自然言語処理 67－6
（１９８８．７．２２）

日本語文の複雑さの定性的・定量的特徴抽出

石崎　俊　　井佐原　均

電子技術総合研究所

日本語文の複雑さを客観的に評価するために小学校低学年・小学校高学年・中学・高校のそれぞれにおける４教科（国語・数学・理科・社会）の教科書（２６種）の文章を対象に解析を行った。ここでは、文の複雑さの定量化、文の複雑さや文体の定性的特徴の抽出、学年及び教科ごとの標準的例文の収集を行った。

文の定量的特徴としては、文の長さ、用言等の数、係り受けの数及び並列構造の数等を調査した。定性的特徴としては、多義性、形態素の曖昧性、文体の特徴、並列表現、省略、照応を解析した。各教科書から無作為に抽出した１００文を定量的に解析した結果に基づいて、標準的な例文を５文ずつ選択した。

ここで抽出した多くの例文は、現在及び近い将来の機械翻訳システムの技術レベルを客観的に評価するための評価基準として、役立つことが期待される。

Extraction of Qualitative and Quantitative characteristics of Complexity included in Japanese Sentences.

Shun Ishizaki and Hitoshi Isahara

Electrotechnical Laboratory,
1-1-4, Umezono, Tsukuba, Ibaraki, 305 Japan

The purpose of this research is to extract objectively the characteristics of complexity included in Japanese sentences. About 2700 sentences are sampled from 26 kinds of textbooks which include elementary school(second grade and fifth grade), junior highschool, and highschool textbooks. The subject of the textbooks consists of language, mathematics, science, and social studies.

They are analyzed into qualitative and quantitative characteristics of Japanese sentence complexity and sentence style characteristics. The quantitative characteristics include length of a sentence, number of verbs, adjectives and adjective verbs, number of modifying phrases and number of parallel structures.

The qualitative features include multiplicity of meaning, morphological ambiquities, sentence style features, parallel structures, ellipsis and anaphora. Most standard sentences are extracted from the textbooks(5 sentences each) and those of 16 textbooks are listed in this monuscript.

These data obtained here will be useful for evaluating machine translating systems(current and future ones).

1　はじめに

我が国では、コンピュータ関連会社を初めとして多くの企業で機械翻訳の研究が進められている。特に最近では、幾つかの機械翻訳システムが市販されており、機械翻訳の実用化の時代に入ったということができよう。

しかし、コンピュータが自動的に文章を翻訳することは、小説などの芸術作品は言うまでもなく、我々が日常目にする新聞などの自然言語の文章についても、まだ難しい段階にある。現在の機械翻訳システムでは、人間が前処理や後処理を行って、システムに足りない面を補っているわけで、むしろ人間の翻訳を機械が支援する段階と言っても良いかも知れない。

機械翻訳システムが対象とする例文には、次のような多くの種類の文体が考えられる。

（１）教科書
（２）新聞
（３）特許、法律、ＪＩＳ，ＩＳＯ
（４）論文、論文抄録、解説文
（５）マニュアル、カタログ、広報・ＰＲ
（６）手紙
（７）契約書
（８）対談、話し言葉、演説
（９）小説、随筆、詩歌
（１０）その他

また、文体のほかに、文章の内容の専門分野についても、情報処理、医学、数学、物理学、政治、経済、法律など多くの分野がある。機械翻訳システムはそれらの分野の専門用語を登録した辞書を前もって用意する必要がある。さらに、文体と分野を決めた後でも、翻訳の難易度による例文の分類が考えられる。

このような観点を考慮し、現在および近い将来の機械翻訳システムの技術レベルを評価するためにふさわしい対象として、小学校から高校までの教科書を対象とすることにした。

本研究では、表１に示す２６冊の教科書（小学校、中学、高校の国語、数学、理科、社会）に現われる文章を対象とし、日本語文章の複雑さの解析および例文抽出を昭和６０年度から６２年度までの３年間にわたって（社）日本電子工業振興協会機械翻訳調査専門委員会例文評価ワーキンググループで行ったものである。

この研究の観点は三つある。一つは、文の複雑さを定量的に表すためのものである。具体的には、大きく分けて次の項目が対応する。

（１）文の長さ
（２）用言等の数
（３）係り受けの数および並列構造の数

これらの量は、定義することが比較的容易であり、多くの例文を対象として解析、定量化が可能なものある。また、小学校から高校までの生徒が学ぶ文の複雑さの変化と、国語、社会、数学、理科という学科別の特徴がわかりやすく比較できる可能性がある。

