07207102　A01

外国人のための効果的な漢字・読解教育を支援する教育データベースの開発

研究代表者：カイザー　シュテファン（筑波大学・文芸・言語学系・教授　）

研究分担者：加納千恵子（筑波大学・文芸言語学系・助教授）
市川　保子（筑波大学・文芸言語学系・助教授）
小林　典子（筑波大学・文芸言語学系・講師）
酒井たか子（筑波大学・文芸言学語系・講師）
山本　啓史（筑波大学・文芸言語学系・講師）
吉岡　亮衛（国立教育研究所・教育情報資料センター・主任研究員）

交付金額　：2,000千円

代表者連絡先：〒305 つくば市天王台 1-1-1
筑波大学・留学生センター
Tel. 0298-53-6242　Fax. 0298-53-6204
E-mail：kaiser@intersc.tsukuba.ac.jp

重点領域では5つのデータベースを開発した他，電子化テキストを日本語の教材や教育の資料として加工する　ツールの開発，導入，調査をおこなった。一部は，日本語教育，特に本研究でテーマとする漢字・読解を訓練する各種データベースとその周辺にあたる評価問題，誤用例に関する調査データベースの開発について述べる。二部では，日本語教育を支援するデータベース用各種ツールの入手，導入の調査，および開発したシステムについて述べる。最後に次年度の計画についてその概略を述べる。

1.　漢字教育・読解教育に役立つデータベースの開発と関連データベース

本研究課題において，本研究班は勤務先に蓄積する秩序のない資料を整理し，適宜判断基準とデータ構造を選択しつつ，データベースの構築を行った。研究はもっぱら，テキスト処理を後に行うことを前提に，漢字・読解教育のように文字列を主に扱う教科の教育支援をテーマとした。同時に，小林担当分のような学生の音声テープの書き取り課題をワープロで行うことにより，外国人の聞き誤りをデータベ

ース化し，誤用研究に利用したり，酒井・市川担当分のような診断テストの問題データバンクや学習指針の提示にデータベースを利用したりするなど現実的かつ効率のよい計算機利用を実践し，人文領域の計算機利用が有意義であることを証明することができた。本年度，本研究課題で開発した，あるいは開発中のデータベースは次の通りである。

表 1 本研究課題で開発・開発中のデータベース

No.	データベース名(仮命名のものを含む)
1	非漢字圏学習者のための漢字学習情報データベース
2	論文読解のための表現文型データベース
3	漢字学習方法処方のための漢字力診断問題データベース
4	学習者の誤用実態を明らかにする誤聴解データベース
5	事前診断テスト問題データベース

上記 1 ～ 3 のデータ整理の終ったデータベースについて，目的とデータ形式，アクセス方法，応用例にわけ，以下に解説する。

1.1 「非漢字圏学習者のための漢字学習情報データベース」

(1) 目的

非漢字圏日本語学習者の漢字学習支援システムを開発する目的で，このデータベースは開発された。最終的な目的は，システムを使用する学習者がどういう機能をどの程度利用するか記録し，データを収集し，そのデータの分析に基づいて漢字学習に有効な学習・指導方法があるかを検討することである。

(2) データ形式

表 2 KID(Kanji Information Database のデータ形式

```
No.34 円
=========
ON : EN
Kun : maru(i)
Core meaning(s) : EN circle,yen,maru(i) round
Related Meaning Kanji : 丸 (GAN : round)
Ready made story :
Your own story :
Same phonetic[same ON] : --
Same phonetic[similar ON] : 員 (IN : member)
                            韻 (IN : rhyme)
Same phonetic[different ON] : 損 (SON : loss)
Similar shape phonetic : --
Similar Shape Kanji : 内 , 再 , 同
Same Radical Kanji : 内 , 冊 , 同 , 再 , 周 , 岡
Simple Examples in Context : 十円玉がたくさんある
Word building[compounds] : 百円 , 円形
Compounds Examples in Context : この広場は円形だ
Your Comments for Improvement :
```

