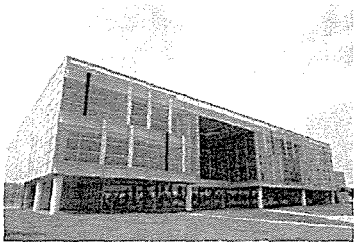
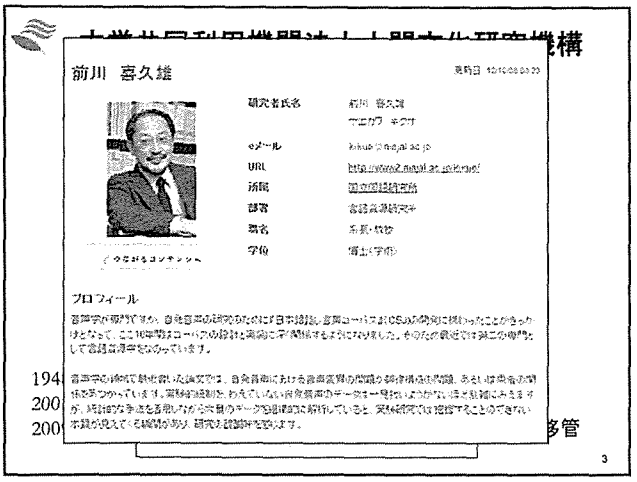
国立国語研究所における言語資源開発の近況

前川 喜久雄
大学共同利用機関法人人間文化研究機構
国立国語研究所
言語資源研究系・コーパス開発センター

大学共同利用機関法人人間文化研究機構 国立国語研究所 (東京都立川市)



1948年文部省直轄研として設立(1968年より文化庁)
2001年独立行政法人化
2009年独立行政法人廃止 ⇒ 大学共同利用機関法人へ移管



今日のトピック

近年、自然言語処理技術(NLP)の発展が言語研究にも大きな影響を与えるようになってきた。言語研究は文学部の占有事項ではなくなってきている

ただしNLPがそのまま言語研究に使えるわけではない。今日はコンピュータを使って日本語を語学的にきちんと研究できるようにするためにどのようなインフラ整備が必要であるかについて述べ、講演者の研究グループがそのなかで果たしてきた役割を紹介する

講演者の専門は音声の研究だが、この領域の例はわかりにくいので、今日は音声学の話は封印する。

母語研究のひとつの問題

■「言語直観」の仮定

日本語の母語話者は現代日本語の文法を完全に「知って」いる。従って与えられた例文が文法的(適格)かどうかを即座に判断できる (理由は説明できなくともよい)

例:
*「山がそびえた」
*「反対しないべきだ」
*「反対しなそうだ」
*「ジョンはメアリーを泣いた」
*「ジョンは命を死んだ」
*「この壁紙は黄色いでした」

母語研究の問題

■「言語直観」の仮定

日本語の母語話者は現代日本語の文法を完全に「知って」いる。従って与えられた例文が文法的(適格)かどうかを即座に判断できる (理由は説明できなくともよい)

例:	言語学的な説明
*「山がそびえた」	• 常に「そびえている」の形で用いる
*「反対しないべきだ」	• 命題中に否定をふくむことは不可
*「反対しなそうだ」	• 同上
*「ジョンはメアリーを泣いた」	• 自動詞だからヲ格をとらない
*「ジョンは命を死んだ」	• 同上
*「この壁紙は黄色いでした」	• 形容詞述語に「です」は不要

金田一春彦「国語動詞の一分類」(1950)

動詞として挙げたいものは、(中略)、ある状態を表す動詞と言いたいものであるが、例えば「山が聳える」がこれである。

いつも「ーている」の形で状態を表すのに用い、ただ「聳える」だけの単独の形で動作・作用を表すために用いることがないのを特色とする。

7

寺村秀夫「ムードの形式と否定」(1979)

