EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Funded by
the European Union

14. Text Annotation Structure (Advanced)

This series of tutorials is based upon work from COST Action
Multi3Generation CA18231, supported by COST
(European Cooperation in Science and Technology).

COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation, cf. www.cost.eu

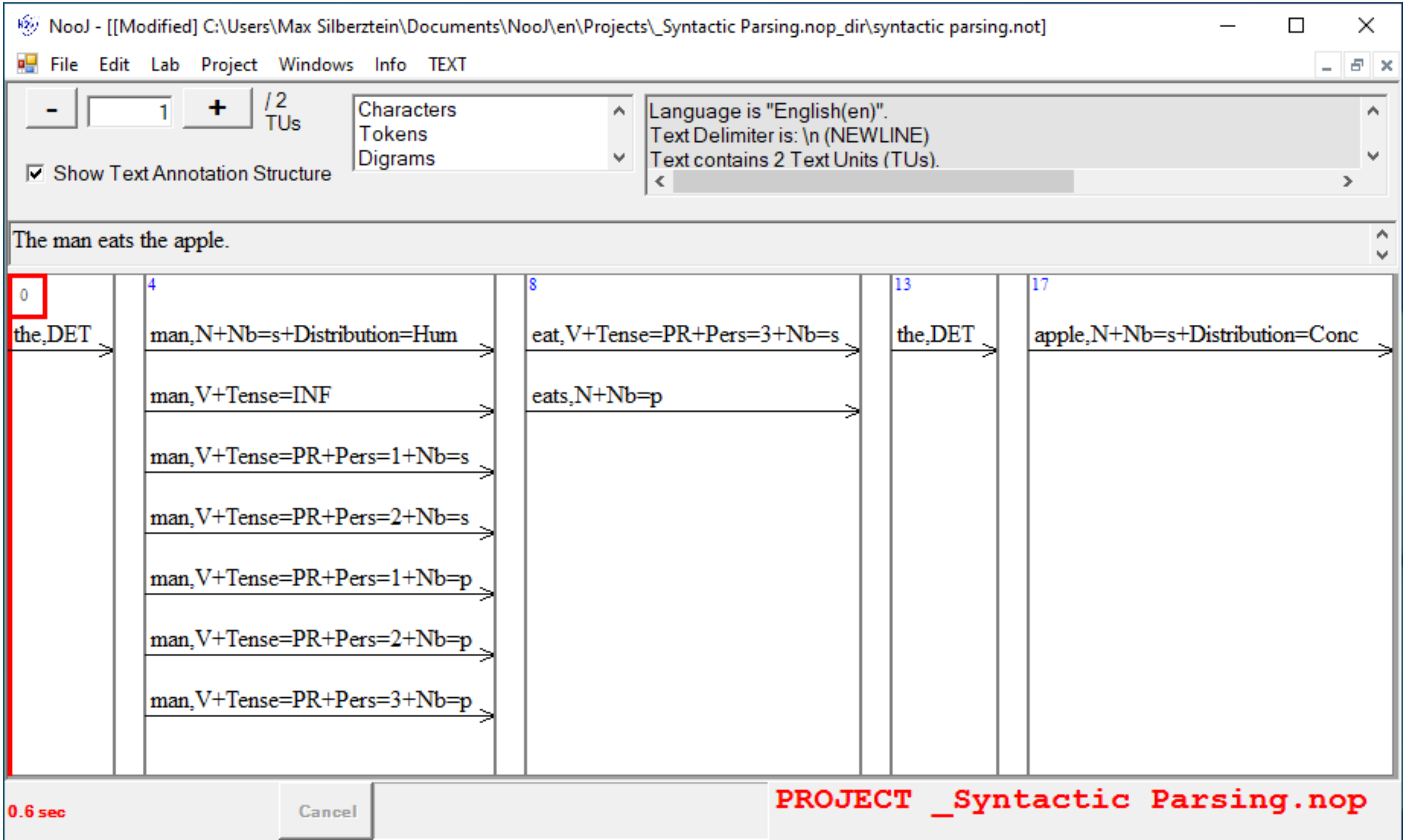
Text Annotation Structure

- When applying any linguistic resource to a text, NooJ adds, or removes annotations to the Text Annotation Structure (TAS)
- During the lexical analysis, NooJ applies dictionaries and morphological grammars to the text, to add annotations that represent ALUs into the TAS
- When applying syntactic grammars to a text, NooJ can add annotations (e.g., structural), or remove annotations (*e.g.*, lexical hypotheses).

Syntactic Analysis

Text, right after the lexical analysis

man: noun or verb (several forms); *eat*: noun or verb


NooJ - [[Modified] C:\Users\Max Silberztein\Documents\NooJ\en\Projects_Syntactic Parsing.nop_dir\syntactic parsing.not]
File Edit Lab Project Windows Info TEXT
-
1
+
/ 2 TUs
☑ Show Text Annotation Structure
Characters
Tokens
Digrams
Cancel

Language is "English(en)".
Text Delimiter is: \n (NEWLINE)
Text contains 2 Text Units (TUs).

The man eats the apple.

0
the,DET

4
man,N+Nb=s+Distribution=Hum
man,V+Tense=INF
man,V+Tense=PR+Pers=1+Nb=s
man,V+Tense=PR+Pers=2+Nb=s
man,V+Tense=PR+Pers=1+Nb=p
man,V+Tense=PR+Pers=2+Nb=p
man,V+Tense=PR+Pers=3+Nb=p

8
eat,V+Tense=PR+Pers=3+Nb=s
eats,N+Nb=p

13
the,DET

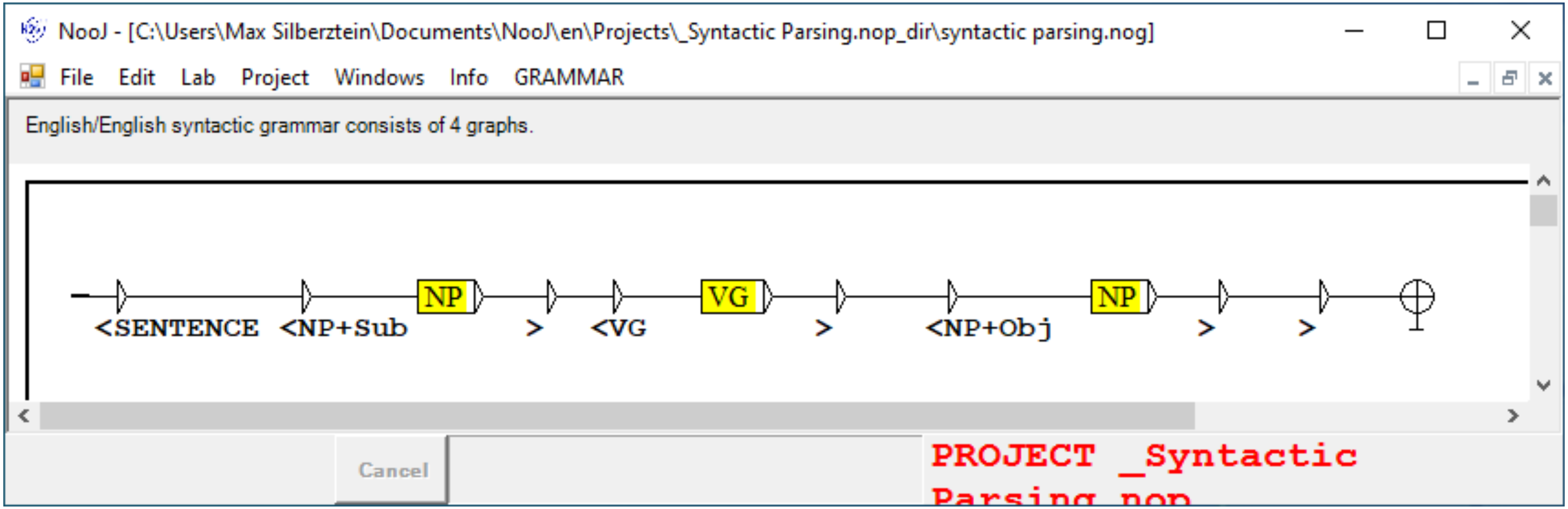
17
apple,N+Nb=s+Distribution=Conc

0.6 sec

PROJECT _Syntactic Parsing.nop

Syntactic Analysis

Syntactic grammar produces structural annotations



NooJ - [C:\Users\Max Silberztein\Documents\NooJ\en\Projects_Syntactic Parsing.nop_dir\syntactic parsing.nog]
File
Edit
Lab
Project
Windows
Info
GRAMMAR

