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」公開シンポジウム
（国立国語研究所、2012年12月15日）

日本語コーパスと複文の研究
――BCCWJの特性と利用の方法――

田野村 忠温
（大阪大学）

概要

コーパスは強力な言語研究の資料・手法であるが、あらゆる種類の研究にコーパスを容易に生かせるわけではない。コーパスの苦手とする重要な言語研究領域の１つが構文論である。これは、コーパスは少なくとも現状では基本的に言語表現に関する線状的なデータに過ぎず、表現の構造に関わる情報の利用が限定的であることによる。

昨年（2011年）、国立国語研究所の開発による「現代日本語書き言葉均衡コーパス(BCCWJ)」が完成し公開された。綿密な設計に基づくBCCWJは今後の現代日本語研究において標準的なコーパスとして広範に用いられるはずのものであるが、従来日本語のコーパスとして広く使われてきた文学作品や新聞記事などの電子テキストと比べると、その利用者に対する要求のレベルがすこぶる高い。

ここでは、BCCWJの特性の分析結果の一端を紹介したうえで、BCCWJから複文の研究のための用例を収集する方法の基礎的知識を解説する。

1　BCCWJの資料的特性［拙論(2012a, 2012b, 2013?)］

1.1　従来のコーパスとの大きな違い

・複雑な内部構成を有する。

・従来日本語研究に広く使われてこなかった種類の言語データを含む。

1.2　各種日本語コーパスの規模

・BCCWJは１億語のコーパスと説明されるが、それはどのくらいの規模だろうか？

表1　各種の日本語コーパスの規模の比較

日本語データの種類	字数（データ量）	単行本換算量
小説単行本	20万字	1
CD-ROM版 新潮文庫の100冊	2,000万字（40MB）	100
新聞記事１年分	6,000万字（120MB）	300
BCCWJ	2億字（400MB）	1,000
Yahoo!知恵袋ベータ版データ	16億字（3.2GB）	8,000
国会会議録のデータ	35億字（7GB）	18,000
拙作Webコーパス	750億字（150GB）	375,000

1.3　BCCWJのサブコーパスとそのサイズ

・BCCWJは３つのサブコーパスから構成され、それぞれがさらに下位区分される。

表2　BCCWJのサブコーパスとその下位区分

サブコーパス	サブコーパス・下位区分
出版	書籍(PB)、雑誌(PM)、新聞(PN)
図書館	書籍(LB)
特定目的	白書(OW)、教科書(OT)、広報紙(OP)、ベストセラー(OB)、Yahoo!知恵袋(OC)、Yahoo!ブログ(OB)、韻文(OV)、法律(OL)、国会会議録(OM)

・サブコーパス、固定長・可変長ごとに見たサンプルのサイズは次の通り。

表3　サブコーパスごと、固定長・可変長サンプルごとのサイズ

サブコーパス		サンプル数	固定長		可変長		固定長＋可変長	
			総字数	単行本	総字数	単行本	総字数 (比率%)	単行本
出版	書籍(PB)	10,117	11,802,588	59	50,148,895	251	52,766,227 (27.0)	264
	雑誌(PM)	1,996	2,310,715	12	8,369,425	42	8,835,299 (4.5)	44
	新聞(PN)	1,473	1,697,189	8	1,606,936	8	2,504,698 (1.3)	13
図書館	書籍(LB)	10,551	12,347,986	62	53,165,672	266	55,899,439 (28.6)	279
特定目的	白書(OW)	1,500	1,807,002	9	8,183,753	41	8,476,458 (4.3)	42
	教科書(OT)	412			*1,744,451*	9	*1,744,449* (0.9)	9
	広報紙(OP)	354			7,001,594	35	7,001,594 (3.6)	35
	ベストセラー(OB)	1,390			6,875,567	34	6,875,567 (3.5)	34
	Yahoo!知恵袋(OC)	91,445			19,300,579	97	*19,300,548* (9.9)	97
	Yahoo!ブログ(OY)	52,680			20,868,316	104	*20,868,243* (10.7)	104
	韻文(OV)	252			376,811	2	376,811 (0.2)	2
	法律(OL)	346			1,786,081	9	1,786,081 (0.9)	9
	国会会議録(OM)	159			8,887,321	44	8,887,321 (4.6)	44
全体		172,675	29,965,480	150	188,315,401	942	195,322,735 (100.0)	977

