

Disambiguation Tools for NooJ

Max Silberztein

LASELDI, MSH Ledoux
Université de Franche-Comté

www.nooj4nlp.net

NooJ v2.1

- NooJ recognizes and represents 5 types of text units:
 - morphemes, affixes (*dis-*, *-ization*) and contracted words (*cannot*)
 - simple words and their morphological variants (a laugh, two laughs, to laugh, laughable)
 - multi word units, semi-frozen terms and their variants
 - local syntactic units, e.g. complex determiners, dates
 - discontinuous expressions, e.g. collocations, support verb constructions
- NooJ provide various tools to identify these units: morphological FST and CF grammars, unified dictionaries for simple words, multi-word units and discontinuous expressions and syntactic grammars (FSTs, CFs and RTNs).

8 levels of formalization

- Linguists formalize 8 levels of linguistic phenomena: orthography, inflectional and derivational morphology, productive morphology, local syntax, structural syntax, transformational syntax, semantic extraction and analysis
- Each formalization level is an autonomous module, e.g. in the orthographical level: missing punctuation in Armenian, missing vowels in Arabic and Hebrew, character variants in Chinese, missing accents in French, etc.
- NooJ processes each level of analysis one after another in cascade, and thus must produce a result with 100% recall

an ideally tagged 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/PREP 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's/N Kensetsu Engineering
Inc./ORG unit/N.

A reliable parser with a 100% recall

... should represent unsolved ambiguities, because it is not always possible to remove all ambiguities automatically

There is a round table in room A32

Ambiguities are generated by each of the 8 levels of analyses: between morphemes and simple words, between simple words and multi-word units, between simple words and frozen expressions, etc.

All types of units are represented by annotations

The screenshot shows the NooJ Community Edition interface. The main window displays the text "He cannot take the round table into account". Below the text, a detailed annotation structure is shown, consisting of a grid of cells. The first cell (index 0) contains "he,PRO" and is highlighted with a red box and a red arrow. The second cell (index 3) contains "can,V". The third cell (index 3,1) contains "not,ADV+Neg". The fourth cell (index 10) contains "take into account,V+CNP2+INF". The fifth cell (index 15) contains "the,DET". The sixth cell (index 19) contains "round table,N+XN+Conc+z1+s". The seventh cell (index 25) contains "round,A+N". The eighth cell (index 25) contains "table,N+s". The ninth cell (index 31) contains "into,PREP". The tenth cell (index 36) contains "account,N+s". Arrows indicate the flow of annotations between cells. A "Cancel" button is visible at the bottom.

NooJ Community Edition - [take X into account.not]

File Edit Lab Project Windows Info TEXT

· 1 + / 2 TUs

Characters
Tokens
Digrams
Annotations
Unknowns

Show Text Annotation Structure

Language is "English (United States)(en)".
Text Delimiter is: \n (NEWLINE)
Text contains 2 Text Units (TUs).
8 tokens including:
8 word forms
Text contains 12 annotations (34 different)

He cannot take the round table into account

0	3	3,1	10	15	19	25	31	36
he,PRO	can,V	not,ADV+Neg	take into account,V+CNP2+INF	the,DET	round table,N+XN+Conc+z1+s
			take,V+INF		round,A+N	table,N+s	into,PREP	account,N+s

Cancel

Syntactic and Semantic annotations

Nool Community Edition - [La femme de trente ans.not [Modified]]

File Edit Lab Project Windows Info TEXT

95 / 926 TUs

Characters
Tokens
Digrams
Annotations
Unknowns

Language is "French (France) (fr)".
Text Delimiter is: \n (NEWLINE)
Text contains 926 Text Units (TUs).
Text contains 395126 characters.