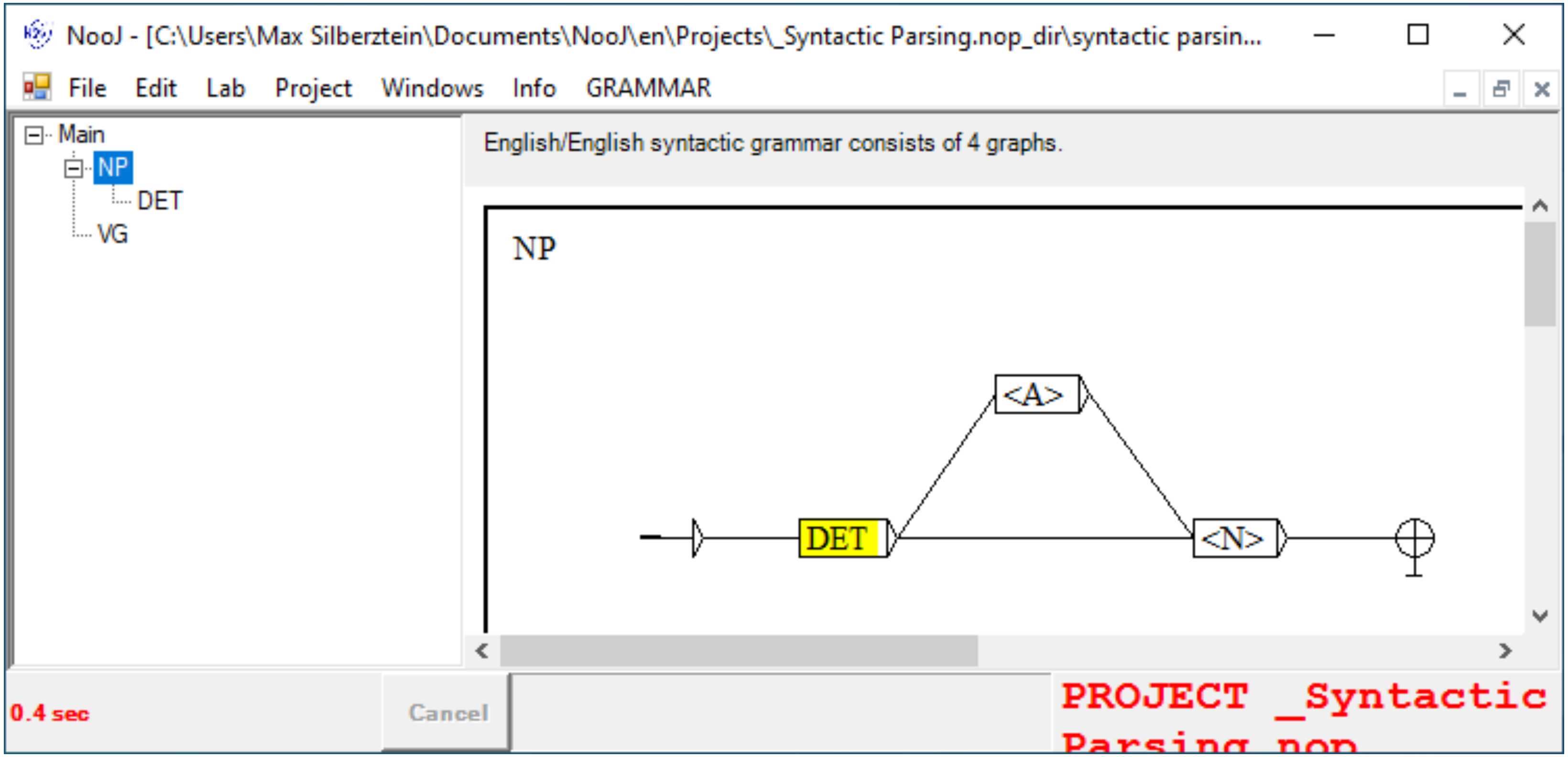
English/English syntactic grammar consists of 4 graphs.

<SENTENCE
<NP+Sub
NP
>
<VG
VG
>
<NP+Obj
NP
>
>
Cancel
PROJECT _Syntactic
Parsing.nop

- Annotation <SENTENCE>
- Annotation <NB+Sub>
- Annotation <VG>
- Annotation <NB+Obj>

Syntactic Analysis

Syntactic grammar produces structural annotations



NooJ - [C:\Users\Max Silberztein\Documents\NooJ\en\Projects_Syntactic Parsing.nop_dir\syntactic parsin...
File
Edit
Lab
Project
Windows
Info
GRAMMAR
Main
NP
DET
VG
NP
DET
<A>
<N>
0.4 sec
Cancel

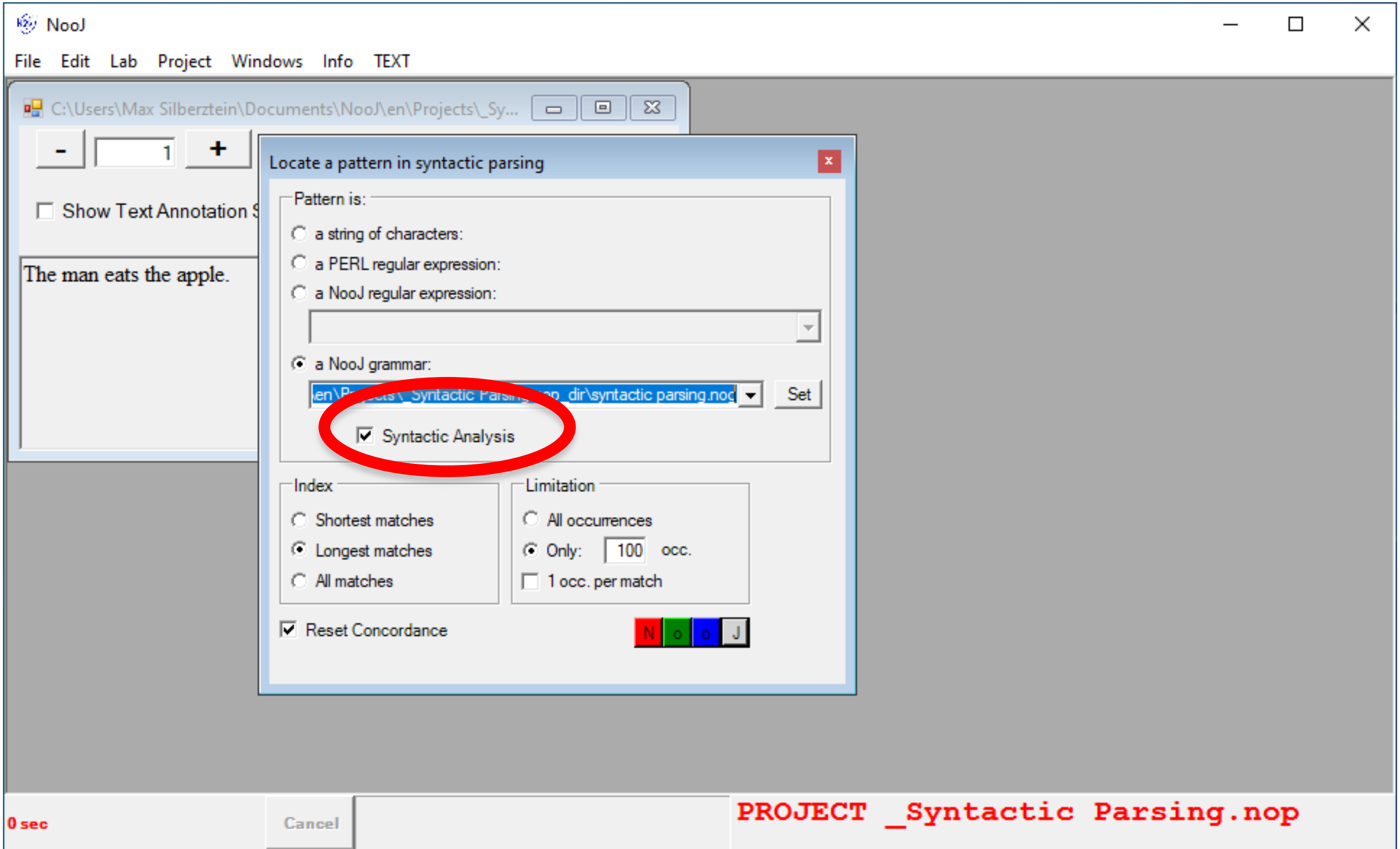
English/English syntactic grammar consists of 4 graphs.

PROJECT _Syntactic Parsing.nop

- Several embedded grammars

Syntactic Analysis

Applying grammar to text



☐ Show Text Annotation S

The man eats the apple.

Locate a pattern in syntactic parsing

Pattern is:

○ a string of characters:

○ a PERL regular expression:

○ a NooJ regular expression:

● a NooJ grammar: \en\Projects_Syntactic Parsing.nop_dir\syntactic parsing.nod

☑ Syntactic Analysis

Index

○ Shortest matches

● Longest matches

○ All matches

Limitation

○ All occurrences

● Only: 100 occ.

