クラウドソーシングによる大規模読み時間データ収集

浅原 正幸
国立国語研究所
masayu-a@ninjal.ac.jp

1 はじめに

本研究では、クラウドソーシングにより、日本語を対象とした大規模読み時間データを構築したので報告する。我々は、視線走査装置[1)]や自己ペース読文法 [1] を用いて、新聞記事を対象として読み時間データを収集 [2] し、さまざまな分析 [3, 4, 5, 6, 7] をおこなってきた。また、教科書・書籍に対しても視線走査装置による小規模の読み時間データを収集した [8]。しかしながら、被験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) [9] 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。本稿では読み時間収集手法を示すとともに、簡単な統計分析結果を報告する。

2 読み時間データの収集手法

2.1 刺激文

刺激文は『現代日本語書き言葉均衡コーパス』(BCCWJ) [10] のOW（白書）・OT（教科書）・PB（書籍）を用いた。OW（白書）は著作権フリーのもの 1 サンプルを、実験環境とともに結果を公開する。OT（教科書）は、国語科（現代文）のもの 38 サンプル（小 17、中 9、高 12）を用いた。これは、日本で国語教育を受けた方であれば過去に読んだことがある刺激文として、読み時間データを収集した。今後、国外で国語教育を受けた方や日本語学習者のデータを取得し、国語教育・日本語教育におけるリーダビリティ評価の対照データとして利用することを想定する。PB（書籍）はコアデータ 83 サンプルを用いた。これは、係り受けなど多様なアノテーションに基づく検討を行うほか、PN（新聞）の読み時間データで制限されている商用利用を想定する。

表 1 に、刺激文のサンプル数・文数・文節数（および募集被験者数：次節で言及）を示す。対照のために NSC のサンプル数・文数・単語数を示す。NSC はおよそ 1,000 語程度の文章をサンプリングしたが、我々のデータもOT（教科書）・PB（書籍）サンプルについては平均 1,000 文節以上のデータを刺激として用いた。

各刺激文章に対して、内容をきちんと読んでいるかを確認するための YES/NO で回答する質問を 2 つずつ（YES が正解 1 つ、NO が正解 1 つ）設定した。

表 1 刺激文の統計

レジスタ		サンプル	文	文節 (語)	被験者
OW	白書	1	36	462	427
OT	教科書	38	9,521	50,606	200
		(平均)	250.6	1331.7	
PB	書籍	83	10,075	84,736	200
		(平均)	121.4	1,020.9	
NSC [9]		10	485	10,245	181 †
		(平均)	48.5	1,024.5	

† 本研究と NSC の集計方法は異なる（2.2 節参照）

2.2 自己ペース読文法

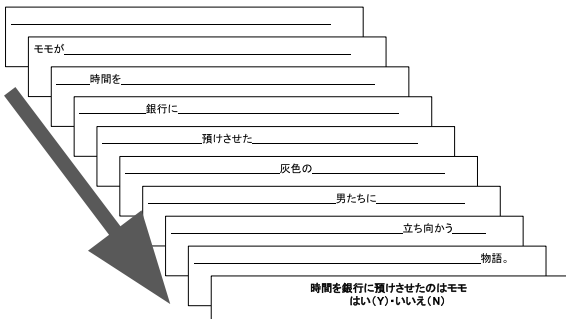

図 1 移動窓方式による自己ペース読文法

自己ペース読文法は、移動窓を用いて疑似的な視線移動環境を設定しながら、部分的に呈示された単

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

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

語や文節の表示時間により読み時間を計測する手法である。図 1 に例を示す。スペースキーを押すたびに逐次的に文節が表示され、スペースキーを押した時間間隔をミリ秒単位で記録することで読み時間を計測する。1 文章を読んだ最後に「はい」「いいえ」の 2 択で答える内容把握の設問を設定した。ブラウザ上で自己ペース読文法による被験者実験を行うため、ibexfarm [3] を用いた。日本語対応は文献 [11] に倣った。

