日本語の語彙力と読み時間について

浅原 正幸
国立国語研究所
masayu-a@ninjal.ac.jp

概要

本研究では、ヒトの語彙力の差が読み時間にどのように影響を与えるのかについて検討を行う。具体的には、単語親密度調査に参加した方の評定値に対して一般化線形混合モデルを適用した際の調査協力者のランダム効果を語彙力と仮定する。その後、同調査協力者に自己ペース読文課題に参加してもらい読み時間を収集し、読み時間に対して一般化線形混合モデルを適用した際の語彙力の固定効果により、語彙力の差が読み時間に与える傾向を明らかにした。また、得られた傾向からサプライザルや言語モデルにおけるパープレキシティとの関連について考察を行う。

1 はじめに

本研究では、日本語の語彙力と読み時間の関係について検討する。以前の調査では、視線走査装置[1)]や自己ペース読文法 [1] を用いた、新聞記事を対象として読み時間データを収集 [2] し、記憶力を計測するリーディングスパンテスト [3] と平成年代に構築された語彙数判定テスト [4, 5] による実験協力者の能力差に基づく分析を行った [6]。得られたは語彙力が高い実験協力者の読み時間が長い傾向があるという結果であった。24 人の実験協力者の語彙力の分散が小さいことが考えられるが、実験協力者を研究室に呼んで、視線走査装置を用いて実験を行うことが困難となった。一方、英語においては、Amazon Mechanical Turk で被験者を募集して、読み時間を収集した Natural Story Corpus (NSC) [7] が構築されている。同様の試みとして、Yahoo! Japan の Yahoo! クラウドソーシングを用い被験者を募集したうえで、ibex [2)] を利用して、ウェブブラウザを介して自己ペース読文法により大規模に読み時間データを収集した [8]。読み時間収集対象者には、事前に単語親密度判定タスクに参加し、そのモデル化の副産物として得られる対象者の語彙力を得た [9]。これらの二つの調査結果を線形混合モデルで対照させたところ、**「語彙力が高い実験協力者の読み時間が短い」**傾向が確認された。従前の結果が語彙力の高い 1 人の実験協力者の行動に由来する可能性があり、大規模化した行動実験に基づく今回の分析により異なる結果が得られた。

また、サプライザル [10, 11] や言語モデルのパープレキシティとの関連 [12, 13, 14, 15, 16, 17, 18] についても考察を行う。

2 データの収集手法

2.1 語彙力データの構築

語彙力データは単語親密度データの収集 [19, 9] 時に得られる実験協力者の偏りを数値化したものを用いる。具体的には、2018 年より毎年『分類語彙表』の見出し語に対して、「知っている」「書く」「読む」「話す」「聞く」の程度を収集した。収集したデータを一般化線形混合モデルで単語親密度を 84,114 語のランダム効果として、語彙力を 6,732 人の実験協力者のランダム効果としてベイジアン線形混合モデルによりモデル化した。モデル化する際に、単語側の標準偏差を 1.0、実験協力者の標準偏差を 0.5 とした正規分布によりモデル化したものを用いる [3)]。

後述するとおり、単語親密度調査の協力者の一部の方が、次節に述べる読み時間データの収集に参加した。読み時間データの収集に参加した方の語彙力の分布を図 1 に示す。

2.2 読み時間データの収集

前節に示した単語親密度データ収集の 2020 年 10 月の調査に協力された方のうち、語彙数判定の分散が適切な 2,092 人を対象に、自己ペース読文法に

1） https://www.sr-research.com/eyelink-1000-plus/
2） https://github.com/addrummond/ibex/

