

システミック文法に基づく書きことばの複雑さ測定 —日本語大規模コーパスを用いた語彙密度計測—

佐野 大樹

丸山 岳彦

独立行政法人 国立国語研究所 研究開発部門

1 導入

ある文章の言語的な特徴を表す指標の一つとして、その文章が持つ「複雑さ」が挙げられる。文章の複雑さを数量的に把握するための方法はこれまでに多く提案されてきたが [4, 9]、M.A.K. Halliday の主導するシステミック文法 (Systemic Functional Grammar) の枠組みで提案されている手法として、「Lexical Density (語彙密度)」の計測がある。これは、文章に含まれる内容語の数と、述語をもつ節 (システミック文法でいうところの “ranking clause”) の数との関係に着目した、文章の複雑さの計測方法である。

語彙密度の計測は、特に英語の分析において、平均文長や文章に含まれる品詞の割合などからだけでは計測し得ない複雑さを測定する方法として、コーパスを利用した言語研究のみならず、教育学 [5] や翻訳研究 [10] でも利用されてきた。しかしながら、語彙密度計測が日本語の分析に応用されたことはほとんどなく、また計測方法自体も確立されていない。そこで本稿では、日本語の文章を対象とした語彙密度計測法を提案する。また、日本語のコーパス分析に応用することで、この計測法が文書分類を行う上での新たな指標となる可能性について検討する。

以下ではまず、英語における語彙密度計測の方法について概説する。次に、現在国立国語研究所を中心に開発が進められている『現代日本語書き言葉均衡コーパス (BCCWJ)』[2] に含まれる「白書」と「書籍」のサンプルを対象に、語彙密度計測を日本語に適用するための方法について示す。その後、語彙密度を計測した結果を報告し、提案する語彙密度計測の手法が日本語の書きことばの複雑さを適切に表現できるかを検討する。

2 文章の「複雑さ」と語彙密度

書きことばが複雑となる原因には、コンテキスト的、意味的、語彙-文法的なものなど様々あるが、Halliday

は、その要因の一つは節 (clause) に多くの情報を詰め込む言語行為にあると考え、このような情報の詰め込み (information packing) によって起こる複雑さを “Lexical Complexity” と定義した [6]。この Lexical Complexity を計測する方法として提案されたのが、語彙密度計測である。語彙密度は、次の式によって求めることができ、語彙密度が高いほど、Lexical Complexity は高くなるという。

$$\text{語彙密度} = \frac{\text{テキストに含まれる内容語の数}}{\text{テキストに含まれる述語をもつ節 (ranking clause) の数}}$$

例えば、以下の文章の語彙密度は、次のように計測される [8]。

In bridging river valleys, the early engineers built many notable masonry viaducts of numerous arches.

まず、述語を持つ節 (ranking clause) の数を計測すると、‘bridging’ を述語とするものと、‘built’ を述語とするものの二つがある。前者には ‘in’、後者には ‘the’、‘many’、‘of’ という機能語が含まれ、それ以外の 11 語が内容語である。ここから

$$\text{語彙密度} = 11 / 2 = 5.5$$

となり、この文章の語彙密度は 5.5 となる。英語の場合、インフォーマルな話しことばでの語彙密度は約 2、一般的な書きことばでは約 6、科学的文章では約 13 程度になるという [6, 7]。

Halliday が Lexical Complexity を計測する方法として内容語の数および節 (ranking clause) の数に着目するのは、i) Lexical Complexity が高い文章では、動詞群 (verbal group) の名詞化 (nominalisation) などにより、内容語の割合が機能語の割合に比べて高くなる傾向があること、また、ii) システミック文法では、文 (sentence) はあくまで書きことばの形式的な単位であり、ことばを機能的かつ構造的に捉えた場合、語彙-文

法単位としては節 (clause) が適切であるという考えが強いことに由来する [7]。

3 日本語の文章に対する語彙密度計測

3.1 分析対象データ

ここでは分析対象として、『現代日本語書き言葉均衡コーパス (BCCWJ)』[2]の一部を使用する。BCCWJは、2006年より国立国語研究所を中心に開発が進められている1億語規模のコーパスで、2011年以降に全体の公開が予定されている。ここでは、構築中のBCCWJデータのうち、「白書」「書籍」のサンプルを用いた。データサイズは以下の通りである。

