レベル分けされた市販の英語学習単語帳を使用した「単語の難易度」と「読みやすさ」によるディズニーの英語教材 15 冊のランク付け

寺田　捷人†　　加藤　直孝‡

†久留米高専 専攻科 機械・電気システム工学専攻
‡久留米高専 電気電子工学科

1. はじめに

英語学習者が効率的に英語を学習するためには、自分のレベルに合った英語教材を選定する必要がある。しかし、自分に適した英語教材を選ぶことは必ずしも容易ではない。評価の高い英語教材を選択しても、それが自分のレベルに合致しているとは限らない。また、英文の難易度を測定する際、単語の難易度と文章の読みやすさを同時に評価した英語教材のレベル分けは少ない。

本研究では、「単語の難易度」だけではなく、文の量や音節数(「読みやすさ」)をも考慮して、英語教材の難易度を判定する。特に、日本の英語学習者のために、レベル分けされた市販の英語学習単語帳を基準とした英語教材の難易度を分析する手法を提案する。この手法を用いることで、英語学習者が自分に適したレベルの単語が使われている英語教材を選定しやすくなる。

この手法には次に挙げるメリットがある。

1. あらかじめ英語学習単語帳のデータを準備しておくことで英語教材のテキストデータから簡単に英語教材の難易度を評価できる。
2. 日本の英語学習単語帳を使用することで日本人の学習に適した英語教材を選定できる。
3. 英検学習や TOEIC 学習といった目的に応じて測定に使用する英語学習単語帳を変えることで、その目的に適した英語教材を選定できる。

我々は、上記の手法を試すために、英語学習単語帳のデータを収集し、データベースを作成した。このデータベースをもとに本研究では、株式会社 KADOKAWA の「CD 付 Disney の英語」シリーズ 15 冊[1](以後、ディズニーの 15 冊と呼ぶ)の英語教材を、「単語の難易度」及び「読みやすさ」の 2 つの観点から分析し、ランク付けを行った。そして、英語学習者の教材選択を手助けできるように、ランク付けの結果を二次元テーブルに示した。結果である二次元テーブルを 5 の Figure4 に示す。使用した英語学習単語帳は 3 で述べる。

2. 関連研究

既存の英語の難易度を測定する方法として、Flesch-Kincaid が開発した「Readability Score」[2]と MetaMetrics 社が開発した「Lexile Measure」[3]が挙げられる。

Flesch-Kincaid による「Readability Score」は、各英文の長さと各単語の音節数をもとに、英文の読みやすさを測定する指標である。そのための計算式は次(右上)のように示され、Readability Score の値が 0 に近づくにつれて難解な英文であることを表す。

Fresch の Readability Score = 206.835-1.015*(一文当たりの単語数)
-84.6*(一単語当たりの音節数)

この方法は、英文の読みやすさを評価するための方法として優れている。しかし、単語の難易度によらずスコアが決まるため、語彙の限られた英語学習者にとっては必ずしも役立たない。

MetaMetrics 社による「Lexile Measure」は、英文の読みやすさに加えて、英単語の難易度をもとにして測定される。このスコアは通販サイト「Amazon」でも英語の本の難易度を示すために使われており、世界的に利用されている手法といえる。しかし、世界共通の指標であるために、このスコアが日本人学習者の語彙や目的に応じた教材の選択に適するとは限らない。また、スコアの計算方法は公開されておらず、測定のためには MetaMetrics 社に依頼するしか方法がないといった難点がある。

3. 手法

英語教材の難易度を測定するための手法を説明する。難易度測定の際には、「単語の難易度」を第一の尺度とする。そして、それらの評価が近いものを「読みやすさ」の観点からランク付けを行う。「読みやすさ」の評価には、Flesch-Kincaid による「Readability Score」による「英文の読みやすさ」及び「文章全体の長さ」を用いた。文章全体が長くなると、英語学習者が途中で挫折し学習の継続を妨げる可能性があるので、文章全体の長さをも考慮することにした。

英語教材にはディズニーの 15 冊を用いた。3.1 で研究に使用したデータベースの概要について説明し、3.2 で各分析手法について説明する。

3.1. データ収集用データベースの概要

本研究におけるデータベースは worddata データベースと呼び、2 つのテーブルで構成する。1 つはディズニーの 15 冊の英語教材の disney テーブルで、もう 1 つは英語学習単語帳の words テーブルである。次に、それぞれのテーブルの概要を説明する。

disney テーブルは、英語教材のテーブルである。ディズニーの 15 冊を OCR にかけた後、プログラムで処理し準備した。disney テーブルのデータ名と日本語のタイトルのリストを Table1 に示す。disney テーブルには、Apache OpenNLP と Word Net を用いてそれぞれの単語の原形と品詞の他、その単語はどの英語教材の何段落目の何文目の何単語目に記されているかまで細かくデータを収集している。disney テーブルのフィールドの一覧を

