【原著】

自然言語処理技術を用いたセンター試験問題の統計的解析

—英語および国語の試験問題を対象として—

石岡 恒憲，橘本 貴充，大津起夫（大学入試センター）

平成17年から平成20年度のセンター試験（本試験）の英語および国語の試験問題を対象として，その問題文における語彙の難しさや文章の難読性が得点率に影響を与える様子について調査した。その結果，センター試験で用いる範囲においては，これらは互いに無相関であることがわかった。また英語の文章は，4年次（9歳）から8年次（13歳）のレベルの英語によって構成されていることがわかった。

1 はじめに

平成9年以降における大学入試センター試験の解答データについての統計情報の整備は，研究開発部において着々とおこなわれてきた。これにより，今や，必要に応じて過去の関連問題を検索し，その詳細な統計情報を獲得することが可能となっている。

次に我々が目指すべきものは，自然言語処理技術を用いた試験問題文そのものがもつ属性に踏み込んだデータ解析とそのデータ提供である。たとえば，国語や英語の長文読解問題などで用いられる問題文そのものの難読性に関する指標や出題形式といった試験問題そのものがもつ属性が，受験者の得た得点とでどのような関係にあるかがわかれば，今後の作題において得点予測の十分な資料となることが期待される。難読性に関する指標には，たとえば（1）語彙の多様性を示すユール(Yule)のKなどの幾つかの指標，(2) 文の長さ，(3) 用いられて語彙それ自体の難易度，などを挙げることができる。また読みやすさの指標としても，英語などについては，Flesch Reading Easeなどがある（Baayen, 2008）。

このようなアプローチを取ることの技術的背景としては，従来，知識工学的なアプローチの多かった自然言語処理の分野に，1990年代頃よりコーパスと呼ばれる言語集体を用いた確率・統計的なアプローチが成功を収め，その有効性が多くの研究者や技術者に広く認知されてきたことがある。またこのような自然言語を取り扱うための道具立ても，整ってきている。事実，今回我々が用いる解析ツールは，GoogleDocs，統計言語 R およびその上で動作する languageR ライブラリ，形態素解析 MeCab(和布蕪)，および JACET 8000 レベルメーカであり，全てネット上からフリーで利用できるものである。

本稿では，平成17（2005）年から平成20（2008）年度のセンター試験（本試験）のうち，その問題文における語彙の難しさや文章の難読性が得点率に影響を与えやすいと考えられる英語と国語を取り上げ，その相互の影響について調査することを試みる。

2節では，語彙の難しさや難読性に関する指標の定義について整理しておく。3節では，これら指標が歴史的に整理されている英語を例に解析を試みる。4節では国語を取り扱う。日本語形態素解析を用いて，各種指標の適用を試み，その結果を報告する。5節には，問題文で用いられている語彙のネットワーク分析を試み，単なる語彙頻度だけでなく語彙の繋がりの様子が一目で見られることを示す。これら一連の解析では，解析のソースコードを示し，興味のある方には追試ができるよう配慮した。6節はまとめである。

2　文章の難読性や語彙の困難度を示す各種指標

2.1　GoogleDocs

GoogleDocsのワードカウント機能はGoogleドキュメント英語版で利用できる(http://docs.google.com/)。文書内の英単語数をカウントするには，[Tools] タブをクリックして [Word count...] を選択する。すると，ウィンドウに次の項目が表示される。

Counts
- Word count (単語数)
- Character count (with spaces) (スペースを含めた文字数)
- Character count (without spaces) (スペースを除いた文字数)
- Number of paragraphs (段落の数)
- Number of sentences (文の数)
- Approximate number of pages (おおよそのページ数)

Readability statics (文書全体の統計):
- Average sentences per paragraph (1段落あたりの平均文数)
- Average words per sentence (1文あたりの平均単語数)
- Average characters per word (1単語あたりの平均文字数)
- Average words per page (1ページあたりの平均単語数)
- Flesch Reading Ease
- Flesch-Kincaid Grade Level
- Automated Readability Index

