Introduction to Vectorial representations for NLP

Didier Schwab (Didier.Schwab@imag.fr)
LIG-GETALP

Vectors to represent Meaning

- Basically, integer/double vectors may permit to represent meaning
  - [0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
  - [1,0,0,0,1,0,0,0,0,1,0,0,0,0,0,1,0,1,0,1,0,0]
  - [0,0.24,0,0,1,0,0,0.12,0,0,0.25,0,0,0,0.9,0.8,0,0.6,0]
  - [0.1,-0,2,0.3,0,1,-0.8,0.7,0.1,0.5,-0.5,0.8,0.3,0.2,-0.3]
  - [843,900,1045,24,234,123,983,813,452,574,276]
- Meaning of
  - Words
  - Sentences
  - Texts
  - …

Types of vectors

- Two types of vectors inspired by two linguistic theories
  - Distributional linguistics
  - Componential linguistics

Distributional linguistics

- Represents linguistic objects with the associability possibilities they share or not
- Linguistic items with similar distributions have similar meanings
- « You shall know a word by the company it keeps » (John Ruppert Firth, 1957)
- Meaning of a word is represented with all contexts where it can be find in texts.
  - Milk : {cow, milk, white, cheese, mammal,…}
  - Computer{school, electronic, machine, programmable,…}
- Distributionnal vectors

Distributional Vectors

- Built from corpora
- Each component corresponds to words in a corpus
  - Directly : Saltonian vectors
  - Indirectly : Latent Semantic Analysis, word embeddings

Componential Semantics

- Represent linguistic objects with semantic components (primitives, primes [Wierzbicka], constituents [Greimas], attributes, semes, ideas,…)
- Examples :
  - man : [+ male], [+ mature]
  - woman : [– male], [+ mature]
  - boy : [+ male], [– mature]
  - girl : [– male] [– mature]
  - Child : [+/– male] [– mature]

Componential Vectors (Idea Vectors)

- Each Component corresponds to ideas
  - Directly : Semantic Vectors [Chauché] 1992->2005
  - Indirectly : Conceptual Vectors [Lafourcade] 1999 → ?
- Some experiments about Conceptual vectors
  - How to build lexical bases and process semantic analyses ?

Text Semantic Analysis

Identification/resolution of a set of semantic phenomena

Computable representations

Thanks to Lexical Functions

« *Jack gave me a precious advice*. »

Bon

« He *saw the girl with a telescope*. »

Instr

« *John had a strong fear*. »

Magn

« The *cat climbed onto the chair. The animal began to sleep*. »

Gener

Lexical Functions

LF formalise linguistic relations between terms [Mel'čuk]

Paradigmatic LF (semantic relations)

synonymy	Syn('*plane*') = '*airplane*', '*aeroplane*', ...
antonymy	Anti('*uncertain*') = '*certain*', '*sure*'
generic	Gener('*trout*') = '*fish*' Gener('*siakap*') = '*fish*'
	Gener('*dog*') = '*animal*' Gener('*cat*') = '*animal*' ≠ '*mammal*'

Syntagmatic LF (collocations)

intensification	Magn('*fear*') = '*numbing*', '*strong*'
	Magn('*love*') = '*tremendous*', '*big*'
laudative	Bon('*advice*') = '*precious*', '*good*'
	Bon('*choice*') = '*fortunate*', '*good*'
confirmation	Ver('*argument*') = '*valuable*', '*admissible*'
	Ver('*fear*') = '*justified*'

Semantic Analysis

1) Lexical ambiguity

« *The mouse is eating the cheese.* »

mouse/computer or mouse/animal ?

2) Interpretation paths

« *The sentence is too long.* » 2 probable interpretations, not 4

sentence/phrase **sentence/condemnation** **long/duration** **long/length**

Semantic Analysis

3) Reference

Anaphora resolution

« *The cat climbed onto the seat, then it began to sleep.* »

Identity relations

« *The cat climbed onto the seat. The animal began to sleep.* »

4) Prepositional attachments

« *He saw the girl with a telescope.* »

Applications

Information Retrieval

Direct effects (equality of values)

« *numbing fear* » ≡ « *strong fear* »
« *vast majority* » ≡ « *strong majority* »
« *The cat has gone* » ≡ « *The tabby has gone* »
« *This number is not even* » ≡ « *This number is odd* »

Indirect effects (lexical ambiguity, prep attach, references)

⇒ precision +, recall +

Machine Translation

Direct effects (lexical transfer)

« *grosse fièvre* » = « *high fever* »
« *grosse pluie* » = « *heavy rain* »
« *L'appareil s'est posé.* » ≡ « *The plane has landed.* »

Indirects Effects on the overall phenomena

Semantic Lexical Base

Modelling lexical functions

Three problems

- Discovery of as many lexical items as possible
- Acquisition of information about their meanings
- Fabrication of lexical objects representing these meanings

Three questions

- How to represent meaning?
- How to compute it?
- How to obtain a generic and evolutive system?

Which hypotheses have we taken?

Hypothesis I

Hybrid representation of the meaning

Hypothesis I

For the lexical objects

Lexical functions (discrete, symbolic connectionnist)

- modelling relations between lexical objects

Internal information

- symbolic
  - Morphology (*noun, adj, verb, masc, fem, ...*)
  - etymological information, level of language, field, ...
- numeric
  - usage frequency
- vectorial
  - thematic information (conceptual vectors)

Conceptual Vectors

- Thematic representation [Chauché, Lafourcade]
  - Lexical item = Ideas = Conceptual Vector
  - For example, 873 component (concepts from Larousse thesaurus)
    - (1) existence, (2) inexistence, (3) matérialité, …, (516) liberté, …, (872) jeux, (873) jouets
  - A vector component corresponds to the activation of a concept.
- V taken from a thesaurus hierarchy (Larousse)
  - translation of Roget's thesaurus, 873 leaf nodes
  - the word 'peace' has non zero values for concept PEACE and other concepts

Our conceptual vectors

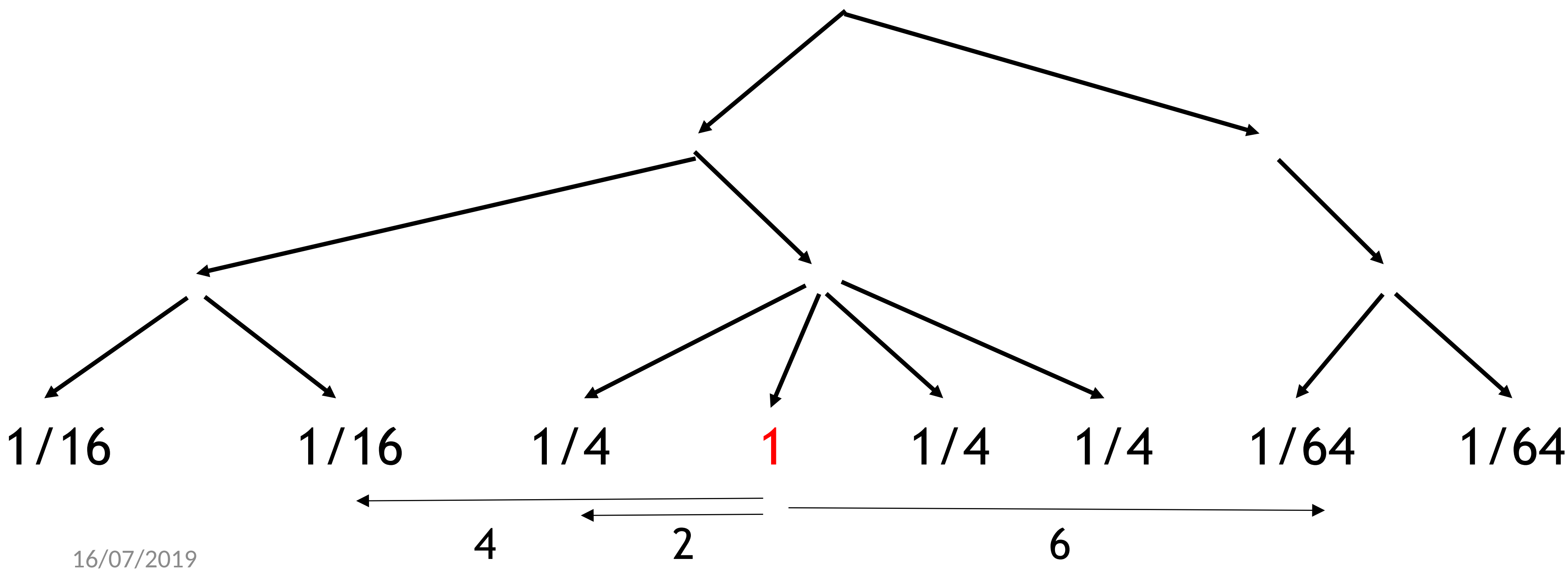
Thesaurus

- H : thesaurus hierarchy — K concepts

  Thesaurus Larousse = 873 concepts (leafs)

- $V(C_i)$: $\langle a_1, \ldots, a_i, \ldots, a_{873} \rangle$

  $a_j = 1/\ (2\ \hat{}\ D_{um}(H, i, j))$

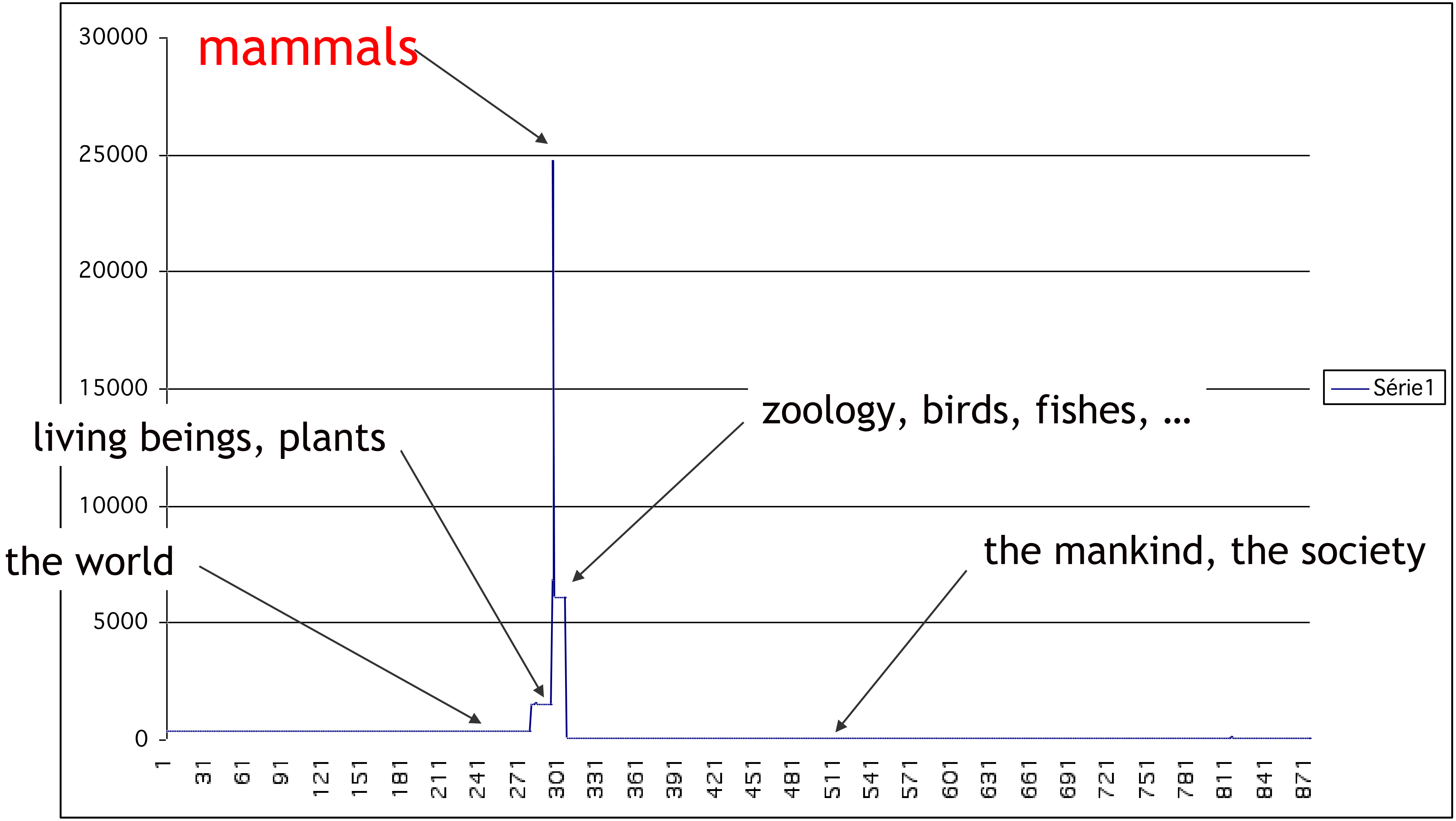


Vector construction
Concept vectors

- C : mammals
  - L4 : zoologie, mammals, birds, fish, …
  - L3 : animals, plants, living beings
  - L2 : … , time, movement, matter, life , … ,
  - L1 : the society, the mankind, the world

Vector construction
Concept vectors

mammals



Vector construction
Term vectors

- Example :   cat (chat)
  - Kernel
    - manually built : relevent vectors

c:mammal (mammifère), c:stroke (caresser)

$$\nu(\text{mammal}) + \nu(\text{stroke})$$

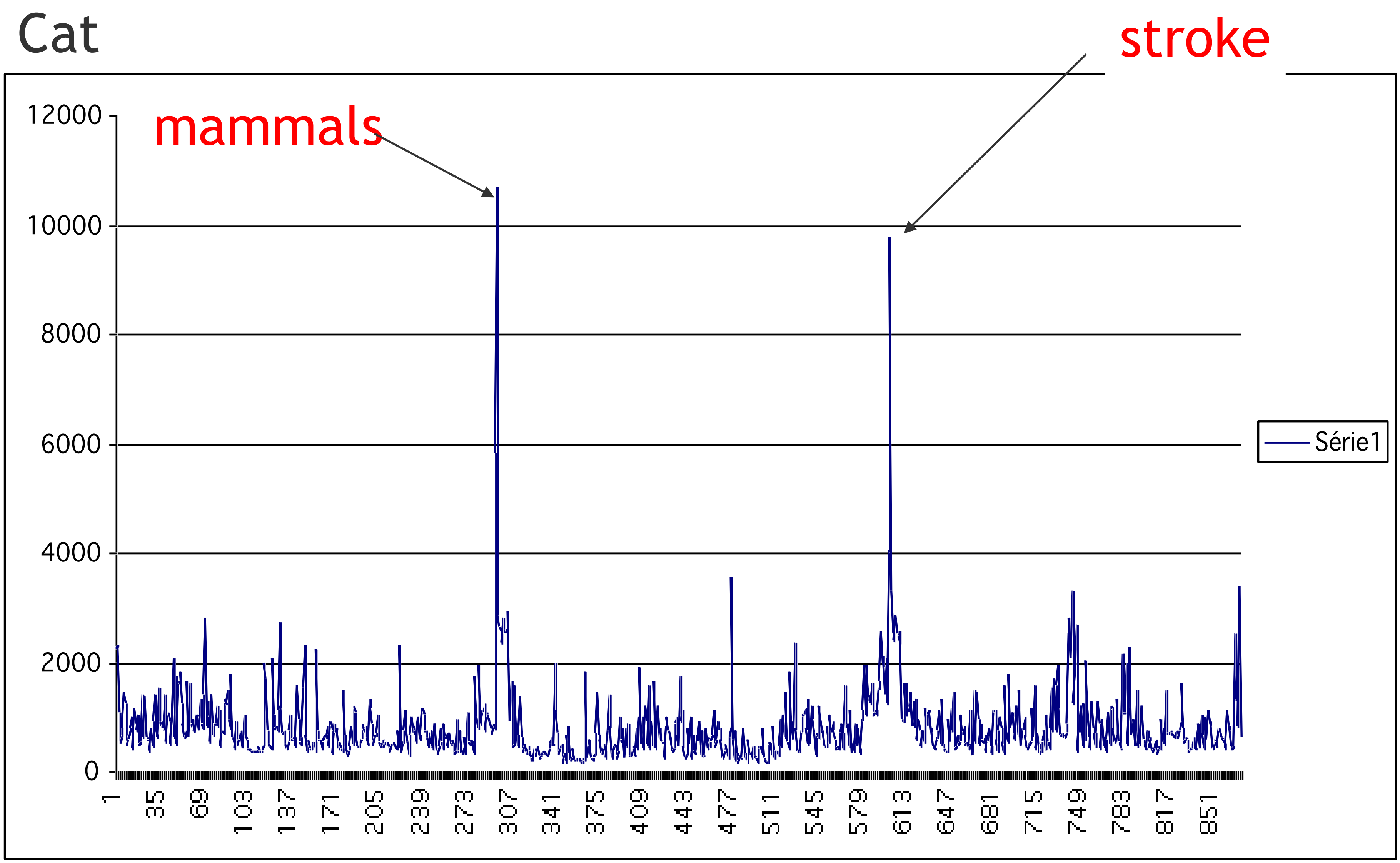
- Augmented with weights

  c:mammal, c:stroke, 0.75*c:zoology, 0.75*c:love …

  $\nu(\text{zoology}) + \nu(\text{mammal}) + 0.75\ \nu(\text{stroke}) + 0.75\ \nu(\text{love})$ …