2.3 被験者

被験者は Yahoo! クラウドソーシング[4]により募集した。2020 年 10 月に、あらかじめ「単語親密度」の評定実験 [12] を行い、単語親密度の回答において、回答の分散が大きい 2,092 人を対象に募集した。これは、適切に作業に取り組む方をあらかじめ選別するほか、単語親密度推定時に被験者毎の語彙力も推定できるために、事前実験として行った。

OWに対しては、試験的に 500 人募集して実験を行った。OT・PBに対しては、200 人募集した。内容確認質問の正解が YES である実験 100 人、NO である実験 100 人募集した[5]。表 1 に示す本研究の被験者数はサンプル毎の募集人員だが、NSC の被験者数はいずれかのサンプルを確認した異なり人数である。NSC においては、各被験者は最低 5 サンプルを確認した。NSC 論文で言及されているように本実験の作業者の異なりを集計すると 574 人（延べ 24,727 人）であった。

3 実験データの整理

3.1 不適切なデータの排除

本データの収集は、オンラインで実施したため、データの品質が対面で収集した際よりも品質が悪いことが想定される。そこで、不適切なデータの排除を試みる。まず、(1) 実験開始時に実験データの取り扱いおよび謝金の支払方法に同意していないものを排除した。次に、(2) 同じ被験者が同じサンプルを複数回実施した場合、2 回目以降の試行データを削除した。さらに、(3) サンプル単位の平均読み時間が 150ms 未満もしくは 2,000ms 超過のもの、(4)YES/NO 質問を誤答しているものを排除した。

3） https://spellout.net/ibexfarm/
4） https://crowdsourcing.yahoo.co.jp/
5） OT・PBにおいて実験設定のミスにより 200 人以上募集したサンプルもある。

表 2 不適切なデータの排除（サンプル数）

摘要	合計	OW	OT	PB
(1) 不同意	531	54	278	199
(2) 重複回答	110	0	46	64
(3) <150 or 2,000<	4,329	15	1,439	2,875
(4) 誤答	2,963	48	825	2,090
適切なサンプル	16,498	308	4,865	11,325
合計	24,431	425	7,453	16,553

最後にデータポイント単位に不適切なデータを排除する。表 3 の「適切なサンプル」の行に、適切なサンプルに含まれるデータポイント数を示す。これらから、さらに (5) 100ms 未満もしくは 3,000ms 超過のものを排除した。結果、表中「分析対象」に示す件数を適切なデータポイントとして扱うことにした。[6]NSR のデータポイント数が 848,857 と比しても、大規模な読み時間データが構築できたといえる。

表 3 不適切なデータの排除（データポイント数）

摘要	OW	OT	PB
適切なサンプル	140,449	6,114,814	11,309,783
(5) <100 or 3,000<	3,653	409,916	540,403
分析対象	136,797	5,704,898	10,769,380

3.2 実験データの形式

表 4 に実験データの形式について示す。データは TSV の帳票形式で 1 行目がヘッダ行で列名を記載してある。

BCCWJ_Sample_ID と BCCWJ_start が BCCWJ 上の位置情報で、この情報により BCCWJ 上の形態論情報や各種アノテーションデータと突合ができる。

SPR_sentence_ID と SPR_bunsetsu_ID が実験時の文 ID と文節 ID で、実質的に実験内の呈示順を表す。SPR_surface は呈示した表層形であるが、OT と PB はマスクして文字数のみ (SPR_word_length) がわかるようにした。

SPR_reading_time が分析対象の読み時間である。分析時には、対数読み時間 (SPR_log_reading_time) も生成した。SPR_instructiontime が最初の実験教示の時間である。2 回目以降の実験においては読み飛ばされる傾向がある。

SPR_QA_question が実験時の確認用の YES/NO 質問で、SPR_QA_answer がその正解である。SPR_QA_correct が被験者が YES/NO 質問に正解したか否かのフラグで、公開データは基本的には正

6） 基準として、NSC [9] が 100ms 未満もしくは 3,000ms 超過のデータポイントを削除しているのを参考にした。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 4　実験データの形式

