

電子コーパスを使ったフランス語研究 今だからできることー

1 . 英語研究と電子コーパス :

<http://www.lang.nagoya-u.ac.jp/~takizawa/KokaiCorpusIntro.html>

<http://www.lang.nagoya-u.ac.jp/~takizawa/KokaiCorpusTakizawa.pdf>

- 始まり (1960 年代)
 - Brown Corpus (アメリカ英語)
 - LOB Corpus (イギリス英語)
 - 1961 年に出版された 15 のジャンルを代表する各 2000 語の 500 のサンプルからなる約 100 万語のコーパス
- 大規模なコーパス (1990 年代)
 - The bank of English: 4 億 8 千万語
 - The BNC World Edition: 9000 万語の書き言葉と 1000 万語の話し言葉 (2001 年)
- コーパスを活用した文法書・辞書
 - 1998 年: コーパス言語学関係の書物の大量出版 (英語)
 - 1999 年: 本格的な文法書
- コーパスは記述研究むけ
 - 言語学における経験主義と演繹主義, 実例と作例, 観察と直観
 - コロケーション
- コンピュータを使った研究: データの**大量・高速・正確**な処理

2 . フランス語の問題 (フランス語と日本語の文字コード)

日本語環境の中でフランス語を使うということにまつわるさまざまな問題がある.

情報が少ない. うまくいかない理由がわからない. OS によって動き方が違うなどの個別的な問題がある. 英語で可能であっても, フランス語では不可能なのが普通. さまざまな試行錯誤, 工夫が必要
例えば: Frantext を使っても, コピーができなければ役に立たない.

フランス語と UNIX:

http://www.lang.nagoya-u.ac.jp/~fujimura/memo/french_unix.html

3 . フランス語研究とコーパス

Benoît HABERT, Adeline NAZARENKO, André SALEM : Les linguistiques de corpus, Armand Colin (1997)

3.1.本来型のコーパス

A. FRANTEXT:

Centre National de la Recherche Scientifique(CNRS) に属する Institut National de la Langue Francaise (INaLF) が、1957 年から "Tresor de la Langue Francaise" の編纂のために作成したテキストデータベース . 中世から 20 世紀までのおよそ 2000 タイトルを含む。大学などの研究機関が利用契約を結ぶ。

FRANTEXT peut se définir comme un vaste corpus, à dominante littéraire, constitué de textes français qui s'échelonnent du XVIe au XXe siècle. Sur l'intégralité du corpus, il est possible d'effectuer des recherches simples ou complexes (base non-catégorisée). Sur un sous-ensemble comportant des oeuvres en prose des XIXe et XXe siècles, les recherches peuvent en outre répondre à des critères syntaxiques (base catégorisée).

テキストの種類に偏りがある： 「立派な」書き言葉、8割は文学作品
書簡集，弁論，隨筆，論文集，回想録，風刺文，詩，紀行文，散文，小説，戯曲，概論，韻文：

総語数: 187429462

ex. 1950 年以降のテキスト：475 点

Nombre d'occurrences dans le corpus : 30961078.

cf: Cobuild Direct

<http://www.cobuild.collins.co.uk/cdguide/svenguide.html>

Type the names of any number of subcorpora, taken from the following list:

frbks 3033197 04 Miscellaneous books

frmcc 3088622 03 Marie Claire magazine
 fract 1975508 01 Actuel magazine
 lm00 26068585 08 Le Monde 2000
 lm97 5672024 05 Le Monde 1997
 lm01 7263867 09 Le Monde 2001
 lm98 8227023 06 Le Monde 1998
 lm99 21158225 07 Le Monde 1999
 frlib 1487725 02 Liberation magazine

