語彙の規格化とそれに基づく用言の言い換え支援

村山 賢洋　麻岡 正洋　土屋 雅稔　佐藤 理史

京都大学大学院 情報学研究科

1. はじめに

家電製品やコンピュータのマニュアルなどの中には、読んでもすぐに内容を理解できないものがある。以下に示す文は、ある携帯電話のマニュアルに実際に記述されていた文である。

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。

このような文章は、非常に重要な情報を伝えており、できるだけ多くの人々がその内容を容易に理解できるように、平易な表現で書かれるべきである。

このような考えに基づき、佐藤ら[1]は、日本語の平易度の規格を定め、それに基づいて平易な文章を作成する方法を提案している。この提案は、日本語文の平易度を、漢字 (K)、語彙 (V)、文法 (G)、量的複雑さ (C) の4部門に分け、次の3段階の平易度を定義している。

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など、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 (基本情報) 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。その他の情報を記述するのに用いるレベル。

それと同時に、この規格が実効的なものとなるように、自然言語処理の手法を用いて自動的に文の平易度を判断するシステムと、文を指定された平易度を満たす文に書き換えること支援するシステムの実現を目指している。

本論文では、上記の提案の一環として、語彙部門の平易度の定義と、それに基づく単語の平易度の自動判定の実現法について述べる。また、用言を、より平易な用言へ書き換えることを支援するシステムについても説明する。

2. 語彙の規格とその自動判定

2.1 規　　格

語彙の規格の作成には、「日本語能力試験 出題基準」[2] (以下、出題基準) を利用する。日本語能力試験は、日本語を母国語としない人を対象にして、日本語の能力を測定し、認定するための試験である。日本語能力試験は1～4級の4段階に分かれており、1級が一番難しく2級、3級、4級となるに従って易しくなる。出題基準は、日本語能力試験の問題作成者のために、問題作成の指針をまとめたもので、「文字・語彙」「文法」「聴解」「読解」の4つに分かれている。

表1　語彙の規格と「日本語能力試験 出題基準」の級との対応表

語彙の規格	出題基準	単語数 (語)
V1	1級	8,009
V2	2級	5,035
V3	3級	1,409
V4	4級	728

語彙の規格では、出題基準の「文字・語彙」の部分を利用する。この部分には、日本語能力試験の各級ごとに使用することができる語彙が語彙表という形で提示されている。

本研究では、この語彙表に基づき、各単語の平易度を設定する。**表1**に、平易度、級、各級に対して提示された単語数を示す。但し、4級の語彙は数が非常に少ないので、実際に使用する規格は、平易度3から平易度1の3段階とする。また、平易度1の語彙に含まれないものは、平易度0 (最も難しい) とする。

なお、以下では、語彙の規格を、語彙部門表すアルファベット (V) と平易度を表す数字 (0,1,2,3) を組み合わせた記号で表現する。

2.2 単語の平易度の自動判定

与えられた文に含まれる単語の平易度を判定するためには、まず、文を単語に分割し、それぞれの単語がどの規格の語彙表に含まれているかを調べればよい。本研究では、形態素解析システムとして、Juman[3] を利用する。

ここで一つ大きな問題が存在する。上記の方法がうまく機能するためには、形態素解析システムが認定する形態素（単語）と、語彙表に登録されている単語がきちんと整合している必要がある。しかしながら、Jumanの形態素と出題基準の語彙表の単語において、このことは一般に成り立たない。

例えば、Jumanでは形態素を品詞、品詞細分類、読み、表記、活用型、活用形の6つで定義しており、「きれいだ」と「綺麗だ」は別の形態素として定義されている。また、「美しい」と「美しさ」のように派生関係にあるものも別の形態素となっている。

ところが、出題基準の語彙表では、漢字表記とひらがな表記は同一の単語として扱われ、エントリとしては一つしか存在しない。また、形容動詞は語幹で表記されている。さらに、形容詞・形容動詞の語幹に「み、さ」が

ついてできた派生名詞は、元の形容詞・形容動詞に含まれるとして、語彙表には明示的には示されていない。このようにJumanの形態素と出題基準の単語は、必ずしも整合しないので、平易度を機械的に判定する際に、出題基準の語彙表を、そのまま用いることができない。

この問題を解決するためには、Jumanの形態素と出題基準の単語のどちらかを、もう一方に合わせることが必要となる。本研究では、出題基準をJumanに合わせる。具体的には、Jumanの辞書の各形態素エントリの「意味情報」の項目に、その形態素の平易度を記述する。この情報は、形態素解析の結果として出力される。すなわち、このような方法をとることにより、形態素解析と同時に各単語の平易度を判定することが可能となる。

