International Technology and Education Journal
Vol. 2 No. 1; June 2018
ISSN: 2602-2885
URL: http://itejournal.com/

Use of Open-Ended Questions in Measurement and Evaluation Methods in Distance Education

İbrahim BENLİ*, ibrahimbenli@hotmail.com, TURKEY, https://orcid.org/0000-0001-8316-0875
Rita İSMAİLOVA, rita.ismailova@manas.edu.kg, KYRGYZSTAN, https://orcid.org/0000-0003-0308-2315

SUMMARY

In the 21st century, which is characterized as the Information Age, information access, knowledge quick learning is vital to the development of individuals and societies. With the use of technological innovations in the field of education in the information society, it will be possible to acquire a lasting place in the globalizing world. Distance Education refers to a model of the education system that students and teachers carry out through their learning and teaching activities by communicating via technologies and postal services. According to the year 2015 data, 68 out of 184 higher education institutions in Turkey offer an open and distance learning program. 47 of them are undergraduate, 17 are graduate, 11 are graduate completion and 56 are in master degree level. In total there are 505 different programs. The measurement and evaluation process of results of training is as important as developing content in distance education applications. When question types used in Distance Education Measurement and Assessment are examined, the use of Open-ended Questions is less than other methods. However, it is well-known fact that these type of questions are good predictors of the students' knowledge. The biggest problem that arises with the usage of open-ended questions is the evaluation part. The different answers that students will give to the questions, their personal narrative skills, or the interpretations they will answer in response to the questions make the evaluation process difficult. At this point, it would be better to interpret the answers recorded in the database with Text Mining methods and Natural Language Processing techniques. In this work, we implement an algorithm for evaluation of open-ended question. The experimental results showed that a correlation was found between 0,89 - 0,96 when evaluating open-ended questions of our system by teacher evaluation.

Keywords: Distance Education, Open Ended Questions, Natural Language Processing

INTRODUCTION

In the 21st century, which is characterized as the Information Age, information access, knowledge quick learning is vital to the development of individuals and societies. It is almost impossible to follow developments in information technology in the century we are living. With the rapid development of information technology, a communication network that surrounds the whole world has been established. With the use of technological innovations in the field of education in the information society, it will be possible to acquire an important place in the globalizing world.

Table 1. World Internet Users and 2017 Population Stats

World Regions	Population (2018 Est.)	Population % of World	Internet Users 30 Jun 2017	Penetration Rate (% Pop.)	Growth 2000-2017	Internet Users %
Africa	1.246.504.865	16.6 %	388,376,491	31.2 %	8,503.1%	10.0 %
Asia	4.148.177.672	55.2 %	1,938,075,631	46.7 %	1,595.5%	49.7 %
Europe	822.710.362	10.9 %	659,634,487	80.2 %	527.6%	17.0 %
Latin America / Caribbean	647.604.645	8.6 %	404,269,163	62.4 %	2,137.4%	10.4 %
Middle East	250.327.574	3.3 %	146,972,123	58.7 %	4,374.3%	3.8 %
North America	363.224.006	4.8 %	320,059,368	88.1 %	196.1%	8.2 %
Australia	40.479.846	0.5 %	28,180,356	69.6 %	269.8%	0.7 %
World Total	7.519.028.970	100.0 %	3,885,567,619	51.7 %	976.4%	100.0 %

* Kyrgyz Turkish Manas University, Department of Computer Engineering, Bishkek, Kyrgyz Republic

As shown in Table1, according to the statistics dated June 30, 2017, 51.7% of the world population 3.885.567.619 people is able to reach internet (Internet World Stats, 2017).

This global network is the main source of information on scientific research, productivity, cultural change, world trade and world-wide education. This network is a global hub for the communication of written, verbal and visual communication among all people living in the world (Çallı, İşman, & Torkul, 2002). At this point, the concept of distance education is at the forefront. There are various definitions of this concept. This concept is also expressed as "Distance Learning" and "e-Learning".

