Funded by
the European Union

10. Text Annotation Structure

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
- During the syntactic analysis, NooJ applies syntactic grammars to the text to add structural annotations, or remove ALU annotations (*e.g.*, when solving ambiguities).

Text Annotation Structure

NooJ lexical analyzer can annotate any ALU and represent them in the TAS:

- Agglutinated or contracted forms:

  ... *cannot* ... ⟶ **<can,V> <not,ADV>**
- Simple words:

  ... *is* ... ⟶ **<be,V+PR+3+s>**
- Multiword units:

  ... *Blue collars* ... ⟶ **<blue collar,N+Hum+p>**
- Discontinuous expressions:

  turns the light *off* ⟶ **<turn off,V+PR+3+s>**

Text Annotation Structure

NooJ uses linguistic resources to recognize ALUs:

- Morphological grammars:

cannot/<can,V><not,ADV>

... *cannot* ... ⟶ **<can,V> <not,ADV>**

- Dictionaries associated with inflectional/derivational grammars:

`be,V+FLX=BE+Aux`

`blue collar,N+Hum+FLX=TABLE`

... *is* ... ⟶ **<be,V+Aux+PR+3+s>**

... *Blue collars* ... ⟶ **<blue collar,N+Hum+p>**

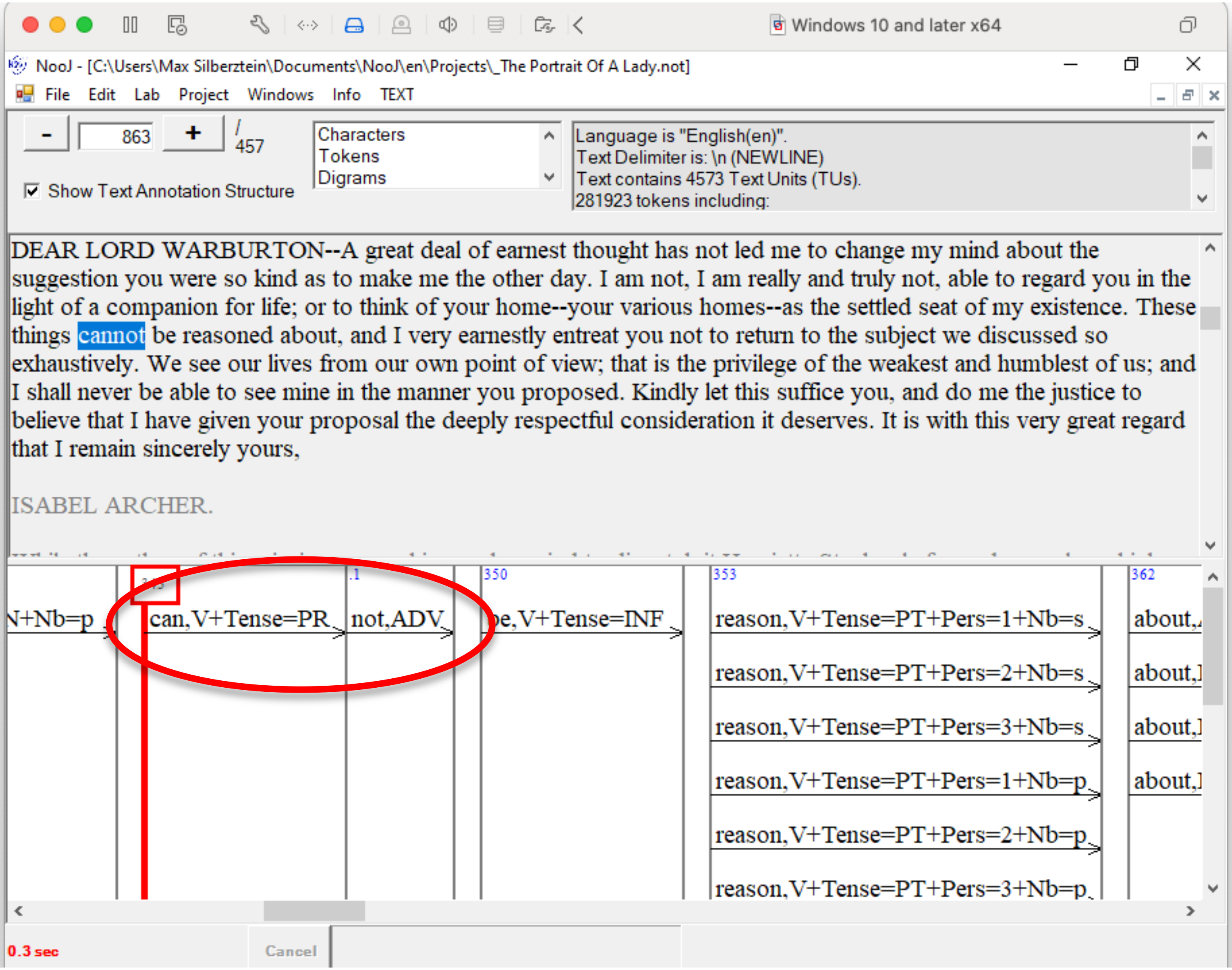
- Dictionary/grammar pairs (see later)

