日本語歴史コーパス『中納言』の使い方

2013 年 7 月 28 日　国立国語研究所コーパス開発センター・小木曽智信

凡例

検索例①　検索例

☞　参考情報

㊟　注意点

1. ３つの検索方法

（一番上のタブ）

短単位検索
長単位検索
} 形態論情報を使った検索

※長単位データは準備中

文字列検索 } 形態論情報を使わない検索

３つの検索方法の使い分け

短単位検索

- 名詞＋「めく」のような組み合わせ検索で「〇〇めく」を一度に検索できる

㊟ 「冬めく」で検索してもヒットしない（短単位では「冬」＋接尾「めく」）

長単位検索

- 「冬めく」でヒットする

㊟ ただし、自分が検索したいものと一致するとは限らない 例：「昔物語めく」

文字列検索

- 単位を気にせずに文字列で検索できる（たとえば「冬めきて」など）

㊟ あくまでも検索対象は表記なので「冬めく」は「ふゆめく」にヒットしない

2. 形態論情報を利用した検索

2.1. 形態論情報利用の長所

活用語の一括検索

検索例① 語彙素「読む」（終止形）

→ 「読ま」「読み」「読む」「読め」「読もう」（各活用形）

異表記の一括検索

検索例② 語彙素読み「ウグイス」

→ 「うぐひす」「鶯」「鴬」

異語形の一括検索

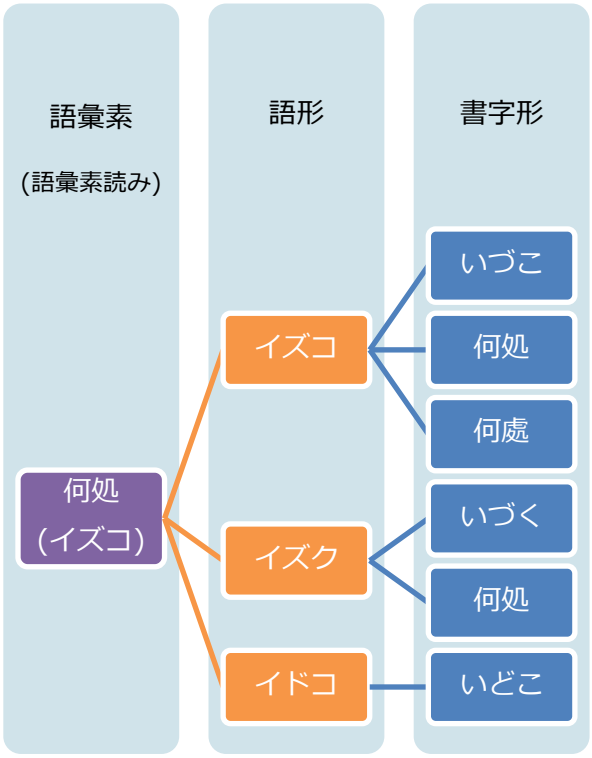
検索例③ 語彙素読み「イズコ」

→ 「イズコ」「イズク」「イドコ」

2.2. 形態論情報の階層構造

BCCWJ の形態論情報の見出し語は次のような階層構造を持っている

（UniDic の見出し階層）



（発音形は省略）

語彙素：辞書の見出しのレベル

語　形：異語形を区別するレベル

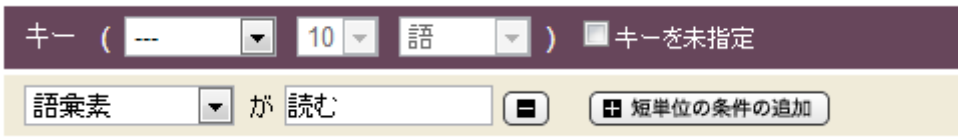
書字形：異表記を区別するレベル

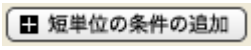
☞ <u>語彙素（見出し語の代表表記）が分からないときは「文字列検索」で検索して確認してみる</u>　（例：何処？いずこ？イズク？ → 「いづこ」で文字列検索、表示される語彙素「何処」を見て確認）

