# Conceptual Vectors, Lexical Networks, Morphosyntactic Trees and Ants : a bestiary for Semantic Analysis

Mathieu Lafourcade

# Conceptual Vectors, Lexical Networks, Morphosyntactic Trees, and Ants

## a bestiary for Semantic Analysis


Mathieu LAFOURCADE

LIRMM-INFO
Equipe TALN

# Outline

| | |
|---|---|
| Semantic Analysis | What to look at? – Applications |
| Casting | Conceptual Vectors – Analysis Trees – Lexical Network |
| Ant algorithm | Principles - Without lexical network - With lexical network |
| Ressource Production | Lexical Network - Conceptual Vectors |
| Future directions | Toward holistic analysis: syntax + semantic |

# Outline

| | |
|---|---|
| Semantic Analysis | What to look at? – Applications |
| Casting | Conceptual Vectors – Analysis Trees – Lexical Network |
| Ant algorithm | Principles - Without lexical network - With lexical network |
| Ressource Production | Lexical Network - Conceptual Vectors |
| Future directions | Toward holistic analysis: syntax + semantic |

# Semantic Analysis

What do we look at?

## Selection/weighting of acceptions (WSD)

"L'avocat plaide" The lawyer pleads

avocat/fruit or avocat/justice ?

## Prepositional group attachment

"He saw the girl with a telescope"

The man sees (the girl with a telescope)
/ The man (sees [the girl] with a telescope)

"Il (voit avec un télescope ) la fille" or
"Il voit la (fille avec un télescope)" ?

## Interpretation trails

"L'avocat est véreux" 2 trails but not 4

The avocado is rotten
/ The lawyer is corrupted

# Semantic Analysis

What do we look at?

## Anaphora/reference resolution

"The lawyer defended well his client. He was acquitted."

he = lawyer or he = client ?

## Recognition of occurrences of lexical function [meltchuk] [schwab]

"Il a une forte fièvre"

Magn(fièvre) = forte ?

He has a high fever

Magn(fever) = high ?

# Semantic Analysis Applications

## Indexing for information retrieval [jaillet, prince, chauché, teissere]

WSD (mouse)
increase in precision

Synonymy (cat/true cat/feline) / semantic fields (horse/saddle)
increase in recall

## Machine Translation [prince, delorme]

Reference resolution (il => he/she/it ? Son => his/her/its ?)

Contrastive phenomena (river <= rivière/fleuve ?)
(abats => ofals/giblets ?)

Lexical functions (forte fièvre <=> high fever) but (forte <=> high)

# Outline

| | |
|---|---|
| Semantic Analysis | What to look at? – Applications |
| Casting | Conceptual Vectors – Analysis Trees – Lexical Network |
| Ant algorithm | Principles - Without lexical network - With lexical network |
| Ressource Production | Lexical Network - Conceptual Vectors |
| Future directions | Toward holistic analysis: syntax + semantic |

# Casting Conceptual Vectors

## Thematic representation [chauché, lafourcade]

Lexical item = set of ideas => vector

An idea => a concept

For example, 873 concepts in Larousse thesaurus (French)
more than 1000 in Roget

(1)existence, (2)inexistence, (3)matérialité, ..., (516)liberté, ...,
... (872)jeux, (873)jouets

A component of the vector is the activation of the related concept

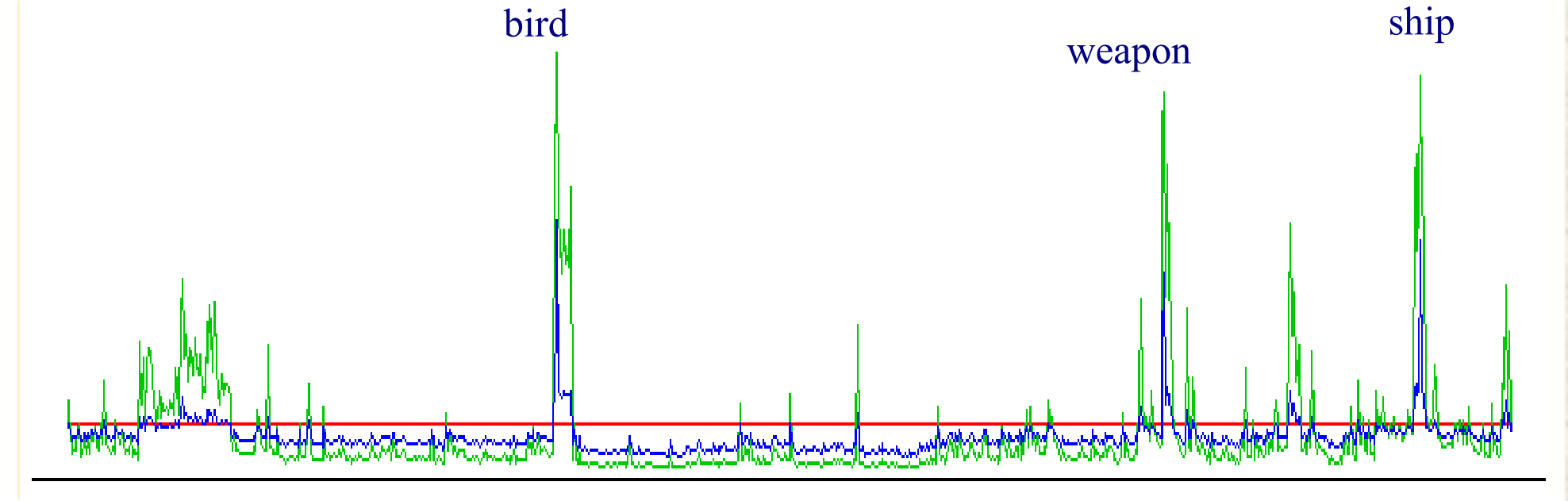
A concept is itself a vector on the same space (neighbourhood of concepts)

Concept are interdependent => « Conceptual » vectors

Vector combinations : addition, contextualisation…
[lafourcade, prince, schwab]

# Conceptual Vectors Example

Frigate
- Ancient ship
- Modern ship
- Bird

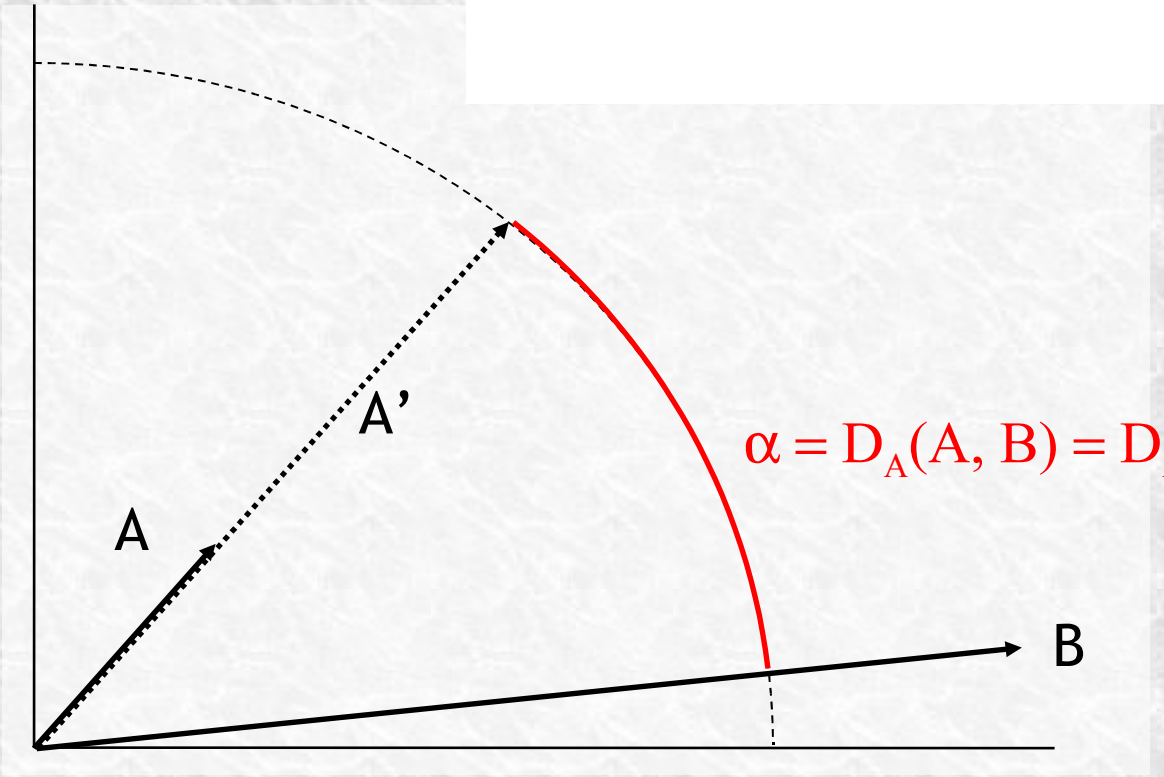


# Conceptual vectors

## Angular distance $D_A$

$$D_A(A, B) = \arccos(Sim(A, B))$$

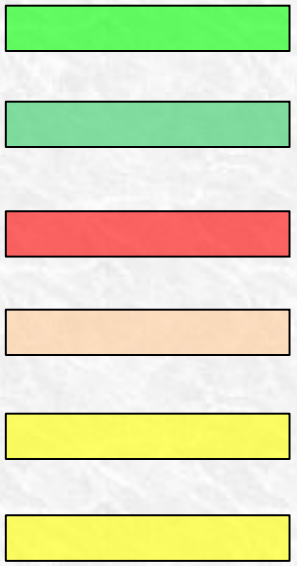
$$Sim(A, B) = \cos(\widehat{A, B}) = \frac{A \cdot B}{\|A\| \times \|B\|}$$




$\alpha = D_A(A, B) = D_A(A', B)$

# Conceptual Vectors

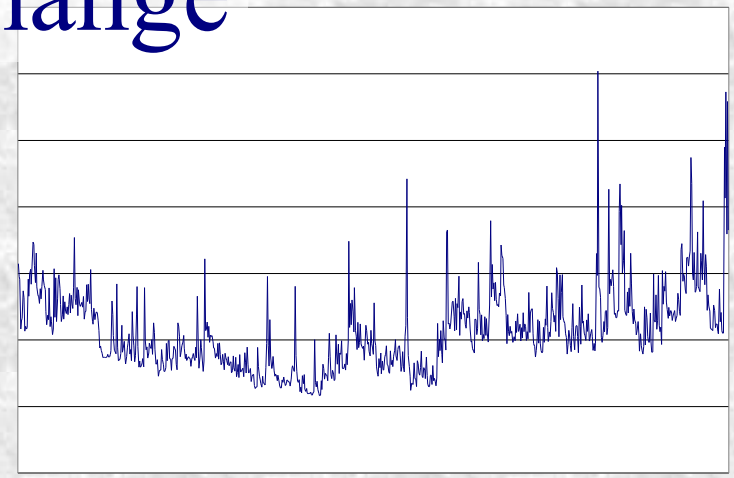
$D_A$(*pelican*, *pelican*) = 0 (0°)
$D_A$(*pelican*, *white pelican*) = 0,2 (11°)
$D_A$(*pelican*, *train*) = 1,22 (70°)
$D_A$(*pelican*, *bird*) = 0,46 (26°)
$D_A$(*pelican*, *seagull* ) = 0,4 (23°)
$D_A$(*pelican*, *fish* ) = 0,35 (20°)

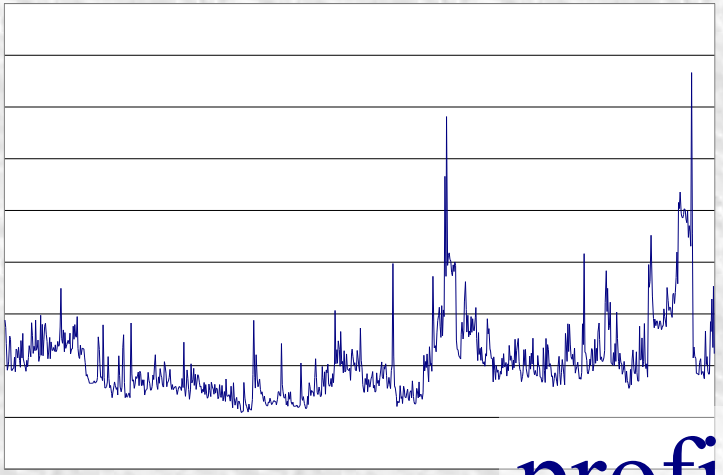


Thematic distance ≠ Ontologic distance (of type *is-a*)
but
Thematic distance ⊃ Ontologic distance

