自 然 言 語 処 理 153－18
（2003．1．21）

日本語文の規格化

佐藤 理史†, 土屋 雅稔†, 村山 賢洋†, 麻岡 正洋†, 王 晴晴‡

†京都大学大学院情報学研究科知能情報学専攻
‡京都大学工学部電気電子工学科

重要な情報の伝達を目的としたテキストは、読みやすさを保証するようなガイドライン、あるいは、規格にしたがって書かれるべきである。本論文では、3段階、4部門（漢字、語彙、文法、量的複雑さ）からなる日本語文の平易度の規格を提案する。同時に、文がその規格を満たすかどうかを判定するプログラムの実現法について述べる。

Standardization of Japanese Sentences

Satoshi Sato†, Masatoshi Tsuchiya†, Masahiro Murayama†,
Masahiro Asaoka†, QingQing Wang‡

†Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
‡School of Electrical and Electronic Engineering, Kyoto University

The text that informs crucial information should be written under a guideline or standard that guarantees the readability. This paper proposes the standard of Japanese sentences, which defines three-level readability for four factors (Kanji, vocabulary, grammar, and quantitative complexity), and a method that determines whether a sentence satisfies a given level of standard.

1. はじめに

コンピュータや電化製品のマニュアルを読んでいて、「何をいっているのだかさっぱりわからない」、あるいは、「もっと簡単に、普通の言葉を使って書いてくれればわかりやすいのに」と思ったことがある人は少なくないだろう。次の文は、ある携帯電話のマニュアルにおいて、「危険」表示が何を意味するかを説明する文として実際に記述されているものである。

(1) 危険：この表示は、取扱いを誤った場合、「死亡または重傷を負う危険が切迫して生じることが想定される」内容です。

この文を一回ですんなりと理解できる人は、はたしてどれくらいいるだろうか。この文の「危険が切迫して生じることが想定される」という表現が回りくどく、わかりにくい表現であることに、ほとんどの人は同意するのではないかと思う。一方、別のマニュアルでは、同じ「危険」表示を以下のように説明している。

(2) 危険：人が死亡または重傷を負うおそれが高い内容を示しています。

「危険」表示が何を意味するかは、その内容の重要性から、できるだけ多くの人に理解される必要があるだろう。文(2)は、文(1)に比べて、より多くの人が容易に理解できるという点において優れているのは明らかである。逆に、文(1)は、ほとんどの人が理解にいささかの困難を感じるという点で、生命の安全に直結する情報を伝達する文としては不適切である。

この簡単な例からわかることは、情報の伝達を目的としたテキストは、想定する読み手が困難なく理解できるように書かれるべきであるということである。特に、伝達すべき情報が重要であればあるほど、このことは重要になる。

このような想定読者に応じた配慮は、まったくなされていないわけではない。たとえば、幼児向けのおもちゃの注意書きでは、難しい単語や漢字を使わずに説明がなされていたり、漢字にルビが振られていることも多い。製造物責任法(PL法)の施行以降、危険・警告・注意などの表示や説明に、工夫が見られるマニュアルも多くなってきている。

しかしながら、伝えるべき情報をテキストとしてどの

ように表現するかは、結局のところ、書き手の直観に頼っており、守るべき明示的な基準や規格が存在し、それにしたがって書かれることはほどんどない。この状態を比喩的にいうならば、テキストの作成は、家内制手工業のレベルにあり、大量生産される工業製品のレベルには達していないということである。

多くの工業製品には、その製品が機能し、かつ、安全に使えることを最低限保証するための規格や仕様が存在する。これをテキストには当てはめるならば、「テキストが情報伝達という機能を果たすことを最低限保証する規格または仕様」、すなわち、「想定する読み手が理解できることを最低限保証するような規格または仕様」があってしかるべきである。すべてのテキストがこのような規格や仕様にしたがって書かれる必要はないが、そのテキストが非常に重要な情報の伝達を目的としている場合は、このような規格や仕様を明確化し、伝達すべき情報が確実に伝わるように最大限努力すべきである。

このような規格に近いものの一つに、Easy Japanese(やさしい日本語) がある [1]。これは、阪神・淡路大震災のとき、在日外国人に重要な情報がなかなか伝わらなかったという反省に基づき、災害時の情報提供に用いることを想定して作られたもので、日本語能力試験 3・4 級のレベルの日本語に相当する。しかしながら、これに合致したテキストを作成しようとする際に利用できる道具は、日本語能力試験 3・4 級の語彙表と災害時によく使う表現のやさしい表現への言い換え例のリストだけであり、書き手である人間がそれらの道具を参照しながら、やさしい日本語のテキストを作成することを前提としている。

