文章ブロックの構造化における形態素タグとXMLの活用

ー年少者日本語教育への応用にむけた小学校教科書の最小文章単位のパターン化ー

中尾桂子，森下淳也
神戸大学大学院総合人間科学研究科
EMAIL:{nakao@ccs2000.cla.jm@}kobe-u.ac.jp

あらまし:本研究は，形態素解析処理を行った一文を，XMLにより単語単位で構造化し，単語の配列パターンによって文の並びの形状パターンを検出するものである．小学校理科教科書における談話の構成パターンに対して，単語の配列パターンの検出を試みることにより，この文脈パターンの構造化の手法を評価し，そのパターン検出の手段として用いたXML，形態素解析結果の利用を考察する．
キーワード: 日本語教育,文脈,文構造パターン,形態素解析,XML,系列検索

A Method for Sructuring Context with XML and Morphological Analyzer.

K. NAKAO, J. MORISITA
Graduate School of Cultural Studies and Human Science,Kobe University

Summary: We discuss the method for the context structuring in the textbooks of primary school, using the sequential patterns of the words between the sentences in small stream. Depends on the Instructical intension the stream has different construction patterns of sentence. It will be an important signature towards understanding the learning issues in each stream, especially for foreigner's pupils in Japan, who has no idea for the schem of context in textbook. We also discuss about exammining of the analysis with XML and Morphorogical Analyzer.

0.はじめに

ここ10年，外国人児童・生徒の日本の公立小中学校への編入が相次いだ．当初，現場は混乱したが，現在では最小限の日本語を教えるノウハウが充実してきている．しかし，ある程度日本語を覚えた外国人児童・生徒に教科学習等，専門性が高い日本語を指導するには，教科の日本語に不明な部分が多く，依然，教育現場ではその指導に苦慮している．

基礎学力の保証を目指す日本語指導においては，主に教科の内容を理解すること，また，学習概念を理解するために必要な日本語を習得させることが目的となる．

このような現状を受け，教科学習に必要な日本語を明らかにするために，算数，理科，社会科の教科書の語彙調査が行われ[1-3]，これらの語彙調査の結果をもとに基本的な語彙を中心にして文章を簡略化した算数教材が提案されている[4]．これは進学のための暫定的な措置であり，日本語能力が十分でなくても教科の概念が理解できることを第一目的としている．

外国人児童・生徒は進学するにつれ，さらに専門性が高くなる教科書から，自力で必要な情報を得，概念を習得していかなければならない．そのため，教科学習のための日本語指導においては，読解能力も合わせて養う必要がある．

一般に，成人に対する語学教育としての読解指導では，一字一句逐次理解する「精読」の他に，文章の筋やテーマを掴むことを目的として「多読」や「速読」が指導される[5]．特に，大要を把握する「掴み読み」や一定の情報だけを探して拾い読みする「検索読み」などが情報収集能力の向上につながるとして，このターゲットを検索するための語彙力と文法知識の指導，また，大要を掴むための指導として，文章内容のまとまりを認識するために，単語間や文どうしの関係性など，文章の構造を把握すること等，様々な実践活動に結び付けられている[6]．

外国人児童・生徒の場合，母語においても，また，それ以上に日本語においても，文章読解に既

有の知識として利用する一般的な「文脈」や「テキスト型」に触れる経験が少なく，一般的な文脈やテキスト型自体が認識できない．

そのため，外国人児童・生徒の読解教育としては，第一言語発達における言語教育の過程を踏まえて語学指導的な読解指導を盛込む必要がある．具体的には，語彙の体系的な理解や，文化・社会や著者の意図など，具体的な内容の背景に流れる文脈の認識能力，含意の論理性を認識する訓練が必要になる[7-10]．

