LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

2014/09/02
Workshop on Finite-State
Language Resources
Sofia

Acyclic automaton of a text

Éric Laporte

Outline

Word lattices

Lexical analysis with several solutions

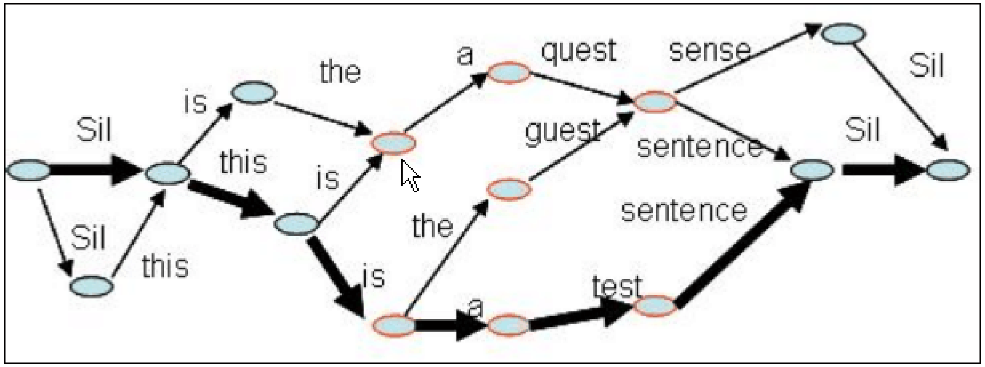
Information retrieval

Hybrid tagging

Agglutinative languages

More contextual constraints

Word lattice



A string of words and a finite set of variants

Speech recognition

Terminology

Lattice

Other mathematical meaning in order theory

Ordered set where each pair has a sup and an inf

Acyclic automaton

Automata theory

Determinization and minimization of finite automata

My choice

Directed acyclic word graph

Formal language theory

Directed acyclic graph

Graph theory

Trellis

Information theory

Seldom used

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

Word lattices

Lexical analysis with several solutions

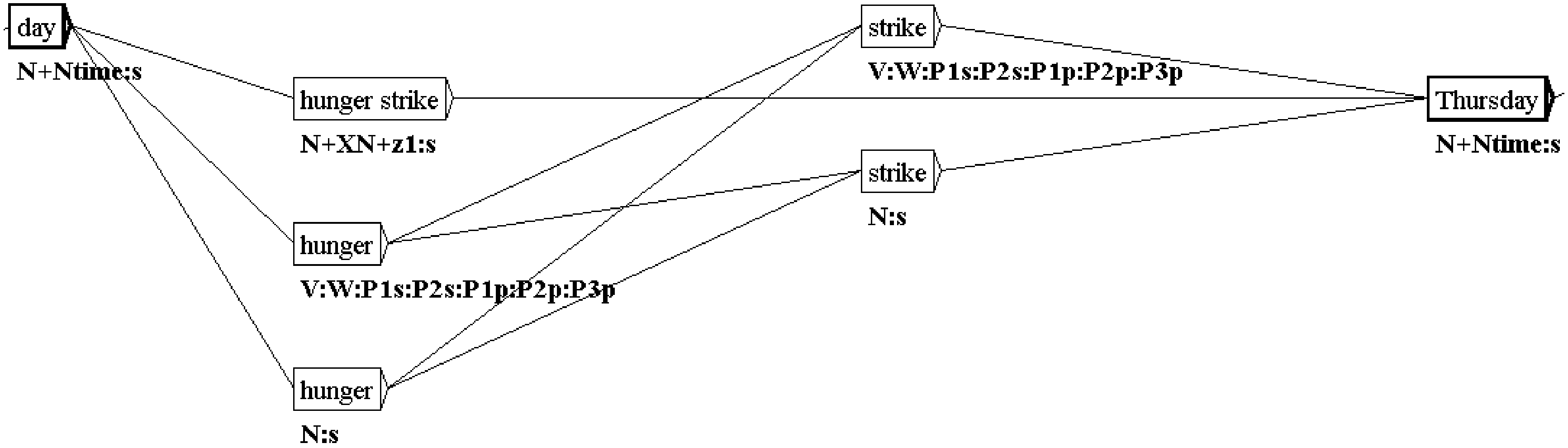
Information retrieval

Hybrid tagging

Agglutinative languages

More contextual constraints

Lexical analysis with several solutions



... her husband would start a two-day hunger strike Thursday...

Multiword units

hunger strike

Lexical ambiguity

These stories have left me hungering for more
Urbanisation drives hunger for luxury goods