What is Distance Education?

Distance Education; refers to a model of the education system that students and teachers in different settings carry out through their learning and teaching activities, communication technologies and postal services (İşman, 2005).

e-Learning; is a tool or system that enables you to learn at any time or anywhere, with an intuitive computer based tutorial system. Today, e-learning activities are mostly carried out via the Internet (Epignosis, 2014).

Distance education in Turkey, the first time in the 1950s began to be made by letter within the scope of in-service training for bank employees in Ankara University Faculty of Law. In the 1960s, distance education activities started with letters at the level of higher education, and in 1982, Anadolu University has implemented the open and distance education applications. The first of the online computer-assisted online program was given by the 2000 Bilgi University e-MBA. This was followed in 2001 by ODTÜ's online learning master program in the field of information (Kaya, 2002).

According to the year 2015 data, 68 out of 184 higher education institutions in Turkey offer an open and distance learning program. 47 of them are undergraduate, 17 are graduate, 11 are graduate completion and 56 are in master degree level. In total there are 505 different programs (Koçdar & Görü Doğan, 2015).

In distance education applications, which are getting more and more varied day by day, the measurement and evaluation process as a result of the education applied as well as the content development are also different.

Assessment methods in distance education are examined in two different groups. Called traditional evaluation methods, oral exams, written exams, multiple choice exams, Gap Fill-exams, True False Tests and Alternative assessment methods referred to as portfolio assessment, project, an open book, Concept Maps, Authentic Assessment, Tree, Peer Review (Balta & Türel, 2013).

When question types used in Distance Education Measurement and Assessment are examined, the use of Open-ended Questions is less than other methods. However, it is well-known fact that these types of questions are good predictors of the students' knowledge. The biggest problem that arises with the usage of open-ended questions is the evaluation part. The different answers that students will give to the questions, their personal narrative skills, or the interpretations they will answer in response to the questions make the evaluation process difficult. At this point, it would be better to interpret the answers recorded in the database with Text Mining methods and Natural Language Processing techniques.

Recently, some studies have been done to automatically evaluate students' answers to open-ended questions. In a web-based learning environment called Apex, students' answers are automatically assessed by LSA (Latent Semantic Analysis) technique (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Dessus, Lemaire, & Vernier, 2000; Foltz, Kintsch, & Landauer, 1998). With the BLEU (Bilingual Evaluation Understudy) method, the answers of the students with the reference words were used to determine the accuracy of the translation process, and this method helps in the scoring process (Pérez, Alfonseca, & Rodríguez, 2004). With the M-BLUE (Modified BLUE) method, the closest reference answer to the student answer was found and then the student answer and the selected reference answer were compared (Noorbehbahani & Kardan, 2011). The Atenea evaluation system is integrated into the Tangow system, a Web-based Learning tool, and short answers of the students are automatically scored (Carro, Pulido, & Rodríguez, 2000; Pérez-Marín, Alfonseca, & Rodríguez, 2006). The Atenea system was developed and the Willow system was proposed, in which student answers and answers given by the teacher are compared with Natural Language Processing techniques (Perez-Marin, Pascual-Nieto, Alfonseca, Anguiano, & Rodriguez, 2007). In order to increase the accuracy of the assessment, some researchers have obtained the word of POS (part-of-speech) and combined it with the LSA technique (Kanejiya, Kumar, & Prasad, 2003; Wiemer-Hastings & Zipitria, 2001). An automatic evaluation process was performed

with the Crater system, in which the frequency similarity of the terms was combined with the frequency analysis of the terms (Leacock & Chodorow, 2003) .

The above mentioned studies should be evaluated in English, French and Spanish languages. In this study, it will be checked whether the keywords that will be determined by the teacher using the Natural Language Processing Techniques in Turkish language are in the student answers and a result will be obtained as a percentage. Thus, the teacher may have an idea about the level of the students according to the key words he / she determines. At the same time, students will be able to evaluate their own levels.

