**Speech Deterioration in Alzheimer’s disease: Case Study of Agatha Christie**

Statistics show that signs of Alzheimer’s in Agatha Christie can be seen in her works from her 70’s. So I take the works after 70 (from the year 1960) as training set and before 70 (before 1960) as testing set.

**Steps followed by me:-**

1. Collection of texts of Agatha Christie
2. Removing the punctuations and converting it to lower case(cleaning data)
3. Un-stemmed document was used
4. Features extraction
5. Analysis using Weka
6. Used java for coding

**Text features I have used:-**

**Lexical**

1. Unique word count ( using Hashmap for this)
2. Frequency of nouns (POS tag NN and NNS)
3. Frequency of pronouns(POS tag PRP)
4. Frequency of conjunction(denoting coordinated sentences POS tag CC)
5. Frequency of Indefinite terms(thing, something, stuff)
6. Reduced sentences (POS tags VBG and VBN)
7. Interjections (POS tag UH)
8. Total word count
9. Count of Adjectives(POS tag JJ)
10. Count of verbs(POs Tags VB,VBD,VBG,VBN,VBP,VBZ)
11. Total sentences count
12. Normalized sentences count
13. Noun to verb ratio
14. Adverb
15. TTR
16. Brunet’s index
17. Honore’s Statistic
18. Mean length of words
19. Lexical density (ratio of words providing meaning to the text like nouns, adjectives, verb and adverbs to the total number of words, Lexical words give a text its meaning and provide information regarding what the text is about. More precisely, lexical words are simply nouns, adjectives, verbs, and adverbs. Nouns tell us the subject, adjectives tell us more about the subject, verbs tell us what they do, and adverbs tell us how they do it.)
20. Function words(Non- informative words in the text)
21. Use of punctuations
22. Count of trigrams
23. Year

**Machine Learning Models Used**

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. KNN
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms
10. Gradient Boost & Adaboost

**Results from the Linear regression Model**

1. UNIQUE\_WORD = 1.5359 \* NOUNS +

-5.1076 \* PRONOUN +

18.0233 \* INDEFINITE\_TERMS +

1.9614 .