Table 1. disney テーブルのデータ名とタイトル一覧

データ名	日本語名
D01_Pooh	くまのプーさん
D02_Nemo	ファインディング・ニモ 他2話
D03_TOYSTORY	トイ・ストーリー
D04_SleepingBea	眠れる森の美女
D05_Frozen	アナと雪の女王
D06_Mickey	ミッキーマウス
D07_Cinderella	シンデレラ
D08_Beautyandth	美女と野獣
D09_BigHero6	ベイマックス
D10_INSIDEHEAD	インサイド・ヘッド
D11_Frozen_Coll	アナと雪の女王 ショートストーリー
D12_THELITTLEME	リトル・マーメイド
D13_Aladdin	アラジン
D14_ZOOTOPIA	ズートピア
D15_FindingDory	ファインディング・ドリー

Table 3. words テーブルのフィールド一覧

フィールド名(列名)	データ型	備考
textbook	varchar(8)	単語が記載されている媒体
page	int	単語が記載されているページ
id	varchar(5)	媒体ごとに単語に振られた番号
chapter	text	単語が記載されている章
word	text	単語
sentence	text	単語に記載されている例文

Table2 に示す。

words テーブルは、英語学習単語帳の単語データを収集したテーブルである。本研究で用いた英語学習単語帳は株式会社ジャパンタイムズの「出る順で最短合格！EX」シリーズ[4]と株式会社旺文社の「出る順パス単」シリーズ[5]の 2 種類である。これらの英語学習用単語帳はともに英検 1 級、準 1 級、2 級、準 2 級、3 級、4 級、5 級の 7 つのレベルに分けられている。words テーブルには、英語学習単語帳のそれぞれの単語の記載ページと単語番号、例文のデータを収集している。words テーブルのフィールドの一覧を Table3 に示す。

worddata データベースのデータベース管理システム(DBMS)には PostgreSQL を使用し、データを入出力する際には、JDBC(Java Database Component)を使用した。入出力プログラムの言語には Java を使用した。

3.2. 英語教材の分析手法

3.2.1.「単語の難易度」の分析手法

各ディズニーの英語教材に対して、各レベル(英検 1~5 級)の英語学習単語帳の単語がどれだけ使用されているかをもとに、それらの教材のランク付けを行った。具体的には、作成したデータベースと分析用のプログラムを使用し、英語教材の各単語が英検学習単語帳のどのレベルの単語なのかを調べ、英語教材の全単語をレベル分けした。その際、 ”the” や ”I” 等のあまりにも簡単な単語は最もレベルの低い英語学習単語帳でもカバーされておらず、それが測定結果に大きな影響を与えることがあった。それを解決する手段として、「JACET8000」[6][1]の上位 200 単語は「基本単語」として分類し、予め英検学習用単語帳による分類の前に除外した。同様に、固有名詞も英検学習用単語帳に記載されていないものが多く、分析結果に大きな影響を与えてしまうため、固有名詞として分類し予め除外した。

全単語を分類した後、英語教材の各レベルの単語の比率をもとに英語教材全体の単語の難易度を決定した。難易度決定は、簡単なレベルに分類された単語には低い点数を、難しいレベルに分類された単語には高い点数をつけ、全体の平均値を求めることで行った。

Table 2. disney テーブルのフィールド一覧

フィールド名(列名)	データ型	備考
word_number	integer	単語の通し番号
word	text	単語
word_class	text	単語の品詞
base_form	text	単語の原形
book_id	text	単語が記載されている教材
story	text	単語が記載されている話 (話を複数収録している場合使用)
chapter	text	単語が記載されている章
detect_page	text	単語が記載されているページ
paragraph_number_ofpage	integer	単語が記載されている段落がそのページの何段落目か
sentence_number_ofparagraph	integer	単語が記載されている文がその段落の何文目か
sentence_number_ofpage	integer	単語が記載されている文がそのページの何文目か
sentence_number_ofbook	integer	単語が記載されている文がその教材の何文目の文か
word_number_ofsentence	integer	その単語はその文の何単語目か
word_number_ofbook	integer	その単語はその教材の何単語目か
sentence	text	単語が記載されている文

