Journal of Knowledge Information Technology and Systems(JKITS), Vol. 10, No. 2, pp. 163~171, April 2015

Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

Development of an Algorithm for Testing Vocabulary-Based Reading-Level Suitability of English Book Readers

Hangyeol Song[1], Eunwoo Joung[1], Sukyeong Cho[1], Yong Ho Choi[2], Hyunbean Yi[*1]

[1]*Department of Computer Engineering, Hanbat National University*
[2]*Research Institute, Ubiquitous Core System*

ABSTRACT

It is very important for a book reader to choose an appropriate book which is suitable for the reader's reading-level for a successful reading. Especially, when a reader choose a foreign book, more attention needs to be paid because the book's readability-level and the reader's reading-level have to be considered simultaneously. In case of English books that are sold or published in Korea, somewhat objective readability levels of them can be estimated to some degree by referring to the levels of difficulty which are developed by institutes in the U.S.A. or Korea. However, because reader's reading-levels can be estimated through a long English test, many readers avoid taking such a test and subjectively estimate their reading-level. They are therefore likely to choose an English book which are not suitable for their reading-level. This paper, for Korean readers whose mother tongue is not English, proposes an algorithm which can help a reader find out his/her reading-level suitability for an English book by a short time English vocabulary test. We present our systematic process for extracting test words from a target English book and for estimating the level suitability between the book and the reader. In order to evaluate our algorithm, we implement it in a Web application and show survey results.

KEYWORDS: Readability, eBooks, Reading level, Suitability, Vocabulary tests

ARTICLE INFO: Received 23 February 2015, Revised 10 April 2015, Accepted 10 April 2015.

[*]Corresponding author is with the Department of computer Engineering, Hanbat National University, 125, Dongseo-daero, Yuseong-gu, Daejeon, 305-719, KOREA.

E-mail address: bean@hanbat.ac.kr

1. 서 론

많은 정보가 영어로 전달되는 현대 글로벌 시대에 영어의 필요성은 갈수록 커지고 있고 그에 따라 영어 실력 향상을 위한 다양한 교육 방법이 등장하고 있다. 하지만, 글쓰기, 듣기, 말하기등 전반적인 영역의 실력 향상을 위해서는 무엇보다도 읽기가 기본이 된다. 따라서, 정부의 영어교육 강화 추세에 맞추어 국공립, 민간 영어 도서관이 늘어나고 있으며, 그에 따라 영어 도서 시장이 큰 신장세를 보이고 있다[1].

성공적인 독서를 위하여 도서 선정 시 고려해야 할 주요한 사항 두 가지는 “흥미”와 “수준”이다. 영어 실력 향상을 위한 영어 도서 선정 시에도 역시 이 두 가지 사항을 고려해야 한다[2]. “흥미”는 광고, 제목, 도서의 구성, 독자의 관심사등에 의해서 매우 주관적으로 쉽게 결정되지만 “수준”은 영어 도서의 수준과 독자의 수준을 동시에 고려해야 하기 때문에 상호간의 적합성을 쉽게 파악하기 어렵다. 영어 도서의 수준은 학력별, 나이별, 특정 지수로 제시되기도 하지만, 대부분의 독자들은 자신의 객관적인 수준을 정확히 모르고 있거나 객관적인 수준을 파악하기 위해서는 장시간의 테스트를 거쳐야 한다. 또한, 어느 정도 객관적인 수준을 알고 있다하더라도 분야에 따라서 사용하는 어휘가 다르기 때문에 시간을 들여 도서의 여러 곳을 훑어보기 전에는 그 도서가 독자가 읽기에 얼마나 적합한지 제대로 파악하기 어렵다. 게다가, 인터넷의 발달로 대다수의 사람들은 도서 선정 시 서점에 직접 방문하기 보다는 인터넷 서점을 사용한다. 인터넷 서점의 등장으로 도서 구매가 편리해지긴 했지만 도서를 직접 훑어 볼 수 없기 때문에 영어 도서 선정 시 특정 도서에 대한 독자의 수준 적정성을 파악하기 어렵다. 지금까지 도서의 가독성 분석이나 독자의 읽기 능력을 평가하기 위한 지수 개발에 관한 연구는 있었으나 특정 영문 도서에 대한 독자의 적합성 파악에 관한 연구는 없었다. 따라서, 영어 도서의 수준과 그 도서에 대한 독자의 수준 적합성을 복잡한 절차 없이 빠르게 제시 해 줄 수 있는 시스템이 제공된다면 영어 도서 선정의 실패율을 낮추어 줄 수 있을 것이다.