3） 他のハイパーパラメータも試してみたが、この設定でのみ収束した。

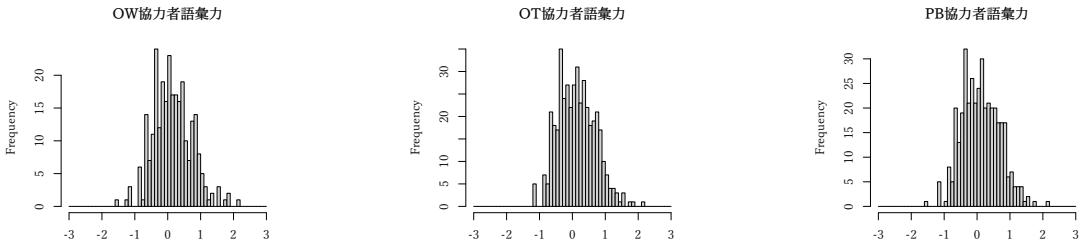

図 1　語彙力の分布

レジスタ		サンプル	文	文節 (語)	協力者
OW	白書	1	36	462	277
OT	教科書	38	9,521	50,606	422
		(平均)	250.6	1331.7	
PB	書籍	83	10,075	84,736	388
		(平均)	121.4	1,020.9	

表 1　刺激文と実験協力者の統計

サンプル × 協力者数	OW	OT	PB
BCCWJ-SPR2	308	4,865	11,325
（内）分析対象	277	4,685	10,932
データポイント数	OW	OT	PB
BCCWJ-SPR2	136,797	5,704,898	10,769,380
（内）分析対象	124,502	5,490,977	10,484,300

表 2　分析対象

よる読み時間データ収集実験参加を募った [8]。最初に『現代日本語書き言葉均衡コーパス』の白書 (OW)1 サンプルについて試験的に 500 人の協力者を募集した。その後、教科書 (OT)38 サンプルと書籍 (PB)83 サンプルについて、1 サンプル当たり 200 人の協力者を募集した。各実験協力者は 2020 年 11 月〜12 月にかけて、白書 → 教科書 → 書籍の順に自由に実験に参加でき、読むサンプルの数も自由に設定した。従前の発表 [8] と同じスクリーニングを行った後、単語親密度調査 200 回答以上・読み時間実験 5 サンプル以上に参加した実験協力者のデータを分析対象とした。表 1 に刺激文と実験協力者の統計を、表 2 に分析対象となるデータ数を示す。

3　統計分析

統計分析は対数読み時間に対する線形混合モデルにおいて、各固定効果における有意差が確認できるか否かにより行う。呈示順[4]・文字列長[5]・試行順[6]・係り受け[7]・実験協力者の語彙力[8]を固定効果とし、実験協力者[9]・サンプル ID[10]をランダム効果とした。本稿では lme4[20, 21] を用いた一般化線形混合モデルの結果を示す。一度モデルを推定したうえで、3SD よりも外側の値のデータポイントを排除し、再推定を行った結果を表 3 に示す。なお、対数読み時間を頻度主義的な線形混合モデルで分析したものを付録の表 4 に、同様のモデルをベイジアン線形混合モデル（対数正規モデル）で分析したものを付録の表 5 に示す。

まず、同一サンプル内の呈示順については、実験が進むについて文脈を得るために読み時間が短くなる傾向が確認された。文字列長は長いほど認識に時間が長くなる傾向が確認された。試行順については教科書においては実験協力者が慣れて短くなる傾向が確認された。書籍においては、反対に繰り返し実施することによる疲れからか、長くなる傾向が確認された。最後に係り受けの数が多いほど、予測が効くために、読み時間が短くなる傾向が確認された。ここまでは既存の分析結果 [8] と同様の結果であった。

