コーパスを利用した近代語研究
～太陽コーパスと近代文語UniDic

人間文化研究機構 国立国語研究所 言語資源研究系
小木曽 智信
togiso@ninjal.ac.jp

2009.12.2
於 東京外国語大学

コーパスと日本語研究

- コーパスとは
  - 実際に用いられた言語資料を、その言語の実態を正確に反映するように組織的かつ大量に収集してコンピューターで検索できるようにしたもの
- コーパス日本語学は発展途上
  - コーパス整備の遅れ
    - 英語Brown corpus：1960年代
    - 日本語の表記の多様性，分かち書きされない
- 国語研による日本語コーパス
  - 日本語話し言葉コーパス(CSJ)：現代語話し言葉(2004)
  - 太陽コーパス：確立期現代語≒近代語(2005)
  - 現代日本語書き言葉均衡コーパス(BCCWJ)：現代語書き言葉(2011完成予定)
  - 通時コーパス(計画中)

言語の歴史的研究とコーパス

- 歴史的研究：残された文献資料に頼らざるを得ない
  - ×内省
  - ×インフォーマントテスト
  - ×意識調査・アンケート
  - ○使用例に基づく分析

- 歴史的研究はコーパス言語学と親和性が高い
- もっといえば、歴史的研究は本質的にコーパス言語学にならざるを得ない

近現代語のコーパス

- コーパスとは
  - 実際に用いられた言語資料を、その言語の実態を正確に反映するように組織的かつ大量に収集してコンピューターで検索できるようにしたもの
- 全てを満たす十分な近代語コーパスはまだないが、利用可能なデータは増えている
  - 新潮文庫の百冊：明治の文豪、大正の文豪
  - 青空文庫
  - 国会会議録
  - 太陽コーパス
    - 雑誌だけが対象だがコーパスとして十分な質と量

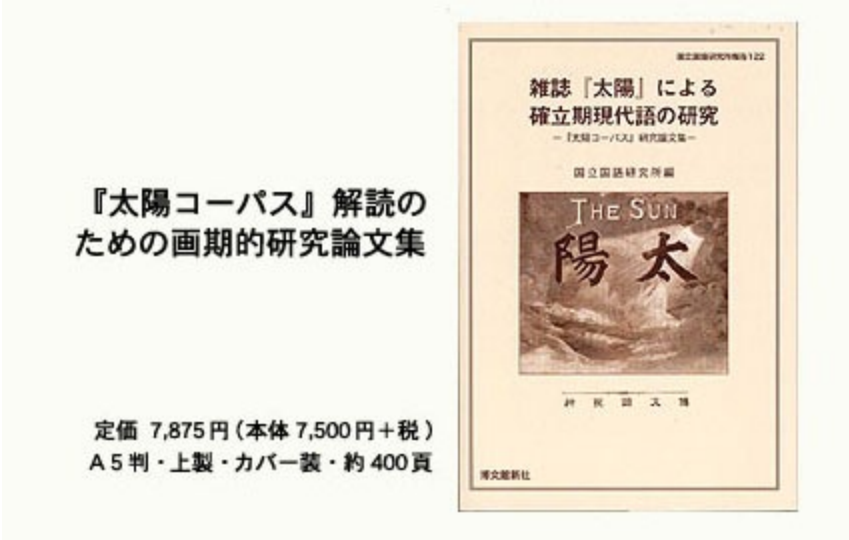
太陽コーパス

『太陽コーパス』

- 『太陽コーパス ―雑誌「太陽」日本語データベース―』
  - 国立国語研究所編（国立国語研究所資料集15）
  - 2005年3月31日／博文館新社／CD-ROM1枚＋解説書A5判横組み56ページ／税込み9,975円

「太陽コーパス」研究論文集

- 『雑誌「太陽」による確立期現代語の研究
  ―「太陽コーパス」研究論文集―』
  - 国立国語研究所編（国立国語研究所報告122）
  - 2005年3月31日／博文館新社／A5判横組み414ページ／税込み7,875円



「太陽コーパス」の収録資料

- 博文館刊
  総合雑誌『太陽』
- 収録範囲
- 創刊から8年おき5年分
  （1895, 1901, 1909, 1917, 1925年）
  の各年12冊、計60冊
- 特集号などがあるため各年の全号ではない
- 広告・英文・漢文などは未収録

- データの規模

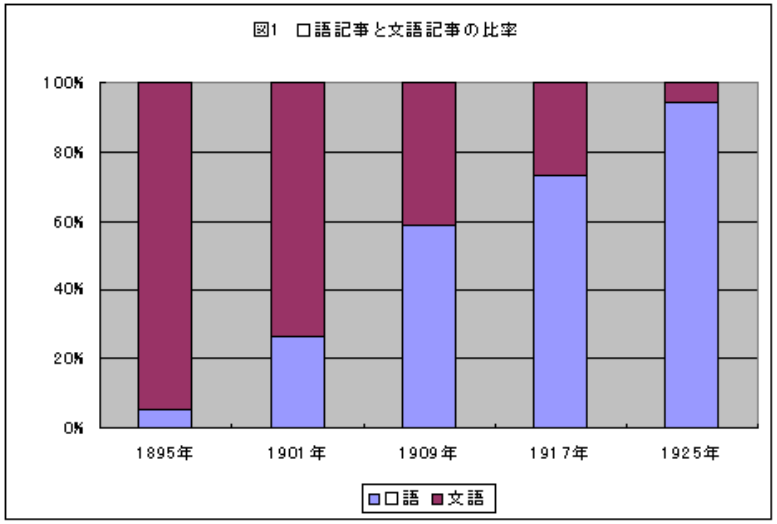
年	記事数	文字数
1895年	729	3335367
1901年	635	3154563
1909年	652	2860352
1917年	503	2647455
1925年	889	2453905
計	3408	14451642

※ タグを除いた本文の文字数
※ ちなみに源氏物語は約91万文字

『太陽』の資料としての特徴

- 当時もっとも良く読まれた総合雑誌
- 広汎な記事ジャンル
  - 政治・経済・世界情勢から科学・思想、文学作品まで
- 著名人を含む多彩な執筆陣

- 文語文から口語文への過渡期



『太陽』の本文

- 不統一な句読法
- 多様なふりがな（臨時的熟字訓）
- 数多い誤植（誤字・衍字・脱字・転倒など）
- 仮名遣いの不統一
- 濁点付与の不徹底
- 難解な漢字、異体字
- 非規範的語形

  →電子化に際して（できるだけ元の情報を残しつつ）検索に適した形に修正

太陽

明治卅四年 第七卷第壹號 一月五日發行 (禁轉載)

