

# 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](http://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 Silberstein\Documents\NooJ\en\Projects\\_Syntactic Parsing.nop\_dir\syntactic parsing.not]

File Edit Lab Project Windows Info TEXT

- 1 + / 2 TUs

Characters Tokens Digrams

☒ Show Text Annotation Structure

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

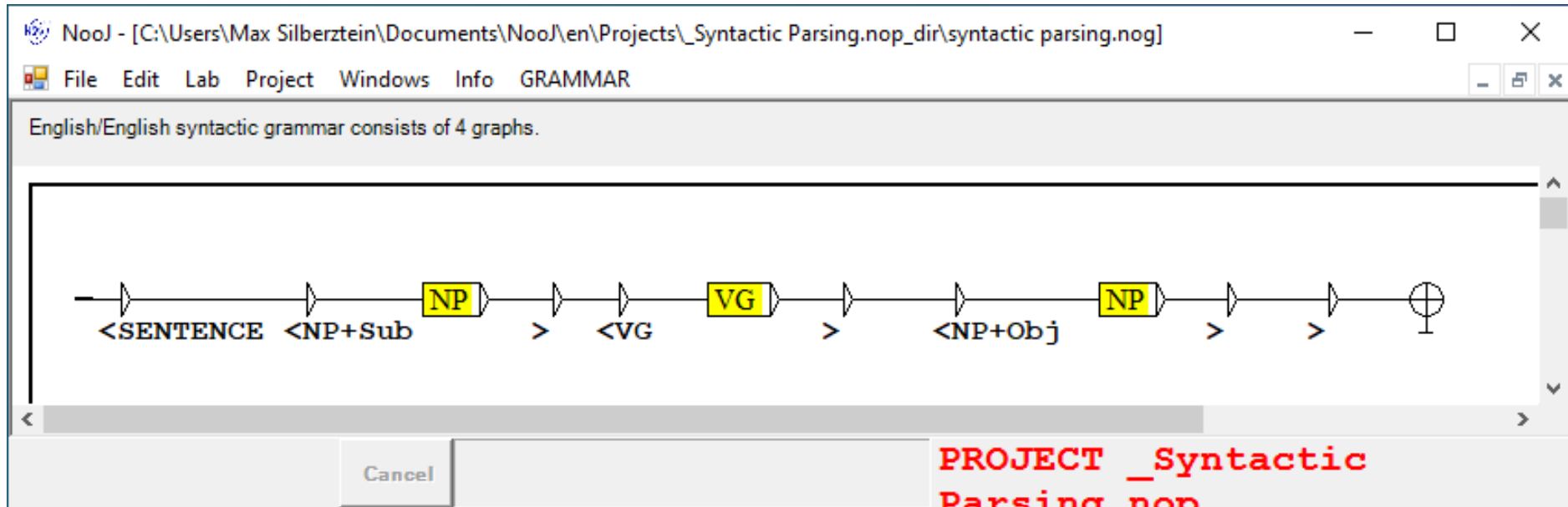
The man eats the apple.

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 Cancel PROJECT \_Syntactic Parsing.nop

# Syntactic Analysis

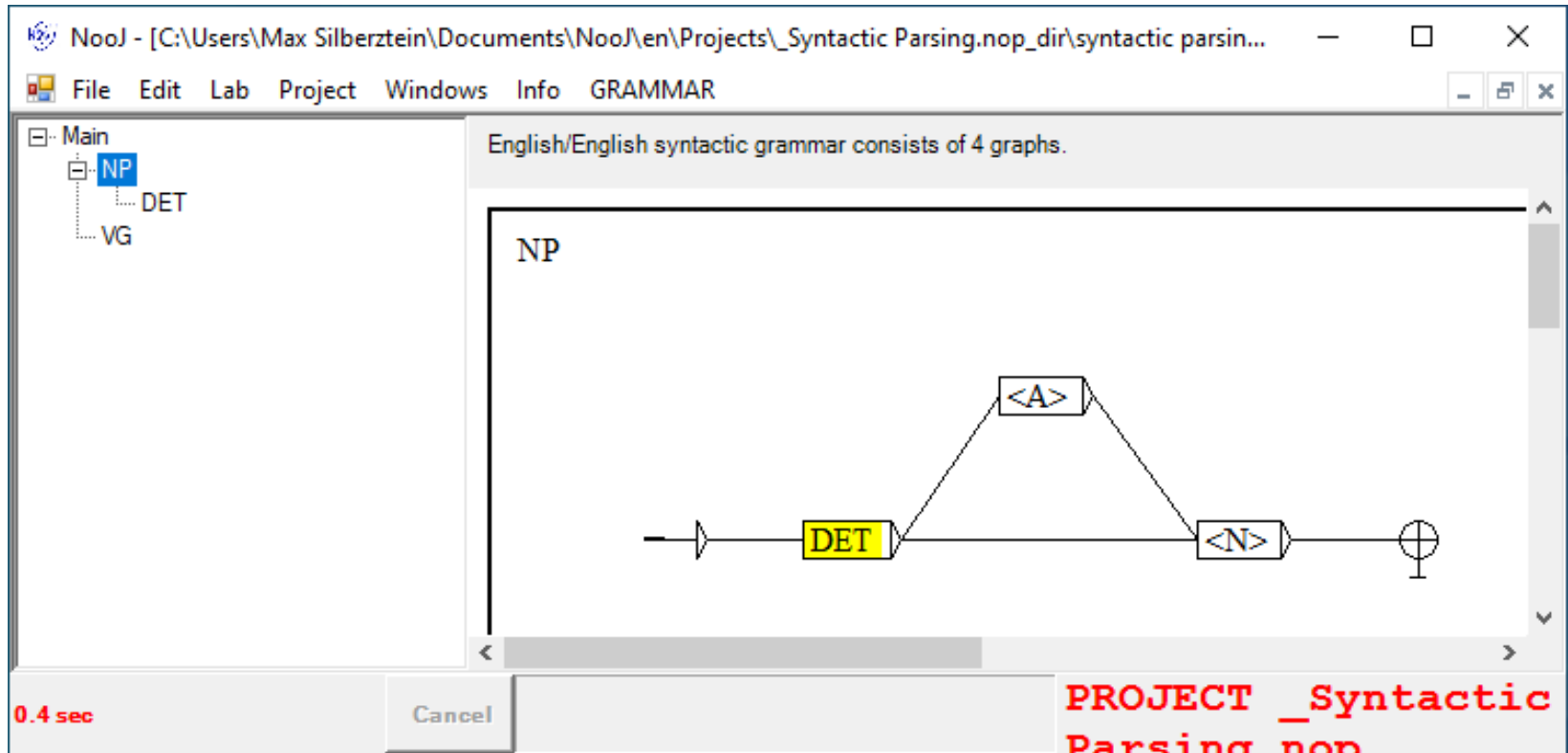
Syntactic grammar produces structural annotations



