Modeling Comma Placement in Chinese Text for Better Readability using Linguistic Features and Gaze Information

Tadayoshi Hara\textsuperscript{1}  Chen Chen\textsuperscript{2*}  Yoshinobu Kano\textsuperscript{3,1}  Akiko Aizawa\textsuperscript{1}
\textsuperscript{1}National Institute of Informatics, Japan  \textsuperscript{2}The University of Tokyo, Japan  \textsuperscript{3}PRESTO, Japan Science and Technology Agency
{harasan, kano, aizawa}@nii.ac.jp

Abstract

Comma placements in Chinese text are relatively arbitrary although there are some syntactic guidelines for them. In this research, we attempt to improve the readability of text by optimizing comma placements through integration of linguistic features of text and gaze features of readers.

We design a comma predictor for general Chinese text based on conditional random field models with linguistic features. After that, we build a rule-based filter for categorizing commas in text according to their contribution to readability based on the analysis of gazes of people reading text with and without commas.

The experimental results show that our predictor reproduces the comma distribution in the Penn Chinese Treebank with 78.41 in F\textsubscript{1}-score and commas chosen by our filter smoothen certain gaze behaviors.

1 Introduction

Chinese is an ideographic language, with no natural apparent word boundaries, little morphology, and no case markers. Moreover, most Chinese sentences are quite long. These features make it especially difficult for Chinese learners to identify composition of a word or a clause in a sentence.

Punctuation marks, especially commas, are allowed to be placed relatively arbitrarily to serve as important segmentation cues (Yue, 2006) for providing syntactic and prosodic boundaries in text; commas indicate not only phrase or clause boundaries but also sentence segmentations, and they capture some of the major aspects of a writer’s prosodic intent (Chafe, 1988). The combination of both aspects promotes cognition when reading text (Ren and Yang, 2010; Walker et al., 2001).

However, although there are guidelines and research on the syntactic aspects of comma placement, prosodic aspects have not been explored, since they are more related with cognition. It is as yet unclear how comma placement should be optimized for reading, and it has thus far been up to the writer (Huang and Chen, 2011).

In this research, we attempt to optimize comma placements by integrating the linguistic features of text and the gaze features of readers. Figure 1 illustrates our approach. First, we design a comma predictor for general Chinese text based on conditional random field (CRF) models with various linguistic features. Second, we build a rule-based filter for classifying commas in text into ones facilitating or obstructing readability, by comparing the gaze features of persons reading text with and without commas. These two steps are connected by applying our rule-based filter to commas predicted by our comma predictor. The experimental results for each step validate our approach.

Related work is described in Section 2. The functions of Chinese commas are described in Section 3. Our CRF model-based comma predictor is examined in Section 4, and our rule-based comma filter is constructed and examined in Section 5 and 6. Section 7 contains a summary and outlines future directions of this research.
There are five main uses of commas in Chinese text, as shown in Table 1. Cases 1 to 4 are from ZDIC.NET (2005), and Case 5 obviously exists in Chinese text. The first three serve the function of emphasis, while the latter two indicate coordinating or subordinating clauses or phrases.

In Cases 1 and 2, a comma is inserted as a kind of pause between a short subject and a long predicate, or between a short remainder predicate, such as see/know, indicate, etc.
that Chinese sentence segmentation can be viewed as detecting loosely coordinated clauses separated by commas.

4 CRF Model-based Comma Predictor

We first predict comma placements in existing text. The prediction is formalized as a task to annotate each word in a word sequence with an IOB-style tag such as I-Comma (following a comma), B-Comma (preceding a comma) or O (neither I-Comma nor B-Comma). We utilize a CRF model for this sequential labeling (Lafferty et al., 2001).

4.1 CRF Model for Comma Prediction

A conditional probability assigned to a label sequence \( Y \) for a particular sequence of words \( X \) in a first-order linear-chain CRF is given by:

\[
P_{\lambda}(Y|X) = \frac{\exp(\sum_{w} \sum_{i} \lambda_i f_i(Y_{w-1}, Y_w, X, w))}{Z_0(X)}
\]

where \( w \) is a word position in \( X \), \( f_i \) is a binary function describing a feature for \( Y_{w-1}, Y_w, X, w \), \( \lambda_i \) is a weight for that feature, and \( Z_0 \) is a normalization factor over all possible label sequences.

