Coh-Metrix とパターン認識を用いた課題英作文の自動評価

小林 雄一郎
日本学術振興会

金丸 敏幸
京都大学

本研究は，Coh-Metrix とパターン認識の技術を用いて，英語学習者による課題英作文の自動評価を試みるものである。自動評価の基準は，既存の評価システムによる分類結果とし，出力結果を比較，検討することによって，よりよい採点システムの可能性を模索する。

Automated essay scoring using Coh-Metrix and pattern recognition

Yuichiro Kobayashi
Japan Society for the Promotion of Science

Toshiyuki Kanamaru
Kyoto University

The aim of the present study is to explore the possibility of computer-based automated essay scoring system. By applying pattern recognition and Coh-Metrix to the assessment of English writings, the present paper proposes a new method for automated essay scoring, and the result is verified by the comparison with that of the existing writing-scoring system.

1. はじめに

英語教育の分野では，数多くの英語力テストが存在し，中学校や高校でのカリキュラムに組み込まれている場合もある。これらのテストの多くは，熟練した教師や採点者が，学習者の英作文や発話を評価するという形式を取っている。しかし，熟練した採点者を育成するには，かなりの時間が必要とされるであろう。また，いかに熟練した採点者たちが厳密な基準に基づいて評価を下したとしても，複数の採点者間の評価が一致しないこともある。そのような状況において，客観的な評価基準と統計モデルを用いて習熟度を推定する技術を開発することは，言語教育分野にとって非常に有意義なことである。

本研究では，習熟度の情報が付与された課題英作文を対象とし，Coh-Metrix (Graesser, *et al.,* 2004) の出力を説明変数の候補とするランダムフォレストを用いて，書き手の習熟度を推定する。また，目的変数となる習熟度は，Educational Testing Service (ETS) が開発した自動採点システム e-rater による推定結果とする。

2. 先行研究

英作文の評価は，テキスト全体に対する評定者の印象に基づく全体的評価 (holistic scoring)，作文を語彙や文法などの項目別に評定していく分析的評価 (analytical scoring)，そして，ある特定の要因（修辞的特徴など）がどの程度作文に反映されているかを基準とする特定要因の評価 (primary trait scoring) の 3 つに分けられる (Perkins, 1983)。このうち，比較的実用性が高く，多くのテストや評価で使用されているのは全体的評価であり (Weigle, 2002)，1960 年代から始まった自由英作文の自動採点システムにおいても，表面的な特徴を用いて，自由英作文の全体的評価を行ってきた (Shermis & Burstein, 2003)。また，自動採点システムで使われることが多い説明変数は，単語の出現頻度ベクトル，文や句の長さ，代名詞や助動詞の数，議論を深めるための手がかり語や修辞句などである。

そして，自動採点システムには，パターン認識の技術が適用されることがある。パターン認識は，音声認識，手書き文字認識，X 線画像・CT 画像からの病気の診断，指紋・静脈・虹彩などによる本人識別などを含む様々な分野で用いられている。これらは全て，対象の特徴を表す何らかの

量を手がかり（説明変数，特徴量，素性）とし，対象の属性を表す識別子（目的変数，クラス）を推定するという形で定式化される (e.g. Bishop, 2006)。

言語データに対するパターン認識の適用例としては，検索キーワード（説明変数）から適切なウェブサイトであるか（目的変数）を判定したり，テキスト中のキーワード（説明変数）からスパムメールやスパムブログ（目的変数）を自動選別したりするテキスト分類の技術が知られている。自動採点システムでは，その頻度に書き手の習熟度が如実に反映される言語項目を手がかり（説明変数）とし，作文の質（目的変数）を推定する。自動採点システムで用いられるテキスト分類の技術としては，ナイーブベイズや k 近傍法などが知られている (Larkey & Croft, 2003)。

パターン認識の技術を英作文評価に応用した先行研究の 1 つとして，小林・金丸 (2012) が挙げられる。この論文は，習熟度の情報が付与された課題英作文を対象とし，12 種類の言語的特徴を説明変数とするランダムフォレストを用いて，課題英作文における書き手の習熟度 (5 段階) を推定した。その精度は 62.32%で，主に総語数や異語数が分類に寄与していることが分かった。

