NooJ's Text Annotation Structure

An Alternative Approach to Tagging

Workshop on Language Resources for Teaching and Research

Max Silberztein
Université de Franche-Comté

www.nooj4nlp.net

NooJ

- a tool used to formalize natural languages at the orthographical, morphological, syntactic and semantic levels.
- a corpus processing tool used to apply complex query to large texts; results are displayed as concordances, indices, statistical results, etc.
- a development environment used to build various NLP applications: search engine, question answering, automatic abstracts, competitive intelligence, terminology extractors, semi-automatic translation,

NooJ modules

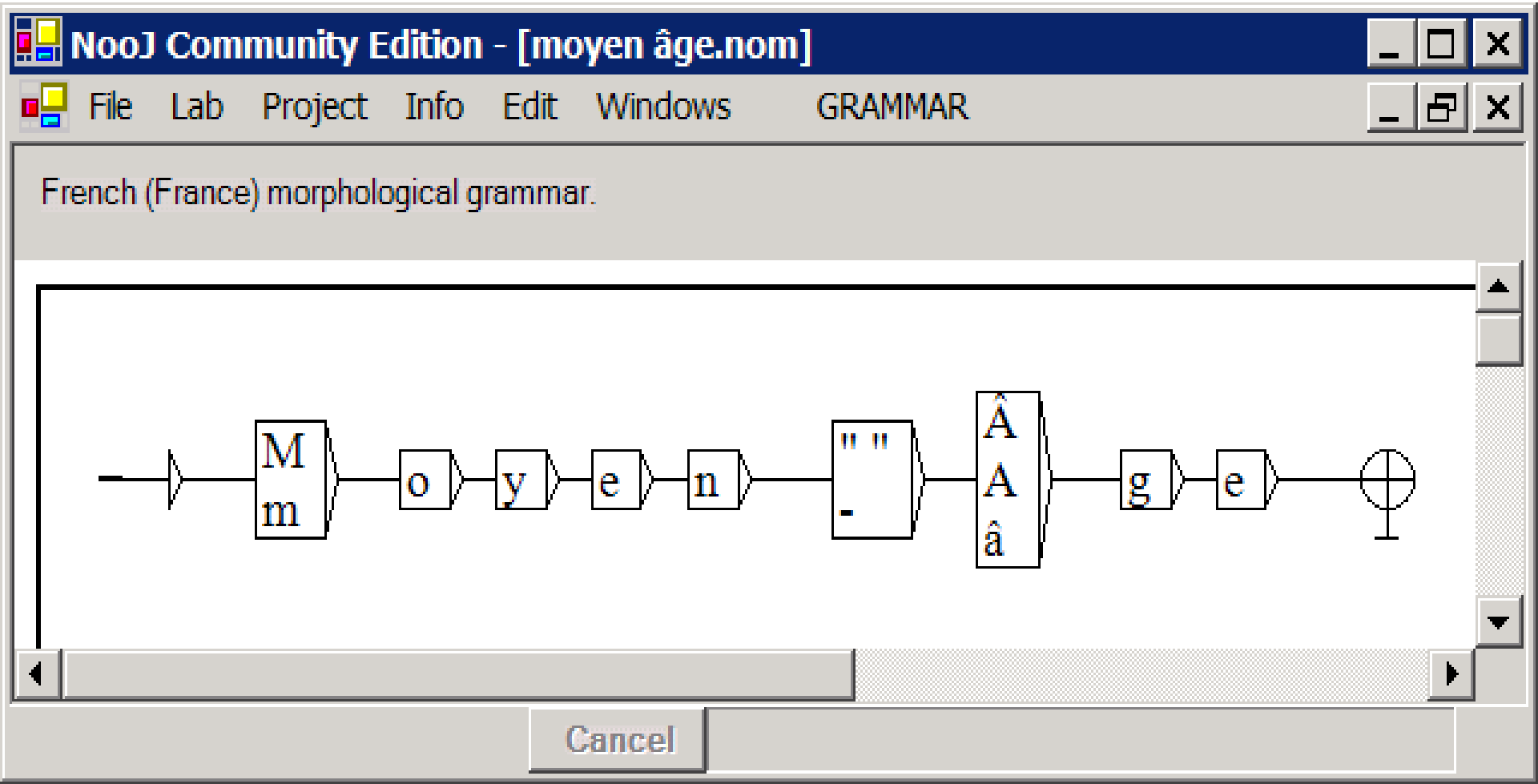
- Acadian (Univ. de Moncton) Arabic (UFC), Armenian (INALCO), Bulgarian (BACL), Chinese (UFC), English (UFC), French (UFC), Hebrew (Neguev Univ.), Hungarian (Academy of Sciences, Budapest), Italian (Univ. of Salerne), Portuguese (Linguateca), Serbo-Croatian (Univ. of Belgrade), Spanish (UAB); more modules are being built (Greek, Korean, Latin, Russian, Vietnamese)
- French Technical Medical Terminology (Univ. de Rouen), English Epidemiology Expressions (CHU de Marseille) Proper Names (Univ. de Tours)
- Various private companies have developed specialized modules for NooJ (spam, real estate, medecine description, civil engineering, etc.)

Computational Devices

- Finite-State Automata (Regular Expressions)
- Finite-State Transducers (Regular Expressions with outputs)
- Recursive graphs (Context-Free Grammars)
- Recursive Transition Networks
- Enhanced RTNs with variables (Turing Machines)

A Finite-State Automaton has : (1) an alphabet, (2) a set of states, including at least one initial state and at least one terminal state, (3) transitions (state, letter, state)

A NooJ Graph has : (1) an alphabet or a vocabulary, (2) labeled nodes, including one initial node and one terminal node, (3) connections (node, node).



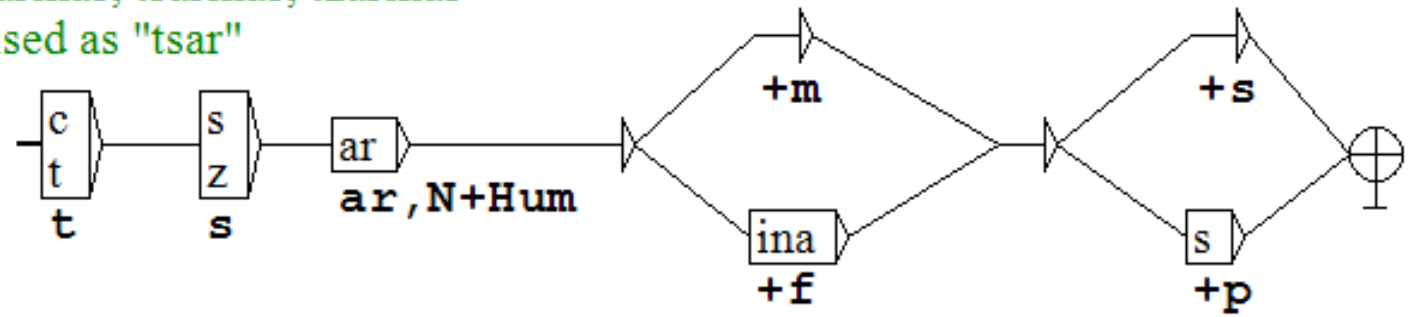
Labels may produce an output

NooJ Community Edition - [_Tsar.nom]

File Lab Project Info Edit Windows GRAMMAR

English (United States) morphological grammar.

The 16 forms:
csar, czar, tsar, tzar,
csarina, czarina, tsarina, tzarina
csars, czars, tsars, tzars,
csarinas, czarinas, tsarinas, tzarinas
are lemmatised as "tsar"