2.3 辞書への平易度割り当て

先にのべたように、Jumanの形態素と出題基準の単語との間には、ある種の不整合が存在する。出題基準の級に基づいて、Jumanの辞書に平易度を割り当てるためには、この不整合を解消する必要がある。以下に、その際の問題点とその解決策を示す。

(1) Juman辞書では、各形態素エントリに必ず品詞がついているが、出題基準では、ほとんどの単語に品詞がついていない。このため、出題基準のエントリとJuman辞書のどのエントリが対応するのかわからない。
→ 語彙表で品詞が明示されていたものは、その品詞をもつもののみ対応づける。品詞が明示されていないものは、読みと漢字が合うものすべてを対応づける。

(2) Juman辞書では、形容動詞は「きれいだ」のように「語幹 + だ」の形で存在するが、出題基準では「きれい」のように「語幹」の形で存在している。
→ プログラムで不整合を吸収する。

(3) Juman辞書では、カタカナ語のエントリは存在せず未知語となるが、出題基準にはカタカナ語も含まれている。
→ 出題基準のカタカナ語をJuman辞書に追加する。

(4) 出題基準では「造語成分 + 語」となっているものが、Juman辞書では1語となっているものがある。またその逆で、出題基準では1語となっているものが、Juman辞書では2語以上になっているものがある。
→ 前者は、対応するJumanの1語に、造語成分と語の平易度のうち低い方を割り当てる。後者の場合は、Juman辞書に語を追加する。

(5) 出題基準では動詞の派生語(可能・使役)は元の動詞に含まれるが、Juman辞書では別語となっている。
→ IPALの動詞辞書から、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。

(6) Jumanでは、動詞の補助的用法(「～やすい」「～にくい」など)が1語になる場合と2語になる場合がある。
→ (4)と同じ方法で平易度を割り当てる。

(7) 出題基準では、「美しい」→「美しさ」のように形容詞・形容動詞の語幹に「み、さ」がついてできた派生名詞は、元の形容詞・形容動詞に含まれるが、Jumanでは別語となる。
→ プログラムで形容詞・形容動詞の派生名詞を見つけ、平易度を割り当てる。

(8) 出題基準の語彙表には、助詞・助動詞などの機能語が含まれていない。
→ 出題基準の「文法」の部分を参考に、Juman辞書のエントリに人手で平易度を割り当てる。

これらの処理を行ない、Juman辞書の全エントリ231,593語中、20,752語に平易度(V1～V4)を割り当てた。各規格別のエントリ数を、表2に示す。

こうして作成した辞書をJumanに組み込むことにより、語彙の規格判定システムを実現した。実行例を図1に示す。最後の要素がNILであるものは規格外(V0)であり、それ以外は、VL=nという形式で平易度が出力される。

表2 Juman辞書への平易度割り当ての結果

	Juman辞書のエントリ数	累計	出題基準
V4	1,520	1,520	728
V3	2,054	3,574	1,409
V2	10,313	13,887	5,035
V1	6,865	20,752	8,009
V0(規格外)	210,841	-	-
合計	231,593	-	-

```
赤ちゃん あかちゃん 赤ちゃん 名詞 6 普通名詞 1 * 0 * 0 "VL=3"
の の の 助詞 9 接続助詞 3 * 0 * 0 "VL=4"
寝顔 ねがお 寝顔 名詞 6 普通名詞 1 * 0 * 0 NIL
は は は 助詞 9 副助詞 2 * 0 * 0 "VL=4"
とても とても とても 副詞 8 * 0 * 0 * 0 "VL=4"
愛らしい あいらしい 愛らしい 形容詞 3 * 0 イ形容詞イ段 19 基本形 2 NIL
EOS
```

図1 語彙の規格判定システムの実行例

3. 用言の言い換え支援

前節で述べた語彙の規格判定システムを用いることにより、文中に含まれる難しい単語を特定することが可能となる。しかし、平易な文を作成するためには、ここで見つかった難しい単語を、より平易な単語に置き換えて、文を書き直す必要がある。

この作業を支援するために、本研究では、単語の言い換え候補を提示するシステムを実現する。なお、ここでは、対象とする単語を、用言(形容詞、形容動詞、動詞)に限定している。

表 3　言い換え表現抽出パターン

品詞	パターン数	パターン例
形容詞	26	用言 1 語 ガ格＋用言 副詞＋用言
形容動詞	23	用言 1 語 副詞 1 語 ニ格＋用言
動詞	34	用言 1 語 ヲ格＋用言 用言＋用言

