『日本語話し言葉コーパス』RDB の構築

小磯 花絵 (国立国語研究所理論・構造研究系)†
伝 康晴 (千葉大学文学部/国立国語研究所言語資源研究系)
前川 喜久雄 (国立国語研究所言語資源研究系)

Construction of Relational Database for the *Corpus of Spontaneous Japanese*

Hanae Koiso (Dept. Linguistic Theory and Structure, NINJAL)
Yasuharu Den (Faculty of Letters, Chiba University/Dept. Corpus Studies, NINJAL)
Kikuo Maekawa (Dept. Corpus Studies, NINJAL)

1. はじめに

『日本語話し言葉コーパス』（*Corpus of Spontaneous Japanese*, 以下 CSJ）は，1999 年から 5 年間かけ，国立国語研究所・情報通信研究機構（旧通信総合研究所）・東京工業大学が共同で開発した，約 640 時間の日本語自発音声からなるデータベースである。2004 年に公開を開始して以降，音声言語情報処理，自然言語処理，日本語学，言語学，音声学，心理学，社会学，日本語教育，辞書編纂など，幅広い領域で利用されており，第 2 刷（2008 年），第 3 刷（2011 年）と順調に版を重ねている。

このように CSJ は様々な分野の研究者の関心を集める一方で，多くの（主に人文系）ユーザから，具体的にどのように利用してよいか分からないといった相談が数多く寄せられているのも事実である。CSJ には多種多様な研究用付加情報が付されており, 各種情報を統合して表現した XML 文書も提供されているが，XML を操作する技術を持たないユーザには手も足も出ないのである。

そこで筆者らは，CSJ 第 3 刷に基づき，XML 文書で表現された情報をもとに各種情報を相互に関連付けて表現した RDB を構築した。本稿ではその設計について紹介する。

2. 『日本語話し言葉コーパス』の概要

CSJ には，転記情報，文節情報，形態論情報（長単位・短単位），節単位情報，分節音情報，韻律情報，係り受け構造情報，談話境界情報，要約・重要文情報，印象評定データなど，多様な研究用付加情報（アノテーション）が付されている。ただし，これらの情報は CSJ の全体に対して斉一的に付されているわけではなく,「コア」と呼ばれるデータ範囲（約 50 万語，約 45 時間）を対象に集中的に付与されている。図 1 はコアとそれ以外における情報付与の異同の概念図である。