Show Text Annotation Structure

"C'est un jeune Anglais, un gentilhomme, l'honorable Arthur Ormond, fils aîné de lord Grenville. Son histoire est intéressante. Il est venu à Montpellier en 1802, espérant que l'air de ce pays, ou il était envoyé par les médecins, le guérirait d'une maladie de poitrine à laquelle il devait succomber. Comme tous ses compatriotes, il a été arrêté par Bonaparte lors de la guerre, car ce monstre-là ne peut se passer de guerroyer. Par distraction, ce jeune Anglais s'est mis à étudier sa maladie, que l'on croyait mortelle. Insensiblement, il a pris goût à l'anatomie, à la médecine; il s'est passionné pour ces sortes d'arts, ce qui est fort extraordinaire chez un homme de qualité; mais le Régent s'est bien occupé de chimie ! Bref, M. Arthur a fait des progrès étonnants, même pour les professeurs de Montpellier; l'étude l'a consolé de sa captivité, et, en même temps, il s'est radicalement guéri. On prétend qu'il est resté deux ans sans parler, respirant rarement, demeurant couché dans une étable, buvant du lait d'une vache venue de Suisse, et vivant de cresson. Depuis qu'il est à Tours, il n'a vu personne, il est fier comme un paon; mais vous avez certainement fait sa conquête, car ce n'est probablement pas pour moi qu'il passe sous nos fenêtres deux fois par jour depuis que vous êtes ici... Certes, il vous aime."

Ces derniers mots réveillèrent la comtesse comme par magie. Elle laissa échapper un geste et un sourire qui surprirent la marquise. Loin de témoigner cette satisfaction instinctive ressentie même par la femme la plus sévère quand elle apprend qu'elle fait un malheureux, le regard de Julie fut terne et froid. Son visage indiquait un sentiment de répulsion voisin de

688	691	698	700	704	709
N+Sujet+Hum		si, CONJS	est, A+z1+m+s	bien, A+z1+m+s	occuper, V+OPER="attention D"+P1c+P10b0+PP+m+s
le, PRO+z1+3+m+s	régent, A+z2+m+s	se, PRO+PPV+3+s	est, A+z1+f+s	bien, A+z1+f+s	V+Psy+OP=attention D
le, DET+z1+m+s	régent, N+z2+m+s	se, PRO+PPV+3+p	est, A+z1+m+p	bien, A+z1+m+p	

16 sec Cancel

All parsers process all types of units in one unified way

The screenshot displays the NooJ Community Edition interface. The main window shows a text editor with the text "He cannot take the round table into account". A search dialog titled "Locate a pattern in _en take X into account" is open, showing the pattern "<ADV> <V> <DET>". The dialog also includes options for "Index" (Shortest matches, Longest matches, All matches) and "Limitation" (Only: 100 matches, All matches, 1 example per match). A concordance window at the bottom shows the text "He cannot take the round table into account" with the words "cannot", "take", "the", "round", "table", "into", and "account" highlighted in red. The concordance window also includes a "Clear Concordance" button and a "Reset Concordance" checkbox.

NooJ Community Edition
File Edit Lab Project Windows Info TEXT

_en take X into account.not

2 / 2 TUs

Language is "English (United States) (en)".
Text Delimiter is: \n (NEWLINE)

Show Text Annotation Structure

Characters
Tokens
Digrams
Annotations
Unknowns

He cannot take the round table into account

Locate a pattern in _en take X into account

Pattern is:

a string of characters:
 a PERL regular expression:
 a NooJ regular expression:
<ADV> <V> <DET>
 a NooJ grammar:

