



Text processing with NooJ

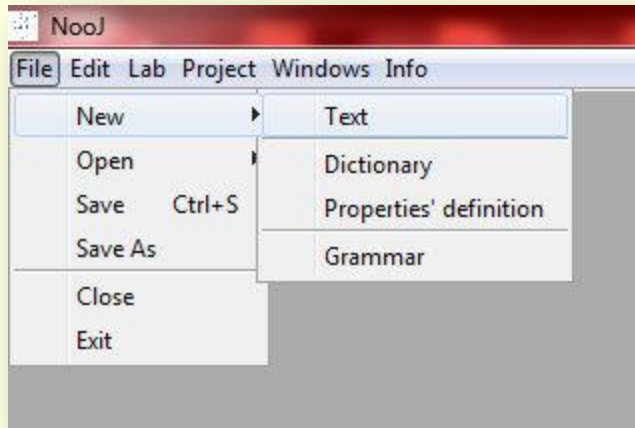
Dhekra Najjar

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

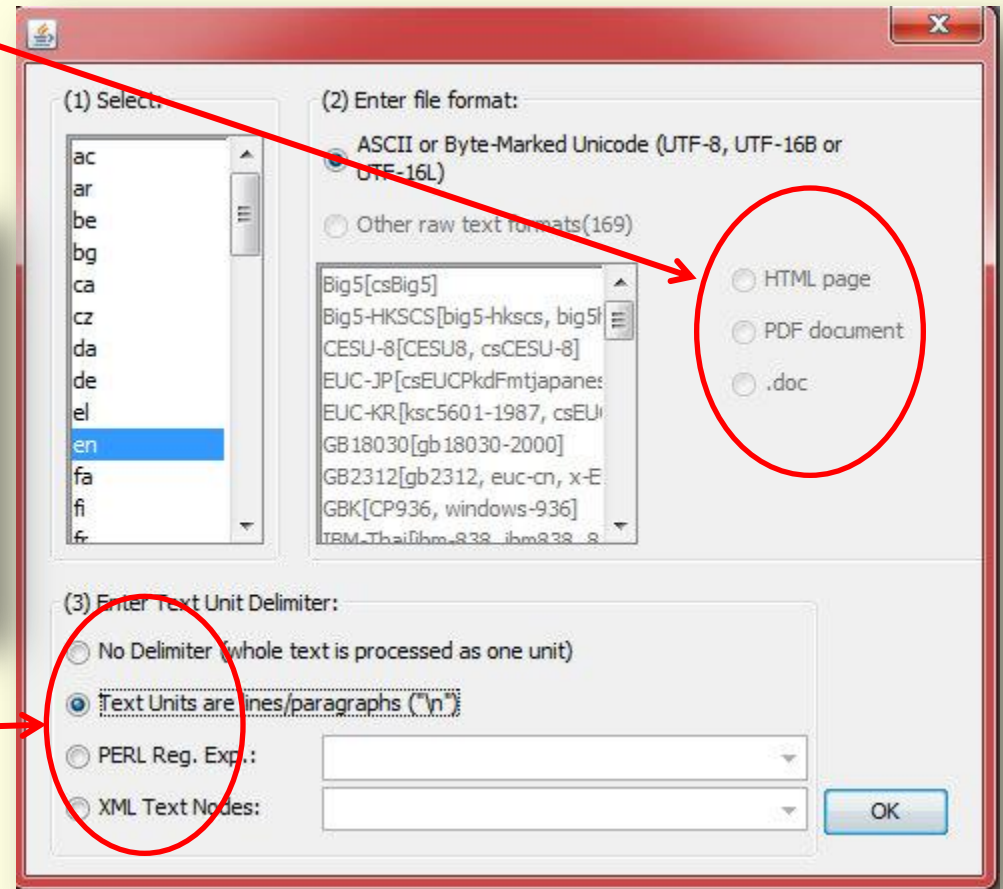
NooJ'16 conference
České Budějovice, 9-11 June 2016

