近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
（国立国語研究所）

2008.5.18 日本語学会春季大会（於 日本大学）

近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（国立国語研究所）
小椋秀樹（国立国語研究所）
近藤明日子（国立国語研究所）

2008.5.18 日本語学会春季大会
於 日本大学

1. はじめに

形態素解析とは

- コンピュータを使って、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。
  （コンピュータに品詞分解をさせる）
- 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用化されている。
- 日本語の形態素解析ソフトとしては、奈良先端科学技術大学院大学松本研究室によるフリーウェア「茶筌」が広く用いられている。

1.1. 形態素解析と文語文

- これまでは現代語を対象とした形態素解析辞書しか存在しなかった。
- 文語文を対象とした場合には、十分な精度が出ない。

- 次の例文を解析してみると・・・
  - こゝに漢字の利害と題するは、即ち聊か袈裟の眞價を問はんとするなり。

（『太陽コーパス』「漢字の利害」より）

従来の解析辞書による解析結果例①

（茶筌2.4.2とIPADIC2.7.0の組み合わせ）

IPADIC 2.7.0/ChaSen 2.4.2				
出現形	読み	品詞	活用型	活用形
こ	コ	名詞-一般		
↘	↘	記号-一般		
に	ニ	助詞-格助詞-一般		
漢字	カンジ	名詞-一般		
の	ノ	助詞-連体化		
利害	リガイ	名詞-一般		
と	ト	助詞-並立助詞		
題	ダイ	名詞-一般		
する	スル	動詞-自立	サ変・スル	基本形
は	ハ	助詞-係助詞		
、	、	記号-読点		
即ち	スナワチ	副詞-一般		
聊か	イササカ	副詞-一般		
袈裟	ケサ	名詞-一般		
の	ノ	助詞-連体化		
眞	マコト	名詞-固有名詞-人名-名		
價		未知語		
を	ヲ	助詞-格助詞-一般		
問	トイ	名詞-一般		
はん	ハン	名詞-接尾-人名		
と	ト	助詞-格助詞-一般		
する	スル	動詞-自立	サ変・スル	基本形
なり	ナリ	名詞-一般		
。	。	記号-句点		

従来の解析辞書による解析結果例②

（茶筌2.4.2とUniDic1.3.5の組み合わせ）

UniDic 1.3.5 /ChaSen 2.4.2					
出現形	代表形	代表表記	品詞	活用型	活用形
こ	コ	小	接頭辞		
ゝ		ゝ	補助記号-一般		
に	ニ	に	助詞-格助詞		
漢字	カンジ	漢字	名詞-普通名詞-一般		
の	ノ	の	助詞-格助詞		
利害	リガイ	利害	名詞-普通名詞-一般		
と	ト	と	助詞-格助詞		
題	ダイ	題	名詞-普通名詞-一般		
する	スル	磨る	動詞-一般	五段-ラ行-一般	連体形-一般
は	ワ	は	助詞-係助詞		
、		、	補助記号-読点		
即ち	スナワチ	即ち	接続詞		
聊か	イササカ	些か	副詞		
袈裟	ゲサ	袈裟	名詞-普通名詞-一般		
の	ノ	の	助詞-格助詞		
眞	マコト	マコト	名詞-固有名詞-人名-名		
價	價		未知語		
を	オ	を	助詞-格助詞		
問は	トワ	問う	動詞-一般	文語四段-ハ行+ふ	未然形-一般
ん	ン	ず	助動詞	助動詞-ヌ	助動詞-ヌ
と	ト	と	助詞-格助詞		
する	スル	為る	動詞-非自立可能	サ行変格	連体形-一般
なり	ナリ	成る	動詞-非自立可能	五段-ラ行-一般	連用形-一般
。		。	補助記号-句点		

1.1. 形態素解析と文語文

- 形態素解析の仕組み自体は、データさえ用意すれば文語にも対応可能。
- 文語文の形態素解析が行えれば、品詞を考慮した検索や、テキストの語彙比較、通時的な研究が可能になる。

まず、近代の文語論説文を対象に、
文語文を対象とした形態素解析辞書を作る。

近代文語UniDicによる解析結果例

（茶筌2.4.2と近代文語UniDic0.7の組み合わせ）

近代文語UniDic 0.7 /ChaSen 2.4.2							
出現形	発音形	代表形	代表表記	品詞	活用型	活用形	語種
こゝ	ココ	ココ	此処	代名詞			和
に	ニ	ニ	に	助詞-格助詞			和
漢字	カンジ	カンジ	漢字	名詞-普通名詞-一般			漢
の	ノ	ノ	の	助詞-格助詞			和
利害	リガイ	リガイ	利害	名詞-普通名詞-一般			漢
と	ト	ト	と	助詞-格助詞			和
題する	ダイスル	ダイスル	題する	動詞-一般	文語サ行変格	連体形-一般	混
は	ワ	ハ	は	助詞-係助詞			和
、			、	補助記号-読点			記号
即ち	スナワチ	スナワチ	即ち	接続詞			和
聊か	イササカ	イササカ	些か	副詞			和
袈裟	ケサ	ケサ	袈裟	名詞-普通名詞-一般			外
の	ノ	ノ	の	助詞-格助詞			和
眞價	シンカ	シンカ	真価	名詞-普通名詞-一般			漢
を	オ	ヲ	を	助詞-格助詞			和
問は	トワ	トウ	問う	動詞-一般	文語四段-ハ行	未然形-一般	和
ん	ン	ム	む	助動詞	文語助動詞-ム	連体形-撥音便	和
と	ト	ト	と	助詞-格助詞			和
する	スル	スル	為る	動詞-一般	文語サ行変格	連体形-一般	和
なり	ナリ	ナリ	なり-断定	助動詞	文語助動詞-ナリ-断定	終止形-一般	和
。			。	補助記号-句点			記号

1.2. なぜ近代文語(論説)文か

- 残された資料が多い。
  - 応用の幅が広がる。
  - 論説文は比較的均質的。
- 著作権の問題が少なく、電子化・公開されている資料が多い。
  - 青空文庫・太陽コーパスなどが利用可能。
- 現代語との比較がしやすい。
  - 現代語に直接つながる時代。
  - 現代語UniDicと同じ単位にそろえてあるので、解析結果を比較可能。
- いきなり全時代に対応した辞書は作れない。

