Text processing with NooJ

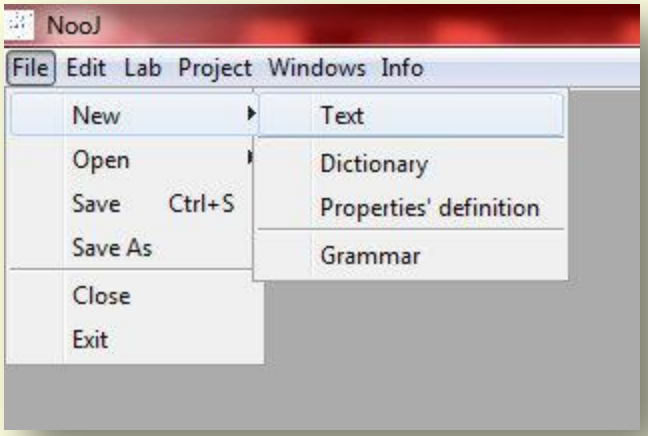
Dhekra Najar

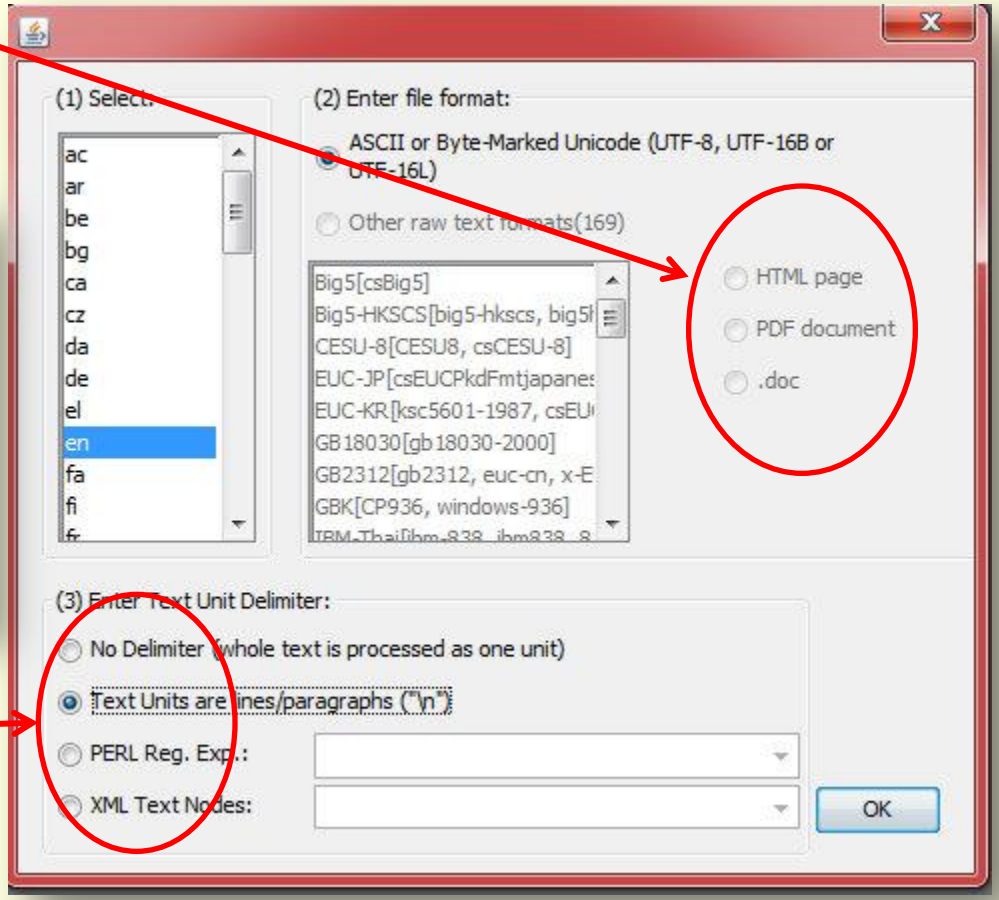
RIADI Laboratory, University of Manouba, Tunisia
dhekra.najar@gmail.com