3.2.2.「文章全体の長さ」の分析手法

英語教材の文章全体の長さを評価するために、前述の disney テーブルを用いて各英語教材の単語の全体数を数えた。その際、disney テーブルには、「,」(コンマ)や「.」(ピリオド)といった記号も一つの単語として登録されているため、プログラムで単語数を数える際に除外し、正確な単語数を数えることができるようにした。

3.2.3.「英文の読みやすさ」の分析手法

英語教材の英文の読みやすさを評価するための指標として、2 で説明した Flesch-Kincaid による「Readability Score」を使用した。Readability Score を算出するためには、英語教材の「一文当たりの単語数」及び「一単語当たりの音節数」を求める必要がある。

まず、一文当たりの単語数は disney テーブルの word_numberofsentence フィールドとして disney テーブルに収集済みのデータから求めることができる。

次に、一単語当たりの音節数の求め方を説明する。英単語の音節数は「一音節＝一つの母音(a,e,i,o,u)」が基本とされているが、実際には幾つものルールに基づいて決定される。そのため、それらのルールからできるだけ正確な単語の音節数を求めるためのプログラムを開発する必要がある。本研究では、音節数を正確に求めるために 2 つのプログラムを開発した。次に、それぞれのプログラムを説明する。

[1] 大学英語教育学会が作成した日本人向けの英単語の基本語彙リスト。使用頻度が高い順に 1～8000 までの番号がふられている。

1 つ目のプログラムは、単純に単語の母音(a,e,i,o,u)の数を数えるわけではなく、母音として働く文字を選定して数えるプログラムである。以下にこのプログラムで使用する母音選定のルールの一覧を示す[7]。

- 単語最初以外の y は母音として働く。
- 単語末の e は母音として働かない。ただし、e の前が「子音+l」の場合は母音として働く。
- 単語末の ed は母音として働かない。ただし、ed の前が「t」、「d」、「ch」の場合は母音として働く。
- 母音が 2 つ連続している場合、合わせて 1 つの母音として働くことがある。
- 「-」(ハイフン)が単語に使われている場合、「-」の前後それぞれの母音数を数えて足し合わせる。

2 つ目のプログラムは 1 つ目のプログラムを改善したものである。はじめに、作成した英単語の接頭辞・接尾辞のリストを基に単語をいくつかの部分に分解し、それぞれの部分の音節数を 1 つ目のプログラムと同様のプロセスで数えることで、より正確な音節数の値を得ることができるようになっている。ディズニーの英語教材の冒頭部分を使用して行った各プログラムの精度テストの結果を Table4 に示す。

Table 4. 音節数 精度テスト結果

	母音数カウントプログラム	1つ目のプログラム	2つ目のプログラム
教材の単語数	172	172	172
正当数	108	167	171
正当率[%]	62.790698	97.093023	99.418605

Table4 より、音節数を求めるプログラムの精度が母音数のみをカウントするプログラムでは、約 63%であったものが、開発した 2 つ目のプログラムでは約 99%まで改善されていることが確認できる。本研究では、この 2 つ目のプログラムを使用して各教材の音節数を測定した。

4. 分析

今回の分析では英検準 2 級レベル以上の単語に注目し、単語の難易度を後述する方法 1 と方法 2 の 2 つの方法で設定して、各英語教材の単語難易度スコアを算出した。ただし、単語帳に出現しない単語と 3.2.1 で述べた「基本単語」及び固有名詞は予め除外した。これらの単語は、単語帳をもとにしたレベル分けには影響は少ないからである。

この分析では「読みやすさ」を測定するために「読みにくさ」の指数を導入する。Table5 の「読みにくさ指数」とは、「文章全体の長さ」及び「英文の読みやすさ」の値の逆数のそれぞれの平均値を 100 に換算し、足し合わせて 2 で割ったものである。Fresch-Kincaid の Readability Score は値が 0 に近づくにつれて難解な文であることを示すため、難解であるほど値が大きくなるように逆数にした。この「読みにくさ指数」は、英文全体の読みにくさを求めたものであり、値が大きいほど読みにくい教材(難易度が高い)であることを示す。方法 1 と方法 2 の分析方法を説明する。

方法 1 は、英検準 2 級以上のレベルの単語すべてに対して 1 点をつけて、単語の難易度を求めたものである。このスコアが低いものであれば、英検 3 級程度レベルの英語学習者にとって読みやすい英語教材ということになる。