Several countries are stricken by the epidemic
There is a strike

Lexical analysis with several solutions

A space-efficient storage structure

Number of lexical tags per word: *a*

Number of words: *n*

Number of analyses: about a^n

Number of transitions: *an*

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

Word lattices

Lexical analysis with several solutions

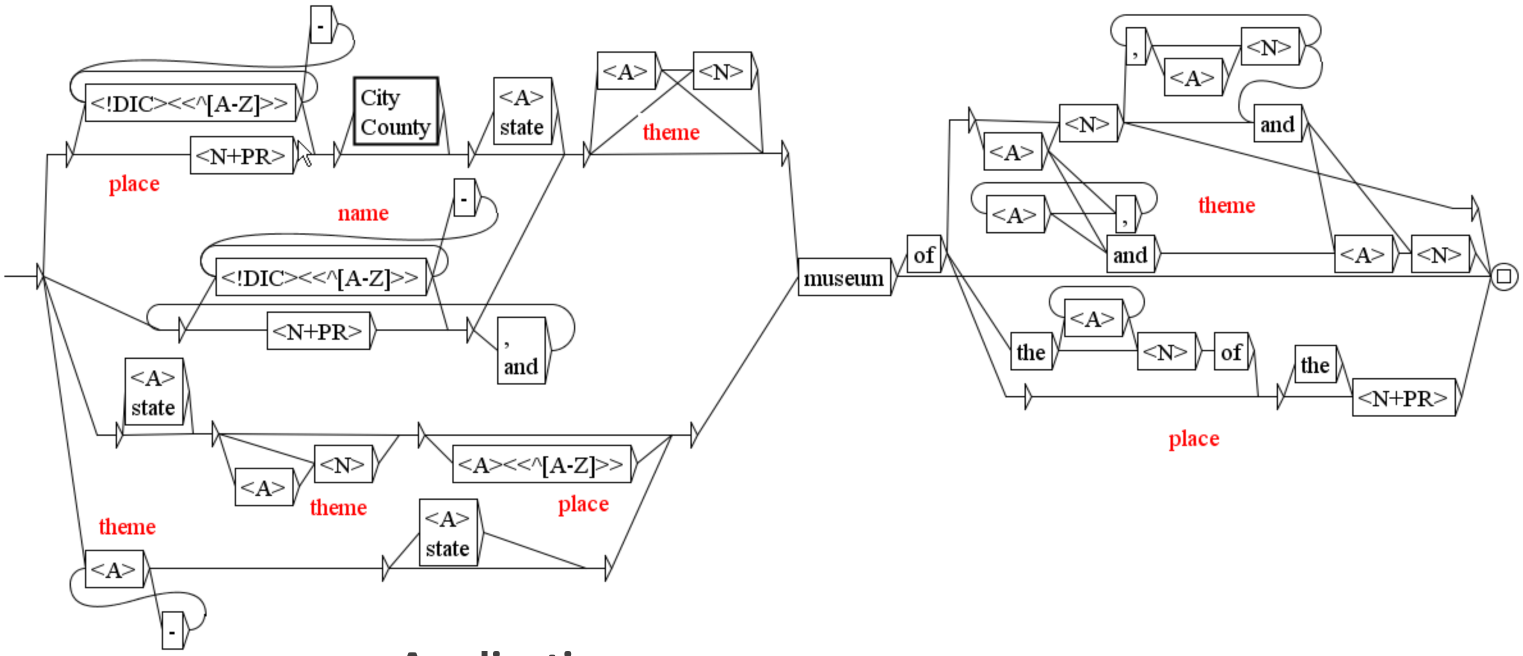
Information retrieval

Hybrid tagging

Agglutinative languages

More contextual constraints

Use with local grammars



Applications

Information retrieval and extraction

Indexing

Example: extraction of names of museums

Lexical masks are matched with dictionary-based tags of words

Use with local grammars

Acyclic automaton compared with uniquely tagged text

Higher recall

The correct lexical analysis of the text is present in the acyclic automaton more often than in uniquely tagged text

Lower precision

Paths parallel to the correct analysis may match with the local grammar

... can now be seen, heard, and even touched in a museum that was opened here...

touched,.A	*This guy here is a little touched*
in,.A	*This is really in now*
a,.N	Woman *is spelt with an a*

Use with local grammars

Lower precision

Paths parallel to the correct analysis may match with the local grammar

This effect is limited (Fairon *et al.*, 2005)

Local grammar paths usually have at least 5 words

Matches with parallel paths are usually partially correct

Syntactic parsing

Similar situation

Cédrick Fairon, Sébastien Paumier, and Patrick Watrin. 2005. Can we parse without tagging ? In Zygmunt Vetulani (ed.), *Language & Technology Conference (LTC)*, pp. 473–477.

