Using Psycholinguistic Features for the Classification of Comprehenders from Summary Speech Transcripts

Santosh Kumar Barnwal(✉) and Uma Shanker Tiwary

Indian Institute of Information Technology, Allahabad, India
iis2009002@gmail.com

Abstract. In education, some students lack language comprehension, language production and language acquisition skills. In this paper we extracted several psycholinguistics features broadly grouped into lexical and morphological complexity, syntactic complexity, production units, syntactic pattern density, referential cohesion, connectives, amounts of coordination, amounts of subordination, LSA, word information, and readability from students' summary speech transcripts. Using these Coh-Metrix features, comprehenders are classified into two groups: poor comprehender and proficient comprehender. It is concluded that a computational model can be implemented using a reduced set of features and the results can be used to help poor reading comprehenders for improving their cognitive reading skills.

Keywords: Psycholinguistics · Natural language processing
Machine learning classification

1 Introduction

Reading is a complex cognitive activity where learners read texts to construct a meaningful understanding from the verbal symbols i.e. the words and sentences and the process is called as reading comprehension. In Reading process, the three main factors - the learner's context knowledge, the information aroused by the text, and the reading circumstances together construct a meaningful discourse. Previous researches claim that in academic environment several reading and learning strategies including intensive reading and extensive reading [2], spaced repetition [7] and top-down and bottom-up processes [1] play vital role in students developing comprehension skills.

Intensive Reading: It is the more common approach, in which learners read passages selecting from the same text or various texts about the same subject. Here, content and linguistic forms are repeated themselves, therefore learners get several chances to comprehend the meaning of the textual contents. It is usually classroom based and teacher centric approach where students concentrate on linguistics, grammatical structures and semantic details of the text to retain in

P. Horain et al. (Eds.): IHCI 2017, LNCS 10688, pp. 122–136, 2017.
https://doi.org/10.1007/978-3-319-72038-8_10

memory over a long period of time. Students involve themselves in reading passages carefully and thoroughly again and again aiming to be able translating the text in a different language, learning the linguistic details in the text, answering comprehension questions such as objective type and multiple choice, or knowing new vocabulary words. Some disadvantages are - (a) it is slow, (b) needs careful reading of a small amount of difficult text, (c) requires more attention on the language and its structure, including morphology, syntax, phonetics, and semantics rather than the text, (d) text may be bored to students since it was chosen by the teacher, and (e) because exercises and assessments are part of comprehension evaluation, students may involve in reading only for the preparation for a test and not for getting any pleasure.

Extensive Reading: On the other hand, extensive reading provides more enjoyments as students read big quantities of own interest contents; focus on to understand main ideas but not on the language and its structure, skipping unfamiliar and difficult words and reading for summary [12]. The main aim of extensive reading is to learn foreign language through large amounts of reading and thus building student confidence and enjoyment. Several Research works claim that extensive reading facilitating students improving in reading comprehension to increase reading speed, greater understanding of second language grammar conventions, to improve second language writing, and to motivate for reading at higher levels [10].

The findings of previous researches suggest that extensive and intensive reading approaches are beneficial, in one way or another, for improving students' reading comprehension skills.

Psycholinguistic Factors: Psycholinguistics is a branch of cognitive science in which language comprehension, language production and language acquisition are studied. It tries to explain the ways in which language is represented and is processed in the brain; for example, the cognitive processes responsible for generating a grammatical and meaningful sentence based on vocabulary and grammatical structures and the processes which are responsible to comprehend words, sentences etc. Primary concerned linguistic related areas are: Phonology, morphology, syntax, semantics, and pragmatics. In this field, researchers study reader's capability to learn language for example, the different processes required for the extraction of phonological, orthographic, morphological, and semantic information by reading a textual document.

More recent work, Coh-Metrix [5] offers to investigate the cohesion of the explicit text and the coherence of the mental representation of the text. This metrix provides detailed analysis of language and cohesion features that are integral to cognitive reading processes such as decoding, syntactic parsing, and meaning construction.