The weight \( \lambda_i \) for each \( f_i \) is learned on training data. For \( f_i \), the linguistic features shown in Table 2 are derived from a syntactic parse of a sentence\(^4\). The first three were used initially; the rest were added after we got feedback from construction of our rule-based filters (see Section 5). Figure 2 shows an example of a parsing tree and its corresponding training data.

\(^4\)Some other features or tag formats which worked well in the previous research, such as bi-tri-gram, a preceding word (L-1) or its POS (POS-1), and IO-style tag (Leaman and Gonzalez, 2008) were also examined, but they did not work that well, probably because of the difference in task settings.

Table 3: Performance of the comma predictor

| Feature | \( f_i \) | \( P \) | \( R \) | \( A \) |
|---------|---------|------|------|------|
| WS      | 59.32 (72.67/50.12) | 95.45 |
| POS     | 32.51 (69.06/21.26) | 94.08 |
| DIP     | 34.14 (68.65/22.72) | 94.13 |
| STAG    | 22.44 (64.00/13.60) | 93.67 |
| OIC     | 9.27 (66.56/4.98) | 93.42 |
| WL      | 10.70 (75.24/5.76) | 93.52 |
| LOD     | 35.32 (59.20/25.17) | 93.81 |
| WL+LOD  | 75.43 (84.76/67.95) | 97.11 |
| WS+POS+DIP+STAG+OIC+WL+LOD | 77.58 (83.66/72.27) | 97.01 |

Table 4: Materials assigned to each subject

| Article ID | Characters | (A) #Punctuations | (B) #Commas | (C) #Commas | Subjects |
|------------|------------|--------------------|-------------|-------------|----------|
| 6          | 692        | 49                 | 28          | 4.04%       | L, T, C  |
| 7          | 335        | 30                 | 15          | 4.48%       | L, T, C  |
| 10         | 346        | 18                 | 7           | 2.02%       | L, T, C  |
| 12         | 221        | 18                 | 7           | 3.17%       | L, T, C  |
| 14         | 572        | 33                 | 14          | 2.45%       | L, T, C  |
| 18         | 471        | 36                 | 13          | 2.76%       | C, Z     |
| 79         | 655        | 53                 | 28          | 4.27%       | Z         |
| 82         | 471        | 30                 | 13          | 2.76%       | Z         |
| 121        | 629        | 41                 | 19          | 3.02%       | Z         |
| 294        | 608        | 50                 | 24          | 3.95%       | Z         |
| 401        | 567        | 43                 | 21          | 3.70%       | L, T, C  |
| 406        | 558        | 39                 | 18          | 3.23%       | Z         |
| 413        | 552        | 52                 | 22          | 3.99%       | T, C, Z  |
| 423        | 580        | 49                 | 26          | 4.48%       | L, C, Z  |
| 438        | 674        | 46                 | 28          | 4.15%       | L, T, C  |

Average | 528.73 | 39.13 | 18.87 | 3.57% | 48.22% |

Figure 4: Obtained eye-movement trace map

Figure 5: Total viewing time
4.2 Experimental Settings

The Penn Chinese Treebank (CTB) 7.0 (Naiwen Xue and Palmer, 2005) consists of 2,448 articles in five genres. It contains 1,196,329 words, and all sentences are annotated with parse trees. We selected four genres for written Chinese (newswire, news magazine, broadcast news and newsgroups/weblogs) from this corpus as our dataset. These were randomly divided into training (90%) and test data (10%). We also corrected errors in tagging and inconsistencies in the dataset, mainly by solving problems around strange characters tagged as PU (punctuation). The commas and characters after this preprocessing numbered 63,571 and 1,533,928 in the training data and 4,116 and 111,172 in the test data.

MALLET (McCallum, 2002) and its application ABNER (Settles, 2005) were used to train the CRF model. We evaluated the results in terms of precision (P = tp/(tp + fp)), recall (R = tp/(tp + fn)), F₁-score (F₁ = 2PR/(P + R)), and accuracy (A = (tp + tn)/(tp + tn + fp + fn)), where tp, tn, fp and fn are respectively the number of true positives, true negatives, false positives and false negatives, based on whether the model and the corpus provided commas at each location.

4.3 Performance of the CRF Model

Table 3 shows the performance of our CRF model². We can see that WS contributed much more to the performance than other features, probably because a word surface itself has a lot of information on both prosodic and syntactic functions. Combining WS with other features greatly improved performance, and as a result, with all features (WS + POS + STAG + DIP + OIC + LOD + WL), precision, recall, F₁-score and accuracy were 83.97%, 73.54%, 78.41 and 97.36%.