In this study, we will be discussed from the evaluation techniques in Section 2, the method we used in Section 3, from the findings obtained in the research result in Section 4, the comparison of results and the recommendations for future work in Section 5.

Evaluation Techniques

Natural Language Processing

The answers to open-ended questions are constructed using natural language and are usually not structured. For this reason, in order to apply the techniques of Natural Language Processing on texts, there is a need for a preprocessing step which transforms non-structural data structure. Tokenization, stopword filtering, stemming, and term weighting are some commonly used methods in text mining to preprocess text documents (Hotho, Nurnberger, & Paas, 2005).

Information Extraction and Pattern Matching

Information extraction (IE) techniques pull out pertinent information from syntactically analyzed pieces of text answers by applying a set of patterns. Patterns are defined either on surface text (words, phrases) or structural elements such as parts of speech (PoS) tags. In the case of short free-text answers, they are typically created by subject matter experts to indicate important concepts which should be present in answers (Roy, Narahari, & Deshmukh, 2015).

Text Similarity

Finding similarities between two texts to be compared is a frequently used method. Problem-based results can be obtained in dictionary-based comparisons. For example, automobile and cars have the same meaning in Turkish, but in the dictionary they are two different words. In this case, better performance can be achieved by using LSA techniques (Roy et al., 2015).

METHOD

Case Study

In this study, the answers given by 8 students in the written exam for the 11th grade History lesson were used at Kyrgyz Turkish Anatolian High School. A pool of 6 to 10 keywords was created for the evaluation of the 3 selected open-ended questions. The responses of students to open-ended questions were recorded in the database and evaluated using our method.

Table 2. Data of Case Study

No	Questions (Turkish)	Keywords (Turkish)
1	Atatürk ilkelerini yazınız.	Cumhuriyetçilik, Milliyetçilik, Halkçılık, Devletçilik, Laiklik, inkılapçılık
2	Şeyh Sait olayı hakkında bilgi veriniz.	Doğu Anadolu, Musul, İngilizler, Terakkiperver Cumhuriyet Fırkası, istiklal mahkemeleri, isyan, 1925
3	Türk Hava Kurumu hakkında bilgi veriniz.	Atatürk, 1925 Türk Tayyare Cemiyeti, 1935 Türk Kuşu

Table 2 shows Turkish questions and keywords produced for our case study.

Evaluation

We use in the developed system for programming language Java (build 1.8.0_144-b01) and MySQL (version 5.7.14) is used as the database. Figure 1 shows Evaluation Algorithm.

Preprocessing

Firstly, database connection is performed and data is collected. First the toLowercase() function is used to prevent upper case for keywords. After the "," characters have been deleted and cleared. In the next step, the repeated words are cleared. Stopwords with no meaning of Turkish language (ve, ancak, bazı, etc.) have been cleaned. The cleaned data is transferred back into the ArrayList and the data is updated in the program. The same operations are performed in the answers given by the students and transferred to the ArrayList.

Processing

A StringTokenizer () operation has been performed to handle the word or vocabulary of the obtained cues or word groups. Stemming has been performed for each word. For stemming process Turkish Natural Language Processing Library zemberek-nlp (version 0.13.0) was used (Akın & Akın, 2007). In the stemming process if stem isn't found in Turkish language, the word is accepted as a special name (Wilson, TBMM, etc.) and taken as it is.

Interpretation / Evaluation

The answers given by the students are compared with the determined key words. In calculating the success rate; If the keyword is in the answer, the value of Kt (Keyword true) is incremented by one. While the percentage is calculated as the result of the evaluation process O (Percentage Ratio) and Ks(Total Keywords) values with O = (Kt/Ks)*100 formula , the accuracy rate for the student response was calculated.

