形態素解析結果を Access で集計する

2009/07/21 小木曽 智信 ogiso@ogiso.net

1. 形態素解析を行う

・解析結果をファイルに出力する

複数ファイルを一度に解析する場合

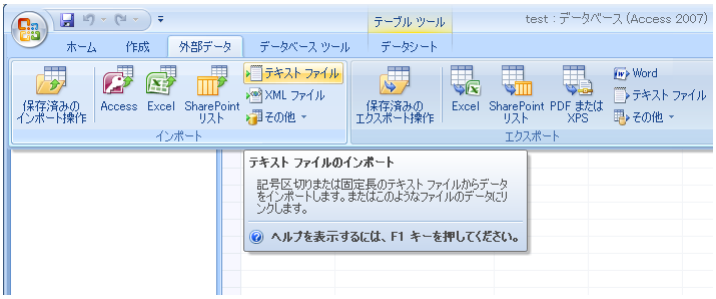
（青空文庫の夏目漱石のテキストを全部解析してみる）

2.　データベースに取り込む

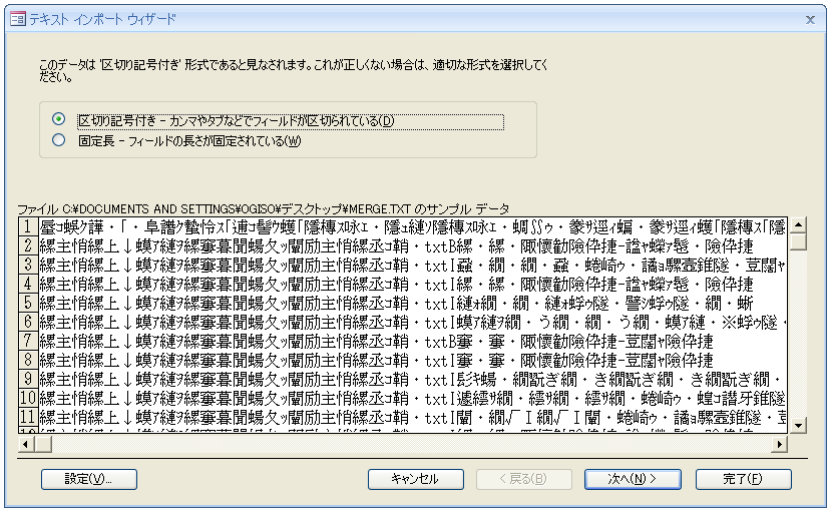
1.　Access で新規データベースを作成(mdb ファイル)

