様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ９ 月　 ８ 日現在

機関番号：１３１０２
研究種目：挑戦的研究（萌芽）
研究期間：2017～2019
課題番号：１７Ｋ１８４８１

研究課題名（和文）やさしい日本語化実証実験による言語資源構築と自動平易化システムの試作

研究課題名（英文）Constructing simplified Japanese corpus and prototyping automatic text simplification

研究代表者
山本　和英（Yamamoto, Kazuhide）

長岡技術科学大学・工学研究科・准教授

研究者番号：４０３５９７０８

交付決定額（研究期間全体）：（直接経費）　　4,800,000円

研究成果の概要（和文）：(1)やさしい日本語かどうかを自動判定するやさしい日本語チェッカーを作成してWebアプリとして一般に公開した
(2)このツールを使って50,000文規模からなる入力文とやさしい日本語テキストの対訳コーパスを収集した。これは現時点で平易化コーパスとして日本語唯一であると同時に世界最大規模である。同時に、2000語から構成されるやさしい日本語の語彙・文法を定義した。さらに、クラウドソーシングを利用して前述のやさしい日本語コーパスを拡張し、35000文(両者で85000文)のコーパスを構築した。
(3)以上のコーパスを用いて、日本語平易化に関する各種研究を実施した。

研究成果の学術的意義や社会的意義
やさしい日本語に対する潜在需要と一般の関心は高く、NHK News Web Easy や自治体などで徐々に社会的に認知される段階に入りつつある。そのような中で、日本語で唯一の平易化コーパスを構築して公開した意義は非常に大きい。日本語の自動平易化研究は語彙平易化を除けば現時点で本研究課題のみであり、自然言語処理への貢献も非常に大きいと考える。政府や自治体からのお知らせがもし自動でやさしい日本語に変換することができれば情報保障の観点から非常に有益で、本研究課題はそのための基礎を構築できたと考えている。

研究成果の概要（英文）：(1) A simple Japanese checker is developed that automatically determines whether it is simple Japanese, and it is released to the public as a Web application.
(2) Using this tool, a Japanese simplified corpus is created which includes 50000 original sentences and their corresponding simplified ones. At present, this is the only simplified Japanese corpus and is also the largest in the world. At the same time, a simple Japanese vocabulary of 2000 words and grammar are defined. In addition, the simplified Japanese corpus described above is expanded by using crowdsourcing, and a corpus of 35,000 new sentences is created.
(3) Using the above corpora, various studies on Japanese simplification were conducted.

研究分野：自然言語処理

キーワード：やさしい日本語　平易化　対訳コーパス

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

【背景と経緯】

やさしい日本語に対する潜在需要と一般の関心は高く、『やさしい日本語』(庵功雄著，岩波新書，2016 年）や NHK News Web Easy、各種講演会など多くの場で耳にする。また科研費においてもいくつかの研究課題が採択されている。しかし、(1)やさしい日本語の具体的な語彙・文法項目の提案はほとんどなく、(2)やさしい日本語に書き換えるためのツールは全く整備されていない。このように、やさしい日本語に対して学術が一般の潜在需要に応えられていない。

【研究の意義】

本研究の特徴は(1)実証実験であること、及び(2)工学（自然言語処理）の観点からやさしい日本語に取り組む点である。まずチェッカーを作成してやさしい日本語書き換え作業の効率化を実現し、この環境下で被験者がやさしい日本語語彙・文法の(仮)策定をする。これをたたき台にしてチェッカーを公開することで実証実験を行い、やさしい日本語化作業の省力化を実現する。同時に、従来研究の最大の問題であったやさしい日本語テキストを大規模に収集することを可能にする。

すなわち、

A)　チェッカーの公開で、ある基準でやさしい日本語に書き換えたいという需要に応える
B)　平易化テキストを収集することでより高度な言語分析・システム構築を実現する
C)　将来の自動平易化のために試作システムを構築、公開して技術の底上げを狙う

以上の問題を同時に実現することを目指すという意味で本課題は意欲的・挑戦的と考える。

日本語の語彙制限についてはいくつかの提案がされている。土居(1933)の 1,000 語、旧日本語能力試験３級の 1,500 語と同２級の 6,000 語、国立国語研究所の「基本語二千」「基本語六千」が主なものである。しかし、これらはいずれも少数の研究者が提案した演繹的なリストであり、（次項で述べるようなチェッカーもないので）これら語彙がどの程度実用性があるのかは全く不明である。