☐ 1 occ. per match

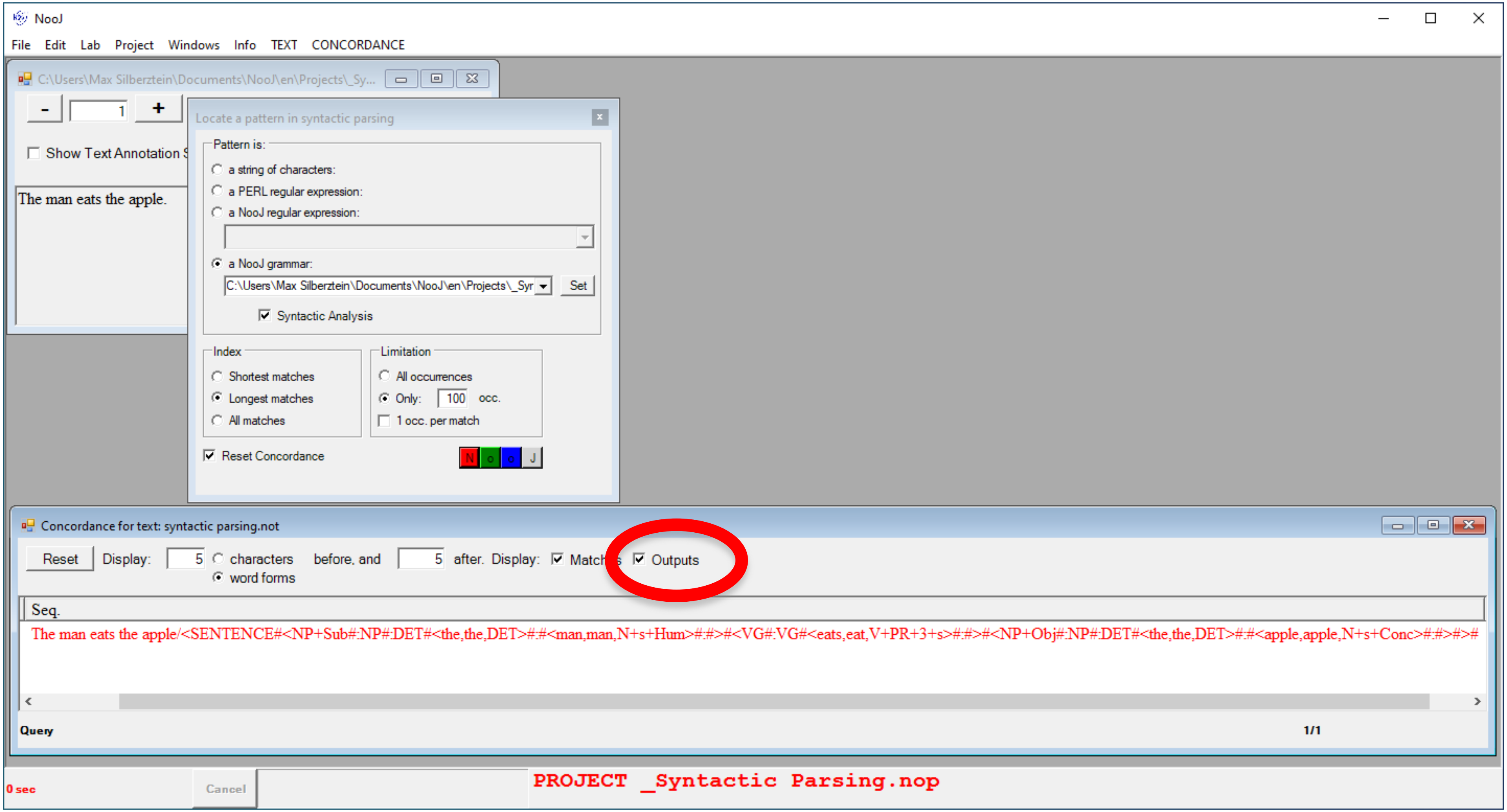
☑ Reset Concordance

0 sec

PROJECT _Syntactic Parsing.nop

Syntactic Analysis

Applying grammar to text


NooJ
File Edit Lab Project Windows Info TEXT CONCORDANCE
C:\Users\Max Silberztein\Documents\NooJ\en\Projects_Sy...
- 1 +
Set
Reset
Query
1/1
0 sec
Cancel

☐ Show Text Annotation S

The man eats the apple.

Locate a pattern in syntactic parsing

Pattern is:

○ a string of characters:

○ a PERL regular expression:

○ a NooJ regular expression:

● a NooJ grammar: C:\Users\Max Silberztein\Documents\NooJ\en\Projects_Syr

☑ Syntactic Analysis

Index

○ Shortest matches

● Longest matches

○ All matches

Limitation

○ All occurrences

● Only: 100 occ.

☐ 1 occ. per match

☑ Reset Concordance

🟥 N 🟩 o 🟦 o J

Concordance for text: syntactic parsing.not

Display: 5 ○ characters ● word forms before, and 5 after. Display: ☑ Matches ☑ Outputs

Seq.

The man eats the apple/<SENTENCE#<NP+Sub#:NP#:DET#<the,the,DET>#:#<man,man,N+s+Hum>#:#>#<VG#:VG#<eats,eat,V+PR+3+s>#:#>#<NP+Obj#:NP#:DET#<the,the,DET>#:#<apple,apple,N+s+Conc>#:#>#>#

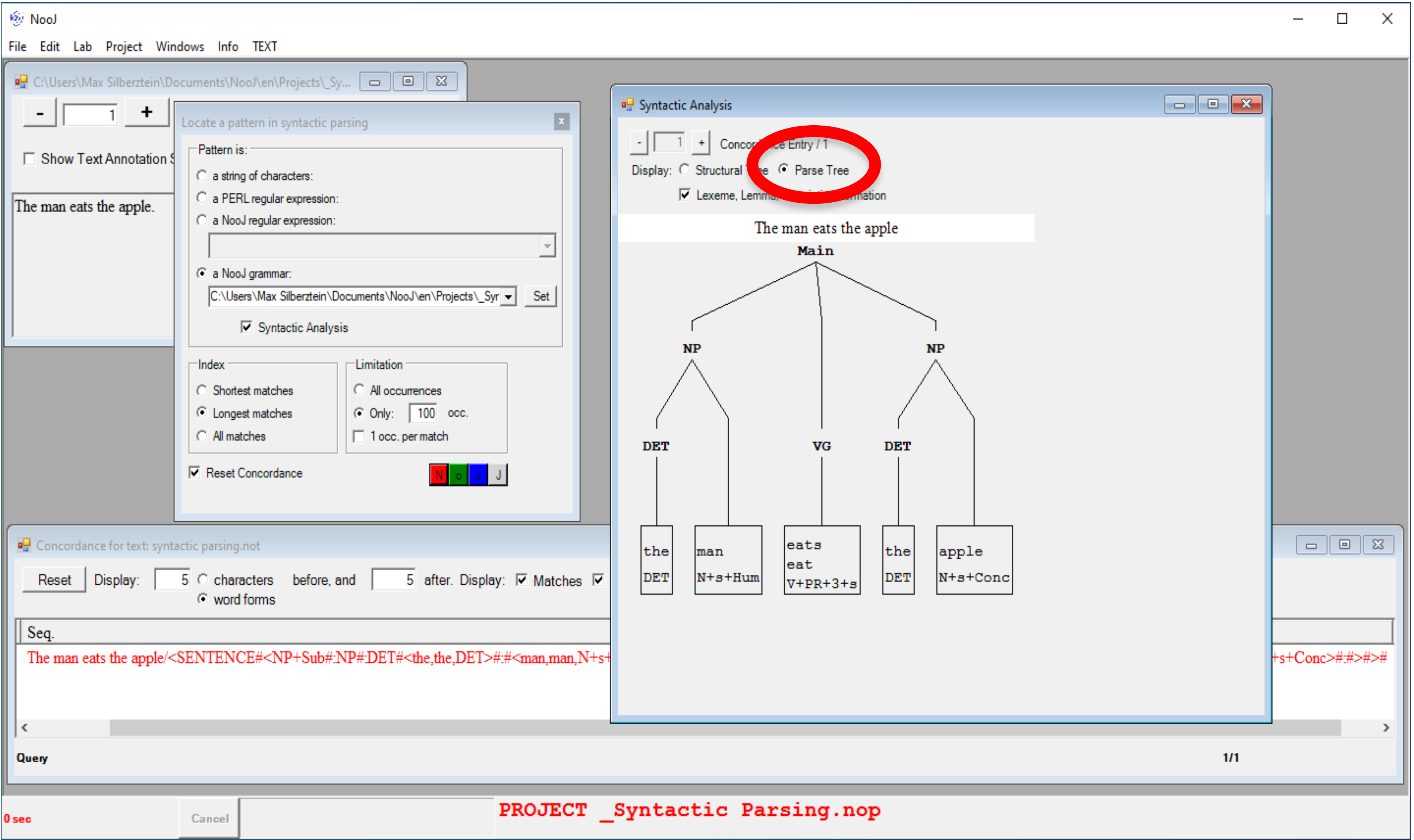
PROJECT _Syntactic Parsing.nop

The output represents the structure produced by the annotations

Syntactic Analysis

CONCORDANCE > Display Syntactic Analysis

NooJ displays the parse tree:



C:\Users\Max Silberztein\Documents\NooJ\en\Projects_Sy...

- 1 +

☐ Show Text Annotation S

The man eats the apple.

Locate a pattern in syntactic parsing

Pattern is:

○ a string of characters:

○ a PERL regular expression:

○ a NooJ regular expression:

● a NooJ grammar:

C:\Users\Max Silberztein\Documents\NooJ\en\Projects_Syn Set

☑ Syntactic Analysis

Index

○ Shortest matches

● Longest matches

○ All matches

Limitation

○ All occurrences

● Only: 100 occ.

☐ 1 occ. per match

☑ Reset Concordance

Syntactic Analysis

- 1 + Concordance Entry / 1

Display: ○ Structural Tree ● Parse Tree

☑ Lexeme, Lemma, ... Information

The man eats the apple

the DET	man N+s+Hum	eats eat V+PR+3+s	the DET	apple N+s+Conc

Concordance for text: syntactic parsing.not

Display: 5 ○ characters before, and 5 after. Display: ☑ Matches ☑

● word forms

Seq.