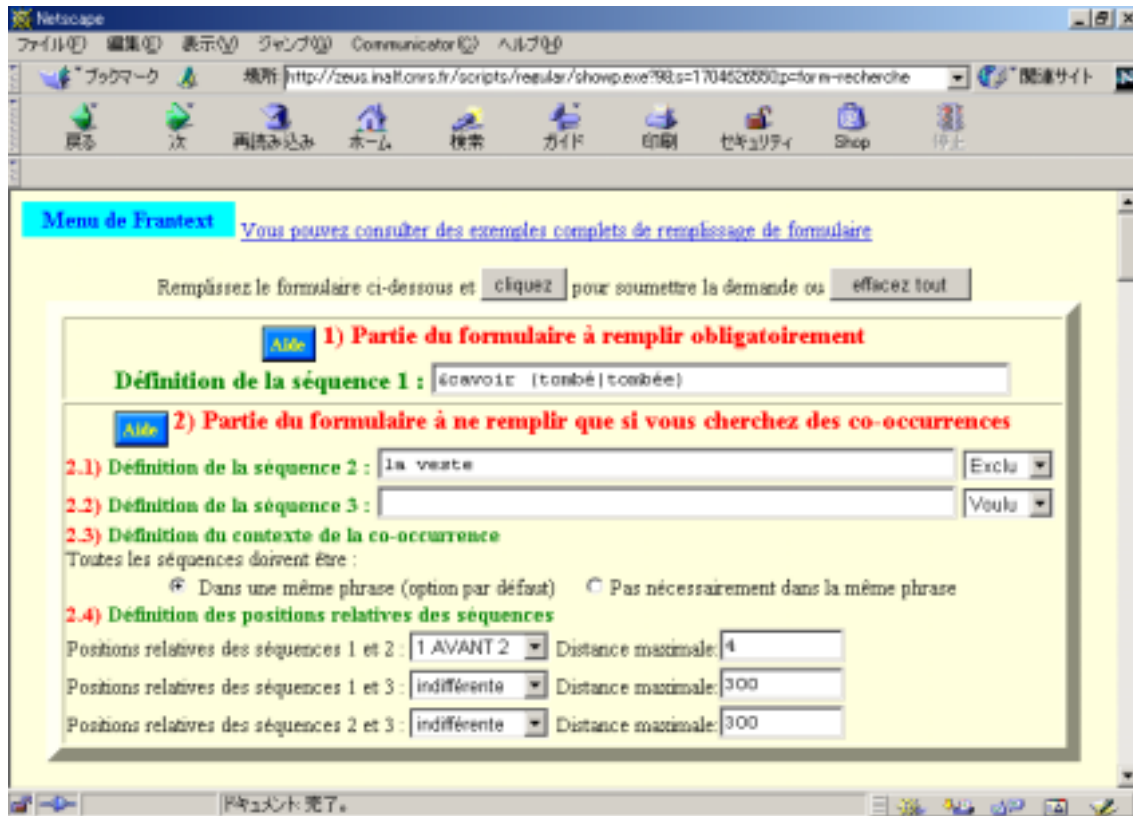
<http://www.collinswordbanks.com/>

Full title	Size (million words)
Australian news	5.3
UK ephemera	3.1
UK magazines	4.9
UK spoken	9.3
US ephemera	1.2
BBC World Service	2.6
National Public Radio	3.1
UK books	5.3
US books	5.6
Times newspaper	5.7
Today newspaper	5.2