Index

Shortest matches
 Longest matches
 All matches

Limitation

Only: 100 matches
 All matches
 1 example per match

Reset Concordance

Concordance for Text _en take X into accou

Clear Concordance 20 characters before, and 60

Text	Before	Seq.	After
		He cannot take the round table into account	

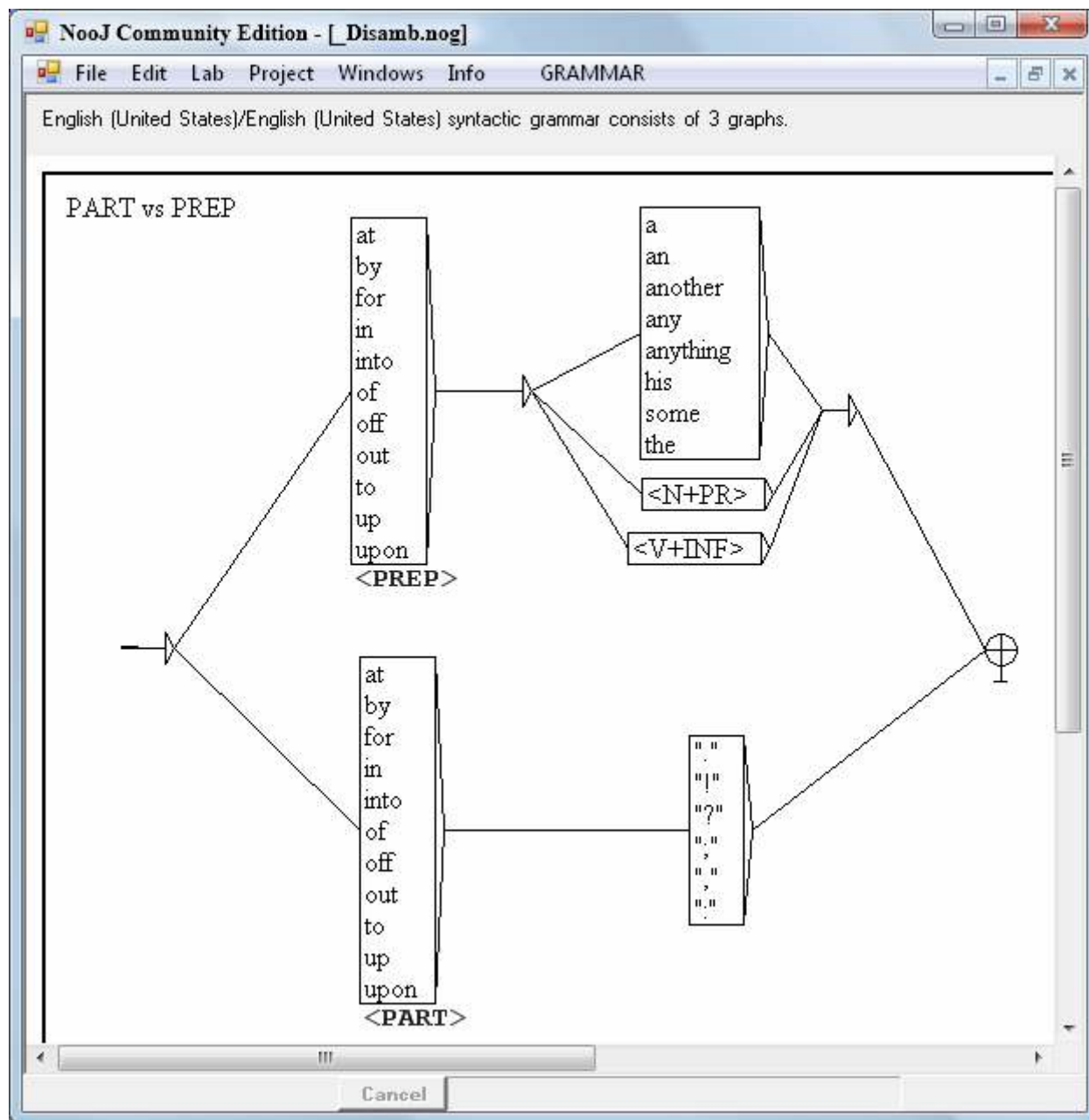
Query 1/1

Cancel

But there are needs for unambiguous texts

- construct « filtering » grammars
- manually
 - navigate the TAS
 - remove annotations (and undo)
- semi-automatically with concordances
 - Locate and disambiguate a pattern, eg.:
all/<DET> the
 - Locate all types of ambiguities
 - Locate all unambiguous forms

A Disambiguation Grammar



CONCORDANCE & TREE annotations

NooJ Community Edition - [Concordance for Text _The portrait of a lady.not]

