英文リーダビリティ測定システムの開発

福井正康，小篠敏明

福山平成大学経営学部経営情報学科

概要

英文リーダビリティの測定にはこれまで様々な指標が用いられて来たが、多くは単語数と音節数によるものであった。我々はこれらに単語と熟語の難易度を加え、新たな指標を開発したが、これには単語と熟語の難易度を与える辞書とこの辞書を利用して文を解析するプログラムが不可欠であった。我々はこのプログラムの主な３つの機能、リーダビリティの測定、新しい指標作成のためのデータ出力、新しい辞書の作成方法について説明する。

キーワード

英文，リーダビリティ，プログラム

１．はじめに

英文リーダビリティの測定については、これまで様々な指標が与えられて来たが、ほとんどは単語数と音節数によるものであり、文の構造や単語の難しさなどを考慮したものはあまり見られなかった。最近、熟語の難易度を取り入れた指標の開発も進められているが [1)-3)]、まだ具体的なシステムとしては完成していない。さらにこれらの指標は英語を母国語とする人のための指標であり、英語教育を受ける我々にとっては実感と異なる場合も少なくないと思われる。

この状況をふまえ、我々は英語学習者のための英文リーダビリティ指標の開発を始めた。手始めとして単語と熟語の難易度を加えた新たな指標作りを目指し、単語辞書と熟語辞書の作成、及び英文の難易度を測定するプログラムの製作を開始した。単語辞書には出発点として JACET 8000 を使い、単語の難易度はその中にある単語のランクをそのまま利用することにした。また単語の変化形については、JACET 8000 の各単語にその変化形を加え、原形を見出し語とした。単語の音節数については変化形ごとに異なるため、プログラムにより計算することにし、確認のため単語辞書に見出し語の音節数を加えた。

熟語辞書については、雛形とするものが見つけにくいため、参考文献 4) に現れる 1000 の熟語を利用することにした。これは重要度別に 3 段階に分かれているため、最頻出を難易度 10、最重要を難易度 20、重要を 30 と難易度を 3 段階に分類した。熟語の当てはめについては、例えば「look at」の場合、間に副詞などを含む可能性を考えて、「look * at」として、「*」に複数の任意の単語を適合させるワイルドカードの役割を持たせた。このワイルドカードにはデフォルトで 3 単語の上限を設け、プログラム中で変更できるようにした。また、「*ing」のように 1 つの単語の部分的なワイルドカードも使えるようにした。

難易度の判定プログラムはまず、文章を段落に分け、段落を文に分けるところから始まる。各種の出力データはこれらの段落か文ごとに表示される。これらが完了すると、1 文の単語数と各単語の音節数を計算する。音節数を計算するプログラムはよく利用される単語と判定の難しそうな単語を 900 程度選んで正当性を評価している。その後、文単位で各単語を原型に変形する準備を始める。これには大文字と小文字の問題、「It's」や「I'd」などの短縮形の問題、単語と「,」や「.」などが結合している問題、また「.」は「Mr.」や「Prof.」などの省略形の一部になっている問題、強調の「' '」や会話部分の「" "」の問題など多くの問題があるが、詳細は後に述べる。

これらの処理の後、単語辞書を用いて各単語を原形に変形し、同時に難易度の割付けも行う。原形への変換が終わるとそれを利用して熟語辞書の検索を行う。熟語辞書は原形で書かれた部分が多いが、比較級や進行形なども含まれているため、この検索は原形に変形した単語と元の

変化形の単語を２重に調べる。熟語辞書は件数が 1000 件程度であるのであまり高速化を気にかけていないが、単語の検索には今後辞書が拡大する可能性も考えて独特の方法を用いた。これによって中学１年から高校２年までの一連の教科書を 10 秒程度で解析できるようになった。

この解析から新しい評価式を作り出すことは難しくない。各文の単語数と音節数はすでに調べてあり、それに各単語に当てられた難易度の１文当たりの合計と各文の熟語難易度の合計を加えてデータとして出力できる。もし、上に述べた中学１年から高校２年までの教科書を用いて、各文の出現学年をデータとして利用できるならば、それを目的変数とし、単語数、音節数、単語難易度、熟語難易度を説明変数とする重回帰分析が実行でき、重回帰式はこの教科書にそった難易度の評価指標となる。

さらに上のように文中に学年データを埋め込んでおけば、検索過程を逆に利用して単語の出現学年が分かることになる。これを繰返し、より低学年で現れた場合のみ書き直しをすれば単語の初出現学年も分かる。またこの考え方は熟語にも利用できる。英語学習者にとっては単語や熟語の難易度をどの時期に習ったものかで判断する場合が多いと思われるので、この初出現学年を難易度にすることは意味のあることと考えられる。そこで我々はプログラムにこの難易度を持った新しい辞書を出力する機能を加えた。これらの処理は教科書１種類では偏りがあるため、中学１年から高校２年まで通して出版している出版社３社を選んで、それらの初出現学年の平均値を新しい難易度とする辞書を作った。指標作りのためのデータはこれらの学年のデータを埋め込んだ３社の全教科書とした。指標の詳細については参考文献 5) に詳しい。

以後、プログラムの流れにそって処理やそれらのアルゴリズムを説明する。このプログラムの開発には日本学術振興会科学研究費補助金 基盤研究Ｃ（19520535）の援助を受けており、謝意を表したい。

２．辞書ファイルの構造

このプログラムでは単語辞書と熟語辞書を利用するが、最初にこれらの構造について説明する。単語辞書は JACET 8000 をベースに作成されており、以下のような構造になっている。

