Reading-Time Annotations
for *Balanced Corpus of Contemporary Written Japanese*

Masayuki Asahara
National Institute of Japanese
Language and Linguistics,
National Institute for the Humanities, Japan
masayu-a.ninjal.ac.jp

Hajime Ono
Faculty of Liberal Arts,
Tsuda College.

Edson T. Miyamoto
Graduate School of
Humanities and Social Science,
University of Tsukuba

Abstract

The *Dundee Eyetracking Corpus* contains eyetracking data collected while native speakers of English and French read newspaper editorial articles. Similar resources for other languages are still rare, especially for languages in which words are not overtly delimited with spaces. This is a report on a project to build an eyetracking corpus for Japanese. Measurements were collected while 24 native speakers of Japanese read excerpts from the *Balanced Corpus of Contemporary Written Japanese* Texts were presented with or without segmentation (i.e. with or without space at the boundaries between *bunsetsu* segmentations) and with two types of methodologies (eyetracking and self-paced reading presentation). Readers' background information including vocabulary-size estimation and Japanese reading-span score were also collected. As an example of the possible uses for the corpus, we also report analyses investigating the phenomena of anti-locality.

1 Introduction

Corpora of naturally-produced texts such as newspapers and magazines marked with detailed morphological, syntactic and semantic tags, are often used in human language-production research. In contrast, texts created by psycholinguists exclusively for research purposes, are commonly used in language-comprehension research.

We introduce a reusable linguistic resource that can help bridge this gap by bringing together techniques from corpus linguistics and experimental psycholinguistics. More concretely, we have collected reading times for a subset of texts from the *Balanced Corpus of Contemporary Written Japanese* (BCCWJ) (Maekawa et al., 2014), which already contains syntactic and semantic types of annotations. The goal is to produce a resource comparable to the *Dundee Eyetracking Corpus* (Kennedy and Pynte, 2005), which contains reading times for English and French newspaper editorials from 10 native speakers for each language, recorded using eyetracking equipment. The English version of the *Dundee Eyetracking Corpus* is composed of 20 editorial articles with 51,501 words.

The *Dundee Eyetracking Corpus* does not target a specific set of linguistic phenomena; instead, it provides naturally occurring texts for the testing of diverse hypotheses. For example, Demberg and Keller (2008) used the corpus to test Gibson's Dependency Locality Theory (DLT), (Gibson, 2008), and Hale's surprisal theory (Hale, 2001). The corpus also allows for replications to be conducted, as in Roland et al. (2012), who concluded that previous analyses (Demberg and Keller, 2007) had been distorted by the presense of a few outlier data points.

Our goal is to produce a similar resource that can serve as a shared, available foundation for research in Japanese text processing. Once completed, the corpus will allow us to address two issues that are specific to Japanese. The first issue is related to two types of reading-time measurements commonly used, namely, eyetracking and self-paced reading. Although eyetracking provides detailed recordings of eye movements, it requires specialized equipment. Self-paced reading requires only a regular computer to collect button presses, which have been shown to be an effective alternative that correlates well with

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,
pages 684–694, Osaka, Japan, December 11-17 2016.

eyetracking data in English (Just et al., 1982). However, to date, no similar correlation analyses have been conducted for Japanese. A second issue related to Japanese is that its texts do not contain spaces to mark boundaries between words (or other linguistic units such as *bunsetsu*, in other words, a content word plus functional morphology), and the question arises as to the best way to show segments in self-paced reading presentations.

Here, we present specifications and basic statistics for the *BCCWJ Eyetracking Corpus*, which makes available reading times for BCCWJ texts that have been previously annotated with syntactic and semantic tags. This should allow for detailed analyses of human text processing having a diverse range of purposes (e.g., readability measurements, evaluations of stochastic language models, engineering applications).

The rest of the paper is organized as follows. Section 2 provides basic information about the reading-time annotations, participants (§2.1), articles (§2.2), apparatus/procedure (§2.3), and data format and basic statistics (§2.4). Section 3 presents an analysis investigating a phenomenon of anti-locality (Konieczny, 2000). These are followed by the conclusion and future directions.

2 Method

2.1 Participants

Twenty-four native speakers of Japanese who were 18 years of age or older at the time, participated in the experiment for financial compensation. The experiments were conducted from September to December 2015. Profile data collected included age (in five-year brackets), gender, educational background, eyesight (all participants had uncorrected vision or vision corrected with soft contact lenses or prescription glasses), geographical linguistic background (i.e. the prefecture within Japan where they lived until the age of 15), and parents' place of birth (See Table 1 for a summary).

