LABORATOIRE D'INFORMATIQUE GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

2014/09/01
Workshop on Finite-State Language Resources
Sofia

Local Grammars 1

Éric Laporte

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

Concordance

Local grammar of dates

Invoking a subgraph

Lexical masks

Dictionaries of a text

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

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 Da
forces is one of the major problems which the army

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Concordance: diversity

ters, 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} Lieutenant	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,

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Concordance: sequences

Text:{S} A senior	government official	charged Wednesda
Abel{S} Olopai, a	government official	who was 10 years
regrets," a German	government official	said.{S} Key Mor
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 Rw
ollapse in 1991, a	government official	said Sunday.{S}
pens next month, a	government official	said.{S} Until t
ound," said a U.S.	government official	.{S} He said Wash
WILLIAM J. (U.S.	GOVERNMENT OFFICIAL)	{S} Internationa
of the century, a	government official	said.{S} "We wan

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

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

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

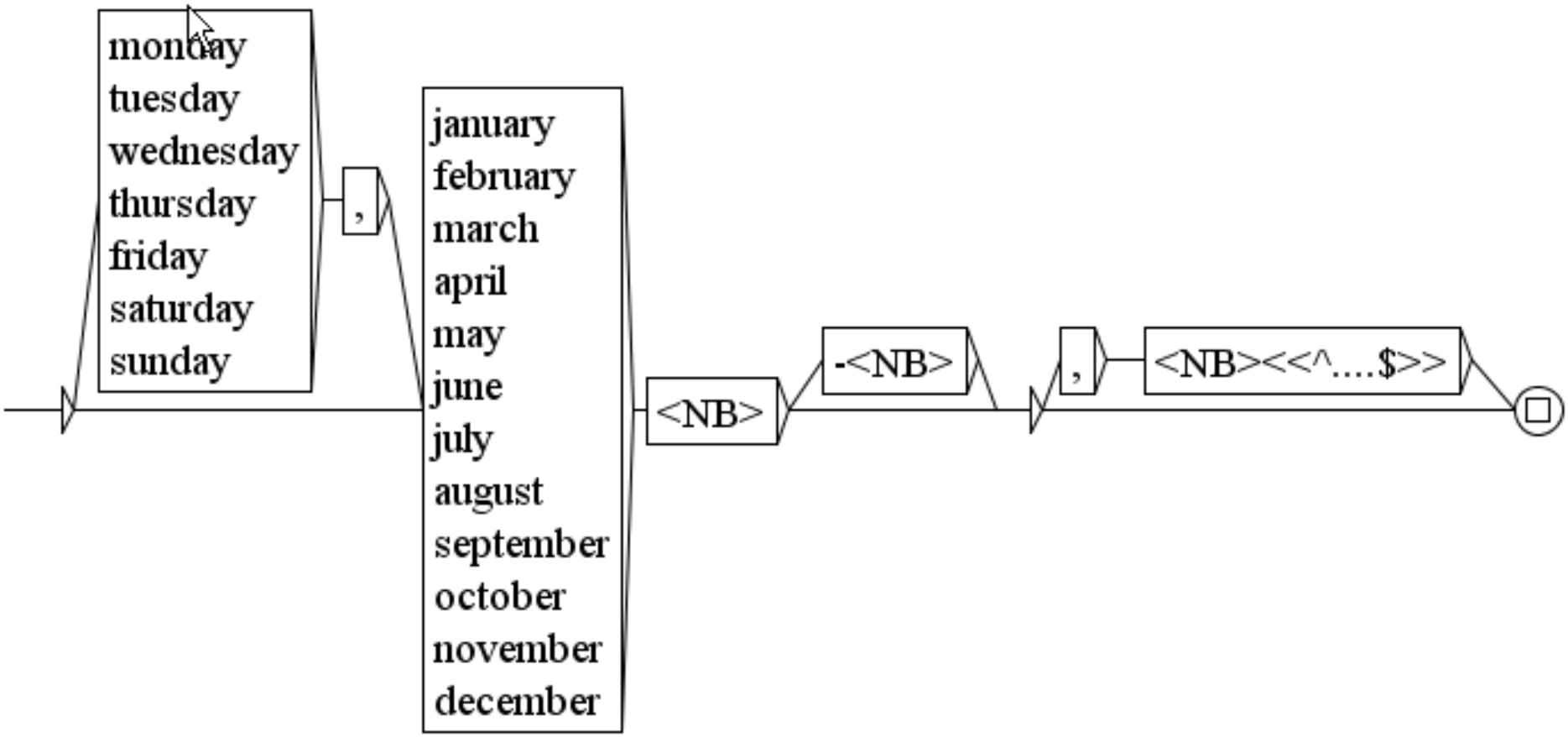
Dates in English

t Hong Kong after the July 1, 1997, handover to Chinese
ole candidate for the April 1995 presidential poll was
up of drivers for the July 3 race would be made until t
asily blurred, as the June 6 anniversary showed (even t
as been as bad as the April 26 crash of the Taiwan-base
which organizes the June 26 regional elections, said
chairmen have set the July 4 recess as the deadline to
airlines have put the July 4th weekend on sale for trav
been mandated by the June 26 election, the Group of Se
worked, helping make June 6, 1944, the biggest D-Day o
began a hunger strike May 25 after two of them were fir