본 논문에서는 영어가 모국어가 아닌 한국인이 영어 도서 선정 시 쉽고 빠르게 영어 도서에 대한 독자 자신의 읽기 수준 적합성을 판별할 수 있는 방법을 제시한다. 2장에서 기존 영어 도서 수준 평가 방법 및 관련 연구를 살펴보고, 3장에서 본 논문에서 제안하는 어휘기반 영어 도서 읽기 수준 판단 알고리즘을 설명한다. 제안한 알고리즘의 신뢰성을 파악하기 위하여 4장에서 구현 및 설문을 통한 평가 결과를 제시하고, 5장에서 결론을 맺는다.

2. 기존 방법 및 연구

미국의 MetaMetrics사에서 미국 학생들의 독서능력 신장을 위해 렉사일 독자 지수(Lexile reader measure)와 렉사일 문서 지수(Lexile text measure)를 개발하였다[3]. 렉사일 독자 지수는 몇 가지 영역의 테스트를 통하여 독자의 읽기 능력을 측정한 것이고, 렉사일 문서 지수는 도서의 난이도를 제시한다. 하지만, 렉사일 지수는 영어가 모국어인 미국 학생들을 위한 지수로, 영어가 모국어가 아닌, 즉, English as Foreign Language(EFL) 환경인 나라의 사람들에게 적용하기에는 무리가 있다[4]. 미국의 렉사일 지수와 비슷하게 한국 독자의 한글 도서에 대한 지수로써 ㈜낱말에서 개발한 한국 독자의 한글 도서에 대한 READ(Reading Environment & Ability Degree) 지수가 있다[5]. 리드 지수는 개인의 독서에 필요한 제반 능력을 측정하는 도구로 READ 검사와 READ 지수 도서로 구성된 독서 프로그램이다. 즉, 도서의 난이도를 과학적·객관적

으로 측정한 후 숫자로 표시하여 개인의 독서능력에 맞는 도서를 선정하여 읽을 수 있도록 도와주는 프로그램이다. 텍스트만을 고려하느냐, 텍스트뿐만 아니라 사용빈도, 문장의 길이, 어절의 수, 글자의 크기 등과 같은 기타 이독성 요소들을 고려하였다. 두 가지 방법 모두 독자의 능력을 수치화하기 위하여 어느 정도 시간을 들여 테스트를 거쳐야 하고, 일반적인 지수이기 때문에 특정 도서와 독자 간의 수준 적정성을 파악하기에는 부족하다.

Chall과 Dale은 도서의 난이도 구별 방법을 일반인들의 설문을 통하여 조사하였다[6]. 결과적으로 압도적인 숫자의 사람들이 도서의 난이도 기준의 첫 번째 요소로 어휘를 꼽았다. 또한, 이독성 측정에서 어휘 요소가 문장 요소보다 더 강력한 요인으로 작용하며 상급수준의 독해 수준에서는 학습자의 선행지식 및 문화 지식이 주요한 요인이 된다. 하지만, 그 이하의 수준에서는 단어의 난이도 및 문장의 복합성이 주요한 요인이 된다[7].

본 논문에서도 어휘 분석 및 선정에 초점을 둔다. 적절한 어휘 선정 및 독자의 짧은 테스트를 통하여 간단하고 신속하게 영어 도서에 대한 독자의 읽기 수준 적합성 판단하는 알고리즘을 제시한다.

3. 어휘기반 영어 도서 읽기 수준 적합성 판단 알고리즘

특정 영어 도서에 대한 독자의 읽기 적정성을 어휘를 기반으로 판단하는 방법으로써 본 논문에서는 어휘 테스트 방법을 사용한다. 만일, 독자가 인터넷 서점을 통하여 자신의 수준에 적절한 영어 도서를 선정하고자 할 때, 그 도서에 등장하는 주요 어휘들을 살펴볼 수 있다면 어느 정도 적정성을 파악할 수 있을 것이다. 하지만, 짧은 시간에 파악할 수 있도록 하기 위해서 가능한 한 적은 수의 어휘로 그 도서의 난이도와 내용을 충분히 드러낼 수 있는 어휘 선정 전략이 필요하다.

