テキストの難易度と語の分布

Text Readability and Word Distribution

名古屋大学大学院工学研究科教授 **佐藤　理史**

PROFILE: 京都大学大学院工学研究科電気工学第二専攻博士課程研究指導認定退学。博士（工学）。北陸先端科学技術大学院大学、京都大学を経て、2005 年 6 月より名古屋大学大学院工学研究科教授

1 はじめに

情報伝達を主目的とするテキストにおいて最も重要なことは、「正確で読みやすいこと」である。「正確で読みやすい」ことには、多くの要因が関係しているが、「使用する日本語」という観点においては、「平易である」ことに尽きる。しかし、「平易な日本語」とはどのような日本語であるか、その指針はかならずしも明確ではない。

「平易な日本語」の指針を作るためには、実際の日本語の使用についての量的な把握が必要である。たとえば、基本語彙を決めるには、それを何語とすべかの判断が必要となり、その判断は、どのぐらいの語彙で、どのぐらいの割合のトークン（語の出現）をカバーできるのかといった調査に基づいて決定する必要がある。

我々は、「平易な日本語」の指針の中核要素となる基本語彙を、『現代日本語書き言葉均衡コーパス（BCCWJ）』[1] に基づいて編纂することを計画している。BCCWJ を基礎資料として用いるのは、次の理由による。

（1）BCCWJ は、日本語初の均衡コーパスであり、そこから得られる統計量は、母集団である日本語全体の統計量の良い近似となっていることが期待できる。

（2）BCCWJ では、短単位と長単位という 2 つの語の単位が採用されており、これまでの日本語の語彙において問題であった、単位の不明確さの問題を回避できると考えられる。

（3）BCCWJ の各サンプルに、9 段階のテキスト難易度を付与したデータが存在する [2]。この難易度データを利用することにより、テキスト難易度と語の分布の関係を観察することができる。

本稿では、基本語彙の編纂の準備段階として、BCCWJ に対して行なった、テキストの難易度と語の分布の関係に関する調査について報告する。なお、本稿は、文献 [3] のダイジェスト版である。

2 調査対象

『現代日本語書き言葉均衡コーパス（BCCWJ）』（DVD 版）[1] から、2 つのサブコーパスを抽出・編纂し、それらを調査対象とした。その概要を表 1 に示す。

コーパス A は、BCCWJ に含まれる書籍の固定長サンプル（長さ 1000 字）のうち、若干の特異なサンプルを除去したものである。これらのサンプルには、すべて obi2/B9 難易度が付与されている [2]。この難易度は、1 から 9 までの 9 段階で、1 が最もやさしく、9 が最も難しい。その難易度分布は、ほぼ正規分布に従う。たとえば、難易度 5 のサンプル数は全体の約 20% を占

表 1　調査対象の概要

		コーパス A	コーパス B
サンプル数		20,544	7,200
総文字数		約 2,000 万	約 720 万
短単位語	トークン数	12,962,906	4,519,110
	タイプ数	107,243	69,135
長単位後	トークン数	10,534,524	3,642,366
	タイプ数	515,203	235,836

めるが、難易度 1 と 9 のサンプル数は、それぞれ全体の約 4% に過ぎない。

異なる難易度間でサンプルの統計を比較するためには、それぞれの難易度でサンプルの数を揃えておくことが望ましい。そのため、各難易度のサンプル群からそれぞれ 800 サンプルを選んで新たなコーパスを作成した。これがコーパス B である。

BCCWJ の DVD 版には、短単位語（SUW）および長単位語（LUW）の解析結果を格納した形態論情報付きデータ（TSV データ）が含まれている。短単位語と長単位語は、この TSV データから抽出した。この際、語種が和語、漢語、外来語、混種語、固有語のいずれかであるもののみを残し、記号や語種が記述されていないもの（未知語）を削除した。