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

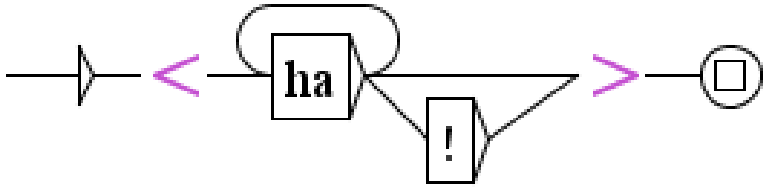
Query for dates in English



LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Finite automata and regular expressions



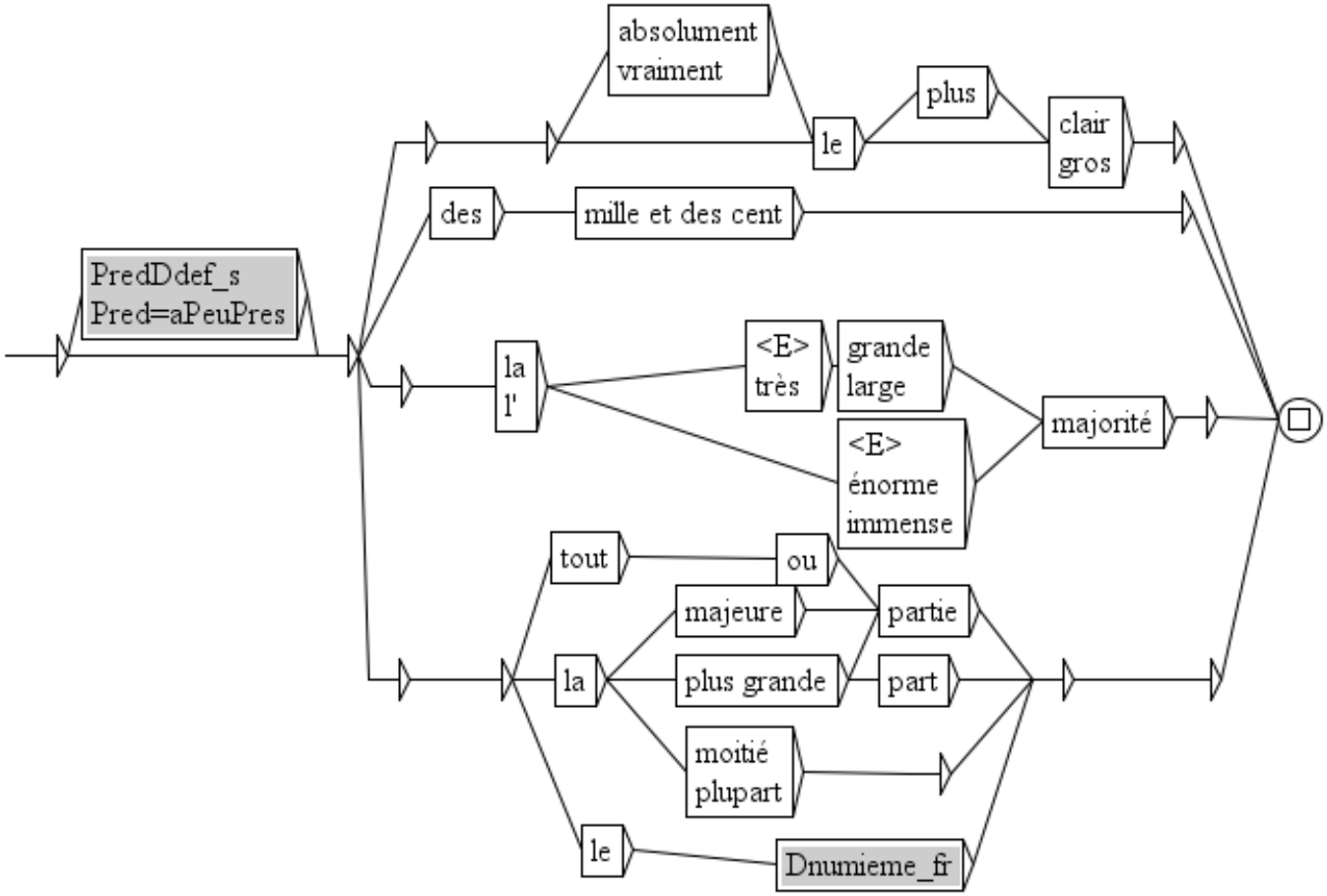
Equivalent regular expression:

ha(*ha*)**!*?