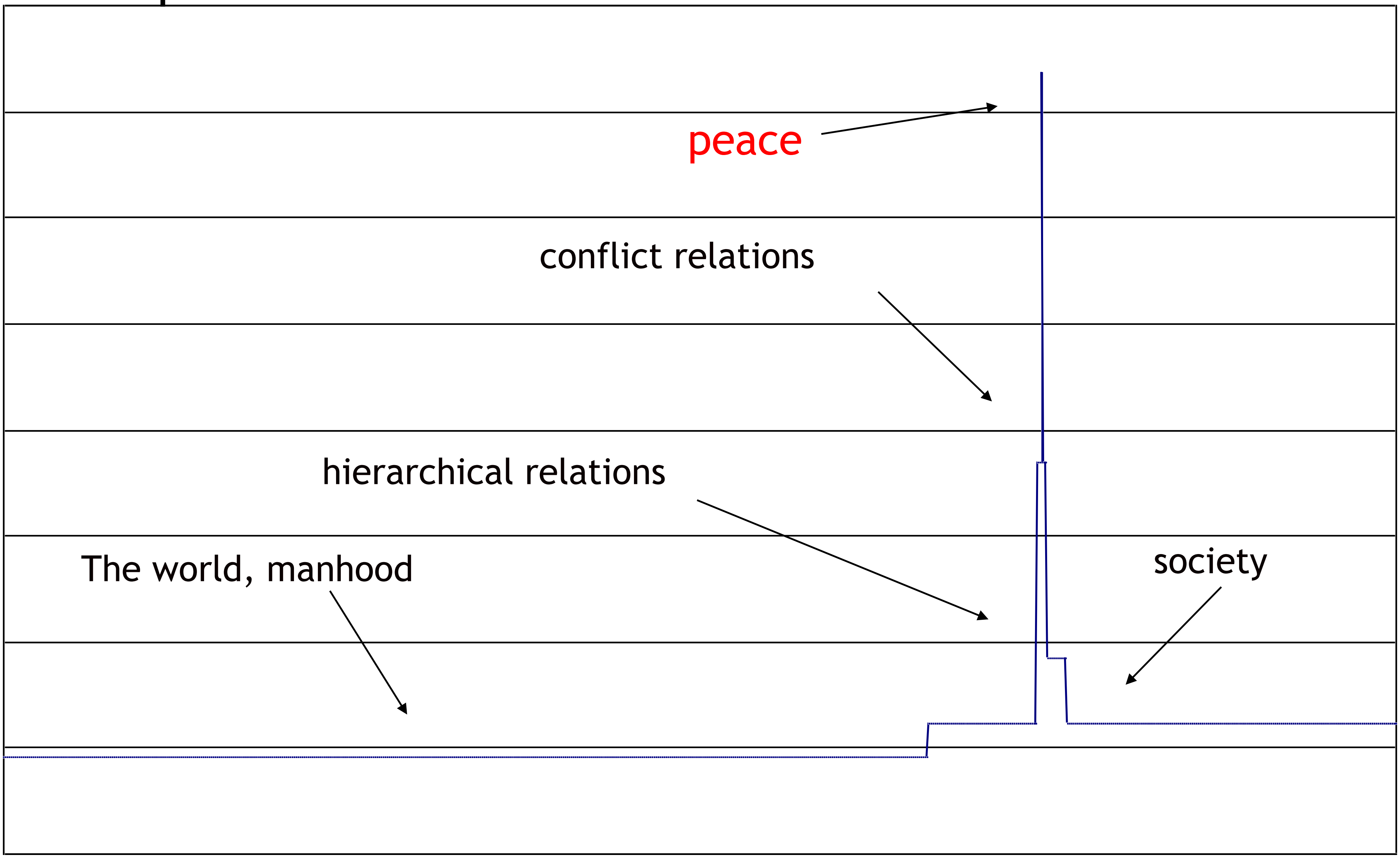
  - Learning phase

Vector construction
Term vectors

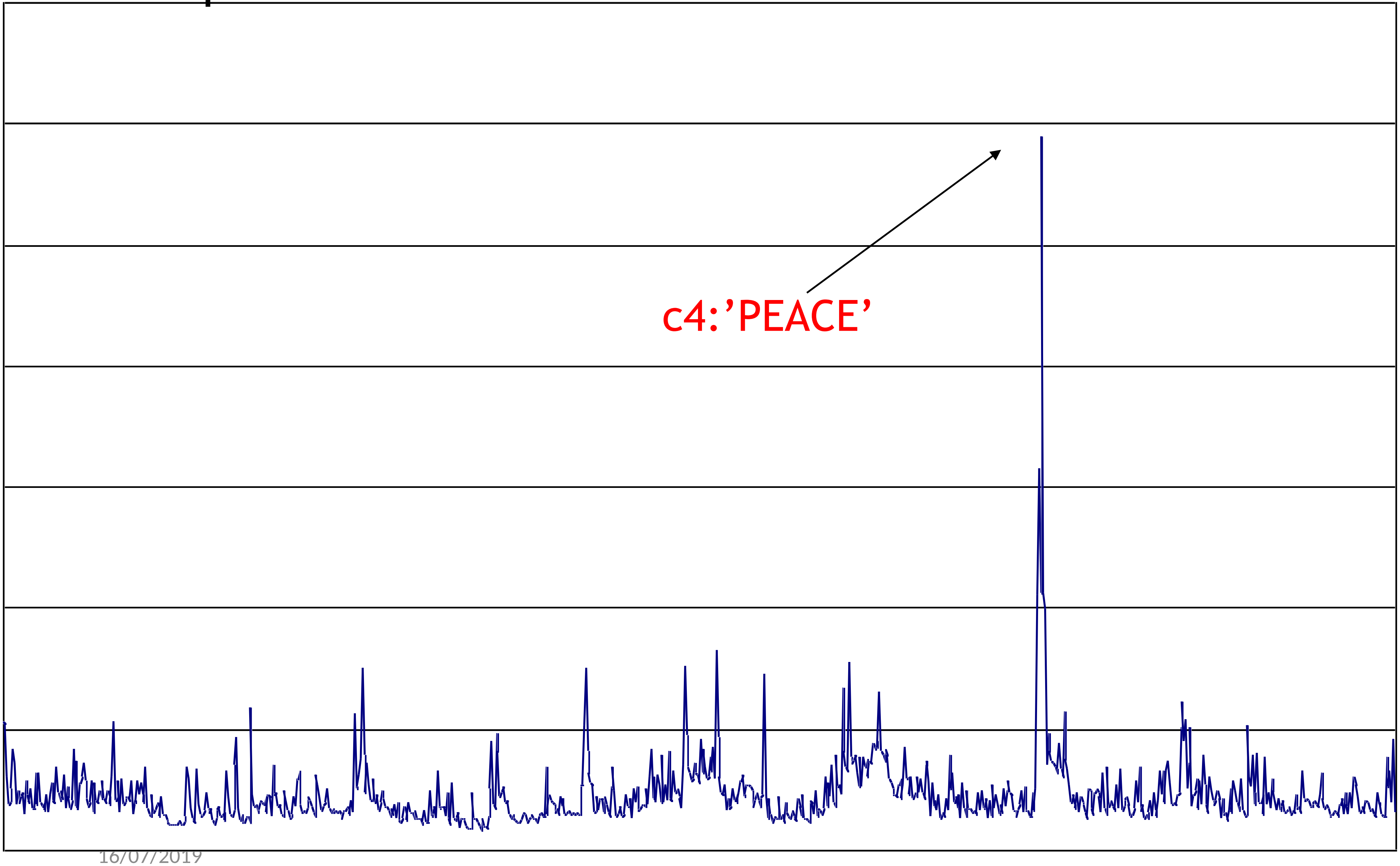


Conceptual vectors

Concept c4: ‘PEACE’



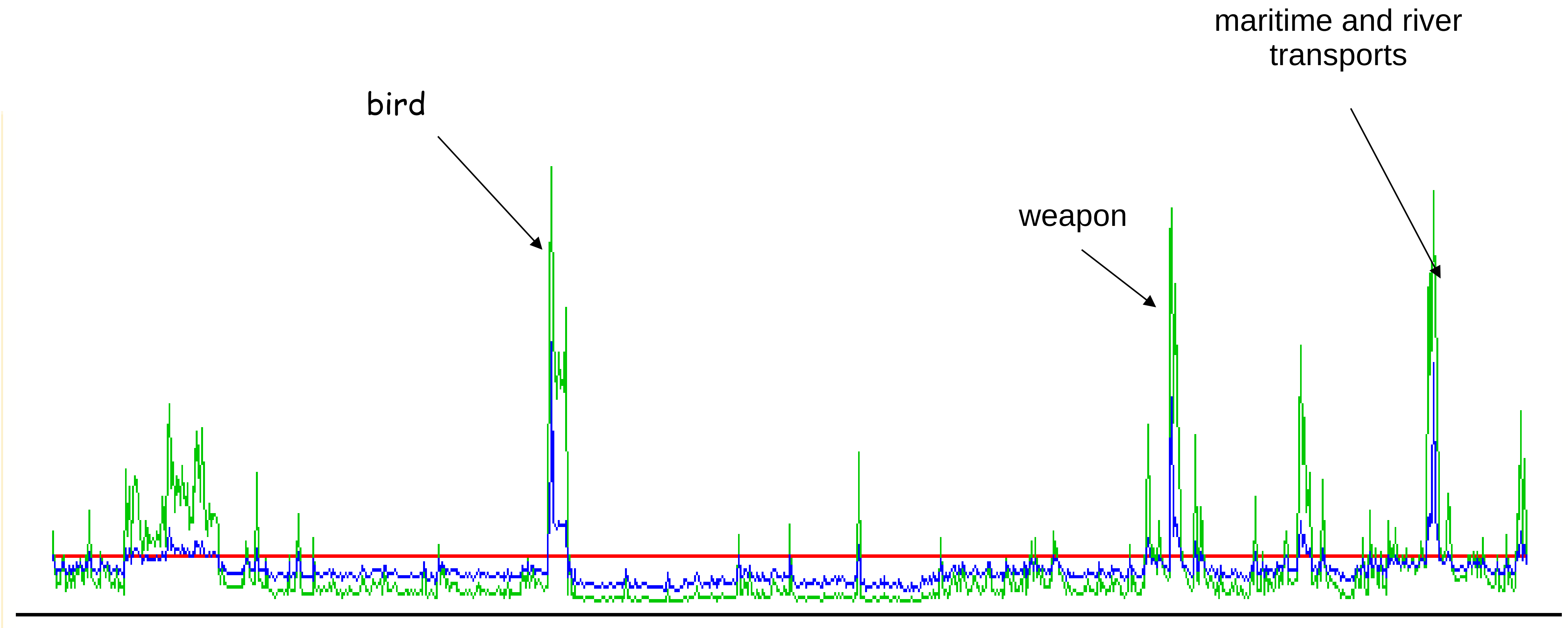
Conceptual vectors
Term “peace”



Conceptual Vectors

Conceptual vector of *frégate*

(polysemic : frigate/frigatebird)



Conceptual Vectors

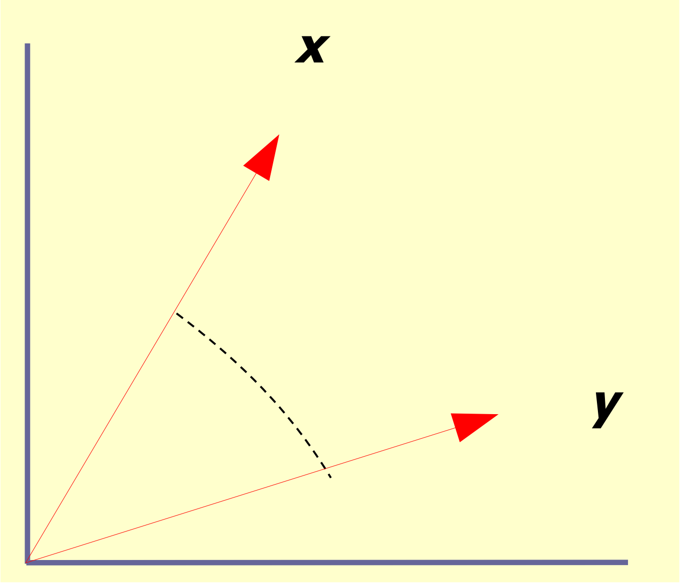
Thematic distance

$$D_A(x,y) = angle(x,y) = \arccos(similarity(x,y)) = \arccos\left(\frac{x.y}{|x||y|}\right)$$

$$0 \leq D_A(x,y) \leq \frac{\pi}{2} \text{(positive components)}$$

if 0 then x and y are collinear : same idea

if $\frac{\pi}{2}$: nothing in common



Conceptual Vectors

Thematic distance (examples)

D_A('*anteater*', '*anteater*') = 0 (0°)

D_A('*anteater*', '*animal*') = 0.45 (26°)

D_A('*anteater*', '*train*') = 1.18 (68°)

D_A('*anteater*', '*mammal*') = 0.36 (21°)

D_A('*anteater*', '*quadruped*') = 0.42 (24°)

D_A('*anteater*', '*ant*') = 0.26 (15°)

thematic distance ≠ ontological distance (*is-a*)

but thematic distance ⊃ ontological distance

Vector Proximity (Neighbourhoud)

- Function V gives the vectors closest to a lexical item
- Allow the database to be explored continuously
- V(life) = life, alive, birth…
- V(death) = death, to die, to kill…
- V(vie) = vie quotidienne , VIE, s'animer , demi-vie , survivant
- V(ranger) = trier , cataloguer , sélectionner , classer
- V (D_A,'*death*', 7)=('*death*', 0) ('*murdered*', 0.367) ('*killer*',0.377) ('*age of life*', 0.481) ('*tyrannicide*', 0.516) ('*to kill*', 0.579) ('*dead*', 0.582)

