複文構文プロジェクト 中間報告「コーパス言語学」

丸山 岳彦 *
国立国語研究所 言語資源研究系

2012 年 12 月 16 日（日）
シンポジウム「複文構文の意味の研究」† （於：国立国語研究所）

1 はじめに

本発表では、コーパス言語学の立場から、現代日本語の複文構文に関する研究を見渡してみたい。ここでは、実際に書かれたり話されたりした日本語の原資料を大量に収集し、組織的に整理したものを「日本語コーパス（Japanese Corpus）」と呼ぶことにする。そして、日本語コーパスを用いて日本語を研究する方法論を、「コーパス日本語学（Japanese Corpus Linguistics）」と呼ぶことにする。

コーパス日本語学がその射程に収める研究分野には、文字・表記、語彙・意味、文法、文章・談話、会話分析、言語変異、言語変化など多くのものが含まれる。いずれの場合でも、当該の研究を行なうために適切な日本語コーパスを準備し、かつ適切な方法によって用例を検索し、分析を行なう必要がある。さらに、定量的な分析と考察の結果を言語学的な観点から一般化し、一言語としての日本語の特性を記述する研究へと昇華できることが望ましい。すなわち、コーパス日本語学の発展には、高品質な日本語コーパスの整備と、よりよい分析方法の開発が常に求められることになる。

そこで以下では、日本語コーパスの開発史を辿りながら、コーパス日本語学の研究、および複文構文の研究がどのように行なわれてきたかについて見ていくことにする。さらに、現段階におけるコーパス日本語学をめぐる状況と、今後の見通しについて述べる。

2 日本語コーパスの開発史と複文構文の研究

2.1 黎明期の日本語コーパスと複文構文の研究

日本語コーパスに基づく複文研究の嚆矢の一つに、三尾 (1942) のいわゆる「丁寧化百分率」調査を挙げることができる。よく知られているように、三尾は 13 編の戯曲を調査対象として、「です体」の中に現れた従属節の諸形式が丁寧化する割合を調べ、表 1 のようにまとめた。

表 1: 三尾 (1942) による「丁寧化百分率」調査

	総数	だ体	です体	半終止等	丁寧化率
から	550	63	169	318	73%
が	327	13	223	91	94.5%
ら	177	160	10	7	6%
けれど	120	13	83	24	86%
と	112	102	8	2	7.3%
ので	103	56	22	25	28%
もの	71	5	0	66	—
り	67	67	0	0	0%
し	52	20	28	4	58%

※三尾 (1942) pp.280–281 より、一部改変

調査対象全体の規模（語数）は不明であるが、各接続助詞の総数から見ても、決して小規模の調査であるとは言えないだろう。戯曲中の「です体」とは、当然、話し言葉を想定したものである。この点を考えると、三尾の調査は、現代で言う話し言葉コーパスを対象として数量的な分析を行なった複文研究の先駆として位置づけることができる。

*maruyama@ninjal.ac.jp
†国立国語研究所共同研究プロジェクト「複文構文の意味の研究」（プロジェクトリーダー：益岡隆志）

1948 年、国立国語研究所が設置された。初期の研究活動の中で実施されたのは、雑誌や新聞など生活に密着した書き言葉、あるいは日常の話し言葉から一定量のサンプルを収集し、そこに見られる用語・用字、語彙、文法・文型を記述する仕事であった。当時の研究報告書を 2 つ挙げておく[1]。

(1) a. 国立国語研究所 (1951)『現代語の助詞・助動詞 —用法と実例—』

b. 国立国語研究所 (1960、1963)『話しことばの文型 (1) (2)』

(1a) は、1949 年から 1950 年に発行された 6 種の新聞、28 種の雑誌から約 48,000 例の助詞・助動詞の実例を抽出してカード化し、その用法を子細に分類した研究である。執筆者の永野賢は、翌 1952 年に「「から」と「ので」とはどう違うか」という著名な論文 (永野, 1952) を発表するが、本報告書の時点ですでに「から」と「ので」の違いに関する記述が見られる。いわゆる「カラとノデ」問題を扱った古典的な論文が、大量の実例の観察から得られたものであるという点に注目しておきたい。