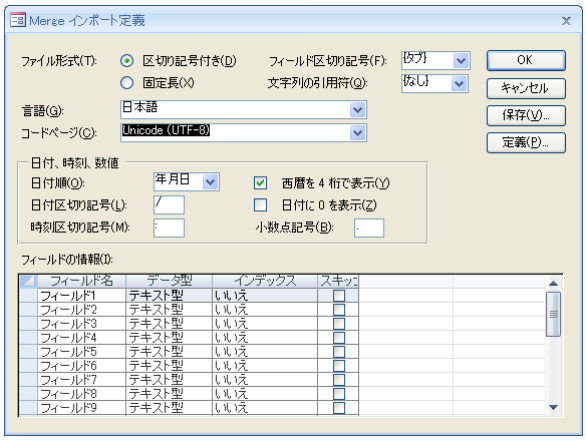
2.　インポート

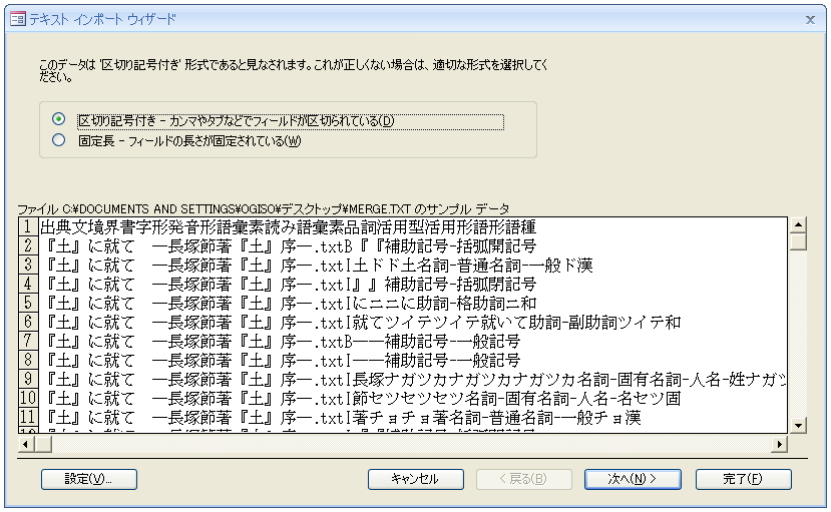
テキストファイルのインポート



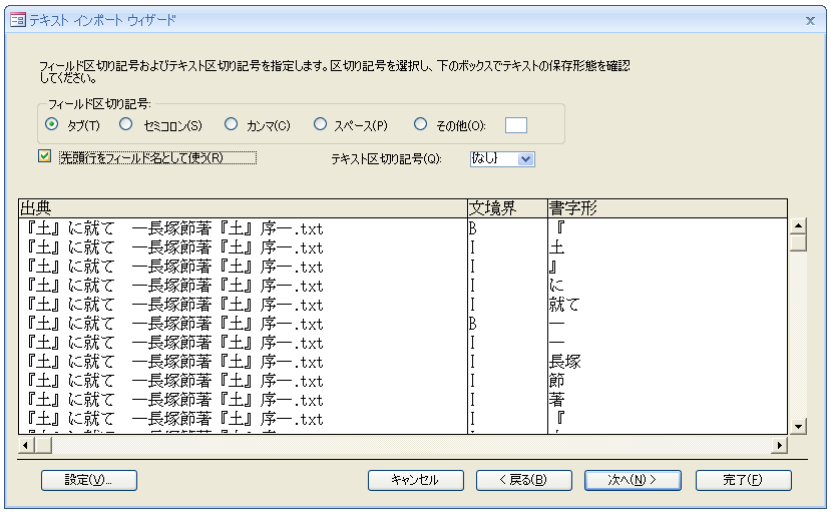
- 「設定」で文字コード（コードページを UTF-8 に指定する）



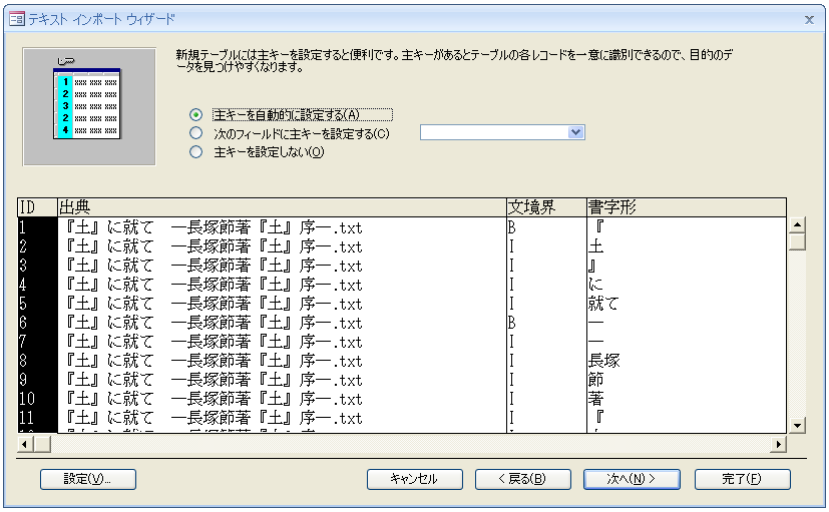




