同一見出し語の出現間隔の分布と文体差

山崎　誠（国立国語研究所言語資源研究系）†

Distribution of Gaps between Successive Occurrences of the Same Word and Stylistic Differences

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)

１．はじめに

計量語彙論の基本的なテーマのひとつに同一の見出し語の出現間隔の問題がある。水谷(1983)によれば、古くは、Epstein(1953)がプーシキンの『大尉の娘』をデータとしてロシア語の前置詞 κ の出現間隔の分布を、 $F(x) = 1 - e^{-\lambda x}$ で近似した例がある（ただし、x は出現間隔、λ は当該の語の使用率。この場合 λ は、8.3×10^{-3}）。Herdan(1966:127-130)は、これをさらに進めて機能語の出現はポワソン分布に従うのではないかということを示唆した。

本稿では、同一見出し語の分布がテキストの持つ特性、とくに、文体ないしはレジスターと呼ばれるものと何らかの関係を持つのではないかという想定のもとに 2 つの調査でそれを確かめることを目的とする。ひとつは、高頻度語の出現間隔の分布、もうひとつは、ひとつのテキストに現れるすべての同一見出し語の出現間隔の分布である。

２．出現間隔の測り方

テキスト中に現れる語は、使用頻度 1 の語以外は、2 回以上繰り返して出現することになる。その際、同一の語が 2 回出現する場合、それらの間の距離を、間に含まれる言語単位の数で測った値を出現間隔とする。本稿では、同一語の 2 回出現の間に他の語が x 語存在している場合、その 2 語の間の出現間隔を x+1 とする。したがって、同一語が隣り合って出現する場合はその出現間隔は 1 となる。また、出現間隔は、着目している出現と直前あるいは直後の出現との距離を測り、間にひとつ以上同じ語をはさんだ距離は出現間隔の測定の対象外とする。以上の手続きにより、例えば、あるテキストで使用頻度 x(x≧2)の語の持つ出現間隔の総数は、x－1 回となる。

３．データ

本稿で使用するデータは、『現代日本語書き言葉均衡コーパス』(BCCWJ)のコアデータである。意味的なまとまりを重視するため、可変長データを用い、言語単位は短単位を利用した。また、語数のカウントにあたって品詞が空白ないしは品詞欄に「記号」の文字列を持つ語[1]を除外した。以下、本稿では語数という場合はこの方法による。

４．高頻度語（機能語）の出現間隔

ひとつめの調査は、コーパス開発センターで公開されている、BCCWJ の短単位語彙表データにおいて BCCWJ 全体の順位の上位 10 語を高頻度語とし、調査対象とした。具体的には、以下の 10 語である。「の（格助詞）、に（格助詞）、て（接続助詞）、は（係助詞）、だ（助動詞）、を（格助詞）、た（助動詞）、する（動詞）、が（格助詞）、と（格助詞）」

表 1 は、高頻度における出現間隔の分布のようすである。いずれも平均よりも中央値が小さく、分布が低いほうに偏っていることが分かる。例として格助詞「の」の出現間隔のヒストグラムを図 1 に挙げた。表 1 に挙げた語はいずれもこのような分布を示している。

† yamazaki@ninjal.ac.jp

1 品詞が補助記号であるものと記号であるものとが該当する。

ただし、表 2 に示したように、出現間隔の値を対数（自然対数）で表すと、正規分布に近くなるように見える[2]。

表 1　高頻度語の出現間隔の分布

語	出現間隔数	平均値	最小値	中央値	最大値
の（格助詞）	46,485	18.848	1	13	315
に（格助詞）	31,007	27.401	1	19	461
て（接続助詞）	27,465	30.314	2	20	1289
は（係助詞）	26,043	32.115	2	23	784
だ（助動詞）	24,395	43.610	1	20	2212
を（格助詞）	27,288	30.245	1	20	481
た（助動詞）	22,982	35.068	2	20	1196
する（動詞）	22,845	35.663	1	23	745
が（格助詞）	20,820	39.350	2	26	737
と（格助詞）	18,439	43.438	1	29	776