＜単語難易度＞,＜見出し語音節数＞,＜見出し語（原形）＞,＜変化形１＞,＜変化形２＞, ･･･

単語難易度については、始めは JACET 8000 のランク（1〜8000）を利用するが、後に教科書中の初出現学年を利用することも考える。見出し語音節数については、辞書から求めた見出し語の音節数が記入してある。しかし音節数はプログラムで計算するので、ここで入力された音節数はあくまで確認用であるため、現在は全体で 900 程度の入力になっている。それ以外の音節数については、-1 が記入されている。実際の辞書ファイルの内容を表 2.1 に示す。

表 2.1 単語辞書ファイルの例

0,-1,¥s	31,1,my
15,-1,¥d	32,1,all
1,1,the	33,1,will,would,wills,will's,wills'
2,1,and	34,1,by
:	35,1,me
15,1,have,has,had,having	36,1,can,could,couldn't,cans,can's,cans'
16,1,on	37,1,or
17,1,they	38,2,about
18,1,we	39,1,when
19,1,with	40,1,go,goes,went,gone,going
20,1,do,does,did,done,doing	41,1,if,ifs
21,1,but	42,1,would
22,1,as	43,1,an
23,1,she	:
24,1,his	3496,-1,resort,resorts,resort's,resorts'
25,1,this	3497,-1,confront,confronts,confronted,confronting
26,1,at	3498,-1,chancellor,chancellors,chancellor's,chancellors'
27,1,from	3499,-1,substitute,substitutes,substitute's,substitutes',・・・
28,1,her	3500,3,particle,particles,particle's,particles'
29,1,what	3501,-1,incentive,incentives,incentive's,incentives'
30,1,there	:

先頭から２つの「¥s」と「¥d」については、文中に現れる「's」や「'd」のような省略形を表す文の変形の過程で現れる記号で、JACET 8000 の見出し語以外のものである。

熟語辞書については参考文献 4) をベースにしており以下の形式である。

＜熟語難易度＞，＜熟語構文＞

熟語辞書については、最初の辞書では 10, 20, 30 の３段階である。後にこれにも教科書の初出現学年を利用することを考える。実際の熟語辞書ファイルの内容の一部を表 2.2 に示す。熟語構文中の「*」はワイルドカードで、0 から指定した数までの任意の個数の単語と一致する。

表 2.2 熟語辞書ファイルの例

10,look * at	10,listen * to
10,think * of	10,give * up
10,go * on	10,find * out
10,be * interested in	10,bring * up
10,depend * on	10,grow * up
10,depend * upon	10,used * to
10,come * to	10,look * for
10,come * from	:

３．データの前処理

最初にプログラムを起動し、解析するデータを読み込んだ初期画面を図 3.1 に示す。

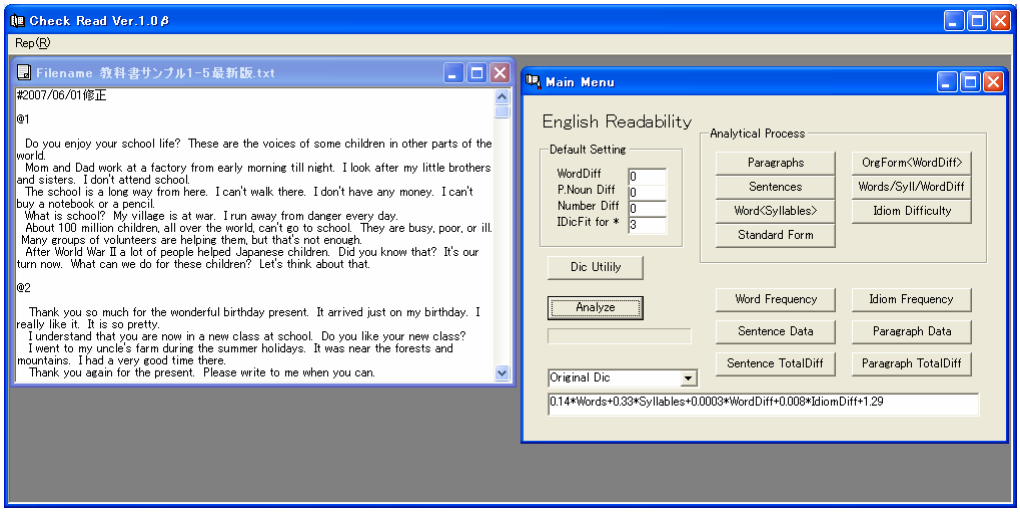


図 3.1 データ入力後の初期画面

左のウィンドウがサンプルデータで、右がメインメニューである。サンプルデータの先頭にある「#2007/06/01 修正」の部分はコメント行である。各行の先頭に「#」の付いた文字列はコメントと判定する。また「@1」や「@2」は、それに続く文章の学年を表す。学年は中学１年から高校２年まで１から５の数字で示し、「@1」，「@2」はそれぞれ中学１年と中学２年を表す。この学年指定によって指標作成のためのデータ出力、学年を難易度とした新しい単語辞書や熟語辞書の作成が可能となる。

解析を始めるにはまず右側のメインメニューの「Analyze」ボタンをクリックする。これにより文の解析が始まる。この解析は大きく分けて、①段落の分割、②文の分割、③音節数の計算、④文の変換、⑤原形変換と難易度設定、⑥単語難易度の計算、⑦熟語難易度の計算、⑧式による難易度の計算の過程よりなるが、これら各段階の結果は解析終了後、「Analytical Process」グループボックス内の各ボタンやその下のボタンによって表示可能である。説明は分かり易くするため解析過程と平行してデータの表示について行うが、その部分についてはすべての解析が終わった後に表示可能となる。この節では①から④までを説明する。

