

教科書コーパスから何が見えるか

— 高等学校英語教科書の場合 —

山添孝夫

中学校の教科書に続いて高校の教科書ではどうなるかを見てみます。このコーパスに取り組む以前から高校の教科書の本文データを個人的に集めておりました。そのデータを利用してコンコーダンスラインから練習問題を作って授業中に使っていました。BNCなどの本格的なコーパスのコンコーダンスラインでは生徒が英文を理解できないので、生徒がわかるようなデータを中心にという観点でデータを集めてきていました。Willisがpedagogic corpusと呼んでいる「学習に使用した言語材料を集めたコーパス」を作っていたことになります。ただ高校用の教科書全部を集めていたわけではないので、この機会に全教科書が入ったコーパスが作れないかと考えていましたところ、立命館で勉強するチャンスをいただき今回高校英語教科書コーパスに本格的に取り組むことができました。

1. コーパス化した教科書と処理

今回、高校の教科書でコーパスの材料として使用したのは「英語」35冊「英語」37冊です。「英語」は2種類の教科書が2分冊になっているため全体では37冊ということになっています。教科書名でまとめると35種類の教科書としてまとめられると思います。具体的な教科書名は別に表にまとめましたので、そちらをご覧ください(Appendix 1)

高校の教科書は大別するとA5判、B5判、B5変形判の3種類の判型に分けることができます。英語担当教員の間では「判型の大きいものが、小さいものよりも易しい英文で書かれている教科書」(易B5 > B5変型 > A5難) という認識があるようです。この直感的分類が本当に正しいといえるのか、また、たとえばB5判サイズの教科書群で、それぞれの語彙の違いや文の長さの違いなども統計的に比べてみようと思いました。

以前に教科書の語彙を調査されたものがいくつかあります。一部をあげれば、垣田・三浦(1987)や塩見(2002)、杉浦(2002)などがあります。杉浦(2002)は2000年度発行の「英語」「英語」「リーディング」の教科書本文をコーパス化しています。レマ化された語彙リストがインターネット上に公開されています。

コーパスデータの収集と分析は、先ほどの中村先生の中学校教科書の場合と同じような手順を踏んでいきました。図1のような流れで本文を電子化し、Special Marksをつけました。マークの一覧を図2にあげます。

固有名詞は語頭に@マークを付けました。たとえば国名を例に取るとAmericaは@Americaのようになります。固有名詞の形容詞形について今回はAmerican, Japanese, Chineseなどには@マークを付けずにそのままにしました。New YorkなどのMulti-Wordはスペースで分割されると具

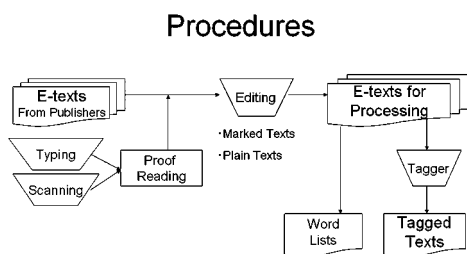


図1 編集作業の流れ



図2 編集時に付加した記号

I'm @Catherine=Taylor. I come from @Melbourne, @Australia.

I like music. I play the drums, and now I'm interested in Japanese &taiko drums.
I'd like to play the &taiko myself.

In @Australia I have a father, a mother, and a brother. My father is a flying doctor.
Flying doctors visit patients in distant, lonely places by airplane. My father is very busy.
Perhaps he's flying in the air now. I'm very proud of him.

In @Australia most schools don't have after-school clubs. This school has a lot. I'd like to join the music club.

I like this school very much.

図3 マークを付加した教科書本文の例

合が悪いので、=マークを付けて結びつけました。New Yorkなどの例は短くて良いのですが、映画監督のスピルバーグを扱った課の中に University of Southern California Film School という学校名が出てくるのですが、これも multi-word のひとつのまとまりとして扱いました。New York などと比べるとかなり文字数の多いものとなりましたが、この場合も、統計処理上は1語の扱いになります。後で述べます一文に何 word あるかという値に若干影響を与えている可能性があったかもしれません。実際の教科書本文にマークを付けた例をあげます(図3)。

それぞれの教科書本文を英語 と英語 をつないで一本のファイルにし、流れ図にあるような処理を Frequency Analysis Toolkit によって行いました(図4, 図5)。処理ツールについては中村先生から先ほど説明がありましたので割愛させていただきます。ツールの簡単な説明は Appendix 2 をご覧ください。特徴的な点は、ある段階の処理で作成されたファイルが次の処理の入力ファイルになっていることです。また作成されたファイルは ASCII テキストファイルですので、内容の確認が簡単にできます。

この処理によってできあがってきた語彙のリストの一部が表1です。行はそれぞれの語彙項目、列は教科書名を表す略号とそれぞれの語彙の頻度を表しています。この表は固有名詞を除

Frequency Analysis Toolkit (中村: 2005)による処理 (1)

- テキストファイルから単語頻度情報、分割表作成、数量化などの処理およびグラフィック表示を行うプログラム群(中村純作教授作成)

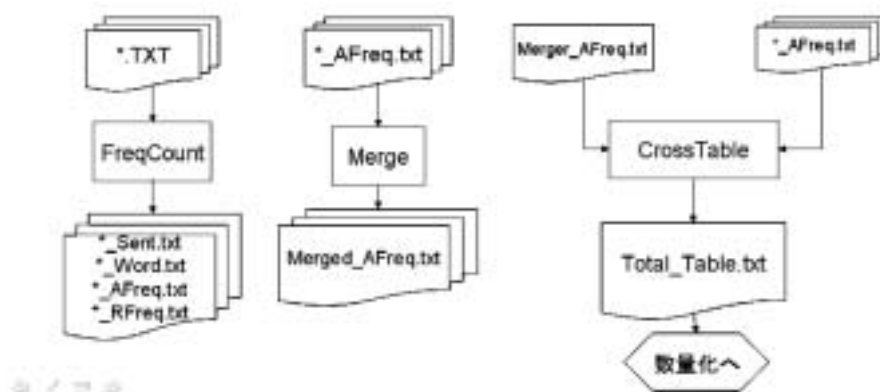


図4 FreqAnalysisToolKitの処理(1)

Frequency Analysis Toolkit (中村: 2005)による処理 (2)

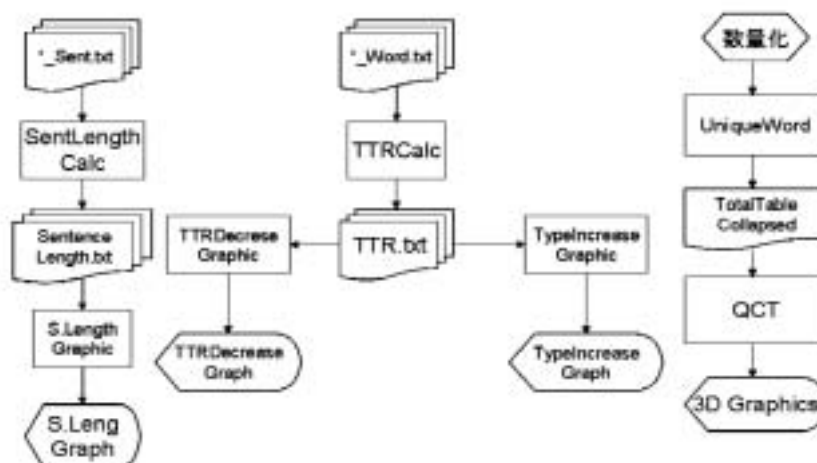


図5 FreqAnalysisToolKitの処理(2)