2. 形態素解析辞書の作成

解析辞書作りに必要なもの

学習用コーパス

正しく品詞づけされた文章データ

生起コスト
連接コスト

機械学習処理を行うプログラム

学習器

辞書データ

解析に使う語のリスト

活用表

活用形展開語のリスト

形態素解析辞書

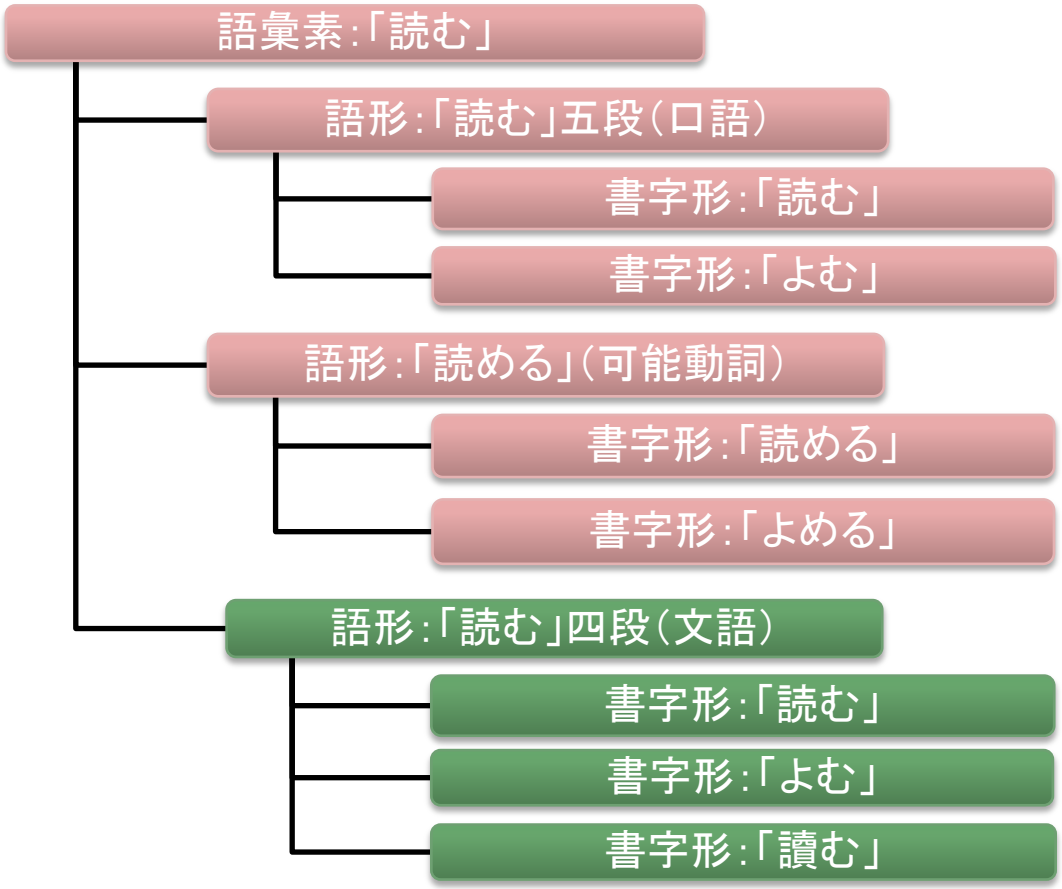
形態素解析器が解析時に使うデータ

2.1. 辞書データの整備

- UniDicの特長を活かして近代語の見出し語を整備
  - **階層化された見出し**
    →口語・文語を統一的に扱える
  - **斉一な単位（短単位）**
    →現代語と近代語の語彙比較が可能
  - **音声研究に利用可能**
    →△（音声情報などは現代語での読み）

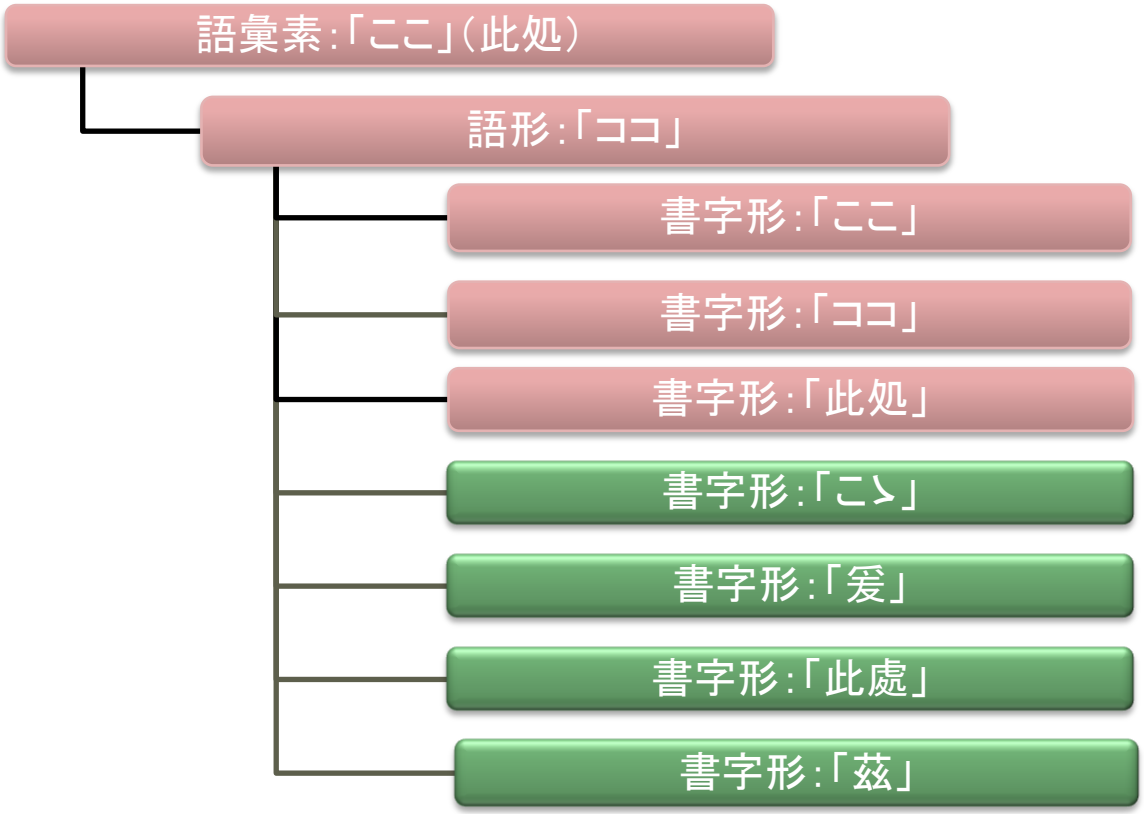
UniDicの階層と近代語用の見出し語①

- 文語形を語形レベルで追加して口語形と統一的に扱う



UniDicの階層と近代語用の見出し語②

- 現代語では用いられない表記を書字形レベルで追加して統一的に扱う



見出し語の追加

- 自動生成と手作業による修正
  - 旧字形　12,000語※
  - 文語形　18,000語
- 用例からの追加
  - 『太陽』スカウト式用例採集データ
  - 学習用コーパスの未知語
  - 計 7,000語
- 計３.７万語を近代語用として追加
  （現代語用の約１５万語+3.7万語＝18.7万語に）

