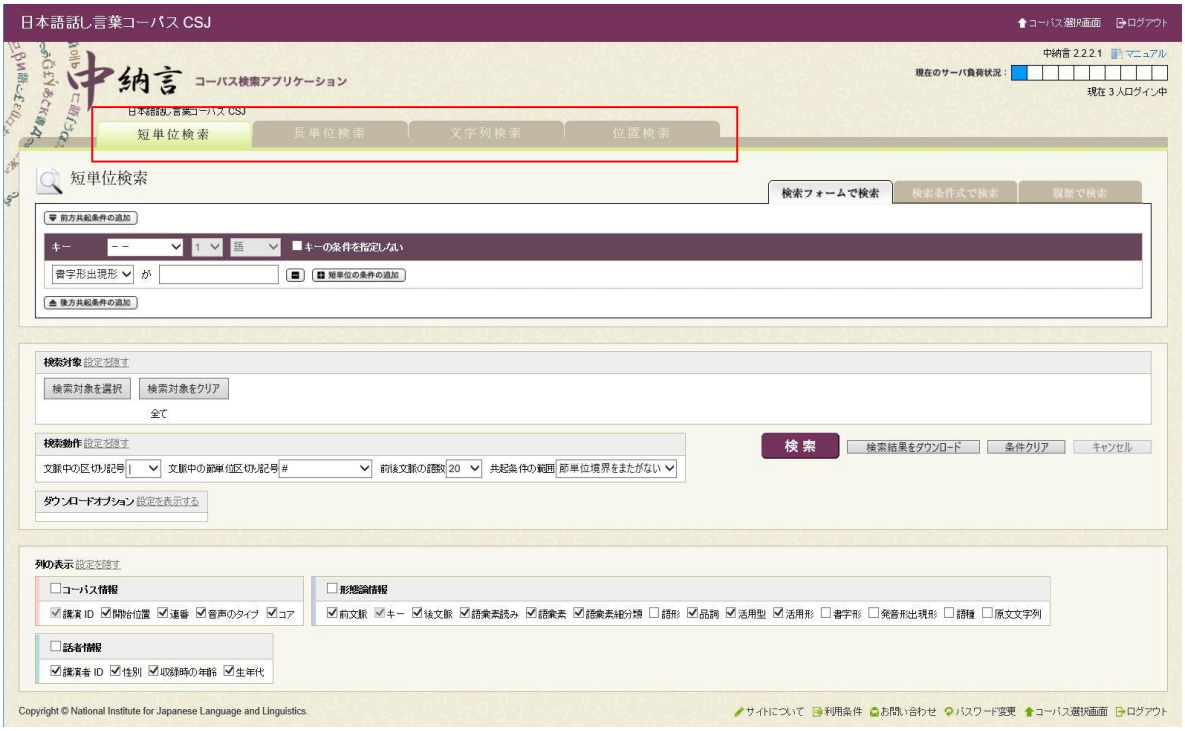
『中納言』概説・実習　CSJ・名大会話

2018 年 3 月 19 日

ログイン画面　https://chunagon.ninjal.ac.jp/auth/login

コーパス選択画面　（どれか１つを選ぶ）

検索画面　（初期画面）



1.　四つの検索方法

（一番上の赤で囲んだタブ）

短単位検索
長単位検索　} 形態論情報を使った検索　※今回主に解説するのは短単位検索

文字列検索　}　形態論情報を使わない検索
位置検索　　}　語の出現位置を指定する検索

検索方法の使い分け

短単位検索

- 名詞＋「的」のような組み合わせ検索で「oo的」を一度に検索できる
  「基本的」で検索してもヒットしない（短単位では「基本」＋「的」）

長単位検索

- 「基本的」でヒットする
  ただし，「基本的人権」に含まれる「基本的」はヒットしない。

文字列検索

- 単位を気にせずに任意の文字列で検索できる（たとえば「と言った」など）

- 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：「プリン」で検索すると「サンプリング」「チャップリン」などもヒットする。
- 語ではなく，表記を検索するので「おこなう」で「行う」は検索できない。
- 語彙素が分からないときに利用すると効果的。