外した後の値になっています。総語数で36万語強になります。異語数では12,000語弱になります。

表1. 語彙頻度表の一部

Word	Total	AAB	PON	PRM	SSH	NOW	EXC	CRW
the	21029	151	558	821	851	311	744	821
to	11175	109	285	528	456	141	338	495
and	9152	96	222	476	378	119	233	425
of	8204	56	205	383	351	105	279	400
in	8089	79	207	385	267	128	269	353
i	6160	93	131	315	275	40	136	305
was	4784	43	140	205	170	69	119	177
you	4291	68	97	174	168	44	135	244
is	4178	67	80	219	123	58	174	232
it	4085	55	90	160	165	52	156	216
he	4023	54	154	183	121	62	181	149
that	3992	18	81	185	170	37	119	208
for	3458	30	74	156	128	47	102	150
they	2651	14	66	101	100	44	95	114
on	2599	23	79	109	102	28	85	104
we	2585	29	50	147	96	31	76	166

2. Token数とType数の比較

次に先に述べました教科書の判型サイズでグループ化して比較を続けます。語彙数の面から比較を始めます。まず図6のB5判サイズの教科書比較です。Token数では約4,000から約6,300とばらつきがあります。Type数は約1,100から約1,500語で揃っています。

次に図7のB5変形版ではToken数は最も少ないSTPの約4,000から最も多いPOLの約16,000と、ばらつきが非常に大きくなっています。Type数で見えていくと最も少ないSTPの約1,100と最も多いCSMの約2,900と、こちらもB5判グループの場合よりもばらつきが見られます。STPのToken数4,057Type数1,102という値は、先のB5判グループと同じ範囲に入ることになります。この点でSTPはサイズB5変型判ですが、B5判グループとして扱った方が良いのかもしれない。

次に図8のA5判サイズの教科書の比較です。このグループには伝統的なCRWやUNIなどが含まれます。入力チェックの段階で、他の教科書と雰囲気が違うと感じたものがありました。ENCです。通常高校1年生の最初の3レッスン程度は中学の復習で、新しい文法は出てきません。新しい語彙もあまりたくさん出てきませんが、この教科書はいきなりnew wordsがずらり並んでいて、本文もビッシリ英語が入っています。第一印象で圧倒されるような感じの教科書でした。後でご覧頂く数量化の画面でも他教科書と違った場所にポツンと出てきますので、頭の片隅に入れておいてください。Token数では最も少ないSTGで約11,100語、最も多いUNIで約17,800語。Type数ではSTGの約2,300語からENCの約3,100語です。Token数ではばらつきがありますが、先ほどのB5変型判グループほどではありません。Type数から見ていくと、約2,500語程度のあたりに平均がありそうです。

教科書コーパスから何が見えるか：高等学校英語教科書の場合（山添）

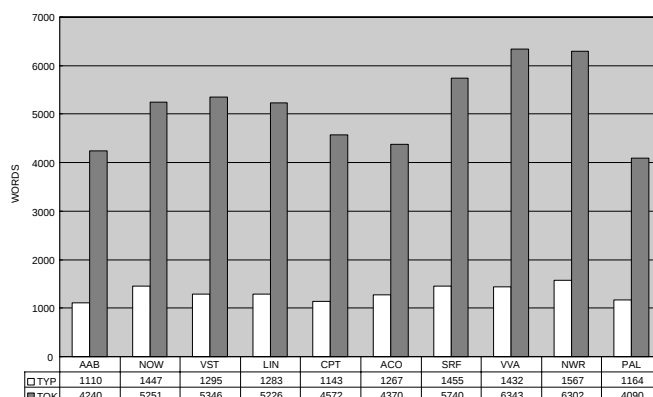


図6 B5判教科書のToken数Type数比較

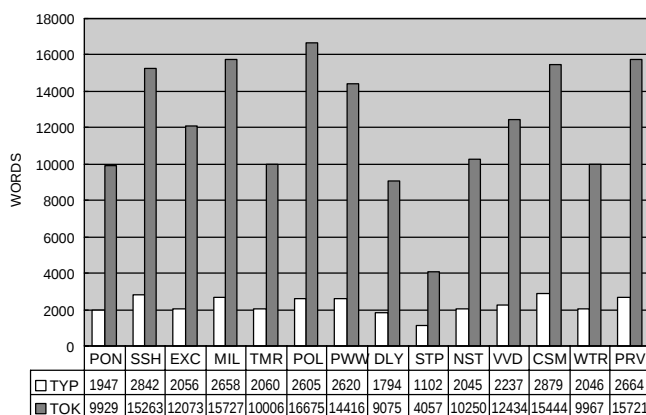


図7 B5変形判教科書のToken数Type数比較

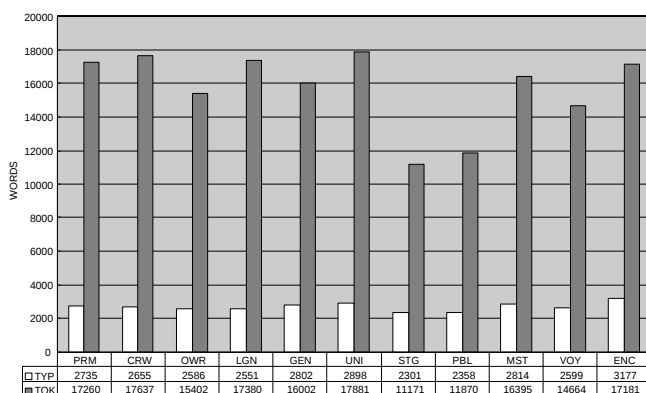


図8 A5判教科書のToken数Type数比較

3. 文の累積平均長

次に文の累積平均長を比べていきました。教科書ごとに文の総数が違うのですが、グラフ化して一定の部分、たとえば400文での累積平均文長を比較することができます。

B5判サイズの教科書について図9をご覧ください。400文を超えたあたりから、2つのグループに分かれているのが見えます。NOW, SRF, VSTの3種類が1文あたり9語から10語の範囲にあります。もう一つのグループはVVA, NWR, LIN, CPT, PAL, ACO, AABが1文あたり7語から8語の範囲にあります。

B5変型判については図10をご覧ください。このグループも400文を超えたあたりから、2つのグループに分かれています。2種類の教科書SSH, MILが最終的に1,200文で1文あたり12語から13語という長さになっています。残りの教科書は900文程度でグラフが終わっているものと、その後1,000文、1,200文、1,400文を超えるものもあります。1種類の教科書POLが1,600文を超えています。文の総数は異なるのですが、1文あたり10語から11語の範囲にあります。文の総数が1,400文を超える教科書は、文の総数という点ではこの後述べるA5判サイズの教科書と同じ値です。

A5判サイズの教科書の平均文長の変化は図11をご覧ください。文の総数が1,000文に達しないのはSTG, PBLの2種類、1,200文から1,300文の範囲に入る教科書がVOY, OWR, GEN。1,300文を超える教科書はENC, UNI, LGN, PRM, CRW, MST。ENCが1文あたり13語あたりで振れていることと、MSTがコンスタントに1文あたり11語のまま推移していることを除くと、他の教科書は1文あたり11語から12語の長さの範囲に入っています。

