

英文解析プログラムから得られる各種指標を使った

テキスト難易度の推定

—教材作成への適用可能性—

水本 篤

関西大学

概要

本稿では、従来から使われているリーダビリティ指標に加えて、英文解析プログラムから得られる各種指標を使いテキスト難易度の推定を試みた。その結果、Coh-Metrix のような新しい指標を用いることにより、精緻な分析を行い、教材開発にも適用可能であることが示された。

Keywords: リーダビリティ、テキスト難易度、Coh-Metrix、多変量解析、教材開発

1. はじめに

学習者のレベルに合ったテキストの難易度の推定は、リーディング教材の選定だけでなく、テスト作成や研究での実験題材選択の際にも重要になる。また、ライティングやスピーキングの評価においても、パフォーマンスの結果として算出されたテキストの難易度を調べることによって、客観的な指標として自動採点システムに利用される場合もあるため (水本, 2008)、テキストの難易度を示すリーダビリティ指標は外国語教育においては大きな関心事である。

従来用いられてきた Flesch-Kincaid Grade Level のようなリーダビリティ指標や、TTR (Type-Token Ratio) のように、使用されている語彙の豊富さを示す指標 (Koizumi & In' nami, 2012) では、総語数や一文に含まれる平均語数など、英文テキストの表面的な特徴を用いて算出されている。しかし、最近では、英文全体でどの程度、文や句の結束性 (cohesion) があるかを示す指標も得る事ができる Coh-Metrix (<http://cohmetrix.memphis.edu/cohmetrixpr/index.html>) というプログラムが開発され、Crossley, Salsbury, and McNamara (2012) や小林・金丸 (2012) に見られるようにライティング評価にも利用されつつある。図 1 はデータベースの Proquest Basic Search を使い、国際誌掲載論文でどれだけ Coh-Metrix が近年使用されているかの推移を示したものである (2013 年 3 月 11 日検索)。この図を見ても、明らかにここ数年で

Coh-Metrix を使用した研究が増加していることがわかる。

Coh-Metrix は比較的簡易に数多くの指標を得る事ができるため、適用範囲は広い。そこで本研究では、Coh-Metrix を含む英文解析プログラムから得られた指標を用いてテキスト難易度の推定を行い、教材作成にそのような手法が適用できるかを検討することとした。

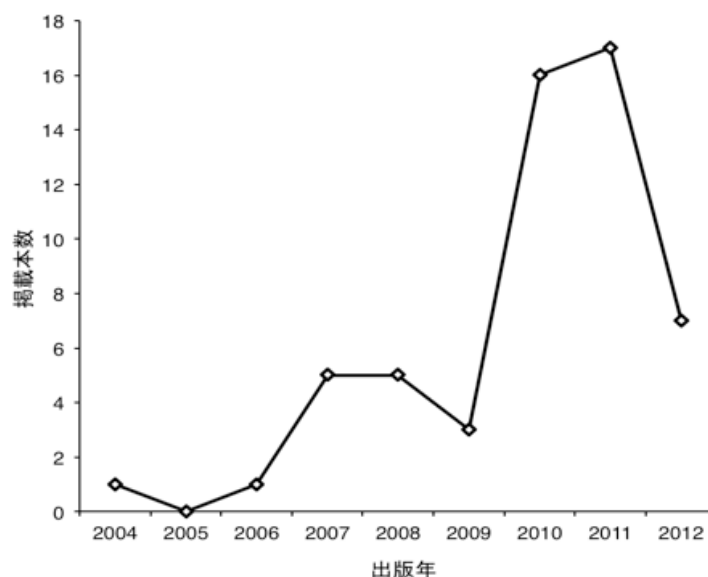


図 1. 国際誌掲載論文における Coh-Metrix の使用推移

2. 方法

2.1 分析対象とした教科書

ある出版社が発行予定の高等学校英語教科書（コミュニケーション英語Ⅲ）の1冊を分析対象とした（この教科書は出版前のものであるため、詳細については本稿では触れない）。教科書は全15レッスンで構成されており、多様なジャンルを取り上げ、受験対策も可能にするというのが編集者の意図であった。また、各レッスンの総語数は後半に行くほど長くなり、テキストの難易度も上がっていくという前提で作成されている。全15レッスンの平均語数は636.53語、最小語数は421語、最大語数は887語であった。

