

3 言語データの整理 <Word>

最近の言語研究のひとつの流れとして、電子化された膨大な言語資料に既存の分析用プログラムを適用する方法があります。資料が大きくなればなるほど一般的な解釈が可能になり、また、データがカバーする範囲が広がるため、このような大規模な研究の成果は看過できません。これは「エクステンシブな研究」と呼ぶことができます。

一方、従来の文献学的研究では特定の文献資料を精読することから始めます。そこでは資料の一字一句の解釈が必要です。従来このタイプの研究ではノートやカードの作業を通して言語の記述と説明がなされてきましたが（「インテンシブな研究」と呼ぶことができるでしょう）、最近ではここでもコンピュータが利用されることが多くなっています。コンピュータを用いて資料を電子化することで検索性や保存性を高めることができます¹。このように文学、文献学、言語学など文系の分野でもコンピュータが有用なツールとなっています。スペインの国立図書館では、利用者が資料を汚す可能性があるため、ボールペンやインクの使用が禁じられ、鉛筆とコンピュータの持ち込みだけが許されています

言語データをコンピュータで分析する場合、まずは電子化された既存の資料を整理する必要があります。分析の対象となる資料は、目的に応じてさまざまに加工されます。たとえば文を単位にしてテキストを分析するとき、すべてのピリオドの後で改行する、という操作が必要になります。この作業をピリオドの後で一つ一つ改行コードを入れる（[Enter]キーを打つ）ということをしていると、大変な手間がかかります。しかし、Wordの置換機能を利用して、ピリオドを「ピリオド+改行コード」に置き換えれば、瞬時に作業が完了します。

また、分析の対象となる文字を色付けすると、見落としを避けることができます。たとえば定冠詞の the を黄色に、不定冠詞の a/an を緑にマークするといったことも、Wordの機能で簡単に実現できます。ワイルドカードを使えば、より複雑な処理が可能です。たとえば、「4

¹ テキストの分析のためには、その転写・文字化の規則を一定にしなければなりません。音声表記、文字表記、句読点、改行、段落の形式、ページの形式など、さまざまな規則があります。あらかじめ規則を定めておかないと、後で統一するのが困難になります。

桁の数字」や「b で始まり t で終わる単語」を検索したり、マークしたり、置換えたりすることができます。

以下では、分析対象の資料を、Word を使って効率的に整理していく方法を見ていきます。Word を使いこなし、味方につければ、分析の準備時間を短縮し、より正確で効果的な分析が可能になります。

◇22 データをマークする

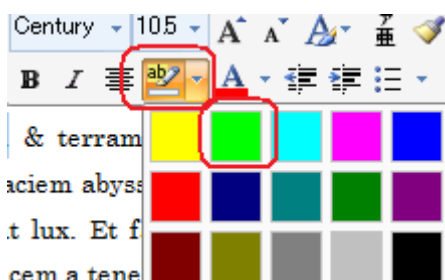
まず、電子データに対して分析の対象となる単語や表現をマークアップする方法を見ていきます。ここでは Word の蛍光ペンを利用して色をつけていく方法を紹介します。

[1] 目視によるマーク

次のテキスト（ラテン語『旧約聖書』「創世記」の冒頭）の中で a という文字を探してみましよう。

(1:0) Cap. 1 (1:1) In principio creavit Deus caelum & terram. (1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas. (1:3) Dixitque Deus. Fiat lux. Et facta est lux.

手作業で（目視で）マークするときは該当部分を選択し、右クリックして、蛍光ペンの色を選びます。ここでは緑色にします。



文字の選択（ドラッグ）のショートカットは[Shift]+[→]です。単語の末尾までの選択のショートカットは[Ctrl]+[Shift]+[→]です。蛍光ペンのショートカットは[Ctrl]+[Alt]+H です。

(1:0) Cap. 1 (1:1) In principio creavit Deus caelum & terram. (1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas. (1:3) Dixitque Deus. Fiat lux. Et facta est lux.

最初に（選択しないで）「蛍光ペンの色」を指定すれば、カーソルが次のように蛍光ペンの形になり、直接マークしていくことができます。



カーソルの形を元に戻すには[Esc]キーを押してください。

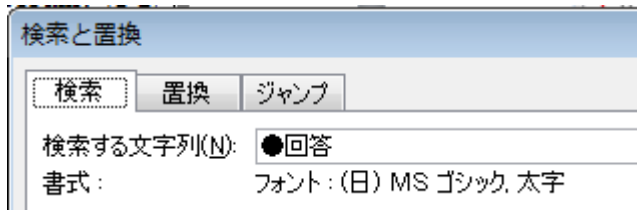
[2] 検索によるマーク

手作業でマークすると時間もかかりますし、見落としが出てしまうかもしれませんが、Word の検索を使えば、効率的にマークすることができます。ここでは先ほどと同様に全体のテキストで a のすべてを検索しましょう。テキストを選択し、「ホーム(H)」→「検索(FD)」→「高度な検索(A)」の「検索する文字列(N)」を a とし、「検索された項目の強調表示(R)」ボタンを押し、「すべて強調表示(H)」を選択します。検索の対象となる範囲を選択しておくこと、選択された部分だけが検索されます。

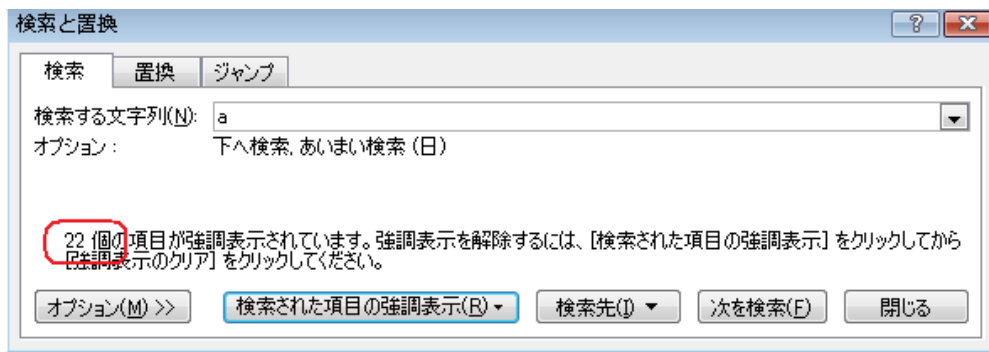


「検索する文字列」にフォントが指定されていると、そのフォントを使った検索文字列だけが検索されます。フォントの指定が必要でないときは、◆「オプション(M)」→「書式の削除(T)」²

² T のキーは検索と置き換えのオプション内の「接尾辞に一致する(T)」にも使われているので、注意してください。「書式の削除(T)」が有効なのは、書式が設定されているときだけです。それ以外のときは、「書式の削除(T)」の表示が薄くなり、無効になります。



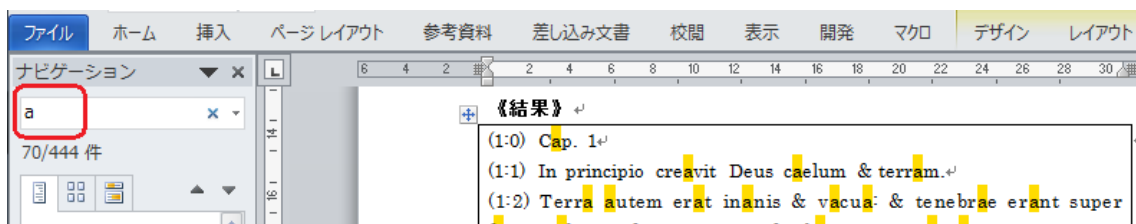
《結果》



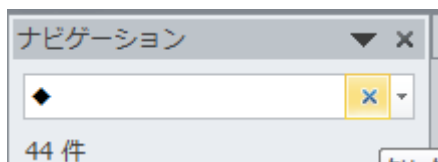
(1:0) Cap. 1 (1:1) In principio creavit Deus caelum & terram. (1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas. (1:3) Dixitque Deus. Fiat lux. Et facta est lux.

上の図で囲んだように、強調表示されている項目の個数が計算されています。頻度を調べるときはこれをメモしておきましょう。

Word 2010 では「ナビゲーションウィンドウ」が使われます。



検索のボックスの「×」を押すと、マークが消えます。



* Word 2007 以前の検索方法を使うには、◆「ホーム(H)」→「検索(FD)」→「高度な検索(A)」

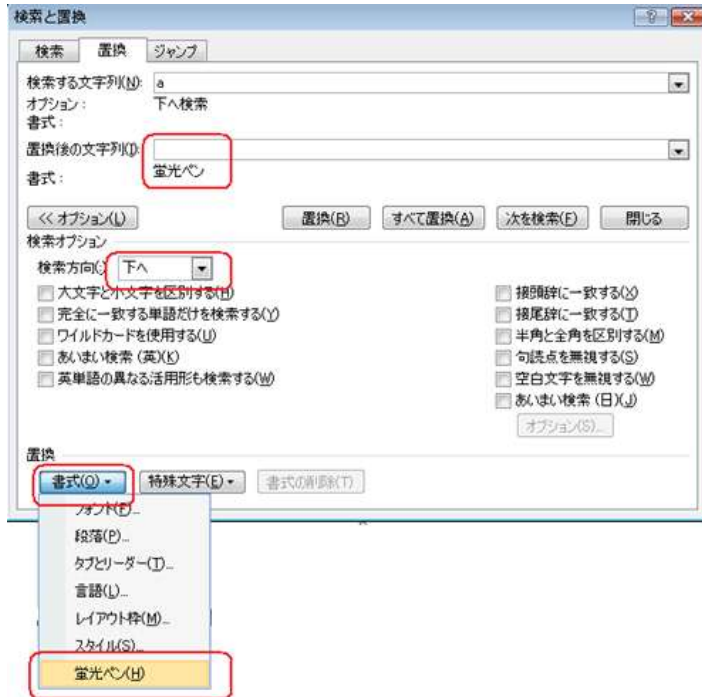
<TIPS> Word 2010 で「ナビゲーションウィンドウ」を使わずに、Word2007以前の検索をショートカットキー[Ctrl]+Fで常時利用する場合には以下のように設定します。◆「ファイル(F)」→「オプション(T)」→「リボンのユーザー設定」→「ショートカットキー：ユーザー設定(T)」ボタン→(キーボードのユーザー設定)の「分類(C)」で[ホーム]タブを選択し、「コマンド(O)」:EditFind→「割り当てるキーを押してください(N)」:[Ctrl]+F(キーイン)→「割り当て」ボタン

[3] 置換によるマーク

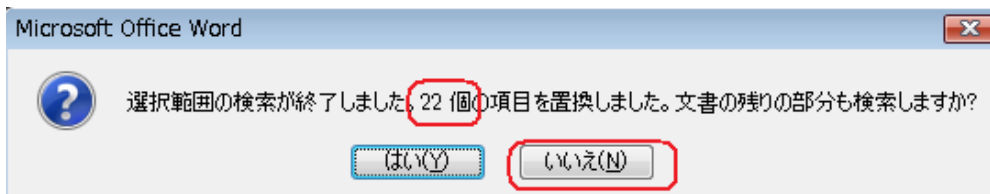
「検索」による強調表示の方法では一時的にハイライトされるだけで、テキストに文字を書き込むなどの操作をするとハイライトは消えてしまいます。一方、「置換」の機能(→参照)を使って蛍光ペンの色をつければ、ハイライトをテキスト中に残しておくことができます。aの文字をマークする操作を見てみましょう。

◆対象テキストを選択します。「ホーム(H)」→「編集」グループ→「置換(R)」→「置換後の文字列(I)」にカーソルを置き、「オプション(M)」ボタンを押し、「検索方向」で「下へ」を選択し、「書式(O)」で「蛍光ペン(H)」を選択します。「置換後の文字列」の入力欄は空欄のまま構いません。「すべて置換(A)」ボタンを押すと、蛍光ペンの色は直前に使った色になります³。

³ 置換のダイアログでは蛍光ペンの色を指定することはできません。なお、蛍光ペンで「色なし」を選んだ後にこの操作を行ってもマークされませんから注意してください。もし、「色なし」になっている場合には、適当な文字を入力し、一旦何らかの色を使って削除するなどの操作をすればよいでしょう。



◆ 次のメッセージが現れたら個数を確認して「いいえ」と答えてください。「はい」を選択した場合は、選択範囲だけではなく文章全体で置換を実行します。



[4] あいまい検索によるマーク

Word には「あいまい検索」という機能が用意されています⁴。この機能を有効に使うと、目的の文字列に変異形（「ウィンドウ」と「ウインドウ」など）がある場合、見落としを減らすことができます。

英語のあいまい検索

英語のあいまい検索では、ミススペルと思われるようなケースも一緒に検索できます。たとえば、apple を「検索する文字列」にしてあいまい検索を有効にすると、次の結果になります。

⁴ 「あいまい検索（英）（K）」 「あいまい検索（日）（J）」では「特殊文字(E)」は使えません。

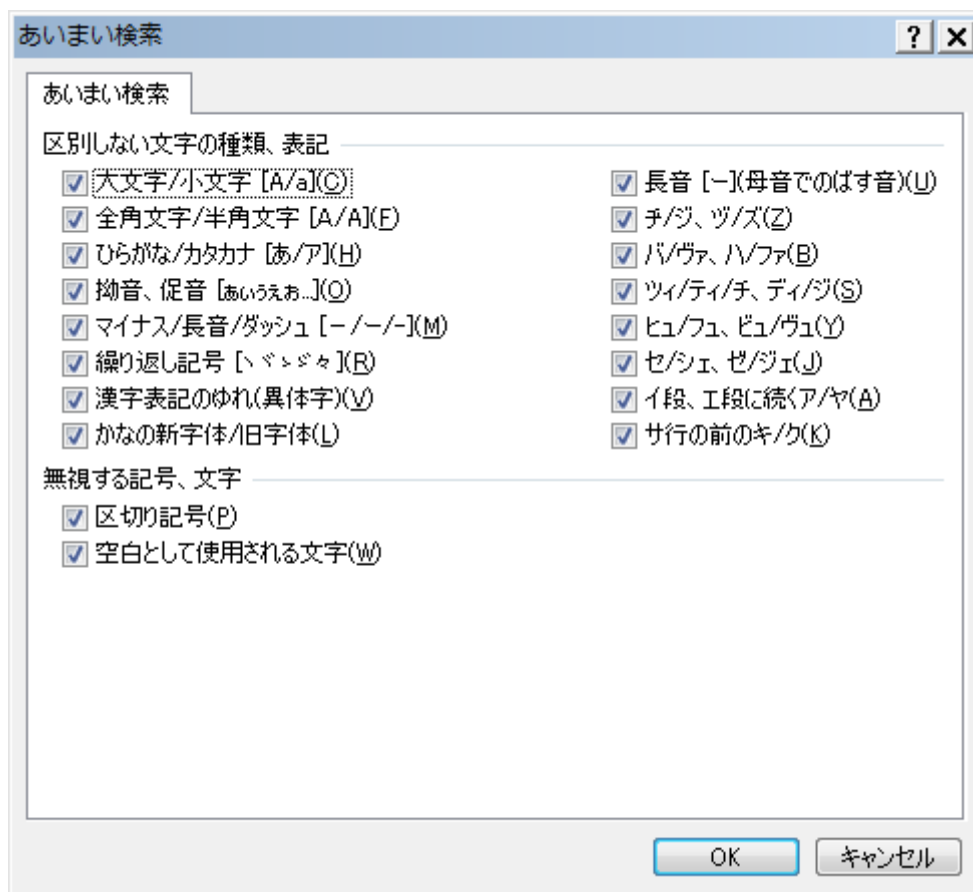
- ◆ 「ホーム(H)」 → 「検索(FD)」 → 「高度な検索(A)」 → 「あいまい検索(英)(K)」をチェック

apple, aple, upple, uple, upl, ap, up, apple を⁵

日本語のあいまい検索

日本語のあいまい検索では、ひらがなとカタカナを区別せずに「テレビ」と「てれび」を同時に検索したり、「ジ」と「ヂ」などの区別をせずに検索することができます。その他の項目に関しては以下の操作で確認できます。

- ◆ 「ホーム(H)」 → 「検索(FD)」 → 「高度な検索(A)」 → 「あいまい検索(日)(J)」をチェック → 「オプション(S)」をクリック



このオプションを調整することで思い通りの検索が可能になります。たとえば、「拗音、促音(Q)」のチェックを外すと、「ウィンドウ」と「ウインドウ」は区別される

⁵ 「apple を」のように、英字と日本語がつながっていると正しく検索されません。

ようになります。

<Tips> 「英単語の異なる活用形も検索する(W)」にチェックを入れると、動詞、形容詞、名詞の変化形も同時に検索できるようになります。たとえば、go で検索すると、go・went・gone、small で検索すると、small・smaller・smallest、window で検索すると、window・windows がヒットします。

◇23 高度な検索と置換

言語データを整理するにあたって、文単位にデータを区切ろうと思うと、ピリオドや読点を「改行」に置き換える必要があります。改行は普通の文字列とは異なるため、特殊な記号を使って表します。また、-tion で終わる単語や un-で始まる単語といった高度な検索が必要となることがあります。Word のワイルドカードの機能を使えば、このような検索もできます。

[5] 特殊文字による検索と置換

改行やタブなどの編集記号は「特殊文字」で表します。よく用いられる特殊文字には以下のようなものがあります。

特殊文字名	記号	特殊文字名	記号
段落記号	^p	三点リーダー	^j
タブ文字	^t	文末脚注記号	^e
コメント記号	^a	フィールド	^d
任意の一文字	^?	脚注記号	^f
任意の数字	^#	グラフィックス	^g
任意の英字	^\$	セクション区切り	^b
段区切り	^n	全角または半角の空白	^w
省略記号	^i	クリップボード	^c

また、「検索と置換」のオプションを開き、「特殊文字(E)」から一覧を見ることができます⁶。

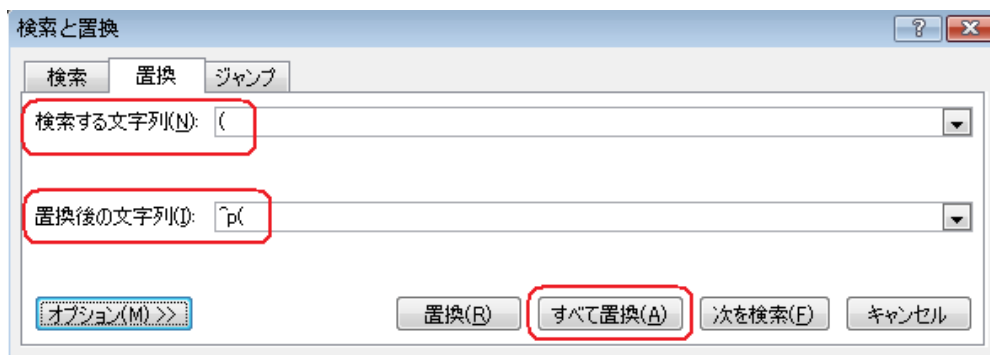
【例】：以下のテキストを「(」の前で改行してみましよう。

⁶ 「あいまい検索(英)(K)」と「あいまい検索(日)(J)」、「英単語の異なる活用形も検索する(W)」、「ワイルドカードを使用する(U)」のチェックが外れていることを確認して下さい。

(1:1) In principio creavit Deus caelum & terram.(1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas.(1:3) Dixitque Deus. Fiat lux. Et facta est lux.