表 1: 分析対象データ

媒体	サンプル数	総語数
白書	28	23,850
書籍	28	24,269

なお、使用したサンプルはBCCWJで提供される「固定長サンプル」であり、各サンプルに含まれる語数はほぼ均等になっている。

各サンプルが抽出された書誌タイトル (異なり) を以下に示す。なお、白書・書籍いずれも、2005年に出版されたものに限った。また書籍は、日本十進分類法 (NDC) で「9 (文学)」に分類されるものに限った。

白書： 『エネルギー白書』『科学技術白書』『外交青書』『環境白書』『観光白書』『経済財政白書』『警察白書』『原子力白書』『交通安全白書』『厚生労働白書』『国土交通白書』『国民生活白書』『首都圏白書』『消防白書』『食料・農業・農村白書』『森林・林業白書』『水産白書』『政府開発援助白書』『青少年白書』『地方財政白書』『中小企業白書』『土地白書』『独占禁止白書』『犯罪白書』『文部科学白書』『防衛白書』『労働経済白書』

書籍： 『こぶしの上のダルマ』『もったいないじいさん』『ハマヒルガオ』『ホンマに運命?』『マリッジ』『ヤーンの朝』『ユカのこころの旅』『ラスト・ワルツ』『雲の都』『花雪の降る場所で』『強情・彦左』『鏡よ、鏡』『狐官女』『再恋』『彩雲』『私は美人』『女王様と私』『人生ベストテン』『青いあひる』『旦那さまとウェディングベル』『地果て海尽きるまで』『透明な貴婦人の謎』『別れごっこ』『忘れても、しあわせ』『無私と我欲のはざまで』『沃野の伝説』『恋占い』

Halliday は、文章は、より綿密に計画され、また、よりフォーマルなものであるほど、語彙密度が高くなる傾向にあると述べている [7]。これに従うと、政府系刊行物である白書のサンプルと文学小説が主体である書籍のサンプルを比較した場合、文体の違いや抽象度の度合いなどから、白書の方が語彙密度が高くなる傾向にあると予測される。

3.2 計測方法

内容語と機能語の判別には、BCCWJで提供される形態論情報の一つである「短単位」[1]を利用した¹。

ここでは一般的な理解にしたがって、「名詞」「動詞」「形容詞」「形状詞」「副詞」「連体詞」の品詞が付与された語 (=短単位) は、実質的な意味を表す語として「内容語」と認定した。それ以外の品詞 (「助詞」「助動詞」「接続詞」など) が付与された語は、機能語とした。ただし、以下の語は特別に機能語と見なすこととした²。

1. 連体詞および副詞のうち、「あの」「この」「こう」「そう」のように指示表現として機能するもの。
2. 動詞および形容詞のうち、自立できないもの。「～する・始める・くる、ない・ほしい・よい」など、他の内容語に後接して補助的に利用されるもの。

また、句読点などの記号類は計測から除外した。

節 (ranking clause) の数は、日本語をシステミック文法の視点から記述した Teruya(2007) に基づき、人手で分析・計測した。Teruya(2007) によると、日本語のシステミック文法では、語彙-文法の構造ユニットとして、形態素 (morpheme)、語 (word)、群 (group/phrase)、節 (clause) の4種類があり、これらはrankという概念に基づき階層的に位置づけられる。つまり、下位層のユニットが一つ上位の層の構成要素として機能するという関係にあると考えられている。例えば、「イラクへの主権移譲が行われた」という一つの節は、図1のように階層的に区分 (segment) される。

イラクへの主権移譲が行われた			
	イラクへの主権移譲が行われた		
clause	clause		
group/phrase	group		group
word	word	word	word
morpheme	mor.+mor. +mor.	mor.+mor.	mor.+mor. +mor.

図 1: 語彙-文法の構造ユニットの階層 1

なお、語彙密度計測に関わる節 (ranking clause) とは、一般的な意味での節の中でも特に、図1の“clause”rankで機能するユニットのことを指す。つまり、体裁として節の構造を持っていたとしても、それが“clause”

¹ 品詞体系は、形態素解析辞書「UniDic」による。

² 内容語と機能語の厳密な峻別には、形式名詞の扱いや、文脈によって内容語か機能語かを判別しなければならないケースの扱いなどを解決する必要がある。これは今後の課題としておく。

rank よりも下位で機能していれば、ranking clause とは見なされない。図 2 には、「別途計画されている」など、節構造をもつユニットがあるが、これらはより上位にある一つの群・句 (group) を構成する要素として機能しているため、“clause”rank より下位層のユニットであると考え、ranking clause とは見なさない。

