英語例文データベースのための基本システムの作成

渡辺　雅仁

Production of the Basic System for the English Sentence Database

Masahito WATANABE

Abstract

It is important to collect huge amount of linguistic data and analyze them in a proper way for any linguistic research or language teaching.

Recent technological development enables us to approach these date in a drastically simpler way than before. However, just collecting or searching language data is not sufficient for the better linguistic or educational purposes. We have to store the selected date systematically, tagging a lot of relevant information with them. In other words, we have to reorganize the selected data into a systematic database.

We aimed to construct a highly workable system for the English sentence database. Specifically, our system has following advantages:

1. We can reduce and share most of the laborious work of data input.
2. We can integrate the performance of the several ready-made programs.
3. We adopt a data format adaptable to almost any kinds of database software for the further develpment of the data.

0．はじめに

膨大な言語資料より、文や文章を採集しデータベース化することは、英語の各種学際的研究と教育活動には欠かせない。過去このデータベースの作成は膨大な人的資源の投入なしには成し得なかった。しかし、近年の科学技術の進歩によりこの煩雑な作業の多くの部分を機械化することが可能となった。本研究は最新の情報機器の援助により、英語例文のデータベースための操作性と汎用性の高いシステムを構築することを目的として行われた。

1．データベース化に関して考慮すべき点

① データベースの用途・目的

→外国語教育という目的に有効な例文のデータベースを作成する。

② 情報の収集方法

→①の目的を効率よく果たすことのできる情報源から収集する。

③ 情報の入力形式

→品詞、関連語句、文法事項など複数の付加情報を与える。

④ 物的な制約

→費用合計は100万円を越えないようにする。

⑤ 入力作業上の制約

→煩雑な入力作業の省力化と分散化を図る。

(1) Leech(1991)より

Clearly, a corpus, however large, when stored orthographically in machine-readable form, is of no use unless the information it contains can become available to the user.

2．研究計画

① 購入機材

→文字認識ソフト（ＯＣＲ：Optical Character Recognizer、PCR-English*）

→Lotus 1-2-3, 1-2-3 Card, PC-9801 DA + 8MB RAM

② ソフトウェア開発

→特定プログラム＋汎用プログラム

③ 情報の収集

→本学の過去の入試問題

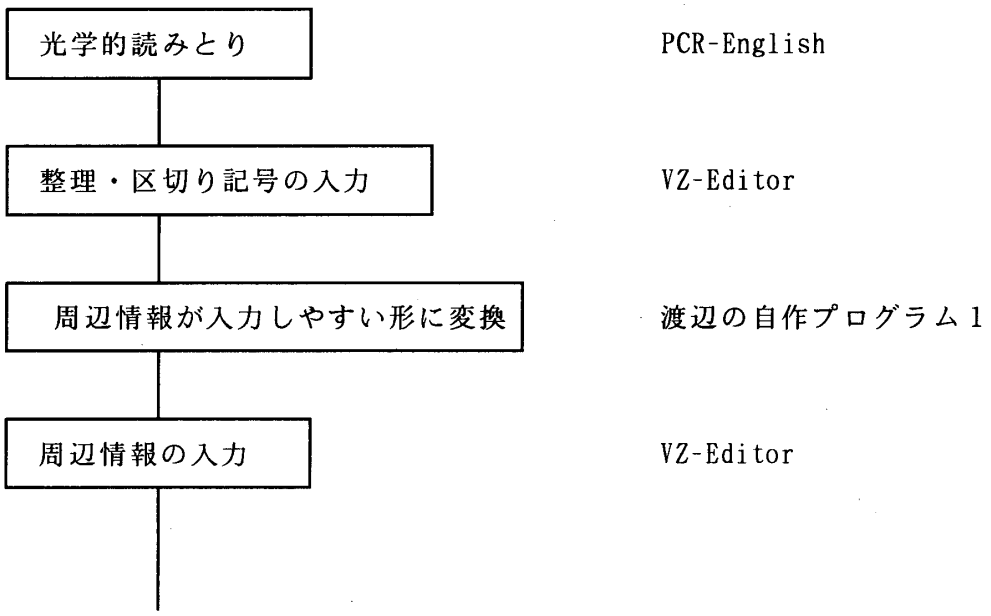
④ 情報の入・出力形式

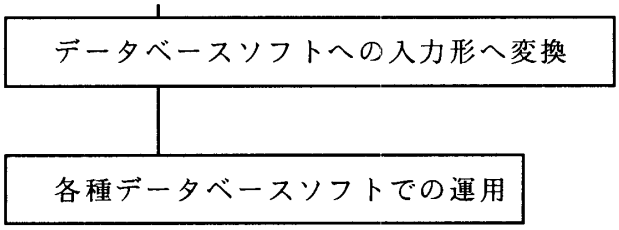
→最終的には以下のようなカード形式で出力できるよう入力を行う。

(2)

年度/90　試験種/二期外国語　問題/C　　形式/文の書き換え問題
主題/文の意義　　難易度/普通　正答/3　　問題番号/21
ｷｰﾜｰﾄﾞ/
not ... any more　　never
no more
内容/
21.　He never eats cake any more.
(1)　He doesn't eat as much cake as he used to.
(2)　He has never eaten any more cake.
(3)　He has stopped eating cake.
(4)　He used to eat cake, but not as much as he does now.

