〈共同研究プロジェクト紹介〉コーパスアノテーションの基礎研究；コーパス日本語学の創成
「コーパスアノテーションの基礎研究」および「コーパス日本語学の創成」

国語研プロジェクトレビュー
NINJAL Project Review　Vol.3 No.2 pp.63–83（October 2012）

「コーパスアノテーションの基礎研究」および「コーパス日本語学の創成」

"Basic Research on Corpus Annotation" and "Foundation of Corpus Japanese Linguistics"

前川 喜久雄（MAEKAWA Kikuo）

1. はじめに

言語資源研究系では現在 3 本の基幹型共同研究プロジェクトを実施している。ここではそのうち「コーパスアノテーションの基礎研究」および「コーパス日本語学の創成」を紹介する。ただし後者の成果の一部については既に本誌 5 号で紹介しているので，ここでは前者に紙面の多くを割くことにする。執筆にあたっては成果の宣伝だけでなく，現在直面している問題点にも触れることによって研究の現状を率直に伝えることを心掛けた。

2. 「コーパスアノテーションの基礎研究」

2.1 アノテーションの必要性

コーパスにはさまざまなものがある。書き言葉の場合，もっとも単純なのは単にテキストを電子化しただけのコーパスであるが，この種のコーパスには実用上さまざまな制約がともなう。特に日本語では分かち書きの習慣がないことと表記の複雑性とが仇となって語の検索が思うようにならない。

例えば感動詞の「ああ」を検索したいとする。この語には何種類の表記があるだろうか。『現代日本語書き言葉均衡コーパス』の形態論情報解析のために整備した解析用電子辞書 UniDic には 33 種類の表記が登録されている。一部を示すと「ああ, あ～, あ～あ, あ～っ, あぁ, ああ，… アヽ，アァ，アア，アー，アーア，アーッ，吁嗟，嗚乎，嗚呼，嗟，嗟呼，噫，…」といった調子である。一層すさまじいのは人名で，例えば「ひろし」と読む人名には 75 種類の表記がある。「ひろし，ヒロシ，博，博史，博司，博士，博師，博志，博至，博資，啓，啓史，啓志，坦，大，大志，央，宏，宏史，宏司，宏士，…」と延々とつづく。

表記の複雑さに語形変化が重なるともう大変である。UniDic には「沸き起こる」に関する活用形と表記の組み合わせが「わきおこれ，わきおこりゃ，わきおころ，…　わき起これ，わき起こりゃ，わき起ころ，…　沸きおこれ，沸きおこりゃ，沸きおころ，…　涌きおこれ，涌きおこりゃ，涌きおころ，…」等，324 通り登録されている。「飛び上がる」は 277 通り，「引き下がる」は 226 通りである。試みに計算してみると 1 個の動詞には平均して 38.6 個の組み合わせがあった。

例はこれぐらいにしておこう。要するにユーザーがこうした可能性のすべてを考慮したう

えで文字列検索を行わなければならないのであれば，そのコーパスの可用性は非常に低いといわねばならない。

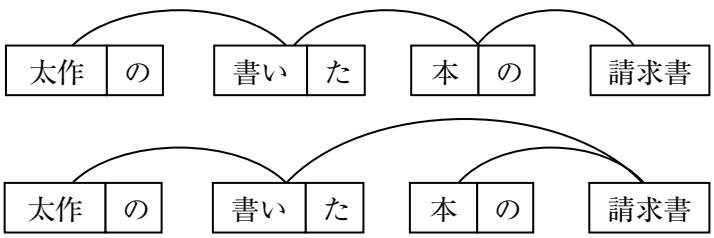
上で論じた問題を解決するために必要になるのがいわゆる形態素解析である。『日本語話し言葉コーパス』（以下 CSJ）や『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）のように形態素解析され形態論情報が付与されたコーパスであれば，語彙素を指定して検索することによって上例のような語を苦もなく一網打尽に検索できる[1]。このようにコーパスからの情報抽出を助けるために情報を付加する作業がコーパスアノテーションである。

形態論情報はもっとも基本的なアノテーションであり，研究の目的に応じて他にもさまざまなアノテーションがある。本プロジェクトでは，コーパスの価値はコーパスの代表性とアノテーションの積として定まるという信念のもとに，BCCWJ のコアと呼ばれる約 130 万語からなるミニコーパスを共通の素材として，プロジェクトメンバーが各自の必要と興味に応じてさまざまなアノテーションを試みている[2]。そのうちのいくつかを紹介しよう。

2.2　係り受け構造

形態論情報は個々の語についての情報を提供するだけで，語の修飾関係についての情報は提供してくれない。現代日本語では動詞の終止形と連体形が合一してしまっているので，連体形の動詞の直後に名詞があっても動詞が名詞を連体修飾しているとはかぎらない。「太作の書いた本」という文字列中の「書いた」は「太作の書いた本の評判」という文脈ならば「本」を連体修飾している可能性だけを考えればよいだろうが,「太作の書いた本の請求書」ならば,「本」を修飾している可能性に加えて「請求書」を修飾している可能性も考慮すべきであろう。

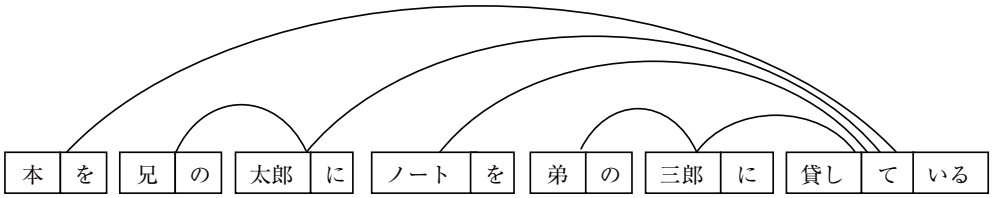
係り受け構造は，このような修飾関係を明らかにするためのアノテーションである。記述の単位としては文節が用いられることが多い。以下の例では矩形が文節，その中の縦線が語境界を表わしており，修飾関係が弧によって示されている。係り受け構造が上のようであれば「書いた」は「本の」を修飾しているが，下のようであれば「書いた」は「請求書」を修飾している。



次の例は自動構文解析が困難なことで知られている並列構造を含む文の係り受けアノテーションである。このような構造に対しては，処理の目的によってこれとは異なるアノテーショ

[1] ちなみに BCCWJ には感動詞「ああ」が 11934 個含まれている。
[2] すべてのアノテーション作業で 130 万語全体を対象としているわけではない。