このように CSJ には多様な研究用付加情報が付されているが，例えば，節単位末尾の短単

† koiso@ninjal.ac.jp

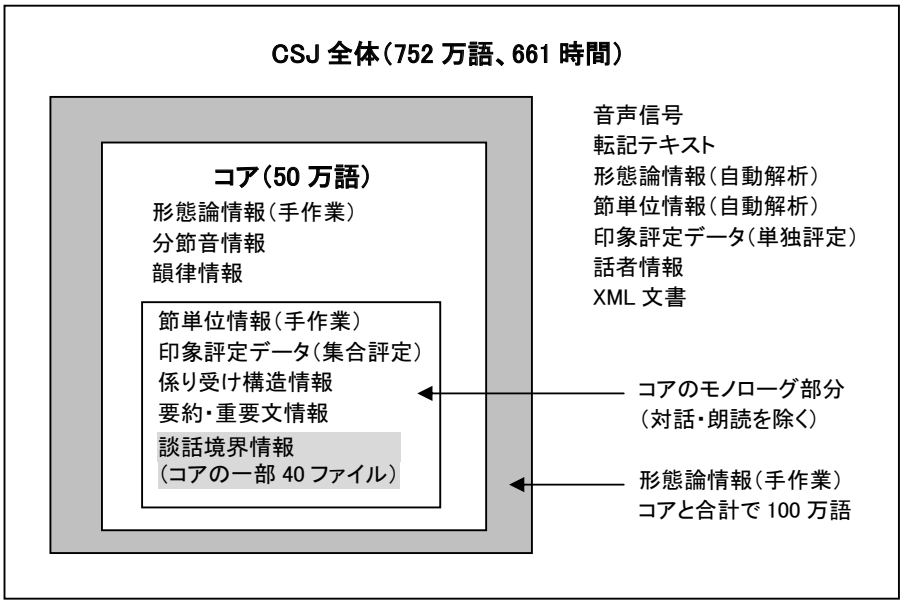

図 1　CSJ の研究用付加情報の階層構造（(国立国語研究所 2006) より）

位冒頭のモーラの時間長を知りたいといったように，複数の言語単位に関わる分析を効率的に行うためには，各種情報を相互に関連付けて表現することが求められる。また，コーパス構築の観点からは，情報間の整合性を検証・修正した上で精度の高いデータを作り上げる必要がある。そこで CSJ では，多様な情報の記述に適し，かつ，構造の妥当性を検証する仕組みを持つ XML を用いて，各種情報を統合した文書（以下，CSJ-XML）を作成し，ユーザに提供している。CSJ-XML 文書の例を図 2 に示す。

これを見ると分かるように，各種情報が多種多様なタグで表現された複雑な文書である。ここから必要な情報を効率的に抽出するには，XSLT という XML 文書の書式変換用言語を用いる必要があるが，プログラミングの経験のないユーザには敷居が高く，CSJ の豊富な研究用付加情報を高度に利用した研究の推進を阻んでいたといえる。

3.　『日本語話し言葉コーパス』RDB

この状況を踏まえ，筆者らは，図 1 のコアと呼ばれるデータ範囲を対象に，CSJ-XML で表現された情報をもとに各種情報を相互に関連付けて表現した RDB を試作した（以下，CSJ-RDB）。RDB（リレーショナルデータベース）とは，相互に関連付けられた複数のテーブルから構成されるデータベースであり，複数の言語単位に関する研究用付加情報を有する CSJ の表現に適したデータ構造である。個々の情報はテーブル，つまり行と列で構成される表の形式で記述され，直感的に把握しやすい。また，各テーブルは相互に関連付けられ，ばらばらにデータが提供される場合とは異なり，複数の言語単位に関わる検索も比較的容易に行うことができる。

3.1 では，CSJ-XML のデータ表現方式の概要を述べ，その問題点を指摘する。3.2 で CSJ-RDB における各種単位の基本構成について言及した上で，3.3 で具体的なデータベース構成について述べる。

```
<IPU IPUID="0065" IPUStartTime="00142.148" IPUEndTime="00142.524" Channel="L">
 <LUW LUWID="1" LineID="001" IsNewLine="1" LUWDictionaryForm="マタ" LUWLemma="又" LUWPOS="接続詞">
  <SUW SUWID="1" ColumnID="001" OrthographicTranscription="また"
       PlainOrthographicTranscription="また" PhoneticTranscription="マタ"
       SUWDictionaryForm="マタ" SUWLemma="又" SUWPOS="接続詞"
       ClauseUnitID="12" ClauseBoundaryLabel="&lt;接続詞&gt;" Dep_BunsetsuUnitID="0"
       SE_Subject1_50p="1" SE_Subject2_50p="1" SE_Subject3_50p="1">
   <TransSUW TransSUWID="1">
    <Mora MoraID="1" MoraEntity="マ">
     <Phoneme PhonemeID="1" PhonemeEntity="m">
      <Phone PhoneID="1" PhoneStartTime="142.184389" PhoneEndTime="142.227423"
             PhoneEntity="m" PhoneClass="consonant"/>
     </Phoneme>
     <Phoneme PhonemeID="2" PhonemeEntity="a">
      <Phone PhoneID="1" PhoneStartTime="142.227423" PhoneEndTime="142.305318"
             PhoneEntity="a" PhoneClass="vowel">
       <XJToBILabelTone Time="142.230742" ToneClass="ibt" F0="234.8590">%L</XJToBILabelTone>
      </Phone>
     </Phoneme>
    </Mora>
    <Mora MoraID="2" MoraEntity="タ">
     <Phoneme PhonemeID="1" PhonemeEntity="t">
      <Phone PhoneID="1" PhoneStartTime="142.305318" PhoneEndTime="142.343411"
             PhoneEntity="SclS" PhoneClass="others"/>
      <Phone PhoneID="2" PhoneStartTime="142.343411" PhoneEndTime="142.357172"
             PhoneEntity="t" PhoneClass="consonant"/>
     </Phoneme>
     <Phoneme PhonemeID="2" PhonemeEntity="a">
      <Phone PhoneID="1" PhoneStartTime="142.357172" PhoneEndTime="142.486984"
             PhoneEntity="a" PhoneClass="vowel">
       <XJToBILabelTone Time="142.486954" ToneClass="fbt" F0="243.6610">L%</XJToBILabelTone>
       <XJToBILabelWord Time="142.486984" PerceivedAccPos="0">mata</XJToBILabelWord>
       <XJToBILabelBreak Time="142.486984" FillerStart="1">3</XJToBILabelBreak>
      </Phone>
     </Phoneme>
    </Mora>
   </TransSUW>
  </SUW>
 </LUW>
</IPU>
```