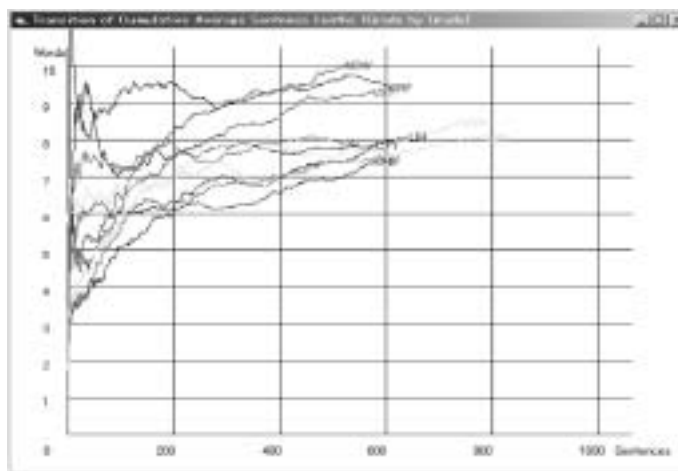


図9 B5判教科書の累積平均文長推移

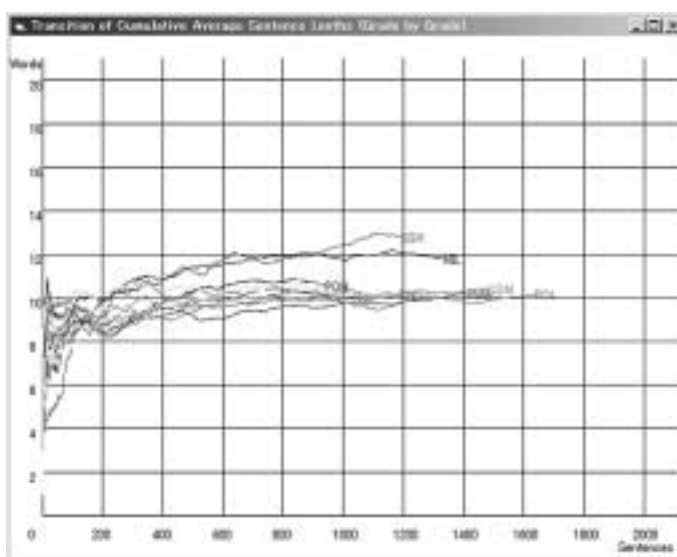


図10 B5変形判教科書の累積平均文長推移

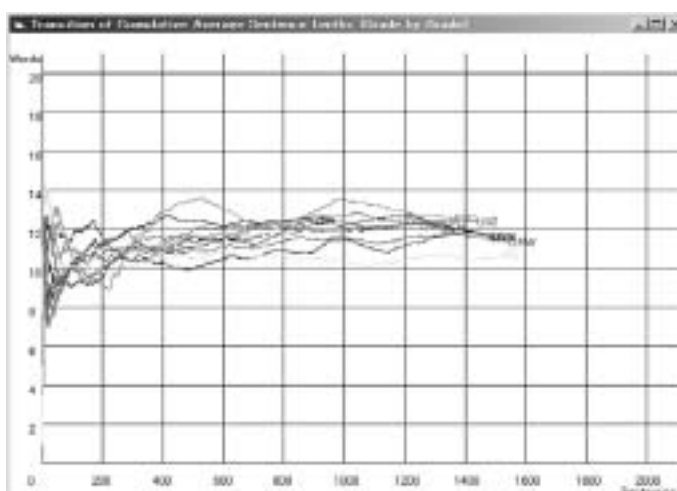


図11 B5変形判教科書の累積平均文長推移

4 . Type Token Ratio の推移

続いてToken数の伸びに伴うType Token Ratio (TTR) の変化をグラフで見えます。一般的にTTRが1に近い値をとるとテキスト内に含まれる語のバラエティーが豊富であると考えられます。学習者の立場で考えると新しい単語が多く含まれているということになります。

まずB5判サイズの教科書のTTR変化は図12をご覧ください。すべての教科書が通過している

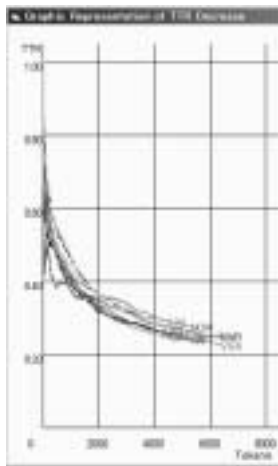


図12 B5判教科書TTR推移

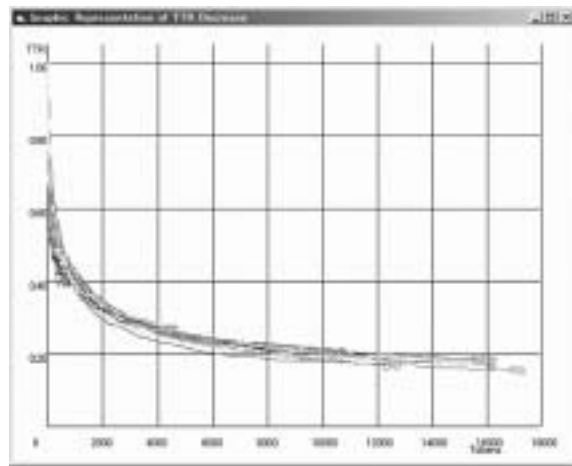


図13 B5変形判教科書TTR推移

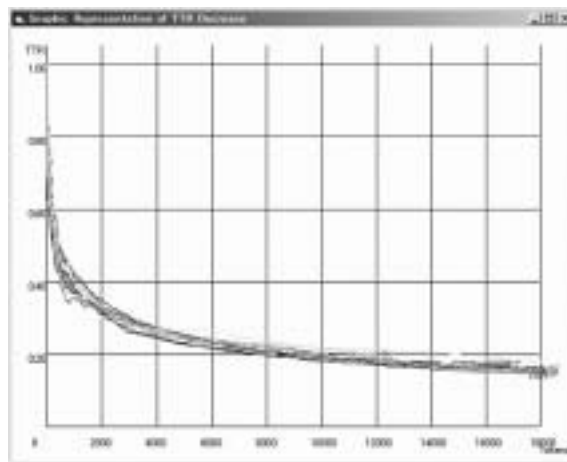


図14 A5判教科書TTR推移

4,000 Tokenで比較しますと、約0.3程度で、それほどのバラツキは見られません。

B5変形判については図13をご覧ください。Token数・Type数の棒グラフからもわかっていたことですが、STPは4,000 Token程度しか語数がありませんので、グラフには描かれているのですが、ここでは除外して考えます。総Token数は10,000から16,000超まで幅がありますが、TTRは0.2を少し下がる程度になっています。

A5判教科書のTTR推移は図14をご覧ください。各教科書がほぼ同じ曲線を示しています。PBLが総Token数が12,000弱ですので、全体の比較は11,000 Tokenでしか行えませんが、TTRは0.2を少し下回る値になっています。ENCは、これよりも少し高い値を示しています。ENCは他の分野の比較でも他の教科書と違う傾向が見られましたが、ここでもその傾向がみられます。ENCのグラフは3,000から6,000 Tokenにかけて平板状になっています。この部分は新出語

が次から次へと導入されていることを示しています。各教科書のTTRは最終的には16,000Tokenを超えた辺りまでTTR下がりますが、ENCのTTRは他の教科書群よりも少し高い値のままで推移しています。

