To appear in the Proceedings of the 19th Annual Meeting of the
Cognitive Science Society (1997).

Adding Spaces to Thai and English: Effects on Reading

Chananda Kohsom **Fernand Gobet**
ESRC Centre for Research in Development
Instruction and Training
Department of Psychology
University of Nottingham
Nottingham NG7 2RD
United Kingdom
{cgk, frg}@psyc.nott.ac.uk

Abstract

Most research on reading has used Western languages, which have the property of being spaced. This paper examines how spacing and meaning affect reading in Thai, a modern, alphabetic and unspaced language. Results show that subjects were faster in reading and made less errors when spaces were added. Meaning facilitates reading as well, and does not interact with spacing. Finally, ability to read unspaced texts in Thai does not transfer to English. The results support the hypothesis that spaces, when present at all, offer perceptual cues that facilitate reading. Efficiency considerations raise the question of whether Thai should follow the example of Western languages and incorporate spaces and punctuation.

Introduction

Western readers are so used to seeing spaces in texts that it may come as a surprise that spaces are a relatively late feature of Western written languages. As a matter of fact, it was not until late in the eighth century that written Latin incorporated spaces and punctuation in order to improve the copying of texts (Boorstin, 1983). Even nowadays, quite a few modern written languages, such as Chinese, Japanese, and Thai, do not use spaces as a mean for separating lexical units.

Most research on reading in general and on the role of spaces in particular has been carried out using spaced languages, mainly English (e.g., McConkie et al., 1988, 1989; for experiments using Finnish, see Hyönä, Niemi, & Underwood, 1989; for a review, see Underwood & Everatt, 1992). It is known that removing spaces from normal, meaningful texts reduces reading speed by about 30-50% (Rayner & Pollatsek, 1996) with readers having little practice with such material. In general, however, saccade size and landing position in a word are similar in both types of text (Epelboim, Booth, & Steinman, 1994). Finally, Booth, Epelboim and Steinman (1995) have shown that removing spaces has more effect with meaningless material than with meaningful material.

Classical accounts of reading in English assume that spaces are important cues, which readers use in planning where to move their eyes; for example, word recognition and saccade programming are both influenced by the removal of space information (Pollatsek & Rayner, 1982; Rayner, 1993; Rayner & Pollatsek, 1996). This view has been recently challenged by Epelboim et al. (1994, p. 1735), who stress that "words, not spaces, may serve as the perceptual units that guide the line of sight through the text."

The ratio of number of spaces to number of words varies across Western languages. In particular, Dutch and German tend to collapse syntactic units into long words, although spacing still plays an important role in these two languages (for example, articles or adverbs are never compounded). Epelboim et al. (1994) report that one Dutch reader (CE) showed little decrement in reading Dutch when the spaces were removed. They suggested that CE was relatively unaffected by the removal of spaces when reading Dutch because, as other native German or Dutch speakers, CE could use the strategies he had developed to cope with the long compound words in his language.

Up to now, research has focused on the removal of spaces in spaced languages (e.g., the seminal work of Fisher, 1976). But what happens in the converse situation, where spaces are added in unspaced languages? Two competing predictions can be generated from general considerations. On the one hand, since spaces add information to a text, one could expect that readers can use this information in order to read faster and with less errors. This would support Rayner and Pollatsek's (1996) contention that spaces play a key role in reading. On the other hand, adult native speakers of an unspaced language have probably tuned their reading mechanisms, including their oculomotor strategies and lexical access processes, to the morphological characteristics of their language. Addition of spaces should disrupt these mechanisms and, as a consequence, slow down reading and increase the number of errors. This outcome would undermine Rayner and Pollatsek's theoretical position.

The case for Thai

Written Thai offers an ideal case for studying these competing hypotheses. Thai is a modern, alphabetic language, which has the peculiarity of not using spaces to segment syntactic units—spaces are used only to delimit sentences—and of rarely using punctuation signs. (Even though words run together, Thai native speakers do have the notion of words, as can been seen in the word-by-