※BCCWJ-DVD版から拙作ソフトウェアbccwj2textによって抽出・復元したテキストに基づく。

1.4　サブコーパスの性質の差

1.4.1　述語の有標率

・簡易な分析によるサブコーパスごとの述語有標率：

表4　サブコーパスごとの述語有標率

サブコーパス		否定率	過去率	丁寧率	終助詞率
出版	書籍(PB)	3.88	30.81	26.16	3.54
	雑誌(PM)	2.89	27.86	34.57	8.83
	新聞(PN)	2.47	46.92	9.04	1.11
図書館	書籍(LB)	4.15	35.00	22.16	4.22
特定目的	白書(OW)	1.16	37.38	1.76	0.00
	教科書(OT)	2.22	34.65	17.12	3.24
	広報紙(OP)	1.19	34.80	72.40	1.23
	ベストセラー(OB)	3.86	36.65	22.11	4.43
	Yahoo!知恵袋(OC)	2.13	25.81	79.04	13.94
	Yahoo!ブログ(OY)	1.43	42.20	55.05	9.77
	韻文(OV)	4.10	31.74	9.56	1.71
	法律(OL)	11.41	0.00	0.00	0.00
	国会会議録(OM)	1.67	24.17	75.23	24.08

1.4.2　語彙の使用状況――書籍サブコーパスPB（出版）とLB（図書館）の比較

・頻度の和が5,000件以上である名詞の頻度の相関の一部：

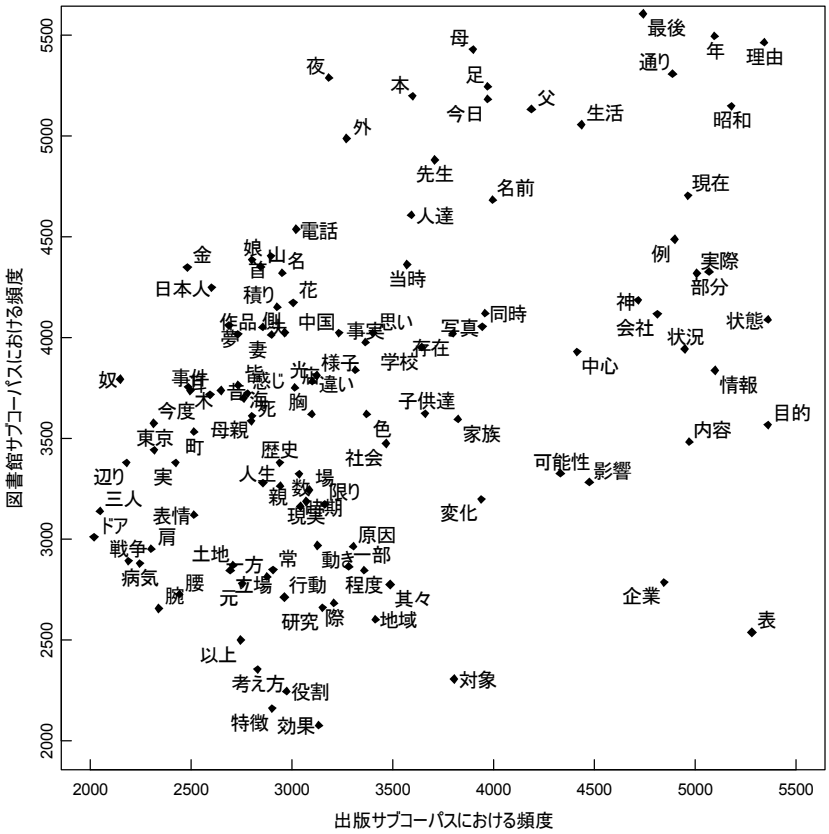


図1　PB・LBサブコーパスにおける名詞の出現頻度の相関

・出現頻度の和が1,000以上の名詞を偏りの順に並べたとき両端に位置する名詞：

表5　出版・図書館サブコーパスにおける名詞の出現の偏り

名詞	LB率	名詞	LB率		名詞	LB率	名詞	LB率
第一項	12.45	式	24.45		猿	63.94	雪	65.75
ファイル	19.58	数値	24.49		持ち主	64.21	台所	66.01
市町村	20.31	規定	24.84		中国人	64.22	女房	66.02
事項	20.66	金額	25.16		死体	64.28	自殺	66.06
図表	20.92	項目	25.36		ユダヤ人	64.34	紐	66.14