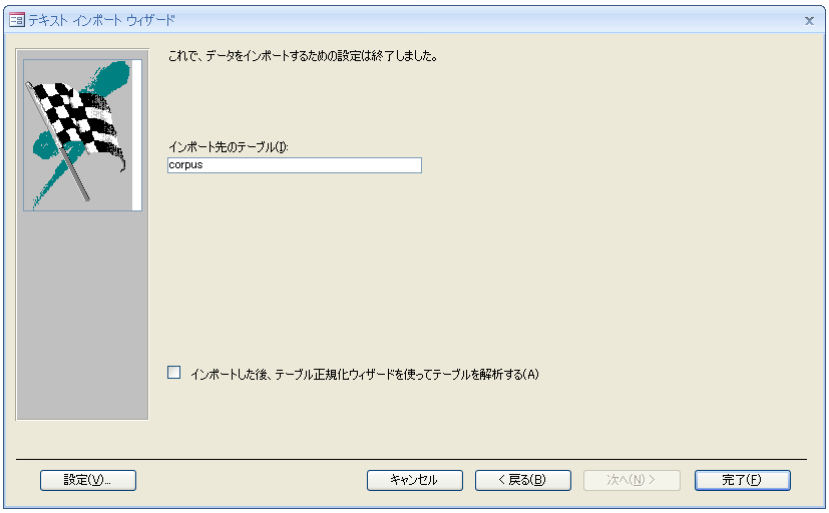
- 先頭行をフィールド名として使う



3. 連番を付ける



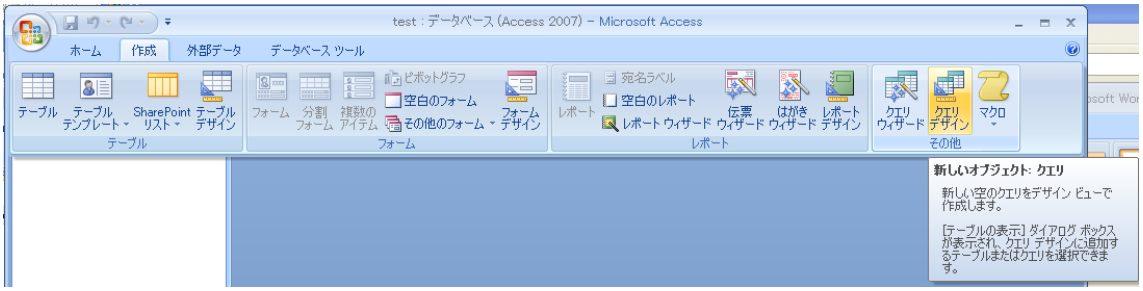
インポート完了



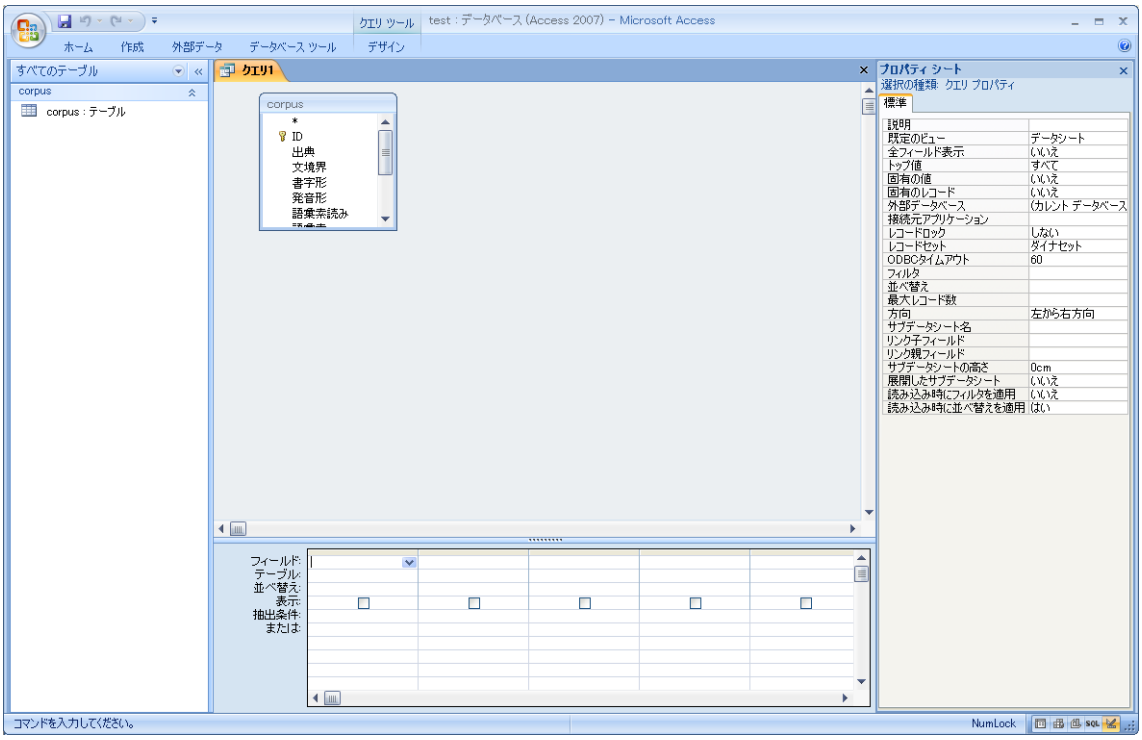
※ここでエラーが出てしまう場合は、解析結果を別の文字コード（Unicode(UTF-16LE)など）で保存し直してためしてみてください。（Access2007 のバグ？）

