第7回産業日本語研究会・シンポジウム
(20160229) 於東京・丸ビルホール

日本語の全体像を知るために
―国立国語研究所による言語資源整備―

前川 喜久雄
国立国語研究所 言語資源研究系・コーパス開発センター

国立国語研究所（東京都立川市）

国立国語研究所は1948年に設置された国立の研究機関。独立行政法人を経て、現在は大学共同利用機関法人人間文化研究機構が設置する大学共同利用機関の一つ。「国語及び国民の言語生活並びに外国人に対する日本語教育に関する科学的な調査研究並びにこれに基づく資料の作成及びその公表」を目的とする。

自己紹介

- 本来の専門は音声学・音声科学
- 90年代末から「言語資源」学に深入り
- 設計・開発・分析に携わってきた言語資源
  - 『日本語話し言葉コーパス』
  - 『現代日本語書き言葉均衡コーパス』
  - 『国語研日本語ウェブコーパス』

なぜ今日私がここにいるか？

- 単なる人選ミス
- 「人工知能」で日本語（特許文書）を扱う

⇒ 自然言語処理

⇒ 現在の自然言語処理は統計ベース

⇒ 言語資源

⇒ 前川

自然言語処理と言語資源の関係

当初の自然言語処理は「ことばとは、このようなものだ」という規則を書き連ねることで実現していた。しかし、ことばは極めて多様で、常に変化し、人や文脈によって解釈が異なりうる。そのすべてを規則として記すのは、まったく現実的ではなかった。

中谷秀洋「特集統計的自然言語処理—ことばを扱う機会」
IWANAMI DATA SCIENCE, Vol.2, p. 4, 2016.

言語内多様性

- 多様性の源
  - 歴史的多様性
    - 内発的変化
    - 外国語の影響
  - 地理的多様性
  - 創造的使用
  - 確率的変動(言語変異)
  - その他
    - 非母語話者
    - 機械翻訳

言語内多様性

- 多様性の源
  - 歴史的多様性
    - 内発的変化
    - 外国語の影響
  - 地理的多様性
  - 創造的使用
  - 確率的変動（言語変異）
  - その他
    - 非母語話者
    - 機械翻訳

言語内多様性を把握する手段

- 直観だけでは把握できない（例は後で）
- 客観的なデータが必要
- 対象を偏りなく代表するデータ
- できるだけ大量に
- 検索用の情報をつけて
- コンピュータで利用できる形式で
- 誰でも利用可能な形で

コーパス ～ 言語資源(language resources)の整備

コーパス開発センターHP

国語研による言語資源開発の経緯

- 1950年代から各種「語彙調査」を実施してきたがデータは公開しなかった
- 1990年代末にコーパス開発始動（KOTONOHA計画）
  - 『日本語話し言葉コーパス（CSJ）』（構築 1999~2003,公開2004）
  - 『太陽コーパス』（構築 1995~2005, 公開2005）
  - 『現代日本語書き言葉均衡コーパス（BCCWJ）』（構築 2006-2011, 公開2011）
  - 『日本語歴史コーパス（CHJ）』（構築2010～, 段階的に公開）
  - 『国語研日本語ウェブコーパス』（構築2011～2015, 公開2016予定）
  - 『多言語母語の日本語学習者横断コーパス（I-JAS）』（構築2012～, 部分試験公開2016）

