一般のニュースからやさしい日本語ニュースへの書き換えの分析

後藤 功雄　　熊野 正　　田中 英輝

NHK 放送技術研究所

1　はじめに

日本在住の外国人へニュースを分かりやすく伝えるために，ニュースをやさしい日本語で提供する研究を進めている．NHK ではこれまで，やさしい日本語ニュースへの書き換えの基準を作成し [2]，NHK のインターネットサービス「NEWS WEB EASY」[*1]でこの基準に沿って書き換えたニュースの配信を開始している．現在の書き換えは，支援システム [4] を利用して人手で行っており，一日あたり 5 記事のニュースを配信している．この支援システムに自動処理技術を導入することで作業を効率化して，より多くのやさしい日本語ニュースを配信することを目指している．

自動処理技術では具体的にどのような書き換え処理が必要であるかを調べるために，実際にやさしく書き換えられたニュースを分析した．本稿は，この分析結果を報告する．以下，やさしい日本語ニュースへの書き換え作業の実施方法，今回分析の対象とした書き換え，用いたデータ，そして分析結果について述べる．

2　やさしい日本語ニュースへの書き換え作業

やさしい日本語ニュースは，日本語をやさしく書き換える他に，読む量が少なくてすむように重複や周辺的な情報を削除したり，ニュースとして適切な構成となるように編集して作成する．この作業をやさしい日本語を学んだ日本語教師と記者（経験者）が共同で実施している．日本語教師が主に表現をやさしくし，記者が内容に関わる書き換えや削除などの編集を行っている．

3　分析の対象

普通のニュースからやさしい日本語ニュースに自動的に書き換える処理を導入することで作業を支援するためには，前記の 2 者の書き換えの自動処理を開発する必要がある．このうち，まず日本語教師が行っている書き換えに対して取り組むことにした．そこで，日本語教師が行った書き換え内容を分析する．人手でどのような書き換えが実際に行われたのかを分析して把握できれば，自動処理の要件が分かるため有用である．

書き換えの自動化処理では，はじめに，大きな語順の変更を伴う書き換えを主にルールを用いて行い，次に機械翻訳の技術などを用いて表現をやさしく書き換えるという 2 段階の処理[*2]を考えている．この 1 段階目の処理に関わる語順の変更を伴う書き換えを主に分析する．また，2 段階目の処理に関わる書き換えについても調べる．

4　データ

「NEWS WEB EASY」でやさしい日本語ニュースを配信するための作業で利用または構築された以下のデータを用いた．

- 書き換え元の一般のニュース（以下，ORG と

[*1] http://www3.nhk.or.jp/news/easy/

[*2] このように処理を 2 段階にすることは統計的機械翻訳でのプレオーダリング手法 [1] と同様な考え方である．

表 1　語順が変わった書き換えの分類と件数

分類	件数
連体節を文へ変更	39
格の順番	21
修飾方向の反転	13
文分割での補完	13
格の変更	10
名詞連続	8
複合名詞	7
品詞の変更	7
連体節（形式名詞）	5
サ変名詞の動詞化	4
量の概念	3
難しい表現の抽出	3
文末から文頭への移動	3
難しい節	2
その他	66

表記する.）

- 一般のニュースを日本語教師がやさしく書き換えたニュース（以下，EASY と表記する.）

これらのニュース 50 記事対の本文に対して，人手で記事対間の単語対応を付与した．ORG の 50 記事中の文数は 309，EASY の 50 記事中の文数は 530 であった.

5　分析結果

語順に関わる書き換えおよびその他の書き換えを単語対応を利用して調べた結果を示す.

5.1　語順に関わる書き換え

50 記事対において，語順が入れ替わっているものを調べて分類した．この際に要約や再編が大幅に行われているものは除外した．各分類の説明と例を以下に示す．例において，各分類の要因による語順が入れ替わった部分を赤文字と青文字で表している．また，各分類の書き換え頻度を表 1 に示す.

連体節を文へ変更　連体節が独立文になる場合.

（被修飾名詞が連体節の動詞より前に複製される.）

[例 1] NASA は去年１１月に**打ち上げた火星探査機「キュリオシティー」**を日本時間の来月６日、午後２時半過ぎに火星に着陸させます。

↓

NASA は去年１１月に**「キュリオシティー」という火星探査機を打ち上げました**。
「キュリオシティー」は、日本時間の来月６日、午後２時半すぎに、火星に着きます。

（さらに，文頭以外の連体節の場合は，連体節部分が元の文の後または前へ移動することが多い.）

[例 2] この遺伝子をマウスの脳の**記憶などをつかさどる海馬という部分に大量に組み込みました**。

↓

その遺伝子をマウスの脳の中の**海馬という部分にたくさん入れました。**
海馬は記憶などをコントロールする働きがあります。

格の順番　格の順番が変わる場合.

[例 3] **富士山が大規模に噴火した場合、山梨県は**・・・

↓

山梨県は、富士山が大規模に噴火した場合・・・

修飾方向の反転　修飾する方向が異なる語彙への変更に伴う語順変更の場合.

[例 4] **半年余り**で → **約半年**で

格の変更　格の変更を伴う語順変更の場合.

