愛知淑徳大学論集―人間情報学部篇　第 8 号　2018 年 3 月，pp. 31–39

原著

テキストの時系列的特徴の定量分析：談話的まとまりの長さ

Quantitative Analysis of Time Series Feature of Text: The Length of Discourse Unit

浅　石　卓　真*

Takuma ASAISHI

要　旨

本論文では中学・高校の理科教科書を対象として，テキストを読み進めていく過程での内容の部分的まとまり（談話的まとまり）の長さを定量的に分析した。具体的にはテキストを専門用語の bag of words とみなした上で，専門用語がランダムに出現した場合と実際のテキストとの差が最大になるように，換言すれば専門用語が最も秩序を保って出現するようにテキストを分割した時の各部分の長さを計量した。その結果，談話的まとまりは教科書ごとの違いが大きいが，学年が上がるにつれ長くなる傾向が見られた。また，個々の専門用語に着目した時の談話的まとまりも教科書間の違いが大きいが，談話的まとまりが非常に長い一部の専門用語は教科書内での離れた部分同士をつなぎ，テキスト全体を一つにまとめる役割を担っている可能性が示唆された。

キーワード：教科書，専門用語，段落，相互情報量

1. はじめに

本論文は，学習を可能にする条件としてのテキストの特徴を明らかにする研究の一部であり，ある領域の知識または特定の概念を論述していく際の部分的まとまり（以下，談話的まとまり）の長さを分析する。本論文では学習用に編纂された代表的なテキストであり，扱われる領域や対象とする学年が明確である教科書を分析対象とする。これにより，将来的に専門書や教養書など様々なテキストをどの教科・学年に近いかという観点から分類し，それぞれの特徴を直感的に把握するための参照点となるデータを得ることを目的とする。

教科書に限らずテキストを分析対象とする研究は多く，それらはテキスト分析と総称されることもある。テキスト分析は大きく「何が記述されているか」に着目する内容分析と「どのように記述されているか」に着目する文体分析に分けられる。前者は新聞・雑誌，教科書，アンケートなど，後者は小説や古典文学，政治文書などを対象とした研究がある（上田ほか，1998; 鈴木・影浦，2008; 陳，2012）。また近年では著者推定や真贋判定，難易度推定といった応用研究もなされている（金，2016）。

テキストを分析対象とした上記の研究では，テキストを特徴づける何らかの指標（以下，テキスト特徴量）が利用されるのが一般的である。内容分析では何らかの概念やカテゴリーを表す語の出現頻度が，文体分析では品詞別・語種別の相対頻度や指示語，色彩語などの出現頻度が，著者推定や真贋判定では Type-Token

*　愛知淑徳大学人間情報学部

Ratio や Yule の K など出現頻度分布の要約統計量が利用されている（石田ほか，2004）。浅石（2017）はテキスト特徴量を文字，語，文節，文，段落といった言語単位ごとに整理しているが，特に語の出現頻度に関わるテキスト特徴量は多くの研究で利用されている。

上記のテキスト特徴量は，いずれもテキスト全体の静的な特徴を示す指標である。しかし，テキストを学習の対象と捉えて分析する際にはテキストの動的な特徴，または時系列的な特徴を示す指標が必要になる。学習者である読み手はテキストを「読み進めていく」過程で学習するからである。個々の語の出現に着目する場合も，テキスト全体の出現頻度だけでなく出現過程を示す指標が必要である。本論文ではそのようなテキストの時系列的な特徴を示す指標の一つとして，テキスト中で語が特徴的に出現する範囲に着目する。

本論文の構成は以下の通りである。続く第２章では，Montemurro and Zanette（2010）で提案された方法を改変する形で，テキストの談話的まとまりの計算方法を説明する。また，分析対象とする教科書とデータも提示する。第３章では分析結果を記述する。ここではまずテキスト全体の談話的まとまりを，次に個々の専門用語についての談話的まとまりを記述する。第４章では，分野・学年・時代間の比較結果と，談話的まとまりが特に長い（または短い）専門用語の役割を考察する。第５章では本論文のまとめと今後の課題を述べる。