詳しくは原稿記載のURLを参照してください

日本語コーパス整備の経緯Ⅰ:1990年代

それ以前

近　代

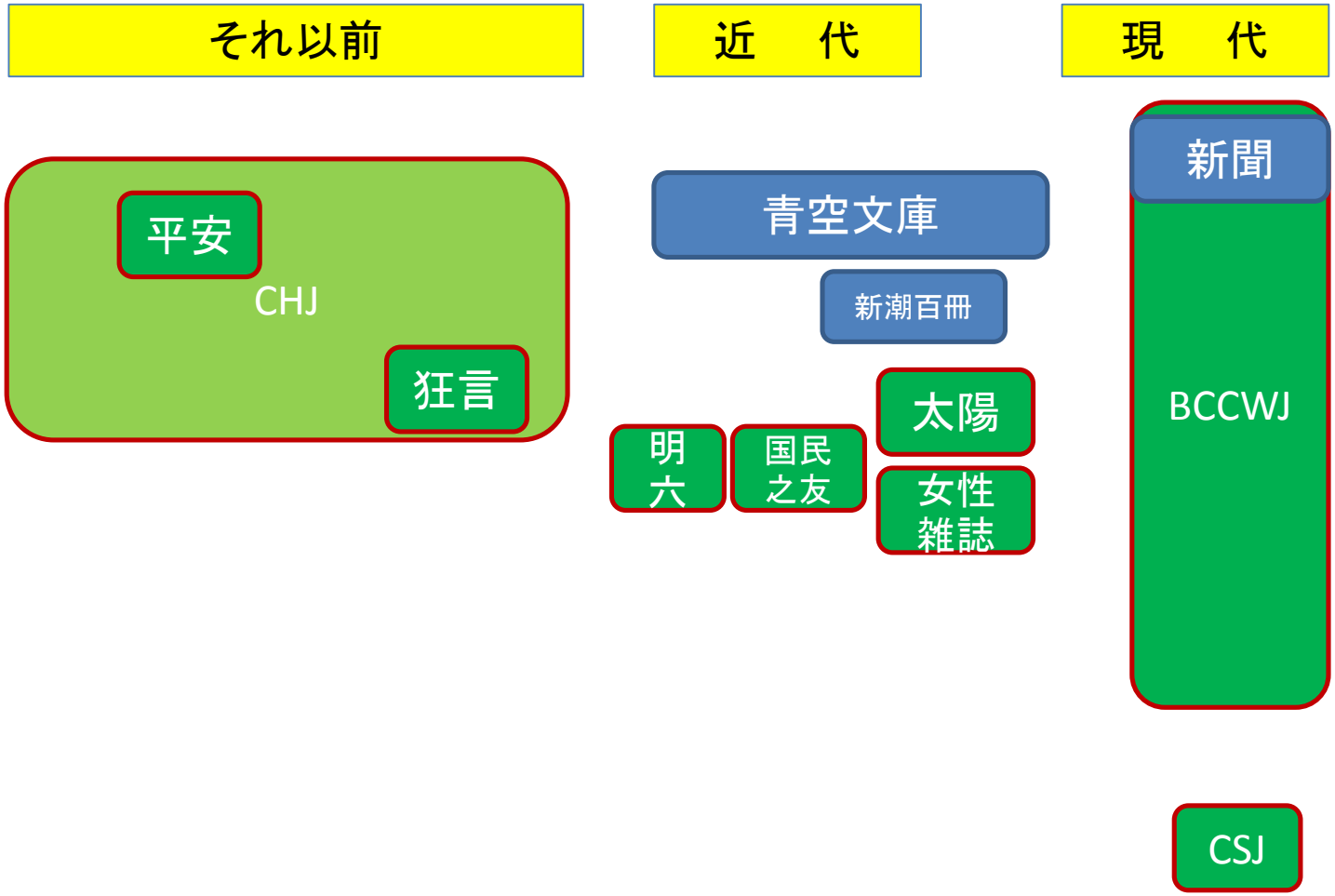
現　代

新聞

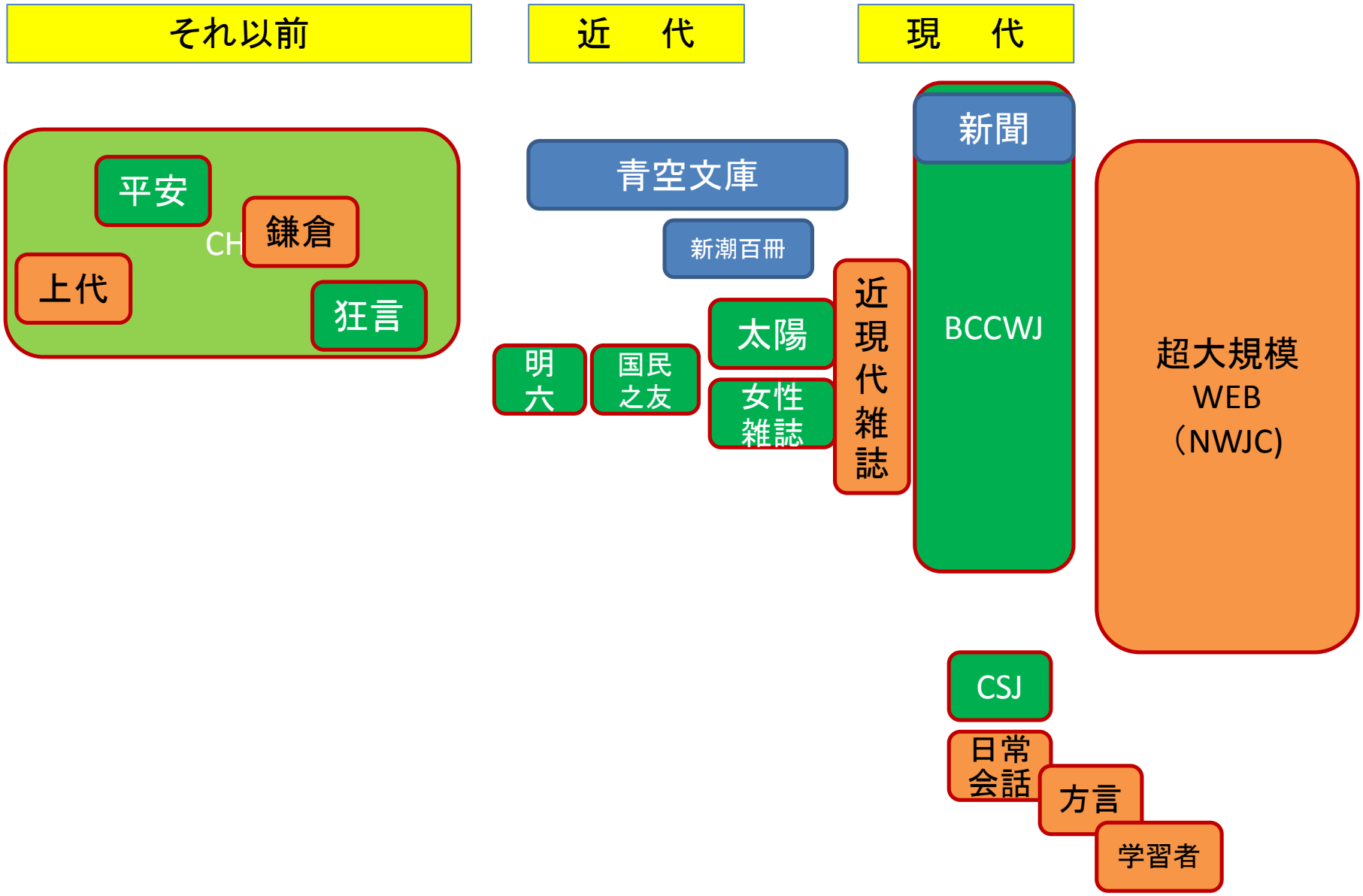
青空文庫

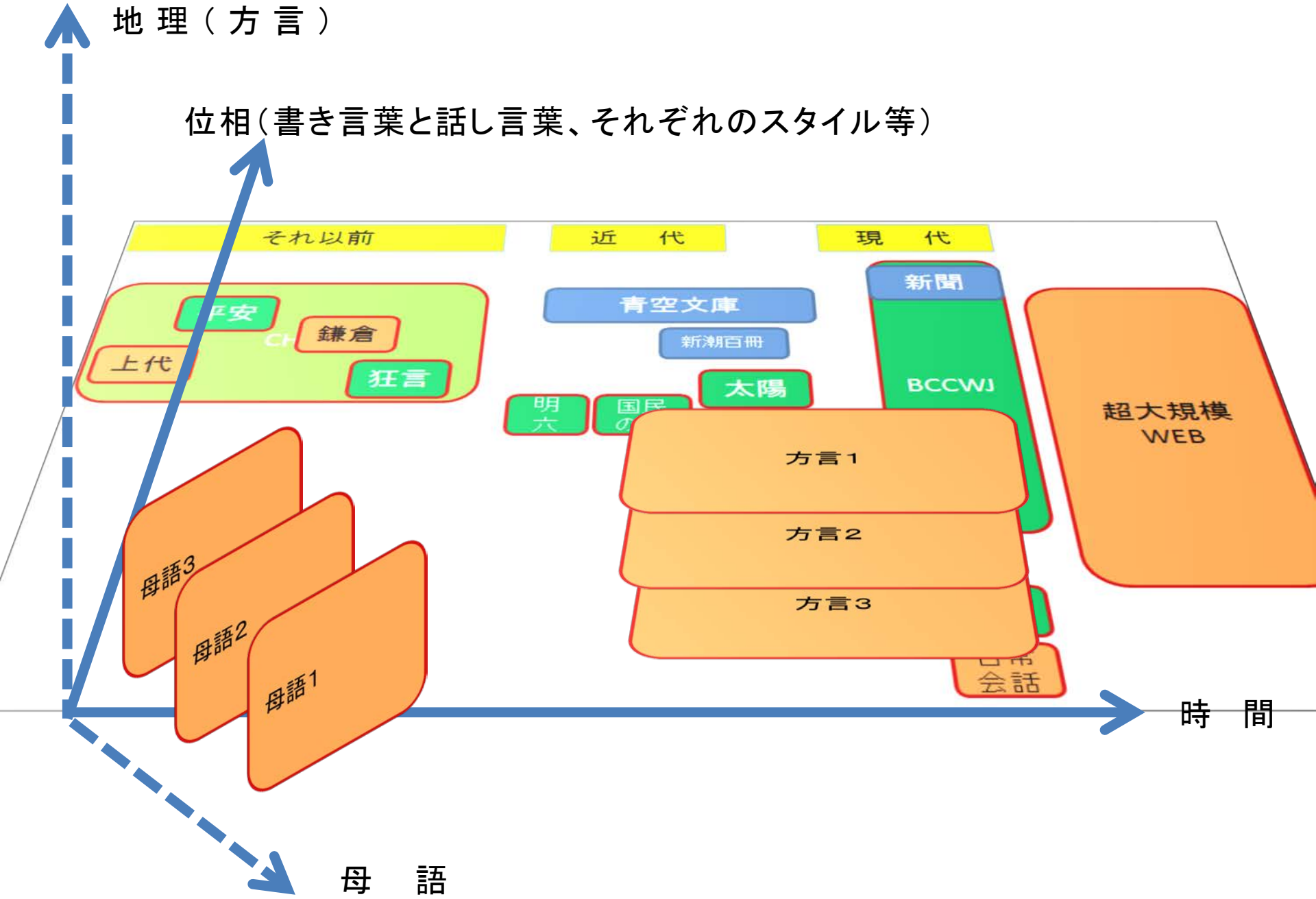
新潮百冊

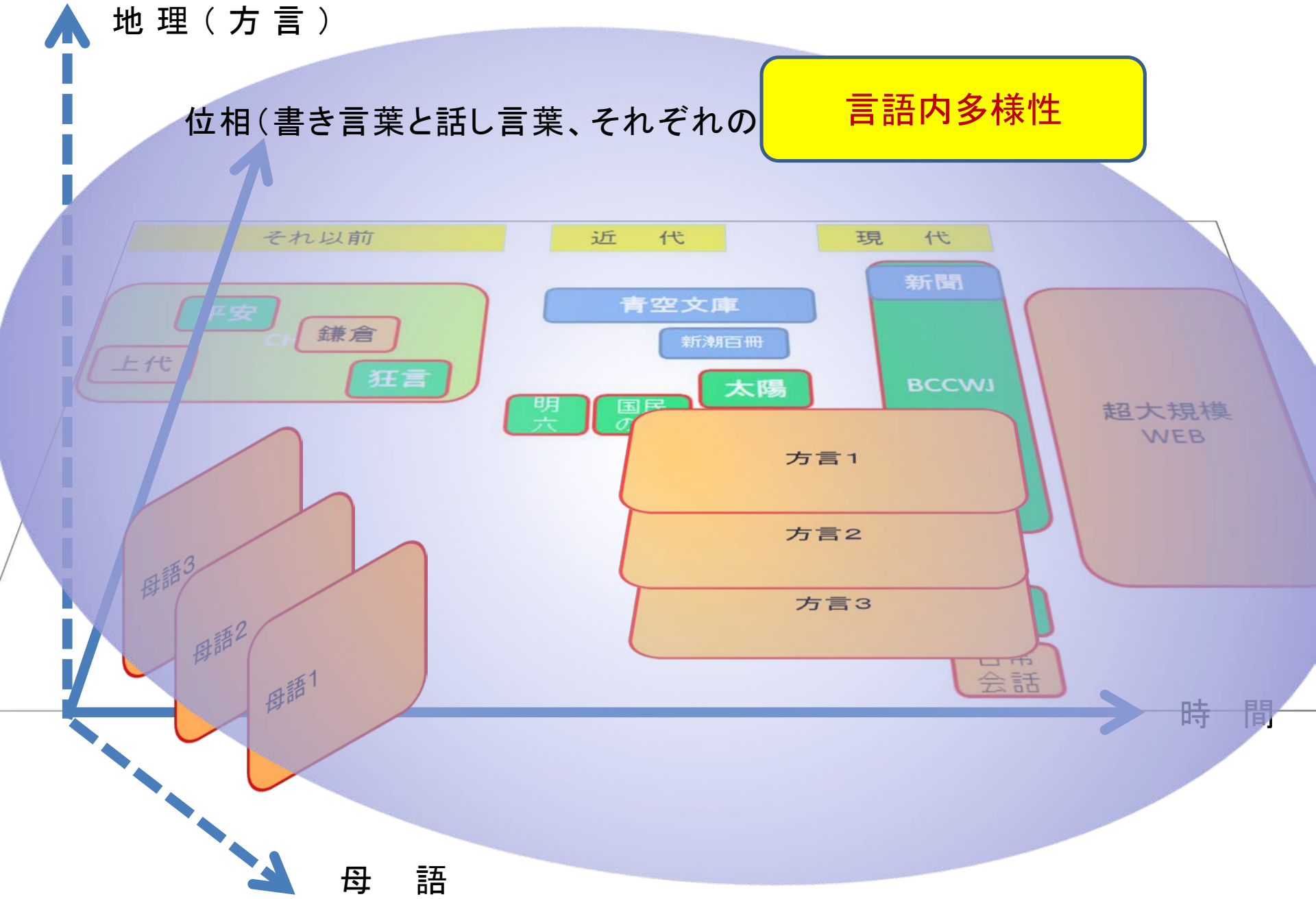
日本語コーパス整備の経緯Ⅱ：現状



日本語コーパス整備の経緯Ⅲ：2021年度末




地理（方言）
位相（書き言葉と話し言葉、それぞれのスタイル等）
それ以前
近　代
現　代
平安
鎌倉
上代
狂言
青空文庫
新潮百冊
明六
太陽
新聞
BCCWJ
超大規模
WEB
方言1
方言2
方言3
母語3
母語2
母語1
時　間
母　語


地理（方言）
位相（書き言葉と話し言葉、それぞれの
言語内多様性
それ以前
近　代