方法 2 は、英検準 2 級のレベルの単語に 1 点、2 級レベルの単語に 2 点、準 1 級レベルの単語に 3 点、1 級レベルの単語に 4 点をつけて、全体の単語の難易度を求めたものである。このスコアが高くなるにつれて、難しい単語が多くなり、よりレベルの高い英語教材であると判断する。

5. 結果

各英語教材の「単語の難易度」、「文章全体の長さ」、「英文の読みやすさ」の分析結果についてまとめたものを Table5 に示す。Table5 の各英語学習単語帳の方法 1、方法 2 に対する単語難易度スコアの値を基に、15 冊の英語教材の単語難易度スコアの平均値を 100 に換算し、各英語教材の単語の難易度及び 3.2.3 で述べた Readability Score と読みにくさ指数の分析結果をグラフにしたものを Figure1 に示す。

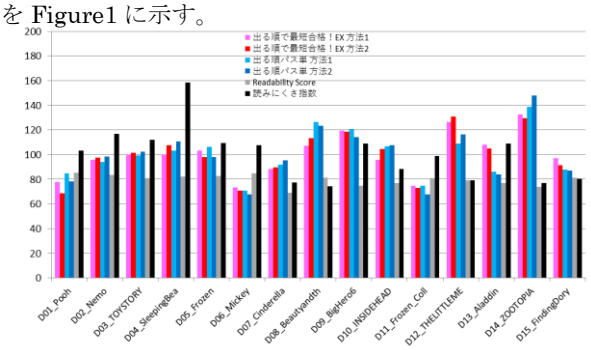

Figure 1. 英語教材の「単語の難易度」の比較と「読みにくさ指数」の比較

Table5 英語教材の「単語の難易度」、「文章全体の長さ」、「英文の読みやすさ」 の分析結果

教材名	総単語数	ジャパンタイムズ「出る順で最短合格！EX」シリーズ										旺文社「出る順パス単」シリーズ										readability score	読みにくさ指数
		準2級レベル		2級レベル		準1級レベル		1級レベル		単語難易度スコア		準2級レベル		2級レベル		準1級レベル		1級レベル		単語難易度スコア			
		単語数	単語比率[%]	単語数	単語比率[%]	単語数	単語比率[%]	単語数	単語比率[%]	方法1	方法2	単語数	単語比率[%]	単語数	単語比率[%]	単語数	単語比率[%]	単語数	単語比率[%]	方法1	方法2		
D01_Pooh	6168	213	3.4533	77	1.2484	27	0.4377	15	0.2432	0.0538	0.0824	186	3.0156	49	0.7944	46	0.7458	8	0.1297	0.0469	0.0736	85.3687	103.1488
D02_Nemo	7532	276	3.6644	94	1.2480	100	1.3277	30	0.3983	0.0664	0.1174	194	2.5757	108	1.4339	69	0.9161	20	0.2655	0.0519	0.0925	83.5073	116.7253
D03_TOYSTORY	6855	270	3.9387	84	1.2254	81	1.1816	39	0.5689	0.0691	0.1221	197	2.8738	93	1.3567	67	0.9774	19	0.2772	0.0549	0.0963	81.0778	111.9230
D04_SleepingBea	11980	440	3.6728	153	1.2771	146	1.2187	91	0.7596	0.0693	0.1292	348	2.9048	168	1.4023	99	0.8264	67	0.5593	0.0569	0.1043	82.1440	158.4111
D05_Frozen	6685	287	4.2932	102	1.5258	54	0.8078	34	0.5086	0.0714	0.1180	240	3.5901	97	1.4510	41	0.6133	15	0.2244	0.0588	0.0923	82.8207	109.3310
D06_Mickey	6610	199	3.0106	68	1.0287	50	0.7564	19	0.2874	0.0508	0.0849	149	2.2542	63	0.9531	37	0.5598	9	0.1362	0.0390	0.0638	84.6154	107.6262
D07_Cinderella	2172	64	2.9466	41	1.8877	24	1.1050	4	0.1842	0.0612	0.1077	54	2.4862	31	1.4273	21	0.9669	4	0.1842	0.0506	0.0898	69.2203	77.2399
D08_Beautyandth	2765	108	3.9060	44	1.5913	31	1.1212	22	0.7957	0.0741	0.1363	110	3.9783	49	1.7722	23	0.8318	11	0.3978	0.0698	0.1161	81.1883	74.2503
D09_BigHero6	6054	274	4.5259	108	1.7839	100	1.6518	18	0.2973	0.0826	0.1424	226	3.7331	121	1.9987	45	0.7433	12	0.1982	0.0667	0.1075	74.7064	108.7280
D10_INSIDEHEAD	3994	116	2.9044	83	2.0781	43	1.0766	23	0.5759	0.0663	0.1259	125	3.1297	63	1.5774	34	0.8513	13	0.3255	0.0588	0.1014	76.8640	88.2975
D11_Frozen_Coll	5428	157	2.8924	71	1.3080	31	0.5711	21	0.3869	0.0516	0.0877	133	2.4503	66	1.2159	20	0.3685	5	0.0921	0.0413	0.0636	80.6779	99.0446
D12_THELITTLEME	3157	145	4.5930	67	2.1223	38	1.2037	26	0.8236	0.0874	0.1574	92	2.9142	56	1.7738	26	0.8236	16	0.5068	0.0602	0.1096	79.0053	79.2038
D13_Aladdin	6216	254	4.0862	122	1.9627	64	1.0296	24	0.3861	0.0746	0.1264	156	2.5097	92	1.4801	40	0.6435	8	0.1287	0.0476	0.0792	76.8543	108.7345
D14_ZOOTOPIA	2533	133	5.2507	48	1.8950	38	1.5002	13	0.5132	0.0916	0.1559	98	3.8689	52	2.0529	26	1.0265	18	0.7106	0.0766	0.1390	73.6605	77.1070
D15_FindingDory	3420	138	4.0351	49	1.4327	32	0.9357	11	0.3216	0.0673	0.1099	92	2.6901	42	1.2281	24	0.7018	8	0.2339	0.0485	0.0819	81.2612	80.2290