規格や基準は、それを満たすかどうかを判定する客観的あるいは機械的な方法があって初めて意味をなすものである。この点において、Easy Japanese は規格というよりも、より緩やかなガイドラインと呼ぶのが適切であろう。ガイドラインではなく、真に「日本語の規格」と呼ぶにふさわしいものを作成するためには、規格を定義するのと同時に、与えられたテキストがその規格を満たすかどうかを判定する客観的あるいは機械的方法を考案・実現する必要がある。

我々が本年度より開始した**醍醐プロジェクト**は、このような日本語の規格を策定するおそらく最初の試みである。理想的には、規格を満たさないテキスト (規格外テキスト) を規格を満たすようなテキスト (規格内テキスト) に自動的に書き換えるソフトウェアを実現することが望ましいが、現在の自然言語処理の技術レベルでは、これを実用的な高い精度 (たとえば、95%以上) で実現することは、おそらく不可能である。本プロジェクトでは、到達目標を、現在の技術レベルでおそらく実現可能な、次の 2 点に設定している。

目標 1 日本語の規格を作成し、規格外表現を自動的に発見するソフトウェアを実現する。

目標 2 規格外表現に対する代替表現を自動提案するソフトウェアを実現する。

この 2 点の目標を達成することができれば、規格外テキストを規格内テキストに書き換えることを支援するかなり強力なソフトウェアシステムを実現することができる。

本稿では、目標 1 に対する研究の現状について報告する。

2. 日本語規格の基本設計

日本語の規格を作るということは、日本語のあるサブセットを設定するという問題に他ならない。ここで注意しなければならないのは、次の 2 点である。

- 規格を満たすテキストで、伝達すべき情報を十分に表現できること。(**<表現力の観点>**)
- テキストが規格を満たしているかどうかを、十分な精度で機械的に判定する方法が実現可能なこと。(**<判定方法の実現可能性の観点>**)

言語の豊かさ (語彙や文法の豊かさ) とそれによって表現 (伝達) 可能な情報の豊かさの間には、正の相関がある。たとえば、設定するサブセットに「仮定する」という語や「もし〜ならば」という表現が含まれないならば、「あることを仮定して、それについて述べる」ことは不可能になってしまう。一般に、単純な情報の伝達には小さなサブセットで十分であるが、複雑な情報を伝達するためには、より大きなサブセットが必要となる。このため、「適切なサブセットを設定する」ことは非常に難しい問題となる。

一方、判定方法の実現可能性の観点からは、規格が明快であり、現在の自然言語処理の技術レベルで規格判定が実現可能であることが要請される。

これらの点を考慮し、我々は次の 4 つの基本方針を立てた。

方針 1 日本語能力試験「出題基準」[2] を出発点とする

方針 2 3 段階の平易度を設ける

方針 3 文を単位とする規格 (日本語文の規格) とする

方針 4 4 つの部門を設ける

まったく何もないところから規格を作ることは非常に難しい。そこで、我々は、日本語能力試験「出題基準」にその出発点を求めることにした (方針 1)。日本語能力試験は、4 級から 1 級の 4 段階に分かれている。このうち、4 級は語彙や文法がかなり限られており、十分な表現力を有していないと判断したため、3 級、2 級、1 級を出発点とした 3 段階のレベルを設けることとした (方針 2)。それぞれレベルは、次のような情報を記述するのに用いることを想定して策定する。

平易度 3 最もやさしいレベル。生命の安全に直結する情報等、できるだけ多くの人々に最優先で伝達すべき情報 (最優先情報) を記述するのに用いるレベル。

平易度 2 中間レベル。基本的な社会生活を営むのに不可欠な情報 (重要情報) 等を記述するのに用いるレベル。

平易度 1 最上位レベル。その他の情報を記述するのに

用いるレベル。コンピューターの使い方など、比較的複雑な情報を記述するのに用いるレベル。

現在の自然言語処理の技術で安定して利用できる解析技術は、形態素解析と構文解析である。これらは文を単位とした解析技術である。これに対して、2文以上のつながりや構造を把握するための文章解析・文脈解析の技術は、実用レベルには達していない。このため、日本語の規格は、現時点では、文の規格として設定するのが適切であると考えた (方針 3)。