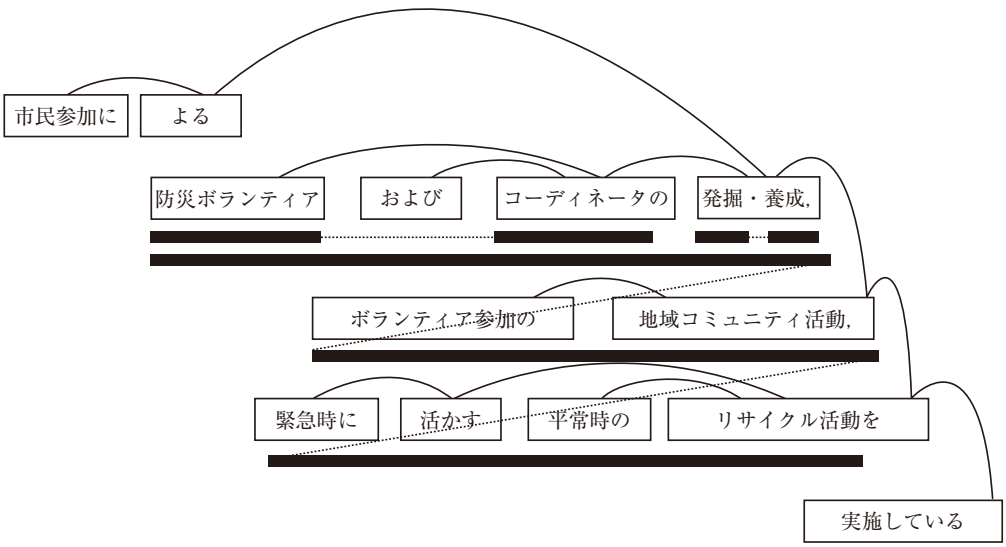
ンの可能性も考えられる（浅原・岩立・松本 2011 参照）。



最後に複雑な実例を示しておく。BCCWJ の白書のサンプルに含まれていた用例である。図中の太線で示されているように「防災ボランティアおよびコーディネータの発掘・養成,」「ボランティア参加の地域コミュニティ活動,」「緊急時に活かす平常時のリサイクル活動」の三者が並列構造をなしており，そのうち最初のものの内部にも 2 種類の並列構造が認められる（並列構造の入れ子）。また最後の要素の末尾にある格助詞「を」は，先行するふたつの並列要素にも共有されていると考えている。

係り受け構造の記述単位は文節に限られているわけではない。語を単位とした分析も可能であるが，煩雑になるのでこれまでは実施されてこなかった。ただし後述する長単位の内部構造の分析（2.7 節参照）を実施するのであれば，係り受けもまとめて語を単位として実施した方がよいかもしれない。このあたりの問題は現在検討中である。

BCCWJ 全体に対する係り受け構造アノテーションは終了している。もちろんコンピュータを用いた自動解析である。問題はデータが巨大すぎて（21 GB ほどある），効率的に検索できない点である。アノテーション情報を実用化するためにはこういう運用面の問題もどこかで検討する必要がある [3]。



[3] 係り受け構造アノテーションの担当は松本裕治（奈良先端大）と浅原正幸（国語研）である。

2.3　拡張モダリティ

日本語の文は命題とモダリティから構成されているとされる。「健太郎は元気じゃないらしいね」であれば,「健太郎が元気」という命題に, 断定（だ）, 否定（ない）, 推量（らしい）, 同意要求（ね）などのモダリティ要素が付加されたと分析される。ここで命題は価値中立的な事象であり，モダリティはそれに対する話し手ないし書き手の意思・判断・態度等であるとされる。

この例がそうであるように, 日本語学ではモダリティを言語形式とむすびつけて分析する。しかし，それだけでは情報が十分でないことが多い。「あの寿司屋はうまい」と「あの寿司屋はうまいと聞く」は語学的にはともに断定のモダリティをもっているが，前者が話し手自身の判断であるのに対して，後者は伝聞した情報を伝達している。聞き手の反応に，前者なら信用するが後者については保留するという違いが生じてもおかしくない。自然言語による情報検索システムや対話理解システムを作る立場からすれば，語学的なモダリティよりも一層広い観点からモダリティに関わる情報を収集する必要があり，そのような情報は人間の情報処理行動の理解にもつながる可能性がある。

このような問題意識から松吉・佐尾・乾・松本（2011）は拡張モダリティのアノテーションを考案している[4]。拡張モダリティは以下の各項目から構成されている。

- **態度表明者**：対象とする事象の成否の判断や，他者への働きかけや問いかけをしている人物, もしくは, 団体。多くの場合, この態度表明者は書き手である。ラベルは「wr：筆者」「wr：筆者 _arb：特定」など 5 種類。
- **相対時**：態度表明時から見た，対象事象の相対的な時間関係。過去・現在のことであるのか，それとも，未来のことであるのかを表す。ラベルは「未来」と「非未来」の 2 種類。
- **仮想**：仮定された条件の有無。仮想世界の話であるのか，そうでないのかを表す。ラベルは「条件」「帰結」「0」の 3 種類。「0」は文に条件が関わっていない場合。
- **態度**：叙述，意志，働きかけ，問いかけなどの伝達的態度。ラベルは「叙述」「意志」「欲求」「働きかけ _ 直接」「働きかけ _ 間接」「働きかけ _ 勧誘」「許可」「問いかけ」の 8 種類。
- **真偽判断**：態度表明者による対象事象の真偽判断。対象事象が成立か不成立かを確信度とともに表す。ラベルは「成立」「不成立」「不成立から成立」「成立から不成立」「高確率」「低確率」「低確率から高確率」「高確率から低確率」「0」の 9 種類。「X から Y」のように判断の変化も記述する。「0」は態度表明者にとって真偽が不明であり，態度表明者の判断が文中に表明されていないことを示す。
- **価値判断**：態度表明者による対象事象の価値判断。対象事象の成立が望ましいことであるかどうかを表す。ラベルは「ポジティブ」「ネガティブ」「0」の 3 種類。

[4] 拡張モダリティの担当は松吉俊（山梨大）と乾健太郎（東北大）である。

以下若干の実例を示す。アノテーションの対象はいずれも BCCWJ からとった実例である[5]。用例中の下線はアノテーションの対象とする事象（命題）の核となる述語を示している。

A）音楽を、ダウンロードしたいのですが、信頼できる、ページを教えていただけますでしょうか？

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr：筆者	未来	0	欲求	0	ポジティブ

※態度表明者「wr：筆者」：態度を表明しているのは書き手（writer）である。
※相対時「未来」：実行の時期は未来である。
※仮想「0」：条件は表明されていない。
※態度「欲求」：態度表明者が自分自身の行為の実行を望んでいる。
※真偽判断「0」：表明されていない。
※価値判断「ポジティブ」：態度表明者が事象成立は望ましいと判断している。

B）日本大学文理学部の佐藤秀夫教授によると、家庭訪問は明治初期に、不就学児を学校に通わせるように親を説得する目的で始まったという。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr：筆者 _arb：不特定	非未来	0	叙述	成立	0

※態度表明者「wr：筆者 _arb：不特定」：対象事象に対する態度表明者が不特定の（arbitrary）個人や集団であると，文章の書き手（writer）が述べている。
※相対時「非未来」：過去の事象である。
※態度「叙述」：事象やそれに対する判断などを情報の受け手に伝えている。
※真偽判断「成立」：態度表明者が事象（命題）が成立すると判断している。

C）気が向いたら戻ります…。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr：筆者	非未来	条件	叙述	成立	0

※仮想「条件」：事象が条件として仮想的に述べられている。

D）参考で試聴できますので確かめてください。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr：筆者	未来	0	働きかけ _ 直接	0	ポジティブ

※態度「働きかけ _ 直接」：態度表明者が直接相手に対して行為の実行もしくは非実行を求めている。
※価値判断「ポジティブ」：事象の実行が望ましいと判断している。

[5] 用例は BCCWJ の表記のまま示す。

E）軽く考えないでください。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr：筆者	未来	0	働きかけ _ 直接	0	ネガティブ

