段落間の類似度を利用したテクストの結束性の測定

山崎　誠（国立国語研究所言語資源研究系）†

Measurement of Textual Cohesion Using Similarity between Paragraphs

Makoto Yamazaki (Dept. Corpus Studies,NINJAL)

１．はじめに

テクストの結束性（cohesion）は、一貫性（coherence）とともにテクストの基本的な性質とされる重要な概念である。とくに結束性は Halliday&Hasan（1976）以来、多くの研究が行われている。本稿は、テクストを構成する段落間の類似度を使ってテクストの結束性を計量的に明らかにしようとするものである。

２．テクストにおける結束性とその現れ方

結束性とは、文章をひとつの統一体としてまとめあげるために必要な性質のひとつとされる。結束性について最初に詳細に研究を行ったのは Halliday&Hasan(1976)である。それによると、結束性について次のように紹介されている。

「結束性が生じるのは，談話のある要素の解釈（INTERPRITATION）が別の要素の解釈に依存する場合である。一方を効果的に解釈するためには他方に頼らなければならないという意味で，一方は他方を前提（PRESUPPOSE）とする。こういうことが生じるとき，結束関係が成立する。その結果，前提語と被前提語という　2 つの要素が、少なくとも潜在的には，統合されて 1 つのテクストになるのである。」（邦訳 p.5）

また、結束性には文法的結束性と語彙的結束性とがあり、前者の手段として「指示」「代用」「省略」「接続」が、後者には「再叙（reiteration）」と「コロケーション」がある。再叙には以下の 4 つのタイプがある。

(a)同一語（繰り返し）

(b)同義語（または近似同義語）

(c)上位語

(d)一般語

平林（2003:72）によれば、日本人高校生の英作文（1 作文あたり平均約 100 語、10 文）の分析から 1 作文あたり平均 12.87 個の結束数が現れ、その内訳は、指示 4.9 個、接続 3.62 個、語彙的結束性 3.36 個、代用 1.00 個、省略 0 個だという[1]。

Károly（2002:162）では、英語の作文においては、(a)の同一語の繰り返しよりは(b)〜(d)を合わせた「異なる語の繰り返し」の方が多く用いられるという報告がある。しかし、本稿では、同義語（類義語）や上位語の判断を自動的に行うことが難しいため、(a)の同一語の繰り返しのみを観察対象とする。

† yamazaki@ninjal.ac.jp

1 作文における結束性の割合はひとつの目安となるが、今回例として取り上げた白書のデータでは指示(1 個)、接続(1 個)よりも語彙的結束性(181 回:繰り返し使用された語における繰り返しの数）のほうが多く、やや異なる出現状況であった。

３．データと方法

３．１ データ

本発表では、2011 年 12 月にリリースされた『現代日本語書き言葉均衡コーパス』の DVD 版を使用した。Disk1 の CORE/M-XML フォルダに含まれる白書の xml ファイル 62 個が対象である。これらの xml ファイルは可変長サンプルと固定長サンプルを統合したもので、短単位、長単位の形態論情報のタグのほか可変長部分には文章構造のタグを含んでいる[2]。本稿では結束性を適切に観察するため、文章構造のタグを含まない固定長部分を対象外とし、可変長部分のみを用いる。

対象となるデータには文章構造のタグとして<paragraph>が使われており、このタグで囲まれた部分を一つの段落としてテクストの構成を考えることにする。なお、<title>のタグで囲まれた見出し部分や注記、図表のキャプションなどは対象外とした。

３．２ 結束性の測定方法

結束性の測定方法は 2 節で挙げた語彙的結束性のタイプのうち(a)の同一語の繰り返しを利用する。各段落をひとつの語彙とみなし、語彙の類似度を利用して、各段落間の関連を観察する。

