Excel VBA-L, LETRAS による語彙データ分析

ver. 2014.2.16.

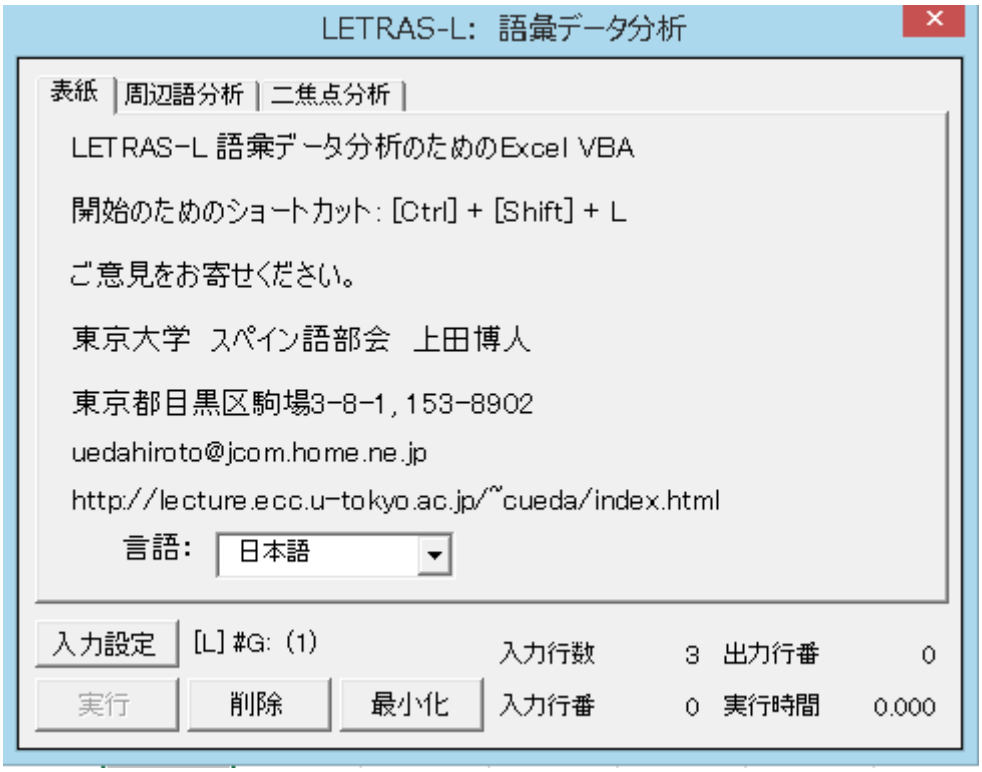
この文書は LETRAS.xlsm (以下では LETRAS とします) を簡単に解説したものです。LETRAS は随時改訂していますので、この文書も予告なしに改訂していきます。常に最終バージョンを次のサイトにアップロードします。ご使用になられた方はぜひご意見をお寄せください。私のメールアドレスは LETRAS の開始ページをご覧ください。参考にさせていただき、よりよいものを目指したいと思います。よろしくお願いいたします。

http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/

(東京大学・上田博人)

1. 開始

LETRAS のファイルを開き、マクロを有効にした後、ショートカット[Ctrl] + [Shift] + L(エル)を押すと LETRAS のマクロが起動します。次は「説明」のタブを開いたところです。



このフォームのタイトルバーに、プログラム名(LETRAS ...)、最小化ボタン（—）、最大化ボタン（□）、終了ボタン（×）があります。分析中にこのフォームが邪魔になったときは最小化ボタンを押してください。再び最大化ボタンを押すと元の大きさに戻ります。「終了」ボタンを押すとフォームが消えます。再度立ち上げるときはショートカット[Ctrl]+[Shift]+Lで起動してください。

【表紙】の下に作成者のメールアドレスが載せられています。プログラムの不具合や改善点などのご意見をお寄せください。なるべく多くの人に回答を差し上げられるようにいたします。