◆対象を選択し、「ホーム(H)」→「置換」の「検索する文字列」を「(」、置換後の文字列に「^p(」とします⁷。

「^p」は改行コードを示す特殊文字です。「^」は「キャレット」と呼びます。



「すべて置換」のボタンを押すと、次のようになります。以後このような置換の手続きを、置換：「(」→「^p(」と表記します。

《結果》

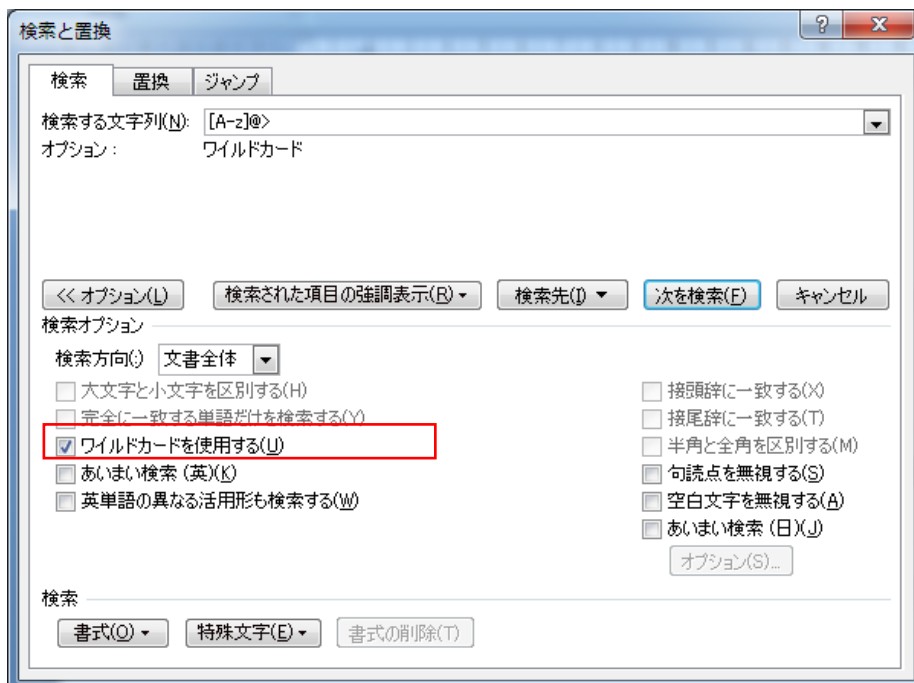
(1:1) In principio creavit Deus caelum & terram.
(1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas.
(1:3) Dixitque Deus. Fiat lux. Et facta est lux.

[6] ワイルドカードによる検索と置換

「ワイルドカード」とは、任意の文字列を指定したり、文字の繰り返しなどを指定することができる特殊な記号です。これを使うことによって具体的な個別の文字（列）を検索するばかりでなく、検索を一般化し、様々な形の文字列を一度に検索ことができます。たとえば、a で

⁷ 蛍光ペンやイタリックなどの書式を削除しておきましょう。「検索と置換」の画面で最下段に「書式の削除(T)」をクリックします。また、IMEの全角・半角の区別に注意しましょう。ここでは半角文字を使います。

終わる単語をすべてマークする、sc ではじまる単語をすべてマークする、漢字をすべてマークするといった操作が可能になります。ワイルドカードは、検索のオプションで「ワイルドカードを使用する(U)」をチェックすると使用可能になります。



ワイルドカードの規則

ワイルドカードの規則は次のとおりです。

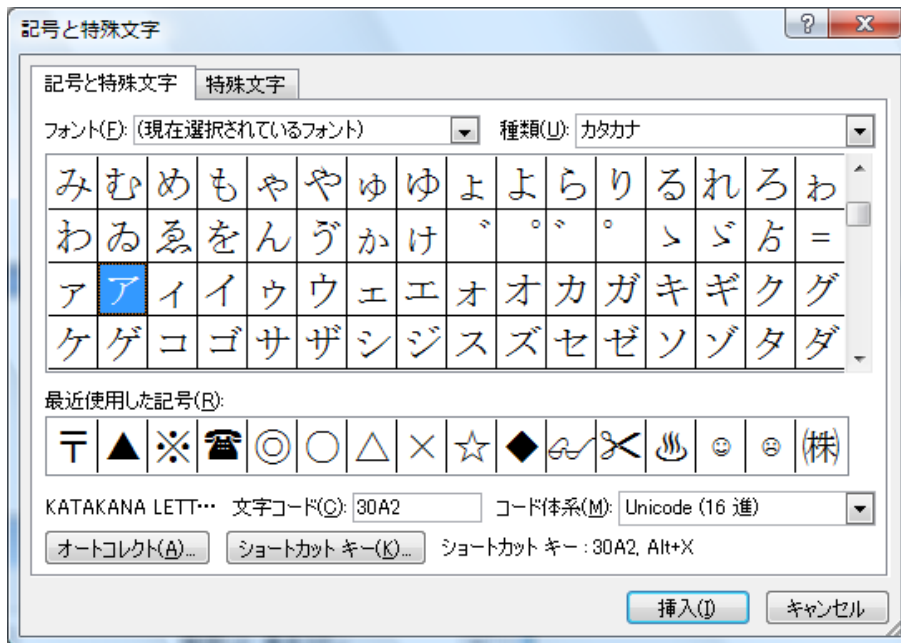
?	任意の 1 文字を検索します。例：「p?t」で pat, pet, pit, pot, put などを検索します。
*	任意の文字列を検索します。例：「b*d」で, bd, bad, bed, beed, bread などを検索します。任意の文字列は空白も含まれます。
<	単語の先頭に一致します。例：「<in」で inactive, interesting, inside の in などを検索します。begin の in は検索しません。
>	単語の末尾に一致します。例：「ed>」で、called, surprised, visited などの ed を検索します ⁸ 。edit の ed は検索しません。

⁸ 日本語などの全角文字では単語の先頭と末尾を一致させることができません。

[...]	指定した文字のいずれかに一致します。例：「<b[ai]t>」で bat と bit を検索します。but は検索しません。
[-]	指定した範囲内の任意の 1 文字に一致します。文字の範囲は、ユニコードの順番に従います。例：「<[n-t]ight>」で night, right, sight, tight などを検索します。might は検索しません。
[!...]	[...]内の文字や「.-」で示す文字間の範囲に含まれる文字を除く任意の 1 文字に一致します。例「t[!a-m]ck」で tock, tuck を検索します。tack, tick は検索しません。
{n}	直前の文字または式を n 個に一致します。n は 1 以上の数字です。例：「Ah{2}」で Ahh を検索します。A, Ah は検索しません。
{n,}	直前の文字または式を n 個以上に一致します。n は 1 以上の数字です。例：「fe{1,}d」で fed や feed などを検索します。fd は検索しません。
{n,m}	直前の文字または式を n ~ m 個に一致します。n, m は 1 以上の数字です。例：「Ah{1,3}」で Ah, Ahh, Ahhh を検索します。Oh, Oh, Ohh は検索しません。
@	直前の文字または式 1 個以上に一致します。「fe@d」で fed, feed を検索します。

ワイルドカードを使用する際の注意点

- [*-*]の範囲は昇順で指定してください。たとえば、×[t-n]は検索しません。範囲は文字コードの順番に従います。日本語で検索する場合も同様です。たとえば、カタカナを検索するときは[ア-ン]を使います。文字コードを知るには、「挿入(N)」→「記号と特殊文字(U)」→「その他の記号(M)」。



- 英字以外の ß や ñ などは、[A-zßñ]のように[...]の中に書き加えることができます。ただし、ハイフンの前後には書かないで下さい。
- ワイルドカードは大小文字を区別します。たとえば、a と A をどちらも検索するときは[Aa]と指定します。the, The, THE など全部検索するときは、[Th][Hh][Ee]と指定しなければなりません。
- ワイルドカードで「改行」を検索するときは特殊文字の改段落のコード^pではなく、^13を使います。

【例】ここでは例として「aで終わる単語」を、ワイルドカードを使ってマークしてみましょう。

◆まず、「ワイルドカードを使用する(U)」にチェックが入っているか確認します。次に、任意のアルファベットを指定するために[A-z]を指定します⁹。これがaの前で1つ以上連続していればよいので、[A-z]@となります。しかし、このままではeratなどaの後ろにアルファベットが続く場合も拾ってしまいます。そこで、>を使って単語区切りであることを明示します。

「検索と置換」の画面で、検索する文字列に「[A-z]@a>」を書き込み、(1)「オプション」を開き、(2)「ワイルドカードを使用する」にチ

⁹ [A-z]は正確には[A-Za-z]としなければなりません。ユニコードの順番で、Zとaの間に[,], ^, _などが含まれるためです。

エックを入れ、(3)「置換後の文字列」で「書式」から蛍光ペンを選択します。そして、(4)「すべて置換」のボタンを押します。

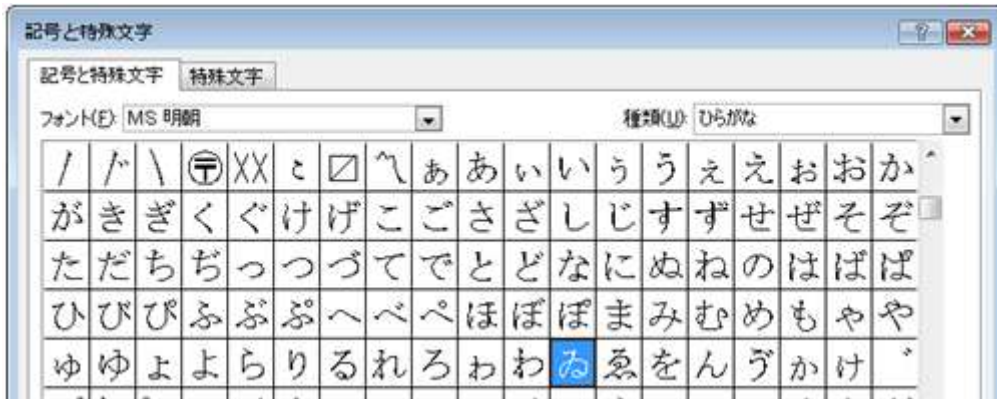


《結果》

(1:1) In principio creavit Deus caelum & terram.
(1:2) **Terra** autem erat inanis & **vacua**: & tenebrae erant super faciem
abyssi: & spiritus Dei ferebatur super aquas.
(1:3) Dixitque Deus. Fiat lux. Et **facta** est lux.

<TIPS> ¥, (,), {, }, ?, *, [,], !など、ワイルドカードとして使われている文字を検索するときは、「¥」をつけて「エスケープ」します。たとえば「?」を検索するときは検索する文字列に「¥?」を記入すれば「?」にヒットするようになります。

<TIPS> 旧字体の「ゐ」や「ゑ」の位置を「記号と特殊文字」で確認してみましょう。



上の図を見ると、「ゐ」と「ゑ」は「わ」と「を」の間にありますから、[わ-を]で検索できます。「い」と「え」は「あ」行だけで「や」行にはありませんから、「やいゆえよ」は[やいゆえよ]、または[いえや-よ]で検索します。

繰り返し一致

繰り返しの{1,}と@の機能は類似していますが、次の違いがあります。{1,}は、先行するパターンがマッチする範囲で一番長い部分で一致し（最長一致）、@は、全体のパターンがマッチする一番短い部分で一致します（最短一致）。どちらも左の位置から検索を始めます。まず、「aで始まる単語」を検索してみましょう。

検索(1): <a[A-z]{1,}

(1:1) In principio creavit Deus caelum & terram.
 (1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas.

「語頭にある a に続く英文字 1 個以上」というパターンがマッチする一番長い位置まで検索をします。

検索(2): <a[A-z]@

(1:1) In principio creavit Deus caelum & terram.
 (1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas.

一方、「@」を使って検索すると autem で <a が a に一致し、[A-z]@ が u に一致して、ここで検索をやめてしまいます。したがって、「aで始まる単語」を検索するには{m,n}の繰り返しを用いる必要があります。

次に、「mで終わる単語」を検索してみましょう。

検索(3): [A-z]{1,}m>

(1:1) In principio creavit Deus caelum & terram.
(1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem
abyssi: & spiritus Dei ferebatur super aquas.

最初に[A-z]{1,}で caelum まで最長で一致してしまい、その後に m を検索するので見つかりません。

検索(4): [A-z]@m>

(1:1) In principio creavit Deus caelum & terram.
(1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem
abyssi: & spiritus Dei ferebatur super aquas.

一番左の位置から語末の m までパターンが最短で一致します。よって、「mで終わる単語」を検索するには、「@」を使った繰り返しが適切です。

【例】任意の文書で「母音+m」で終わる単語を検索する方法を考えてみましょう。まず、「母音+m」の部分は[aeiou]m で表します。語末は「>」を用いましょう。母音の前にはアルファベットがありますので、[A-z]とします。また、「mで終わる」単語なので繰り返し記号として「@」を使います。

◆「検索する文字列(N)」 : [A-z]@[aeiou]m>を入力します。

《結果》

1	0	Cap. 1
1	1	In principio creavit Deus caelum & terram.
1	2	Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas.

TIPS 「X 回以下の繰り返し」を検索するには、{m,n}を使用して表現します。このとき、m の部分の数字は省略できません。たとえば、「edで終わる5文字以

下の単語を検索するとき、「<[a-z]{,3}ed」とはできず、「<[a-z]{1,3}ed」とします¹⁰。

<TIPS> 1010, 101010 などの 10 の繰り返しを検索するときは検索式を (10)@とします。カッコ(...)で囲んだ部分が繰り返し@の対象となります。これは、(10){1,}としても同じです。

否定

上の例でワイルドカード D*i を適用すると次のような結果になります。

(1:1) In principio creavit Deus caelum & terram.
(1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas.
(1:3) Dixitque Deus. Fiat lux. Et facta est lux.

これはワイルドカードの*が空白も改行コードも飛び越えて一致するためです。それを防ぐためには、D と i の間に空白と改行コードを含まない、という条件を入れなければなりません。このとき、役に立つのが否定を表す「!」です。「!」は「それ以外」を表すので、たとえば[!a-c]は a, b, c 以外 (d, e, f, ...) を表します。

D で始まり、i で終わる単語を検索するには、「空白と改行以外」が D と i の間に連続していればよいので、D[!^13]@i>とします。または、空白や改行を含んでしまう*を使わずに D[A-z]@i>と書くこともできます。

(1:1) In principio creavit Deus caelum & terram.
(1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas.
(1:3) Dixitque Deus. Fiat lux. Et facta est lux.

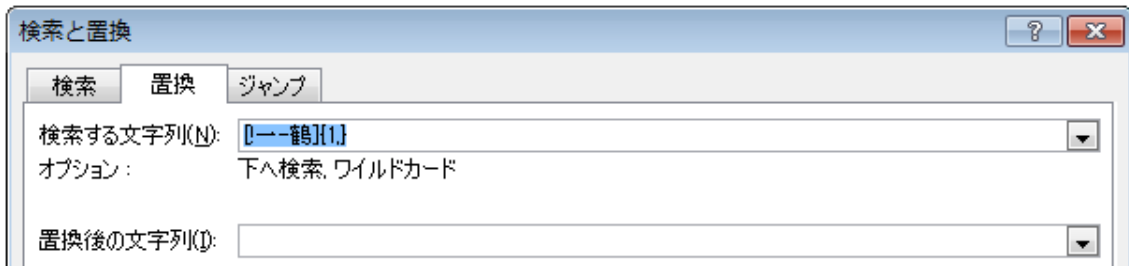
【例】以下のデータに対して、漢字以外の文字を削除します。「漢字以外」を表すには否定を用います。

¹⁰ ed も含めて 5 文字以下なので、ed の前にアルファベット 3 文字ということになります。

目的：漢字以外の文字を削除する。

◆次のように設定して検索を実行します。

- 「検索する文字列(N)」：[!一-鶴¹¹]{1,}¹²
- 「置換後の文字列(I)」：（なし）
- 「オプション」：ワイルドカード



《結果》

目的漢字以外文字削除

参照

ワイルドカードの「参照」の機能によって、高度な置換ができます。「参照」とは、検索の際にヒットした文字列を、置換のときにその部分を「参照」して利用する機能です。たとえば、「2010」「2011」などの4桁の数字を、「2010年」のように置換えたいときを考えてみてください。数字4桁を検索するには、ワイルドカードを有効にして、[0-9]{4}（あるいは[0-9][0-9][0-9][0-9]）と書くことができます（年代を抜き出すには1桁目を[1-2]としたほうが正確です）。これに「年」をつけて置換えるには、ヒットした数字が置換後も必要になります。このとき活躍するのが「参照」の機能です。

参照を利用する部分は、検索する文字列で括弧をつけておきます。この括弧をつけた部分は、「置換後の文字列」の指定で¥1と指定することで「参照」することができます。¥2とすれば、2つ目の括弧内の文

¹¹ ユニコードで漢字の先頭は「一」、最後は「鶴」なのでこのように設定します。

¹² 直前のもの（漢字）が一回以上繰り返すことを表します。この[!a]や[!a]@を指定すると1文字ずつ処理をするためページ数が多くなると時間がかかります。

字列を参照します。設定は次のようにします。

検索する文字列(N)	([0-9]{4})
置換後の文字列(I)	¥1 年
検索オプション	ワイルドカードを使用する(U)

これで 2010、2011 などそれぞれ 2010 年、2011 年に置き換わります。

【実習編】

【使用データ】 ラテン語版『創世記』冒頭部分

(1:1) In principio creavit Deus caelum & terram. (1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas. (1:3) Dixitque Deus. Fiat lux. Et facta est lux.

