

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

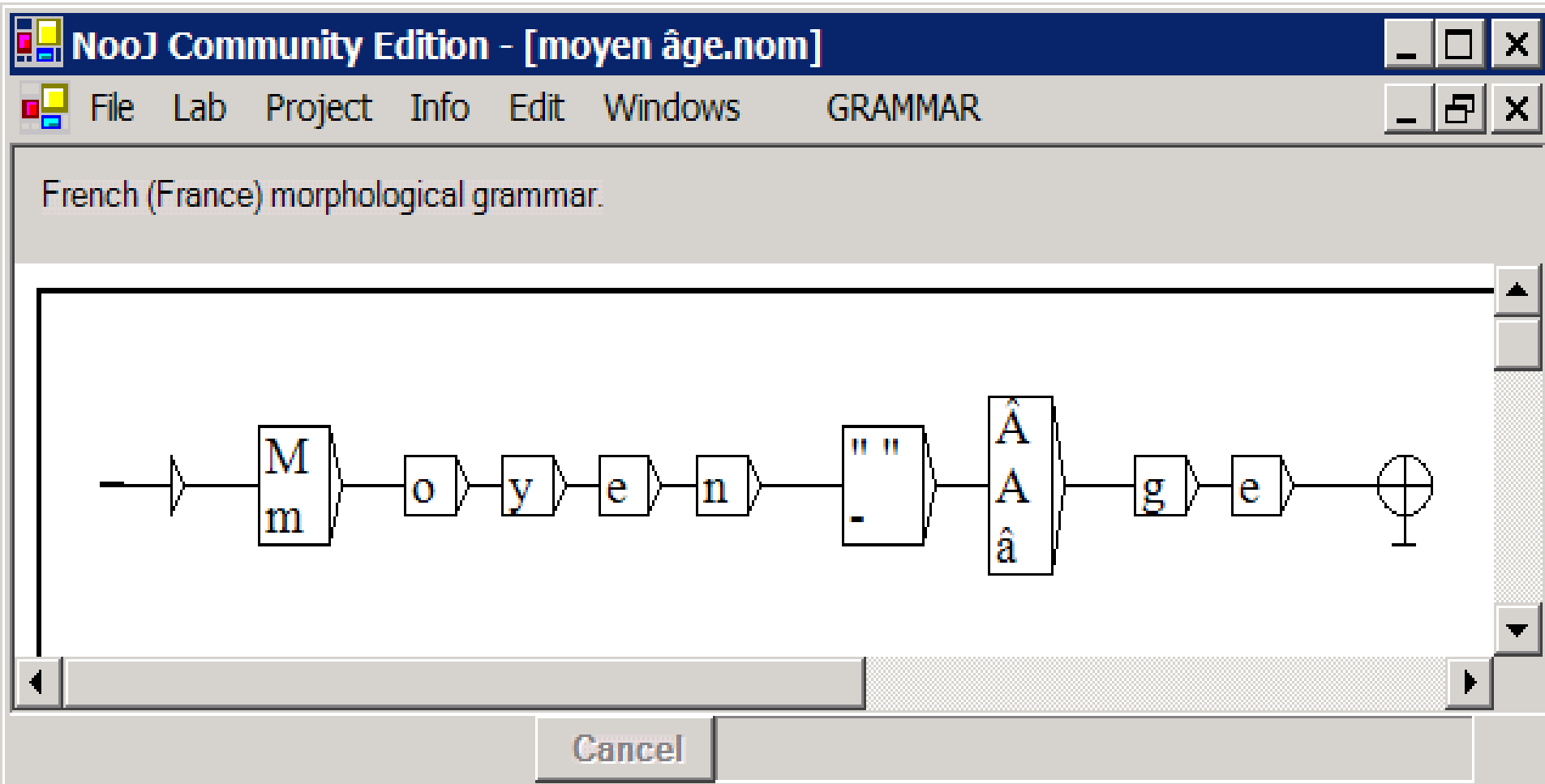
- Acadian (Univ. de Moncton) Arabic (UFC), Armenian (INALCO), Bulgarian (BACL), Chinese (UFC), English (UFC), French (UFC), Hebrew (Neguev Univ.), Hungarian (Academy of Sciences, Budapest), Italian (Univ. of Salerne), Portuguese (Linguatca), Serbo-Croatian (Univ. of Belgrade), Spanish (UAB); more modules 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"