(1b) は、約 29 時間分の対話、約 9.5 時間分の独話を録音し、そこに見られる表現意図、構文、イントネーションなどを整理して「総合文型」を網羅的に記述することを試みた研究である。ここで着目すべきは、2 巻の「III 構文」（南不二男と鈴木重幸が執筆）の中で、後年の南 (1974, 1993) などで提示される、南不二男の従属句の研究（通称「南モデル」）の原型がすでに示されている点である。現在、南モデルの文献としては南 (1974) が挙げられることが多く、時に南 (1964) が挙げられることもあるが、その原型は (1b) にまで遡ることができるわけである。表 2 に、その記述の一部を示す。なお、R は (1b) で規定された「連用語」、J は「状況語」、T は「陳述的成分」、K は「従属句」を指す。

表 2: 「南モデル」の原型（『話しことばの文型 (2)』p.93、一部改変）

句		R		J													T			K							
句内部の形式		テ	ナガラ	連用形	ズ・ズニ	テ	バ	タラ	ナラ	テモ	ト1	ト2	ナガラ	ノデ	ノニ	ガ	ケレドモ	ト2	ガ	カラ	ケレドモ	シ	連用形	テ	ノデ
句の述語の形	デス	×	×	×	×	○	×	○	×	×	×	○	×	○	○	○	○	○	○	○	○	○	×	○	○
	マス	×	×	×	▽	○	△	○	△	○	×	○	×	○	○	○	○	○	○	○	○	○	×	○	○
	ナイ	×	×	×	×	×	○	○	○	○	×	○	○	○	○	○	○	○	○	○	○	○	×	×	○
	タ	×	×	×	×	×	×	×	○	×	×	×	×	○	○	○	○	×	○	○	○	○	×	×	○
	ダロウ	×	×	×	×	×	×	×	×	○	×	×	△	△	△	△	×	○	○	○	○	×	×	△
	ウ・ヨウ	×	×	×	×	×	×	×	×	○	×	×	×	×	△	△	×	○	○	○	○	×	×	×
	マイ	×	×	×	×	×	×	×	×	○	×	×	×	×	△	△	×	○	○	○	○	×	×	×
述語以外の部分	キット	×	×	×	×	×	×	○	×	×	×	×	×	×	×	×	×	○	○	○	○	△	△	×
	ケッシテ	×	×	×	○	×	○	○	○	×	×	×	▽	○	○	×	×	×	○	○	○	○	×	×	○
	セメテ	×	×	×	×	○	×	○	×	×	×	×	×	○	○	×	×	×	○	○	○	○	×	○	○
	タブン	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	○	○	○	○	×	×	×
	サゾ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	○	○	○	○	×	×	×
	マサカ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	○	○	○	○	×	×	×
	～ガ	×	×	○	○	○	○	○	○	○	○	○	○	○	○	×	×	×	○	○	○	○	○	○	○
	～ハ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	○	○	○	○	○	○	○

△は、ありうると思われるけれども、やや疑問のあるものを示す。
▽は、ないと思われるけれども、ばあいによってはあるかもしれないもの。

後の南 (1974, 1993) の表と比べるといささかの違いがあるものの、従属句内部における各種要素の包含可能性という基本的な考え方自体は、すでにこの時点で出来上がっていたと見てよい。重要な点は、後年、日本語文法の根幹をなす文構造モデルとして大いに利用される研究が、話し言葉コーパスの観察の中から生まれていたという事実である。膨大な量の発話データの文型を整理するという作業を通じて、言語学的に一般化できる文の階層構造を描き出すモデルが作り出されたという点において、「南モデル」は、コーパス言語学の手法が極めてうまく機能した例として評価することができる。

[1] これらの報告書は、国立国語研究所のウェブサイト上で PDF ファイルとして公開されている。

その後の国立国語研究所では、語彙調査が継続的に実施され、その調査対象としていくつかの日本語コーパスが作成された。国立国語研究所 (1962、1963、1964)『雑誌九十種の用語用字 I〜III』は、さまざまなジャンルに分類される雑誌 90 種から、統計的に厳密な手続きで約 44 万語のサンプルを採取し、そこに現れる文字や用語を集計した研究である。さらに、収集された用例カードを再利用して、宮島 (1972)『動詞の意味・用法の記述的研究』、西尾 (1972)『形容詞の意味・用法の記述的研究』といった精緻な記述文法書が執筆された。コーパスに基づく記述文法書（レファレンスグラマー）は、英語では Quirk et al. (1972) *A Grammar of Contemporary English* や、その改訂版である Quirk et al. (1985) *A Comprehensive Grammar of the English Language* がつとに有名であるが、日本語でも早い時期からコーパスに基づく記述文法書が編纂されていたことになる。