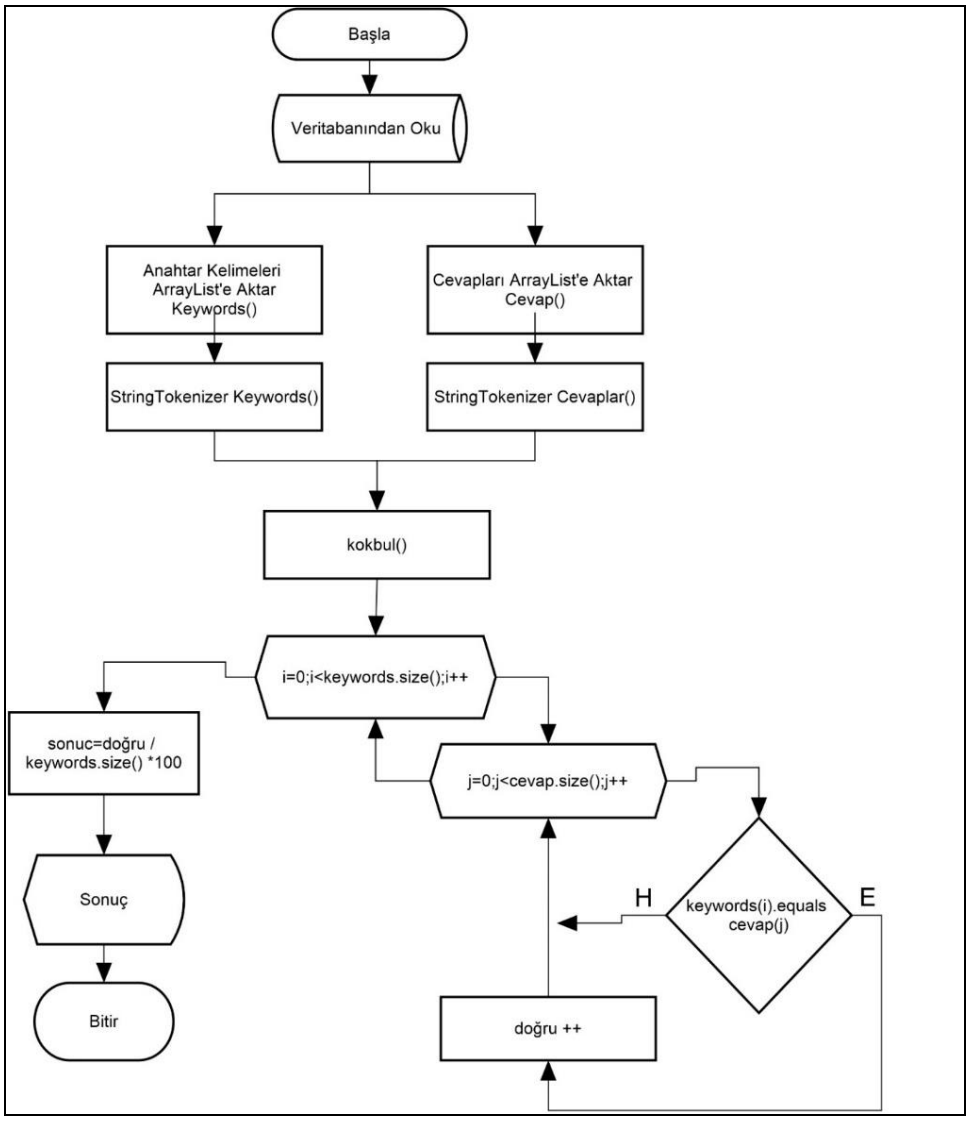

Figure 1. Evaluation Algorithm (Some terms are written in Turkish)

FINDINGS

In this study, the analysis of the open-ended questions evaluation algorithm has been analyzed. In Table 3 evaluation results were given. The responses of 8 students to 3 open ended questions were checked for algorithm operation control. Zemberek-nlp 0.13.0 library was used for stemming process. 91487 registered word is separated from the inserts examined continued for 8 minutes on our computer (HP dv6 – 1378nr, Windows 7 64 Bit). In order to compare the results of the evaluation process with the scores given by the teachers, the scores given by 2 teachers to the answers are also given in Table 3. Scores are based on 100 points for each question.
Scorer Teacher 1: Experience: 12 years, Profession: History
Scorer Teacher 1: Experience: 16 years, Profession: Social Studies

