Between Reading Time and Information Structure

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

Abstract

This paper presents a contrastive analysis between the reading time and information structure in Japanese. We overlaid the reading time annotation BCCWJ-EyeTrack and an information structure annotation on the Balanced Corpus of Contemporary Written Japanese. Statistical analysis based on a mixed linear model showed that the "specificity," "sentience," and "commonness" of the Japanese information structure affect the reading time. These three characteristics produce different patterns of delay in the reading time. Especially, the reading time patterns differ depending on the commonness such as new information or bridging. The results suggest that new information and bridging can be classified by the reading time pattern.

1 Introduction

Information structures[1] such as definiteness and specificity affect the selection of articles. However, some languages such as Japanese and Russian do not have articles for noun phrases (NPs). The definiteness and specificity in such languages are not overtly marked in their surface form.

Nagata et al. (Nagata et al., 2005) proposed a statistical model to detect article errors made by English learners. They constructed the model by using an enormous text produced by English native speakers. However, when we consider the choice of articles in some languages by native speakers of a language with no articles, careful attention must be paid to the information structures in their NPs. Moreover, when texts in language without articles are translated by humans or machine translation systems into a target language with articles, we should consider the information structure of the source language.

Other aspects of information structures are the information status and commonness. The information status indicates whether the co-referred mention appears in the preceding discourse (`discourse-old`) or not (`discourse-new`). Commonness is defined as whether the speaker assumes that the hearer already knows the information (`hearer-old`), can estimate the information (`bridging`), or does not know the information `hearer-new`. Bridging reference cannot be resolved purely from the surface forms of written texts. Previous research (Hou et al., 2013) tried to resolve bridging references by using annotated corpora and world knowledge. However, bridging is an information structure for the language recipient (hearer). It should incorporate recipient features such as the reading time to estimate whether an NP is `bridging` or not (i.e., `hearer-new`). Moreover, identifying if readers can resolve a bridging reference with their own knowledge is important for user-oriented information extraction and document summarization.

These information structures are correlated with animacy, sentience, and agentivity. The features of information structures may affect the reading time; there are various ways to monitor the reading time, such as eye tracking to obtain gaze information.

This paper presents a contrastive analysis between the reading time and information structures in Japanese in order to classify information structures according to the gaze information. We over-

[1] We focus on the information status which is a fundamental notion to define the *theme* and *rheme* in the information structure.

laid the reading time annotation BCCWJ-EyeTrack (Asahara et al., 2016) and information structure annotation on the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014). We performed statistical analysis on the overlaid data. The results showed that different patterns of reading time can be observed to determine variations in the information structure.

2 Related Work

First, we present related work on eye tracking. The Dundee Eyetracking Corpus (Kennedy and Pynte, 2005) contains reading times for English and French newspaper editorials from 10 native speakers of each language that were recorded by using eye-tracking equipment. The corpus does not target a specific set of linguistic phenomena but instead provides naturally occurring texts for testing diverse hypotheses. For example, Demberg and Keller (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; for example, Roland et al. (Roland et al., 2012) concluded that previous analyses (Demberg and Keller, 2007) had been distorted by the presence of a few outlier data points.

Second, we present related work on information structure annotation. Götze et al. (Götze et al., 2007) devised criteria for annotating the information status (given/accessible/new), topic (aboutness/frame setting), and focus (new-information focus/contrastive focus) independently of languages and linguistic theories. Prasad et al. (Prasad et al., 2015) discussed the bridging annotation standard of the Penn Discourse Treebank (Miltsakaki et al., 2004) and PropBank (Palmer et al., 2005).

Third, we present language analyses or models with reading time or eye tracking gaze information. Barret et al. (Barrett et al., 2016) presented a POS tagging model with gaze patterns. Klerke et al. (Klerke et al., 2015) presented a grammaticality detection model for machine processed sentences. Iida et al. (Iida et al., 2013) presented an analysis of eye-tracking data for the annotation of predicate–argument relations.

