Vol.2013-NL-213 No.6
2013/9/12

テキストの難易度と語の分布

佐藤 理史[1,a)]

概要：現代日本語書き言葉均衡コーパスに対して実施した，テキストの難易度と語の分布に関する一連の調査の結果について報告する．

キーワード：現代日本語書き言葉均衡コーパス，テキストの難易度，語の分布

Text Readability and Word Distribution

Satoshi Sato[1,a)]

Abstract: This paper reports the result of investgation that examines the relation between text readability and word distribution on the Balanced Corpus of Comtemporary Written Japanese.

Keywords: Balanced Corpus of Contemporary Written Japanese, text readability, word distribution

1. はじめに

テキストからの情報取得は，我々の日常生活において，重要な役割を果たしている．新聞やポータルサイトから世の中の動きを知ること，説明書や解説を読んで装置やサービスの使い方を知ること，何かしら新しいことを書籍や論文等から学ぶことなどは，その典型例である．

そのような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」には，多くの要素が関係するが，「使用する日本語」という観点においては，「平易である (plain)」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は，かならずしも明確ではない．

日本語テキストの難易度を測定する方法の一つに，我々が実現したツール obi2 [1] がある．このツールは，与えられた日本語テキストに対し，教科書コーパスに基づく 13 段階の難易度 (T13 スケール) と，『現代日本語書き言葉均衡コーパス (BCCWJ)』[2] に基づく 9 段階の難易度 (B9 スケール)[3] のそれぞれの値を出力する．このうち，後者は，人間の判断と高い相関を示すことが，被験者実験により確認されている [4].

しかし，このツールが提供する機能は，与えられたテキストの難易度評価だけであり，「平易な日本語とはどのようなものか」，あるいは，「文章をより平易にするためにはどうすればよいか」といった問いに対しは，直接的な答えを提供しない．これらの問いに答えるためには，「平易な日本語とはどのような日本語か」ということを明確化・明示化した指針が不可欠である．

「平易な日本語」の指針を作ろうとする場合，その指針の中心は，語彙表となる．言語を構成する二大要素は語彙と文法であるが，日本語では，文法事項の大部分が機能語と結び付いているため，どのような機能語を語彙表に含めるかを定めれば，間接的に，「平易な日本語」に含めるべき文法事項を指定したことになるからである．

「平易な日本語」の指針の中心となる語彙表を，仮に「基本語彙表」と呼ぶことにしよう．我々は，この基本語彙表を，『現代日本語書き言葉均衡コーパス (BCCWJ)』[2] 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

1 名古屋大学大学院工学研究科
Graduate School of Engineering, Nagoya University
2 名古屋大学工学部
School of Engineering, Nagoya University
a) ssato@nuee.nagoya-u.ac.jp

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．これは，基本語彙表を定めるための資料として望ましい性質である．

(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表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 [5]．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，基本語彙表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一連の調査 (上記の (3)) について報告する．この調査の背後にある基本的なアイディアは，

基本語彙の選定において，各サンプルテキストに
付与された難易度を利用できないか

というものである．たとえば，

- 難易度の低いテキストに頻出する語を，基本語の候補とする
- 頻出語であっても，出現が難易度の高いテキストに偏る語は，基本語の候補としない

といった方針の採用が考えられる．さらに，学習者用 (初心者用) の基本語彙表，母語話者用 (一般用) の基本語彙表といった，用途別の基本語彙表の編纂を視野に入れるのであれば，

- あらかじめ定めた難易度レベルに対して，基本語を選定することができないか

という可能性についても，検討の余地がある．

以下，本稿は，次のように構成されている．まず，2 節で，本稿で用いる用語を説明し，それらにまつわる 2 つの事項について確認する．3 節では調査対象に使用するデータについて述べ，4 節では本稿で検証する事項を提示する．5 節から 7 節にかけては，実際の調査結果とその分析について述べる．最後に，8 節で，本研究をまとめる．

2. 用語と確認事項

2.1 用語

本稿では，次の用語を用いる．

(1) 語の出現をトークン (token) と呼ぶ．トークン数とは，語の出現数 (のべ数) である．

(2) 語の異なりをタイプ (type) と呼ぶ．タイプ数とは，語の種類の数である．

(3) 語の単位として，短単位語 (**SUW**) と長単位語 (**LUW**) の 2 種類を設定する．これらの定義は，BCCWJ の定義 [6], [7] に従う．

(4) 語種として，和語，漢語，外来語，混種語，固有語の 5 種類を設定する．これらの定義は，BCCWJ の定義 [7] に従う．なお，BCCWJ には語種「記号」が含まれるが，本研究では，そのような語をあらかじめ除外するため，語種に「記号」を含めない．

表 **1** BCCWJ の固定長サンプルの概要

レジスタ	#	$c \geq 500$	use
PB (出版・書籍)	10,117	10,026	√
LB (図書館・書籍)	10,551	10,518	√
PM (出版・雑誌)	1,996	1,938	
PN (出版・新聞)	1,473	1,457	
OW (白書)	1,500	1,491	

2.2 確認事項

ほとんど自明ではあるが，次の 2 つの事項を確認しておく．

(1) トークン数は，テキストサイズに比例するが，タイプ数は比例しない．つまり，トークン数とタイプ数の比は，テキストサイズに依存する．

(2) 任意のテキストに対して，

$$\text{“SUW のトークン数”} \geq \text{“LUW のトークン数”} \quad (1)$$

が成り立つ．しかし，

$$\text{“SUW のタイプ数”} \leq \text{“LUW のタイプ数”} \quad (2)$$

は，必ずしも成り立たない．

最後の事項については，若干説明が必要かもしれない．一般に，十分に大きなテキストに対しては，式 (2) が成立する．しかしながら，たとえば，1,000 字のテキストにおいては，これが成り立たない場合がある．なぜならば，ある LUW が複数の SUW から構成されており，それらの SUW がその LUW においてしか出現しない場合，タイプ数の逆転が起こるからである．

3. 調査対象

本研究では，『現代日本語書き言葉均衡コーパス (BCCWJ)』(DVD 版)[2] を使用する．加えて，BCCWJ の各サンプルに付与された難易度データ [5] も使用する．

