



Sous la co-tutelle de :

CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Local Grammar Graphs

Eric Laporte



Sous la co-tutelle de :

CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Outline

Concordance

Local grammar of dates

Invoking a subgraph

Lexical masks

Dictionaries of a text



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Concordance of a word

rman Fowler, said Mr. Major's job was safe.{S} "He
licy changes."{S} Mr. Major's hold on power may hir
RITAIN--PARLIAMENT{S} MAJOR, JOHN{S} CHIDGEY, DAVII
lots at SAS and other major companies were not taki
th insurance took two major strides forward in the
health advocates as a major setback for the tobacc
yamira of Burundi and Major General{S} Juv?nal Haby
an War{S} Headline 2: Major Jean-Guy Plante, left,
rror show continues," Major General Romeo Dallaire
vy fighting along the major highways.{S} General D
forces is one of the major problems which the army



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Concordance: diversity

cers, Thomas found the commanders a bit disturbed.{S} "
en found.{S} "And, mon colonel," the officer said, "eve
ghting on Omaha Beach, Captain Joseph{S} Dawson, introd
ere as a veteran.{S} A captain in the 101st{S} Airborne
t D. Eisenhower, their commander, the day's accomplishm
ecalled the feat of{S} Lieutenant Colonel James E. Rudd
feat of{S} Colonel James E. Rudder and his
a Long, Long Time."{S} General Orwin Clark Talbott, who
ne attack, the militia commander, Lam Horm, and his men
ter, in the morning, a general, one of 2,000 in the 140
eeing soldiers killing generals, generals selling food,



Sous la co-tutelle de:
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Concordance: sequences

Text: {S} A senior government official charged Wednesday
Abel {S} Olopai, a government official who was 10 years
regrets," a German government official said. {S} Key Word
WILLIAM J. (U.S. GOVERNMENT OFFICIAL) {S} U.S.--ARMED
WILLIAM J. (U.S. GOVERNMENT OFFICIAL) {S} U.S.--ARMED
te Gatete, a local government official from southern Rwanda
collapse in 1991, a government official said Sunday. {S}
pens next month, a government official said. {S} Until the
bund," said a U.S. government official. {S} He said Washington
WILLIAM J. (U.S. GOVERNMENT OFFICIAL) {S} International
of the century, a government official said. {S} "We want



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Concordancers

Objective

Explore a linguistic phenomenon

Explore a text

Query language

A major criterion of quality

Allow diversity and sequences at the same time: graphs



Sous la co-tutelle de :

CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Outline

Concordance

Local grammar of dates

Invoking a subgraph

Lexical masks

Dictionaries of a text



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

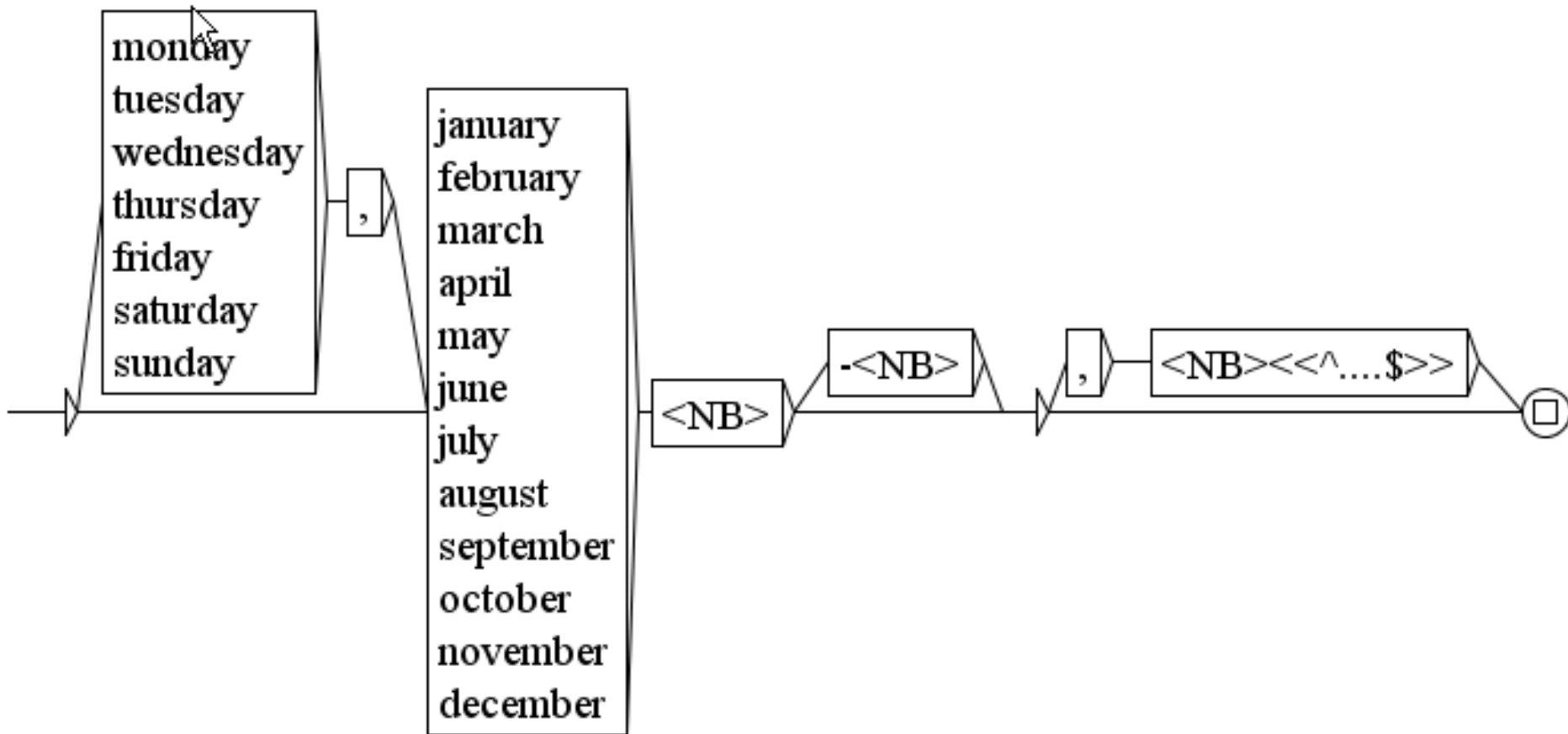
Dates in English

: Hong Kong after the July 1, 1997, handover
ble candidate for the April 1995 presidential
up of drivers for the July 3 race would be ma
asily blurred, as the June 6 anniversary show
as been as bad as the April 26 crash of the T
which organizes the June 26 regional electi
chairmen have set the July 4 recess as the de
airlines have put the July 4th weekend on sal
been mandated by the June 26 election, the G
worked, helping make June 6, 1944, the bigge
began a hunger strike May 25 after two of the



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

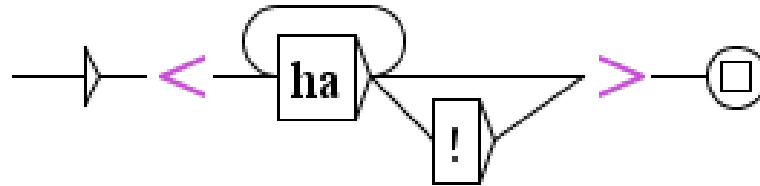
Query for dates in English





Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Finite automata and regular expressions