LETRAS を立ち上げていれば、他の Book も分析できます。Book 内で LETRAS のユーザーフォームを表示してください。

シート[L] の列[A]のデータを次のように[A6]=3, [A8]=PT, [A10]=CM に変更し、[A12]にテーマの色を使って塗りつぶしてください。

English	Español	日本語
LETRAS for textual data analysis **ver. 2013.10.5**	**LETRAS para análisis de datos textuales** «	**LETRAS: テキストデータ分析用プログラム集** «
Select laguaguage in the cell [A6]: English=1; Spanish=2; Japanese=3, and restart LETRAS.	**Seleccione el idioma en la celda [A6]: inglés = 1; español = 2; japonés = 3, y reinicie LETRAS.**	**言語を選択してください。英語=1; スペイン語=2; 日本語=3 をセル[A6]に書き込み再度LETRASを起動してください。**
3	«	«
Select decimal separator in the cell [A8]: PT (point) or CM (comma), and restart LETRAS.	**Seleccione el separador decimal en la celda [A8]: PT (punto) o CM (coma), y reinicie LETRAS.**	**小数点を選択してください。(点)=PTまたはCM(コンマ)をセル[A8]に書き込み、再度LETRASを起動してください。**
PT	«	«
Select thousands separator in the cell [A10]: PT (point), CM (comma) or BL (blank), and restart LETRAS.	**Seleccione el separador de miles en la celda [A10]: PT (punto), CM (coma) o BL (blanco), y reinicie LETRAS.**	**千位点を選択してください。PT(点)、CM(コンマ)またはBL(ブランク)をセル[A10]に書き込み、LETRASを再起動してください。**
CM	«	«
Select background color in the cell [A12].	Seleccione el color de fondo en la celda [A12].	背景色を[A12]に指定してください。
Background color Color de fondo **背景色**	«	«

[Ctl]+[Shift]+L で日本語バージョンの LETRAS を起動します。

このフォームは次のような共通のベースの上に、さまざまなタブのついたページが載せられています。

入力設定	現在選択されているシートと列を入力データに設定します。
実行	処理を実行します。
削除	選択されているシートを削除します。複数選択することもできます。開始時のシートを削除しようとすると確認を求められます。
入力行数	入力データの全行数が示されます。
入力行番	実行中に入力データの行番が順次表示されます。
出力行番	実行中に出力データの行番が順次表示されます。
実行時間	実行時間がミリセカンド単位で表示されます。

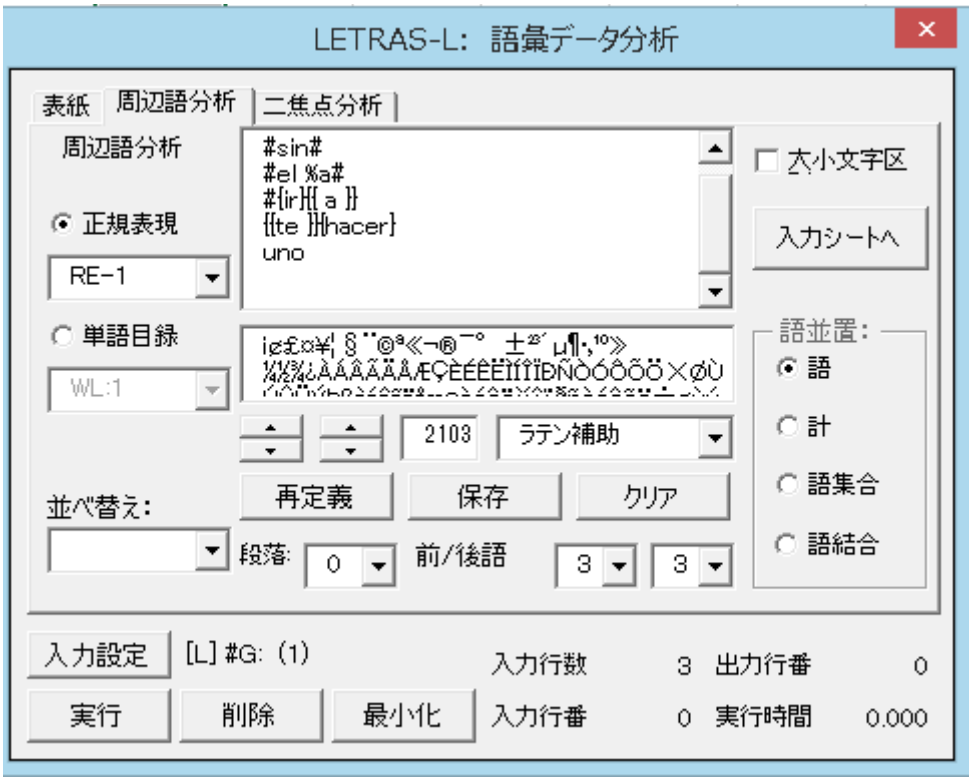
次は LETRAS の Excel シートに載せたサンプルデータ Sample です。

テキスト	見出し:1	見出し:2	行
A la recepción de un hotel madrileño llega un profesor extranjero para participar como conferenciante en un seminario sobre Nutrición organizado por una universidad de verano con sede en El Escorial. El profesor hablará con el conserje, pidiéndole información sobre los servicios del hotel, así como sobre posibles visitas turísticas por la región.	[A] Hotel	(a) Madrid	1
- ¡Buenos días! Desearía una habitación individual para estar tres noches. ¿Qué precio tiene?	[A] Hotel	(a) Madrid	2