Our paper is slightly different from these preceding papers. We present corpus-based psycholinguistic research on the relationship between the information structure and reading time, including gaze information, of the language recipient.

Table 1: Data format

Column Name	Type	Description
`surface`	factor	Word surface form
`time`	int	Reading-time
`logtime`	num	Reading-time (log)
`measure`	factor	Reading time type
`sample`	factor	Sample Name
`article`	factor	Article information
`metadata_orig`	factor	Document structure tag
`metadata`	factor	Metadata
`length`	int	Number of characters
`space`	factor	Boundary with space or not
`subj`	factor	Participant ID
`setorder`	factor	Segmentation type order
`rspan`	num	Result of reading span test
`voc`	num	Result of vocabulary test
`dependent`	int	Number of Dependents
`sessionN`	int	Session order
`articleN`	int	Article display order
`screenN`	int	Screen display order
`lineN`	int	Line display order
`segmentN`	int	Segment display order
`is_first`	factor	The leftmost segment
`is_last`	factor	The rightmost segment
`is_second_last`	factor	The second rightmost segment
`infostatus`	factor	Information status
`definiteness`	factor	Definiteness
`specificity`	factor	Specificity
`animacy`	factor	Animacy
`sentience`	factor	Sentience
`agentivity`	factor	Agentivity
`commonness`	factor	Commonness

3 Data and Method

We used the overlaid data of BCCWJ-EyeTrack and information structure annotations, as given in Table 1. We present the data below in detail.

3.1 Reading Time Data: BCCWJ-EyeTrack

We now explain the two measurement methods for estimating the reading time: eye tracking and self-paced reading. The order of tasks was fixed with eye tracking in the first session and a self-paced reading method in the second session. Each participant saw each text once with the task and segmentation of the texts counterbalanced across participants.

Eye tracking was recorded with a tower-mounted EyeLink 1000 (SR Research Ltd). The view was binocular, but data were collected from each participant's right eye at a resolution of 1000 Hz. Participants looked at the display by using a half-mirror; their heads were fixed with their chins on a chin rest. Unlike self-paced reading, during eye tracking all segments were shown Simultaneously. This allowed more natural reading because each participant could freely return and reread earlier parts of the text on the same screen However, participants were not allowed to return to previous screens. Stimulus texts were shown in a fixed full-width font (MS Mincho 24 point) and displayed horizontally as is customary with computer displays for Japanese; there were five lines per screen on a 21.5-in display.[2] Under the segmented condition, a half-width space was used to indicate the boundary between segments. In order to improve the vertical tracking accuracy, three empty lines were placed between lines of text. A line break was inserted at the end of a 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 under both conditions. Figure 1 shows the screen dump of the eye tracking results.

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 by using a non-cumulative self-paced moving-window presentation. This had the best correlation with eye-tracking 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 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.

Twenty-four native Japanese speakers 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. The collected profile data included the age (in 5-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. Table 2 shows the profile data for the participants. The vocabulary size of the participants was measured by using a Japanese language vocabulary evaluation test (Amano and Kondo, 1998). Participants indicated words they knew from a list of 50 words, and scores were calculated by taking word-familiarity estimates into consideration. As a measure of the 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. Table 3 shows the results of the tests. 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; a half point was added if half of the words were recalled in the last trial.

Reading times were collected for a subset of the core data of the BCCWJ (Maekawa et al., 2014), which consisted 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.

The 21 newspaper articles[4] chosen were divided into four datasets containing five articles each: A, B, C, and D. Table 4 presents the numbers of words, sentences, and screens (i.e., pages) for each dataset. Each article was presented as starting on a new screen.

Articles were shown segmented or unsegmented (i.e., with or without a half-width space to mark the boundary between segments). Segments con-

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

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

[4]The original BCCWJ-EyeTrack paper (Asahara et al., 2016) presented 20 articles. However, there were two consecutive articles in dataset C. These two articles were presented on separate screens. Thus, we split them into two for statistical analysis.