1966 年には、国立国語研究所にコンピュータ（中型事務用計算用機 HITAC3010）が導入され、新聞の語彙調査における集計処理に用いられた。その成果は国立国語研究所 (1970)『電子計算機による新聞の語彙調査』などで見ることができる。これらは、コンピュータを利用して日本語コーパスが処理された、最も初期の研究成果として位置づけられる。

ここまで見たように、1950 年代から 1970 年代にかけて、国立国語研究所において日本語コーパスの組織的な開発と利用が積極的に進められていた。特に 1960 年代前半まではすべてのデータが用例カードの上に記録されていたため、現代の視点から見れば極めて使いにくい「電子化されていないコーパス」の状態であったが、それらを人手で集計し、極めて質の高い日本語研究が実践されていたという事実は、記憶に値すると言えるだろう。実際、当時の報告書を読んでみると、現代のコーパス日本語学で突き当たるさまざまな問題（例えば、単語の認定、文/発話の認定、係り受け構造の捉え方など）の多くが、すでに詳細に検討されていたことが分かる。

ただし、当時公開されたのは最終的な調査結果としての語彙頻度表や研究報告書のみであり、調査の過程で収集された日本語コーパス（用例カード）を一般に公開して研究者間で共有するという機運は、残念ながら生まれなかった[2]。また、1970 年代に入って以降は国立国語研究所の語彙調査も徐々に下火になり、日本語コーパスの開発にはブレーキがかかることになる。

2.2 1990 年代の日本語コーパスと複文構文の研究

時代が下り、1980 年代の後半から 1990 年代に入ると、徐々にパーソナル・コンピュータが普及し始め、個人でコンピュータが使える状況が生じてきた。同時に、電子テキストを収録したフロッピーディスクや CD-ROM などが電子出版物として（あるいは出版物の付録として）出版されるようになった。また、電子ブック版の新聞記事データベース、そして当時パソコン通信によって広がったインターネット上の電子テキストなど、さまざまな種類の電子テキストが流通し始めるようになった。そこで、電子テキストを自分のパソコンに取り込み、言語分析を試みる研究者が現れるようになった。

後藤 (1993) は、パソコン通信で「朝日新聞記事データベース」にアクセスして得たデータから「神話」という語を抽出し、比喩的な用法を分類したものである。近藤 (1993)、荻野・塩田 (1994)、田野村 (1994) などは、CD-ROM として発売された『朝日新聞 — 天声人語・社説 1985–1989』『CD-HIASK 朝日新聞記事データベース』『CD-毎日新聞』などに収録された新聞記事テキストデータを検索して得られた大量の用例をもとに、副助詞、語順、類義表現の分析を行なったものである。これらの萌芽的な研究によって、現代におけるコーパス日本語学の先鞭が付けられたと言ってよい。

また、出版された CD-ROM から電子テキストを抽出するプログラムを作成する研究者が現れた。上述の朝日新聞・毎日新聞の CD-ROM や、新潮文庫の小説など 100 作品を収録した『CD-ROM 版新潮文庫の 100 冊』（1995 年）から抽出した電子テキストを、各研究者や大学院生がコピーして使うというケースが広がった（朝日新聞・毎日新聞については、新聞社との協議の上、利用者には誓約書を提出させる措置が取られた）。これによって、電子テキストが徐々に研究者の間に普及し始めた。

[2]後年、国立国語研究所 (1987) によって調査の原資料である用例カードが公開されたものの、媒体がマイクロフィッシュであったため、必ずしも使い勝手のよいものではなかった。さらに 1997 年には、国立国語研究所 (1997) によって『現代雑誌九十種の用語用字』の調査結果が電子化された。ここから得られた統計情報は、宮島 (1997) にまとめられている。