※書字形レベル。以下同じ。

2.2. 活用表の整備

- 形態素解析辞書の活用表は、基本形（終止形）から各活用形を生成するためのもの。
- もともとUniDicは文語の活用表を持っていたが、不足する部分を追加。
- 次のような表記・語法上の問題に対応。
  - 現代仮名遣いの文語形
  - 濁点無表記の活用形
  - 送り仮名省略
  - ク語法

2.3.各種表記と辞書の対応

- 辞書で対処しきれない部分は解析前処理で対応
  - 漢字カタカナ交じり文
    - カタカナ→ひらがな変換の前処理で対応
  - 踊り字
    - 語中の踊り字については辞書で対応
    - 語の境界をまたぐ踊り字は前処理で対応
- 前処理はGUI「茶まめ」に実装（後述）

2.4.学習用コーパスの整備

- 「青空文庫」などで公開されているテキストデータと「太陽コーパス」から選定。
  - 総語数：約175,000語

- 整備に要する時間
  - 専用に開発したアプリケーションを利用、熟練した大学院生アルバイトが作業して、１日（7時間）あたり2000～3000語程度（未知語の辞書登録を含む）

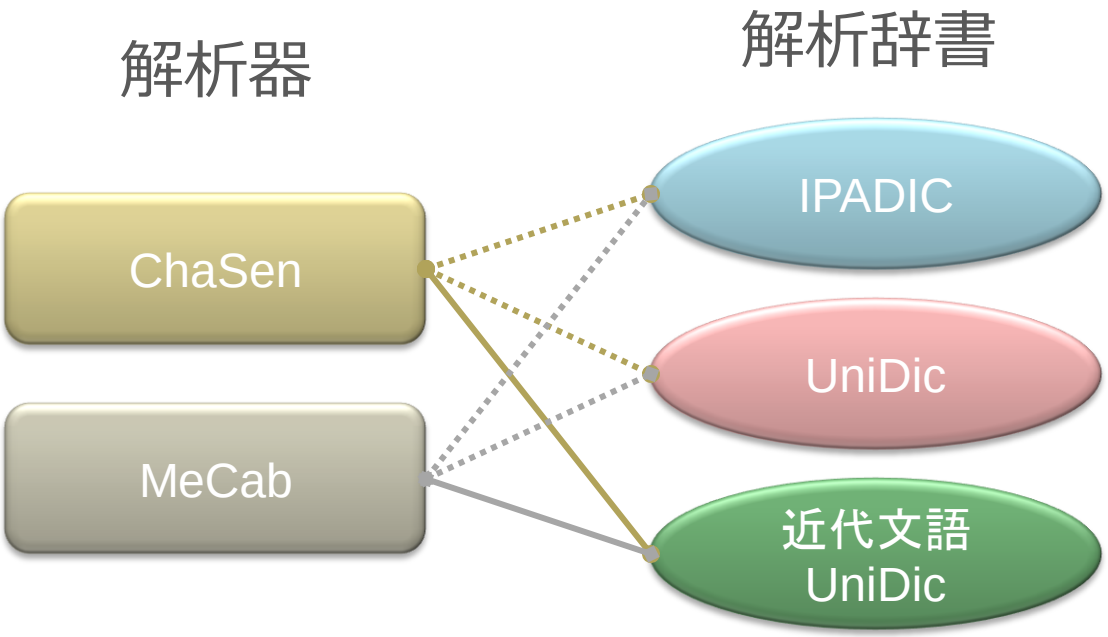
2.4.学習用コーパスの整備

青空文庫	綱島梁川	「国民性と文学」	新字旧かな
	高山樗牛	「一葉女史の「たけくらべ」を讀みて」	旧字旧かな
	山路愛山	「信仰個条なかるべからず」「唯心的、凡神的傾向に就て（承前）」	新字旧かな
	田中正造	「公益に有害の鉱業を停止せざるの儀に付質問書」	旧字旧かな
	内村鑑三	「ネルソン伝に序す」「時事雑評二三」「問答二三」	新字旧かな
	二葉亭四迷	「小説総論」	新字新かな
	福沢諭吉	「教育の目的」「新女大学」「中津留別の書」	新字新かな
	北村透谷	「各人心宮内の秘宮」「頑執妄排の弊」「実行的道徳」「人生に相渉るとは何の謂ぞ」「人生の意義」「熱意」	新字旧かな
文明論之概略	福沢諭吉	文明論之概略　緒言、巻之一第一章～巻之二第四章	新字旧かな
法律・公文書		「教育勅語」「軍人勅諭」「終戦の詔勅」「大日本帝国憲法」	旧字旧かな
		「皇室典範」「褒章条例」「民法」第一編・第二編	新字旧かな 無濁点
近代詩		『藤村詩集』序（島崎藤村）・そぞろごと（与謝野晶子）・荒城の月（土井晩翠）・初恋（島崎藤村）・千曲川旅情の歌（島崎藤村）・落葉（上田敏訳）・椰子の実（島崎藤村）	旧字旧かな （一部新字）
太陽コーパス		1901年1号（記事番号01～06, 08～14, 40, 50）	旧字旧かな

