

語彙的側面からみた文章の難易度の判定

Measuring the level of difficulty in Japanese Textbooks using the "Vocabulary Level Checker"

川村 よし子

KAWAMURA Yoshiko

東京国際大学 商学部

Tokyo International University

kawamura@tiu.ac.jp

日本語の文章の難易度と語彙の難易度との関係を調べるため、日本語教科書を対象に語彙調査を行った。調査には語彙のレベル判定システム「語彙チェッカー」を用いて、初級から上級までの教科書に含まれる語彙のレベル別含有率を調べた。その結果文章の難易度と語彙の難易度には強い相関があることが明らかとなった。

キーワード： 日本語学習者 難易度判定 読解教材 語彙調査 CALLシステム

はじめに.

文章の難易度を決定する要素は一つではない。語彙・文法・内容等、様々な要素がある。これらの要素が複雑に絡み合っている文章の難易度を決定しているわけだが、例えば、英語の読解能力は読み手の語彙力の影響を強くうけることが Nation (1990) 等で指摘されている。また、山内 (1996) は読解力の 64% を語彙力で説明できている。文章の難易度を決定する要因の中でも語彙の果たしている役割は大きいと言えよう。

本研究では、日本語の文章の難易度を決定する諸要素のうち特に語彙に焦点を当て、語彙の難易度と文章の難易度とがどのような関係にあるのかを調査した。日本語学習者の視点から文章の難易度を判定することを目標に置き、日本語の教科書を分析の対象にした。一般に教科書は学習レベルにあわせて編集されている。初級から上級までの各段階の教科書に、どのようなレベルの語彙がどのくらい含まれているかを分析することによって、教材の難易度と語彙の難易度との関係を調べることができるはずである。

分析には日本語能力試験の 1 級から 4 級までの語彙リストに準拠した語彙のレベル判定システム「語彙チェッカー」⁽¹⁾を用いた。

1. 調査方法

対象とする教科書は、1) 日本で出版された

もの、2) 日本の大学への進学希望者を対象とした日本語教育の現場で用いられているもの、3) 読解教材にあたる部分のあるものを基準に選定した。その結果、今回の調査では次の教科書を用いることにした。

初級：国際交流基金日本語国際センター編

『日本語初歩』

文化外国語専門学校編

『文化初級日本語Ⅰ・Ⅱ』

海外技術者研修協会編

『新日本語の基礎』

名古屋大学総合言語センター編

『A Course in Modern Japanese』

中級：国際交流基金日本語国際センター編

『日本語中級』

東京外国語大学日本語教育センター編

『中級日本語』

名古屋大学総合言語センター編

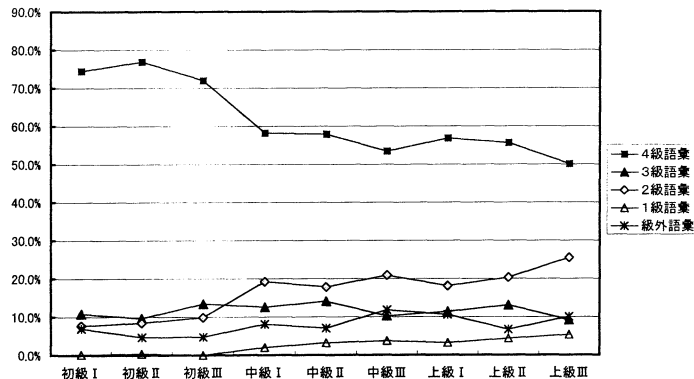
『現代日本語コース中級Ⅰ・Ⅱ』

上級：東京外国語大学付属日本語学校編

『日本語Ⅲ』

データとしては、初級に関しては本文、中級・上級に関しては読解教材部分を取りあげ、各級の始め・中央・終わりの部分を対象とした。但し、級によって対象となる文の長さが大きく異なっているため、各々の該当部分に関して初級は 3 課、中級は 2 課、上級は 1 課分を分析す

図1 レベル別語彙含有率の変化(延べ語彙数)



ることとした。各教材を電子情報化したものを語彙チェッカーにかけて、それぞれの教材に含まれるすべての語彙のレベル判定を行った。

2. 語彙チェッカーの仕組み

分析に用いた「語彙チェッカー」は、教材に含まれるすべての語彙のレベルを日本語能力試験の語彙レベルを規準に判定するシステムである。入力されたテキストに関して、コンピュータが次の作業を自動的に行い、テキストに含まれるすべての語彙のレベルを判定する。