2.2 使用した指標

今回の分析で使用した指標を表1にまとめた。また、表2はこれらの指標間の相関係数を示している。Microsoft Wordで算出される指標のうち、Flesch-Kincaid Grade Levelは、アメリカの小学生・中学生のどの学年にふさわしい英文であるかが、次の式から計算される。

$$GradeLevel = (\text{一文に含まれる平均語数}) \times 0.39 + (\text{一単語あたりの平均音節数}) \times 11.8 - 15.59$$

Coh-Metrix 3.0 を利用して算出される指標は、Narrativity, Syntactic Simplicity, Word Concreteness, Referential Cohesion, Deep Cohesion, Verb Cohesion, Connectivity, Temporality の 8 つであった（それぞれの指標の詳細については Coh-Metrix のホームページを参照）。Coh-Metrix 3.0 では 108 の指標が算出されるが、Coh-Metrix Easability components と呼ばれている難易度を示す指標がこの 8 つであり、テキストの表面的特徴ではなく、言語的特徴を捉えた難易度が確認できるとされている (Graesser, McNamara, & Kulikowich, 2011)。

Word Level Checker (<http://someya-net.com/wlc/>) では、TTR, Average Word Level (SVL), Average Word Level (JACET) の 3 つを算出した。Word Level Checker は、Automated Readability Index (ARI) や Coleman-Liau Index が算出されるが、Flesch-Kincaid Grade Level との相関係数が高く (染谷, 2009), 後の主成分分析でもこれらのリーダビリティ指標を追加しても、結果は大きく変わらないことを確認したため、今回は分析には含めなかった。

表 1

使用した各種指標のまとめ

指標	説明	算出プログラム
1 Total Words	総語数	
2 Words Per Sentence	一文に含まれる平均語数	MS Word
3 Flesch-Kincaid Grade Level	一文に含まれる平均語数と一語の平均音節数を基に計算	
4 Narrativity	物語や会話調で読み手が親しみのある英文かどうかを示す	
5 Syntactic Simplicity	文体が簡潔でわかりやすいかを示す	
6 Word Concreteness	語が具体的でイメージしやすいかを示す	
7 Referential Cohesion	文や句の結束性が英文全体でどの程度あるかを示す	Coh-Metrix
8 Deep Cohesion	因果・論理関係を示す接続語がどの程度含まれているかを示す	
9 Verb Cohesion	同じ動詞がどの程度繰り返し使われているかを示す	
10 Connectivity	反対, 追加, 比較などの関係がどの程度明確にわかるかを示す	
11 Temporality	時制などが文中でどの程度同じように使われているかを示す	
12 TTR (Type-Token Ratio)	総語数に対する異なり語の数で, 1 に近くなるほど異なり語が多く難しい	
13 Average Word Level (SVL)	SVL12000 の 12 レベルでの平均レベル	Word Level Checker
14 Average Word Level (JACET)	JACET8000 の 8 レベルでの平均レベル	

表 2

使用した指標間の相関係数

指標	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 Total Words	-													
2 Words Per Sentence	.27	-												
3 Flesch-Kincaid Grade Level	.11	.62	-											
4 Narrativity	.00	-.42	-.72	-										
5 Syntactic Simplicity	-.21	-.82	-.37	.09	-									
6 Word Concreteness	-.09	.11	-.30	.24	-.30	-								
7 Referential Cohesion	-.21	.34	-.27	.23	-.42	.59	-							
8 Deep Cohesion	.17	.13	.11	.05	-.05	.09	-.14	-						
9 Verb Cohesion	.21	.12	.24	-.71	.10	-.47	-.23	-.16	-					
10 Connectivity	-.51	.03	.19	-.32	-.14	-.24	.06	-.67	.24	-				
11 Temporality	.27	-.12	-.17	.52	.02	.10	-.13	.54	-.48	-.69	-			
12 TTR	-.46	-.09	.46	-.35	.17	-.17	-.51	.34	-.04	.09	-.01	-		
13 Average Word Level (SVL)	-.30	.43	.50	-.62	-.38	.24	.04	.06	.22	.28	-.24	.37	-	
14 Average Word Level (JACET)	.04	.17	.26	-.36	-.12	-.09	-.57	.18	.16	.06	-.05	.40	.52	-