明治三十四年

太陽

人生には限りありて寸望には際涯なし、有限の生を以て、無限の望を懐く、其發して理想となる者、千萬年後の光景を豫測すべし、四圍暗黑の社會に、哲人が黃金世界を豫言するは之が爲めなり、春秋の時代は、臣其君を弑し子其父を弑する暗黑世界なりしも、孔子は別に道義の世界を開拓して、苟も道を踐み義を行はゝ、四海の內皆兄弟なりと道破せり、プラトの理想的社會を說けるは、希臘國民が大哲ソクラテスを刑殺せし昏昧時代に在り、トーマスモーアのユトピアを草せるは、英國の暴王が無罪の彼を刑せる時代に於てせり、哲人の永遠に涉る寸望は、時代の空氣と懸隔すると此の如し、此光明の暗中に閃くは、社會の漸次に進步する所以にして、數千歲後の今日、未だ哲人の寸望に達せずと雖、之を春秋の時代と希臘の當時とに、比するに、絕大に變化を生じ、ヘンリー八世の英國を以て、之をヴヰクトリアの今代に較ぶるに、人道の開發文運の進步、殆ど別社會の觀あり、此變化此進步は、遂次古哲人が冥想せし社會に近づくの段階なるを見る、如此にして止まざれば、終に黃金世界を現出せんと、決して之れ無きとせず、否な吾人は其必然を信せざるを得ざるなり、永年の寸望あり、百年の寸望あり、一年の寸望あり、瞬時は是れ永年の一部分なりとせば、一年の寸望は社會進步の動機として至大の關係あり、況や百年の寸望に於てをや、吾人明治三十四年の初に筆を操りて、未來の一年を臆想し、又千九百一年に際して、百年以後の世界を豫測するに、好望奉海を眺むるが如き感無くんばあらず、四圍の事情は不滿足の者多く、現實の社會は傷心の事多し、哲人が所謂缺陷の世界といへる者其異を得たりとす、萬物を支配する一貫の主力は、吾人の社會を化して醇より醇に赴かしむ、此間一點の疑を容る可からず、此信仰を懷きて社會を觀る者、誰か無限の快感を生せざらんや。

明治の世となつて以來、三十四回の年を迎へり、個人より之

『太陽』の電子化（１）

- 漢字の処理
  - JIS X0208（JIS第1・第2水準漢字）の包摂基準を拡張して適用
  - 外字は『今昔文字鏡』番号（諸橋大漢和と互換）で対処
  - それでも残る外字は画像を収録
- ふりがな
  - 文学作品中のルビのみ収録
  - 『太陽』の中には、総ルビ・パラルビ・ルビ無しの記事があるが、文学作品以外では削除されている
  - 傍点・傍線もすべて削除されている

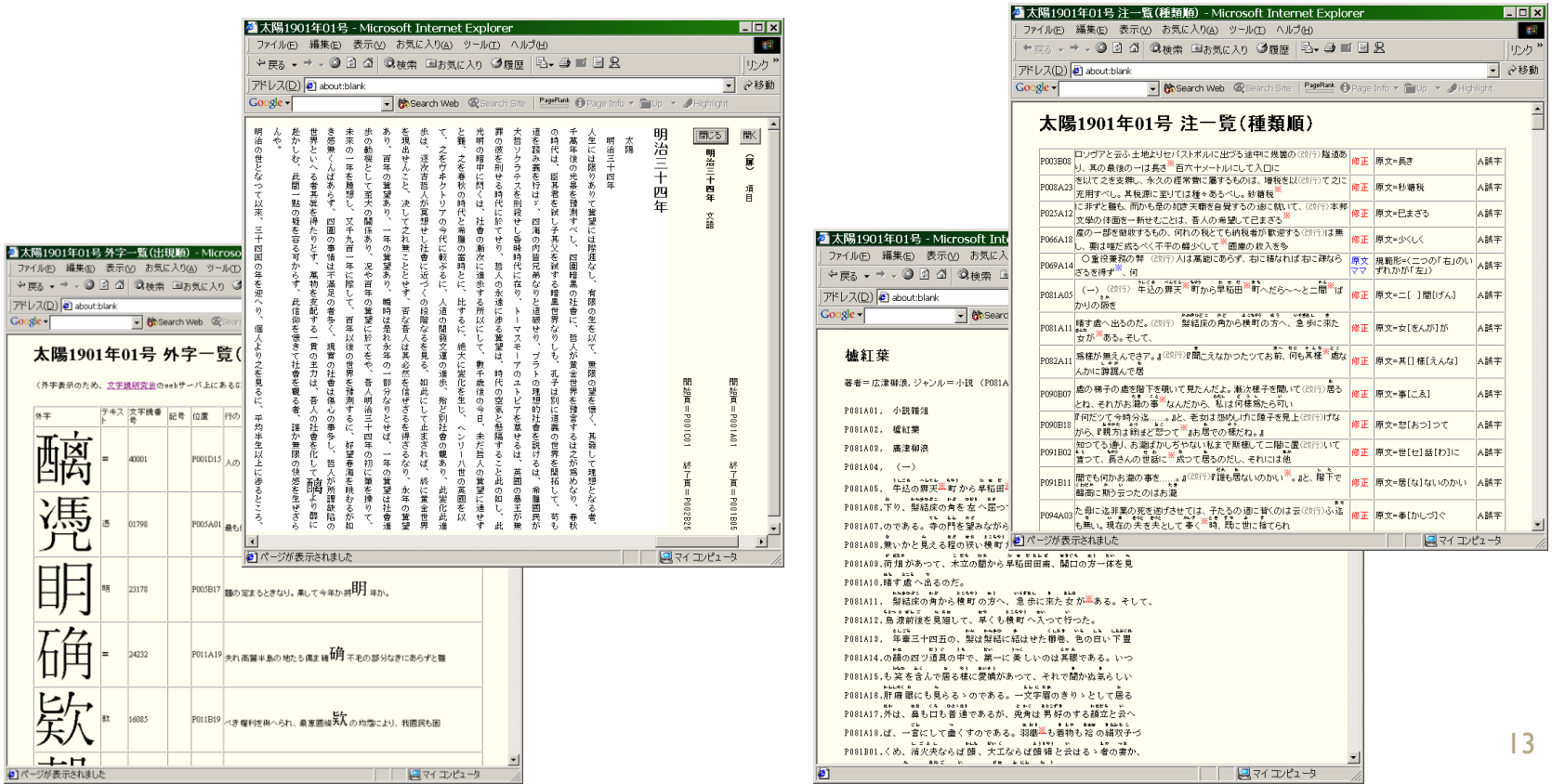
『太陽』の電子化（２）

- 注の付与と本文の修正
  - 検索しやすい標準的な形に本文を修正
  - 原文の情報はタグに記述

    例：戴冠<注 原文="紀念" 分類="A誤字通用">記念</注>祭
- XMLによる構造化
  - XMLによって文書を構造化
  - 著者・ジャンル、引用情報などの属性を記述

