特定領域研究「日本語コーパス」平成20年度研究成果報告書

語彙密度を利用した『現代日本語書き言葉均衡コーパス』テクスト分類の試み

佐野 大樹　丸山 岳彦　山崎 誠　柏野 和佳子
秋元 祐哉　稲益 佐知子　田中 弥生　大矢内 夢子

平成21年3月

文部科学省科学研究費特定領域研究
「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」
データ班
JC-D-08-02

特定領域研究「日本語コーパス」平成 20 年度研究成果報告書
（JC-D-08-02）

語彙密度を利用した『現代日本語書き言葉均衡コーパス』テクスト分類の試み

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

平成 21 年 3 月

目 次

表 目 次

図 目 次

はじめに

国立国語研究所では『現代日本語書き言葉均衡コーパス (Balanced Corpus of Contemporary Written Japanese)』の構築が行われている。構築期間は 2006〜2010 年度であり，現在，サンプリング・電子化・形態論情報付与・著作権処理などの作業が進められている。本報告書はサンプリングに関する研究報告である。

データ班サンプリングサブグループでは,『現代日本語書き言葉均衡コーパス』におけるサンプルの抽出作業を行っている。また，サンプリング作業の実践から，コーパスに含まれるサンプルの特徴を表す情報として有用だと考えられる分類指標の人手，及び，自動付与について検討している。

本報告書では，この作業の一環として行われている研究である，語彙密度 (Lexical Density) を用いたテクスト分類の試みについて説明する。BCCWJ に付与が予定されている情報を利用し，語彙密度を用いて現代日本語書き言葉をどのように分類できるのかについて検討した結果について示す。

2008 年度のサンプリング作業は，国立国語研究所研究開発部門言語資源グループの，山崎誠，柏野和佳子，丸山岳彦，佐野大樹，秋元祐哉，稲益佐知子，田中弥生，大矢内夢子が中心となり行った。このうち「語彙密度 (Lexical Density) を用いたテクスト分類の試み」は，佐野大樹，及び，丸山岳彦が中心となり行った。本報告書の執筆は，佐野大樹と丸山岳彦が担当した。

第1章　語彙密度を用いたテクスト分類の背景と目的

佐野大樹・丸山岳彦

コーパスを用いた言語研究においては，コーパスに付与された種々の情報を利用して分析が行われる。例えば,「形態論情報」を用いて語彙調査を行ったり,「ジャンル情報」を用いて異なるジャンル間の言語的変異を記述したりする。このうち後者のような情報，すなわち，コーパスに収録されたテクストを分類し，その性質を示す情報は，コーパスから特定の条件を満たすテクストを抽出したり，分析対象となるデータの特徴について把握したりする場合に必要である。

『現代日本語書き言葉均衡コーパス』(以下，BCCWJ) では，収録されるサンプルのそれぞれに，ジャンル，書誌，著者情報などが付与される。しかし，同一のカテゴリに属しながら異なる性質を有するテクストの個別的な差を捉えるには，連続的な尺度が必要になる。

そこで本研究では，語彙密度（Lexical Density）という概念を提案する。語彙密度は，選択体系機能言語理論 (Systemic Functional Theory，以下，システミック理論) の枠組みで提案された，情報の詰め込み度を計測するための手法である。本章では，その前提として，コーパスに付与される情報の種類を整理し，多様なテクストの特徴を捉えることができる分類法の必要性について述べる。

1.1　BCCWJ のドキュメントアノテーション

コーパスには，コーパスの本体であるテクストを収録した言語データ以外にも，様々な情報が付与される。コーパスに付与される情報には，大別して以下の二つがある。

- セグメントアノテーション（segment annotation）
- ドキュメントアノテーション（document annotation）

セグメントアノテーションとは，テクストの一部を対象として付与される情報のことである。例えば，形態素，文節，文境界，段落境界，係り受け構造，談話構造情報などが該当する。一方，ドキュメントアノテーションとは，テクスト全体を対象として付与される情報のことである。書名のタイトル，出版年，著者の氏名や性別，生年，ジャンルなどが該当する。

研究者は，セグメントアノテーションの情報を手がかりとして言語データの分析を進めつつ，ドキュメントアノテーションの情報を手がかりとしてその特徴を分類・把握していくことになる。

BCCWJ において提供されるドキュメントアノテーションは以下のように整理できる。

- 書誌情報（書名タイトル，著者，発行年，ジャンルなど）
- サンプル情報（取得されているページ数など）
- 著者情報（氏名，性別，生年など）

『BCCWJ 領域内公開データ (2008 年度版)』(以下、BCCWJ2008) に付与された書誌情報・サンプル情報・著者情報の一部を表 1.1 に例示する[1]。

表 1.1: 書誌・サンプル・著者情報

書誌情報	
タイトル （Title）	原本のタイトル
副題 （Subtitle）	原本の副題
巻号 （Number）	原本の巻号
責任表示 （Bib_author）	原本の責任表示（著者，編者，監修者など）
出版者 （Publisher）	原本の出版者（出版社）
出版年 （Year）	原本の出版年
ISBN （ISBN）	原本に付された ISBN（国際標準図書番号）
ジャンル （Genre）	原本のジャンルに関する情報
責任表示 ID （Bib_author_ID）	原本の責任表示に対応する ID
サンプル情報	
サンプリングページ（Sampling_page）	「サンプル抽出基準点」を取得したページ
サンプリングポイント（Sampling_point）	「サンプル抽出基準点」を取得した交点
著者情報	
性別（Sex）	著者の性別
生年（BirthYear）	著者の生年

BCCWJ のように様々なタイプのサンプルが収集された大規模均衡コーパスでは，研究用途上，ドキュメントアノテーションが豊富に付与されていることが望ましい。しかしながら，日本語研究において，どのような情報がどのような形で付与されていれば，コーパスを十全に活用できるのかという点は未だ明らかでなく，今後試行錯誤を積み重ねながら考察されていかなければならない。

大規模均衡コーパスに含まれるテクストを種々の観点によって分類したり，研究者の分析目的に合うテクストを自在に抽出したりするために必要なドキュメントアノテーションの検討

[1] 現在，データ班サンプリングサブグループでは，これらの情報以外にも，日本図書コード (以下，C コード) をコーパスに付与することを検討している。C コードとは，販売対象，形態（単行本，文庫，絵本など），内容に関する情報を 4 桁の数字により分類するものである。C コードを用いた分析については，4 章以降を参照。「ジャンル」は，書籍は日本十進分類法 (以下，NDC)，白書はその内容，Yahoo!知恵袋は投稿カテゴリの情報が付与されている。すなわち，Martin[13] が定義するような，テクストの社会的目的を考慮したものではなく，British National Corpus(以下，BNC) における Domain 分類に近いものである。以下,「ジャンル」という言葉は，BCCWJ で定義されている意味合いで用いる。

は，今後のコーパス研究において重要な役割を担うことが予想される。

1.2 ドキュメントアノテーションとコンテクスト情報

BCCWJ にドキュメントアノテーションとして付与される情報の多くは，テクストのコンテクスト情報—テクストが書かれたり読まれたりする環境に関連する情報—である。コンテクスト情報がコーパスに付与されることは一般的であり，Brown Corpus， BNC， Child Language Data Exchange System (CHILDES)，日本語話し言葉コーパス (CSJ) など様々なコーパスに付与されている。

コンテクスト情報の種類は多々あるが，システミック理論の状況コンテクスト（context of situation）の枠組みでは，**Field**(活動領域)，**Tenor**(役割関係)，**Mode**(伝達様式) の 3 種に大別されている [9]。

Field はテクストによって表現されている社会行為のことをさし，テクストの分野や主題などを含む。Field 分類には，文学作品をミステリー，サイエンスフィクション，ラブストーリーなどに分類する Brown Corpus の創作散文 (imaginative prose) 分類や，経済，自然科学，ビジネスなどのカテゴリを用いる BNC の分野分類などがある [1, 4]。

Tenor は話し手と聞き手，もしくは書き手と読み手の社会的関係などをさす。その要因としては，話し手と聞き手の距離 (social distance)，習熟度や専門性の差 (expertise)，フォーマリティ (formality) などがある。Tenor 分類には，BNC Web Indexer にある読み手の対象レベル，年齢，性別 (Audience Level， Age， Sex)，著者の年齢や性別 (Author Age， Sex) の分類などがある [12]。

Mode はテクストにおける言葉の役割のことで，テクストの修辞的目的 (rhetorical purpose)，聞き手や読み手がテクスト生成に参加しやすいかどうか (process sharing)，話し言葉か書き言葉か (channel) などによって決定される。Mode 分類には，BNC の話し言葉と書き言葉の分類などがある [12]。

この枠組みを用いると，BCCWJ のドキュメントアノテーションは，Field，Tenor，Mode 別に，表 1.2 のように分類することが可能である。Field については，サンプルの主題・内容を示すジャンルと出版年がある。Tenor については，著者性別，著者年齢，C コードが表す販売対象 (一般，専門，教養など) がある。Mode については，C コードの形態 (文庫，新書など) やテクスト中のサンプルの位置 (冒頭部分など) がある。

表 1.2: BCCWJ における現行のドキュメントアノテーションと状況コンテクスト

状況コンテクスト	要因
Field	ジャンル，出版年
Tenor	著者性別，著者生年，販売対象 (C コード)
Mode	形態 (C コード)，テクスト中のサンプルの位置

したがって，BCCWJ に含まれるサンプルは，Field，Tenor，Mode の全ての観点から，サンプルの特徴について把握することが可能である。しかしながら，これらのアノテーションのみでは把握することが難しいサンプルの特徴もある。

1.3　テクスト分類とサンプルの多様性

BCCWJ に含まれる書籍サンプルの多様性について考察した柏野他 [16] では,「多様性をとらえる観点」として，以下の 18 項目をあげている。

(1) NDC 分類の 1〜3 次区分 (本の内容や主題)
(2) 種類 小説 (物語)，手紙，日記，論説文，紀行文，ルポルタージュ，韻文，翻訳，戯曲 (シナリオ)，マニュアル，ガイドブック，辞書，事典
(3) 形式 座談，対談，インタビュー，パネル討論，講演，会話形式，往復書簡形式，リレー執筆形式，Q&A 形式，投稿形式，辞書・事典形式，見本・用例形式
(4) 場面設定 時代 (現代，江戸時代，平安時代，未来)，場所 (日本国内，国外，仮想世界)
(5) 著者の属性 年代，性別，出身地
(6) 対象読者の属性 年代，性別，好み
(7) 視点 人称，人間以外
(8) 硬軟 難解，堅い，平易，くだけている
(9) 論理構成・紙面構成 章節，キャプション，注記，コラム，引用，ブロック割り構成，図説，カタログ的構成
(10) 文体 口語文，文語文，候文，和漢混淆文，条文
(11) 文末・調子 デスマス調，デアル調，ゴザイマス調，体言止め，語りかけ口調，演説調
(12) 文長 長短
(13) 修辞・比喩 種類，使い方
(14) オノマトペ 種類，使い方
(15) 語彙 語彙の選択，特に位相の異なる語彙の選択 (古語，俗語，幼児語，方言など)，語種の選択
(16) 表記 文字種の選択 (漢字，カタカナ，ひらがな)，表外漢字の使用，仮名遣い，(現代仮名遣い，歴史的仮名遣い)，ローマ字や外国語の使用
(17) 記号類 句読点，記号類の使い方
(18) ルビ・注記 使用量 (多少)，使用目的 (読み，原語，別の言い換え語，注釈，参考文献)

ここでは,「形式」,「場面設定」,「硬軟」,「文体」などがあげられているが，現行のドキュメントアノテーションから，サンプルごとにこれらの特徴を把握することは難しい。例えば，形式や文体が他の媒体に比べ統一されていると考えられる白書のサンプルのうち，同ジャンルに分類され，同一タイトルをもつサンプルであっても，その言語的特徴が大きく異なる場合がある。例として以下に，ジャンルが共に「福祉」，タイトルが『国民生活白書』のサンプルをあげる。

四季折々の風景，農村の行事，跳ね回る子供たちや動物たち…。その素朴な画風を愛され，アメリカの国民的画家となった「グランマ・モーゼス（モーゼスおばあさん）」ことアンナ・メアリ・ロバートソン・モーゼスが画筆を握ったのは７０を過ぎてからのこと。１２歳から農場で働き，結婚して１０人の子供をもうけ，農場の主婦としての切り盛りから手が離れてからであった。画家としての訓練はもちろんのこと，学校にすらたまにしか行けなかった彼女が孫娘のために刺で絵を作ってあげたのがきっかけとなった。高齢期を迎える準備は不可欠であり，当然，趣味も必要である。しかし，それで身を立てるつもりでないなら身構えて無理に作る必要はなく，仕事と子育てに明け暮れる日々の中で「時間ができたらやってみたい」と思ってきたことに取り組んでみようとするくらいの感覚で良いのではないだろうか。例えば，現役を退いた後，北上する桜前線を追いかけている夫婦がいる。桜の微妙な表情を墨絵で描き，エッセーや句にする。「ゆっくり」を合言葉にハンドルを握り,「若い日本人は旅の心をどこかに置き忘れてしまったのではないか」と問いかける。

『国民生活白書 平成 6 年版』 経済企画庁 OW4X_00397

省資源・省エネルギーを着実に推進するため，省エネルギー・省資源対策推進会議において，６３年６月２８日夏季の省エネルギー対策について，また，同年１１月２２日冬季の省エネルギー対策について決定し，関係各省庁は決定事項の周知徹底を行った。また，毎月１日の「省エネルギーの日」，１２月１日の「省エネルギー総点検の日」，２月の「省エネルギー月間」等の機会を利用して，関係各省庁，地方公共団体，省エネルギー推進関係団体等はパンフレットの配布，ポスター等による広報，集会，展示会，講習会，表彰，作文募集等，各種行事を実施した。総理府においては，テレビ等の媒体を利用した政府広報により，省エネルギーの普及広報活動を行った。警察庁は,（財）全日本交通安全協会に対し，経済運転の広報を行うよう要請した。経済企画庁は，省エネルギー啓発パンフレット及びポスターを作成し，省資源国民運動参加団体等に配布するとともに，省エネルギー月間（２月）に当庁において，懸垂幕の掲示を行った。通商産業省は,（財）省エネルギーセンターを通じて，ポスター，パンフレット等の配布などにより普及広報活動を行うとともに，全国数か所で省エネルギー展，省エネルギー講演会を開催し，また，エネルギー管理優良工場等の表彰を行った。運輸省は，運輸部門におけるエネルギー政策に関するパンフレットの作成・配布等により，普及広報活動を行うとともに，運輸部門におけるエネルギー政策に関する講演会を行った...