【課題 1】 語頭に pl, pr, tl, tr などの 2 つの子音連続がある単語を、#...# で囲み、子音連続と後続する文字の間にハイフン(-)を入れる¹³。

【方針】 子音連続部分（参照 1）とその他の部分（参照 2）を参照で置換え、ハイフンをその間に入れる。

検索する文字列

はじめの子音を[ptcbdgf]、2 番目の子音を[lr]と指定し、参照できるように全体を括弧で結びます。後続するのは任意のアルファベット文字列なので、[A-z]{1,}と指定します。これで 1 回以上のアルファベットの繰り返しを最長一致でヒットすることができます。この部分も参照で用いるので括弧で括りましょう。また、単語の境界として最初と最後にそれぞれ「<」、「>」を指定します。検索のオプションで「ワイルドカードを使用する(U)」のチェックを入れます。

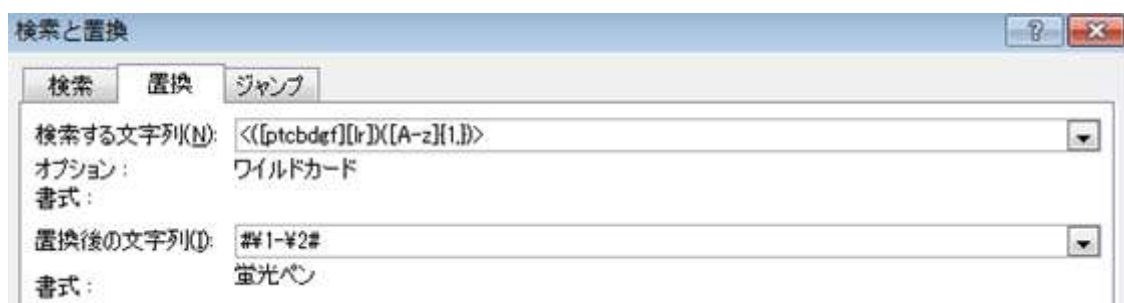
置換後の文字列

子音部分は¥1 で、後続のアルファベットは¥2 で参照できるので、それ

¹³ 検索する文字列と置換後の文字列の「書式の削除」ボタンを押します。

らをハイフンで結びます。また、#で囲むには最初と最後に直接記号を入力します。

検索する文字列(N)	<([ptcbdgf][lr])([A-z]{1,})>
置換後の文字列(I)	#¥1-¥2#
置換後の文字列(I)の書式(O)	蛍光ペン(H)
検索オプション	ワイルドカードを使用する(U)



《結果》

(1:1) In #pr-incipio# #cr-eavit# Deus caelum & terram.
 (1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem
 abyssi: & spiritus Dei ferebatur super aquas.
 (1:3) Dixitque Deus. Fiat lux. Et facta est lux.

【課題 2】 使用データの行頭にある数字（章番号と節番号）をタブで分離する（括弧とコロンは削除する）。

【方針】 数字部分は参照を利用して置き換える。括弧部分はエスケープして検索対象とし、置換後は指定しない。

検索する文字列

検索の対象となる文字列は(1:0)のように、「開きカッコ、数字連続、コロン、数字連続、閉じカッコ、スペース」です。これをワイルドカードにすると、¥([0-9]@[0-9]@¥)になります。カッコに¥をつけるのは、括弧がワイルドカードでは「参照」の対象になる、という機能があるため、文字通りの括弧であることを明示する必要があります。このため、¥を前につけてエスケープします。

次に、それぞれの数字の連続を、置換後の文字列から参照するために、([0-9]@)のように、カッコ (...) で囲みます。

置換後の文字列

置換後の文字列は、参照する 2 つの部分 を ¥1 と ¥2 で再生し、それぞれ後にタブコード (^t) を入れます。

検索する文字列(N)	¥([0-9]@):([0-9]@)¥
置換後の文字列(I)	¥1^t¥2^t
検索オプション	ワイルドカードを使用する(U)

《結果》

1	0	Cap. 1
1	1	In principio creavit Deus caelum & terram.
1	2	Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas.
1	3	Dixitque Deus. Fiat lux. Et facta est lux.

<TIPS>このようにタブコードで分離されたデータは Excel のシートにコピーすると、セルに分離して配置されます。

	A	B	C
1	1	0	Cap. 1
2	1	1	In principio creavit Deus caelum & terram.
3	1	2	Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas.
4	1	3	Dixitque Deus. Fiat lux. Et facta est lux.

【課題 3】置換の機能を使って、『創世記』の冒頭部分を単語リストにする。

【方針】(1)置換で記号部分を削除し、(2)空白を基準にそれぞれの単語を改行して表示する。

置換え(1)

はじめに、アルファベット以外の記号を全部半角の空白にします。「アルファベット以外」は否定を使って[!A-Za-z]のように表します。ただし、文章中に ç や ñ などの文字が使われている場合は、その文字を[!A-Za-zçñ]のように否定要素の中に書き込んでください。

「置換後の文字列(I)」は、半角スペースを入れます。以下の表の“_”は半角の空白（空白）を示します。アンダースコアを入力するのではなく、スペースキーで空白を入力して下さい。

検索する文字列(N)	[!A-Za-z]
置換後の文字列(I)	_
検索オプション	ワイルドカードを使用する(U)

《結果》

Cap	In	principio	creavit	Deus	caelum
terram	Terra	autem	erat	inanis	vacua
tenebrae	erant	super	faciem	abyssi	spiritus Dei
ferebatur	super	aquas	Dixitque	Deus	Fiat lux
Et	facta	est	lux		

置換え(2)

次に、1 つ以上の連続する空白を改行コードに置換します。「検索する文字列(N)」の“_”は半角のブランク（空白）を示します。「1 つ以上」を最長一致で指定することにより、半角スペースが2つある場合でも、それらをまとめて1つの改行に置き換えることができます。

検索する文字列(N)	_ {1,}
置換後の文字列(I)	^p
検索オプション	ワイルドカードを使用する(U)

《結果》

Cap
In
principio
creavit
Deus
caelum
(...)

言語の特徴を見る

スペイン語は母音をはっきりとした言語です。それは、英語やフランス語と比較すると、語尾の母音が強く表れていることで確認されます。たとえば、英語の pianist 「ピアニスト」と boat 「ボート」には、

pianista, bote が対応します。語尾の下線を引いた母音は、[pianísta ピアニスタ], [bóte ボテ]のように、しっかりと発音され、アクセントがなくても一般に弱音化したり、脱落したりすることはありません。

実はこの現代スペイン語の特徴は、中世スペイン語の一部の資料では揺れていました。次は 13 世紀の Valladolid で発行された文書の一部です。

A la por cima despues que ouiemos uistas las cartas de la una **part** & de la otra & despues de muchos razonamientos (10) todo el pleyto fue librado desta guisa. (...) Et est es el termino de villa nueva. a la **parte** de Montiel que de villa nueva fasta Montiel que la quarta **parte** sea termino de villa nueva (13) & (...)

ここに、**parte**「部分」という単語が 3 回使われていますが、語尾が一定していません。この現象については、当時（13 世紀前半）のカスティーリャ地方をはじめとするスペインにフランス人が移民したことが影響したと考えられたことがありました。フランス語では語末の母音が弱化します。13 世紀後半にその影響が少なくなって母音が回復したそうです。

現在では多くの中世スペイン語文献を資料として分析対象にすることができます。次の表はその資料を使って年代と地方ごとに計算したものです。それぞれ語尾が脱落するパーセンテージを示します。

年代	レオン地方			新カスティーリャ地方			旧カスティーリャ地方			ナバラ地方		アラゴン地方			
	ゼロ	-e	%	ゼロ	-e	%	ゼロ	-e	%	ゼロ	-e	ゼロ	-e	%	
1100								2	0.0%						
1180								1	0.0%						
1220		4	0.0%				1	1	50.0%						
1240		17	0.0%				10	12	45.5%			4		100.0%	
1260		13	0.0%	2		100.0%	4	9	30.8%	37	12	75.5%	10	2	83.3%
1280		6	0.0%	4		0.0%	3	78	3.7%	17	1	94.4%	8		100.0%
1300	1	22	4.3%	4		0.0%	4	111	3.5%	36	7	83.7%	6		100.0%
1320		6	0.0%	5		0.0%		6	0.0%	4	7	36.4%	5	2	71.4%
1340		23	0.0%	27		0.0%	1	39	2.5%	22		100.0%	28	3	90.3%
1360		27	0.0%	6		0.0%	1	66	1.5%	12		100.0%	34	1	97.1%
1380	2	25	7.4%		112	0.0%	3	61	4.7%	2		100.0%	97		100.0%
1400		50	0.0%	16		0.0%	2	120	1.6%	35	1	97.2%	165	5	97.1%
1420		120	0.0%	29		0.0%		19	0.0%				46	2	95.8%
1440		54	0.0%	1	33	2.9%		73	0.0%	15		100.0%	30	1	96.8%
1460		132	0.0%	28		0.0%		72	0.0%				27		100.0%
1480	1	15	6.3%		29	0.0%	1	87	1.1%	4		100.0%	36	1	97.3%
1500		80	0.0%	54		0.0%	1	134	0.7%	6	2	75.0%	19	28	40.4%
1520		1	0.0%	1	56	1.8%		53	0.0%		2	0.0%	17	26	39.5%
1540		1	0.0%		81	0.0%	3	12	20.0%						
1560		1	0.0%		61	0.0%		82	0.0%				3		0.0%
1580		24	0.0%		67	0.0%		3	0.0%						
1600		1	0.0%		14	0.0%		1	0.0%						
1620		3	0.0%		50	0.0%		9	0.0%						
1640		4	0.0%		10	0.0%							4		0.0%
1660		1	0.0%		58	0.0%							1		0.0%
1680					7	0.0%									
Total	4	630	0.6%	4	751	0.5%	34	1051	3.1%	190	32	85.6%	532	79	87.1%

この表を見ると、とくにナバラ地方とアラゴン地方で、13世紀に限らず、さらに15世紀にいたるまで語尾の脱落が多かったことがわかります。一方、レオン地方や新・旧カスティーリャ地方ではきわめて少数です。

どうやら母音の弱化はカスティーリャ地方ではなく、むしろ東部のナバラ地方とアラゴン地方の特徴であって、西部のレオン地方と中央部のカスティーリャ地方では、母音が比較的よく保持されていたようです。そして、イベリア半島最東部に位置するカタルーニャ語では、母音が脱落するのが普通ですし、また、最西部のポルトガル語では母音がよく保たれています。

このように、多くの資料で調べてみると、中央部のカスティーリャ地方では母音が脱落することは少なく、むしろそれは東部地域の特徴であったことがわかります。現代スペイン語はイベリア半島の中央部を占めた中世カスティーリャ地方の話し言葉が基礎になっています。この資料によって、スペイン語の安定した母音の特徴が歴史的にも地理的にも確かめられます。

4 Excel 入門

【目標】 Excel の使い方をマスターし、「集計」ができるようになる。また、「分析」のためのデータの視覚化やデータの選択（フィルタリング）の方法を身につける。

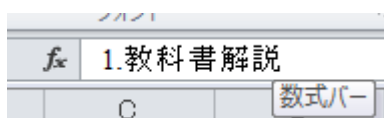
本章では Excel の使い方を見ていきます。Excel は「表計算ソフト」と呼ばれるアプリケーションで、非常に多くの機能を有し、言語データ分析においても有益なツールですが、文系の人だとあまり使ったことがないという人も多くいます。以下、最低限身につけておくべき Excel の基本的な使い方からはじめ、分析に役立つ実践的な使い方を紹介していきます。用いるデータは、分かりやすさを優先して身近な題材を使い、徐々に言語分析に応用できるよう足場を固めていきます。

◇24 Excelの基礎

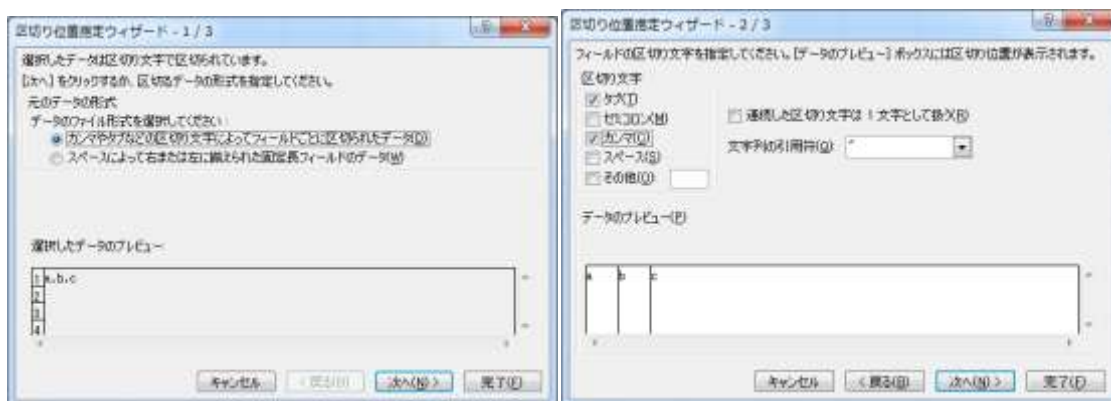
[1] Excel ファイルの構成

Excel のファイルは「ブック」と呼ばれます。ファイルの拡張子は xls (Excel 2003) またはxlsx (Excel 2007, 2010) です。ブックにはデフォルトの状態では 3 枚の「シート」があります (Sheet1, Sheet2, Sheet3)。シート名をダブルクリックすると、シートの名前を変更することができます。また、右クリック→「シート見出しの色」で、色を変更することも可能です。それぞれのシートには「セル」Cell と呼ばれる升目があります。縦の列全体を「列」(column) と呼び、横の行全体を「行」(row) と呼びます。列は A, B, C, ... というアルファベットで表示し、行は 1, 2, 3, ... という整数で示します。たとえば C2 という表示は C 列と 2 行が交差する位置のセルを示します。

TIPS セルを選択した状態で数字や文字を入力すると、上書きされてしまい、元の値は書き換えられてしまいます。セルの値を編集するには、セルをダブルクリックするか、選択後に「数式バー」で編集をします。



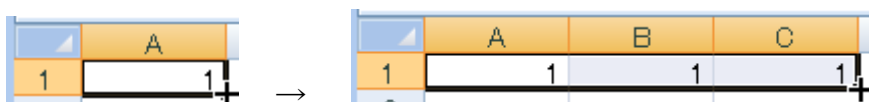
<TIPS> 外部からデータを入力する場合、タブ区切りのものであれば自動的にセルに分割して貼り付けられます。タブ以外のものでデータを区切りたいときは、「データ」→「区切り位置」で、「カンマやタブなどの区切り文字によってフィールドごとに区切られたデータ」を選択し、対象となる記号を選択するか入力して下さい。



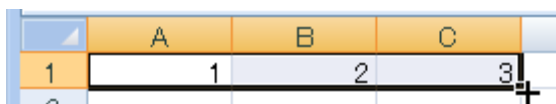
[2] オートフィルによる入力

連番を入力する

規則的に配列するデータを入力するときはオートフィルの機能を使うと便利です。同じデータを繰り返してコピーするにはコピー元のセルの右下に現れるプラス (+) のハンドルをドラッグします。たとえば、A1に「1」と入力してそれを右のセルにドラッグしてみましょう。

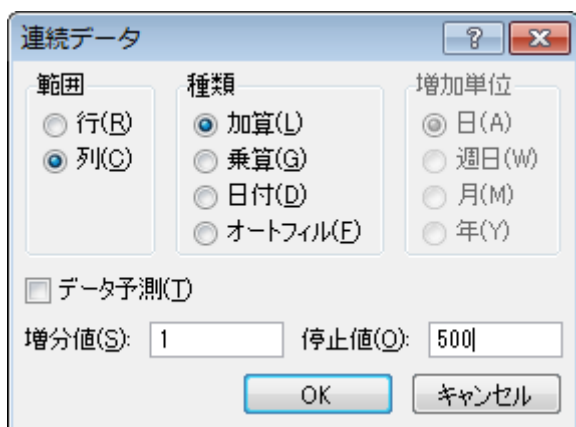


[Ctrl]キーを押しながらドラッグすると連番になります。



◆ドラッグによらないオートフィルは、初期値があるセルを選択し、「ホーム(H)」→「フィル(FI)」→「連続データの作成(S)」¹⁴

¹⁴ 非常に多くのオートフィルをするときに使います。



文字列と数字の組み合わせ

B1に「第1回」を入力します。このように文字と数字が組み合わさったデータのハンドルをドラッグすると自動的に連番がつきます。逆に、[Ctrl]を押しながらハンドルをドラッグするとセルの内容がそのままコピーされます。

	A	B	C	D	E	F
1	ID	第1回	第2回	第3回	第4回	第5回

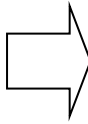
<Tips> 右下に現れるプラス(+)のハンドルをダブルクリックすると、隣接する列（一番長い列）にあるデータの分だけオートフィルが実行されます。

日付の入力

セルに4/6と書くと、自動的に「4月6日」という表示になります¹⁵。これはスラッシュを含む数字列が日付として認識されるためです。次のセルを4/13にし、そのセルの+ハンドルをダブルクリックすると、以下は等差的にオートフィルされます。

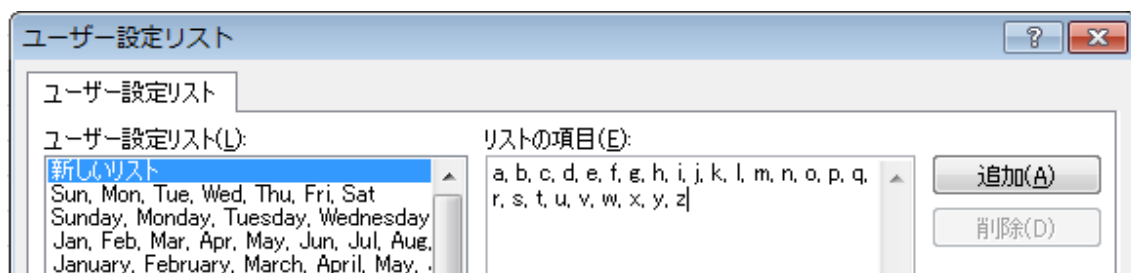
¹⁵ このとき入力時の年の情報も追加されます（2011年など）。セルを右クリック→「セルの書式設定(F)」で表示方法を変えることができます。たとえば6-Aprや年も表示させて2011/4/6などのようにすることもできます。

	A	B
1	回	月日
2	第1回	4月6日
3	第2回	4月13日
4	第3回	
5	第4回	



	A	B
1	回	月日
2	第1回	4月6日
3	第2回	4月13日
4	第3回	4月20日
5	第4回	4月27日
6	第5回	5月4日
7	第6回	5月11日
8	第7回	5月18日
9	第8回	5月25日
10	第9回	6月1日
11	第10回	6月8日
12	第11回	6月15日
13	第12回	6月22日
14		

<TIPS> アルファベットはオートフィルされず、a,b を選択してオートフィルにすると、a,b,a,b,a,b...の繰り返しになります。a-z のオートフィルを設定するには「ユーザー設定のリスト」を作ります。「ファイル(F)」→「オプション(T)」→「詳細設定」→「全般」の「ユーザー設定リスト」の「新しいリスト」で「a, b, c, ...」を用意します。



[3] 選択

それぞれの対象を選択するには、次の操作をします。

セル	セルをクリックします。
セル内の文字	位置をダブルクリックします。(あるいは、F2 キー)
列	シートの上にある A, B, C, ...などの列記号をクリックします。
行	シートの左にある 1, 2, 3, ...などの行番号をクリックします。
シート全体	シートの左上の三角形 (列文字と行番号がぶつかる A1 セルの左上側にあたる位置) をクリックします。

