Item Difficulty Analysis of English Vocabulary Questions

Yuni Susanti[1], Hitoshi Nishikawa[1], Takenobu Tokunaga[1] and Obari Hiroyuki[2]
[1]*Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan*
[2]*College of Economics, Aoyama Gakuin University, Tokyo, Japan*

Keywords: English Vocabulary Test, Item Difficulty, Multiple-choice Question.

Abstract: This study investigates the relations between several factors of question items in English vocabulary tests and the corresponding item difficulty. Designing the item difficulty of a test impacts the quality of the test itself. Our goal is suggesting a way to control the item difficulty of questions generated by computers. To achieve this goal we conducted correlation and regression analyses on several potential factors of question items and their item difficulty obtained through experiments. The analyses revealed that several item factors correlated with the item difficulty, and up to 59% of the item difficulty can be explained by a combination of item factors.

1 INTRODUCTION

English proficiency tests such as TOEFL® and TOEIC® are imperative in measuring English communication skills of non-native English speakers. Manual construction of questions for such tests, however, requires high-level skills, and is a hard and time-consuming task. Recent research has investigated how natural language processing (NLP) can contribute to automatically generating such questions, and more generally research on Computer-Assisted Language Testing (CALT) has received immense attention lately. Open-ended question asking for the “why”, “what” and “how” of something, and vocabulary questions are two of the most popular types of questions for evaluating English proficiency. Figure 1 shows an example of a TOEFL-like multiple-choice vocabulary question, asking an option with the closest meaning to the target word in the reading passage.

Automatic question generation for evaluating language proficiency is an emerging application since it has been made possible only recently with the availability of NLP technologies and resources such as word sense disambiguation (WSD) techniques (McCarthy, 2009) and WordNet (Fellbaum, 1998), a machine-readable lexical dictionary. To generate a question as in Figure 1, one needs to produce four components: (1) a target word, (2) a reading passage, (3) a correct answer and (4) distractors. Susanti et al. (2015) generated *closest-in-meaning* vocabulary questions employing Web news articles for the reading passage and WordNet for the correct answer. The distractors or incorrect options were generated by using both the retrieved reading passage and WordNet lexical dictionary. Brown et al. (2005) generated multiple-choice questions by taking their components from WordNet, including the reading passage from the example sentences in the dictionary for their *cloze questions* (fill-in-the-blank questions). Lin et al. (2007) also adopted WordNet to produce English adjective questions from a given text. The candidates of options (a correct answer and distractors) were taken from WordNet and filtered by Web searching.

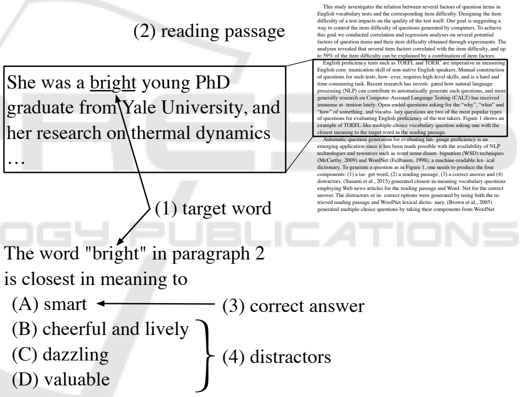

(2) reading passage
She was a bright young PhD graduate from Yale University, and her research on thermal dynamics …
(1) target word
The word "bright" in paragraph 2 is closest in meaning to
(A) smart
(3) correct answer
(B) cheerful and lively
(C) dazzling
(D) valuable
(4) distractors

Figure 1: Four components in a multiple-choice question asking for closest-in-meaning of a word.