『国民生活白書 平成元年版』 経済企画庁 OW3X_00294

『国民生活白書 平成 6 年版』は，柏野他 [16] の「硬軟」から見れば,『国民生活白書 平成元年版』に比べ「くだけている」テクストであると考えられる。また,「種類」からいえば，前者が物語的であるのに対して，後者は報告書的である。

これらのサンプルは，現行のドキュメントアノテーションから得られる情報に基づけば，同じ特徴をもつテクストだと推測されるものである。しかしながら実際は，内容のみならず利用されている語彙や文法の利用傾向に多くの違いが認められる。現行のドキュメントアノテー

ションだけを用いた場合，研究目的にそぐわないサンプルをコーパスから抽出してしまう可能性や，コーパスから得られた分析結果を意図としたものとは異なるテクストの特徴と関連づけてしまう可能性がある。

1.4　語彙密度を用いたテクスト分類の試み

現行のドキュメントアノテーションからは把握することが困難だと考えられるサンプルの特徴を捉えるために，データ班サンプリングサブグループでは，以下の三つの試みを行っている。

人手によるアノテーション 1 場面設定や形式，硬軟などの情報の付与

人手によるアノテーション 2 過程構成 (TRANSITIVITY) 分析を用いた修辞的目的に関する情報の付与

自動解析に基づくアノテーション 語彙密度を用いた情報の詰め込み (information packing) 度に関するデータの付与

本報告書で述べるのは，自動解析に基づくアノテーション「語彙密度を用いた情報の詰め込み (information packing) 度に関するデータの付与」についてである。語彙密度は，英語コーパス研究では頻繁に利用される概念であるが，日本語では計測方法が確立されておらず，活用には至っていない [2, 3, 11]。日本語への応用が可能であれば，日本語コーパス研究にとって，テクストの特徴を把握する手段として有用な概念となり得る。そこで，以下の二つについて検討を行った。

1. 日本語における語彙密度の計測方法
2. 語彙密度と現行のドキュメントアノテーションが示すテクストの特徴との関係

1. については，語彙密度の計測に必要となる節数の計測方法と語彙密度の計測対象となる語の特定を行った。2. については，表 1.2 に示した語彙密度と Field，Tenor，Mode 要因との関係について調べた。また，BCCWJ に含まれるサンプルの特徴を表す指標の一つとしての，語彙密度の可能性について検討した。

以下，第 2 章で語彙密度の概念について述べ，第 3 章で BCCWJ における語彙密度の計測方法を説明する。続いて第 4 章で語彙密度と現行のドキュメントアノテーションが表すテクストの特徴がどのような関係にあるかを，分析結果に基づき示す。さらに第 5 章でサンプルの特徴を把握するため指標としての，語彙密度の可能性について述べる。

第2章　語彙密度の概要と枠組み

佐野大樹

本章では，語彙密度の概念について説明する。まず，語彙密度が提案された理論的枠組みである，システミック理論について概説し，その後，語彙密度の計測に関わる，名詞化，節 (ranking clause)，内容語と機能語の位置づけについて説明する。

2.1　システミック理論

語彙密度はシステミック理論の枠組みから，Halliday[5] によって提示された概念である。システミック理論では，言語は選択体系 (system) ，つまり，'a set of interrelated choices for making meaning' (意味を創るための選択体系) と考えられている [8]。Halliday and Matthiessen[10] では，この考えを以下のように説明している。

> A characteristic of the approach we are adopting here, that of systemic theory, is that it is *comprehensive*: it is concerned with language in its entirety, so that whatever is said about one aspect is to be understood always with reference to the total picture. At the same time, of course, what is being said about any one aspect also *contributes* to the total picture; but in that respect as well it is important to recognize where everything fits in. There are many reasons for adopting this systemic perspective; one is that languages evolve - they are not designed, and evolved systems cannot be explained simply as the sum of their parts. Our traditional compositional thinking about language needs to be, if not replaced by, at least complemented by a 'systems' thinking whereby we seek to understand the nature and the dynamic of a semiotic system as a whole. (pp.19-20)

> システミック理論の特徴の一つは包括的であるということである。この理論では，ある一つの言語の側面は常に全体像の中で位置づけられることによって理解される。無論，一つの側面は全体像を構成する一要因であるわけだが，この意味からしても，どの要因がどこに位置づけられるのかを認識することは重要である。
>
> このような包括的な観点を適用する必要性は多数あるが，その一つは，言語は計画的に構築されてきたものではなく，進化の結果，今の状態にあるということである。進化によって形成されるシステムは，単に，構成要素の集合としては説明

> することができない。構造主義的観点はシステム主義的観点と補完的に考慮されるべきである。これにより，意味体系の性質とダイナミックを言語体系全体の中で捉えることができる。
>
> (訳: 筆者)

ここに示されているように，システミック理論では，テクストは「そこで何が話されているのか (what is being said)，書かれているのか」を解釈するだけでなく,「何を話すことができたのか (what could be said)，書くことができたのか」を考慮することで，ある特定の表現を言語の全体像 (with reference to the total picture) の中で理解することが重要であると考えられている。

このような理論に必要となるのはテクストの特徴を多角的に解釈するための分析方法である。システミック理論では，ある表現を言語システムの相互関係の中で捉えるために，system(選択体系)，structure(構造)，metafunctional diversification(メタ機能)， stratification(分層化)，instantiation(インスタンス化) という五つの概念をベースにした様々な分析方法が用意されている。詳細を説明することは避けるが，表 2.1 では，一つの節に対して，TRANSITIVITY(過程構成)，MOOD(叙法)，THEME(主題) の三つの方法を用いて分析が行われている [10]。

表 2.1: システミック理論における分析

these containers have been washed in the dishwasher

分析法	these containers	have been washed	in the dishwasher
主題分析	Theme	Rheme	
叙法分析	Subject	Finite / Predicator	Adjunct
	indicative: declarative		
過程構成分析	Goal	Process	Circumstance: Place
	material:transitive:receptive:non-agentive		

例えば，過程構成分析では，‘material: transitive: receptive: non-agentive’ という解釈が一つの表現に対して与えられている。この分析は，言語表現として選ぶことが可能であった他の選択肢と，システムネットワークと呼ばれる言語記述法によって体系づけられて考えられる。過程構成のシステムネットワークの一部を図 2.1 に示す。

表 2.1 の分析では，図 2.1 に示されたシステムネットワークの左から選択が始まり，まず，‘clause’ の選択肢として，‘material’ が選ばれ，次の選択肢から，‘transitive’ が選ばれ，さらに，‘receptive’，‘non-agentive’ が選ばれたことが，‘material: transitive: receptive: non-agentive’ として分析に示されている。なお，四角で囲まれた部分は，分析の判断基準として用いられる構造的特徴 (relalisation statement) である。このような方法で分析することで，‘material’ という言語表現を選ぶことで，選ばれなかった他の選択肢との関係について示し，なぜ特定の選択肢が選ばれ，他が選ばれなかったのかを考慮する。これにより，‘material: transitive: receptive: non-agentive’ という選択を他の選択体系の中に位置づけて解釈することができる。

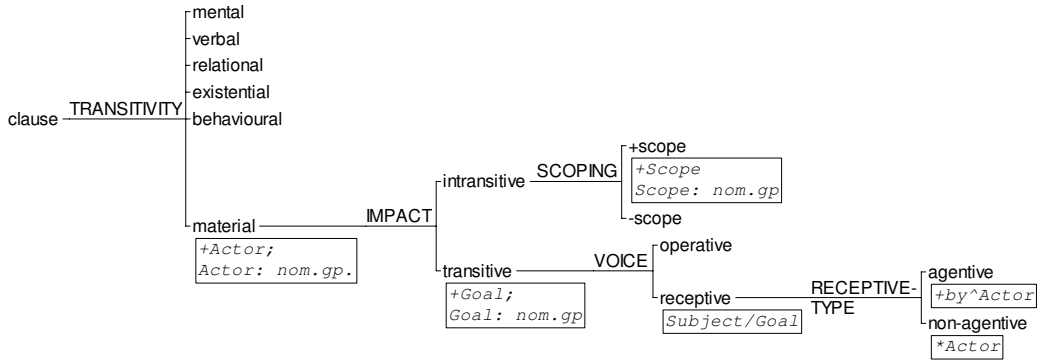

図 2.1: TRANSITIVITY システムネットワーク

このように，複数の分析方法を用い，また，システムネットワークからどの選択肢が選ばれたのかを特定することで，一つの言語現象に対して多角的にアプローチする。このことによって「instance」(what is being said) としての言語を「potential」(what could be said) としての言語の中で捉えようとする [10]。

2.2 語彙密度の概要

このような枠組みの分析方法の一つとして，語彙密度の概念が提案された。語彙密度はテクストにおける情報の密度を示すものであり，以下のように定義づけられている [6]。

> This [Lexical density] is a measure of the density of information in any passage of text, according to how tightly the lexical items (content words) have been packed into the grammatical structure [clause].(p.22)
>
> これ [語彙密度] は，あるテクスト (全体，もしくは一部) における情報の密度を計測する方法であり，当該の文法ユニット [節] にどれだけの内容語が詰め込まれているのかによって計測することができる。
>
> (訳: 筆者)

語彙密度の値が表すのはテクストにどれだけの情報が詰め込まれているかである。なお，テクストの情報の密度を高める言語行為は information packing と呼ばれている。基本的に，語彙密度は以下の式によって求められる [10]。

$$\text{語彙密度} = \frac{\text{テクストに含まれる内容語の総数}}{\text{テクストに含まれる節 } (\textit{ranking clause}) \text{ の総数}}$$

例えば，以下の文章の語彙密度は，次のように計測される [10]。

> In bridging river valleys, the early engineers built many notable masonry viaducts of numerous arches.(p.654)

まず，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

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言葉での語彙密度は約 2，一般的な書き言葉では約 6，科学的文章ではさらに上がるという [5, 6]。

2.3　語彙密度と動詞群の名詞化

先述したように，語彙密度計測では，情報の詰め込み度を計測するために，内容語と節 (ranking clause) 数を計測する。情報の密度の計測に，内容語，及び，節 (ranking clause) 数を指標として用いるのは，これらが情報の詰め込み度を高くする「名詞化」(nominalisation) と強い関係があるためである。この関係を Butt 他 [2] の 2 文を比較することで示す。

例 2-1 If you drink too much alcohol //when you drive your car, //you are likely to have an accident.//

例 2-2 Excessive consumption of alcohol is a major cause of motor vehicle accidents.//(p.60)

*「//」は節境界 (ranking clause) を示す。

これらの文は，ほぼ同じ情報を提示するものであるにも関わらず，一節あたりに含まれる情報量は，例 2-1 に比べ，例 2-2 のほうが多い。このような情報の詰め込みを可能とする方法の一つが，文法的比喩 (Grammatical Metaphor) の一種である，動詞群 (verbal group) の名詞化である。

2 文を比較すると，図 2.2 に示すように，例 2-1 では動詞群として表されていた内容が，名詞化によって，例 2-2 では名詞群 (nominal group) の一部として表されていることがわかる[1]。

このように動詞群が名詞化されることで，例 2-1 で動詞群として表されていた内容は節の述部として表現される必要がなくなる。名詞となることで，述部としての役割から逸し，節構造を保持する形式的制約がなくなる。例 2-1 で節として表されていた情報は内容語を名詞群に詰め込むことで表現できるようになる。このため，一名詞群あたりの内容語数が増える代わりに節数が少なくなり，結果的に節あたりの内容語数は増加する。このような言語的特徴が科学的文章のような情報が詰め込まれているテクストでは頻出する傾向にあったため，Halliday[5] は内容語数と節数を指標として語彙密度の概念を提案した[2]。

[1] 名詞化の詳細については，佐野 [17] 参照。
[2] 同様の傾向が，日本語においても見られると考えられる [14]。

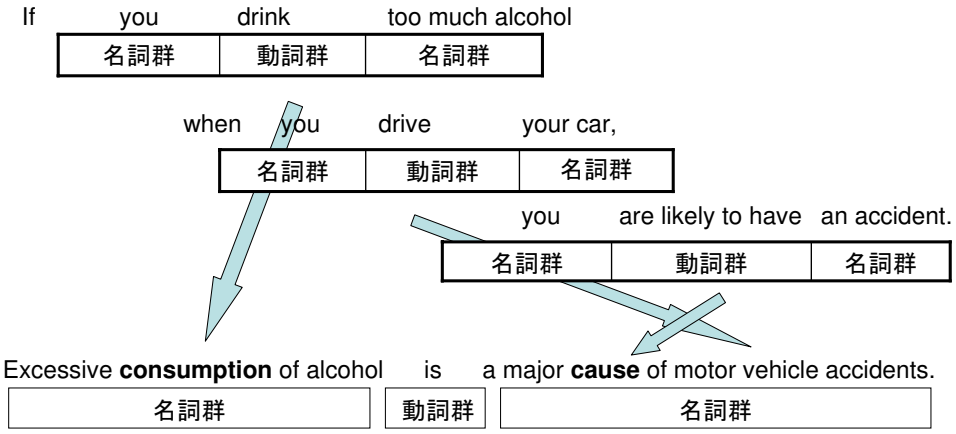

図 2.2: 名詞化による節数の減少，及び，内容語数の増加

2.4 節とランク

語彙密度を計測するためには，節 (ranking clause) 数を計測する必要があるが，一般に節（clause）と呼ぶ対象と語彙密度で計測する対象の節（ranking clause）との間には差があるので注意が必要である。

ranking clause とは，節の中でも特に，clause ランクで機能する節のことをさす。システミック理論では，語彙–文法に関するユニットとして以下の四つが設定され，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。

- clause(節)
- group/phrase(群/句)
- word(語)
- morpheme(形態素)

例えば，‘The eldest oyster winked his eye’ の節–群/句–語の階層構造は以下のようになる [7]。

The eldest oyster winked his eye.

節	clause					
群/句	nom.gp			verbal gp.	nom.gp	
語	The	eldest	oyster	winked	his	eye

ここで，ある節構造をもつユニットがあったとして，それが clause ランクで機能する節であれば，ranking clause として認められる。一方，ある節構造をもつユニットが group/phrase ランクの要素，もしくは group/phrase ランクを構成する一要素として機能するのであれば，それは ranking clause としては認められない。以下に例を示す [7]。

(a) Since Caesar was dead, everyone cried.

(b) That Caesar was dead was obvious to all.(p.266)

この 2 文において,「Caesar was dead」というユニットはいずれも節構造をもつ。しかしながら，(a) の場合，このユニットが機能するのは clause ランクであるのに対して，(b) の場合，このユニットが機能するのは group/phrase ランクである，という違いがある。

(a) **Since Caesar was dead, everyone cried.**

節	clause	clause	
群/句	Since Caesar was dead,	everyone	cried

(b) **That Caesar was dead was obvious to all.**

節	clause			
群/句	That Caesar was dead	was	obvious	to all

(b) の「Caesar was dead」は group/phrase ランクで機能するものである以上，ranking clause であるとは認められない。それゆえ，語彙密度を計測する際，節 (ranking clause) 数の計測対象とはならない。(b) の場合,「That Caesar was dead was obvious to all」全体が ranking clause 数として計測される対象となる。

ranking clause のみを計測対象とすることで，文（sentence）とは異なり，文法機能として同一のランクで機能するユニットあたりの内容語数平均値を求め，テクストの情報の密度について測定することができる。

2.5　内容語と機能語

システミック理論 [5] において，内容語は以下のように定義されている。

> Lexical items are often called ‘content words’. Technically, they are ITEMS (i.e. constituents of variable length) rather than words in the usual sense, because they may consist of more than one word: for example, *stand up, take over, call off*, and other phrasal verbs all function as single lexical items. They are LEXICAL because they function in lexical sets not grammatical systems: that is to say, they enter into open not closed contrasts.(p.63)

> 語彙アイテムは，一般的に，内容語と呼ばれる。厳密には，内容語はアイテム (すなわち，長さの異なる構成要素) であるといえる。このように呼ぶのは，例えば，stand up，take over，call off などのような複数の語からなる句動詞 (phrasal verb) も，他の動詞と同様に，一つのアイテムとして機能するためである。内容語が語彙的であるというのは，文法システムの中でなく，語彙セットの中で機能するため，つまり，閉鎖したシステムの中ではなく，オープンセットの中に位置づけられるためである。
>
> (訳: 筆者)

つまり，内容語は「open contrasts」—class membership に限りがない— という特徴をもつ語，もしくは,「ITEM」のことをさす。すなわち，ある語を基準に，それに何らかの形で関連する語をあげた場合，関連する語の語数が限定されないもののまとまりを内容語というわけである。この点を補足する形で，機能語は以下のように定義づけられ，内容語と区別される。

> A grammatical item enters into a closed system. For example, the personal pronoun *him* contrasts on one dimension with *he, his*; on another dimension with *me, you, her, it, us, them, one*; but that is all. There are no more items in these classes and we cannot add any. With a lexical item, however, we cannot close off its class membership; it enters into an open set, which is indefinitely extendable. So the word *door* is in contrast with *gate* and *screen*; also with *window, wall, floor* and *ceiling*; with *knob, handle, panel* and *sill*; with *room, house, hall*; with *entrance, opening, portal* – there is no way of closing off the sets of items that it is related to, and new items can always come into the picture.(p.63)
>
> 文法アイテムは，閉鎖したシステムの中に位置づけることができる。例えば，代名詞 him は，一つの観点からは，he，his と，他の観点からは，me，you，her，it，us，them，one と対照をなすが，これで全てである。これら以外に him と対照をなすものはなく，また，既存のリストに追加することもできない。しかしながら，語彙アイテムの場合，このようにクラスメンバーシップを閉じることはできない。オープンセットに位置づけられる語は，拡張がどこまでも可能である。ゆえに，例えば door という語は，gate，screen とも対照をなすし，また，window，wall，ceiling とも対照をなす。さらに，knob，handle，panel，sill とも，room，house，hall とも，entrance，opening，portal とも対照をなす。当該の語に関連するセットを閉じることができず，また，新たなアイテムを加えることもできる。
>
> (訳: 筆者)

機能語は「closed contrasts」—class membership に限りがあるもの— という特徴をもつ。この点で open constracts である内容語と異なる。機能語は，ある語を基準とした場合，関連する語が限定されているもののまとまりである。

また，内容語と機能語の特徴として，以下のように説明されている。

> Another aspect of the distinction between lexical and grammatical words is that grammatical items tend to be considerably more frequent in occurrence. A list of the most frequently occurring words in the English language will always be headed by grammatical items like *the* and *and* and *it*. Lexical items are repeated much less often.(p.64)
>
> もう一つの内容語と機能語の違いは機能語は内容語に比べ出現頻度が高いということである。英語における語彙頻度上位リストには，the，and，it などといった機能語が入る。これに比べ，内容語の出現頻度は高くない。
>
> (訳: 筆者)

このように，内容語が機能語との比較によって位置づけられるわけであるが，語彙密度の計測対象となる内容語について，Halliday[5] は以下のように述べている。

> As you would expect, there is a continuum from lexis into grammar: while many items in a language are clearly of one kind or the other, there are always likely to be intermediate cases. In English, prepositions and certain classes of adverb (for example, MODAL adverbs like *always, perhaps*) are on this borderline.(p.63)
>
> 予想されるとおり，語彙と文法は連続体である。明確にどちらかに分類できる語もあれば，どちらにも分類できそうな語もある。英語では，前置詞，及び，副詞の一部，例えば，for example，alway や perhaps などのモーダルな副詞がこれにあたる。
>
> (訳: 筆者)

