# Evaluation of Text Reuse Corpora for Text Alignment Task of plagiarism Detection

Vahid Zarrabi, Javad Rafiei, Khadijeh Khoshnava, Habibollah Asghari, Salar Mohtaj

**ICT Research Institute**

Academic Center for Education, Culture and Reseach (ACECR), Iran

## Task

**Text Alignment Corpus Review**

**Peer-review:** Assess and analyze the corpora of all participants of text alignment corpus construction task in order to determine corpus quality. All corpora are available to the other participants of this task to the subject of peer-review.

## Corpora Statistical Information

- For evaluation of corpora based on statistical information, we categorized the statistical information in three different aspects:
  - The first view describes the numerical information about corpora such as number and length of documents and suspicious cases (The first Table)
  - In the second view, the distributions of obfuscation strategies are demonstrated (The second Table)
  - In the last view, we have calculated some ratios for demonstrating a better statistical picture of corpora (The third Table)

### The Statistical Information of the 5 Corpora

| | Cheema15 | Hanif15 | Kong15 | Alvi15 | Palkovskii15 |
|---|---|---|---|---|---|
| Number of Docs | | | | | |
| Suspicious Docs | 248 | 250 | 4 | 90 | 1175 |
| Source Docs | 248 | 250 | 78 | 70 | 1950 |
| Length of Docs (in chars) | | | | | |
| Min Length | 2263 | 361 | 394 | 514 | 519 |
| Max Length | 22471 | 74083 | 121829 | 45222 | 517925 |
| Average Length | 7239 | 4382 | 42839 | 7718 | 6512 |
| Length of Plagiarisms Cases (in chars) | | | | | |
| Min Length | 134 | 78 | 62 | 259 | 157 |
| Max Length | 2439 | 849 | 2748 | 1160 | 14336 |
| Average Length | 503 | 361 | 423 | 464 | 782 |

### Obfuscation Strategies Employed by Participants

| Obfuscation Strategies | Cheema15 | Hanif15 | Kong15 | Alvi15 | Palkovskii15 |
|---|---|---|---|---|---|
| Simulated | 123 | 135 | - | - | - |
| Real | - | - | 109 | - | - |
| Automatic | - | - | - | 25 | - |
| Retelling-Human | - | - | - | 25 | - |
| Character-Substitution | - | - | - | 25 | - |
| Translation | - | - | - | - | 618 |
| Summary | - | - | - | - | 1292 |
| Random | - | - | - | - | 626 |
| None | - | - | - | - | 624 |

### Relative Statistical Information of Corpora

| Number/Percent | Cheema15 | Hanif15 | Kong15 | Alvi15 | Palkovskii15 |
|---|---|---|---|---|---|
| Plagiarism Cases | 123 | 135 | 109 | 75 | 3160 |
| Plagiarism Cases per Suspicious Document | 0.49 | 0.54 | 27.25 | 0.83 | 2.68 |
| Share of plagiarism cases in Suspicious documents | 3.4% | 4.4% | 26.9% | 4.9% | 32.18% |