อำกำศในฤดูใบไม้ผลีชวนให้ทำตัวขี้เกียจได้สบำยๆ หรือขำดเรียนในวันรุ่งขึ้น เพรำะเมื่อคืนว่ำนไปดื่มกับเพื่อนร่วมชั้นแล้วยังเมำค้ำง หรือใช้เวลำเพื่อคิดหำวีธีรับน้องใหม่ จัดกิจกรรมหำเงินเข้ำสโมสรและอื่นๆที่มักพบเป็นปกติในกลุ่มนักเรียนไทย แต่กระนั้นก็ตำม ในช่องว่ำงแห่งหนึ่งในนีสัยกำรตรงต่อเวลำที่กำหนดโดยสังคมนี้ ถูกใช้เป็นข้ออ้ำงของกลุ่มนักเรียนที่ขี้เกียจได้เช่นกัน พบว่ำนักเรียนที่มำโรงเรียนทุกวัน

Figure 1: An example of Thai text. This passage describes Thai student life in Japan.

word method used to teach Thai in primary school and in the presence of dictionaries using lexical entries.) There are 44 consonants in Thai and 32 basic vowels which are used with consonants to form words but which are not counted as part of the 44-letter Thai alphabet. In the Thai alphabet consonants and vowels are considered separately because they are governed by different rules. Thai vowels are written in a number of positions in relation to the consonant or consonants (before, after, above or under), so the position of a certain vowel is fixed and unchangeable (for more about the Thai language, see, for example, Allison, 1994).

In Thai, spaces are used only for delimiting sentences and, rarely, for emphasising words. Note that sometimes the lack of spacing may lead to ambiguity, for example when the same string of characters may be parsed into words in different ways, which may result in different meanings. To distinguish the meaning of those words, Thai readers may have to read all along the sentence until they find out what those words really mean. In general, no punctuation mark is used in Thai. (See Figure 1 for an example of Thai text).

We were interested in three questions in relation to spacing in a reading-aloud task. First, we wanted to test the two competing hypotheses mentioned above. Second, we wanted to see whether spacing and the degree of meaning of the text interact in the same way in Thai readers as it does in English readers (Booth et al., 1995). Third, we wanted to see whether Thai readers adapt easily to the situation where spaces are removed in a spaced foreign language, in this case, English.

Methods

Subjects

We studied five female and three male Thai native speakers studying at the University of Pittsburgh (mean age 26 years; mean TOEFL 566 points), with educational background ranging from undergraduate degree (5 subjects) to Masters' degree (3 subjects).

Design

The within-subject design included three independent variables: (1) Language: text in Thai (subjects' native language) or in English; (2) Spacing: spaced or unspaced texts; (3) Coherence: coherent or incoherent texts. Dependent variables were the time taken to read a paragraph, as well as the type and number of errors in reading.

Material

Each cell (Language x Coherence x Spacing) consisted of 4 paragraphs of around 10 lines each. Two different texts were chosen, one in Thai, the other in English. Both texts dealt with Japanese politics, were semi-technical and were judged of roughly the same difficulty. Eight consecutive paragraphs were selected for each language (sometimes changing the way the original text was organized in paragraphs). For each paragraph, we tried to keep the average number of lines (English = 10.4, Thai = 9.1), the number of words (English = 86.9, Thai = 98.2), and the number of characters (English = 602, Thai = 621) similar between the two languages. The small differences are explained mainly by the presence of accents in Thai but not in English and by the fact that Thai words are on average smaller than English words (cf. Allison, 1994).

Manipulation of spacing and creation of incoherent text was carried out in a way similar to Booth et al. (1995), both for the English and Thai texts. In the unspaced condition with English, all spaces were simply removed. To keep the unspaced conditions consistent across languages, we also removed the spaces at the end of Thai sentences. In the spaced condition with Thai, one space was added after each word; thus, the end of a sentence was now delimited by two spaces. Incoherent paragraphs were built by randomizing coherent paragraphs, replacing randomly a word by a word of the same size. For the English text, capital letters were maintained at the beginning of a sentence and for proper nouns. (Thai does not differentiate between uppercase and lowercase letters). Words within a paragraph were randomized differently for each subject. Both with coherent and incoherent texts, the paragraphs were presented in the same order as in the original document. Half the subjects received the English texts first, and half the subjects received the Thai texts first. Within each language, the presentation of the 4 possible permutations spacing x coherence was counterbalanced. Note that the length of the text varies between the spaced and unspaced conditions. This is necessary in order to keep the number of words in a line constant between the conditions. Figure 2 illustrate, for Thai and English, the eight experimental conditions.