In the broader area of vocabulary question, many studies have been done, e.g. generation of cloze questions for completing a sentence, word collocation, synonym, antonym, etc. Vocabulary questions have been generated to evaluate test takers’ knowledge of

Susanti, Y., Nishikawa, H., Tokunaga, T. and Hiroyuki, O.
Item Difficulty Analysis of English Vocabulary Questions.
In *Proceedings of the 8th International Conference on Computer Supported Education (CSEDU 2016) - Volume 1*, pages 267-274
ISBN: 978-989-758-179-3

English in correctly using verbs (Sakaguchi et al., 2013), prepositions (Lee and Seneff, 2007) and adjectives (Lin et al., 2007). Concerning their target languages, many attempts have focused on the English language.

The CALT research focuses mainly on question item generation, scoring, providing feedback to test takers and the like; yet research on test design, especially concerning the difficulty of question items is rather rare. The difficulty of question items in a test greatly impact the difficulty of the whole test. According to Bachman (1990), a too easy or too difficult test for a particular group generally results in a limited range of scores, or very little variance. For example, a test with all items at the same level of difficulty would not be a very accurate measure for individuals whose abilities are either greatly above or below that level, i.e. the test has low discrimination ability. A test that contains too many easy or too many difficult question items would result in a skewed score distribution. Therefore only when the difficulty of each question item in a test is set at an appropriate level, can the test scores be informative. That being the case, controlling the difficulty of each item is the first important step for designing a high quality test. Thus the present study focuses on the analysis of factors of question items affecting the item difficulty.

Studies of factors that affect question item difficulty are relatively few. Brown (1989) conducted an experiment on non-native speakers of English to measure the correlation between various linguistic features and item difficulty of cloze questions and identified that word classes, local word frequency, and readability measures correlated with the item difficulty. Sigott (1995) examined word frequency, word classes, and constituent types of the gap for the C-test[1] and found high correlation only with the word frequency. Beinborn et al. (2014) introduced a model predicting gap difficulty of the C-test and they found that combinations of macro and micro-level cues such as spelling, phonetic difficulties, and text complexity contributed to the gap difficulty.

The present study investigates factors affecting question item difficulty toward controlling the item difficulty of machine-generated questions. Unlike most past studies that dealt with cloze questions, we work on vocabulary-type questions asking for *closest-in-meaning* of an English word as shown in Figure 1, which is different from cloze questions in that it is necessary to generate a correct answer, a synonym of the target word[2]. Our ultimate goal is to develop a method of automatically generating vocabulary-type questions with the ability of controlling item difficulty. Toward this goal, this study explores factors that influence the item difficulty of vocabulary questions, and on the basis of the findings, to suggest possible ways to control the item difficulty in machine-generated questions.

We start with introducing potential factors affecting item difficulty (section 2), then explain the design of the experiments for data collection (section 3), followed by an analysis of the impact of each potential factor on item difficulty (section 4). Finally we conclude the paper and look at future directions (section 5).

2 POTENTIAL FACTORS AFFECTING ITEM DIFFICULTY

Considering that automatic question generation requires generation of the four question components as shown in Figure 1, it is natural to investigate the relations of the difficulty of each component and the overall question item. Having understood the relations, we might be able to control the item difficulty by controlling the difficulty of the dominant components. In the following sections, potential factors affecting the difficulty of each component are described.

2.1 Target Word (TW)

The first factor to be considered is the target word, which is the word being asked in the vocabulary question. It is natural to assume that item difficulty is, to a certain degree, related to the difficulty level of the target word. There are a number of studies on determining the difficulty level of an English word (or reading difficulty), and they are based on various word features such as *word frequency* (frequency of occurrence of the word in certain corpora) and *word length* (the character length of the word) (Heilman et al., 2008; Petersen and Ostendorf, 2009). Medero and Ostendorf (2009) compared articles in standard and simple English Wikipedia and found that words that appear in standard but not simple English tend to have shorter definitions, fewer part-of-speech types, word senses, and so on.

JACET 8000 (Uemura and Ishikawa, 2004) is a radically new word list designed for Japanese English

[1] C-test involves a piece of text from which a number of words have been removed.