※価値判断「ネガティブ」：事象の不実行が望ましいと判断している。

F）サックスを買ったら何が必要ですか？

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr：筆者	非未来	0	問いかけ	0	0

※態度「問いかけ」：態度表明者にとって不明なことがあるために，その事象に対して態度表明者の判断が成り立たないことを表す。いわゆる疑問文以外に確認要求や疑いの文を含む。

G）ホイットニーの声とか、初心者は出ません。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr：筆者	非未来	0	叙述	不成立	0

※真偽判断「不成立」：態度表明者が事象（命題）が不成立であると判断している。

H）税理士役も俳優さんらしいし、完全なフィクションでしょう…。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr：筆者	非未来	0	叙述	高確率	0

※真偽判断「高確率」：態度表明者が想像や思考により，もしくは外部に存在する情報を観察したり取り入れたりすることにより，事象が成立していると推測していることを表す。

I）彼女がおれを訴えるとさえ言わなかったならば、殺さずにすんだ。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr：筆者	非未来	0	叙述	成立	ネガティブ

※反実仮想の例。態度表明者は事象の成立を望んでいない（価値判断「ネガティブ」）が実際には成立していると判断している（真偽判断「成立」）。

J）なにか、黒二十九の石を挟んで打つべきだった、と。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr：筆者	非未来	0	叙述	不成立	ポジティブ

※事象不成立の後悔・不満の例。成立しなかった事象（真偽判断「不成立」）に関してその成立が望ましいと判断している（価値判断「ポジティブ」）。

2.4　時間情報

情報検索や情報抽出においては，人名・地名・組織名・商品名などからなる固有表現

（named entity）が重要な役割を果たすことは当然であるが，テキストに記載されている事象を時間軸に結びつけて解釈することが必要となることも多い。以下，小西・淺原・前川（2012）で提案されている時間情報のタグを紹介する。

タグづけの対象は日付表現・時刻表現・時間表現・頻度集合表現の 4 種類である。日付表現は「一九二九年二月」「前日」のような日暦に焦点をあてた表現である。時刻表現は「午前十時ごろ」「午後六時ごろ」「昼」「九日昼」のような一日のうちのある時点に焦点をあてた表現である。日付表現と時刻表現の区別は時間軸上の粒度の区別でしかない。時間表現は「その間」のような時間軸上の両端に焦点をあてておらず，期間を表すことに焦点をあてている表現である。頻度集合表現は「毎日」のような複数の日付・時刻・時間に焦点をあてた表現である。この分類は解析の方便のために導入したものである。時間軸上 1 つもしくは複数の時点・時区間を表現するものをタグづけ対象である時間情報表現とする。以下に BCCWJ のサンプルにタグを付与した例を示す。読み易くするために，時間構造タグ（<TIMEX3> と </TIMEX3> で囲われた部分）が付与された部分は改行して表示している。

<sentence type="quasi">
<TIMEX3 tid="t1" type="DATE" value="2003-10-20" valueFromSurface="2003-10-20">
二〇〇三年十月二十日 </TIMEX3>
<TIMEX3 tid="t2"type="DATE" value="2003-10-W3-1" valueFromSurface="XXXX-WXX-1">
月曜 </TIMEX3>
</sentence><br type="automatic original" /> <sentence type="quasi">
<TIMEX3 tid="t3"type="TIME" value="2003-10-20T17:30:XX" valueFromSurface="XXXX-XX-XXT17:30:XX">
午後五時三十分 </TIMEX3>
</sentence> <br type="automatic original" /> <blockEnd /> <paragraph> <sentence> ステイシーはだらけた姿勢でモニターの前に陣取り、白黒の画像に見入っていた。</sentence> <sentence> 彼女は伸びをし、腕時計に目をやった。</sentence> <sentence>
<TIMEX3 tid="t4" type="DURATION" value="PT2H30M" valueFromSurface="PT2H30M">
二時間半 </TIMEX3> で収穫ゼロ </sentence>

タグの属性（<TIMEX3 …> タグの…部分に記入されている情報。tid，type など）についてごく簡単に説明する。XML の約束にしたがって属性の名前は冒頭に @ をつけて表記することにする。

属性 @tid はひとつの文書中におけるタグの通し番号である。@type 属性は DATE，TIME，DURATION，SET の 4 つの値を持つ。それぞれ日付表現・時刻表現・時間表現・頻度集合表現を意味する。

@value 及び @valueFromSurface 属性は時間情報表現が含意する日付・時刻・時間の値を表す。このうち @value は文脈情報を用いて正規化を行った値を付与し，@valueFromSurface 属

性は文脈情報を用いずに文字列の表層表現のみから判定できる値を付与する。

この例では用いられていない要素に @mod があり「時間表現のモダリティ」を表わす。例えば「2000 年以前」をタグづけするために @mod 属性に ON OR BEFORE という値をわりあてることにより「以前」というモダリティを表現する。属性にわりあてる値を表 1 に示す[6]。

表 1　@mod 属性に対する値

値	定義	例
@mod = START	日付時刻表現の初期	「初め」「初頭」
@mod = MID	日付時刻表現の中期	「半ば」「中ごろ」
@mod = END	日付時刻表現の後期	「末」「暮れ」
@mod = APPROX	近似表現	「ごろ」
@mod = BEFORE	日付時刻表現より前	「前」
@mod = AFTER	日付時刻表現より後	「過ぎ」
@mod = ON OR BEFORE	日付時刻表現以前	「以前」
@mod = ON OR AFTER	日付時刻表現以後	「以降」「以来」
@mod = EQUAL OR LESS	時間表現の範囲以下	「以内」
@mod = EQUAL OR MORE	時間表現の範囲以上	「以上」
@mod = LESS THAN	時間表現の範囲未満	「未満」「近く」
@mod = MORE THAN	時間表現の範囲超過	「余り」「過ぎ」

2.5　語義

世界的にみてコーパスアノテーション作業の主要な対象となっているのは形態論情報と統語情報(上に紹介した係り受け構造はそのひとつ)である。反対に意味に関するアノテーションはあまり行われていない。意味にはさまざまなレベルがあるが出発点となるのはやはり語の意味，すなわち語義のアノテーションである。語の意味の捉え方については多くの議論があるが，ここではある語を国語辞典でひいたときに，番号を付けて順に記述されている意味のひとつひとつが語義であると考えることにする。例えば「モデル」という語には

- **模範・手本または標準となるもの。「緑化____地区」**
- **模型。また展示用の見本。「プラスチック____」**
- **ある事象について, 諸要素とそれらの相互の関係を定義化して表わしたもの。「計量経済____」**
- **美術家・写真家が制作の対象とする人やもの。「ヌード____」**

[6] 時間表現の担当は淺原正幸（国語研）である。

等々の項目が並んでいるはずである。コーパスに現れたひとつひとつの「モデル」がこのうちどの意味で用いられているかを決めるのが語義アノテーションである。その際，上に挙げられていない語義が新しくみつかることもある。そのときは例えば「『ファッションモデル』の略」というように新語義を決めることもアノテーション作業の一部に含まれる[7]。

上記のような語義レベルのアノテーションは実質語を対象としたものだが，機能語についても同様の語義アノテーションを行うことができる。例えば日本語の助動詞「レル・ラレル」には受身・尊敬・可能・自発の 4 種の意味があるとされているが，形態論情報のレベルではこれらを区別することができない。「可能」表現の研究や受動文の研究は日本語学上重要な研究テーマであるから，その点でもこれらの語義の区別が望まれる。試行的なアノテーションの結果を表 2 に示す（小山田・柏野・前川 2012）。