* 選択範囲を変更するショートカットキー

複数列・行	Ctrl キーを押しながら選択
連続列・行	先頭の位置をクリックし、Shift を押して最後の位置をクリック
行全体に拡張	Shift+Space
列全体に拡張	Ctrl+Space
シート全体に拡張	Shift+Space, Ctrl+Space
選択範囲を先頭まで拡張	Ctrl+Shift+Home
選択範囲を最後まで拡張	Ctrl+Shift+End
入力された範囲の連続	Ctrl+A ¹⁶ , Ctrl+Shift+*
入力された範囲の先頭行まで	Ctrl+Shift+↑
入力された範囲の末尾行まで	Ctrl+Shift+↓
入力された範囲の左端列まで	Ctrl+Shift+←
入力された範囲の右端列まで	Ctrl+Shift+→
選択範囲の一部を修正	Shift を押しながら、カーソル移動キー（↑ ↓ ← →）
選択範囲を解除	いずれかのセルをクリック
最初のセル	Ctrl+Home
最後のセル	Ctrl+End
前のシート	Ctrl+PageUp
次のシート	Ctrl+PageDown

[4] ステータスバー

データを選択すると、その範囲の簡単な統計量がステータスバーに表示されます。表示内容はステータスバーを右クリックして、選択することができます。

¹⁶ 入力されていないならば、全体の範囲になります。

s	3	5	3	3	4	4	3
t	4	4	2	4	3	5	3
u	3	3	4	4	4	5	2
v	4	5	3	2	2	2	3
w	5	5	4	4	4	5	3
x	2	4	3	4	3	3	2
値	24	24	24	24	24	24	24
	5	5	5	5	5	5	3

2.3 / 2.1a / 2.1b / 2.1c / 2.1d / 2.1e / 2.1f / 2.2 / 2.3 / 2.3a / 2.3b / 教材 / 2.3 (2)

平均: 3.548611111 データの個数: 175 数値の個数: 144 最小値: 1 最大値: 5 合計: 511

<TIPS> 平均は、選択した範囲の書式に応じて、小数点以下が表示されます。書式を「標準」または「数値」で小数点以下を1以上に指定すれば、オートカルクの平均の小数点以下も表示されます。以下の図では、A1~A4は小数点以下の表示は1桁ですので、オートカルクの結果も同じように表示されます。一方、A5にはそのような設定をしていないので、小数点以下2桁まで表示されています。

	A	B	C	D
1	4.0			
2	5.0			
3	3.0			
4	1.0			
5	3.25			
6				
7				
8				

平均: 3.3 データの個数: 4 合計: 13.0

[5] 書式

Word またはエディター使ってタブ区切りのデータを作り、それをExcelで読み込みましょう。

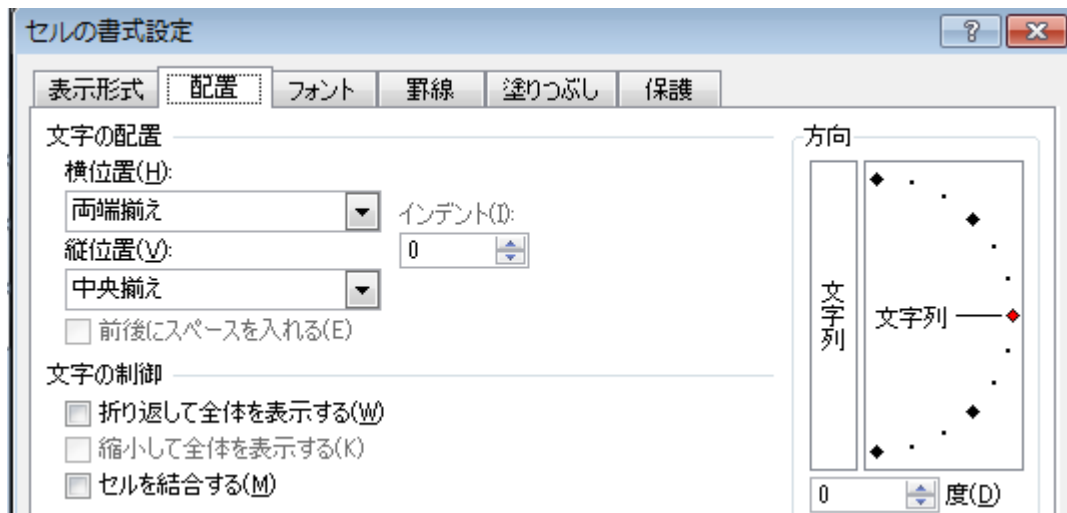
例：

Reverse	Word	Pos	Sum	Disp	Usage
a	a	Prep	22814	0.963	21971.788
abart	traba	N	11	0.613	6.741

	A	B	C	D	E	F
1	Rever	Word	Pos	Sum	Disp	Usage
2	a	a	Prep	22814	0.963	21971.79
3	abart	traba	N	11	0.613	6.741

◆書式を設定するには、セルを選択し、右クリック→「セルの書式設定(F)」 (あるいは、[Ctrl] + [1])

「セルの書式設定(F)」には、次のタブがあります。

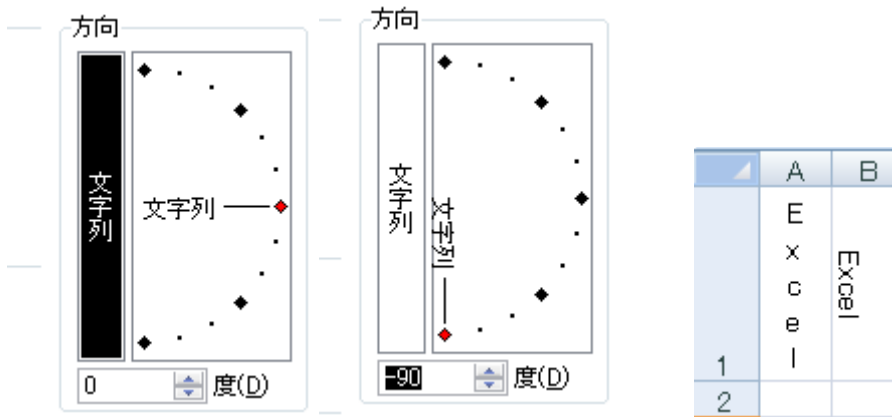


セルからはみ出した文字列を表示するには 2 つの方法があります。1 つは「折り返して全体を表示する(W)」 (A2) で、もう 1 つは「縮小して全体を表示する(K)」です (A3)。

	A	B
1	Excelを使った分析	
2	Excelを 使った分 析	
3	Excelを使った分析	
4		
5		

文字列が長い場合は、縮小すると見えなくなることもありますので、セルの幅を調整するとよいでしょう。

TIPS 文字を縦書きにするには、「セルの書式設定」→「配置」→「方向」で設定ます。



縦書きで「文字列」と書かれているところをクリックすると、A1 のようになり、分度器のようなところにある「文字列」をドラッグすると、B1 のようになります。

<TIPS> 8.60423E-8 などのような数字は 8.60423×10^{-8} (= 0.00000000860423) という意味で、ほとんどゼロに近い数字になります。セルの書式を小数点以下 3 にすると 0.000 になります。Excel ではある一定の値以上の数値はこのような指数表記になります。指数の部分をプラスにしても同じです。ためしにセルに 5E+3、または 5E3 と書き込むと 5000 が表示されます。E は Exponential (べき乗) の頭文字で、たとえば 5E3 は 5×10^3 を示します。

[6] Word へコピー & ペースト

Excel のデータを Word へコピーするには、いくつかの方法があります。

そのまま貼り付ける

Word にそのまま貼り付けるには、Excel で範囲を選択し、そのままコピー・アンド・ペーストします。この場合、Word では表として扱われるようになります。色や枠線などもコピーされます。

Chapter	Section	Sentence
1	1	In principio creavit Deus caelum & terram.
1	2	Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas.
1	3	Dixitque Deus. Fiat lux. Et facta est lux.

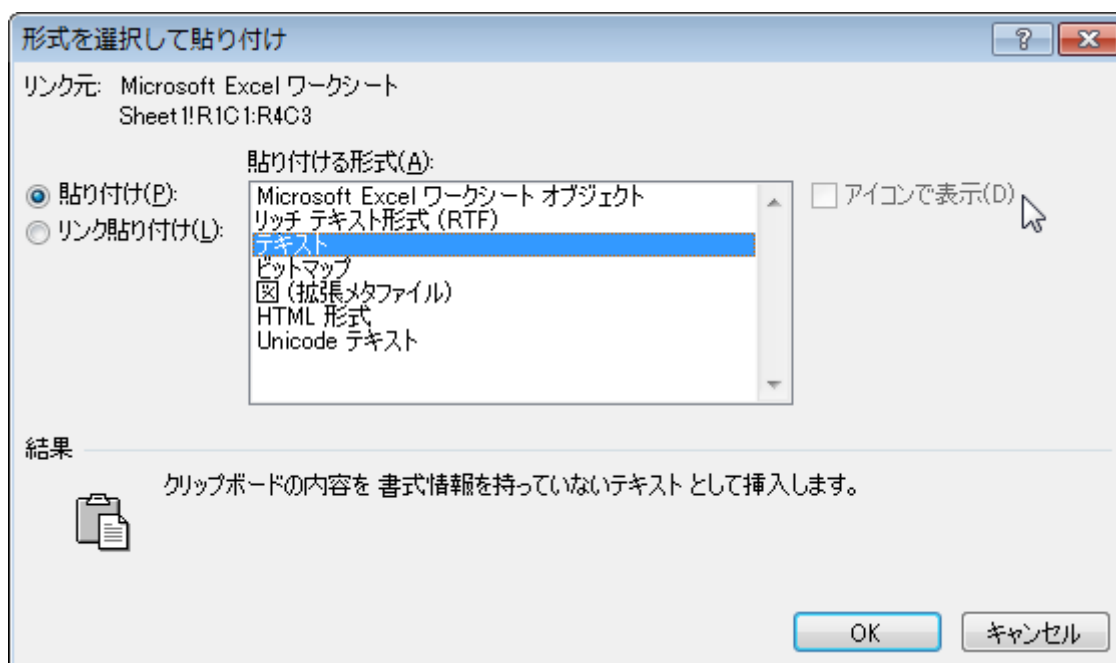
<TIPS> 表全体の大きさをページの範囲に合わせるときは、表の一部を選択し、→「表ツール(JL)」→「レイアウト(F)」→「セルのサイズ」グループの「自動調整」→「ウインドウのサイズに合わせる(W)」

Cha	Sect	Sentence
pter	ion	

1	1	In principio creavit Deus caelum & terram.
1	2	Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas.
1	3	Dixitque Deus. Fiat lux. Et facta est lux.

テキストとして貼り付ける

Word にテキストとして貼り付けるには、「ホーム(H)」→「貼り付け(V)」→「形式を選択して貼り付け(S)」→「貼り付ける形式(A)」を「テキスト」とします。



次のように、タブコードで区切られたテキスト形式で貼り付けられます。

- (1:1) In principio creavit Deus caelum & terram.
 (1:2) Terra autem erat inanis & vacua: & tenebrae erant super faciem
 abyssi: & spiritus Dei ferebatur super aquas.
 (1:3) Dixitque Deus. Fiat lux. Et facta est lux.

画像として貼り付ける

◆ Word に画像（「ビットマップ」）として貼り付けるには、「ホーム(H)」→「貼り付け(V)」→「形式を選択して貼り付け(S)」→「貼り付ける形式(A)」を「ビットマップ」とします。ビットマップとして貼り付けると、その内容は編集できなくなりますが、大きさや場所は比較的自由に変更できます。

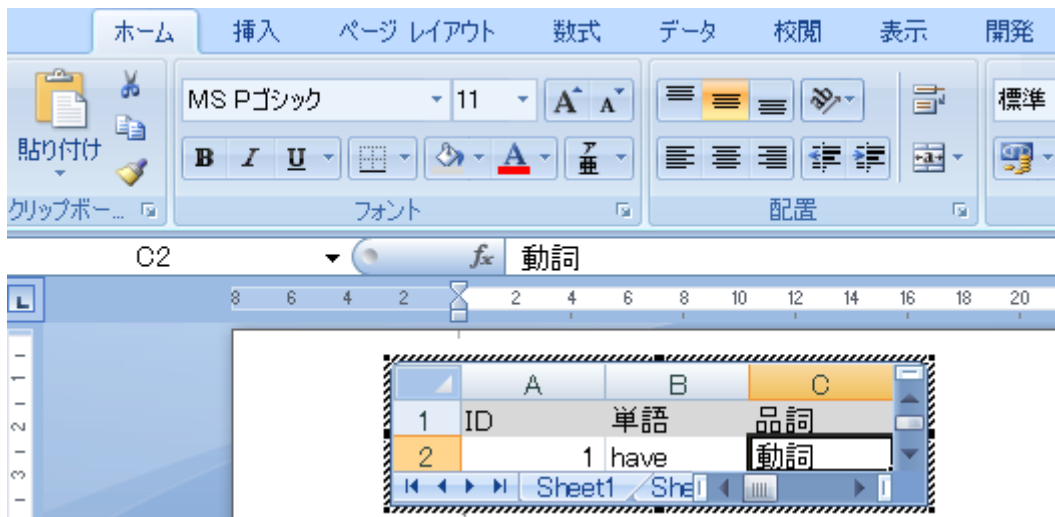
1	0	Cap. 1
1	1	In principio creavit Deus caelum & terram.
1	2	Terra autem erat inanis & vacua: & tenebrae erant super faciem abyssi: & spiritus Dei ferebatur super aquas.
1	3	Dixitque Deus. Fiat lux. Et facta est lux.

Excel のシートとして貼り付ける

最後に、「Microsoft Excel ワークシートオブジェクト」で Word に入りつけてみましょう。「ホーム(H)」→「貼り付け(V)」→「形式を選択して貼り付け(S)」→「貼り付ける形式(A)」→「Microsoft Excel ワークシートオブジェクト」で貼り付けます。

ID	単語	品詞
1	have	動詞

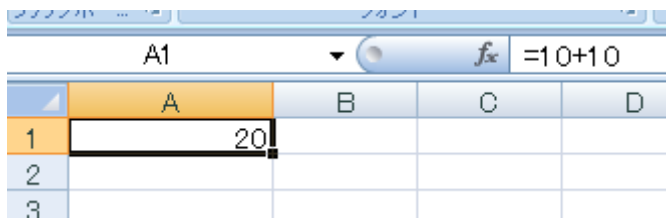
一見、ビットマップと同じように見えますが、ダブルクリックをすると編集できます。また、エクセルの機能もそのまま使えます。つまり、Word の中に Excel を埋め込むようなイメージです。



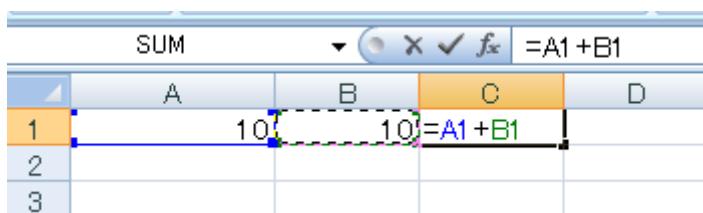
データを変更する可能性がある場合はこの形式を使うと便利です。

◇25 四則の計算

Excel の表の中では様々な計算をすることができます。計算式を入力するときは、「=」で始めます。たとえば、A1に「=10+10」を入力して[Enter]を押すと、20という値が表示されます。なお、掛け算は「*」を使い、割り算には「/」を使います。



セル同士の値を使って演算することも可能です。「=」入力後に、計算したいセルを選択するか、「A1」などのように入力します。



TIPS セル同士の文字を結合するには、「&」を用います。C1にA1とB1を合わせた内容を入力するには、=A1 & B1とします。

	C1		f_x	=A1 & B1
	A	B	C	D
1	言語情報	分析	言語情報分析	
2				
3				

語彙学習の実験

外国語の語彙はどのように学習・獲得されるのでしょうか？それにはいろいろな条件が関係しているようです。たとえば、語形の難易度、使用頻度、既習項目、意味などが考えられます。

次の図は 20 の「芸術」に関連するスペイン語単語(*rima* ‘ryme’, *novela* ‘novel’, *tragedia* ‘tragedy’, *comedia* ‘comedy’, *ópera* ‘opera’, *ritmo* ‘rythm’, *dúo* ‘duo’, *trío* ‘trio’, *cuarteto* ‘quartet’, *orquesta* ‘orchestra’, *violín* ‘violin’, *órgano* ‘organ’, *flauta* ‘flute’, *arpa* ‘harp’, *armónica* ‘harmonica’, *castañuela* ‘castanet’, *acuarela* ‘watercolor’, *pincel* ‘paintbrush’, *lienzo* ‘canvas’, *escultura* ‘sculpture’)について 3/4 年生 25 名の(1)第 1 回テスト(初期状態), (2)第 2 回テスト(10 分の記憶練習), (3)第 3 回テスト(1 週間後の再テスト)の結果です。それぞれのテストでは意味を示し、語形を答えてもらいました。

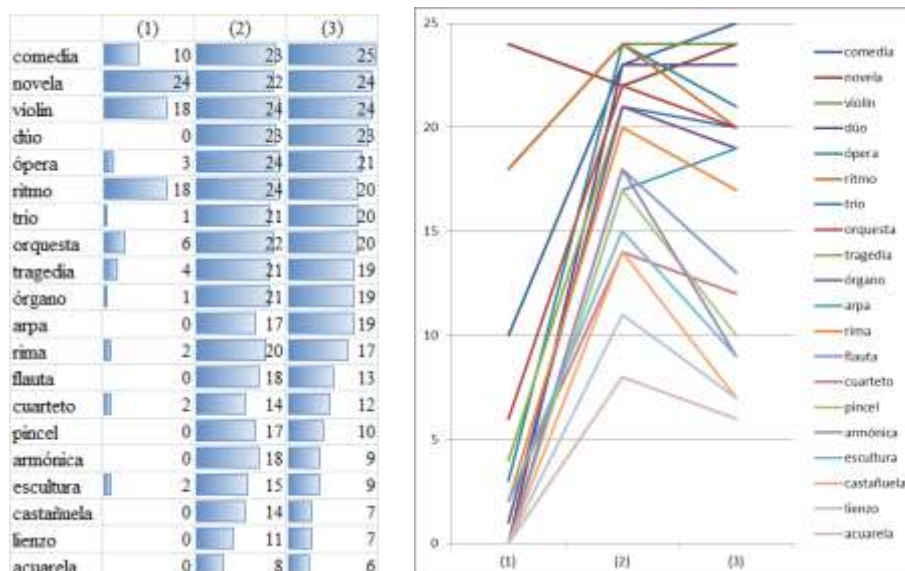


Fig. 1. 芸術

最終結果である(3)の列でソートしてありますが、これらは易から難の順に並んでいると思われれます。既習の英語や外来語に形が近い語は容易になるようです。

次は、同じテストを「食品・食事」に関連する語彙で行った結果で

す。 *sopa* 'soup', *consomé* 'consome', *potaje* 'potage', *apio* 'celery', *espárrago* 'asparagus', *perejil* 'parsley', *pimienta* 'pepper', *berenjena* 'eggplant', *cebolla* 'onion', *calabaza* 'pumpkin', *col* 'cabbage', *zanahoria* 'carrot', *espinaca* 'spinach', *higo* 'fig', *sandía* 'watermelon', *pera* 'pear', *jamón* 'ham', *queso* 'cheese', *sardina* 'sardine', *bacalao* 'cod fish'. 今回は 35 名が参加しました。ここでは(2)で相互教育学習法 (2名の学習者による質問・解答練習) を適用しました。(2)の成績がかなり高くなっていることがわかります。しかし、(3)での成績低下も目立ちます。