NooJ'16 conference
České Budějovice, 9-11 June 2016

Creating new text

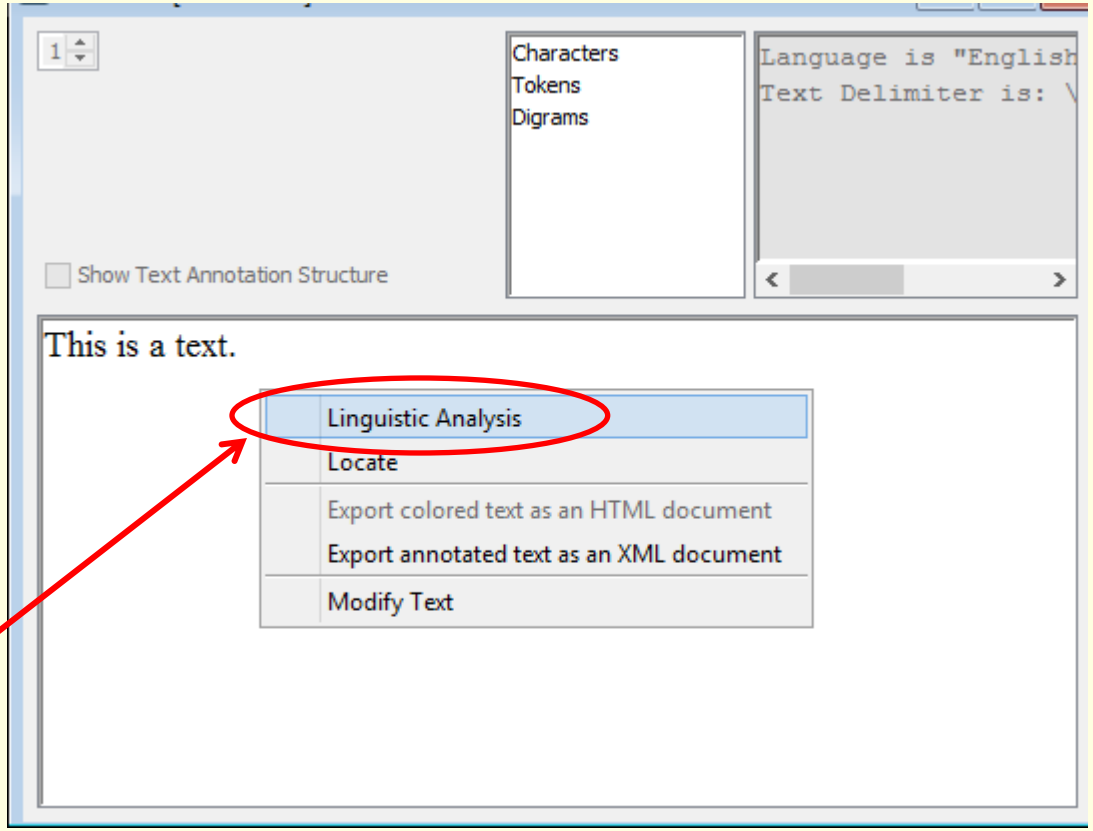
Different input formats





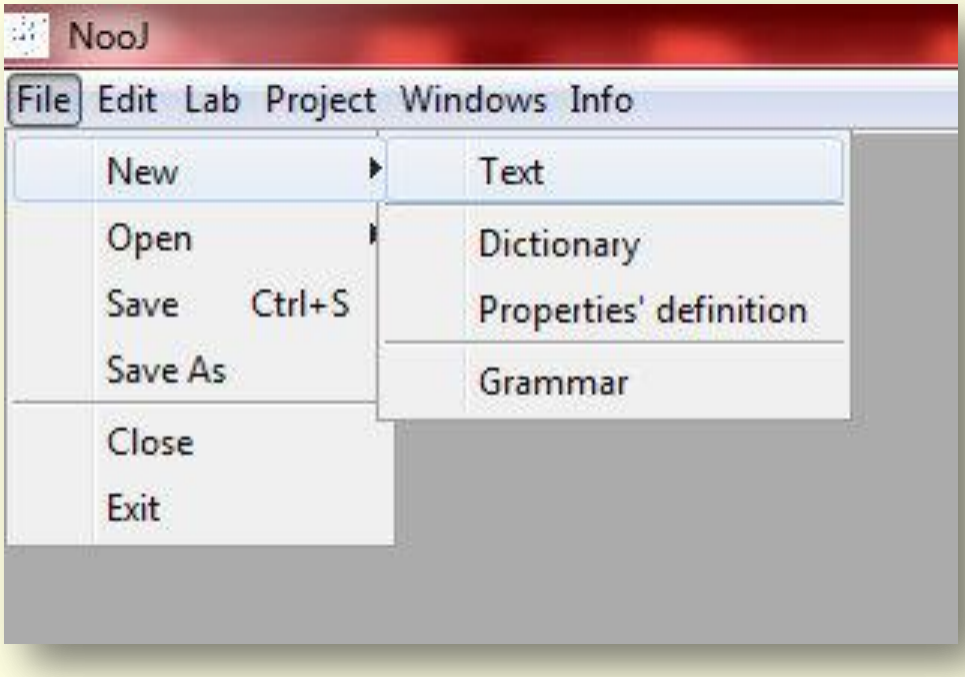
Different text delimiters