Outline

Word lattices

Lexical analysis with several solutions

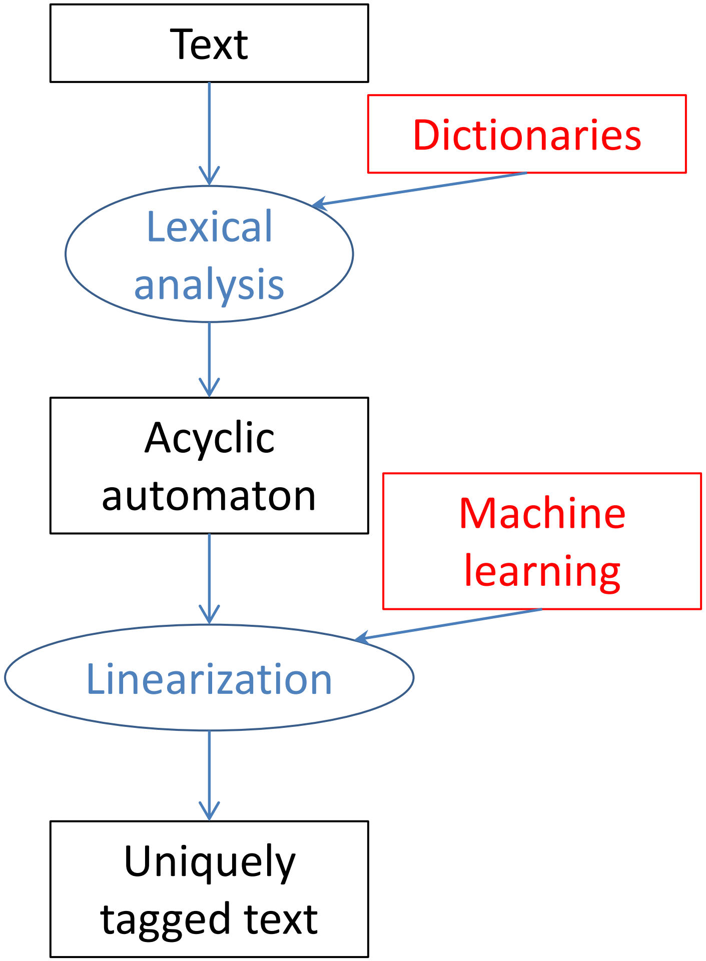
Information retrieval

Hybrid tagging

Agglutinative languages

More contextual constraints

Use for hybrid one-solution tagging



Contribution of dictionaries

Good-quality dictionaries are rich in **multiword expressions** and **rare uses of words** (annotated corpora have data sparseness): *hunger, touched*

Good-quality dictionaries provide information on **words not found in the corpus** (more reliable than guessing methods)

Contribution of supervised tagging

Select likely analyses

Sigogne, 2010

Outline

Word lattices

Lexical analysis with several solutions

Information retrieval

Hybrid tagging

Agglutinative languages

More contextual constraints

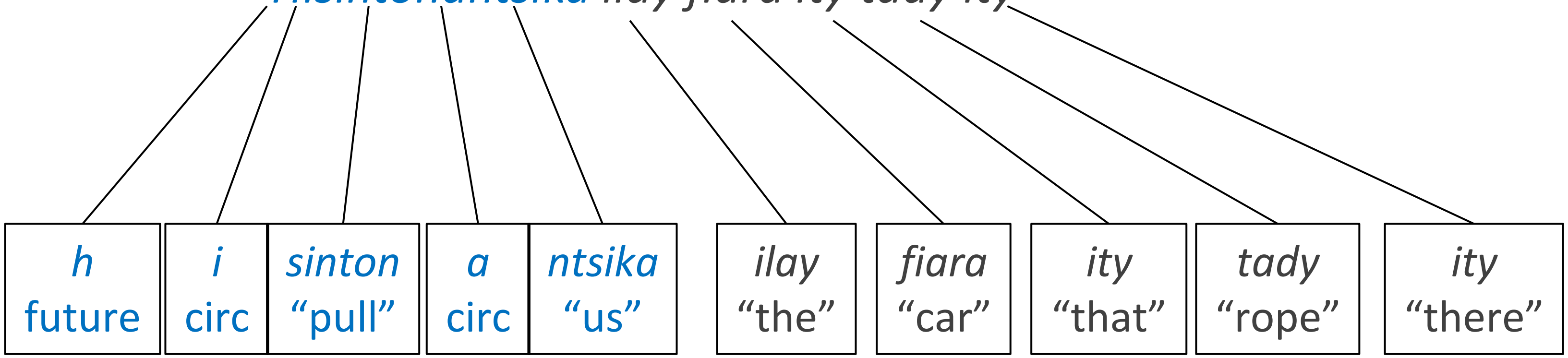
Agglutinative languages

In agglutinative languages, a word is often a sequence of morphemes with separate functions or meanings

Example: Malagasy

“That rope is what we will pull the car with”

Hisintonantsika ilay fiara ity tady ity



h	*i*	*sinton*	*a*	*ntsika*	*ilay*	*fiara*	*ity*	*tady*	*ity*
future	circ	“pull”	circ	“us”	“the”	“car”	“that”	“rope”	“there”

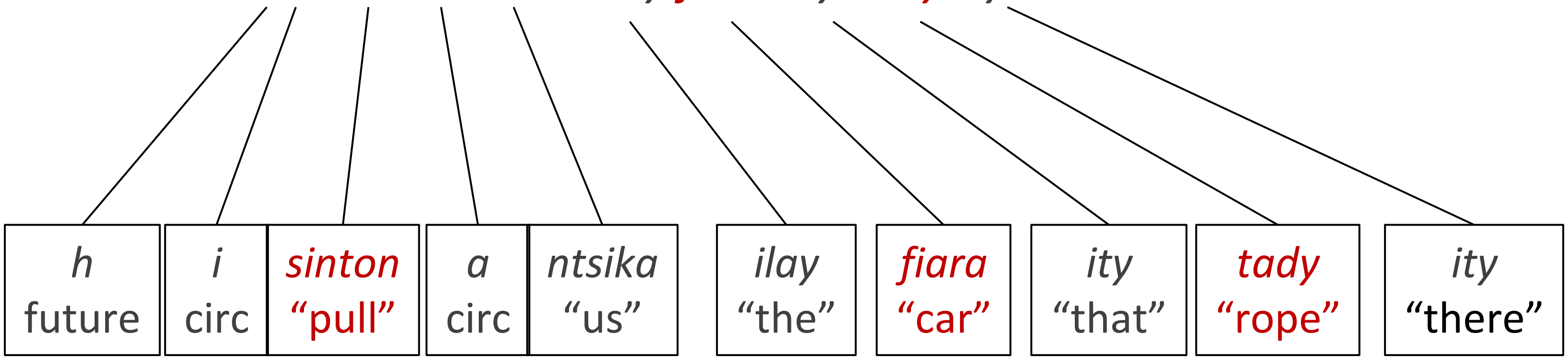
Source: Ranaivoarison *et al.*, 2013

Words behave this way for derivation, inflection and part of syntax

Agglutinative languages

"That rope is what we will pull the car with"