学習者によってはシステムで用意されているデータは不十分であったり，デザイン上漏れてしまった，他の機能が欲しかったりする可能性があり，そのためにはユーザ側のコメントをできる限りとりこむようなデータのフィールドをあらかじめ用意しておいた。

(3) アクセス方法

現在はデータの入力にマッキントッシュを使った。今後は WWW ブラウザで使用するようにデータをコンバートした。

漢字そのもののデータだけでなく，漢字を取り扱っているときに学習者はどのような試みを行っているかを調査するための基礎資料となるであろう。

2.2 論文読解のための表現文型データベース

(1) 目的

科学技術の読解支援システムに利用するためのデータベースを開発した。とくに，本研究では，科学技術文献における重要表現の訓練のためのデータベースを開発した。開発にはサンマイクロシステム社のワークステーション Sun SPARCstation 5 を開発マシンとし，C，AWK，SED などのプログラミング言語，スクリプト言語などを使い，開発した。

(2) データ形式

日本語の科学・技術論文，雑誌の記事を収集し，それぞれ機能別に例文と文型を整理し，機能，日本語，英語，翻訳のインデックスをつけたデータベースを作成した。

表 3 論文読解のための表現文型データベースのデータ構造

```
機　能：問題の核心・要点
日本語：これは問題の核心にふれる / これは問題
　　を解く鍵である
英　語：～ is, in fact, the key to ～
翻　訳：～は，実際，～を解く鍵である。
```

(3) アクセス方法

現在は awk のスクリプトで対話的に検索・提示している。

(4) 応用例

たとえば，問題の核心や要点を述べる表現を日本語でなんというかを調べたい時は key,glimpse,core,heart などの用語で検索し，“これは問題の核心にふれる / これは問題を解く鍵である”という日本文を提示してくれる。これは作文にも使えるデータベースである。しかし，本研究ではこのデータベースを利用して，電子化された専門文献から問題点を指摘する，問題を考察する観点を示す，問題を後回しに

する，問題の考察を始める，問題の核心・要点を述べる，例をあげる，理由や根拠を述べる，要約する，由来を示す，本論へ導入する，分類する，付け加える，表現法を限定する，表現される，反論する，反対の趣旨に導く，対照的なことへつなぐ，反対の趣旨に導く，などの機能を含む日本語文を自動抽出し，単文，複文，名詞修飾節を持つ文，またその複文のように文の複雑さ，文の長さでソーティングを行い，単文読解訓練用の文章を作成するシステムを作った。今後の課題は，このシステムが出力する単文に対する回答方法を受理し，その正誤の判定をする部分を開発すること，システムの評価があげられる。

2.3 漢字学習方法処方のための漢字力診断問題データベース

(1) 目的

初級の漢字学習を終えた外国人学習者の漢字力を測定するためには，文中の漢字語の読みをひらがなで書かせたり，ひらがなで書かれた語を漢字で書き換えさせたりする，いわゆる「漢字の読み書きテスト」が用いられてきた。しかし，この従来の方法では，表面的な漢字の読み書きの到達度を測ることはできても，その背後にある漢字の習得状況を把握したり，学習者の抱えている漢字学習上の困難点を発見したりすることは難しい。なぜなら，テストを厳格に採点すると，その漢字を全く知らないために解答できなかった場合の結果と，ひらがな表記の誤りや発音の間違い，細かい点画の間違いなどによる減点の結果と区別がつかないからである。

そこで，漢字力を構成する概念を分析し，その構造にしたがった要因についての問題とその形式を厳しく検討し，時には実験によって検討し，適切な測度をもった漢字診断問題を作成し，データベースとして蓄積，利用することが必要となる。

本研究におけるデータベースは漢字の診断に用いるのに役立つ問題のデータベースとその構造の開発である。

(2) データ形式

データの形式は従来は1問題1レコードの形式を採用していたが，SGML化をおこない，テストシステムとして用いる時には，HTMLにコンバートし，プリントテストとして用いる時にはLaTeXとしてきれいに清書されて出力できるようにした。