英語教材の単語の難易度の分析結果に有意な差が認められることを確かめるために、t 検定を行った。Figure2、Figure3 は、それぞれ「出る順で最短合格！EX」シリーズ、「出る順パス単」シリーズを使用し、分析方法 2 で求めた単語難易度スコアにおいて、1%水準または 5%水準で有意な単語難易度の差が認められるかを確かめるために、各英語教材で相互に有意差検定を行った結果である。

Figure2、Figure3 の見方を説明する。縦軸(縦方向)が難易度を測定した英語教材を、横軸(横方向)が有意差を調べるための比較・検定に使用した英語教材を示している。縦軸から英語教材を選び、検定結果を横に見ていった際に青色が多ければ他の英語教材と比べて難しい英語教材であり、赤色が多ければ易しい英語教材であると判断できる。

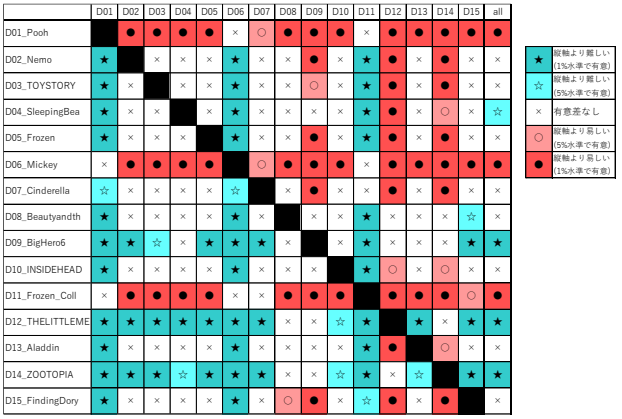

Figure 2. 「出る順で最短合格！EX」シリーズ
分析方法 2 による単語難易度スコアの
有意差検定結果

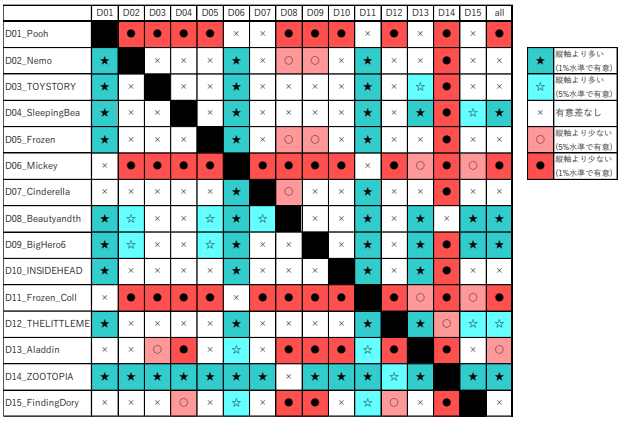

Figure 3. 「出る順パス単」シリーズ
分析方法 2 による単語難易度スコアの
有意差検定結果

6. 考察

Figure2 の単語難易度スコアの有意差検定の結果から、「 D14_ZOOTOPIA 」、「 D12_THELITTLEME 」、「D09_BigHero6」の 3 冊は特に単語の難易度が高く、「D01_Pooh」、「D06_Mickey」、「D11_Frozen_Coll」の 3 冊は単語の難易度が低いということが読み取れる。Figure3 の検定結果からも同様の傾向が見て取れることから、これらの 6 冊の英語教材は他の英語教材と比較して、単語の難易度に大きな差があると判断できる。

さらに、Table5 の読みにくさ指数の値を追加した分類を示す。単語難易度スコアの有意差検定の結果から、単語の難易度が特に難しくも易しくもないと判断された 9 冊の英語教材をもとに読みにくさ指数を 3 つに分類した。「D04_SleepingBea」は、単語のレベルも高く、読みにくさ指数が 158.5 と非常に大きな値となっているため、「難しい」に分類した。それ以外の 8 冊は、読みにくさ指数が 100 程度またはそれ以上のものを「普通」、90 程度以下のものを「易しい」に分類した。

分類結果を Figure4 に示す。以上のようにして、15 冊の英語教材を「単語の難易度」及び「読みやすさ」に応じて二次元テーブル上に分類することができた。このテーブルの使用例として、英語学習者は、単語の難易度が「易しい」の 3 つの英語教材は非常にレベルが低いのでスキップし、「普通」のものを「読みやすさ」が「易しい」ものから順に読んでいくといった判断が可能になる。

		単語の難易度					
		易しい	読みにくさ指数	普通	読みにくさ指数	難しい	読みにくさ指数
読みやすさ	難しい			D04_SleepingBea	158.4		
	普通	D06_Mickey	107.6	D02_Nemo	116.7	D09_BigHero6	108.7
