UNIVERSITY OF EASTERN FINLAND

Doing Lexical and Stylistic research using corpora

Corpus Linguistics Association of Nigeria (CLAN) workshop

Temitayo Olatoye

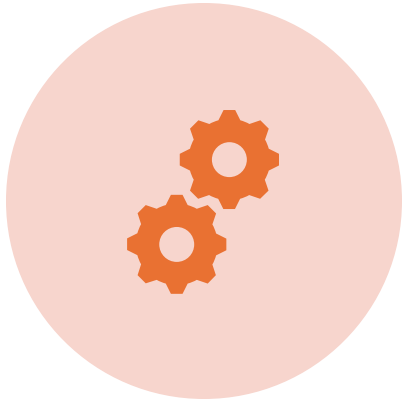
temitayo.olatoye@uef.fi

Nov 21, 2025

Outline

TOOLS

USEFUL CONCEPTS



ADVANCED TEXT PROCESSING

Tools

- ANTCONC: A corpus analysis toolkit for concordancing and text analysis

https://www.laurenceanthony.net/software/antconc/

- R and R Studio: Software and user-friendly integrated development environment (IDE)

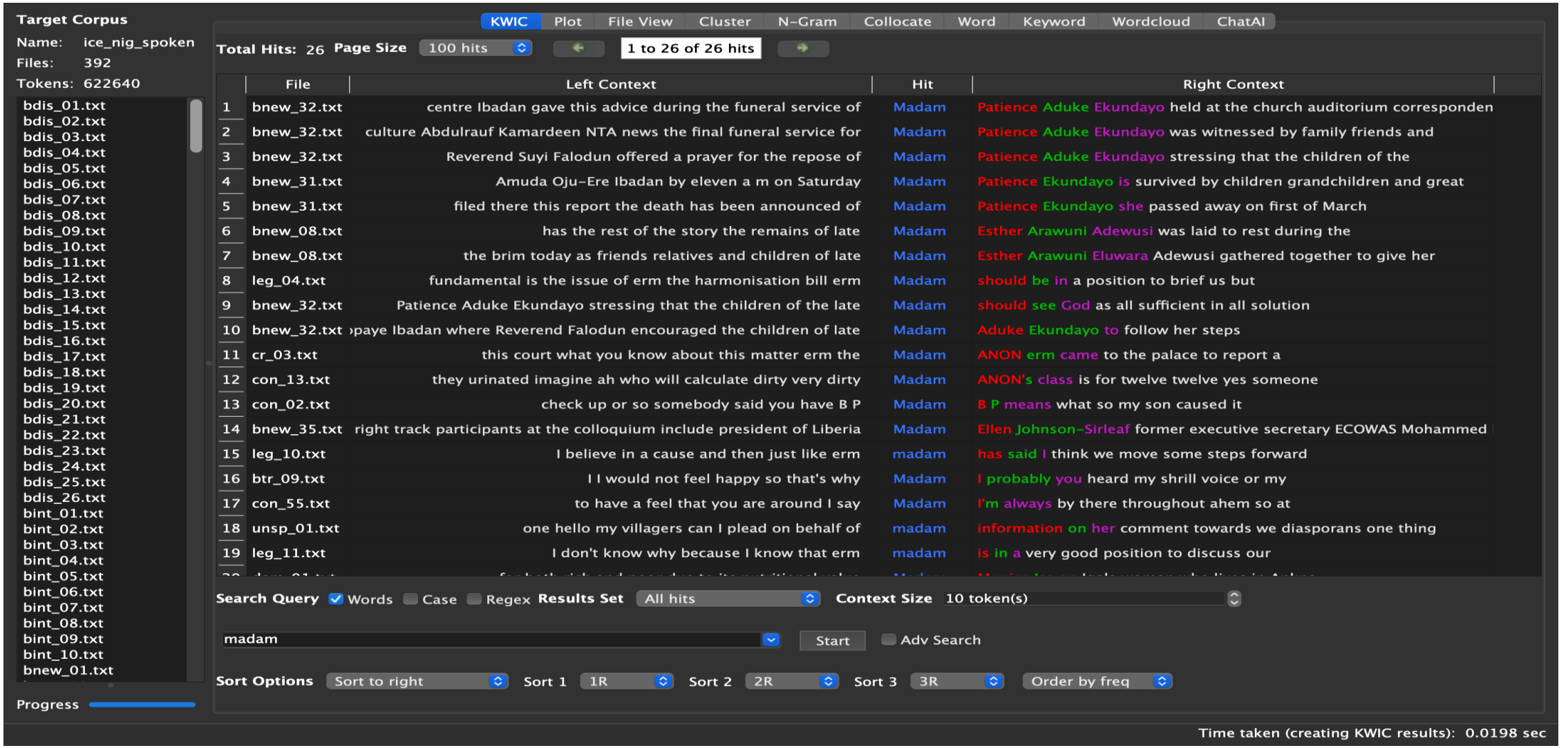
https://cran.r-project.org/

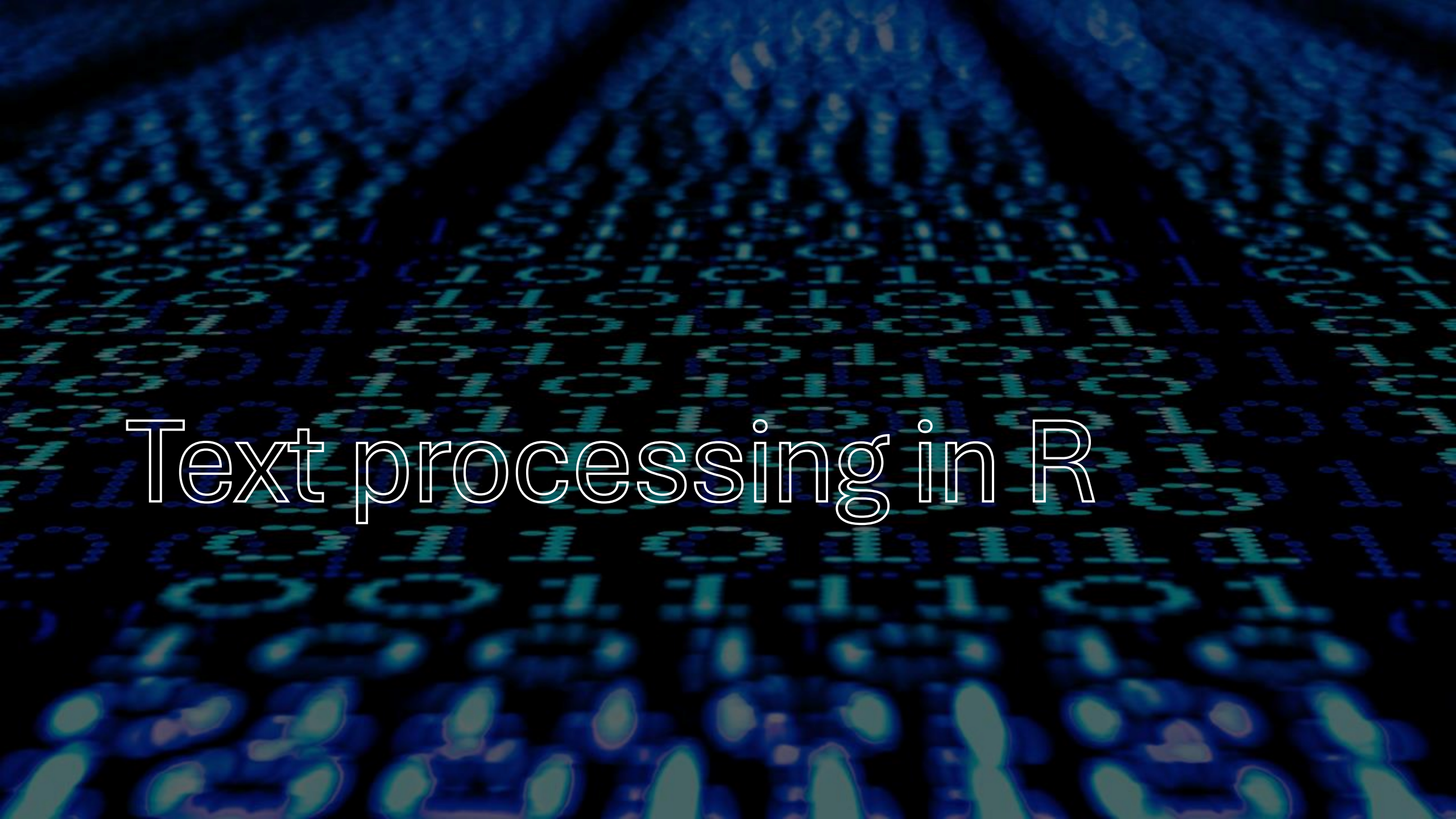
https://posit.co/download/rstudio-desktop/

These tools are open-access.

Concordancing in AntConc

A list of all the cases of a word, phrase or expression in a corpus




Text processing in R

Natural Language Processing

- **1. Tokenization**

This is the process of breaking text into smaller pieces called *tokens*.
Tokens can be words, sentences, or even characters.
Example: “I like oranges” → “I”, “like”, “oranges”.

- **2. Lemmatization**

This means converting a word to its *dictionary form* (called a lemma).
It uses context and grammar. Thus, it produces real words.
Example: “running”, “ran” → “run”.

- **3. Stemming**

This is a simpler, rougher way of chopping words down to their *root form*.
The result is not always a real word.
Example: “running”, “runner” → “run” or “runn”.

