

コーパス言語学のための Perl入門

Windows環境でのコーパス分析

英語コーパス学会 第22回大会 ワークショップ
於: 明海大学 浦安キャンパス
講師: 赤瀬川 史朗

aka-san@mx.biwa.ne.jp

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

1. GUIとCUI

■ 2つのインターフェイス

■ GUI (Graphical User Interface)

マウス、アイコン、メニュー、ダイアログ、ウィンドウ

Windows, MacOS, X Window System(Unix)

■ CUI (Character User Interface)

キーボード、コマンドライン

コマンドプロンプト、Unix、MS-DOS

コマンドプロンプトはCUIのインターフェイスをもつ

1.1. GUIは万能か?

- マウスやアイコンを使うので操作が直感的
- WYSIWYG(What You See Is What You Get)
- アプリケーションソフトの操作が統一される

メリットばかりのように思えるが...

- アプリケーションソフトは小回りがきかない
- アプリケーションソフト同士の連携が難しい
- 自動処理・一括処理が苦手

1.2. CUIとテキスト処理

CUIの特長は...

- プログラムを組み合わせるので小回りがきく
- 自動処理・一括処理ができる



テキスト処理の要件は...

- データにあわせたケースバイケースの対応
- 大量の一括処理

CUIはテキスト処理に適したインターフェイス

1.3. コマンドプロンプトの3つの道具

■ コマンド

コマンドプロンプトで標準で提供される「命令」、ファイル操作・ディレクトリ操作が主体 (2.1. ~ 2.15.)

■ コンソールプログラム

コマンドプロンプト上で動作する既成のプログラム、ふつう単独で動作する (2.22 ~ 2.23.)

■ スクリプト言語

スクリプト(プログラムの一種)を実行・開発するための環境、スクリプトを実行するには言語本体のインストールが必要 (3.1 ~)

1.4. 道具を連携させるための仕組み

- リダイレクト

データをファイルから読み書きする (2.17. ~ 2.19.)

- パイプ

データをプログラムからプログラムへ直接受け渡す (2.20.)

- バッチファイル

コマンド、コンソールプログラム、スクリプト言語を一括実行、複数ファイルの処理もできる (2.21. ~ 2.22.)

2. コマンドプロンプト

- 起動画面 (2.1.)
- カレントディレクトリ (2.2.)
- ファイルの作成 (2.3.)
- ファイルの一覧 (2.4.)
- ファイルの内容の表示 (2.5.)
- サブディレクトリの作成 (2.6.)
- サブディレクトリへ移動 (2.7.)
- ディレクトリツリーの表示 (2.8.)
- ファイルのコピー (2.9.-10.)
- ファイルの移動 (2.11.)
- ファイル名の変更 (2.12.)
- コマンドのオプション (2.14.-15.)
- よく使うその他のコマンド (2.16.)
- 標準入力・標準出力 (2.17.)
- リダイレクト (2.18.-20.)
- パイプ (2.21.)
- バッチファイル (2.22.-23.)
- コンソールプログラム (2.24.-25.)
- コマンドサーチパス (2.26.)

2.1. 起動画面

起動

[スタート|すべてのプログラム|アクセサリ|コマンドプロンプト]



The image shows a screenshot of a Windows XP Command Prompt window. The title bar reads "コマンド プロンプト". The window content displays the following text:

```
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

c:¥corpus>
```

Annotations in Japanese are overlaid on the screenshot:

- A horizontal line is drawn under the prompt "c:¥corpus>".
- A vertical line descends from the center of this horizontal line to the text "プロンプト...現在、直接操作できるディレクトリ (フォルダ) を表す".
- Another horizontal line is drawn under the text "プロンプト...".
- A vertical line descends from the center of this second horizontal line to the text "カレントディレクトリ".

2.2. カレントディレクトリ

cd

カレントディレクトリの表示

```
d: ¥corpus>cd
```

```
d: ¥corpus   カレントディレクトリが表示される
```

```
d: ¥corpus>   コマンド入力待ち状態
```

2.3. ファイルの作成

copy con ファイル名

画面からの入力をファイルに書き込む

```
d: ¥corpus>copy con first.txt    conはコンソール
This is a text file.
^Z    [Ctrl-Z]を入力
      1 個のファイルをコピーしました。
      ファイルが作成された
```

2.4. ファイルの一覧

dir

カレントディレクトリのファイルの一覧

```
d: ¥corpus>dir
```

```
ドライブ C のボリューム ラベルがありません。  
ボリューム シリアル番号は xxxx-xxxx です
```

```
d: ¥corpus のディレクトリ
```

```
2003/10/25  10: 30    <DI R>          .  
2003/10/25  10: 30    <DI R>          ..  
2003/10/25  10: 35                22 first.txt
```

先ほどのファイルが表示された

:

2.5. ファイルの内容の表示

type ファイル名

ファイルの内容の表示

```
d: ¥corpus>type first.txt
```

```
This is a text file.
```

ファイルの内容が表示される

2.6. サブディレクトリの作成

md ディレクトリ名

ディレクトリの作成

```
d: ¥corpus>md sub
```

```
d: ¥corpus>di r
```

```
                :  
2003/10/25  10: 30  <DI R>      .  
2003/10/25  10: 30  <DI R>      ..  
2003/10/25  10: 30                22 first.txt  
2003/10/25  10: 30  <DI R>      sub
```

作成したディレクトリが表示された

2.7. サブディレクトリへ移動

cd ディレクトリ名

ディレクトリの移動

```
d: ¥corpus>cd sub
```

```
d: ¥corpus¥sub>
```

subディレクトリに移動

```
d: ¥corpus¥sub>cd .. .. は親ディレクトリを表す
```

```
d: ¥corpus>
```

親ディレクトリに移動

2.8. ディレクトリツリーの表示

tree ディレクトリ名

ディレクトリツリーの表示

```
d: ¥corpus>md sub¥sub2    subの下にsub2を作成
```

```
d: ¥corpus>tree
```

フォルダ パスの一覧

ボリューム シリアル番号は xxxxxxxx xxxx:xxxx です

```
C: . . (ピリオド)はカレントディレクトリ
```

```
    sub
```

```
        sub2
```

2.9. ファイルのコピー 1

copy コピー元 コピー先 ファイルのコピー

```
d: ¥corpus>copy first.txt second.txt  
1 個のファイルをコピーしました。
```

dir /b ファイル名 ファイル名・ディレクトリ名だけを表示

```
d: ¥corpus>dir /b *.txt  
テキストファイルをまとめて指定  
(ワイルドカード)  
  
first.txt  
second.txt
```

2.10. ファイルのコピー 2

copy ファイル名 ディレクトリ名

ファイルをディレクトリにコピー

```
d: ¥corpus>copy *.txt sub
```

テキストファイルをsubにコピー

```
fi rst.txt
```

```
second.txt
```

2 個のファイルをコピーしました。

```
d: ¥corpus>di r /b sub¥*.txt
```

```
fi rst.txt
```

```
second.txt
```

2.11. ファイルの移動

move ファイル名 移動先ディレクトリ

ファイルの移動

```
d: ¥corpus>move sub¥*.txt sub¥sub2
```

subのテキストファイルをsub¥sub2に移動

```
d: ¥corpus¥sub¥fi rst.txt
```

```
d: ¥corpus¥sub¥second.txt
```

```
d: ¥corpus>di r /b sub¥sub2¥*.txt
```

```
fi rst.txt
```

```
second.txt
```

2.12. ファイル名の変更

ren 元のファイル名 変更後のファイル名

ファイル名の変更

```
d: ¥corpus>ren fi rst. txt 1st. txt
```

fi rst. txtを1st. txtに変更

```
d: ¥corpus>ren second. txt 2nd. txt
```

second. txtを2nd. txtに変更

```
d: ¥corpus>di r /b *. txt
```

1st. txt

2nd. txt

2.13. 基本コマンドのまとめ

cd	c hange d irectory	ディレクトリの移動*
copy		ファイルのコピー
dir	d irectory	ファイルの一覧
type		ファイルの内容の表示
md	m ake d irectory	ディレクトリの作成
tree		ディレクトリツリーの表示
move		ファイルの移動
ren	r ename	ファイル名の変更

* 引数なしのcdはカレントディレクトリの表示

2.14. dirコマンドのオプション 1

dir /b /ad ディレクトリだけを表示

```
d: ¥corpus>dir /b /ad  
sub
```

dir /b /aa ファイルだけを表示

```
d: ¥corpus>dir /b /aa  
1st.txt  
2nd.txt
```

2.15. dirコマンドのオプション 2

`dir /b /aa /s`

サブディレクトリのファイルも表示

```
d: ¥corpus>dir /b /aa /s
d: ¥corpus¥1st. txt
d: ¥corpus¥2nd. txt
d: ¥corpus¥sub¥sub2¥fi rst. txt
d: ¥corpus¥sub¥sub2¥second. txt
```

2.16. よく使うその他のコマンド

■ 削除

del

delete

ディレクトリの移動*

rd

remove directory

ファイルのコピー

■ ファイル

fc

file compare

ファイルの内容の比較

find

ファイルから文字列を検索

findstr

find string

正規表現を使ってファイルから文字列を検索

■ 環境設定

path

コマンドサーチパスの設定

set

環境変数の設定

2.17. 標準入力・標準出力

標準入力・標準出力

コマンドやコンソールプログラムの結果はふつう画面(コンソール)に出力される。これは、標準の出力先として画面が指定されているためである。これを**標準出力**という。同様に、ふつうキーボードから入力する標準の入力先のことを**標準入力**という。

```
d: ¥corpus>sort      データを並べ替えるコマンド
abc                  これが標準入力
abb
aad
^Z                  [Ctrl-Z]を入力、入力の終わり
```

2.18. リダイレクト 1

> (標準出力のリダイレクト)

リダイレクトを使うと、標準の出力先を画面からファイルに切り替えることができる。標準出力のリダイレクトには>を使う。リダイレクトしてファイルに出力することを「**ファイルに落とす**」という。

```
d: ¥corpus>echo redi rect test      画面に文字列を表示  
redi rect test
```

```
d: ¥corpus>echo redi rect test > redi rect. txt  
          今度はリダイレクトしてファイルに落とす
```

```
d: ¥corpus>type redi rect. txt  
redi rect test
```

2.19. リダイレクト 2

< (標準入力のリダイレクト)

標準入力もリダイレクトすれば、入力先をキーボードからファイルに切り替えることができる。記号は<を使う。

```
d: ¥corpus>copy on before.txt
```

```
abc
```

```
abb
```

```
aad
```

```
^Z    [Ctrl-Z]を入力
```

```
1 個のファイルをコピーしました。
```

```
d: ¥corpus>sort < before.txt
```

```
aad
```

```
abb
```

```
abc
```

2.20. リダイレクト 3

フィルタ < 入力ファイル > 出力ファイル

標準入力から読み込み標準出力に書き出すsortのようなプログラムは、リダイレクトを使えばファイルから読み込んでファイルに書き出すことができる。こうしたタイプのプログラムを**フィルタ**という。

```
d: ¥corpus>sort < before.txt > after.txt
```

```
d: ¥corpus>type after.txt
```

```
aad
```

```
abb
```

```
abc
```

2.21. パイプ

| (パイプ)

あるプログラムの標準出力を別のプログラムの標準入力に連結することをパイプという。リダイレクトとは違い、標準出力をいったんファイルに落とす必要がない。複数のフィルタをパイプで連結すれば、あたかも1つのプログラムのように実行することができるので、柔軟で高度なテキスト処理が可能になる。パイプの記号には|を使う。

```
d: ¥corpus>type before.txt | sort
```

```
aad
```

```
abb
```

```
abc
```

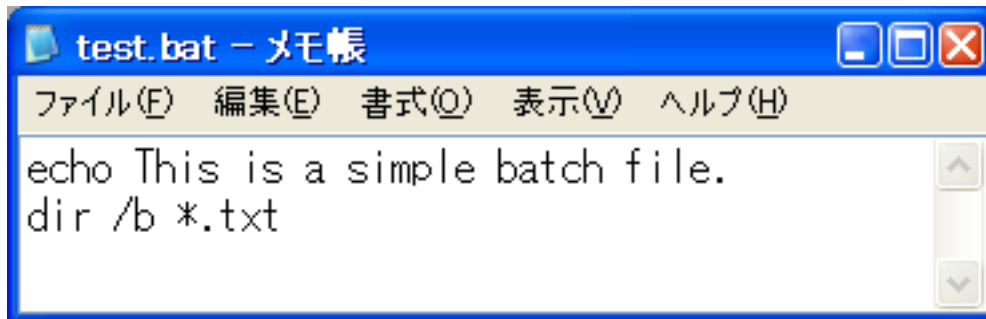
```
d: ¥corpus>type before.txt | sort > after.txt
```