Correlation coefficient 0.969

Mean absolute error 4.3384

Root mean squared error 5.4776

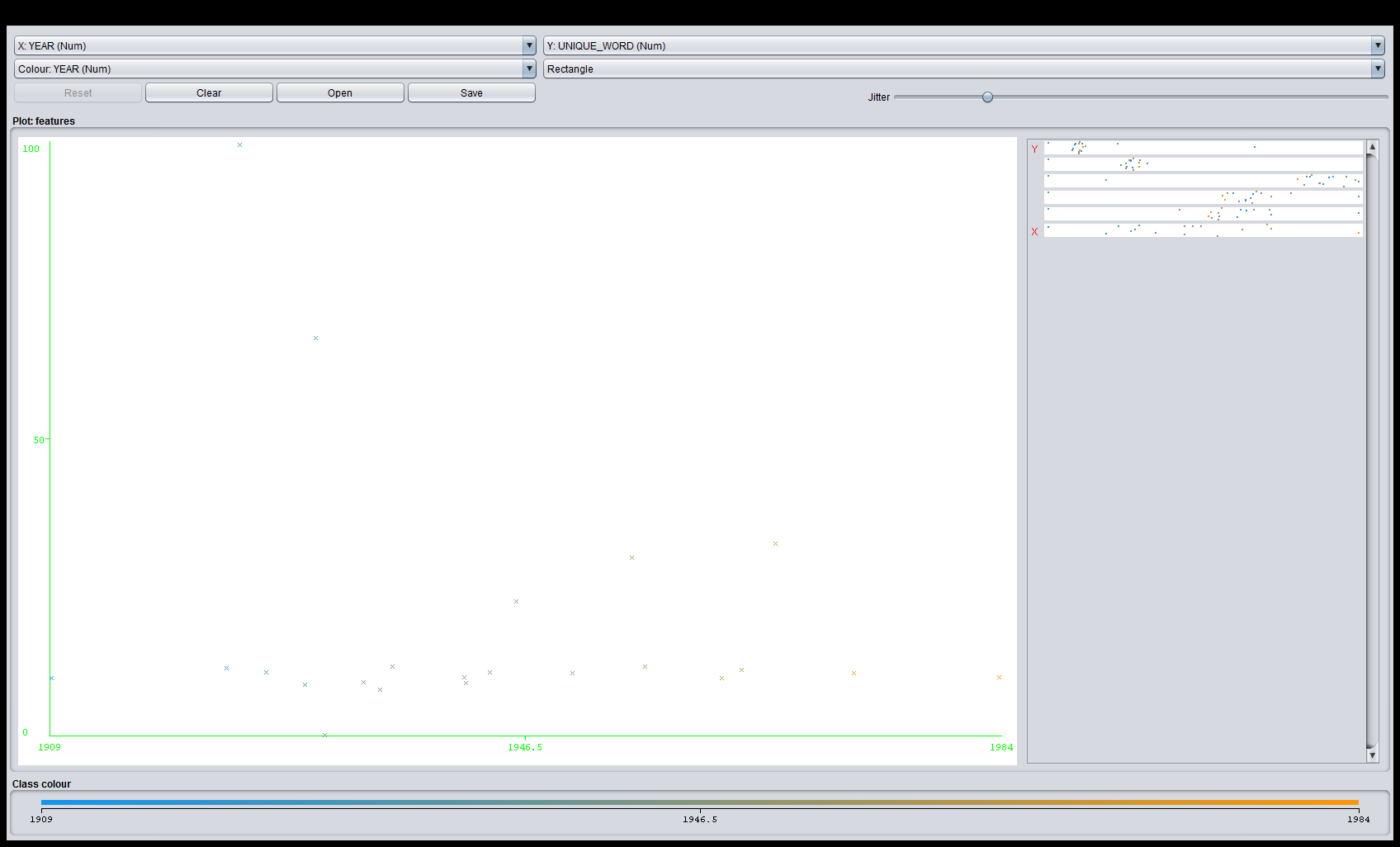
Relative absolute error 30.3777 %

Root relative squared error 24.7163 %

Total Number of Instances 22

So the unique count of words is independent of year and inversely dependent on the count of pronouns.

As seen in the screenshot below the number of unique words remain almost constant over the years.



