パソコンと汎用テキスト処理ツールによるドイツ語研究の可能性

城　岡　啓　二
(jjksiro@hss.shizuoka.ac.jp)

0．はじめに

かつてはコンピュータをドイツ語学・文学の研究に利用するなどというと，情報処理センターのようなところの高性能なコンピュータを利用してかなり高度なプログラムを購入するか，自分で書くか，他人に書いてもらうしかなかったようだ（植田 (1982)，菊池 (1983)，小林 (1988)，森 (1991)，河野 (1995)）．誰にでもおいそれと真似のできない大がかりなプロジェクト研究だったわけだ．その意味ではパソコンを使用した藤澤 (1984, 1985) や米井 (1986, 1989) の研究は研究活動へのパソコン利用の有効性をしめし，パソコンによるコンピュータ利用研究を身近なものにしたと言える．藤澤や米井は語彙や文字の頻度を出したり，語彙を逆引き配列にして派生語や複合語などの語彙家族を抽出する目的でパソコンを利用しているのだが，プログラミング言語として BASIC が使われている．したがって，それなりに高度なプログラミング技術が必要だったわけで，ちょっと使い方をおぼえてというわけにはいかなかった．しかし，パソコンの分野の発展は急速で，現在ではテキストデータの処理なら

汎用性の高いテキスト処理ツール群を活用するだけでかなりのことができるようになっている．事実，藤澤や米井のおこなったような研究は，現在の高性能なパソコンならこれらの汎用ツールを組み合わせるだけでほとんど実現可能である．

汎用テキスト処理ツールというのはテキストファイルを扱って各種の処理をしてくれる汎用性の高いツールで，もともとUNIXのコマンドやツールだったものが多く，現在ではMS-DOSなどに移植され，パソコンでも気軽に使えるようになっている．こういったツールの利点をまずあげるなら，使い方が易しく複雑なプログラムを組む必要がないということと，インターネットやCD-ROMで容易に入手可能ということが指摘できるだろう．一般的なものを列挙するならCUTやFOLDやSORTやUNIQやREVやCOMMやWCやDDやNLやPASTEやSPLITやTRやGREPやSEDやAWKなどがある．これらのツールは多くの一般の利用者を予想しているワープロソフトのようなものではなく．言わばプロのひとの使う小道具である．だから取っ付きにくい面もないわけではないが，使い方のコツさえつかんでしまえば，研究補助手段として極めて有効な道具になってくれるものである．使い方といってもSEDとAWK以外はMS-DOSのコマンドを利用するのとなんら違いはない．ただし，SEDやAWKを使う場合は，最低でも数行程度のスクリプトを自分で書ける必要がある．また，他人の書いたあまり難しくないスクリプトであれば意味がある程度は理解でき，必要に応じて書きかえができるぐらいの知識は身につけたほうがいいだろう．私自身はようやくこのレヴェルでしかない．

現在，汎用テキスト処理ツールはどの程度言語研究に利用されているだろうか．日本ではSORTFやMCLやKKCなどが文系の研究者によって開発されたりしているし[1]「パソコンを使う日本語研究」（明治書院）などを読めば分かるがかなり有効に利用されつつあるようだ．ドイツでは，どういうわけか，汎用テキスト処理ツールは人文系研究者にはほとんど利用されていないのか，「人文系科学におけるコンピュータの利用」というような題の本を見てもまったく触れられていない．W. Lenders (1993)：「応用言語学におけるコンピュータの利用」はコーパスの利用が中心の論文集で，コーパス利用のためのツールという章が設けられているにもかかわらず，扱われているツールはインデクスやコンコーダンスや語彙統計の統合ソフトのWordCruncherやMicro-OCPやTACTなどである．特定の使用目的に縛られない汎用テキスト処理ツールはまったく扱われていない．しかし，D. Hein (1995) はUNIXやLINUX上で

の使用であるが，汎用テキスト処理ツールを活用した語彙や形態素の研究という内容で，ドイツ語圏でもいずれ汎用テキスト処理ツールがパソコン上で利用されるようになってくるだろう．いずれにしても，汎用テキスト処理ツールの利用状況は，おそらく，日本語に対応した語学研究用の市販ソフトが未発達なこともあってか現段階では日本のほうがはるかに進んでいるようだ．

さて，本稿では，私自身の研究内容から具体例をひきながら，パソコンと汎用テキスト処理ツールがひらく可能性を紹介していくつもりである．しかし，この種のツールは汎用であっても万能ではない．また，私自身の関心分野が限られているということもある．したがって，以下で汎用テキスト処理ツールで可能になるドイツ語研究としてとりあげているのは，主として文字から語句レヴェルまでの文字列が関係した研究である．

1．データの整理と頻度調査

1．1 SORT で数値データを整理する方法

SORT はテキストファイルの行をいろいろな方法で並べかえるツールだ．MS-DOS にも付いてくるが，これは扱えるテキストファイルの大きさも制限されるうえ，並べかえの種類も少なく，非力である．幸い，UNIX の SORT に準じたものがかなり作られていて，フリーウェアとして提供されている．SORT は語学研究でもいろいろと役に立つツールで，城岡（1996）では一行一語の語彙リストから逆引き辞典を作成するのに利用している．この節では SORT の利用例として数値をソートして数値データを整理し，分析に役立てる例を示そう．

次のページのデータは E. Rosch のプロトタイプとよばれる意味の理論に基いたもので，「家具（furniture, Möbel）」の各種の成員について「家具らしさ」の評点をしめしている．被験者に 1 から 7 の 7 段階で典型性を評価させ，その平均をとった数値である．4 が平均的成員をあらわし，「ふつうの家具」という意味だ．数値が 1 に近いほどより典型的な成員ということで，逆に数値が 7 に近づけばより非典型的な成員ということになる．イスやテーブルはアメリカ人にとってもドイツ人にとってももっとも典型的な家具であり，電話や灰皿は非典型的な家具であり，アメリカ人もドイツ人も 6 以下の数値をつけている．データの読み方だが，アメリカ人学生（200 人）による数値が(a)だ．これは E. Rosch 自身の研究にある数値である．(b) は詳細は不明だが，アメリカ留学中のドイツ人学生（複数）がドイツ語に対して出した結果として J. R. Taylor（1995, p. 57）に紹介されている数値である．ただし，実際にどういうドイツ語の単語

に対してドイツ人学生の判断を調査したのかは不明である．最後の (c) はアメリカ人とドイツ人の数値の差の絶対値を出したものだ．なお，各行はアメリカ人のデータで数値の小さいものから並べてある．

■ プロトタイプとしての「家具」

(a) アメリカ人のつけた評点
(b) ドイツ人のつけた評点
(c) 評点の差の絶対値

	(a)	(b)	(c)
chair	1.04	1.20	0.16
sofa	1.04	1.13	0.09
table	1.10	1.00	0.10
desk	1.54	1.20	0.34
bed	1.58	1.00	0.58
bookcase	2.15	1.47	0.68
cabinet	2.49	1.73	0.76
lamp	2.94	4.79	1.85
cupboard	4.27	1.20	3.07
stereo	4.32	6.64	2.32
TV	4.41	6.07	1.66
shelf	4.52	2.00	2.52
closet	5.95	1.20	4.75
ashtray	6.35	6.80	0.45
telephone	6.68	6.80	0.12

それでは，アメリカ人とドイツ人の各種家具の典型性の判断の差の大きいものについてしらべたい場合はどうすればいいだろう．まず，知りたいのは評点の差の大きい「家具」がどんなものかということであろう．(c) の数値の大きいほうから順に見ていくのが研究方法の手順としてはよさそうだ．15 行程度のデータでもこれを正確にやろうとすると，手作業ではけっこう難しいのではないだろうか．また，15 行程度ならやる気はおこるが，対象の量がふえ，50 行や 100 行のデータに順番をつける作業となると，これはもう人間業ではない．こう

いう仕事を文句も言わずにやってくれるのがSORTというツールで，各行を差の大きいほうから何万行でも並べかえることができる．chairからtelephoneまでの15行をテキストファイルにしてkagu.datという名前にしておこう．差異の大きい順に並べ替えて結果をkagu.srtに出力するには，UNIX仕様のSORT（MS-DOSに付属のsort.exeではこういう使い方はできない）の場合だとコマンドラインに次のように打ち込む．

```
sort -nr +3 kagu.dat > kagu.srt
```