5 . Type 数の推移

TTRの推移を逆の視点から見たものが次のToken数の伸びに伴うType数の推移です。新しいTokenが導入されると、Type数がどう変化するかをグラフで示しました。

まずB5判教科書について図15をご覧ください。1,000から2,000 Token辺りまでは各教科書のグラフが重なるようになっていますが、2,500 Tokenを超える辺りから少しグラフが分かれ2つのグループが見られるようになります。

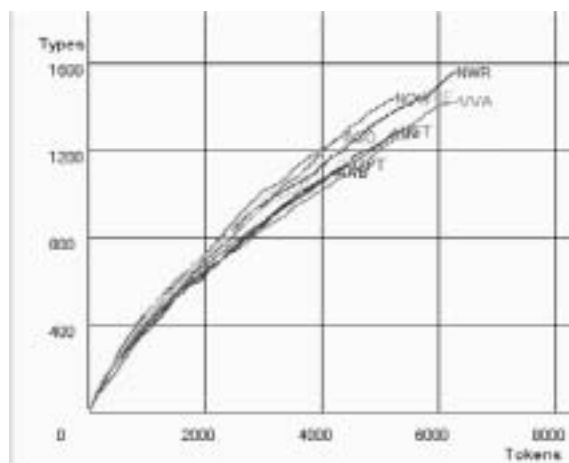


図15 B5判教科書Type数推移

B5変型判教科書のグラフは図16をご覧ください。こちらも3,000 Token辺りまではすべての教科書のグラフが重なっています。4,000 Token辺りでENCの傾きが他の教科書よりも少なくなり、その傾向は変わりません。他の教科書も6,000 Tokenを超えた辺りから更にグラフが分散するようになります。Type数の伸びが他の教科書よりも速いものはSSHとCSMの2種類です。

A5判教科書のグラフは図17です。2,000 Token辺りまではグラフが重なっています。3,000から4,000Token辺りをすぎるとグラフが分かれ始めます。このグラフでもENCが他の教科書と比べてType数の伸びが著しいことがわかります。先ほどのTTRの推移のグラフでは平板状になって現れていた部分は、このグラフではグラフが急に立ち上がっていて新語彙の導入ペースが他の部分よりも速くなっていることを示しています。他の教科書群はグラフが分散し始めているとはいえ、12,000 Token辺りまでは、それほど大きく分散しないで同じような傾きで伸びています。このグループはグラフが急に立ち上がる部分が見られないので、新しい語彙の導入ペースは一定であると予測できます。

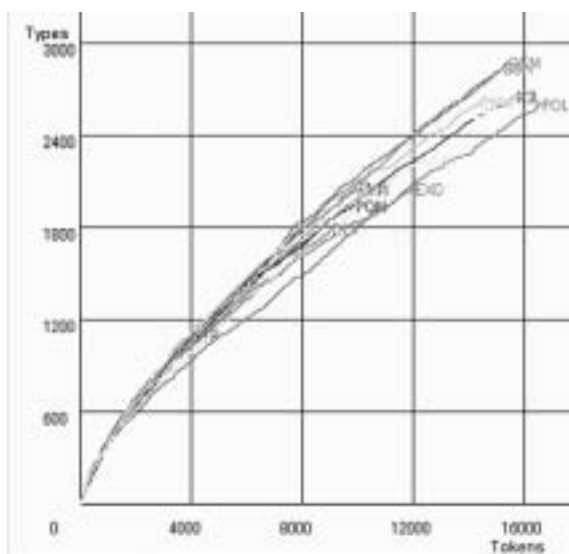


図16 B5変形判教科書Type数推移

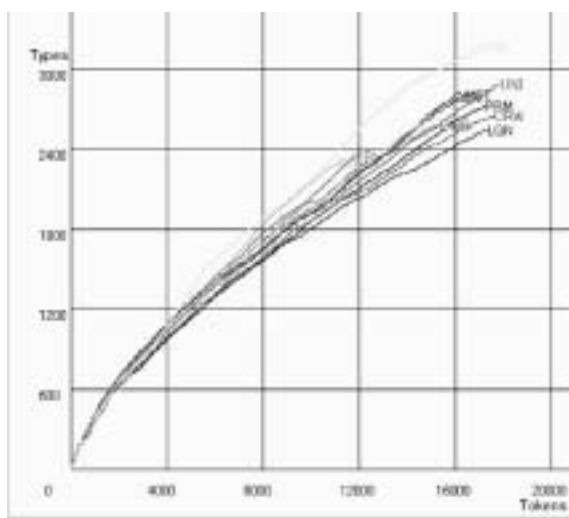


図17 A5判教科書Type数推移

6．語彙頻度表と数量化

語彙頻度表から英語の全教科書に共通の語を抽出してみました。83 Type ありました，その一部を表2にあげます。

語彙頻度表に対して数量化を行うと，ある教科書に特有の語はすべて同じ数量を与えられます。それらの語は三次元スペースに表示しますと，すべてが同じ位置に重なり非常に見づらく

表2．英語 全教科書に共通の語彙項目（一部分）

WORD	TOTAL	AAB1	PON1	PRM1	SSH1	NOW1	EXC1
a	4139	45	120	214	128	55	126
about	735	7	20	43	23	9	24
all	614	2	18	24	18	7	16
an	405	3	10	14	11	6	8
and	3886	37	82	242	137	49	91
are	1111	11	16	42	25	24	37
as	744	2	19	37	40	4	21
at	915	9	20	43	35	9	20
be	668	9	10	29	37	5	22
because	277	6	10	11	10	6	10
but	979	8	29	40	35	14	38
by	481	1	9	24	11	8	12
can	668	15	22	33	21	3	24

	AAB	PON	PRM	SSH	NOW	EXC
1 AAB	35	88	185	231	73	116
2 TYPES	68	136	253	322	91	181
3 TOKENS	1	1	1	1	1	1
4 anti-laminar	1	1	1	1	1	1
5 astronomer	1	1	1	1	1	1
6 bass	1	1	1	1	1	1
7 bear's	1	1	1	1	1	1
8 calligrapher	1	1	1	1	1	1
9 calligraphy	1	1	1	1	1	1
10 composition	1	1	1	1	1	1
11 corsets	1	1	1	1	1	1
12 croissants	1	1	1	1	1	1
13 cucumbers	1	1	1	1	1	1
14 disk-type	1	1	1	1	1	1
15 dream-mate	1	1	1	1	1	1
16 ex-boyfriend	1	1	1	1	1	1
17 explorers	1	1	1	1	1	1

図18 各教科書に特有の語（一部）

なります。それを防ぐために事前に各教科書に独自の語を抽出し、それらの語の頻度合計を教科書名でまとめ1つの語彙項目として扱います。この処理は‘UniqueWord’ toolで行います。各テキストに特有な語のリストの一部を図18にあげます。

先ほど中学校の教科書の分析で説明と同様に、35種類の教科書がCategory、各語彙項目がSampleになります。数量化の結果得られるCategory Weights、Sample Scoresは34組の数値が得られますが、それぞれ10組の数値を図19、図20にあげます。これらの数値のうちCategoryとSampleの間に最も高い相関係数を得られる上位3組を用いてグラフィック表示を行います。