図 2　CSJ-XML 文書の例

3.1　CSJ-XML 文書のデータ表現方式

CSJ-XML では，原則として階層構造が認められる 6 つの単位「転記基本単位（一定以上のポーズで区切られた単位)」「長単位」「短単位」「モーラ」「音素」「分節音」を設定して各種情報を表現している (国立国語研究所 2006)。しかし，ポーズを基準に認定される転記基本単位と言語的に認定される長単位はつねに階層関係を維持するわけではなく，「国立 ⟨ポーズ⟩ 国語研究所」のように，長単位内にポーズが生じて複数の転記基本単位に長単位がまたがることも少なからず存在する。このような場合，XML 表現の階層性を保つために長単位をいったんポーズで分割して表現した上で，分割されたことを示す情報を記すなど，複雑な記述となっている。

長単位よりも上位の（より長い）単位である節単位や文節では，同様の問題がより頻繁に生じる。これらは言語単位として直接的には表現されず，その境界の情報が短単位の情報に埋め込まれている。アクセント句と長単位の間にも同様の問題が生じる。このように，節単位，文節，アクセント句では言語単位が CSJ-XML 文書中で直接的に表現されないため，扱いがより複雑化する。

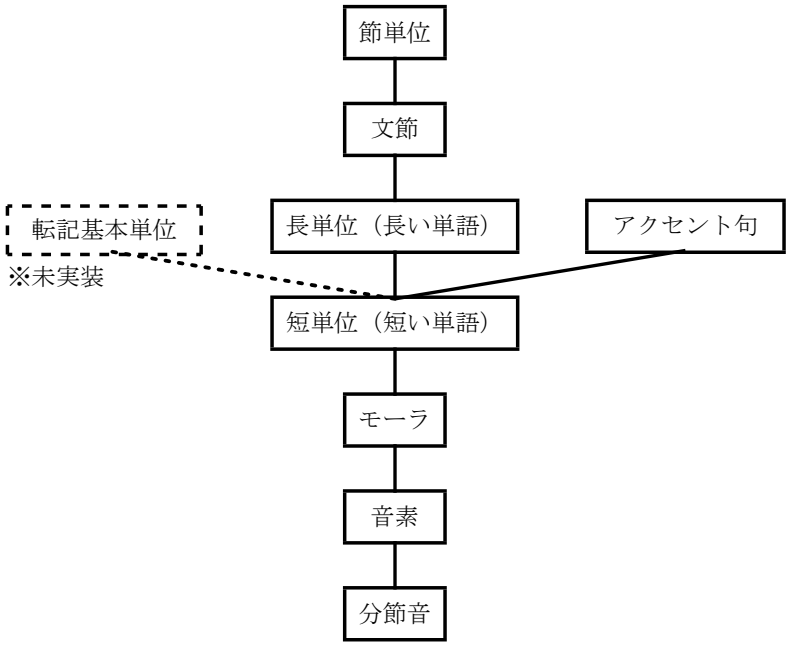

図 3　CSJ-RDB のデータ表現方式

なお CSJ では，節単位，文節，長単位，短単位といった統語形態論的な階層関係の認められる言語単位に特化した別の XML 文書も提供されているが，これとベース XML 文書にあるアクセント句の情報とを組み合わせて使うといったようなことは容易ではない。