このうち，Flesch Reading Ease(FRE)のスコアは１文あたりの単語数や１文あたりの音節数を考慮した読み易さの指標であり，スコアが大きいほど読み易いと判定される。スコアの概ねの目安として90.0–100.0は平均的な5年次の学生が容易に理解できるレベル，60.0–70.0 は8年次から9年次の学生が容易に理解できるレベル，0.0–30.0 はカレッジ卒業生が容易に理解できるレベルであるとしている。

またFlesch-Kincaid Grade Level(FKGL)は，FREと同様，１文あたりの単語数や１文あたりの音節数を考慮した読み易さの指標であるが，こちらはどの学年によって理解できるかを示すものである。たとえば8.2という数値が出たならば，それは平均的な8年次の学生（アメリカでは通常13歳から14歳）の学生にとって理解できるレベルであることを示している（www.readabilityformulas.com/)。GoogleDocsではこの数字は整数に丸められる。

一方，Automated Readability Index(ARI)は，ワードあたりの文字数や１文あたりの文字数を考慮した読み易さの指標であり，FKGL同様，対応する学年レベルを示す。これもまたGoogleDocsで数値が整数に丸められる。

2.2　langaugeR ライブラリ

Rは統計とグラフィックスのためのフリーのプログラミング環境である。Rにはbaseと呼ばれる標準ライブラリに加え，数多くの投稿されたライブラリ（拡張パッケージ）が用意され，ユーザは必要に応じて追加してインストールすることができる。

いま，RのライブラリであるlangaugeRをインストールすれば，述べ語数(Tokens)や異なり語数(Types)に加えて，トークン比やZipfの法則のパラメータの他，語彙の豊富さを示す指標であるYule's K(ユールのK), Herdan's C, Guiraud's R（ギロー指数），Sichel's Sなどの値を得ることができる（Baayen, 2008)。

ユールのK(Yule, 1944)は単語の頻度スペクトルを用いた指標である。出現頻度がポアソン分布に従うと仮定して，語彙が多様なほど 小さな値を取るように作られている。式で示せば，ある文章にx_i回現れた単語がf_i個あるとすれば延べ語数Nは$N=\sum x_i f_i$となり，そのときK特性値は

$$K=10^4\frac{\sum x_i^2 f_i - N}{N^2}$$

となる。いま1000語からなるテキストがあるとして，この1000 語が全て異なっているのならば， K特性値は 0 となる；一方，これが全て同じ単語であれば，K特性値は 9990 となる。

Herdan's C (Herdan, 1960)とGuiraud's R (Guiraud, 1954)は，延べ語数Nと異なり語数V を用いた指標であり，それぞれ次式で示される：

$$C=\frac{\log V}{\log N},\quad R=\frac{V}{\sqrt{N}}$$

Sichel's Sは1度だけ現れた単語（hapax dislegomena）を異なり語数V で割った指標である（Sichel, 1986)。

2.3 JACET 8000

JACET 8000は「大学英語教育学会基本語改訂委員会」（委員長：村田年 千葉大学名誉教授・和洋女子大教授）が，2003年3月に制定した「大学英語教育学会基本語リスト」の通称である。これは「日本人英語学習者のための教育語彙表」であり，英語学習の指針になることを目的としている。Level 1からLevel 8まで各レベルに1000語が割り当てられ，それぞれの意味づけは以下の通りである（相澤ほか, 2005)。

・Level 1 〔順位1000位まで〕
中学校英語教科書に頻出する基本語。一般英文の70%をカバー。

・Level 2 〔順位1001～2000位〕
高校初級。英字新聞の75%をカバー。英検準2級に相当。

・Level 3 〔順位2001～3000位〕
高等学校英語教科書・大学入試センター試験は，ほぼこのレベルの単語で作成。 英検2級に相当。社会人は教養として必要なレベル。