第二点は、文の複雑さや文体の特徴を定性的に抽出するもので、次の項目が対応する。

（１）多義性
（２）形態素の曖昧性
（３）文体の特徴
（４）並列表現
（５）省略
（６）照応

これら諸項目は、現在の自然言語処理技術でネックになっているものであり、したがって、機械翻訳システムでも難しい問題である。それらの課題が、小、中、高と学年が上がるに従ってどのように変化し、学科別にどの様な特徴があるかを調査する。

第三点は、例文の収集である。各教科書から無作為に１００文ずつ抽出し、さらにその中から、その教科書の特徴を、上記の定量的な観点から代表している標準的な文を５文ずつ選び整理した。

2　定量的解析

各教科書からほぼ無作為に選んだ１００文について、文の複雑さを表す定量的なデータを収集した。主要な項目について、データを学年別・教科別にグラフ化したものを図１に示す。以下では、各項目について、順次検討を加える。

１）文字数

各文について、句読点・空白・記号も含むすべての文字を数える。

表１　教科書リスト

科目	学年	教科書名	出版社
国語	小(低)	こくご二上たんぽぽ	光村図書
	小(高)	国語五上銀河	光村図書
	中	改訂新しい国語三	東京書籍
	高	高等学校国語Ⅱ	三省堂
数学	小(低)	新版たのしいさんすう２年上	大日本図書
	小(高)	新版たのしい算数５年上	大日本図書
	中	改訂中学校数学３	大日本図書
	高	高等学校数学Ⅱ[改訂版]	旺文社
	高	高等学校数学確率・統計	旺文社
理科	小(低)	新訂小学理科２	教育出版
	小(高)	新訂小学理科６(下)	教育出版
	中	改訂理科１分野上	新興出版社啓林館
	中	改訂理科２分野上・下	新興出版社啓林館
	高	高等学校物理改訂版	学校図書
	高	高等学校化学改訂版	新興出版社啓林館
	高	高等学校生物改訂版	新興出版社啓林館
	高	改訂版高等学校地学	数研出版
	高	改訂精選理科Ⅰ[力学・地球編]	東京書籍
社会	小(低)	しょうがくせいのしゃかいかはたらく人２	中教出版
	小(高)	小学生の社会科 国民生活と生産５上	中教出版
	中	中学校社会地理的分野	学校図書
	中	中学校社会公民的分野	学校図書
	高	詳説世界史改訂版	山川出版
	高	新版倫理	教育出版
	高	改訂政治・経済	東京書籍
	高	新訂現代社会	教育出版