	別途計画されている同業他社の統合が実行された場合を	想定している。
clause	clause	
group/phrase	group	group

図 2: 語彙-文法の構造ユニットの階層 2

上記の基準により、「白書」「書籍」サンプルの語彙密度を計測した。

3.3 計測結果

語彙密度を計測した結果を、図 3 に示す。縦軸は語彙密度を表す。

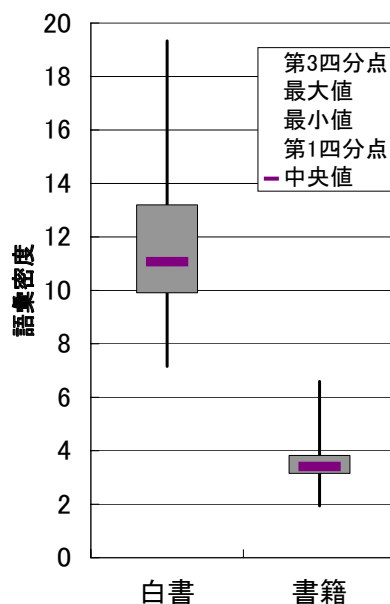


図 3: 白書と書籍の語彙密度の分布

白書サンプルの語彙密度は比較的高い位置に分布しているのに対し、書籍は低い位置に分布していることが分かる。これは、語彙密度は「白書」で高く「書籍」で低いという、3.1 で示した予測と一致する結果である。また、「白書」「書籍」それぞれのサンプル間を比較すると、特に白書において、語彙密度に大きなばらつきが認められる。

次に、「白書」と「書籍」の 1 サンプルごとの語彙密度を計測した結果について、詳細を表 2 に示す。

表 2: 1 サンプルあたりの語彙密度の計測結果

	白書	書籍
内容語数 平均値	320.17	242.07
機能語数 平均値	243.35	384.67
内容語数/機能語数	1.33	0.63
節数 平均値	28.28	70.64
語彙密度 中央値	11.07	3.41
語彙密度 最大値	19.33	6.59
語彙密度 最小値	7.15	1.94

白書サンプルの語彙密度の中央値が 11.07 であったのに対し、書籍サンプルの中央値は 3.41 であり、3 倍以上の開きがあった。白書サンプルの語彙密度の最小値が書籍サンプルの語彙密度の最大値を上回っているところから、白書サンプルは総じて書籍サンプルよりも複雑な書きことばであると判定されたことになる。また、内容語の割合は白書のほうが書籍より 2 倍以上高いが、節 (ranking clause) の平均数については、白書で 28.28、書籍で 70.64 と、書籍のほうが 2.5 倍も多くなっている。

4 考察

分析結果に基づき、「白書」「書籍」ごとに、語彙密度の高いサンプルと低いサンプルの比較を行った。

白書サンプルのうち、語彙密度が高いものは、都市計画、施策、構想、制度など、抽象的で一般性の高い事象に関する記述が多かった。例えば以下のようなものである。

● 語彙密度が高い白書サンプル

都市計画における土地利用計画の基本となる用途地域制度を始め、特別用途地区、特別用途制限地域、高度地区、風致地区等の地域地区の活用を推進した。また、平成十四年の都市再生特別措置法の制定により創設された既存の用途地域等の規制を受けず自由度の高い計画を定めることができる都市再生特別地区制度、都市計画法等の改正により... (『国土交通白書』語彙密度 19.33)

一方、語彙密度が低いサンプルは、コラム記事や解説など、具体的な事例を読者に伝えるようなタイプの記述が多かった。例を以下に示す。

● 語彙密度が低い白書サンプル

コラム 事例 災害復旧活動を支える道路ネットワーク「稲むらの火」をご存知ですか？ 今からおよそ五十年前、安政元年（千八百五十四年）十一月（旧暦）、紀州広村（現在の和歌山県広川町）は大きな地震（安政南海地震）とそれに伴う津波に見舞われました。村は、三十六名の死者を出し、被害にあわなかった家は 1 軒もないなど... (『土地白書』語彙密度 8.14)