また，以下のようにも述べている。

> We have been assuming a simple measure in which all lexical items count the same. But the actual effect that we are responding to is one in which the relative frequency of the item plays a significant part. The relative frequency of grammatical items can be ignored, since all of them fall into the relatively frequent bracket. But the relative frequency of lexical items is an important factor in the situation.
>
> The vocabulary of every language includes a number of highly frequent words, often general terms for large classes of phenomena. Examples from English are *thing, people, way, do, make, get, have, go, good many.* These are lexical items, but on the borderline of grammar; they often perform functions that are really grammatical – for example *thing* as a general noun (almost a pronoun) as in *that's a thing I could well do without ...*They therefore contribute very little to the lexical density.(pp.64-65)

> ここまでは，全ての語彙アイテムが等価に計測されることを仮定してきた。しかしながら，[情報の詰め込み度を計測する上では]，語彙頻度が重要な役割を果たす。機能語の頻度に関しては問題ないが，語彙密度計測にとって，内容語の頻度は重要な要因である。
>
> 全ての言語のボキャブラリーには，様々な事象を表す際に利用できる一般的で使用頻度が高い語というものが存在する。英語の場合では，thing，people，way，do，make，get，have，go，good，many などがそれにあたる。これらの語は，内容語ではあるが，文法的でもある。例えば，that's a thing I could well do without ... という文において，thing という語は，文法的である。これらの語は，情報の密度にほとんど貢献しない。
>
> (訳: 筆者)

内容語と一般的に認識される語であっても，文脈によっては文法的振る舞いをするものもあり，情報の密度への影響という観点から見れば，ほとんど貢献しない語もある。計測対象に含める必要のない，もしくは，重みづけによって語彙密度計測における扱いを考えるべき内容語もあるというわけである。このような観点から，英語においては，語彙密度計測の対象外となる語がリスト化されている [11]。

語彙密度の概念を日本語に適用するには，日本語における節 (ranking clause) 数の計測方法，及び，語彙密度計測で対象とする内容語の特定が必要となる。次章では，BCCWJ に含まれるサンプルの語彙密度を計測するために，どのように節 (ranking clause) 数を計測し，語彙密度の計測対象となる内容語を特定したかを説明する。

第3章　語彙密度の計測方法

佐野大樹

本章では，BCCWJ における語彙密度の計測方法について示す。語彙密度計測には，1) 語彙密度計測の対象となるサンプル内の要素を特定し，2) そこに含まれる節 (ranking clause) 数を計測し，3) 語彙密度の計測対象となる内容語数を計測することが必要となる。それぞれの過程について以下に示す。

3.1　語彙密度計測の対象となる要素の特定

第 2 章で説明したとおり，語彙密度計測の対象となるのは clause ランクに属するユニット (ranking clause) である。したがって，組織名の羅列など，group ランク以下の構造をもつユニットによってのみ構成される要素は計測の対象とならない。

そこで本研究では，語彙密度の計測対象となる要素を BCCWJ 電子化フォーマットに基づいて付与されている XML タグを利用して特定した [24]。BCCWJ 電子化フォーマットで定義する XML タグには，サンプルに関するタグ，文字・表記に関するタグ，文書構造に関するタグの 3 種があり，これらを用いることで, ある言語表現のテクスト中の役割などについて同定することが可能である。

これらの XML タグを利用して，group ランク以下に属するユニットで構成されることが多い表 3.1 の要素を計測対象から除外した。表 3.1 にある要素中には，noteBody など，節構造をもつ文法ユニットが含まれる可能性があるものもあるが，節構造をもたないユニットを含むことも多いため，計測対象外として扱った。これらの要素を除外した残りの言語表現—`sentence` タグが付与された言語表現を中心とする— が，語彙密度の計測対象要素となる[1]。

3.2　節数の計測

節 (ranking clause) 数の計測には，節境界検出プログラム CBAP(Clause Boundaries Annotation Program) を用いた。CBAP は，局所的な形態素の連接を対象としたパタンマッチによって，節境界の位置と種類を自動認定するもので，節境界の直後に，以下の例文のような，「連体節」,「並列節デ」,「理由節ノデ」など，147 種の節境界ラベルを付与する [23]。

> 身長の二乗掛ける/**連体節**/ 二十二が標準体重ということになっとりまして/**テ節**
> / 私の標準体重は/**主題ハ**/ 六十四キロなんですが/**並列節ガ**/ それから見ると/**条**

[1] setence タグが付与されるもののうち，属性が「quasi」でないもの

表 3.1: 語彙密度計測の対象から除外される要素

要素名	説明
authorsData	記事構造上，著作者表示・署名に当たる要素
caption	図表についてのタイトルや説明を表す要素
contents	目次に相当する文書要素
delete	抹消線などによって削除された本文要素
figureBlock	図表・写真などの要素と，それに付随する文書要素をまとめた要素
noteBody	脚注・後注など，本文と区別して記述される注記
noteBodyInline	傍注など，行外に付随する形式で現れる注記
noteMarker	他の文章要素を参照する際の目印として機能する文字列
orphanedTitle	不特定範囲の文書要素を代表する記述
profile	文書要素著者や登場人物のプロフィールに相当するもの
sentence type 属性 quasi	文区切り文字以外の基準により自動付与された sentence 要素
speaker	話者を明示的に表した文字列やマーク
table	表を表すもの
title	特定範囲の文書要素の内容を代表する記述
titleBlock	title 要素とそれに付随する要素全体
verse	詩や和歌などの韻文を表すもの

*各要素の詳細は，山口他 [24] を参照。

> **件節ト**/ 約七キロぐらいは/**主題ハ**/ 減量が必要ということで/**並列節デ**/ 運動をする/**連体節**/ 方がいいことになってまして/**テ節**/ 本当は食事量を減らすという/**連体節トイウ**/ ことなんでしょうけど/**並列節ケレドモ**/ なかなかそれは/**主題ハ**/ 難しいので/**理由節ノデ**/ 私は/**主題ハ**/ 専ら運動の方で健康を維持しようという/**連体節トイウ**/ ことに努めとります。/**文末**/ (p.47)

ただし，CBAP における「節」の定義と語彙密度の計測に用いられる節 (ranking clause) は必ずしも同義ではない。CBAP では，節は「「述語を中心としたまとまり」と定義される文の構成要素」と捉えられている [23]。一方，語彙密度の計測では，第 2 章で説明したとおり，語彙–文法単位のうち，clause ランクで機能するものを節 (ranking clause) とする。

例えば，CBAP で連体節と解析されたユニットは group/phrase ランクの一構成要素として機能するものと考えられ，語彙密度計測において節 (ranking clause) 数としてはカウントされない。

この違いを踏まえて，CBAP で検出される節境界ラベルを，group/phrase ランクで機能すると考えられるものと，clause ランクで機能すると考えられるものとに分類し，節 (ranking clause) 数の計測に用いる対象を絞り込んだ。CBAP が付与する節境界ラベルのうち，語彙密度の計測対象，及び，計測対象外となる節 (ranking clause) の一部を表 3.2 に示す[2]。

[2] 節境界ラベルの詳細については，丸山他 [23] を参照。

表 3.2: 節数計測の対象，及び，対象外とする節境界ラベル

計測対象	計測対象外
連用節	間接疑問節
並列節ケレドモ	連体節-形式名詞
並列節タリ	連体節タメノ
条件節カギリ	連体節トイウ
条件節タラ	連体節ヨウナ
譲歩節テモ	形容詞連体節
理由節カラ	形容詞連体節-形式名詞
理由節ノデ	形容動詞連体節
時間節アト	感嘆詞
時間節イマ	間投句
時間節トキ	談話標識
テ節	主題ハ

例えば，以下のようなテクストを解析した場合，

自身のもつ自然の回復力を待っているのである。これを「自然治癒力」という。

『噛み合わせバカにしてると恐ろしい』 山田敏輔 LBp4_00025

CBAP の解析結果は，以下のようになる。

自身のもつ/**連体節**/自然の回復力を待っているのである。/**文末**/
これを「自然治癒力」という。/**文末**/

「自身のもつ」という連体節は節構造をもつが,「自然の回復力」の修飾部であり,「自身のもつ自然の回復力を」が一つの名詞群と見なされる。つまり，このユニットは，節構造をもっていたとしても,「自身のもつ自然の回復力を待っているのである」と同じ clause ランクでなく，下位階層で機能している。したがって，節 (ranking clause) 数にはカウントされない。このような基準で，各サンプルごとの節数を計測した。なお，CBAP の検出精度は 97%以上であるが，以下の四つの検出誤りの可能性があることが確認されていることを述べておく [23]。

1 形態素解析の誤りに起因する検出誤り

2 そもそも検出することが困難な節境界

3 節境界検出ルール自体の問題

4 複合動詞に関する検出誤り

3.3　語彙密度計測の対象となる内容語の特定

語彙密度の計測対象となる内容語を特定するために，計測対象となる内容語の条件を規定した。第 2 章で説明したとおり，ある種の内容語は，文脈によっては機能語的に振る舞うため，情報の密度にはほとんど貢献しないと考えられる。そこで，計測対象とする内容語について条件を付しておくことが必要となる。

本研究では，形態素解析辞書 UniDic の品詞ごとに，語彙密度の計測対象とするかどうかを検討した[3]。情報の密度への貢献が少ないと考えられる品詞を対象語から除外し，計測対象とする内容語の範囲を限定した。以下，特定方法の詳細について示す。

まず，第 2 章で述べたように，内容語と機能語は以下のように定義される。

内容語 関連した語の範囲が限定できないもののまとまり。機能語に比べ，出現頻度が低い。

機能語 関連した語の範囲が限定できるもののまとまり。内容語に比べ，出現頻度が高い。

この定義に基づけば，大規模コーパスで各品詞ごとの異なり数と述べ数を計測した場合，以下のような違いが認められるはずである。

内容語 異なり数が多く，また，異なり数/延べ数 (Type/Token Ratio 以下，TTR) は機能語に比べ高くなる。

機能語 異なり数が少なく，また，TTR は内容語に比べ低くなる。

この違いに着目し，山崎他 [25] の語彙頻度データを利用して，品詞ごとの異なり数と，延べ数，及び，TTR を計測した。計測結果を，表 3.3 に示す[4]。

表 3.3 を見ると，延べ数，異なり数，TTR 全てに，品詞ごとに顕著な差が認められる。品詞ごとの延べ数，異なり数，TTR の分布を，図 3.1，図 3.2，図 3.3 に示す。図 3.1 は延べ数の分布，図 3.2 は異なり数の分布，図 3.3 は TTR の分布を示す。

[3] UniDic の品詞体系には，大分類，中分類，細分類がある。本研究では，細分類まである品詞に関しては，細分類までを考慮した。UniDic の品詞体系に関する詳細については小椋他 [15] を参照。

[4] なお，辞と解析されるユニットは機能語と考えられるため，表には含まれていないことを述べておく。

表 3.3: 品詞別延べ数，異なり数，TTR

品詞	延べ数	異なり数	TTR
感動詞–フィラー	10,040	25	0.00249004
感動詞–一般	15,560	1,089	0.069987147
形状詞–タリ	2,339	207	0.088499359
形状詞–一般	53,908	870	0.016138607
形状詞–助動詞語幹	49,173	5	0.000101682
形容詞–一般	95,584	694	0.007260629
形容詞–非自立可能	71,969	6	8.33692E-05
助詞–格助詞	2,081,759	23	1.10483E-05
助詞–係助詞	512,720	7	1.36527E-05
助詞–終助詞	81,636	33	0.000404233
助詞–準体助詞	112,662	2	1.77522E-05
助詞–接続助詞	536,400	27	5.03356E-05
助詞–副助詞	156,553	50	0.000319381
助動詞	1,073,874	89	8.28775E-05
接続詞	55,656	46	0.000826506
代名詞	174,670	110	0.000629759
動詞–一般	786,272	6,307	0.008021397
動詞–非自立可能	816,887	114	0.000139554
副詞	194,132	1,532	0.007891538
補助記号–一般	13	7	0.538461538
名詞–固有名詞–一般	16,140	962	0.05960347
名詞–固有名詞–人名–一般	26,623	3,069	0.115276265
名詞–固有名詞–人名–姓	58,596	3,893	0.066437982
名詞–固有名詞–人名–名	64,261	3,270	0.05088623
名詞–固有名詞–組織名	8,803	815	0.092582074
名詞–固有名詞–地名–一般	67,471	4,779	0.070830431
名詞–固有名詞–地名–国	41,241	424	0.010281031
名詞–助動詞語幹	1,277	2	0.001566171
名詞–数詞	345,469	139	0.000402352
名詞–普通名詞–サ変可能	700,443	10,355	0.014783501
名詞–普通名詞–サ変形状詞可能	11,264	82	0.00727983
名詞–普通名詞–一般	1,922,957	55,278	0.028746353
名詞–普通名詞–形状詞可能	117,596	1,802	0.01532365
名詞–普通名詞–副詞可能	284,857	625	0.002194083
連体詞	116,539	60	0.000514849

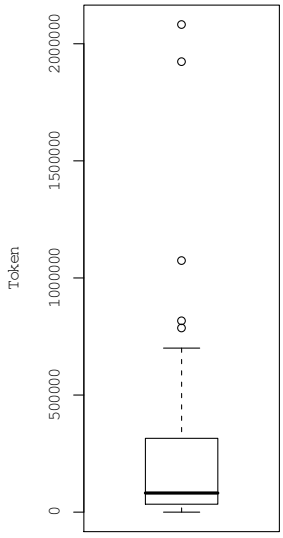

図 3.1: 延べ数の分布

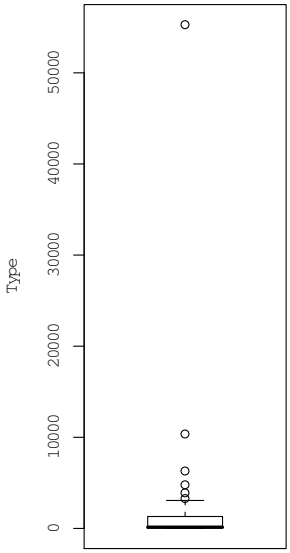

図 3.2: 異なり数の分布

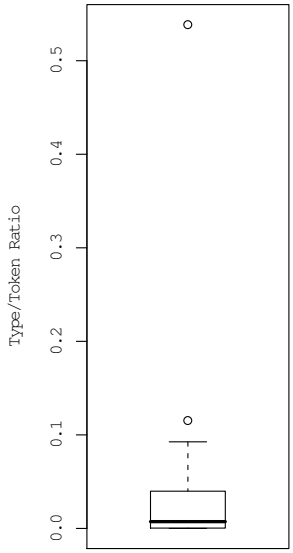

図 3.3: TTR の分布

延べ数上位 3 品詞は「助詞–格助詞」(2,081,759)「名詞–普通名詞–一般」(1,922,957)「助動詞」(1,073,874) である。一方，下位 3 品詞は「名詞–助動詞語幹」(1,277),「形状詞–タリ」(2,339),「名詞–固有名詞–組織名」(8,803) である。

異なり数上位 3 品詞は「名詞–普通名詞–一般」(55,278),「名詞–普通名詞–サ変可能」(10,355),「動詞–一般」(6,307) である。下位 3 品詞は「名詞–助動詞語幹」(2),「助詞–準体助詞」(2),「形状詞–助動詞語幹」(5) である。

TTR 上位 3 品詞は「名詞–固有名詞–人名–一般」(0.115276265),「名詞–固有名詞–組織名」(0.092582074),「形状詞–タリ」(0.088499359) である。一方下位 3 品詞は「助詞–格助詞」(1.10483E-05),「助詞–係助詞」(1.36527E-05),「助詞–準体助詞」(1.77522E-05) である。

この結果を踏まえて，本研究では，延べ数，異なり数と TTR を条件として以下の二つの条件をたて，いずれか一方の条件にあてはまるものは，語彙密度の計測対象外となる品詞とした。

条件 1 使用頻度が 1 万以上であり，かつ，異なり数が 150 以下の品詞

条件 2 TTR が 0.002 以下の品詞

条件 1

表 3.3 の異なり数のデータを図 3.4，及び，図 3.5 に示す。図 3.4 は全体像を示すものであり，図 3.5 は異なり数 0–300 までの違いに着目したものである。

先述したように，異なり数には品詞ごとに大きな差があるが，図 3.5 を見ると，異なり数 300 以上のものと 300 未満のものに分けることができる。300 以上のものと 300 未満のものの境界には「形状詞–タリ」,「名詞–数詞」,「動詞–非自立可能」などがある。

ここで注意しておきたいのは，異なり数は延べ数によって大きく影響される可能性があるということである。境界付近に位置するものの述べ数を確認すると,「名詞–数詞」は，345,469 であり,「動詞–非自立可能」は 816,887 であるのに対し,「形状詞–タリ」は 2,339 と他の二つに比べて極端に少ない。「形状詞–タリ」については，頻度の低さによる異なり数への影響を否定できない。

そこで本研究では,「形状詞–タリ」を除く，異なり数 300 未満のものを計測対象外品詞とし，条件 1「使用頻度が 1 万以上であり，かつ，異なり数が 150 以下の品詞」を設定した。条件 1 で計測対象外となる品詞の一覧を表 3.4 に示す。