		D01_Pooh	103.1	D03_TOYSTORY	111.9		
		D11_Frozen_Coll	99.0	D05_Frozen	109.3		
				D13_Aladdin	108.7		
	易しい			D10_INSIDEHEAD	88.3	D12_THELITTLEME	79.2
				D15_FindingDory	80.2	D14_ZOOTOPIA	77.1
				D07_Cinderella	77.2		
				D08_Beautyandth	74.3		

Figure 4. 「CD 付 Disney の英語」シリーズの
「単語の難易度」及び「読みやすさ」による分類結果

7. おわりに

本研究では、レベル分けされた市販の英語学習単語帳を使用した「単語の難易度」と、文章全体の長さや単語の音節数から計算した「読みやすさ」の 2 つの観点からの分析によって、ディズニーの 15 冊の英語教材の難易度の相互評価を行い、ランク付けを行った。今後、英語教材や英語学習単語帳の数を増やしていくことで、英語教師や英語学習者が自分のニーズに合った英語教材を選定することを手助けすることが期待できる。

今回は、OCR を用いてテキストファイル化された英語教材の英文データをデータベースに入力するためのソフトウェアを開発したが、今後は、英語教材と英語学習単語帳を指定して英語教材の難易度を簡単に分析し結果を示すことができるソフトウェアの開発を進める。多くのデータを集めた後、より簡単に難易度評価ができるシステムを開発し、実用化を目指す。

参考文献

[1] 石原真弓(2016)「CD 付 Disney の英語」シリーズ 15 冊, 株式会社 KADOKAWA

[2] Rudolf Flesch(2017)『How to Write Plain English』, <https://web.archive.org/web/20171115193605/http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml> (参照 2018-9-25)

[3] MetaMetrics (2017)『Understanding Lexile® Measures』, <https://lexile.com/educators/understanding-lexile-measures/> (参照 2018-9-25)

[4] 小笠原敏晶(2016),『出る順で最短合格！EX』シリーズ, 株式会社ジャパンタイムズ

[5] 生駒大壱(2013),『出る順パス単』シリーズ,株式会社旺文社

[6] 大学英語教育学会基本語改訂特別委員会(2016),『大学教育学会基本語リスト 新 JACET8000』,桐原書店

[7] 『英語の「シラブル」(音節) を理解するための基礎知識 』<https://eikaiwa.weblio.jp/column/study/pronunciation/syllable-basic-rules>(参照 2018-12-17)