全教科を収録対象とした日本語教科書コーパスの構築

松吉　俊[†,‡]　　近藤　陽介[‡]
橋口　千尋[‡]　　佐藤　理史[‡]

[†] 京都大学大学院 情報学研究科, [‡] 名古屋大学大学院 工学研究科

1. はじめに

テキストの難易度を自動的に推定する技術は、そのテキストが、想定する読者に相応しいものであるかどうかを知る助けとなる。これまで提案されてきた難易度推定手法は、大きく次の 2 種類に分類できる。

(1) 語長や文長など、いくつかの特徴量の値を公式に代入して難易度を計算する方法（公式法）
(2) 複数の言語モデルに対する尤度を計算し、分類問題として難易度を計算する方法（言語モデル法）[1)]

われわれは、昨年、言語モデル (文字ユニグラム) を用いて日本語テキストの難易度を推定する方法を提案した[2)]。この方法を用いて、中学と高校の社会科の教科書および新聞の社説の 3 種類のテキストを規準コーパスとして言語モデルを構築し、中学・高校・一般の 3 段階の難易度推定を行なうシステムを実現した。このシステムは、規準コーパスと同じ分野のテキスト (社会科に関するテキスト) に対しては、高い精度で難易度推定を行なうことができたが、分野の異なるテキストに対しては、十分な性能が得られなかった。われわれは、その原因を、規準コーパスが内容的に狭く偏っており、難易度という観点の類似性よりも、内容の類似性が強く現れてしまうためと考えた。

この問題を解決するために、われわれは、より広い分野をカバーする規準コーパスとして、全教科を対象とした日本語教科書コーパスを構築した。本論文では、このコーパスとその構築法について報告する。

2. 難易度推定の規準コーパス

公式法および言語モデル法のどちらの方法を採用する場合も、公式のパラメーターまたは言語モデルを推定するためのコーパス（規準コーパス）が必要である。このコーパスは、次の要件を満たす必要がある。

要件 1　コーパスを構成するそれぞれのテキストに、難易度が付与されていること
要件 2　コーパスは、色々なレベルの難易度のテキストを含んでいること

これに加えて、昨年の研究の結果より、次の要件を追加する。

要件 3　コーパスは、さまざまな分野のテキストを含んでいること

このような条件を満たすコーパスのテキスト収集源として、われわれは、教科書に着目した。すなわち、

(1) ほとんどの教科書は、使用する学年が明示されている。この学年を難易度とみなせば、要件 1 を満たすことになる。
(2) 小・中・高の教科書は、文部科学省の検定を受けており、その内容・記述が統制されている。このことにより、学年と記述の難易度に強い相関があることが期待できる。
(3) 小・中・高の教科書により、12 段階の難易度を設定できる。
(4) 小・中・高の教科は多岐に渡っており、これらをすべて網羅すれば、要件 3 を満たすことが期待できる。

3. コーパス構築の実際

実際にコーパスを構築する作業過程において、さまざまな問題に直面した。ここでは、これらの問題とそれらに対してわれわれがとった対策について述べる。

3.1 小・中・高の教科書の選定

われわれが比較的簡単に入手できる教科書を用いた。具体的には、小・中・高のすべての学年のすべての教科（科目）に対して、愛知県名古屋市で採用されている教科書を 1 冊ずつ入手し使用した。例えば、小 1 算数の教科書としては啓林館の「わくわくさんすう 1」を、中 1 国語の教科書としては教育出版の「伝え合う言葉 中学国語 1」を、高校世界史 B の教科書としては山川出版社の「詳説世界史 B 改訂版」を入手した。これらの教科書のうち、英語科を除く全教科（国語科、社会科、数学科、理科、音楽・美術・書道などの芸術科、保健・情報を含む技術家庭科）の合計 111 冊をコーパス作成に用いた。

3.2 大学の教科書の選定

小 1 の教科書に含まれる文字は、そのほとんどがひらがなであり、これよりやさしいレベルを導入することは不要であると判断した。その一方で、高 3 の教科書より難しいテキストは実際に存在すると考えられるため、本研究では、小・中・高の 12 段階に加えて、高 3 より難しいレベルを導入することとした。そのレベルの規準テキストとして、昨年用いた新聞記事 (社説) のような一般的