-nrはオプション指定と言われるものだが，ここでnは「数字としての並べ替え」を指定している（実はこの例ではnはあってもなくても結果は同じ）．rは「大きい順から並べかえ（降順）」という意味だ．-nrの次の+3は「行（レコード）の先頭から並べ替えるのではなく，3つ先の項目（フィールド）から並べ替え」という指定だ．kagu.srtという名前のファイルに結果は保存されるが，内容は次のようになっている．

```
closet      5.95    1.20    4.75
cupboard    4.27    1.20    3.07
shelf       4.52    2.00    2.52
stereo      4.32    6.64    2.32
lamp        2.94    4.79    1.85
TV          4.41    6.07    1.66
cabinet     2.49    1.73    0.76
bookcase    2.15    1.47    0.68
bed         1.58    1.00    0.58
ashtray     6.35    6.80    0.45
desk        1.54    1.20    0.34
chair       1.04    1.20    0.16
telephone   6.68    6.80    0.12
table       1.10    1.00    0.10
sofa        1.04    1.13    0.09
```

結果を見てみると，家具性の判断の違いが大きいのはなんらかの収納家具である closet や cupboard や shelf や cabinet や bookcase でいずれの場合もドイツ人の方がより典型的だと判断している．また，電化製品(stereo, lamp, TV)についてはアメリカ人のほうはまあ家具として認知しているようであるが，ドイツ人は家具とは見なしていないことも上の結果から読み取れる．それから，ベットであるが，アメリカ人にとってもドイツ人にとっても典型的な家具である点には違いがないが，ドイツ人にとっては bed は table と並んでもっとも典型的な家具であるのにたいして，アメリカ人のほうはそれほどでもないという結果になっている点が目につく．closet と cupboard と shelf については，「収納家具」という観点以外にも家具の可動性[2]に関する判断のずれがドイツ語と英語のあいだにあると思われるが，詳細は注で述べることにして，ここでは SORT がこのような数値データの整理に大いに役立つものだということを確認して，先に進みたい．

1．2 SORT と UNIQ で語彙や文字の頻度をしらべる方法

SORT と UNIQ の-c オプションでテキストファイルの各行の頻度が簡単にしらべられる．1行めが B，2行めが A，3行めが B とならんだテキストファイルを考えてみよう．B-A-B は SORT で並べかえると A-B-B になる．さらに UNIQ に-c オプションをつけてソート済みファイルを処理すると，連続する同一行をかぞえ，A や B の頻度をしらべ，1 A や 2 B のように同一行数を前につけて打ち出してくれる．UNIQ の行頭の数字の付け方は，通常，行頭にいくつかスペースをおいてから数字を出力し，その後にまたスペースをいくつか入れるようになっているようだ．ただし，次の例にあるように数字をかっこ付きで出力するものもある．

抽象的な説明だけでは分かりづらいと思うので，3つのステップを具体例で示しておこう．

a．行の内容が数えたい内容になっているテキストデータを用意する．

```
ja,ja,nein,nein
ja,ja,ja,nein
ja,ja,nein,nein
ja,nein,ja,ja
```

b．SORT で並べかえ，同一行が連続するようにする（2 行めと 3 行め）．

```
ja,ja,ja,nein
ja,ja,nein,nein
ja,ja,nein,nein
ja,nein,ja,ja
```

c．UNIQ -C で同一行をまとめ，同一行が何行あったか計算させる．

```
[1]ja,ja,ja,nein
[2]ja,ja,nein,nein
[1]ja,nein,ja,ja
```

簡単なアンケートの集計のようなことをしているわけだが，行単位で完全に同一の行だけをかぞえていて，たとえばコンマ区切りデータの 2 つめに ja がいくつで nein がいくつというような各項目ごとの集計はできない[3]．そういう目的なら AWK という使い方のやや難しいツールなどを使う必要がある．とはいえ，同一行をかぞえてくれるだけでも語彙や文字の頻度などをかぞえるには十分な機能である．

文字（列）や語彙の頻度調査のポイントは，行の内容がかぞえたい単位になるようなデータをつくることである．通常のテキストデータから一行一単位のデータをつくるのに役に立つのがやはり汎用テキスト処理ツールの FOLD や CUT や REV や WORD などである．汎用テキスト処理ツールを使った文字頻度の調査法を簡単にまとめておこう．

テキストファイルであれば何であれ行を指定した桁数で折り返してくれるのが FOLD だ．ふつうは，-70 などを指定して 70 桁で文章を折りかえすのに利用するツールだが，-1 を指定すれば 1 桁めで折りかえしてくれ，テキストデータを一行一字の文字リストに変換することができる．たとえばファイル test.dat の内容が「Tokio」で「fold -1 test.dat > fold.dat」とプロンプトで打ち込めば，fold.dat の内容は次のように一行に一字ずつになる．

T
o
k
i
o

あとは，上で見たように，SORT と UNIQ で文字頻度が計算できる．もちろん，test.dat の中身は「Tokio」のように短いものでなく，長ければ長いほどコンピュータのありがたみが分かる．

また，すでに一行一語のデータが作成してあって，語頭の文字の頻度がしらべたいような場合もあるだろう．この場合は CUT というツールで-c1 と指定すれば一字めだけが切り出せ，語頭の文字頻度がしらべられる．また，語頭ではなく語末の文字の頻度がしらべたいなら，まず REV というツールで各行の内容を反転させて語末の文字を語頭に持っていってから CUT で一字めを切り出せばよい．

文字頻度の調査の具体例は掲載しないが，実際に約 5.5 万語の辞書の見出し語データ（[3] 参照）や約 100 万字のフロッピー版実用文例集のテキストデータで文字頻度をしらべた結果が城岡（1996）に出してある．

文字頻度ではなく，語彙の頻度がしらべたいなら，テキストファイルを一行一語に変換するツールがいくつか出回っているようである．WORD という名前になっていたりするが，この種の単語抽出ツールは研究目的の使用にはあまり向かないと思う．なぜなら，単語の識別はかなり難しい面もあって，うまく単語に切り分けられなかったり，ひとつの単語が分断されてしまったりするからである．というのも，WORD などの自動単語切り分けツールの標準の単語の定義は「英字か "_" で始まり英数字か "_" の続いたもの」という英語かつプログラマ向けの仕様になっているらしく，たとえばウムラウトを Schlo¥s のように¥をつけてあらわしていると，これは一語とは認識されず，¥は削除され，Schlo と s の 2 語に分割されてしまうのである．したがって，研究目的の厳密な仕事に使うなら，原始的な方法だが，テキストエディタなどでスペースを改行に一括置き換えしたうえで（WORD でも-s オプションを指定すれば可能），対象のテキストデータの内容をその都度実際に見ながら不要な記号を削除したり改行に置き換えたりして一行一語の語彙リストを整理したほうが確実だろう．ただし，テキスト中の全スペースを改行に一括置き換えした場合でも，余計な

削除がされないのはいいのだが，ピリオドやコンマなどはそのまま残されてしまうからあとでこれらを一括削除する必要がある．また，たとえば「z. B.」が真ん中にスペースをあけて書いてあると，これは当然2語に分析されてしまうからこういう切り分け過ぎの例を探し出して対応する必要もある（もっともz. B. の場合はもともと2語からなる略語で，1単位なのか2単位と見なすべきなのか判断の難しい場合なのだが）．反対に der/die などと斜線をはさんで続けて書かれている個所があれば，全体が1語という扱いになる．したがって，どういう方法で文章を単語に切り分けたとしても，厳密に調査する場合は後処理が欠かせないようだ．

アメリカのパソコン通信の Compuserve 経由で（サービス内容は同一ではないがインターネットでも提供されている）SPIEGEL の記事が現在入手可能だが，これはテキストファイルとして保存可能だ．週刊誌 Spiegel の 1996 年 29 号から 33 号までの 5 冊分の記事をあつめると，大小 74 本の記事になった（現状ではペーパー版のすべての記事が提供されているわけではなくセレクトされている）．合計すると約 9 万 3000 語のデータになった（[3.1] 参照）．必要な処理をほどこして，頻度順の語彙表を作成してみたのが次のページの上位 60 位の語彙表である．一番頻度の高かったのは die でテキスト中に 2922 回出てきている．なお，文頭の大文字はそのままにしたので die と Die を区別している．だから 18 位のところに Die がふたたび出てきている．文頭の大文字を本来小文字のものは小文字にして計量したい場合は元のデータに手を入れる必要があるが，かなり面倒な作業になりそうだ．