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

# Syntactic Analysis

Syntactic grammar produces structural annotations



- Several embedded grammars

# Syntactic Analysis

## Applying grammar to text

The screenshot shows the NooJ application window. The main text area contains the sentence "The man eats the apple." A dialog box titled "Locate a pattern in syntactic parsing" is open. In the dialog, the "Pattern is:" section has three radio buttons: "a string of characters:", "a PERL regular expression:", and "a NooJ regular expression:". Below these is a text field. The "a NooJ grammar:" section is selected, showing a dropdown menu with the path "C:\Users\Max Silberstein\Documents\NooJ\en\Projects\\_Syntactic Parsing\PROJECT\_Syntactic Parsing.nop" and a "Set" button. A red circle highlights the "Syntactic Analysis" checkbox, which is checked. The "Index" section has three radio buttons: "Shortest matches", "Longest matches" (selected), and "All matches". The "Limitation" section has three radio buttons: "All occurrences", "Only: 100 occ." (selected), and "1 occ. per match". At the bottom of the dialog, there is a "Reset Concordance" checkbox (checked) and a small icon with the letters N, o, o, J in colored boxes. The bottom status bar shows "0 sec" and a "Cancel" button. The file name "PROJECT\_Syntactic Parsing.nop" is displayed in red text at the bottom right.

NooJ

File Edit Lab Project Windows Info TEXT

C:\Users\Max Silberstein\Documents\NooJ\en\Projects\\_Syntactic Parsing\PROJECT\_Syntactic Parsing.nop

- 1 +

☐ Show Text Annotation

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 Silberstein\Documents\NooJ\en\Projects\\_Syntactic Parsing\PROJECT\_Syntactic Parsing.nop Set

☒ 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

0 sec

Cancel

PROJECT\_Syntactic Parsing.nop

# Syntactic Analysis

## Applying grammar to text

The screenshot displays the Nool software interface. A dialog box titled "Locate a pattern in syntactic parsing" is open, showing the following settings:

- Pattern is:
  - ☐ a string of characters:
  - ☐ a PERL regular expression:
  - ☐ a Nool regular expression:
  - ☒ a Nool grammar:
    - Path: C:\Users\Max Silberstein\Documents\Nool\en\Projects\\_Syr...
    - Set button
- ☒ Syntactic Analysis
- Index:
  - ☐ Shortest matches
  - ☒ Longest matches
  - ☐ All matches
- Limitation:
  - ☐ All occurrences
  - ☒ Only: 100 occ.
  - ☐ 1 occ. per match
- ☒ Reset Concordance

Below the dialog box, the main window shows the text "The man eats the apple." and a concordance window titled "Concordance for text: syntactic parsing.not". The concordance window has the following settings:

- Reset button
- Display: 5 characters before, and 5 after. Display: ☒ Matches ☒ Outputs
- Seq. button
- Output text: 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>#.#>#
- Query button
- 1/1

The output text is highlighted in red. At the bottom of the window, the status bar shows "0 sec", a "Cancel" button, and the file name "PROJECT \_Syntactic Parsing.nop".

