Data Sorting and Analysis With K-Means to Evaluate Language Efficiency of Teacher Trainees

Abirami Kanagarajan [1] ,School of Arts, Sciences, Humanities and Education, SASTRA Deemed University, Thanjavur, Email:abirami@src.sastra.edu

Subha S [2] , Department of English, M Kumarasamy College of Engineering, Karur, Email:subhaprasath2012@gmail.com

Vijayakumar M [3] ,School of Social Sciences and Languages, Vellore Institute of Technology, Vellore, Tamil Nadu, India -632014, Email:vijayakumar.muthu@vit.ac.in

Anu Baisel [4] ,School of Social Sciences and Languages, Vellore Institute of Technology, Vellore, Tamil Nadu, India – 632014, Email: anu.baisel@vit.ac.in

B. Mariappan [5] ,Department of English, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu, India,Email:marssrec@gmail.com

Abstract. The present paper analyses the English language skills of final year teacher trainees of three multiple branches of under graduate course in the delta regions of Tamil Nadu, with reference toKumbakonam. Whatever is the major programme of a student communication skills and proficiency in English is considered to be a minimum qualifying criterion for any candidate who aspire to become a teacher.The presentresearch analysed the English proficiency of teacher trainees by conducting a test on Cloze passage by usingthe K-means clustering technique indata mining and analyzed the results to predict the performance of various students and the factors that influence their English proficiency. The current study tookthe school environment of the student as a crucialinfluencing element when they learn and acquire the English language as language cannot be learnt in isolation. The research proved the correlation of language with the learning environment of the teacher trainees from primary school till secondary education,and on the basis of thetrainees's score in the cloze test conducted at three different difficulty levels.

INTRODUCTION

English language proficiency is required to excel in any profession and it is inevitable for a teacher in the era of globalization and cultural merge. Students across the country come to study in a place based on their merit and cultural exchange takes place. But it requires a common medium to transact what one mean and helps the flow of communication till one get used to another person's culture. Communication not only means speaking but is a combination of all the four skills – Listening, Speaking, Reading and Writing. The present study investigates the correlation among the trainees study environment, their place, their board, their parent's educational qualification, their medium of instruction and their location of schools when they were studying from Grade one to Grade twelve at different junctures. The study revealed poignant relationshipsexist in acquiring language proficiency in relation to the students' educational background and the results have been reported.

LITERATURE REVIEW

The research article by SangitaOswal proposed that an intelligent evaluation method based on clustering is evolved and it could be applied to mine different groups of teachers and evaluate their quality of teaching automatically. They proposed a model to evaluate teaching quality using K-Means clustering. The paper by Ji Lixia estimated the importance of data mining in education and Ji Lixia and et al. introduced common data mining algorithms and their applications in educational data mining. Palwinder Kaur Mangat and Dr. Kamaljit Singh Saini in their article discusses the process of cleaning the data for accurate results to predict syudents' success rate. Tomkins and et. Al. predicts the success of students who take online MOOC courses by proposing a general socio- modeling framework.

MATERIALS AND METHODS

The students taken as sample were at the final year of their post graduation course. They have learnt English as their second language for almost twelve years at school, and have done four courses in English Communication at their under graduate level. The present paper scrutinized their proficiency in English language. The present is conducted to know the level of their proficiency in English language and to predict their placement ration as they write aptitude in English and also attempt other tests that have English included in the test syllabus. The selected students belong to Maths, English and Computer Science streams. They were given a worksheet that contained questions pertaining to their educational and demographic profile was registered. Their performance in the Clozetest was analyzed using K-means clustering method and the papers are evaluated. The final clustering helped to predict therespective students' proficiency of English to help them to be aware of their language to get placement in better companies.

THE CLOZE PROCEDURE

Among the other procedures available in English like reading comprehension, paraphrasing, summarizing, The Cloze test is the most comprehensive test which helps to determine the language skills of a teacher trainee. There are many methods available to test a student in Cloze writing but Taylor's cloze procedure is an efficient tool to test thelinguistic competence. To define the term cloze, the word'Cloze' is derivedfrom the domain psychology where it remained as'clozure', which was propounded by Gestalt. A passage with blanks will be given and the testee is expected to complete the blank with the appropriate word whci was omitted and given as a blank. Usually, in a cloze test, every nth word will be removed and given as a blank and the examinee is to accurately write the removed language component using context clues.

