『日本経済新聞記事オープンコーパス』に対する言語受容情報付与
Collecting Language Reception Information for the 'Nikkei Newspaper Article Open Corpus'

浅原 正幸♭, 加藤 祥♯, 高松 純子♮
Masayuki Asahara, Sachi Kato, Junko Takamatsu
♭ 国立国語研究所, ♯ 目白大学, ♮ 日本経済新聞社
♭ National Institute for Japanese Language and Linguistics,
♯ Mejiro University, ♮ Nikkei Inc.
♭masayu-a@ninjal.ac.jp

概要

本研究では、オープンデータである『日本経済新聞記事オープンコーパス』に対して、言語受容情報を収集した。一般の方が読んで語・文節単位にどのような印象を受けるかをアンケート調査するとともに、テキストの読み時間を収集した。さらに印象評定情報と読み時間を対照し、自然さ・わかりやすさ・古さ・新しさ・比喩性の印象と読み時間の関係について検討を行った。

キーワード：オープンコーパス, 読み時間, 印象評定

1. はじめに

認知科学の分野でも研究データのオープンアクセス化が進められ、論文の投稿と同時にデータの共用化が求められている。しかし、テキストなどを刺激とする言語受容過程の分析においては、刺激となるテキストコーパスの著作権の問題があり、データの共用化が困難であった。著作権法には、著作物を公開する場合には著作権者の許諾が必要である旨が規定されており、良質なテキストコーパスの共用化には、元テキストの著作権者の合意が必要である。『日本経済新聞記事オープンコーパス』は、2023 年 3 月に日本経済新聞社により公開された本文記事データも CC BY-NC-SA で公開されているオープンデータである。2013 年 1-2 月の日本経済新聞社朝夕刊 96 記事に対して、国立国語研究所により UniDic 形態論情報（短単位・長単位）[1]・文節係り受け情報 [2] が付与され、Universal Dependencies 互換の単語係り受け情報 [3] も付与されている。

本研究では同データに対して、印象評定情報と読み時間を付与した。いずれも Yahoo! クラウドソーシングにより実験協力者を募集した。印象評定情報は、一般の方が新聞記事データをどのように捉えるかをレーティング情報で収集した。読み時間は、逐次的に呈示する自己ペース読文法により、文節単位の呈示時間を収集した。本稿では、これら 2 つのデータを対照することにより、印象評定と読み時間の相関について検討する。

2. 『日本経済新聞記事オープンコーパス』

表 1 『日本経済新聞記事オープンコーパス』の統計

短単位形態素数	33,346
長単位形態素数	24,379
文節数	10,627
文	1,333
記事数	96

『日本経済新聞記事オープンコーパス』は 2013 年 1-2 月の新聞記事 96 記事を CC BY-NC-SA 4.0 ライセンスで利用可能にしたテキストコーパス[1]である。同データに国立国語研究所により UniDic 形態論情報（短単位・長単位）・文節係り受け情報・Universal Dependencies に基づく単語係り受け情報を付与され、これらの言語情報アノテーションは CC BY 4.0 ライセンスで利用できる。表 1 に『日本経済新聞記事オープンコーパス』の基礎統計を示す。

3. 言語受容情報

3.1 印象評定情報

印象評定情報は加藤らの先行研究 [4] にならい、国語研長単位・文節単位に呈示した言語表現について、自然さ・わかりやすさ・新しさ・古さ・比喩性の程度を確認し、0: まったく違う〜5: そう思うの 6 段階の評定情報を 1 表現あたり 20 人分収集した。長単位自立語は 11,074 表現、文節は 10,627 表現を対象とした。全体で異なり 3,499 人、延べ 431,160 人分収集した（調