2 Brief Description of Coh-Metrix Measures

Coh-Metrix is an automatic text analysis tool forwarding traditional theories of reading and comprehension to next higher level and therefore, can plays important role in different disciplines of education such as teaching, readability, learning etc. The tool analyses and measures features of texts written in English language through hundreds of measures, all informed by previous researchers in different disciplines such as computational linguistics, psycholinguistics, discourse processes and cognitive sciences. The tool integrates several computational linguistics components including lexicons, pattern classifiers, part-of-speech taggers, syntactic parsers, semantic interpreters, WordNet, CELEX Corpus etc. Employing these elements, Coh-Metrix can analyze texts on multi levels of cohesion including co-referential cohesion, causal cohesion, density of connectives, latent semantic analysis metrics, and syntactic complexity [5].

All measures of the tool have been categorized into following broad groups:

1. **Descriptive measures:** These measures describe statistical features of text in form of total number of paragraphs, total number of sentences, total number of words, average length of paragraphs with standard deviation, average number of words with standard deviation, mean number of syllables in words with standard deviation etc.
2. **Easability components:** For measuring text easability score, the tool provides several scores including text narrativity, syntactic familiarity, and Word Concreteness.
3. **Referential Cohesion:** It is a linguistic cue that helps readers in making connections between different text units such as clauses, and sentences. It includes Noun overlap (words overlap in terms of noun), and Argument overlap (sentences overlap in terms of nouns and pronouns). Coh-Metrix measures semantically similar pairs such as car/vehicle etc.
4. **Latent Semantic Analysis:** It is used to implement semantic co-referentiality for representing deeper world knowledge based on large corpora of texts.
5. **Lexical Diversity:** It is the variety of unique words (types) in a text in relation to number of words (tokens). It refers to variation of Type-token ratio (TTR).
6. **Connectives:** It provides clues about text organization and aid reader in the creation of cohesive links between ideas and clauses. It measures the cohesive links between different conceptual units using different types of connectives such as causal (because, so), logical (and, or), adversative (whereas), temporal (until) and additive (moreover). In addition to this, there is a difference between positive connectives (moreover) and negative connectives (but).
7. **Situation Model:** It refers to the level of reader's mental representation for a text when a given context is activated.
8. **Syntactic Complexity:** It is measured using NP density, mean number of high-level constituents per word, and the incidence of word classes that indicate analytical difficulty (e.g. and, or, if-then, conditionals).
9. **Syntactic Pattern Density:** It refers to the density of particular syntactic patterns, word types, and phrase types. The relative density of noun phrases, verb phrases, adverbial phrases, and prepositions can affect processing difficulty of text, especially with respect to other features in a text.

10. **Word Information:** It provides density scores for various parts of speech (POS), including pronouns, nouns, verbs, adjectives, adverbs, cardinal numbers, determiners, and possessives.
11. **Readability:** It provides the readability formulas of Flesch Reading Ease and Flesch-Kincaid Grade Level [4,9]. Both are two readability tests designed to indicate how difficult a passage in English is to understand. These tests use word length and sentence length as core measures; however they have different weighting factors.

The aim of present work is to identify the linguistic features that can classify students into two groups - students having proficient comprehension skills and students with poor comprehension skills from their summary speech transcripts.

3 Participants and Method

A brief description of the participants, materials, and procedure that we used in this study are described here.

Participants: Twenty undergraduate students (mean age (SD)- 21.4(0.86)) in information technology major; studied in same batch and performed all academic activities only in English, whereas their primary languages were different; participated in this experimental sessions. Students were told that they would be awarded some course credits for participating in the research. Based on their academic performance in last four semesters, these students were divided into two groups - ten as proficient and others as poor comprehenders.

Materials: The reading materials consisted of two passages. One passage (total sentences- 38, total words- 686, sentence length (mean)- 18.0, Flesch-Kincaid Grade level- 13.3) had been selected from students' course book whereas other was a simple interesting story (total sentences- 42, total words- 716, sentence length (mean)- 17.0, Flesch-Kincaid Grade level- 3.9). Both passages were written in English and were unread until the experiment began. Reading story passage was simulated extensive reading experience and reading course passage was simulated intensive reading experience.

Procedure: All experimental sessions were held in a research lab in a set of 5 students. The experiment consisted of two tests. In each test, student had instructed to read a given passage and then to solve a puzzle and lastly to tell summary as much detail as they can. Both tests were similar except the reading material - the story passage was given in first test and the course passage was given in second test. Students were informed to read the passage on computer screen as they would normally read. The speech were recorded using a digital audio recorder software installed in the computer system. The puzzle task was useful to erase students' short term memory of read text to ensure that the summary would come from their long term memory.