WHY THE CLOZE TEST?

Cloze test shows a concrete method of measurability among other written tests as it measures the grammar, spelling, punctuation, cohesion, coherence and deep understanding of concepts. The major advantages in using Clozetest to measure linguistic ability are:

- They are economical,
- easy to administer,
- simple to score/correct,
- Cloze yield valid results
- results are objective and so remain unbiased,
- measures the linguistic competenceat semantic and syntactic level.

HYPOTHESIS

I. The understanding of English passages will not vary among teacher trainees of differentcourses.
II.The Trainee teacherswere able to interpret sample passages from different intermediate sources only at levels below the expected level in placement exams.
III. The understanding capability of sample passages is not influenced by their educational background
IV. The understanding of sample passages is not influenced by their medium of study.
V. The level of understanding of sample passages is not influenced by their board of study.
VI. The level of understanding of sample passages is not influenced by their parents' educational background.

TABLE 1. Survey Data

This test is conducted at three levels - preliminary, intermediate and higher level (inference level) passages with

PROGRAMME	MATRICULATION MARK	HIGHER SECONDARY SCORE	UNDER GRADUATE	POST GRADUATE	LOCATION OF SCHOOL
English	85	77.19	64.11	76.67	URBAN
English	94	85.633	66	77	URBAN
English	92.6	82.56	69.34	79.44	MIXED
English	87.8	67.85	62.41	67.22	URBAN
Maths	97	84	74.5	48.5	URBAN
Maths	84.2	71.5	82	50.5	URBAN
Maths	89	72.75	82	59	URBAN
Maths	93.6	86.67	75.16	69	URBAN
Maths	94.8	84.4	74.77	65	RURAL

blanks. The teacher trainees are instructed to read the entire passage and to write the appropriate word that would fill the blanks. Hence, the skill of a learner to fill the exact word depends on their ability to the context clues which develops in a due course by constant exposure to good language in all the four major skills, LSRW.

DESCRIPTION

The study was conducted for 27 student teacher trainees of English, Mathematics and Computer Science Course. Each question paper contained a passage with 20 blanks, and they have to take three levels of test. These three levels are categorized based on the level of the cloze passage as easy, intermediate and higher level passages. The demographic profile of the examinees were received and it had nine attributes. Theattributes are Department, Medium of instruction at elementary level, Medium of instruction at matriculation level, Medium of instruction at thepost matriculation level,Matriculation board, Type of post matriculation board, Location of schools at different levels of classes, and Parents Educational background.

TEST ADMINISTRATION

The test was administered to final year students of Five-year integrated progamme teacher trainees of three different programmes and they were in the tenth semester of their programme. The duration ofthe test was 60 minutes. Instructions to fill demographic profile isgiven. It is to be noted that other than the English major students, Maths and Computer science students had no familiarity about the Cloze test. The instructions for the test were read to the participants before the start of the test.Then, the teacher trainees answered the question paper. The trainees were under supervision during the test session to ensure credibility and to clarify any doubts.

EXPERIMENTAL RESULTS

The proposed model gave the following results. The K-Means clustering algorithm is used to find out the level of the teacher Trainees'linguistic competence.Figure 1 showed the teacher traineesacademic aspects in collaboration with different attributes using learning tool WEKA. Figure 2 showed the various clusters of students. The association rules are generated out of the clustering behavior. Among the results, four rules were generated and shown below:

(i) The first rule exhibits the poor performance of the student.
(ii) The second rule shows the average performance of the student.
(iii) The third rule categories the good performance in trainees
(iv)The fourth rule exhibits the very good performance

These patterns of different kind were generated using the machine learning tool WEKA.They were in turnconverted as predication rules against the respective types of students' performance as poor, average, good and very good.