語彙の類似度を表す主な指標には C（宮島（1970））と D（水谷（1980））とがあるが、本稿では D（以降、単に「類似度」と言う）を用いる。この類似度は、非対称的であることに特徴があり、語彙 a の語彙 b に対する類似度と語彙 b の語彙 a に対する類似度がそれぞれ別の値をとることができる。見方を変えれば語彙の「依存度」（水谷（1980:147））と考えることもできるものであり、順を追って構成されるテクストの内容の関係をさぐる上で適切な指標となるものである。語彙 a の語彙 b に対する類似度は次の式で表される[3]。

$$Da|b \stackrel{def}{=} \sum_{M \in V_a \cap V_b} P_a(M) = \frac{1}{N_a} \sum_{M \in V_b} F_a(M)$$

V_x：表現域 x の上の語彙
$P_x(M)$：見出し語 M の表現域 x における使用率
N_x：表現域 x の延べ語数
$F_X(M)$：表現域 x での見出し語 M の使用度数

類似度の測定にあたっては、短単位を用い、品詞が空白、補助記号、助詞・助動詞であるものを除外した。これらは結束性を表しているわけではないからである。同様に、結束性への貢献が低い語群についても除外した。この語群の候補として田中（1973）の「無性格語」を利用した。無性格語とは、田中によれば「これらの単語は，どんな文章にも現われるようなものであって，ある特定な文章や文献の性格とか特徴とかを反映することは，ほとんどない。いわば無性格な語群であろう。」（田中（1973:157））とされるものである。田中（1973）では 108 語が無性格語としたリストアップされている。本稿ではこの無性格語を短単位での実現形に合わせて適宜修正して用いた[4]。また、田中（1973）の趣旨を汲み、リストに上がっていない数詞についても無性格語として処理した。

---

[2] タグの詳細については小木曽ほか（2011）を参照。
[3] 式は水谷（1983）より引用。
[4] 本稿で用いた無性格語のリストを付表に掲げた。

４．分析例

４．１ データ

表 1 は『平成 16 年度文部科学白書』（サンプル ID:OW6X_00000）の可変長部分を段落に分けて示したものである[5]。見出し部分は本文とは別立てにして、その見出しが及ぶ範囲が明らかになるように示した。このサンプルは、接続詞が第 5 段落の「さらに」の 1 つのみ、指示詞が第 3 段落の「これ（まで)」の 1 つのみであり、文法的結束性が少ないという特徴を持っている。

表 1　白書データ（OW6X_00000）の構造

見出し1	見出し2	見出し3	段落番号	テクスト
1　日本文化の発信による国際文化交流の推進	(1)文化庁文化交流使事業	①　文化庁文化交流使事業	P1	文化庁文化交流使事業は，芸術家，文化人等，文化に携わる人々に，一定期間「文化交流使」として世界の人々の日本文化への理解の深化や，日本と外国の文化人のネットワークの形成・強化につながる活動を展開してもらうことを目的として，平成15年度から始めた事業です。
			P2	「文化交流使」の活動には，(ⅰ)日本在住の芸術家，文化人が海外に一定期間滞在し，日本の文化に関する講演，講習や実演などを行う「海外派遣型」，(ⅱ)海外在住の日本文化に深い知見を持つ芸術家，文化人が，講演，講習，現地メディアへの投稿，出演等を行う「現地滞在者型」，(ⅲ)講演等で来日する諸外国の著名な芸術家が，日本滞在期間を利用して学校などを訪問して実演・講演等を行う「来日芸術家型」の三つの類型があります。
			P3	平成16年度は，「海外派遣型」文化交流使として11名，「現地滞在型」文化交流使として4名，「来日芸術家型」文化交流使として4組の指名を行いました。重要無形文化財保持者，写真家や音楽家など様々な分野で活躍中の方々の活動を通じて，日本文化のこれまで紹介されていなかった一面や，日本文化になじみの薄かった国や地域での日本文化の紹介などの活動を行っています。
		②　文化庁文化交流使活動報告会	P4	平成15年度に文化庁文化交流使として海外で活動した人々による報告会を，東京国立博物館平成館大講堂にて開催しました。
			P5	笑福亭鶴笑氏(落語家)，田中千世子氏(映画評論家)，バロン吉本氏(漫画家)，三浦尚之氏(福島学院大学教授)，渡辺洋一氏(和太鼓奏者)の5名が活動報告を行うとともに，国際文化交流について討論し，さらに笑福亭鶴笑氏によるパペット落語(笑福亭鶴笑氏が自ら考案した落語形式で，足や膝につけた人形を操りながら演じる。)の実演が行われました。
	(2)国際文化フォーラムの開催		P6	「国際文化フォーラム」は，国際的に著名な国内外の芸術家・文化人などを招聘し，座談会，講演などの形式により，世界の文化芸術の最新の諸相や動向について語り合ってもらうことを目的として，平成15年度から開始した事業です。
			P7	平成16年度も15年度に引き続き，11月に関西地区で，「文化の多様性」の共通テーマの下に，「国際情勢における『文化の多様性』の意義」，「シルクロードと仏教文化」などについて話し合い，世界に向け，文化のメッセージを強く発信しました。
	(3)国際芸術見本市		P8	舞台芸術のブース設置や実演を行うことにより，国内外の劇場関係者に，我が国の新進の舞台芸術作品などを紹介する国際芸術見本市を，平成16年8月に東京芸術劇場にて開催しました。文化庁から国際舞台芸術交流センターへ委任を行い，実演デモンストレーション(ショーケース(＊))を制作しました。