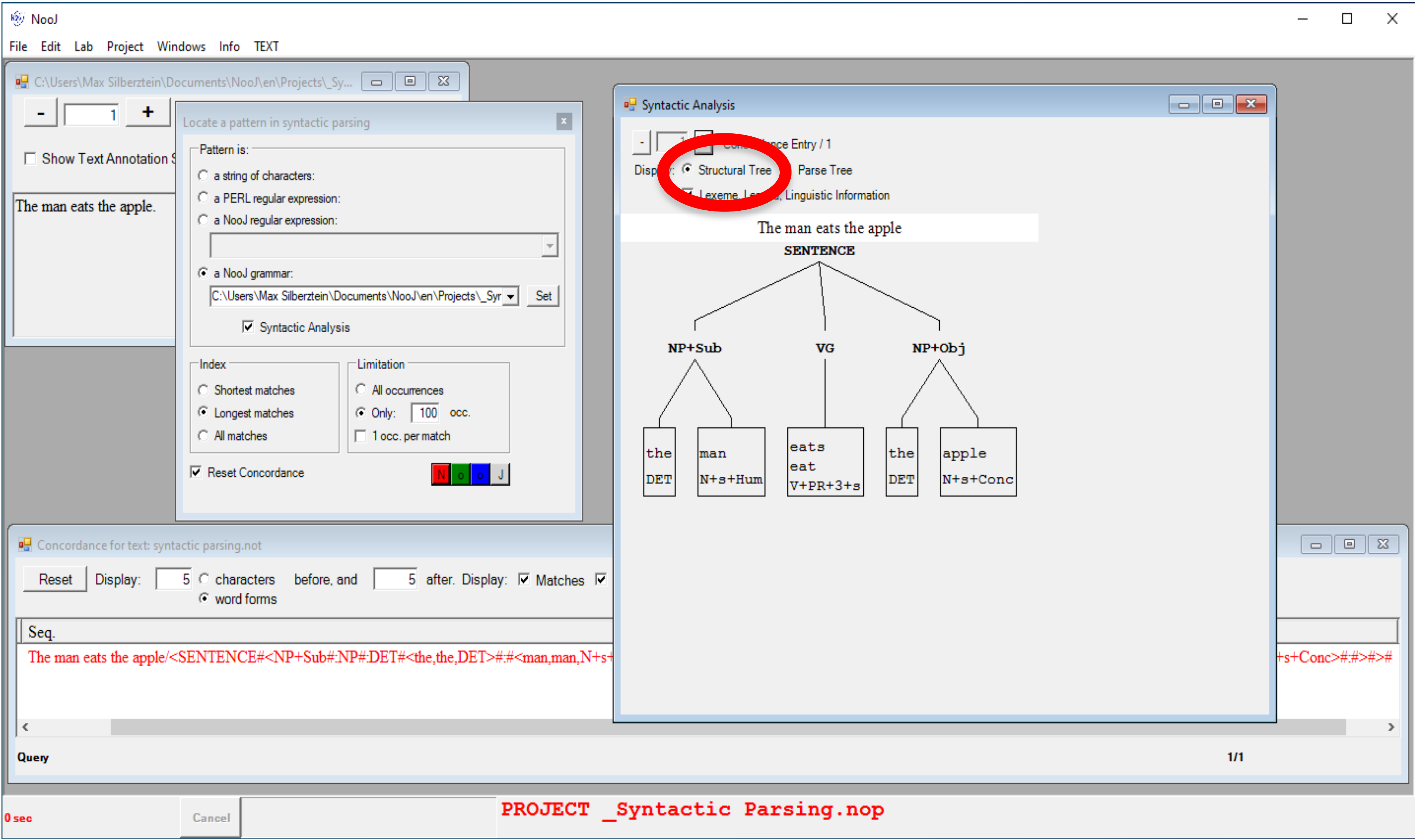
The man eats the apple/<SENTENCE#<NP+Sub#:NP#:DET#<the,the,DET>#:#<man,man,N+s+ ... +s+Conc>#:#>#>#

PROJECT _Syntactic Parsing.nop

Syntactic Analysis

CONCORDANCE > Display Syntactic Analysis

The TAS is displayed as a tree:


NooJ
File Edit Lab Project Windows Info TEXT
C:\Users\Max Silberztein\Documents\NooJ\en\Projects_Sy...
- 1 +
Locate a pattern in syntactic parsing
Syntactic Analysis
Concordance for text: syntactic parsing.not
Reset
Seq.
Query
1/1
0 sec
Cancel
N o o J
SENTENCE
NP+Sub
VG
NP+Obj
the
DET
man
N+s+Hum
eats
eat
V+PR+3+s
the
DET
apple
N+s+Conc

☐ Show Text Annotation S

The man eats the apple.

Pattern is:

○ a string of characters:

○ a PERL regular expression:

○ a NooJ regular expression:

● a NooJ grammar: C:\Users\Max Silberztein\Documents\NooJ\en\Projects_Syr

☑ Syntactic Analysis

Index: ○ Shortest matches ● Longest matches ○ All matches

Limitation: ○ All occurrences ● Only: 100 occ. ☐ 1 occ. per match

☑ Reset Concordance

Display: ● Structural Tree ○ Parse Tree

☑ Lexeme, Lemma, Linguistic Information

The man eats the apple

Display: 5 ○ characters ● word forms before, and 5 after. Display: ☑ Matches ☑

The man eats the apple/<SENTENCE#<NP+Sub#:NP#:DET#<the,the,DET>#:#<man,man,N+s+ … +s+Conc>#:#>#>#

PROJECT _Syntactic Parsing.nop

Syntactic Analysis
Parse *vs.* Structural trees

- The **parse tree** represents the structure of the grammar, rather than the structure of the sentence.

➢ It is useful to debug a grammar, as it shows how the grammar was explored during parsing

- The **structural tree** represents the TAS. It is produced by annotations in the grammar, and is independent from the structure of the grammar.

➢ It is useful to accumulate and share grammars, as it is independent from how linguists want to organize their grammar; it also allows NooJ to optimize grammars (*e.g.*, remove useless rules and recursions, etc.) without any consequence

Syntactic Analysis
Atomic Linguistic Units

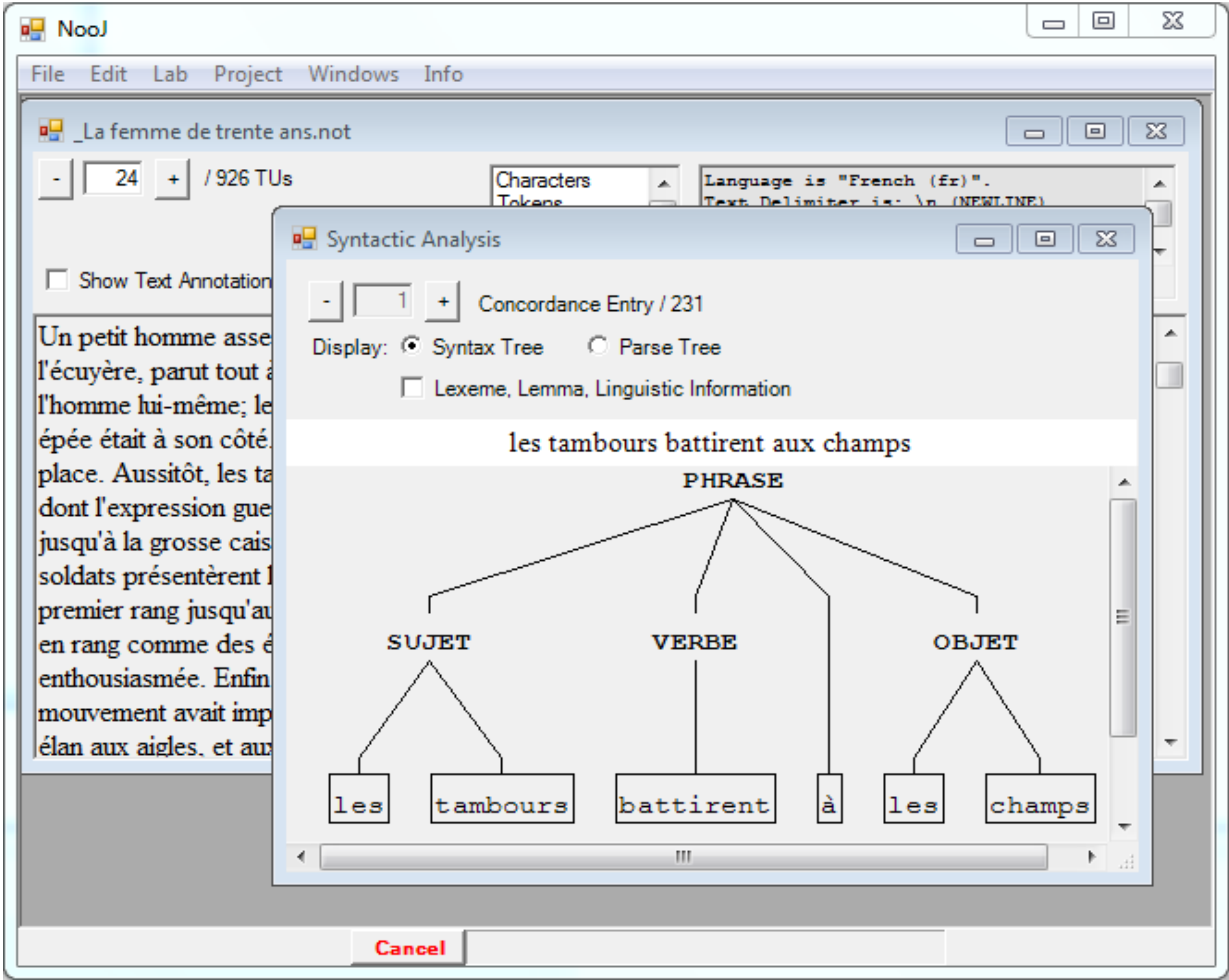
- There are four types of ALUs: affixes, simple words, multiword units and discontinuous expressions.