Operations

Vectors combinations

Operations $\Rightarrow$ reasonable linguististic interpretations

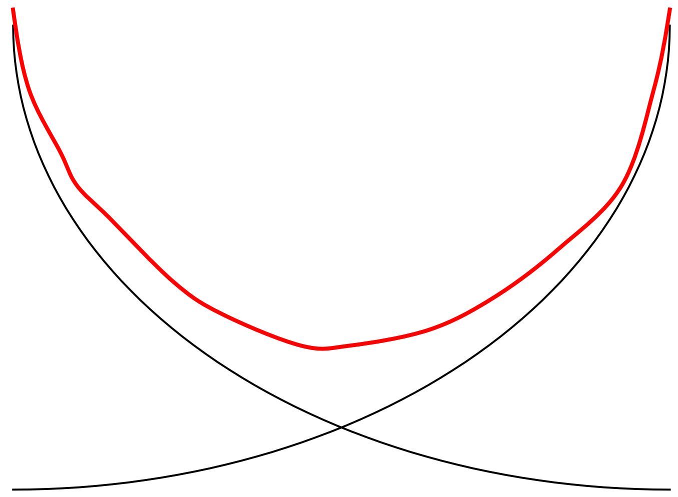
normalised sum $\oplus$: union of ideas

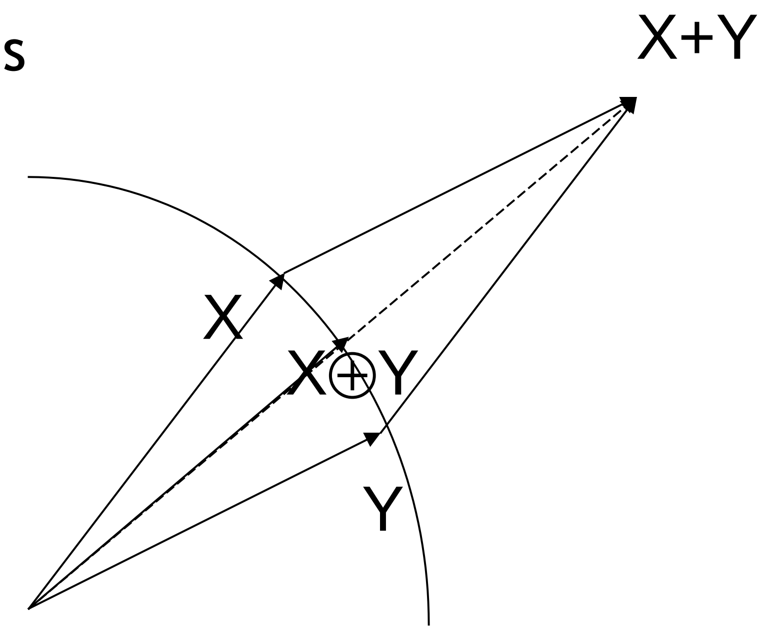
term to term product $\otimes$: intersection of ideas

week contextualisation : $\gamma(A,B) = A \oplus (A \otimes B)$

Vector operations

- Sum
  - $V = X + Y \Rightarrow V_i = X_i + Y_i$
  - Neutral element : **0**
  - Normalization of sum : $V_i \, / |X+Y|$
  - Average of normalized vectors
  - Interpretation : Union of ideas



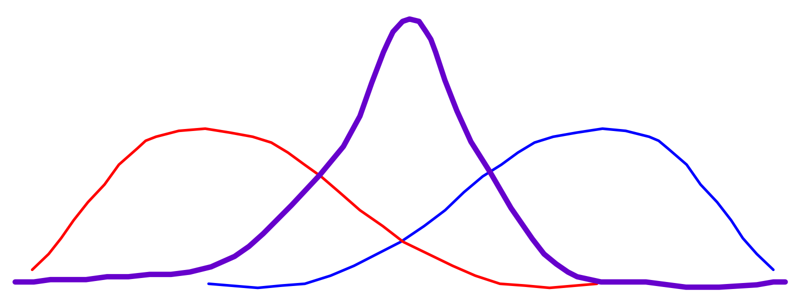


Vector operations

- Term to term product

$$V = X \otimes Y \Rightarrow \sqrt{X_i Y_i}$$

  - Neutral element : **1**
  - Interpretation : Intersection of ideas



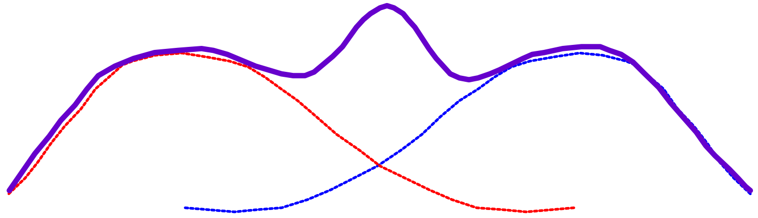
Kind of intersection

Vector operations

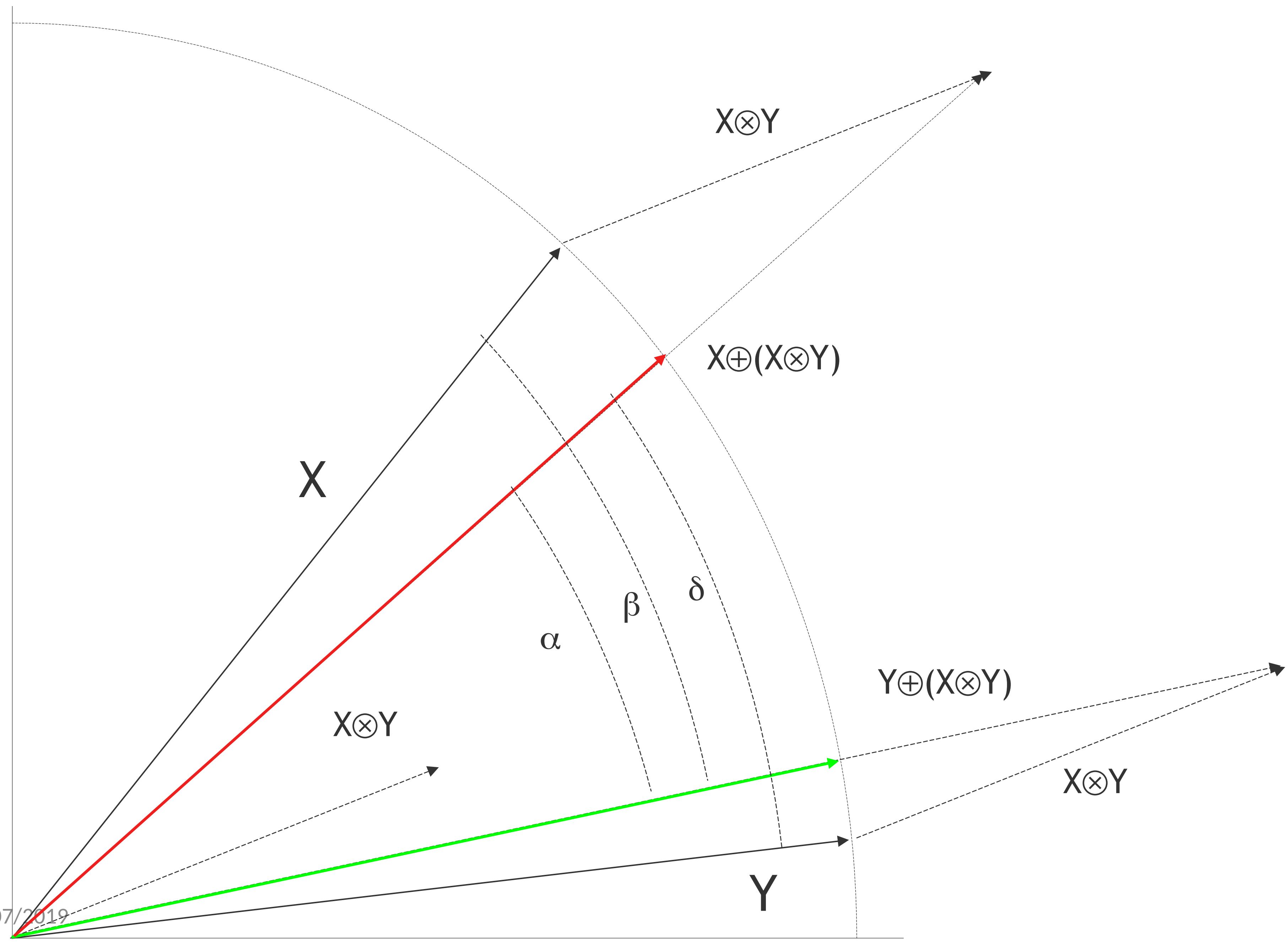
weak contextualisation Γ : Product + sum

$$Z = \Gamma(X,Y) = X + Y + (X \otimes Y)$$

- Z is X augmented by its mutual information with Y



2D view of weak contextualization



Vector operations

- Subtraction
  - $V = X - Y \Rightarrow v_i = x_i - y_i$
- Dot subtraction
  - $V = X \dot{-} Y \Rightarrow v_i = \max(x_i - y_i, 0)$
- Complementary
  - $V = C(X) \Rightarrow v_i = (1 - x_i/c) * c$
- etc.

Set operations

Hypothesis I

For the lexical objects

Lexical functions (discrete, symbolic connectionnist)

modelling relations between lexical objects

Internal information

- symbolic
  - Morphology (*noun, adj, verb, masc, fem, ...*)
  - etymological information, level of language, field, ...
- numeric
  - usage frequency
- vectorial
  - thematic information (conceptual vectors)

Why ?

Limitation of CV for lexical functions modelisation

paradigmatic

hyperonymy [Lafourcade et Prince, 2003]

synonymy (relative, subjective) [Lafourcade et Prince, 2001]

antonymies (complementar, scalar, dual)

[COLING'2002, JADT'2002, TALN'2002]

syntagmatic

collocations

Mixing high recall of CV to the high precision of relations

Cognitive model adequacy

3 areas in the brain

- area 1 : fabrication and classification of concepts
- area 2 : management of the language "surface" (syntax, lexical associations)
- area 3 : combination of information from the 2 other areas

Hypothesis II

Joint Usage of lexical objects of type ACCEPTION and LEXICAL ITEM

Hypothesis II

Lexical item, entrance point to the meaning

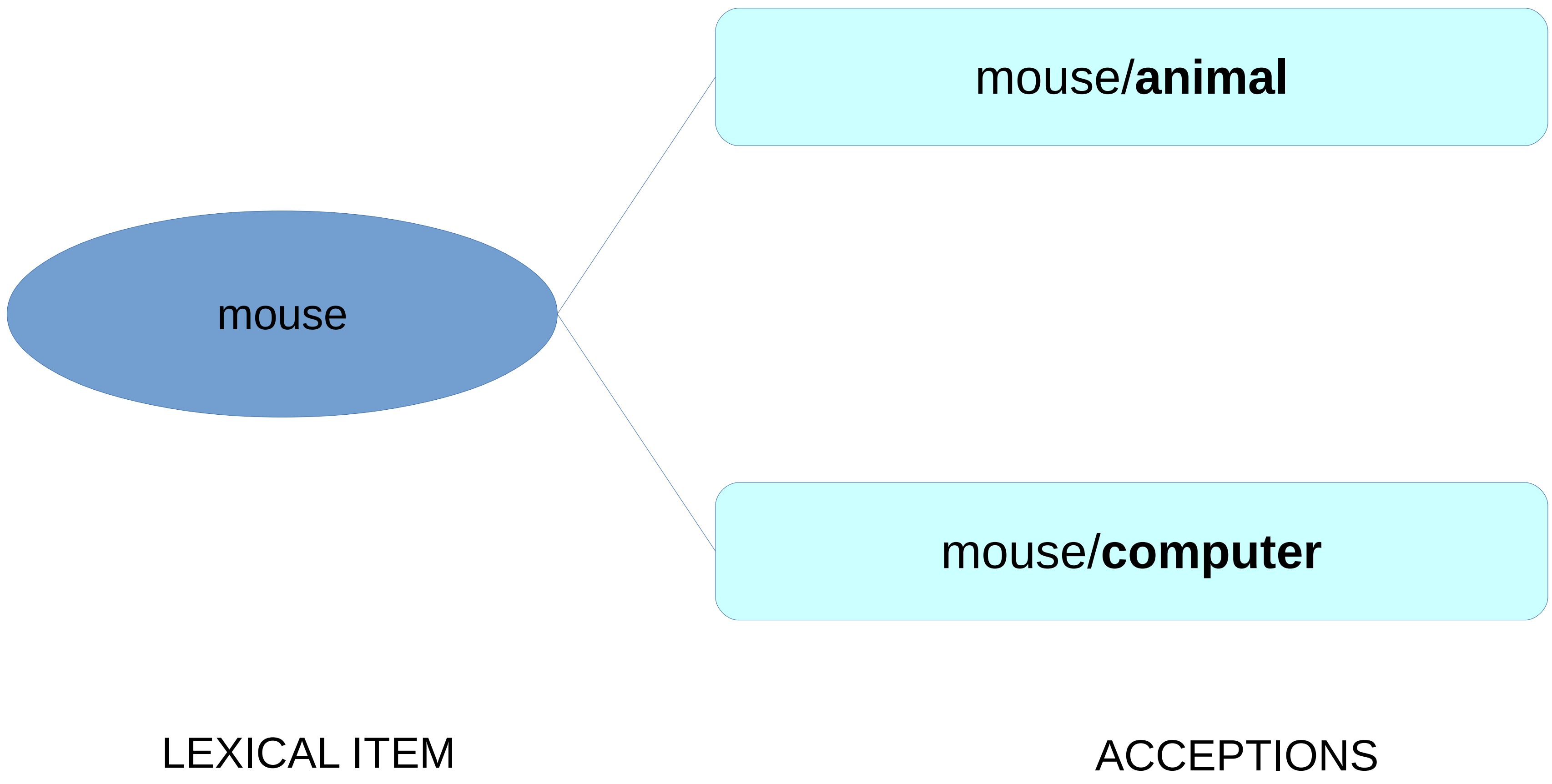
Terms are monosemic or polysemic

'*cashew*', '*neuroleptic*', '*daucus carota*', '*mouse*', '*rabbit*', '*carot*'

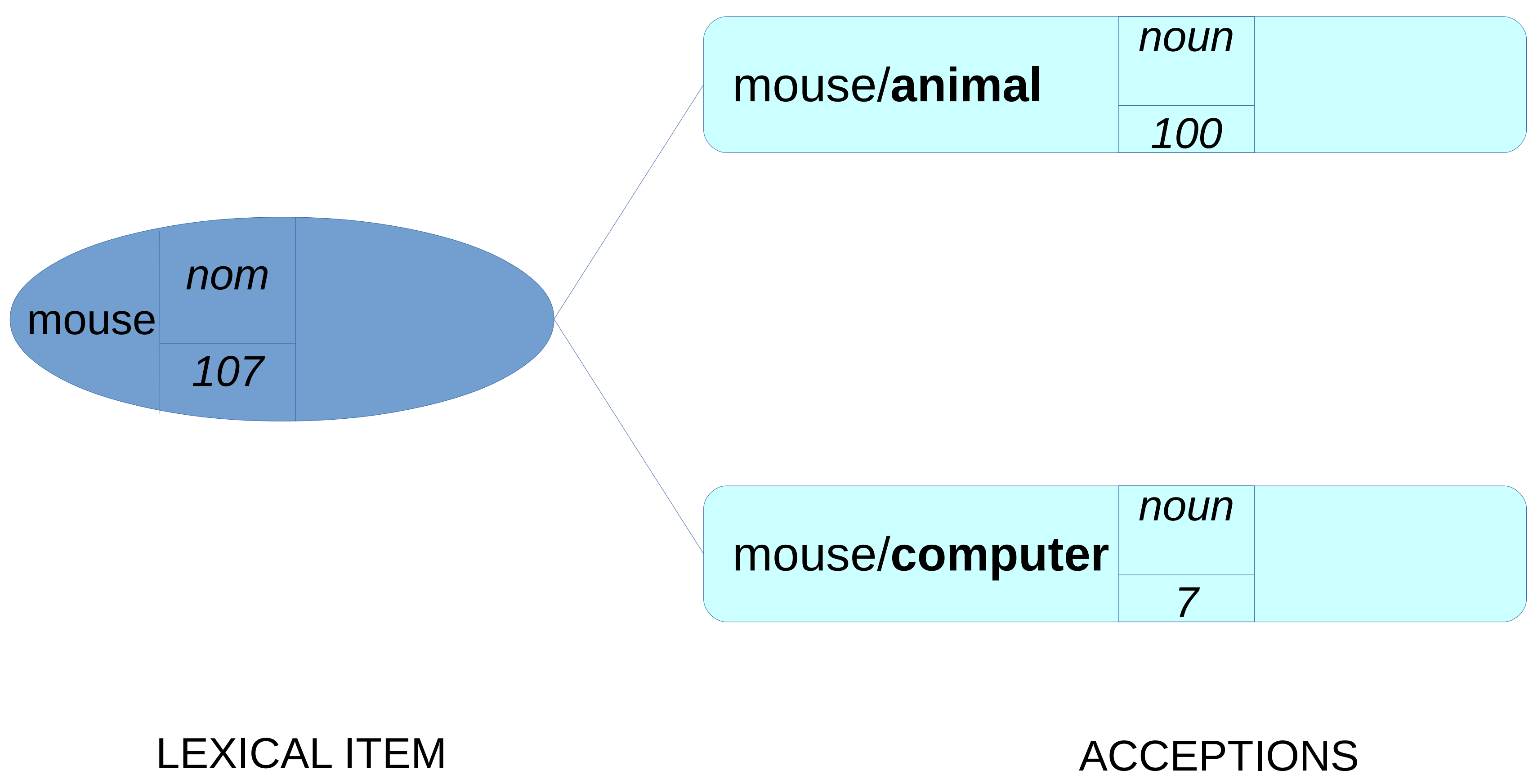
Acception : particular meaning of an item which is accepted by usage

The meaning comprehension is not only to select a good acception but also to etablish relations between surface structure and deep structure.