サンプルはスペイン語圏各地の会話例です。データは、このように、初めの 1 行をタイトル行とします。A 列がテキスト、B 列以降はそれぞれの行についての付加情報です。

「表紙」以外のタブで実行が可能です。時間がかかる処理を中止するときは、[Esc]キーを押してください。

2. 周辺語分析



「連続」では、焦点と一緒に現れる語をその位置にそって集計したり、集めたり、連続して扱ったりして、その関係を探ります。ここでは単語を単位として、その前後の数語との連続関係を分析します。出力の形式には「語形」「集計」「合同」「結合」がありますが、この中で「語形」だけが文の横のつながりを保持します。一方、「集計」「合同」「結合」では、横のつながりを切って分析します。「並べ替え」は「文字順」と「出現順」が選択できます。

2.1.1. 語形

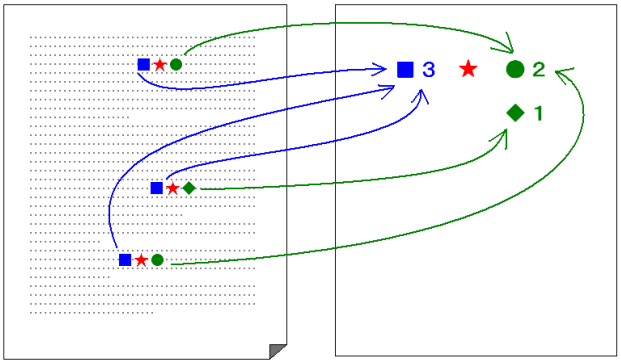
単語を単位としていることを除いて、「内置」とほぼ同じ機能を持ちます。たとえば、上の設定で実行すると次のように出力されます。

por (単語目録)

語-3	語-2	語-1	**焦点**	語+1	語+2	語+3
hacen	al	va	**por**	,	y	en
blanco	al	va	**por**	.	¿	Está
un	taxi	y	**por**	1000	pelas	te
¿	Qué	haces	**por**	acá	?	
qué	la	trae	**por**	acá	?	

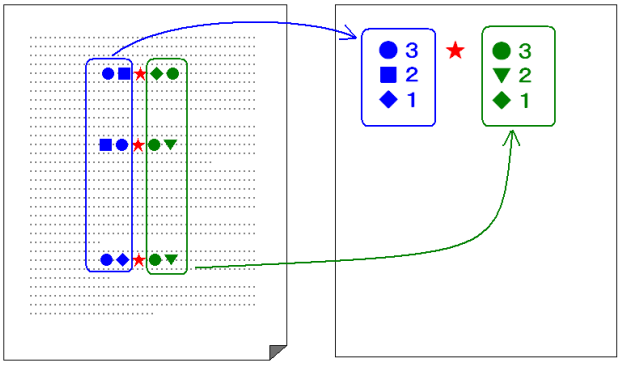
2.1.2. 集計

上の「語形」の前語(W-1, W-2, ...)、鍵語、後語(W+1, W+2, ...)のそれぞれの列の単語をまとめて集計します。まとめているので、横の関係は切れています。焦点のそばにある語の頻度を縦の列だけを区別して調べたいときに役に立ちます。出力は「文字順」「出現順」「頻度順」が選択できます。



2.1.3. 集合

前語の列(W-1, W-2, ...)、後語の列(W+1, W+2, ...)、両者(W-1, W-2, ..., W+1, W+2, ...)をまとめてそれぞれを合計列に出力し、その中の単語を合同して、その頻度を集計します。焦点のそばで連続する語の集合を見るときに使います。



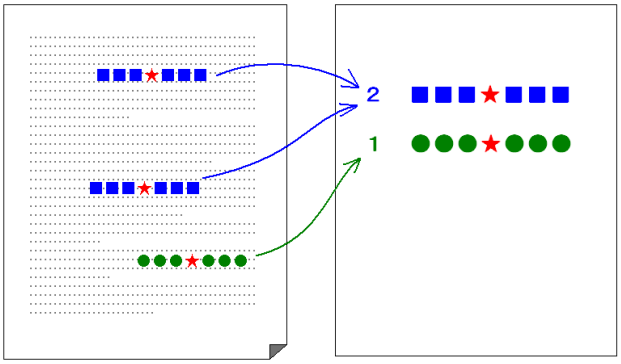
連続：合同のイメージ

por (単語目録)　並べ替え：頻度順

計	-3 語	計	焦点	計	+3 語	計	計: 6 語
103	,	348	**por**	79	,	182	,
34	.			70	que	93	que
28	de			53	.	87	.
23	que			49	la	71	la
22	la			25	favor	42	de
21	no			24		42	
20	–			23	y	42	y
19	y			22	el	34	no