短単位と長単位という語の単位を簡潔に説明するのは難しいが、おおよそ、短単位は国語辞典の見出しとなるような語の単位（複合語を含まない）、長単位は文節を内容部と機能部に分割した単位、と捉えればよい（表 2 参照）。それらの認定規準は、文献 [4、5] に詳細に述べられている。

表 2　文節・長単位・短単位の例（文献 [6]）

文節	日本語コーパスについて						解説した		
長単位	日本語コーパス			について			解説し		た
短単位	日本	語	コーパス	に	つい	て	解説	し	た

本稿では、「トークン」を語の出現を表す用語として、「タイプ」を語の異なり（語彙素が一致するものを同一タイプとみなす）を表す用語として用いる。あるテキストまたはテキスト群において、トークン数は語の総出現数を表し、タイプ数は何種類の語が現れたかを表す。

3 頻出語の累計カバー率

一般に、言語において、比較的少数の頻出語が、テキストに現れるトークンの大半を占めることは、よく知られた事実である。しかしながら、日本語においては、語の単位の問題があり、何タイプぐらいで、どのぐらいの割合をカバーするのかは、文献 [7] に示されている新聞（短単位）と雑誌（β 単位）の調査結果を除き、かならずも明確には示されてきたとは言いがたい。

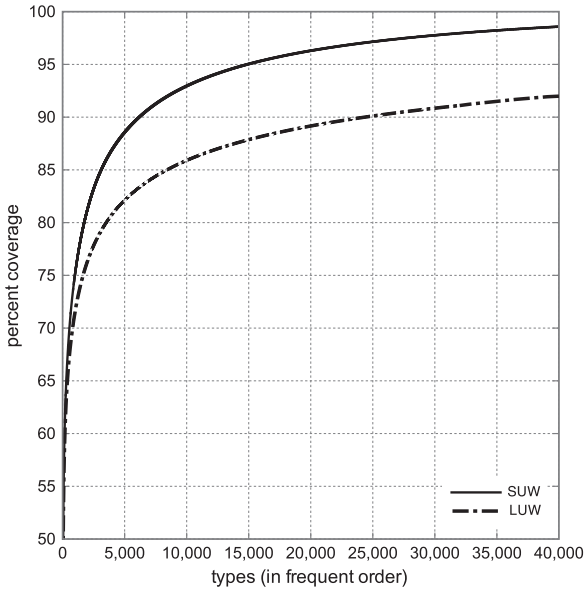

図 1　累計カバー率（コーパス A）

図 1 に、コーパス A 全体の累計カバー率のグラフを示す。このグラフの X 軸は、高頻度順に並べた語のタイプ数を表し、Y 軸は、それらの語のトークン数の総計（頻度累計）が全体に占める割合をパーセントで示している。

このグラフより、短単位語と長単位語では、カバー率の上昇に大きな差があることが確認できる。主要なカバー率達成に必要なタイプ数を表 3 に示す。

表 3　主要なカバー率達成に必要なタイプ数（コーパス A）

	短単位語	長単位語
75%	987	1,649
80%	1,736	3,509
85%	3,108	8,396
90%	6,165	24,323
95%	14,860	98,314

ここで、累計カバー率を難易度別に求めたら、どのような結果が得られるであろうか。この計測には、各難易度のサンプル数を揃える必要があるので、コーパス B を用い、全体の傾向が知りたいので、難易度 1 から 3 を難易度 Easy（E）、難易度 4 から 6 を難易度 Moderate（M）、難易度 7 から 9 を難易度 Difficult（D）として、3 段階に集約し、累計カバー率を求めた。得られた結果を図 2 と図 3 に示す。