…思弁の遊びのように思われるかもしれないが必ずしも…日本語を外国語として習得しようとする者にとっては、…にはダロウの前の部分を否定にする、ダロウは否定の形…る、逆に、〜ベキダのように、その否定は〜ベキ(デ)…イベキダという形にない、一方、ハズダのように、内面…の判断を否定にできるものもある、等々といったことは必要な文法的知識であるに相違なく、また「内と外」両面の否定があり得るとき、それはどう意味的に違うのか、といったこともまた必要な文法的知識に属するであろう。これらの問いに多少とも統一的に答えるためには、どうしても上のような議論を避けて通るわけにはいかないのである。

8

さて、実際は

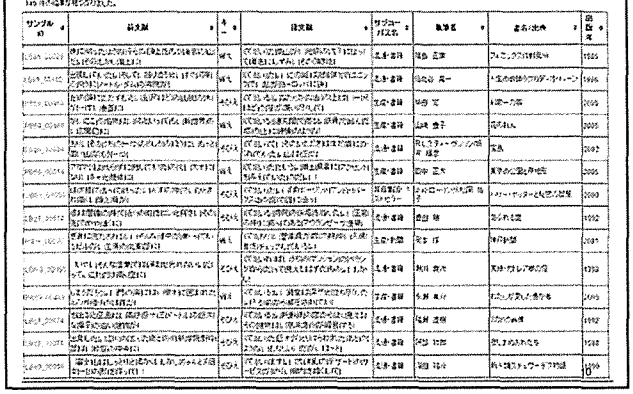
理論をうのみにするのではなく、世の中で実際に用いられている用例を偏りなく、大量に観察してみる

⇒ 国立国語研究所が昨年公開した『現代日本語書き言葉均衡コーパス』(BCCWJ,1億語規模)を検索

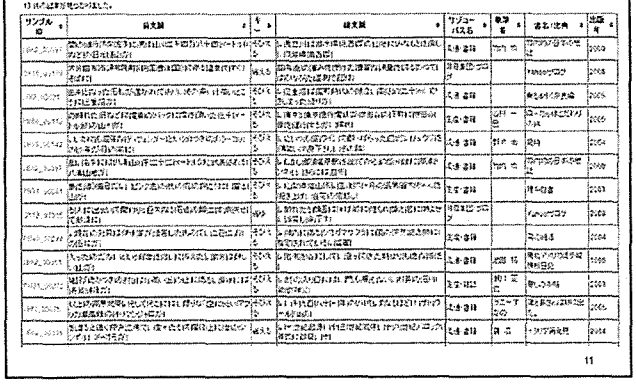
まずは「そびえている」から

9

「そびえている」が多数派



しかし「そびえる」も13例みつかる



11

「そびえる。」

- 右手には八溝山(千二十二メートル)に代表される八溝山地がそびえる。(LBo2_00097 竹内均)
- いかにも屈強なディフェンダーという体つきの「ラーヨ」の少年が目の前にそびえる。(LBs9_00142 野沢尚)
- そしてそこには、限りなく空に近いアフリカ最高峰のキリマンジャロがそびえる。(LBt2_00075 ダニーすがの)

やはり金田一が第四種動詞とする「優れる」の場合、「優れている。」123件に対して「優れる。」が17件。

12

「～しないべき」

- 脱ぐべきか、脱がないべきか。(PM31_00671 安田拡了ほか)
- それとも一人では生きていかないべきですか？(OC14_04650 Yahoo!知恵袋)

「～すべきでない」468例に対し「～いないべき」11例
文型に制約？

13

「～を泣く」

- それから、こゝに、弟がおれの不仕合せを泣いてゐる。(PB29_00682坪内逍遙)

1億語中に1例のみ
戦前(明治末？)の用例
意味はわかる

14

「～を死ぬ」

- 土にかえるまえにかれらは土の色に一つの死を死んだ(OV2X_00035田村隆一)
- そのとき私たちは地獄にもまさる苦しみの果てに、悲惨きわまりない死を死んでゆかねばならないだろうということである。(OB0X_00017五島勉)
- 西郷さんは、「自分の死」を死んだ経験者だった。(LBl2_00090南伸坊)