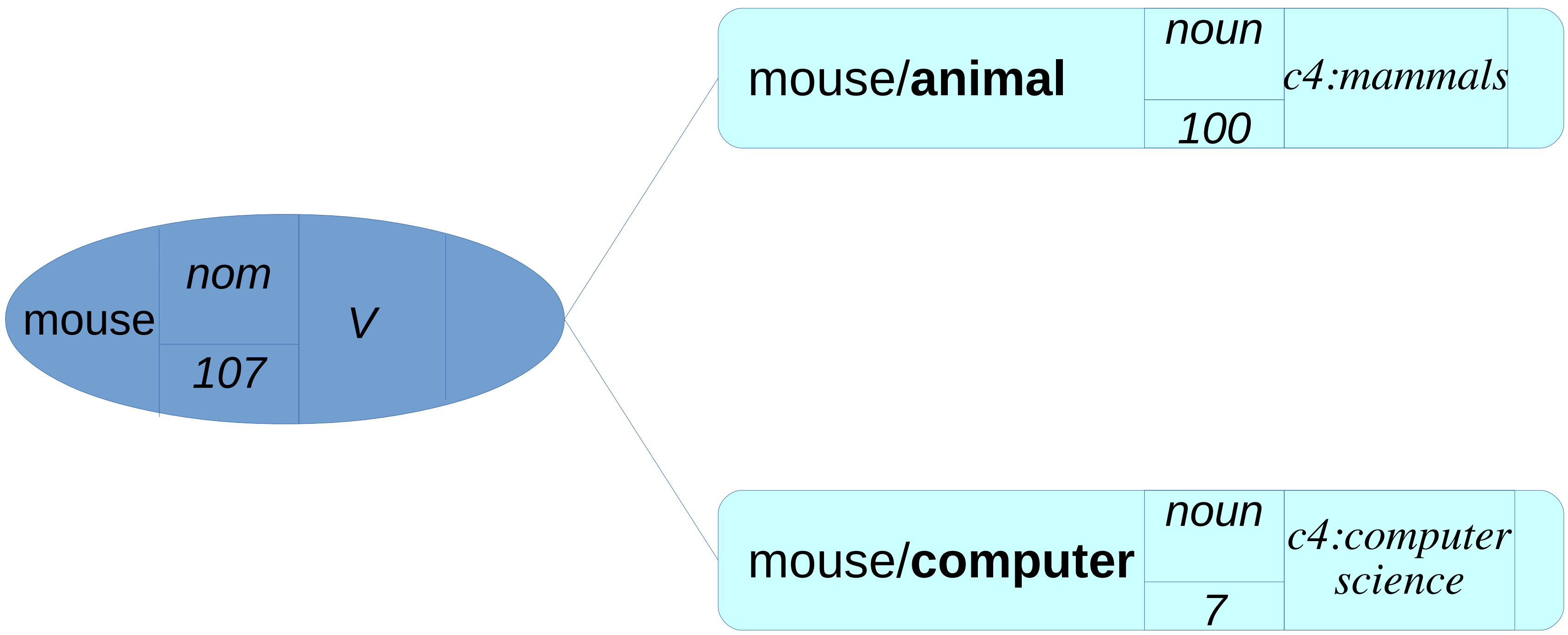
Hypothesis II



Hypothesis II



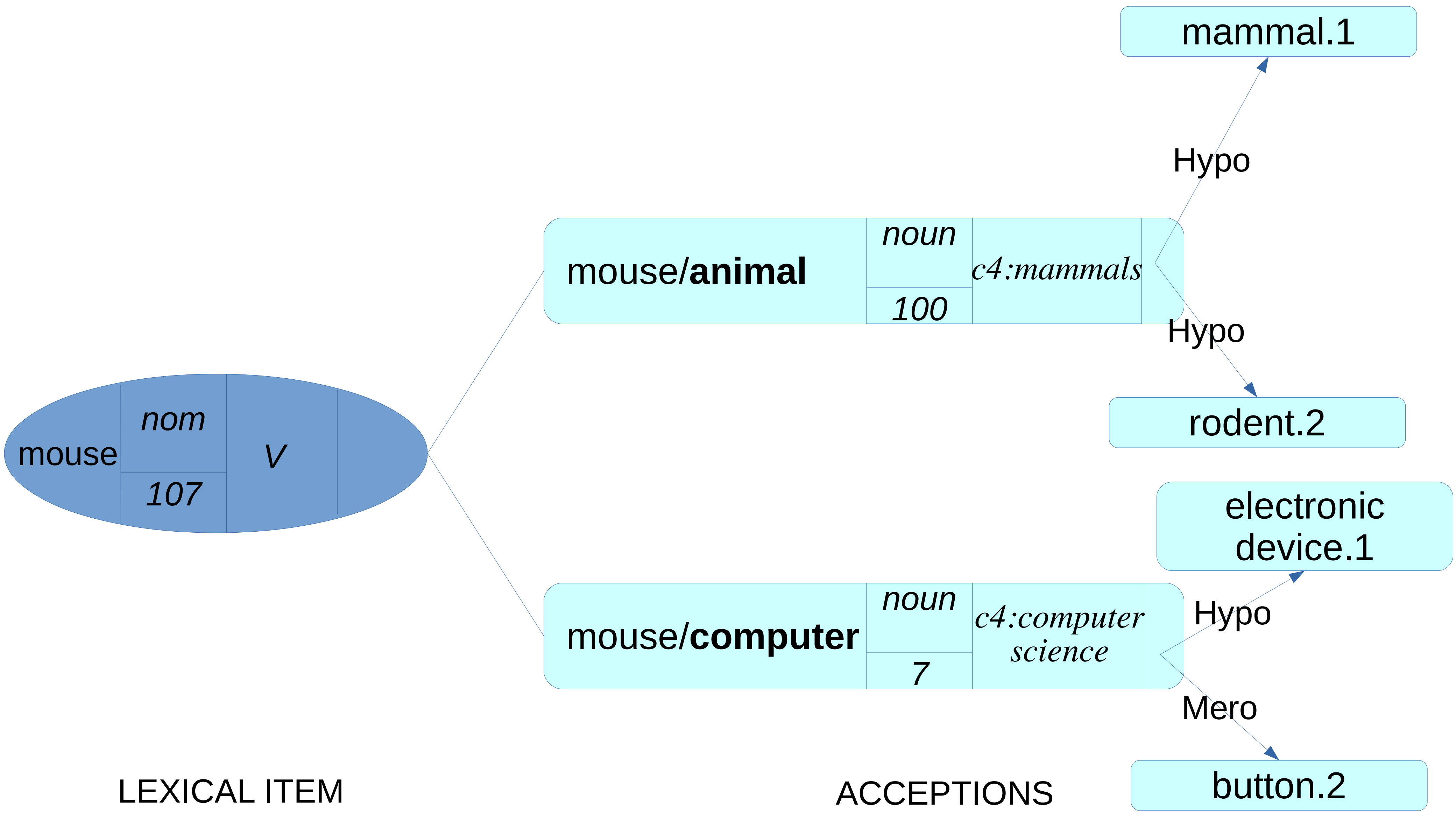
Hypothesis II



LEXICAL ITEM

ACCEPTIONS

Hypothesis II



Hypothesis III

Automatic Generation of Lexical Objects

Hypothesis III

Objective : to build a database to store lexical objects
ACCEPTIONS and LEXICAL ITEMS

For French, on more than 100 000 entries, polysemy rate of 61%

Average of 5 definitions, 400 000 lexical objects

Impossible to manually index

Hypothesis III

How ?

- from a reduced kernel of relevant terms (1000-2000) manually indexed
- automatic indexing of others

Utilisation of information extracted from diverse sources

dictionaries (semantic analysis)

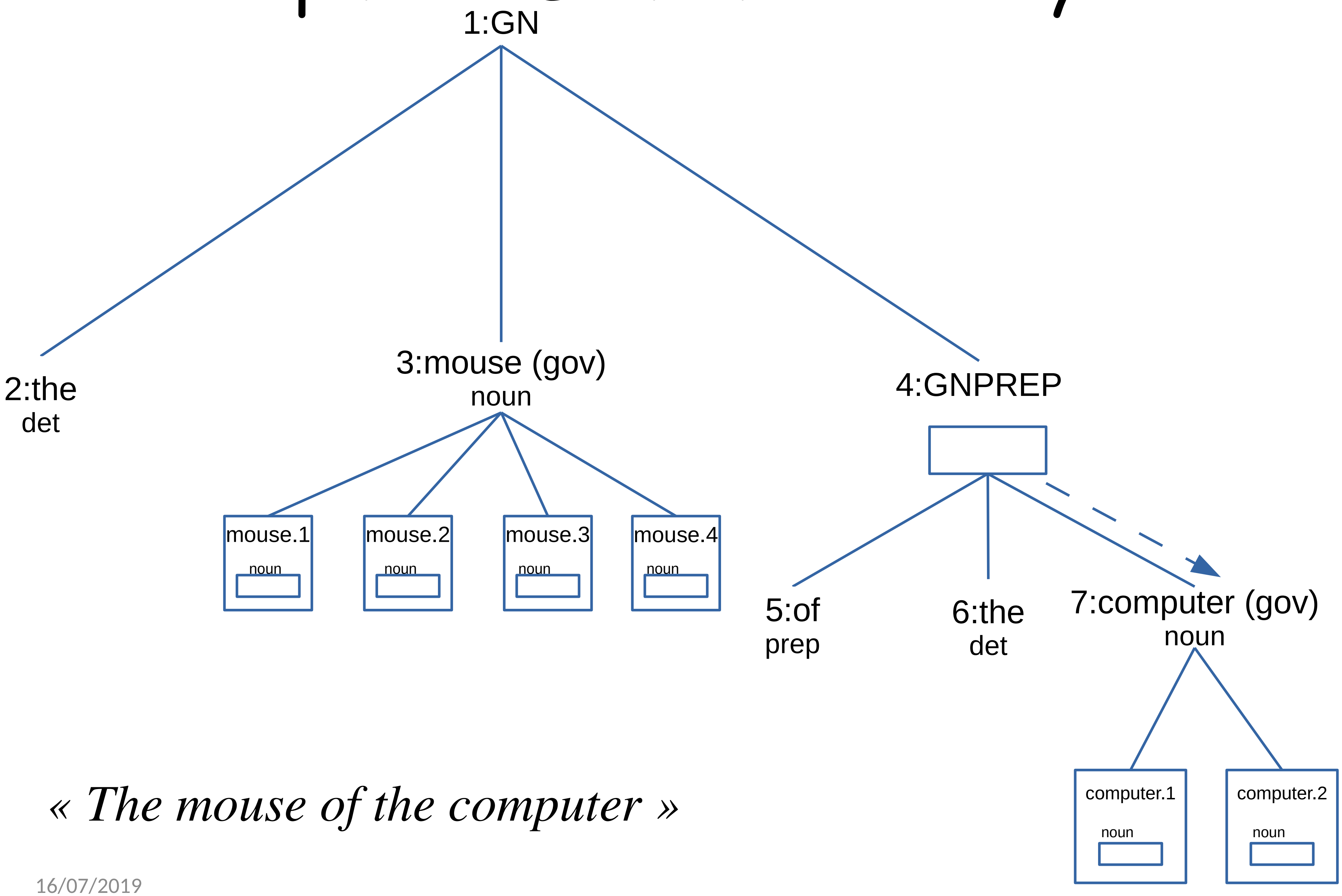
synonyms (vectors + morphology)

antonyms (vectors (antonymy function) + morphology)

Web (information site, Google, ...)

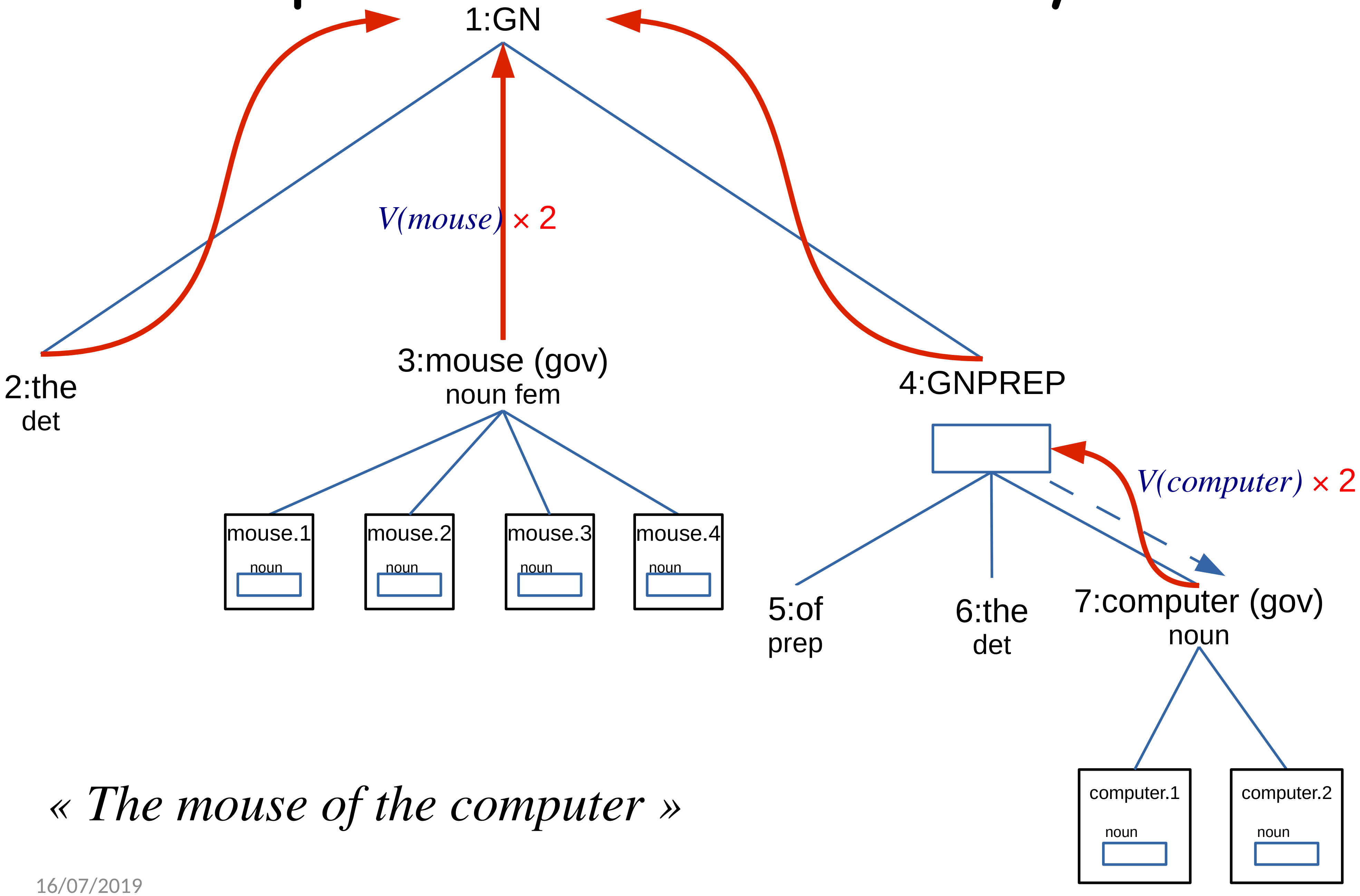
Corpora, ...

