Essay Length, Lexical Diversity and Automatic Essay Scoring
by
Andrew MELLOR
Department of Media Science
Faculty of Information Science and Technology
(Manuscript received September 30, 2010)

Abstract
A simple two-dimensional model of essay quality consisting of essay length and lexical diversity is investigated and tested on a set of essays written by second language (L2) learners of English. The results from the model are compared to the ratings of a human judge. A number of different measures of lexical diversity are compared.

Keyword : second language acquisition, vocabulary, automatic assessment, testing, CALL
Introduction

This experiment explores a very simple model for automatic essay assessment. Plenty of studies have shown that both essay length and lexical diversity can be strongly correlated with L2 essay ratings, for example, Larsen-Freeman & Strom (1977), Ferris (1994) and Engber (1995). However, despite significant correlations, these two features are by themselves inadequate both empirically and theoretically to account for essay ratings. This experiment investigates whether together they may be able to better account for essay ratings. To this end, a simple two-dimensional quantity/content model of essay assessment is employed in which quantity is represented by essay length and content by lexical diversity.

In addition, the performances of various measures of lexical diversity are compared. The following measures of lexical diversity are considered:

- $TTR(N)$ – the type-token ratio for a sample of $N$ words
- Guiraud’s Index (Guiraud, 1960)
- Yule’s $K$ (Yule, 1944)
- The $D$ estimate (Malvern et al., 2004)
- Hapax ($N$) – the number of hapax legomena in a sample of $N$ words
- An estimate of Advanced Guiraud (Daller, van Hout & Treffers-Daller, 2003)

Aims of the experiment

The aims of this experiment are to see:

1) if a two-dimensional quantity/content model employing essay length and lexical diversity can predict human assessment of essays better than a single dimension of either quantity or content.

2) which measure of lexical diversity works best as the content dimension alongside the quantity dimension of essay length for this set of essays.

Experimental design

Tasks

A set of 34 timed essays was collected from a group of third year English majors at a Japanese university. The task involved writing an essay in thirty minutes given the following prompt:

Watching television is bad for children. Do you agree or disagree with this statement? Use specific reasons and examples to support your answer.

(See Appendix for sample essays)

Essay ratings

The quality of the essays were assessed by a native speaker judge and rated as: good, above average, average, below average, or poor. The number of essays for each rating is shown in Table 1.

<table>
<thead>
<tr>
<th>Rating</th>
<th>good</th>
<th>above average</th>
<th>average</th>
<th>below average</th>
<th>poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of essays</td>
<td>4</td>
<td>5</td>
<td>18</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

Analysis

Selection of quantity and content measures

A simple two-dimensional model of quantity and content was used. The two dimensions were given equal weight. Quantity was represented by essay length in words and content was represented by lexical diversity. In addition, several lexical diversity measures were evaluated. There were
two considerations in selecting these measures. One was a correlation with essay ratings in other studies. Another was ease of calculation enabling incorporation into a simple computer program. Malvern & Richards’ D estimate involves a complex calculation. However, it was included in this analysis for comparison as the most commonly used measure of lexical diversity. The measures considered in this analysis were as follows:

1) TTR(100), the number of word types in a hundred word sample
2) Guiraud Index
3) Yule’s K
4) the D estimate
5) Hapax(100), the proportion of hapax legomena in a hundred word sample
6) an estimate of Advanced Guiraud

The first four measures are all alternatives to standard TTR and attempt to get around the effect of essay length. Standard TTR tends to fall as essay length increases and so can penalize longer essays. TTR of a fixed sample compensates for length effects but the drawback of this measure is that it does not use all the word types for many of the longer essays. Guiraud Index (GI) is a variant of the TTR and is the number of word types divided by the square root of the number of word tokens as follows:

\[ GI = \frac{\sqrt{v}}{\sqrt{N}} \]

where \( v \) is the number of word types and \( N \) is the number of word tokens. Guiraud Index is also affected by essay length but in a different way to TTR. Malvern et al. (2004) found that Guiraud Index rises with increasing essay length for the first few hundred words but then decreases steadily with increasing essay length.

Yule’s K is calculated using the following formula:

\[ K = 10^4(\sum r^2V_r - N)/N^2 \quad (r = 1, 2, \ldots) \]

where \( V_r \) is the number of word types occurring \( r \) times in a text consisting of \( N \) word tokens. Yule’s K is independent of text length but usually only for longer texts over one thousand words or so. However, it was found to be remarkably robust even for some short L2 essays (Mellor, 2010). The D estimate (Malvern et al., 2004) is extracted from the following formula:

\[ \text{TTR} = \frac{D}{N} \left[ 1 + 2 \frac{N^{\frac{1}{2}}}{D} \right] - 1 \]

where \( N \) is the number of tokens in the essay.