位置検索

- [会話 ID],[語の開始位置]の指定で検索
- 複数の同時検索が可能（ひと組ずつの改行が必要）

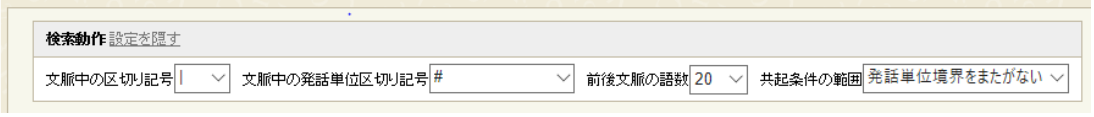
2. 検索の前にやること（検索対象の選択，検索動作の指定）

検索対象の選択

例：コアを選択

コアとは，CSJ の一部で人手により解析精度を高め，韻律情報，係り受け構造情報などの付加情報を追加した部分（約 50 万語）。

検索動作の指定



節単位（発話単位）境界をまたぐ／またがない：＃で表示される部分をまたいで検索するかどうか。

3. 形態論情報を利用した検索

今回は短単位検索を例に説明するが，長単位でも方法は同じ。

3.1. 形態論情報利用の長所

活用語の一括検索

①検索例：語彙素「読む」（終止形）
→ 「読ま」「読み」「読む」「読め」「読もう」（各活用形）

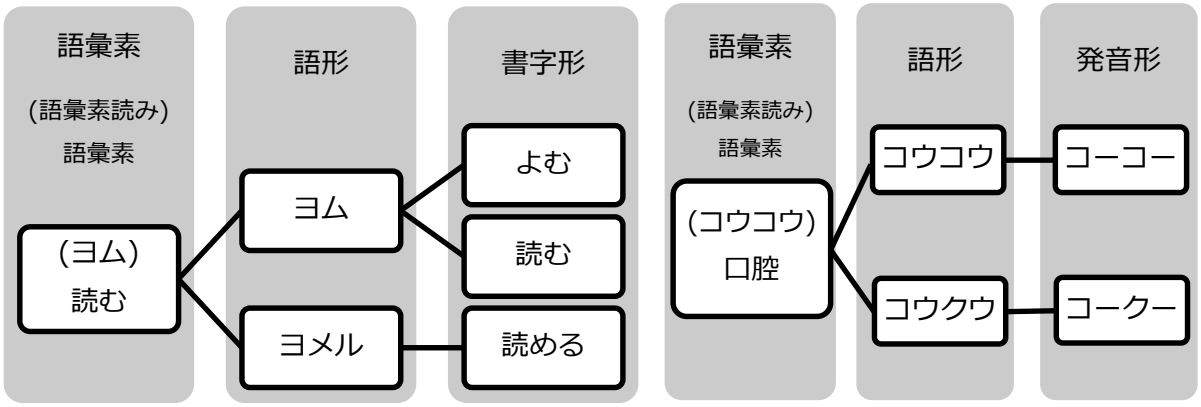
異語形の一括検索

②検索例：語彙素読み「ヤハリ」
→ 「やはり」「やっぱり」「やっぱ」「やっぱし」

異表記の一括検索
（BCCWJ の場合有効。CSJ では異表記は原則ないため関係しない。名大会話も異表記はあまりない。）

③検索例：語彙素読み「タマネギ」
→ 「玉葱」「玉ねぎ」「たまねぎ」「タマネギ」「玉ネギ」
（CSJ はすべて「玉ネギ」，名大会話はすべて「タマネギ」）

3.2.　形態論情報の階層構造

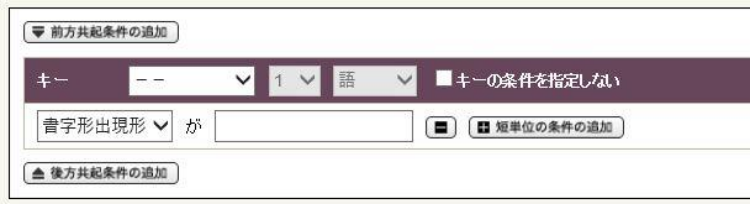


語彙素：辞書の見出しのレベル
語　形：異語形を区別するレベル
書字形：異表記を区別するレベル（CSJ・名大会話ではあまり関係しない）
発音形：語形の下にあり，語形を表音的に表す

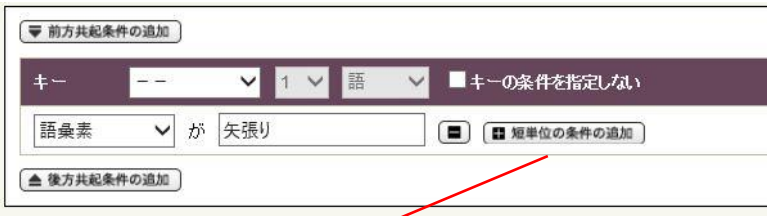
☞ **語彙素（見出し語の代表表記）が分からないときは「文字列検索」で検索して確認してみる。**
（「やはり」？「やっぱり」？）→（「やはり」で検索，表示される語彙素は「矢張り」）

3.3.　検索語の詳細な条件指定

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。



条件を指定する属性はデフォルトで「書字形出現形」になっている。
検索したい属性（「語彙素」「品詞」など）をプルダウンメニューから選び，右の空欄にその中身を入力する。