2.22. バッチファイル 1

バッチファイル

複数のコマンドやプログラムを連続して実行することのできるファイル。拡張子には.batを使う。

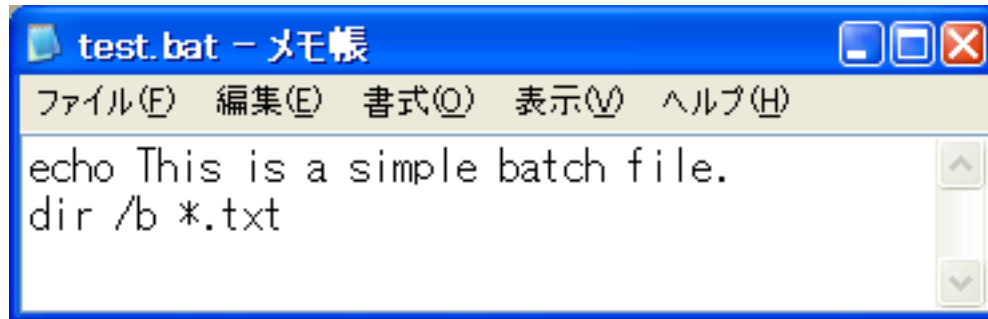
```
d: ¥corpus>notepad test.bat
```



```
d: ¥corpus>di r test.bat
```

```
2003/10/25 10:50          47 test.bat
```

2.23. バッチファイル 2



```
test.bat - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
echo This is a simple batch file.
dir /b *.txt
```

```
d: ¥corpus>test      拡張子. batは省略できる
d: ¥corpus>echo This is a simple batch file.
This is a simple batch file.

d: ¥corpus>dir /b *.txt
1st.txt
2nd.txt
redirect.txt
before.txt
after.txt
```

2.24. コンソールプログラム 1

■ GNU Utilities for Win32

Unix上のGNU UtilitiesのWindows版(日本語非対応)

<http://unxutils.sourceforge.net/>

TextUtilsの一部

cat	con cat enate	ファイルの表示と連結
head	dir ectory	ファイルの先頭を表示
nl	n umber l ine	行番号を振る
tail		ファイルの末尾を表示
uniq	unique	ソート済みのファイルから 重複行を削除
wc	w ord c ount	文字、単語、行のカウント

2.25. コンソールプログラム 2

■ テキストファイル変換プログラム

RTFコンバータ (針谷壮一氏)

- ・RTF (リッチテキストフォーマット) をテキストファイルに変換
- ・テキストファイルの文字コード変換 (80種類の文字コードに対応)

<http://www5b.biglobe.ne.jp/~harigaya/rftcnv.html>

HtoX32c (T-Matsuo氏)

HTMLファイルをテキストファイルに変換 (欧文は英語のみ)

<http://win32lab.com/fsw/htox.html>

pdftotext

PDFファイルをテキストファイルに変換

<http://www.foolabs.com/xpdf/download.html>

2.26. コマンドサーチパスの設定

コンソールプログラムを実行するには、あらかじめその実行ファイルがあるディレクトリをコマンドサーチパスに登録しておく必要がある。この作業を「**パスを通す**」という。パスが通っていないと、プログラムをインストールしても利用することはできない。

ここでは、abc.exeというコンソールプログラムがc:¥abcにインストールされていると仮定して、パスを通してみよう(Windows XP)。

- ・[スタート | マイコンピュータ]をクリック
- ・[システムの情報を表示する]をクリック
- ・[詳細設定]タブをクリック
- ・[環境変数]をクリック
- ・[システム環境変数]から[path]をクリックし、[編集]をクリック
- ・[変数値]の最後にc:¥abcを追加する

2.27. コマンドプロンプトの参考文献

岡田庄司『プチリファレンス Windows DOSプロンプト』、秀和システム、2002

天野司『Windows XP/2000 コマンドプロンプト』、技術評論社、2002

中尾浩・赤瀬川史朗・宮川真悟『コーパス言語学の技法 I - テキスト処理入門』、夏目書房、2002

市川昭彦『Windows XPのコマンドプロンプト入門』、ディーアート、2001

3. Perl

Perlの特徴

- 最も広く使われているスクリプト言語
- 強力な正規表現、テキスト処理に最適

コンソールプログラムとの違い

- スクリプトと呼ばれるプログラムを用意する
- スクリプトには拡張子.plがつく
- スクリプトの実行には、Perl本体をインストールしておく必要がある

3.1. インストール

1. 次のURLにアクセスし、ユーザ登録する(無料)。

<http://www.activestate.com/Products/Download/Register.plex?id=ActivePerl>

2. 次の画面に進み、バージョン5.8.0(2003年10月現在)のWindows用(MSI)をダウンロードする。Microsoft Installerがインストールされていない機種(98/Me, NT)では、同じページからInstallerをダウンロードしインストールしておく必要がある。

ActivePerl 5.8.0 build 806	Windows	MSI ← Click	11.5MB
	Windows	AS package	11.4MB

3. ダウンロードしたファイルActivePerl-5.8.0.806-MSWin32-x86.msiを実行し、指示に従ってインストールを行う。

3.2. インストールの確認

`perl -v` バージョンの表示

```
d: ¥corpus>perl -v
```

```
This is perl, v5.8.0 built for MSWin32-x86-multi-thread
```

```
(with 1 registered patch, see perl -V for more detail)
```

```
Copyright 1987-2002, Larry Wall
```

```
Binary build 806 provided by ActiveState Corp.  
http://www.ActiveState.com
```

```
Built 00:45:44 Mar 31 2003
```

3.3. ワンライナー

ワンライナー コマンドラインに書いた1行プログラム

perl -e "コード" ワンライナーの実行 (Windowsではコードはダブルクォート(")で囲む)

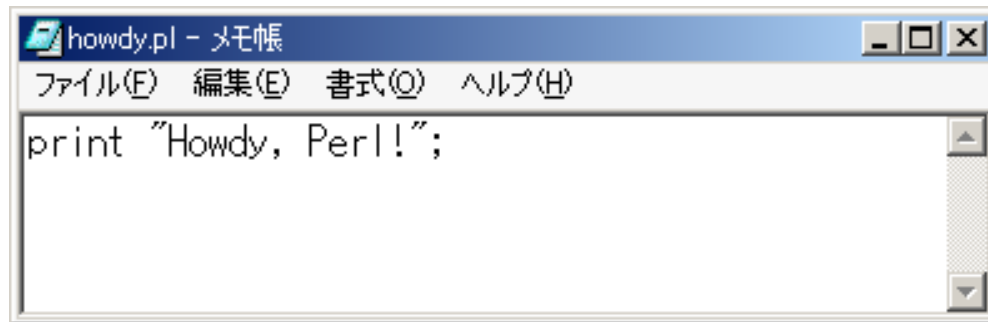
エスケープ コード中に"を使うときは¥"のようにする。

```
d: ¥corpus>perl -e "print ¥"Howdy, Perl!¥n¥" "
Howdy, Perl!
```

3.4. スクリプト

スクリプト プログラムの設計図となるテキストファイル

```
d: ¥corpus>notepad howdy. pl
```



[ファイル|メモ帳の終了]
をクリックして、ファイルを
保存する

perl スクリプト名 スクリプトの実行

```
d: ¥corpus>perl howdy. pl  
Howdy, Perl !
```

3.5. Perlの参考文献

深沢千尋『すぐわかるPerl』、技術評論社、1999

Randal L. Schwartz 『初めてのPerl 第3版』、オライリージャパン、2003

Andrew L. Johnson 『プログラミング言語Perlマスターコース』、ピアソン・エデュケーション、2000

中尾浩・赤瀬川史朗『コーパス言語学の技法 II - 実践テキスト処理』、夏目書房、2003(近刊)

4. CAT (Corpus Analysis Toolkit)

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

■ XML文書コーパスの作成

センテンス、パラグラフの区切りの明確化

■ 基本的なコーパス処理を網羅

単語・n-gram頻度, grep, KWIC, コロケーション, 統計値

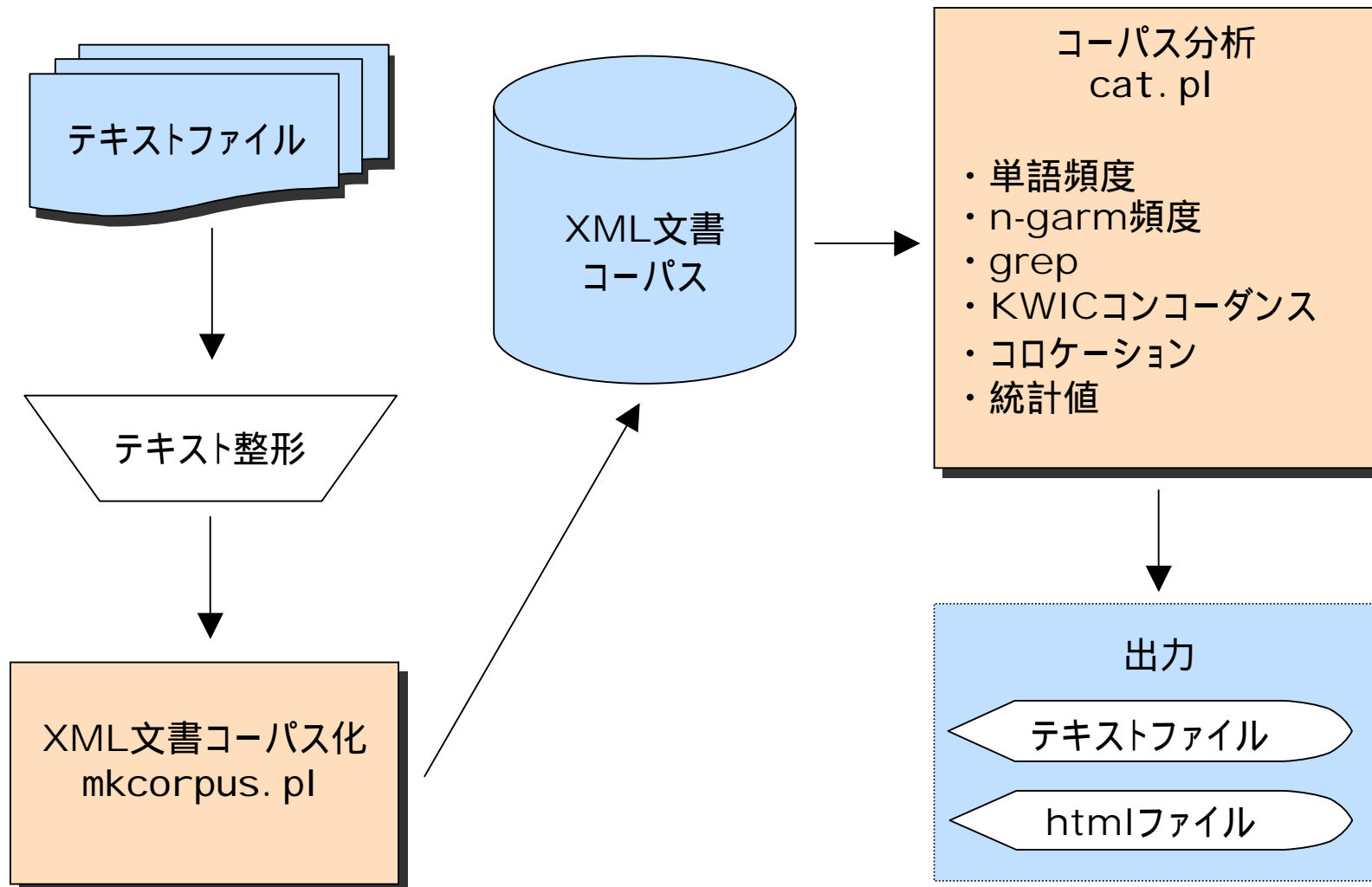
■ 柔軟な出力

ファイルごとの個別処理、複数ファイルの一括処理

プレーンテキスト、HTMLファイルの2つの出力形式

■ フリーウェア

4.1. CATの概要



4.2. 準備 (モジュールのインストール)

文書コーパス作成スクリプトmkcorpus.plで使用するモジュール
Lingua::EN::Sentenceをインストール

ppm3

ActiveState Perlに付属するモジュールインストーラ

```
d: ¥corpus>ppm3
PPM - Programmer's Package Manager version 3.0.1.
:
ppm> search sentence
Searching in Active Repositories
:
3. Lingua-EN-Sentence [0.25] Module for splitting text
:
ppm> install 3
Package 3:
=====
Install 'Lingua-EN-Sentence' version 0.25 in ActivePerl 5.8.0.805.
=====
Downloaded 5614 bytes.
:
Successfully installed Lingua-EN-Sentence version 0.25 in ActivePerl
5.8.0.805.
ppm> quit
```

4.3. XML文書コーパスの作成

■ サンプルテキスト

Project Gutenberg <http://promo.net/pg/>

George Bernard Shaw "*Man and Superman*", etc.