`turn,V+PV+FLX=HELP+Part="off”`

<V+PV> <WF>* $PV$Part

... *turns* the light *off* ... ⟶ **<turn off,V+PR+3+s>**

Annotation of the wordform *cannot*


Windows 10 and later x64
NooJ - [C:\Users\Max Silberztein\Documents\NooJ\en\Projects_The Portrait Of A Lady.not]
File
Edit
Lab
Project
Windows
Info
TEXT
-
863
+
/ 457
☑ Show Text Annotation Structure
Characters
Tokens
Digrams

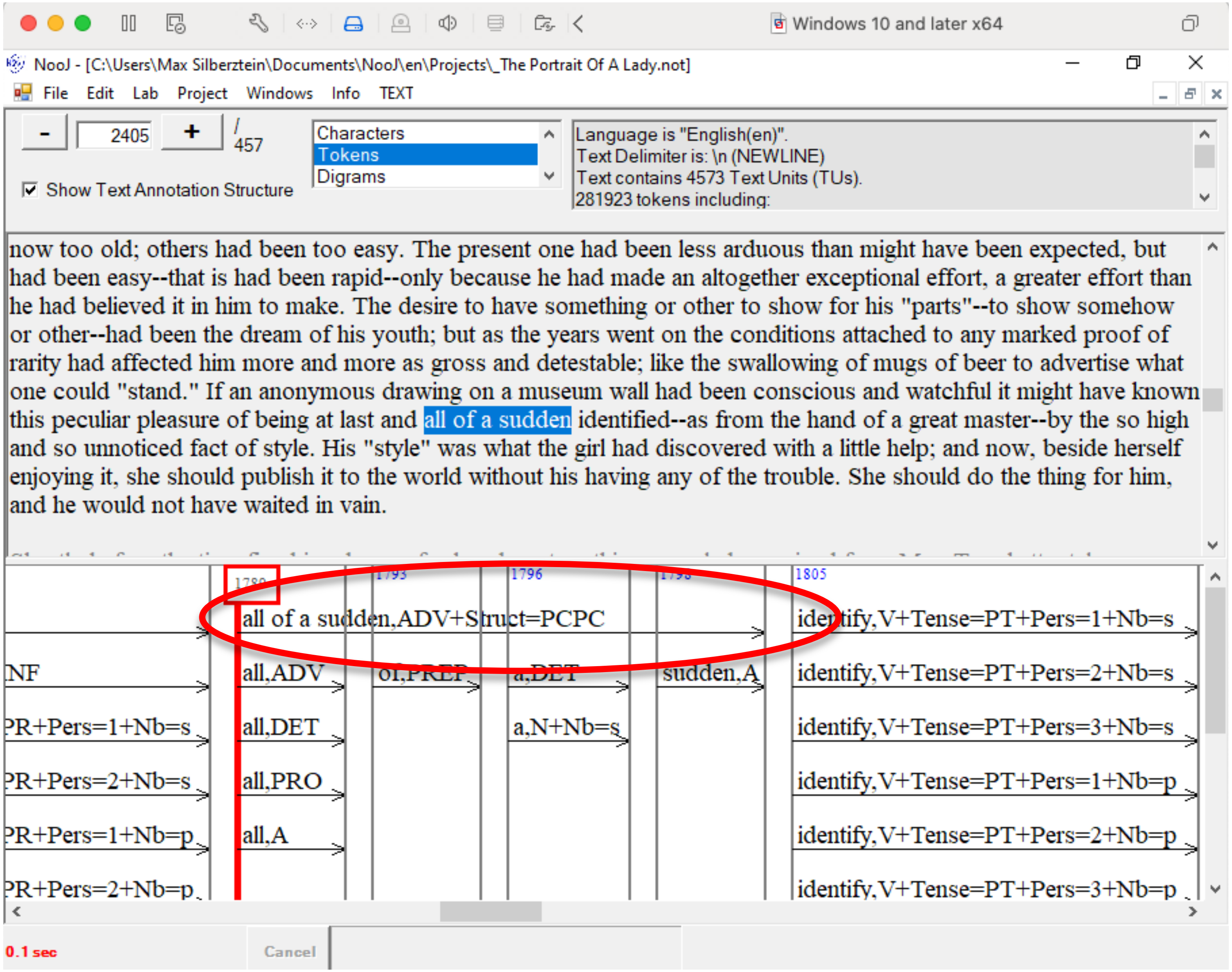
Language is "English(en)".
Text Delimiter is: \n (NEWLINE)
Text contains 4573 Text Units (TUs).
281923 tokens including:

DEAR LORD WARBURTON--A great deal of earnest thought has not led me to change my mind about the suggestion you were so kind as to make me the other day. I am not, I am really and truly not, able to regard you in the light of a companion for life; or to think of your home--your various homes--as the settled seat of my existence. These things cannot be reasoned about, and I very earnestly entreat you not to return to the subject we discussed so exhaustively. We see our lives from our own point of view; that is the privilege of the weakest and humblest of us; and I shall never be able to see mine in the manner you proposed. Kindly let this suffice you, and do me the justice to believe that I have given your proposal the deeply respectful consideration it deserves. It is with this very great regard that I remain sincerely yours,

ISABEL ARCHER.

N+Nb=p
can,V+Tense=PR
.1
not,ADV
350
be,V+Tense=INF
353
reason,V+Tense=PT+Pers=1+Nb=s
reason,V+Tense=PT+Pers=2+Nb=s
reason,V+Tense=PT+Pers=3+Nb=s
reason,V+Tense=PT+Pers=1+Nb=p
reason,V+Tense=PT+Pers=2+Nb=p
reason,V+Tense=PT+Pers=3+Nb=p
362
about,
about,
about,
about,
0.3 sec
Cancel

Annotation of the multiword unit *all of a sudden*

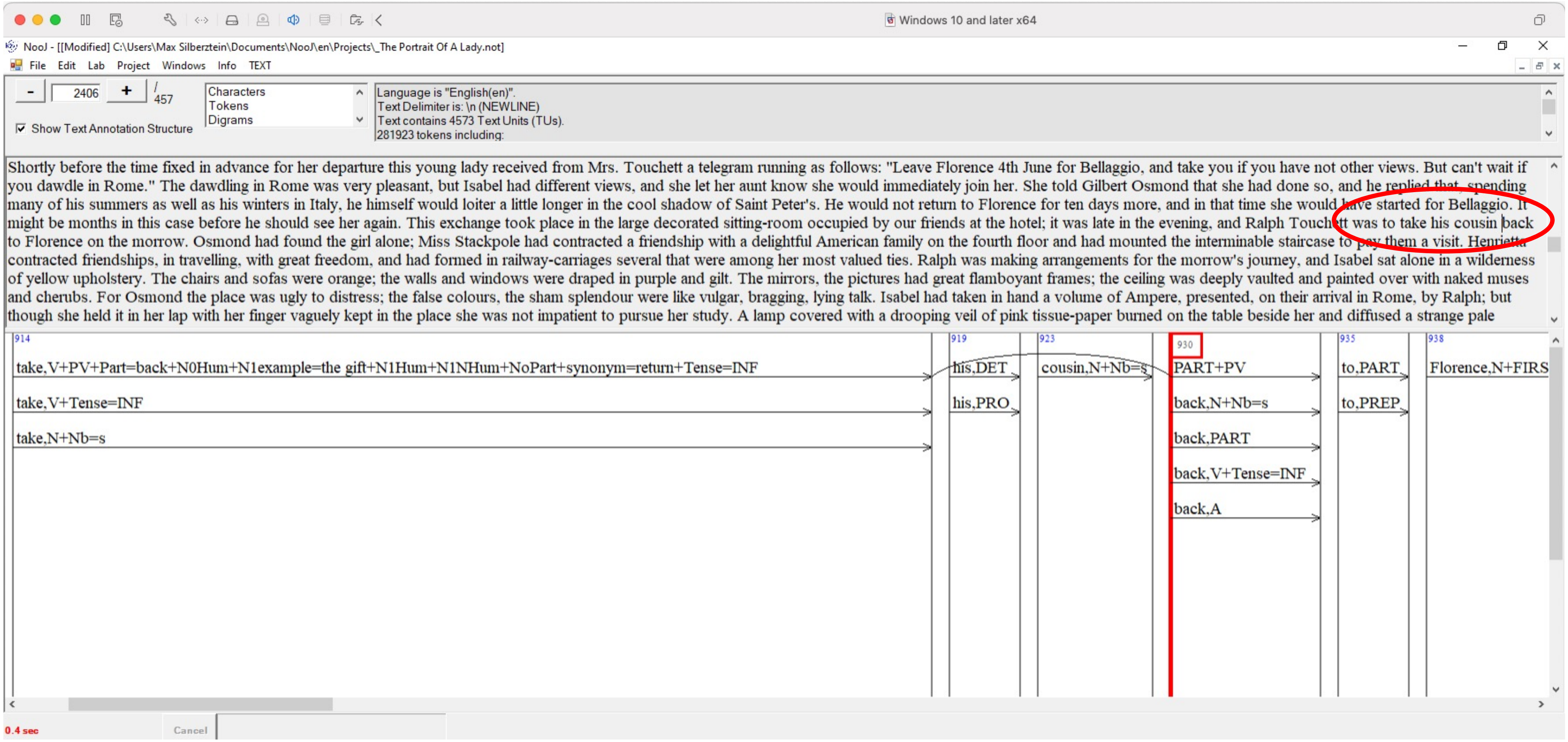

Windows 10 and later x64
NooJ - [C:\Users\Max Silberztein\Documents\NooJ\en\Projects_The Portrait Of A Lady.not]
File
Edit
Lab
Project
Windows
Info
TEXT
-
2405
+
/ 457
Characters
Tokens
Digrams
☑ Show Text Annotation Structure
Cancel
0.1 sec

Language is "English(en)".
Text Delimiter is: \n (NEWLINE)
Text contains 4573 Text Units (TUs).
281923 tokens including:

now too old; others had been too easy. The present one had been less arduous than might have been expected, but had been easy--that is had been rapid--only because he had made an altogether exceptional effort, a greater effort than he had believed it in him to make. The desire to have something or other to show for his "parts"--to show somehow or other--had been the dream of his youth; but as the years went on the conditions attached to any marked proof of rarity had affected him more and more as gross and detestable; like the swallowing of mugs of beer to advertise what one could "stand." If an anonymous drawing on a museum wall had been conscious and watchful it might have known this peculiar pleasure of being at last and all of a sudden identified--as from the hand of a great master--by the so high and so unnoticed fact of style. His "style" was what the girl had discovered with a little help; and now, beside herself enjoying it, she should publish it to the world without his having any of the trouble. She should do the thing for him, and he would not have waited in vain.

1789
1793
1796
1798
1805
all of a sudden,ADV+Struct=PCPC
NF
all,ADV
of,PREP
a,DET
sudden,A
PR+Pers=1+Nb=s
all,DET
a,N+Nb=s
PR+Pers=2+Nb=s
all,PRO
PR+Pers=1+Nb=p
all,A
PR+Pers=2+Nb=p
identify,V+Tense=PT+Pers=1+Nb=s
identify,V+Tense=PT+Pers=2+Nb=s
identify,V+Tense=PT+Pers=3+Nb=s
identify,V+Tense=PT+Pers=1+Nb=p
identify,V+Tense=PT+Pers=2+Nb=p
identify,V+Tense=PT+Pers=3+Nb=p

Annotation of the discontinuous expression *take … back*



NooJ - [[Modified] C:\Users\Max Silberztein\Documents\NooJ\en\Projects_The Portrait Of A Lady.not]

2406 / 457

☑ Show Text Annotation Structure

Language is "English(en)".
Text Delimiter is: \n (NEWLINE)
Text contains 4573 Text Units (TUs).
281923 tokens including:

Shortly before the time fixed in advance for her departure this young lady received from Mrs. Touchett a telegram running as follows: "Leave Florence 4th June for Bellaggio, and take you if you have not other views. But can't wait if you dawdle in Rome." The dawdling in Rome was very pleasant, but Isabel had different views, and she let her aunt know she would immediately join her. She told Gilbert Osmond that she had done so, and he replied that, spending many of his summers as well as his winters in Italy, he himself would loiter a little longer in the cool shadow of Saint Peter's. He would not return to Florence for ten days more, and in that time she would have started for Bellaggio. It might be months in this case before he should see her again. This exchange took place in the large decorated sitting-room occupied by our friends at the hotel; it was late in the evening, and Ralph Touchett was to take his cousin back to Florence on the morrow. Osmond had found the girl alone; Miss Stackpole had contracted a friendship with a delightful American family on the fourth floor and had mounted the interminable staircase to pay them a visit. Henrietta contracted friendships, in travelling, with great freedom, and had formed in railway-carriages several that were among her most valued ties. Ralph was making arrangements for the morrow's journey, and Isabel sat alone in a wilderness of yellow upholstery. The chairs and sofas were orange; the walls and windows were draped in purple and gilt. The mirrors, the pictures had great flamboyant frames; the ceiling was deeply vaulted and painted over with naked muses and cherubs. For Osmond the place was ugly to distress; the false colours, the sham splendour were like vulgar, bragging, lying talk. Isabel had taken in hand a volume of Ampere, presented, on their arrival in Rome, by Ralph; but though she held it in her lap with her finger vaguely kept in the place she was not impatient to pursue her study. A lamp covered with a drooping veil of pink tissue-paper burned on the table beside her and diffused a strange pale

914
take,V+PV+Part=back+N0Hum+N1example=the gift+N1Hum+N1NHum+NoPart+synonym=return+Tense=INF
take,V+Tense=INF
take,N+Nb=s

919
his,DET
his,PRO

923
cousin,N+Nb=s

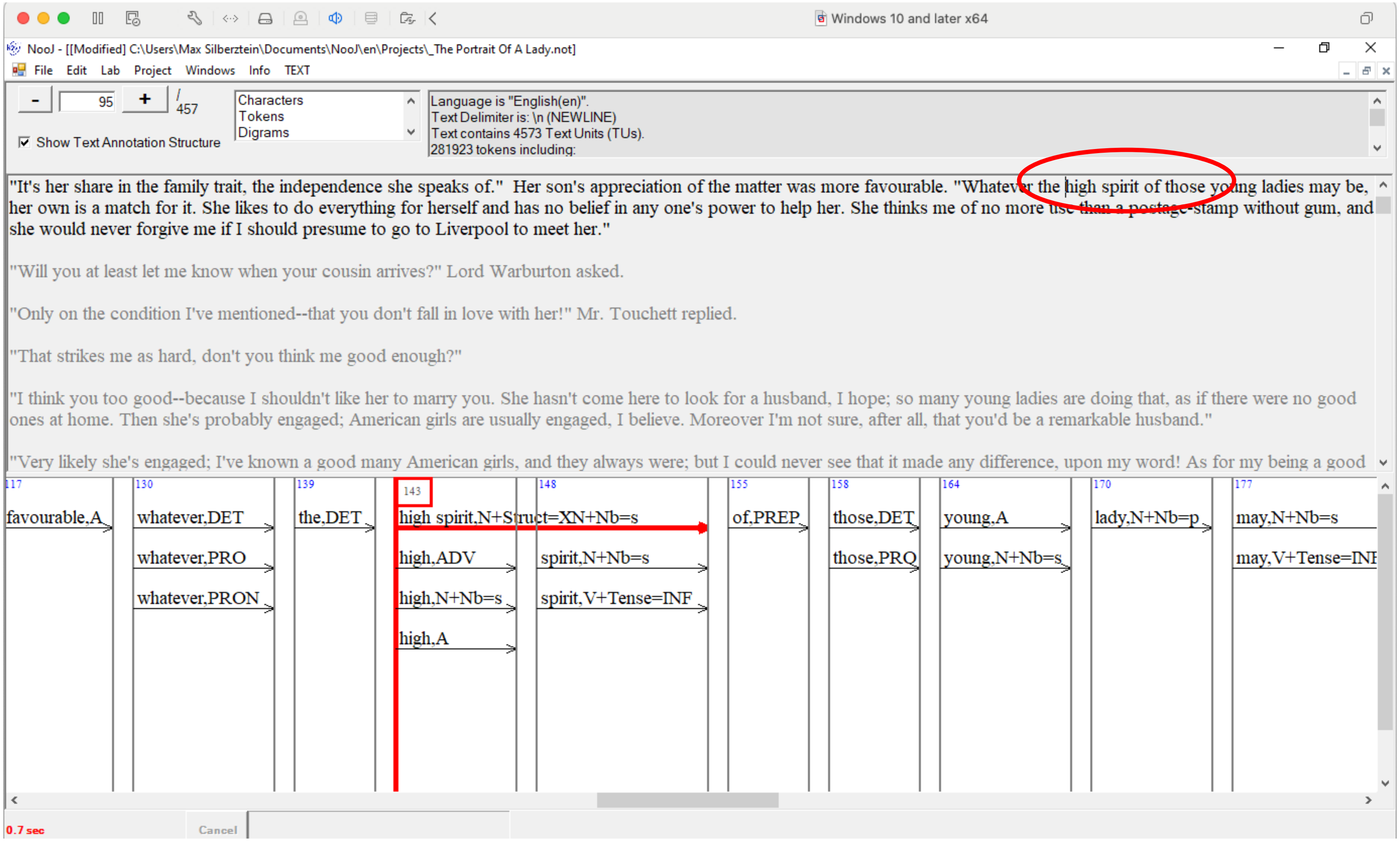
930
PART+PV
back,N+Nb=s
back,PART
back,V+Tense=INF
back,A