Hisintonantsika ilay fiara ity tady ity



h	*i*	*sinton*	*a*	*ntsika*	*ilay*	*fiara*	*ity*	*tady*	*ity*
future	circ	"pull"	circ	"us"	"the"	"car"	"that"	"rope"	"there"

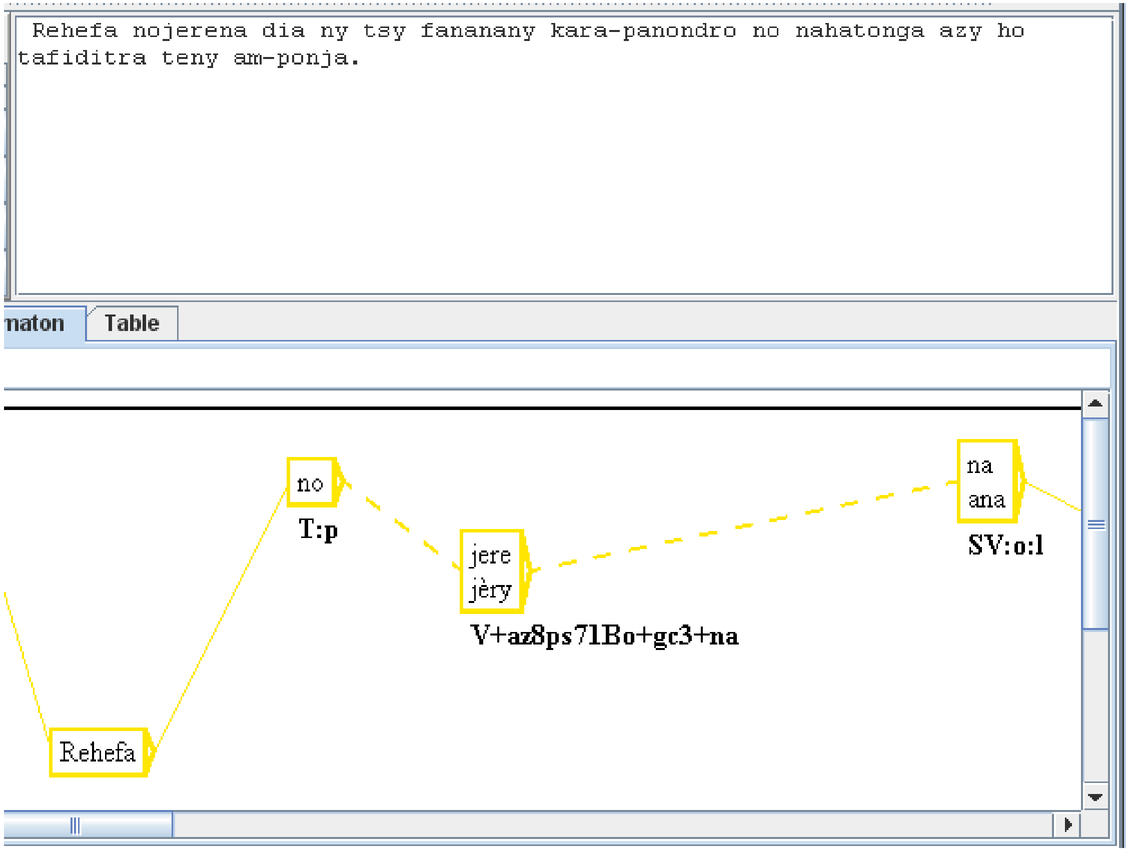
In most agglutinative languages, morphemes inside a word are not graphically delimited

