教育情報学研究　第14号（2015）
Educational Informatics Research, 2015, No. 14, 53-58

文章一貫性測定へのLatent Semantic Analysisの適用可能性 ―説明文による検討―

牛　娜*，尹　得霞**，北村　勝朗***

*深圳大学外国語学院
**東北大学大学院教育情報学研究部
***東北大学大学院教育情報学研究部

要旨：文や節などのつながりを意味する「一貫性（coherence）」のある文章は，理解が容易である．しかし一貫性が欠如した文章は論理性が乏しいため読み手に意味が伝わりにくい．したがって，一貫性は文章の内容の一致性や質を判断するための重要な基準である．しかし，定量的に一貫性を測定する手法は存在しない．本研究では，説明文における一貫性の定量的な測定の観点から，Landauer & Dumais（1998）による"潜在意味分析（Latent Semantic Analysis：LSA）"に基づき，小学校国語教科書の説明文を題材とし，説明文の一貫性の自動測定方法の検討を行った．分析の結果，LSAによる一貫性の測定が定量的に示せることが明らかとなった．

キーワード：一貫性，自然言語処理，潜在意味分析，特異値分解

1. はじめに

一貫性とは，談話における個々の発話と文脈を関係付けることによって作り出される解釈可能な意味的まとまりを指す．首尾一貫性とも言われている．結束性などをもたらす言語的な要素だけでなく，話し手と聞き手，書き手と読み手が共有する常識や推論，また連想などの非言語的要素なども含めた談話の意味的なつながりの善し悪しのことを言う（亀山，1999）．一貫性は，文章や談話の内容の連続性を重視するものである．一貫性が欠けると，話がずれたり，意味がわからなくなったりすることが多い．したがって，一貫性は文章の内容の一致性や質を判断する重要な基準と考えられる．しかし，定量的に一貫性を測定する手法は存在しない．

一方，LSAは，複数の文章間及び単語間，及び文章と単語間の意味的な類似度を検出する方法である．更にLSAを用いることによって，文章の内容の理解を測定することができ，テキストの主観評定（小論文の成績等）も可能となる．

そこで，本研究では，Deerwester, S., et al. による潜在意味分析（LSA : Latent Semantic Analysis）[4]に着目し，一貫性測定への適用可能性を検討する[1]．本研究を基礎とし，信頼性の高い一貫性自動測定システムが開発されることにより，従来の一貫性測定に一定の尺度基準が提供可能となるとともに，読解指導に定量な方法が提供可能となり，日本語教師の教育活動支援につながると考えられる．

2. LSAを用いた一貫性の測定実験

2.1 実験に使用した材料

本研究では，日本の小学校国語教科書に掲載されている説明文を測定の対象として使用した．5編の説明文[2]を表1に示した．5編の説明文の平均総文字数は922文字であった．

表1．説明文タイトル一覧

文章	学年	タイトル	総文字数
1	国語3年	ヤドカリのすみかえ	467
2	国語3年	おおばこ	419
3	国語4年	カブトガニを守る	540
4	国語4年	恐竜を探る	1231
5	国語6年	人類は滅びるか	1956

2.2 実験の手続き

2.2.1 内容を表す品詞の抽出

単語・文（センテンス）行列を作成する前に，内容を表す品詞の抽出を行う必要がある．本研究では，日本語形態素解析システムChaSen（茶筌）[3]を用いて，文章に対して形態素解析を行い，得られた品詞のうち一般名詞，サ変接続名詞，副詞，説明文の内容に特徴的な4種類の単語（未知語）を抽出した．表2は文章1における1番目の文（ヤドカリはまき貝のから中に住んでいます．）の形態素解析の結果である．

表2．形態素解析の結果例

词语	词性
ヤドカリ	名詞-一般
は	助詞-係助詞
まき	名詞-一般
貝	名詞-一般
の	名詞-非自立-一般
から	名詞-一般
中	名詞-非自立-副詞可能
に	助詞-格助詞-一般
住ん	動詞-自立　五段
で	助詞-接続助詞
い	動詞-非自立
ます	助動詞

2.2.2 単語・文（センテンス）の対応行列の作成

次に，単語・文（センテンス）の対応行列を作成した．作成された行列を表3に示す．

文章1には，15の文（センテンス）が出現した．

単語については，例えば「ヤドカリ」という単語は，センテンス1（以下，"センテンス"は"S"と略記），S3，S14にそれぞれ1回づつ，S15に3回出現していることがわかる（S4〜S13は，紙幅の関係から省略）．

表3．単語・文（センテンス）の対応行列（X）

	1	2	3	…	14	15
ヤドカリ	1	0	1	…	1	3
まき	1	0	0	…	0	0
貝	1	0	0	…	0	0
から	1	0	0	…	0	0
中	1	0	0	…	1	0
貝がら	0	1	0	…	1	1
それで	0	0	1	…	0	0
体	0	0	1	…	0	1
成長	0	0	1	…	0	0
なんど	0	0	1	…	0	0
とき	0	0	0	…	0	0
ほか	0	0	0	…	0	0
海	0	0	0	…	0	0
観察	0	0	0	…	0	0
相手	0	0	0	…	1	0
はじめ	0	0	0	…	0	0
入り口	0	0	0	…	0	0
きず	0	0	0	…	0	0
次	0	0	0	…	0	0
自分	0	0	0	…	0	1
そして	0	0	0	…	0	0
空き	0	0	0	…	1	0