見出し語 [平易度]	<=>	(語義) 言い換え表現 [平易度]
明るい [4]	<=	(1) はっきり見える [3]
明るい [4]	<=	(3) ほがらかだ [2]
温かい [2]	=>	(1) ちょうどよい [3]
温かい [2]	<=	(2) 情けぶかい [1]
暖かい [3]	<=	(2) ゆたかだ [2]
疎い [0]	=>	(1) 親しくない [2]
親しい [2]	=>	(1) なかがいい [3]

図 2　言い換え辞書の例

3.1　言い換え辞書の自動生成

言い換え候補を提示するためには、それぞれの単語に対して、その単語の言い換え表現を定義した辞書（言い換え辞書）が必要である。本研究では、この言い換え辞書を、国語辞典を利用して自動生成する。国語辞典としては、三省堂の例解小学国語辞典を用いた。

国語辞典では、それぞれの見出し語に対して、その語を説明する文（定義文）が与えられている。この定義文から、見出し語に対する言い換え表現を抽出する。これを実現するために、定義文の調査を行ない、言い換え表現を抽出するパターンを作成した。それぞれの品詞に対して作成したパターン数とパターン例を表 3 に示す。

これらのパターンを用いて、見出し語と言い換え表現の対を総計 4771 個抽出した。こうして作成した言い換え辞書の一部を図 2 に示す。

この図より、定義文から抽出した言い換え表現は、必ずしも見出し語より平易な表現となっているわけでなないことがわかる。このような場合は、全体の約 18%であった。このうち、言い換え表現が一語である場合は、その逆方向（言い換え表現 → 見出し語）が、平易な語への言い換えとして利用できる。

3.2　用言の言い換え支援システム

前節で作成した言い換え辞書をもとに、用言の言い換え支援システムを作成した。作成したシステムの動作概要を図 3 に示す。以下では、この図に従って、システムの動作を説明する。

(1)　入力
　文とその文が満たすべき平易度を入力する。

(2)　言い換え部分の特定と候補の提示
　(a)　言い換え部分の特定
　　語彙の規格判定システムを用いて、入力文に含まれる各単語の平易度を判定し、与えられた平易度を満たさない単語を特定する。その単語が用言である場合は、次のステップである、候補の生成を実行する。
　(b)　候補の生成 (図 4)
　　用言を原形に戻したのち、言い換え辞書から、用言の言い換え候補を取得する。取得した言い換え候補を、表 4 に示すようなルールに従って変形する。その後、あらかじめ用意してある活用形のマッピングテーブルに従って、元の単語の活用形と言い換え候補の活用形を合わせる。

(3)　候補の選択
　ユーザーに候補を選択してもらう。

(4)　出力
　選択された候補で該当単語を置き換える。

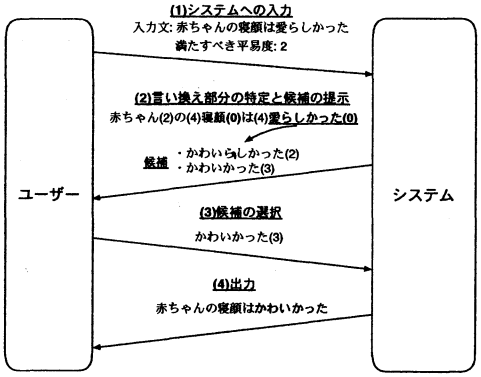

図 3　システムの動作概要

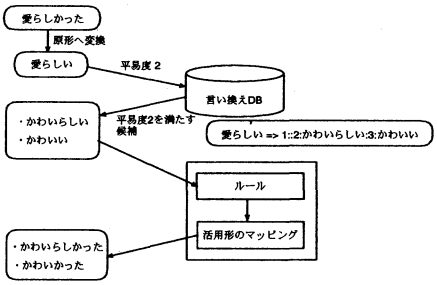

図 4　言い換え候補生成の例

4.　実験と検討

まず、語彙の規格判定の精度を求めるために、以下の実験を行なった。

(1)　国語辞典の形容詞・形容動詞・動詞を無作為に 100 語ずつ選び、その用例を取り出す。

(2)　取り出された用例 (形容詞 121 文、形容動詞 108 文、動詞 153 文) を実験文として規格判定を行なう。

表 5　語彙の平易度判定の実験結果

	形容詞	形容動詞	動詞	全体
正しく判定	607(94%)	497(96%)	727(94%)	1883(95%)
判定失敗	34(6%)	23(4%)	41(6%)	98(5%)
合計	641(100%)	520(100%)	768(100%)	1929(100%)