現　代
上代
鎌倉
狂言
青空文庫
新潮百冊
太陽
新聞
BCCWJ
超大規模
WEB
方言1
方言2
方言3
母語3
母語2
母語1
時　間
母　語

コーパスが捉えた日本語の変異

内省できますか？

- 「NHK」はどのように発音されているか？
- 「日本」は？
- 「来られる」と「来れる」は話し言葉でどちらが多い？
- 「読むです」「行くです」等と書く人はいるか？
- 「読めれる」「行けれる」は？
- 「すらさえ」と「さえすら」はどちらが正用？
- Etc.

発　音	頻　度	
エヌエチケー	132	
エネーチケー	24	
エヌエッチケー	9	← 発音辞書の見出し
エヌエイチケー	7	← 発音辞書の見出し
エヌエチケ	3	
エネーチケ	3	
エネエチケー	2	
エヌエスケー	1	
エヌチケー	1	
エネーシケー	1	

『日本語話し言葉コーパス』(講演などのモノローグ中心、752万語)の検索結果

LUW	ニッポン	ニホン	総計	%Nippon
日本一	9	31	40	22.5
日本代表	7	29	36	19.4
日本列島	1	24	25	4.0
日本	122	3108	3230	3.8
西日本	1	30	31	3.2
日本語教育	2	64	66	3.0
日本人	19	1019	1038	1.8
日本語	8	1591	1599	0.5
現代日本語		20	20	0.0
中世末期日本語		25	25	0.0
日本円		20	20	0.0
日本海		26	26	0.0
日本海側		30	30	0.0
日本語らしい		26	26	0.0
日本語らしさ		26	26	0.0
日本語学習者		23	23	0.0
日本語教師		21	21	0.0
日本語話者		148	148	0.0
日本酒		43	43	0.0
日本中		23	23	0.0
日本的		27	27	0.0

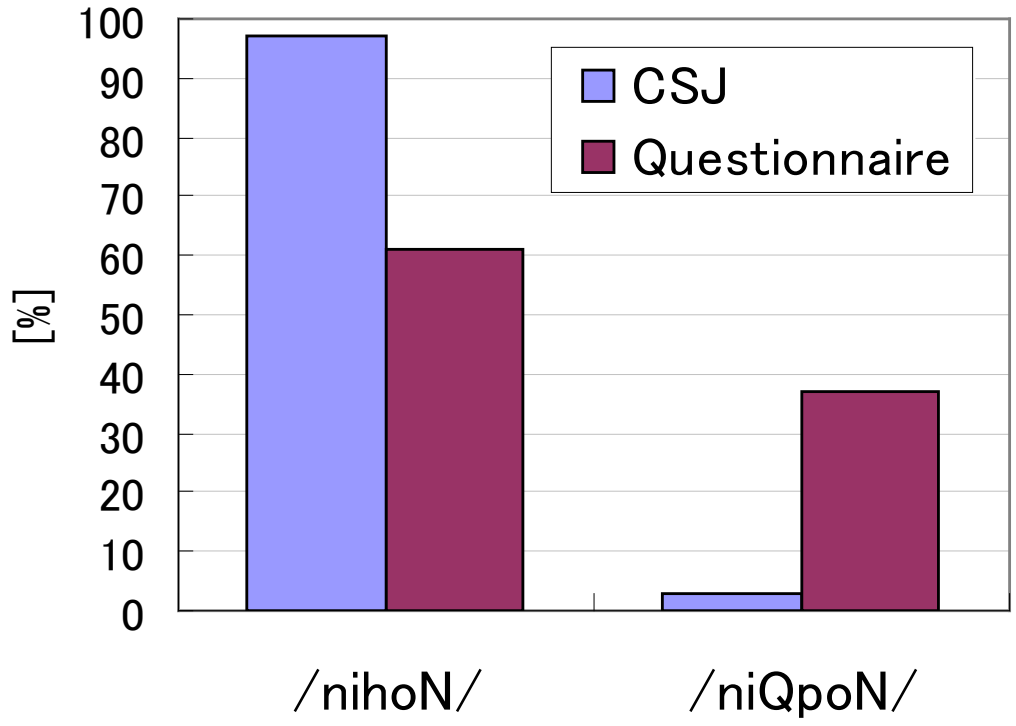
- CSJに関する限りニホンの圧勝。ニッポンは全体として3%以下
- 「日本一」「日本代表」では比較的ニッポンになりやすい