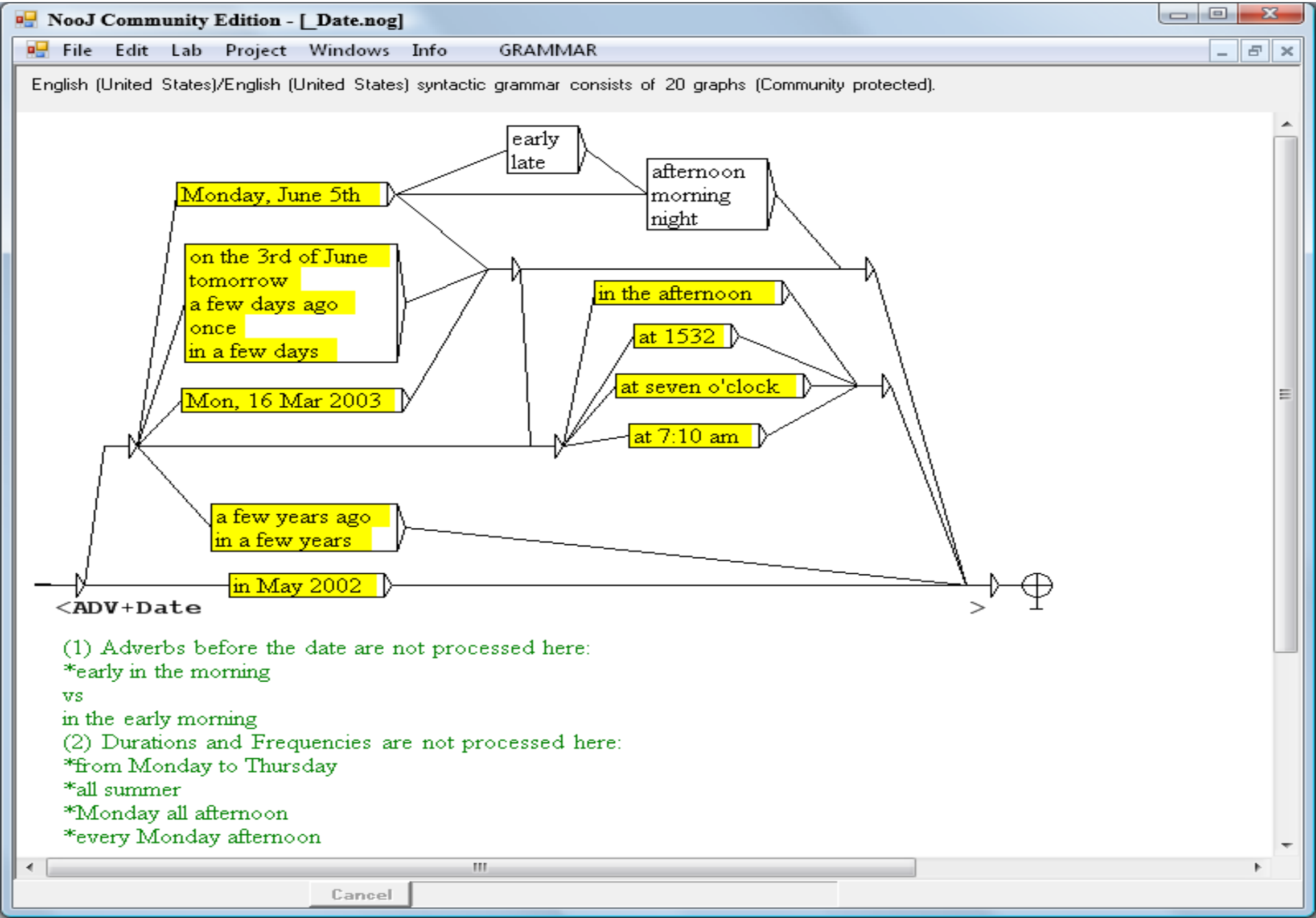


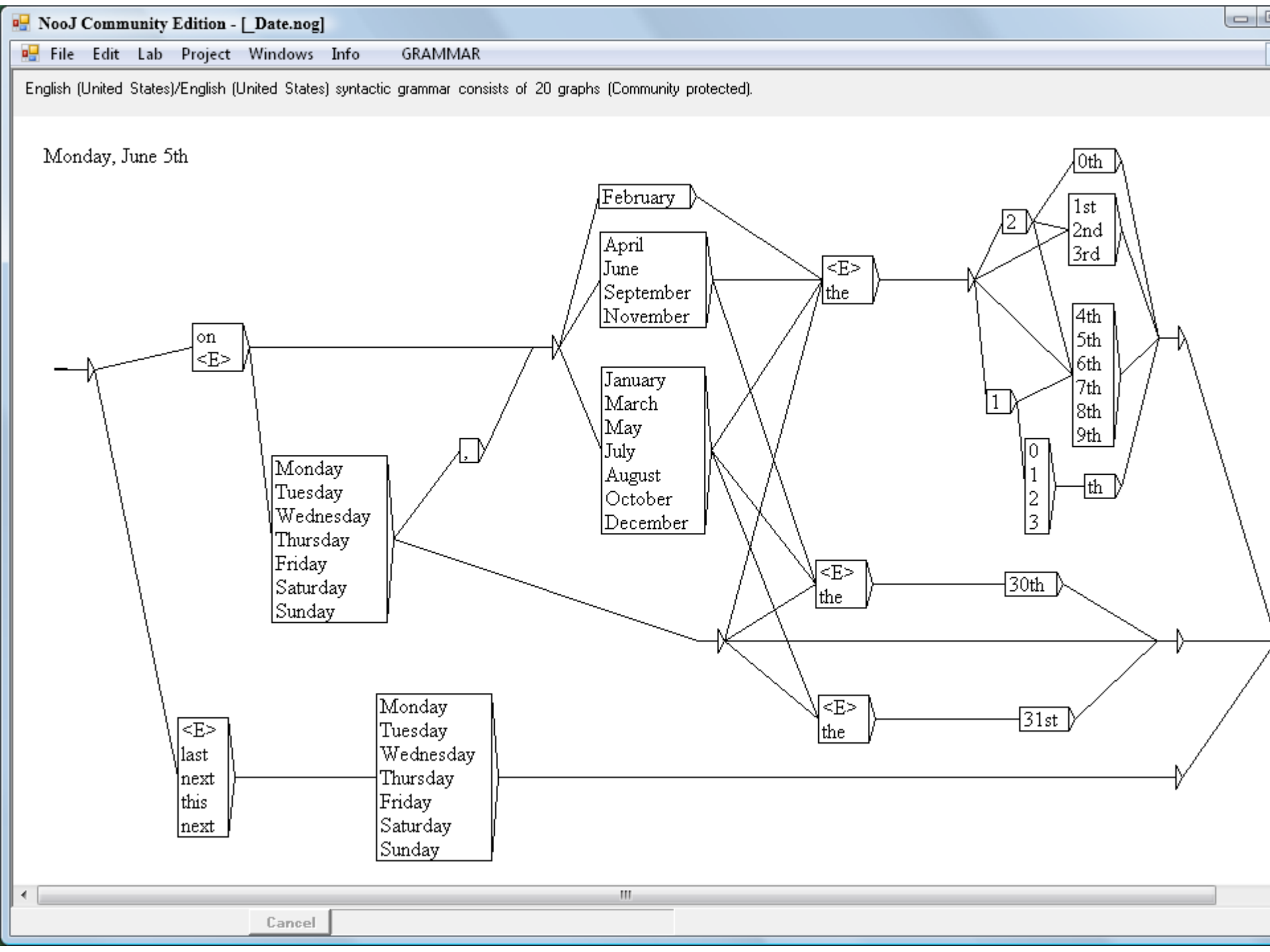
+Hum : Human Noun
+m : masculine
+s : singular
+p : plural

It is best to give a higher priority
to this grammar than to the general dictionary,
so that the 16 forms will be properly analyzed