① 段落の分割

段落の分割には以下の処理が含まれる。但し、以後「□」は全角の空白を表し、「_」は半角の空白を表す。

(1) 特殊文字の全角は半角にする。

「□」,「“」,「”」,「‘」,「’」は日本語ワープロで英語を入力した場合に全角として残

る可能性がある文字である。これをまず半角に変換する。

(2) 段落末の「¥」＋改行を取り除く。

段落末の「¥」は元々１行のものを便宜的に２行に分けたものと解釈するが、これを元に戻す処理である。

(3) タブは空白にする。

(4) 空白３つ以上は空白２つにする。

(5) colon の後ろの空白を１つにする。

(6) 最終行の最後に改行を付ける。

これは改行のところで分割し易くするためである。

(7) 段落の先頭が、「#」「@」「%」なら特殊処理をする。

「#」はコメント行で段落に含めず、段落表示からも消える。

「@」は難易度（学年）を表す行で段落番号に含めないが、段落表示では残る。

「%」は単語登録を表す行で段落に含めるが、文章の解析用データには表示されない。

実際の変換は例えば、以下のようになるが、変換後の #1, #2, #3 は段落番号である。先頭が「%」の段落は単語の登録のみに使われ、解析用のデータとしては表示されないので、段落の難易度が表示される際は #2 の段落からである。

変換前

```
@1
% game, apple, bag, city, car, desk, evening, egg, family
# Program 1
Are you a junior high school student?
Yes, I am.
```

変換後

```
@1
#1
% game, apple, bag, city, car, desk, evening, egg, family
#2
Are you a junior high school student?
#3
Yes, I am.
```

(8) 改行部分で段落に分割する。その際ヌル文字列の段落（改行だけの段落）は除く。

(9) 段落の先頭と最後の空白は除く。

(10) 表示用に段落に分けた形を保存する。

メインメニューの「Paragraphs」ボタンをクリックすると段落に分けた図 3.2 が表示される。

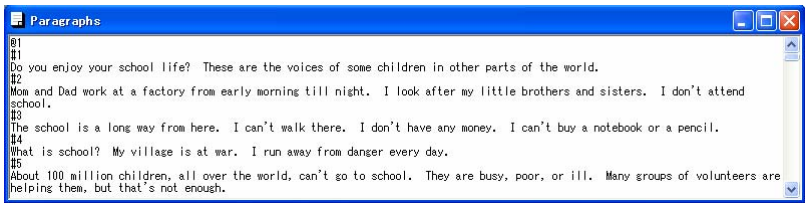


図 3.2 段落分割画面

この出力結果では文中に段落番号がコメント行として挿入してある。これを新たにデータとすることもできるが、これは文の位置を分かり易くするための方法である。次にこの段落を文に分ける過程に入る。

② 文の分割

文の分割には以下の処理が含まれる。

(1) 空白２つで文の区切りとみなし、文を分割する。

(2) 表示用に文に分割した形を保存する。

メインメニューの「Sentences」ボタンをクリックすると図 3.3 の文の分割画面が表示される。

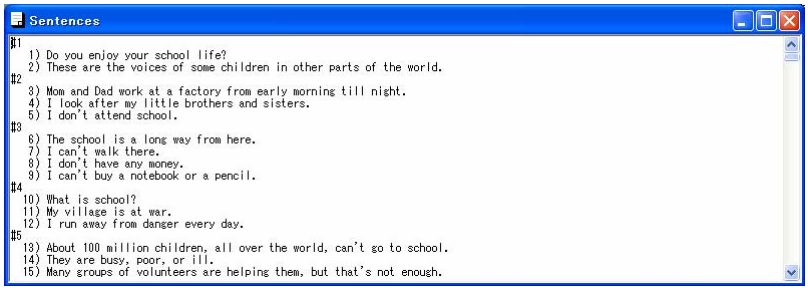


図 3.3 文の分割画面

③ 音節数の計算

音節数はプログラムによって計算する。辞書を用いると単語の変化形ごとに音節数を割り当てなければならず、また当然辞書にない単語については求められないからである。しかし計算が正しいかどうか調べる必要もあり、現在 916 個の見出し語について、音節数を単語辞書の中に書き込んである。この 916 個は、使用頻度の高い最初の 200 個と 200 個目から 10 個飛ばしに選んだ 660 個、推定が難しいと思われる 50 個程度の見出し語を選んだ。これらの見出し語の音節数の正当性の確認には後に述べる辞書ユーティリティを利用する。この過程では音節数を計算し保存するだけでなく、１文当たりの単語数についても保存する。

