現代日本語書き言葉均衡コーパスのサンプル長と言語的特徴

－固定長サンプルと可変長サンプルの質的な違い－

山崎誠　丸山岳彦　柏野和佳子　佐野大樹　秋元祐哉　稲益佐知子　田中弥生　大矢内夢子

{yamazaki,maruyama,waka,toki.sano,akimoto,inamasu,yayoi,yume_o}@kokken.go.jp

独立行政法人　国立国語研究所

1. はじめに

国立国語研究所が中心となって2006年度から構築している「現代日本語書き言葉均衡コーパス」（Balanced Corpus of Contemporary Written Japanese、以下BCCWJと略す）に格納されるサンプルには、固定長サンプルと可変長サンプルという2種類の異なるサンプルのタイプがある。この2つのタイプは言語的に見てどのような違いがあるか、文字及び語彙の面から計量的に分析する。

2. 固定長サンプルと可変長サンプル

固定長のサンプルは、統計的に厳密な分析を目的として設計されている。1サンプルは句読点などの記号類を除く1,000字から構成される。1,000字はBCCWJで採用している短単位に換算するとおよそ590語である 。

可変長サンプルは文章の長さではなく、内容的なまとまりを基準にして１サンプルの範囲を決定する。具体的には新聞・雑誌の１記事、書籍における章・節などのまとまりが１サンプルに該当する。ただし、無制限に長いサンプルが出来るとコーパスの分析に影響を与えるため上限を１万字としている。

３．データ

「『現代日本語書き言葉均衡コーパス』モニター公開データ（2008 年度版）」[1]に収録された書籍及び白書のサンプルのうち固定長サンプルと可変長サンプルの両方があるものを対象とした。具体的には以下のとおり。

書籍　3,773 サンプル[2]

白書　1,500 サンプル

分析に使用した形態素解析環境は、MeCab ver.0.97＋UniDic-1.3.9 である。なお、本稿では形態素解析の結果得られた短単位を便宜上「語」とみなして分析を行う。また、特に断らない限り分析には助詞・助動詞を含み、記号・符号を含まないこととする。品詞体系は UniDic に従う。

1 BCCWJ に格納するサンプルのうち、著作権処理の済んだものを配布している。詳しくは以下のサイトを参照。
http://www.kokken.go.jp/kotonoha/ex_8.html

2 候補となるサンプルは 3,795 サンプルあったが、後述の separated パターンの 22 サンプルは比較の条件を備えていないため分析の対象としなかった。

4. 固定長サンプル、可変長サンプル全体の比較

表１に対象となった各サンプル全体の延べ語数及び異なり語数と個々のサンプルごとの延べ語数及び異なり語数の平均値を示した。

表１ 固定長サンプルと可変長サンプルの概観

	書籍		白書	
	固定長	可変長	固定長	可変長
全体延べ語数	2,455,558	10,217,488	1,035,345	4,685,128
全体異なり語数	52,550	82,800	15,780	26,748
個別延べ語数(n)	647.2	2671.9	690.2	3213.4
個別異なり語数(k)	247.2	638.1	225.0	534.8
n/k値の平均	2.66	3.89	3.14	5.55

個々のサンプルの値で見ると、可変長サンプルは書籍で固定長サンプルの約4.1倍、白書で4.6倍の長さを持つ。延べ語数では固定長・可変長ともに書籍より白書の方が値が大きいが、異なり語数では逆に書籍の方が大きな値になっている。これは、白書の方が延べ語数の伸びに比べて異なり語数の伸びが鈍いことを意味している。1 語あたりの平均使用度数を表す n/k 値（延べ語数/異なり語数