これに対して本課題では多くの実証実験を重ねることで帰納的な方法で語彙制限を目指す。作業の都合上最初に語彙集合は定義するが、これを自由に変更することを許し、多くの作業者の感覚を反映させることで初めての帰納的な制限語彙を作成する。

また、本課題が目指す「やさしい日本語チェッカー」に該当するツールは全く存在しない。これに類するツールとして任意の入力文に対して何らかの難易度を表示するツールは「リーディング チュウ太」や日本語文章難易度判別システムがある。これらツールは主に日本語学習者への情報提供が主眼となっているため「入力文を平易な表現に変換する」ことが目的の場合は使いにくい。また与えられる難易度を入力者が調整するという発想がないため、このような作りになっていない。

これに対して本課題で作成するチェッカーは、入力文が現在登録されている平易語彙に含まれるかどうかだけの表示に特化した専用ツールであり、当該作業には当然使いやすい。また、作業者自身が平易語彙の追加・削除ができるので、非常に柔軟性が高い。

２．研究の目的

「やさしい日本語」は近年重要性を増している考え方である。日本語初学者、子供、障害者などに向けた情報提供の一手段として重要であり、例えば関連する書籍もいくつか出版されている。これに伴って（日本語の）自然言語処理においても自動平易化研究が行われている。

ここで、これら自動平易化研究に必要な言語資源に目を向けると、英語では Simple English Wikipedia というサイトがある。ここでは各記事が平易な英語（後述）によって書き換えられており、原稿執筆時で約 12 万記事が収録されている。この一方で、日本語については Simple English Wikipedia に相当するサイトは存在せず、今後も期待できない。一方 NHK が News Web Easy というサイトを運営している。ここでは日々のニュースが平易な日本語で記述されており非常に有益なデータであるが、言語資源としては公開されている訳ではない。

以上のように、日本語を対象にした自動平易化の研究は我々も含めて少しずつ行われてきているが、言語資源が増える気配を何も感じない。この状況が続くようでは、日本語の自動平易化の研究は英語などの他言語と比較していつまでも遅れを取るだろうという危機感を持っている。そこで本研究では、やさしい日本語に関する言語資源構築の観点から我々自身で何ができるかを考え、限られた時間と労力で実現できるぎりぎりの品質・規模の言語資源構築を行ったので報告する。研究成果物として下記を作成し、一般に公開する。

- ５万文の（日本語、やさしい日本語、英語）対訳コーパス
- ２千語のやさしい日本語辞書（基礎語彙）
- やさしい日本語チェッカー

（関連研究）
平易な日本語を用いたコーパス作成に関する関連研究として[李 2013]を挙げる。ここでは約４万文の公的文書（市役所から市民へのお知らせ文書）に対して文単位でやさしい日本語に書き換えた。このコーパスの大きな特徴はほぼ実在の文書に対して日本語教師が書き換えを行っている点で、長文や難解な概念を含んだ文に対しても分かりやすく書き換えられていたり、場合によっては冗長な言い回しという理由で全文削除されていたりする。また、様々な分野、話題を含んでいる。当初はこれを利用することを検討したが、語彙数や文長、分野などの理由からこのコーパスを用いた処理は難解と判断し、より処理しやすい（つまり現象として簡単化した）コーパスを作ることにした。以上をまとめて本研究と比較した表を表１に示す。

表１：関連研究と本研究との比較

	[李 2013]	本研究
文数	42,274 文	50,000 文
使用語彙	(6,000 語相当)	2,000 語
作業者	日本語教師	一般（学生）
分野	公的文書	（非限定）
平易化	３種類(逐語訳/意訳/要約)	１種類
英訳	なし	あり

３．研究の方法

（使用テキスト）
まず、書き換えを行う原テキストとして、small_parallel_enja: 50k En/Ja Parallel Corpus for Testing SMT Methods を採用した。このテキストは田中コーパスの部分集合で日英機械翻訳のために作成された小規模対訳コーパスである。このテキストを我々が採用した理由は下記の通りである。

1. 我々にとって適度な作業規模である
2. （田中コーパスの性格上）短い文が多い
3. 直ちに機械翻訳の研究が可能である
4. Creative Commons CC-BY である田中コーパスの一部であり、書き換え前の対訳テキストはすでに公開されている

