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, 橋本 貴充, 大津起夫

(大学入試センター)

背景・研究の位置づけ

- 平成2年以降のセンター試験解答データについての統計情報の整備
  - 研究開発部において着々と進展
  - 過去の関連問題の検索
  - 詳細な統計情報の獲得が可能
- 自然言語処理技術の活用
  - 試験問題文そのものがもつ自然言語の属性に踏み込んだデータ解析
  - そのデータ提供

言語的解析の有用性

- 国語や英語の長文読解問題
- 試験問題がもつ属性と試験得点との関係
  - 難読性に関する指標、出題形式
- 今後の作題において得点予測の十分な資料

言語的解析の視点

- 難読性に関する指標
  - 語彙の多様性を示すユールのKなどの指標
  - 文の長さ
  - 用いられて語彙それ自体の難易度
- 読みやすさの指標
  - 英語などについては Flesch Reading Easeなど
  - Harry Potter vs Kane and Abel

言語的解析の技術的基盤

- 自然言語処理の分野
- 1990年代頃よりコーパス(言語集体)を用いた確率・統計的なアプローチの成功
- 有効性が多くの研究者や技術者に広く認知
- 道具立ての整備
  - GoogleDocs
  - 統計言語RおよびlanguageRライブラリ
  - 形態素解析MeCab(和布蕪)
  - JACET 8000レベルメーカ
  - 全てネット上からフリーで利用可能

本研究でおこなったこと

- 平成17－20(2005－2008)年度センター試験(本試験)
- 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
- その相互の影響について調査

構成

- 語彙の難しさや難読性に関する指標の定義
- 英語：指標が歴史的に整理されている
- 国語：日本語処理、形態素解析
- 語彙のネットワーク分析の例

GoogleDocsのワードカウント機能

- Googleドキュメント英語版 http://docs.google.com
- [Tools] タブをクリックして [Word count...] を選択
- **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)

GoogleDocsのワードカウント機能（続き）

- **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 (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、0－100)
  - Flesch-Kincaid Grade Level (学年に換算)
  - Automated Readability Index (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、学年に換算)

langaugeR ライブラリ

- R：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
- baseと呼ばれる標準ライブラリ＋投稿された多くのライブラリ (拡張パッケージ)
- ユーザは必要に応じて追加インストール
- langaugeR (Baayen 2008)
  - 述べ語数(Tokens)や異なり語数(Types)
  - トークン比やZipfの法則のパラメータ
  - 語彙の豊富さを示す指標：Yule's K (ユールのK), Herdan's C, Guiraud's R (ギロー指数), Sichel's Sなど

語彙の豊富さを示す指標

- ユールのK；ある文章に X_i 回現れた単語が f_i 個
  - 延べ語数N

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

- Herdan's C と Guiraud's R
  - 延べ語数 N と異なり語数 V

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

JACET 8000

- 「大学英語教育学会基本語改訂委員会」, 2003年3月に制定
- 「大学英語教育学会基本語リスト」の通称
- 日本人英語学習者のための教育語彙表」→英語学習の指針
- Level 1からLevel 8まで各レベルに1000語が割り当て

レベルの意味づけ

- 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。

レベルの意味づけ(続き)

- 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。
- 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。

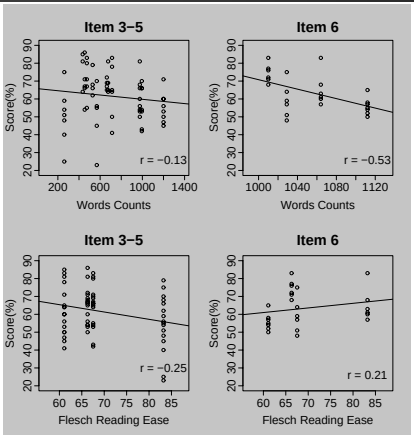
JACET8000レベルメーカ

- 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
  - http://www01.tcp-ip.or.jp/~shin/J8LevelMarker/j8lm.cgi

センター試験・英語の構成

- 平成4年 〔1992年〕 以降，出題形式はほぼ一定
- 6つの大問より構成
- 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
- 読解問題にはグラフの説明や料理レシピの説明も
- 第6問は，例年，比較的分量がある