規格にどのような部門を設定するかも、規格設計の重要なポイントである。これは、日本語文のやさしさ・読みやすさがどのような要因によって決定されると捉えるかに依存する。日本語能力試験「出題基準」は、文字・語彙、文法、聴解、読解、の 4 つのテストに対応する形で書かれているが、このうち文の規格に直接関係するのは、文字・語彙に含まれる漢字と語彙、および、文法の 3 つである。この他に、文の量的複雑さ（文の長さ、埋め込みの深さ等）も文の平易度に関係すると考え、漢字 (K)、語彙 (V)、文法 (G)、量的複雑さ (C) の 4 つの部門を設定することとした (方針 4)。

上記の方針に従って、現在、規格の第 1 版と、それに対応した規格判定ソフトウェアを作成中である。以降の節では、これらについて、各部門別に説明する。

なお、以下では、規格の部門・レベル表示には、部門を表すアルファベット 1 文字 (上記の K, V, G, C のいずれか) と数字 (3, 2, 1, 0, 9) を組み合わせたものを用いる。数字の 3, 2, 1 はそれぞれのレベルに対応し、0 は規格外 (平易度 1 よりも難しい)、9 は判定対象外を表す。

3. 漢字の規格とその自動判定

3.1 規　　格

日本語能力試験「出題基準」改訂版（以下、「出題基準」と略記）では、各級で使用してよい漢字を**漢字表**という形で提示している。漢字数は、以下のとおりである。

3 級　284 字
2 級　1023 字（このほかに「付表」に 54 語）
1 級　第 1 水準漢字 1926 字、第 2 水準漢字 114 字

3 級、2 級、および、1 級の第 1 水準漢字は、いずれも常用漢字であり、「『常用漢字表』に挙げられている音訓の範囲内で用いること」との留意事項が付いている。また、1 級の第 2 水準の漢字表は、漢字に、使用できる音訓が付加された形で提示されている。さらに、2 級では、「2 級漢字・付表」として、いわゆる当て字や熟字訓を漢字表記をすることを例外的に許すものが示されている。これに従うと、「眼 (がん)」、「鏡 (かがみ)」はいずれも 1 級であるが、「眼鏡 (めがね)」は「2 級漢字・付表」により 2 級となる。つまり、単に文字として級を判定することはできず、単語を単位に読みを考慮して判定することが必要となる。

もう一つの問題は、固有名詞を表記する漢字である。留意事項に、「固有名詞を表記する漢字は、漢字表に挙げたものに限らない」との記述がある。ある語が固有名詞かどうかを判定することは、技術的にはできないこともない。しかし、精度の点で問題が残るとともに、そのままにしておくよりは、読み仮名を付加する方がよいと思われる。

これらの点を考慮して、漢字の規格を次のように定めた。

(1) 漢字、レベル、許された音訓のリスト、許された単語のリスト、の 4 項目からなる表で、各漢字のレベルを定義する。これを**漢字定義表**と呼ぶ。
(2) 各漢字の具体的なレベルは、「出題基準」に準じる。
(3) 単語に読み仮名をつけた場合は、その単語内では、自由に漢字を使ってよいとする。(読み仮名をつけた単語は、レベル判定の対象外 (K9) と判定する。)

3.2 自動判定の方法

漢字のレベル判定は、文を単語に分割（形態素解析）した後、単語単位に行い、与えられた平易度を満たさない単語を規格外表現として出力する。

各単語に対する漢字のレベル判定は、以下の手順で行なう。

- 漢字を含まない単語の漢字レベルは、対象外 (K9) とする。
- 単語に使われている漢字が漢字定義表で許された範囲（読み、単語）で使われているかどうかを調べ、許されている場合は、その漢字の使用を、漢字定義表で与えられたレベルと判定する。許されていなかった場合は、K0 とする。単語に複数の漢字が使われている場合は、単語の漢字レベルは、使われている漢字のもっとも平易度の低いレベルとする。

形態素解析には Juman[3] を用いる。Juman は、単語に対する「読み（ひらがな表記）」を出力するが、各漢字の「読み」は出力しない。このため、漢字が複数ある場合には、漢字定義表で与えられる読みを組み合わせたものの一つが Juman が出力する読みと一致した場合に、漢字定義表で許された範囲での使用と判定する☆。

