1日目

2010/1/23-24
名古屋大学

日本語コーパスの最新の検索エンジンと分析

李　在鎬（り　じぇほ）

JAPANFOUNDATION 国際交流基金

全体スケジュール

一日目	前半	後半
時間(目安)	13:00～14:30	14:45～18:00
区分	講義	実習
全体テーマ	コーパス分析の理論と方法	文字列基盤のコーパスツール。コーパスツールとしてのエクセル
主な項目	コーパス基盤の言語研究の特徴、良いコーパス調査のために、コーパス調査の一般的なサイクル、操作のデモ	正規表現を覚えよう、 KWIC Finderを使う。テキスト抽出用の関数を使う、 条件式を使う、ピボットテーブルを使う

二日目	前半	後半
時間(目安)	10:00～12:30	13:30～15:20
区分	講義・実習	実習
全体テーマ	コーパスツールの紹介	コーパスデータ紹介
主な項目	形態素解析、ChaSen, KH Coder, E-Kwicなど	様々なコーパスの紹介

配布データ

- ¥nagoya¥ソフトウェア
  - khcoder-2b22-f.exe /56.4M
  - kwic329.exe /2.3MB
  - hm711_signed.exe /1.8M
- ¥nagoya¥コーパス 終了後は削除
  - 書籍サンプル, 国会議事録サンプル, 白書サンプル, Yahoo知恵袋, 読売新聞サンプル, KYc.zip
- ¥nagoya¥temp
  - SQL.txt, hinshi_chasen

- 2010.1.23-24名古屋大チュートリアル資料.pptx
- http://jhlee.sakura.ne.jp/trans/1.pdf
- ID: misokatsu, PW: daisuki

1／23のお品書き

- 自己紹介
- 前半
  - 全体的な話
  - 理論，実践，方法
- 後半
  - 実習：「漢語+する」の調査
  - Kwic Finderや正規表現

自己紹介

- 李在鎬（りじぇほ）
  - 日本語学（助詞・深層格）→認知言語学→理論言語学→構文研究→コーパス研究・応用言語学
  - 国際交流基金日本語試験センター：出題基準、コーパスに基づく作題支援システム作成
  - 最近の主な関心分野
    1. コーパス基盤の語彙意味論
    2. テキストの難易度抽出、テキスト比較
    3. 構文使用に関する計量言語学的研究
    4. 実験的手法による言語研究
    5. 学習者コーパス

別の説明

- コーパス：辞書と並んで、もっとも基本的かつ中核的な言語資源。辞書は言語の知識に関わる言語資源だが、コーパスは言語が実際に書かれて話されたりして結果が記録されたもの。言語の使用実態に関わる言語資源

言語処理学会(2009)『言語処理学事典』共立出版 pp.58

今日の話では

Before

- 消極的コーパス言語学：PCに適当に単語を入れてみたら用例が出てきた．良いなあ～

After

- **積極的**コーパス言語学：コーパス言語学のロジックと技術を知ろう．自分の研究に応用してみよう！

コーパス言語学の導入 (1/2)

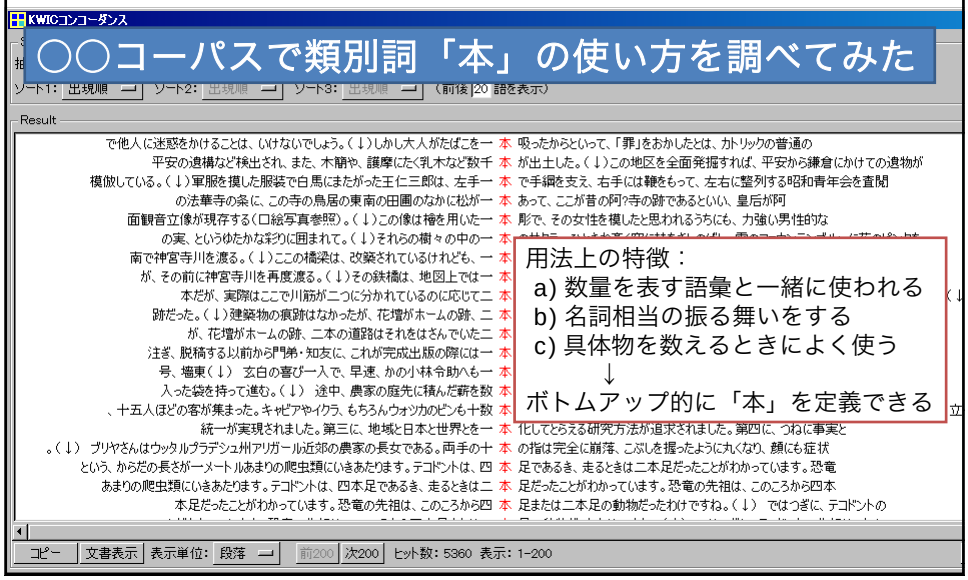
- 考え方(Leech 1992)
  1. 言語能力より言語運用に中心をおくアプローチ
  2. 言語の普遍的特徴の解明より、個別言語の記述に中心をおくアプローチ
  3. 質的な言語モデルのみならず、数量的な言語モデルにも中心をおくアプローチ
  4. 言語研究における合理主義的な立場よりも、経験主義的立場に立つアプローチ

コーパス言語学の導入 (2/2)

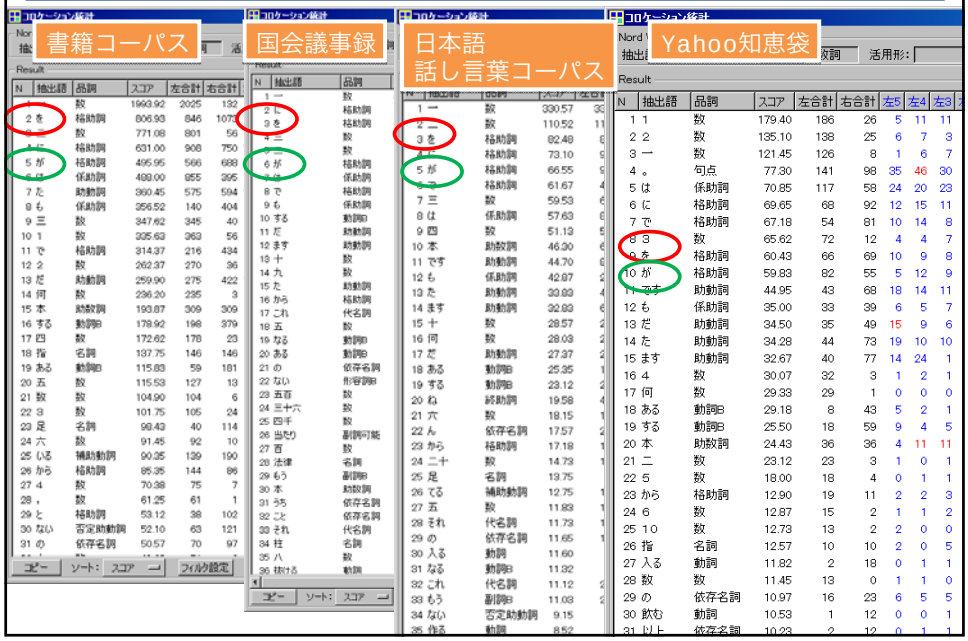
- メリット(Biber, Conrad and Reppen 1998)
  1. 具体的な言語の使用実態をパターン化し、調べることができる。
  2. 網羅的に現象を収集することができ、データの偏りが解消できる。

→ コーパスによる語彙研究の手法は、その生起文脈を客観的に記述することができることから言語研究の単なる手法の域を超え、従来の言語研究では認識されてこなかった問題や現象を発掘し、解明していく新たな言語研究のパラダイムとして確立しつつある(Tognini-Bonelli 2001)。

例えば,



共起する語を調べてみたら，こうだった



暗黙知

1. 近くで表れることは意味がある

→共起関係重要視

例）他動詞の近くには「を」が表れることが多い

→ 他動詞と「を」は関係が深い

一般的にはキーワードの右５、左５を見る

キーワード

小沢氏が土地購入代に充てたとする資金について東京地検

左5　左4 左3　左2 左1　右1　右2 右3　右4　右5

2. よく使われるということには意味がある→頻度重視

コーパス言語学の実践

- 良い調査をするために、調査をデザイン！
  1. どういう目的で、
  2. どのようなコーパスを利用し、
  3. どのようなデータ処理を行うのか、
  4. その結果に対してどのような予測や仮説を持っているか　（皆が頷くような）

  →しっかりしたストーリー構成が重要

コーパス言語学の実践

- 良い調査をするため心掛けるべきこと（伊藤2002）
  1. 妥当性：まさに知ろうとする目的を正しくつく調査方法であること
  2. 信頼性：同じ対象に同じ操作を加えた結果があまり大きく変動するようなものでないこと
  3. 客観性：結果が主観によってまちまちになるような操作ではないこと
  4. 再現性：調べる対象の実際の姿が操作を加えた結果から正しく見極められること
  5. 適応性：その操作が実際に無理なく行え、かつ妥当な結果をもたらすようなものであること

ちょっと横道

- コーパスを使った研究の二分類
  1. コーパス駆動型研究（Corpus-driven Approach）：工学的な言語処理。全面的にコーパスに依拠。コーパスの中ですべての問題解決をはかる、コーパスで始まってコーパスで終わるタイプの研究→億単位のデータ
  2. コーパス基盤型研究（Corpus-based Approach）：人文科学的な研究。研究の補助的手段としてコーパスを利用。疑問に思う事項を検証する立場

コーパス分析の実践

コーパス選択
- どの程度の規模
- どのような内容

データ抽出
- KWIC形式
- 日本語を扱う場合の問題

データ整形
- エクセルの関数
- コーディング

データ分析
- 数える
- 共起スコア

コーパス分析の実践

コーパス選択
- どの程度の規模
- どのような内容

データ抽出
- KWIC形式
- 日本語を扱う場合の問題

データ整形
- エクセルの関数
- コーディング

データ分析
- 数える
- 共起スコア

コーパス言語学の実践

- 代表性の問題
  1. コーパスは、人の言語行動の結果を記録したもの
  2. 分析に用いるコーパスが人の言語行動の実態を正しく反映したもの．例えば、新聞コーパスを利用→新聞は日本語の使用態を代表しうることが前提

  → 「代表性(representativeness)」の問題

コーパス言語学の実践

- 代表性の問題
  1. 代表性の問題がクリアされない研究→違うデータを使えば、調査結果が変わる、結果の信頼性・妥当性は得られない
  2. コーパスを用いた研究のあらゆるステップにおいて問題視される
  3. きちんと答えてないといけない．
     しかし・・・・

困った！！！

どんなに大きなコーパスであれ、どんなに多様なデータを含むコーパスであれ、現実問題として当該言語をそっくり代表しうるコーパスは存在しない。なぜなら、コーパスデータの母集団となる当該言語の使用の総体は潜在的には無限だからである。

代表性の問題をクリアする

- 英語などの一部の言語に関しては（代表性の要請を満足する）均衡コーパスを用意されているので、それを使えば良い。
- 日本語の研究では
  1. 自らがデータを用意する
  2. 自らが下処理を行った上で、サンプルをとる

  →何らかの戦略を立てるべし。

代表性の問題をクリアする

量の面から

- 収集データの量を増やすことで、代表性の要請に答えていこうとする考え方

→　とりあえず沢山！

→　自然言語処理の考え方

質の面から

- 書き言葉や話し言葉を共に含めたり、様々なジャンルのデータをバランスよく取り入れることで代表性の要請に答えていこうという考え方

→ とりあえず色々なもの

→ 人文系のコーパス分析

代表性の問題をクリアする

- 絶対的な基準はない。
- 量に関して
  - 大きければ無条件に良いのか：違う。大きいコーパスを使えば、得られるデータの量も大きくなり、作業に時間がかかる。データの規模が大きくなれば、人為的なミスの可能性も大きくなる。
  - 示唆：5万語程度の小規模のコーパスで助詞「に」の用法を収集し、クラスタ分析を行ったが、そこで得た結果は、後に行った別の調査で、100万語のコーパスで調査した内容と大きく違わない．李(2003)

→　必要最適な量を見極めることが重要。数百程度

大規模でなければ
できないものもある

	ウェブコーパス	日英新聞記事対応付けデータ
恐れる	1010	11
積む	1362	4
冷やす	664	3
寄与する	582	9
固まる	473	3
散る	289	1
心掛ける	178	0
賛成する	185	12
逃げる	388	0
逃れる	504	1
躍進する	70	2
合計	5705	46

黒田・李(2008)

質も大事

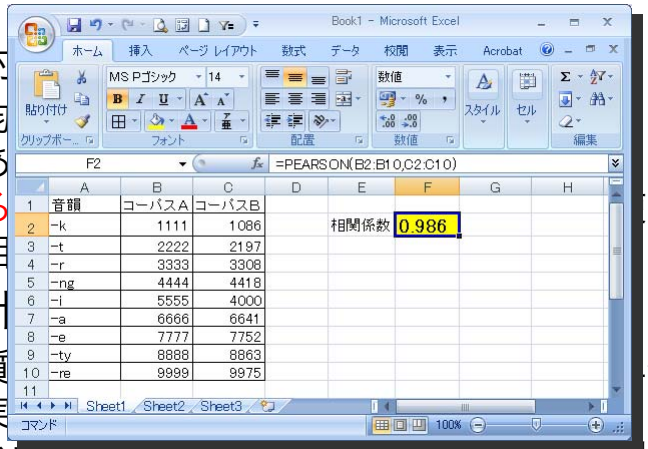
(1) a. 地震に驚いた人々が部屋から廊下に出た。 →位置変化
b. 太郎が会議に出た。 →出現
c. 我が校から優勝者が出た。 →発生
d. 太郎は住み慣れた街を出た。 →離脱

李（2008）

		文意				合計
		位置変化	出現	発生	離脱	
コーパス	新潮新書	45	30	17	16	108
	読売新聞	14	79	391	5	489
合計		59	109	408	21	597

代表性の問題をクリアする

- 絶対
  - 均
    あ
    る
    相
- 統計
  - 質
    実



	A	B	C
1	音韻	コーパスA	コーパスB
2	-k	1111	1086
3	-t	2222	2197
4	-r	3333	3308
5	-ng	4444	4418
6	-i	5555	4000
7	-a	6666	6641
8	-e	7777	7752
9	-ty	8888	8863
10	-re	9999	9975
11			

ので
ている
用し、
い。
位の
。高
い相関が得られた→どのコーパスから調べて
も結果は変わらない（李・玉岡・林2008）

コーパス分析の実践

コーパス選択
- どの程度の規模
- どのような内容

データ抽出
- KWIC形式
- 日本語を扱う場合の問題

データ整形
- エクセルの関数
- コーディング

データ分析
- 数える
- 共起スコア

KWICって？

- KWIC形式
  - 「クイック」と読む.
  - **K**ey**W**ord **I**n **C**ontext
    → 「文脈の中の鍵語」のこと

KWICの見本

先行文脈	キーワード（鍵語）	後続文脈
に21年に準急として復活して江ノ島線にも	走る	ようになったが、
ンスカーによる臨時特急（特急料金不要）も	走る	ようになった。
区間は3線線路になっており、上下いずれも	走る	ようにして、
街なかをしばらく	走る	と片瀬江ノ島となる。
小田原線とほぼ同様な車両が	走る	。
車エクセは10両編成を分割した4両編成が	走る	。
特急「えのしま」、急行、各停の３種が	走る	。
庫した4両を増結して新宿まで10両編成で	走る	。

後ほど実習を！

日本語KWIC問題

1. 異表記の問題

2. 異形の問題

3. 複数の要素にまたがっている場合の問題

4. 一部が省略されている場合の問題

日本語KWIC問題

1. 異表記の問題
   ⇒ 友達，友だち，ともだち
2. 異形の問題

3. 複数の要素にまたがっている場合の問題

4. 一部が省略されている場合の問題

日本語KWIC問題

1. 異表記の問題

2. 異形の問題
   ⇒ 走らない，走ろう，走った，走り，・・・.
3. 複数の要素にまたがっている場合の問題

4. 一部が省略されている場合の問題

日本語KWIC問題

1. 異表記の問題

2. 異形の問題

3. 複数の要素にまたがっている場合の問題
   ⇒ 日本，本を読む，三本
4. 一部が省略されている場合の問題

日本語KWIC問題

1. 異表記の問題

2. 異形の問題

3. 複数の要素にまたがっている場合の問題

4. 一部が省略されている場合の問題
   ⇒助詞や主語など，ハ・モとガの問題など

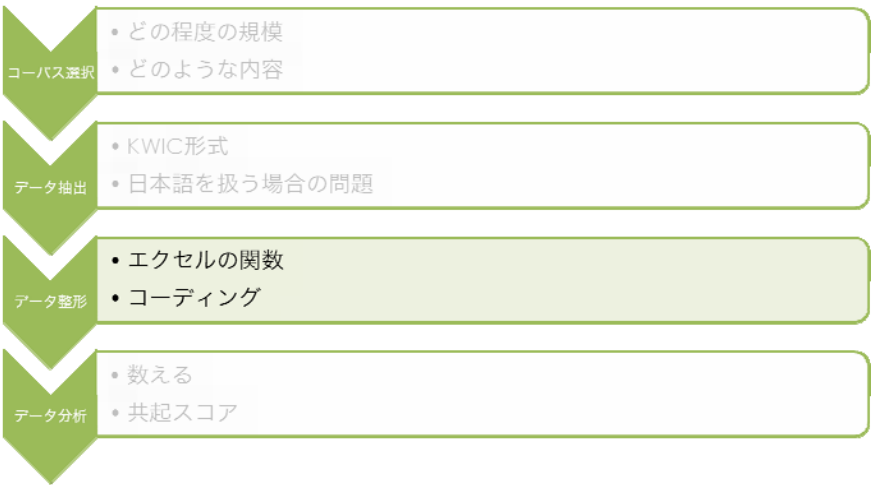
日本語KWIC問題

1. 異表記の問題
2. 異形の問題
3. 複数の要素にまたがっている場合の問題
4. 一部が省略されている場合の問題

形態素解析

形態素解析
の
実演

コーパス分析の実践



意外とすごいEXCEL

- **Excelと言えば,**
  1. 表組にしてくれる
  2. グラフを書いてくれる
  3. 決まった順番に並び変えてくれる
  4. 足し算や引き算などの計算ができる
  5. 文字数を数えてくれる

     ・・・・・・・

- それだけじゃないぞ！

意外とすごいEXCEL

- 文字列を切り取ってくれる．⇒探索の手助けをしてくれる
  - 太郎が学校に行った.
  - 太郎の自動車が壊れた.
  - 太郎の家のトイレがきれいだ.

  ⇒主語だけ切り取ってデータを見やすくしたい.

	A	B	C
		が格を取り出す	被修飾名詞を取り出す
1	太郎が通った学校。	**=LEFT(A1,FIND("が",A1))**	**=MID(A1,FIND("た",A1)+1,2)**

	A	B	C
		が格を取り出す	被修飾名詞を取り出す
1	太郎が通った学校。	太郎	学校

色々できる；
高度なフィルターリングをする，ピボットテーブルなどで，集計する．
詳しくはのちほどの実習で

コーディング

- テキストデータをより構造化し、ターゲットとなる言語現象を動的に理解し、把握するためのもの

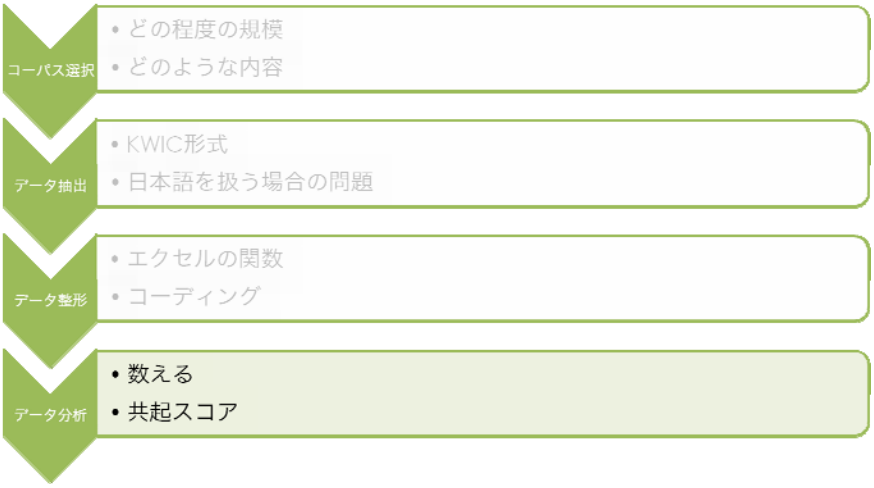
言語表現	変数１	変数２	変数３	変数４	・・・
・・・					
・・・・					
・・・・・・・・					
・・・					

言語表現	主語	主語の特徴	動詞の語形	助詞	・・・
太郎が走る	太郎	人	る形	が	
花子も走っている	花子	人	ている形	も	
激痛が走った	激痛	抽象物	た形	が	
犬も走っている	犬	人以外の生物	ている形	も	

コーディング

- コーディングにおけるポイント
  1. どういう変数を立てるか
     - 調査全体に成否に関わる要因
     - 選択に関して責任を持つこと
     - 伊藤2002が言う「信頼性、再現性の高いもの」
  2. どう判断するか
     - 外部言語資源を利用する
     - 複数の被験者に判断させる

コーパス分析の実践



コーパス研究における分析

- 分析の基本
  - 数える
  - 多かった，高かった

  ⇒　きっと理由がある　⇒　考察する
- 数える際の鉄則
  - 手で数えない。必ずミスる、不必要に時間がかかる
  - ソフトを使う。エクセル
    - 数えるための関数やピボットテーブル
    - 詳細は後半で。

コーパス研究における分析

- 色々な数え方
  - 語の数える；短単位か長単位か，延べか異なりか
  - 共起を数える；ｔスコア，MIスコア，エントロピなどなど→目的に合わせて使う。

→ 玉岡先生に！

論文化する

- 結果を報告する際、留意すべき点
  - 調査の手順や各ステップにおける結果を可能な限り、クリアに報告する
  - 同じ解析方法を用いて調査を行えば、概ね同じ結果が再現できることが期待される
  - どのようなコーパスに、どのような下処理を行ったのか、どのようなデータ抽出方法を用いたのか

目次の見本

1. 導入：なぜその研究が行われなければならないのか
2. 先行研究と問題提起：その研究が答えようとしている問いは何なのか
3. データと調査方法：追試可能な形で調査の手順などを明記する
   1. 使用データ（コーパス）
   2. 解析方法
4. 結果：得られた結果を客観的事実として報告する
5. 考察：「結果」から「問題提起」に対してどのような答えが得られたかを述べる
6. まとめと課題

- お疲れ様でした。
- 休み時間に¥nagoya¥ソフトウェア
  - kwic329.exe /2.3MB
  - hm711_signed.exe /1.8M

後半

- ソフトウェアのインストール＆確認
  - コンコーダンスソフト：Kwic Finder
  - テキストエディタ：秀丸エディタ

- 秀丸エディタ
  - 機能・目的：テキストファイルを開くためのもの。超多機能の編集機！！！。
  - 入手：ttp://hide.maruo.co.jp/software/hidemaru.html
  - インストール：http://vision.ameba.jp/watch.do?movie=610709

- Kwic Finder
  - 機能・目的：キーワードと一致した文字列を集めるためのもの
  - 入手：http://www31.ocn.ne.jp/~h_ishida/KWIC.html

『現代日本語書き言葉均衡コーパス』(BCCWJ)

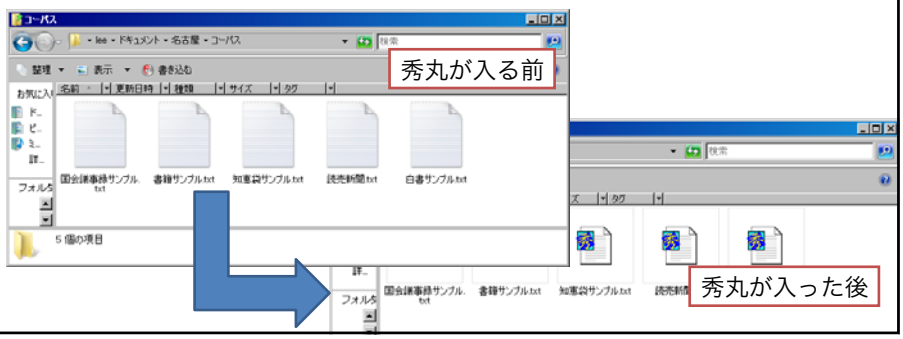
- 書き言葉の均衡コーパス。特定領域研究
- 書籍、国会議事録、白書、Yahoo知恵袋のデータがバランスよく入っている
- チュートリアルのため少し加工した
  - ９０００字～１００００字の間で字数を統一
  - 改行コードを調整。一文に一改行
  - テキストＩＤやタグを削除

- コーパスの確認
  1. BCCWJから抽出
  2. 一文一行に整形
  3. 空行の削除
  4. 特殊記号の削除

  など

ちょっと横道

- どうしてもメモ帳ファイルなのか。
  - 正確にはテキストファイル（.txt）
    - マシン環境を選ばない、Win, Mac, Unix…
    - 編集・検索が楽、処理が速い→何かと便利。



サンプル調査

- 漢語＋するの共起に関するコーパス調査
  - 勉強する、変形する
- 言語研究的観点から
  - 一つの形式でありながら多様な性質を持つ
    - 自他の混在、非対格性・意図性の有無、助詞の介在(勉強する・勉強をする/変形する・*変形をする）など（影山1996）

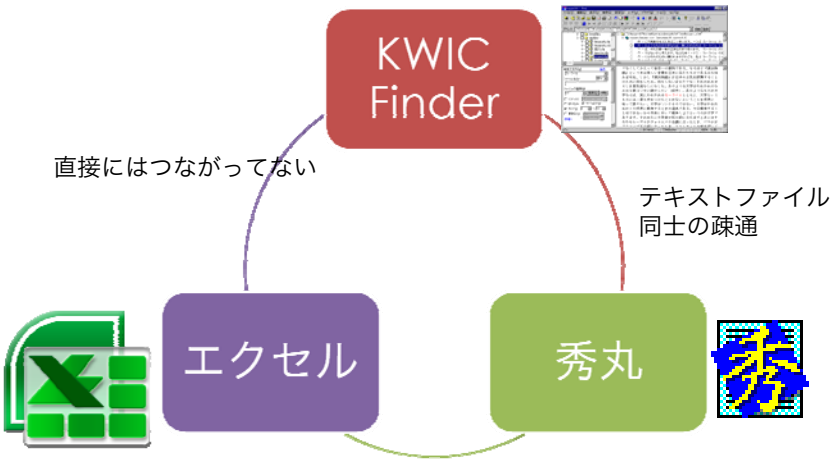
サンプル調査

- 概要
  - 「二文字の漢字＋する」でなる用例をコーパスベースにみる。
  - コーパス間で分布の違いが見られるかどうか
- 使うソフトウェア
  - Kwic Finder, 秀丸, エクセル
- 使う技術
  - 正規表現　＝＞データ収集
  - エクセルのテキスト関数　＝＞データ整形
  - ピボットテーブル　＝＞データ集計

・・お詫び・・

- エクセルの操作を覚えてもらうため、過剰に関数を使っている部分がある。
- 練習だと思ってやってみてください。
- Kwic finderのタスクも操作を覚えてもらうためのもので、検索対象によっては別の操作を加えたほうが良い場合もある

操作イメージ



KWIC Finder とは

- KWIC Finder
  - もともとは複数テキストからの文字列検索ソフト
  - 検索結果に **KWIC 形式**を採用
  - **正規表現**を使用した検索に対応
  - テキストファイルの検索のみであれば，フリー：http://www31.ocn.ne.jp/~h_ishida/KWIC.html

