「やさしい日本語」のための語彙制限の検討

李 真奈見　　山本 和英
長岡技術科学大学 電気系
{moku, yamamoto}@jnlp.org

1 はじめに

現在、日本に在住する外国人は200万人を超え、その中でも日常生活に必要な日本語能力を持たない外国人は数十万人に及ぶ。しかし、一般的に日本社会で日本語以外は使用されない。よって外国人が日本で生活していくために日本語の知識は必要不可欠である。

外国人のために、必要最低限の日本語を提示する「やさしい日本語」[1]がある。「やさしい日本語」とは、日本語母語話者が日本語の文法や語彙に制限をかけて難しい日本語から「やさしい日本語」へ変換を行ったものである。ここでの難しいとは、最低限の文法と語彙を習得した日本語初学者の外国人が理解できないこと、やさしいとは日本語初学者でも理解できることである。

本研究の「やさしい日本語」で対象としている文書は公的文書である。公的文書とは、市役所や病院、学校等の公共施設で配られる文書を指しており、これらの文書は生活するために重要な情報を多く含んでいる。しかし、日本語初学者が学習する文に比べ理解が困難であり、特有な表現も含むため、「やさしい日本語」へ換言する必要がある。

そこで我々は自然言語処理から見た「やさしい日本語」の調査を行うこととした。本研究では関連研究より語彙の制限により文書をやさしくすることを目的とした。語彙の制限は機械処理による容易な制限を目指すために日本語 WordNet[1)]を使用した。評価実験は日本語初学者を対象として行った。書き換えは公的文書と一般文書に対して行い、このうち一般文書は新聞記事を使用した。

2 関連研究

関連研究として、制限言語[2]がある。制限言語とは特別な目的のために制限された言語である。例として AECMA 簡約英語がある。これは主に宇宙産業で使用される。制限言語で書かれた文書は通常の文書よりもわかりやすいため翻訳を必要としない。制限言語は宇宙関係や乗り物の取り扱い説明書だけでなく、電話、ソフトウェア等の重要な例を引用する取り扱い説明書に使用される。また、国境警察や航空機の操縦士などの危険かつ重要な指示のための対話にも応用される。

現在は様々な形式の簡約英語が幅広くつくられている。英語における制限言語をつくるときの仮定は次の2つである。(1)英語圏で非専門家の人々のために専門用語や見慣れない書き方は通訳する必要がある。(2)非英語圏の人々が読みやすくなるように定められた語彙で書く。これらのことから以下の2つの目標がたつ。(1)専門用語を一般的な標準語彙へ換言する。(2)語彙と構文の制限により文書を単純な形式へ換言する。

3 日本語 WordNet

本研究では語彙の制限のために日本語 WordNet を用いた。これは日本語の概念辞書である。個々の概念はそれぞれ synset という単位にまとめられており、それらが他の synset と意味的に結びついている。英語 WordNet を基に構築されている。英語において名詞、動詞、形容詞、形容動詞に区分される単語の 57,238 概念(synsets)、93,834 語が収録されている[3]。

本研究では内容語を上位語へ置き換えることにより、語彙数を減少させることを考えた。日本語 WordNet を始めとした概念辞書は上位語になるほど抽象的な語彙が多くなる。我々は意味を包括でき、内容が伝わる意味の階層を探す必要があった。そこで外国人に必要な内容を伝えることを目的とし、上位語へ置き換えることによって意味が落ち、正確な意味が伝わらなくても良いこととした。そして最初の足掛かりとして各単語を一つ上位の語へと置き換えた。

4 「やさしい日本語」コーパス

「やさしい日本語」には、2名の日本語教師が公的文書の日本語を「やさしい日本語」に訳したコーパスがある。このコーパスは原文である公的文書と共に逐語訳、意訳、要約という3段階の訳を含んでいる。これらは一定の文法基準[1]と日本語能力試験2級レベルの語彙のみに制限されている。