図 3.4: 異なり数による認定 (0-60000)

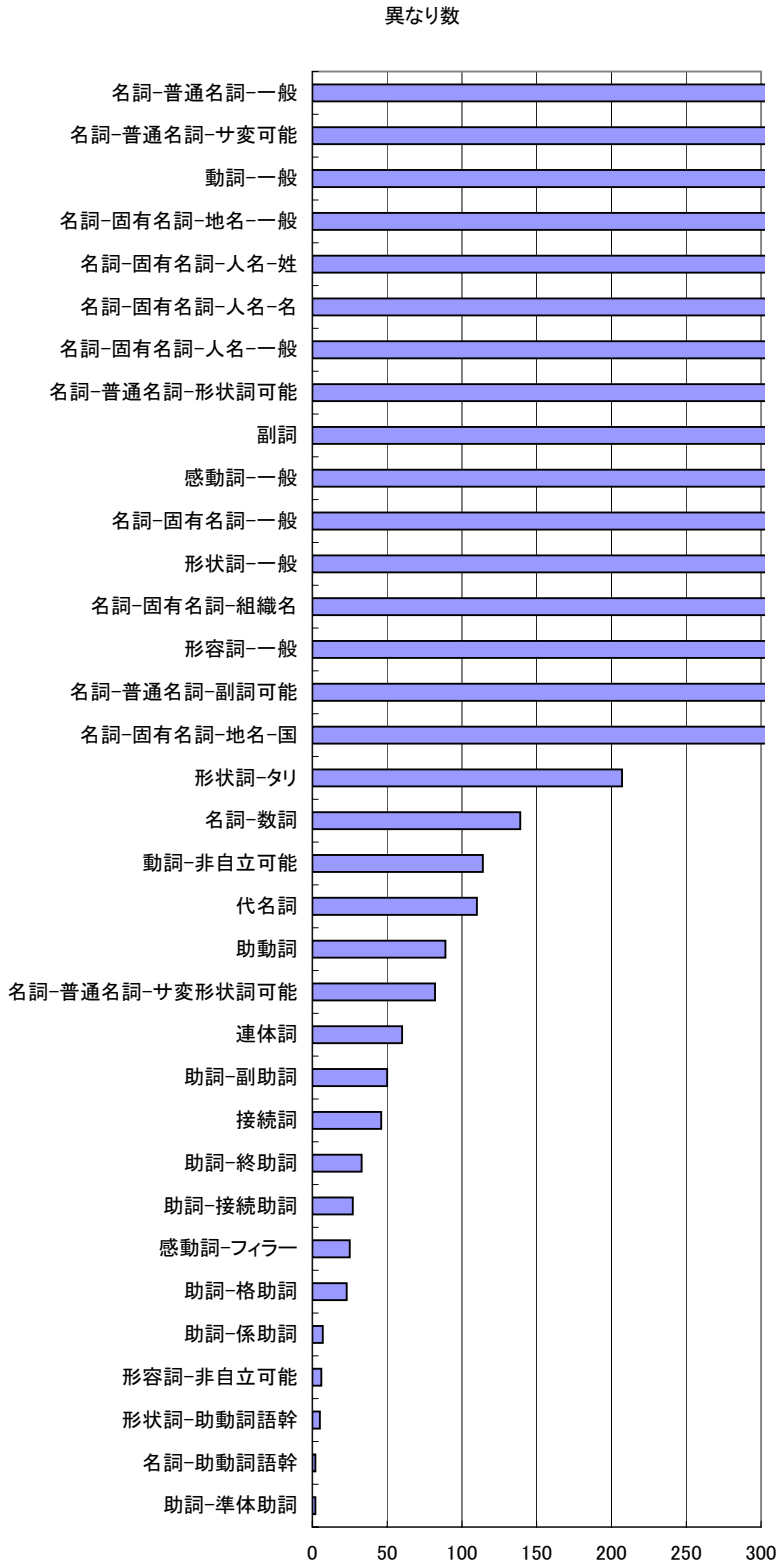

図 3.5: 異なり数による認定 (0-300)

表 3.4: 条件 1 により計測対象外となる品詞

品詞	延べ数	異なり数
感動詞–フィラー	10,040	25
形状詞–助動詞語幹	49,173	5
形容詞–非自立可能	71,969	6
助詞–格助詞	2,081,759	23
助詞–係助詞	512,720	7
助詞–終助詞	81,636	33
助詞–準体助詞	112,662	2
助詞–接続助詞	536,400	27
助詞–副助詞	156,553	50
助動詞	1,073,874	89
接続詞	55,656	46
代名詞	174,670	110
動詞–非自立可能	816,887	114
名詞–数詞	345,469	139
名詞–普通名詞–サ変形状詞可能	11,264	82
連体詞	116,539	60

条件 2

条件 2 は TTR に基づくものである。先述したように，TTR が低い品詞は，機能語的な性質が高いと考えられる。図 3.6，及び，図 3.7 に，表 3.3 の TTR に着目した図を示す。図 3.6 は全体の傾向を表し，図 3.7 は TTR が 0 から 0.004 までの範囲を表す図である。

図 3.7 を見ると，0.004 未満の品詞とそれ以上の品詞で，TTR に大きな違いがあることが認められる。その境界にあるのが,「感動詞–フィラー」,「名詞–普通名詞–副詞可能」と「名詞–助動詞語幹」である。このうち，条件 1 によって計測対象外となっていないものは,「名詞–普通名詞–副詞可能」と「名詞–助動詞語幹」である。

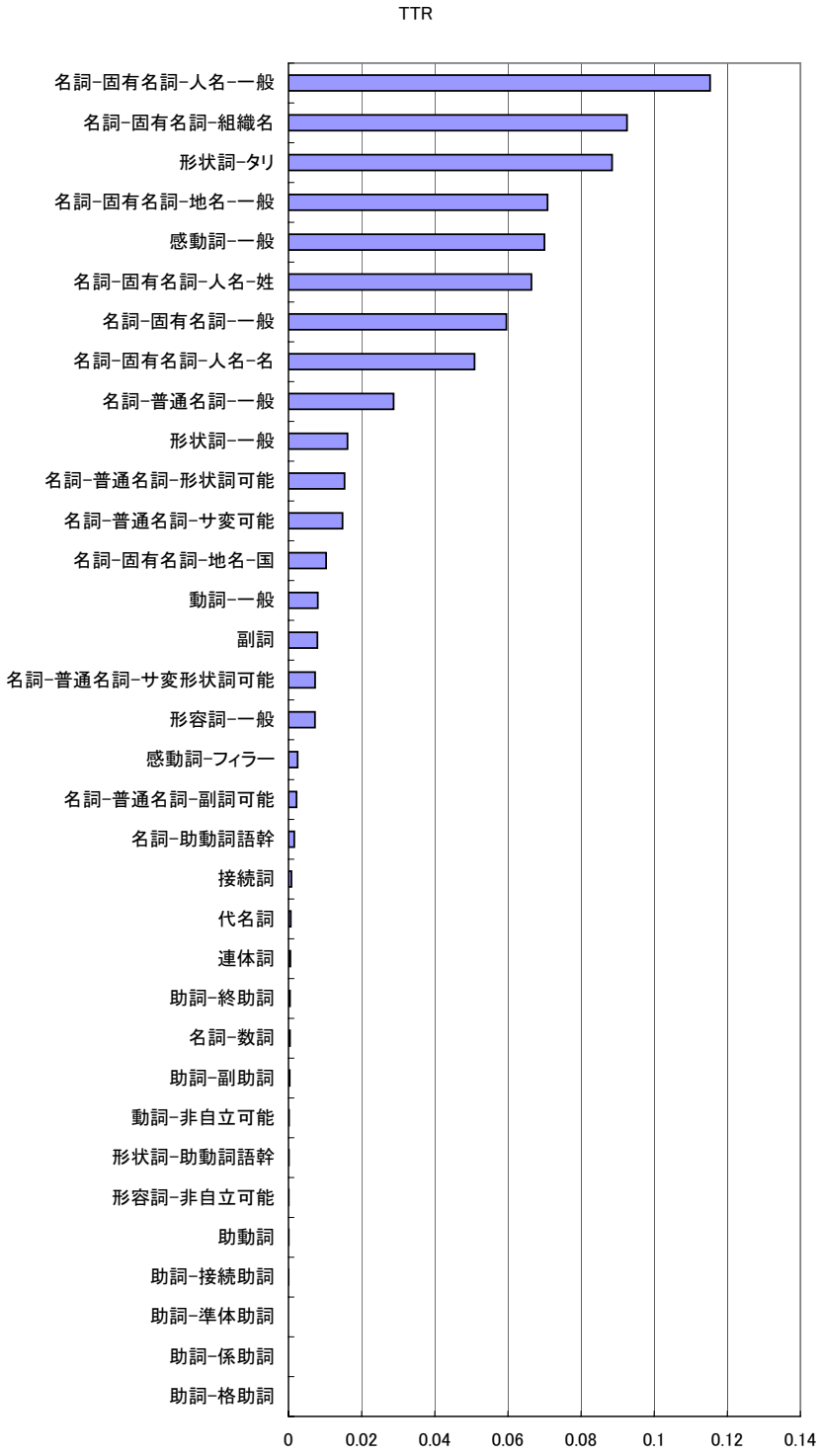

図 3.6: TTR による認定 (0-0.14)

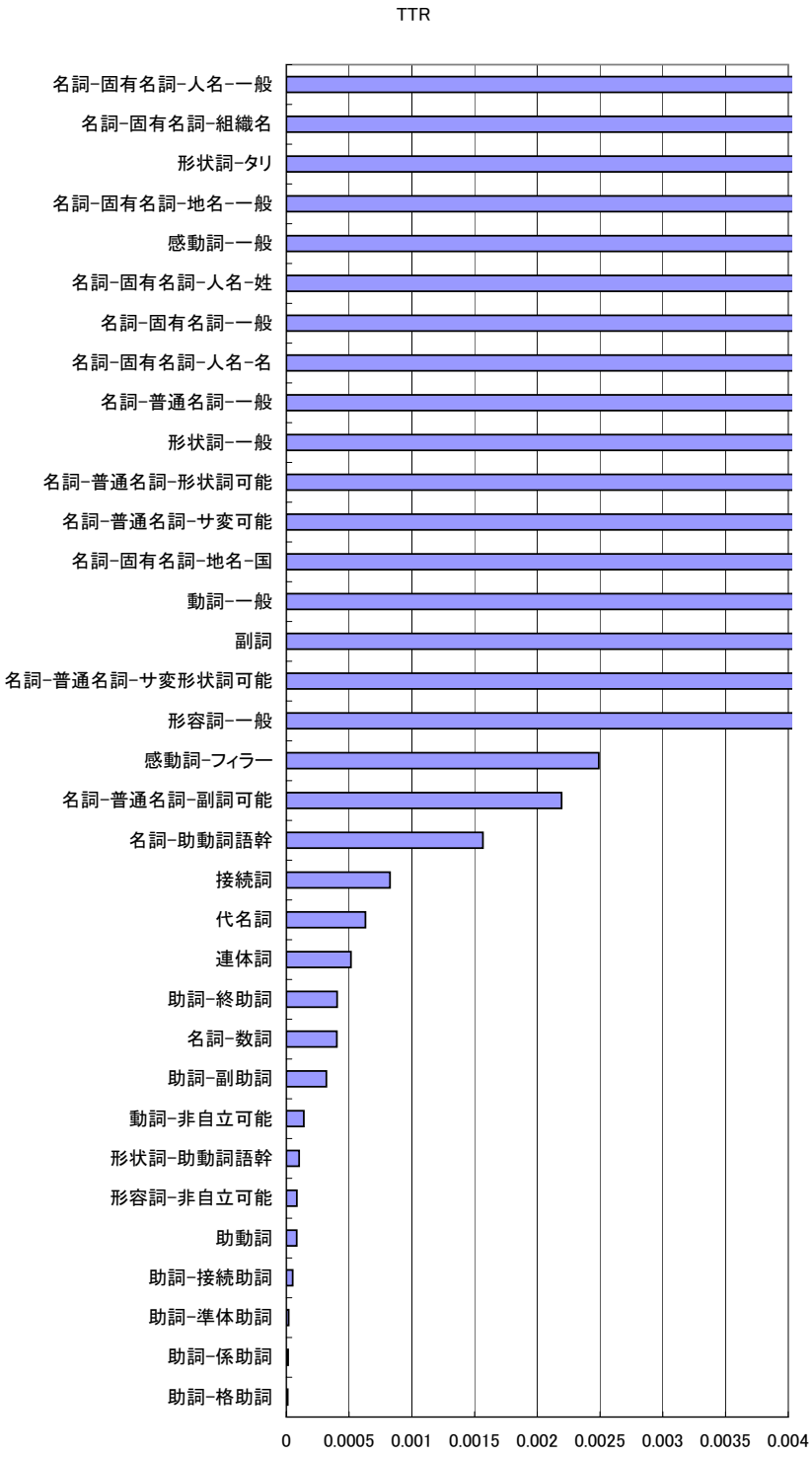

図 3.7: TTR による認定 (0-0.004)

これら二つの品詞の延べ数，及び，異なり数を見ると,「名詞-普通名詞-副詞可能」の延べ数は 284,857 であり，異なり数は 625 である。一方「名詞-助動詞語幹」では，延べ数が 1,277 とそれほど多くないが，異なり数が 2 と極めて少ない。これは，class membership が限られていることを意味し，機能語として性質が高いことを示唆する。これを踏まえて,「名詞-普通名詞-副詞可能」と「名詞-助動詞語幹」の間に境界を認め，条件 2「TTR が 0.002 以下の品詞」を設定した。

語彙密度の計測対象となる品詞

条件 1,2 に基づくと，語彙密度の計測対象となる品詞は表 3.5 の 17 項目となる。「名詞-助動詞語幹」を除き，条件 2 にあてはまるものは条件 1 にもあてはまる。

3.4 語彙密度の計測

表 3.5 にある 17 項目に該当する品詞の語数，及び，節 (ranking clause) 数をサンプルごとに計測し，語数を節 (ranking clause) 数で割ることで，サンプルの語彙密度を計測した。

したがって，サンプルの語彙密度は以下の式によって求めることができる。

$$\text{サンプルの語彙密度} = \frac{\text{サンプルに含まれる 17 品詞に属する語の総数}}{\text{サンプルに含まれる節 } (ranking\ clause) \text{ の総数}}$$

次章では，BCCWJ2008 書籍サンプルの語彙密度を計測した結果について述べる。

表 3.5: 語彙密度の計測対象，及び，計測対象外となる品詞

品詞	計測対象	条件 1	条件 2
名詞–普通名詞–副詞可能	○		
形容詞–一般	○		
副詞	○		
動詞–一般	○		
名詞–固有名詞–地名–国	○		
名詞–普通名詞–サ変可能	○		
名詞–普通名詞–形状詞可能	○		
形状詞–一般	○		
名詞–普通名詞–一般	○		
名詞–固有名詞–人名–名	○		
名詞–固有名詞–一般	○		
名詞–固有名詞–人名–姓	○		
感動詞–一般	○		
名詞–固有名詞–地名–一般	○		
形状詞–タリ	○		
名詞–固有名詞–組織名	○		
名詞–固有名詞–人名–一般	○		
助詞–格助詞		対象外	対象外
助詞–係助詞		対象外	対象外
助詞–準体助詞		対象外	対象外
助詞–接続助詞		対象外	対象外
助動詞		対象外	対象外
形容詞–非自立可能		対象外	対象外
形状詞–助動詞語幹		対象外	対象外
動詞–非自立可能		対象外	対象外
助詞–副助詞		対象外	対象外
名詞–数詞		対象外	対象外
助詞–終助詞		対象外	対象外
連体詞		対象外	対象外
代名詞		対象外	対象外
接続詞		対象外	対象外
名詞–助動詞語幹			対象外
感動詞–フィラー		対象外	
名詞–普通名詞–サ変形状詞可能		対象外	

第4章　語彙密度計測の結果とその分析

佐野大樹

本章では，第3章で提示した計測法を用いて，BCCWJ2008の生産実態サブコーパス (Publication Sub-Corpus以下，PSC)，及び，流通実態サブコーパス (Library Sub-Corpus以下，LSC) の書籍可変長サンプルの語彙密度を計測した結果について述べる。特に，語彙密度と現行のドキュメントアノテーションによって示されるコンテクスト要因との関連性について検討した結果を示す。まず，語彙密度とコンテクスト要因との関係の検証方法について述べ，次に，分析データの概要を説明する。その後，計測結果について述べる。

4.1　語彙密度とコンテクスト要因の関係の分析

以下の分析では，語彙密度と第1章の表1.2にあげた，Field，Tenor，Modeの要因との関係について検討していく。表4.1に，表1.2のコンテクスト要因を再掲しておく。各コンテクスト要因には，さらに細分化されたカテゴリが設けられている。例えば書籍のジャンルであれば,「文学」や「社会科学」などの10のNDCカテゴリに分類されている。

