神戸女子短期大学　論攷　49巻　69－79（2004）

―ノート―

コーパスに基づく言語研究
－コーパス分析ソフトを利用した語彙リスト作成の試み－

岩中貴裕

A Language Study Based on a Corpus:
An Attempt to Make a Vocabulary List with a Corpus Analyzer

Takahiro IWANAKA

要　旨

英語学習者の多くは語彙の知識を内容理解の第一の拠り所にすると考えられている。そのため語彙不足は学習者にとって読解の際に大きな障害になる。読解を効率よく行わせるためには学習者が困難を感じると思われる語彙，表現についての知識を前もって与えておくことが望ましい。この作業は電子化されたテキストとコーパス分析ソフトを利用することによって，誰にでも簡単に行うことができる。本稿ではその具体的な方法を紹介する。

キーワード：コーパス　corpus，　語彙リスト　vocabulary list，
コロケーション　collocation，　リーディング　reading，
ワードラボ　WordLab

1．はじめに

世界最初のコーパスである Brown Corpus[1)] がアメリカで誕生したのは1960年代初期のことである。それから約40年の歳月が経過しているわけであるが，特に最近，コンピュータの性能の向上と使用の手軽さが引き金となり多くの人々が研究，あるいは教育のためにコーパスを活用している。コーパスの規模も以前とは比べ物にならない規模のものが入手可能になっている。British National Corpus[2)] や Bank of English[3)] のようなコーパスを利用すれば，延べ語数で億単位の電子テキストを入手することができる。Project Gutenberg[4)] を利用すれば膨大な量の文学作品電子テキストを入手することが可能である。今後も技術の進歩に伴ってコーパスはますます充実していくものと思われる。

コーパスを利用した研究のこれまでの歩みと今後の可能性について八木（2003：248）は次のように述べている。

> 一般に，新しい分野を開拓する時期には，研究者はそれぞれ自らの研究推進に没頭す

るものである。これを初期段階と考えることができるであろう。ある程度研究の成果が出てきて成熟期になると，その研究によって得られた知見を何らかの形で普及したい，あるいは，普及することによって社会に還元したいと考えることは自然の成り行きである。

コーパスを利用した言語研究は初期段階から，次の段階へ移行しつつあると言っていいだろう。今後は，コーパスがそれぞれの研究領域でどのように有効利用できるのかについて具体的な議論がなされなくてはならない。本稿ではコーパスが英語教育にどのように貢献できるのかについて考察し，その一例を示すことを目的とする。コーパスの英語教育への貢献については様々な方法が考えられるが，本稿ではリーディング教材に焦点をあてて議論を進めていく。電子テキスト化されたリーディング教材と言語処理ソフトを利用することによってどのような教材提示が可能になるのか考えていく。

2．使用するテキストと言語分析ツールについて

2.1. 使用テキスト

本稿では大学レベルで使用されるテキストの一例として *The Universe of English* に収録されている ‘How Our Ancestors Survived the Ice Age’ を利用する。最近は難易度や使用する語彙のレベルを下げた教材が大学レベルでも使用されているが，この傾向を安易に受け入れることは慎まなければならない。むしろ，教材のレベルをなるべく落とさずにわかりやすい授業を展開する方法を模索しなくてはならない。*The Universe of English* に収録されているテキストは確かに難解なものが多い。高校を出たばかりの学生にとってはこのレベルのテキストを理解することはかなりの困難を伴うであろう。簡単な教材を使用して，楽しくわかりやすい授業を展開したほうが望ましいという意見ももっともである。しかし難しいテキストも，事前にキーワードを提示する，学生が苦労するであろうと思われる点について説明を加えておく等の配慮をすることによって問題なく授業で使用することができると筆者は考えている。そしてテキストを電子化しておけば，その作業は誰にでも簡単に行えるものなのである。以下その手順を説明していく。