[2] Note that a correct answer comes from the original passage in cloze questions.

learners. JACET 8000 ranks the word list based on the word frequency in the British National Corpus supplemented with six million tokens of texts targeted at the needs of Japanese students. The 8,000 words in the list are divided into eight groups of 1,000 words based on their word difficulty level.

Throughout this study, we use the JACET 8000 level system to assign a word difficulty level to words in a question item, as participants of our experiments are all Japanese university students. JACET 8000 uses the 1–8 levelling system in which level 1 is the easiest word. A special level *Other* or *O* is defined for words over level 8, which include non-English or misspelling words. The JACET 8000 difficulty level of the target word (TW_J8) is the first factor to be examined.

Another factor considered is the number of word senses of the target word (TW_WS). An ambiguous word (many word senses) tends to be difficult because its correct word sense in a given context should be identified before answering the vocabulary question.

2.2 Reading Passage (RP)

The difficulty of a reading passage might influence the item difficulty since test takers need to understand the context to answer the question. It is safe to assume that a reading passage composed of simple and easy words would be easier to understand than a passage with lots of difficult words.

We suspect, however, when working on vocabulary questions, test takers might not read the whole reading passage, but often only neighbouring portions around the target word. Hence we consider the average difficulty level of the words appearing in the sentence containing the target word as the difficulty of the whole reading passage. This reading passage difficulty is one of the potential factors affecting the item difficulty (RP_J8_1s_ave). For example, given a target word "authentic" in the sentence "*The journalist sent an authentic report on poverty in Africa.*", after removing the stopwords[3] the average of the difficulty level of "journalist", "sent", "report", and "poverty" is calculated to define the difficulty level of the whole reading passage. In addition to this one-sentence-average, we also calculate an average over narrower neighbouring words, i.e. the adjacent two words on both sides of the target word (RP_J8_2w_ave). In the example above, they would be "sent" and "report". When the target word appears at the beginning or the end of the sentence, the two following or preceding words of the target word are taken as the adjacent words. Instead of an average, we can consider different factors by taking the maximum difficulty level among the words in question for both RP_J8_1s_ave and RP_J8_2w_ave. We name them RP_J8_1s_max and RP_J8_2w_max respectively.

Table 1: Potential factors of question items.

Factor	Description
TW_J8	the difficulty level of the target word
TW_WS	the number of word senses of the target word
RP_J8_1s_ave	the average difficulty level of words in a sentence containing the target word in the reading passage
RP_J8_1s_max	the maximum difficulty level of words in a sentence containing the target word in the reading passage
RP_J8_2w_ave	the average difficulty level of two adjacent words of the target word in the reading passage
RP_J8_2w_max	the maximum difficulty level of two adjacent words of the target word in the reading passage
CA_J8_ave	the average difficulty level of words comprising the correct answer
CA_J8_max	the maximum difficulty level of words comprising the correct answer
DS_J8_ave	the average difficulty level of the distractors
DS_J8_max	the maximum difficulty level of the distractors

2.3 Correct Answer (CA)

The correct answer here is the option with the closest meaning to the target word used in the reading passage. The difficulty of the correct answer also has a possibility of affecting the question item difficulty. Since the correct answer can be composed of more than one word (multiple-word correct answer), we average the difficulty level of the words comprising the multiple-word correct answer (CA_J8_ave). Similar to the reading passage difficulty level, we consider the maximum difficulty level among words comprising the correct answer (CA_J8_max) as well.

2.4 Distractors (DS)

Distractors are the incorrect (or less correct) options in a question. There are three distractors for a single question item used in our experiments.

The factor of distractors to be examined is their difficulty level. Since we have three distractors, and each of them can be composed of more than one word (multiple-word distractor), we average those difficulty levels to obtain the difficulty level of the distractors (DS_J8_ave). Another possible factor is the maximum difficulty level among those distractor-composing words instead of their average

[3] Words bearing less information such as function words.

Table 2: Configuration of evaluation sets (Exp. 1).