・Level 4 〔順位3001～4000位〕
大学受験，大学一般教養初級。日本人が単語力の有無を問われるレベル。英検2級に相当。

・Level 5 〔順位4001～5000位〕
難関大学受験，大学一般教養。英検準1級のレベル。TOEICでは，おおよそ400点から500点前後に相当。

・Level 6 〔順位5001～6000位〕
英語専門外の大学生やビジネスマンが目標とするレベル。英検準1級，TOEICでは600点に相当。

・Level 7 〔順位6001～7000位〕
英語専門の大学生，英語教師，仕事で英語を使うビジネスマンの到達目標。英検1級やTOEICでは95%以上の単語をカバー。

・Level 8 〔順位7001～8000位〕
日本人英語学習者の最終目標。英語を仕事して使う場合，95%の単語を知っていることに。英検1級やTOEICでは95%以上の単語をカバー。

あるテキスト文書を入力して，そこで用いられている単語にレベル付けを行うツールとしてJACET8000 レベルメーカがあり，ネット上で利用できる (http://www01.tcp-ip.or.jp/~shin/J8LevelMarker/j8lm.cgi)。

3 センター試験・英語

3.1 試験問題の構成

平成 4 年（1992 年）以降，英語の出題形式はほぼ一定で 6 つの大問より構成される。第 1 問がアクセント問題，第 2 問が単文の穴埋め問題（文法問題），および短い会話文の穴埋め問題，第 3 問から第 6 問が読解問題である。読解問題にはグラフの説明や料理レシピなどの説明文などが含まれる。4 コマ漫画の説明として適当なものを選ぶ問題の年もあり，書かれている内容については一定ではない。ただ第 6 問は，例年，比較的分量のある読解問題が出題される。平成 18 年度からはリスニングテストが導入された。

3.2 語数や読み易さに対する評価

本稿では，平成 17 年度から平成 20 年度ま

での 4 年間の試験問題を対象とした。第 3 問から第 5 問までを併合したものと第 6 問の 2 つのデータ対象に対して，総語数や Flesch Reading Ease(FRE)が大問得点率にどのように影響を与えるのかを示したのが図 1 である。上段の 2 つが総語数を説明変数とする得点率のグラフ，下段の 2 つが FRE を説明変数とする得点率のグラフである。図中の直線は回帰直線であり，グラフの枠内の右下には相関係数を記してある。ここで FRE を用い（学年を示す）FKGL を用いないのは，FKGL は数字が丸められているために，関係性をみるためには FRE の方がより良いと考えるからである。

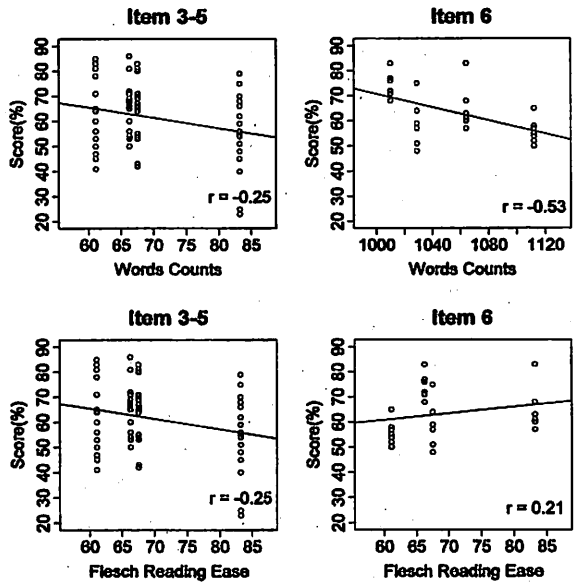

図 1 ：語数や読み易さに対する得点率の変化

これより，総語数が増えるとわずかではあるが，正解率が低下する傾向のあることがわかる。一方，読み易さの指標と得点率の相関は少ないように感じられる。なお，FRE の値が 60 から 85 程度であることからわかるように，センター試験の英語は 4 年次（9 歳）から 8 年次（13 歳）程度のレベルの英語が使われていることがわかる。勿論，常識的に考えれば，総語数や FRE の値は正解率に影響を与えるはずのものであるが，センター試験ではそのレンジの幅が作題者の調整により小さいために相関が現れないことは容易に推察できる。