Language processing requires delimiting meaningful units

Morphological analysis

Morphological dictionary-graphs

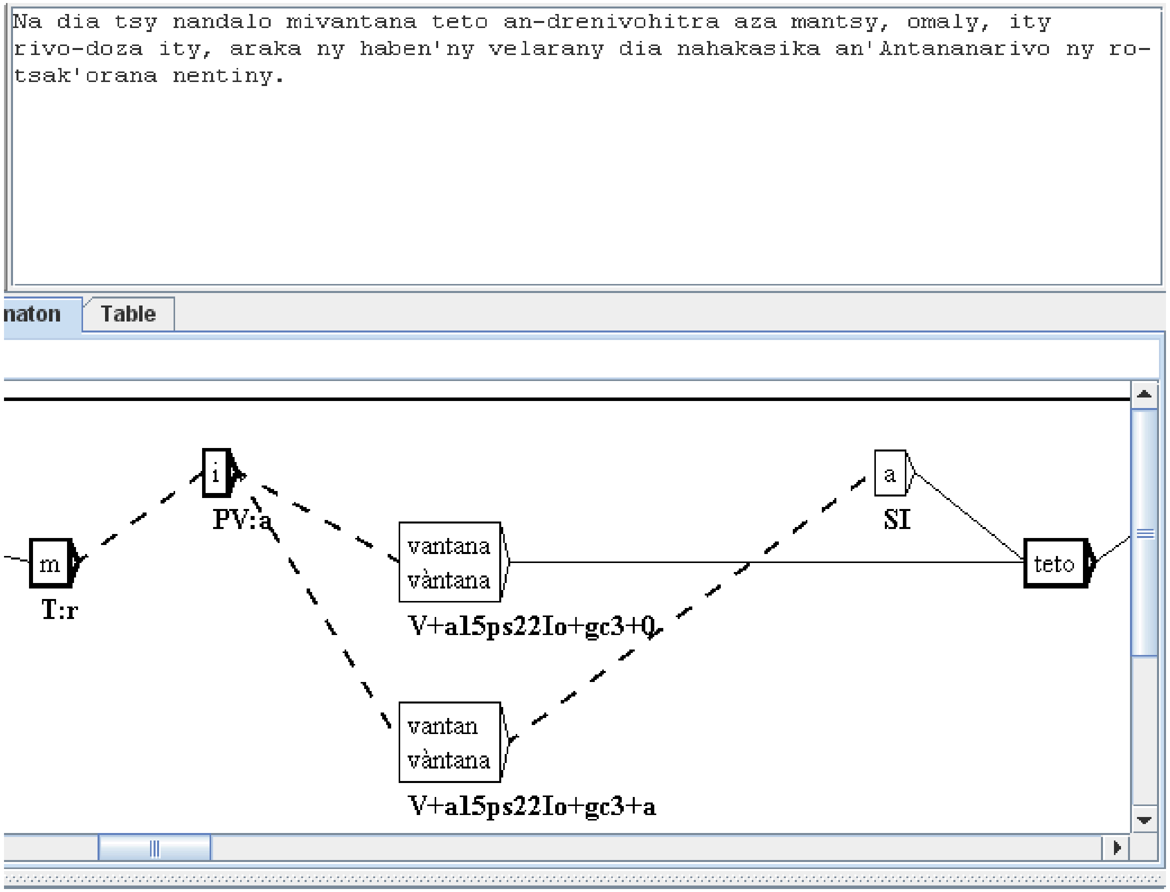
Agglutinative languages



```
 Rehefa nojerena dia ny tsy fananany kara-panondro no nahatonga azy ho
tafiditra teny am-ponja.
```

Analysis of *nojerena* "has been watched": one solution

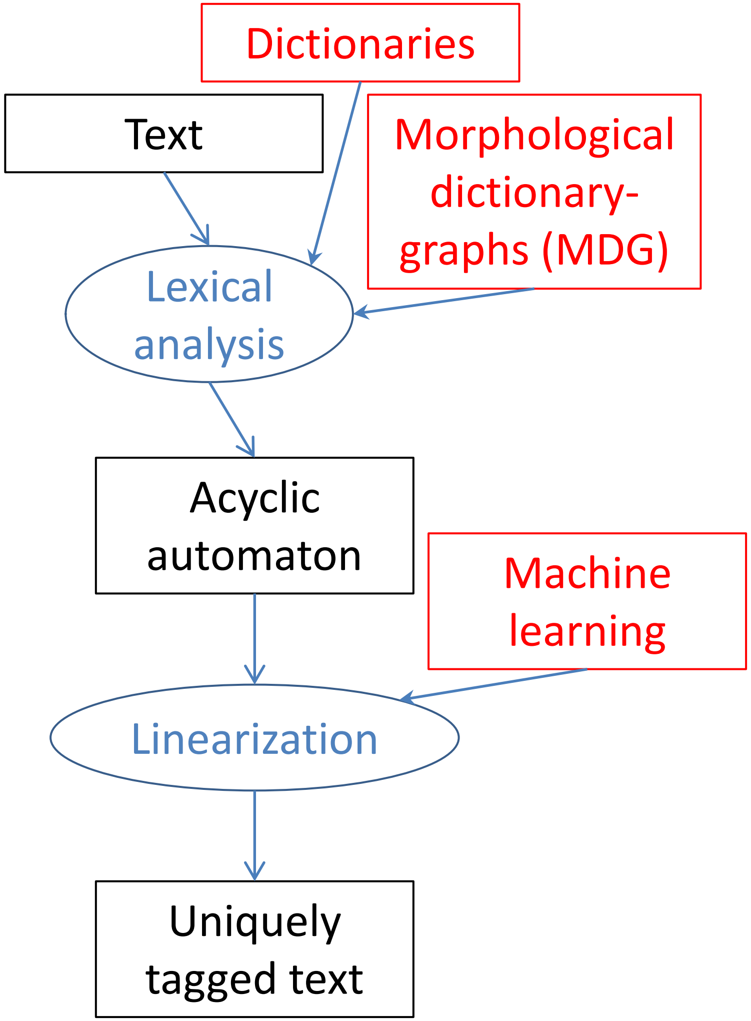
Agglutinative languages



```
Na dia tsy nandalo mivantana teto an-drenivohitra aza mantsy, omaly, ity
rivo-doza ity, araka ny haben'ny velarany dia nahakasika an'Antananarivo ny ro-
tsak'orana nentiny.
```

Analysis of an ambiguous form, *mivantana* “go direct to”: two solutions

Use for hybrid, one-solution morphological analysis



Dictionaries

Text

Morphological dictionary-graphs (MDG)

Lexical analysis

Acyclic automaton

Machine learning

Linearization

Uniquely tagged text

Contribution of dictionaries and graphs

Dictionaries provide accurate information on morphological variations: *jery, jere*