3. 実験手法

3.1. 説明変数

本研究で用いる説明変数は，Coh-Metrix 2.0 によって定量化された 56 の指標である（本稿末尾の付録 1 を参照）。

Coh-Metrix は，語彙の数や文の長さのようなテキストの形式的な面だけではなく，文と文，あるいはパラグラフとパラグラフの間の結束性 (cohesion) や一貫性 (coherence) などを視野に入れた，新しいリーダビリティ指標である (Crossley, *et al.,* 2008)。この指標は，これまでレジスター分析 (e.g. Louwerse, *et al.,* 2004)，著者推定 (e.g. McCarthy, *et al,* 2006)，英作文の評価 (e.g. Crossley, 2009; Crossley, *et al.,* 2012)，話し言葉の分析 (e.g. Crossley, *et al.,* 2011) など，様々な言語分析に応用されてきた。

従来の自動評価システムでは，語数や文数という英作文の表面的な特徴しか扱っていないことが多かったが，Coh-Metrix の指標を用いることによって，英作文を多角的に評価することが可能になる。

3.2. 相関係数

相関係数とは，2 変数間の直線的関係の強さおよびその方向を表す指標である。本研究では，Coh-Metrix 2.0 によって定量化された 56 種類の説明変数に，書き手の習熟度 (Level) という目的変数を加えた，全 57 変数の相関行列を作成し，目的変数と関係の強い説明変数を探る。なお，相関係数に解釈にあたっては，相関係数の t 検定を用いる。

3.3. ランダムフォレスト

習熟度推定に用いる手法は，Breiman (2001) によって提案されたランダムフォレストである。端的に言えば，ランダムフォレストとは，決定木のアンサンブル学習である。決定木とは，非線形判別分析の 1 つとして位置付けられ，説明変数の値を何らかの基準で分岐させ，分類モデルを構築する。分岐の過程は，木構造で図示することができ，IF-THEN のような簡単なルールで表すこともできる。また，アンサンブル学習とは，必ずしも精度の高くない複数の分類器の結果を組み合わせ，精度を向上させるパターン認識の手法である。

ランダムフォレストでは，まず，与えられたデータセットから，N 組のブートストラップサンプルを作成する。次に，各々のブートストラップサンプルデータを用いて，未剪定の最大の決定木を生成する（但し，分岐のノードは，ランダムサンプリングされた説明変数のうち最善のものを使用する）。そして，全ての結果を多数決で統合し，新しい分類器を構築する (Hastie, *et al.,* 2009)。

ランダムフォレストの長所としては，精度が高いこと，非常に多くの説明変数を扱うことができること，それぞれの説明変数が予測に寄与する度合いが分かること，などが挙げられる。

なお，小林ほか (2011) は，英語科学論文の質判定を行う際に，線形判別分析，ナイーブベイズ，決定木，k 近傍法，ニューラルネットワーク，学習ベクトル量子化，サポートベクターマシン，バギング，ブースティング，ランダムフォレストという 10 種類の分類器の精度比較を行った。その結果，アンサンブル学習を用いた分類器（バギング，ブースティング，ランダムフォレスト）の精度が高く，その中でランダムフォレストによる精度が最も高かった。

4. 結果と考察

4.1. 実験データ

本研究の実験データは, 国立大学の学部２年生を対象に英作文（エッセイ）課題を与えて収集したものである。データの収集は，2011 年の 10 月から 11 月にかけて，２回に分けて収集した。

英作文課題には，ETS が運営している Criterion のサイトで提供されている TOEFL 形式の問題サンプルから，“expository essay”と“persuasive essay”の問題を１問ずつ選出して使用した。両課題とも，指定語数は 300 語から 350 語である。“expository essay”の課題は, “New Product”というものであり，提出された英作文の数は 34 本であった。“persuasive essay”の課題は, “Money on technology”を選択し，提出された英作文の数は 35 本であった (課題の詳細については, 本稿末尾の付録２を参照)。一般的に学習データは 300 本以上，そして，各レベルに 20 本以上が必要であると言われているが (Elliot, 2003), 今回のデータセットには 2 つの英作文課題に対して合計で 69 本の英作文しか含まれていない。