2.2. 言語分析ツール

本稿では，言語分析ツールとして WordLab（English Corpus All-round Analyzer）[5] を使用する。一般には TXTANA[6] や WordSmith[7] の方がよく知られているかもしれないが，WordLab はこれらと比較しても遜色ない機能を持ち合わせている。WordLab が備えている語彙統計作成機能，コロケーション統計作成機能はテキストを読む前の活動として学習者にどのような言語材料を提示すればいいのか，客観的な根拠に基づいて決定することを可能にして

くれる。また高度な検索条件を簡単に設定できるので，これまで言語分析ツールを使用したことがない人でも簡単に使用することが可能である。

3．テキストと語彙

3.1. テキストにおける語彙の役割

テキストの難しさを決定する要因には様々なものがあるが本稿では語彙に焦点をあてる。一般的にはあるテキストを理解するためにはテキスト内の語を95%知っておく必要があるとされている。もちろん語の意味は文脈の中で変わることもあり自分が知っていると思っていた語がまったく別の意味で使用されている可能性もある。しかし，一般的にはテキスト内で使用される語彙の情報について事前に提示しておけば読解の際にそれが役に立つことは明白である。次節では WordLab を利用して語彙リストの作成を試みる。

3.2. WordLab を使用した語彙リスト作成

WordLab には語彙統計を作成する機能が備わっている。品詞分類についてはそのまま信頼することができない[8)]がテキスト中でどのような語彙がどれだけの頻度で使用されているかについて短時間で知ることができる。スペースの都合上，すべてのリストをここで紹介することはできないので名詞のリストのみ紹介する。テキスト内で使用されている名詞は下記の通りである[9)]。

表 1

word	単数	複数	所有形	計	%
111語	124	82	0	206	
ability	1	0	0	1	0.48
absorber	1	0	0	1	0.48
acre	0	1	0	1	0.48
advantage	2	0	0	2	0.97
ago	5	0	0	5	2.42
alternation	1	0	0	1	0.48
ancestor	1	3	0	4	1.94
animal	1	2	0	3	1.45
area	0	1	0	1	0.48
arrival	1	0	0	1	0.48
attachment	1	0	0	1	0.48
Australopithecus	5	0	0	5	2.42
baboon	0	3	0	3	1.45
band	1	0	0	1	0.48
being	4	0	0	4	1.94
box	0	1	0	1	0.48
buttock	1	0	0	1	0.48
century	1	0	0	1	0.48

chimp	1	2	0	3	1.45
climate	3	0	0	3	1.45
combination	1	0	0	1	0.48
comeback	0	1	0	1	0.48
competition	2	0	0	2	0.97
cousin	1	0	0	1	0.48
cup	1	0	0	1	0.48
day	1	0	0	1	0.48
degree	0	1	0	1	0.48
descent	1	0	0	1	0.48
diet	3	1	0	4	1.94
digger	0	1	0	1	0.48
eater	1	0	0	1	0.48
equator	1	0	0	1	0.48
evolution	1	0	0	1	0.48
explosion	1	0	0	1	0.48
extinction	1	0	0	1	0.48
eye	0	1	0	1	0.48
fact	2	0	0	2	0.97
family	1	3	0	4	1.94
fang	0	1	0	1	0.48
finger	0	1	0	1	0.48
food	1	0	0	1	0.48
foot	2	0	0	2	0.97
footnote	1	0	0	1	0.48
forebear	0	1	0	1	0.48
forest	0	1	0	1	0.48
Fossil	1	0	0	1	0.48
four	3	0	0	3	1.45
fruit	1	0	0	1	0.48
gait	1	0	0	1	0.48
game	1	0	0	1	0.48
glacier	0	2	0	2	0.97
gorilla	0	2	0	2	0.97
grassland	1	0	0	1	0.48
greenhouse	1	0	0	1	0.48
hallmark	1	0	0	1	0.48
hands	2	0	0	2	0.97
hardship	1	0	0	1	0.48
hip	0	2	0	2	0.97
hominid	1	13	0	14	6.79
Homo	4	0	0	4	1.94
hunter	0	1	0	1	0.48
ingredient	0	1	0	1	0.48
insect	0	1	0	1	0.48
intelligence	1	0	0	1	0.48
interruption	0	1	0	1	0.48