We also found that a large number of false positives seemed helpful according to native speakers (see the description of the subjects in Section 5 and 6). Although these commas do not appear in the CTB text, they might smoothen the reading experience. We constructed a rule-based filter in order to pick out such commas.

5 Rule-based Comma Filter

We constructed a rule-based comma filter for classifying commas in text into ones facilitating (positive) or obstructing (negative) the reading process as follows:

[Step 1]: Collect gaze data from persons reading text with or without commas (Section 5.1).
[Step 2]: Compare gaze features around commas to find those features that reflect the effect of comma placement. (Section 5.2).
[Step 3]: Annotate commas with categories based on the obtained features (Section 5.3), and devise rules to explain the annotation (Section 5.4).

5.1 Collecting Human Eye-movement Data

Eye-movements during reading contain rich information on how the document is being read, what the reader is interested in, where difficulties happen, etc. The movements are characterized by fixations (short periods of steadiness), saccades (fast movements), and regressions (backward saccades) (Rayner, 1998). In order to analyze the effect of commas on reading the features, we collected gaze data from subjects reading text in the following settings.

[Subjects and Materials] Four native Man-
Table 5: Comma categories

| Subject | Positive (⃝) | Semi-positive (△) | Semi-negative (Ⅳ) | Negative (×) | Adjustment formula |
|---------|--------------|------------------|------------------|-------------|-------------------|
| L       | ∆FT > 800   | 500 < ∆FT ≤ 800 | -100 < ∆FT ≤ 500 | ∆FT < -100  | ∆FT = ∆FT + ∆RT ≥ 200 |
| C       | ∆FT > 900   | 600 < ∆FT ≤ 900 | -200 < ∆FT ≤ 600 | ∆FT < -200  | ∆FT = ∆FT + ∆RT ≥ 275 |
| T       | ∆FT > 600   | 300 < ∆FT ≤ 600 | -300 < ∆FT ≤ 300 | ∆FT < -300  | ∆FT = ∆FT + ∆RT ≥ 250 |
| Z       | ∆FT > 600   | 350 < ∆FT ≤ 650 | -250 < ∆FT ≤ 350 | ∆FT < -250  | ∆FT = ∆FT + ∆RT ≥ 250 |

\[ \Delta RT = \text{fixation time (without commas)} \text{ [ms]} - [ \text{fixation time (with commas)} \text{ [ms]} ] \]

Table 6: Estimation formula for judging the contribution of commas to readability

| ID | ⃝ | △ | △ | × | ⃝ | △ | △ | × |
|----|---|---|---|---|---|---|---|---|
| 6  | 13| 6 | 4 | 5 | 121| 11| 2 | 6 | 0 |
| 7  | 8 | 6 | 1 | 0 | 294| 9 | 9 | 4 | 1 |
| 10 | 5 | 1 | 1 | 1 | 401| 10| 7 | 2 | 2 |
| 12 | 1 | 4 | 2 | 0 | 406| 5 | 6 | 5 | 2 |
| 14 | 4 | 4 | 5 | 1 | 413| 8 | 5 | 6 | 3 |
| 18 | 5 | 1 | 4 | 3 | 423| 11| 4 | 7 | 4 |
| 79 | 11| 4 | 9 | 4 | 438| 6 | 16| 6 | 0 |
| 82 | 5 | 6 | 2 | 0 | Total| 112| 80| 64| 26 |

Table 7: Categories of annotated commas

darin Chinese speakers (graduate students and researchers) read 15 newswire articles selected from CTB 7.0 (included in the test data in Section 4.2). Table 4 and Figure 3(a) show the materials assigned to each subject and a screenshot of one material. Each article was presented in 12-15 points of bold-faced Fang-Song font occupying 13 × 13, 14 × 15, 15 × 16 or 16 × 16 pixels along with a line spacing of 5-10 pixels.

[Apparatus] Figure 3(b) shows a scene of the experiment. An EyeLink 1000 eye tracker (SR Research Ltd., Toronto, Canada) with a desktop mount monitored the movements of a right eye at 1,000 Hz. The subject’s head was supported at the chin and forehead. The distance between the eyes and the monitor was around 55 cm, and each Chinese character subtended a visual angle 1°. Text was presented on a 19” monitor at a resolution of 800 × 600 pixels, with the brightness adjusted to a comfortable level. The displayed article was masked except for the area around a gaze point (see Figure 3(c)) in order to confirm that the gaze point was correctly detected and make the subject concentrate on the area (adjusted for him/her).