2. NOUNS = 0.5745 \* UNIQUE\_WORD +

2.7934 \* PRONOUN +

-10.7821 \* INDEFINITE\_TERMS +

2.5463

Correlation coefficient 0.9491

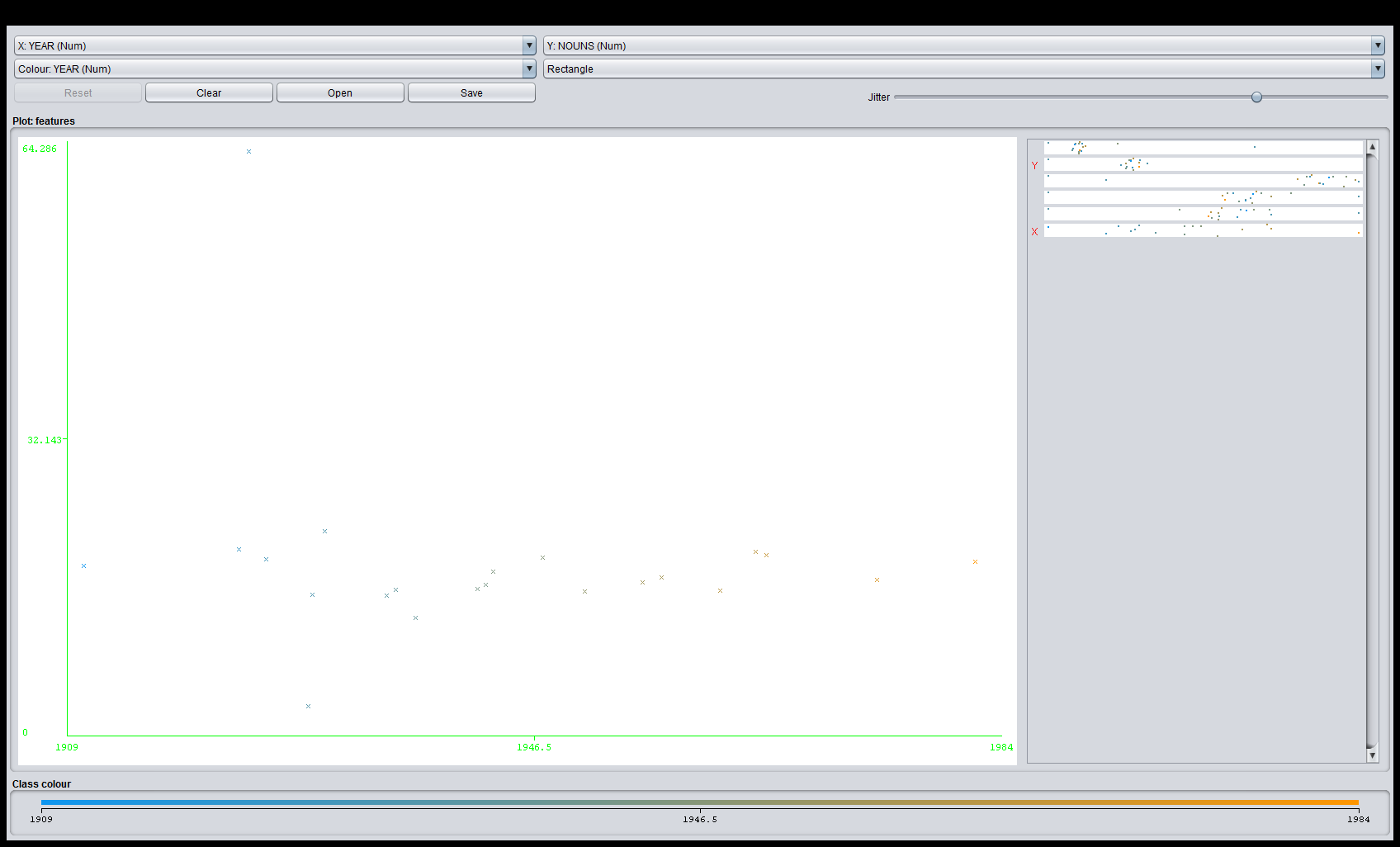
Mean absolute error 2.7575

Root mean squared error 3.3503

Relative absolute error 63.326 %

Root relative squared error 31.4883 %

Total Number of Instances 22



1. PRONOUN = -0.1498 \* UNIQUE\_WORD +

0.2251 \* NOUNS +

0.4113 \* CONJUNCTIONS +

2.4852 \* INDEFINITE\_TERMS +

0.0182 \* YEAR +

-34.3791