■ SPIEGEL の語彙の頻度順生起度数表（文頭の大文字は大文字として計量）

1位：die	(2922)	21位：er	(528)	41位：Sie	(271)
2位：der	(2829)	22位：eine	(522)	42位：über	(270)
3位：und	(1784)	23位：als	(502)	43位：sind	(269)
4位：in	(1427)	24位：es	(462)	44位：einen	(268)
5位：den	(1201)	25位：aus	(445)	45位：haben	(266)
6位：das	(887)	26位：an	(407)	46位：vor	(253)
7位：zu	(871)	27位：auch	(396)	47位：SPIEGEL	(248)
8位：mit	(869)	28位：wie	(378)	48位：werden	(243)
9位：sich	(793)	29位：nur	(373)	49位：Der	(238)
10位：von	(762)	29位：hat	(373)	50位：ich	(230)
11位：nicht	(733)	31位：daß	(364)	51位：wird	(226)
12位：ein	(705)	32位：war	(355)	52位：zum	(220)
13位：ist	(672)	33位：so	(318)	53位：am	(204)
14位：im	(620)	34位：nach	(316)	54位：hatte	(202)
15位：auf	(605)	35位：noch	(315)	55位：oder	(200)
16位：sie	(594)	36位：Das	(280)	56位：schon	(194)
17位：dem	(561)	37位：einer	(274)	57位：aber	(187)
18位：für	(556)	38位：um	(271)	58位：mehr	(177)
19位：des	(556)	39位：einem	(271)	59位：sein	(161)
20位：Die	(556)	40位：bei	(271)	60位：In	(156)

実は，ここで述べたような頻度調査は AWK だけでも実現できる．文字頻度については植村・富永(1993)に symbol.awk というスクリプトがあるし，Aho/Kernighan/Weinberger (1989) や Stallman 他 (1993) に語彙頻度を計算するスクリプトの例が出ている．しかし，研究目的での使用を考えると既に述べた単語認定の問題があるから，やはり，AWK は使わずに言わば手作業で単語を切って，SORT と UNIQ を使って段階的にしらべるほうが時間はかかるが正確だろう．

1．3 一行一単位のリストを COMM で比較する方法

なんらかの共通語彙を出すのにコンピュータを利用するケースはこれまで多

かったように思う(植田(1982)，菊池(1983)，米井(1986)，森(1991))．COMMという汎用ツールを使えば2つの語彙リストが簡単に比較できる．共通語彙だけを求めるなら，3つ以上の語彙リストでも作業を繰り返すだけだ．実際に2つの語彙リストを比較してみよう．重藤(1990)の301語のリストをテキストファイルで一行一語の形式で入力し，ファイル名はshigetou.datとしておこう．重藤のリストに対抗するかたちで発表された近藤・川崎（1991）の324語（女性形も別の語として計算)のリストはkonkawa.datとしよう．いずれもドイツ語教育のための最重要語彙を策定するという背景のもとで発表されたものだ．COMMで比較するには，一行一語でつくったテキストファイルのリストの各行をあらかじめソートしておく．ふたつのリストの共通部分や差異部分をしらべるには，

comm shigetou.dat konkawa.dat > kekka.dat

とするだけである．重藤だけにある語彙，近藤・川崎にだけある語彙，共通語彙の3つがkekka.datに書き込まれている．重藤だけが110語で，共通語彙が191語で，近藤・川崎だけが133語だった．近藤・川崎の語彙リストが重藤の語彙リストに対抗して発表された経緯を考えると191語という共通語彙は意外に多い感じがする．中身を比較すると，近藤・川崎は「学問的な所産としての基礎語彙」ではなく「意志表示の手段」としての「共通基本単語」を模索しているわりにはButter, Eis, Fisch, Geld, Name, Wasser, freiなどが欠けている．また，近藤・川崎にあって重藤にない語彙[4]にはArzt, Ausstellung, Bus, Café, Durst, Hunger, Kaufhaus, Konzert, Museum, Rathaus, Platz, U-Bahn, billig, falschなど，たとえばドイツ旅行などの際には知っていてよかったと言えそうな語彙が目立つ．

■ 重藤（1990）と近藤・川崎（1991）に共通する191語

Abend, Auto, Berg, Bett, Bier, Bild, Blume, Brief, Brot, Bruder, Buch, Eltern, Familie, Fenster, Frau, Freund, Garten, Glas, Haus, Heft, Herr, Jahr, Junge, Küche, Kaffee, Kind, Kino, Kirche, Kuchen, Land, Leute, Mädchen, Mann, Meer, Milch, Monat, Morgen, Mutter, Nacht, Onkel, Schuh, Schwester, Stadt, Straße, Student, Stunde, Suppe, Tag, Tante, Tee, Tisch, Tochter, Uhr, Vater, Wagen, Wein, Wetter, Zeit, Zimmer, Zug, über, aber, als, alt, an, antworten,

arbeiten, auch, auf, aus, bald, bei, bleiben, dürfen, da, daß, danken, dann, denn, doch, dort, durch, einmal, essen, für, fahren, finden, früh, fragen, geben, gehören, gehen, genug, gerade, gern, gestern, glauben, gleich, groß, gut, hören, haben, halb, halten, heißen, helfen, heute, hier, hoch, immer, in, ja, jetzt, jung, kalt, kaufen, kennen, klein, kommen, krank, kurz, lang, lassen, laufen, legen, lernen, lesen, letzt, liegen, mögen, müssen, machen, morgen, nächst, nach, nehmen, nein, neu, nicht, noch, nur, ob, oder, oft, regnen, sagen, schön, schlecht, schnell, schon, schreiben, schwer, sehen, sehr, seit, setzen, sollen, spät, spielen, sprechen, stehen, studieren, teuer, tragen, trinken, tun, um, und, verstehen, viel, vielleicht, von, vor, während, wann, warm, warten, warum, weil, weit, wenn, werden, wie, wieder, wissen, wo, wohnen, wollen, zeigen, ziehen, zu

■ 重藤（1990）にあって近藤・川崎（1991）にない 110 語

Apfel, Arm, Auge, Baum, Butter, Dorf, Ei, Eis, Erde, Feld, Feuer, Fisch, Fuß, Geld, Haar, Hand, Herz, Himmel, Hund, Katze, Klasse, Kleid, Kopf, Kuh, Licht, Luft, Mensch, Mond, Name, Papier, Rose, Schiff, Schnee, Seite, Sonne, Stück, Stein, Tür, Tier, Vogel, Volk, Wald, Wasser, Weg, Wind, öffnen, all, also, ander, arm, beide, bis, brechen, dick, einander, einige, erst, etwas, fallen, fliegen, frei, ganz, gegen, gesund, heiß, hell, her, hinter, jeder, jemand, kaum, klug, lachen, leben, müde, nah, neben, nichts, nun, ohne, plötzlich, recht, reich, reisen, rufen, schlafen, schwach, sich, sicher, singen, sitzen, so, sondern, stark, statt, sterben, tanzen, tief, tot, trotz, unter, wahr, was, waschen, weg, wer, wohl, zurück, zusammen, zwischen

■ 近藤・川崎（1991）にあって重藤（1990）にない 133 語

Anzug, Arzt, Ausflug, Ausland, Ausstellung, Bahnhof, Bibliothek, Bluse, Brille, Bus, Café, Chemie, Durst, Elektronik, Fahrkarte, Fahrrad, Ferien, Fieber, Film, Firma, Fluß, Fräulein, Freundin, Fußball, Geburtstag, Geschäft, Geschenk, Geschichte, Gitarre, Hemd, Hunger, Jura, Kaufhaus, Klavier, Konzert, Kopfschmerzen, Kugelschreiber, Lehrer, Lehrerin, Mensa, Mittag, Motorrad, Museum, Musik, Nachmittag, Platz, Prüfung, Rathaus, Referat, Reise, Rock, Roman, Rundfahrt, Schüler, Schülerin, Schauspiel, Schule, See,

Seminar, Sohn, Studentin, Stuhl, Tasche, Tasse, Theater, U-Bahn, Universität, Unterricht, Vormittag, Wand, Woche, Wohnung, Wurst, Zeitschrift, Zeitung, abholen, anfangen, ankommen, anrufen, aufstehen, aussehen, bekommen, bestehen, besuchen, billig, bitten, brauchen, bringen, dauern, denken, donnern, einladen, entschuldigen, erinnern, falsch, fernsehen, freuen, freundlich, gefallen, interessant, interessieren, können, kennenlernen, kochen, kosten, langsam, leicht, leider, lieben, links, mit, rauchen, rechts, richtig, ruhig, schenken, schicken, schmecken, schwimmen, sein, stattfinden, stellen, teilnehmen, treffen, umsteigen, unterhalten, verheiratet, vorstellen, wünschen, wandern, wenig, woher, wohin