- 『日本語話し言葉コーパス』に記録された行動

  「ニホン」(97%)「ニッポン」(3%)

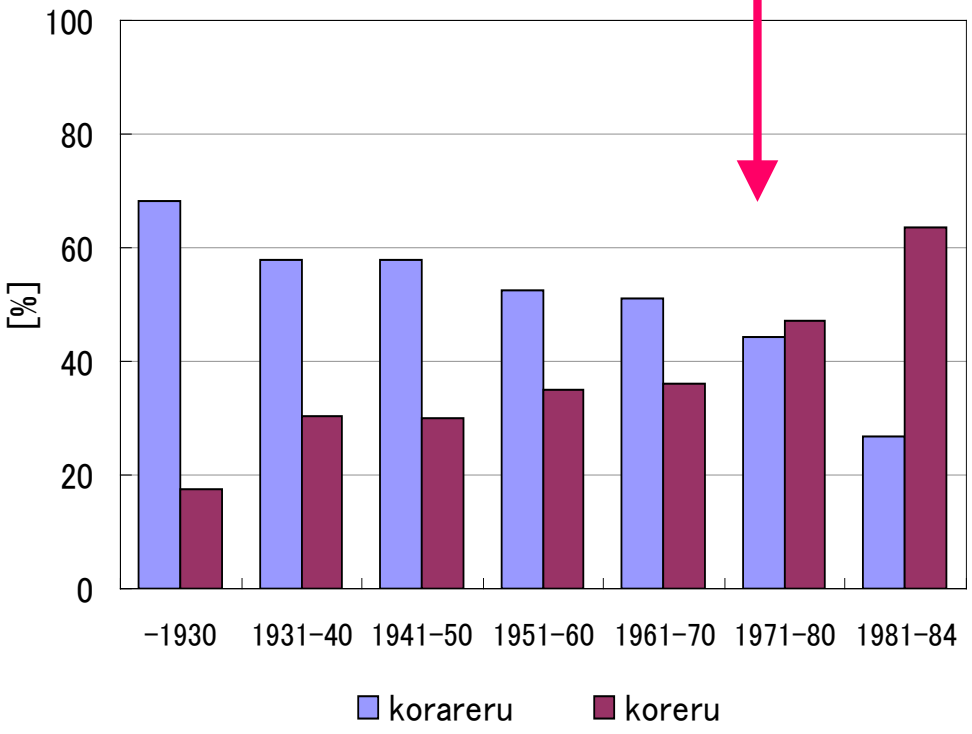
- NHK放送文化研究所によるアンケート(2004)

  「ニホン」(61%)「ニッポン」(37%)

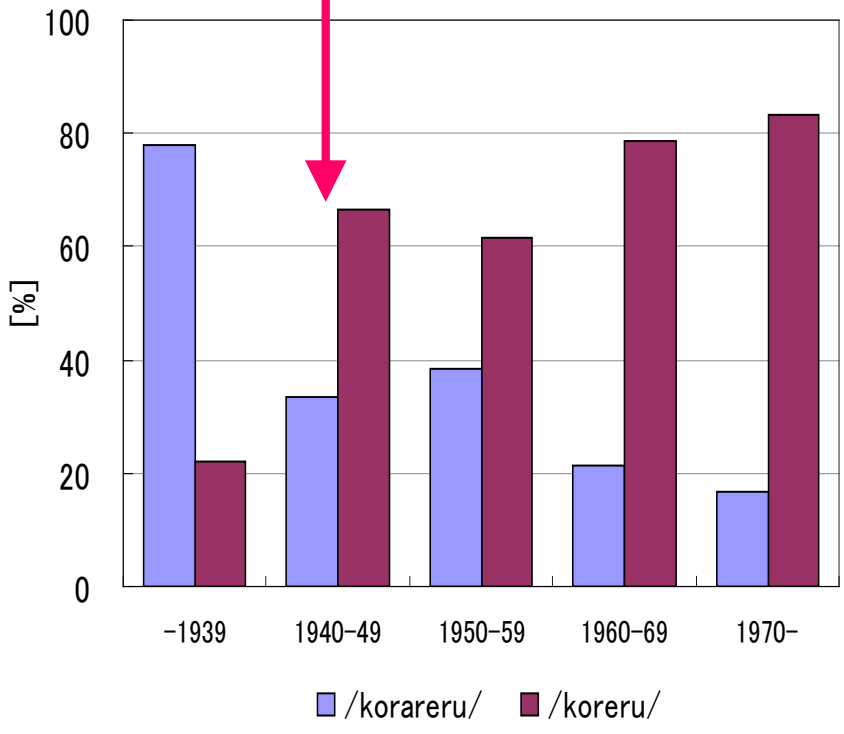


「コラレル」か「コレル」か

逆転のタイミングに30年のずれ



文化庁国語課による世論調査
2001



『日本語話し言葉コーパス』における行動

観察結果

表現	BCCWJ（1億語）	NWJC（200億語）
形容詞＋デス	11,000	393,619
動詞＋デス	232	7,172
～シナ（サ）ソウ	75	312
～シナイベキ	11	205
～サセラレ／ラレサセ	479/0	7450/6
～スラサエ／サエスラ	0/0	1/2
～サエホド／ホドサエ	0/0	0/2

拙稿「コーパスの存在意義」（2013, 朝倉講座1巻）などでとりあげた誤用と正用の境界線上の例をBCCWJと開発中の『国語研日本語ウェブコーパス』で検索した。BCCWJは1億語、『国語研日本語ウェブコーパス』は現状で200億語以上。