4. 語彙の規格とその自動判定

4.1 規　　格

「出題基準」では、各級で使用することが許された単語の集合を語彙表という形で提示している。大筋では、この語彙表をもって、語彙の規格を定義すればよいということになるが、機械的に厳密に判定するためには、いくつかの問題がある。そのうち最も大きな問題は、なにをもって一語（一単語；同一語）とみなすか、という問題である。

文に含まれる単語を認定することは、形態素解析によって行なわれる。たとえば、Juman では、形態素 (単語)

☆ 『常用漢字表』の音訓には、音便化されたもの「学 (がっ)」や濁音化されたもの「やけ食 (ぐ) い」は記述されていないので、これらの変形を考慮する必要がある。

を、形態品詞、品詞細分類、活用型、活用形、表層形の5つ組によって定義しており[3]、表層形が異なるものは、すべて別の形態素として取り扱われる。たとえば、「食べる」と「たべる」は別の形態素として扱われる。もちろん、「赤い」と「赤み」のような派生語関係にあるものも、別の形態素となる。

一方、「出題基準」の語彙表では、漢字表記とひらがな表記は、同一語として扱われている。また、形容動詞の語幹に「み」や「さ」がついてできた名詞や、動詞から派生した可能動詞や使役動詞は、元の単語に含まれるとしており、語彙表には含まれていない。このように、語彙表と形態素解析システムには同一語の扱いに関する不整合が存在するため、「出題基準」の語彙表を、そのまま機械的に判定可能な語彙の規格とすることはできない。

この問題を解決するためには、どちらか一方をもう一方に合わせるということが必要となるが、我々は、語彙表を形態素解析システムに合わせるという現実的な選択肢を採用する。すなわち、形態素解析システムの辞書の各形態素エントリーに対して語彙表に基づいて平易度を割り当て、この辞書をもって規格の定義とするという方法をとる。この方法の最大の利点は、単語の平易度判定のための特別なプログラムを必要とせず、既存の形態素解析システムが利用できるという点である。

4.2 語彙表に基づく辞書エントリーへ平易度割当

実際に、「出題基準」の語彙表で与えられた級に基づき、形態素解析辞書の各エントリーに平易度を割り当てる際には、多くの細かな問題を解決する必要がある。以下に、そこで生じる主要な問題とその解決策を示す。

(1) 語彙表には助詞、助動詞、動詞性接尾辞などが含まれていない。→ それらには平易度を設定しない。

(2) 語彙表のほとんどの単語には、品詞が明示されてないため、品詞が確定できない。→ 品詞が明示されているものは、その品詞を持つもののみに対応させる。品詞が明示されていないものは、読みと漢字が整合するものすべてに対応させる。

(3) 語彙表の形容動詞は語幹で示されているが、Jumanの形容動詞(正確には、形容詞のうち、ナ形容詞、ナノ形容詞、タル形容詞のいずれかの活用型をとるもの)は基本型「〜だ」で書かれている。→ プログラムで不整合を機械的に吸収する。

(4) Jumanにはカタカナ語のエントリーがない → 語彙表に含まれるカタカナ語をJuman辞書に追加する。

(5) 語彙表で「造語成分+語」となっているものが、Jumanでは1語となる。あるいは、その逆 → 対応するJumanの1語に、造語成分と語の平易度の低い方の平易度を与える。逆の場合は、Juman辞書に語を追加する。

(6) 動詞の派生語(可能、使役)が語彙表には含まれていない → IPALから、動詞と可能動詞、使役動詞の対応表を作成し、この表に基づいて平易度を割り当てる。

(7) 動詞の補助的用法(「〜やすい」)が、Jumanでは1語扱いのものがある → 5. と同様の方法で処理

(8) 形容詞、形容動詞の語幹から派生した名詞が、Jumanでは別語 → プログラムで該当する名詞をみつけ、平易度を割り当てる

表1 語彙の規格
Table 1 Standards of vocabulary

	形態素数	累計	(語彙表)
V4	1438	1438	(728)
V3	1405	2843	(1,409)
V2	7527	10370	(5,035)
V1	5293	15663	(8,009)
V0(規格外)	215490	–	–
V9(対象外; 助詞、助動詞など)	268	–	–
総計	231421	–	–