Thai, Spaced, Coherent

นับ ตั้งแต่ ที่ พรรค เสรี ประชำธีปไตย ได้ สูญเสีย อำนำจ เบ็ดเสร็จ ใน ฐำนะ
พรรค รัฐบำล เป็น ต้น มำ กำร เมือง ญี่ปุ่น ก็ อยู่ ใน สภำพ ยุ่ง เหยิง มำ โดย
ตลอด ภำพ ที่ ปรำกฏ ขึ้น เห็น เด่น ชัด คือ พรรค ข้ำรำชกำร ต้อง เข้ำ

Thai, Spaced, Incoherent

ที่ ตั้งแต่ นี้ หลวง เล่น นักกำรเมือง ไหน หลังจำก เมือง ข้ำรำชกำร ใน เด่น
เป็น ปรียำย เพำะ นี้ ใน ให้ อย่ำง ญี่ปุ่น ก็ ผนึก จะ สมัย เป็น เมือง ผล หำด
เสนอ ได้ และ อำนำจ เช่น เห็น อยู่ ปัน ของ เล่น ข้ำรำชกำร ต้อง เข้ำ

Thai, Unspaced, Coherent (normal text without end-of-sentence spaces)

นับตั้งแต่ที่พรรคเสรีประชำธีปไตยได้สูญเสียอำนำจเบ็ดเสร็จในฐำนะ
พรรครัฐบำลเป็นต้นมำกำรเมืองญี่ปุ่นก็อยู่ในสภำพยุ่งเหยิงมำโดย
ตลอดภำพที่ปรำกฏขึ้นเห็นเด่นชัดคือพรรคข้ำรำชกำรต้องเข้ำ

Thai, Unspaced, Incoherent

และกระทรวงนี้เป็นนำยกประชำธีปไตยจึงหน้ำที่กลุ่มศูนย์กลำงในจริง
เสรีนโยบำยเสียตำมในว่ำเรืองรีบร้อนใจคว่ำมก็เข้ำเข้ำระบอบจะของ
เช่นไหนตำมเพื่อซึ่งแปลกเป็นดันที่ร่ำงอย่ำงยิ่งเช่นเด่น

English, Spaced, Coherent (normal text)

In industrial democracies, governments differ with respect to
their welfare policies. In Europe, governments at times respond
directly to organized groups demanding changes in welfare

English, Spaced, Incoherent

As autonomous bureaucrats, bureaucrats offers been despite to
seeks welfare specific. At policy, democracies is focus paradox
programs of dominance social emergence explain of explain

English, Unspaced, Coherent

Inindustrialdemocracies,governmentsdifferwithrespectto
theirwelfarepolicies.InEurope,governmentsattimesrespond
directlytoorganizedgroupsdemandingchangesinwelfare

English, Unspaced, Incoherent

Toapproachesexplanation,responsiblesociallackwelfareor
needsresolveresearch.Topolicy,bureaucratsispointwelfare
policiesofprecedentsocialelectoralWesterniswelfare

Figure 2: Examples of Thai and English texts used in the experiment, in the four spaced x coherent permutations. (Actual texts were on average 10 lines long).

The texts were presented on a Macintosh Quadra with a 13-inch viewable diagonal screen, using a program written in Hypertalk. The texts were in black font; the background screen was light blue. Thai texts were presented with the Thonburi Font, 18 point, and English texts were presented with the Times font, 18 point. Texts in both languages had roughly the same appearance, to the extent that this is possible with alphabets so different.

Procedure

In the introduction to the experiment, subjects were required to speak loudly and clearly, so that they could be clearly recorded. A single paragraph appeared on the screen. Subjects pressed the space key to indicate that they had finished reading a paragraph. There were two practice paragraphs (one spaced and one unspaced) before each language x coherence block. Within each such block, spaced and unspaced paragraphs were presented in pairs. Before each paragraph, subjects received a message (always in English), either "Thenextparagraphwillbeunspaced", or "The next paragraph will be spaced". Subjects were recorded when they read the text. The experiment lasted on average about one hour and a half.

Results

Results were analysed separately for each language using an Analysis of Variance with repeated measures on Coherence and Spacing. Except for planned comparisons, we refrained of carrying a three-way analysis of variance, because the third independent variable (Thai vs. English) was not under direct experimental control. The outcome of such a three-way analysis of variance would be hard to interpret, as several variables would be confounded, mainly: language, average length of words, differential experience with each language.