■ テキスト整形

不要な情報を削除し、段落の終わりに空行を挿入する

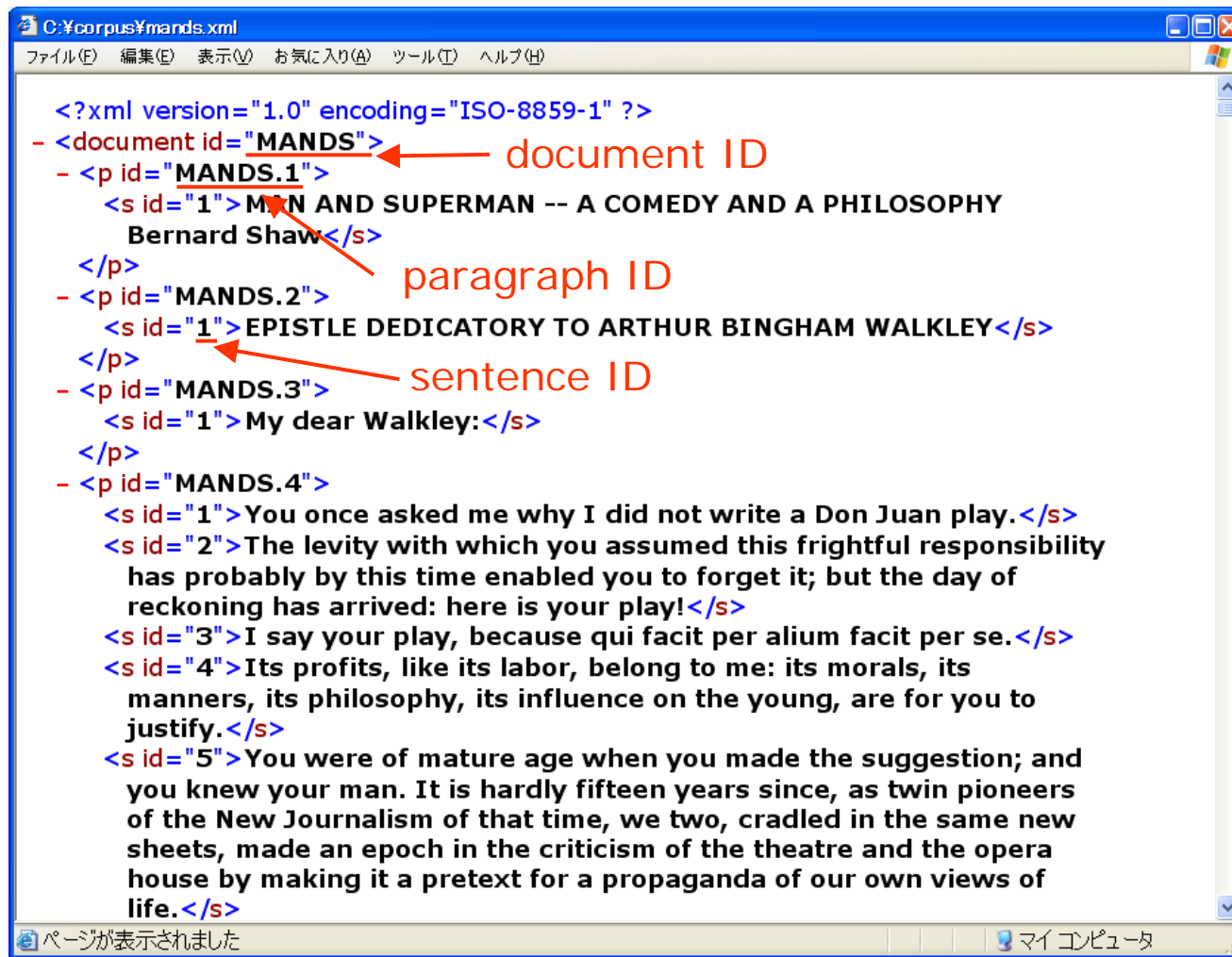
■ XML文書コーパス作成

```
perl mkcorpus.pl -i=ファイル名
```

```
d: ¥corpus>perl mkcorpus.pl -i=mands.txt
Done: mands.txt => mands.xml , mands.html
```

* mands.txtは整形済みテキスト

4.4. XML文書コーパスの構造



The screenshot shows a text editor window titled "C:\corpus\mands.xml". The menu bar includes "ファイル(F)", "編集(E)", "表示(V)", "お気に入り(A)", "ツール(T)", and "ヘルプ(H)". The main text area contains XML code with several annotations in red text and arrows:

- `<?xml version="1.0" encoding="ISO-8859-1" ?>`
- `- <document id="MANDS">` ← **document ID**
- `- <p id="MANDS.1">` ← **paragraph ID**
- `<s id="1">MAN AND SUPERMAN -- A COMEDY AND A PHILOSOPHY
 Bernard Shaw</s>`
- `</p>`
- `- <p id="MANDS.2">` ← **paragraph ID**
- `<s id="1">EPISTLE DEDICATORY TO ARTHUR BINGHAM WALKLEY</s>` ← **sentence ID**
- `</p>`
- `- <p id="MANDS.3">` ← **paragraph ID**
- `<s id="1">My dear Walkley:</s>`
- `</p>`
- `- <p id="MANDS.4">`
- `<s id="1">You once asked me why I did not write a Don Juan play.</s>`
- `<s id="2">The levity with which you assumed this frightful responsibility
 has probably by this time enabled you to forget it; but the day of
 reckoning has arrived: here is your play!</s>`
- `<s id="3">I say your play, because qui facit per alium facit per se.</s>`
- `<s id="4">Its profits, like its labor, belong to me: its morals, its
 manners, its philosophy, its influence on the young, are for you to
 justify.</s>`
- `<s id="5">You were of mature age when you made the suggestion; and
 you knew your man. It is hardly fifteen years since, as twin pioneers
 of the New Journalism of that time, we two, cradled in the same new
 sheets, made an epoch in the criticism of the theatre and the opera
 house by making it a pretext for a propaganda of our own views of
 life.</s>`

The status bar at the bottom shows "ページが表示されました" and "マイコンピュータ".

4.5. 処理名の指定

```
perl cat.pl --処理名
```

スクリプト名の次に処理名を指定する。
処理名の先頭にはハイフンを2つ置く。

単語頻度

```
perl cat.pl --freq
```

n-gram頻度

```
perl cat.pl --ngram
```

grep

```
perl cat.pl --grep
```

KWICコンコーダンス

```
perl cat.pl --kwi c
```

コロケーション(頻度)

```
perl cat.pl --col
```

コロケーション(統計値)

```
perl cat.pl --stat
```

4.6. 共通オプション 1

- -i=入力ファイル名 (必須)

入力先となるコーパスファイルを指定する。コーパスファイルはmkcorpus.plで作成したXML文書コーパスに限る。複数のファイルを指定するときはワイルドカードを使う。

- -o=出力ファイル名 (省略可)

出力先のファイル名のベースネーム(ファイル名の拡張子を除いた部分)を指定する。テキストファイルとして出力するときは拡張子.txt、htmlファイルのときは.htmlが付加される。指定がないときは、入力ファイルのベースネームに各処理ごとの拡張子と.txtまたは.htmlが付加される。たとえば、入力ファイルがmands.xmlで単語頻度を行なった場合、出力ファイル名はmands.frq.txtまたはmands.frq.htmlになる。

4.7. 共通オプション 2

- -1 (省略可)

コーパスファイルごとに1つの出力ファイルを作成する。指定がないときは、複数のコーパスファイルに対し、1つの出力ファイルが作成される。

- -h (省略可)

htmlファイルを出力する。指定がないときはプレーンテキストを出力する。

4.8. 単語頻度 1 頻度順

--freq -i=ファイル名

```
d: ¥corpus>perl cat.pl --freq -i=mands.xml
```

```
-----  
CAT (Corpus Analysis Toolkit) ver. 1.00 Copyright  
-----
```

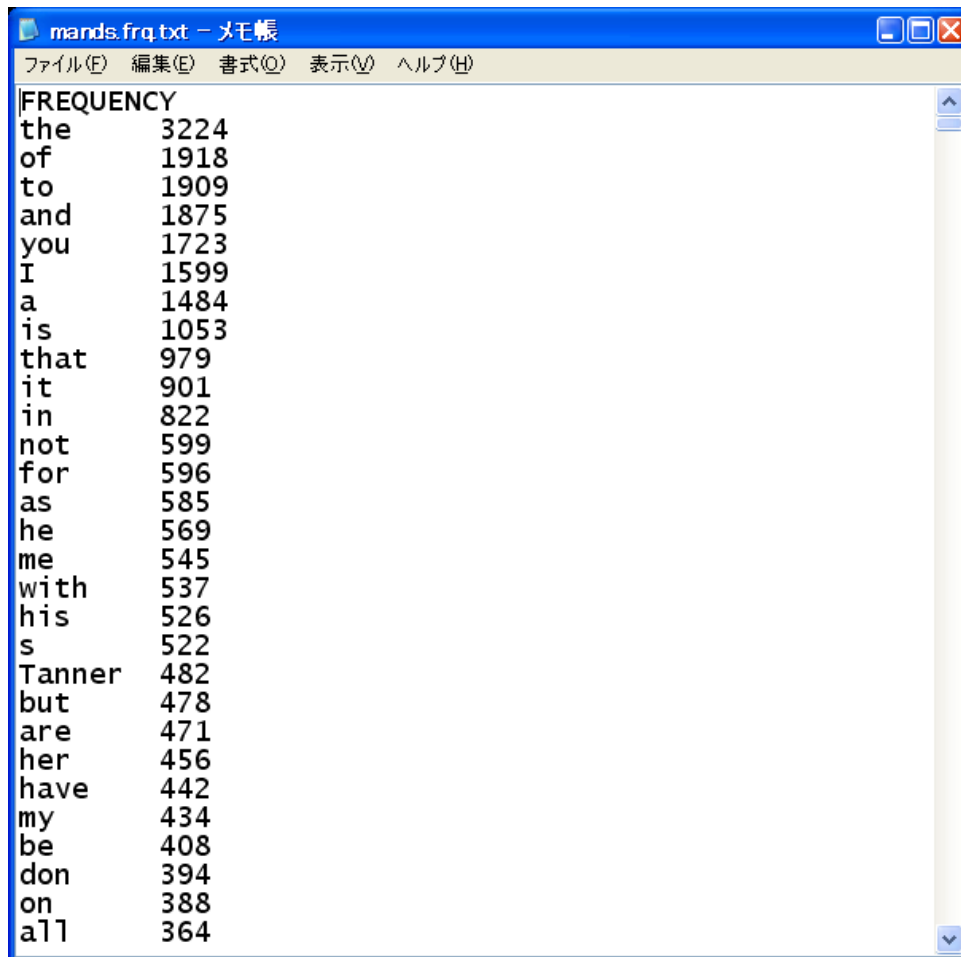
```
INPUT: c: ¥corpus¥mands.xml
```

```
OUTPUT: c: ¥corpus¥mands.frq.txt
```

```
d: ¥corpus>notepad mand.frq.txt
```

4.9. 単語頻度 2 頻度順

perl cat.pl --freq -i=mands.xml の結果



```
mands.freq.txt - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
FREQUENCY
the      3224
of       1918
to       1909
and      1875
you      1723
I        1599
a        1484
is       1053
that     979
it       901
in       822
not      599
for      596
as       585
he       569
me       545
with     537
his      526
s        522
Tanner   482
but      478
are      471
her      456
have     442
my       434
be       408
don      394
on       388
all      364
```