2.3 分析方法

すべての分析は R version 2.15.2 を使用して行った。まず、分析の対象とした教科書の確定されている全 15 レッソンの配置順と、14 の各種指標との順位相関係数（スピアマンの順位相関係数）を確認した。これにより、もし、高い相関係数が得られた場合には、レッスンの配置とその指標との関連があると考えられる。また、今回の研究の目的は、14 の指標を対象として「テキストの難易度」を示すことであるため、指標の 1 つずつを検討していくことではない。そこで、たくさんの指標の持つ情報量を圧縮して、1 つの合計点のようなものを作成することができる主成分分析を用いた（主成分分析の詳細は、水本、2010 を参照）。また、主成分分析で 15 のレッスンがどのような特徴で分類できるかを示すために、クラスター分析も行った。

これらの分析を再現可能にするために、使用したデータと R のコードはこちらで公開している (<http://rpubs.com/mizumot/metho2012>)。

3. 結果

全 15 レッソンの配置順と 14 の各種指標との順位相関係数は表 3 のとおりである。もともと総語数が後のレッスンになるほど長くなるように作成されている教科書であるため、Total Words との関係が強いことがわかる。また、Flesch-Kincaid Grade Level や、Syntactic Simplicity, TTR などとも中程度の相関があるため、後のレッスンになるほどある程度難しくなる傾向があることが確認できる。

表 3

教科書のレッスン配置順と指標の順位相関係数

指標	順位相関
Total Words	.83
Words Per Sentence	.62
Flesch-Kincaid Grade Level	.31
Narrativity	-.20
Syntactic Simplicity	-.43
Word Concreteness	.01
Referential Cohesion	.01
Deep Cohesion	.11
Verb Cohesion	.29
Connectivity	-.18
Temporality	.16
TTR	-.50
Average Word Level (SVL)	-.06
Average Word Level (JACET)	-.08

次に、主成分分析の結果を図2に示す。最終的に7つの指標しか使われていないのは、14の指標を使い主成分分析をしたところ、第一主成分か第二主成分どちらかへの寄与率が低かった(0.3以下の)指標を除外して、再度、主成分分析を行ったためである。この結果から、第一主成分(x軸)の右側(プラス)に行くほど難易度では易しくなり、左側(マイナス)に行くほど難しいテキストであるということが推測できる。第一主成分で考えると、レッスンは1, 8, 10, 11, 2, 9, 5, 14, 3, 13, 4, 6, 7, 12, 15の順に難しくなると考えられる。また、第二主成分(y軸)の上側(プラス)には、Coh-MetrixのReferential CohesionやWord Concretenessがあるため、こちらに位置しているレッスンは、具体的で結束性の強いテキストであり、下側(マイナス)では、そのような特徴が見られないテキストを含むレッスンが位置しているのではないかとわかる。