[Procedure] Each article was presented twice (once with/once without commas) to each subject.

The one without commas was presented first (not necessarily in a row). We did not give any comprehension test after reading; we just asked the subjects to read carefully and silently at their normal or lower speed, in order to minimize the effect of the first reading on the second. The subjects were informed of the presence or absence of commas beforehand. The apparatus was calibrated before the experiment and between trials. The experiment lasted around two hours for each subject.

[Alignment of eye-tracking data to text] Figure 4 shows an example of the obtained eye-movement trace map, where circles and lines respectively mean fixation points and saccades, and color depth shows their duration. The alignment of the data to the text is a critical task, and although automatic approaches have been proposed (Martínez-Gómez et al., 2012a; Martínez-Gómez et al., 2012b), they do not seem robust enough for our purpose. Accordingly, we here just compared the entire layout of the gaze point distribution and that of the actual text, and adjusted them to have relatively coherent positions on the x-axis; i.e., the beginning and end of the gaze point sequence in a line were made as close as possible to those of the line in the text.

5.2 Analysis of Eye-movement Data

The gaze data were analyzed by focusing on regions around each comma or where each one should be (three characters left and right to the comma).
1. If $L_{\text{Seg}}$ and $R_{\text{Seg}}$ are both very long, a comma must be put between them.
2. If two $\triangle$ appear serially, one is necessary whereas the other might be optional or judged negative, but it still depends on the lengths of the siblings.
3. If two neighboring commas appear very close to each other, one of them is judged as negative whereas judgment on the other one is reserved.
4. If several (more than 2) $\times$ appear continually, one or more $\times$ might be reserved in consideration of the global condition.
5. A comma is always needed after a long sentence or clause without any syntactically significant punctuation with the function of segmentation.
6. If a $\triangle$ appears near a $\bigcirc$, it might be judged as negative with a high probability. However, the judgment process is always from the bottom up, which means $x \rightarrow \bigcirc \rightarrow \triangle \rightarrow \bigcirc$. For example, if a $\bigcirc$ appears near a $\triangle$, we judge $\bigcirc$ first (to be positive or negative), then judge the $\triangle$ in the condition with or without the comma of $\bigcirc$.

#### Table 8: General rules for reference

Figure 5, 6 and 7 respectively show the total viewing time, fixation time (duration for all fixations and saccades in a target region) per comma, and number of regressions per comma to each trial. We can see a general trend wherein the former two were shorter and the latter was smaller for the articles with commas than without. The diversity of the subjects was also observed in Figure 6.

Figure 8 and 9 show the saccade length per comma for different measures. The former (latter) figure considers a saccade in which at least one edge (both edges) was in the region. We cannot see any global trend, probably because of the difference in global layout of materials brought by the presence or absence of commas.

### 5.3 Categorization of Commas

Using the features shown to be effective to represent the effect of comma placement, we analyzed the statistics for each comma in order to manually construct an estimation formula for judging the contribution of each comma to readability. The contribution was classified into four categories (Table 5), and the formula is described in Table 6.

The adjustment formula was based on our observation that the number of regressions could only be regarded as an aid. For example, for subject C, if $\Delta FT=200\text{ms}$ and $\Delta RT=-2$, $\Delta FT'=-350$, and therefore, the comma is annotated as negative.

All parameters were decided empirically and manually checked twice (self-judgment and feedback from the subjects).

On the basis of this estimation formula, all articles in Table 4 were manually annotated. Table 7 shows the distribution of the assigned categories.

$^6$Calculated by counting the instances where the $x$-position of [a fixation / end point of a saccade] was ahead of [the former fixation / its start point]. Although the counts of these two types were almost the same, by counting both of them, we expected to cover any possible regression.

$^7$One or two features are used to judge the category of a comma. We will explore more features in the future.

$^8$In the case of severe contradictions, the annotators discussed them and resolved them by voting.