3.1 使用するサンプル

BCCWJ のサンプルには，固定長サンプルと可変長サンプルが存在する．このうち，計量的分析に向いているのは，固定長サンプルである．本研究では，固定長サンプルを使用する．

BCCWJ の固定長サンプルの概要を表 **1** に示す．固定長サンプルの長さは 1,000 字であるが，実際にサンプルを調べてみると，常識的には日本語テキストの構成文字とは考えにくい文字を多数含むサンプルも存在する．そこで，有効文字[*1]が 500 文字以上含まれているサンプルのみを使用

*1 ひらがな，カタカナ，JIS 第一水準の漢字．

表 **2** 調査対象

		コーパス A	コーパス B
サンプル数		20,544	7,200
総文字数		約 2000 万	約 720 万
SUW	token	12,962,906	4,519,110
	type	107,243	69,135
LUW	token	10,534,524	3,642,366
	type	515,203	235,836

する．

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 [5]．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，BCCWJ の書籍レジスタ (PB と LB) において，ほぼ，正規分布に従う[*2]．これに対して，他のレジスタ (PM, PN, OW) の難易度分布には，大きな偏りが存在する．このため，本研究では，書籍レジスタ (PB と LB) のみを使用する．使用するサンプル数は，PB が 10,026 サンプル，LB が 10,518 サンプルの，計 20,544 サンプルである．これを，**コーパス A** (all) と呼ぶ．なお，このコーパスのサンプルの分布は，「均衡」という性質を受け継いでいる．

すでに 2.2 節で述べたように，トークン数はテキストサイズに比例するが，タイプ数は比例しない．そのため，タイプ数を比較する際には，比較対象のテキストサイズを揃えておく必要がある．この条件を満たすために，各難易度のサンプル群から，それぞれランダムに 800 サンプルずつ選び，難易度毎のテキストサイズ (サンプル数) が同一となるようなコーパスを新たに構成する．これを，**コーパス B** (balanced) と呼ぶ．

3.2 SUW と LUW の抽出

BCCWJ の DVD 版には，SUW および 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．SUW と LUW は，この TSV データから抽出した．この際，語種が前述の 5 種類以外のものを除外した[*3]．それぞれのコーパスの SUW と LUW の総トークン数および総タイプ数を表 **2** に示す．

4. 検証する仮説

本稿で検証するのは，次のような直感に基づく仮説である．

仮説 1 テキストの難易度が高いほど，
- **仮説 1A** 色々な種類の語が出現する
- **仮説 1B** 複合語の割合が増える
- **仮説 1C** 1 つの語からの，より多くの複合語の生産が観察される
- **仮説 1D** 漢語の割合が増える

仮説 2 テキストの難易度が高いほど，頻出語の累計カバー率の上昇が遅い (全体に対する，頻出語が占める割合が低い)

仮説 3 テキストの難易度によって，頻出語が異なる

ここでは，直感性を重視し，あえてあいまいな書き方を採用している．そのため，実際の検証では，これらの仮説が意味するところを明確にした後，実際のデータと照合する．

[*2] 難易度スケール obi2/B9 は，難易度の値の分布が，書籍レジスタに対して正規分布をとるように設計されている [3].

[*3] 語種が記述されていないもの (未知語)，および，「記号」が除外される．

表 **3** トークン数とタイプ数

		サンプル単位	セグメント単位
SUW	token	so	$SO\ (=\sum so)$
	type	sy	$SY\ (\neq \sum sy)$
LUW	token	lo	$LO\ (=\sum lo)$
	type	ly	$LY\ (\neq \sum ly)$
計測数		800	1

5. トークン数・タイプ数に基づく分析

仮説 1 を検証するために，SUW と LUW のトークン数とタイプ数を調べる．この調査では，難易度間の比較が必要となるので，コーパス B を用いる．

5.1 計測する値

SUW と LUW のトークン数とタイプ数の計測には，サンプル単位で計測するか，あるいは，特定の難易度を持つサンプル群 (以下，**セグメント**と呼ぶ) 単位で計測するか，の選択肢が存在する．サンプル単位の計測は，「それぞれのテキストで，どのような値をとるか」を調べることに相当し，セグメント単位の計測は，「あるテキスト群全体で，どのような値をとるか」を調べることに相当する．本研究では，これらの両方を計測し，検証すべき仮説に応じて，参照すべき計測単位を定める．

表 **3** に計測するトークン数とタイプ数の一覧を示す．本論文では，以下のような記号を使用する．

(1) 語の単位を表す 1 文字に，s (SUW) と l (LUW) を採用する．
(2) トークンとタイプを表す 1 文字に，o (token) と y (type) を採用する．
(3) サンプル単位の計測値を小文字で，セグメント単位の計測値を大文字で表す．

たとえば，so は，SUW のトークン数をサンプル単位で計測した値，LY は，LUW のタイプ数をセグメント単位で計測した値，を表す．

5.2 計算する値

それぞれの単位で計測した 4 つの値から，表 **4** に示す 4 つの比を計算する．この表は，サンプル単位の場合を示したが，セグメント単位も同様である．

表 4　計算する値

	SUW	LUW	SUW/LUW
token	so	lo	so/lo
type	sy	ly	sy/ly
token/type	so/sy	lo/ly	

それぞれの比は，以下のことを表す．

- トークン数とタイプ数の比 (so/sy, lo/ly)：
  1 タイプ当たり，平均何トークン出現したかを表す．
- トークン数における，SUW と LUW の比 (so/lo)：
  1 つの LUW が，平均いくつの SUW から構成されているかを表す．つまり，テキスト中に複合語がどれだけ多いか・長いかを示す．
- タイプ数における，SUW と LUW の比 (sy/ly)：
  この比の逆数 (ly/sy) は，SUW の 1 タイプ当たり，何種類の LUW が存在するかを示す．つまり，1 つの SUW から作られる複合語の種類が，多いか少ないかを示す．

サンプル単位の測定値および計算値は，各難易度セグメントに対して，それぞれ 800 件存在する．これらの平均値を $\overline{so}$, $\overline{so/sy}$ のように，バーによって表す．

