# Diploma thesis on IntrinsicPD – Notes from Reading

However, linguistically, it is unknown how simulated plagiarism is

different from real cases[54]

Which threshold methodology captures more suspicious passages ?

- Two thresholds were instituted, as already explained, 90th percentile

and 0.85 as absolute confidence which are changing and unchanging

for each document. The goal is to see which threshold gets a better

score.

* What about setting these values based on an experiment???

„an atom type containing more characters increases the chances of suspicious

passages being identified

„Another important finding would be most of the single stylistic feature

detection had high precision with low recall.“

„The assumption was made that the bigger cluster is to be considered plagiarized and the

smaller is the non-plagiarized cluster[14].“

„After some basic tests conducted by Plagcomps team , it was found that k>2 did not better the results.“ – so why do we have such an option?

# Stylometric features

* Text statistics- which involve individual characters such as number of punctuation marks, length of words, etc.
* Syntactic features- which involve sentences such as length of sentences, no. of function words, etc.
* Part-of-speech features- which involve the type of words used in a text such as no. of adjectives or pronouns.
* Closed class words- which consists of special words such as foreign words, dialects, no. of stop words and complex vocabulary.
* Structural features- which involve the organization of text in a documents such as paragraphs, chapter, etc.

## Structure in the thesis

**Vocabulary richness**

honore\_r\_measure

hapax\_legomena

hapax\_dislegomena

simpsons\_d

sichels\_s

brunets\_w

yule\_k\_characteristic

yule\_i

type\_token\_ratio

**Syntactic features**

stopword\_percentage

punctuation\_percentage

avg\_internal\_word\_freq\_class

avg\_external\_word\_freq\_class

average\_word\_length

average\_sentence\_length

syntactic\_complexity

average\_syllables\_per\_word

alpha\_chars\_ratio

digit\_chars\_ratio

upper\_chars\_ratio

white\_chars\_ratio

**Readability features**

coleman\_liau\_index

automated\_readability\_index

linsear\_write\_formula

gunning\_fog\_index

dale\_chall\_readability\_score

polysyllablcount

smog\_index

flesch\_reading\_ease

flesch\_kincaid\_grade

## Explanations

**Vocabulary richness**

|  |  |  |
| --- | --- | --- |
| honore\_r\_measure | *R = 100 logN / (1 - V\_1 / V) where V\_1 = # words appearing only once, V = total vocab size, N = number of words* | Share of unique words in the atom |
| hapax\_legomena | Number of words occurring once/ number of all words |  |
| hapax\_dislegomena | Number of words occurring twice / number of all words |  |
| simpsons\_d |  | Value is based on the word frequency vector (how many times each frequency occurs in given atom)  Strange measure… could not find any reliable source. Originally designed as a biodiversity measure??? |
| sichels\_s | Hapax\_dislegomena/number of words |  |
| brunets\_w | Number of words ^ (number of unique words ^ -0,172) |  |
| yule\_k\_characteristic | 10000/yule\_i | The larger value, the less diverse vocabulary |
| yule\_i | (m1\*m1)/(m2-m1) | m1 = number of all word forms the text contains  m2 = sum for all wordcounts(wordcount^2)  The larger value, the more diverse vocabulary |
| type\_token\_ratio | number of unique words / number of all words |  |

**Syntactic features**

|  |  |  |
| --- | --- | --- |
| stopword\_percentage | Number of stopwords / number of all words |  |
| punctuation\_percentage | Number of punctuation symbols / number of characters | LEX! |
| avg\_internal\_word\_freq\_class | Average of frequency classes | Same as external\_..., but frequencies are taken from the text itself, not from an external corpus |
| avg\_external\_word\_freq\_class | Average of frequency classes. | <https://webis.de/downloads/publications/papers/stein_2006d.pdf>  Frequency class is defined as floor(log(f(most\_frequent\_word)/f(w)))  (the most frequent word “the” has fr. Class 0)  Frequencies are taken from Brown corpus  This feature corresponds to complexity and vocabulary richness (author is using less common terms) |
| average\_word\_length | average number of characters per word | LEX! |
| average\_sentence\_length | Number of words per sentence | LEX! |
| syntactic\_complexity | 2 \* num\_conjunctions + 2 \* num\_wh\_pronouns + num\_verb\_forms | *This feature is a modified version of the "Index of Syntactic Complexity" taken from Szmrecsanyi, Benedikt. "On operationalizing syntactic complexity." Jadt-04 2 (2004): 1032-1039.  found at http://www.benszm.net/omnibuslit/Szmrecsanyi2004.pdf it tallies various part-of-speech counts which are approximated by NLTK's POS tags. the original version accounts for Noun Phrases, which are not counted in the NLTK tagger, and so are ignored.* |
| average\_syllables\_per\_word | Number of syllables / number of words | LEX! |
| alpha\_chars\_ratio | Number of alphabetic chars / number of all chars | LEX! |
| digit\_chars\_ratio | Number of digits / number of all chars | LEX! |
| upper\_chars\_ratio | Number of uppercase chars / number of all chars | LEX! |
| white\_chars\_ratio | Number of white space chars / number of all chars | LEX! |

**Readability features**

|  |  |  |
| --- | --- | --- |
| coleman\_liau\_index | (0.058 \* avg word length) - (0.296 \* avg sentence length) - 15.8 |  |
| automated\_readability\_index | 4,71 \* (characters / words) + 0,5 \* (words / sentences) – 21,43 | https://en.wikipedia.org/wiki/Automated\_readability\_index |
| linsear\_write\_formula | (number of easy words + 3 \* number of difficult words) / number of sentences  If ≤20, subtract 2  Then divide by 2 | <https://en.wikipedia.org/wiki/Linsear_Write>  Difficult words = 3 or more syllables |
| gunning\_fog\_index | 0,4 \* (avg words per sentence + 100\*(number of complex words / number of words)) | complex word = 3 or more syllables. More details: https://en.wikipedia.org/wiki/Gunning\_fog\_index |
| dale\_chall\_readability\_score | 0,1579 \* (number of difficult words / number of words \*100) + 0,0496 \* avg\_words per sentence | Difficult words = not in a white list |
| polysyllablcount | Number of words of ≥3 syllables | LEX! |
| smog\_index | 1,043 \* sqrt(30\*difficult words / sentences ) + 3,1291 | <https://en.wikipedia.org/wiki/SMOG>  Difficult words = 3 or more syllables |
| flesch\_reading\_ease | 206,835 – 1,015 \* (words / sentences) – 84,6 \* (syllables/words) | Paragraph feature |
| flesch\_kincaid\_grade | 0,39 \* (words/sentences) + 11,8 \* (syllables/words) – 15,59 | Paragraph feature |

Pairs of highly correlated features (does not make sense to evaluate both):

* Type\_token\_ratio; Hapax\_legomena
* Flesch\_reading\_ease; Flesch\_kincaid\_grade
* Yule\_k = 10000/yule\_i