3.　データを取り出す・集計する

クエリを新規作成



コーパステーブルを表示



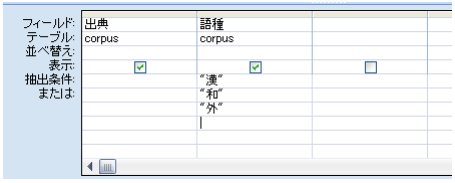
選択クエリ

- 列を選択

（出典と語種だけを選択）

ドラッグアンドドロップで列を追加できる

- 行を選択

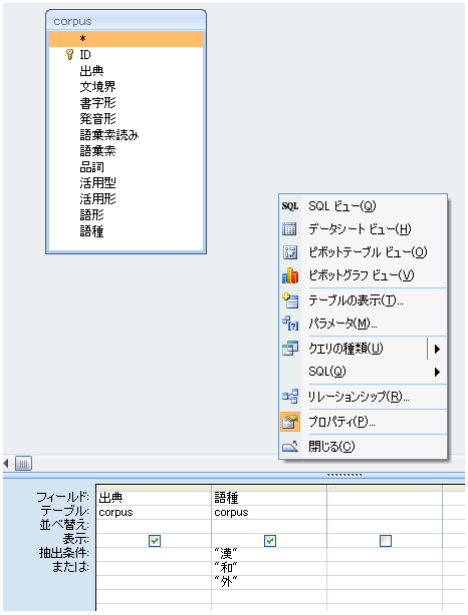


（語種が「漢 or 和 or 外」の行だけを選択）

部分一致は"*"を使う。自動的に like がつく（例 : Like "名詞*"　で名詞で始まる）

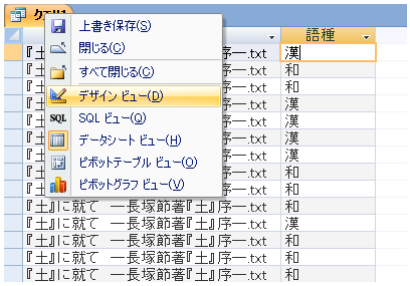
複数の列に抜き出す条件を指定することもできる

Access のビュー

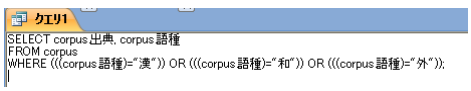


・データシートビューでクエリの実行結果を見ることができる

・デザインビューでクエリのデザインに戻ることができる

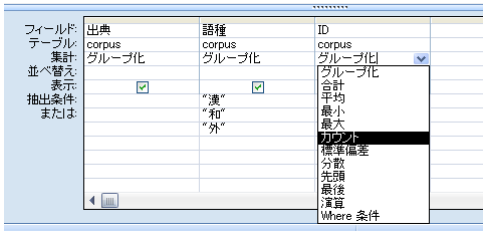


・SQL ビューでクエリの中身（SQL 文）を見ることができる



集計

デザインビューのフィールド上で右クリック→集計　で項目がグループ化される



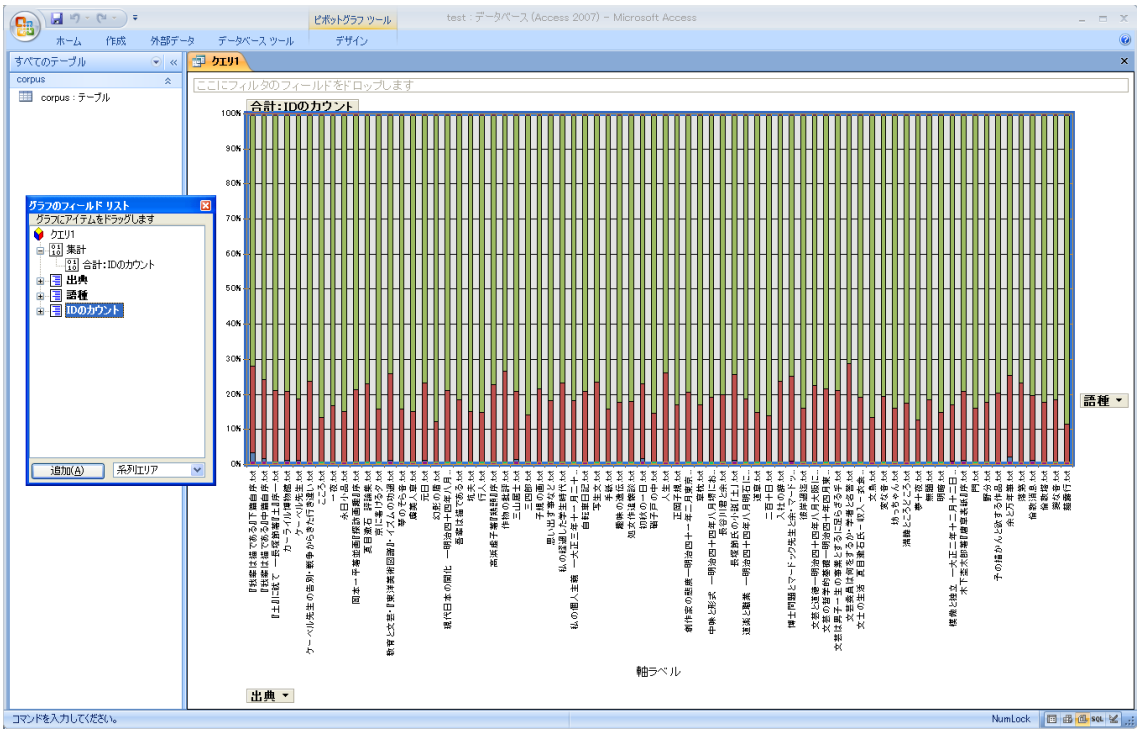
- ID 列を追加してカウント（何例あるか）

クエリ1

出典	語種	IDのカウント
『我輩は猫である』下篇自序.txt	外	12
『我輩は猫である』下篇自序.txt	漢	91