Category	Weight
3 AAB	-0.881296227125279
4	-0.394865153020962
5	-0.266197827802931
6	-0.615612200464488
7	-1.53346370815962
8	-1.63119540515387
9	-1.36091304172042
10	-2.64687309397559
11	-0.68583750601993
12	-0.865424565482745
13	

図19 Category Weights

Sample	Score
1228 a	-2.24601324175849E-02
1229	-4.92881115965628E-02
1230	-3.09652434980883E-02
1231	-3.45914001418743E-02
1232	-7.33736728269858E-03
1233	-0.134416113892136
1234	-4.39748317604308E-02
1235	-2.15244651812906E-02
1236	-5.75367109969643E-03
1237	-8.39521456593072E-03
1238	

図20 Sample Score

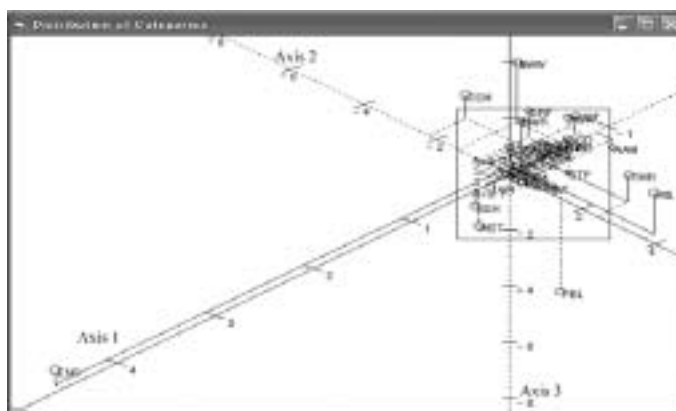


図21 数量化の結果の三次元表示（教科書の分布）

7. 数量化の結果図示（語彙頻度情報に基づく分布）

図21に各教科書の分布を三次元座標で表示しました。座標の原点付近に教科書名が現れている教科書はすべての語彙項目を満遍なく用いており中立的な教科書といえます。それに対して各軸上で原点から離れた位置に名前が出ている教科書はその語彙使用が独特のものとなっています。先に述べました各種統計から得られた情報ではENCが他の教科書と違った様子が見られましたが、この座標上では1軸のかかなり離れたところに位置しています。教科書本文のチェックをされていて直感的に感じた「他の教科書と違う」という感覚が実際に目に見える形で示されていると思います。1軸の負の側にはAABが原点付近の集団から少し離れて位置しています。

2軸上での位置関係に目を向けると、MIL, TMR, PBLが2軸の正の側にあり、CSM, PWWな

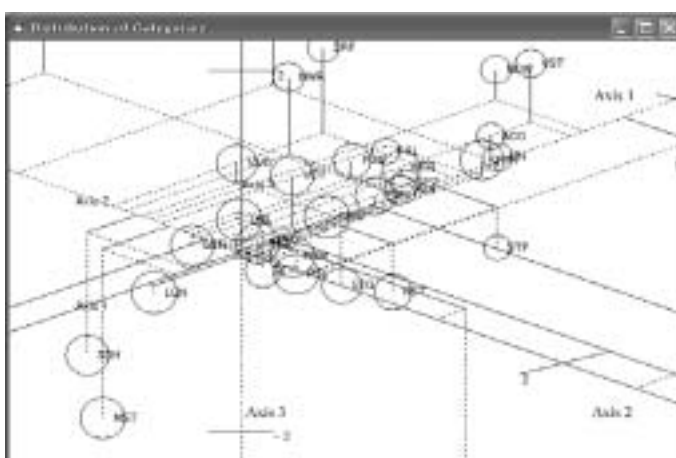


図22 図21の原点付近拡大図

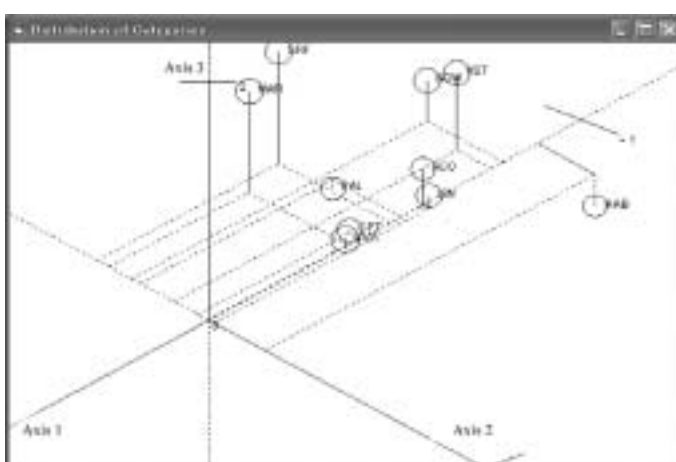


図23 B5判教科書の分布

どが負の側に離れて位置しているのが見えます。3軸上の位置関係はPWV, CSM, TMR, MILが正の側に、PBL, MST, SSHなどが負の側に離れて位置しているのがわかります。原点付近の分布状況を拡大して見てみると、図22のようになります。さらにこの部分に位置しているB5判サイズの教科書だけを表示してみると、図23のようにすべての教科書が1軸上の負の側に位置していることがわかります。2軸での位置関係をみると、AABは2軸の正の側に位置し他の教科書は全て2軸の負の側に位置していることがわかります。

B5変型判教科書を表示したものが図24です。ほとんどの教科書は原点付近に集まっていますが、MIL, TMRが2軸上で正の位置に、CSM, PWV, VVD, SSHなどが2軸上で負の位置にあります。A5判サイズの教科書の分布は図25です。原点付近の拡大図が図26です。先にも述べましたENCが大きく離れている他にOWRが1軸上で負の値を取っています。PBL, MSTが3軸上で負の値を取り他の教科書から離れたところに位置しています。

