文法の規格とその自動判定

土屋 雅稔　　佐藤 理史
京都大学大学院情報学研究科
tsuchiya@pine.kuee.kyoto-u.ac.jp, sato@i.kyoto-u.ac.jp

1 はじめに

ほとんどの市販薬には、その薬の服用方法についての説明書が添えられている。もしも、その説明書の記述が非常に難しく、購入者が服用方法を理解できない場合は、正しい薬効が得られないばかりか、場合によっては人命にも関わる可能性がある。このような重要情報を伝達するためのテキストは、想定する読み手が困難なく理解できるように配慮して記述されるべきである。テキストの伝達すべき情報が重要であればあるほど、このような配慮は必要不可欠なものとなる。

現在は、想定読者に対する配慮は、書き手の個人的な直観に完全に依存している。しかし、書き手によって分かり易さに関する考え方が異なるため、テキストの品質が安定せず、分かり難いテキストが存在する原因となっている。つまり、重要な情報を伝達するためのテキストは、そのテキストに記述されている情報が読者に確実に伝わることを保証するために、読み易さに関するガイドラインまたは規格にしたがって記述されるべきである。

佐藤ら [1] は、そのような規格として、4部門 (漢字、語彙、文法、量的複雑さ)、3段階の平易度からなる規格を提案している。

平易度3 最も易しいレベル。生命の安全に直結する情報など、できるだけ多くの人々に最優先で伝達すべき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などを記述するのに用いるレベル。

平易度1 最上位レベル。その他の比較的複雑な情報を記述するのに用いるレベル。

このような規格を実効性あるものとするためには、規格を定義するだけでは不十分であり、与えられたテキストがその規格を満たしているか否かを判定する客観的あるいは機械的方法が不可欠である。

本研究では、この提案の一環として、規格の文法部門の内容と、文の文法的な平易度を自動的に判定するプログラムの実現方法について述べる。

2 文法に関する規格

2.1 規格化の方針

日本語の規格を作るということは、日本語のサブセットを定義するという問題に他ならない。この問題を解くには、表現力と実現可能性の2つの観点を考慮する必要がある。非常に制約が厳しい規格を定義すると、規格内表現の変化は少なくなるから、与えられたテキストが規格の範囲内に含まれるか否かを判定することは容易になる。しかし、使える表現が少なくなり過ぎると、伝達すべき内容が伝達できないような規格となってしまう。

規格を零から作り出すことは大変難しいので、我々は、日本語能力試験「出題基準」[2] を出発点として、最初の規格を定義することにした。日本語能力試験は、原則として日本語を母語としない人を対象として日本語の能力を測定・認定することを目的として行う試験であり、文字・語彙、聴解、読解・文法の3科目、1～4級の4段階からなる。ただし、4級は十分な表現力をもたないと判断して、3級と4級を合併して平易度3 の規格とする。

「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。文法事項については、更に (I) 文型 / 活用等と (II) 助詞 / 指示詞 / 疑問詞などに分けられている。

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。ただし、このリストは、1級・2級のレベルを示すための掲示であり、1級・2級に属する「＜機能語＞の類」を網羅したものではない。

3,4 級

(A) 文法事項
A-I 文型 / 活用等

	[文法事項]	[表現形式]	[例文]
1	疑問詞を含む文 $_1$	…ハ疑問詞	それはなんですか
2	疑問詞を含む文 $_2$	疑問詞ガ…	どれがあなたの靴ですか
	⋮		

A-II 助詞 / 指示語 / 疑問詞等

(B) 表現意図等

	[表現意図等]	[表現形式]	[例文]
1	依頼 $_1$	N ヲクダサイ	あのりんごをください

1,2 級

＜機能語＞の類	用例
～あげく / ～あげくに	困ったあげく
～あまり	考えすぎたあまり / 心配のあまり
⋮	⋮
～に足る	満足するに足る
⋮	⋮

図 1: 出題基準の実例

図 1に実際の出題基準の例を示す。これらの内容から、文法部門の規格は、次に示すような内容を含むことが分かる。

1. 助詞、助動詞などの付属語 (機能語): 単語または文節を単位とする判定
2. 文型 / 活用などの文法事項: 単語、文節、係り受けなどを単位とする判定
3. 表現意図で記述される文法事項: 文節、係り受けなどを単位とする判定
4. ＜機能語＞の類: 単語列、文節列、係り受けなどを単位とする判定

したがって、文法規格の判定単位には、単語または単語列、文節または文節列、および係り受けの 3 種類がある。

機械処理の観点から考えると、文法部門に対応する処理システムは構文解析である。しかし、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。そのため、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。

先に述べたように、文法部門の判定対象には 3 種類の大きさの異なる単位があるため、最初から計算機によって解釈可能な規格を人手で作成することは難しい。そのため、文法部門の規格は、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 (計算機によって解釈可能な) 規則集合の 2 段構えで定義することとする。