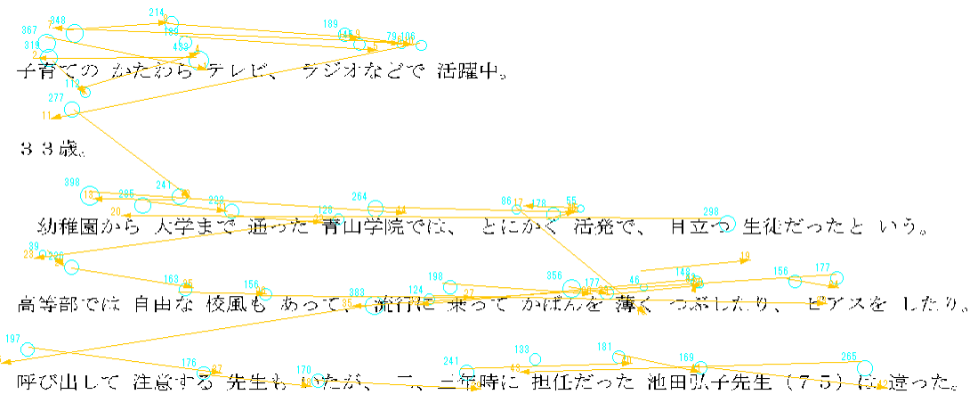

Figure 1: Screen Dump of the Eye Tracking Results

Table 2: 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 3: Results for reading span test and vocabulary-size test

Vocab. size	Reading span test score								Total
	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	
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

formed 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 eight groups of three participants each. Each group was subjected to one of the eight experimental conditions with varying combinations of measurement methods and boundary marking for different datasets presented in different orders.

During 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 eye-tracking data, five types of measurements were used: first fixation time (FFT), first pass time (FPT), regression path time (RPT), second pass time (SPT), and total time (TOTAL). These are explained in Figure 2.

The FFT is the fixation duration measured when the gaze first enters the area of interest. In the figure, the FFT for “the first fiscal year settling of accounts also” (hereafter “the area of interest”) is the duration of fixation 5.

Table 4: Data set sizes

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

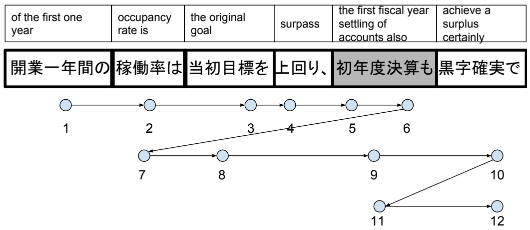

Figure 2: Example of fixations

The 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.

The 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 durations for fixations 5–9. 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).

The SPT is the total span of time the gaze spends in the area of interest excluding the FPT. In the figure, the SPT is the sum of the durations of fixations 9 and 11.

The 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 FPT. In the figure, TOTAL is the sum of the durations of fixations 5, 6, 9, and 11.

Table 1 presents the data. `surface` is the surface form of the word. The reading time (i.e., `time`) is converted into log scale (i.e., `logtime`). `measure` is the reading type {SELF, FFT, FPT, RPT, SPT, TOTAL}. `sample, article, metadata_orig, metadata` are information related to the article. `length` is the number of characters in the surface form. `space` is whether spaces are present between segments. `subj` is the participant ID, which is used as a random effect for the statistical analysis. `setorder` is the presentation order of the space. `rspan, voc` are features of the participants. `dependent` is the number of dependents for the segments. The dependency relation is annotated by humans (Asahara and Matsumoto, 2016). `sessionN, articleN, screenN, lineN, segmentN` are the display order of the elements. `is_first,is_last,is_second_first` are the layout features on the screen.