ここではHTMLの形式を提示する。

(3) アクセス方法

現在はNetscape 2.0などのWWWのブラウザでアクセスできるようになっている。

(4) 応用例

漢字診断のバリエーションとして文法診断のためのデータベースや漢字学習のデータベースもネットワークによる集中管理でデータを整理する形式を採用することにより，総合評価データベースとそのデータ管理の足掛りとして利用できよう。

2. 日本語教育を支援するデータベース関連ツールの調査，導入，開発

ワークステーションにおける言語処理ツールが発達している理由として，工学系の研究者の研究道具の多くがワークステーションであることがあげられよう。しかし，それが直接人文系の研究者にとっては手軽に使える状況でないと言い切るには無理があろう。事実10年前のパソコンには現在のようなGUI(Graphical User Interface)のようなメモリ食いのディスク食いなど資源的に許されないことであり，みなコマンドラインからコマンドを入力することによってプログラムを動かしていたのである。確かにワークステーションが，設定に関してパソコンとは比べものにならないほど管理は大変であることは事実であるが，研究機関に所属する研究者なら，計算機センターにUNIXの入った汎用機が動いているであろう。今日，インターネットによって言語ツールや言語データの共有化がすすめられ，多くの自然言語処理の研究者によって，多彩なツール・データが開発され，これらのものはかなり強力で，組合せて利用すれば多彩なテキスト編集が行える。

ここではワークステーションを中心とする環境で利用できるツール，コーパス，辞書について述べる。また，これらの一部は本研究を進める上でかなり強力な道具であったことを付け加え，多くの日本語教育の研究者や教育者がこれらの便利な道具を使える

表4 漢字学習方法処方のための漢字力診断問題のデータベース構造

```
<TR ALIGN=CENTER><TD></TD></TR>
<OL>
<TR ALIGN=CENTER>
<TD><font size=6, color="green"> 乗 </TD>
<!--Monday--><TD><font size=6><INPUT TYPE="radio" NAME="kanjianswer8"
    VALUE="answer1"> 進 </TD>
```

```
<!--Teusday--><TD><font size=6><INPUT TYPE="radio" NAME="kanjianswer8"
    VALUE="answer2"> 退 </TD>
<!--Wednesday--><TD><font size=6><INPUT TYPE="radio" NAME="kanjianswer8"
    VALUE="answer3"> 降 </TD>
<!--Thursday--><TD><font size=6><INPUT TYPE="radio" NAME="kanjianswer8"
    VALUE="answer4"> 行 </TD>
</font>
</TR>
</OL>
<TR ALIGN=CENTER><TD></TD></TR>
```

ような中心的なサーバーがどこかに提供されることが望ましいと考える。

2.1 OCR... 光学式文字読みとり装置 (Optical Charac te Reader)

筑波大学留学生センターではパソコン (Apple 社 Macintosh) にスキャナで (Hewlett Packerd 社 ScanJet II cx) を接続し，ソフト (MacReaderJapan Ver.2.5) を使用して，印刷物からテキストを読み取っている。読み取り後の文字を日本語辞書・英語辞書と照合することにより辞書にあることばから字形の近似を行い，精度の向上を図っている。これは確かにワークステーションでは作動するものではないが，ペーパーメディアの加工は個々からはじめなければならない。

2.2 JUMAN... 形態素解析プログラム (Japanese Mor phological Analysis System)

松本裕治ら（奈良先端科学技術大学院大学）が開発した形態素解析プログラム。研究目的に利用するなら，配布は自由。BSD,System V 系の OS に対応している。JUMAN は形態素に区切る他，基本形 (辞書見だし)，漢字の読み，品詞情報を出力する。また，JUMAN の出力は次に示す構文解析プログラムに通すことによって文構造データを出力する。日本語教育への応用例としては，ある文章の単語リストを作成したいとき，形態素切りを行った後で grep, awk などのテキスト処理ツールで，独立語だけの抽出をおこなう。そうすることによって助詞，助動詞などを除いた単語リストの一覧がすぐにできあがる。

