Analysis of Eye Movements based on the Entropy of the N-gram Model for the Investigation of Japanese Reading Processes

Akemi Tera† Kiyoaki Shirai‡ Takaya Yuizono† Kozo Sugiyama†
†School of Knowledge Science
Japan Advanced Institute of Science and Technology
‡School of Information Science
Japan Advanced Institute of Science and Technology
{ tera, kshirai, yuizono, sugi }@jaist.ac.jp

Abstract

In order to investigate reading processes of Japanese language learners, we have conducted an experiment to record eye movements during Japanese text reading using an eye-tracking system. We showed that Japanese native speakers use "forward and backward jumping of eye movements" frequently [12] [13]. In this paper, we analyzed further the same eye tracking data. Our goal is to examine whether Japanese learners fix their eye movements at boundaries of linguistic units such as words, phrases or clauses when they start or end "backward jumping". Linguistic boundaries are empirically defined based on the entropy of the N-gram model. Another goal is to examine the relation between the entropy of the N-gram model and the depth of syntactic structures of sentences. Our analysis shows that (1) Japanese learners often fix their eyes around linguistic boundaries, (2) the average of the entropy is the greatest at the fifth depth of syntactic structures.

Keywords: eye tracking, saccade, fixation, N-gram model, entropy

1 Introduction

In the field of Japanese education, many Japanese language learners (Japanese learners) feel that a Japanese text is difficult to read. The reasons are generally summarized as follows: there is no space between words, a Japanese text consists of four different kinds of characters (*Hiragana*, *Katakana*, *Kanji and Roma-ji*), and it is difficult to look up words represented by Kanji in a dictionary since Kanji words have several possible readings. In order to support the learning of Japanese, an investigation of Japanese learners' reading processes plays an important role.

Several studies aiming at analyzing reading processes of Japanese learners have been conducted. Tera et al. [10] reported that "Na-adjectives", "sahen" and complex verbs are often consulted by Japanese learners via a reading support system developed in [5] [6] [7][1]. They indicated that Japanese learners especially want to know information about predicates in the Japanese text. From the beginning of 1990, research on reading processes of Japanese or English texts using an eye-tracking system were initiated. Osaka [2] reported differences in the locations where subjects fix their eyes according to the types of characters in the text. Shigematsu et al. [4] discussed differences of fixation of eyes between Japanese and Chinese students. We also tried to investigate reading processes using an eye-tracking system [12] [13]. We found that (1) when reading text in native or familiar languages, fixation time tended to be short, (2) Japanese native speakers often showed "backward jumping of eye movements" and "forward jumping of eye movements". It is well known that "backward jumping of eye movements" occurs when a reader finds out some discrepancies while reading with respect to the meanings he holds, and is often used by skilled language learners [1].

In this paper, we further analyze the eye tracking data obtained in [12][13]. We consider the entropy of the N-gram model, and examine the correlation between the entropy and the eye movements.

The paper is organized as follows: in Section 2, the entropy, fixation and saccade are briefly introduced, since they are important concepts in

[1] "Na-adjective" and "sahen" are word classes in Japanese. The Na-adjective is one kind of adjective, while sahen is a word which functions both as a noun and as a verb.

this paper, and our goal is explained in more detail. Section 3 describes the detailed procedures of our analysis. Results of our analysis are reported in Section 4. A preliminary survey of the relation between syntactic structures and entropy is presented in Section 5. We conclude the paper in Section 6.

2 Goal

We here briefly introduce the entropy of the N-gram model. The N-gram model is a well-known probabilistic language model widely used in the research field of natural language processing [9]. The N-gram model of characters is the probabilistic model predicting the appearance of a character (c_i) given its preceding N-1 characters ($c_{i-N+1}..c_{i-1}$). It is defined as follows:

$$P(c_i \mid c_{i-N+1} \cdots c_{i-1}) \quad (1)$$

The N-gram model can be automatically trained from a large amount of corpora [9]. Next, the entropy E of the N-gram model is given as below:

$$E = -\sum_{c_i} P(c_i \mid c_{i-N+1} \cdots c_{i-1}) \log P(c_i \mid c_{i-N+1} \cdots c_{i-1}) \quad (2)$$