「動詞＋です。」
『現代日本語書き言葉均衡コーパス』

74 件の結果が見つかりました。

サンプルID	前文脈	キー	後文脈	語彙素	品詞	活用型	活用形	執筆者	書名/出典
LBb7_00014	調|の|ド|レ|ミ|が|読め|たに|と|に|	なる	|(です)|(。)| |だ|って|, |われわれ|日本|人‖は|	成る	動詞-非自立可能	五段-ラ行	終止形-一般	相原　末治(著)	やさしい楽譜の読み方
LBe9_00215	は|観光|ガイド|で|、|お|客|さん‖に|案内|	する	|(です)|(。)|相手|の|お|客|さん‖は|香港|の|	為る	動詞-非自立可能	サ行変格	終止形-一般	笹倉　明(著)	東京難民事件
LBi9_00179	の|スポーツ|カー‖に|も|乗っ|た|と|言っ|て|	いる	|(です)|(。)|その|上|一|山| |三森‖は|ちょっと|	居る	動詞-非自立可能	上一段-ア行	終止形-一般	＝(泡)坂　妻夫(著)	湖底のまつり
LBj9_00039	犯人|を|当て|きる|と|、|あと|が|書け|なく|	なる	|(です)|(。)|それ|で|も|、|あなた|、|僕|の|	成る	動詞-非自立可能	五段-ラ行	終止形-一般	坂口　安吾(著)	不連続殺人事件
LBm3_00025	携帯|電話|)|も|、|また|その|周辺|ビジネス|も|	然り	|(です)|(。)| |利益|の|ある|ビジネス|、|時流|に|	然り	動詞-一般	文語ラ行変格	終止形-一般	新村　貢一(著)	個人事業の始め方・儲け方
LBn9_00104	世間|の|信用|を|墜し‖は|すまい|か|と|	思ふ	|(です)|(。)|又|、|いづれ|憲|君|を|出せ‖ば|	思う	動詞-一般	五段-ワア行	終止形-一般	徳田　秋聲(著)	徳田秋聲全集
LBq1_00036	か|わから|ない|。| |彼女|と|吉田|くん|も|	然り	|(です)|(。)| |ただし|、|ここ|で|確認|し|て|	然り	動詞-一般	文語ラ行変格	終止形-一般	秋元　康(著)	「わがままな女」になろう
OB2X_00153	文明|を|危機|的|状況|に|直面|は|せ|て|	おる	|(です)|(。)|科学|と|は|科学|その|もの|の|ため|	居る	動詞-非自立可能	五段-ラ行	終止形-一般	村上　春樹(著)	世界の終りとハードボイルド・ワンダーランド
OB2X_00153	それぞれ|の|骨‖に|は|それぞれ|固有|の|音|が|	ある	|(です)|(。)|それ‖は|まあ|言う|なれ|ば|隠さ|れ|	有る	動詞-非自立可能	五段-ラ行	終止形-一般	村上　春樹(著)	世界の終りとハードボイルド・ワンダーランド
OB2X_00153	ない‖はず|です|」| |「私‖は|知っ|て|	おる	|(です)|(。)|「組織」|の|上層|部|と|は|	居る	動詞-非自立可能	五段-ラ行	終止形-一般	村上　春樹(著)	世界の終りとハードボイルド・ワンダーランド
OB2X_00153	と|か|、|そう|いう|こと|を|研究|し|て|	おる	|(です)|(。)|まあ|これ|を|ごらん|なさい|」| |彼|	居る	動詞-非自立可能	五段-ラ行	終止形-一般	村上　春樹(著)	世界の終りとハードボイルド・ワンダーランド
OB2X_00153	に|戦車|を|並べ|て|喜ん|で|いる|金持|も|	おる	|(です)|(。)|私‖は|頭骨|を|集め|て|おる|です|	居る	動詞-非自立可能	五段-ラ行	終止形-一般	村上　春樹(著)	世界の終りとハードボイルド・ワンダーランド
OB2X_00153	を|収集|できる|と|なる|と|拷問|の|必要|も|	なくなる	|(です)|(。)|相手|を|殺し|て|肉|を|そぎおとし|、|	無くなる	動詞-一般	五段-ラ行	終止形-一般	村上　春樹(著)	世界の終りとハードボイルド・ワンダーランド
OB2X_00153	研究|を|奴ら‖は|悪用|しよう|と|し|て|	おる	|(です)|(。)|たとえば|骨|から|記憶|を|収集|できる|と|	居る	動詞-非自立可能	五段-ラ行	終止形-一般	村上　春樹(著)	世界の終りとハードボイルド・ワンダーランド