Table 3. Evaluation Results

Question No	Student Name	Teacher 1	Teacher 2	Average Score (Teachers)	Our System
1	S1	67	67	67	66,67
	S2	83	83	83	83,33
	S3	67	67	67	50
	S4	50	50	50	50
	S5	67	67	67	66,67
	S6	100	100	100	100

		S7	100	100	100	83,33
		S8	100	100	100	100
	Average		79,25	79,25	79,25	75,00
2		S1	100	100	100	80
		S2	100	95	97,50	80
		S3	80	86	83	60
		S4	33	40	36,50	30
		S5	100	100	100	90
		S6	67	60	63,50	50
		S7	80	85	82,50	50
		S8	85	80	82,50	80
	Average		80,63	80,75	80,69	65,00
3		S1	100	100	100	100
		S2	33	25	29	16,66
		S3	53	50	51,50	50
		S4	100	100	100	83,33
		S5	50	50	50	33,33
		S6	100	100	100	100
		S7	100	100	100	100
		S8	100	100	100	100
	Average		79,50	78,13	78,81	72,92

When the first question evaluation is examined, there is a parallel between the teacher scores and the system scores. While the teacher average score was 79.25, the system average score was calculated as 75.00. In this question, the answer consists of six items. When we examined student answers, it was determined that the difference between system and teacher scores in S3 and S7 student evaluation was due to the synonyms of "İnkılapçılık" and "Devrimcilik".

When the evaluation of the second question is examined, it appears that the teacher scores are higher than the system scores. While the teacher average score was 80.69, the system average score was 65.00 . The answers to this question depend on the comments of the students. The answer can be explained in three or four sentences, as well as in ten sentences. The "Doğu Bölgesi" word was accepted correctly by the teachers, but the system could not detect the "doğu Anadolu" expression in the keywords in some answers. It has been determined that the difference in the results is caused by this.

When the evaluation of the third question is examined, there is a parallel between the teacher scores and the system scores, too. While the teacher average score was 78.81, the system average score was 72.92 . Located two questions in this keyword phrase dates in 1925 and 1935 to compare the accuracy of the assessment process to facilitate the positive effect was seen.

According to the results, when teachers' average scores and system scores are examined, correlation between two variables were calculated; r_1= 0,92, r2= 0,89, r3= 0,96. Accordingly, it can be said that the relationship between the two variables is high. There was no correlation between the number of keywords and the evaluation result. However, it was determined that there was a correlation between the answers length and the evaluation result. It was determined that the probability of finding a keyword is high if the student answer is long.

CONCLUSION AND DISCUSSION

When the types of questions used in Distance Education and Assessment are examined, the use of Open-Ended Questions is less than that of other methods when considering the above-mentioned advantages and disadvantages (Scalise & Gifford, 2006).

In open-ended questions, the emotions and thoughts of the source can be reached (analysis, synthesis and evaluation ability). It is thought that open-ended questions can improve the free thought that the individual can

reveal creative answers without being limited by options. It is suitable for measuring advanced behaviors. There is no chance of luck (Gelbal & Kelecioğlu, 2007)

Today, open-ended questions have begun to be used in ÖSYM (Student Selection And Placement Center) and MEB (The Ministry of National Education) examinations in order to make the measurement and evaluation more reliable (MEB, 2017; ÖSYM, 2017).

The most important problem in the use of open-ended questions is the evaluation part. The different answers that students will give to the questions, their personal narrative skills, or the interpretations they will answer in response to the questions make the evaluation process difficult. At the same time, it is possible that the evaluator teachers may score different points in the answers during the evaluation process.