表中の数字は生起数である。レジスターを問わず受身が最も多いが，生起数には 6893 から 3914 までかなりの開きが認められる。国会会議録では尊敬が図抜けて多いこと，CSJ の学会講演では可能の生起率が高いことなどが注目される。

日本語の受動文の特色とされる「被害の受身」（典型的には「母に死なれた」「巨人に優勝された」など自動詞の受動文）の生起状況なども今後アノテーションの対象に追加したいと考えている。

表 2 助動詞レル・ラレルの意味アノテーションの試行結果

レジスター		受身	尊敬	可能	自発	決定不能
BCCWJ（コア）	Yahoo！知恵袋	4876	596	447	0	11
	Yahoo！ブログ	3914	248	668	11	0
	白書	6893	5	1233	0	0
	書籍	6861	211	912	44	0
	雑誌	6022	163	836	25	5
	新聞	6399	42	784	13	3
	国会会議録	4408	1447	651	20	8
CSJ	学会講演	5030	175	2087	8	87
	模擬講演	4063	267	809	8	67

2.6 節境界

節（clause）は統語論上の単位であり，ひとつの述語を核としてそれをいくつかの補語や副詞類が修飾している構造をさす。コーパス言語学で節が重要視されるのは，話し言葉コーパス，特に自発音声コーパスにおいては，書き言葉と同様の文を認定することがしばしば困

[7] 実質語の語義は奥村学（東京工業大），レル・ラレルの語義は筆者と柏野和佳子（国語研）が担当している。

難であるのに対して，節については安定して認定することが可能だからである。国語研が開発したコーパスではCSJに節境界アノテーションが施されている。

CSJの節単位アノテーション方式は朗読音声に近いTVの論説番組などのアノテーションのために開発されたものを自発音声用に拡張したものであり（丸山・高梨・内元2006），以下の49種類の節境界を認定したうえで，それらを節境界の統語的従属度の強さに従って「絶対境界」「強境界」「弱境界」の3段階に分類している。詳細は上記文献参照。

・**絶対境界**：[文末]，[文末候補]，[と文末]
・**強境界**：/並列節ガ/，/並列節ケレドモ/，/並列節ケレド/，/並列節ケドモ/，/並列節ケド/，/並列節シ/，/ヨウニ節/
・**弱境界**：＜条件節タラ＞，＜条件節タラバ＞，＜条件節ト＞，＜条件節ナラ＞，＜条件節ナラバ＞，＜条件節レバ＞，＜理由節カラ＞，＜理由節カラニハ＞，＜理由節カラ−助詞＞，＜理由節ノデ＞，＜タリ節＞，＜タリ節−助詞＞，＜テ節＞，＜テハ節＞，＜テモ節＞，＜テカラ節＞，＜テカラ節−助詞＞，＜テ節−助詞＞，＜トカ節＞，＜トカ節−助詞＞，＜ノニ節＞，＜連用節＞，＜引用節＞，＜引用節−助詞＞，＜引用節トノ＞，＜トイウ節＞，＜間接疑問節＞，＜間接疑問節−助詞＞，＜連体節テノ＞，＜並列節ダノ＞，＜並列節デ＞，＜並列節ナリ＞，＜フィラー文＞，＜感動詞＞，＜接続詞＞，＜接続詞C>，＜接続詞L>，＜接続詞CL>，＜接続詞M>

節境界ラベルの例をいくつか示す。すべてCSJからの実例である。例文中のすべての節境界ラベルを表示し，注目しているクラスのラベルを太字（下線）で示している。

A）絶対境界
・簡単に最初に復習をしておきたいと＜引用節＞思います［**文末**］
・あたしニワトリ歩いてるのって見たことが今までなかったんですね［**文末**］
・それにしても＜テモ節＞私が本で読んだあの情報は一体何だったのでしょうか［**文末候補**］
・で＜接続詞＞総論では賛成なんだけれども/並列節ケレドモ/ 少し煮詰める必要があるんではないだろうかと［**と文末**］

B）強境界
・で＜接続詞＞結果ですが/**並列節ガ**/まず絶対音感群の結果から見ていきたいと＜引用節＞思います［文末］
・まずその教室の様子ですけれども/**並列節ケレドモ**/ 教室の名前は俳句文法教室と言います［文末］
・で＜接続詞＞その豪華な船に乗れたっていう＜トイウ節＞体験もできたし/**並列節シ**/楽しかったと＜引用節＞思っています［文末］
・先程申し上げましたように/**ヨウニ節**/ 条件付き確率の式はこのように異なっています［文末］

C）弱境界
・で＜**接続詞**＞その時にちょうど先生がもしかしたら＜**条件節タラ**＞私は東京都指定の特殊な難病かもしれないと＜**引用節**＞言われました［文末］
・翌日朝食を済ませると＜**条件節ト**＞私達は旅行代理店へ向かいました［文末］
・もし形式的にこの会話の中で起きた他の相づちに習うならば＜**条件節ナラバ**＞ここではそうかなそうだよといったものになるはずなるのではないでしょうか［文末候補］

・で＜**接続詞**＞ついでにこれもいい日本語に替えてくれれば＜**条件節レバ**＞いいなと＜引用節＞そんな気がしてます［文末］
・どうしてもこの遊びの方が先に入ってたから＜**理由節カラ**＞そっちを優先しちゃったっていう＜**トイウ節**＞ことでした［文末］
・縮むことがないので＜**理由節ノデ**＞安心して＜**テ節**＞何度でも洗います［文末］
・ある時は失恋して＜**テ節**＞泣いている私に寄り添ってもくれました［文末］
・もう初日から彼らを探しては＜**テハ節**＞きゃあきゃあ騒ぎ過ぎて＜**テ節**＞もう声枯れしちゃって＜**テ節**＞大変でした［文末］

節境界の情報は，話し言葉同様，書き言葉の研究にも有用である。BCCWJ のサンプルのなかにはそのままでは 1 文が長すぎて情報抽出・情報検索に不適当なものがある。以下は白書の一例である。

(1) 農業用水の確保及び水利用の安定と合理化を図るとともに、水田の汎用化に資する観点から、ほ場条件の整備の前提である基幹かんがい排水施設の体系的な整備を進めるため、国営かんがい排水事業を実施し、都道府県営かんがい排水事業の実施の推進等を行うほか、特に、食料の安定供給の確保を図るため、優良農業地域における複数の基幹的水利施設について最適な整備年次計画を策定するとともに、農業水利施設の計画的・機動的な整備・更新を実施し、広域の優良農業地域の持続的な保全を図る食料供給広域基盤確立対策を推進した。

またウェブから収集したサンプルには句点を用いずに書かれているものが多数見つかる。

(2) スゴく気になる..（笑） カワイイ声してたからねぇ.. 最初の頃は小倉さんにメッチャ泣かされたらしいけどホントに何やってんでしょ？
(3) カレーやシチューの缶詰の方がよほどうれしいな〜なんて、、、ずうずうしいですねｗ
(4) どうでしょうで、ミスターはいつも甘い物が嫌いなのに無理矢理食べていますが、大泉さんも少し食べていますが、甘い物がきらいっぽい？