# Conceptual Vectors Examples

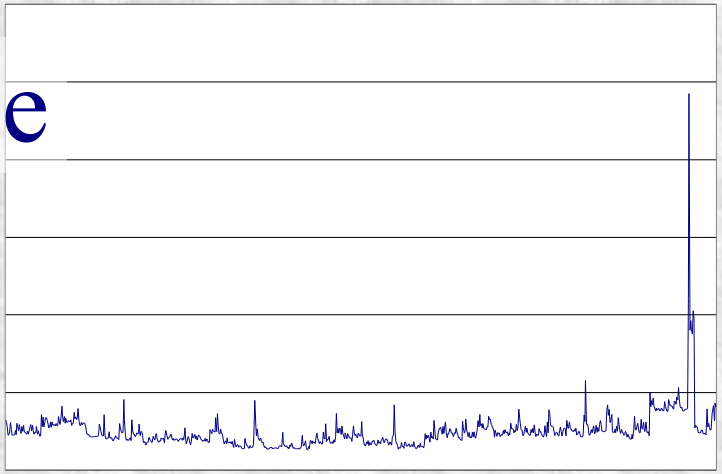
exchange





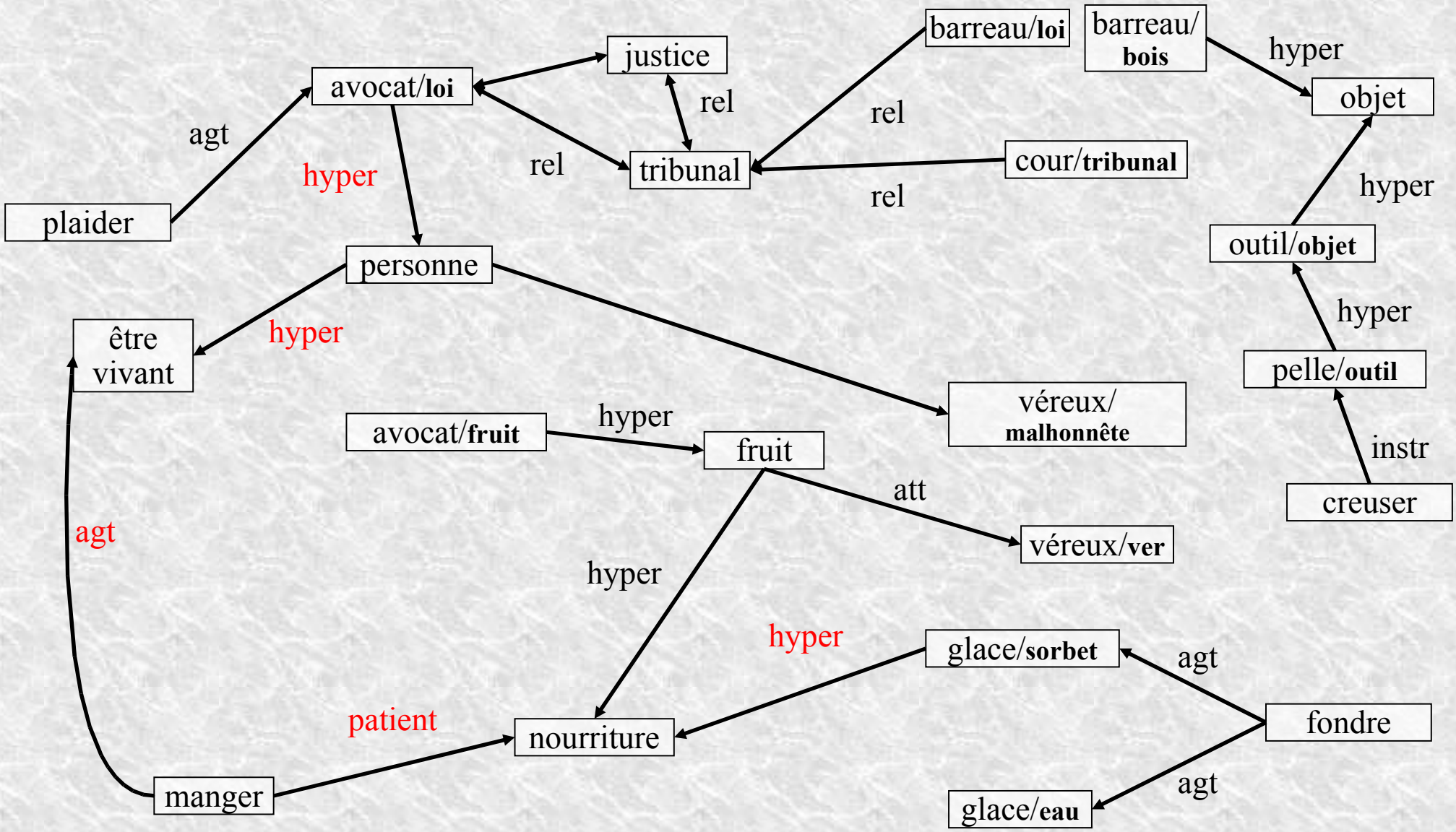
profit

finance



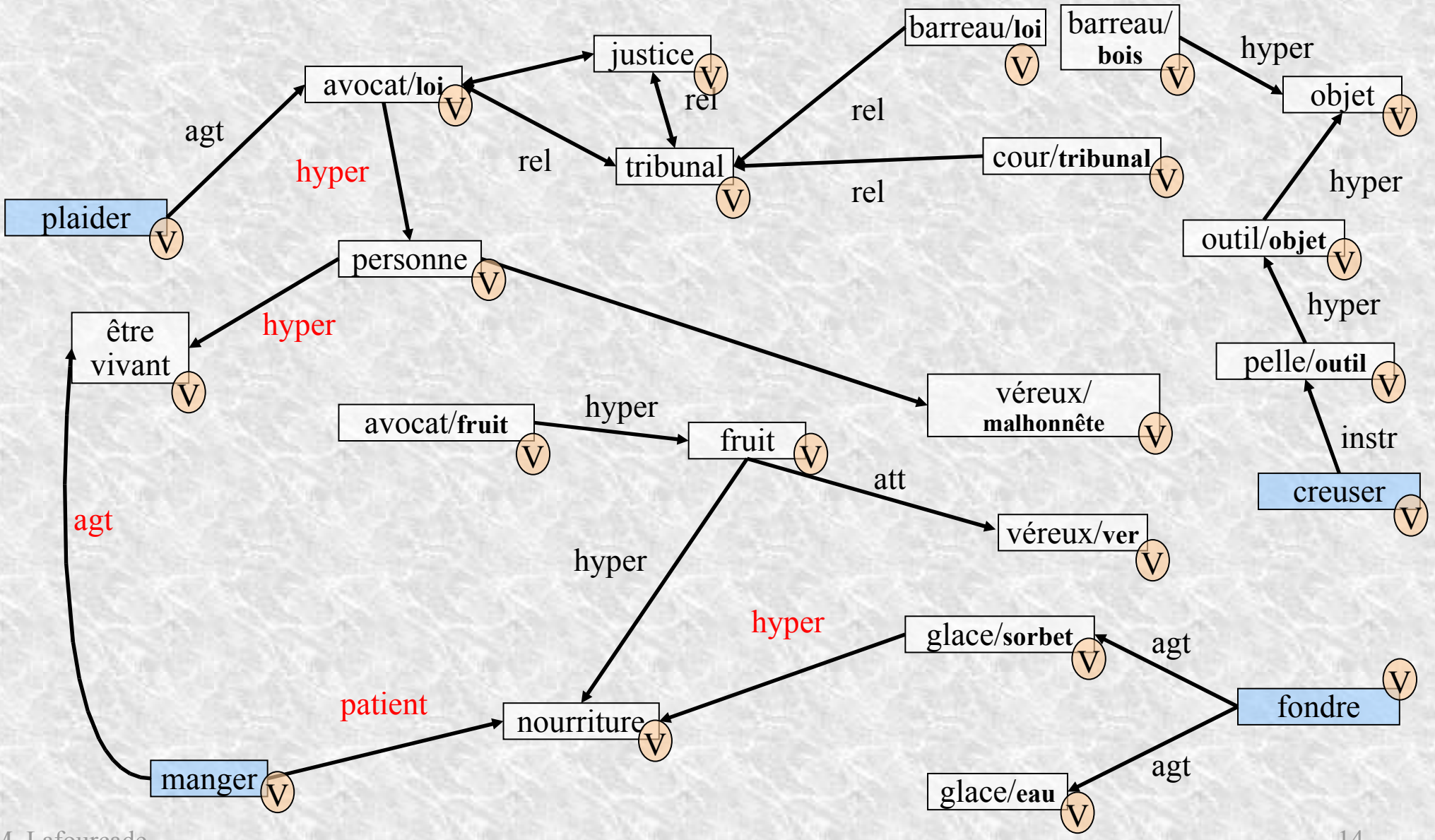
# Casting

## Lexical network



# Casting

## Lexical network with conceptual vectors



## Adding vectors to graph nodes

=> a shortcut for term proximity

graph searching algorithm cost > $D_A$ function cost

=> vectors can be computed according to points of view

=> can be done incrementally

Graphs tend to be a **precision** oriented data structure

Vectors tend to be a **recall** oriented data structure

We need both!

# Casting

## Morphosyntactic trees

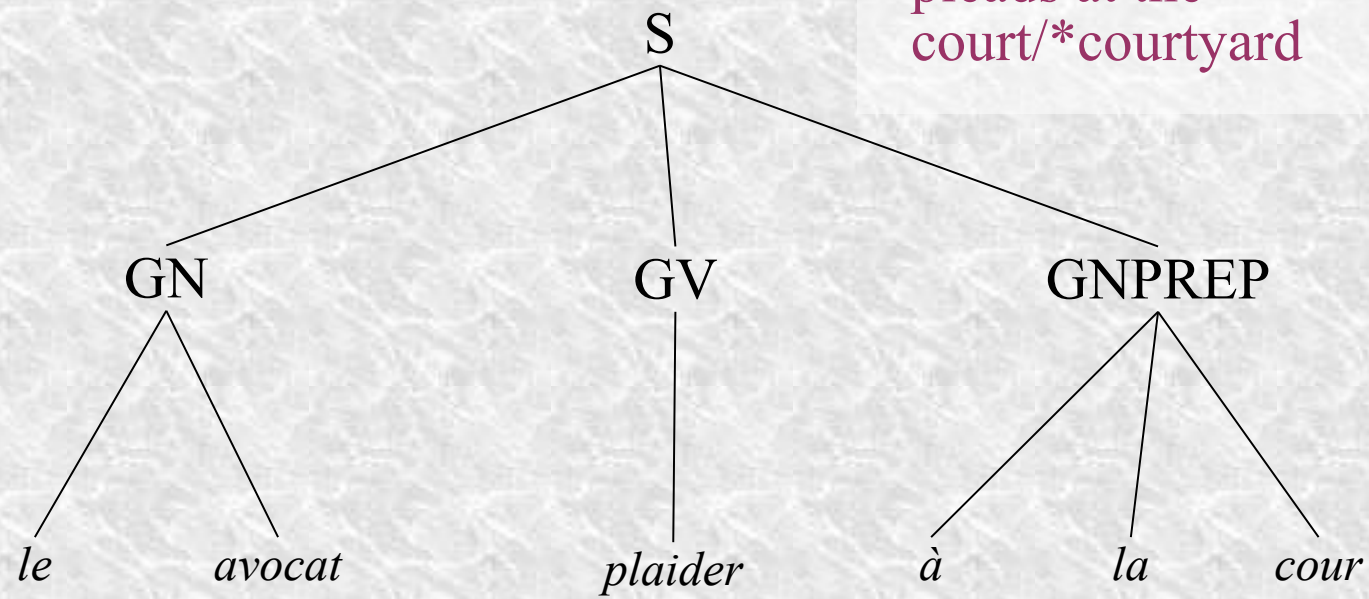
### Constituent tree structure

geometry + set of instantiated variables

SYGMART application to French (SYGFRAN) **[chauché]**

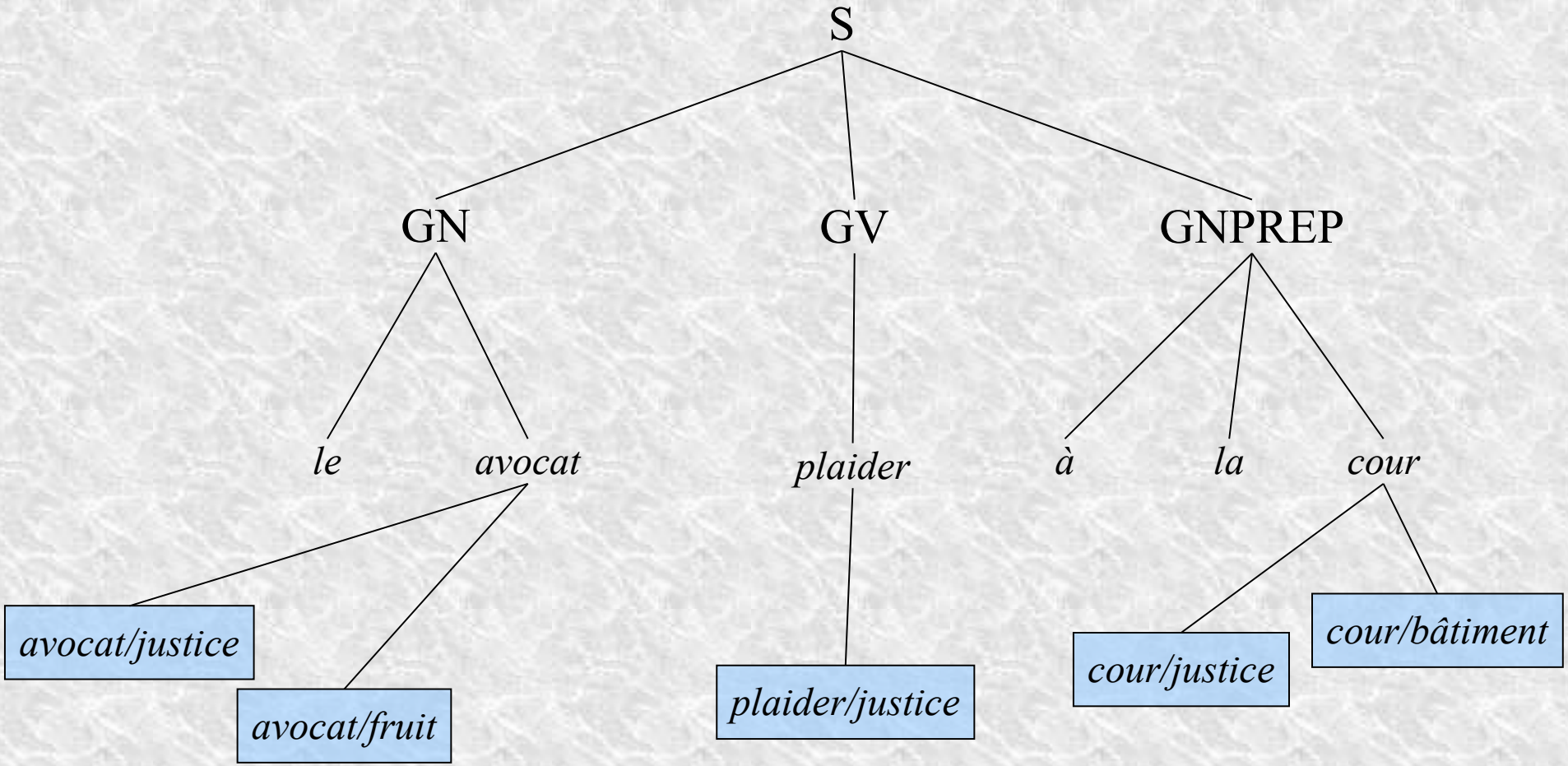
"L'avocat plaide à la cour"

The laywer/*avocado pleads at the court/*courtyard



# Morphosyntactic trees

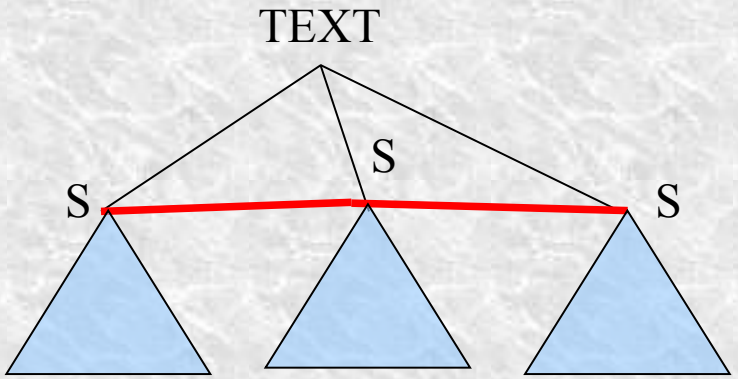
## Adding acceptions



# Morphosyntactic trees Adding links => Graph

## Transforming the tree into a graph

Links between sentence to reconstitute text flow



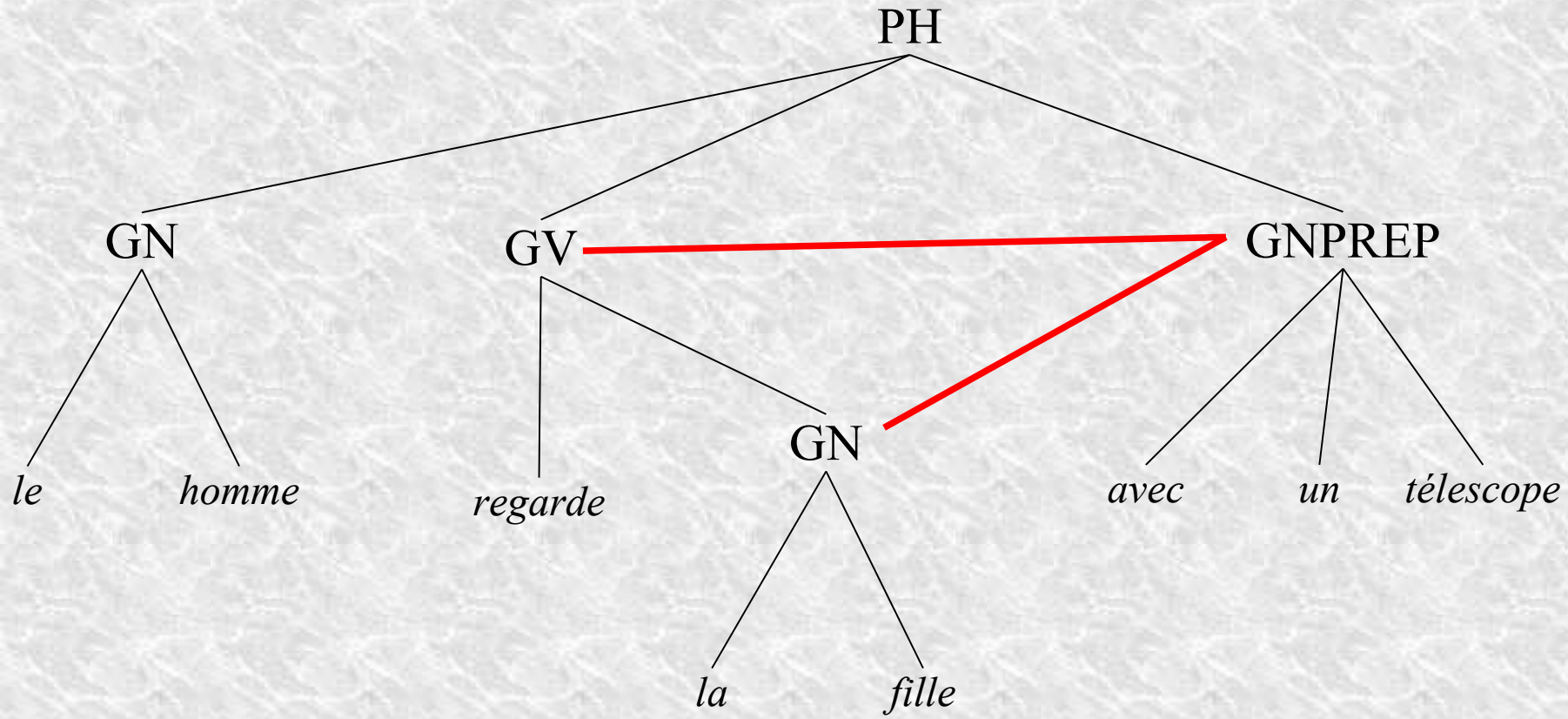

=> allows the computation of distance between sentences

Combination between structural and surface information

# Morphosyntactic trees Adding links => Graph

## Transforming the tree into a graph

Enumeration of syntactically possible attachments (GNPREP)



# Ant algorithms

## Multi-agent system – reactive agents

TSP – operational research - network – bio informatic
[dorigo] [bertelle] [bonabeau, théraulaz] [bruten] [costa, hertz]

## Principle

Stygmergy = indirect communication
through modification of the environment

Pheromones drops – Slow decay over time
=> handling of an ever-changing environment

## NLP

Cognitive agents [stéphanini] [sabah]

