言語単位と文の長さが品詞比率に与える影響

山崎　誠（国立国語研究所言語資源研究系）†

Influence of Word Unit and Sentence Length on the Ratio of Parts of Speech

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)

１．はじめに

語類構成比の問題は語彙の分布問題の重要なテーマとして 1950 年代から研究が行われているが、大野(1956)、樺島(1954)(1955)などの初期の研究以降、目立った進展がないようである。水谷(1983)における語類構成比の記述と計量国語学会編(2009)の品詞構成比率の記述(pp.95-96)の間に進歩が見られないことはこの分野の研究が停滞していることを示していると言ってよいだろう。

２．品詞構成比

語類構成比について唯一研究が行われているのが品詞の比率である。テキストにおける品詞構成比を異なり語数のレベルで分析し定式化につながる傾向を見いだしたのが大野(1956)であり[1]、延べ語数のレベルで分析し定式化したのが樺島(1954)(1955)である。大野は古典、樺島は現代文という違いはあるものの、どちらも品詞の比率がテキストのジャンルによって異なり、一定の法則が見いだせるということを指摘している。

本稿はこれまで十分検討されてこなかった、文の長さが品詞比率とどう関わっているか[2]という点について調査したものである。日本語では文の長さは文字数で計測することが多いが、本稿では操作的に作成した言語単位によって計測する。具体的には『現代日本語書き言葉均衡コーパス』（以降、BCCWJ と略す）で用いられている短単位(SUW)及び長単位(LUW)である。この 2 つの単位の違いにより、品詞構成比がどのように違ってくるかということも併せて調査する。

３．データと方法

本稿では、BCCWJ の DVD 版に納められている M-XML ファイルを利用した。M-XML ファイルには<sentence>というタグが埋め込まれており、これを元に文の候補となるデータ抽出した[3]。さらに<sentence>タグの中に含まれる<suw>タグで短単位の情報を、<luw>タグで長単位の情報を抽出した。

品詞の認定は BCCWJ で用いられている UniDic の品詞体系に従った。なお、本稿では樺島(1954)(1955)(1979)に基づいて品詞を類別する[4]が、その際の基準は接尾辞を除き、UniDic の大分類を利用した。接尾辞については UniDic の中分類までの情報を利用した。品詞分類

† yamazaki@ninjal.ac.jp

1 いわゆる「大野の法則」を数学的に定式化したのは水谷(1965)(1981)である。

2 樺島(1979:218)に「名詞の比率が大きくなるにつれて、文の長さの平均値も次第に大きくなっているようである。」という指摘がある。

3 小木曽他(2011)によれば、「サンプルは sentence の集合として捉えられる」ということなので、抽出の際に漏れた部分は存在しない。

4 類別した品詞の各類の名称は樺島(1954)では、名詞、動詞、形・形動・副・連体詞、感動・接続詞、樺島(1955)では略称を用いて、N、V、Ad、I、樺島(1979)では、N、V、M、I となっている。いずれも 4 類に分類し、所属する品詞は同じである。本稿では樺島(1979)の名称を用いる。

の詳細を表 1 に示す。短単位、長単位ともこの基準を用いた。本稿の分析ではことわりのない限り、品詞の属性が「空白」および句読点などの「補助記号」を除外している。

表 1　本稿での品詞の類別

分類	UniDic の品詞
M（名詞類）	名詞
	代名詞
	接尾辞－名詞的
	記号[5]
V（動詞類）	動詞
	接尾辞－動詞的
M（形容詞・形状詞・副詞類）	形容詞
	形状詞
	副詞
	連体詞
	接頭辞
	接尾辞－形容詞的
	接尾辞－形状詞的
I（接続詞・感動詞類）	接続詞
	感動詞
P（助詞・助動詞類）	助詞
	助動詞
O（その他）	（その他）[6]

４．文の長さの分布

４．１　全体について

BCCWJ 全体における<sentence>タグで定義された「文」の数は 5,515,952 個であった。ただし、この全てがいわゆる通常の文というわけではない。間淵(2011:237)が指摘しているように、<sentence>は自動的に付与されているため、「文と認定されるべき要素であっても、sentence と認定されないものがある」または「文には相当しない sentence や，明らかに文中である不適切な位置で sentence が分断されているものがある」からである。そのことはさておき、上記の「文」について、長さの分布を見てみよう。

表 2　文の長さの分布（<sentence>タグ全体）

言語単位	最小値	第 1 四分位数	中央値	平均値	第 3 四分位数	最大値
短単位(SUW)	1.0	113.2	225.5	275.0	352.5	5527.0
長単位(LUW)	1.0	89.8	178.5	232.0	283.5	5019.0