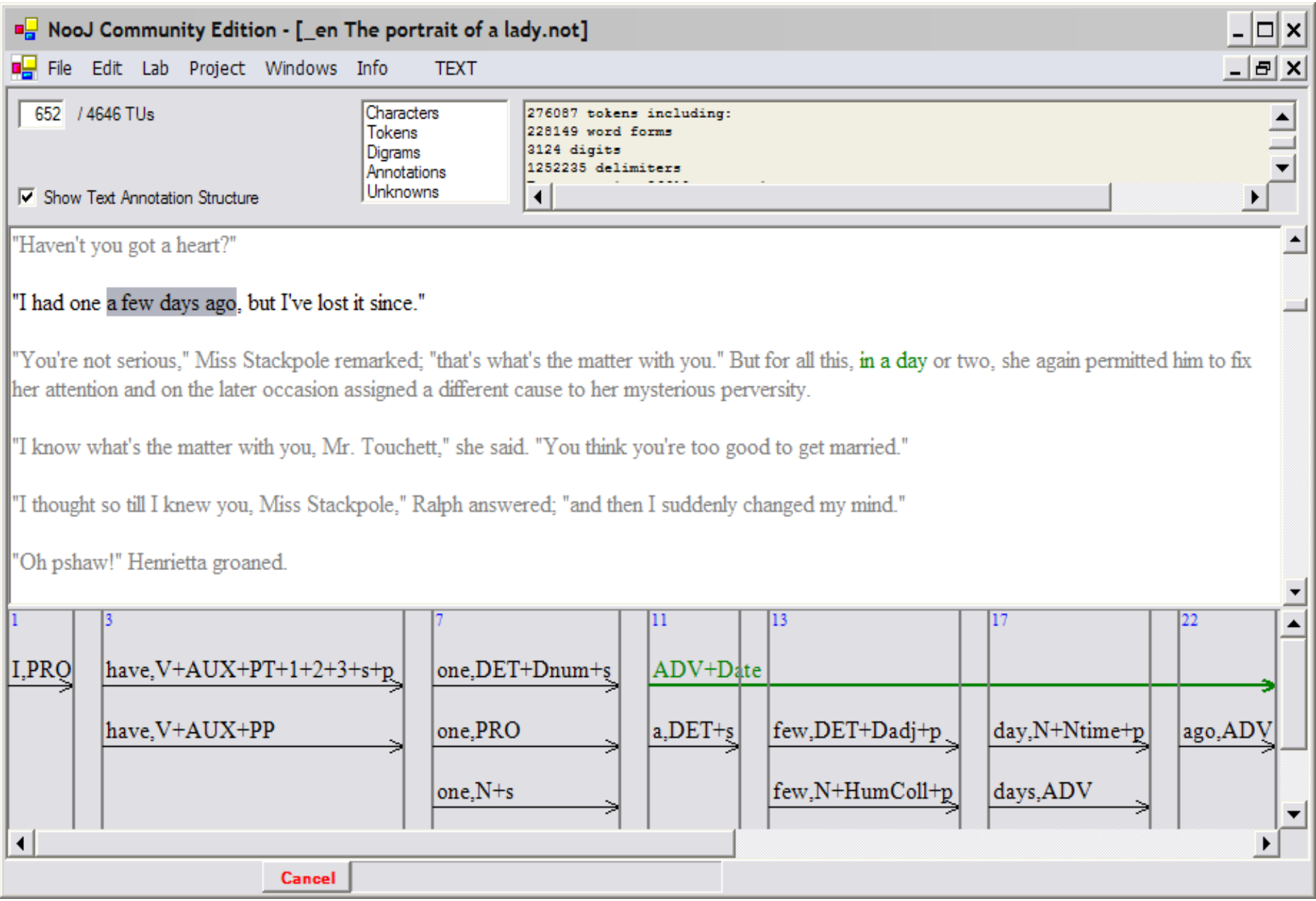
Cancel

NooJ grammars are structured sets of embedded graphs




NooJ Community Edition - [_Date.nog]
File Edit Lab Project Windows Info GRAMMAR
English (United States)/English (United States) syntactic grammar consists of 20 graphs (Community protected).
Monday, June 5th
on
<E>
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
,
February
April
June
September
November
January
March
May
July
August
October
December
<E>
the
2
1
0th
1st
2nd
3rd
4th
5th
6th
7th
8th
9th
0
1
2
3
th
<E>
the
30th
<E>
the
31st
<E>
last
next
this
next
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Cancel

Annotating large texts



Need for annotations words ≠ linguistic units

- Wedon’tpronounceblanksbetweenwords

- spaces are not always reliable:

  cannot, don’t, give in

  parce que, *afin* (vs. *à fin*), *au jour d’hui*

- Semitic languages (e.g. *andinthehouse*), Germanic languages (e.g. Unternehmensangebote), Chinese (no space)

A journalistic text, tagged

Battle/N tested/PP Japanese/A industrial/A managers/N here/ADV always/ADV buck/V up/PART nervous/A newcomers/N with/PREP the/DET tale/N of/PREP the/DET first/A of/PREP their/DET countrymen/N to/PART visit/V Mexico/N, a/DET boatload/N of/PREP samurai/N warriors/N blown/PP ashore/ADV 375/NB years/N ago/PART. From/PREP the/DET beginning/N, it/PRO took/V a/DET man/N with/PREP extraordinary/A qualities/N to/PART succeed/V in/PREP Mexico/LOC, says/V Kimihide/NPR Takimura/NPR, president/N of/PREP Mitsui/NPR group/N ’s/PART Kensetsu/N Engineering/VG

Another journalistic text, translated

"Bonne année". L'an 1994 aura commencé samedi à zéro heure, qui en douterait ? Et pourtant, cette évidence est pour le moins trompeuse. C'est ainsi que, décalage horaire aidant, les Australiens de Sydney fêtent la Saint-Sylvestre avec neuf heures d'avance sur nous et les Américains de Los Angeles avec neuf heures de retard. Un détail ? Admettons. Mais que dire, alors, des musulmans pratiquants qui, eux, se considèrent à la mi-1414, et célébreront leur "nouvel an" (l'hégire) le 10 juin ? Quant aux juifs, leur année 5754 a débuté le 16 septembre. Incontestablement, le temps est une notion toute relative. Les physiciens savent cela depuis longtemps. Les politiques et les religieux aussi. Mais, grands amateurs de certitudes, ces derniers se sont alliés aux scientifiques pour tenter d'offrir aux hommes un guide chronologique susceptible de régler une vie sociale qu'ils voulaient aussi harmonieuse que possible. Ce n'est pas un hasard si le calendrier grégorien que nous utilisons

amikai.com, freetranslation.com, officeupdate.lhsl.com, reverso.net, systransoft.com, t-mail.com, tranexp.com, translationwave.com:

" Good year ". Will the year 1994 have begun Saturday at zero hour, which would doubt it? And yet, this obviousness is at the very least misleading. Thus, time shift helping, the Australian ones of Sydney celebrate New Year's Eve with nine hours in advance on us and the Americans of Los Angeles with nine hours of delay. A detail? Let us admit [it]. But what to say, then, of the Moslems practise who, them, are considered with semi-1414, and will celebrate their " new year " (the hégire) on June 10? As for the Jews, their year 5754 began on September 16. Incontestably, time is a quite relative concept. The physicists [have known] that for a long time. Policies and monks too. But, large amateurs of certainty, the latter were combined to the scientists to try to offer to the men a chronological guide suitable for regulate a social life which they wanted possible as harmonious as. It is not a chance if the Gregorian

amikai.com, freetranslation.com, officeupdate.lhsl.com, reverso.net, systransoft.com, t-mail.com, tranexp.com, translationwave.com:

" ~~Good year~~ ". ~~Will~~ ~~the year 1994~~ have begun Saturday ~~at zero hour~~, ~~which would doubt it~~? And yet, this obviousness is at the very least misleading. Thus, ~~time shift~~ helping, the Australian ~~ones~~ of Sydney celebrate New Year's Eve ~~with nine hours in advance on~~ us and the Americans of Los Angeles ~~with nine hours of delay~~. A detail? Let us admit [it]. But what to say, then, of the ~~Moslems practise~~ who, ~~them~~, ~~are considered with~~ ~~semi-1414~~, and will celebrate their " new year " (the hégire) on June 10? As for the Jews, their year 5754 began on September 16. Incontestably, time is a quite relative concept. The physicists [have known] that for a long time. ~~Policies~~ and ~~monks~~ too. But, ~~large amateurs~~ of certainty, the latter ~~were combined to~~ the scientists to try to offer to ~~the~~ men a chronological guide suitable ~~for~~ regulate a ~~social life~~ which they wanted ~~possible as~~ harmonious ~~as~~. ~~It is not a chance if~~ the Gregorian calendar that we currently use ~~in Occident~~ ~~were~~ imposed (under penalty of excommunication!) four centuries ago by a pope ~~grained~~ to see ~~the Easter Day~~ moving away ineluctably from the period that its predecessors ~~in year 325~~ had fixed ~~to him~~!

"Bonne année". **L'an 1994** aura commencé **samedi à zéro heure**, qui en douterait ? Et pourtant, cette évidence est pour le moins trompeuse. C'est ainsi que, décalage horaire aidant, les Australiens de Sydney fêtent la Saint-Sylvestre **avec neuf heures d'avance sur** nous et les Américains de Los Angeles **avec neuf heures de retard**. Un détail ? Admettons. Mais que dire, alors, des musulmans pratiquants qui, eux, se considèrent **à la mi-1414**, et célébreront leur "nouvel an" (l'hégire) **le 10 juin** ? Quant aux juifs, leur **année 5754** a débuté **le 16 septembre**. Incontestablement, le temps est une notion toute relative. Les physiciens savent cela depuis longtemps. Les politiques et les religieux aussi. Mais, grands amateurs de certitudes, ces derniers se sont alliés aux scientifiques pour tenter d'offrir aux hommes un guide chronologique susceptible de régler une vie sociale qu'ils voulaient **aussi** harmonieuse **que possible**. Ce n'est pas un hasard si le calendrier grégorien que nous utilisons actuellement **en Occident** fut imposé (sous

"Bonne année". **L'an 1994** aura commencé **samedi à zéro heure**, qui en douterait ? Et pourtant, cette évidence est pour le moins trompeuse. C'est ainsi que, décalage horaire aidant, les Australiens de Sydney fêtent la Saint-Sylvestre **avec neuf heures d'avance sur** nous et les Américains de Los Angeles **avec neuf heures de retard**. Un détail ? Admettons. Mais que dire, alors, des musulmans pratiquants qui, eux, se considèrent **à la mi-1414**, et célébreront leur "nouvel an" (l'hégire) **le 10 juin** ? Quant aux juifs, leur **année 5754** a débuté **le 16 septembre**. Incontestablement, le temps est une notion toute relative. Les physiciens savent cela depuis longtemps. Les politiques et les religieux aussi. Mais, grands amateurs de certitudes, ces derniers se sont alliés aux scientifiques pour tenter d'offrir aux hommes un guide chronologique susceptible de régler une vie sociale qu'ils voulaient **aussi** harmonieuse **que possible**. Ce n'est pas un hasard si le calendrier grégorien que nous utilisons actuellement **en Occident** fut imposé (sous peine d'excommunication !) **il y a quatre siècles** par un pape chagriné de voir le jour de Pâques s'éloigner de la période que lui avaient fixée ses prédécesseurs **en l'an 325** !