5.3　計測・計算結果

コーパス B に対する計測・計算結果を表 **5** に示す．表の左半分はサンプル単位の平均値，右半分はセグメント単位の値を示している．表の下半分は，難易度による値の変化をわかりやすくするために，対応する難易度 5 の値との比を示している．この比が，1.0 から 10%以上変動した値[*4] を，ボールド体で示した．さらに，難易度に対する $\overline{sy}$, $\overline{so/sy}$, $\overline{ly}$, $\overline{lo/ly}$ の変動の様子を図 **1** に，$\overline{so/lo}$, $\overline{sy/ly}$, SY/LY の変動の様子を図 **2** に示した．

5.4　仮説 1A の検証

仮説 1A は，「テキストの難易度が高いほど，色々な種類の語が出現する」である．この仮説が意味するところは，「それぞれのテキストにおいて，そのテキストの難易度が高いほど，出現するタイプ数が多い」である．つまり，サンプル単位で計測したタイプ数を調べればよい．

(1) サンプル単位の SUW タイプ数の平均値 ($\overline{sy}$) は，難易度 1–4 において緩やかに上昇するが，難易度 6–9 においては逆に減少する (図 1).

(2) LUW タイプ数の平均値 ($\overline{ly}$) も，ほぼ同様の傾向を示す (図 1).

これらからわかることは，「タイプ数は，難易度の高いところで，難易度の上昇とともに減少する」ということである．サンプル単位のトークン数の平均値 ($\overline{so}$ と $\overline{lo}$) は，難易度に対して若干変動するが，タイプ数をトークン数で正規化した値，すなわち，$\overline{so/sy}$ の逆数と $\overline{lo/ly}$ の逆数も，タイプ数と同様に傾向を示す (図 1)．以上のことより，仮説 1A が成立しないことは明らかである．

では，なぜ仮説 1A が成り立たないのであろうか．仮説 1A の背後にある直感は，おそらく，「難易度が低いところでは語彙は限られているので，出現する語の種類は限定されるはずだ」というものであろう．難易度低位から中位にかけてのタイプ数の増加は，この直感を裏付けるものである．つまり，難易度低位から中位にかけては，仮説 1A は成立する．

仮説 1A が成り立たないのは，難易度中位から高位にかけてである．おそらく，ここでは，もはや語彙は制限されず，別の要因が支配的になるのであろう．ここで観察されることは，$\overline{sy}$ や $\overline{ly}$ が減少傾向に転じ，$\overline{so/sy}$ や $\overline{lo/ly}$ の値が増加することである．後者の 2 つの値は，1 タイプが平均的に何トークン出現するかを表す．これらの値の増加が意味することは，同一語 (タイプ) の繰返し使用回数が増えるということである．

難易度高位の 1,000 字のテキストにおいて，同一語の繰返し使用回数が増えるということは，どのように解釈できるのであろうか．その一つの説明は，「ある特定の事柄に対して，濃密に (集中して詳しく) 書かれる」ということであろう．ある事柄について狭く深く記述する場合，広く浅く記述するのに比べ，使用される語のタイプ数は相対的に少なくなると考えられる．これが，難易度高位のテキストでタイプ数が減少する原因と考えられる．

5.5　仮説 1B の検証

仮説 1B は，「テキストの難易度が高いほど，複合語の割合が増える」である．複合語の割合は，SUW と LUW の比で観測することができる．ここでの複合語の割合は，トークン数での割合と考えるのが妥当であろう．

トークン数の難易度による変動は，サンプル単位で計測しても，セグメント単位で計測しても，ほぼ同じとなる[*5]．それゆえ，以下では，サンプル単位のみを取り上げる．

(1) サンプル単位の平均値である $\overline{so/lo}$ は，難易度 1–2 を除き，難易度の上昇に従って一貫して上昇する (図 2).

図 2 より，全体の傾向は明らかであり，仮説 1B「テキストの難易度が高いほど，複合語 (複数の SUW から構成される LUW) トークンの割合は増える」は成立するとみなしてよい．

5.6　仮説 1C の検証

仮説 1C は，「テキストの難易度が高いほど，1 つの語からの，より多くの複合語の生産が観察される」である．これは，SUW と LUW のタイプ数の比を観察すればよい．

[*4] これらの値は比なので，10%の変動は，$1.1 (= 1.0 \times 1.1)$ 以上，$0.909 (= 1.0/1.1)$ 以下とした．

[*5] サンプルの大きさがほぼ均一であるからである．

表 **5** 難易度別のトークン数・タイプ数—コーパス B

	サンプル単位の平均値								セグメント単位							
	$\overline{so}$	$\overline{sy}$	$\overline{so/sy}$	$\overline{lo}$	$\overline{ly}$	$\overline{lo/ly}$	$\overline{so/lo}$	$\overline{sy/ly}$	SO	SY	$\frac{SO}{SY}$	LO	LY	$\frac{LO}{LY}$	$\frac{SO}{LO}$	$\frac{SY}{LY}$
1	583.3	224.3	2.634	507.9	217.9	2.349	1.153	1.028	466,651	20,741	22.50	406,334	33,813	12.02	1.148	0.613
2	613.2	238.9	2.589	550.3	235.2	2.354	1.115	1.015	490,547	21,051	23.30	440,263	30,104	14.62	1.114	0.699
3	632.4	248.7	2.570	555.2	241.1	2.318	1.140	1.030	505,958	24,560	20.60	444,145	36,457	12.18	1.139	0.674
4	635.6	252.1	2.559	538.4	240.0	2.262	1.183	1.049	508,479	27,669	18.38	430,718	44,096	9.77	1.181	0.627
5	634.2	250.6	2.568	514.6	233.3	2.222	1.236	1.072	507,347	29,647	17.11	411,689	50,573	8.14	1.232	0.586
6	644.1	247.6	2.642	498.1	228.4	2.195	1.299	1.082	515,258	26,436	19.49	398,456	51,337	7.76	1.293	0.515
7	628.8	229.8	2.774	474.1	215.5	2.217	1.330	1.065	503,034	21,021	23.93	379,295	48,430	7.83	1.326	0.434
8	598.5	204.3	3.000	437.2	194.6	2.282	1.373	1.045	478,803	15,376	31.14	349,742	45,840	7.63	1.369	0.335
9	678.8	200.4	3.468	477.2	194.7	2.487	1.430	1.028	543,033	11,516	47.15	381,724	38,234	9.98	1.423	0.301
1	0.92	**0.90**	1.03	0.99	0.93	1.06	0.93	0.96	0.92	**0.70**	**1.32**	0.99	**0.67**	**1.48**	0.93	1.05
2	0.97	0.95	1.01	1.07	1.01	1.06	**0.90**	0.95	0.97	**0.71**	**1.36**	1.07	**0.60**	**1.80**	**0.90**	**1.19**
3	1.00	0.99	1.00	1.08	1.03	1.04	0.92	0.96	1.00	**0.83**	**1.20**	1.08	**0.72**	**1.50**	0.92	**1.15**
4	1.00	1.01	1.00	1.05	1.03	1.02	0.96	0.98	1.00	0.93	1.07	1.05	**0.87**	**1.20**	0.96	1.07
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	1.02	0.99	1.03	0.97	0.98	0.99	1.05	1.01	1.02	**0.89**	**1.14**	0.97	1.02	0.95	1.05	**0.88**
7	0.99	0.92	1.08	0.92	0.92	1.00	1.08	0.99	0.99	**0.71**	**1.40**	0.92	0.96	0.96	1.08	**0.74**
8	0.94	**0.82**	**1.17**	**0.85**	**0.83**	1.03	**1.11**	0.97	0.94	**0.52**	**1.82**	**0.85**	**0.91**	0.94	**1.11**	**0.57**
9	1.07	**0.80**	**1.35**	0.93	**0.83**	**1.12**	**1.16**	0.96	1.07	**0.39**	**2.76**	0.93	**0.76**	**1.23**	**1.16**	**0.51**