公的文書は日本語初学者が学習する文に比べ理解が困難であり、特有な表現も含むため、「やさしい日本語」へ換言する必要がある。以下に「やさしい日本語」へ換言した例を示す。

例 1)
公的文書:　　予防接種
やさしい日本語:　予防注射

「予防接種」は重要な情報だが、日本語の学習内容として一般的ではないため、理解できない外国人が多い。しかし、「接種」を一般的な語彙である「注射」に換言することによって、意味を理解しやすくなる。

本研究ではこのコーパスのうち逐語訳のみを評価実験で使用した。逐語訳は日本語文の難しい表現をやさしい表現に、忠実に訳したものと定義している。コーパスにおける難しい、やさしいの基準は日本語教師の主観である。「やさしい日本語」コーパスに含まれる原文に対する逐語訳の例を示す。

例 2)
原　文:ニュース等で報道されておりますように、世界的に新型(豚)インフルエンザの流行が危惧されています。
逐語訳:ニュースなどにもあるように、世界中で新型インフルエンザの流行が心配されています。

このコーパスから原文を無作為に 15 文抽出し、それぞれに対応する逐語訳と上位語へ置き換えた文を評価した。

5 評価実験

上位語へ置き換えることが日本語初学者に有効であるかを評価するために実験を行った。

本研究は留学生や外国人労働者等を対象にしているため、文書のやさしさについての判断は日本語を学習している 7 名の外国人が行った。「やさしい日本語」は日本語能力試験2級レベルの語彙で制限していることから、日本語能力試験の N2

(旧日本語能力試験の2級)保有者を評価者とした。評価者は全員がマレーシア国籍で日本語歴が3年から4年である。うち5名は現在、日本で留学中であり、残りの2名は現地で日本語教育を受けている。

5.1 評価文

評価実験に使用する文は「やさしい日本語」コーパスに含まれる公的文書とアサヒコム[2]にあげられていた新聞記事とした。公的文書からは15文、新聞記事からは10文無作為に抽出した。それらを形態素解析器[3]によって分かち書きを行い、含まれる内容語を日本語 WordNet と照合した。日本語 WordNet にその内容語が記載されている場合は上位語へ置き換え、記載されていない場合はそのままとした。複合名詞は複合名詞の状態で日本語 WordNet に記載されていない場合、分かち書きを行い前方または後方の形態素を削り、組み合わせて照合した。また内容語によっては多義性のあるものも含まれたため、人手で最も近い語義の上位語を選択した。以下にそれぞれの評価文の例を示す。3a が公的文書を用いた評価文、3bが新聞記事を用いた評価文である。下線は書き換えが行われた部分を示す。

例 3a)

原　文: かかりつけ医などの医療機関でお申込みください。

置換文: かかりつけ医などのトリートメント施設でお申込みください。

逐語訳: いつも行く病院など近くの病院で申し込んでください。

例 3b)

原　文: 日本は、自国内の絶滅危惧(きぐ)種についてレッドデータブックの整備を進めてきた実績があり、そのノウハウを計画に生かす。

置換文: 日本は、ステート内の種についてレッドデータブックの修正を進めてきた実績が存在、その才幹を行動計画に生かす。

5.2 評価方法

評価方法は、原文と上位語へ置き換えた文、または「やさしい日本語」コーパスの逐語訳の比較で行った。各評価者には3種類の評価をしてもらった。①それぞれの意味がわかるかは、文末に○をつけてもらい、②やさしいと感じる順位は、評価者がやさしいと感じた順位で番号をふってもらった。③わからない語彙は、評価者が意味がわからないと判断した語を[ ]で区切ってもらった。評価方法の例を表1に示す。

表1 評価方法の例

評価文	評価①	評価②
かかりつけ医などの医療機関でお申込みください。	○	2
かかりつけ医などの[トリートメント施設]でお申込みください。		3
いつも行く病院など近くの病院で申し込んでください。	○	1