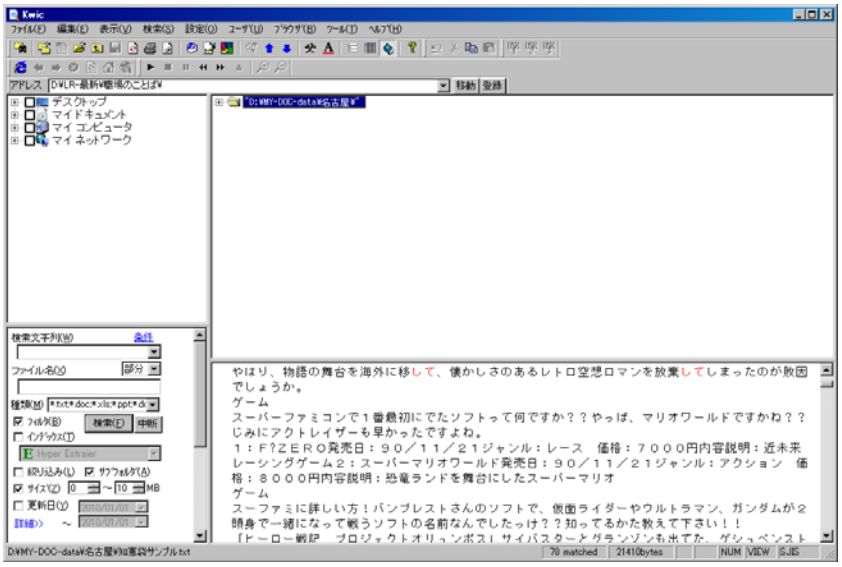
正規表現とは

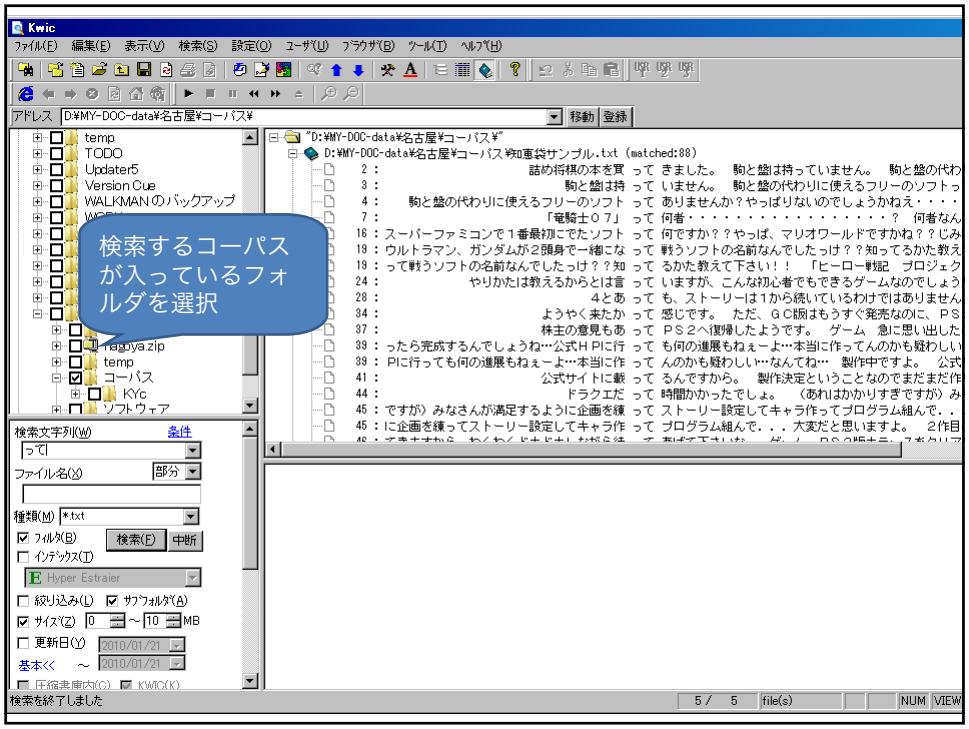
- せいきひょうげん、regular expression）とは，文字列の集合を一つの文字列で表現する方法の一つである。
- ある要素（文字とか単語など）の前後に特定のパターンを持ったものが連接していく形の記述
  - 戻って　買って　貼って　いって
  - → [戻買]って
  - → .って

正規表現の記号

項目	意味
[ ]	キャラクタクラス
( )	パターンをグループ化する
^	行頭
$	行末
.	改行を除く任意の１文字
*	直前のパターンの０回以上の繰り返し
＋	直前のパターンの１回以上の繰り返し
?	直前のパターンが０回または１回現われる
|	パターンの論理和
￥	エスケープ文字

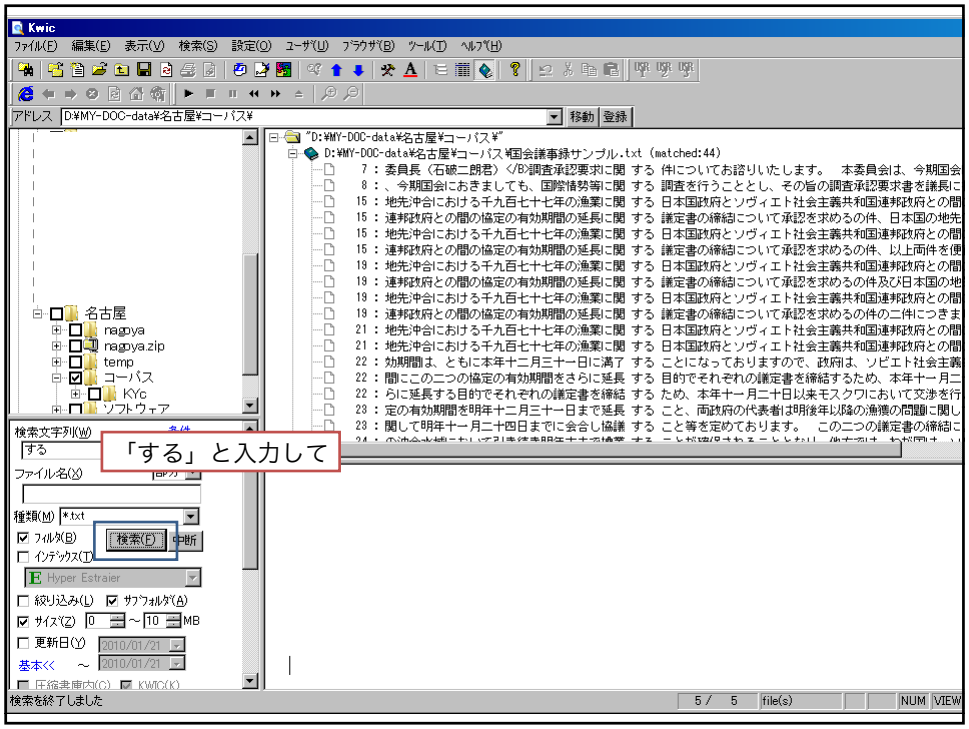
KWIC Finder の画面

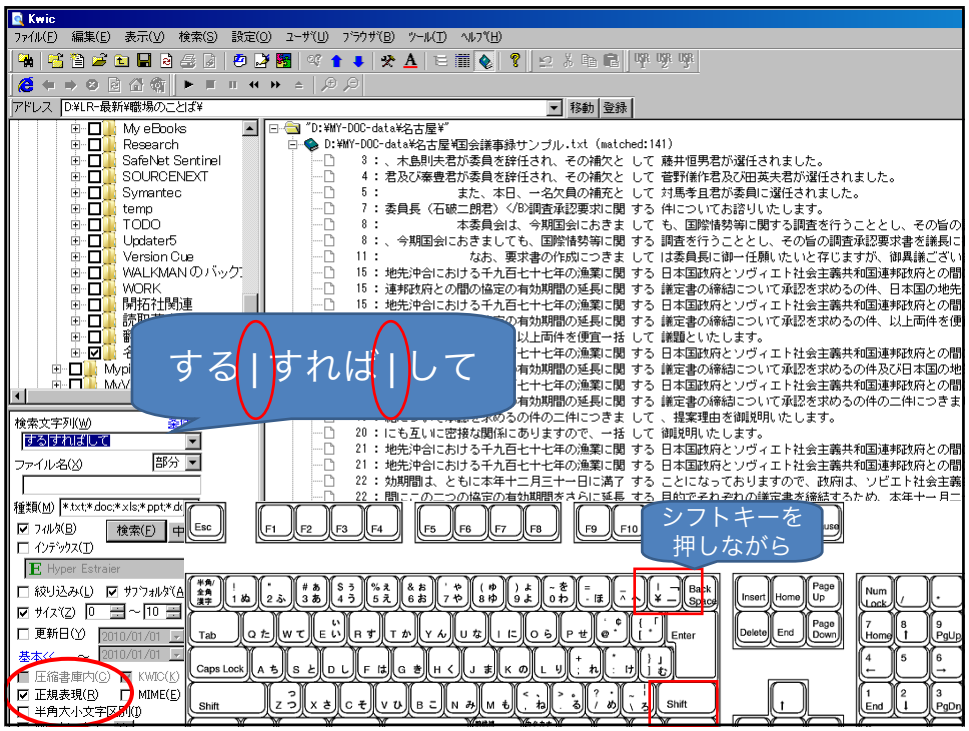


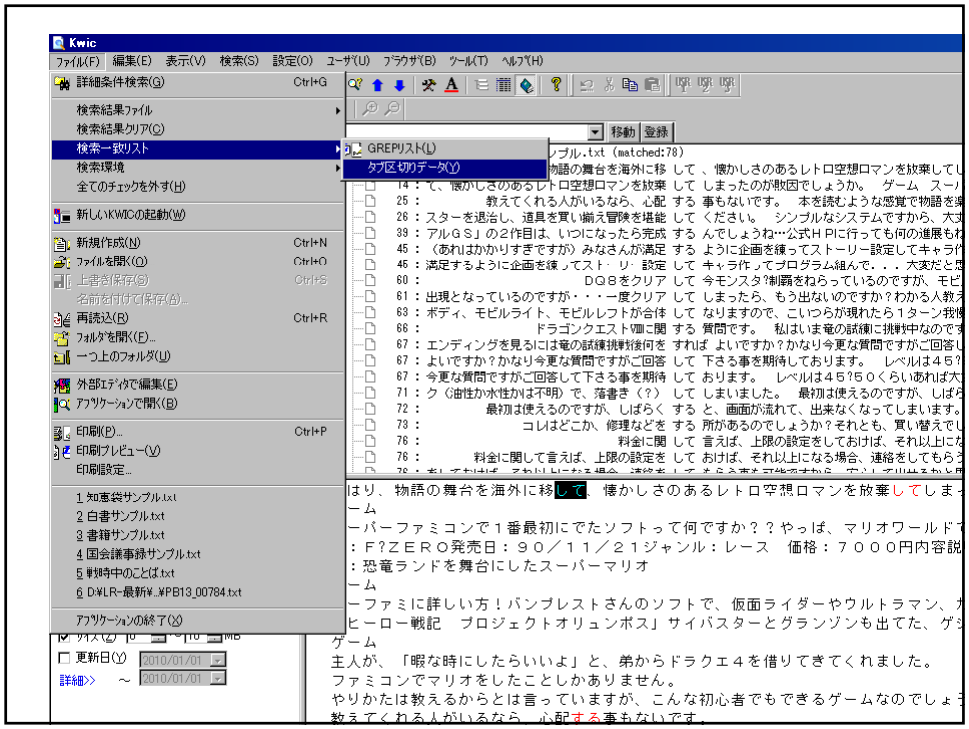


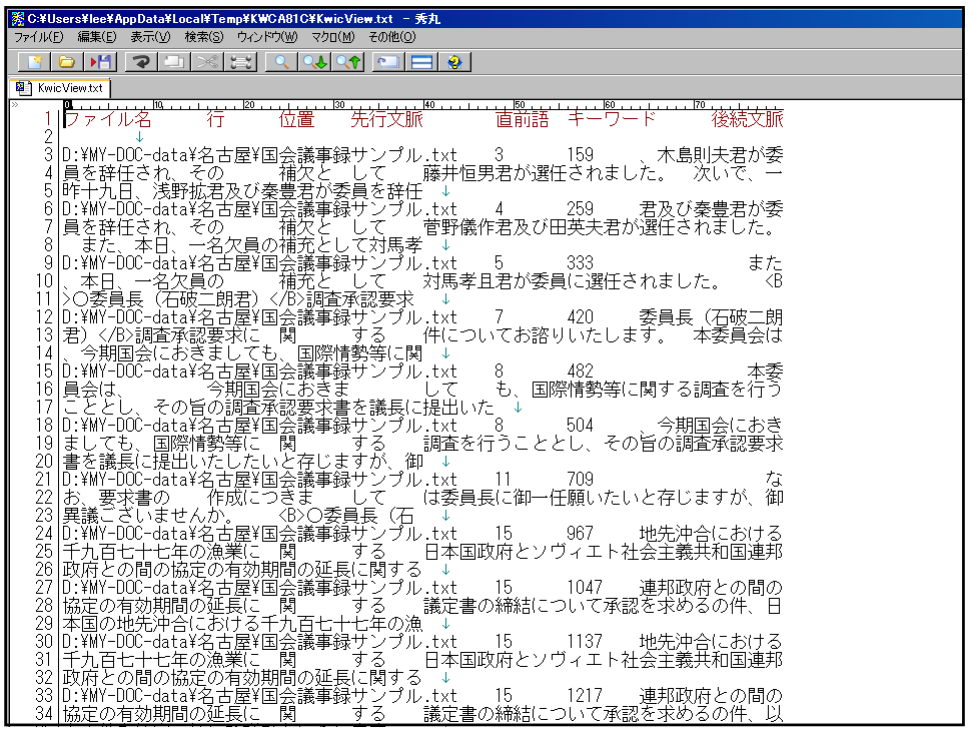
正規表現とは

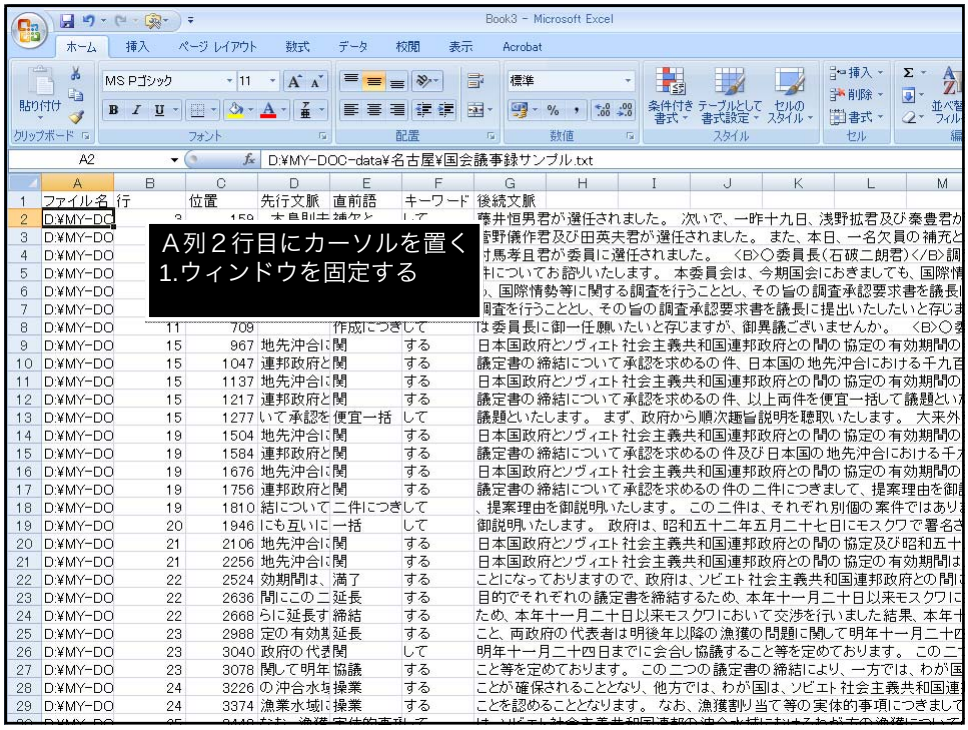
- regular expression
  列で表現する方法の一つである。
- ある要素（文字とか単語など）の前後に特定のパターンを持ったものが連接していく形の記述
  - 戻って　買って　貼って　いって
  - → [戻買]って
  - → .って

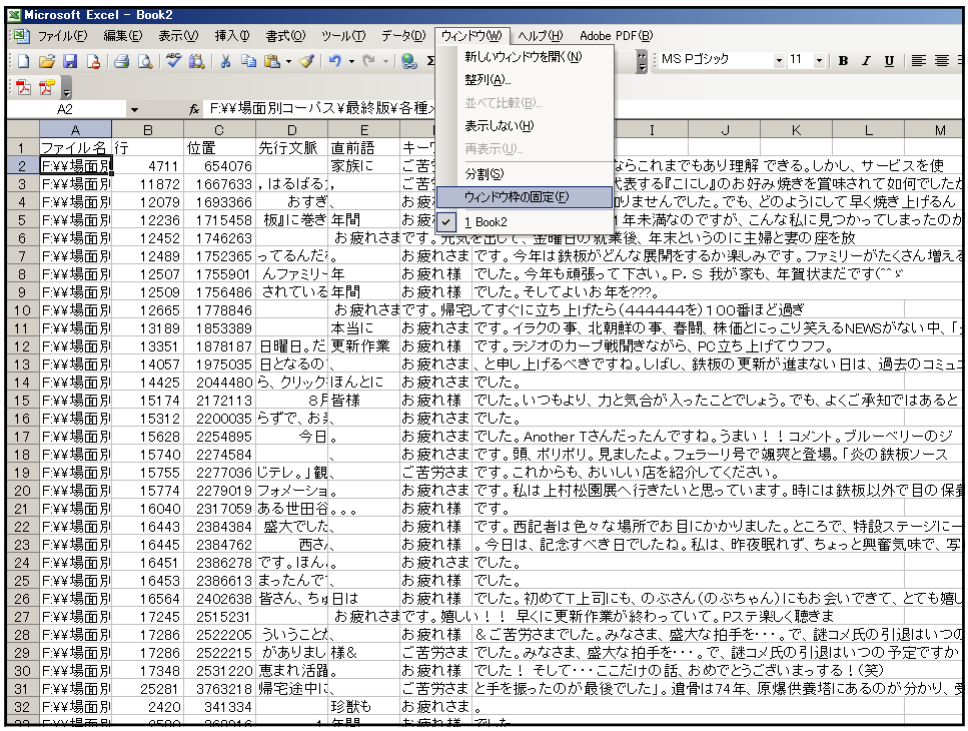



する | すれば | して
シフトキーを
押しながら

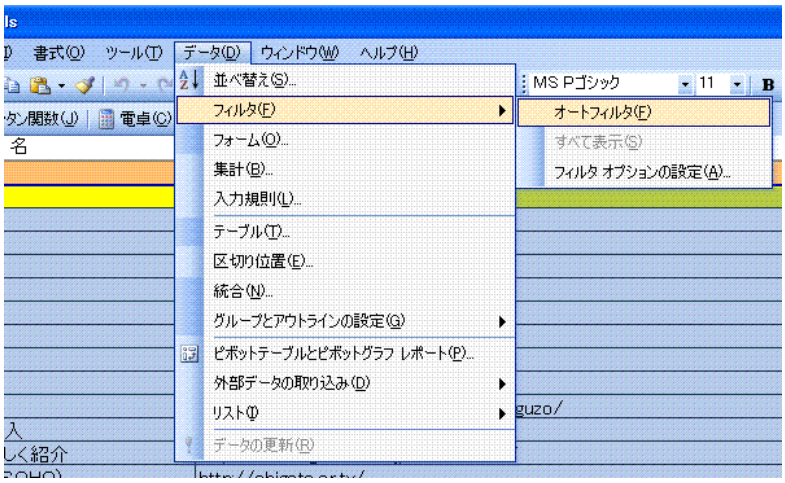


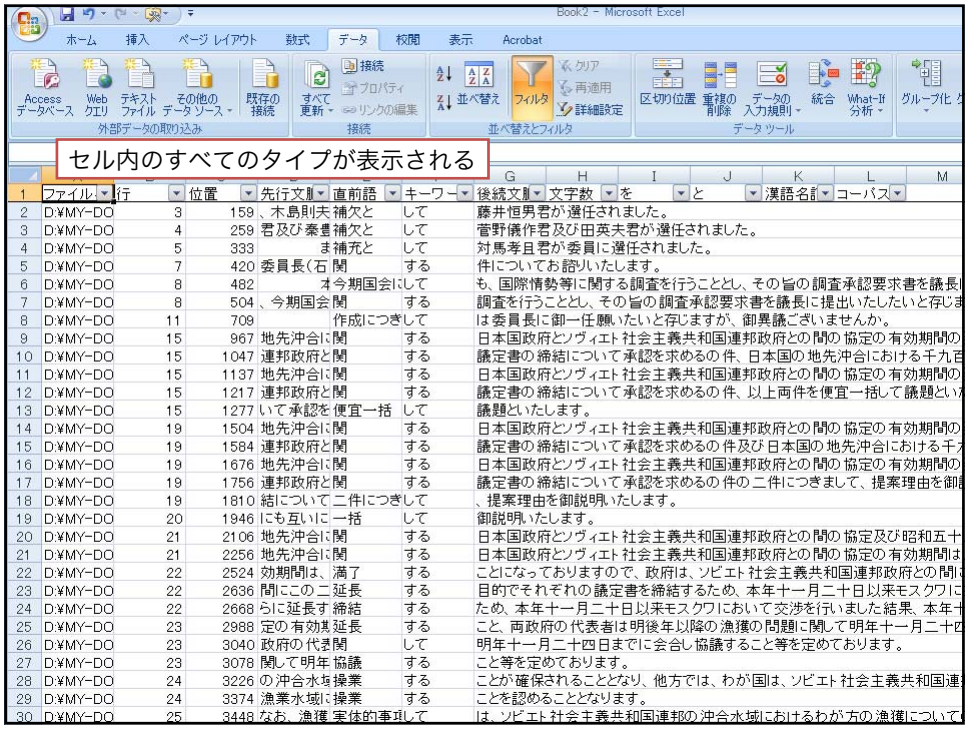

C:¥Users¥lee¥AppData¥Local¥Temp¥KWCA81C¥KwicView.txt - 秀丸
ファイル(F) 編集(E) 表示(V) 検索(S) ウィンドウ(W) マクロ(M) その他(O)
KwicView.txt
ファイル名 行 位置 先行文脈 直前語 キーワード 後続文脈
D:¥MY-DOC-data¥名古屋¥国会議事録サンプル.txt 3 159 、木島則夫君が委員を辞任され、その 補欠として 藤井恒男君が選任されました。 次いで、一昨十九日、浅野拡君及び秦豊君が委員を辞任
D:¥MY-DOC-data¥名古屋¥国会議事録サンプル.txt 4 259 君及び秦豊君が委員を辞任され、その 補欠として 菅野儀作君及び田英夫君が選任されました。 また、本日、一名欠員の補充として対馬孝
D:¥MY-DOC-data¥名古屋¥国会議事録サンプル.txt 5 333 また、本日、一名欠員の 補充として 対馬孝且君が委員に選任されました。 <B>○委員長（石破二朗君）</B>調査承認要求
D:¥MY-DOC-data¥名古屋¥国会議事録サンプル.txt 7 420 委員長（石破二朗君）</B>調査承認要求に 関する 件についてお諮りいたします。 本委員会は、今期国会におきましても、国際情勢等に関
D:¥MY-DOC-data¥名古屋¥国会議事録サンプル.txt 8 482 本委員会は、 今期国会におきまして も、国際情勢等に関する調査を行うこととし、その旨の調査承認要求書を議長に提出いた
D:¥MY-DOC-data¥名古屋¥国会議事録サンプル.txt 8 504 今期国会におきましても、国際情勢等に 関する 調査を行うこととし、その旨の調査承認要求書を議長に提出いたしたいと存じますが、御
D:¥MY-DOC-data¥名古屋¥国会議事録サンプル.txt 11 709 なお、要求書の 作成につきま して は委員長に御一任願いたいと存じますが、御異議ございませんか。 <B>○委員長（石
D:¥MY-DOC-data¥名古屋¥国会議事録サンプル.txt 15 967 地先沖合における千九百七十七年の漁業に 関する 日本国政府とソヴィエト社会主義共和国連邦政府との間の協定の有効期間の延長に関する
D:¥MY-DOC-data¥名古屋¥国会議事録サンプル.txt 15 1047 連邦政府との間の協定の有効期間の延長に 関する 議定書の締結について承認を求めるの件、日本国の地先沖合における千九百七十七年の漁
D:¥MY-DOC-data¥名古屋¥国会議事録サンプル.txt 15 1137 地先沖合における千九百七十七年の漁業に 関する 日本国政府とソヴィエト社会主義共和国連邦政府との間の協定の有効期間の延長に関する
D:¥MY-DOC-data¥名古屋¥国会議事録サンプル.txt 15 1217 連邦政府との間の協定の有効期間の延長に 関する 議定書の締結について承認を求めるの件、以


Book3 - Microsoft Excel
ホーム 挿入 ページ レイアウト 数式 データ 校閲 表示 Acrobat
A2 D:¥MY-DOC-data¥名古屋¥国会議事録サンプル.txt
A列２行目にカーソルを置く
1.ウィンドウを固定する



- 入力する
  - Ｈ１セル：文字数
  - Ｉ１セル：を
  - Ｊ１セル：と
  - Ｋ１セル：漢語名詞
  - Ｌ１セル：コーパス
- フィルターをクリック（データ→フィルター）





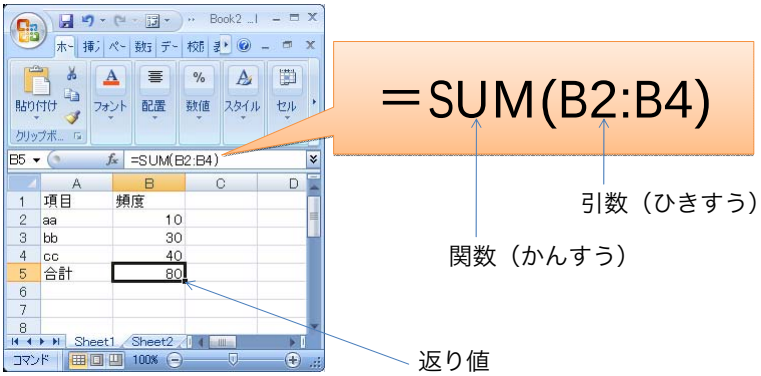
	ファイル	行	位置	先行文脈	直前語	キーワー	後続文脈	文字数	を	と	漢語名詞	コーパス
2	D:¥MY-DO	3	159	、木島則夫	補欠と	して	藤井恒男君が選任されました。					
3	D:¥MY-DO	4	259	君及び秦豊	補欠と	して	菅野儀作君及び田英夫君が選任されました。					
4	D:¥MY-DO	5	333	ま	補充と	して	対馬孝且君が委員に選任されました。					
5	D:¥MY-DO	7	420	委員長（石	関	する	件についてお諮りいたします。					
6	D:¥MY-DO	8	482	オ	今期国会に	して	も、国際情勢等に関する調査を行うこととし、その旨の調査承認要求書を議長					
7	D:¥MY-DO	8	504	、今期国会	関	する	調査を行うこととし、その旨の調査承認要求書を議長に提出いたしたいと存じま					
8	D:¥MY-DO	11	709		作成につき	して	は委員長に御一任願いたいと存じますが、御異議ございませんか。					
9	D:¥MY-DO	15	967	地先沖合に	関	する	日本国政府とソヴィエト社会主義共和国連邦政府との間の協定の有効期間の					
10	D:¥MY-DO	15	1047	連邦政府と	関	する	議定書の締結について承認を求めるの件、日本国の地先沖合における千九百					
11	D:¥MY-DO	15	1137	地先沖合に	関	する	日本国政府とソヴィエト社会主義共和国連邦政府との間の協定の有効期間の					
12	D:¥MY-DO	15	1217	連邦政府と	関	する	議定書の締結について承認を求めるの件、以上両件を便宜一括して議題とい					
13	D:¥MY-DO	15	1277	いて承認を	便宜一括	して	議題といたします。					
14	D:¥MY-DO	19	1504	地先沖合に	関	する	日本国政府とソヴィエト社会主義共和国連邦政府との間の協定の有効期間の					
15	D:¥MY-DO	19	1584	連邦政府と	関	する	議定書の締結について承認を求めるの件及び日本国の地先沖合における千					
16	D:¥MY-DO	19	1676	地先沖合に	関	する	日本国政府とソヴィエト社会主義共和国連邦政府との間の協定の有効期間の					
17	D:¥MY-DO	19	1756	連邦政府と	関	する	議定書の締結について承認を求めるの件の二件につきまして、提案理由を御					
18	D:¥MY-DO	19	1810	結について	二件につき	して	、提案理由を御説明いたします。					
19	D:¥MY-DO	20	1946	にも互いに	一括	して	御説明いたします。					
20	D:¥MY-DO	21	2106	地先沖合に	関	する	日本国政府とソヴィエト社会主義共和国連邦政府との間の協定及び昭和五十					
21	D:¥MY-DO	21	2256	地先沖合に	関	する	日本国政府とソヴィエト社会主義共和国連邦政府との間の協定の有効期間は					
22	D:¥MY-DO	22	2524	効期間は、	満了	する	ことになっておりますので、政府は、ソビエト社会主義共和国連邦政府との間					
23	D:¥MY-DO	22	2636	間にこの二	延長	する	目的でそれぞれの議定書を締結するため、本年十一月二十日以来モスクワに					
24	D:¥MY-DO	22	2668	らに延長す	締結	する	ため、本年十一月二十日以来モスクワにおいて交渉を行いました結果、本年十					
25	D:¥MY-DO	23	2988	定の有効期	延長	する	こと、両政府の代表者は明後年以降の漁獲の問題に関して明年十一月二十四					
26	D:¥MY-DO	23	3040	政府の代表	関	して	明年十一月二十四日までに会合し協議すること等を定めております。					
27	D:¥MY-DO	23	3078	関して明年	協議	する	こと等を定めております。					
28	D:¥MY-DO	24	3226	の沖合水域	操業	する	ことが確保されることとなり、他方では、わが国は、ソビエト社会主義共和国連					
29	D:¥MY-DO	24	3374	漁業水域に	操業	する	ことを認めることとなります。					
30	D:¥MY-DO	25	3448	なお、漁獲	実体的事項	して	は、ソビエト社会主義共和国連邦の沖合水域におけるわが方の漁獲について					