Correlation coefficient 0.9414

Mean absolute error 0.683

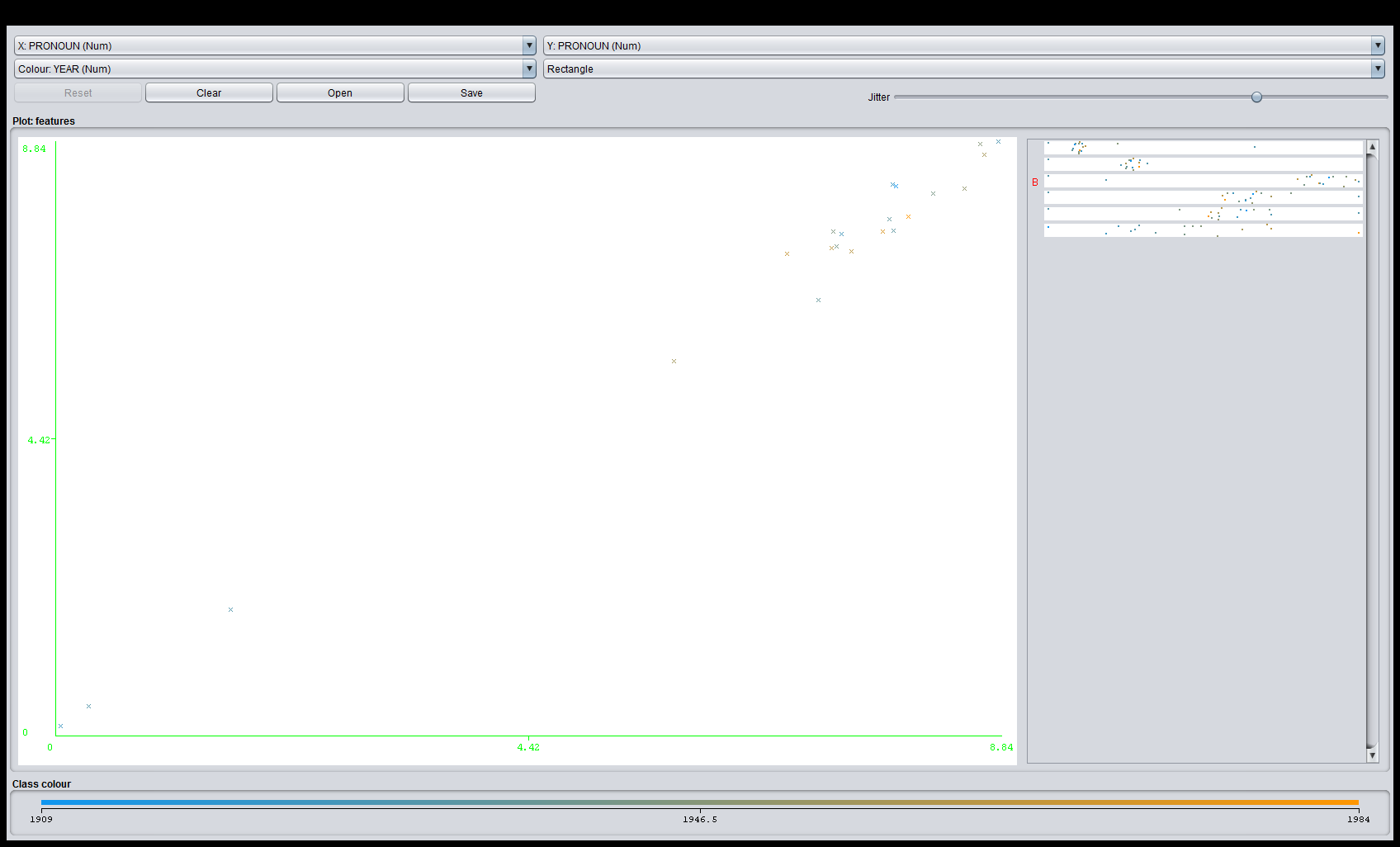
Root mean squared error 0.8584

Relative absolute error 38.7725 %

Root relative squared error 33.7269 %

Total Number of Instances 22

So we see that the frequency of pronouns increases with the increase in the year (that is increase in the age of the author). As we know that Alzheimer’s patients use less of nouns and more of pronouns so the author used more pronouns with age.



