ニュースの文末表現調査と外国人の理解度に関する大規模アンケート

田中 英輝　　熊野 正　　後藤 功雄
NHK 放送技術研究所
{tanaka.h-ja, kumano.t-eq, goto.i-es}@nhk.or.jp

1 はじめに

NHK ではニュースサイト NEWSWEB EASY でやさしい日本語のニュースを提供している．このニュースは通常のニューステキストを旧日本語能力試験の 3 級合格から 2 級準備程度の範囲になるように人手で書き換えて作成している．著者らはその効果を確認するため，外国人[1]を使ったニュース記事全体の理解度調査を行ってきた [1][2]．これらはサービスの効果を確認するには有効だが，やさしい日本語の書き方の参考となる情報を得るには表現に着目した調査を行っていく必要がある．そこで今回，特にニュースの特徴の出やすい文末表現に着目し，これを 16 年分のニュースコーパスから収集すると共に，分かりやすさに関する大規模アンケート調査を行った．

2 文末表現の収集

本稿では文末表現を文の末尾の付属語列とする．これを，以下の手順に従って，通常ニュースの原稿データベース（1997 年 11 月から 2013 年 7 月まで）および，やさしい日本語ニュースの原稿データベース（2011 年 4 月から 2012 年 4 月まで）から抽出した（表 1）．

- 文末表現抽出
  上記のニュースを MeCab で解析し，品詞情報を使って文末から最長の付属語列を抽出した．このとき,「議長には○○氏が 就任する見込みです」のような「動詞 ＋ 名詞＋です」の形を取る「名詞＋です」構文はニュースに特徴的なため収集することにした．なお通常ニュースからは頻度 100 以上の文末だけを，やさしい日本語ニュースからはすべての文末を抽出した．通常ニュースに着目すると，約 400 万の文の文末は頻度 100 以上に限ると 477 種類しかなかった．このことからニュースの文末はかなり定型的なことがわかる（表 1）．

表 1: データベース規模と文末表現数

	普通	やさしい	合計
データベースの文数	3,937,214	20,616	-
文末表現数	477	775	1,063
文法タグ列付与数	-	-	841

- 文法タグ列の付与
  表層の文末の文法機能を分析するため，各文末表現に文法的な機能を表すタグ（文法タグ列）を付与した．個別の文法タグには表 2 に示す 28 個を設定した．また形態素解析結果に文法タグ列を自動的に付与するため，正規表現からなるスクリプトを作成した．普通のニュースとやさしいニュースの文末表現 1,063 個の一部は単純な名詞述語（テストです）や本動詞に丁寧の助動詞だけがついた表現（行きます）であった．これらは単純すぎるため調査対象外と考え，タグ列を付与しなかった．この結果 1,063 表現のうち 841 表現にタグ列が付与された（表 1）．

3 アンケート

3.1 調査方針

外国人日本語学習者が，文末表現をどれくらい理解できるか調査するにあたり次のような方針を立てた．

- 分析の視点
  1. 文末の理解度は文法機能ごとに議論できると便利である．そこで文法タグ列の情報を使って調査表現を決めることにした．文末表現の文法タグ列を観察すると，通常ニュー

[1]日本語を学習している日本語非母語話者を外国人と呼ぶ．

表 2: 個別文法タグと表層表現例（辞書形）

文法タグ	表層表現例	文法タグ	表層表現例	文法タグ	表層表現例
parallel	たりする	hope	たい	need	なければならない
order	してください	will	しよう	percept.	と見ている
amb.-percept.	としている	selection	だけ	pas.-percept.	と見られる
prohibition	てはいけない	invitation	ましょう	guess	そうだ
change	ことになる	example	など	give-get	てもらう
probability	はず	noun	見込みだ	pas.-amb.-percept.	とされている
aspect	ている	reason	ためだ	objectivity	ものだ
hearsay	ということだ	question	ではないか	explanation	のだ
change-guess	ことになりそうだ	causative	させる	pas.	られる
nominalization	もの				

amb. = ambiguous, pass. = passive, percept. = perception を表す

スとやさしい日本語ニュースでかなり傾向が異なっていた．そこでタグ列をどちらのニュースに出やすいかで分類した上で表現を抽出することにした．

2. 一般に文末の文法タグ列には個別のタグが複数含まれる．このタグの数を文法タグ列の長さとする．これは文末表現の長さにも対応する．タグ列は短い方が分かりやすいなど長さに依存することが予想される．そこで文法タグ列の長さも一つの視点とした．