これらの処理によって、Juman辞書の全エントリー231,421語に対して、15,663語に平易度(3〜1)を割り当てた。レベル毎の内訳を**表1**に示す☆。こうして作成した辞書をJumanシステムに組み込むことにより、語彙の規格判定プログラムを実現した。

5. 文法の規格とその自動判定

5.1 規　　格

4部門の規格の内、最も難しいのが文法部門の規格の設定である。その問題の根幹は、文法部門が判定すべき対象が、他の部門ほど明確ではない点にある。既に述べたように、漢字と語彙は、単語が判定の単位となる。また、次節で述べる複雑さは、文が判定の単位となる。これに対して、文法は、何を単位として判定すればよいか明確ではない。

「出題基準」では、文法テストの出題基準として、次のような内容が示されている。

3・4級 (A) 文法事項と (B) 表現意図等に分け、それぞれ、表現形式と例文が示されている。文法事項は、さらに、(I) 文型／活用等と (II) 助詞／指示詞／疑問詞等に分けられている。

1・2級 「文法的な<機能語>の類」のサンプルが、用例とともに示されている。

これらの内容から、

- 助詞・助動詞等の付属語(機能語):単語または文節を単位とする判定
- 文型／活用等の文法事項:単語、文節、係り受け、文などを単位とする判定
- 表現意図で記述される文法事項:文節、係り受け、文などを単位とする判定
- <機能語>の類:単語列、文節列、係り受けなどを

☆ 規格として最終的に定義するのは平易度3〜1であるが、この辞書においては、「出題基準」4級に対応する単語には、平易度4(V4)という記号を付加した。

単位とする判定

などが、文法部門で扱うべき内容であることが想定される。

機械処理の側面から考えると、文法部門に対応する処理システムは、構文解析である。もし、構文解析が、宣言的規則の集合によってのみ規定されるシステムで実行可能であれば、それぞれの規則に平易度を設定し、それによって、文の構文的平易度を判定するという方法が妥当なものとなろう。しかしながら、現実的には、選好知識 (プリファレンス) や手続き的知識を導入しなければ、一つの構文木に絞り込むことはできないため、単純にはこのような方法をとることはできない。また、係り受け解析の場合は、解析処理の目的 (それぞれの文節の係り先を決定すること) と、使用されている文法事項の発見は、完全に一致するわけではない。これらの理由から、今回は、文法部門の平易度判定を、構文解析システムとは独立に実現するという方針を採用した。

文法部門、特に、「<機能語>の類」におけるもう一つの問題は、「出題基準」の記述が網羅的でなく、かつ、要素合成に関して非単調性がみられる点である。ここでいう非単調性とは、「か」、「の」、「ようだ」はそれぞれ G3 であるのに対して、「～かのようだ」は G2 となることを意味する。つまり、正確に平易度判定を行なうためには、「<機能語>の類」の網羅的リストを作成し、規格外 (G0) に対しても「<機能語>の類」を定義する必要がある。

これは、かなり大変な作業となるので、我々は、最初の版では、G1、G2、G3 に対して、できるだけ網羅的なリストを作成する方針を立てた。「出題基準」に記載されている「<機能語>の類」だけでは不足なので、日本語能力試験に対応した教科書[4),5)]でその不足を補うこととした。また、他の部門の規格は、完全に機械的に判定可能な規格として定義したが、文法部門に関しては、人間が理解可能なリスト（マスター規格）と、それに対応する (機械が解釈可能な) 規則集合の 2 段構えで規格を定義することとした。

5.2 自動判定の方法

自動判定は、それぞれの文法事項に対して、発見規則を定義することによって実現する。先に述べたように、文法部門の規格の判定単位は 1 つには定まらない。このため、各種の単位に対する発見規則が必要となる。

ほとんどの文法事項の発見は、形態素列パターンによる発見規則によって実現できる。たとえば、形容詞の丁寧な現在形の否定「Aクナイデス」(p123) に対する発見規則は、次のような記述となる。

```
ge 4, 'Aクナイデス',
    '<形容詞基本連用形><基本形:ない><::です>';
```

この規則は、文法事項「Aクナイデス」が規格 G3☆に含まれることを定義しており、最後の要素が、この文法事項を発見するための形態素列パターンである。このパターンは、『形容詞の基本連用形、「ない」の基本形、「です」』という長さ 3 の形態素列とマッチする。