XMLの利用

- XML関連技術を用いることでさまざまな応用が可能。
  - 「太陽コーパス」はXMLを用いて人文科学系資料の電子化を行った試みとして、世界的に見ても早いもの。



太陽コーパスの利用

主な使い方と想定される利用者

決まった使い方があるわけではないが、次のような使い方が考えられる

1. 「ひまわり」で検索してその結果を使う
   - とにかく試してみたい人、初心者
2. 「プリズム」「たんぽぽ」でXMLを使う
   - 構造化文書（XML）の機能を試してみたい人
3. XMLからテキストファイルを生成して使う
   - 旧来のテキスト処理（grep, sort, awk等）に慣れている人
4. XMLをそのままXMLアプリケーションで使う
   - 簡単なプログラムが書ける人、上級者

ここでは１．の方法を説明する

全文検索システム「ひまわり」

1. いちばん上の空欄に検索語を入力
2. 旧字に変換したい場合は「字体変換」ボタンをクリック
3. 「検索」ボタンをクリック

↓

検索結果が表示される

※検索結果のダブルクリックで本文表示

デモ

「ひまわり」でできること

ただ調べたい文字列を入れて検索するだけでも次のような目的に簡単に利用できる

- 初出例の検索
- 用例数の把握

※表記に関わる問題、語形の揺れなどは

- コーパス本文に注が付いていないか
- 電子化の際の基準によって見落としや過剰なカウントが生じていないか

といった点に常に気を配る必要はある。

テーマ１
「拉致」という言葉の意味変化

- 特定の語の古い用例を検索する

デモ

テーマ１
「拉致」という言葉の意味変化

- 研究會の全部を政友會に拉致するを得ん
  - 研究会の全員を政友会に拉致することができるだろう
- 即ち大に門戸を開いて、眞個政治家の資質ある人士を拉致するの外に無いのである
  - 門戸を開いて政治家の資質のある人を**拉致する**
- 海内の學者を闕下に拉致したるを以て、濟々たる多士は翰林に充滿せり
  - 天下の学者を皇帝のもとに**拉致し**たので、多くの優秀な人物でいっぱいだった

- 当時は「引っ張ってくる」「連れてくる」「招き入れる」といった意味だった。少なくとも、現代語のような「無理矢理連れてくる」「誘拐する」というような意味ではない。

テーマ２
「婦人」と「女性」の用例数推移

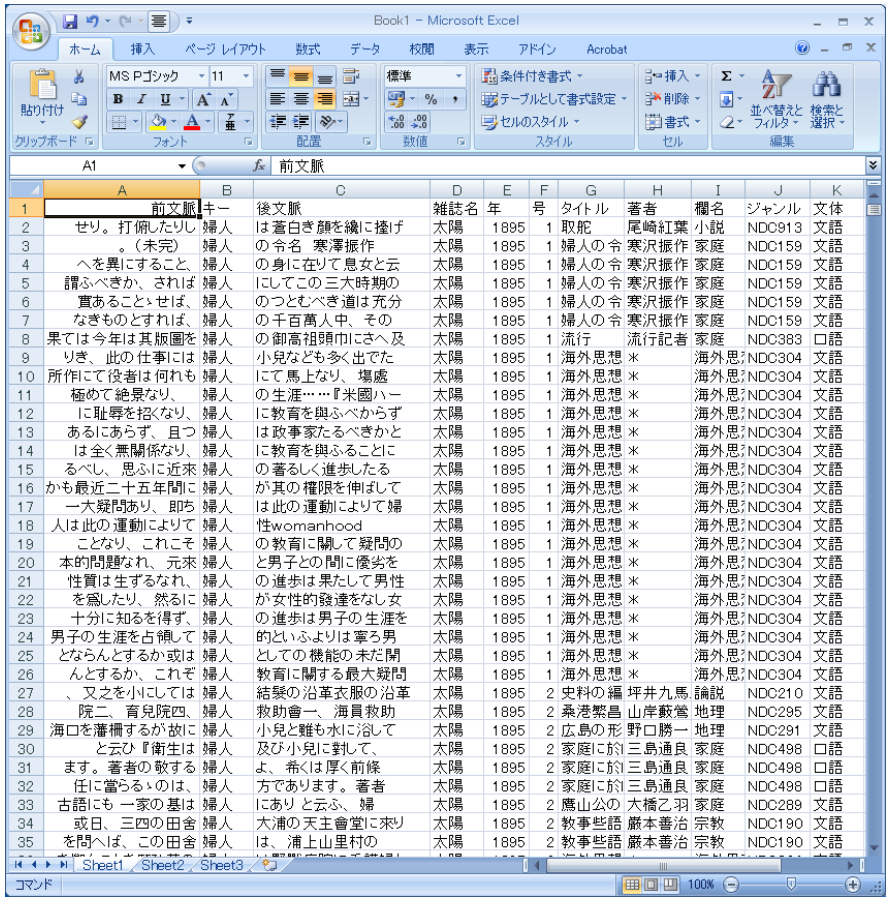
- 婦人
  - 現代語では複合語や雑誌名などに残るのみだが、かつては「成人女性」の意味で広く使われていた。現代語では「女性」に取って代わられた。