- 調査対象者
外国人の文末表現の理解度を調査するには，外国人を対象に調査をしたい．しかし彼らを多数，特に様々な日本語レベルにわたって集めるのは難しい．さらに，調査対象の文末表現は名詞などに較べて複雑な文法機能を持つため，特に初級レベルの人にその理解を問うことは難しいだろう．以上の理由から，日本語教育に経験の深い日本語教師などを対象にアンケートの形で調査することにした．またできるだけ多くのサンプルを収集するためネット調査会社経由で調査した．

3.2 調査表現の抽出

分析の視点の 1 と 2 を反映するため，文法タグ列を「出現しやすいニュースのタイプ」と「長さ」で分類して表現を抽出した．以下に手順を示す．

- 文法タグ列のニュースタイプ決定
841 の文末表現に対して付与された文法タグ列は 146 種類であった．これらのタグ列が出やすいニュースのタイプ（通常ニュース，やさしい日本語のニュース）を決定した．これを「文法タグ列のニュースタイプ」と呼ぶ．文法タグ列の出現頻度はそれが付与された文末表現の出現頻度を元に計算できる．そこで各文法タグ列に対して 2 つのニュースタイプでの相対出現頻度を計算しオッズ比を求めれば，ニュースタイプを決める指標が得られる [2]．今回はオッズ比が 1 より大きい文法タグ列のニュースタイプを通常ニュース，1 以下の文法タグ列のニュースタイプをやさしい日本語ニュースとした．

表 3: 文法タグの数の分布

長さ	1	2	3	4
通常ニュース	9	35	21	0
やさしい日本語ニュース	13	43	22	3

- タグ列の長さとニュースタイプによる分類
146 種類の文法タグ列の分布を長さとニュースタイプで分類する．結果を表 3 に示す．

- 文末表現の抽出
表 3 の 2 つのニュースタイプ（通常ニュースとやさしい日本語ニュース）の長さ 1 および 3 の部分に着目する．そして，それぞれの部分に分類される文法タグ列が付与された文末表現を各 13 個ずつ，合計 52 個選択した．表 3 の「ニュースタイプが通常ニュースで長さ 1」以外の部分では文法タグ列から 13 個をランダムに選択した．そしてこれらのタグ列を持つ文末表現を頻度の高い順に一つずつ選択した．表 3 の長さ 1 の通常ニュース部分はタグ列が 9 種類しかないため，前記の操作を繰り返して 13 個の表現を得た [3]．

[2] あるタグ列の普通のニュースでの相対頻度を p，やさしい日本語ニュースでの相対頻度を q とするとオッズ比は $\frac{p}{1-p}/\frac{q}{1-q}$ で計算する．また値は $[0, \infty)$ の範囲を取る．∞ は通常ニュースのみで出現，0 はやさしいニュースのみで出現することを示す．

[3] このため，この部分は同一の表現が選択されている．ただし，アンケート時の例文は異なるようにした．

表 4: アンケートに使った文末表現（表層形）

長さ 1 の表現		長さ 3 の表現	
普通	やさしい	普通	やさしい
としています	になりそうです	ている もの と見られます	になっ ていた のです
方針です	はずでした	したりする ことにし ています	たりし ている ためです
とされています	てもらいました	させ たい としています	になる などし ています
と見ています	などをしました	させ ていた ということです	ならない ようにする ためです
させます	でしょうか	になっ ている ものです	なっ たりした ためです
とされています	てはいけません	になっ ていた ということです	られ ている ためです
ということです	てあります	していく ことにし ています	てみ てもらっ ています
と見ています	になりました	ている ことなどによる ものです	なっ ている ためです
見通しです	だけでした	れ ている ものです	になる だろう ということです
と見られています	てください	になる もの と見られます	てもらう ことになっ ています
としています	たりします	れ ていた ということです	られる ことになっ たのです
ものです	からです	れる ことになっ ています	にする ことになっ ています
されませんでした	なければなりません	など となっ ています	てはいけない ことになっ ています