File Edit Lab Project Windows Info TEXT CONCORDANCE

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

Text	Before	Seq.	After
platonc praise of the "distractions"	of Paris/<PREP>		--they were his great word
had come into the world	in Brooklyn/<PREP>		--though one could doubtless not
believe you're supposed not	to care/<PREP>		-through being so clever-for
exactly what she had expected	of Isabel/<PREP>		-to give it form and
long as you should be	in England/<PREP>		-to my care," said Goodwood
please me. She does it	to please/<PREP>		--to please-" And he lay
to feel attractions. You mean	to stay/<PREP>		--to settle? That would be
incense to be a compound	of long/<PREP>		-unanswered prayers. There was no
penetrate you! What am I	to believe/<PREP>		-what do you want me
sure of his coming down	to Gardens court/<PREP>		--which he will do the
like to like him." "Liking	to like/<PREP>		'--why, it makes a passion
quite capable of living there-	in summer/<PREP>		--with a maid-of-all
rich experienced, so easily come	by!/<PART>		--with a modesty at times
considerate than she now desired	to be/<PREP>		--would in fact be uproariously
You don't know what	to do/<PREP>		-you don't know where
know what you're going	to say/<PREP>		--you've had almost no

Query 14980/14980

11 sec Cancel

CONCORDANCE & TEXT annotations

The screenshot shows the NooJ Community Edition interface. The main window is titled "NooJ Community Edition - [Concordance for Text _The portrait of a lady.not]". The menu bar includes File, Edit, Lab, Project, Windows, Info, TEXT, and CONCORDANCE. The CONCORDANCE menu is open, showing options: Filter out selected lines, Filter out unselected lines, Select all, Unselect all, Annotate Text (add/remove annotations), Color matching sequences in text, Export Concordance, Export Index, Extract Matching Text Units, Extract Non Matching Text Units, and Build Statistical Report for matches. The "Annotate Text" option is highlighted. The main text area shows a concordance for the word "like". The text is displayed in red. The concordance entries are:

platonc praise of the "distrac		
had come into the v		
believe you're suppose		
exactly what she had exp		
long as you shou		
please me. She d		
to feel attractions. You		
incense to be a comp		
penetrate you! What		
sure of his coming		
like to like him." "Liking	to like/<PREP>	'--why, it makes a passion
quite capable of living there--	in summer/<PREP>	--with a maid-of-all
rich experienced, so easily come	by!/<PART>	--with a modesty at times
considerate than she now desired	to be/<PREP>	--would in fact be uproariously
You don't know what	to do/<PREP>	-you don't know where
know what you're going	to say/<PREP>	--you've had almost no

The bottom of the window shows a "Query" field with the text "14980/14980" and a "Cancel" button. The status bar at the bottom left indicates "11 sec".

filter annotations



TEXT

a lady.not

before, and 5 after. Display: Matches Results

Text	Before	Seq.	After
please me. She does it	to please/<PREP>		--to please-" And he lay
to feel attractions. You mean	to stay/<PREP>		--to settle? That would be
incense to be a compound	of long/<PREP>		-unanswered prayers. There was no
penetrate you! What am I	to believe/<PREP>		-what do you want me
sure of his coming down	to Gardencourt/<PREP>		--which he will do the
like to like him." "Liking	to like/<PREP>		'--why, it makes a passion
quite capable of living there--	in summer/<PREP>		--with a maid-of-all
rich experienced, so easily come	by!/<PART>		--with a modesty at times
considerate than she now desired	to be/<PREP>		--would in fact be uproariously
You don't know what	to do/<PREP>		-you don't know where
know what you're going	to say/<PREP>		--you've had almost no

GRAM = _Disamb 14980/14980

The portrait of a lady.not [Modified]

433 / 4646 TUs

Characters
Tokens
Digrams
Annotations
Unknowns

Language is "English (United States)(en)".
Text Delimiter is: \n (NEWLINE)
Text contains 4646 Text Units (TUs).
276087 tokens including:
... ..

Show Text Annotation Structure

English life. Isabel was often amused at his explicitness and at the small allowance he seemed to make either for her own experience or for her imagination. "He thinks I'm a barbarian," she said, "and that I've never seen forks and spoons"; and she used to ask him artless questions for the pleasure of hearing him answer seriously. Then when he had fallen into the trap, "It's a pity you can't see me in my war-paint and feathers," she remarked; "if I had known how kind you are to the poor savages I would have brought over my native costume!" Lord Warburton had travelled through the United States and knew much more about them than Isabel, he was so good as to say that America was the most charming country in the world, but his recollections of it appeared to encourage the idea that Americans in England would need to have a great many things explained to them.