4　考察

以下、結果についての考察を示す。

4.1　語彙力の影響

以前の調査 [6] では語彙力が高い実験協力者の読み時間が長い傾向が確認された。しかしながら、調

4）　BCCWJ-SPR2 における文呈示順 SPR_sentence_ID (sentid)・文節呈示順 SPR_bunsetsu_ID (bid)。
5）　BCCWJ-SPR2 における表層形文字列長 SPR_word_length (length)。
6）　BCCWJ-SPR2 における実験協力者ごとの試行順（サンプル単位に集計） SPR_trial。OW については 1 サンプルのみのために常に試行順が 1 のため未定義。
7）　BCCWJ-SPR2 における当該文節に対する係り受けの数 SPR_DepPara_depnum (dependent)。BCCWJ-DepPara から復元。OT には係り受けが付与されていないために未定義。
8）　WLSP-Familiary3.1 における実験協力者のランダム効果 WFR_subj_rate。
9）　BCCWJ-SPR2 における実験協力者固有の ID SPR_subj_ID_factor。
10）　BCCWJ におけるサンプル ID BCCWJ_Sample_ID。

	Dependent variable:					
	SPR_reading_time					
	OW (白書)		OT (教科書)		PB (書籍)	
SPR_sentence_ID (呈示順)	−6.087***	(0.051)	−0.127***	(0.0004)	−0.142***	(0.001)
SPR_bunsetsu_ID (呈示順)	−1.501***	(0.049)	−2.046***	(0.011)	−0.856***	(0.006)
SPR_word_length (文字列長)	24.820***	(0.170)	5.170***	(0.021)	6.798***	(0.015)
SPR_trial (試行順)			−0.757***	(0.005)	0.382***	(0.006)
DepPara_depnum (係り受け)	−15.310***	(0.591)			−5.258***	(0.034)
WFR_subj_rate (語彙力)	−81.239***	(21.227)	−16.169*	(9.087)	−18.405**	(8.731)
Constant	558.984***	(12.936)	353.723***	(6.548)	306.631***	(5.425)
データポイント数	121,769		5,407,252		10,321,560	
3SD より外側の削除数 (削除率)	2,732	(0.0219)	83,724	(0.0152)	162,740	(0.0155)
対数尤度	−818,815.100		−32,796,021.000		−62,393,234.000	
注:					*p<0.1; **p<0.05; ***p<0.01	

回帰式：

```
SPR_reading_time ~ SPR_sentence_ID + SPR_bunsetsu_ID + SPR_word_length + DepPara_depnum
+ SPR_trial + WFR_subj_rate + (1 | SPR_subj_ID_factor) +  (1 | BCCWJ_Sample_ID)
```

但し、OT は 1 回の試行のみのため SPR_trial は未定義、OT は係り受け情報が付与されていないため DepPara_depnum は未定義。

表 3　一般化線形混合モデルに基づく分析結果（読み時間）

査人数が少ないうえ語彙力が高い群が実質 1 名であったために、この 1 名のデータに依存し、一般性のない分析である可能性があった。本調査では、読み時間計測タスクを自己ペース読文課題に限定しながらも、数百人規模のデータを用い、読む文章量も増やした設定で再検証を行った。また語彙力の測定も NTT の平成版語彙数判定テストではなく、新たに単語親密度調査に基づく副産物としての協力者の語彙力を用いた。結果、語彙力が高い群の読み時間がより短くなる傾向が観察された。

4.2　レジスタの影響

語彙力の影響を有意差の観点で見ると、OT(教科書) が $p < 0.1$、PB(書籍) が $p < 0.05$、OW(白書) が $p < 0.01$ であった。教科書は小・中・高の国語の教科書のサンプルであり、協力者は一度は読んだことがある可能性がある。書籍と白書を比べると、一般に書籍のほうが協力者が読みなれている可能性もある。このレジスタについてのなじみの程度が、語彙力の影響の有意差に表れている可能性がある。なじみのある教科書では語彙力の影響の差が小さいが、なじみのない白書では語彙力の影響の差が大きくなった可能性がある。