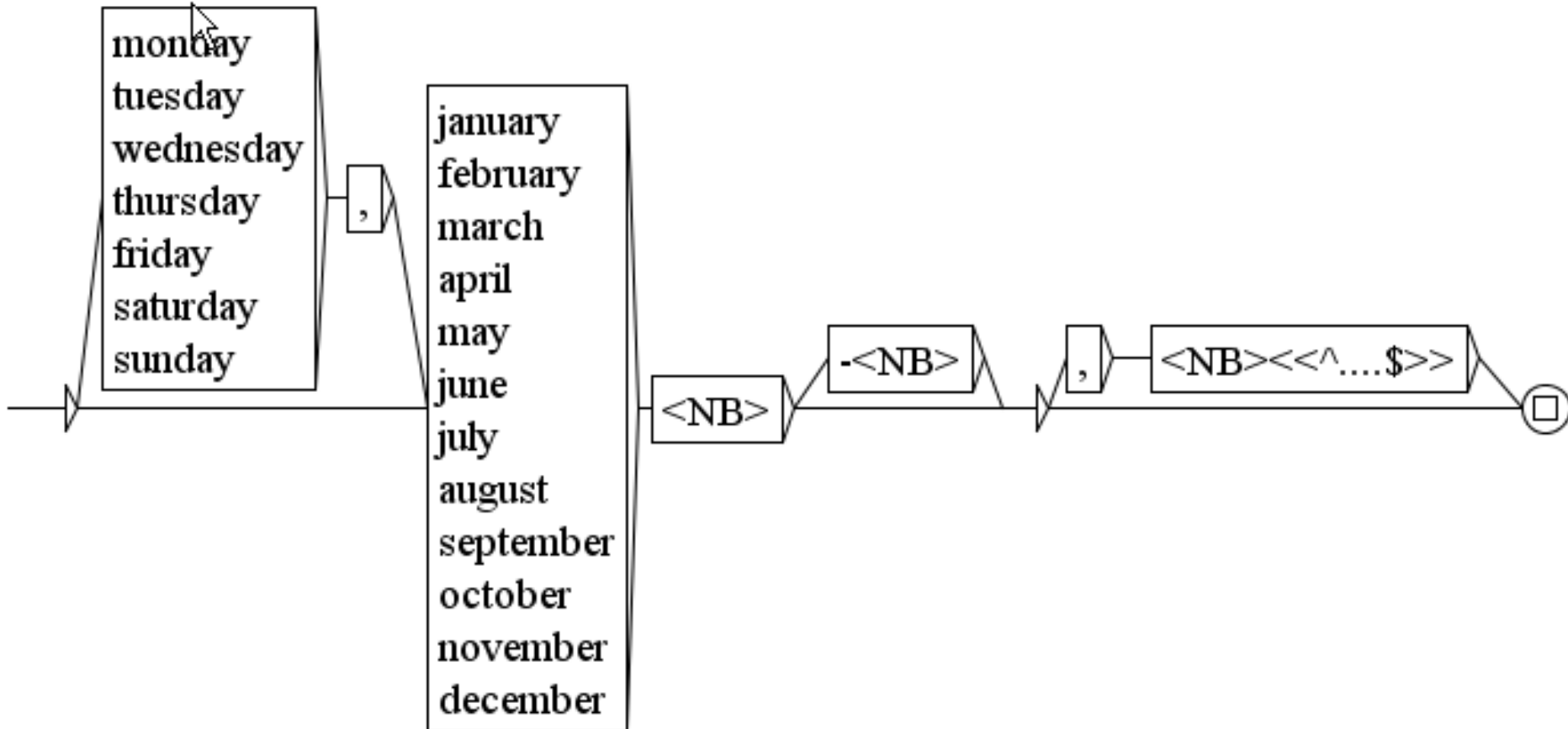


Equivalent regular expression:
 $(ha)^*!?$



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Imagine the regular expression for this one





Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Beyond concordances

Such graphs are also used in NLP applications:

- indexation

- information extraction

- annotation



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

How to edit a graph with Unitex-Gramlab

So that you can play around if I'm boring
Menu FSGraph > New

Create a box

Ctrl-click where you want to create the box
Fill in the field above the graph
Validate with Enter

Create a transition

Click once on the source box
Click once on the target box



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

How to edit a graph with Unitex-Gramlab

Select a box

Click once on it
It becomes blue

Unselect a box

Click on the white background of the graph

Double-clicking on a box

Same as clicking twice
Creates a transition from the box back to it (loop)

Select several boxes

Draw a rectangle around them
You can move, delete, copy, paste them together



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

How to edit a graph with Unitex-Gramlab

Delete a box

Select it

Press Delete

Validate with Enter

Initial and final boxes

You cannot remove them

You cannot create new ones

Delete a transition

Click once on the source box

Click once on the target box



Sous la co-tutelle de :

CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Outline

Concordance

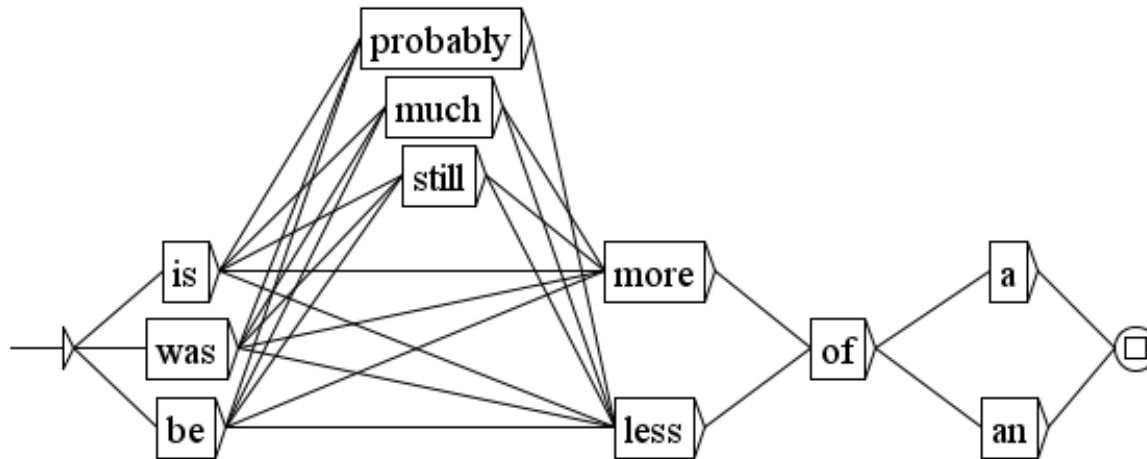
Local grammar of dates

Invoking a subgraph

Lexical masks

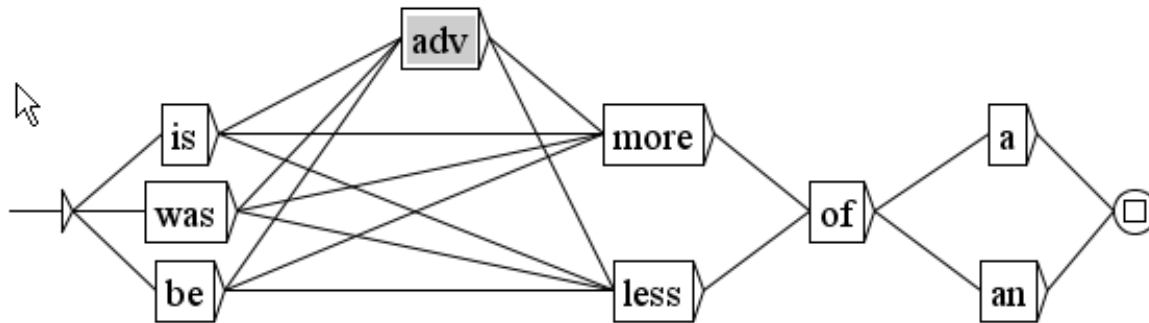
Dictionaries of a text

A graph



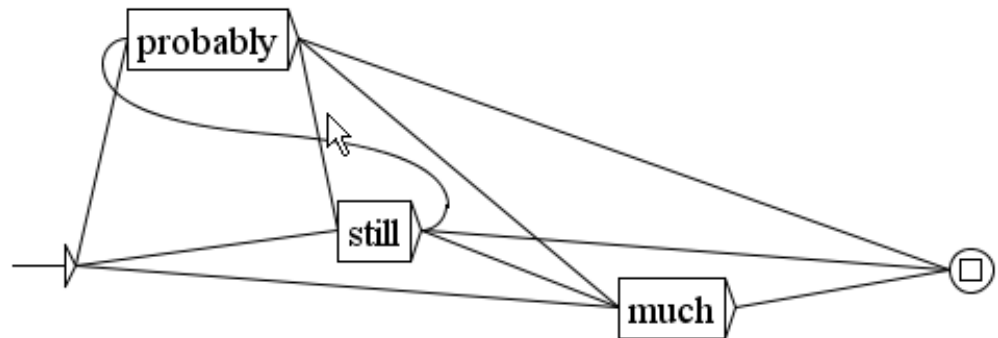
...l rising, it is probably less of a problem politically than it was
The new guru was less of a showman and more of a technician who be
& "and there is much less of a feeling of impending crisis here th
said, "There is less of a sense among juries that it is other peop
sun proved to be more of a menace than soccer hooligans at the firs
change would be more of a threat.{S} But the prevailing belief in
ts.{S} K mart is more of a runaway scraper; its stock is off 40 perc
on, he would be more of a provocation to the British, officials sa
or individual is still more of an art than a science," the former C