短単位の条件の追加　ボタンで１つの単位について詳細な条件指定を**追加**できる。

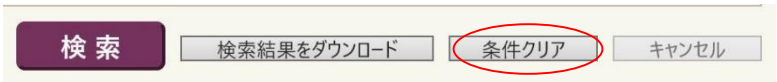
④検索例：語彙素「読む」　＋　活用形（の大分類）「連体形」

活用形など選択肢が決まっている属性はドロップダウンメニューから選択できる。

㊟　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（但し，検索エラーにはならない）

☞　短単位条件の追加で追加した条件を削除するには，検索ボタンの右にある「条件クリア」をクリック。



3.4.　複数単位の組み合わせ（共起・連接）

前方共起条件の追加　ボタンで前方に出現する単位を指定。

後方共起条件の追加　ボタンで後方に出現する単位を指定。

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

- 前方と後方とで合わせて最大 10 個まで共起条件を追加できる。

⑤検索例：「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>

検索条件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。

⑥検索例：修飾表現「ooな言葉」→
「品詞：形状詞」「書字形出現形：な」＋「語彙素：言葉」

3.5. ワイルドカード

語彙素などの検索指定では次のワイルドカード（記号部分は半角）も利用できる。

%	任意の文字列… 0 文字以上の文字列，何でも良い。
_	任意の 1 文字　（一部利用に制限がある）
[abc]	文字クラス… 括弧内の文字いずれか一文字。
[^abc]	文字クラスの補集合… 括弧内の文字以外のいずれか一文字。

⑦検索例：心%　　「心」ではじまる「心がけ」「心地」etc.

⑧検索例：[はがも]　「は」または「が」または「も」

3.6. 特殊な検索指定

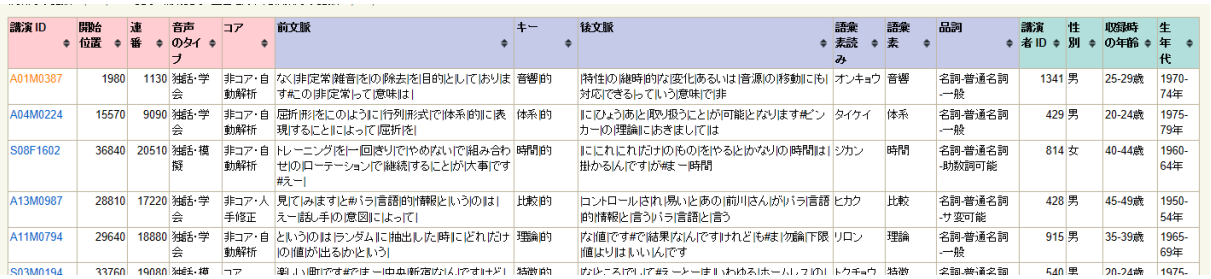
前方・後方共起が指定されている場合，キーに隣接する共起条件の「キーと結合して表示」をチェックすることでキーに隣接する前後条件を含めることができる。

⑨検索例：　キー　＝　語彙素「的」
　　前方　＝　品詞「固有名詞」（キーと結合して表示）

講演 ID	開始位置	連番	音声のタイプ	コア	前文脈	キー	後文脈	語彙素読み	語彙素	品詞	講演者 ID	性別	収録時の年齢	生年代
A07M0395	25290	14550	独話・学会	非コア・自動解析	はまー|例えば|あの|ベイズ|の|あのー|定理|で|決まって|くる|という|風|に|かうん|考え|られ|	ベイズ|的	|に|考えます|と#え|こう|いう|風|に|えー|にっち|と|L|から|F|って|いう|の|は|こう	テキ	的	接尾辞-形状詞的	1212	男	40-44歳	1960-64年
S00M0008	28330	17230	独話・模擬	非コア・人手修正	いう|シーン|を|必ず|撮って#後|神社|と|か|お|墓|と|か|そう|いう|もの|と|か|	日本|的	|な|もの|と|か|を|よく|使って|る|から#そう|いう|の|も|本当|本人|に|会ったら|聞い|て	テキ	的	接尾辞-形状詞的	168	男	25-29歳	1970-74年
A09M0407	67350	40090	独話・学会	非コア・自動解析	の|ところ|は|決して|本来|的|な|意味|で|達成|され|る|こと|の|ない|ま|一種|の|その|	ユートピア|的	|な|形式|で|しか|ありません#だ|からにそ|ま|こう|いった|テーマ|群|と|いう|の|は	テキ	的	接尾辞-形状詞的	1177	男	30-34歳	1965-69年
A07M0368	18390	11030	独話・学会	非コア・自動解析	なお|予測|の|評価|は|従来|の|単一|の|モデル|に|基づき|モデル|平均|を|用い|ない|場合|の|	ベイズ|的	|な|方法|と|比較|しました#この|段階|に|した|沿って|説明|します#まず|それぞれ|の|モデル	テキ	的	接尾辞-形状詞的	149	男	25-29歳	1975-79年
S06M1006	33620	19640	独話・模擬	非コア・自動解析	まー|経営|的|に|も|凄く|素晴らしい|と|は|思う|んです|が#まー|そこ|をまー|凄く|もう|	ガリバー|的	|存在|です|よ|ね#通信|では|で|そこ|で|DDI|と|か|IDO|と|か|後|KDD|という	テキ	的	接尾辞-形状詞的	118	男	30-34歳	1965-69年