2.2.3 特異値分解と対応行列の再構成

次に，表3の「単語と文（テキスト）の対応行列（X）」に対して，特異値分解（SVD：Singular Value Decomposition）を行い，行列（X）を三つの積の形（①単語の特徴を表す行列（単語の特徴ベクトル：T）②固有値を表す対称行列（固有値ベクトル：S）③文章の特徴を表す行列（文の特徴ベクトル：D'））に分解した．

特異値分解は式（1）による．

$$X=TSD' \quad (1)$$

SVDによって得られた各行列より，固有値が大きい順に二つの要素を選択し，単語及び文の意味空間を再構成し，再構成された行列（X'）を算出した（表4）．

表４．単語・文（センテンス）の対応行列（X'）

	1	2	3	…	14	15
ヤドカリ	1.00	0.05	1.28	…	0.88	3.05
まき	0.05	0.00	0.07	…	0.04	0.16
貝	0.05	0.00	0.07	…	0.04	0.16
から	0.05	0.09	0.02	…	0.18	0.14
中	0.08	0.18	0.01	…	0.35	0.21
貝がら	0.32	0.61	0.10	…	1.24	0.85
それで	0.07	-0.02	0.10	…	0.03	0.22
体	0.24	0.00	0.31	…	0.19	0.73
成長	0.07	-0.02	0.10	…	0.03	0.22
なんど	0.06	0.09	0.04	…	0.19	0.18
とき	0.05	0.06	0.04	…	0.13	0.15
ほか	0.05	0.06	0.04	…	0.13	0.15
海	0.06	-0.02	0.08	…	0.02	0.17
観察	0.06	-0.02	0.08	…	0.02	0.17
相手	0.06	0.35	-0.10	…	0.62	0.13
はじめ	0.04	0.05	0.03	…	0.11	0.11
入り口	0.03	0.15	-0.03	…	0.28	0.07
きず	0.00	0.03	-0.02	…	0.05	-0.01
次	-0.01	0.10	-0.06	…	0.16	-0.04
自分	0.16	0.10	0.16	…	0.28	0.46
そして	-0.01	0.07	-0.05	…	0.10	-0.04
空き	0.04	0.05	0.03	…	0.11	0.12

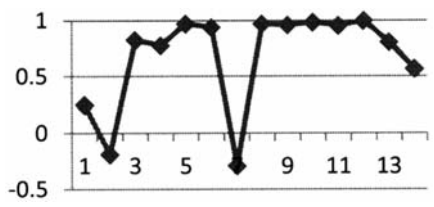
※　縦軸：単語，横軸：文（センテンス）

※　S4～S13は，紙幅の関係から省略

2.2.4 文と文の相関係数の算出

相関からセンテンス間の意味的な類似度，すなわち文脈同士の意味的な結びつきの強弱の程度を調べるために，センテンス間の相関係数を算出した．センテンス間の相関係数の結果を表5に示す．文と文の意味的なつながりを図1に示す．

図１．文と文の意味的なつながり



2.2.5 一貫性の計算

表5に示されている通り，赤の太数字は隣接の文の相関係数である．それぞれの平均値は文章の一貫性の結果と捉えられる．LSAによる文章1の一貫性は0.68215であった．

3. 結果と考察

3.1 結果

以上の計算方法による5篇の説明文の一貫性の結果は表6に示す．表6より，5篇の説明文の一貫性の結果は0.5～0.8の区間に集中していることが明らかとなった．5篇の説明文は国語教科書に掲載されているものであり，模範的で，つながりのよい文章であると考えられる．文章の一貫性測定へのLSAの適用可能性が高いと考えられる．

表６．一貫性の結果

文章	coherence
1. ヤドカリのすみかえ	0.682
2. おおばこ	0.543
3. カブトガニを守る	0.709
4. 恐竜を探る	0.704
5. 人類は滅びるか	0.794

3.2 考察

本論文では，小学校国語教科書の説明文を題材に，説明文の一貫性の自動測定方法を検討した．LSAによる一貫性の測定に定量的に示すことが可能である点が明らかとなった．一貫性の結果に影響を与える要素は以下のようにまとめられる．

１）文章の文字数はひとつの要素であると考えられる．本論文では，一貫性の結果が最も低い文章は［おおばこ］であり，文字数は最も少なかった．結果が最も高い文章は人［類は滅びるか］であり，文字数は最も多かった．こうしたことから，文字数が多いほど，LSAによる一貫性の結果が精確となる点が推察される．今後，文字数が多い文章を厳選した上で，実験による再検討が必要であると考えられる．

２）本研究では，5篇の模範説明文を用いて一貫性の測定を試みた．一貫性の結果が0.5～0.8という区間に集中していた．今後，文章の数を増や

し，実験を行うことで，つながりのよい模範的な説明文の一貫性の結果がどの区間に集中するか検討を行うことが課題として残される．

3）LSAによる一貫性の計算前に，測定の精度を高めるために予め大量の言語集体であるテキストコーパスを作成する必要がある[4]．今後，テキストコーパスを作成し，一貫性を再計算することが必要である．