3例すべて「死を死ぬ」

15

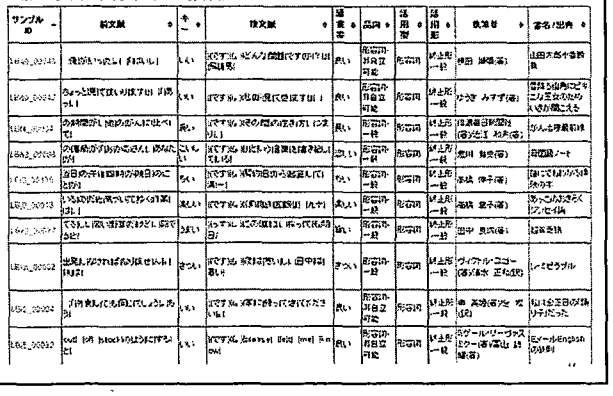
「～を生きる」

- 「起承転転、すなわち人生は最後までころがりっ放し」を頭の中に置きながら、毎日を生きている。(LBn1_00043 童門冬二)
- 薔薇の力も、千年を生きる力も、その一部。(LBp9_00116 三田誠)
- 他の政治家も、自分のドラマを生きたいという欲望はもっているでしょう。(LBt3_00101福田和也)

20,654例中847例がヲ格
ヲ格名詞の性格もかなり多様

16

「形容詞＋です。」

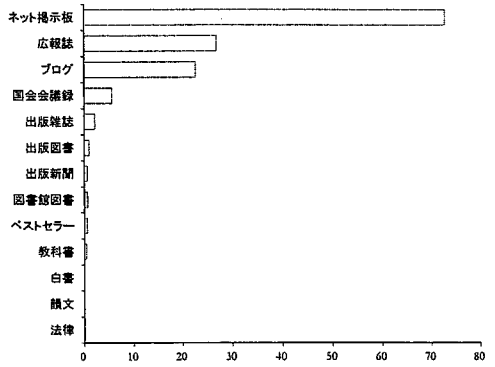


「形容詞＋です。」率の変動

略号	レジスター	A_	A+です。	%A+です。
OL	法律	296	0	0.0
OV	韻文	39	0	0.0
OW	白書	1772	0	0.0
OT	教科書	1015	4	0.4
OB	ベストセラー	5917	35	0.6
LB	図書館書籍	44058	296	0.7
PN	出版新聞	1676	11	0.7
PB	出版書籍	35634	353	1.0
PM	出版雑誌	5724	131	2.2
OM	国会会議録	924	55	5.6
OY	ブログ	7879	2314	22.7
OP	広報誌	197	72	26.8
OC	ネット掲示板	2950	7883	72.8

18

「形容詞＋です。」率の変動



19

ここで、ちょっと注意

- ここまで挙げてきた例は「使われているかもしれないが、誤った日本語だ」という立場
- 規範を示す立場（国語教育）としてはありうる
- しかし「正しい」日本語だけを研究していたのでは、日本語の情報処理はできない。外国人に対する日本語教育にも支障がでる
- 規範を離れた（非情な）立場での研究が必要
  肺魚も魚、ダチョウも鳥という立場

20

ここまでをまとめると

- ある表現を「言うか言わないか」あるいは「どう言っているか」の判断は予想外に難しい。ケースによっては内省に著しい限界がある
- 問題はなにが「ケースによる」かがわからないこと
- 非情に世界を見渡してデータを集める必要がある

21

非情な言語研究のために何が必要か

- 偏りのない大量の電子化データ構築　**コーパス**(corpus)
- データへの検索用情報付与　**形態素解析、係り受け解析** etc.
- データの分析技術　**種々の統計手法等**