3.解析辞書と解析用GUI

解析器と解析辞書の組み合わせ

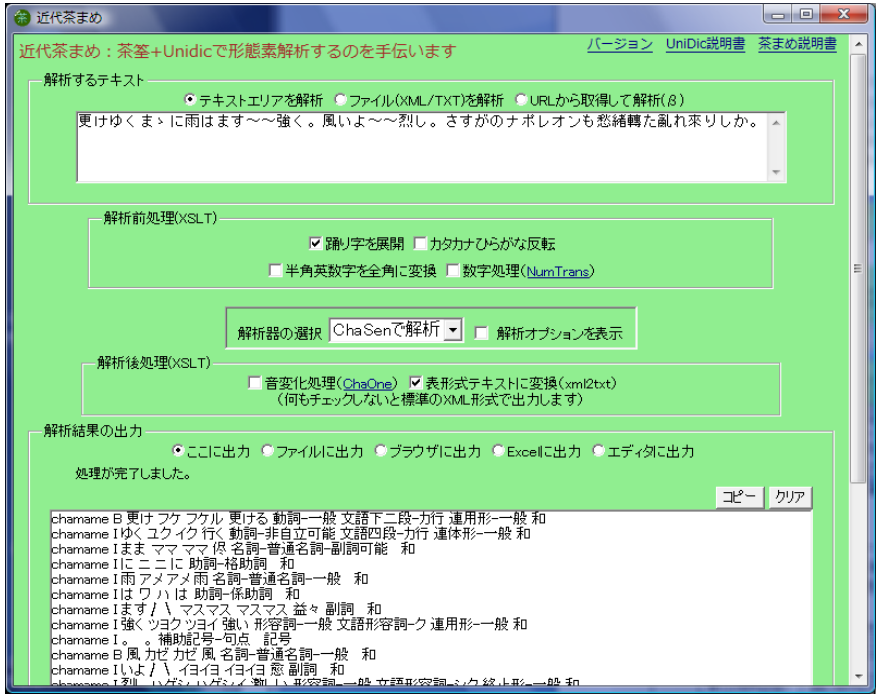
- 解析器（解析処理プログラム）と解析辞書はそれぞれ独立。



- 近代文語UniDicもChaSen版とMeCab版を用意。

解析用ツール「茶まめ」と解析前処理

- 初心者でも簡単に形態素解析ができるインターフェイス「茶まめ」を近代語用に拡張。
- クリック一つで近代語用の解析前処理ができる。
  - カタカナ→ひらがな変換
  - 踊り字変換
- 解析器の切り替えが可能。



4.解析精度

精度評価

- 人手で修正したデータ（学習用には利用しない）を使って、解析辞書の精度を測る。
- 「未知語なし」（テキストに出現するすべての語を解析辞書に登録した状態）で計測。
- 評価用のデータ（約3.5万語）

福澤諭吉	「経世の学、また講究すべし」「物理学の要用」
山路愛山	「北村透谷君」「透谷全集を読む」
太陽	「明治三十四年の経済界」「昨年の経済問題」「経済時評」（いずれも1901年1号）
民法	第三編

ChaSen版の解析精度

		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	4202	3074	6196	21334	34806
	正解	4170	3022	6117	21148	34457
	再現率	0.994751	0.988227	0.989165	0.994638	0.993111
	適合率				0.991281	0.989972
	F値				**0.992956**	**0.991538**
品詞認定	テスト				21262	34696
	解析結				21334	34806
	正解				20883	33901
	再現率				0.982174	0.977086
	適合率				0.97886	0.973998
	F値	**0.973552**	**0.963469**	**0.965751**	**0.980513**	**0.975539**
語彙素認定	テストデータ語数	4192	3058	6184	21262	34696
	解析結果語数	4202	3074	6196	21334	34806
	正解	4060	2943	5935	20864	33801
	再現率	0.968511	0.962393	0.959734	0.981281	0.974204
	適合率	0.966206	0.957384	0.957876	0.977969	0.971125
	F値	**0.967356**	**0.959881**	**0.958803**	**0.979621**	**0.972661**

語彙素認定で

約96～97%

MeCab版の解析精度

		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				.998072	0.996485
	F値				**.998237**	**0.996686**
品詞認定	テスト				21262	34696
	解析結				21269	34710
	正解				21080	34199
	再現率				0.99144	0.985676
	適合率				.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**

語彙素認定で

約97～98%

精度について

- 未知語なしの解析結果は現代語の解析辞書の精度とほぼ同等。
- 未知語があるテキストでは精度が下がるおそれがある（近代語のテキストは未知語が発生しやすい）。
- 利用方法として
  - この精度でも研究可能な分野で使う
  - 手を加えて100%に近づけて使う

未知語ありテキストの解析

福澤諭吉「学問のすすめ（初編）」	約96.2%
北村透谷「内部生命論」	約96.4%
三宅雪嶺「漢字の利害」太陽1985年1号	約92.6%
添田壽一「経済上の病原」太陽1901年2号	約97.3%
「歩兵操典（綱領）」	約97.7%