ふたつの語彙リストから共通部分と差異部分を出したわけだが，たったこれだけのことでもパソコンとCOMMがなければかなり大変な作業になってしまうに違いない．

2．検索ツールを利用する

もとはUNIXのGREPを意識してつくられたテキストデータ検索ツールにGREPがある．名称のどこかにGREPが含まれるツールがMS-DOS用やWindows用としてかなり出回っている．私の手元にあるものでは，AGREP, CGREP, DDJGREP, DGREP, FGREP, FZGREP, GREP32, GREPFV, MGREP, PSGREP, QGREP, QTGREP, SGREP, VGREP, WGREP, XGREP, YGREPなどがある．加えて，同等のツールでも名称にGREPが含まれないこともあるわけだからパソコン通信やインターネットや雑誌の付録として出回っているテキストファイルから文字列を検索するツールの数はかなり膨大なものである．これだけ多くの検索ツールがつくられているのは，この種の検索ツールの需要の高さを示しているわけだが，また，供給過多とも思えるくらい提供されているのは，使用目的や検索の多様性に対応しているものと見ることができる．圧縮ファイルに対応しているもの，VZ Editorなどで可能なタグジャンプ（後述）に対応したタグ出力の可能なものなどあるから，自分の目的にあったものを探す必要がある．一般的に言って，この種の検索ツールでは正規表現[5]が使えるものとそうでないものがあるが，正規表現の使えるGREPでは正規表現の本領を発揮するような使い方をしたい．たんなる語句の検索ならKWIC形式の出力が可能なもののほうが語学研究には使いやすいだろう．KWICという

のは Key Word in Context の略でもともと IBM で開発された用語索引作成システムだったようだ（伊藤（1996））．この形式の出力はキーワードの位置を固定しているので出力データが見やすく，とくに大量のデータの場合はたんに該当行を表示する GREP にはない見通しのよさが有り難い．正規表現は扱えないが，KWIC 形式で出力してくれる検索ツールに浜口崇さんの KKC がある．語学研究・学習用に開発されたソフトで，MS-DOS 版はフリーソフトだが，Windows 版がシェアウェアとなっている．

viel とともに単数形で使われる名詞について考えてみようとしたことがある．実証的な研究には用例が欠かせない．そもそもどんな名詞が単数で無変化の viel と使われるのか．次にあげるのは，テキストファイルにした Universalwörterbuch[6]を対象に KKC で viel を検索した出力の一部だ．

■ KKC の出力例，テキストファイル版 Universalwörterbuch で viel を検索

（左下がりの斜線は元のデータに改行があったところで，右下がりの斜線はその次の文字とふたつで一組でウムラウトやエスツェットをあらわしている）

```
UNI.A:  26840 | nh\aufen <hat>: der/Wind hat viel Schnee, Sand, viele Bl\atter
UNI.A:  26841 | aufen <ist>: hier weht immer viel/Sand an.//anweisen/<st. V.;
UNI.A:  26875 | e List, Gewalt a.; wir haben viel Sorgfalt,/M\uhe auf die Sach
UNI.B:   4781 | t); c) ben\otigen, brauchen: viel Raum, Zeit b.;//Beanspruchun
UNI.B:   5461 | etwas b.; sein/Name bedeutet viel in der Fachwelt; Geld bedeut
UNI.B:   5811 | it dem Akk.:> dazu bedarf es viel Geld;//Bed\urfnis,/das; -ses
UNI.D:   1890 |  Angriff nehmen: es ist noch viel Post zu erledigen,/ich werde
UNI.D:   1896 | einsetzen,/aufbieten: er hat viel M\uhe, Zeit darangewandt, di
UNI.D:   3017 | viele Geld (er bedauerte, so viel Geld ausgegeben zu/haben).//
UNI.F:  13134 |  Wein schmeckt f. (hat nicht viel Geschmack). 4./(abwertend) o
UNI.F:  13199 | //flachk\opfig/<Adj.>: nicht viel Geist besitzend; geistlos;//
UNI.F:  13771 | immer wieder zur F. greifen (viel Alkohol/trinken, Alkoholiker
UNI.G:  28879 | ampfung von Sch\adlingen. 2. viel Raum bietend od./beanspruche
UNI.G:  29041 | o\sverdiener,/der: jmd., der viel Geld verdient, der ein gro\s
UNI.G:  30984 |  Pl.> (ugs.) Verstand: nicht viel G. haben.//Gr\utzwurst,/die
UNI.H:   6168 | fer,/der (salopp): jmd., der viel Schnaps trinkt;//hartumk\amp
UNI.H:   6316 | in alter H. sein (ugs.; sehr viel Erfahrung/[in einer bestimmt
UNI.H:   6640 | ube: die H. schlie\st nicht; viel Kraft/unter der H. haben ([v
UNI.I:   3239 | h, da\s; damit, da\s: er hat viel/Geld sparen k\onnen, i. er e
UNI.I:   4184 | dustriestadt,/die: Stadt mit viel Industrie;//Industrieunterne
UNI.I:   9787 | : in etw. seine ganze Kraft, viel Zeit i.; in/jmdn. sein Gef\u
```

出力結果をみると，viel が必ず一定位置（用例部分の中央）に来るようになっていて，viel の部分にすぐ目がいき，viel について他の語との結びつきをしらべるには便利なかたちだ．これが KWIC 形式である．UNI.A というのはファイル名で，一行めの冒頭の「UNI.A: 26840」という部分は UNI.A の 26840 行めという意味であり，これがいわゆるタグ出力と呼ばれるもので，対応したエディタソフトだったらこの情報をもとに一定の操作でファイル UNI.A を自動的に開

いてくれ，該当個所までジャンプしてくれる．このタグジャンプ機能は，検索ツールの出力行がふつう一行なので，前後の文脈が見たい場合には非常に有効な機能だ．

用例を探す目的なら用例辞典も使える．現代ドイツ語の用例辞典で代表的なものといえばDuden Stilwörterbuchと東ドイツから出ていたWörter und Wendungenがあるが，vielと単数の名詞の組み合わせをしらべてみると，それぞれvielの項目にある例文の中からさがして，Aufheben, Blut, Geld, Gemüse, Glück, Humor, Liebe, Mühe, Milch, Spaß, Vergnügen, Wesen, Zeit, Wille, Schmuckの合計15種類の名詞が語尾の無いvielと使われていた．おそらく，辞書の中をくまなく探すことができれば，viel以外の項目の解説文や例文にvielと単数名詞の例がもっともっと見つかるはずだ．しかし，よほど暇を持て余しているのでなければ辞書の中に散らばっている用例をしらみつぶしに見ていくという方法は実用性がない．紙の辞書の限界だ．上記のKKCとUniversalwörterbuchでしらべた結果は，2冊の用例辞典あわせても15例しかなかったのに，なんと一冊の辞書で163種類の「語尾をとらないviel＋単数名詞」が見つかった．以下がその163語だ．

■ 無語尾のvielと使われていた163語の単数名詞

Alkohol, Alkoholgenuß, Anlauf, Applaus, Arbeit, Aufheben, Aufwand, Ausdauer, Ausdruck, Beifall, Benzin, Betrieb, Bewegung, Blut, Blutvergießen, Bravour, Brimborium, Busen, Butter, Champagner, Druck, Einsatz, Elan, Energie, Erde, Erfahrung, Erfolg, Federlesen, Feingefühl, Fett, Fettgewebe, Feuer, Fleiß, Fleisch, Freude, Futter, Gas, Gaudi, Geächze, Gedöns, Geduld, Gefühl, Geist, Gelaufe, Geld, Gemüt, Gepäck, Geräusch, Geröll, Gerede, Geschick, Geschirr, Geschmack, Geschrei, Gesums, Getöse, Gewicht, Glück, Grütze, Gutes, Härte, Harz, Helligkeit, Holz, Ignoranz, Industrie, Information, Körper, Kümmel, Kaffee, Kalk, Kapital, Karamel, Kiki, Klamauk, Kleidung, Kohle, Komfort, Kompost, Kraft, Kram, Kritik, Kummer, Lärm, Lametta, Land, Langmut, Laub, Leid, Licht, Mühe, Macht, Mark, Meerrettich, Milch, Mut, Pech, Pfeffer, Pflege, Phantasie, Platz, Post, Power, Raum, Ruhe, Saft, Salz, Sand, Schönheit, Schatten, Scheiß, Schlacke, Schlacken, Schlackerschnee, Schlamm, Schmalz, Schmutz, Schnaps, Schnee, Schrott, Schußkraft, Schweiß, Schweres, Schwung, Sekt, Sinn, Sonne, Sorge, Sorgfalt, Spaß,