The diagram illustrates a morphological grammar for the word 'tsar'. It starts with a root 't' and branches into 'c' and 's'. The 'c' branch leads to 't' and 's', which then lead to 'ar'. The 's' branch leads to 'z' and 's', which then lead to 'ar'. The 'ar' branch leads to 'ar, N+Hum'. From 'ar, N+Hum', the grammar branches into two paths: one for '+m' (masculine) and one for '+f' (feminine). The '+m' path leads to 'ina' and '+f' leads to 'ina'. The 'ina' branch leads to 's' and '+p' (plural). The 's' branch leads to 's' and '+p' leads to 's'. The final output is a circle with a plus sign, representing the lemmatised form '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).

<ADV+Date

early
late

Monday, June 5th

on the 3rd of June
tomorrow
a few days ago
once
in a few days

Mon, 16 Mar 2003

a few years ago
in a few years

in the afternoon

at 1532

at seven o'clock

at 7:10 am

>

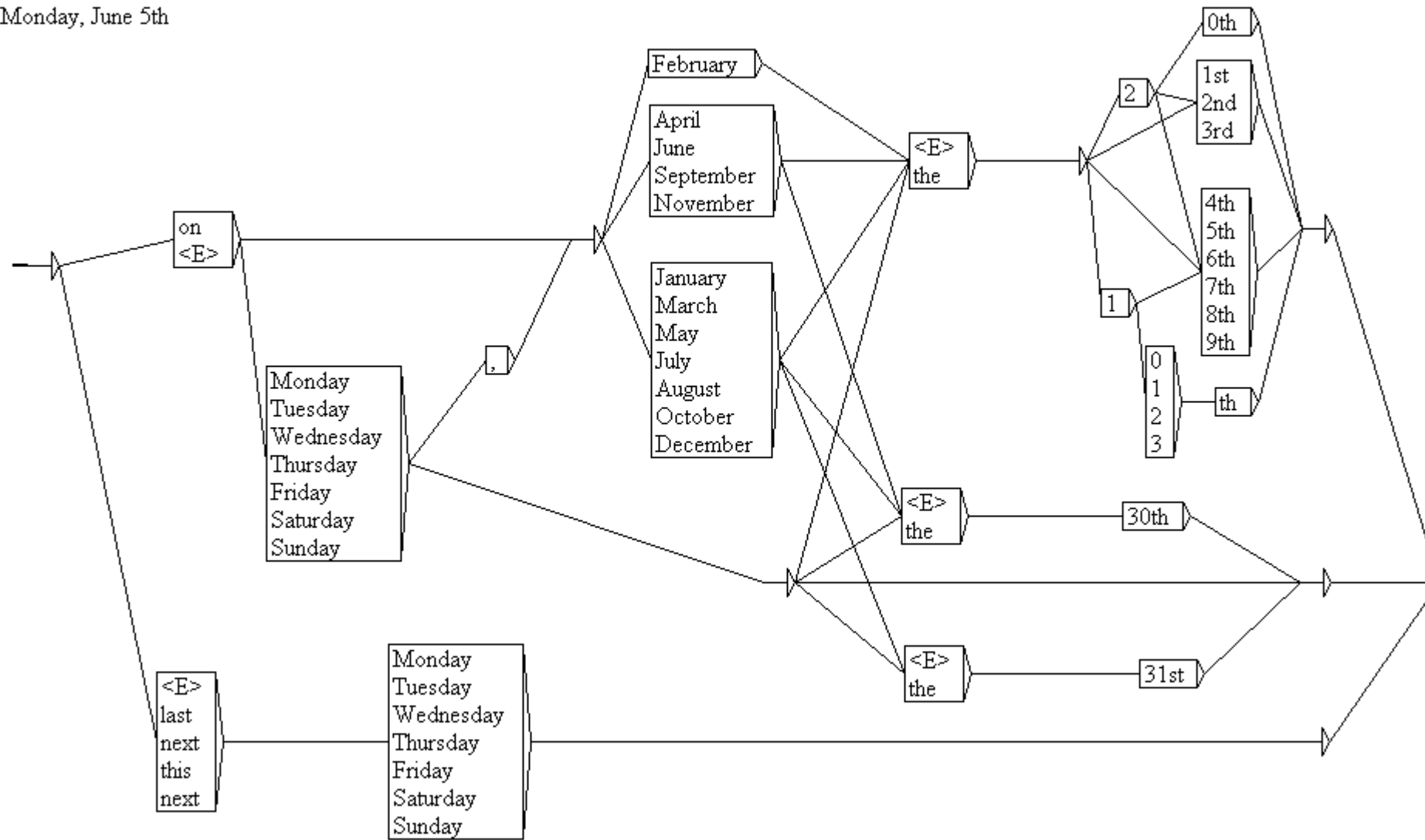
(1) Adverbs before the date are not processed here:
*early in the morning
vs
in the early morning