In general, the entropy of the N-gram model is relevant to linguistic boundaries such as boundaries of words, phrases or clauses. For example, let us consider the cases where characters $c_{i-N+1}...c_i$ are in a word. In such cases, only a limited number of possible characters would appear after the string c_{i-N+1} ... c_{i-1}. Thus the entropy would be low since the probabilistic distribution $P(c_i \mid c_{i-N+1} \cdots c_{i-1})$ is not uniform, or is skewed. On the other hand, if there is a linguistic boundary between c_{i-1} and c_i, various characters could appear after the string $c_{i-N+1}...c_{i-1}$. In such cases, the entropy would be high since $P(c_i \mid c_{i-N+1} \cdots c_{i-1})$ tends to be uniform. To summarize, if the entropy of the N-gram model is high, we can assume that there is a linguistic boundary at a position between characters c_{i-1} and c_i.

Next, we introduce "fixation" and "saccade" in eye movements. When we read a text, we use our eyes to get information from the text. It is already known that humans do not move their eyes smoothly or constantly while reading, but repeat fixation and saccade. "Fixation" means that the eyeballs stop moving for the eye to glance at the same point for a while in order to carefully read a text, while "saccade" means that the eyeballs move from a fixation point to the next fixation point. We call a saccade in the forward direction a "forward saccade", while a saccade in the backward direction is a "backward saccade".

The goal of this paper is to investigate the reading process of Japanese learners. More specifically, we investigate where backward saccades occur in the text. As described in Section 1, a backward saccade (backward jumping of eye movement) occurs when a Japanese learner cannot understand the text smoothly and wants to read previous sentences again to make sure that what he/she understood before is correct. If this is correct, we suppose that Japanese learners would start or end a backward saccade not within a word but at linguistic boundaries. So we propose the following hypothesis:

[Hypothesis]

When reading a Japanese text, a backward saccade starts or ends around linguistic boundaries.

Furthermore, as we mentioned before, linguistic boundaries can be empirically defined by measuring the entropy of the N-gram model. In this paper, we will empirically prove the above hypothesis through an analysis of real eye-tracking data as well as the entropy of the N-gram model trained from a large amount of text.

Another goal is to examine the relation between the backward saccade and syntactic structures of sentences. We would like to know if Japanese learners start or end backward saccades at deep positions in syntactic structures. As a preliminary survey for this investigation, we examine the correlation between the entropy of the N-gram model and the depth of syntactic structures.

3 Methodology

In this section, procedures to verify our hypothesis are described. The experiment for recording eye-tracking data is briefly described in Subsection 3.1 [12][13], while the verification of our hypothesis by use of the obtained eye-tracking data is discussed in Subsection 3.2.

3.1 Collecting Eye Tracking Data

The following equipment is used to obtain eye-tracking data of Japanese learners:

1) An eye-mark recorder EMR-8 (NAC)
2) A 44-degree lens
3) An analyzing system (NAC)

4) A windows NT, 17 inches monitor
5) A head stand

Since subjects are required to have eyesight over 1.0 to record eye movements with our eye-mark recorder, an eyesight check is done before the experiment. An illustration of the recording of the eye tracking data with the above equipment is shown in Figure 1.

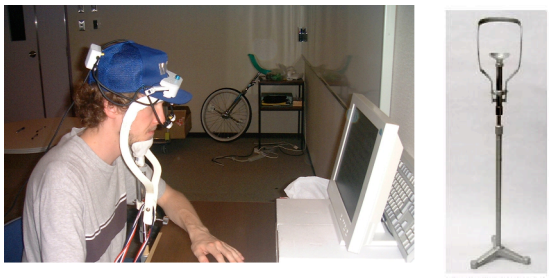

Figure 1. Experimental set-up

The procedures for obtaining eye-tracking data consist of the following 7 steps:

1) Choose intermediate learners as subjects according to the results of the Standard Japanese Ability Test and the English TOEFL / TOEIC.
2) The explain on overview of the experiments and instructions to the subjects, then let them do preliminary experiments.
3) Ask the subjects to read texts displayed randomly on a PC screen one by one. Time for reading is not limited.
4) Record eye tracking data during reading.
5) Carry out an examination to check the subjects' comprehension of texts, and obtain a score for each subject.
6) Ask the subjects which texts are most difficult and easiest to understand, and the reason why.
7) Analyze the obtained eye tracking data as well as other data (Questionnaires, etc.).