Invoking a subgraph from a node



The adv node invokes the adv.grf graph
Equivalent to substituting the graph for the node

adv.grf:





Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Invoking a subgraph from a node

Objectives

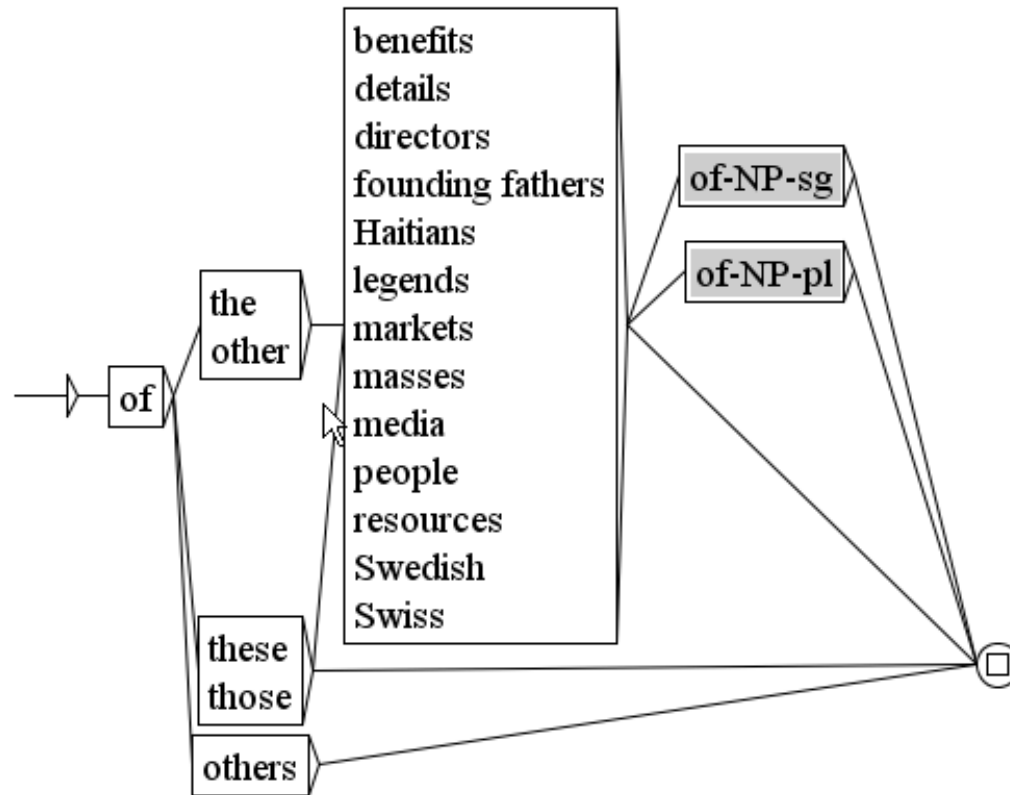
Use the same graph in various contexts

Distribute a description in several graphs



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Invoking a subgraph from a node



A graph can invoke itself



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Local grammars

A graph and its subgraphs describing a set of expressions
little dependent from the rest of syntax

Local grammar graphs (LGG)

A type of language resources

Used to recognise and process the expressions

Can be included in a production chain



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

A vintage area of NLP

GROSS, Maurice. 1997. The Construction of Local Grammars. *Finite-State Language Processing*, The MIT Press, pp. 329-352.