본 논문에서 제안하는 어휘기반 영어 도서 읽기 수준 판단 알고리즘은 크게 1) 대상 영어 도서에서 독자의 검토를 위해 제시할 어휘 선정 및 레벨(난이도) 분류 방법과 2) 독자의 검토 결과를 바탕으로 읽기 수준 적합성 정량화 방법으로 나뉜다.

3.1 어휘 선정 및 레벨 분류

3.1.1 교과과정 기반 어휘 선정 및 레벨 분류

어휘 수집 및 선정을 위하여 다음 세 어휘 데이터를 참조한다.

A) 한국 교육과학기술부 2009 개정교육과정 영어과 교육과정의 초·중등 필수 어휘[8],
B) 한국 중·고등학생 교과서에 수록된 어휘[9],
C) 옥스퍼드 선정한 자주 사용되는 어휘 3,000단어[10].

A는 초·중등 추천 어휘를 여러 학년으로 묶어 대략적으로 수준을 분류하고 있으며 C는 어휘만 제시할 뿐 레벨 분류는 제시하고 있지 않다. B는 어휘별 레벨을 제시하기 보다는 어떤 어휘가 중·고등학교의 어느 학년, 어떤 출판사의 교과서에 등장하는지 검색할 수 있는 기능을 제공한다. 따라서, 본 논문에서는, 레벨 분류를 위하여 A와 C에 제시된 어휘 총 6306개 중 중복되는 어휘를 제거한 4272개의 어휘에 대해 B의 검색과정과 A와 C간 포함관계를 분석하여 <표 1>과 같이 분류한다. 기본적으로는 교육과정 단계 정보에 따르되, A, B, C에 공통으로 등장하는 어휘일수록 레벨이 낮게 분류되고, 레벨 11은 난이도가 가장 높은 그룹으로써 공통으로 등장하지 않는 어휘와 고등레벨의 어휘를 포함한다.

표 1. 어휘 레벨 구분 기준
Table 1. Criteria of vocabulary level division

레벨 (난이도)	포함 관계	어휘 개수
1	초등 ⊂ A, 초등 ∩ B ∩ C	722
2	초등 ⊂ A, (초등 ∩ B) ⊄ C	274
3	초등 ⊂ A, 초등 ⊄ (B ∪ C)	408
4	중1 ⊂ B, A ∩ 중1 ∩ C	430
5	중2 ⊂ B, A ∩ 중2 ∩ C	485
6	중3 ⊂ B, A ∩ 중3 ∩ C	296
7	중등 ⊂ A, 중등 ∩ B ∩ C	304
8	중1 ⊂ B, 중1 ⊄ (A ∪ C)	112
9	중2 ⊂ B, 중2 ⊄ (A ∪ C)	120
10	중3 ⊂ B, 중3 ⊄ (A ∪ C)	93
11	Etc.	1028

3.1.2 도서 구문분석 및 난이도 결정

<그림 1>은 독자 검토용 어휘 데이터베이스 생성 과정을 보여준다. 도서 구문분석, 어휘 데이터 전처리는 영문도서에서 사용된 어휘를 분석하는 단계로, 도서내의 어휘들을 분석하고 전처리를 통하여 고유명사와 같은 이독성에 필요 없는 요소를 걸러내고. 그 후, 기준 어휘와의 비교를 통해 어휘난이도를 부여하고 독자가 검토 할 어휘를 추출한다.

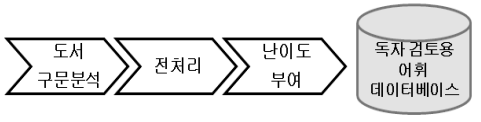


그림 1. 어휘 데이터베이스 생성 과정
Figure 1. Vocabulary database generation process

3.1.3 어휘 추출

사용자가 읽기 수준 판단을 원하는 도서를 선택하면 해당 도서의 난이도별로 분류된 어휘 데이터베이스에서 독자에게 제시할 어휘를 추출해야 하는데, 단순히 무작위로 추출하여 자칫 자주 등장하지 않는 어휘를 많이 추출하거나, 낮거나 높은 수준의 어휘 위주로 추출하면 독자의 판단 결과가 실제 도서의 수준과 거리가 멀 수 있다. 따라서, 레벨(Level)별 어휘 개수의 분포를 고려하여 어휘의 개수가 많은 레벨에서는 상대적으로 많은 어휘를, 어휘의 개수가 적은 레벨에서는 상대적으로 적은 어휘의 개수를 테스트 어휘로 추출한다.