4.10. 単語頻度 3 アルファベット順

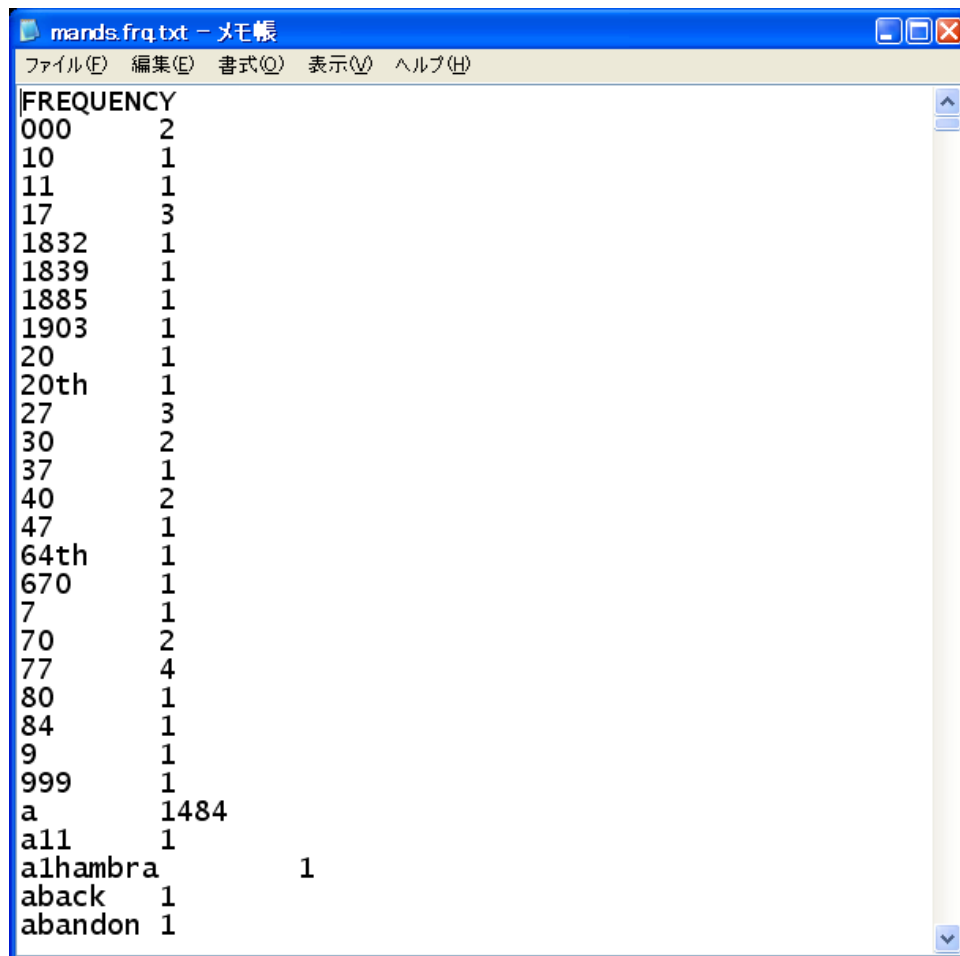
`--freq -i=ファイル名 -a`

```
d: ¥corpus>perl cat.pl --freq -i=mands.xml -a  
[ ]キーを押すと下線のコマンドが出てくる  
:  
d: ¥corpus>notepad mand.s.frq.txt [ ]キーを押す
```

* オプションの指定の順序は任意

4.11. 単語頻度 4 アルファベット順

perl cat.pl --freq -i=mands.xml -a の結果



```
mands.freq.txt - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
FREQUENCY
000      2
10       1
11       1
17       3
1832     1
1839     1
1885     1
1903     1
20       1
20th     1
27       3
30       2
37       1
40       2
47       1
64th     1
670      1
7        1
70       2
77       4
80       1
84       1
9        1
999      1
a        1484
a11      1
alhambra      1
aback     1
abandon   1
```

4.12. 単語頻度 5 数字排除

```
--freq -i=ファイル名 -a -n
```

```
d: ¥corpus>perl cat.pl --freq -i=mands.xml -a -n
```

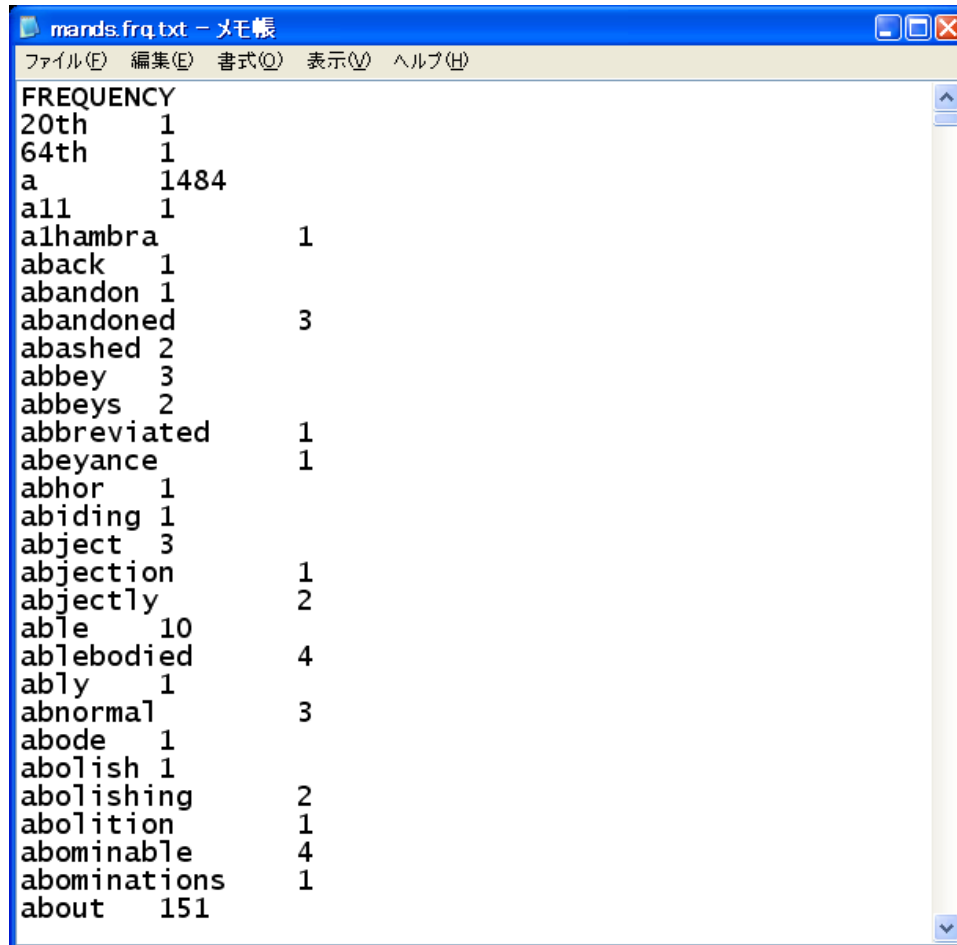
[]キーを押す

:

```
d: ¥corpus>notepad mand.s.frq.txt [ ]キーを押す
```

4.13. 単語頻度 6 数字排除

perl cat.pl --freq -i=mands.xml -a -n の結果



```
mands.freq.txt - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
FREQUENCY
20th      1
64th      1
a         1484
a11       1
alhambra      1
aback      1
abandon     1
abandoned   3
abashed    2
abbey      3
abbeys     2
abbreviated 1
abeyance   1
abhor      1
abiding    1
abject     3
abjection  1
abjectly   2
able       10
ablebodied  4
ably       1
abnormal   3
abode      1
abolish    1
abolishing  2
abolition  1
abominable  4
abominations 1
about     151
```

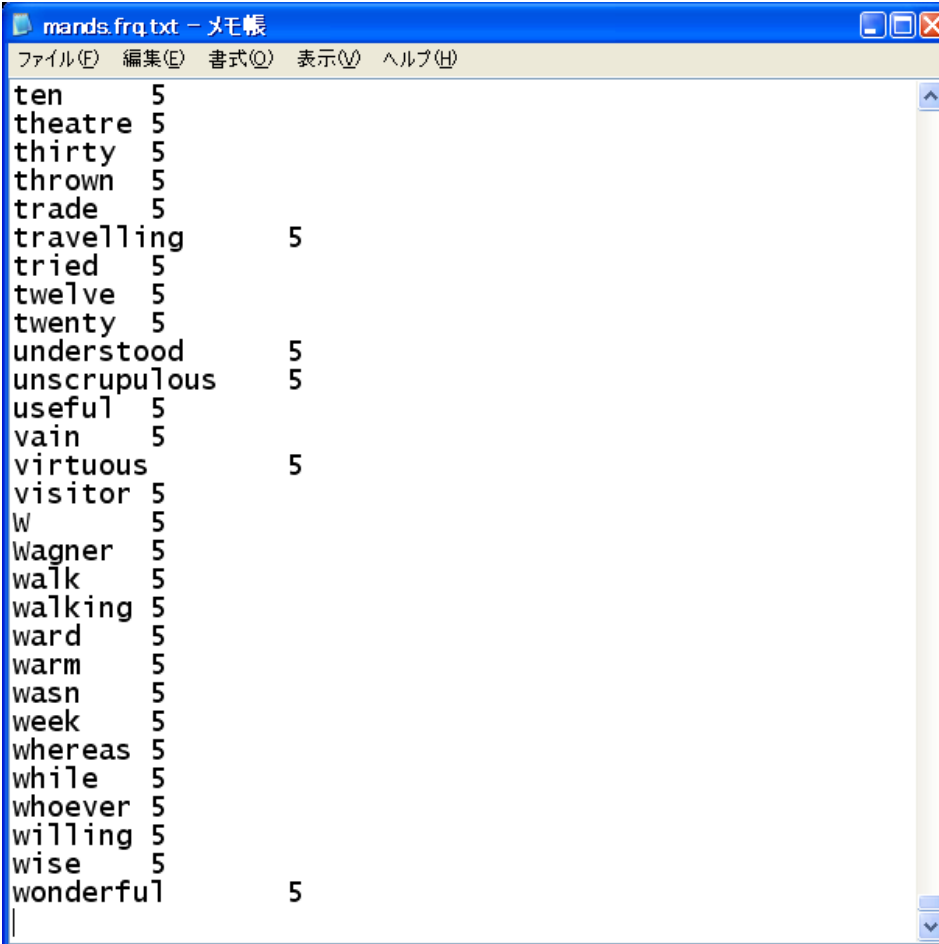
4.14. 単語頻度 7 低頻度語排除

`--freq -i=ファイル名 -n -f=5` 低頻度語排除

```
d: ¥corpus>perl cat.pl --freq -i=mands.xml -n -f=5
:
d: ¥corpus>notepad mand.frq.txt
```

4.15. 単語頻度 8 低頻度語排除

perl cat.pl --freq -i=mands.xml -n -f=5 の結果



```
mands.freq.txt - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
ten      5
theatre  5
thirty   5
thrown   5
trade    5
travelling      5
tried    5
twelve   5
twenty   5
understood      5
unscrupulous    5
useful    5
vain      5
virtuous   5
visitor   5
W         5
Wagner    5
walk      5
walking   5
ward      5
warm      5
was       5
wasn      5
week      5
whereas   5
while     5
whoever   5
willing   5
wise      5
wonderful      5
|
```

[Ctrl-End]でファイル
の末尾にジャンプ

4.16. 単語頻度 9 htmlファイル出力

`--freq -i=ファイル名 -n -f=5 -h` htmlファイル出力

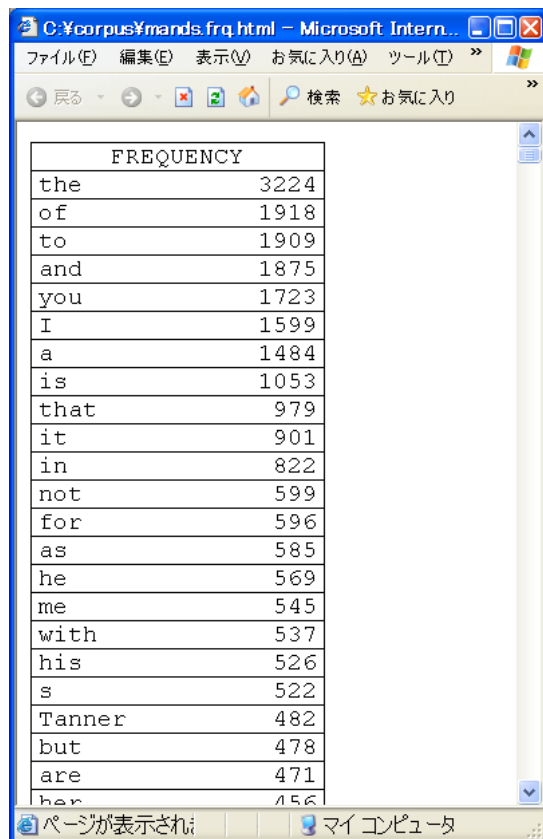
```
d: ¥corpus>perl cat.pl --freq -i=mands.xml -n -f=5 -  
h
```

```
OUTPUT: c: ¥corpus¥mands.frq.html
```

```
d: ¥corpus>mands.frq.html ブラウザの起動
```

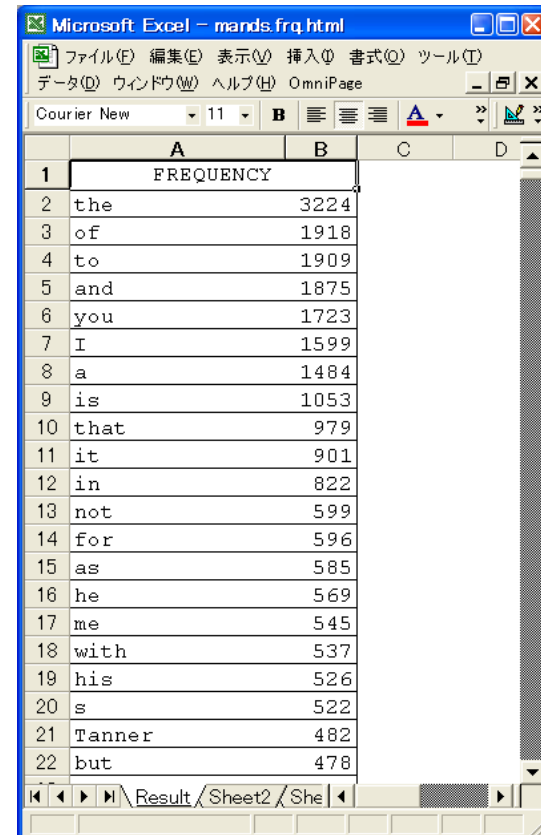
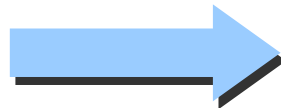
4.17. 単語頻度 10 htmlファイル出力

KWICコンコーダンス以外のhtmlファイルはExcel互換なので、[ファイル|Microsoft Excel ...で編集]を選ぶとExcelで編集できる



A screenshot of a Microsoft Internet Explorer window showing a web page with a table of word frequencies. The table has two columns: the word and its frequency. The words listed are: the, of, to, and, you, I, a, is, that, it, in, not, for, as, he, me, with, his, s, Tanner, but, are, her.