2.3. 検索語の条件指定

形態論情報を使った検索では、次の画面で検索条件を設定する

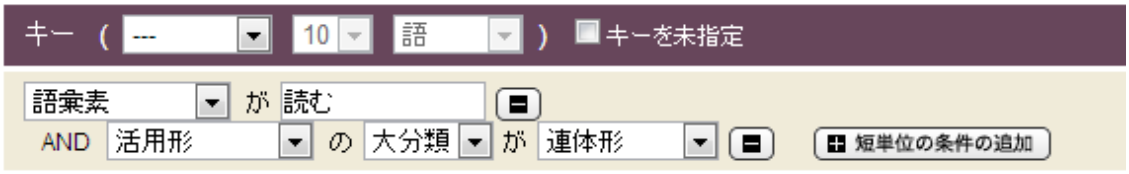
「---選択---」とある部分で条件指定する属性（「語彙素」「出現書字形」など）を選び、右の空欄でその中身を指定する





ボタンで一つの単位について詳細な条件指定を追加できる

検索例④　　語彙素「読む」　+　活用形（の大分類）「連体形」



活用形など選択肢が決まっているものはドロップダウンメニューから選択する

㊟　ここで追加される条件は AND 指定（この画面上では OR 指定はできない）

♣　誤った検索例：語彙素「読む」　+　語彙素「書く」　　→用例数 0 件

2.4. 複数単位の組み合わせ（共起・連接）

ボタンでキーの前方に出現する単位を指定

ボタンでキーの後方に出現する単位を指定

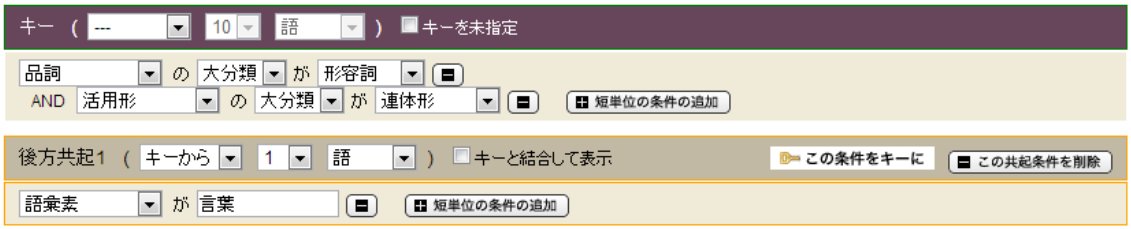
- 「キーから or 文頭から」「N 語 or N 語以内」のように、共起位置を指定可能

- 前方後方合わせて最大 10 個まで共起条件を追加できる

検索例⑤　　「言葉」を連体修飾する形容詞

キー　＝　品詞「形容詞」+ 活用形（の大分類）「連体形」

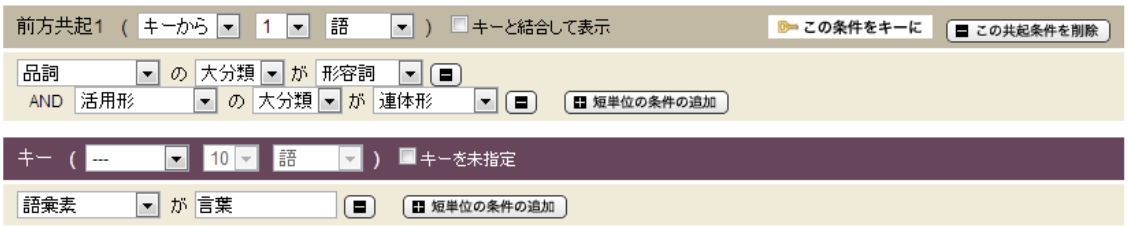
後方共起 1（キーから 1 語）＝ 語彙素「言葉」



☞ <u>集計に使いたいもの（この場合は形容詞）をキーの位置に持ってくるのがコツ</u>

条件指定を入力したあとでも 

ボタンで、キーの位置を移動できる。



㊟ 係り受け情報はアノテーションされていないので、離れた場所にあるものを修飾している例は取得できないし、直後に来ていても必ず修飾しているとは限らない

㊟ 短単位と長単位を組み合わせた検索はできない

2.5. ワイルドカード

語彙素などの検索指定では通常の文字の代わりに、次のワイルドカードが利用できる。

%	任意の文字列	0 文字以上の文字列、何でも良い
[abc]	文字クラス	括弧内の文字いずれか一文字

検索例⑥　心%　「心」ではじまる「心ばえ」「心地」etc.