キーを未指定にすることで広く「OO的」を検索することも可能。

⑩検索例：　キー　＝　未指定
　　後方共起１　＝　語彙素「的」（キーと結合して表示）



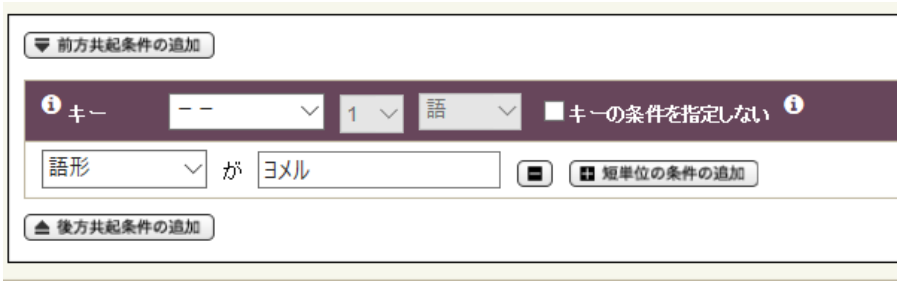
講演 ID	開始位置	連番	音声のタイプ	コア	前文脈	キー	後文脈	語彙素読み	語彙素	品詞	講演者 ID	性別	収録時の年齢	生年代
A01M0387	1980	1130	独話-学会	非コア-自動解析	なく|非|定常|雑音|を|の|除去|を|目的|として|あります#この|非|定常|って|意味|は|	音響|的	|特性|の|経時|的|な|変化|あるいは|音源|の|移動|にも|対応|できる|って|いう|意味|で|非	オンキョウ	音響	名詞-普通名詞-一般	1341	男	25-29歳	1970-74年
A04M0224	15570	9090	独話-学会	非コア-自動解析	|屈折|形|を|この|ように|行列|形式|で|体系|的|に|表現|する|ことの|によって|屈折|を|	体系|的	|に|ひょう|あと|取り|扱う|こと|が|可能|となります#ビンカーの|理論|におきまして|は	タイケイ	体系	名詞-普通名詞-一般	429	男	20-24歳	1975-79年
S08F1602	36840	20510	独話-模擬	非コア-自動解析	トレーニング|を|一|回|ごり|で|やめ|ない|で|組み合わせ|の|ローテーション|で|継続|することの|が|大事|です#えー|	時間|的	|にこれこれ|だけの|もの|を|やる|とかなり|の|時間|は|掛かる|んです|が#ま|一時間	ジカン	時間	名詞-普通名詞-助数詞可能	814	女	40-44歳	1960-64年
A13M0987	28810	17220	独話-学会	非コア-入手修正	見|て|みます|と#バラ|言語|的|情報|という|のは|えー|話し手|の|意図|によって|	比較|的	コントロール|され|易い|とあの|前川さん|が|バラ|言語|的|情報|と|言う|バラ|言語|と|言う	ヒカク	比較	名詞-普通名詞-サ変可能	428	男	45-49歳	1950-54年
A11M0794	29640	18880	独話-学会	非コア-自動解析	という|のは|ランダム|に|抽出|した|時|に|どれ|だけ|の|値|が|出る|か|という|	理論|的	な|値です#で|結果|なん|ですけれど|も#ま|ク論|下限|値より|は|いいん|です	リロン	理論	名詞-普通名詞-一般	915	男	35-39歳	1965-69年
S03M0194	33760	19080	独話-模	コア	測|い|間|です#で|ま|一|中央|新宿|なん|ですけど|	特徴|的	な|ことで|ない|ので#えー|とーまる|いわゆる|ホーム|しる|の|	トクチョウ	特徴	名詞-普通名詞	540	男	20-24歳	1975-

3.7.　検索にあたっての注意点

㊟　語彙素だけでは一意にならない（辛い→　つらい・からい）。
　　語彙素レベルで例外なく完全に一意にするには　語彙素・語彙素読み・品詞・語彙素細分類　の四つを指定する必要がある。但し語彙素細分類は中納言上は指定できない。