2.1.4. 語結合

前の数語、または後の数語の連続を切らずに、その連続の頻度を集計します。焦点に隣接するまとまった語数の特徴を調べることができます。



por (単語目録)　並べ替え：頻度順

Sum	-3 語	Sum	焦点	Sum	+3 語
6	__–	348	**por**	4	favor_,_dónde
4	si_no_fuera			4	la_cuestión_y
3	tenido_que_venir			4	favor_._
3	para_divertirse_.			3	casualidad_dos_amigos
3	precios_del_trans			3	ejemplo_,_hay
3	de_crédito_,			3	teléfono_a_su

後期ラテン語から初期スペイン語を想像する

大学で習うラテン語はカエサルやキケロなどの紀元前後に書かれた文章を対象とする**古典ラテン語**です。一方、ここでテキスト例として見ている『創世記』のラテン語は紀元 5 世紀の「**後期ラテン語**」Late Latin とよばれるものです。

ラテン語は名詞が格変化し、それによって主語や目的語の関係がわかるので、とくに動詞の目的語の位置が定まっているわけではありませんが、ふつうは動詞の前におきます。そして動詞はふつう文末に置かれます。一方、現代スペイン語などラテン語から派生した言語では目的語は動詞の後に置くのがふつうです。

さて、次は名詞の対格の例として aquam, arcam, terram を選び、その直前の語を頻度順に並べたときの出力です。

SUM	W-1	SUM	aquam arcam terram
42	super	121	terram
38	in	12	aquam
8	omnem	11	arcam
7	et		
6	ad		
4	universam		
3	tibi		
3	dabo		
3	hauriendam		
2	replete		
2	operaretur		
1	aridam		
1	illuminent		
1	deus		
1	ingredieris		
1	elevaverunt		
1	aquae		
1	dissipans		
1	inter		
1	exercere		
1	abram		
1	perambula		
1	te		
1	hydriam		
1	hauriam		
1	hauriret		
1	hausit		
1	reppererunt		
1	invenimus		
1	possideas		
1	isaac.		
1	habebat		
1	sunt		
1	attulit		
1	eis		
1	fames		
1	praeter		

ここで興味深いのは、『創世記』ラテン語の対格（目的語）の位置が、まるで現代スペイン語のように、ほとんど例外なく動詞の直後になっていることです。このように当時の話し言葉が反映していると思われる後期ラテン語の様子から文献によって記録されていない**原始スペイン語** Proto-Spanish のシンタックスを想像することができます。

2.2.【補説】正規表現

2.2.1. 一般の正規表現

正規表現は複雑な文字列処理に適しています。正規表現の規則は非常に単純ですが、使い方については練習が必要です。何度でも実験して確認してください。

特殊文字

¥t	水平タブに一致します。
¥b	任意の英単語の境界に一致します。
¥B	任意の英単語境界以外の位置に一致します。
¥n	改行に一致します。

入力文：

The Universal Declaration of Human Rights Article 1. All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

正規表現:¥bin¥b:単語境界に挟まれた in

The Universal Declaration of Human Rights Article 1. All human beings are born free and equal {*in*} dignity and rights. They are endowed with reason and conscience and should act towards one another {*in*} a spirit of brotherhood.

beings の中の in は、単語境界に挟まれていないので一致しません。

文字クラス

[*xyz*]	文字セットに含まれている任意の 1 文字に一致します。[...]の中では., ? , *などをエスケープする必要はありません。
[^*xyz*]	文字セットに含まれていない任意の 1 文字に一致します。
.（ピリオド）	改行(¥n)以外の任意の文字に一致します。
¥w	英単語に使用される任意の文字（アルファベット、数字、アンダースコア）[a-zA-Z0-9_]に一致します。
¥W	英単語に使用される文字以外の任意の文字に一致します。[^a-zA-Z0-9[a-zA-Z_0-9]と同じです。
¥d	任意の数字に一致します。[0-9]と同じです。
¥D	任意の数字以外の文字に一致します。[^0-9]と同じです。
¥s	任意のスペース文字に一致します。[¥t¥r¥n¥f]と同じです。

¥S	任意の非スペース文字に一致します。[^ ¥t¥r¥n¥f]と同じです。

入力文：

All human beings are born free and equal in dignity and rights.