[5] 該当箇所は文部科学省のホームページでも確認することができる。URL は次のとおり。
http://www.mext.go.jp/b_menu/hakusho/html/hpab200401/hpab200401_2_277.html

４．２ 各段落間の類似度

４．２．１ 全体の傾向

OW6X_00000 を構成する 8 個の段落相互の類似度を表 2 に挙げた（同一段落どうしの類似度は必ず 1 になるので除く）。値は 0.0364～0.5614 の間に分布し、平均値は 0.280 である。類似度の分布の様子を図 1 に示した[6]。

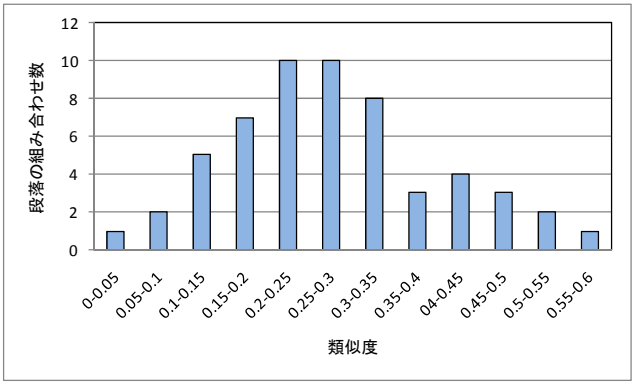

図 1 類似度の分布

表 2 はすべての段落間の類似度であるが、ここからある段落 a から他の段落 b への類似度において、対象とする相手方の段落 b との類似度が最も高い段落（表の太字のセル）がほとんど第 1 段落（P1）～第 3 段落（P3）に集中しており、このサンプルは前方の段落に依存する傾向があることが見て取れる。

表 2 段落間の類似度

	P1	P2	P3	P4	P5	P6	P7	P8	平均
P1		**0.5128**	0.4103	0.4359	0.2821	0.4615	0.2564	0.3077	0.0440
P2	0.3906		**0.5156**	0.1719	0.25	0.3125	0.0781	0.2344	0.0335
P3	0.4118	**0.5686**		0.3529	0.3333	0.2549	0.1765	0.3333	0.0476
P4	**0.5**	0.3	0.45		0.25	0.25	0.25	0.4	0.0571
P5	0.12	**0.18**	0.16	0.08		0.12	0.04	0.12	0.0171
P6	**0.4815**	0.3333	0.2963	0.1852	0.2593		0.2963	0.3333	0.0476
P7	0.2857	0.1429	0.25	0.25	0.1786	**0.3214**		0.2143	0.0306
P8	0.2368	0.2895	**0.3158**	0.1579	0.2105	0.2632	0.1053		0.1429
平均	0.0338	0.0414	0.0451	0.0226	0.0301	0.0376	0.0150	0.1429	