$$Level\,x\text{의 테스트 어휘 개수} = \frac{Level\,x\text{ 어휘 개수}}{\text{전체 어휘 개수}(n)} \times \text{테스트 어휘 총 개수}(m) \quad (1)$$

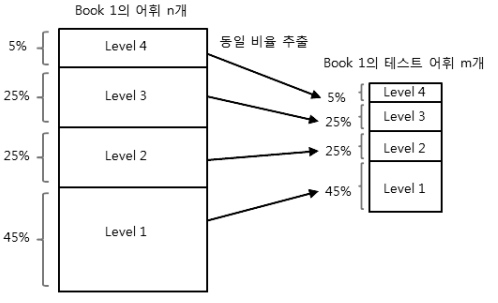


그림 2. 레벨별 테스트 어휘 개수 산출
Figure 2. Computation of the number of test words for each Level

레벨별 테스트 어휘 추출을 위하여 두 단계를 거친다. 첫 번째 단계에서 레벨별 테스트를 위해 선택할 어휘의 개수 결정하고, 두 번째 단계에서 그 개수에 맞추어 어휘를 선택한다. 첫 번째 단계에서 레벨별 어휘 개수의 분포를 고려하여 테스트 어휘 개수를 산출하기 위해서는 식(1)을 이용하여 쉽게 구할 수 있으며, <그림 2>에서 레벨별 비율에 따른 테스트 어휘 개수 할당 예를 보여준다. 두 번째 단계의 어휘 선택과정에서는 각 어휘의 등장 횟수, 즉, 빈도수를 고려한다. 전처리 과정 중에 도서 내에 등장하는 모든 어휘의 빈도수를 계산하여 빈도수의 비율에 따라 선택 확률을 할당한다. <표 2>에서 간단한 예를 보여준다. 총 5개의 어휘 중 빈도수 10을 가지는 Apple의 경우 문제에 출제될

확률은 50%이다. 반면에 빈도수 1을 가지고 있는 Egg의 경우 5%의 낮은 확률을 가지고 있다.

표 2. 어휘 빈도수에 따른 선택 확률 예
Table 2. An example of selection probability according to word frequency

어휘	빈도수	확률(%)
Apple	10	50
Bee	5	25
Cat	3	15
Dog	1	5
Egg	1	5

3.2 읽기 수준 적합성 정량화

해당 도서에 대한 독자의 읽기 수준을 판단하기 위해서는 독자의 테스트 어휘 검토 결과를 점수화 할 필요가 있다. 간단한 방법으로써, 독자가 테스트 어휘 검토 시 알고 있는 어휘에 체크하도록 하고 제시된 총 어휘의 개수와 체크된 어휘의 개수의 비율을 점수화 할 수 있을 것이다. 하지만, 단순히 알고 있는 어휘의 개수만으로 점수화 하면 어휘의 레벨이 고려되지 않아 정확한 읽기 수준 판단 결과를 얻을 수 없다. 따라서, 본 논문에서는 난이도가 높은 어휘에 가중치를 높게 두어, 즉, 높은 레벨에 높은 가중치를 할당하였다.

점수화 과정은 다음과 같다. 레벨별 가중치를 할당하고, 만점, 즉, 모든 어휘를 알고 있을 경우의 점수와 독자의 점수를 계산하여 100분위로 환산한 결과를 도서와 독자간 수준 적합성 점수로 한다. 레벨의 개수를 N_L, 각 레벨을 $L(x)$ (x = 1, 2, 3, ..., N_L)로 나타낼 때, 각 레벨의 테스트 어휘 개수를 $N_{L(x)}$, 레벨별 가중치를 $W_{L(x)}$, 각 레벨의 어휘 중 독자가 알고 있다고 체크한 어휘의 개수를 $N_{Chk(x)}$로 놓으면, 만점 $S_{perfect}$, 독자 점수 S_{reader}, 적합성 점수 S_{suit} ($suit$ = suitability(적합성))는 각각 식(2), (3), (4)과 같이 구할 수 있다.