しかしながら，現在，文中での単語の連続性や共起関係に基づく文章構造，また，その文章構造に基づいた論理的な含意の構造については基礎研究が少ない．

そこで，専門毎にまとめられた文章が見られる小学校教科書の構成パターンを文単位以上の単位において調べようと考える．

調査は，大きく，外国人児童への「教科学習につなげる日本語教育」を念頭に，情報収集攻略法の指導方法の可能性をさぐることが目標である．これを目標として，まず，文章構造を指導するために，理科の教科書に見られる学習の流れである文脈とそのパターンについて調査した[11]．これは，文末表現パターンに基づく学習の流れを，文章の流れから見い出すものであった．さらに具体的な情報を得るためには，文中の語の文脈情報を把握し，その並びの形状パターンを解析する必要がある．

そこで，文の形態素解析から得られる品詞情報を同時に格納できるように，XMLに基づいたデータ形式を拡張し，品詞情報に基づく視点から，再度，文の流れを文の形状パターンとして見直すことにした．

引き続き，本稿では，さらに具体化をはかる目的で，小学校の理科の教科書[12]を取り上げ，理科の文章の最小単位における文の関係について，その構成要素である単語を指標にして文の並びを調べる．なお，理科の教科書を利用するのは，談話内の文章があまり長くないため，談話毎の関係性が把握しやすいことから，調査手法の確認という目的に適うためである．

1.構造解析における形態素の役割

詳細な場面情報を文章から手に入れるには，談話の場面や著者の意図等の文脈を知る指標が必要である．

場面を知るには，語彙が重要であるが，語彙を取出し，頻度や類型を調べるのみでは，意味が一意に決定できない．単語の意味を一意に判断するには，文のならびにおいて単語が文の一要素として構成されている意図を捉える必要がある．

また，語どうしの関係を把握するためには意味だけではなく，各語が文を構成する上での機能的な性質を考慮する必要がある．

つまり，単語の意味機能と，文として構成されている状態との関係をもとに，文章のどの部分に文脈を知る指標があるかを明らかにすれば，語学教育としての読解指導に応用できる．

そこで，本の階層構造や文章と，その構成要素である各単語を，それらの出現する位置により関係づけ，その文章構造のパターンを明らかにしたい．

構造との関係に基づいて，教科書の構成要素である談話や文の特性を明らかにするためには，タグを付けることで階層構造を表現する汎用のデータ形式であるXML(eXtensible Mark-up Language)を利用する[13]．XMLはデータを階層化し，データの持つ上下関係や順序関係を木構造で表現するデータ形式であるが，その構造の解析や変換を行なうソフトウェアが充実しているため，階層に依存する探索や順序関係を抽出するのに適している．

そのため，[11]で文章単位にXMLタグをつけて構造を持たせたように，各単語にも構造をもたせて位置を示す意味を付与する．

日本語の文章において単語を認識するには，形態素解析が不可欠である．また，単語どうしの関係を把握するためには単語の機能的な性質を規定した品詞情報を利用したい．

そこで，形態素解析を行い，その解析時に得られる品詞情報を各文中の単語に付加し，その一文の各単語に付加した形態素解析時の品詞タグを利用して，一文中の単語にXMLタグをつける．これにより，一文中における単語単位まで構造化する．