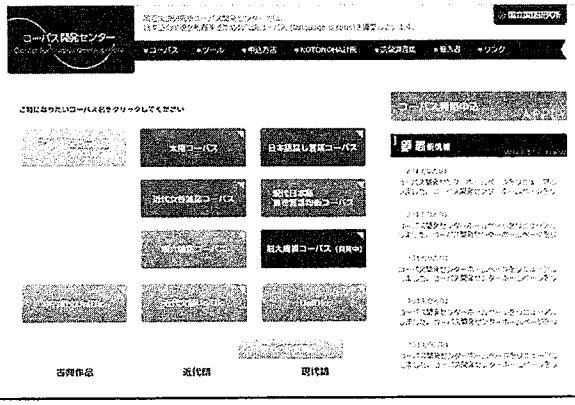
言語資源学

22

国立国語研究所による日本語コーパスの開発

23

言語資源研究系が開発してきたコーパス



『日本語話し言葉コーパス』
Corpus of Spontaneous Japanese (CSJ)
2004年公開

25

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

利用目的：音声認識研究と言語研究
対象：現代日本語の知的なモノローグ
規模： 662時間（752万語）
学会での研究発表と一般的なスピーチが中心（330H）
対照用に対話音声(12H）と朗読音声(20H）も
付加情報：非常に豊富
音声信号、話者情報、転記テキスト、形態論（品詞）情報、節単位情報、係り受け情報、要約・重要文情報、談話構造情報、分節音ラベル、韻律ラベル(X-JToBI)、講演が与える印象の評価
⇒ XML文書として統合
開発：国立国語研究所、情報通信研究機構、東京工業大学

世界の主要話し言葉コーパスとの比較

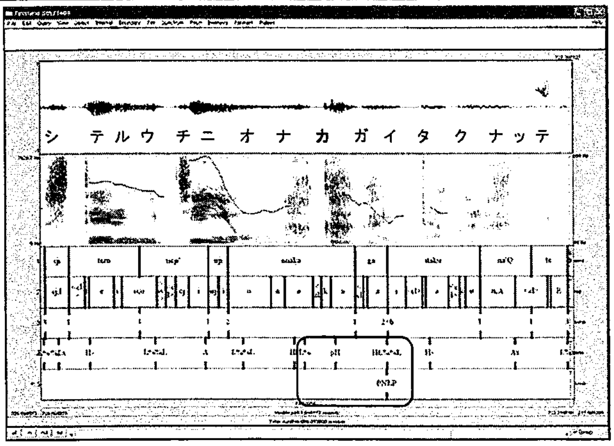
名称	音声タイプ	音声公開	話者	時間	語数(万)	転記	形態	分節	イントネ	言語
CSJ	独話中心	○	1408	661	752	○	○	○	○	日本語
Switchboard Corpus	対話	○	543	250	---	○	○	○	×	米語
CSAE	対話	○	---	6	5	○	×	×	○	米語
CSPAE	演説・記者会見、教授会	×	400	---	200	○	○	×	×	米語
Map Task Corpus	対話	○	64	34	---	○	×	×	×	英語
BNC	独話,対話,朗読,放送	×	2700	1220	1000	○	△	×	×	英語
現代漢語口語對話語料庫	対話	○	60	27	---	○	○	○	×	中国語

音声を公開しているコーパスとしては世界最大、付加情報は抜群に豊富

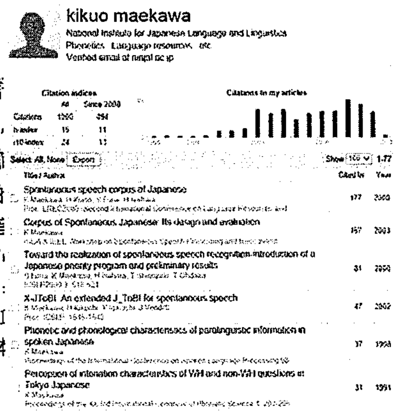
X-JToBI

- 東京語の自発音声を対象とした韻律ラベリングスキーム
- J_ToBI（実際上朗読音声用）を拡張
- 音声信号との対応づけが可能
- Word層、Segment層、Tone層、BI層、Prominence層、Misc層にわけて情報を記述
  ⇒ BI: Break Index 後半ではただBreakと呼んでいる