正規表現検索:[e-h]（e, f, g, h, i に一致）

All {*h*}uman b{*e*}in{*g*}s ar{*e*} born {*f*}r{*e*}{*e*} and {*e*}qual in di{*g*}nity and ri{*g*}{*h*}ts.

正規表現検索:[ˆa-v]（a-v 以外に一致）

All{* *}human{* *}beings{* *}are{* *}born{* *}free{* *}and{* *}equal{* *}in{* *}dignit{*y*}{* *}and{* *}rights{*.*}

正規表現検索:[c-i]（c, d, e, f, g, h, i に一致）

All human beings are born free and equal in dignity and rights.

結果　正規表現検索:[^c-i](c, d, e, f, g, h, i 以外に一致)

All human beings are born free and equal in dignity and rights.

選択、グループ化、繰り返し

|	複数の句を 1 つの正規表現にまとめ、これらのうちの任意の句に一致します。たとえば、d(os|a)は dos または da に一致します。¥ba¥b|¥bthe¥b のように(…)でも使うことができます。
(…)	複数の句をグループ化して 1 つの句を作成します。(ab)*c は abc または c に一致します。
+	1 個以上の直前の文字に一致します。{1,}と同じです。e+で e, ee, eee, …に一致します。
*****	ゼロ個以上の直前の文字またはぐるに一致します。{0,}と同じです。ah*で a, ah, ahh, …に一致します。
?	ゼロ個または 1 個の直前の文字に一致します。{0,1}と同じです。books?で book と books に一致します。
{*a*}	先行する正規表現 *a* 個に一致します。[aeoiu]{2}で 2 母音の連続(ei, ee, ua など)に一致します。
{*a*,}	先行する正規表現 *a* 個以上の直前の文字に一致します。[aeoiu]{3,}で 3 母音の連続(aei, uai, auuu など)に一致します。

{*a*,*b*}	先行する正規表現 *a* 個以上、*b* 個以下に一致します。[aeoiu]{2, 4}で2-4 母音の連続（ei, aei, uai, auuu など）に一致します。

正規表現検索　(free|equal)　（free と equal に一致）

All human beings are born {*free*} and {*equal*} in dignity and rights.

正規表現検索　[e-h]+　（[e-h]の連続に一致）

All {*h*}uman b{*e*}in{*g*}s ar{*e*} born {*f*}r{*ee*} and {*e*}qual in di{*g*}nity and ri{*gh*}ts.

正規表現検索　[aeoiu]{2}（2 母音の連続）

All human b{*ei*}ngs are born fr{*ee*} and eq{*ua*}l in dignity and rights.

結果 3　正規表現検索　[^aeoiu]{2,}　（母音以外の文字 2 個以上の連続に一致）

A{*ll h*}uma{*n b*}ei{*ngs *}are{* b*}o{*rn fr*}ee a{*nd *}equa{*l *}i{*nd*}i{*gn*}i{*ty *}a{*nd r*}i{*ghts.*}

エスケープ文字

特殊文字の検索　　(,),　[,], {, }, ?, !, .(ピリオド), +, *, |, ¥を探すときは，その前に¥をつけてエスケープします。たとえば¥?でクエスチョンマークを検索します。

入力文：

¿Cómo está usted?

正規表現　　¥?（クエスチョンマーク）

¿Cómo está usted?

置換文字

正規表現の後方参照を使うと、検索式の一部を参照することができます。句を括弧で囲み、$の後に 1 つの数字を続けることによってその句を指定します。

$*n*	検索パタンの *n* 番目の(…)に一致した文字列
$$	$という文字

入力文：

Rumi: Hola, profesor.Prof. Rubio: Buenos di/as.Rumi: Buenos di/as. Nos encontramos ahora en la Universidad [[Complutense]] de Madrid. ¿Dónde nos vamos ahora?
Prof.: Bueno, vamos a iniciar hoy el [[Camino]] del [[Cid]], la primera parte.

正規表現：HTML コードを作成します。

a/=>á
e/=>é
i/=>í
ó=>ó
ú=>ú

正規表現:([aeiou])/=>&$1acute;:上の連立式を折りたたみます。

Rumi: Hola, profesor.Prof. Rubio: Buenos días.Rumi: Buenos días. Nos encontramos ahora en la Universidad Complutense de Madrid. ¿Dónde nos vamos ahora? Prof.: Bueno, vamos a iniciar hoy el Camino del Cid, la primera parte.

