英語辞書の構造

——電子テキスト化した辞書を利用して——

高橋　涉

1．目　的

本稿の目的は，高校生・大学生を主たる対象にした中辞典クラスの学習英和辞典の一つを選び，そこに含まれるほぼすべての見出し語を電子テキスト化することによって，英語の語彙部門がどのような内部構造を持つのかを計量分析することである。この計量分析は，学内LANによって結ばれたUNIXシステムで提供される各種の汎用コマンドを援用することによって行われた。

2．分　析

2．1　電子テキストの作成

高性能パーソナル・コンピュータの普及と，さらにはかつては極めて高価だったCD-ROMドライブの急速な低価格化によって，電子ブックやCD-ROMなど様々なメディア形態をとった辞書が以前に比べて容易に利用できるようになってきている。これらの機器の普及によって，従来は逆引き辞典のような「本」の形態でしか利用できなかった逆引き機能（たとえば語末が-ityで終わる単語のリストアップなど）が，CD-ROM版辞書や電子ブックに備えられた逆引き機能を実現するソフトウェアによって手軽に実行できるようになった。

しかしながら，これらの既成の「電子辞書」には様々な使用上の制約が存在することも事実である。たとえば，たとえ辞書から-ityで終わる単語のすべてをリストアップできたにせよ，それらのリストアップされた語群を別ファイルにして再利用することが不可能であることが多いこと，あるいは実行結果をリストアップする際の最大語数に制限があり，一定数（たとえば400件以上）が該当する場合は正確なヒット数が不明になることが多いこと，該当語が何語存在するのか即座にはカウントできない場合があることなどである。

さらに，上記のようなデメリットに加え，既成の電子辞書の持つ最大の弱点は，正規表現の利用がほとんどの場合不可能であることであろう。正規表現とは，特定の文字列パターンを一般的な表現で示す書式のことで，文字列の検索や置換時には極めて強力な手段となるものである。ことに以下に論じるような，何万行にも及ぶ膨大なデータ群から目的の文字列を高速に検索する際には，正規表現を利用した検索機能は必須のものであると言えよう。

下記(1)に述べられた条件を満たす文字列の検索は正規表現を用いずにどのようにして実行可能であろうか？

(1) a　5文字からなる単語をリストアップする。
　b　行の先頭が数字で，そのあとに英語大文字が2つ続く行をリストアップする。
　c　un で始まり，ity で終わる単語のすべてをリストアップする。
　a'　^..... $
　b'　^[0-9][A-Z][A-Z]
　c'　^un.* ity$

上記（1a-c）のような内部構造を持つ複雑な文字列はそれぞれ（1a'-c'）のような正規表現で表されるが，これを正規表現以外の手段で高速に検索することは不可能とは言わないまでも著しい困難を伴うことは予想に難くない[1]。

以上のような理由から，CD-ROM など様々なメディアで多くの辞典類が出版されている現状ではあるが，あえて自力での英和辞典の電子テキスト化を試みたわけである。選択した英和辞典は，たまたま筆者の手もとにあった辞典のなかから，ジーニアス英和辞典（大修館書店　1988）を選んだ[2]。

同辞典の見出し語をすべてエディタ（= MIFES）によって入力したが，入力の際以下に述べるようないくつかの項目はあえて入力しなかった。

(2) a．名詞句が見出し語となっている場合（blue ribbon, good afternoon, Sancho Panza, etc.）。ただし blue-ribbon のようにハイフンで結ばれている場合は1語とみなして入力した。
　b．略語が見出し語となっている場合（dB, lb., Rb., etc.）。ただし Mt., Mr., Mrs., rasar のような略語や頭字語は高頻度に使用される単語として入力したものもある。
　c．接頭辞，接尾辞のような単語より小さい要素が独立して見出し語となっている場合。
　d．外国語から借入されている単語で，もとの言語に存在する綴りの上での補助記号（senõr の ~，fiancé の ´ などの記号類）も省略した。

また本電子テキスト辞書ファイルは，英単語の見出し語のみを入力したことからも明らかなように，将来的には形態論，音韻論の研究用基礎データとしての利用を念頭においており，意味論や語義の研究に利用することは当初から考慮していない。したがって，英語語彙に多数存在する同音異義語で，辞書の中では別見出し語としてリストされている語群，たとえば bank（土手，銀行），bode（～の前兆となる，bide の過去形）はすべて1語のみを入力した。また英国式・米国式の区別やその他の理由による綴りの上での変異体はそれぞれ別語として入力した。