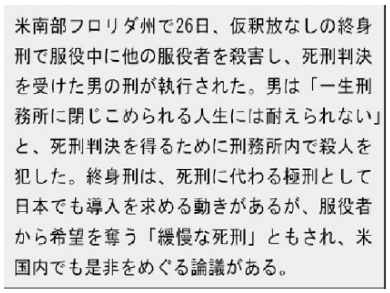
米南部フロリダ州で26日、仮釈放なしの終身刑で服役中に他の服役者を殺害し、死刑判決を受けた男の刑が執行された。男は「一生刑務所に閉じこめられる人生には耐えられない」と、死刑判決を得るために刑務所内で殺人を犯した。終身刑は、死刑に代わる極刑として日本でも導入を求める動きがあるが、服役者から希望を奪う「緩慢な死刑」ともされ、米国内でも是非をめぐる論議がある。

Figure 2. Japanese Text Used for the experiment

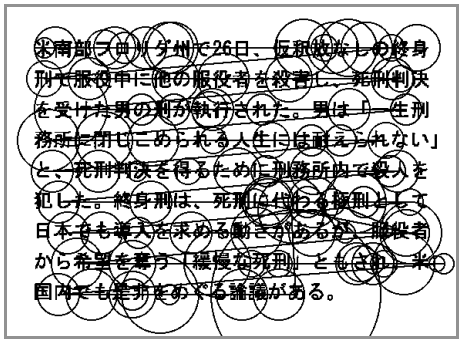

Figure 3. Saccades and Fixations on the Text

The Japanese text used in the experiment is a newspaper article excerpted from the Chunichi Newspaper [11], shown in Figure 2. It contains 178 characters. Figure 3 shows a schematic representation of the obtained eye tracking data overlapped with the text. In Figure 3, circles indicate fixations, while lines indicate saccades. The size of the circles represents the time of the fixations.

We collected eye-tracking data for 20 subjects. They are classified into the following 5 groups according to their native language and familiarity with Kanji [8]:

- **Japanese-Native-Speaker(J)**
  Subjects who are Japanese native speakers.
- **Kanji-area(K)**
  Subjects who use Kanji in their native language.
- **Middle-area(M)**
  Subjects who know Kanji but do not usually use them.
- **Non-Kanji-area-Asia(NA)**
  Subjects who do not use Kanji and are from Asia.
- **Non-Kanji-area-Europe(NE)**
  Subjects who do not use Kanji and are from Europe.

The Numbers of subjects in each group as well as their nationalities are summarized in Table 1.

Table 1 Nationalities of Subjects

J	Japanese (4)
K	China (5)
M	Korea (4)
NA	Nepal (1), Vietnam (1), Thailand (1)
NE	Belgium (1), Germany (1), Hungary (1), Spain (1)

3.2 Analysis of Eye Tracking Data

First, we train the N-gram model of characters. We set N equal to 5, that is, we estimate a 5-gram model which predicts the probabilities of

appearance of characters given 4 preceding characters. The model is trained by maximum likelihood estimation from newspaper articles published over 13 years. Then, the entropy of the 5-gram model is calculated at all positions in the text that subjects read in our experiment. The 'Position in the text' refers to a position between characters, and the entropy at a position between c_{i-1} and c_i is the entropy of the probability distribution $P(c_i \mid c_{i-4}c_{i-3}c_{i-2}c_{i-1})$, where c_{i-4}, c_{i-3}, c_{i-2}, and c_{i-1} are the 4 characters appearing before that position. Unfortunately, the entropy cannot be calculated for all positions due to the data sparseness, since the entropy is not determined when the string $c_{i-4}c_{i-3}c_{i-2}c_{i-1}$ does not occur in the training corpus. Hereafter we call such positions "uncertain entropy positions".

