The effects of space insertion into Japanese texts

Michiko Nakamura

Computational Linguistics Laboratory

Outline of the talk

- Study question
- Motivation & previous studies
- Eye-tracking experiment
- Results
- Discussion
- Conclusion

Research question

Will space-insertion into Japanese texts facilitate reading?

- Japanese written texts have no spaces between words

  昨日山田社長の秘書がスーパーで封筒を買った。

- If spaces are inserted,

  昨日　山田社長の　秘書が　スーパーで　封筒を　買った。

Is this easier to read?

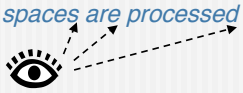
Motivation

- On-going project on determining the optimal format of a text
  - To make information on the Internet accessible to various users (e.g., elderly people, people with a disability)
- Automatic space insertion into texts can be done by using parsing tools
  - ChaSen (Matsumoto et. al 2003)
  - CaboCha (Kudo & Matsumoto 2002, 2003)

What do spaces do in reading?

- Spaces mark word boundaries

  spaces are processed in peripheral vision



- In order for an eye to extract information

  eye must land on appropriate points in a word

How about in other languages?

- Thai uses alphabetic scripts, but no spaces between words

  การปรับสำนวนแปลเช่นในหัวข่าวก็นำมาใช้ในประโยคแรกของย่อหน้านี้ได้เช่นกัน

- Kohsom & Gobet (1997)
  - Spaced texts were read faster than unspaced texts
- Will Japanese readers use these cues when present?

Experiment: overview

- Participants read spaced and unspaced texts
- Eye movements were recorded during reading
- Examined if spacing has positive effects on reading
- Results from 16 native speakers of Japanese (students from NAIST)

Apparatus (EyelinkII)

- Head-mounted eye tracker with high speed eye cameras
- Binocular recording of eye movements during reading
- Sampling rate: 500 data samples per second
- Spatial resolution: < 0.5º

Materials

- 20 (unspaced) texts
  - 13 texts from a science magazine (science articles translated into Japanese)
  - 7 texts from a newspaper corpus
- Length of texts
  - 286-386 characters
  - 8-10 lines
  - 1-3 paragraphs

Space insertion into texts

- Used parsing tools developed at CL-lab
  - ChaSen added POS-tags into texts
  - CaboCha (dependency analyzer) determined *bunsetsu* boundaries
  - One segmentation error corrected manually (ー番は なれた → ー番 はなれた)
- Spaces
  - one half-width space between *bunsetsu*
  - one full-width space at the beginning of each paragraph

Spaced and unspaced texts

- The amount of information per line was the same
  - 40 characters per line
  - Line breaks appear at the same position
- Line length is shorter in unspaced texts than in spaced texts

Procedure

- Each participant had two reading sessions (separated by a short break)
- 8 spaced texts in one session and 8 unspaced texts in another session
  - Never read the same text in two different spacing versions
  - Spacing-type and session-order were equally distributed across participants

Procedure (cont.)

- Each session begins with 2 practice texts
- In the test trials
  - A text is displayed on the screen
  - A participant reads the text silently for comprehension
  - Three questions asked after each text

    Q1: "Was the text easy to read?" [Yes or No]

    Q2 & Q3: About the content of the text [Yes or No]

Types of data

- Data from comprehension questions
  - Q1: text readability
  - Q2 & Q3: content of the text
- Eye movement data
  - Fixations
  - Total reading times
  - Regressions
- Significance level set at $p<0.05$

Comprehension questions

No difference in

- Q1: text readability [$Fs < 1$]
  - Unspaced 75.0%
  - Spaced 77.3%
- Q2 & 3: comprehension accuracy [$Ps > 0.23$]
  - Unspaced 64.4%
  - Spaced 68.7%

Fixations

The number of times an eye is fixated within a text, and its average duration

- Spaced texts (209ms) had shorter fixations than unspaced texts (216ms) [$Ps<0.005$]
  - Tendency in Session 2 only
  - Trade-off between duration and number "many short fixations vs. fewer long fixations"

Total reading times

- No difference [$p_1>0.22$; $p_2>0.4$]
  - Unspaced 62.8 sec
  - Spaced 58.9 sec

Regressions (on the same line)

eye's backward movements to earlier positions in a text

- No difference in latency and number
- Length
  - Spaced texts (151) had longer regressions than unspaced texts (133) [$p_1>0.27$; $p_2<0.05$]
  - Could be due to spaces, but see…

Distribution of regressions

The number of regressions per length range

Length	Spaced	Unspaced
0- 19	35.31	48.81
20- 39	50.50	57.38
40- 59	53.88	62.31
60- 79	53.75	54.31
80- 99	41.63	44.63
100-199	104.75	96.63
200+	84.13	84.44

Fewer regressions in spaced texts (47) than in unspaced texts (53) [$p_1<0.07$; $p_2<0.05$]

No difference over 100 ranges

Regressions (to previous lines)

- No difference in length, latency, and number

Summary of the results

- Spaced texts produced
  - Higher rating on text readability
  - Higher comprehension accuracy
  - Shorter fixations (session 2 only)**
  - Fewer short-range (<100) regressions**
- Space insertion seems to have some positive effects on reading
- but the effects are relatively weak

Discussion

- Spaces provide visual cues for word boundaries
- Japanese has other cues
  - Three sets of characters
    - Katakana → loan words
    - Hiragana → grammatical info (e.g., verb-endings, case markers)
    - Kanji → content words (e.g., nouns, verbs)
- Segmentation cues are already available in normal texts, and readers may be using them
- Then, space insertion may not have that large effects

Discussion (cont.)

- Lack of statistical power
- Comprehension questions slowed down overall reading speed
  - This may obscure the subtle but significant effects of space insertion

Conclusions

- Examined the effects of space insertion on the reading of Japanese texts
- The results suggest some positive effects of space insertion
- If space insertion does facilitate reading, potential applicability to natural documents on the Internet