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Imagine the regular expression for this one

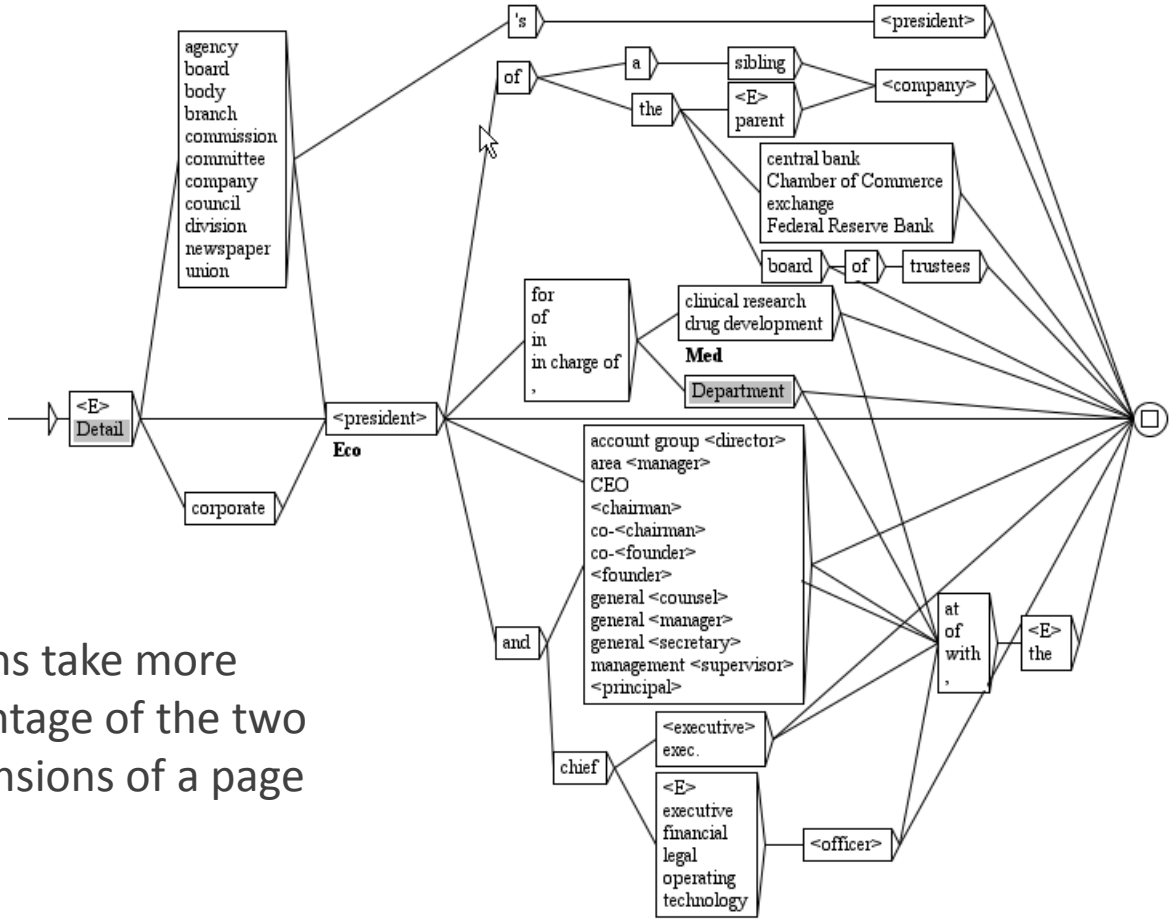


Source: Éric Laporte, 2006

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Now for this one



Graphs take more advantage of the two dimensions of a page

Source: Maurice Gross

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Beyond concordances

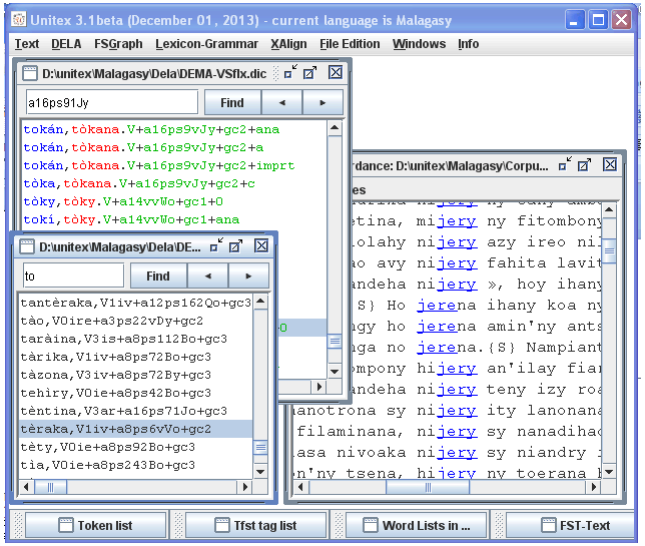
Such graphs are also used in NLP applications:

- indexation
- information extraction
- annotation

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

How to edit a graph with Unitex-Gramlab



An open-source corpus processor based on language resources
- generation of concordances
- automatic annotation
- dictionary management
- dictionary-based morphological analysis
- graphical grammar editor

Now 22 languages

Runs on Linux, Windows, OS X

Paumier (2002-2014)

http://igm.univ-mlv.fr/~unitex

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

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

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

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

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

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

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

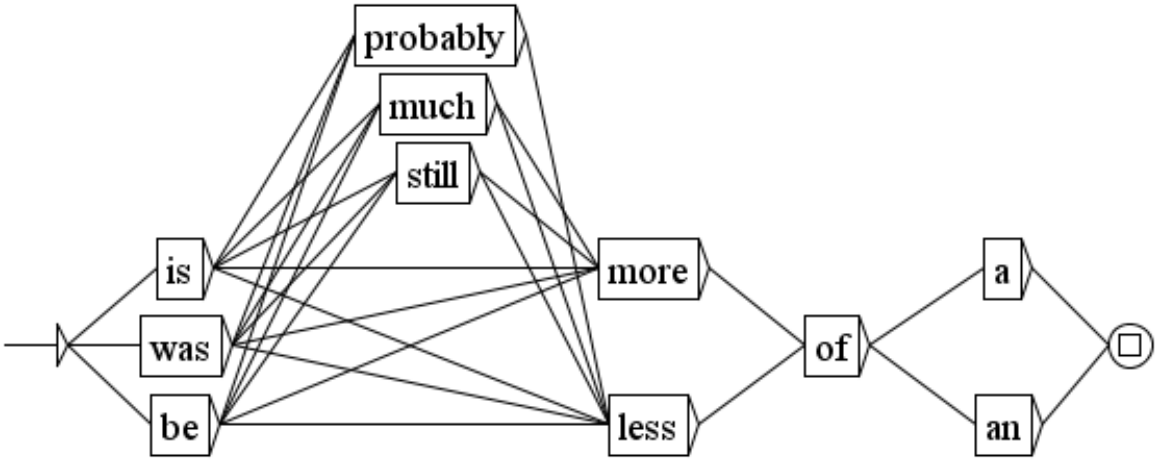
Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

A graph

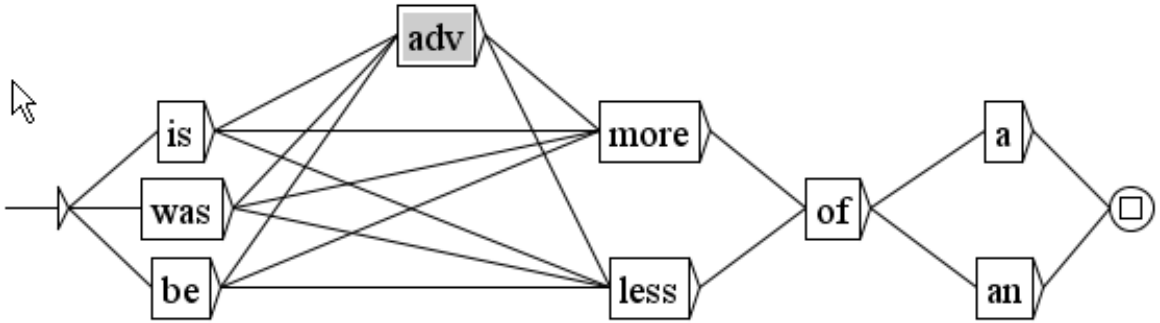


.l rising, it is probably less of a problem politically than it was
The new guru was less of a showman and more of a technician who be
"and there is much less of a feeling of impending crisis here th
said, "There is less of a sense among juries that it is other peop
sun proved to be more of a menace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

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

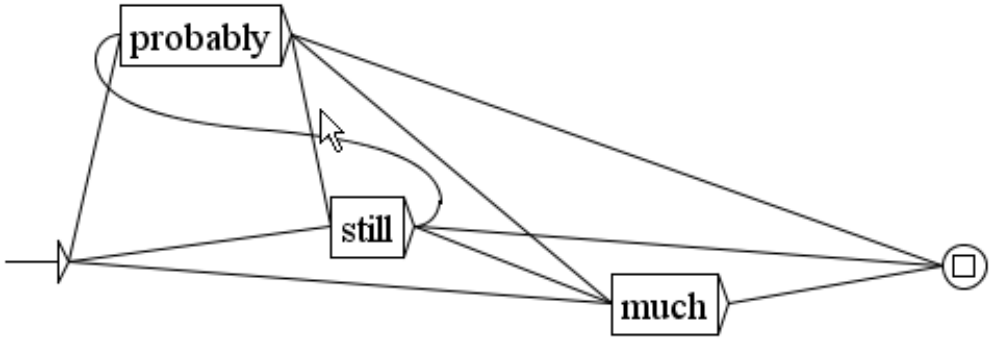
Invoking a subgraph from a node



The adv node invokes the adv.grf graph

Equivalent to substituting the graph for the node

adv.grf:



LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Invoking a subgraph from a node

Objectives

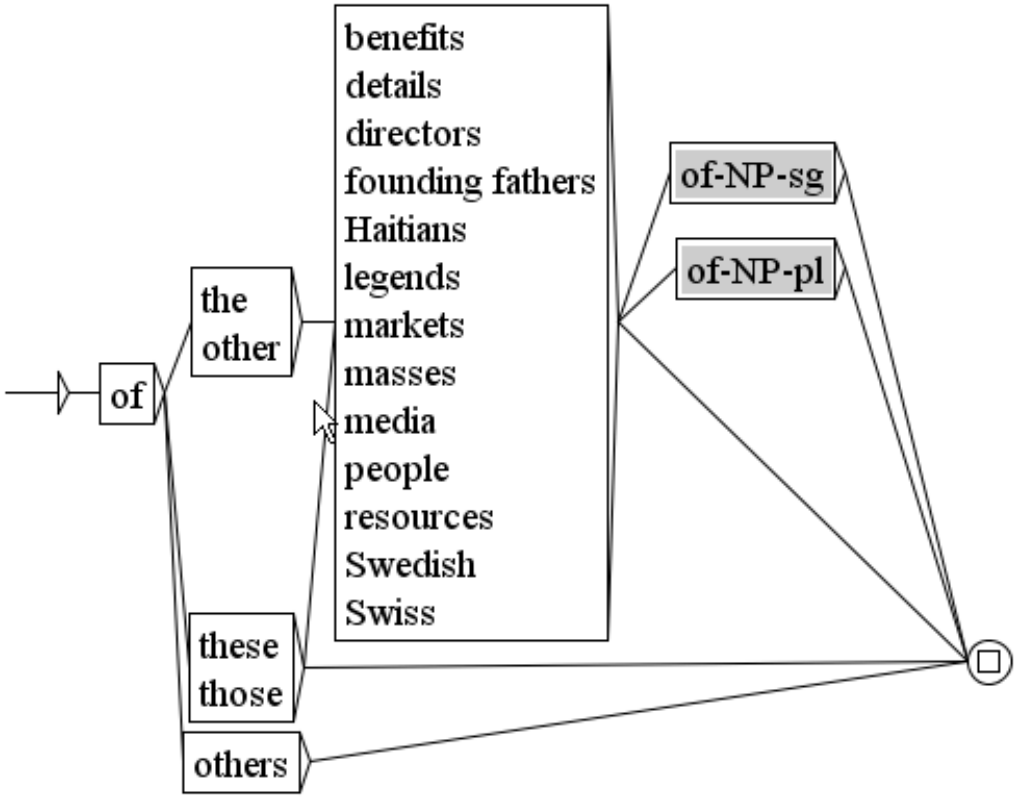
Use the same graph in various contexts

Distribute a description in several graphs

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Invoking a subgraph from a node



A graph can invoke itself

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

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

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)