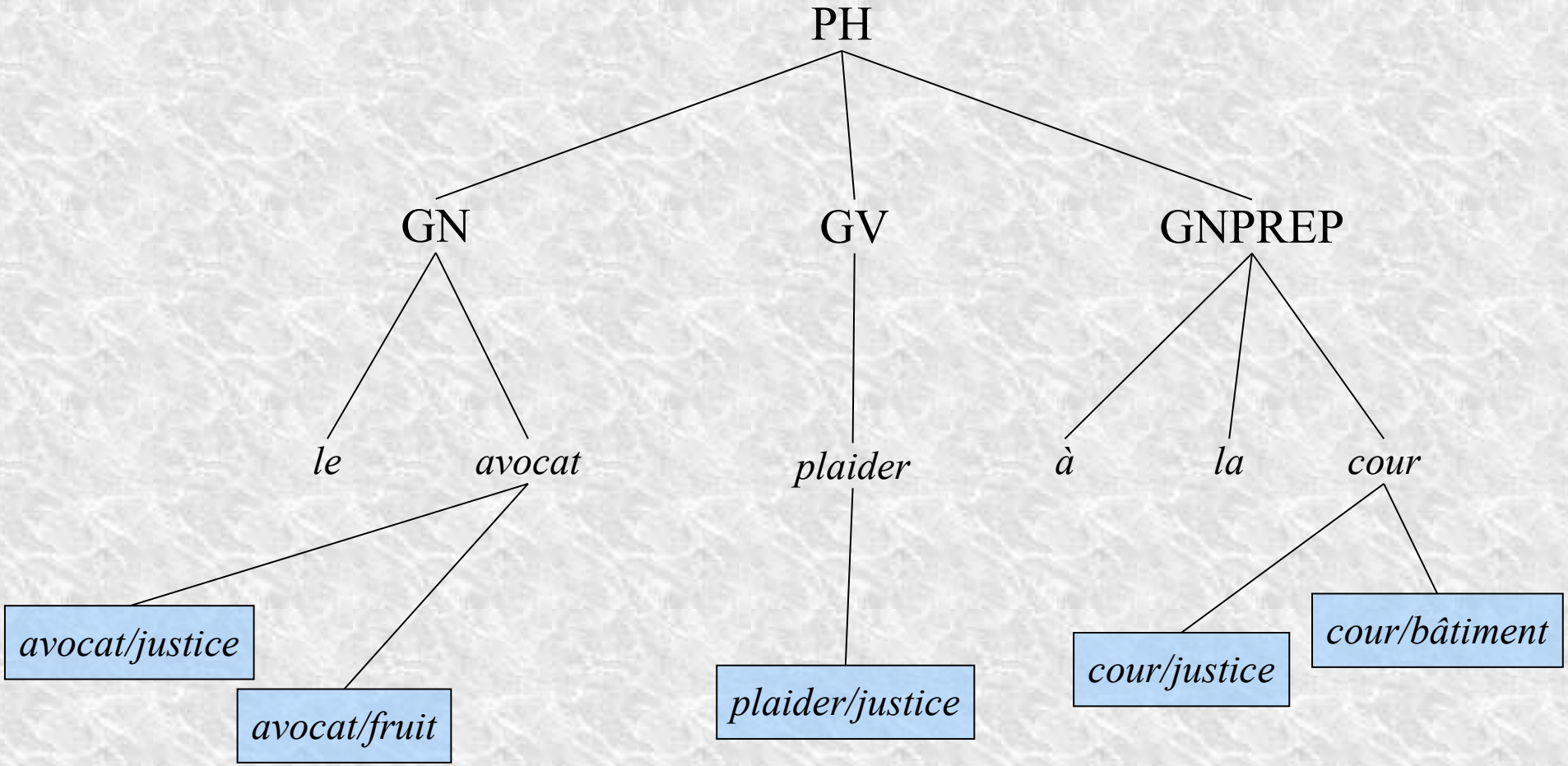
otherwise reactive [lafourcade, guinand] [+cunningham] [zamora]

# Outline

| | |
|---|---|
| Semantic Analysis | What to look at? – Applications |
| Casting | Conceptual Vectors – Analysis Trees – Lexical Network |
| Ant algorithm | Principles - Without lexical network - With lexical network |
| Ressource Production | Lexical Network - Conceptual Vectors |
| Future directions | Toward holistic analysis: syntax + semantic |

# Ant algorithms

## Environment

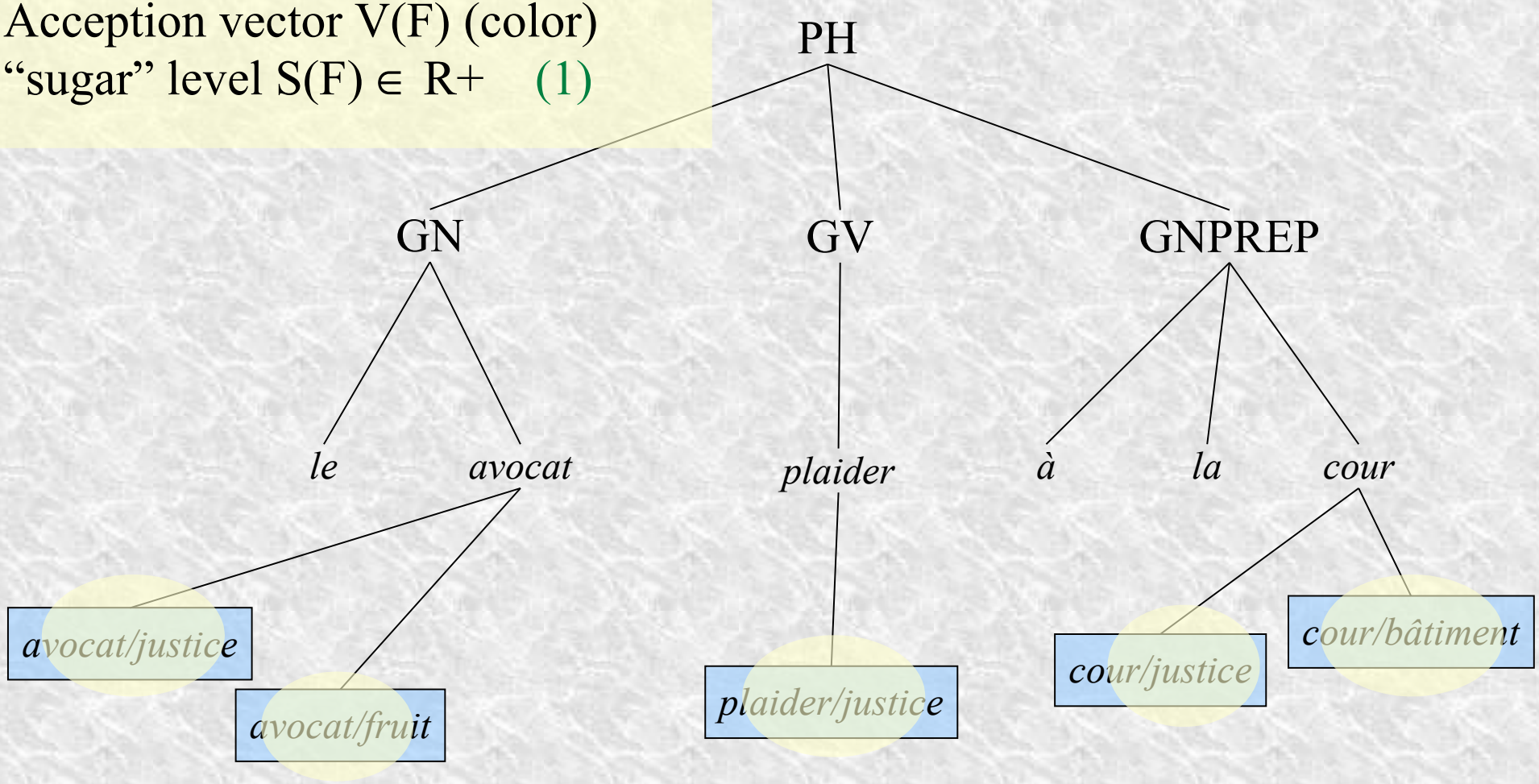


# Ant algorithms

## Environment

**Anthill / Acception nodes**

Acception vector V(F) (color)
"sugar" level S(F) ∈ R+ (1)

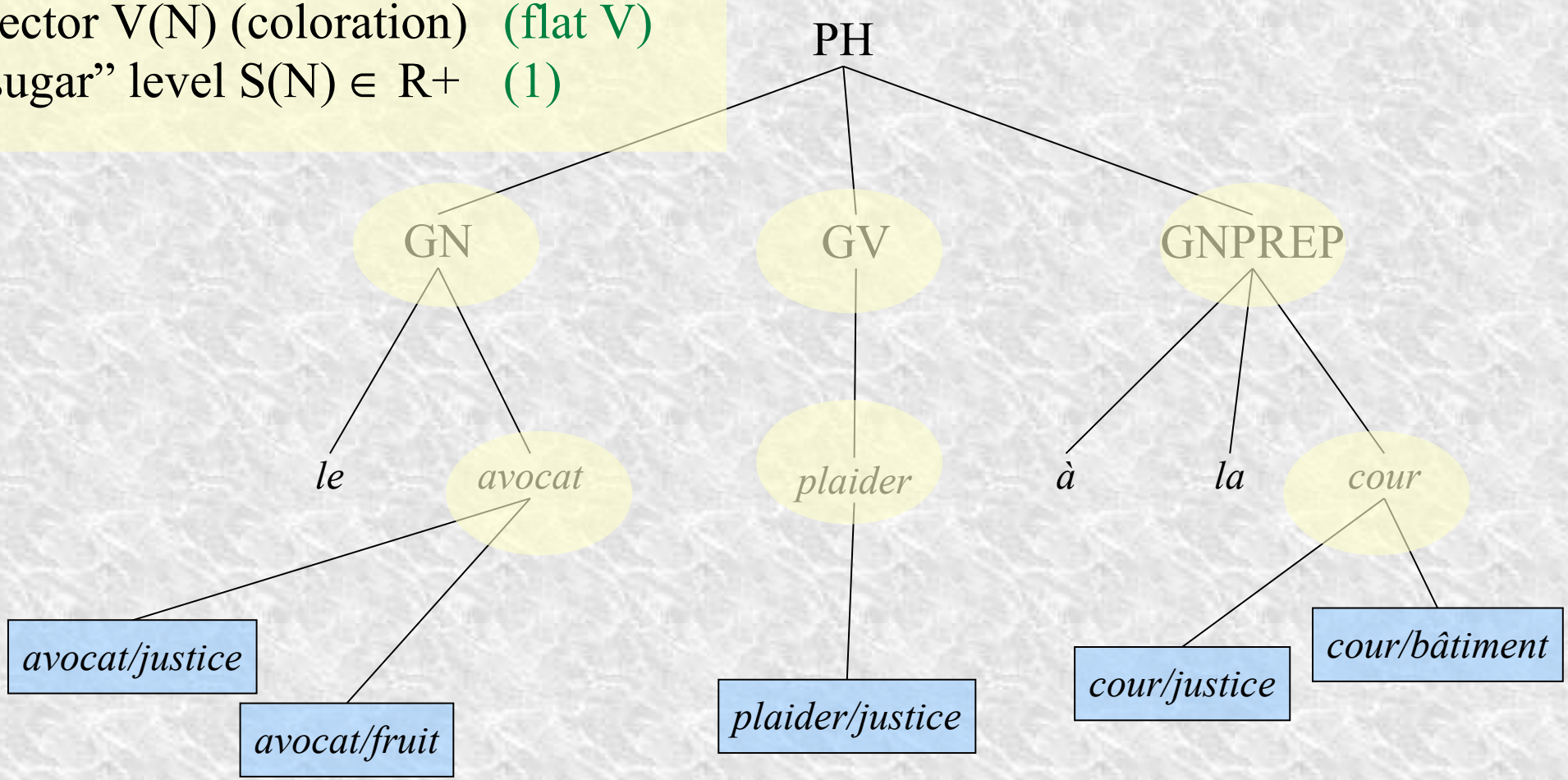


# Ant algorithms

## Environment

**Standard node**

Vector V(N) (coloration) (flat V)
"sugar" level S(N) ∈ R+ (1)



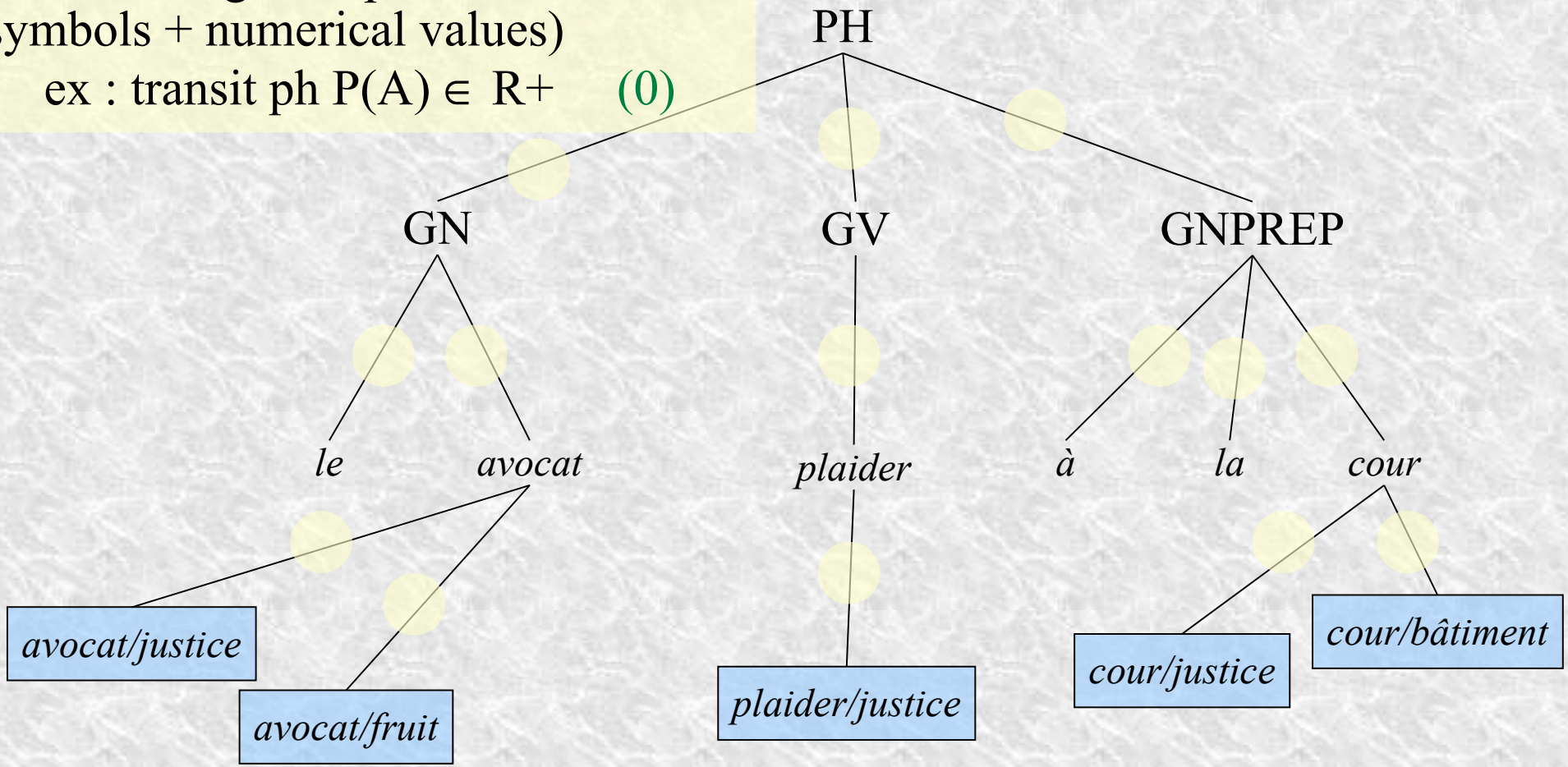
# Ant algorithms

## Environment

**Arc A**

- type
- Valuated signal = pheromones (symbols + numerical values)

ex : transit ph P(A) ∈ R+ (0)



# Ant algorithms

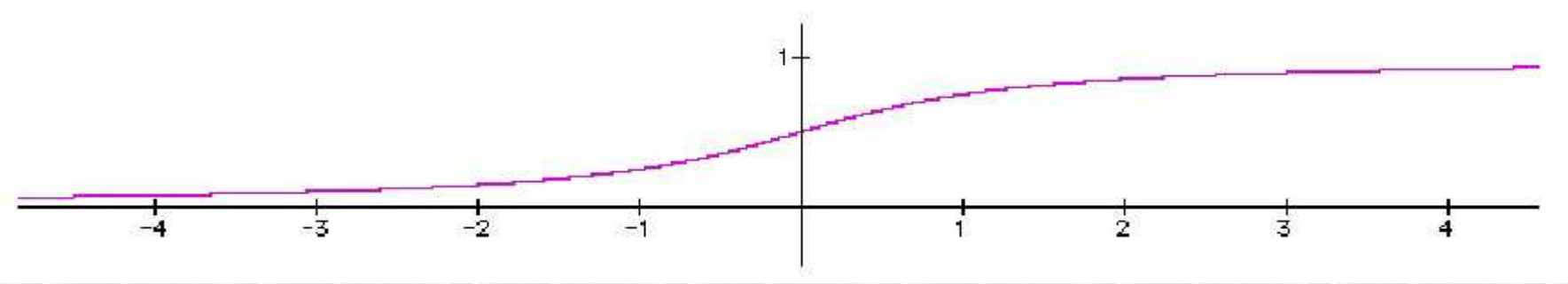
Environment

## Ant production

one or several ants at each cycle

with a probability equal to the sugar level S(F)

"sigmoïde" function




Production cost (experiment with 1/10)

# Ant algorithms Agents

## Attributes of an ANT

| | |
|---|---|
| « sugar » load | Q real in [0,1] |
| sugar dropped at death | |
| Life span | 20 cycles |
| Reference to its home F | V(*f*) = V(F) |
| Current moving strategy | Prob = Q |

## Two moving strategies

Looking for « sugar » - **search mode**

Bringing the « sugar » back home – **return mode**

# Ant algorithms

## Ant moves

### Pseudo-random moves

Evaluation of possible destinations + pseudo taboo list (k = 1)