データ整形

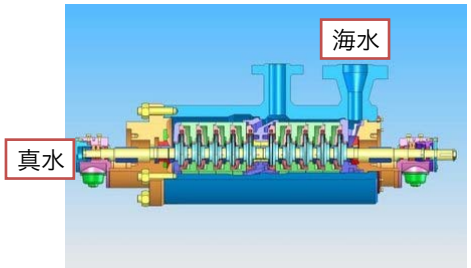
- 目的
  - 見やすくする、数えやすくする
  - 統計処理(SPSS,JMPなど)にもかけやすくする
- 技術
  - テキスト関数：={find, len, left, mid}
    - 目で見て作業可能なデータ(?数百程度)では手でやるのが良いが、
    - 千や万単位のデータを扱う際には強力
    - データの特徴を素早く把握するためには不可欠
    - 人為的なミスを防げる

用語について

- 関数＝いろいろな機能を実現してくれる、便利な命令文



エクセル関数とは

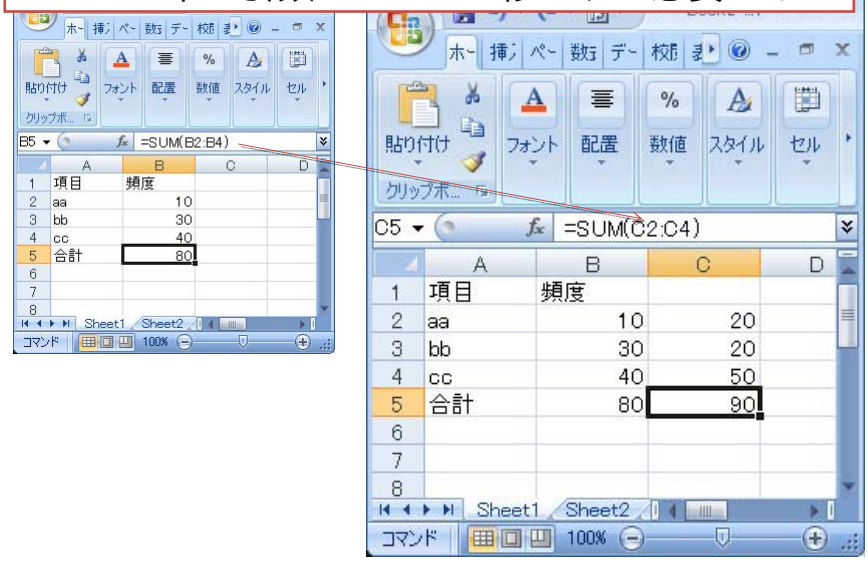


海水を真水に変える装置

海水を投入すると、真水が得られます

この装置に投入する**海水**が**引数（ひきすう）**にあたります。
海水を真水にする装置が**関数**です。
その結果出てきた**真水**が**返り値**です。

コピペで中の引数をいちいち修正する必要がない



エクセル関数とは

- 値同士の関係を命令(指定)するための書式

＝関数（引数）

- 共通記号
  1. ＝　　関数であることの指定
  2. ＄　　セルの固定する
  3. ＆　　セルを結ぶ
  4. “ “　　文字列であることの印

- H2のセルに
  「＝b2+c2」と入力してみよう

  次に、
  「＝b2＆c2」と入力してみよう

- H2の内容をH3, H4にコピペしてみよう
- H2を 2，2，3，3にコピペ

- ＄記号を入れてみよう
  ＝C2 → =$C2, =C$2, =C2

  コピペしても、中の数字が変わらない！

- “ ”記号を入れてみよう
  K2のセルに
  ＝”aaaa”&C2

- ここから本番です

関数の説明

- ファインド
  =FIND(検索文字列,対象,開始位置)：文字列が開始位置から何番目にあるかを返します
- レン
  =LEN(文字列）：文字列の文字数を返します。半角文字と全角文字との区別はされません

関数の説明

- レフト
  =LEFT(文字列,[*文字数*])：文字列の左端から指定した文字数だけ取り出す。
- ミッド
  =MID (文字列,　開始位置,文字数)：文字列の指定した位置から指定した文字数だけ取り出す。
- ライト
  =RIGHT(文字列,[*文字数*]）：文字列の右端から指定した文字数だけ取り出す。

入れてみよう

- H2セルに **=len(e2)**と入力
- I2セルに **=find(“を“,e2)**と入力
- j2セルに **=find(“と“,e2)**と入力

- #VALUE!：関数が正しくない

- あとはコピペ

- Find(i2,e2), Find(j2,e2) → i$2, j$2

- フィルターを活用しよう
  - 文字数「2」
  - 「と」が「#VALUE!」

- 漢語名詞に
  ＝e19

あとはコピペ。。。。

- フィルターを活用しよう
  - 文字数「3」
  - 「を」が「3」

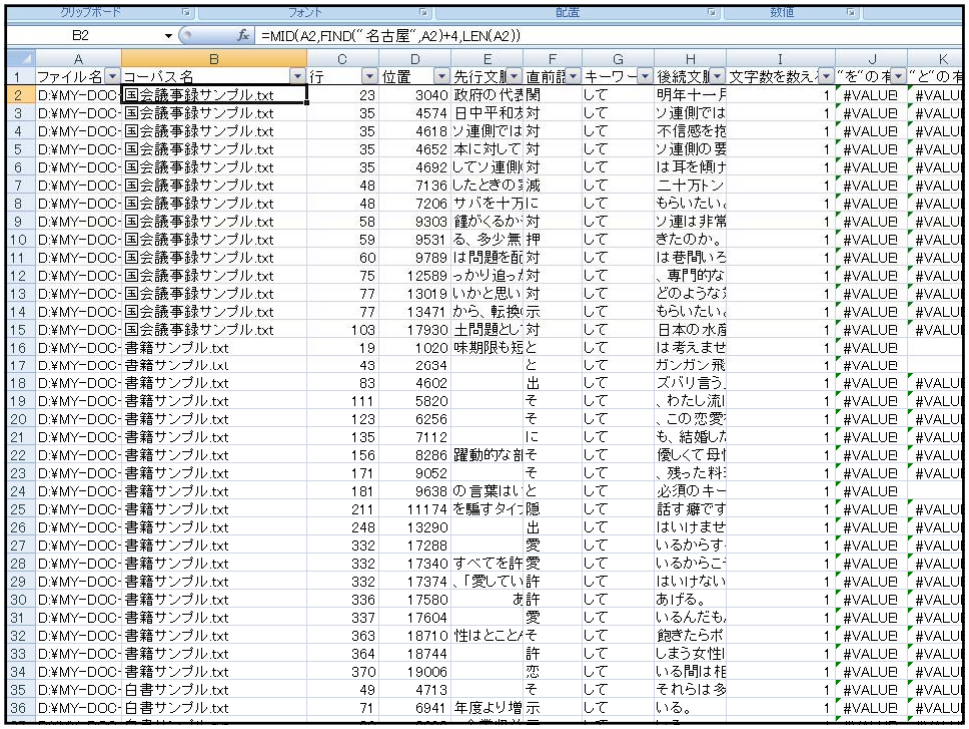
- 漢語名詞　　K53
  =LEFT(E53,2)

あとはコピペ。。。。

入れてみよう

- Ｌ２に

=MID(A2,FIND("コーパス",A2)+5,LEN(A2))



B2 =MID(A2,FIND("名古屋",A2)+4,LEN(A2))

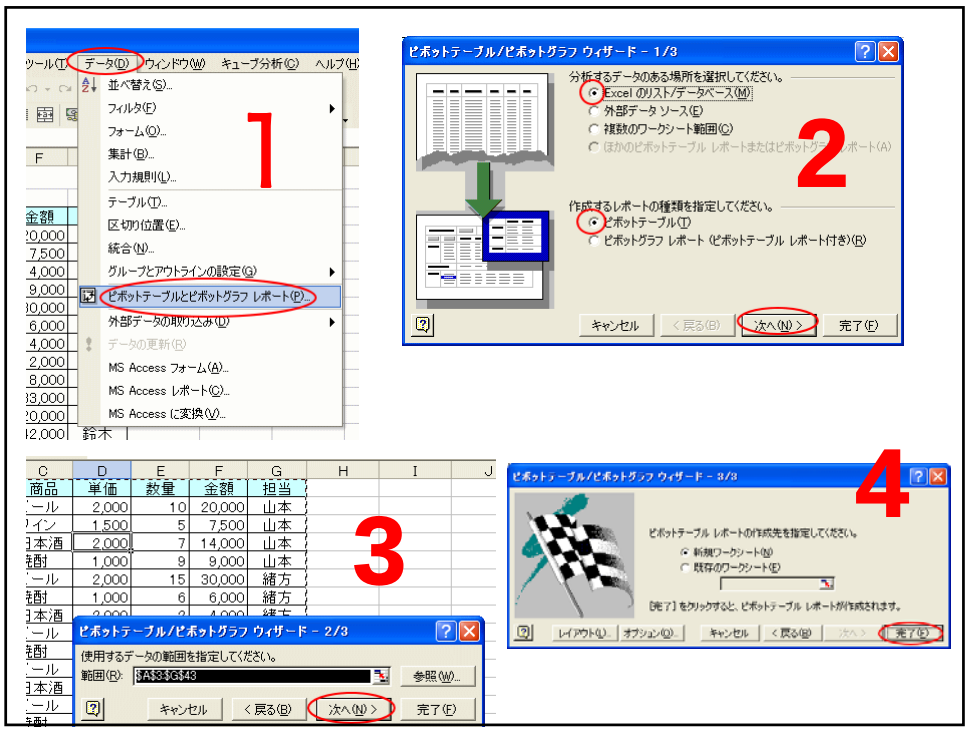
	A	B	C	D	E	F	G	H	I	J	K
1	ファイル名	コーパス名	行	位置	先行文脈	直前語	キーワー	後続文脈	文字数を数え	"を"の有	"ど"の有
2	D:¥MY-DOC-	国会議事録サンプル.txt	23	3040	政府の代表	関	して	明年十一月	1	#VALUE	#VALU
3	D:¥MY-DOC-	国会議事録サンプル.txt	35	4574	日中平和友	対	して	ソ連側では	1	#VALUE	#VALU
4	D:¥MY-DOC-	国会議事録サンプル.txt	35	4618	ソ連側では	対	して	不信感を抱	1	#VALUE	#VALU
5	D:¥MY-DOC-	国会議事録サンプル.txt	35	4652	本に対して	対	して	ソ連側の要	1	#VALUE	#VALU
6	D:¥MY-DOC-	国会議事録サンプル.txt	35	4692	してソ連側	対	して	は耳を傾け	1	#VALUE	#VALU
7	D:¥MY-DOC-	国会議事録サンプル.txt	48	7136	したときの	減	して	二十万トン	1	#VALUE	#VALU
8	D:¥MY-DOC-	国会議事録サンプル.txt	48	7206	サバを十万に		して	もらいたい。	1	#VALUE	#VALU
9	D:¥MY-DOC-	国会議事録サンプル.txt	58	9303	鐘がくるか	対	して	ソ連は非常	1	#VALUE	#VALU
10	D:¥MY-DOC-	国会議事録サンプル.txt	59	9531	る、多少無	押	して	きたのか。	1	#VALUE	#VALU
11	D:¥MY-DOC-	国会議事録サンプル.txt	60	9789	は問題を面	対	して	は巷間いろ	1	#VALUE	#VALU
12	D:¥MY-DOC-	国会議事録サンプル.txt	75	12589	っかり追っ	対	して	、専門的な	1	#VALUE	#VALU
13	D:¥MY-DOC-	国会議事録サンプル.txt	77	13019	いかと思い	対	して	どのような	1	#VALUE	#VALU
14	D:¥MY-DOC-	国会議事録サンプル.txt	77	13471	から、転換	示	して	もらいたい。	1	#VALUE	#VALU
15	D:¥MY-DOC-	国会議事録サンプル.txt	103	17930	土問題とし	対	して	日本の水産	1	#VALUE	#VALU
16	D:¥MY-DOC-	書籍サンプル.txt	19	1020	味期限も短	と	して	は考えませ	1	#VALUE	
17	D:¥MY-DOC-	書籍サンプル.txt	43	2634		と	して	ガンガン飛	1	#VALUE	
18	D:¥MY-DOC-	書籍サンプル.txt	83	4602		出	して	ズバリ言う、	1	#VALUE	#VALU
19	D:¥MY-DOC-	書籍サンプル.txt	111	5820		そ	して	、わたし流	1	#VALUE	#VALU
20	D:¥MY-DOC-	書籍サンプル.txt	123	6256		そ	して	、この恋愛	1	#VALUE	#VALU
21	D:¥MY-DOC-	書籍サンプル.txt	135	7112		に	して	も、結婚した	1	#VALUE	#VALU
22	D:¥MY-DOC-	書籍サンプル.txt	156	8286	躍動的な部	そ	して	優しくて母	1	#VALUE	#VALU
23	D:¥MY-DOC-	書籍サンプル.txt	171	9052		そ	して	、残った料	1	#VALUE	#VALU
24	D:¥MY-DOC-	書籍サンプル.txt	181	9638	の言葉はい	と	して	必須のキー	1	#VALUE	
25	D:¥MY-DOC-	書籍サンプル.txt	211	11174	を騙すタイ	隠	して	話す癖です	1	#VALUE	#VALU
26	D:¥MY-DOC-	書籍サンプル.txt	248	13290		出	して	はいけませ	1	#VALUE	#VALU
27	D:¥MY-DOC-	書籍サンプル.txt	332	17288		愛	して	いるからす	1	#VALUE	#VALU
28	D:¥MY-DOC-	書籍サンプル.txt	332	17340	すべてを許	愛	して	いるからこ	1	#VALUE	#VALU
29	D:¥MY-DOC-	書籍サンプル.txt	332	17374	、「愛してい	許	して	はいけない	1	#VALUE	#VALU
30	D:¥MY-DOC-	書籍サンプル.txt	336	17580	友	許	して	あげる。	1	#VALUE	#VALU
31	D:¥MY-DOC-	書籍サンプル.txt	337	17604		愛	して	いるんだも	1	#VALUE	#VALU
32	D:¥MY-DOC-	書籍サンプル.txt	363	18710	性はとことん	そ	して	飽きたらポ	1	#VALUE	#VALU
33	D:¥MY-DOC-	書籍サンプル.txt	364	18744		許	して	しまう女性	1	#VALUE	#VALU
34	D:¥MY-DOC-	書籍サンプル.txt	370	19006		恋	して	いる間は相	1	#VALUE	#VALU
35	D:¥MY-DOC-	白書サンプル.txt	49	4713		そ	して	それらは多	1	#VALUE	#VALU
36	D:¥MY-DOC-	白書サンプル.txt	71	6941	年度より増	示	して	いる。	1	#VALUE	#VALU

その他便利な関数

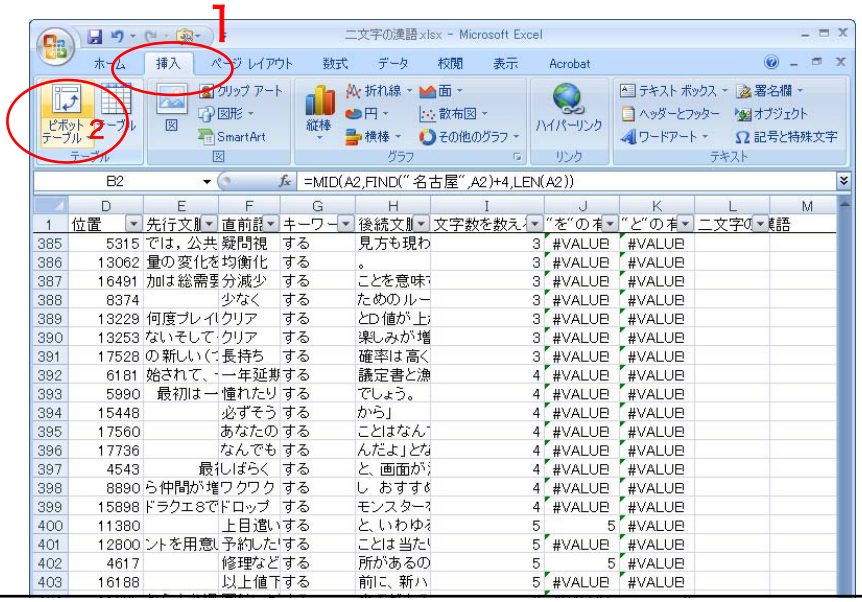
- **=セル番地＆セル番地**：文字列をくっ付ける
- **=セル番地＆“文字列 “**：セルに文字列を追加する
- **=PHONETIC(セル番地)**：発音をかえす
- **=EXACT(文字列1,文字列2)**：文字列の異同をチェックする

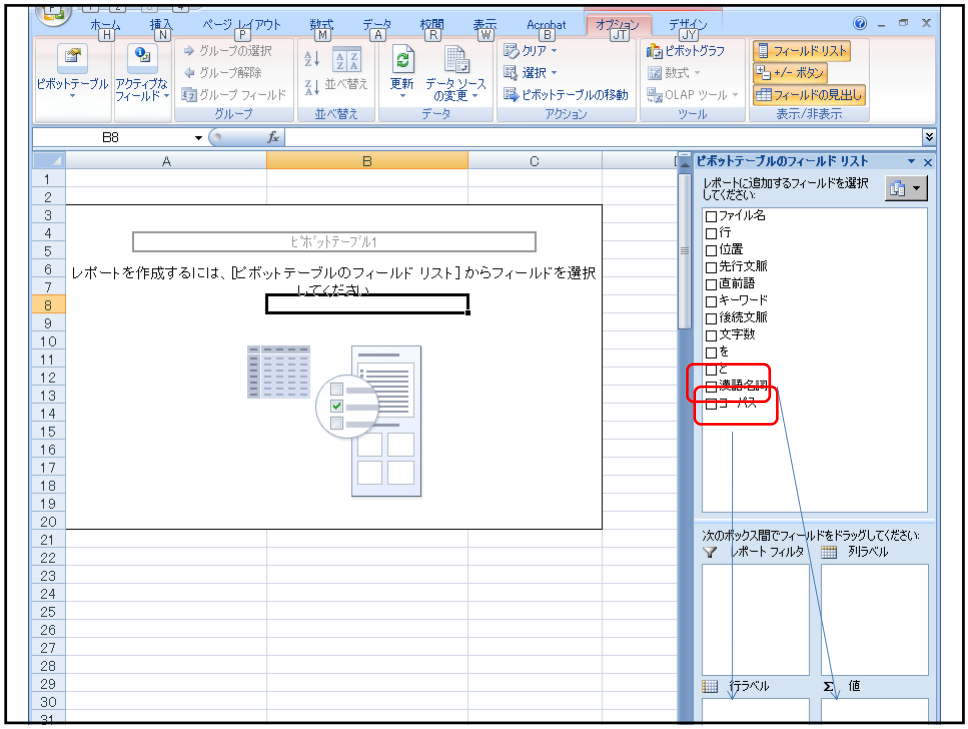
ピボットテーブル

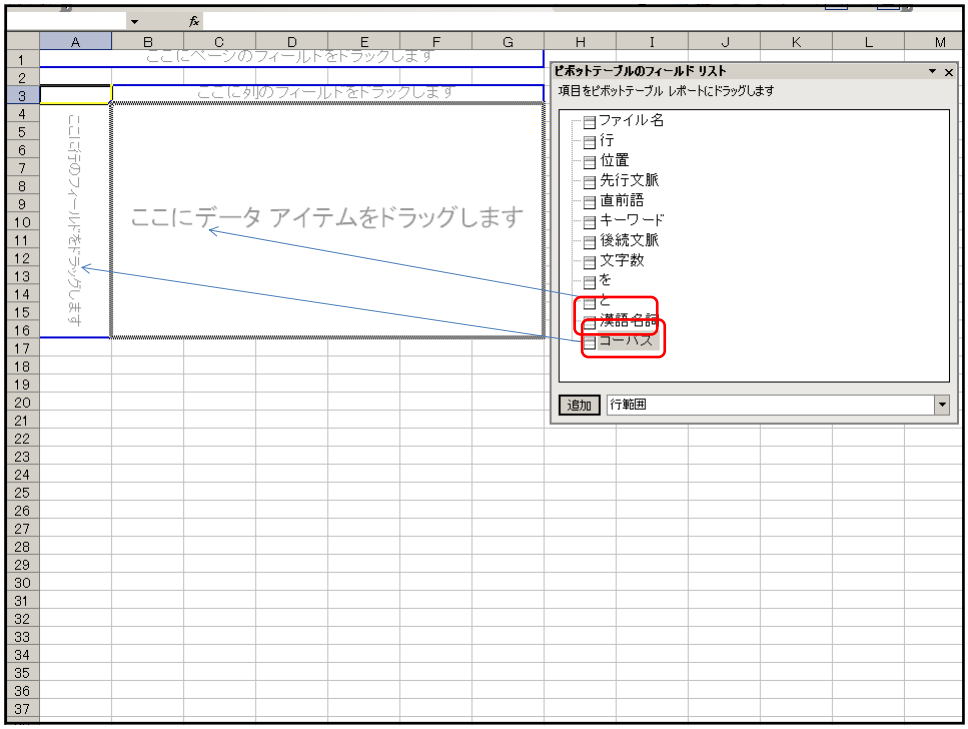
- エクセルに備わっている データの集計や分析をマウス のドラッグ操作で簡単に行うことが出来る機能
- 主な用途
  - 度数表：タイプ、トークン頻度
  - クロス集計表
  - 平均や合計なども計算可能

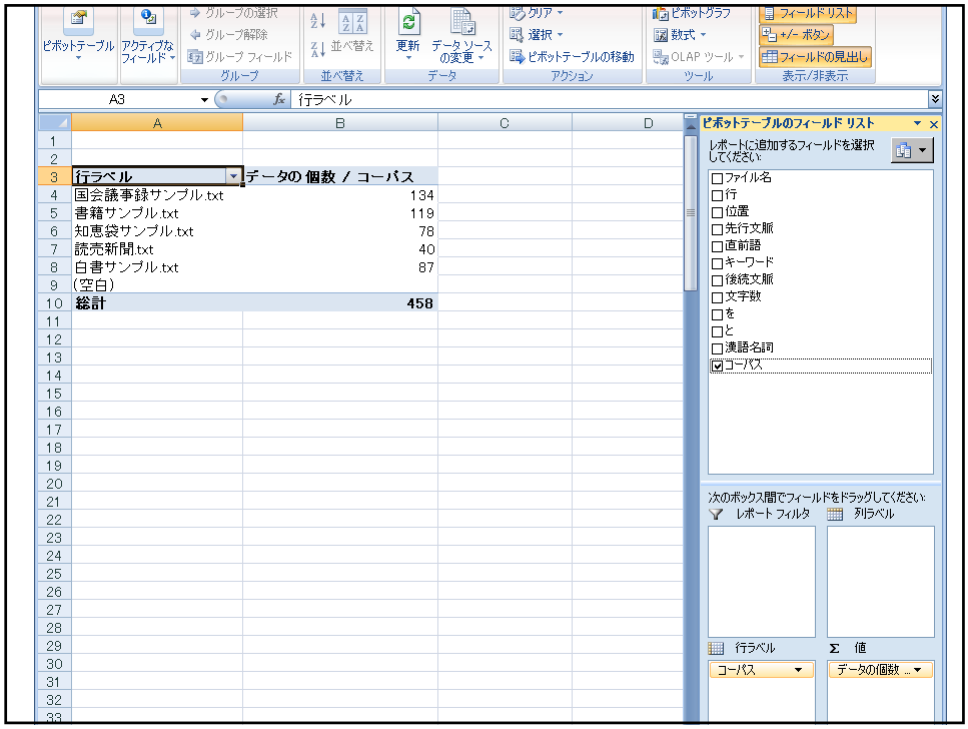


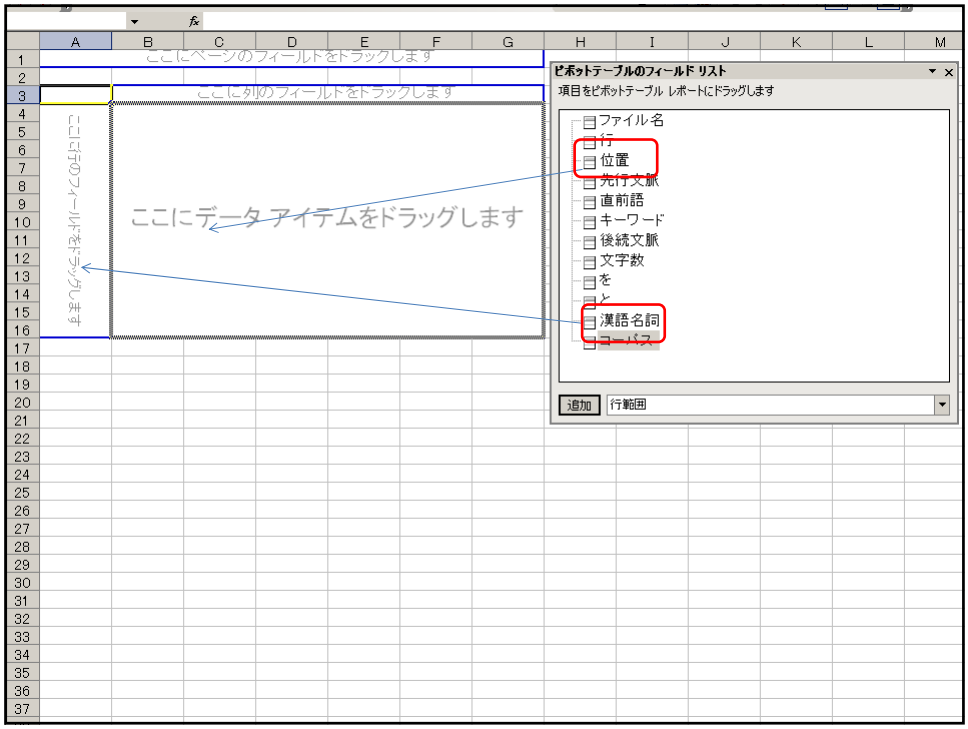
オフィス2007の場合

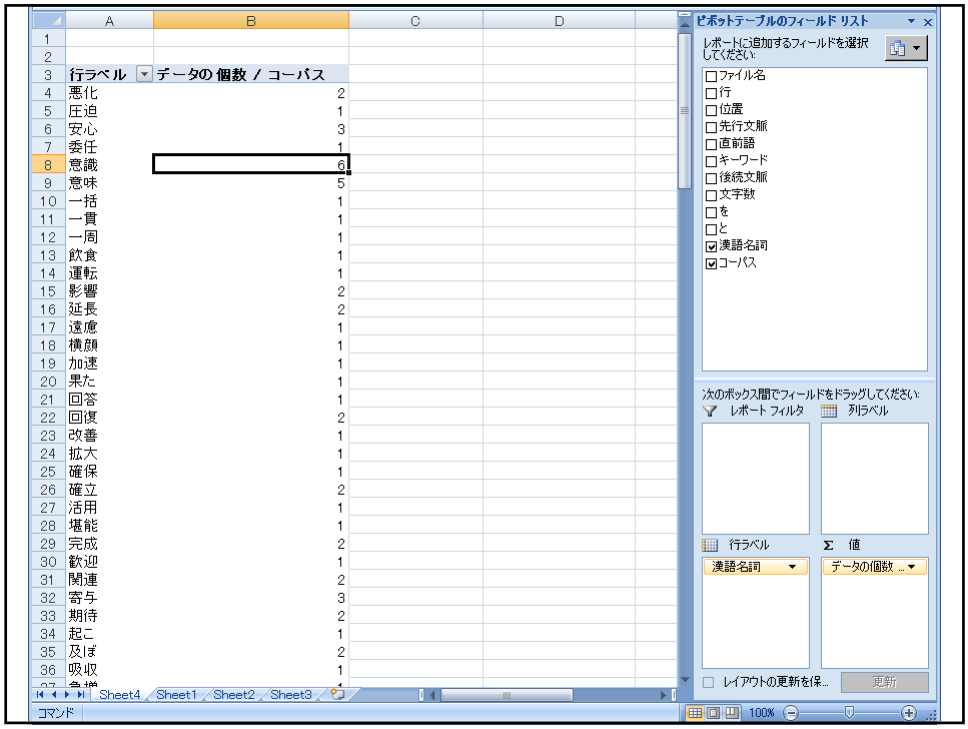



ピボットテーブル1
レポートを作成するには、[ピボットテーブルのフィールド リスト] からフィールドを選択してください
ピボットテーブルのフィールド リスト
レポートに追加するフィールドを選択してください:
ファイル名
行
位置
先行文脈
直前語
キーワード
後続文脈
文字数
を
と
漢語名詞
コーパス
次のボックス間でフィールドをドラッグしてください:
レポート フィルタ
列ラベル
行ラベル
値


ここにページのフィールドをドラッグします
ここに列のフィールドをドラッグします
ここにデータ アイテムをドラッグします
ここに行のフィールドをドラッグします
ピボットテーブルのフィールド リスト
項目をピボットテーブル レポートにドラッグします
ファイル名
行
位置
先行文脈
直前語
キーワード
後続文脈
文字数
を
と
漢語名詞
コーパス
追加
行範囲


ピボットテーブル
アクティブな フィールド
グループの選択
グループ解除
グループ フィールド
グループ
並べ替え
並べ替え
更新
データ ソースの変更
データ
クリア
選択
ピボットテーブルの移動
アクション
ピボットグラフ
数式
OLAP ツール
ツール
フィールド リスト
+/- ボタン
フィールドの見出し
表示/非表示
A3
行ラベル
行ラベル
データの 個数 / コーパス
国会議事録サンプル.txt 134
書籍サンプル.txt 119
知恵袋サンプル.txt 78
読売新聞.txt 40
白書サンプル.txt 87
(空白)
総計 458
ピボットテーブルのフィールド リスト
レポートに追加するフィールドを選択してください:
ファイル名
行
位置
先行文脈
直前語
キーワード
後続文脈
文字数
を
と
漢語名詞
コーパス
次のボックス間でフィールドをドラッグしてください:
レポート フィルタ
列ラベル
行ラベル
値
コーパス
データの個数 ...