Creating new text

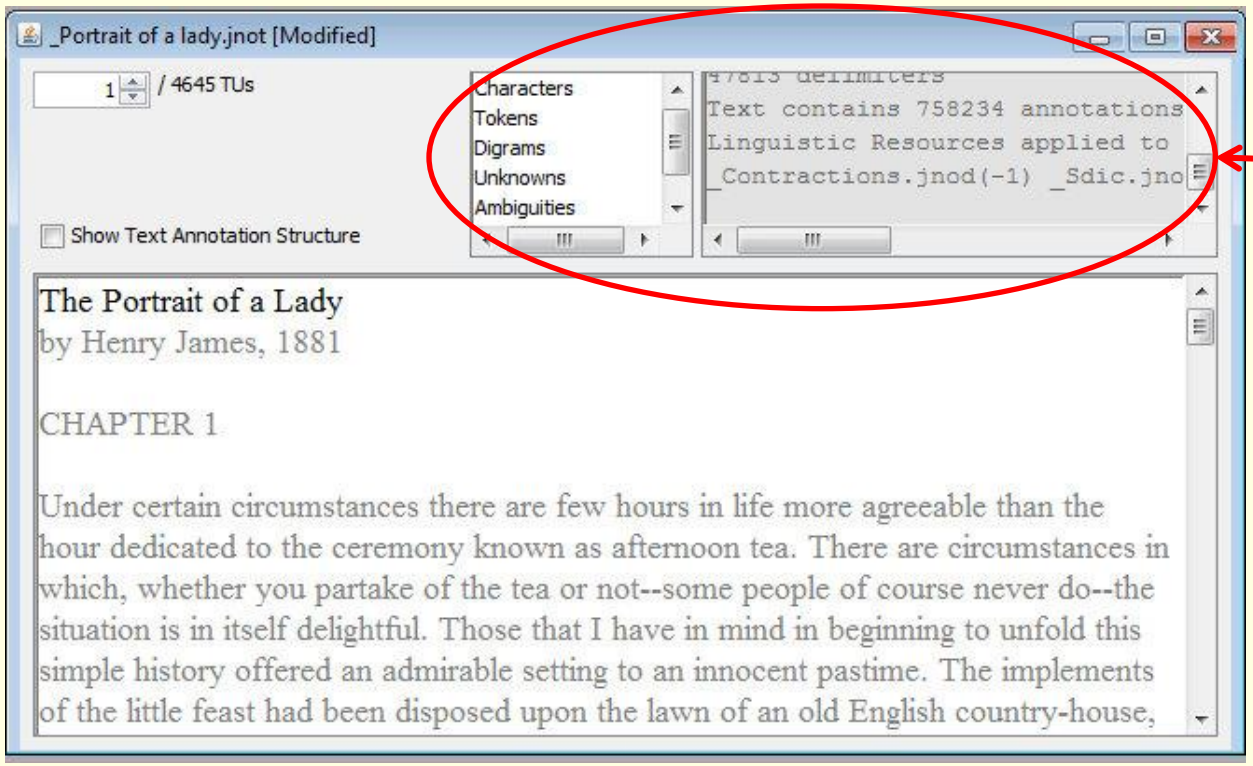


Linguistic analysis

Loading a text



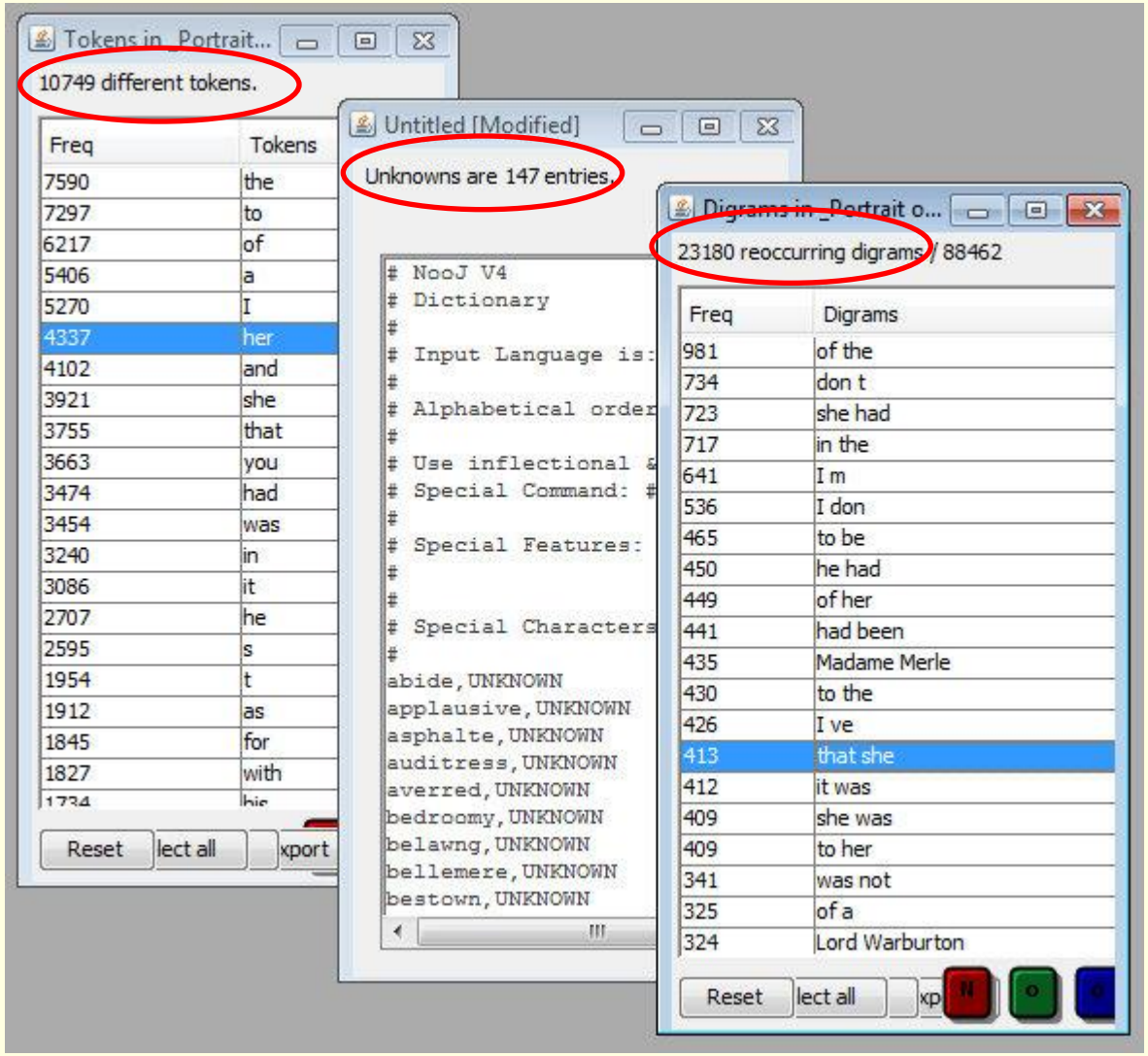
Loading a text



Different indications of the text.