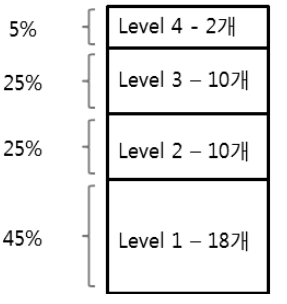
도서마다 레벨별 어휘의 수가 다를 것이기 때문에 $S_{perfect}$는 도서마다 다르게 나올 것이기 때문에 절대 점수를 사용할 수는 없으므로 식(4)을 이용하여 100분위로 환산한 S_{suit}를 적합성 점수로 사용한다.

$$S_{perfect} = \sum_{x=1}^{N_L} (N_{L(x)} \times W_{L(x)}) \quad (2)$$

$$S_{reader} = \sum_{x=1}^{N_L} (N_{Chk(x)} \times W_{L(x)}) \quad (3)$$

$$S_{suit} = \frac{S_{reader}}{S_{perfect}} \times 100 \quad (4)$$

Book 1의 테스트 어휘 40개



N_L = 40,
$N_{L(1)}$ = 18, $N_{L(2)}$ = 10, $N_{L(3)}$ = 10, $N_{L(4)}$ = 2,
$W_{L(1)}$ = 1, $W_{L(2)}$ = 2, $W_{L(3)}$ = 3, $W_{L(4)}$ = 4.
∴ $S_{perfect}$ = (18*1)+(10*2)+(10*3)+(2*4) = 76

Case 1 : 독자가 모든 어휘를 알고 있을 경우
$N_{L(x)} = N_{Chk(x)}$ 이므로 $S_{reader} = S_{perfect}$ = 76
∴ S_{suit} = 76/76 * 100 = 100

Case 2 : 독자가 Level 1과 2의 어휘만 알고 있을 경우
S_{reader} = (18*1)+(10*2) = 38
∴ S_{suit} = 38/76 * 100 = 50

그림 3. 적합성 점수 계산 예
Figure 3. Examples of calculating suitability scores

<그림 3>은 간단한 예를 보여준다. 총 4개의 레벨이 있고, 테스트 어휘의 총 개수는 40개이며, 레벨과 가중치를 같은 값으로 할당하고 두 가지

Case에 대해서 S_{suit}를 구하였다. Case 2의 결과에서 독자가 총 테스트 어휘 중 70% 이상을 알고 있다 하더라도 낮은 레벨의 어휘만 알고 있으므로 S_{suit}는 50으로 낮게 나왔음을 확인 할 수 있다.

4. 구현 및 평가

본 논문에서 제안한 영어 도서와 독자 간 수준 적합성 판단 알고리즘의 효용성을 평가하기 위하여 웹 형태로 구현하고 중·고등학생들을 대상으로 설문조사를 실시하였다. Windows 7 이상의 버전에서 JAVA, JSP, Tomcat을 이용하여 구현하였으며, 도서 구문분석을 위하여 Stanford Parser[11]를 사용하였다. <그림 4>와 <그림 5>는 구현한 시스템의 기능 및 구현 흐름도를 보여준다. 관리자 부분에서 도서 내의 어휘를 구문 분석, 원형으로 변환하는 전처리 과정과 한국 독자들의 학습 단계를 고려하여 수집된 어휘 데이터를 이용한 도서 내의 어휘 레벨 결정 과정이 이루어진다.

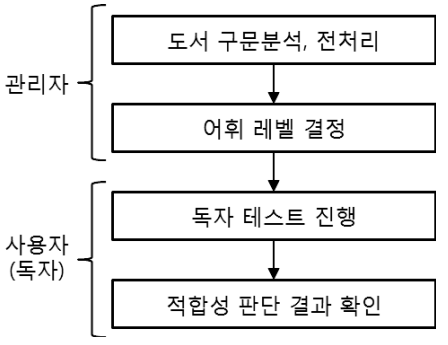


그림 4. 시스템 기능 플로우
Figure 4. System functional flow

즉, eBook에서 텍스트(TXT)를 추출하고 파서를 통하여 원형 변환 및 등장 횟수를 계산하고 기준 어휘 데이터를 이용하여 파싱된 eBook의 어휘를 레벨별 분류 및 필터링 하여 테스트용(문제출제용) 어휘 데이터베이스를 구축한다. 사용자 부분은, 테스트지 생성(Create Quiz), 독자가 알고 있는 어휘를 선택하는 독자 테스트, 테스트 결과에 따른 적합성 제안 메시지 출력으로 구성된다.