順序は前後するが、まず、長単位語の累計カバー率（図 3）からみていこう。長単位語の累計カバー率は、難易

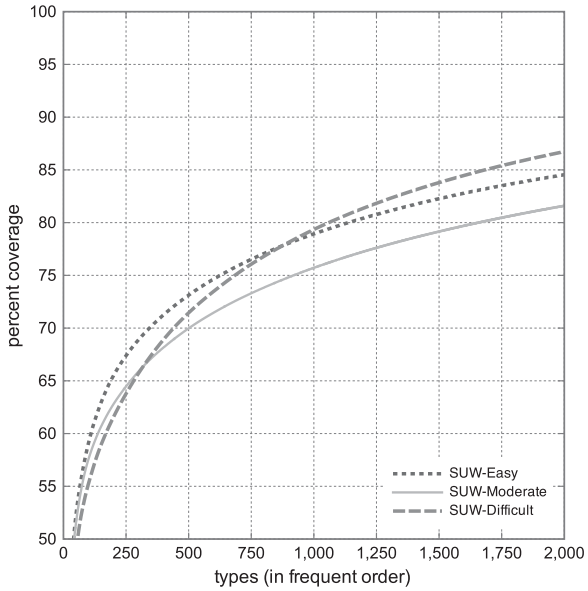

図２　短単位語の累計カバー率（コーパス B）

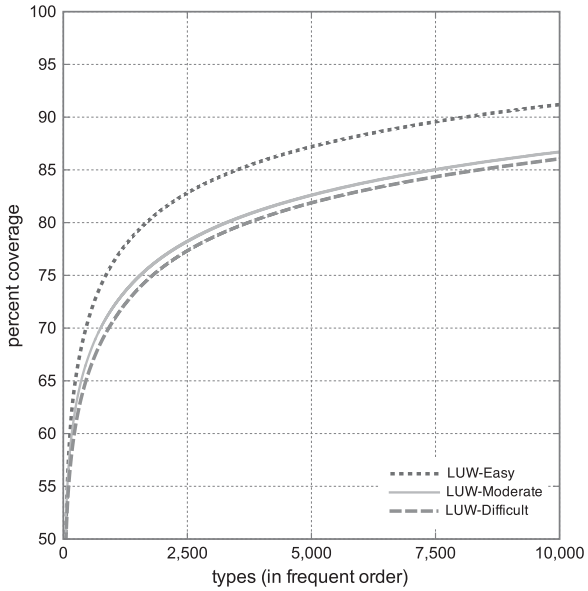

図３　長単位語の累計カバー率（コーパス B）

度が上昇するにつれて、立上りが遅くなり、上昇がより鈍化するのがわかる。これは予想通りの結果である。

一方、短単位語の累計カバー率（図２）は、異なる振舞をみせる。難易度 E と難易度 M の関係は、長単位語と同じである。しかし、難易度 D は、立上りこそ遅いが、319 位で難易度 M を越え、876 位で難易度 E を越える。これは、予期しなかった結果である。

4　難易度と頻出語

先の結果は、難易度によって頻出語が異なる可能性を示唆する。これを確認するために、まず、コーパス A 全体に対して、各タイプのトークン数（頻度）を計測し、各タイプに頻度順の順位をつける（頻度が同じ場合は、同位とする）。次に、コーパス A を難易度 E、難易度 M、難易度 D の 3 セグメントに分割し、それぞれのセグメントに対して、同様に順位付けを行なう。最後に、コーパス A 全体の頻出語上位 N 位までと、各セグメントの頻出語上位 N 位までを比較し、両方に含まれるタイプがどのぐらいあるか（積集合の要素数）を調べる。得られた結果を表４に示す。この表において、記号∩は積集合の要素数を表し、∪は和集合の要素数を表す。

この表からわかるように、難易度 M のセグメントの頻出語は、コーパス全体の頻出語に対して 85% 以上重複する。これは、難易度 M セグメントが、コーパス全体の約半分を占めることによる。その一方で、難易度 E と難易度 D のセグメントでは、全体に対する重複度は 60 ～ 70% 程度に低下する。この表には、難易度 E と難易度 D のセグメント間の頻出語の重複度も示したが、これは、おおよそ 30 ～ 40% 程度である。以上のことから、テキストの難易度が異なれば、頻出語の集合が異なることが確認できる。