列名	摘要	OW	OT	PB
BCCWJ_Sample_ID	BCCWJ のサンプル ID	✓	✓	✓
BCCWJ_start	BCCWJ の開始位置	✓	✓	✓
SPR_sentence_ID	実験時の文 ID	✓	✓	✓
SPR_bunsetsu_ID	実験時の文節 ID	✓	✓	✓
SPR_surface	実験時に呈示した表層形（文節）	✓	masked	masked
SPR_word_length	実験時に呈示した表層形の文字数	✓	✓	✓
SPR_sentence	実験時に呈示した文	✓	masked	masked
SPR_reading_time	読み時間	✓	✓	✓
SPR_log_reading_time	読み時間	✓	✓	✓
SPR_instruction_time	実験教示時間	✓	✓	✓
SPR_QA_question	実験時の確認用 YES/NO 質問	✓	✓	✓
SPR_QA_answer	実験時の確認用 YES/NO 質問の正解	✓	✓	✓
SPR_QA_correct	実験時の確認用 YES/NO 質問に正解したか	✓	✓	✓
SPR_QA_qa_time	実験時の確認用 YES/NO 質問の回答時間	✓	✓	✓
SPR_subj_ID	被験者 ID	✓	✓	✓
SPR_averageRT	実験時のサンプル単位平均読み時間	✓	✓	✓
SPR_timestamp	実験時刻	✓	✓	✓
SPR_trial	被験者毎の実験試行回数	N/A	✓	✓
SPR_control	実験時の呈示方法	✓	✓	✓
DepPara_bid	BCCWJ-DepPara の文節 ID	✓	N/A	✓
DepPara_depid	BCCWJ-DepPara の係り先 ID	✓	N/A	✓
DepPara_depnum	BCCWJ-DepPara の係り受け数	✓	N/A	✓
BCCWJ_OT_school_type	教科書の校種 (OT01:小・OT02:中・OT03: 高等学校)	N/A	✓	N/A

解したもののみとする。SPR_QA_qatime が被験者が YES/NO 質問に要した回答時間である。

SPR_subj_ID が被験者 ID である。SPR_averageRT が実験時のサンプル単位の平均読み時間である。SPR_timestamp が実験を実施した時刻、SPR_trial が当該ジャンルの読み時間計測を行うのが何回目かを表す。SPR_control が実験時の呈示方法（文節間に空白があるかないか）である。

読み時間の係り受けの影響を検討するために、コアデータの OW（白書）・PB（書籍）に対しては、 BCCWJ-DepPara [13] の情報を付与した。BCCWJ-DepPara は BCCWJ と異なる文境界を定義しているため、呈示した際の文節 ID との齟齬が生じる (DepPara_bid, DepPara_depid) が、係り受けの数 (DepPara_depnum) は BCCWJ-DepPara の基準に基づく。

また、OT（教科書）については、分析対象として教科書の校種 (BCCWJ_OT_school_type, OT01: 小・OT02:中・OT03:高等学校) を設定した。

4　実験データの分析

4.1　分析方法

本稿では、頻度主義的な分析手法 [14, 15] による結果を示す。読み時間の検討を一般化線形混合モデル (R[16], lme4[17], stargazer[18]) を用いて行う。

固定効果として、呈示順の情報である SPR_sentence_ID (実験時の文 ID)・SPR_bunsetsu_ID (実験時の文節 ID) と、表層形の文字数である SPR_word_length を考慮した。OW（白書）・PB（書籍）については、当該文節の係り受けの数 DepPara_depnum を考慮した。OT（教科書）については、校種 SPR_OT_school_type を考慮した。OT（教科書）・PB（書籍）については、同一の被験者が複数のサンプルを読んだ際の試行順序 SPR_trial も固定効果としてモデル化した。

また、被験者間の個人差をモデル化するために SPR_subj_ID (被験者 ID) をランダム効果として考慮した。OT（教科書）・PB（書籍）については、サンプル間の個体差をモデル化するために BCCWJ_Sample_ID (BCCWJ のサンプル ID) もランダム効果として考慮した。分析式は次のとおり：

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

以下では、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示す。なお、対数読み時間の頻度主義的な分析は付録に示す。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 5　一般化線形混合モデルの基づく分析結果（読み時間）