㊟　可能動詞は（短単位では）語形レベルで定義されている（「読める」の語彙素は「読む」だが，語形は「ヨメル」）。

⑪検索例：可能表現「読める」→　語彙素「読める」でなく語形「ヨメル」を指定。

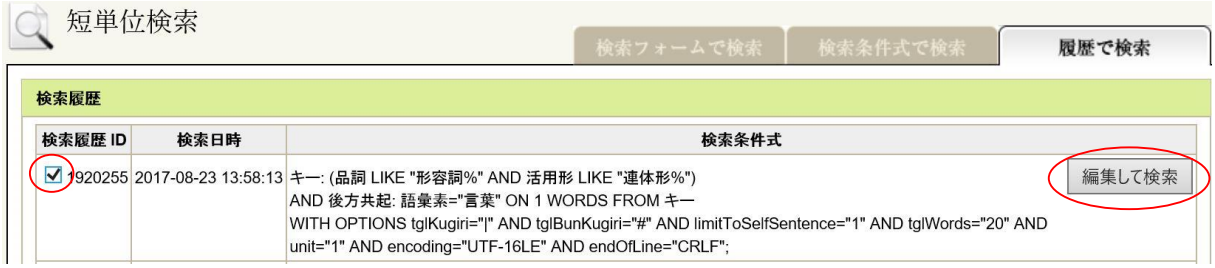


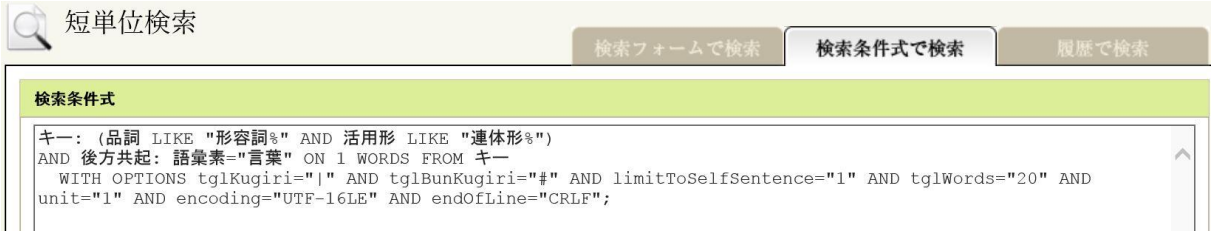
☞ どのように指定したらいいか分からないときは「文字列検索」で該当する例を検索して確認してみる（用例のサンプルIDをクリックすると前後の単位にどのような形態論情報が付与されているか確認できる）ただし，「語形」は表示されないので注意。

4. 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。検索履歴 ID にチェックを入れ，「編集して検索」で再検索ができる。

⑤の検索履歴





この状態で，下の「検索」ボタンをクリックすると再検索ができる。

☞ 最初の２行だけでも検索可能。

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

最小限「キー: 語彙素 LIKE "言葉"」だけでも OK。

- 「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。

☞ 例えば，単に「言葉」を「人」に修正するなど，同じ条件で別の語を検索したい場合にも便利に使える。

➢ 検索例⑤の修正版（「言葉」を「人」へ）：

```
キー: 品詞 LIKE "形容詞%"
AND 後方共起: 語彙素="人" ON 1 WORDS FROM キー
WITH OPTIONS tglKugiri="|" AND tglBunKugiri="#" AND limitToSelfSentence="1" AND
tglWords="20" AND unit="1" AND encoding="UTF-16LE" AND endOfLine="CRLF"
```

☞ **OR 指定も可能。**

➢ 検索例⑤の修正版（「言葉」または「人」へ）：

```
キー: 品詞 LIKE "形容詞%"
AND 後方共起: (語彙素="言葉" OR 語彙素="人") ON 1 WORDS FROM キー
WITH OPTIONS tglKugiri="|" AND tglBunKugiri="#" AND limitToSelfSentence="1" AND
tglWords="20" AND unit="1" AND encoding="UTF-16LE" AND endOfLine="CRLF"
```