3.3 語彙の多様性に対する評価

Yule's K（ユールの K），Herdan's C，Guiraud's R（ギロー指数）を説明変数として，大問得点率がどのように変化するのかを示したのが図 2 である。上段の 2 つが Yule's K に対するグラフ，中段の 2 つが Herdan's C に対するグラフ，下段の 2 つが Guiraud's R に対するグラフである。図 1 同様，左側の 3 つが大問 3 から大問 5 までをまとめたもので，右側の 3 つが大問 6 に対するものである。

これらを見ると，大問 3 から大問 5 までをまとめたものに対しては，これら代表的な語彙の多様性を示すいずれの指標においても，得点率には影響がない（無相関である）ことがわかる。大問 6 についても，サンプル数が少ないために確定的なことはいえないが，相関がないように推察される。

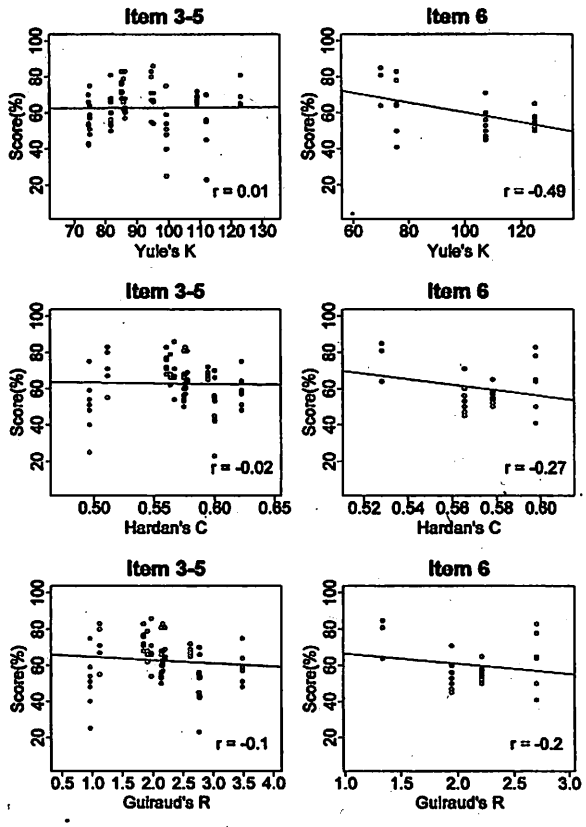

図 2 ：語彙の多様性に対する得点率の変化

3.4 JACET 8000 を用いた評価

表 1 は，平成 17 年度から平成 20 年度までの各大問における用いられた単語の Level 別頻度の平均である。その他は Level 8 までに含まれない単語の他，試験問題という特質のためにたとえば「①」のような英語以外の文字が含まれている。単語には問題文の英文も含まれる。

表 1 : JACET8000 のレベル分類（平成 17 年度から 20 年度までの平均)による語数と比率

Level	大問3 (%)	大問4 (%)	大問5 (%)	大問6 (%)
Level 1	703.5 (80.8)	419.0 (77.5)	465.0 (80.9)	885.3 (83.9)
Level 2	68.8 (7.9)	54.5 (10.1)	34.0 (5.9)	62.3 (5.9)
Level 3	25.3 (2.9)	17.3 (3.2)	11.8 (2.1)	23.5 (2.2)
Level 4	14.0 (1.6)	17.0 (3.1)	4.8 (0.8)	16.0 (1.5)
Level 5	10.3 (1.2)	6.0 (1.1)	11.0 (1.9)	3.0 (0.3)
Level 6	7.5 (0.9)	0.8 (0.1)	1.0 (0.2)	5.3 (0.5)
Level 7	4.5 (0.5)	1.3 (0.2)	4.2 (0.7)	1.8 (0.2)
Level 8	3.3 (0.4)	2.3 (0.4)	0.8 (0.1)	1.3 (0.1)
その他	33.3 (3.8)	22.5 (4.2)	42.3 (7.4)	56.5 (5.4)
計	870.3 (100)	540.5 (100)	574.8 (100)	1054.8 (100)