前述の基準に従って，最終的には55,188語，583,781バイトにも及ぶ膨大なデータを独力で手作業のみによって入力したのであるが，その際には数多くのミスタイプや同一単語の重複入力があった。これらのミスの多くは，Word Perfect のスペリングチェック機能によるミスタイプの訂正や，UNIX の uniq コマンドによる重複行チェックなどにより可能な限り

の修正を行った[3]。このようにして完成した電子テキスト化された辞書ファイルを利用して，UNIX で提供されるいくつかの文字列処理コマンドを用いつつ，以下に英語のレキシコンの内部構造を計量分析していくことにしよう。

2. 2 英語辞書の構造

まずこのファイルに含まれる単語群の語頭・語末のアルファベット順の生起回数を計量してみよう[4]。

(3)

	G				W			
	語 頭	%	語 末	%	語 頭	%	語 末	%
a	3163	5.73	1273	3.31	1679	6.83	969	3.96
b	2928	5.31	199	0.36	1639	6.70	128	0.52
c	5012	9.09	1349	2.45	2425	9.91	739	3.02
d	3073	5.57	3905	7.08	1323	5.41	1242	5.07
e	2116	3.84	9692	17.57	1059	4.33	5409	22.10
f	2243	4.07	272	0.49	993	4.06	147	0.60
g	1721	3.12	1970	3.57	888	3.63	775	3.17
h	2170	3.93	1266	2.29	1036	4.23	769	3.14
i	2425	4.40	253	0.46	1020	4.17	188	0.77
j	511	0.93	5	0.01	304	1.24	4	0.02
k	472	0.86	1019	1.85	308	1.26	548	2.24
l	1791	3.25	2825	5.12	855	3.49	1481	6.05
m	2879	5.22	1491	2.70	1409	5.76	704	2.88
n	1346	2.44	5020	9.10	510	2.08	2716	11.10
o	1359	2.46	672	1.22	477	1.95	433	1.77
p	4368	7.92	790	1.43	1907	7.80	327	1.34
q	267	0.48	4	0.01	125	0.51	3	0.01
r	2827	5.12	4772	8.65	1022	4.18	1474	6.02
s	6520	11.82	4490	8.14	2716	11.10	1346	5.50
t	3005	5.45	4598	8.33	1298	5.31	2328	9.51
u	2129	3.86	113	0.20	197	0.86	72	0.29
v	949	1.72	22	0.04	403	1.65	18	0.07
w	1525	2.76	314	0.57	697	2.85	188	0.77
x	37	0.07	210	0.38	12	0.05	127	0.52
y	196	0.36	8589	15.57	106	0.43	2279	9.31

z	133	0.24	57	0.10	56	0.23	59	0.24
全	55165		55170		24464		24473	

表(3)におけるGは分析対象のジーニアス英和辞典を示す略号であり（以下すべてGと略記），表の右半分にWとして表示したものは，UNIXに標準で備わっているスペルチェック機能を実現する際に参照されるテキストファイルwords（以下Wと略記）に対して同様の分析を行なった結果を示した。普通/usr/dictのディレクトリーに存在するこのwordsというファイルは，（筆者の利用するサイトでは）総語数24,474語からなる英単語のリストである。

UNIXのコマンドラインからspell *filename*と入力すると，*filename*で指定されたファイル中のすべての単語のスペリングをチェックし，words中に存在しない単語（すなわちスペリングを間違えている単語）を標準出力へ出力する。ところがこのUNIXのスペルチェック機能はいくつかの単語をそれらが本来の英単語ではないにもかかわらず，正しい単語と判断するようである。

(4) goed (go+ed), goment (go+ment), disgo (dis+go), unsad (un+sad), unbad (un+bad), etc.

これらの単語は英語には存在しないはずのものであり，かつwordsファイル中にも存在しないのであるが，UNIXのスペルチェック機能はこれらを可とする。すなわちspellコマンドはソフトウェアでいくつかの接辞と語基（base）とを結びつけ，その結合が本来の英語には存在しなくても接辞と語基それぞれが正しいスペリングならば，全体も正しいスペリングと見なしているようである。したがってこれも一種の電子テキストであるwordsという辞書ファイルは，いくつかの接辞が付加された結果として存在する単語はファイル中に存在しないという点で，一般的な辞書とは性質が異なっていることになる。しかし今回われわれの検討する辞書Gとの比較のために，Wも語頭・語末の文字別生起数という面でのみ，分析対象とした。