Sport, Staub, Stoff, Stroh, Sums, Tageslicht, Takt, Trubel, Umsicht, Unsinn, Unverstand, Verdruß, Vergnügen, Verkehr, Verständnis, Verstand, Verve, Vieh, Wärme, Wasser, Wert, Wesen, Wucht, Zündstoff, Zank, Zeit, Zeug, Zucker, Zuneigung, Zuwendung, Ärger, Öl, Übung

私の関心は，viel とともに単数形でつかわれる上記の名詞の複数形がそもそも可能なのかどうか，また可能だとしてどのような場合につかわれるのかということである．まだ，結論は出していないが，KKC のおかげで具体的に問題を考える準備はできたわけである．

3．AWK を使ったやや高度なテキスト処理

AWK は検索や削除や変換だけでなく計算機能も備わった複合ツールで，使いこなすのが上で述べた他のツールとくらべてやや難しいが，語学研究でテキストデータを扱う際には極めて有効な手段になる．複数の検索が一度にできるだけでなく，集計作業も同時にこなすことができるし，特定の文字（列）を削除したうえで削除後の語形をかぞえることも可能だ．私自身の取り組みから二つの事例を紹介してみよう．いずれも調査の対象として三冊の辞書から取り出して整理したデータを利用しているので，これについて手短かに説明しておこう．これは 3 冊の電子ブック（「クラウン独和辞典」，Langenscheidts Eurowörterbuch Italienisch，Langenscheidts Data Disc Wörterbuch Französisch）から抜き出したドイツ語の見出し語を一行一語に整理したデータ（詳しくは城岡（1996）を参照）で現時点で 55464 語になっている．ここでは jisho.dat と名付けておく．

3．1 ドイツ語のつづりの読み方の規則を頻度から検討する

文法項目の扱いはともすれば網羅的になってしまう．文法には空欄のない表が多いのはそのためだろう．私は城岡（1994）で動詞の人称変化形の頻度を各種のテキストで調査して頻度のいちじるしい差を確認している．つづりの扱い方も一般にかなり網羅的である．

ドイツ語では dt というつづりはたとえば[t]という発音をする．これは規則として覚えないとひとりでに読めるようにはならない．だから，教科書などではこのようなローマ字とは異なるドイツ語固有の読み方をする文字列（以下これを学習つづりと呼ぶ）の読み方を最初に学習することになっている．しかし，

学習つづりはすべて同じように重要なのだろうか．2字以上の学習つづりの主なものについてjisho.datをAWKで処理して，頻度をしらべてみた．調査では語中の位置による区別はしなかった．だから，spやstは語頭のものもあれば語中や語末のものも混じっている．また，音節内，音節間という区別もしていない．たとえばdtを[t]と発音するのは同一音節にdtがあるときだけで，Bestandteilや Landtagのように音節をまたがる場合にはこの規則はあてはまらない．だから，dtの読み方の規則を厳密に考える場合は，音節間のものは除外するなり，別にかぞえるなりしたほうがいい．しかし，これを簡単に実現する方法は思いつかなかったので行っていない．作業に見合うだけの結果が得られるかどうか分からないが，このような簡便な調査ではなく本格的な調査が必要であろう．なお，下の結果はAWKの出力をSPLITというツールで3つのファイルに切り分けてからPASTEというツールで横につないだものをもとにしている．また，ここでかぞえているのは語数であって，つづりの出現回数ではない．したがって，erが一語に4回出てくるSonderberichterstatterも一語としかかぞえていない．

■ 学習つづりは辞書の語彙にどれだけ含まれているか（55464語中）

er	17299語(31.190%)	tz	1555語(2.804%)	ds	426語(0.768%)
ch	14565語(26.260%)	nk	1482語(2.672%)	ph	381語(0.687%)
ei	8890語(16.028%)	ah	1385語(2.497%)	qu	350語(0.631%)
sch	7862語(14.175%)	eu	1317語(2.375%)	uh	321語(0.579%)
st	7358語(13.266%)	pf	1068語(1.926%)	aa	250語(0.451%)
ng	6411語(11.559%)	chs	666語(1.201%)	ih	200語(0.361%)
ie	5537語(9.983%)	oh	660語(1.190%)	ai	191語(0.344%)
au	4308語(7.767%)	th	657語(1.185%)	ps	165語(0.297%)
ig	3558語(6.415%)	rh	589語(1.062%)	dt	163語(0.294%)
ck	2194語(3.956%)	tsch	588語(1.060%)	öh	127語(0.229%)
ss	2129語(3.839%)	üh	489語(0.882%)	oo	79語(0.142%)
sp	1910語(3.444%)	ee	460語(0.829%)	ay	28語(0.050%)
ts	1884語(3.397%)	äu	451語(0.813%)	ey	15語(0.027%)
eh	1871語(3.373%)	äh	427語(0.770%)		

31.190%の er から 0.027%の語彙にしか出てこない ey までつづりの頻度差はいちじるしい．あきらかに重要な学習つづりとそれほど重要でない学習つづりがあるようだ．それから，結果をながめていて気になったのが rh である．この学習つづりは扱わない教材も多いが，かなり頻度は高いようだ．qu や dt や ai などのほうが頻度は低い．もちろん，ここで計量しているのは辞書の見出し語データである．ふつうの文章での頻度とは違うかもしれない．ふつうの文章では短い語が繰り返し使われるわけで学習つづりのパーセントは全体的にかなり下がるはずだ．また，頻度の低い外来語に典型的な学習つづりの頻度や順位も下がるだろう．反対にふつうの文章では頻度や順位が上がりそうな学習つづりも予想される．日常よく使われる語彙がドイツ語の語彙としてむしろ奇妙な作り方をしているという ih のような例があるからだ．ih を含む語は jisho.dat のデータで 0.361%の語彙にしか含まれていなかった．しかし，この中には ihm や ihn や ihr や Ihr や ihnen や Ihnen などの基本的語彙がはいっているわけだ．

次にあげるのは，[1.3]でも利用した雑誌 SPIEGEL のデータを対象に学習つづりについて調べた結果だ．SPIEGEL のデータについて補足しておくと，つづりの読み方とは関係のない数字や略語や記号などはまず削除した．同じことは jisho.dat のデータ作成の際にもおこなっているから，基準をそろえたわけだ．整理後の語数は延べ 92952 語になった．

■ SPIEGEL における学習つづりの頻度順生起度数表（延べ 92952 語中）

(a) 生起度数　　(b) 語彙総数（延べ語数）に対するパーセント

(c) jisho.dat での順位　　(d) 順位の差

(e) jisho.dat との順位差が 5 位以上の場合の jisho.dat におけるパーセント

		(a)	(b)	(c)	(d)	(e)
1.	er	18749語	(20.171%)	: 1位	0	
2.	ch	13455語	(14.475%)	: 2位	0	
3.	ie	9595語	(10.323%)	: 7位	＋4	
4.	ei	9453語	(10.170%)	: 3位	－1	
5.	st	5735語	(6.170%)	: 5位	0	
6.	sch	4771語	(5.133%)	: 4位	－2	
7.	au	4128語	(4.441%)	: 8位	＋1	
8.	ng	3637語	(3.913%)	: 6位	－2	