2.3 KNP... 構文解析プログラム (KN parser (Kuroha sh-Nagao parser)

京都大学長尾真研究室で開発された構文解析プログラム。JUMAN と同じく UNIX ワークステーション上でコンパイルし利用することができる。KNP は後に述べる EDR 電子化辞書の日本語単語辞書や分類語彙表のデータ，および JUMAN の結果を受け取り，木構造を出力することができる。このような頻繁な入出力データのやりとりを柔軟に行うためには GUI の環境ではなくパイプ処理の行える CUI(Character User Interface) のほうが便利である。

日本語教育への応用としては，単文，複文，名詞修飾文，入れ子の深さなど読解教育で使用する文構造パターンの練習教材を作成する際，同じ形式の同じ難易の文型を計算処理によって抽出できるというおそろしく便利な使い方ができる。

2.4 KAKASI... "KAKASI" は "kanji kana simple in veter"

作者は高橋裕信 takahasi@tiny.or.jp (foreign: hirono bu@netcom.co) さんで，KAKASI は漢字かなまじり文をひらがな文やローマ字文に変換することを目的として作成したプログラムと辞書の総称。プラットホームは sunos4.1.3 (cc, gcc)，solaris 2.3 ibm rs6000 (cc) でコンパイル，実行できる。日本語教育の応用は辞書の見出し語作成，活用形音変化テーブルの作成，ローマ字記述教科書の作成などさまざま。

2.5 日本経済新聞記事データ CD-ROM 版

日経総合販売㈱の厚意により有料ではあるが研究目的での記事データが使用できる。多くの新聞記事データは著作権の問題でなかなか研究目的に利用できないという問題点があった。利用条件については「日経新聞データ使用許諾に関する覚書」を交わす必要がある。

詳しくは http://cactus.aist-nara.ac.jp:80/lab/resource/cdrom/Nikkei/NKS.html。

2.6 朝日新聞 WWW home page

1995 年 8 月 10 日より朝日新聞（総合，社会，経済，スポーツ，社説，天声人語，特集）がほぼ新聞誌面と同じ内容で掲載されている。また，ニュース速報，求人情報，天気予報，リリース，広告一覧のほか，Asahi Evening News の記事も掲載されている。著作件上このような著作物を利用して作成した教材使用は個人の使用に限定される。

詳しくは http://www.asahi.com。

2.7 IPAL 辞書

情報処理振興事業協会 (IPA) の開発・編集の辞書で，計算機用日本語基本動詞辞書 IPAL(Basic Verbs) 付録：サ変動詞辞書，基本形容詞辞書 IPAL(Basic Adjectives) 付録：形容動詞辞書の他，1995 年 10 月には基本名詞辞書 (800 語) をリリース。すべてインターネット経由で利用できる。ftp://ftp.mgt.ipa.go.jp から利用することができる。ただし，利用には，利用申請書を提出する必要がある。この辞書を利用して JUMAN,KNP を動かして，解析をする研究者も多い。

2.8 EDR 辞書

基盤技術研究促進センターとコンピュータメーカ 8 社との共同出資のもとに，9 年間のプロジェクト (1985 年度～ 1994 年度) により得られた成果。単語辞書中で定義した概念の類義を記述する概念体系（シソーラス），辞書記述の典拠としてのコーパス DB(例文集) を統合した日本語と英語の機械処理用の電子化辞書。いくつかの自然言語処理ソフトウェアはこの辞書のデータを基に動くようになっている。1995 年 4 月より (株) 日本電子化辞書研究所から市販されている。詳しくは http://www.iijnet.or.jp:80/edr/Intro.html。

2.8 フロッピー版分類語彙表

1994 年国立国語研究所編，秀英出版より出版された刊行物で，国立国語研究所資料集 6「分類語彙表」のフロッピー版。日本語教育の教材を作る時にはこのような語彙シソーラスをつかって関連語彙表を作ったり，新聞記事などから検索した例文に含まれる語彙が教育的に適切でない場合，入れ替える際，便利である。