英作文課題の提出に際して，受講生は，制限時間 30 分以内に英作文を作成し，自動評価システムである e-rater による評価を受けた。また，評価に満足しない場合には，一度だけ同じ条件で e-rater による評価を参考に英作文を書き直すことが認められた。

今回の実験データを収集する際に利用した Criterion とは，非営利団体である ETS が提供しているライティング指導におけるフィードバック支援ツールのことである。Criterion は，教師の管理下で使用する指導者向けのフィードバックツールであり，ETS により開発された TOEFL テストのライティング自動採点システム e-rater と連携している (Attali & Burstein, 2006)。Criterion では利用者の英作文を e-rater が 10 秒以内に採点し，英作文中の誤りを「構成 (organization & development)」，「文体 (style)」，「構造 (mechanics)」，「語法 (usage)」，「文法 (grammar)」の 5 つの観点から分析し，一度に提示する。Criterion の与えるフィードバックと教師によるフィードバックの関連性を検討した先行研究では，0.97 という高い相関 (Attali & Burstein, 2006) や0.64～0.67 という中程度の相関 (Weigle, 2010) を示すという報告がある。

収集した英作文を e-rater によって自動評価した結果を表１に示す。e-rater の評価は１点から６点までの点数で行われる。ここでは e-rater の点数を，そのまま英作文の習熟度の指標として使用した。但し，今回提出された英作文の中には評価が１点のものはなかった。

表１ e-rater による習熟度の評価結果（作文数）

L2	L3	L4	L5	L6
3	8	17	33	8

4.2. 変数間の関係

表１にある 69 本の英作文を対象に, Pearson の積率相関係数を用いて，56 種類の説明変数に，書き手の習熟度 (Level) という目的変数を加えた，全 57 変数の相関行列を作成した。そして，相関係数の t 検定を用いて，対象数が 57 の場合の相関の基準を求めた。その結果，２変数間の相関係数が 0.39 以上であれば，0.1%水準の有意差があると判断し得ることが分かった。

全 57 変数の相関行列において，目的変数との相関係数が 0.39 以上の説明変数は，異語数 (READNW) の 0.83 と文数 (READNS) の 0.54 のみであった。

4.3. 習熟度の推定

ランダムフォレストによる習熟度推定にあたって，ランダムサンプリングする説明変数の数は, 説明変数の数の正の平方根を取り，木の数は 500 とした。

表２は，69 本の英作文の書き手の習熟度をランダムフォレストで推定し， OOB (out-of-bag) による交差妥当化を行った結果である。因みに，OOB とは，ブートストラップサンプルの３分の２で分類モデルを作成し, 残りの３分の１を用いて評価を行った結果の誤判別率である。表中の accuracy は，そのレベルにおける推定精度を表している。この図を見ると，69 本のうち 37 本が正しく推定されているため，全体の精度 (exact agreement) が 53.62%であることが分かる。また，レベル 4 をレベル 3 やレベル 5 と推定した例のように，１つまでの誤差を正答とみなした場合の精度 (adjacent agreement) は，95.65%である。そして，本実験の推定結果と e-rater による評価結果の相関係数は，0.63 である。

表２　ランダムフォレストによる習熟度推定の結果

	L2	L3	L4	L5	L6	accuracy
L2	0	1	2	0	0	0.00
L3	0	1	6	1	0	12.50
L4	0	2	7	8	0	41.18
L5	0	0	4	29	0	87.87
L6	0	0	0	8	0	0.00

そして，図１は，56 種類の説明変数に関して，本実験における寄与度 (MeanDecreaseGini) の大きい順にプロットしたものである。この図を見ると，英作文の書き手の習熟度を推定するにあたって，異語数 (READNW) が大きく寄与していることが分かる。英作文の評価にあたって，熟練した採点者もまずは語数を見るという報告 (e.g. Erdosy, 2004) もあることから，これは極めて妥当な結果である。また，それ以外では，文数 (READNS)，平均文長 (READASL)，場所を表す前置詞と動作を表す前置詞の比率 (SPATC)，意図的な動作や出来事の数 (INTEi)，名詞句の数 (DENSNP)，などが推定に寄与している。