2. 分析方法

2.1　分析手続き

本論文では，Montemurro and Zanette（2010）（以下，先行研究とする）での提案方法を改変して，テキストの談話的まとまりを計算する。先行研究では書き言葉において「最も情報量のある区分」（most informative segments）の大きさを求めることを目的として，テキスト全体を bag of words とみなし，各語の出現に関する情報量が最大となる（より正確には，実際のテキストと，語をランダムに出現させた場合との相互情報量の差が最大になる）ようにテキストを等分割した時の各部分の長さを求める，という方法を提案している。

本論文で先行研究の提案方法を改変した点は以下の３点である。第一に，先行研究ではテキストを一定の延べ語数で分割しているのに対して，本論文では一定の段落数でテキストを分割する。これは，学習対象としてのテキストの特徴を分析するという本研究の問題意識に合わせるためである。段落はトピックセグメンテーションの単位となる（Hearst, 1997）など内容の部分的まとまりに相当する言語単位であると同時に，テキストのリーダビリティ改善実験で段落の順序や段落間の関係が改善項目に挙げられる（酒井，2011）など読み手が内容を理解していく上での言語単位である。

第二に，先行研究ではテキスト中の全ての語の出現過程を考慮しているが，本論文では内容を学習する上で特に重要な専門用語に限定する。実際の専門用語には巻末索引の見出し語（以下，索引語）を利用した。テキスト中の専門用語としては，日本語の形態素解析器（ChaSen，MeCab）の解析結果から抽出した内容語やその中の一部を利用する方法も考えられる。しかし，それらの中には明らかに科目の概念を表さないものが無視できない程度に多く入っていたため，上記の方法を選択した。

第三に，先行研究ではジャンルの異なる３点のテキストを分析しているのに対して，本研究では対象を教科書に限定し，談話的まとまりの長さとテキストの社会的属性（分野・学年・時代）との関係を考察する。また，個々の専門用語について特徴的なものはテキストにおける論述上の役割についても考察する。

ここから具体的な分析手順を説明する。はじめに分析に必要な記号を次のように定義する。

N：テキスト長（延べ語数）

V(N)：語彙量（異なり語数）

P：段落数

W：専門用語の確率変数（確率変数の値は 1, 2, ... , V(N)）

J：段落の確率変数（確率変数の値は 1, 2, ... , P）

n：専門用語 w の出現頻度

n_j：冒頭から j 番目の段落（j = 1, 2, … , P）での w の出現頻度

N_j：j 番目の段落の長さ（j 番目の段落に出現する専門用語の延べ語数）

そして、テキストの談話的まとまりを以下のステップで求めていく。

1. 連続する複数の段落を１ユニットとして，テキストを P ユニットに分割する。
2. 分割したテキストに対して，専門用語とユニットの相互情報量 MI(W, J)を計算する[1)]。ただし，相互情報量 MI(W, J)は次のように定義される。

$$MI(W,\ J)=\sum_{w=1}^{V(N)}p(w)\sum_{j=1}^{P}p(j|w)\log_2\frac{p(w)p(j|w)}{p(w)p(j)}$$
$$=\sum_{w=1}^{V(N)}p(w)\sum_{j=1}^{P}p(j|w)\log_2\frac{p(j|w)}{p(j)}$$

3. テキスト中の専門用語をランダムに並び替えた場合の相互情報量〈MI(W, J)〉を計算する。〈MI(W, J)〉は以下のように定義される。ただし〈p(j|w)*〉は，専門用語をランダムに並び替えた全ての場合にわたる p(j|w)の平均を表す。

$$<MI(W,\ J)>=\sum_{w=1}^{V(N)}p(w)\sum_{j=1}^{P}<p(j|w)^*>\log_2\frac{<p(j|w)^*>}{p(j)}$$

4. MI(W, J)と〈MI(W, J)〉との差を計算する
5. １ユニットの段落数を変えて１～４を実施し，MI(W, J)と〈MI(W, J)〉との差が最大となる時の段落数を談話的まとまりの長さとする

以上の手順で計算された談話的まとまりは，専門用語がランダムに出現した場合と実際のテキストとの相互情報量の差が最大になる時，換言すれば専門用語が最も秩序を保って出現するようテキストを分割した時の１ユニットの段落数を示している。ここで，MI(W, J) − 〈MI(W, J)〉を１ユニットの段落数 s の関数 ΔI(s)とみると，先行研究から ΔI(s)は以下のように変形できる。

$$\Delta I(s)=MI(W,\ J)-<MI(W,\ J)>$$
$$=\sum_{w=1}^{V(N)}p(w)[<H(J|w)^*>-H(J|w)]$$