MDGs describe restrictions on morpheme combinations

This includes **rare uses of words** (annotated corpora have data sparseness) and **words not found in the corpus** (more reliable than guessing methods)

Contribution of supervised tagging

Select likely analyses

Outline

Word lattices

Lexical analysis with several solutions

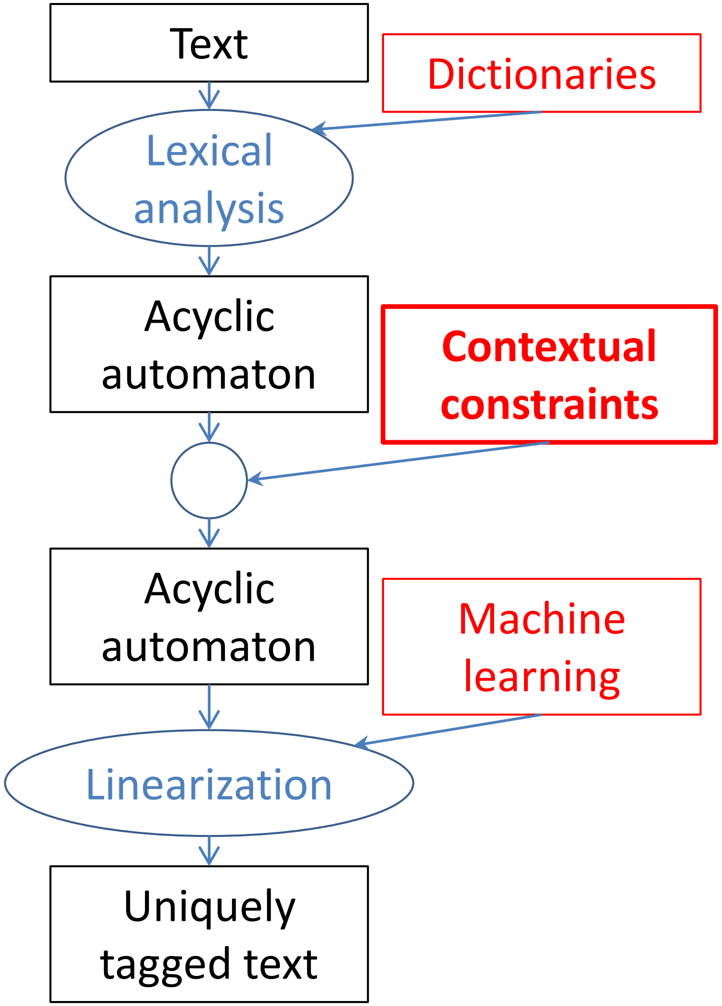
Information retrieval

Hybrid tagging

Agglutinative languages

More contextual constraints

Describing more contextual constraints with Unitex



Enhance the contribution of the symbolic approach to hybrid one-solution tagging

Describe combinatorial constraints between words

Remove analyses from the acyclic automaton

Describing more contextual constraints with Unitex

Unitex has 2 versions of the acyclic automaton of the text
Contents may be different
The updatable version allows for removing analyses

	read-only version	updatable version
Graphical display	no	Text > Construct FST-Text menu
Update after dictionary application	no	Elag program or manually
Search	Locate program (Paumier, 2003)	LocateTfst program, slower
Available with Gramlab	yes	no
Affected by MDGs	no	yes

Two types of contextual constraints

Lax constraints

At the beginning of a sentence, a subject personal pronoun is often followed by a verb

We smile for pictures

A counter-example in the type of text to be processed

We usually smile for pictures

An *<A><N>* analysis is more likely than an *<A><A>* analysis

...the current round of food shortages...

but:

...the fugitive German real-estate tycoon...

Symbolic grammars might not be a good choice for checking plausibility and preferences

Two types of contextual constraints

Strict constraints

At the beginning of a sentence, if a subject personal pronoun is followed by a verb in the present or preterit, they agree in person and number