設定	21.82	株式	25.49	中略	ロンドン	64.72	連中	66.29
資産	21.88	要件	25.51		女達	64.76	小説	66.60
顧客	22.16	値	25.78		パパ	64.82	作者	68.69
適用	22.58	平成	25.80		辺	65.02	犯人	69.43
キー	22.73	取り組み	26.16		銃	65.20	御ばあちゃん	69.56
変更	23.05	上記	26.44		アパート	65.55	刑事	72.98
業務	23.38	当事者	27.37		赤ん坊	65.56	ソ連	73.53

※LB率＝LBにおける頻度÷PB・LBにおける頻度の和×100（調整）

・動詞、形容詞・形状詞についても類似の傾向が見られる。

1.4.3　データの重複

・Yahoo!ブログサブコーパスにはデータの重複（サンプル間、サンプル内）が多い。

・法律サブコーパスにはサンプル内のデータの重複が多い。

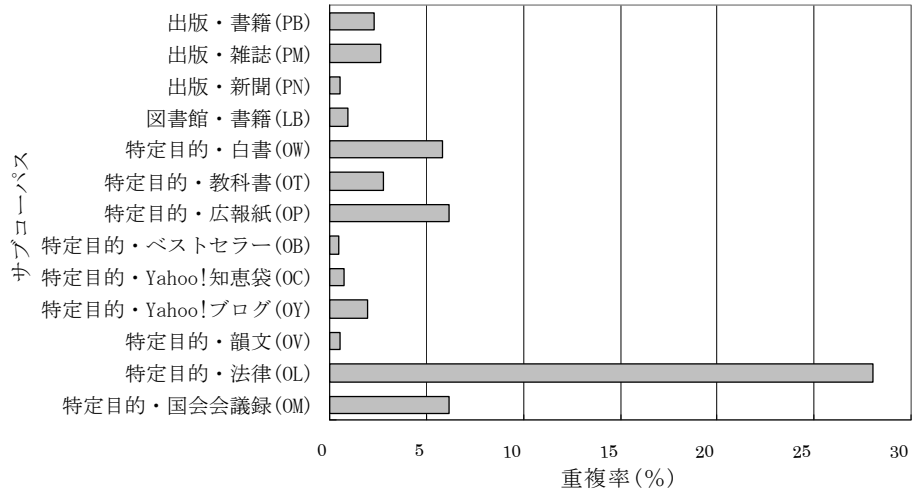


図2　各サブコーパスのサンプル内データ重複率

2　BCCWJの３つの利用形態

1) 「少納言」……単純な文字列検索のみ
2) 「中納言」……文法情報に基づく検索が可能
3) 「BCCWJ-DVD版」……BCCWJの全情報を利用可能（ソフトウェアを用意する必要あり）

3　中納言による用例収集

3.1　「短単位検索」「長単位検索」における３通りの検索方法

1) 「検索フォームで検索」

……検索条件をボタン、ドロップダウンメニュー、テキストボックスで指定して検索

※長所：　フォームに頼って手軽に検索条件を指定できる。
　短所：　条件指定の作業が煩雑。複雑な検索には向かない。

2) 「検索条件式で検索」

……検索条件を式で指定して検索

※長所：　一貫した方法で条件を指定できる。類似条件による検索を合理的に反復できる。
　短所：　式の書き方を覚え、文法を間違えずに指定する必要がある。

3) 「履歴で検索」

……過去の検索を再実行

※検索フォームで検索したあと、「履歴で検索」のタブをクリックすれば、その検索がどのような検索条件式で表されるかを見ることができる。
※表示された式を修正して「検索条件式で検索」にコピーして検索し直すこともできる。