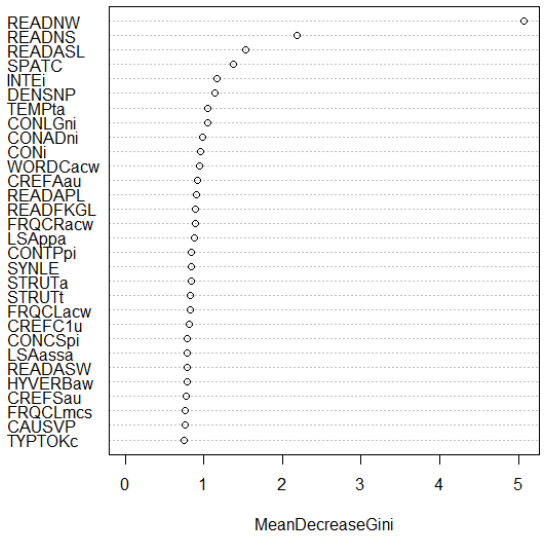

図１　各説明変数の寄与度

以下の図２〜4 は，ランダムフォレストにおける寄与度が特に高かった異語数 (READNW)，文数 (READNS)，平均文長 (READASL) に関して，習熟度による分布の違いを箱ひげ図で示したものである。

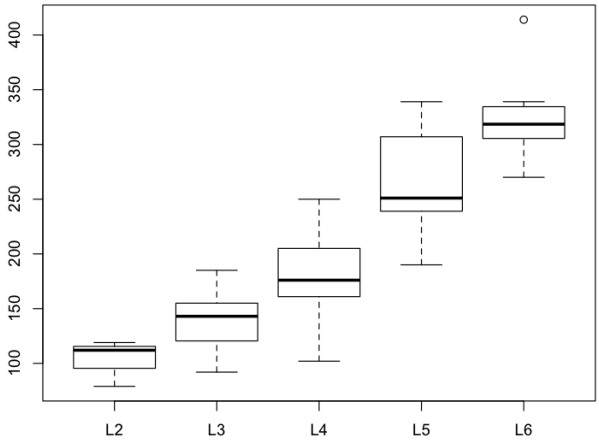

図２　異語数の分布

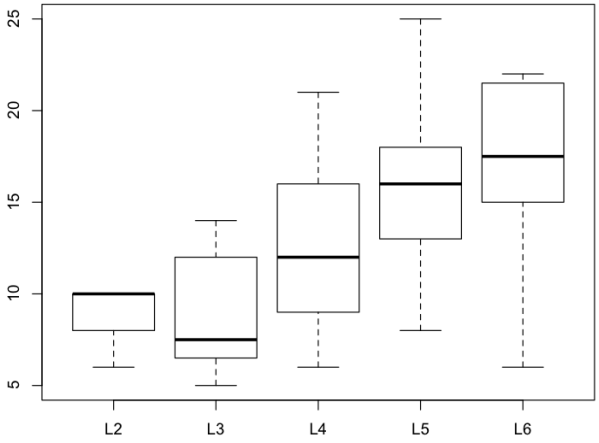

図３　文数の分布

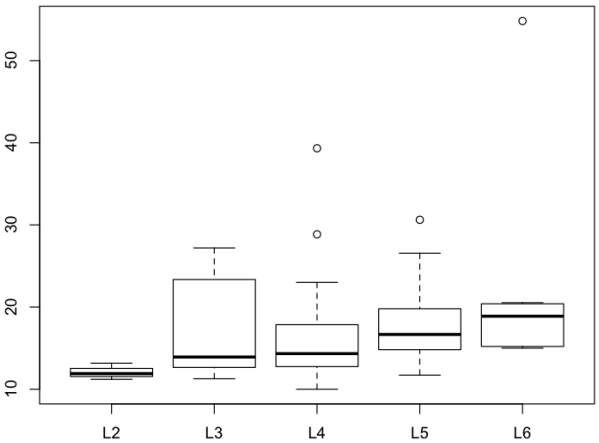

図４　平均文長の分布

図２を見ると，異語数が今回の実験データにおける習熟度の推定に極めて有効な指標であることが分かる。各習熟度における異語数の平均値に一元配置分散分析を行うと，0.1%水準での有意差が見られる（F = 51.94, df = 4, p-value = 0.00)。