Table 5: Basic statistics of information structure labels

`infostatus`	discourse-new		discourse-old	
	1345		678	
`definiteness`	definite	indefinite	either	-
	1122	899	2	-
`specificity`	specific	unspecific	either	neither
	1157	749	116	1
`animacy`	animate	inanimate	either	-
	342	1680	1	-
`sentience`	sentient	insentient	either	-
	337	1678	8	-
`agentivity`	agent	patient	both	neither
	192	338	2	1491
`commonness`	hearer-new	hearer-old	bridging	neither
	494	1036	489	4

Table 6: Parameters of the linear mixed model for the self paced reading time (SELF) (`logtime`)

	Estimate	Std. Error	t value
`(Intercept)`	2.893	0.062	46.51
`length`	0.102	0.002	**42.31**
`space=T`	0.003	0.004	0.86
`dependent`	-0.005	0.003	-1.61
`sessionN`	-0.021	0.022	-0.94
`articleN`	-0.023	0.007	**-3.23**
`screenN`	-0.032	0.002	**-11.19**
`lineN`	-0.014	0.002	**-6.10**
`segmentN`	-0.005	0.001	**-4.83**
`is_first=T`	0.047	0.006	**7.19**
`is_last=T`	0.040	0.008	**4.71**
`is_second_last=T`	-0.011	0.005	**-2.11**
`space=T:sessionN`	-0.019	0.044	-0.43
`is=discourse-old`	-0.005	0.005	-0.98
`def=indefinite`	0.004	0.015	0.30
`spec=specific`	0.044	0.016	**2.78**
`spec=unspecific`	0.001	0.010	0.16
`ani=inanimate`	-0.000	0.050	-0.02
`sent=insentient`	-0.105	0.067	-1.56
`sent=sentient`	-0.098	0.050	-1.94
`ag=both`	-0.058	0.049	-1.18
`ag=neither`	-0.004	0.007	-0.69
`ag=patient`	-0.013	0.008	-1.63
`com=hearer-new`	0.025	0.007	**3.59**
`com=hearer-old`	-0.020	0.009	**-2.11**
`com=neither`	0.000	0.025	0.01

45 data points (0.69%) were excluded in the 3-SD trimming.

Table 7: Parameters of the linear mixed model for the first pass time (FPT) (`logtime`) (only information structure related)

	Estimate	Std. Error	t value
`(Intercept)`	2.303	0.102	22.53
`is=discourse-old`	0.005	0.010	0.50
`def=indefinite`	0.024	0.026	0.90
`spec=specific`	0.064	0.028	**2.26**
`spec=unspecific`	0.031	0.018	1.70
`ani=inanimate`	0.210	0.104	**2.01**
`sent=insentient`	-0.001	0.129	-0.01
`sent=sentient`	0.194	0.086	**2.25**
`ag=both`	-0.050	0.087	-0.57
`ag=neither`	0.014	0.012	1.19
`ag=patient`	-0.006	0.015	-0.43
`com=hearer-new`	0.024	0.012	1.95
`com=hearer-old`	0.000	0.017	-0.03
`com=neither`	0.002	0.043	0.05

13 data points (0.24%) were excluded in the 3-SD trimming.

Table 8: Parameters of the linear mixed model for the regression path time (RPT) (`logtime`) (only information structure related)

	Estimate	Std. Error	t value
`(Intercept)`	2.188	0.118	18.48
`is=discourse-old`	-0.003	0.011	-0.30
`def=indefinite`	0.041	0.030	1.34
`spec=specific`	0.095	0.032	**2.95**
`spec=unspecific`	0.038	0.020	1.82
`ani=inanimate`	0.112	0.119	0.94
`sent=insentient`	0.248	0.150	1.65
`sent=sentient`	0.345	0.102	**3.37**
`ag=both`	-0.054	0.100	-0.54
`ag=neither`	0.013	0.014	0.91
`ag=patient`	-0.000	0.017	-0.01
`com=hearer-new`	0.001	0.014	0.09
`com=hearer-old`	-0.018	0.019	-0.94
`com=neither`	0.042	0.049	0.86

43 data points (0.81%) were excluded in the 3-SD trimming.