FREQUENCY	
the	3224
of	1918
to	1909
and	1875
you	1723
I	1599
a	1484
is	1053
that	979
it	901
in	822
not	599
for	596
as	585
he	569
me	545
with	537
his	526
s	522
Tanner	482
but	478
are	471
her	456



A screenshot of a Microsoft Excel window showing the same frequency list data from the browser. The data is organized into a table with columns A and B. The words are in column A and the frequencies are in column B.

	A	B	C	D
1	FREQUENCY			
2	the	3224		
3	of	1918		
4	to	1909		
5	and	1875		
6	you	1723		
7	I	1599		
8	a	1484		
9	is	1053		
10	that	979		
11	it	901		
12	in	822		
13	not	599		
14	for	596		
15	as	585		
16	he	569		
17	me	545		
18	with	537		
19	his	526		
20	s	522		
21	Tanner	482		
22	but	478		

4.18. 単語頻度 11 複数ファイル集計処理

```
--freq -i=*.xml -n -f=5
```

```
d: ¥corpus>perl cat.pl --freq -i=mand.xml -n -f=5  
[ ]キーを押して、以前のコマンドを表示
```

```
d: ¥corpus>perl cat.pl --freq -i=*.xml -n -f=5  
カーソルを戻して、mand.xml を*.xml に変える
```

```
INPUT: c: ¥corpus¥pygml.xml
```

```
INPUT: c: ¥corpus¥mands.xml
```

```
OUTPUT: result.frq.txt
```

2つのファイルの読み込み

4.19. 単語頻度 12 複数ファイル個別処理

```
--freq -i=*.xml -n -f=5 -1
```

```
d: ¥corpus>perl cat.pl --freq -i=*.xml -n -f=5 -1
```

```
:
```

```
Done: c: ¥corpus¥pygml.xml => c: ¥corpus¥pygml.frq.txt
```

```
Done: c: ¥corpus¥mands.xml => c: ¥corpus¥mands.frq.txt
```

入力ファイルごとにファイル出力

-a

アルファベット順に表示。指定がないときは頻度順。

-f=<n>

表示する最低頻度を指定。

-n

数字を排除。

4.20. n-gram頻度 1 bigram

You once asked me why I did not write a Don Juan play.

文頭の単語の正規化

you once asked me why I did not write a Don Juan play.

bigramの抽出

you once, once asked, asked me, me why, why I, I did ...

--ngram -z=2 -i=ファイル名

```
d: ¥corpus>perl cat.pl --ngram -z=2 -i=mands.xml
```

```
:
```

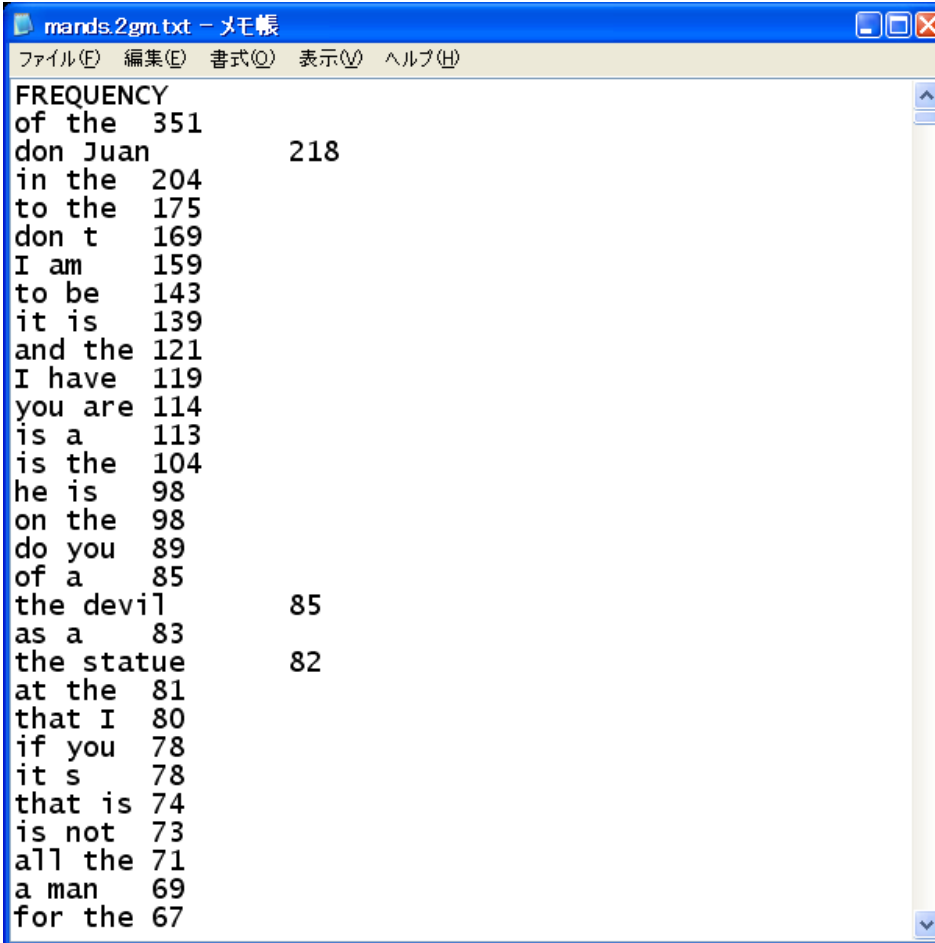
```
INPUT: c: ¥corpus¥mands.xml
```

```
OUTPUT: c: ¥corpus¥mands.2gm.txt
```

```
d: ¥corpus>notepad mand.2gm.txt
```

4.21. n-gram頻度 2 bigram

perl cat.pl --ngram -z=2 -i=mands.xml の結果



```
mands.2gm.txt - メモ帳
ファイル(E) 編集(E) 書式(O) 表示(V) ヘルプ(H)
FREQUENCY
of the 351
don Juan 218
in the 204
to the 175
don t 169
I am 159
to be 143
it is 139
and the 121
I have 119
you are 114
is a 113
is the 104
he is 98
on the 98
do you 89
of a 85
the devil 85
as a 83
the statue 82
at the 81
that I 80
if you 78
it s 78
that is 74
is not 73
all the 71
a man 69
for the 67
```

4.22. n-gram頻度 3 trigram

--ngram -z=3 -i=ファイル名 trigram

```
d: ¥corpus>perl cat.pl --ngram -z=3 -i=mands.xml
:
INPUT: c: ¥corpus¥mands.xml
OUTPUT: c: ¥corpus¥mands.3gm.txt
d: ¥corpus>notepad mand.3gm.txt
```

-a

アルファベット順に表示。指定がないときは頻度順。

-f=<n>

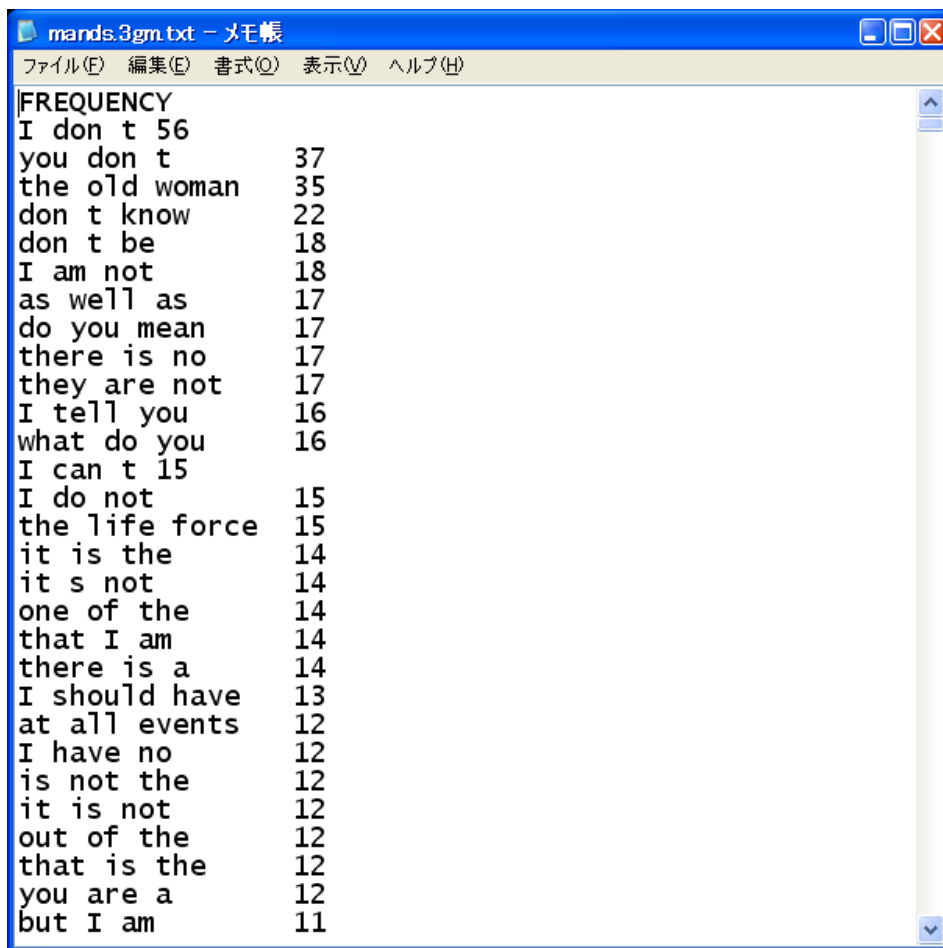
表示する最低頻度を指定。指定がないときはすべて表示。

-z=<n>

n-gramのサイズ。2～10まで指定可能。

4.23. n-gram頻度 4 trigram

perl cat.pl --ngram -z=3 -i=mands.xml の結果



```
mands.3gm.txt - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
FREQUENCY
I don t 56
you don t      37
the old woman  35
don t know    22
don t be      18
I am not      18
as well as    17
do you mean   17
there is no   17
they are not  17
I tell you    16
what do you   16
I can t 15
I do not      15
the life force 15
it is the    14
it s not     14
one of the   14
that I am    14
there is a   14
I should have 13
at all events 12
I have no    12
is not the   12
it is not    12
out of the   12
that is the  12
you are a    12
but I am     11
```