Eval.	Contents		Test
set	HQs	MQs	taker
A1	TW#01–13	TW#14–25	C_A
B1	TW#14–25	TW#01–13	C_B
A2	TW#26–37	TW#38–50	C_A
B2	TW#38–50	TW#26–37	C_B

(DS_J8_max). Table 1 summarises the potential factors introduced in this section.

3 EXPERIMENTAL DESIGN

Two experiments were conducted to collect item difficulty data of the vocabulary questions. We used two kinds of materials (question sets) in the experiments: machine-generated questions (MQs) created by an automatic question generation method (Susanti et al., 2015), and human-generated questions (HQs) taken from the real TOEFL iBT® tests and preparation books. The aim of utilising two kinds of questions is to see if there is any difference between HQs and MQs in terms of their difficulty. Fifty target words were compiled from TOEFL® sample questions[4] and official preparation books (ETS, 2007), and other preparation books (Sharpe, 2006; Phillips, 2006; Gear and Gear, 2006). The target sites for retrieving reading passages for the MQs were the NY Times[5], CNN[6] and Science Daily[7] websites.

Two kinds of experiments were conducted; from each of them, a different kind of item difficulty was induced: one is based on the achievement of the test takers who answered the questions, and other is based on the subjective evaluation of the questions by English teachers. They provide different views of the same entities, question items, thus we can compare their difficulty from different perspectives.

3.1 Experiment 1: Student-based Item Difficulty (ID_S)

We prepared two kinds of question item datasets: 50 HQs and 50 MQs. The target words of these two datasets are the same. Given a certain target word, however, other components of the question item would be different across the datasets since one was human-made and the other was machine-made.

From these two question item datasets, we created four evaluation sets (A1, B1, A2 and B2) by mixing HQs and MQs as shown in Table 2. The target words (TW#01-13) of 13 HQs in Set A1 and that of 13 MQs in Set B1 were identical, and so did for the others. The order of target words in the evaluation sets was randomised and the same between sets A1 and B1, and between sets A2 and B2.

We recruited 79 Japanese university undergraduate students (46 first year, 20 third year and 13 fourth year students) and randomly divided them into two classes C_A (40 students) and C_B (39 students) with keeping closer distribution of student years across classes. The ratio between male and female students was roughly 2:1. We assigned the evaluation set A1 and A2 to the class C_A, and B1 and B2 to the class C_B. Thus the students of different classes answered different question items (HQs and MQs) for the same 50 target words. The time taken for completing each evaluation set was roughly 20 minutes, and there was one week interval between answering the set A1/B1 and set A2/B2.

Based on the student responses, we calculated the difficulty index for each item. The difficulty index P is the proportion of students who correctly answered a question item (Brown, 2012). The range of P spans between 0 and 1, and the lower the value, the more difficult an item is. We induce the student-based item difficulty ID_S from P by inverting the scale with equation (1). Thus, a greater ID_S indicates more difficult item.

$$ID_S = 1 - P \qquad (1)$$

3.2 Experiment 2: Teacher-based Item Difficulty (ID_T)

We asked 8 English teachers (non-native English speakers: 4 Japanese and 4 Filipinos) to judge the item difficulty of each question item on a scale 1–5, with 5 being the most difficult. We used the same question items as the experiment 1, but only half of them (set B1 and B2). The order of question items in a set was kept as the same as in the experiment 1. In total we had 25 HQs and 25 MQs to be evaluated by each teacher in this experiment. The teacher-based item difficulty ID_T was calculated by averaging the teachers' responses and then normalising it into the range between 0 and 1.

[4] www.ets.org

[5] www.nytimes.com

[6] www.cnn.com

[7] www.sciencedaily.com

Table 3: Statistics of item difficulties.