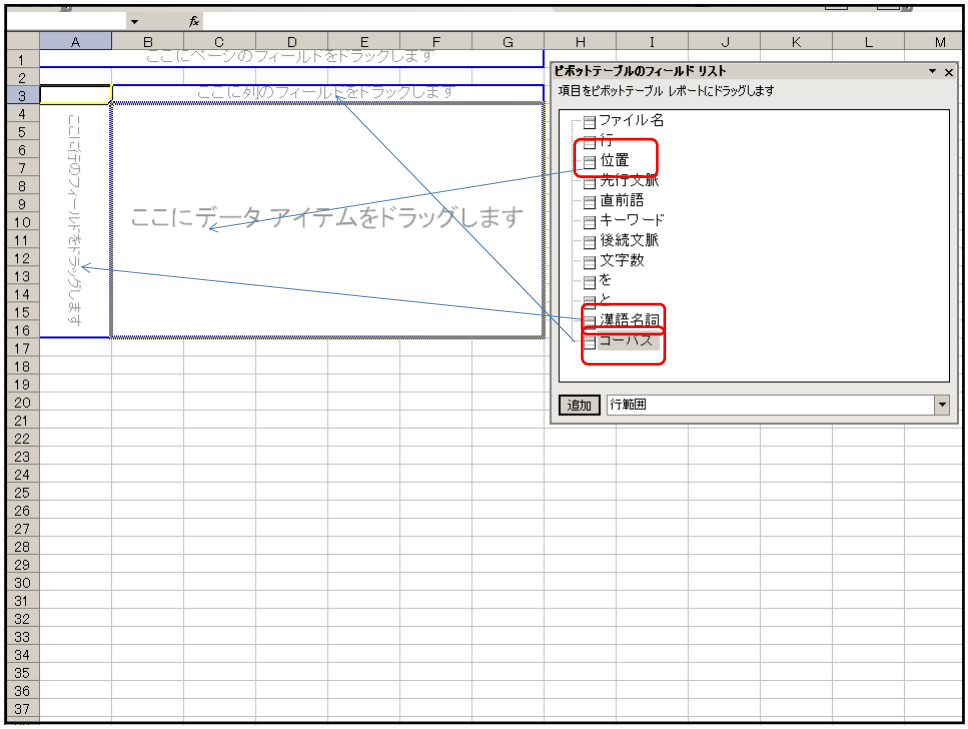

ここにページのフィールドをドラッグします
ここに列のフィールドをドラッグします
ここにデータ アイテムをドラッグします
ここに行のフィールドをドラッグします
ピボットテーブルのフィールド リスト
項目をピボットテーブル レポートにドラッグします
ファイル名
行
位置
先行文脈
直前語
キーワード
後続文脈
文字数
を
と
漢語名詞
コーパス
追加
行範囲

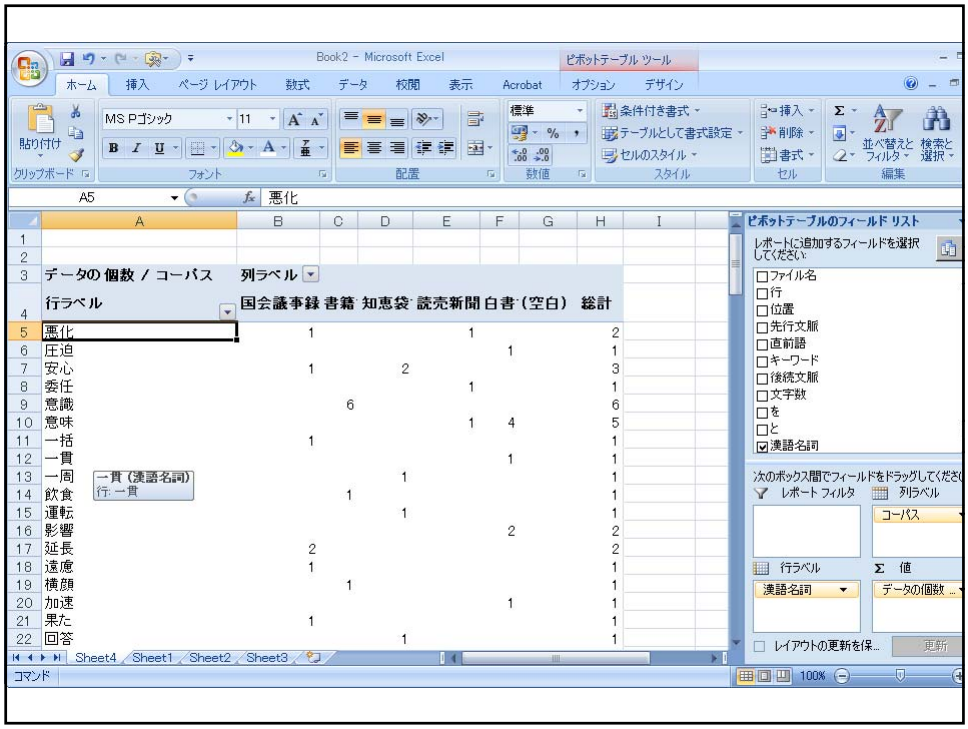


行ラベル	データの個数 / コーパス
悪化	2
圧迫	1
安心	3
委任	1
意識	6
意味	5
一括	1
一貫	1
一周	1
飲食	1
運転	1
影響	2
延長	2
遠慮	1
横顔	1
加速	1
果た	1
回答	1
回復	2
改善	1
拡大	1
確保	1
確立	2
活用	1
堪能	1
完成	2
歓迎	1
関連	2
寄与	3
期待	2
起こ	1
及ぼ	2
吸収	1

サ変名詞	トークン頻度
存在	4131
説明	3807
利用	3636
理解	3396
紹介	2943
確認	2383
発見	2293
結婚	2218
主張	2089
発生	2084
参加	2051
使用	2019
意味	2007
指摘	1960
成功	1861
登場	1817
実現	1815
判断	1775
展開	1764
表現	1718
成立	1706
用意	1660
発表	1645
心配	1634

←**BCCWJ**の書籍コーパス全体で調べると、


ここにページのフィールドをドラッグします
ここに列のフィールドをドラッグします
ここにデータ アイテムをドラッグします
ここに行のフィールドをドラッグします
ピボットテーブルのフィールド リスト
項目をピボットテーブル レポートにドラッグします
ファイル名
行
位置
先行文脈
直前語
キーワード
後続文脈
文字数
を
と
漢語名詞
コーパス
追加
行範囲


Book2 - Microsoft Excel
ピボットテーブル ツール
ホーム
挿入
ページ レイアウト
数式
データ
校閲
表示
Acrobat
オプション
デザイン
A5
悪化
データの個数 / コーパス
列ラベル
行ラベル
国会議事録 書籍 知恵袋 読売新聞 白書 (空白) 総計
悪化 1 1 2
圧迫 1 1
安心 1 2 3
委任 1 1
意識 6 6
意味 1 4 5
一括 1 1
一貫 1 1
一周 1 1
飲食 1 1
運転 1 1
影響 2 2
延長 2 2
遠慮 1 1
横顔 1 1
加速 1 1
果た 1 1
回答 1 1
一貫 (漢語名詞)
行: 一貫
ピボットテーブルのフィールド リスト
レポートに追加するフィールドを選択してください:
ファイル名
行
位置
先行文脈
直前語
キーワード
後続文脈
文字数
を
と
漢語名詞
次のボックス間でフィールドをドラッグしてください:
レポート フィルタ
列ラベル
コーパス
行ラベル
値
漢語名詞
データの個数
レイアウトの更新を保...
更新
Sheet4
Sheet1
Sheet2
Sheet3
コマンド
100%

正規表現とは

- 正規表現（せいきひょうげん、regular expression）とは，文字列の集合を一つの文字列で表現する方法の一つである。
- ある要素（文字とか単語など）の前後に特定のパターンを持ったものが連接していく形の記述
  - 戻って　買って　貼って　いって
  - → [戻買]って
  - → .って

正規表現

- 「秀丸」でできる
- ワードではできない色々な操作が可能
  - 改行やタブを消したり、入れたり。
  - 空白行を見つけて消したり、
  - 特定の条件にマッチした表現だけを集めたり
    - 詳しくは：http://www.shuiren.org/chuden/teach/hidemaru/seiki/index-j.html
- ワードに比べ、非常に高速であり、動作が軽い。

記号の説明

- ［］キャラクタクラス
  - 一文字単位で指定するときに使用.
  - abc，abb，aba とあるとき，ab の後ろには，a, b, c のどれかがくるので，これらをまとめるときには，ab[abc] のように記述する.
  - 便利な方法
    - ひらがな全部：[ぁ-ん]
    - 漢字全部：[亜-黒]
    - 「-」は，その間にあるもの全てを表す.

記号の説明

- **()** パターンをグループ化する
  - 同じ種類の表現をまとめたいときに使用.
  - 今日，明日，明後日 のどれかを検索したいときは，[]は使えない． [今明明後]日では，(明後日が拾えず)「後日」が検索されてしまう.
  - そこで，それぞれをまとめて(今日|明日|明後日)とする． 「|」は区切りの記号と覚える.
  - もう少しまとめて「(今|明|明後)日」としてもよい.

正規表現はすごい！

1.<E>.*</E>

2.<T.*¥n

3.<E>.*</E> | <T.*¥n

- お疲れ様でした！
- 明日は、
  - 形態素解析器をインストールします。
  - PC操作に自身のない方は、9:30に来ていただければ、李がインストール・形態素解析器の動作確認します。

2日目

2010/1/23-24
名古屋大学

日本語コーパスの最新の検索エンジンと分析

李在鎬（りじぇほ）
国際交流基金

JAPANFOUNDATION 国際交流基金

1/24のお品書き

- 復習
- 前半
  - コーパスツール紹介：形態素解析
  - 実習
- 後半
  - タグ付きＫＹコーパス
  - コーパスデータ紹介：ウェブからアクセス可能なフリーデータ

復習

- コーパスとは構造化された言語データの集積。
- コーパス言語学では、共起関係と頻度を重視。
- 調査を開始する際には、
  - コーパス選び：代表性のあるデータ選択
  - 用例抽出：KWIC形式。正規表現を活用→Kwic finder
  - 用例集計：エクセルを活用→関数・ピボットテーブル

- 正規表現：さるでもわかる正規表現
  - http://www.mnet.ne.jp/~nakama/
- エクセル関数：
  - http://www.eurus.dti.ne.jp/~yoneyama/Excel/kansu/itiran.html

KWIC Finderの問題点

- コーパス大きさが分からずに検索をする→得られたデータの位置づけがよくわからない。
  - １億語の中で１００回使われている表現
  - １万語の中で１００回使われている表現
- Kwic finderとエクセルで数える方法ではいずれも、同じくらいの表現として解釈！
- ウェブ検索と結果的には変わらない→コーパス研究とは言えない。

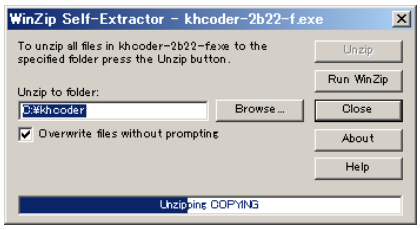
- 何とかしなければ・・・・。
- コーパスの大きさを知る
  - 文字数
  - 語数（延べか異なりか）
  - 一文の平均文字数や語数

→　形態素解析を使って、語数を調べる。

ソフトのインストール

– KH Coder

- これで、KH Coder(プログラム本体)とMySQL(データベース用ソフト), ChaSen(形態素解析ソフト), R(統計ソフト)が同時にインストールされたことになる。



形態素解析に基づくデータ加工と抽出

形態素解析とは（

- 形態素解析（**Morphological Analysis**）
  - コンピュータ等の計算機を用いた自然言語処理の基礎技術のひとつ。
  - 対象言語の文法の知識（文法のルールの集まり）や辞書（品詞等の情報付きの単語リスト）を情報源として用い、自然言語で書かれた文を形態素（Morpheme; 言語で意味を持つ最小単位）の列に分割し、それぞれの品詞を判別する作業を指す。
  - 具体例：「お待ちしております」

形態素解析とは（

文字列	読み	原形	品詞の種類	活用の種類	活用形
お待ち	オマチ	お待ち	名詞-サ変接続		
し	シ	する	動詞-自立	サ変・スル	連用形
て	テ	て	助詞-接続助詞		
おり	オリ	おる	動詞-非自立	五段・ラ行	連用形
ます	マス	ます	助動詞	特殊・マス	基本形
。	。	。	記号-句点		

*茶筌の場合

形態素解析とは

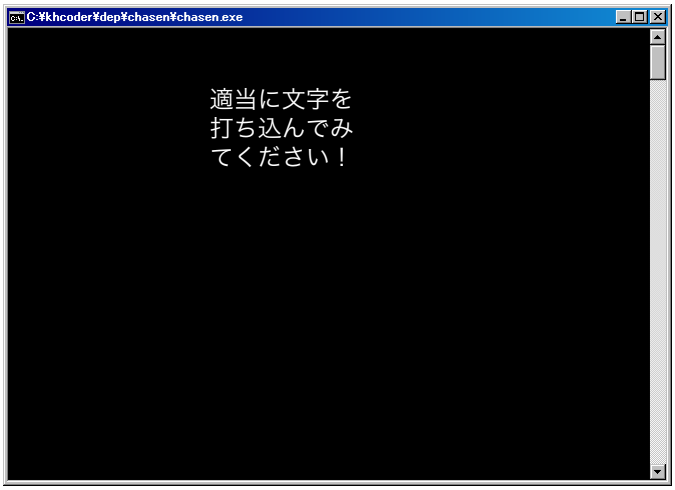
- どんなものがあるのか。
  - ChaSen（茶筌(×茶筅)）：奈良先端大学
    - http://chasen.naist.jp/hiki/ChaSen/
  - Juman：京大で開発
    - http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
  - Mecab（和布蕪）：京都大学・NTT
    - http://mecab.sourceforge.net/
  - 色々ある。。。

- やってみよう！　形態素解析

ソフトを立ち上げよう

- 手順1：マイコンピューターをクリック
- 手順2：C:¥ドライブをクリック
- 手順3：c:¥KH Coder→dep →chasen → chasen.exe

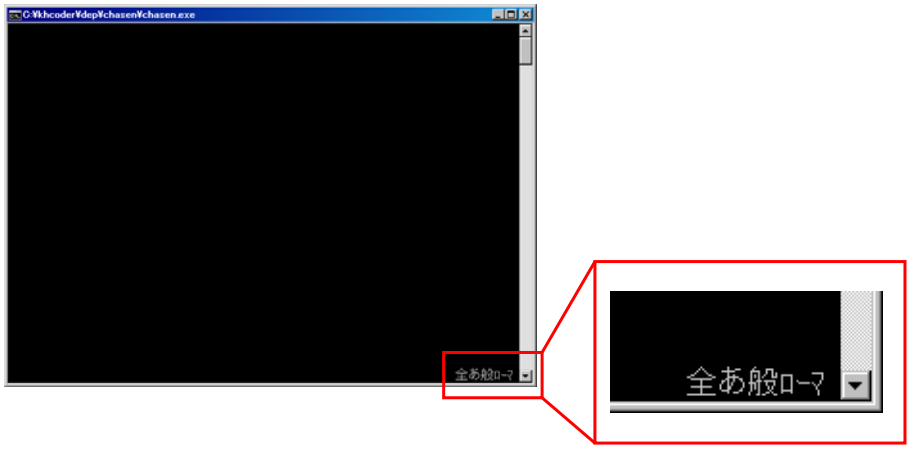
この画面が出てればOK



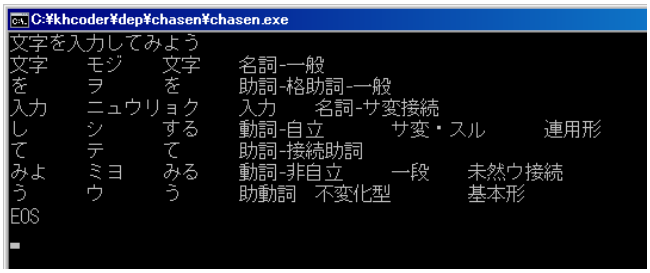
日本語を出すには？

「Alt」キーを押しながら「半角/全角」キーを押す！

この画面が出てればOK

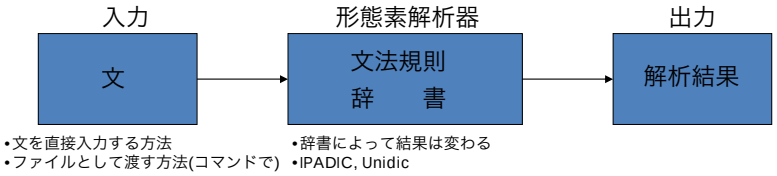


- 文字を入力してみよう！
  - 「文字を入力してみよう」と入力



結果と簡単な仕組み

- 最初のウィキペディアの説明に戻ると、
  - 対象言語の文法の知識（文法のルールの集まり）や辞書（品詞等の情報付きの単語リスト）を情報源として用い、自然言語で書かれた文を形態素の列に分割し、それぞれの品詞を判別する作業を指す。



- 何に使える？、どんなことに役立つ？
  - 直接的な利用法：解析結果をエクセルファイルなどに読み込んで品詞や語単位で数える
    →計量的調査に利用
  - 二次的な利用法：コンコーダンサー(concordancer)を使って高度な検索を行うことができる
- 例えば、

何ゆえ形態素解析か

- タスク
  - 動詞の用法確認とデータの収集
    e.g., サ変名詞＋する
- **Before**
  - Googleで文字列検索
  - 秀丸で「検索」
  - KWIC Finder
- **After**

⇒ゴミが大量。
活用形別に正規表現を組まないといけないので面倒

KWICコンコーダンス

Search Entry

抽出語: する　品詞:　活用形:　追加条件＊　検索

ソート1: 出現順　ソート2: 出現順　ソート3: 出現順　（前後 20 語を表示）

Result

“警察メディア”が	成立	する	（↓）『無尽蔵殺人事件』と『フォーカス』『週刊文春』（↓） “無尽蔵殺人事件”なるもの
古美術商「無尽蔵」の店主・長尾泰次氏が八二年の二月ごろから	密発	し	た。（↓）殺人事件と判断した警察が、横領容疑の別件で「無尽蔵」の
店主・長尾泰次氏が八二年の二月ごろから密発した。（↓）殺人事件と	判断	し	た警察が、横領容疑の別件で「無尽蔵」の店員を逮捕、殺害を自供
た警察が、横領容疑の別件で「無尽蔵」の店員を逮捕、殺害を	自供	さ	せたもので、現在公判中である（ちなみに横領容疑などという代物を絶対的
ある（ちなみに横領容疑などという代物を絶対的に免れうる生活者などこの世に	存在	し	ないだろう）。（↓） わたしはシロ・クロを判定する立場にないが、この
生活者などこの世に存在しないだろう）。（↓） わたしはシロ・クロを	判定	する	立場にないが、この事件は冤罪事件のさまざまなケースを、過不足なく思い起こさせる
冤罪事件のさまざまなケースを、過不足なく思い起こさせる。（↓）問題は、“冤罪を	創作	する	”ためにマスコミ・メディアが重要な役割を意図してはたす、という新しい状況が
問題は、“冤罪を創作する”ためにマスコミ・メディアが重要な役割を	意図	し	てはたす、という新しい状況が派生したことである。（↓）見のがされやすいが、ほんとう
ためにマスコミ・メディアが重要な役割を意図してはたす、という新しい状況が	派生	し	たことである。（↓）見のがされやすいが、ほんとうに危険なのはこの“新しい現象
商・無尽蔵」ホモ店主殺害事件／緊迫の法廷シーン》と題して、冤罪	逮捕	さ	れた店員をめぐる公判の現状を伝えた。（↓）《「生きている姿を見
の官憲がどれほどお手軽に重大事件を立件しているかを、如実に	暴露	し	たケースといえるだろう。（↓） この記事によると、証人は三人だ
ため、殺意をもって、手近にあった鉄棒で同人の頭部を多数回にわたり	殴打	し	、よって、そのころ、同所において、同人をして頭蓋内損傷により死亡
、よって、そのころ、同所において、同人をして頭蓋内損傷により	死亡	する	に至らせ…》と起訴状にあるように、“犯人”笹川竜則氏
の五月二十八日に、土曜ポエムナイトという集会で冤罪ルポライター・佐藤友之と	対談	し	たとき、わたしは次のようにのべている。（↓） 《ぼくなんかが
方になっているが、引用末尾は、捜査当局が立件にあたって物的証拠を	提出	し	なくてすむ、という意味である。（↓）きわめて似かよったケースである六甲山事件の冤罪被告
死体を捨てたんだろう、どこか水のあるところだろう…”と	誘導	さ	れた、とのべている。（↓） さらにいえば、わたしは笹川氏の逮捕以前
いる。（↓） さらにいえば、わたしは笹川氏の逮捕以前に彼の無実を	確信	し	、その時点で“無尽蔵殺人事件”に冤罪の“構造”をかぎとってい
のである。（↓） 逮捕以前に有罪であるとはいかなる意味か。（↓）法に	先行	する	法、制度に超越する制度が彼を犯人と特定し、その有罪を一方的
以前に有罪であるとはいかなる意味か。（↓）法に先行する法、制度に	超越	する	制度が彼を犯人と特定し、その有罪を一方的に裁いていた、
意味か。（↓）法に先行する法、制度に超越する制度が彼を犯人と	特定	し	、その有罪を一方的に裁いていた、ということである。（↓）すなわち、
”をめぐる、ほんとうに裁かれねばならない犯罪は、まさにそこにこそ	存在	する	。（↓）さきほどの引用に続き、わたしは次のようにものべた。（↓） 《それ
ある（笑）と一言でいっていいと思うんですけど、無尽蔵事件で	自白	する	前に、（「文春」にものったけれども）「フォーカス」と「新潮」
。（↓）それをつくられた“像”であると対象化する読者は、ぬったに	存在	し	ない。（↓）事実そのものと受けとめるのが大方だろう。（↓）そして公判に入るとすぐ自白を
。（↓）事実そのものと受けとめるのが大方だろう。（↓）そして公判に入るとすぐ自白を	否定	し	て冤罪を主張しているという事実などは、意図して殺人犯の“像
のが大方だろう。（↓）そして公判に入るとすぐ自白を否定して冤罪を	主張	し	ているという事実などは、意図して殺人犯の“像”をつくりあげたメディア
入るとすぐ自白を否定して冤罪を主張しているという事実などは、	意図	し	て殺人犯の“像”をつくりあげたメディアの、同じ必然性によって黙殺される
意図して殺人犯の“像”をつくりあげたメディアの、同じ必然性によって	黙殺	さ	れるから、いったんつくりあげられた像によって、犯人はいつまでも犯人であり続けね
あり続けねばならないのだ。（↓） 今回のケースも、彼のアリバイが	立証	さ	れたぐらいでは“黙殺”は続いていただろう。（↓）しかし殺された
が、警察はいっこうに私の話を聞こうとしなかった》という声を紹介		し	、片岡証人の《池袋署の刑事は『証言を変えろ。（↓）お前の証言は間違い
、片岡氏と出版元の社長が警視庁に呼ばれ、「本の内容を訂正		し	なくともいいが、『本に書いたことは事実と違う』という一札をいれ
に書いたことは事実と違う』という一札をいれてくれ。（↓）そのことは公表		し	ないから」と要求され、片岡氏らが拒否するという一幕もあった》

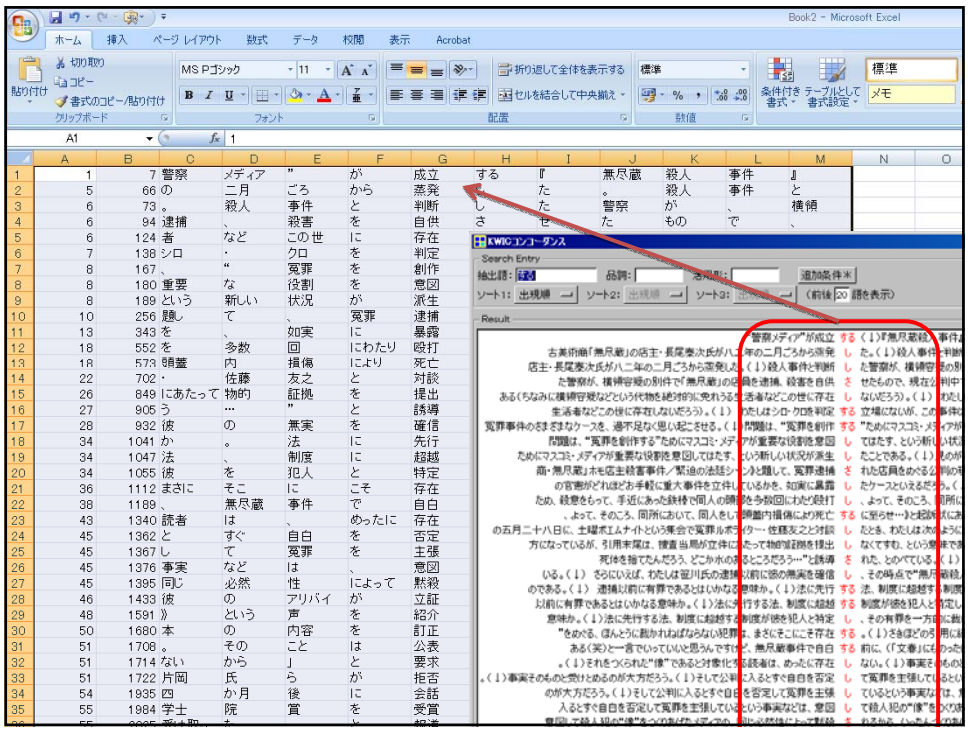
コロケーション統計

Nord Word

抽出語: する　品詞:　活用形:　ヒット数: 486468

Result

N	抽出語	品詞	スコア	左合計	右合計	左5	左4	左3	左2	左1	右1	右2	右3
1	た	助動詞	118487.08	21406	164789	8317	9588	3157	344	0	84131	29664	245
2	を	格助詞	86272.63	166286	43246	13516	17111	8656	127003	0	50	7007	112
3	に	格助詞	59948.47	103168	47073	11754	12858	8740	69816	0	1982	12060	104
4	が	格助詞	45362.47	78455	39281	9921	11375	5543	51616	0	36	11137	87
5	は	係助詞	35406.27	54895	56044	14547	12988	8489	18871	0	38	10339	140
6	こと	依存名詞	31259.92	13872	40302	2133	1795	9889	55	0	19750	7566	54
7	いる	補助動詞	31074.80	6439	67175	2697	2902	807	33	0	1	46479	154
8	だ	助動詞	24688.22	23792	57336	9149	11044	2908	691	0	1070	12233	150
9	ない	否定助動詞	18646.30	7541	30709	2598	3136	1638	169	0	10941	2727	43
10	の	依存名詞	18398.90	3850	34538	1212	1245	1392	1	0	7831	9681	76
11	で	格助詞	16957.90	33818	10870	4507	5410	2853	21048	0	25	1655	29
12	ます	助動詞	14612.27	2436	24312	1235	806	395	0	0	9269	2538	57
13	する	動詞B	14010.42	22449	28770	8656	7176	5831	786	0	56	2342	93
14	も	係助詞	13961.98	20502	18121	3848	3456	2458	10740	0	45	5698	41
15	よう	依存名詞	8837.35	6517	12184	2050	507	3270	690	0	3903	3162	20
16	と	格助詞	8658.80	12328	10708	2586	1539	1213	6990	0	18	4454	19
17	う	助動詞	8473.38	2347	13003	757	570	1010	10	0	5392	1854	20
18	ため	依存名詞	7248.75	2723	8628	996	403	1139	185	0	5316	1212	7
19	から	格助詞	6805.25	12607	5715	2199	2065	971	7372	0	14	1504	12
20	もの	依存名詞	6495.83	3532	10207	758	648	2121	5	0	2581	3488	19
21	ある	助動詞	4737.35	4516	14033	1830	1732	948	6	0	0	187	37
22	存在	サ変名詞	4532.52	4802	587	162	128	381	0	4131	41	121	1
23	べし	助動詞	4282.42	761	4809	210	407	51	93	0	3646	534	2
24	なる	動詞B	3966.02	3827	10425	1487	1290	996	54	0	16	538	42
25	説明	サ変名詞	3952.05	3989	294	74	51	56	1	3807	16	49	5
26	たい	助動詞	3882.17	619	4928	202	236	179	2	0	3078	444	9
27	利用	サ変名詞	3824.43	3952	324	94	138	82	2	3636	13	56	1
28	理解	サ変名詞	3553.80	3624	313	83	58	85	2	3396	11	39	1
29	ぬ	否定助動詞	3536.83	1984	4356	581	508	582	313	0	2266	600	3
30	です	助動詞	3520.40	2031	9864	1012	668	351	0	0	112	1589	25

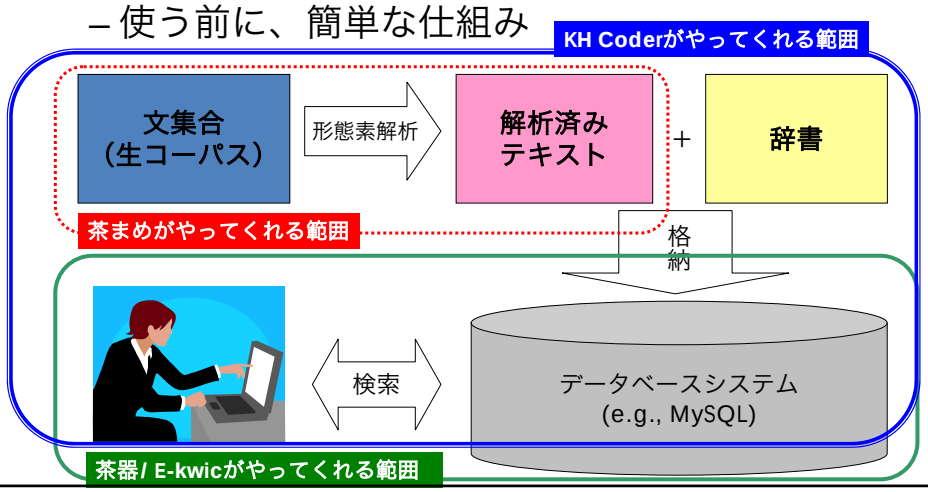


形態素解析を使う

- フリーで利用可能なツール
  - KH Coder
    - 立命館大学の樋口耕一氏が製作
    - 内容分析に特化。集計機能が充実
  - 茶器
    - 奈良先の浅原正幸氏が製作
    - 多言語コーパスに特化。カスタマイズ機能が充実
  - 茶まめ
    - 国研の日本語コーパス伝(他)が製作
    - 人文系への親和性重視。音声言語処理に有効。
  - E-Kwic

形態素解析を使う

- 茶まめ、茶器、KH Coder
  - 使う前に、簡単な仕組み



ツールの詳細情報

- 茶まめ
  1. 国研日本語コーパス電子辞書班
  2. 形態素解析補助ツール
  3. 形態素解析は茶筌かMecab
  4. 入手方法：
     http://www.tokuteicorpus.jp/dist/index.php
  5. 関連情報： 使用レポートなど
     http://www30.atwiki.jp/corpus-ling/pages/36.html
  6. 関連論文：伝 康晴(他)(2006)
     「多様な目的に適した形態素解析システム用電子化辞書の開発」
     (「日本語コーパス」全体会議電子化辞書班報告)

ツールの詳細情報

- KH Coder
  1. 樋口耕一（立命館大学）作成
  2. 計量社会学ツール
  3. 分析機能やデータ抽出機能が充実
  4. 入手方法：
     http://khc.sourceforge.net/
  5. 関連情報： 使用レポートなど
     http://www30.atwiki.jp/corpus-ling/pages/28.html
  6. 関連論文：佐野香織・李在鎬(2007)「KH Coderで何ができるか~日本語習得・日本語教育研究利用への示唆~」、『言語文化と日本語教育』(No.33),pp.94-95.