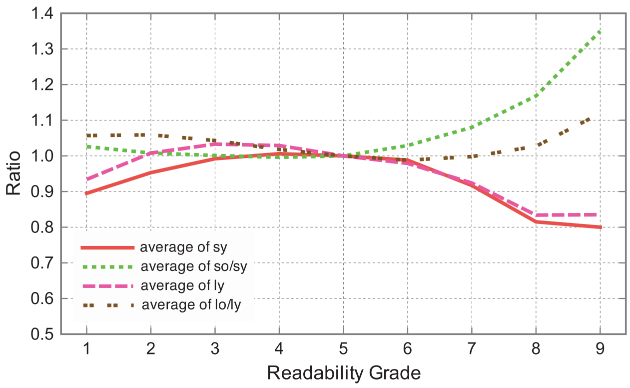

図 **1** $\overline{sy}$, $\overline{so/sy}$, $\overline{ly}$, $\overline{lo/ly}$ の変動

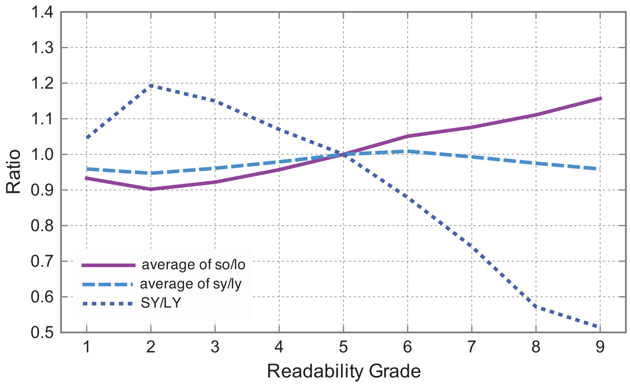

図 **2** $\overline{so/lo}$, $\overline{sy/ly}$, SY/LY の変動

計測単位としては，サンプル単位とセグメント単位の両方を観察する．

まず，サンプル単位の場合を見てみよう．

(1) サンプル単位の平均値 $\overline{sy/ly}$ の値は，難易度にはあまり依存せず，変動は ±10% の範囲内に収まる (図 2)．しいて言うのであれば，難易度 6 をピークとする，上に凸の線を描く．

サンプル単位では，仮説 1C が成立しないのは明らかである．しかしながら，より詳細に見れば，難易度中位から高位にかけては，$\overline{sy/ly}$ の低下が見られる．仮説 1C の「難易度が高いほど」は，おそらく，難易度中位から高位の範囲を想定している．その範囲においては，仮説 1C は，あながち的外れではない．

一方，先に議論したように，難易度低位から中位にかけては，語彙 (SUW) の制限が想定される．その影響を受けて，難易度低位において，$\overline{sy/ly}$ が相対的に小さな値をとるのだと考えられる．

次に，セグメント単位の場合を見てみよう．

(1) セグメント単位の SY/LY の値は，難易度 1 を除けば，難易度の上昇につれて大幅に減少する．

難易度 1 を除外すれば，セグメント単位では，仮説 1C が成立している．すなわち，1,000 字単位のテキストにおいては，仮説 1C は成立しないが，それらを同一難易度で束ねた難易度セグメント単位でみれば，仮説 1C が成立する．つまり，仮説 1C の成立・不成立は，テキストの測定単位に依存する．

5.7 仮説 1D の検証

仮説 1D は，「テキストの難易度が高いほど，漢語の割合が増える」である．この仮説を検証するために，サンプル単位で各語種のトークン数およびタイプ数の割合を調べ，それらの難易度別の平均値を求めた．その結果を表 **6** に示