『我輩は猫である』下篇自序.txt	和	264
『我輩は猫である』中篇自序.txt	外	24
『我輩は猫である』中篇自序.txt	漢	300
『我輩は猫である』中篇自序.txt	和	1020
『土』に就て　一長塚節著『土』序一.txt	外	11
『土』に就て　一長塚節著『土』序一.txt	漢	604
『土』に就て　一長塚節著『土』序一.txt	和	2315
カーライル博物館.txt	外	57
カーライル博物館.txt	漢	950
カーライル博物館.txt	和	3827
ケーベル先生.txt	外	29
ケーベル先生.txt	漢	413
ケーベル先生.txt	和	1914
ケーベル先生の告別・戦争からきた行き	外	8
ケーベル先生の告別・戦争からきた行き	漢	298
ケーベル先生の告別・戦争からきた行き	和	981
こころ.txt	外	181
こころ.txt	漢	14230
こころ.txt	和	93140
一夜.txt	外	22
一夜.txt	漢	762
一夜.txt	和	3875
永日小品.txt	外	236
永日小品.txt	漢	4967
永日小品.txt	和	29185
岡本一平著並画『探訪画趣』序.txt	外	6
岡本一平著並画『探訪画趣』序.txt	漢	242
岡本一平著並画『探訪画趣』序.txt	和	911
夏目漱石　評論集.txt	外	19
夏目漱石　評論集.txt	漢	1651
夏目漱石　評論集.txt	和	5621
京に着ける夕.txt	外	8
京に着ける夕.txt	漢	443
京に着ける夕.txt	和	2412