メインメニューの「Word<Syllables>」ボタンを押すと、以下の単語別の音節数を表す図 3.4 が表示される。

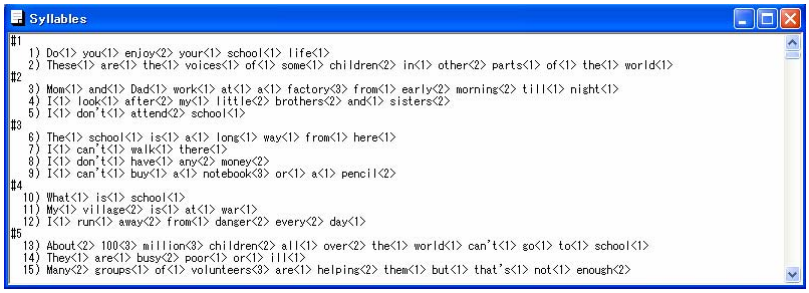


図 3.4 音節数表示画面

④ 文の変形

次に単語辞書と熟語辞書を検索するための最終的な変形を行う。この変形には以下の処理が含まれる。ここで「_」は半角の空白を意味する。

(1) 数字以外の「+」,「–」を分離する。

ここでは以下の例のように、数字でない単語に付いた「+」や「−」を分離したり、空白に変えたりする。例えば以下である。

「-abc」→「-_abc」,「rh-」→「rh -」,「abc-def」→「abc_def」

数字や電話番号などの「-123」や「084-972」などはそのまま残しておく。

(2) 「1st」,「2nd」,「3rd」の形を変換する。

ここでは上記の表記を通常の英語表記にする。これ以上の数字について、例えば「21th」などは数字と「th」に分ける。

(2) 括弧を分離する。

ここでは文中の以下の括弧について分離を行う。分離された括弧の記号は単語としても難易度としても評価しない。

「(」→「(_」,「)」→「_)」,「[」→「[_」,「]」→「_]」,「{」→「{_」,「}」→「_}」,「<」→「<_」,「>」→「_>」

(3) 省略形を分離する。

ここでは以下のような省略形の処理を行う。省略形が分かる場合はそのまま元の形に戻し、［Bob's］などのように文脈でしか分からない場合は、「¥s」や「¥d」として分離する。

「Let's」→「Let us」,「cannot」→「can not」,「can't」→「can not」,「n't」→「_not」,「'll」→「_will」,「'm」→「_am」,「're」→「_are」,「've」→「_have」,「's」→「_¥s」,

「'd」→「_¥d」

(4) 「"」または「'」を消去する。

ここでは文中の引用や強調などの「"」や「'」の消去を行う。

(5) 「,」 などを分離する。

「,_」→「_,_」,「:」→「_:」,「;」→「_;」 「/」→「_/_」

(6) 文末の「.」などを分離する。

「!」→「_!」,「?」→「_?」,「.」→「_.」,「¥.」→「_.」

Double quotation 中の２つ以上の文は、２つの空白を空けると文を分けることになるが、文の間が「¥._」の場合は一続きの文と解釈される。

(7) 数字を「¥n」に変換し、単位付き数字を分離する。

「5」→「¥n」, 「30cm」→「¥n_cm」

(8) もう一度２つ以上の空白を１つにする。

この文の変形を行った後の形式は、メインメニューの「Standard Form」ボタンで見ることができる。その例を図 3.5 に示す。

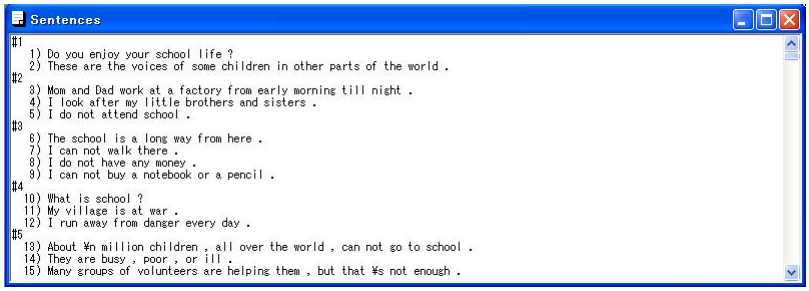


図 3.5 文の変形画面

４．原形変換・単語難易度の計算

辞書検索のスピードはこのソフトの使い勝手の重要な要素である。検索スピードの向上には２つのアイデアが含まれている。一つは辞書を読み込む際にアルファベット順に並べ替えて読み込み、検索される単語の先頭文字と同じアルファベットのところだけを読み込むことである。しかし、単語の中には、「be」や「eat」などのように変化形の先頭文字が原形と異なるものもあるので、このような単語だけを集めて副単語辞書を作り、検索は最初にこの副単語辞書を調べ、後に単語辞書の先頭文字を対象として検索する。この検索では副単語辞書の単語数が限られているので時間的な損失は殆どない。副単語辞書の作成は単語辞書を読み込む際に行われる。

この方式によって使用頻度順に単語が並んだ場合に比べても数倍スピードが向上する。

単語は一度すべて小文字に直して原形から変化形へと検索を行う。どこかで検索対象と一致した場合はその単語の原形と辞書に含まれる難易度を保存する。単語辞書にない場合は、小文字に変換する前の単語を見て、先頭が大文字の場合は固有名詞のデフォルト設定、それ以外は一般の難易度のデフォルト設定を行う。このデフォルト値はそれぞれメインメニューの「PNoun Diff」、「WordDiff」のテキストボックスに書かれた値であり、初期値はそれぞれ 0 になっている。これ以外にも数字のデフォルト設定も指定できるが、数字は文の変形のところで「¥n」に変換されており、容易に判定できる。また記号類は難易度 0 となっている。

この原形変換を行った後の文は、メインメニューの「OrgForm<WordDiff>」ボタンで表示できる。その例を図 4.1 に示す。

図 4.1 原形変換と難易度

難易度設定を行った後、前にデータとして格納していた単語数と音節数も合わせ、これらの各文当たりの合計を計算する。結果はメインメニューの「Words/Syll/WordDiff」ボタンをクリックすると表示される。その例を図 4.2 に示す。