検索例⑦　[をはも]　「を」または「は」または「も」

2.6. 検索にあたっての注意点

㊟ 「語彙素」の指定だけでは一意に決まらない場合がある（例：辛い→　つらい・からい）

㊟ 語彙素レベルで例外なく一意にするには　語彙素・語彙素読み・品詞・語彙素細分類　の４つを指定する必要がある

㊟ 可能動詞は語形レベルで定義されている（「読める」の語彙素は「読む」）

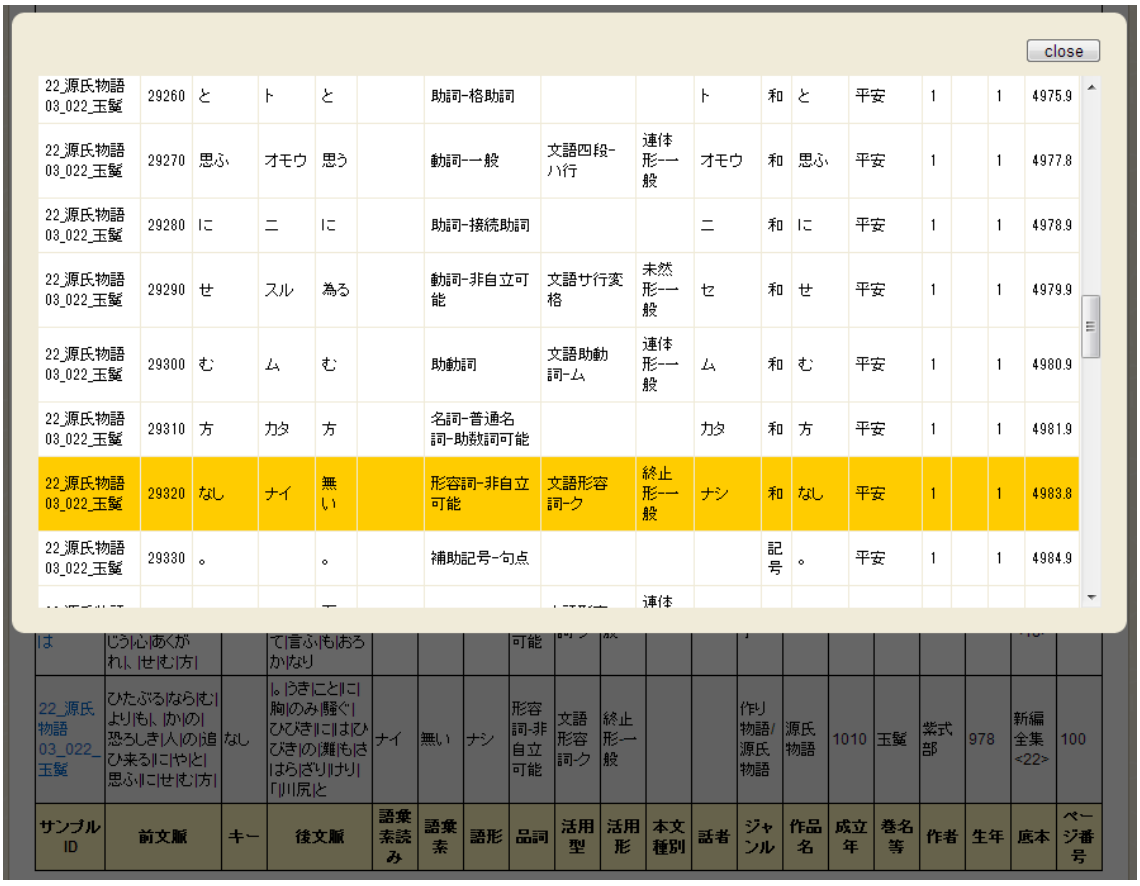
☞ **どう指定したらいいか分からないときは「文字列検索」で該当する例を検索して確認してみる（用例のサンプル ID をクリックすると前後の単位にどのような形態論情報が付与されているか確認できる）**