このようなサンプルに対して節境界アノテーションが施されていれば，検索が非常に楽になる。BCCWJ のコアだけにでも節境界アノテーションを施せるとよいのだが，そのためには解決しなければならない問題がある。ひとつは日本語の書き言葉の述語句末がきわめて多様であるため，節境界の認定規則を書きにくいという問題，もうひとつは書き言葉に特有の挿入構造をどう処理するかという問題である。以下の白書の例では名詞句（「首位事業者を含む２以上の主要事業者」）と格助詞「が」の間に丸括弧に入れて名詞句を説明する文が挿入されている。このような構造は学術論文のなかではしばしば用いられるものだが，全体を単一の階層構造にまとめあげることが困難である。言語学が前提とする言語の線状性が成立しない例とみなしてよいのかもしれない[8]。この種の「破格な」構造を処理するためのガイドラインを作ることが当面の課題であるが，これがなかなか難問であり，今後の進展にまつと

[8] こういう文が文法理論でほとんど論じられることがないのはどうしてだろうか。

ころが大きい[9]。

(5) 独占禁止法第１８条の２の規定により、年間国内総供給価格が６００億円超で、かつ、上位３社の市場占拠率の合計が７０％超という市場構造要件を満たす同種の商品又は役務につき、首位事業者を含む２以上の主要事業者（市場占拠率が５％以上であって、上位５位以内である者をいう。）が、取引の基準として用いる価格について、３か月以内に、同一又は近似の額又は率の引上げをしたときは、当委員会は、当該主要事業者に対し、当該価格の引上げ理由について報告を求めることができる。

2.7　長単位と短単位

ここでもう一度形態論情報について考える。日本語はいわゆる膠着語タイプの言語なので，語とは何かがはっきりしない面がある。「現代日本語書き言葉均衡コーパス」は 1 語だろうか。そうでないならば何語だろうか。唯一の正解はない。研究者にきいても十人十色の回答になる。自分の研究目的に都合のよい単位を考案するからである。

実際，国立国語研究所による長年にわたる計量的語彙調査の歴史をふりかえってみると，調査の目的によって短めの単位を設計したり，長めの単位を設計したりしている。日本語の基礎語彙を調べたければ短めの単位が好適だろうし，人名・地名・組織名・商品名などの固有表現（named entity）を調べたければ長い単位が好適である。

繰り返すが，短い単位と長い単位に理論上の優劣はない。問題が生じるのはこれらを無反省に混在させた場合である。現在よく利用されている形態素解析用辞書のなかには，「国立国会図書館」を 1 語に認定するのに，「国立公文書館」は 3 語に分析するようなものがある。

言うまでもないことだが，こういうシステムを使って日本語学を研究することはできない。そこで国立国語研究所では BCCWJ の開発と並行して UniDic と呼ばれる形態素解析用辞書を構築した（伝ほか 2007）。これは短単位と呼ばれる比較的に短めの単位（およそ国語辞典の見出し語程度の長さである）に一貫して依拠した辞書であり，「国立国会図書館」も「国立公文書館」もともに 4 単位に分析する。

近年国語研で開発した CSJ と BCCWJ の特徴は，短単位に依拠した形態素解析を施しているだけでなく，さらに長単位と呼ばれる長めの単位を用いた二重の形態素解析を実施している点である。「国立国語研究所」は短単位では「国立｜国語｜研究｜所」の 4 単位に分割される。「国立国会図書館」は「国立｜国会｜図書｜館」，「国立公文書館」は「国立｜公｜文書｜館」でともに 4 単位である。そして長単位としてはこれらの語はすべて 1 語に認定される。

このような二重形態素解析は日本語学研究における形態論情報の利便性を大幅に向上させたと自負しているが，問題もある。「公害紛争処理法における公害紛争処理の手続は」という例を考えてみよう（やはり BCCWJ からの実例である）。この例を解析すると，「公害紛争処理法」と「公害紛争処理」がそれぞれ別の長単位として析出される。さらにこの例を含むサンプル全体では「公害紛争」「公害紛争処理制度」「公害紛争事件」「公害紛争処理機関」「公

[9] 節境界情報は丸山岳彦（国語研），文の認定は柏野和佳子（国語研）の担当である。

害紛争処理情報」等々の長単位が得られる。

問題はこれらの語に共有されている「公害紛争」と残りの文字列との関係が不明である点である。「公害紛争処理」と「公害紛争事件」とでは明らかに関係が異なっているが，現在の二重解析結果はそれを捉えることができていない。今後は長単位を構成する短単位の相互関係（長単位の内部構造）のアノテーションを考案することが望まれる。「アノテーション」プロジェクトでもこの問題にとりくんでいるので，いずれ成果を報告できると期待している[10]。

2.8 その他のアノテーション

本稿では紹介しきれなかったアノテーションも多数ある。小原京子（慶応義塾大）を中心とした日本語フレームネットのアノテーション（小原 2011），竹内孔一（岡山大）による述語項構造のアノテーション（上野・竹内 2012），宇津呂武仁（筑波大）と松吉俊（山梨大）による複合辞（拡張機能表現）のアノテーション（宇津呂ほか 2010），佐野大樹（情報通信研究機構）による日本語アプレイザル辞書の開発（佐野 2012），そして小磯花絵（国語研），伝康晴（千葉大）と筆者による日本語自発音声アノテーション体系の見直し作業（小磯・伝・前川 2012）などである。これらは機会をあらためて紹介することにしたい。

2.9 アノテーションデータの公開

本プロジェクトで整備した各種アノテーションデータは，作業用マニュアルとともにプロジェクト終了時までに公開する予定である。データは無償公開する予定であるが，BCCWJのサンプルは著作権の関係で無償公開できないので，利用希望者には別途 BCCWJ-DVD 版を購入してもらい，両者をあわせて利用してもらうことになる予定である。

3. 「コーパス日本語学の創成」

残された紙幅で筆者がリーダーを務めるもうひとつの基幹型プロジェクト「コーパス日本語学の創成」に触れる。

3.1 「講座日本語コーパス」と「コーパス日本語学ワークショップ」

プロジェクト名称に端的に示されているように，これは種々のコーパスを用いた研究を日本語研究の世界に定着させることを目標とした戦略的プロジェクトであり，特定の科学研究上の目標を達成しようとする通常の研究プロジェクトとは趣を異にしている。

この戦略的な目標を達成するために，共同研究者は大きく「語彙・文法・文体・表記研究」（歴史研究を含む），「音声・対話研究」の 2 グループに分かれて，KOTONOHA 計画で開発した一連のコーパスを初めとする各種コーパスを利用した先進的な日本語研究を推進している。

[10] 長単位・短単位の関係の分析は小椋秀樹（立命館大）と森信介（京都大）が担当している。

ただし，それだけでコーパス日本語学という新領域が確立されるとは考えられないので，この領域に興味をもつ若手研究者を読者に想定した講座本の刊行を企画した。幸い『講座日本語コーパス』（全 8 巻）として朝倉書店からの刊行が決まり，現在来年度からの刊行にむけて準備を進めている。