(2) Durations and Frequencies are not processed here:
*from Monday to Thursday
*all summer
*Monday all afternoon
*every Monday afternoon

Cancel

English (United States)/English (United States) syntactic grammar consists of 20 graphs (Community protected).

Monday, June 5th



Annotating large texts

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

File Edit Lab Project Windows Info TEXT

652 / 4646 TUs

Characters
Tokens
Digrams
Annotations
Unknowns

276087 tokens including:
228149 word forms
3124 digits
1252235 delimiters

Show Text Annotation Structure

"Haven't you got a heart?"

"I had one **a few days ago**, but I've lost it since."

"You're not serious," Miss Stackpole remarked; "that's what's the matter with you." But for all this, **in a day** or two, she again permitted him to fix her attention and on the later occasion assigned a different cause to her mysterious perversity.

"I know what's the matter with you, Mr. Touchett," she said. "You think you're too good to get married."

"I thought so till I knew you, Miss Stackpole," Ralph answered; "and then I suddenly changed my mind."

"Oh pshaw!" Henrietta groaned.

1	3	7	11	13	17	22
I,PRO	have,V+AUX+PT+1+2+3+s+p	one,DET+Dnum+s	ADV+Date			
	have,V+AUX+PP	one,PRO	a,DET+s	few,DET+Dadj+p	day,N+Ntime+p	ago,ADV
		one,N+s		few,N+HumColl+p	days,ADV	

Cancel

Need for annotations

words \neq linguistic units

- We don't pronounce blanks between words
- spaces are not always reliable:
cannot, don't, give in
parce que, afin (vs. à fin), au jour d'hui
- Semitic languages (e.g. *and in the house*),
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. Politicians 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 soeial 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***

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

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

File Edit Lab Project Windows Info TEXT

1 / 2 TUs

Characters
Tokens
Digrams
Annotations
Unknowns

Language is "English (United States) (en)".
Text Delimiter is: \n (NEWLINE)
Text contains 2 Text Units (TUs).
Text contains 44 characters.
8 tokens including:
8 word forms

Show Text Annotation Structure

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+PR+1+2+3+p	the,DET	round table,N+XN+Conc+z1+s	...		
			take into account,V+CNP2+PR+1+2+s		round,PREP	table,N+s	...	
			take into account,V+CNP2+INF		round,V+INF	table,V+INF	...	
			take into account,V+CNP2+s		round,V+PR+1+2+s	table,V+PR+1+2+s	...	
			take,N+s		round,V+PR+1+2+3+p	table,V+PR+1+2+3+p	into,PART	account,N+s
			take,V+INF		round,A+N		into,PREP	account,V+t+INF
			take,V+PR+1+2+s		round,ADV			account,V+t+PR+1+2+s