9.	ig	2093語	(2.252%) : 9位	0	
10.	eh	1867語	(2.009%) : 14位	+ 4	
11.	ts	1504語	(1.618%) : 13位	+ 2	
12.	sp	1403語	(1.509%) : 12位	0	
13.	eu	1387語	(1.492%) : 18位	+ 5	(2.375%)
14.	ss	1365語	(1.468%) : 11位	− 3	
15.	ah	1116語	(1.201%) : 17位	+ 2	
16.	tz	1073語	(1.154%) : 15位	− 1	
17.	ck	903語	(0.971%) : 10位	− 7	(3.956%)
18.	ih	883語	(0.950%) : 34位	+16	(0.361%)
19.	nk	853語	(0.918%) : 16位	− 3	
20.	tsch	642語	(0.691%) : 24位	+ 4	
21.	oh	496語	(0.534%) : 21位	0	
22.	rh	473語	(0.509%) : 23位	+ 1	
23.	th	437語	(0.470%) : 22位	− 1	
24.	pf	411語	(0.442%) : 19位	− 5	(1.926%)
25.	chs	364語	(0.392%) : 20位	− 5	(1.201%)
26.	üh	353語	(0.380%) : 25位	− 1	
27.	äh	328語	(0.353%) : 28位	+ 1	
28.	ai	257語	(0.276%) : 35位	+ 7	(0.344%)
29.	aa	230語	(0.247%) : 33位	+ 4	
30.	ds	213語	(0.229%) : 29位	− 1	
31.	äu	210語	(0.226%) : 27位	− 4	
32.	ee	208語	(0.224%) : 26位	− 6	(0.829%)
33.	dt	152語	(0.164%) : 37位	+ 4	
34.	ph	143語	(0.154%) : 30位	− 4	
35.	uh	129語	(0.139%) : 32位	− 3	
36.	qu	111語	(0.119%) : 31位	− 5	(0.631%)
37.	öh	108語	(0.116%) : 38位	+ 1	
38.	ay	93語	(0.100%) : 40位	+ 2	
39.	oo	84語	(0.090%) : 39位	0	
40.	ps	72語	(0.077%) : 36位	− 4	
41.	ey	62語	(0.067%) : 41位	0	

ihについては予想どおり辞書データでは34位だったものが18位になっている．また，外来語に出てくるつづりが軒並み順位を下げていることも確認できる．

th：22位 → 24位	ph：30位 → 34位
qu：31位 → 36位	ps：36位 → 40位

しかし，多少の順位の異同はあるが，辞書の見出し語データの結果とくらべて著しくことなる結果になっているわけではない．これは相関係数を出してみるとはっきりする．要素の順位だけをもとにふたつの項目間の相関をもとめるスピアマンの順位相関係数（A. Kenny (1996) pp. 92, 93 参照）というのがある．jisho.datとSPIEGELの結果は0.941289になる．したがって，ふたつの結果のあいだには強い正の相関があるという統計的結果になる．

さて，2種類のデータをもとに学習つづりの頻度を出してみた．結果はどうだろうか．一般に文法での学習つづりの扱い方は重要度の区別がつけられることはまずない．これは動詞の人称変化形の扱いが頻度を無視して一般に各人称変化形の扱いが同じなのと通じている．しかし，動詞の各人称変化形の頻度が著しく差があり，重要な3人称単数形とほとんど使われていない2人称複数親称形があるように，学習つづりにも頻度のきわめて高いものと低いものがあることが確認できた．この結果をもとに学習つづりの教育方法について考えることも可能だろう?)

ちなみに，学習つづりについてではなくあらゆる可能な2文字の部分文字列の頻度をjisho.datとAWKでしらべると，もっとも頻度の高いのがerで，その次はchではなく，ドイツ語に固有の読み方はしないので学習つづりではないenが来るようだ[8] 次に頻度が1位のerから60位のnaまであげておく．なお，ここでは一語に複数回出現する場合もそれぞれかぞえ，語数ではなく，回数をもとに順位を出した．

■ 頻度の高い2文字の部分文字列（1位から60位まで）

1		5			10				15									20
er	en	ch	ei	sc	te	st	ge	re	un	ng	in	be	el	ie	he	an	le	nd	au
de	se	it	ra	is	li	ri	ti	ig	ic	ar	la	al	es	ke	ne	ha	ve	me	ns
nt	at	or	tr	fe	us	ll	ta	rt	we	on	rs	hr	ck	ze	ss	ht	ru	hl	na

ところで，jisho.datのデータはウムラウトなどを¥を付加する独文学会データベース委員会方式であらわしているのだが，この書き方だと2文字のöhが¥ohの3文字になってしまうし，ohが本当は含まれていないのに含まれていることになってしまう．実は，jisho.datやSPIEGELの学習つづりの調査でもohの頻度に¥ohのものが含まれないようにするために/(^|[^¥¥])[Oo]h/という正規表現を使って工夫していたのである．正規表現に慣れないひとがみるときわめて複雑に見えると思うが，「行頭のOhあるいはoh，行頭にない場合は直前に¥ではない文字が付いているOhあるいはoh」という意味である．2文字の部分文字列の調査では計量に植村・富永（1993，p. 200）のdegree.awkというスクリプトを一部変更して利用した．¥ohなどを2文字としてかぞえるために特殊文字は未使用記号一字に割り当てる方式でしらべた．たとえば¥ohは$hとしてしらべたからちゃんと2文字の部分文字列として扱うことができるわけだ．もっとも，結果を見れば分かるが，ウムラウトやエスツェットなどの特殊文字の入った文字列は60位以内には登場していない．

3．2 ドイツ語でb/w, l/rの一字違い語の分布を調査する

HundとBundは語頭の一字だけで区別されているし，dannとdenn，wannとwenn，denkenとdankenは二字めの文字でだけ区別されている．一字違い語はErsterとElsterのように他の部分の発音の違いから音素のミニマルペアとは対応しない場合もあるが，たいていはミニマルペアでもある（bittenとbietenのように一字違い語ではないミニマルペアも存在する）．ミニマルペアは音韻論で音素を確定する際には重要な概念であるが，理論を離れても，一般に紛らわしい語として要注意であるし，外国語教育の立場からは習得困難な外国語の音素の区別（日本人なら/l/と/r/や/b/と/v/など）を学習する際に聞き取り練習や発音練習が対立する音の区別を学習するのに効果的だとされている（ミニマルペア・ラーニング）．AWKを使うと一字違い語の調査は容易だ．たとえば2字めがlとrの違いだけの一字違い語をjisho.datの5万5千語の語彙リスト[9]から抽出するには，まず，正規表現/^.[lr]/で2字めがlかrの行だけを処理対象として絞り込む．次に2字めを削除した語形「substr($0, 1, 1)substr($0, 3)」を記録し，語彙リスト中の他の行に同じ語形が出てくればかぞえるようにする．lとrの一字違い語があれば最終的に同じ語形が2回記録されているはずだ．具体例を添えておくと，blauはbauになり，brauも同じbauになり，最終的にbauという語形が2回記録されるということである．こういう考え方

でAWKのスクリプトをつくりlとrの一字違い語とbとw（ドイツ語としての発音は[v]）の一字違い語をしらべてみた．

■ b/wとl/rの1字めから9字めまでの一字違い語のペア数

	1	2	3	4	5	6	7	8	9	合計
b/w:	59-	00-	10-	12-	00-	01-	00-	00-	00	82ペア
l/r:	87-	38-	38-	44-	27-	36-	15-	14-	04	303ペア

結果を見ると，面白い点がふたつある．まず，語頭の一字違いという意味はbとwでもlとrでもどちらの場合も大きいらしく，語頭の一字違い語のペアが圧倒的に多いことに気づく[10]もうひとつ面白い事実はb/wとl/rの語頭以外での分布の違いである．bとwでは語頭以外の一字違い語は-biegenと-wiegenをふくむ複合動詞だったり，TischbeinとTischweinやGewißやGebißのような複合名詞だった．つまり，すくなくとも形態素の先頭でなければbとwの区別はドイツ語で利用されていないということである．語頭から9文字めまでの一字違い語をしらべたが，通常の語彙に例外はない．ミニマルペアとしての一字違い語ではErbinと人名のErwinが唯一の例外だ．また，発音の関係でミニマルペアでないものならJubelとJuwelがあった．一方，lとrの一字違い語は語頭ほどでないにしても語中でもかなり利用されている事実が上の表から分かる．lとrの一字違い語でしかも音素のミニマルペアになっている例をひとつずつあげておこう．

■ l/rの一字違い語の例

1字め：Leiche, Reiche
2字め：Flucht, Frucht
3字め：melken, merken
4字め：spülen, spüren
5字め：Schal, Schar
6字め：Schmelz, Schmerz
7字め：schwielig, schwierig
8字め：Salzsäule, Salzsäure
9字め：Nachtmahl, Nachtmahr

日本人に難しい発音の区別として[l]と[r]が有名だが，同じように区別の難しい[b]と[v]の区別についてそれほど指摘されないことが私にはこれまで不思議だったが，ここでのドイツ語の結果をもとに考えると，ミニマルペアの頻度や分布の違いが関係している可能性がありそうだ．lとrのほうがミニマルペア