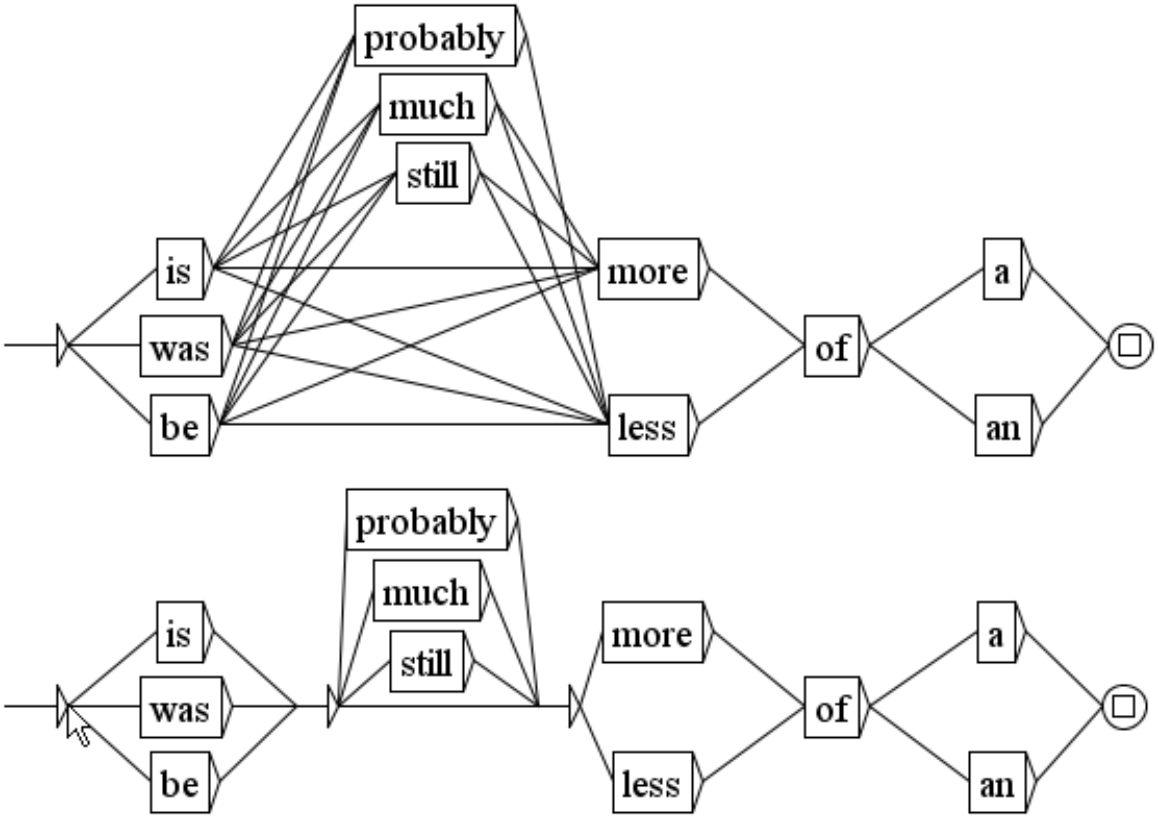
GROSS, Maurice. 2000. A Bootstrap Method for Constructing Local Grammars. *Contemporary Mathematics. Proceedings of the Symposium*. University of Belgrade, p. 231-249.