	ID_S		ID_T	
	HQs	MQs	HQs	MQs
n	50	50	25	25
$\overline{x}$	.47	.49	.55	.57
sd	.23	.20	.19	.18
max	1	.82	.88	.84
min	.08	.10	.19	.16

4 ANALYSIS OF RELATIONS BETWEEN ITEM DIFFICULTY AND POTENTIAL FACTORS

This section describes the analysis of the relations between the following two kinds of item difficulties and potential factors of question items summarised in Table 1.

ID_S : item difficulty from students' perspective, is calculated by one minus the proportion of students who correctly answered the question item as in equation (1).

ID_T : item difficulty from teachers' perspective, is calculated by averaging the teachers' difficulty judgements of the question item and then normalising into the range between 0 and 1.

Table 3 shows the descriptive statistics of the item difficulties, including the number of question items (n), the mean ($\overline{x}$), the standard deviation (sd), as well as the maximum (max) and minimum (min) values. The overall values are very similar between HQs and MQs, and between ID_S and ID_T. The means are close to 0.5 and the maximum and minimum values stretch out to almost both extremes: .08 and 1. As far as looking at these numbers, our test sets are not so skewed and favourable for investigating the relations between item difficulties and various factors of question items. The following sections describe the correlation and regression analyses performed between the item difficulties and the potential factors. All numbers were calculated using R[8] (version 3.2.1).

4.1 Correlation Analysis

The Pearson correlation coefficient was calculated between ID_S and ID_T to see to what extent both item difficulties from different perspectives correlated to each other. This resulted in positive correlation with .69 of correlation coefficient for HQs and .56 for MQs (p-value < 0.05). We can conclude that there is no big difference between the item difficulty of the student viewpoint and that of the teacher viewpoint.

Table 4: Pearson correlation coefficients for HQs and MQs.

Factor	ID_S		ID_T	
	HQs	MQs	HQs	MQs
`TW_J8`	.26	.22	**.66**	**.30**
	(.063)	(.12)	(.0003)	(.14)
`TW_WS`	.27	−.11	.03	−.17
	(.063)	(.45)	(.90)	(.42)
`RP_J8_1s_ave`	−.10	.26	.15	.14
	(.49)	(.07)	(.48)	(.48)
`RP_J8_1s_max`	−.04	.16	**.35**	**.44**
	(.79)	(.26)	(.086)	(.03)
`RP_J8_2w_ave`	.01	.23	.17	.19
	(.92)	(.10)	(.42)	(.36)
`RP_J8_2w_max`	.05	.18	.17	.19
	(.71)	(.21)	(.42)	(.37)
`CA_J8_ave`	**.38**	.19	**.37**	**.48**
	(.006)	(.19)	(.071)	(.01)
`CA_J8_max`	**.38**	.18	**.37**	**.44**
	(.006)	(.20)	(.068)	(.02)
`DS_J8_ave`	−.10	**.54**	.15	**.52**
	(.50)	$(4\cdot10^{-5})$	(.46)	(.008)
`DS_J8_max`	−.004	**.48**	.16	**.52**
	(.97)	(.0004)	(.47)	(.008)

Boldface indicates numbers more than or equal to .3.
P-values are enclosed with parentheses under correlation coefficients.

Table 4 shows the Pearson correlation coefficients between each of the potential factors and one of item difficulties (ID_S and ID_T) with each p-value in the underneath parentheses. Comparing the effective factors between ID_S and ID_T, the ID_T columns have high-correlation factors from all question components (TW, RP, CA and DS), but the ID_S columns do not. Only the difficulty level of correct answers and distractors (`CA_J8_*` and `DS_J8_*`) show salient correlation for ID_S. This means that ID_S is more difficult than ID_T to be characterised in terms of the potential factors under consideration. This is probably because the evaluation by the English teachers is more consistent and they refer to all components of a question item for difficulty judgement. On the other hand, each student has their own strategy for answering question items, thus the components they cared about would tend to be diverse over both individuals and question items.