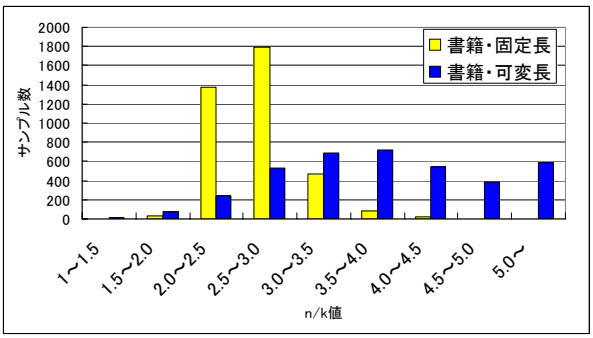


図１ n/k 値の分布（書籍）

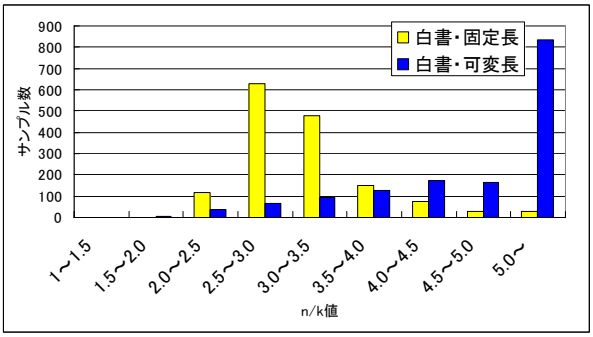


図２ n/k 値の分布（白書）

の値)も固定長・可変長ともに白書の方が大きくなっていることもそのことの表れである。固定長サンプルと可変長サンプルにおけるn/k値の分布のようすを図1、図2に示した。

次に文字種・語種・品詞の割合を見てみよう。表2〜4を見ると分かるように、書籍・白書ともに固定長サンプルと可変長サンプルの差異は無視できるほど小さい。これらのカテゴリの使われ方はサンプルの長さに影響を受けにくいと言える。

表2 固定長・可変長サンプルにおける文字種の割合

	書籍		白書	
	固定長	可変長	固定長	可変長
英数字	0.61	0.66	1.24	1.24
平仮名	59.19	59.21	38.66	38.87
片仮名	5.88	5.82	4.77	4.77
漢字	34.31	34.31	55.33	55.12

表3 固定長・可変長サンプルにおける語種の割合(延べ)

	書籍		白書	
	固定長	可変長	固定長	可変長
和語	72.28	72.32	46.37	46.13
漢語	21.90	21.83	48.36	48.59
外来語	1.87	1.86	2.21	2.18
混種語	0.92	0.93	1.45	1.47
固有名	2.57	2.60	1.49	1.51
不明	0.16	0.16	0.06	0.06
なし	0.30	0.30	0.07	0.07

表4 固定長・可変長サンプルにおける品詞の割合(延べ)

	書籍		白書	
	固定長	可変長	固定長	可変長
名詞	31.19	31.14	47.68	47.90
代名詞	1.64	1.64	0.27	0.27
動詞	14.61	14.60	10.71	10.65
形容詞	1.60	1.61	0.49	0.48
形状詞	0.98	0.97	0.60	0.60
連体詞	1.06	1.07	0.71	0.71
副詞	1.84	1.84	0.46	0.45
接続詞	0.44	0.44	0.95	0.95
感動詞	0.24	0.24	0.01	0.01
助詞	31.55	31.57	24.31	24.20
助動詞	10.05	10.07	4.26	4.25
接頭辞	0.74	0.73	1.25	1.25
接尾辞	4.07	4.07	8.29	8.29

表5 上位語の比較(書籍)

	固定長			
順位	語彙素	表記	品詞	使用率
1	ノ	の	格助詞	0.05003
2	ニ	に	格助詞	0.03615
3	テ	て	接続助詞	0.03587
4	ハ	は	係助詞	0.03408
5	ダ	だ	助動詞	0.03354
6	タ	た	助動詞	0.03229
7	ヲ	を	格助詞	0.03178
8	ガ	が	格助詞	0.02441
9	スル	為る	動詞	0.02439
10	ト	と	格助詞	0.02362
11	モ	も	係助詞	0.01289
12	デ	で	格助詞	0.01235
13	イル	居る	動詞	0.01218
14	アル	有る	動詞	0.01105
15	ノ	の	準体助詞	0.01091
16	イウ	言う	動詞	0.00883
17	コト	事	名詞	0.00820
18	ナイ	ない	助動詞	0.00662
19	レル	れる	助動詞	0.00640
20	マス	ます	助動詞	0.00639
21	ナル	成る	動詞	0.00601
22	デス	です	助動詞	0.00517
23	ナイ	無い	形容詞	0.00511
24	カラ	から	格助詞	0.00475
25	ソノ	其の	連体詞	0.00443
26	ヨウ	様	形状詞	0.00409
27	ガ	が	接続助詞	0.00369
28	カ	か	副助詞	0.00339
29	ソレ	其れ	代名詞	0.00324
30	イチ	一	名詞	0.00314