[5] この「記号」は、「M氏」の「M」や「Ｆ１５」の「F」のような部分に付けられた品詞名である。機能的に名詞として位置付けられるとしてここに置いた。

[6] その他に該当するのは、品詞欄に以下の属性を持つものである。各属性の後ろの括弧内の数字は、(短単位/長単位)の順に該当する言語単位数を示した。「言いよどみ」(2931/318)、「web 誤脱」(173/166)、「英単語」(2931/3093)、「カタカナ文」(65450/44093)、「漢文」(551/555)、「ローマ字文」(21/0)、「新規未知語」(2/0)、「方言」(236/232)、「未知語」(194825/194756)、「URL」(18465/18232)。なお、品詞欄が空欄のものが長単位に 1 つあったが、これも（その他）に含めている。（その他）の占める割合は、短単位・延べ語数で約 0.24%、長単位・延べ語数で約 0.27%である。この数字は補助記号、空白も含めた全ての言語単位に占める割合である。

短単位、長単位とも最大値となっているのは、ジェイムズ・ジョイスの『ユリシーズ』からのサンプル（LBr9_00057、丸谷才一他訳、集英社、2003 年刊）で、この小説の特異な文体の特徴であり、かなり例外的なものであろう[7]。文の長さと文数の関係を図 1 に示した。

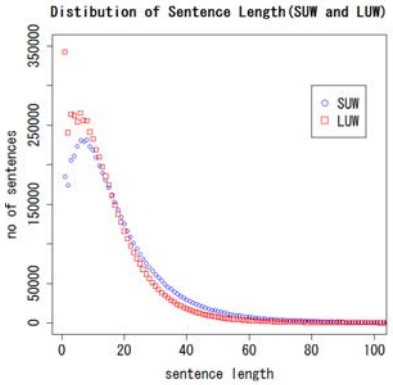


図 1　文の長さの分布

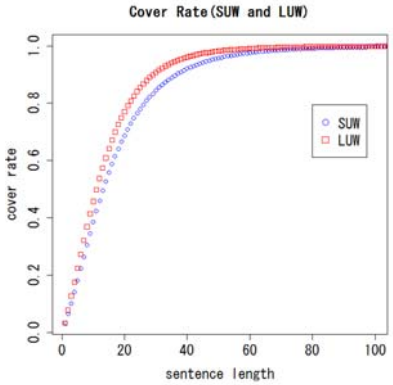


図 2　カバー率

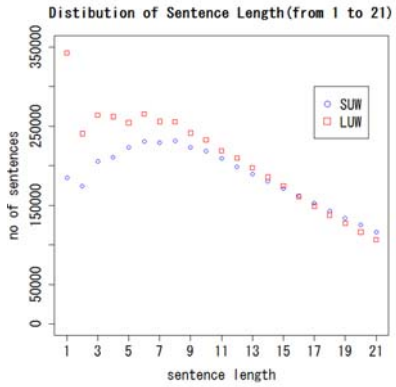


図 3　文の長さの分布（1～21）

図 1 から、文の長さが短い値を示す部分、とくに 1～10 の付近が変則的な動きになっていることが見て取れる。

図 2 のカバー率から、短単位は文の長さが 37 語で全体の 90%をカバー、長単位は 29 語で全体の 90%をカバーしていることが分かる。

図 3 は、文の長さが 20 語以下の分布を示したものである。これを見ると、短単位、長単位ともに文の長さが 1 から 2 にかけて下降、2 から上昇に転じている。短単位では、文の長さが 3～5 にかけて下降し、その後 8 まで上下を繰り返している。長単位ではやはり 8 まで上昇し、そこから下降に転じている。この不規則な動きの原因は後述のように、文の認定にある。

４．２　通常の文について

前節では<sentence>タグで認定された文全体を対象としたが、これには新聞の見出しや書籍の章のタイトル、図表のキャプションのような、通常は文とみなさないものも含まれる。このような「通常でない文」を排除するために、<sentence>タグで囲まれた文の末尾[8]の要素の語彙素が「。」「.」「！」「？」「」」で終わるものを本稿で言う「通常の文」と見なして抽出した。「」」以外は、間淵(2011:236)によると、<sentence>タグの自動認定の際の文終止マーカーである。「」」をそこに加えたのは、小説などの会話文では「」」で終わるものが多いのではないかと予想したからである。通常の文は 4,374,273 個であり、<sentence>でタグ付けされた文の約 80%である。