Another interesting observation is the difference between HQs and MQs in distractor correlation (`DS_J8_*`) with both ID_S and ID_T. The item difficulties more highly correlated with the difficulty level of distractors in MQs than in HQs. This difference suggests that composing distractors would be a key to control item difficulty in automatically generating question items. In contrast, the difficulty level of correct answers behaves differently in ID_S. The factor

[8] https://www.r-project.org

Table 5: Results of multiple regression.

No.	Dependent var.	Independent variables	R^2	adjusted R^2
1	ID_S (HQs)	CA_J8_ave	.15	.13
2	ID_S (HQs)	CA_J8_ave + TW_WS	.24	.21
3	ID_S (HQs)	CA_J8_ave + TW_WS + TW_J8	**.41**	**.38**
4	ID_S (MQs)	DS_J8_ave	.30	.28
5	ID_S (MQs)	DS_J8_ave + RP_J8_1s_ave	.32	.29
6	ID_S (MQs)	DS_J8_ave + RP_J8_1s_ave + TW_J8	**.35**	**.31**
7	ID_T (HQs)	TW_J8	.43	.41
8	ID_T (HQs)	TW_J8 + RP_J8_1s_max	.60	.57
9	ID_T (HQs)	TW_J8 + RP_J8_1s_max +CA_J8_ave	**.64**	**.59**
10	ID_T (MQs)	DS_J8_ave	.27	.24
11	ID_T (MQs)	DS_J8_ave + CA_J8_ave	.43	.39
12	ID_T (MQs)	DS_J8_ave + CA_J8_ave + TW_J8	.50	.43
13	ID_T (MQs)	DS_J8_ave + CA_J8_ave + TW_J8 + RP_J8_1s_max	**.59**	**.50**

Boldface indicates maximum numbers in the section.

CA_J8_* shows high correlation with ID_S on HQs but not on MQs. This means that there is quite a lot of room for improvement in composing correct answers in the automatic question generation method adopted in the present study.

Surprisingly, the target word factors (TW_*) do not necessarily have a great impact on item difficulty. The only exception is the difficulty level of the target word (TW_J8) against ID_T, the item difficulty by teacher evaluation. The number of word senses (TW_WS) particularly does not correlate quite well with both item difficulties. One possible explanation is that the target words used in the question items were likely the ones with the most common meanings. Therefore, even if the target word has many senses to be ambiguous, it might not really matter. As a matter of fact, the method we adopted in this study for generating question items tries to use more common word senses for generating question items (Susanti et al., 2015).

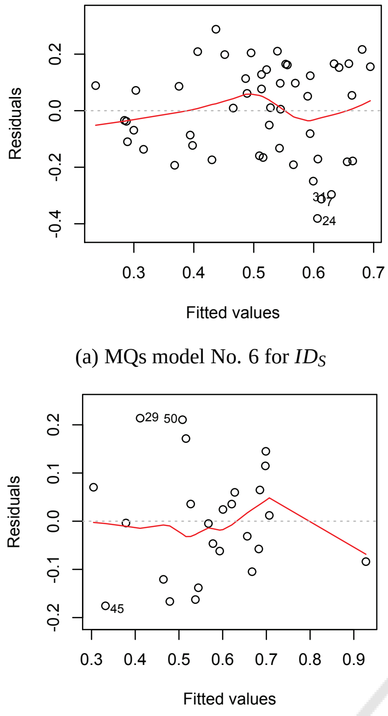
4.2 Regression Analysis

The results of the correlation analysis in Table 4 lead us to investigate the degree to which various combinations of these potential factors could explain item difficulties. Several mixtures of the potential factors were analysed using regression analysis to determine which combination best predicts the item difficulties.

We added factors one by one to a set of independent variables starting from more highly correlated factors for each item difficulty until gaining no improvement in terms of the coefficient of determination (R-squared, R^2). Only a single factor concerning the word difficulty level was employed from each item component, e.g. either CA_J8_ave or CA_J8_max is adopted as a factor from the correct answer component of question items. Promising sets of independent variables are shown in Table 5, with R-squared (R^2) and adjusted R-squared which takes into account the number of independent variables. All coefficients of the factors in Table 5 were positive and all these regressions were significant at p-value < 0.05.