Vocabulary size was measured using a Japanese language vocabulary evaluation test (Amano and Kondo, 1998). Participants indicated the words they knew from a list of 50 words and scores were calculated taking word-familiarity estimates into consideration.

As a measure of working-memory capacity, the Japanese version of the reading-span test was conducted (Osaka and Osaka, 1994). Each participant read sentences aloud, each of which contained an underlined content word. After each set of sentences, the participant recalled the underlined words. If all words were successfully recalled, the set size was increased by one sentence (sets of two to five sentences were used). The final score was the largest set for which all words were correctly recalled, with a half point added if half of the words were recalled in the last trial (See Table 2 for the scores in the vocabulary and working memory tests).

Table 1: Profile data for the participants

Age range (years)	Females	Males	Gender not given	Total
-20	1	1		2
21-25		2		2
26-30	2			2
31-35	3			3
36-40	9		1	10
41-45	3			3
46-50	1			1
51-		1		1
total	19	4	1	24

Table 2: Results for reading span test and vocabulary-size test

Vocab.	Reading span test score								
size	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	Total
36,000 -		1	1						2
38,000 -		4		1					5
40,000 -	1	1							2
42,000 -		1							1
44,000 -						1			1
46,000 -									0
48,000 -			1						1
50,000 -			4	1	1		1		7
52,000 -			1					1	2
54,000 -	1								1
56,000 -									0
58,000 -			1						1
60,000 -		1							1
Total	2	8	8	2	1	1	1	1	24

2.2 Texts

Reading times were collected for a subset of the core data of the *Balanced Corpus of Contemporary Written Japanese* (BCCWJ) (Maekawa et al., 2014), consisting of newspaper articles (PN: published newspaper) samples. Articles were chosen if they were annotated with information such as syntactic dependencies, predicative clausal structures, co-references, focus of negation, and similar details, following the list of articles that were given annotation priority in the BCCWJ.[1]

The 20 newspaper articles chosen were divided into four sets of data containing five articles each: sample sets A, B, C, and D. Table 3 shows the numbers of words, sentences, and screens (i.e. pages) for each set of data. Each article was presented starting on a new screen.

Articles were shown segmented or unsegmented, that is, with or without a half-width space to mark the boundary between segments. Segments conformed to the definition for *bunsetsu* units (a content word followed by functional morphology, e.g., a noun with a case marker) in the BCCWJ as prescribed by the National Institute for Japanese Language and Linguistics. Each participant was assigned to one of the eight groups of three participants each, one group for each of the eight experimental conditions with varying combinations of measurement methods and boundary marking for different data sets presented in different orders (see Table 4). The next section provides explanations for the two measurement methods (eyetracking and self-paced reading). Order of the tasks was fixed with eye movements collected in the first session, and keyboard presses recorded during a self-paced reading presentation in the second session. Each participant saw each text once, with task and segmentation for the texts counter-balanced across participants.

Table 3: Data set sizes

Data set	Segments	Sentences	Screens
A	470	66	19
B	455	67	21
C	355	44	16
D	363	41	15

Table 4: Experimental Design

Group	Eye tracking		Self-paced reading	
1	A unseg	B seg	C unseg	D seg
2	A seg	B unseg	C seg	D unseg
3	C unseg	D seg	A unseg	B seg
4	C seg	D unseg	A seg	B unseg
5	B unseg	A seg	D unseg	C seg
6	B seg	A unseg	D seg	C unseg
7	D unseg	C seg	B unseg	A seg
8	D seg	C unseg	B seg	A unseg

'seg' stands for with spaces, and 'unseg' stands for without spaces.

2.3 Apparatus and Procedure

Eye movements were recorded using a tower-mounted EyeLink 1000 (SR Research Ltd). View was binocular but data were collected from each participant's right eye using 1000-Hz resolution. Participants looked at the display by way of a half-mirror as their heads were fixed with their chins resting on a chin rest. Unlike self-paced reading, in eyetracking all segments are shown simultaneously thus allowing more natural reading as the participant can freely return and reread earlier parts of the text on the same screen (but, participants were not allowed to return to previous screens). Stimulus texts were shown in a fixed full-width font (MS Mincho 24 point), displayed horizontally as is customary with computer displays for Japanese, with five lines per screen on a 21.5-inch display.[2] In the segmented condition, a half-width space was used to indicate the boundary between segments. In order to improve vertical tracking accuracy, three empty lines intervened between lines of text. A line break was inserted at the end of sentence or when the maximum 53 full-width characters per line was reached. Moreover, line breaks were inserted at the same points in the segmented and unsegmented conditions to guarantee that the same number of non-space characters were shown in both conditions.