Bonne année = Happy new year; qui en douterait ? = who would doubt it?; et pourtant = and yet; pour le moins = at the very least; C'est ainsi que = thus; décalage horaire = time difference; la Saint-Sylvestre = New Year's Eve; musulman pratiquant = practising

"Bonne année". **L'an 1994** aura commencé **samedi à zéro heure**, qui en douterait ? Et pourtant, cette évidence est pour le moins trompeuse. C'est ainsi que, décalage horaire aidant, les Australiens de Sydney fêtent la Saint-Sylvestre **avec neuf heures d'avance sur** nous et les Américains de Los Angeles **avec neuf heures de retard**. Un détail ? Admettons. Mais que dire, alors, des musulmans pratiquants qui, eux, se considèrent **à la mi-1414**, et célébreront leur "nouvel an" (l'hégire) **le 10 juin** ? Quant aux juifs, leur **année 5754** a débuté **le 16 septembre**. Incontestablement, le temps est une notion toute relative. Les physiciens savent cela depuis longtemps. Les politiques et les religieux aussi. Mais, grands amateurs de certitudes, ces derniers se sont alliés aux scientifiques pour tenter d'offrir aux hommes un guide chronologique susceptible de régler une vie sociale qu'ils voulaient **aussi** harmonieuse **que possible**. Ce n'est pas un hasard si le calendrier grégorien que nous utilisons actuellement **en Occident** fut imposé (sous peine d'excommunication !) **il y a quatre siècles** par un pape chagriné de voir le jour de Pâques s'éloigner de la période que lui avaient fixée ses prédécesseurs **en l'an 325** !

" ~~Good year~~ ". ~~Will~~ ~~the year 1994~~ have begun Saturday ~~at zero hour~~, ~~which would doubt it~~? And yet, this obviousness is at the very least misleading. Thus, ~~time shift~~ helping, the Australian ~~ones~~ of Sydney celebrate New Year's Eve ~~with nine hours in advance on~~ us and the Americans of Los Angeles ~~with nine hours of delay~~. A detail? Let us admit [it]. But what to say, then, of the ~~Moslems practise~~ who, ~~them~~, ~~are considered with~~ ~~semi-1414~~, and will celebrate their " new year " (the hégire) on June 10? As for the Jews, their year 5754 began on September 16. Incontestably, time is a quite relative concept. The physicists [have known] that for a long time. ~~Policies~~ and ~~monks~~ too. But, ~~large amateurs~~ of certainty, the latter ~~were combined to~~ the scientists to try to offer to ~~the~~ men a chronological guide suitable ~~for~~ regulate a ~~social life~~ which they wanted ~~possible as~~ harmonious ~~as~~. ~~It is not a chance if~~ the Gregorian calendar that we currently use ~~in Occident~~ ~~were~~ imposed (under penalty of excommunication!) four centuries ago by a pope ~~grained~~ to see ~~the Easter Day~~ moving away ineluctably from the period that its predecessors ~~in year 325~~ had fixed ~~to him~~!

Bonne année = Happy new year; qui en douterait ? = who would doubt it?; et pourtant = and yet; pour le moins = at the very least; C'est ainsi que = thus; décalage horaire = time difference; la Saint-Sylvestre = New Year's Eve; musulman pratiquant = practising Muslim; se considérer = to considere oneself to be; nouvel an = new year; depuis longtemps (+Présent) = for a long time (+Present Perfect); grands amateurs = big fans; s'allier à = to ally oneself with; vie sociale = life in society; Ce n'est pas un hasard si = it is not by accident if; calendrier grégorien = Gregorian calendar; sous peine de = under penalty of; jour de Pâques = Easter

A Reliable tokenizer

... must process 4 types of **Atomic Linguistic Units** (not only simple words):

- affixes, e.g.: ***dis-, -ization***
- simple words, their inflection and their derivation, e.g. ***help, helped, helpful***,
- multi-word units, e.g. ***road map, round table***,
- discontinuous expressions, e.g. ***take ... into account***

Another way to tag the text

Battle-tested/A Japanese/A industrial managers/N here/ADV always/ADV buck up/V nervous/A newcomers/N with/PREP the/DET tale/N of/PREP **the first of their**/N countrymen/N to/PART visit/V Mexico/LOC, a boatload of/DET samurai warriors/N blown ashore/VPP **375 years ago**/DATE. From the beginning/DATE, it took/EXP1 a/DET man/N with/PREP extraordinary/A qualities/N to/EXP1 succeed/V in/PREP Mexico/LOC, says/V Kimihide Takimura/NPR, president/N of/PREP Mitsui/NPR group/N ’s Kensetsu Engineering Inc./ORG unit/N.

A Reliable Lexical Parser

... must represent ambiguities, because it is not always possible to correctly disambiguate at the early lexical stage:

There is a round table in room A32

It must represent all types of ambiguities: between simple words and sequences of affixes, between multiword units and sequences of simple words, etc. (100% recall)

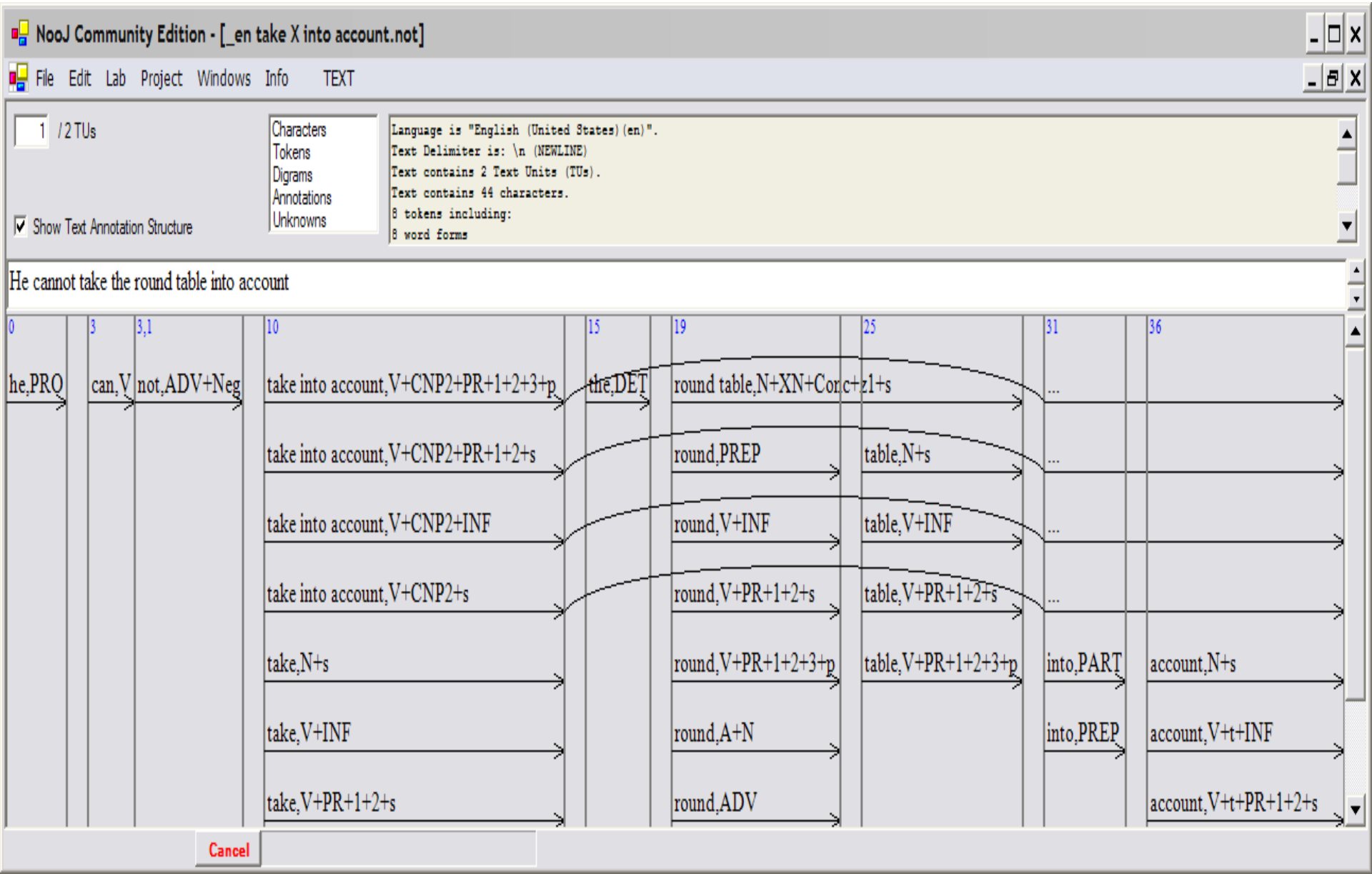
He cannot take the round table into account

