追手門学院大学文学部紀要　33号　1997年12月30日

パーソナルコーパス作成――整形上の諸問題

稲　木　昭　子

Constructing Procedure of a Personal Corpus

Akiko INAKI

Ⅰ．コーパスの定義

「英語コーパス研究」における「コーパス」は従来一般的に規模の大きいものをいう。このようなコーパスを大量検索し，その結果を考察することにより，これまでみえてこなかった言語事実が明らかにされ，語法研究，辞書編纂等の分野で大きな貢献がなされてきている。

Renouf（1987, p. 1）は，コーパスを，'a collection of texts, of the written or spoken word, which is stored and processed on computer for the purposes of linguistic research' と定義し，赤野（1997, p. 115）では，さらにこれに説明を加え，「言語研究のためにコンピュータで収集，蓄積された言語データの集合体のことで，具体的にはさまざまなジャンルの，単一もしくは複数の言語がテキストファイルとして電子化されたものをいう」と定義する。

そこでこの解釈を活かしつつ，ここでは一般的にいうコーパスとは異なり，特定の個別的な研究，たとえばある作家の，あるいはある作品の文体論研究などで使用されるパーソナルコーパスを作成する上での，問題点を述べてみたいと思う。量的には比べものにならないほど小さいものであるので，個人でも短時間で構築可能である。構築過程としては，①特定の作品を電子テキスト化する。②このテキストファイルを，言語処理プログラムが有効であるような形にエディターで整形する，の2段階が考えられる。

①の段階には主として4通りの方法が考えられる。

①-1. キーボードで入力する。

①-2. オリジナルテキストをスキャナーで読み取り，OCR（Optical Character Recognition）でテキストデータ化し，のちオリジナルとの照合をする。

①-3. インターネットからダウンロードし，テキストファイル化する。

①-4. CD-ROM からとる。

キーボードで入力する方法は，時間がかかり，また正確性という点からも問題が残る。OCR の場合，現在ではかなり認識率があがっているものの，依然として，もとのテキストの状態いかんによっては，'l' を 'I' あるいは 'i' と認識したり，'e' を 'c' と，'m' を 'ni' と認識する間違いがよく見受けられるので，修正は不可欠の操作になる。インターネット，CD-ROM を利用するやり方は，テキストファイル作成の時間的短縮，正確性という点からもかなり有効である。このようにしてとられたいずれの場合も，ファイルが整形されることにより，検索プログラムがかけられる状態になる。ここでは 2 および 3 の電子テキスト化の際の注意すべき点を先ず述べ，次に整形の段階においてでてくる諸問題を考えることにする。

Ⅱ. 電子テキスト化

具体的な作品として，Agatha Christie と Lewis Carroll のものをとりあげる。前者の作品として，インターネット（ソフトは，Netscape Navigator Gold 3 を使用）による検索では，*The Mysterious Affair at Styles*（以下 *Styles* と略記する）1 つが見つかった。この作品は著者の最初の作品で，名探偵エルキュール・ポワロ誕生の記念すべき作品である。これを次のアドレスからダウンロードすることにする。

http://www.columbia.edu/acis/bartleby/christie

もう一つの作品として，*Witness for the Prosecution*（以下 *Witness* と略記する）を取り上げる。この作品は，批評家，および作者自身も最高傑作（the best play）と認めている戯曲であるが，戯曲ということで，特に整形上問題になることが生じる。この作品は OCR で認識し，英文のチェックをかけて，電子テキスト化したものである。

Lewis Carroll の作品例として，*Alice's Adventures in Wonderland*（以下 *Alice* と略記する）と，*Through the Looking-Glass and What Alice Found There*（以下 *Through* と略記する）をとりあげる。これらの作品はインターネット上で数多く提供されている。

http://www-cgi.cs.cmu.edu/cgi-bin/book/authorrrstart?C

しかし，インターネットからダウンロードする際に共通して注意しなければいけない点は，電子化されたファイルがオリジナルのテキストに忠実なもの，正確なものかどうかということである。次は，*Through* の 7 章をダウンロードした html ファイルの一部である。

> 'Not at all,' said the King. 'He's an Anglo-Saxon Messenger—and those are Anglo-Saxon attitudes. He only does them when he's happy. His name ia Haigha.'(He pronounced it so as to rhyme with 'mayor.'