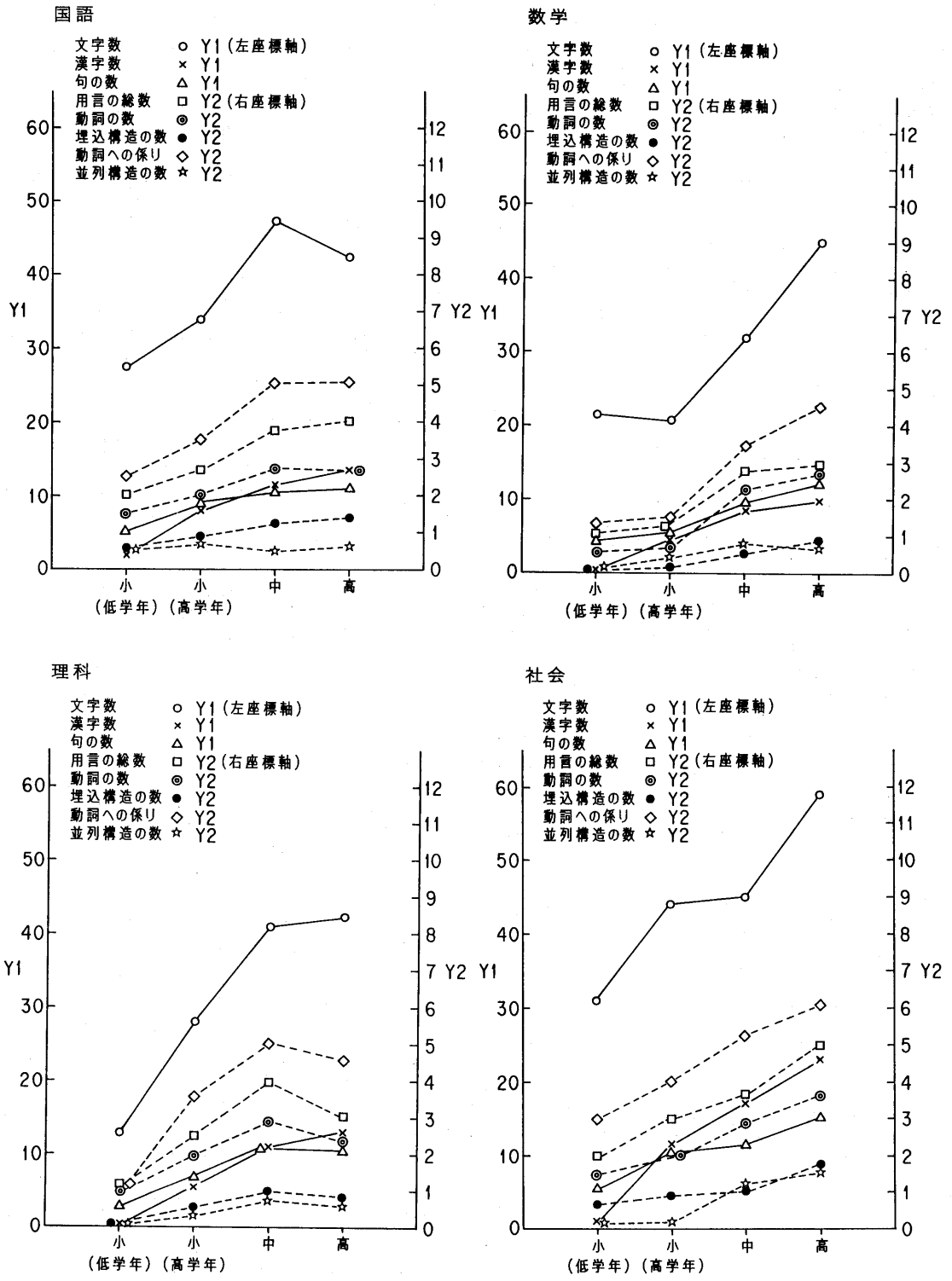

図1　教科書中の文の複雑さに関する定量的評価（学年別、学科別表示）

各教科とも、学年が進むにつれて、文字数が増加する傾向にある。中学国語の文字数が高校よりも多いのは、このデータについてのみ言えることだが、教科書全体からランダムに選んだものではないために、教材とされた特定の作家の文章の特徴が出ているものと思われる。教科別に見ると、社会の文字数が突出している。

2）漢字数

各教科とも、学年が進むにつれて、漢字数が増加する傾向が見られるが、中学と高校においては、あまり差が見られない。図には現われていないが、高校における教科内部での科目による差異が顕著になる。なお、社会は文字数と同様に、漢字数でも他を引き離している。なお、文字数、漢字数が学年が進むにつれて共に増加することから、もしすべてを平仮名表記にするならば、その増加はさらに顕著になると思われる。

3）句の数

句の数の数え方は概略次の規則による。

名詞と助詞で一つの句
用言は助動詞も含めて一つの句
接続詞・副詞等は一つの句
連体詞は一つの句
名詞＋断定の助動詞は二つの句

句の数は、学年が進むにつれて増加する傾向があるが、理科の中学と高校には差がなく、社会の小学校高学年と中学にも差がない。また、社会を除いては教科による差はほとんどない。社会では、小学校高学年で他の教科の中学高校と肩を並べる程度にまで句の数が増大する。

4）用言等の総数

ここで数えたものは、動詞・形容詞・形容動詞・断定の助動詞である。文末の体言止め・連用中止の数も含める。

用言等の総数は、学年が進むにつれて増加する傾向があるが、中学と高校にはあまり差がなく、特に理科においては逆転している。全般に社会のデータの値が大きい傾向は、用言の総数にもあてはまる。

5）動詞の数

動詞（接合動詞を含む）の数は用言の総数とほぼ同様の傾向を持つ。

6）埋込構造の数

ここでは、用言の連体修飾を埋込構造とする。従って、「赤い本」も埋込構造となる。

国語と社会においては、学年が進むにつれて増加する傾向が現われている。数学も同様であるが、数が少し減少する。理科は中学においてもっともその数が増える。このように中学がピークとなるような特徴は、理科において見られることが多いが、この特徴にどのような意味があるかは、さらに検討する必要がある。

名詞への埋込構造の数、形式名詞への埋込構造の数、および準体助詞への埋込構造の数について、それぞれ調べてみると、全体としての特徴は埋込構造の数とほぼ同様であるが、数量的には、名詞、形式名詞、準体助詞の順に少なくなっていく。

7）動詞への係り

ここで係り（受け関係）とは、前から後への係りを考える。従って、埋込構造は名詞への係りと考えて、動詞への係りには含めない。

国語、数学、社会においては、学年が進むにつれて、係り受けの総数は増加する。理科においては、中学と高校では顕著な差はない。いずれも、教科内部の科目による差はない。数量的には社会が幾分多い。