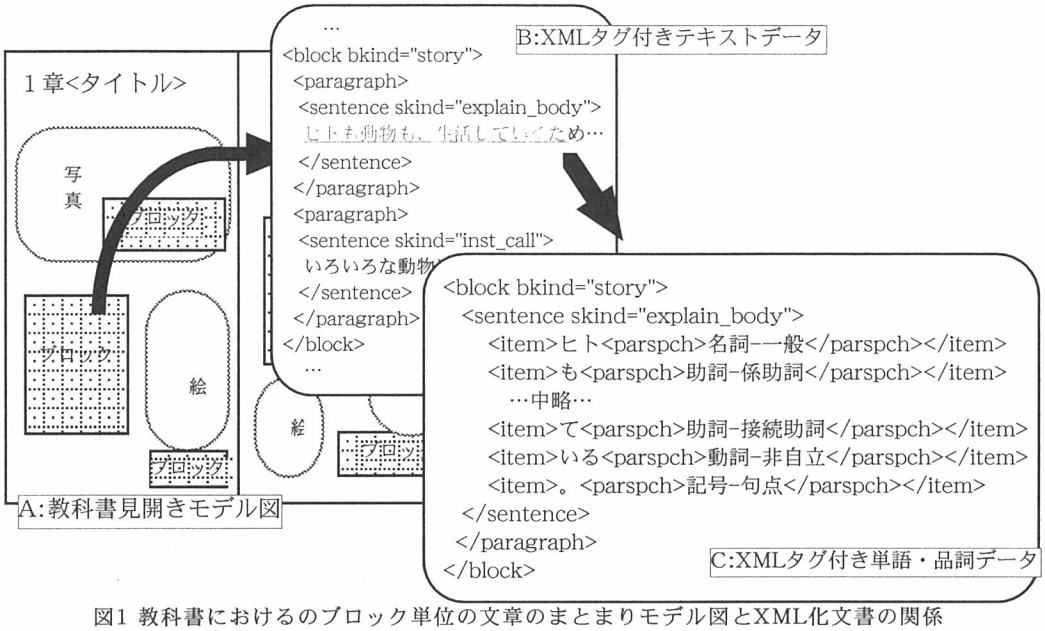

図1 教科書におけるのブロック単位の文章のまとまりモデル図とXML化文書の関係

このような方法を利用することで，文中の出現位置を正確に保持したまま単語単位まで構造化された文データが得られる．

今回利用するchasen2.2.8[14]は，解析結果が向上するようにIPA辞書の品詞定義が改良されており，さらに，解析精度向上のために新聞記事や文学作品などの解析を重ねて統計的な学習を辞書に施したものである．様々な分野の単語が現れる小学校教科書の解析にも対応できると考える．

2.形態素解析に基づくパターン検索

検索は，以下の手順で行なった．図1に教科書，ブロック，そのXML化について示す．

(1)図1のAのモデル図のように，教科書の二次元的な広がりを表現するために「ブロック」を定義し，複数のストリームを文書構造の階層下に位置付ける．

(2)節以下に「ブロック」という単位を設けて構造化することにより，図1のBのように，理科の教科書[12]をXML文書にする．

(3)章節構造を持たせた文書から，内容のひとまとまりであるブロック中の各文を，形態素解析システムchasen2.2.8[14]を利用して解析する．

(4)図1のCのように，解析過程で得られる品詞情報を各単語の下位層に付加し，文中の単語に構造をもたせる．

(5)ブロック内の文の配列に基づき，単語間の関係を単語の配列パターンから調べる．

(6)単語の並び方のパターンによって特徴付けられた文の並びの形状パターンを教科書の構造中で評価する．

(1)(2)は，前回の調査で作成した[11]．前回の調査では，教科書の構造における最小談話単位であるブロック自体の出現パターンを調べた．また，パターンの出力において，テキスト構造の中のブロックは全て同列に扱った．しかし，ブロックは全て同じレベルではなく，主流となる話の流れと副次的な話の流れがある．

今回の調査は，より具体的で忠実な教科書の特性を明らかにするために，前回の調査とは異なる観点から，ブロックの特性を調べるものである．前回同様，出現パターンを調べる形状検索ではあるが，今回は，ブロックの特性をブロックの内部の構成要素である単語の出現パターンから明らかにする．

なお，教科書は授業の資料という性質もあるため，写真や絵等が多い．しかし，一目でわかる情報は，日本語がわからない外国人児童・生徒でも，情報として受け取れると考え，写真の説明も含む日本語の文のみを調査の対象とする．文は末尾に句点が１つある文字列とする．