このファイルは，'s' を 'a' と入力している等，おそらく手作業で打ち込みをしたものと思われる。さらに closing parenthesis が忘れられている。この他にも closing quotation mark がないことが多い。このような場合，結果的には検索プログラムをかける時に問題が生じることになるので，ダウンロードには注意を要する。

ダウンロードに関する注意点を次にまとめることにする。

① インターネットからダウンロードする時に，画面を｛オプション――文書の文字コードセット――欧米｝に設定すると文字化けしない。これを｛ファイル――名前を付けて保存｝で，一旦フロッピーディスクにおとす。.html の拡張子がつく。のちに行う整形の途中で，文書ソース確認のため，度々html ファイルにもどることがあるので，ディスクにおとしておいたほうがよい。

② ①をエディター（WZ Editor version 3.0）で開く。これを｛書式――画面モード――テキストモード｝｛編集――全てを選択｝｛書式――段落書式――タグを削除｝でタグを削除する。タグを削除したものを，テキストファイルとして保存する。

この点に関しては，html ファイルを Netscape Navigator Gold で開いて（文書の文字コードを確認），｛編集――全てを選択｝｛ファイル――として保存｝で，テキストファイルとして保存することも可能である。ところが *Styles* の場合，タグ以外に文字でもって，たとえば（ ）（&# 151 ;）（&agrave ;）等で文書ソースが表示される。

```
〈html〉
〈title〉 Christie, Agatha.  1920.  The Mysterious Affair at Styles: Chapter 8: Fresh
Suspicions. 〈/title〉
〈body bgcolor="#ffffff" text="#000020"
LINK="#000050" VLINK="#000050" ALINK="#000050"〉

〈center〉
〈b〉〈font color="#0000FF"〉〈font size="+2"〉8
〈br〉 Fresh Suspicions 〈/font〉〈/font〉〈/b〉
〈/center〉

〈p〉
〈table cellspacing=1 cellpadding=1〉
〈tr〉〈td〉 T 〈font size="-1"〉 HERE 〈/font〉 was a moment's stupefied silence.  Japp, who
```

was the least surprised of any of us, was the first to speak. 〈/td〉〈td valign＝top〉〈font size＝“−1”〉〈a name＝“1”〉   ; 〈/a〉〈/font〉〈/td〉〈/tr〉

〈tr〉〈td〉   ;   ; “My word,” he cried, “you’re the goods! And no mistake, Mr. Poirot! These witnesses of yours are all right, I suppose?” 〈/td〉〈td valign＝top〉〈font size＝“−1”〉〈a name＝“2”〉   ; 〈/a〉〈/font〉〈/td〉〈/tr〉

〈tr〉〈td〉   ;   ; 〈i〉 “Voil&agrave ; ! 〈/i〉 I have prepared a list of them&#151 ; names and addresses. You must see them, of course. But you will find it all right.” 〈/td〉〈td valign＝top〉〈font size＝“−1”〉〈a name＝“3”〉   ; 〈/a〉〈/font〉〈/td〉〈/tr〉

〈tr〉〈td〉   ;   ; “I’m sure of that.” Japp lowered his voice. “I’m much obliged to you. A pretty mare’s nest arresting him would have been. “ He turned to Inglethorp. “But, if you’ll excuse me, sir, why couldn’t you say all this at the inquest?” 〈/td〉〈td valign＝top〉〈font size＝“−1”〉〈a name＝“4”〉   ; 〈/a〉〈/font〉〈/td〉〈/tr〉

〈tr〉〈td〉   ;   ; “I will tell you why,” interrupted Poirot. “There was a certain rumour&#151 ; &#151 ;” 〈/td〉〈td valign＝top〉〈font size＝“−1”〉〈a name＝“5”〉〈i〉 5 〈/i〉〈/a〉〈/font〉〈/td〉〈/tr〉

そこでこのファイルを Netscape Navigator Gold 上で，テキストファイルにすると，（ ;）が（・ ）に，（&# 151 ;）が次のアルファベット文字一文字とともに，日本語化けする。