表 3 は「通常の文」の長さの分布を示したものである。表 2 の<sentence>タグ全体に比べると第 1 四分位数～最大値の値が低くなっていることが分かる。

---

[7] 短単位で 2 番目に長い文（文長 3299）および長単位で 2 番目に長い文（文長 2917）も『ユリシーズ』からのサンプルである。

[8] 品詞の属性が「空白」と「補助記号」を除いた末尾の要素を指す。

表 3　文の長さの分布（「通常の文」）

言語単位	最小値	第 1 四分位数	中央値	平均値	第 3 四分位数	最大値
短単位(SUW)	1.0	105.0	209.0	242.2	323.0	3279.0
長単位(LUW)	1.0	84.5	168.0	204.1	261.5	2917.0

図 4 は通常の文について、文の長さの分布を示したものである。図 1 と違って文の長さが少ないあたりのカーブがなだらかになっていて不規則性が解消されていることが分かる。図 5 は通常の文のカバー率である。カバー率が 90%に達する文の長さは図 3 と比べると若干長くなり、短単位で 40 語、長単位では 31 語である。

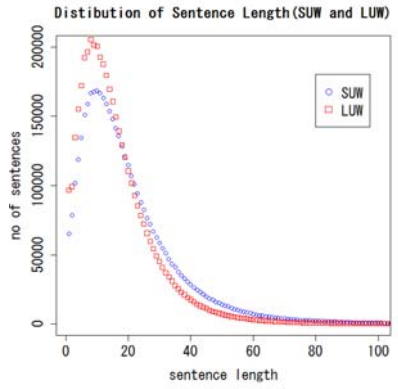


図 4　文の長さの分布（通常の文）

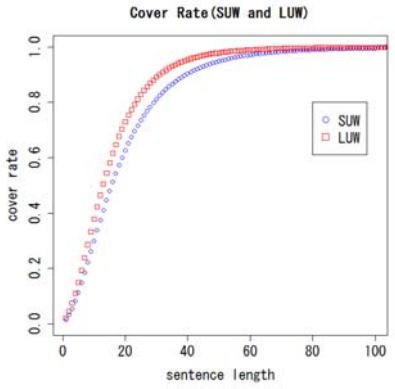


図 5　カバー率（通常の文）

４．３　通常でない文について

表 4 は、「通常でない文」すなわち、「。．！？」」で終わっていない文の長さの分布、図 6 は「通常でない文」における文の長さの分布、図 7 はそのカバー率である。ここには、見出し相当の語句などが含まれるため、文の長さが短い方に偏っていることが分かる。また、図 6 の分布の形も図 4 と異なる形になっている。90%のカバー率になるのは短単位で 18 語、長単位で 13 語であり、こちらも前記 2 つのカバー率よりも急峻であることが分かる。

表 4　文の長さの分布（「通常でない文」）

言語単位	最小値	第 1 四分位数	中央値	平均値	第 3 四分位数	最大値
短単位(SUW)	1.0	72.3	143.5	194.1	217.8	5527.0
長単位(LUW)	1.0	53.0	105.0	160.2	164.0	5019.0

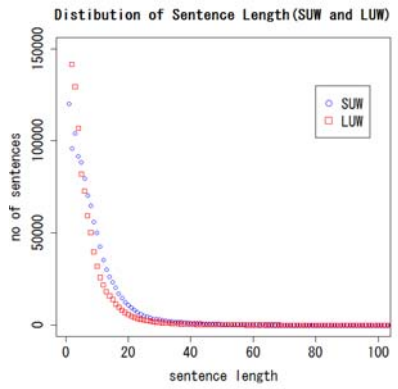


図 6　文の長さの分布（通常でない文）

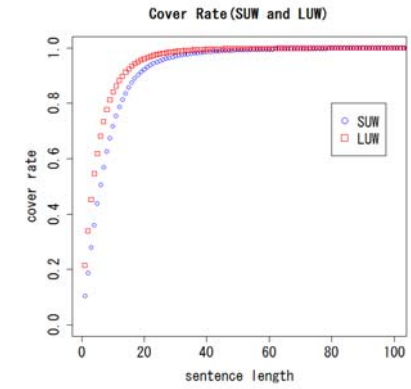


図 7　カバー率（通常でない文）

４．４　文末の記号別の分布

本稿では、「。．！？」」を「通常の文」の文末を示す記号として文の抽出を行った。その各記号別に文の長さの分布を見てみよう。表 5 は、「通常の文」に占める各文末の記号の数である。また、表 6、表 7 に分けて言語単位別の分布の様子である。最小値と最大値を除いて文の長さは短単位も長単位も＜。　」　．　？　！＞の順になっていることが分かる。

表 5　文の長さの分布

文末の記号	。	.	!	?	」
文数	3643999	62346	118680	163037	386211
割合	83.30	1.43	2.71	3.73	8.83