表 1 教科書の冊数

	小学校	中学校	高等学校	大学	計
本研究	53	25	33	16	127
全数 [4)]	293	134	979	–	1,406

なテキストを利用する方法も考えられるが、教科書というカテゴリーを重視し、大学の教養課程の教科書を利用することとした。

大学の教科書の選択には、次の 2 つの問題がある。

(1) 小・中・高のように、網羅的な教科書リストが存在しない

(2) 専門化が進むため、科目数が非常に多い

これらの問題に対して、われわれは次のように対処した。

(1) 京都大学もしくは名古屋大学において、主に 1, 2 年生を対象とした講義で用いられている指定教科書・指定参考文献のリストの中から、サンプルとして相応しいものを選択する

(2) 高等学校における各教科各科目 (全 16 科目) に対して、それに相当する分野の講義のみを対象とし、それぞれに対して 1 冊ずつ、計 16 冊を選択する

この方針に従い、例えば、国語に対しては「認知言語学原理」(山梨正明) を、数学に対しては「基礎課程 線形代数」(吉野雄二) を選択した。

表 1 に、本コーパスのテキスト収集源として使用した教科書の分布を示す。なお、この表の「全数」は、平成 19 年度検定済み教科書の総数である。

3.3 抽 出 単 位

一般に、コーパスは大きければ大きいほど良いので、選択したすべての教科書の全文を電子化することが望ましい。その一方で、テキストの電子化には、その量に応じた時間と費用がかかる。そのため、われわれは、あらかじめ定めた抽出単位と抽出数に基づいて、教科書からその一部分 (サンプル) を抽出することにした。

抽出単位としては、学年・教科を通して一定量の大きさのテキストが望ましい。この方針に基づき、当初は、抽出単位を、一律 1,000 から 1,500 字程度の文字列と定めた。しかしながら、小 1、小 2 など低学年の教科書や、芸術科などの教科書においては、写真や図などのビジュアルエイドが多いため、それほどまとまった量のテキストを抽出することができないことが分かった。このような理由により、本研究では、教科書のさまざまな記述形式を考慮して、以下に示す 6 種類の抽出単位を採用した。

千五百字 1,000 から 1,500 字程度のテキスト

見出し 大半の教科書においては、章・節の下に話題ごとに見出しが立てられている。この 1 つの見出しのもとに記述された 1, 2 ページ程度のテキスト

見開き 教科書の見開き (2 ページ) に記述される、切りのよい段落で区切ったテキスト。1 つの話題に対する長い文章から、その一部分を抽出するときに用いた

数段落 数段落分のテキスト。判が大きく字が小さい教科書において、「見開き」の代わりに用いた

表 2 各抽出単位を採用した教科書の数と例

抽出単位	冊数	例
千五百字	7	中学社会、中学技術、中学家庭
見出し	52	中学理科、高校社会、高校理科
見開き	35	中学国語、高校国語、大学国語
数段落	2	大学地学、大学音楽
章	27	小学理科、小学音楽、中学美術
作品	4	小 1 国語上・下、小 2 国語上・下

章 1 つの章に含まれるすべてのテキスト。写真や図などが多く、地の文が少ない教科書に対して用いた

作品 1 つの作品に含まれるすべてのテキスト。作品あたりの文字数が非常に少ない教科書に対して用いた

それぞれの抽出単位を採用した教科書の数と例を表 2 に示す。本コーパスでは、抽出単位として、主に「見出し」と「見開き」を用い、低学年の教科書や芸術科の教科書に対して、適宜、「章」や「作品」を用いた。

3.4 抽 出 数

抽出単位と同様に、抽出数も学年・教科を通して一定であることが望ましい。この方針に基づき、当初は、抽出数を一律 10 と定めた。しかしながら、先ほどと同様の理由により、抽出数を固定することは難しいことが分かった。そこで、本研究では、原則として、抽出数を (総ページ数)/10 と定め、その最大値を 20 に設定した。これは、教科書内の文字全体の 1 割を抽出の目安とし、かつ、コーパス全体に占める、ページ数が多い教科書のテキストの割合が大きくなりすぎないようにするためである。

教科書の中には、複数の学年を通して利用されるものがある。例えば、中学歴史の教科書は中 1 と中 2 で、小学家庭科の教科書は小 5 と小 6 で利用される [4)]。このような教科書に対しては、抽出したテキスト集合のうち、開始ページが若い半数に低いほうの学年を、残りの半数に高いほうの学年を割り当てた。

3.5 テキストの記述形式

本研究では、次の 2 種類の形式を採用した。

(1) **“as is”**
原文の改行箇所に従って改行を入れる

(2) **一行一文**
句点などの文区切りに従って改行を入れる

当初は、前者の形式のみの予定であったが、「1 文内の平均語数」のような、文という単位を考慮した統計情報を用いる難易度推定手法との比較を可能とするために、後者の形式を別ファイルとして作成した。