- **4. Stop-word removal**

This is the process of removing words that do not add much meaning.
Examples: “the”, “is”, “and”, “a”.

- **5. Part of Speech (POS) tagging**

This means labelling each token with its grammatical role.
Examples: noun, verb, adjective, adverb.
Cats sleep often. → *Cats* (**noun**), *sleep* (**verb**), *often* (**adverb**), .(**punctuation**)

Regular Expressions (REGEX)

Regular expressions (often called *regex*) are special patterns used to search for, match, or replace text.

- They work like advanced search tools that let you describe what you want to find using symbols.

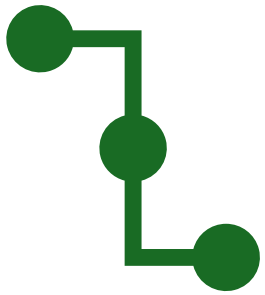
Examples:

- The pattern [A-Z] matches any capital letter.
- The pattern cat|dog matches either “cat” or “dog”.
- The pattern \w+ matches a whole word.
- "(\\w+)/VERB(?:\\s+\\w+/\\w+){0,2}\\s+(\\w+)/ADP"

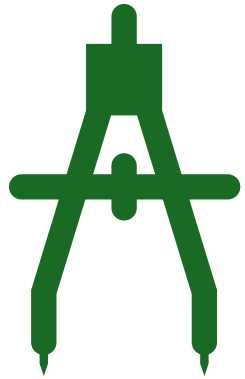
#any verb with up to 2 intervening elements followed by a preposition

Regular expressions are rules that help you find specific patterns in text, not just exact words.

Turning texts into social networks



How are characters in a text connected?



How can we visualize a complex object by its core structure?

Network Analysis

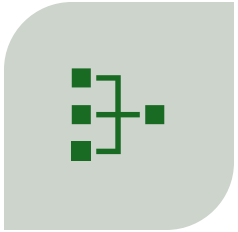
- Constructing and visualizing relationships in digital texts.
- Networks consist of nodes (dots) and edges (lines).

Network Analysis

Occurrences – mentions of/references to a character

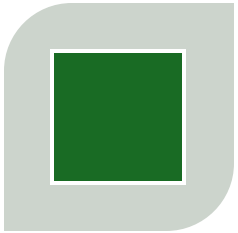
Conversations – character contributes/performs an event

- Visualization of a whole - framing
- In sections – development over time

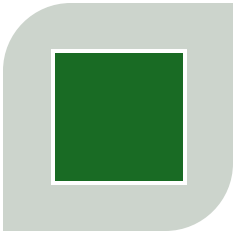


CO-OCCURRENCE NETWORK

CONVERSATIONAL NETWORK



STATIC NETWORK



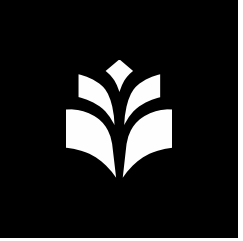
DYNAMIC NETWORK

Network Analysis

- Directed: the direction of the interaction matters.
- Undirected

Directed network

Undirected network



Named Entity Recognition (NER)

- We tag the entities in a text.

PERSON, LOCATION etc.

- We group the co-referents in various parts of the text.

- We can also assign genders to the characters.

Elizabeth, Lizzy, Miss Bennet = Elizabeth Bennet

Annotation

Mr. Bennet
She
Mrs. Long
His wife
Mr. Morris
~~His servants~~
Bingley
~~Our girls~~
he

Mr. Bennet replied that he had not.

“But it is,” returned she; “for Mrs. Long has just been here, and she told me all about it.”

Mr. Bennet made no answer.

“Do not you want to know who has taken it?” cried his wife, impatiently.

“_You_ want to tell me, and I have no objection to hearing it.”

This was invitation enough.

“Why, my dear, you must know, Mrs. Long says that Netherfield is taken by a young man of large fortune from the north of England; that he came down on Monday in a chaise and four to see the place, and was so much delighted with it that he agreed with Mr. Morris immediately; that he is to take possession before Michaelmas, and some of his servants are to be in the house by the end of next week.”

“What is his name?”

“Bingley.”

“Is he married or single?”

“Oh, single, my dear, to be sure! A single man of large fortune; four or five thousand a year. What a fine thing for our girls!”

References

- Anthony, Laurence. 2024. AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc
- Ardanuy, Mariona Coll & Sporleder, Caroline. 2014. Structure-based clustering of novels. Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL) @ EACL 2014, pages 31–39.
- Bamman, David, Lewke, Olivia & Mansoor, Anya. 2020. An Annotated Dataset of Coreference in English literature. *Proceedings of the Twelfth Language Resources and Evaluation Conference*", 44 -54. https://aclanthology.org/2020.lrec-1.6/