The same procedure was adopted for the self-paced reading presentation, except that the chin rest was not used and participants could move their heads freely while looking directly at the display. Doug Rohde's Linger program, Version 2.94[3] was used to record keyboard-press latencies while sentences were shown using a non-cumulative self-paced moving-window presentation, which had the best correlation with eyetracking data when different styles of presentation were compared for English (Just et al., 1982). Sentence segments were initially shown masked with dashes. Participants pressed the space key of the

[1] https://github.com/masayu-a/BCCWJ-ANNOTATION-ORDER

[2] EIZO FlexScan EV2116W (resolution 1920 × 1080), set 50 cm from the chin rest.

[3] http://tedlab.mit.edu/~dr/Linger/

keyboard to reveal each subsequent segment of the sentence while all other segments reverted to dashes. Participants were not allowed to return to reread earlier segments.

Figure 1 shows two types of segmentations in the self-paced reading setting. In order to illustrate the difference between full-width underscore and half-width underscore, their heights are slightly altered in the figure. In the original Linger software presentation, these are shown at the same height.

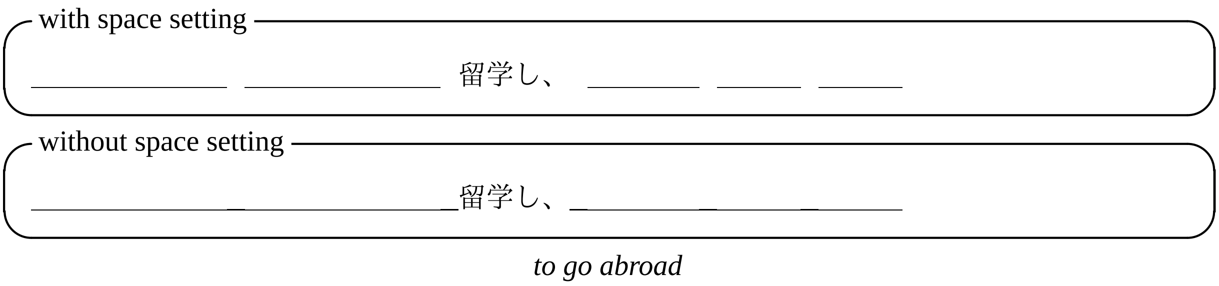


Figure 1: Types of segmentations in the self-paced reading experiment

2.4 Analysis

2.4.1 Reading-Times Tabulation

In the self-paced reading session, each segment was displayed separately, and participants could not return to reread earlier parts of the text. Therefore, the latencies for the button presses are straightforward measures of the time spent on each segment.

For the eyetracking data, five types of measurements are included, namely, First Fixation Time (FFT), First-Pass Time (FPT), Regression Path Time (RPT), Second-Pass Time (SPT), and Total Time (TOTAL), which will be explained using Figure 2.

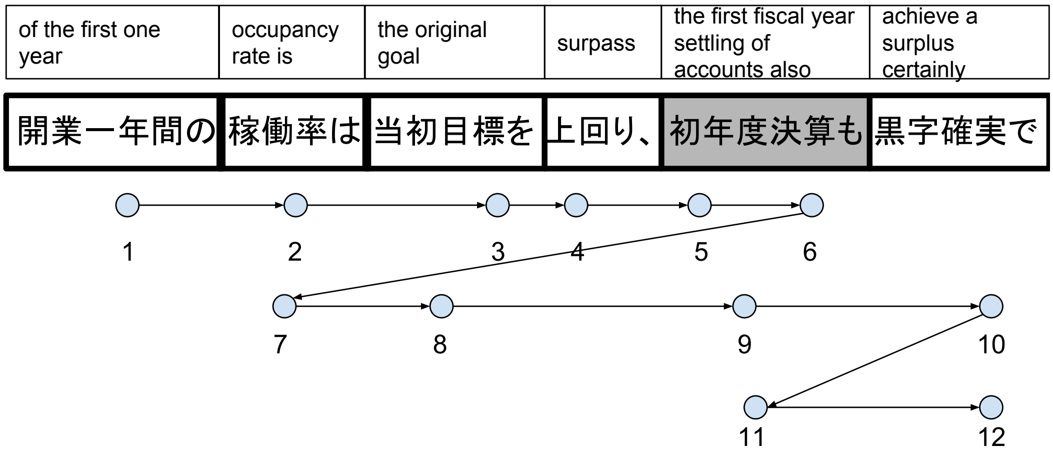


Figure 2: Example of fixations

First Fixation Time (FFT) is the fixation duration measured when the gaze first enters the area of interest. In Figure 2, the FFT for ‘the first fiscal year settling of accounts also’ (hereafter ‘the area of interest’) is the duration of fixation 5.