935
to,PART
to,PREP

938
Florence,N+FIRS

0.4 sec

Automatic Lexical Analysis
Representing lexical and morphological ambiguities


Windows 10 and later x64
NooJ - [[Modified] C:\Users\Max Silberztein\Documents\NooJ\en\Projects_The Portrait Of A Lady.not]
File
Edit
Lab
Project
Windows
Info
TEXT
-
95
+
/ 457
☑ Show Text Annotation Structure
Characters
Tokens
Digrams
0.7 sec
Cancel

Language is "English(en)".
Text Delimiter is: \n (NEWLINE)
Text contains 4573 Text Units (TUs).
281923 tokens including:

"It's her share in the family trait, the independence she speaks of." Her son's appreciation of the matter was more favourable. "Whatever the high spirit of those young ladies may be, her own is a match for it. She likes to do everything for herself and has no belief in any one's power to help her. She thinks me of no more use than a postage-stamp without gum, and she would never forgive me if I should presume to go to Liverpool to meet her."

"Will you at least let me know when your cousin arrives?" Lord Warburton asked.

"Only on the condition I've mentioned--that you don't fall in love with her!" Mr. Touchett replied.

"That strikes me as hard, don't you think me good enough?"

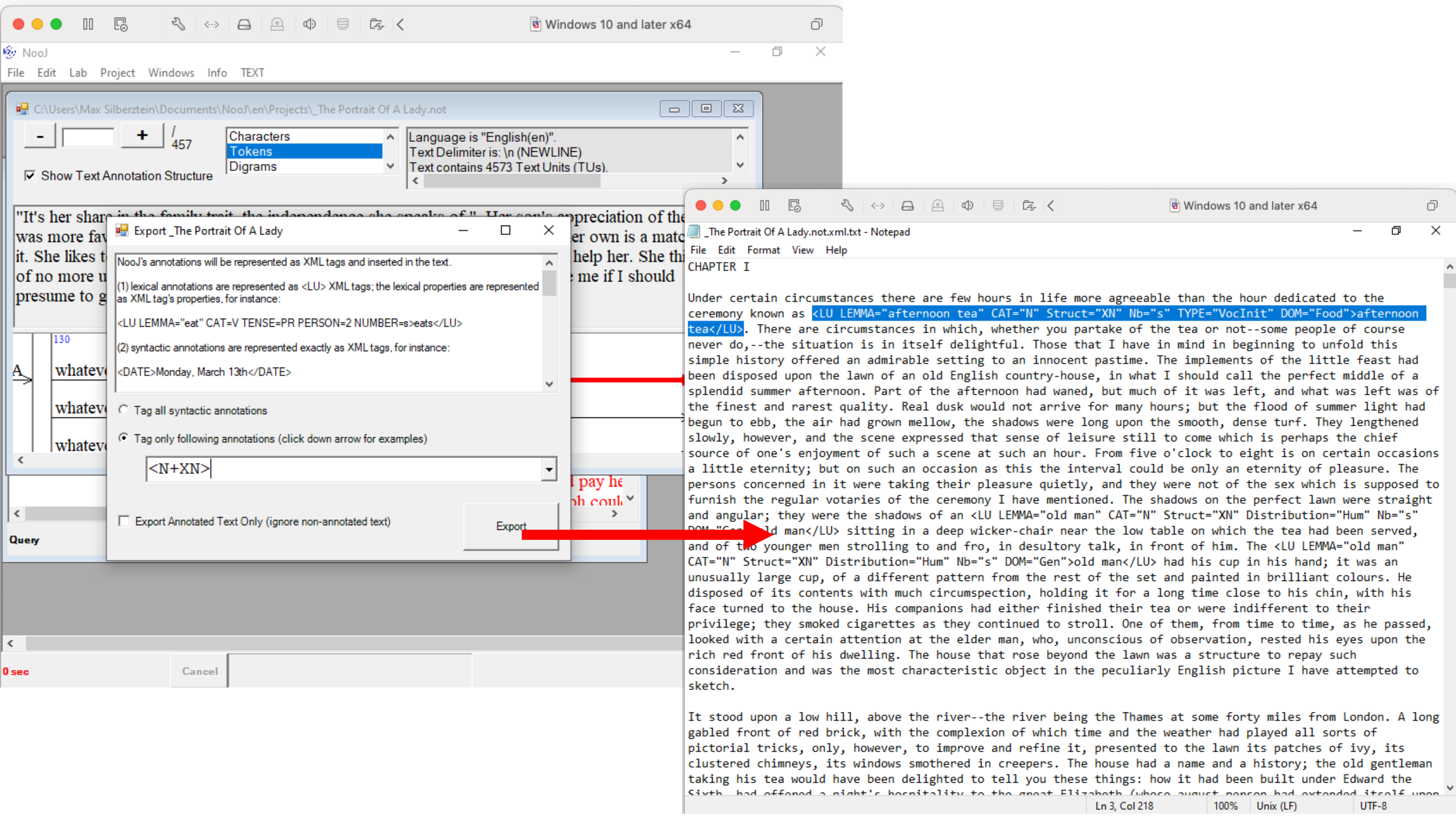