jaw	3	2	0	5	2.42
leopard	1	1	0	2	0.97
life	2	0	0	2	0.97
male	0	1	0	1	0.48
mammoth	0	1	0	1	0.48
man	3	0	0	3	1.45
member	0	1	0	1	0.48
menu	1	0	0	1	0.48
mile	0	1	0	1	0.48
nature	1	0	0	1	0.48
niche	1	0	0	1	0.48
organization	1	0	0	1	0.48
pelvis	1	0	0	1	0.48
pemman	0	1	0	1	0.48
period	0	3	0	3	1.45
plain	0	6	0	6	2.91
planet	1	0	0	1	0.48
pollution	1	0	0	1	0.48
predator	0	1	0	1	0.48
rainfall	1	0	0	1	0.48
recipe	1	0	0	1	0.48
scientist	0	1	0	1	0.48
shard	0	1	0	1	0.48
shellfish	1	0	0	1	0.48
side	2	0	0	2	0.97
size	2	0	0	2	0.97
specialist	0	1	0	1	0.48
species	6	0	0	6	2.91
spine	1	0	0	1	0.48
stance	1	0	0	1	0.48
stardom	1	0	0	1	0.48
story	2	0	0	2	0.97
success	1	0	0	1	0.48
teeth	5	0	0	5	2.42
temperature	0	1	0	1	0.48
ten	2	0	0	2	0.97
thighbone	0	1	0	1	0.48
tiger	0	1	0	1	0.48
time	2	0	0	2	0.97
upon	1	0	0	1	0.48
vulnerability	1	0	0	1	0.48
ways	2	0	0	2	0.97
weapon	1	1	0	2	0.97
wildlife	1	0	0	1	0.48
world	3	0	0	3	1.45
year	0	9	0	9	4.36

テキストの意味を理解するためには実語（full word）[10] の理解が必要である。実語（full word）は文法上の概念を表す機能語（function word）と異なり，独立した語彙的意味を持っている。テキストの意味を理解するためにはその意味の理解が必要不可欠である。難易度の高いテキストを読解教材として使用するためには事前に，そこで使用されている実語（full word）の使用状況を明らかにしておく必要がある。

3.3. 実語（full word）の使用状況

今回使用のテキスト，‘How Our Ancestors Survived the Ice Age’ 中で使用されている実語（full word）の中で，学習者にとって難解であると思われる語の一覧は下記の通りである。研究社の *NEW COLLEGIATE ENGLISH-JAPANESE DICTIONARY* を利用して語彙レベルを確認し，高校レベルで学習されていないと思われる語をテキストからリストアップした[11]。テキスト内で使用された形のままで表記してある。

表 2

accelerating	alternation	anchor	ape	attachment	baboons
buttock	chimps	chomp	coarse	compensating	conserve
crest	descent	diggers	disrupting	drastic	drought
equipped	evolution	evolutionary	exposed	extinction	fangs
flourished	flowing	footnote	forebears	fossil	fused
gait	glaciers	gradual	greenhouse	grinding	guts
half-crouch	hallmark	hominids	ingredients	knock-kneed	knuckle
lengthened	leopards	lush	mammoths	massive	niche
nimble	overthrown	pelvis	pitted	pluck	poles
predators	prey	respites	retreat	ridge	saber
shaggy	shards	shuffling	sketchy	skull	spine
stance	stardom	stick	stricken	sturdy	swiveled
tempting	thaws	thighbones	trek	tripling	unappetizing
vegetarian	vulnerability				