4 Feature Analysis

Feature Extraction: The recorded audio files were transcripted in English where brief pauses were marked with commas, while long pauses were marked with full stops (end of sentence) if their places were according to semantic, syntactic and prosodic features. Repetitions, incomplete words and incomprehensible words were not included in transcription. In the experiment, two sets of transcripts were generated - (a) **story transcripts** had texts of story summary audio files and (b) **course transcripts** had texts of course summary audio files. Both sets had twenty texts, ten of proficient comprehenders' audio files and the other ten of poor comprehenders' audio files.

For analysing the texts of both sets of transcripts, we used the computational tool Coh-Metrix. Coh-Metrix 3.0 (http://cohmetrix.com) provided 106 measures; which were categorized into eleven groups as described in Sect. 2.

Feature Selection: In machine learning classifiers including too many features may lead to overfit the classifier and thus resulting in poor generalization to new data. So, only necessary features should be selected to train classifiers.

We applied two different approaches for the selection of necessary features improving the accuracy of the classifiers.

Approach-1: Coh-Metrix provides more than hundreds of measures of text characteristics and several of them are highly correlated. For example, Pearson correlations demonstrated that *z score of narrativity* was highly correlated ($r = 0.911, p < 0.001$) with *percentile of narrativity.* Of 106 measures of the tool, 52 variables were selected on the basis of two criteria. First, all such variables which had high correlations with other variables ($|r| \geq 0.80$) were discarded for handling the problem of collinearity. Remaining measures were grouped in feature sets. Thus, after removing all such redundant variables, the feature set of story transcripts had 65 measures whereas the feature set of course transcripts had 67 measures. In Table 1, superscripts 1, 2 and 3 indicate measures presented in only story transcripts, in only course transcripts and in both transcripts respectively. Therefore, in first step, measures indicated with superscripts 1 and 3 were selected for the classification of story transcripts; whereas measures indicated with superscripts 2 and 3 were selected to classify the course transcripts. In next step, we had selected only those measures which were presented in both feature sets. Therefore, in second step, 52 common measures indicated with superscript 3 in Table 1, were selected for the classifications.

Pairwise Comparisons: Pairwise comparisons were conducted to examine differences between proficient comprehenders' text and poor comprehenders' text of both sets of transcripts (story and course). These results are reported below.

1. Descriptive measures: Co-Metrix provided eleven descriptive measures in which six measures were selected as features. Paragraph count, Paragraph length, Sentence length and Word length had significant difference between

Table 1. A comparison of proficient and poor comprehenders' transcripts features. Values shown are mean (standard deviation).

Description	Story transcript		Course transcript	
	Proficient comprehender	Poor comprehender	Proficient comprehender	Poor comprehender
1. Descriptive				
[3]Paragraph count	19.7(5.47)	10.5(4.03)	10.9(2.64)	6.2(2.09)
[3]Paragraph length (μ)	1.202(0.14)	1.088(0.11)	1.405(0.26)	1.239(0.23)
[3]Sentence length (μ)	20.21(3.76)	16.98(3.05)	22.52(5.19)	17.45(5.10)
[3]Sentence length (SD)	10.84(2.42)	9.461(2.33)	10.79(2.92)	8.443(3.25)
[3]Word length (μ, syllables)	1.215(0.04)	1.196(0.04)	1.596(0.05)	1.568(0.08)
[1]Word length (μ, letters)	3.841(0.13)	3.750(0.12)	4.834(0.17)	4.753(0.24)
2. Text Easability Principle Component Scores				
[3]Narrativity(z score)	1.779(0.59)	1.840(0.50)	−0.00(0.46)	−0.01(0.57)
[3]Syntactic simplicity (z score)	−0.04(0.55)	0.072(0.60)	−0.51(0.83)	−0.55(0.71)
[3]Word concreteness (z score)	0.781(0.82)	0.409(1.14)	−0.26(0.82)	0.118(1.36)
[3]Referential cohesion(z score)	2.452(0.82)	3.274(1.58)	1.146(0.95)	1.316(1.06)
[3]Referential cohesion(percentile)	98.29(1.51)	97.55(4.52)	80.46(12.5)	81.97(14.5)
[3]Deep cohesion (z score)	1.590(1.05)	3.188(2.01)	−0.24(0.70)	1.308(2.38)
[1]Deep cohesion (percentile)	87.49(15.0)	95.00(9.63)	41.24(24.3)	70.75(43.1)
[3]Verb cohesion (z score)	0.225(0.93)	1.066(1.18)	0.714(1.15)	1.582(1.35)
[3]Connectivity (z score)	−4.07(1.79)	−3.81(1.40)	−3.52(0.86)	−3.90(0.53)
[1]Connectivity (percentile)	1.392(3.45)	1.062(2.25)	0.169(0.24)	0.014(0.02)
[3]Temporality (z score)	0.443(0.55)	1.039(0.88)	−0.05(1.30)	0.787(0.90)
3. Referential Cohesion				
[3]Noun overlap, adjacent sent. (μ)	0.555(0.14)	0.509(0.33)	0.535(0.18)	0.557(0.24)
[3]Argument overlap, adj. sent. (μ)	0.681(0.08)	0.653(0.30)	0.727(0.16)	0.686(0.16)