また，図３を見ると，レベル３からレベル６にかけて，文数の中央値が上がっていくことが分かる。各習熟度における文数の平均値に一元配置分散分析を行うと，0.1%水準での有意差が見られる（F = 9.01, df = 4, p-value = 0.00)。

そして，図４を見ると，平均文長の中央値と習熟度の間にも相関関係がわずかに見られる。各習熟度における平均文長の平均値に一元配置分散分析を行うと，0.1%水準での有意差が見られる（F = 9.54, df = 4, p-value = 0.00)。

5. おわりに

本研究では，ランダムフォレストを用いて，課題英作文における書き手の習熟度を推定した。その精度は53.62%で，異語数，文数，平均文長，場所を表す前置詞と動作を表す前置詞の比率，意図的な動作や出来事の数，名詞句の数，などが分類に寄与していることが分かった。

今後の課題としては，実験データを増やし，レベル間における作文数の偏りを減らすことが挙げられる。今回の実験に使用した英作文の評価が全体的に高かったのも，Criterion のフィードバックの利用を許可したことによるところが大きい。その意味で，今回のデータは学習者の実力以上の英作文である可能性が高く，必ずしも受講者の実状を反映しているものとは言えない側面もある。学習者が誤りやすい冠詞などの文法誤りや単純なスペリングミスなどが含まれた英作文を対象としたときに，より適切な分類が可能かどうか，また，そのような誤りが分類に影響を与えるか否かについては，一般に公開されている学習者コーパスなどを用いた分類実験との比較検討が必要である。さらに，CEFR (Common European Framework of Reference) における基準特性 (Hawkins & Filipović, 2012) など，他の習熟度判定指標との関係も明らかにされなければならない。

註

本研究の成果の一部は，科学研究費補助金（特別研究員奨励費（PD 実験））「パターン認識と自然言語処理の技術を用いた習熟度判定」（研究代表者：小林雄一郎）（2012-2014 年度），科学研究費補助金（基盤研究（B））「総合研究大学における英語学術論文作成技能の育成に向けた全学共通教育のコース設計」（研究分担者：金丸敏幸）(2010〜2013 年度)，科学研究費補助金(基盤研究(B))「第二言語ライティング研究の現代的課題と解決のための将来構想―東アジアからの発信」（研究協力者：小林雄一郎）(2012〜2015 年度) によって行われたものである。

参考文献

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. T*he Journal of Technology, Learning, and Assessment,* 4(3), 1-30.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York: Springer.

Breiman, L. (2001). Random forests. *Machine Learning* 24, 123-140.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly,* 42(3), 475-493.

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing,* 18, 119-135.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly,* 45(1), 182-193.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing,* 29(2), 243-263.

Elliot, S. (2003). IntelliMetric: From here to validity. In Shermis, M. D., & Burstein, J. C. (Eds.),

Automated essay scoring: A cross-disciplinary perspective (pp. 71-86). Hillsdale: Lawrence Erlbaum Associates.

Erdosy, M. U. (2004). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions. *TOEFL Research Reports*, 70, 1-62.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Intruments, and Computers,* 36, 193-202.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Second Edition. New York: Springer-Verlag.

Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework.* Cambridge: Cambridge University Press.

小林雄一郎・金丸敏幸 (2012). 「パターン認識を用いた課題英作文の自動評価の試み」『電子情報通信学会技術研究報告』112(103), 37-42.

小林雄一郎・田中省作・冨浦洋一 (2011). 「メタ談話標識を素性とするパターン認識を用いた英語科学論文の質判定」『人文科学とコンピュータシンポジウム論文集—「デジタル・アーカイブ」再考』(pp. 51-58) 情報処理学会.

Larkey, L. S. & Croft, W. B. (2003). A text categorization approach to automated essay grading. In Shermis, M., & Burstein, J. (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 55-70). Hillsdale: Lawrence Erlbaum Associates.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In Forbus, K., Gentner, D. & Regier, T. (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 843-848). Mahwah: Erlbaum.