8

Fresh Suspicions

THERE was a moment’s stupefied silence. Japp, who was the least ・
surprised of any of us, was the first to speak.
・・“My word,” he cried, “you’re the goods! And no mistake, Mr. Poirot! ・
These witnesses of yours are all right, I suppose?”
・・“Voil・ I have prepared a list of them溶ames and addresses. You must ・
see them, of course. But you will find it all right. ”

・・"I'm sure of that." Japp lowered his voice. "I'm much obliged to ・
you. A pretty mare's nest arresting him would have been." He turned to
Inglethorp. "But, if you'll excuse me, sir, why couldn't you say all
this at the inquest?"
・・"I will tell you why," interrupted Poirot. "There was a certain 5
umour覧"

従ってこの時点から，コーパス用に整形するよりも，エディター上で，テキストファイルに変換して，コーパステキストを作成した方が都合がよい。

Ⅲ．整形の諸問題

各電子化されたテキストを，エディターを使い，整形を試みることにする。全般的な注意点をまずあげる。

① テキストの加工上の記号としては，MicroConcord の MCA, MCB のマニュアルが参考になる。[1]（たとえば，〈 〉，[ ]，{ }，# などの記号の使い方）

② タイトルは，テキストの先頭にもってきて，〈title〉と〈/title〉で括る。検索プログラムを使用するときは，検索テキストのパラメータで，たとえば，〈*〉を'ignore'と指定しておけば，検索対象にはならない。

③ テキストは全て左寄せにする。

④ WZ Editor の正規表現，メタキャラクターをうまく使用する。

⑤ 改行記号は，MicroConcord, WordSmith 等の検索プログラムの場合，問題はない。[2] しかし段落などを問題にする検索のときには重要になるので，このときには段落単位で改行記号をつけるように整形する。

⑥ スペル・チェック，行末ハイフンの処理等，個別的な単語関係の整形を早い段階で行う。

④に関連して，WZ Editor version 3.0 の {ヘルプ} 機能の，「正規表現――使用可能なメタキャラ」の説明事項を以下に引用する。

ファイル名を指定する部分では「*」など任意の文字列を表す記号を用いることができますが，
これは MS-DOS がサポートする『ワイルドカード』という表現方法です。
これに対して，文字列検索で使われる表現方法が『正規表現』です。正規表現検索では，次の
拡張表現が使用可能になります。

^	パターンの先頭にあるときのみ，行の先頭を表す。
$	パターンの末尾にあるときのみ，行の末尾を表す。
.	改行を除いた任意の 1 文字。（全角も 1 文字）
[……]	[ ] 内に含まれる任意の 1 文字を表す。
[^……]	[ ] 内に含まれない任意の 1 文字を表す。
*	直前文字の 0 回以上のくり返し
+	直前文字の 1 回以上のくり返し
?	直全文字が 0 個または 1 個であることを表す。
¥（……¥）	くくられた範囲をタグに記憶する。（最大 9 個）
¥N	タグに記憶した N 番目の文字列を表す。
|	和演算子（A または B）
&	積演算子（A かつ B）
¥c	正規表現 c の意味を打ち消し，文字 c 自身を表す。 「.」は「¥.」，「+」は「¥+」，そして「¥」自身は「¥¥」で表す。

正規表現の「^」や「$」はそれぞれ行頭や行末であることを表す記号であり，改行を表す文字ではありません。正規表現は 1 つの行の中での文字列パターンを特定する表現ですので，改行をまたぐ検索はできません。正規表現検索では改行を表す「¥n」は使えません。

以下個別に整形上の問題を取り上げる。

III–A *Styles* の場合

① html ファイルを WZ Editor で開く。

```
〈html〉
〈title〉 Christie, Agatha.  1920.  The Mysterious Affair at Styles: Chapter 1: I Go to
Styles. 〈/title〉
〈body bgcolor="#ffffff" text="#000020"
LINK="#000050" VLINK="#000050" ALINK="#000050"〉
```