[1] https://nkbb.nikkei.co.jp/alternative/corpus/

図 1　印象評定情報調査画面

査期間：2023 年 3 月 15-17 日)。図 1 に印象評定情報調査画面を示す。なお、本調査は次節の読み時間の調査の後に行った。

分析対象として、実験協力者ごとの各調査事項の評定値の標準偏差の平均を計算し、0.25 以下の方のデータを排除した[2]。

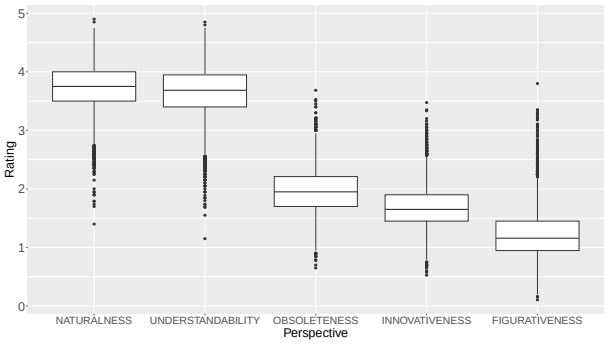

図 2　印象評定（自立語長単位の平均値）の分布（箱ひげ図）

[2]調査の際にすべて同じ値を入力する方を排除するため。

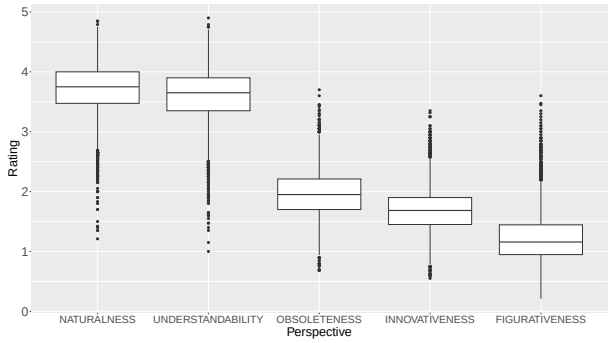

図 3　印象評定（文節単位の平均値）の分布（箱ひげ図）

図 2, 3 に収集した評定情報の分布を箱ひげ図を用いて示す。NATURALNESS が自然さを、UNDERSTANDABILITY がわかりやすさを、OBSOLETENESS が古さを、INNOVATIVENESS が新しさを、FIGURATIVENESS が比喩性を示す。新聞記事の特性から、自然さ・わかりやすさが高い傾向にある。

3.2　読み時間

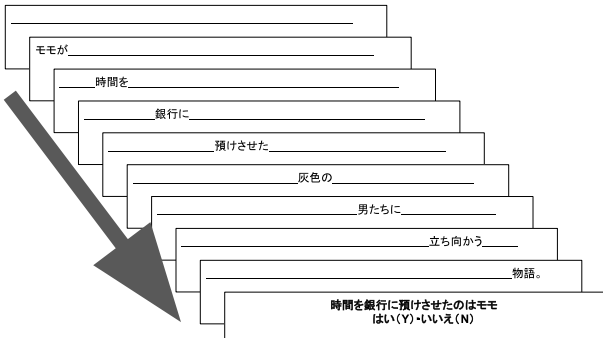

図 4　自己ペース読文法

また、文節単位に基づく自己ペース読文法 [5] により、文節読み時間を収集した（調査期間：2023 年 3 月 1 日-15 日：印象評定情報の調査より先に実施）。自己ペース読文法（図 4）は、実験協力者がスペースキーなどを押すことにより逐次的に文節を表示することにより、キーを押す間隔を文節呈示時間として評価し、読み時間を測る方法である。1 記事あたり 50-200 人分、全体で異なり 585 人、延べ 6,828 人分の読み時間データ、782,475 データポイントを収集した。

分析対象として、次の処理を行った。まず、調査の最後に新聞記事の内容把握の YES/NO 質問に答えてもらい、不正解だったものを分析対象外とした。次

に、実験試行ごとの平均読み時間が 150ms 以下もしくは 2000ms 以上のものを分析対象外とした。さらに、データポイントごとの読み時間が 100ms 以下もしくは 3000ms 以上のものを分析対象外とした。なお、ウェブを介しての調査のために通信時のトラブルにより欠損値（値が負になるもの）も存在する。欠損値も分析対象外とした。結果、562,719 データポイントを対照分析用のデータとした。