ツールの詳細情報

- E-KWIC
  1. 茶まめによる形態素解析済みのデータを検索
  2. エクセル上で動作する。インストール作業は不要
  3. 入手方法：
     http://www30.atwiki.jp/corpus-ling/pages/63.html

形態素解析のツールを使う

- 実習１：KH Coder
- 実習２：E-Kwic

形態素解析のツールを使う

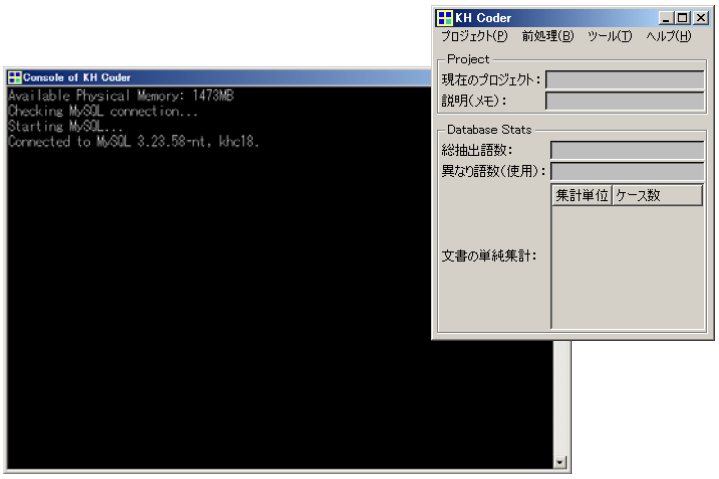
- 実習１：KH Coder
- 実習２：E-Kwic

ソフトのインストール

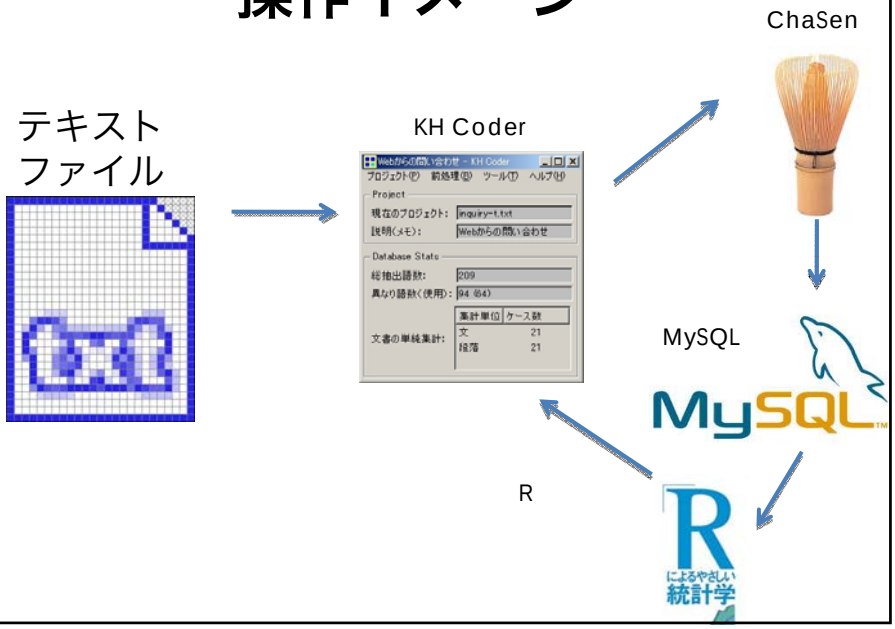
– KH Coder設定ファイルの差し替え
  - 皆さんの研究テーマに合わせて抽出可能な形態素の設定を書き換える
  - 方法：¥nagoya¥tempのhinshi_chasenをコピーして、c:¥kh coder¥configに張り付ける。
  - 上書きするかどうか聞いてくるが、上書きを選択する。
  - これで、格助詞や補助動詞、終助詞などが検索できるようになる。

起動してみよう

– C:¥khcoder¥kh_coder.exe をクリック！



操作イメージ

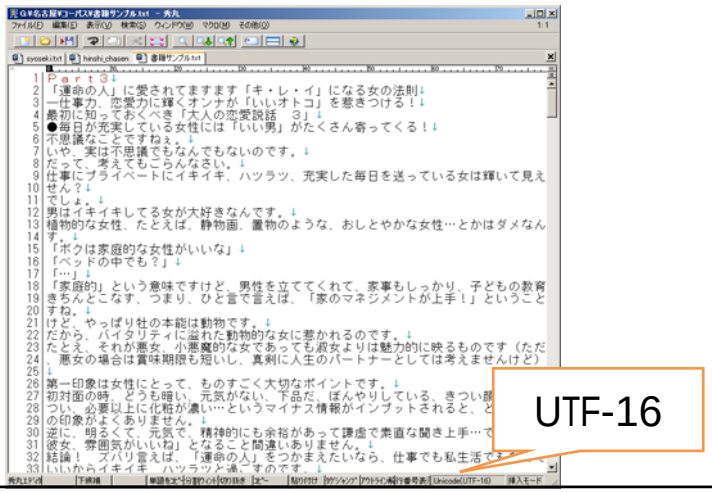


メニューバー

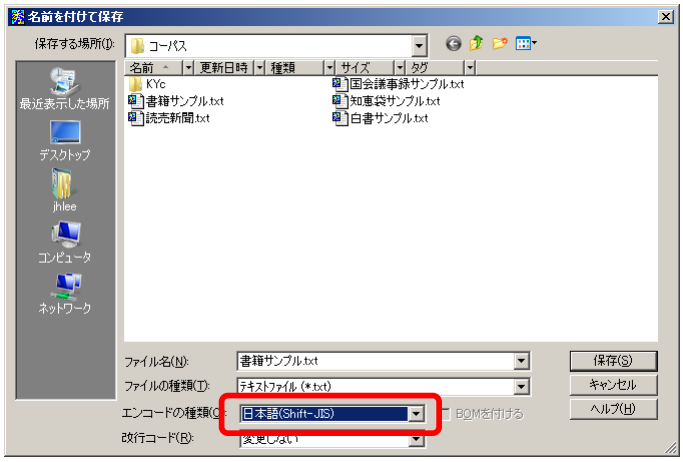
- 四つの基本操作
  - プロジェクト：コーパスの登録など
  - 前処理：「茶筌」による解析
  - ツール：データ抽出系や集計
  - ヘルプ

ごめんなさい

- テキストファイルの文字コードを変更

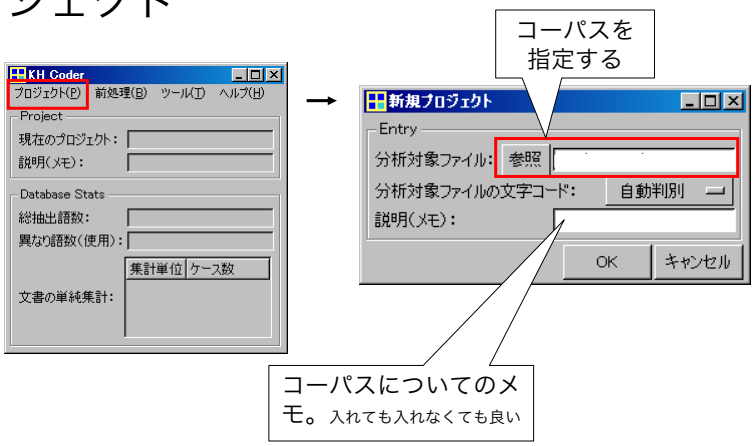


名前付けて保存



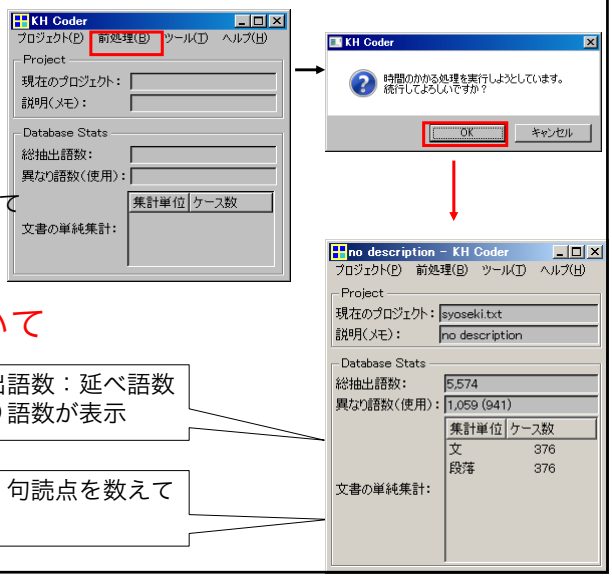
分析データの登録

- プロジェクト



前処理

- 茶筌の実行



KH Coderが分析対象として認識しているものの語数

異なり語数について
1059(941)

ChaSenによる解釈

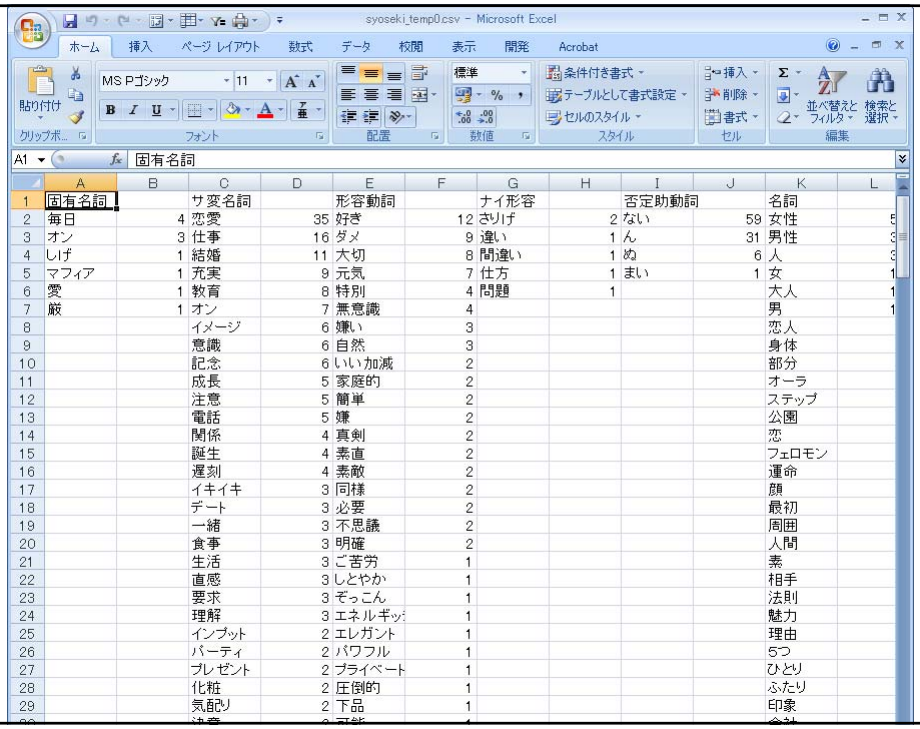
総抽出語数：延べ語数
異なり語数が表示

文数：句読点を数えている

結果を確認

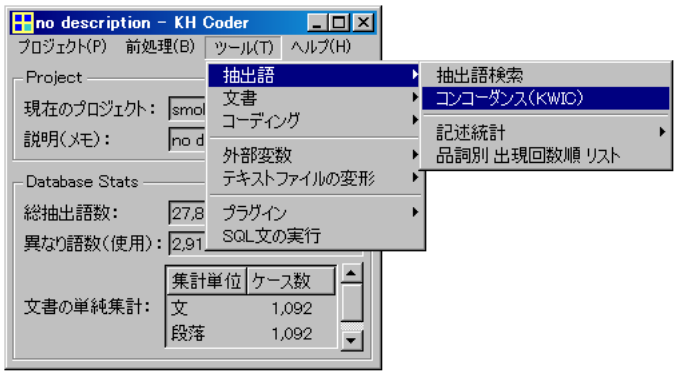
- 解析の結果を確認しよう
  - ツール > 抽出語 > 抽出語リスト

エクセルへ



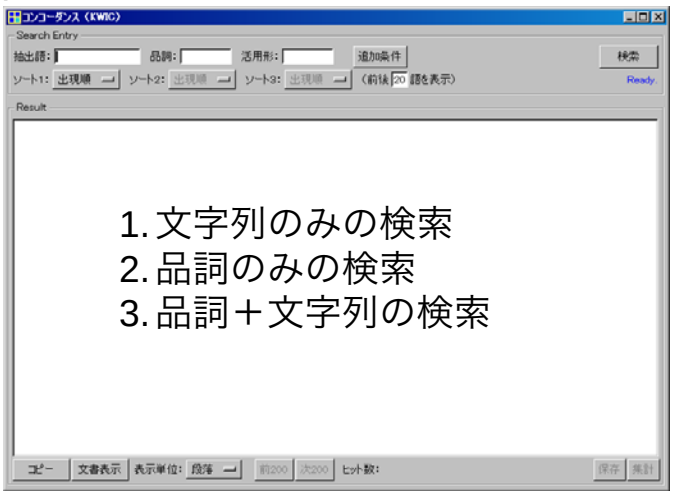
コンコーダンサーを起動

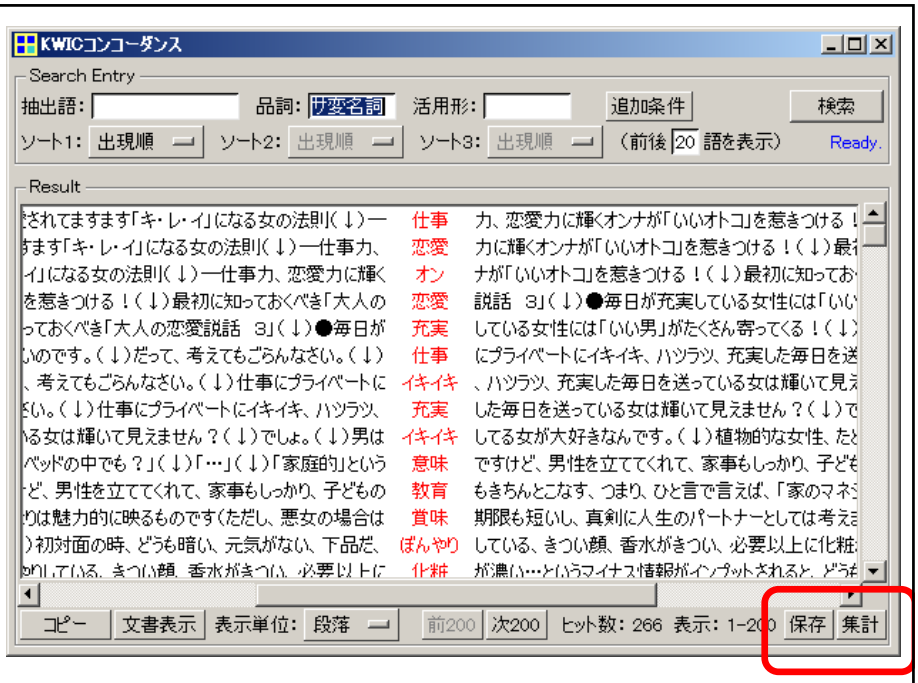
- KWIC
  - ツール＞抽出語＞コンコーダンス



コンコーダンサーを起動

- 検索しよう

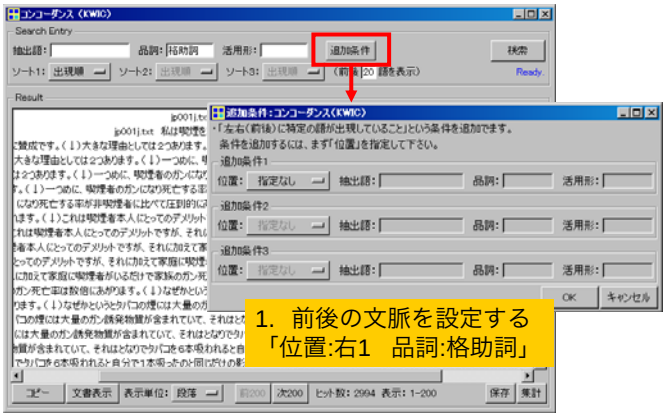




N	抽出語	品詞	スコア	左合計	右合計	左5	左4	左3	左2	左1	右
1	する	動詞	106.92	21	120	7	7	4	3	0	9
2	。	句点	57.53	82	55	13	14	27	8	20	
3	は	係助詞	52.82	45	43	3	7	6	9	20	1
4	に	格助詞	38.68	26	41	8	1	4	2	11	1
5	を	格助詞	38.23	24	34	1	5	2	1	15	1
6	が	格助詞	37.35	22	37	3	5	3	1	10	1
7	て	接続助詞	30.70	25	56	9	7	5	2	2	
8	です	助動詞	25.08	38	29	7	12	11	8	0	
9	で	格助詞	19.12	15	15	4	2	0	1	8	
10	も	係助詞	18.85	21	21	3	5	5	4	4	
11	だ	助動詞	17.25	25	18	5	7	5	3	5	
12	ます	助動詞	11.37	22	16	8	8	2	4	0	
13	た	助動詞	11.20	10	22	1	4	4	0	1	
14	ない	否定助動詞	10.63	9	19	1	2	2	2	2	

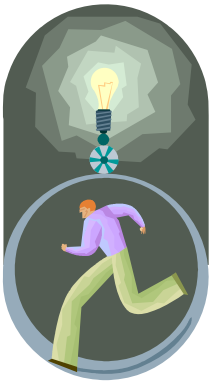
パターン検索

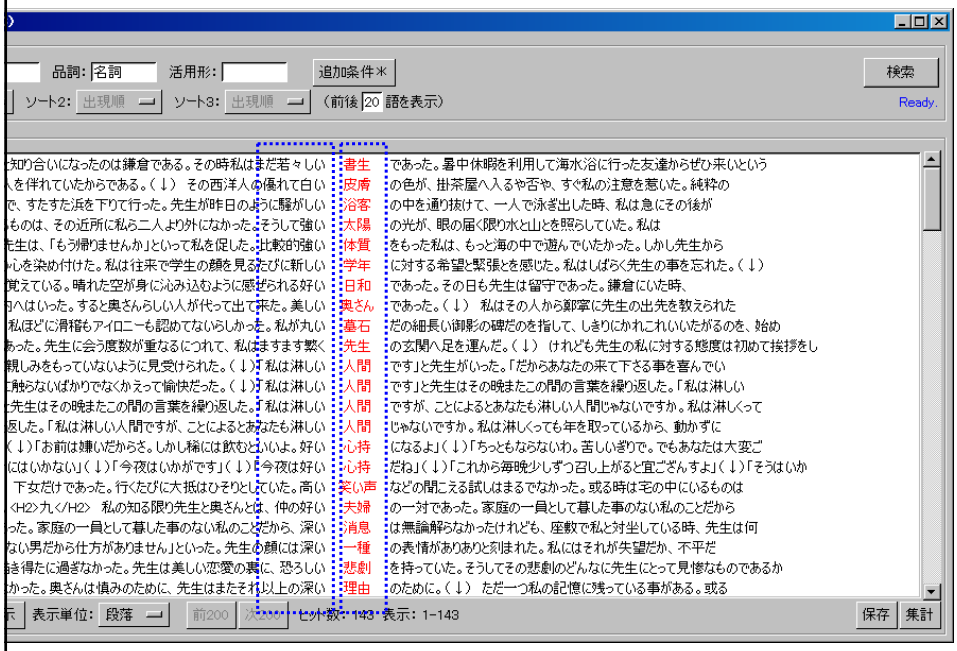
- 追加条件



パターン検索

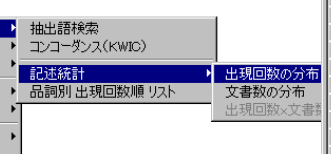
- 追加条件
  - 形容詞の連体修飾表現
  - やってみてください！





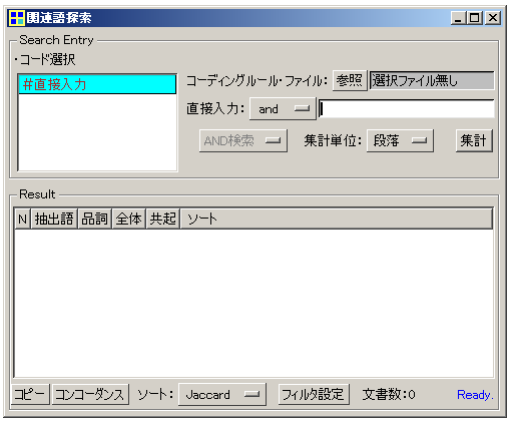
便利な機能

- 集計
  - ツール > 抽出語 >記述統計 > 出現回数の分布



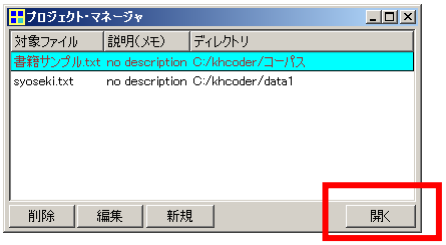
出現数	度数	パーセント	累積度数	累積パーセント
1	1334	52.58	1334	52.58
2	433	17.07	1767	69.65
3	209	8.24	1976	77.89
4	116	4.57	2092	82.46
5	68	2.68	2160	85.14
6	51	2.01	2211	87.15
7	36	1.42	2247	88.57
8	48	1.89	2295	90.46
9	22	0.87	2317	91.33
10	22	0.87	2339	92.20
11	18	0.71	2357	92.91
12	17	0.67	2374	93.58
13	11	0.43	2385	94.01
14	11	0.43	2396	94.44
15	8	0.32	2404	94.76
16	8	0.32	2412	95.07
17	8	0.32	2420	95.39
18	6	0.24	2426	95.62

- 関連語探索



ファイルを呼び出す

- プロジェクト > 開く

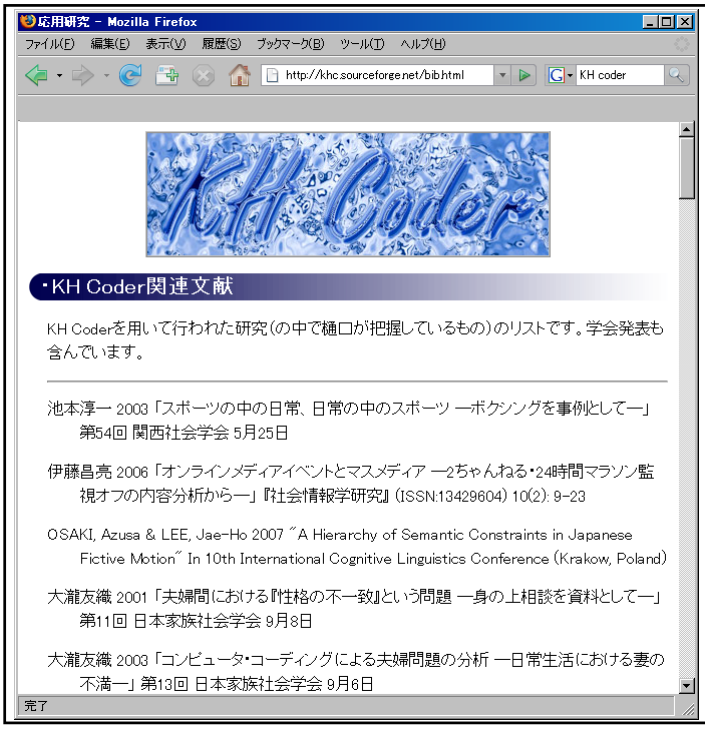


よくあるトラブル

- 一行が長すぎると処理が止まる（茶筌の問題）
- テキストの最初の一行目が空白だとエラーが出る。無駄な空白は消しておこう
- 文字コードは「Shift-JIS」「EUC」「JIS」のみ対応。Unicodeは非対応。Chasenの処理が止まる！

最後に

- 色々な分析機能がついている。Rによる統計処理なども可能だが、注意して使うべし
- テキストサイズ：無制限だが、大きいファイルを分析する場合は、十分なハードディスクの容量(100倍)と時間が要る
- KH Coderを使って論文を書いたら作成者に報告



- お疲れ様でした。
- ちょっと休憩＆質問お受けします。

2日目

2010/1/23-24
名古屋大学

日本語コーパスの最新の検索エンジンと分析

李在鎬（りじぇほ）
国際交流基金

JAPANFOUNDATION 国際交流基金

使い分ける

- Kwic Finder：長い単位に強い。
- KH Coder: 長い単位は検索が面倒。
  - お疲れ様。ご苦労様
  - 併用しましょう。

形態素解析のツールを使う

- 実習１：KH Coder
- 実習２：E-Kwic

開発背景

- 研究パラダイムのシフト
  - 誤用分析→中間言語：誤用のみならず、正用をも含めた包括的観点から学習者言語の実態を知ることの重要性が強調 （長友1993、長友1999、迫田2001、鎌田2006)
  - データの収集と体系的分析が必要：コーパス化(鎌田1999、大曽1999、宇佐美2005)

開発背景

- KYコーパス
  - OPIに基づく学習者コーパス
  - もっとも多くの研究において利用
  - フリーで利用可能
- 現状確認
  - 限定的調査：文字列を基盤にしたもの
  - 手作業による事例収集：ミスの可能性あり。計量的調査としての信頼性に関して必ずしも十分ではない