⇒ 自発音声の研究には非常に有益



- 日本言
- CSJを,
- 博士言 …など）
- 代表白
  - 母音
  - /z/
  - /b, :
  - PNI …能の解明
  - 日本



『太陽コーパス』
Taiyo Corpus
2005年公開

31

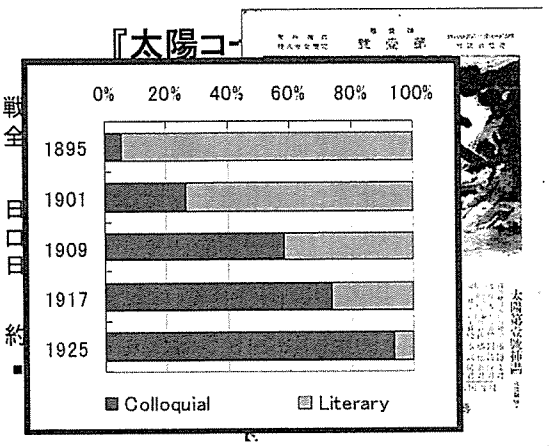
『太陽コーパス』

- 戦前の有名な総合雑誌である『太陽』(1895-1925)の全文テキストコーパス
- 日本語書き言葉の文法が、いわゆる文語文法から口語文法へと急速に変化した時期(言文一致期)の日本語の記録
- 約700万語(形態素解析されていないので推計)
  - 近い将来に形態素解析データを公開予定

32

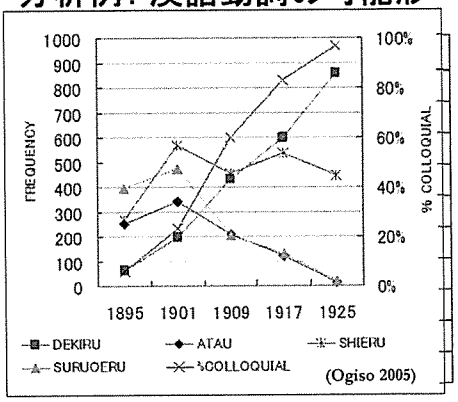
『太陽コー

- 戦
- 日
- 約



33

分析例: 漢語動詞の可能形



34

『現代日本語書き言葉均衡コーパス』

出版(生産実態)サブコーパス
2001～2005年に出版された書籍、雑誌、新聞
3500万語

図書館(流通実態)サブコーパス
3500万語

特定目的(非母集団)サブコーパス
ウェブ上の文書、白書、教科書、国会会議録、ベストセラー等
対象期間はさまざま、最長30年。3500万語

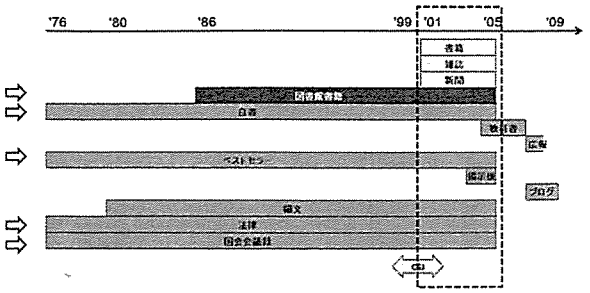
35

BCCWJを構成するレジスター

サブコーパス	レジスター	サンプル(個)	語数(万語)
出版サブコーパス	書籍(PB)	10,117	2,855
	雑誌(PM)	1,996	444
	新聞(PN)	1,473	137
図書館サブコーパス	書籍(LB)	10,551	3,038
特定目的サブコーパス	白書(OW)	1,500	488
	教科書(OT)	412	93
	広報紙(OP)	354	376
	ベストセラー(OB)	1,390	374
	Yahoo!知恵袋(OC)	91,445	1,026
	Yahoo!ブログ(OY)	52,680	1,019
	韻文(OV)	252	25
	法律(OL)	346	108
	国会会議録(OM)	159	510
合計		172,675	10,493