We smile for pictures

Strongly consistent with the type of text to be processed

**We is who we is*

Few strict constraints

- in free-word-order languages
- in very informal styles

Symbolic grammars are appropriate for strict constraints

Describing more contextual constraints with Unitex

Elag

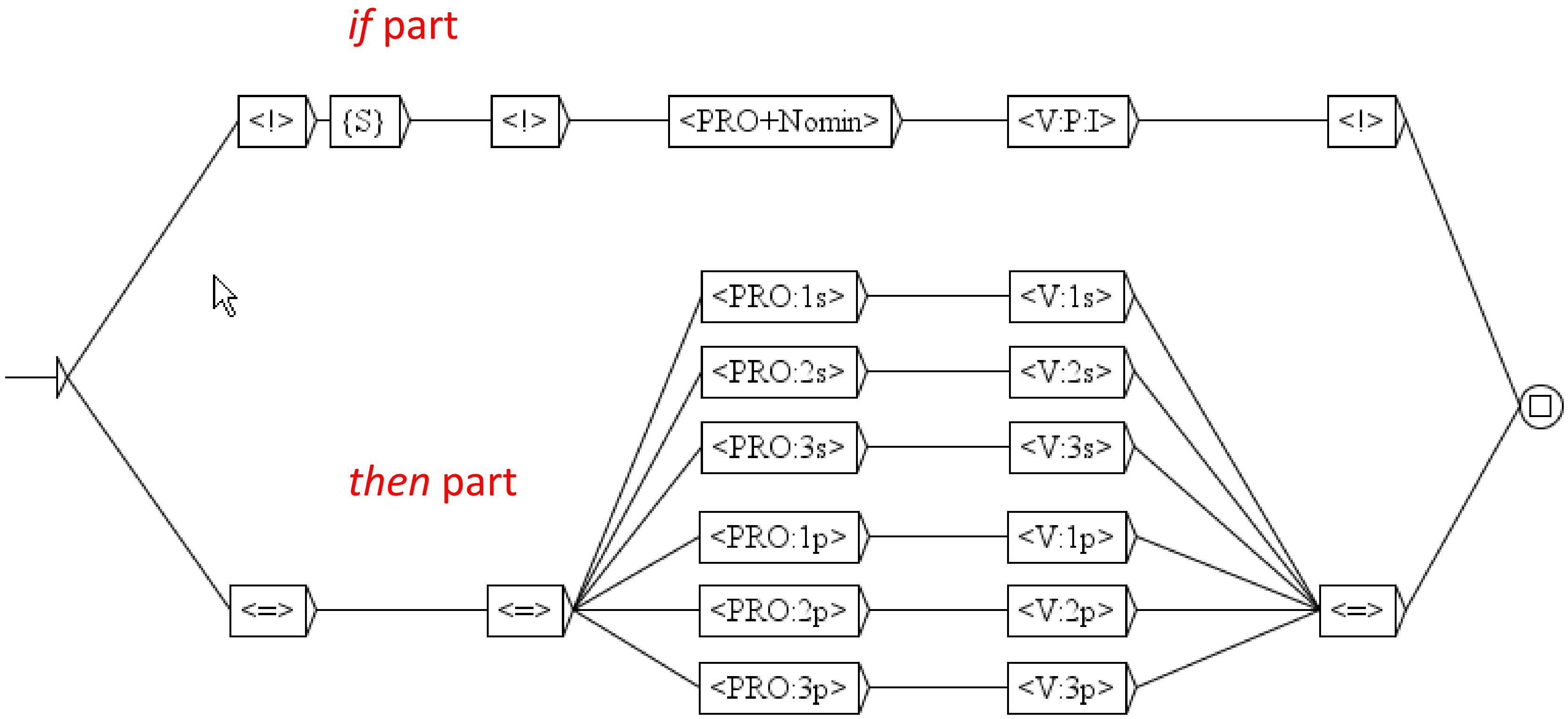
Symbolic description of contextual constraints

Focuses on strict constraints

Unitex-compatible

Éric Laporte, Anne Monceaux, 1999. Elimination of lexical ambiguities by grammars. The ELAG system, *Lingvisticae Investigationes* XXII, pp. 341-367.

Describing more contextual constraints with Elag



if part

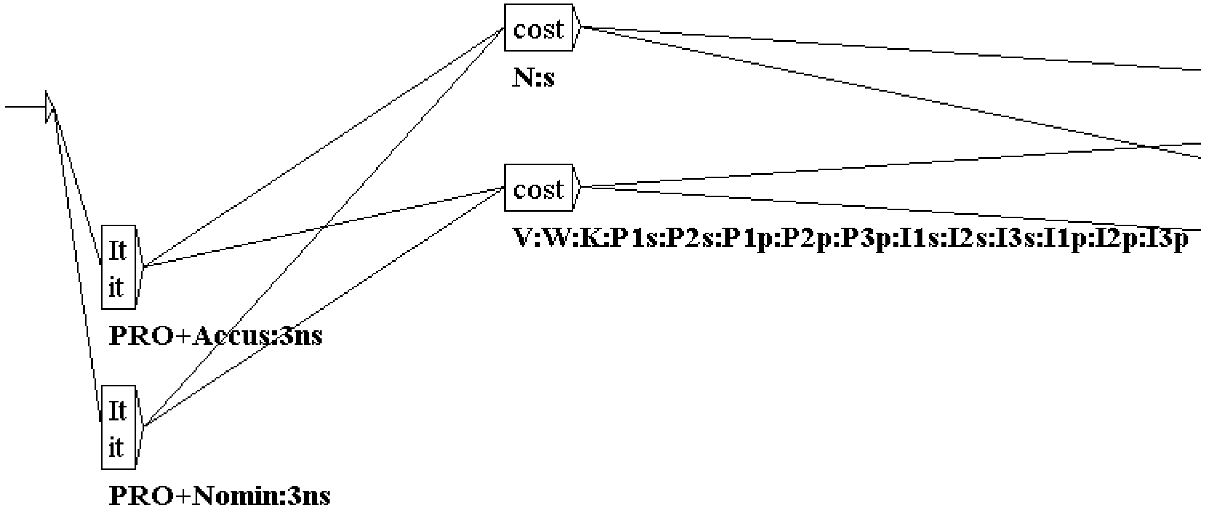
then part

At the beginning of a sentence, if a subject personal pronoun is followed by a verb in the present or preterit, they agree in person and number