「動詞＋デス」
『国語研日本語ウェブコーパス』

1. [デュエルマスターズ 名詞 や 助詞] [ポケモン 名詞 の 助詞] [カード 名詞 も 助詞] [描か 動詞 れ 動詞 てる 助動詞 です 助動詞 ねー 助詞]
2. [デモ 名詞 が 助詞] [頻発 名詞 し 動詞 て 動詞 て 助詞] [危ない 形容詞 、 記号] [大手 名詞 メディア 名詞 の 助詞] [嘘 名詞 も 助詞] [感覚 名詞 的 名詞 に 助詞 かも 助詞 しれ 動詞 ない 助動詞 が 助詞 、 記号] [中学生 名詞 、 記号] [見破っ 動詞 てる 動詞 ん 名詞 じゃ 助詞 ない 助動詞 か 助詞 と 助詞] [思う 動詞 です 助動詞 よ 助詞]
3. [デメリット 名詞 を 助詞] [教える 動詞 です 助動詞 か 助詞 ・ 記号 ・ 記号 ・ 記号 良い 形容詞 かも 助詞 しれ 動詞 ませ 助動詞 ん 助動詞 ね 助詞 ★ 記号]
4. [デメリット 名詞 は 助詞] [やり 動詞 すぎる 動詞 と 助詞 、 記号] [くらくら 名詞 する 動詞 です 助動詞]
5. [デメリット 名詞 は 助詞 、 記号] [両方 名詞 お金 名詞 が 助詞] [いる 動詞 です 助動詞]
6. [デメリット 名詞 、 記号] [日本 名詞 の 助詞] [裕福 名詞 で 助動詞] [安定 名詞 し 動詞 た 助動詞] [生活 名詞 に 助詞] [変化 名詞 が 助詞] [起こる 動詞 です 助動詞]
7. [デミオ 名詞 だ 助動詞 と 助詞] [インパクト 名詞 不足 名詞 か 助詞 な 助詞 、 記号 って 助詞] [感じ 名詞 です 助動詞 が 助詞] [扱い 名詞] [安い 形容詞 と 助詞 は 助詞] [思う 動詞 です 助動詞 が 助詞 ね 助詞] [w 名詞]
8. [デポ 名詞 の 助詞] [周り 名詞 に 助詞] [また 副詞] [ショッピング 名詞 センター 名詞 が 助詞] [建つ 動詞 です 助動詞 か 助詞]
9. [デブ 名詞 ちん 名詞 は 助詞] [何 名詞 を 助詞] [さ 動詞 れ 動詞 て 助詞 も 助詞] [ジーッ 名詞 と 助詞] [し 動詞 てる 動詞 です 助動詞]
10. [デフラグ 名詞 統合 名詞 版 名詞 Win 名詞 の 助詞] [チェック 名詞 ディスク 名詞 と 助詞 デ 助詞] [フラグ 名詞 を 助詞] [有効 名詞 に 助詞] [使う 動詞 です 助動詞]
11. [デフラグ 名詞 について 助詞 は 助詞 、 記号] [諸般 名詞 諸々 名詞 の 助詞] [件 名詞 が 助詞] [ある 動詞 です 助動詞 が 助詞]
12. [デフラグ 名詞 なる 動詞 を 助詞] [知ら 動詞 ん 助動詞 でも 助詞 、 記号] [HDD 名詞 内 名詞 の 助詞] [ファイル 名詞 を 助詞] [整理 名詞 整頓 名詞 し 動詞 て 助詞 くれる 動詞 です 助動詞]

「二重可能」(レ足す言葉)

BCCWJ

動詞	レル	レナイ	レタ
行ケ	3	2	1
聞ケ	0	0	0
書ケ	0	0	0
遊べ	0	0	0
歩ケ	0	0	0
出来	0	0	0
描ケ	0	0	0
飛べ	0	0	0
聴ケ	0	0	0
読メ	0	0	0

NWJC

動詞	レル	レナイ	レタ
行ケ	32	28	2
聞ケ	5	1	2
書ケ	4	1	0
遊べ	4	0	0
歩ケ	3	0	0
出来	2	0	0
描ケ	2	2	1
飛べ	1	0	1
聴ケ	1	1	0
読メ	1	1	0

100億語規模のコーパスではじめて把握できるようになる低確率の(しかし実際に生起する)言語現象が少なくない。

観察結果

表現	BCCWJ（1億語）	NWJC（200億語）
形容詞＋デス	11,000	393,619
動詞＋デス	232	7,172
～シナ(サ)ソウ	75	312
～シナイベキ	11	205
～サセラレ／ラレサセ	479/0	7450/6
～スラサエ／サエスラ	0/0	1/2
～サエホド／ホドサエ	0/0	0/2

「～シナイベキ」

1. [デフラグ について は 、] [早く] [完了 さ せる に は] [スクリーンセイバー は] [同じ よう に] [起動 さ せ ない べき です]
2. [デイトレ の] [場合 、] [サラ 場 で] [「 エントリー する べき か 、] [し ない べき か 、] [それ が] [問題 だ]
3. [テレビ の] [前 で] [「 光 ちゃん は] [可愛い けど 、] [録画 する べき か] [し ない べき か ・ ・ ・ 」 と] [葛藤 する] [
4. [テレビ と] [現実 を] [一緒 に し て しまう] [人 は] [見 ない べき です]
5. [テレビ って 、] [何 を] [伝える べき で] [何 を] [伝え ない べき な の か を] [わきまえ て いる と] [思い ます か]
6. [チョコレート を] [あげる べき か 、] [あげ ない べき か で] [悩ん で いる] [人 が] [多い の で は ない でしょ う か]
7. [チャンス は] [逃さ ない べき だろ う な ー と] [思い まし た]
8. [チャリ で] [たばこ 吸わ ない べき です よ ね ・ ・ ・]
9. [チャイナ エアライン は] [乗ら ない べき]
10. [チビ 太 に] [見せる べき か] [見せ ない べき か] [悩ん だ まま 、]
11. [ダッシュ する] [時 も] [ダッシュ す べき か] [し ない べき か の] [迷い も] [見て取れ た し 、]
12. [タル は 、] [合理 的 に は] [3つ を] [越え ない べき で ある]
13. [タク ： タクシー に] [乗ら ない べき] [日]

観察結果

表現	BCCWJ（1億語）	NWJC（200億語）
形容詞＋デス	11,000	393,619
動詞＋デス	232	7,172
～シナ（サ）ソウ	75	312
～シナイベキ	11	205
～サセラレ／ラレサセ	479/0	7450/6
～スラサエ／サエスラ	0/0	1/2
～サエホド／ホドサエ	0/0	0/2

拙稿「コーパスの存在意義」（2013, 朝倉講座１巻）などでとりあげた誤用と正用の境界線上の例をBCCWJと開発中の日本語ウェブコーパスで検索した。
BCCWJは1億語、超大規模Webコーパスは現状で200億語以上。

「～ラレサセ」

(surface:られ)[0,0] (surface:させ)[1,1]

1 - 6 / 6 (0.969801902771 sec)