Upward-Downward Analysis



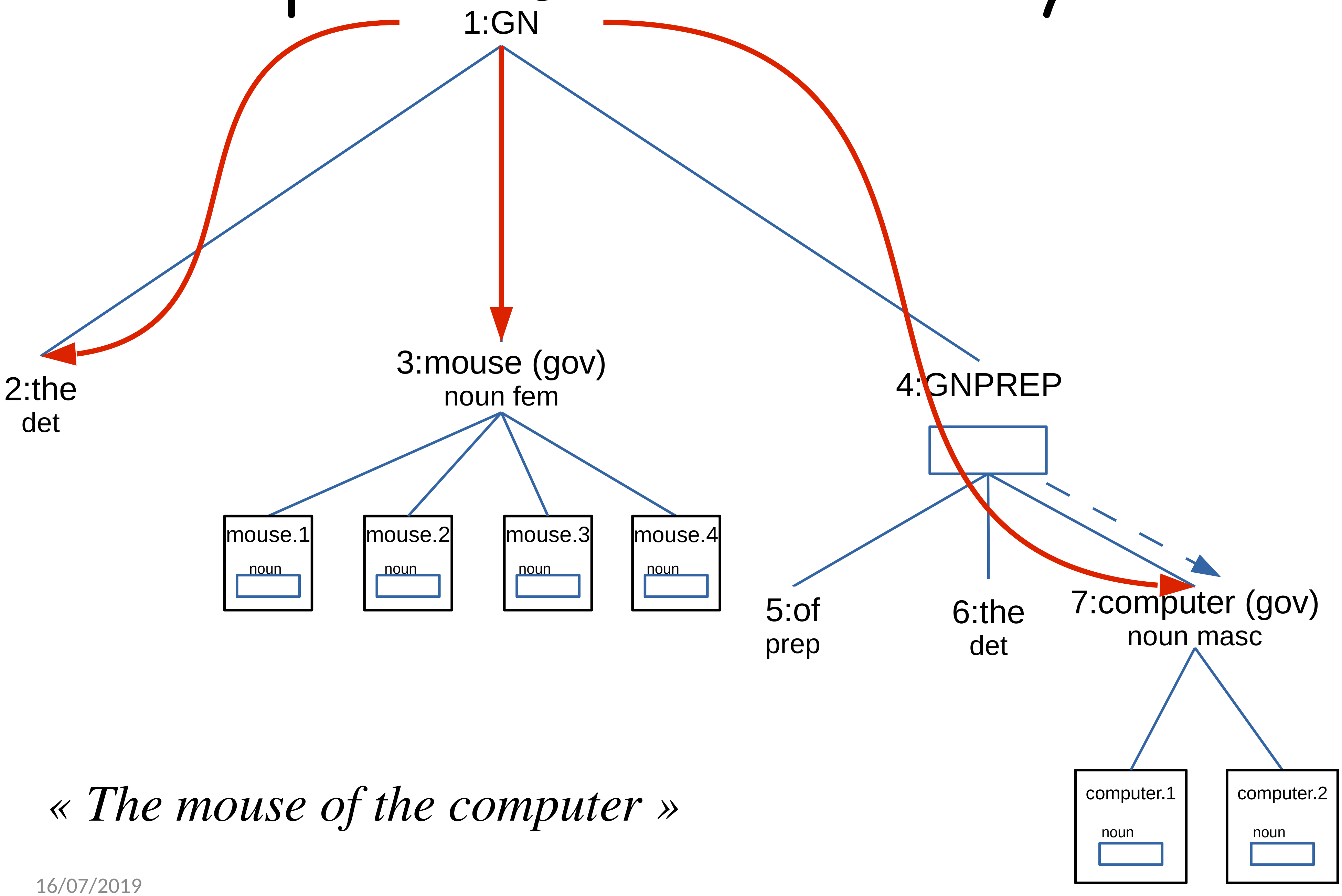
« *The mouse of the computer* »

Upward-Downward Analysis



« The mouse of the computer »

Upward-Downward Analysis



« The mouse of the computer »

Hypothesis III

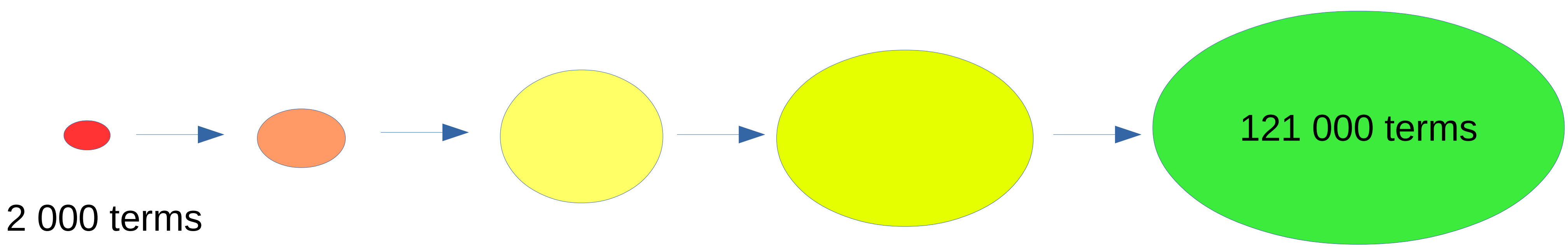
The kernel of lexical objects O is relevant

The learning must be coherent

Relevance (O) + Coherence (L)

Relevance (L) + Coherence (O)

End of 2005 : 121 000 terms automaticaly indexed



Hypothesis IV

Multi-source Analysis

Hypothesis IV

Metalanguage : *refer to, term for, plural of…*

luftwaffe : « *is the commonly used term for the German Air Force*. »

men : « *plural of man*. »

Lexicon coverage

constant evolution

« incompleteness » of dictionaries

'*liturgiste*' $\in$ Robert

$\notin$ Larousse

Solution

Construction of one LEXIE for one definition

LEXIE = atom of our database

Example

botte-1 : #nf# Réunion de végétaux de même nature liés ensemble. (Une botte de paille, de radis, de fleurs) . [Hach]
botte-2 : #nf# En escrime, coup porté à l'adversaire avec un fleuret ou une épée. (Pousser, porter, parer une botte) (Botte secrète.). [Hach]
botte-3 : #nf# Chaussure de cuir, de caoutchouc ou de plastique qui enferme le pied et la jambe, parfois la cuisse. (Des bottes de cavalier) Chaussure d'extérieur basse. (Botte d'hiver, de ski, de marche) . [Hach]
botte-4 : #nf# (néerl. bote, touffe de lin) . Assemblage de végétaux de même nature liées ensemble : (Botte de paille. Botte de radis.) . [Lar]
botte-5 : #nf# (#ethym-it# botta, coup) . Coup de pointe donné avec le fleuret ou l'épée . [Lar]
botte-6 : #nf# (p.-ê. de bot) . Chaussure à tige montante qui enferme le pied et la jambe généralement jusqu'au genou : (Bottes de cuir) . [Lar]

Example

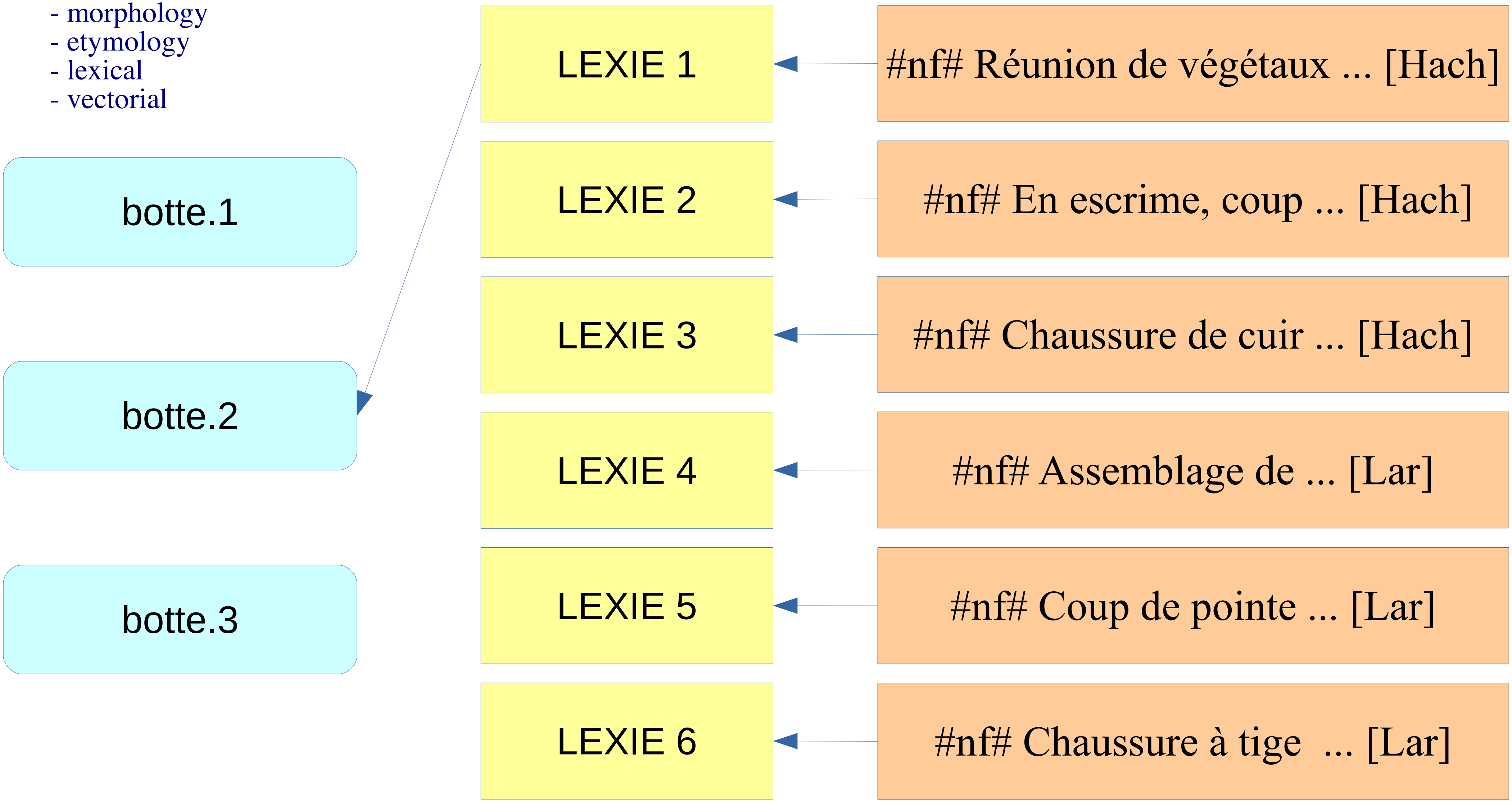
Collection of lexical information and conceptual vectors computation

LEXIE 1 ←	**botte-1** : #nf# Réunion de végétaux ... [Hach]
LEXIE 2 ←	**botte-2** : #nf# En escrime, coup ... [Hach]
LEXIE 3 ←	**botte-3** : #nf# Chaussure de cuir ... [Hach]
LEXIE 4 ←	**botte-4** : #nf# Assemblage de ... [Lar]
LEXIE 5 ←	**botte-5** : #nf# Coup de pointe ... [Lar]
LEXIE 6 ←	**botte-6** : #nf# Chaussure à tige ... [Lar]

Example

Senses categorisations
function of
- morphology
- etymology
- lexical
- vectorial

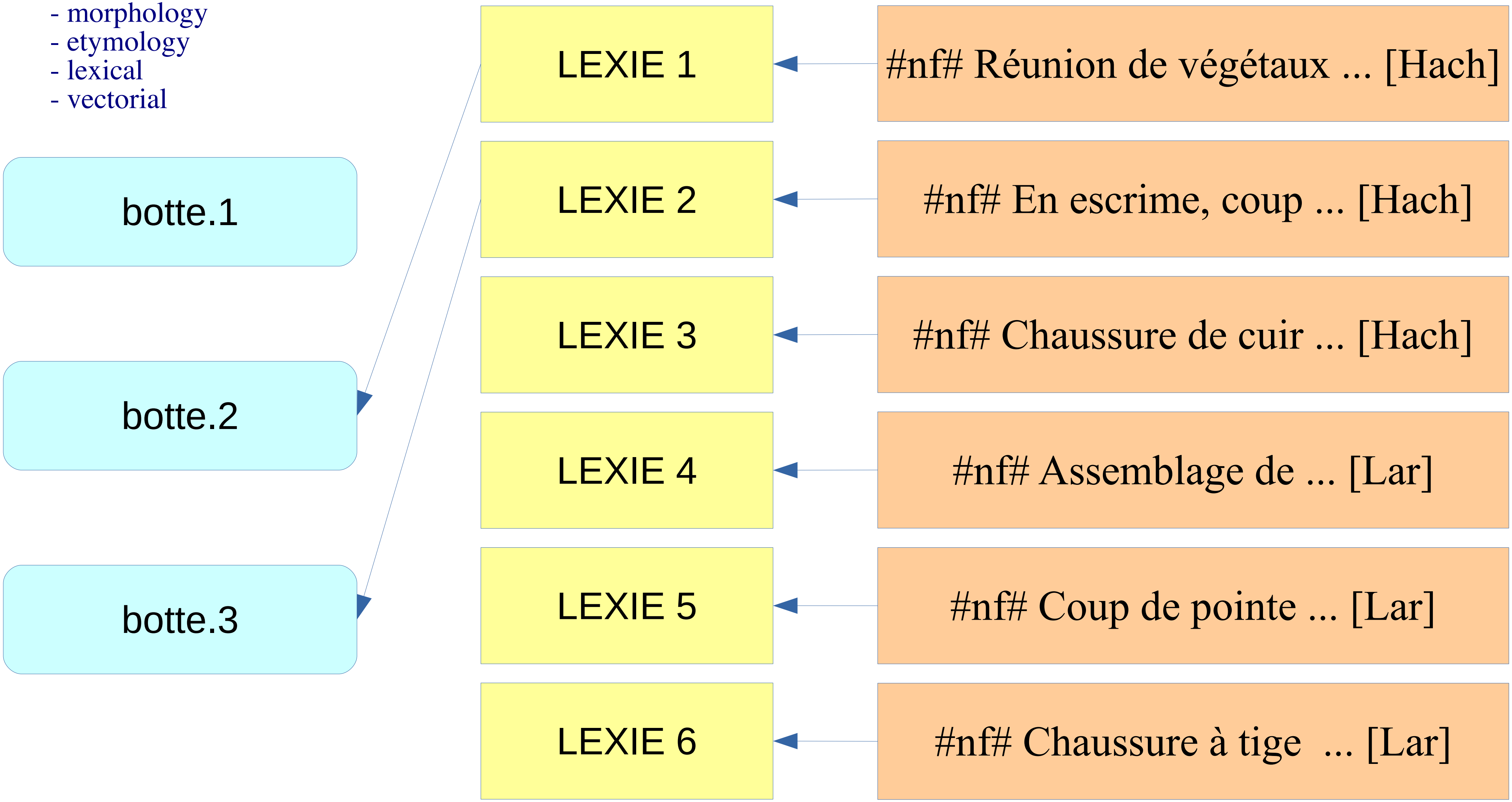
[Jalabert, Lafourcade] [Schwab]