表 6　用言の言い換えの実験結果

	形容詞	形容動詞	動詞	全体
正しい言い換え候補を出力	73(60%)	66(61%)	68(44%)	207(55%)
候補を生成せず	33(29%)	32(30%)	42(28%)	107(28%)
その他 (失敗)	14(13%)	10(19%)	43(28%)	68(17%)
合計	112(100%)	108(100%)	153(100%)	382(100%)

表 4　言い換え候補の変形ルール

ルール名	説明と例
付属語	元の単語に付属語があり、言い換え表現に同様の付属語がついていた場合は削除する。 例) みずぼらしい → 言い換え：貧しそうだ **みずぼらし** そうな 格好 → **貧し** そうな 格好
動詞	元の単語に続く動詞が、言い換え表現の末尾の動詞と重複した場合は削除する。 例) 冷やか → 言い換え：冷えていると感じる **冷やかに 感じる** → **冷えていると 感じる**
格の変更	元の単語の直前格と言い換え表現の格が重複した場合は、どちらかの格を変更する。 例) 分厚い → 言い換え：暑みがある **本 が 分厚い** → **本 に 暑みがある**
格要素の埋め込み	格の変更が必要な状況で、元の単語の直前格の格要素と言い換え表現の格要素が似ている場合は、言い換え表現の格要素を埋め込む処理をする。 例) 重たい → 心が晴れない 気分 が **重たい** → 気分 が **晴れない**

実験結果を表 5 に示す。この結果は、形態素を単位として評価している。この表より、語彙の規格判定の精度は95%であり、ほぼ実用的な精度となっていることがわかった。判定が失敗した原因の大半は、Juman の解析ミスであり、それらは、a) 単語分割に失敗した場合と、b) 平仮名表記の単語を読みが同じ他の単語と誤認識した場合、の 2 つの場合に分けられる。後者の失敗については、辞書のエントリを整理することで回避できると考えられる。

次に、同じ実験文を用いて、用言に対して、どの程度正しい言い換え候補を生成できるかを調べた。実験結果を表 6 に示す。ここで、「正しい言い換え候補を出力」とは、生成した候補の中に、言い換えとして適切な候補が少なくとも一つは含まれていた場合を示す☆。55%の用言に対して正しい言い換え候補を出力できたことは、不十分ではあるが、それほど悪い値ではない。正しい候補を出力できなかった場合うち、過半の場合が候補を全く出力できなかった。これは、言い換え辞書が不十分であることを意味しており、それを強化することよって、改善できると考えられる。また、その他の失敗のうち、過半は、その語義に対する適切な言い換え候補が存在しなかった場合であり、これも言い換え辞書の不備である。すなわち、言い換え辞書をより豊かなものにすることにより、本システムをより強化することができると考えられる。

☆ 本研究では、語義の曖昧性の解消を行なっていないので、不適切な言い換えを排除することは原理的に不可能である。

5. 関連研究

語彙の規格判定には、日本語読解学習システム「リーディングチュウ太」[4] がある。このシステムの一部である、語彙のレベル判定ツールは、本研究と同様に「出題基準」に基づいて、語彙のレベル判定を行なうものである。しかし、このシステムは形容詞からの派生名詞や複合語の扱いが十分ではない。

文を平易に言い換える研究には、乾らの研究[5] や鍛治らの研究[6] がある。乾らの研究では、対象を聴覚障害者として、聴覚障害者が理解しづらい表現を平易な表現へと言い換えることを目標としている。鍛治らの研究では、本研究と同様に辞書の定義文を利用して用言を平易な表現へと言い換えることを目標としている。どちらの研究も「平易さ」を計る基準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とするか」が不明確である。これに対して、本研究では、先に平易度を定義し、それに従って平易化する方法を与えている。また、本研究では多段階の平易度を定義したので、それぞれの平易度への多段階の平易化が可能である。

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, 自然言語処理研究会, 2003-NL-153, pp133-140,2003
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】
3) 黒橋禎夫, 長尾眞: 形態素解析システム JUMAN version 3.6 使用説明書, 1998
4) 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
5) 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 8 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, 2002
6) 鍛治伸裕, 河原大輔, 黒橋禎夫, 佐藤理史: 国語辞典とコーパスを用いた用言の言い換え規則の学習, 言語処理学会第 8 回年次大会, pp331-334, 2002