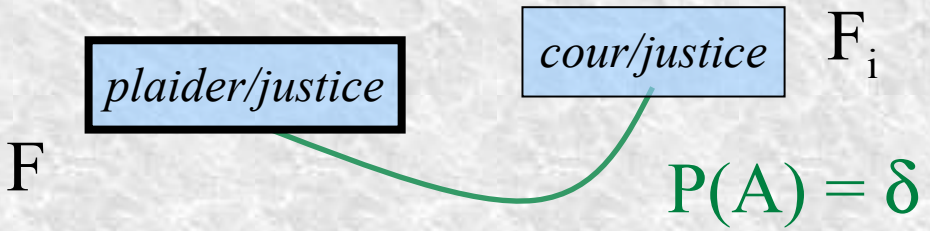
### Environment modification

Drop of pheromones of the arc $P(A) = P(A) + \delta$

Modification of the vector of the reached node $V(N) = V(N) + \alpha V(f)$

normed sum with $\alpha$ and $\delta$ small

### Creation of a bridge towards F



# Ant algorithms

## Ant moves

### General principle

Current position = a node N with k neighboors $N_i$ through arcs $A_i$

Evaluation function for each $N_i$ = Eval($N_i$)

Evaluation function for each $A_i$ = Eval($A_i$)

Global evaluation function of destination $dest_i$

$$Eval(dest_i) = \max \begin{cases} Eval(N_i) + Eval(A_i) \\ \epsilon \end{cases}$$

$$Prob(dest_i) = \frac{Eval(dest_i)}{\sum_{k=1}^{n} Eval(dest_k)}$$

# Ant algorithms

## Sugar foraging and sharing

### Search mode

Specific action: taking sugar from the node

$$Eval(N_i) = S(N_i)$$

$$Eval(A_i) = 1 - P(A_i)$$

### Return mode

Specific action: dropping sugar on the node if its looks like home

$$Eval(N_i) = 1 - \frac{2}{\pi} D_A(V(N_i), V(f)))$$

$$Eval(A_i) = P(A_i)$$

# Ant algorithms

## Signal Decay

At each simulation cycle, pheromone signals decay slightly

$P(A) = \text{Min}\,(P(A) - d * \eta, 0)$

$d = \text{dist}(N_a, N_b)$ and $\eta$ a decay factor

Arc of the tree between $N_a$ and $N_b$

$P(A) = \text{Min}\,(P(A) - \eta, 0)$ because $\text{dist}(N_a, N_b) = 1$

Bridge between $N_a$ and $N_b$

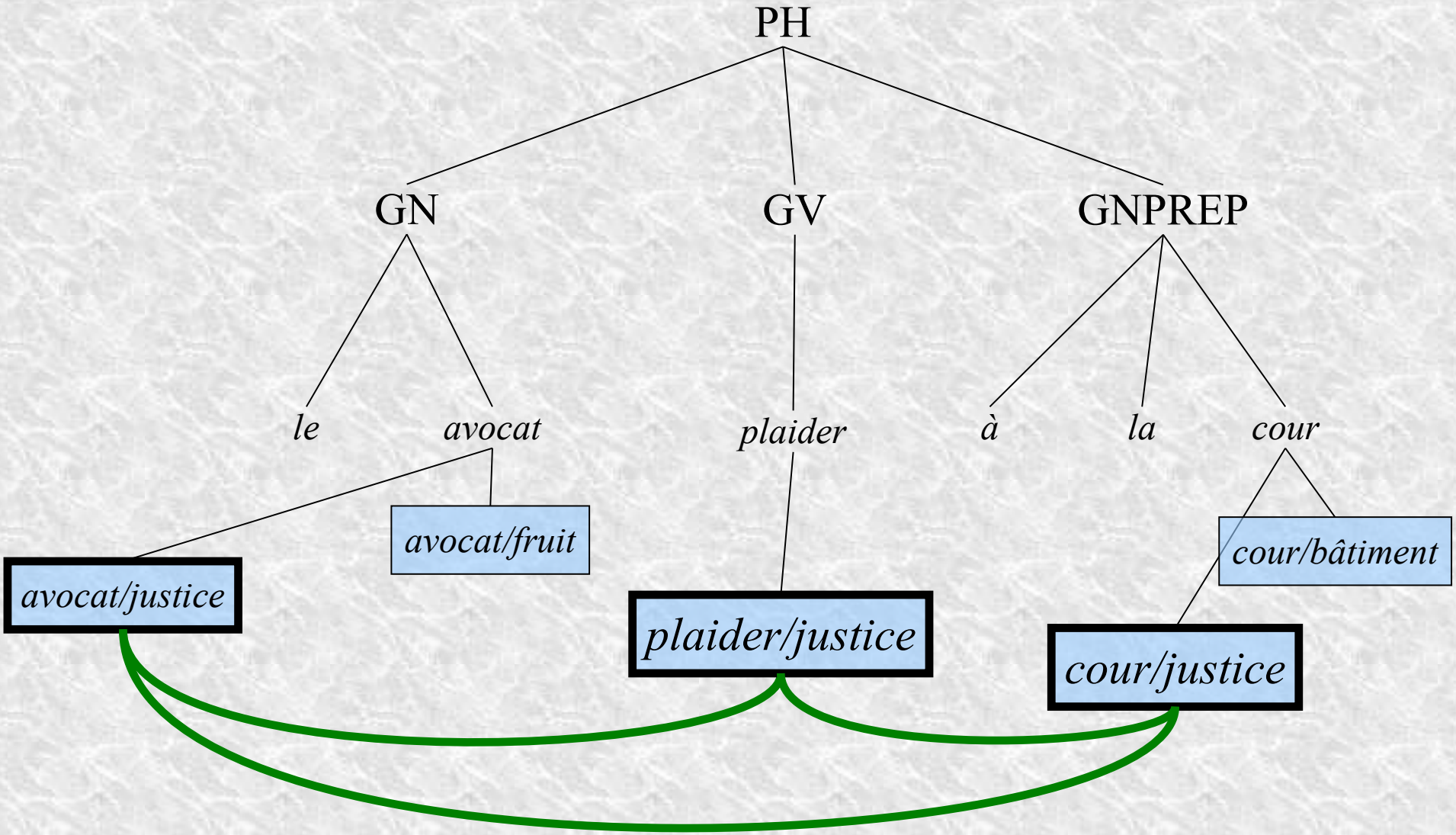
Deleted if $P(A) = 0$

The longer the bridge, the harder it is to keep

# Ant algorithms

## Simple example

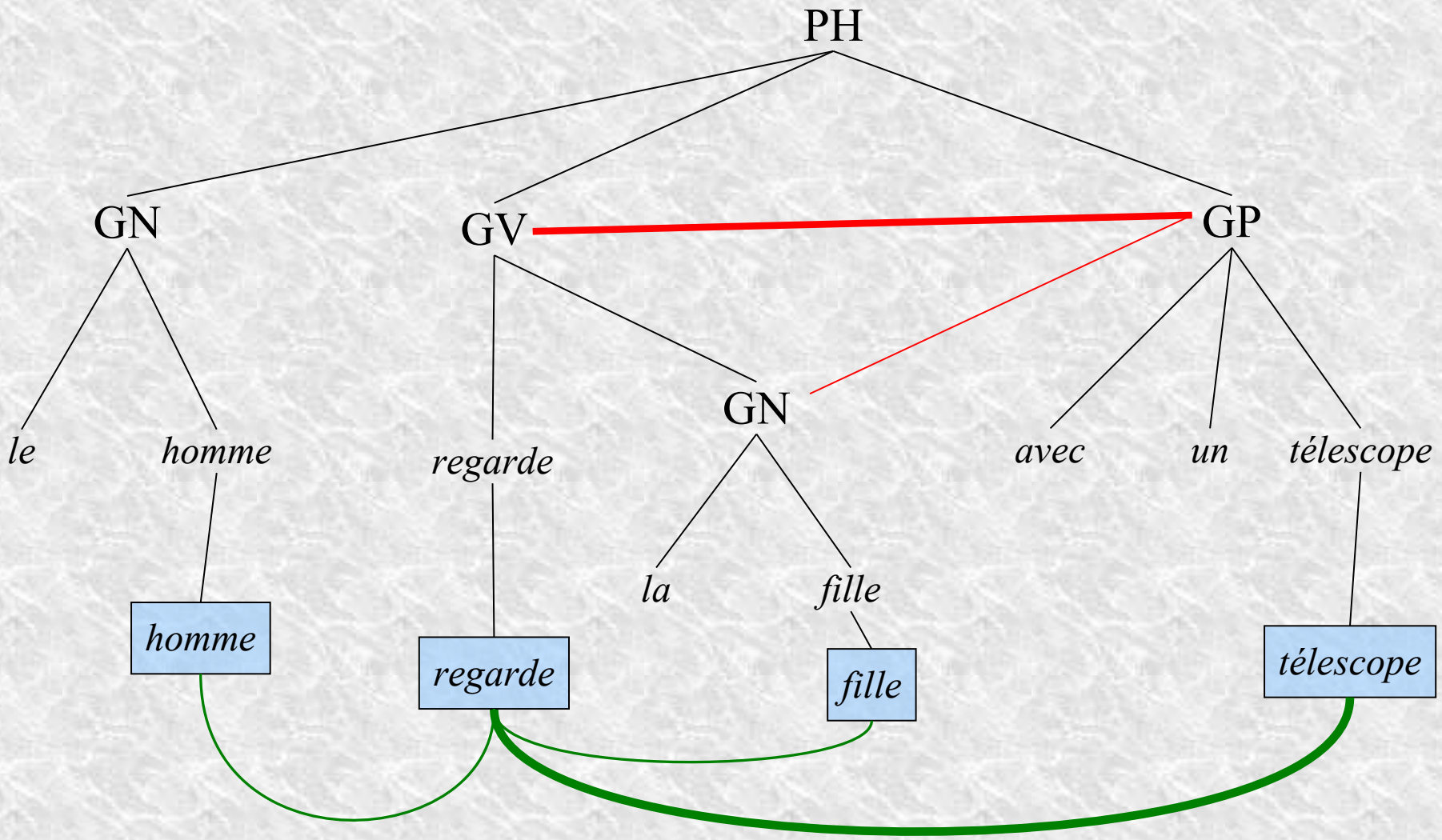
The laywer/*avocado pleads at the court/*courtyard



# Ant algorithms

## Simple example for attachment

The man sees (the girl with a telescope)
/ The man (sees [the girl] with a telescope)



# Ant algorithms

## Results for the simple approach

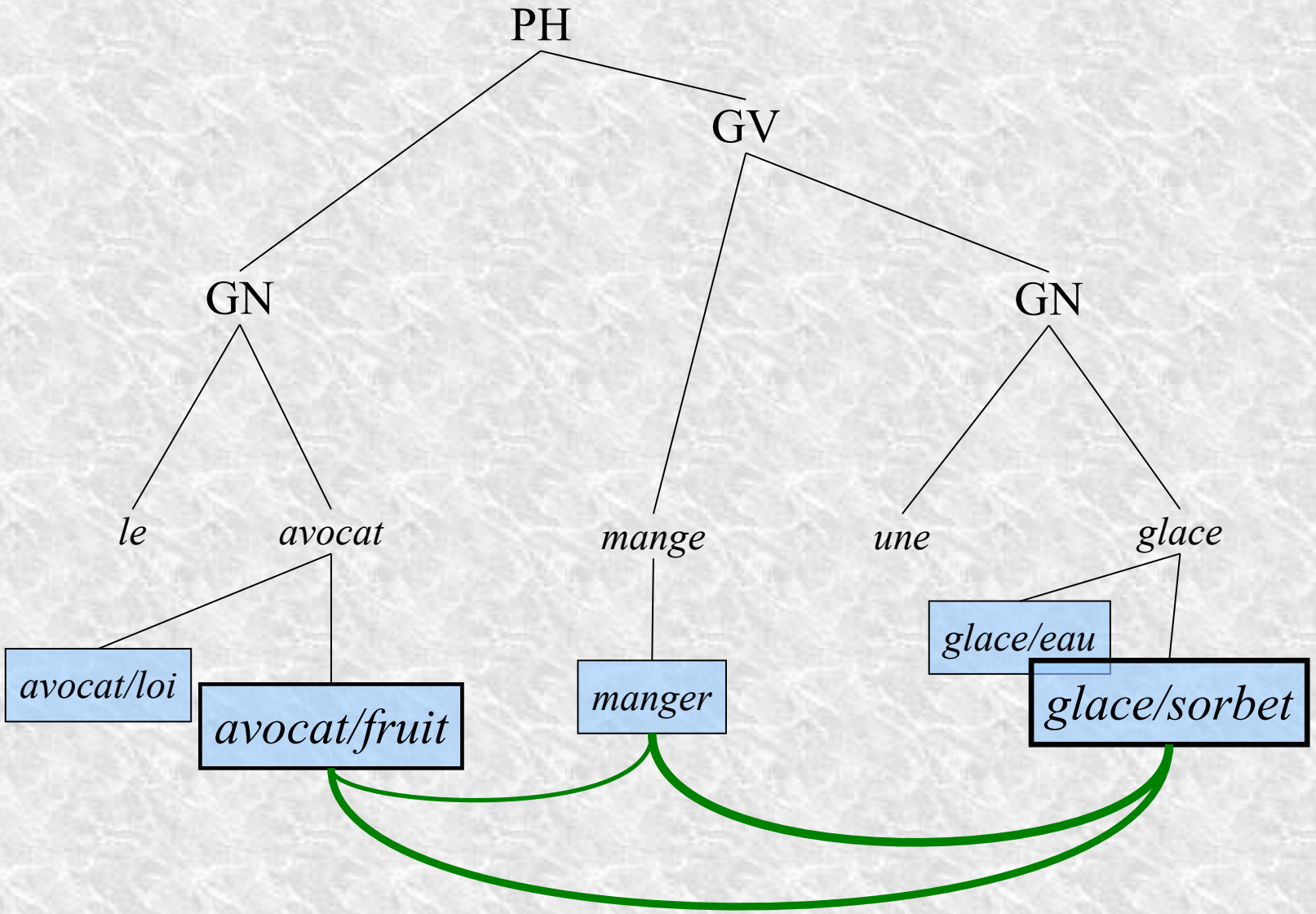
Thematic information is enough for WSD in 73% of the cases

| Scoring scheme | All terms | Nouns | Adjectives | Verbs | Adverbs |
|---|---|---|---|---|---|
| Fine Grain Scoring | 0.73 | 0.76 | 0.78 | 0.61 | 0.85 |
| Coarse Grain Scoring | 0.85 | 0.88 | 0.9 | 0.72 | 0.95 |
| Base Line with POS | 0.33 | 0.30 | 0.4 | 0.17 | 0.60 |
| Base Line no POS | 0.25 | 0.24 | 0.25 | 0.12 | 0.52 |

Room for improvement

# Ant algorithms

*Simple example



# Ant algorithms Ant Caste

## Ant castes

Ants with some constraints on the evaluation functions

## Castes exploiting the lexical network

Predicate => agent

Prédicate => patient

Prédicate => instrument …

## Local copy

Acceptions and relations of the networks

## Local fusion

of nodes that seems to be identical

# Ant algorithms caste Predicate => Patient

Ant caste created by a verb (predicate)

Search mode

$$Eval(N_i) = S(N_i)$$

$$Eval(A_i) = 1 - P(A_i) + \max \begin{cases} 2 & si \quad type(A_i) = patient \\ 0 & sinon \end{cases}$$

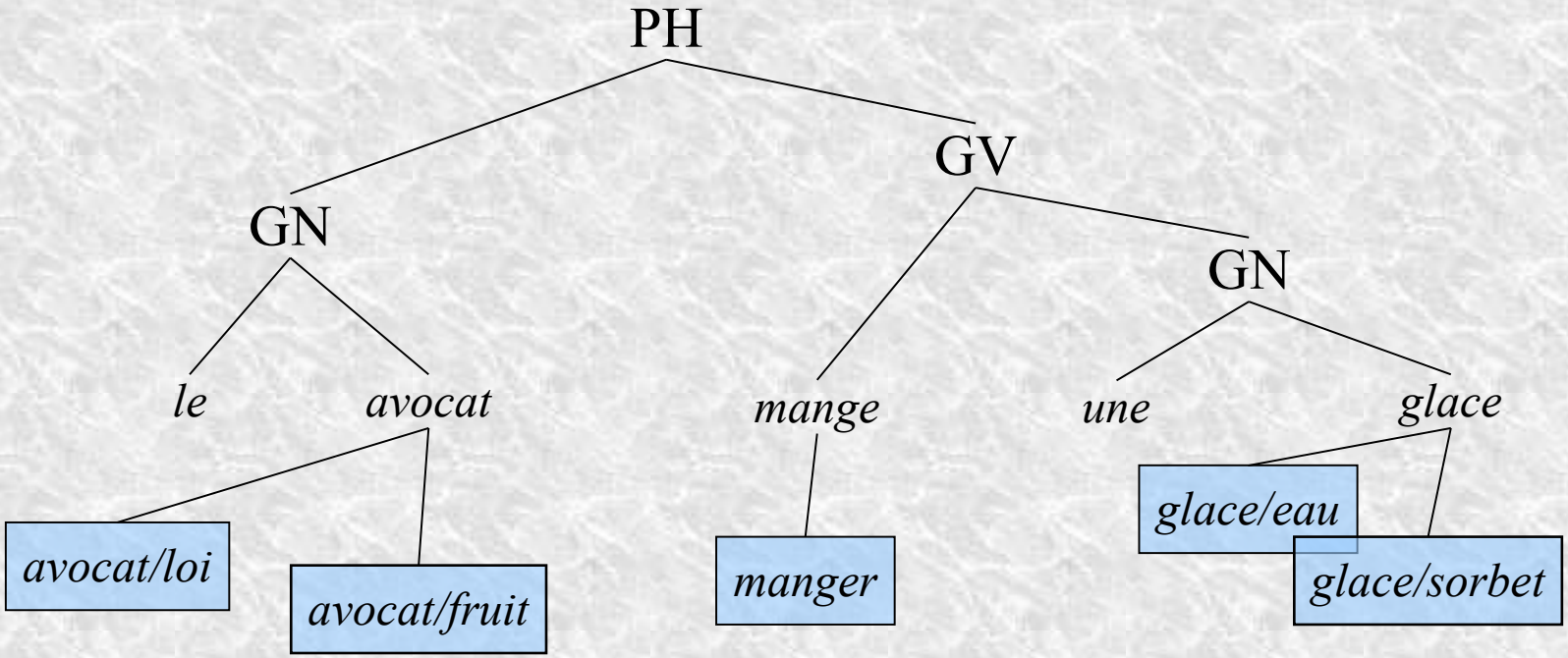
Return mode

$$Eval(N_i) = 1 - \frac{2}{\pi} D_A(V(N_i), V(f)))$$

$$Eval(A_i) = P(A_i) + \max \begin{cases} 2 & si \quad type(A_i) = patient \\ 0 & sinon \end{cases}$$