正規表現：#(c%)=>[[$1]]（c で始まる語を[[...]]でマークします。）

Rumi: Hola, profesor.Prof. Rubio: Buenos días.Rumi: Buenos días. Nos encontramos ahora en la Universidad [[Complutense]] de Madrid. ¿Dónde nos vamos ahora?
Prof.: Bueno, vamos a iniciar hoy el [[Camino]] del [[Cid]], la primera parte.

次は中世スペイン語の資料を文字化した資料です。

Otro(22)ssí mando que los menestrales non echen suerte en el judgado por seer juezes, ca el juez deve tener la seña, e tengo que si <a> afruenta viniesse o a logar de periglo e omne vil o rafez toviesse la seña que podrié (23) caer el concejo en grant onta e en grant vergüença.

(22)は語の途中で改行され ssí 以下が 22 行目になることを示しています。文法研究のためには、これを Otrossí (22)にする必要があります。これは次の置換式によって実現できます。

正規表現：(¥(¥d+¥))(&)=>$2 $1

Otrossí (22) mando que los menestrales non echen suerte en el judgado por seer juezes, ca el juez deve tener la seña, e tengo que si <a> afruenta viniesse o a logar de periglo e omne vil o rafez toviesse la seña que podrié (23) caer el concejo en grant onta e en grant vergüença.

後方参照

後方参照を使うことで式の内容を記憶させ、それを後から参照させることができます。

(…)¥*n*	検索文字列の(…)の式に一致した文字列が記憶され、それを *n* 回繰り返して参照します。
(…)…(…)=>$*n*	検索文字列の(…)の式に一致した文字列が記憶され、置換文字列でそれを参照して再生します。*n* は(…)の順番に対応する番号です。

入力文 ：

どんどんテーマが広がって、ますます興味がわいてきた。

正規表現：(..)¥1:2 文字が 2 回繰り返す文字列

どんどんテーマが広がって、ますます興味がわいてきた。

参照する文字（列）がわかっているときは、検索式をたとえば「(どん){2}」のようにすることができますが、ここでは他にも「ますます」「ぐんぐん」のように、さまざまに変化する場合を想定しています。¥1 が先行する(..)を後方から参照しています。

2.2.2. 拡張正規表現

特殊文字を再定義

LETRAS.xlsm では一般の正規表現を拡張して次の検索字を再定義します。

#	単語の境界：#b%は b で始まる単語を検索します。
&	単語文字 1 個以上 ¥l+
%	単語文字 0 個以上 ¥l*
¥l	西欧語単語文字 [A-Za-zÀ-ǿ]

¥L	西欧語単語文字以外 [^A-Za-zÀ-ǿ]
¥i	キリル文字
¥I	キリル文字以外
¥g	ギリシャ文字
¥G	ギリシャ文字以外
¥e	ハングル
¥E	ハングル以外
¥y	CJK 互換漢字、統合漢字、漢字拡張文字
¥Y	CJK 互換漢字、統合漢字、漢字拡張文字以外
¥v	母音文字 [aeiouÀ-Æà-æÈ-Ëè-ëÌ-Ïì-ïÒ-Öò-öÙ-Üù-ü]
¥V	母音文字以外 [^aeiouÀ-Æà-æÈ-Ëè-ëÌ-Ïì-ïÒ-Öò-öÙ-Üù-ü]
¥c	子音文字 [bcdfghj-np-tv-zÇçÑñß]
¥C	子音文字以外 [^bcdfghj-np-tv-zÇçÑñß]

単語の境界

拡張正規表現：#(m%)=>[$2]（語頭の単語境界#を使った置換式の置換文字列のトークンの数字は+1 としてください。）

Otro(22)ssí [mando] que los [menestrales] non echen suerte en el judgado por seer juezes, ca el juez deve tener la seña, e tengo que si <a> afruenta viniesse o a logar de periglo e omne vil o rafez toviesse la seña que podrié (23) caer el concejo en grant onta e en grant vergüença.

拡張正規表現：(%[rs])#=>[$1]（語尾の単語境界#を使った置換式の置換文字列のトークンの数字はそのままにしてください。）

Otro(22)ssí mando que [los] [menestrales] non echen suerte en el judgado [por] [seer] [juezes], ca el juez deve [tener] la seña, e tengo que si <a> afruenta viniesse o a [logar] de periglo e omne vil o rafez toviesse la seña que podrié (23) [caer] el concejo en grant onta e en grant vergüença.