Example

Senses categorisations
function of
- morphology
- etymology
- lexical
- vectorial

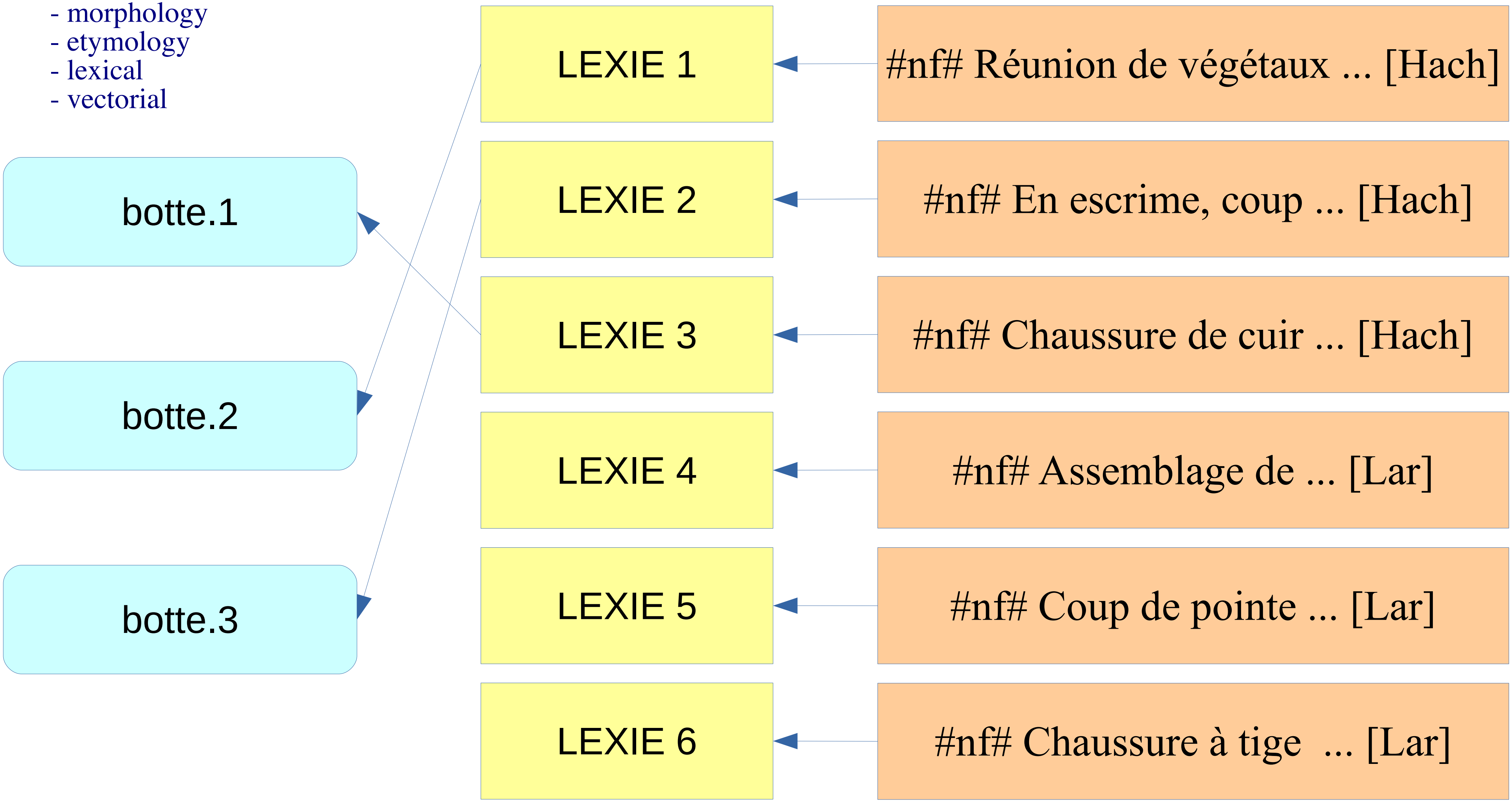
[Jalabert, Lafourcade] [Schwab]



Example

Senses categorisations
function of
- morphology
- etymology
- lexical
- vectorial

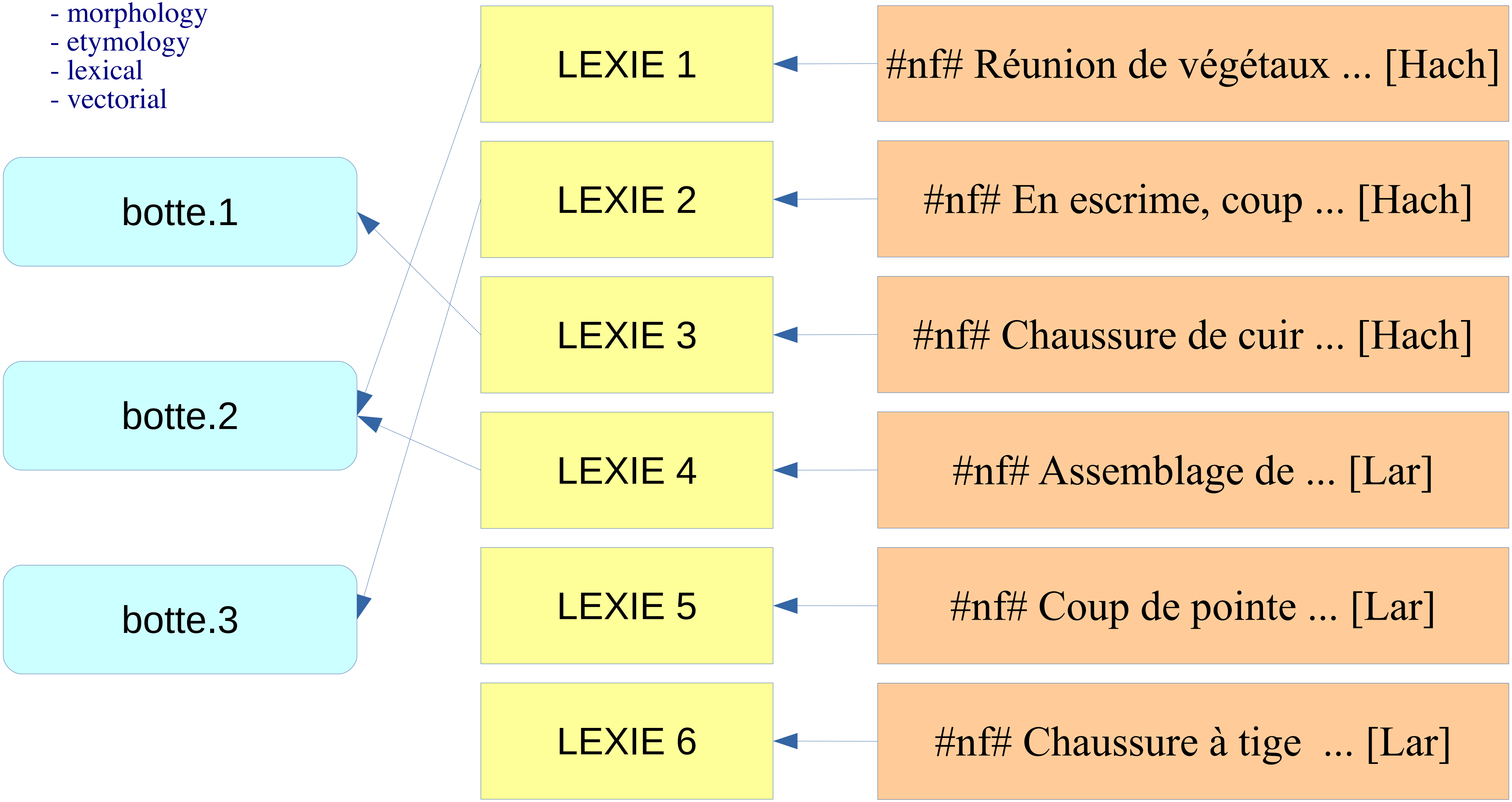
[Jalabert, Lafourcade] [Schwab]



Example

Senses categorisations
function of
- morphology
- etymology
- lexical
- vectorial

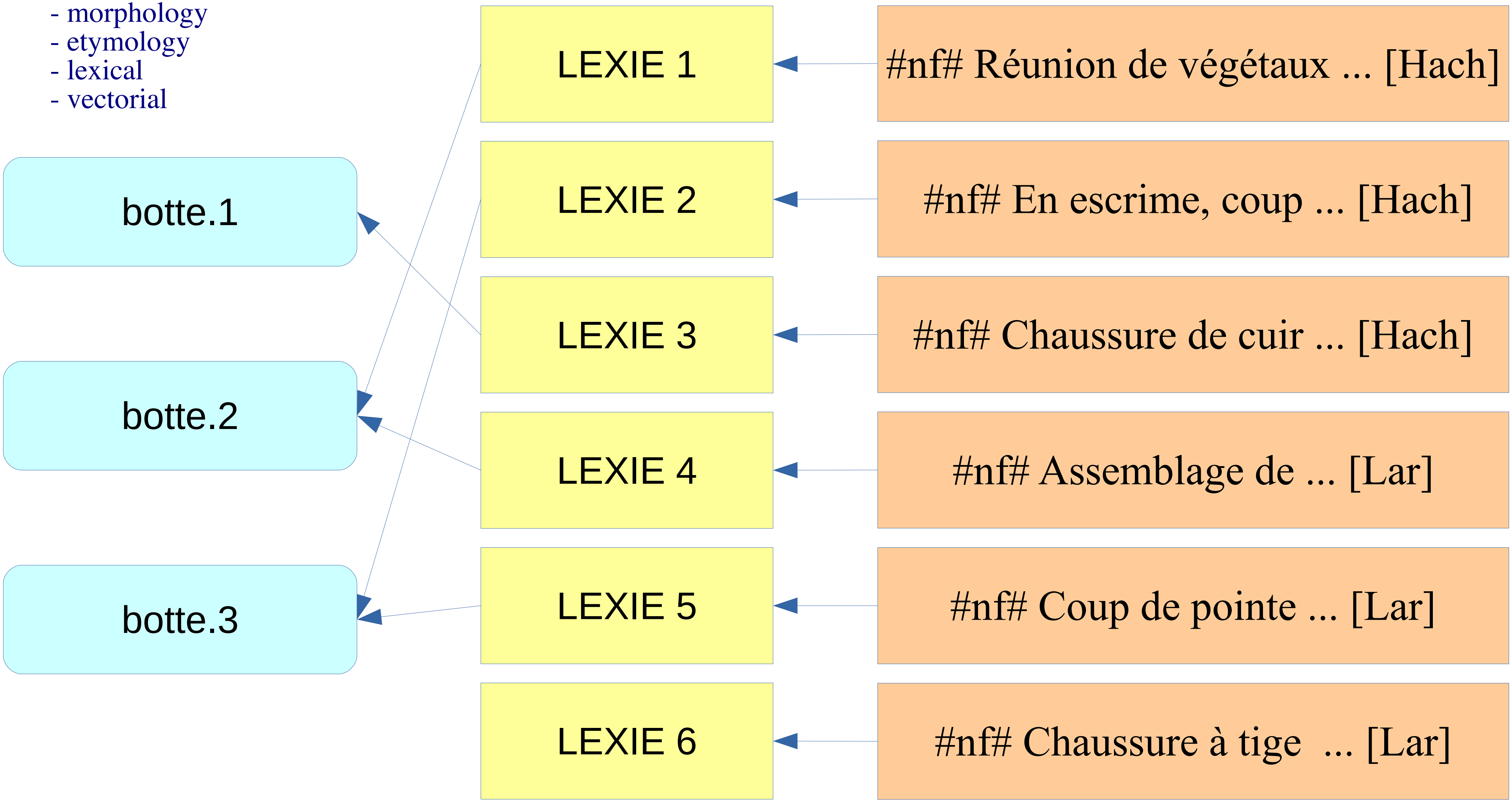
[Jalabert, Lafourcade] [Schwab]



Example

Senses categorisations
function of
- morphology
- etymology
- lexical
- vectorial

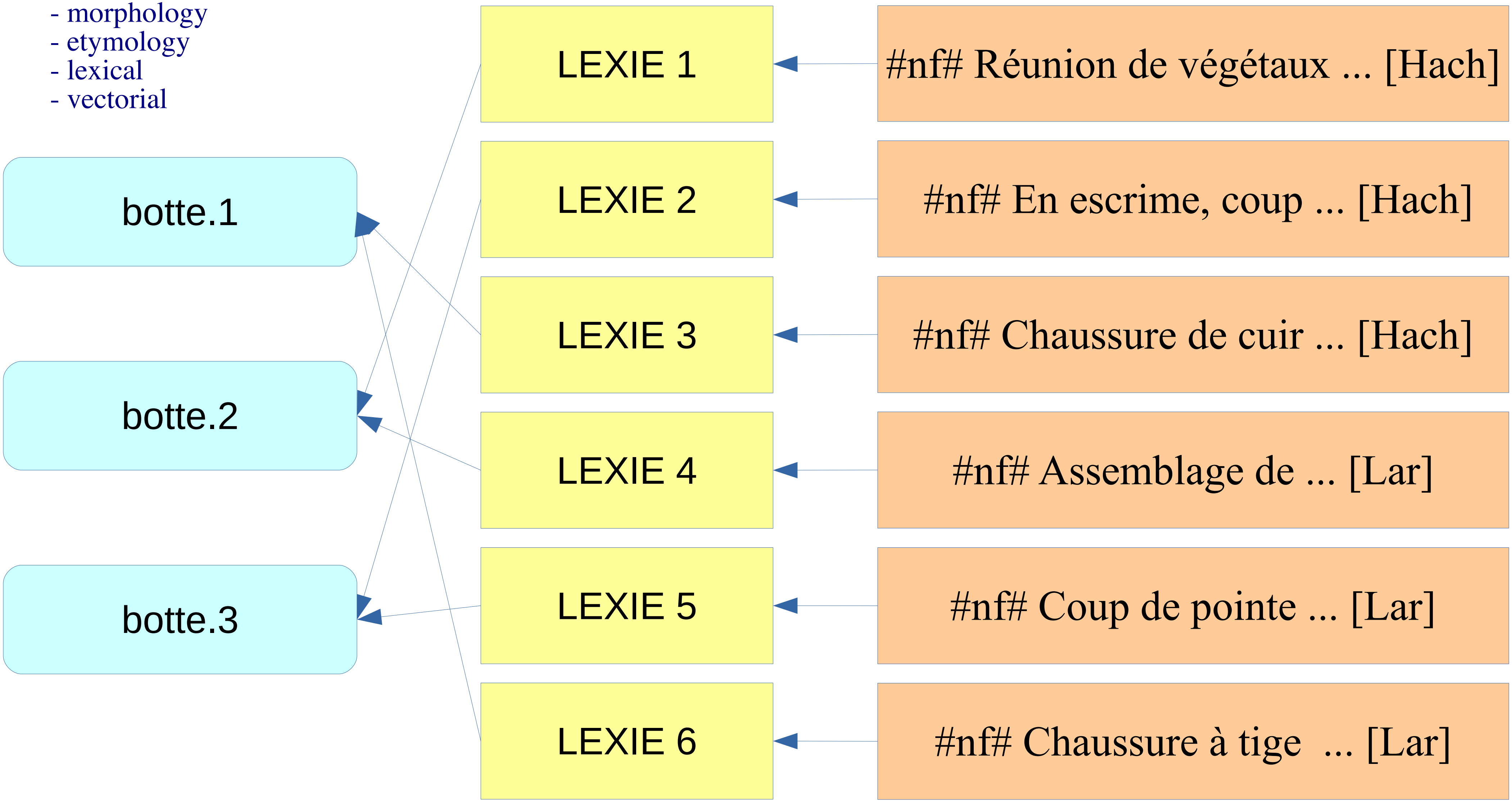
[Jalabert, Lafourcade] [Schwab]