36

各レジスターの時間区分



37

日本語コーパス構築の問題点

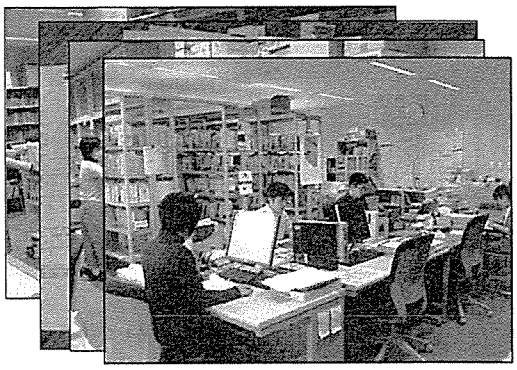
38

日本語コーパス開発上の困難

- 日本の著作権法
  権利者の保護に手厚い。ユーザーとしてはきわめて不便。
- 日本語表記の特異性
  多くの文字を併用(漢字、ひらがな、カタカナ、Roman alphabet 等)
- いわゆる膠着語(こうちゃくご)的性格
  語の境界がはっきりしない ⇒ 分かち書きの習慣がない

39

著作権処理の努力



40

日本語検索の難しさ

- 「国語」 ⇒ 国語、中国語、韓国語、外国語, etc.
- 「リズム」⇒ リズム、アルゴリズム、フォルマリズム, etc.
- 「明かり」⇒ 「明り」「灯り」「燈」「あかり」, etc.
- 「やはり」⇒ 「やはり」「やっぱり」「やはし」, etc.

41

感動詞「嗚呼」

- BCCWJには33種類の表記がある

あ～　あぁ　ああ　あゝ　あァ　あー　アゝ　アァ
アア　アー　吁嗟　嗚呼　嗟　嗟呼　噫　於戯
臆　あ～っ　ああっ　あーっ　アーッ　嗚乎　ぁああ
あああ　あああ　ああー　アアア　あ～あ　あーあ
あーア　アーア　嗚～呼　あああ～

42

人名「ヒロシ」

■ BCCWJ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 宏志 宏至 容 寛 寛志 広 広史 広士 広志
廣 廣司 廣志 弘 弘史 弘司 弘士 弘志 弘至 弘視
恕 日呂志 昊 普 比呂志 比露思 汎 汪 洋 洋司
洋士 洋志 洪 洽 浩 浩嗣 浩士 浩志 浩至 溥 滉
熙 碩 礼嗣 秦 紘 緯 裕 裕史 裕士 裕志 飛呂志
鴻 凞 Hiroshi

43

複合動詞「沸き起こる」

■ 終止形だけで20種

わきおこる わき起こる わき起る 沸きおこる 沸き起こる 沸き起る
沸起こる 涌きおこる 涌き起こる 涌き起る 涌起こる 湧きおこる
湧き起こる 湧き起る 湧起こる 湧起る わきおこれる 沸き起これる
沸き起れる 沸起れる

■ すべての活用形を考慮すると324種

沸きおこれ 沸きおこりゃ 沸きおころ 沸きおこれ 沸きおころう
沸きおころっ 沸きおころ 沸きおこら 沸きおこん 沸きおこる 沸きおこん
沸きおこる 沸きおこん 沸きおこり 沸きおこっ 沸きおこん 沸きおこりゃ
沸き起これ 沸き起こりゃ 沸き起ころ 沸き起これ 沸き起ころう
沸き起ころっ 沸き起ころ 沸き起こら 沸き起こん 沸き起こる 沸き起こん
沸き起こる 沸き起こん 沸き起こり 沸き起こっ 沸き起こん 沸き起こりゃ
Etc.

44

日本語検索の難しさについての結論

■ 日本語のコーパスは単にテキストを電子化しただけでは使い物にならない

■ テキストを単語に区切って、表記の多様性と用言の活用による多様性を吸収し、いわゆる「原形」で検索できるように加工する必要がある

■ 形態素解析（形態論情報アノテーション）

45

日本語の形態素解析

■ 情報科学の研究者達の努力により、1990年前後に実用化

京都大学　JUMAN
奈良先端大　ChaSen, MeCab

■ ただし言語研究に使うには解析用辞書に問題が

辞書の内部で「単語」の定義に一貫性がない

46

Yahoo!	IPA	JUMAN
国立国会図書館	国立	国立
	国会図書館	国会
		図書
		館
国立	国立	国立
公文書館	公文書	公文書
	館	館
国立	国立	国立
歴史	歴史	歴史
民俗	民俗	民俗
博物館	博物館	博物
		館