㊟ OR, AND は半角大文字で，前後に半角スペースを入れる，括弧()も半角。

5.　検索結果

5.1.　検索結果の表示

検索結果は，見つかった件数と，500 件までの該当例が画面上に表示される。
500 件以上ある場合には表示は 500 件まで。並び順は無秩序。

5.2.　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。

5.3.　検索結果に関する注意

共起検索で同じレコードが複数表示されることがある。

- たとえば，前方共起条件がキーから３語以内にある名詞である場合，同一のキーから３語以内に２つの名詞があれば，同じ ID の行が２度表示される

- 並び順は無秩序
- ランダム性は保証されない。ランダムサンプリングする際は全件取得後に別途，並べ替えが必要。

検索結果が 10 万件以上ある場合，10 万件で切れる。

- ダウンロードしたファイルには特に警告がないので注意。

6. Excel による集計

6.1. Excel で保存

検索結果は，見つかった件数と，500 件までの該当例が画面上に表示される。
500 件以上ある場合には表示は 500 件まで。並び順は無秩序。

⑫検索例：（全対象で）語彙素「一番」＋形容詞 **1,001** 件
（同様に，語彙素「最も」＋形容詞 **310** 件）

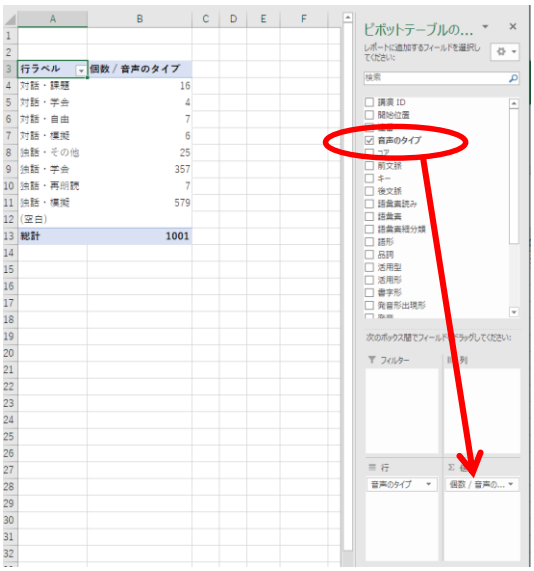
1. 検索結果をダウンロード ※「開く」ではなく「保存」を推奨。

デフォルトは，kwic-1620452.csv　のようなファイル名で保存される。

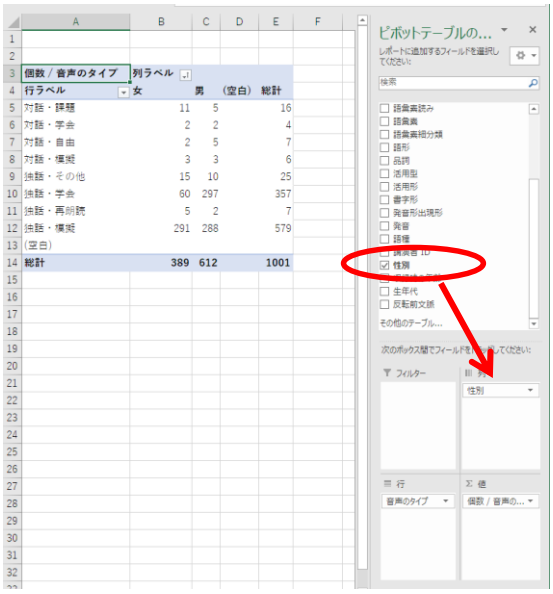
2. Excelファイル形式で保存する。
例：一番形容詞CSJkwic-2403305.csv，最も形容詞CSJkwic-2403311.csv

6.2. ピボットテーブルの作成

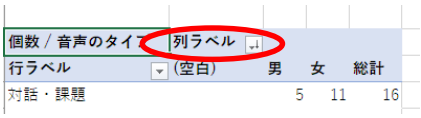
1. Excelの「挿入」→「ピボットテーブル」（一番左）→「OK」。
2. 「音声のタイプ」にチェックを入れる。
3. 「音声のタイプ」を「値」のところへドラッグ＆ドロップする。　※個数が集計される。



4. 「性別」にチェックを入れる。
5. 「性別」を「行ラベル」のところへドラッグ＆ドロップする。　※クロス集計される。



6. 「列ラベル」を「降順」にすると，左から男，女の順になる。



6.3. ピボットテーブルの保存

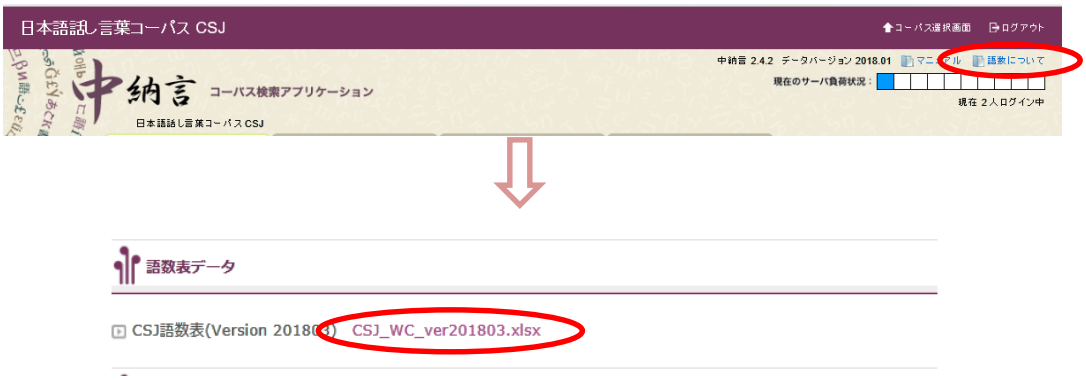
1. テーブル全体を選択して，「コピー」する。

2. 別シートを開き，「貼り付ける」。このとき，必ず形式を「値」にする。

個数 / 音声	列ラベル			
行ラベル	(空白)	男	女	総計
対話・課題		5	11	16
対話・学会		2	2	4
対話・自由		5	2	7
対話・模擬		3	3	6
独話・その他		10	15	25
独話・学会		297	60	357
独話・再朗読		2	5	7
独話・模擬		288	291	579
(空白)				
総計		612	389	1001