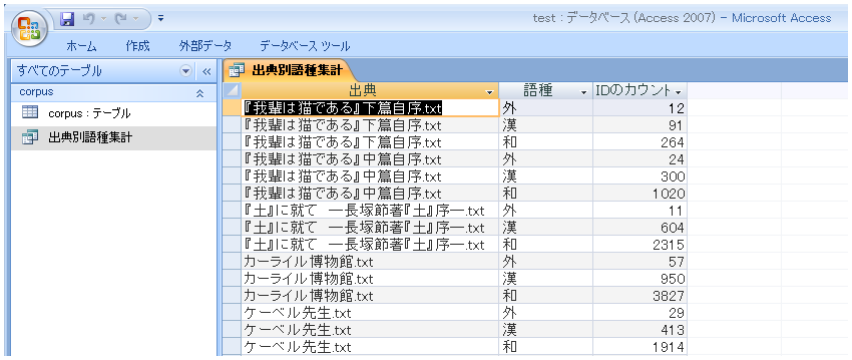
実行結果（データシートビュー）

- 元のデータが数字であれば、合計・平均などを出すこともできる



ピボットグラフビュー（このように Access 上でピボットテーブル・ピボットグラフも作れるが、Excel にコピーしてから処理した方が小回りがきく）

- クエリは名前を付けて保存しておくことができる
- 保存されたクエリは、あたかも新しいテーブル（表）であるかのように利用することができる。



出典	語種	IDのカウント
『我輩は猫である』下篇自序.txt	外	12
『我輩は猫である』下篇自序.txt	漢	91
『我輩は猫である』下篇自序.txt	和	264
『我輩は猫である』中篇自序.txt	外	24
『我輩は猫である』中篇自序.txt	漢	300
『我輩は猫である』中篇自序.txt	和	1020
『土』に就て　一長塚節著『土』序一.txt	外	11
『土』に就て　一長塚節著『土』序一.txt	漢	604
『土』に就て　一長塚節著『土』序一.txt	和	2315
カーライル博物館.txt	外	57
カーライル博物館.txt	漢	950
カーライル博物館.txt	和	3827
ケーベル先生.txt	外	29
ケーベル先生.txt	漢	413
ケーベル先生.txt	和	1914

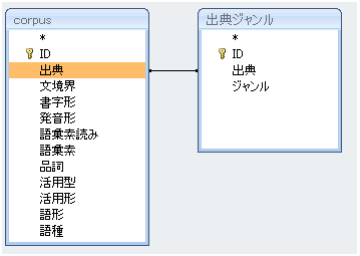
表の結合

出典とジャンルの対照表を作っておく

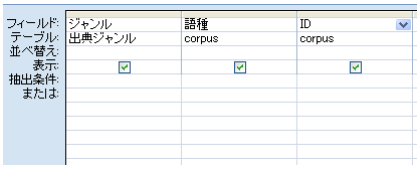
出典ジャンル

ID	出典	ジャンル	新しいフィールドの追加
1	『我輩は猫である』下篇自序.txt	その他	
2	『我輩は猫である』中篇自序.txt	その他	
3	『土』に就て　一長塚節著『土』序一.txt	その他	
4	カーライル博物館.txt	小説	
5	ケーベル先生.txt	小説	
6	ケーベル先生の告別・戦争からきた行き違い.txt	その他	
7	こころ.txt	小説	
8	一夜.txt	小説	
9	永日小品.txt	小説	
10	岡本一平著並画『探訪画趣』序.txt	その他	
11	夏目漱石　評論集.txt	評論	
12	京に着ける夕.txt		
13	教育と文芸・『東洋美術図譜』・イズムの功過.txt	評論	
14	琴のそら音.txt	小説	
15	虞美人草.txt	小説	
16	元日.txt	小説	
17	幻影の盾.txt	小説	
18	現代日本の開化　一明治四十四年八月和歌山において述一.txt	評論	
19	吾輩は猫である.txt	小説	
20	坑夫.txt	小説	
21	行人.txt	小説	
22	高浜虚子著『鶏頭』序.txt	その他	
23	作物の批評.txt	評論	
24	三山居士.txt	小説	
25	三四郎.txt	小説	

