月までに出版された論文（口頭発表論文や商業誌への寄稿も含む）で CSJ に言及している論文は 456 本あり、そのうち 109 本が英文論文である。筆頭著者が文科系研究者であるものが 137 本、理科系が 319 本である。表 4 に CSJ 関係の発表が多く発表されている国内学会を上位 6 位まで示す。論文数には査読論文だけでなく解説論文類も含めている。

Table 4 Speech Type in CSJ

学会	論文	口頭発表
音響学会	4	140
言語処理学会	3	36
電子情報通信学会	4	25
音声学会	4	10
情報処理学会	3	7
人工知能学会	0	7

4 言語研究での利用

理工系での利用については本セッションでの古井、河原両先生の報告と文献[2]に譲ることとし、以下では言語研究での利用例を筆者らの研究によって紹介する。

4.1 語形の変異

CSJ のような自発音声コーパスがもっとも直接に役立つ言語研究の分野のひとつは言語変異の研究である。同じ意味を表す言語形式が複数存在している（揺れている）とき、その揺れの背景に潜む要因を解明しようとする研究である。CSJ は転記テキストが精密であり、データ量も多く、さらに或る程度のスタイル差を包含しているためにこの領域の研究にとってほぼ理想的なデータベースとなっている[3,4]。

ところで言語変異の研究では、他の社会調査同様、調査行為自体が回答にバイアスを与えることが知られている。このバイアスをできるだけ抑えた状態で、自然な言語行動を大量に観察することが望まれるのだが、それが難しいために、従来はアンケート調査というバイアスのかかりやすい研究方法を用いることが多かった。

図 1 は動詞「来る」の可能形を「コラレル」と言うか、いわゆるラ抜き言葉である「コレル」と言うかについて、文化庁国語課が 2001 年度に実施した全国規模のランダムサンプリングによるアンケート調査結果と CSJ の分析結果とを比較したものである。横軸は回答者ないし話者の生年である。

いずれの調査結果も若い層ほどいわゆるラ抜き言葉であるコレルの率が上昇している点は共通している。しかし伝統的なコラレルの使用率をコレルが逆転するタイミング（生年代）に注目すると、アンケート調査では 1970 年代生まれの層で始めて逆転が生じているのに対し、CSJ のデータでは 1940 年代で既に逆転が生じていることがわかる。

アンケート調査の実施と解釈には慎重を期すべきことがよく納得できる結果である。本件の場合、ラ抜きが乱れた語形であるという意識が回答におけるバイアスとなっていることは間違いないだろう。

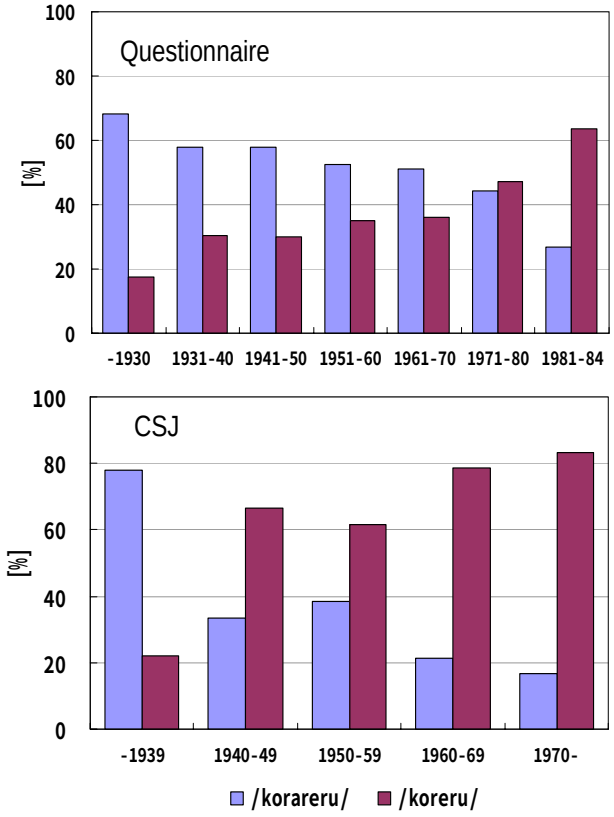

Figure 1. Comparison of language behavior and its retrospection. Case of potential verb.