Table 5 shows a tendency where adding factors generally improves the R-squared values. That means the factors used as independent variables here complementarily contribute to the item difficulty. It is also observed that ID_T is better fitted than ID_S, i.e. almost 60% of the ID_T data can be explained by the best models, while at most 41% of the ID_S data can. As we have discussed in the correlation analysis concerning Table 4, the English teachers presumably look at all components of question items for evaluation, while the students look only at the minimum necessary components for answering each question. For instance, if students know the meaning of the target word, they might not care about the reading passage; they would rather directly move to the question options and look for the correct answer. In contrast, if the students do not have any idea about the target word, they would read the passage, or try to investigate the distractors one by one, or even choose a choice randomly. Thus, the difficulty level of a reading passage is important for students depending on if they know the word or not. Hence, what makes a question item difficult for each student is different, depending on their knowledge.

Another thing to be investigated is the *outlier* of the results. An outlier is a point that lies outside the overall distribution pattern, or is far from the fitting line. Analysing outliers might give some insights for creating better question items for both human and ma-



(a) MQs model No. 6 for ID_S

(b) MQs model No. 13 for ID_T

Figure 2: Scatter plot of residuals.

chine. Residual scatter plots depicting three outliers[9] for the best regression models of MQs (No. 6 for ID_S and No. 13 for ID_T) are shown in Figure 2.

Three outlier items are shown as numbered points in each figure. Taking Figure 2 (a) as an example, the items 7, 24, and 31 are revealed to be outliers for the ID_S regression analysis on the MQs data. The target word of the item 7 is “serve”, which is an easy word (TW_J8 = 0) but has lots of word senses (TW_WS = 16), hence the question became difficult for the students. Since the regression model (No. 6) happens not to include the factor on the number of word senses, this item was predicted to be easy.

Items 29, 45, and 50 are revealed to be the outliers in the ID_T regression analysis on the MQs data as shown in Figure 2 (b). Item 29 has “step” as its target word, and this word belongs to the easiest level (TW_J8 = 0). However, according to ID_T, it is considered to be a difficult question. It could be due to the fact that “step” has many word senses (TW_WS = 21), and the word senses used in the question item happened to be not the most common word sense. Therefore, it would be considered as a difficult item by the teachers. Since the regression model (No. 13) does not include the number of word senses in the independent variables, the model predicted this item to be easy. These two outliers taken as examples here accused the model for not including the number of word senses to make them outliers. However, introducing the number of word senses into the model might harm the prediction for other items, since the correlation analysis (Table 4) has shown that the number of word senses does not correlate with both item difficulties.

5 CONCLUSION AND FUTURE WORK

Targeting English vocabulary questions, the present study investigated the relations between potential factors of each component of a question item and its item difficulty. Aiming at controlling item difficulty of automatically generated questions, we conducted the correlation and regression analyses on several potential factors of question items and their item difficulty obtained through experiments. Two kinds of question items were utilised: machine generated (MQs) and human generated (HQs) questions. Two kinds of experiments were conducted for obtaining item difficulties from different perspectives: from test takers (ID_S) and from English teachers as a human expert (ID_T). The correlation analysis revealed the following tendencies.

- The two item difficulties from different perspectives (ID_S and ID_T) correlated quite well with coefficient .56 $\sim$ .69. The ID_T item difficulty has high-correlation factors from all question components (TW, RP, CA and DS), while ID_S does not. It means that ID_S is more difficult than ID_T to be characterised in terms of the potential factors considered in this study. This would be due to the difference of focal points in their task between test takers and teachers. The teachers tend to refer to all components of question items for evaluation, while the test takers only look at the necessary components for answering the question.
- The difficulty level of distractors correlated higher with the item difficulties in MQs than those in HQs. This result suggests that composing distractors would be an important key to control item difficulty in automatically generating question items.
- The number of word senses does not correlate quite well with both item difficulties. This would be explained by the fact that most of the question items adopted the most common meaning of the target word in the reading passage context. We need to take into account familiarity or usability of each word sense as well as their numbers.