〈center〉

〈b〉〈font color＝“#0000FF”〉〈font size＝“＋2”〉 1

〈br〉 I Go to Styles 〈/font〉〈/font〉〈/b〉

〈/center〉

〈p〉

〈table cellspacing＝1 cellpadding＝1〉

〈tr〉〈td〉 T 〈font size＝“−1”〉 HE 〈/font〉 intense interest aroused in the public by what was known at the time as “The Styles Case” has now somewhat subsided. Nevertheless, in view of the world-wide notoriety which attended it, I have been asked, both by my friend Poirot and the family themselves, to write an account of the whole story. This, we trust, will effectually silence the sensational rumours which still persist. 〈/td〉〈td valign＝top〉〈font size＝“−1”〉〈a name＝“1”〉   ; 〈/a〉〈/font〉〈/td〉〈/tr〉

〈tr〉〈td〉   ;   ; I will therefore briefly set down the circumstances which led to my being connected with the affair. 〈/td〉〈td valign＝top〉〈font size＝“−1”〉〈a name＝“2”〉   ; 〈/a〉〈/font〉〈/td〉〈/tr〉

〈tr〉〈td〉   ;   ; I had been invalided home from the Front ; and, after spending some months in a rather depressing Convalescent Home, was given a month’s sick leave. Having no near relations or friends, I was trying to make up my mind what to do, when I ran across John Cavendish. I had seen very little of him for some years. Indeed, I had never known him particularly well. He was a good fifteen years my senior, for one thing, though he hardly looked his forty-five years. As a boy, though, I had often stayed at Styles, his mother’s place in Essex. 〈/td〉〈td valign＝top〉〈font size＝“−1”〉〈a name＝“3”〉   ; 〈/a〉〈/font〉〈/td〉〈/tr〉

〈tr〉〈td〉   We had a good yarn about old times, and it ended in his inviting me down to Styles to spend my leave there. 〈/td〉〈td valign＝top〉〈font size＝“−1”〉〈a name＝“4”〉 ;〈/a〉〈/font〉〈/td〉〈/tr〉

〈tr〉〈td〉   ;   ; “The mater will be delighted to see you again&#151 ; after all those years,” he added. 〈/td〉〈td valign＝top〉〈font size＝“−1”〉〈a name＝“5”〉〈i〉 5 〈/i〉〈/a〉

〈/font〉〈/td〉〈/tr〉

② これを {表示——テキスト・モード} {編集——すべてを選択} {書式——タグのクリア} で，すべてのタグを削除する。テキストファイルで保存する（ファイル名を打ち込み，ファイルの種類を「テキストファイル」にすると，自動的に .txt の拡張子がつく）。

Christie, Agatha. 1920. The Mysterious Affair at Styles : Chapter 1 : I Go to Styles.

LINK＝"#000050" VLINK＝"#000050" ALINK＝"#000050"〉

1
I Go to Styles

THE intense interest aroused in the public by what was known at the time as "The Styles Case" has now somewhat subsided. Nevertheless, in view of the world-wide notoriety which attended it, I have been asked, both by my friend Poirot and the family themselves, to write an account of the whole story. This, we trust, will effectually silence the sensational rumours which still persist.   ;

  ;   ; I will therefore briefly set down the circumstances which led to my being connected with the affair.   ;

  ;   ; I had been invalided home from the Front ; and, after spending some months in a rather depressing Convalescent Home, was given a month's sick leave. Having no near relations or friends, I was trying to make up my mind what to do, when I ran across John Cavendish. I had seen very little of him for some years. Indeed, I had never known him particularly well. He was a good fifteen years my senior, for one thing, though he hardly looked his forty-five years. As a boy, though, I had often stayed at Styles, his mother's place in Essex.   ;

  ;   ; We had a good yarn about old times, and it ended in his inviting me down to Styles to spend my leave there.   ;

    "The mater will be delighted to see you again— after all those years," he added. 5

このテキストファイル上にいくつかの特徴が見られる。先ず，すべてのタグをクリアしたあとに，数字が残っているが，これは，元々のhtmlファイル上についている行番号で，これがそのまま残ったものである。

次に特殊な文字列が見られるが，これらはアンパサント（&）で始まり，セミコロン（;）で終わる一連のエスケープシーケンス（escape sequences）である。( ) はスペースを，(&# 151;) はハイフンをあらわす。[3]

また *Styles* では，主人公のことばの中にフランス語も出てくるが，これらの一部も特殊な文字列でててくる。

&_grave; (grave accent)