ただし〈H(J|w)*〉と H(J|w)の定義はそれぞれ以下の通りである。なお，先行研究ではテキストを一定の延べ語数で等分割するため各ユニットの相互情報量を P 倍しているが，本論文では各ユニットの延べ語数が異なるため，各ユニットについて計算した相互情報量を足し合わせる。

$$<H(J|w)^*>=\sum_{j=1}^{P}\left(\sum_{m_j=1}^{min\{n_j,\ N_j\}}p(m_j)\frac{m_j}{n}\log_2\frac{m_j}{n}\right)$$
$$H(J|w)=-\sum_{j=1}^{P}\frac{n_j}{n}\log_2\frac{n_j}{n}$$

1）相互情報量は, 2 変数の関連の強さを測る自己相互情報量の全ての対にわたる平均である。ここでは，専門用語 w とユニット j とが共起しやすい（ある専門用語 w が出現したことで, そこが j 番目のユニットであると予測しやすい）ほど, MI(W, J)は大きくなる。

$\langle H(J|w)^*\rangle$を求める際の$p(m_j)$は，分割したテキストのｊ番目のユニットで専門用語ｗがm_j回出現する確率であり，次のように定義される。ただしｎが大きい場合，組み合わせ計算の量が非常に大きくなり計算不能になってしまう。そのため，計算の途中で出現する階乗計算はスターリングの公式($\log_e(n!) \fallingdotseq n(\log_e n-1)$)で近似した。

$$p(m_j)=\frac{\binom{n}{m_j}\binom{N-n}{N_j-m_j}}{\binom{N}{N_j}}$$

上述したように，１ユニットの段落数ｓを変えていく（２段落を１ユニット，３段落を１ユニット……としていく）中で，$\Delta I(s)$が最大になるときのｓをテキストの談話的まとまりとする。これをｕとするとｕは以下のように定義される。

$$u=\arg\max \Delta I(s)$$

2.2　分析対象とデータ

中学・高校の理科教科書を分析対象とした。複数の理科教科書を分析することで，多くの教科書に共通する一般的傾向と分野・学年・時代別の傾向を明らかにする。具体的には1998年告示（高校は1999年）の学習指導要領の科目から，中学では「理科（第１分野）」「理科（第２分野）」，高校は「物理Ⅰ」「化学Ⅰ」「生物Ⅰ」「地学Ⅰ」「物理Ⅱ」「化学Ⅱ」「生物Ⅱ」「地学Ⅱ」，５期の学習指導要領のもと作成された中学理科教科書を対象とする（以下，告示年に応じて「1958年」「1969年」「1977年」「1989年」「1998年」とする。科目は「理科（第１分野）」に限定した）。

各教科書の本文をテキストデータ化して直接の分析対象とする。本文には学習すべき内容の主たる部分が記述されており，読み進めていく過程に応じた特徴を分析する上でも本文に限定する（「実験・観察」「問題」などを含めない）方が良いと考えたためである。本文をテキストデータ化する際には，最低限の修正を施して入力した。具体的には独立した化学式や数式，漢字の読み仮名などは入力対象から除外した。また，丸囲みの文字や化学式の価数・添字などそのままではテキストデータ化できない部分は最低限の修正を加えて入力した。詳細は浅石（2016）を参照のこと。

表１に各教科書における索引語とテキストデータの基本統計量を示す。段落は行頭の空白部分を区切りとみなした。表１から，個々の索引語の本文中での平均出現頻度N/V(N)や，一段落あたりの専門用語数N/Pも教科書間の差が大きいことが分かる。専門用語の出現に関して，複合語の一部として出現する場合もカウントした。例えば中学の「理科（第１分野）」では「エネルギー」と「運動エネルギー」が共に索引語になっているが,「運動エネルギー」が出現した場合は「エネルギー」も出現したと考える。また「水（みず）」と「水（すい）」,「力（ちから）」と「力（りょく）」のように読み方が異なる場合も，概念上は少なくとも重複していると考え区別なくカウントした。