	可変長			
語彙素	表記	品詞	使用率	
ノ	の	格助詞	0.05028	
ニ	に	格助詞	0.03620	
テ	て	接続助詞	0.03597	
ハ	は	係助詞	0.03446	
ダ	だ	助動詞	0.03367	
タ	た	助動詞	0.03279	
ヲ	を	格助詞	0.03192	
ガ	が	格助詞	0.02444	
スル	為る	動詞	0.02433	
ト	と	格助詞	0.02360	
モ	も	係助詞	0.01291	
イル	居る	動詞	0.01234	
デ	で	格助詞	0.01219	
アル	有る	動詞	0.01097	
ノ	の	準体助詞	0.01096	
イウ	言う	動詞	0.00884	
コト	事	名詞	0.00812	
ナイ	ない	助動詞	0.00679	
レル	れる	助動詞	0.00635	
マス	ます	助動詞	0.00607	
ナル	成る	動詞	0.00595	
ナイ	無い	形容詞	0.00522	
デス	です	助動詞	0.00505	
カラ	から	格助詞	0.00467	
ソノ	其の	連体詞	0.00447	
ヨウ	様	形状詞	0.00410	
ガ	が	接続助詞	0.00362	
カ	か	副助詞	0.00343	
ソレ	其れ	代名詞	0.00330	
カ	か	終助詞	0.00313	

表5に示したのは書籍の固定長・可変長のそれぞれにおける使用頻度の上位語30語の比較であるが、順位・使用率ともにほぼ同じである。表は掲載しないが、白書においても同様である。

固定長サンプルと可変長サンプルの範囲が異なるため、固定長サンプルにしか出現しない語あるいは可変長サンプルにしか出現しない語が存在する。表6、表7はそれらを異なり語数レベル、延べ語数レベルで集計したものである。

表6 固定長・可変長の一方にしか出現しない語(異なり)

	書籍		白書	
	語数	割合	語数	割合
固定長のみ	1,963	3.74	416	2.64
可変長のみ	32,214	38.91	11,384	42.56

表7 固定長・可変長の一方にしか出現しない語(延べ)

	書籍		白書	
	語数	割合	語数	割合
固定長のみ	2,649	0.11	532	0.05
可変長のみ	79,170	0.77	27,682	0.59

固定長サンプルにしか出現しない語は、使用率の高い順に書籍で「滌除、*トワダ、*ナミエ、*クツナ、*兼六、*サクノスケ、スピリチュアリズム、スウェデン、邢、黄斑、*キリコ、tell」(頻度7以上)、白書では「スラッジ、*マレ、Publication、主査、*セントルシア、Drugs、サージ、*ダルフール、船溜、*アンティグア、稲叢、嗅覚、錘数、線分、*全労連、*テーラー、*フィゲーレス、領収、Dangerous」(頻度3以上)である。*を附した語は形態素解析結果において固有名詞となっているものである。

同様に、可変長サンプルにしか出現しない語は、書籍で「*コウダユウ、蘇芳、シリウス、春雨、山車、*ジュネーブ、*足羽、*タカツネ、ストーマ、フルート、コウスケ、水虫、フランシーヌ、ダンベル、体節、景勝、思量、夕霧、*ライ

ル」(頻度 36 以上)、白書では「燻蒸、県庁、通、氏名、着陸、*JAS、通数、パン-pao、字幕、同感、定係、*JIS、商船、方位、ボーナス、ミスマッチ」(頻度 30 以上)である。

両者を比較すると、可変長サンプルの方によりなじみのある語が含まれているようであるが、今後サンプル数が増えてくれば、可変長サンプルにしか出現しない語も固定長サンプルにしか出現しない語と同様の傾向を示すのではないかと思われる。

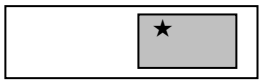
5. 固定長サンプルと可変長サンプルの位置的関係

5. 1. パターン

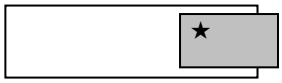
4節で見たように、固定長サンプルと可変長サンプルは文字種、品詞、語種の割合や上位語の使用率においてほぼ同じであることが確認できたが、その最大の理由は、両者は包含関係にあるものが多いということだろう。以下その事情を説明する。