※既婚女性の意味ではない（≠夫人）

- 女性向けの雑誌名
  - 婦人画報　1905年創刊
  - 婦人公論　1916年創刊
  - 週刊女性　1957年創刊
  - 女性セブン　1963年創刊

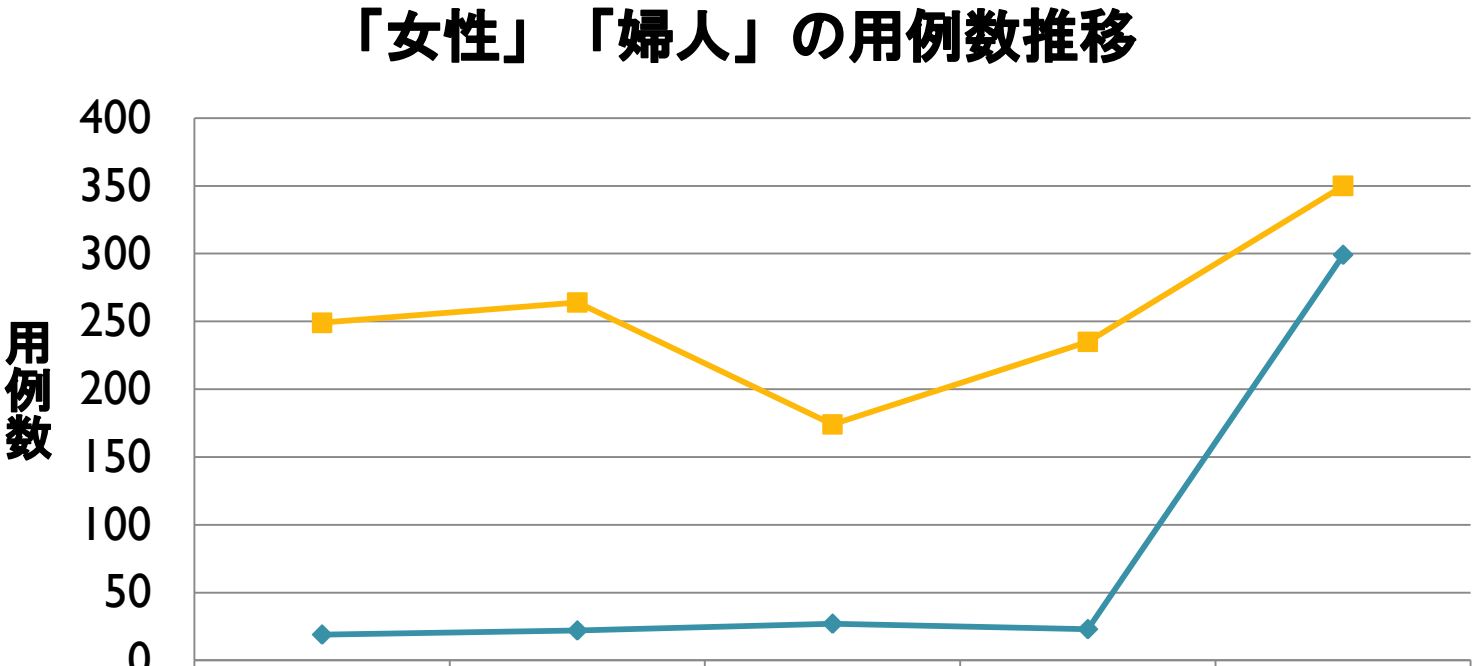
テーマ２
「婦人」と「女性」の用例数推移

- 「婦人」と「女性」の検索結果を集計



デモ

テーマ２
「婦人」と「女性」の用例数推移



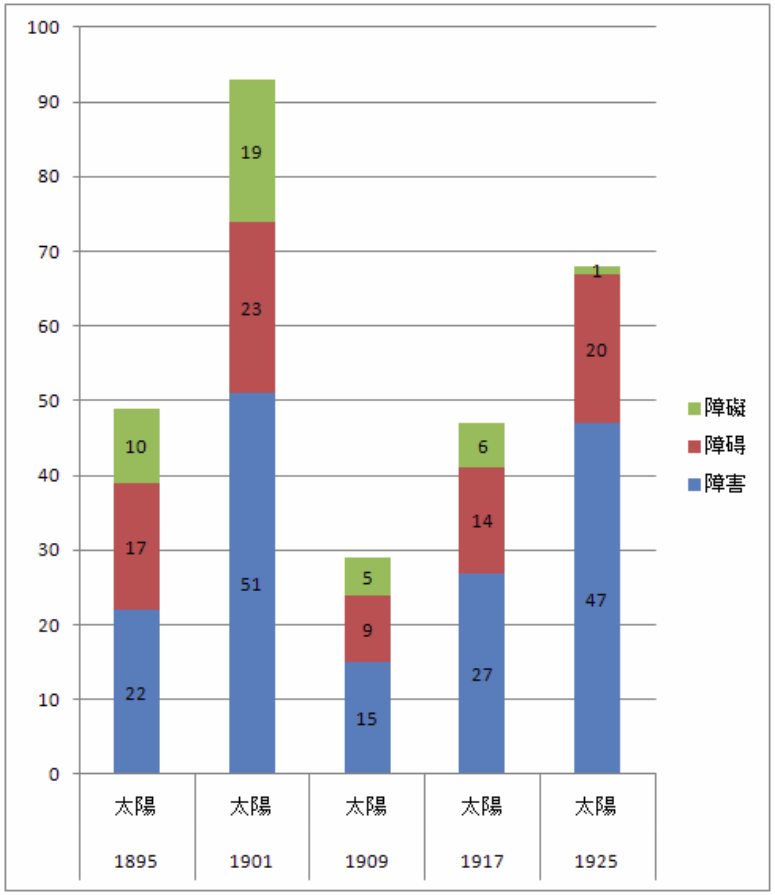
	1895	1901	1909	1917	1925
女性	19	22	27	23	299
婦人	249	264	174	235	350

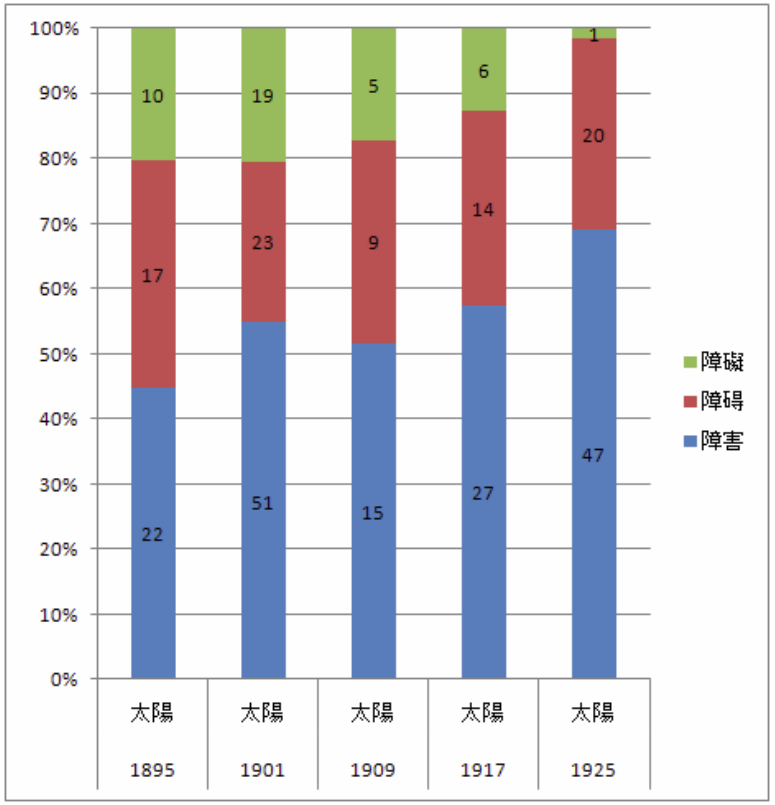
※「女性」は「じょせい」「にょしょう」を含む

テーマ３
「障がい」の表記

- “戦前は「障害」と書かず「障碍（障礙）」と書いていた”（当用漢字・常用漢字の施行後に変化した）という話は本当か？
- 太陽コーパス以外のデータでは、表記が改められている（元の表記がわからない）可能性があり、信用できない。