3.学習活動の文脈と教科書の構造

理科の教科書の文章は，学習活動の流れに基づいた事実や実験，それらのまとめなどで構成されている．さらに，これらの視覚的に目を引く効果や章節構造とは別に，目標とする学習内容へと学習活動を通じて学習者を導く流れがある[11]．この流れは，指示や注意等の表現等によって示されている学習活動への誘導という著者の意図に基づいている．

このような作者の意図等，表立って文章構成に現れ出てこない文脈といった談話の流れは，教科書の構造，単語の配列や文の並びなどから把握するものである．

本章では，読解時，パターン検索の手掛かりとなるこれらの文脈判断情報と教科書やその文章の構成との関係について述べる．

3.1.理科の教科書の構造とブロック

教科，学年の違いは多少あるが，小学校の教科書は文書であることから基本的には章節構造順に並んでいる．しかし，教科書は学習目的に応じた学習活動の手引きや資料といった役割を担う面があり，見やすさ，ポイントのフォーカスに配慮した構成となっている．このため，教科書の構造は二次元的な広がりを持つ．

教科にもよるが，基本的に小説等と同様に教科書にも談話の流れがある．しかし，実際には，各教科の１単元に相当する「章」の中には，「本文」，「囲み記事」や「写真説明」などが並列されており，異なる談話のまとまりが複数並べられた配置となっている．

内容の異なる別々のストリームと見られるひとまとまり毎の文章は，通常の小説等に見られる章，節，部，段落という文書構造木にはない単位である．この単位を考慮した教科書文書の構成を定義するため，この物理的なまとまりは，教科書の章節以下の構成単位の１つであるとして，「ブロック」と名付けた．

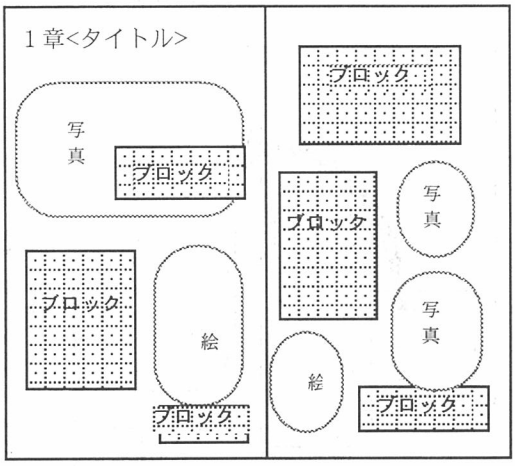

図2:教科書の見開き図

ブロックの性質は，1)文章談話の最小単位である．2)内容のまとまりを示す視覚的な単位である．3)ブロックを構成する文は，同様の文体で表されている，と定義できる．

図2にブロック毎に分かれた構成の一例を挙げる．図2は，見開きで見た小学校教科書が，視覚的にも，内容的にもブロックにまとまっていることを表した概念図である．

3.2.ブロックの出現パターン

ブロックは，その文体や位置により，機能や内容に特徴がある．啓林館新訂理科の場合，ブロックは，全部で7種類あった．各々，(1)ストーリーブロック，(2)課題提示ブロック，(3)実験ブロック，(4)見出しブロック，(5)囲み記事ブロック，(6)取り扱い説明ブロック，(7)まとめブロック，と呼ぶ．図3に章節構造下のブロックの出現の様子を表す．図3は，啓林館新訂理科[12]においては，全ての学年の教科書で見られる章節構造下の基本的なパターンをモデル化した図である．

章のはじまりには「ストーリーブロック」が現れ，その章の導入となる状況が説明されている．節内でも，通常，まず，節毎の導入部分となる「ストーリーブロック」があり，次に，考察ポイントでもある作業課題が提示され，さらに，実験手順が説明される．そして，最後に，作業の結果

