やさしい日本語対訳コーパスの構築

山本 和英, 丸山 拓海, 角張 竜晴, 稲岡 夢人,
小川 耀一朗, 勝田 哲弘, 髙橋 寛治
長岡技術科学大学
yamamoto@jnlp.org

1．自動平易化の言語資源が少なすぎる

「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（例えば[庵 2013]）。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（例えば[梶原 2015]）。

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[1] 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[2] 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

2．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[杢 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

	[杢 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]

[1] https://simple.wikipedia.org/
[2] http://www3.nhk.or.jp/news/easy/
[3] 3.1 節で述べるように、英訳は使用テキストに付与されており、我々は

3．作業

3.1　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[4] を採用した。このテキストは田中コーパス[5]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。作業は学生５人（山本、髙橋を除く本原稿著者５名）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。書き換え作業は主に 2016 年 11 月～12 月の２か月間で実施した。

3.2　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、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[6]
(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[7]

3.3　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、

付与していない。

[4] https://github.com/odashi/small_parallel_enja

[5] http://www.edrdg.org/wiki/index.php/Tanaka_Corpus

[6] 句読点などの記号は UniDic 上では「単語」であるが、このような議論をする際には単語と考えないほうが自然であろう。

[7] 例えば、例文中に含まれる（UniDic 辞書にない）英単語などがこれに該当する。

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。

3.4　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。

図 1 ：やさしい日本語チェッカー

4．所感

本対訳コーパスを用いることで様々な言語研究での分析及び言語処理が可能であるが、これらは今後改めて報告することとし、本稿では所感を述べる。

表 2 に、今回の作業で追加された単語と書き換え文の例を示す。表中の順位とは、BCCWJ における頻度順位である。3.4 節で述べたように、初期状態として頻度上位 2,000 語を登録し、ここから追加／削除を行っていったが、表 2 に示すように頻度上位語でなくても日本語表現の根幹をなす語があることが分かる。従って、基礎語彙の選定を頻度のみで行うことは不可能であると同時に、本稿で行ったような手作業がどうしても必要と考える。この観察は、[梶原 2014]で議論しているように高頻度語が必ずしも平易な語ではなく、低頻度であっても平易な語が多く存在するという結果を追認している。

なお、この傾向は頻度を取った BCCWJ に何らかの偏りや問題があることを意味するものではなく、おそらくどのようなコーパスを用いたとしても（傾向は違うであろうが）何らかの重要語の漏れが一定割合存在するであろう。従って、基礎語彙選定に際して[松田 2010]で議論しているように、頻度情報は基礎語彙選定の大きな情報源にはなるが、単独で用いるのは不十分である。

一方、作業において初期高頻度 2,000 語から削

表２：やさしい語彙に追加された単語の例

順位	単語	書き換え前の文
3169	青い	彼女の青い靴は服によく合っている。
3321	貸す	彼女はあなたに本を貸すだろう。
4628	泳ぐ	彼は上手に泳げる。
5370	アレルギー	魚アレルギーなんです。
6481	こんにちは	「世界のみなさん、こんにちは」
7565	宿題	あなたは英語の宿題をもう終えましたか。

除された単語は、3.2 節で述べたような固有名詞等の対象外となる語か、他の語句で代替できる同義語や表記ゆれなどであった。

５．おわりに

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準備ができ次第一般公開を行う。また機会があれば規模拡大を目指す。

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早期に公開する。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。

謝　辞

本研究は、平成 27～31 年科学研究費補助金基盤（B）課題番号 15H03216、課題名「日本語教育用テキスト解析ツールの開発と学習者向け誤用チェッカーへの展開」の助成を受けています。

使用したツールと言語資源

1. 現代日本語書き言葉均衡コーパス（BCCWJ), Ver.1.1, 国立国語研究所
2. 形態素解析器 MeCab Ver.0.996, http://taku910.github.io/mecab/
3. UniDic: http://osdn.jp/projects/unidic/

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)

[杢 2013] 杢真奈見, 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, 桾澤 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)