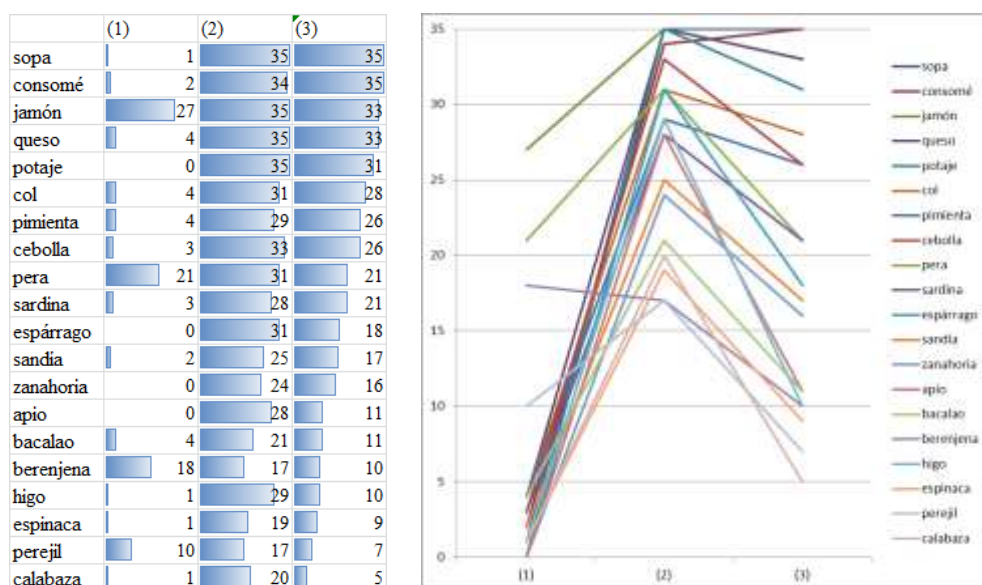


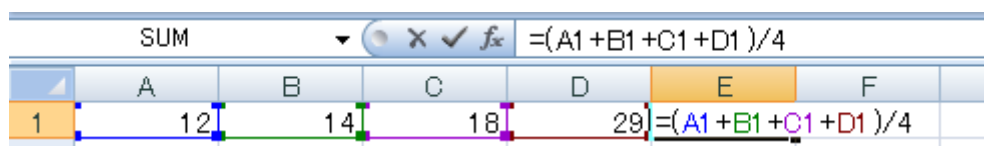
Fig. 2. 食品・食事

ここで *col* 'cabbage', *pera* 'pear', *apio* 'celery', *higo* 'fig' に注目してみましよう。どれも 1-2 音節の短い単語ですから、記憶の負担は少ないはずですが。しかし、*col* に比べて、*pera*, *apio*, *higo* の成績がこの順番で悪くなっています。この原因は単語の意味にある、という仮説を立てることができるでしょう。学習者にとって意味があまり重要でない単語・日常生活であまり出会うことのない意味の単語は記憶されにくい、ということがあるのかもしれませんが。逆に *pimienta* 'pepper' や *cebolla* 'onion' の成績がよいことも、その裏付けになっていると思われます。

◇26 関数

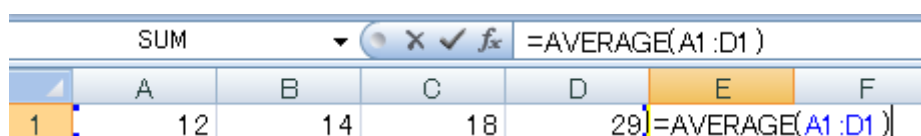
Excel には多くの「Excel 関数」と呼ばれる機能が用意されています(以下では簡単に「関数」と呼びます)。関数は、複雑な計算手順をひとまとめにして行うことができます。たとえば次のような 12, 14, 18, 29 というデータの平均を求める場合、E1 のところにセルをクリックしな

から「=(A1+B1+C1+D1)/4」と入力すると値を求めることができます。



	A	B	C	D	E	F
1	12	14	18	29	= (A1+B1+C1+D1)/4	

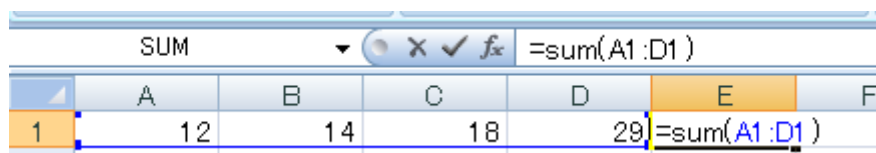
関数を使うと、より簡単にできます。E1に「=average(」までを入力し、A1 から D1 までを選択し、カッコを閉じれば値が出てきます (=average(A1:D1)となります)。



	A	B	C	D	E	F
1	12	14	18	29	=AVERAGE(A1:D1)	

このように関数は、(...)の中に値を入れれば、その関数の値（ここでは平均）を返します。（...）の中の値を「引数」（ひきすう）と呼びます。「返す」という意味は、「その引数で計算した結果を出す」という意味です。

【例】合計を求める SUM 関数を使ってみましょう¹⁷。



	A	B	C	D	E	F
1	12	14	18	29	=sum(A1:D1)	

計算式の入力には 3 つの方法があります。

(1) クリックで範囲を指定

「=SUM(」と書いてセルを選択し、括弧を閉じエンターキーを押します。

(2) コロンで範囲を指定

セル E1 を選択し、セルに=SUM(A1:D1)と書き込み、エンターキーを押します。またはセルを選択した状態で上にある「数式バー」に入力することもできます。

(3) コンマで範囲を指定

17

セル E1 を選択し、=SUM(A1,B1,C1,D1)と書き込みます。クリックでセルを1つずつ選択し、コンマで区切ることもできます。この方法を使うと、離れたセルを計算の対象に含めることができます。

数式を書き込んだセルには、計算の結果が表示されますが、クリックすると「数式バー」に数式が表示され、ダブルクリックまたは[F2]キーで数式がセルに表示され編集が可能になります。

代表的な Excel 関数は次のようなものがあります。

= AVERAGE (...)	平均を求める
= COUNT (...)	個数を求める
= MAX (...)	最大値を求める
= MEDIAN (...)	中央値を求める
= MIN (...)	最小値を求める
= MODE (...)	最頻値を求める
= STDEVP (...)	標準偏差を求める
= SUM (...)	和を求める

<TIPS> これらの関数は「オート SUM」を使うこともできます。「オート SUM (U)」のアイコンは「ホーム(H)」のリボンの右の方にある「編集」グループの中にあります。



対象のセル範囲と和を書き出すセルを含めて選択し、「オート SUM(U)」のアイコンをクリックします。

コピーによる関数の入力

次のようなデータのあるシートで E1 のセルを E2 にコピーしてみましよう。

	A	B	C	D	E	F
1	12	14	18	29	73	
2	13	15	20	31	79	
3						

すると、`=SUM(A1:D1)`が`=SUM(A2:D2)`になっていることがわかります。関数は、このように相対的にコピーすることができます。多くのデータが対象となる場合、セルのコピーをうまく利用することで作業を効率化できます。

IF 関数

IF 関数を使えば、セルの値を判断して、数式や文字列を入力できます。この関数は、基準に合うデータを見つけたり、成績を判定したりするのに大変便利です。IF 関数は次のような引数を持ちます。

`=IF(条件, “真の場合”, “偽の場合”)`

次の表のデータで、頻度が 300 以上の場合、「*」を表示する式を考えてみましょう。

	A	B	C
1	単語	頻度	
2	appear	225	
3	belong	259	
4	cough	48	
5	cry	154	
6	fall	325	
7	fry	87	
8	laugh	87	
9	live	517	
10	open	425	
11	smile	98	

判定結果は C 列に出しますので、ここに数式を入力します。条件は、`B2>=300` で、真の場合は「*」、偽の場合は何も表示しないので、次の式が立ちます。

C2=IF(B2>=300, “*”, “”)

	A	B	C	D	E
1	単語	頻度	ランク		
2	appear	225	=IF(B2>=300, “*”, “”)		
3	belong	259	[IF(論理式, [真の場合], [偽の場合])]		
4	cough	48			
5	cry	154			
6	fall	325			
7	fry	87			
8	laugh	87			
9	live	517			
10	open	425			
11	smile	98			

これをコピーすると、次のようにすべてのセルに関して判定ができます。

	A	B	C	D
1	単語	頻度	ランク	
2	appear	225		
3	belong	259		
4	cough	48		
5	cry	154		
6	fall	325	*	
7	fry	87		
8	laugh	87		
9	live	517	*	
10	open	425	*	
11	smile	98		

IF 関数の埋め込み

「*」だけではなく、「***」、「**」、「*」のように3段階で表示したい場合は、IF 関数を埋め込みます。ここでは頻度 300 以上は「***」、200 以上は「**」、100 以上は「*」とします。まず、「頻度 300 以上は***」から考えます。

C2=IF(B2>=300, “***”, “”)

真の場合は「***」を入力して終わりですが、偽の場合、さらに判定を続ける必要があります。次は、セルの値が 200 以上かどうかを判定する必要がありますので、入力する式は次のようになります。

C2=IF(B2>=300, “***”, IF(B2>=200, “**”, “”))

↑ 偽の場合、更に判定を続ける

閉じ括弧が 2 つになることに注意してください。さらに、100 以上ならば「*」、それ以外は空欄となるよう、もう 1 つ式を埋め込みます。

C2=IF(B2>=300, “***”, IF(B2>=200, “**”, IF(B2>=100, “*”, “”)))

この数式を C11 までコピーすると次のようになります。

	A	B	C	D
1	単語	頻度	ランク	
2	appear	225	**	
3	belong	259	**	
4	cough	48		
5	cry	154	*	
6	fall	325	***	
7	fry	87		
8	laugh	87		
9	live	517	***	
10	open	425	***	
11	smile	98		
12				

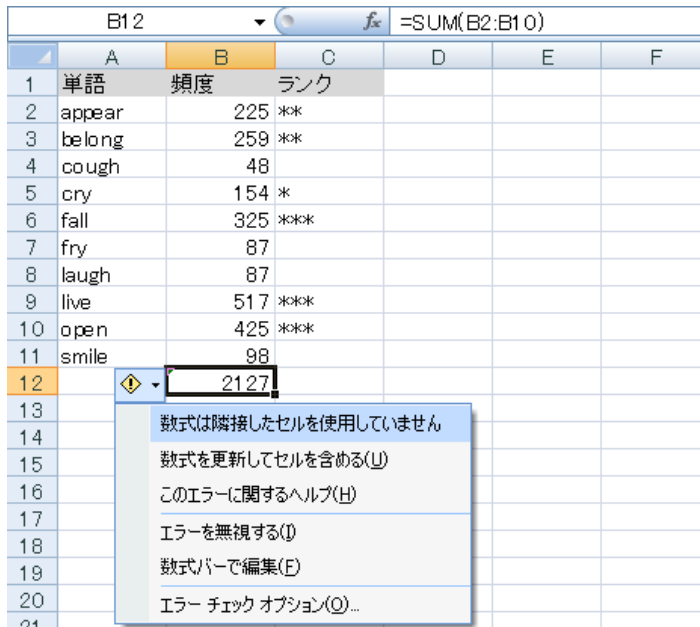
TIPS 「A1 が 60 以上かつ B1 が 60 以上」のように、複数の条件を設定するには、AND 関数を用います。AND 関数は AND(条件 1, 条件 2 …)のように、括弧の中にコンマ区切りで条件を並べていきます。

C1=if(AND(A1>=60, B1>=60), “○”, “”)

C1		fx =IF(AND(A1>=60, B1>=60), “○”, “”)					
	A	B	C	D	E	F	G
1	65	66	○				
2	58	64					
3	90	88	○				
4	45	56					
5	77	59					
6	57	60					
7	92	93	○				

<TIPS> 関数を使用すると、セルの左上に緑の三角形が出る場合があります。これは Excel 関数を使用したときにエラーの可能性あることを警告しているものです。たとえば、次の場合、B2:B10 の範囲を SUM 関数で求めています、B11 にもデータがあるので、警告が出ています。ここでは「数式は隣接したセルを使用していません」という

警告ですが、これは「隣接したセル全部を…」と解釈すればよいでしょう。しかし、これは意図的にいくつかのセルを省いたのであれば問題ないので、「エラーを無視する(I)」をクリックしてください。



Excel の主なエラーメッセージは以下のとおりです。

#####	数値をすべて表示するにはセル幅がたりません。
#DIV/0! (ディバイド・パー・ゼロ)	ゼロや空欄のセルで割り算をしています。
#N/A (ノットアベイラブル)	指定した値が適切でないか有効な値が見つかりません。
#NAME? (ネーム)	関数名や範囲名が間違っています。
#NULL! (ヌル)	存在しないセル範囲を参照しています。
#NUM! (ナンバー)	引数の数値が適切な範囲を超えています。
#REF! (レファレンス)	参照先のセルが存在しません。
#VALUE! (バリュー)	引数の種類が間違っています。

文字列関数

Excel では数値の計算だけではなく、文字列を操作する関数も用意

されています。

=LEN (文字列)	文字数を求める
=LENB (文字列)	文字のバイト数を求める
=UPPER (文字列)	すべて大文字に変換する
=LOWER (文字列)	すべて小文字に変換する
=PROPER (文字列)	語頭だけ大文字に変換する
=LEFT (文字列, 文字数)	語頭から指定分の文字数を抜き出す
=RIGHT (文字列, 文字数)	語末から指定分の文字数を抜き出す
=MID (文字列, 開始位置, 文字数)	開始位置から指定分の文字数を抜き出す

これらの関数を「abcdefg」と「あいうえお」を対象に計算してみます。

	LEN	LENB	UPPER	LOWER	PROPER
abcdefg	7	7	ABCDEFG	abcdefg	Abcdefg
あいうえお	5	10	あいうえお	あいうえお	あいうえお

LEN は半角全角の区別なくその文字数を、LENB は全角の文字は 2 バイトなので「あいうえお」は 10 になっています。また、UPPER、LOWER、PROPER は日本語には無効なため、結果に変化はありません。

	LEFT-3	RIGHT-3	MID-3, 2
abcdefg	abc	efg	cd
あいうえお	あいう	うえお	うえ

上の結果は、LEFT と RIGHT はそれぞれ左と右から 3 文字ずつ抜き出したものです。また、MID に関しては「3 文字目」（それぞれ「c」と「う」）から、それを含めて 2 文字抜き出した結果です。

【例】 下のようなデータで、「文字数」、「語頭の文字」（小文字に変換）、「語末の文字」、「語末の 2 文字」を抜き出してみましょう。

E2		fx			=RIGHT(A2,2)
	A	B	C	D	E
1	単語	文字数	語頭	語末	語末-2
2	Cap	3	c	p	ap
3	In	2	i	n	In
4	principio	9	p	o	io
5	creavit	7	c	t	it

- B2 =LEN(A2)
- C2 =LOWER(LEFT(A2))
- D2 =RIGHT(A2)
- E2 =RIGHT(A2,1)

これをそれぞれ下にコピーすれば完了です。

関数

言語データ分析に限らず一般にデータ分析をするとき「関数」はとても幅広く使われます。Excelにははじめから多くの関数が用意されているので、それを知っておくと便利です。和(SUM)や平均(AVERAGE)は関数を使わなくてもどうにか計算できますが、最大値(MAX)、最小値(MIN)、中央値(MEDIAN)などは複雑な手順が必要になります。ところが、関数を使えばカッコ(...)の中に引数を入れると簡単に一義的に値を返してくれます。まるでお金を入れると商品を返す自動販売機のようなのです。

このように関数はとても便利なのですが、はじめから関数を使って操作ばかり覚えてしまうと、操作方法はわかっても、操作そのものがブラックボックスになってしまい、操作の意味がわからなくなります。和や平均ならば簡単にイメージできるのですが、少しずつ複雑な概念を扱うようになると、これも同じように関数に頼ることになり、実際に何が行われているのかわからないまま、その結果だけを見ることになりかねません。ちょうど、自動販売機から出てきた商品をそのまま消費するような感じです。

しかし、数式の導出過程を紙に書いて納得することも重要です。そして納得したら、こんどはできるだけ関数を使わないで、紙と鉛筆、電卓、または Excel で数式どおりの方法を実験します。その結果と関数が出す結果が同じであることを確認します。

数学の公式の導出方法を忘れても、とにかく公式の適用のスキルだけで問題を解くことができます。このほうが、間違いがなく、効率的です。しかし、結果の説明の中にブラックボックスがひそんでいて、その説明を依頼されても、導出方法を知らなければ、公式がこうなのだから、と言うだけで説明ができません。

研究をする上で、効率的な手順・方法というのはもちろん重要な側面です。しかし、その根底にある仕組みを理解していなければ、その結果に完全に責任がもてません。Excel 関数を利用するときはちょっと立ち止まってその意味と導出過程を考えてみましょう。そうしなければ、自動販売機（関数）でしか買い物ができないという状態になってしまうかもしれません。

◇27 相対参照と絶対参照

数式は繰り返し同じものを入力することがあります。たとえば、次のようなデータでは、ID の 1 の F2 に「B2+C2+D2+E2」という数式が入っていますが、F3 と F4 にも合計を計算する式を入れることになります。

F2							fx	=B2+C2+D2+E2
	A	B	C	D	E	F		
1	ID	項目A	項目B	項目C	項目D	合計		
2	1	5	4	4	3	16		
3	2	3	2	4	3			
4	3	1	2	1	1			

このとき、F2 の数式を F3 にコピーすると、次のように自動的に「B3+C3+D3+E3」となります。

F3							fx	=B3+C3+D3+E3
	A	B	C	D	E	F		
1	ID	項目A	項目B	項目C	項目D	合計		
2	1	5	4	4	3	16		
3	2	3	2	4	3	12		
4	3	1	2	1	1			

これは「相対参照」と呼ばれるもので、数式の参照先が、入力箇所に応じて相対的に変更されていきます。前節では関数の入力でも同じ方法で計算することを見ました。

一方、計算内容によっては、参照先を固定したい場合があります。その場合は、「\$」をつけると参照先を固定することができます。これを「絶対参照」といいます。この概念を理解するため、次の簡単な例を見