注

(1)縦の系列の段落が横の系列の段落に対してとる類似度の表。例えば、P3 の P2 に対する類似度は 0.5686（この値は P2 への P3 からの類似度と解することもできる）。

(2)太字は、当該段落から他の段落への類似度のうちもっとも値が高いもの。P4 の段落を例にとると、P4 の横の列（0.5,03,0.45,0.25,0.25,0.25,0.4）の中でいちばん高い値の 0.5 になる。

(3)下線は、他の段落から当該段落への類似度のうちもっとも値が高いもの。P5 の段落を例にとると、P5 の縦の列（0.2821,0.25,0.3333,0.25,0.2593,0.1786,0.2105）の中でいちばん高い値の 0.3333 になる。

[6] 例えば、階級 0.1-0.15 は 0.1 より大きく 0.15 以下であることを示す。そのほかも同様。

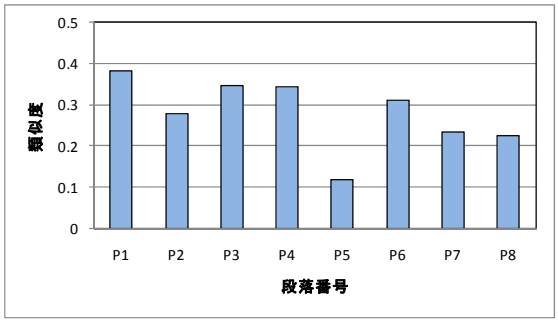

図 2 他段落への類似度

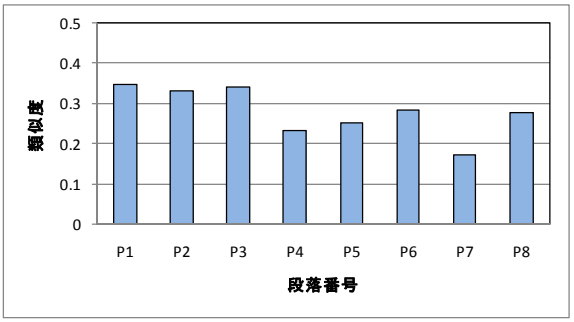

図 3 他段落からの類似度

図 2 は、ある段落の他の段落に対する類似度の平均、図 3 はある段落の他の段落からの類似度の平均である。他段落からの類似度の平均は全体的に同じような値を示しているが、他段落への類似度の平均では、第 5 段落の値がほかとくらべて低くなっていることが分かる。このことは第 5 段落と他の段落とで共通して用いられる語について、第 5 段落での使用度数は少ないが、他の段落での使用度数が多いことを示唆する。例えば、第 5 段落と、(第 5 段落への類似度がもっと高い) 第 3 段落の場合は「行う、家 (か)、活動、交流、文化、名 (めい)」の 6 語が共通して現れる語であるが、「文化」は第 5 段落に 1 回しか使用されていないのに対して第 3 段落では 7 回使用されている。同様に「交流」は 1 回に対して 3 回、「行う、名 (めい)」はそれぞれ 1 回に対して 2 回であった。この共通出現語の使用の不均衡が類似度の非対称性に影響していると考えられる。このことを踏まえて第 5 段落と第 3 段落を比較してみると、テクストの表層的な構造では第 5 段落は第 4 段落に従属するものであるが、語の分布状況から見ると第 3 段落とも関係が深いことになる。

４．２．２ 類似度からみた全体の構成

図 4 は、段落間の類似度の平均値 0.280 の 1.5 倍 (0.420) 以上の値を持つ段落の組み合わせを図示したものである。図の見方は、例えば P1→P2 であれば P1 の P2 に対する類似度が高かったことを表している。両矢印は相互に類似度の値が高かったことを示す。図 4 から、第 1 段落から第 4 段落までは結束性が強いことが伺える。一番多くの矢印が出入りしている第 1 段落がこのテクストの中心的な位置を占めていると言えよう。