Voilà / là là / là -bas / rè gle

&_acute; (acute accent)

proté gé e / Sacré / Ohó / ré union / dé nouement

&_circ; (circumflex accent)

fê te / fâ chez / tê te / rô le / châ teau

&_cedil; (cediilla)

Ç a

外国語の場合の整形として，いくつかの方法がある。たとえば，〈fr〉...〈/fr〉で括るとか，テキストファイル作成者が同一化できる文字を適当に入れる（Voilà を Voila とか Voila#，あるいは，全く省略して Voil にする)。

さらに *Styles* の中では温度も次のような文字列で表される。

° (degree) 80 °

③　テキストファイルを開いて，{検索——置換} 機能を使って整形していく。整形前のテキストファイルの状態により，当然整形順序は変わってくる。置換機能を使用する際には前述した，メタキャラクター，正規表現が役に立つ。

1．空白行の処理（置換機能，通常モード)。

置換前：¥n¥n（¥n は改行記号を表す）

置換後：¥n

あるいは,

置換前：＾ $（空白行を表す）

置換後：なにも入れない

2．文頭のスペースを削除して，テキストをすべて左寄せにする。この場合は文頭にある（ ）をとる（置換機能，通常モード）。

置換前：

置換後：なにも入れない

3．段落末の をとる。

置換前：

置換後：なにも入れない

4．—をハイフンにかえる（置換機能，通常モード）。

置換前：—

置換後：－

5．元々の html ファイルについている行番号をとる（置換機能，正規モード）。

置換前：[0-9]＋$

置換後：スペースをうつ

6．外国語を検索し，適当な文字に変える（検索機能，正規モード）

{検索}：&.＋;（'&　　;' に注目して検索する。'.' は任意の 1 文字，'＋' は直前文字の一回以上の繰り返しを表す）

7．テキストファイルの最初の個所，最終行の整形をする。

テキストの最初と最後の〈著者，タイトル〉――〈title〉...〈/title〉で括る

章の表記は，〈chp id＝8〉

整形済みのテキストファイルは以下のようになる。

〈title〉 Christie, Agatha. 1920. The Mysterious Affair at Styles: Chapter 1: I Go to Styles. 〈/title〉

〈chp id＝1〉

THE intense interest aroused in the public by what was known at the time as "The Styles Case" has now somewhat subsided. Nevertheless, in view of the world-wide notoriety which attended it, I have been asked, both by my friend Poirot and the family themselves, to write an account of the whole story. This, we trust, will effectually silence the sensational rumours which still persist.

I will therefore briefly set down the circumstances which led to my being connected with the affair.

I had been invalided home from the Front; and, after spending some months in a rather depressing Convalescent Home, was given a month's sick leave. Having no near relations or friends, I was trying to make up my mind what to do, when I ran across John Cavendish. I had seen very little of him for some years. Indeed, I had never known him particularly well. He was a good fifteen years my senior, for one thing, though he hardly looked his forty-five years. As a boy, though, I had often stayed at Styles, his mother's place in Essex.

We had a good yarn about old times, and it ended in his inviting me down to Styles to spend my leave there.

"The mater will be delighted to see you again—after all those years," he added.

III-B *Witness* の場合

Witness は前述したように，OCR でテキストデータ化したものにスペルチェックをかけたものであるが，特に人名の不完全なもののスペルチェックがこぼれていることが多い。また英文字ながら，全角で入っていることも多く，元のテキストとの照合が大切なポイントになる。また，元の頁番号，行番号が途中に入っていたり，さらにその数字が全角で入っていることもある。戯曲という点で，整形上問題になるのは，登場人物を，〈 〉で括ることである。またト書きの部分はそのまま丸括弧付きで残すことにする。

9

Act One

SCENE: The chambers of Sir Wilfrid Robarts, Q. C.

The scene is Sir Wilfrid's private office. It is a narrow room with the door L. and a window R. The window has a deep built-in window seat and overlooks a tall plain brick wall. There is a fireplace C. of the back wall, flanked by bookcases filled with heavy legal volumes. There is a desk R. C. with a swivel chair R. of it and a leather-covered upright chair L. of it. A second upright chair stands against the bookcases L. of the fireplace. In the corner up R. is a tall reading desk, and in the corner up L. are some coat-hooks attached to the wall. At

night the room is lit by electric candle-lamp wall-brackets R. and L. of the fireplace and an angle-poise lamp on the desk. The light switch is below the door L. There is a bell push L. of the fireplace. The desk has a telephone on it and is littered with legal documents. There are the usual deed-boxes and there is a litter of documents on the window seat.