3.　分析結果

3.1　テキスト全体の談話的まとまり

図１に段落数ｓに応じた$\Delta I(s)$の推移を示す。図１を見ると，いずれの教科書でもｓが大きくなるにつれ$\Delta I(s)$も増加していくが，$\Delta I(s)$が最大値を迎えた後は一様に減少することが分かる。表２に$\Delta I(s)$が最大値をとる時の段落数ｕを示す。表２から，談話的まとまりが最も長いのは「1998年」（19段落）で最も短いのは「物

表１　索引語とテキストデータの基本統計量

	V(N)	N	P	N/V(N)	N/P
理科（第１分野）	200	2647	345	13.23	7.67
理科（第２分野）	180	1408	289	7.82	4.87
物理Ⅰ	289	1408	289	7.82	4.87
物理Ⅱ	354	3722	504	10.51	7.38
化学Ⅰ	552	7955	675	14.41	11.79
化学Ⅱ	261	1604	505	6.15	3.18
生物Ⅰ	520	4487	465	8.63	9.65
生物Ⅱ	431	3035	472	7.04	6.43
地学Ⅰ	493	2660	441	5.40	6.03
地学Ⅱ	286	1453	450	5.08	3.23
1958 年	360	5436	915	15.10	5.94
1969 年	365	5832	945	15.98	6.17
1977 年	169	2564	482	15.17	5.32
1989 年	136	1596	485	11.74	3.29
1998 年	163	1843	412	11.31	4.47

理Ⅰ」(6 段落）であること，その他の教科書は概ね 10 ～ 15 段落の範囲に収まっていることが分かる。

表 3 に，分野・学年・年代間での比較結果を示す。左の教科書ほど談話的まとまりが長く，差が 5 以上ある場合は≫で強調した。なお，中学の「理科（第 1 分野)」では物理と化学領域,「理科（第 2 分野)」では生物と地学領域と複数の領域が扱われているため分野間の比較には利用しなかったが，学年間の比較では利用している。

表 3 を見ると，分野間で比較すると上級学年（「Ⅱ」のつく科目）と下級学年（「Ⅰ」のつく科目）とで傾向が一貫していない。例えば地学は下級学年では談話的まとまりが最も長いが，上級学年では最も短い。一方，学年間で比較すると概ね中学より高校の談話的まとまりが長く，特に生物分野では学年が上がるにつれて談話的まとまりが長くなっている。高校の上級学年と下級学年を比較した場合，地学分野を除いては上級学年の方が談話的まとまりは長い。時代間で比較すると「1998 年」が最も談話的まとまりが長く,「1969 年」「1958 年」「1977 年」「1989 年」と続いており，過去（または近年）になるほど長い（または短い）といった傾向は見られない。

3.2　個別の専門用語の談話的まとまり

次に，個々の専門用語の談話的まとまりを計算する。前項で計算したテキスト全体の談話的まとまりが出現頻度分布の要約統計量に相当すると考えると（正確には，個々の専門用語の談話的まとまりの期待値となっている)，本節で計算するのは特定の語の出現頻度に相当する。ただしここでは提案手法の応用可能性を検討するため，全ての専門用語の談話的まとまりを計算した上で特徴的なものを探索的に分析する。2.1 を振り返ると，ΔI(s)は全ての専門用語に関する以下の式の総和として定義されていた。

$$p(w)[<H(J|w)^{*}>-H(J|w)]$$

ここから，個々の専門用語の談話的まとまりは，上記の式を最大化する s として定義できる。相対出現頻度 p(w)は s と無関係なので，以下の v が各専門用語の出現する談話的まとまりとなる。

$$v=\arg\ \max[<H(J|w)^{*}>-H(J|w)]$$

各専門用語の談話的まとまり v の度数分布の要約統計量を表 4 に示す。表 4 から個々の専門用語の談話的まとまり v は平均で 13 ～ 27 程度と幅があり，中央値で比較しても最小の 2（「地学Ⅱ」）から最大の 13（「1969 年」）まで幅があることが分かる。ただし，いずれの教科書も談話的まとまりの短い専門用語が多い一方，談