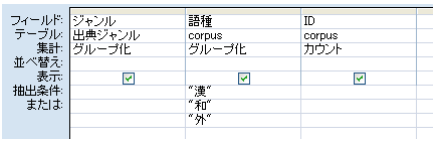
クエリデザインでふたつのテーブルを表示し、同じ内容の列をドラッグアンドドロップで結ぶ



こうすることで複数の表から列を選択できるようになる



これで、ジャンル別の語種集計が可能になる



クエリ1

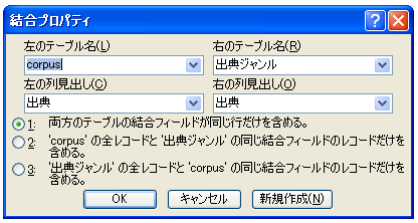
ジャンル	語種	IDのカウント
評論	和	22233
評論	漢	6636
評論	外	120
小説	和	719002
小説	漢	129966
小説	外	3077
その他	和	9135
その他	漢	2597
その他	外	79
	和	936028
	漢	183048
	外	4371

- 内部結合と外部結合

・　内部結合では、対応する出典がない場合、結果に何も表示されない

・　外部結合では、対応する出典がない場合、規準となるテーブルのレコードは全て表示

される



図中の 1 が内部結合、2 が corpus 規準の外部結合、3 が出典ジャンル規準の外部結合

4. 連続する形態素を取り出す・集計する

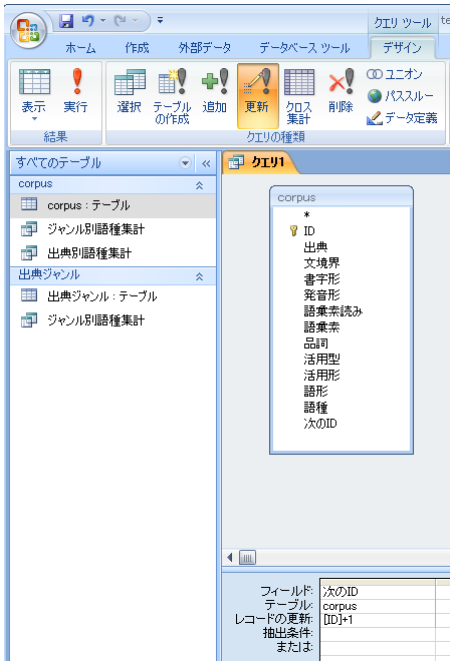
・ 次の形態素の連番を付ける

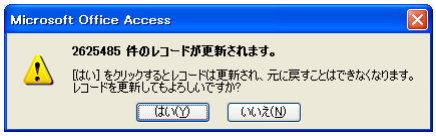
➢ 列を用意する

corpus テーブルをデザインビューで開き「次の ID」列を追加、保存して閉じる

➢ 更新クエリを作成、実行する

「次の ID」列を[ID]+1 に更新する





corpus

ID	出典	文境界	書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種	次のID
1	『土』に就て -B	『				『	補助記号-括弧				記号	2
2	『土』に就て -I	土	ド	ド		土	名詞-普通名詞			ド	漢	3
3	『土』に就て -I	』				』	補助記号-括弧				記号	4
4	『土』に就て -I	に	ニ	ニ		に	助詞-格助詞			ニ	和	5
5	『土』に就て -I	就て	ツイテ	ツイテ		就いて	助詞-副助詞			ツイテ	和	6
6	『土』に就て -B	―				―	補助記号-一般				記号	7
7	『土』に就て -I	―				―	補助記号-一般				記号	8
8	『土』に就て -I	長塚	ナガツカ	ナガツカ		ナガツカ	名詞-固有名詞			ナガツカ	固	9
9	『土』に就て -I	節	セツ	セツ		セツ	名詞-固有名詞			セツ	固	10
10	『土』に就て -I	著	チョ	チョ		著	名詞-普通名詞			チョ	漢	11