4.3　サプライザルや言語モデルのパープレキシティとの関係

サプライザル [10, 11] は、文処理の負荷に対する情報量基準に基づいた指標で、当該単語の文脈中の負の対数確率（次式）によりモデル化される文処理の困難さを表す：

$$\mathrm{Surprisal}(w) = -\log_2 P(w|\mathrm{context})$$

似た概念の言語モデルの評価指標としてパープレキシティ（平均分岐数）$\mathrm{perplexity(context)} = \Pi_{i=1}^{N} P(w_i)^{1/N}$ がある。これを対数を取った形式 (cross entropy) で表記すると $H(\mathrm{context}) = \frac{1}{N}\Sigma_{i=1}^{N} \log_2 P(w_i)$ となるが、言語モデルとして次にくる単語 w を予測する際の仮定としては、$\arg\max \log_2 P(w|\mathrm{context})$ と、この cross entropy の最大化問題に帰着でき、実質的には符号が反転したサプライザルの式の最小化と等価の問題と捉えることができる。

このパープレキシティ (cross entropy) が低いほどヒトらしいという研究は英語については確認されている [12, 13, 14, 15, 16] 一方、英語と日本語における読み時間と Transformer 系の言語モデルの対照分析 [17, 18] において、日本語においては必ずしもそうとは言えない報告がなされた。

ここで、サプライザル・パープレキシティとは何

か、さらには「ヒトらしい」とは何かについて考えたい。過去の研究においては、**言語モデルの能力差**によるパープレキシティの検討がなされてきた。本研究では**ヒトの能力差**と**レジスタの親密性**の観点から検討を行いたい。

まず、**ヒトの能力差**の観点からは、仮定として語彙力が高いヒトは文脈から単語を推測する確率が高いとするならば、この能力差がサプライザルの指標を下げ、全体的に読み時間を短くした可能性がある。サプライザルと言語モデルのパープレキシティの等価性から、言語モデルのパープレキシティが高くなるにつれ、より語彙力の高い群に親和性がある予測を行う可能性がある。

次に、**レジスタの親密性**の観点からは、仮定として、OT(教科書) は一度読んだ可能性があるテキストであり、PB(書籍) は普段読んでいるものに近い新しいテキストであり、OW(白書) は普段読み慣れないテキストであると想定する。親密性の観点からは、OT>PB>OW の順で次にくる単語が予測しやすい可能性があり、サプライザルはこの逆順で小さい（パープレキシティの観点からは正順で大きい）レジスタであると考える。表 3 の語彙力の有意差の傾向から、OT<PB<OW の順で語彙力の差が出やすくなっていることがわかる。本データ (BCCWJ-SPR2) の PB の NDC（日本十進分類法）に基づく分析 [22] では、多くの人が親しみやすい芸能やスポーツなどを含む「7. 芸術・美術」が読み時間が短くなる傾向が確認されている。

以上の観点を総合的に考えると、「ヒトらしい」言語モデルとは、適切なレベルの語彙力を持ち、レジスタの親密性の差異により予測可能性が変化しうるものであろうと考える。後者のレジスタの親密性については、Hale らが、サプライザルをジャンル横断的に調査し、EEG データと推定したサプライザルとを対照することで、同一ジャンルで訓練した言語モデルのほうがより EEG データに適合することを報告している [23]。

5 おわりに

本研究では、日本語の読み時間と語彙力の関係について検討した。本研究の貢献は以下の 3 点である：

- 対面で実験できないなか、オンラインで日本語の読み時間と語彙力調査を大規模に行う方法を確立した。
- 大規模調査により、日本語において語彙力の読み時間に与える影響について、複数レジスタの資料に対して調査した。
- 調査結果として、「語彙力の高い」群が読み時間が短くなることを確認した。

分析結果から以下の 2 点について考察を行った：

- ヒトの能力差において、「語彙力の高い」ことと「言語モデルのパープレキシティの高い」ことのアナロジーを仮定すると、言語モデルのパープレキシティが高くなるにつれてより語彙力の高い群に親和性がある予測を行う可能性がある。
- レジスタの親密性において、サプライザルが全体的に低いであろう読みなれているレジスタにおいては、語彙力の差が出にくい。