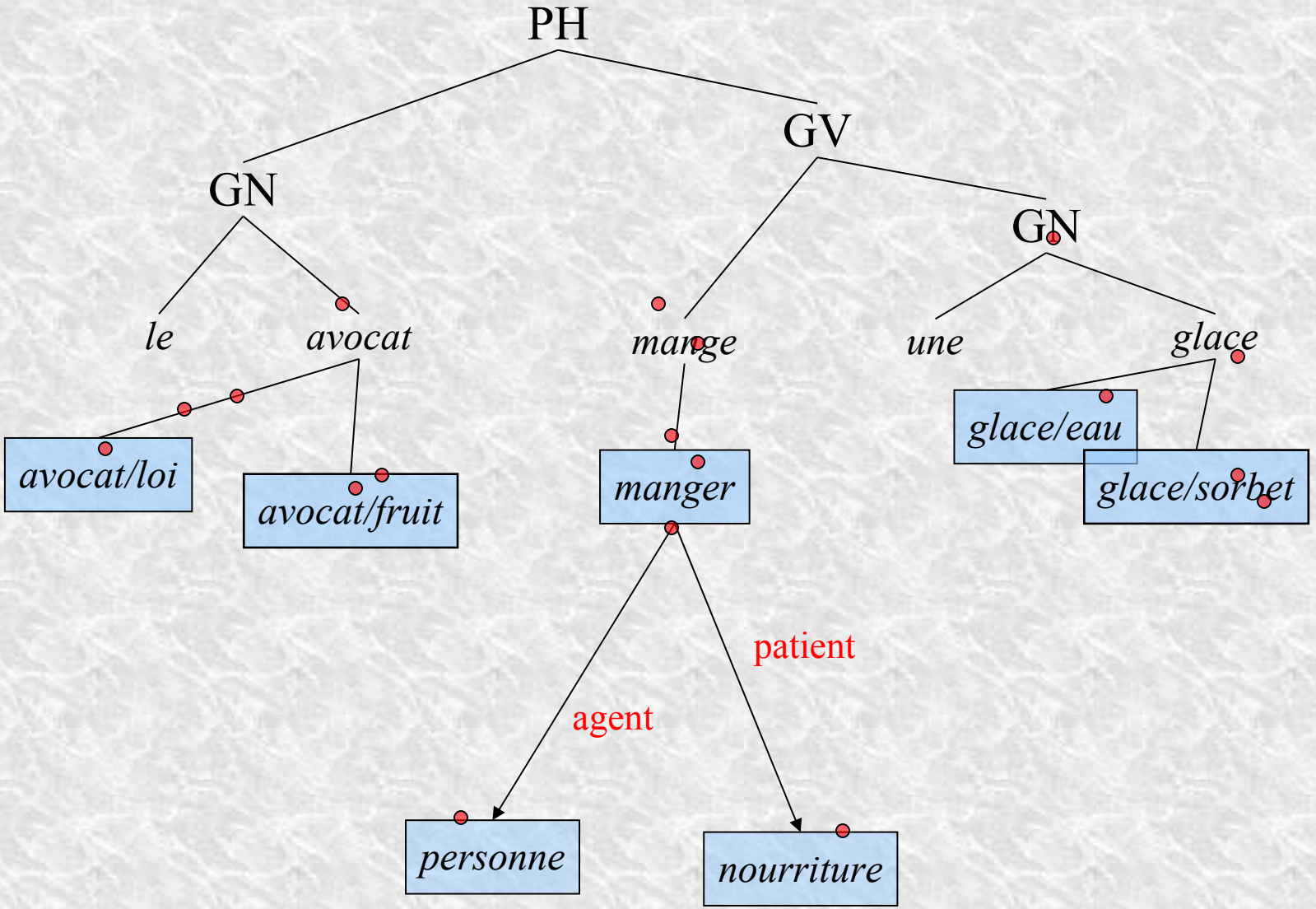
# Ant algorithms

## Example with lexical network



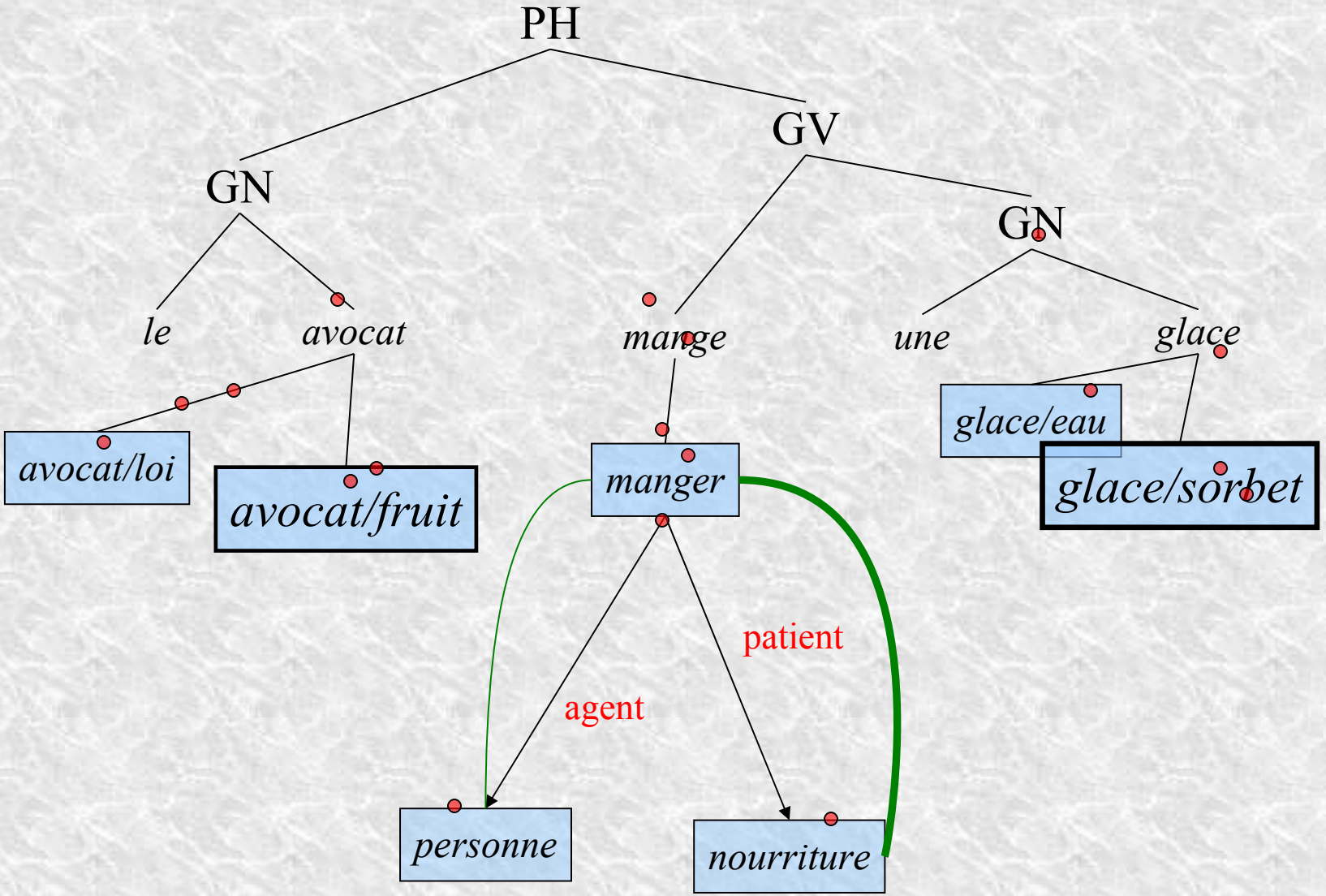
# Ant algorithms

## Example with lexical network



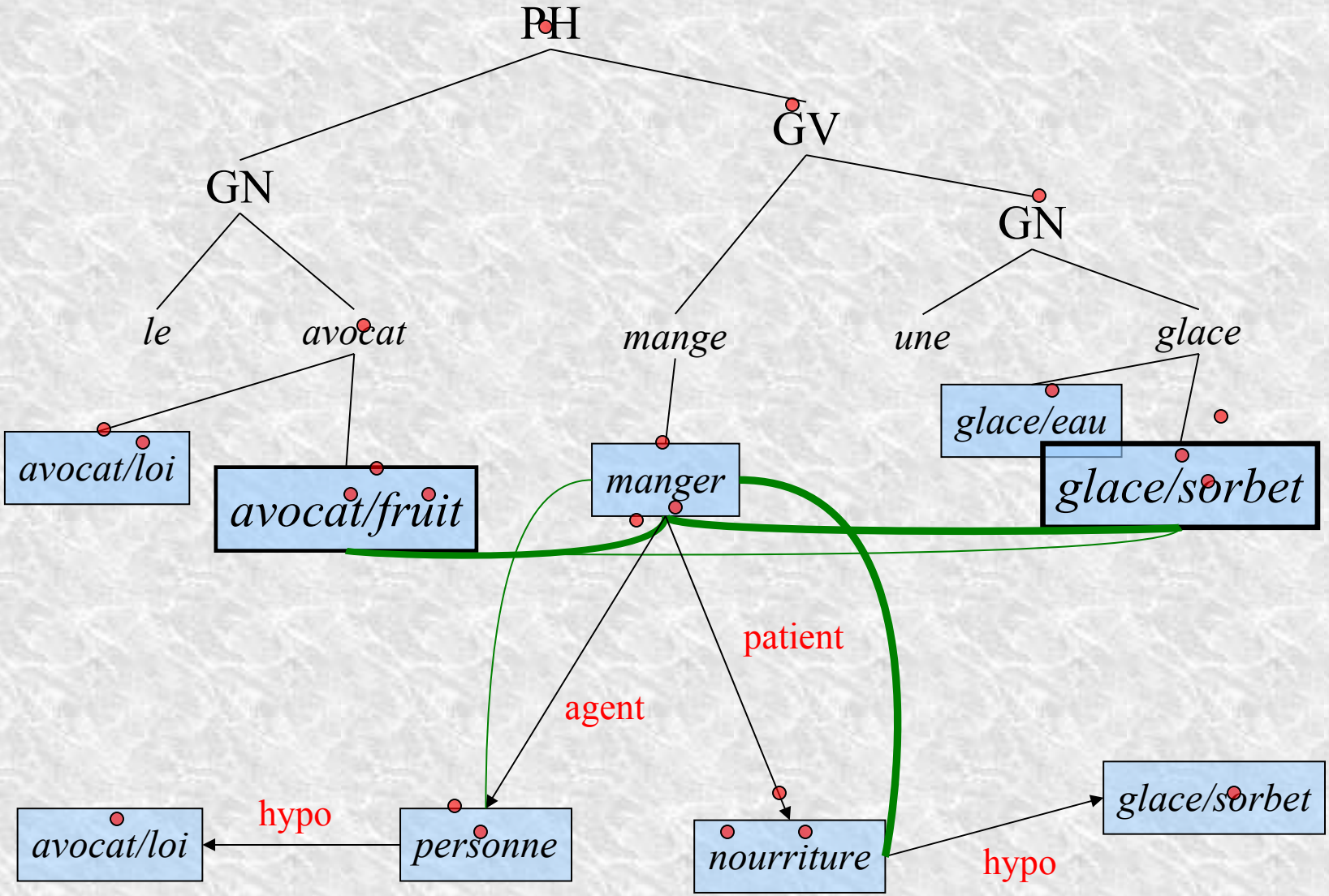
# Ant algorithms

## Example with lexical network



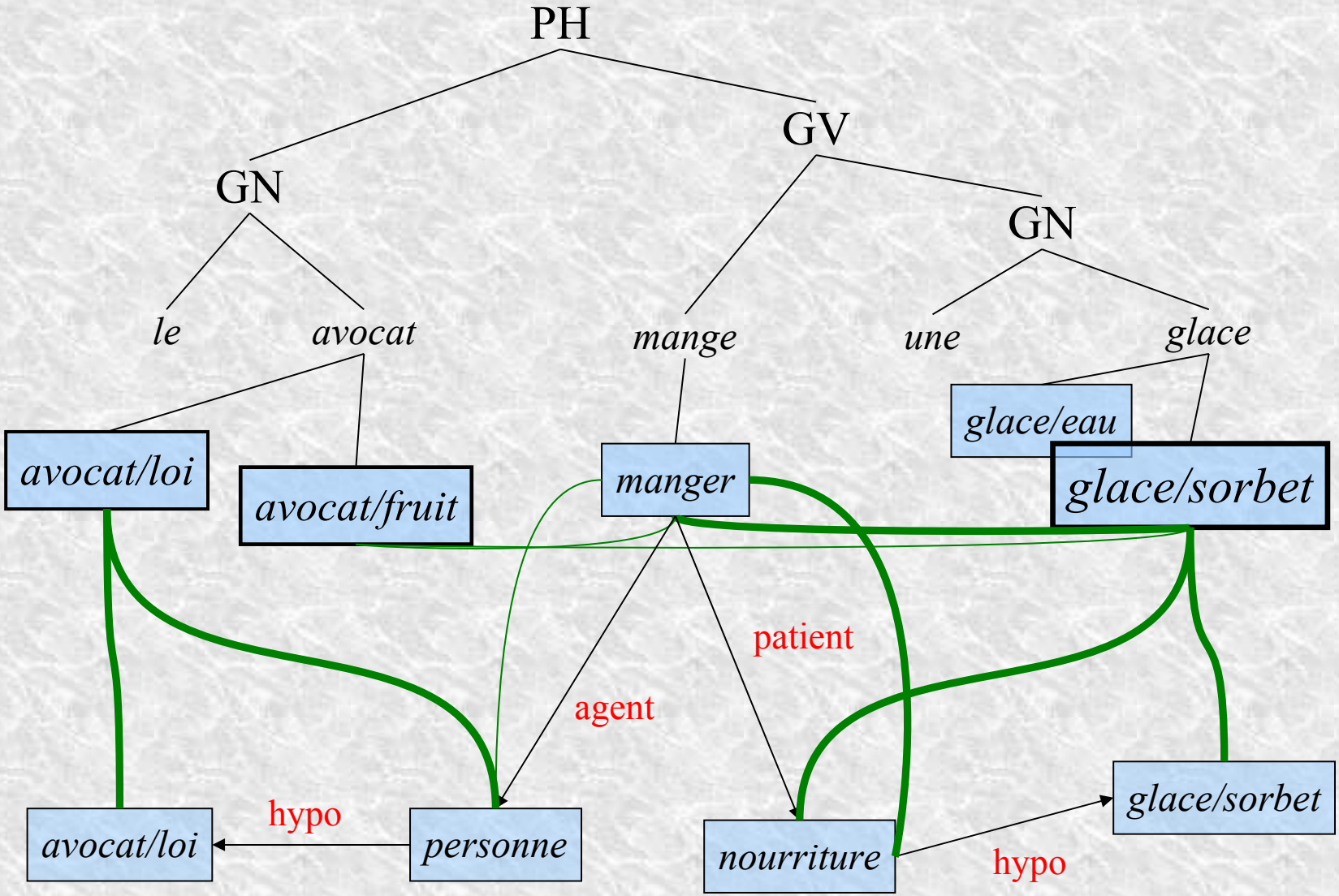
# Ant algorithms

## Example with lexical network



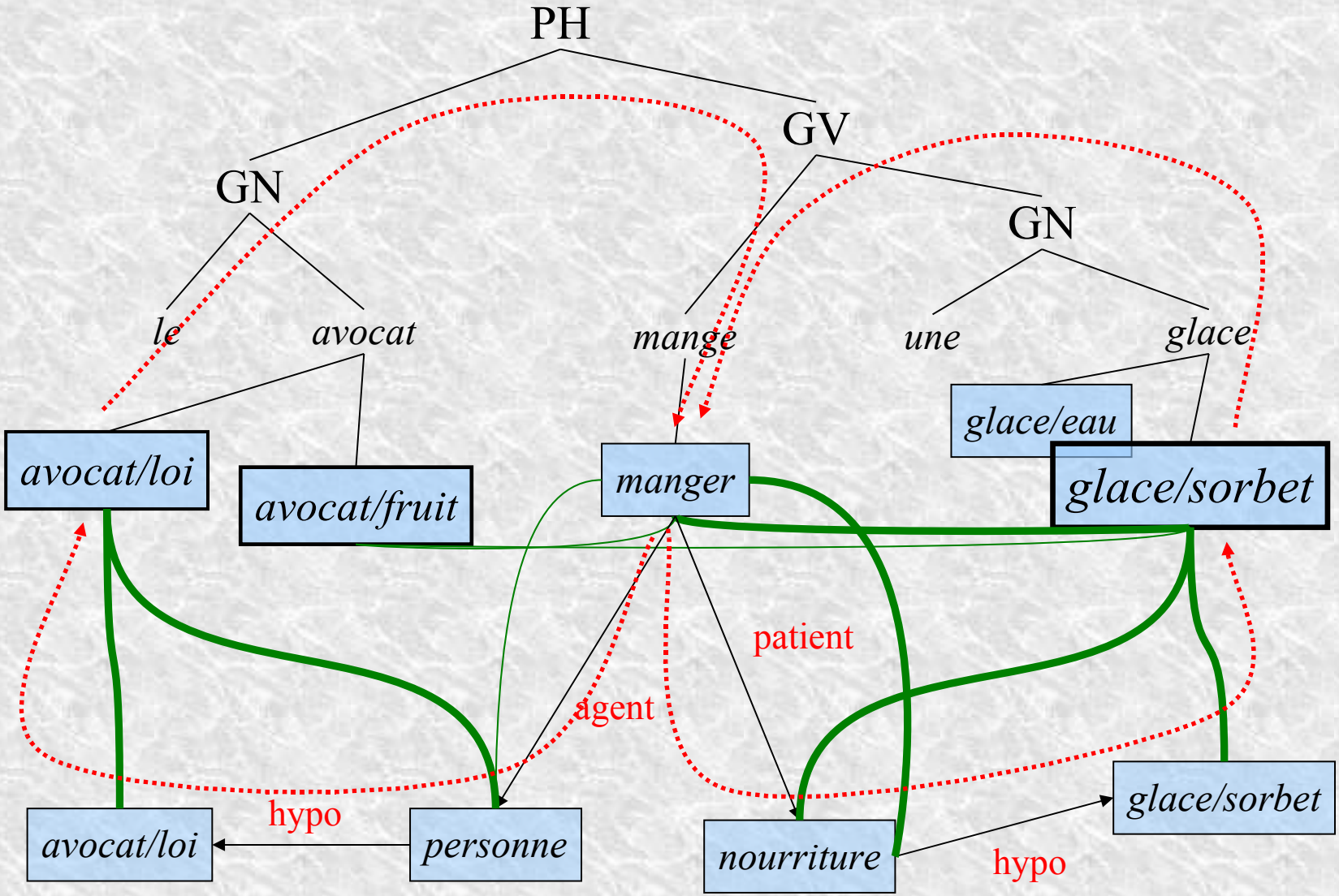
# Ant algorithms

## Example with lexical network



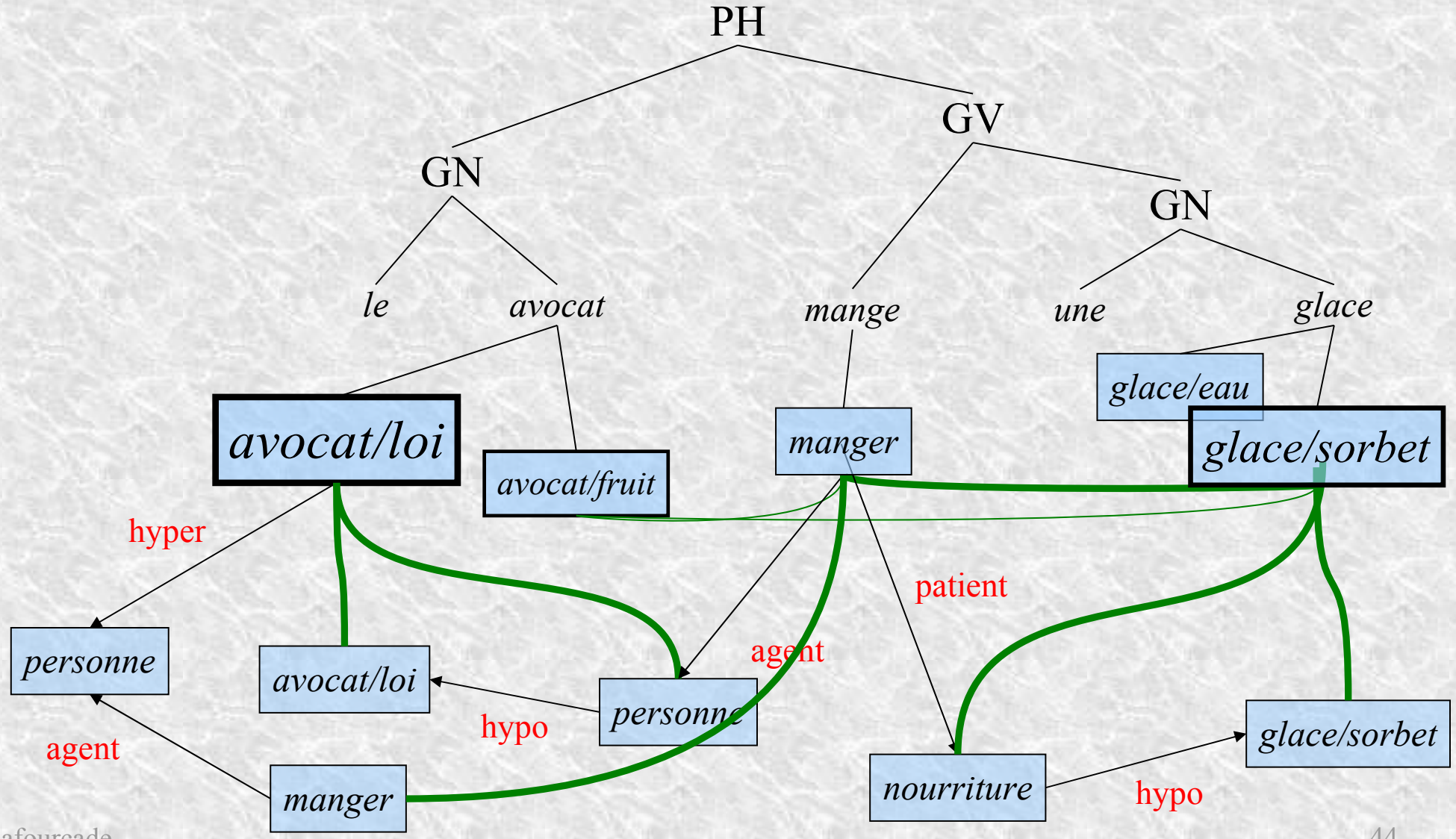