文法部門、特に「＜機能語＞の類」の平易度判定を実現するためには、「出題基準」に提示された「＜機能語＞の類」のリストが網羅的でなく、かつ、要素合成に関して非単調性が見られる点を解決しなければならない。ここで言う要素合成の非単調性とは、「か」「の」「ようだ」がそれぞれ独立に現れる場合は、接続助詞、判定詞、助動詞として平易度 3 の文法事項として判定されるにも関わらず、これらの 3 形態素が連続し「～かのようだ」という表現となって現れた場合には平易度 2 の文法事項として判定されることを指す。また、複数の文法事項が連続して現れることによって、より難しい文法事項に常に変化するとも限らない。例えば、以下の例文で現れる「～ように」は平易度 2 の「＜機能語＞の類」に含まれているのに対して、より長い「～ようにする」という表現は平易度 3 の規格に含まれている。

熱が下がる ように 注射をする　　(平易度 2)
傘を忘れない ようにして ください (平易度 3)

つまり、正確に平易度判定を行うためには、「＜機能語＞の類」の網羅的リストを作成し、規格外の表現に対しても「＜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、最初の版では平易度 1 ～ 3 の文法規格として出来るだけ網羅的な「＜機能語＞の類」のリストを作成することにした。

2.2 マスター規格

1 つの文法事項に対するマスター規格は、平易度と規格名および例文と正解の対のリストからなる。例として、形容詞の丁寧な現在形の否定「Aクナイデス」(p123) に対するマスター規格を以下に示す。

平易度　4
規格名　Aクナイデス
例文 $_1$　この部屋は広くないです
正解 $_1$　広くないです

ここで、4 級の出題基準に由来する文法事項については、マスター規格でも平易度 4 と記述してあるが、実

際には、平易度4の文法事項は平易度3の規格に含める。例文は、「出題基準」に掲載されていたものをそのまま用いているが、一部の例文については、形態素解析誤りを避けるために平仮名を漢字に書き換えた。また、文法事項によっては、例文と正解の対が複数含まれることがある。

2.3 規則集合

規則集合は、マスター規格の例文を形態素解析・構文解析した後、規則名に含まれている「~」や名詞1語を表す記号である「N」などを手がかりとしてできるだけ機械的に生成した規則を雛形として、人手で修正を行って作成する。

ほとんどの文法事項の判定は、形態素列パターンによる判定規則によって実現できる。たとえば、「Aクナイデス」(p123)に対する判定規則は、次のような記述となる。

```
np( 4, 'Aクナイデス',
    Dm({ H1=>'形容詞', K2=>'基本連用形' },
       { G=>'ない', H1=>'接尾辞', K2=>'基本形' },
       { G=>'です', H1=>'助動詞' }) );
```

ここで、np()は1つの規則を記述するための関数、Dm()は形態素列に一致するパターンを記述するための関数である。この記述は、文法事項「Aクナイデス」の平易度が4であり、この文法事項は、形容詞の基本連用形、接尾辞「ない」の基本形、助動詞「です」という長さ3の形態素列と一致するパターンによって判定できることを示している。

付属語を含まない文法事項は、文節を単位とする規則として記述する。たとえば、形容詞の普通の現在形の肯定「A(辞書形)」(p123)に対する判定規則は、次のようになる。

```
np( 4, 'A!辞書形',
    Db( { H1=>'形容詞', K2=>'基本形' } ) );
```

ここで、Db()は文節に一致するパターンを記述するための関数である。この記述は、文法事項「A!辞書形」の平易度が3であり、形容詞の基本形と句読点のみからなる文節に一致するパターンによって判定できることを示している。

係り受け単位でしか記述できない文法事項もある。例えば、形容詞の連用形+動詞である「Aク+V」(p123)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。このような判定規則を記述するために、係り元文節に一致するパターンと、係り先文節に一致するパターンを引数としてとる関数Dk()を導入した。

表 1: 規則集合

平易度	規則数
1	134
2	322
3	97
4	95
計	648

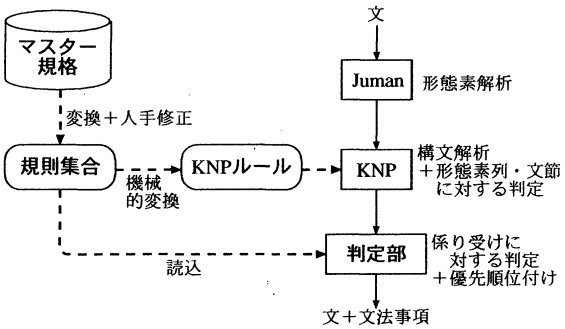

図 2: システム構成

```
np( 4, 'Aク+V',
    Dk( Db({ H1=>'形容詞', K2=>'基本連用形' }),
        Dm({ H1=>'動詞' }) ) );
```