アンケート調査の限界についてもうひとつ例を挙げておこう。図 2 は我国の国名「日本」をどう発音するかについて、NHK 放送文化研究所が 2004 年に実施したアンケート調査（対象はアナウンサーと有識者）と CSJ を比較したものである。ニホンとニッポンとの比率はアンケートでは約 3 対 2 であるが、CSJ に記録された言語行動では 97 対 3 である[5]。

この場合、ラ抜き言葉とは違って、言葉の乱れに関するバイアスがあるとは考えにくい。人間は自身の言語行動を量的に内省することがそもそも不得手なのだと思わせられる事例である。

実際 CSJ からは図 2 の類例を数多く抽出することができる。例えば「NHK」に多くの語形が存在することは広く認識されているだろうが、そのなかで一番よく用いられる語形は何であるかを内省してみていただきたい（正解は文献[6]参照）。

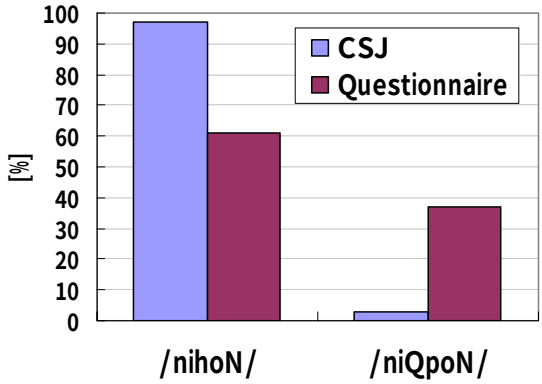

Figure 2. Comparison of language behavior and its retrospection. Case of /nihon/ vs. /niQpon/.

以上は個別的な語例についての解析結果であったが、CSJ を利用した言語変異研究の最大の利点は、多数の語の変異を同時に観測できること、つまり日本語の変異の全体像に肉薄できることにある。

CSJ に記録された 752 万語の短単位全体を解析してみると、イ）形態論的ないし音声的な変異は延べ 40 万件以上記録されており、発話のスタイルと相関を示すこと（図 3 参照）、ロ）そのうち 1 回でも変異を示す語は異なり語数で 1 万語以上におよぶこと。しかし、ハ）変異の大部分は少数の語によって生み出されており、変異の生起数がもっとも高い 20 語の変異形が変異全体の 77%を占め、また、ニ）各語について最大 3 個までの変異形リストを作成することによって変異全体の 90%以上を網羅できることなどが判明してきた[4]。

秋田と河原による発音辞書の研究においても同様の結果が報告されている[7]。

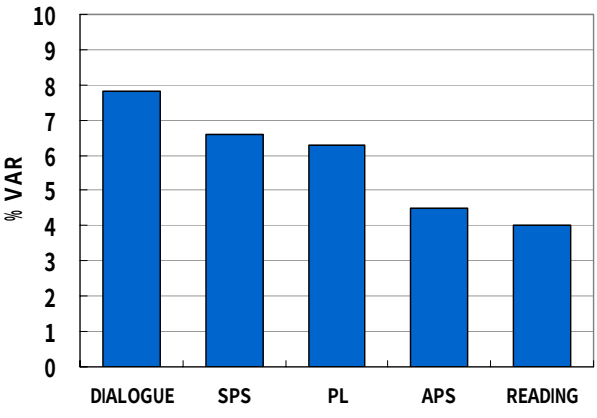

Figure 3. Correlation between the speech type (abscissa) and the rate of word-form variation. See table 1 for speech types shown on abscissa.