The output represents the structure produced by the annotations

# Syntactic Analysis

CONCORDANCE > Display Syntactic Analysis

NooJ displays the parse tree:

The screenshot displays the NooJ software interface with the following components:

- Main Window:** Shows the sentence "The man eats the apple." and a "Syntactic Analysis" window.
- Syntactic Analysis Window:** Displays a parse tree for the sentence. The root node is "Main", which branches into three nodes: "NP", "VG", and "NP". The first "NP" branches into "DET" (the) and "N+s+Hum" (man). The "VG" node branches into "eat" (V+PR+3+s). The second "NP" branches into "DET" (the) and "N+s+Conc" (apple). A red circle highlights the "Concordance Entry / 1" label in the top right corner of this window.
- Locate a pattern in syntactic parsing window:** Shows options for pattern matching, including "a string of characters", "a PERL regular expression", "a NooJ regular expression", and "a NooJ grammar". The "a NooJ grammar" option is selected, and the path "C:\Users\Max Silberstein\Documents\NooJ\en\Projects\\_Syr..." is entered. The "Syntactic Analysis" checkbox is checked.
- Concordance for text: syntactic parsing.not window:** Shows a concordance table with columns for "Seq.", "Query", and "Display". The "Seq." column contains the sentence "The man eats the apple/<SENTENCE#<NP+Sub#<NP#<DET#<the,the,DET>#<man,man,N+s+<eat,V+PR+3+s>#<the,the,DET>#<apple,N+s+Conc>#>#>#". The "Query" column contains the text "The man eats the apple/<SENTENCE#<NP+Sub#<NP#<DET#<the,the,DET>#<man,man,N+s+<eat,V+PR+3+s>#<the,the,DET>#<apple,N+s+Conc>#>#>#". The "Display" column contains the text "The man eats the apple/<SENTENCE#<NP+Sub#<NP#<DET#<the,the,DET>#<man,man,N+s+<eat,V+PR+3+s>#<the,the,DET>#<apple,N+s+Conc>#>#>#".



# Syntactic Analysis

CONCORDANCE > Display Syntactic Analysis

The TAS is displayed as a tree:

The screenshot displays the Nool software interface. The main window shows a concordance entry for the sentence "The man eats the apple". The "Display" tab is set to "Structural Tree", which is circled in red. The tree structure is as follows:

```
graph TD
    SENTENCE[SENTENCE] --> NP+Sub[NP+Sub]
    SENTENCE --> VG[VG]
    SENTENCE --> NP+Obj[NP+Obj]
    NP+Sub --> the1[the]
    NP+Sub --> man[man]
    VG --> eats[eats]
    NP+Obj --> the2[the]
    NP+Obj --> apple[apple]
```

Each word is shown in a box with its part of speech and grammatical function below it: "the" (DET), "man" (N+s+Hum), "eats" (V+PR+3+s), "the" (DET), and "apple" (N+s+Conc).

In the bottom left, the "Concordance for text: syntactic parsing.not" window shows the sequence: "The man eats the apple/<SENTENCE#<NP+Sub#<NP#<DET#<the,the,DET>#<man,man,N+s+>#<eats,V+PR+3+s>#<the,the,DET>#<apple,apple,N+s+Conc>#>#>#".

The bottom status bar shows "0 sec", a "Cancel" button, and the file name "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*

The screenshot shows the NooJ software interface. The main window displays a text file named "\_La femme de trente ans.not" with 24 tokens and 926 TUs. The language is set to "French (fr)". A "Syntactic Analysis" dialog box is open, showing a concordance entry for "Concordance Entry / 231". The display is set to "Syntax Tree" and "Lexeme, Lemma, Linguistic Information" is unchecked. The phrase "les tambours battirent aux champs" is analyzed, showing a syntax tree structure:

```
graph TD
    PHRASE --> SUJET
    PHRASE --> VERBE
    PHRASE --> OBJET
    SUJET --> les1[les]
    SUJET --> tambours[tambours]
    VERBE --> battirent[battirent]
    OBJET --> a[à]
    OBJET --> les2[les]
    OBJET --> champs[champs]
```

The tree structure shows the phrase "les tambours battirent aux champs" branching into three main components: SUJET (Subject), VERBE (Verb), and OBJET (Object). The SUJET branches into "les" and "tambours". The VERBE branches into "battirent". The OBJET branches into "à", "les", and "champs".

At the bottom of the dialog box, there is a "Cancel" button.

# Syntactic Analysis

## Contracted and agglutinated forms

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

The screenshot shows the NooJ software interface. The main window displays a text file named "\_La femme de trente ans.not" with 24 tokens and 926 TUs. The language is set to "French (fr)". A "Syntactic Analysis" dialog box is open, showing a concordance entry for "les tambours battirent aux champs". The dialog box has a "Display" section with "Syntax Tree" selected and "Lexeme, Lemma, Linguistic Information" unchecked. The syntax tree is displayed below the text. The root node is "PHRASE", which branches into "SUJET", "VERBE", and "OBJET". "SUJET" branches into "les" and "tambours". "VERBE" branches into "battirent". "OBJET" branches into "à" and "les champs". A red arrow points from the text "aux" in the phrase to the "à" node in the syntax tree, which is circled in red. The "Cancel" button is visible at the bottom of the dialog box.

NooJ

File Edit Lab Project Windows Info

\_La femme de trente ans.not

- 24 + / 926 TUs

Characters Tokens

Language is "French (fr)".