の頻度がはるかに高い（l/r が合計 303 ペアに対して b/w が 82 ペア）からかもしれないし，l と r が語頭以外でも多用されている点に注目すると，語頭のミニマルペアと語中のミニマルペアの相対的な聞き分けの難易度の差ということにその理由があるかもしれない．

なお，ペア数の調査にあたっては，jisho.dat のデータをあらかじめすべて小文字に換えて整理した別のデータを作成した．こういう配慮をしないと，上で述べた単純な方法では Floh と froh や klug と Krug（2 字めを削除すると Foh と foh，kug と Kug）が一字違い語として検出できない．また，jisho.dat のデータをすべて小文字に換えると，Essen が essen になって最初からの essen と合わせて essen を内容とする行が重複してしまうが，これは UNIQ で重複行を一行にまとめて整理している．

4．最後に

実際に各種の汎用テキスト処理ツールを利用する場合は，たとえ同じ名前のツールであっても使い方が微妙に違っていることもある．したがって，実際に使用するにあたっては，それぞれのツールに添付されている説明に従うのは当然だ．また，2 バイトの日本語の文字に対応したものやそうでないものがあるということも知っておきたい．たとえば REV というツールは日本語に対応していないのがふつうで，REV で日本語を処理しようとしても当然うまく行かない．それから，同じ種類のツールがいくつもあるときには，テストしてみて，自分の目的にあったものを少なくとも自分の使用する目的ではバグが出ないということを確認してから使うようにするといいだろう．

注

1）SORTF は平仮名と片仮名を同一文字の 2 形態として並べ替えること（日常的には当たり前のことであるが，コンピュータにとっては「あ」と「ア」は本来コード番号のことなるまったく別の文字である）も可能な SORT ツールであるが，作者の豊島正之さんは人文系の研究者であるし，多数の語学調査用小プログラム集の MCL（Method in Computational Linguistics）の作者の中野洋さんも国語学の研究者である．また，本文でも紹介している KWIC 出力のテキスト検索ツール KKC の作者の浜口崇さんは英語学の研究者である．

2）closet や cupboard や shelf における著しい判断の違い（4.75，3.07，2.52）

は，おそらく収納家具がドイツ語で英語よりも高い家具性が付与されるということのほかに，「可動性」についての Möbel と furniture の違いを反映していると考えられる．ふつう典型的な「家具」とされるのは英語でもドイツ語でも建物や部屋の一部ではなく，内部の比較的大型の移動可能なモノである．「家具」というのは家や部屋の一部とまでは言えないまでもそれに準ずるものという中途半端な存在だとも言える．小さな灰皿や花瓶では家や部屋との関係が希薄になってしまい，Möbel や furniture とされにくいのはそのためであろう．Longman Dictionary of American English (1983) で furniture をしらべると，all large or quite large movable articles, such as beds, chairs, tables, etc. と movable が語義の一部にはいっている．ドイツ語の Möbel となると，語源は mobil やこの movable と同じでまさに「可動性」に由来しているから本来は「動かせる財産」の意味だったわけだし，Der Sprach-Brockhaus (1984, 9. Aufl.) をみると，bewegliches Einrichtungsstück eines Zimmers と，今日でも beweglich (「可動性」) が語義の一部にあると感じられているようである．とはいえ，Einbaumöbel や Einbauschrank という語があることからも分かるが，ドイツ語の Möbel では取り付けられたり，固定されたり，作り付けのものであってもちゃんとした家具ということになるようだ．アメリカ英語で closet というのは言ってみれば作り付けのドア付きの収納スペースである．小部屋や押し入れや物置のようなものとも言える．また，Longman の米語辞典にある図解をみると，キッチンの流しの上についている食器などを仕舞い込むための場所が cupboard の例としてあげてある．取り外し可能であっても，固定されていて，ふつう，動かされることはないだろう．shelf にしても，これは「棚板」のことであるから壁に固定して使われるわけだから英語ではやはり典型的な家具とは考えにくいのだろう．

3）各項目ごとの ja と nein の集計ではなく，たんに ja と nein の合計が数えたい場合は SORT と UNIQ で簡単に実現できる．まず，コンマを改行に置き換えて (エディタなどで一括置き換えできる) 一行に ja や nein がひとつずつ並ぶデータにして，それから SORT と UNIQ を使えばよい．ja が 10 個で，nein が 6 個という結果になる．

```
[10]ja
[6]nein
```

10 個の ja をかぞえるぐらいの作業なら，手作業でも難しくないが，処理する

データの量がふえると手作業では不可能になり，人間とコンピュータの処理能力の違いが顕著になる．

4）ここでふたつの語彙リストの優劣は論じるつもりはないが，「資料すべてに共通する語彙の抽出」という重藤の方法は語彙リストを比較する際の一般的な問題をふくんでいそうである．4種類の単語集すべてで最重要語とされていた語彙が重藤のリストであるが，こういう資料全体に共通する語彙を求めるやり方は植田（1982）や米井（1986）でも見られる．全体に共通する語彙を抜き出すということは，多数決ではなく全員一致でものごとを決める場合とおなじ弱点があると言えそうである．ひとりでも反対したら採用されないというのでは，ひとりの誤った判断でも全体に影響が出てしまうわけだ．おそらく，5種類の単語集から4種類以上で最重要語とされていたものを抜き出していたならより合理的な語彙リストになっていただろう．「日本語教育基本語彙七種比較対照表」にある語彙の共通度（p. 10）を見ても，2種のみ共通から6種のみ共通する語彙までは14.3%，8.8%，7.8%，7.0%，9.1%と推移するのが，7種全部に共通となると4.6%で極端に落ち込んでいる．また，重要語彙集ではなく教材のテキストの使用語彙について共通語彙を出す場合も同種の問題が出るようだ．菊池（1983）では5種類のドイツ語の教科書すべてに共通する語彙は229語に過ぎないが，4種類以上に出てくる語彙ではかなり増えて407語という結果を出している．しかし，「調査対象すべてに共通する語彙」をもとめることに全員一致の欠点があるといっても，この欠点は調査対象によってはまったく見られないようである．大野（1987）によると源氏物語や土佐日記や枕草紙や大鏡など古典16作品の共通語彙についての調査結果があるが，この場合はすべての作品に共通する語彙が極端に少なくなるという結果にはなっていない．

■ 日本の古典16作品における共通語彙数の推移（大野，p. 258）

7作品：359語	10作品：191語	13作品：137語	16作品：145語
8作品：297語	11作品：151語	14作品：111語	
9作品：236語	12作品：157語	15作品： 97語	

人為的に語彙を決める重要語彙集や人為的に語彙を制限する外国語の教材に特有の現象である可能性もあるだろう．無作為の文学作品などにはこの種の現象は起こらないのかもしれない．また，吉岡（1996，pp. 208-210）に英語についての6つの語彙調査の表が出ている．LOB Corpusという大規模な語彙調査における上位50位の単語について他の5つの調査の結果と比較している．それに

よれば，同様に大規模な Brown Corpus と LOB Corpus の上位 50 語の種類や順位はかなり一致しているが，日本の中学生用英語教科書や高校生用英語教科書を対象にした調査もふくめて，調査対象が違えば上位の語彙や順位はかなり異なるという結果になっている．吉岡は「英語教育において基本語彙の設定が重要な課題であるが，その場合どのような資料に基ずくかによって相当な違いが生じてくるであろう」(p. 210) とまとめている．

5）正規表現は GREP や SED や AWK などの汎用テキスト処理ツールを使いこなすためにも必要な知識である．実際的な知識を得るには市販のコンピュータ関連図書がわかりやすい．植村・富永 (1993)，SE 編集部 (1992)，羽山 (1991)，Aho/Kernighan/Weinberger (1989)，Dougherty (1991) などが参考になる．動作原理や成立事情については石川 (1995) が詳しい．

6）ここで利用している辞書データは電子ブック版の Duden Universalwörterbuch の内容をテキストファイルで取り出したものである．草本和馬さんの DDwin というフリーソフトでこれが可能になるのだが，これについては，見出し語だけを取り出す方法や辞書の中身をすべて取り出す方法について城岡 (1996) で詳しく書いている．