これらの語は学習者にとって未習である可能性が高い。これらの語をリストにしたものを事前に学習者に提示し意味を確認しておけば，実際に passage を読む際に学習者の負担を軽減することができる。また，どういう文脈で使用されているのかを事前に把握しておけば，授業を効率よく行うことができる。では次節では WordLab（English Corpus All-round Analyzer）を使用することによってどのようなことができるのかについていくつか具体的に説明していく。本稿で紹介するのは「語彙集計・品詞分類機能」,「All Search 機能」,「コロケーション統計機能」の 3 つである。

4．WordLab（English Corpus All-round Analyzer）の機能

4.1. 語彙集計・品詞分類機能

3.1. で説明したように WordLab の語彙集計機能を使用すれば短時間で，テキスト中で使用されているすべての語彙のリストを作成することができる。語彙集計を行った後で品詞分類を行えば，各品詞別の語彙リストを作成することができる。3.1.では名詞のリストのみを紹介したが，他の品詞についても同様のリストを作成することが可能である。WordLab で作成できる品詞リストは下記の通りである。

表 3

助動詞	一般動詞	副詞	否定詞	形容詞	冠詞
名詞	代名詞	疑問詞[12)]	接続詞	前置詞	間投詞

テキスト中でどのような語彙がどのような頻度で使用されるかを把握するためには非常に便利な機能である。

4.2. All Search 機能

テキスト中でどのような語がどれくらいの頻度で使用されているかが明らかになっても，それがどのようなコンテキストで使用されているか[13)]がわからなければあまり意味がない。ある語がどういう文の中で使われているかがすぐにわからなければテキストを電子化する意味はあまり無いと言わざるを得ない。WordLab では All Search 機能を使用することによって，問題となっている語がどこでどういう文中で使用されているかを即座に明らかにできるようになっている。例えば3.2.で示した表 2 に“hominids”という学習者にとっておそらく意味がわからないであろうと思われる語がある。All Search 機能を使用してテキスト内を検索すれば次のような結果が即座に得られる。

（1）It is often thought that man’s ancestors stood upright to free their hands for using weapons and tools, but the first hominids or upright “ape-men” appeared about four million years ago-several million years before the great explosion in intelligence and use of tools took place. [How Our Ancestors Survived the Ice Age1]2[14)]

（2）The first hominids had brains little larger than any other ape and the only immediate advantage that walking gave them was that they could trek miles across the landscape each day in search of food and shelter. [How Our Ancestors Survived the Ice Age1]3

（3）For the hominids to survive their early years on the exposed plains, they must have depended upon having close-knit social groups, compensating for their vulnerability by being alert and organized. [How Our Ancestors Survived the Ice Age1]6

（4）It was what allowed monkeys, Like baboons, successfully to join the hominids on the plains. [How Our Ancestors Survived the Ice Age1]8

（5）Whatever the combination of factors that allowed early hominids to survive on the plains, they moved in a relatively short time from an awkward chimplike shuffle to a flowing stride. [How Our Ancestors Survived the Ice Age1]11

（6）The thighbones sloped together until they almost touched at the knee, giving a knock-kneed look and allowing hominids to put one foot efficiently in front of the other. [How Our Ancestors Survived the Ice Age1]14

（7）In next to no time, the first hominids had human hips and legs to match the human arms and shoulders gained in the tree-swinging period. [How Our Ancestors Survived the Ice Age1]17

（8）The first hominids had human legs and bodies but small apelike heads. [How Our Ancestors Survived the Ice Age1]22

（9）Although they had slightly larger brains than the average ape, the real increases in brain size did not take place until long after the hominids had made a success of walking. [How Our Ancestors Survived the Ice Age1]23

（10）Indeed, even when hominids had got this close to being human, it was still possible that the human tripling in brain size that was to come might never have happened. [How Our Ancestors Survived the Ice Age1]24