表 6　文末の記号別の文の長さの分布（短単位）

文末の記号	最小値	第 1 四分位数	中央値	平均値	第 3 四分位数	最大値
。	1.0	102.5	204.0	238.3	313.5	3279.0
.	1.0	45.8	90.5	105.2	136.5	461.0
!	1.0	30.8	60.5	66.1	92.3	182.0
?	1.0	33.0	65.0	73.0	97.0	433.0
」	1.0	51.0	101.0	113.7	158.0	430.0

表 7　文末の記号別の文の長さの分布（長単位）

文末の記号	最小値	第 1 四分位数	中央値	平均値	第 3 四分位数	最大値
。	1.0	82.3	163.5	200.8	252.8	2917.0
.	1.0	36.8	72.5	85.8	114.5	367.0
!	1.0	27.5	54.0	58.5	82.5	150.0
?	1.0	29.5	58.0	65.7	86.5	403.0
」	1.0	44.3	87.5	97.5	131.8	354.0

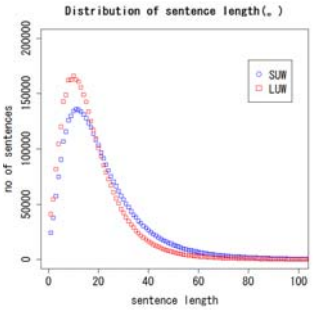


図 8 文の長さの分布（。）

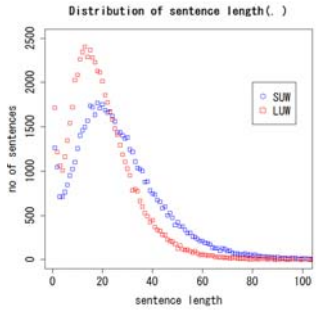


図 9 文の長さの分布（.）

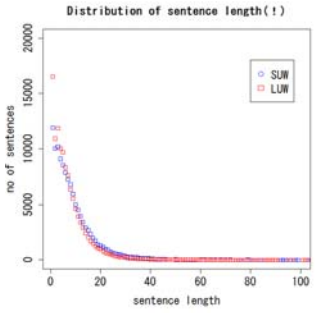


図 10 文の長さの分布（！）

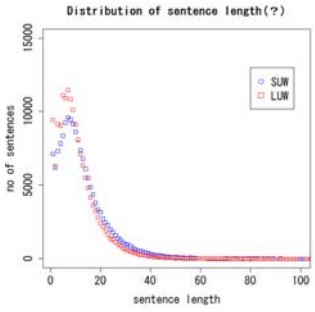


図 11 文の長さの分布（？）

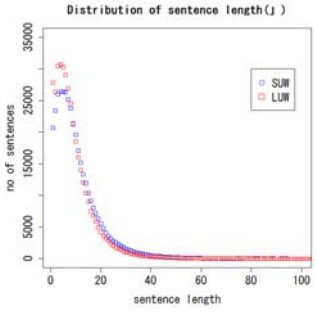


図 12 文の長さの分布（」）

図 8～図 12 は各文末記号別に文の長さの分布を観察したものである。図 10 の感嘆符の場合が他と異なる分布になっていること、また、図 9 のピリオドの場合、短単位と長単位の差が大きいが、感嘆符、疑問符、カギ括弧の場合は短単位と長単位の差が少ないことが分かった。これらは具体的にどのような文が用いられているかを見ないとこれらの違いの言語学的な解釈ができない。今回はあくまで現象の指摘にとどまる。

５．文の長さと品詞類の比率

５．１　全体の傾向

文の長さにより品詞類の比率がどのように変化するか（しないか）を観察した。品詞類は表 1 に挙げたように、N（名詞類）、V（動詞類）、M（形容詞・形状詞・副詞類）、I（接続詞・感動詞類）、P（助詞・助動詞類）の 5 つに分類した。P 以外は樺島(1954)以降の一連の分類と同じものである。樺島の分類には助詞・助動詞を含めていないが、延べ語数の水準での品詞類の構成比を見るには助詞・助動詞は言語量が多いため重要な要素であると考え、考察の対象とすることにした。表 8 は BCCWJ 全体の「通常の文」における各品詞類の割合である。短単位と長単位の違いとしては、N の比率が長単位では相対的に低く、P の比率が相対的に高くなることが挙げられる。V、M、I の比率は短単位でも長単位でもあまり変わらない。表 9、表 10 はそれぞれ短単位、長単位での文末記号別の品詞類の比率である。短単位では N と P の変動が大きいが、長単位ではどの品詞類も変動が相対的に小さくなる。また、句点で終わる文に比べて疑問符、かぎ括弧で終わる文で、N の割合が低く、P の割合が多くなっているのは話し言葉的な要因が関係している可能性がある。