コーパス構築に当たって固定長と可変長のサンプルを別々に取得するのは作業コストがかかりすぎるため、BCCWJ では1回のサンプリングで当たった同一箇所から固定長と可変長の2つのサンプルを取得している。そのため、固定長サンプルと可変長サンプルの間には包含関係を基本とする3種類の「パターン」が生じる。以下の図3に示す included, overflow, separated である。

included : 固定長サンプルが可変長サンプルに包含される場合



overflow : 固定長サンプルの一部が可変長サンプルからはみ出す場合



separated : 固定長サンプルが可変長サンプル包含されない場合

図 3 固定長サンプルと可変長サンプルの関係

included は、固定長サンプルが可変長サンプルの中に完全に納まる場合である。図 3 の★は、ランダムに決められる「サンプル抽出基準点」を表すが、この文字を含む後続の 1,000 字が可変長サンプルの終端位置を超える場合は、overflow になる。separated は、2. で述べた可変長サンプルの長さの制限(1万字)のため、強制的に打ち切った可変長サンプルの終端が固定長サンプルの開始位置に届かなかった場合である。

今回の対象データにおけるパターンの分布は表 8 のとおりである。

表 8 サンプルのパターンの分布

	included	overflow	separated
書籍	2,319	1,454	22
白書	1,018	482	0

5. 2. 固定長サンプルの開始位置と言語的特徴

固定長サンプルの大多数は可変長サンプルの中にサンプルの開始位置を持つ。開始位置を決める「サンプル抽出基準点」はランダムに当てているため、固定長サンプルの開始位置は可変長サンプル内に均等に分布しているはずである。そのことを確認したのが図 4 である。図 4 は可変長サンプルをその長さにかかわらず 10 等分してその 10 個の区画のどこに固定長サンプルの開始位置が来るかを調べたものである。

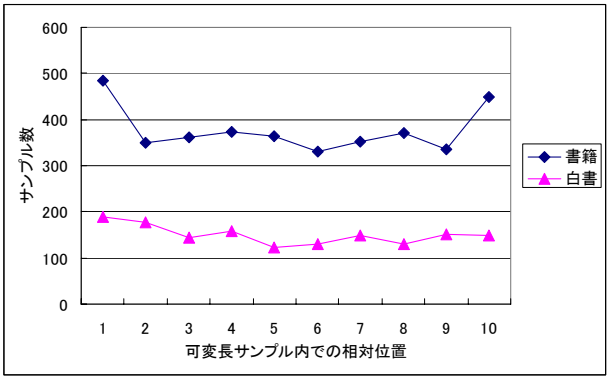


図 4 可変長サンプル内での相対位置で示した固定長サンプルの開始位置の分布

固定長サンプルが可変長サンプル内のどの位置から開始するかによって、固定長サンプルの言語的特徴に影響があると考えられる。例えば、上述の overflow パターンの場合、固定長サンプルには、一つの完結したまとまりの最後の部分とその次のまとまりの最初の部分とから構成されることになる。そのような状況で影響を受ける可能性がある指標は n/k 値である。同一の内容よりも異なる2つの内容の方が異なり語数を増やしやすいからである。そのことを確かめるために、可変長サンプル内での相対的位置と n/k 値との関連を調査した。

結果を図5、図6に示す。書籍、白書ともに、可変長サンプル内での相対位置が後ろの方になるにつれて、n/k 値

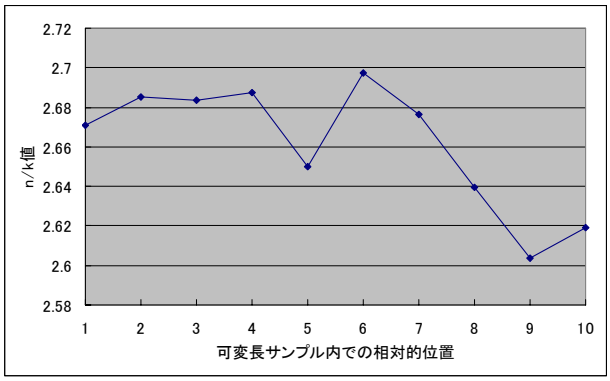


図 5 可変長サンプルにおける相対位置による固定長サンプルの n/k 値の分布(書籍)