語数や読み易さに対する得点率の変化

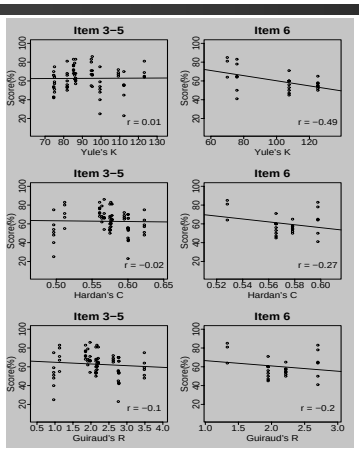


- 上段：総語数を説明変数とする得点率のグラフ
- 下段：FREを説明変数

図よりわかること

- 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
- 読み易さの指標と得点率の相関は少ない
- FREの値が60から85程度⇒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
- 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⇒センター試験ではそのレンジの幅が小さいために相関が現れない

語彙の多様性に対する評価



- 上段：Yule's K (ユールのK)を説明変数とする得点率のグラフ
- 中段：Herdan's C
- 下段：Guiraud's R (ギロー指数)

図よりわかること

- 大問3から大問5：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)
- 大問6：サンプル数が少ないために確定的なことはいえない；相関がないように推察される

JACET 8000を用いた評価

Level	大問3	大問4	大問5	大問6
Level 1	703.5	419.0	465.0	885.3
Level 2	68.8	54.5	34.0	62.3
Level 3	25.3	17.3	11.8	23.5
Level 4	14.0	17.0	4.8	16.5
Level 5	10.3	6.0	11.0	3.0
Level 6	7.5	0.8	1.0	5.3
Level 7	4.5	1.3	16.2	1.8
Level 8	3.3	2.3	0.8	1.3
その他	33.3	22.5	42.3	56.5
計	870.3	540.5	547.8	1054.8

- Level 1中学校英語教科書に頻出する基本語 (70%) →80%
- Level 2高校初級で (英字新聞の75%) →87%
- Level 4以降の単語は少なくない；各問で平均して30単語

英文の分量

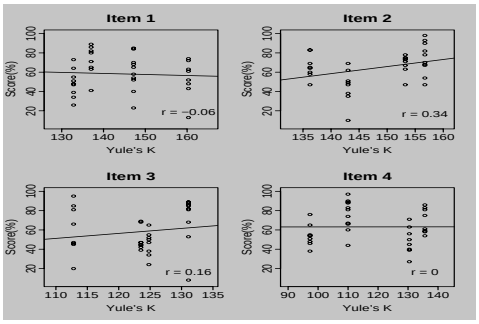
- 大問3から大問6までのいわゆる読解で毎年3000語
- 共通一次時代の読解量は約1300語

センター試験・国語 の構成

- 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
- 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

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
- langaugeRライブラリを用いてYule's Kを算出
- Yule's Kを説明変数とする国語の得点率

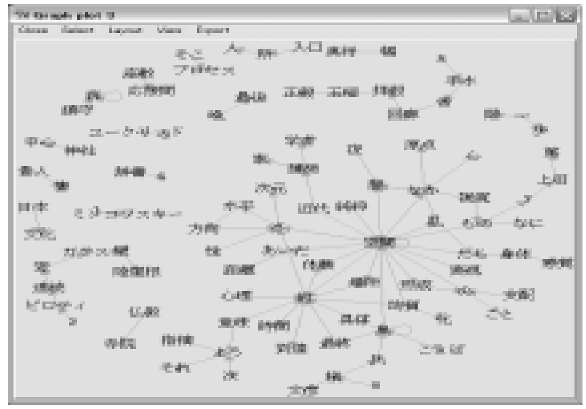


図よりわかること

- 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い
- 読解における理解の難しさと，得点とは別
- 作題者は概ね60%を目標として試験問題を作成
- 素材文の難しさを設問の易しさで相殺している（？）

語彙のネットワーク分析

平成20年度・国語・第1問、狩野俊次「住居空間の心身論—『奥』の日本文化」



ネットワーク分析のRによるプログラム

```
• #  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 <- kekkaDF[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)
```


まとめ

- センター試験は大問形式
- 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
- 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
- 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
- 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

まとめ（続き）

- 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
- 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いる
- これらを超える難しさの単語は全体のわずか1.5%

ご清聴ありがとうございました