(*continued*)

Table 1. (*continued*)

Description	Story transcript		Course transcript	
	Proficient comprehender	Poor comprehender	Proficient comprehender	Poor comprehender
[2]Noun overlap, all sent. (μ)	0.528(0.13)	0.416(0.29)	0.415(0.14)	0.400(0.15)
[2]Argument overlap, all sent. (μ)	0.681(0.07)	0.566(0.24)	0.583(0.16)	0.488(0.13)
[2]Stem overlap, all sent. (μ)	0.572(0.11)	0.504(0.28)	0.526(0.13)	0.522(0.14)
[2]Word overlap, adjacent sent. (μ)	0.192(0.05)	0.231(0.08)	0.147(0.05)	0.149(0.03)
[3]Word overlap, adjacent sent.(SD)	0.139(0.02)	0.141(0.04)	0.117(0.04)	0.133(0.04)
[2]Word overlap, all sentences (μ)	0.184(0.03)	0.190(0.06)	0.107(0.03)	0.112(0.02)
[3]Word overlap, all sentences (SD)	0.155(0.01)	0.159(0.02)	0.107(0.02)	0.125(0.02)
4. LSA				
[2]LSA overlap, adjacent sent. (μ)	0.356(0.05)	0.382(0.16)	0.282(0.06)	0.261(0.10)
[3]LSA overlap, adjacent sent.(SD)	0.218(0.04)	0.197(0.06)	0.170(0.03)	0.182(0.07)
[3]LSA overlap, all sent. (μ)	0.271(0.14)	0.232(0.21)	0.256(0.15)	0.175(0.19)
[3]LSA overlap, all sent. (SD)	0.180(0.10)	0.066(0.14)	0.168(0.10)	0.103(0.16)
[2]LSA overlap, adj. paragraph (μ)	0.409(0.03)	0.391(0.16)	0.321(0.08)	0.330(0.13)
[2]LSA overlap, adjacent para. (SD)	0.219(0.02)	0.195(0.06)	0.143(0.04)	0.127(0.04)
[3]LSA, sentence (μ)	0.440(0.04)	0.390(0.10)	0.351(0.03)	0.274(0.09)
[1]LSA, sentence (SD)	0.180(0.02)	0.200(0.04)	0.143(0.01)	0.159(0.05)
5. Lexical Diversity				
[3]Lexical diversity (MTLD)	41.25(3.99)	38.53(8.87)	44.70(12.9)	40.63(10.0)
[3]Vocabulary Diversity (VOCD)	44.08(5.99)	30.10(18.9)	55.21(13.2)	35.30(20.5)
6. Connectives				
[1]All connectives	132.8(21.3)	147.8(28.3)	96.39(17.0)	125.0(33.0)
[3]Adversative and contrastive conn.	12.91(8.25)	13.11(8.02)	19.03(7.52)	16.17(10.1)

(*continued*)

Table 1. (*continued*)