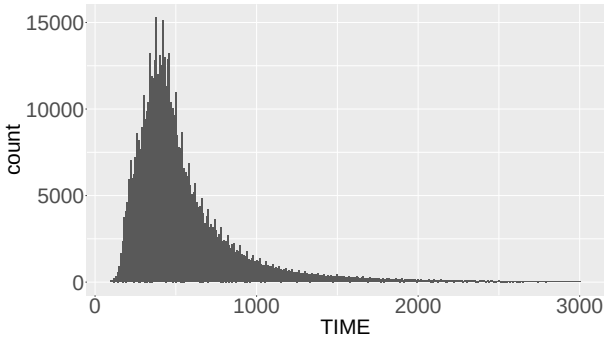

図 5　文節単位の読み時間の分布（分析対象のみ）

図 5 に、分析対象である文節単位の読み時間の分布を示す。データポイントが数百 ms に分布していることから、文節単位の読み時間が適切に収集できていることがわかる。

4. 対照分析

次にテキスト中の表現が読むヒトに与える印象が、読み時間とどのような関係にあるのかを検討するために、線形混合モデルに基づく分析を行う。対数読み時間 (LOGTIME) を次に示す要因により線形回帰を行う。当該文節の印象評定値（平均評定値）との関係を明らかにするために、NATURALNESS を自然さ、UNDERSTANDABILITY をわかりやすさ、OBSOLETENESS を古さ、INNOVATIVENESS を新しさ、FIGURATIVENESS を比喩性とする固定要因の傾きを検討する。同一文書内で実験が進むにつれて慣れていく効果として文書内の文節順 BIDC (Bunsetsu ID Cumulative) を、文節の文字数として CHARNUM を、当該文節に係る文節数として DEPNUM を固定要因とする。また、実験協力者ごとの揺れを統制するために、実験協力者 ID SUBJ を因子型のランダム要因として導入し、記事ごとの揺れを統制するために記事 ID ARTICLE を因子型のランダムとして導入する。以下に分析式を示す。

$$\text{LOGTIME} \sim \text{BIDC} + \text{CHARNUM} + \text{DEPNUM} + \text{NATURALNESS} + \text{UNDERSTANDABILITY} + \text{OBSOLETENESS} + \text{INNOVATIVENESS} + \text{FIGURATIVENESS} + (1|\text{SUBJ}) + (1|\text{ARTICLE})$$

表 2　線形混合モデルに基づく分析

	Dependent variable:
	LOGTIME
BIDC （進捗順）	−0.001*** (0.00000)
CHARNUM （文字数）	0.045*** (0.0002)
DEPNUM （係り受けの数）	−0.031*** (0.0004)
NATURALNESS （自然さ平均評定値）	−0.015*** (0.002)
UNDERSTANDABILITY （わかりやすさ平均評定値）	−0.027*** (0.002)
OBSOLETENESS （古さ平均評定値）	0.006*** (0.001)
INNOVATIVENESS （新しさ平均評定値）	−0.002* (0.001)
FIGURATIVENESS （比喩性平均評定値）	−0.023*** (0.001)
Constant	6.238*** (0.020)
Observations	556,214
Log Likelihood	−183,728.500
Akaike Inf. Crit.	367,481.100
Bayesian Inf. Crit.	367,615.800
Note:	*p<0.1; **p<0.05; ***p<0.01

分析は一旦回帰したうえで、3SD より外側のデータを排除したあとに再度回帰した。いずれもモデルも収束した。この外れ値処理により、562,719 データポイント中 6,605 データポイント (1.16%) を排除し、556,214 データポイントの結果を表 4. に示す。推定された係数とともに括弧内に標準誤差を示す。p 値の情報は *により示す (*p<0.1; **p<0.05; ***p<0.01)。

まず、前提として、進捗順 (BIDC) については、負の値を持つことから、進めば進むほど実験に慣れるとともに文脈が累積されることにより理解が促進し、読み時間が短くなる効果が確認された。文字数 (CHARNUM) については、正の値を持つことから、文字数が長くなればなるほど読み時間が長くなる効果が確認された。係り受けの数 (DEPNUM) については、負の値を持つことから、係り受けの数が多いほど予測が効いて読み時間が短くなる効果が確認された。この傾向は先行研究 [5] と同じであった。

次に印象評定と読み時間の関係について検討する。自然さ (NATURALNESS)・わかりやすさ (UNDERSTANDABILITY)・比喩性 (FIGURATIVENESS) に