[9] These three are provided by applying the lm procedure of the R software.

The results of regression analysis indicates that even the best combination of factors for predicting item difficulty is only able to predict about 59% of the data (HQs model for ID_T). There is still 41% of data to be explained. There are many other factors affecting item difficulty that have not yet been investigated in this study, and they are left to future investigation.

The subjects of the experiment in this study were rather homogeneous; they were all Japanese students. The human-experts who evaluated the question items were also all non-native speakers of English, despite them being English teachers. When investigating what causes a question item to be difficult or easy, conducting experiment on subjects with different backgrounds might provide different useful insight.

REFERENCES

Bachman, L. F. (1990). *Fundamental Consideration in Language Testing*. Oxford University Press.

Beinborn, L., Zesch, T., and Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. In *Transactions of the Association for Computational Linguistics*, volume 2, pages 517–529. Association for Computational Linguistics.

Brown, J. C., Frishkoff, G. A., and Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 819–826.

Brown, J. D. (1989). Cloze item difficulty. In *Japan Association for Language Teaching Journal*, volume 11, No.1, pages 46–67. JALT.

Brown, J. D. (2012). Classical test theory. In Fulcher, G. and Davidson, F., editors, *The Routledge Handbook of Language Testing*, chapter 22, pages 323–335. Routledge.

ETS (2007). *The Official Guide to the New TOEFL iBT Internation edition*. Mc Graw-Hill.

Fellbaum, C. (1998). *WordNet: A lexical database for English*. A Bradford Book.

Gear, J. and Gear, R. (2006). *Cambridge Preparation for the TOEFL Test 4th Edition*. Cambridge University Press;.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, EANL '08, pages 71–79, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lee, J. and Seneff, S. (2007). Automatic generation of cloze items for prepositions. In *Proceedings of Interspeech 2007*, pages 2173–2176.

Lin, Y.-C., Sung, L.-C., and Chen, M. C. (2007). An automatic multiple-choice question generation scheme for English adjective understanding. In *Proceedings of Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, the 15th International Conference on Computers in Education (ICCE 2007)*, pages 137–142.

McCarthy, D. (2009). Word sense disambiguation: An overview. *Language and Linguistics Compass*, 3(2):537–558.

Medero, J. and Ostendorf, M. (2009). Analysis of vocabulary difficulty using wiktionary. In *Proceedings of the Speech and Language Technology in Education Workshop (SLaTE)*.

Petersen, S. E. and Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Comput. Speech Lang.*, 23(1):89–106.

Phillips, D. (2006). *Longman Preparation Course for the TOEFL Test: iBT*. Pearson Education Inc.

Sakaguchi, K., Arase, Y., and Komachi, M. (2013). Discriminative approach to fill-in-the-blank quiz generation for language learners. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistic*, pages 238–242. Association for Computational Linguistic.

Sharpe, P. J. (2006). *Barron's TOEFL iBT Internet-Based Test 2006-2007 12th Edition with CD-ROM*. Barron's Educational Series Inc.

Sigott, G. (1995). The c-test: some factors of difficulty. In *AAA: Arbeiten aus Anglistik und Amerikanistik, 20(1)*, volume 20(1), pages 43–53. Narr Francke Attempto Verlag GmbH Co. KG.

Susanti, Y., Iida, R., and Tokunaga, T. (2015). Automatic generation of english vocabulary tests. In *Proceedings of the 7th International Conference on Computer Supported Education*, pages 77–87.

Uemura, T. and Ishikawa, S. (2004). JACET 8000 and asia TEFL vocabulary initiative. In *Journal of ASIA TEFL*, volume 1(1), pages 333–347. ASIA TEFL).