このコーパス中にある５万文すべてをやさしい日本語に変換することとした。作業は本研究室学生５人で行った。コーパスは配布時においてすでに train.ja.000～train.ja.004 の５ファイルに分割されており、この１ファイルをそのまま１作業者分の担当とした。作業内容は常に作業者間で相互に閲覧できる状態にし、作業者間での相談や調整も緊密に行った。

（語彙規模）
やさしい日本語として用いる語彙の規模は、過去の事例を参考にして 2,000 語とした。日本語では、旧日本語能力試験３級語彙が 1,500 語、同２級が 6,000 語の規模であり、これ以外では[土居 1933]において 1,000 語、[国立国語研究所 1984] の「基本語六千」が 6,000 語規模である。また、英語においては Ogden の BASIC English が 850 語、Simple English Wikipedia で 利用できる語彙はこの 850 語と VOA Special English 1500 語、及び固有名詞であり、また Longman Dictionary of Contemporary English では約 2,000 語、Oxford Advanced Learner's Dictionary では約 3,000 語、Macmillan English Dictionary for Advanced Learners では約 2,500 語で語釈文が記述されている。以上から、厳密な比較はできないが日本語でも 2,000 語規模でかなりの説明能力を持つのではないかと予想している。
日本語語彙は、UniDic の単語分割基準に従った。ただし同一用言で活用のみ違うもの（例：行く、行か、行け、行こ）は同一単語とした。多品詞語については別の単語として考え、単一品詞で多義があるものは（単語解析上区別できないので）１単語と見なした。すなわち、

図１：やさしい日本語チェッカー

本作業で定義したのは 2,000 語義ではなく UniDic 上の 2,000 語である。なお、本研究では雪だるま[山本 2016]などによる表記ゆれ解消処理は行っていない。語彙数について、以下の語は定義の 2,000 語に含めず、従って書き換え対象外とした

(a) 記号
(b) 固有名詞、及び固有性の高い一部の単語
(c) UniDic 上で未知語となる語句

（作業方針）

本研究では既提案の基礎語彙を一切使わず、我々自身の手で 2,000 語の基礎語彙選定を行うこととした。

この理由は、すでに提案されている基礎語彙は日本語教育用である可能性があり、我々が目指す語彙集合と一致しない可能性があるためである。また仮にこれが誤解であったとしても、本作業のように具体的にどのような作業を経て選定されたかが不明である。

このことから、本作業の最大の目的は我々自身が利用するためであるが、日本語教育・日本語学における基礎語彙研究への貢献も視野に入れて、あえて他の研究成果とは独立に語彙集合を選定することとした。従って、今回の作業において作業者（学生）には従来研究での語彙集合を一切提示していないし、作業者は日本語学や日本語教育の教育を一切受けていない。（工学系の）一般成人が 2,000 語を選定したらこういう結果になったという意味で他の基礎語彙と比較することは言語研究上非常に興味深いが、この分析については本稿の範囲を超えるので今後の研究を待ちたい。

（作業手順）

次に、具体的な作業は以下のように行った。なお、以下では便宜上平易な語で構成される 2,000 語の語彙集合のことを「やさしい語彙」、これ以外の単語を「やさしくない語」と表記する。

1. BCCWJ における UniDic 高頻度 2,000 語を初期のやさしい語彙として選定する。
2. 入力文に対して単語解析を行い、やさしくない語を含む場合は何らかの書き換えを行う。書き換えは単語単位でなく文単位で行う。やさしい語彙のみでできるだけ同義となるよう努力する前提で、原文中の一部情報の欠落を許す。
3. 作業途中に、作業者にやさしい語彙への追加、及び削除を許す。一定のタイミングで追加語、削除語を収集して、やさしい語彙の定義の修正を行う。なお、この作業過程において、一時的に単語数が 2,000 語より多くなる、または少なくなることを許す。
4. やさしい語彙の定義を修正した場合は、上記ステップ 2 から作業を繰り返す

この作業を効率的に行うために、研究室において Web ベースの簡単な単語チェッカーを自作した。この外観を図１に示す。図１において、青色の単語はやさしい語彙中に含まれている単語であり、白色の単語はやさしくない語である。作業者はこのチェッカーを用いるこ