When the Curtain rises it is afternoon and there is sunshine streaming in through the window R. The office is empty. GRETA, Sir Wilfrid's typist, enters immediately. She is an adenoidal girl with a good opinion of herself. She crosses to the fireplace, doing a "square dance" step, and takes a paper from a box-file on the mantelpiece. CARTER, the Chief Clerk, enters. He carries some letters. GRETA turns,sees CARTER, crosses and quietly exits. CARTER crosses to the desk and puts the letters on it. The TELEPHONE rings. CARTER lift the receiver.

10

CARTER. (Into the telephone.) Sir Wilfrid Robart's Chambers... Oh, it's you, Charles... No, Sir Wilfrid's in Court... Won't be back just yet... Yes, Shuttleworth Case... What-with Myers for the prosecution and Banter trying it?... He's been giving judgment for close on two hours already... No, not an earthly this evening. We're full up. Can give you an appointment tomorrow... No, couldn't possibly. I'm expecting Mayhew, of Mayhew and Brinskill you know, any minute now... Well, so long. (He replaces the receiver and sorts the documents on the desk.)

GRETA. (Enters. She is painting her nails.) Shall I make the tea, Mr. Carter?

CARTER. (Looking at his watch) It's hardly time yet, Greta.

1．'Act One' を〈A 1〉に，'Scene 1' を〈S 1〉に，検索機能を使用して，換える。

2．登場人物名（CARTER.）を〈C CARTER〉（C : Character）に変換する。

置換前：CARTER.

置換後：〈C CARTER〉

ただし文尾に人の名前がきて，ピリオドでおわっている場合も置換してしまうので，注意を要する。話し手名はほとんどが行頭にきているので，置換前：＾CARTER.　置換後：〈CARTER〉も有効である。

登場人物名に，誤字もなく，大文字小文字の間違いもなく，また常に行頭にきている場合には，一括変換が可能である。

置換前：＾¥([A-Z]＋¥)¥.

置換後：〈C ¥1〉

3．頁番号，行番号をとる。場所は一定していないうえ，時々全角で入っているため，まず，行頭の数字を削除する（正規モード）。

置換前：＾[0-9]＋

置換後：何も入れない

あるいは二文字目の数字を削除する。

置換前：＾[0-9]＋

置換後：何も入れない

さらに文中にある数字を確認しながら削除していく。

{検索}：[0-9]＋（正規モード）

4．左寄せにする。

5．空白行をとる。

6．段落の単位で改行記号をつけ，段落内の改行記号を削除する。文字などとリターン記号の間にスペースが入っている場合があるので，文字のすぐ後ろにリターン記号がくるようにまず整形する（置換前：スペース¥n　置換後：¥n　を繰り返すことにより可能）。段落単位の改行記号の場合，'.?!' の次に改行記号がきていることに注目し，まず改行記号を他の記号に置き換える（置換前：.¥n　置換後：.### / 置換前：?¥n　置換後：?### / 置換前：!¥n　置換後：!###)。段落内の改行記号を，スペースを入れて削除する（置換前：¥n　置換後：スペース入れる)。再び .###，?###，!###をそれぞれ，.¥n ?¥n !¥n に変換する。ただし，たまたま .¥n であっても段落内の改行記号の時もあるので注意を要する。

7．ハイフンの処理：原文の行またがりのハイフンの場合，ダッシュがハイフンに化けている場合，元来必要なハイフンの場合などさまざまであるので，{検索} を利用して整形する。

〈A1〉

SCENE : The chambers of Sir Wilfrid Robarts, Q. C.