1)est tombé : 1331, a tombé :33

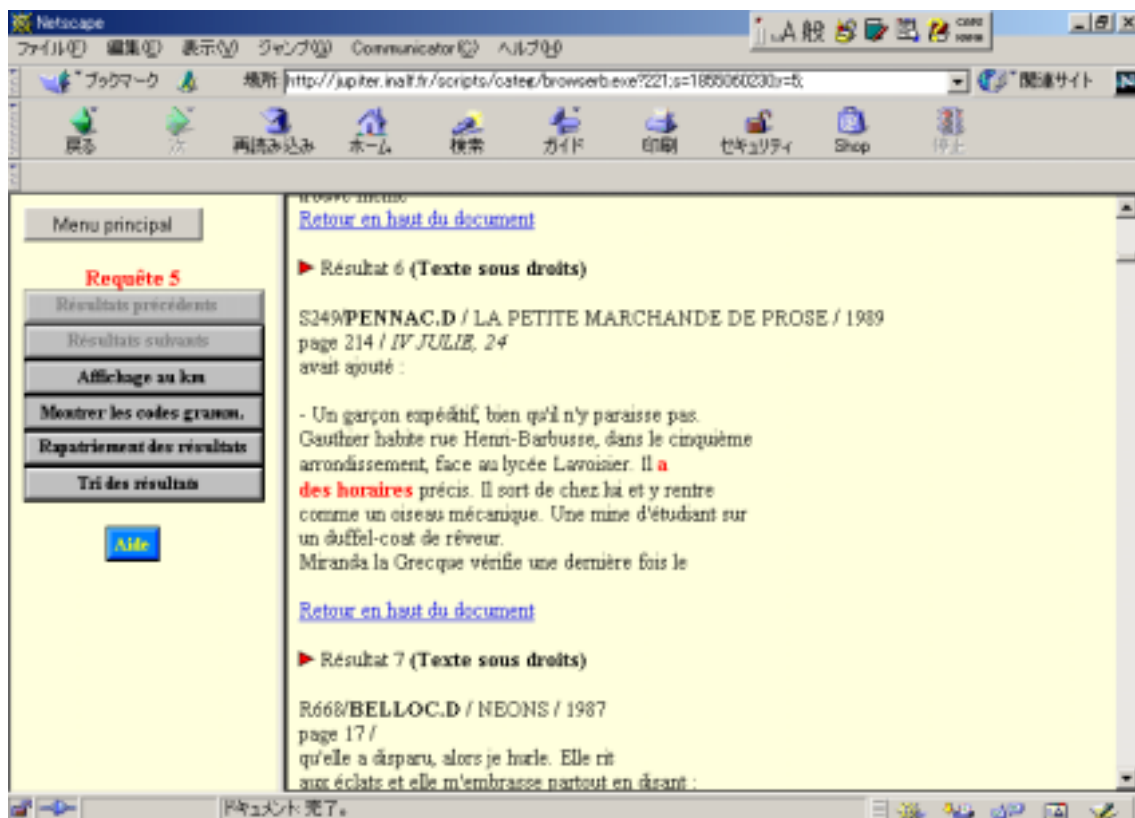
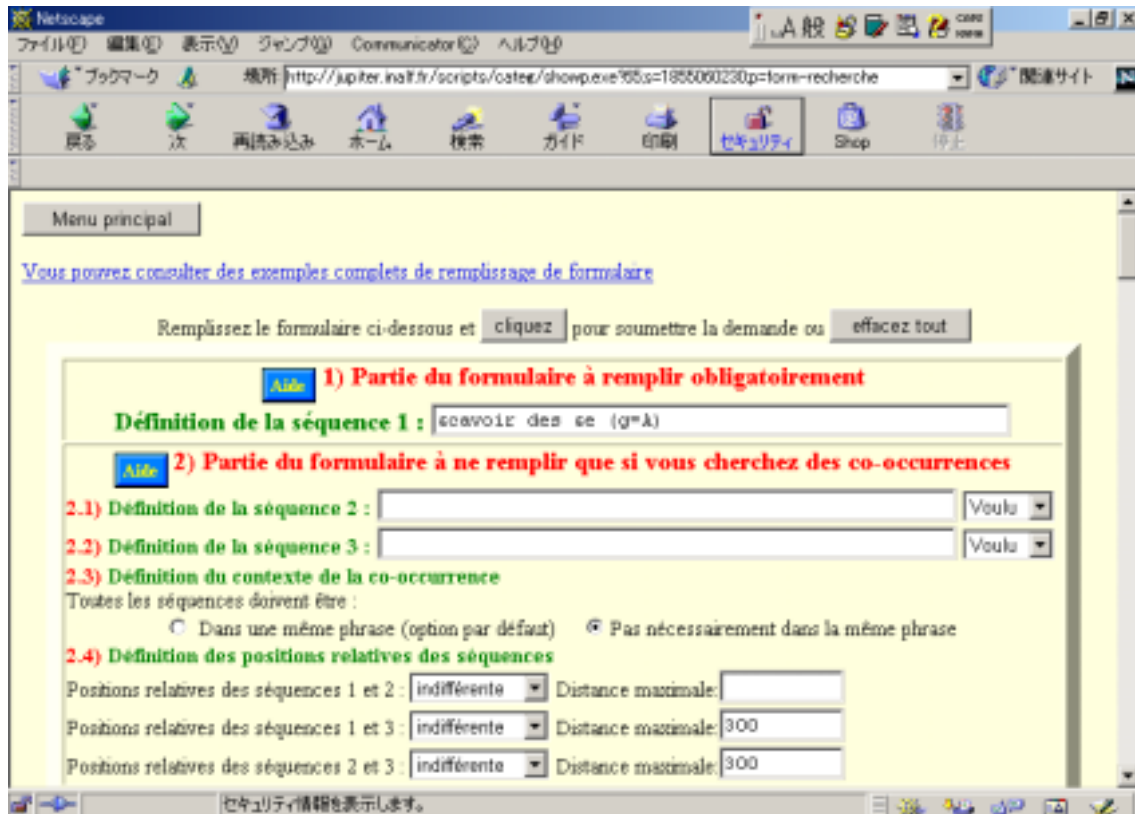
<http://lang.nagoya-u.ac.jp/~fujimura/shuchu/atombe.crp>