Cancel

Locate sequences of ALUs inside word forms

The screenshot shows the NooJ Community Edition interface. The main window displays the text "He cannot take the round table into account" with a search pattern "<ADV> <V> <DET>" applied. A dialog box titled "Locate a pattern in _en take X into account" is open, showing the search options and results. The search options include "a string of characters", "a PERL regular expression", "a NooJ regular expression" (selected), and "a NooJ grammar". The search results show the text "He cannot take the round table into account" with the words "cannot", "take", and "the" highlighted in red. The dialog box also includes options for "Index" (Shortest matches, Longest matches, All matches) and "Limitation" (Only: 100 matches, All matches, 1 example per match). A "Reset Concordance" checkbox is checked, and a "Cancel" button is visible at the bottom.

NooJ Community Edition
File Edit Lab Project Windows Info TEXT

_en take X into account.not
2 / 2 TUs
Characters
Tokens
Digrams
Annotations
Unknowns
 Show Text Annotation Structure

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

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

Locate sequences of discontinuous AIIIs

The screenshot shows the NooJ Community Edition interface. The main window displays the text "He cannot take the round table into account" with a concordance search overlay. The search dialog is titled "Locate a pattern in _en take X into account" and contains the following elements:

- Pattern is:**
 - a string of characters:
 - a PERL regular expression:
 - a NooJ regular expression:
 - a NooJ grammar:
- Index:**
 - Shortest matches
 - Longest matches
 - All matches
- Limitation:**
 - Only: matches
 - All matches
 - 1 example per match
- Reset Concordance
- Buttons: N (red), o (green), o (blue), J (grey)

The concordance table below the search dialog shows the following results:

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

At the bottom of the interface, there is a "Query" section showing "1/1" and a "Cancel" button.

Morphological & Syntactic Resources

The screenshot displays the NooJ Community Edition interface with two open grammar files:

- cannot.nom**: English (United States)/English (United States) morphological grammar. The diagram shows a morpheme `cannot` with its internal structure `<can, V><not, ADV+Neg>`.
- _take X into account.nog**: English (United States)/English (United States) syntactic grammar consists of 2 graphs. The diagram shows a syntactic structure for the phrase "take into account":
 - A root node `<$V_# $N1, V` branches into `<take>` and a node `(`.
 - The node `(` branches into `V+CNP2` and `VALLF#+XREF`.
 - `VALLF#+XREF` branches into `N1` and `>`.
 - `N1` branches into `into account` and `<XREF`.
 - `>` branches into `>` and a terminal node `⊕`.

A **Cancel** button is visible at the bottom of the window.

Discontinuous expressions e.g. English Phrasal Verbs

Phrasal verb table without up and out.xls [Mode de compati... M

Accueil Insertion Mise en page Formules Données Révision Affichage

C378

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

	No =; N-hum	No =; N-hum	Verb	Particle	N ₁	Ni =; N-hum	Ni =; N-hum	without particle	Ni V Part	Synonym
2	+	+	bring	about	the accident	-	+	-	-	cause
3	+	-	drag	along	a friend	+	-	-	+	accompany
4	+	-	have	along	the blanket	+	+	+	-	have
5	+	+	move	along	the people	+	+	-	+	move/progress
6	+	-	push	along	the child	+	+	+	+	hurry
7	+	-	string	along	Max	+	-	-	-	deceive
8	+	-	take	apart	the computer	-	+	-	-	disassemble
9	+	-	boss	around	Max	+	-	-	-	give orders to
10	+	-	carry	around	the picture	+	+	+	-	carry
11	+	-	kick	around	the idea	-	+	-	-	entertain
12	+	+	throw	around	the ball	-	+	+	-	toss
13	+	+	turn	around	the situation	-	+	-	+	change
14	+	-	elbow	aside	the passengers	+	-	-	-	push
15	+	-	motion	aside	the speaker	+	+	-	-	signal to move aside
16	+	+	push	aside	the books	+	+	+	-	displace
17	+	-	throw	aside	the dress	-	+	-	-	toss
18	-	+	blow	away	the leaves	+	+	+	+	blow