分析では各要因のカテゴリごとに，語彙密度の最大値，最小値，中央値，第1四分点，第3四分点を計測し，カテゴリ間における語彙密度の分布状況の相違について示していく。また，カテゴリ間の平均値の差についても検討する。

表 4.1: BCCWJにおける現行のドキュメントアノテーションと状況コンテクスト (表1.2再掲)

状況コンテクスト	要因
Field	ジャンル，出版年
Tenor	著者性別，著者生年，販売対象 (Cコード)
Mode	形態 (Cコード)，テクスト中のサンプルの位置

4.2　分析データの概要

分析データには，先述したように，BCCWJ2008に含まれる，PSC，及び，LSCの書籍可変長サンプルを用いた。

PSCは書き言葉の生産力に着目したサブコーパスで，2001年から2005年の間に国内で出版された全ての書籍・雑誌・新聞を対象とした母集団からランダムサンプリングされたデータである [21, 22]。

LSC は書き言葉の流通・流布に着目したサブコーパスで，2007 年の時点で東京都内の公共図書館に所蔵されている書籍を対象とした母集団からランダムサンプリングされたデータである [21, 22]。

BCCWJ2008 には，PSC，及び，LSC の書籍データの一部が収録されている。分析データのサンプル数は表 4.2 のとおりである[1]。

表 4.2: 分析データのサンプル数

サブコーパス	総サンプル数	計測可サンプル数
PSC	5,365	5,298 (99.2%)
LSC	4,429	4,397 (98.8%)
合計	9,794	9,695 (99.0%)

PSC には 5,365 サンプル，LSC には 4,429 サンプル，計 9,794 サンプル収録されている。このうち，第 3 章で示した語彙密度計測の対象外要素を除いた結果，節 (ranking clause) 数が 15 以下であるものが，PSC で 1.2%(67)，LSC で 0.7%(32) あった。これに該当するサンプルは計測対象となる言語表現がそもそも少ないため，語彙密度の計測結果がサンプルの特徴を表す情報としては不適切であると考え欠測値とした。なお，PSC，及び，LSC の語彙密度の平均値は表 4.3 のとおりである。

表 4.3: 分析データの語彙密度の平均値

サブコーパス	平均値
PSC	4.7
LSC	4.1

以下，表 4.1 に示したコンテクスト要因ごとに，語彙密度の分布と平均値を示し，その結果を分析する。

4.3　Field 情報から見た語彙密度

4.3.1　ジャンル

BCCWJ のジャンル情報は，書籍の場合，NDC が 3 次区分まで付与されている [16]。分析データの NDC1 次区分カテゴリ別サンプル数を表 4.4 に示す。9. 文学が最も多く，それに 3. 社会科学が続く。

[1] なお，本報告書ではサンプルに対する情報付与・分類を目的とするため，語数ではなく，サンプル数を用いる。

表 4.4: NDC 別サンプル数

NDC	サンプル数	割合
0. 総記	235	2.4%
1. 哲学	561	5.8%
2. 歴史	902	9.3%
3. 社会科学	2,336	24.2%
4. 自然科学	480	5.0%
5. 技術 工学	404	4.2%
6. 産業	275	2.8%
7. 芸術 美術	384	4.0%
8. 言語	207	2.1%
9. 文学	3,787	39.2%
データなし	124	1.3%
合計	9,695	100.0%

NDC カテゴリ別語彙密度の計測結果を図 4.1，及び，図 4.2 に示す。y 軸は語彙密度 (LD) を示す。x 軸は NDC の 1 次区分を示す。図 4.1 は語彙密度の分布を示す。図 4.2 は平均値を示す。エラーバーは標準誤差である。

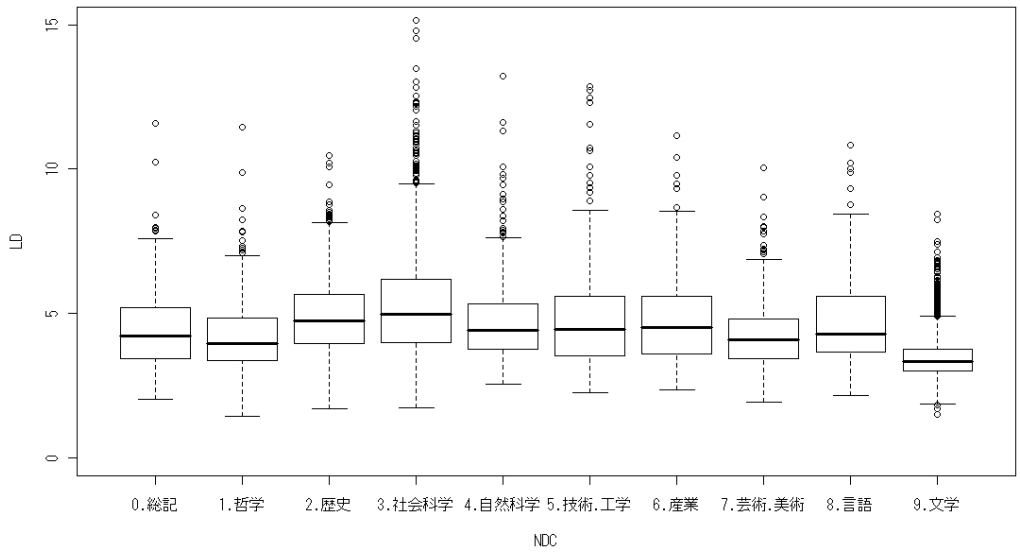

図 4.1: 語彙密度:NDC 別

図 4.1 は NDC のカテゴリ間，及び，カテゴリ内に，語彙密度に大きな差があることを示す。特に，サンプル数が多い 3. 社会科学 (2,336 サンプル) で，語彙密度の分布にばらつきが見られる。しかし，3. 社会科学よりもサンプル数が多い 9. 文学 (3,787) はばらつきが最も小さい。同分類内でばらつきが大きいものとそうでないものが混在していることがうかがえる。

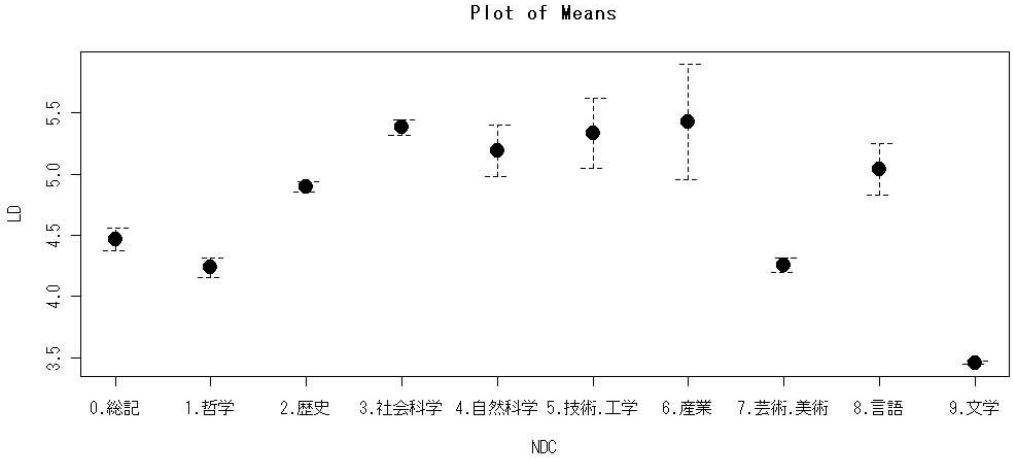

図 4.2: 語彙密度平均値:NDC 別

カテゴリ間の平均値の差に着目すると，図 4.2 に示すように，9. 文学，1. 哲学，7. 芸術. 美術の語彙密度の平均値は他の分類に比べ低く，一方，6. 産業，3. 社会科学，5. 技術. 工学で高いことがわかる。表 4.4 にこの傾向について示す。

表 4.4: 語彙密度の傾向:ジャンル (NDC)

NDC 分類	語彙密度 平均値	
6. 産業	5.4	↑情報の詰め込み度が高い
3. 社会科学	5.4	(語彙密度高)
5. 技術. 工学	5.3	
4. 自然科学	5.2	
8. 言語	5.0	
2. 歴史	4.9	
0. 総記	4.5	
7. 芸術. 美術	4.3	
1. 哲学	4.2	(語彙密度低)
9. 文学	3.5	↓情報の詰め込み度が低い

9. 文学や 7. 芸術. 美術など創作的な分野に属する書籍は語彙密度は低く，情報の詰め込みはあまり行われない傾向があることがうかがえる。一方，6. 産業や 3. 社会科学など，(社会) 科

学的・工学的分野に属するものは語彙密度が高く，情報が詰め込まれたテクストが多いと考えられる。

4.3.2 出版年

語彙密度と出版年との関係の分析にはLSCのデータを用いた。PSCは，BCCWJの設計上，収録期間が2001〜2005年と短く，LSCと母集団の性質が異なる[22]。

そこでここでは，収録期間が1986年〜2005年と長く，PSCに比べ通時的分析に適していると考えられるLSCのデータのみを用いて，語彙密度と出版年との関係について調べた。出版年ごとのサンプル数を表4.5に示す。

表 4.5: 出版年代別サンプル数

出版年	サンプル数	割合
1986	78	1.8%
1987	109	2.5%
1988	101	2.3%
1989	128	2.9%
1990	158	3.6%
1991	179	4.1%
1992	167	3.8%
1993	162	3.7%
1994	210	4.8%
1995	230	5.2%
1996	226	5.1%
1997	257	5.8%
1998	262	6.0%
1999	234	5.3%
2000	244	5.5%
2001	293	6.7%
2002	350	8.0%
2003	319	7.3%
2004	351	8.0%
2005	339	7.7%
合計	4,397	100.0%

1993年以前のサンプル数は200以下と比較的少ないものの，NDCカテゴリほどの数量的偏りは見られない。出版年別の語彙密度の分布傾向を図4.3に，平均値を図4.4に示す。

図4.3，及び，図4.4からは，NDCカテゴリ間に見られたような顕著な差は，出版年の間では見られない。NDCカテゴリでは，平均値の最大値 (6. 産業) と最小値 (9. 文学) に2程度の

差が見られたが，出版年では，0.6 程度の差しかない。但し，1997 年を境にして，徐々にではあるが，語彙密度が減少している傾向があることがうかがえる。

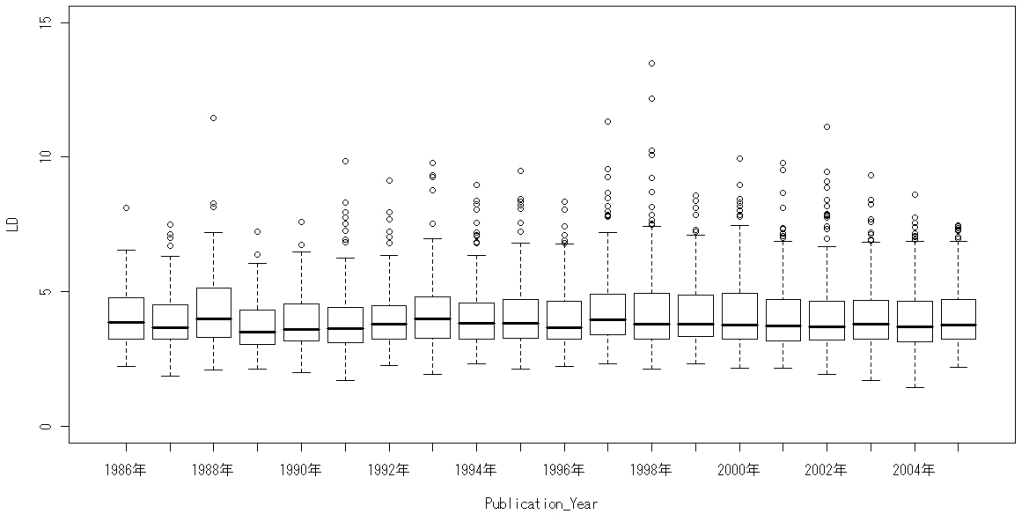

図 4.3: 語彙密度:出版年別

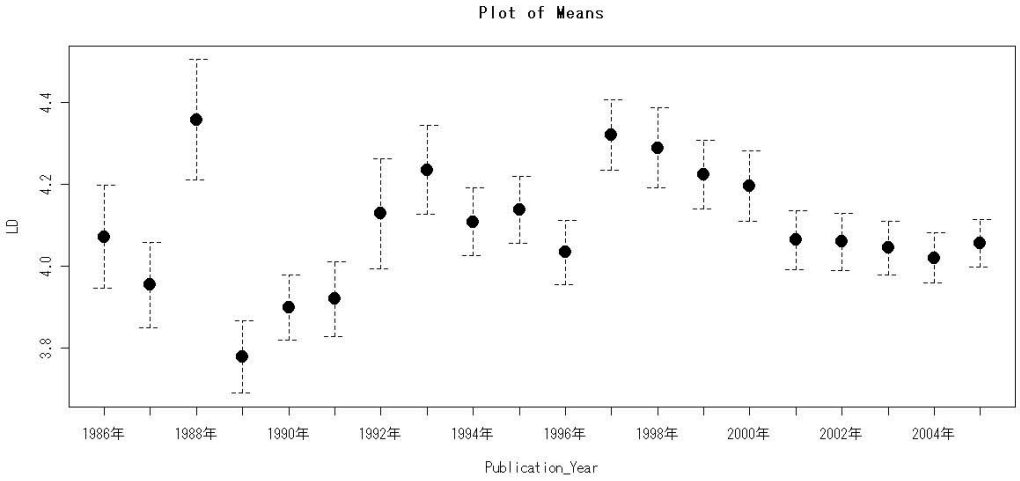

図 4.4: 語彙密度平均:出版年別

これらの結果から，Field 要因としては，出版年よりもむしろテクストのジャンルによって，語彙密度の分布，及び，平均値に差があることがわかった。語彙密度は，9. 文学や 7. 芸術. 美術など，創作的なテクストでは低くなる傾向があるが，6. 産業，3. 社会科学や 4. 自然科学などの内容を扱うテクストでは高くなる傾向がある。このことは，創作散文 (imaginative prose) と情報散文 (informative prose) で語彙密度に違いがあることを示唆するものでないかと考える。

一方，出版年ごとの語彙密度の計測結果からは，語彙密度平均値の通時的変化はほとんど見られなかった。1986 年から 2005 年では平均値の差は小さい。

4.4 Tenor 情報から見た語彙密度

4.4.1 書き手 (1):著者性別

分析データは男性著者のサンプルが占める割合が多いが，女性著者が書いたサンプルも 840 サンプル収録されている。性別という観点からは，データに偏りが見られる。

表 4.6: 著者性別サンプル数

性別	サンプル数	割合
女	840	8.7%
男	3,603	37.2%
データなし	5,252	54.2%
合計	9,695	100.0%

著者の性別に語彙密度の分布，及び，平均値を計測した結果を，それぞれ図 4.5，図 4.6 に示す。図 4.5 によると，男性の方が女性に比べ語彙密度が高くなる傾向があるものの，語彙密度の分布のばらつきも広い。サンプル数の違いがばらつきの違いとして表れている可能性がある。図 4.6 に示した平均値を見ると，男性著者の平均値は 4.7，女性著者の平均値は 3.9 であり，約 0.8 の差がある。

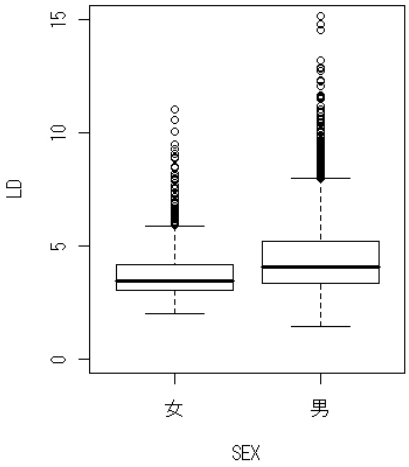

図 4.5: 語彙密度:性別

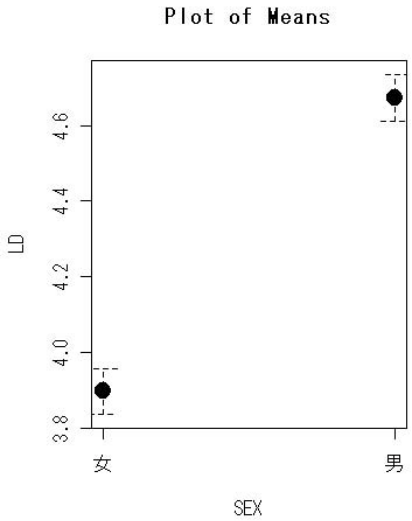

図 4.6: 語彙密度平均:性別

4.4.2　書き手 (2):著者生年

分析データに含まれる著者生年別サンプル数を表 4.7 に示す。1870 年代や 1880 年代に生まれた著者のデータがあるものの，サンプル数が極端に少ない。そこでここでは，100 サンプル以上数が確保されている，1910 年〜1970 年代までのサンプルを利用して，生年ごとの語彙密度を計測した。