なお、文字はすべて全角で入力した。

3.6 ル ビ

難解な漢字にルビ (ふりがな) が付いているかどうかは、そのテキストの難易度を決定する重要な要素の一つであり、教科書コーパスにこの情報を記載することは有用であると考えられる。コーパス構築当初、われわれは、ルビを SGML に則ったタグで記述していた。しかしながら、小学校の教科書や社会科の教科書においては、ルビの使用がかなり多く、予想以上の作業時間がかかることが分

かった。このような理由により、今回のコーパス構築作業においては、ルビ情報の付与を断念した。

3.7 太字、下線、傍点、下付き、上付き

教科書においては、ルビ以外の文字装飾要素として、主に、太字、下線、傍点、下付き、上付きが使用される。当初は、ルビと同様に、これらの要素を SGML に則ったタグで記述していた。しかしながら、教科書では、重要語句を強調するための太字、化学の分子式 (例えば、H_2O) 中の下付き、数式 (例えば、x^3+2) における上付きの出現率が非常に高く、これらの電子化にはかなりの時間がかかることが分かった。文字装飾要素の有無は、テキストの難易度の決定にほとんど影響を与えないと思われる。これらの理由により、本コーパスでは、文字装飾要素に関する記述を省略した。

3.8 入力できない漢字と記号

入力作業とその後の利用環境を考慮して、テキストはシフト JIS で電子化した。この文字コードに存在しない漢字については、その読みをひらがな、もしくはカタカナ (中国の人名や地名) で入力した。$\notin$ や $\subseteq$ など、シフト JIS に存在しない記号については、その箇所を空白とした。

3.9 除 外 箇 所

次の箇所は、抽出対象から除外した。

絵や写真とその説明、独立した数式、欄外の表、
目次、索引、奥付

3.10 構 築 手 順

各教科書に対して、以下の手続きでサンプルを選び、電子化した。

(1) 教科書を k (= (総ページ数)/(抽出数)) ページごとに分割する
(2) それぞれの断片から、抽出対象とする部分を決定する。この決定においては、3.9 節で述べた除外箇所を含む量が最も少ないものを優先する
(3) その部分のテキストを OCR ソフトを用いて電子化する
(4) 人手で OCR ソフトの認識誤りを修正する (→ “as is” 形式)
(5) 自動的に一行一文に変換する
(6) 人手で文区切りを修正する (→ 一行一文形式)

コーパス構築の実際の作業は、1 人の作業者がおよそ半年をかけて行なった。

4. 現状とまとめ

構築した教科書コーパスの概要を表 3 に示す。本コーパスのテキスト数は 1,478、総文字数は 1,073,263 字である。なお、表 3 には、昨年われわれが使用したコーパス、および、最近の 2 つの難易度推定のための規準コーパスの概要も同時に示した。

本コーパスにおける学年別・教科別統計量を表 4 に示す。この表の「常用漢字」は、学習漢字以外の常用漢字を指す。

表 4 より、次の 3 つのことが分かる。

(1) テキスト数と文字数において、学年間に大きな差がある。これは、各学年において利用した教科書の数と、それぞれの教科書に対して選択した抽出単位が異なることによる。
(2) 学年が上がるにつれて、テキストのひらがな含有率が減少し、代わりに漢字含有率が増加する。しかし、学年を固定して見ると、それらの値は、教科ごとにばらついている。例えば、中学のテキストにおけるひらがな含有率は、国語科はおよそ 61%、社会科はおよそ 50%、数学科はおよそ 44%である。したがって、単純にひらがな含有率や漢字含有率を用いて難易度を推定する手法には、高い性能を期待することができないと思われる。
(3) 全体のひらがな含有率は、高 3 まで単調に減少していくが、大学で増加に転じる。学習漢字含有率も、高 3 まで増加傾向にあるが、大学で著しく減少する。本研究では、高 3 より難しいレベルのテキストとして、大学の教科書から 3.2 節で述べた方法によって収集したテキストを用いた。しかしながら、このテキストは、小・中・高の教科書の延長として期待される性質を持っていなかったと言える。

本研究で構築したコーパスを用いて言語モデルを学習し、日本語テキストの難易度を推定する実験を行なったところ、良好な結果が得られ、本稿の冒頭で述べた目的を達成することができた。この実験の詳細については、別稿 [3] で述べる。