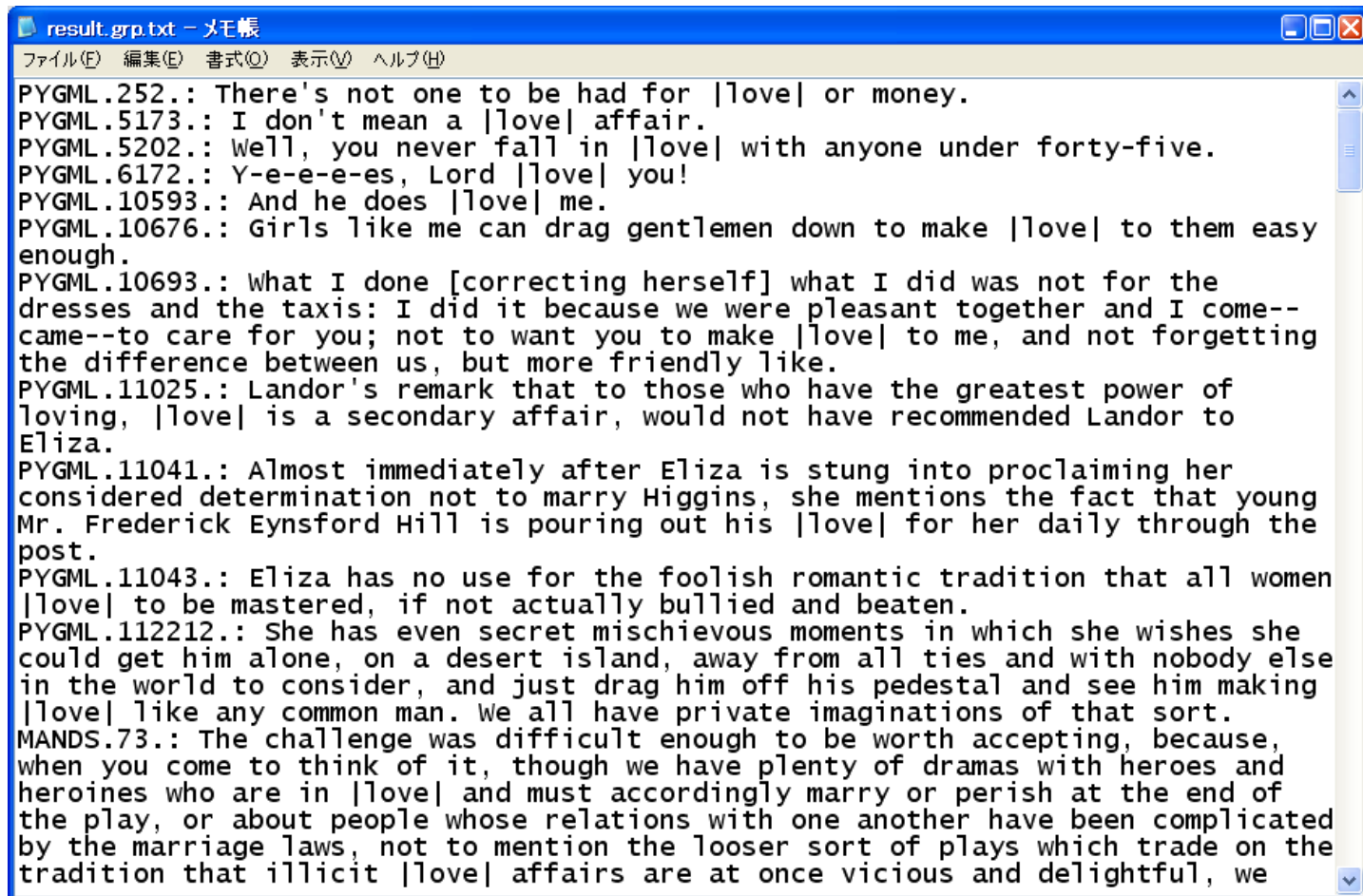
4.24. grep 1 センテンス単位

--grep -k=キーワード -i=ファイル名

```
d: ¥corpus>perl cat.pl --grep -k=love -i=*.xml
:
INPUT: c: ¥corpus¥pygml.xml
INPUT: c: ¥corpus¥mands.xml
OUTPUT: result.grp.txt
d: ¥corpus>notepad result.grp.txt
```

4.25. grep 2 センテンス単位

perl cat.pl --grep -k=love -i=*.xml の結果



```
result.grp.txt - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
PYGML.252.: There's not one to be had for |love| or money.
PYGML.5173.: I don't mean a |love| affair.
PYGML.5202.: Well, you never fall in |love| with anyone under forty-five.
PYGML.6172.: Y-e-e-e-es, Lord |love| you!
PYGML.10593.: And he does |love| me.
PYGML.10676.: Girls like me can drag gentlemen down to make |love| to them easy
enough.
PYGML.10693.: What I done [correcting herself] what I did was not for the
dresses and the taxis: I did it because we were pleasant together and I come--
came--to care for you; not to want you to make |love| to me, and not forgetting
the difference between us, but more friendly like.
PYGML.11025.: Landor's remark that to those who have the greatest power of
loving, |love| is a secondary affair, would not have recommended Landor to
Eliza.
PYGML.11041.: Almost immediately after Eliza is stung into proclaiming her
considered determination not to marry Higgins, she mentions the fact that young
Mr. Frederick Eynsford Hill is pouring out his |love| for her daily through the
post.
PYGML.11043.: Eliza has no use for the foolish romantic tradition that all women
|love| to be mastered, if not actually bullied and beaten.
PYGML.112212.: She has even secret mischievous moments in which she wishes she
could get him alone, on a desert island, away from all ties and with nobody else
in the world to consider, and just drag him off his pedestal and see him making
|love| like any common man. We all have private imaginations of that sort.
MANDS.73.: The challenge was difficult enough to be worth accepting, because,
when you come to think of it, though we have plenty of dramas with heroes and
heroines who are in |love| and must accordingly marry or perish at the end of
the play, or about people whose relations with one another have been complicated
by the marriage laws, not to mention the looser sort of plays which trade on the
tradition that illicit |love| affairs are at once vicious and delightful, we
```

4.26. grep 3 パラグラフ単位

`--grep -k=キーワード -i=ファイル名 -p`

```
d: ¥corpus>perl cat.pl --grep -k=love -i=*.xml -p
```

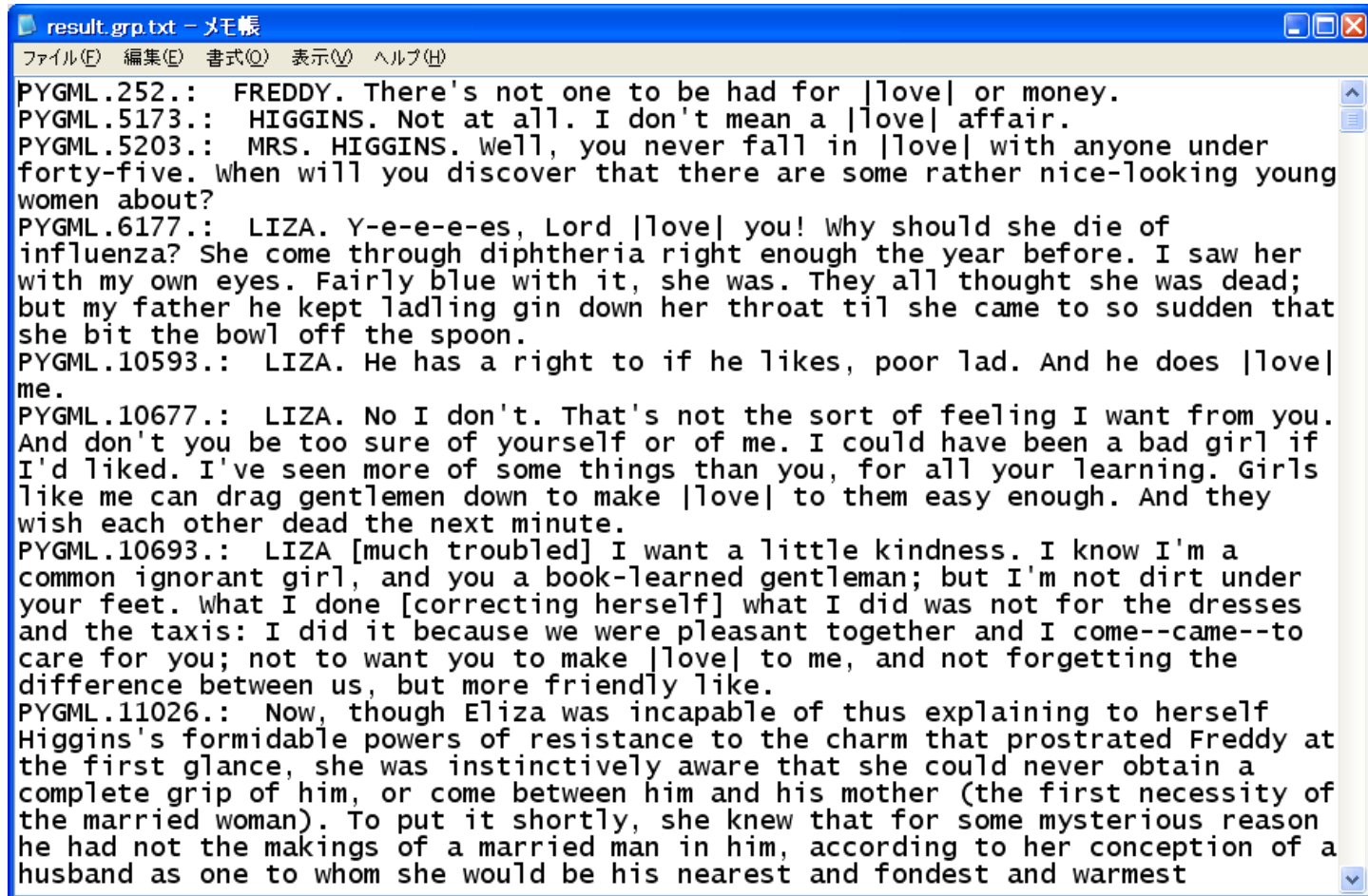
```
:
```

```
OUTPUT: result.grp.txt
```

```
d: ¥corpus>notepad result.grp.txt
```

4.27. grep 4 パラグラフ単位

perl cat.pl --grep -k=love -i=*.xml -p の結果

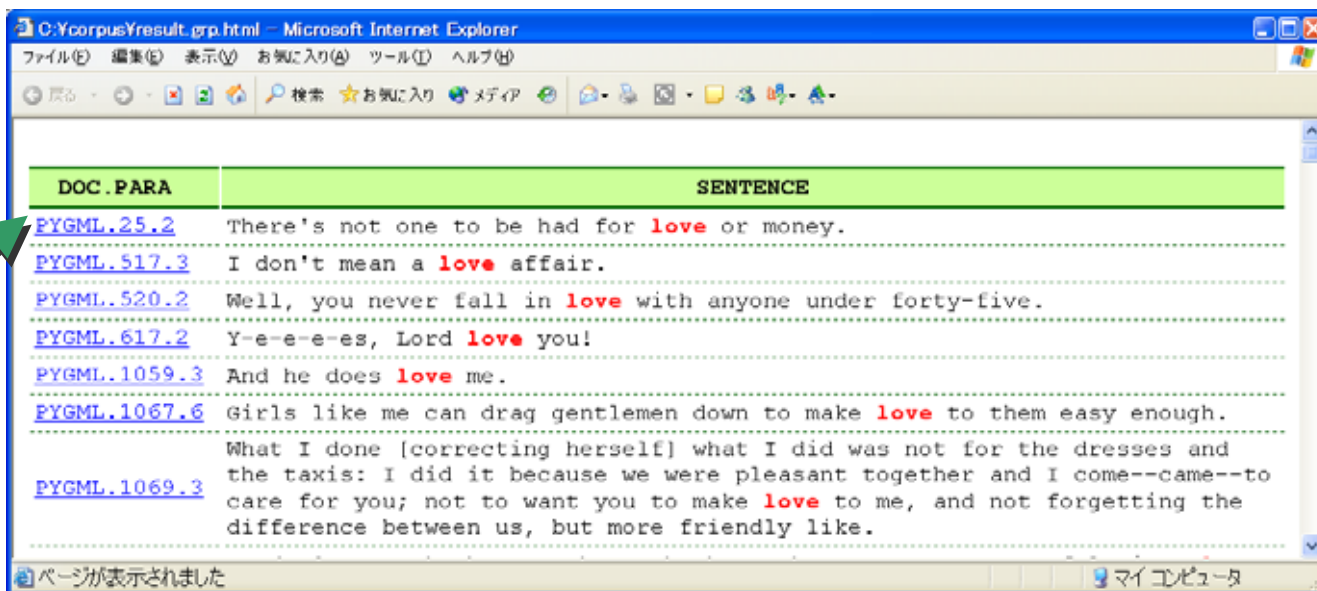


```
result.grp.txt - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
PYGML.252.: FREDDY. There's not one to be had for |love| or money.
PYGML.5173.: HIGGINS. Not at all. I don't mean a |love| affair.
PYGML.5203.: MRS. HIGGINS. Well, you never fall in |love| with anyone under
forty-five. When will you discover that there are some rather nice-looking young
women about?
PYGML.6177.: LIZA. Y-e-e-e-es, Lord |love| you! why should she die of
influenza? She come through diphtheria right enough the year before. I saw her
with my own eyes. Fairly blue with it, she was. They all thought she was dead;
but my father he kept ladling gin down her throat til she came to so sudden that
she bit the bowl off the spoon.
PYGML.10593.: LIZA. He has a right to if he likes, poor lad. And he does |love|
me.
PYGML.10677.: LIZA. No I don't. That's not the sort of feeling I want from you.
And don't you be too sure of yourself or of me. I could have been a bad girl if
I'd liked. I've seen more of some things than you, for all your learning. Girls
like me can drag gentlemen down to make |love| to them easy enough. And they
wish each other dead the next minute.
PYGML.10693.: LIZA [much troubled] I want a little kindness. I know I'm a
common ignorant girl, and you a book-learned gentleman; but I'm not dirt under
your feet. What I done [correcting herself] what I did was not for the dresses
and the taxis: I did it because we were pleasant together and I come--came--to
care for you; not to want you to make |love| to me, and not forgetting the
difference between us, but more friendly like.
PYGML.11026.: Now, though Eliza was incapable of thus explaining to herself
Higgins's formidable powers of resistance to the charm that prostrated Freddy at
the first glance, she was instinctively aware that she could never obtain a
complete grip of him, or come between him and his mother (the first necessity of
the married woman). To put it shortly, she knew that for some mysterious reason
he had not the makings of a married man in him, according to her conception of a
husband as one to whom she would be his nearest and fondest and warmest
```