6.4. 頻度の処理

1. CSJ語数表をダウンロードする。



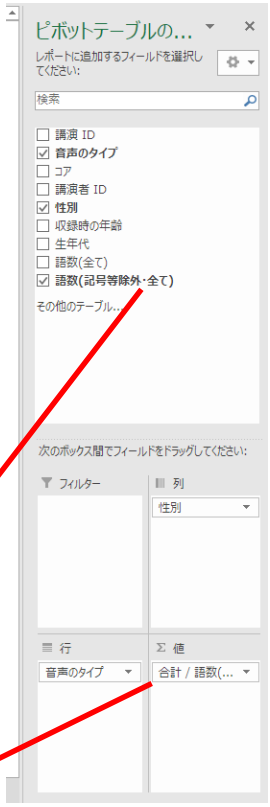
2. ピボットテーブルを用いて，必要な語数表を作成する。

語数表（冒頭一部）

	講演 ID	音声のタイプ	コア	講演者 ID	性別	収録時の年齢	生年代	語数(全て)	語数(記号等除外・全て)
2	A01F0001	独話・学会	非コア・人手修正	72	女	20-24歳	1975-79年	2559	2553
3	A01F0019	独話・学会	非コア・人手修正	76	女	30-34歳	1965-69年	3031	2959
4	A01F0034	独話・学会	非コア・人手修正	465	女	35-39歳	1960-64年	2374	2368
5	A01F0055	独話・学会	コア	459	女	35-39歳	1960-64年	2492	2476
6	A01F0063	独話・学会	非コア・人手修正	467	女	20-24歳	1975-79年	1901	1877
7	A01F0066	独話・学会	非コア・人手修正	42	女	25-29歳	1970-74年	1992	1958
8	A01F0067	独話・学会	コア	426	女	25-29歳	1970-74年	3210	3189
9	A01F0085	独話・学会	非コア・人手修正	84	女	20-24歳	1975-79年	1889	1877
10	A01F0090	独話・学会	非コア・人手修正	84	女	20-24歳	1975-79年	2234	2216

◆例１　音声タイプ別，性別

合計 / 語数(記号等除外・全て)2	列ラベル			
行ラベル	(空白)	男	女	総計
対話・課題		6880	23476	30356
対話・学会		9556	19383	28939
対話・自由		16763	31302	48065
対話・模擬		14303	28835	43138
独話・その他		149988	132812	282800
独話・学会		2707993	591275	3299268
独話・再朗読		25934	23400	49334
独話・模擬		1792999	1843772	3636771
独話・朗読		79484	77900	157384
(空白)				
総計		**4803900**	**2772155**	**7576055**



語数は，語数（記号等除外-全て）の選択を推奨。

「データ個数」となっている場合は，右の▼をクリック，一番下の「値フィールドの設定」をクリックし，一番上の「合計」に変更する。

◆例２　音声タイプ別，年齢別

合計 / 語数(記号等除外・全て)2	列ラベル																
行ラベル	15-19歳	20-24歳	25-29歳	30-34歳	35-39歳	40-44歳	45-49歳	50-54歳	55-59歳	60-64歳	65-69歳	70-74歳	75-79歳	80-84歳	nr	(空白)	総計
対話・課題		1129	7045	13355	3162	1427	2450		1013	775							30356
対話・学会			7500	14673	1445		3196		2125								28939
対話・自由		2682	8614	20146	5168	2435	4394		2178	2448							48065
対話・模擬		2789	7344	18633	4912	2275	3797		1301	2087							43138
独話・その他			22804	30404	21952	41447	27209	96383	26434	7113			9054				282800
独話・学会		517172	709017	536361	407842	409525	212146	208112	148395	48653	31440	11018	4783	16495	38309		3299268
独話・再朗読		3436	6760	16536	6517	2207	4796		6747	2335							49334
独話・模擬	7305	486317	558702	424725	384144	365927	257778	308035	283940	264301	291217	4380					3636771
独話・朗読		12600	27781	18366	13407	15102	11642	15529	14061	17270	11020	606					157384
(空白)																	
総計	**7305**	**1026125**	**1355567**	**1093199**	**848549**	**840345**	**527408**	**628059**	**486194**	**344982**	**333677**	**16004**	**13837**	**16495**	**38309**		**7576055**

3. 粗頻度を調整頻度にする。語数表を用い，「100万語あたり調整頻度」(PMW)へ。
例： =B19/H19*1000000

粗頻度

例：B19

	男	女	総計
対話・課題	5	11	16
対話・学会	2	2	4
対話・自由	5	2	7
対話・模擬	3	3	6
独話・その他	10	15	25
独話・学会	297	60	357
独話・再朗読	2	5	7
独話・模擬	288	291	579
独話・朗読			
総計	612	389	1001

語数表

例：H19

	男	女	総計
対話・課題	6880	23476	30356
対話・学会	9556	19383	28939
対話・自由	16763	31302	48065
対話・模擬	14303	28835	43138
独話・その他	149988	132812	282800
独話・学会	2707993	591275	3299268
独話・再朗読	25934	23400	49334
独話・模擬	1792999	1843772	3636771
独話・朗読	79484	77900	157384
総計	4803900	2772155	7576055

検索結果が 0 だった音声タイプの行を追加している。

4. 調整頻度にした結果。(PMW)

調整頻度(PMW)

=B19/H19*1000000

	男	女	総計
対話・課題	726.7	468.6	527.1
対話・学会	209.3	103.2	138.2
対話・自由	298.3	63.9	145.6
対話・模擬	209.7	104.0	139.1
独話・その他	66.7	112.9	88.4
独話・学会	109.7	101.5	108.2
独話・再朗読	77.1	213.7	141.9
独話・模擬	160.6	157.8	159.2
独話・朗読	0.0	0.0	0.0
総計	7699.7	4993.6	6360.2