テーマ３
「障がい」の表記





明治のころから「障害」が多く使われている。

文字列検索の限界

- 特定の語の用例を探し、集計することはできるが…
  - そもそも総語数がわからない
  - 不特定の語は探せない
    - 形容動詞の連体形
    - 動詞の異なり語数
    - 五段活用動詞に助動詞「れる」がついたもの
    - Etc…

- →形態素解析が必要

「太陽コーパス」で

できること

- 文字列での検索
  - 正規表現を使ったパターンによる検索
- 文書構造を活かした検索
  - 年別、ジャンル別、著者別の用例カウント
  - 引用情報の利用
  - XSLTの活用
- 文字数の比較
  - 文字数、漢字含有率…

できないこと

- 語としての検索
  - 品詞別の検索や絞り込み
  - 前後の語と組み合わせたコロケーション検索
- 語彙の比較
  - 語種の割合
  - 延べ語数・異なり語数…

（形態素解析が必要）

近代語テキストの形態素解析

形態素解析とは

- コンピュータにより、文章を自動で単語に区切り、品詞や読みなどの情報を付与する自然言語処理の基礎技術。
- インターネットの検索サイトをはじめ、さまざまな分野で実際に用いられている。
- 「現代日本語書き言葉均衡コーパス」でも単語情報の付与に利用。

解析結果の例

出現形	読み	代表形	代表表記	品詞	活用型	活用形
私	ワタシ	ワタシ	私	代名詞		
は	ワ	ハ	は	助詞-係助詞		
国立	コクリツ	コクリツ	国立	名詞-普通名詞-一般		
国語	コクゴ	コクゴ	国語	名詞-普通名詞-一般		
研究	ケンキュー	ケンキュウ	研究	名詞-普通名詞-サ変可能		
所	ショ	ショ	所	接尾辞-名詞的-一般		
に	ニ	ニ	に	助詞-格助詞		
勤め	ツトメ	ツトメル	勤める	動詞-一般	下一段-マ行	連用形-一般
て	テ	テ	て	助詞-接続助詞		
い	イ	イル	居る	動詞-非自立可能	上一段-ア行	連用形-一般
ます	マス	マス	ます	助動詞	助動詞-マス	終止形-一般
。			。	補助記号-句点		

形態素解析の仕組み
（１）解析辞書の作成

1. 単語の一覧（語彙表）を用意する
2. 正しく単語に区切り、情報を付けた大量の文章データ（学習用コーパス）を用意する
3. 学習用コーパスを元に、機械学習の方法で単語ごとの頻度（生起コスト・連接コスト）を計算

↓

解析用辞書

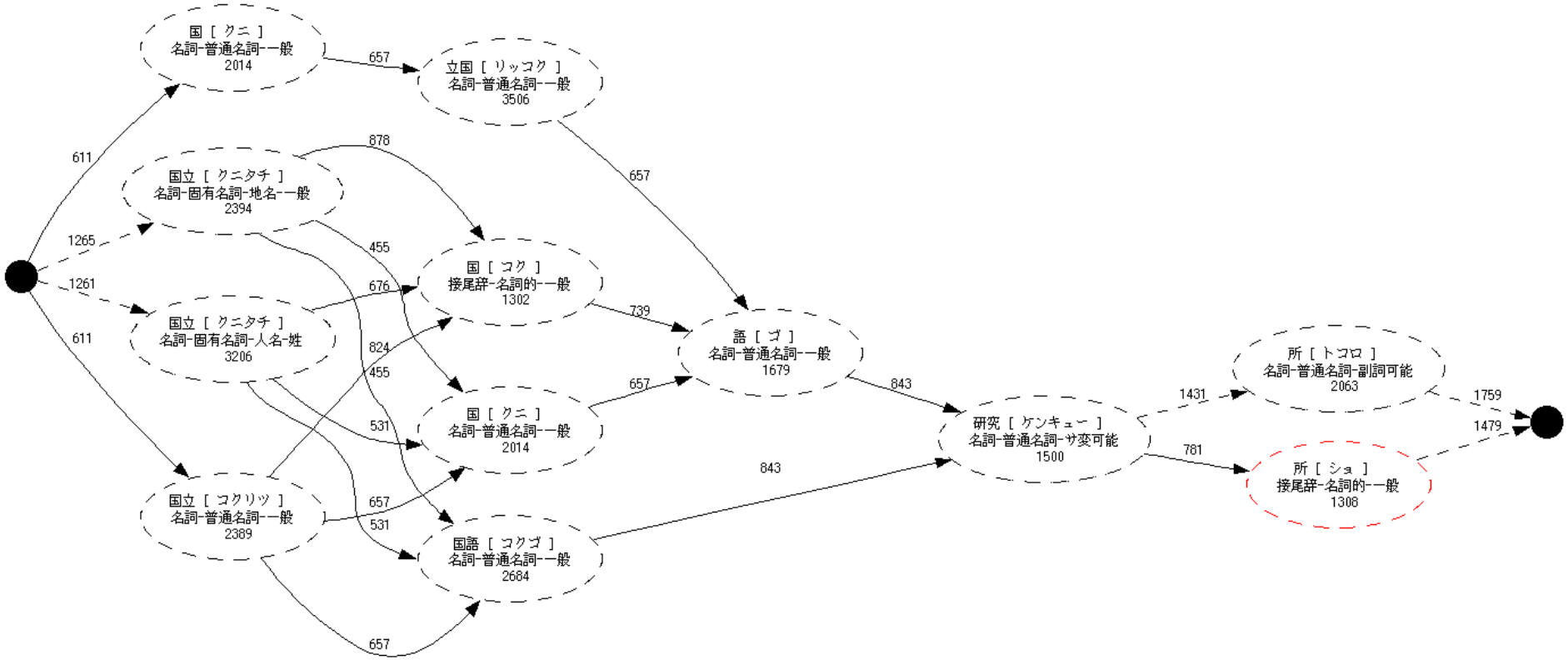
形態素解析の仕組み（２）解析器の処理

1. 入力文を辞書にある単語の可能な組み合わせに切っていく。
2. その組み合わせの中で、もっともコスト（出現コスト・連接コスト）の合計が低いものを最適の解として出力する。