4.28. grep 5 htmlファイル出力

`--grep -k=キーワード -i=ファイル名`

```
d: ¥corpus>perl cat.pl --grep -k=love -i=*.xml -h
:
d: ¥corpus>result.grp.html
```



DOC.PARA	SENTENCE
PYGML.25.2	There's not one to be had for love or money.
PYGML.517.3	I don't mean a love affair.
PYGML.520.2	Well, you never fall in love with anyone under forty-five.
PYGML.617.2	Y-e-e-e-es, Lord love you!
PYGML.1059.3	And he does love me.
PYGML.1067.6	Girls like me can drag gentlemen down to make love to them easy enough.
PYGML.1069.3	What I done [correcting herself] what I did was not for the dresses and the taxis: I did it because we were pleasant together and I come--came--to care for you; not to want you to make love to me, and not forgetting the difference between us, but more friendly like.

htmlファイル
にジャンプ

4.29. grep 6 キーワードを指定するときの注意

- `-k="good enough"`

2語以上の連続したパターンを指定するときはスペースを含むのでダブルクォートで囲む。

- `-k="love?(|d|s|ing)"`

正規表現の選択を表す|を使うときは、パイプと解釈されないようにダブルクォートで囲む。

-k=<s>

キーワードを指定する。

-p

パラグラフ単位で検索する。指定のないときは
センテンス単位。

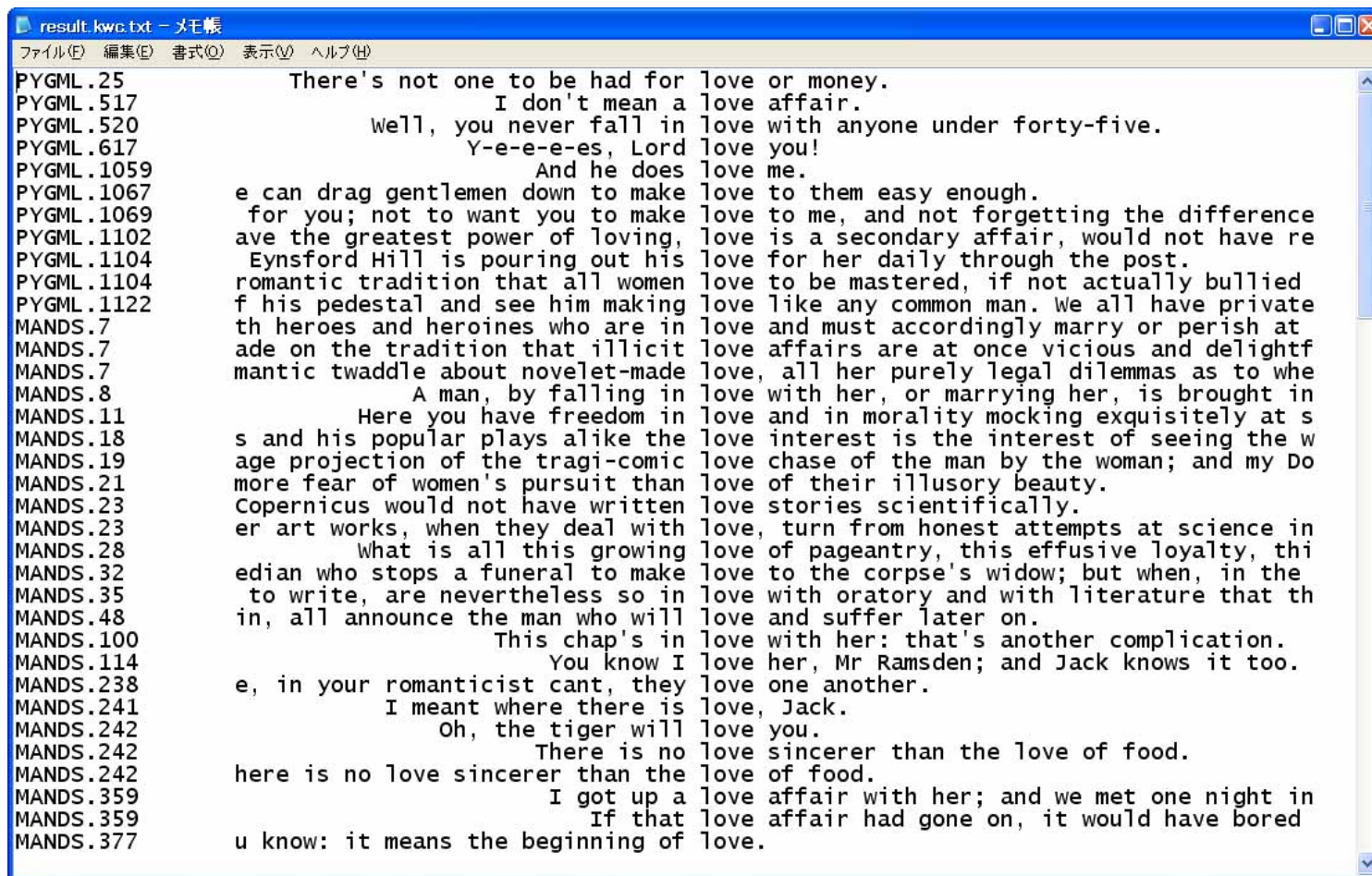
4.30. KWICコンコーダンス 1 センテンス単位

--kwic -k=キーワード -i=ファイル名

```
d: ¥corpus>perl cat.pl --kwi c -k=l ove -i =*. xml
:
INPUT: c: ¥corpus¥pygml . xml
INPUT: c: ¥corpus¥mands. xml
OUTPUT: resul t. kwc. txt
d: ¥corpus>notepad resul t. kwc. txt
```

4.31. KWICコンコーダンス 2 センテンス単位

perl cat.pl --kwic -k=love -i=* .xml の結果



The screenshot shows a text editor window titled "result.kwic.txt - メモ帳". The window contains the following KWIC output:

```

PYGML.25          There's not one to be had for love or money.
PYGML.517          I don't mean a love affair.
PYGML.520          Well, you never fall in love with anyone under forty-five.
PYGML.617          Y-e-e-e-es, Lord love you!
PYGML.1059         And he does love me.
PYGML.1067         e can drag gentlemen down to make love to them easy enough.
PYGML.1069         for you; not to want you to make love to me, and not forgetting the difference
PYGML.1102         ave the greatest power of loving, love is a secondary affair, would not have re
PYGML.1104         Eynsford Hill is pouring out his love for her daily through the post.
PYGML.1104         romantic tradition that all women love to be mastered, if not actually bullied
PYGML.1122         f his pedestal and see him making love like any common man. We all have private
MANDS.7            th heroes and heroines who are in love and must accordingly marry or perish at
MANDS.7            ade on the tradition that illicit love affairs are at once vicious and delightf
MANDS.7            mantic twaddle about novelet-made love, all her purely legal dilemmas as to whe
MANDS.8            A man, by falling in love with her, or marrying her, is brought in
MANDS.11           s Here you have freedom in love and in morality mocking exquisitely at s
MANDS.18           s and his popular plays alike the love interest is the interest of seeing the w
MANDS.19           age projection of the tragi-comic love chase of the man by the woman; and my Do
MANDS.21           more fear of women's pursuit than love of their illusory beauty.
MANDS.23           Copernicus would not have written love stories scientifically.
MANDS.23           er art works, when they deal with love, turn from honest attempts at science in
MANDS.28           what is all this growing love of pageantry, this effusive loyalty, thi
MANDS.32           edian who stops a funeral to make love to the corpse's widow; but when, in the
MANDS.35           to write, are nevertheless so in love with oratory and with literature that th
MANDS.48           in, all announce the man who will love and suffer later on.
MANDS.100          This chap's in love with her: that's another complication.
MANDS.114          You know I love her, Mr Ramsden; and Jack knows it too.
MANDS.238          e, in your romanticist cant, they love one another.
MANDS.241          I meant where there is love, Jack.
MANDS.242          Oh, the tiger will love you.
MANDS.242          There is no love sincerer than the love of food.
MANDS.242          here is no love sincerer than the love of food.
MANDS.359          I got up a love affair with her; and we met one night in
MANDS.359          If that love affair had gone on, it would have bored
MANDS.377          u know: it means the beginning of love.
```

4.32. KWICコンコーダンス 3 パラグラフ単位

`--kwic -k=キーワード -i=ファイル名 -p`

```
d: ¥corpus>perl cat.pl --kwic -k=love -i=*.xml -p
```

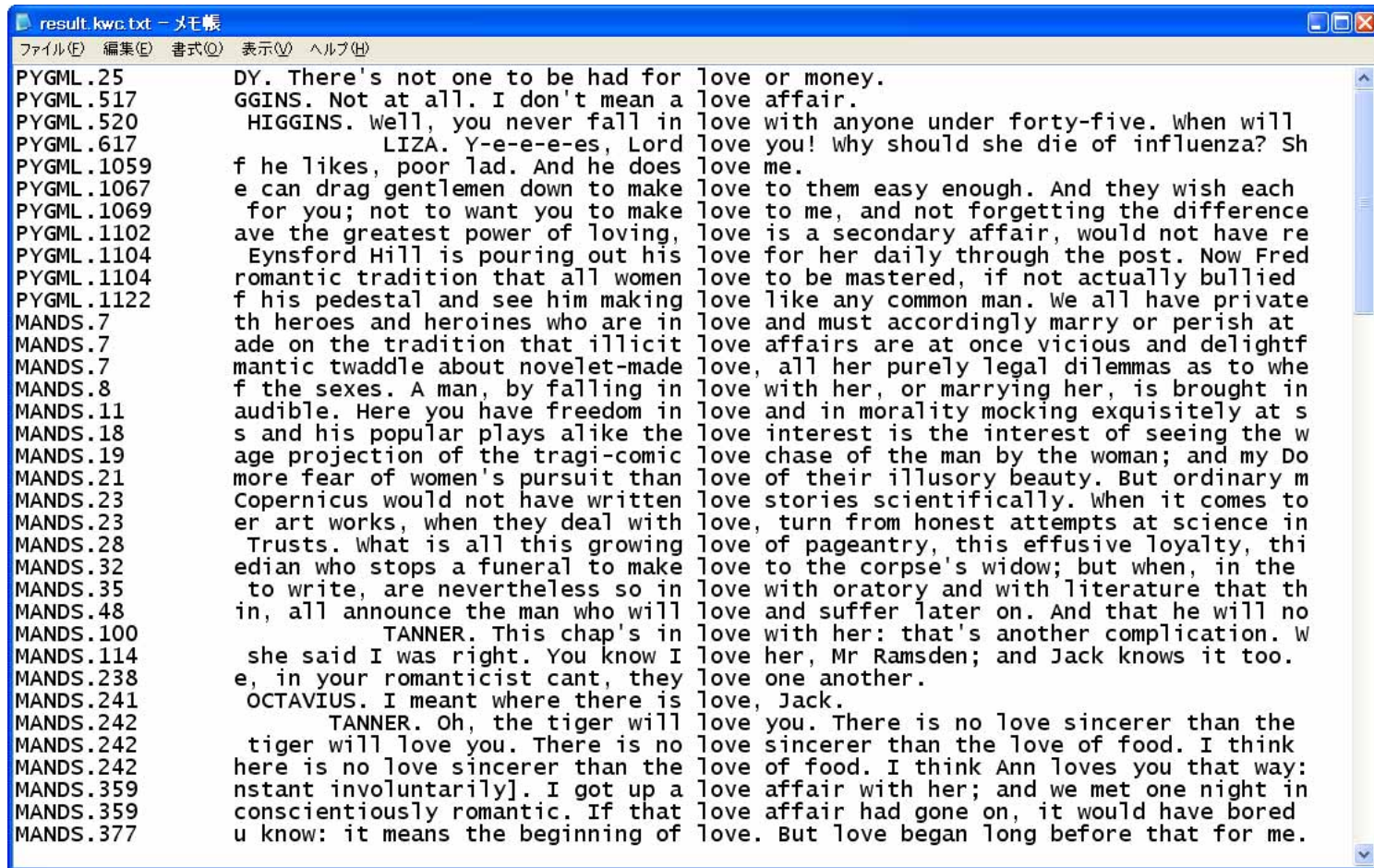
```
:
```

```
OUTPUT: result.kwic.txt
```

```
d: ¥corpus>notepad result.kwic.txt
```