動詞への係りの数も、係り受けの総数とほぼ同様の特徴を持つが、理科において中学と高校とでは逆転している。

8）並列構造の数

並列構造は、連用中止・体言止め・読点・「・」・接続詞・接続助詞等で複数の構造（単語、句、文）を並列にしたものとする。

並列構造の数は、国語を除いて、小学校から中学にかけて増加し、中学高校では、教科内の科目によるばらつきが多い。国語には特徴はない。

3　定性的データ解析

2章で示した定量的なデータの解析で用いた同じ例文から下記のような項目について、文の複雑さを定性的に解析した。

1）多義語の定性的評価

（1）学年による違い

今回調査した教科書の中でも小学校と中学校・高校とではかなり違っている。これは主に表記の違いから生じるものである。小学校では一般的な文書において漢字表記される語もその教育段階に応じてひらがな表記されており、漢字表記であれば問題にならない語が同音語であるために複数発生することから、多義になってしまうものもある。

例）しき（式、指揮、四季）をかい（書い、欠い）てみよう。
よ（夜、世）があけ（明け、開け）ました。

中学校・高校においてはひらがな表記の同音語は小学校に比べるとかなり少なくなっており、文中の漢字の比率は一般の文書と同じようになってきている。ひらがな表記になりやすいのは動詞複合語の後置の語である。しかし、後置の語には普通、補助的な意味の和語動詞が連なり、前置の語によって多義を解消できるのであまり問題にならない。しかし、単独の動詞であっても、ひらがな表記の語もある。名詞のひらがな表記はほとんどなくなっている。

（2）教科による違い

教科の違いは一般的に言われる分野の違いと同じである。物理、数学などの理数系と社会科系統のものとでは同表記であっても意味が違ってくる。

例）くらい（数学では桁、社会では地位）をそろえてたてにかく。

社会科の教科書は理数系の教科書に比べ、言い回しが難しく、そのために多義になる場合がある。特に比喩的な表現に多く見られる。

例）この地方出身のエラスムスは、１６世紀最大の人文主義者として国際的に活躍し、「愚神礼賛」で協会の腐敗を攻撃した（攻めた、非難した）。

2）形態素解析上の曖昧性の定性的評価

（1）補助動詞

「動詞＋て＋動詞」「動詞＋動詞」という形式で、動詞が他の動詞の後に付けて用いられることがある。そして、後の動詞が前の動詞にある一定の意味を付け加える場合、後の動詞を補助動詞という。後の動詞が本来の意味を保っている場合は、補助動詞とは言わない。補助

動詞は、この両方の場合に解釈されるため、以下のように曖昧性を生じるものが多い。

例）～ている　←→　～て；居る／て・要る
～てみる　←→　～て；見る
～ていく　←→　～て；行く
～てくる　←→　～て；来る

（2）多品詞語

ある単語が複数の品詞に属する場合、解析の際に曖昧性を生じる。たとえば、「ある」のように動詞と連体詞という解釈ができるような場合、形態素解析は曖昧性を生じる。

例）わたしの　ふねは、ながさ　千メートルもある
あみを　海に　入れます

（3）助詞

ひらがなの単語の一部が助詞として解釈されることによって曖昧性が生じる。

例）接続詞「そこで」「それで」は、代名詞「そこ」「それ」に助詞「で」がついたものとの間に曖昧性を生じる。

3）文体の特徴の定性的評価

（1）学年による違い

小学校では、「ですます調」が基調になっており、特に、小学校低学年では、命令文に「～してみましょう」、疑問文に「～かな（？）」といった話し言葉のような形で書かれていることが多い。また、ひらがな表記が中心で、空白による分かち書きがなされている。中学校、高校になると、「である調」が普通になるが、命令文は依然として「～してみよう」といった婉曲の表現が用いられている。

（2）教科による違い

国語の教科書では、サンプル文のジャンルや書き手によって文体はかなり異なる。そのため小学校の教科書では本文以外の説明の部分の特徴を抽出している。

数学では、「小学校５年生用算数」で見られるように、問題などを与える命令文で「～しなさい」といった、はっきりした命令口調が現れる。

理科では、「小学校６年生用理科」で見られるように、実験の手順などで「～る」といった終止形で命令を表現する例が現れている。

社会では、いわゆる教科書的な堅い表現が多く現れる。

4）並列表現の定性的評価［1］［2］［3］

並列表現に関しては、学年の違いは文字数や文節数などの文章そのものの違いに起因する違いであって、特に学年による違いは見受けられない。

（1）並列表現の構文

①文の並列

文の並列とは、並列要素が述語を含む形式になっているもので、次のようなものがある。

例）「お米を食べるのがへり、副食をたくさんとるようになったのは、栄養のことを考えるようになったからだと思う。」（小５社会）
「わたしは、まい朝　六じに　おきて、いちばへ魚を　しいれに　いきます。」（小２社会）