The scene is Sir Wilfrid's private office. It is a narrow room with the door L. and a window R. The window has a deep built-in window seat and overlooks a tall plain brick wall. There is a fireplace C. of the back wall, flanked by bookcases filled with heavy legal volumes. There is a desk R. C. with a swivel chair R. of it and a leather-covered upright chair L. of it. A second upright chair stands against the bookcases L. of the fireplace. In the corner up R. is a tall reading desk, and in the corner up L. are some coat-hooks attached to the wall. At night the room is lit by electric candle-lamp wall-brackets R. and L. of the fireplace and an angle-poise lamp on the desk. The light switch is below the door L. There is a bell push L. of the fireplace. The desk has a telephone on it and is littered with legal documents. There are the usual deed-boxes and there is a litter of documents on the window seat.

When the Curtain rises it is afternoon and there is sunshine streaming in through the window R. The office is empty. GRETA, Sir Wilfrid's typist, enters immediately. She is an adenoidal girl with a good opinion of herself. She crosses to the fireplace, doing a "square dance" step, and takes a paper from a box-file on the mantelpiece. CARTER, the Chief Clerk, enters. He carries some letters. GRETA turns,sees CARTER, crosses and quietly exits. CARTER crosses to the desk and puts the letters on it. The TELEPHONE rings. CARTER lift the receiver.

〈C CARTER〉(Into the telephone.) Sir Wilfrid Robart's Chambers... Oh, it's you, Charles ... No, Sir Wilfrid's in Court... Won't be back just yet... Yes, Shuttleworth Case... What —with Myers for the prosecution and Banter trying it?... He's been giving judgment for close on two hours already... No, not an earthly this evening. We're full up. Can give you an appointment tomorrow... No, couldn't possibly. I'm expecting Mayhew, of Mayhew and Brinskill you know, any minute now... Well, so long.(He replaces the receiver and sorts the documents on the desk.)

〈C GRETA〉(Enters. She is painting her nails.)Shall I make the tea, Mr. Carter?

〈C CARTER〉(Looking at his watch)It's hardly time yet, Greta.

〈C GRETA〉It is by my watch.

III–C *Alice, Through* の場合

二つのアリス物語は，電子テキストとして，かなりポピュラーなものであり，いろいろな場

所で提供されているが，ものによっては，かなり間違いの見受けられるものがある。したがってしっかりしたテキストをダウンロードする必要がある。次にある版では，*Alice* の最初の献呈詩や，*Through* の最後の詩なども含まれ，完全な形のテキストになっている。ただ，Tenniel の絵も含まれているため，ダウンロードには時間がかかる。

http://www.students.uiuc.edu/~jbirenba/carroll.html

http://surf.germany.EU.net/bookland/classics/carroll/

このファイルでは，*Styles* とは異なり，通常の改行と段落内改行の区別分けをつけ，通常の改行のみ有効に変換することが容易に可能である。この点に重点をおき，*Alice* の 1 章の整形を試みる。

① WZ Editor で，html ファイルを開く。

```
〈HTML〉
〈!—This HTML file has been created by Lionel Con's texi2html 1.30.1 j (additions
by-joke)
    from alice. texi on 29 December 1994—〉
〈HEAD〉

〈TITLE〉 Alice's Adventures in Wonderland : I. Down The Rabbit-hole 〈/TITLE〉
〈/HEAD〉

〈BODY〉
〈P〉〈A HREF＝"alice_3. html"〉〈IMG SRC＝"rsc/btn.prev.gif" ALT＝"previous"〉〈/A〉〈A
HREF＝"alice_toc.html"〉〈IMG SRC＝"rsc/btn.toc.gif" ALT＝", contents"〉〈/A〉〈A HREF
＝"alice_5.html"〉〈IMG SRC＝"rsc/btn.next.gif" ALT＝", next"〉〈/A〉〈IMG SRC＝"rsc/hr.
gif" ALT＝"—"〉〈H1〉〈A NAME＝"SEC7" HREF＝"alice_toc.html#SEC7"〉 I. Down The
Rabbit-hole 〈/A〉〈/H1〉
〈P〉
〈IMG SRC＝"images/alice01.gif" ALT＝"[PICTURE]"〉
〈P〉
ALICE was beginning to get very tired of sitting by her sister on the
bank, and of having nothing to do : once or twice she had peeped into
the book her sister was reading, but it had no pictures or
```

conversations in it, “and what is the use of a book,” thought Alice,
“without pictures or conversation?”
〈P〉
So she was considering in her own mind(as well as she could, for the hot day made her feel very sleepy and stupid)whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.
〈P〉