- フリーの解析器としてChaSenとMeCabが著名。
- UniDicはChaSenとMeCabに対応。

形態素解析の仕組み（３）解析の例

- ○の中の数字＝生起コスト（その語がどれだけ現れやすいか）
- 枝の数字＝連接コスト（どのような品詞とつながりやすいか）



近代文語UniDic

現代語の形態素解析辞書UniDicをもとに、太陽コーパスの論説記事のような近代文語文（明治普通文）を解析できるようにした形態素解析辞書

近代文語UniDicの利用

- 「茶まめ」を使った解析

デモ

解析精度

表の数字は近代文語UniDic 0.97 MeCab版にもとづく

		福澤諭吉	山路愛山	太陽	民法	全体
単位境界	テストデータ語数	4192	3058	6184	21262	34696
	解析結果語数	4193	3057	6191	21269	34710
	正解	4184	3032	6144	21228	34588
	再現率	0.998092	0.991498	0.993532	0.998401	0.996887
	適合率	0.997854	0.991822	0.992408	0.998072	0.996485
	F値	**0.997973**	**0.99166**	**0.99297**	**0.998237**	**0.996686**
品詞認定	テストデータ語数	4192	3058	6184	21262	34696
	解析結果語数	4193	3057	6191	21269	34710
	正解	4097	2981	6041	21080	34199
	再現率	0.977338	0.97482	0.976876	0.99144	0.985676
	適合率	0.977105	0.975139	0.975771	0.991114	0.985278
	F値	**0.977221**	**0.97498**	**0.976323**	**0.991277**	**0.985477**
語彙素認定	テストデータ語数	4192	3058	6184	21262	34696
	解析結果語数	4193	3057	6191	21269	34710
	正解	4071	2973	6003	21064	34111
	再現率	0.971135	0.972204	0.970731	0.990688	0.983139
	適合率	0.970904	0.972522	0.969633	0.990362	0.982743
	F値	**0.97102**	**0.972363**	**0.970182**	**0.990525**	**0.982941**