The D estimate has been found to be independent of essay length for shorter essays (Malvern et al., 2004). However, it requires a complex calculation and in this experiment was calculated using \text{D-Tools} software (Meara & Miralpeix, 2007).

Results of an earlier experiment (Mellor, 2008) suggested that the proportion of hapax legomena (words occurring only once in a text) may be a good discriminator between learner and native essays. The proportion is included here to see how it discriminates between learner essays. In the earlier experiment, the comparisons were made on essay samples of equal length. As this measure may also be affected by length, it was calculated for a hundred word sample. The final measure, Advanced Guiraud, was proposed by Daller, van Hout & Treffers-Daller (2003). Advanced Guiraud \( (A_c) \) is the number of advanced word types divided by the square root of the total number of tokens as follows:
where $v_a$ is the number of advanced word types and $N$ is the total number of word tokens. As with Guiraud Index, Advanced Guiraud may be affected by essay length but to a lesser extent than TTR. Since identification of advanced types is not a simple procedure for a computer application, an estimate of Advanced Guiraud is used in this experiment. This involves an estimate of advanced word types calculated by subtracting frequent types from total types. For this study, frequent types were defined as word types from a list based on the first 1000 words of the JACET word list (Ishikawa et al., 2003). A concern is that this Advanced Guiraud estimate may be contaminated by error variants of frequent types.

Sampled features
The effect of essay length on TTR is well documented. As essay length increases, TTR tends to decrease. This makes it problematic for comparisons of essays of varying length. There is also a suggestion that the proportion of hapax legomena may decrease with essay length. Therefore, both of these features were calculated for a hundred word sample from each essay. Only TTR and hapax legomena were calculated for sampled values. The notation $\text{TTR}(100)$ and $\text{Hapax}(100)$ is used to show these measures are based on a sample of 100 words. All other features were calculated using raw data.

However, six essays were shorter than one hundred words. These essays were of length 95, 95, 86, 79, 68 and 50 words. These six essays were treated differently for these two sampled measures. To make up the shortfall of words, additional words were sampled from each essay and added to the essay to bring the word count to a hundred. Therefore, the word tokens were increased to a hundred but the word types were not increased. This procedure was incorporated into the computer program.

Guiraud measures have also been found to be affected by essay length. Given the range of essay length in this set of essays, these Guiraud measures are likely to be rising against essay length. Features with a positive relationship with essay length are easier to incorporate into an analysis than features with a negative relationship. Therefore, the Guiraud measures are not controlled for essay length but the relationship with essay length needs to be taken into account in interpretation.

Calculation of quantity and content dimensions
For each essay, the measures of lexical diversity were calculated along with essay length. To ensure an equal weighting and easy comparison between essay length and the measure of lexical diversity, the values for each measure were standardized over the 34 essays by transforming into z scores. Z scores are calculated by subtracting the mean from each value and then dividing by the standard deviation. Standardization ensures that the mean of each measure over the essays is zero with a standard deviation of one.

Comparison of content variables
The two-dimensional model using z scores was graphed for each measure of lexical diversity. The use of z scores is quite conducive to easy interpretation. On a simple graph, four quadrants can be seen as shown in Figure 1.
Quantity as represented by essay length is graphed on the x-axis while content represented by a measure of lexical diversity is graphed on the y-axis. Quadrants A and B are of particular interest. Quadrant A indicates a domain where essays are above average on both dimensions. If this quantity/content model is sound then highly rated essays should appear in this domain. Conversely, Quadrant B indicates a domain where essays are below average on both dimensions. We expect essays in this domain to be rated low. Quadrant C shows essays that are above average on one dimension, in this case essay length, but below average on the second dimension, lexical diversity. Quadrant D shows essays that are above average on the second dimension, lexical diversity, but below average on essay length. The quality of essays appearing in Quadrants C and D is likely to be more difficult to predict. In this graphical representation, essay points close to the origin will also be close to other quadrants and any prediction based on position in the quadrants would be potentially unreliable. Therefore, we would hope that essays assessed high or low would be clearly defined in the respective quadrants not close to the origin or either axis.