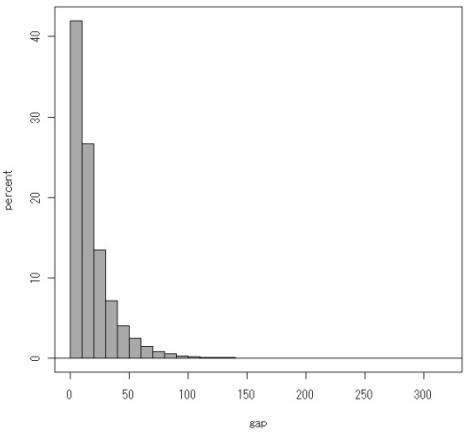

図 1 格助詞「の」の出現間隔の分布

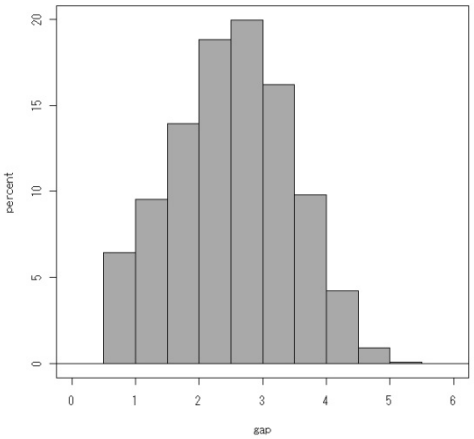

図 2 格助詞「の」の出現間隔の分布（対数）

出現間隔を対数に変換したものを使い、表 1 の 10 語について、R の多重比較（チューキーの HSD 検定）を行った。表 2 は、各語ごとに、コアデータを構成する各レジスターの組み合わせのどれに有意差があったかを示したものである。表 2 から、以下のようなことが見てとれる。(1)多くの組み合わせに有意差が認められること。(2)「の」「を」の出現間隔の分布は、他と比べて有意差の組み合わせが少ない。(3)PB（出版書籍）と OY（Yahoo!ブログ）の組み合わせはここで挙げた高頻度語の出現間隔の分布では大きな差がない。

表 2 レジスターの組み合わせと有意差

組合せ	の	に	て	は	だ	を	た	する	が	と
OW-OC	***		***	***	***		***	**	***	***
OY-OC		***	***	***		*	**	***	***	***
PB-OC		***	***	***	***		***	***	***	***
PM-OC		***	***	***	***		***	***	***	***

[2] ただし、バートレット検定を行うと各群（OC、OW、OY、PB、PM、PN）の分散は等しくない。(df=5、p=2.2e-16)

PN-OC		***	***	***	***		***	***	***	***
OY-OW	***	***	***	***	***		***	***	***	***
PB-OW	***	***	***	***	***		***	***	***	***
PM-OW	***	***	***	***	***		***	***	***	***
PN-OW	***	***		***	***	***	***	***	***	***
PB-OY					**	*	*			
PM-OY			***		***		***	*	*	***
PN-OY			***		***	***	***			***
PM-PB		***	***	***	***		***			***
PN-PB		**	***	***	***	***	***	***		***
PN-PM	***		***		***	***		***		

（注）***：0.1%水準で有意差、**：1%水準で有意差、*：5%水準で有意差

５．全出現間隔の分布

テキストに出現する語は、頻度 1 以外は出現間隔を持つが、それらすべての出現間隔の分布はどのようになっているだろうか。LB（図書館書籍）、OW（白書）、OY（Yahoo!ブログ）の 3 つのレジスターについて観察した。それぞれのレジスターにおいて可変長部分で短単位の語数が 2000 語～2099 語のサンプルを選んだ。OW と OY は該当するサンプルのすべて（26 サンプルずつ）、LB は、ランダムに抜き出した 22 サンプルである。その一覧を表 3 に挙げる。

表 3 全出現間隔の分布

サンプル ID	Token[3]	Type[4]	総個数[5]	平均[6]	標準偏差	ジャンル（NDCなど）
LBd0_00011	2078	623	1473	136.3252	251.0421	0 総記
LBe0_00003	2049	607	1458	130.7236	245.2134	0 総記
LBl1_00007	2069	418	1663	105.8443	215.3948	1 哲学
LBq1_00035	2086	417	1676	120.3365	232.6101	1 哲学
LBe2_00056	2052	536	1528	128.0118	224.7713	2 歴史
LBr2_00028	2096	619	1490	113.943	224.3516	2 歴史
LBg3_00067	2073	525	1770	166.4633	307.9518	3 社会科学
LBo3_00020	2021	390	1643	112.5526	229.1028	3 社会科学
LBh4_00036	2059	414	1658	111.7835	214.3982	4 自然科学
LBo4_00014	2022	522	1531	137.1633	245.3251	4 自然科学
LBm5_00009	2061	470	1601	132.6552	245.3482	5 技術･工学
LBm5_00011	2047	562	1502	129.8755	252.5623	5 技術･工学
LBb6_00012	2021	535	1516	133.0389	234.9972	6 産業
LBj6_00011	2095	605	1507	132.8653	249.6481	6 産業
LBo7_00002	2025	589	1449	154.8075	271.912	7 芸術･美術
LBs7_00075	2039	676	1384	130.2536	250.1484	7 芸術･美術
LBb8_00005	2030	325	1749	90.52144	196.4048	8 言語
LBs8_00014	2079	656	1454	125.3521	221.0104	8 言語
LBr9_00016	2047	657	1404	152.6489	282.0645	9 文学