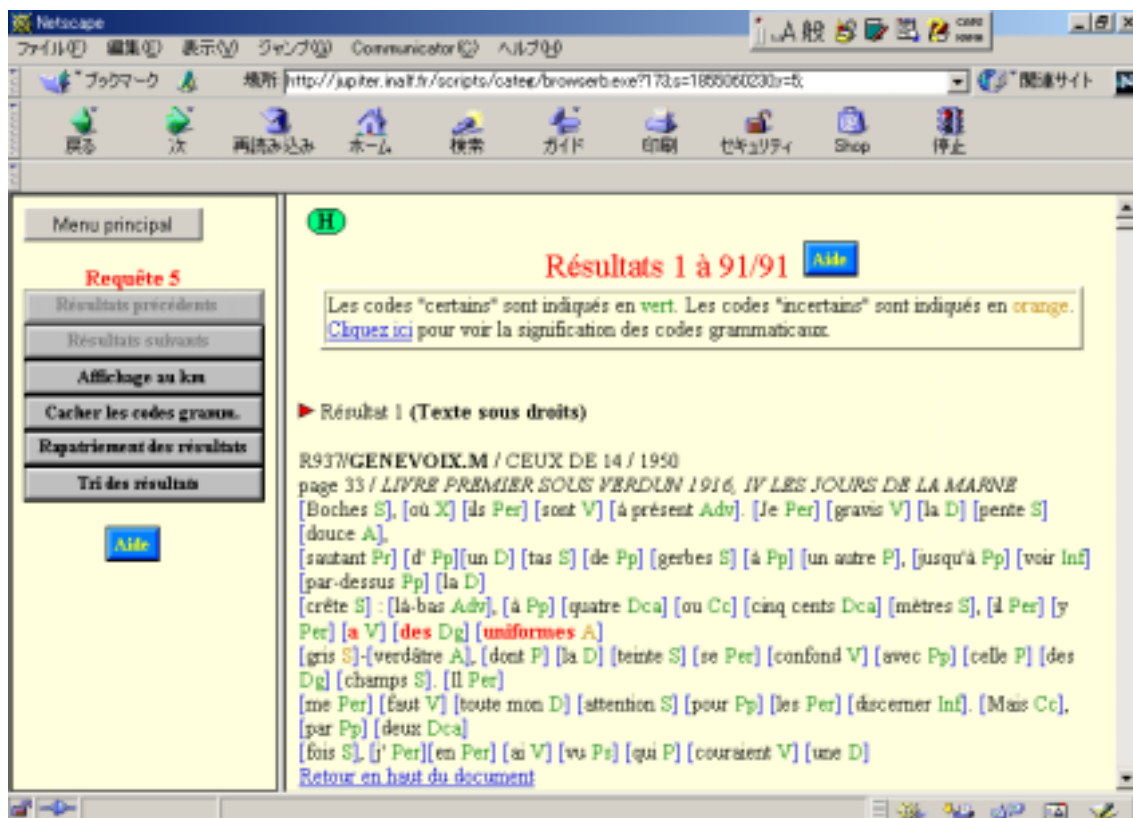
<http://lang.nagoya-u.ac.jp/~fujimura/shuchu/avoirtombe.crp>



Rôle du catégoriseur

Un catégoriseur est un programme ayant pour fonction de découper un texte en une suite de segments auxquels il va attribuer une catégorie grammaticale. Généralement un segment comprend un seul mot, mais il peut arriver qu'un segment en contienne plusieurs, dans la mesure où ce segment peut être considéré comme une entité grammaticale insécable (par exemple une locution adverbiale). Aucune norme ne définit l'ensemble des catégories attribuées: chaque catégoriseur possède donc son propre ensemble de catégories (voir l'ensemble des catégories définies dans le cas de Frantext).