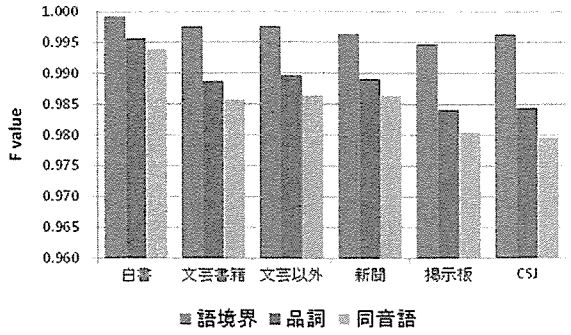
47

Yahoo!	IPA	JUMAN	UniDic
国立国会図書館	国立	国立	国立
	国会図書館	国会	国会
			図書
			館
国立			国立
公文書			公
	館	館	文書
			館
国立	国立	国立	国立
歴史	歴史	歴史	歴史
民俗	民俗	民俗	民俗
博物館	博物館	博物	博物
		館	館

UniDicはCSJの経験を踏まえて
BCCWJ用に開発した電子化辞書。
同じ粒度で「語」を切り出せる

48

UniDic + MeCabによる解析精度



49

二重形態素解析

短単位	短単位品詞	長単位	長単位品詞
公害	名詞-普通名詞-一般	公害紛争処理法	名詞-普通名詞-一般
紛争	名詞-普通名詞-サ変可能		
処理	名詞-普通名詞-サ変可能		
法	名詞-普通名詞-一般		
に	助詞-格助詞	における	助詞-格助詞
おけ	動詞-一般		
る	助動詞		
公害	名詞-普通名詞-一般	公害紛争処理	名詞-普通名詞-一般
紛争	名詞-普通名詞-サ変可能		
処理	名詞-普通名詞-サ変可能		
の	助詞-格助詞	の	助詞-格助詞
手続	名詞-普通名詞-サ変可能	手続	名詞-普通名詞-一般
は	助詞-係助詞	は	助詞-係助詞

短単位～辞書見出し語　長単位～複合語(複合辞)

50

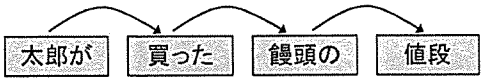
形態素解析の次のステップ

- 係り受け構造のアノテーション
  - 語と語の修飾関係
- 動詞項構造のアノテーション
  - 動詞と目的語などの関係
- 文末・節末のアノテーション
  - 文末が句読点などで明示されているとは限らない
- 意味のアノテーション
  - 多くの語には複数の用法がある
- 拡張モダリティのアノテーション
  - 書き手の意図や判断
- その他、時間表現、長単位内部構造、等々