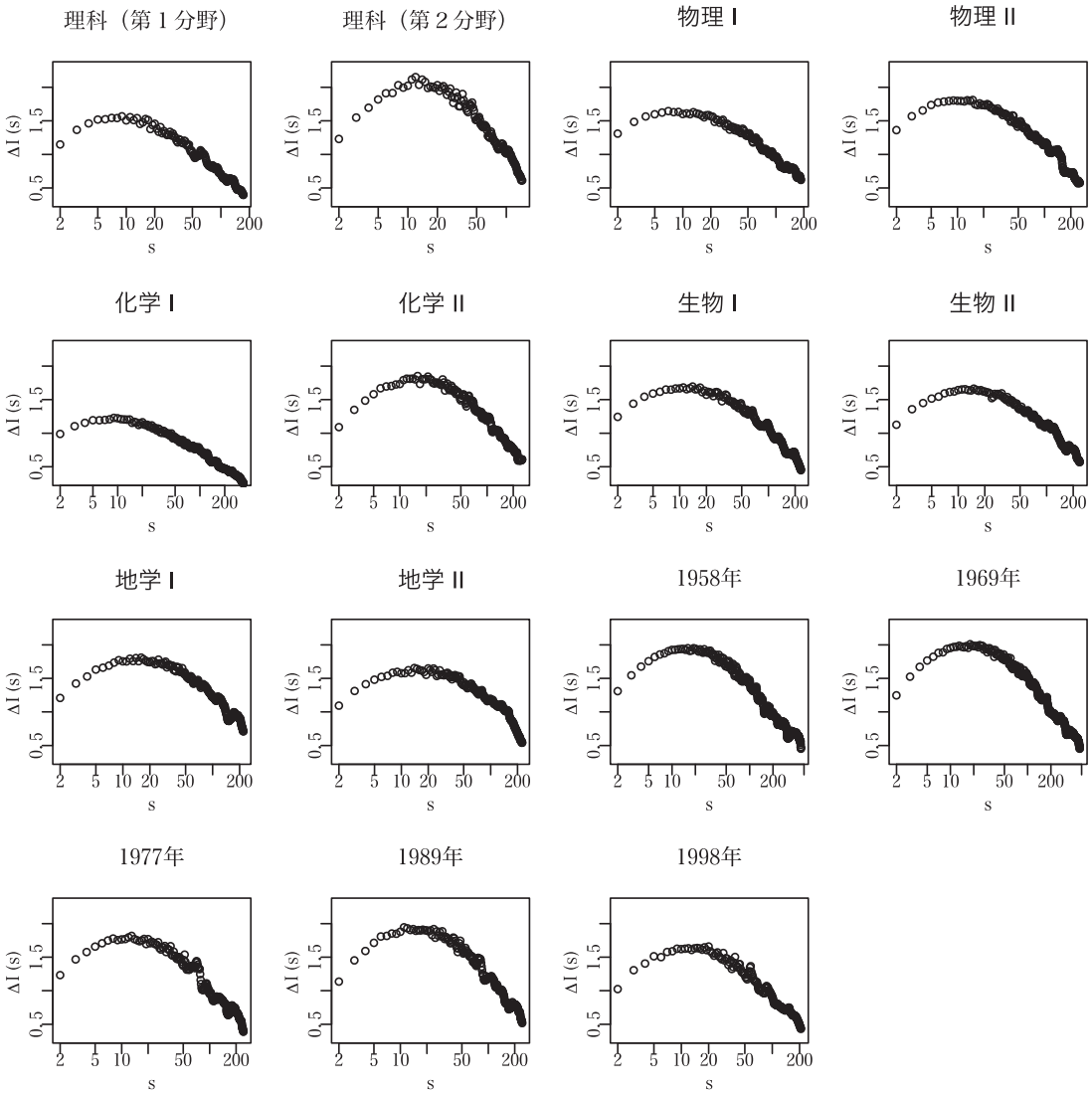

図１　段落数ｓに応じた相互情報量の差ΔI(s)の推移

表２　各教科書の談話的まとまりの長さ

教科書	u	教科書	u	教科書	u
理科(第１分野)	8	化学Ⅱ	15	1958 年	15
理科(第２分野)	11	生物Ⅰ	13	1969 年	17
物理Ⅰ	6	生物Ⅱ	14	1977 年	12
物理Ⅱ	14	地学Ⅰ	15	1989 年	10
化学Ⅰ	8	地学Ⅱ	13	1998 年	19