- ①与えられたテキストの形態素解析を行う
- ②各形態素をレベル別語彙リストに照合する
- ③テキスト内の語彙にレベル表示を行う
- ④レベル別の語彙表を作成する
- ⑤語彙のレベル別含有率を算出する

「語彙チェッカー」は入力された文章を「茶筌」⁽²⁾を用いて形態素に分析し、その分析結果を日本語能力試験の1級から4級までの語彙表をもとに作成したレベル別語彙リストと照合してレベル判定を行っている。

3. 語彙チェッカーによる教科書の分析

3.1 テキストの難易度と語彙の難易度

テキストの難易度と語彙の難易度との間に相関があるとすれば、次のような関係があるはずである。

仮説1：易しい語彙の含有率は学習段階が進むにつれて減少する。

仮説2：難しい語彙の含有率は学習段階が進むにつれて増加する。

以上の2点に関して、教科書の本文が実際にその傾向を示すかどうかの検証作業を行った。

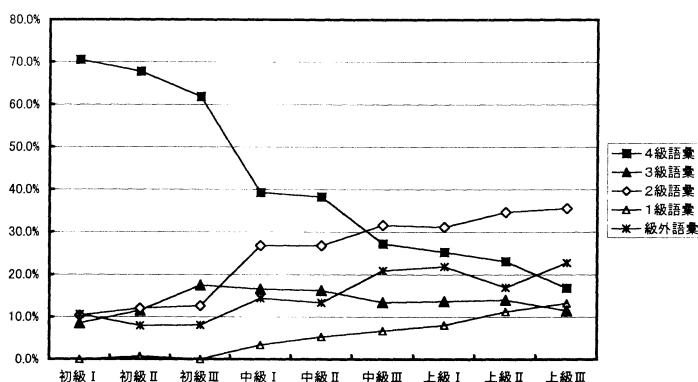
3.2 調査結果

調査対象となった各級の文章は、初級 7,311 字、中級 14,067 字、上級 16,749 字である。形態素解析の結果、総語彙数は、初級 3,068 語、中級 6,547 語、上級 7,776 語であった。調査では、各々の教材を語彙チェッカーに入力し、個々の語彙のレベル判定を行うとともに、各教材に占めるレベルごとの語彙含有率を算出した。この場合、語彙の含有率は延べ語彙数で数えた場合と異なり語彙数の場合とでは異なる。文章の難易度との関連はどちらが強いのかも調査する必要があるため、それぞれ別個に調べることにした。

3.2.1 延べ語彙数による調査

学習段階ごとのレベル別語彙含有率の平均を示したものが図1である。ここでの語彙数は延べ語彙数で算出してある。図1から、次のような傾向を読みとることができる。

図2 レベル別語彙含有率の変化(異なり語彙数)



① 4 級レベルの語彙の占める割合は学習段階が進むにつれて減少傾向にある。

② 3 級レベルの語彙の含有率は初級から上級まで、ほぼ 10%と一定である。

③ 2 級レベルの語彙の含有率は学習段階が進むにつれて増加傾向にある。

④ 1 級レベルの語彙は初級ではほとんど含まれず、学習段階が進むにつれ増加傾向にある。

⑤ いずれの級の教材においても、級外の語彙が多少含まれている。初級では相対的に少ないものの、平均 5%程度あり、中・上級ではいずれも平均 10%程度含まれている。

この結果をもとに、初級 I から上級 III までの各教材について語彙レベルとの相関を調べたところ次のようなことがわかった。4 級レベルの語彙に関して、ピアソンの相関係数は $r = -0.74$ (Pearson correlation coefficient r) ($N=34$) であり、負の相関が高いことが認められた。つまり、4 級レベルの語彙に関しては、仮説 1 の「易しい語彙の含有率は学習段階が進むにつれて減少する」傾向があることが示された。一方、3 級レベルの語彙に関しては相関係数も低く、学習段階との相関関係は認められなかった。

2 級および 1 級レベルの語彙に関する相関係数は各々 0.82 と 0.86 であり、ともに高い正の相関が認められた。以上のことから、仮説 2 の「難しい漢字の含有率は学習段階が進むにつれ

て増加する」傾向にあることも明らかになった。

一方、1 級から 4 級までの語彙リストに含まれない級外語彙の相関係数は 0.21 と、きわめて低い値を示している。級外語彙は、単純に考えれば、1 級以上のレベルの語彙で、難易度の高いものとみなしたくなる。もし難易度が高いのであれば、増加傾向を示さなければ仮説はなりたたなくなる。ところが、今回の分析結果は、級外語彙は、初級の教科書にもある程度含まれていることを示している。なぜこのような現象が起こるのだろうか。