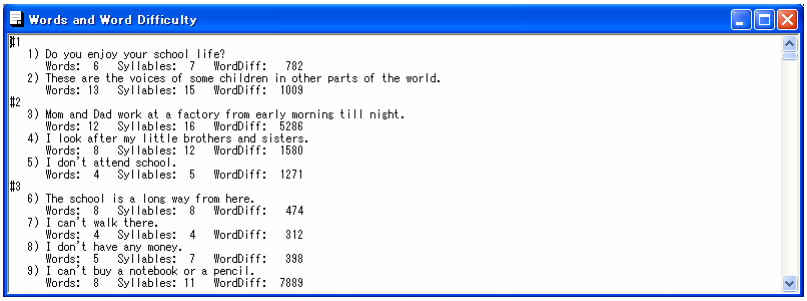


図 4.2 文別単語数・音節数・単語難易度表示画面

５．熟語難易度の計算

熟語難易度の設定には、原形に変換する前の文と原形に変換した後の文２つを利用する。まず原形変換後の単語を対象に熟語辞書の検索を行い、その後原形変換前の単語を用いた検索を行う。すべてこの順番で行うので、以後２つを区別しないで説明する。原形と元の変化形を両方用いる理由は、熟語辞書に「better」や「～ing」のような変化形が含まれるからである。

対象の単語を熟語辞書の先頭単語（見出し単語）から検索した後、次の単語と熟語辞書の次の単語が一致するか調べる。一致する場合は次の単語に進むが、一致しない場合はそこで検索を中止する。但し熟語辞書の単語がワイルドカード（このシステムでは「*」を指定）の場合は、次の単語どうしを比較する。これで一致しない場合は、文の次の単語を比較する。この処理を指定の回数だけ繰り返す。この途中で一致する単語が見つかった場合は、またもとの比較に戻って処理を続ける。見つからない場合は検索を中止する。これらの処理を熟語の終わりまで行うことができればそこに熟語が含まれると解釈する。熟語が確認されると、その文に熟語難易度が加えられる。１文に複数の熟語が含まれるとその合計が計算される。熟語難易度の設定状況はメインメニューの「Idiom Difficulty」ボタンによって見ることができる。その例を図 5.1 に示す。

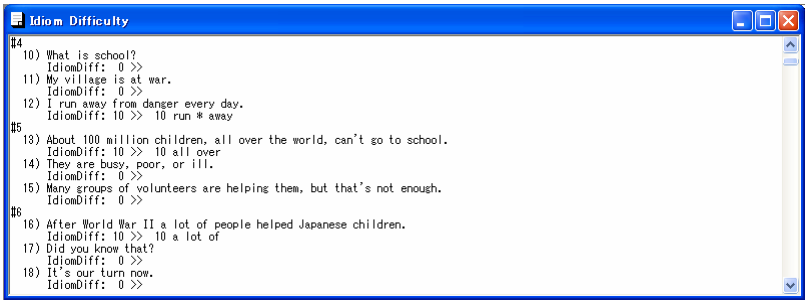
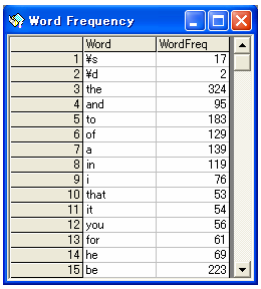


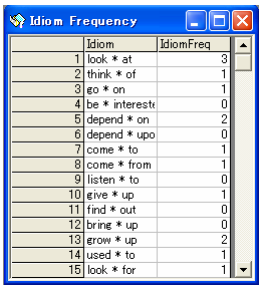
図 5.1 熟語難易度設定状況確認画面

これまでの過程で単語辞書と熟語辞書を検索したが、その結果それらの単語及び熟語の出現件数も数えることができる。この結果はそれぞれメインメニューの「Word Frequency」ボタンと「Idiom Frequency」ボタンで見ることができる。その例を図 5.2a と図 5.2b に示す。



	Word	WordFreq
1	¥s	17
2	¥d	2
3	the	324
4	and	95
5	to	183
6	of	129
7	a	139
8	in	119
9	i	76
10	that	53
11	it	54
12	you	56
13	for	61
14	he	69
15	be	223

図 5.2a 単語出現頻度



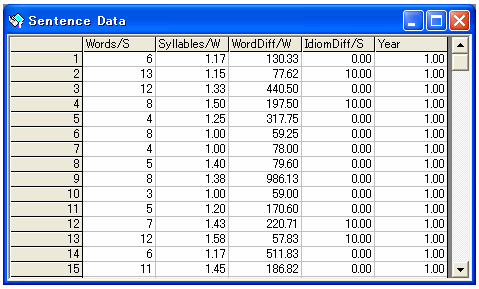
	Idiom	IdiomFreq
1	look * at	3
2	think * of	1
3	go * on	1
4	be * interest	0
5	depend * on	2
6	depend * upo	0
7	come * to	1
8	come * from	1
9	listen * to	0
10	give * up	1
11	find * out	0
12	bring * up	0
13	grow * up	2
14	used * to	1
15	look * for	1

図 5.2b 熟語出現頻度

６．評価式による難易度の計算

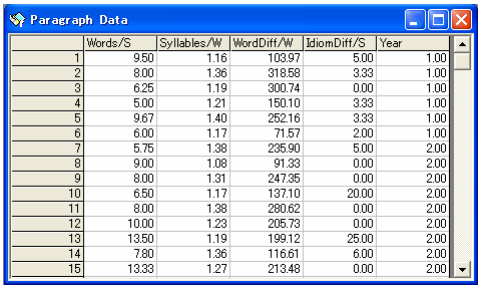
熟語難易度を設定した後、１文当たりの単語数、音節数、単語難易度、熟語難易度が計算され保存される。