このように日本語の読み時間において、ヒトの能力差やレジスタ差の検討が可能になった。今後、より精緻な数理モデルによる検討を期待する。

今回、ヒト側の能力として語彙力について検討を行い、上記の傾向について明らかにした。日本語においては、再帰的ニューラルネットワーク文法による読み時間の検討 [24, 25] も進められている。文法に基づくヒトの文処理過程をつかさどる重要な概念として、実験協力者の作業記憶容量（パージングアルゴリズムにおけるビーム幅相当）がある。BCCWJ-EyeTrack においては、日本語リーディングスパンテスト [3] を行い、成績の優劣による検討も行った [6]。本データにおいても、オンラインで作業記憶容量を評価する方法について検討し、パージングに基づく文処理過程のモデリングについて評価可能なデータを構築したい。

謝辞

本研究は国立国語研究所コーパス開発センター共同研究プロジェクトの成果です。また、科研費 17H00917, 18H05521 の支援を受けました。

参考文献

[1]Marcel Adam Just, Patricia A. Carpenter, and Jacqueline D. Woolley. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 3:228–238, 1982.

[2]浅原 正幸, 小野 創, and 宮本 エジソン 正. BCCWJ-EyeTrack–『現代日本語書き言葉均衡コーパス』に対する読み時間付与とその分析–. **言語研究**, 156:67–96, 2019.

[3]苧阪満里子, editor. **ワーキングメモリー脳のメモ帳**. 新曜社, 2002.

[4]Shigeaki Amano and Tadahisa Kondo. Estimation of Mental Lexicon Size with Word Familiarity Database. In *Proceedings of International Conference on Spoken Language Processing*, volume 5, pages 2119–2122, 1998.

[5]天野 成昭 and 近藤 公久, editors. **単語親密度，*NTT* データベースシリーズ　日本語の語彙特性第 *1* 巻**. 三省堂, 1999.

[6]浅原 正幸, 小野 創, and 宮本 エジソン 正. 『現代日本語書き言葉均衡コーパス』の読み時間とその被験者属性. In **言語処理学会第 *23* 回年次大会発表論文集**, pages 273–276, 2017.

[7]Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. The Natural Stories Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[8]浅原 正幸. クラウドソーシングによる大規模読み時間データ収集. In **言語処理学会第 *27* 回年次大会発表論文集**, pages 1156–1161, 2021.

[9]浅原 正幸. クラウドソーシングによる単語親密度データの構築 (2021 年版). In **言語処理学会第 *28* 回年次大会発表論文集**, 2022.

[10]John Hale. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Conference of the North American Chapter of the Association for Computational Linguistics*, volume 2, pages 159–166, 2001.

[11]Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008.

[12]Stefan L. Frank and Rens Bod. Insensitivity of the human sentence-processing system to hierarchical structure. *Journal of Psychological Science*, 22(6):829–834, 2011.

[13]Victoria Fossum and Roger Levy. Sequential vs. Hierarchical Syntactic Models of Human Incremental Sentence Processing. In *Proceedings of CMCL*, pages 61–69, 2012.

[14]Adam Goodkind and Klinton Bicknell. Predictive Power of Word Surprisal for Reading Times is a Linear Function of Language Model Quality. In *Proceedings of CMCL*, pages 10–18, 2018.

[15]Adam Goodkind and Klinton Bicknell. Comparing Gated and Simple Recurent Neural Network Architectures as Models of Human Sentence Processing. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 112–118, 2019.

[16]Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 1707–1713, 2020.

[17]Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguisticsand the 11th International Joint Conference on Natural Language Processing*, pages 5203–5217, 2011.

[18]栗林 樹生, 大関 洋平, 伊藤 拓海, 吉田 遼, 浅原 正幸, and 乾 健太郎. 予測の正確な言語モデルがヒトらしいとは限らない. In **言語処理学会第 *27* 回年次大会発表論文集**, pages 267–272, 2021.

[19]浅原 正幸. 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**, 27(1):133–150, 2020.

[20]Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67:1–48, 2015.