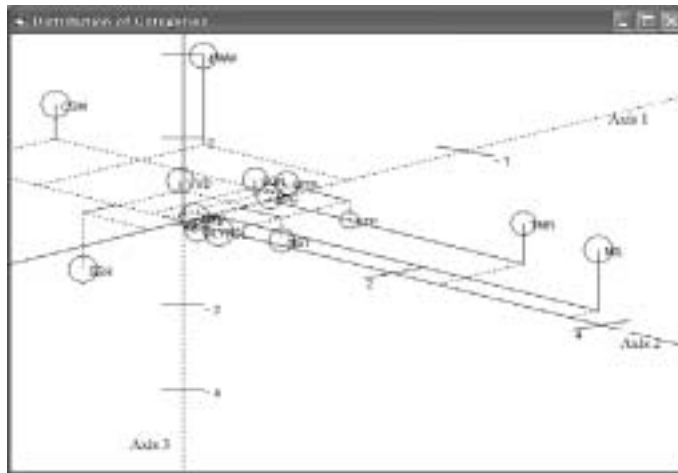


図24 B5変形判教科書の分布

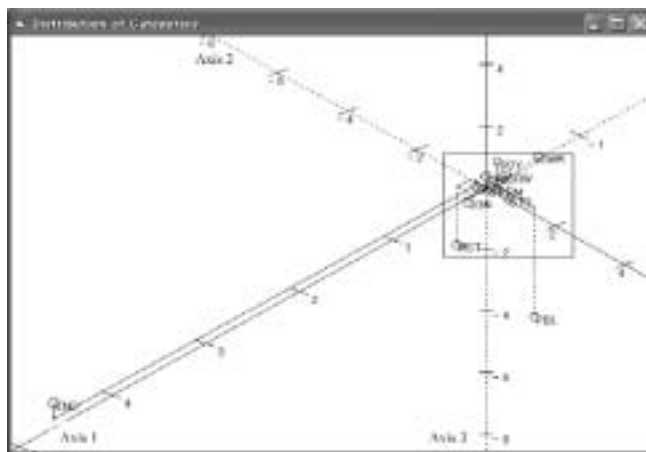


図25 A5判教科書の分布

全てのサンプルを表示したものが図27です。原点付近に来ている語彙項目は各教科書で満遍なく使用されている語彙項目になります。またこの図からはわかりませんが、使用頻度の高い語は画面上に表示される球の大きさが大きくなっています。また画面内に教科書の略号、たとえば画面左下にENC右下にPBLという2つの教科書の略号が見えますが、これらはその教科書にだけ使用されている語彙項目をまとめたものが表示されているものです。ENCの位置にはENCでのみ用いられた語彙の全ての頻度を合計したものが、語彙全体の中での比率に応じて小さな球が線の先端に表示されています。その付近に見られる他の語を示す線では線の太さと同じ程度の点で、その語の相対頻度が表現されています。また、ENCとPBLを結ぶ線分を仮定すると、その線上に表示されている単語はENCとPBLにだけ用いられている語彙ということになります。その2つの教科書でどちらに多く用いられているかによって、PBLに近い側あるいはENCに近い側に、それぞれ該当する単語が表示されています。同様にENCとPWWを結ぶ線分

が画面には
も可能です。
の語が原点
とがわかり
の大きさも

書に固有の
けられてい

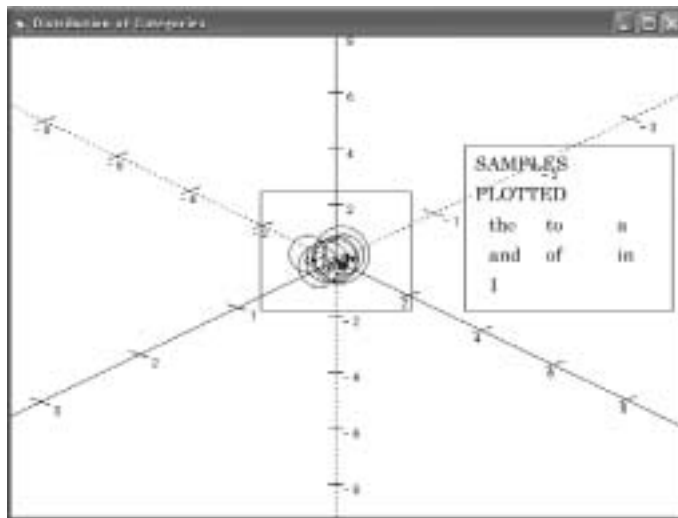


図28 頻度5,000以上の語彙項目の分布

8. 数量化の結果図示（代名詞の頻度情報に基づく分布）

語彙頻度表から代名詞のすべての変化形を抜き出して、それらを一人称単数・一人称複数・二人称・三人称単数男性・三人称単数女性・三人称単数中性・三人称複数のcategoryにまとめました。それぞれのcategoryはアルファベット大文字で 'I,' 'WE,' 'YOU,' 'HE,' 'SHE,' 'IT,' 'THEY,' と表現しました。Sampleは各教科書になります。このデータを使って同じように数量化を行いました。この分布図は図29です。

一人称単数の'mine'が曲者でした。「私のもの」の意味で使用されているものと、「地雷」や「炭坑」の意味で使われているものも'mine'の頻度の中に含まれています。特に今回扱った教科書ではクリス・ムーンという地雷除去に関わって自らも犠牲となりながらも義足でマラソンに参加している人物の話題がいくつかの教科書で扱われていました。コンコードダンサーで'mine'を含む行を検索して1つ1つ文を読んで「地雷」や「炭坑」の意味の例は除外して頻度数を調整しました。

HEとITが原点付近に来ています。これは各教科書に満遍なく使用されていることになります。原点から離れたところに位置する代名詞はある教科書に特徴的に使われていることになります。

'I,' 'WE,' 'YOU,'は1軸の正の側に来ています。逆に'THEY,' 'SHE,'は1軸の負の側に来ています。これらの2つのグループは教科書によって使われ方が違うということを示しています。特に'SHE'については原点からかなり離れたところに位置しているので、他の代名詞と扱われ方がかなり違うことを予測させます。

'I,' 'WE,' 'YOU,'はパーソナルな話題に関わる代名詞であるので、会話文や日記など個人に関わった内容を扱っていることが予想できます。'IT,' 'THEY,'は第三者的な視点で物事を述べているような解説文のような内容を扱っているのではと予想できます。'SHE'については今のところ詳



次に教科書の分布を見てみます。図30が代名詞の使用に基づく教科書の分布です。代名詞と教科書の分布位置が似ているものは 'I', 'WE,' 'YOU'のあたりに NWR, UNI, CRW, SSH, POL, CPTなどが出てきています。'THEY'の位置には OWR, WTR, LINなどが来ています。図29と図30を重ねてみると、その傾向がわかりやすくなります。図31に2つの図を重ねてみました。

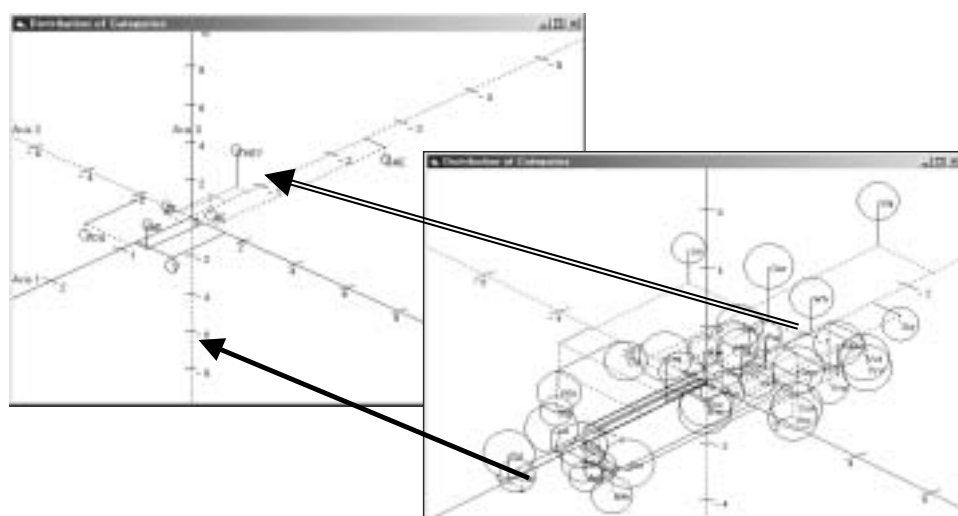


図31 代名詞と教科書の分布の対応

表3 . 4-gram の例

頻度	4-gram	頻度	4-gram
85	all over the world	38	do n't want to
52	at the end of	37	for a long time
51	for the first time	36	at the same time
50	on the other hand	35	one of the most
45	the end of the	34	i do n't think
40	i 'd like to	32	what do you think
40	i do n't know	32	i 'm going to
39	in the united states	30	why do n't you