3．実際の入力作業





(3) 光学的読みとり

21 .
He never eats cake any more.
(1) He doesn't eat as much cake as he used to.
(2) He has never eaten any more cake.
(3) He has stopped eating cake.
(4) He used to eat cake, but not as much as he d
oes now.

(4) 整理・区切り記号の入力

```
**
::
21.     He never eats cake any more.
        (1)  He doesn't eat as much cake as he used to.
        (2)  He has never eaten any more cake.
        (3)  He has stopped eating cake.
        (4)  He used to eat cake, but not as much as he does now.
```

(5) 周辺情報が入力しやすい形に変換

```
:::年度/
:::試験種/
:::問題/
:::形式/
:::主題/
:::難易度/
:::正答/
:::番号/21
:::ｷｰﾜｰﾄﾞ/
:::
:::内容/
21.     He never eats cake any more.
        (1)  He doesn't eat as much
        (2)  He has never eaten any
        (3)  He has stopped eating c
        (4)  He used to eat cake, bu
:::
```

(6) 周辺情報の入力

```
:::年度/90
:::試験種/5
:::問題/C
:::形式/6
:::主題/5
:::難易度/2
:::正答/3
:::番号/21
:::ｷｰﾜｰﾄﾞ/
not ... any more
never
no more
:::
:::内容/
21.     He never eats cake any more.
        (1)  He doesn't eat as much
        (2)  He has never eaten any
        (3)  He has stopped eating c
        (4)  He used to eat cake, bu
:::
```

(7) データベースソフトへの入力形へ変換

```
 90 ,"5","C","6","5", 2 ,"3","21","not ... any more","never","no more","","","",
"21.     He never eats cake any more.","       (1)  He doesn't eat as much cake a
s he used to.","       (2)  He has never eaten any more cake.","       (3)  He h
as stopped eating cake.","       (4)  He used to eat cake*| but not as much as h
e does now.    ","ENDOFKWD"
```

4．自作ソフト解説

(8) 定義ファイルによるソフトの汎用化

	タイプ	項目名	項目属性	桁数／文字数	個数
:::A/年度/N/3/	項目	年度	数字	3	1
:::A/試験種/C/2/	項目	試験種	文字	2	1
:::A/問題/C/2/	項目	問題	文字	2	1
:::A/形式/C/2/	項目	形式	文字	2	1
:::A/主題/C/2/	項目	主題	文字	2	1
:::A/難易度/N/2/	項目	難易度	数字	2	1
:::A/正答/C/3/	項目	正答	文字	3	1
:::A/番号/N/2/	項目	番号	数字	3	1
:::K/ｷｰﾜｰﾄﾞ/C/30/6/	キーワード	ｷｰﾜｰﾄﾞ	文字	３０	6
:::K/内容/C/80/10/	キーワード	内容	文字	８０	１０

5．まとめ

(9) A. 市販ソフトの良い部分だけをつなぎあわせるので、高度な操作性と実行結果を得ることができる。

B. 複数の市販ソフトで利用できるデータ形式を採用しているので、データの汎用性が高い。

C. 複雑なアルゴリズムを必要とする処理は市販ソフトで行う。自作ソフトは市販ソフトを操作性よくつなぐことだけを目的とするので、ソフト開発が簡単に行える。

＊それぞれのソフトはそれぞれの会社の登録商標です。

PCR-English:　（株）バーズ情報科学

VZ-Editor:　（株）ビレッジハウス

Lotus 1-2-3:　ロータス（株）

1-2-3 Card:　（株）ダットジャパン

参照文献

Leech, G. (1991), 'The State of the Art in Corpus Linguistics', in Aijmer and Altenberg (eds) 1991: 8-29.

淀縄ほか(1992),　「英語例文データベースのための基本システムの作成」, 明海大学外国語学部,『外国語学部論集』第５集.

渡辺雅仁(1992),　「英語例文データベースのためのソフトウェア開発」, 明海大学外国語学部,『外国語学部論集』第５集.

明海大学 外国語学部 英米語学科 講師　渡辺　雅仁

Masahito WATANABE

Meikai University

Faculty of Foreign Languages and Cultures--Department of English

PC-VAN TBB14330 WATA