RESULTS AND DISCUSSION

The results arrived after evaluating the scripts were interpreted using the maching learning tool WEKA and the clusters were formed to make predictions. The first hypothesis is that the degrees of understanding of English passages do not vary among teacher trainees of different courses. This is proved from the results which tell that the students of all the three streams will perform only at the below average level. The second hypothesis is that the trainee teachers were able to interpret sample passages from different intermediate sources only at levels below the expected level. This is proved that in separation of from their designed syllabus, their practice in general reading is minimum and they could not excel at language tests. The next hypothesis is the level of understanding of sample passages is not influenced by their educational background and this hypothesis is not true as the students from English medium performed better than the others and their score is better than students who are from regional medium. The fourth hypothesis stated that the level of understanding of sample passages is not influenced by their medium of study. This again is found to be false as the students from regional medium scored less than English medium students. The fifth hypothesis is about The level of understanding of sample passages and it is said that it is not influenced by their board of study. But the students from the matriculation board and central board performed well. The final hypothesis stated that the level of understanding of sample passages is not influenced by their parents' educational background. But the students who have educated parents have better language skills.

CONCLUSION

Textbooks in English play an important role in knowledge transfer but English language learning is a skill. The quality of textbooks and the understanding level from the primary classes till the final year ofpost-graduation influence the results of cloze procedure. Textbooks with a higher level linguistic structure will have less educational value as the students have different environment at home and school. The purpose of this research is to introduce a psycholinguistic technique, the Cloze Procedure, to trainee teachers and to demonstrate how trainee teachers can effectively use this technique to improve theirEnglish language competence. The evaluation of the test in correlation with the demographic profile made clear that the Cloze Procedure can be used to test higher level of languagecompetence. One advantage of Cloze test is its cost effectiveness and objectivity in giving scores and the wide coverage of language aspects needed to get success in communication,placement exams and other arenas where English is predominantly in use.

REFERENCES

1. H. Adelberg, "The accounting syntactic complexity formula: a new instrument for predicting the readability of selected accounting communications," Accounting and Business Research **13**, 163–175 (1983).
2. E. Sambasivan and S. Gunapalan, "An analysis of factors influencing in social media and e-communication among younger generation," (2015).
3. S. NO, "School of engineering and technology sripadmavathimahilavisvavisyalayam (scheme of instruction and evaluation of b. tech (ece) department of electronics and communication engineering i year–i semester (2016-17),".
4. L. Ji, X. Zhang, and L. Zhang, "Research on the algorithm of education data mining based on big data," in 2020 IEEE 2nd International Conference on Computer Science and Educational Informatization (CSEI) (IEEE, 2020) pp. 344–350.
5. Patel and R. Day, "The influence of cognitive style on the understandability of a professional accounting pronouncement by accounting students," The British Accounting Review **28**, 139–154 (1996).
6. Palwinder Kaur Mangat,Dr. Kamaljit Singh Saini. (2020). Educational Data Mining Tools and Framework for Predicting Students Academic Performance. International Journal of Advanced Science and Technology, 29 (10s), 2525-2533. Retrieved From: http://sersc.org/journals/index.php/IJAST/article/view/16915
7. E. Shannon, The mathematical theory of communication, by CE Shannon (and recent contributions to the mathematical theory of communication), W. Weaver (University of illinois Press Champaign, IL, USA, 1949).
8. J. Pasek, E. Hargittai, et al., "Facebook and academic performance: Reconciling a media sensation with data," First Monday (2009).
9. G. S. Narayana, M. D. Ansari, V. K. Gunjan, et al., "Instantaneous approach for evaluating the initial centers in the agricultural databases using k-means clustering algorithm," Journal of Mobile Multimedia, 43–60 (2022).
10. M. Sánchez-Martínez and A. Otero, "Factors associated with cell phone use in adolescents in the community of madrid (spain)," CyberPsychology& Behavior **12**, 131–137 (2009).
11. S. Tomkins, L. Getoor, et al., "Understanding hybrid-mooc effectiveness with a collective socio-behavioral model," Journal of Educational Data Mining **11**, 42–77 (2019).

12. W. W. Weaver and A. J. Kingston, "A factor analysis of the cloze procedure and other measures of reading and language ability." Journal of Communication (1963).
13. J. L. Martin and K.-T. Yeung, "Persistence of close personal ties over a 12-year period," Social Networks **28**, 331–362 (2006).