#### Table 8: General rules for reference

5.4 Implementation of Rule-based Filter

The annotated commas were classified into Cases 1 to 5 in Table 1, based on the types of left and right segment conjuncts ($L_{\text{Seg}}$ and $R_{\text{Seg}}$, which were obtained from the parse trees in CTB). For each of the five cases, the reason for the assignment of a category ($\bigcirc$, $\triangle$, $\square$ or $\times$) to each comma was explained by a manually constructed rule which utilized information about $L_{\text{Seg}}$ and $R_{\text{Seg}}$. The rules were constructed so that they would cover as many instances as possible. Table 8 shows the general rules utilized as a reference, and Table 9 shows the finally obtained rules. The rightmost column in this table shows the number of commas matching each rule. These rules were then implemented as a filter for classifying commas in a given text.

For several rules ($\bigcirc10$, $\square8$, $\square10$, $\square11$ and $\square12$), there were only single instances. In addition, although our rules were built carefully, a few exceptions to the detailed threshold were found. Collecting and investigating more gaze data would help to make our rules more sophisticated.

### 6 Performance of the Rule-based Filter

We assumed that our comma predictor provides a CTB text with the same distribution as the original one in CTB (see Figure 1). Accordingly, we examined the quality of the comma categorization by our rule-based filter through gaze experiments.

#### 6.1 Experimental Settings

Another five native Mandarin Chinese speakers were invited as test subjects. The CTB articles assigned to the subjects are listed in Table 10. These articles were selected from the test data in Section 4.2 in such a way that 520<$\#\text{characters}$<700, $\#\text{commas}>17$, $\#\text{commas}/\#\text{punctuations}>38\%$, and $\#\text{commas}/\#\text{characters}>3.1\%$, since we needed articles of appropriate length with a fair number of commas. After that, we manually chose articles that seemed to attract the subjects' interest from those that satisfied the conditions.

54
Case 1: LSubject + RPredicate #commas

Case 2: LPredicate + RObject #commas

Case 3: LModifier #commas

Case 4: Lc + Rc #commas

Case 5: Lp + Rp #commas

Table 9: Entire classification of rules based on traditional comma categories

Table 10: Materials assigned to each subject

Our rule-based filter was applied to the commas of each article, and the commas were classified into two distributions: a positive one (positive + semi-positive commas) and a negative one (negative + semi-negative commas). Two types of materials were thus generated by leaving the commas in one distribution and removing the others. manually checked and corrected.
The apparatus and procedure were almost the same as those in Section 5.1, whereas, on the basis of the feedback from the previous experiments, the font size, number of characters in a line, and line spacing were fixed to single optimized values, respectively, 14-point Fang-Song font occupying $15 \times 16$ pixels, 33 characters and 7 pixels.

6.2 Evaluation Metrics

We examined whether our positive/negative distributions really facilitated/obstructed the subjects’ reading process by using the following metrics:

- $TT$, $EM_{FFT} = \frac{FFT}{FT}^{10}$, $EM_{FT} = \frac{FT}{TT}^{11}$, $EM_{RT} = \frac{RT}{2CN}^{12}$, $EM_{SLO} = \frac{SLO}{TT}$,

where TT, FT, RT and CN are total viewing time, fixation time, number of regressions, and number of commas respectively, as described in Section 5.2. FFT and SLO are additionally introduced metrics respectively for the “total duration for all first-pass fixations in a target region that exclude any regressions” and for the “length of saccades from inside a target region to the outside”\textsuperscript{13}. All of the areas around commas appearing in the original article were considered target areas for the metrics. The other settings were the same as in Section 5.

6.3 Contribution of Categorized Commas

Figure 10, 11, 12, 13 and 14 respectively show TT, $EM_{FFT}$, $EM_{FT}$, $EM_{RT}$ and $EM_{SLO}$ for two types of comma distributions in each trial.

For TT, we cannot see any general trend, mainly because this time, the reading order of the text was random, which spread out the second reading effect evenly between the two distributions. For $EM_{FFT}$, we cannot reach a conclusion either. In contrast, in more than half of the trials, $EM_{FFT}$ was larger for positive distributions, which would imply that the positive commas helped to prevent the reader’s gaze from revisiting the target regions. For most trials, except for subject S whose calibration was poor and reading process was poor in M56, $EM_{FT}$ and $EM_{RT}$ decreased and $EM_{SLO}$ increased for positive distributions, which implies that the positive commas smoothed the reading process around the target regions.

7 Conclusion

We proposed an approach for modeling comma placement in Chinese text for smoothing reading. In our approach, commas are added to the text on the basis of a CRF model-based comma predictor trained on the treebank, and a rule-based filter then classifies the commas into ones facilitating or obstructing reading. The experimental results on each part of this approach were encouraging.