Next, we extract fixations at the start and at the end of backward saccades from the eye tracking data. The positions of extracted fixations in the text area should be identified, since our current goal is to examine if fixations happen at linguistic boundaries or at high entropy positions. However, it is rather difficult to decide the exact positions of fixations. Osaka [2] reported that readers saw between 2 and 5, mostly 3 and 4 characters when they fixed their eyes. This means that Japanese learners may see not a point but an area including several characters at fixations. Therefore, we suppose that subjects see 3 or 4 characters when they fix their eyes, and identify characters they glance at fixations as shown in Figure 4.

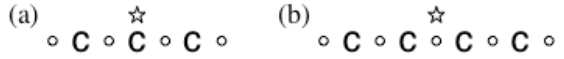

Figure 4 Fixation Area

In this figure, a star shows for the center of eye glance at fixation, while 'C' stands for characters supposed to be seen by subjects. That is, when the center of eye glance at fixation is on a character, we suppose that subjects see 3 characters, the character on eye glance and its previous and succeeding characters (Figure 4 (a)). Alternately, when the center of eye glance is between characters, we suppose that subjects see 4 characters around the center of eye glance (Figure 4 (b)). Hereafter we call the area including characters glanced by subjects "fixation area".

After identifying the fixation area for each fixation, we manually check if linguistic boundaries exist in the fixation area. More concretely, we look for linguistic boundaries among positions indicated by circles in Figure 4. Linguistic boundaries are defined by the entropy of the 5-gram model: if the entropy at a position is greater than a certain threshold T, we regard that position as a linguistic boundary. Then the linguistic boundary ratio LBR, defined by the formula below, is calculated.

$$LBR = \frac{\text{No. of fixations such that at least one linguistic boundary exists in the fixation area}}{\text{Total No. of fixations}}$$

LBR evaluates how likely fixations happen around linguistic boundaries. If LBR is high enough, our hypothesis proposed in Section 2 is verified.

4 Experiment

4.1 Results of our Analysis

We found 367 backward saccades from the eye tracking data of 20 subjects. The LBR is calculated for fixations at the start and at the end of these backward saccades. Results are summarized in Table 2. It shows LBR for several cases classified according to the following aspects:

- **Threshold of entropy (T)**

Since linguistic boundaries are defined according to entropy, the number of linguistic boundaries in the text can be controlled by the threshold T. In this experiment, we set the threshold T to 2.5 and 3. Among 177 positions in the text[2], 53 (29.9%) and 36 (23.3%) positions are regarded as linguistic boundaries when T=2.5 and T=3.0, respectively.

- **Start or End**

Fixations are distinguished if they are at the start or at the end of backward saccades. The LBR for start fixations is shown in the left tables, and the LBR for end fixations in the right tables.

- **Groups of subjects**

We separately calculate the LBR for fixations of subjects in the 5 different groups described in Subsection 3.1. In Table 2, the average and standard deviation (SD) of LBR for each group is shown in each row. The same data for all subjects is also shown in the 'All-ave' row.

- **Uncertain entropy position**

As described in Subsection 3.2, the entropy at some positions cannot be calculated due to data

[2] Note that the text used in this experiment contains 178 characters.

Table 2 Results of Analysis

< T = 2.5 >

< Start >	Certain	(SD)	All	(SD)
J-ave	0.883	(0.082)	0.894	(0.102)
K-ave	0.829	(0.115)	0.810	(0.124)
M-ave	0.778	(0.114)	0.806	(0.081)
NA-ave	0.783	(0.132)	0.868	(0.081)
NE-ave	0.855	(0.129)	0.906	(0.118)
All-ave	0.828	(0.112)	0.854	(0.105)
E	< 0.815 >		< 0.786 >	