B. Discotext: 1827年から1923年までのフランス語の文学作品300点を収録したCD-ROM。1992年。MS-DOS対応。FRANTEXTの一部を収録したもの。

C. European Language Resources Association

<http://www.icp.grenet.fr/ELRA/home.html>

3.2. 流用型のコーパス:

3.2.1. CD-ROM

Le Monde 1997-98: 320MB

Le Monde Diplomatique 1984-98: 約1320万語, 82.6MB

3.2.2. Web上のコーパス

3.2.2.1. Bibliothèque nationale de France

http://gallica.bnf.fr/textesListe.htm#_F

フランスのBibliothèque nationaleの電子化された資料(印刷刊行された著作と

画像) を公開する圧倒的な規模をもつ画期的なサイト。当初はフランス 19 世紀を対象としていたが、"Gallica Classique" が新たに登場し、中世から 1914 年にまで至るフランスの著作の相当部分をカバーするようになった。

3.2.2.2.Le project Gutenberg

<http://promo.net/pg/index.html>

英語のコーパスが中心だが、フランス語の作品も現在のところ 44 点ある。他の言語のものもある。

3.2.2.3.GOOGLE:

<http://www.google.com>

"aller +au +coiffeur": 26	"aller +chez +le coiffeur" :	396
"aller +au docteur": 47	"aller +chez +le docteur":	153
"aller au professeur": 1	"aller +chez le professeur":	18
"aller +à +mes parents":0	"aller +chez +mes parents":	61
"travaille à Renault": 6	"travaille chez Renault":	43
"travaille à IBM": 6	"travaille chez IBM"	73
"+à Coubertin": 499	"chez Coubertin":	2
"aller +en pied": 4	"aller +à pied":	1230
"aller +en vélo": 97	"aller +à vélo":	119
"aller +en voiture": 415	"aller +à voiture":	0

Leeman-Bouix, Danielle, 1994, "Les fautes de français existent-elles?", Seuil

4. ツール：フランス語の分析のためのプログラム

4.1.TEXTANA

<http://www.biwa.ne.jp/~aka-san/>

TEXTANA 用フランス語動詞活用辞書

<http://taweb.aichi-u.ac.jp/hnakao/txtana.html>

De 複数形形容詞 名詞 s

(@avoir|@être) des* (bons|bonnes|beaux|belles)

Au pluriel, de bons fruits est le tour habituel dans la langue écrites; il s'entend couramment chez les gens qui ont un langage soigné; mais des bons fruits prévaut dans la langue parlée et se répend dans la langue écrite. (*Le Bon Usage*, 659)

本当にことばのレベルの問題だけがファクターなのだろうか？

corpus

Le Monde97-98

Le Monde Diplomatique 84-98

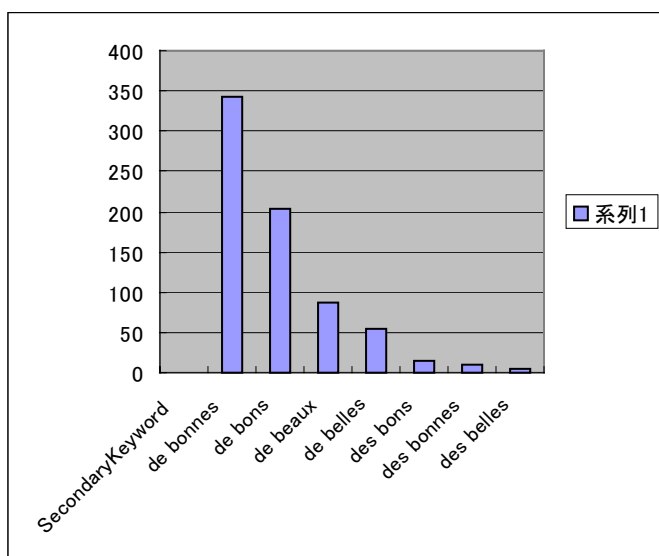
L'Humanité 99

Les dernières nouvelles d'Alsace 99

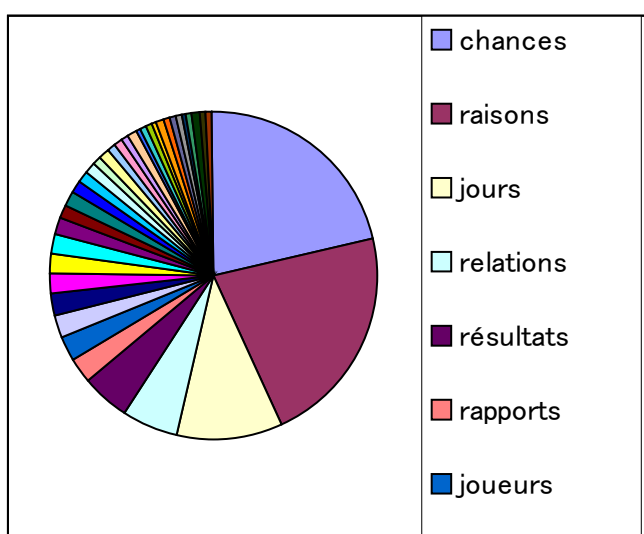
Le télégramme 96

Proust (1871-1922) A LA RECHERCHE DU TEMPS PERDU, TOME I, DU COTÉ
DE CHEZ SWANN

総数：718 occurrences



< 後続する名詞の分布 >



```

3R      Count
chances      107
  @avoir de bonnes chances :107
raisons      106
  avoir de bonnes raisons :105
jours         52
  avoir de beaux jours devant :50
relations     27
  