表 8　品詞類の比率（通常の文）

	N	V	M	I	P
短単位	0.376	0.141	0.067	0.006	0.410
長単位	0.297	0.129	0.078	0.010	0.485

表 9　文末の記号別の品詞類の比率（短単位）

文末の記号	N	V	M	I	P
。	0.380	0.142	0.066	0.006	0.407
.	0.460	0.125	0.058	0.006	0.350
!	0.373	0.124	0.090	0.015	0.398
?	0.312	0.124	0.078	0.007	0.480
」	0.300	0.142	0.086	0.013	0.459

表 10　文末の記号別の品詞類の比率（長単位）

文末の記号	N	V	M	I	P
。	0.300	0.130	0.077	0.009	0.484
.	0.351	0.117	0.069	0.010	0.454
!	0.296	0.123	0.101	0.022	0.458
?	0.255	0.120	0.088	0.011	0.527
」	0.254	0.133	0.093	0.020	0.500

図 13、図 14 はそれぞれ短単位、長単位における文の長さによる品詞類の比率の推移を示したものである。ここでは 4．2 節で採り上げた「通常の文」を対象としている。図 13、図 14 からは、文の長さが 10 を超えたあたりから各品詞類の比率が一定化することが見て取れる。ただし、短単位では長さ 43 から N（名詞類）の比率が P（助詞・助動詞類）の比率を上回る。また、特徴的なのは文の長さ 1 における分布で、短単位では I(0.3997)、N(0.342)、

M(0.1645)、P(0.0469)、V(0.0468)、長単位では N(0.4974)、I(0.293)、M(0.1275)、V(0.0492)、P(0.033)となっている。括弧内の数字は構成比の値である。I（接続詞・感動詞類）の比率が高いのは、長さ 1 の文にはそれらだけで成り立っているものが多いということを示している。

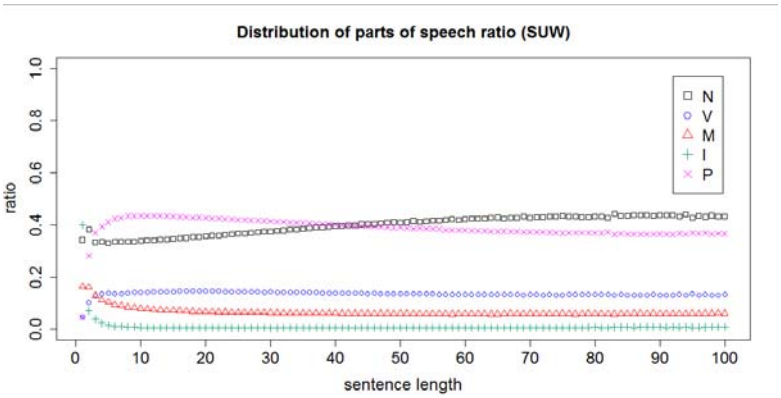


図 13　文の長さによる品詞類の比率の推移（短単位）

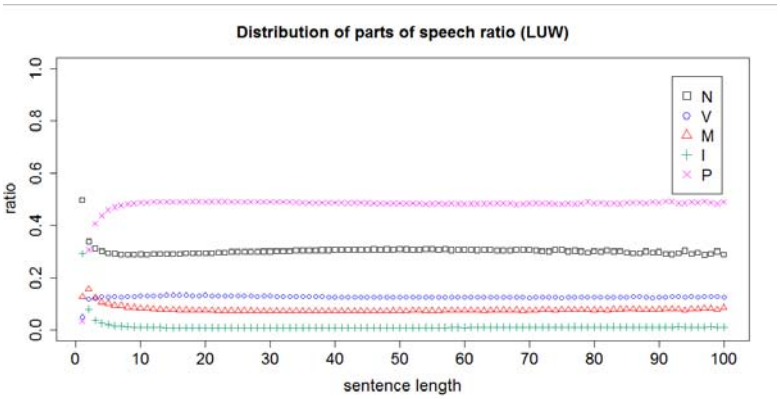


図 14　文の長さによる品詞類の比率の推移（長単位）

５．２　文の長さと品詞類の相関

樺島(1979:218)には、名詞の比率が高くなると文の長さの平均値も大きくなるという指摘があるが、今回のデータで文の長さと各品詞類および各品詞類どうしの相関がどうなっているかを調査した。文の長さが 1～100 について相関を調べた結果を表 11、表 12 に示す。相関係数（ピアソンの積率相関係数）が絶対値で 0.7 以上の部分のセルを網掛けで示したが、文の長さとの相関があったのは、短単位の N（名詞類）のみであった。各品詞類どうしの相関は、短単位で V と P、V と I、I と P の３組、長単位で I と P、N と V、V と I、N と I、N と P、V と P、M と P の 7 組であった。