35冊の教科書を大ざっぱに見てきました。学校現場では教科書の判型で慣例的に教科書の難易度を分類していますが、同じ判型の教科書であっても文の平均長や総数・総語数などにも違いが見られました。グラフから読み取った新出語彙の導入スピードも教科書によって違っていました。今回の調査で得られた情報を今後の教科書選定の際の判断基準の参考にできればと考えています。とはいいいましても、高校の教科書は通例4年で新版が出版されます。今回作成した教科書コーパスの中身も再来年には「現行」ではなくなります。終わることのない作業になるかもしれませんが、今後も教科書コーパスの作業を続けていこうと考えています。また著作権の絡みで難しいことも多いのですが、教科書コーパスを 'pedagogic corpus' として、生徒たちの発見学習に利用できる工夫をしていきたいと考えております。

参考文献

赤瀬川史朗・中尾浩(2004)『コーパス言語学の技法 : 言語データの収集とコーパスの構築』夏目書房
 垣田直巳, 三浦省吾(編纂)(1987)『英語教科書使用語彙高校編』1987年版 溪水社

- 斉藤俊雄・中村純作・赤野一郎（編）（2005）『英語コーパス言語学 改訂新版』研究社
- 塩見知之（2002）『文部省検定済中学校・高等学校教科書に現れた英語の語彙』北星堂
- 杉浦千早（2002）「杉浦リスト（eigo1, Ver. 2）」<http://grad.nufs.ac.jp/sugiura/sugiuralist.html>
- 杉浦千早（2002）「杉浦リスト（eigo2, Ver. 1）」<http://grad.nufs.ac.jp/sugiura/sugiuralist.html>
- 鷹家秀史・須賀広（1998）『実践コーパス言語学 - 英語教師のインターネット活用』桐原ユニ.
- 投野由紀夫（2003）「コーパスを英語教育に活かす」『英語コーパス研究』10, 249-64.
- 中尾浩・赤瀬川史朗・宮川進悟（2002）『コーパス言語学の技法：テキスト処理入門』夏目書房
- 水谷静夫（1982）『数理言語学』培風館
- 山口和範・高橋淳一・竹内光悦（2004）『図解入門 よくわかる多変量解析の基本と仕組み—巨大データベースの分析手法入門』秀和システム.
- 山添孝夫（2005）「教育現場におけるコーパスの利用」『立命館言語文化研究』16（4）185-195.
- Aston, G. (1997) Enriching the learning environment: corpora in ELT. In Wichmann et al. (1997) *Teaching and language corpora* (pp. 51-64). London: Addison Wesley Longman.
- (2001) Learning with corpora: an overview. In G. Aston (Ed.), *Learning with corpora* (pp. 7-45). Houston, TX: Athelstan.
- Aston, G. (Ed.) (2001) *Learning with corpora*. Houston, TX: Athelstan.
- Bernardini, S. (1996) Corpora in the classroom: an overview and some reflections on future developments. In E. Tognini-Bonelli (General Ed.) & J.M. Sinclair (Vol. Ed.) *Studies in corpus linguistics: Vol. 12. How to use corpora in language teaching* (pp. 15-36). Amsterdam: John Benjamins.
- Carter, R. (1998) *Vocabulary: applied linguistic perspectives* (2nd ed.). London: Routledge.
- Manning, C. and Schütze, H. (2000) *Foundations of statistical natural language processing* (2nd printing with corrections). Cambridge, MA: MIT Press.
- McEnery, T. and Wilson, A. (1996) *Corpus linguistics*. Edinburgh: Edinburgh University Press.
- Mindt, D. (1996) English corpus linguistics and the foreign language teaching syllabus. In J. Thomas and M. Short(Eds.), *Using corpora for language research* (pp. 232-247). London: Longman.
- . (1997) Corpora and the teaching of English in Germany. In Wichmann et al. *Teaching and language corpora* (pp. 40-50). London: Addison Wesley Longman.
- Nakamura, J.(1993) Statistical methods and large corpora: a new tool for describing text types. In M. Baker, G. Francis, and E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 293-312). Amsterdam: John Benjamins.
- . (1994) Extended Hayashi's quantification method type III and its application in corpus linguistics. In *Journal of language and literature, Faculty of Integrated Arts and Sciences, The University of Tokushima*, 1, 141-192.
- Oakes, M.P.(1998) *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
- Reppen, R. and Simpson, R.(2002) Corpus linguistics. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 92-111). London: Arnold.
- Schmitt, N. (2000) *Vocabulary in language teaching*. Cambridge: Cambridge University Press.
- Schmitt, N. (Ed.) (2003). *An introduction to applied linguistics*. London: Arnold.
- Sinclair, J.M. (Ed.)(1996) *Studies in corpus linguistics: Vol. 12. How to use corpora in language teaching*. Amsterdam: John Benjamins.
- Thomas, J. and Short, M. (Eds.)(1996). *Using corpora for language research*. London: Longman.
- Trumble, C. and Jones, G. (1997) *Concordances in the classroom*. Houston, TX: Athelstan. (Original work published in 1990 by Longman)

- Tsui, A.B.M. (1996) What teachers have always wanted to know—and how corpora can help. In E. Tognini-Bonelli (General Ed.) & J.M. Sinclair (Vol. Ed.) *Studies in corpus linguistics: Vol. 12. How to use corpora in language teaching* (pp. 39-61). Amsterdam: John Benjamins.
- Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (Eds.)(1997) *Teaching and language corpora*. London: Addison Wesley Longman.
- Willis, D. (2003) *Rules, patterns and words: grammar and lexis in English language teaching*. Cambridge: Cambridge University Press.

Appendix 1. The Textbooks Included in the Corpus.

出版社	CODE	TITLE	判型	頁	教科書番号
東京書籍	AAB1	All Aboard! English I	B5	116	英語 001
東京書籍	AAB2	All Aboard! English II	B5	124	英語 001
東京書籍	PON1	Power On English I	B5変型	148	英語 002
東京書籍	PON2	Power On English II	B5変型	172	英語 002
東京書籍	PRM1	PROMINENCE English I	A5	164	英語 003
東京書籍	PRM2	PROMINENCE English II	A5	164	英語 003
開隆堂	SSH1	SUNSHINE ENGLISH Course I	B5変型	138	英語 004
開隆堂	SSH2	SUNSHINE ENGLISH Course II	B5変型	160	英語 005
開隆堂	NOW1	ENGLISH NOW I	B5	114	英語 005
開隆堂	NOW2	ENGLISH NOW II	B5	126	英語 004
三省堂	EXC1	EXCEED English Series I	B5変型	136	英語 006
三省堂	EXC2	EXCEED English Series II	B5変型	152	英語 006
三省堂	CRW1	CROWN English Series I	A5	172	英語 007
三省堂	CRW2	CROWN English Series II	A5	192	英語 007
三省堂	VST1	VISTA English Series I	B5	112	英語 008
三省堂	VST21	VISTA English Series II Step One	B5	64	英語 008
三省堂	VST22	VISTA English Series II Step Two	B5	64	英語 009
教育出版	LIN1	Lingua-Land English Course I	B5	108	英語 009
教育出版	LIN2	Lingua-Land English Course II	B5	118	英語 010
教育出版	OWR1	ONE WORLD English Course I	A5	130	英語 010
教育出版	OWR2	ONE WORLD English Course II	A5	162	英語 011
開拓社	LGN1	NEW LEGEND ENGLISH I	A5	156	英語 011
開拓社	LGN2	NEW LEGEND ENGLISH II	A5	184	英語 012
大修館	CPT1	Captain English Course I	B5	112	英語 012
大修館	CPT2	Captain English Course II	B5	112	英語 013
大修館	GEN1	Genius English Course I	A5	160	英語 013
大修館	GEN2	Genius English Course II	A5	160	英語 014
啓林館	MIL1	MILESTONE English Course I	B5変型	140	英語 014
啓林館	MIL2	MILESTONE English Course II	B5変型	152	英語 017
啓林館	ACO1	ACORN English Course I	B5	104	英語 015
啓林館	ACO2	ACORN English Course II	B5	104	英語 016
啓林館	TMR1	Tomorrow English Course I	B5変型	128	英語 016
啓林館	TMR2	Tomorrow English Course II	B5変型	140	英語 015
数研出版	POL1	POLESTAR English Course I	B5変型	148	英語 017
数研出版	POL2	POLESTAR English Course II	B5変型	158	英語 018
文英堂	PWW1	POWWOW ENGLISH COURSE I	B5変型	144	英語 018
文英堂	PWW2	POWWOW ENGLISH COURSE II	B5変型	160	英語 019
文英堂	SRF1	Surfing English Course I	B5	112	英語 019
文英堂	SRF2	Surfing English Course II	B5	136	英語 020
文英堂	UNI1	UNICORN ENGLISH COURSE I	A5	160	英語 020
文英堂	UNI2	UNICORN ENGLISH COURSE II	A5	192	英語 021
池田書店	DLY1	DAILY ENGLISH COURSE I	B5変型	132	英語 021
池田書店	DLY2	DAILY ENGLISH COURSE II	B5変型	132	英語 022
池田書店	STG1	NEW STAGE English Course I	A5	148	英語 022
池田書店	STG2	NEW STAGE English Course II	A5	148	英語 023