➢ ALUs are represented by annotations in the TAS

- Syntactic trees must represent the ALUs

Syntactic Analysis
Contracted and agglutinated forms

In French, *aux* is a contracted form of *à les*


NooJ
File Edit Lab Project Windows Info
_La femme de trente ans.not
- 24 + / 926 TUs
Characters
Tokens
☐ Show Text Annotation
Syntactic Analysis
- 1 + Concordance Entry / 231
Display: ● Syntax Tree ○ Parse Tree
☐ Lexeme, Lemma, Linguistic Information
PHRASE
SUJET
VERBE
OBJET
les
tambours
battirent
à
les
champs
Cancel

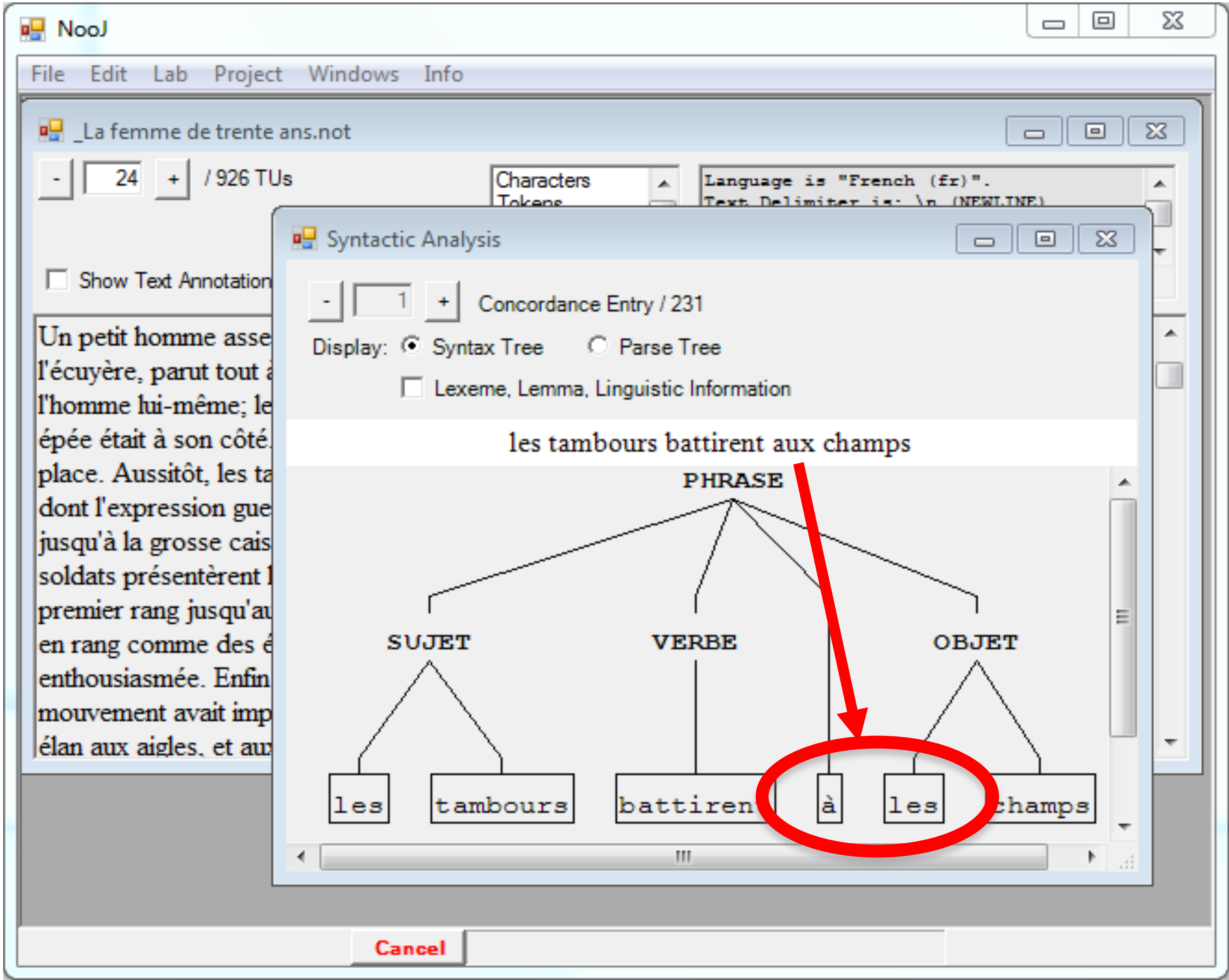
Language is "French (fr)".
Text Delimiter is: \n (NEWLINE)

Un petit homme asse
l'écuyère, parut tout à
l'homme lui-même; le
épée était à son côté
place. Aussitôt, les ta
dont l'expression gue
jusqu'à la grosse cais
soldats présentèrent l
premier rang jusqu'au
en rang comme des é
enthousiasmée. Enfin
mouvement avait imp
élan aux aigles, et aux

les tambours battirent aux champs

Syntactic Analysis
Contracted and agglutinated forms

In French, *aux* is a contracted form of *à les*



- 24 + / 926 TUs

Language is "French (fr)".
Text Delimiter is: \n (NEWLINE)

☐ Show Text Annotation

Un petit homme asse
l'écuyère, parut tout à
l'homme lui-même; le
épée était à son côté.
place. Aussitôt, les ta
dont l'expression gue
jusqu'à la grosse cais
soldats présentèrent l
premier rang jusqu'au
en rang comme des é
enthousiasmée. Enfin
mouvement avait imp
élan aux aigles, et aux

- 1 + Concordance Entry / 231

Display: ● Syntax Tree ○ Parse Tree

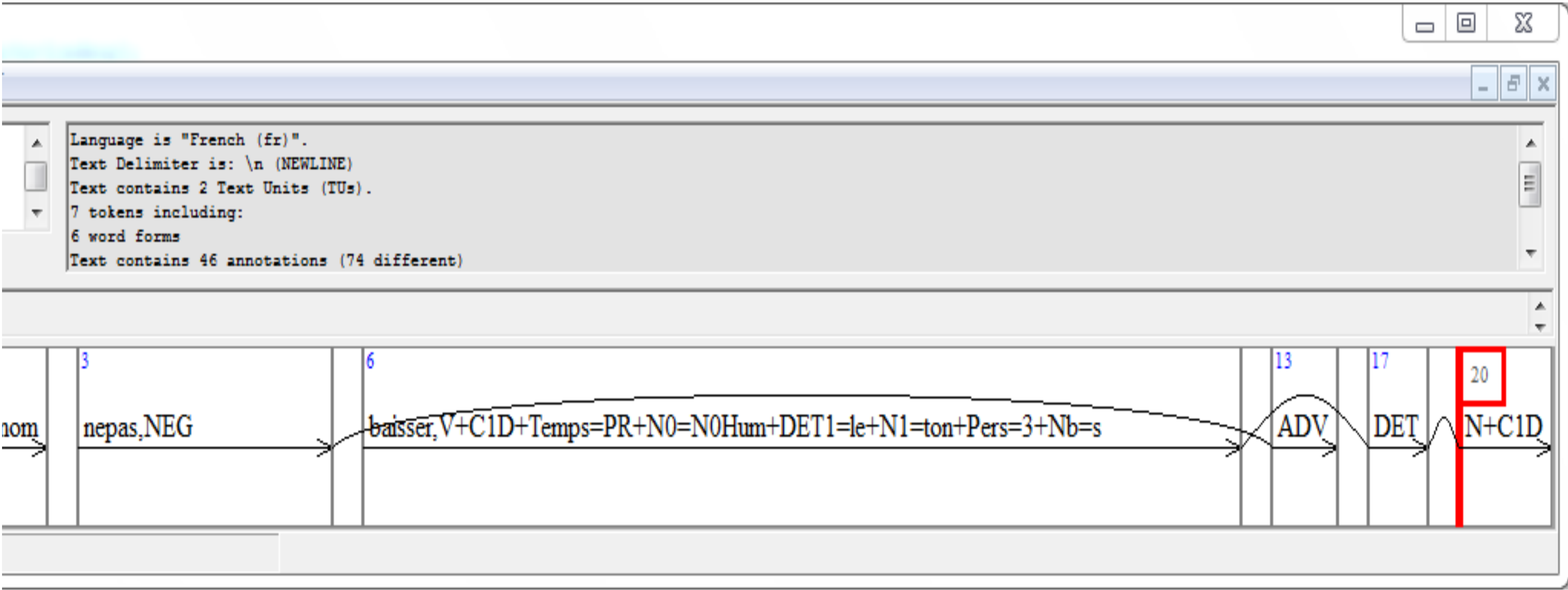
☐ Lexeme, Lemma, Linguistic Information

les tambours battirent aux champs

Syntactic Analysis
Discontinuous expressions

- *Il ne baisse pas le ton* [he does not lower his voice]
- In French, the negation *ne ... pas* is discontinuous.
- In French, the frozen expression *baisser ... le ton* is discontinuous.