表 11　文の長さと品詞類との相関行列（短単位）

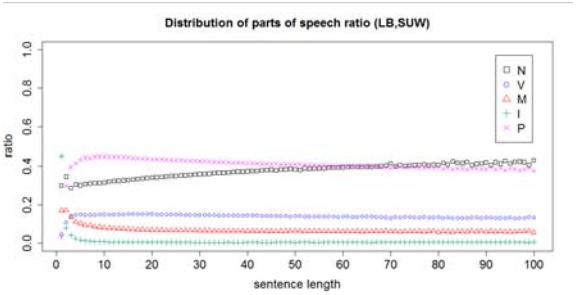
	文の長さ	N	V	M	I	P
文の長さ	1.000	0.951	-0.157	-0.575	-0.210	-0.307
N		1.000	-0.210	-0.566	-0.187	-0.358
V			1.000	-0.568	-0.885	0.975
M				1.000	0.681	-0.438
I					1.000	-0.832
P						1.000

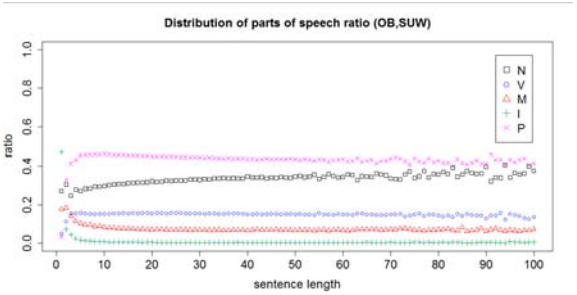
表 12　文の長さと品詞類との相関行列（長単位）

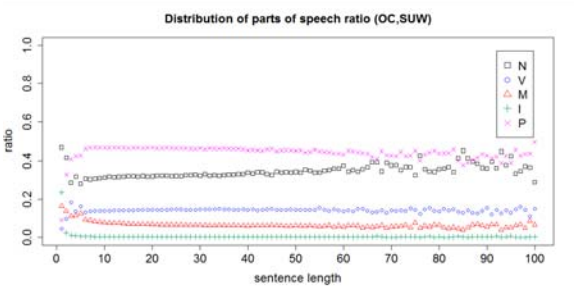
	文の長さ	N	V	M	I	P
文の長さ	1.000	-0.156	0.033	-0.306	-0.213	0.260
N		1.000	-0.965	0.454	0.940	-0.931
V			1.000	-0.493	-0.954	0.928
M				1.000	0.637	-0.726
I					1.000	-0.986
P						1.000

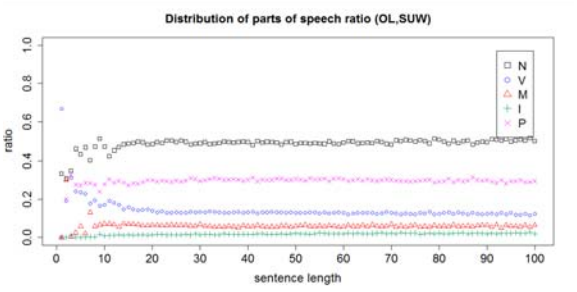
５．３　レジスターによる違い

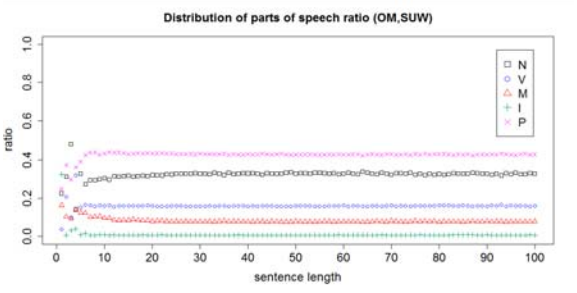
文の長さによる品詞類の比率はレジスターによって違いがあるかを見てみよう。図 15 に短単位における文の長さ 1～100 の推移を示した。