とで、簡単に書き換え対象語を特定することができ、効率的な作業が可能となった。

４．研究成果

日本語で初めての言語資源であるやさしい日本語対訳コーパス 50,000 文と 2,000 語辞書を作成した。これらの言語資源によって日本語の自動平易化、及び平易な文と英語との機械翻訳の研究は大きく進展するものと確信している。これら言語資源はいずれも一般公開を行った。またクラウドソーシングによってコーパスの規模拡大も行った。

さらに、本研究の副産物として図１で示したやさしい日本語チェッカーも公開した。これに類するツールは少なくとも日本語では存在しないが、ある一定の基準に従って入力文をやさしい日本語に書き換えたいという潜在需要は膨大と予想する。チェッカーは本来我々の作業のための内部ツールであるが、自然言語処理の社会への貢献としてコーパス、辞書と同様に公開する。

この他、本報告では触れなかったが、本研究課題において日本語の平易化に関する様々な研究を実施し、いずれも国際会議または全国大会において研究発表を行った。この結果学会発表（国際会議における英語発表を含む）を１２件行うことができた。

（参考文献）

[庵 2013] 庵 功雄，イ ヨンスク，森 篤嗣 (編集)．「やさしい日本語」は何を目指すか: 多文化共生社会を実現するために．ココ出版 (2013)

[梶原 2014] 梶原 智之，山本 和英．高頻度語は平易語なのか？ NLP 若手の会 第 9 回シンポジウム，(発表 P02) (2014)

[梶原 2015] 梶原 智之，山本 和英．語釈文を用いた小学生のための語彙平易化．情報処理学会論文誌，Vol.56，No.3，pp.983-992，情報処理学会 (2015)

[国立国語研究所 1984] 国立国語研究所．日本語教育のための基本語彙調査．秀英出版 (1984)

[土居 1933] 土居光知．基礎日本語．六星館 (1933)

[松田 2010] 松田 真希子，児玉 茂昭，竹元 勇太，石坂 達也，森 篤嗣，川村 よし子，山本 和英．コーパスの異なりと単語親密度を活用した日本語共通基礎語彙の抽出．言語処理学会第 16 回年次大会，pp.579-582 (2010)

[李 2013] 李真奈見，山本 和英．「やさしい日本語」変換システムの試作．言語処理学会第 19 回年次大会，pp.678-681 (2013)

[山本 2016] 山本 和英，高橋 寛治，椙澤 優希，西山 浩気．日本語解析システム「雪だるま」第 2 報 ～ 進捗報告と活用形態素の導入 ～．電子情報通信学会 テキストマイニングシンポジウム，信学技報，Vol.116，No.213，pp.63-68 (2016)

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 Takumi Maruyama and Kazuhide Yamamoto
２．発表標題 Lexical Substitution is Practical for Rare Word Simplification
３．学会等名 The 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32), no page numbers（国際学会）
４．発表年 2018年

１．発表者名 Takumi Maruyama and Kazuhide Yamamoto
２．発表標題 Simplified Corpus with Core Vocabulary
３．学会等名 The 11th International Conference on Language Resources and Evaluation (LREC 2018), pp.1153-1160（国際学会）
４．発表年 2018年

１．発表者名 Akihiro Katsuta and Kazuhide Yamamoto
２．発表標題 Crowdsourced Corpus of Sentence Simplification with Core Vocabulary
３．学会等名 The 11th International Conference on Language Resources and Evaluation (LREC 2018), pp.461-466（国際学会）
４．発表年 2018年

１．発表者名 稲岡 夢人，山本 和英
２．発表標題 日本語文法平易化コーパスの構築
３．学会等名 言語処理学会第25回年次大会，pp.375-378
４．発表年 2019年

１．発表者名
角張 竜晴，山本 和英

２．発表標題
クラウドソーシングによる大規模なやさしい日本語換言辞書の構築

３．学会等名
言語処理学会第24回年次大会

４．発表年
2018年

１．発表者名
角張 竜晴，山本 和英

２．発表標題
やさしい日本語格フレームの構築による係り受け解析

３．学会等名
言語処理学会第24回年次大会

４．発表年
2018年

１．発表者名
Takumi Maruyama and Kazuhide Yamamoto

２．発表標題
Sentence Simplification with Core Vocabulary

３．学会等名
Proceedings of the International Conference on Asian Language Processing (IALP 2017)（国際学会）

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考