He/PRO can/V not/ADV take into account/V the/DET round table/N

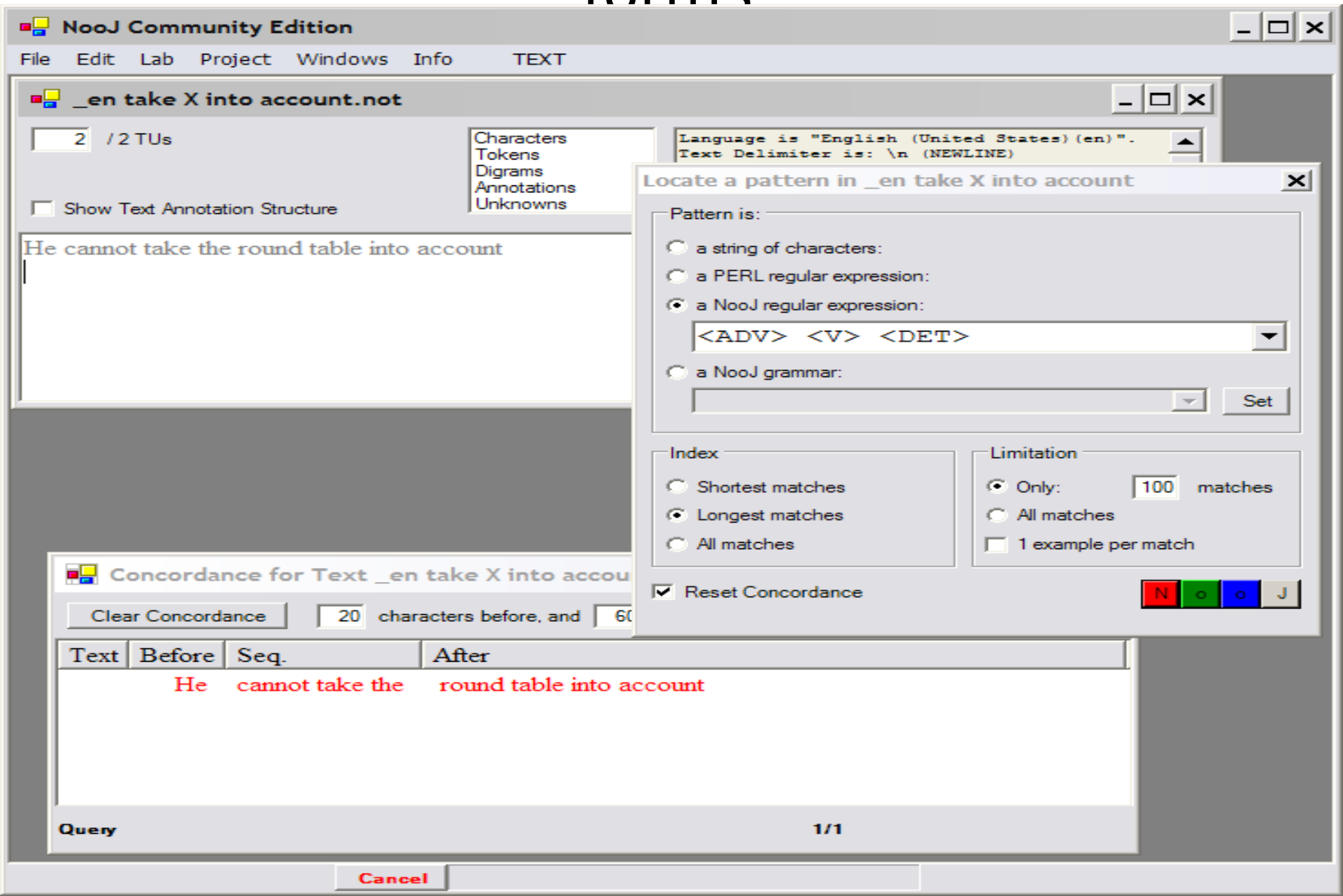
He cannot take the round table into account

He/PRO can/V not/ADV
(take into account/V + take/V
into/PREP account/N)
the/DET
(round table/N + round/A table/N)