表 4.7: 著者生年別サンプル数

生年	サンプル数	割合
1870 年代	1	0.0%
1880 年代	8	0.1%
1890 年代	35	0.4%
1900 年代	48	0.5%
1910 年代	101	1.0%
1920 年代	502	5.2%
1930 年代	1,006	10.4%
1940 年代	1,139	11.7%
1950 年代	934	9.6%
1960 年代	530	5.5%
1970 年代	108	1.1%
1980 年代	3	0.0%
データなし	5,280	54.5%
合計	9,695	100.0%

著者生年ごとの語彙密度の分布を図 4.7 に，平均値を図 4.8 に示す。

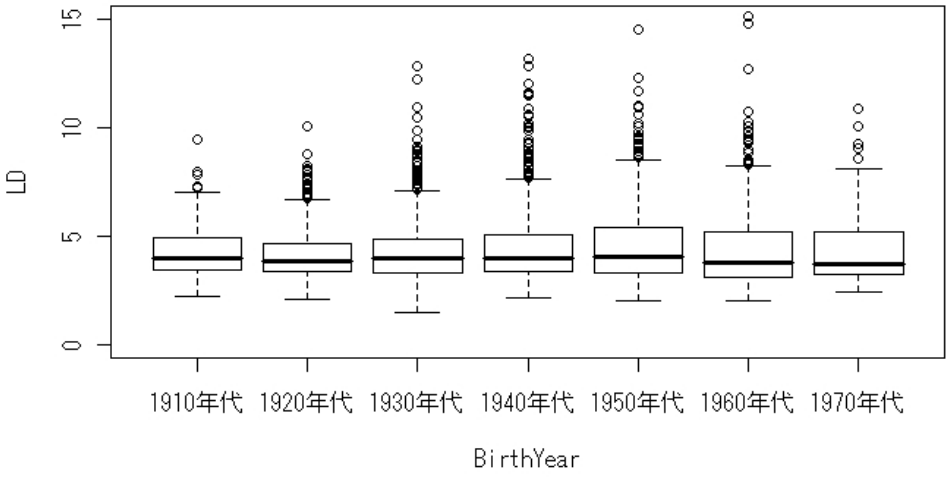

図 4.7: 語彙密度:生年別

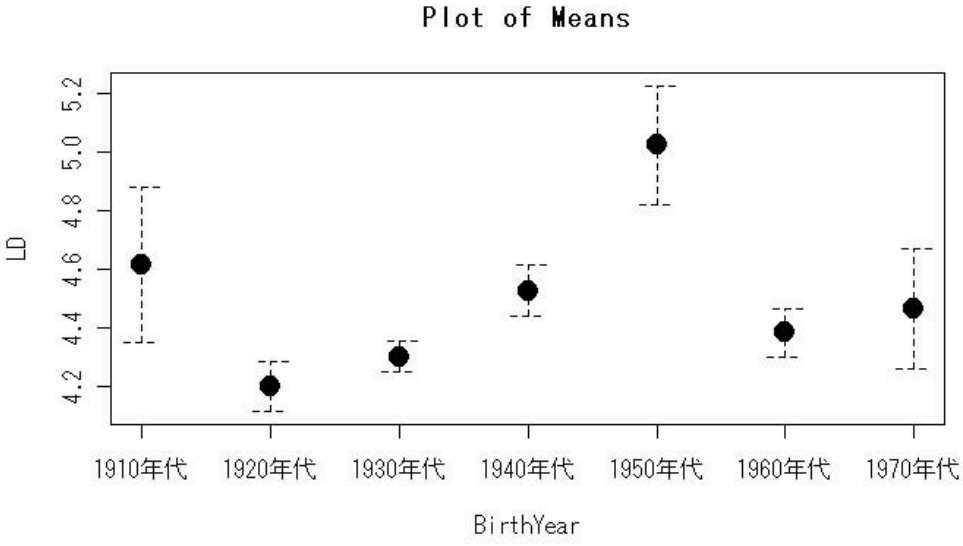

図 4.8: 語彙密度平均:生年別

1950 年代の著者の語彙密度平均が約 5.0 と高いが，その他の年代に関しては，平均値にあまり差がない。平均値が最も低い 1920 年代と最も高い 1950 年代の差は約 0.8 であり，著者生年による語彙密度の平均値への影響はあまり見られない。

4.4.3　読み手:販売対象

C コード (1 桁目) によって表される「販売対象」の情報を用いると，分析データのサンプルは表 4.8 のように分類できる。C コードでは，販売対象が，一般 (0)，教養 (1)，実用 (2)，専門 (3)，婦人 (5)，学参 I（小中）(6)，学参 II（高校）(7)，児童 (8)，雑誌扱い (9) に分類されている。

表 4.8: 販売対象別サンプル数

販売対象	サンプル数	割合
一般	7,504	77.4%
学参 I（小中）	6	0.1%
学参 II（高校）	1	0.0%
教養	525	5.4%
雑誌扱い	61	0.6%
児童	187	1.9%
実用	310	3.2%
専門	817	8.4%
婦人	12	0.1%
データなし	272	2.8%
合計	9,695	100.0%

販売対象によってサンプル数が大きく異なり,「学参 I」や「学参 II」などはサンプル数が極めて少ない。一方「一般」は 77.4%あり，分析データの大部分を占めていることがわかる。そこで，先ほどと同様に，サンプル数が 100 以上のものに限り，語彙密度を計測した。販売対象別に語彙密度を計測した結果を，図 4.9，及び，図 4.10 に示す。

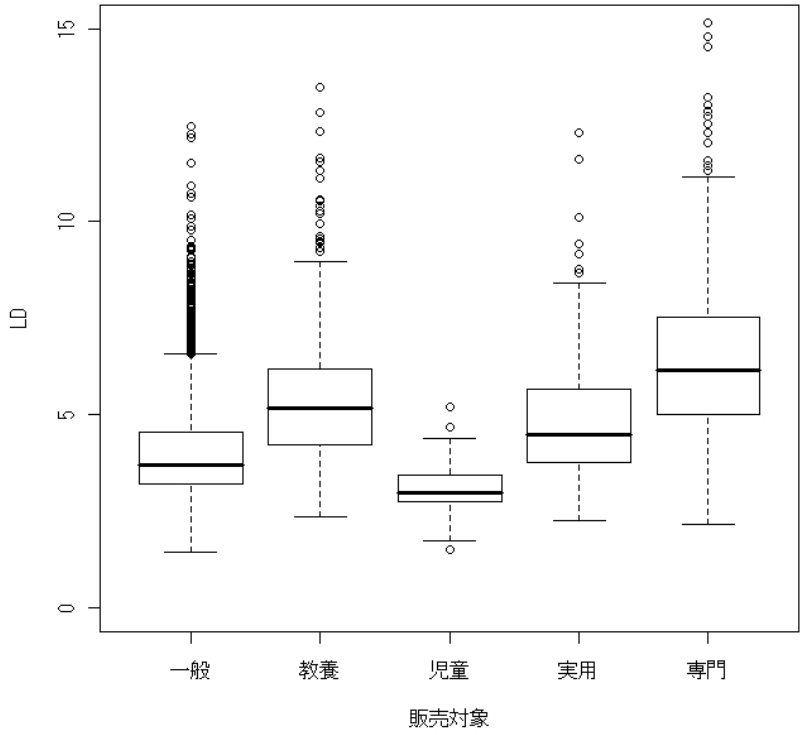

図 4.9: 語彙密度:販売対象別

図 4.9 からも，図 4.10 からも，カテゴリ間の語彙密度の差が顕著であることがわかる。語彙密度の平均値が最も高いカテゴリは「専門」であり，平均値は 7.3 である。この値は，Field, Tenor，Mode の他分類と比べても，最も高い平均値である。これに,「教養」,「実用」が続く。一方，語彙密度の平均値が最も低かったのは「児童」であり，語彙密度平均値は 3.1 であった。次に語彙密度平均が低かったのは「一般」で，平均値は 4.0 である。

したがって，販売対象のカテゴリと語彙密度には表 4.9 に示したような関係があると考えられる。販売対象が「児童」や「一般」などであれば語彙密度は低く，情報の詰め込みはあまり行われない。一方，テクストが「専門」であれば語彙密度は高く，情報の詰め込みが多く行われる傾向があると考えられる。

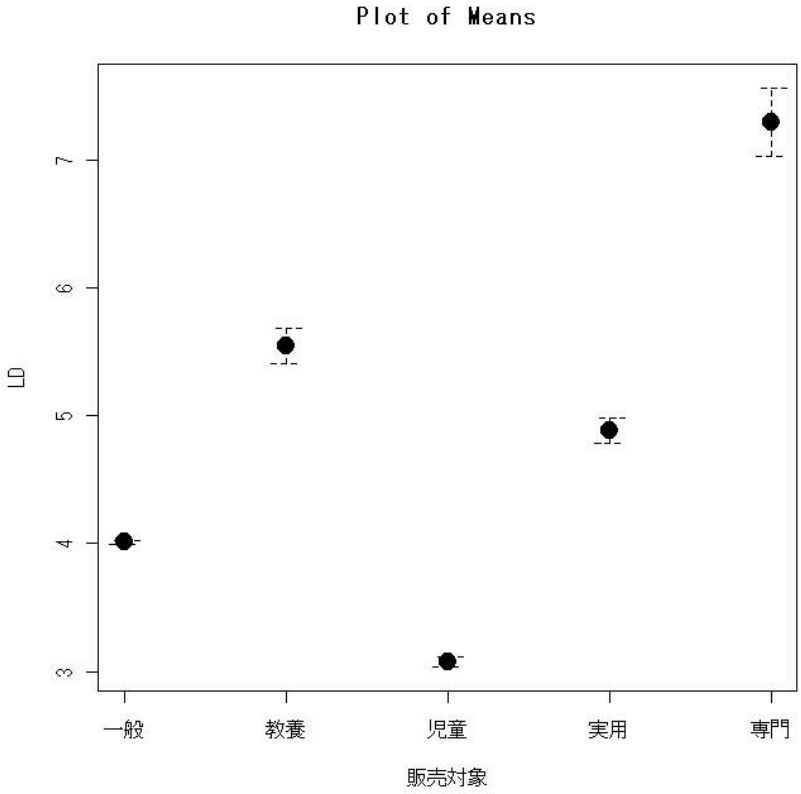

図 4.10: 語彙密度平均:販売対象別

表 4.9: 語彙密度の傾向:販売対象

販売対象	語彙密度 平均値	
専門	7.3	↑情報の詰め込み度が高い
教養	5.5	(語彙密度高)
実用	4.9	
一般	4.0	(語彙密度低)
児童	3.1	↓情報の詰め込み度が低い

以上，Tenor の観点から，著者性別，著者生年，及び，販売対象のカテゴリ別に語彙密度の分布と平均値を分析した。結果，語彙密度の平均値は著者性別，著者生年のカテゴリ間ではあまり変わらないことがわかった。著者生年で，1950 年代に生まれた著者から語彙密度の平均値が徐々に減っていることは，1997 年以降に出版された書籍の語彙密度が下降傾向にあることと関連があるのかもしれない。しかし，いずれにせよカテゴリ間で平均値に大きな差は見られない。

一方，販売対象では，カテゴリ間における語彙密度の平均値の差が大きい。販売対象カテゴ

リ間のテクストの性質が語彙密度に影響していると考えられる。販売対象が「児童」や「一般」の場合，語彙密度は低くなる傾向がある一方,「専門」の場合，語彙密度が高くなる。この傾向から，テクストが想定する読み手のレベル，もしくは，テクストの専門性と語彙密度には関係があると推測できる。

4.5 Mode 情報から見た語彙密度

4.5.1 形態

C コードの形態 (2 桁目) の情報から，分析データのサンプルは表 4.10 のように分類することができる。C コードは書籍の形態を「コミック」,「ムック・その他」,「絵本」,「事・辞典」,「新書」,「図鑑」,「全集・双書」,「単行本」,「文庫」に分類する。販売対象の場合と同様，サンプル数が 100 以上ある分類に関してのみ，語彙密度の分布，及び，平均値について調べた。結果を図 4.11，及び，図 4.12 に示す。

表 4.10: 形態別サンプル数

形態	サンプル数	割合
コミック	3	0.0%
ムック・その他	59	0.6%
絵本	4	0.0%
事・辞典	27	0.3%
新書	800	8.3%
図鑑	8	0.1%
全集・双書	826	8.5%
単行本	5,462	56.3%
文庫	2,234	23.0%
データなし	272	2.8%
合計	9,695	100.0%

語彙密度の平均値が最も高いのは「全集・双書」(平均値 5.1) である。一方，平均値が最も低いのは,「文庫」(平均値 3.5) である。NDC や販売対象ほどの差は見られないが，表 4.11 に示すとおり,「全集・双書」と「文庫」の平均値には 1.6 の差がある。このことから，形態も語彙密度に影響する要因の一つとして考えられる。

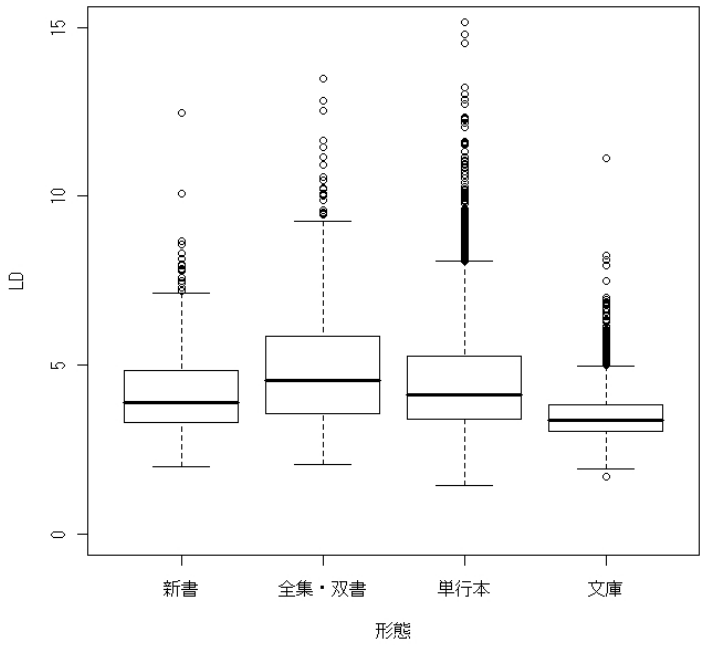

図 4.11: 語彙密度:形態別

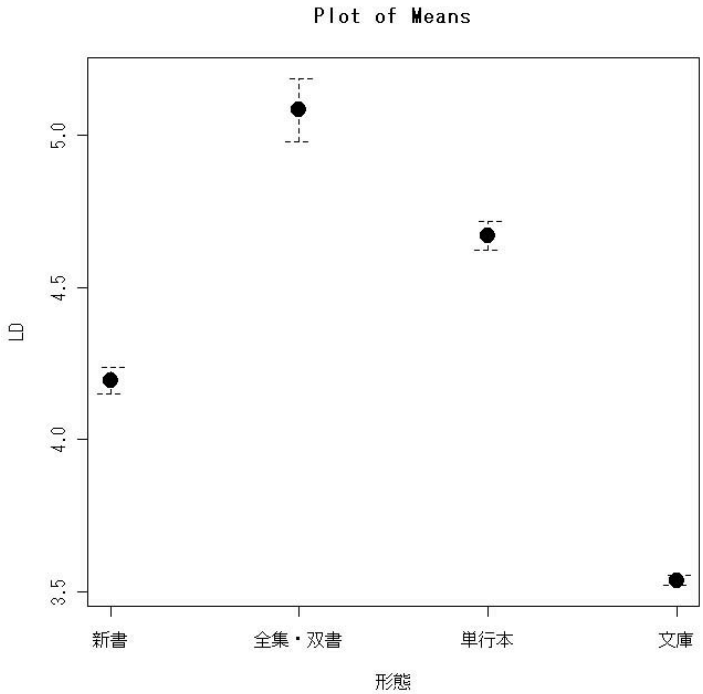

図 4.12: 語彙密度平均値:形態別

表 4.11: 形態に基づく語彙密度の傾向

形態	語彙密度 平均値	
全集・双書	5.1	↑情報の詰め込み度が高い
単行本	4.7	(語彙密度高)
新書	4.2	(語彙密度低)
文庫	3.5	↓情報の詰め込み度が低い

4.5.2 テクスト中のサンプルの位置

サンプルの抽出開始頁，終了頁，及び，サンプルが取得された書籍の総頁数から，サンプルのテクスト中の位置を特定し，語彙密度とテクスト中のサンプルの位置関係を調べた。サンプルによっては，総頁中の全頁など，広範囲が取得されているものなどがあったため，表 4.12 にある 10 の範囲に該当するサンプルのみを対象とした。0%が 1 頁目にあたり，100%が最終頁である。

表 4.12: テクスト中の位置ごとのサンプル数

開始位置〜終了位置	サンプル数	割合
0%〜10%台	390	4.0%
10%〜20%台	319	3.3%
20%〜30%台	332	3.4%
30%〜40%台	347	3.6%
40%〜50%台	355	3.7%
50%〜60%台	309	3.2%
60%〜70%台	350	3.6%
70%〜80%台	342	3.5%
80%〜90%台	328	3.4%
90%〜100%台	145	1.5%
合計	3,217	33.2%