この時期におけるコンピュータと電子テキストの普及は、コーパス日本語学の展開に新たな可能性をもたらしたと言える。上述の近藤 (1993)、荻野・塩田 (1994)、田野村 (1994) らによる研究は、ある言語表現が実際に出現する際の文法的・文体的な偏りについて、大量の電子テキストを用いて実証的に明らかにしたものであった。母語話者の内省では予想できない言語使用の実態を、個人の研究者が手元のコンピュータで分析できるようになったことの意義は大きい。しかしながら、当時は「研究者がたまたま入手できるようになった電子テキストを研究に役立てているに過ぎない (田野村, 2000)」状態であり、そこで得られた分析結果が現代日本語の全般に通用する結果なのか、あるいは特定の分野（例えば新聞）に限定された結果なのかを、明確に位置づけることができないという問題があった。さらに、当時の日本語研究者には形態素解析されたデータが出回ることがなかったため、文字列検索でしか用例を検索できない状態にあった。

次に、この時期における複文構文の研究に目を向けてみたい。1980 年代の後半以降、記述的な日本語文法研究の中では、命題（格、ヴォイス、テンス・アスペクト）からモダリティへと研究の関心が移行し、さらに複文構文が研究対象として盛んに取り上げられるようになった。この時期、

(2) a. 益岡 (1993)『日本語の条件表現』くろしお出版
b. 田窪 (1994)『日本語の名詞修飾表現』くろしお出版
c. 仁田 (1995a, 1995b)『複文の研究（上）（下）』くろしお出版

などの論文集や益岡 (1997) などの研究書が相次いで刊行されたことは、その隆盛ぶりを示している。

さて、当時の記述的な文法研究の中では、新聞や書籍などに現れた実例を研究者自身が目で拾い、論文に引用することが多く行なわれてきた（「目視スカウト方式」と呼んでおく）。上記 4 冊の論文集に収録された 37 本の論文のうち、目視スカウト方式によって実例を引用している論文は 18 本あった。一方、電子テキストから検索した実例を引用している論文は 1 本のみであった。電子テキストの利用が徐々に始まっていたとは言え、その普及には多少の時間がかかっていたことがうかがえる。

1990 年代の後半に入ると、インターネット上に「青空文庫」が開設され、著作権の切れた比較的古い時代の文学作品のテキストファイルが入手できるようになった。また、市販の研究書にコーパスが同梱されて出版されるケースも見られるようになった (現代日本語研究会, 1997)。さらに、日本語教育の現場から、教材開発用に複数の種類の書き言葉を集めた「Castle/J」や、日本語学習者 90 人分の OPI を文字化した学習者コーパス「KY コーパス (カッケンブッシュ, 1999)」などが公開された[3]。

2.3 2000 年代の日本語コーパスと複文構文の研究

2000 年代に入ると、多様な日本語コーパスの整備が進んできた状況を背景に、言語学系・日本語学系の学会の年次大会や論文誌で、日本語コーパスを使った研究が多く見られるようになってきた。また、学会誌や商業誌で日本語コーパスに関する特集号が企画されるケースが相次いだ。

(3) a. 『日本語学』20 巻 13 号「特集 コンピュータによる日本語研究の新展開」2001 年
b. 『日本語学』22 巻 4 月臨時増刊号「コーパス言語学」2003 年
c. 『日本語教育』130 号 【特集】コーパスと日本語教育 —現状と課題— 2006 年
d. 『日本語科学』22 号「特集 コーパス日本語学の射程」2007 年
e. 『日本語学』27 巻 2 号「特集 WWW を対象にした日本語研究」2008 年
f. 『国文学 解釈と鑑賞』74 巻 1 号「特集 日本語研究とコーパス」2009 年
g. 『人工知能学会誌』24 巻 5 号「特集 日本語コーパス」2009 年
h. 『言語研究』138 号「特集 コーパスを活用した言語研究」2010 年

[3]学習者コーパス構築の先駆的な試みとしては、寺村 (1990)『外国人学習者の日本語誤用例集』がある。

以下では、2001 年に設立された「日本語文法学会」の学会誌『日本語文法』に収録された論文における日本語コーパスの使用状況について、調査した結果を報告する。調査対象は、『日本語文法』の第 1 巻 1 号（2001 年 9 月）から第 12 巻 2 号（2012 年 9 月）までの 23 冊である。ここから特集論文、研究論文、研究ノートを抜き出したところ、215 本の論文が得られた。このうち、既存の日本語コーパスを利用（ウェブ検索を含む、古典資料・方言資料の利用は除く）した論文は 63 本、全体の 29.3%であった。約 1/3 の論文が、現代日本語のコーパスを利用していることになる。ここで使われている日本語コーパスの種類を集計したところ、表 3 のような結果が得られた。なお、1 本の論文で複数の日本語コーパスを使っている場合がある。