< T = 2.5 >

< End>	Certain	(SD)	All	(SD)
J-ave	0.767	(0.097)	0.787	(0.164)
K-ave	0.855	(0.144)	0.849	(0.182)
M-ave	0.867	(0.103)	0.875	(0.106)
NA-ave	0.754	(0.126)	0.789	(0.139)
NE-ave	0.829	(0.089)	0.867	(0.090)
All-ave	0.817	(0.119)	0.835	(0.140)
E	< 0.815 >		< 0.786 >	

< T = 3.0 >

< Start >	Certain	(SD)	All	(SD)
J-ave	0.617	(0.170)	0.638	(0.178)
K-ave	0.675	(0.237)	0.710	(0.213)
M-ave	0.644	(0.253)	0.677	(0.249)
NA-ave	0.623	(0.219)	0.698	(0.186)
NE-ave	0.658	(0.087)	0.719	(0.074)
All-ave	0.649	(0.189)	0.695	(0.186)
E	< 0.669 >		< 0.611 >	

< T = 3.0 >

< End>	Certain	(SD)	All	(SD)
J-ave	0.617	(0.111)	0.617	(0.202)
K-ave	0.573	(0.149)	0.575	(0.216)
M-ave	0.578	(0.160)	0.600	(0.152)
NA-ave	0.464	(0.231)	0.456	(0.209)
NE-ave	0.658	(0.055)	0.750	(0.053)
All-ave	0.578	(0.146)	0.597	(0.186)
E	< 0.669 >		< 0.611 >	

sparseness. If such uncertain entropy positions are located in the fixation area, the judgment of whether a linguistic boundary exists in the fixation area is also uncertain. So we calculated the *LBR* when the entropy at all positions in the fixation area can be calculated ('Certain' column in Table 2), and the *LBR* for all fixations, including ones such that uncertain entropy positions exist in the fixation area ('All' column). In the latter case, *LBR* is just an approximation. Note that it is underestimated since uncertain entropy positions may be real linguistic boundaries.

The expectation of the *LBR* is also shown in the 'E' row in Table 2. It is defined as the proportion of points such that linguistic boundaries exist in the neighborhood of all points in the text. Here, points in the text mean both centers of characters as indicated by the star in Figure 4 (a), and positions between characters as indicated by the star in Figure 4 (b). That is, the expectation of the *LBR* represents a probability such that when a point in the text represents randomly chosen as the center of fixation, one or more linguistic boundaries exist around that point.

4.2 Discussion

When *T* is set to 2.5, the averages of *LBR* for all subjects are higher than the expectation of the *LBR* for both start and end fixations of backward saccades, as shown in Table 2. Furthermore, the averages *LBRs* for most groups are also higher than the expectation. This is almost true for start fixations when *T*=3. These results indicate that our hypothesis, namely Japanese learners tend to start or end their backward saccades at linguistic boundaries, is valid to some degree. On the other hand, since the averages of the *LBR* are less than the expectation, our hypothesis is not valid for end fixations when *T*=3.

Next, we will discuss differences among 5 groups of Japanese learners. Before the experiment, we expected that Japanese native speakers might start or end their backward saccades around linguistic boundaries more often than foreign students, since foreign students did not have a good knowledge of Japanese. However, such a tendency is not observed. When T=2.5, the *LBR* of group J (Japanese-Native-Speaker) for start fixations is higher than that of other groups, but this is not true for end fixations or when *T* is set equal to 3. Surprisingly, the *LBR* of group NE (Non-Kanji-area-Europe) is relatively high for all cases. Although Europeans might not be familiar with Kanji characters, they often start or end their backward saccades around linguistic boundaries. The above observations may suggest that the reading process of Japanese learners is not strongly related to their familiarity with Japanese. However, the size of the experiment is not sufficient in terms of both the number of texts and the number of subjects. It is necessary to conduct a larger experiment and to analyze the results further in order to reveal the reading process of Japanese learners.