Tokens are the basic linguistic objects processed by NooJ. They are classified into three types: **Word Forms** are sequences of letters between two delimiters; **Digits**; and **Delimiters**. **Digrams** are pairs of word forms (we ignore the delimiters between them).

Loading a text



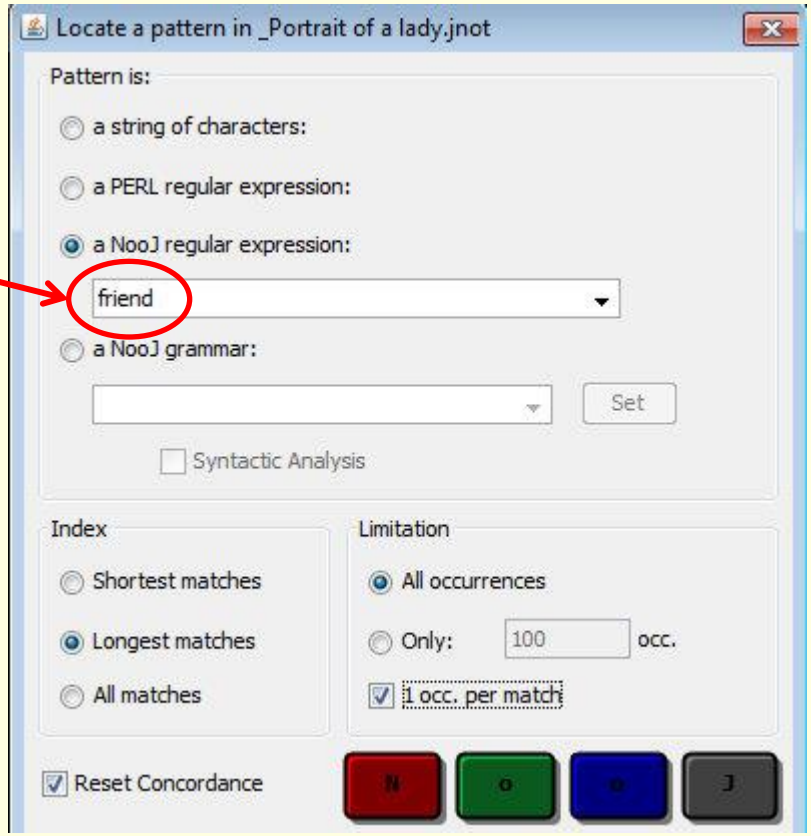
The text's tokens, digrams and unknowns.

The digrams of tokens and digrams and characters can be sorted alphabetically or according to the frequency of each item.