1. [コイツ だけ は] [ヤバ い 、 と] [サンジ は] [自分 の] [身 を] [切り裂く] [巨 き 過ぎる] [男根 に 、] [脳 まで] [突き通さ れる よう な] [感覚 に] [一瞬] [息 を] [止め られ させ ながら 、] [そう] [思う]
2. [クラブ の] [存続 の] [ため に] [何 か] [手伝える] [こと は] [ない か 、 と] [考え られ させ まし た]
3. [オザンナ] [(主 の] [名 により] [来 たれる] [者 は] [祝せ られ させ] [給え 、] [最 高き] [ところ まで)]
4. [イエズス・キリスト 、] [祝せ られ させ] [賜え]
5. [アルツハイマー の] [人 は 、] [人 の] [生き方 を] [考え られ させ られ まし た ね] [(*´ エ `*)]
6. [それに 、] [すっごく] [謎 って] [かんじ で] [読ん で い たら] [よく] [考え られ させ られ ます] [w w]

レジスターの重要性

表現	BCCWJ	NWJC
～ヲ泣ク	1	0
～ヲ死ヌ	3	0

BCCWJで得られた用例のレジスターは、PV（書籍）、LB（図書館書籍）、OB（ベストセラー）、OV（韻文）。先の検索結果は、Webのデータには、相対的に見てこの種の文書が少ないことを示唆しているように思える。

⇒　BCCWJのような均衡コーパスの必要性

『特許ライティングマニュアル』より

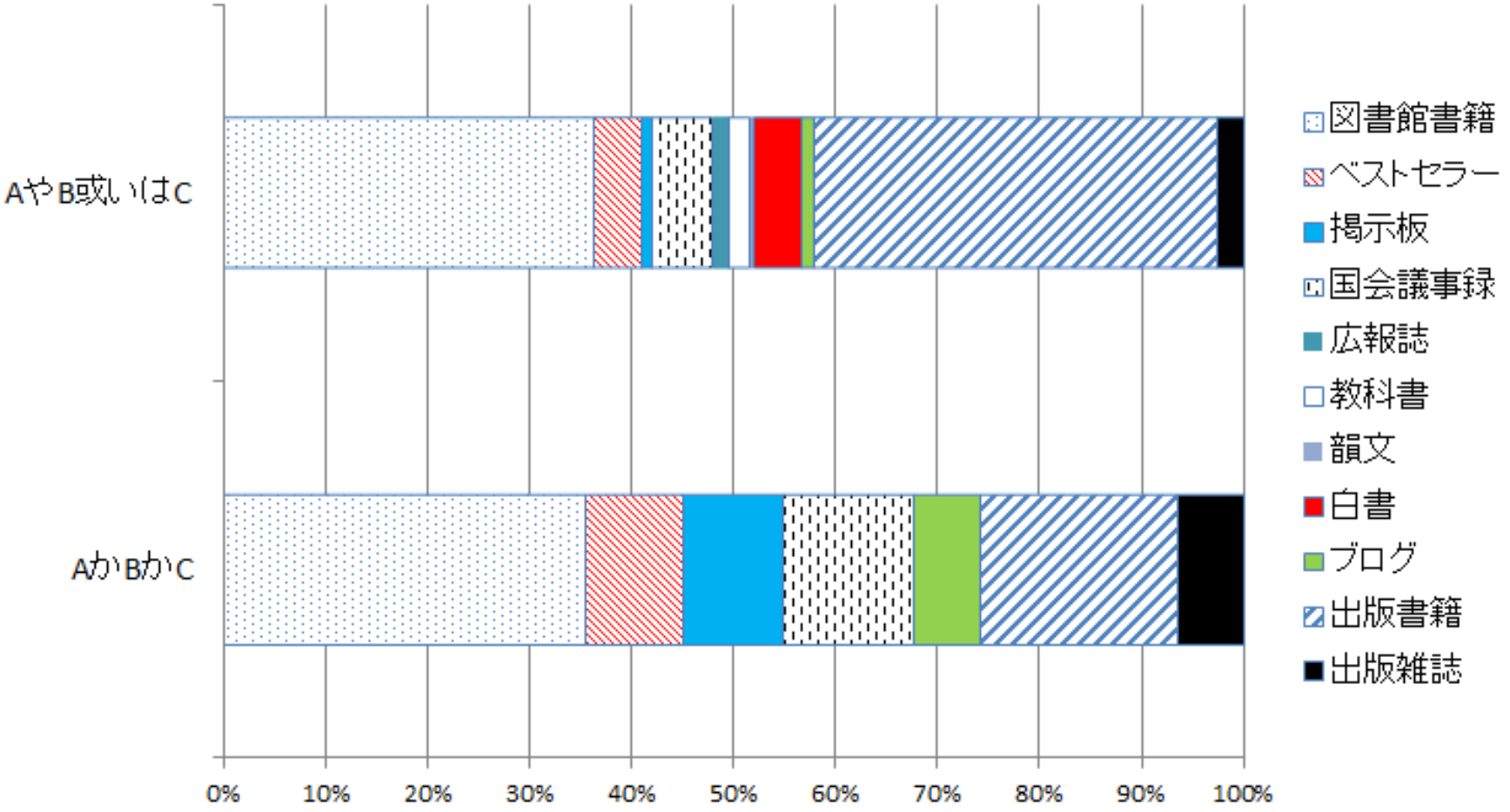
ルール第F条の１：並立表現の並立要素が同じ表現になるように整える。

改善例：
「水やスチーム、或いは、薬品など」
⇒「水かスチームか薬品など」

BCCWJの検索結果

AやB(、)或いはC	⇒	288例
AかBかC	⇒	31例

レジスター差


AやBあるいはC
AかBかC
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
図書館書籍
ベストセラー
掲示板
国会議事録
広報誌
教科書
韻文
白書
ブログ
出版書籍
出版雑誌

おわりに

言語研究と情報科学の連携

- 現代の言語資源開発は、情報科学なくして考えられない。機械学習（人工知能）の研究とも関係が深い
- 言語研究からの情報科学への貢献は、日本語の全体像を伝える良質な言語資源の提供
- 今後、複雑多様な言語を過度に単純化するのではなく、複雑さを保ってモデル化する研究を進めるにあたっては、両者の協力が不可欠