"I think you too good--because I shouldn't like her to marry you. She hasn't come here to look for a husband, I hope; so many young ladies are doing that, as if there were no good ones at home. Then she's probably engaged; American girls are usually engaged, I believe. Moreover I'm not sure, after all, that you'd be a remarkable husband."

"Very likely she's engaged; I've known a good many American girls, and they always were; but I could never see that it made any difference, upon my word! As for my being a good

117	130	139	143	148	155	158	164	170	177
favourable,A	whatever,DET	the,DET	high spirit,N+Struct=XN+Nb=s		of,PREP	those,DET	young,A	lady,N+Nb=p	may,N+Nb=s
	whatever,PRO		high,ADV	spirit,N+Nb=s		those,PRO	young,N+Nb=s		may,V+Tense=INF
	whatever,PRON		high,N+Nb=s	spirit,V+Tense=INF					
			high,A						

Text Annotation Structure

The TAS can be exported as an XML file


Windows 10 and later x64
NooJ
File Edit Lab Project Windows Info TEXT
C:\Users\Max Silberztein\Documents\NooJ\en\Projects_The Portrait Of A Lady.not
- + / 457
Characters
Tokens
Digrams
☑ Show Text Annotation Structure
Query
0 sec
Cancel
Export _The Portrait Of A Lady
Export
_The Portrait Of A Lady.not.xml.txt - Notepad
File Edit Format View Help
Ln 3, Col 218
100%
Unix (LF)
UTF-8

Language is "English(en)".
Text Delimiter is: \n (NEWLINE)
Text contains 4573 Text Units (TUs).

NooJ's annotations will be represented as XML tags and inserted in the text.