Locating a word in text

In the **TEXT menu, click "Locate". The "Locate Panel" window will show up.**

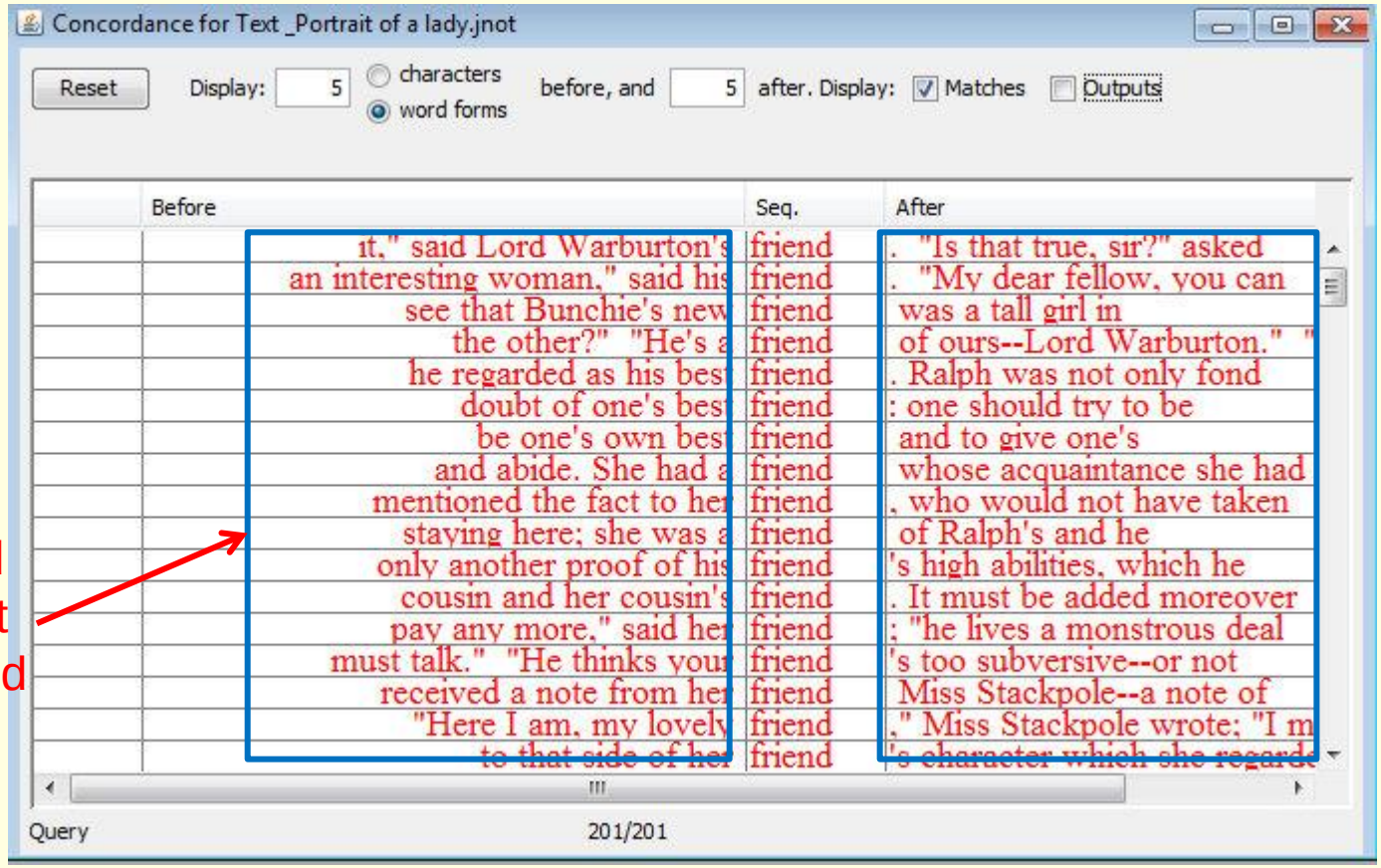
The word to locate



Click a colored button to launch the search operation.

Locating a word in text

NooJ lets you know that it found 201 matches for the query, and then displays a **concordance** in the selected color.



The left and right context Of the located word

Locating an expression in text

Basics

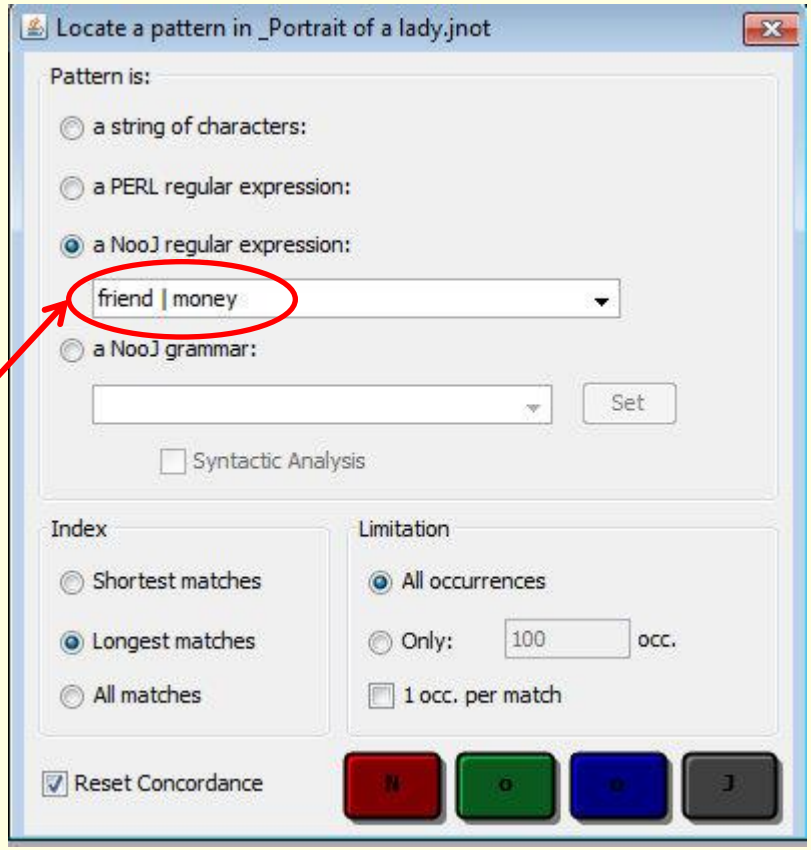
- **Disjunction (OR)**: is symbolized by the “|” character;
- **The blank**: concatenation operator;
- **WF**: any word form;
- **The « * » (Kleene operator)**: is used to mark unlimited repetitions;
- **<E>**: empty string.

Exemple: a (credit | <E>) card = a credit card | a card;

- **Parentheses** can be used to change the order of priority.