図3　「検索フォームで検索」の様子

図4　「検索条件式で検索」の様子

3.2　検索条件式による検索の基礎

・検索条件式の一般形（多少簡略化）：

キーの条件 [AND 共起単位の条件 …]
[IN subcorpusName=サブコーパス名 [OR subcorpusName=サブコーパス名 …]]
[WITH OPTIONS …]
[OUTPUT INTO ファイル名]

※条件検索式については中納言オンラインマニュアルに一通りの説明がある。ここでの記述は実験に基づく推測・判断を含む。
※[　]内の指定は任意。全体を１行で書いてもよいし、適宜途中で改行してもよい。
※キーの条件は「キー: …」、共起単位の条件は「前方(or後方)共起: … ON (or WITHIN) 数 WORDS FROM キー」の形で指定する。
※「WITH OPTIONS…」は短単位検索・長単位検索の別、文脈中の区切り記号、前後文脈の語数、検索対象（固定長・可変長)を指定するものであるが、ここではそれらはフォーム上で指定することとし説明を省く。

・検索条件式の例：

1) 単純な条件

キー: 書字形出現形="熱い"　　→熱い
キー: 語彙素="熱い"　　→熱い、熱く、熱かっ、アツイ、あつク
キー: 発音形出現形="アツイ"　　→熱い、暑い、厚い、篤い、あつい、アツイ、アツい

2) ワイルドカードの使用――「_」（＝1字)、「%」（＝0字以上の文字連鎖)、「[…]」、「[^…]」

キー: 語彙素="日本_的"　　→日本人的、日本語的、日本学的
キー: 語彙素="日本%的"　　→加えて、日本的、日本語史的、日本社会的
キー: 語彙素="日本_%的"　　→以上から 日本的 を除いたもの
キー: 語彙素="_%日本語_%"　　→古代日本語文法、初級日本語講座、ＢＢＣ日本語放送
※オンラインマニュアルではLIKE演算子を使うとされているが、「=」でも問題なく機能する。
※検索語例は長単位検索の場合のもの。以後も同様。

3) キーに関わる複合的な条件

キー: (語彙素="大%" AND 品詞="動%")　　→大別する、大別さ、大爆笑し、大喜びする、おとなび
キー: (語彙素="大%" AND (品詞="動%" OR 品詞="形%"))　　→上記に加えて、大切、大好き、たいへん
※必ず「(A AND B)」「(A OR B)」のように括弧で囲んで指定する。

4) 共起単位に関わる条件

キー: 語彙素="大阪" AND 後方共起: 語彙素="から" ON 1 WORDS FROM キー　　→大阪から
キー: 語彙素="大阪" AND 後方共起: (語彙素="から" OR 語彙素="まで") ON 1 WORDS FROM キー
→大阪から、大阪まで

5) 検索対象とするサブコーパスの指定

キー: 語彙素="日本語" IN subcorpusName="PB"
キー: 語彙素="日本語" IN subcorpusName="PB" OR subcorpusName="LB"
※「subcorpusName="出版・書籍"」のようにサブコーパスの日本語名で指定することもできる。
※ここでは「(A OR B)」のように括弧で囲まなくても問題ない。

6) 複数の検索を行い、結果をダウンロード（出力ファイル名を指定）

キー: 語彙素="日本語" IN subcorpusName="PB" OUTPUT INTO PB.txt;
キー: 語彙素="日本語" IN subcorpusName="PM" OUTPUT INTO PM.txt;
キー: 語彙素="日本語" IN subcorpusName="PN" OUTPUT INTO PN.txt;
キー: 語彙素="日本語" IN subcorpusName="LB" OUTPUT INTO LB.txt