2.9 Edict... 電子和英辞書

Jim Breen(オーストラリア，モナッシュ大学）がネットワークハッカーに呼びかけて作成した 3MB ほどの Freeware の英和・和英辞典。ftp://ftp.cc.monash.edu.au/pub/nihongo にある。使用は一切無料。アップデートも随時行われている。

2.10 Lookup... ルーキャップ

Jeffrey E.F. Friedl(オムロン株式会社 jfriedl@nff.ncl.omron.co.jp) が開発した検索システムで Edict をワークステーション上で高速に検索することができ，なおかつ，日本語文字列の正規表現検索ができる。また，同時に複数の辞書を使用でき，サブストリング，類似文字列提示，さらにいかにも外国人が作ったシステムらしい，長音，濁音を無視したり補完したりして検索する機能がある。日本語教育関係者よりも日本語学習者にとってありがたいシステムといえよう。ワークステーションがない場合，Netscape WWW browser により，http://www.wg.omron.co.jp/cgi-bin/j-e/jfriedl.html から辞書を引くことができる。

日本語教育での応用として，このような freeware の英和辞書があることは非常にありがたく JUMAN で切り出した単語を見出しとして検索し，その項目を取り出して一覧表を作ることで日英対訳単語帳が簡単にできてしまう。

2.11 LaTeX... タイプセッティングプログラム

数学者，Knuth が自分の論文を清書するために作ったといわれるプログラム TeX に Lamport がさらに論理的な概念を記述するだけで，視覚的にきれいに出力できるよう，マクロ・パッケージを加えたもの。もともと UNIXのシステムで提供されていたが，プログラムの移植が行われ，現在では多くのパソコンでもつかえるようになっている。スタイルファイルを作ることにより定型的な文書の処理に威力を発揮する。科学研究費申請書マクロ・パッケージはインターネットによって配布され，多くの研究者によって利用されているのは周知の事実。多くの日本語教育の漢字ルビは上ルビではなく，下ルビが使われているが，ほとんどのワープロには下ルビの機能を有していない。TeX は簡単な命令の書き換えで下ルビをつけることもできる。また，地名，登場人物名をユーザーが定義しておくことで，それらの一括置き換えが可能である。UNIX，DOS，Macintosh いずれの処理系で作成されたソースファイルも相互にやりとりでき，出力ファイルは DVI(Device Independent) とよばれ，いかなるプリンタにおいても同じレイアウトで出力することができる。ファイルのやりとりに伴う危険な illegal copy の問題の解消にも役立つ。

2.12 Grep,Sed,AWK... いずれも UNIX のテキストツール

正規表現検索を行い，任意の文型を含む行を抜き出して表示する Grep(Global RegularExpression Printer)，任意の文型を見つけたら，それに編集を加え，出力する Sed(Stream EDitor)，任意の文型を見つけたら，それを数えたり，編集したりできる簡易テキ

スト編集プログラミング言語ツールAWK(AWKは開発者，A.V.Aho, P.J.Weinberger, B.W.Kernighamの3人の頭文字をとってつけられた名前)。簡単な小規模のプレーンテキストのデータベースであれば，これらツールを使って作ることもできる。実際に本研究のデータベースはこれらのツールを使って整備された。

2.13 **Uniq...UNIXのテキストツール。**

隣り合う同じ文字列を数えて結果を出力する。語彙の出現頻度計算などに利用する。

2.14 **Sort...UNIXのテキストツール。**

テストの各行をアスキーコード順に昇順，降順に並べ替えるツール。

3. 日本語教育を支援するデータベース関連文献

3.1 日本語学連載「わたしのパソコン言語学」1991年9月～1994年2月，30回連載。大修館書店

言語学の専門家がパソコンをどのように使っているかを報告した記事。中には，パソコンと悪戦苦闘しているなどというつまらない記事もあるが，多くの報告は実務に関する有益な報告記事である。使い方，ソフトの入手法，使用例，研究例などが記載されている。