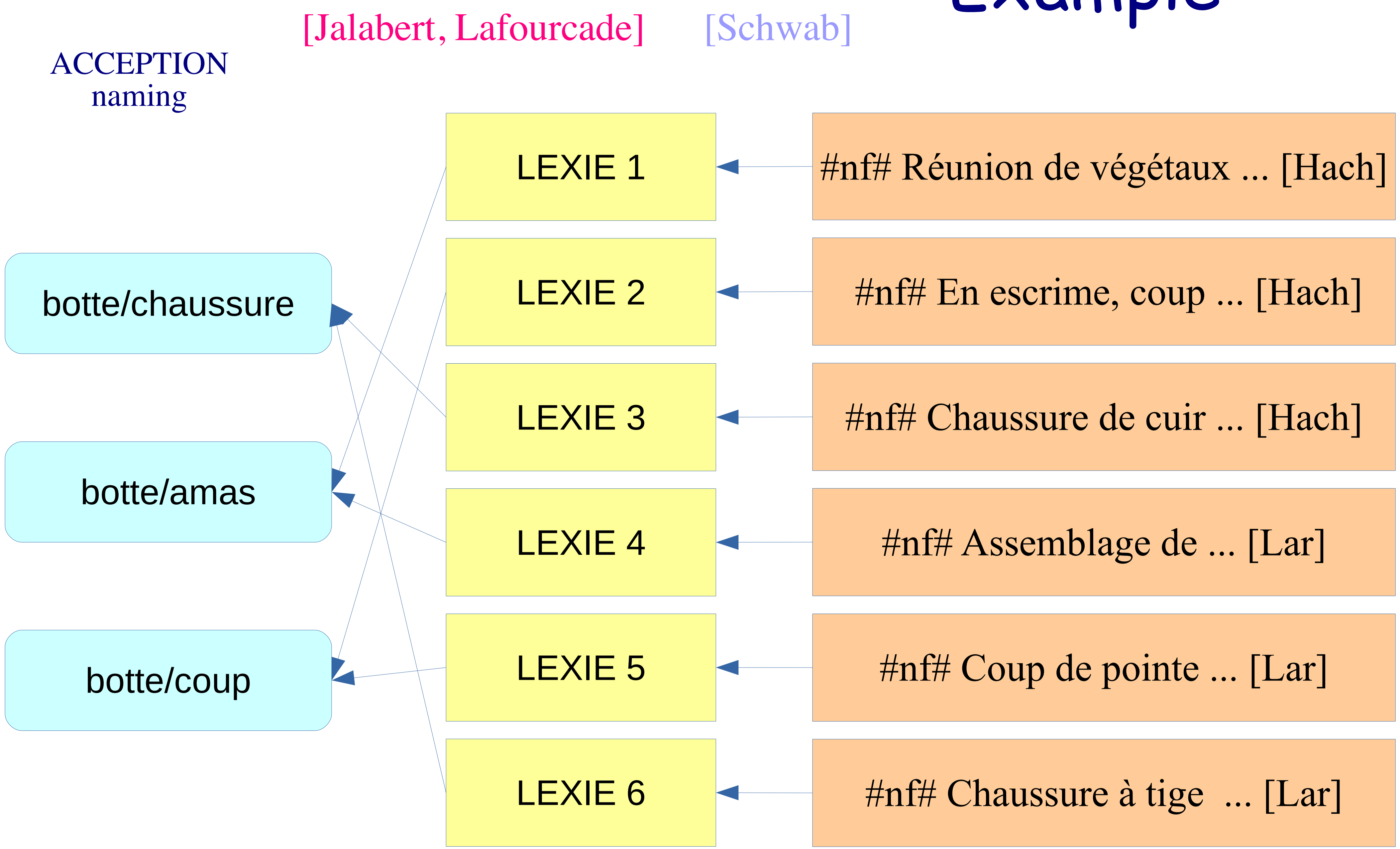
Example

Senses categorisations
function of
- morphology
- etymology
- lexical
- vectorial

[Jalabert, Lafourcade] [Schwab]



Example


[Jalabert, Lafourcade]
[Schwab]
ACCEPTION naming
botte/chaussure
botte/amas
botte/coup
LEXIE 1
LEXIE 2
LEXIE 3
LEXIE 4
LEXIE 5
LEXIE 6
#nf# Réunion de végétaux ... [Hach]
#nf# En escrime, coup ... [Hach]
#nf# Chaussure de cuir ... [Hach]
#nf# Assemblage de ... [Lar]
#nf# Coup de pointe ... [Lar]
#nf# Chaussure à tige ... [Lar]

Hypothesis V

Continuous Learning

Continuous Learning

Analysis of newspaper articles, crowsourcing

- New words, new senses
- Named entities
  - Entities : Podemos, Engie (former GDR Suez), ...
  - People : Peter Dinklage, Nabilla, Emmanuel Macron, ...
  - → Web pages, Wikipedia, wiktionaries

For database coherence

- Base is not coherent during the first cycles
- Vector convergence to a quasi-stable position after a certain number of cycle (experimentally at least 10)
- This number of cycle is function of the learning order and function of definitions.

Hypothesis VI

Double Loop

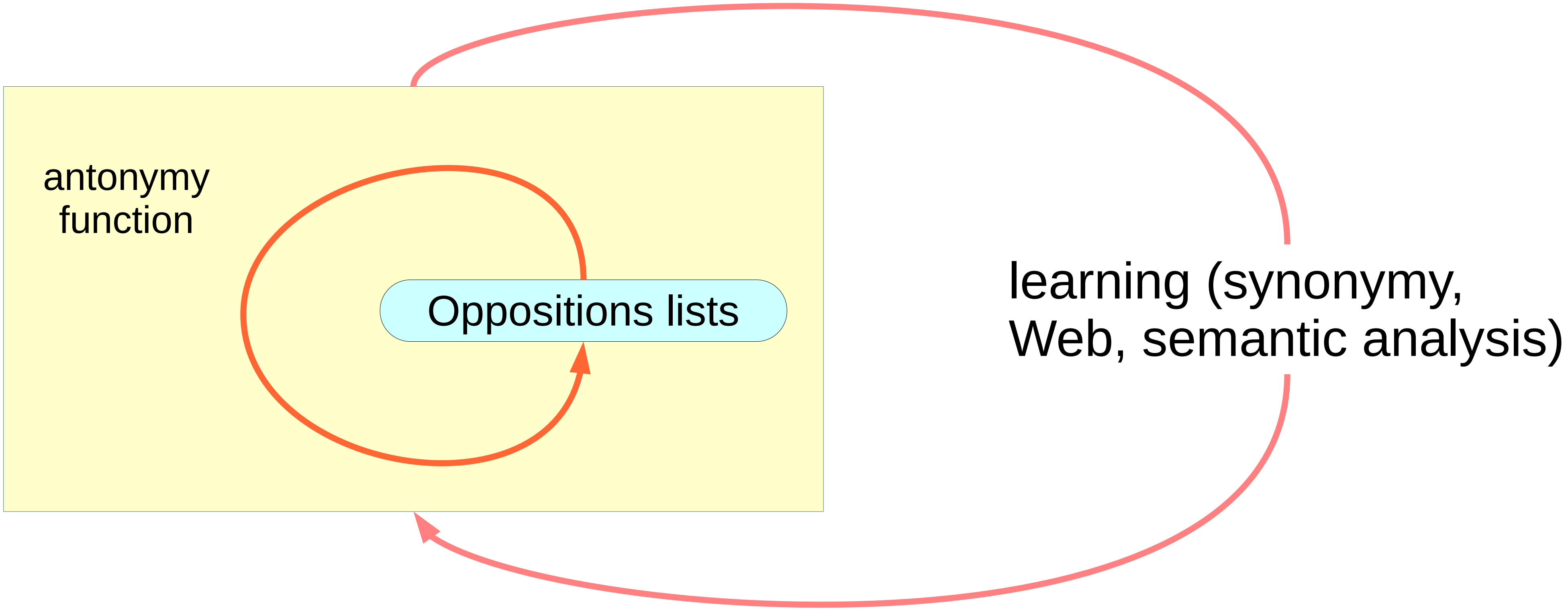
Double Loop

From biology [Lecerf]

Invariant structural element of organism

Permit action on its environment and is its product

Example : antonymy function [COLING'2002, JADT'2002, TALN'2002]



Experiment (2004-2005)

115 agents (1 base, up to 10 of each type)

5 machines (PC Linux, Sun Unix)

5 sources (Larousse, Robert, thésaurus Larousse, synonyms, antonyms dictionnaries from Caen)

French data base

121 000 LEXICAL ITEMS

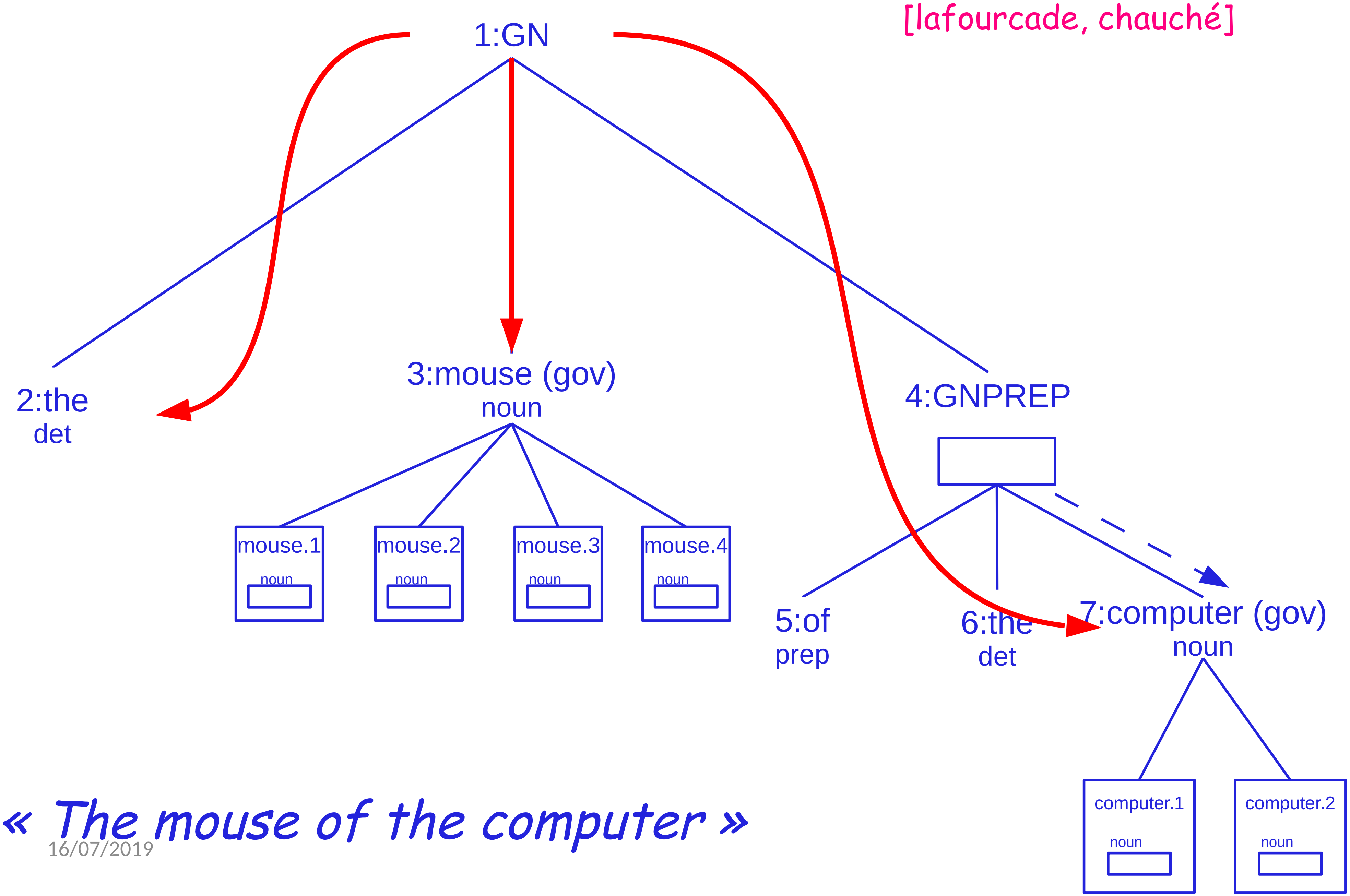
276 000 ACCEPTIONS

842 000 LEXIES

Cycle (around 4 days)

Upward-Downward Analysis

[lafourcade, chauché]



« The mouse of the computer »

Upload-Download Analysis : Outcome

Lexical Disambiguation : Yes
References : No
Prepositional Attachments : No
Lexical Functions Detection : No
Interpretation path : No

Experiments After 2005
Penang, Malaysia, 2006-2007
Grenoble, France, 2007-2012

Conceptual vectors, a complementary tool to lexical networks

Lexico-semantic Network

From *Ross Quillian*'s work during the 60's

Psycholinguistic experiments about organisation of concepts and words in the mind

Task : lexical disambiguation (≅ Word sense disambiguation), categorisation, ...

Applications : Machine Translation, Automatic Summarization, Information Retrieval, message composition, ...