明らかになったことが，「ストーリーブロック」か「見出しブロック」の形式により提示される．それらの流れの合間に，補足事項や応用的な内容が「囲み記事ブロック」や「見出しブロック」，「取扱い説明ブロック」などにより補われている．「見出しブロック」は，他のブロックのさらに下位の階層に組み込まれることがある．また，「ストーリーブロック」や「見出しブロック」は，章の前半に現れるか後半に現れるかにより，その機能が異なるが，基本的に文体が同じであることから，1種類のものだと考える．さらに，基本的なブロックの節内での配列は，若干の例外もある，図3のようである．章末には，啓林館新訂理科の場合，必ず，各単元の内容をダイジェストにまとめ，列挙したまとめブロックがある．

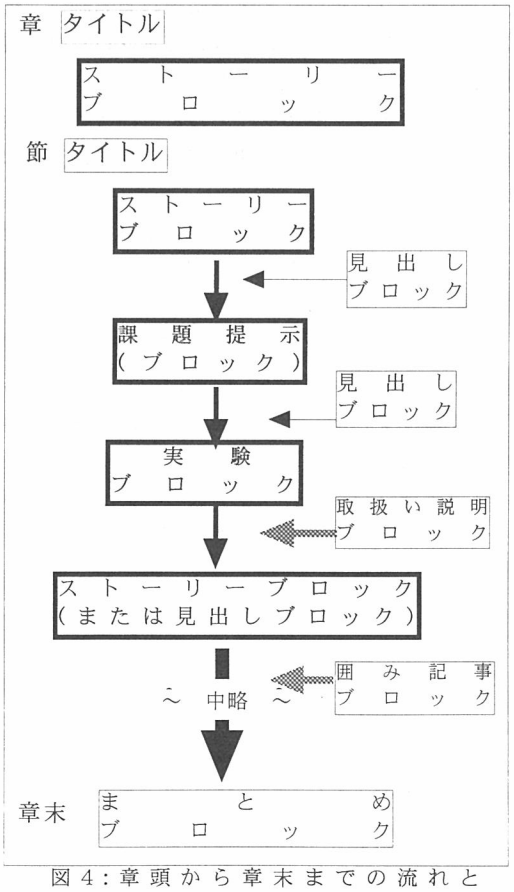

図４：章頭から章末までの流れと節内の主流ブロックの出現モデル

3.3.ブロック内の文章

教科にもよるが，ブロック内の文は，理科，社会，生活では，事象や事物の状況説明，概念や事物の呼び方とその理由，活動目的と活動方法の説明，因果関係の説明，作業手順，学習指示(要点指示) といった表現が多い．

算数は，学習指示(命令)，状況・場面説明，仮定と帰結，概念や事物の呼び方を説明する表現が多い．

国語は，題材となる文章そのものを各単元のブロックと考えると，基本的には算数と同様である．

一方，教科書における文章の文型を計量的に調査した結果から，教科毎に使用されている文型が異なることがわかっている[15]．

また，ブロックは，学習指示を意図する表現の流れと関係が深く，学習活動の流れを形成する要素となっている[11]．

これらのことから，ブロック内に見られる文章は，ブロック自体の意味や機能により異なった特性が見られると仮定できる．

通常，談話構造に関する研究では「主題」「指示詞」「接続詞」や語彙の関連性に着目して構造化されることが多い[16]．本稿でも基本的には「主題」等の語彙の関連性に着目する．しかし，本稿の目的は，教科学習で扱われる学習内容と，教科書の最小談話単位であるブロック内の構造との関係に，「構造」を規定している要因を見い出すことである．また，そのため，文脈を把握する指標となるなんらかの要素の出現位置を明らかにすることが重要である．したがって，通常の談話分析で行われるように段落単位の結束性のみを言及しない．つまり，ブロック間の結束性の考察には，単語，文，ブロックの出現パターンのみを扱い，談話内の文の単語配列の形状に，共通部分を探すという方法を試みる．本の階層構造と文章とを位置により関係づけ，その文章の並びを，単語を指標に明らかにするためにこの方法を採る．