Text Delimiter is: \n (NEWLINE)

[- 1 +] Concordance Entry / 231

Display: ☒ Syntax Tree ☐ Parse Tree

☐ Lexeme, Lemma, Linguistic Information

les tambours battirent aux champs

PHRASE

SUJET VERBE OBJET

les tambours battirent à les champs

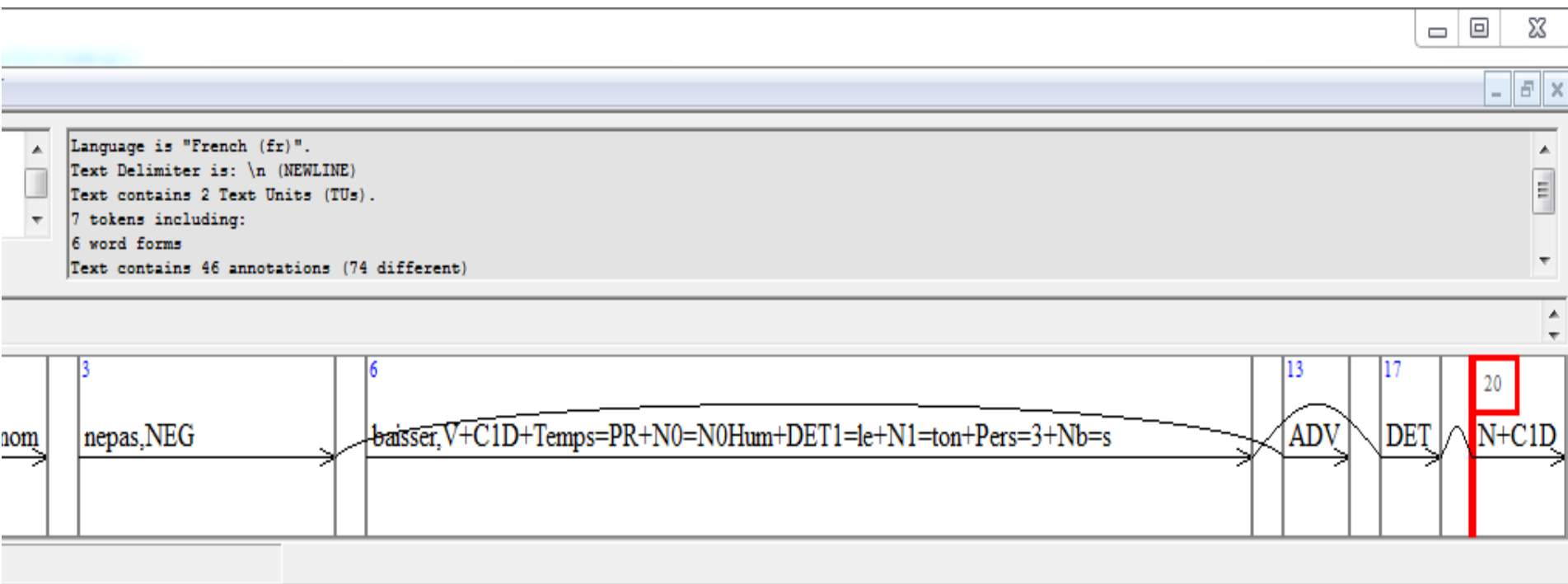
Cancel

# 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.

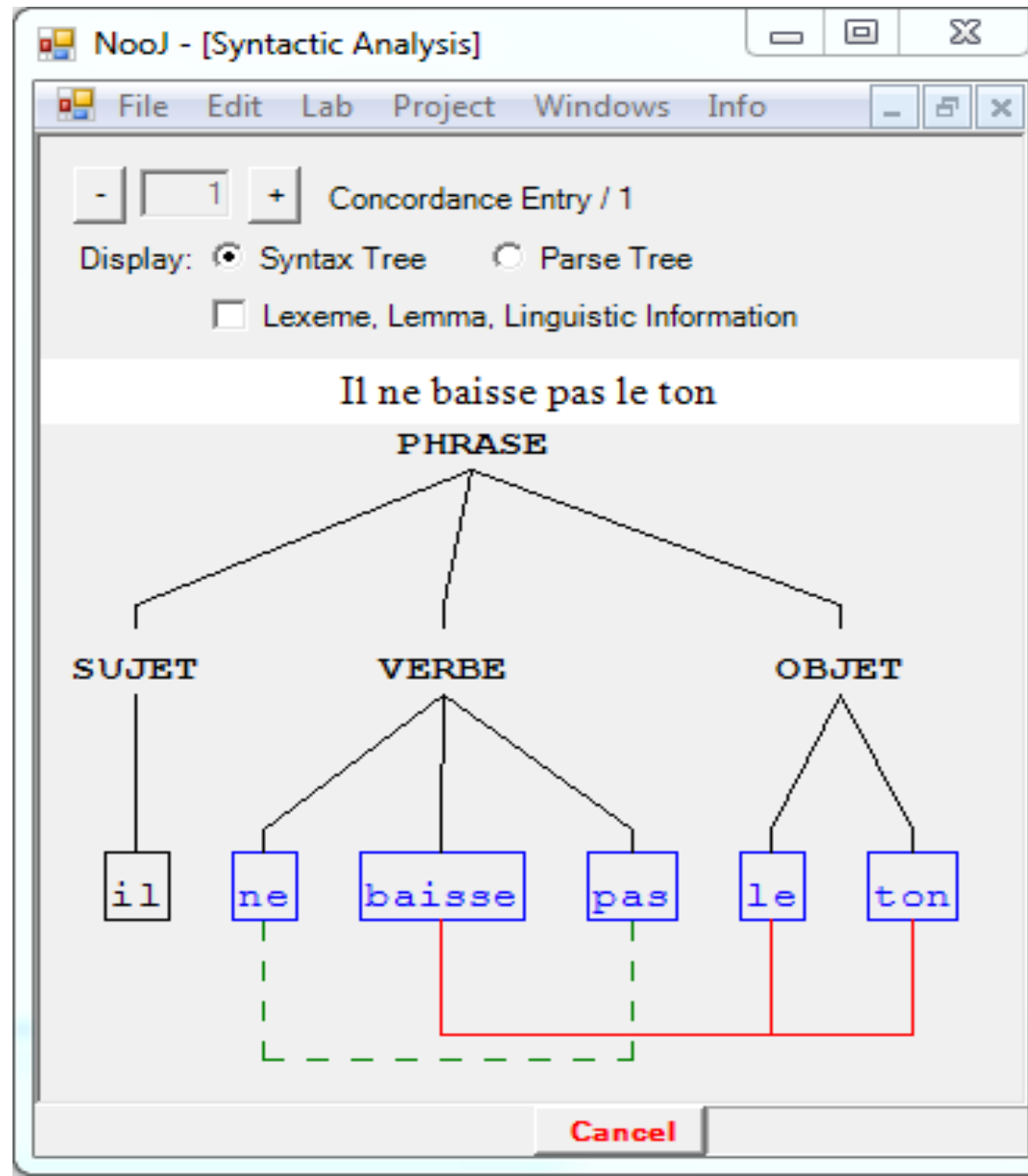
### The TAS:



# Syntactic Analysis

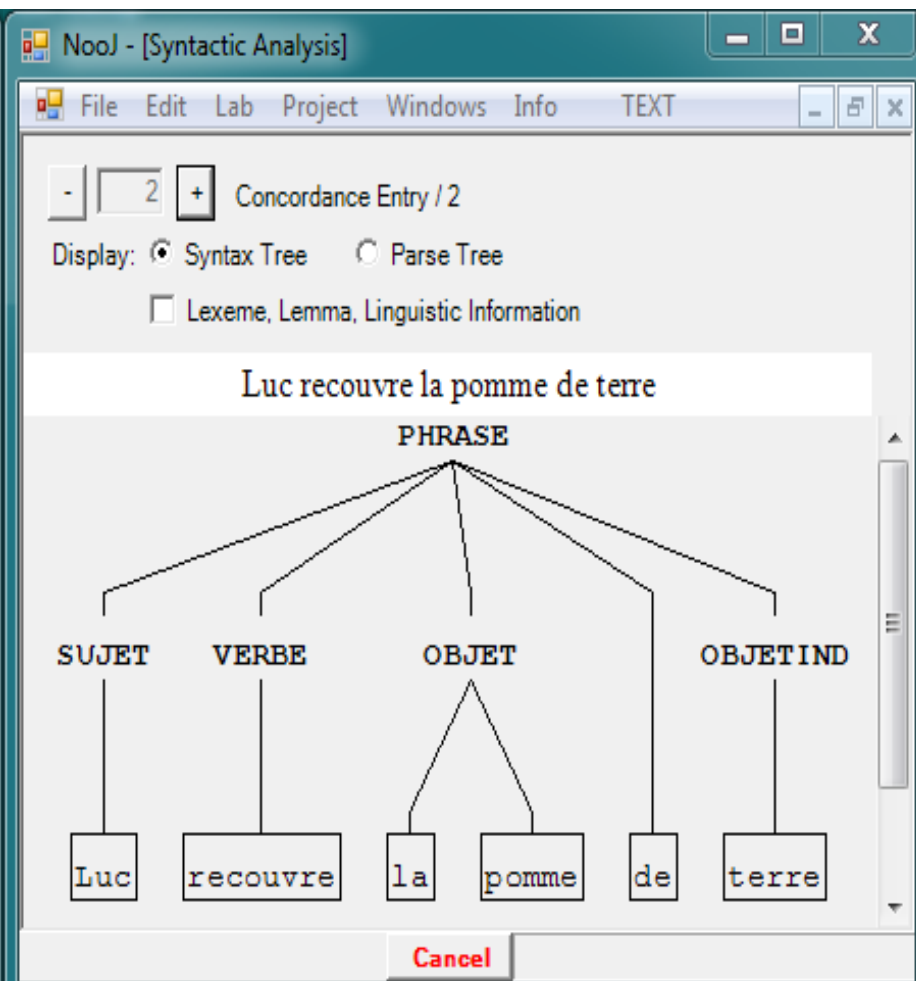
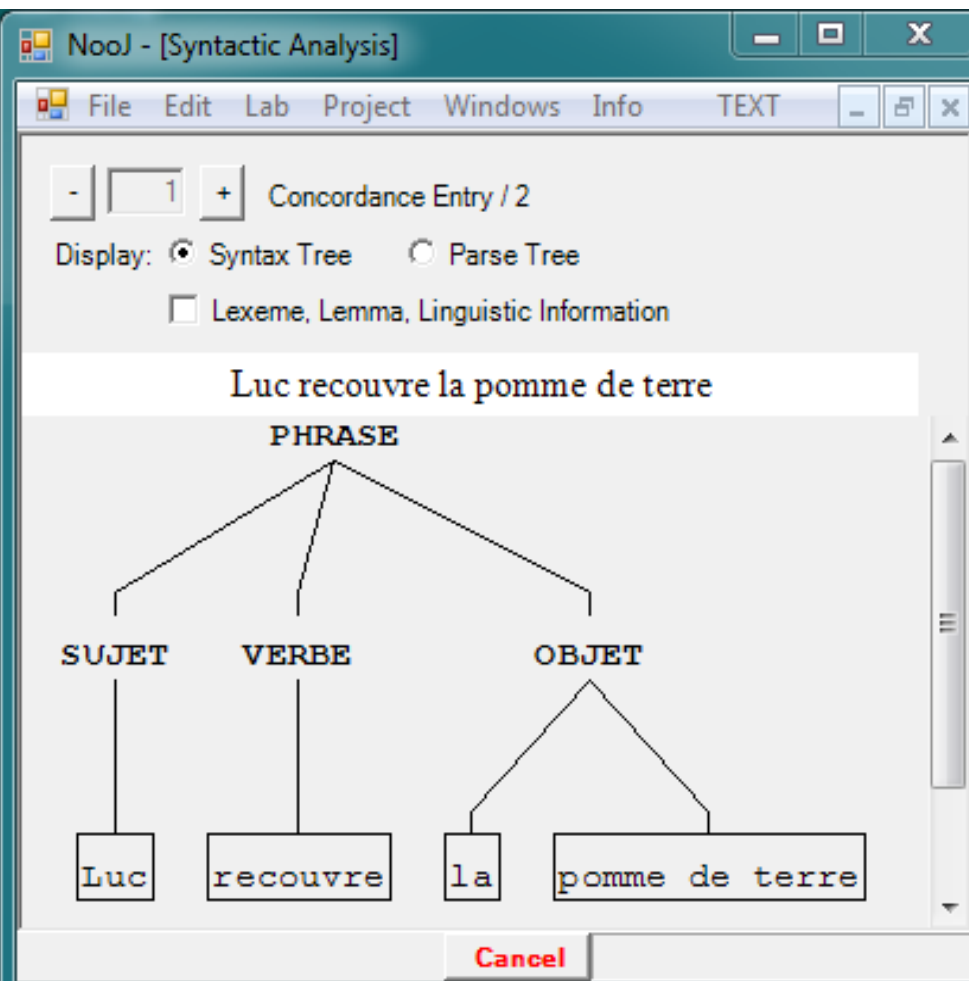
## Discontinuous expressions