Yule’s K is a special case of lexical diversity because its values have an inverse relationship with diversity. Lower values of Yule’s K indicate higher lexical diversity and higher values indicate lower lexical diversity. To counter this and align Yule’s K with the other variables, a negative value of the measure was graphed.

Results

Basic characteristics
The values of the various base measures before standardization are shown in Table 2.

Table 2: Essay characteristics of 34 learner essays

|       | length | TTR (100) | G/ | K    | D    | Hapax (100) | A
g |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>146.3</td>
<td>64</td>
<td>7.1</td>
<td>121.7</td>
<td>75.9</td>
<td>46.6</td>
<td>5.2</td>
</tr>
<tr>
<td>SD</td>
<td>56.6</td>
<td>9</td>
<td>0.9</td>
<td>34.6</td>
<td>25.3</td>
<td>8.8</td>
<td>0.7</td>
</tr>
<tr>
<td>High</td>
<td>328</td>
<td>78</td>
<td>9.1</td>
<td>192</td>
<td>153.0</td>
<td>65</td>
<td>6.6</td>
</tr>
<tr>
<td>Low</td>
<td>50</td>
<td>34</td>
<td>5.0</td>
<td>58.7</td>
<td>38.3</td>
<td>25</td>
<td>3.5</td>
</tr>
</tbody>
</table>

All the measures showed a wide range of values. For example, essay length ranged from 50 words to 328 words.

Quantity/content graphs
Sample word types TTR(100)
Figure 2 shows a quantity/content graph using TTR(100) as the content dimension. Essays rated good are indicated by 1, above average by 2, average by 3, below average by 4 and poor by 5.
In this graph, Quadrant A contains three of the four most highly rated essays but only two of the above average essays. It also contains four average essays but also one below average essay. However, these essays are not clearly defined. In Quadrant B, the one poor essay is well isolated as are two below average essays. However, this clear definition is deceptive. TTR (100) was based on a hundred word sample but these clearly defined essays all contained less than a hundred words. The topping up of tokens may have emphasized the dearth of types in these essays so the poor performance of these essays probably owes itself as much to the shortage of words as to the content dimension.

Overall, the two-dimensional model seems inferior to a one dimensional model using only essay length which would clearly identify some good and above average essays at the top end of the essay rating spectrum.

Figure 3 shows Guiraud Index as the content dimension. This seems to perform better than TTR (100) with two good essays and one above average essay well defined in Quadrant A. Similarly, the poor essay and two below average essays are well-defined in Quadrant B. This suggests that a two-dimensional model based on Guiraud Index would be able to clearly isolate essays at the top and bottom of the rating spectrum. The model seems to indicate essays at the bottom end of the scale more clearly than essay length by itself although essay length may still be a superior discriminator at the top end.

Yule’s K

Figure 4 shows Yule’s K as the content element. As with TTR (100), essays are not very well defined at the top of the range but the measure seems better at the bottom end distinguishing the poor essay and some below average essays. However, it does not seem to be an obvious improvement over simple essay length.
Hapax(100)

Figure 5 shows Hapax (100), the proportion of hapax legomena in a 100 word sample, as the content element. As with TTR (100), values at the bottom end may be emphasized by the short length of the essays rather than the lexical diversity measure. At the top end, the performance is inferior to essay length by itself.

The D estimate

Figure 6 shows that the D estimate isolates essays at the bottom end of the scale well with several poor and below average essays clearly defined. In Quadrant A, the D estimate does not define highly rated essays so clearly. D seems the most conservative estimator in that no major misclassifications occur. No below average or poor essays appear at all in Quadrant A and no good or above average essays appear in Quadrant B. However, it does not highlight essays at the top end of the scale as well as Guiraud Index.

Advanced Guiraud

Figure 7 shows an estimate of Advanced Guiraud as the content dimension. Advanced Guiraud seems to perform well at the extreme of both ends of the scale clearly identifying two good essays, one poor and one below average essay. However, other essays are less clearly identified.
Andrew MELLOR

The graphical evidence suggests that for the essays in this study, essay length was a very strong predictor of overall essay ratings. Certainly, the content dimension of the model seems to take second place to the quantity dimension. This means that perhaps we need to rethink the equal weighting of the model to one that offers a greater contribution to quantity. Guiraud Index and the Advanced Guiraud estimate seem to perform better than the other lexical measures in terms of clearly defining a small number of high quality or low quality essays. However, the fact that these may both still depend on length may be accounting for their performance to some extent.