検索例⑧　文字列検索で「せむ方なし」を検索

2 件の結果が見つかりました。　テーブルの幅を固定 短

サンプルID	前文脈	キー	後文脈	語彙素読み	語彙素	語形	品詞	活用型	活用形	本文種別	話者	ジャンル	作品名	成立年	巻名等	作者	生年	底本	ページ番号
18_枕草子_039_三九_鳥は	て、夜深くうち出でたる声のいとらうらうじう愛敬づきたる、いみじう心あくがれ、せむ方	なし	。六月になりぬれば、音もせずなりぬる、すべて言ふもおろかなり	ナイ	無い	ナシ	形容詞-非自立可能	文語形容詞-ク	終止形一般			随筆/枕草子	枕草子	1001	鳥は	清少納言	966	新編全集<18>	97
22_源氏物語03_022_玉鬘	ひたぶるならむよりも、かの恐ろしき人の追ひ来るにやと思ふにせむ方	なし	。うきにとに胸のみ騒ぐひびきには雷もはばかりけり「川尻と	ナイ	無い	ナシ	形容詞-非自立可能	文語形容詞-ク	終止形一般			作り物語/源氏物語	源氏物語	1010	玉鬘	紫式部	978	新編全集<22>	100
サンプルID	前文脈	キー	後文脈	語彙素読み	語彙素	語形	品詞	活用型	活用形	本文種別	話者	ジャンル	作品名	成立年	巻名等	作者	生年	底本	ページ番号

↓　上の赤丸部分をクリックすると下のように文脈が表示される



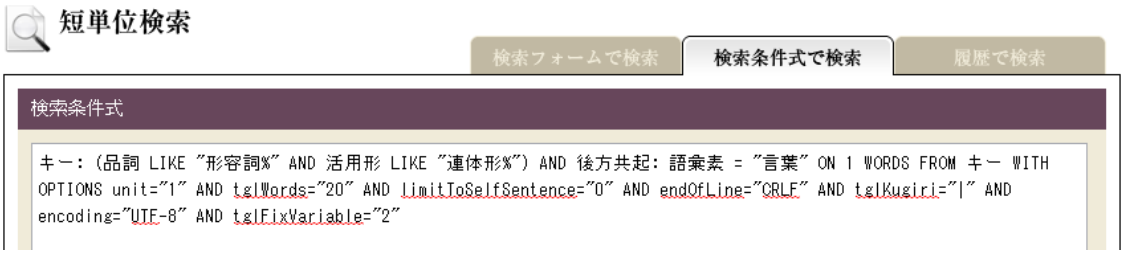
3.　検索条件式

検索画面で指定した検索条件は、「検索条件式」として履歴に記録される（「履歴で検索」で再検索可能）

検索例⑤の検索条件式：

```
キー: (品詞 LIKE "形容詞%" AND 活用形 LIKE "連体形%") AND 後方共起: 語彙素 = "言葉" ON 1 WORDS FROM キー WITH OPTIONS unit="1" AND tglWords="20" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-8" AND tglFixVariable="2"
```

- 「検索条件式」を使うことで、中納言ユーザーなら誰でも、同じ検索を行うことができる
- 研究の再現性のために論文などで使用した検索条件式を明記するとよい



- 検索条件式を修正することで条件の OR 指定が可能

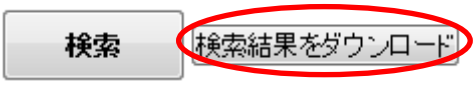
  検索例⑤の修正版（「言葉」または「言語」）：

```
キー: (品詞 LIKE "形容詞%" AND 活用形 LIKE "連体形%") AND 後方共起: ( 語彙素 = "言葉" OR 語彙素 = "言語" ) ON 1 WORDS FROM キー WITH OPTIONS unit="1" AND tglWords="20" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-8" AND tglFixVariable="2"
```

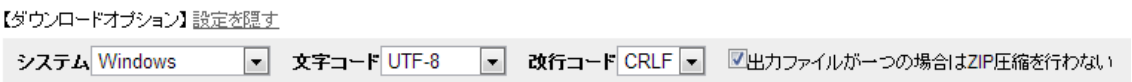
  ㊟ OR, AND は大文字で、前後に半角スペースを入れる．括弧( )も半角
- 検索条件を複数並べると、複数の検索条件を一括して検索、ダウンロードできる

4. 検索結果のダウンロード

中納言自身には集計機能はないので、検索結果をダウンロードして利用する



- 検索画面では 500 例までしか表示されないが、ダウンロード時には最大 10 万件まで一度にダウンロードできる
- 検索画面の【ダウンロードオプション】で、文字コード等を指定できる