Creating new text

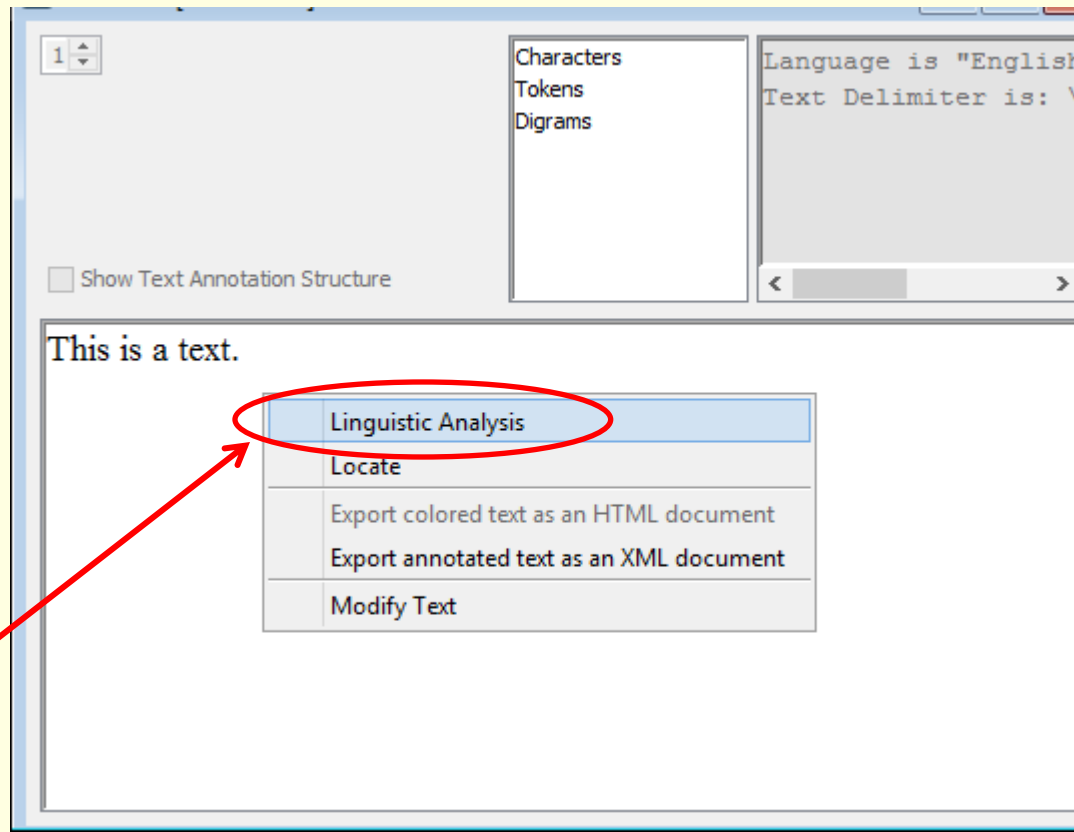
Different input formats



Different text delimiters

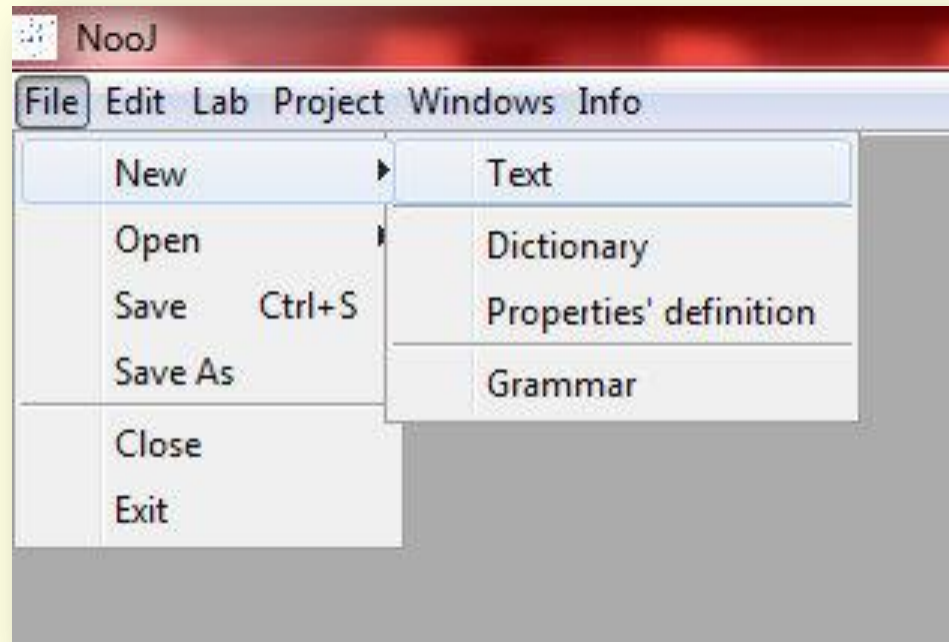


Creating new text

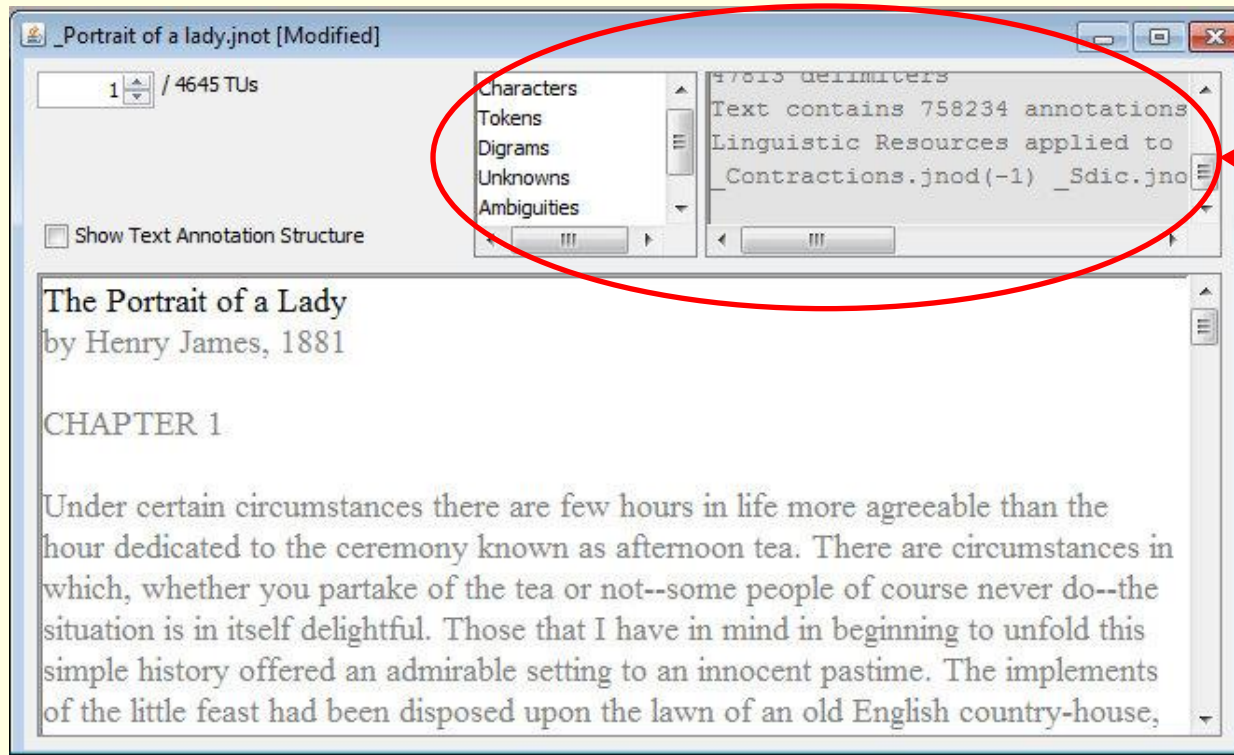


Linguistic
analysis

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 screenshot displays three windows from the NooJ V4 software. The 'Tokens in Portrait...' window shows a list of tokens with their frequencies. The 'Untitled [Modified]' window shows the NooJ V4 Dictionary configuration. The 'Digrams in Portrait o...' window shows a list of digrams with their frequencies.

Tokens in Portrait...

Freq	Tokens
7590	the
7297	to
6217	of
5406	a
5270	I
4337	her
4102	and
3921	she
3755	that
3663	you
3474	had
3454	was
3240	in
3086	it
2707	he
2595	s
1954	t
1912	as
1845	for
1827	with
1734	his

Untitled [Modified]

```
# NooJ V4
# Dictionary
# Input Language is:
# Alphabetical order
# Use inflectional &
# Special Command: #
# Special Features:
# Special Characters
#
abide, UNKNOWN
applausive, UNKNOWN
asphalte, UNKNOWN
auditress, UNKNOWN
averred, UNKNOWN
bedroomy, UNKNOWN
belawng, UNKNOWN
bellemere, UNKNOWN
bestown, UNKNOWN
```