てみましょう。

	A	B	C	D	E
1	商品	定価	販売価格		割引率
2	A	1200	960		0.8
3	B	1000	=B3*E3		
4	C	580			
5	D	300			

「販売価格」は「定価」×「割引率」で計算します。C2の数式「=B2*E2」を、C3にコピーすると、「割引率」の参照先がE3に移動してしまいます。これを防ぐには、E2を\$E\$2と絶対参照にしてコピーします。すると、「定価」の方は相対参照で動いていきますが、「割引率」はE2の内容に固定されます。

	A	B	C	D	E
1	商品	定価	販売価格		割引率
2	A	1200	960		0.8
3	B	1000	800		
4	C	580	464		
5	D	300	240		

絶対参照は列または行のみを指定することができます。この例では、列は動きませんので、E\$2のように行だけ固定した形式でも入力できます。

「\$」を挿入するショートカットは[F4]キーです。「数式バー」の該当する文字の位置にカーソルを置いて、押すと1回目は\$E\$2と行列の両方に、2回目はE\$2となり行だけに、3回目は\$E2と列方向だけが絶対参照になります（4回目では相対参照に戻ります）。

◇28 データの視覚化

Excelでは表に入力されたデータをわかりやすく視覚化する機能があります。ここでは条件付き書式とグラフについて簡単に見ていきます。

[7] 条件付き書式

データバー

数値だけでは目で比較しにくいときに、データバーが役立ちます。

◆数値のあるセルを選択し、「ホーム(H)」→「条件付き書式(L)」→「データバー(D)」



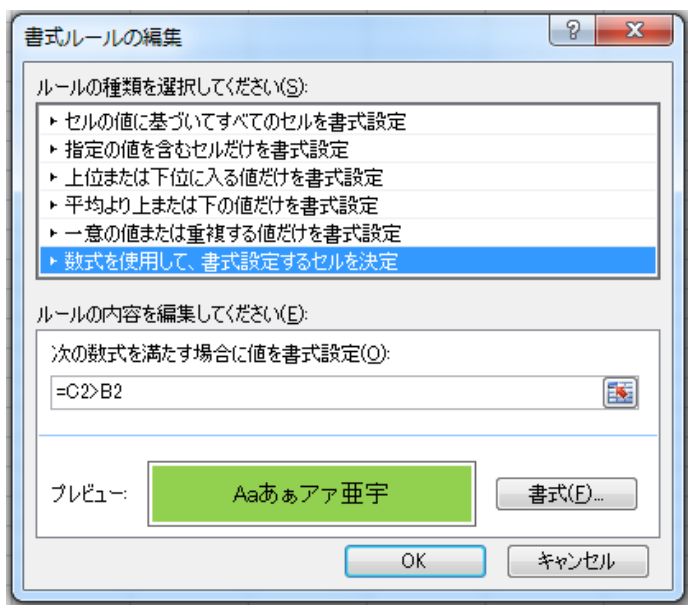
	A	B	C	D	E
1	単語	頻度		単語	頻度
2	appear	225		appear	225
3	belong	259		belong	259
4	cough	48		cough	48
5	cry	154		cry	154
6	fall	325		fall	325
7	fry	87		fry	87
8	laugh	87		laugh	87
9	live	517		live	517
10	open	425		open	425
11	smile	98		smile	98

他のセルの値を参照して条件を作る

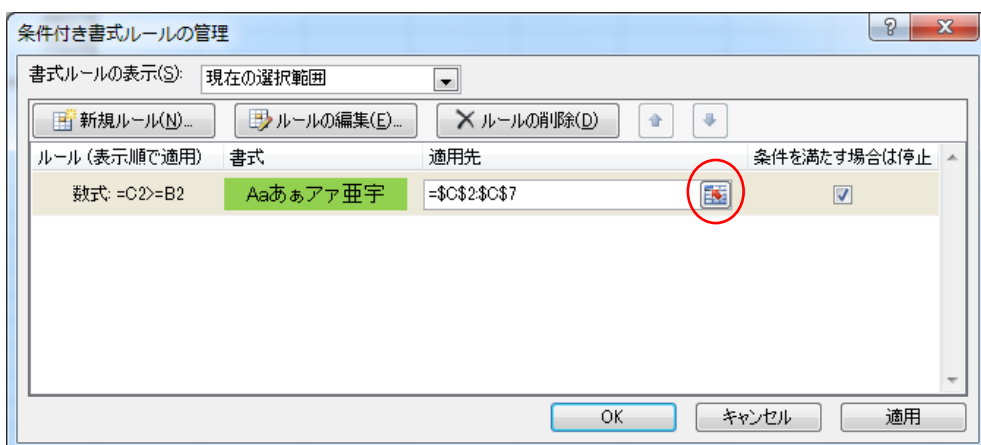
条件付き書式は、他のセルの値を参照してルールを作ることができます。たとえば、次のデータでコーパス B のほうがコーパス A よりも頻度が高い場合、セルを塗りつぶす方法を見てみましょう。

	A	B	C	D
1		コーパスA	コーパスB	
2	する	6724	6661	
3	いく	9780	6954	
4	くる	1226	1916	
5	なる	1242	4627	
6	みる	7855	7529	
7	とる	3269	5704	

◆セル C2 を選択し、「ホーム(H)」→「条件付き書式(L)」→「新しいルール(N)」。「数式を使用して、書式を設定するセルを決定」を選択し、「次の数式を満たす場合に値を書式設定(O)」に=C2>B2 と入力します。その後、「書式(F)」から「塗りつぶし」を設定します。



次に、適応範囲を拡大します。C2 に条件を設定しましたが、これを C2 から C7 に変更します。先ほどの数式は相対参照でしたので、拡大後は条件が相対参照で適応されます。範囲を変更するには、「ホーム(H)」→「条件付き書式(L)」→「ルールの管理(R)」から、先ほどのルールの「適用先」で直接範囲を入力するか、マウスで選択します（マウスを使用した場合、自動的に絶対参照になります）。



「OK」ボタンを押すと、次のようになります。

H35				
	A	B	C	D
1		コーパスA	コーパスB	
2	する	6724	6661	
3	いく	9780	6954	
4	くる	1226	1916	
5	なる	1242	4627	
6	みる	7855	7529	
7	とる	3269	5704	

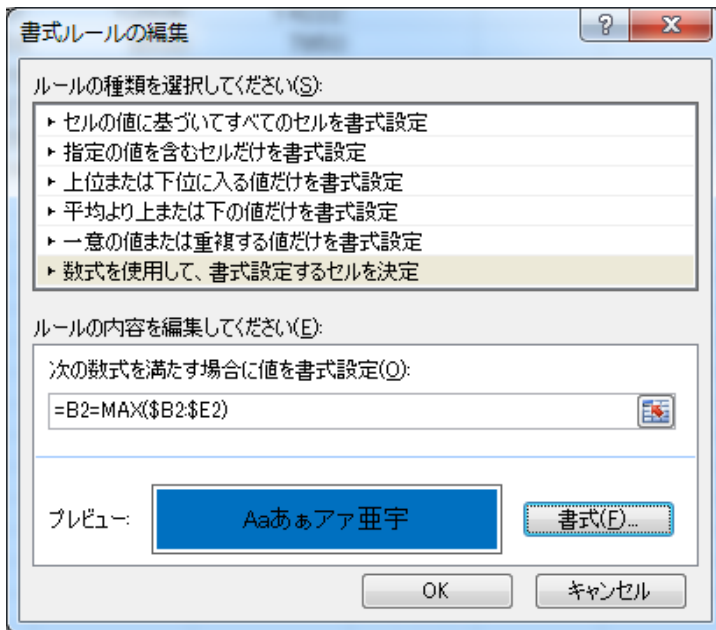
【例】次のデータでそれぞれの行の中で最大の値をもつセルに色を塗ってみましょう。

	A	B	C	D	E
1		コーパスA	コーパスB	コーパスC	コーパスD
2	する	6724	6661	12505	14222
3	いく	9780	6954	8500	7950
4	くる	1226	1916	2522	1850
5	なる	1242	4627	2028	4327
6	みる	7855	7529	6505	5584
7	とる	3269	5704	4245	4270

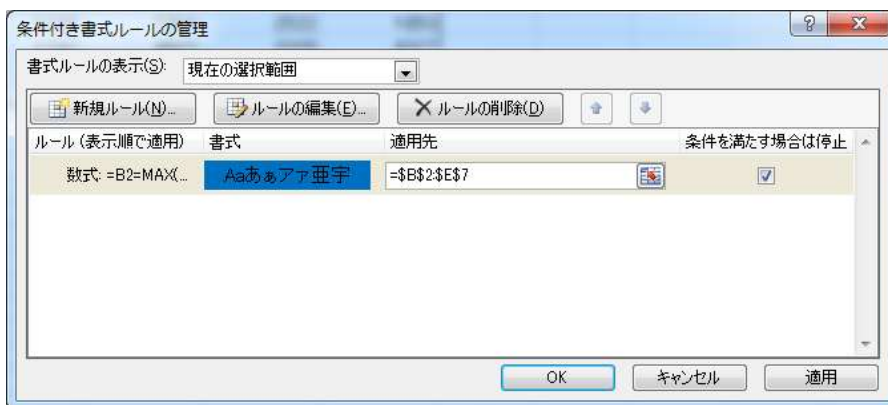
まず、B2にルールを作成します。最大値を求める関数はMAXです。行の中の最大値であれば色をつければよいこととなりますので、ルールとして用いる数式は次のようになります。

=B2=MAX(\$B2:\$E2)

このとき、列方法は絶対参照で固定して下さい。そうしないと、たとえばC2にルールをコピーしたとき、範囲が1つずつ右にずれてしまいます。



適応範囲は「ホーム(H)」→「条件付き書式(L)」→「ルールの管理(R)」から表全体を指定します。










結果は次のようになりました。

	A	B	C	D	E
1		コーパスA	コーパスB	コーパスC	コーパスD
2	する	6724	6661	12505	14222
3	いく	9780	6954	8500	7950
4	くる	1226	1916	2522	1850
5	なる	1242	4627	2028	4327
6	みる	7855	7529	6505	5584
7	とる	3269	5704	4245	4270
8					

[8] グラフ

ある特定項目内でのデータの広がり、あるいはある項目と別の項目の相関を眺めて分析に結びつけるために、グラフは有効な手段です。グラフには種類がたくさんあります。目的に応じて適切なグラフを選

びましょう。たとえば、棒グラフはある言語テキストに単語が何語含まれており、無関係な別の単語が何語含まれていたかを単純にみる場合には適切ですが、ある単語の段落ごとでの増減をみるとときには適切とは言えません。対照的に、折れ線グラフはある単語の段落ごとでの推移を眺めるときには有効な手段ですが、全く無関係なテキストに単語がそれぞれ含まれているかを示すのには不向きと言えます。このように、そのデータの性質と分析に見合ったグラフを選択することが言語分析にとって重要になります。実際の場面ではどのようにグラフを使えばよいか迷うこともありますが、それぞれのグラフについて、基本的に次のように考えるとよいでしょう。

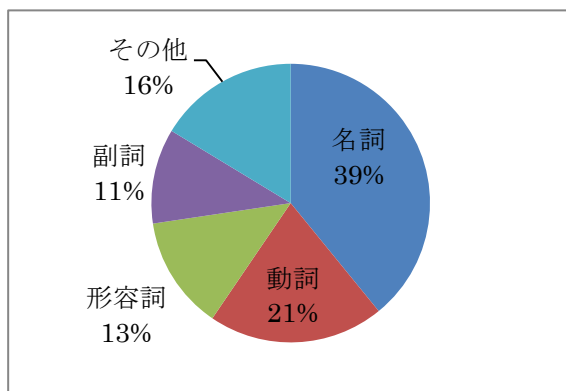
	棒グラフ	それぞれのケース分離して量を比較します。
	折れ線グラフ	連続する変化を観察します。
	円グラフ	1つの現象についてその内訳・割合を観察します。
	レーダー (チャート)	1つの現象について、その属性のバランスがよいか、または一部に偏っているかを観察します。
	面グラフ	量的な変化を連続して観察します。(棒グラフ + 折れ線グラフのような性質)
	散布図	2つの現象(軸)の関係を点として観察します。
	3-D	2つの現象(軸)の関係を量として観察します。

グラフは対象となるデータを選択し、「挿入」→「グラフ」から簡単に作れます。

【例 1】 次の品詞の頻度を割合として円グラフで示してみましょう。

	A	B
1	品詞	頻度
2	名詞	12507
3	動詞	6520
4	形容詞	4202
5	副詞	3525
6	その他	5221
7		

◆A1:B6 を選択します。「挿入(L)」→「グラフ」→「円グラフ(Q)」を選択します。これでグラフは出来上がりますが、割合を示すには、グラフを選択した状態で「グラフツール」→「デザイン」→「グラフのレイアウト」から「%」を含むものを選択します。

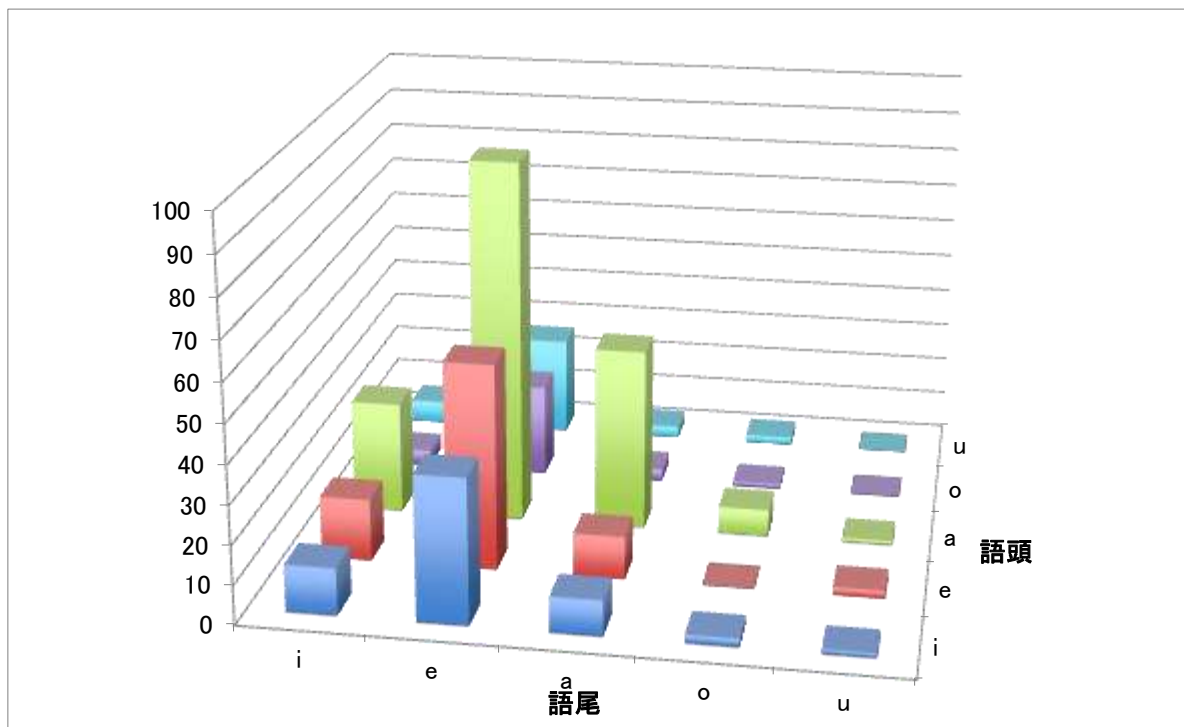


【例 2】 次のデータは聖書の英語訳 (The King James Version) の全単語を対象に母音で始まり母音で終わる単語の分布を表したものです。

		語尾					総計
		i	e	a	o	u	
語頭	i	12	37	9	2	1	61
	e	16	53	11	0	2	82
	a	29	94	47	7	1	178
	o	4	24	3	1	0	32
	u	6	26	3	2	0	37
	総計	67	234	73	12	4	390

このデータを 3D グラフで視覚化してみましょう。

◆次のようにデータの範囲を選択→「挿入」→「グラフ」→「縦棒」→「3-D 縦棒」。「語頭」と「語尾」は「レイアウト(JA)」→「軸ラベル(I)」で指定します。

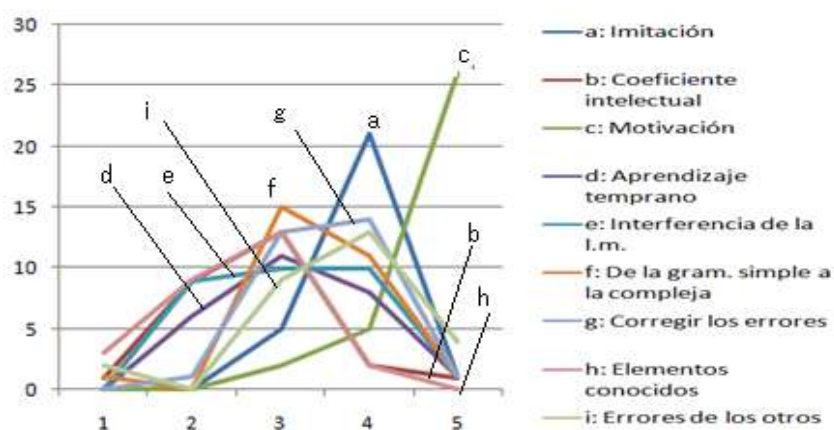


このグラフから、aで始まり eで終わる単語が多いことはとても多いのに対し、uで終わる単語は全体的に少ないということを瞬時に読み取ることができます。

頻度の観察

次は「外国語学習で重要だと思うこと」(cf. Santos, 1999:23)についての意見を、28名の3/4年生に尋ねた簡単なアンケート調査の結果です。「どのような状況・条件が外国語学習に重要か」という質問について、1(まったく重要でない)から5(非常に重要)までを答えてもらいました。3は「どちらでもない」という回答です。折れ線は、(a)「繰り返し練習」、(b)「知能指数」、(c)「モチベーション」、(d)「早期学習」、(e)「母語の干渉の除去」、(f)「段階的な文法学習」、(g)「間違いの訂正」、(h)「既習の内容に沿う」、(i)「他の人の間違いを参照

する」についてのそれぞれの回答の番号の頻度を示します。



このような調査では、ふつう質問項目(a-i)の数値を合計し棒グラフで比較して示しますが、ここでは、1-5の分布の仕方に興味があったため、むしろ折れ線グラフを使ってみました。このグラフを見ると、(c)モチベーションだけが特別な分布をして5に集中しています。他は、おおむね中央の値が最大になっていることがわかります。

Santos Gargallo, I. (2004). *Lingüística aplicada a la enseñanza-aprendizaje del español como lengua extranjera*. Madrid: Arco /Libros.