6.5. グラフの作成

1. 先の表でグレーの部分を選択する。
2. 「挿入」→「グラフ」の「縦棒」の「100%積み上げ縦棒」を選択する。

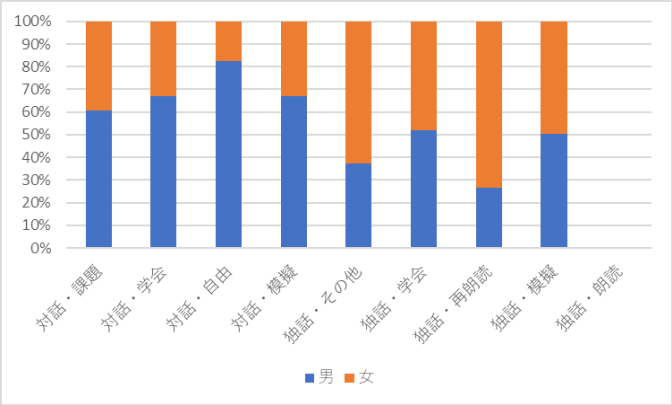


図1　「一番」の音声タイプと性別

3. 「一番」と同様に「最も」でグラフを作成し，比較する。

	男	女	総計
対話・課題	0.0	85.2	65.9
対話・学会	0.0	0.0	0.0
対話・自由	0.0	0.0	0.0
対話・模擬	0.0	0.0	0.0
独話・その他	6.7	7.5	7.1
独話・学会	75.0	87.9	77.3
独話・再朗読	154.2	85.5	121.6
独話・模擬	11.2	7.1	9.1
独話・朗読	75.5	77.0	76.2
総計	48.7	27.4	40.9

検索結果が 0 だった音声タイプの行を追加している。

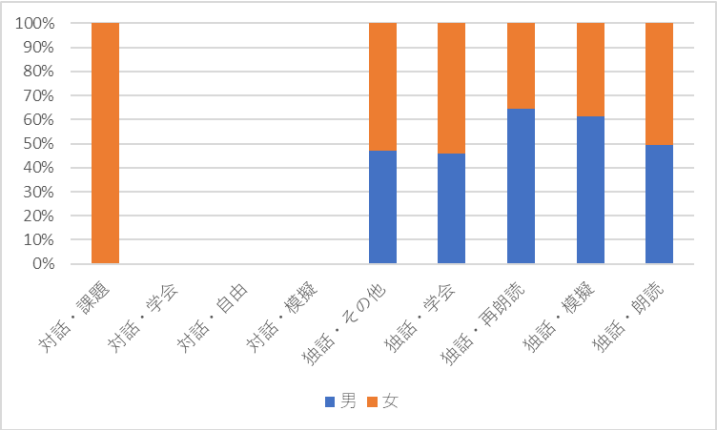


図2　「最も」の音声タイプと性別

	一番	最も
対話・課題	527.1	65.9
対話・学会	138.2	0.0
対話・自由	145.6	0.0
対話・模擬	139.1	0.0
独話・その他	88.4	7.1
独話・学会	108.2	77.3
独話・再朗読	141.9	121.6
独話・模擬	159.2	9.1
独話・朗読	0.0	76.2

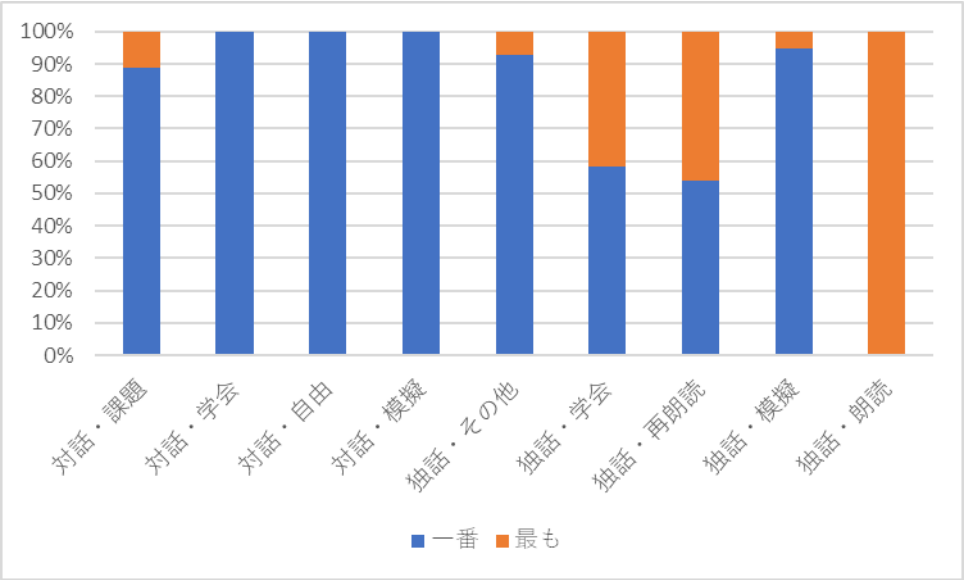


図3　「一番」「最も」の音声タイプの比較

※なお，副詞用法の「一番」と「最も」を正確に比較したい場合は，「一番」「最も」の検索結果より，人手で名詞の用例を除く必要がある（中納言上ではすべて「副詞」となっているため，人手処理が必要）。