First-Pass Time (FPT) is the total duration of fixation from the moment the gaze first stops within the area of interest until it leaves the focus area by moving to the right or left of this area. In the figure, the FPT is the sum of the durations of fixations 5 and 6.

Regression Path Time (RPT) is the total span of from the moment the gaze enters the area of interest until it crosses the right boundary of this area for the first time. In the figure, the RPT is the sum of the

Table 5: Data format

Column name	Type	Description
`surface`	factor	Word surface form
`time`	int	Reading-time
`measure`	factor	Reading time types
`sample`	factor	Sample name
`article`	factor	Article information
`metadata_orig`	factor	Document structure tag
`metadata`	factor	Metadata
`sessionN`	int	Session order
`articleN`	int	Article display order
`screenN`	int	Screen display order
`lineN`	int	Line display order
`segmentN`	int	segment display order
`sample_screen`	factor	Screen identifier
`length`	int	Number of characters
`space`	factor	segment boundary with space or not
`setorder`	int	Segmentation-type order
`subj`	factor	Participant ID
`rspan`	num	Reading-span test score
`voc`	num	Vocabulary-test score
`dependent`	int	Number of dependents

durations for fixations 5, 6, 7, 8 and 9. The RPT can includes fixations to the left of the left boundary (e.g., 7 and 8) and durations of fixations when the gaze returns to the area of interest (e.g., 9).

Second-Pass Time (SPT) is the total span of time the gaze spend in the area of interest excluding the FPT. In the figure, the SPT is the sum of the durations of fixations for 9 and 11.

Total Time (TOTAL) is the total duration that the gaze spends within the area of interest. In other words, it is the sum of the SPT and the FPT. In the figure, TOTAL is the sum of the durations of fixations 5, 6, 9 and 11.

Only fixation times have been tabulated thus far. In the future, saccade information will also be made available.

2.4.2 Data Format and Basic Statistics

Data will be made available in tab-separated valuses (TSV) format for each of the reading-time measurements described in the previous section, along with information about the original articles and profiles of the participant. Table 5 summarizes the data format.

Word surface form (Surface: factor) refers to the text strings shown to the participants. These are organized according to the segment standards of the National Institute for Japanese Language and Linguistics, with full-width blank spaces removed.

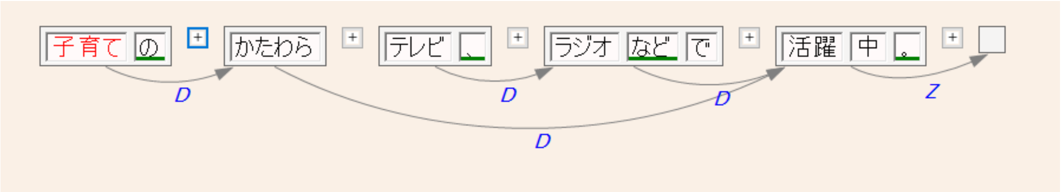
Reading time (time: int) is the `time` measurement expressed in milliseconds. For self-paced reading, this is the button-press latency for a single segment. For eyetracking, numbers are provided for each of the five measurements discussed in the previous section: First Fixation Time (FFT), First-Pass Time (FPT), Second-Pass Time (SPT), Regression Path Time (RPT) and Total Time (TOTAL). The *reading time types* (`measure`: factor) are defined as {'Self-Paced', 'EyeTrack: FFT', 'EyeTrack: FPT', 'EyeTrack: SPT', 'EyeTrack: RPT', 'EyeTrack: Total'}.

There are four types of information provided for the newspaper articles: `sample`, `article`, `metadata_orig` and `metadata`. The *sample name* (`sample`: factor) is derived from the data sets prepared for each session 'A', 'B', 'C', 'D'; each sample consists of five newspaper articles. *Article in-*

formation (`article`: factor) is a unique identifier for the individual articles, which is connected with an underscore to the BCCWJ annotation priority rankings, the BCCWJ internal sample IDs, and the article numbers. *Document structure tag* (`metadata_orig`: factor) is a BCCWJ internal document structure tag, which is connected with the tag information in the BCCWJ XML ancestor axis using a slash. *Metadata* (`metadata`: factor) is generated through the extraction of the properties of the article from the previously mentioned `metadata_orig`. It is set to one of {'authorsData', 'caption', 'listItem', 'profile', 'titleBlock', or 'undefined'}, and indicates manual revisions of mistakes or omissions in the BCCWJ internal document structure tag.