② {置換} を使用し，通常改行を一旦別の記号に換える。
置換前：〈P〉
置換後：##

③ すべてを選択して，タグクリアし，テキストファイルで保存する。

from alice. texi on 29 December 1994—〉

Alice's Adventures in Wonderland : I. Down The Rabbit-hole

##I. Down The Rabbit-hole
##

##
ALICE was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do : once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, “and what is the use of a book,” thought Alice, “without pictures or conversation?”
##
So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid) whether the pleasure of making a daisy-chain would be worth the trouble of getting up and

picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.
##

④　テキストファイル上で，以下の整形をする。

1．先ず空白行を処理する。

置換前：^ $

置換後：なにも入れない

2．段落内改行をスペースに置換する。

置換前：¥n

置換後：スペース

3．##を改行記号に置換する。

置換前：##スペース

置換後：¥n

4．この時点でててくる空白行をさらに処理する（¥n¥n を ¥n に）。

5．ファイルの最初の部分を整形する。

整形ずみのファイルの一部を次に示す。

〈title〉I. Down The Rabbit-hole 〈/title〉
ALICE was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading,

but it had no pictures or conversations in it, "and what is the use of a book," thought Alice, "without pictures or conversation?"
So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid) whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

Ⅳ. ま と め

パーソナルコーパス構築に関して，オリジナルテキストの電子テキスト化，及び整形の段階の問題点をいくつか指摘した。電子テキスト化の段階として，OCR と，インターネットの場合を取り上げたが，前者に関してはオリジナルテキストの状態いかんによって，かなり修正に時間をとる。当然未修正の部分も残っているので整形にも時間がかかる。後者の場合には，特に正確なファイルをダウンロードすることが大切である。また html ファイルをどの時点でテキストファイルにするかということも，のちの整形のことを考えて決める必要がある。

全般的に整形の手順はテキストファイルの状態によって決まる。特に置換機能に関してはエディターの正規表現の項目が重要である。オリジナルテキストとして，*Syles, Witness, Alice, Through* の 4 作品をとりあげたが，*Syles* に関しては，テキストファイル化のときに残っている特殊な文字列，*Witness* では，OCR に付随する問題，登場人物名の処理，ハイフンの問題，*Alice, Through* に関しては，通常の改行と段落内改行の処理等を特にとりあげた。

謝 辞

京都外国語大学の赤野一郎先生には，「パーソナルコーパス構築」に関して，いろいろとご指導いただきました。心より御礼申し上げます。

注

1．MicroConcord は言語処理プログラムで，オックスフォード大学出版局が販売している。
2．WordSmith も言語処理プログラムで，オックスフォード大学出版局が販売している。
3．エスケープシーケンスなどに関する詳しい情報は次のアドレスにある。
http ://www.cc.ukans.edu/info/HTML_quick.html
http://www.w3.org/pub/WWW/MarkUp/html-spec/html-spec toc.html
http://www.ncsa.uiuc.edu/General/Internet/WWW/HTMLPrimerAll.html

参 考 文 献

赤野一郎. 1996.「英語教師のための有益サイト情報とパーソナル・コーパス構築のすすめ」*CHART NETWORK*, No. 20. 数研出版.
赤野一郎・吉村由佳. 1994.「KUFS コーパスの構築について」『京都外国語大学研究論叢』No. 42.
井上永幸. 1995.「MicroConcord—コンコーダンスプログラム—」『英語コーパス研究』第 2 号.
斉藤俊雄編. 1992.『英語英文学研究とコンピュータ』英潮社.
長瀬真理・西村弘之. 1986.『コンピュータによる文章解析入門—OCP への招待—』オーム社.
村山 皓・赤野一郎編. 1992.『〈新版〉異文化を知るための情報リテラシー—外国語と外国文化研究教育におけるコンピュータ利用入門』法律文化社.

村山 皓・赤野一郎編. 1997.『大学生活のためのコンピュータ　リテラシー・ブック』オーム社.
Renouf, Antoinette. 'Corpus Development' in Sinclair, J. (ed.) 1987. *Looking up : An Account of the COBUILD Project in Lexical Computing*. London : Collins.