4.2 イントネーションの変異

CSJ コアのイントネーションラベルを用いたイントネーション研究の例として未発表のデータを示す。

日本語の句末イントネーションのうち上昇下降調には､上昇と下降が句末最終モーラ(ないし音節）内部で生じるものと、句末２モーラにわたって生じ、上昇のピークが次末モーラに位置するものとがある。CSJ では両者に L%HL%という同じトーンラベルを付与し、その上で後者を PNLP(Penult Non-Lexical Prominence)と名づけて区別している。

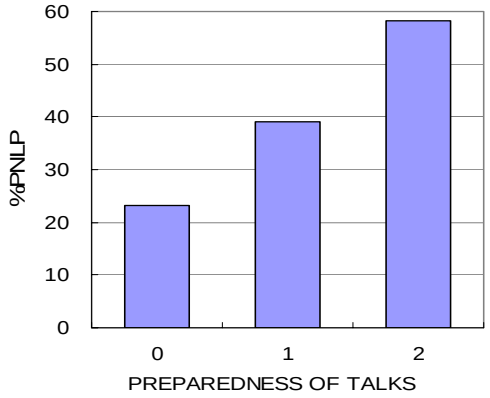

Figure 4. Ratio of PNLP variant to the total occurrence of L%HL% as a function of rated preparedness of APS talks. In abscissa, ‘0’ is spontaneous, ‘1’ is partly prepared, and ‘2’ is mostly prepared.

図 4 は学会講演を対象として、L%HL%全体に占める PNLP 型の生起率と講演の「準備度」に関する印象評定値の相関を示したものである。準備度が高い、すなわち原稿に依存している印象を与える講演では PNLP の生起率が上昇することがわかる。

講演音声を聞いただけで、それがどの程度原稿に依存しているかを、或る程度確信をもって判断できるのには、このような韻律特徴が関係していると思われる。

5　問題点と解決への展望

CSJ にはどのような問題があるだろうか。最大の問題点は話し言葉の全体像を把握していないこと、つまり均衡コーパス(balanced corpus)になっていないことであろう。

話し言葉は書き言葉とちがって便宜的にも母集団を規定することが難しいから、統計学的な意味で真に代表性を有するコーパスを設計することはほぼ不可能である。しかし、現在の CSJ に欠落しているタイプの話し言葉、例えばより親密な間柄での対話や種々の放送音声などを段階的に補足してゆくことは可能である。将来的には是非拡張を図りたい。

自然な対話音声の収録や特定個人の長時間にわたる音声行動の追跡には、マイクロホンアレイやユビキタスコンピューティングの技術が役にたちそうに思う。

最後に CSJ を用いて言語変異現象の研究を始めてから痛感するようになった問題を指摘しておく。現代日本語には書き言葉の均衡コーパスが存在しないことである。そのため、話し言葉の分析結果を書き言葉と比較することが難しい。

この問題を解決するために 2006 年度から現代日本語書き言葉の均衡コーパスを構築しはじめることにした。国語研究所の運営費交付金による研究事業として位置づけると同時に科学研究費も申請中である。できれば５年間で１億語規模の均衡コーパスを構築したい。これが完成すれば、音声研究にも間接的に裨益するものと考えている。

謝辞：CSJ に音声を提供してくださった発話者の方々、CSJ をご利用いただいているユーザー各位に感謝します。

参考文献

[1] 前川「『日本語話し言葉コーパス』の概要」, 日本語科学, 15, pp.111-133, 2004.

[2] Furui et al. “Analysis and recognition of spontaneous speech using Corpus of Spontaneous Japanese,” *Speech Communication*, 47/1-2, pp.208-219, 2005.

[3] 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.

[4] Maekawa. "Quantitative analysis of word-form variation using a spontaneous speech corpus", *Proceedings of Corpus Linguistics 2005*, Birmingham, To be published.

[5] 前川「ニホンかニッポンか」,文化庁月報, 431, p.30, 2004.

[6] 前川「NHK の発音」,情報通信ジャーナル, 23(5), p.40, 2005.

[7] Akita. and Kawahara “Generalized statistical modeling of pronunciation variations using variable-length phone context.” *Proceedings of IEEE-ICASSP*, 1, pp.689-692, 2005.