Level 1〔順位 1000 位まで〕が中学校英語教科書に頻出する基本語で，一般英文の 70%をカバーすることを考えれば，センター試験の英語はこれが 80%を占めており，かなり平易な語彙により構成されていることがわかる。Level 2〔順位 1001～2000 位〕は高校初級で，英字新聞の 75%をカバーし，英検準 2 級に相当するわけだが，センター試験はここまでで約 87%を占めている。

その一方で，センター試験のレベルとされる Level 3 より難しい Level 4 以降の単語も少なくなく，各問で平均して 30 単語が含まれている。大問 6 は分量が多いせいか，他の大問（大問 3 から大問 5）に比べ，易しい単語の比率が大きいこともわかる。ただ全体を通して Level 5（難関大学受験レベル）を超える単語はさすがに少ない。

全体の分量としては，大問 3 から大問 6 までのいわゆる読解で，毎年 3000 語の文章を読んでいることになる。ちなみに共通一次時代の読解量は約 1300 語である。

4 センター試験・国語

4.1 試験問題の構成

国語の試験問題の構成は，第 1 問が評論，第 2 問が小説（近代），第 3 問が古文，第 4 問が漢文で，この構成および配点（各50点）は，共通一次時代より変わっていない。

4.2 語彙の多様性に関する評価

形態素解析にR上で動作するRMeCabを使い，英語の試験問題の解析同様にlangaugeRライブラリを用いてYule's Kを算出した。このYule's Kを説明変数とした場合に，国語の得点率がどのように変化するかを示したのが図 3 である。これより語彙の多様性の指標であるYule's Kの値の違いによって，得点率が変化するという傾向は認められなかった。読解における理解の難しさと，得点とは別ということであると思われる。作題者は概ね60%を目標として試験問題を作成しており，素材文の難しさを設問の易しさで相殺しているのだと推測される。

著者らは，Herdan's CやGuiraud's R（ギロー指数）についても解析結果を得ているが，特に興味深い結果を得ることができなかった。

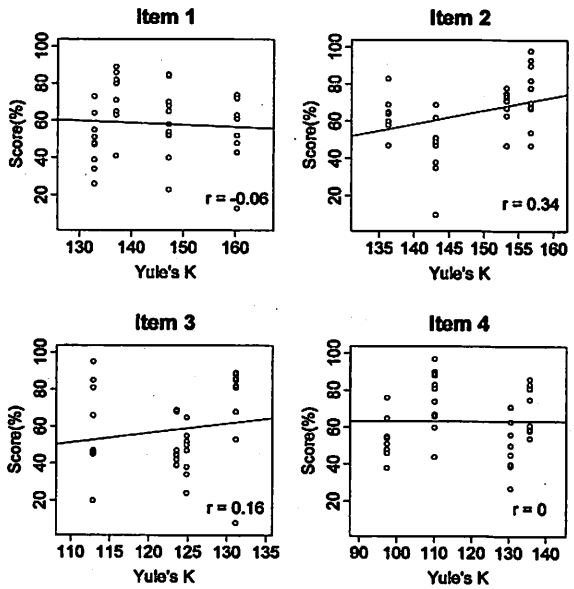

図 3 : Yule's K に対する国語の得点率の変化

5　語彙のネットワーク分析

図４は，平成20年度・国語・第1問の素材文として用いられた狩野俊次「住居空間の心身論―『奥』の日本文化」で用いられた語彙の繋がりを有向グラフで示したものである。形態素解析にR上で動作するRMeCabを使い，igraphライブラリを用いてR上で実行・描画した。これより，「空間」や「的」といった単語が，この素材文のキーワードになっている様子がみてとれる。なお，この図は正味でわずか15行足らずのソースコードで作成することができる（図５）。