The TAS:

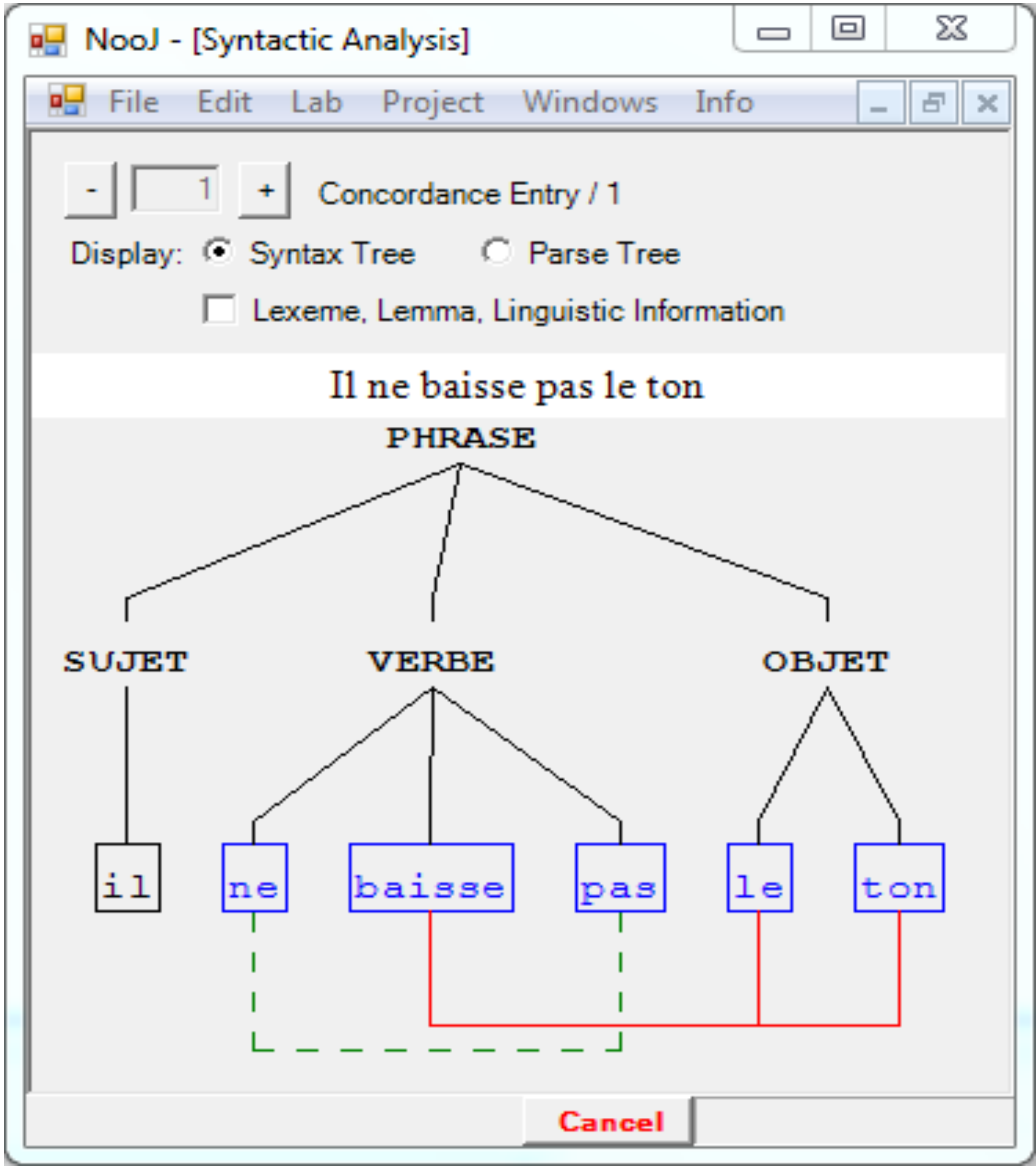


```
Language is "French (fr)".
Text Delimiter is: \n (NEWLINE)
Text contains 2 Text Units (TUs).
7 tokens including:
6 word forms
Text contains 46 annotations (74 different)
```

Syntactic Analysis
Discontinuous expressions

- The negation *ne ... pas* is discontinuous.

- The frozen expression *baisser ... le ton* is discontinuous.



Concordance Entry / 1

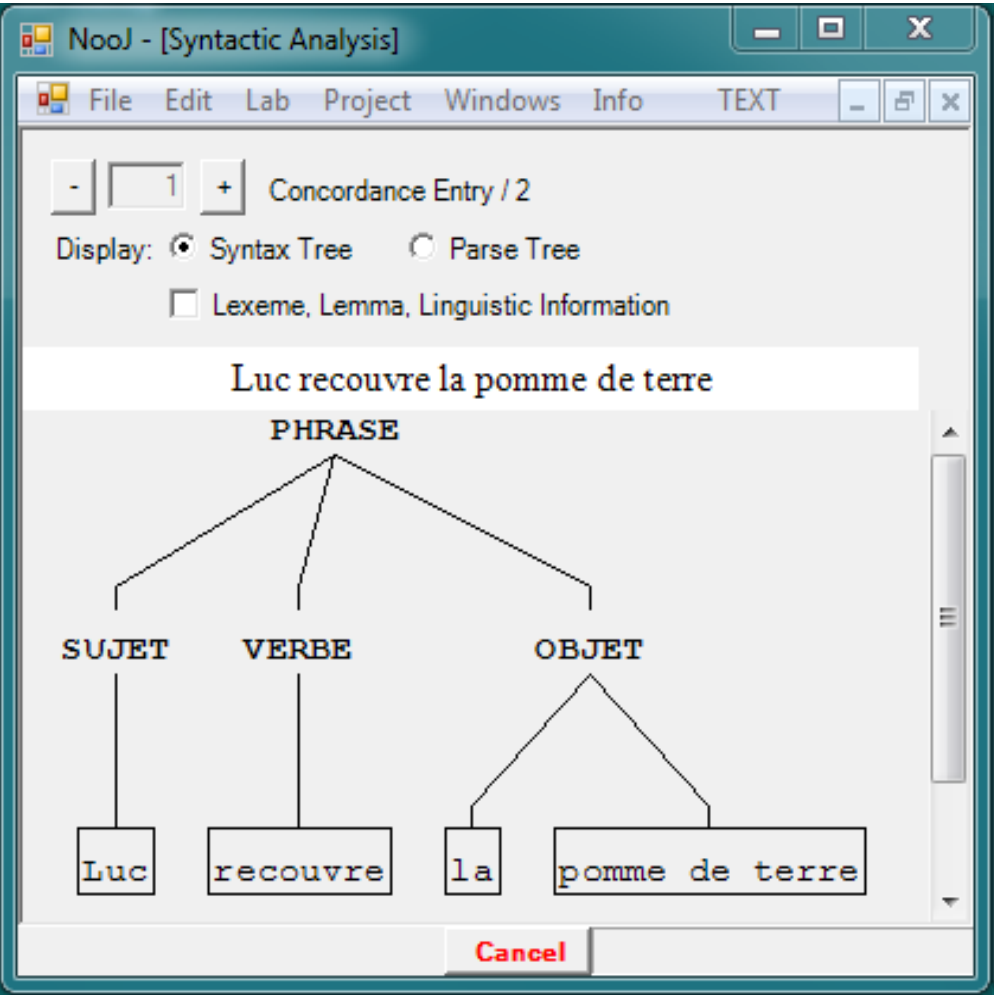
Display: ● Syntax Tree ○ Parse Tree

☐ Lexeme, Lemma, Linguistic Information

Il ne baisse pas le ton

Syntactic Analysis
Ambiguities

Luc recouvre la pomme de terre
[Luc covers the apple with earth] or *[Luc covers the potato]*

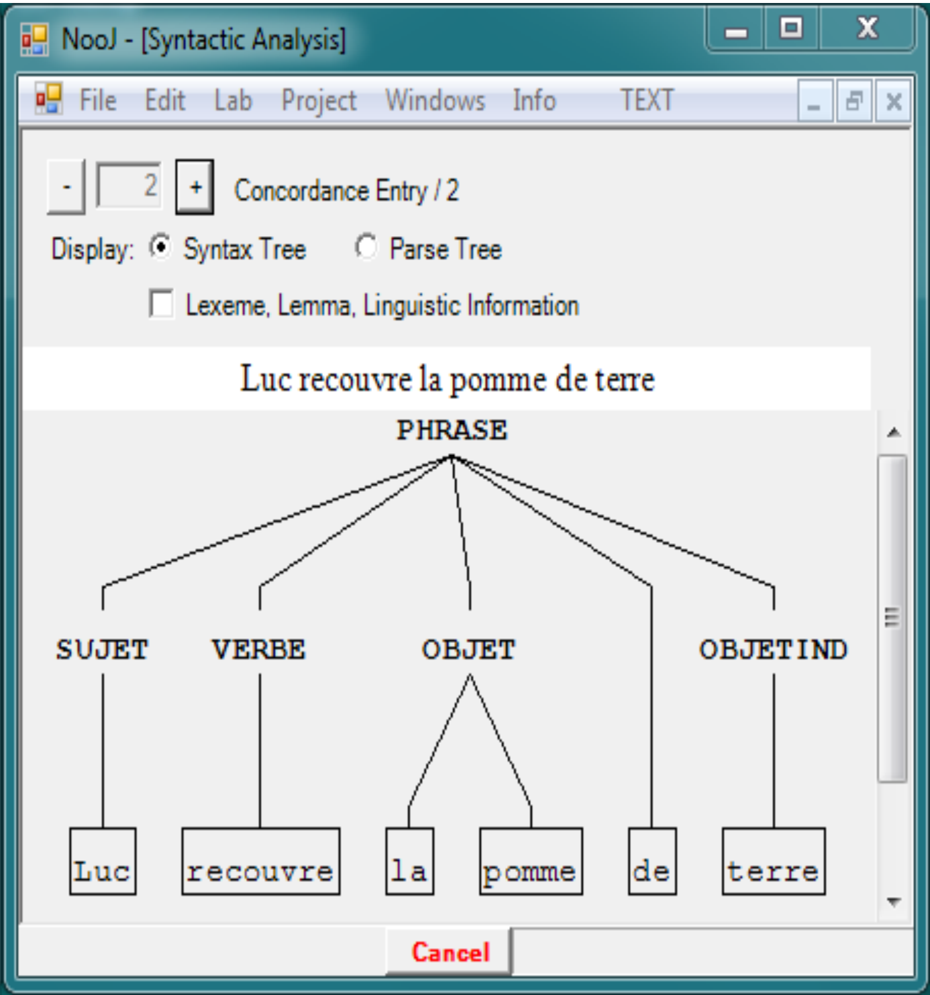

NooJ - [Syntactic Analysis]
File Edit Lab Project Windows Info TEXT
PHRASE
SUJET
VERBE
OBJET
Luc
recouvre
la
pomme de terre
Cancel