語彙素認定（境界・品詞・見出し語認定全て正解）で

約97～98%

太陽コーパス＋形態素解析

形態素解析をすることでできること

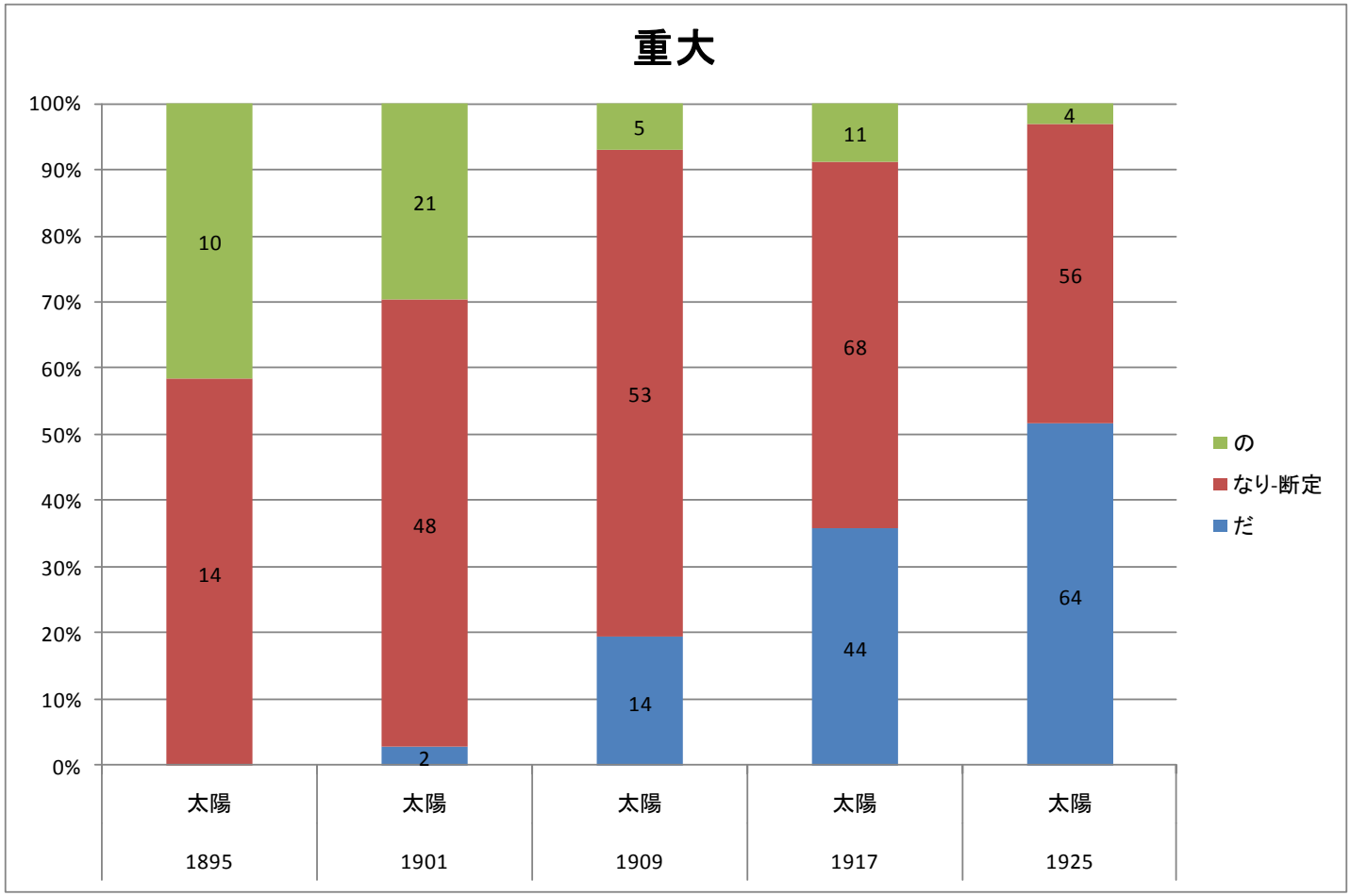
- 形態素解析結果のデータベース

＝文字通りの「語彙」

- 見出し語別、品詞別の延べ語数・事内語数集計
- 前後の語と組み合わせた検索・集計、コロケーション
- Etc…

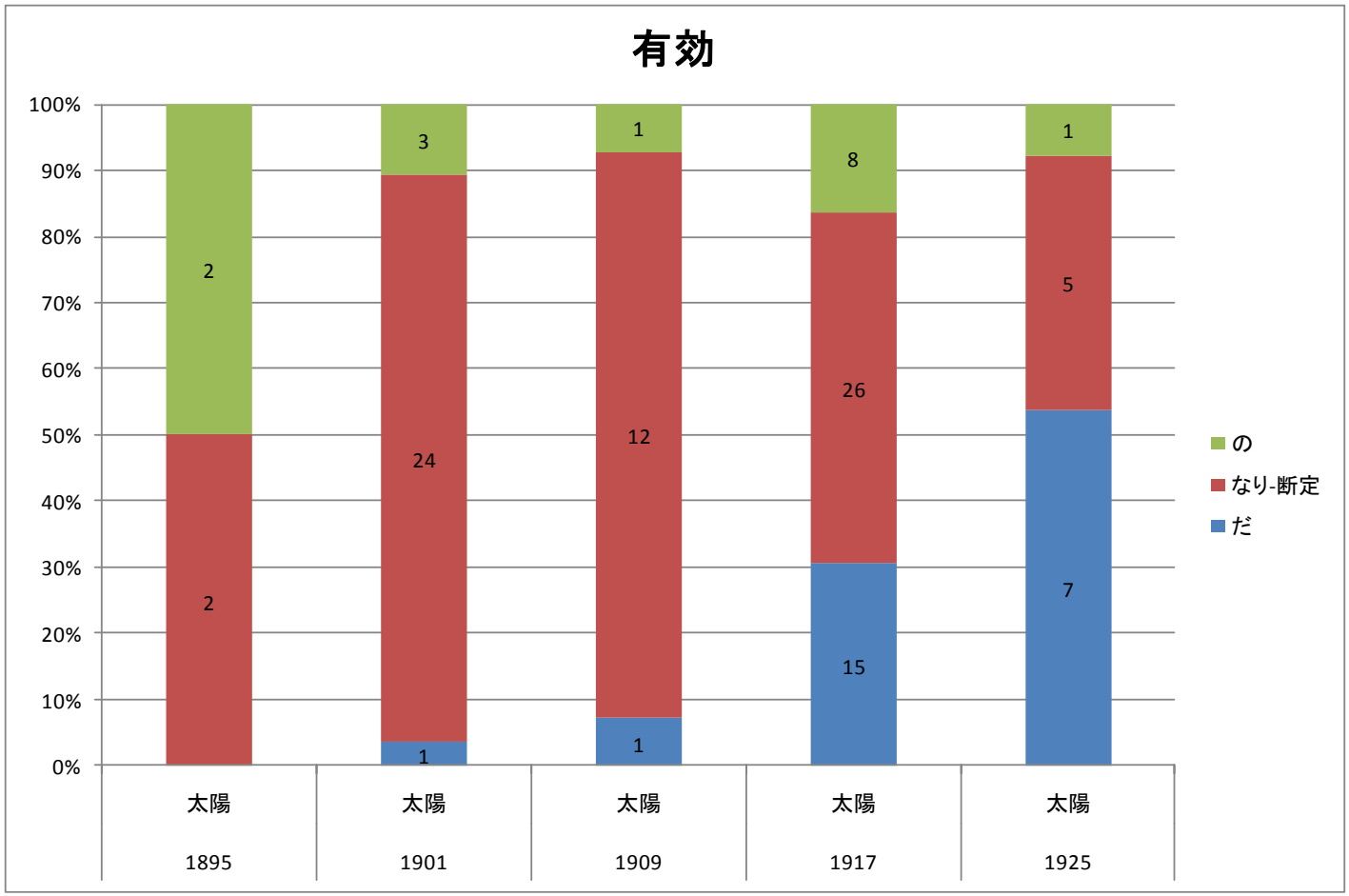
例：形容動詞連体形の推移

- 「重大」の連体の形を集計



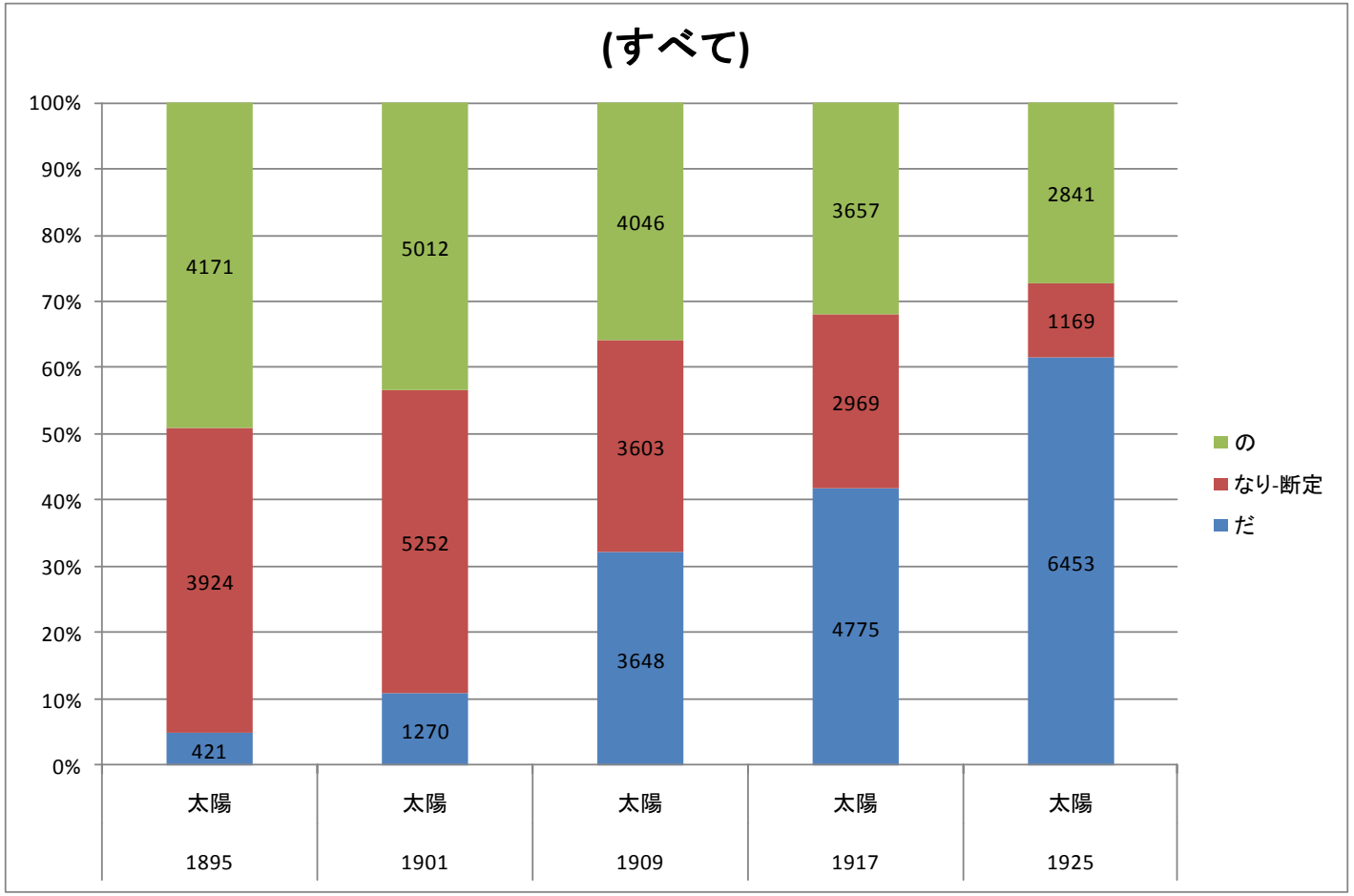
例：形容動詞連体形の推移

- 「有効」の連体の形を集計



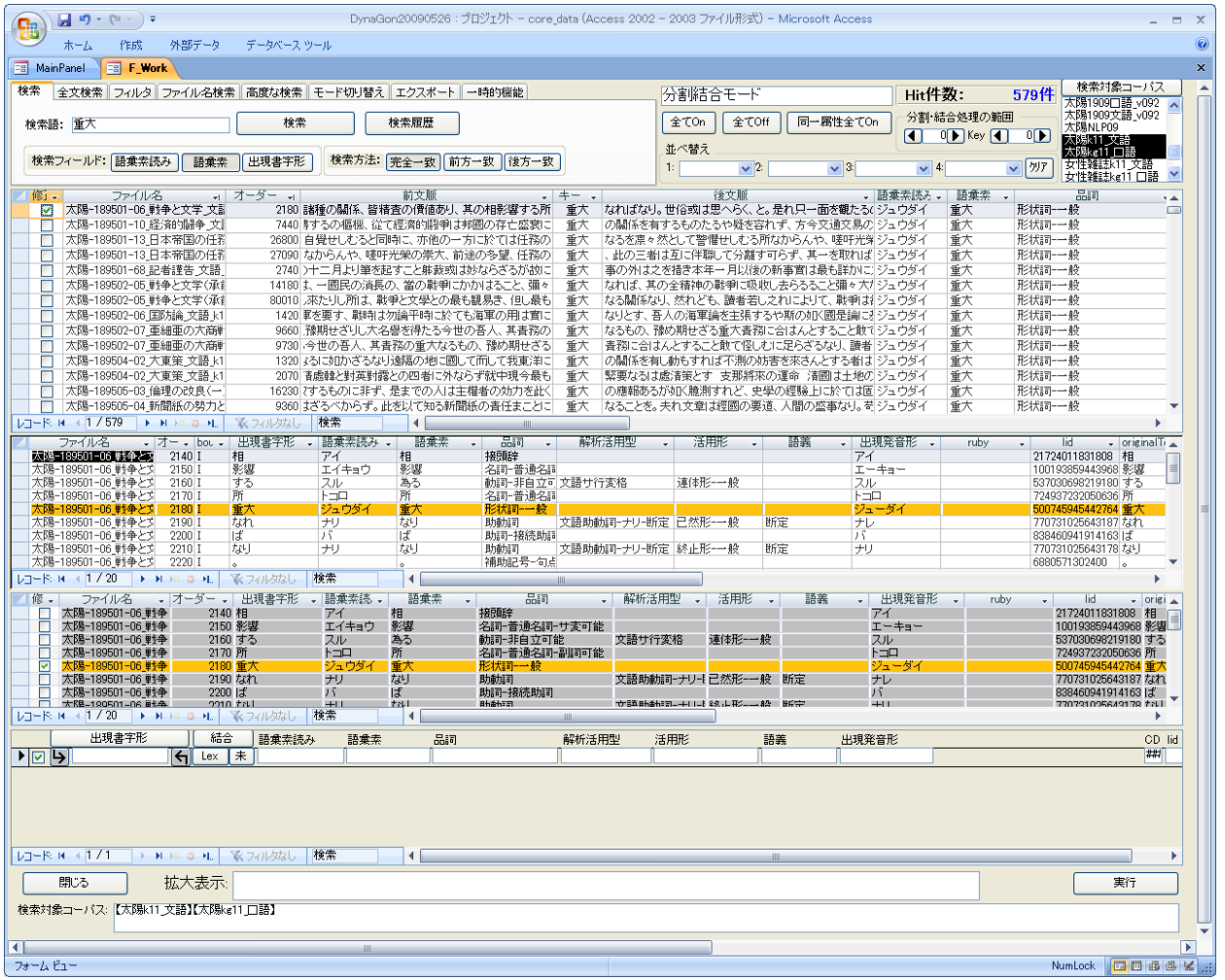
例：形容動詞連体形の推移

- すべての形容動詞の連体の形を集計



コーパス検索ツール「大納言」

- 形態素解析済みデータベースの検索ツール



参考文献

- 太陽コーパス
  - 『太陽コーパス』を利用した近代語法の研究 —形容動詞の連体形を例に—（近代語学会（2007年6月30日・於昭和女子大学） 2007 口頭発表）
  - 『雑誌『太陽』による確立期現代語の研究—『太陽コーパス』研究論文集—』国立国語研究所報告122（共著）(2005)　出版社：博文館新社
  - 『太陽コーパス 雑誌『太陽』日本語データベース』国立国語研究所資料集15（共著）(2005) 出版社：博文館新社