```

avoir de bonnes relations avec :22
 résultats 23
 avoir de bons résultats :23
 rapports 13
 avoir de bons rapports :13
 joueurs 12、restes 11、choses 11、sensations 10、
 yeux 10、idées 9、et 8、arguments8、
 occasions 7、notes 6、contacts 6、amis 5、
 jambes 5、moments 5

他のファクター

de bonnes 331
 de bons 129
 des bonnes 9
 des bons 10

	il y a	他	総数
de	66	394	460
des	9	10	19
	75	404	479

72.03 387.97
 2.97 16.03
 0.000104
 il y a の後は des が多い
 p < 0.0005

des (bons|bonnes)..

1. "IL Y A DES BONNES CHOSES à écouter ce soir ?" Dans son hôtel parisien, Clint Eastwood s'enquiert d'un possible programme auprès de ses visiteurs, "spécialistes" du jazz,
2. mais, "parfois, il a des bonnes idées".
3. "Nous avons l'avantage d'avoir des partenaires qui nous font confiance sur de longues périodes. Ils savent qu'il y a des bons et des mauvais moments. Ils nous ont fait part de leur soutien."

4. Pour moi, il y a des bons et des mauvais livres.
5. J'ai bossé j'ai eu des bonnes notes, explique Sophie, une fille d'ouvrier.
6. Nous avons eu des bons contacts avec la communauté chrétienne.
7. « Dans une négociation difficile, il y a des bons moments et des moments moins faciles.
8. Cette loi a des bons côtés.
9. "Il peut y avoir des bonnes politiques de gauche, des mauvaises politiques de gauche, des bonnes politiques de droite et des mauvaises politiques de droite".

(@avoir|@être)

a	ont	avoir	sont	avait	eu	être	avons	ai
aura	avaient	avais	est	aurait	as	ayant		
étaient	avez	aurez	sommes	eut	ait			
seront	soient	avons	aviez	auront				
eurent	était	aurons	aurions	étions				
étant	soyons	seraient	auraient	soit	soyez			
aurai								

Type39

725

なぞなぞ：あるはずなのにない活用形はなんでしょう？

(教訓：コンピュータに頼るわけにはいかない。フランス語の知識と言語学の知識)

既存のプログラムの問題点：

4.2. エディタの検索機能を使う

秀丸 grep、

4.3. Unix のコマンドの組み合わせ

grep, sort, uniq など

```
% less kaigyo.sed
s/ /¥
/g

$ tr '¥n' ' ' < html/proust/proust1.crp | tr '¥r' ' ' | sed "s/['¥.¥]¥(,!¥?:;-)/ /g" | sed 's/' / /g'
| sed 's` / /g' | sed 's/¥*//g' | sed 's/¥+//g' | sed 's/ * / /g' | sed -f kaigyo.sed | grep -v '^ *$'
| sort -f | uniq -ci > html/proust/motsfreq.txt

% sort -n -r < corpus/proust2.txt > corpus/proust.4.txt
```

語彙頻度表:

アルファベット順 :

<http://lang.nagoya-u.ac.jp/~fujimura/shuchu/proust/proust2.txt>

頻度順 :

<http://lang.nagoya-u.ac.jp/~fujimura/shuchu/proust/motsfreq.txt>

頻度順 (大文字と小文字の区別をしない):

<http://lang.nagoya-u.ac.jp/~fujimura/shuchu/proust/motsfreq1.txt>

1 語 1 文のテキスト

<http://lang.nagoya-u.ac.jp/~fujimura/shuchu/proust/proust3.txt>

http://lang.nagoya-u.ac.jp/~ohna/perl_lesson/locale.html

4.4 表計算ソフト (Excel など) を使った語彙頻度表・n-gram

http://lang.nagoya-u.ac.jp/~ohna/word_list.html

<http://lang.nagoya-u.ac.jp/~fujimura/files/wordlist.xls>

<http://lang.nagoya-u.ac.jp/~fujimura/files/n-gram-pub.xls>

記号を削除し, スペースに置き換える。ハイフンは注意が必要。シングルクォーテーションも注意が必要。スペースが 2 個以上の個所は 1 個にする。タブもとる。ギョメ, NBSP (no-break space) もとる。改行をとる。空行をとる。¥r¥n で改行をとる。もう一度スペースを一つにする。最後にスペースに改行をいれる。それで, やっとできた。