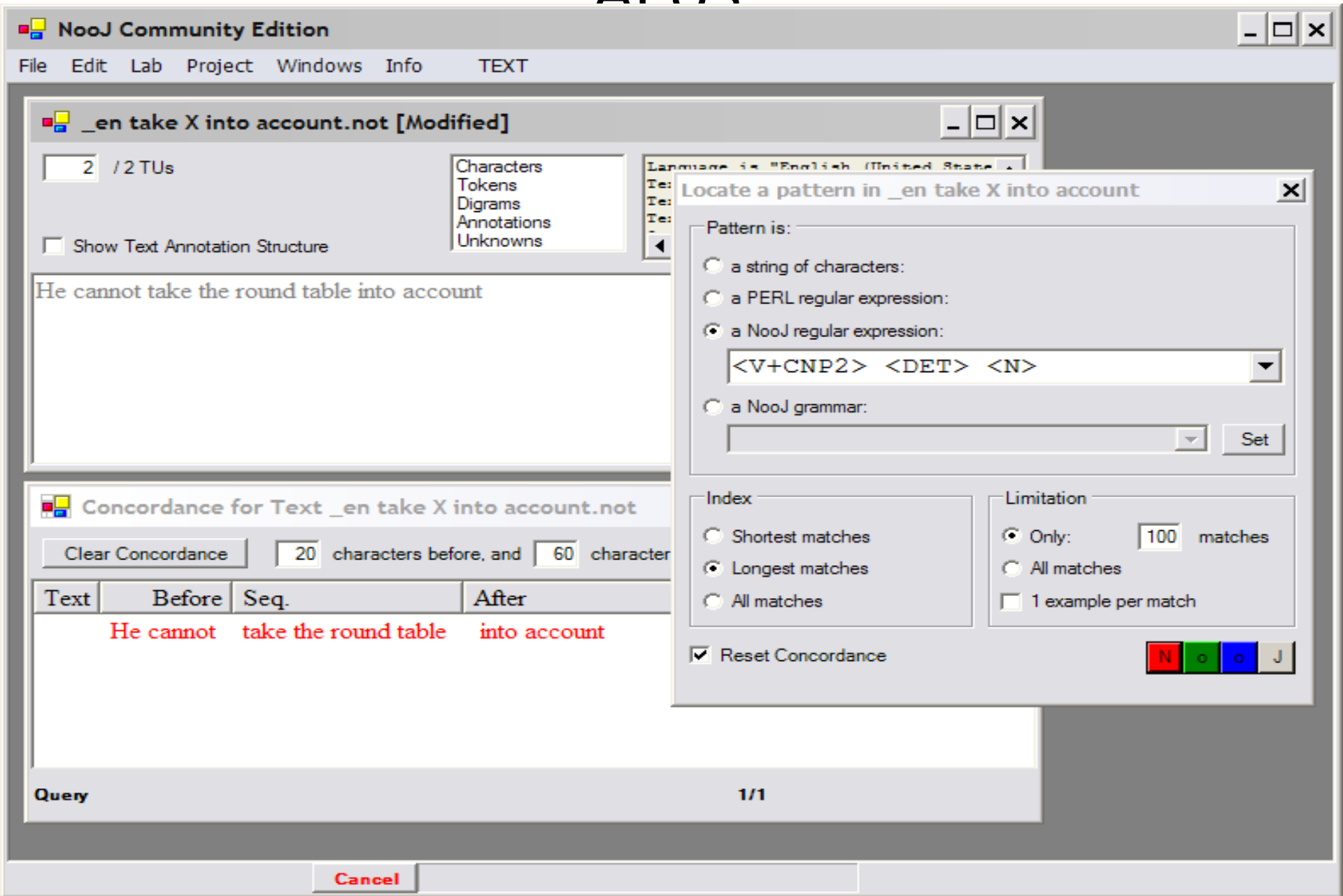
Text's Annotation Structure



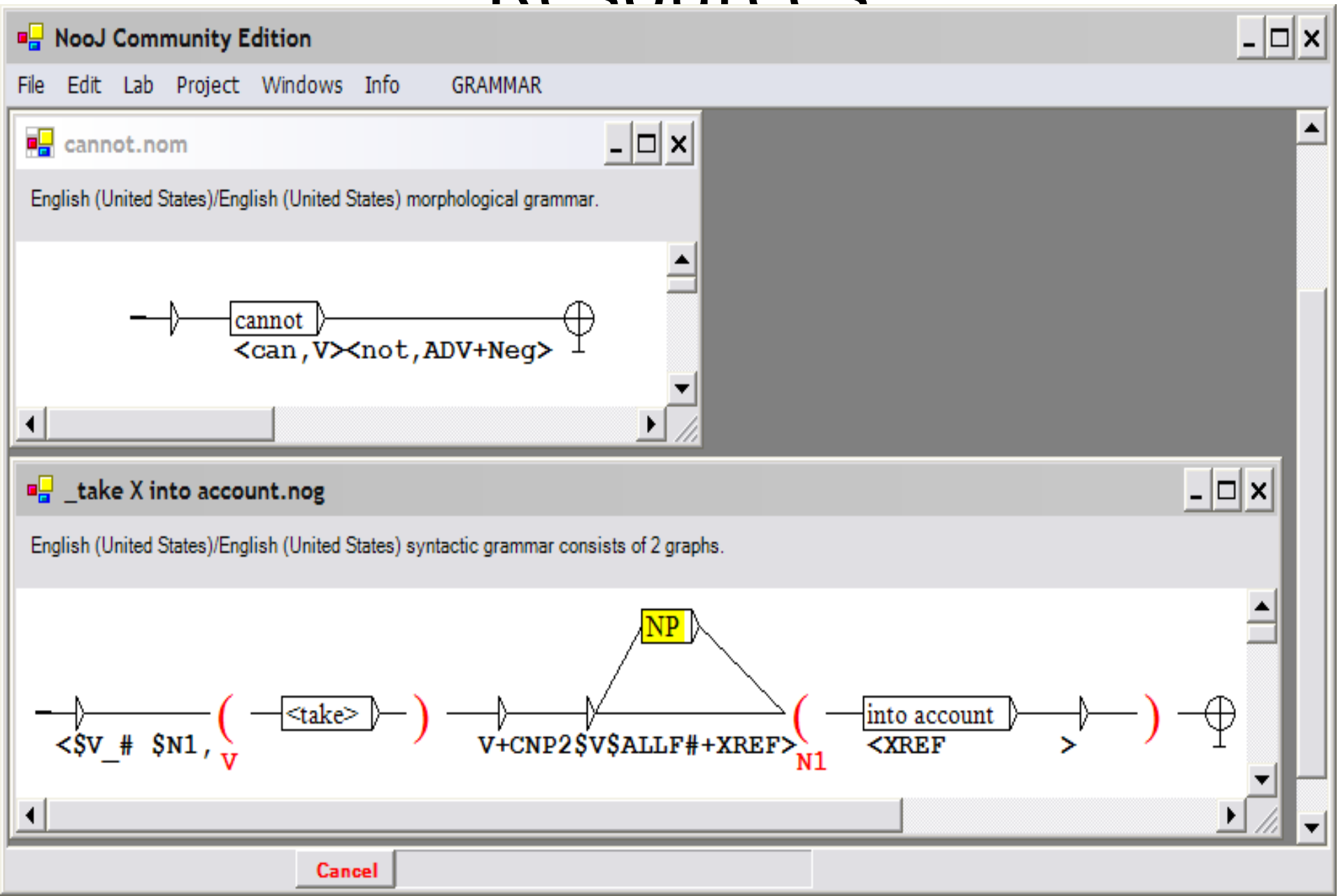
Locate sequences of ALUs inside word forms