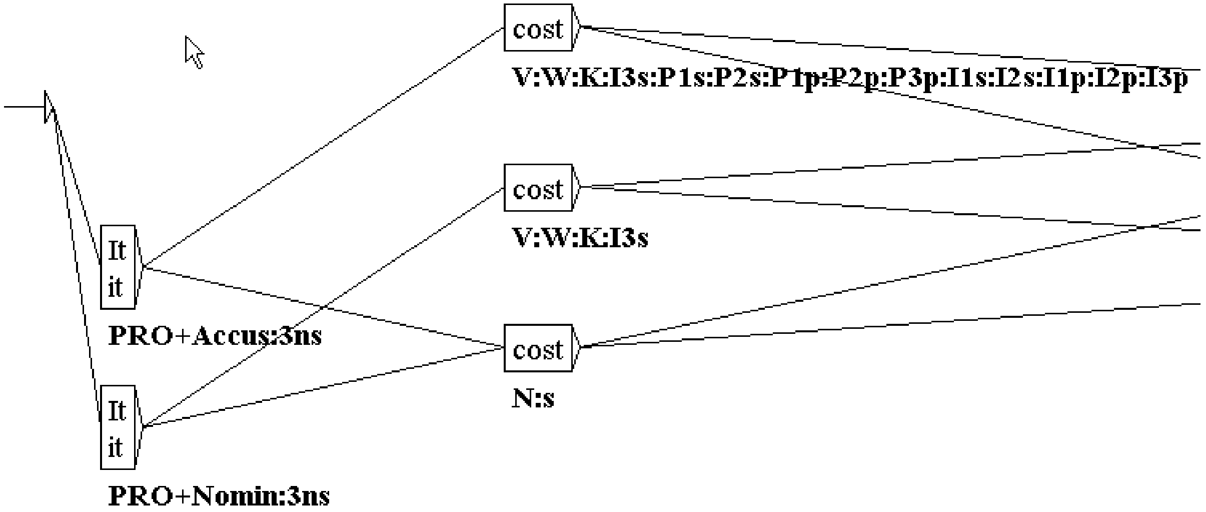
Result on a text

*It cost some exercise of
the white truncheon...*

before



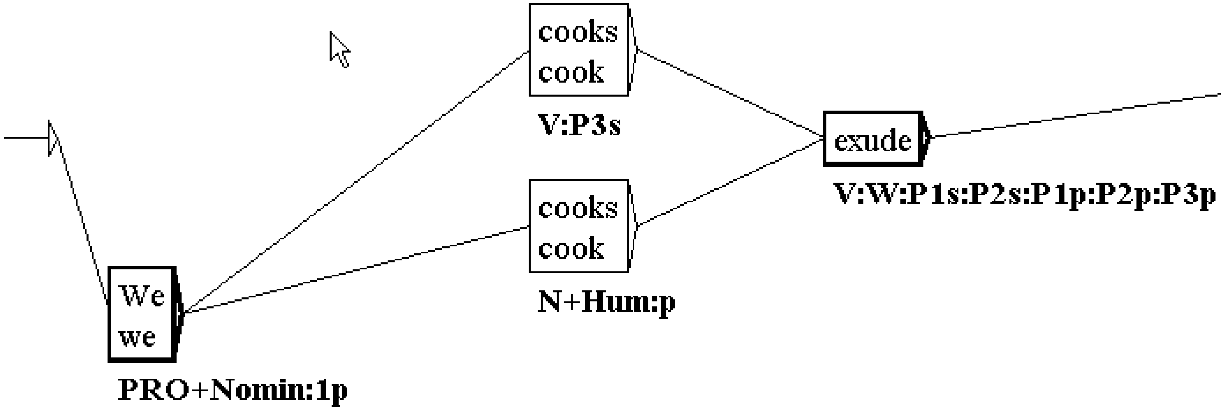
after



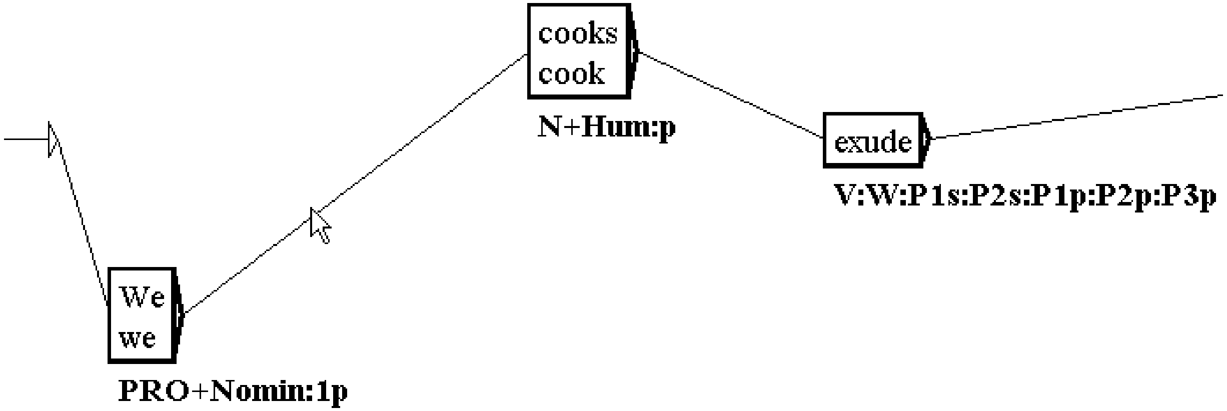
Result on a text

We cooks exude & expend a lot of energy during service

before



after



Describing more contextual constraints with Elag

A specificity of Elag

An analysis can be removed from the acyclic automaton
- on the basis of its own characteristics only
- independently of any parallel analyses

Motivation

Strict constraints on an analysis are unlikely to take into account any characteristics of another

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