3.2 Information Structure Annotation

This subsection presents information structure annotation. The detailed explanation is in (Miyauchi et al., 2017) We set the following seven labelset for the NP segments:

- information status (`infostatus:is`)
- definiteness (`definiteness:is`)
- specificity (`specificity:spec`)
- animacy (`animacy:ani`)
- sentience (`sentience:sent`)
- agentivity (`agentivity:ag`)
- commonness (`commonnness:com`)

The information status is the distinction between new and old/given information. In some discourses, the information that the speaker wants to convey to the hearer is `discourse-new`, and the information that the hearer already knows is `discourse-old`.

Table 9: Parameters of the linear mixed model for the total time (TOTAL) (`logtime`) (only information structure related)

	Estimate	Std. Error	t value
`(Intercept)`	2.500	0.105	23.69
`is=discourse-old`	0.009	0.010	0.89
`def=indefinite`	0.036	0.027	1.32
`spec=specific`	0.070	0.029	**2.39**
`spec=unspecific`	0.016	0.019	0.88
`ani=inanimate`	0.177	0.108	1.63
`sent=insentient`	-0.027	0.133	-0.20
`sent=sentient`	0.130	0.089	1.46
`ag=both`	-0.025	0.091	-0.28
`ag=neither`	0.006	0.013	0.50
`ag=patient`	-0.011	0.015	-0.70
`com=hearer-new`	0.030	0.013	**2.34**
`com=hearer-old`	-0.000	0.017	-0.02
`com=neither`	0.033	0.045	0.74

5 data points (0.09%) were excluded in the 3-SD trimming.

The definiteness is a semantic category about whether it is possible for hearers to identify referents (Lyons, 1999) and (Heim, 2011). An NP referent that a speaker considers to be identifiable for the hearer is `definite`, and an NP referent that the speaker does not consider to be identifiable is `indefinite`. In this research, the scope of definiteness was set to be the range before and after three sentences.

The specificity is a semantic category about whether speakers think of specific referents or not (von Heusinger, 2011). An NP is `specific` when its referent is either regarded to be unique or is specified by speakers. An NP is `unspecific` when its referent is not. Similar to the definiteness, the scope of specificity is set to be the range before and after three sentences for the annotation.

The animacy is a category about whether referents are alive. Living creatures (e.g., human beings, animals) are `animate`, and nonliving objects (including plants) are `inanimate`. In our research, the tags of animacy were annotated judging solely from the NP in question. The sentience may be addressed as a concept of animacy. This parameter expresses whether referents have emotion. An NP is `sentient` when its referent moves of its own free will and `insentient` when its referent does not.

For example, the choice of verbs *aru / iru* (i.e., “exist”) is based not so much on a distinction between animate and inanimate but between sentient and insentient. Thus, we needed to set the parameter of sentience. Because there are cases in which the presence or absence of sentience cannot be determined, the tags of sentience were annotated judging from not only the NP but the predicate of the sentence in question.

The agentivity show the roles played in a situation by those related to the situation. An NP whose referent intentionally performs an act is an `agent`, and an NP whose referent undergoes a change from an act is a `patient`. The tags of agentivity were annotated judging from both the main and subordinate clauses, including the NP in question. Furthermore, we introduced the `both` and `neither` tags.

The commonness is a parameter expressing whether the speaker assumes that the hearer already knows the information. Information that the speaker thinks the hearer already knows is `hearer-old`, and information that the speaker does not think the hearer already knows is `hearer-new`. In other words, `hearer-old` information is common ground for both the speaker and hearer, and `hearer-new` information is not. In addition to these two labels, we introduced the label `bridging` for when the NP is a bridging anaphora. Note that, when annotators judge commonness, they may use their worldly knowledge.

Table 5 presents the basic statistics of information structure labels. We introduced `either` for the definiteness, specificity, animacy, sentience, and agentivity when the annotator cannot judge the label from the limited contextual information. We also introduced `neither` for the specificity, agentivity, and commonness when the concept is not appropriate for the NP.