Description	Story transcript		Course transcript	
	Proficient comprehender	Poor comprehender	Proficient comprehender	Poor comprehender
[3]Temporal connectives	34.95(12.3)	42.72(12.9)	13.53(6.16)	18.49(18.1)
[3]Expanded temporal connectives	28.45(9.37)	30.83(15.5)	3.402(4.34)	9.725(12.2)
7. Situation Model				
[3]Causal verb (CV) incidence	19.97(5.74)	24.65(10.8)	24.19(9.96)	27.42(15.0)
[2]Causal particles (CP) incidence	37.38(14.2)	50.65(20.0)	31.42(8.95)	43.08(12.1)
[3]Intentional verbs (IV) incidence	34.20(6.66)	35.94(11.0)	12.43(4.16)	17.94(12.7)
[2]Ratio of CP to CV	0.811(0.52)	1.144(1.30)	0.333(0.28)	0.941(1.80)
[2]Ratio of intentional particle to IV	0.694(0.40)	0.961(0.84)	1.383(0.81)	2.116(2.00)
[3]LSA verb overlap	0.070(0.03)	0.085(0.04)	0.130(0.07)	0.119(0.08)
[3]WordNet verb overlap	0.679(0.05)	0.581(0.17)	0.448(0.11)	0.477(0.24)
8. Syntactic Complexity				
[3]Words before main verb (μ)	3.999(0.94)	3.860(1.06)	4.814(2.26)	3.371(1.87)
[3]Numbers of modifiers (μ)	0.652(0.12)	0.523(0.11)	0.867(0.17)	0.803(0.18)
[3]Sentence syntax similarity (μ)	0.110(0.03)	0.083(0.02)	0.086(0.02)	0.079(0.03)
9. Syntactic Pattern - Phrase Density (PD)				
[3]Noun PD, incidence	318.3(17.5)	354.0(20.9)	377.0(22.0)	366.7(30.7)
[3]Verb PD incidence	258.4(18.9)	249.6(20.3)	194.2(25.0)	204.7(46.8)
[3]Adverbial PD incidence	52.71(19.3)	41.99(15.4)	28.08(11.9)	19.79(13.4)
[3]Preposition PD incidence	92.28(20.2)	87.94(24.3)	115.4(21.1)	122.6(34.3)
[3]Agentless passive voice density	2.972(3.35)	1.96(4.76)	10.06(8.84)	12.08(13.5)
[3]Negation density incidence	19.62(7.96)	26.21(13.6)	11.60(6.82)	9.287(6.47)
[3]Gerund density incidence	17.34(7.35)	13.60(11.6)	9.926(8.20)	15.85(12.3)
[3]Infinitive density, incidence	22.69(9.36)	16.78(12.0)	14.65(6.50)	12.78(12.2)

(*continued*)

Table 1. (*continued*)

Description	Story transcript		Course transcript	
	Proficient comprehender	Poor comprehender	Proficient comprehender	Poor comprehender
10. Word Information				
[3]Noun incidence	194.9(22.8)	211.2(30.5)	262.4(33.9)	268.5(67.1)
[3]Verb incidence	163.1(11.6)	153.7(23.4)	115.1(11.0)	119.2(27.2)
[3]Adjective incidence	27.02(6.50)	21.28(15.5)	70.28(21.2)	56.61(28.9)
[3]Adverb incidence	87.22(20.9)	100.1(14.2)	52.92(23.0)	43.46(19.3)
[1]Pronoun incidence	94.14(23.4)	107.0(23.8)	57.50(17.8)	51.63(28.6)
[1]1st person sing. pronoun in.	3.603(4.20)	4.647(9.26)	0(0)	0(0)
[1]1st person plural pronoun in.	3.250(4.61)	1.817(3.94)	0.808(1.33)	1.612(5.10)
[1]2nd person pronoun incidence	8.069(8.60)	15.82(13.2)	0(0)	0(0)
[2]3rd person singular pronoun in.	62.65(21.5)	59.28(37.2)	14.43(6.86)	15.04(15.7)
[1]3rd person plural pronoun in.	12.19(6.30)	20.25(20.7)	33.11(14.3)	34.98(24.9)
[1]CELEX word frequency (μ)	2.6(0.10)	2.777(0.13)	2.374(0.14)	2.433(0.19)
[3]CELEX Log frequency (μ)	3.319(0.04)	3.349(0.11)	3.155(0.06)	3.212(0.16)
[3]CELEX Log min. frequency (μ)	1.384(0.18)	1.449(0.25)	1.234(0.26)	0.945(0.76)
[3]Age of acquisition for words (μ)	257.6(8.85)	258.7(35.4)	382.9(17.2)	346.7(123.)
[3]Familiarity for words (μ)	569.3(4.59)	572.2(6.70)	574.1(6.57)	577.7(6.66)
[2]Concreteness for words (μ)	395.5(21.7)	383.0(25.0)	362.7(15.0)	356.4(23.8)
[1]Meaningfulness words (μ)	413.2(7.03)	399.4(14.3)	422.7(13.7)	414.2(17.9)
[3]Polysemy for words (μ)	4.563(0.44)	4.610(0.28)	3.887(0.39)	3.980(0.46)
[3]Hypernymy for nouns (μ)	6.562(0.36)	6.842(0.98)	6.003(0.34)	5.350(0.77)
[3]Hypernymy for verbs (μ)	1.927(0.14)	1.837(0.20)	1.607(0.12)	1.676(0.22)
[2]Hyper. for nouns and verbs (μ)	1.589(0.12)	1.705(0.36)	1.788(0.10)	1.610(0.21)
11. Readability				
[2]Flesch reading ease	83.49(6.46)	88.38(5.52)	48.89(7.28)	56.43(6.85)