NEUTRAL

Prêt 100%

NooJ dictionary

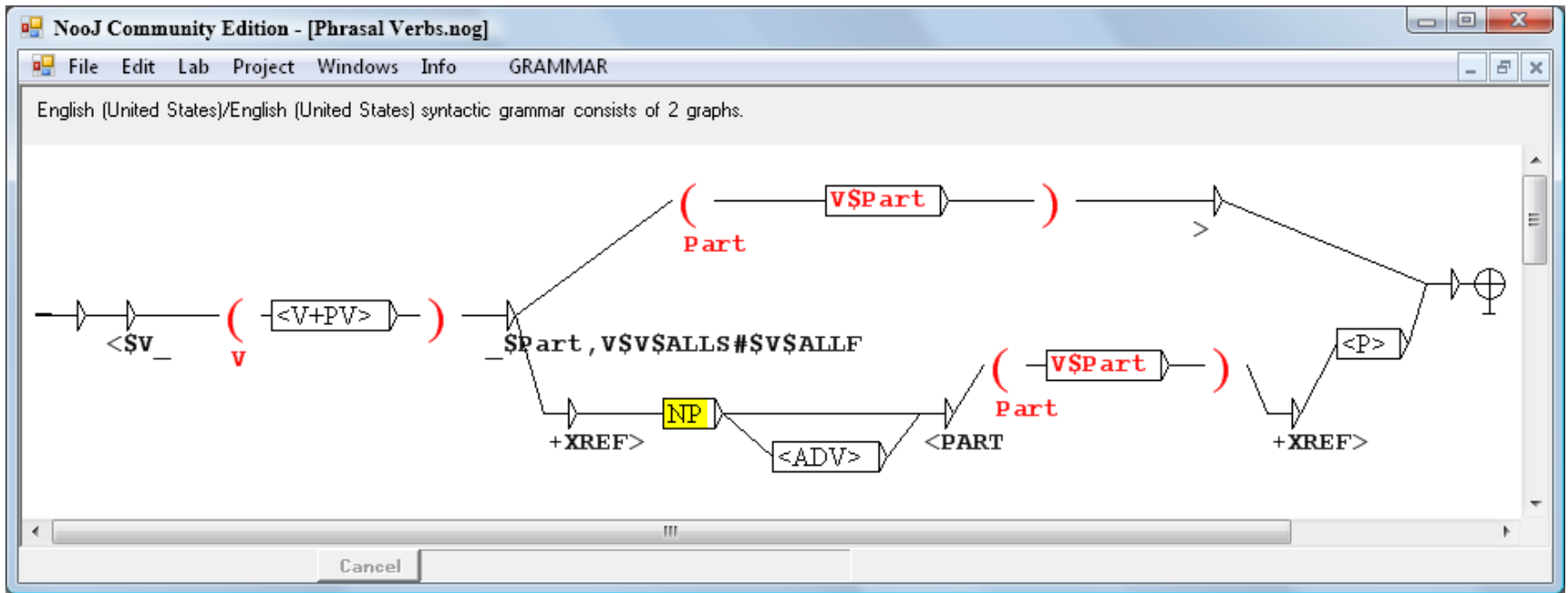
NooJ Community Edition - [phrasal verbs.dic]

File Edit Lab Project Windows Info DICTIONARY

Dictionary contains 1256 entries

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

Cancel



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

NooJ recognizes phrase verbs

The screenshot shows the NooJ Community Edition interface. The title bar reads "NooJ Community Edition - [Concordance for Text_The portrait of a lady.not [Modified]]". The menu bar includes "File", "Edit", "Lab", "Project", "Windows", and "Info". The main window title is "CONCORDANCE".