第 1 段落と相互に類似度が高い 2 つの段落 (第 4 段落と第 6 段落) のうち第 6 段落は、「～は，・・・てもらうことを目的として，平成１５年度から開始した事業です。」という文構造であり、第 1 段落と骨格は同じである。このような「形式の類似性」も結束性に貢献していると考えられる。

この中では、第 5、第 7、第 8 段落が比較的独立性が高く他と分離されているが、後述のように第 5 段落は第 4 段落を具体的に展開したものであり、第 7 段落も事情は同じである。この関係は今回の分析からは読み取れない。

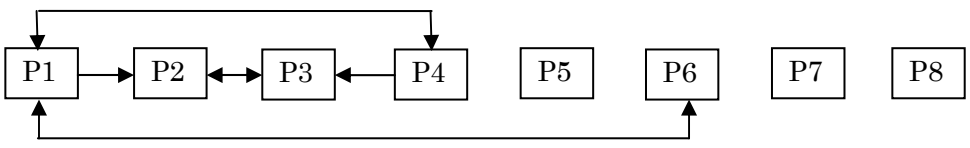

図 4 類似度からみたテクストの構成

４．２．３ 直前・直後の段落との類似度

類似度の値を利用して連続する段落間の切れ続きについて考察する。山崎（2012）では直前の段落への類似度よりも直後の段落への類似度の方が大きいところが内容的な切れ目であり表層的なテクストの構成とも一致する場合が多いと指摘しているが、このサンプルではどうであろうか。結果を図 5 に示す。

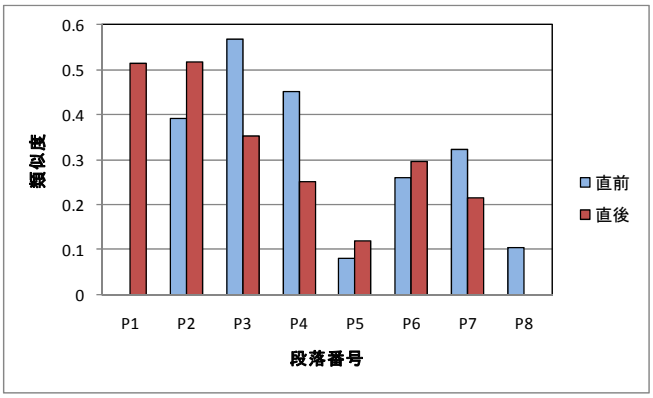

図 5 直前・直後の段落との類似度

図 5 から第 2 段落、第 5 段落、第 6 段落が直後への類似度が高いことが分かる。表層的には第 2 段落は第 1 段落の続きであり、第 1 段落の内容を具体化しているものであるが、具体的な内容が多くなったために第 1 段落への類似度が相対的に低くなったものと思われる。同様の関係は第 5 段落にも見られる。第 5 段落も直前の第 4 段落の内容を具体化したものであるが、第 4 段落が第 5 段落の短いまとめ的な内容であることから直前の段落への類似度が相対的に低くなったものである。本稿の類似度の測定は同一語かどうかによっているので、このようなものごとを具体化して述べるようなつながりについては感度が弱い。

なお、第 8 段落は後続の段落がなく、上記の方法では観察できないが、直前の段落との類似度がかなり低いことからここも内容的な切れ目に相当する可能性が高いと思われる。

５．まとめと今後の課題

本稿では段落間の非対称的類似度を利用して、テクストの結束性のようすを概観した。今回扱ったデータは白書のサンプル 1 つのみであったが、すべての段落間の組み合わせを観察することにより、どの段落とどの段落とが関係が深いのか結束性の一端を伺うことができた。また、隣接した段落以外にも結束性の高い段落があり、それらの関係を利用したテクストの構成の分析への発展の可能性を示唆した。