講座のタイトルが「コーパス日本語学」ではなく「日本語コーパス」となっているのは，コーパスの設計や構築技術など，コーパスそのものが講座の主要な対象であることを示している [11]。本講座の構成を以下に示す。（　）内は各巻の編者である。1 巻「コーパス入門」（前川），2 巻「書き言葉コーパス―設計と構築―」（山崎誠），3 巻「話し言葉コーパス―設計と構築―」（小磯花絵），4 巻「コーパスと国語教育」（田中牧郎），5 巻「コーパスと日本語教育」（砂川有里子），6 巻「コーパスと日本語学」（田野村忠温），7 巻「コーパスと辞書」（伝康晴・荻野綱男），8 巻「コーパスと自然言語処理」（松本裕治・奥村学）。

本講座の刊行のほかに，本プロジェクトの重要な貢献として位置付けられるものにコーパス日本語学に関する実質的な学会機能の提供がある。これは一般からの応募も可能な公開研究会を開催して，幅広くコーパス日本語学の活動を支援しようとする試みである。

プロジェクト開始当初の 2 年間（2009，2010 の両年度）は，本プロジェクトと並行して，文科省科学研究費特定領域研究「日本語コーパス」を推進していたので，特定領域研究の研究成果発表会を優先して開催したが，2011 年 8 月の「『現代日本語書き言葉均衡コーパス』完成記念講演会」をもって特定領域関係の研究成果発表会は打ち止めとなり，2011 年度末からは本プロジェクトが主催する「コーパス日本語学ワークショップ」を年に 2 回，原則として 3 月と 9 月に開催している。

このワークショップでは言語資源研究系で進行中の全共同研究プロジェクトの成果が発表されるが，共同研究に参加しておらずとも，コーパス利用に興味をもつ研究者であれば誰でも応募が可能である。実際には 1 回あたり 40 件から 50 件の研究発表があり，そのうち 4 割前後が一般公募による研究発表である。発表論文の大部分はホームページからダウンロードできるようにしているが，プロジェクトメンバーにはワークショップで発表した研究を最終的には既存各学会の査読誌に投稿することを推奨している。

3.2　コーパスを用いた音声研究

本プロジェクトで実施されている研究のテーマは日本語学の全体におよんでおり，また個々の研究が相互に有機的に関連づけられているわけでもない。そのため本稿前半でとりあげた「アノテーション」プロジェクトのような形で研究を紹介することは困難である。また敢えて試みても本レビューの主旨にかなうとは思えない。そこで以下では筆者自身の最近の研究成果を紹介することでプロジェクト紹介の責を果たすことにしたい。

筆者の本来の専門は音声学であり，近年は『日本語話し言葉コーパス』を用いた日本語自発音声の研究を主要な研究テーマにしている。その成果の一部である日本語有声子音の変異

[11]　もともとは 2006 ～ 2010 年度に筆者が領域代表者となって実施した文科省科学研究費特定領域研究「日本語コーパス」の成果普及のために案出された企画であったが，現在は本プロジェクトの成果として位置付けている。

ないし弱化に関する音声学的分析は本誌第 5 号の記事で紹介しているので（前川 2011b），今回は趣向を変えて韻律特徴の分析について報告する。

3.2.1 Penultimate Non-Lexical Prominence

日本語の句末に多様なイントネーションが生じる。句末イントネーションには下降調，上昇調，上昇下降調などがあり，これらのそれぞれが複数の変種をもっているので，全体としては非常に多彩である。また句末イントネーションは「耳につきやすい」音声現象であり，発話の印象形成に大きく影響することが知られている（例えば前川 2011a の 8 章参照）。

今回とりあげたのは上昇下降調イントネーションの変種で，PNLP（Penultimate Non-Lexical Prominence）と呼ばれるイントネーションである。普通の上昇下降調イントネーションでは上昇と下降が句の最後のモーラ内部で生じるのに対し，PNLP では上昇の起点が次末モーラ（句末から数えて 2 番目のモーラ：penultimate mora）の冒頭付近に，上昇の頂点（下降の開始点）が次末モーラと句末モーラの境界にそれぞれ位置している点が異なっている。

いま「キーワードで」という名詞句を例にとり，ピッチの上昇と下降を記号「と」であらわすことにすれば，通常の上昇下降調は「キーワード「デ」」，PNLP は「キーワー「ド」デ」のように聞こえる。PNLP の存在を最初に報告したのは設立間もない国立国語研究所に勤務していた大石初太郎であったが（大石 1959），大石はこれを発話の局所的な強調（卓立）であるプロミネンスと解釈して，「あと高型」のプロミネンスと命名している。

この観点はその後多くの研究者に継承されたが，大石以来の研究では，PNLP の生起が何に由来しているかという根本問題が検討されていない。ある発話に PNLP が生じたとき，それが他のイントネーション（上昇調や通常の上昇下降調）ではなく PNLP として実現されているのはランダムな変異なのか，それとも何らかの言語的機能の相違に由来する変異なのかという問題である。

3.2.2 CSJ-Core の分析

PNLP の生起条件を東京語の自発音声コーパスである『日本語話し言葉コーパス』（CSJ）を利用して分析した。実際に分析したのは X-JToBI による詳細な韻律アノテーションが施された CSJ-Core（44 時間，50 万語）である。CSJ-Core には上昇下降調イントネーションが 10225 回生じており，そのうち約 1 割にあたる 1026 件が PNLP であった。CSJ の主要レジスターである学会講演と模擬講演（原稿なしでのいわゆる「スピーチ」）との関係をみると，通常の上昇調が模擬講演に多く生じるのに対して，PNLP は学会講演の方に多く生じている。

最初に，発話の長さが変動したとき，1 発話中での各種イントネーションの平均生起数がどのように変動するかを検討した結果を図 1 に示す。横軸は発話を構成するアクセント句の数であり，発話の長さを示している。縦軸は句末イントネーションの平均生起数である。分析対象とした句末イントネーションは，上昇調（図では H%），通常の上昇下降調（図では HL%），そして PNLP である。ここで各イントネーションの生起率の計算法には少し注意が必要である。生起率の分母となっているのは「当該イントネーションが最低 1 個は生起して

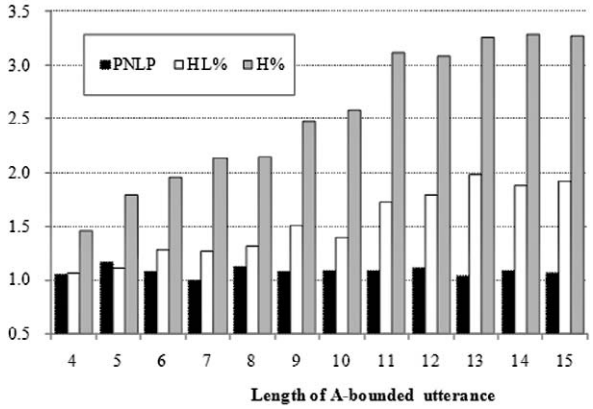


図 1　発話長ごとの句末イントネーションの平均生起数
（各イントネーションの頻度 1 以上の発話が対象。横軸の単位はアクセント句）

いる発話の総数」であり，分子は「それらの発話全体での当該イントネーションの総生起数」である。したがって定義上，平均生起数が 1.0 を下回ることはない。

図 1 で，上昇調と通常の上昇下降調の平均生起率は，発話長の増大とともにほぼ単調に増大している。例えばもっとも長い 15 アクセント句からなる発話であれば，ひとつの発話に上昇調イントネーションならば 3 個以上，上昇下降調イントネーションならば 2 個弱が含まれている。