In our future work, we would like see how commas affect reading by using much more material, and thereby refine our framework in order to bring a better reading experience to readers.

Acknowledgments

This research was partially supported by Kakenhi, MEXT Japan [23650076] and JST PRESTO.
References

Wallace Chafe. 1988. Punctuation and the prosody of written language. Written Communication, 5:396–426.

Yuqing Guo, Haifeng Wang, and Josef van Genabith. 2010. A linguistically inspired statistical model for Chinese punctuation generation. ACM Transactions on Asian Language Information Processing, 9(2):6:1–6:27, June.

Masako Hirotani, Lyn Frazier, and Keith Rayner. 2006. Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. Journal of Memory and Language, 54(3):425–443.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Pause and stop labeling for Chinese sentence boundary detection. In Proceedings of Recent Advances in Natural Language Processing, pages 146–153.

Jing Huang and Geoffrey Zweig. 2002. Maximum entropy model for punctuation annotation from speech. In Proceedings of the International Conference on Spoken Language Processing, pages 917–920.

Mei xun Jin, Mi-Young Kim, Dongil Kim, and Jong-Hyeok Lee. 2002. Segmentation of Chinese long sentences using commas. In Proceedings of the Third SIGHAN Workshop on Chinese Language Processing, pages 1–8.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML ’01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Robert Leaman and Graciela Gonzalez. 2008. BAN-NER: An executable survey of advances in biomedical named entity recognition. In Pacific Symposium on Biocomputing (PSB’08), pages 652–663.

Baolin Liu, Zhongnong Wang, and Zhixing Jin. 2010. The effects of punctuations in Chinese sentence comprehension: An erp study. Journal of Neurolinguistics, 23(1):66–68.

Wei Lu and Hwee Tou Ng. 2010. Better punctuation prediction with dynamic conditional random fields. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP ’10), pages 177–186.

Pascual Martínez-Gómez, Chen Chen, Tadayoshi Hara, Yoshinobu Kano, and Akiko Aizawa. 2012a. Image registration for text-gaze alignment. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (IUI ’12), pages 257–260.

Pascual Martínez-Gómez, Tadayoshi Hara, Chen Chen, Kyohei Tomita, Yoshinobu Kano, and Akiko Aizawa. 2012b. Synthesizing image representations of linguistic and topological features for predicting areas of attention. In Patricia Anthony, Mitsuru Ishizuka, and Dickson Lukose, editors, PRICAI 2012: Trends in Artificial Intelligence, pages 312–323. Springer.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit.

Fu-dong Chiou Naixen Xue, Fei Xia and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. Natural Language Engineering, 11(2):207–238.

Stephan Peitz, Markus Freitag, Arne Mauser, and Hermann Ney. 2011. Modeling punctuation prediction as machine translation. In Proceedings of International Workshop on Spoken Language Translation, pages 238–245.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124(3):372–422.

Gui-Qin Ren and Yufang Yang. 2010. Syntactic boundaries and comma placement during silent reading of Chinese text: evidence from eye movements. Journal of Research in Reading, 33(2):168–177.

Burr Settles. 2005. ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. Bioinformatics, 21(14):3191–3192.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosodically based automatic segmentation of speech into sentences and topics. Speech Communication, 32(1-2):127–154.

Judy Perkins Walker, Kirk Fongemie, and Tracy Daigle. 2001. Prosodic facilitation in the resolution of syntactic ambiguities in subjects with left and right hemisphere damage. Brain and Language, 78(2):169–196.

Nianwen Xue and Yaqin Yang. 2011. Chinese sentence segmentation as comma classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 631–635.

Ming Yue. 2006. Discursive usage of six Chinese punctuation marks. In Proceedings of the COLING/ACL 2006 Student Research Workshop, pages 43–48.

ZDIC.NET. 2005. Commonly used Chinese punctuation usage short list. Long Wiki, Retrieved Dec 10, 2012, from http://www.zdic.net/appendix/f3.htm. (in Chinese).

X. Y. Zeng. 2006. The comparison and the use of English and Chinese comma. College English, 3(2):62–65. (in Chinese).
Kaixu Zhang, Yunqing Xia, and Hang Yu. 2006. CRF-based approach to sentence segmentation and punctuation for ancient Chinese prose. *Journal of Tsinghua Univ (Science and Technology)*, 49(10):1733–1736. (in Chinese).