②名詞句の並列

名詞句の並列とは並列要素が名詞句のもので、並列接続語としては「、」「・」「と」「や」「か」「および」などがある。

例）「わたしたちの　まい日の　くらしと、はたらく人たちの　しごとは、いろいろな　つながりがあります。」（小２社会）

③非名詞句未完形式の並列

非名詞句未完形式の並列とは、述部を含む文末側部分単語列を文形式から取り去ったものを並列要素とするもので、次のようなものがある。

例）「しかし、鉛筆１本のねだんは１００円で、消しゴム１個が５０円というようにちがった使用価値を持つ商品でも、共通の単位で、そのねうちの大小を表すことができる。」（中学公民）
「また、横軸に体積を、縦軸に重さをとり、その関係をグラフにする。」（中学理科）

（2）並列表現における構文の型の割合

国語は他の教科に比べて文の並列の割合が多いことは顕著である。その中でも小２の文は「～て～」の形が大半を占めている。小６になると「～て～」と連用中止の形の並列構造が同じくらいの割合で現れ、それでほとんどを占めている。中学・高校になると文と文の間に「しかも」や「また」、「だけでなく」などの語が挿入されている並列構造の方が多くなる。これは学年が上がると共に文章が複雑になり、また表現法の違いによる微妙なニュアンスを理解できるようになることによるものと思われる。

数学、理科、社会では、ものを数え上げることが多いので、名詞句を並列することが多い。その名詞句も一つの名詞からなるものがほとんどである。

理科や数学は社会よりも文の並列の割合が少ない。その理由として一文が比較的短いこと、言い回しがある程度決まっていることなどが考えられる。また、図や表などにまとめることが多いことも関係あると思われる。

社会における文の並列の割合は、国語と数学・理科との中間に位置する。社会科は文章化して解説し、また、社会の教科書に見られる図や表は統計的資料になるグラフや地図のようなものがほとんどである。使用されている型は連用中止のものが多く、また、「～たり～」のものも多い。前者についてはこの型が簡潔に文をまとめること、後者については数え上げの際、他にもまだ可能性があることを暗示することの働きがあることから多く使用されていると思われる。

国語の中でも説明文などは「～て～」の型よりも連用中止の方が使用頻度が高い。

数学や理科では非名詞句未完形式の並列の割合が多くなっている。これは「直角をはさむ２辺の長さをａ、ｂ、斜辺の長さをｃとする。」（中学数学）のような表現法が多く使われるためである。