・ 連続する形態素を抜き出す

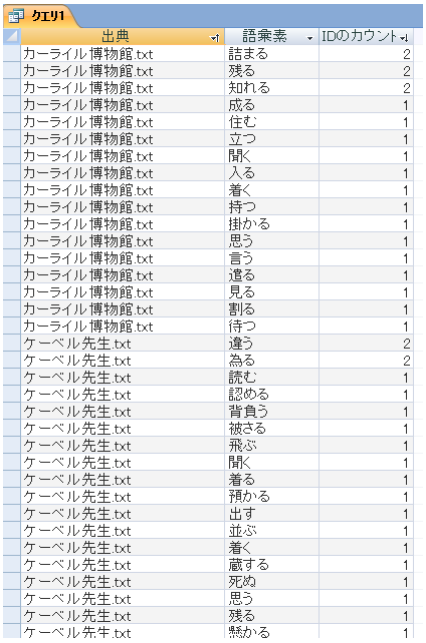
➢ 解析結果テーブル自身をずらして結合する

たくさん結合することもできる

➢ 前後の形態素で条件指定した選択クエリ

動詞＋接続助詞「て」＋「いる」の用例数を出典・動詞別に集計

（抽出条件とするだけで表示しなくてもいい列は「表示」のチェックをはずす）

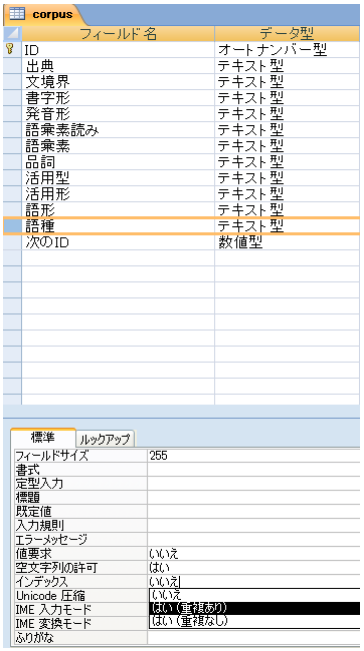


クエリ1

出典	語彙素	IDのカウント
カーライル博物館.txt	詰まる	2
カーライル博物館.txt	残る	2
カーライル博物館.txt	知れる	2
カーライル博物館.txt	成る	1
カーライル博物館.txt	住む	1
カーライル博物館.txt	立つ	1
カーライル博物館.txt	開く	1
カーライル博物館.txt	入る	1
カーライル博物館.txt	着く	1
カーライル博物館.txt	持つ	1
カーライル博物館.txt	掛かる	1
カーライル博物館.txt	思う	1
カーライル博物館.txt	言う	1
カーライル博物館.txt	遣る	1
カーライル博物館.txt	見る	1
カーライル博物館.txt	割る	1
カーライル博物館.txt	待つ	1
ケーベル先生.txt	違う	2
ケーベル先生.txt	為る	2
ケーベル先生.txt	読む	1
ケーベル先生.txt	認める	1
ケーベル先生.txt	背負う	1
ケーベル先生.txt	被さる	1
ケーベル先生.txt	飛ぶ	1
ケーベル先生.txt	聞く	1
ケーベル先生.txt	着る	1
ケーベル先生.txt	預かる	1
ケーベル先生.txt	出す	1
ケーベル先生.txt	並ぶ	1
ケーベル先生.txt	着く	1
ケーベル先生.txt	蔵する	1
ケーベル先生.txt	死ぬ	1
ケーベル先生.txt	思う	1
ケーベル先生.txt	残る	1
ケーベル先生.txt	黙かる	1

その他

・　よく集計に使う列にインデックスを付けることで集計が早くなる



corpus

フィールド名	データ型
ID	オートナンバー型
出典	テキスト型
文境界	テキスト型
書字形	テキスト型
発音形	テキスト型
語彙素読み	テキスト型
語彙素	テキスト型
品詞	テキスト型
活用型	テキスト型
活用形	テキスト型
語形	テキスト型
語種	テキスト型
次のID	数値型

標準　ルックアップ

フィールドサイズ	255
書式	
定型入力	
標題	
既定値	
入力規則	
エラーメッセージ	
値要求	いいえ
空文字列の許可	はい
インデックス	いいえ
Unicode 圧縮	いいえ
IME 入力モード	はい (重複あり)
IME 変換モード	はい (重複なし)
ふりがな	