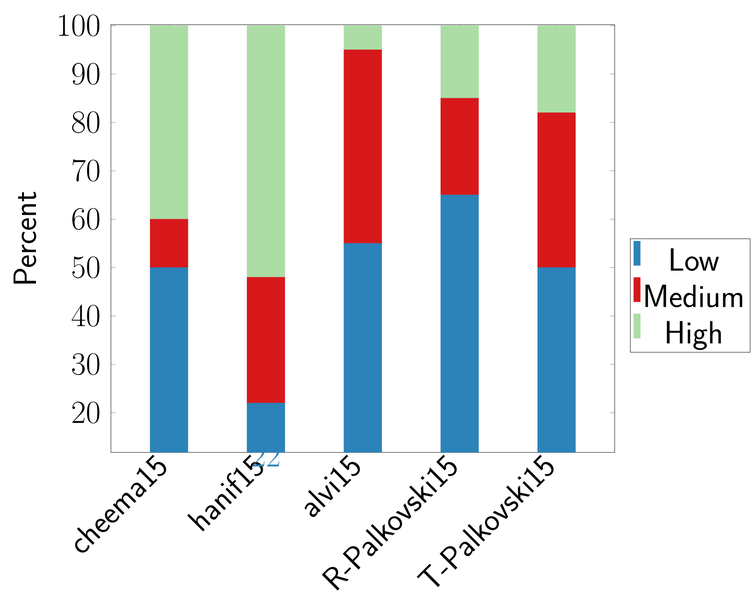
- Some participants have employed one type of obfuscation:
  - . **Cheema15** and **Hanif15** applied simulated obfuscation in their corpora.
  - . **Kong15** corpus includes just real obfuscation strategy.
- Two participants have multiple obfuscation strategies in their corpora:
  - . **Alvi15** corpus has employed three types of obfuscation (retelling-human - character-substitution and automatic - Character-Substitution)
  - . **Palkovskii15** corpus covers four kinds of obfuscation (None - cyclic Translation - summary obfuscation - random obfuscation)

## Manual Evaluation of Corpora

- In this section we manually investigate twenty pairs of corresponding source and suspicious fragments in each corpus.

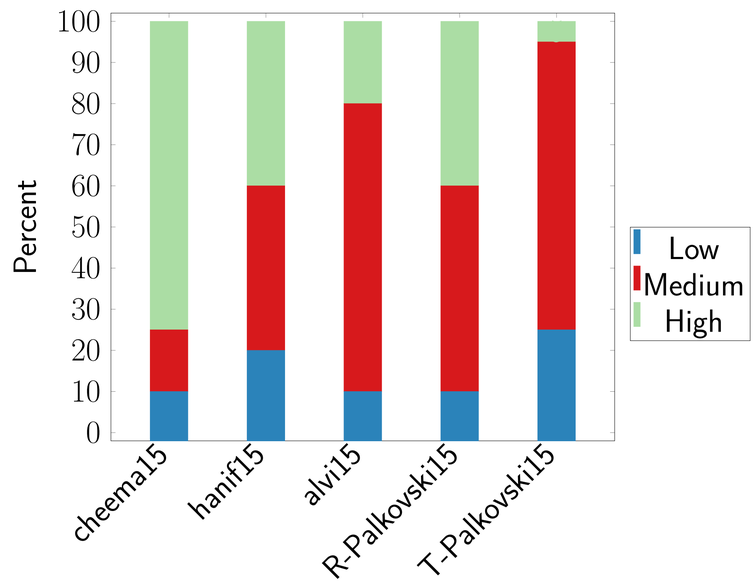
### Changes in syntactic structure

- This measure shows the rate of structural changes in each corpus based on three categories low, medium and high.




### Rate of obfuscation

- This measure expresses the ratio of alternated words (based on 4 types of obfuscations) to total number of the words in source fragments.




* In both tables R-Palkovski15 refer to the Random part and T-Palkovski15 refer to the Translation part of Palkovski15 corpus.

## Automatic Evaluation of Corpora

- In this section, we separately evaluate two remained obfuscation scenarios:
  - Real obfuscation from Kong15 corpus
  - Summary obfuscation from Palkovskii15 corpus
- Automatic Evaluation of **Kong15** corpus.
  - For Kong15 corpus, all source and correspond suspicious fragments are extracted, and the total number of similar characters n-grams between source and suspicious plagiarized passages are calculated for n in range of one to four
  - In the next step, the normalized total numbers (in percent) are clustered using k-means clustering algorithm
- Automatic Evaluation of **Palkovskii15** Corpus.
  - For evaluation of summary obfuscation from the point of concept preserving measure, we have extracted 10% of top words from source fragments based on tf.idf weight.
  - Using k-means clustering algorithm, the suspicious fragments are classified into three clusters.

This work has been accomplished under the support of Vice Presidency for Science and Technology of Iran grant No. 1164331



{vahid.zarrabi, javad.rafiei, khadijeh.khoshnava, habib.asghari, salar.mohtaj}@ictrc.ir