3.3 Statistical Analysis

We investigated the reading time (`logtime`) of NPs that were annotated with the information structure labels. During the preprocessing, we excluded data {`authorsData`, `caption`, `listItem`, `profile`, `titleBlock`} of `metadata`. We also excluded zero-millisecond data points from the eye tracking data. The number of data points were 6444 for SELF; 5268 for FFT, FPT, RPT, and TOTAL; and 2081 for SPT. After model-based trimming was used to eliminate points beyond three standard deviations, the model was rebuilt (Baayen, 2008). `subj` and `article` were considered as random effects, as expressed in the following formula:

```
logtime ~ space * sessionN + length + dependent
+ is_first + is_last + is_second_last  + articleN
+ screenN + lineN + segmentN + infostatus
+ definiteness + specificity + animacy
+ sentience + agentivity + commonness
+ (1 | subj) + (1 | article)
```

4 Results

Tables 6, 7, 8, and 9 present the results for the reading time types of SELF, FPT, RPT, and TOTAL, respectively. The fixed effects other than the information structure labels are omitted in the FPT, RPT, and TOTAL.

When the absolute *t*-value of an effect was larger than 1.96, we regarded the factor as statistically significant and put the sign of the estimate. Otherwise, we placed a value of 0 to indicate nonsignificant factors. The results were very similar to those of the previous report (Asahara et al., 2016). [5] However, *anti-locality* phenomena (Konieczny, 2000) in which a head is read faster if it is preceded by more dependents were not observed for the eye-tracking method with the NPs.

5 Discussion

Table 10 presents the result for the fixed effects of the labels related to the information structure.

The information status `discourse-new`, `old` did not affect the reading time. Whereas the definiteness affected only TOTAL of the eye-tracking data, the specificity affected SELF, FPT, RPT, and TOTAL. Thus, the specificity has a stronger effect than the definiteness for the reading time of Japanese. The animacy (`animate/inanimate`) affected only FPT. However, the sentience (`sentient/insentient`) affected FPT and RPT. These two did not affect the self-paced reading. The agentivitiy showed no significant effect. Finally, the commonness, which is a feature for the hearer, affected SELF and TOTAL. The most important result is that the

[5] Whereas Asahara et al.’s paper was based on `time`, ours was based on `logtime` to reduce the outliers in the model.

Table 10: Summary: reading time and information structures

Fixed Effect		SELF	FFT	FPT	SPT	RPT	TOTAL
infostatus=discourse-old	(vs. discourse-new)	0	0	0	0	0	0
definiteness=indefinite	(vs. definite)	0	0	0	0	0	+
specificity=specific	(vs. either)	+	0	+	0	+	+
specificity=unspecific	(vs. either)	0	0	0	0	0	0
animacy=inanimate	(vs. animate)	0	0	+	0	0	0
sentience=insentient	(vs. either)	0	0	0	0	0	0
sentience=sentient	(vs. either)	0	0	+	0	+	0
agentivity=both	(vs. agent)	0	0	0	0	0	0
agentivity=neither	(vs. agent)	0	0	0	0	0	0
agentivity=patient	(vs. agent)	0	0	0	0	0	0
commonness=hearer-new	(vs. bridging)	+	0	0	0	0	+
commonness=hearer-old	(vs. bridging)	-	0	0	0	0	0
commonness=neither	(vs. bridging)	0	0	0	0	0	0

reading times of SELF and TOTAL may split for hearer-new and bridging.

In sum, the reading time showed that different patterns of reading time to determine the specificity, sentience and commonness in the information structure.

6 Conclusions

This paper presents a contrastive analysis between the reading time and information structure. The results showed the different patterns of promotion or interference of the reading time for the information structure of the NPs. They may lead to the possibility of classifying information structures by the reading time pattern.