開発背景

- 具体的には、
  - 例1）文字列検索で助詞「は」を調査
    - 得られたデータが学習者のものかどうか？
    - 助詞以外のものが混ざっている
  - 例2）文字列検索で指示詞「あの」を調査
    - フィラーの場合と連体詞の場合はどうする？

→人が見て必要なもののみを抽出することも可 能だが、もっと効率よくいろんなデータを解析する

- 形態素解析技術を使うことに
- しかし、・・・。形態素解析器はあまり賢くない。

形態素解析器の問題1

- 例えば、

```
C:¥hcoder¥dep¥chasen¥chasen.exe
ご飯を食べている
ご飯    ゴハン  ご飯    名詞-一般
を      ヲ      を      助詞-格助詞-一般
食べ    タベ    食べる  動詞-自立       一段    連用形
て      テ      て      助詞-接続助詞
いる    イル    いる    動詞-非自立     一段    基本形
EOS
ご飯を食べてる
ご飯    ゴハン  ご飯    名詞-一般
を      ヲ      を      助詞-格助詞-一般
食べ    タベ    食べる  動詞-自立       一段    連用形
てる    テル    てる    動詞-非自立     一段    基本形
EOS
君の元へ走ってく
君      キミ    君      名詞-代名詞-一般
の      ノ      の      助詞-連体化
元      モト    元      名詞-一般
へ      へ      へ      助詞-格助詞-一般
走っ    ハシッ  走る    動詞-自立       五段・ラ行       連用タ接続
て      テ      て      助詞-接続助詞
く      ク      く      動詞-非自立     五段・カ行促音便        基本形
EOS
君の元へ走っていく
君      キミ    君      名詞-代名詞-一般
の      ノ      の      助詞-連体化
元      モト    元      名詞-一般
へ      へ      へ      助詞-格助詞-一般
走っ    ハシッ  走る    動詞-自立       五段・ラ行       連用タ接続
て      テ      て      助詞-接続助詞
いく    イク    いく    動詞-非自立     五段・カ行促音便        基本形
EOS
```

形態素解析器の問題2

- 例えば、

```
C:¥khcoder¥dep¥chasen¥chasen.exe
EOS
それがね、とてもおいしいの
それ    ソレ    それ    名詞-代名詞-一般
が      ガ      が      助詞-格助詞-一般
ね      ネ      ねる    動詞-自立       一段    連用形
、      、      、      記号-読点
とても  トテモ  とても  副詞-助詞類接続
おいしい        オイシイ        おいしい        形容詞-自立     形容詞・イ段
基本形
の      ノ      の      名詞-非自立-一般
EOS
それがですね
それ    ソレ    それ    名詞-代名詞-一般
が      ガ      が      接続詞
です    デス    です    助動詞  特殊・デス      基本形
ね      ネ      ね      助詞-終助詞
EOS
それをですね
それ    ソレ    それ    名詞-代名詞-一般
を      ヲ      を      助詞-格助詞-一般
です    デス    です    助動詞  特殊・デス      基本形
ね      ネ      ね      助詞-終助詞
EOS
```

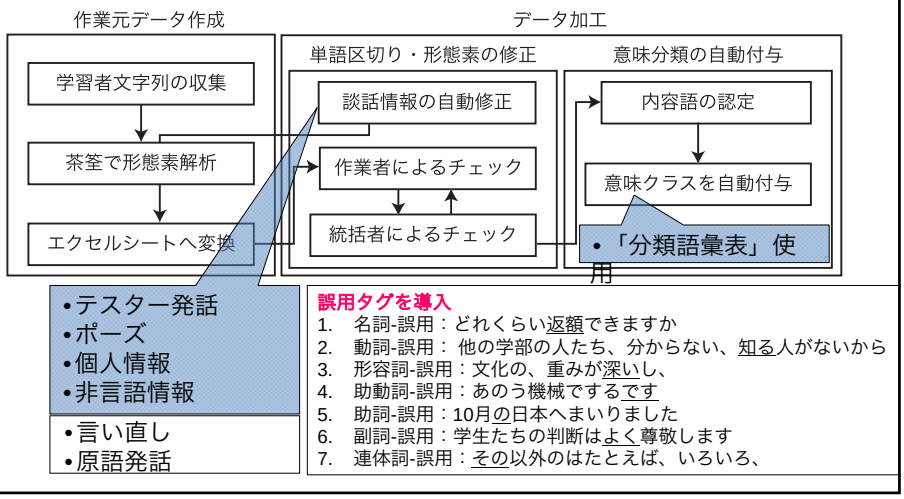
形態素解析器の問題3

- 例えば、

```
C:¥Program Files¥ChaSen¥chasen.exe
早い    ハヤイ  早い    形容詞-自立     形容詞・アウオ段        基本形
の      ノ      の      名詞-非自立-一般
ほう    ホウ    ほう    名詞-一般
が      ガ      が      助詞-格助詞-一般
良い    ヨイ    良い    形容詞-自立     形容詞・アウオ段        基本形
じゃ    ジャ    じゃ    助詞-副助詞
ない    ナイ    ない    助動詞  特殊・ナイ      基本形
の      ノ      の      助詞-終助詞
EOS
たの、楽しい、
た      タ      た      助動詞  特殊・タ        基本形
の      ノ      の      名詞-非自立-一般
、      、      、      記号-読点
楽しい  タノシイ        楽しい  形容詞-自立     形容詞・イ段    基本形
、      、      、      記号-読点
EOS
ですくそうですねで、
で      デ      で      接続詞
す      ス      す      名詞-一般
く      く      く      記号-一般
そうですね      ソウデスネ      そうですね      感動詞
>       >       >       記号-一般
で      デ      だ      助動詞  特殊・ダ        連用形
、      、      、      記号-読点
EOS
```

データ加工

- 人手によるチェックと修正を行う



- テスター発話
- ポーズ
- 個人情報
- 非言語情報

- 言い直し
- 原語発話

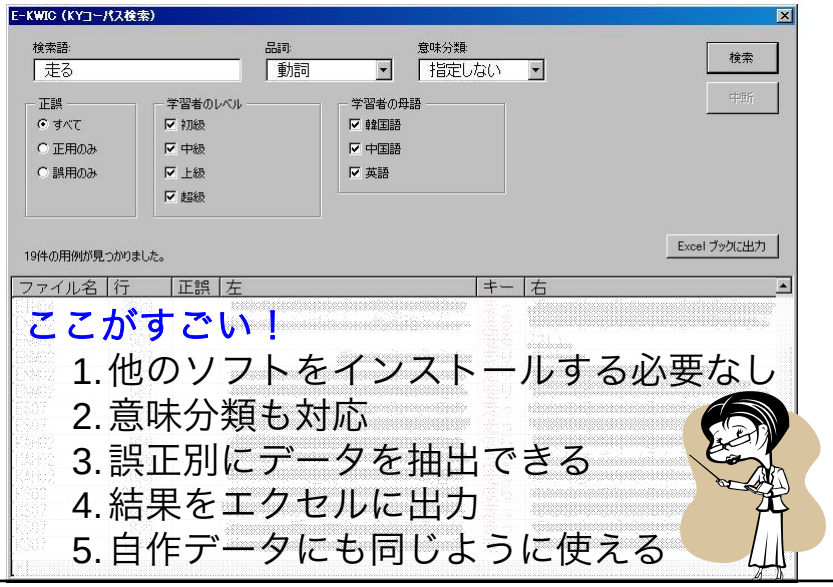
誤用タグを導入

1. 名詞-誤用：どれくらい返額できますか
2. 動詞-誤用： 他の学部の人たち、分からない、知る人がないから
3. 形容詞-誤用：文化の、重みが深いし、
4. 助動詞-誤用：あのう機械ですることです
5. 助詞-誤用：10月の日本へまいりました
6. 副詞-誤用：学生たちの判断はよく尊敬します
7. 連体詞-誤用：その以外のはたとえば、いろいろ、

検索ツールの開発

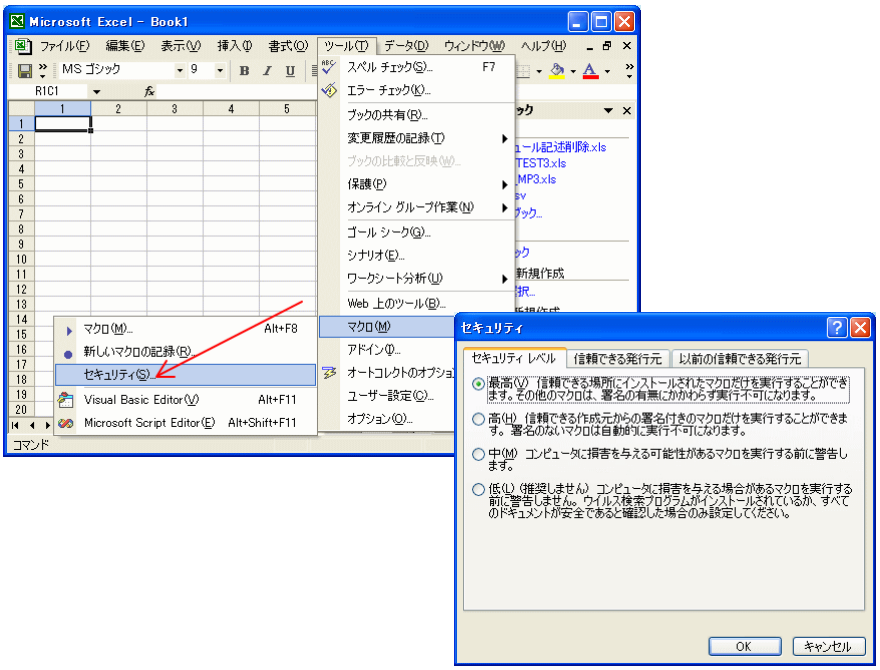
- 既存のツールではうまく動作しない
- 既存のツールはインストールが難しい
- 高機能でなくても良いので、使いやすいツールを作る：エクセルで動作するツールを作成

E-KWIC

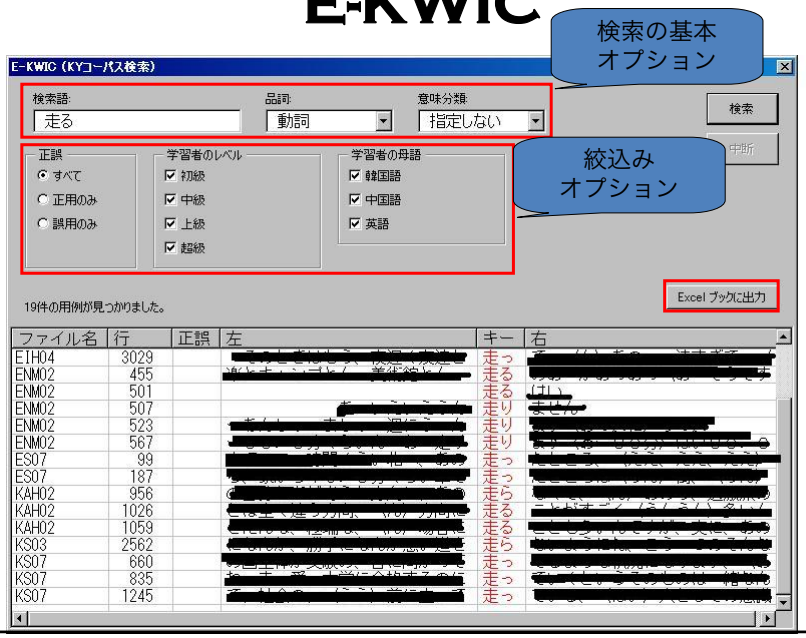


ここがすごい！

1. 他のソフトをインストールする必要なし
2. 意味分類も対応
3. 誤正別にデータを抽出できる
4. 結果をエクセルに出力
5. 自作データにも同じように使える



E-KWIC



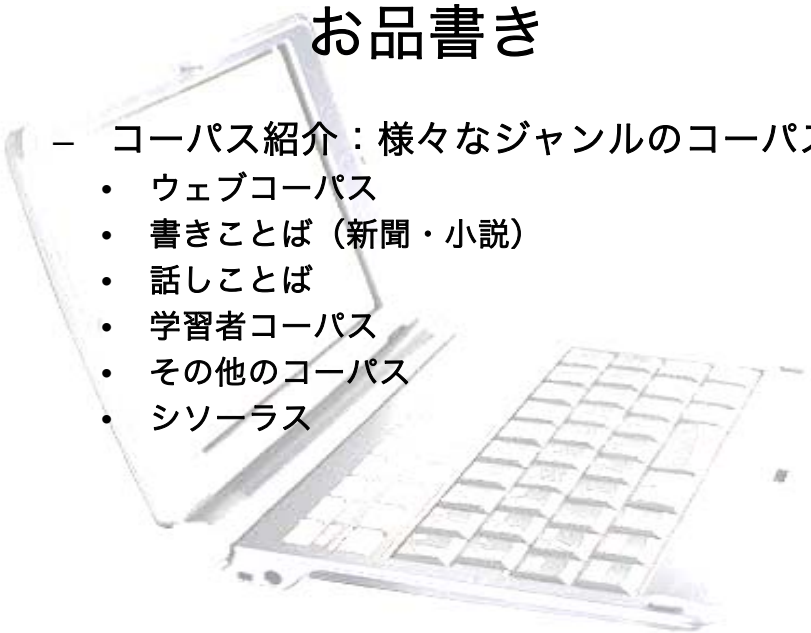
E-KWIC

- dataフォルダ内の学習者データに対して以下の基本検索ができる
  1. 形態素の検索
  2. 形態素と品詞の組み合わせによる検索
  3. 形態素と品詞と意味分類の組み合わせによる検索
  4. 品詞のみの検索
  5. 意味分類のみの検索

E-KWIC

- 絞り込みによる検索
  1. 形態素の検索
  2. 形態素と品詞の組み合わせによる検索
  3. 形態素と品詞と意味分類の組み合わせによる検索
  4. 品詞のみの検索
  5. 意味分類のみの検索

お品書き



- コーパス紹介：様々なジャンルのコーパス
  - ウェブコーパス
  - 書きことば（新聞・小説）
  - 話しことば
  - 学習者コーパス
  - その他のコーパス
  - シソーラス

ウェブ（基盤の）コーパス

1. 位置づけ
   - ウェブ上のデータをコーパス化したもの
   - 自然言語処理の分野で利用。ブログなど

   * ネット検索とは異なることに注意

2. 短所と長所
   - 長所：大規模でかつ、著作権フリー、簡単(ウェブブラウザ上で使えるため)、無料
   - 短所：書き手が不明,メンテなし,ノイズ多し

→ **長所を生かし、(文法研究に特化した)コロケーションを見るなどの目的においては有効**

コーパス紹介

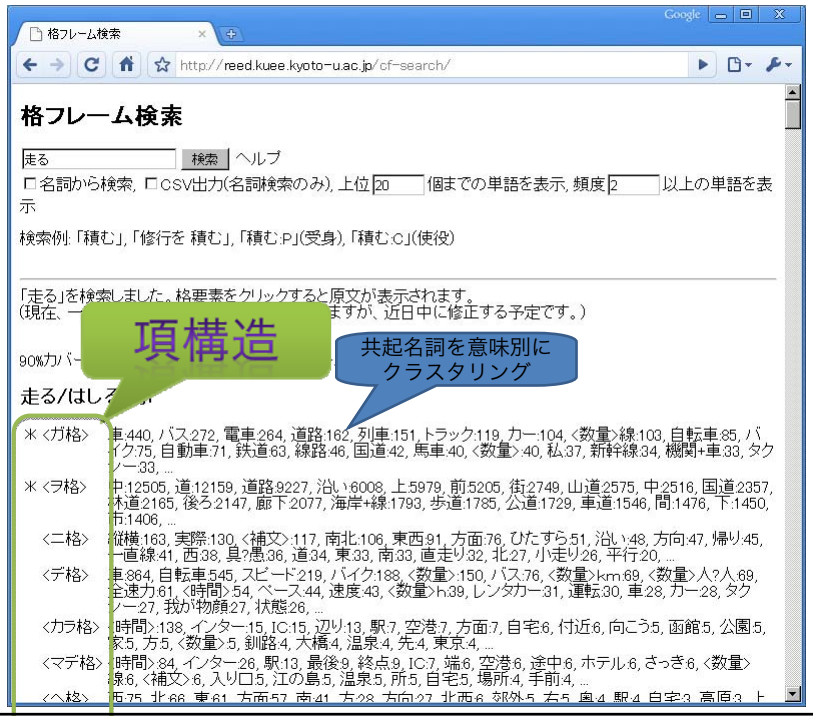
3. 利用可能な資源

ウェブ検索ではないので、見た日によって結果がかわることはない

	格フレーム辞書	Sketch engine	ウィキペディア
内容	表層形としての述語と項を関係づけたデータベース	イギリスで作成された多言語データベース	ウィキペディアのデータをテキスト化し、コーパスとして利用
規模	約5億文 約5万用言	約4億語 (409,384,405語)	時期によって 変動あり
下処理 (コーパス化)	不要 形態素・係受解析済	不要 形態素解析済	要 (テキスト化と形態素解析)
利用形態	ウェブブラウザ 無料	ウェブブラウザ 有料(年間1万円)	ダウンロード後 整形(データは無料)
備考		シソーラス搭載	固有名を多数含んでいる、特殊なデータ
元データ	条件付きアクセス可	アクセス不可	アクセス可能

http://reed.kuee.kyoto-u.ac.jp/cf-search/

動詞による検索：格パターンと共起名詞
名詞による検索：共起する動詞



格フレームID	格	頻度
行う/おこなう?行く/いく:動1	ニ格	2011
住む/すむ:動1	ニ格	1557
来る/くる:動2	ニ格	1395
帰る/かえる:動1	ニ格	1357
戻る/もどる:動5	ニ格	829
行く/いく:動4	ヘ格	716
行う/おこなう?行く/いく:動5	ヘ格	674
着く/つく:動1	ニ格	628
開催/かいさい:動:P1	デ格	621
居る/いる?射る/いる?鋳る/いる:動13	ニ格	555
代表/だいひょう:動1	ヲ格	417
居る/いる?射る/いる?煎る/いる?鋳る/いる?要る/いる:動16	ニ格	406
戻る/もどる:動11	ヘ格	376
住む/すむ+る/る:動1	ニ格	337
到着/とうちゃく:動1	ニ格	336
開催/かいさい:動1	デ格	266

cf-search.csv - Microsoft Excel

A1 格フレームID

	A	B	C
1	格フレームID	格	頻度
2	行う/おこな	ニ格	2011
3	住む/すむ:1	ニ格	1557
4	来る/くる:動	ニ格	1395
5	帰る/かえ1	ニ格	1357
6	戻る/もどる	ニ格	829
7	行く/いく:動	ヘ格	716
8	行う/おこな	ヘ格	674
9	着く/つく:動	ニ格	628
10	開催/かい:デ	格	621
11	居る/いる?1	ニ格	555
12	代表/だい:2	ヲ格	417
13	居る/いる?1	ニ格	406
14	戻る/もどる	ヘ格	376
15	住む/すむ:1	ニ格	337
16	到着/とうち	ニ格	336
17	開催/かい:デ	格	266
18	出発/しゅっ	ヲ格	254
19	行く/いく:動	ニ格	239
20	時間/じか:1	カラ格	226
21	引っ越す/ひ	ニ格	212
22	行う/おこな	マデ格	186
23	出る/でる:1	ニ格	186
24	居る/いる:1	ニ格	175
25	着る/きる?1	ニ格	175
26	時間/じか:1	カラ格	168
27	出張/しゅっ	ニ格	163
28	分/ふん:判	カラ格	151
29	働く/はたら	デ格	125

http://www.sketchengine.co.uk/

LEX COM Lexical Computing

Sketch Engine

user: leejaeho

Help Change passwd WebBootCaT CorpusBuilder Bug reporting News Logout

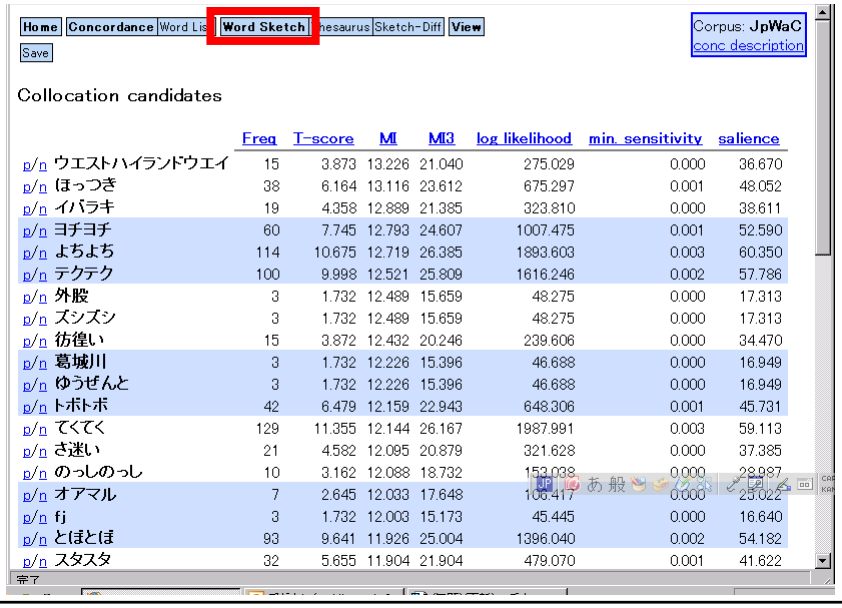
Corpora

Language	Name	Tokens [?]	
Chinese	Chinese GW, simpl	706 427 624	info
Chinese	Chinese GW, trd	706 428 333	info
English	British Academic Spoken English Corpus (BASE)	1 252 256	info
English	British Academic Written English Corpus (BAWE)	7 474 757	info
English	British National Corpus	111 244 375	info
English	ukWaC	2 035 621 120	info
French	French web corpus	126 850 281	info
German	deWaC	1 644 785 836	info
Greek	GkWaC	149 067 023	info
Italian	itWaC	1 909 535 984	info
Japanese	JpWaC	409 384 405	info
Persian	WBC-Per	6 375 735	info
Portuguese	Cetenfolha, Cetempublico	66 319 147	info
Russian	Russian Web Corpus	187 965 822	info
Slovenian	Fida PLUS 620m	738 503 185	info
Spanish	Spanish web corpus	116 900 060	info

<< less corpora

レンマレベルで検索し、用法を確認

共起語を調べる。様々な共起指標を表示



		Freq	T-score	MI	MI3	log likelihood	min. sensitivity	salience
p/n	ウエストハイランドウエイ	15	3.873	13.226	21.040	275.029	0.000	36.670
p/n	ほっつき	38	6.164	13.116	23.612	675.297	0.001	48.052
p/n	イバラキ	19	4.358	12.889	21.385	323.810	0.000	38.611
p/n	ヨチヨチ	60	7.745	12.793	24.607	1007.475	0.001	52.590
p/n	よちよち	114	10.675	12.719	26.385	1893.603	0.003	60.350
p/n	テクテク	100	9.998	12.521	25.809	1616.246	0.002	57.786
p/n	外股	3	1.732	12.489	15.659	48.275	0.000	17.313
p/n	ズシズシ	3	1.732	12.489	15.659	48.275	0.000	17.313
p/n	彷徨い	15	3.872	12.432	20.246	239.606	0.000	34.470
p/n	葛城川	3	1.732	12.226	15.396	46.688	0.000	16.949
p/n	ゆうぜんと	3	1.732	12.226	15.396	46.688	0.000	16.949
p/n	トボトボ	42	6.479	12.159	22.943	648.306	0.001	45.731
p/n	てくてく	129	11.355	12.144	26.167	1987.991	0.003	59.113
p/n	さ迷い	21	4.582	12.095	20.879	321.628	0.000	37.385
p/n	のっしのっし	10	3.162	12.088	18.732	153.038	0.000	28.987
p/n	オアマル	7	2.645	12.033	17.648	106.417	0.000	23.022
p/n	fj	3	1.732	12.003	15.173	45.445	0.000	16.640
p/n	とぼとぼ	93	9.641	11.926	25.004	1396.040	0.002	54.182
p/n	スタスタ	32	5.655	11.904	21.904	479.070	0.001	41.622

シソーラスとしての使う：(生起文脈に基づく)類義語を調べる

Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff

Save

歩く JpWaC freq = 42732

走る	0.448
乗る	0.361
行く	0.361
通る	0.357
帰る	0.356
見る	0.344
探す	0.332
飲む	0.332
立つ	0.33
読む	0.33
思う	0.329
出る	0.325
話す	0.321
入る	0.32
言う	0.32
食べる	0.319
来る	0.319
戻る	0.318
みる	0.318
進む	0.313

似たような名詞や構文パターンが見られた場合、類義語と判定

語彙の類似性・非類似性を調べる

Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff

Word Sketch Differences Entry Form

Corpus:	JpWaC
First lemma:	歩く
Second lemma:	乗る
Sort grammatical relations:	☑
Separate blocks:	○ all in one block ◉ common/exclusive blocks
Minimum frequency:	5
Maximum number of items in a grammatical relation of the common block:	25
Maximum number of items in a grammatical relation of the exclusive block:	12

Show Diff

歩く/乗る

preloaded/jpwac freq = 42732/37665

change options

Common patterns

歩く 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 乗る

nounまで	1829	444	11.4	2.9
駅	212	33	7.2	4.6
ホテル	29	9	5.1	3.5
どこ	26	18	3.4	2.9
いつ	13	6	3.4	2.3
そこ	21	7	2.6	1.0
ここ	29	11	2.6	1.2
今	30	34	2.2	2.4
最後	12	7	2.3	1.6

nounに	2867	26425	1.1	10.2
一緒	372	193	7.9	6.4
片手	43	7	7.9	3.0
いっしょ	46	18	7.4	4.2
久しぶり	38	54	6.4	5.5
風	7	148	2.2	6.0
久し振り	7	8	5.9	3.3
後ろ	5	73	3.2	5.8
帰り	24	33	5.8	4.8
方向	91	11	5.7	2.2
久々	11	32	5.5	5.0
為	36	11	5.4	2.8
背	5	28	4.3	4.8
とき	28	89	2.8	4.3
絶対	7	24	3.5	4.2
夜	22	11	4.1	2.5
こっち	7	5	4.0	2.1
前	49	117	2.8	3.9

nounへ	900	256	7.3	2.1
駅	23	5	4.1	1.9
学校	7	6	1.6	1.4
どこ	6	6	1.3	1.3
そこ	7	6	1.0	0.8

ある/いる

preloaded/jpwac freq = 2954750/3574782

緑：「ある」の共起語
赤：「いる」の共起語

Common patterns

ある 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 いる

nounが	763890	90712	12.7	0.5
こと	69993	219	9.7	1.8
もの	25686	506	9.3	4.6
人	1276	19826	4.6	9.2
性	20073	20	9.0	0.2
ところ	9462	35	8.2	1.7
の	12026	962	7.9	5.0
子供	77	1350	1.5	7.9
事	6630	23	7.9	1.5
効果	5218	5	7.7	0.1
興味	4975	11	7.7	1.5
子	72	868	1.5	7.6
機会	4571	9	7.5	1.1
人間	143	1174	2.4	7.5
感	4768	10	7.5	0.7
奴	33	507	0.5	7.4
友達	25	606	0.0	7.4
部分	4371	15	7.4	1.3

V_bound	54322	4345340	0.5	12.3
する	8367	1048404	4.5	11.0
れる	1562	418244	4.1	10.9
書く	15116	38530	10.1	8.1
なる	652	196557	2.8	9.8
置く	3275	8007	9.3	5.9
思う	90	99612	1.1	9.2
られる	333	75023	3.8	8.9
持つ	100	60775	2.7	8.7
言う	323	58503	3.7	8.6
考える	113	56541	2.5	8.5
貼る	852	1141	8.5	3.1
やる	134	45141	2.9	8.2
見る	168	45040	2.9	8.2
知る	63	38965	2.5	8.1
行う	36	37671	1.2	8.0
使う	188	35140	4.0	7.9
くる	120	36376	2.1	7.8

nounは	182575	37481	5.0	0.3
の	11800	1732	8.5	5.9
こと	21324	431	8.3	2.8
これ	4531	373	7.9	4.6
人	1257	7311	5.1	7.9
理由	1339	32	7.5	3.4
事	1859	49	7.4	2.9
それ	3307	347	7.3	4.4
もの	3392	547	7.2	4.9
奴	17	203	1.5	7.2
私	3178	1230	7.1	6.1
問題	2145	43	7.1	1.9
今日	1072	266	7.0	6.1
方法	1075	8	7.0	1.0
ところ	1555	40	6.9	2.2
原因	682	10	6.7	2.2
人間	201	460	4.5	6.6
僕	413	313	5.7	6.5