謝辞 本研究は、科学研究費補助金 基盤研究 (A) 「円滑な情報伝達を支援する言語規格と言語変換技術」(課題番号 16200009) の支援を受けた。

参 考 文 献

1) Collins-Thompson, K. and Callan, J.: Predicting Reading Difficulty with Statistical Language Models, *Journal of the American Society for Information Science and Technology*, Vol. 56, No. 13, pp. 1448–1462 (2005).
2) 近藤陽介, 佐藤理史: 多項ナイーブベイズ分類を用いた日本語テキストの難易度判定手法の検討, 言語処理学会 第 13 回年次大会発表論文集, pp. 534–537 (2007).
3) 近藤陽介, 松吉俊, 佐藤理史: 教科書コーパスを用いた日本語テキストの難易度推定, 言語処理学会 第 14 回年次大会発表論文集, D5-5 (2008).
4) 文部科学省初等中等教育局: 教科書制度の概要 (2007). http://www.mext.go.jp/a_menu/shotou/kyoukasho/gaiyou/04060901.htm.
5) 柴崎秀子, 沢井康孝: 国語教科書コーパスを応用した日本語リーダビリティー構築のための基礎研究, 信学技報 NLC2007-32, pp. 19–24 (2007).

表 3 本研究と先行研究のコーパスの概要

	言語	テキスト収集源	対象	難易度	テキスト数	文字数	語数	文数
本研究	日本語	全教科教科書	小中高大	13 段階	1,478	1,073,263	–	22,250
近藤ら [2)]	日本語	社会科教科書と社説	中高般	3 段階	121	388,834	–	–
柴崎ら [5)]	日本語	国語科教科書	小	6 段階	155	284,426	121,144	10,753
Collins ら [1)]	英語	ウェブ文書	小中高	12 段階	550	–	415,331	–