※冒頭約1000語を調査した結果。
精度は全て語彙素レベルのF値。

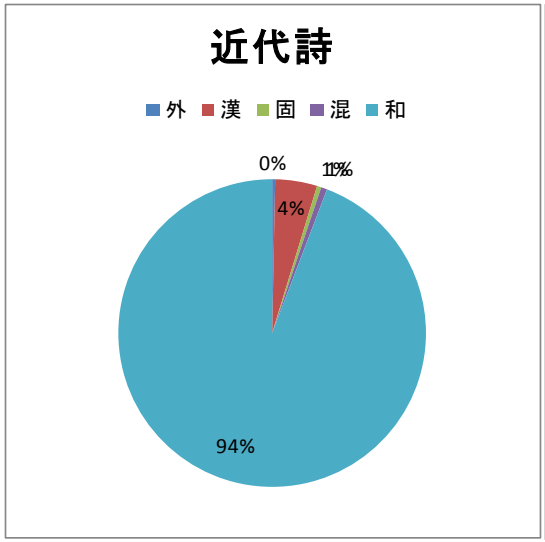
- **Excel ファイル**

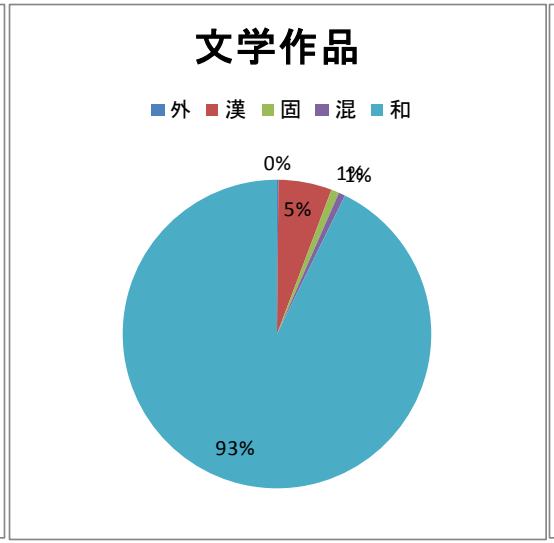
5.解析結果の利用

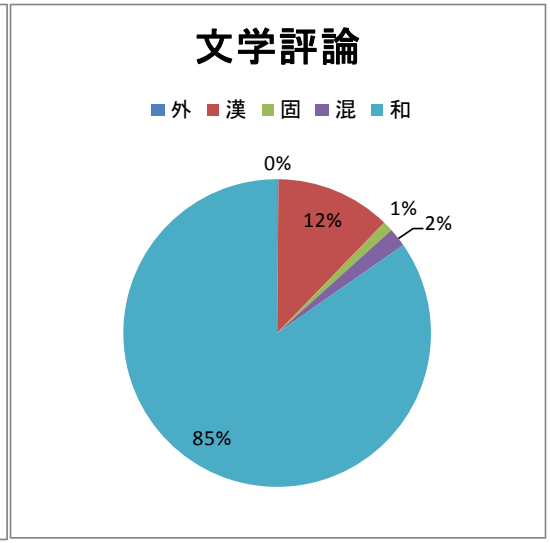
解析結果の利用（デモ）

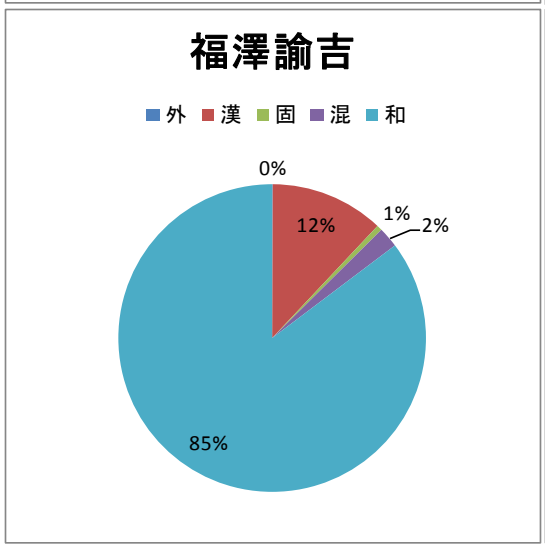
1. **茶まめ**でテキストを解析
2. **Excel**で検索（オートフィルタ）
3. **Excel**で集計（ピボットテーブル）