[例 5] **笑った顔に見える埴輪が、パリで来月開かれる展覧会で**展示されることになり、

↓

パリで来月開かれる展覧会に笑った顔に見える埴輪を出すことになりました。

[例 6] **難しい役柄を表現力豊かに**演じ

↓

すばらしい表現力で難しい役を演じて

補完　文の分割に伴い，提題などの要素が複製されて補完される場合．

[例 7] **この有料サービスは、・・・の現在地を地図で把握できるというもので**、１０日から運用が始まりました。

↓

この有料サービスは、・・・が今いる場所を地図で知ることができるというものです。
このサービスは１０日から始まりました。

複合名詞　複合名詞の変更に伴う語順変更の場合．

[例 8] 家庭の**電力消費** → 家庭で**使う電力**

[例 9] **新たな被害**想定を

↓

被害について、新しい予想を

名詞連続　名詞間の関係の明示に伴う語順変更の場合．

[例 10] **火星探査機「キュリオシティー」**

↓

「キュリオシティー」という火星探査機

品詞の変更　品詞と修飾先が異なる語への変更に伴う語順変更の場合．

[例 11] **具体的な場所を**尋ね → **場所を細かく**聞くと

連体節（形式名詞）　形式名詞を修飾する連体節の動詞が後ろへ移動する場合．

[例 12] 研究を**行ったの**は、**・・・のグループ**です。

↓

この研究は**・・・のグループ**が**行いました。**

サ変名詞の動詞化　サ変名詞を動詞に変更する場合．

[例 13] **着陸**まであと**２０日余り**

↓

約２０日で**着きます。**

量の概念　量の概念を表す表現の変更に伴って語順が変わる場合．

[例 14] **避難が必要な人数**の試算を進め

↓

どのくらいの人が避難する必要があるか計算を進め

文末から文頭への移動　文末表現（の一部）が文頭に移動する場合．

[例 15] **・・・に提供することを明らかにし**、

↓

発表によると、・・・に渡したりしました。

[例 16] **・・・を考えているという調査結果がまとまりました。**

↓

その結果、・・・を考えてみようとしていることが分かりました。

難しい表現の抽出　文中の難しい表現の意味を別の文で説明する場合．

[例 17] と述べ、**異例の謝罪を行いました**が、

↓

と言って**謝りました。**
このようなことは今までにありませんでした。

難しい節　難しい節の概念をやさしく変更する場合．

[例 18] １１年８か月ぶりの**円高水準を更新しました。**

↓

２０００年１１月から今までで**いちばんの円高**です。

その他　複雑または背景知識を要する高度な書き換えで分類が難しい場合，または該当する分類の頻度が 1 回の場合など．

[例 19] IT **企業の間では**、**この分野を強化する**動きが広がっています。

↓

地図に力を入れる IT **企業が**増えています。

[例 20] **電力会社が提供する**需給状況のデータに基づいて、**電力需要が少なく**価格が安い**時間**帯に

↓

電力が足りている時間は、**電力会社の**データから分かります。

5.2　その他の書き換え

文がどの程度分割されているかを調べた．結果を表 2 に示す．[*3]

Easy に出現する語[*4]のうち，元の表現と一致する語，元の表現と一致しない語，新たに追加された語の割合を表 3 に示す．また，Org から削除された語[*5]の Org 中の割合は 0.08 であった．[*6]

受動態が能動態に変わった割合および，使役[*7]が使役でなくなった割合を表 4 に示す．[*8]

表 2　1 文が何文に書き換えられたか

	割合
1 文が 1 文	0.40
1 文が 2 文	0.46
1 文が 3 文	0.11
1 文が 4 文以上	0.02

表 3　書き換えられたニュース中の語の種類

	割合
元の表現と一致する語	0.56
元の表現と一致しない語	0.37
新たに追加された語	0.07

表 4　元ニュース中の受動態と使役の変更割合

	変更割合（頻度）
受動態から能動態	0.92 (139/151)
使役から非使役	1.0 (11/11)

6　おわりに

NEWS WEB EASY サービスのための作業で構築されたデータを用いて，ニュースをやさしく書き換える際にどのような書き換えが行われたかを分析した．語順の変更を伴う書き換えで最も頻度が高かった要因は連体節の文への変換であることが分かった．今後は，この結果を参考にして書き換えの自動処理を検討する予定である．

参考文献

[1] Fei Xia and Michael McCord. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of Coling*, pp. 508–514, 2004.

[2] 田中英輝, 美野秀弥．やさしい日本語によるニュースの書き換え実験．情報処理学会研究報告，Vol.2010-NL-199, No.11, 2010.

[3] 田中英輝, 美野秀弥, 越智慎司, 柴田元也．やさしい日本語ニュースの公開実験．NHK 技研 R&D, No.139, 2013.

[4] 美野秀弥, 田中英輝．ニュース原稿のやさしい日本語ニュースへの書き換え支援ツール―日本在住外国人のために―．映像情報メディア学会年次大会, No.18-6, 2012.

*3 この結果は日本語教師による書き換えでの文の分割であるが，さらに記者による書き換えも行った記事の分析結果もある [3]．それによると，元の記事に対して書き換え後の記事の平均文数は 1.13 倍，記事の平均文字数は 0.73 倍，文の平均文字数は 0.64 倍である．

*4 形態素を語の単位とした．形態素解析には IPA 品詞の辞書と形態素解析器の MaCab を用いた．

*5 内容にあまり関わらない表現は削除されることがある．[例 20] はそのような例でもある．

*6 リード文（ニュースの第 1 文で，概要を説明している．）は本文と内容が重複しているために削除される場合がある．主に重複以外の理由で削除される語の割合を調べるため，リード文である第 1 文以外の本文で削除された語の割合を調べた．

*7 IPA 品詞の辞書を使って MeCab で解析した結果において，品詞が「動詞, 接尾」で基本形が「せる」または「させる」の形態素のみを対象とした．

*8 Org 中の該当箇所が Easy に含まれる場合での割合．[例 5] は受動態が能動態に変わる例，[例 1] は使役が使役でなくなる例でもある．