表(3)はGと上記の条件を認めた上でのWに対しての，語頭・語末の文字別頻度一覧表である。26文字からなるアルファベットがすべて同一の確率で語頭ないしは語末に生起するとすれば，その確率は約3.846%（=1/26）となるはずであるが，当然ながら実際の文字別生起率は語頭・語末のそれぞれの位置において著しい偏りを示す。以下にG，Wそれぞれのアルファベット26文字の生起率順リストを提示する。

(5)

G語頭

s, c, p, a, d, t, b, m, r, i, f, h, u, e, l, g, w, o, n, v, j, k, q, y, z, x

W語頭

s, c, p, a, b, m, d, t, e, h, r, i, f, g, l, w, n, o, v, k, j, u, q, y, z, x

G語末

e, y, n, r, t, s, d, l, g, m, c, a, h, k, p, o, w, f, i, x, b, u, z, v, j, q

W語末

e, n, t, y, l, r, s, d, a, g, h, c, m, k, o, p, w, i, f, b, x, u, z, v, j, q

表(3), (5)からわかることは，Wのやや特殊な構造を考慮してもGとWの語頭，語末それぞれの文字別生起率は極めて類似していることである。語頭においてはまったく同一の順位である文字は11文字（42.3%）であり，順位が±2の差しかない文字は13文字（50%）で，全体の92.3%の文字が±2の順位差のうちにおさまっている。

これに対し語末では，同一順位の文字は8文字（30.7%），±2以内に順位差がおさまる文字は15文字（57.7%）で，全体の88.4%が±2の順位差の中に入っている。±3までの順位差を考慮すると，語頭におけるeとuの相違以外のすべての生起率がその誤差の中におさまり，英語の辞書（すなわち我々の言語能力の一部としてのレキシコン部門）は語頭・語末の文字別生起率という点で極めて大きな規則性がひそんでいることがまったく別個の2つの辞書に対する計量分析からわかった[5]。

次にそれぞれの辞書の語頭・語末における文字の生起率について考えてみよう。我々が日常使用する辞書は単語の先頭からのアルファベット順に語彙は配列されているから，s，c，pなどの項目のページ数の多さ，あるいは逆にy，z，xなどの項目のページ数の少なさから，アルファベット26文字のそれぞれについての語彙数が不均一であることは経験的に知っている。しかしながら語末の文字別生起率に関しては，逆引き辞典をたんねんに数えるなら別だが，一般に語末における文字別生起数（生起率）を経験的に知る機会は少ないように思われる。表(3)や(5)のリストから，語末の文字別生起率は語頭のそれとはまったく異なった様相を示していることがわかる。そしてその理由は多く，英語の正書法上のシステムや形態論の観点から説明できるものであるように思われる。

たとえばG，Wとも語末における生起率第1位のeは英語の語彙に多数見られる語末の黙字eの存在が理由であろうし，Gで第2位を占めるyは，英語形態論において極めて生産的な派生接尾辞のひとつである -ly の存在がその理由であろう。Wにおいてはyの順位がGより下位なのは，前述のようにWは派生形をほとんどリストせずソフトウェアで処理しているからであろう。また周知のとおり英語の接辞は接頭辞より接尾辞のほうがはるかにその数が多く，また生産性も接尾辞のほうが高い。したがってそれだけ語頭より語末において多様な（語頭とは異なった）パターンが見られることになる。

次にGのファイル中に含まれるアルファベット別の全文字数を計量してみよう。この計量においては大文字と小文字は同一視しており，アルファベット以外の文字種（たとえば数字，ハイフン，アポストロフィなどの文字）は計量の対象としていない。

(6)

	生起回数	%		生起回数	%		生起回数	%
a	38468	8.21	j	864	0.18	s	29162	6.23
b	9354	2.00	k	4278	0.91	t	32612	6.96
c	20130	4.30	l	29077	6.21	u	17258	3.68
d	15460	3.30	m	1336	2.85	v	5188	1.11
e	52463	11.20	n	33384	7.13	w	4475	0.96
f	6872	1.47	o	31179	6.66	x	1342	0.29
g	10745	2.29	p	13743	2.93	y	11412	2.44
h	12130	2.59	q	835	0.18	z	1531	0.33
i	39252	8.03	r	33802	7.22			

(7)