At this point computer based evaluation process gives more objective results. The Apex application found a correlation between teacher ratings and computer-based system scores ranging from 0,59 to 0,68 (Dessus et al., 2000). Foltz and his colleagues found a correlation between scores of 0,80-0,86 in their study (Foltz, Laham, & Landauer, 1999). Landauer found a correlation of 0.86 between the scores in the Automatic Assessment in Education study (Landauer, 2003). When the BLUE algorithm was applied, a correlation of 0.76 was observed between the scores (Pérez et al., 2004). In the M-BLUE algorithm, a correlation of 0,85 was observed between the scores (Noorbehbahani & Kardan, 2011). In the system called CRATER, Leacock and his colleagues found 85% similarity between scoring (Leacock & Chodorow, 2003). In a study conducted using the kind of words (POS) was found correlation between 0.56 and 0.60 (Kanejiya et al., 2003; Wiemer-Hastings & Zipitria, 2001). In our study, the correlation was between 0.89-0.96 and similarities with the studies in the literature. When the questions and answers are examined, the accuracy of the evaluation is higher when the answers are written in the form of a short answer or items.

In order to obtain better results, the answer "Ankara" while the "Anakra" written when the zemberek-nlp normalization process is applied. For synonyms and homologous words, key words must be chosen properly. In this study, calculations are made on text similarity. Only the similarity of texts will not always give the correct result when considering the multitude of synonyms and homonyms in Turkish. Using semantic analysis algorithms will increase the accuracy rate. It is stated that zemberek-nlp version 0.13.0 used in our work will become stable with 1.0 version later. Thus, the accuracy rate of Natural Language Processing studies in Turkish will increase. The duration of the stemming process is directly proportional to the number of words used in the keywords and answers. The more words there are, the longer it takes the program to run and evaluate. The average stemming in our study is around 2000 ms per word.

As in other languages in Turkish, the order in which the words are used influences the meaning. For example; Even though they have different meanings of "dog man bite" sentence and "man dog bite" sentence, they will get the same score when the word roots are examined and evaluated. In this case, analyzes should be made according to the structure of the subject + verb + predicate in Turkish. Islam, in his work, calculated the relation of the words in English, the characters used in words, the type and order of the word , found 82% similarity between two texts with the same meaning(Islam & Inkpen, 2008) .

Future works will focus on text similarity, sequence word type (Name, Adjective, Verb, etc.), and the meaning of the phrase.

REFERENCES

Akın, A. A., & Akın, M. D. (2007). Zemberek, an open source nlp framework for turkic languages. *Structure*, *10*, 1–5.

Balta, Y., & Türel, Y. K. (2013). Çevrimiçi uzaktan eğitimde kullanılan farklı ölçme değerlendirme yaklaşımlarına ilişkin bir inceleme. *Electronic Turkish Studies*, *8*(3),37-45.

Çallı, İ., İşman, A., & Torkul, O. (2002). Sakarya üniversitesi'nde uzaktan eğitimin dünü bugünü ve geleceği. *Sakarya Üniversitesi Eğitim Fakültesi Dergisi*, *3,1-7*.

Carro, R. M., Pulido, E., & Rodríguez, P. (2000). Adaptive internet-based learning with the Tangow system: *Computers and Education in the 21st Century* (pp. 127–135). Dordrecht:Springer. doi:10.1007/0-306-47532-4_12

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391.

Dessus, P., Lemaire, B., & Vernier, A. (2000). Free-text assessment in a virtual campus. *Proc. 3rd International Conference on Human System Learning (CAPS'3)* (pp. 61–76). Paris, France:Learning's W.W.W.

Epignosis, L. L. C. (2014). E-learning concepts, trends, applications. [*Version 1.1]*. Retrieved from https://www.talentlms.com/elearning/elearning-101-jan2014-v1.1.pdf

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*(2–3), 285–307.doi: 10.1080/01638539809545029

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. *EdMedia: World Conference on Educational Media and Technology* (pp. 939–944). Seattle, WA USA:Association for the Advancement of Computing in Education (AACE).