- 1 + Concordance Entry / 2

Display: ● Syntax Tree ○ Parse Tree

☐ Lexeme, Lemma, Linguistic Information

Luc recouvre la pomme de terre


NooJ - [Syntactic Analysis]
File Edit Lab Project Windows Info TEXT
PHRASE
SUJET
VERBE
OBJET
OBJETIND
Luc
recouvre
la
pomme
de
terre
Cancel

- 2 + Concordance Entry / 2

Display: ● Syntax Tree ○ Parse Tree

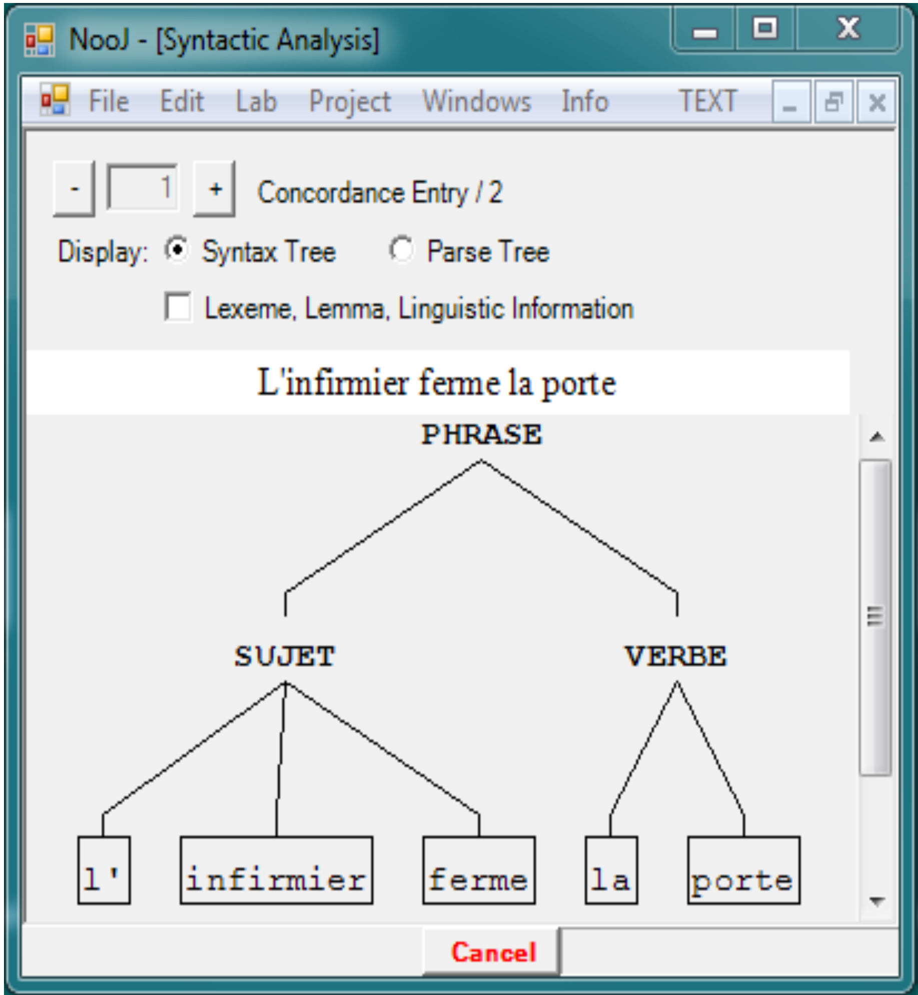
☐ Lexeme, Lemma, Linguistic Information

Luc recouvre la pomme de terre

Syntactic Analysis
Ambiguities

L'infirmier ferme la porte

[The nurse closes the door] or *[the firm nurse carries her]*



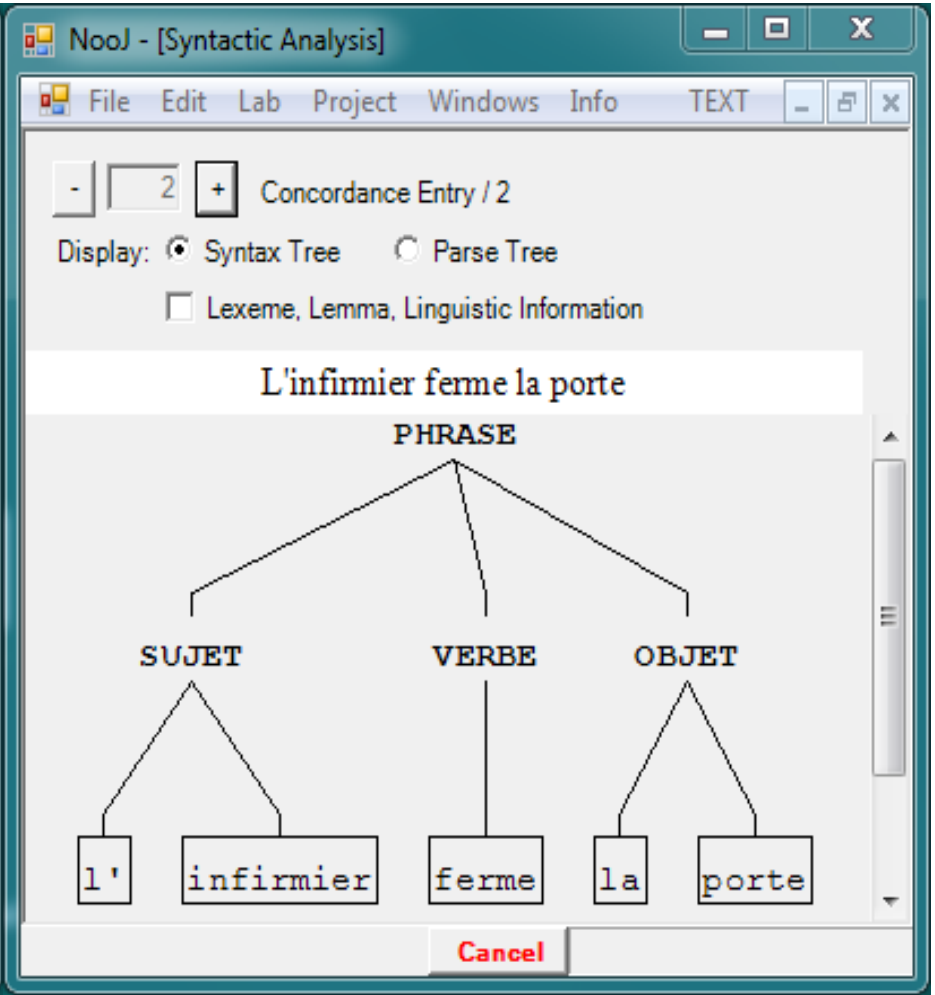
Concordance Entry 1 / 2

Display: ● Syntax Tree ○ Parse Tree

☐ Lexeme, Lemma, Linguistic Information

L'infirmier ferme la porte

NooJ - [Syntactic Analysis]
File Edit Lab Project Windows Info TEXT
PHRASE
SUJET
VERBE
l'
infirmier
ferme
la
porte
Cancel



Concordance Entry 2 / 2

Display: ● Syntax Tree ○ Parse Tree

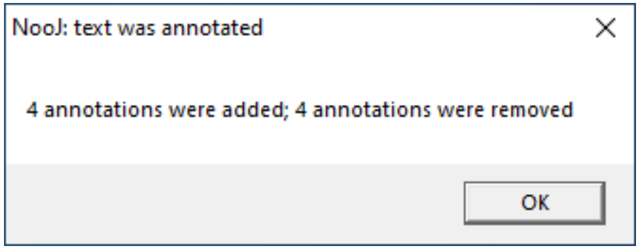
☐ Lexeme, Lemma, Linguistic Information

L'infirmier ferme la porte

NooJ - [Syntactic Analysis]
File Edit Lab Project Windows Info TEXT
PHRASE
SUJET
VERBE
OBJET
l'
infirmier
ferme
la
porte
Cancel