4.33. KWICコンコーダンス 4 パラグラフ単位

perl cat.pl --kwic -k=love -i=* .xml -p の結果



The screenshot shows a text editor window titled "result.kwc.txt - メモ帳". The window contains a list of KWIC concordance results for the word "love". Each line consists of a line number, a snippet of text containing the word "love", and the full sentence it belongs to. The results are as follows:

```
PYGML.25          DY. There's not one to be had for love or money.
PYGML.517        GGINS. Not at all. I don't mean a love affair.
PYGML.520        HIGGINS. Well, you never fall in love with anyone under forty-five. When will
PYGML.617        LIZA. Y-e-e-e-es, Lord love you! Why should she die of influenza? Sh
PYGML.1059       f he likes, poor lad. And he does love me.
PYGML.1067       e can drag gentlemen down to make love to them easy enough. And they wish each
PYGML.1069       for you; not to want you to make love to me, and not forgetting the difference
PYGML.1102       ave the greatest power of loving, love is a secondary affair, would not have re
PYGML.1104       Eynsford Hill is pouring out his love for her daily through the post. Now Fred
PYGML.1104       romantic tradition that all women love to be mastered, if not actually bullied
PYGML.1122       f his pedestal and see him making love like any common man. We all have private
MANDS.7          th heroes and heroines who are in love and must accordingly marry or perish at
MANDS.7          ade on the tradition that illicit love affairs are at once vicious and delightf
MANDS.7          mantic twaddle about novelet-made love, all her purely legal dilemmas as to whe
MANDS.8          f the sexes. A man, by falling in love with her, or marrying her, is brought in
MANDS.11         audible. Here you have freedom in love and in morality mocking exquisitely at s
MANDS.18         s and his popular plays alike the love interest is the interest of seeing the w
MANDS.19         age projection of the tragi-comic love chase of the man by the woman; and my Do
MANDS.21         more fear of women's pursuit than love of their illusory beauty. But ordinary m
MANDS.23         Copernicus would not have written love stories scientifically. When it comes to
MANDS.23         er art works, when they deal with love, turn from honest attempts at science in
MANDS.28         Trusts. What is all this growing love of pageantry, this effusive loyalty, thi
MANDS.32         edian who stops a funeral to make love to the corpse's widow; but when, in the
MANDS.35         to write, are nevertheless so in love with oratory and with literature that th
MANDS.48         in, all announce the man who will love and suffer later on. And that he will no
MANDS.100        TANNER. This chap's in love with her: that's another complication. W
MANDS.114        she said I was right. You know I love her, Mr Ramsden; and Jack knows it too.
MANDS.238        e, in your romanticist cant, they love one another.
MANDS.241        OCTAVIUS. I meant where there is love, Jack.
MANDS.242        TANNER. Oh, the tiger will love you. There is no love sincerer than the
MANDS.242        tiger will love you. There is no love sincerer than the love of food. I think
MANDS.242        here is no love sincerer than the love of food. I think Ann loves you that way:
MANDS.359        nstant involuntarily]. I got up a love affair with her; and we met one night in
MANDS.359        conscientiously romantic. If that love affair had gone on, it would have bored
MANDS.377        u know: it means the beginning of love. But love began long before that for me.
```

4.34. KWICコンコーダンス 5 桁数指定

--kwic -k=キーワード -i=ファイル名 -w=桁数 桁数指定

```
d: ¥corpus>perl cat.pl --kwic -k=love -i=*.xml -w=90
:
OUTPUT: result.kwc.txt
```

--kwic -k=キーワード -i=ファイル名 -h htmlファイル出力

```
d: ¥corpus>perl cat.pl --kwic -k=love -i=*.xml -h
:
OUTPUT: result.kwc.html
d: ¥corpus>result.kwc.html
```

4.35. KWICコンコーダンス 6 htmlファイル出力

ソートキー2つで左右5語の並べ替え

DOC.PARA	CONCORDANCE
PYGML.25.2	There's not one to be had for love or money.
PYGML.517.3	I don't mean a love affair.
PYGML.520.2	Well, you never fall in love with anyone under forty-five.
PYGML.617.2	Y-e-e-e-es, Lord love you!
PYGML.1059.3	And he does love me.
PYGML.1067.6	e can drag gentlemen down to make love to them easy enough.
PYGML.1069.3	for you; not to want you to make love to me, and not forgetting the difference
PYGML.1102.5	ave the greatest power of loving, love is a secondary affair, would not have re
PYGML.1104.1	Eynsford Hill is pouring out his love for her daily through the post.
PYGML.1104.3	romantic tradition that all women love to be mastered, if not actually bullied

ただし、実際に使えるのは数百件程度まで

4.36. KWICコンコーダンス 7

-k=<s>

キーワードを指定する。

-p

パラグラフ単位でコンコーダンスラインを作成する。指定のないときはセンテンス単位。

-w=<n>

コンコーダンスラインの桁数を指定する。40～100までの範囲で指定。指定のないときは80。

4.37. コロケーション(頻度順) 1

--col -s=数字 -k=キーワード -i=ファイル名 -h

```
d: ¥corpus>perl cat.pl --col -s=5 -k=love -i=*.xml -h
```

:

```
OUTPUT: result.col.html
```

```
d: ¥corpus>result.col.html
```

--col -s=数字 -k=キーワード -i=ファイル名 -h -f=最低頻度

```
d: ¥corpus>perl cat.pl --col -s=5 -k=love -i=*.xml -h -f=5
```

:

```
OUTPUT: result.col.html
```

4.38. コロケーション(頻度順) 2

5L ~ 5Rまえの位置ごとのcollocateが頻度順で表示される。印刷するときにはExcelを使えば縮小印刷ができる。

3L	2L	1L	NODE	1R	2R	3R
5 the	9 to	5 in	16 love 122	and	12 beauty	7 and
5 you	8 Louisa	4 I	15	with	10 her	7 the
3 is	4 nothing	4 of	12	you	7 and	6 or
3 I	3 that	4 but	5	me	6 is	5 I
3 Louisa	2 is	3 a	4	of	6 a	4 but
3 a	2 and	2 for	4	three	6 Louisa	3 it
2 am	2 cant	2 the	4	to	5 the	3 that
2 his	2 do	2 you	4	for	4 he	2 to
2 never	2 don	2 about	3	is	4 its	2 you
2 nothing	2 else	2 wake	3	affair	3 me	2 Ann
2 of	2 fall	2 not	3	her	3 romance	2 Jack
2 that	2 know	2 that	3	compact	2 tavy	2 Juan
2 these	2 not	2 with	3	him	2 with	2 Louisa
2 anarchism	1 you	2 and	2	it	2 you	2 Ramsden
2 Ann	1 alike	1 his	2	or	2 I	1 accordingly
2 Jack	1 all	1 honorable	2	will	2 Mr	1 at
2 about	1 am	1 my	2	Ann	1 all	1 beauty
1 according	1 an	1 t	2	Hector	1 another	1 before
1 all	1 are	1 they	2	Jack	1 any	1 common
1 and	1 art	1 will	2	affairs	1 anyone	1 corpse
ce 1 art	1 been	1 Lord	1	all	1 are	1 daily
ce 1 artist	1 beginning	1 as	1	as	1 but	1 disappointments
1 as	1 but	1 comic	1	beauty	1 chastity	1 don
1 be	1 by	1 does	1	because	1 everything	1 dream
1 beauty	1 can	1 emotions	1	began	1 food	1 dupe
1 begin	1 cloud	1 free	1	but	1 from	1 easy
1 bring	1 cynically	1 growing	1	chase	1 good	1 emotion
1 by	1 deal	1 have	1	ducts	1 had	1 gone
1 chap	1 difference	1 if	1	even	1 happiness	1 good
1 did	1 disappointments	1 illicit	1	has	1 in	1 have
1 does	1 ears	1 in	1	interest	1 in	1 her

4.39. コロケーション (頻度順) 3

■ スパンの指定のしかた

■ -s = 数字のみ

数字で指定した前後の共起語を計数する。

■ -s = ± 数字

+ の場合はノードの右の共起語のみ、- の場合は左の共起語のみを計数する。

-f = <n>

表示する最低頻度を指定する。

-k = <s>

キーワードを指定する。

-p

パラグラフ単位で共起語を採取する。指定のないときはセンテンス単位。

4.40. コロケーション(統計値) 1

```
--stat -s=数字 -m=統計値 -k=キーワード -i=ファイル名 -h
```

```
d: ¥corpus>perl cat.pl --stat -s=3 -m=FMT -k=love -  
i=*.xml -h -f=5
```

統計値として、単純頻度、MI スコア、Tスコアを指定

OUTPUT: result.stat.html

```
d: ¥corpus>result.stat.html
```

F	単純頻度	T	Tスコア
M	MIスコア	Z	Zスコア
3	MI3スコア	L	Log-log

4.41. コロケーション(統計値) 2

Overall Rankとは、使用した統計値(以下ではMI ScoreとT-score)のランクの高い順に重みをつけて、総合評価した数値。

25項目

Frequency	MI Score	T-score	Overall Rank
and 29	thee 6.973923	beauty 3.094042	beauty 49
you 25	beauty 5.534300	Lou 433	I 7
of 22	Lou 3094	wit 383	t 7
the 22	not 2696	the 704	r 3
I 21	mal 3496	nothing 2.306369	with 43
with 18	with 1.721783	in 2.033115	make 39
in 17	or 1.269327	is 1.845725	in 37
is 17	but 1.113457	make 1.837619	is 35
to 14	in 0.980235	and 1.813835	but 34
that 12	is 0.856356	but 1.700719	or 33
her 11	me 0.816071	me 1.366146	me 30
Louisa 10	for 0.715357	or 1.308428	and 29
a 10	her 0.705647	her 1.282978	her 26
beauty 10	and 0.592529	you 1.241264	for 25
but 10	not 0.478318	for 1.236278	you 22
for 10	you 0.411680	of 0.851282	not 19
me 10	of 0.288935	I 0.820180	of 19
not 8	my 0.288824	not 0.798142	I 15
it 7	that 0.288006	that 0.626887	my 14
nothing 7	I 0.284507	my 0.405691	that 14
thee 7	his -0.046779	his -0.073693	his 10
his 5	it -0.450244	it -0.969062	it 8
make 5	the -0.543612	to -1.736940	to 5
my 5	to -0.550129	a -1.924562	the 4
or 5	a -0.685806	the -2.146417	a 3

4.42. コロケーション(統計値) 3

-f=<n>	表示する最低頻度を指定する。
-k=<s>	キーワードを指定する。
-m=<s>	使用する統計値を指定。指定できる統計値の種類については前項(4.27.)を参照。
-p	パラグラフ単位で共起語を採取する。指定のないときはセンテンス単位。
-s=<n>	スパンを指定する。指定のしかたはコロケーション(頻度順)と同じ(4.25.)。

統計値の算出方法については、以下のURLを参考にした。

<http://homepage.mac.com/bncweb/manual/bncwebman-collocation.htm>

4.43. Perlによるコーパス処理 参考URL

- NSP (N-gram Statistics Package), Ted Pedersen
N-gramプログラム

<http://www.d.umn.edu/~tpederse/nsp.html>

- Serge Sharoff's personal page Tools for working with parallel corpora and studying contrastive semantics, Serge Sharoff

コーパス作成・分析プログラム

<http://www.comp.leeds.ac.uk/ssharoff/concordance-fr.html>

- Dan Melamed's NLP Tools

テキスト処理・統計処理プログラム

<http://www.cs.nyu.edu/~melamed/software.html>

4.44. コーパス処理 参考文献

Erik T. Ray. 2002. *Perl & XML*. オライリージャパン.

Oakes, Michael P. 1998. *Statistics for Corpus Linguistics*. Edinburgh University Press.

Manning, Christopher and Hinrich Schutze. 1999. *Foundation of Statistical Natural Language Processing*. Cambridge: MIT Press.