一方，PNLP の平均生起率は発話長の影響を全くと言ってよいほど蒙っていない。5 アクセント句から 15 アクセント句まで，どの長さの発話においても PNLP の生起率は常に 1.1 前後である。これは PNLP が原則として 1 個の発話に 1 個しか生じないことを示しており，PNLP には言語学で言う頂点機能（ある言語単位のまとまりを示す機能）が認められることを示している。

それでは PNLP は発話中のどのような位置に生じているだろうか。図 2 は長さが 5 アクセント句から 10 アクセント句までの発話における各アクセント句位置での PNLP の生起率を百分率で示したものである。どの長さの発話においても，PNLP 生起率は第 1 アクセント句においては低く，第 2 アクセント句以降において漸次上昇して次末アクセント句において最高値に達し，最終アクセント句ではゼロに近い数字まで急下降するというパタンを示している。図 3 は長さが 11 から 15 アクセント句の発話について同一の分析を施した結果であるが，同一のパタンが生じていることがわかる。図 3 では見やすさのために第 1〜第 5 アクセント句の生起率を表示していないが，この部分のデータにも上記の傾向に反する挙動は全く観察されない。

図 2，3 は PNLP に発話の終端を予告する機能が備わっている可能性を示唆している。図 1 の結果とあわせて解釈すると，発話の頂点を示すと同時に発話の終端を予告することが

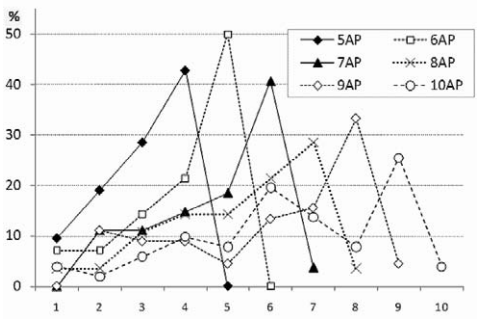


図 2　発話における PNLP の位置別生起率
（発話長 5 から 10 アクセント句まで。
横軸はアクセント句の位置）

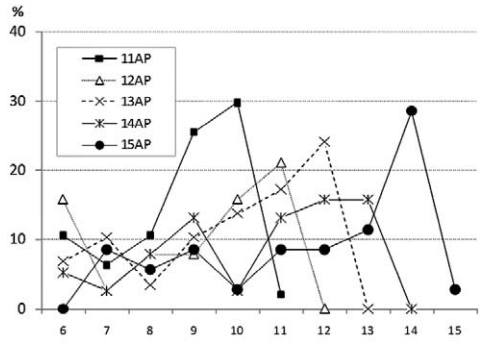


図 3　発話における PNLP の位置別生起率
（発話長 11 から 15 アクセント句まで。
横軸はアクセント句の位置）

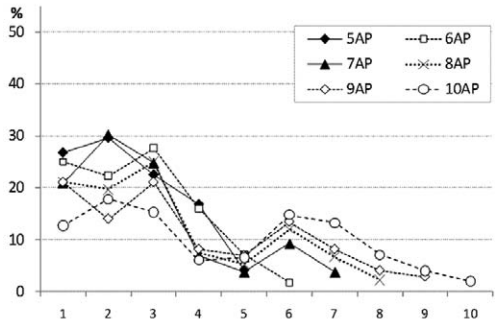


図 4　発話における上昇下降調句末イントネーション（HL%）の位置別生起率
（発話長 5 から 10 アクセント句まで。
横軸はアクセント句の位置）

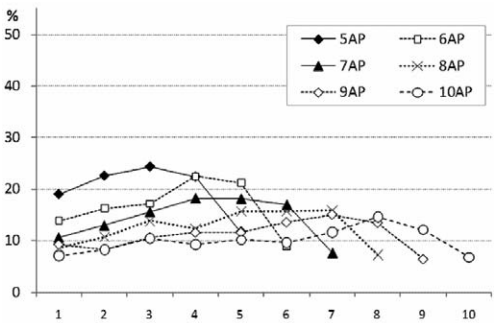


図 5　発話における上昇調句末イントネーション（H%）の位置別生起率
（発話長 5 から 10 アクセント句まで。
横軸はアクセント句の位置）

PNLP の機能であるということになる。これが PNLP 固有の機能だとするならば，他の句末イントネーションには，その種の機能が備わっていないことを示す必要がある。

図 4 は通常の上昇下降調（HL%）について，図 2 と同一の分析を施した結果である。生起率の分布は PNLP とは全く異なっている。次末アクセント句で生起率がピークに達する傾向は全く認められず，長い発話では 2 峰性の分布が観察される。図 5 は上昇調（H%）について，図 2 と同一の分析を施した結果である。ここでは発話長によらず一様に近い分布が観察され，PNLP とは全く異質の分布パタンである。

3.2.3　PNLP に関する結論

CSJ-Core の分析から PNLP は自発音声中の発話単位の頂点を示すと同時に発話の終端が近いことを予告する緩やかな境界表示機能をもった句末イントネーションであることがわかっ

た。このような機能は通常の上昇下降調にも上昇調にも全く認められないことから，PNLP を上昇下降調の自由異音と認定するのは正しくないとの結論が導かれる。PNLP は談話に関わる機能をもって東京語の音声コミュニケーションに独自に寄与している特殊な句末イントネーションである。

なお PNLP を生成している話者は，通常の句末イントネーションの場合とはちがって，自分がそのようなイントネーションを用いていることを意識していない可能性が高い。そのため朗読音声を用いて PNLP を検討することは，データ収録の時点で既に大きな困難にぶつかることが予想される。そのため自発音声の分析が必要になるのだが，自発音声の分析にはコーパスが必須であることを考えれば，PNLP の分析は音声研究における自発音声コーパスの必要性を示す典型的な事例であると言ってよかろうと思われる [12]。

4. おわりに

以上，筆者がリーダーを務めるふたつの基幹型プロジェクトの進捗状況を紹介した。片やコーパスの可用性向上を目標とするプロジェクト，片やコーパス日本語学全体の振興を目標とするプロジェクトで，両者の色合いは著しく異なっているが，日本語のコーパス言語学的研究の現状においては，ともに欠くことのできないプロジェクトだと考えている。

これらのプロジェクトでカバーできていない問題としては，いわゆる文系の言語研究者を対象としたコンピューターリテラシーの教育プログラムの立案がある。この方面の活動についても，国語研のチュートリアルを利用するなどして，試行錯誤を始めたところである [13]。

●参照文献●

浅原正幸・岩立将和・松本裕治（2011）「BCCWJ コアデータへの係り受け・並列構造アノテーション」『「特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』317−324.

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22: 101−122.

小磯花絵・伝康晴・前川喜久雄（2012）「『日本語話し言葉コーパス』RDB の構築」『第 1 回コーパス日本語学ワークショップ予稿集』393−400.（http://www.ninjal.ac.jp/event/specialists/project-meeting/m-2011/jclws01/）

小西光・浅原正幸・前川喜久雄（2012）「『現代日本語書き言葉均衡コーパス』に対する時間情報アノテーション」『第 2 回コーパス日本語学ワークショップ予稿集』317−324.

前川喜久雄（2011a）『コーパスを利用した自発音声の研究』東京工業大学博士論文.