表 3: 『日本語文法』で使われている日本語コーパス

18	『新潮文庫の 100 冊』	3	青空文庫	1	ひつじ書房「女性のことば」
15	毎日新聞 CD-ROM	2	KY コーパス	1	学会抄録コーパス
12	朝日新聞 CD-ROM	2	科研音声コーパス	1	学習者データ
11	ウェブ検索（Google 等）	2	ひつじ書房「男性のことば」	1	国会会議録検索システム
8	BCCWJ モニター公開	2	日経新聞 CD-ROM	1	上村コーパス
5	Castle/J	1	『CD-ROM 大正の文豪』	1	中日対訳コーパス
5	CSJ	1	『CD-ROM 明治の文豪』	1	電子ブック

表 3 を見ると、『新潮文庫の 100 冊』のような市販の CD-ROM、そして新聞記事を収録した CD-ROM（朝日新聞のオンライン記事 DB「聞蔵」なども含む）の利用例が多いことが分かる。BCCWJ や CSJ、KY コーパスなど、言語研究用に特化して構築されたコーパスの利用があまり進んでいないが、特に BCCWJ は本公開が 2011 年であったことが影響しているだろう。今後の利用例の増加を期待したい。

また、目視スカウト方式によって実例を引用している論文は 57 本、全体の 26.5% であった。日本語コーパスを利用した論文は 63 本、29.3%であるから、日本語コーパスの利用が以前より普及してきていると見てよい。なお、日本語コーパスを利用している 63 本の論文のうち、定量的な分析を行なっている論文は 10 本に過ぎなかった。日本語コーパスを定量的な分析に用いるよりは、適切な例文を探すための情報源として使っている傾向が強いことがうかがえる。

さらに、全 215 本の論文のうち複文構文を扱った論文は 38 本、全体の 17.7%であった。このうち 4 本は、古典資料に基づく研究である。残る 34 本のうち、日本語コーパスを利用した論文は 17 本であった。現代日本語の複文構文を扱った論文の 50% が、日本語コーパスを利用していることになる。

例えば、ナロック (2006) は、各種電子ブックや「Castle/J」などから 10 数万の接続助詞の用例を採集し、従属節内部に現れる各モダリティ形式の分布を詳細に分析した研究である。また清水 (2009) では、毎日新聞 9 年分の電子テキストから「V1 テハ V2」という構文を抽出し、周辺に共起しやすい語や、テハ構文の形式と用法の関連について分析を行なっている。

さて、2000 年代以降の大きな研究動向として、国立国語研究所の主導により、話し言葉・書き言葉の大規模コーパスが開発・公開されたことが挙げられる。中でも『日本語話し言葉コーパス（CSJ）』および『現代日本語書き言葉均衡コーパス（BCCWJ）』の開発と公開は、膨大な量の言語データと質の高いアノテーション情報を備えた大規模コーパスが一般に出回ったという点で、特筆に価する。以下では、これら 2 つのコーパスの概要と、それらを用いた複文研究の可能性について述べる。

『日本語話し言葉コーパス（CSJ: Corpus of Spontaneous Japanese）』

CSJ は、国立国語研究所、情報通信研究機構、東京工業大学により共同開発された自発音声コーパスである。1999 年から 5 年間の構築期間を経て、2004 年に 18 枚組の DVD で一般公開された。CSJ に収録されているのは、発話状況の異なる 2 種類の独話（学会講演・模擬講演）を中心とする、約 661 時間、752 万語分の自発音声である。音声データに対して、発話内容を忠実に書き起こした転記テキスト、形態論情報、節単位情報、分節音情報、韻律情報、係り受け構造情報、談話構造情報、要約・重要文情報、印象評定データ、話者情報など、豊富な研究用情報が提供されている。

さて、CSJ にアノテーションされている「節単位情報 (丸山・高梨・内元, 2006)」の中では、発話中に現れる従属節の終端境界に「節境界ラベル」が付与されている。例を以下に示す。