本稿で利用した「無性格語」のリストは雑誌九十種調査の結果から作られたもので、異なるレジスターの分析に耐えるかどうかは検証が必要であろう[7]。例えばリストには固有名詞「日本」が含まれているが、白書の分析には「日本」は重要な話題として必要な語であり、必ずしも無性格とは言えないだろう。

今後の予定としては、指示詞や接続詞などのほかの結束性を表す手段との関連も視野に入れて語彙的結束性の現れ方を総合的に記述したいと考えている。

---

7 今回使用したサンプルについては無性格語を排除しなくてもほとんど同じ結果であったが、どのような場合にこのリストが有効かは確認が必要である。

付表 本稿で用いた「無性格語」

語彙素	語彙素読み	品詞
余り	アマリ	副詞
余り	アマリ	形状詞-一般
有る	アル	動詞-非自立可能
言う	イウ	動詞-一般
行く	イク	動詞-一般
一	イチ	名詞-数詞
今	イマ	名詞-普通名詞-副詞可能
居る	イル	動詞-非自立可能
内	ウチ	名詞-普通名詞-副詞可能
円	エン	名詞-普通名詞-助数詞可能
円	エン	名詞-普通名詞-一般
御	オ	接頭辞
多い	オオイ	形容詞-一般
大きい	オオキイ	形容詞-一般
大きな	オオキナ	連体詞
置く	オク	動詞-非自立可能
於く	オク	動詞-一般
同じ	オナジ	形状詞-一般
同じ	オナジ	連体詞
思う	オモウ	動詞-一般
居る	オル	動詞-非自立可能
会	カイ	名詞-普通名詞-一般
方	カタ	接尾辞-名詞的-一般
方	カタ	名詞-普通名詞-助数詞可能
月	ガツ	名詞-普通名詞-助数詞可能
彼	カレ	代名詞
考える	カンガエル	動詞-一般
聞く	キク	動詞-一般
九	キュウ	名詞-数詞
位	クライ	名詞-普通名詞-副詞可能
来る	クル	動詞-非自立可能
五	ゴ	名詞-数詞
こう	コウ	副詞
五十	ゴジュウ	名詞-数詞
事	コト	名詞-普通名詞-一般
此の	コノ	連体詞
此れ	コレ	代名詞
三	サン	名詞-数詞
さん	サン	接尾辞-名詞的-一般
三十	サンジュウ	名詞-数詞
氏	シ	接尾辞-名詞的-一般
氏	シ	名詞-普通名詞-一般
四	シ	名詞-数詞
然し	シカシ	接続詞
七	ナナ	名詞-数詞
七	シチ	名詞-数詞
自分	ジブン	名詞-普通名詞-一般
仕舞う	シマウ	動詞-非自立可能
者	シャ	接尾辞-名詞的-一般
知る	シル	動詞-一般
十	ジュウ	名詞-数詞
十	トオ	名詞-数詞
為る	スル	動詞-非自立可能
生活	セイカツ	名詞-普通名詞-サ変可能