We first present the analysis of reading times, and then the analysis of reading errors. In both cases, we first discuss the Thai condition, and then the English condition.

Reading Time

Figure 3 shows the average reading times for the Thai and English texts. (We have presented the results as the time to read a paragraph instead of as the time to read a word, because of the different average length of words in English and in Thai). With Thai texts, subjects were slower to read meaningless texts [$F(1, 7) = 38.68$, $p < .001$], while the addition of spaces made reading faster (the effect is marginally significant: $F(1, 7) = 4.21$, $p < .08$). With coherent texts, 6 subjects out of 8 were faster when reading spaced texts, and with incoherent texts, 7 subjects out of 8 were faster when reading spaced texts. There was no interaction between Coherence and Spacing [$F(1, 7) = 0.94$, ns.]. Subjects' restrospective reports indicate that, if anything, they thought that spaces were making reading harder. In addition, no Subject made comments on the absence of spaces at the end of sentences. This suggests that they were not negatively affected by this feature of our material.

With English texts, the two main effects were significant: coherent texts [$F(1, 7) = 14.65$, $p < .01$] and spaced texts [$F(1, 7) = 50.75$, $p < .001$] were read faster than incoherent and unspaced texts, respectively. There was no interaction [$F(1, 7) = 2.35$, ns].

Subjects show then the same general pattern of results with the two languages: they were faster with coherent texts than with incoherent texts, and they were faster with spaced texts than with unspaced texts. There are, however, two clear differences between the languages. First, Subjects were faster [$t = 4.77$, $p < .001$] to read texts in Thai, their native language (52.95 sec, on average) than texts in English, their second language (76.48 sec, on average). For the condition English/Coherent/Spaced, it took .66 sec per word on average (with a minimum average of .53 sec per word for the faster Subject and a maximum average of .91 sec per word for the slowest Subject), or 57.4 sec per paragraph. For the condition Thai/Coherent/Unspaced, which approximates written Thai, it took 0.47 sec per word on average (with a minimum of .36 and a maximum of .61 sec per word), or 46.2 sec per paragraph. Second, the effect of spacing is stronger with English (on average, a difference of 33.72 sec between the spaced and unspaced conditions) than with Thai (on average, a difference of 1.37 sec) texts [$t = 6.76$, $p < .001$].

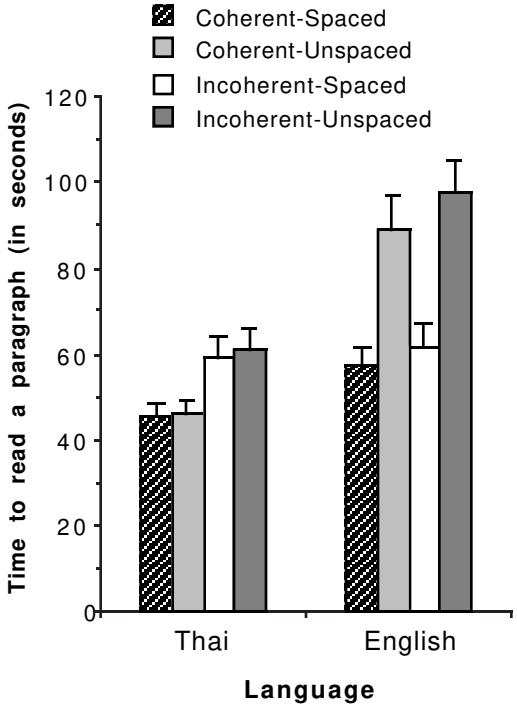

Figure 3: Average time to read a paragraph, as a function of language, coherence and spacing. The error bars indicate the standard errors of the mean.

	Coherent Spaced	Coherent Unspaced	Incoherent Spaced	Incoherent Unspaced
			Thai	
Word pronounced incorrectly	0.3	0.7	1.3	1.1
Word repeated	0.2	0.2	0.2	0.2
Word skipped	0.5	0.4	0.6	0.8
Word added	0.3	0.5	0.5	1.0
Order of words reversed	0.1	0.0	0.1	0.2
Total errors	**1.4**	**1.7**	**2.7**	**3.4**
			English	
Word pronounced incorrectly	1.8	4.3	2.0	6.1
Word repeated	0.2	0.4	0.4	0.5
Word skipped	0.1	0.5	0.4	0.9
Word added	0.2	0.4	0.4	0.5
Order of words reversed	0.0	0.0	0.0	0.0
Total errors	**2.4**	**5.6**	**3.2**	**8.0**

Table 1: Average number of errors in reading a paragraph per language and per type of error.