  ➢ 使っているパソコンに合わせて自動選択されるが、自分で変更することも可能

  ➢ そのまま Excel に読み込ませる場合はシステム「Excel(Windows)」が便利

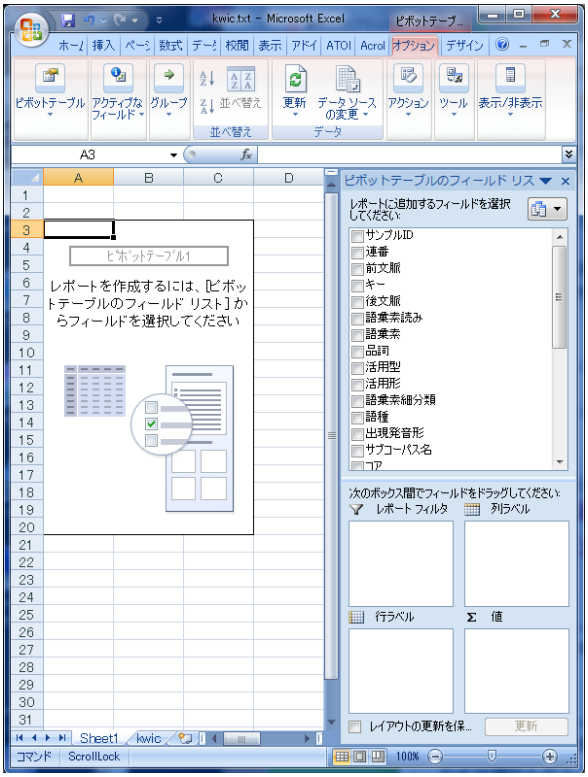
5. Excel による集計（ピボットテーブル）

5.1. Excel での集計の流れ

1. 検索結果をダウンロード
   ㊟ 複数検索時など、検索結果ファイルが zip 圧縮されている場合はダウンロード完了後、自分で解凍しておく
2. 検索結果をインポート
   ㊟ 「ファイルを開く」ダイヤログでファイルの種類を「テキストファイル」または「全てのファイル」にしてファイルを表示後、選択する
   ㊟ システム「Excel(Windows)」にしておけばドラッグアンドドロップで開ける
3. ピボットテーブルの挿入
   ㊟ Excel2007 以降では「挿入」タブ左端のボタンをクリック（バージョンによって大きく違う）
4. ピボットテーブルの作成・集計
   4.2 参照
5. ピボットグラフの作成