3 延べ語数。

4 異なり語数。語彙素、語彙素読み、語彙素細分類、品詞の 4 つの属性が一致したものを同一の語として集計した。

5 当該サンプルの可変長部分における同一語の出現間隔の全個数。

6 当該サンプルの可変長部分における同一語の出現間隔の総和を全個数で割ったもの。

LBr9_00093	2078	557	1976	143.3011	295.4928	9 文学
LBmn_00016	2050	485	1681	130.8917	247.1816	N 分類なし
LBsn_00024	2019	565	1624	142.3565	296.3724	N 分類なし
OW1X_00089	2056	301	1755	116.159	186.6517	科学技術
OW1X_00457	2053	421	1632	119.4424	203.7048	安全
OW1X_00540	2086	350	1736	119.4764	217.0649	国土交通
OW2X_00030	2033	357	2179	140.8123	256.6013	農林水産
OW2X_00949	2008	489	1637	151.843	269.7998	福祉
OW2X_00960	2016	272	1744	96.06479	159.5637	安全
OW3X_00044	2062	450	1825	142.5326	290.3834	国土交通
OW3X_00046	2083	582	1706	179.2093	315.1336	外交
OW3X_00198	2000	508	1809	152.382	278.9863	農林水産
OW3X_00205	2076	451	1625	138.4732	246.2405	安全
OW3X_00359	2051	452	1599	138.9962	236.5418	環境
OW3X_00435	2059	419	1640	113.1311	196.4604	安全
OW3X_00562	2069	479	1590	110.4755	204.0817	安全
OW4X_00145	2060	416	1940	143.2495	264.8059	環境
OW4X_00197	2061	325	1736	108.8278	183.9437	福祉
OW4X_00238	2031	343	1942	122.7549	239.4015	環境
OW4X_00536	2098	416	1682	125.9417	228.489	安全
OW4X_00648	2084	318	1766	106.9468	188.3023	国土交通
OW5X_00025	2081	266	1815	105.4997	160.1334	福祉
OW5X_00170	2093	624	1469	150.8754	263.0187	環境
OW5X_00663	2033	585	1448	149.9841	269.2003	外交
OW5X_00853	2032	481	1551	113.0129	206.1083	福祉
OW5X_01851	2012	508	1730	145.9046	284.9876	安全
OW6X_00035	2007	264	2103	140.8783	242.0626	経済
OW6X_00090	2004	447	1557	129.9332	249.6389	環境
OW6X_00190	2082	467	1615	124.2533	240.2149	環境
OY01_02573	2063	400	1663	120.9784	223.2255	ビジネスと経済
OY01_02830	2099	525	1574	155.1652	244.3484	ビジネスと経済
OY04_02358	2096	623	1473	123.7502	236.6436	エンターテインメント
OY04_02691	2046	428	1618	116.2979	210.2759	エンターテインメント
OY04_03782	2077	532	1545	116.2979	210.2759	エンターテインメント
OY04_04825	2067	529	1538	128.4018	247.8479	エンターテインメント
OY04_06866	2013	513	1500	145.018	259.8423	エンターテインメント
OY06_01586	2094	625	1469	140.998	275.7961	政治
OY13_00240	2006	611	1395	131.0616	244.4963	芸術と人文
OY13_03195	2039	508	1531	127.1346	238.8284	芸術と人文
OY13_03612	2090	671	1419	142.561	260.089	芸術と人文
OY14_03639	2090	478	1612	127.6917	222.3718	Yahoo!サービス
OY14_06486	2009	617	1392	119.403	227.533	Yahoo!サービス
OY14_07107	2000	503	1497	134.5137	248.3512	Yahoo!サービス
OY14_12047	2063	675	1388	149.438	258.3771	Yahoo!サービス
OY14_12048	2059	675	1384	149.8215	258.5773	Yahoo!サービス
OY14_36191	2026	651	1375	125.8211	250.4715	Yahoo!サービス
OY14_37664	2045	616	1429	155.3611	288.0154	Yahoo!サービス
OY14_50772	2008	438	1570	135.235	234.4837	Yahoo!サービス
OY15_00485	2069	757	1312	148.3941	259.358	趣味とスポーツ