(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

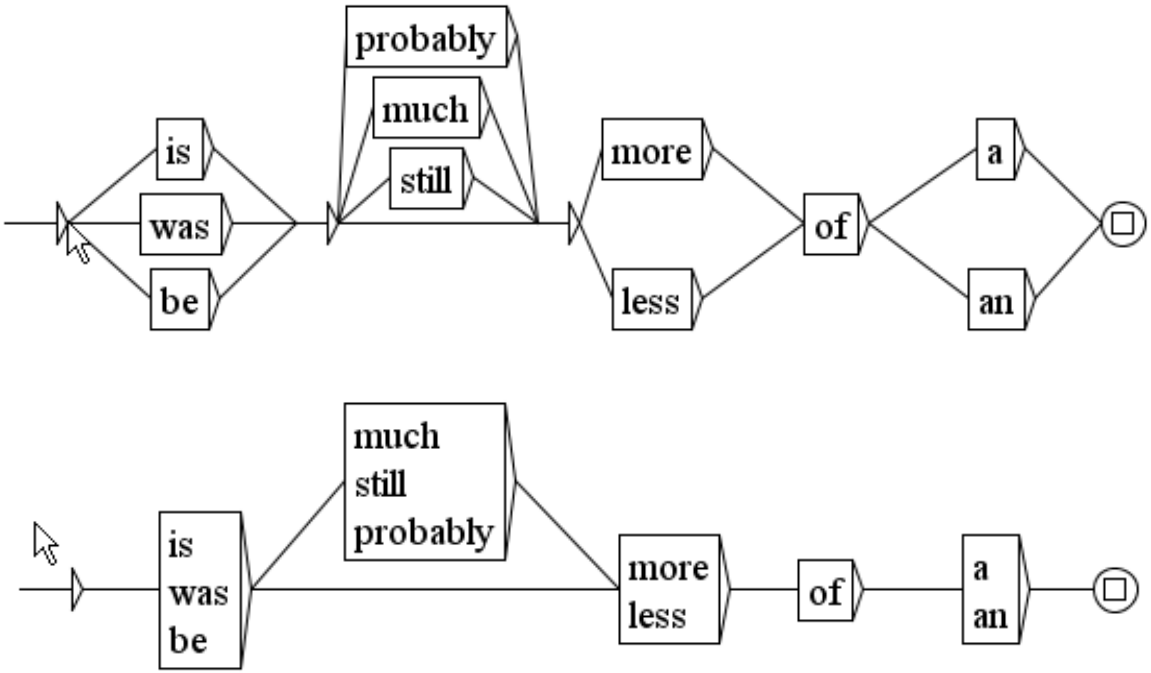


An empty node can serve as an intermediate step between two others

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

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

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

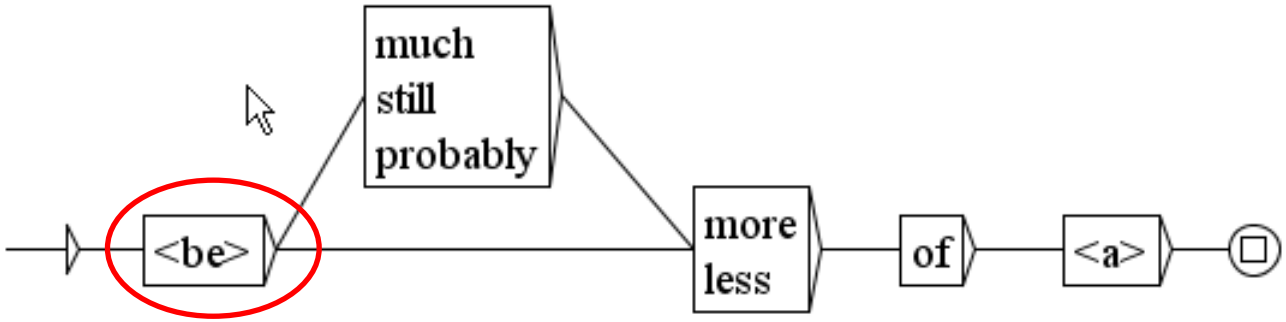
Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Inflection and derivation

The same meaning conveyed by words with different suffixes

collect collects collected collecting collector collection

Increase the recall automatically



LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

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

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Lexical masks

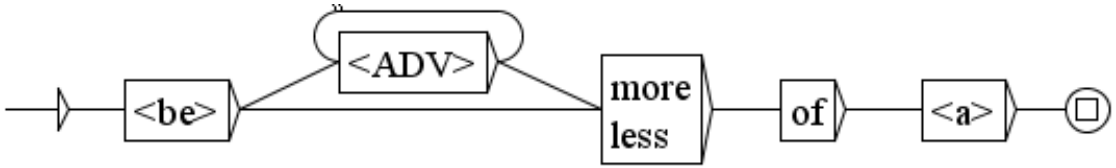
<collect> recognises automatically any inflected form of *collect*

<MOT> any simple word

<PRE> any simple word with capitalized first letter

<NB> any contiguous sequence of digits

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



LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Lexical masks

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

Can be used in graphs and regular expressions

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

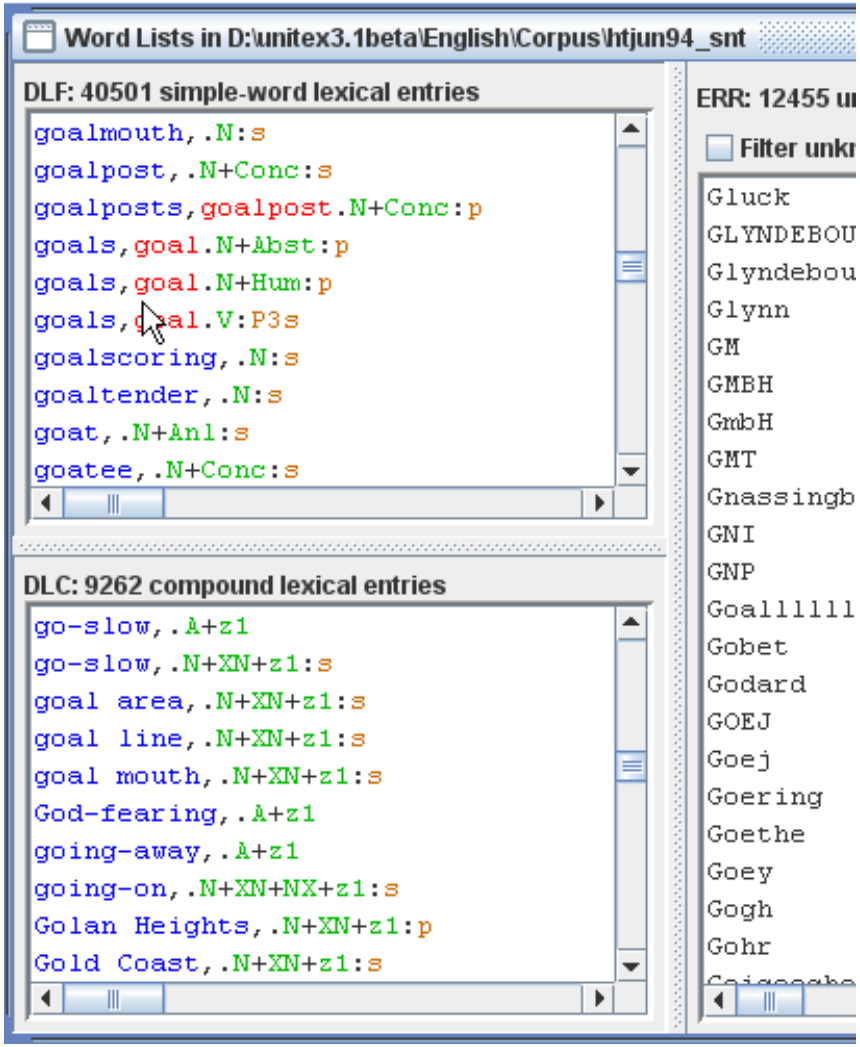
LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Dictionaries of a text