- The negation *ne ... pas* is discontinuous.
- The frozen expression *baisser ... le ton* is discontinuous.



# Syntactic Analysis Ambiguities

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

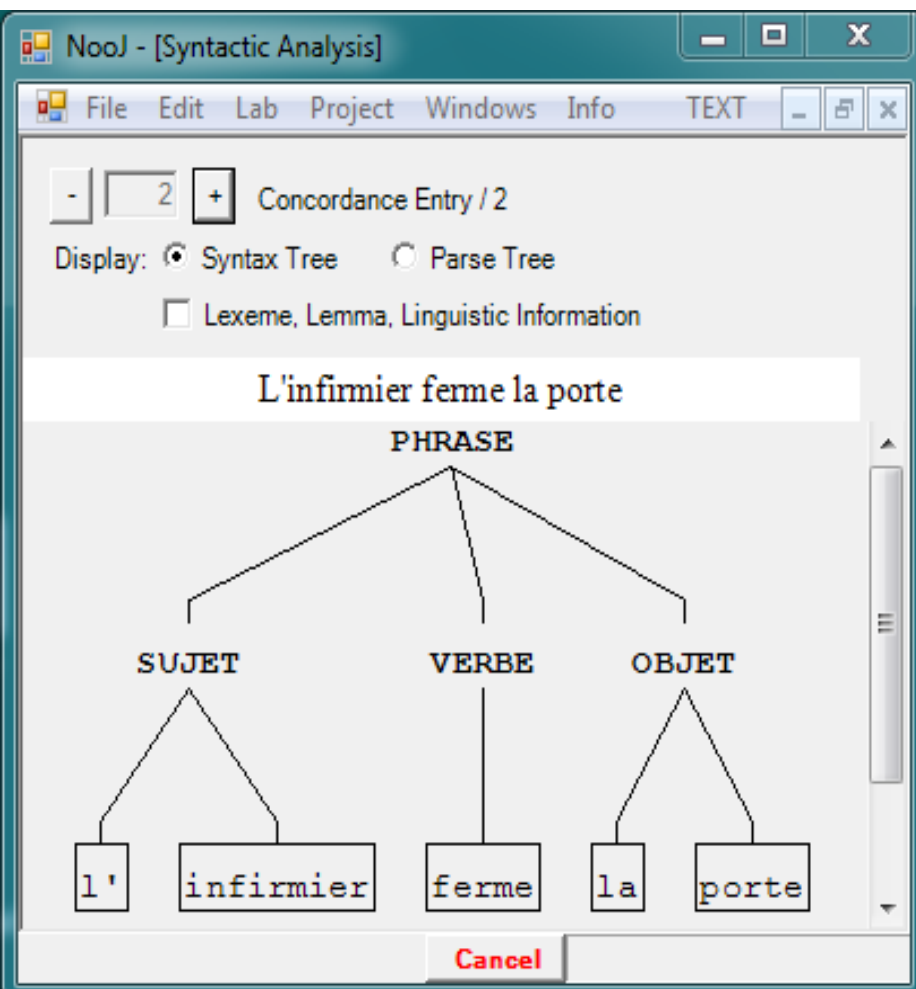
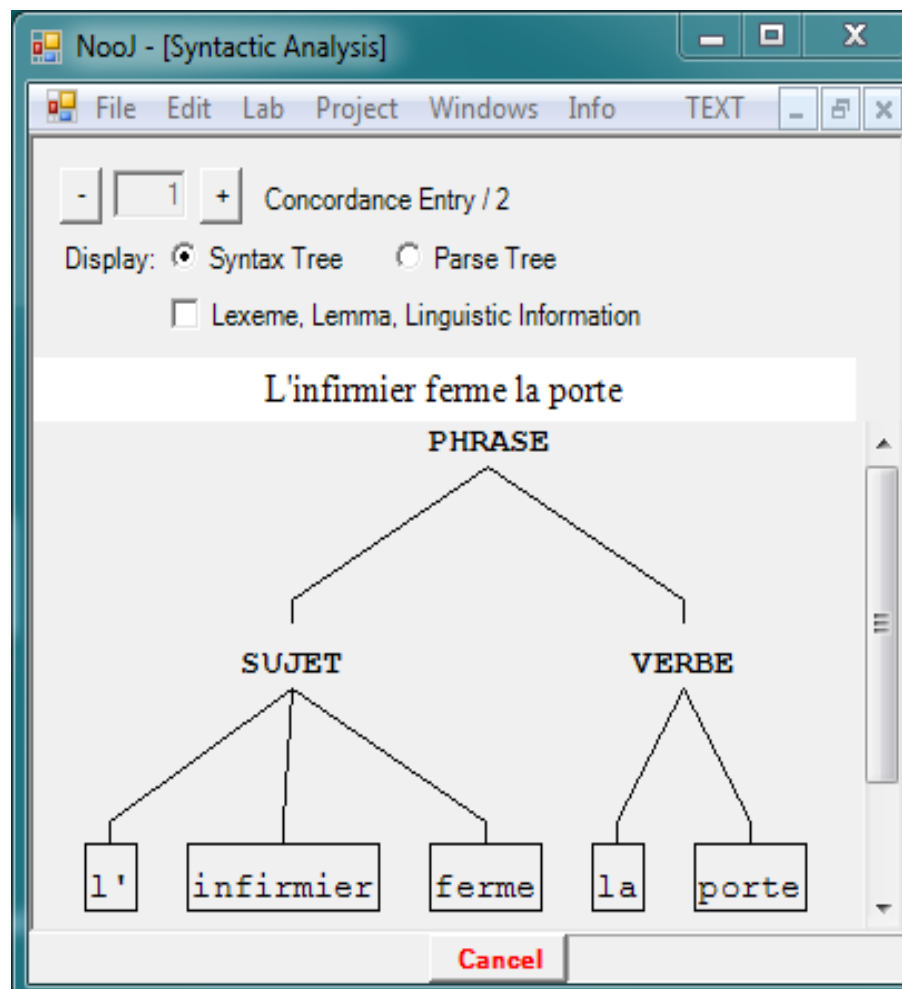




# Syntactic Analysis Ambiguities

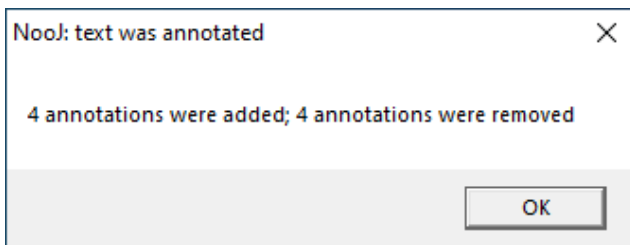
*L'infirmier ferme la porte*

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



# Syntactic Analysis Disambiguation

CONCORDANCE > Annotate Text (add/remove annotations)



NooJ - [C:\Users\Max Silberstein\Documents\NooJ\en\Projects\\_Syntacti...