これらの評価結果より、原文と上位語へ置き換えた文、または「やさしい日本語」コーパスの逐語訳のどれがやさしいかを総合的に判断した。ここでのやさしいとは意味がわかるようになったか、または順位が向上したかとする。評価①を用いた文書のやさしさについての判断方法を表2に示す。

表2 評価①を用いた文書のやさしさの判断方法

		原文	
		意味がわかる	意味がわからない
置換文または逐語訳	意味がわかる	変化なし	やさしい
	意味がわからない	やさしくない	やさしくない

5.3 公的文書による評価実験

始めに「やさしい日本語」コーパスの公的文書を評価対象とした。公的文書である原文と上位語へ置き換えた文と「やさしい日本語」コーパスの逐語訳のどれがやさしいかについて人手で評価した。評価結果の例を表3に示す。表中の a は置換文が最もやさしいと判断された例、b は逐語訳が最もやさしいと判断された例、c は 置換文も逐語訳もやさしくないと判断された例である。

表3 評価結果の例

	評価文	評価①	評価②
a.	(本人確認済みの[口座]を利用した送金や送金の受領を除きます)		3
	(個人判定済みのアカウントを使い行った支払金額や支払金額の取得を消去します)	○	1
	(本人確認済みの口座を使った送金や送金のもらいは問題ない)	○	2
b.	母子家庭の母で 20 歳未満のお子さんを[扶養]している方で、次の用件をすべて満たす方		3
	母子環境の母で 20 歳未満の子孫を補助している方で、次の目当てをすべて満たす方	○	2
	母子家庭の母で 20 歳未満の子どもを育てている人で、次の条件に全部合う人	○	1
c.	口座振替申込書をお渡しいたしますので、ガス料金のお支払いには、便利な口座振替をご利用ください。	○	1
	アカウント振替用紙をお渡しいたしますので、状態料金のご貿易には、[安楽なアカウント振替]をお使いください。	○	3
	口座振替申込書を渡します。ガス料金の支払いには口座振替が便利です。	○	2

評価実験の結果を表 4、5 に示す。

表 4 評価①それぞれの意味がわかるかの結果

	やさしい	変化なし	やさしくない
置換文	17.1%	15.2%	67.6%
逐語訳	33.3%	22.9%	43.8%

表 5 評価②やさしいと感じる順位の結果

	やさしい	変化なし	やさしくない
置換文	24.8%	46.7%	28.6%
逐語訳	51.4%	46.7%	1.9%

評価①の結果、約 17.1%の置換文がやさしいと評価された。また評価②の結果、原文よりも置換文の方がやさしいと評価されたものは 24.8%であった。評価①と②のそれぞれのやさしいという判断は逐語訳の評価の半分だが、やさしくすることに置き換えが有効であった。

ただし、評価②で順位をつけられないと評価者が判断した場合、評価者は同じ数字を選んでいる。例を表 6 に示す。この評価者は 3 種類の文の全てに 1 をつけた。

表 6 順位がつけられなかった例

評価文	評価①	評価②
その他の給付金等の支給を証明する書類		1
あれやこれやの支払金額金等の消費を裏付行う著作		1
その他の給付金等をもらったことを証明する書類		1

この例の場合、評価①と評価②の双方において置換文と逐語訳は変化なしと判断することになる。しかしこれはどの文の意味も分からず、順位も変化しなかったためにやさしくないという判断が相応しいと考える。そこで評価①と②を総合して新たに表 7 の判断方法を使用した。

表 7 評価①と②を総合したやさしさの判断方法

		評価①		
		やさしい	変化なし	やさしくない
評価②	やさしい	やさしい	変化なし	変化なし
	変化なし	変化なし	変化なし	やさしくない
	やさしくない	変化なし	変化なし	やさしくない

表 7 の判断方法を用いて評価①と②を総合した結果を表 8 に示す。

表 8 評価①と②を総合した結果

	やさしい	変化なし	やさしくない
置換文	1.0%	52.4%	46.7%
逐語訳	18.1%	55.2%	26.7%

置換文でやさしいと評価された文は 1.0%となった。このやさしいの評価は各評価文においてやさしく感じる順位が向上し、意味がわかるようになったものである。これの数値が評価①や②と比較すると減少していることから、順位が向上していてもどちらも意味がわからないような文が多いことがわかった。意味がわからなくなった文は逐語訳の倍ではあるが、置き換える語を制限するなどの処理を行えば改善される余地があると考える。