proficient comprehenders' text and poor comprehenders' text of both sets of transcripts.

2. Easability components: The tool provided sixteen easability measures in which eleven measures were selected as features. Deep cohesion, Verb cohesion, Connectivity and Temporality had significant difference between proficient comprehenders' text and poor comprehenders' text of both sets of transcripts.
3. Referential Cohesion: The tool provided ten referential cohesion measures in which nine measures were selected as features. The findings from different overlap measures demonstrated that proficient comprehenders used more *co-referential nouns, pronouns, or NP phrases* than poor comprehenders.
4. Latent Semantic Analysis: The tool provided eight LSA measures and all were selected as features. LSA overlap measures had significant difference between proficient comprehenders' text and poor comprehenders' text of both sets of transcripts.
5. Lexical Diversity: The tool provided four lexical diversity measures in which two measures were selected as features. MTLD and VOCD had more significant difference between proficient comprehenders' text and poor comprehenders' text of both sets of transcripts.
6. Connectives: The tool provided nine lexical connective measures in which four measures were selected as features. The findings from different connective measures demonstrated that proficient comprehenders used more connectives, such as *in other words, also, however, although* etc. than poor comprehenders; whereas poor comprehenders used comparatively more logical operators such as *and, then* etc. as well as more temporal connectives, such as *when* etc.
7. Situation Model: The tool provided eight situation model measures in which seven measures were selected as features. Causal verb measures had significant difference between proficient comprehenders' text and poor comprehenders' text of both sets of transcripts.
8. Syntactic Complexity: The tool provided seven syntactic complexity measures in which three measures were selected as features. Words before main verb (mean), Number of modifiers per noun phrase (mean), and Sentence syntax similarity (mean) had less significant difference between proficient comprehenders' text and poor comprehenders' text of both sets of transcripts.
9. Syntactic Pattern Density: The tool provided eight syntactic pattern density measures and all were selected as features. Noun phrase density, Verb phrase density, Adverbial phrase density, Preposition phrase density, Agentless passive voice density, Negation density, Gerund density, and Infinitive density had high significant difference between proficient comprehenders' text and poor comprehenders' text of both sets of transcripts.
10. Word Information: The tool provided twenty two word information measures in which twenty one measures were selected as features. Noun incidence, Verb incidence, Adjective incidence, and Adverb incidence were highly

significant. Poor comprehenders' transcripts had a comparatively greater proportion of pronouns compared to that of proficient comprehenders.

11. Readability: The tool provided three readability measures in which one measure was selected as feature. Flesch Reading Ease had significant difference between proficient comprehenders' text and poor comprehenders' text of both sets of transcripts.

Table 2. A comparison of proficient and poor comprehenders' features extracted from story transcripts. Values shown are mean (standard deviation).