File Edit Lab Project Windows Info TEXT

- 1 + / 2 TUs

Characters Tokens Digrams

☒ Show Text Annotation Structure

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

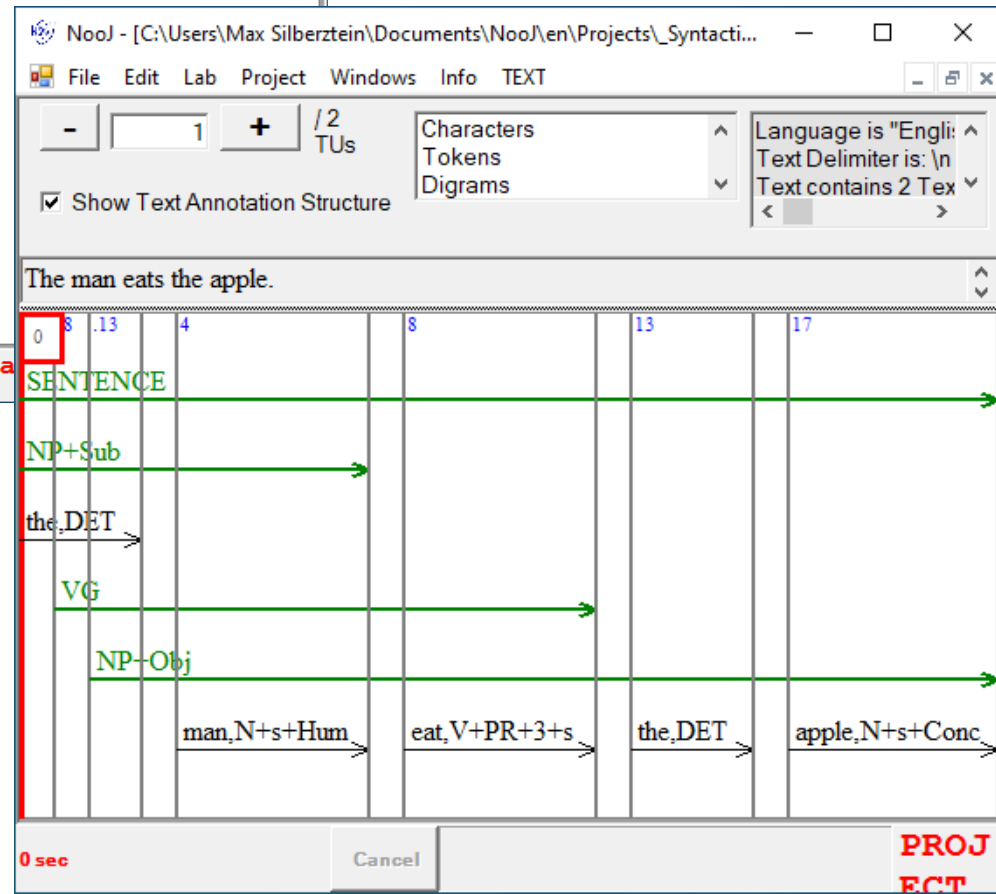
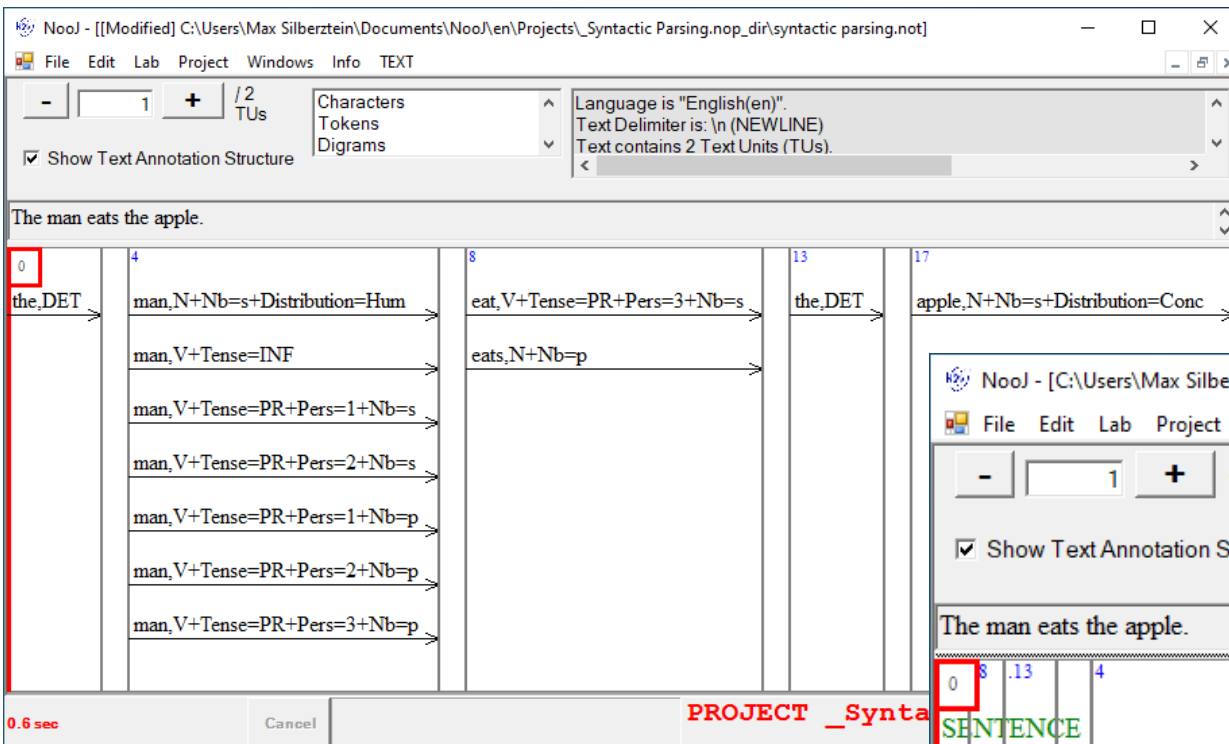
The man eats the apple.

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 Cancel PROJECT

# Syntactic Analysis

## Disambiguation: before and after





# CONGRATULATIONS



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