日本語文字の再定義

日本語文字を検索するときは、¥h（ひらがな）, ¥k（カタカナ）, ¥z（漢字）, ¥j（日本語文字）を使用します。

¥h	ひらがな [ぁ-んー]
¥H	ひらがな以外 [^ぁ-んー]
¥k	カタカナ [ァ-ンー]

¥K	カタカナ以外 [^ァ-ンー]
¥z	漢字 [一-龥々〆]
¥Z	漢字以外 [^一-龥々〆]
¥j	日本語文字 [ぁ-んァ-ンー一-龥々〆]
¥J	日本語文字以外 [^ぁ-んァ-ンー一-龥々〆]

入力文：

```
親譲りの無鉄砲で小供の時から損ばかりしている。
```

拡張正規表現 **¥z{2}**（漢字 2 文字の連続）

```
{*親譲*}りの{*無鉄*}砲で{*小供*}の時から損ばかりしている。
```

拡張正規表現:**¥z+**（1 個以上の漢字）

```
{*親譲*}りの{*無鉄砲*}で{*小供*}の{*時*}から{*損*}ばかりしている。
```

入力文：

```
『坊っちゃん』夏目漱石
　親譲りの無鉄砲で小供の時から損ばかりしている。小学校に居る時分学校の二階から飛び降りて一週間ほど腰を抜かした事がある。なぜそんな無闇をしたと聞く人があるかも知れぬ。別段深い理由でもない。新築の二階から首を出していたら、同級生の一人が冗談に、いくら威張っても、そこから飛び降りる事は出来まい。弱虫やーい。と囃したからである。小使に負ぶさって帰って来た時、おやじが大きな眼をして二階ぐらいから飛び降りて腰を抜かす奴があるかと云ったから、この次は抜かさずに飛んで見せますと答えた。
```

拡張正規表現 **¥z*¥h+**（漢字とひらがな：±漢字＋ひらがな）

```
『{*坊っちゃん*}』夏目漱石
　{*親譲りの*}{*無鉄砲で*}{*小供の*}{*時から*}{*損ばかりしている*}。{*小学校に*}{*居る*}{*時分学校の*}{*二階から*}{*飛び*}{*降りて*}{*一週間ほど*}{*腰を*}{*抜かした*}{*事がある*}。{*なぜそんな*}{*無闇をしたと*}{*聞く*}{*人があるかも*}{*知れぬ*}。{*別段深い*}{*理由でもない*}。{*新築の*}{*二階から*}{*首を*}{*出していたら*}、{*同級生の*}{*一人が*}{*冗談に*}、{*いくら*}{*威張っても*}、{*そこから*}{*飛び*}{*降りる*}{*事は*}{*出来まい*}。{*弱虫や*}ー{*い*}。{*と*}{*囃したからである*}。{*小使に*}{*負ぶさって*}{*帰って*}{*来た*}時、{*おやじが*}{*大きな*}{*眼をして*}{*二階ぐらいから*}{*飛び*}{*降りて*}{*腰を
```

*}{*抜かす*}{*奴があるかと*}{*云ったから*}、{*この*}{*次は*}{*抜かさずに*}{*飛んで*}{*見せますと*}{*答えた*}。

単語境界のある検索式を使用するときは次のような検索置換式に変換します。

#a% → (^|$|¥L)a%=>$1{*a%*}
%r# → %r(^|$|¥L)=>%r¥1

外国語文字の再定義

「再定義」というシートには次のような設定をしています。これは自由に変えることができます。変更したときは「更新」ボタンを押してください。ここで使用している「/, ", ~, `, ^」という特殊記号を検索するときは前に¥をつけてエスケープしてください。

¥/	x128$
a/	á
e/	é
i/	í
ó	ó
ú	ú
A/	Á
E/	É
I/	Í
Ó	Ó
Ú	Ú
¥"	x128$
a"	ä
e"	ë
i"	ï
o"	ö
u"	ü
A"	Ä
E"	Ë
I"	Ï
O"	Ö
U"	Ü

x128$	"
¥~	x128$
a~	ã
e~	e
i~	i
o~	õ
u~	u
A~	Ã
E~	E
I~	I
O~	Õ
U~	U
x128$	~
¥`	x128$
a`	à
e`	è
i`	ì
o`	ò
u`	ù
A`	À
E`	È
I`	Ì
O`	Ò
U`	Ù
x128$	`
¥^	x128$
a^	â
e^	ê
i^	î
o^	ô
u^	û
A^	Â
E^	Ê
I^	Î
O^	Ô

U^	Û
x128$	^

前後の条件

-{{正規表現}} 検索文字列の前後に付加して検索の条件とします。
たとえば、{{te }}va%は te と空白に続くデータを検索します。

2.3. 大小文字区別

「大小文字区別」をチェックして次の置換式を使うと小文字ではじまる語だけにマッチします。

(%e%)=><$1> (大小区別なし)

A la <recepción> <de> un <hotel> <madrileño> <llega> un <profesor> <extranjero> para participar como <conferenciante> <en> un <seminario> <sobre> Nutrición organizado por una <universidad> <de> <verano> con <sede> <en> <El> <Escorial>.