The previous co-reference resolution methods addressed the issue of the information status based on whether or not an NP is mentioned in the previous discourse. However, these methods cannot identify the information structure for a recipient, such as the commonness (hearer-new, -old or bridging). It is necessary to introduce a recipient reaction feature to identify the commonness. The results in this paper showed that the reading time is a potential feature that can be used to solve bridging.

Our future work will involve building a classifier to split bridging and hearer-new based on the reading time for each subject participant. The classifier will be able to detect bridging that cannot be resolved by the subject participant's knowledge. This will enable us to develop user-oriented information extraction or document summarization models that incorporate present hearer-new Information.[6]

Acknowledgement

The work reported in this article was supported by the NINJAL research project of the Center for Corpus Development. This work was also supported by JSPS KAKENHI Grant Number JP25284083 and JP17H00917.

[6]The subject participants of BCCWJ-EyeTrack wrote a summary of the presented text (BCCWJ-SUMM).

References

Shigeaki Amano and Tadahisa 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.

Masayuki Asahara and Yuji Matsumoto. 2016. BCCWJ-DepPara: A Syntactic Annotation Treebank on the 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of the 12th Workshop on Asian Langauge Resources (ALR12)*, pages 49–58.

Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto. 2016. Reading-Time Annotations for 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 684–694.

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

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584.

Vera Demberg and Frank 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.

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

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

Michael Götze, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, and Ruben Stoel. 2007. Information structure. In Stefanie Dipper, Michael Götze, and Stavros Skopeteas, editors, *Information structure in cross-linguistic corpora: annotation guidelines for phonology, morphology, syntax, semantics and information structure*, volume 7, pages 147–187. Universitätsverlag Potsdam.

John 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.

Irene Heim. 2011. Definiteness and indefiniteness. In Klaus von Heusinger, Claudia Maienborn, and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, volume 2, pages 996–1025. Mouton de Gruyter.

Yufang Hou, Katja Markert, and Michael Strube. 2013. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 907–917, Atlanta, Georgia, June. Association for Computational Linguistics.

Ryu Iida, Koh Mitsuda, and Takenobu Tokunaga. 2013. Investigation of annotator's behaviour using eye-tracking data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 214–222.

Marcel A. Just, Patricia A. Carpenter, and Jacqueline D. Woolley. 1982. Paradigms and Processes in Reading Comprehension. *Journal of Experimental Psychology: General*, 3:228–238.

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

Sigrid Klerke, Héctor Martínez Alonso, and Anders Søgaard. 2015. Looking hard: Eye tracking for detecting grammaticality of automatically compressed sentences. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 97–105.

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

Christopher Lyons. 1999. *Definiteness*. Cambridge University Press, Cambridge.

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

Eleni Miltsakaki, Rashmi Prasad, Aravind K Joshi, and Bonnie L Webber. 2004. The penn discourse treebank. In *LREC*.

Takuya Miyauchi, Masayuki Asahara, Natsuko Nakagawa, and Sachi Kato. 2017. Annotation of Information Structure on "The Balanced Corpus of Contemporary Written Japanese. In *Proceedings of 2017 Conference of the Pacific Association for Computational Linguistics*.

Ryo Nagata, Tatsuya Iguchi, Fumito Masui, Atsuo Kawai, and Isu Naoki. 2005. A statistical model based on the three head words for detecting article errors. *IEICE TRANSACTIONS on Information ans Systems*, E88-D(7):1700–1706.

Mariko. Osaka and Naoyuki 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.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Conputational Linguistics*, 31(1):71–105.

Rashmi Prasad, Bonnie Webber, Alan Lee, Sameer Pradhan, and Aravind Joshi. 2015. Bridging sentential and discourse-level semantics through clausal adjuncts. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 64–69.

Douglas Roland, Gail Mauner, Carolyn O'Meara, and Hongoak Yun. 2012. Discourse expectations and relative clause processing. *Journal of Memory and Language*, 66(3):479–508.

Klaus von Heusinger. 2011. Specificity. In Klaus von Heusinger, Claudia Maienborn, and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, volume 2, pages 1058–1087. Mouton de Gruyter.