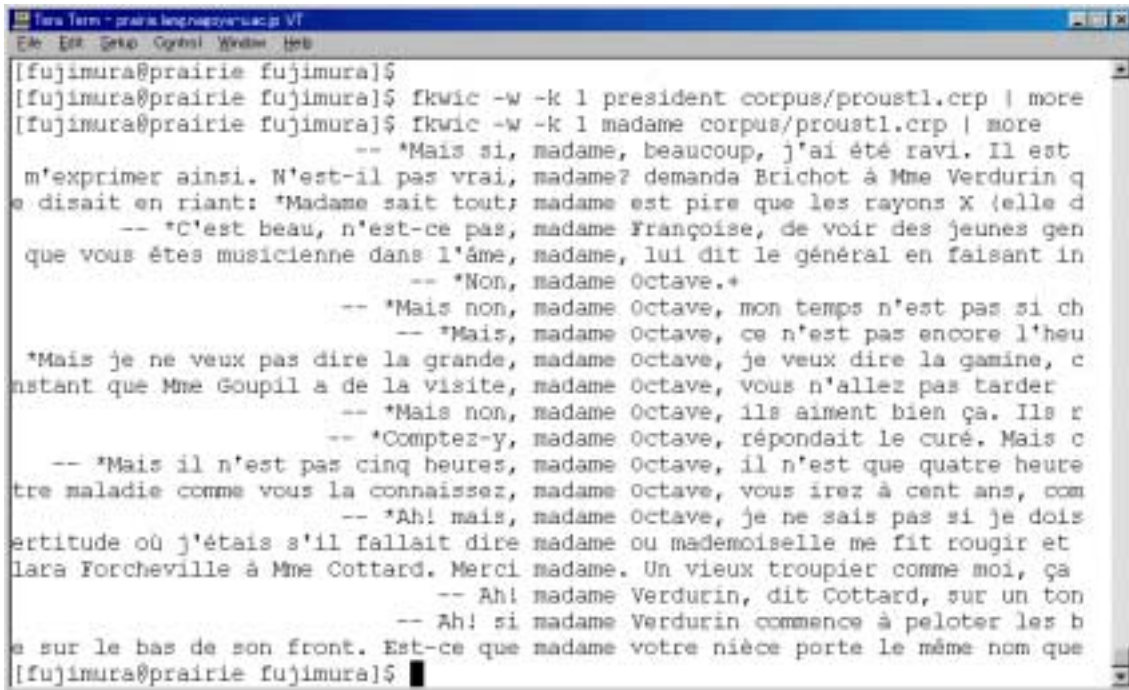
```
perl -pe 's/[;!¥¥(¥)¥?:-]/ /g;' lemonde2000-12-1-1.crp | perl -pe "s/¥' /g;" | perl -pe 's/ + /g;' | perl -pe 's/¥t/g;' | perl -pe 's/[¥xAO¥xAB¥xBB]/g;' | perl -pe 's/¥n/g;' | perl -pe
```

```
's/^$/g;'> lemonde2000-12-2.crp
```

```
$ perl -pe 's/[¥r¥n]/g;' lemonde2000-12-2.crp | perl -pe 's/  */g;' > yakekuso.crp
```

```
$ perl -pe 's/  */¥n/g;' yakekuso.crp > yakekuso-1.crp
```