表３　談話的まとまりの長さによる分野・学年・時代間比較の結果

分野間比較	高校（下級学年） 高校（上級学年）	地学Ⅰ＞生物Ⅰ＞＞化学Ⅰ＞物理Ⅰ 化学Ⅱ＞生物Ⅱ＝物理Ⅱ＞地学Ⅱ
学年間比較	物理分野 化学分野 生物分野 地学分野	物理Ⅱ＞理科（第１分野)＞物理Ⅰ 化学Ⅱ＞＞化学Ⅰ＝理科（第１分野） 生物Ⅱ＞生物Ⅰ＞理科（第２分野） 地学Ⅰ＞地学Ⅱ＞理科（第２分野）
時代間比較		1998 年 >1969 年 >1958 年 >1977 年 >1989 年

話的まとまりが非常に長い専門用語が少数存在する。談話的まとまりに応じた専門用語の異なり語数を図２に示した。ただし V(v)は談話的まとまりの長さが v である専門用語の異なり語数を表す。

表４　各専門用語の談話的まとまりの度数分布の要約統計量

	V(N)	平均	標準偏差	最大値	3/4 点	中央値	1/4 点	最小値
理科（第１分野）	200	13.64	23.99	157	16	5	1	1
理科（第２分野）	180	10.17	13.25	121	13	5	1	1
物理Ⅰ	289	19.40	30.50	159	21	6	1	1
物理Ⅱ	354	18.14	28.55	195	22	6	1	1
化学Ⅰ	552	27.73	52.50	330	29	7	1	1
化学Ⅱ	261	16.49	31.48	224	20	4	1	1
生物Ⅰ	520	13.81	24.95	228	15	3	1	1
生物Ⅱ	431	17.19	29.60	199	19	4	1	1
地学Ⅰ	493	14.46	29.27	189	14	2	1	1
地学Ⅱ	286	20.09	35.34	217	20	4	1	1
1958 年	360	20.86	44.38	358	22	6	1	1
1969 年	365	26.49	41.99	429	35	13	1	1
1977 年	169	20.30	28.18	185	25	11	1	1
1989 年	136	14.37	22.71	184	20	6	1	1
1998 年	163	15.32	23.90	153	20	6	1	1