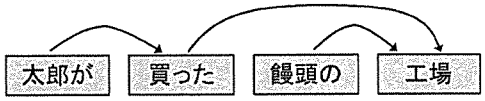
51

係り受け構造解析

- 太郎が買った饅頭の値段



- 太郎が買った饅頭の工場

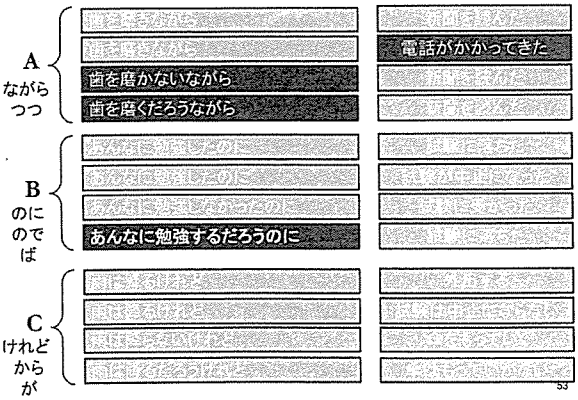


※音声合成にも有用。

52

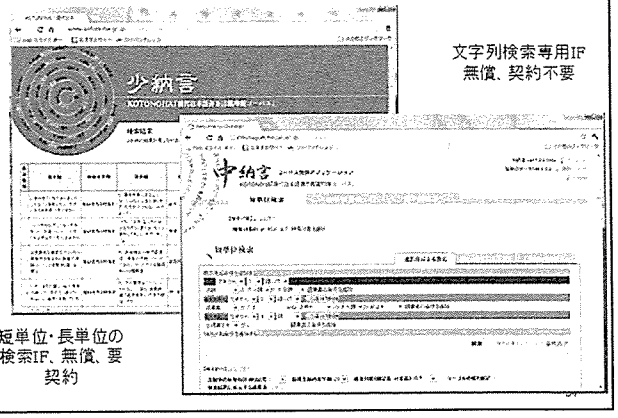
節(従属句)の特徴

Cf. 南不二男(1974)『現代日本語の構造』



53

BCCWJの一般公開



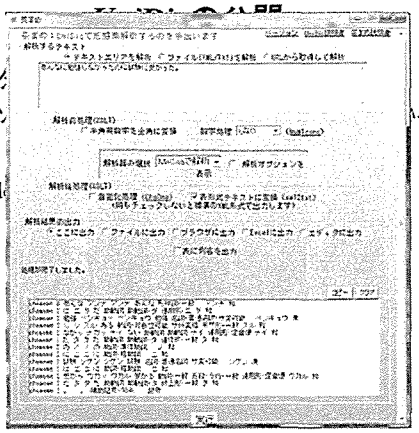
文字列検索専用IF
無償、契約不要

短単位・長単位の
検索IF、無償、要
契約

UniDicの公開

- データ全体を無償公開
  パソコンやスマートフォンのOSで既に利用されている。
- Windows PC用インタフェース「茶まめ」

55



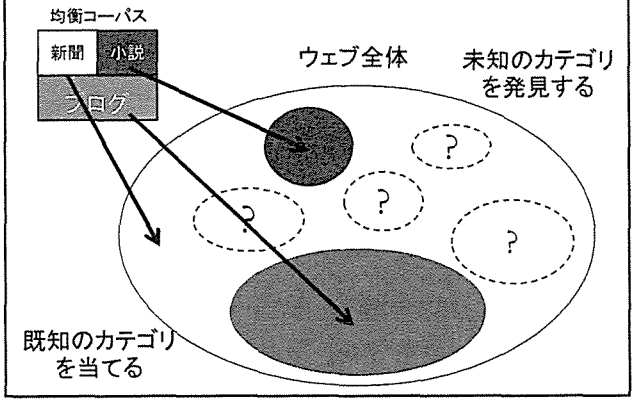
56

いまはどこに向かっているか

- 各種アノテーションの基礎研究
  多くの大学と共同研究
- 日本語歴史コーパスの準備研究
  平安から江戸時代までの代表的文学作品
  形態素解析も
- 超大規模コーパスの構築
  Webを母集団として100億語規模
  サンプルのレジスター推定

57

ウェブのレジスター推定問題



レジスターの推定実験

正解＼予測	掲示板	白書	ブログ	書籍	雑誌	新聞
掲示板	42	0	5	2	1	0
白書	0	49	0	0	0	1
ブログ	3	0	40	3	0	4
書籍	2	0	0	47	1	0
雑誌	0	1	3	12	27	7
新聞	0	0	0	3	5	42

BCCWJの6レジスターからランダムに選んだ300サンプルをSVM (support vector machines)という手法で分類。データは文長、品詞比率、語種比率

59

まとめ

コンピュータで日本語をきちんと研究できるようにするためには、
- 各種コーパスの構築
- 検索用情報(形態素、係り受け等)の付与
- 検索用インターフェースの開発
- 著作権処理

などの壁を超える必要があった。

これらの仕事を一応終えたいま、過去の日本語を対象とした歴史コーパスや、100億語規模の超大規模コーパスの開発を進めている。また遠くない時期に、日常会話を対象とした新しい話し言葉コーパスにも着手したいと考えている。

60