4.XMLによる文データの解析

前節で述べた教科書の構造をXMLで表現し，文書構造の中に埋め込まれたブロックを探索して

そのブロック内の文を構成する単語の配列順序を抽出する．XMLの構造検索はXSLT (eXtensible Stylesheet Language Transformations)[17]による．教科書のブロック内テキストのXML化タグ付けの手順は以下の通りである．

(1) 章(section)や節(subsection)などの，文書の構成単位の区別のためのXMLタグを付加したテキストデータ内の，テキスト部分にchasen2.2.8をかける．

(2) 形態素解析処理の結果，文(sentence)中の各単語に，「単語」として構造を与えるXMLタグ<item>を付ける．

(3) 同時に，解析過程で得られる単語の品詞情報を取出し，各単語の下位の位置にXMLタグ<parspch>のデータとして付加する．

以上の手順を経て，文中の単語の位置を文章構造に基づいた形で定義する．例を図4に示す．

このような形式でXML文書化したテキストに対して，XSLTで名詞(名詞-一般)，動詞(動詞-自立)を検索し，その配列を調べる．品詞はchasen

図3:文中の単語毎のタグ

で利用するIPA辞書に準拠するものである．

5.XML構造検索結果

名詞と動詞の検索により，それらの単語の配列，章節構造との関係を調べた結果，(1)名詞が概念理解への視点の移動を示す道標となっていること，(2)名詞，動詞の出現分布が学習概念の性質を説明する文脈であるか，学習概念の意味を説明する文脈であるかに関係があること，(3)章節構造とブロックの関係が単語からも明らかになること，(4)文脈を主に形成する文脈判断の指標ブロックとそうでないものがあること，が明らかになった．以下，順に述べる．

(1)名詞の並びについて

各ブロックの末尾にくる名詞は次のブロックの先頭にも現れる．特に，節のタイトルや「課題提示ブロック」の前にある「ストーリーブロック」や「見出しブロック」の末尾に現れる名詞は，その次に続く節のタイトルや課題提示ブロックに現れる．しかし，その他の「見出しブロック」や「実験ブロック」等のタイトルと，前後のブロック間では，名詞どうしが関連することはあまりない．

このことから，文脈の流れは，章節構造と「課題ブロック」，「ストーリーブロック」，「見出しブロック」で形成されていると考えられる．

また，図5に名詞の配列に見られる関連性の例を示す．図5のように，タイトルから「ストーリーブロック」へ，そこからさらに，節タイトルへ，そして，次の「ストーリーブロック」からさらに「課題ブロック」に至るまでの構造上の流れと名詞(名詞-一般)の出現の仕方を関連させてみていくと，タイトルの概念「物，燃え方，空気」を具体化した単語「薪，煙，並べ方，火」を「ストーリーブロック」で展開し，その後，節でその一部を抽象化し，また，節内の「ストーリーブロック」で具体化している．そして，それら具体化した際に利用した題材（単語）を利用して，さらに，考察すべき抽象性の高い題材をあげ，その考察から，抽象的な概念の思考を促すといった流れで，物事を見る視点を徐々に考察対象である抽象的な事象に結び付ける視点の移動が促されてい

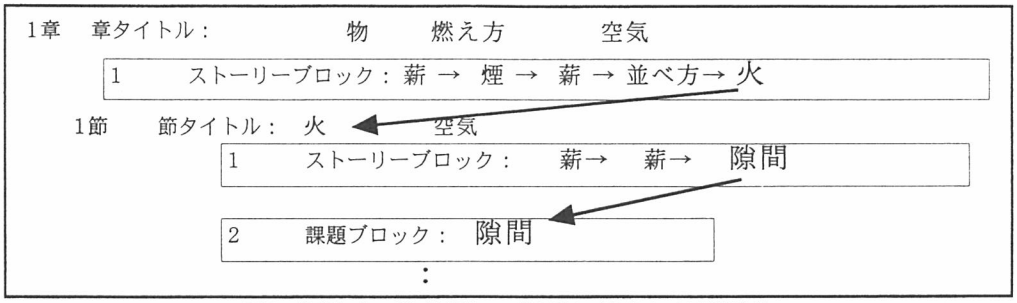

図5:名詞（一般名詞）の配列に見られる関連性

る.