メインメニューの「Sentence Data」ボタンをクリックすると、この過程で保存されたデータから、単語当たりの音節数と単語難易度を計算し、１文当たりの単語数と熟語数と合わせて表示する。その例を図 6.1a に示す。メインメニューの「Paragraph Data」ボタンをクリックすると、これらの値を段落ごとに平均し表示する。その例を図 6.1b に示す。



	Words/S	Syllables/W	WordDiff/W	IdiomDiff/S	Year
1	6	1.17	130.33	0.00	1.00
2	13	1.15	77.62	10.00	1.00
3	12	1.33	440.50	0.00	1.00
4	8	1.50	197.50	10.00	1.00
5	4	1.25	317.75	0.00	1.00
6	8	1.00	59.25	0.00	1.00
7	4	1.00	78.00	0.00	1.00
8	5	1.40	79.60	0.00	1.00
9	8	1.38	986.13	0.00	1.00
10	3	1.00	59.00	0.00	1.00
11	5	1.20	170.60	0.00	1.00
12	7	1.43	220.71	10.00	1.00
13	12	1.58	57.83	10.00	1.00
14	6	1.17	511.83	0.00	1.00
15	11	1.45	186.82	0.00	1.00

図 6.1a 文当たりの難易度項目



	Words/S	Syllables/W	WordDiff/W	IdiomDiff/S	Year
1	9.50	1.16	103.97	5.00	1.00
2	8.00	1.36	318.58	3.33	1.00
3	6.25	1.19	300.74	0.00	1.00
4	5.00	1.21	150.10	3.33	1.00
5	9.67	1.40	252.16	3.33	1.00
6	6.00	1.17	71.57	2.00	1.00
7	5.75	1.38	235.90	5.00	2.00
8	9.00	1.08	91.33	0.00	2.00
9	8.00	1.31	247.35	0.00	2.00
10	6.50	1.17	137.10	20.00	2.00
11	8.00	1.38	280.62	0.00	2.00
12	10.00	1.23	205.73	0.00	2.00
13	13.50	1.19	199.12	25.00	2.00
14	7.80	1.36	116.61	6.00	2.00
15	13.33	1.27	213.48	0.00	2.00

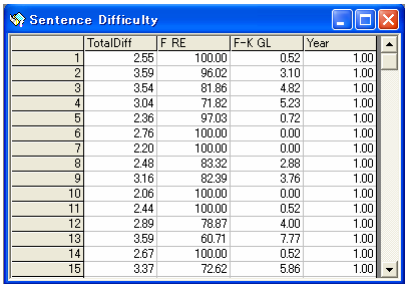
図 6.1b 段落当たりの難易度項目

次の過程ではメインメニュー下のテキストボックスに書かれた数式を使って、各文の難易度を計算する。この例で使われたのは以下の式である。

難易度=0.14*Words+0.33*Syllables+0.0003*WordDiff+0.008*IdiomDiff+1.29　　(1)

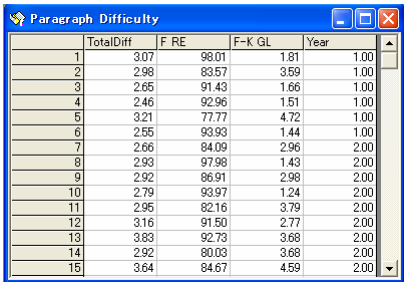
これらの変数には図 6.1a の各レコードデータが代入される。すなわち Words には Words/S、Syllables には Syllables/W、WordDiff には WordDiff/W、IdiomDiff には IdiomDiff/S の値である。我々が定義した難易度の他に、１文当たりの単語数と１単語当たりの音節数を用いて各文の Fresh Reading Ease の値と Fresh-Kincade Grade Level の値も計算しておく。これらの各文の値を計算した後、段落ごとにこれらの平均を計算する。以上で解析過程は終了する。

１文ごとのこれらの値はメインメニューの「Sentence TotalDiff」ボタンで、段落当たりのこれらの値は同じく「Paragraph TotalDiff」ボタンで表示される。その例をそれぞれ図 6.2a と図 6.2b に示す。



	TotalDiff	F RE	F-K GL	Year
1	2.55	100.00	0.52	1.00
2	3.59	96.02	3.10	1.00
3	3.54	81.86	4.82	1.00
4	3.04	71.82	5.23	1.00
5	2.36	97.03	0.72	1.00
6	2.76	100.00	0.00	1.00
7	2.20	100.00	0.00	1.00
8	2.48	83.32	2.88	1.00
9	3.16	82.39	3.76	1.00
10	2.06	100.00	0.00	1.00
11	2.44	100.00	0.52	1.00
12	2.89	78.87	4.00	1.00
13	3.59	60.71	7.77	1.00
14	2.67	100.00	0.52	1.00
15	3.37	72.62	5.86	1.00

図 6.2a 文当たりの難易度



	TotalDiff	F RE	F-K GL	Year
1	3.07	98.01	1.81	1.00
2	2.98	83.57	3.59	1.00
3	2.65	91.43	1.66	1.00
4	2.46	92.96	1.51	1.00
5	3.21	77.77	4.72	1.00
6	2.55	93.93	1.44	1.00
7	2.66	84.09	2.96	2.00
8	2.93	97.98	1.43	2.00
9	2.92	86.91	2.98	2.00
10	2.79	93.97	1.24	2.00
11	2.95	82.16	3.79	2.00
12	3.16	91.50	2.77	2.00
13	3.83	92.73	3.68	2.00
14	2.92	80.03	3.68	2.00
15	3.64	84.67	4.59	2.00