5.4 新聞記事による評価実験

次に一般的な文書でも有効であるか確かめるため、アサヒコムの新聞記事を評価対象とした。原文と上位語へ置き換えた文のどちらがやさしいかについて人手で評価した。公的文書における置換文の評価結果と合わせて新聞記事の評価結果を表 9、10 に示す。

表 9 置換文に対する評価①の結果

	やさしい	変化なし	やさしくない
公的文書	17.1%	15.2%	67.6%
新聞記事	31.4%	8.6%	60.0%

表 10 置換文に対する評価②の結果

	やさしい	変化なし	やさしくない
公的文書	24.8%	46.7%	28.6%
新聞記事	38.6%	51.4%	10.0%

公的文書の評価結果と比較すると、評価①の結果、置換文のやさしいの評価が 14.3%増加し、やさしくないの評価が 17.6%減少した。また評価②の結果、置換文のやさしいの評価が 13.8%増加し、やさしくないの評価が 18.6%減少した。これらのことから、公的文書よりも新聞記事のような一般文書の方が上位語へ置き換える処理が有効であった。公的文書の置き換えにおいてやさしいという評価が少ない原因は、公的文書に特有の言語表現として頻出する語彙[4]が影響していると考える。以下に公的文書に頻出する語彙を示す。

例 4)
手続き, 登録, 申請, 機関, 場合, 証明, 緊急,
書類, 地域, 年度, 保護者までに

このような語彙が日本語 WordNet に含まれていなかったことから置き換えが行われずに変化なしと判断された、または誤った上位語への置き換えを引き起こしてやさしくないと判断されたと考える。

公的文書による評価と同様に、表 7 の判断方法を用いて評価①と②を総合した結果を表 11 に示す。

表 11 評価①と②を総合した結果

	やさしい	変化なし	やさしくない
公的文書	1.0%	52.4%	46.7%
新聞記事	4.3%	64.3%	31.4%

総合した結果についても比較すると、やさしいの評価と変化なしの評価がそれぞれ 3.3%、11.9%増加し、その分やさしくないの評価が 15.3%減少した。このことからも、公的文書よりも新聞記事のような一般文書の方が上位語へ置き換える処理がやさしくすることに有効であった。

6 考察

本研究では内容語を日本語 WordNet に登録されている限り上位語へ置き換えた。しかし制限をかけすぎると抽象的な語

彙や多義性の多い語彙が残りやすく、Caterpillar Fundamental English(CFE)[5]という制限言語はこれにより伸び悩んだ背景がある。そこで評価者が意味がわかる、またはわからないと判断した語彙から置き換え可能な語彙を調査した。

上位語へ置き換えることで意味がわかるようになった語彙の一部を表 12 に示す。逐語訳の記入がないものは新聞記事から抽出した語彙である。

表 12 意味がわかるようになった語彙

原文	置換文	逐語訳
口座	アカウント	口座
受領	取得	もらい
給付金	支払金額金	給付金
扶養している	補助している	育てている
売却	販売	−
焦点	対象	−

これらの語彙は上位語へ置き換えることが有効であった。このうち受領、扶養、売却はサ変名詞である。サ変名詞は動詞に換言することが可能なものは動詞から学習する傾向がある。またサ変名詞は数も多いため、学習数も限定されている。よって動詞に換言可能なサ変動詞を換言することはやさしくすることにつながると考える。