Description	Proficient comprehender (Story transcript)	Poor comprehender (Story transcript)	p-value < 0.05
Descriptive			
Number of paragraphs	19.7(5.47)	10.5(4.03)	0.001
Number of sentences	23.1(4.70)	11.7(5.37)	0.00
Number of words	453.6(59.1)	197.8(86.4)	0.00
Number of sentences in a paragraph (SD)	0.383(0.17)	0.203(0.19)	0.041
Deep cohesion (z score)	1.590(1.05)	3.188(2.01)	0.044
Lexical Diversity			
Type-token ratio (all words)	0.318(0.02)	0.422(0.11)	0.022
Lexical diversity	44.08(5.99)	30.10(18.9)	0.049
Connectives			
Logic connectives	52.51(14.8)	81.49(20.8)	0.002
Syntactic Complexity			
Mean number of modifiers per noun-phrase	0.652(0.12)	0.523(0.11)	0.029
Minimum editorial distance score for words	0.758(0.26)	0.433(0.39)	0.049
Minimum editorial distance score for lemmas	0.738(0.26)	0.407(0.37)	0.035
Syntactic Pattern Density			
Noun phrase density	318.3(17.5)	354.0(20.9)	0.001
Word Information			
Average word frequency for content words	2.6(0.10)	2.777(0.13)	0.004
Meaningfulness content words (mean)	413.2(7.03)	399.4(14.3)	0.017
Readability			
Second language readability score	29.53(3.27)	33.88(4.30)	0.021

Approach-2: In this approach, we selected appropriate features from all 106 Coh-Metrix measures by applying Welch's two-tailed, unpaired t-test on each measure of both types of comprehenders' transcripts. All features that were significant at $p < 0.05$ were selected for classification. Thus, the feature set of story transcripts had 15 measures (Table 2) whereas the feature set of course transcripts had 14 measures (Table 3).

Table 3. A comparison of proficient and poor comprehenders' features extracted from course transcripts. Values shown are mean (standard deviation).

Description	Proficient comprehender (Course transcript)	Poor comprehender (Course transcript)	p-value < 0.05
Descriptive			
Number of paragraphs	10.9(2.64)	6.2(2.09)	0.00
Number of sentences	15.5(5.40)	7.7(2.90)	0.001
Number of words	336.9(101.0)	133.3(52.1)	0.00
Sentence length(mean)	22.52(5.19)	17.45(5.10)	0.041
LSA			
Latent Semantic Analysis (mean)	0.351(0.03)	0.274(0.09)	0.028
Lexical Diversity			
Type-token ratio (content word lemmas)	0.618(0.07)	0.741(0.11)	0.009
Type-token ratio (all words)	0.425(0.04)	0.552(0.12)	0.012
Lexical diversity	55.21(13.2)	35.30(20.5)	0.021
Connectives			
All connectives, incidence	96.39(17.0)	125.0(33.0)	0.03
Situation Model			
Causal verbs and causal particles incidence	31.42(8.95)	43.08(12.1)	0.026
Word Information			
Hypernymy for nouns (mean)	6.003(0.34)	5.350(0.77)	0.03
Hypernymy for nouns and verbs (mean)	1.788(0.10)	1.610(0.21)	0.038
Readability			
Flesch reading ease	48.89(7.28)	56.43(6.85)	0.028
Flesch-Kincaid grade Level	12.03(2.19)	9.724(1.81)	0.02

5 Classification

We examined several classification methods such as Decision Trees, Multi-Layer Perceptron, Naïve Bayes, and Logistic Regression using Weka toolkit [6]. 10-fold cross-validation method had been applied to train these classifiers. The results of these classifiers are reported in Table 4 in terms of classification accuracy and root mean square error (RMSE). The classification accuracy refers to the percentage of samples in the test dataset that are correctly classified (true positives plus true negatives). Root-mean-square error (RMSE) is frequently used as measure of the differences between values predicted by a classifier and the values expected. In this experiment, it provided the mean difference between the predicted students' comprehension level and the expected comprehension level. The baseline accuracy represents the accuracies that would be achieved by assigning every sample to the larger training size of the two classes. In this experiment, both classes had 10 training samples, therefore, the baseline accuracy for poor vs. proficient comprehenders' transcripts would be achieved by assigning all the samples in any one group and thus the baseline accuracy of the experiment would be 0.5 (10/20 = 0.5).

6 Result and Discussion

Table 4 shows the accuracies for classifying poor vs. proficient comprehenders' transcripts. The classifier accuracies were not as high for approach-1 compared to approach-2; however, they were above or equal to the baseline for all four classifiers. Also, common features provided better accuracies as compared to first

Table 4. Accuracies for the four classifiers.