	Dependent variable:					
	SPR_reading_time					
	OW (白書)		OT (教科書)		書籍 (PB)	
SPR_sentence_ID	−6.042***	(0.048)	−0.125***	(0.0004)	−0.143***	(0.001)
SPR_bunsetsu_ID	−1.477***	(0.046)	−2.047***	(0.011)	−0.856***	(0.006)
SPR_word_length	24.311***	(0.160)	5.113***	(0.021)	6.705***	(0.014)
DepPara_depnum	−15.048***	(0.555)			−5.225***	(0.033)
SPR_trial			−0.760***	(0.005)	0.379***	(0.006)
BCCWJ_OT_school_typeOT02			−7.721	(8.898)		
BCCWJ_OT_school_typeOT03			−25.228***	(8.352)		
Constant	540.050***	(11.951)	361.566***	(7.155)	306.887***	(5.321)
データポイント数	133,806		5,617,794		10,701,504	
3SD より外側の削除数 (削除率)	1,726	(0.0126)	87,103	(0.0152)	168,671	(0.0155)
対数尤度	−897,781.800		−34,071,483.000		−64,679,004.000	
注:					*p<0.1; **p<0.05; ***p<0.01	

4.2　分析結果

表 5 に分析結果を示す。表中、固定効果に対する推定値を有意差の有無ともに示す。カッコ内の値は標準誤差である。

サンプル内の呈示順 (SPR_sentence_ID, SPR_bunsetsu_ID) については、いずれのレジスタにおいても読み時間が短くなる傾向が確認された。文字数が長くなるにつれて読み時間が長くなる傾向も確認された (SPR_word_length)。

OW・PB(白書 書籍)については、当該文節の係り受けの数 (DepPara_depnum) との相関を検討した。いずれの結果も係り受けの数が多い文節の読み時間が短くなる傾向が確認された。

OT・PB(教科書 書籍)については、同一の被験者が複数のサンプルを読んだ際の試行順序 SPR_trial も固定効果として検討した。OT(教科書)においては試行回数を重ねるにあたり読み時間が短くなるが、PB(書籍)においては試行回数を重ねるにあたり読み時間が長くなる傾向がみられた。OT(教科書)については、校種 SPR_OT_school_type を考慮した。小学校のテキストに比べて、高等学校(教科書)の読み時間のほうが短くなる傾向がみられた。OT(教科書)においては、実験選択画面において小学校⇒中学校⇒高等学校という順で進めやすいようになっており、実際 SPR_trial と小中高等学校の校種順に相関がある状態であった。このことが読み時間に対する試行回数の影響のレジスタ間の齟齬をもたらした可能性がある。

最後にレジスタ間の関係について言及する。被験者は基本的にはOW(白書)・OT(教科書)・PB(書籍)の順に試行しており、最初に実験を行ったOW(白書)では、他のレジスタと比べて長い傾向がみられる。OT(教科書)は日本国内で国語の教育を受けた場合には一度は目にしている可能性が高いテキストである。しかしながら、呈示順が影響して、PB(書籍)より読み時間が長い傾向がみられた。今後、書籍に対しては、日本十進分類法などによるジャンル間比較を行うことを検討する。

5　おわりに

本発表では、クラウドソーシングを用いて構築した大規模な読み時間データについて紹介した。ブラウザ上で動作する自己ペース読文法による実験環境 ibexfarm を用い、Yahoo! クラウドソーシングで被験者を募集することにより、短期間・低価格に大規模読み時間データを構築することができた。実験データの分析においては、線形混合モデルによる試行順・係り受けの数の影響を検討し、過去の実験室で収集したデータによる結果と同様の結果が得られることを確認した。さらに、OT(教科書)の分析においては、小学校の教科書よりも高等学校の教科書のほうが成人日本語話者の読み時間が短くなることを確認した。

本データは https://github.com/masayu-a/BCCWJ-SPR2 で公開する。今後、BCCWJ のさまざまなアノテーションとの重ね合わせを行い様々な分析を行うとともに、ベイズ主義的な分析 [19, 20] も進めたい。また、L1 学習者・L2 学習者の読み時間を収集することにより、言語の読解能力の習得過程のデータ化を進めたい。

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 18H05521, 18K18519 の支援を受けました。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

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]浅原 正幸. 名詞句の情報の状態と読み時間について. **自然言語処理**, 25(5):527–554, 2018.