Digrams in Portrait o...

Freq	Digrams
981	of the
734	don t
723	she had
717	in the
641	I m
536	I don
465	to be
450	he had
449	of her
441	had been
435	Madame Merle
430	to the
426	I ve
413	that she
412	it was
409	she was
409	to her
341	was not
325	of a
324	Lord Warburton

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

Locate a pattern in _Portrait of a lady.jnot

Pattern is:

- a string of characters:
- a PERL regular expression:
- a NooJ regular expression:
- a NooJ grammar:

friend

Set

Syntactic Analysis

Index

- Shortest matches
- Longest matches
- All matches

Limitation

- All occurrences
- Only: 100 occ.
- 1 occ. per match

Reset Concordance

N O B J

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.

Concordance for Text_Portrait of a lady.jnot

Reset Display: 5 characters before, and 5 after. Display: Matches Outputs
 word forms

Before	Seq.	After
it." said Lord Warburton's	friend	. "Is that true, sir?" asked
an interesting woman," said his	friend	. "My dear fellow, you can
see that Bunchie's new	friend	was a tall girl in
the other?" "He's a	friend	of ours--Lord Warburton." "
he regarded as his best	friend	. Ralph was not only fond
doubt of one's best	friend	: one should try to be
be one's own best	friend	and to give one's
and abide. She had a	friend	whose acquaintance she had
mentioned the fact to her	friend	, who would not have taken
staying here; she was a	friend	of Ralph's and he
only another proof of his	friend	's high abilities, which he
cousin and her cousin's	friend	. It must be added moreover
pay any more," said her	friend	: "he lives a monstrous deal
must talk." "He thinks you	friend	's too subversive--or not
received a note from her	friend	Miss Stackpole--a note of
"Here I am, my lovely	friend	," Miss Stackpole wrote; "I m
to that side of her	friend	's character which she regard

Query 201/201

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

Locate a pattern in _Portrait of a lady.jnot

Pattern is:

- a string of characters:
- a PERL regular expression:
- a NooJ regular expression:
- a NooJ grammar:

friend | money

Set

Syntactic Analysis

Index

- Shortest matches
- Longest matches
- All matches

Limitation

- All occurrences
- Only: 100 occ.
- 1 occ. per match

Reset Concordance

N O B J

Locating an expression in text

NooJ finds 266 matches for this query.