◇29 フィルタと並べ替え

[9] フィルタ

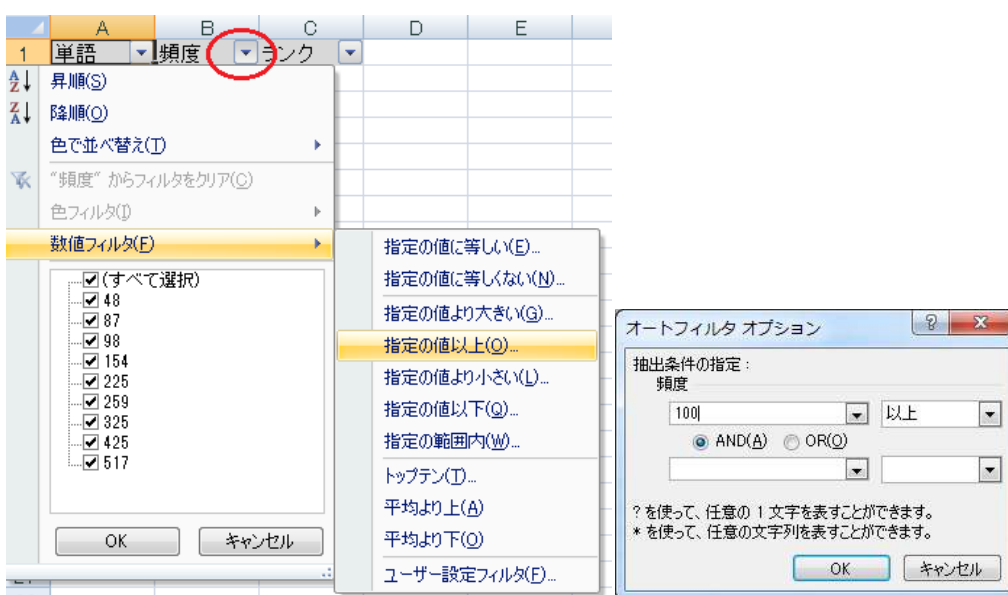
フィルタとは文字や数字を基準に目的の値を含むデータを取り出す機能です。次のデータから頻度 100 以上のものだけを抜き出してみましよう。

	A	B	C
1	単語	頻度	ランク
2	appear	225	**
3	belong	259	**
4	cough	48	
5	cry	154	*
6	fall	325	***
7	fry	87	
8	laugh	87	
9	live	517	***
10	open	425	***
11	smile	98	

A1 を選択した状態で「データ」→「フィルタ」を選択します。



これで表全体にフィルタが設定されました。「頻度」のフィルタを選択し、「数値フィルタ(F)」をクリックし、「指定の値以上(O)」を選びます。



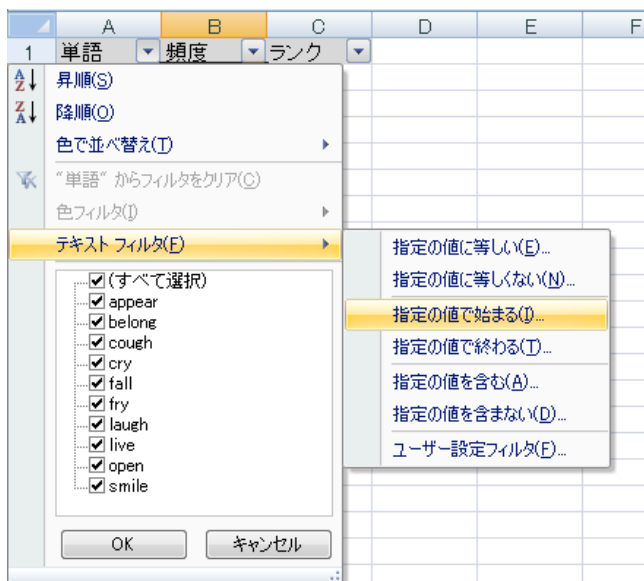
ここで「100 以上」とすると、次のように頻度が 100 以上のものだけを抽出することができます。

	A	B	C	D
1	単語	頻度	ランク	
2	appear	225	**	
3	belong	259	**	
5	cry	154	*	
6	fall	325	***	
9	live	517	***	
10	open	425	***	

また、各列のデータを昇順や降順で並べ替えたり、チェックボックスのオン/オフを切り替えることで各データを選択したりすることができます。フィルタを外すときは、「データ(A)」→「フィルタ(T)」をクリックします。もう一度クリックするとフィルタが設定されます。

<TIPS> 文字列データを含む列には「テキストフィルタ」を設定する

ことができます。この機能を用いると、たとえば「ghで終わる単語」や「lを含む単語」などの指定ができます。



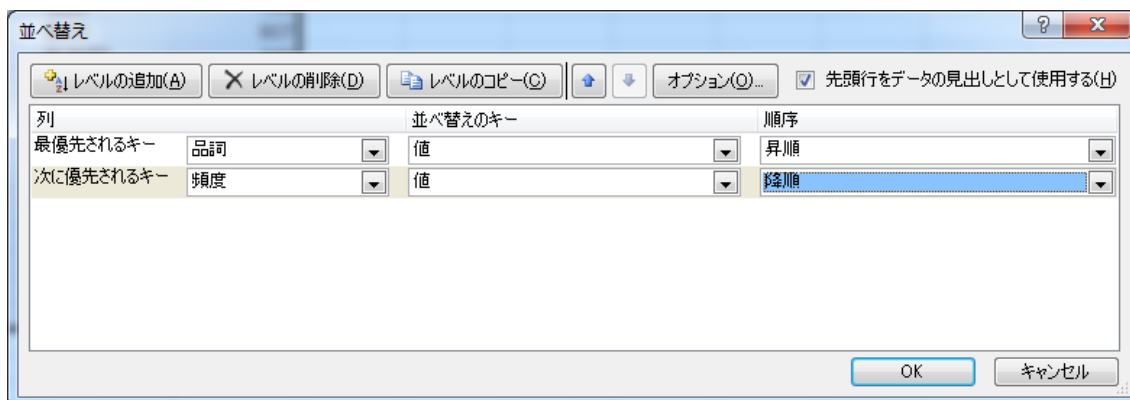
[10] 並べ替え

フィルタを使うことによってそれぞれの列のデータにしたがって、全体のデータを並べ替えることができます。このときの並べ替えの「キー」は選択した列一つになります。

◆「データ(A)」→「並べ替えとフィルタ」→「並べ替え(SA)」を使うと複数のキーを設定することができます。たとえば、次のデータで最優先されるキーを「品詞」とし、次に優先されるキーを「頻度」としてみましょう。

	A	B	C
1	単語	品詞	頻度
2	big	形容詞	654
3	do	動詞	1354
4	fact	名詞	726
5	get	動詞	807
6	good	形容詞	365
7	have	動詞	1297
8	large	形容詞	348
9	little	形容詞	365
10	make	動詞	1339
11	number	名詞	255
12	really	副詞	787
13	small	形容詞	420
14	take	動詞	1055
15	thing	名詞	982
16	very	副詞	435

キーを追加するときは「レベルの追加」ボタンを押します。ここでは「頻度」の「順序」を降順にします。なお、「先頭行をデータの見出しとして使用する」にチェックを入れると、先頭行がキー名として表示されますが、このチェックを外すと、キー名は列 A、列 B などになり、1 行目も含めて並び替えられます。



これで OK ボタンを押すと次のように品詞ごとに頻度順で並び替えることができます。

	A	B	C
1	単語	品詞	頻度
2	big	形容詞	654
3	small	形容詞	420
4	good	形容詞	365
5	little	形容詞	365
6	large	形容詞	348
7	do	動詞	1354
8	make	動詞	1339
9	have	動詞	1297
10	take	動詞	1055
11	get	動詞	807
12	really	副詞	787
13	very	副詞	435
14	thing	名詞	982
15	fact	名詞	726
16	number	名詞	255

◇30 集計

[11] ピボットテーブル

「ピボットテーブル」は、さまざまな方法でデータを集計することができます。この機能を使うと、クロス集計が一瞬でできます。

ここでは次のデータを例に見ていきましょう。

	A	B	C
1	単語	品詞	頻度
2	big	形容詞	654
3	do	動詞	1354
4	fact	名詞	726
5	get	動詞	807
6	good	形容詞	365
7	have	動詞	1297
8	large	形容詞	348
9	little	形容詞	365
10	make	動詞	1339
11	number	名詞	255
12	really	副詞	787
13	small	形容詞	420
14	take	動詞	1055
15	thing	名詞	982
16	very	副詞	435

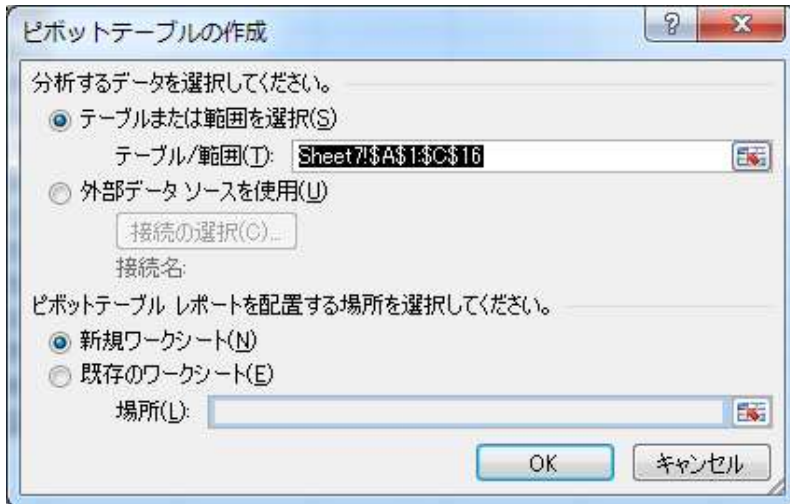
このデータは架空の言語資料でのそれぞれの単語の頻度を示したものです。品詞ごとの頻度の合計を調べたいとき、どのようにすればよいでしょうか。並べ替えて合計したり、関数を使ったり様々な方法がありますが、データが大きくなったとき、これらの方法では太刀打ちできません。そこで登場するのがピボットテーブルです。

- (1) 任意のセルを1つ選び¹⁸、「挿入(V)」→「ピボットテーブル(T)」
→「ピボットテーブル(T)」



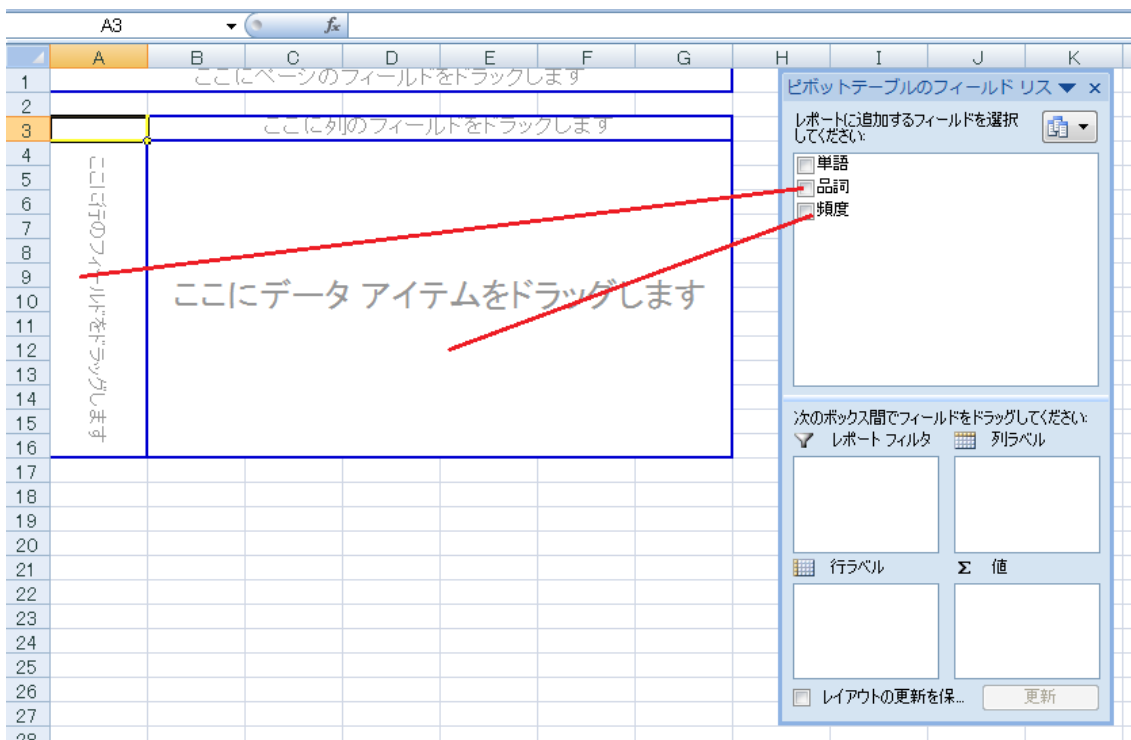
自動的に範囲が選択されます。必要に応じて変更して下さい。

¹⁸ 列や行を選択しないで、セルを選択してください。



「OK」 ボタンを押すと次のような画面が現れます¹⁹。

- (2) 「ピボットテーブルのフィールドのリスト」から、「品詞」を行に、「頻度」をデータアイテム領域にドラッグします。

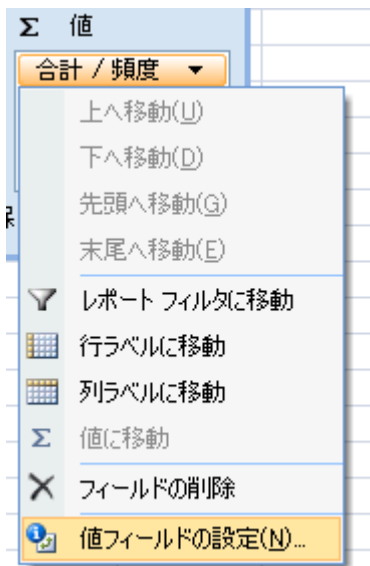


すると、次のように集計の結果が出てきます。

¹⁹ Excel 2010 では「ピボットテーブルツール」→「デザイン(JY)」→「レポートのレイアウト(P)」→「アウトライン形式で表示(O)」とします。

	A	B	C
1	ここにページのフィールドをドラッグします		
2			
3	合計 / 頻度		
4	品詞	集計	
5	形容詞	2152	
6	動詞	5852	
7	副詞	1222	
8	名詞	1963	
9	総計	11189	

データが数字の場合、合計（SUM）がデフォルトに設定されていますが、「ピボットテーブルのフィールドのリスト」の「値」の部分を変更することで、他の統計量を求めることも可能です。次の例は「データの個数」に変更した場合を示しています。



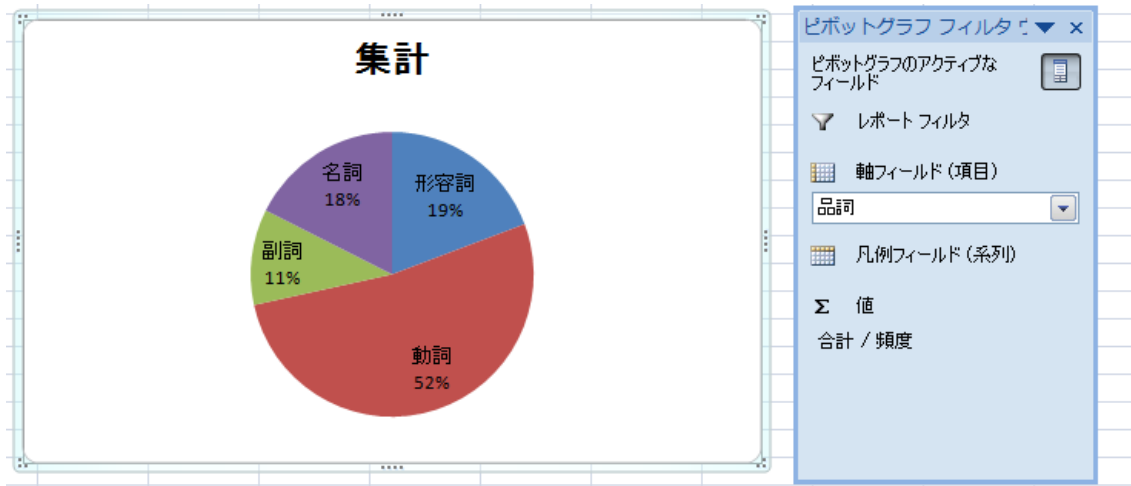
	A	B
1	ここにページのフィールドをドラッグします	
2		
3	データの個数 / 頻度	
4	品詞	集計
5	形容詞	5
6	動詞	5
7	副詞	2
8	名詞	3
9	総計	15

[12] ピボットグラフ

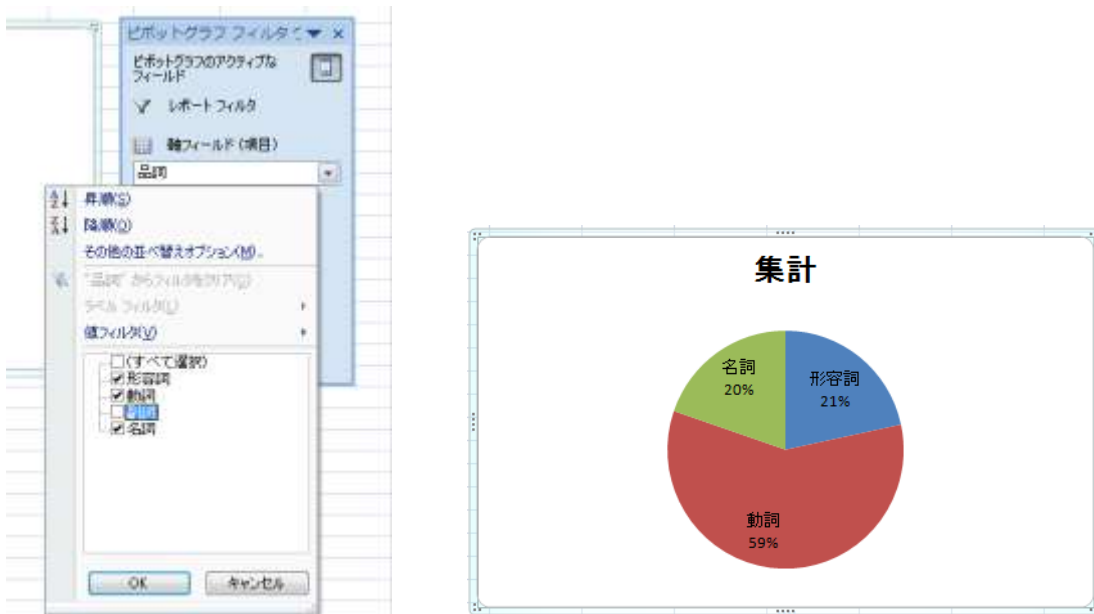
ピボットグラフの機能を使えば、このデータのグラフを簡単に作ることができます。

◆ピボットテーブルを選択した状態で「ピボットテーブルツール」→「オプション(JT)」→「ツール」→「ピボットグラフ(C)」

ここでは「円」を選択します。



ピボットグラフフィルタを使うと、特定の値だけを選んでグラフを作成することもできます。たとえば、「副詞」を外すと以下のようなグラフになります。



【実習編】

【使用データ】 ラテン語版『創世記』第1章

	A	B
1	ID	Text
2	(1:0)	Cap.1
3	(1:1)	In principio creavit Deus caelum et terram.
4	(1:2)	V
5	(1:3)	Dixitque Deus. Fiat lux. Et facta est lux.
6	(1:4)	Et vidit Deus lucem quod esset bona: et divisit lucem a tenebris:
7	(1:5)	appellavitque lucem diem et tenebras noctem. Factumque est vespere et mane dies unus.
8	(1:6)	Dixit quoque Deus. Fiat firmamentum in medio aquarum: et dividat aquas ab aquis.
9	(1:7)	Et fecit Deus firmamentum: divisitque aquas quae erant sub firmamento ab his quae erant super firmamentum. Et factum est ita.