語彙素	語彙素読み	品詞
千	セン	名詞-数詞
そう	ソウ	副詞
そう	ソウ	名詞-助動詞語幹
そう	ソウ	形状詞-助動詞語幹
そして	ソシテ	接続詞
其の	ソノ	連体詞
其れ	ソレ	代名詞
第	ダイ	接頭辞
対する	タイスル	動詞-一般
出す	ダス	動詞-非自立可能
達	タチ	接尾辞-名詞的-一般
為	タメ	名詞-普通名詞-副詞可能
つく	ツク	動詞-一般
付く	ツク	動詞-非自立可能
強い	ツヨイ	形容詞-一般
的	テキ	接尾辞-形状詞的
出来る	デキル	動詞-非自立可能
出る	デル	動詞-一般
度	ド	名詞-普通名詞-助数詞可能
どう	ドウ	副詞
時	トキ	名詞-普通名詞-副詞可能
所	トコロ	名詞-普通名詞-副詞可能
所	トコロ	名詞-普通名詞-一般
共	トモ	名詞-普通名詞-一般
共	トモ	接尾辞-名詞的-副詞可能
取る	トル	動詞-一般
無い	ナイ	形容詞-非自立可能
中	ナカ	名詞-普通名詞-副詞可能
何	ナニ	代名詞
何	ナン	名詞-数詞
成る	ナル	動詞-非自立可能
二	ニ	名詞-数詞
日	ニチ	名詞-普通名詞-助数詞可能
二十	ニジュウ	名詞-数詞
日本	ニッポン	名詞-固有名詞-地名-国
人	ニン	接尾辞-名詞的-一般
年	ネン	名詞-普通名詞-助数詞可能
はいる	ハイル	動詞-一般
場合	バアイ	名詞-普通名詞-副詞可能
八	ハチ	名詞-数詞
日	ヒ	名詞-普通名詞-副詞可能
人	ヒト	名詞-普通名詞-一般
一	ヒト	名詞-数詞
ひとり	ヒトリ	名詞-普通名詞-副詞可能
百	ヒャク	名詞-数詞
方	ホウ	名詞-普通名詞-一般
僕	ボク	代名詞
程	ホド	名詞-普通名詞-副詞可能
前	マエ	名詞-普通名詞-副詞可能
また	マタ	接続詞
また	マタ	副詞
万	マン	名詞-数詞
見る	ミル	動詞-非自立可能
目	メ	名詞-普通名詞-一般

語彙素	語彙素読み	品詞
目	メ	接尾辞-名詞的-一般
持つ	モツ	動詞-一般
物	モノ	名詞-普通名詞-一般
物	モノ	名詞-普通名詞-サ変可能
問題	モンダイ	名詞-普通名詞-一般
遣る	ヤル	動詞-非自立可能
行く	ユク	→行く(イク)
良い	ヨイ	形容詞-非自立可能
様	ヨウ	形状詞-助動詞語幹
因る	ヨル	動詞-一般

語彙素	語彙素読み	品詞
四	ヨン	名詞-数詞
四十	ヨンジュウ	名詞-数詞
等	ラ	接尾辞-名詞的-一般
零	レイ	名詞-数詞
六	ロク	名詞-数詞
分かる	ワカル	動詞-一般
訳	ワケ	名詞-普通名詞-一般
私	ワタクシ	代名詞
私	ワタシ	代名詞

付表の注

「余り」「円」「同じ」「方（かた）」「七」「十」「そう」「所」「共」「何」「まだ」「目」「物」「私」については、短単位での品詞が複数に渡っているため、該当するものを列挙した。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「テキストにおける語彙の分布と文章構造」による研究成果の一部である。データとして利用した BCCCWJ 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

Halliday, M.A.K. and Hasan, R.(1976) *Cohesion in English.* Longman（邦訳『テクストはどのように構成されるか』、大修館書店、1997 刊）

Károly, Krisztina.(2002) *Lexical Repetition in Text.* Peter Lang.

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 回年次大会予稿集、pp.352-355.

田中章夫(1973)「自動抄録処理におけるキー・ワードの性格」『電子計算機による国語研究 V』秀英出版、pp.141-184.

平林健治(2003)「日本人初級学習者の英文ライティングの結束性の視点からみた分析」、愛知新城大谷短期大学研究紀要、2-4、pp.67-76.

水谷静夫(1980)「用語類似度による歌謡曲仕分『湯の町エレジー』『上海帰りのリル』及びその周辺」、計量国語学、12(4)、pp.145-161.

水谷静夫(1983)『朝倉日本語新講座 2 語彙』朝倉書店

宮島達夫(1970)「語いの類似度」、国語学、82、pp.42-64.

山崎誠(2012)「共起語率の分布からみるテキストの語彙的特徴」、第 1 回コーパス日本語学ワークショップ予稿集、国立国語研究所、pp.221.226.