(1) lexical annotations are represented as <LU> XML tags; the lexical properties are represented as XML tag's properties, for instance:

<LU LEMMA="eat" CAT=V TENSE=PR PERSON=2 NUMBER=s>eats</LU>

(2) syntactic annotations are represented exactly as XML tags, for instance:

<DATE>Monday, March 13th</DATE>

○ Tag all syntactic annotations

● Tag only following annotations (click down arrow for examples)

<N+XN>

☐ Export Annotated Text Only (ignore non-annotated text)

```
CHAPTER I

Under certain circumstances there are few hours in life more agreeable than the hour dedicated to the
ceremony known as <LU LEMMA="afternoon tea" CAT="N" Struct="XN" Nb="s" TYPE="VocInit" DOM="Food">afternoon
tea</LU>. There are circumstances in which, whether you partake of the tea or not--some people of course
never do,--the situation is in itself delightful. Those that I have in mind in beginning to unfold this
simple history offered an admirable setting to an innocent pastime. The implements of the little feast had
been disposed upon the lawn of an old English country-house, in what I should call the perfect middle of a
splendid summer afternoon. Part of the afternoon had waned, but much of it was left, and what was left was of
the finest and rarest quality. Real dusk would not arrive for many hours; but the flood of summer light had
begun to ebb, the air had grown mellow, the shadows were long upon the smooth, dense turf. They lengthened
slowly, however, and the scene expressed that sense of leisure still to come which is perhaps the chief
source of one's enjoyment of such a scene at such an hour. From five o'clock to eight is on certain occasions
a little eternity; but on such an occasion as this the interval could be only an eternity of pleasure. The
persons concerned in it were taking their pleasure quietly, and they were not of the sex which is supposed to
furnish the regular votaries of the ceremony I have mentioned. The shadows on the perfect lawn were straight
and angular; they were the shadows of an <LU LEMMA="old man" CAT="N" Struct="XN" Distribution="Hum" Nb="s"
DOM="Ge[illegible]ld man</LU> sitting in a deep wicker-chair near the low table on which the tea had been served,
and of two younger men strolling to and fro, in desultory talk, in front of him. The <LU LEMMA="old man"
CAT="N" Struct="XN" Distribution="Hum" Nb="s" DOM="Gen">old man</LU> had his cup in his hand; it was an
unusually large cup, of a different pattern from the rest of the set and painted in brilliant colours. He
disposed of its contents with much circumspection, holding it for a long time close to his chin, with his
face turned to the house. His companions had either finished their tea or were indifferent to their
privilege; they smoked cigarettes as they continued to stroll. One of them, from time to time, as he passed,
looked with a certain attention at the elder man, who, unconscious of observation, rested his eyes upon the
rich red front of his dwelling. The house that rose beyond the lawn was a structure to repay such
consideration and was the most characteristic object in the peculiarly English picture I have attempted to
sketch.

It stood upon a low hill, above the river--the river being the Thames at some forty miles from London. A long
gabled front of red brick, with the complexion of which time and the weather had played all sorts of
pictorial tricks, only, however, to improve and refine it, presented to the lawn its patches of ivy, its
clustered chimneys, its windows smothered in creepers. The house had a name and a history; the old gentleman
taking his tea would have been delighted to tell you these things: how it had been built under Edward the
Sixth, had offered a night's hospitality to the great Elizabeth (whose august person had extended itself upon
```

Text Annotation Structure

The TAS can be exported as an XML file, but...

- Consider the two multiword units, represented in NooJ dictionary:

```
all of a sudden,ADV
sudden death syndrome,NOUN
```

... all of a sudden death syndrome ...

- When applying this dictionary to the above text, NooJ will represent the ambiguity in this TAS, as any of these two multiword units might be occurring.
- But XML cannot handle crossed-scoped annotations such as:

... <ADV>all of a <NOUN>sudden</ADV> death syndrome</NOUN> ...

- **Solution**: Either remove all ambiguities before exporting the TAS, or give priority to the first occurring multiword unit, e.g.,

... <ADV>all of a sudden</ADV> death syndrome ...

Text Annotation Structure

Removing ambiguities

- Removing ambiguities = deleting annotations from the TAS
- By exploring contexts. For example “can” is ambiguous, but not in these two contexts:

... They can (VERB) *open the can* (NOUN) *of beer ...*

- We will use syntactic or semantic grammars that describe contexts in which some ambiguities can be solved.

- However, in general, many ambiguities will not be solved, e.g.,

... There is a round table in room A32 ...

(a meeting, or a round piece of furniture?)

CONGRATULATIONS

You know how to automatically annotate a text by applying dictionaries and morphological grammars to the text.