(%e%)=><$1> (大小区別あり)
A la <recepción> <de> un <hotel> <madrileño> <llega> un <profesor> <extranjero> para participar como <conferenciante> <en> un <seminario> <sobre> Nutrición organizado por una <universidad> <de> <verano> con <sede> <en> El Escorial.

(%E%)=><$1> (大小区別あり)
A la recepción de un hotel madrileño llega un profesor extranjero para participar como conferenciante en un seminario sobre Nutrición organizado por una universidad de verano con sede en <El> <Escorial>.

2.4. 単語目録

「単語目録」を選択すると単語を単位にして置換します。次の例では、(1:1)の In は P に置換されますが、(1:2)の inanis の in は置換されません。大量の単語を置換したり検索するときは単純一致や正規表現と比べて処理が高速になります。

a=>A
de=>DE
en=>EN

A la recepción DE un hotel madrileño llega un profesor extranjero para participar como conferenciante EN un seminario sobre Nutrición organizado por una universidad DE verano con sede EN El Escorial.

この検索式は正規表現ではないので、(a|de|en)のようにまとめることができません。正規表現では、#(a|de|en)#=>$1 <$2> $3 とします。

3. 二焦点分析

正規表現で指定する 2 つの要素の結合度を探ります。語形変化が多いラテン語などでは正規表現を工夫することで、さまざまな語の組み合わせを実現できます。語形変化が比較的少ない英語でも、たとえば#ha(ve|s|d|ving)#によって要素の 1 つに have を指定することができます。結合度を示すさまざまな係数を同時に出力します。

3.1. 共起係数

「共起係数」のオプションを選択すつと、2 つの要素は同じセルの中にある、という条件を満たしているかぎり、どれだけ間隔が空いてもかまいません。

二焦点分析：共起係数	全体
頻度 X: #(por|para)#	521
頻度 Y: &[aei]r#	882
a (X:+ / Y:+): #(por|para)#.*?&[aei]r#	395
b (X:+ / Y:-): #(por|para)#	126
c (X:- / Y:+): &[aei]r#	487
Total (N)	26,581
相互情報量	4.5140
Dice 2a/(2a+b+c)	0.5631
Jaccard a/(a+b+c)	0.3919
Ochiai a/√[(a+b)(a+c)]	0.5827
Ueda (2a-b-c)/(2a+b+c)	0.1262

3.2. 間隔係数

「間隔係数」は 2 つの要素の間に指定した数字以下の文字数のデータがあるときの回数を計算し、それを共起回数とします。

二焦点分析：間隔係数・文字数100	全体
頻度 X: #(por|para)#	521
頻度 Y: &[aei]r#	882
a (X:+ / Y:+): #(por|para)#.{0,100}?&[aei]r#	387
b (X:+ / Y:-): #(por|para)#	134
c (X:- / Y:+): &[aei]r#	495
Total (N)	26,581
相互情報量	4.4845
Dice 2a/(2a+b+c)	0.5517
Jaccard a/(a+b+c)	0.3809
Ochiai a/√[(a+b)(a+c)]	0.5709
Ueda (2a-b-c)/(2a+b+c)	0.1033

3 つの係数を比較すると、接係数よりも共起係数のほうが「共起回数」が多く、間隔係数は、両者の間になることがわかります。間隔語数を増やすと、共起回数が増加する可能性が高くなります。

隣接係数 <= 間隔係数 <= 共起係数

3.3. 順番を区別

ラテン語は語順が比較的自由な言語です。「順番区別」のチェックの有無による出力を比較しましょう。次は「文」の中で順番を区別したときと、

区別しないときの出力を比べたものです。

二要素共起分析:共起係数	順番区別	区別なし
X: #domin(us|i|um)# の度数	164	164
Y: #de(us|i|um)# の度数	200	200
a (++) 共起回数(XY): #domin(us|i|um)#+#de(us|i|um)#	29	33
b (+-) 排他回数(X): #domin(us|i|um)#	135	131
c (-+) 排他回数(Y): #de(us|i|um)#	171	167
全度数 (N)	25,569	25,569
相互情報量(MI)	4.4987	4.6851
Dice 係数 2a/(2a+b+c)	0.1593	0.1813
Jaccard 係数 a/(a+b+c)	0.0866	0.0997
Cosine 係数 a/√[(a+b)(a+c)]	0.1601	0.1822
相対優先係数 [a/(a+b) + a/(a+c)])/2	0.1609	0.1831