☆ 文法に対する規格においても、4 級に対応する文法事項には、平易度 4 という記号を付加しているが、これは G3 に含まれるものとしている。

表 2 文の長さ
Table 2 Average length of sentences

級	1 文の平均長
1	40 字～65 字
2	30 字～45 字
3	25 字～30 字
3	20 字～25 字

付属語を含まない文法事項は、文節区切り位置の指定を含む形態素列パターンとして記述する。たとえば、形容詞の普通の現在形の肯定「A (辞書形)」(p123) に対する発見規則は、次のようになる。

```
ge 4, 'A!辞書形', '<形容詞基本形>\b'
```

この規則の発見パターンの\b という記号は、この場所に文節区切り位置が来ることを要求する。したがって、このパターンは、形容詞の基本形が文節の末尾に現れた場合にのみマッチする。

係り受け単位で記述するのが適切な文法事項もある。たとえば、形容詞の連用修飾を表す「Aク＋V」は、形容詞と動詞の間に文節が入りうるため、形態素列パターンとしては記述できない。このような文法事項に対する発見規則は、係り文節に対する形態素列パターンと、受け文節に対する形態素列パターンの組によって、以下のように記述する。

```
ge 4, 'Aク＋V',
    '\k{<形容詞基本連用形>\b}{<動詞>}';
```

ここで、\k の 2 引数が係り文節パターンと受け文節パターンを表している。

実際の自動判定では、まず、文を形態素解析 (Juman)・構文解析 (knp) し、その結果の形態素列および文節係り受け構造に対して、上記の発見規則群を適用して、使用されている文法事項を決定する。その後、与えられた平易度を満たさない文法事項を出力する。

6. 量的複雑さの規格

文の平易度に、文の長さや埋め込みの深さなどの量的な複雑さが関係していることはおそらく間違いない。ここでは、比較的簡単に計ることができる文の長さについて、規格を設定する方針を立てた。なお、埋め込みの深さについては、長さを制限すれば、おのずと制限されるだろうという考えにより、今回は、規格化を見送った。

実際に文の長さの規格を決めるには、その長さを何を単位として測るかに、いくつかの選択肢（文字、単語、文節）が存在する。「出題基準」(p224) には、読解テストで使用する文の長さ（句読点を含まない）が**表 2** のように示されている。そこで、今回は、これに基づき文字を単位とした長さの規格を設定することにした。具体的な値は、この表の大きい方の値に約 20%のマージンを付加し

た値の、36 字 (平易度 3)、55 字 (平易度 2)、80 字 (平易度 1)、を採用した。

7. 規格判定プログラムの実行例

現在までに、漢字の規格判定プログラムと語彙の規格判定プログラムの実装が、ほぼ完了している。文法の規格判定プログラムは、一部未実装の部分を残している。これらのプログラムは、現時点では統合されておらず、独立に動作する。

これらのプログラムは、emacs クライアントと通信するサーバーとして設計・実装されている。emacs のクライアントは、スペルチェッカー ispell のクライアントプログラムと良く似たプログラムとなっており、指定されたリージョンに含まれる文をサーバーに送り、サーバーから得られた情報、すなわち、どの部分が規格外であるかをユーザーに表示し、書き換えを促すことを行なう☆。クライアントとサーバー間の通信には、Interactive Rewriting Protocol (IRP) と名付けたプロトコルを定義し、使用している。

図 1 に、漢字の規格判定（読み候補生成を含む）プログラムの実行例を示す。2 行目がサーバーへの要求であり、「死亡または重傷を負う危険が切迫して生じることが想定される内容です」という文を、漢字の平易度 3 という規格 (K3) でチェックし、規格を満たさないものに対しては、その代替表現 (この場合は「読み」) を出力することを要求している。3 行目の 210 から最後のピリオドまでの行が、その要求に対するサーバーの応答である。それぞれの行は、先頭からの文字数、対象文字列、メッセージ (理由説明)、候補の数、候補リスト、を表す。

図 2 に、語彙の規格判定プログラムの実行例を示す。ここでも、2 行目がサーバーへの要求であり、この場合は、語彙の平易度 3 の規格 (V3) でチェックすることを要求している。3 行目以下がサーバーからの応答である。この応答から、7 つの単語 (「死亡」、「重傷」、「負う」、「切迫」、「生じる」、「想定」、「内容」) が規格 V3 を満たさないことがわかる。