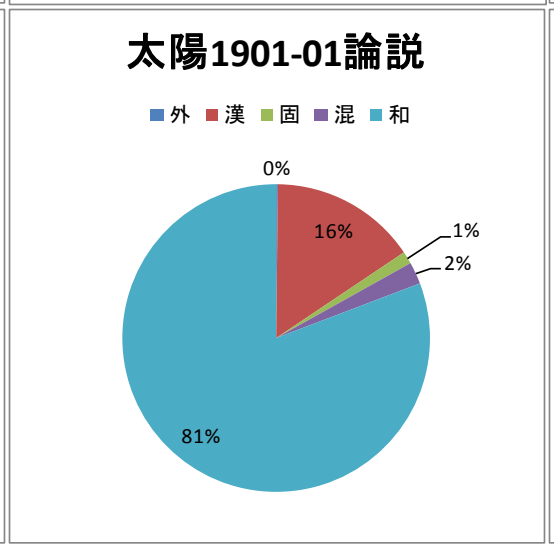
語種比率の比較
（のべ語数・記号を除く）

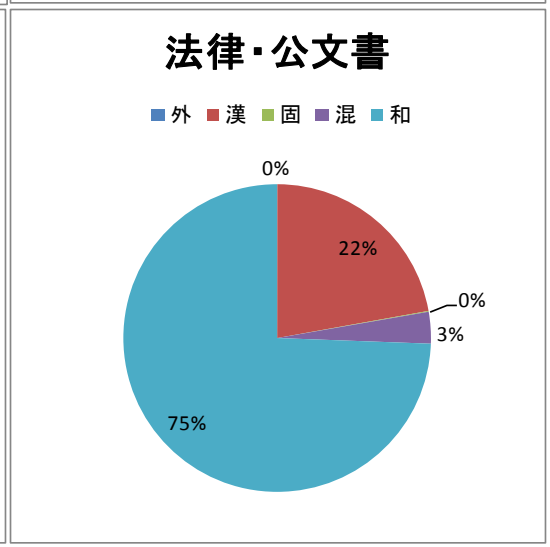






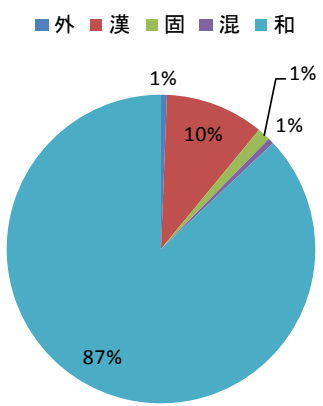




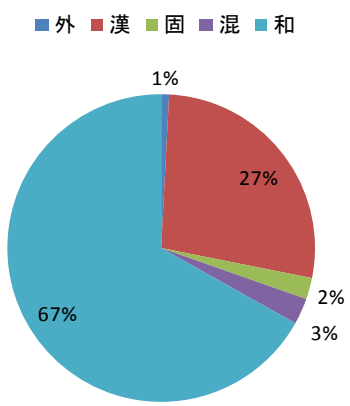


語種比率の比較
（異なり語数・記号を除く）

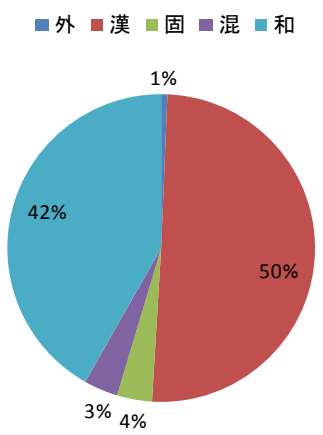
近代詩



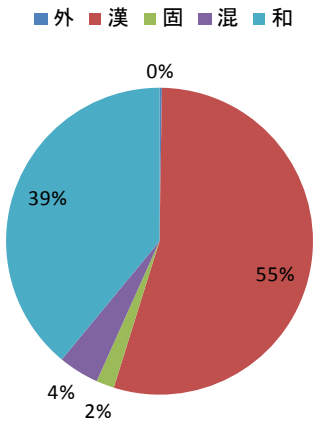
文学作品



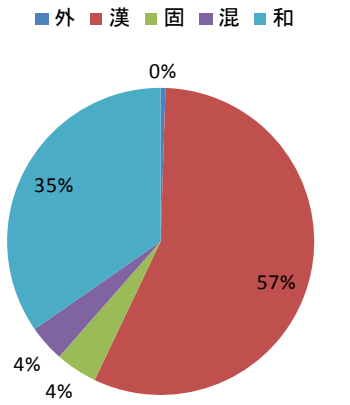
文学評論



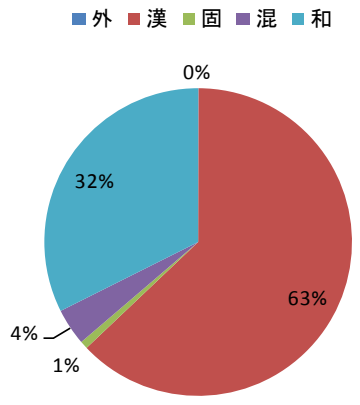
福澤諭吉



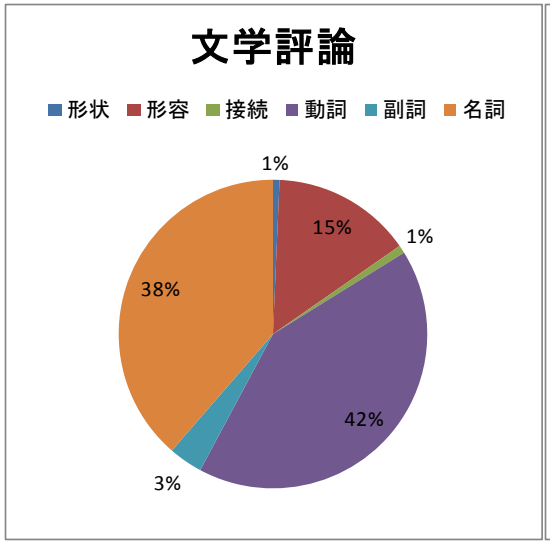
太陽1901-01論説

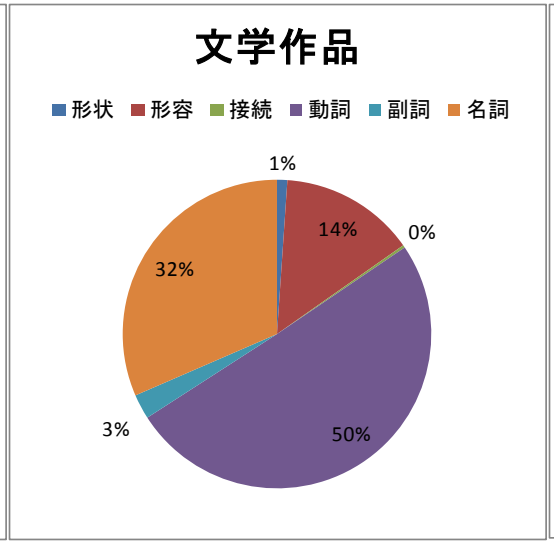


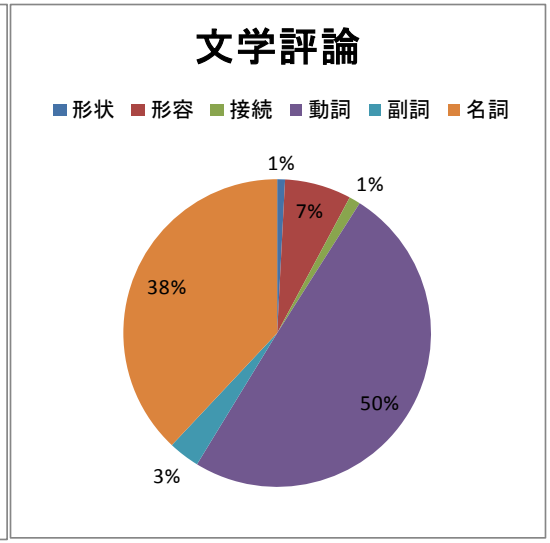
法律・公文書

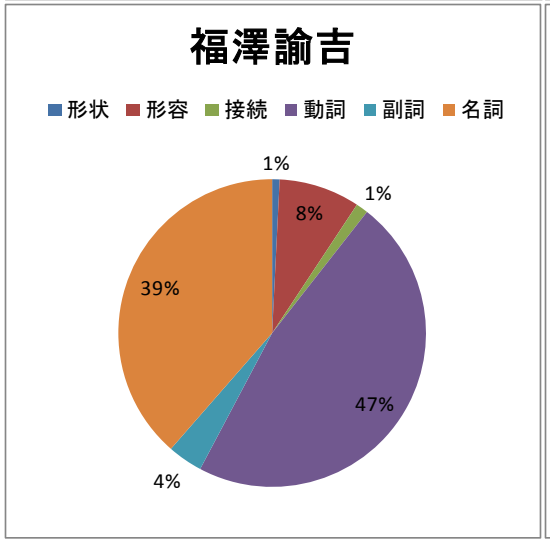