この定義によると、基本連用形の形容詞と句読点のみからなる文節と動詞を含む文節の間に係り受けが存在しているときに、文法事項「Aク+V」が判定されたことになる。

3 規格の自動判定

2.3節で作成した規則集合に基づいて、文中に使われている文法事項を検出するシステムを実装した。規則集合の内訳を表1に、全体の構成を図2に示す。形態素解析にはJuman[3]を、構文解析にはKNP[4]を利用する。規則集合は、KNP用のルール集合に変換した上で、KNPの標準ルール集合に追加した。このようにルール集合を変更しておくと、構文解析と同時に形態素列および文節を単位とする文法事項の検出を行うことができる。係り受けを単位とする検出については、条件に合致する文節の存在をKNPを利用して検出しておき、その結果を後段の判定部で統合する。さらに、文法事項の平易度を考慮して順位付けを行い、最終的な検出結果とする。

文法事項の自動判定を行うためには、出題基準の記

述に用いられている単語区切りと、Jumanによる形態素区切りが一致しないことがある点に注意が必要である。この点には、2種類の問題がある。第1の場合は、出題基準の記述に用いられている区切りが、Jumanによって出力される形態素区切りよりも大きく、出題基準の1単語が複数の形態素からなっている場合である。そのような単位としては、出題基準において名詞1語を表すために用いられている記号Nや、形容動詞1語を表す記号ANなどがある。例えば、平易度3の規格に含まれている「AN・Nデゴザイマス」という文法事項の例文は「この/くつ/は/イタリア/製/で/ございます。」と形態素解析され、記号Nにあたる部分「イタリア製」は名詞と接尾辞の2形態素からなっている。この他、接頭辞や複合名詞などが現れる場合も、同様の問題が生じる。したがって、名詞・動詞・形容詞については、前後の接頭辞、接尾辞、付属動詞を1つの塊と見なして判定を行うようにした。

第2の場合は、出題基準の記述に用いられている単位が、Jumanによって出力される形態素よりも小さい場合である。例えば、「〜っぱなし」という文法事項(平易度1)の例文は「開けっぱなし」である。この文をJumanによって解析すると、「開けっぱなしだ」という形容詞の語幹となるため、「〜っぱなし」に一致する形態素列パターンを単純に書くことはできない。このような一語からなる文法事項に対処するため、パターンに一致する形態素を列挙した辞書を用意することにした。

図3に、本システムの動作例を示す。2行目がシステムの要求であり、「地図はおろか、略図さえも配られなかった。」という文を平易度3の文法規格に基づいて検査することを要求している。3行目以下がシステムからの応答であり、「〜はおろか」と「〜さえ」の2つの「<機能語>の類」が指定された平易度3を満たしていないことが分かる。

4 実験

作成したシステムを評価するため、2つの実験を行った。まず最初に、人手で修正した規則集合の正確さを確認するために、マスター規格(出題基準)の例文656文に対して文法事項の検出を行った。例文から正しく文法事項が検出されなかった例は39例あった。この誤りのほとんどは、形態素解析の誤りに原因がある。たとえば、「〜でなくてなんだろう」(p171)という表現の例文は「これが愛でなくてなんだろう」だが、「愛で」の部分が「愛でる」という動詞の未然形

```
200 grammer-check-0.0.1 is running
check 3 G 地図はおろか、略図さえも配られなかった。
210
0 地図 *
2 はおろか 「〜はおろか」はG1です 0
6 、 「読点」はG4です 0
7 略図 *
9 さえ 「〜さえ」はG2です 0
11 も 「も!副」はG4です 0
12 配ら 「Vレル」はG3です 0
14 れ 「Vレル」はG3です 0
15 なかった 「〜ない」はG4です 0
19 。 「句点」はG4です 0
.
```

図3: 自動検出例

として解析されてしまっているため、「名詞+でなくてなんだろう」という検出規則では検出できなかった。

次に、能力試験に準拠した教科書[5]に記載されている例文1110個を対象として文法事項の判定を行い、それぞれの例文の掲載されているページで説明されている文法事項が検出できるかを調べた。567個の例文からは正しく文法事項が検出され、71個の例文からは間違った文法事項が検出された。残りの例文からは、相当する文法事項は検出されなかった。よって、本システムの再現率は51%、適合率は89%である。

実験の結果、システムの現在の課題は再現率の向上にあることが分かる。今後は、機械解釈可能な規則集合を改善することによって、より頑健なシステムとする予定である。

参考文献

[1] 佐藤理史, 土屋雅稔, 村山賢洋, 麻岡正洋, 王晴晴. 日本語文の規格化. 情報処理学会研究報告, 第2003-NL-153巻, pp. 133–140, 2003.

[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, 1994.

[3] 黒橋禎夫, 長尾真. 日本語形態素解析システムJUMAN version 3.6 使用説明書. 京都大学大学院 情報学研究科, 11 1998.

[4] 黒橋禎夫. 日本語構文解析システムKNP version 2.0 b6 使用説明書. 京都大学大学院 情報学研究科, 6 1998.

[5] 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.