これらの問題は，利用面での困難さを招くだけでなく，データの整合性の検証作業を過度に複雑化し，検証が不十分なままデータを提供することにつながりかねない。

3.2　CSJ-RDB のデータ表現方式

前節で指摘した問題点を踏まえ，CSJ-RDB では，CSJ-XML で採用されていた入れ子構造による階層関係の表現をやめ，多層的アノテーションの表現方法として主流となっているスタンドオフ形式による表現を採用した。この方式では，各要素の談話中での生起位置を開始時間と終了時間の対で表し，時間区間の包含関係によって階層関係を表現する。ただし，以下に述べるように，実際には利用の簡便のため，単位間の親子関係を別テーブルで陽に表現している。この結果，図 3 に示すように，節単位～短単位の系列からなる統語形態論的階層関係と，アクセント句～分節音の系列からなる韻律音韻論的階層関係を同時に表現できるようになった。

3.3　CSJ-RDB の構成

CSJ-RDB は，伝・小磯 (2012) で提案したコーパス管理用 RDB を研究用 RDB に作り変えることで構成されている。コーパス管理用 RDB では，談話中の要素を記述するセグメントと，セグメント間の関係を記述するリンクによって，アノテーションを極めて一般的に表現している。セグメントは，図 3 のように層化されており，どの層とどの層の間に親子（先祖・子孫）関係があるか指定されている。研究用 RDB では，利用の簡便を考え，層（単位）ごとに別々のテーブルでセグメントを表現している。層（単位）間の親子（先祖・子孫）関係もそれぞれ

表 1　セグメント・テーブルの個別記載情報

テーブル名	列名	説明	取りうる値 or 例
segClause （節単位）	OrthographicTranscription ClauseBoundaryLabel CU_ObligateComment	基本形 節境界ラベル 節単位義務的コメント	そこに行きましたが /並列節ガ/ 引用節構造
segBunsetsu （文節）	OrthographicTranscription	基本形	国立国語研究所では
segLUW （長単位）	OrthographicTranscription	基本形	国立国語研究所
segSUW （短単位）	OrthographicTranscription word nMorae	基本形 音素記号列 モーラ数	国語 kokugo 3
segAP （アクセント句）	OrthographicTranscription break fbt prm misc	基本形 Break Index 句末境界音調 プロミネンス 注釈情報	これが 2+bp HL% PNLP AYOR
segMora （モーラ）	MoraEntity PerceivedAcc	モーラ記号 アクセント核の有無	ユ 0/1
segPhoneme （音素）	PhonemeEntity	音素記号	kj
segPhone （分節音）	PhoneEntity PhoneClass Devoiced StartTimeUncertain EndTimeUncertain	分節音記号 分節音記号のクラス 無声化の有無 開始位置不明 終了位置不明	kj consonant 0/1 0/1 0/1
pointTone （トーン）	tone F0Uncertain CategoryUncertain PositionUncertain	トーン記号 F0 の不明確さ カテゴリーの不明確さ 位置の不明確さ	HL% 0/1 0/1 0/1

の層（単位）ごとに別々のテーブルで表現している。

RDB 全体は，セグメント・テーブル，サブセグメント・テーブル，親子関係テーブル，リンク・テーブル，メタ情報テーブルの 5 種類のテーブルから構成される。以下，各テーブルの概要について説明する。

3.3.1　セグメント・テーブル

セグメント・テーブルは，図 3 の各単位ごとに談話中の要素を記述したテーブルである。すべてのセグメント・テーブルに共通する属性として，

談話 ID，セグメントの ID，開始時間，終了時間，話者ラベル

の 5 つがある。これらによって，各セグメントの生起位置が一意に特定される。特別な場合として，トーン（アクセントや句末境界音調）のようにある瞬間に生起する（開始時間と終了時間が等しい）要素もある。