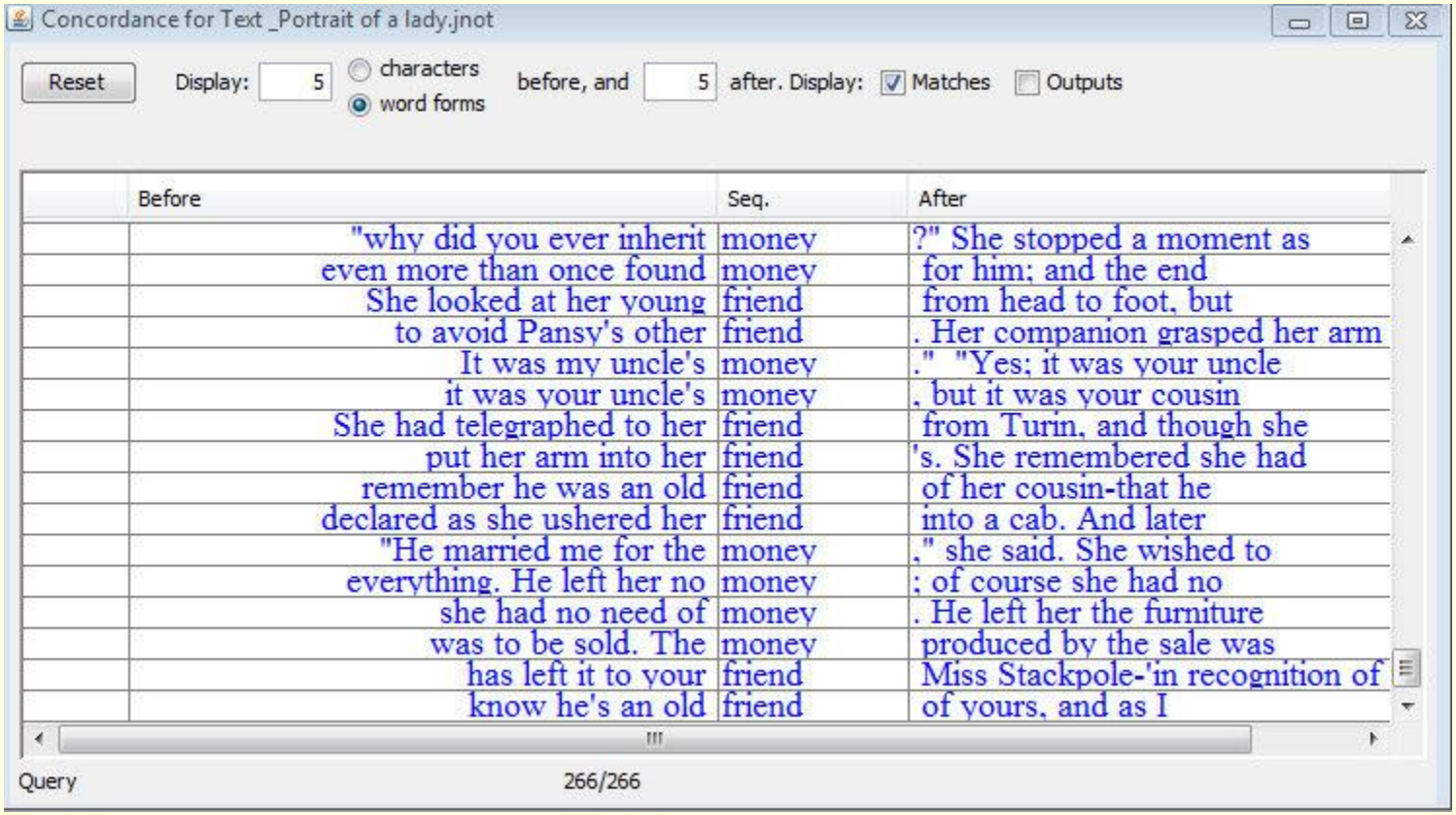
Locating an expression in text

locate all of the utterances for “friend” or “money” in the text



Locating an expression in text

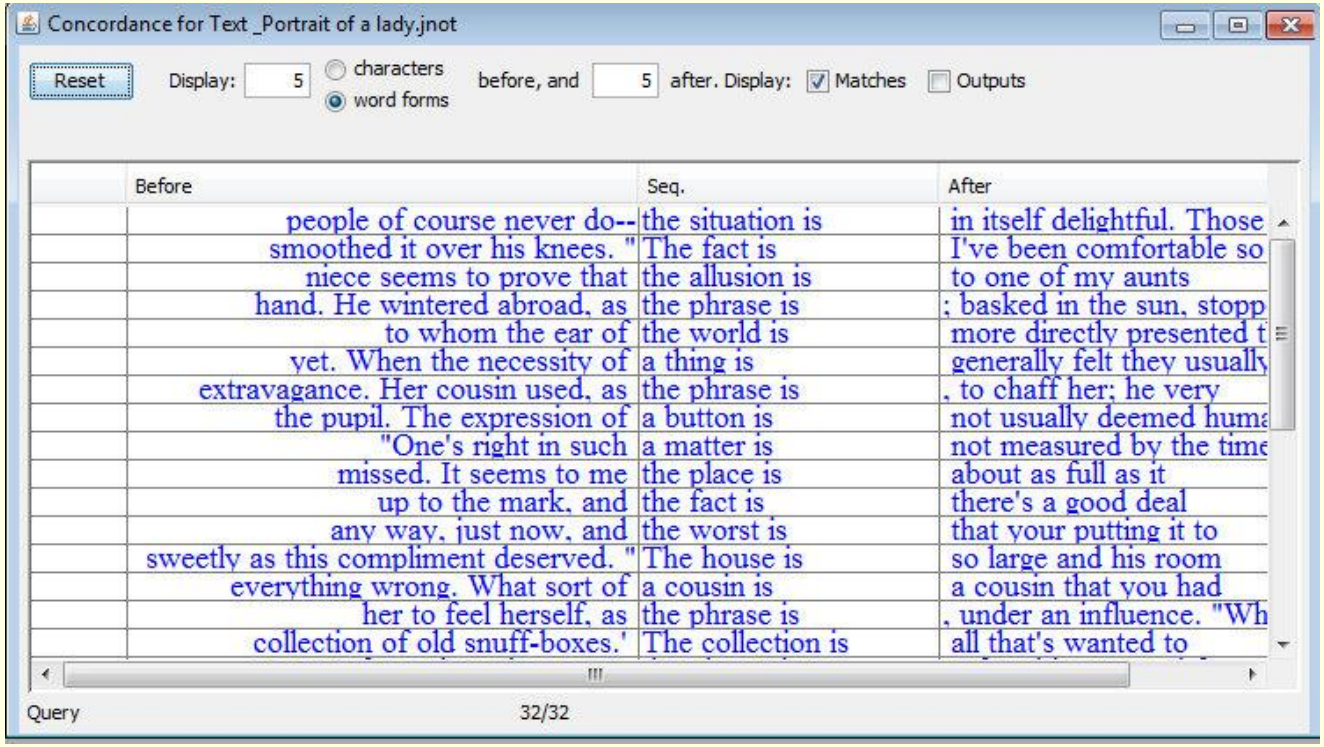
NooJ finds 266 matches for this query.