5　まとめ

本稿では、BCCWJ に対して行なった、テキストの難易度と語の分布の関係に関する調査について報告した。今回の調査で判明した最も大きな発見は、「テキストの難易度セグメントによって、頻出語の集合はかなり異なる」ということであり、これは、「たとえ、1,000 ～ 2,000 語の基本語彙を選定するのであっても、頻度を計測するコーパスは注意深く選定しなければならない」ことを意味する。

謝辞：本研究では、『現代日本語書き言葉均衡コーパス（BCCWJ）』（DVD 版）を使用した。

表4　頻出語の重なり（コーパス A）

	短単位語			長単位語		
N	∩	∩/N	∪	∩	∩/N	∪
Easy vs 全体						
50	40	0.80	60	41	0.82	59
100	72	0.72	128	80	0.80	120
250	182	0.73	318	190	0.76	310
500	337	0.67	664	360	0.72	642
1,000	645	0.65	1,358	720	0.72	1,281
2,000	1,312	0.66	2,703	1,436	0.72	2,572
4,000	2,698	0.67	5,323	2,768	0.69	5,294
8,000	5,532	0.69	10,740	5,405	0.68	10,668
Moderate vs 全体						
50	49	0.98	51	49	0.98	51
100	94	0.94	106	91	0.91	109
250	221	0.88	279	228	0.91	272
500	446	0.89	554	453	0.91	547
1,000	854	0.85	1,148	891	0.89	1,111
2,000	1,726	0.86	2,282	1,759	0.88	2,243
4,000	3,492	0.87	4,517	3,504	0.88	4,515
8,000	7,031	0.88	9,104	6,923	0.87	9,251
Difficult vs 全体						
50	40	0.80	60	42	0.84	58
100	78	0.78	122	70	0.70	130
250	155	0.62	345	166	0.66	335
500	318	0.64	682	306	0.61	695
1,000	647	0.65	1,360	600	0.60	1,405
2,000	1,325	0.66	2,688	1,202	0.60	2,802
4,000	2,703	0.68	5,346	2,493	0.62	5,531
8,000	5,501	0.69	10,811	4,886	0.61	11,193
Easy vs Difficult						
50	30	0.60	70	34	0.68	66
100	52	0.52	148	53	0.53	147
250	96	0.38	404	113	0.45	388
500	177	0.35	824	189	0.38	814
1,000	331	0.33	1,677	357	0.36	1,647
2,000	705	0.35	3,317	709	0.35	3,299
4,000	1,576	0.39	6,480	1,436	0.36	6,616
8,000	3,464	0.43	13,006	2,777	0.35	13,365

参考文献

[1] 国立国語研究所．現代日本語書き言葉均衡コーパス（BCCWJ）、国立国語研究所、http://www.ninjal.ac.jp/corpus_center/bccwj/.

[2] 佐藤理史．現代日本語書き言葉均衡コーパスに対する難易度付与、第２回コーパス日本語学ワークショップ予稿集、pp.175-184（2012）.

[3] 佐藤理史．テキストの難易度と語の分布、情報処理学会研究報告、2013-NLP-213 No.6、情報処理学会（2013）.

[4] 小椋秀樹、小磯花絵、冨士池優美、宮内佐夜香、小西光、原裕．『現代日本語書き言葉均衡コーパス』形態論情報規定集 第４版（上）、LR-CCG-20-05-01、国立国語研究所（2011）.

[5] 小椋秀樹、小磯花絵、冨士池優美、宮内佐夜香、小西光、原裕．『現代日本語書き言葉均衡コーパス』形態論情報規定集 第４版（下）、LR-CCG-20-05-02、国立国語研究所（2011）.

[6] 小椋秀樹．『現代日本語書き言葉均衡コーパス』の短単位・長単位、『中納言』講習会、国立国語研究所 言語資源系・コーパス開発センター（2013）.

[7] 林大（監修）、宮島達夫、野村雅昭、江川清、中野洋、真田信治、佐竹秀雄（編）．図説日本語、角川書店（1982）.