見る/見える preloaded/jpwac freq = 404633/85244

change options

Common patterns

見る 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 見える

nounが	27716	27534	1.5	8.1
姿	391	943	7.2	8.4
傾向	464	64	8.2	5.3
顔	186	696	5.9	7.8
兆し	37	178	5.3	7.6
星	8	177	2.6	7.1
景色	18	132	4.1	7.0
海	38	201	4.4	6.8
変化	257	43	6.8	4.2
光	10	183	2.6	6.8
先	7	392	0.9	6.7
動き	217	99	6.6	5.5
光景	93	31	6.5	4.9
様子	119	154	6.0	6.4
影	6	92	2.4	6.3
現象	107	11	6.3	3.0
周り	13	119	3.1	6.3
僕	246	11	6.3	1.8

nounから	25405	4197	6.3	5.8
観点	920	6	9.3	2.9
窓	296	376	7.8	9.1
車窓	59	69	6.2	8.8
視点	649	26	8.6	4.6
遠く	265	54	8.1	7.4
角度	268	16	8.0	5.4
外	681	284	7.9	7.1
立場	587	15	7.9	3.1
隙間	24	50	4.8	7.7
傍	171	9	7.7	5.5
正面	165	12	7.5	5.4
側	724	35	7.4	3.2
はた	129	5	7.3	5.1
端	153	8	7.2	4.3
横	171	9	7.0	3.8
他人	151	19	6.8	4.7
ベランダ	19	18	4.5	6.6

modifier_Adv	28139	5215	5.1	5.3
よく	3057	298	10.9	8.4
初めて	1440	32	10.2	6.0
はっきり	145	557	6.6	9.4
実際	685	18	9.1	5.1
どう	2378	163	9.1	5.5
まだ	771	153	9.0	7.6
あまり	638	83	8.8	6.8
ほとんど	578	154	8.8	8.0
ちょっと	662	74	8.6	6.2
ずっと	431	15	8.5	5.0
もっと	410	17	8.4	5.0
ちゃんと	413	85	8.4	7.3
全く	242	210	7.5	8.3
しっかり	409	48	8.2	6.1
まったく	142	136	6.9	8.2
ふと	269	6	8.1	4.4
いつも	306	18	8.0	5.3

歩く/乗る preloaded/jpwac freq = 42732/37665

change options

Common patterns

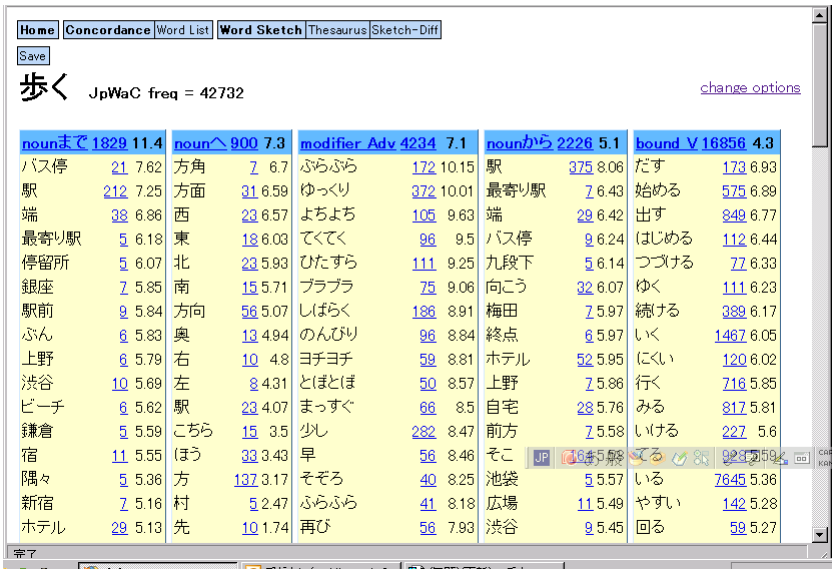
歩く 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 乗る

nounまで	1829	444	11.4	2.9
駅	212	33	7.2	4.6
ホテル	29	9	5.1	3.5
どこ	26	18	3.4	2.9
いつ	13	6	3.4	2.3
そこ	21	7	2.6	1.0
ここ	29	11	2.6	1.2
今	30	34	2.2	2.4
最後	12	7	2.3	1.6

nounに	2867	26425	1.1	10.2
一緒	372	193	7.9	6.4
片手	43	7	7.9	3.0
いっしょ	46	18	7.4	4.2
久しぶり	38	54	6.4	5.5
風	7	148	2.2	6.0
久し振り	7	8	5.9	3.3
後ろ	5	73	3.2	5.8
帰り	24	33	5.8	4.8
方向	91	11	5.7	2.2
久々	11	32	5.5	5.0
為	36	11	5.4	2.8
背	5	28	4.3	4.8
とき	28	89	2.8	4.3
絶対	7	24	3.5	4.2
夜	22	11	4.1	2.5
こっち	7	5	4.0	2.1
前	49	117	2.8	3.9

nounへ	900	256	7.3	2.1
駅	23	5	4.1	1.9
学校	7	6	1.6	1.4
どこ	6	6	1.3	1.3
そこ	7	6	1.0	0.8

共起語をまとめる。格フレーム辞書同様



Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff

Save

歩く JpWaC freq = 42732

change options

nounまで 1829 11.4			noun へ 900 7.3			modifier Adv 4234 7.1			nounから 2226 5.1			bound V 16856 4.3		
バス停	21	7.62	方角	7	6.7	ぶらぶら	172	10.15	駅	375	8.06	だす	173	6.93
駅	212	7.25	方面	31	6.59	ゆっくり	372	10.01	最寄り駅	7	6.43	始める	575	6.89
端	38	6.86	西	23	6.57	よちよち	105	9.63	端	29	6.42	出す	849	6.77
最寄り駅	5	6.18	東	18	6.03	てくてく	96	9.5	バス停	9	6.24	はじめる	112	6.44
停留所	5	6.07	北	23	5.93	ひたすら	111	9.25	九段下	5	6.14	つづける	77	6.33
銀座	7	5.85	南	15	5.71	ブラブラ	75	9.06	向こう	32	6.07	ゆく	111	6.23
駅前	9	5.84	方向	56	5.07	しばらく	186	8.91	梅田	7	5.97	続ける	389	6.17
ぶん	6	5.83	奥	13	4.94	のんびり	96	8.84	終点	6	5.97	いく	1467	6.05
上野	6	5.79	右	10	4.8	ヨチヨチ	59	8.81	ホテル	52	5.95	にくい	120	6.02
渋谷	10	5.69	左	8	4.31	とぼとぼ	50	8.57	上野	7	5.86	行く	716	5.85
ビーチ	6	5.62	駅	23	4.07	まっすぐ	66	8.5	自宅	28	5.76	みる	817	5.81
鎌倉	5	5.59	こちら	15	3.5	少し	282	8.47	前方	7	5.58	いける	227	5.6
宿	11	5.55	ほう	33	3.43	早	56	8.46	そこ	16	5.58	くる	928	5.59
隅々	5	5.36	方	137	3.17	そぞろ	40	8.25	池袋	5	5.57	いる	7645	5.36
新宿	7	5.16	村	5	2.47	ふらふら	41	8.18	広場	11	5.49	やすい	142	5.28
ホテル	29	5.13	先	10	1.74	再び	56	7.93	渋谷	9	5.45	回る	58	5.27

Wikipediaのコーパス化

書きことばコーパス

1. 書きことばコーパスの現状と位置づけ

- 新聞、小説など複数のジャンルのテキストあり
- 代表性の問題をクリアするのが難しい
- 利用には制限が多いため、要確認

2. 短所と長所

- 長所：書き手・出典が明確、表現の規範性が高い
- 短所：手間・コスト高、データのバイアス強

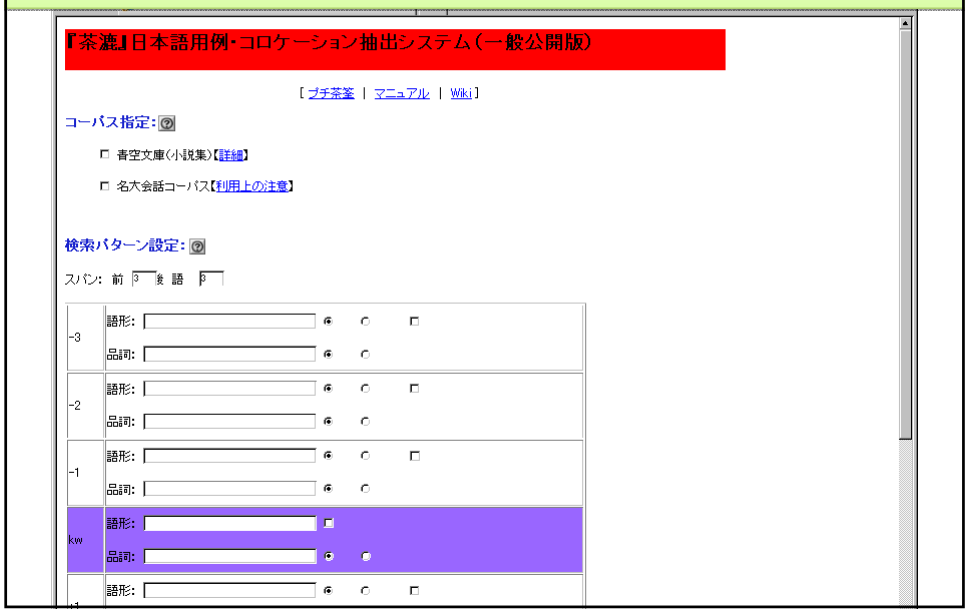
→ **複数のデータを比較評価しながら利用するのが望ましい**

コーパスの紹介

1. 書き言葉（1）小説

	青空文庫	新潮文庫の100冊
内容	ネット上の電子図書館『青空文庫』の公開作品を一枚に収録	新潮文庫で絶版となった、入手不可能な100作品を収録
入手	『インターネット図書館　青空文庫』を購入するとDVD-ROM(¥1,500)	CD-ROMを購入（¥2,600)
コーパス化	「青空文庫」変換スクリプトパッケージ（山口@国研）	エキスパンドブックテキスト化ツールをウェブでDL
その他	茶漉による青空文庫の検索	

http://tell.fll.purdue.edu/chakoshi-wiki/



コーパスの紹介

1. 書き言葉（2）新聞

	新聞記事データベース	日英新聞記事対応付けデータ	京都テキストコーパス
内容	毎日、朝日、読売新聞記事のデータ版	読売新聞と The Daily Yomiuri から自動作成された日英対応付けコーパス(NICT、内山氏)	毎日新聞1995年度版の記事と社説に各種言語情報を人手で付与したテキストコーパス（4万文）
入手	利用許諾契約後、研究目的での利用	ウェブから検索。フリーで利用可	ウェブよりダウンロード。フリーで利用可
コーパス化	テキストデータ	自分で加工する必要あり	毎日新聞データは別に購入し、対応づけが必要
その他	高価なため個人での購入は向かない		茶器で使用可能

http://mastarpj.nict.go.jp/~mutiyama/jea/index-ja.html

English page

日英新聞記事対応付けデータ（JENAAD）

―― 読売新聞と The Daily Yomiuri から自動作成された日英対応付けコーパス ――

文対応のサンプル

- 一対一対応の日英文 (shift_jis-dos)
- 一対多もしくは多対一対応の日英文 (shift_jis-dos)

利用に際しては覚書提出
規模：180万(文)対
日本語は約400万形態素

覚書

日英新聞記事対応付けデータをダウンロードするには，覚書(日本語版/
で郵送して下さい．

〒619-0288　「けいはんな学研都市」光台2-2-2
(独) 情報通信研究機構　知識創成コミュニケーション研究センター
言語翻訳グループ　内山 将夫

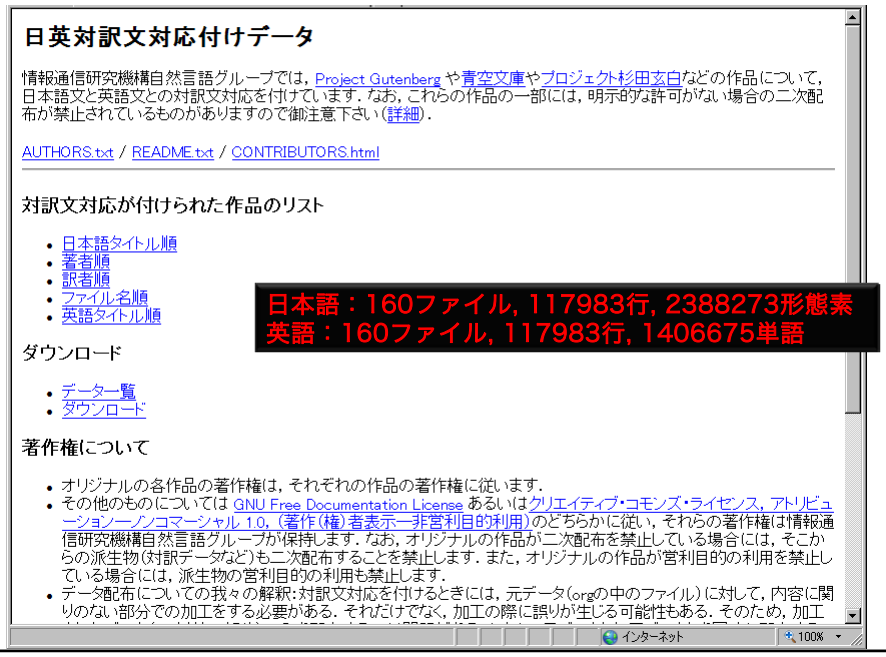
覚書が届きしだい，アクセスのためのアカウントとパスワードをメールします．なお，アカウントは一定期間が経過したら削除しますので御注意下さい．

ダウンロード

- p11-1989-2001.txt.gz (150000 一対一文対応, 23118450バイト, euc-jp-unix)
- p11-1989-2001.zip (150000 一対一文対応, 23335741バイト, shift_jis-dos)
- pnm-1989-2001.txt.gz (30000 一対多文対応, 7417253バイト, euc-jp-unix)
- pnm-1989-2001.zip (30000 一対多文対応, 7468055バイト, shift_jis-dos)
  pnm-1989-2001＊の78対応は，日英文のどちらかもしくは双方が欠けています．
- split.zip (22919410バイト) には以下のファイルがあります．
  - p11-1989-2001-j.txt (日本語 150000行 , 19751012バイト, shift_jis-dos)
  - p11-1989-2001-e.txt (日本語 150000行 , 22522594バイト, ascii-dos)
  - pnm-1989-2001-j.txt (英語 29922行 , 6458527バイト, shift_jis-dos)

```
<T ID="19930709J1TYEUB0400010/19930710E1TDY03C000010/21" NM="1-1" SCORE="1.710796165">
<J>欧州は、エディンバラにおいて合意され、コペンハーゲンにおいて強化された成長イニシアチブを精力的に実行しつ
<E>Europe is carrying out vigorously the Growth Initiative agreed in Edinburgh and strengthened in Copenhage
</T>
<T ID="19930709J1TYEUB0400010/19930710E1TDY03C000010/71" NM="1-1" SCORE="1.539716548">
<J>我々は、ロシアの経済発展にとって、改善された市場アクセスが重要であることを認識する。</J>
<E>We recognize the importance of improved market access for economic progress in Russia. </E>
</T>
<T ID="19920709J1TYMCD1300010/19920709E1TDY09C000040/158" NM="1-1" SCORE="1.347403181">
<J>法人レベルでのパートナーシップ及びマネージメント支援は、特に効果的であり得る。</J>
<E>Partnerships and management assistance at corporate level can be particularly effective. </E>
</T>
<T ID="20000503J1TYMCA1200020/20000503E1TDY15B000010/10" NM="1-1" SCORE="1.31897521">
<J>違憲の問題については、連邦憲法裁判所が決定する。</J>
<E>The Federal Constitutional Court decides on the question of unconstitutionality. </E>
</T>
<T ID="19920629J1TYMAA1400090/19920629E1TDY06C000010/61" NM="1-1" SCORE="1.315772446">
<J>統合の実際のプロセスは、経済分野から始めねばならない。</J>
<E>The practical process of integration must begin in the economic sphere. </E>
</T>
<T ID="19930708J1TYEUB0400090/19930709E1TDY03C000020/36" NM="1-1" SCORE="1.308626617">
<J>関連する安全保障理事会決議の諸条件が満たされるまで、制裁は維持されるべきである。</J>
<E>Sanctions should be upheld until the conditions in the relevant Security Council resolutions are met. </E
</T>
<T ID="20010928J1TYMAC1400030/20010928E1TDY08D000060/9" NM="1-1" SCORE="1.30335977">
<J>国際テロは、世界の平和と安全に対する重大な脅威だ。</J>
<E>International terrorism is a grave threat to world peace and security. </E>
</T>
<T ID="19920709J1TYMCD1300010/19920709E1TDY09C000040/87" NM="1-1" SCORE="1.251160096">
<J>貧困、人口政策、教育、保健、女性の役割、及び児童の福祉は、特別の注意に値する。</J>
<E>Poverty, population policy, education, health, the role of women and the well-being of children merit spe
</T>
<T ID="19930416J1TYMAD1400020/19930416E1TDY03A000020/33" NM="1-1" SCORE="1.213566872">
<J>国際市場に対するロシア産品のアクセス改善は、ロシアの構造改革を大いに強化する。</J>
<E> Improvement of access for Russian products to international markets strongly reinforces Russian structur
</T>
<T ID="19931013J1TYEUB0400410/19931014E1TDY12C000010/6" NM="1-1" SCORE="1.16174498">
<J>２　日本国政府及びロシア連邦政府は、市場経済への移行を目指すロシア連邦の努力に対する支援に当たって、戦後の
<E>2. The Government of Japan and the Government of the Russian Federation share the perception that Japan's
</T>
<T ID="19950709J1TYMAA1400040/19950709E1TDY05C000010/43" NM="1-1" SCORE="1.126793067">
<J>また若い世代に自由の必要性を教えこまなければならない。</J>
<E>He must teach the younger generations the need for the freedom. </E>
```

http://mastarpj.nict.go.jp/~mutiyama/align/index.html



日英対訳文対応付けデータ

情報通信研究機構自然言語グループでは，Project Gutenberg や青空文庫やプロジェクト杉田玄白などの作品について，日本語文と英語文との対訳文対応を付けています．なお，これらの作品の一部には，明示的な許可がない場合の二次配布が禁止されているものがありますので御注意下さい(詳細)．

AUTHORS.txt / README.txt / CONTRIBUTORS.html

対訳文対応が付けられた作品のリスト

- 日本語タイトル順
- 著者順
- 訳者順
- ファイル名順
- 英語タイトル順

日本語：160ファイル，117983行，2388273形態素
英語：160ファイル，117983行，1406675単語

ダウンロード

- データ一覧
- ダウンロード

著作権について

- オリジナルの各作品の著作権は，それぞれの作品の著作権に従います．
- その他のものについては GNU Free Documentation License あるいはクリエイティブ・コモンズ・ライセンス，アトリビューション―ノンコマーシャル 1.0，(著作(権)者表示―非営利目的利用)のどちらかに従い，それらの著作権は情報通信研究機構自然言語グループが保持します．なお，オリジナルの作品が二次配布を禁止している場合には，そこからの派生物(対訳データなど)も二次配布することを禁止します．また，オリジナルの作品が営利目的の利用を禁止している場合には，派生物の営利目的の利用も禁止します．
- データ配布についての我々の解釈:対訳文対応を付けるときには，元データ(orgの中のファイル)に対して，内容に関りのない部分での加工をする必要がある．それだけでなく，加工の際に誤りが生じる可能性もある．そのため，加工

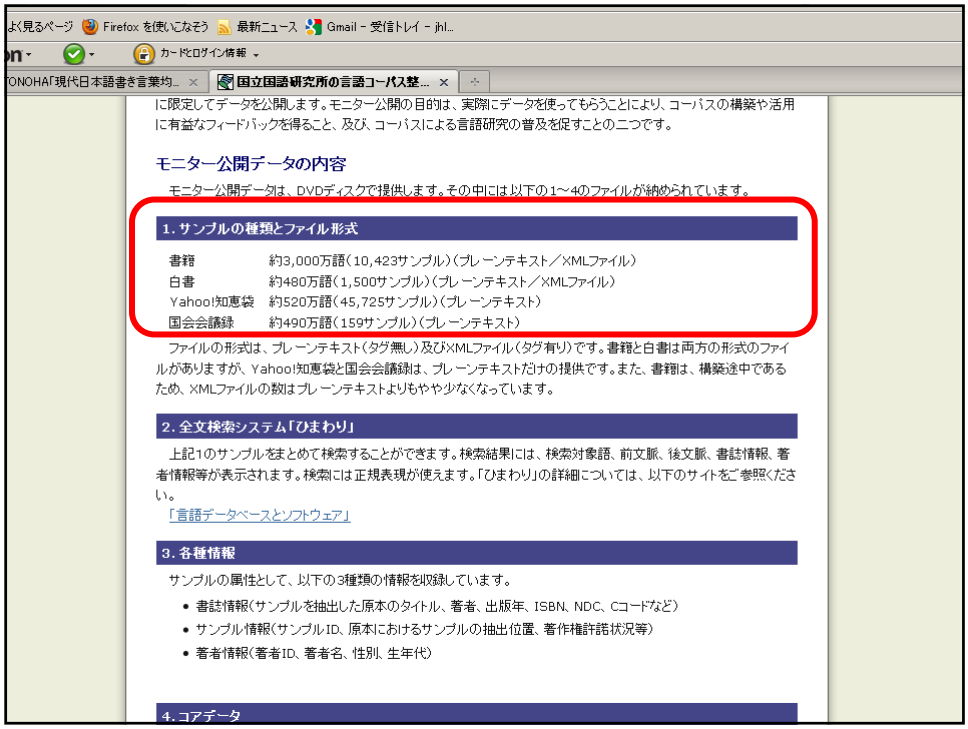
コーパスの紹介

1. 書き言葉（3）その他のデータベース

- テキスト検索専用のデータベース
  1. 国会会議録検索システム
  2. 判例検索システム
  3. 日本語書き言葉均衡コーパスモニター公開
     （BCCWJ）

http://www30.atwiki.jp/corpus-ling/

KOTONOHA「現代日本語書き言葉均衡コーパス」 検索デモンストレーション - 検索条件 - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.kotonoha.gr.jp/cgi-bin//search_form.cgi?viaTopPage=1
よく見るページ Firefox を使いこなそう 最新ニュース Gmail - 受信トレイ - jhl...
Norton
カードとログイン情報
KOTONOHA「現代日本語書き言葉均...
独立行政法人 国立国語研究所
KOTONOHA
「現代日本語書き言葉均衡コーパス」
検索デモンストレーション
検索条件
検索文字列:
こちらをクリックすると正規表現を使用して前後の文脈を指定できます。
検 索
キャンセル
ジャンル
(検索対象とするジャンルを選択できます。+をクリックすると細かく指定できます。)
白書
Yahoo!知恵袋
書籍


よく見るページ Firefox を使いこなそう 最新ニュース Gmail - 受信トレイ - jhl...
カードとログイン情報
国立国語研究所の言語コーパス整...
に限定してデータを公開します。モニター公開の目的は、実際にデータを使ってもらうことにより、コーパスの構築や活用に有益なフィードバックを得ること、及び、コーパスによる言語研究の普及を促すことの二つです。
モニター公開データの内容
モニター公開データは、DVDディスクで提供します。その中には以下の1～4のファイルが納められています。
1. サンプルの種類とファイル形式
書籍 約3,000万語(10,423サンプル)(プレーンテキスト/XMLファイル)
白書 約480万語(1,500サンプル)(プレーンテキスト/XMLファイル)
Yahoo!知恵袋 約520万語(45,725サンプル)(プレーンテキスト)
国会会議録 約490万語(159サンプル)(プレーンテキスト)
ファイルの形式は、プレーンテキスト(タグ無し)及びXMLファイル(タグ有り)です。書籍と白書は両方の形式のファイルがありますが、Yahoo!知恵袋と国会会議録は、プレーンテキストだけの提供です。また、書籍は、構築途中であるため、XMLファイルの数はプレーンテキストよりもやや少なくなっています。
2. 全文検索システム「ひまわり」
上記1のサンプルをまとめて検索することができます。検索結果には、検索対象語、前文脈、後文脈、書誌情報、著者情報等が表示されます。検索には正規表現が使えます。「ひまわり」の詳細については、以下のサイトをご参照ください。
「言語データベースとソフトウェア」
3. 各種情報
サンプルの属性として、以下の3種類の情報を収録しています。
書誌情報(サンプルを抽出した原本のタイトル、著者、出版年、ISBN、NDC、Cコードなど)
サンプル情報(サンプルID、原本におけるサンプルの抽出位置、著作権許諾状況等)
著者情報(著者ID、著者名、性別、生年代)
4. コアデータ

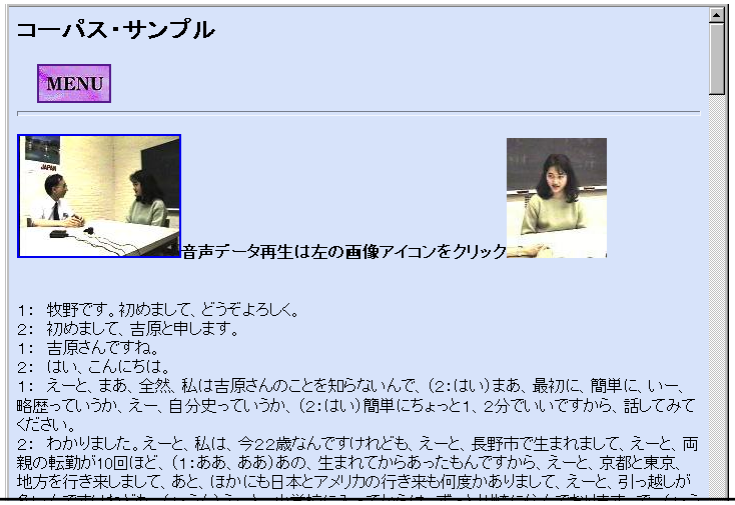
コーパスの紹介

2. 話しことば

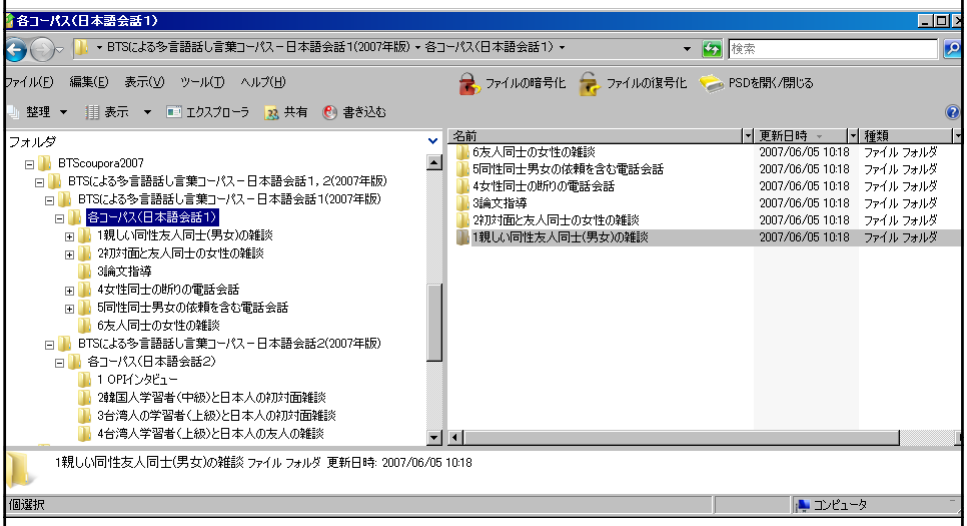
	日本語話し言葉コーパス	多言語話し言葉コーパス	日本語会話データベース
内容	国立国語研究所・情報通信研究機構・東京工業大学が共同開発した話し言葉データベース	日本語母語話者同士の会話と、日本語母語話者と日本語学習者の会話の文字化資料(CD-ROM)	日本語母国語話者と非母国語話者の発話パターンの比較分析と日本語教育向けの基礎資料となる言語データベース
入手	利用申込後、DVD18枚	利用申込書による申請。無料	ウェブからダウンロード
コーパス化	ツール付き	エクセル形式	*.raによる音声データとHTMLのテキストデータ