（11）Scientists have only a sketchy picture of the several species of hominids that lived on the African plains about three to four million years ago. [How Our Ancestors Survived the Ice Age1]29

（12）One line of hominids was slimly built and quickly developed a weak jaw yet large brain. [How Our Ancestors Survived the Ice Age1]32

（13）Both ways of life were fine for a while and the two families of hominids appear to have lived side by side. [How Our Ancestors Survived the Ice Age1]39

この機能を使用すれば問題となっている語がテキストのどこでどういうコンテキストで使用されているかがすぐにわかる。本稿で扱っているテキストでは“hominids”という語が13回も使用されている。このように高頻度で使用される語彙は実際にテキストを読む前に例文を与えながら学習者に提示しておくことが望ましい。

4.3. コロケーション統計機能

英語を学習する際にはひとつひとつの語を独立して覚えていくのではなくコロケーションに

注意を払うことが必要である。このことについては多くの研究者がすでに指摘している。例をいくつか挙げる。

> 語と語の正確な共起関係の知識を学習者が身につけているということは，英文法の正確な知識を持っているという事実に勝るとも劣らない重要な事柄である（稗島　1990: 198)。

> Learners of English as a foreign or second language, like learners of any language, have traditionally devoted themselves to mastering words...their pronunciation, forms, and meanings. However, if they wish to acquire active mastery of English, that is, if they wish to be able to express themselves fluently and accurately in speech and writing, they must learn to cope with the combination of words into phrases, sentences, and texts.
> Students must learn how words combine or 'collocate' with each other. In any case, certain words regularly combine with certain other words or grammatical constructions. (Benson et al. 1997:ix)

WordLab にはコロケーション統計を行う機能がついている。この機能を用いることによってある語が他の語とどのような組み合わせで使用されるのかを明らかにすることができる[15)]。表 4 は“hominids”を KWIC 軸として作成したコロケーション統計の結果である。

表 4

近接語	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	左方	右方	計
the	0	2	0	4	3	0	1	0	2	0	9	3	12
to	0	0	1	0	0	3	1	0	0	1	1	5	6
had	0	0	0	0	0	5	0	0	0	0	0	5	5
and	1	0	0	1	0	0	0	0	3	0	1	3	4
first	0	0	0	0	4	0	0	0	0	0	4	0	4
of	2	0	0	0	3	0	0	0	0	1	3	1	4
on	0	0	0	0	0	1	0	2	0	0	0	3	3
human	0	0	0	0	0	0	2	0	0	0	0	2	2
survive	0	0	0	0	0	0	2	0	0	0	0	2	2

表の見方を簡単に説明しておきたい。L1 と“first”がクロスする部分が 4 になっている。これは“first hominids”という組み合わせが4例あったことを示している。表中の L は左を，R は右を表している。この機能を用いることによってキーワードの前後，つまり左右 5 語ずつの範囲でどのような語が用いられているかを明らかにすることができる。本テキストの場合であれば「first hominids＝最初の人類」という形で提示することが望ましい。

本稿で扱っているような小規模なデータではコロケーション統計を行っても意義のある結果は得られないが，コーパスの規模が大きくなればなるほどそこから得られる情報は価値のあるものになってくる。

5．まとめ

本稿では電子化したテキストと言語分析ツールを使用することによってどのようなことが可能になるのかについて具体例を示すことを試みた。もちろん，本稿で示したことはごく一部であり他にも様々な可能性が考えられる。また本稿で示したことは，時間と労力を惜しまなければすべて手作業で行うことが可能である。事実，コンピュータが普及する以前は，コンコーダンスの作成等の作業はすべて時間と労力をかけて手作業で行われていたのである。それがコンピュータと言語処理ツールの普及によって誰でも短時間で簡単に行えるようになってきたのである。本稿で示したような使用方法は誰にでも簡単に行えることである。