# Ant algorithms

## Example with lexical network



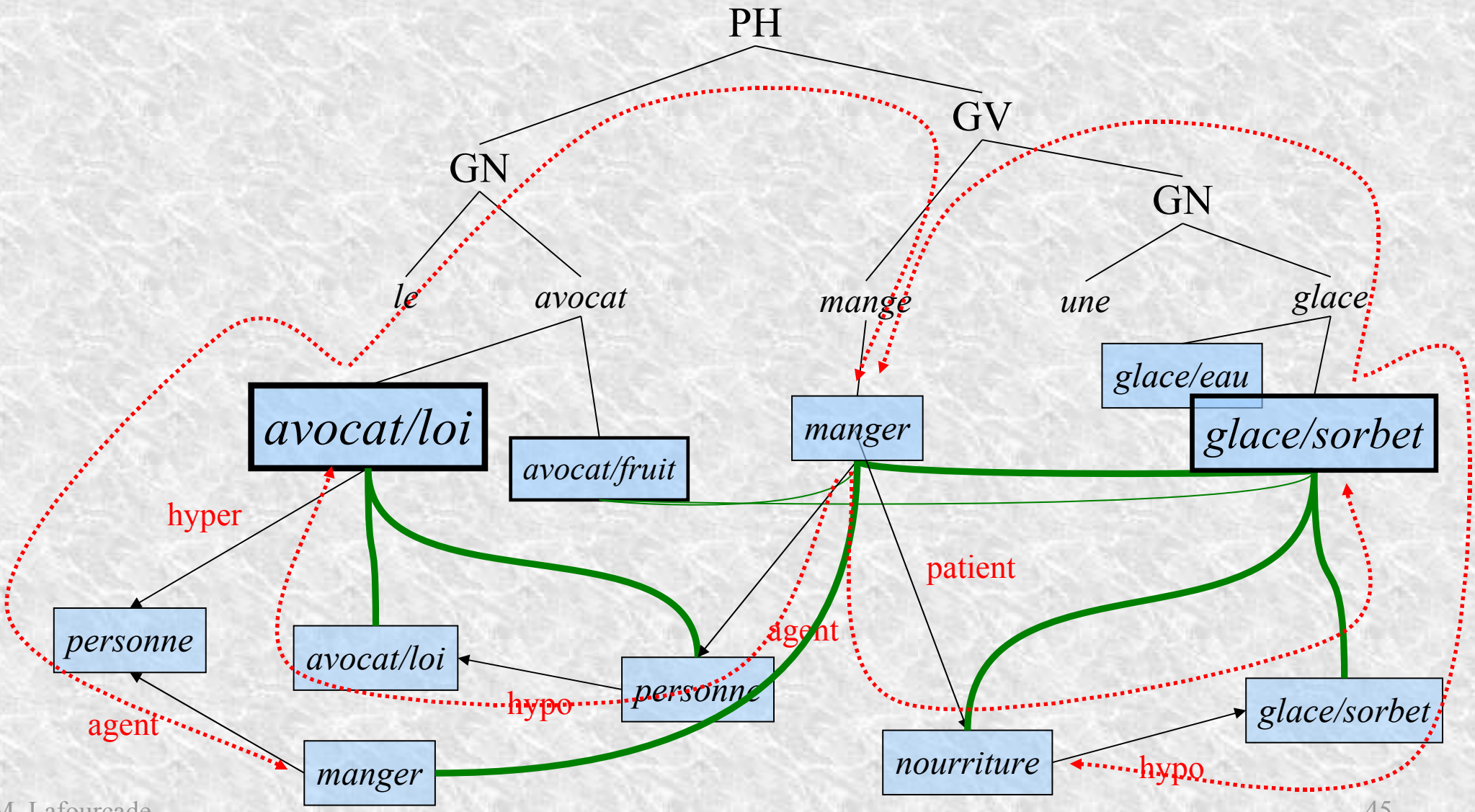
# Ant algorithms

## Example with lexical network



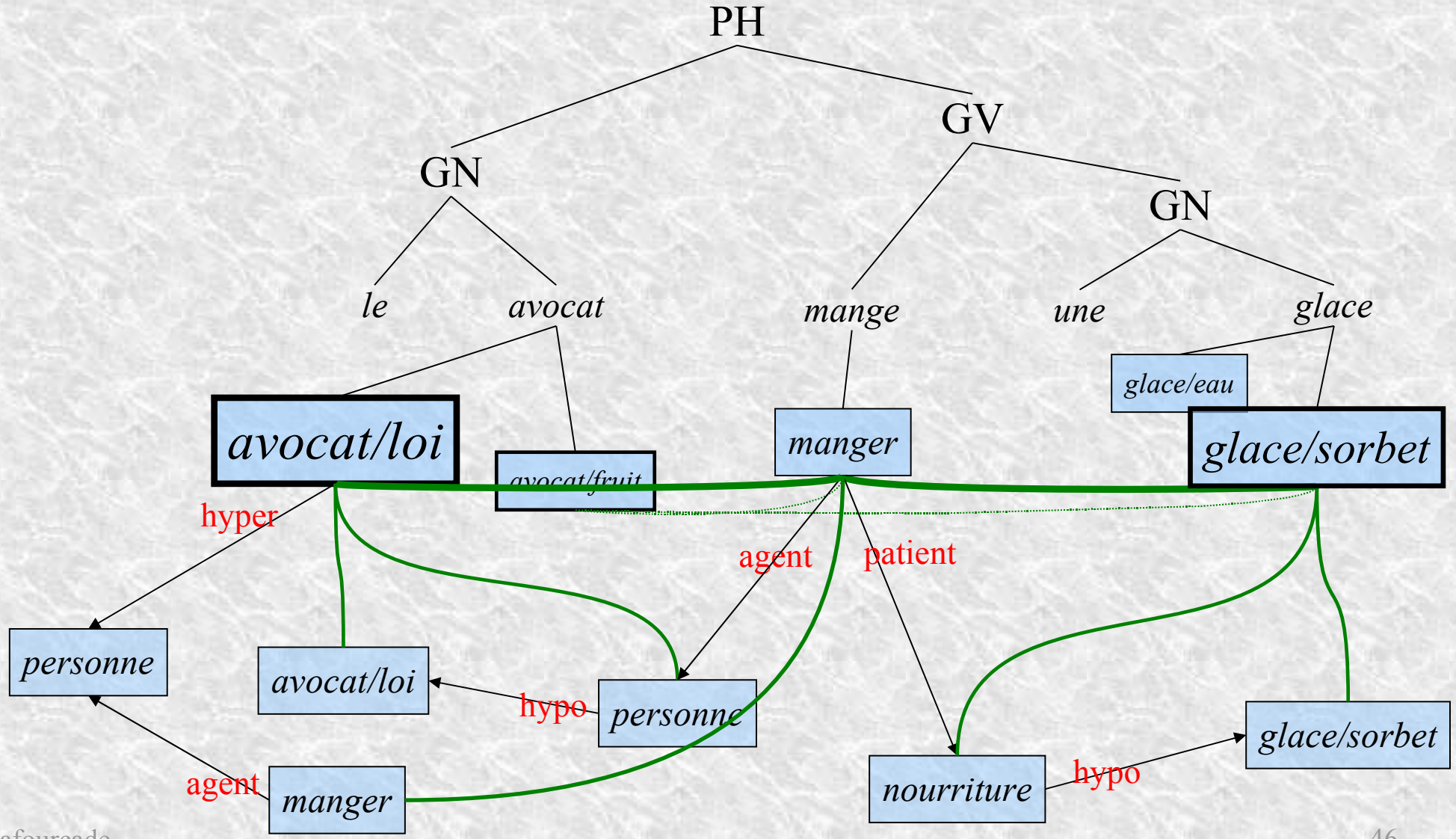
# Ant algorithms

## Example with lexical network



# Ant algorithms

## Example with lexical network



# Lexical Network

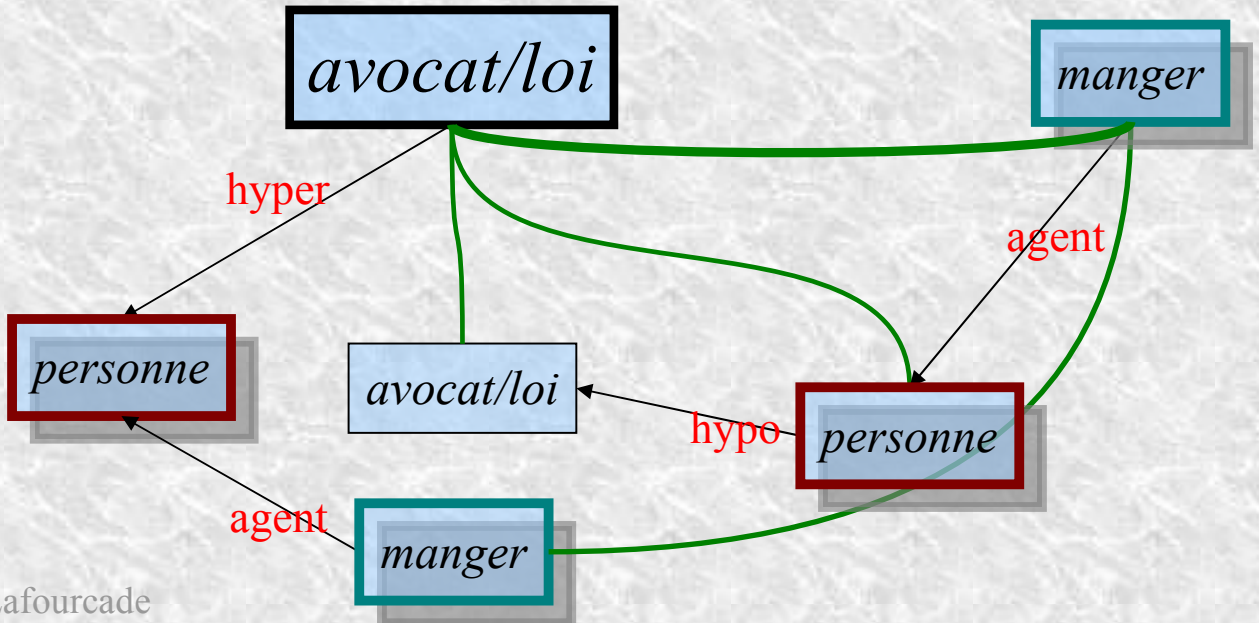
## Node fusion

### Node fusion between identical nodes

same label (simplified view here)