5.2. ピボットテーブル

ユーザーの指示により動的にクロス集計表を作る機能



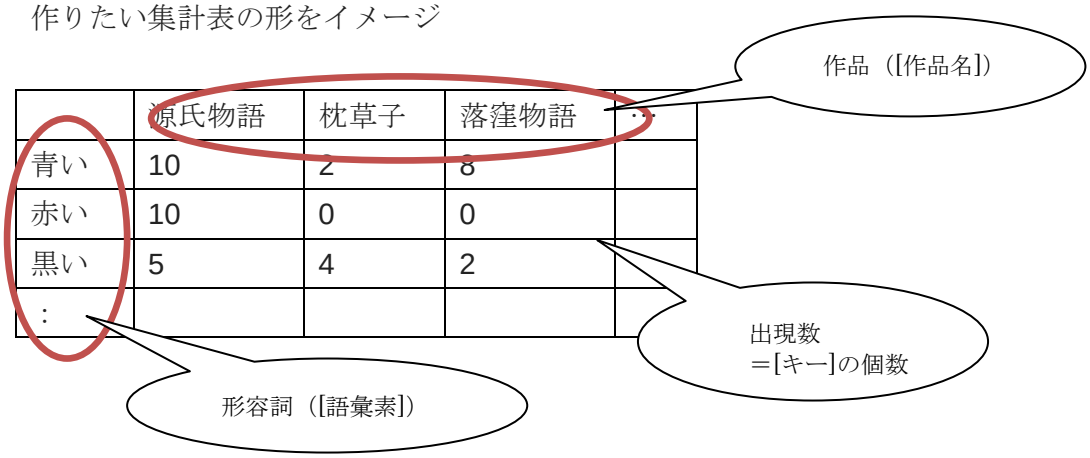
ピボットテーブルの作り方

形容詞の一覧を例に

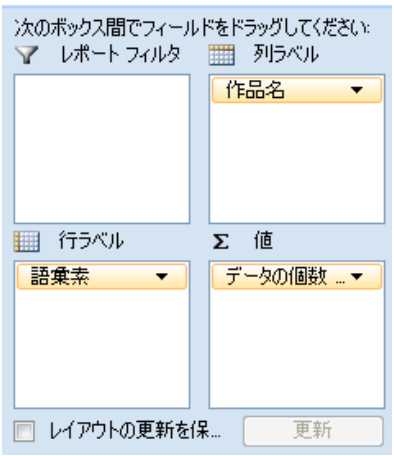
検索例⑨　　形容詞の一覧

キー: 品詞 LIKE "形容詞%" WITH OPTIONS unit="1" AND tglWords="10" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-8" AND tglFixVariable="2"

1. 作りたい集計表の形をイメージ



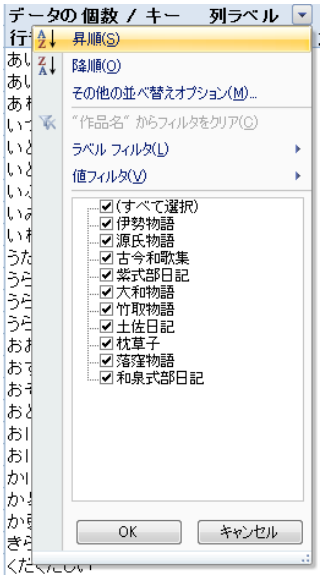
2. イメージに合わせて行ラベル/列ラベルをドラッグアンドドロップで指定
（ドロップすべき場所はイメージした表と位置関係が同じ）



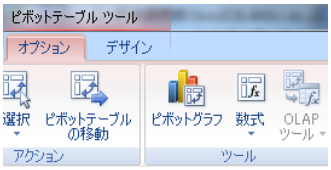
3. これだけで自動的に集計表ができる

データの個数 / キー	列ラベル							
行ラベル	伊勢物語	源氏物語	古今和歌集	紫式部日記	大和物語	竹取物語	土佐日記	枕草子
あいだちない		3						
あいない		105		4				9
あわつけい		22		1				
いつかしい		6						
いとおしい		387		7	4	2		24
いとどしい	1	34		1	1			
いぶせい		50	1				1	2
いみじい	6	700		22	32	9		345
いわけない		60						
うたてい		26						2
うら若い	1				1			
うら寂しい		2	1					
うら悲しい		2						
おおけない		29						1
おずましい		1						
おぞい		2						
おどろおどろしい		68		5		1		10

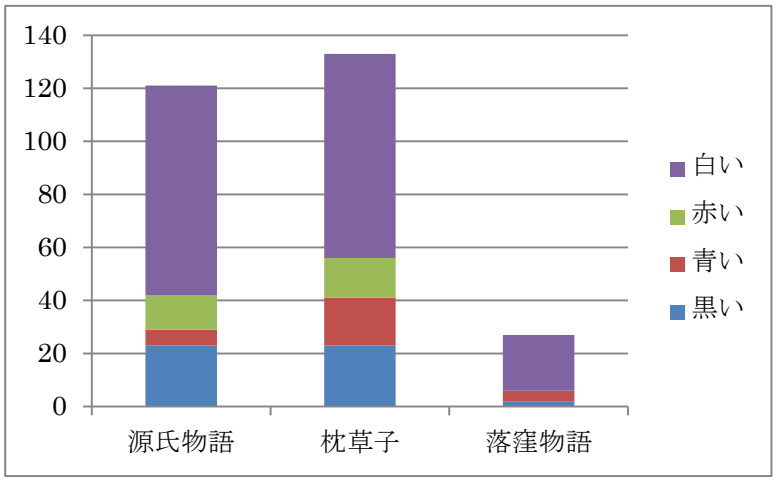
4. 「▼」ボタンで表示する列・行（＝作品名・語彙素）を絞り込むことができる



	A	B	C	D	E
1					
2					
3	データの個数 / キー	列ラベル			
4	行ラベル	源氏物語	枕草子	落窪物語	総計
5	黒い	23	23	2	48
6	青い	6	18	4	28
7	赤い	13	15		28
8	白い	79	77	21	177
9	総計	121	133	27	281



5. ピボットグラフボタンでグラフにまとめることもできる



※行と列を入れ替えている