G	I	II	III	IV	V	VI
e	e	e	e	e	e	e (12000)
i	t	t	t	t	t	t (9000)
a	a	i	a	a	a	a (8000)
r	o	a	i	o	o	i (8000)
n	n	o	o	h	i	n (8000)
t	i	n	n	n	n	o (8000)
o	s	s	s	i	s	s (8000)
s	r	r	r	s	r	h (6400)
l	h	h	h	r	h	r (6200)
c	l	l	l	d	l	d (4400)
u	d	d	c	l	d	l (4000)
d	c	c	d	u	c	u (3400)
p	m	u	u	w	u	c (3000)
m	u	m	m	m	m	m (3000)
h	f	f	f	c	f	f (2500)
y	p	p	p	g	p	w (2000)
g	g	y	g	f	g	y (2000)
b	w	w	y	y	w	g (1700)
f	y	g	b	p	y	p (1700)
v	b	b	w	b	b	b (1600)
w	v	v	v	k	v	v (1200)
k	k	k	k	v	k	k (800)
z	j	x	x	j	x	q (500)
x	x	j	q	x	j	j (400)
j	q	q	j	z	q	x (400)
q	z	z	z	q	z	z (200)

G：ジーニアス英和辞典
I：press reporting
II：religious writing

Ⅲ：scientific writing
Ⅳ：general fiction
Ⅴ：総計100万語以上におよぶ15のカテゴリーの文の中の平均的順序
Ⅵ：S. Morse が Morse 符号を作成したときに参考にした順序
（　）内の数字はその基礎となったある印刷所の所有する活字数

表(6)はGに生起する文字のアルファベット順生起回数と生起率である。また表(7)は(6)を頻度順に並べたものである。表(7)中のⅠ～ⅥはCrystal（1987：86）に提示された別記の基準で調査された文字別頻度順一覧であるが，これもGとの比較のために併記してみた。

英語の中で用いられる文字は，その生起率が大きく異なることは周知の事実であり，この相違はコンピュータやタイプライターのキーボード配列（我々になじみのQWERTY配列とは異なるDVORAK配列は使用頻度の高い文字を力の入れやすい指で打てるように文字別生起率の相違を考慮してキーボード上の文字を配列したものとして有名である）という身近な分野から，S. Morseのモールス符号の策定時の事情や，暗号解読といったやや我々にはなじみの無い分野にいたるまでいたるところに反映されている。本稿でコンピュータによって計量された文字別生起率は辞書という，表(7)のⅠ～Ⅵのような実際のテキストの分析とは明らかに異なったデータを基にはしているが，これは我々の内在化した語彙部門の（一部の）データに対する計量という意味で，興味深い事実を示しているものと言えよう。

まずGに生起する全468,352文字の実に1割強（11.2%）を占める文字がｅである。実際のテキストを計量した結果を示すⅠ～Ⅵでもやはり1位を占めている。しかしながらこれらのカテゴリーは当然のことながら，我々の分析したGとは異なり，Type/Tokenの区別でいうところのTokenとしての語彙を計量している。したがって英語の実際のテキストに極めて多用される定冠詞theの中に含まれるｅが高頻度に出現するその回数分だけ計量されていることになる。Type/Tokenの区別におけるTypeを計量したGにおいては，当然ながら実際には極めて高頻度に出現するはずの定冠詞theは全データ中たった一回しか出現しない。それにもかかわらずGにおいても文字ｅが第1位にランクされるということは，英語の語彙の中にはいかに多くのｅという文字が使用されているかを示す証拠といえよう。

さらにⅠ～Ⅵで第2位にランクされている文字ｔは，Gにおいては第6位である。これも前述のtheのType/Tokenの相違から説明できようし，同じくｈの順位の相違（Gで15位，Ⅰ～Ⅵで6つのクラス中4クラスで9位）もここに起因するものと思われる。実際のテキスト中（特に書き言葉において）多用される前置詞ofのType/Tokenの区別は，Gにおいての方がⅠ～Ⅵより文字ｏの生起率が低下していることに反映されているだろう。

一方低頻度使用文字については次のようなことが言えよう。22位から26位までの5文字を仮に低頻度使用文字群とすると，Gにおける低頻度使用文字ｋ，ｚ，ｘ，ｊ，ｑはクラスⅣにおいてｋのかわりにｖが存在することを除けば，順位の細かな変動はあるがすべて共通している。すなわちGというTypeを基準としているデータ中で低頻度に生起する文字群は，Ⅰ～ⅥというTokenを基準とするデータ中でもやはり使用頻度は極めて低いことがわかっ