> 今日お話しさせていただく内容なんですけれども **/並列節ケレドモ/** (F えーっとー) (F ま) 特に珍しいことではないと **<引用節>** 思うのですが **/並列節ガ/** (F あのー) 自分は (F あの) 体は (F えー) 元から強い方ではなかったのですが **/並列節ガ/** (F あの) いわゆる病気らしい病気ということはしたことがなくて **/テ節/** て (F その) (F え) (D い) 一か月程ずっと寝たきりと言うか **<文末候補>** 家におりまして **/テ節/** (F あの) 病気をしておりましたもので **<並列節デ>** 自分にとっては **<テハ節>** 生まれて **<テ節>** 初めてのことだったので **<理由節ノデ>** (F えー) そのことについて **<テ節>** お話しさせていただきたいと **<引用節>** 思います **[文末]**

太字で示されている部分が節境界ラベルである。特に自発音声においては、上記のように連綿と長く続く発話スタイルがよく観察される。このような発話では、ある種の節境界（およびそのイントネーション）が談話構造の切れ目や話題の展開に関与していると考えられる。そこで、発話データに付与された節境界ラベルの連鎖パタンを検討することによって、多重的な節連鎖の構造と、発話の大局的な構造の分析が可能になるだろう。これは、複文研究の応用領域の一つと言える。

『現代日本語書き言葉均衡コーパス（BCCWJ: Balanced Corpus of Contemporary Written Japanese）』

BCCWJ は、国立国語研究所を中心に開発された、現代日本語の書き言葉の大規模コーパスである。2006 年度から 5 年間の開発期間を経て、2011 年に一般公開された。1 億語超の書き言葉を収録したこのコーパスは、日本語では初となる均衡コーパスであり、綿密な調査に基づいた設計が施され、母集団に対する統計的な代表性を有するサンプリングが実施されている点に最大の特徴がある。

BCCWJ に収録されているテキストの種類（メディア）と語数（短単位）の一覧を、表 4 に挙げる。

表 4: BCCWJ に収録されたテキストと語数

メディア	サンプル数	語数	メディア	サンプル数	語数
出版書籍	10,117	28,552,283	ベストセラー	1,390	3,742,261
雑誌	1,996	4,444,492	Yahoo!知恵袋	91,445	10,256,877
新聞	1,473	1,370,233	Yahoo!ブログ	52,680	10,194,143
図書館書籍	10,551	30,377,866	韻文	252	225,273
白書	1,500	4,882,812	法律	346	1,079,146
教科書	412	928,448	国会会議録	159	5,102,469
広報紙	354	3,755,161	合計	172,675	104,911,464

BCCWJ は、ウェブ上の検索サイト「少納言」および「中納言」から検索できるようになっている。誰でも利用できる少納言は文字列検索に限定されるの対して、登録制の中納言では形態素解析されたデータを検索することが可能である。また、データ全体を収録した DVD が、有償で公開されている。

さて、2006 年から開始された BCCWJ の構築プロジェクト[4]では、コーパス構築の進捗報告や研究発表ための公開ワークショップが毎年度末に開催されていた。そこでの発表件数をカウントしてみると、図 1 のようになる。5 年間の構築期間が進むにつれて、BCCWJ を利用したコーパス日本語学の実践例が増加していった様子を見て取ることができる。

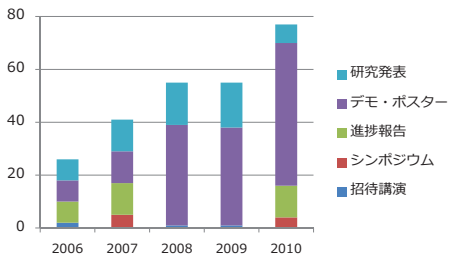

図 1: 公開 WS での発表件数の推移

これら一連のワークショップの中で、複文構文に関する研究は必ずしも多くなかったが、その中で茂木 (2008) は「ないために」という従属節の形式を取り上げ、「目的」「理由」という 2 つの用法と出現数、およびタメニ節全体に占める割合について、表 5 のようにまとめている（p.125）。

[4]文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（2006～2010 年度、研究代表者 前川喜久雄）。