5）省略の定性的評価

ここで、「省略」現象を次に示すように分類する。

（1）文脈的な省略

前後の文脈あるいは一般常識などによって省略されているものが捕えるもの。

1-a）特定化できない必須要素の省略

一般常識やテキスト全体を通して前提となっている背景などから必然的に省略されるもの。

例）〈ー が〉数列を表すのに、２の代わりに、その一般項を用いて｛an｝の形に〈ー が〉表すこともある。

1-b1）「が」格の省略
1-a以外の「が」格の省略

例）〈ひまわりのたねが〉早くめを出すといいな。

1-b2）「を」格の省略
1-a以外の「を」格の省略

例）そして、〈ー が〉作文を　かくときに　〈作文ノートを〉やく立てましょう。

1-b3）「に」格の省略
1-a以外の「に」格の省略

例）せっけんも〈水に〉とけるでしょうか。

1-b4）上記以外の必須要素の省略

例）〈いまは〉朝九時です。

（２）構文的な省略
前後の文脈あるいは一般常識などを使わなくても構文的な関係から省略されているものが捕えるもの。
2-a）提題化に伴うもの

例）しかし、当時の後進国ドイツは、先進工業国であるイギリスからの商品の流入をくいとめるため、〈ドイツは〉保護貿易政策をとった。

2-b）並立表現に伴うもの

例）〈ー が〉電流を流したり、〈電流を〉切ったりしてみる。

2-c）主文、従属文の関係によるもの［４］

例）〈明美が〉かしわもちの葉っぱをはがしながら、明美が（「わたし、かたみがせまくて」）といった。

（３）定性的評価
①全体的な傾向
全体として主格の省略が多く、「を」格や「に」格やその他の必須要素の省略はあまりみられない。
「が」、「を」、「に」以外の省略は小学校５年生用国語と小学校２年生用社会にそれぞれ１例づつみられるだけである。前者は、話法にともなうもの、後者は「〈いまは〉、朝九時です。」で、ともにレトリカルな省略である。
②学科による傾向
数学や理科などで、1-aの省略が多く見られる。数学や理科などでは計算や実験などの手続や説明を示す文が多いためであろう。数学や理科などでは、問題文などをのぞき動作主体は普通省略される。また、数学や理科などに現れる手続き文の多くは、一種の命令文と考えることもできる。
国語の場合は、抽出された題材に依存し、随筆文では1-aや1-b1（筆者主語）の省略が多くなっている。物語や小説でも、主人公を中心に文章が進む場合は、主格（私が／は）の省略が多く現れている。
数学で、「を」格の省略が多くなっているが、そのほとんどは、特定の語（例えば、「けいさんを」）の省略である。
国語や社会で、他の教科に比べ構文的な省略が比較的多くなっているのは、重文が多く使われているためと思われる。
政治経済で2-aの省略が多いのは、ある項目を取り立てて説明する文が多いためである。
③学年による傾向
小学校低学年で、意外に前文の文脈を参照しないと省略内容が同定できない省略がみられる。高学年になると、前文の文脈を参照する省略は少なくなり、構文的な省略や文内省略が多くなる。高学年では叙述的で長い文が多いが、小学校低学年では単文で、しかも問い掛け的な文が多いためであろう。
小学校低学年の社会で1-b1の省略が多いのは、中学公民や高校政治経済などの叙述的な文章と違い、読み手に親密感を与えるために動作主体が具体化されているためであろう。

６）照応の定性的評価
（１）学年による傾向
一般に学年が増すごとに照応表現の出現回数、種類が多くなる。また、学年が増すごとに照応表現の難易度が増している。
（２）学科による傾向
小学校低学年では、国語以外は照応表現が少ない。
小学校高学年では、国語、社会に照応表現が多い。
中学、高校では、学科ごとの照応表現の出現回数にあまり差はなくなる。
社会、理科では、「このような」「このように」という照応表現が多い。特に社会では、文頭に「これから」「これまで」のような照応表現が多い。
社会、国語では、以前の文脈を指す照応表現がよく出現する。これに対し、数学や理科などでは、計算や実験などの手続や説明を示す文が多いため、このような照応表現は少なく、直前の１文またはその１部を指し示す場合が多い。
（３）照応の種類の傾向
一般に全ての学年、全ての教科において、照応表現の出現回数は、（ア）前出の名詞を指すもの（イ）前出の文を指すもの（ウ）指示するものが明示されておらず、推論の必要なものの順である。
（ウ）の出現回数は、（ア）（イ）に比べて非常に少ない。
「こそあ」の「こ」と「そ」の照応表現がほとんどである。
後方照応はほとんど出現しなかった。

４　例文の抽出
２６冊の教科書からそれぞれ無作為に１００文ずつの例文をまず抽出する。次に、それらの例文の中から、最も標準的な文を次のような手続きで５文ずつ抽出し、各教科書の代表的な文とする。
教科書ごとの定量的な文の複雑さがそれぞれのパラメータの平均値として得られているので、１００文の中から文字数、動詞数などの主要パラメータについて、その平均値に最も近い文を５文選びだす。
表２に４教科４学年計１６冊の教科書における５例文の一覧表を掲げる。

表２　標準的な５例文（小学校低学年・小学校高学年・中学・高校について国語・数学・理科・社会の各教科より一種類ずつ）

こくご二上たんぽぽ
1．つぎの　かん字を　つかって、ことばを　つくりましょう。
2．下の　ひょうを、こえを　出して　はっきりと　よみましょう。
3．こうして、たんぽぽは、たねを　どんどん　太らせるのです。
4．こんな　日には、わた毛が　しめって、おもく　なります。
5．小さい　おうちは、しずかな　おかの　上に　ありました。

国語五上銀河
1．東京から南へ約千キロメートルの所に、西之島新島とよばれる小さな火山島がある。
2．この出来事は、日本じゅうの人々が熱心に見守る中で進行した。
3．夏休みに、物語・伝記・科学読み物など、いろいろな種類の本を読んでみよう。
4．ぬま地にやってくるガンのすがたが、かなたの空に黒く点々と見えだしました。
5．みそやたくあんが長く保存できることはだれでも知っています。

改訂新しい国語三
1．むしろ、コンピューター時代になり数字が氾濫してくると、ますます数感覚が必要となるのである。
2．アジアの子供も、アメリカの子供もアフリカの子供も、同じ数字を学び同じところでつまずいたりしくじったりする。
3．覚えの遅い子供が先生にしかられて泣きべそをかいている光景などを想像すると、ただもううれしくなってしまう。
4．それぞれどんな考えをどのように述べているかをとらえ、意見や主張を述べるときの参考にしましょう。
5．右の例文では、問題提起の部分にそれがはっきり示されているので、そこをよくつかみます。