図 3 に、文法の規格判定プログラムの実行例を示す。2 行目がサーバーへの要求であり、「日本語の規格化について検討するかたわら、その実現を試みた」という文を文法の平易度 3 の規格 (G3) でチェックすることを要求している。3 行目以下がサーバーからの応答であり、「～について」と「～（する）かたわら」という 2 つの「＜機能語＞の類」が、規格 G3 を満たしていないことを示している。

8. 議　　論

日本語に対して、文のやさしさ・わかりやすさの規格を明確に定め、その規格を満たすかどうかを自動的に判定することを実現しようという試みは、我々のこのプロジェクトがおそらく最初のものである。このような規格の策定においては、表現力の観点と判定方法の実現可能性の観点をうまくバランスさせ、最終的に実用に耐えうる有用な規格に収束させていく必要がある。これは、一朝一夕になし得るものではなく、長期に渡る継続的な改良と実証試験が必要である。そのためには、まずは、不十分ではあっても、出発点となる規格の第 1 版を定め、実際に動作する規格判定プログラムを実現することが重要と考え、本論文で報告した内容の研究を行なっている。

すでに 2 節で述べたように、規格立案においては、判定方法の機械的実現可能性の観点を考慮したが、実際に、判定プログラムの実現に取り組むと、各種の問題の存在が明らかになってきた。たとえば、形態素解析は、技術的にほぼ確立され、精度的にも問題ないと一般に認識されつつある。しかし、ここでの精度は、単語区切位置の判定精度であり、形態素の認定精度ではない。語彙の規格判定では、形態素の認定精度が問題となるため、この精度を向上させることが必要となる。このためには、形態素解析用辞書のエントリーの整理が不可欠である。

文法の規格判定プログラムは、今回は、構文解析システムと独立に作成することにしたが、将来的には、統合することが望ましい。特に、「＜機能語＞の類」を把握してそれを係り受け解析に利用することは、係り受け解析の向上にもつながると考えられる。

今回、「出題基準」を出発点として規格を作成したが、我々の用途においては、各部門が十分にバランスしているとはいいがたい。文法部門は、平易度 3(G3) で基本的な文法事項がほとんどカバーされているのに対し、語彙の平易度 3(V3) は、かなり語彙が少なく窮屈な感じが否めない。また、漢字の平易度 3(K3) は、ほとんど漢字が使えないに等しい。これは、日本語能力試験が外国人のための試験であることに強く関連している。今後、「出題基準」から離れ、我々の用途に合うように、平易度規格を見直していく必要がある。

規格判定プログラムがある程度動くようになると、これらのプログラムは、次の研究を進めるためのツールとして利用できる。たとえば、現在、規格外の単語に対する代替表現の候補を提案するシステム (1 節の目標 2) を実現するために、辞書の定義文から代替表現を抽出することを行なっているが、見出し語より平易な代替表現を定義文から抽出するためには、規格判定プログラムを利用する必要がある☆☆。また、今後、オリジナルの文と、それを平易度規格を満たすように書き換えた文の組からなる書き換えコーパスを作成することを計画しているが、この際にも、これらのプログラムを利用する予定である。

☆ サーバーが代替表現の候補を出力した場合は、その候補を提示する。

☆☆ 辞書の定義文に現れる代替表現は、かならずしも見出し語より平易度が高い (やさしい) わけではない。

```
200 qingqingyomi-withpronounnew-0.1 is running
do 3 K 死亡または重傷を負う危険が切迫して生じることが想定される内容です
210
0 死亡 『亡(K2)』 は規格を満たしていません 1 しぼう
5 重傷 『傷(K1)』 は規格を満たしていません 2 じゅうしょう おもで
8 負う 『負(K2)』 は規格を満たしていません 1 おう
10 危険 『危(K2)険(K2)』 は規格を満たしていません 1 きけん
13 切迫 『迫(K1)』 は規格を満たしていません 1 せっぱく
23 想定 『想(K2)定(K2)』 は規格を満たしていません 1 そうてい
28 内容 『内(K2)容(K2)』 は規格を満たしていません 1 ないよう
.
```

図 1 判定結果 (漢字)
Fig. 1 System output (Kanji)

```
200 vocabulary-0.1 is running
check 3 V 死亡または重傷を負う危険が切迫して生じることが想定される内容です
210
0 死亡 V2(言い換え必要です)
2 または -
5 重傷 V0(言い換え必要です)
7 を -
8 負う V1(言い換え必要です)
10 危険 V3
12 が -
13 切迫 V0(言い換え必要です)
15 して V4
17 生じる V2(言い換え必要です)
20 こと V3
22 が -
23 想定 V0(言い換え必要です)
25 さ V4
26 れる -
28 内容 V2(言い換え必要です)
30 です V4
.
```