4）今回は説明文を題材とし，一貫性の測定の検討を行った．今後，他の体裁の文章を用いて，一貫性の測定を試みる必要があると考えられる．

注：

[1] 本論文は深圳大学人文社会科学基金項目"日本語読解力自動測定方法についての研究"（項目号11QNCG10）の研究成果に基づいて作成したものである．

[2] 5篇の説明文の作者は以下の通りである．

	作者
文章1： ヤドカリのすみかえ	今福道夫
文章2： おおばこ	真船和夫
文章3： カブトガニを守る	土屋圭示
文章4： 恐竜を探る	小畠郁生
文章5：人類は滅びるか	日高 敏隆

[3] 奈良先端科学技術大学院大学自然言語処理学講座が開発した日本語形態素解析システム（http://chasen-legacy.sourceforge.jp/）．無償のソフトウェアとして公開されている．

[4] 牛娜・村木英治（2008）Latent Semantic Analysisを用いた要約文の自動評価についての研究．日本テスト学会第6回大会発表論文抄録集，58-59（※「テキストコーパス（資料の総体）」とは，大規模な電子テキスト（何らかの文字集合で定義された文字と，特定のビット列との対応のルールに従い，0と1との組み合わせのみで構築された文字列データ）の集合を意味する．

参考文献

[1] Deerweter S, Dumais S T, Furnas G W et a1. Indexing by Latent Semantic Analysis [J]. Journal of the American Society for Information Science, 1990 ; 41 (6) : 391-407

[2] Kintsch W. Text comprehension, memory, and learning [J]. American Psychologist, 1994, (49) : 294-303.

[3] Landauer T K, et al. A Solution to Plato's Problem : The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge [J]. Psychological Review, 1997, 104 : 211-240.

[4] Landauer, T. k., Foltz, p. w., & Lahman, D. 1998. An Introduction to Latent Semantic Analysis [J]. Discourse Processes, 25, 259-284.

[5] Peter W Foltz, Darrell Laham, Thomas K Landauer. The Intelligent Essay Assessor : Applications to Educational Technology [J]. Interactive Multimedia Electronic Journal of Computer-Enhanced, Learning, 1999 ; 1 (2).

[6] Rehder B, Schreiner M E, Wolf B W, Laham D, Landauer T K, Kintsch W. Using Latent Semantic Analysis to assess knowledge : Some technical considerations [J]. Discourse Processes, 1998, (25) : 337-354.

[7] 亀山恵，談話分析：整合性と結束性．田窪行則・三藤博・片桐恭弘・西山佑司・亀山恵，岩波講座　言語の科学7　談話と文脈談話と文脈．第3章，岩波書店，P93－124，1999.

[8] 木村宗男．日本語教授法 [M]．凡人社，1982.

[9] 南之園博美.読解ストラテジーの使用と読解力との関係に関する調査研究　―外国語としての日本語テキスト読解の場合― [J]．世界の日本語教育7，1997.

[10] 中村・椿本・岸．文章評価へのLSAの適用可能性 [J]．日本教育心理学会第46回総会論文集，2004.

[11] 牛娜等．潜在意味分析を用いた日本語要約文自動評価に関する基礎研究 [J]．コンピュータ＆エデュケーション学会誌，56-60，2009.

[12] 椿本弥生，赤堀侃司．主観的レポート評価の系列効果を軽減するツールの開発と評価 [J]．日本教育工学学会論文誌30 (4)，275-282，2007.

[13] 椿本弥生，中村光伴，岸学，赤堀侃司（2004）文章読解指導へのLatent Semantic Analysis の適用可能性．日本教育心理学会第46回総会論文

集，549.
[14] 吉本清久．要約力を磨く説明文の指導 [M]．明治図書，2005.
[15] 牛娜・村木英治（2006）Latent Semantic Analysis を用いた要約文の評価の試み．日本テスト学会第 4 回大会発表論文抄録集，96-97.

The Possibility to measure coherence by latent semantic analysis

Na NIU*, Dexia YIN, Katsuro KITAMURA*****

* School of foreign Languages, Shenzhen University
** Graduate School of Educational Informatics, Tohoku University
*** Graduate School of Educational Informatics, Tohoku University

ABSTRACT

Textual coherence defines the connexity in a sense between different parts of a paragraph of texts or a discourse. When these parts are out of the connexity, there would be that utter words do not hang together, it makes people unintelligible. Only coherent discourse or texts can be read and understood. Therefore, coherence is the key criteria to measure the integrity, consistency and quality of a discourse. However, there is no quantitative measuring coherence approach for Japanese language. So, based on the latent semantic analysis, this research aims to present the viewpoint and method of quantitative measurement in textual coherence area, and also discuss the textual coherence auto-measuring approach with the theme of primary school Chinese notebook expository article.

Key words: Textual coherence, Natural language processing, Latent semantic analysis, Singular values decomposition (SVD)