高等学校国語Ⅱ
1．立派に見える父親、美しく見える母親というものも、今日ではほとんど存在し得なくなった。
2．私は民族派ではないから、今日の若者の気風や風俗を嘆いたりする気持はさらさらない。
3．類義語というのは、その意味や用法がほとんど同じであるか、近似している語群をいう。
4．日本の外にいて日本のことを考えていると、生まれ育った国がさらに美しく、よく思えてくる。
5．私の通った漫才の小屋は京都の場末にあり、そこに出てくる漫才師はあまりうまくなかった。

新版たのしいさんすう２年上
1．ひろしさんの　せきは　まえから　２ばんめです。
2．たかしさんの　くつばこに　○を　つけましょう。
3．本を　かりた　２年生は　みんなで　なん人でしょうか。
4．つぎの　けいさんを　あんざんで　しましょう。
5．しきを　かいて、けいさんの　しかたを　かんがえましょう。

新版たのしい算数５年上
1．整数の計算と少数の計算のしかたを比べましょう。
2．次の計算の答えの見当をつけましょう。
3．このりんご１個の重さは約何ｇでしょう。
4．この食用油０．３ℓの代金は何円でしょう。
5．ある数の少数倍のときも、かけ算が使われます。

改訂中学校数学３
1．式の計算も、数の場合と同様に、次の計算法則をもとにしている。
2．タイルの１辺を１としたときの正方形ＡＢＣＤの面積を求めよう。
3．上の結果から、どんな直角三角形にも、次の性質のあることが推定される。
4．上の定理で、$a^2+b^2=c^2$の代わりに、$BC^2+CA^2=AB^2$と書くこともある。
5．上の例２の角度を、実際に縮図を書いて求めよ。

高等学校数学Ⅱ［改訂版］
1．下の図のように、ジュースのかんを１段、２段、３段、…に積んだものを順に並べてみよう。
2．一般の数列を表すには、文字を使ってa_1、a_2、a_3、…、an、…のような記号が用いられる。
3．数学Ⅰで学んだように、ｍ、ｎが整数であるとき、次の指数法則がなりたっている。
4．このように、その大きさと同時に、向きを考えなければならない場合がある。
5．線分ＡＢについて、ＡからＢへの向きをあわせて考えたものを有向線分という。

新訂小学理科２
1．日なたと日かげをくらべよう。
2．たいようがうごいたからかな。
3．土とすなをくらべてみましょう。
4．ヒマワリのたねとりをした。
5．アブラナのたねをまきましょう。

新訂小学理科６
1．大きなろうそくに火をつけて、燃えている様子を調べてみよう。
2．写真のようなしかけには、かん電池が１個だけ使われている。
3．コイルにかん電池をつないで、方位磁針などを近づける。
4．２回まきのコイルに電流を流して、１回まきのときと比べる。
5．２回まき、３回まきとまき数が増えると、ふれが大きくなる。

改訂理科１分野上
1．どちらも無色の透明な液体であるが、水の分量の方が少ないのにつりあっている。
2．こうしてはかった体積が、計算で求めた結果と一致するかどうか確かめてみよう。
3．しかし、どちらで表しても、同じ分量なら、いつも正確に同じ数値になるだろうか。
4．それで、物体の重い軽いをいうときには、同じ体積にして比べる必要がある。
5．図のようにして、操作１で重さをはかった鉄を、メスシリンダーの水の中に沈めて体積をはかる。

高等学校物理改訂版
1．しかし、現代の私達は、走行中の自動車のエンジンを切っても「クルマは急には止まらない」ことを知っている。
2．それで、力を合成するときには、大きさだけでなく、力の方向・向きも考えなければならない。
3．図３（ｂ）のように、２力Ｆ１とＦ２を合成するときには、平行四辺形の法則を用いる。
4．鉄球Ｂを投げ出した点を原点にとり、水平方向にｘ軸、鉛直下向きにｙ軸をとる。
5．（６）式と（９）式とからｔを消去すると、鉄球Ｂの描く軌道の式が得られる。

しょうがくせいのしゃかいか はたらく人２
1．わたしは　しいれた　魚を、すぐに、みせの　れいとうこに　します。
2．ふねには、私たちが　魚のむれを　さがす　きかいが　ついて　います。
3．〈係りの人が〉小むぎこに　水と　イーストきんを　入れ、きかいで　よく　ねります。
4．このころには、ほかの　えきいんも　しゅっきんして　きます。
5．この　しごとは、大きな　ゆうびんきょくでは、ぜんぶ　きかいが　します。