図 6.2b 段落当たりの難易度

7．評価式の決定

式 (1) の決定には図 6.1b のデータが重要である。ここでは、１文当たりの単語数、単語当たりの音節数、単語当たりの難易度、１文当たりの熟語難易度の値の段落ごとの平均が示されている他、その段落が現れた学年が Year の変数名で示されている。これは英文を段落単位に分割する際、各段落に設定されたデータである。我々の目的は学年を他の変数で予想するのであるから、この学年を目的変数に、残りの４つの変数を説明変数にして重回帰分析を実行すればよい。まず我々はある教科書シリーズの中から文をいくつか抜き出し、抜き出した学年を文の中に「@2」などの形で加え、ここで用いた学年を含むデータを作成した。これにより、図 6.1b のデータが作成され、このデータを用いて重回帰分析を実行した。計算に使用する段落当たりのデータは同じ学年でも大きくゆらぐため、寄与率は 0.35 程度とあまり高くなかったが、一応の予測式が完成した。社会システム分析ソフト College Analysis [6] を用いた計算結果を図 7.1 に示す。

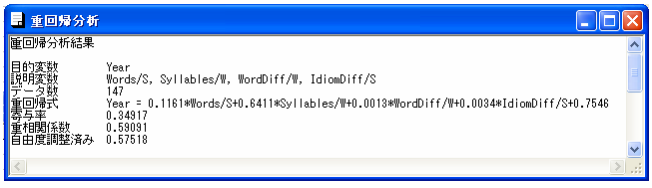


図 7.1 重回帰分析結果

しかし、我々が選んだ文を使ったこの予測式は一般性を欠くため、中学１年から高校２年ま

での教科書を販売している出版社３社を選び、すべての教科書について学年を入力したデータを作成した。これは日本の英語教育の現状を示しているものと考え、このデータを用いて重回帰分析を行った。この最終的な結果については、新しい辞書作成法を紹介した後に述べる。

８．新しい辞書の作成

我々は、英文の解析過程において、文中の単語に難易度を割り当てられるのなら、逆に単語に学年が割り当てられると考えた。これによって容易に単語の初出現学年を見つけることができる。そこでこの初出現学年を逆に単語の難易度に設定することを考えた。同様の設定は熟語についても適用できる。そこである会社の中学１年から高校２年までのテキストを用いて、単語や熟語の初出現学年を難易度とする新しい辞書の作成を試みる。

辞書に関する操作は辞書ユーティリティメニューに含まれるので、ここでは最初にこのメニューの機能を紹介する。メインメニューの「Dic Utility」ボタンをクリックすると図 8.1 の辞書ユーティリティメニューが表示される。

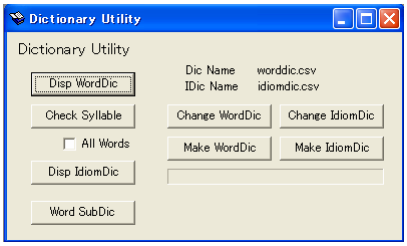


図 8.1 辞書ユーティリティメニュー

この中で、「Disp WordDic」ボタンをクリックすると現在使われている単語辞書の内容が表示され、「Disp IdiomDic」ボタンをクリックすると現在の熟語辞書の内容が表示される。その例を図 8.2a と図 8.2b に示す。

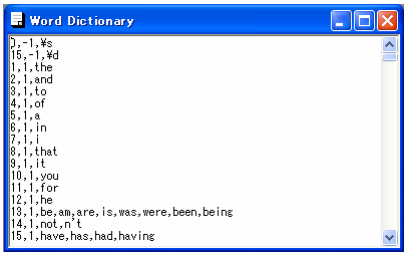


図 8.2a 単語辞書表示画面

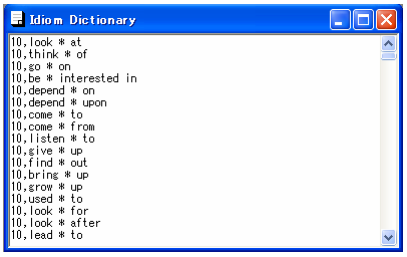


図 8.2b 熟語辞書表示画面

単語辞書の中で見出し語の先頭文字と変化形の先頭文字が異なる単語で、副単語辞書を作ることを前に述べたが、この辞書は辞書ユーティリティメニューの「Word SubDic」ボタンをクリックすると表示される。その例を図 8.3 に示す。

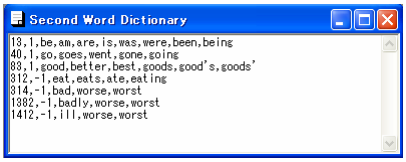
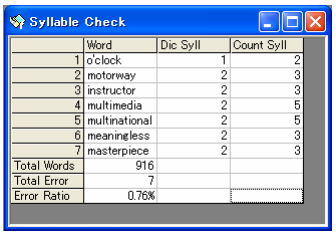


図 8.3 副単語辞書表示画面

辞書ユーティリティメニューの「Check Syllable」ボタンをクリックすると、辞書の中に書かれた見出し語の音節数と計算で求めた音節数の比較が行われ、結果が表示される。「All Words」チェックボックスですべての単語を表示するか、違っている単語だけにするか決めることができる。ここでは違っている単語のみ表示する。図 8.4 にその例を示す。結果は複合語について違いが見られるが、我々の計算がより現実的であると考え、変更は行っていない。



	Word	Dic Syll	Count Syll
1	o'clock	1	2
2	motorway	2	3
3	instructor	2	3
4	multimedia	2	5
5	multinational	2	5
6	meaningless	2	3
7	masterpiece	2	3
Total Words	916		
Total Error	7		
Error Ratio	0.76%		