There are five types of information related to presentation order. *Session order* (`session`: int) indicates the session number (1 or 2). *Article display order* (`articleN`: int) indicates the article display sequence (1–5) within each session. *Screen display order* (`screenN`: int) indicates the screen's display sequence number within each article. *Line display order* (`lineN`: int) indicates the line number within each screen (1–5). Segment display order (`segmentN`: int) indicates the segment sequence number within each line.

Screen identifier (`sample_screen`: factor) is a unique identifier for the screens displayed to the participants. *Number of characters* (`length`: int) is the number of characters in the segment. *Segmented or unsegmented* (`space`: factor) indicates whether there is a half-width space between segment units ('1'), or not ('0'). *Segmentation-type order* (`setorder`: factor) is set to '0-1' if the participant saw unsegmented texts followed by segmented texts, and it is set to '1-0' otherwise. *Number of dependents* (`dependent`: int) is the number of segment units that are syntactically dependent on the current segment. Segment dependency relationships were annotated manually. Figure 3 shows an example of a dependency annotation on segments. Note, the dependency arcs are written from dependent to head following convention in Japanese annotations.



She raises her twins and is also active as a broadcaster of TV and radio programs.

Figure 3: Example of Dependency Annotation

There are three types of information assigned for each participant. *Experiment participant ID* (`subj`: factor) is a unique identifier for each participant, and is associated with two pieces of information. The first is the *reading span test score* (`rspan`: num), ranging from 1.5 to 5.0 in gradations of 0.5. The second is the *Vocabulary test score* (`voc`: num), which is the original result divided by 1,000 (37.1-61.8).

Table 6 shows means, standard deviations (SDs), and quartiles for each measurement. For eyetracking, the numbers shown exclude reading times of zero milliseconds (i.e. instances where segments were not fixated).

After each article, a simple yes-no question verified readers' comprehension. Overall accurary was 88.5% (ranging from 70%-100%). Accuracy was higher in eye-tracking (99.2%; 238/240) than in self-paced reading (77.9%; 187/240: $p < 0.001$). One possible factor favoring eye-tracking is that participants could reread texts freely. Another factor is that self-paced reading data was always collected in the second session, therefore participants may have been more tired.

3 Example Analysis

3.1 Anti-locality

As an example of the possible uses for the corpus, we conducted analyses investigating *anti-locality* phenomena, in which a head is read faster if it is preceded by more dependents as first reported for

Table 6: Mean reading times (in milliseconds)

	Mean	SD	Min.	1st Qu.	Median	3rd Qu.	Max.
Self-paced	699	506	62	415	550	798	9454
Eye tracking (excl. 0)							
First Fixation Time	235	142	12	162	219	292	1700
First-Pass Time	475	497	14	205	321	548	7340
Second-Pass Time	330	253	20	173	258	418	2553
Regression Path Time	698	1013	19	235	391	745	21577
Total Time	597	589	18	247	416	721	8397

German (Konieczny, 2000; Konieczny and Döring, 2003) and later replicated for Japanese (Uchida et al., 2014). Such speedu gains in head-final constructions are not easily explained by working-memory models, which either predict that attaching a large number of dependents should be costly (Gibson, 2008) or predict that the cost of an upcoming head should not be affected by the number of dependents preceding it (Nakatani and Gibson, 2010). Although the phenomenon is compatible with surprisal (Hale, 2001; Levy and Gibson, 2013), previous results were limited to reports that a ditransitive verb was read more quickly when preceded by two dependents (an accusative-marked argument and a dative-marked argument) than when preceded by just one dependent (the accusative argument). Therefore, these results do not necessarily support *anti-locality*, instead they may be related to ditransitive verbs being more natural when the dative noun phrase is expressed overtly. We report corpus analyses that show that *anti-locality* holds more generally.

3.2 Modeling Results

Linear mixed models were constructed for reading times to the main texts of the articles (i.e. excluding reading times that had the metadata field labelled as authorsData, caption, listItem, profile, or titleBlock). The first and last segments on a line may be exceptional as they may be affected by large eye movements going from the end of the line to the beginning of the following line, or backtracking to reread content at the end of the previous line. Therefore, factors were included in the analyses encoding whether the segment is the first (`is_first`), second to last (`is_second_last`) or last (`is_last`) on a line. We also excluded zero-millsecond data points from the eyetracking data. Because the models with maximal or close-to-maximal random structure did not converge, we performed forward model selections for each time setting and report the model with the smallest AIC that converged. After model-based trimming was used to eliminate points beyond three standard deviations, the model was rebuilt (Baayen, 2008). Tables 7, 8, 9, 10, 11, and 12 show the results of the smallest-AIC models for ‘Self-Paced Reading (Self)’, and eyetracking data for ‘First Fixation Time (FFT)’, ‘First-Pass Time (FPT)’, ‘Second-Pass Time (SPT)’, ‘Regression Path Time (RPT)’, and ‘Total Times (Total)’, respectively.