OY15_01982	2074	803	1271	148.8891	279.3363	趣味とスポーツ
OY15_02404	2015	697	1318	148.9484	278.0393	趣味とスポーツ
OY15_07093	2074	691	1383	121.5582	220.2368	趣味とスポーツ
OY15_17336	2084	749	1335	153.4022	285.8187	趣味とスポーツ
OY15_17721	2008	628	1380	151.259	259.3086	趣味とスポーツ
OY15_18790	2037	515	1522	119.4087	238.4768	趣味とスポーツ

前節で高頻度語においては、レジスターによって出現間隔の分布が異なっているものが多いことが確認されたが、テキスト全体の出現間隔はレジスターにより異なるだろうか。出現間隔の値を対数に変換して[7]、R の多重比較を用いたところ、表 4 のような結果を得た。どの組み合わせにおいても 5%水準で有意差は認められなかった。しかし、表 5 に示したように、出現間隔の総個数[8]で比較したところ、いずれも組み合わせも 1%の水準で有意差があった。また、表 6 は Type（異なり語数）による比較であるが、OW（白書）と LB（図書館書籍）、OY（Yahoo!ブログ）と白書（OW）に 1%水準で有意差が認められた。出現間隔全体の分布はレジスターによる特徴がないことから、むしろテキストの普遍的な属性として考えられる可能性がある。この点についてはさらに多くのレジスターのテキストについて考察が必要である。

表 4 出現間隔の平均値による比較

組み合わせ	P 値
OW-LB	0.08873
OY-LB	0.13782
OY-OW	0.13463

表 5 出現間隔の総個数による比較

組み合わせ	P 値
OW-LB	0.0023
OY-LB	0.0062
OY-OW	$<1e^{-04}$

表 6 Type による比較

組み合わせ	P 値
OW-LB	0.000402
OY-LB	0.185692
OY-OW	$<1e^{-04}$

ところで、出現間隔の総個数が多くなると、理屈の上では、平均出現間隔が小さくなると期待されるが、実際にデータを見てみると図 3 に示したように、出現間隔の総個数と平均出現間隔との間には相関は見られない（相関係数は－0.1867）。一方、出現間隔の総個数

図 3 出現間隔総個数と平均出現間隔

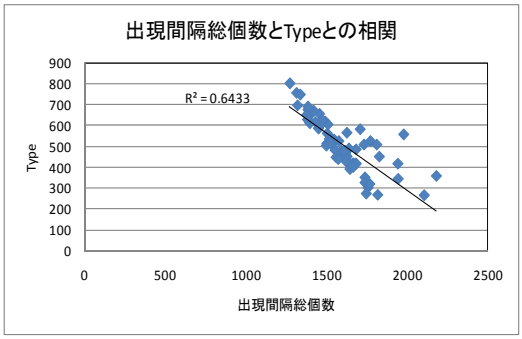

図 4 出現間隔総個数と Type

[7] バートレット検定を行ったところ各群（LB、OW、OY）の分散には有意差がなかった（df=2、p=0.1411）。

[8] 対数に変換した値を用いた。表 6 の Type も同じ処置である。

は、図 4 に示すように Type（異なり語数）と強い負の相関がある。出現間隔が増えることは個々の見出し語の使用頻度が多くなることを意味する。今回扱ったサンプルは延べ語数がほぼ一定なので、使用頻度の多い見出し語が増えれば、相対的に Type が少なくなるものと想定される。また、当然のことであるが、出現間隔の総個数と Token（延べ語数）との間には相関はない。

６．まとめと今後の課題

本稿では、テキストにおける見出し語の出現間隔の分布と文体（レジスター）との関係を考察した。その結果、高頻度語（主に機能語）の出現間隔は、BCCWJ のレジスターによって違いがあることが分かった。また、テキストにおける出現間隔の平均は、レジスターによる違いは見られないが、出現間隔の総個数がレジスターにより違いがあった。

今回は、出現間隔と文との関係は考慮しなかった。出現が 1 文の中で起きたものか、それとも文を超えるものであるか、その違いも考察の対象となるだろう。また、テキストを文の連続と考え、何番目の文に同じ見出し語が出現しているかというとらえ方も可能であろう。

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

Epstein, B. (1953) "Some Remarks on the Length of Gap Between Successive Occurrences of High Frequency Words" in Josselson, H.H.(1953) "Russian Word Count", Wayne University Press.

Herdan, G. (1966) The Advanced Theory of Language as Choice and Chance. Springer-Verlag

水谷静夫(1983)　『朝倉日本語講座 2 語彙』朝倉書店