5 Syntactic Structure and Entropy

Another question about backward saccades is whether Japanese learners tend to start or end their backward eye movements at deep positions in syntactic structures of sentences. As a preliminary survey, we investigate the relationship between depth of syntactic structures and the entropy of the N-gram model. First, we obtained syntactic structures of sentences in the text used in our experiment with the Japanese CALL system *Asunaro*[3]. For each character in sentences, we obtained the entropy of the 5-gram model at the position after the character and its depth in a syntactic tree, where depth is the distance from the root to the character. We obtained the statistics shown in Table 3.

Table 3 Depth in Structures and Entropy

Depth	Average of entropy	No. of Characters
3	1.9314	15
4	1.3876	5
5	2.4844	20
6	1.9822	19
7	1.3822	49
8	1.8715	30
9	1.6572	29
10	1.0735	11
All	1.8599	178

The correlation between the depth and the average of the entropy is - 0.0941. This means that there is no relationship between the two variables.

6 Conclusion

In this paper, we analyzed eye tracking data to investigate the reading process of Japanese learners. First, we proposed the hypothesis that learners tend to start and end their backward saccades around linguistic boundaries. The hypothesis is empirically verified by analyzing eye tracking data when learners read a Japanese text and by identifying linguistic boundaries according to the entropy of the N-gram model. The results of our analysis indicate that our hypothesis is valid to some degree.

In future research, large-scale experiments are required. We have already obtained eye-tracking data for 3 other texts with the same subjects. The analysis of these 3 texts will be performed shortly. A detailed comparison between nationalities and languages of Japanese learners will also be made. Furthermore, we plan to empirically investigate the relation between backward saccades and the depth in syntactic structures.

[3] http://hinoki.ryu.titech.ac.jp/

References

[1] Ford, Bresnan, and Kaplan. A competence-based theory of syntactic closure. In Joan Bresnan, editor, The Mental Representation of Grammatical Relations, MIT Press, 1982.

[2] Naoaki Osaka. Size of saccade and fixation duration of eye movements during reading. Psychophysics of Japanese text processing, Journal of the Optical Society of *America*, A (9): 5-13, 1992.

[3] Ryoji Osaka, Yukio Nakazawa and Kazuo Koga. Experimental psychology of eye movement. Nagoya University press, Tokyo, Japan, 1993.

[4] Jun Shigematsu and Tsutomu Konosu. The research on the Reading Processes by Non-Native Speakers using on Eye-Camera. *Spring Proceedings of the KNG*: 31-42, 1993. (in Japanese)

[5] Akemi Tera, Tatsuya Kitamura and Koichiro Ochimizu. Japanese Reading Support System "dict-linker". *Autumn Proceedings of the KNG*: 43-48, 1996. (in Japanese)

[6] Akemi Tera. Extensive Reading Support System for Learning Kanji. Meiji Shoin, 16(6): 101-108, 1997. (in Japanese)

[7] Akemi Tera, Tatsuya Kitamura, Koichiro Ochimizu, Tomoko Graham and Ann Lavin. Japanese Reading Support System (DL) Evaluation - Results from MIT User Survey. JLEM, 4(1): 26-27, 1997. (in Japanese)

[8] Akemi Tera, Tatsuya Kitamura and Manabu Okumura. The Verification of Japanese Reading Support System "DL" -The Research of The Process of Japanese learner's reading. JLEM, 6(1): 20-21, 1999. (in Japanese)

[9] Christopher D. Manning, Heinrich Schutze. Foundations of Statistical Natural Language Processing. The MIT Press, 1999.

[10] Akemi Tera, Hajime Motizuki and Akira Shimazu. An analysis of the words that Japanese Learners need Information during Reading. *Autumn Proceedings* of the 27th JSISE Conference: 259-260, 2002. (in Japanese)

[11] Chunichi Newspaper. Chunichi Newspaper Company, 2004.

[12] Akemi Tera and Kozo Sugiyama. Eye-Tracking Analyses of Japanese Reading Processes (IV) - Forward and Backward Jumping of Eye Movements-. *Proceedings of the IEICE*, TL2005-35-48: 43-48, 2006. (in Japanese)

[13] Akemi Tera and Kozo Sugiyama. Eye-Tracking Analyses of Japanese Reading Processes. *Proceedings* of the KICSS*: 108-114,* 2006.