表 **6** 難易度別の語種の比率 (サンプル単位の平均値)—コーパス B

	token					type				
	和語	漢語	外来	混種	固有	和語	漢語	外来	混種	固有
	SUW									
1	79.49	13.75	3.84	1.12	1.80	70.90	20.32	5.05	1.80	1.92
2	84.17	10.04	2.04	0.92	2.83	73.89	18.81	3.37	1.74	2.19
3	81.48	13.01	1.89	0.97	2.65	68.03	24.17	3.20	1.90	2.71
4	76.39	17.48	2.05	0.98	3.10	60.49	30.41	3.34	1.89	3.87
5	69.97	23.08	2.44	1.02	3.50	52.43	37.23	3.67	1.93	4.73
6	64.53	28.93	2.06	0.94	3.54	45.67	44.65	3.18	1.80	4.70
7	60.88	33.16	3.34	0.90	1.73	42.91	48.45	4.47	1.71	2.46
8	56.12	35.89	6.38	0.88	0.73	39.54	49.90	7.82	1.61	1.12
9	54.89	41.80	1.26	1.45	0.60	39.05	55.60	1.76	2.53	1.06
	LUW									
1	81.29	9.50	2.64	4.90	1.66	71.60	14.72	3.60	8.44	1.63
2	84.91	7.64	1.63	2.96	2.87	75.88	13.70	2.54	5.74	2.14
3	82.74	9.61	1.46	3.67	2.51	70.72	17.35	2.32	7.15	2.47
4	78.79	12.26	1.50	4.80	2.66	63.68	21.46	2.36	9.14	3.36
5	73.84	15.40	1.68	6.27	2.80	55.77	26.20	2.50	11.62	3.90
6	69.46	19.04	1.35	7.30	2.85	49.11	31.66	1.97	13.42	3.83
7	66.11	22.03	2.05	8.32	1.50	45.43	34.94	2.79	14.80	2.03
8	61.50	22.99	3.90	10.95	0.67	40.54	35.50	4.88	18.15	0.92
9	61.53	27.90	0.64	9.31	0.62	40.10	42.29	0.90	15.87	0.83

す．数字は，すべて，パーセントで表示している．

この表より，SUW と LUW，トークンとタイプの 4 つの組み合わせのいずれにおいても，例外的な難易度 1 を除けば，難易度の上昇について，漢語の割合が上昇し，和語の割合が減少することがわかる．すなわち，仮説 1D は成立すると考えてよい．

5.8 難易度と品詞分布

語種について調べたので，品詞についても調べておこう．サンプル単位で各品詞のトークン数およびタイプ数の割合を調べ，それらを難易度別に平均を計算した結果を表 **7** に示す．数字は，すべて，パーセント表示である．

この表から，数的に主要な品詞は，名詞，動詞，助動詞，助詞の 4 品詞であることがわかる．これらの品詞の難易度に対する変動は，次のとおりである．(ただし，難易度 1 は除外する．)

(1) 名詞の割合は，トークン数・タイプ数ともに，難易度の上昇につれて，一貫して増加する．
(2) 動詞の割合は，トークン数・タイプ数ともに，難易度の上昇につれて，一貫して減少する．
(3) 助動詞のトークン数の割合は，難易度の上昇につれて，一貫して減少する．一方，タイプの割合，ほぼ一定である．
(4) 助詞の SUW のトークン数・タイプ数の割合は，難易度の上昇につれて，緩やかに減少する．一方，LUW のトークン数・タイプ数の割合は，ほぼ一定である．

名詞や動詞の変動と比較して，機能語 (助動詞および助詞) の難易度変動は，やや複雑である．この点については，ここでは追求しないが，もう少し詳細に検討する必要がある．

5.9 この節のまとめ

本節では明らかになった主要な事実は，以下のとおりである．

(1) 仮説 1A「1000 字のテキストにおいて，テキストの難易度が高いほど，出現するタイプ数が多い」は，難易度中位から高位にかけて，成立しない．
(2) 仮説 1B「1000 字のテキストにおいて，テキストの難易度が高いほど，複合語の割合が増加する」は，成立する．
(3) 仮説 1Ca「1000 字のテキストにおいて，テキストの難易度が高いほど，1 つの語からの，より多くの複合語の生産が観察される」は，成立しない．1 つの語からの複合語の生産は，難易度に関わらず，ほぼ一定である．
(4) 仮説 1Cb「ある難易度をとるテキスト群 (難易度セグメント単位) において，テキストの難易度が高いほど，1 つの語からの，より多くの複合語の生産が観察される」は，成立する．
(5) 仮説 1D「1000 字のテキストにおいて，テキストの難易度が高いほど，漢語の割合が増加する」は，成立する．同時に，「和語の割合は減少する」．
(6)「1000 字のテキストにおいて，テキストの難易度が高いほど，名詞の割合が増加し，動詞の割合が減少する」．