【課題】 『創世記』 第 1 章の単語の頻度を集計する。

【方針】 まず Word を使ってデータを整理し（単語単位で切り出す）、Excel のピボットテーブルを使って集計する。

Word による整形

3 章の実習（課題(3)）を参考に、単語単位で改行した一覧を作りましょう。ワイルドカードを使った置換えを 2 回行います。

置換え(1)

まずはデータを Word にテキスト形式で貼り付けて記号や数字を取り除きます（「貼り付け」→「形式を選択して貼付け(S)」→「テキスト」）。

検索する文字列(N)	[!A-Za-z]
置換え後の文字列(I)	_ ※半角ブランク
検索オプション	ワイルドカードを使用する(U)

< 結果 >

In	principio	creavit	Deus	caelum	terram	Terra
autem	erat	inanis	vacua	tenebrae	erant	super
faciem	abyssi	spiritus	Dei	ferebatur	super	aquas
Dixitque	Deus	Fiat	lux	Et	facta	est lux

置換え(2)

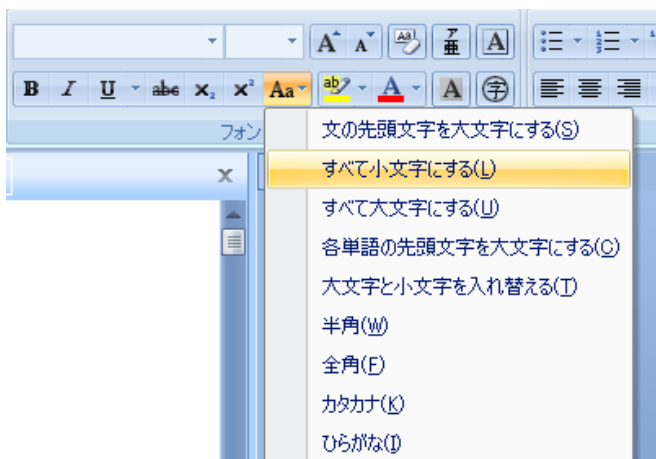
次に 1 個以上のスペース（の連続）を改行に置換えます。

検索する文字列(N)	_ {1,} ※半角空白が先頭
置換後の文字列(I)	^p
検索オプション	ワイルドカードを使用する(U)

< 結果 >

In
 principio
 creavit
 Deus
 caelum
 (...)

大文字と小文字の区別をなくすため、Word の「フォント」から「すべて小文字にする」を選択します。

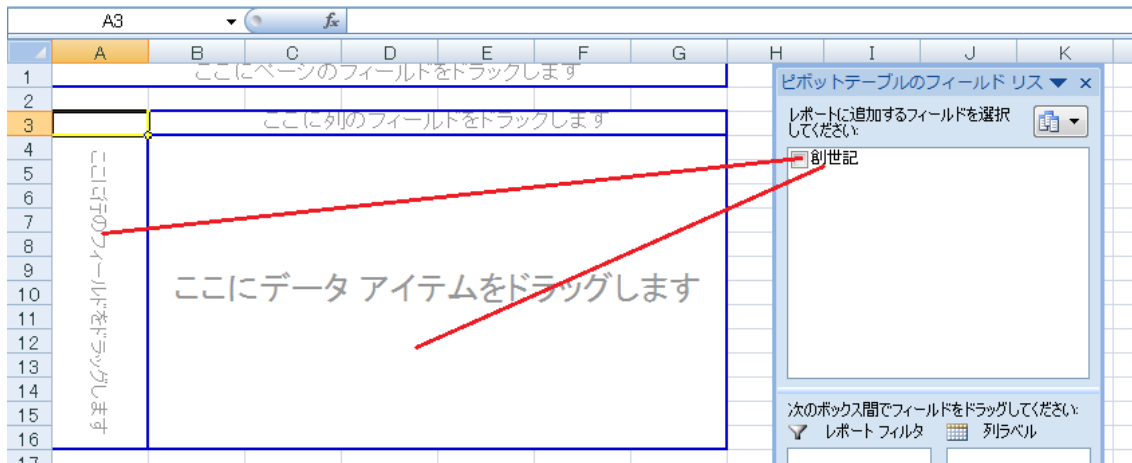


エクセルで集計

Word のデータを Excel に貼り付けます。一行目に「創世記」と入力しておきましょう。

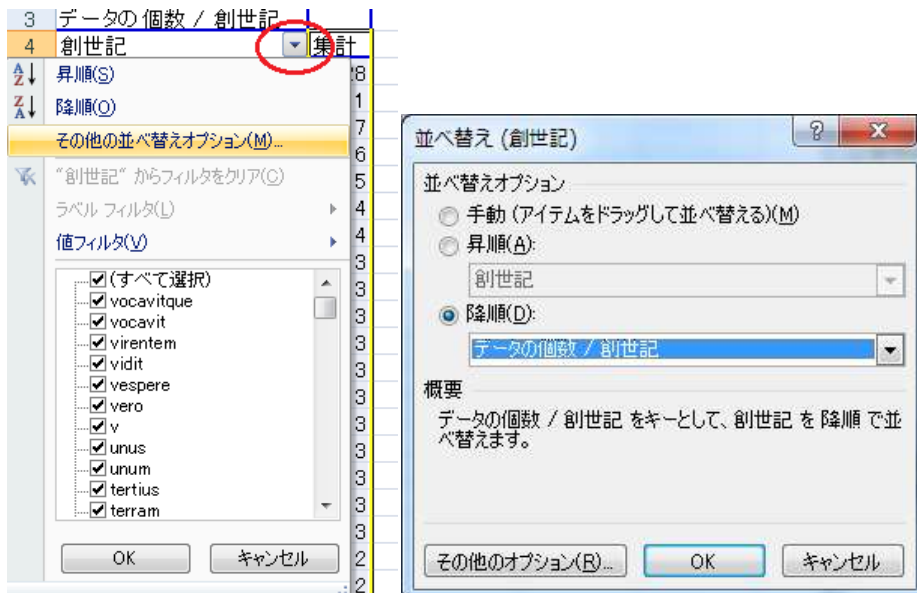
	A	B	C	D
1	創世記			
2	in			
3	principio			
4	creavit			
5	deus			
6	caelum			
7	et			
8	terram			

次に、「挿入」→「ピボットテーブル」をクリックし、範囲が正しいか確認します。次の画面が出たら、「創世記」を「行」と「データアイテム」の両方にドラッグします。



文字データなので、「値」の部分は自動で「データの個数」となります。念のため確認しておきましょう。

仕上げとして、データを頻度順で並べ替えます。ピボットテーブルの▽を選択し、「その他の並べ替えオプション(M)」→「降順」→「データの個数／創世記」を選択してOKを押します。

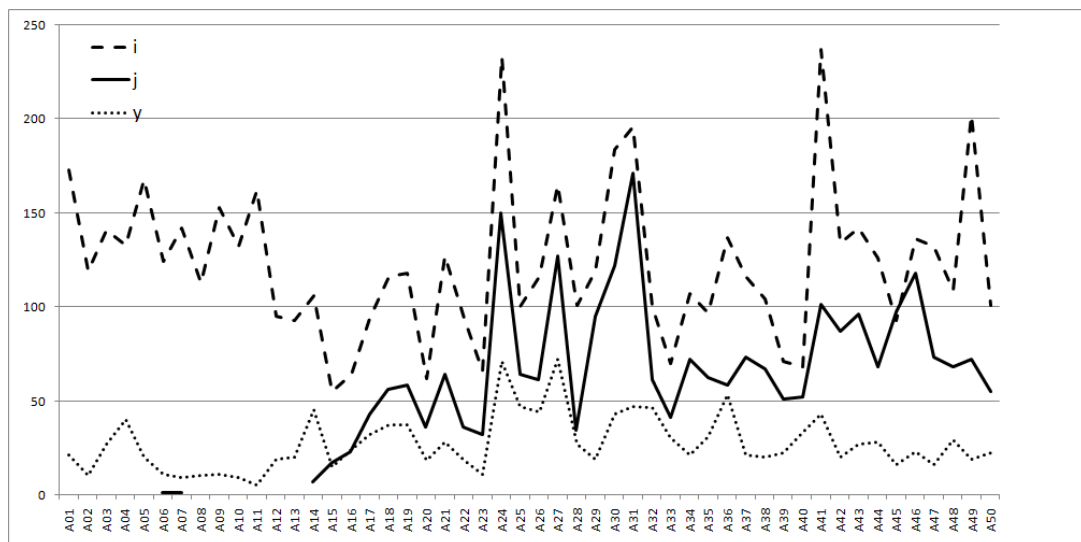


以下のような一覧になれば完成です。頻度順に et, deus, est, in... と続くことがわかります。

	A	B	C	D
1				
2				
3	データの個数 / 創世記			
4	創世記	▼ 集計		
5	et	28		
6	deus	11		
7	est	7		
8	in	6		
9	factum	5		
10	dies	4		
11	firmamentum	4		
12	semen	3		
13	esset	3		
14	vespere	3		

対象の頻度を数える

中世スペイン語に翻訳された『旧約聖書・創世記』の中に使われている <i>, <j>, <y> の文字の頻度を数えてみました。次のグラフは 3 つの文字が各章の中で使われている頻度の分布を示しています。



【図 a】 <i>, <j>, <y> in Genesis, Biblia de Alba

『創世記』全体の 50 章の中で、とくに気になるのは、実線で示した <j> の文字の分布です。最初の 13 章まではほとんど使われていないのに 14 章を境に急に出現しているのです。このことは漫然とテキストを

読んでいたときにはまったく気がつきませんでした。

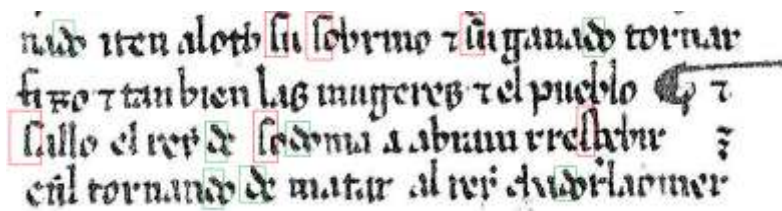
最初の 13 章で<j>という文字がまったく使われなかったのは、そもそも<j>という文字がラテン語にはなくて、後から<i>の文字の下を長くして使いはじめたからです。次の写真を見てください。



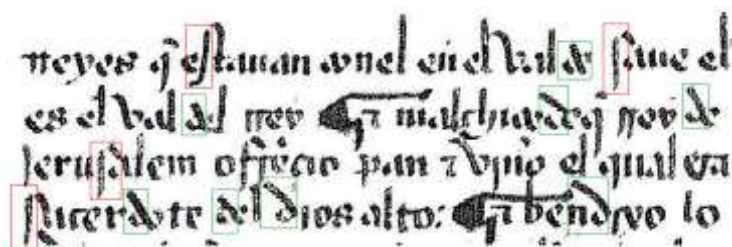
【図 b】 tus ene<mi>gos 【図 c】 firma<mi>ento

これらは 14 章以降に使われた<j>の形です。当時の文字は羽根ペンで上から下に向けてグイグイと縦に書いていたので、<m>(や<n>, <u>)などの文字と<i>が続くとまるで同じ縦の棒が並んだようになって読みにくかったのです。そこで、<m>の後では、<i>ではなくて<j>のように下の部分を長くした文字が使われていました。

それでは、なぜこの『創世記』の 13 章までの部分でそのような工夫がなされなかったのでしょうか。そこには写字生の交替があったと考えられます。14 章の途中を境にして、その前半部と後半部の文字の書き方を比べると、明らかに違いがわかります。



【図 d】 前半の一部



【図 e】 後半の一部

【図 d】と【図 e】を比べると、たとえば枠で示した文字<s>と<d>の形が違います。どうやら<i>の代わりに用いられた<j>の使用は書き手によって異なっていたようです。

次の表は語の末尾で使われている<i>と<y>の文字の分布を示してい

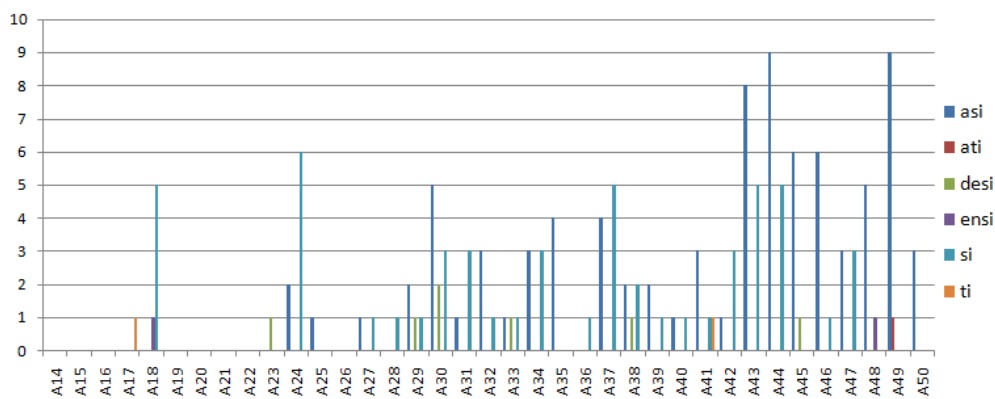
ます。

Cap.	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23	A24	A25	A26	A27	A28	A29	A30	A31	A32	A33	A34	A35	A36	A37	A38	A39	A40	A41	A42	A43	A44	A45	A46	A47	A48	A49	A50		
asi											2	1		1		2	5	1	3	1	3	4		4	2	2	1												
ati																																							1
desi									1							1	2			1					1						1								
ensi				1																																		1	
si			5								6			1	1	1	3	3	1	1	3		1	5	2	1	1	1	3	5	5		1	3					
ti				1																									1										
asy		2	2	2	4	6	2	2	2	2	9		3	5	2	1	3	2	2							1		1	1								2	1	
aty			1	2	1								1	1	1	1								1	1				1										
desy												1	1						1																			1	
ensy																			1																				
sy		1			2	1	1	1		2	3	1	1	1			1	3	1	2				1					2	1								1	
ty	1		1	10		1	2			1	3		1	2	3			5				2		1		1	1	1	1	1	1				4				

【図 f】 語末<i> <y>の分布

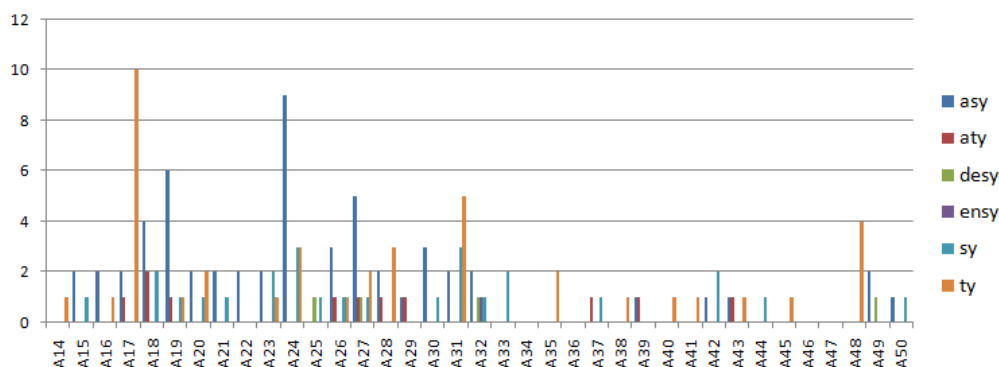
先ほど見たように 14 章の途中を境にして写字生が異なるので、ここでは 14 章以降だけに限って見ることにしましょう。ここまではすべて同じ写字生の文字だと思われます。ところが同じ写字生でも場所によって asi という語が asy と書かれていたり、ati が aty となったりしてバラバラなのです。これはテキストを読んでいて気づいたことなのですが、それでも 2 つの文字の分布の傾向は、実際に測定してみるまでわかりませんでした。

次は<i>の頻度分布です。『創世記』全体の後の方に集中していることがわかります。



【図 g】 語末<i>の分布

次は<y>の頻度分布です。今度は前の方に集中しています。



【図 h】 語末<y>の分布

中世スペイン語の<y>は語末だけでなく、語頭でも<i>の代わりに使われていました。



【図 i】 las <y>magenes

これは語の境界を明示するための工夫だと思われます。この写字生は<y>で語の境界を示そうと意識していたようですが、それがコンスタントではなかったのです。それでもグラフからわかるようにそれぞれ一定の集中が見られます。テキストの中で近い位置にある場合は同じ基準を保つ傾向があったようです。

言語データがデジタル化されると、そのすべてがコンピュータで処理されてしまうので、元のテキストから離れた作業を繰り返すこととなります。しかし、写真、できれば現物などのオリジナルデータを常に参照すべきです。文字化されたデータはコード化され抽象化されているので言語の実態と離れています。次の写真を見てください。



【図 j】 ojos & vjo abenjamjn su hermano fijo

ここで、緑枠で示した<j>と赤枠で示した<j>は大きさが異なります。緑枠の<j>は最初からしっかりと<i>とは異なる文字として書かれているように思えます。一方、赤枠の<j>は先ほど見たように<i>の一種なのですが、隣の文字と混同されないように多少下に向かって長くして

いるような感じですが、これまで両者の違いは気づかれなかったのですが、1993年にスペインのアルカラ大学文学部の学生だったマリアデルカルメン・フェルナンデスの広範な文献学的調査によって、その分布と音声の違いが確認されました。赤枠の<j>は母音[i イ]ですが、緑枠の<j>は[ʒ ジュ]という子音です。

Fernández López, María del Carmen. 1996. “Una distinción fonética inadvertida en el sistema gráfico medieval”, in *Actas del III Congreso Internacional de Historia de la Lengua Española : Salamanca, 22-27 de noviembre de 1993*. edited by A. Alonso González, 112-123.

上田博人. 2011.「スペイン語の変化と変異—音声・文字・文法・語彙」(東京大学言語情報科学専攻編)『言語科学の世界へ—ことばの不思議を体験する45題』東京大学出版会、pp. 133-147.