また口座と給付金は原文と逐語訳が等しい。これらは公的文書内で頻出する語彙であり、「やさしい日本語」では学習する必要がある単語としている。しかし、本実験の評価者は「やさしい日本語」の日本語初学者の前提にある公的文書に頻出する語彙を学習していないため意味がわからなかった。またそれらが上位語へ置き換えることによって、日常生活でも使われる単語となって一時的に理解を促した。しかしこれらは「やさしい日本語」では置き換える必要のない言語表現である。よってこれらは置き換えを必要としない語彙として登録することにより、誤変換を防ぐこともできると考える。

対して上位語へ置き換えることで意味がわからなくなった語彙の一部を表 13 に示す。

表 13 意味がわからなくなった語彙

原文	置換文	逐語訳
医療機関	トリートメント施設	病院
ご心配	お患い	ご心配
生年月日	滑出し	生年月日
都市ガス警報器	薪炭機器	ガスのベル

医療の上位語であるトリートメントを和訳すると、手当や治療などがある。このことから間違いではないが、医療機関の置き換えとしては不適切である。医療機関という複合名詞は日本語 WordNet に含まれていなかったため、医療と機関それぞれで照合したことが問題であった。都市ガス警報器も都市ガスを薪炭、警報器を機器と置き換えたことにより意味が伝わらなかった。複合名詞は日本語初学者の読解を妨げている背景もある[6]。しかし、複合名詞を分かち書きした上での置き換えは表 12 における給付金が支払金額金へ置き換えられたように日本語として不自然であっても意味がわかるようになる例もある。表 13 における医療機関がトリートメント施設へ置き換えられた例も、評価者 7 人のうち 3 人は語彙の意味がわからなくなったと判断したが、評価者 7 人のうち 5 人は文全体を見たときのやさしさの順位が原文よりやさしくなったと判断している。よって、複合名詞はその形態を維持できなければ置き換えすべきではないと一概にいうことは難しい。

ご心配の心配は置き換えるべきと前述したサ変名詞である。しかしこの心配という語彙は日常生活で頻出する語彙であり置き換えをしなくても評価者は意味がわかった。生年月日についても同様に日常生活で頻出する語彙である。

7 おわりに

日本語初学者のために「やさしい日本語」に自動的に翻訳する研究を進めている。本研究では日本語 WordNet を用いて語彙を制限し、実験によってその有効性を確認した。そして頻出しないサ変名詞と複合名詞が日本語初学者の読解を妨げていることを実験で明らかにした。公的文書だけでなく一般的な文書でも日本語 WordNet を用いた語彙制限がやさしくすることに有効であった。今後は置き換えの制限の制定を行いたい。

参考文献

[1]庵功雄.「やさしい日本語」をめぐって.多文化共生社会における日本語教育研究会 第4回研究会, pp.1-12 (2008)

[2]Richard I. Kittredge. Sublanguages and Controlled Languages. The Oxford Handbook of Computational Linguistics, pp.430-447 (2005)

[3]Francis Bond, Timothy Baldwin, Richard Fothergill, Kiyotaka Uchimoto. Japanese SemCor: A Sense-tagged Corpus of Japanese. The 6th International Conference of the Global WordNet Association (GWC), no page number (2012)

[4]筒井千絵.試用版書き換えコーパスの作成.日本語教育学会大会 2009(平成 21)年度春季大会予稿集, pp.86-87 (2010)

[5]Christine Kamprath, Eric Adolphson, Teruko Mitamura, Eric Nyberg. Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English. Proceedings of the Second International Workshop on Controlled Language Applications (CLAW), pp.51-61 (1998)

[6]杢真奈見,山本和英.公的文書に対する「やさしい日本語」換言辞書作成のための調査.言語処理学会第 17 回年次大会発表論文集, pp.376-379 (2011)

使用した言語資源及びツール

1) 日本語 WordNet, Ver.1.1,
独立行政法人情報通信研究機構(NICT),
http://nlpwww.nict.go.jp/wn-ja/
2) アサヒコム (2010 年 10 月 4 日)
http://www.asahi.com/
3) 形態素解析器 ChaSen, Ver.2.3.3,
奈良先端科学技術大学院大学松本研究室,
http://chasen.naist.jp/hiki/ChaSen/,