た。

最後に母音と子音という二分法からデータを検討してみよう。Crystal は(8a)のように述べている。

(8)

a. We can say with confidence that if we write a *q* in English, it is almost always going to be followed by *u* (though not always, because of *Iraq*, and other exceptions). Less obviously, but equally confidently, it emerges that just over 60% of everything we say will be made up of consonants, and just under 40% of vowels. Crystal (1987 : 86)

b.	Vowels (a, e, i, o, u)	178,620文字	38.14%
	Consonants (残り21文字)	289,732文字	61.86%
	G 全体の文字数	468,352文字	100.00%

Crystal が(8a)後半で述べていることは，おそらく実際のテキストの分析に基づいて述べているのであろうが，前述のように Type としての語彙をリストしたGを計量しても彼の主張は妥当であることが(8b)から明確にわかる。彼が主張するように，また他でもしばしば指摘されるように，我々の使用する言語はふつうは気づかれることの少ない，数々の統計的規則性に満ちていることを示す充分な証拠といえるであろう。

また(8a)前半で彼が指摘している事実も我々が日常それとなく実感していることであるが，このことも我々のファイルGをコンピュータを用いて計量すれば実際の確率まで正確に計量できるのである。

表(6)からファイルGに存在する文字qは全部で835あることがわかる。そのうち語末に生じるqは表(3)から4つ存在することがわかる。さらに文字qのあとにu以外の文字が続くケースは，Iraqi，Qantas，Qatar，Qiana など全部で8語存在することが正規表現を用いた検索から容易にわかる。すなわち全835回という文字qの生起回数から，語末のqの生起回数と前述のu以外の文字が続く8回を減じた結果である823回という回数がqの後に彼の指摘したuが生起している回数なのである。つまり英語の単語において文字qの後にuが後続する確率は98.56%（823/835）と計算できるのである。これもデータが明確に示す言語の統計的規則性の一つと言えるはずであり，これらの計量はコンピュータの手助け無くてはとても実行が難しいものであると考えられる。

3．結　　語

本稿では，一冊の英和辞典の見出し語を電子テキスト化する作業の完成に伴って，ふだんあまり指摘される機会のないいくつかの計量分析を，パーソナルコンピュータやそれを端末として利用する UNIX システムを援用することによって行った結果を報告したものである。単語の語頭・語末における文字別生起数の相違や，英語辞書の中に存在する文字の計量などはこのように電子テキスト化したファイルを利用して初めて容易に実現できるものである。

本稿では，この分析法の実際の過程にはほとんど言及していないが，高橋・野澤（出版予定）などには，比較的その分析法そのものを紹介した部分もあるので参考となるはずである。英語学の各分野に限らず，ほかの様々な分野でこのような分析法を活用できる可能性は極めて高いと思われる。

註

1　正規表現の簡単な解説は高橋（1992）を参照のこと。また参考文献ほかのUNIXの解説書に正規表現の詳しい解説がある。

2　この入力はすべて手作業かつ独力で行ったので，2年以上の期間がかかってしまった。1994年には同辞典は改訂版が出版されているが，資料の均一性を考慮してあえて旧版を最後まで入力した。

3　UNIXにはスペリングチェック機能が備わっているが，後述の理由からその機能はやや完全とは言えない。

4　単語の合計がG，Wそれぞれ語頭・語末において本来の総語数より少ないのは，英字以外（アポストロフィ，ピリオド，数字など）が語頭・語末に生じているケースをカウントしていないためである。

5　語頭のu（Gで13位，Wで23位）の2つの辞書間の生起率の差は容易に説明が可能である。すなわちWは前述の理由から接頭辞un-を含む単語のほとんどをリストしていないためである。語頭のe（Gで14位，Wで9位）についてはその生起率の差を説明する理由が筆者には不明であった。

参考文献

Bauer, Laurie. 1983. English word-formation. Cambridge University Press.

Crystal, David. 1987. The Cambridge encyclopedia of language. Cambridge University Press.

竝木　崇康．1985．語形成．新英文法選書第2巻．大修館書店．

高橋　　渉．1992．「パーソナル・コンピュータを利用した英語語彙分析法」　信州大学教育工学センター紀要第8号．pp. 61-72.

高橋　　渉，野澤重典．1995．「UNIXを利用した英語語彙分析」　信州大学教育学部附属教育実践指導研究センター紀要第3号掲載予定．

山口　和紀　他．1992．The UNIX super text（上，下）．技術評論社．

（1995年4月26日　受理）