(<http://books.google.fr/books?id=q4URKd5XKo0C&pg=PA329>)

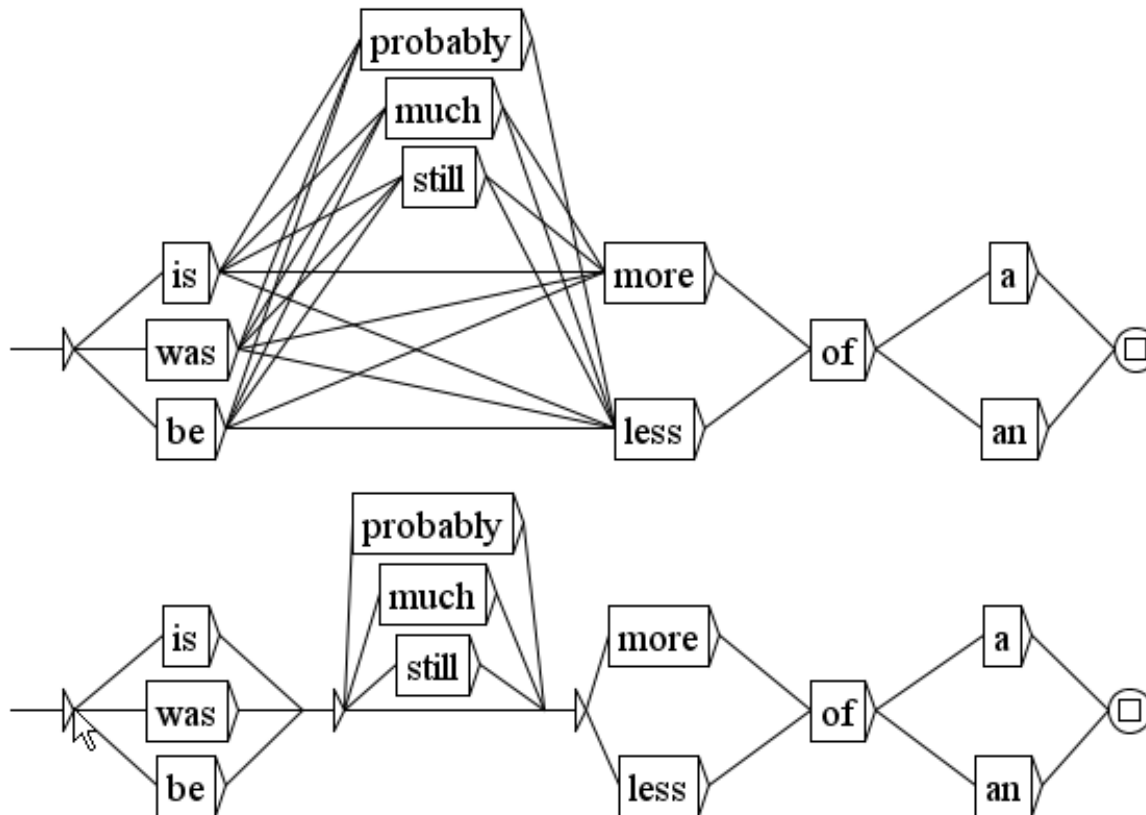
GROSS, Maurice. 2000. A Bootstrap Method for Constructing Local Grammars. *Contemporary Mathematics. Proceedings of the Symposium*. University of Belgrade, p. 231-249.

(<http://halshs.archives-ouvertes.fr/docs/00/27/83/19/PDF/BELG.pdf>)

PAUMIER, Sébastien. 2002. *Unitex. User Manual*. Revision: 2014. <http://igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf>

SILBERZTEIN, Max. 1993. *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*. 240 p., Paris : Masson.