出版社	CODE	TITLE	判型	頁	教科書番号
旺文社	STP1	Step English I	B5変型	92	英語 023
旺文社	STP2	Step English II	B5変型	108	英語 024
旺文社	PBL1	Planet Blue English I	A5	124	英語 024
旺文社	PBL2	Planet Blue English II	A5	138	英語 025
増進堂	NST1	NEW STREAM I	B5変型	144	英語 025
増進堂	NST2	NEW STREAM II	B5変型	152	英語 027
増進堂	MST1	MAIN STREAM I	A5	176	英語 026
増進堂	MST2	MAIN STREAM II	A5	192	英語 026
第一学習社	VOY1	Voyager English Course I	A5	152	英語 027
第一学習社	VOY2	Voyager English Course II	A5	152	英語 028
第一学習社	VVD1	Vivid English Course I	B5変型	136	英語 028
第一学習社	VVD2	Vivid English Course II	B5変型	144	英語 029
第一学習社	VVA1	Viva English! I	B5	112	英語 029
第一学習社	VVA21	Viva English! II (Unit 1)	B5	64	英語 030
第一学習社	VVA22	Viva English! II (Unit 2)	B5	64	英語 031
三友社	CSM1	NEW COSMOS ENGLISH COURSE I Revised Edition	B5変型	128	英語 030
三友社	CSM2	NEW COSMOS ENGLISH COURSE II Revised Edition	B5変型	160	英語 033
三友社	NWR1	NEW WORLD ENGLISH COURSE I Revised Edition	B5	112	英語 031
三友社	NWR2	NEW WORLD ENGLISH COURSE II Revised Edition	B5	128	英語 032
桐原書店	WTR1	WORLD TREK ENGLISH COURSE I	B5変型	142	英語 032
桐原書店	WTR2	WORLD TREK ENGLISH COURSE II	B5変型	158	英語 034
桐原書店	PAL1	NEW English PAL I	B5	116	英語 033
桐原書店	PAL2	NEW English PAL II	B5	124	英語 035
桐原書店	PRV1	PRO-VISION ENGLISH COURSE I	B5変型	156	英語 034
桐原書店	PRV2	PRO-VISION ENGLISH COURSE II	B5変型	166	英語 036
秀文館	ENC1	ENCOUNTER English Series I	A5	144	英語 035
秀文館	ENC2	ENCOUNTER English Series II	A5	172	英語 037

Appendix 2. Descriptions of Analysis Tools

TOOL	INPUT(S)	FUNCTION(S)	OUTPUT(S)
FreqCount	TBK.txt	準備されたテキストから単語を切り出し、頻度数をカウント。 他のプログラムで使用する基本的なデータも出力。	*_Sent.txt *_Word.txt *_Afreq.txt *_Rfreq.txt
Merge	*_Afreq.txt Merged_Afreq.txt	アルファベット順の頻度リストをいくつか併合して大きな頻度リストにする。	Merged_Afreq.txt
CrossTable	Merged_Afreq.txt *_Afreq.txt	Merged_Afreq.txtをもとに幾つかの*_Afreq.txtを参照しながら分割表を作成 第1段階：Mergeで生成されたMerged_Afreq.txtを語彙総目録とする 第2段階：各教科書の*_Afreq.txtを教科書ごとの語彙頻度情報として全教科書の語彙頻度表を生成	Total_Table.txt
UniqueWordExtraction	Total_Table.txt	Total_Table.txtから各テキストに固有の単語群を抽出、削除し、代わりにテキストに固有な単語群という項目を立て、その頻度の合計を表の頭に挿入。	Total_Table_Collapsed.txt Unique.txt
QTC	Total_Table.txt	Total_Table_Collapsed.txtを読み込んで、固有方程式の係数行列Aを計算し、固有方程式を解き、得られた固有ベクトルからカテゴリーとサンプルに与えられる数量を計算した上、その各々を平均値0、分散1に正規化したものを結果として返す。 固有値が得られた段階で、各軸の固有値、理論的な相関係数、寄与率、累積寄与率も出力。	Category_Weight.txt Sample_Score.txt QTC_Summary_Excel.txt
Ratio	Total_Table_Collapsed.txt	Total_Table_Collapsed.txtを読み、総合計から各Categoryと各Sampleの割合を計算する。結果は、CategoryRatio.txt、SampleRatio.txtに保存される。 この結果は数量化の結果を図示する際に、CategoryあるいはSampleの相対頻度を表す円、あるいは球体の大きさを決めるのに用いられる。	CategoryRatio.txt SampleRatio.txt
PLT2 PLT3	Category_Weight.txt CategoryRatio.txt Sample_Scores.txt SampleRatio.txt	数量化の結果を2次元あるいは3次元空間に図示する。PLT2は2次元表示、PLT3は3次元表示のプログラム。	Graphics
SentenceLengthCalc	*_Sent.txt	*_Sent.txtを参照しながら文の平均長に関する基礎データの計算と統計値を出力。	SentenceLength.txt Word_Sentence.txt
SentenceLengthGraphic	SentenceLength.txt	Sent.txtを参照しながら文の平均長の推移を図示。	Graphics
TTRCalc	*_Word.txt	*_Word.txtを参照しながら各Token番号に見合ったType番号と、その時点でのTTRの計算を行い結果を保存。	TTR.txt
TTRDecreaseGraphic	TTR.txt	TTR.txtを参照しながらToken数の増加に従って、TTRがどのように変化するかを図示。	Graphics
TypeIncreraseGraphic	TTR.txt	Token数の増加につれてType数がどのように変化するかを観点から図示。	Graphics