3.2 白井 英俊：計算言語学とインターネット，名古屋大学大型計算機センターニュース，Vol. 26. No 1, pp. 53-60

自然言語理解研究に従事する筆者が計算機を利用する人向けの講座読みものとして，綴った記事。7ページの短い記事ではあるが，その中には1.ニュースグループと電子メールによる情報収集，プログラム付き本，ソフトウェアやコーパスの蓄積，出版者案内などフリーで手に入る言語資源のURLを数多く記述している。こんなに研究に使えるものがあるのか，と思うなっとくのリストである。

3.3 松本 裕司，他：自然言語資源の共有化 Home-Page

平成6年6月に行われた「言語データ共有化計画」シンポジウムに関連して，自然言語資源共有化の重要性の認識の向上を目指し，研究調査の結果が公表されている。前述の項と同じく数多くの自然言語処理資源が整備されていることに驚かされる。関連のWWWサイトをごらんになることを勧める。

URLは http://cactus.aist-nara.ac.jp/lab/resource/resource.html。

3.4 前川 守：文学編 文章を科学する 1000万人のコンピュータ科学3, 岩波書店,1995.

作家と文体，文章を作る，暗号の話，文の構造など文学の分析に関する記事で興味深く書かれている。大学の教科書として書かれているようで，文の解析に関する基本的な概念や手法，問題点について述べられているだけでなく，C言語によるプログラム例が親切なコメントをちりばめて掲載されている。日本語教育の文の分析と教材としての適切さなどを計算機で処理する研究を行う際，有益である。

3.5. 坂口ら：コンコーダンスを指向したテキストデータベースの研究，情報処理学会研究報告 人文科学とコンピュータ，Vol. 96,No. 15, pp. 13-18,1996.

コンコーダンスプログラム(総語彙環境一括提示プログラム)の研究例。上記研究では，パソコンNECPC98で動作するプログラムを開発。文学作品「坊っちゃん」，琉球王国外交文書「歴代宝案」を解析実験している。パソコンで教材開発をしたい人には有益な情報。

3.6 Ken Lunde: Understanding Japanese Informatio Processing, O'Reilly & Associates,Inc,1993 「日語情報処理」

作者のKen Lundeは Adobeに勤める研究者で，この本のほとんどが日本語の文字とその計算機上でのデータの処理の方法についてかかれている。特に海外にいる日本語教育研究関係者にとって，日本語の情報処理は多くの点で問題を含んでいる。壊れたJIS漢字の修復の仕方や簡単な検索プログラム，文字種変換プログラム例，文字テーブルなど，親切に解説している。最近，同じ日本語タイトルで日本語訳が出版された。

4. 今後の課題

データは確実に仕上ってきている。つぎはこれらデータのドライバを何にするか，ユーザにとってやさしいインタフェースはなにか，使用している場面での新たな発見や問題点はどんなことがらであるかを評価し，言語教育資源の共有化の提言を行いたい。

5. 発表

カイザー シュテファン「非漢字圏学習者向けの漢字学習データベースのデザインにおける表音声情報について - 押韻の観点から -」情報処理学会研究報告 人文科学とコンピュータ,Vol. 95,No. 91, pp. 13-20,1995.

カイザー シュテファン，山元啓史「漢字学習情報データベース(KID)のデザインとネットワーク上使用の可能性について」The first international con

ference on computer assisted system for teaching and learning / Japanese 4-6 September, Pavia, Italy,pp. 78-89,1995.

加納千恵子「コンピュータによる漢字力テスト (CAT-K) の開発」 The first international conference on computer assisted system for teaching and learning / Japanese, 4-6 September, Pavia Italy,pp. 69-77,1995.

加納千恵子「科学・技術日本語の読解教育におけるシステム・授業・評価」情報処理学会研究報告人文科学とコンピュータ, Vol. 95,No. 91, pp. 1-6,1995

山元啓史「専門文献読解のための教材作成支援システムの開発 (1)」 The first international conference on computer assisted system for teaching and learning /Japanese 4-6 September, Pavia Italy,pp. 129-135,1995.