表 5: タメニ節に占める「ないために」の割合

	ないために	ために	割合
目的	258	8,917	2.89%
理由	154	1,354	11.37%

茂木はここから、「ないために」の目的用法はまれ、という従来の記述の妥当性を確認した上で、それでもなお、BCCWJ に現れた実例には典型的な例から非典型的（周辺的）な例への連続性が認められることを指摘し、記述対象の設定の難しさを論じている。

また小西 (2009) は、原因・理由を表す接続助詞「から」と「ので」について、各ジャンルにおける文末と従属節における丁寧体の現れ方、およびその多寡について、表 6 のようにまとめている。

表 6: 「から」「ので」と文末に現れる丁寧体/普通体のパタン（上位 3 位）

文学	普+から+普。	(4,224)	>	普+から。	(2,950)	>	普+ので+普。	(1,489)
文学以外	普+から+普。	(6,904)	>	普+ので+普。	(5,261)	>	普+ので+丁。	(2,648)
新聞	普+から+普。	(98)	>	普+ので+普。	(69)	>	普+から。	(44)
雑誌	普+から+普。	(965)	>	普+ので+普。	(750)	>	普+から。	(374)
Y!知恵袋	普+ので+丁。	(8,201)	>	普+から+丁。	(2,381)	>	丁+ので+丁。	(2,124)
白書	普+ので+普。	(325)	>	普+から+普。	(126)	>	普+から。	(78)

※ pp.136-137 より一部抜粋、改変。「普」は普通体、「丁」は丁寧体を表す。

小西はここから、原因・理由表現として真っ先に提示される文型である「丁+から+丁。」の出現が少ないこと（例えば文学では 264 例しかない）などを指摘し、日本語教育において文型を提示する際、学習者の目的に応じた選択をすることが必要であると論じている。

さて、CSJ・BCCWJ が一般に公開されたことの第一の意義は、現代日本語の大規模コーパスが研究者間で共有される状態を生み出したことにあると言ってよい。例えば従来の書き言葉研究では、新聞記事や比較的古い時代の文学作品など、偏ったテキストしか入手できなかった。これに対して、さまざまな使用域（レジスター）を持つ現代の書き言葉が 1 億語超という規模で公開されたことにより、これまでは考えられなかった日本語研究用のインフラストラクチャが整備されたと言える。例えば、上述の小西 (2009) のような研究は、従来、国立国語研究所のように組織的な研究体制を組み、多くの人手と時間をかけないと行ない得なかったものである。BCCWJ の公開により、このような研究が個人のコンピュータ内でさほど時間をかけずに検索できるようになっている。

さらに、同じデータが共有されたことにより、ある研究者の分析結果を別の研究者が追試することが可能になった。すなわち、厳密な意味での実証的・客観的な日本語研究が可能な状況が生じたことになる。CSJ・BCCWJ が公開された現在、コーパス日本語学は新たな段階に入ったと言えるだろう (丸山・田野村, 2007)。

3　今後のコーパス日本語学と複文構文の研究

本発表では、日本語コーパスの開発史を紹介しながら、その中で実践されてきたコーパス日本語学の研究、特に複文構文の研究の一端について見てきた。紙幅の都合上、取り上げることができた研究事例はごく少数のものに限られたが、特に BCCWJ が一般公開された現在、日本語コーパスを用いた複文構文の研究は、今後ますます増えていくものと予想される。

そこで今後必要となるのは、よりよい分析手法の開発であると考えられる。現時点では、接続形式の種類と出現数など、複文構文の形態的な側面に着目した定量的な研究は可能であるが、接続形式と文末形式の呼応関係のように、統語的な側面に着目した研究は、大規模には実施しにくい状況にある。これは、コーパス本体に対して統語解析を実施し、係り受け関係をアノテーションすることによって、部分的に解決されるだろう（BCCWJ に対する統語解析は、現在作業が進行中である）。よりよいアノテーション情報の整備と、それを用いた研究手法の開発、そして複文構文の定量的な記述を推し進めることが、今後の課題であると考えられる。

参考文献

現代日本語研究会（編） (1997). 『女性のことば（職場編）』. ひつじ書房.

後藤斉 (1993). 「『神話』の比喩的用法について—コーパス言語学からのアプローチ—」. 『東北大学言語学論集』, **2**, 1–16.

カッケンブッシュ・寛子 (1999). 『第 2 言語としての日本語習得に関する総合研究』. 平成 8 年度～平成 10 年度科学研究費補助金研究成果報告書.