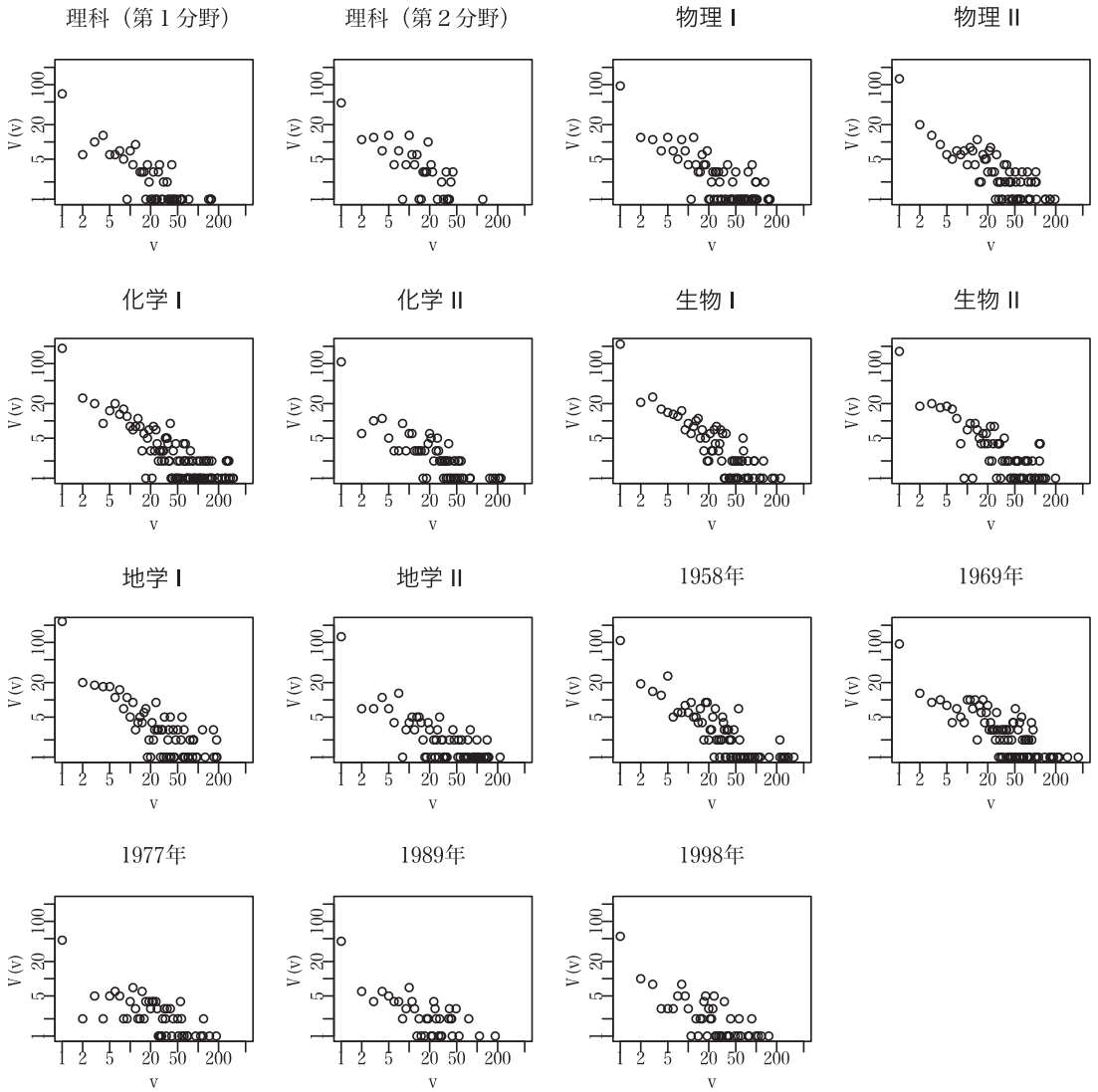

図２　個々の専門用語の談話的まとまり

4. 考察

テキスト全体の談話的まとまりは，概ね学年が上がるにつれて長くなる。この要因としては，中学の教科書は複数の領域の和集合的な性格が強いのに対して，高校では一つの分野としての輪郭がはっきりしており，個々の内容が丁寧に説明されるためと解釈できる。また，分野や時代に応じた傾向は見られなかったが，これは分野や時代よりも出版社の編集方針など他の属性が強く影響している可能性が考えられる。ただしこれらを検証するには，さらに多くの出版社の教科書，または他のジャンルのテキスト（新聞や雑誌，小説など）と比較する必要がある。

個別の専門用語に関しては，談話的まとまりが長いものを具体的に見ると，大きく間隔を空けて出現するものが見られた。例えば「1969 年」で談話的まとまりが最も長い「クロマトグラフィー」（v＝429）は，同教科書の「単元１（物質の特性）」と,「単元５（化学変化と分子・原子）」で１回ずつ出現しており，単元５では「……1 年「1 物質の特性」で行なったクロマトグラフィーも，少量の混合物から純物質を分離する方法である」のように前の文脈を振り返る形で出現している。このように談話的まとまりが長い専門用語は教科書内の離れた部分をつなぎ，一つの科目としてのまとまりを生む機能を持っている可能性が示唆される。

一方で談話的まとまりが短い専門用語には，専門用語全体の多くを占める低頻度語が含まれる。特に出現頻度が１～２程度の低頻度語の殆どは特定の段落（または連続した少数の段落）でしか出現しないため，談話的まとまりの長さも１となる。その一方で，語彙量は少ないが教科書全体を通じて出現する高頻度語も，集中的に出現する段落がある場合は談話的まとまりが短くなる（例えば「生物Ⅰ」「化学Ⅰ」の最高頻度語はそれぞれ「細胞」「酸」で，いずれも談話的まとまりは１である）。これらに対して談話的まとまりが 10 ～ 20 程度の専門用語の中には，各章（または節）で重点的に出現し，そこでの核となる概念が含まれる。