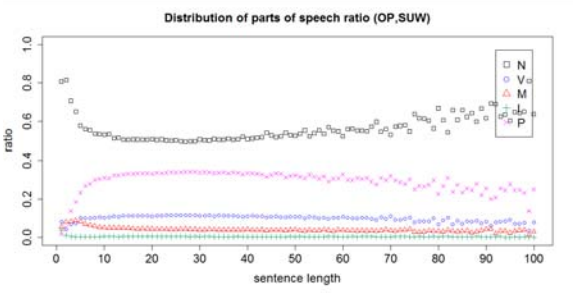


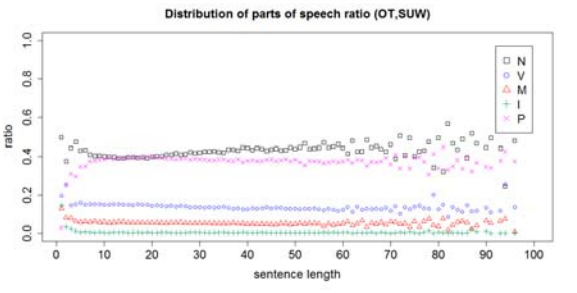


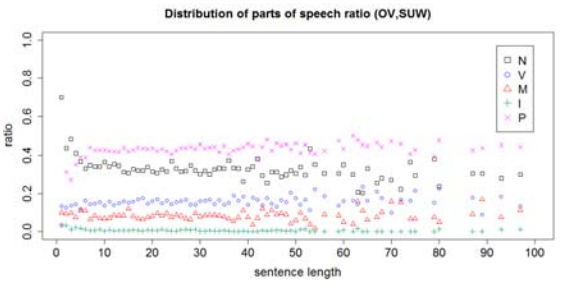












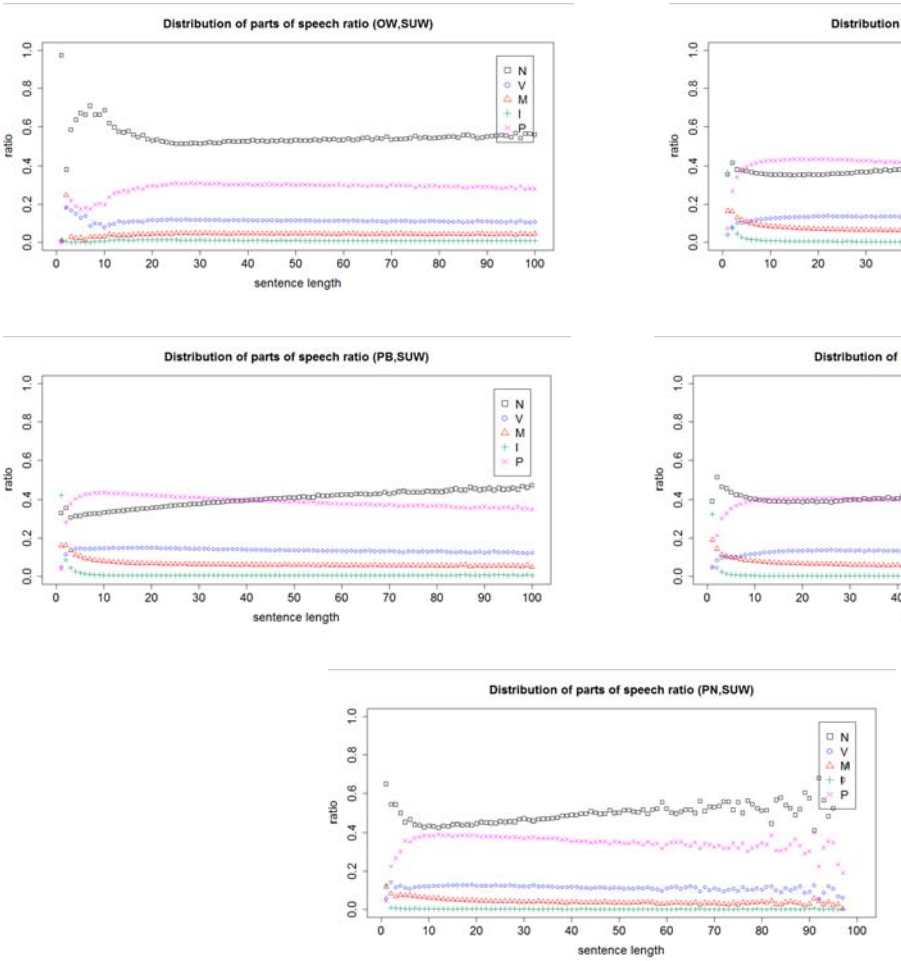


図 15　レジスターごとの文の長さによる品詞類の比率の推移（短単位）

N（名詞類）と P（助詞・助動詞類）との関係が各レジスターで違いが見られた。全体の傾向と同じような、最初 P が大きく、途中で N が上回るようになるのは、LB（図書館・書籍）、PB（出版・書籍）である。OY（Yahoo!ブログ）もその傾向に近いが、N が P を上回るところまでいかず、ほぼ同じ値に収束している。OT（教科書）、PM[9]（雑誌）は最初 N と P がほぼ同じ値であり、途中から N が P より大きくなる傾向がある。OB（ベストセラー）、OC（Yahoo!知恵袋）、OM（国会会議録）は OV(韻文)P が終始 N を上回っている。OL（法律）、OP（広報紙）、 OW（白書）、PN（新聞）は終始 N が P を上回っている。N と P の関係は長単位ではほとんどのレジスターで P が N を上回っている。ただし、OL では N と P の比率がほぼ同じであった。紙幅の関係で対照的な分布を示す LB と OL の 2 つ