テクスト中のサンプルの位置ごとに見た，語彙密度の分布，及び，平均値を図 4.13，図 4.14 に示す。図 4.13 が示すように，テクスト中の位置の違いによって，語彙密度の分布には大きな差は認められない。しかしながら，平均値の差は小さいが，図 4.14 を見ると，語彙密度は，テクストの始めと終わりに高くなる傾向があり，テクストの中心部分に向けて低くなる傾向があるようである。

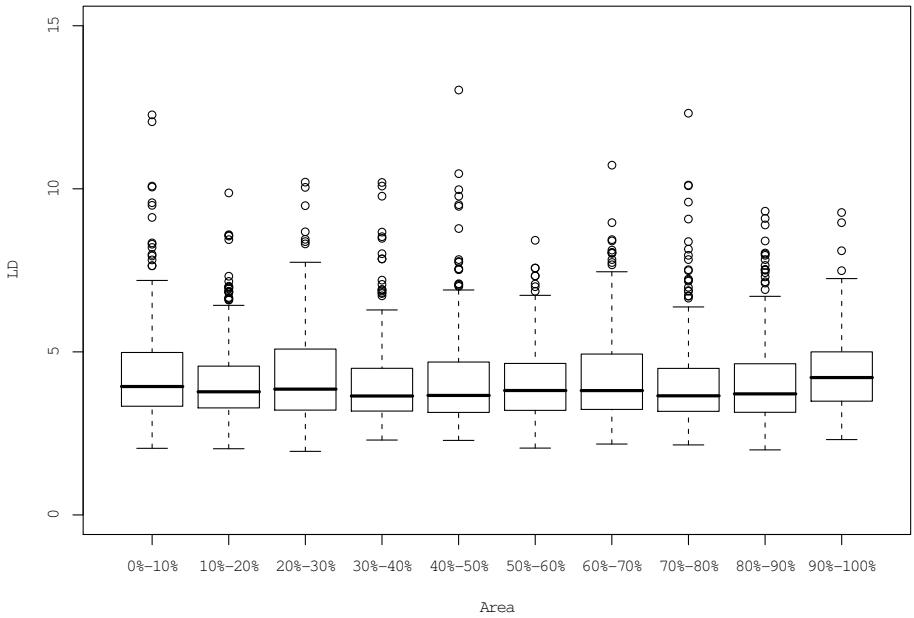

図 4.13: 語彙密度:テクスト中の位置別

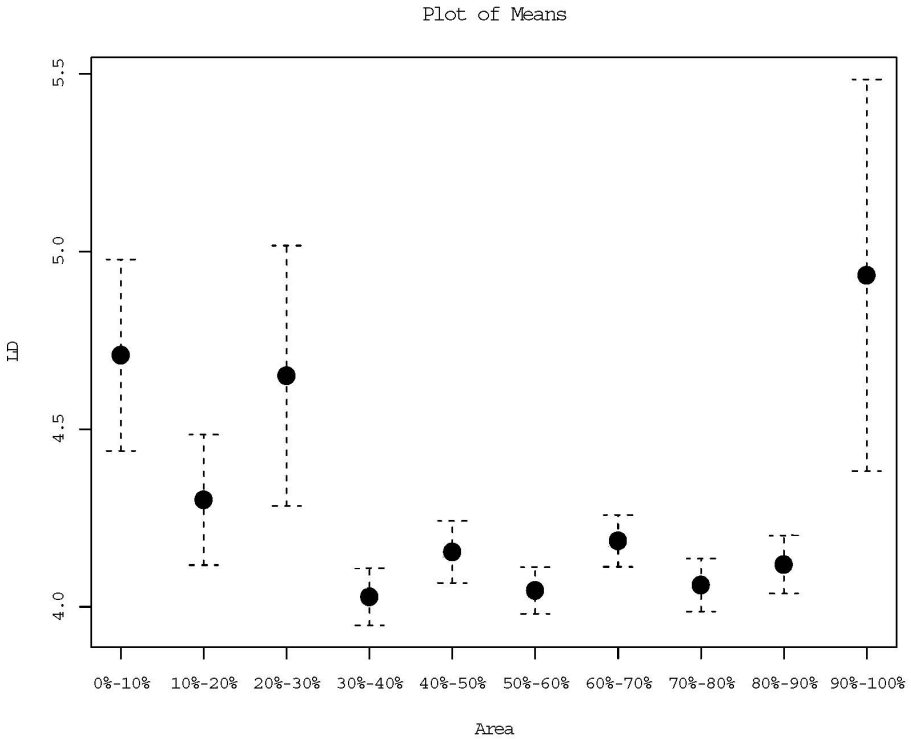

図 4.14: 語彙密度平均値:テクスト中の位置別

表 4.13 に示すとおり，語彙密度の平均値が高いのは「90%〜100%」「0%〜10%」である。逆に低いのは「30%〜40%」に位置するサンプルであり，これに「50%〜60%」が続く。

表 4.13: 語彙密度の傾向:テクスト中の位置

テクスト中の位置	語彙密度 平均値	
90%〜100%台	4.93	↑情報の詰め込み度が高い
0%〜10%台	4.71	(語彙密度高)
20%〜30%台	4.65	
10%〜20%台	4.30	
60%〜70%台	4.19	
40%〜50%台	4.15	
80%〜90%台	4.12	
70%〜80%台	4.06	
50%〜60%台	4.05	(語彙密度低)
30%〜40%台	4.03	↓情報の詰め込み度が低い

以上，Mode に関して，形態カテゴリ，及び，テクスト中のサンプルの位置別に，語彙密度の分布傾向と平均値について調べた。分析結果は，サンプルの位置の違いによっては，語彙密度の平均値に大きな差は見られないことを示した。しかし，形態に関しては，カテゴリ間で差があり,「文庫」として発行されるテクストは語彙密度が低く,「全集・双書」として発行されるテクストは語彙密度が高い傾向があることがわかった。

4.6 語彙密度とコンテクスト要因との関係

本章では，BCCWJ2008 の語彙密度計測の結果から，語彙密度とコンテクスト要因との関係について分析を行ってきた。Field については，ジャンルと出版年，Tenor については，著者性別，著者生年，販売対象，Mode については，形態とテクスト中のサンプルの位置との関係を調べた。表 4.14 に，各コンテクスト要因ごとに，語彙密度の平均値が最も高かったカテゴリと最も低かったカテゴリ，及び，その平均値，平均値の差を示す。

語彙密度の平均値に最も大きな差があったのは「販売対象」のカテゴリ間であり,「児童」で最も平均値が低く,「専門」で最も高い。これにジャンルが続き,「9. 文学」で平均値が最も低く，一方「6. 産業」で最も高い。このことから，日本語書き言葉において語彙密度に大きな差が認められるコンテクスト要因は，Field のジャンル，及び，Tenor の販売対象であると考えられる。

以上を踏まえると，語彙密度は，創作的なジャンルに属するテクストや児童向け・非専門的なテクストで低く，一方 (社会) 科学的なジャンルに属するテクストや専門的読者を対象とするテクストでは高くなる傾向があると考えられる。販売対象と NDC カテゴリを語彙密度の観点から位置づけると，図 4.15 のようになる[2]。「⇔」上の数値は語彙密度である。

[2] 販売対象のカテゴリは**太字**。

表 4.14: 分類別語彙密度平均値とコンテクスト

状況コンテクスト	分類	最高値のカテゴリ	最高値	最低値のカテゴリ	最低値	差
Field	ジャンル	6. 産業	5.4	9. 文学	3.5	2.0
	出版年	1988 年	4.4	1989 年	3.8	0.6
Tenor	著者性別	男	4.7	女	3.9	0.8
	著者生年	1950 年代	5.0	1920 年代	4.2	0.8
	販売対象	専門	7.3	児童	3.1	4.2
Mode	形態	全集・双書	5.1	文庫	3.5	1.6
	テクスト中の位置	90%〜100%台	4.9	30%〜40%台	4.0	0.9

次章では，この傾向を踏まえて，語彙密度がどのようにサンプルの特徴を表す指標として利用できるのか，その可能性について検討した結果について述べ，本報告書をまとめる。

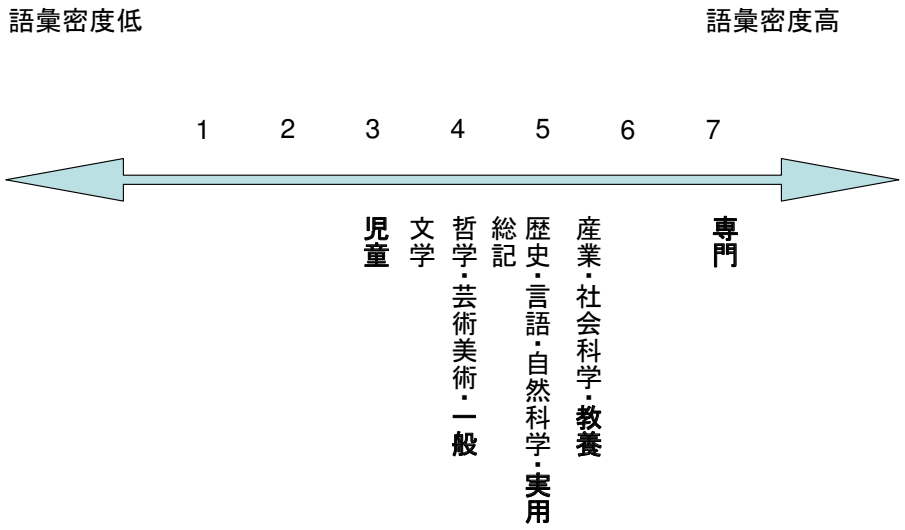

図 4.15: 語彙密度から見た書籍サンプル販売対象，及び，NDC 別カテゴリの位置づけ

第5章　テクスト分類における語彙密度の可能性

佐野大樹・丸山岳彦

第 4 章では，BCCWJ2008 を用いた分析を行い，語彙密度は，専門的で (社会) 科学的なテクストでは高く，一方，児童向け・非専門的で創作的なテクストでは低くなる傾向があることを明らかにした。

本章では，この傾向を踏まえて，語彙密度が BCCWJ に含まれるサンプルを分類する，もしくは，サンプルの特徴を把握するための情報として，現行のドキュメントアノテーションと補完的な形で，どのように利用できるのか，その可能性について考察する。

5.1　クラインとして表されるテクスト情報の必要性

BCCWJ に付与されている，もしくは付与が検討されているドキュメントアノテーションは，全てタイポロジカルな観点にたった分類である。例えば，C コードの「販売対象」であれば,「児童」,「一般」,「専門」などのカテゴリが一つ付与される。タイポロジカルな分類は区分が明確であるため，カテゴリ間の比較の際には有効な手段となる。一方，カテゴリとカテゴリの境界に位置するようなテクストの相違の比較や，同一カテゴリ内に属するテクストの特徴の異なりを検討する際には，その性質上，利用することができないという短所がある。

ここで問題となるのが，BCCWJ は現代日本語書き言葉の縮図となるコーパスを作成するため，層別ランダムサンプリングを用いており，この結果，抽出比の高い層に属するサンプルの割合がコンテクスト要因のカテゴリ別割合にも反映されることである[1]。例えば，第 4 章で示したとおり，BCCWJ2008 に含まれるサンプルの 77.4%は，C コードの販売対象で「一般」と分類されるものである。このことは，77.4%のサンプルの販売対象別の違いは C コードでは判別できないことを意味する。同様に,「NDC」,「著者性別」,「形態」などの分類で，データの割合に偏りがある。

BCCWJ を様々な研究で活用できる環境を整備するためには，現行のドキュメントアノテーションに加え，同一カテゴリ内に属するテクストの特徴の異なりを把握できる，補完的な情報が必要となる。特に，他のドキュメントアノテーションと同様に，テクストの特徴をタイポロジカルに捉えるものでなく，同一カテゴリ内の異なりを漸次的連続体 (cline) として把握できる情報が必要になると考える。

[1] コーパスの構成については，[21, 22] を参照。

5.2　語彙密度によるテクストの特徴把握

本報告書で述べた分析結果は，語彙密度がそのようなクラインとして表される情報を提供する一指標として利用できることを示唆するものである。語彙密度は連続尺度であり，テクストのコンテクスト情報ではなく，テクストで表現されている言語情報を用いて計測が可能である。

このため，異なるカテゴリに属するテクスト間の相違だけでなく，同一カテゴリに属するテクスト間の相違も，第 4 章図 4.15 に示したような語彙密度が低いものと高いものとを両極とするクラインの中に位置づけることができる。語彙密度のみで全てのテクストの特徴を把握できるわけではないが，語彙密度と他の指標を組み合わせることで，現行のドキュメントアノテーションでは把握することのできないテクストの特徴を捉えることができるのではないかと考える。

例えば，第 1 章で示した以下の同ジャンル，同タイトルをもつ二つの白書サンプルの違いは，語彙密度を用いることで捉えることができる。『国民生活白書 平成 6 年版』の語彙密度が 3.9 であるのに対し,『国民生活白書 平成元年版』の語彙密度は 12.7 で，8.7 の差がある。

語彙密度 3.9

四季折々の風景，農村の行事，跳ね回る子供たちや動物たち…。その素朴な画風を愛され，アメリカの国民的画家となった「グランマ・モーゼス（モーゼスおばあさん)」ことアンナ・メアリ・ロバートソン・モーゼスが画筆を握ったのは７０を過ぎてからのこと。１２歳から農場で働き，結婚して１０人の子供をもうけ，農場の主婦としての切り盛りから手が離れてからであった。画家としての訓練はもちろんのこと，学校にすらたまにしか行けなかった彼女が孫娘のために刺で絵を作ってあげたのがきっかけとなった。高齢期を迎える準備は不可欠であり，当然，趣味も必要である。しかし，それで身を立てるつもりでないなら身構えて無理に作る必要はなく，仕事と子育てに明け暮れる日々の中で「時間ができたらやってみたい」と思ってきたことに取り組んでみようとするくらいの感覚で良いのではないだろうか。例えば，現役を退いた後，北上する桜前線を追いかけている夫婦がいる。桜の微妙な表情を墨絵で描き，エッセーや句にする。「ゆっくり」を合言葉にハンドルを握り,「若い日本人は旅の心をどこかに置き忘れてしまったのではないか」と問いかける。

『国民生活白書 平成 6 年版』 経済企画庁 OW4X_00397

語彙密度 12.7
省資源・省エネルギーを着実に推進するため，省エネルギー・省資源対策推進会議において，６３年６月２８日夏季の省エネルギー対策について，また，同年１１月２２日冬季の省エネルギー対策について決定し，関係各省庁は決定事項の周知徹底を行った。また，毎月１日の「省エネルギーの日」，１２月１日の「省エネルギー総点検の日」，２月の「省エネルギー月間」等の機会を利用して，関係各省庁，地方公共団体，省エネルギー推進関係団体等はパンフレットの配布，ポスター等による広報，集会，展示会，講習会，表彰，作文募集等，各種行事を実施した。総理府においては，テレビ等の媒体を利用した政府広報により，省エネルギーの普及広報活動を行った。警察庁は,（財）全日本交通安全協会に対し，経済運転の広報を行うよう要請した。経済企画庁は，省エネルギー啓発パンフレット及びポスターを作成し，省資源国民運動参加団体等に配布するとともに，省エネルギー月間（２月）に当庁において，懸垂幕の掲示を行った。通商産業省は,（財）省エネルギーセンターを通じて，ポスター，パンフレット等の配布などにより普及広報活動を行うとともに，全国数か所で省エネルギー展，省エネルギー講演会を開催し，また，エネルギー管理優良工場等の表彰を行った。運輸省は，運輸部門におけるエネルギー政策に関するパンフレットの作成・配布等により，普及広報活動を行うとともに，運輸部門におけるエネルギー政策に関する講演会を行った...