Locating an expression in text

- find all of the sequences made up of the determinant “the” or “a”, followed by any word form (<WF>), followed by the form “is”.

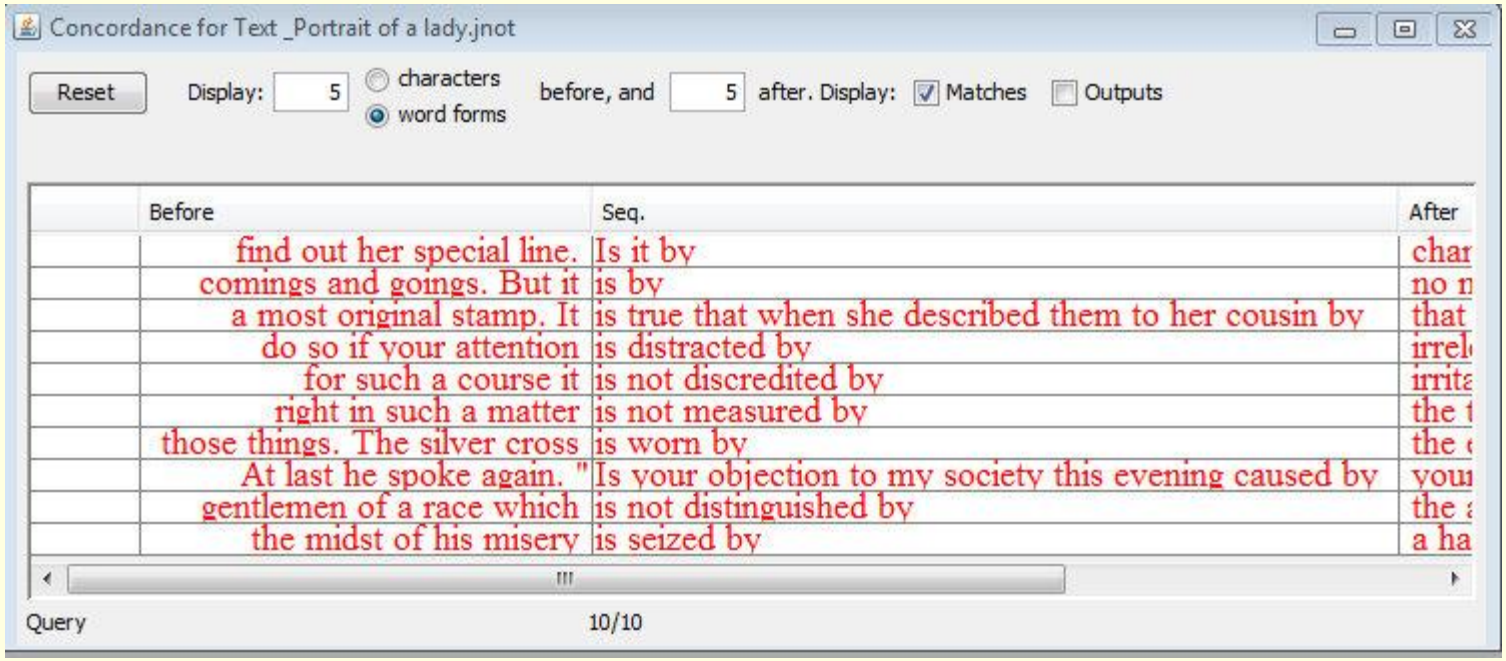
(the | a) <WF> is



Locating an expression in text

- find all of the sequences that begin by the word form “is”, followed by any word form (<WF>) and end by the word form “by”.