[4]浅原 正幸. 日本語の読み時間と節境界情報—主辞後置言語における wrap-up effect の検証—. **自然言語処理**, 26(2):301–327, 2019.

[5]浅原 正幸 and 加藤 祥. 読み時間と統語・意味分類. **認知科学**, 26(2):219–230, 2019.

[6]浅原 正幸. 単語埋め込みに基づくサプライザル. **自然言語処理**, 26(3):635–652, 2019.

[7]浅原 正幸. 読み時間と述語項構造・共参照情報について. In **言語処理学会第 *25* 回年次大会発表論文集**, pages 249–252, 2019.

[8]森山 奈々美, 荻原 亜彩美, 近藤 森音, 浅原 正幸, and 相澤 彰子. BCCWJ-EyeTrack-2: 書籍と教科書の読み時間データ. In **言語処理学会第 *25* 回年次大会発表論文集**, pages 699–702, 2019.

[9]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).

[10]Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 48:345–371, 2014.

[11]中谷 健太郎. 自己ペース読文課題を使った実験：ウェブ編. In 中谷 健太郎, editor, **パソコンがあればできる！ことばの実験研究の方法　容認性調査、読文・産出実験からコーパスまで**, chapter 4, pages 81–106. ひつじ書房, 東京, 2019.

[12]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.

[13]浅原 正幸 and 松本 裕治. 『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**, 25(4):331–356, 2018.

[14]R. Harald Baayen. *Analyzing Linguistic Data: A practical Introduction to Statistics using R*. Cambridge University Press, 2008.

[15]Shravan Vasishth, Daniel Schad, Audrey Bürki, and Reinhold Kliegl. *Linear Mixed Models in Linguistics and Psychology: A Comprehensive Introduction*. (近刊).

[16]R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

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

[18]Marek Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Central European Labour Studies Institute (CELSI), Bratislava, Slovakia, 2018. R package version 5.2.2.

[19]Tanner Sorensen, Sven Hohenstein, and Shravan Vasishth. Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*, 12:175–200, 2016.

[20]Bruno Nicenboim, Daniel Schad, and Shravan Vasishth. *An Introduction to Bayesian Data Analysis for Cognitive Science*. (近刊).

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A 付録

表 6 に対数読み時間の分析結果を示す。外れ値とみなす 3SD より外側の削除数が若干減ったが、基本的に読み時間と同様の結果が得られた。

表 6 一般化線形混合モデルに基づく分析結果（対数読み時間）

	Dependent variable:					
	SPR_log_reading_time					
	OW (白書)		OT (教科書)		書籍 (PB)	
SPR_sentence_ID	−0.012***	(0.0001)	−0.0004***	(0.00000)	−0.0004***	(0.00000)
SPR_bunsetsu_ID	−0.003***	(0.0001)	−0.006***	(0.00003)	−0.003***	(0.00002)
SPR_word_length	0.036***	(0.0002)	0.010***	(0.0001)	0.014***	(0.00004)
DepPara_depnum	−0.022***	(0.001)			−0.012***	(0.0001)
SPR_trial			−0.002***	(0.00001)	0.001***	(0.00002)
BCCWJ_OT_school_typeOT02			−0.025	(0.024)		
BCCWJ_OT_school_typeOT03			−0.078*	(0.022)		
Constant	6.216***	(0.023)	5.826***	(0.020)	5.664***	(0.016)
データポイント数	135,070		5,623,067		10,707,218	
3SD より外側の削除数 (削除率)	1,726	(0.0126)	81,830	(0.0143)	162,957	(0.0150)
対数尤度	−38,248.550		−630,606.800		−780,846.500	
注:					*p<0.1; **p<0.05; ***p<0.01	

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).