最後に，リーディングの指導の際に，なぜ本稿で説明したような教材提示が必要なのかについて述べておきたい。学習者のレベルが低くなってきた場合，教える側はどのように対応すればいいのであろうか。教材そのもののレベルを下げるのが一番簡単な方法であろう。しかし，これは他の策が尽きてしまった時の最終手段であると筆者は考えている。安易に教材のレベルを下げることは結果として教育のレベルそのものを下げることに他ならないのである。教材のレベルを下げなくても，事前にテキスト内で使用される語彙を学習者に提示しておく，学習者にとって理解が困難であろうと思われる文については事前の説明を加える等の手段を取れば難易度の高い教材であっても使用可能である。そしてこのような作業を行う際に，言語分析ツールは大きな力となってくれるのである。

本稿で紹介したのは，WordLab の持つ様々な機能のごく一部である。本稿で扱わなかった高度な検索機能については稿を改めて扱いたい。

註

1）1961年に出版された文献から各2000語のテキスト500，総数約100万語を集めたアメリカ英語のコーパス。

2）詳細については http://info.ox.ac.uk/bnc/index.html を参照。

3）詳細については http://titania.cobuild.collins.co.uk/ を参照。

4）2002年11月現在で，校正未完了文を含めて5500作品が登録されている。2004年までに計6500になる予定である。詳細については http://gutenberg.net/ を参照。

5）WordLab の入手方法については http://www.wordlab21.com/ を参照。

6）詳細については http://www.biwa.or.jp/~aka-san/index.html を参照。

7）詳細については http://www.liv.ac.uk/~ms2928/wordsmit.htm を参照。

8）WordLab に備わっている品詞分類機能は，原文の前後関係に基づいた品詞分類を行なっていないた

め，その分類が正確ではない。いずれにしても使用語彙のリストを作成する上では大きな問題にはならない。必要があれば，原文に戻って確認することが簡単にできるようになっている。

9）WordLab を使用して分析した結果を SDF 形式で保存したものである。アルファベット順に並べ替えてある。

10）一般的には名詞，代名詞，形容詞，動詞，副詞のことと考えて問題ない。

11）今回はこの作業を手作業で行ったが，単語レベルをチェックするソフトも入手可能になっている。㈲イー・キャスト（http://www.e-cast）の「単語レベルチャック」を使用すれば作業時間を大幅に短縮することができる。

12）関係詞もここに含まれる。

13）秀丸のような Editor の検索機能を使用してもこの作業を行うことができる。しかし Editor ではコロケーションの処理等を行うことができない。

14）数字はテキスト内の何番目の文であるかを示している。

15）コロケーションを調べる際は，大規模コーパスを用いて行わなければ本来は有効な結果が得られない。本稿の目的はコロケーション統計を行う機能の紹介をすることでありコロケーション研究を行うことではない。

一次資料

McCrone, J. 1991. ‘How Our Ancestors Survived the Ice Age.’ in Department of English, The University of Tokyo, Komaba (ed) 1993. *The Universe of English*. Tokyo: University of Tokyo Press. 72-79.

参考文献

1） Benson, N., E. Benson, and R. Ilson. (eds.) 1997. *The BBI Dictionary of English Word Combinations*. (Revised edition) Amsterdam/Philadelphia: John Benjamins Publishing Company.

2） Granger, S. (ed.) 1998. *LEARNER ENGLISH ON COMPUTER*. London and New York: LONGMAN.

3）中條清美・長谷川修二 2003.「時事英語の授業で用いられる英文素材の語彙レベル調査－BNC（British National Corpus）を基準にして－」『時事英語研究』XLII，51-62.

4）齊藤俊夫・中村純作・赤野一郎（編）1998.『英語コーパス言語学－基礎と実践－』東京：研究社出版

5）八木克正 2003.「コーパスを利用した英語教育と英語・英文学研究指導－実践報告と今後の可能性」『英語コーパス研究』第10号，247-248.