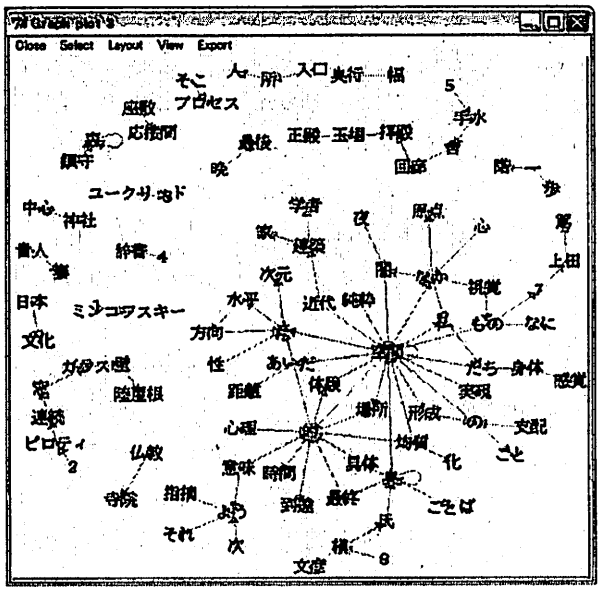

図４：平成 20 年度・国語・第 1 問，素材文に含まれる語彙のネットワーク分析

```
#  語彙のネットワーク分析
#
#  ライブラリーとデータの読み込み
library(igraph)
library(RMeCab)
targetText<-
"F:/Ctr/DNC.doc/2008H.doc/2008-A1-H-01ed.txt"
# 名詞のバイグラムをとる
kekkaDF <- NgramDF(targetText, type = 1, N = 2, pos="名詞")
# 集計した結果を度数 (Freq) の降順に
sortlist <- order(kekkaDF[,3], decreasing = TRUE)
fwn <- ekkaDF[sortlist,]
# 頻度2以上を取り出す
y <- fwn$Freq
freq <- length (y[y>=2])
fwn[1:freq,]
# ネットワークマップデータに置き換える
wng <- graph.data.frame(fwn[1:freq,])
# ネットワークマップを作成
tkplot(wng,vertex.label=V(wng)$name, layout=
    layout.fruchterman.reingold,vertex.size=1)
```

図５：ネットワーク分析の R によるプログラム

6　おわりに

センター試験は大問形式で構成されているために，同じ素材文に対して易しい設問から難しい設問まで幾つかが設定される。その得点率のバラツキの方が，素材文自体の読みやすさや難しさのバラツキに比べて大きいために，結果として素材文の難しさと得点率には相関が現れないことがわかった。逆にいえば，出題者は素材文の読みやすさ/難しさにかかわらず，同程度（約 60%）の得点率を目指しており，それがほぼ実現している様子が確認された。これは英語と国語とで共通である。

また英語では，素材文で用いられている単語の読み易さのレベルは，4 年次から 8 年次である；単語の難しさについても大学受験レベル（Level 4 および Level 5)以下に概ね収まっており，これらを超える難しさの単語は全体のわずか 1.5%であることがわかった。

ただ語彙の多様性については、単語の出現しにくさの度合いを考慮すべきであり、その定式化を含め、今後検討したいと考えている。

参考文献

相澤 一美，ほか(2005). JACET8000 英単語―「大学英語教育学会基本語リスト」に基づく―，桐原書店.

Baayen, R. H. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics Using R, Cambridge Univ. Press (Sd).

Guiraud, P. (1954). Les Caractères statistiques du vocabulaire. Paris : P.U.F. Massonie.

Herdan, G. (1960). *Type-token mathematics*. A textbook of mathematical linguistics, Gravenhage, the Netherlands: Mouton.

Sichel, H. S. (1986). Word frequency distributions and type-token characteristics. *Math. Scientist*, 11, 45–72.