Concordance for Text _Portrait of a lady.jnot

Reset Display: characters before, and after. Display: Matches Outputs
 word forms

Before	Seq.	After
"why did you ever inherit	money	?" She stopped a moment as
even more than once found	money	for him; and the end
She looked at her young	friend	from head to foot, but
to avoid Pansy's other	friend	. Her companion grasped her arm
It was my uncle's	money	." "Yes; it was your uncle
it was your uncle's	money	, but it was your cousin
She had telegraphed to her	friend	from Turin, and though she
put her arm into her	friend	's. She remembered she had
remember he was an old	friend	of her cousin-that he
declared as she ushered her	friend	into a cab. And later
"He married me for the	money	," she said. She wished to
everything. He left her no	money	; of course she had no
she had no need of	money	. He left her the furniture
was to be sold. The	money	produced by the sale was
has left it to your	friend	Miss Stackpole-"in recognition of
know he's an old	friend	of yours, and as I

Query 266/266

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

The screenshot shows a window titled "Concordance for Text_Portrait of a lady.jnot". The search query is "(the | a) <WF> is". The window displays a table with three columns: "Before", "Seq.", and "After". The search results are as follows:

Before	Seq.	After
people of course never do-- smoothed it over his knees. "	the situation is	in itself delightful. Those
niece seems to prove that	The fact is	I've been comfortable so
hand. He wintered abroad, as	the allusion is	to one of my aunts
to whom the ear of	the phrase is	; basked in the sun, stopp
yet. When the necessity of	the world is	more directly presented t
extravagance. Her cousin used, as	a thing is	generally felt they usually
the pupil. The expression of	the phrase is	, to chaff her; he very
"One's right in such	a button is	not usually deemed huma
missed. It seems to me	a matter is	not measured by the time
up to the mark, and	the place is	about as full as it
any way, just now, and	the fact is	there's a good deal
sweetly as this compliment deserved. "	the worst is	that your putting it to
everything wrong. What sort of	The house is	so large and his room
her to feel herself, as	a cousin is	a cousin that you had
collection of old snuff-boxes.'	the phrase is	, under an influence. "Wh
	The collection is	all that's wanted to

At the bottom of the window, the status bar shows "Query" on the left and "32/32" in the center.

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

The screenshot shows a window titled "Concordance for Text _Portrait of a lady.jnot". The interface includes a "Reset" button, a "Display:" field set to "5", and radio buttons for "characters" and "word forms" (the latter is selected). It also has fields for "before, and" and "after" both set to "5", and checkboxes for "Matches" (checked) and "Outputs".

Before	Seq.	After
find out her special line.	Is it by	char
comings and goings. But it	is by	no n
a most original stamp. It	is true that when she described them to her cousin by	that
do so if your attention	is distracted by	irrel
for such a course it	is not discredited by	irrite
right in such a matter	is not measured by	the t
those things. The silver cross	is worn by	the c
At last he spoke again. "	Is your objection to my society this evening caused by	you
gentlemen of a race which	is not distinguished by	the c
the midst of his misery	is seized by	a ha

At the bottom, there is a "Query" field containing "is" and a page indicator "10/10".