書籍サンプルのうち語彙密度が高いものは、思想や主義等に関する意見や考えを主張するスタイルの記述が多かった。例えば、以下のようなものである。

● 語彙密度が高い書籍サンプル

いかにも湛山らしい先方の急所を穏やかに鋭く攻める論法であるが、当然ながらこの種の討議は片鱗たりとも実現しなかった。共産主義の「理想の実現」が「可能か不可能か」との詰問は思想家や運動家の最も弱い部分に向けられている。況んや更に念を押して「人類の性質上」と畳みかけられるのは極度に痛いところであった。彼等はこの現実主義的な設問に答える術を知らず、ただもう... (『無私と我欲のはざま』 語彙密度 6.59)

一方、語彙密度が低いものは、日常的な会話部分や具体的な出来事を叙述する場面を含むサンプルが多かった。以下のようなものが例である。

● 語彙密度が低い書籍サンプル

「仕方ないよ。お母さんのせいじゃないんだから」
「もっと早く話しておくべきだったわ…」
雅代はため息をついた。今日は話そう。一明日は必ず。そう思っている内に、日がたっていた。
「一私のことは？」
「芸能活動のこと？ 何も話してないわ」
「私、今日話す」と、沙也は言った。...
「隠しておく方が... (『鏡よ、鏡』 語彙密度 1.94)

上記の観察から、語彙密度の計測結果は、文章の性格の違いや複雑さの度合いを反映していると考えられる。白書、あるいは書籍という同一の媒体から取られたサンプル間でも、語彙密度の計測結果の違いによって、どの程度の複雑さを備えた文章を含むかを推測することができ、書きことばとしての「複雑さ」を数値で表現できると考えられる。

本稿で示した語彙密度の計測に加えて、語彙自体の難易度に関する意味的な指標、係り受けの深さなど統語的な指標などを利用することによって、書きことばの「複雑さ」を多角的に検討することが可能になると思われる。

5 まとめ

本稿では、システミック文法の枠組みに基づいて、日本語の書きことばの「複雑さ」を測定する方法を提案した。文章中に含まれる内容語と節 (ranking clause) の数から語彙密度を計測する方法を示し、「白書」「書籍」という性格の異なる書きことばの分析を行った。結果として、白書と書籍の語彙密度は異なる分布を見せること、同じ「白書」「書籍」のサンプル間でも、語彙密度の違いによってテキストとしての性格が異なること、などが観察された。

このような計測を大規模な日本語コーパスの分析に用いることにより、文書分類を行う上での新たな指標を得ることができると考えられる。また、バランストコーパスから目的に合わせたサンプルを抽出するための手がかりとしての活用も期待できる。

今後の課題として、より多くのタイプの文章を対象として語彙密度を計測し、その分布を利用して書きことばの特徴抽出を行うこと、節境界解析処理 [3] を利用して大規模コーパスの文書分類を行うこと、文章のリーダビリティの検証などへ応用すること、などが考えられる。

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

- [1] 小椋秀樹 (2007) 『『現代日本語書き言葉均衡コーパス』短単位規程集 Version 1.2』, 独立行政法人国立国語研究所.
- [2] 前川喜久雄 (2007) コーパス日本語学の可能性 —大規模均衡コーパスがもたらすもの—. 『日本語科学』22, 国立国語研究所.
- [3] 丸山岳彦・柏岡秀紀・熊野正・田中英輝 (2004) 日本語節境界検出プログラム CBAP の開発と評価 『自然言語処理』. Vol.11, No.3. 39–68. 言語処理学会.
- [4] 森岡健二 (1988) 『文体と表現』. 明治書院.
- [5] Butt, D., R. Fahery, et al. (1995) Using Functional Grammar: an explorer's guide. Sydney, National Centre for English Language Teaching and Research: Macquarie University.
- [6] Halliday, M.A.K. (1985) Spoken and Written Language. Victoria: Deakin University.
- [7] Halliday, M.A.K. (1990) Some grammatical problems in scientific English. Annual Review of Applied Linguistics (6): 13–37.
- [8] Halliday, M.A.K. and Matthiessen (2004) An Introduction to Functional Grammar. (3rd ed) London: Arnold.
- [9] Hasan, R. (1984) Coherence and cohesive harmony. In: James Flood (ed). Understanding Reading Comprehension: Cognition, Language, and the Structure of Prose. Newark: International Reading Association, 181–219.
- [10] Laviosa, S. (1998) Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. Meta XLIII (4): 557–570.
- [11] Teruya, K (2007) A Systemic Functional Grammar of Japanese. London: Continuum.