In the tables, the baseline (i.e. the intercept) encodes the False value for binary factors. Therefore, a factor name followed by ‘1’ (e.g., `is_first1`) indicates what happens to the model prediction when the factor changes from FALSE to TRUE. For example, in Table 7, the intercept is the baseline (634.08 ms) which excludes reading times to the first, penultimate, and last segments (i.e. `is_first=FALSE`, `is_second_last=FALSE`; `is_last=FALSE`). Therefore, the row starting with `is_first1` (i.e. it is the first segment of the line) indicates that the first segment was 69.73 ms slower than the segments included in the baseline.

Table 13 shows the summary of the results from the linear mixed model. In this table, if the absolute *t*-value of the effect is larger than 1.96, we regard the factor as statistically significant and put the sign of the estimate. Otherwise, we put 0, indicating nonsignificant factors.

Texts presented that were segmented with a blank space had shorter first pass times, second pass times, total-reading times than unsegmented texts (factor `space`). These results are interesting because texts are usually unsegmented in Japanese writing, therefore the result is the opposite of what would be expected based on participants’ reading habits. The result is also not compatible with previous results,

Table 7: Parameters of the linear mixed model for the self-paced reading data

	Estimate	Std. Err.	*t* value
Intercept	634.08	31.05	20.42
space1	3.04	4.03	0.75
sessionN	-46.50	27.10	-1.72
length	159.35	2.22	71.74
dependent	-27.00	2.26	-11.96
is_first1	69.73	6.56	10.63
is_last1	-37.93	6.61	5.73
is_second_last1	-8.22	5.88	-1.40
articleN	-40.84	14.45	-2.83
screenN	-42.56	2.74	-15.49
lineN	-19.36	2.14	-9.06
segmentN	-11.76	3.56	-3.31
space1:sessionN	2.34	54.04	0.04

316 data points (1.79%) were excluded in the 3-SD trimming. We choose the converged model with smaller AIC as follows: `lmer (time ~ space * sessionN + length + dependent + is_first + is_last + is_second_last + articleN + screenN + lineN + segmentN + (1 + articleN + segmentN | subj) + (1 + articleN | article)`

Table 9: Parameters of the linear mixed model for first-pass time

	Estimate	Std. Err.	*t* value
Intercept	421.54	24.32	17.33
space1	-18.04	4.97	-3.73
sessionN	-24.51	17.31	-1.42
length	171.74	2.70	63.55
dependent	-32.16	2.71	-11.85
is_first1	83.53	7.62	10.96
is_last1	9.06	7.95	1.14
is_second_last1	23.14	7.12	3.25
articleN	-1.81	8.91	-0.20
screenN	-19.38	3.21	-6.15
lineN	-19.86	2.55	-7.81
segmentN	-4.26	5.58	-0.76
space1:sessionN	21.47	34.43	0.62

234 data points (1.76%) were excluded in the 3-SD trimming. We choose the converged model with smaller AIC as follows: `lmer (time ~ space * sessionN + length + dependent + is_first + is_last + is_second_last + articleN + screenN + lineN + segmentN + (1 + articleN + segmentN | subj) + (1 + articleN + segmentN | article)`

Table 8: Parameters of the linear mixed model for first fixation time

	Estimate	Std. Err.	*t* value
Intercept	227.00	7.39	30.86
space1	-3.01	1.70	-1.77
sessionN	-11.81	6.70	-1.76
length	-1.38	0.88	-1.57
dependent	-4.81	0.93	-5.15
is_first1	13.18	2.60	5.07
is_last1	-3.11	2.71	-1.15
is_second_last1	3.35	2.43	1.38
articleN	-0.57	1.81	-0.32
screenN	-1.15	1.09	-1.06
lineN	-5.24	0.87	-6.00
segmentN	3.436	1.67	2.06
space1:sessionN	22.32	13.29	1.68

170 data points (1.28%) were excluded in the 3-SD trimming. We choose the converged model with smaller AIC. `lmer (time ~ space * sessionN + length + dependent + is_first + is_last + is_second_last + articleN + screenN + lineN + segmentN + (1 + articleN + segmentN | subj) + (1 + articleN | article)`

Table 10: Parameters of the linear mixed model for second-pass time

	Estimate	Std. Err.	*t* value
Intercept	317.14	12.07	26.28
space1	-26.73	5.86	-4.56
sessionN	-17.87	10.19	-1.75
length	16.72	3.02	5.54
dependent	-13.52	3.30	-4.09
is_first1	-21.05	8.34	-2.52
is_last1	-15.79	9.88	-1.60
is_second_last1	38.30	8.97	4.27
articleN	1.17	4.57	0.26
screenN	-9.29	3.54	-2.63
lineN	-12.30	2.97	-4.13
segmentN	-18.02	3.77	-4.78
space1:sessionN	28.49	19.88	1.43