以上，名詞の並びからわかったことは以下の4点にまとめられる.

①名詞の並びとブロックの流れには強い継続性がある.

②章頭の名詞は具体化と抽象化の間で視点を移動させる筋道に利用されている.

③具体化～抽象化という視点移動に利用される一連の名詞は，類義語ではないが，１つの概念を理解するための語彙としてグループ化できる.

④ブロックには主として文脈を形成するものと情報を与えることに徹する副次的なものとに分かれる.

(2)名詞と動詞の関係

上であげたような「名詞」の並びとブロックや章節間との関連性は，学年の別，上下の別なく，本としての教科書の1章めから2，3章めで顕著である．また，章内においても，同様に，1，2節めで顕著になっている.

しかし，名詞で新概念へと意味理解を誘導する筋道がつけられているのは，最初の章や節だけで，章や節が進むに連れて，身近な話から徐々に抽象度の高い単語に視点を移すといった視点の誘導は少なくなっていく.

そして，各章の中程や，教科書の本自体の中程になると，同じ単語の出現が非常に高い節が出現している．そのような節では，専門用語の多用と，動詞，形容詞の多様化が進む.

動詞においては，視点の誘導等，文脈との関係は顕著には見られない．行動か現象か，つまり，意思性の動詞か，状態性の動詞かの違いがあるのみである．意思性のある動詞か否かで出現配列に特徴はなかった．また，ブロックの機能によっても，特に顕著な違いは見られなかった．全体的に見て，意思性の動作動詞と無意思性の状態性の動詞はほぼ均等に交互に現れており，その現れ方に特徴的なパターンはない.

しかし，動詞が使用されている数には差が見られる．動詞がほとんど使用されていない節と多様な動詞が数多く使用されている節とに分かれる．動詞が数多く使用されている節では，同じ名詞の出現が多くなる．専門用語の多用と呼応している.

以上の数量的な増減について，まとめると，抽象性の高い概念がトピックになれば，名詞は同じ単語が頻繁に出現するようになる．また，その分，動詞や形容詞の利用が多くなり，それらの情報量が高くなる．名詞が様々出てきて視点を次々に移動させていく場合は，同じ動詞が頻繁に使われるようになる．したがって，視点の移動は名詞により誘導されると考えられる.

このように，動詞や形容詞が多様される場面で同じ名詞が何度も出現するのは，その単語をトピックとしてそのトピックとなった事物の性質を理解するための説明や作業が多くなるためであると考えられる.

6.文脈の指標

名詞のみを抜き出してみると，同じ単語のみが並ぶ箇所があった．１つの単語しか表れないということから，その単語の重要性は推測できる．しかし，数量的な多さだけでは，どのように重要であるのかは不明である．出現回数が多いという単

語を見るだけでは単語の使用状況は把握し切れない．このことから，視点の移動の方向性がわかるものが文脈の流れを掴むのに必要であることが指摘できる．

また，文脈の流れは，章節構造と課題ブロック，ストーリーブロック，見出しブロックで形成されている．

7.文の構造化と形態素の利用について

元の情報を損なわず，1つのデータから文章の構造情報と，文単位の形態素情報を引き出すことができるため，読解時の視点の移動をそのままに再構成することができる．

また，教科書全体の構造において，文章の並びに対する形状分析を目的としていたが，そのために順序の体系を壊すことなく，データを記述できるXMlは，たいへん目的に適ったデータ形式であると言える．