図 2 判定結果 (語彙)
Fig. 2 System output (vocabulary)

```
200 grammer-check-0.0.1 is running
check 3 G 日本語の規格化について検討するかたわら、その実現を試みた。
200
0 日本語 -
3 の 「の!接」は G3 です
4 規格化 -
7 について 「～について」は G2 です 0
11 検討する 「Sする」は G3 です
15 かたわら 「～かたわら」は G1 です 0
19 、 「読点」は G3 です
20 その 「その!指示詞」は G3 です
22 実現 -
24 を 「を!格」は G3 です
25 試みた 「Vタ」は G3 です
28 。 「句点」は G3 です
.
```

図 3 判定結果 (文法)
Fig. 3 System output (grammar)

9. 関連研究

本稿で述べた内容と最も関連が深い研究は、日本語読解学習支援システム「リーディング チュウ太」の研究[6]である。このシステムの一部であるレベル判定ツールでは、与えられた文の漢字と語彙の、「出題基準」に基づくレベル判定を実現している（文法のレベル判定は実現されていない)。このレベル判定ツールの機能は、今回我々が作成した漢字規格の自動判定、および、語彙規格の自動判定にほぼ対応する。しかしながら、「チュウ太」の漢字レベル判定は、文字を単位として行なわれているようで、「眼鏡」のような例を正しく判定できない。また、語彙レベル判定でも、複合語等に適切に対応していない。

言語のサブセットを定義することは、制限言語の研究

として歴史がある (たとえば、文献 [7])。これまで研究されてきたのは、主に、機械翻訳等の機械処理を容易にするための制限言語である。これに対して、我々の研究は、人間の理解を容易にすることを目的として、多段階のレベルを設定する点に特徴がある。

他に関連する研究としては、言い換え（パラフレーズ）の研究がある。我々が目指していることは、テキストの平易化と見なすこともできる。テキストをより平易に言い換える研究としては、乾らの研究 [8] や鍛治らの研究 [9] がある。乾らの研究では、その対象を聴覚障害者としており、聴覚障害者が理解しづらい表現形式にターゲットを絞って、それらを平易化することに主眼を置いている。また、鍛治らは、辞書の定義文を利用した単語の言い換えを提案している。しかしながら、いずれの研究も、テキストのやさしさ・むずかしさを計る尺度を提示していないため、「なにをもってテキストが平易化されたか」ということが不明確である。本研究で提案する規格は、この問題に対する解答を与えるものである。

謝辞 本研究の一部は、科学研究費補助金基盤研究 (B)(2)「言い換えを中心としたテキスト自動編集技術とその機械翻訳への応用」(課題番号 13480097) によって実施した。

参 考 文 献

1) 佐藤和之: 外国人のための災害時のことば, 月刊言語, Vol. 25, No. 2, pp. 94–101 (1996).
2) 国際交流基金, 財団法人日本国際教育協会: 日本語能力試験出題基準【改訂版】, 凡人社 (2002).
3) 黒橋禎夫, 長尾真: 日本語形態素解析システム JUMAN version 3.61, 京都大学大学院情報学研究科 (1998).
4) 友松悦子, 宮本淳, 和栗雅子: どんな時どう使う日本語表現文型 500, アルク (1996).
5) 友松悦子, 宮本淳, 和栗雅子: どんなときどう使う日本語表現文型 200, アルク (2000).
6) 川村よし子: 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 (1999).
7) Mitamura, T. and Nyberg, E.: Automatic Rewriting for Controlled Language Translation, *Automatic Paraphrasing: Theories and Applications, NLPRS2001 Workshop*, pp. 1–12 (2001).
8) 乾健太郎: コミュニケーション支援のための言い換え, 言語処理学会第 7 回年次大会ワークショップ「言い換え／パラフレーズの自動化」, pp. 71–76 (2001).
9) Kaji, N., Kawahara, D., Kurohashi, S. and Sato, S.: Verb Paraphrase based on Case Frame Alignment, *Proceedings of 40th Annual meeting of the association for computational linguistics*, pp. 215–222 (2002).