77 data points (1.61%) were excluded in the 3-SD trimming. We choose the converged model with smaller AIC as follows: `lmer (time ~ space * sessionN + length + dependent + is_first + is_last + is_second_last + articleN + screenN + lineN + segmentN + (1 + articleN + segmentN | subj) + (1 + is_last + is_second_last + articleN | article)`

in which segmentation did not have a reliable effect in texts mixing *kanji* and *kana* characters (Sainio et al., 2007), but that may have been due to lack of statistical power or perhaps because the segmented texts were too short for participants to accommodate to this type of presentation and use segmentation information effectively.

An unsurprising finding is that longer segments (i.e. segments having more characters) took longer to read (factor `length`) except for the first fixation time. The result suggests that longer segments do not require longer first fixation, but nevertheless affect later measures as they may require further fixations.

Compared to the intermediate segments (i.e. second segment to the antepenultimate segments) on each line, longer reading times were observed for the first segment (`is_first=TRUE`; in self-paced reading, first fixation, first pass, and total reading time), for the penultimate segment

Table 11: Parameters of the linear mixed model for regression path time

	Estimate	Std. Err.	*t* value
Intercept	560.07	28.63	19.57
space1	-10.29	9.89	-1.04
sessionN	-57.59	23.14	-2.49
length	173.18	5.21	33.22
dependent	-10.86	5.49	-1.98
is_first1	8.37	15.13	0.55
is_last1	205.88	16.04	12.84
is_second_last1	7.38	14.27	0.52
articleN	2.35	13.61	0.17
screenN	-13.75	6.13	-2.24
lineN	21.90	5.09	4.31
segmentN	-34.21	11.22	-3.05
space1:sessionN	59.45	45.55	1.31

219 data points (1.65%) were excluded in the 3-SD trimming. We choose the converged model with smaller AIC as follows:

```
lmer (time ˜ space * sessionN + length + dependent
+ is_first + is_last + is_second_last + articleN +
screenN + lineN + segmentN + (1 + articleN + segmentN
| subj) + (1 + articleN | article)
```

Table 12: Parameters of the linear mixed model for total time

	Estimate	Std. Err.	*t* value
Intercept	549.09	29.74	18.46
space1	-36.16	6.35	-5.70
sessionN	-24.76	20.96	-1.18
length	198.62	3.41	58.21
dependent	-41.04	3.45	-11.90
is_first1	-79.43	9.66	8.23
is_last1	-11.08	10.08	-1.10
is_second_last1	43.58	9.04	4.82
articleN	-7.90	12.87	-0.61
screenN	-31.50	4.13	-7.62
lineN	-23.60	3.24	-7.29
segmentN	-28.21	6.06	-4.65
space1:sessionN	6.68	42.65	0.16

232 data points (1.75%) were excluded in the 3-SD trimming. We choose the converged model with smaller AIC as follows:

```
lmer (time ˜ space * sessionN + length + dependent
+ is_first + is_last + is_second_last + articleN +
screenN + lineN + segmentN + (1 + articleN + segmentN
| subj) + (1 + articleN | article)
```

Table 13: Summary of the results from the linear mixed models

	Self	FFT	FPT	SPT	RPT	Total
space=True	0	0	-	-	0	-
length	+	0	+	+	+	+
is_first=True	+	+	+	-	0	+
is_last=True	+	0	0	0	+	0
is_second_last=True	0	0	+	+	0	+
articleN	-	0	0	0	0	0
screenN	-	0	-	-	-	-
lineN	-	-	-	-	-	-
segmentN	-	+	0	-	-	-
dependent	-	-	-	-	-	-

(`second_last_bunsetsu=TRUE`; in first pass time, second pass time, and total time), and for the last segment (`last_bunsetsu=TRUE`; in self-paced reading and regression path time).

Within a session, reading times from self-paced reading became faster with each article (`articleN`); however, the effect was not reliable in any of the eye-tracking measures. Within an article, reading times got faster with each screen (`screenN`) in all measures except for first fixation time. In the vertical ordering within a screen, all reading times got faster with each line (`lineN`). In the horizontal ordering within a line, reading times except for first fixation time and first pass time became faster with each segment (`segmentN`). These speed gains are expected as readers gain speed as they process more information.