国立国語研究所 (1951). 『現代語の助詞・助動詞 —用法と実例—』. 国立国語研究所報告 3. 国立国語研究所.

国立国語研究所 (1960、1963). 『話しことばの文型 (1)～(2)』. 秀英出版.

国立国語研究所 (1962、1963、1964). 『現代雑誌九十種の用語用字 第 1 分冊～第 3 分冊』. 秀英出版.

国立国語研究所 (1970). 『電子計算機による新聞の語彙調査 I』. 国立国語研究所報告 37. 秀英出版.

国立国語研究所 (1987). 『現代雑誌九十種の用語用字 五十音順語彙表・採集カード』. 国立国語研究所言語処理データ集 3.

国立国語研究所 (1997). 『現代雑誌九十種の用語用字 全語彙・表記【FD 版】』. 国立国語研究所言語処理データ集 7. 三省堂.

近藤泰弘 (1993). 「文法研究における大量言語データ —副助詞研究を例にして—」. 『武蔵野文学』, **40**, 12–16.

小西円 (2009). 「「から」「ので」の形態的特徴と使用ジャンル—日本語教育における類義表現の扱いを考える—」. 『特定領域「日本語コーパス」平成 21 年度公開ワークショップサテライトセッション 予稿集』, pp. 131–138. 特定領域研究「日本語コーパス」総括班.

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』, 国立国語研究所報告書 124, pp. 255–322. 国立国語研究所.

丸山岳彦・田野村忠温 (2007). 「コーパス日本語学の射程」. 『日本語科学』, **22**, 5–12.

益岡隆志（編） (1993). 『日本語の条件表現』. くろしお出版.

益岡隆志 (1997). 『複文』. くろしお出版.

南不二男 (1964). 「複文」. 『口語文法の問題点』, 講座現代語 6. 明治書院.

南不二男 (1974). 『現代日本語の構造』. 大修館書店.

南不二男 (1993). 『現代日本文法の輪郭』. 大修館書店.

三尾砂 (1942). 『話言葉の文法 言葉遣篇』. 復刊 くろしお出版 1995 年.

宮島達夫 (1972). 『動詞の意味・用法の記述的研究』. 国立国語研究所報告 43. 秀英出版.

宮島達夫 (1997). 「雑誌九十種表記表の統計」. 『日本語科学』, **1**.

茂木俊伸 (2008). 「目的を表す「ないために」の実態」. 『特定領域「日本語コーパス」平成 20 年度公開ワークショップサテライトセッション 予稿集』, pp. 119–126. 特定領域研究「日本語コーパス」総括班.

永野賢 (1952). 「「から」と「ので」とはどう違うか」. 『国語と国文学』, **29** (2), 30–41.

ナロック・ハイコ (2006). 「従属節におけるモダリティ形式の使用」. 『日本語文法』, **6** (1), 21–37.

西尾寅弥 (1972). 『形容詞の意味・用法の記述的研究』. 国立国語研究所報告 44. 秀英出版.

仁田義雄（編） (1995a). 『複文の研究（上）』. くろしお出版.

仁田義雄（編） (1995b). 『複文の研究（下）』. くろしお出版.

荻野綱男・塩田雄大 (1994). 「朝日新聞データベースを利用した言語研究」. 『日本語学』, **13** (5), 28–39.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1972). *A Grammar of Contemporary English*. Longman.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

清水由貴子 (2009). 「反復の意味を表す「V1 テハ V2」文の分析—形式的側面を中心に—」. 『日本語文法』, **9** (1), 54–70.

田窪行則（編） (1994). 『日本語の名詞修飾表現』. くろしお出版.

田野村忠温 (1994). 「丁寧体の述語否定形の選択に関する計量的調査—「～ません」と「ないです」—」. 『大阪外国語大学論集』, **11**, 51–66.

田野村忠温 (2000). 「用例に基づく日本語研究 —コーパス言語学—」. 『日本語学』, **19** (5), 192–201.

寺村秀夫 (1990). 『外国人学習者の日本語誤用例集』. 1985～1989 年度科学研究費補助金特別推進研究「日本語の普遍性と個別性に関する理論的及び実証的研究」分担研究「外国人学習者の日本語誤用例の収集・整理と分析」資料.