小学生の社会科 国民生活と生産５上
1．ところで、日本人は、むかしから、いまと同じような食事をとっていたわけではありません。
2．とくに九州地方では、各地に大がかりなみかん園がつくられ、みかんの生産はめざましくふえました。
3．しかし、田植えをする時期が同じになってきたので、一度にたくさんの水が必要となりました。
4．少ない人手で米をつくるためには、地区の人々の協力や共同作業などのくふうが必要といえましょう。
5．このように、いまの社会に見られる工業製品の原料には、おもに鉄や石油が使われています。

中学校社会 公民的分野
1．このように、自分たちの生活に必要な物を、自分たちの生産だけで満たす経済を自給自足の経済という。
2．したがって、労働者が賃金で購入できるのは、かれと家族の生活に必要な消費材、つまり必要生産物である。
3．このような家計の収支をつうじて上の図のように、家庭は、国全体の経済とさまざまな形で結びついている。
4．価格が下がりすぎると、利潤が得られなくなるから、生産をへらしたりやめたりする企業がでてきて、供給がへる。
5．また銀行や大企業が中心になって、ほかの企業の株式を所有して、これを支配するコンツェルンもでてきた。

改訂政治・経済
1．人間が社会生活を営むうえで政治のはたらきは欠かすことはできないが、政治に関する次の事項をそれぞれ４００字以内で説明しよう。
2．また、経済の動きを短期的にとらえ、現在の経済活動が活発であるか停滞しているか、あるいはそのどちらに向かいつつあるかを見ることもできる。
3．戦後のドル本位制は、１９５０年代の末からアメリカの国際収支の慢性的赤字にともなう金流出によって、動揺をつづけることになった。
4．これは、対立する国家または国家群の双方を含めた集団安全保障機構をつくり、いっさいの国際紛争を平和的に解決していこうとするものである。
5．未来の日本は、単に国内の国民福祉にとどまらず、人類福祉を目標とする国際協力を積極的におしすすめなければならない。

5　終わりに

教科書では、学年の違いと教科の違いに関する文の複雑さを解析・抽出した。教科書は、専門家が時間をかけて練り上げた文章ばかりであるため、安心して採用できる。逆にそのような教科書の例文がもし理解できなければ、人間ならば「勉強が足りない」といえるし、機械翻訳システムがうまく翻訳できなければ「システムの性能がいまいちである」といえるであろう。

現在の翻訳システムは、例えばコンピュータマニュアルだとか、輸出製品用解説書、学術論文の抄録など、かなり限定した対象に特化しているため、必ずしも小学校低学年の教科書の翻訳が得意ではないであろう。人間なら常識として知っている小学校程度の言語的・非言語的知識も、現在の機械翻訳システムが良く知っているとは限らない。また、教科書に良く見られる文体も、システムに馴染がないかも知れない。

このように、人間なら小学生でも分かる文章がコンピュータには難しいことは決して驚くべきことではない。これは、機械翻訳に限らず、コンピュータ全体についていえることであり、コンピュータの知能化の研究の進歩が待たれる所以である。

文章の内容をしっかりと理解してから翻訳するという「翻訳の基本」は忘れてはならない大切な指針である。そのためには、長期的な観点から健全な機械翻訳システムの研究開発に取組む必要がある。本報告書の例文集はそのような立場に立つとき、システムの基本的な性能をチェックして、システムの改良や新しい機能を研究するための指針を与えるために役立つことを期待している。

謝辞

本研究は（社）日本電子工業振興協会の機械翻訳用例文評価ワーキンググループでの調査結果に基づいたものである。ワーキンググループのメンバーであった以下の諸氏に感謝します。

小関正彦（日本電気（株））
梶　博行（（株）日立製作所）
加藤安彦（沖電気工業（株））
清野正樹（松下電器産業（株））
重永信一（松下電器産業（株））
鈴木　等（シャープ（株））
西田行輝（三洋電機（株））
西野文人（（株）富士通研究所）
野上宏康（（株）東芝）
藤田　稔（キャノン（株））
望主雅子（（株）リコー）
余田直之（三洋電機（株））
（五十音順敬称略）

また、（社）日本電子工業振興協会の担当者である斎藤、宮川両氏には様々な便宜を図っていただき深謝します。

参考文献

［1］首藤、吉村、津田：「日本語技術文における並列構造、情報処理学会論文誌Vol.27 No.2（1983）
［2］長尾、辻井、田中、石川：「科学技術論文における並列句とその解析」自然言語処理研究会36-4（1983）
［3］田中　章夫：「岩波講座　日本語7　文法Ⅱ　助詞(3)」岩波書店(1977)
［4］南不二男：「現代日本語の構造」pp105-182　大修館書店(1986)
［5］「中学校教科書の語彙調査Ⅰ，Ⅱ」「高校教科書の語彙調査（1），（2）」など　国立国語研究所