表 **7** 難易度別の品詞の割合 (サンプル単位の平均値)—コーパス B

	名詞	代名	形状	連体	副詞	接続	感動	動詞	形容	助動	助詞	接頭	接尾
	SUW token												
1	30.86	1.63	1.27	0.86	2.56	0.27	0.48	15.00	2.04	9.99	31.23	0.85	2.98
2	23.80	2.97	1.39	1.02	2.96	0.28	0.68	15.73	2.28	12.68	33.73	0.55	1.94
3	25.75	2.48	1.33	1.10	2.54	0.31	0.34	15.57	2.05	12.17	33.61	0.57	2.16
4	29.48	1.95	1.30	1.11	2.13	0.37	0.15	14.97	1.84	10.83	32.56	0.63	2.68
5	34.24	1.33	1.29	1.09	1.69	0.43	0.05	13.93	1.59	9.52	31.08	0.64	3.13
6	36.75	1.14	1.26	1.12	1.42	0.53	0.02	13.42	1.15	8.79	30.07	0.66	3.67
7	39.04	0.83	1.42	1.02	1.14	0.59	0.01	13.05	1.10	7.81	29.33	0.76	3.90
8	43.01	0.56	1.37	0.88	0.79	0.65	0.01	12.47	0.81	6.93	27.93	0.82	3.78
9	43.94	0.46	1.03	1.01	0.56	1.05	0.01	12.14	0.57	6.48	27.03	1.16	4.55
	SUW type												
1	44.70	2.05	2.28	1.31	5.26	0.54	0.94	20.91	3.49	3.62	10.36	0.79	3.73
2	39.69	3.24	2.46	1.48	5.94	0.54	1.34	22.83	3.52	4.05	11.36	0.66	2.89
3	44.07	2.81	2.36	1.48	5.24	0.62	0.73	21.30	3.20	3.68	10.39	0.71	3.41
4	48.80	2.33	2.34	1.41	4.50	0.73	0.31	18.89	2.86	3.46	9.54	0.87	3.95
5	53.68	1.78	2.31	1.30	3.66	0.83	0.10	16.47	2.54	3.17	8.66	1.01	4.49
6	56.57	1.59	2.31	1.27	3.19	1.04	0.05	14.90	1.85	3.05	8.27	1.12	4.79
7	57.70	1.30	2.54	1.27	2.68	1.17	0.03	13.93	1.90	3.04	8.46	1.23	4.75
8	59.91	1.04	2.51	1.25	1.98	1.29	0.02	12.45	1.55	3.34	8.72	1.32	4.62
9	60.19	0.94	2.05	1.22	1.54	1.74	0.02	12.98	1.07	3.63	8.28	1.33	5.03
	LUW token												
1	28.95	1.70	1.83	0.98	3.24	0.57	0.54	14.54	2.08	12.94	32.59	0.01	0.02
2	22.40	3.10	2.09	1.13	3.70	0.68	0.78	14.58	2.16	16.03	33.33	0.00	0.01
3	23.77	2.64	2.07	1.25	3.37	0.65	0.40	14.08	1.97	16.06	33.71	0.00	0.01
4	26.39	2.14	2.08	1.31	3.03	0.69	0.18	13.47	1.81	14.94	33.94	0.00	0.01
5	29.73	1.52	2.24	1.34	2.53	0.71	0.06	12.50	1.62	13.67	34.06	0.01	0.02
6	31.08	1.35	2.33	1.44	2.31	0.85	0.03	11.80	1.18	13.16	34.45	0.01	0.02
7	32.50	1.00	2.67	1.35	1.92	0.94	0.02	11.70	1.19	12.13	34.56	0.00	0.02
8	34.88	0.70	2.52	1.19	1.43	0.98	0.01	11.90	0.92	11.05	34.39	0.00	0.02
9	36.87	0.61	1.87	1.43	1.12	1.62	0.01	11.51	0.60	10.46	33.86	0.00	0.03
	LUW type												
1	42.34	2.20	3.18	1.34	6.16	1.07	0.99	21.94	3.54	6.24	10.95	0.01	0.04
2	35.03	3.47	3.71	1.50	7.05	1.23	1.43	23.91	3.56	7.15	11.93	0.01	0.03
3	38.58	3.01	3.65	1.53	6.49	1.22	0.78	22.97	3.25	7.05	11.44	0.01	0.03
4	42.68	2.49	3.64	1.47	5.81	1.27	0.33	21.28	2.94	7.02	11.03	0.01	0.03
5	47.11	1.92	3.88	1.38	4.86	1.25	0.10	19.60	2.64	6.53	10.70	0.01	0.03
6	48.96	1.76	4.02	1.36	4.41	1.46	0.05	18.83	1.91	6.23	10.98	0.01	0.03
7	49.72	1.43	4.43	1.34	3.68	1.57	0.03	18.36	1.94	6.07	11.38	0.01	0.03
8	52.13	1.12	4.16	1.30	2.76	1.58	0.02	17.81	1.53	6.00	11.57	0.00	0.03
9	53.75	0.99	3.27	1.27	2.31	2.16	0.02	17.13	1.00	6.25	11.78	0.00	0.06

6. 頻出語の累計カバー率

次に，仮説 2 を検証するために，頻出語の累計カバー率を調べる．

6.1 短単位と長単位のカバー率

一般に，言語において，比較的少数の頻出語が，テキストに現れる語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，何語ぐらいで，どのぐらいの割合をカバーするのかは，必ずしも明確には示されてきたとは言いがたい[*6]．累計カバー率を計測する対象は，均衡性を有したコーパスが望ましいので，ここでは，コーパス A を用いる．

図 **3** 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，SUW と LUW では，カバー率の上昇

[*6] 文献 [8] に，新聞 (短単位) と雑誌 (β 単位) の調査結果が示されている．

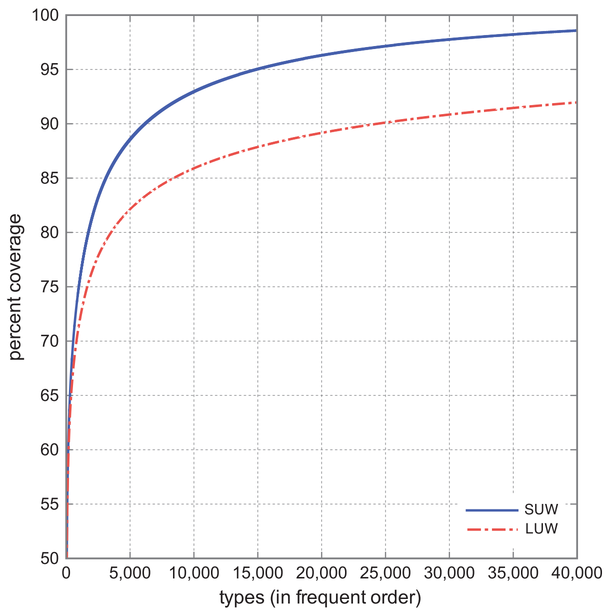

図 3　語の頻度順位と累計カバー率—コーパス A

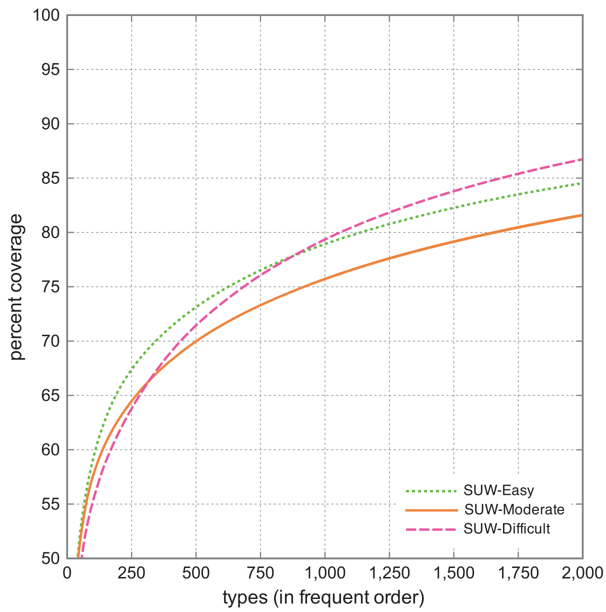

図 4　SUW の頻度順位と累計カバー率—コーパス B

表 8　主要なカバー率達成に必要なタイプ数

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

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 **8** に示す．この表からわかるように，もし，基本語彙の選定の根拠を累計カバー率に求めるのであれば，どちらの単位を採用するかによって，基本語彙の語数が大きく異なることになる．