表 2　サブセグメント・テーブルの個別記載情報

テーブル名	列名	説明	値の例
subsegLUW	LUWDictionaryForm	代表形	イク
（長単位）	LUWLemma	代表表記	行く
	LUWPOS	品詞	動詞
	LUWConjugateType	活用の種類	カ行五段
	LUWConjugateForm	活用形	連用形
	LUWMiscPOSInfo1	その他情報 1	格助詞
	LUWMiscPOSInfo2	その他情報 2	促音便
	LUWMiscPOSInfo3	その他情報 3	連語
subsegSUW	PlainOrthographicTranscription	タグ無し出現形	行き
（短単位）	PhoneticTranscription	発音形	イキ
	SUWDictionaryForm	代表形	イク
	SUWLemma	代表表記	行く
	SUWPOS	品詞	動詞
	SUWConjugateType2	活用の種類 2	カ行五段 2
	SUWConjugateForm2	活用形 2	連用形 2
	SUWMiscPOSInfo1	その他情報 1	副助詞
	SUWMiscPOSInfo2	その他情報 2	語幹
	SUWMiscPOSInfo3	その他情報 3	言いよどみ
	ClauseBoundaryLabel	節境界ラベル	<テ節>
	CU_preBracket	節単位前ブラケット	<<
	CU_postBracket	節単位後ブラケット	>>
	CU_OperationSign	節単位操作記号	-
	CU_ObligateComment	節単位義務的コメント	体言止め

これらの共通情報に加えて，各単位に固有の情報が記されている。表 1 に各テーブルの個別情報を挙げる。

3.3.2　サブセグメント・テーブル

自発音声では，複数の語が融合して，分割できない一つの要素を形成することがしばしばおこる。例えば，「僕は」が融合して「ボカー」のように発音される場合である。CSJ では，これを「(W ボカー; ボクワ)」のように転記したうえで，形態論情報としては「僕」と「は」の 2 つの要素に分けて記述している。しかし，「僕」と「は」の境界は実際の音声中には存在しないため，開始・終了時間に依拠したセグメントとしては表せない。そこで CSJ-RDB では，単語の階層（長単位と短単位）は一般に，時間的に分節化できる部分をセグメントで表し，時間的に分節化できない部分はその下位にあるサブセグメントとして表している。

サブセグメント・テーブルは，

談話 ID，サブセグメントが帰属するセグメントの ID，

セグメント中のサブセグメントの位置，セグメント中のサブセグメントの総数

の 4 つの属性を共通に持つ。これらの共通情報に加えて，各単位に固有の情報が記されている。表 2 に各テーブルの個別情報を挙げる。

表 3　リンク・テーブルの記載情報

テーブル名	列名	説明	値の例
linkDepBunsetsu （文節係り受け関係）	SourceID DestinationID Dep_Label	係り文節 ID 受け文節 ID 係り受けラベル	3 10 D
linkTone2AP （トーンの帰属先）	SourceID DestinationID	トーン ID 帰属先アクセント句 ID	3 1

3.3.3　親子関係テーブル

親子関係テーブルとは，図 3 に表された階層関係に従って，セグメント間の親子（先祖・子孫）関係を ID の対で表現したものである。例えば節単位を親（先祖）とする親子関係テーブルは，文節，長単位，短単位，モーラ，音素，分節音のいずれかを子（子孫）とするものが合計 6 種類作成される。一方，節単位とアクセント句とは階層関係をなさないため，これらの間には親子関係テーブルは作成されない。

親子関係テーブルには，

談話 ID，親（先祖）セグメントの ID，子（子孫）セグメントの ID，

親セグメント中の子セグメントの位置，親セグメント中の子セグメントの総数

が共通して記されている。

セグメントに基づくアノテーション表現では，親子（先祖・子孫）関係は時間的包含関係から導出できる。しかし，CSJ-RDB では利用の簡便から，親子（先祖・子孫）関係をあらかじめ導出し，テーブルとして提供している。この親子関係テーブルを利用することで，複数の単位に関わる分析（例えば，節単位末尾の短単位冒頭のモーラの時間長）が容易に行える。