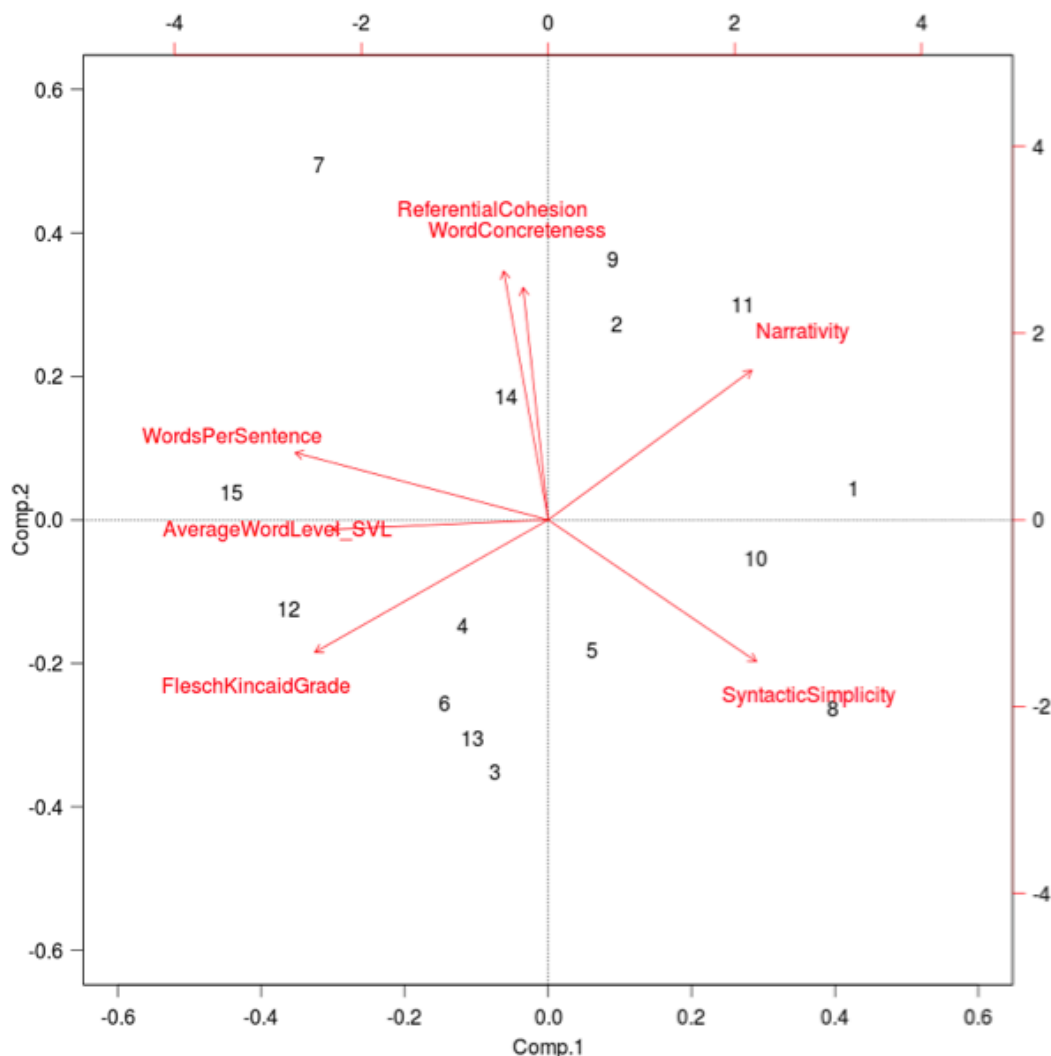


図2. 主成分分析の結果 (累積寄与率 74%)

上記の主成分分析の結果から得られた第一主成分得点と第二主成分得点を使い、クラスター分析（ワード法・ユークリッド距離）を行った結果を図3に示し、それぞれのレッスンの特性を書き加えた。基本的には、図2で近くに位置している変数が、クラスター分析でも近くに分類されることがわかるが、分析に使用した指標の特性を考えると、このようにレッスンごとの特徴を考察する材料となることがわかる。