Errors in Reading

We coded reading errors into five categories: 1. Word pronounced incorrectly (e.g., "theories" read as "theory"); 2. Word repeated (e.g., "in" read as "in in"); 3. Word skipped (e.g., "sponsorship to a abandon" read as "sponsorship to abandon"); 4. Word added (e.g., "one welfare" read as "one of welfare"); 5. Words in wrong order (e.g. "to the" read as "the to"). Table 1 gives the results for each language and each type of error.

With Thai, all types of errors, except for "Word added" did not show any main effect or interaction. With "Word added," spaced texts [$F(1,7) = 6.44$, $p < .05$] and coherent texts [$F(1,7) = 5.72$, $p < .05$] showed less errors. There was no interaction.

With English, there was no significant main effect nor interaction for the following types of errors: Word repeated, Word added, and Word in wrong order. With the type of error "Word pronounced incorrectly," spaced texts [$F(1,7) = 56.42$, $p < .001$) and coherent texts (the effect is marginally significant: $F(1,7) = 4.11$, $p = .082$) had fewer errors. There was no interaction. With the type of error "Skipped word", both the effect of Coherence [$F(1,7) = 5.46$, $p = .052$] and of Spacing [$F(1,7) = 4.63$, $p = .068$] were marginally significant, and there was no interaction. In both cases, the absence of meaning and spaces increased the number of errors in reading.

On average, subjects made more errors for English texts than for Thai texts. Interestingly, the total number of errors mirrors the reading time data in Figure 2 almost exactly. Comparisons of the results for the two languages in Table 1 shows that the main source of errors lies in the number of word pronounced incorrectly.

Discussion

We addressed three questions in this study. First, we tested two competing hypotheses: does adding spaces to Thai texts slow down readers, because it interferes with the well tuned mechanisms that these readers have developed over the years, or does it help reading, because it adds information to the text? Data on reading time and errors give support to the latter hypothesis. Thai readers were able to use the visual cues signalled by spaces to enhance their reading speed and diminish errors even in Thai, although spaces were more useful for them with English texts. As proposed by Rayner (1993), spaces, if present at all, seem to play a key, universal role in reading.

With errors, we found that Spacing and Coherence had a reliable effect on the way subjects made errors of pronunciation and skipped words with English, and on the way they added words with Thai. Both the deletion of words in the second language and the addition of words in the native language may be explained by subjects' knowledge of conditional probabilities for each language. With the second language, such knowledge is weak, so subjects do not easily access words in difficult (incoherent or/and unspaced) texts, and therefore tend to skip words. With their native language, such knowledge is strong, so subjects access easily words suggested by the previous words (and sometimes by the next words), and therefore, when reading difficult texts, tend to add words likely to be present. Finally, the effect of Spacing and Coherence on the errors of pronunciation in English, but not in Thai, may be explained by the fact that the mechanisms to generate a correct pronunciation are not developed in English as well as in Thai. While subjects could control pronunciation to some extent when reading normal texts, their attention was used to other goals when reading the unspaced and incoherent texts, which produced more errors of pronunciation.

Second, we were interested in the interaction between Coherence and Spacing. As mentioned earlier, Booth et al. (1995) found such an interaction, and, based on this result, concluded that semantics, rather than spaces, are the most important determinants of reading speed and errors (Incidently, this conclusion does not follow logically from their data, nor do their data validate the prediction of a clearly stated theoretical model). In our results, there was no such interaction either for reading time or for errors in reading, either with English texts or with Thai texts. Thus, the effects of Coherence and Spacing were additive, which runs against the prediction of Booth et al. (1995).

Third, we tested whether there is a transfer from reading a native unspaced language to reading a second, spaced language where spaces have been added. We found that there was no such transfer. With coherent texts, the proportion of increased time in reading unspaced texts as compared to spaced texts is even higher for our Thai native speaker subjects (55 %) than for English native speaker subjects (from 30% to 48%, cf. Rayner & Pollatsek, 1996).