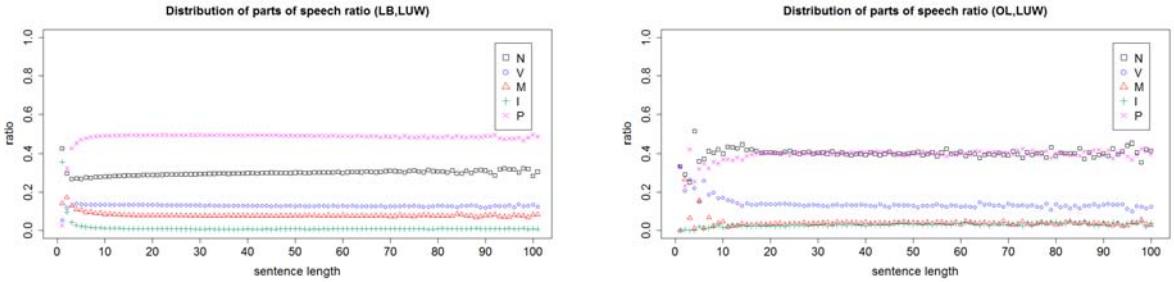


図 16　レジスターごとの文の長さによる品詞類の比率の推移（長単位）

[9] PM は OP と同じグループとも考えられる。

を図 16 に挙げた。

６．まとめと今後の課題

本稿では言語単位と文の長さによる品詞類の構成比を概観した。品詞類の分類は樺島(1954)等にしたがって、N（名詞類）、V（動詞類）、M（形容詞・形状詞・副詞類）、I（接続詞・感動詞類）を設けさらに P（助詞・助動詞類）の追加した 5 分類とした。結果としては以下の傾向が観察された。

（１）短単位と長単位とでは N と P の比率に変化が見られた。N は短単位より長単位の方が低く、逆に P は短単位より長単位のほうが高い。

（２）文末の記号による品詞類の比率に変化が見られた。N の値が「」、？」で低く、「.」で高い。「。！」はその中間であった。

（３）文の長さによる品詞類の比率の推移については、短単位、長単位ともに長さ 10 くらいから値が一定化する傾向が見られた。ただし、短単位ではそれまで P＞N だった傾向が長さ 43 から N＞P となり逆転する。

（４）文の長さとの相関が見られた品詞類は短単位の N の場合のみであった（長さ 1～100 における相関係数において係数 0.951）。

（５）レジスターによる違いについては､短単位において N と P について違いが見られた。

常に N＞P：OL　OP　OW　PN
常に N＜P：OB　OC　OM　OV
N＜P から N＞P へ変化：LB　PB
N＜P から N＝P へ変化：OY
N＝P から N＞P へ変化：OT　PM

今後は個別の品詞についての分析を進めるとともに、今回観察された現象の言語学的な解釈について考察を進めたい。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「コーパス日本語学の創成」による研究成果の一部である。データとして利用した BCCWJ は、国立国語研究所のプロジェクト及び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）による補助を得て構築したものである。

参考文献

大野晋(1956),「基礎語彙に関する二三の研究―日本の古典文学作品における」, 国語学, 24, pp.296-329.
小木曽智信・間淵洋子・前川喜久雄(2011),「『現代日本語書き言葉均衡コーパス』における形態論情報付き XML フォーマット」, 言語処理学会第 17 回大会発表論文集, pp.352-355.
樺島忠夫(1954)「現代文における品詞の比率とその増減の要因について」，国語学，18，pp.15-20.
樺島忠夫(1955)「類別した品詞の比率に見られる規則性」，国語国文，24(6)，pp.55-57.
樺島忠夫(1979)『日本語のスタイルブック』大修館書店
計量国語学会(2009)『計量国語学事典』朝倉書店
間淵洋子(2011)「自動認定によって付与されるタグ」,『現代日本語書き言葉均衡コーパス』における電子化テキストの構築，国立国語研究所内部報告書(LR-CCG-10-03)
水谷静夫(1965)「大野の語彙法則について」，計量国語学，35，pp.1-13.
水谷静夫(1981)「構成比の線型回帰調整，併せて再び大野の語彙法則」計量国語学，13(2), pp.92-97.
水谷静夫(1983)『朝倉日本語講座 2 語彙』朝倉書店