※各式をセミコロンで区切る。ファイル名に全角文字は使用不可。

※この４行から成る検索条件式を指定して［検索結果をダウンロード］のボタンを押すと保存ファイル名を尋ねて来るので、適当なファイル名（例えば、「日本語」）を指定する。これにより、PB.txt 以下4つのファイルを含む日本語.zipというファイルがディスク上に作られる。

3.3 複文に関わる用例の検索例

・「～う＋名詞」の用例をBCCWJ全体から得る。

キー: 活用形="意志推量形" AND 後方共起: 品詞="名詞%" ON 1 WORDS FROM キー
→社長ともあろう者、企業が得るであろう利潤、悪かろうはずがない

・「～う＋名詞」の用例をPB、PM、PN、LBの各サブコーパスから得る。

キー: 活用形="意志推量形" AND 後方共起: 品詞="名詞%" ON 1 WORDS FROM キー
IN subcorpusName="PB" OUTPUT INTO PB.txt;
キー: 活用形="意志推量形" AND 後方共起: 品詞="名詞%" ON 1 WORDS FROM キー
IN subcorpusName="PM" OUTPUT INTO PM.txt;
キー: 活用形="意志推量形" AND 後方共起: 品詞="名詞%" ON 1 WORDS FROM キー
IN subcorpusName="PN" OUTPUT INTO PN.txt;
キー: 活用形="意志推量形" AND 後方共起: 品詞="名詞%" ON 1 WORDS FROM キー
IN subcorpusName="LB" OUTPUT INTO LB.txt

・「動詞・助動詞＋の＋を＋動詞」の用例を13の各サブコーパスから得る。

キー: 品詞="動詞%" AND 前方共起: 語彙素="を" ON 1 WORDS FROM キー AND 前方共起: 語彙素="の" ON 2 WORDS FROM キー AND 前方共起: 品詞="%動詞%" ON 3 WORDS FROM キー
IN subcorpusName="PB" OUTPUT INTO PB.txt;
キー: 品詞="動詞%" AND 前方共起: 語彙素="を" ON 1 WORDS FROM キー AND 前方共起: 語彙素="の" ON 2 WORDS FROM キー AND 前方共起: 品詞="%動詞%" ON 3 WORDS FROM キー
IN subcorpusName="PM" OUTPUT INTO PM.txt;

（中略）

キー: 品詞="動詞%" AND 前方共起: 語彙素="を" ON 1 WORDS FROM キー AND 前方共起: 語彙素="の" ON 2 WORDS FROM キー AND 前方共起: 品詞="%動詞%" ON 3 WORDS FROM キー
IN subcorpusName="OM" OUTPUT INTO OM.txt;