Empty nodes

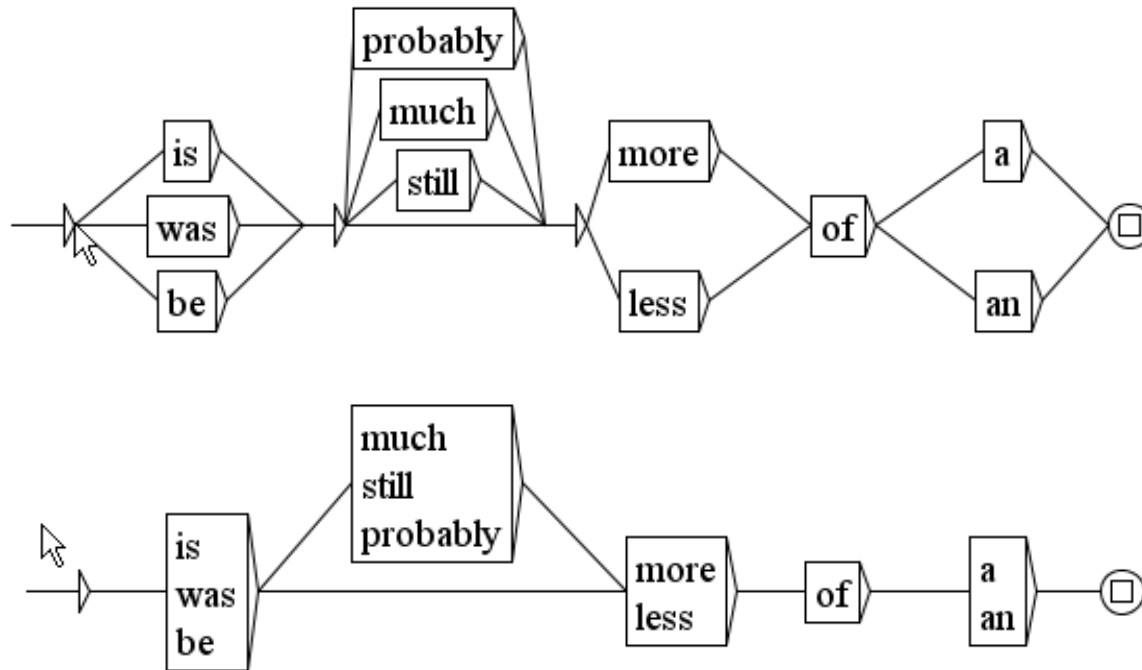


An empty node can serve as an intermediate step between two others



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Forms in parallel in a node



A graphical convention

Select the box, type the content of the first line, type "+"

Type the content of the next line, etc.

Validate with Enter



Sous la co-tutelle de :

CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Outline

Concordance

Local grammar of dates

Invoking a subgraph

Lexical masks

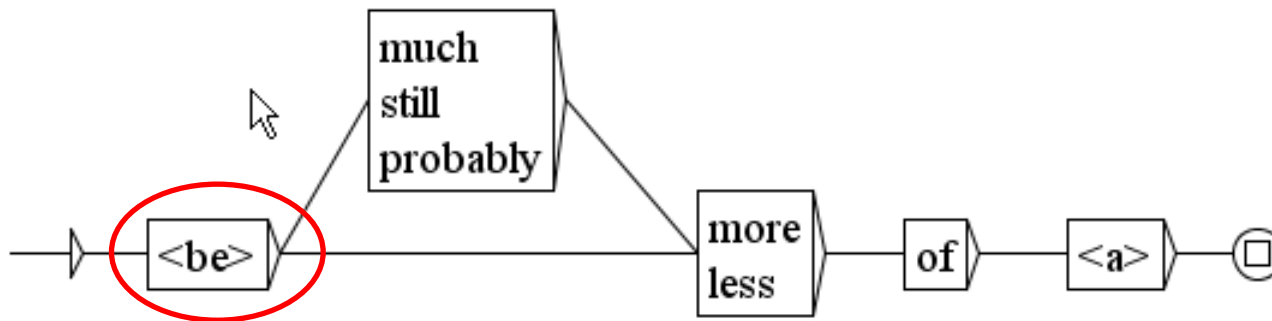
Dictionaries of a text



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Inflection and derivation

The same meaning conveyed by words with different suffixes
collect collects collected collecting collector collection
Increase the recall automatically





Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Inflection and derivation

Inflection

collect collects collected collecting

<collect> automatically includes the inflected forms

Derivation

collection collector collective collectivism

Automatically including derived forms would be more
adventurous



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Lexical masks

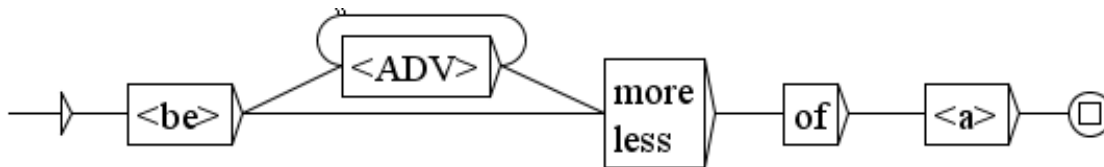
<collect> recognises automatically any inflected form of *collect*

<WORD> any simple word

<FIRST> any simple word with capitalized first letter

<NB> any contiguous sequence of digits

<V> any word marked as *V* in the lexicon (verbs)





Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Lexical masks

Key that recognises a set of linguistic forms on the basis of formal features

Can be used in graphs and regular expressions



Sous la co-tutelle de :

CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Outline

Concordance

Local grammar of dates

Invoking a subgraph

Lexical masks

Dictionaries of a text



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Dictionaries of a text

Lexical entries for the
simple words in the text

Lexical entries for the
multiword units
occurring in the text

Word Lists in D:\unitex3.1beta\English\Corpus\htjun94_snt

DLF: 40501 simple-word lexical entries

goalmouth, .N:s
goalpost, .N+Conc:s
goalposts, goalpost.N+Conc:p
goals, goal.N+Abst:p
goals, goal.N+Hum:p
goals, goal.V:P3s
goalscoring, .N:s
goaltender, .N:s
goat, .N+Anl:s
goatee, .N+Conc:s

DLC: 9262 compound lexical entries

go-slow, .A+z1
go-slow, .N+XN+z1:s
goal area, .N+XN+z1:s
goal line, .N+XN+z1:s
goal mouth, .N+XN+z1:s
God-fearing, .A+z1
going-away, .A+z1
going-on, .N+XN+NX+z1:s
Golan Heights, .N+XN+z1:p
Gold Coast, .N+XN+z1:s

ERR: 12455 un

☐ Filter unkr

Gluck
GLYNDEBOU
Glynnebou
Glynne
GM
GMBH
GmbH
GMT
Gnassingb
GNI
GNP
Goalllllll
Gobet
Godard
GOEJ
Goej
Goering
Goethe
Goey
Gogh
Gohr
Gogogoba



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Dictionaries of a text

Word Lists in D:\unitex3.1beta\English\Corpus\htjun94_snt

DLF: 40501 simple-word lexical entries

goalmouth, .N:s
goalpost, .N+Conc:s
goalposts, goalpost.N+Conc:p
goals, goal.N+Abst:p
goals, goal.N+Hum:p
goals, goal.V:P3s
goalscoring, .N:s
goaltender, .N:s
goat, .N+Anl:s
goatee, .N+Conc:s

DLC: 9262 compound lexical entries

go-slow, .A+z1
go-slow, .N+XN+z1:s
goal area, .N+XN+z1:s
goal line, .N+XN+z1:s
goal mouth, .N+XN+z1:s
God-fearing, .A+z1
going-away, .A+z1
going-on, .N+XN+NX+z1:s
Golan Heights, .N+XN+z1:p
Gold Coast, .N+XN+z1:s

ERR: 12455 unknown simple words

☐ Filter unknown words with tags.ind

Gluck
GLYNDEBOURNE
Glyndebourne
Glynn
GM
GMBH
GmbH
GMT
Gnassingb
GNI
GNP
Goal111111
Gobet
Godard
GOEJ
Goej
Goering
Goethe
Goey
Gogh
Gohr
Gogogobos

← The words in the text
not found in the
dictionaries



Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Thanks

CONTACT

ERIC LAPORTE

+33 1 60 95 75 52

ERIC.LAPORTE@UNIV-PARIS-EST.FR