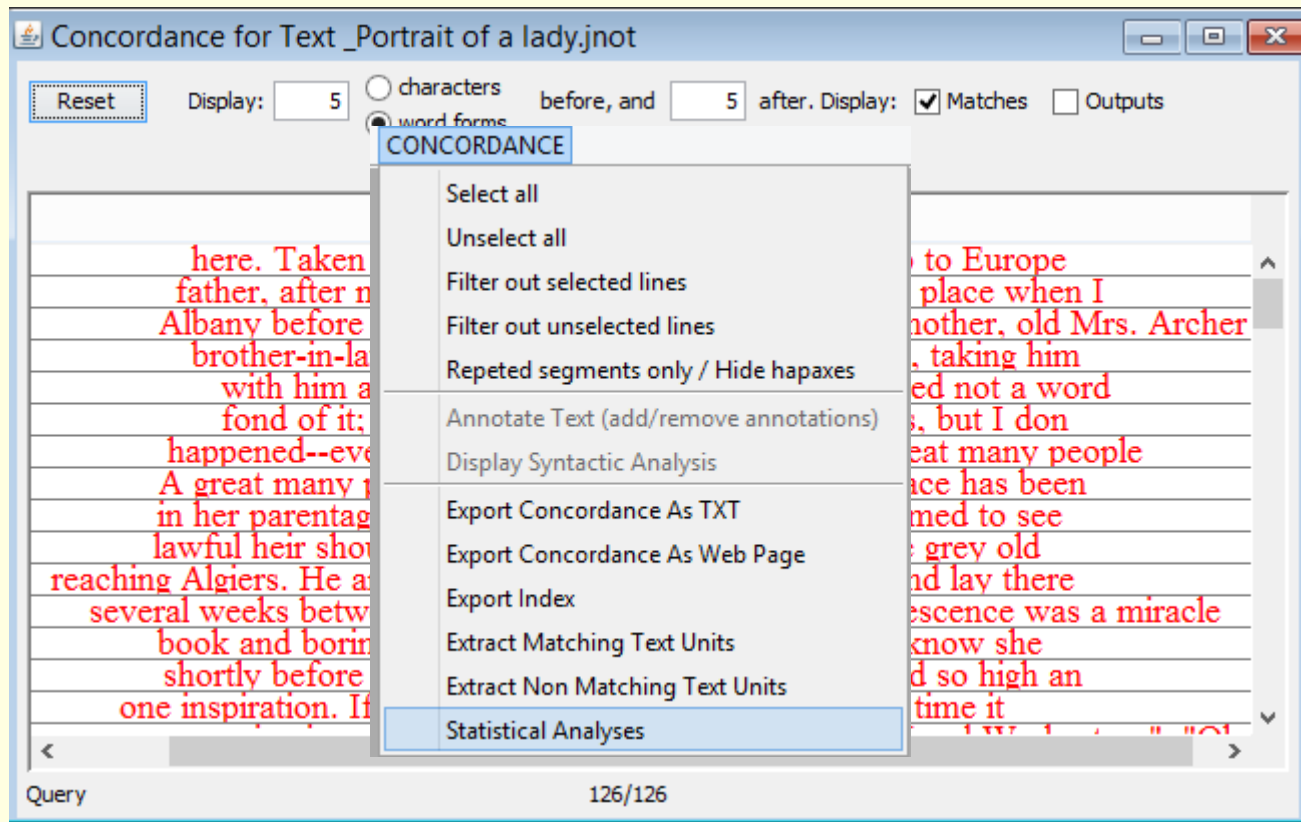
Locating an expression in text

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

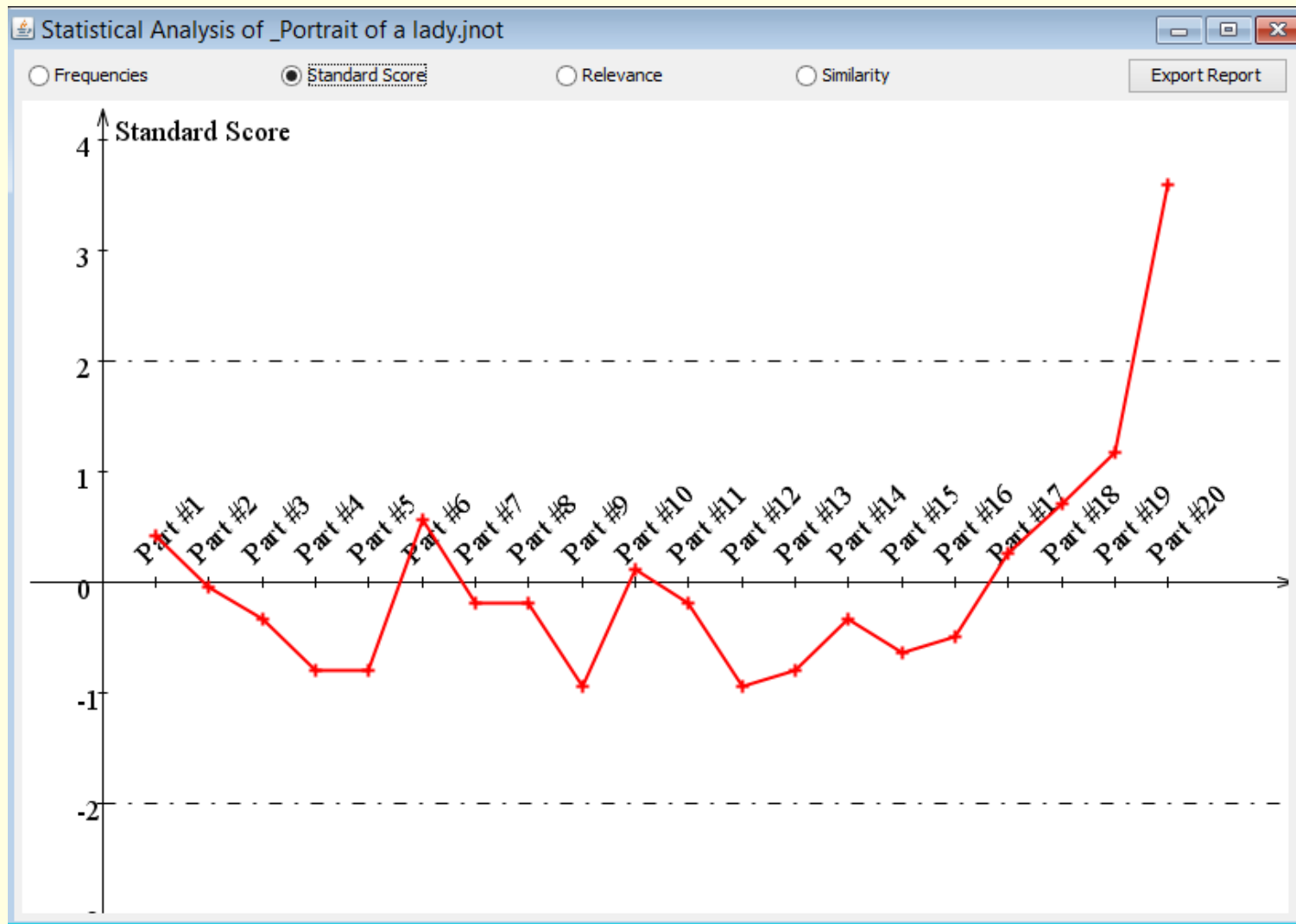
The screenshot shows a window titled "Locate a pattern in _Portrait of a lady.jnot". The interface is divided into several sections:

- Pattern is:** This section contains four radio buttons: "a string of characters:", "a PERL regular expression:", "a NooJ regular expression:" (which is selected), and "a NooJ grammar:". Below the selected option is a text input field containing the pattern "sad | sadness | cry | death | died | dead | die". Below the other options are empty input fields and a "Set" button.
- Syntactic Analysis:** A checkbox labeled "Syntactic Analysis" is currently unchecked.
- Index:** This section contains three radio buttons: "Shortest matches", "Longest matches" (which is selected), and "All matches".
- Limitation:** This section contains three radio buttons: "All occurrences" (which is selected), "Only: 100 occ." (with "100" in a text box), and "1 occ. per match".
- Reset Concordance:** A checkbox labeled "Reset Concordance" is checked.
- Buttons:** At the bottom right, there are four colored buttons: a red button with "N", a green button with "o", a blue button with "o", and a grey button with "J".

Locating an expression in text



Locating an expression in text





THANK YOU