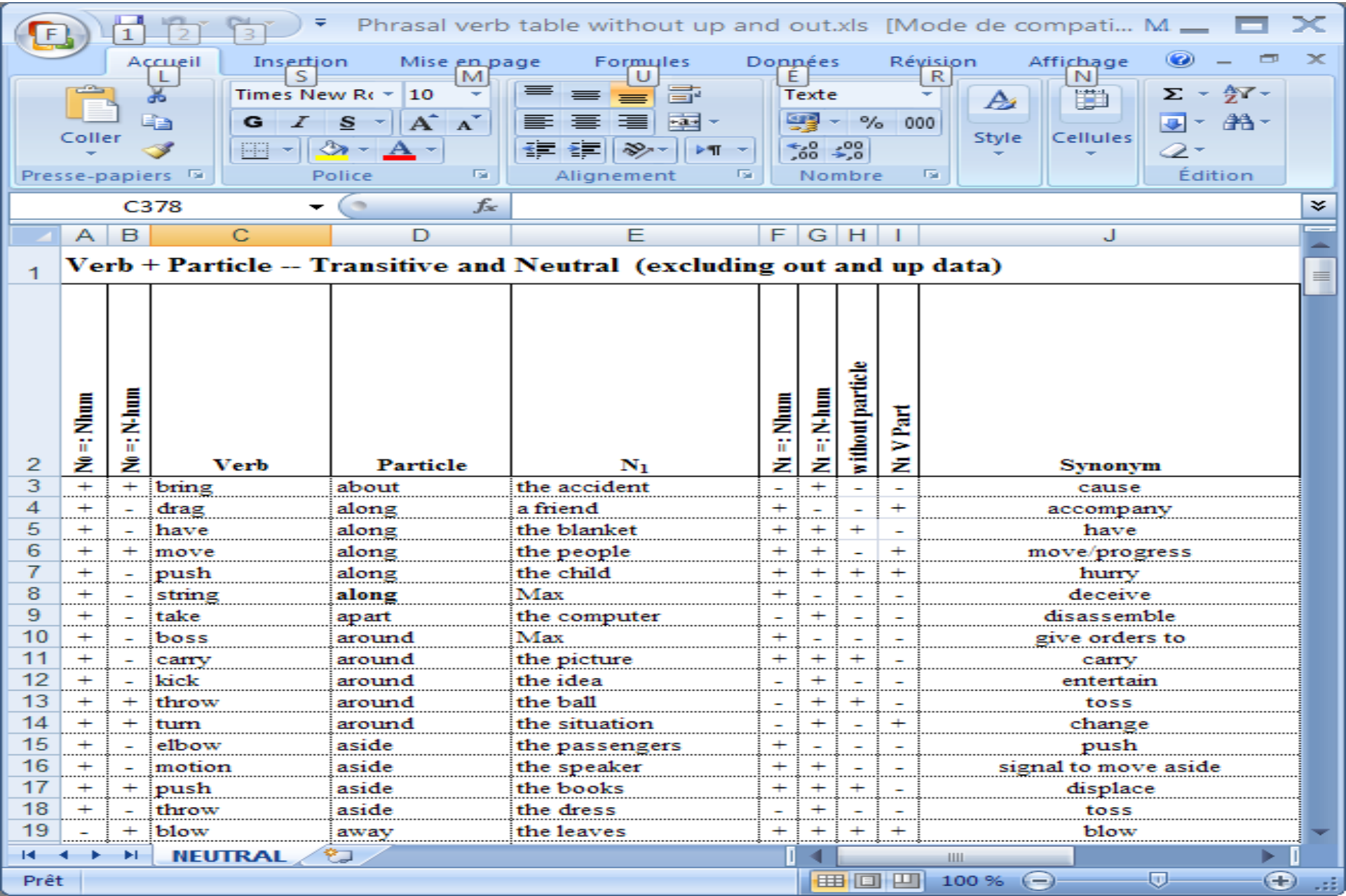
Locate sequences of discontinuous AUs



Morphological & Syntactic Resources



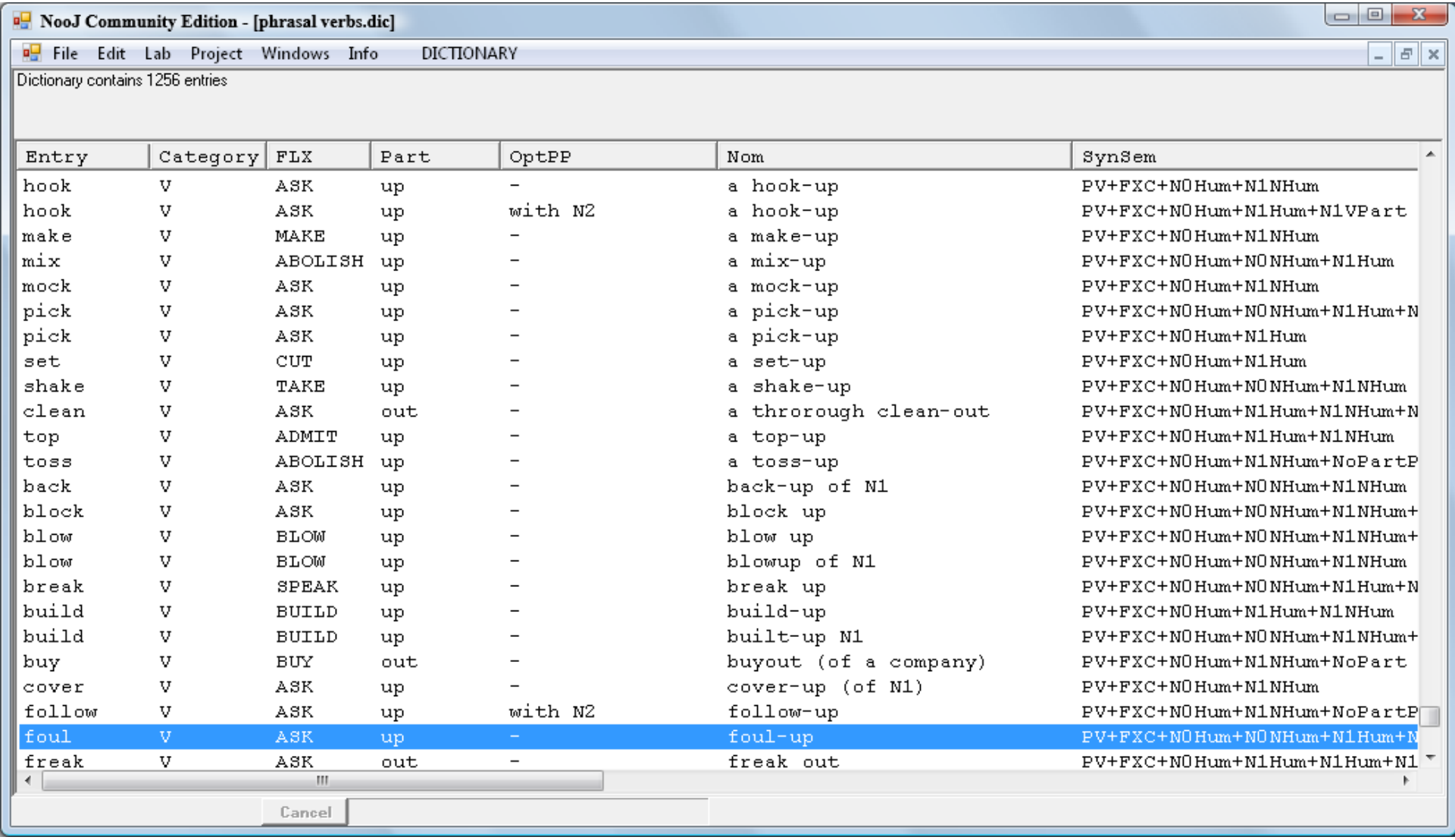
Discontinuous expressions e.g. English Phrasal Verbs



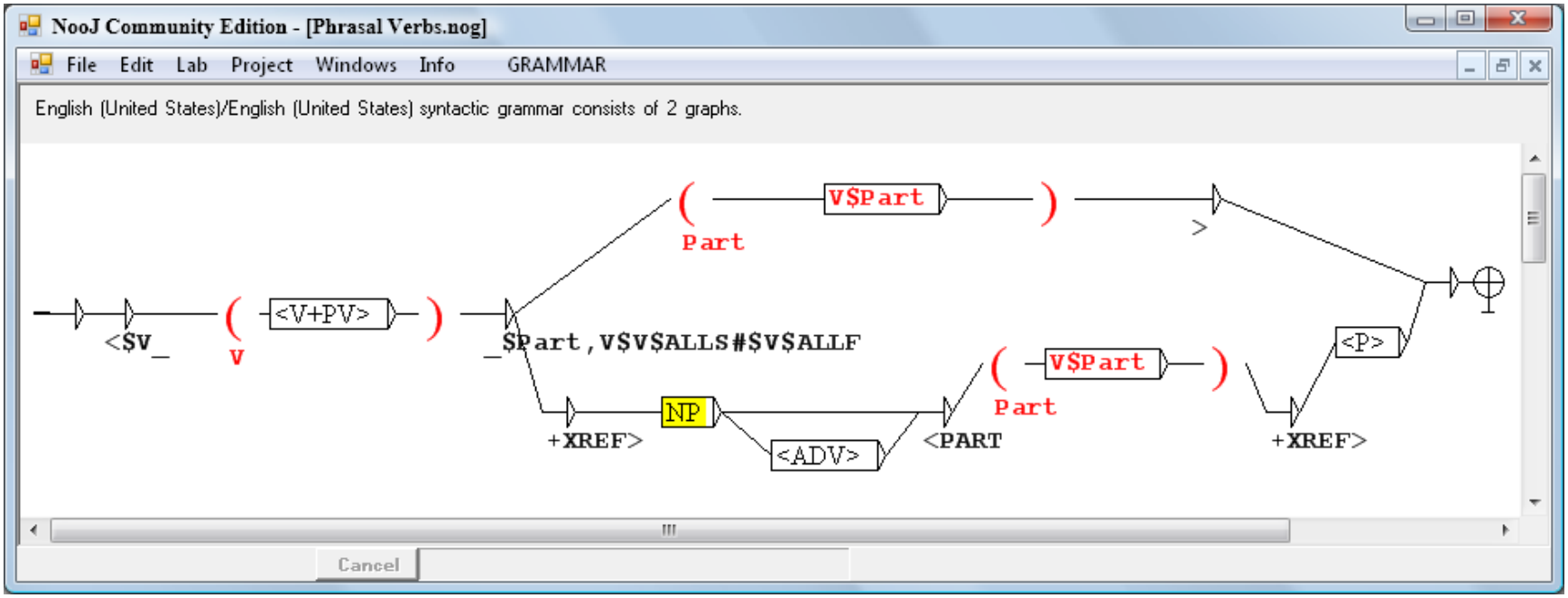
Verb + Particle -- Transitive and Neutral (excluding out and up data)

№ = Nhum	№ = N-hum	Verb	Particle	N_1	N_1 = Nhum	N_1 = N-hum	without particle	N_1 V Part	Synonym
+	+	bring	about	the accident	-	+	-	-	cause
+	-	drag	along	a friend	+	-	-	+	accompany
+	-	have	along	the blanket	+	+	+	-	have
+	+	move	along	the people	+	+	-	+	move/progress
+	-	push	along	the child	+	+	+	+	hurry
+	-	string	**along**	Max	+	-	-	-	deceive
+	-	take	apart	the computer	-	+	-	-	disassemble
+	-	boss	around	Max	+	-	-	-	give orders to
+	-	carry	around	the picture	+	+	+	-	carry
+	-	kick	around	the idea	-	+	-	-	entertain
+	+	throw	around	the ball	-	+	+	-	toss
+	+	turn	around	the situation	-	+	-	+	change
+	-	elbow	aside	the passengers	+	-	-	-	push
+	-	motion	aside	the speaker	+	+	-	-	signal to move aside
+	+	push	aside	the books	+	+	+	-	displace
+	-	throw	aside	the dress	-	+	-	-	toss
-	+	blow	away	the leaves	+	+	+	+	blow

