

UNCOVERING INSIGHTS FROM CORPUS DATA THROUGH UTILIZATION OF ACCESSIBLE ANALYTICAL TOOLS

Comparing selected crime novels of
Arthur Conan Doyle



SUMMARY



What is
Corpus?

A large and structured collection of
written or spoken texts used for
linguistic analysis and research.

Importance
of Corpus

It provides a rich source of language data
for studying language variation,
developing language models, evaluating
language technologies, and more

Metadata
of selected
corpora

Corpora: My Friend The Murderer, A Study of The
Scarlet , The Mystery of Cloomber
Metadata: Title, Genre, Publishing Date

Tools used

- Project Gutenberg for the text files
- Voyant tool for data visualization

Interesting trends

In all 3 of the books, 'Said' was the most
used word and 'man ' came second;
which means most of the charcters were
men, quite contradictory to modern
times literature, and more.

Find the whole
project at

<https://aysa2018.github.io/corpus-assignment/>

Name: Aysa Binte Masud

Course: IM-UH 1511 Introduction
to Digital Humanities

Instructor: David Joseph Wrisley,
Professor of Digital Humanities

جامعة نيويورك أبوظبي



NYU ABU DHABI