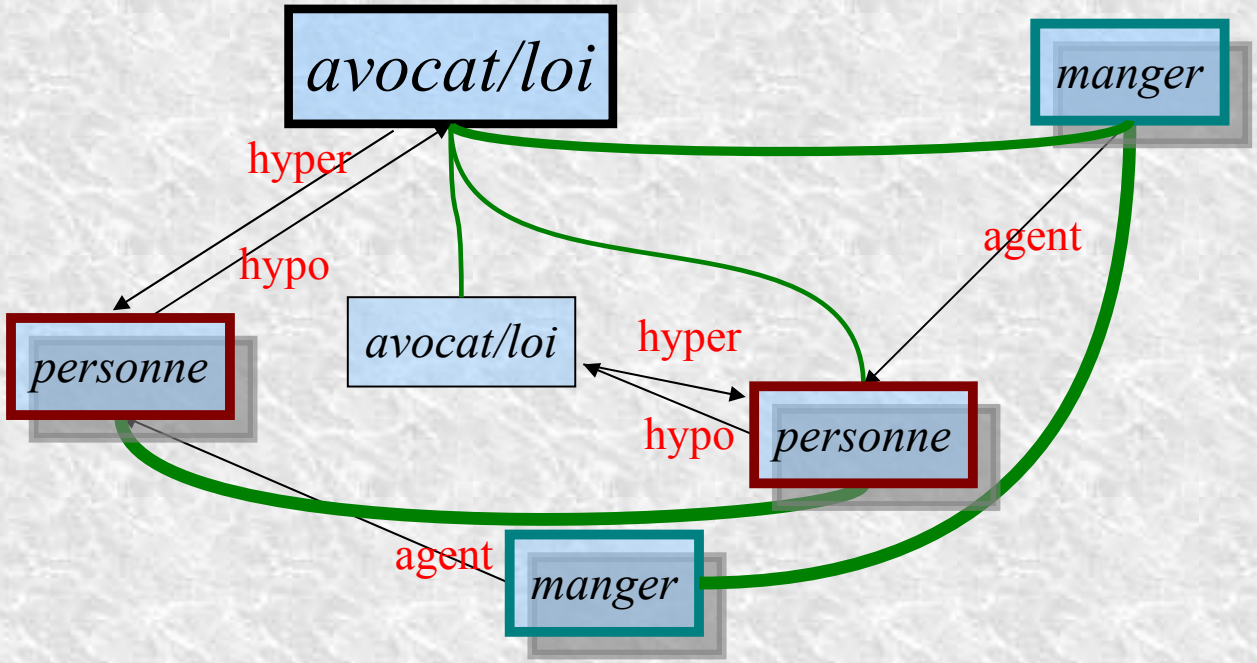
... if they are not too far

... if they are similar regarding their relations



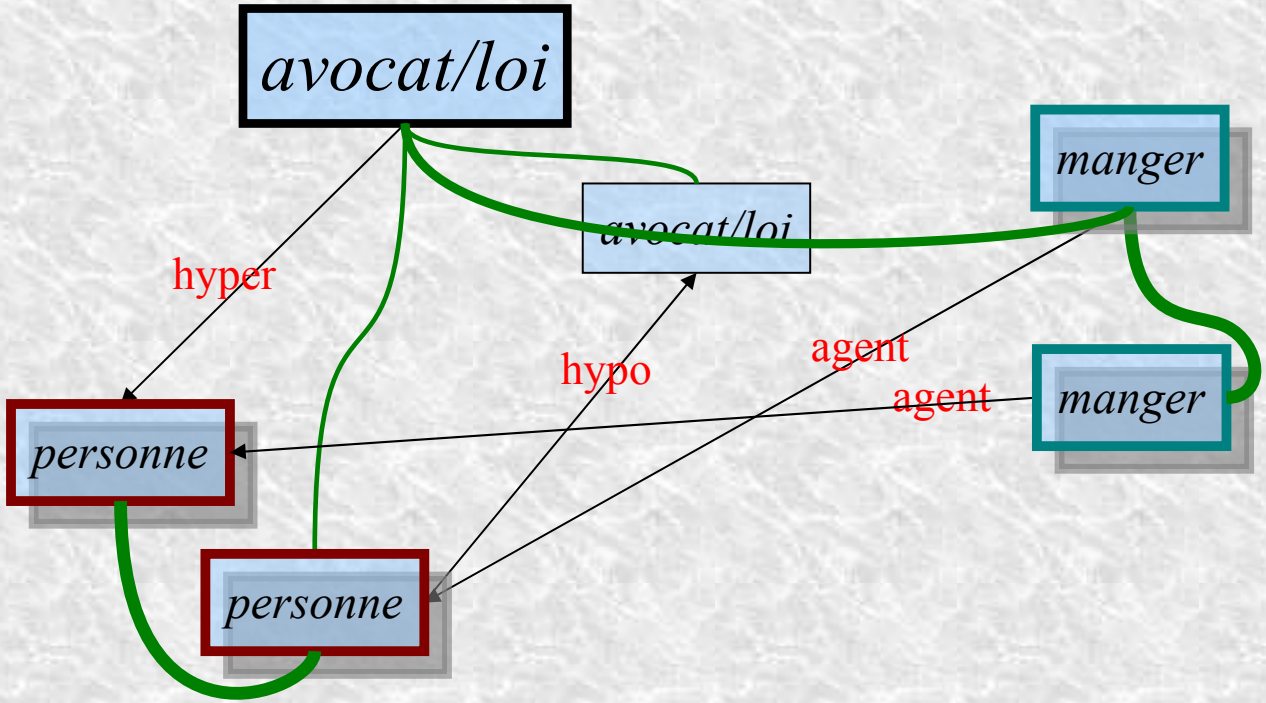
# Lexical Network

## Node fusion



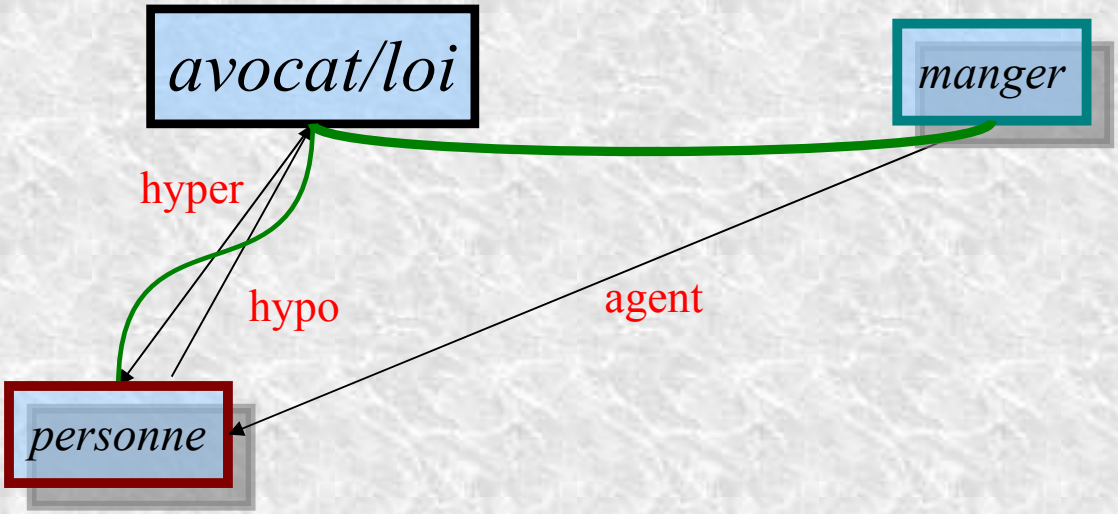
# Lexical Network

## Node fusion



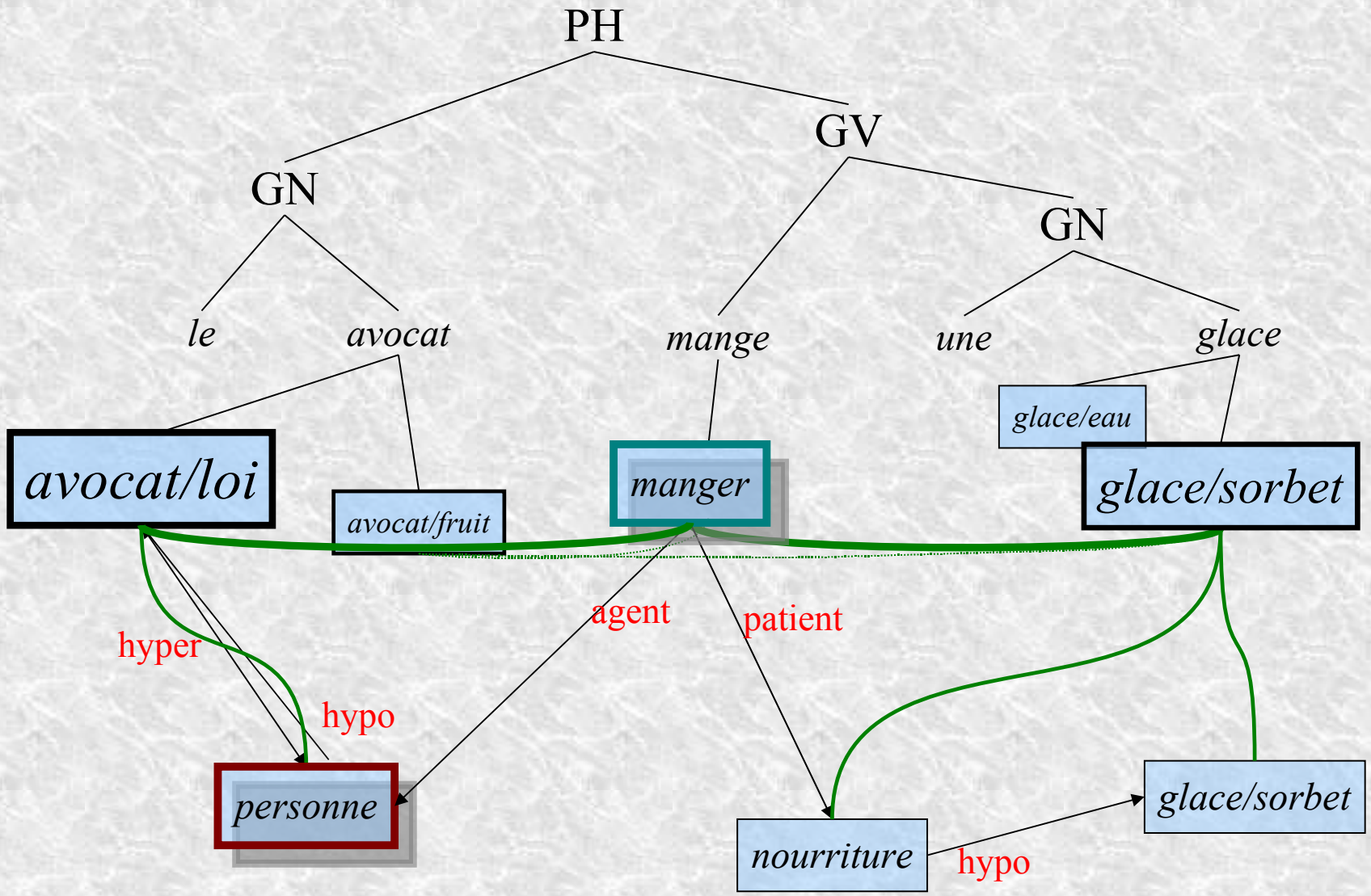
# Lexical Network

## Node fusion



# Ant algorithm

Finally...



# Adding a lexical network with

hyponym/hyperonym
part-of/whole-of
agent/patient/instrument
typical location

# leads to proper WSD in 87% of the cases

Still room for improvement

# Outline

| | |
|---|---|
| Semantic Analysis | What to look at? – Applications |
| Casting | Conceptual Vectors – Analysis Trees – Lexical Network |
| Ant algorithm | Principles - Without lexical network - With lexical network |
| Ressource Production | Lexical Network - Conceptual Vectors |
| Future directions | Toward holistic analysis: syntax + semantic |

# Lexical Network Construction

## Hand made

WordNet – EurowordNet – HowNet

=> precise but not much recall

=> normative

## Automatic Extraction

... (many)

=> many errors mainly due to polysemy
and bad relation identification

There is room for improvement

# Lexical Network Construction


http://www.lirmm.fr/jeuxdemots

There is room for education

# Conceptual Vectors Construction

With a given set of concepts (Thesaurus Larousse – Roget – ... )

=> problem of lexical density

From human usage dictionary, thesaurii, ...

=> problem of WSD and word relation identification

=> problem of saliency

Give up set of concepts => ask only for the set size
Use more explicit ressources => lexical network

# Conceptual Vectors Construction

Emergence - infinite iteration through the lexical network nodes

## Step 1 – Agglomeration

Spring attraction model

V(N) = Weighted means of neighbour vectors $v_i$

If $v_i = 0$ then $v_i$ = random

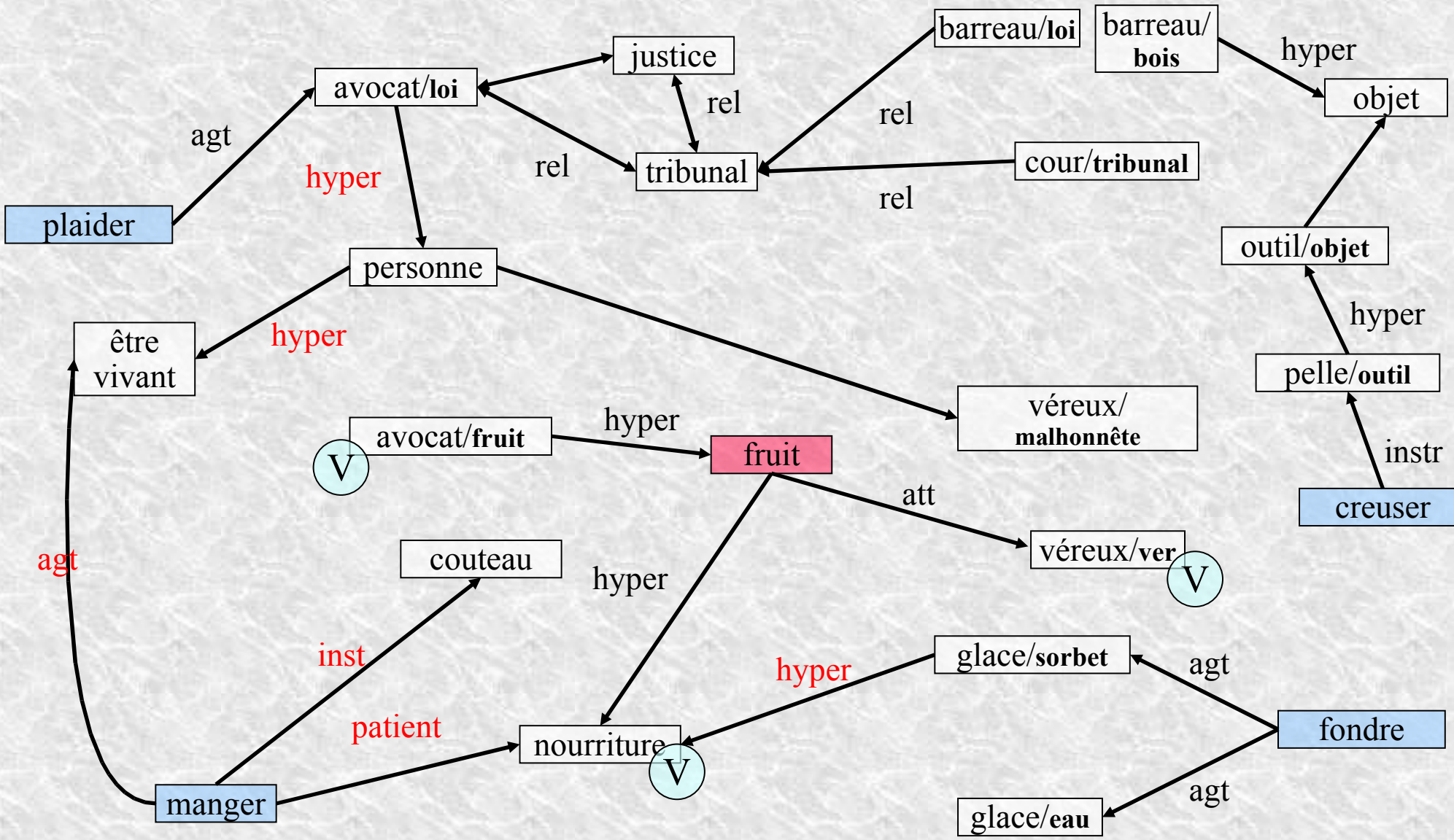
## Step 2 – Separation

Repulsive particular model

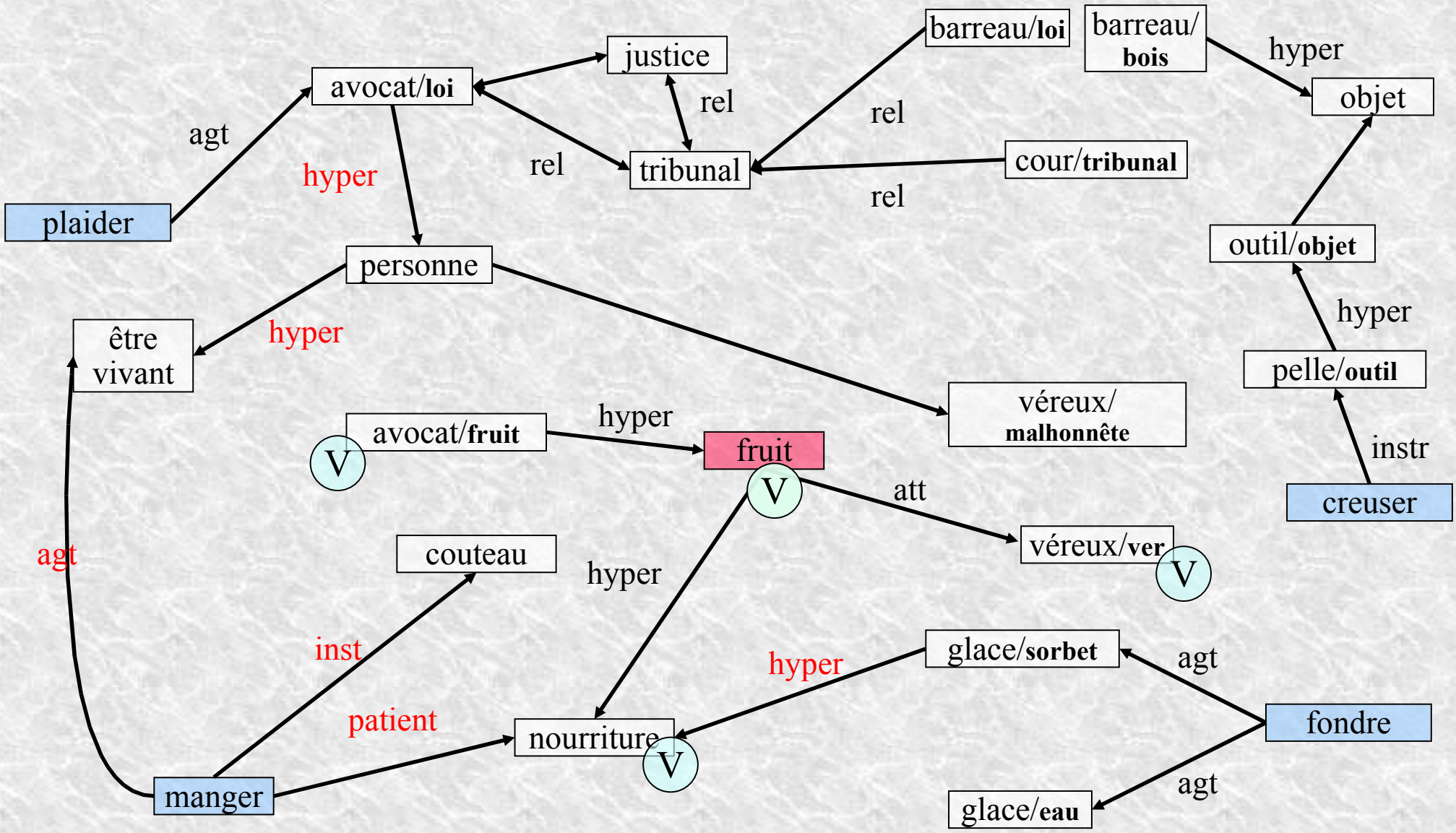
Select k closest neighbours according to $D_A$