NooJ dictionary

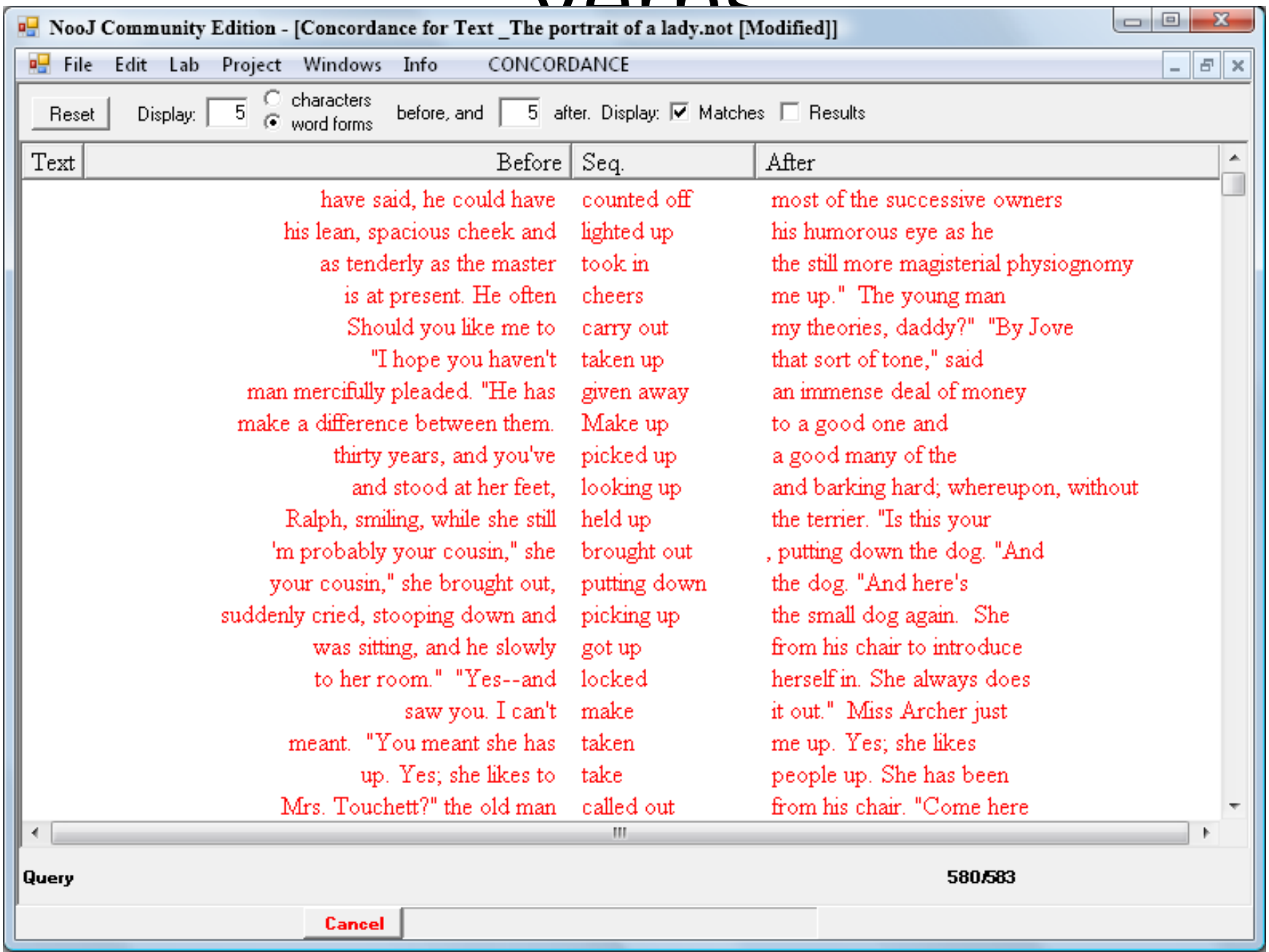


Entry	Category	FLX	Part	OptPP	Nom	SynSem
hook	V	ASK	up	-	a hook-up	PV+FXC+N0Hum+N1NHum
hook	V	ASK	up	with N2	a hook-up	PV+FXC+N0Hum+N1Hum+N1VPart
make	V	MAKE	up	-	a make-up	PV+FXC+N0Hum+N1NHum
mix	V	ABOLISH	up	-	a mix-up	PV+FXC+N0Hum+N0NHum+N1Hum
mock	V	ASK	up	-	a mock-up	PV+FXC+N0Hum+N1NHum
pick	V	ASK	up	-	a pick-up	PV+FXC+N0Hum+N0NHum+N1Hum+N
pick	V	ASK	up	-	a pick-up	PV+FXC+N0Hum+N1Hum
set	V	CUT	up	-	a set-up	PV+FXC+N0Hum+N1Hum
shake	V	TAKE	up	-	a shake-up	PV+FXC+N0Hum+N0NHum+N1NHum
clean	V	ASK	out	-	a throrough clean-out	PV+FXC+N0Hum+N1Hum+N1NHum+N
top	V	ADMIT	up	-	a top-up	PV+FXC+N0Hum+N1Hum+N1NHum
toss	V	ABOLISH	up	-	a toss-up	PV+FXC+N0Hum+N1NHum+NoPartP
back	V	ASK	up	-	back-up of N1	PV+FXC+N0Hum+N0NHum+N1NHum
block	V	ASK	up	-	block up	PV+FXC+N0Hum+N0NHum+N1NHum+
blow	V	BLOW	up	-	blow up	PV+FXC+N0Hum+N0NHum+N1NHum+
blow	V	BLOW	up	-	blowup of N1	PV+FXC+N0Hum+N0NHum+N1NHum
break	V	SPEAK	up	-	break up	PV+FXC+N0Hum+N0NHum+N1Hum+N
build	V	BUILD	up	-	build-up	PV+FXC+N0Hum+N1Hum+N1NHum
build	V	BUILD	up	-	built-up N1	PV+FXC+N0Hum+N0NHum+N1NHum+
buy	V	BUY	out	-	buyout (of a company)	PV+FXC+N0Hum+N1NHum+NoPart
cover	V	ASK	up	-	cover-up (of N1)	PV+FXC+N0Hum+N1NHum
follow	V	ASK	up	with N2	follow-up	PV+FXC+N0Hum+N1NHum+NoPartP
foul	V	ASK	up	-	foul-up	PV+FXC+N0Hum+N0NHum+N1Hum+N
freak	V	ASK	out	-	freak out	PV+FXC+N0Hum+N1Hum+N1Hum+N1

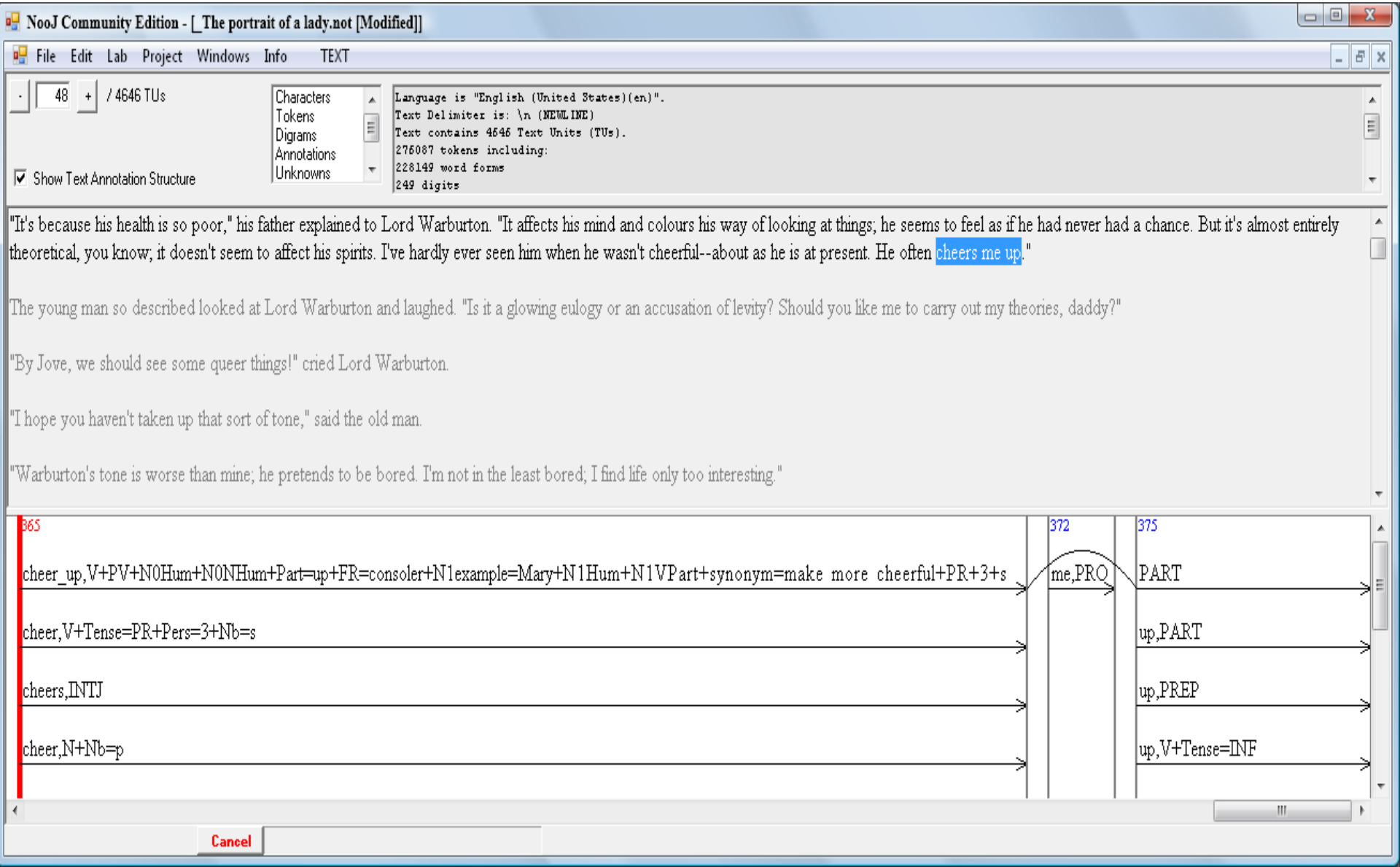


This grammar recognizes and automatically annotates (even discontinuous) phrasal verbs in texts

NooJ recognizes phrasal verbs



... and annotates them



Download NooJ, its manual and its resources from: www.nooj4nlp.net