1. CONJUNCTIONS =2.0209 \* INDEFINITE\_TERMS +0.5447

Correlation coefficient 0.8854

Mean absolute error 0.3753

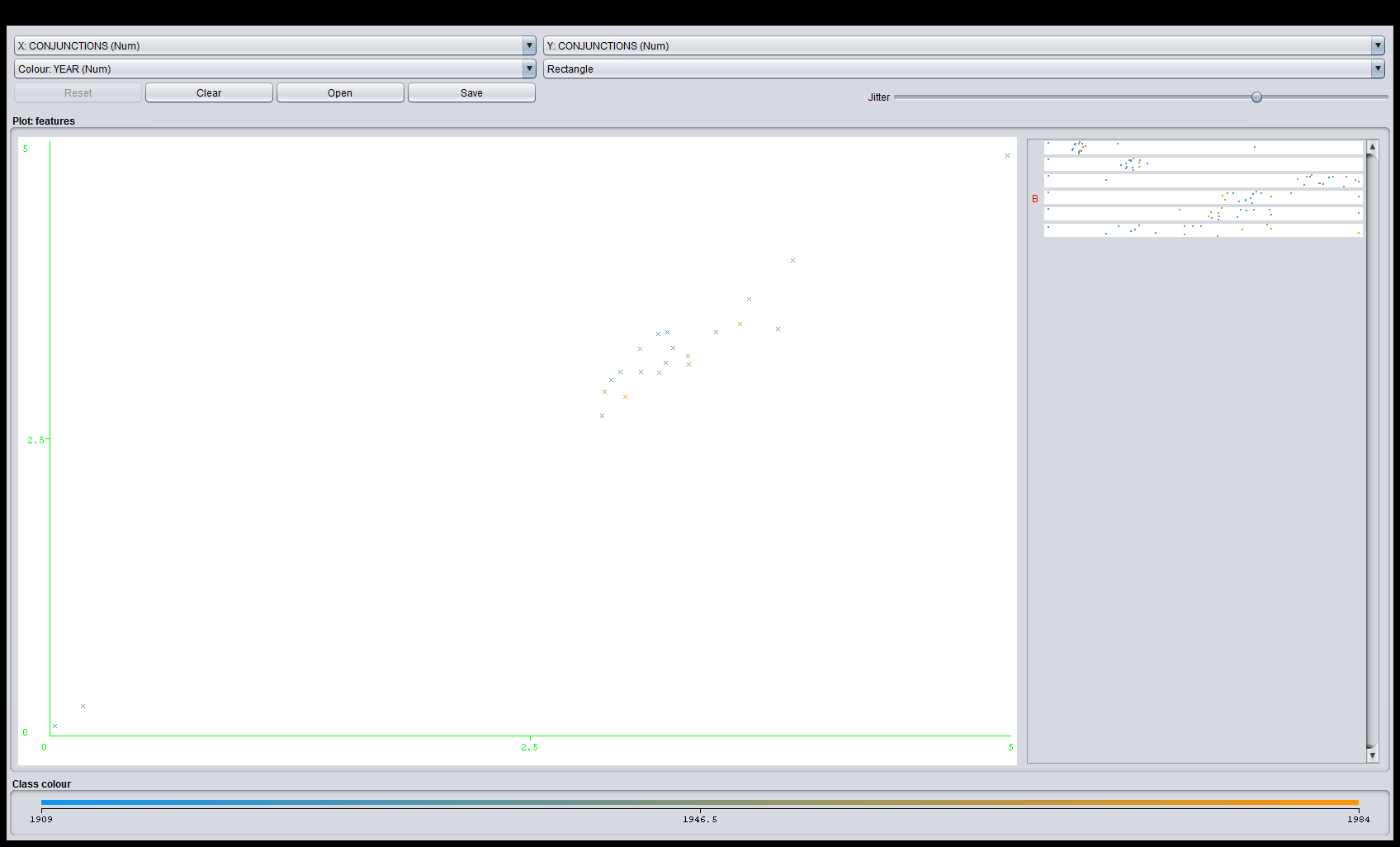
Root mean squared error 0.4926

Relative absolute error 61.0544 %

Root relative squared error 46.4909 %

Total Number of Instances 22

The use of conjunction makes the discourse less coherent so number of conjunctions should increase over the years as is seen from the graph.