6.2　難易度別の累計カバー率

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 1–3 を難易度 E(Easy; やさしい)，難易度 4–6 を難易度 M(Moderate; ふつう)，難易度 7-9 を難易度 D(Difficult; 難しい) として，3 段階に集約し，累計カバー率を求める．得られた結果を図 **4** と図 **5** に示す．

順序は前後するが，まず，LUW の累計カバー率 (図 5) からみていこう．LUW の累計カバー率は，難易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．これは仮説 2 に合致する．ただし，難易度 E と難易度 M との差は大きいが，難易度 M と難易度 D の差は小さい．

一方，SUW の累計カバー率 (図 5) は，異なる振舞をみせる．難易度 E と難易度 M の関係は，LUW 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．すなわち，LUW に対しては，明らかに仮説 2 は成立しない．

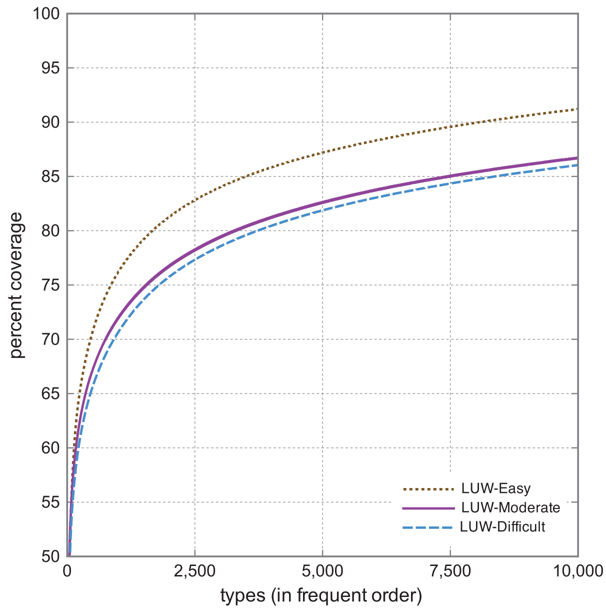

図 5　LUW の頻度順位と累計カバー率—コーパス B

このような現象が起きる一つの原因は，すでに 5.4 節で観察したように，難易度 D における SUW のタイプ数の減少にある．表 **9** に，難易度 E，M，D の 3 つのセグメントの総トークン数と総タイプ数を示す．

この表からわかるように，難易度 D では，SUW のタイプ数の落ち込みが著しい．これが，先ほどの予想外の現象を引き起こす一つの要因と考えられる．

しかしながら，累計カバー率の立上りは，明らかに難易

表 9　3 セグメントのトークン数・タイプ数—コーパス B

	SUW		LUW	
	token	type	token	type
Easy (1–3)	1,463,156	38,639	1,290,742	71,617
Moderate (4–5)	1,531,084	48,592	1,240,863	110,636
Difficult (6–9)	1,524,870	28,599	1,110,761	108,335

度 D が遅い．難易度 M を越えるのが 319 位，難易度 E を越えるのが 876 位であるから，この前後の順位の語の頻度が，他の 2 つのセグメントと比較して多いということである．つまり，テキストの難易度によって，このあたりの順位の頻出語の頻度に差があることを意味する．

6.3　この節のまとめ

本節では明らかになった主要な事実は，以下のとおりである．

(1) SUW と LUW とでは，頻出語の累計カバー率の上昇速度がかなり異なる．
(2) 仮説 2「難易度セグメントにおいて，難易度が高いほど，頻出語の累計カバー率の上昇速度が遅い」は，SUW では成立しない．

7.　難易度と頻出語

最後に，仮説 3 を検証するために，頻出語が難易度によって異なるかどうかを調べる．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[*7]．次に，コーパス A を難易度 E (1–3)，難易度 M (4–6)，難易度 D (7–9) 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 **10** に示す．この表のおいて，記号 $\cap$ は積集合の要素数を表し，$\cup$ 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 (All) 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，仮説 3「テキストの難易度セグメントによって，頻出語の集合が異なる」は成立することが確認できる．

仮説 3 を立てたときの予想は，「テキストの難易度セグメントによって，頻出語の集合は確かに異なるが，80%程度は重複するだろう」というものであった．しかしながら，

*7　トークン数が同じ場合は，同位とする．

表 10　頻出語の重なり—コーパス A

	SUW			LUW		
N	$\cap$	$\cap/N$	$\cup$	$\cap$	$\cap/N$	$\cup$
Easy vs All						
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 All						
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 All						
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

得られた結果は，予想とは異なる．難易度 E と難易度 D のセグメント間を比較する極端な場合では，$N = 100$ であっても，半数強の頻出語しか重複しない．つまり，「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」のである．このことは，次のことを意味する．

> 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．

たとえば，学習者用の基本語彙を定めるために，コーパス A の頻度上位 4,000 位までの LUW を，その候補として採用したとしよう．この場合，難易度 E の上位 4,000 位の LUW のうち，2,768 語しかカバーされず，残りの 1,232 語を落としてしまうことになる．この 1,232 語に含まれる語の

一部を，表 **11** に示す．これらの語には，比較的平易と考えられる「荒い」や「臭い」などの和語形容詞や「焼ける」「冷える」などの和語動詞が含まれる．

以上をまとめると，次のようになる．

(1) テキストの難易度により，頻出語の集合がかなり異なる．(仮説 3 は成立する．)
(2) 基本語彙の候補選択に頻度を用いるのであれば，作成する基本語彙表の用途に合わせ，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．

なお，文献 [9] では，Webster's dictionary の 128,000 語を次のように分類している．

(1) high-frequency words (2,000 words)
(2) university word list (800 words)
(3) technical words (2,000 words)
(4) low-frequency words (123,200 words)

このうち，university word list は，大学で学ぶために必要な語彙である．難易度セグメント間の頻出語集合の差を調べることにより，このような語の候補も抽出できる可能性がある．

8. まとめ

本論文では，現代日本語書き言葉均衡コーパスに対して実施した，テキストの難易度と語の分布に関する一連の調査の結果について報告した．判明した主要な事実は，5–7 節のそれぞれの節の最後に示した．