**Correlation analysis**

Looking at graphs can give us a general idea about relationships between the variables. Correlation analysis may help give some more objective evidence.

<table>
<thead>
<tr>
<th>TTR (100)</th>
<th>GI</th>
<th>K</th>
<th>D</th>
<th>Hapax (100)</th>
<th>A_G</th>
<th>ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>.36</td>
<td>.68</td>
<td>.25</td>
<td>.20</td>
<td>.23</td>
<td>.65*</td>
</tr>
<tr>
<td>TTR</td>
<td>.81</td>
<td>.84</td>
<td>.80</td>
<td>.94</td>
<td>.70</td>
<td>.47*</td>
</tr>
<tr>
<td>GI</td>
<td>.78</td>
<td>.78</td>
<td>.85</td>
<td>.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>.93</td>
<td>.82</td>
<td>.64</td>
<td>.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>.85</td>
<td>.60</td>
<td>.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hapax</td>
<td>.58</td>
<td>.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A_G</td>
<td></td>
<td>.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*significant at the .001 level  **significant at the .01 level

Table 3 shows correlation values between the features and essay ratings. The correlations with essay ratings underline the impact of essay length which has the highest correlation of $r = .79$. The next two highest are Advanced Guiraud and Guiraud Index with $r = .66$ and $r = .61$ respectively. These very high correlations may reflect an indirect influence of essay length. The other measures are broadly similar with TTR(100) performing the best, $r = .47$. This contradicts the graph results somewhat but alerts us to an important aspect of many assessment situations which is decision accuracy. The identification of essays at the top of the range and the bottom of the range may be more important than essays in the middle because there is a chance with these essays to make a larger misclassification error.

The strong correlations of Guiraud Index and Advanced Guiraud with essay length, $r = .68$ and $r = .65$ respectively, support the idea that they are still considerably affected by length of essay. Generally high correlations between the different measures of lexical diversity suggest that they are all measuring broadly the same thing. Very high correlations between pairs of these measures reflect similarities in properties. The Guiraud
Index and Advanced Guiraud, r = .85, are both counts divided by the same denominator and also both depend on essay length. The correlation between Yule’s K and the D estimate is particularly high, r = .93, perhaps reflecting the fact that they both represent parameters of word distributions. The correlation between TTR(100) and Hapax(100) is high, r = .94, probably because they were both measured on the same sample of words.

Partial correlation analysis
Essay length seems to be the best predictor of essay ratings with this set of essays. In addition, we are interested in the measure of lexical diversity that in conjunction with essay length can best improve on this prediction. One way to approach this is to identify the measure which can add something to a one-dimensional essay length model.

Table 4 shows the partial correlations of the measures of lexical diversity with essay ratings when essay length is controlled for. These partial correlations tell us how much each measure correlates with essay ratings if the influence of the length effect is taken out. Guiraud Index and Advanced Guiraud values are interesting. Although both of these have relatively strong correlations with essay ratings, they are both considerably weakened when controlled for essay length. This again underlines the effect of essay length on these measures. Guiraud Index has by far the lowest partial correlation with essay ratings. The other measures are all about the same but TTR(100) and Advanced Guiraud are the strongest. However, excluding Guiraud Index, the measures vary in a short range from r = .27 to r = .32 so it is probably unwise to make any conclusions on which is best.

Multiple regression analysis
A multiple regression analysis was run to see how well each measure of lexical diversity performed alongside essay length in predicting essay ratings. Each measure was used with essay length in a two independent variable model with essay ratings as the dependent variable. Table 5 shows the r values and r² values for each measure of lexical diversity in a two variable model with essay length.

Table 5: R and r² values for two variable model with essay length

<table>
<thead>
<tr>
<th></th>
<th>TTR (100)</th>
<th>GI</th>
<th>K</th>
<th>D</th>
<th>Hapax (100)</th>
<th>A_G</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>.81</td>
<td>.79</td>
<td>.81</td>
<td>.81</td>
<td>.80</td>
<td>.81</td>
</tr>
<tr>
<td>r²</td>
<td>.66</td>
<td>.63</td>
<td>.65</td>
<td>.65</td>
<td>.64</td>
<td>.66</td>
</tr>
</tbody>
</table>

The regression analysis highlights the influence of essay length which alone accounts for 61.7% of the variance. Another dimension of lexical diversity brings only a slight improvement to the model with TTR(100) and Advanced Guiraud performing best and accounting for a further 4% of the variance.