表 5: 選択番号と受検レベル

選択番号	受検レベル	選択番号	受検レベル
1	N5	4	N2
2	N4	5	N1
3	N3	6	上記以上

表 6: 条件ごとの難しさの平均値

	ニュースタイプ	長さ	平均	誤差
1	やさしい	1	3.053	0.019
2	やさしい	3	3.713	0.018
3	通常	1	3.649	0.019
4	通常	3	3.812	0.019

表 4 に今回のアンケートに使った文末表現の一覧を示す．これらの表現に対して「読んですぐに理解できるレベル」を質問した [4]．レベルは日本語能力試験 [5] を受検して合格できる級とした．表 5 で示すように，これらの級の初級から順に 1 から 6 までの選択番号を与え，大きいほど表現が難しくなるようにした．評価者はこの中から番号を一つ選択する．

アンケートの問題は，1) 評価する表現，2）文法特徴，3）例文，の 3 つ組で提示した．評価は文末表現だけを対象として，それ以外の部分の理解は問わないとした．問題例を示す．

> ニュースの文末表現「させることにしています」についての質問です。
> これには次のような文法特徴があります。
> 使役表現「させる」＋変化表現「ことにする」＋アスペクト「ている」
> また、次のような文の中で使います。
> 「環境省はこの指針を来年度予算案の概算要求に反映 させることにしています。」

4 調査と分析

52 個の質問をネット調査会社経由で日本語教育関係者に配布し，390 名から有効と判断できる回答を得た．

本調査は 3.1 節に示した方針に沿った実験計画に従っている．具体的には下記に述べる被験者内 2 要因の実験計画である．

- 従属変数（分析対象）
  各表現に与えられた 6 段階の数値（表 5）．
- 要因
  文法タグ列のニュースタイプ（普通のニュース，やさしい日本語ニュース），タグ列の長さ（1，3）の 2 要因がある．すべて被験者内要因である．各要因は 2 水準である．

要因の水準ごとの平均値と標準誤差を表 6 に，対応するグラフを図 1 に示す．グラフの x 軸は文法タグ列の長さで y 軸は難しさの平均値である．またオレンジのデータは，やさしい日本語に出やすい文末表現，青は通常ニュースに出やすい文末表現を示す [6]．

表 6 の数値を見ると，平均値の最低が 3.05（ニュースタイプ：やさしいニュース，長さ：1）で最大が 3.81

[4] 補助的に「表現を初めて学習する，あるいは教えられるレベルはどれか」という質問も行い同じ 6 段階の評価を得たが今回は分析していない．

[5] 日本語を学習する外国人が受検する試験．入門の N5 から最上級の N1 の 5 段階のレベルがある．

[6] 正確には，付与されたタグ列のニュースタイプがそれぞれやさしい日本語，通常ニュースの表現である．以下の議論では簡便性のため，やさしい日本語ニュースの文末表現，通常ニュースの文末表現という．