図 8.4 音節数の比較

改めてこの章の始めに述べた新しい辞書の作成方法を説明する。ある会社の教科書の初出現学年を難易度とする新しい単語辞書を作るには、その会社の教科書で各学年の文章の先頭に「@2」のような学年を表す段落を入れたデータを作り、それを読み込む。

辞書ユーティリティメニューの「Make WordDic」ボタンで初出現学年を難易度とした新しい単語辞書を作ることができる。新しい熟語辞書についても「Make IdiomDic」ボタンで同様に作ることができる。その例を図 8.4a と図 8.4b に示す。これらはすぐに辞書に使える形をしているので、例えば拡張子を「.csv」として保存するだけでよい。

図 8.4a 新しい単語辞書の作成

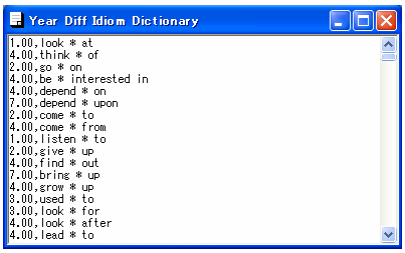


図 8.4b 新しい熟語辞書の作成

これらの辞書は１つの会社の中学１年から高校２年までの教科書を使ったものであるが、我々は３つの会社の教科書で作ったそれぞれの辞書ファイルについて、難易度の平均を取った。但し、ある教科書で出現しない単語がある場合は、出現する教科書の平均を取った。これによって新しい全テキストを平均化する辞書を作った。

この新しい辞書を使うためには、辞書ユーティリティの「Change WordDic」ボタンと「Change IdiomDic」ボタンをクリックし、辞書ファイルを選択する。

９．おわりに

我々はある教科書を元にした単語や熟語の難易度を、その教科書での初出現学年とすることを考えた。複数の教科書を元にする場合は各教科書での初出現学年の平均を用いる。これを使って３社の教科書を解析し、日本の教科書の平均値ともいうべき新しい評価式を作った。それは以下で与えられる。

難易度 = 0.10*Words+0.43*Syllables+0.98*WordDiff+0.063*IdiomDiff+0.28 (2)

この評価式の有効性については、参考文献 5) で詳しく検討される。

この式により３社の教科書を評価した場合、寄与率は 0.41 程度になる。文章は長かったり短かったり、難しかったり易しかったり、１つの学年内で相当揺らぎが大きいものと思われる。今後は、１つ１つの文に専門家によって直感的な学年を割付けてもらい、(1) 式や (2) 式で求めた結果と比較することにより、我々の評価式がより実感に近づくように考えたい。

さて、我々は３社の教科書を使った辞書と評価式を示したが、各社ごとの辞書と評価式も作ることができる。これによって今後、教科書の特性などの調査が行えるかも知れない。また、ここで述べた方法は単に日本だけでなく、他の英語教育が行われている国にも適用できる。それによって各国間の差を見ることができるかも知れない。

教科書の初出現学年を難易度とする辞書の作成には JACET 8000 のランクを難易度にした単語辞書及び 10, 20, 30 の難易度の付いた熟語辞書が用いられる。なぜなら設定する学年は 1 から

5 であり、初出現学年を求めるには使用する辞書の最低の難易度が 6 以上でなければならないからである。JACET 8000 の難易度の場合は、1 から設定されているが、1 から 5 の単語は初出現学年が必ず 1 であるから問題はない。しかし、今後様々な難易度に対応するために十分大きな一律の難易度を持った辞書作成専用の辞書を作っておいてもよい。

我々のプログラムは文の変形などのところで細かい処理を行っているが、これらの方法に合わないケースもあるだろう。特に段落の取り方や会話部分の表示法など、文の書き方によっても難易度は変わってくる。1 つの文の難易度を詳細に見るというよりは、ある程度全体的に平均を取って文を評価する必要があると思われる。

このシステムの作成においてデータ入力で三好文子氏に協力していただいた。心より感謝します。

参考文献

1) Nikolaos K. Anagnostou and George R. S. Weir, Average collocation frequency as an indicator of semantic complexity, ICTATLL Workshop 2007 Preprints, 1-3 August, 2007, 43-48.

2) Nikolaos K. Anagnostou and George R. S. Weir, From corpus-based collocation frequencies to readability measure, ICTATLL Workshop 2006 Preprints, 21-22 August, 2006, 33-46.

3) George R. S. Weir and Calum Ritchie, Estimating readability with the Strathclyde readability measure, ICTATLL Workshop 2006 Preprints, 21-22 August, 2006, 25-32.

4) 赤尾好夫偏, 綿貫陽補訂, 英語基本熟語集 大学入試 1000 熟語, 旺文社, 1991.

5) 小篠敏明・福井正康・細川光浩, 日本人英語学習者のためのリーダビリティ指標の開発 中間報告（1）, 福山平成大学経営研究, 4 号, (2008).

6) 社会システム分析のための統合化プログラム 7 －多変量解析－, 福井正康・細川光浩, 福山平成大学経営情報研究, 7 号, (2002) 85-106.

Development of English Readability Measurement System

Masayasu FUKUI and Toshiaki OZASA

Department of Management Information, Faculty of Management,
Fukuyama Heisei University

Abstract

Almost indices of English readability measurement which have so far been used are due to the number of syllables and word count. We suggested a new index which includes difficulty of words and idioms. For this purpose, we developed dictionaries of word and idiom difficulty and a program of analyzing English sentences with using these dictionaries. In this paper, we explain three main functions of this program that are measurement of readability, creating output data for a new index and making new dictionaries.

Keywords

English sentence, readability, program