初級に含まれる級外語彙を詳しく調べると、それらはすべて名詞であり、しかもその大半が固有名詞である。一方、日本語能力試験の語彙リストにはわずかな例外を除いて固有名詞は載せられていない。そのため級外語彙と分類されたわけである。ところがこれらの固有名詞は、読み方さえわかれば学習者にとって容易に理解できるものである。しかも、教材の内容によっては⁽³⁾こうした固有名詞の使用頻度が高くなる可能性も大きい。語彙の難易度を決定する際、固有名詞の難易度をどう扱うかは、今後解決すべき重要な課題だと言えよう。

3. 2. 2. 異なり語彙数による調査

同一の資料を、異なり語彙数で分析し直したものが図 2 である。学習段階ごとのレベル別語

含有率の変化は、延べ語彙数で見たときに比べて、より顕著になっていることがわかる。4級レベルの語彙は、初級Ⅰでは平均 70%を占めているが、上級Ⅲでは 17%と大幅に減少している。この変化は4級の語彙リストに含まれる語彙が 800 語にすぎないことによるものである。逆に言えば、そのわずかの語彙が延べ語彙数では上級Ⅲの教材でさえ 50%を占めていたわけである。4級語彙の使用頻度の高さを示している。

3級レベルの語彙の含有率に関しては異なり語彙数で見てもごく緩い増加傾向しか見られない。これに対して、1級・2級の語彙に関しては増加傾向がより顕著になっている。ピアソン相関係数も、1級 0.91、2級 0.89 と、延べ語彙数の場合よりも高い相関を示している。級外語彙に関しても、初級ではほぼ 10%の含有率であるのに対して、上級では約 20%と増加している。相関係数に関しても 0.30 と上昇している。

これらを総合して考えると、教材の難易度と語彙の難易度との関係は、異なり語彙数でみた場合、相関関係がより顕著になると言えよう。仮説 1 の「易しい語彙の含有率は学習段階が進むにつれて減少する」という傾向は4級の語彙においてははっきり現れていた。また、仮説 2 の「難しい漢字の含有率は学習段階が進むにつれて増加する」という傾向は2級および1級レベルの語彙において顕著に現れていた。また、級外語彙の扱いに関しては特に固有名詞に関して考慮する必要があるものの、異なり語彙数で見ると限り、難易度の高い語彙として扱うことが可能であることも示された。

4. 今後の課題

以上の調査結果から、各学習段階の日本語教科書にどのレベルの語彙がどの程度含まれているかが明らかになったが、この結果を一般の日本語文章の難易度の判定に応用するにはいくつか問題が残されている。

まず第一に、語彙の難易度の判定に日本語能力試験の語彙レベルを基準として用いることの

妥当性を検証する必要がある。特に、難易度判定の結果が適正かどうかに関する日本語学習者および日本語母語話者による評価は不可欠である。第二に、分析に使用した「語彙チェッカー」の解析精度の問題がある。現在の解析精度は 98%であり⁽⁴⁾、より正確な判定を行うためには解析精度をさらに高めていかななくてはならない。第三に、学習段階の区切り方の問題がある。今回の調査は学習段階を9段階に区分したが、実際の学習段階はより細かく分かれている。資料の数をさらに増やし、きめ細かな分析を進めることによって、より正確な文章の難易度判定が可能となろう。

注

- (1) 語彙チェッカーに関しては川村(1998,1999)に詳しい。また、語彙チェッカーは CGI 化し、インターネット上で公開している。
(<http://www.tiu.ac.jp/~kawamura>)
- (2) 奈良先端科学技術大学院大学の松本研究室によって開発された形態素解析ツールである。
- (3) 例えば自己紹介や出身国の紹介等は初級のはじめの方に現れることも多く今回対象となった教材でも人名や国名が多く用いられていた。
- (4) 解析の問題は川村(1999)で詳しく分析した。

参考文献

- 川村(1998)「読解のためのレベル判定システムの構築－語彙チェッカーの開発と活用」『日本語教育方法研究会誌』Vol.5, No.2, pp.10-11.
- 川村(1999)「語彙チェッカーを用いた読解テキストの分析」『講座日本語教育』Vol.34, pp.1-22
- 松本・北内・山下・今・今村(1997)日本語形態素解析システム「茶筌」ver 1.0 使用説明書, NAIST Technical Report, NAIST-IS-TR97007.
- Nation(1990) *Teaching and Learning Vocabulary*, New York: Newbury House Publishers, pp.115-145.
- 山内(1996)「学習者の語彙レベル判別テストの開発：語彙力と読解力の関係」『関東甲信越英語教育学会研究紀要』10号, pp.1-13.