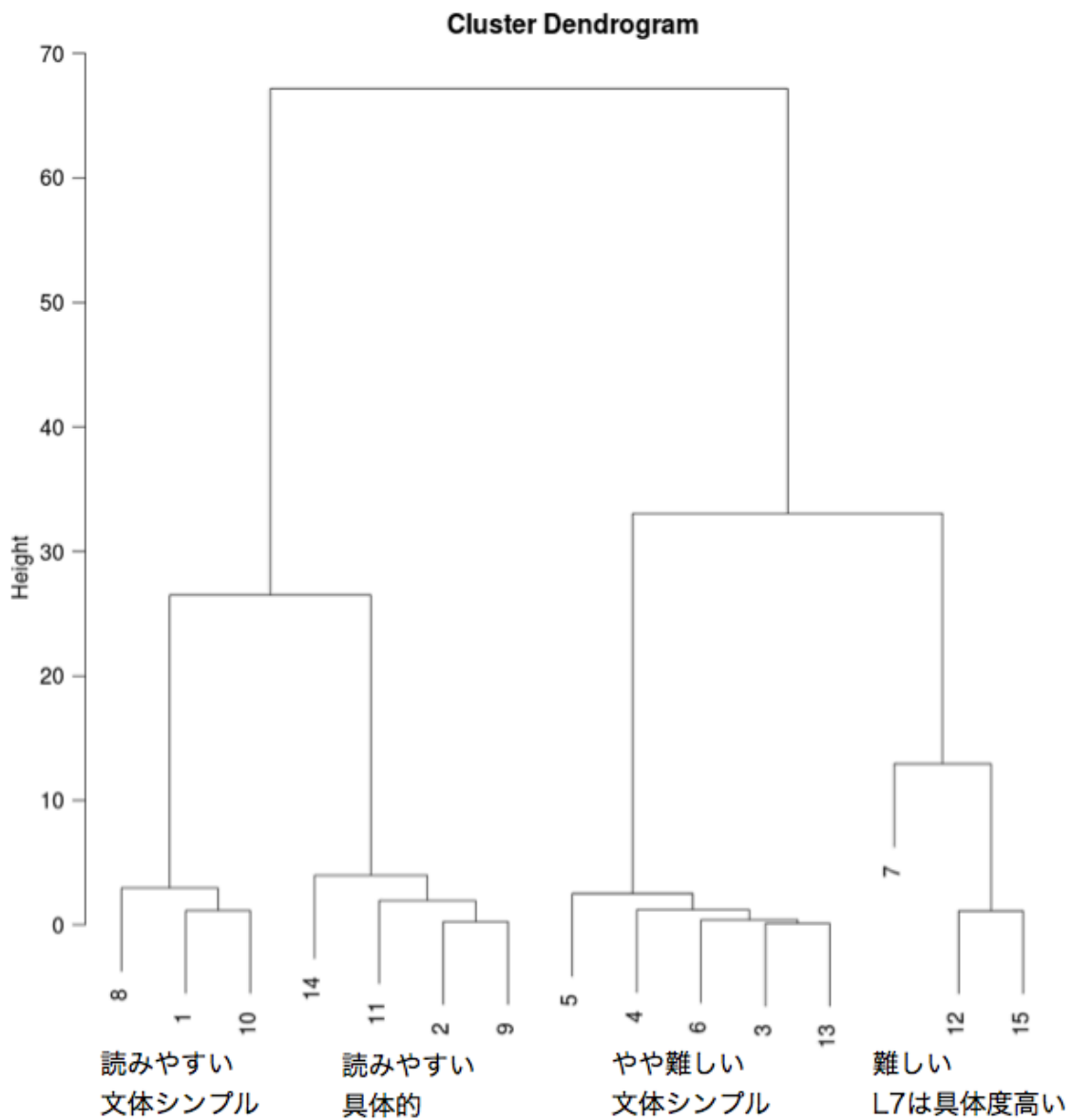


図3. クラスター分析の結果（ワード法・ユークリッド距離）

4. まとめ

本稿では、従来から使われているリーダビリティ指標に加えて、英文解析プログラムから得られる各種指標を使い、テキスト難易度の推定を試みた。その結果として、Coh-Metrixのように新しい指標を用いることにより、より精緻な分析が可能であることが示された。しかし、指標に何を含まかによって、異なる結果が得られるのが主成分分析であるため、今回の分析はあくまで探索的な方法であることに注意しなければならない。

実際に教材作成への適用を考えた場合には、指標の選択方法や、読み手の主観による難易度の判定との相関など、今後、検討しなければならないことは多いが、テキスト難易度の推定は、教材作成、テスト作成、実験（研究）の題材選択、そしてライティング・スピーキングのようなパフォーマンステストに自動採点を組み込むことを考えている者にとってみれば、大きな課題であるため、今後も研究が進むはずである。その過程で、本稿で示したように、得られた知見・データを教材作成などに適用できるかを検討していくことが、理論の応用が求められる外国語教育研究の分野では重要であると言えるだろう。

参考文献

- Crossley, S., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29, 243–263. doi:10.1177/0265532211419331
- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multi-level analyses of text characteristics. *Educational Researcher*, 40, 223–234. doi:10.3102/0013189X11413260
- 小林 雄一郎・金丸 敏幸 (2012). 「Coh-Metrix とパターン認識を用いた課題英作文の自動評価」『人文科学とコンピュータシンポジウム論文集—つながるデジタルアーカイブ』 259–266. Retrieved from <http://image02.wiki.livedoor.jp/l/t/langstat/c47ed88898a6f367.pdf>
- Koizumi, R., & In' nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40, 554–564. doi:10.1016/j.system.2012.10.012
- 水本 篤 (2008). 「自由英作文における語彙の統計指標と評定者の総合的評価の関係」『統計数理研究所共同研究レポート 215 学習者コーパスの解析に基づく客観的作文評価指標の検討』 15–28. Retrieved from <http://www.mizumot.com/files/essay-index.pdf>
- 水本 篤 (2010). 「主成分分析：データの情報を圧縮する」石川慎一郎・前田忠彦・山崎 誠 (編) 『言語研究のための統計入門』 (pp. 193–217). くろしお書店.

染谷泰正 (2009). 「オンライン版「英文語彙難易度解析プログラム」(Word Level Checker)の概要とその応用可能性について」『青山学院大学文学部紀要』51, 97-120. Retrieved from <http://someya-net.com/wlc/readability.pdf>