Results summary
The results suggest that lexical diversity together with essay length can more accurately predict essay ratings than either feature alone with this set of
essays. However, essay length is a very strong predictor as a single dimension. Accordingly, in a two-dimensional model, essay length seems to be the dominant dimension with a greater weighting than lexical diversity.

It is less clear which measure of lexical diversity is the best to use in the two-dimensional model. The results suggest that TTR(100), Yule’s K, Hapax(100) and Advanced Guiraud perform similarly well in correlation and regression analyses. They all perform as well as the standard measure of lexical diversity, the D estimate. Guiraud Index seems to perform less well. The graphical evidence suggests that Advanced Guiraud and Guiraud Index are good at clearly identifying a few essays at the top and bottom end of the scale. TTR(100), Hapax(100), and Yule’s K seem less able to distinguish essays clearly. However, this may be partly due to the fact that the graph model gives the two dimensions an equal weight. The regression analysis suggests that essay length deserves a greater weight. Because the two Guiraud measures are affected by essay length, their inclusion as the content dimension in effect boosts the essay length weighting to a value closer to that estimated by the regression analysis.

Discussion

Overview

The results of this study while being very interesting are still inconclusive. However, there are some observations to be made. One is that with this set of essays, essay length proved to be overwhelmingly the dominant predictor of essay ratings. Lexical diversity had a relatively minor role to play. However, there are likely to be situations where the role of essay length is unlikely to be so great. Higher level learners tend to write more and essays may be less likely easily discriminated by essay length. Also, non-timed tasks may show less dependence on essay length and more on lexical diversity.

Advanced Guiraud

Despite doubts about the accuracy of an automated Advanced Guiraud estimate, it seemed to perform relatively well as a complementary measure to essay length. It was able to clearly identify two highly rated essays and low rated essays.

One doubt about Advanced Guiraud concerned its relationship with essay length. Even though Advanced Guiraud had a significant correlation with essay length, it also had the highest partial correlation with essay ratings when controlled for essay length. This suggests that it is measuring another dimension of vocabulary use. Another doubt about this estimate of Advanced Guiraud centered on whether spelling errors conflated with advanced types would create a major effect. The clear identification of low rated essays suggests this may not have been a major problem. However, this weakness needs to be considered if used with other tasks involving other learners. It may be that in this case, the number of errors was relatively small compared with advanced types. In a situation where there are more spelling errors and/or fewer advanced types, this kind of estimate could be seriously compromised. A superior method of estimating advanced types may still be necessary.

Classifying essays

This experiment has looked at the relationship
between essay length, lexical diversity and essay ratings but it did not actually produce a mechanism for splitting the essays into groups according to quality. A simple split into three groups could be done by isolating essays at the top end of the scale to be part of an above average group and at the bottom end of the scale to be part of a below average group. This would leave essays in the middle range to form an average group. For example, because z scores have an additive property, z scores weighted for the contribution by each dimension to total variance could be used. The weighted z scores for quantity (essay length) and content (lexical diversity) could then be added and the scores ranked. A ranking was calculated using the z scores on the two dimensions weighted for contribution of total variance with essay length as the quantity dimension and Advanced Guiraud as the content dimension. These weighted z scores for each dimension were simply added together. Essays were arranged in descending weighted score order from left to right with the number representing the essay rating as follows:

11122313323243323343334343445

The three essays with the highest weighted z scores were three essays that were assessed as being good essays. The other essay rated good was ranked seventh best according to weighted z scores. At the bottom, the lowest ranked essay was rated poor and the next two lowest ranked essays were rated below average. The ranking seems to progress broadly in line with the ratings but there are some anomalies. An average rated essay ranks sixth best, a below average essay ranks fifteenth best while an above average essay ranks only eighteenth best. These anomalies suggest that this two-dimensional model may be too simple to account for essay quality in some cases.