Apart from the effects described above related to the physical aspects of the presentation of the texts, we also observed a reliable *anti-locality* effect as words were read faster when more dependents preceded them (factor `dependent`). This generalizes previous findings (Konieczny, 2000; Konieczny and Döring, 2003; Uchida et al., 2014) and confirms that dependent phrases provide information that facilitates the processing of an upcoming head.

4 Conclusion

We created a data set with the reading times of 24 native speakers of Japanese. Preliminary analyses illustrate the uses of this type of data. First, although spaces are not commonly used to segment Japanese text, readers were nevertheless faster to read segmented texts. Second, we reported an analysis on *anti-locality* effects, which confirmed previous reports and generalized them to more natural texts.

The reading time data, excluding the original texts, will be licenced through Creative Commons Attribution-Noncommercial 4.0 (CC BY-NC 4.0: `https://creativecommons.org/licenses/by-nc/4.0/`). Apart from the data files described in Section 2.4.2, the original eye-tracking data can be obtained as EyeLink Data Viewer files, by contacting the first author. The original texts can be obtained by purchasing the BCCWJ DVD edition `http://pj.ninjal.ac.jp/corpus_center/bccwj/en/subscription.html`.

Future planned developments are as follows. First, we will extend the corpus with more participants and data. This initial data set was restricted to newspaper articles, and we are currently investigating the possibility of assigning reading times to other texts, such as books and magazines.

Other types of annotations will be added. Apart from information on the number of dependents already available in the current data, we are considering including other types of information such as dependency length, scope of coordinate structure. Other types of information that may be added in the future include morphological information such as word class, vocabulary classification table number, predicate clause structure (*ga*-case:subj, *o*-case:dobj, *ni*-case:iobj), co-reference information, clause boundary information, and information structure.[4]

Finally, we intend to examine possible applications for information processing. Participants were required to write a summary for each text they read. Contrast analysis of the reading times and the summaries may allow us to augment automatic summarization systems tailored to individual readers.

Acknowledgements

This research received assistance from the JSPS Grants-in-Aid for Scientific Research Fund (B) 25284083 (Reading Time Annotation for the Language Corpus and its Uses). We would like to thank Douglas Roland, Manabu Arai and four anonymous reviewers for their comments on the earlier draft.

[4] Parts of such annotation are already available at `https://bccwj-data.ninjal.ac.jp/mdl/`

References

S. Amano and T. Kondo. 1998. Estimation of mental lexicon size with word familiarity database. In *Proceedings of International Conference on Spoken Language Processing*, volume 5, pages 2119–2122.

R. H. Baayen. 2008. *Analyzing Linguistic Data: A practical Introduction to Statistics using R*. Cambridge University Press.

V. Demberg and F. Keller. 2007. Eye-tracking evidence for integration cost effects in corpus data. In *Proceedings of the 29th Meeting of the Cognitive Science Society (CogSci-07)*, pages 947–952.

V. Demberg and F. Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

E. Gibson. 2008. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1–76.

J. Hale. 2001. 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.

M. A. Just, P. A. Carpenter, and J. D. Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 3:228–238.

A. Kennedy and J. Pynte. 2005. Parafoveal-on-foveal effects in normal reading. *Vision Research*, 45:153–168.

L. Konieczny and P. Döring. 2003. Anticipation of clause-final heads. evidence from eye-tracking and srns. In *Proceedings of the 4th International Conference on Cognitive Science*.

L. Konieczny. 2000. Locality and parsing complexity. *Journal of Psycholinguistic Research*, 29(6).

R. Levy and E. Gibson. 2013. Surprisal, the pdc, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology*, 4(229).

K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka, and Y. Den. 2014. Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 48:345–371.

K. Nakatani and E. Gibson. 2010. An on-line study of japanese nesting complexity. *Cognitive Science*, 34(1):94–112.

M. Osaka and N. Osaka. 1994. [working memory capacity related to reading: measurement with the japanese version of reading span test] (in japanese). *Shinrigaku Kenkyu: The Japanese Journal of Psychology*, 65(5):339–345.

D. Roland, G. Mauner, C. O'Meara, and H. Yun. 2012. Discourse expectations and relative clause processing. *Journal of Memory and Language*, 66(3):479–508.

M. Sainio, J. Hyöna, K. Bingushi, and R. Bertram. 2007. The role of inter spacing in reading japanese: An eye movement study. *Vision Research*, 47:2575–2584.

S. Uchida, E. T. Miyamoto, Y. Hirose, Y. Kobayashi, and T. Ito. 2014. An erp study of parsing and memory load in japanese sentence processing – a comparison between left-corner parsing and the dependency locality theory –. In *Proceedings of the Thought and Language/the Mental Architecture of Processing and Learning of Language 2014*.