Search settings at the top: "Reset", "Display: 5", "characters" (radio button), "word forms" (radio button), "before, and 5 after", "Display: Matches Results".

Text	Before	Seq.	After
	have said, he could have	counted off	most of the successive owners
	his lean, spacious cheek and	lighted up	his humorous eye as he
	as tenderly as the master	took in	the still more magisterial physiognomy
	is at present. He often	cheers	me up." The young man
	Should you like me to	carry out	my theories, daddy?" "By Jove
	"I hope you haven't	taken up	that sort of tone," said
	man mercifully pleaded. "He has	given away	an immense deal of money
	make a difference between them.	Make up	to a good one and
	thirty years, and you've	picked up	a good many of the
	and stood at her feet,	looking up	and barking hard; whereupon, without
	Ralph, smiling, while she still	held up	the terrier. "Is this your
	'm probably your cousin," she	brought out	, putting down the dog. "And
	your cousin," she brought out,	putting down	the dog. "And here's
	suddenly cried, stooping down and	picking up	the small dog again. She
	was sitting, and he slowly	got up	from his chair to introduce
	to her room." "Yes--and	locked	herself in. She always does
	saw you. I can't	make	it out." Miss Archer just
	meant. "You meant she has	taken	me up. Yes; she likes
	up. Yes; she likes to	take	people up. She has been
	Mrs. Touchett?" the old man	called out	from his chair. "Come here

Query: 580/583

Cancel

... and annotates them

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

File Edit Lab Project Windows Info TEXT

48 / 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:
228149 word forms
249 digits

Show Text Annotation Structure

"It's because his health is so poor," his father explained to Lord Warburton. "It affects his mind and colours his way of looking at things; he seems to feel as if he had never had a chance. But it's almost entirely theoretical, you know; it doesn't seem to affect his spirits. I've hardly ever seen him when he wasn't cheerful--about as he is at present. He often **cheers me up.**"

The young man so described looked at Lord Warburton and laughed. "Is it a glowing eulogy or an accusation of levity? Should you like me to carry out my theories, daddy?"

"By Jove, we should see some queer things!" cried Lord Warburton.

"I hope you haven't taken up that sort of tone," said the old man.

"Warburton's tone is worse than mine; he pretends to be bored. I'm not in the least bored; I find life only too interesting."

365	372	375
cheer_up,V+PV+N0Hum+N0NHum+Part=up+FR=consoler+N1example=Mary+N1Hum+N1VPart+synonym=make more cheerful+PR+3+s	me,PRO	PART
cheer,V+Tense=PR+Pers=3+Nb=s		up,PART
cheers,INTJ		up,PREP
cheer,N+Nb=p		up,V+Tense=INF

Cancel

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

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

File Edit Lab Project Windows Info TEXT

1 / 2 TUs

Characters
Tokens
Digrams
Annotations
Unknowns

Language is "English (United States) (en)".
Text Delimiter is: \n (NEWLINE)
Text contains 2 Text Units (TUs).
Text contains 44 characters.
8 tokens including:
8 word forms

Show Text Annotation Structure

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+PR+1+2+3+p	the,D	round,table,N+XN+Conc...			
			take into account,V+CNP2+PR+1+2+3+p		round,PREP	table,N+s		
			take into account,V+CNP2+PR+1+2+3+p		round,V+INF	table,V+INF		
			take into account,V+CNP2+PR+1+2+3+p		round,V+PR+1+2+s	table,V+PR+1+2+s		
			take into account,V+CNP2+PR+1+2+3+p		round,V+PR+1+2+3+p	table,V+PR+1+2+3+p	into,PART	account,N+s
			take,V+INF		round,A+N		into,PREP	account,V+t+INF
			take,V+PR+1+2+s		round,ADV			account,V+t+PR+1+2+s

Cancel

Thank you!