1. INDEFINITE\_TERMS = 0.0171 \* UNIQUE\_WORD +

-0.0264 \* NOUNS +

0.0941 \* PRONOUN +

0.2505 \* CONJUNCTIONS +

0.0148

Correlation coefficient 0.9142

Mean absolute error 0.1329

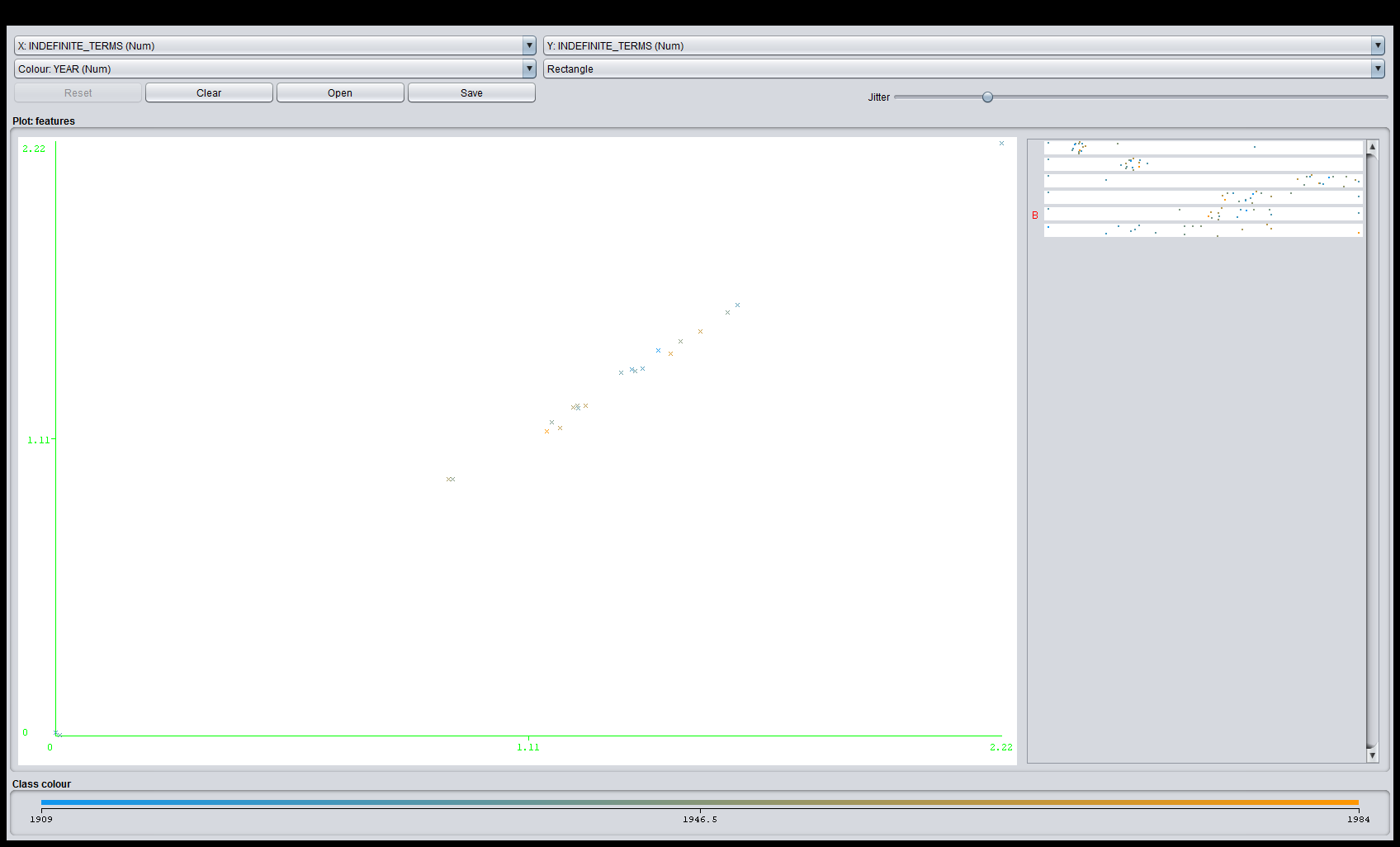
Root mean squared error 0.1881

Relative absolute error 45.3988 %

Root relative squared error 40.5222 %

Total Number of Instances 22

Below is the variation of indefinite terms with year-



YEAR = +1943.5909

Correlation coefficient 0

Mean absolute error 14.7355

Root mean squared error 17.9486

Relative absolute error 100 %

Root relative squared error 100 %

Total Number of Instances 22

The most useful features were:-

Pronouns

Conjunctions

Indefinite terms