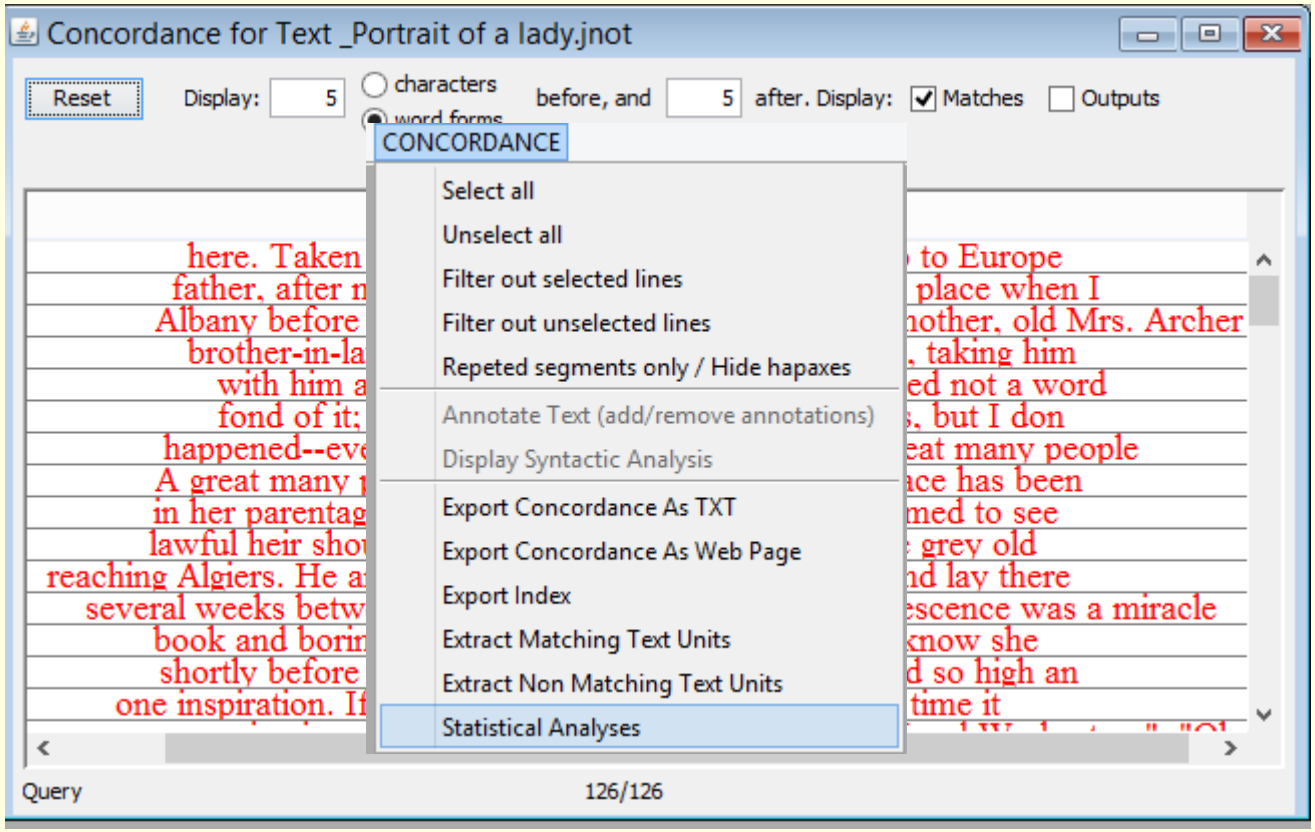
is <WF>* by



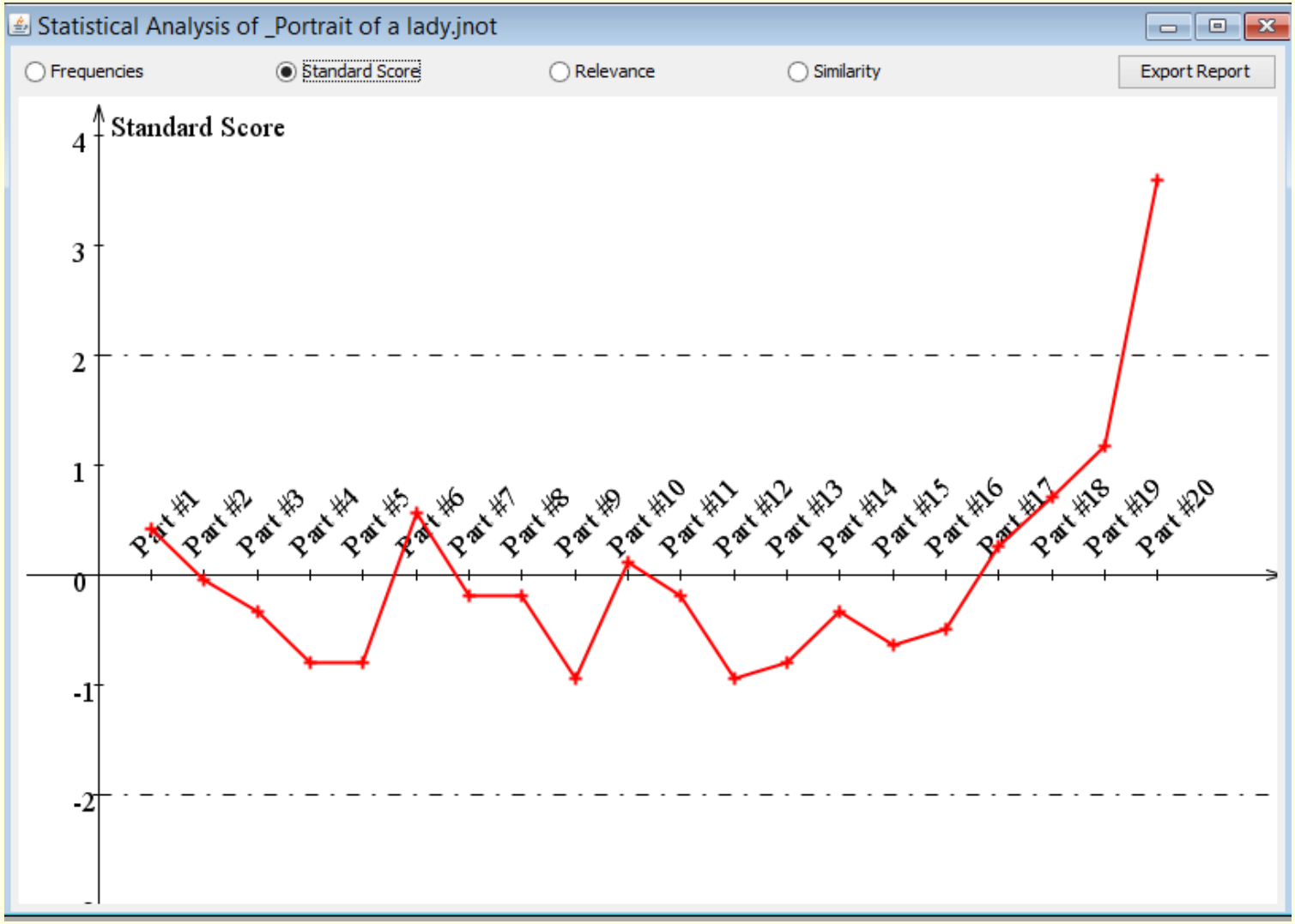
Locating an expression in text

- find all of the words that are related to “death”.

Locating an expression in text



Locating an expression in text



THANK YOU