Compute and apply a repulsive vector with a $1/d^2$ rule

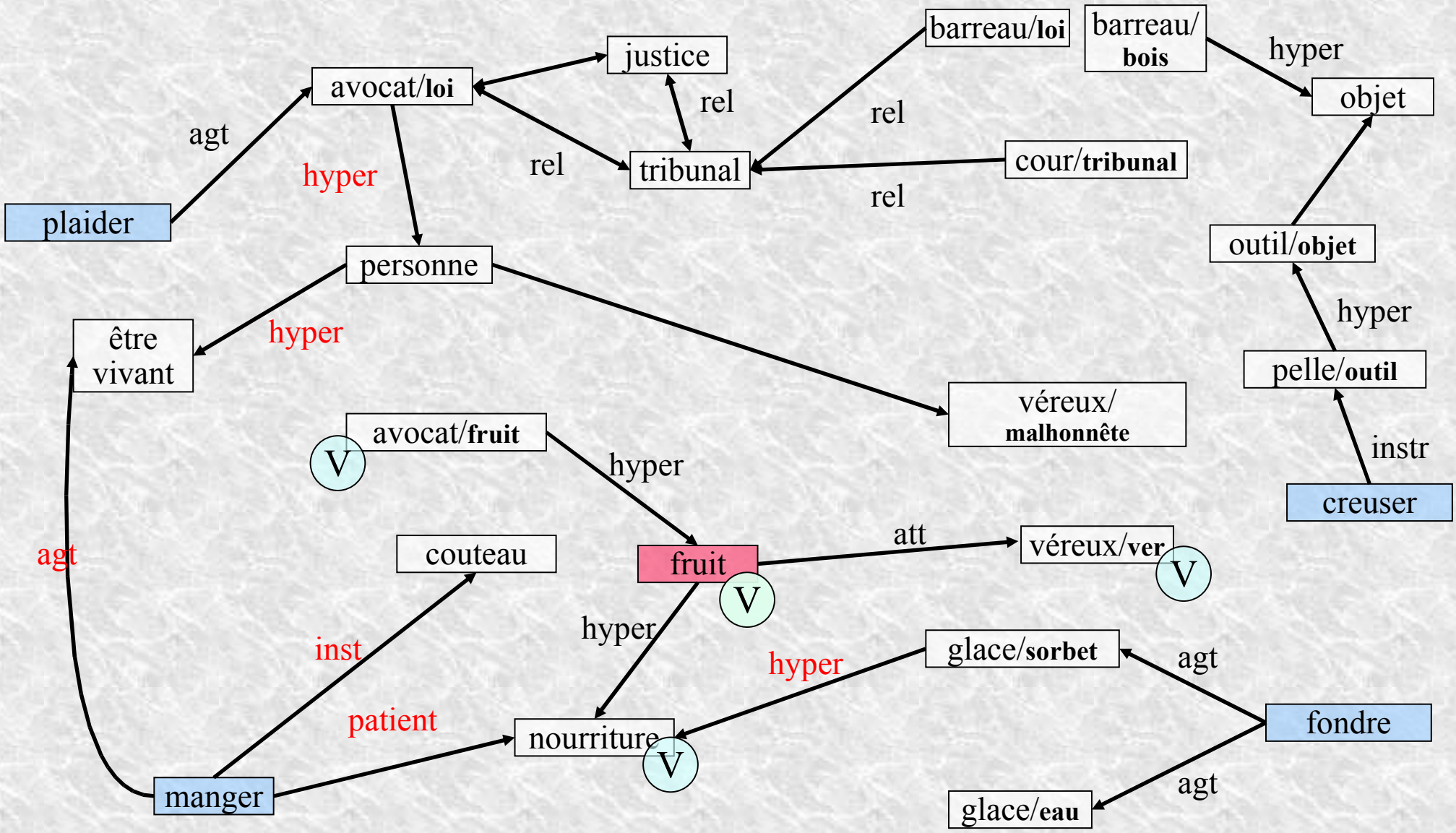
# Conceptual Vectors Construction



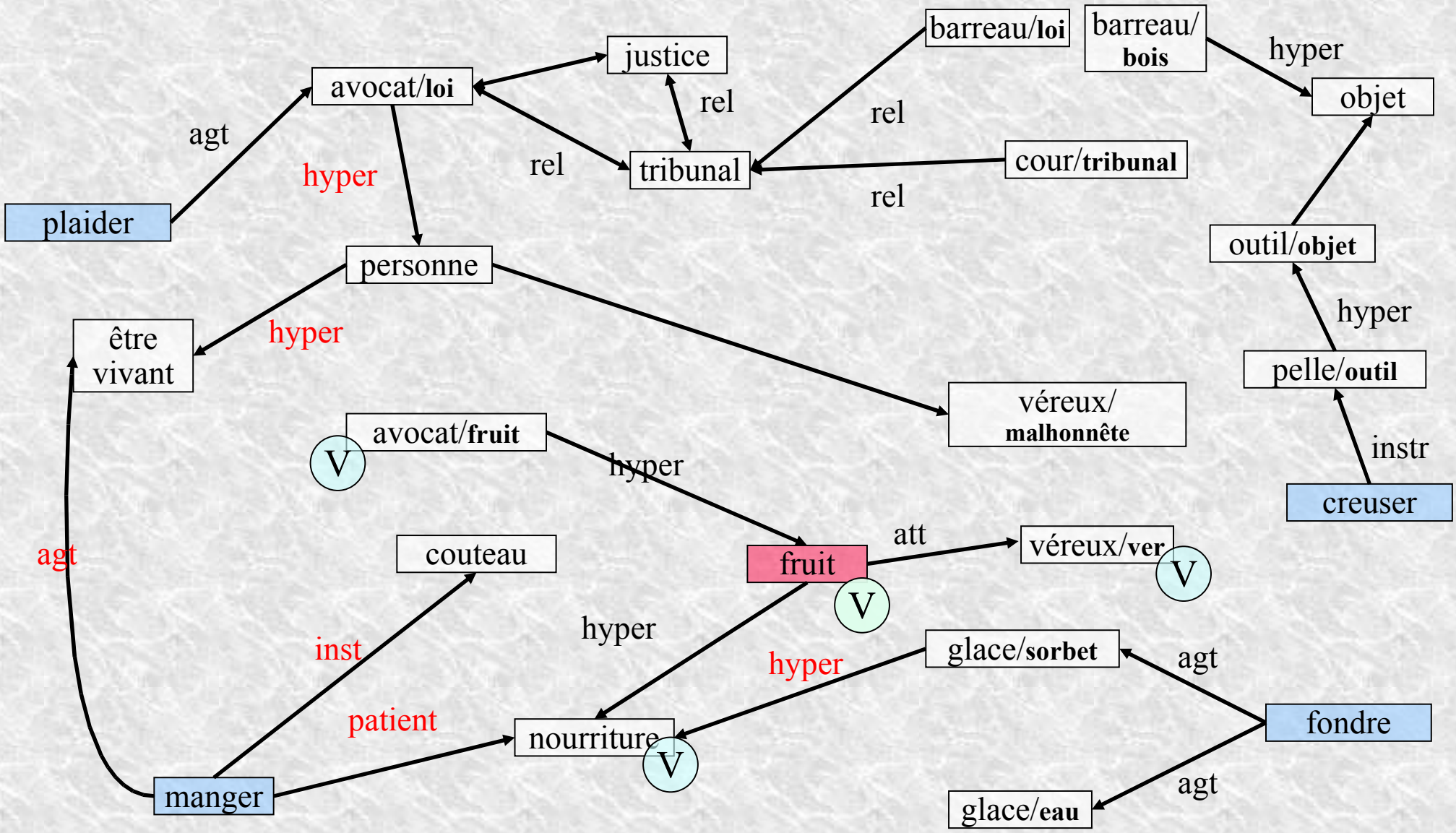
# Conceptual Vectors Construction



# Conceptual Vectors Construction



# Conceptual Vectors Construction



# Ant algorithm

More...

## Production ratio between ant castes

Adjustment through rewards => increase if successfully
=> decrease toward equiprobability with time

## Self termination

Internal factors (≠ simulated annealing)

«freezing ants» - narrow the gap between min and max of pheromones on a given arc

Produced when a stability is perceived locally => contamination

## Inhibition

«killing ants» - alert pheromone –
kill foe ant, compromising the durability of their anthill

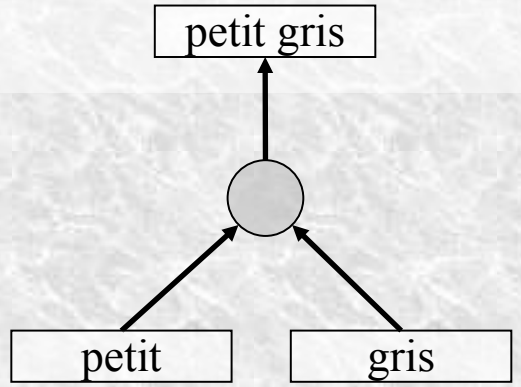
# Outline

| | |
|---|---|
| Semantic Analysis | What to look at? – Applications |
| Casting | Conceptual Vectors – Analysis Trees – Lexical Network |
| Ant algorithm | Principles - Without lexical network - With lexical network |
| Ressource Production | Lexical Network - Conceptual Vectors |
| Future directions | Toward holistic analysis: syntax + semantic |

# Future directions

## Taking into account frequent schemas

Prepositional phrase attachment [gala, lafourcade]




Acceptation distribution

over the general domain

over a specific domain

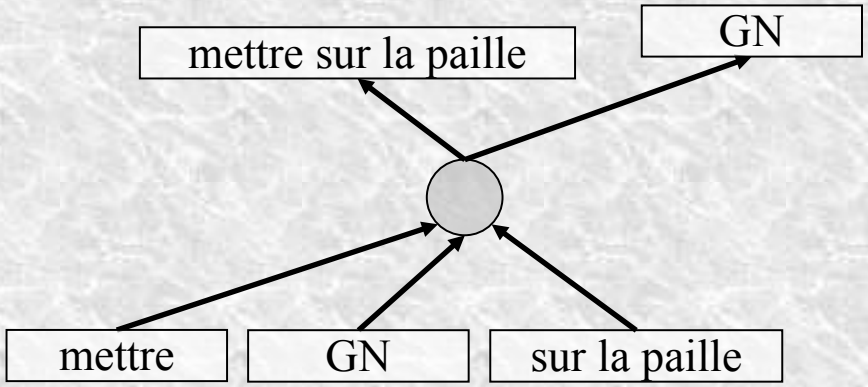
a squirrel – a paintbrush
a mushroom
a snail
a commuter train

Frequency for conpound terms meanings

petit gris - serpent de mer - ...

Frequency for locution meanings

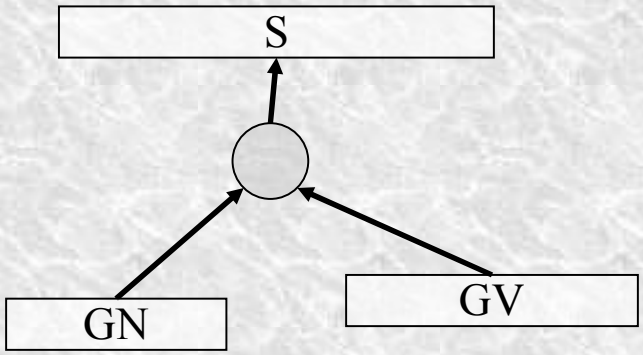
mettre *x* sur la paille

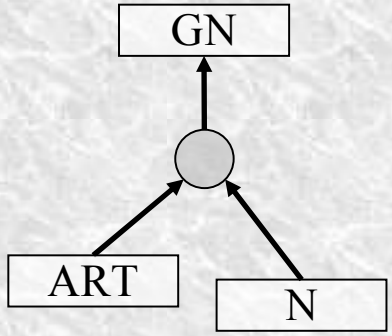


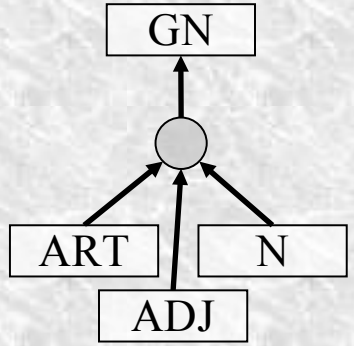

to ruin/bankrupt x
to put x on the hay

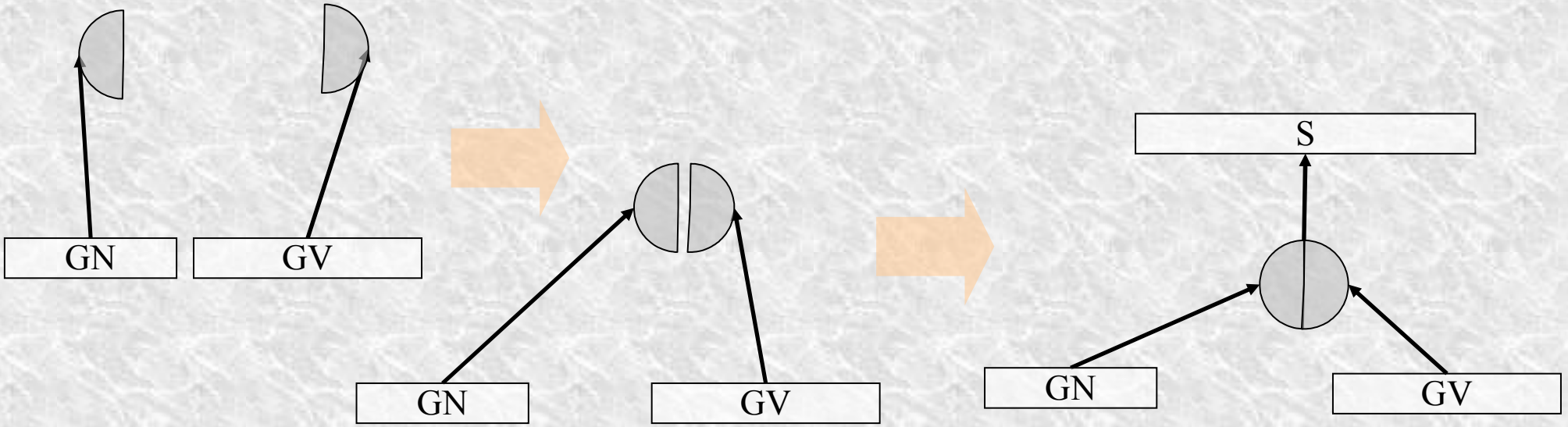
# Future directions holistic analysis

## Syntactic schemas in the lexical network



# Future directions holistic analysis

## All nodes

- produce information
- request information
- Produce agents for these tasks

## Information

- Deduced from the lexical network
- Symbolic (MAS, SING, V, GV, ...)
- Vectors (different types according to network relation labels)

## Ants as agents

- Information carriers
- Structure builders

# Conclusion

Holistic analysis

<= what is important and when?

Emergence of some solutions through errors

<= perfect ants would produce nothing

Coordination problem facilitated

<= no central control but population control

Generic and extensible architecture

<= adding new information type is (relatively) easy

Tuning is difficult => can be assessed through experiment