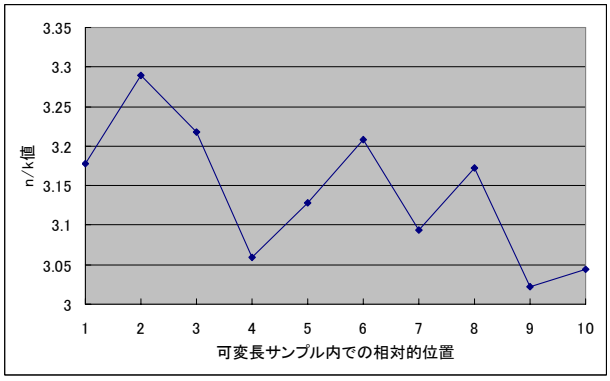


図 6 可変長サンプルにおける相対位置による固定長サンプルの n/k 値の分布(白書)

が低くなる傾向がある。白書では中間でいったん n/k 値がかなり下がるところがあるがどういう理由によるものかは分からないが、n/k 値の変動は語彙で表される同一の話題がどれだけの長さ継続するかということとも関連する。したがって、書籍、白書ともに途中で値が下がっているということは、可変長サンプルの中間付近でそれ以前とは異なる語彙が多く出現しているということを示唆する。また、

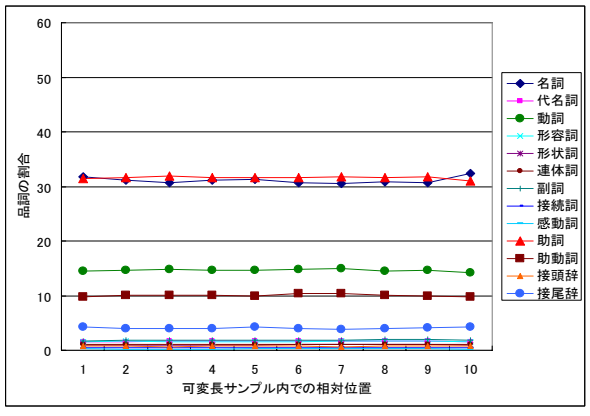


図 7 可変長サンプルにおける相対位置による固定長サンプルの品詞の割合の分布(書籍)

書籍も白書も最後の区画で n/k 値のわずかの上昇が認められるが、これはこの区画においては固定長サンプルの分量のほとんどが開始位置を持つまとまりの次のまとまりの中にあるということになり、2 つの異なる内容が同居する度合いが小さくなったためと思われる。

同様に品詞の割合を書籍、白書で調べたが、可変長サンプル内の相対的位置との関連は見いだせなかった(図 7、図 8)。

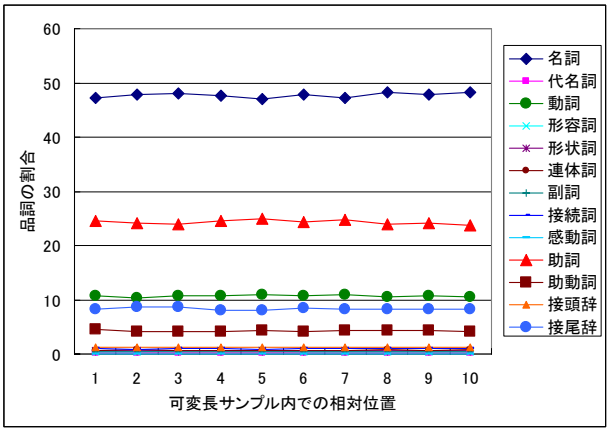


図 8 可変長サンプルにおける相対位置による固定長サンプルの品詞の割合の分布(白書)

6. まとめ

固定長サンプル、可変長サンプルは、文字種、品詞、語種等のマクロな値を見る限りではほぼ同じであり、等質なテキストであると言える。ただし、固定長サンプルの開始位置が可変長サンプルの末尾付近に当たる場合は、延べ語数と異なり語数との関係に変化が見られ、1語あたりの平均使用度数が低くなる傾向があることが分かった。BCCWJ を十分理解して使うために、今後様々な方法でサンプルの検証を行っていくことが必要である。

[**謝辞**] 本研究は、文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築:21 世紀の日本語研究の基盤整備」(平成 18～22 年度, 領域代表者:前川喜久雄)による補助を得た。

[**参考文献**]

柏野和佳子他(2009)「『現代日本語書き言葉均衡コーパス』のサンプル収録方法」、言語処理学会第 15 回年次大会論文集.

丸山岳彦・秋元祐哉(2007)『『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法―現代日本語書き言葉の文字数調査―』(LR-CCG-06-02).

丸山岳彦・秋元祐哉(2008)『『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法(2) ―コーパスの設計とサンプルの無作為抽出法―』(LR-CCG-07-01).