Feature sets	# Features	Logistic regression	Naïve Bayes	Decision tree	Multi-layer perceptron
Approach-1:					
First Step-					
Feature set (Story transcript)	65	60% (0.63)	60% (0.6)	90% (0.3)	80% (0.39)
Feature set (Course transcript)	67	65% (0.59)	75% (0.46)	50% (0.62)	65% (0.5)
Second Step-					
Common feature (Story transcript)	52	85% (0.4)	80% (0.44)	90% (0.3)	80% (0.38)
Common feature (course transcript)	52	75% (0.49)	85% (0.4)	65% (0.48)	65% (0.49)
Approach-2:					
Story transcript	15	*100%* (0)	*95%* (0.22)	*100%* (0)	*90%* (0.28)
Course transcript	14	*90%* (0.31)	75% (0.41)	*95%* (0.23)	80% (0.44)

step features (story or course feature set). In this experiment, the reduced set of features applied in approach-2, provided best results for all four classifiers. However it was observed that selection of features using approach-2 were dependent on the participants involved in the experiment as well as the read text; whereas the features of approach-1 were almost robust against these changes. The major findings of this study demonstrate that three cohesion indices- lexical diversity, connectives, and word information, common in both Tables 2 and 3, played a vital role in the classification of both types of the transcripts. The logistic regression classifier classified story transcripts and course transcripts with accuracies 100% and 80% respectively.

Generally in first attempt of reading a new text, science and technology course does not help most students to develop mental model to represent the collective conceptual relations between the scientific concepts, due to lack of their prior domain knowledge. In contrast, story texts carry some general schema such as name, specific place and chronological details of an event; all these schema help students to develop mental model by integrating these specific attributes of the event described in the story [11]. Therefore, students stored the mental model of story text comparatively in more details in their memory compared to that of course text; which was reflected in their transcripts. Proficient and poor both students' story transcripts contained more noun phrases in comparison to course transcripts.

Poor comprehenders may not benefit as much as good comprehenders from reading a complex text because grammatical and lexical linking within the text increases text length, density, and complexity. As a consequence, reading such text involves creation and processing of more complex mental model. Comprehenders with low working-memory capacity experience numerous constraints on the processing of these larger mental models, resulting in lower comprehension and recall performance [8]. As a result poor comprehenders' transcripts consist of comparatively more sentences with mixed content representing their confused state of mental models. Therefore, as shown in Table 1, values of the measures of situation model index were more in poor comprehenders' transcripts in contrast to proficients' transcripts.

The finding in this study also validates a previous study [3], which demonstrated that less-skilled comprehenders produced narratives that were poor in terms of both structural coherence and referential cohesion.

In short, the Coh-Metrix analysis of transcripts provides a number of linguistic properties of comprehenders' narrative speech. Comprehension proficiency were characterized by greater cohesion, shorter sentences, more connectives, greater lexical diversity, and more sophisticated vocabulary. It is observed that lexical diversity, word information, LSA, syntactic pattern, and sentence length provided the most predictive information of proficient or poor comprehenders.

In conclusion, the current study supports to utilize Coh-Metrix features to measure comprehender's ability.

References

1. Angosto, A., Sánchez, P., Álvarez, M., Cuevas, I., León, J.A.: Evidence for top-down processing in reading comprehension of children. Psicología Educativa **19**(2), 83–88 (2013)
2. Attaprechakul, D.: Inference strategies to improve reading comprehension of challenging texts. Engl. Lang. Teach. **6**(3), 82–91 (2013)
3. Cain, K.: Text comprehension and its relation to coherence and cohesion in children's fictional narratives. Br. J. Dev. Psychol. **21**(3), 335–351 (2003)
4. Flesch, R.: A new readability yardstick. J. Appl. Psychol. **32**(3), 221 (1948)
5. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-metrix: Analysis of text on cohesion and language. Behav. Res. Methods **36**(2), 193–202 (2004)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
7. Hunt, A., Beglar, D.: A framework for developing EFL reading vocabulary. Read. Foreign Lang. **17**(1), 23 (2005)
8. Kendeou, P., Broek, P., Helder, A., Karlsson, J.: A cognitive view of reading comprehension: Implications for reading difficulties. Learn. Disabil. Res. Pract. **29**(1), 10–16 (2014)
9. Kincaid, J.P., Fishburne Jr., R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch (1975)
10. Mason, B., Krashen, S.: Extensive reading in English as a foreign language. System **25**(1), 91–102 (1997)
11. Ozuru, Y., Dempsey, K., McNamara, D.S.: Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. Learn. Instr. **19**(3), 228–242 (2009)
12. Richards, J.C.: Longman Dictionary of Language Teaching and Applied Linguistics (2000)