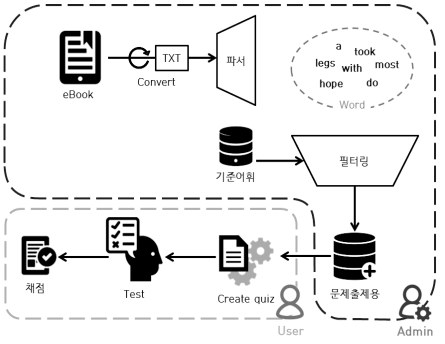


그림 5. 시스템 구현 과정
Figure 5. System implementation process

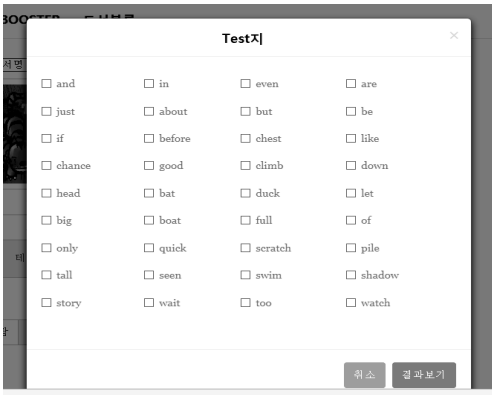


그림 6. 독자 테스트 화면
Figure 6. Screen for reader test

<그림 6>은 실험을 위하여 사용한 eBook 중 하나에 대한 테스트지이다. 사용자가 알고 있는 어휘에 체크를 할 수 있도록 구성되어 있다. 마지막으로 테스트 결과를 기반으로 독자에게 적합성을 판단 할 수 있도록 제안을 해 주어야 하는데, 이를 위하여 (주)낱말이 연구하고 교보문고에서 제공하는 어휘력과 도서 이해도의 상관관계 표를 참고하여 <표 3>과 같이 점수 열을 본 논문에서 구한 Ssuit으로 대체하였다[12].

표 3. 적합성 점수에 따른 제안 문구
Table 3. Suggestions for suitability scores

점수 (S_{suit})	판단	제안 문구
90점 이상	EASY	- 당신의 수준에 비해 쉬운 도서입니다. - 깊이 있는 지적 활동은 이루어지지 않을 것으로 예상됩니다. - 독서에 대한 흥미 저하, 독서효과 저하를 유발할 수 있습니다.
75점 이상	GOOD	- 자신감 있고 편안한 독서가 가능합니다. - 재미있고 효과적인 독서가 가능합니다. - 도전의식과 지적호기심을 자극합니다.
75점 미만	HARD	- 나의 수준에 비해 너무 어려운 도서입니다. - 독서에 대한 좌절감을 느낄 수도 있습니다. - 이 도서에 대한 흥미를 잃을 수도 있습니다.

〈그림 7〉은 구현한 시스템에서 화면으로 보여주는 최종 제안 예이다. 제안한 알고리즘을 평가하기 위하여 난이도가 다른 세 권의 도서를 선정하고 중·고등학생 80명을 대상으로 설문을 통하여 평가하였다. 여러 명에대해서 동시에 테스트 및 설문조사를 하기 위하여, 세 권의 도서에 대해서 내용의 일부분, 생성한 테스트 어휘 리스트, 제안 문구를 문서로 준비하여 설문에 임한 학생이 선택한 서적에 대해서 문서를 나누어주고 시스템에서 제안한 문구의 적절성 여부를 설문을 통하여 조사하였다.

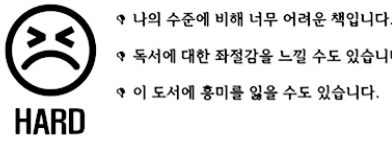
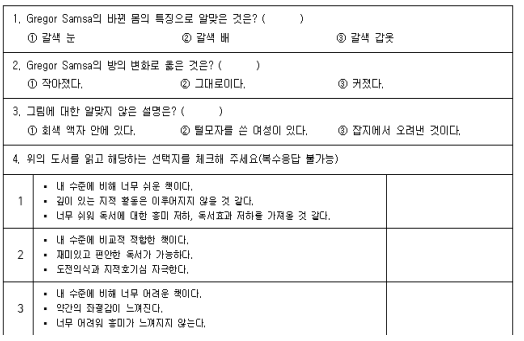


그림 7. 구현한 시스템에서 제시하는 최종 제안 예
Figure 7. A final suggestion example in the implemented system

설문의 신뢰성을 높이기 위해서, 즉, 테스트에 충실히 임하지 않은 학생들의 결과를 걸러내기 위하여 도서의 내용에 대한 간단한 문제도 포함하였다.