これらの節で示したように，4 節で提示した仮説のいくつかは，成立しない．その主要な原因は，おそらく，難易度というものに対する一次元的見方にあると考えられる．我々は，難易度を低位から高位までの一次元のスケールとみなす．しかしながら，難易度低位から中位と，難易度中位から高位では，支配的な現象が異なり，同じように「難易度に差がある」とみなしてはいけないのであろう．多少の飛躍が許されるのであれば，次のように言うこともできよう．

> 難易度が低位のテキストは，意図的にそのように書かれているのに対し，難易度が高位のテキストは，それを意図的して書かれてはいない．

本研究で判明した事実に基づけば，すくなくとも学習者用の基本語彙表の編纂においては，テキストの難易度を考慮した方がよいと考えられる．具体的には，難易度下位と中位のセグメント (それぞれ，全体の約 23%，約 54%) を用い，難易度上位のセグメント (全体の約 23%) を除外するのがよいであろう．同時に，難易度中位の頻度よりも，難易度下位の頻度を重視する方がよいであろう．

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．本研究は，JSPS 科研費 24300052 の助成を受けて実施した．

参考文献

[1] 佐藤理史：日本語テキストの難易度を測る，名古屋大学（オンライン），入手先 ⟨http://kotoba.nuee.nagoya-u.ac.jp/sc/obi2/⟩（参照 2013-07-29）．
[2] 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（オンライン），入手先 ⟨http://www.ninjal.ac.jp/corpus_center/bccwj/⟩（参照 2013-08-09）．
[3] 佐藤理史：均衡コーパスを規範とするテキスト難易度測定，情報処理学会論文誌，Vol. 52, No. 4, pp. 1777–1789 (2011).
[4] 佐藤理史，柏野和佳子：テキストの難易度に対する人間の判断と機械の判断，第 1 回コーパス日本語学ワークショップ予稿集，pp. 195–202 (2012).
[5] 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).
[6] 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)，技術報告 LR-CCG-20-05-01，国立国語研究所 (2011).
[7] 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)，技術報告 LR-CCG-20-05-02，国立国語研究所 (2011).
[8] 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).
[9] Nation, I. S. P.: *Teaching and Learning Vocabulary*, Heinle & Heinle Publishers (1990).

表 11 難易度 E の LUW 頻出語 4000 語で、全体の頻出語 4000 語に含まれないもの (一部)

語彙素	読み	語種	品詞	頻度順位			
				All	Easy	Moderate	Difficult
旦那	ダンナ	外	名詞-普通名詞-一般	4,202	1,545	14,897	182,929
水気	ミズケ	和	名詞-普通名詞-一般	4,504	1,560	38,978	64,849
大根	ダイコン	漢	名詞-普通名詞-一般	4,202	1,749	9,056	23,068
ソファ	ソファ	外	名詞-普通名詞-一般	4,271	1,831	7,637	64,849
居間	イマ	和	名詞-普通名詞-一般	4,157	1,843	7,310	19,180
御願い	オネガイ	和	名詞-普通名詞-一般	4,824	2,053	8,835	64,849
昨夜	サクヤ	漢	名詞-普通名詞-一般	4,603	2,088	6,898	182,929
髪の毛	カミノケ	和	名詞-普通名詞-一般	4,379	2,340	5,254	23,068
そっち	ソッチ	和	代名詞	4,157	1,681	8,835	182,929
私達	アタシタチ	和	代名詞	5,195	1,749	114,293	182,929
あっち	アッチ	和	代名詞	5,195	2,199	9,056	182,929
呆然	ボウゼン	漢	形状詞-タリ	6,007	2,676	9,890	64,849
平然	ヘイゼン	漢	形状詞-タリ	6,719	3,466	8,170	64,849
気の毒	キノドク	混	形状詞-一般	4,077	2,240	4,551	40,224
滑らか	ナメラカ	和	形状詞-一般	4,688	2,240	7,474	11,642
平気	ヘイキ	漢	形状詞-一般	4,309	2,284	4,982	64,849
親切	シンセツ	漢	形状詞-一般	4,179	2,638	4,054	40,224
些とも	チットモ	和	副詞	4,329	2,037	6,055	182,929
こっそり	コッソリ	和	副詞	4,346	2,053	6,055	182,929
さっぱり	サッパリ	和	副詞	4,379	2,102	5,847	182,929
けど	ケド	和	接続詞	4,688	1,733	16,561	182,929
だったら	ダッタラ	和	接続詞	4,478	1,843	9,343	64,849
あれ	アレ	和	感動詞-一般	4,567	1,800	11,259	182,929
ふうん	フウン	和	感動詞-一般	5,263	1,964	16,561	182,929
もしもし	モシモシ	和	感動詞-一般	8,096	3,254	17,506	182,929
歩き出す	アルキダス	和	動詞-一般	4,250	1,749	8,577	182,929
覚ます	サマス	和	動詞-一般	4,102	1,800	7,040	64,849
締まる	シマル	和	動詞-一般	4,229	1,903	7,040	23,068
呆れる	アキレル	和	動詞-一般	4,102	1,950	5,754	64,849
焼ける	ヤケル	和	動詞-一般	4,102	2,053	5,189	64,849
冷える	ヒエル	和	動詞-一般	4,447	2,088	6,898	19,180
情けない	ナサケナイ	和	形容詞-一般	4,127	2,017	5,421	64,849
悔しい	クヤシイ	和	形容詞-一般	4,379	2,199	5,491	64,849
眩しい	マブシイ	和	形容詞-一般	5,263	2,390	7,972	64,849
荒い	アライ	和	形容詞-一般	5,416	2,448	8,170	182,929
臭い	クサイ	和	形容詞-一般	5,377	3,072	6,160	19,180
わけにはいかない	ワケニハイカナイ	和	助動詞	4,688	2,848	5,491	9,132
てらっしゃる	テラッシャル	和	助動詞	7,322	2,965	16,561	64,849
こともない	コトモナイ	和	助動詞	4,398	2,999	4,212	16,419
たら	タラ	和	助詞-副助詞	4,540	1,800	10,866	182,929
にしては	ニシテハ	和	助詞-接続助詞	5,857	3,416	6,055	182,929
たところで	タトコロデ	和	助詞-接続助詞	4,965	3,526	4,982	9,811