7）41 の学習つづりの中には「母音字＋h」や「同一母音字の繰りかえし」があって，つづりの読み方の規則としてはまとめることが可能だが，ここではそういう考慮はしないで，それぞれのつづりの頻度からそれほど重要でないつづりを出しておこう．たとえば 0.9%以下の語彙にしかあらわれないそれほど頻度の高くないつづりには jisho.dat の結果では 25 位以下，SPIEGEL では 20 位以下が該当する．[1.3]で利用した COMM でどちらの調査でも 0.9%以下の頻度のつづりを出すと，aa, ai, ay, ds, dt, ee, ey, ph, ps, qu, uh, üh, äu, äh, öh の 15 だった．これらのつづりについては，最初から一般的な規則として与えるよりは学習する内容に現れたときに個別に対応するか，ある程度学習が進んだ段階であらためて教えるようにしたほうが効率的だろう．

反対に，高頻度のつづりについても教育効果を考えておくと，er と ch の扱いはもっと重視すべきだと結論できるだろう．従来の er の教材での扱われ方は r の語末での母音化の例としてぐらいであるが，Verkehr, Erde, erben, Arbeiter では 4 つのことなる発音があるわけだし，頻度 (jisho.dat で 31.190%，SPIEGEL で 20.171%) を考えれば相応の扱いがふさわしいはずである．また，やはりきわめて高頻度の ch (jisho.dat では 26.260%，SPIEGEL では 14.475%) であるが，伝統的な Ich-Laut と Ach-Laut の区別だけでなく，chs

や sch や tsch も加えた総合的な扱いを受けるだけの重要性はあると思われる．

8）H. Meyer (1964) にもドイツ語の部分文字列の頻度順の表がある (p. 337)．残念ながら，文字の順序は無視してたとえば en と ne をあわせてかぞえるというやり方をしているのでそのまま比較することはできない．なお，Meyer のデータは 10 万字のデータで，詳細は不明だが，辞書の見出し語などではなく，通常のテキストを対象にして文字列の頻度を調査したもの思われる．

■ **頻度の高い 2 文字の部分文字列（H. Meyer から）**

1 .	en	ne	6.1%	6 .	te	et	2.7%
2 .	er	re	5.7%	7 .	in	ni	2.5%
3 .	ei	ie	4.4%	8 .	ge	eg	2.2%
4 .	ch		3.3%	9 .	es	es	2.2%
5 .	de	ed	2.8%	10 .	un	nu	2.2%

9）jisho.dat は一行に一語ずつの形式になっていて，動詞の sein と所有代名詞の sein というような同音異義語，厳密には同字語を一切認めていない．一字違い語の調査では同字語がリストにはないということが重要なポイントである．というのは，語彙から一字削除して，同字語が生じるかどうかしらべることで一字違い語の調査をしているからだ．

10）一字違い語の調査については，b と w のように語頭以外にあまり利用されないのがふつうなのか，それとも l と r のように語頭以外でもかなり利用される対がふつうなのか．あるいは，ドイツ語の b と w，l と r だけを見ていると語頭での一字違い語が圧倒的に多いが，他の言語でもそうなのか，こういった点はさらに調査をすすめてみたい．

参考文献

石井正彦（1992）：プログラムを書かずにできる言語処理「日本語学」1992 年 9 月号，明治書院，pp. 114-124

石川克知（1994）：コンピュータによるテキスト処理「言語文化部紀要」27 号，北海道大学，pp. 77-100

石川克知（1995）：テキスト処理と正規表現「Norden」32 号，pp. 1-30

伊藤雅光（1996）：マルチメディア KWIC 電子文献学「月刊言語」1996 年 9 月号，大修館

植田康成（1982）：計算機によるドイツ語教科書の語彙調査について「広島大学文学部紀要」42巻特輯号3号，pp. 1-162
植村富士夫・富永浩之（1993）：「awkでプログラミング」オーム社
大野晋（1987）：「文法と語彙」岩波書店
片山裕（1993）：「sedパズルブック」インプレス
加藤正隆（1977）：暗号解読「月刊言語」1977年8月号，大修館，pp. 20-30
菊池雅子（1983）：初級ドイツ語教科書の語彙「日吉論文集」32号，慶應義塾大学商学部，pp. 86-113
金水敏（1992）：AWKによるテキスト型データベース活用術「日本語学」1992年6月号，明治書院，pp. 127-135
河野収（1986）：インデックス・コンコーダンスの機械処理「同志社外国文学研究」43・44合併号，同志社大学外国文学会，pp. 1-17
河野収（1995）：カフカの『変身』の語彙分析「同志社外国文学研究」70号，pp. 1-28
国立国語研究所（1982）：「日本語教育基本語彙七種比較対照表」
小林栄三郎（1988）：R. M. リルケの『ドイノの悲歌』と『オルフォイスのソネット』における用語法のいくつかの特色について「日吉紀要」ドイツ語学・文学7号，慶応義塾大学，pp. 1-90
近藤弘・川崎正（1991）：「共通基本単語」の理解と必要性「ドイツ語教育部会会報」39号，pp. 38-44
重藤実（1990）：ドイツ語教育と基礎語彙「ドイツ語教育部会会報」37号，pp. 18-22
志村拓・大池浩一（1990）：「MS-DOS SOFTWARE TOOLS基本セット32」アスキー
城岡啓二（1994）：ドイツ語の動詞の人称変化形をかぞえてみる「研究報告人文・社会科学篇」第30巻第1号，静岡大学教養部，pp. 99-121
城岡啓二（1996）：テキストファイル版ドイツ語逆引き辞典の作成とその利用「人文論集」47号の1，静岡大学人文学部，pp. 277-310
杉本武（1992）：正規表現によるプレーン・テキストの検索「日本語学」1992年7月号，pp. 112-121
SE編集部（1992）：「MS-DOSテキストデータ料理学」翔泳社
中野洋（1994）：「パソコンによる日本語研究法入門―計量語彙論へのいざない」私家版

中野洋(1996)：語彙調査用プログラム集 MCL「計量国語学」20 巻 6 号, pp. 265-279
羽山博 (1991)：「実用 UNIX」アスキー
藤澤正明 (1984)：パーソナルコンピュータによるドイツ語処理「研究報告 (人文・社会科学)」32 号，九州工業大学，pp. 83-98
藤澤正明(1985)：パーソナルコンピュータによるハイネ研究「研究報告(人文・社会科学)」33 号，九州工業大学，pp. 125-145
舟本奨 (1994)：「実用 UNIX ハンドブック [改訂新版]」ナツメ社
古田啓 (1991)：文法研究のためのデータベース利用「日本語学」1991 年 12 月号
室井禎之(1992)：コンピュータ利用によるコロケーション分析の実際とその言語研究上の意義について「言語文化部紀要」22 号，北海道大学，pp. 153-167
森泉 (1991)：パーソナルコンピュータによる初級独作文教材の語彙調査「慶應義塾大学日吉紀要ドイツ語学・文学」12 号，pp. 105-121
吉岡健一 (1996)：付録・計量文体学研究の展望「文章の計量」(A. Kenny 著) 南雲堂，pp. 196-237
米井巌(1986)：パーソナルコンピュータを用いたドイツ語初級教材の語彙調査「研究紀要」32 号，日本大学人文科学研究所，pp. 160-193
米井巌 (1989)：逆引き形式の語彙総覧とその語学教育上の有効性「ドイツ文学論集」10 号，日本大学，pp. 41- 59
明治書院 (1995)：パソコンを使う日本語研究「日本語学」1995 年 7 月臨時増刊号

Aho, A. V./Kernighan, B. W./Weinberger, P. J. (1989)：「プログラミング言語 AWK」トッパン
Dougherty, D. (1991)：「sed & awk プログラミング」アスキー
Hein, D. (1995)：UNIX gestützte maschinelle morphologische Untersuchung zur Komposition. Bonn (Holos Verlag).
Kenny, A. (1996)：「文章の計量」南雲堂
Taylor, J. R. (1995)：Linguistic Categorization. 2nd Edition. Oxford.
Meier, H. (1964)：Deutsche Sprachstatistik. Hildesheim (Olms).
Mocker, U./Mocker H./Werner M. (1993)：PC-Einsatz in den Geisteswissenschaften. (Beck EDV Berater im dtv).

Muster, J./Birns, P. (1992)：「UNIX コマンド活用ハンドブック」パーソナルメディア

Stallman, R. M./Rubin, P. H./Robbins, A. D./Close, D. B. (1993)：「GAWK」アジソン・ウェスレイ

Staubach, G. (1989) : UNIX-Werkzeuge zur Textmusterverarbeitung. Berlin Heidelberg (Springer).

Lenders W. (Hrsg.) (1993) : Computereinsatz in der Angewandten Linguistik. Frankfurt a. M. (Peter Lang)