〈그림 8〉은 설문지의 한 예이며, 결과적으로 〈그림 9〉와 같이 학생들 중 약 80%가 제안한 문구가 적절했다고 답해 주어, 제안한 알고리즘이 적용이 간단하면서도 어느 정도 신뢰할 수 있음을 확인할 수 있다.


1. Gregor Samsa의 바뀐 몸의 특징으로 알맞은 것은? ( )
① 갈색 눈 ② 갈색 벼 ③ 갈색 갑옷

2. Gregor Samsa의 방의 변화로 옳은 것은? ( )
① 작아졌다. ② 그대로이다. ③ 커졌다.

3. 그림에 대한 알맞지 않은 설명은? ( )
① 회색 액자 안에 있다. ② 털모자를 쓴 여성이 있다. ③ 잡지에서 오려낸 것이다.

4. 위의 도서를 읽고 해당하는 선택지를 체크해 주세요(복수응답 불가능)

1	▪ 내 수준에 비해 너무 쉬운 책이다. ▪ 깊이 있는 지적 활동은 이루어지지 않을 것 같다. ▪ 너무 쉬워 독서에 대한 흥미 저하, 독서효과 저하를 가져올 것 같다.	
2	▪ 내 수준에 비교적 적합한 책이다. ▪ 재미있고 편안한 독서가 가능하다. ▪ 도전의식과 지적호기심 자극한다.	
3	▪ 내 수준에 비해 너무 어려운 책이다. ▪ 약간의 좌절감이 느껴진다. ▪ 너무 어려워 흥미가 느껴지지 않는다.	

그림 8. 설문지 예
Figure 8. An example of evaluation form

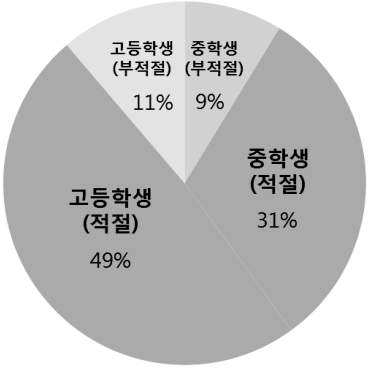


그림 9. 제안한 알고리즘 평가 결과
Figure 9. Evaluation of the proposed algorithm

5. 결 론

본 논문은 영어를 모국어로 사용하지 않는 우리나라 독자들에 대하여 어휘 기반으로 간단하고 빠르게 영문도서와 독자 간 수준 적합성을 판단하는 방법을 제시하였다. 각 영어 도서에 등장하는 어휘를 한국인의 영어 교육과정을 고려한 어휘 목록과

비교하여 테스트 할 어휘를 추출하여 그 도서에 대한 독자의 읽기 수준 적정성을 평가하는 방법을 체계화 하였다. 기존의 여러 가지 영역의 테스트를 통하여 독자의 수준을 파악하는 방식과는 달리 관심 도서의 어휘만을 가지고 몇 분만의 테스트를 통하여 도서에 대한 독자 수준의 적합성을 파악하므로 적용이 훨씬 쉽다. 향후, 신뢰성을 더욱 향상시키기 위해서, 어휘 선정뿐만 아니라 핵심적인 구나 문장을 선정하는 방법 연구와 어린이부터 성인까지 다양한 연령대와 전문서적을 포함한 다양한 분야의 영문도서에 대한 실험이 필요하다.

References

[1] Statistics Korea - National Reading Survey, http://meta.narastat.kr/metasvc/index.do?confmNo=11318&inputYear=2013.

[2] Kwangjin Lee, http://www.bookdaily.co.kr/news/quickViewArticleView.html?idxno=39398.

[3] Lexile, https://www.lexile.com/.

[4] Jeong-ryeol Kim, and Eun Hye Lee, *Development of Korean lexile text measure for differentiated level reading of elementary school students*, Pan-Korea English Teachers Association, English Language Education, Vol. 22, No. 2, pp. 227-254, 2010.

[5] Natmal, Ltd., http://www.natmal.com/.

[6] J. S. Chall, and E. Dale, *Readability revisited: The new dale-chall Readability formula*, Cambridge, 1995.