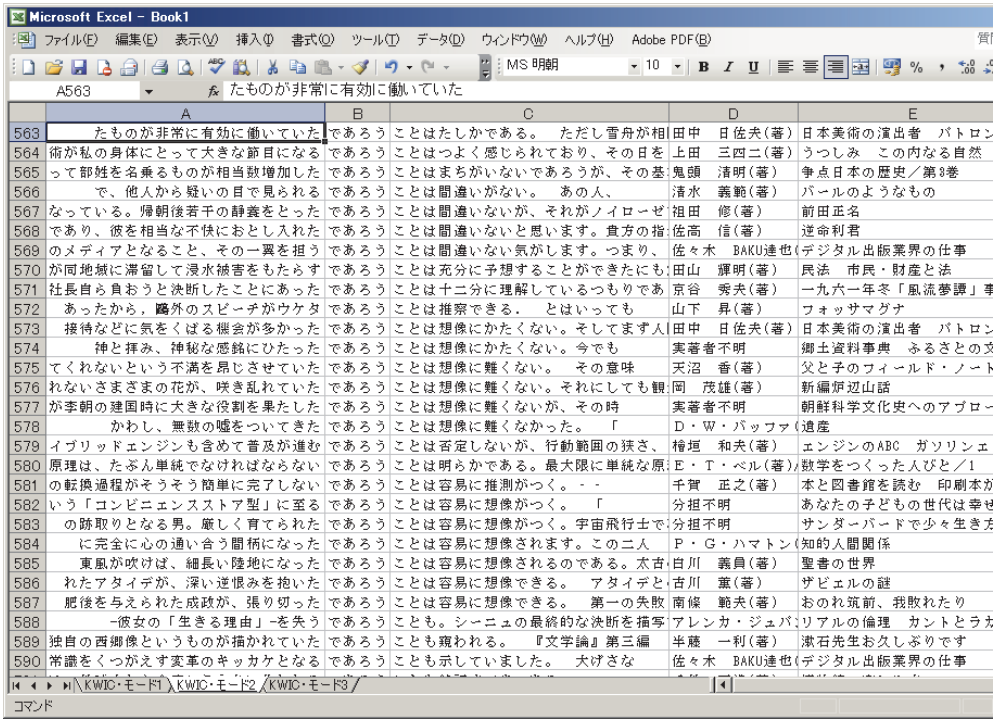
・実際の検索上の課題・問題点：

検索してみてその結果に問題があれば、検索条件を適宜調整する。思い通りの結果を得ることはむずかしい場合もある。

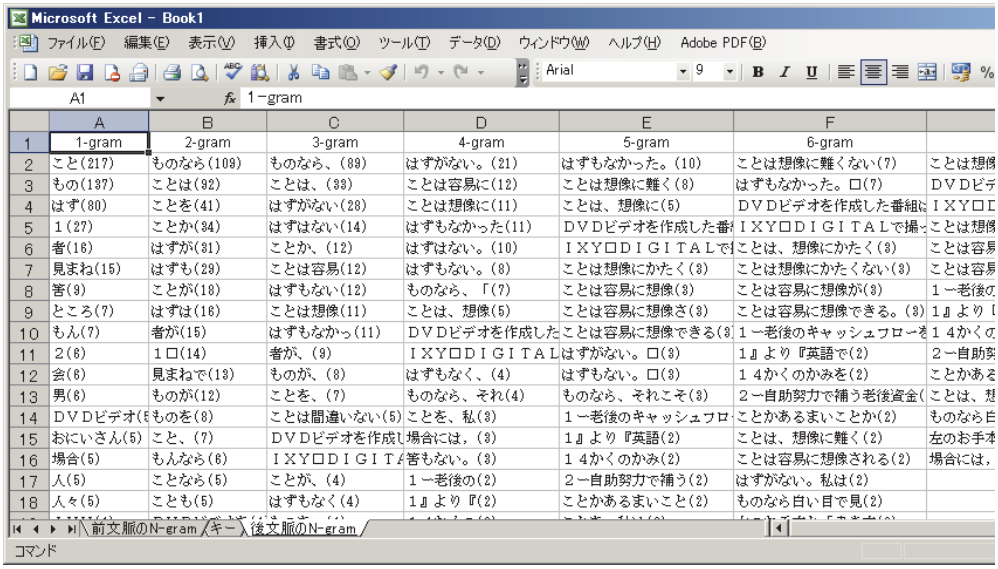
検索の条件によっては反応に長い時間がかかる。ときにはいつまで待っても中納言が反応しないこともある。

4 中納言による検索結果の整理・分析

4.1　拙作ソフトウェアsortKWICによる並べ替え



4.2　拙作ソフトウェアBNAnalyzerによる高頻度の共起語連鎖の分析



文献

国立国語研究所コーパス開発センター(2011)『「現代日本語書き言葉均衡コーパス」利用の手引』第1.0版（国立国語研究所コーパス開発センター）［BCCWJ-DVD版所収の電子文書］

田野村忠温(2012a)「BCCWJに含まれるウェブデータの特性について――データ重複の諸相とBCCWJ使用上の注意点――」『第２回コーパス日本語学ワークショップ予稿集』（国立国語研究所）

田野村忠温(2012b)「BCCWJに収められた新種の言語資料の特性について――データ重複の諸相とコーパス使用上の注意点――」『待兼山論叢』第46号文化動態論篇（大阪大学大学院文学研究科）［拙論(2012a)の増補版］

田野村忠温(2013?)「BCCWJの資料的特性」『講座日本語コーパス６　コーパスと日本語学』（朝倉書店）