『国民生活白書 平成元年版』 経済企画庁 OW3X_00294

語彙密度は，テクストの硬軟，読み手のレベル (Audience level) の推測,「書き言葉らしさ・話し言葉らしさ」を把握する際の一指標として利用できるのではないかと考えている。以下，それぞれの特徴を把握する指標としての語彙密度の可能性について述べる。

5.3 テクストの硬軟の推測

語用論・文体論にとって重要なテクスト情報の一つに，第 1 章で述べた，柏野他 [16] であげられているテクストの「硬軟」がある。テクストの「硬軟」の程度を表す指標が確立できれば，例えば，同義語が「堅い」テクストと「くだけている」テクストでどのように使用傾向が違うのかなどについて検討する際，有効な情報となる。

語彙密度が，専門的で (社会) 科学的なテクストで高く，一方，非専門的で創作的なテクストでは低くなる傾向があるのであれば，テクストの「硬軟」を捉える一指標として利用できるのではないかと考える。例えば，NDC9. 文学のサンプルのうち，語彙密度が高いものと低いものを比べると，高いものは「堅い」テクストであることが多く，一方，低いものは「くだけている」テクストである場合が多い。以下に，一例として，9. 文学のうち，語彙密度が最も高いサンプルと低いサンプルの冒頭部分を示す。

NDC9. 文学 語彙密度:高
二十四孝の成立―全相二十四孝詩選と日記故事―世に二十四孝と称するものがある。古来，二十四箇の孝子譚を総称して，そのように呼んでいるのだが，中国においては，瞿中溶が，と言うように，例えば日記故事の巻頭に置かれた「二十四孝」を通じ，享受されたものらしい。ところが，先の孝子伝と同じく，この二十四孝というものも，見掛けに反し，非常に捉えにくい概念なのであって，文学史的にその具体像を思い描くことは容易でない。その流れから考えて，二十四孝は，孝子伝を母胎として生み出されたものと思われ，二十四孝は，西野貞治氏の,〔孝子伝〕が盛行したことは種々の資料から偲ばれるが，伝存の記録は南宋の鄭樵の通志略を下限とし，稍後の晁公武・陳振孫らの博捜家にも見られていぬようで，或は南宋の兵燹に失われたものが多いかと考えるとの指摘通り 1，宋代を境として姿を消すに至った孝子伝に，取って替わったものと捉えることが出来よう。その二十四孝については，徳田進氏の『孝子説話集の研究―二十四孝を中心に―』が委細を尽くすが 2，ここで再度，近時の説をも参照しながら，主としてテキスト面における概説を試みておきたい。

『孝子伝の研究』黒田彰 PB19_00255

NDC9. 文学 語彙密度:低
いなかのおばあちゃんが，リンゴをたくさんおくってくれました。そのはこのなかに，おばあちゃんがつくったおにんぎょうがはいっていました。おてがみもいっしょにはいっていました。このおにんぎょうは，リンゴちゃんといいます。まちにいけるのでおおよろこびです。マイちゃん，あそんであげてね。マイは，リンゴちゃんをだきあげました。リンゴちゃんは，まんまるのあかいかおで，よそみして，しらんぷりしています。マイは，ぷーっと，おこりたくなりました。「あかいかお…，へんなかお。」マイも，しらんぷりしようとおもいました。

『リンゴちゃん』角野栄子，長崎訓子 PB39_00687

『孝子伝の研究』は文学史的内容を扱うものであり,「堅い」テクストと判断できるだろう。一方『りんごちゃん』は絵本であり，内容から見ても，利用されている語彙・文法的にも「くだけている」テクストと考えられる。

このように，語彙密度を一指標として，9. 文学におけるテクストの特徴の違いを把握することができる。語種の割合など，他の指標と組み合わせれば，テクストの硬軟を表す指標として利用できるのではないかと考える。

5.4　読み手レベル (Audience level) の推測

BNC Web Indexer[12] などでは提供されている読み手のレベルに関する情報が，現行では BCCWJ には付与されていない。BNC Web Indexer では対象となる読み手のレベルを 1) High 2) Medium 3)Low の三つに分類している。C コードの販売対象がこれに類似する情報として

利用可能であるが，先述したように，BCCWJ に含まれるサンプルの大部分は「一般」と分類されており,「販売対象」の情報だけではテクスト間の違いは捉えられない。

語彙密度が (非) 専門性と関連があるのであれば，読み手レベルを推測するための一指標として利用できると考えられる。例として，販売対象が「一般」に属するサンプルのうち，語彙密度が最も高いものと低いものを示す。

販売対象「一般」 語彙密度:高

モータリゼーションが交通行動に与えた変化

モータリゼーションが意味する自動車依存の社会と人々の生活様式の変化は，個人レベルの交通行動に大きな変化をもたらしている．もちろん，これは都市の規模や地域特性にも関係するため，同様の規模の地域を抽出して比較することによって，モータリゼーションが人々の交通行動にもたらした変化を考える．

１就業者の交通行動は変化したか

モータリゼーションにより通勤圏が拡大し，とくに郊外部への移動が増加していることはすでに示した．この就業者が行っている郊外中心の移動はどのようなものであるのか，代表的な交通行動とモータリゼーションの関係を考えてみる．具体的には，１９９０／９１（平成２／３）年の大阪市内（東住吉区）と名古屋市内（守山区）を取り上げて，就業者の交通行動を検討してみよう．これらの地域は，いずれも都心からの距離が７ｋｍ程度の住宅地域である．人々の交通行動は，自宅や勤務先などを活動拠点（ベイス）として，この場所を中心としたトリップの連鎖として表現できる．このような１日の移動をまとめて「トリップ・パターン」あるいは「トリップ・チェイン」と呼ぶ．代表的なパターンが表２・４に示されている．ここで「総チェイン数」は，トリップ連鎖の総数であるから，対象とした交通行動者の数と一致している．

『ポスト・モータリゼーション』 北村隆一 PB16_00034

販売対象「一般」 語彙密度:低

いやなことはやらなくていい自分の思っていることに反する行為をすると，なにかしら抵抗があるでしょう。悔やむこともあるかもしれません。そういうことは，やらないほうがよいのです。やりたいのか，やりたくないのかとまず考えて，やりたい気持ちが多かったらやりましょう。わたしは，やりたくないことは，やらないようにしています。昔は，やりたくないこともやむをえずやりましたが，ろくなことがありませんでした。だから，やらないことにしたのです。それでいいのではないかと思っています。ですから，半々だったらやめたほうがよいと思います。

『上手に生きるルールとコツ』 船井幸雄 LBs1_00002

前者では，モータリゼーションを中心に話題が展開し，テクストを解釈するためには専門的な知識が求められると考えられる。BNC Web Indexer の Audience Level では，High と分類されるものだろうと考えられる。一方，後者は，前者に比べ日常的な言葉が多く使われており，読者のレベルは Low か Medium に属すると考えられる。理解度が低い語は語彙密度が高いテ

クストで利用される傾向があり，一方，理解度が高い語は語彙密度が低いテクストで利用される傾向があることは佐野 [18] でも確認されている。今後，リーダビリティ班の柴崎 [19] で利用されている指標などを参考に，読み手レベルの推測のための指標として語彙密度を利用することが可能ではないかと考える。

5.5　書き言葉らしさ・話し言葉らしさの推測

BCCWJ は書き言葉を収録したコーパスであるが，話し言葉らしい特徴をもつサンプルもあれば，書き言葉らしい特徴をもつサンプルもある。英語においては，情報が音で伝わり，再読が一般的にはできない話し言葉では，情報の詰め込み行為は避けられるため，語彙密度は低くなる傾向が強いと考えられている。一方，書き言葉は再読が可能であり，また，読み手のペースでテクストを読み解くことができるため，語彙密度が高くなる傾向があることが確認されている [5]。

この傾向が日本語でも見られるのであれば，書き言葉における「書き言葉らしさ・話し言葉らしさ」を表す指標として語彙密度を利用することが可能であると考えられる。

語彙密度を用いてテクストの「書き言葉らしさ・話し言葉らしさ」を推測できるかをパイロット的に検討するため，分析データ全てを対象とし，語彙密度の下位 20 位までに属するサンプルが会話文を含むかどうかを人手で確認した結果を表 5.1 に示す。

表 5.1: 語彙密度と話し言葉らしさ

Sample_ID	語彙密度	会話文の有無	形式
LBs1_00002	1.5	×	—
PB39_00687	1.5	○	鍵括弧
LBf9_00041	1.7	○	鍵括弧
LBr2_00019	1.7	○	鍵括弧
PB43_00289	1.7	○	鍵括弧
LBr9_00259	1.8	○	改行
LBb9_00044	1.9	○	鍵括弧
PB39_00297	1.9	○	鍵括弧
LBh9_00116	1.9	○	鍵括弧
LBq7_00006	2.0	×	—
LBr9_00158	2.0	○	鍵括弧
LBe9_00183	2.0	○	鍵括弧
PB39_00070	2.0	○	鍵括弧
PB39_00421	2.0	×	—
PB29_00524	2.0	○	鍵括弧
PB17_00195	2.0	○	鍵括弧
LBs0_00006	2.0	○	改行
PB49_00158	2.0	○	改行
LBhn_00032	2.1	○	改行
LBh9_00139	2.1	○	鍵括弧

表 5.1 に示したとおり，20 サンプル中 17 サンプルに会話文が含まれており，この結果は，語彙密度が「書き言葉らしさ・話し言葉らしさ」の程度を表す一指標として利用できる可能性を示すものである。以下，表 5.1 の中で最も語彙密度の高い LBh9_00139 の会話文部分の一部を示す。

会話文を含むサンプル 語彙密度:低

何日かまえから痛んでた右の奥歯が，夕食のあとに耐えられないくらいの痛さになっちゃったの。鎮痛剤を飲んでも，ぜんぜん効かないほど。だけど。「歯医者さんだけは，いやああ !! 大っ嫌いっっ !! 」「だめよ。診てもらわなくっちゃ」あたふたとコートを肩にかけながら，ママがいった。「真穂。早く」「やだあ !! あそこの先生，痛いから，ぜったい行かなーい !! 」「それはむかしのことでしょ。今はお兄ちゃまだって通ってるんだから，大丈夫よ。上手だっていってたから」「いやいやいやーッ」ふええ。泣きべそかいて抵抗するわたしを，ママが抱くようにして，近所の歯医者さんに連れていったの。「すみません，夜分に。早瀬でございますが。急に痛みだして耐えられないようすなんで。診ていただけませんでしょうか」ここの歯医者さん。

『好きから始まる kiss 物語』青山えりか LBh9_00139

5.6　まとめ

以上，本報告書ではシステミック理論における「語彙密度」という概念を，日本語大規模コーパスに適用し，テクスト分類に役だてる方法について述べてきた。BCCWJ2008 の書籍可変長サンプルを利用して，日本語書き言葉における語彙密度の計測方法を提案し，語彙密度が現行のドキュメントアノテーションとどのような関係にあるのかを調べた。また，語彙密度を指標として用いることで BCCWJ に含まれるサンプルに対して，どのような観点からのテクスト分類ができるのか，その可能性について示した。

分析の結果，語彙密度は，非専門的・創作的なテクストでは低く，専門的・（社会）科学的な内容を扱うテクストで高くなることを確認することができた。この結果から，テクストの硬軟の程度，読み手レベルの推測，及び，書き言葉らしさ・話し言葉らしさの程度の把握に，語彙密度の計測を利用できる可能性があることを述べた。

BCCWJ は，現代日本語の書き言葉の総体に対する縮図として，代表性を備えた均衡コーパスとして構築されている。多種多様なサンプルが統一的な手法により収集されており，Web コーパスや新聞コーパスなどには見られない，多様なサンプルが含まれている点が特徴的である。この多様性という特徴を十全に活用して研究を進めるためには，タイポロジカルな分類だけでなく，クラインとしてサンプルの特徴を位置づけることができるドキュメントアノテーションが必須であると考える。この一つの可能性として，本稿で示した語彙密度の計測がある。

今後，多様性を活用できる環境を整備するために，サンプルの特徴を捉えるためのテクスト分類法について，さらに検討を続けていく予定である。

参考文献

[1] Burnard, L. and Aston, G. (1998) *The BNC Handbook: Exploring the British National Corpus with Sara*. Edinburgh: Edinburgh University Press.

[2] Butt, D., Fahey, R., Feez, S., and Yallop, C.(1995) *Using Functional Grammar: An Explorer's Guide*. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

[3] Clerehan, R., Buchbinder, R., and Moodie, J.(2005) A Linguistic Framework for Assessing the Quality of Written Patient Information: Its Use in Assessing Methotrexate Information for Rheumatoid Arthritis. *Health Education Research*. 20(3)334-344.

[4] Francis, W. N. and Kucera, H. (1979) *Brown Corpus Manual*. Rhode Island: Brown University.

[5] Halliday, M. A. K. (1985) *Spoken and Written Language*. Victoria: Deakin University.

[6] Halliday, M. A. K. (1990) Some Grammatical Problems in Scientific English. *Annual Review of Applied Linguistics*. 6: 13-37.

[7] Halliday, M. A. K. (2007) *An Introduction to Functional Grammar*. 2nd ed. London: Arnold.

[8] Halliday, M.A.K.(1996) On Grammar and Grammatics. *Functional Descriptions: Theory in Practice*. eds. by Ruqaiya Hasan, Carmel Cloran and David G. Butt. Amsterdam: John Benjamins.1-38.

[9] Halliday, M. A. K. and Hasan, R. (1985) *Language, Context and Text* Waurn Ponds: Deakin University Press.

[10] Halliday, M. A. K. and Matthiessen,C. (2007) *An Introduction to Functional Grammar*. 3rd ed. London: Arnold.

[11] Harrison, S. and Bakker,P. (1998) Two New Readability Predictors for the Professional Writer: Pilot Trials. *Journal of Research in Reading*. 21(2): 121-138.

[12] Lee, Y. D. (2001) Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle. *Language Learning & Technology*. 5(3): 37-72.

[13] Martin, J. R. (1993) A Contextual Theory of Language. *The Powers of Literacy: A Genre Approach to Teaching Writing*. eds. by C. Bill and M. Kalantzis. London: Falmer Press. 116-136.

[14] Sano, M., and Maruyama, T.(2008) Lexical Density in Japanese Texts: Classifying Text Samples in the Balanced Corpus of Contemporary Written Japanese (BCCWJ). eds. by C. Wu and C. Matthiessen and M.Herke. *Proceedings of 35th International Systemic Functional Congress*. Sydney: Macquarie University.

[15] 小椋秀樹・小磯花絵・冨士池優美・原裕 (2008)『『現代日本語書き言葉均衡コーパス』形態論情報規程集』国立国語研究所内部報告書 LR-CCG-07-04.

[16] 柏野和佳子，丸山岳彦，秋元祐哉，稲益佐知子，佐野大樹，田中弥生，山崎誠 (2008)『『現代日本語書き言葉均衡コーパス』における書籍サンプルの多様性』特定領域研究「日本語コーパス」平成 19 年度研究成果報告書 (JC-D07-02), 特定領域研究「日本語コーパス」データ班.

[17] 佐野大樹 (2007)「学術的表現への言い換え—教育現場での選択体系機能言語理論—」『日本語学』11 月号, 60-71.

[18] 佐野大樹 (2008)「大規模バランストコーパスにおけるテクスト分類—システミック理論の観点から—」『特定領域研究「日本語コーパス」平成 20 年度全体会議予稿集』83-90.

[19] 柴崎秀子・玉岡賀津雄・山本和英・加納満・原信一郎・李在鎬 (2008)「平成 20 年度研究進捗状況報告：リーダビリティー班日本語コーパスを応用した文章の難易測定の研究」『特定領域研究「日本語コーパス」平成 20 年度全体会議予稿集』77-82.

[20] 伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵 (2007)「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22,101-122.

[21] 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 (1) —現代日本語書き言葉の文字数調査—』 特定領域研究「日本語コーパス」平成 18 年度研究成果報告書 (JC-D-06-02), 特定領域研究「日本語コーパス」データ班.

[22] 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—』 特定領域研究「日本語コーパス」平成 18 年度研究成果報告書 (JC-D-07-01), 特定領域研究「日本語コーパス」データ班.

[23] 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』11(3), 39-68.

[24] 山口昌也・高田智和・北村雅則・間淵洋子・小林正行・西部みちる (2008)『『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.0』 特定領域研究「日本語コーパス」平成 19 年度研究成果報告書 (JC-D-07-03), 特定領域研究「日本語コーパス」データ班.

[25] 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 (2009) 「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—」『言語処理学会第 15 回年次大会発表論文集』.

謝辞 本研究は，文部科学省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) による補助を得ています。また，コーパスに収録されたテクストの著者，出版社をはじめ，国立国会図書館，東京都立中央図書館，東京都立日比谷図書館，東京都立多摩図書館，立川市中央図書館，八王子市中央図書館，横浜市立図書館，一橋大学附属図書館，自治大学図書室，(社) 日本図書館協会，(社) 日本文藝家協会，(社) 日本ペンクラブ，各位よりデータ提供等のご協力を頂いています。記して深く感謝の意を表します。

特定領域研究「日本語コーパス」データ班（サンプリング担当）

山崎 誠　（国立国語研究所グループ長（副））
柏野 和佳子　（国立国語研究所主任研究員）
丸山 岳彦 *　（国立国語研究所研究員）
佐野 大樹 *　（国立国語研究所特別奨励研究員）
秋元 祐哉　（国立国語研究所研究補佐員）
稲益 佐知子　（国立国語研究所研究補佐員）
田中 弥生　（国立国語研究所研究補佐員）
大矢内 夢子　（国立国語研究所研究補佐員）

（＊は主たる執筆者）

特定領域研究「日本語コーパス」平成 20 年度研究成果報告書

語彙密度を利用した『現代日本語書き言葉均衡コーパス』テクスト分類の試み

平成 21 年 3 月 24 日
執 筆 者　佐野大樹 丸山岳彦 山崎誠 柏野和佳子
秋元祐哉 稲益佐知子 田中弥生 大矢内夢子
発行者　文部科学省科学研究費特定領域研究「日本語コーパス」データ班
連絡先　〒 190–8561 東京都立川市緑町 10 番地の 2
独立行政法人国立国語研究所研究開発部門内
文書管理番号： JC-D-08-02