5628	5633	5638	5643	5645	5653	5656
come, V+Tense=INF	by, PART	with, PREP	a, DET+Nb=s	modesty, N+Nb=s	at, PREP	time, N+Nb=s
come, V+Tense=PR+Pers=1+Nb=s						time, V+Tense=...

11 sec Cancel

CLICK IN TAGS & REMOVE an annotation

NooJ Community Edition - [The portrait of a lady.not]

File Edit Lab Project Windows Info TEXT

4 / 4646 TUs

Characters
Tokens
Digrams
Annotations
Unknowns

Language is "English (United States)(en)".
Text Delimiter is: \n (NEWLINE)
Text contains 4646 Text Units (TUs).
276087 tokens including:
...

Show Text Annotation Structure

Under certain circumstances there are few hours in life more agreeable than the hour dedicated to the ceremony known as afternoon tea. 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

Nb=p	28	34	38
→	there,INTJ	are,N+Nb=s+Distribution=Unit	few,DET+Distribution=Dadj+
→	there,PRO	be,V+Tense=PR+Pers=2+Nb=s+Syntax=AUX	few,N+Nb=p+Distribution=H
→	there,ADV	be,V+Tense=PR+Pers=1+Nb=p+Syntax=AUX	few,PRO+Nb=p

Cancel

Rapidly LOCate & INDEX annotations

The screenshot shows the NooJ Community Edition interface. The main window displays a concordance search for the pattern `its/<DET>` in the text `_The portrait of a lady`. The search results are displayed in a concordance window, showing the text with the pattern highlighted in blue. The concordance window also shows the pattern `its/<DET>` and the text `its/<DET>` next to the highlighted text.

The **CONCORDANCE** menu is open, showing the following options:

- Filter out selected lines
- Filter out unselected lines
- Select all
- Unselect all
- Annotate Text (add/remove annotations)**
- Color matching sequences in text
- Export Concordance
- Export Index
- Extract Matching Text Units
- Extract Non Matching Text Units
- Build Statistical Report for matches

The concordance window shows the following text:

had been n
digested, of table-talk that had lost
as to imagine a letter posted without
it with much grumbling at its ugliness,
had happened was out of proportion to
moment, then she dropped six words into
it. I give this little sketch of
privation. Her own large house, remarkable for
very humble servant, and the degree of
its/<DET> actuality. This hint of the
address. The Countess could not
antiquity, its incommidity, and who
appearance; there had really been
aromatic depths. "I love you
articles for what they may
assortment of mosaic tables and
attention was his only measure

RE = `its/<DET>` 100/100

Display the list of all ambiguities

The image shows two windows from the NooJ Community Edition software. The top window, titled 'NooJ Community Edition - [The portrait of a lady.not [Modified]]', has a menu bar with 'File', 'Edit', 'Lab', 'Project', 'Windows', and 'Info'. The 'TEXT' menu is open, and 'Ambiguities' is highlighted with a red circle. The main text area shows a snippet of text from 'The portrait of a lady'. The bottom window, titled 'NooJ Community Edition - [Ambiguities]', has a 'Select Analysis' dropdown menu with '<to,PREP>' selected, also circled in red. Below this is a table of ambiguities.

Freq	Annotations
7383	<to, PART> <to, PREP>
4487	<her, DET> <her, PRO>
4027	<that, CONJ> <that, DET+Distribution=Ddem+Nb=s> <that, PRO> <that, ADV>
3479	<have, V+FLX=HAVE+Tense=PT+Pers=1+Nb=s+Syntax=AUX> <have, V+FLX=HAVE+T
3473	<be, V+FLX=BE+Tense=PT+Pers=1+Nb=s+Syntax=AUX> <be, V+FLX=BE+Tense=PT+
3229	<in, N+FLX=APPLE+Nb=s> <in, PART> <in, PREP> <in, A>
1985	<as, CONJ> <as, PREP> <as, ADV+CmS=A>
1919	<for, CONJ> <for, PREP>
1867	<but, CONJ> <but, PREP> <but, ADV>
1836	<his, DET+Nb=s> <his, PRO>
1236	<say, V+FLX=LAY+Tense=PT+Pers=1+Nb=s> <say, V+FLX=LAY+Tense=PT+Pers=2+
1129	<have, V+FLX=HAVE+Tense=INF+Syntax=AUX> <have, V+FLX=HAVE+Tense=PR+Per
1097	<very, A+FLX=IER+N> <very, ADV+CmS=A>
1089	<on, PART> <on, PREP> <on, A>
1036	<this, DET+Nb=s> <this, PRO+Nb=s> <this, ADV>

At the bottom of the 'Ambiguities' window, there is a checkbox for 'Display Only Categories' and a note '5880 different types of ambiguities'. A 'Cancel' button is at the bottom center, and a color-coded legend (N, S, J) is at the bottom right.