[7] Hung-soo Lee, *Models for measuring sentential readability*, The English Language and Literature Association of Korea, English Language and Literature, Vol. 31, No. 2, pp. 321-337, 1985.

[8] Ministry of Education and Science Technology, 2*009 English Department curriculum in accordance with the revised curriculum ('11 .8.9 notification)*, pp. 55-90, 2011.

[9] NAVER, Middle & High School Textbook Words http://endic.naver.com/lesson.nhn?sLn=kr.

[10] Oxford Learner's Dictionaries, The Oxford 3000™ http://www.oxfordlearnersdictionaries.com/us/wordlist/american_english/oxford3000/.

[11] The Stanford Parser: A statistical parser, The Stanford Natural Language Processing Group, http://nlp.stanford.edu/software/lex-parser.shtml.

[12] Kyobobook, *My correlation between READ score and book READ index*, http://www.kyobobook.co.kr/readinglevel/ReadingLevelGuide.jsp.

영문도서에 대한 독자의 어휘력 기반 읽기 수준 판단 알고리즘 개발

송한결[1], 정은우[1], 조수경[1], 최용호[2], 이현빈[1]

[1] *한밭대학교 컴퓨터공학과*

[2] *(주)유코아시스템 기술연구소*

요 약

성공적인 독서를 위해서는 독자의 읽기 수준에 적합한 도서를 선정하는 일이 매우 중요하다. 특히, 외국어 도서 선정 시에는 도서의 수준과 독자의 읽기 수준을 동시에 고려해야 하기 때문에 더욱 신중을 기해야 한다. 우리나라에서 판매하는 영어 도서의 경우에는, 미국이나 한국의 기관에서 도서 수준 분석 방법을 사용하여 제시하는 도서의 난이도 지수를 참고하여 어느 정도 객관적인 수준을 짐작 할 수 있다. 하지만, 독자의 읽기 수준은 장시간 영어 테스트를 거쳐서 파악해야 하므로 대부분의 독자들은 자신의 읽기 능력을 주관적으로 판단하여 자신의 수준에 적절하지 않은 도서를 선정할 가능성이 높다. 본 논문은 영어를 모국어로 사용하지 않은 우리나라 독자들에 대하여 어휘 기반으로 간단하고 빠르게 영문도서와 독자 간

수준 적합성을 판단하는 알고리즘을 제시한다. 영어 도서에 등장하는 어휘를 한국인의 영어 교육과정을 고려한 어휘 목록과 비교하여 테스트 할 어휘를 추출하여 그 도서에 대한 독자의 읽기 수준 적정성을 평가하는 방법을 체계화 한다. 구현 및 설문을 통하여 제시한 알고리즘의 정확도 분석 결과를 제시한다.

Hangyeol Song is a B.S. candidate in the Department of Computer Engineering, Hanbat National University, South Korea. Her research interests include text mining, readability analysis, and recommendation systems.
E-mail address: ucc5328@naver.com

Eunwoo Joung is a B.S. candidate in the Department of Computer Engineering, Hanbat National University, South Korea. Her research interests include Web search and information retrieval.
E-mail address: dmsdn3234@gmail.com

Sukyeong Cho is a B.S. candidate in the Department of Computer Engineering, Hanbat National University, South Korea. Her research interests include Internet of Things (IoT) infrastructure and IoT service.
E-mail address: madness.sk86@gmail.com

Yong Ho Choi received the B.S. degree in the Department of Computer Engineering from National Institute for Lifelong Education, South Korea, in 2013. He received the M.S. degree in Computer Engineering from Hanbat National University, South Korea, in 2015. Since 2008, he has been working as a senior researcher in the research institute of U-Core system. His research interests include HTML5, intelligent data analysis, and traffic information system.
E-mail address: yhchoi@u-core.co.kr

Hyunbean Yi received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Hanyang University, South Korea, in 2001, 2003, and 2007, respectively. He was with Korea Electronics Technology Institute (KETI) from 2002 to 2007. He had been a Research Scholar at the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, USA from 2007 to 2009. He had been a Postdoctoral Researcher at the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan from 2009 to 2011. Currently, he is a Professor in the Department of Computer Engineering/Graduate School of Information & Communications, Hanbat National University, South Korea. His research interests include topic mining, recommendation systems and design-for-reliability.
E-mail address: bean@hanbat.ac.kr