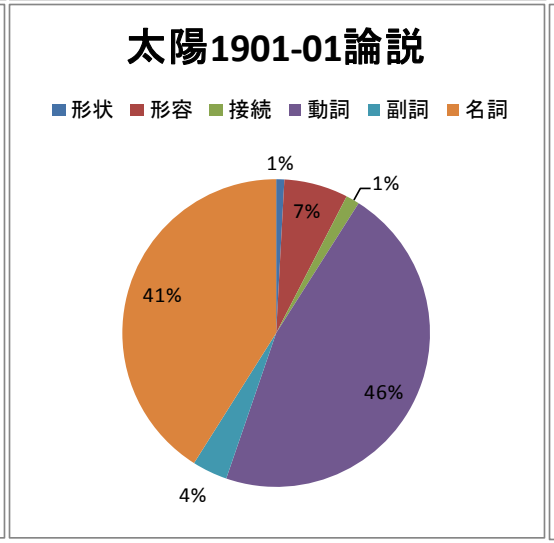
品詞比率の比較
（のべ語数・主な自立語のみ）

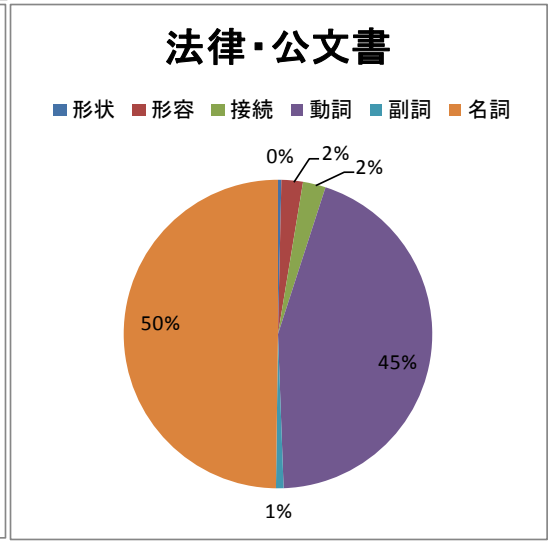




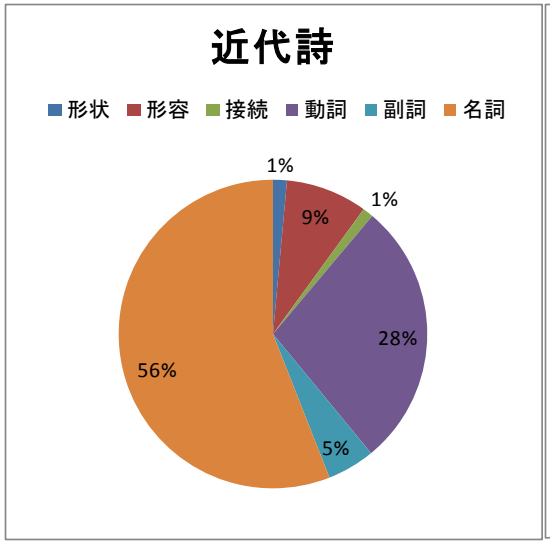


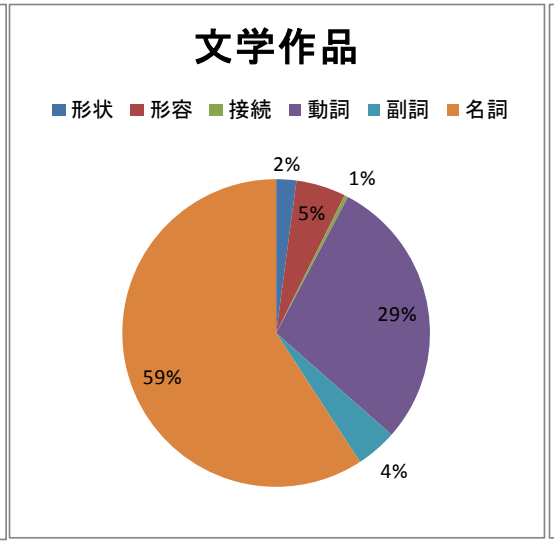


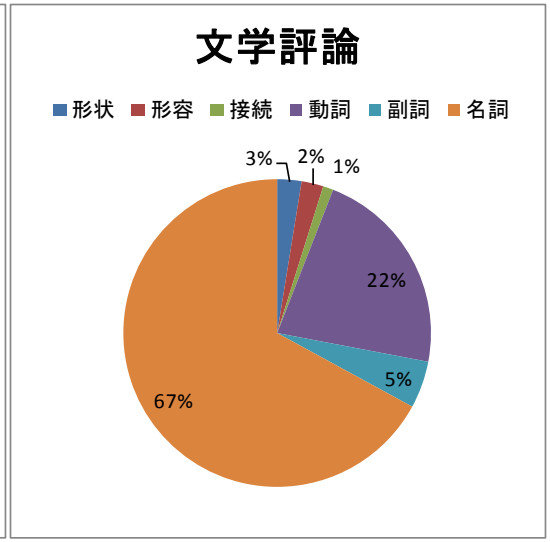


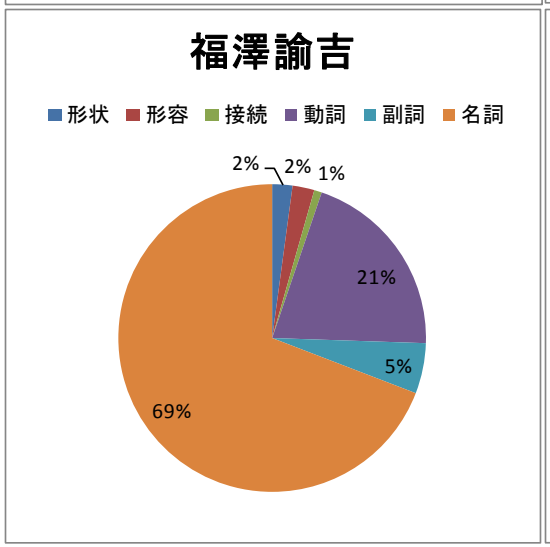


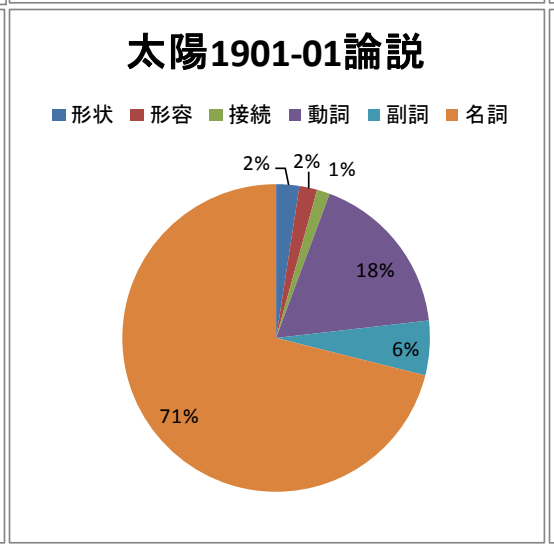
品詞比率の比較
（異なり語数・主な自立語のみ）

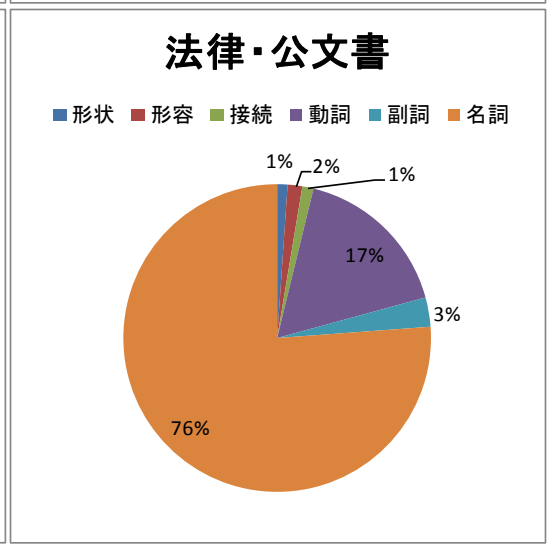






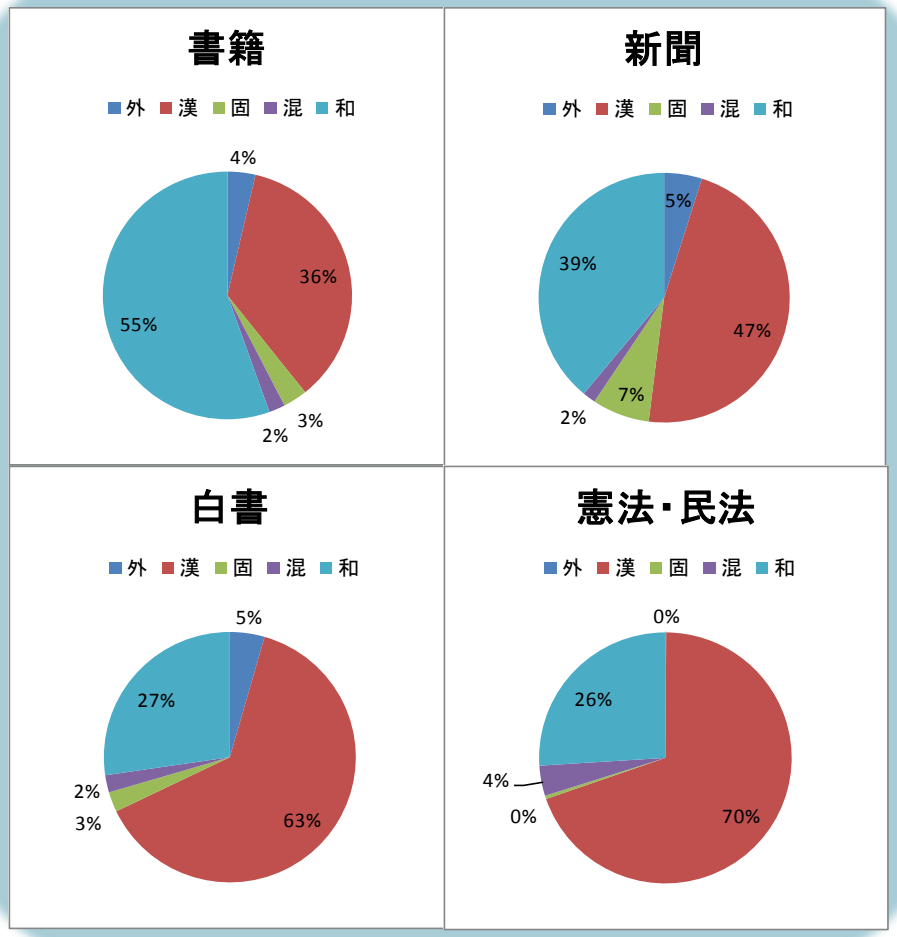




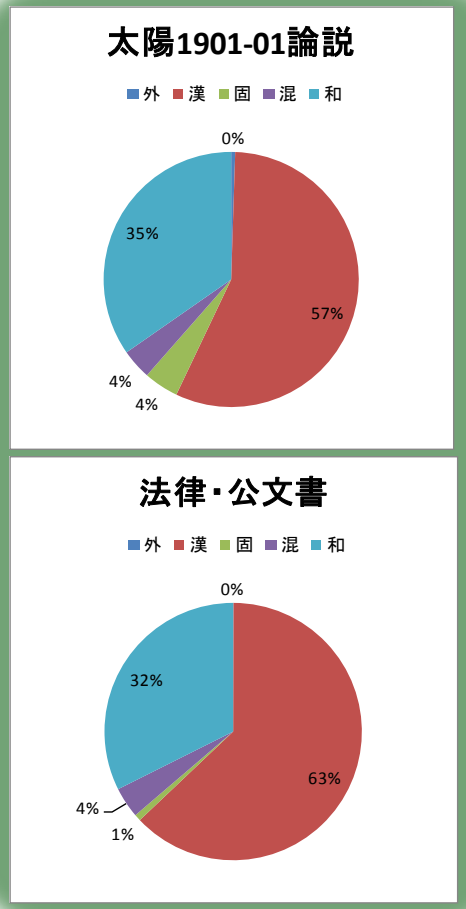


現代語との比較（異なり・語種）

- 同じ「短単位」なので比較が可能。



現代語



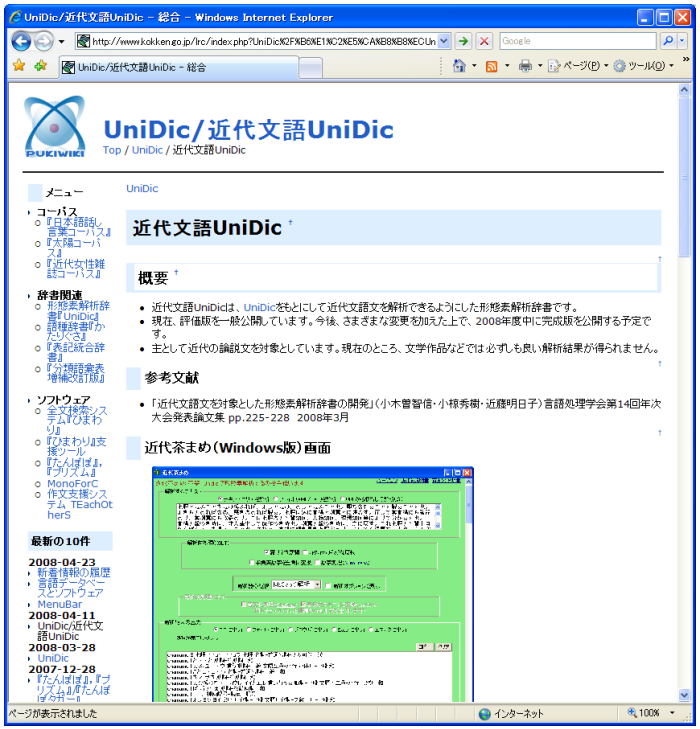
近代語

6.おわりに

近代文語UniDicの入手

- 国語研究所Webサイト「言語データベースとソフトウェア」にて無償公開中
- URL : http://www.kokken.go.jp/lrc/index.php?UniDic

様々な修正を加え精度向上を図ったうえで、2008年度末までに完成版を公開予定。



参考文献

- 国立国語研究所（2005）『太陽コーパス　雑誌『太陽』日本語データベース』博文館新社
- 伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22号 pp.101-122.
- 小木曽智信・小椋秀樹・伝康晴（2007）「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―」『日本語学会2007年度秋季大会予稿集』 pp.255-262.
- 小椋秀樹・小木曽智信・原裕・小磯花絵・冨士池優美（2008）「形態素解析用辞書UniDicへの語種情報の実装と政府刊行白書の語種比率の分析」『言語処理学会第14回年次大会発表論文集』pp.935-938
- 小椋秀樹・小磯花絵・冨士池優美・原裕（2008）『『現代日本語書き言葉均衡コーパス』形態論情報規程集』（国立国語研究所内部報告書LR-CCG-07-04）