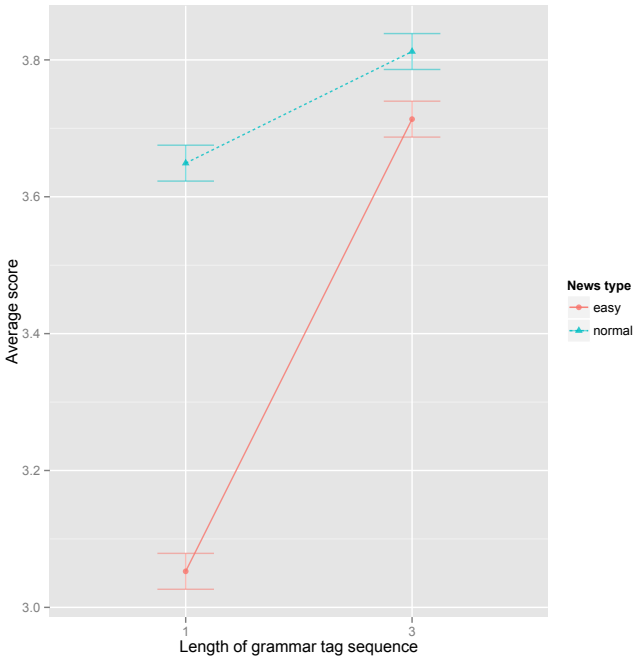

図 1: 条件ごとの難しさの平均値　グラフ

（ニュースタイプ：通表ニュース，長さ：3）であった．著者らのやさしい日本語は，現在の能力試験では N3 合格を少し超える程度と考えている．その意味では，今回の数値が 3.5 以下であればやさしい日本語としてそのまま使えると考えてよい．そうするとこれに合致するのは（ニュースタイプ：やさしいニュース，長さ：1）の文末表現だけとなる．

やさしい日本語に出やすい文末表現であっても長くなる（タグ列の長さが 3 になる）とこの基準を超える．

グラフを観察すると，タグ列の長さが 1 から 3 になる，すなわち文末表現が長くなるとどちらのニュースであっても難しさは増加している．ただし，この傾向はグラフの 2 直線の傾きの差から見えるように，やさしい日本語の文末表現の方で顕著に見える．

このことを検定するために分散分析を行った[7]．表 7 に結果を示すようにすべての項目で有意差が検出された．すなわち上記の観察に相当する表 7 の 3 行目の「ニュースタイプと長さの交互作用」の部分でも有意差が検出された．

以上のことから，やさしい日本語のニュースを書く場合，文末表現に文法機能を重ねることは，どういうものであれ好ましくない可能性が高いと結論できる．

なお通常ニュースでは長さが変化しても難しさがそれほど上昇していない．理由は今後検討したいが，可能性の一つに，通常ニュースの文末表現では，一つの

[7]R の ez パッケージを使用した．

表 7: 分散分析結果

要因	$F(1,\ 389)$	$p \ll .05$
ニュースタイプ	158.31	*
長さ	164.88	*
ニュースタイプ：長さ	173.83	*

文法タグしか持たないような短い表現でもかなり難しいため，それらが重なっても難しさが変化しにくいというのがあるかもしれない．

5　議論

今回の調査では外国人の理解度を日本語教育関係者の直感を通じて間接的に計測した．これが実際の外国人の理解度と一致しているか今後測定する必要があると考えている．

今回得られた平均値の条件による差は，著者らの想定より小さかった．今回はネットを通じた調査であるため，すべての回答を信用できるとは限らない．分析の前段階として，全く同じ選択肢を選んだ人のデータを除外するなど，クリーニングを行ったが，さらなるクリーニングが必要かもしれない．ネット経由の調査の手法についても今後検討したい．

6　おわりに

やさしい日本語を書くための指針を得るためにニュースの文末表現を収集した．またそれらの外国人の理解度を日本語教育関係者のアンケートによって調査した．この結果，やさしい日本語のニュースに使われている文末表現は，短ければ十分やさしいが，長い場合にはかなり難しくなっている可能性があること，通常ニュースに現れている文末表現は短くてもかなり難しい可能性があることが分かった．

参考文献

[1] 田中英輝, 美野秀弥. 「やさしい日本語」ニュースの理解度テスト–ニュースのための「やさしい日本語」の設計に向けて–. 信学技報, No. NLC2011-22, pp. 1–6, 2011.

[2] 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サイト「NEWSWEB EASY」の評価実験. 情報処理学会 研究報告, No. 2012–NL–209, 9, 2012.