McCarthy, P. M., Lewis, G. A., Dufty, D. F., & McNamara, D. S. (2006). Analyzing writing styles with Coh-Metrix. *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)* (pp. 764-769). Menlo Park: AAAI Press.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly,* 17, 651-671.

Shermis, M., & Burstein, J. (Eds.) (2003). *Automated essay scoring: A cross-disciplinary perspective.* Hillsdale: Lawrence Erlbaum Associates.

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing,* 27(3), 335-353.

付録 1：Coh-Metrix 2.0 の指標

1. Incidence of casual verbs, links, and particles (CAUSVP)
2. Ratio of causal particles to causal verbs (CAUSC)
3. Incidence of positive additive connectives (CONADpi)
4. Incidence of positive temporal connectives (CONTPpi)
5. Incidence of positive causal connectives (CONCSpi)
6. Incidence of negative additive connectives (CONADni)
7. Incidence of negative temporal connectives (CONTPni)
8. Incidence of negative causal connectives (CONCSni)
9. Incidence of all connectives (CONi)
10. Argument Overlap, adjacent, unweighted (CREFA1u)

11. Stem Overlap, adjacent, unweighted (CREFS1u)
12. Anaphor reference, adjacent, unweighted (CREFP1u)
13. Argument Overlap, all distances, unweighted (CREFAau)
14. Stem Overlap, all distances, unweighted (CREFSau)
15. Anaphor reference, all distances, unweighted (CREFPau)
16. Noun Phrase Incidence Score (DENSNP)
17. Ratio of pronouns to noun phrases (DENSPR2)
18. Number of conditional expressions, incidence score (DENCONDi)
19. Number of negations, incidence score (DENNEGi)
20. Logical operator incidence score (DENLOGi)
21. LSA, Sentence to Sentence, adjacent, mean (LSAassa)
22. LSA, sentences, all combinations, mean (LSApssa)
23. LSA, Paragraph to Paragraph, mean (LSAppa)
24. Personal pronoun incidence score (DENPRPi)
25. Mean hypernym values of nouns (HYNOUNaw)
26. Mean hypernym values of verbs (HYVERBaw)
27. Number of Paragraphs (READNP)
28. Number of Sentences (READNS)
29. Number of Words (READNW)
30. Average Sentences per Paragraph (READAPL)
31. Average Words per Sentence (READASL)
32. Average Syllables per Word (READASW)
33. Flesch Reading Ease Score (READFRE)
34. Flesch-Kincaid Grade Level (READFKGL)
35. Mean number of modifiers per noun-phrase (SYNNP)
36. Mean number of higher level constituents per word (SYNHw)
37. Mean number of words before the main verb of main clause in sentences (SYNLE)
38. Type-token ratio for all content words (TYPTOKc)
39. Celex, raw, mean for content words (FRQCRacw)
40. Celex, logarithm, mean for content words (FRQCLacw)
41. Celex, raw, minimum in sentence for content words (FRQCRmcs)
42. Celex, logarithm, minimum in sentence for content words (FRQCLmcs)
43. Concreteness, mean for content words (WORDCacw)
44. Incidence of positive logical connectives (CONLGpi)
45. Incidence of negative logical connectives (CONLGni)
46. Ratio of intentional particles to intentional content (INTEC)
47. Incidence of intentional actions, events, and particles (INTEi)
48. Mean of tense and aspect repetition scores (TEMPta)
49. Sentence syntax similarity, adjacent (STRUTa)
50. Sentence syntax similarity, all, across paragraphs (STRUTt)
51. Sentence syntax similarity, sentence all, within paragraphs (STRUTp)
52. Proportion of content words that overlap between adjacent sentences (CREFC1u)
53. Mean of location and motion ratio scores (SPATC)
54. Concreteness, minimum in sentence for content words (WORDCmcs)
55. Genre purity (GNRPure)
56. Topic sentence-hood (TOPSENr)

付録２：英作文の課題

“New Product” (expository essay)

Question: If you could invent something new, what product would you develop? Use specific details to explain why this invention is needed.

“Money on Technology” (persuasive essay)

Question: Some people think that governments should

spend as much money as possible on developing or buying computer technology. Other people disagree and think that this money should be spent on more basic needs. Which one of these opinions do you agree with? Use specific reasons and details to support your answer.