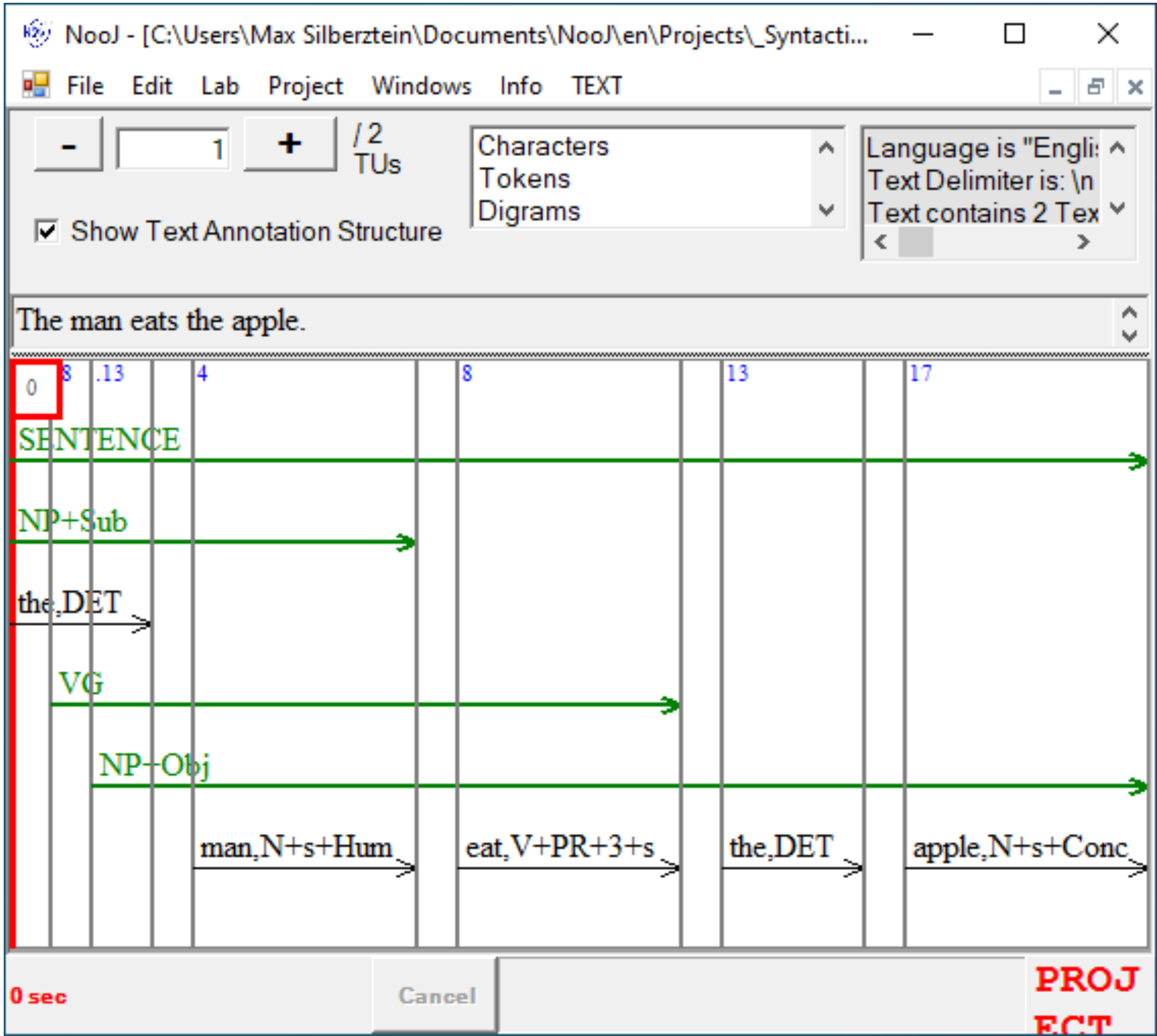
Syntactic Analysis
Disambiguation

CONCORDANCE > Annotate Text (add/remove annotations)



NooJ: text was annotated

4 annotations were added; 4 annotations were removed



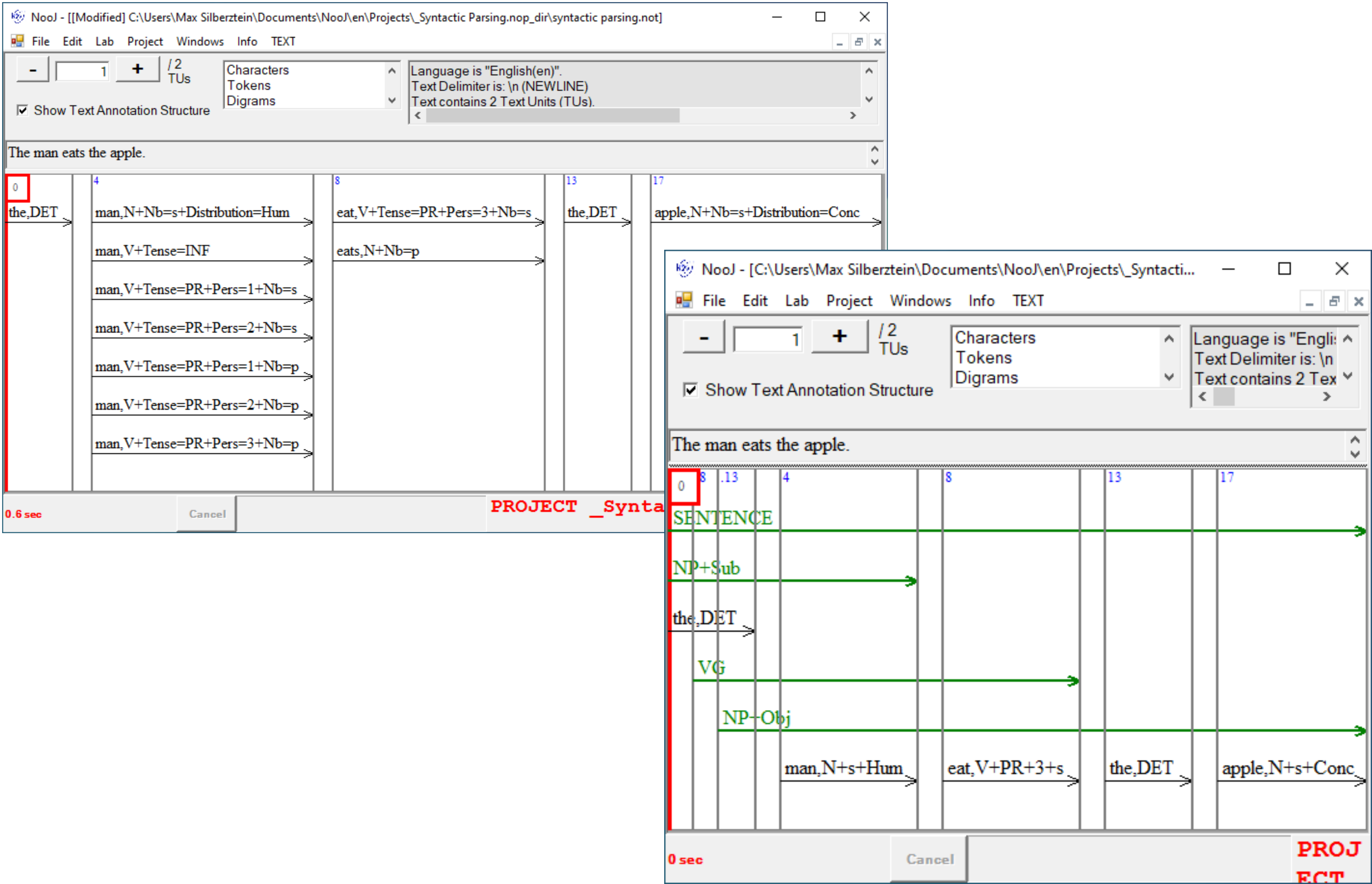
NooJ - [C:\Users\Max Silberztein\Documents\NooJ\en\Projects_Syntacti...

☑ Show Text Annotation Structure

Language is "Engli
Text Delimiter is: \n
Text contains 2 Tex

The man eats the apple.

Syntactic Analysis
Disambiguation: before and after



NooJ - [[Modified] C:\Users\Max Silberztein\Documents\NooJ\en\Projects_Syntactic Parsing.nop_dir\syntactic parsing.not]

Language is "English(en)".
Text Delimiter is: \n (NEWLINE)
Text contains 2 Text Units (TUs).

The man eats the apple.

0	4	8	13	17
the,DET	man,N+Nb=s+Distribution=Hum	eat,V+Tense=PR+Pers=3+Nb=s	the,DET	apple,N+Nb=s+Distribution=Conc
	man,V+Tense=INF	eats,N+Nb=p		
	man,V+Tense=PR+Pers=1+Nb=s			
	man,V+Tense=PR+Pers=2+Nb=s			
	man,V+Tense=PR+Pers=1+Nb=p			
	man,V+Tense=PR+Pers=2+Nb=p			
	man,V+Tense=PR+Pers=3+Nb=p			

0.6 sec

PROJECT _Synta

NooJ - [C:\Users\Max Silberztein\Documents\NooJ\en\Projects_Syntacti...

Language is "Engli
Text Delimiter is: \n
Text contains 2 Tex

The man eats the apple.

0 8 13 4 8 13 17

SENTENCE

NP+Sub

the,DET

VG

NP+Obj

man,N+s+Hum

eat,V+PR+3+s

the,DET

apple,N+s+Conc

0 sec

PROJ
ECT

CONGRATULATIONS

You know how to perform various lexical, morphological, syntactic and semantic analyses by annotating texts with various types of information