8.まとめ

小学校理科教科書をXMLを用いて表現し，階層化された教科書において，文章を構成する文と文の関係を，単語(名詞，動詞)を指標に，文頭からの出現順序を調べることにより，文章構造のパターンをさぐった．

そして，単語の出現パターンと前後の文配列との比較した結果，(1)名詞が概念理解への視点の移動を示す道標となっていること，(2)名詞，動詞の出現分布が学習概念の性質を説明する文脈であるか，学習概念の意味を説明する文脈であるかと関係があること，(3)章節構造とブロックの関係が単語からも明らかになること，(4)文脈を主に形成する文脈判断の指標ブロックとそうでないものがあること，が明らかになった．

今回は，単語単位に形態素解析処理の過程で得られたタグを利用してXML化し，階層化したデータを用いた．この検索でも，ブロック毎の文の出現配列にパターンが見出せることから，テキスト中の文脈を読解する指標を一般化するための手段として用いることが可能であると考えられる．

今回は，文の複雑さ等を見るのではなく，意味的な関係性と文書構造で文脈の指標を調べるものであることから，助詞等の機能語については，検索しなかった．

また，理科の教科書にのみ適応させ，そのブロックとブロック内の配列パターン検索により，教科書毎の特性を明らかにしたが，さらに，他教科のテキストにおいても，配列検索によるパターン化を行い，教科書のテキストにおける配列の一般化を行っていきたいと考えている．

参考文献

[1]外国人子女の日本語指導に関する調査研究協力者会議，“外国人子女の日本語指導に関する調査研究＜最終報告書＞，”東京外国語大学, Jun. 1998.

[2]工藤真由美,"児童生徒に対する日本語教育のための基本語彙調査",1996,横浜国立大学教育学部

[3]白鳥智美，“児童生徒に対する日本語教育のための語彙調査-社会科教科書の語彙-，”日本語教育学会平成１２年度春期大会予稿集, Mar. 2000.

[4]工藤真由美編，“算数・数学教科書の日本語の考察-日本語教育の観点から-”，1997, .横浜国立大学教育学部

[5]日本語教育学会編,"読解指導"日本語教育事典pp.604-605,1982,大修館

[6]松井嘉和,"読解教育"94号,『日本語教育』,1997.10,日本語教育学会

[7]小林葉月,"小学校における語彙教育と単語指導",『国文学解釈と鑑賞』,第64巻1号,1999,至文社

[8]山田みな子,"読解過程に見られる既有知識の影響と文法能力の関係について",『日本語教育』86号,1995.7,日本語教育学会

[9]林部英雄,雨宮朋子,"言語機能が文の"自然さ"の判断に与える影響-発達的観点からの実験的検討-",横浜国立大学教育人間科学部紀要Ⅰ教育科学, 1991.10, 横浜国立大学教育人間科学部

[10]林部英雄,日高聡子,牧野泰美,小野博,"言語に内在する論理の発達に関する研究-いわゆる「含意」の習得を中心として-",横浜国立大学教育人間科学部紀要Ⅰ教育科学,1989.10,横浜国立大学教育学部

[11]中尾桂子,森下淳也,"日本語教育のための表現意図出現パターン調査における文書データベースとＸＭＬの活用",2000.12,情報処理学会人文科学とコンピュータシンポジウム

[12]大木道則他33名"新訂理科3-6年",1996,新興出版社啓林館

[13]http://www.w3c.org/XML/

[14]http://chasen.aist-nara.ac.jp/index.html

[15]中尾桂子,"小学校検定教科書の構文調査 -外国人児童の教科学習支援のための基礎研究-,",論文集7,1999,,小出記念日本語教育研究会

[16]泉子・K・メイナード,"談話の分析の可能性",1997,くろしお出版

[17]http://www.w3.org/TR/xslt.html

[18]http://www.alphaworks.ibm.com/tech/LotusXSL/