前川喜久雄（2011b）「/z/ の調音様式の変異―コーパスによる分析―」『国語研プロジェクトレビュー』5: 21−45.（http://www.ninjal.ac.jp/publication/review/05/）

前川喜久雄（2011c）「PNLP の音声的形状と言語的機能」『音声研究』15（1）: 16−28.

丸山岳彦・高梨克也・内元清貴（2006）「第 5 章 節単位情報」『日本語話し言葉コーパスの構築法』（国

[12] 本節で紹介した PNLP の分析の詳細については前川（2011c）を参照のこと。この論文は平成 24 年度の日本音声学会優秀論文賞を受賞している。

[13] http://www.ninjal.ac.jp/event/specialists/tutorial/ninjal-t005〜6/　参照。

立国語研究所報告 124），255−322.
松吉俊・佐尾ちとせ・乾健太郎・松本裕治（2011）「拡張モダリティ体系の設計と BCCWJ へのアノテーション」『特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集』403−410.
小原京子（2011）「日本語フレームネットの全文テキストアノテーション：BCCWJ への意味フレーム付与の試み」『言語処理学会第 17 回年次大会予稿集』703−704.
大石初太郎（1959）「プロミネンスについて―東京語の観察にもとづく覚書―」『ことばの研究（国立国語研究所論集）』1: 87−102.
小山田由紀・柏野和佳子・前川喜久雄（2012）「助動詞レル・ラレルへの意味アノテーション作業経過報告」『第 2 回コーパス日本語学ワークショップ予稿集』59−68.
佐野大樹（2012）「アプレイザル理論を基底とした評価表現の分類と辞書の構築」『国立国語研究所論集』3: 53−83.（http://www.ninjal.ac.jp/publication/papers/03/）
上野真幸・竹内孔一（2012）「動詞語義及び意味役割付与作業システムの構築」『第 2 回コーパス日本語学ワークショップ予稿集』69−76.
宇津呂武仁・松吉俊・土屋雅稔・鈴木敬文・島内蘭（2010）「自然言語処理における日本語機能表現の解析」『語彙・辞書研究会第 38 回発表会資料集』1−8.

《要旨》 本稿の前半では基幹型研究「コーパスアノテーションの基礎研究」の現状を紹介した。このプロジェクトでは，既存コーパスの利用価値を向上させるために必要とされるさまざまな言語的アノテーションについての研究を進めている。本稿ではそのうち，係り受け構造，拡張モダリティ，時間情報，語義，節境界，形態論情報をとりあげて解説した。

本稿の後半ではもうひとつの基幹型研究「コーパス日本語学の創成」を紹介した。このプロジェクトはコーパス日本語学の振興を直接の目的とする戦略的プロジェクトである。振興のための主要な手段として位置付けている「講座日本語コーパス」と「コーパス日本語学ワークショップ」について説明した後，具体的な研究成果の一例として，『日本語話し言葉コーパス』（CSJ）を用いた日本語イントネーション研究の事例を紹介した。

PNLP と呼ばれる東京語の韻律特徴は，1959 年に発見されて以来現在までその言語的機能が不明のままであった。今回，X-JToBI 韻律アノテーションの施された CSJ-Core のコンピュータ分析によって，PNLP は原則として 1 発話に 1 回だけ生じて発話の頂点を表示するとともに，典型的には発話の次末アクセント句に生じて発話の終端を予告する境界機能をあわせもっていることが判明した。

Abstract: The first half of the paper is devoted to the introduction of a core research project: “Basic Research on Corpus Annotation.” This project is concerned with the development of various linguistic annotations for the improvement of existing corpora, among which the following annotations are discussed in the present paper, viz., dependency-structure annotation, expanded modality annotation, time-information annotation, annotation for word disambiguation, clause boundary annotation, and morphological annotation.

The second half of the paper is concerned with another core research project entitled “Foundation of Corpus Japanese Linguistics.” This is a ‘strategic’ project aiming directly

at the promotion of corpus-based Japanese linguistics. The main promotion methods are the publication of the eight-volume “Japanese Corpus” series and the organization of a semi-annual “Japanese Corpus Linguistics Workshop.” Following the explanation of these two efforts, a study of Japanese intonation is introduced as an example of the work supported by this strategic project.

The prosodic phenomenon known as the PNLP (‘penultimate non-lexical prominence’) was discovered in 1959, but its linguistic function has remained unclear to the present day. Computer-based analysis of the X-JToBI annotated CSJ-Core revealed that PNLP occurred, in principle, only once in an utterance. It also turned out that PNLP occurred typically in the penultimate accentual phrase of an utterance. From these findings it can be concluded that PNLP has both cumulative and delimitative functions.

前川 喜久雄（まえかわ・きくお）

国立国語研究所言語資源研究系教授，系長，コーパス開発センター長。博士（学術）（東京工業大学）。鳥取大学講師，国立国語研究所研究員，主任研究官，室長，領域長等を経て，2009 年 10 月より現職。

主な著書・論文：「日本語有声破裂音における閉鎖調音の弱化」（『音声研究』14(2)，2010），Prominence marking in the Japanese intonation system（共著，*The Oxford handbook of Japanese linguistics*. Oxford University Press，2008），Coarticulatory reinterpretation of allophonic variation: Corpus-based analysis of /z/ in spontaneous Japanese（*Journal of Phonetics*，38(3)，2010）.

社会活動：日本音声学会理事（企画委員長），一橋大学大学院連携教授，Editorial board member of *Phonetica*.

基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」
プロジェクトリーダー　前川喜久雄（国立国語研究所　言語資源研究系　教授）

プロジェクトの概要

1. 目標：既存のコーパスをより高度に活用するために必要とされる研究用付加情報（アノテーション）の基礎研究を行う。
2. 方法：係り受け構造，述語項構造，節境界，時間情報，各種語義タグ，事実性（モダリティ），複合辞タグなどについてタグの仕様を検討するとともに，既存のコーパスを用いたタグ付与実験を行い，自動アノテーションの可能性についても検討する。基本的に書き言葉コーパスを対象とするが，話し言葉のアノテーションについても検討する。
3. 期待される成果：日本語の高水準アノテーションの標準化をめざす。

基幹型共同研究プロジェクト「コーパス日本語学の創成」
プロジェクトリーダー　前川喜久雄（国立国語研究所　言語資源研究系　教授）

プロジェクトの概要

本プロジェクトは，大規模なコーパスを用いた新しい言語研究の方法を日本語研究の世界に定着させることを目標とした戦略的プロジェクトであり，特定の科学研究上の目標を達成しようとする通常の研究プロジェクトとはやや趣を異にしている。この戦略的な目標を達成するために，国語研の KOTONOHA 計画で開発した一連のコーパスを初めとする各種日本語コーパスを利用した先進的な研究を「語彙・文法・文体・表記研究」「歴史研究」「音声・対話研究」の 3 領域で推進するとともに，一般応募も可能な公開研究会を開催して成果を普及する。 コーパスを利用した日本語研究は研究方法が確立されていないために，定量的な言語研究になじんでいない研究者の新規参入が困難である。本研究が着実な成果をあげることによって，この参入障壁を軽減させる効果が期待される。またコーパスの利用は，従来研究の再現性を欠く傾向にあった言語学の研究に追試可能性を導入することになり，その意味で言語学の科学性を高める効果が期待できる。