- 利用例
  - 明治・大正期における形容動詞の連体修飾の形　『ことばのダイナミズム』 (小木曽智信) 333-352 2008 (その他)
  - コラム 現代日本語の確立過程を調べる　国語研の窓 (小木曽智信) /35号, 2008 (その他)
  - 「構造化テキストを直接利用するアプリケーション」『雑誌『太陽』による確立期現代語の研究—『太陽コーパス』研究論文集—』博文館新社 (小木曽智信) pp.83-113 2005 (その他)
  - 総合雑誌『太陽』本文の様態と電子テキスト化　日本語科学 (田中牧郎・小木曽智信) /8号, pp.141-152 2000 (学術雑誌)
- 形態素解析
  - 『近代文語文を対象とした形態素解析のための電子化辞書の作成とその活用』(2009)　出版社：国立国語研究所・科研費報告書19720110
  - 形態素解析を用いた近代文語と現代語の語彙の比較（近藤明日子・小木曽智信）（日本語学会（2009年5月31日・於武庫川女子大学）（予稿集p.200） 2009 ポスター）
  - 代文語文を対象とした形態素解析辞書・近代文語UniDic（小木曽智信・小椋秀樹・近藤明日子）（日本語学会（2008年5月18日・於日本大学） 2008 その他）
  - 日本語研究に適した形態素解析ソフトウェア—「unidic」と「茶まめ」—（小木曽智信・小椋秀樹・伝康晴）（日本語学会（2007年11月17日・於沖縄国際大学） 2007 その他）