5. おわりに

本論文では，学習という行為を考えた時に決定的に重要となるテキストの時系列的な特徴の一つとして，テキストの談話的まとまりを分析した。そのために先行研究の提案方法を一部改変して中学・高校の理科教科書に適用し，テキスト全体の談話的まとまりの長さを分野・学年・時代間で比較するとともに，個々の専門用語の談話的まとまりを計算した。その結果，学年が上がるにつれ談話的まとまりが長くなる傾向が確認できたほか，談話的まとまりが長い専門用語はテキストの異なる部分が結びつける役割を持つ可能性が示唆された。

今後の課題として，まず利用する専門用語の範囲を変えて同様の分析を行い，内容の専門度に応じた談話的まとまりを分析することが考えられる。具体的には ChaSen や MeCab による形態素解析の結果から内容語全体または名詞など一部の品詞に限定する方法，出現状況に応じて付与した重要度の高い語に限定する方法が考えられる。また，テキストを章や節ごとに分割した上で各単位の談話的まとまりを計算することで，トピックごとの傾向を分析することも考えられる。

テキストの時系列的な特徴を示す他の指標の探索も今後の検討課題である。浅石（2016）や Asaishi & Kageura（2016）ではテキストを読み進めていく過程での知識の形成過程を語彙ネットワークの成長として近似し，複数のネットワーク統計量の推移として可視化・分析している。例えばそれらのネットワーク統計量の変動を時系列的特徴とみなし，経済学などで株価や景気の変動を記述するために使われる平均成長率やボラティリティを指標とすることが考えられる。

6. 謝辞

本論文は2016年11月に東京大学大学院教育学研究科に提出した博士論文の一部を加筆・修正したものです。主査である東京大学大学院教育学研究科の影浦峡教授および審査員の先生方にはご指導とご助言を賜りました。また，本研究はJSPS科研費（研究活動スタート支援）26880005の助成を受けました。ここに記して感謝の意を表します。

引用文献

浅石卓真（2016）．高校理科教科書における知識の形成過程：テキストにおける語彙ネットワークの成長過程の分析　日本図書館情報学会誌，*62*，38–53.

Asaishi, T. & Kageura, K. (2016). Growth of the terminological networks in junior-high and high school textbooks. *Proceedings of Joint Second Workshop on Language and Ontology & Knowledge Structures*, 30–37.

浅石卓真（2017）．テキストの特徴を計量する指標の概観　日本図書館情報学会誌，*63*，159–169.

陳 志文（2012）．新聞，週刊誌，高校教科書に見られる文体の類型と特性　現代日本語の計量文体論（pp. 105–122）くろしお出版.

Hearst, M. A. (1997). Texttiling: segmenting text into multi-paragraph subtopic passages,” *Computational Linguistics*, *23*, 33–64.

石田栄美・安形 輝・野末道子・久野高志・池内 淳・上田修一（2004）．文体からみた学術的文献の特徴分析　三田図書館・情報学会研究大会発表論文集，33–36.

金 明哲（2016）．計量文献学の基礎研究とその応用　村上征勝・金　明哲・土山　玄・上阪彩香（編）　計量文献学の射程（pp. 60–84）　勉誠出版.

Montemurro, M. A. & Zanette, D. H. (2010). Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, *13*, 135–153.

酒井由紀子（2011）．健康医学情報を伝える日本語テキストのリーダビリティの改善とその評価：一般市民向け疾病説明テキストの読みやすさと内容理解のしやすさの改善実験　*Library and information science*, (65), 1–35.

鈴木崇史・影浦 峡（2008）．総理大臣国会演説における基本的文体特徴量の探索的分析　計量国語学，*26*，113–122.

上田英代・村上征勝・藤田真理（1998）．源氏物語の会話文と地の文をめぐる数量分析：助動詞を中心に　計量国語学，*21*，193–205.