CONCORDANCE FOR A SELECTED ambiguity

NooJ Community Edition - [Concordance for Text _The portrait of a lady.not]

File Edit Lab Project Windows Info CONCORDANCE

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

Text	Before	Seq.	After
life more agreeable than the hour dedicat...	to/<to,PREP>		the ceremony known as afternoon
that I have in mind in beginning	to/<to,PREP>		unfold this simple history offered
this simple history offered an admirable s...	to/<to,PREP>		an innocent pastime. The implements
the flood of summer light had begun	to/<to,PREP>		ebb, the air had grown
scene expressed that sense of leisure still	to/<to,PREP>		come which is perhaps the
such an hour. From five o'clock	to/<to,PREP>		eight is on certain occasions
not of the sex which is supposed	to/<to,PREP>		furnish the regular votaries of
served, and of two younger men strolling	to/<to,PREP>		and fro, in desultory talk
holding it for a long time close	to/<to,PREP>		his chin, with his face
to his chin, with his face turned	to/<to,PREP>		the house. His companions had

Query 7383/7383

Cancel

CONTINUING SIMILAR AMBIGUITIES to...

The screenshot displays the NooJ Community Edition interface. The top window, titled "Ambiguities", shows a table of analysis results for the category "<DET>". The table has two columns: "Freq" and "Annotations". The first row is highlighted in blue. A red circle highlights the "Display Only Categories" checkbox and the text "5880 different types of ambiguities".

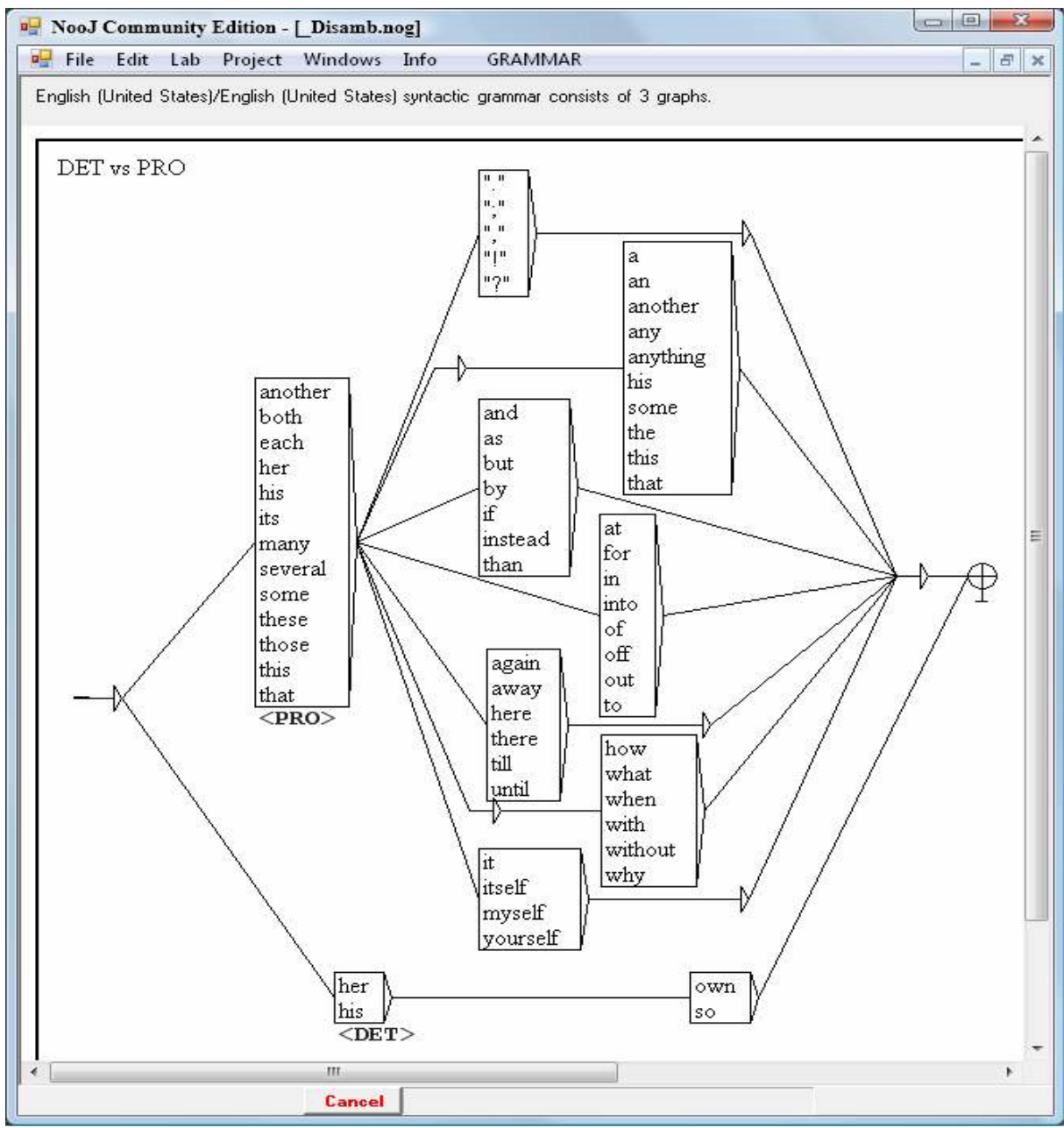
Freq	Annotations
278	<DET> <PRED> <PRO>
1836	<DET> <PRO>
1	<DET> <PRO>
170	<DET> <PRO>
108	<DET> <PRO>
36	<DET> <PRO>
117	<DET> <PRO>
56	<DET> <PRO>
82	<DET> <PRO>
118	<DET> <PRO>
391	<DET> <PRO>
44	<DET> <PRO>
306	<DET> <PRO> <ADV>