表 4 学年別・教科別統計量

		小 1	小 2	小 3	小 4	小 5	小 6	中 1	中 2	中 3	高 1	高 2	高 3	大学
国語科	テキスト数	4	6	6	4	6	7	8	9	10	10	9	7	20
	文字数	1520	4101	3110	2458	3733	4577	7188	8738	10091	7971	9570	7244	21339
	文数	48	115	81	68	88	104	189	233	276	148	176	132	320
	ひらがな率	81.45	73.32	62.96	69.20	60.19	58.42	61.07	61.54	61.67	59.37	59.84	57.44	44.08
	カタカナ率	3.16	4.73	7.56	3.25	3.75	4.35	2.07	3.40	1.51	5.18	4.90	4.61	8.81
	学習漢字率	3.75	10.92	17.40	17.33	23.41	26.17	23.96	21.66	23.59	23.89	23.18	24.59	33.35
	常用漢字率	0.00	0.00	0.10	0.08	0.35	0.44	2.82	1.75	2.31	1.73	2.52	3.12	1.74
	その他の漢字率	0.00	0.02	0.03	0.00	0.00	0.22	0.50	0.21	0.16	0.10	0.64	0.37	0.10
社会科	テキスト数	–	–	10	6	14	9	10	10	10	29	79	50	93
	文字数	–	–	14078	1861	16280	4056	13517	14566	14509	18908	56593	36669	122154
	文数	–	–	441	49	425	85	312	270	285	331	899	593	1415
	ひらがな率	–	–	73.77	62.98	60.48	50.84	52.28	46.83	51.15	43.33	41.32	39.39	49.64
	カタカナ率	–	–	3.59	1.56	3.60	1.58	4.84	3.14	3.48	6.29	7.04	7.17	5.15
	学習漢字率	–	–	12.30	22.51	24.62	33.33	31.42	35.58	33.47	36.14	37.35	38.37	30.75
	常用漢字率	–	–	0.02	0.70	0.39	1.78	1.86	2.88	1.85	3.20	3.33	3.29	2.90
	その他の漢字率	–	–	0.02	0.64	0.52	0.17	0.35	0.79	0.03	0.48	0.43	0.25	0.44
数学科	テキスト数	14	20	20	21	20	22	19	19	19	37	33	34	20
	文字数	3457	7366	5487	6348	6594	7492	9823	11140	11481	20102	16102	21402	26859
	文数	164	333	193	199	210	221	241	253	271	386	345	452	602
	ひらがな率	77.06	68.53	60.74	53.23	52.23	51.99	45.22	44.89	42.81	40.01	42.52	42.95	40.73
	カタカナ率	0.43	2.47	4.45	4.63	3.69	2.96	1.07	1.30	0.97	1.87	1.42	3.08	3.82
	学習漢字率	3.07	8.81	14.45	22.79	23.40	23.17	24.70	24.31	22.35	26.04	26.56	25.57	21.83
	常用漢字率	0.00	0.00	0.00	0.02	0.17	0.04	1.16	0.74	0.57	1.67	1.25	1.72	0.93
	その他の漢字率	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.07	0.09	0.01	0.06	0.01
理科	テキスト数	4	4	13	15	12	10	16	27	13	40	79	79	99
	文字数	1543	1444	3064	4355	6446	5816	7424	11850	5618	33289	42079	45257	80392
	文数	72	52	80	120	179	170	185	262	112	677	822	868	2136
	ひらがな率	86.91	75.48	70.46	64.68	64.06	61.93	52.76	52.71	52.65	48.64	44.63	43.54	45.65
	カタカナ率	1.04	6.51	5.22	2.85	2.75	1.99	4.67	5.57	1.17	7.51	5.96	6.11	5.12
	学習漢字率	1.81	7.89	11.75	20.62	22.51	26.15	29.03	28.90	33.04	30.29	32.74	34.66	29.46
	常用漢字率	0.00	0.00	0.00	0.00	0.00	0.12	1.91	1.96	2.63	2.21	3.83	3.24	3.17
	その他の漢字率	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.12	0.12	0.33	0.32	0.49
芸術科	テキスト数	15	15	14	14	13	13	23	17	16	9	8	7	40
	文字数	4308	5009	4779	6143	7962	8694	21316	17792	17217	8240	5305	7540	53799
	文数	158	176	166	182	246	259	482	388	373	187	109	160	985
	ひらがな率	85.47	78.18	69.47	63.85	59.46	56.08	52.06	49.40	48.19	42.17	40.60	43.53	47.87
	カタカナ率	3.41	5.43	6.05	7.52	7.86	8.17	5.44	5.96	8.12	6.17	7.11	6.54	10.32
	学習漢字率	2.44	6.33	13.89	17.63	20.77	23.86	28.69	30.91	29.81	34.03	31.82	34.77	24.50
	常用漢字率	0.00	0.00	0.00	0.10	0.45	0.90	2.35	3.16	3.14	4.07	4.54	2.44	2.52
	その他の漢字率	0.00	0.00	0.00	0.00	0.15	0.13	0.61	0.49	0.45	1.78	1.55	0.81	0.32
技術家庭科	テキスト数	–	–	7	5	2	1	13	11	10	18	17	16	39
	文字数	–	–	1462	1259	457	237	12486	8439	9458	8403	7913	7266	40718
	文数	–	–	33	25	9	5	269	181	187	145	139	118	580
	ひらがな率	–	–	64.64	64.18	56.89	60.76	52.32	53.12	50.51	52.71	47.73	48.68	42.83
	カタカナ率	–	–	1.16	1.67	0.66	0.00	3.68	3.87	10.08	4.59	6.48	7.42	9.78
	学習漢字率	–	–	23.32	23.35	31.51	23.63	31.76	31.28	26.74	29.64	31.00	30.98	29.31
	常用漢字率	–	–	0.07	0.00	0.22	3.38	1.75	2.10	1.97	2.07	3.80	2.20	2.22
	その他の漢字率	–	–	0.00	0.00	0.00	0.00	0.08	0.04	0.02	0.02	0.08	0.03	0.14
全体	テキスト数	37	45	70	65	67	62	89	93	78	143	225	193	311
	文字数	10828	17920	31980	22424	41472	30872	71754	72525	68374	96913	137562	125378	345261
	文数	442	676	994	643	1157	844	1678	1587	1504	1874	2490	2323	6038
	ひらがな率	82.43	72.89	69.11	61.54	59.46	55.89	52.19	50.63	50.59	46.50	44.10	43.32	46.60
	カタカナ率	2.09	4.14	4.54	4.50	4.28	4.25	4.00	4.06	4.66	5.54	5.87	5.92	6.62
	学習漢字率	2.73	8.53	13.85	20.37	23.33	25.71	28.75	29.44	28.25	30.29	33.11	33.41	28.78
	常用漢字率	0.00	0.00	0.02	0.10	0.30	0.61	1.99	2.24	2.11	2.40	3.26	2.88	2.60
	その他の漢字率	0.00	0.01	0.01	0.05	0.23	0.09	0.34	0.32	0.17	0.32	0.39	0.27	0.35