Lexical entries for the simple words in the text

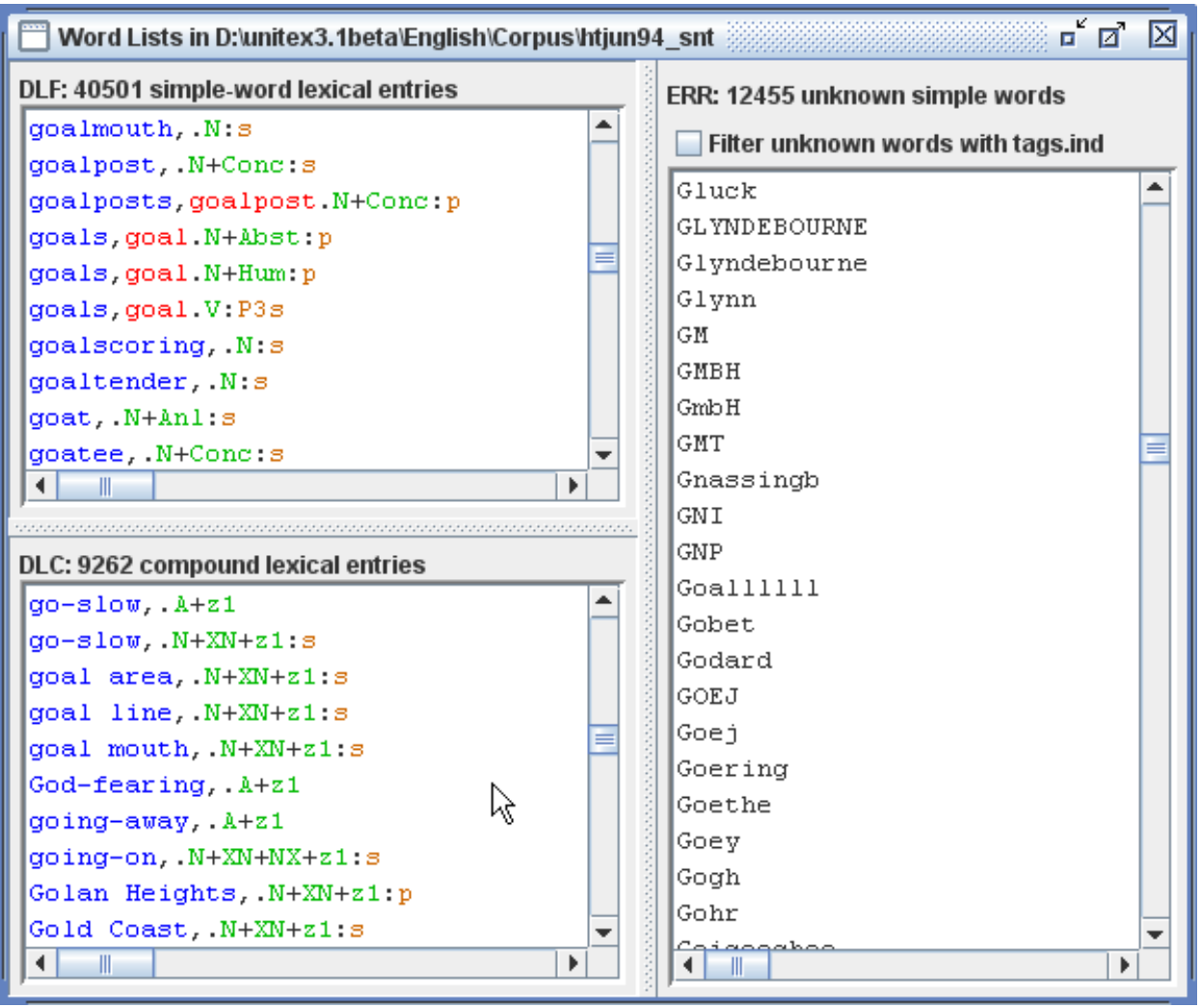
Lexical entries for the multiword units occurring in the text



LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Dictionaries of a text



← The words in the text not found in the dictionaries

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Thanks

CONTACT

ÉRIC LAPORTE

00 +33 (0)1 60 95 75 52

ERIC.LAPORTE@UNIV-PARIS-EST.FR