日本語会話データベース

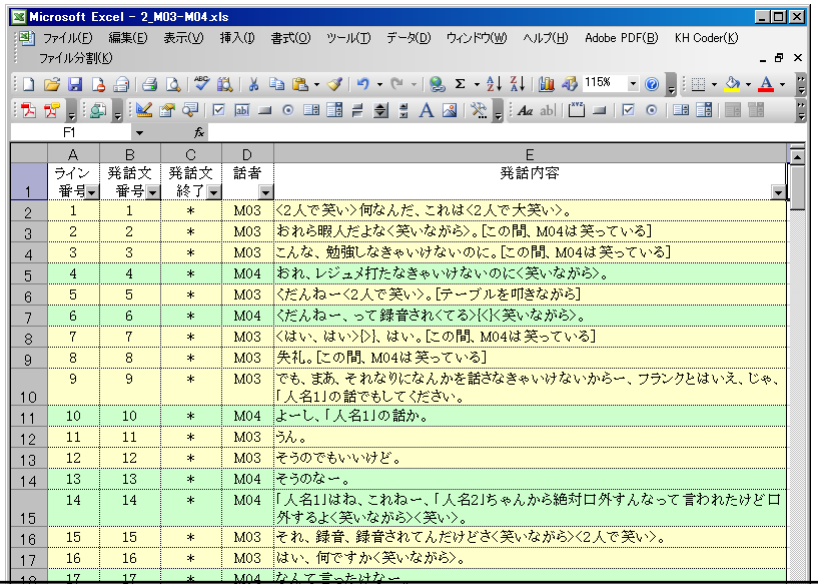
http://www.env.kitakyu-u.ac.jp/corpus/



多言語話し言葉コーパス



多言語話し言葉コーパス

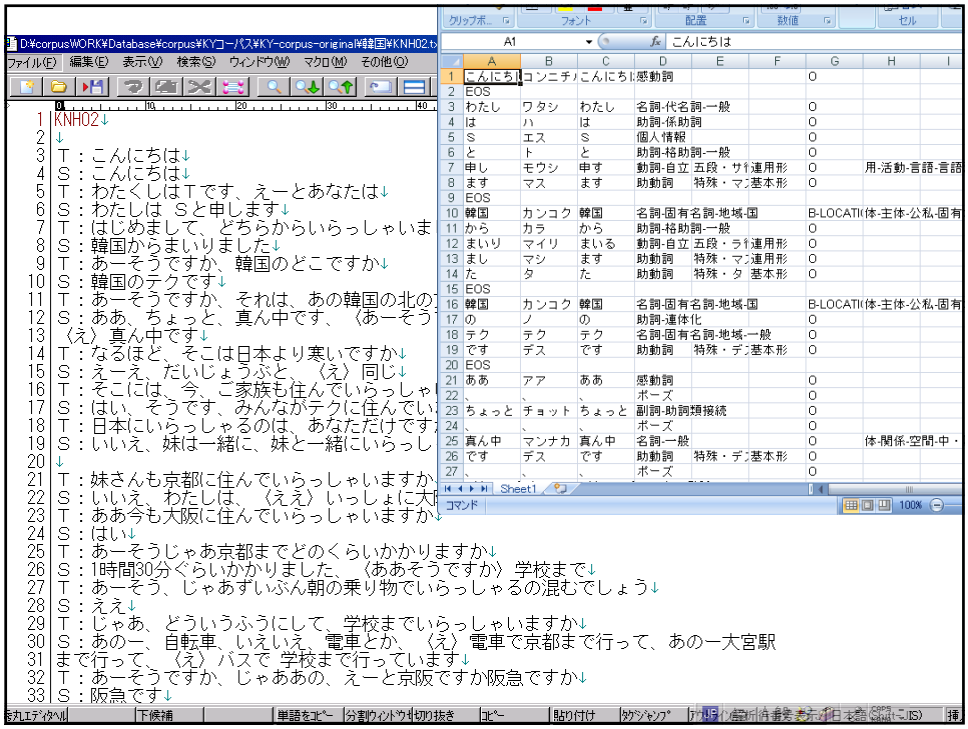


	A	B	C	D	E
1	ライン番号	発話文番号	発話文終了	話者	発話内容
2	1	1	*	M03	<2人で笑い> 何なんだ、これは<2人で大笑い>。
3	2	2	*	M03	おれら暇人だよな<笑いながら>。[この間、M04は笑っている]
4	3	3	*	M03	こんな、勉強しなきゃいけないのに。[この間、M04は笑っている]
5	4	4	*	M04	おれ、レジュメ打たなきゃいけないのに<笑いながら>。
6	5	5	*	M03	<だんねー<2人で笑い>。[テーブルを叩きながら]
7	6	6	*	M04	<だんねー、って録音されてる>[<]<笑いながら>。
8	7	7	*	M03	<はい、はい>[>]、はい。[この間、M04は笑っている]
9	8	8	*	M03	失礼。[この間、M04は笑っている]
10	9	9	*	M03	でも、まあ、それなりになんかを話さなきゃいけないからー、フランクとはいえ、じゃ、「人名1」の話でもしてください。
11	10	10	*	M04	よーし、「人名1」の話か。
12	11	11	*	M03	うん。
13	12	12	*	M03	そうのでもいいけど。
14	13	13	*	M04	そうのなー。
15	14	14	*	M04	「人名1」はね、これねー、「人名2」ちゃんから絶対口外すんなって言われたけど口外するよ<笑いながら><笑い>。
16	15	15	*	M03	それ、録音、録音されてんだけどさ<笑いながら><2人で笑い>。
17	16	16	*	M03	はい、何ですか<笑いながら>。
18	17	17	*	M04	なんて言ったけなー

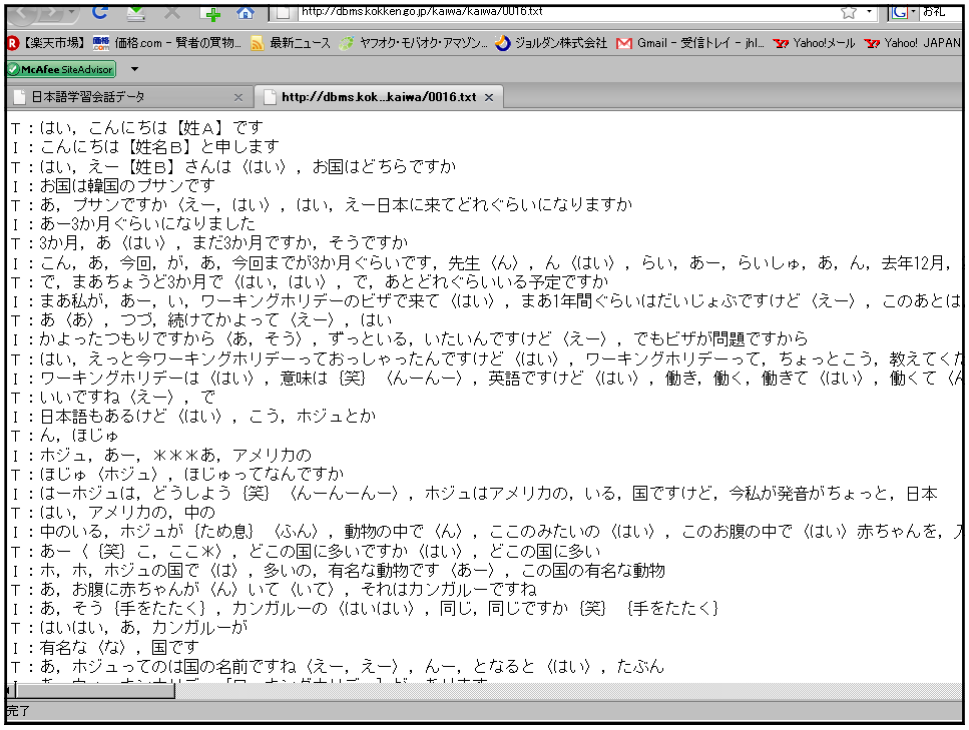
コーパスの紹介

3. 習得コーパス

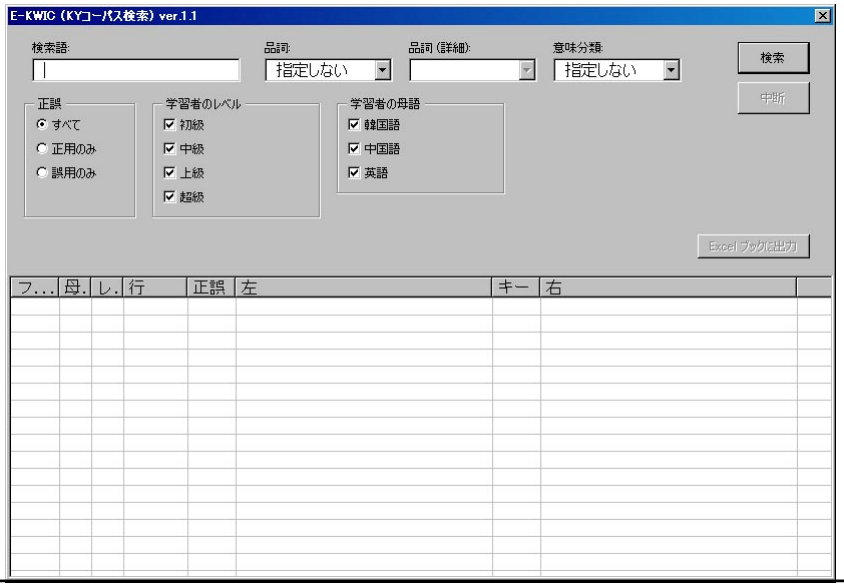
	KYコーパス (第二言語としての日本語学習者の話し言葉)	日本語学習会話データ
内容	90人分のOPIテープを文字化した言語資料である	390名分のOPIデータ
入手	利用申込後、FDが返送。無料	ウェブページからダウンロード
コーパス化	プレーンテキストデータ	プレーンテキストデータ
その他		一部音声データもあり.



http://dbms.kokkengo.jp/kaiwa/
日本語学習会話データ
閉じる
検索
「mp3（データ番号107，169，267）」は音声のサンプルデータです。
項目のうち，「日本語学習期間」と「日本語の能力試験」は参考データです。
「データ概要」を参照してください。
「CTRL」キーを押しながら項目をクリックすると，複数の項目での絞り込みができます。
OPIレベル
性別
年齢
出身国
母語
職業等
日本滞在期間
日本語学習期間
日本語能力試験
検索
検索条件クリア
一覧
データ番号
OPIレベル
年齢
性別
出身国
母語
職業等
日本滞在期間
日本語学習期間(参考)
日本語能力試験(参考)
TXT
表示
DL
完了


http://dbms.kokkengo.jp/kaiwa/kaiwa/0016.txt
日本語学習会話データ
http://dbms.kok...kaiwa/0016.txt
T：はい，こんにちは【姓A】です
I：こんにちは【姓名B】と申します
T：はい，えー【姓B】さんは〈はい〉，お国はどちらですか
I：お国は韓国のプサンです
T：あ，プサンですか〈えー，はい〉，はい，えー日本に来てどれぐらいになりますか
I：あー3か月ぐらいになりました
T：3か月，あ〈はい〉，まだ3か月ですか，そうですか
I：こん，あ，今回，が，あ，今回までが3か月ぐらいです，先生〈ん〉，ん〈はい〉，らい，あー，らいしゅ，あ，ん，去年12月，
T：で，まあちょうど3か月で〈はい，はい〉，で，あとどれぐらいいる予定ですか
I：まあ私が，あー，い，ワーキングホリデーのビザで来て〈はい〉，まあ1年間ぐらいはだいじょぶですけど〈えー〉，このあとは
T：あ〈あ〉，つづ，続けてかよって〈えー〉，はい
I：かよったつもりですから〈あ，そう〉，ずっといる，いたいんですけど〈えー〉，でもビザが問題ですから
T：はい，えっと今ワーキングホリデーっておっしゃったんですけど〈はい〉，ワーキングホリデーって，ちょっとこう，教えてくた
I：ワーキングホリデーは〈はい〉，意味は〔笑〕〈んーんー〉，英語ですけど〈はい〉，働き，働く，働きて〈はい〉，働くて〈
T：いいですね〈えー〉，で
I：日本語もあるけど〈はい〉，こう，ホジュとか
T：ん，ほじゅ
I：ホジュ，あー，＊＊＊あ，アメリカの
T：ほじゅ〈ホジュ〉，ほじゅってなんですか
I：はーホジュは，どうしよう〔笑〕〈んーんーんー〉，ホジュはアメリカの，いる，国ですけど，今私が発音がちょっと，日本
T：はい，アメリカの，中の
I：中のいる，ホジュが〔ため息〕〈ふん〉，動物の中で〈ん〉，ここのみたいの〈はい〉，このお腹の中で〈はい〉赤ちゃんを，
T：あー〈〔笑〕こ，ここ＊〉，どこの国に多いですか〈はい〉，どこの国に多い
I：ホ，ホ，ホジュの国で〈は〉，多いの，有名な動物です〈あー〉，この国の有名な動物
T：あ，お腹に赤ちゃんが〈ん〉いて〈いて〉，それはカンガルーですね
I：あ，そう〔手をたたく〕，カンガルーの〈はいはい〉，同じ，同じですか〔笑〕〔手をたたく〕
T：はいはい，あ，カンガルーが
I：有名な〈な〉，国です
T：あ，ホジュってのは国の名前ですね〈えー，えー〉，んー，となると〈はい〉，たぶん
完了

E-KWIC

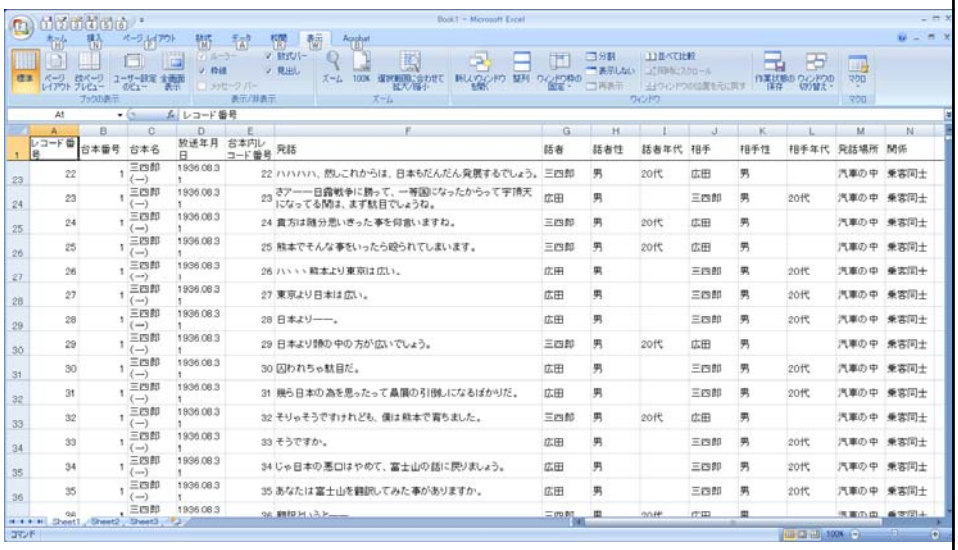


コーパスの紹介

4. 特殊なコーパス

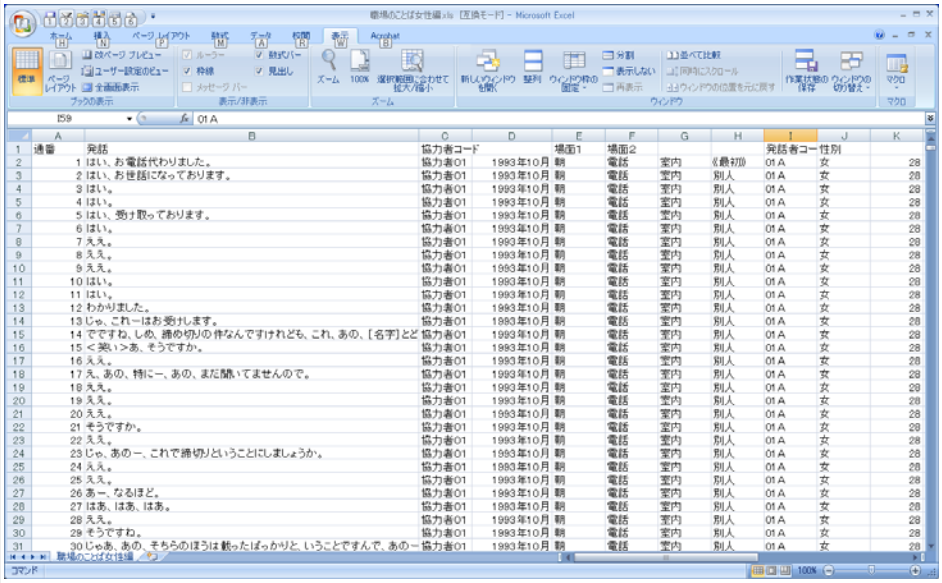
	戦時中のことば	女性のことば・職場編	男性のことば・職場編
内　容	戦時中の1936年から戦後の1955年の間に、日本放送協会からラジオで放送された78冊のラジオ台本を文字化し、Unicodeテキストファイル（タブ区切り）としたもの。	1993年9月～11月に首都圏で収録された音声資料を文字化したもの。対象は有識の20代から50代の女性。19名の協力者に職場でのインフォーマルな場面とフォーマルな場面での自然会話を録音している。	1999年10月～2000年12月に首都圏で収録された音声資料を文字化したもの。対象は有識の20代から50代の男性で、各世代5名で異なる職種・職場からなる19名の協力者に職場でのインフォーマルな場面とフォーマルな場面での自然会話を録音している。
入　手	書籍として購入(¥3,990)。テキストデータCD-ROMが付いてくる	書籍として購入(¥3,675)書籍にフロッピィディスク付。CVS（コンマ区切り）ファイルをLHAで、圧縮した資料を付属する。	書籍として購入(¥2,940)書籍にCD-ROM付。CVS（コンマ区切り） ファイルをLHAで、圧縮した資料を付属する。
規　模	約4万行の発話	約1万行の発話	約1万行の発話
コーパス化	整形化の必要あり	整形化の必要あり	整形化の必要あり

「戦時中のことば」のデータサンプル



レコード番号	台本番号	台本名	放送年月日	台本内レコード番号	発話	話者	話者性	話者年代	相手	相手性	相手年代	発話場所	関係
22	1	三四郎(一)	1936.08.31	22	ハハハハ、然しこれからは、日本もだんだん発展するでしょう。	三四郎	男	20代	広田	男		汽車の中	乗客同士
23	1	三四郎(一)	1936.08.31	23	さアー日露戦争に勝って、一等国になったからって宇宙天になってる間は、まず駄目でしょうね。	広田	男		三四郎	男	20代	汽車の中	乗客同士
24	1	三四郎(一)	1936.08.31	24	貴方は随分思いきった事を何言いますね。	三四郎	男	20代	広田	男		汽車の中	乗客同士
25	1	三四郎(一)	1936.08.31	25	熊本でそんな事をいったら殺られてしまいます。	三四郎	男	20代	広田	男		汽車の中	乗客同士
26	1	三四郎(一)	1936.08.31	26	ハハハ熊本より東京は広い。	広田	男		三四郎	男	20代	汽車の中	乗客同士
27	1	三四郎(一)	1936.08.31	27	東京より日本は広い。	広田	男		三四郎	男	20代	汽車の中	乗客同士
28	1	三四郎(一)	1936.08.31	28	日本より一一。	広田	男		三四郎	男	20代	汽車の中	乗客同士
29	1	三四郎(一)	1936.08.31	29	日本より豚の中の方が広いでしょう。	三四郎	男	20代	広田	男		汽車の中	乗客同士
30	1	三四郎(一)	1936.08.31	30	図われちゃ駄目だ。	広田	男		三四郎	男	20代	汽車の中	乗客同士
31	1	三四郎(一)	1936.08.31	31	我ら日本の為を思ったって亀頭の引倒しになるばかりだ。	広田	男		三四郎	男	20代	汽車の中	乗客同士
32	1	三四郎(一)	1936.08.31	32	そりゃそうですけれども、僕は熊本で育ちました。	三四郎	男	20代	広田	男		汽車の中	乗客同士
33	1	三四郎(一)	1936.08.31	33	そうですか。	広田	男		三四郎	男	20代	汽車の中	乗客同士
34	1	三四郎(一)	1936.08.31	34	じゃ日本の悪口はやめて、富士山の話に戻りましょう。	広田	男		三四郎	男	20代	汽車の中	乗客同士
35	1	三四郎(一)	1936.08.31	35	あなたは富士山を翻訳してみた事がありますか。	広田	男		三四郎	男	20代	汽車の中	乗客同士
36	1	三四郎(一)	1936.08.31	36	難船というと一一	三四郎	男	20代	広田	男		汽車の中	乗客同士

「職場のことば」のデータサンプル



通番	発話	協力者コード		場面1	場面2			発話者コー	性別	
1	はい、お電話代わりました。	協力者01	1993年10月	朝	電話	室内	《最初》	01A	女	28
2	はい、お世話になっております。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
3	はい。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
4	はい。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
5	はい、受け取っております。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
6	はい。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
7	ええ。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
8	ええ。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
9	ええ。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
10	はい。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
11	はい。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
12	わかりました。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
13	じゃ、これーはお受けします。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
14	でですね、しめ、締め切りの件なんですけれども、これ、あの、[名字]と	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
15	<笑い>あ、そうですか。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
16	ええ。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
17	え、あの、特にー、あの、まだ聞いてませんので。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
18	ええ。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
19	ええ。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
20	ええ。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
21	そうですか。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
22	ええ。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
23	じゃ、あのー、これで締切りということにしましょうか。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
24	ええ。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
25	ええ。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
26	あー、なるほど。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
27	はあ、はあ、はあ。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
28	ええ。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
29	そうですね。	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28
30	じゃあ、あの、そちらのほうは載ったばっかりと、ということですんで、あのー	協力者01	1993年10月	朝	電話	室内	別人	01A	女	28

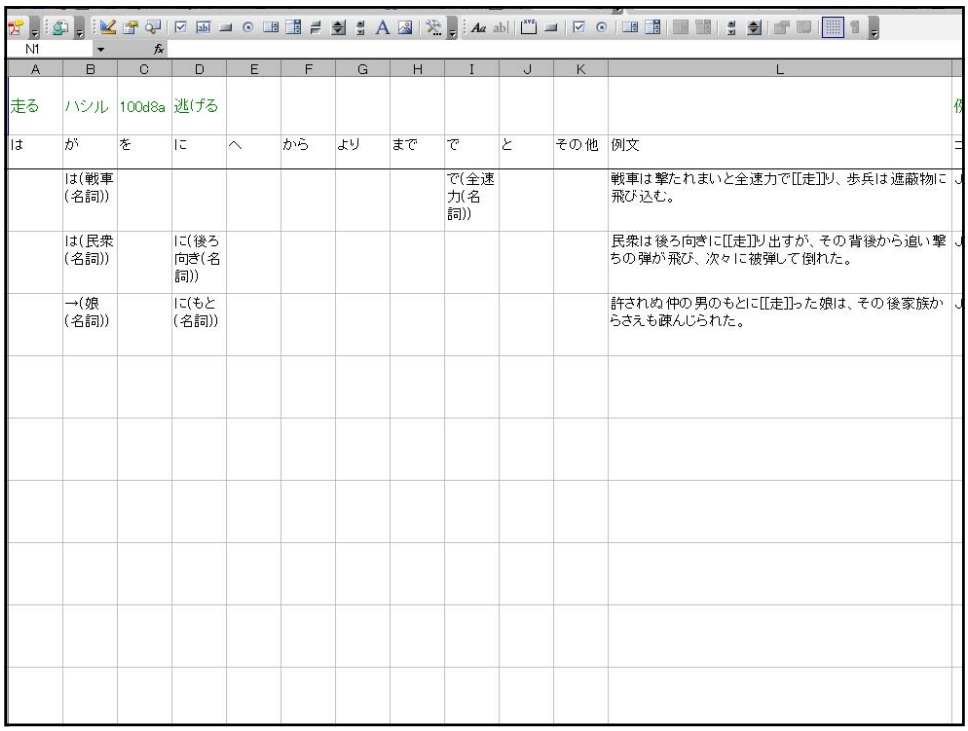
コーパスの紹介

5.　辞書・シソーラス

区分	日本語語彙大系	日本語動詞の結合価	分類語彙表
内容と規模	30万語の収録語は3000種の意味分類を用いて定義されている、大規模の日本語シソーラス。3000種の意味分類を用いて日本語の文型を定義している。	15万文規模の日本語文集合をもとに、「動詞がとりうる格と、その格を担う体言」について整備した、大規模な言語データ集。	延べ約９万６千語の日本語の単語を意味的に分類したもの
入手	有料（無料化予定あり）	有料	有料
検索ソフト	パッケージとして同梱	パッケージとして同梱	ない(PDFの索引のみ)
元の電子データを直接閲覧	基本的には不可能	基本的には不可能	可能(PDFファイル)
製作元	日本電信電話株式会社NTTコミュニケーション科学基礎研究所	日本システムアプリケーション荻野孝野・小林正博・井佐原均 編著	国立国語研究所

日本語動詞の結合価

- **GUI**



走る	ハシル	100d8a	逃げる								
は	が	を	に	へ	から	より	まで	で	と	その他	例文
	は(戦車(名詞))							で(全速力(名詞))			戦車は撃たれまいと全速力で[[走]]り、歩兵は遮蔽物に飛び込む。
	は(民衆(名詞))		に(後ろ向き(名詞))								民衆は後ろ向きに[[走]]り出すが、その背後から追い撃ちの弾が飛び、次々に被弾して倒れた。
	→(娘(名詞))		に(もと(名詞))								許されぬ仲の男のもとに[[走]]った娘は、その後家族からさえも疎んじられた。

日本語語彙大系

- **GUI**

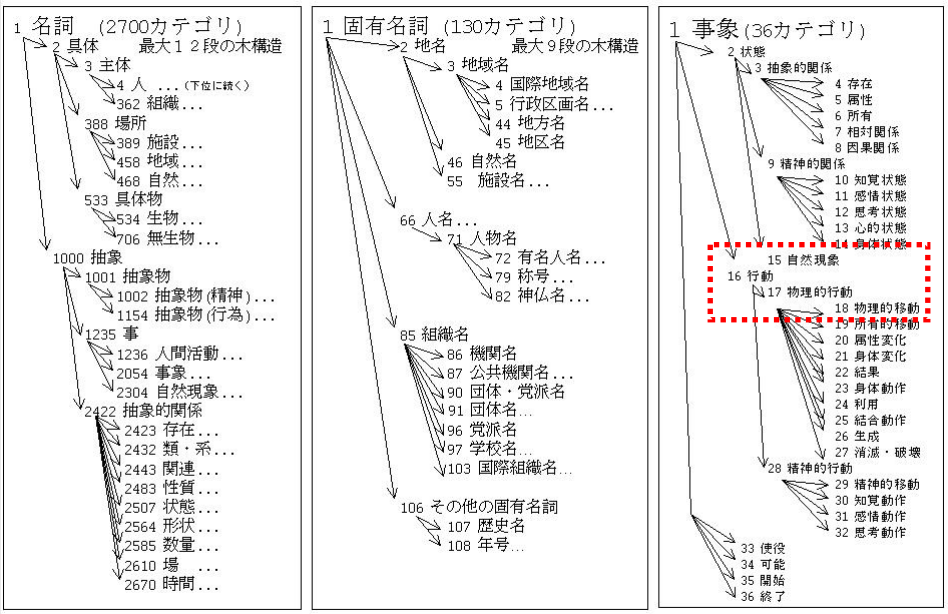


図1　意味の体系

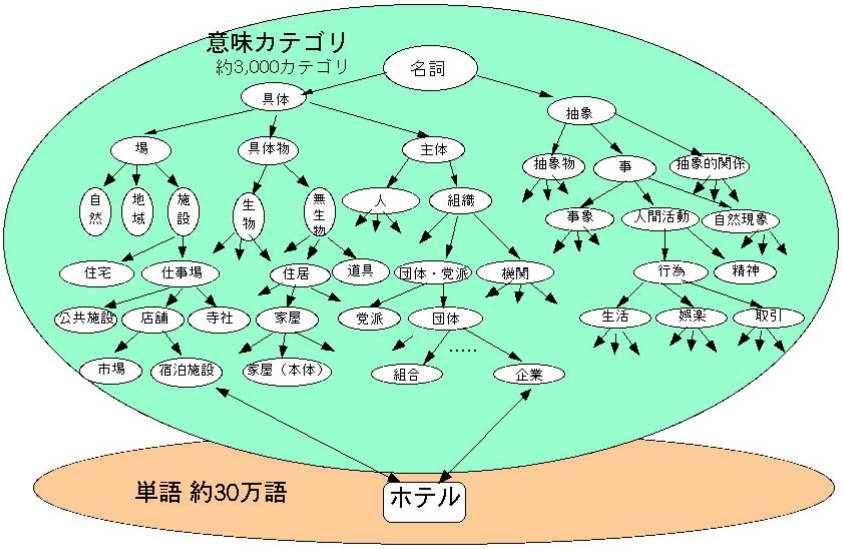


図2　名詞の意味体系と単語(ホテル)の対応関係例

研究紹介

1. 李 在鎬・井佐原 均（2006）「第二言語獲得における助詞「に」の習得過程の定量的分析」、『計量国語学』第二十五巻四号、pp.163-180.
2. 李 在鎬・鈴木 幸平・永田 由香・黒田 航・井佐原 均(2007)「動詞「流れる」の語形と意味の問題をめぐって」『計量国語学』第26巻2号, pp.64-74
3. 李 在鎬(2008)「移動動詞に対する実験的分析」, 『言葉と認知のメカニズム』pp.87-101, ひつじ書房.
4. 黒田 航・李 在鎬（2008）「WWWを丸ごとデータにすると何がわかるか」、『日本語学(2008年2月号)』27-2:pp. 22-37, 明治書院
5. 李 在鎬(2009)「助数詞と多変量解析」, 『日本語学(2009年6月号)』28-7:pp.44-57, 明治書院.
6. 李 在鎬(2009)「タグ付き日本語学習者コーパスの開発」『計量国語学』27-2, pp.60-72.

http://www30.atwiki.jp/corpus-ling/pages/52.html