4.4.fkwc



```
Termin - prairie.lang.nagoya-u.ac.jp VT
File Edit Goto Control Window Help
[fujimura@prairie fujimura]$
[fujimura@prairie fujimura]$ fkwc -w -k 1 president corpus/proust1.crp | more
[fujimura@prairie fujimura]$ fkwc -w -k 1 madame corpus/proust1.crp | more
-- *Mais si, madame, beaucoup, j'ai été ravi. Il est
m'exprimer ainsi. N'est-il pas vrai, madame? demanda Bichot à Mme Verdurin q
e disait en riant: *Madame sait tout; madame est pire que les rayons X (elle d
-- *C'est beau, n'est-ce pas, madame Françoise, de voir des jeunes gen
que vous êtes musicienne dans l'âme, madame, lui dit le général en faisant in
-- *Non, madame Octave.+
-- *Mais non, madame Octave, mon temps n'est pas si ch
-- *Mais, madame Octave, ce n'est pas encore l'heu
*Mais je ne veux pas dire la grande, madame Octave, je veux dire la gamine, c
stant que Mme Goupil a de la visite, madame Octave, vous n'allez pas tarder
-- *Mais non, madame Octave, ils aiment bien ça. Ils r
-- *Comptez-y, madame Octave, répondait le curé. Mais c
-- *Mais il n'est pas cinq heures, madame Octave, il n'est que quatre heure
tre maladie comme vous la connaissez, madame Octave, vous irez à cent ans, com
-- *Ah! mais, madame Octave, je ne sais pas si je dois
ertitude où j'étais s'il fallait dire madame ou mademoiselle me fit rougir et
lara Forcheville à Mme Cottard. Merci madame. Un vieux troupiier comme moi, ça
-- Ah! madame Verdurin, dit Cottard, sur un ton
-- Ah! si madame Verdurin commence à peloter les b
e sur le bas de son front. Est-ce que madame votre nièce porte le même nom que
[fujimura@prairie fujimura]$
```

名古屋大学コーパスプロジェクト :

<http://prairie.lang.nagoya-u.ac.jp/>

4;5.perl を使って自分でプログラミングする

```

Tera Term - prairie.lingvazeyu.ac.jp VT
File Edit Setup Control Window Help
[fujimura@prairie fujimura]$ perl plscript/ezkwic2_2.pl corpus/proust1.crp
SEARCH STRING: monsieur
left length: 10
right length: 10
 1 | proust1.crp |           A |Monsieur| Gaston Ca|           A
 2 | proust1.crp | rai type, |monsieur| Swann!+ C| type, rai
 3 | proust1.crp | *Voyons, |monsieur| Swann, lu| *Voyons,
 4 | proust1.crp | avoir ce |monsieur| à diner, | ce avoir
 5 | proust1.crp | n certain |monsieur| de Charlu| certain n
 6 | proust1.crp | qu'a donc |Monsieur| à pleurer| donc qu'a
 7 | proust1.crp | -- *Mais, |monsieur| Bloch, qu| *Mais, --
 8 | proust1.crp | -- *|Monsieur|, je ne pu| * --
 9 | proust1.crp | -- *Non |monsieur|, mes pare| *Non --
10 | proust1.crp | -- *|Monsieur| le Curé, | * --
11 | proust1.crp | mme?+ -- *|Monsieur| le Curé a| * -- mme?+
12 | proust1.crp | sez-vous, |monsieur| le liseur| sez-vous,
13 | proust1.crp | nnaissez, |monsieur|, la... le| nnaissez,
14 | proust1.crp | quelle un |Monsieur| habillé d| un quelle
15 | proust1.crp | s plus ce |monsieur|. Et puis | ce plus s
16 | proust1.crp | ez bien, *|monsieur|+ Biche, r| * bien, ez
17 | proust1.crp | e de dire |monsieur|, à rendre| dire de e
18 | proust1.crp | férence: *|Monsieur| Swann, vo| * férence:
19 | proust1.crp | ue chose, |Monsieur| Swann?+ | chose, ue
20 | proust1.crp | is. Mais, |monsieur| Swann, vo| Mais, is.

```

ezkwic2_2.pl

.....

```

print STDERR "SEARCH STRING: "; $string = <STDIN>; chomp $string;

print STDERR "left length: "; $left = <STDIN>; chomp $left;

print STDERR "right length: "; $right = <STDIN>; chomp $right;

$/ = "";

while(<>){
s/ *%n */ /g;

while(/$string/ig){

$count++;

$key = $&;

$post = $';

$pre = substr ($`, -$left);

$pre_rev = join (' ', reverse (split (/s+/, $pre)));

$file = $ARGV;

$file =~ s#.*(?:/)+##$1#;

printf "%5d | %-12.12s | %${left}s|s|%-${right}.${right}s| %${left}s%#n",

```

```
$count, $file, $pre, $key, $post, $pre_revd;  
}  
}  
exit;
```

.....

インターネットからのフランス研究---- 社会・文化の一般的情報から人文学研究
まで <http://www.lang.nagoya-u.ac.jp/~ino/index-j.htm>