**Linearity**

A comparison of graph data and correlation ratings suggests that linear methods such as correlation and regression may not always be the best for recognizing the power of some features. Although both TTR (100) and Advanced Guiraud showed similar correlations, Advanced Guiraud seemed better at isolating essays at the top and bottom of the scale. One problem of linearity is that sometimes more of a feature is not necessarily better. Often simple presence or absence is important. Another aspect of the limits of linearity is highlighted in a study by Jarvis, Grant, Bikowski & Ferris (2003). In a study involving clustering of essays according to various essay features, they found that highly rated essays had a variety of profiles. Although many features showed an overall positive correlation with essay ratings, an individual feature often had relations with other features that would override this general relationship in certain situations. They noted two such scenarios, complementarity and compensation. Complementarity refers to a situation where the high presence of one feature may often be accompanied by a low presence of another. For example, in one cluster of highly rated essays, the use of many nouns was accompanied by a relatively low occurrence of pronouns. In another cluster, the opposite pattern was noted. This was despite both features having an overall positive relationship with essay ratings. Compensation refers to a situation where learners who are not good at producing one feature make up for it in another way. In some cases, learners with low values for clausal embedding compensated for
this in other ways, for example, by writing longer essays with higher lexical diversity values. This may come as no surprise to teachers involved with student writing, yet standard statistical analyses have no way of recognizing this kind of relationship.

Conclusion
This experiment has focused on a simple two-dimensional quantity/content model for essay evaluation. The dimension of quantity was represented by essay length and content by various measures of lexical diversity. The two features together appeared to account for quality better than either feature by itself. For this set of essays, essay length was found to be the dominant dimension accounting for over 60% of the variance in essay ratings. The addition of a lexical diversity measure only accounted for a further 4% of variance. However, the relative dominance of the dimensions is likely to vary by task and learners. For some combinations of learner and task, it is likely that lexical diversity will account for quality better than essay length.

The number of word types in a fixed sample (TTR(100)), Advanced Guiraud, Yule’s K, the D estimate and the number of hapax legomena in a fixed sample (Hapax(100)) all accounted for a similar amount of the variance. Advanced Guiraud and Guiraud Index both showed a strong correlation with essay length which reflects their dependence on essay length. However, when essay length was controlled, Guiraud Index lost most of its association with quality while Advanced Guiraud remained strongly associated to essay quality. There was also a very strong association between the D estimate and Yule’s K which was much higher than between other pairs of lexical diversity measures which suggests that these two features may be measuring the same construct.

Graph evidence showed slightly different patterns to the correlation and regression evidence. Whereas the correlation evidence suggested that the measures of lexical diversity had a similar correlation with quality when complementing essay length, graphs showed that some features more clearly identified extreme high quality and low quality essays. In particular, Advanced Guiraud seemed to be the most effective complement to essay length for clearly identifying a small number of highly rated essays and low rated essays.

References


Appendix

Sample essays

Three sample essays are presented. Essay A is the longest, Essay B is the shortest and Essay C is of about average length.

Essay A (328 words)

I don’t think watching television is bad for children. It is true that there are some harmful TV program for children such as programs including sexual or violent expressions. There is an evidence that it causes bad affection for children. However there are a lot of good reasons for watching televisions. Firstly, there are educational programs which treat science, history and mathematics. They are usually broadcasted during day time. Why is it during day time? I guess it’s for the case when children is absent from their school, they can watch and learn something without going to school. Actually the programs are more interesting than a class of school. I used to watch such a program when I was absent from school. Secondly, there are some kinds of news program. For children it’s painful to read a news paper, because they can’t consentrate and there are so many words they don’t know due to a lack of experience of reading books. Therefore watching news on TV can be the important connection between children and world. Phewhaps some people says that news program is also difficult for children, but there is a TV program which teaches children very understandably such as ”kodomo-news (news for children)”. Thirdly, there are variety show and comedy show as a good reason. Some people believe that they are harmful, because in the shows there are only funny things. It’s true that watching too much or only the comedy shows makes children stupid, but it is not because of the program, but a lack of studying. Or rather, the comedy programs can be an important role that children can find a way to express themselves
widely, and make their characters. As a conclusion, I can understand people who say that TV program is harmful, because there are actually bad programs. However, if there no TV program except for bad one, TV industries must haven’t lasted. This is why TV program should be watched by children.

Essay B (50 words)
I disagree with that watching television is bad for children. Although there are many bad program for children, children can learn many things from television. For example, children can learn name of things and language. Also, we can look other country’s scenery and other beautiful things. It grows children heart.

Essay C (147 words)
I agree with the statement. My opinion is children should read good books, such as business books than watching television. This is because, I think, most TV problem is made for all people, from kids to elder people. It should be easier to make people understand. While, books is good materials to get new information. Usually, books are consisted by a lot of ideas of author. We can get those ideas author has gotten through their life by just paying only a few thousand yen. There may be information, which author doesn’t say on TV, in books. Author takes much more time to make their books. Even some people say that watching television is waste of time. I think so too, and children should keep time to watch less than 30 minutes a day. Because of these reasons, I support the statement "Watching television is bad for children."