ついては、平均評定値が高い文節ほど読み時間が短くなる傾向が見られた。また、若干ではあるが古さ (OBSOLETENESS) については、評定値が高いほど読み時間が長くなる傾向が確認された。

いずれの結果も予想可能な結果だと言える。しかしながら、これらの印象評定情報が係り受けなどの統語的な要因に匹敵するほどの要因であったことは重要な観点だと考える。作例に基づく読み時間の検証において、条件を統制した例文を数多く産出する必要がある。作例時には用いる語の頻度や親密度の統制のみならず、一般の方が表現に対してどのような印象を持つかについて統制することも重要であるだろう。

5. おわりに

本研究では、言語の生産実態としてのコーパスに対して言語の受容実態を収集しオープンデータ化した。以前の研究では、語の type に対する親密度として、知っている・書く・読む・話す・聞くの観点を収集した [6] が、文脈を含めた語の token に対する印象評定情報として、自然さ・わかりやすさ・古さ・新しさ・比喩性の情報を収集した。さらに同じコーパスに対して、延べ数千人規模の読み時間データを収集した。

過去の読み時間の分析においては、文法的な特徴量・意味的な特徴量・情報構造に基づく特徴量がどのように読み時間に影響するのかを重点的に分析が進められた。しかしながら、一般の方が感じる表現の印象に基づく読み時間の分析は進められていなかった。そこで、本研究では、文節単位に収集した印象評定情報が読み時間にどのような影響を与えるのかについて検討を行った。分析方法としては、文字数・文節出現順・係り受けの数・印象評定情報を固定効果とし、読み時間参加の実験協力者と記事をランダム効果とした一般化線形混合モデルにより対数読み時間を評価した。結果、自然さ・わかりやすさ・比喩性が読み時間を短くする効果があり、古さが読み時間を長くする効果が確認された。この結果から、文法・意味・情報構造などの差異に関する読み時間の分析においては、そのテキストが表出する印象の統制も重要であることが示唆された。

同様のデータとして、石原ら [7] はサムネイル画像も含めた新聞記事閲覧時間のデータセットと、マルチモーダル情報に基づく読み時間のモデリングを行っている。記事部分について、本データのように言語情報などを用いてより精緻化された分析を組み合わせて行うことにより、ヒトの新聞記事閲覧時の認知処理過程の解明が進むであろうと考える。

謝辞

本研究は国立国語研究所共同研究プロジェクト「実証的な理論・対照言語学の推進」および科研費 22H00663 によるものです。

文献

[1] Den, Y., Nakamura, J., Ogiso, T., Ogura, H., (2008) 'A Proper Approach to Japanese Morphological Analysis: Dictionary, Model, and Evaluation', In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, p.p.1019-1024.

[2] 浅原正幸・松本裕治, (2018) '『現代日本語書き言葉均衡コーパス』に対する文節係り受け・並列構造アノテーション', **自然言語処理**, Vol. 25, No. 4, p.p.331-356.

[3] 浅原正幸・金山博・宮尾祐介・田中貴秋・大村舞・村脇有吾・松本裕治, (2019) 'Universal Dependencies 日本語コーパス ', **自然言語処理**, Vol. 26, No. 1, p.p.3-36.

[4] 加藤祥・浅原正幸, (2021) 'IPAL 用言例文への印象評定情報付与と代表義・典型用例の抽出', **計量国語学**, Vol. 33, No. 3, p.p.178-193.

[5] Asahara, M., (2022) 'Reading Time and Vocabulary Rating in the Japanese Laugnage: Large-Scale Japanese Reading Time Data Collection Using Crowdsourcing', In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p.p.5178-5187.

[6] 浅原正幸, (2020) 'Bayesian Linear Mixed Model による単語親密度推定と位相情報付与', **自然言語処理**, Vol. 27, No. 1, p.p.133-150.

[7] 石原祥太郎, 中間康文, (2023) 'マルチモーダル機械学習によるニュース記事の閲覧時間予測', **第 37 回人工知能学会全国大会論文集**, 3Xin4-58, p.p.1-4.