In their conclusion, Booth et al. (1995) proposed that "reading slows down when spaces are removed because removing spaces impairs word recognition when letter groupings become ambiguous, and not because removing spaces impairs saccadic programming." They used the presence of an interaction between Coherence and Spacing as evidence for their position. The lack of such an interaction in our results raises doubts about the generality of this conclusion. In general, we believe that, instead of opposing word recognition against saccadic programming, as was done in a recent theoretical controversy (Epelboim et al. 1994, 1996; Rayner & Pollatsek, 1996), one should try to develop a theory which explains how these two processes occur together.

The lack of preference for non-spaced texts in reading Thai, even though this is what the readers will have encountered before, as well as the difficulty our Subjects had with unspaced texts in English, run counter to the claims of Epelboim et al. (1994). Similarly, the lack of an interaction between spacing and coherence strongly suggest that readers use reading strategies based on spaces independently from strategies based on meaning.

Thai native speakers, although they are used to reading unspaced text in their own language, can still use the information provided by spaces to improve their reading, both with respect to speed and errors. Presumably, this information is used in addition to the automated processes they have developed in learning how to read. Spacing seems then to be a cue that is universal across languages, which readers can use, in addition to cues specific to each language, to know when words are likely to end. Whether these cues are used to direct eye movements or to facilitate lexical access and whether the advantage offered by spacing extends to punctuation are questions for further research.

It is common knowledge that texts written in a cluttered way tax short-term memory and slow down reading and comprehension (e.g., Burnett, 1990). Taken together with the result of our experiment, this suggests that adding spaces in Thai would have an important impact on Thai education. Should Thai, like Latin about thousand years ago, incorporate spaces to delimit words in order to make reading (and writing) easier? Assuming that our results are robust, this decision depends on a trade-off between the cognitive and educational gain of adding spaces and its cost—cultural, social, and economic.

Acknowledgements

This research was carried out when the first author was at the School of Information Sciences of the University of Pittsburgh. We thank Stephen Hirtle for his advice and his logistic help in running the experiment, as well as Alan Kennedy, Julian Pine, Keith Rayner, and anonymous reviewers for comments on earlier drafts.

References

Allison, G. H. (1994). *Easy Thai. An introduction to the Thai language.* Rutland, Vt.: Tuttle.

Boorstin, D. J. (1983). *The discoverers*. New York: Random House.

Booth, J. R., Epelboim, J., & Steinman, R. M. (1995). The relative importance of spaces and meaning in reading. *Proceedings of the 17th Cognitive Science Society* (pp. 533-538). Mahwah, NJ: Lawrence Erbaum Associates.

Burnett, R. (1990). *Technical communication.* Belmont, CA: Wadsworth.

Epelboim, J., Booth, J. R., & Steinman, R. M. (1994). Reading unspaced text: Implications for theories of eye movements. *Vision Research, 34,* 1735-1766.

Epelboim, J., Booth, J. R., & Steinman, R. M. (1996). Much ado about nothing, The place of space in text. *Vision Research, 36*, 465-470.

Fisher, D. F. (1976). Spatial factors in reading and search: The case for space. In R. A. Monty and J. W. Senders (Eds.), *Eye movements and psychological processes*. New York: Wiley.

Hyönä, J., Niemi, P., & Underwood, G. (1989). Reading long words embedded in sentences: Informativenes of word halves affects eye movements. *JEP: Human Perception and Performance, 15*, 142-152.

McConkie, C. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1988). Eye movement control during reading: I. The location of initial eye fixations on words. *Vision Research, 28*, 1107-1118.

McConkie, C. W., Kerr, P. W., Reddix, M. D., Zola, D., & Jacobs, A. M. (1989). Eye movement control during reading: II. Frequency of refixating a word. *Perception & Psychophysics, 46*, 245-253.

Rayner, K. (1993). Eye movements in reading: Recent developments. *Current Directions in Psychological Science, 2*, 81-85.

Rayner, K. & Pollatsek, S. (1996). Reading unspaced text is not easy: Comments on the implications of Epelboim et al.'s study (1994) for models of eye movement control in reading. *Vision Research, 36*, 461-465.

Underwood, G., & Everatt, J. (1992). The role of eye movements in reading: some limitations of the eye-mind assumption. In E. Chekaluk & K. R. Llewellyn (Eds.), *The role of eye movements in perceptual processes*. Amsterdam: Elsevier Science Publishers.