Gelbal, S., & Kelecioğlu, H. (2007). Öğretmenlerin ölçme ve değerlendirme yöntemleri hakkındaki yeterlik algıları ve karşılaştıkları sorunlar. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 33, 135–145.

Hotho, A., Nurnberger, A., & Paas, G. (2005). A brief survey of text mining. *LDV Forum-GLDV Journal for Computational Linguistics and Language Technology*, *20*(1),19–62.

Internet World Stats. (2017). Internet usage statıstıcs the ınternet big picture world ınternet users and 2017 population stats. Retrieved November 24, 2017, from https://www.internetworldstats.com/stats.htm

Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *2*(2), 10. doi: 10.1145/1376815.1376819

İşman, A. (2005). *Uzaktan eğitim*. Ankara: Öğreti Yayınları.

Kanejiya, D., Kumar, A., & Prasad, S. (2003). Automatic evaluation of students' answers using syntactically enhanced LSA. *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2* (pp. 53–60). Stroudsburg, PA USA:Association for Computational Linguistics. doi:10.3115/1118894.1118902

Kaya, Z. (2002). *Uzaktan eğitim*. Pegem A Yayıncılık.

Koçdar, S., & Görü Doğan, T. (2015). Türkiye'deki açık ve uzaktan öğrenme programlarının bir analizi: Eğilimler ve öneriler. *Eğitim ve Öğretim Araştırmaları Dergisi*, *4*(4), 23–36.

Landauer, T. K. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 295–308.doi: 10.1080/0969594032000148154

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389–405. doi: 10.1023/A:1025779619903

MEB. (2017). TEOG açık uçlu soru örnekleri. Retrieved August 31, 2017, from http://abide.meb.gov.tr/ornek-sorular.asp

Noorbehbahani, F., & Kardan, A. A. (2011). The automatic assessment of free text answers using a modified BLEU algorithm. *Computers & Education*, *56*(2), 337–345. doi: 10.1016/j.compedu.2010.07.013

ÖSYM. (2017). LYS açık uçlu soru örnekleri. Retrieved August 31, 2017, from http://www.osym.gov.tr/TR,12909/2017-lisans-yerlestirme-sinavlari-2017-lys-acik-uclu-sorular-hakkinda-bilgilendirme-ve-acik-uclu-soru-ornekleri-05012017.html

Pérez-Marín, D., Alfonseca, E., & Rodríguez, P. (2006). On the dynamic adaptation of computer assisted assessment of free-text answers. *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 374–377). Berlin:Springer.doi: 10.1007/11768012_54

Perez-Marin, D., Pascual-Nieto, I., Alfonseca, E., Anguiano, E., & Rodriguez, P. (2007). A study on the impact of the use of an automatic and adaptive free-text assessment system during a university course. *Workshop on Blended Learning*, (pp. 186-195). Edinburgh, United Kingdom: The Hong Kong Web Society

Pérez, D., Alfonseca, E., & Rodríguez, P. (2004). Application of the bleu method for evaluating free-text answers in an e-learning environment. *The International Conference on Language Resources and Evaluation* (pp. 1351-1354) .Lisbon, Portugal: European Language Resources Association

Roy, S., Narahari, Y., & Deshmukh, O. D. (2015). A perspective on computer assisted assessment techniques for short free-text answers. *International Computer Assisted Assessment Conference* (pp. 96–109). Cham:Springer. doi:10.1007/978-3-319-27704-2_10

Scalise, K., & Gifford, B. (2006). A framework for constructing "Intermediate Constraint" questions and tasks for technology platforms computer-based assessment in E-learning. *The Journal of Technology, Learning, and Assessment*, *4*(6),1-45.

Wiemer-Hastings, P., & Zipitria, I. (2001). Rules for syntax, vectors for semantics. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Retrieved from https://escholarship.org/uc/item/057457h4