[21]Marek Hlavac. stargazer: Well-formatted regression and summary statistics tables, 2018. R package version 5.2.2.

[22]浅原 正幸 and 加藤 祥. 『現代日本語書き言葉均衡コーパス』書籍サンプルの読み時間の分析. In **日本語学会 *2021* 年度春季大会予稿集**, pages 49–54, 2021.

[23]John Hale, Adhiguna Kuncoro, Keith Hall, Chris Dyer, and Jonathan Brennan. Text genre and training data size in human-like parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5846–5852, Hong Kong, China, November 2019. Association for Computational Linguistics.

[24]Ryo Yoshida, Hiroshi Noji, and Yohei Oseki. Modeling human sentence processing with left-corner recurrent neural network grammars. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2964–2973, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[25]吉田 遼, 能地 宏, and 大関 洋平. 再帰的ニューラルネットワーク文法による人間の文処理のモデリング. In **言語処理学会第 *27* 回年次大会発表論文集**, pages 273–278, 2021.

A 付録

	Dependent variable:					
	SPR_log_reading_time					
	OW (白書)		OT (教科書)		PB (書籍)	
SPR_sentence_ID (呈示順)	−0.012***	(0.0001)	−0.0004***	(0.00000)	−0.0004***	(0.00000)
SPR_bunsetsu_ID (呈示順)	−0.003***	(0.0001)	−0.006***	(0.00003)	−0.003***	(0.00002)
SPR_word_length (文字列長)	0.036***	(0.0002)	0.011***	(0.0001)	0.014***	(0.00004)
SPR_trial (試行順)			−0.002***	(0.00001)	0.001***	(0.00002)
DepPara_depnum (係り受け)	−0.022***	(0.001)			−0.012***	(0.0001)
WFR_subj_rate (語彙力)	−0.180***	(0.040)	−0.052**	(0.025)	−0.052**	(0.026)
Constant	6.255***	(0.025)	5.826***	(0.020)	5.664***	(0.016)
データポイント数	135,070		5,412,398		10,327,584	
3SD より外側の削除数 (削除率)	1,559	(0.0125)	78,578	(0.0143)	156,716	(0.0149)
対数尤度	−38,816.700		−598,180.400		−743,585.500	
注:						*p<0.1; **p<0.05; ***p<0.01

回帰式：

```
SPR_log_reading_time ~ SPR_sentence_ID + SPR_bunsetsu_ID + SPR_word_length + DepPara_depnum
+ SPR_trial + WFR_subj_rate + (1 | SPR_subj_ID_factor) +  (1 | BCCWJ_Sample_ID)
```

但し、OT は 1 回の試行のみのため SPR_trial は未定義、OT は係り受け情報が付与されていないため DepPara_depnum は未定義。

表 4　一般化線形混合モデルに基づく分析結果（OW 白書・OT 教科書・PB 書籍：対数読み時間）

OW(白書)	mean	se_mean	sd
α 切片	6.217	0.064	0.123
β_{sentid} 呈示順	-0.012	0.000	0.002
β_{bid} 呈示順	-0.003	0.000	0.001
β_{length} 文字列長	0.037	0.000	0.006
$\beta_{\text{dependency}}$ 係り受け	-0.024	0.001	0.009
β_{subjrate} 語彙力	-0.118	0.094	0.116
σ	0.988	0.075	0.097
σ_{subj}	2.603	1.020	1.254

```
model {
  real mu;
  gamma_subj ~ normal(0,sigma_subj); // prior
  for (k in 1:N) { //
    mu = alpha +  beta_length * length[k] +
      beta_dependent * dependent[k] +
      beta_sentid * sentid[k] +
      beta_bid * bid[k] +
      beta_subjrate * subjrate[k] +
      gamma_subj[subjid[k]];
    time[k] ~ lognormal(mu,sigma);
  }
}
```

表 5　ベイジアン線形混合モデルに基づく分析結果（OW 白書のみ：対数正規モデル）