3.3.4　リンク・テーブル

セグメント間の関係としては，親子関係以外にも様々なものが考えられる。例えば，文節係り受けは文節同士の間の関係である。CSJ-RDB では，このようなセグメント間の関係をリンク・テーブルで表現している。

現在のところ，リンク・テーブルとしては，「文節係り受け関係」と「トーンの帰属先」の 2 つがある。後者は，韻律ラベルで与えられているアクセントや句末境界音調などのトーンがどのアクセント句に帰属するかを表わしたものである（トーンの生起位置が帰属先アクセント句の範囲をはみ出すことがしばしばあり，時間情報からだけでは帰属先を決められない）。

リンク・テーブルには，

談話 ID，リンク元（source）セグメントの ID，リンク先（destination）セグメントの ID

が共通して記されている。これらの共通情報に加えて，各リンクに付随する情報が記される場合もある。表 3 に各テーブルのリンク元，リンク先の内容と付随情報を挙げる。

3.3.5　メタ情報テーブル

以上のテーブル群に加え，談話や話者など言語単位以外の情報を納めたテーブルが含まれる。表 4 に示すように，談話の基本情報を納めた「談話基本情報」，話者に関する情報を記した「話者基本情報」，個々の談話の各種印象を調査した「印象評定情報」（46 の印象項目を 7 段

表 4　メタ情報テーブルの記載情報

テーブル名	列名	説明	取りうる値 or 例
infoTalk （談話基本情報）	TalkID	談話 ID	S01F0001
	TalkType	談話タイプ	独話/対話/朗読
	Genre	ジャンル	学会/模擬
	SpeakerID	話者 ID	116
	SpeakerAge	話者年齢（5 年刻み）	20to24
infoSpeaker （話者基本情報）	SpeakerID	話者 ID	116
	SpeakerSex	話者性別	男/女
	SpeakerBirthGeneration	話者生年代（5 年刻み）	70to74
	SpeakerBirthPlace	話者出生地	東京都
infoImpression （印象評定情報）	（省略）		

階で評定した結果を収めたもの）がある。

4. おわりに

筆者らは，人文系のユーザが CSJ に付された豊富な情報を使いこなせることを目標に，XML 文書と比べて直感的に構造が把握しやすく，かつ，操作技術の取得がより容易と考えられる RDB によって CSJ のデータを表現することを試みた。

昨年 11 月には，筆者らがそれぞれ代表者として関わる国立国語研究所の共同研究プロジェクトのメンバーを中心に試作版 CSJ-RDB を限定公開し，CSJ-RDB の構成，RDB から情報を抽出するための技法（SQL の書き方）などについて解説する講習会を 2 回開催した。受講者の多くは，いわゆるプログラミングの経験をほとんど持たない人文系の研究者であったが，多くの受講者から「これなら使えそう」という感想を頂いた。

CSJ-RDB を構築する過程で，CSJ 第 3 刷に含まれる様々なバグが発見された。現在，そのバグを修正しつつ，CSJ-RDB の完成に向けて作業を進めているところである。各種検証作業を経た上で，来年度を目途に CSJ 購入者に対する一般公開を予定している。

謝辞　CSJ-RDB の構築に際し，西川賢哉氏（理研），山田篤氏（ASTEM）の協力を得た。ここに記して感謝する。本研究は国立国語研究所萌芽・発掘型共同研究「会話の韻律機能に関する実証的研究」（リーダー：小磯花絵），独創・発展型共同研究「多様な様式を網羅した会話コーパスの共有化」（リーダー：伝康晴），基幹型共同研究「コーパスアノテーションの基礎研究」（リーダー：前川喜久雄），基幹型共同研究「コーパス日本語学の創成」（リーダー：前川喜久雄）による成果である。

参考文献

国立国語研究所 (2006).『国立国語研究所報告書 124: 日本語話し言葉コーパスの構築法』.

伝康晴・小磯花絵 (2012).「RDB と既存のアノテーションツールによる統合的コーパス開発環境」　言語処理学会第 18 回年次大会発表論文集.

関連 URL

『日本語話し言葉コーパス』ホームページ：http://www.ninjal.ac.jp/csj/