The bottom window, titled "Concordance for Text _The portrait of a lady.not", shows search results for the query "another". The results are displayed in a table with columns "Text", "Before", "Seq.", and "After".

Text	Before	Seq.	After
	-front; this was in quite	another	quarter. Privacy here reigned supreme
	believe in anything." "That's	another	sort of joke," said the
	uncommonly good talk." "Is that	another	sort of joke?" asked the
	came. But there had been	another	before, which I think contained
	the dog. "And here's	another	!" she added quickly, as the
	and from one room to	another	, preferring the places where the
	at his watch. "You've	another	quarter of an hour then

At the bottom of the window, there is a "Query" field containing "7402/7402" and a "Cancel" button.

... construct a general disambiguation grammar



Unambiguous words concordance

The screenshot shows the NooJ Community Edition interface. The top window, titled "Unambiguous Words", displays a list of words and their grammatical annotations. The word "of" is highlighted in blue. Below the list, it indicates "4601 unambiguous linguistic units" and shows a small keyboard layout with letters N, o, f, and J highlighted in red, green, blue, and grey respectively.

Freq	Annotation
8066	<the, DET>
6379	<a, DET+Nb=s>
6302	<of, PREP>
4832	<she, PRO>
4723	<I, PRO>
4206	<and, CONJ>

The bottom window, titled "Concordance for Text _The portrait of a lady.not", shows a concordance for the word "of". The search criteria are "characters" (selected), "before, and" (5), and "after" (5). The display options are "Matches" (checked) and "Results" (unchecked). The concordance table shows the following text:

Before	Seq.	After
that Pansy thought Mr. Rosier the nicest	of	all the young men sure
any sort. It's the total absence	of	all these things that pleases
things as you can here? In spite	of	all they say I maintain
knew. Lord Warburton tells me he wants,	of	all things in the world
something else to do. But she talked	of	all things with remarkable vividness

At the bottom of the concordance window, it shows "RE = its/<DET>" and "6302/6302". A "Cancel" button is visible at the bottom of the window.