WordNet

Lexical database for English

Developed since 1985

Under the direction of *George Armitage Miller* by the *Cognitive Science Laboratory* of the U*niversity of Princeton*

Aims to be consistent with the access to the human mental lexicon

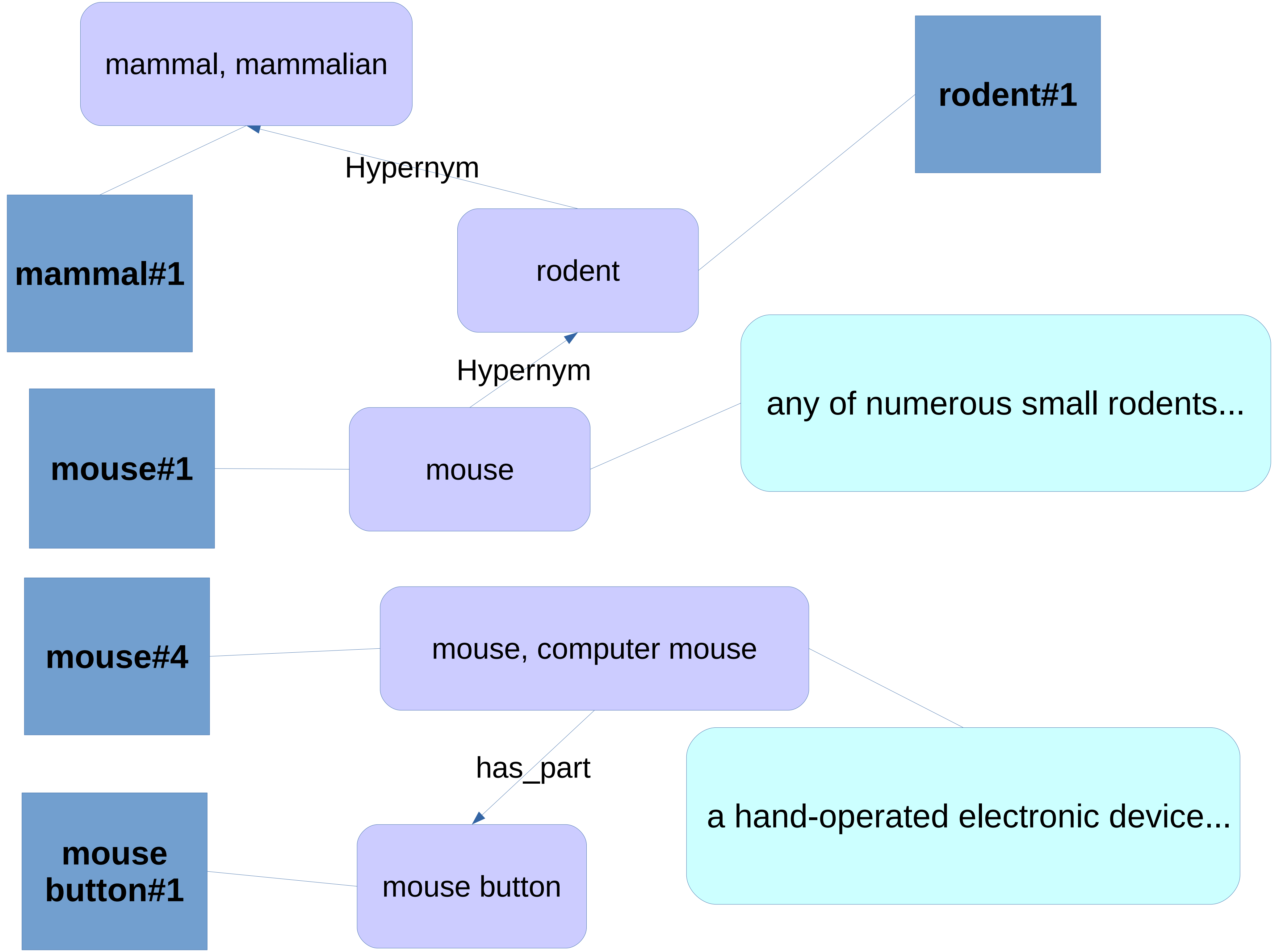
WordNet

Organised in sets of synonyms (synsets)

To each synset corresponds a concept

Meanings are described by 3 means :

- a *definition*
- a *synset*
- some *lexical relations* which link synsets


mammal, mammalian
rodent#1
Hypernym
mammal#1
rodent
any of numerous small rodents...
Hypernym
mouse#1
mouse
mouse#4
mouse, computer mouse
has_part
a hand-operated electronic device...
mouse button#1
mouse button

Some Statistics

POS	**Monosemous**	**Polysemous**
Nouns	101321	15776
Verbs	6261	5227
Adjectives	16889	5252
Adverbs	3850	751
Totals	128321	27006

from http://wordnet.princeton.edu/man/wnstats.7WN

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

Weakness shared with dictionaries

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

Weakness shared with dictionaries

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

derivational forms

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

derivational forms (still seldom)

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

agent, instrument, goal, place,...

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

agent, instrument, goal, place,...
(still missing)

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

no connection between '*doctor*'-'*hospital*',
'*port*'-'*boat*',...

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

no connection between '*doctor*'-'*hospital*',
'*port*'-'*boat*',... (addition of domains)

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

consequence of the three precedents

Known Weaknesses of WordNet

Creators of Wordnet identify 6 weakness (Harabagiu et al, 1999)

1. lack of uniformity and consistency in the definitions
2. some concepts (word senses) and relations are missing
3. the lack of morphological relations
4. the absence of thematic relations/selectional restrictions
5. limited number of connections between topically related words
6. lack of connections between hierarchies

$\Rightarrow$Tennis Problem (Fellbaum, 1998)

Structural Limits

"Messi scored a goal"
semantic field of the football ?
domain ? (football ? sport ? other ?)

How to represent the notion of "semantic field" ?
To introduce such edges would cause 2 problems due to the fuzzy character of this relation :

- how to consider that two meanings are in the same semantic field ? (too many or too few relations)
- how to represent a notion with fuzzy characteristics by a discrete representation ?

Construction by predefined concepts

How ?

- from a reduced kernel of relevant terms (1000-2000) manually indexed
- automatic indexing of other

Advantages ?

- supposed relevance of concepts
- easier "reading" of vectors

Disadvantage ?

- variable lexical density

Construction with predefined concepts

The kernel of lexical objects O is relevant

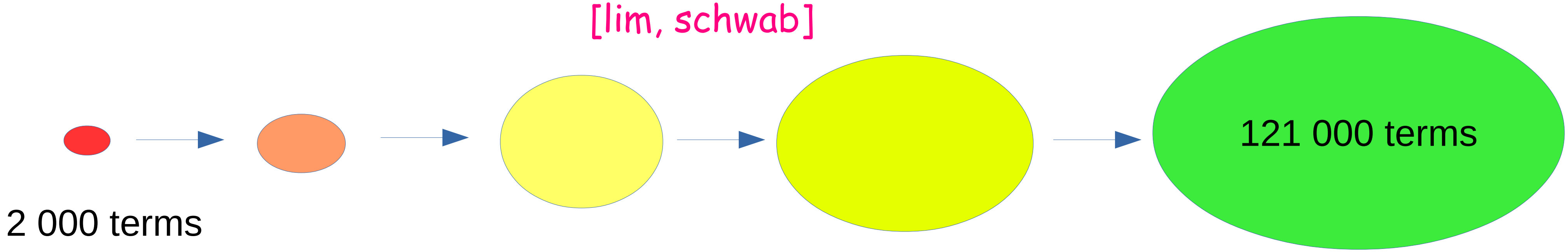
The learning L must be coherent

Relevance (O) + Coherence (L)

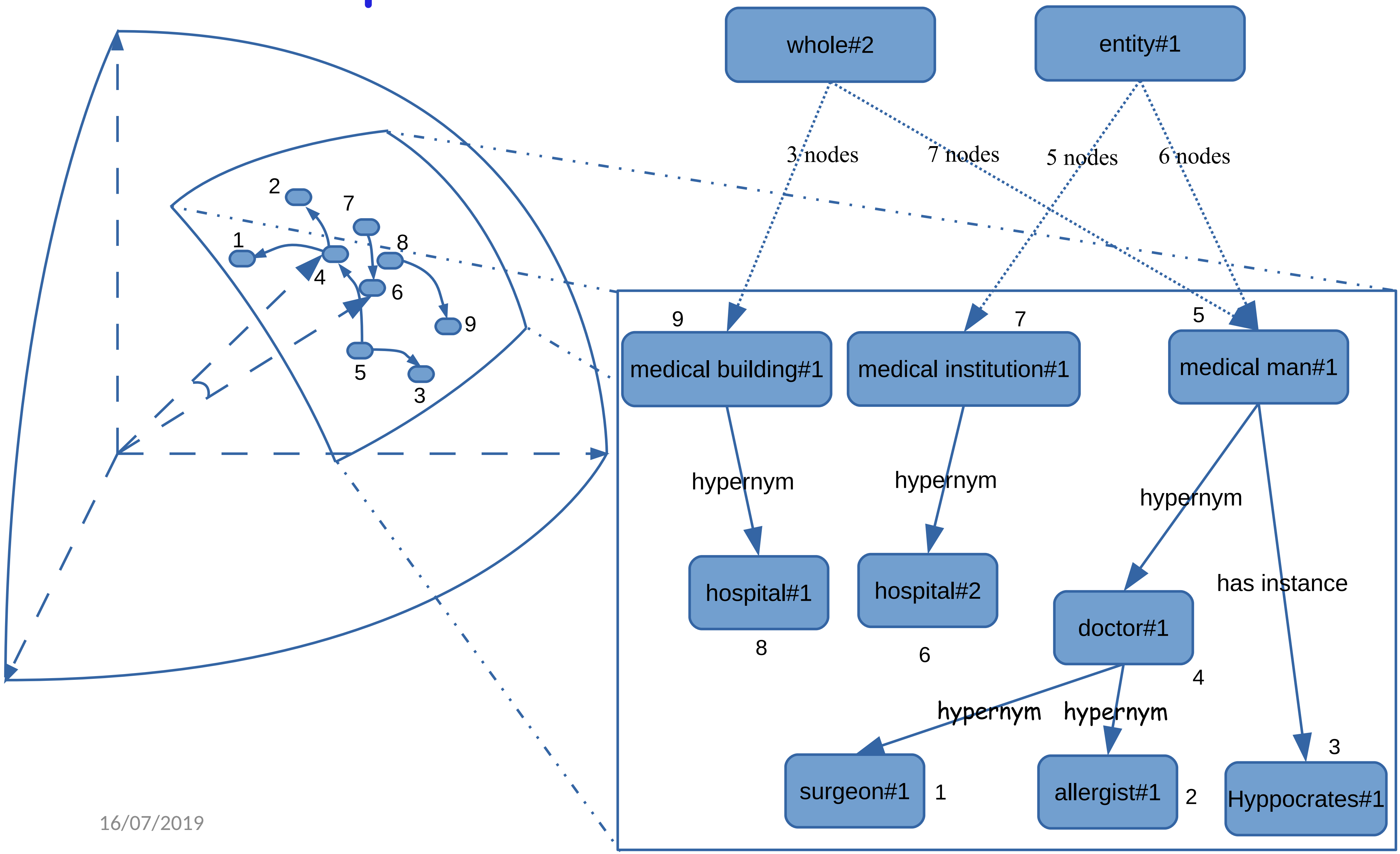
Relevance (L) + Coherence (O)

2 experiments : - Montpellier (Larousse) 121 000 terms automaticaly indexed [schwab, lafourcade]

- Penang (Sumo) indexation of Wordnet [lim, schwab]



Conceptual Vectors and Wordnet



Construction by emergence

How ?

- without hierarchy *a priori defined*
- vector size *a priori* fixed
- randomised vectors
- automatic indexing of terms

Advantages ?

- choice depends on available resources
- lexical density more constant in space

Disadvantage ?

- difficult to "read" a vector

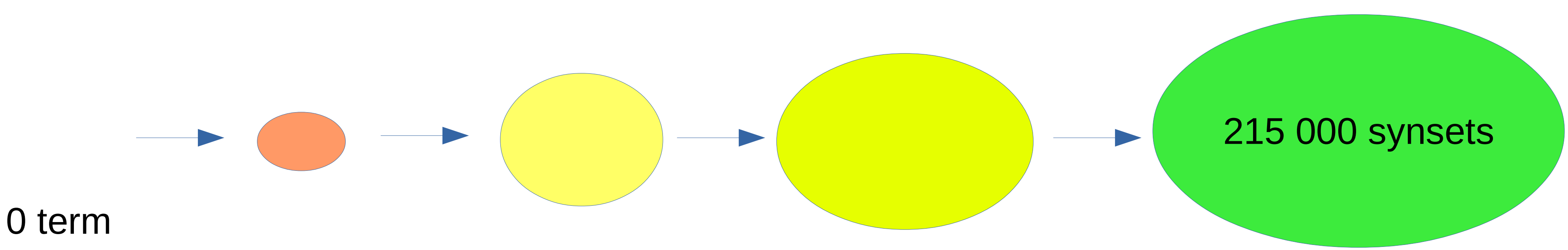
Construction by emergence

The learning must be coherent

Coherence (L)

Relevance (O)

Experiment : on Wordnet, indexation of 215.000 synsets (words meaning)



Conceptual vectors for Word Sense Disambiguation

- resolve examples through thematic
  (75% of ambiguity case)
  "Messi *scored a goal.*"
  "*The lawyer pleads at the court.*"
  same semantic field

- problem for cases as
  "The mouse bit through my LAN cables "

Lexical Function modeling

+ paradigmatics (in part)
  - hypernymy [JADT, 2004]
  - synonymy [Schwab, 2005]
  - antonymy [TALN, 2002; COLING, 2002]

+ syntagmatics (problematic)
  (collocations) [Schwab, 2005]

⇒ need lexical networks

Contribution of Vectors to Networks

Continuous field (flexibility)

any pair of lexical objects easily comparable

Bring closer words on minority but common ideas

Recall $\nearrow$ ('*hospital*' - '*patient*', '*tennis*' - '*ball*')

Vectors allows evaluation of a relation without characterising it (except *Syn* and *Anto*)

Experiment

Aims to a larger objective :

- improve an Example Based Machine Translation System
- semi-automatic creation of a multilingual lexical lexical database

Addition of conceptual vectors to Wordnet

Analysis from :

- definitions under logical form (genus-differentia)
- information from lexical network (lexical functions)

Overview

	Dictionaries	Lexical Networks
Pre-defined Concepts	Montpellier 2000-05	WordNet + Sumo Penang 2007-08
Émergence	WordNet Penang 2007-08 DBNary Grenoble 2010->2012	JeuxDeMots Mtp 08-? Wordnet Pen 07 - 08

Distributional linguistics

- Represents linguistic objects with the associability possibilities they share or not
- Linguistic items with similar distributions have similar meanings
- « You shall know a word by the company it keeps » (John Ruppert Firth, 1957)
- Meaning of a word is represented with all contexts where it can be find in texts.
  - Milk : {cow, milk, white, cheese, mammal,…}
  - Computer{school, electronic, machine, programmable,…}
- Distributionnal vectors

Distributional Vectors

- Built from corpora
- Each component corresponds to words in a corpus
  - Directly : Saltonian vectors
  - Indirectly : Latent Semantic Analysis, word embeddings

Saltonian Vectors

- Given a text corpus containing *n* unique words
- Size of vectors is *n*
- Classic binary word representation : Zeros everywhere but the index of the word
  - [0; 0; 0; 0;….; 0; 0; 1; 0;…; 0; 0]
- Vector of a text : sum of all words
- Vector of a lexical item : sum of all context where it occurs

TF-idf

Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2 + w_2^2 + \ldots + w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N - \mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases} 1 & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/\mathit{CharLength}^{\alpha}$, $\alpha < 1$
L (log ave)	$\frac{1 + \log(\mathrm{tf}_{t,d})}{1 + \log(\mathrm{ave}_{t \in d}(\mathrm{tf}_{t,d}))}$				

Saltionian Vectors

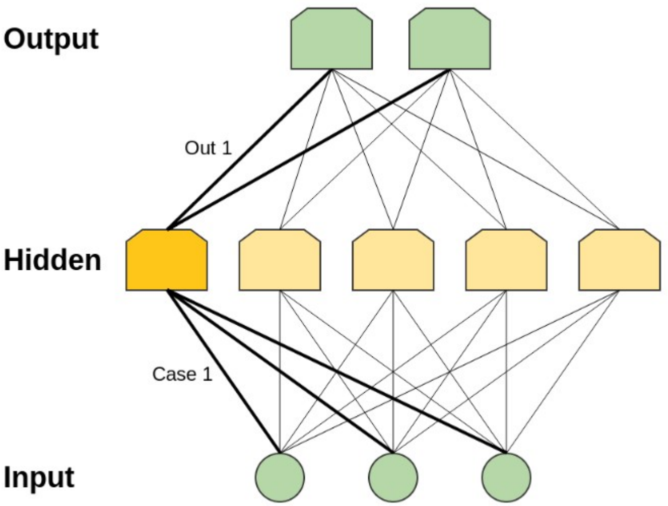
- Problems :
  - Learning has to be done from scratch if texts with new words are added (increase of vector size)
  - Size of vectors is very large and they contain lots of zeros
  - Sizes of databases are huge

Reducing Vector Size

- Given a text corpus containing *n* unique words
- Manually or automatcally define *m* « good » components
- *m<<n (often 100 < m < 500)*
- Size of vectors is *m*
- *Choice of m is empirical*
- *Exemples :*
  - *Matrix reduction : Latent Semantic Indexing* [Deerwester *et* al., 1988]
  - *Neural word embeddings* : *Word2Vec* [Mikolov *et* al., 2013]

Word2Vec

- Automatically learn good features
- Two-layer neural net that processes text
- Input : a text corpus
- Output : a set of vectors
- Very easy to use
  - Set of pre-computed vectors
  - Code in Java, C,...



Word2Vec : Interesting Results

- Cosine distance
- D('Sweden', 'Sweden') = 0
- D('Sweden', 'Norway') = 0.760124
- Neighborhood :

Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408

Word2Vec : Interesting Results

- Trained on 400 million tweets having 5 billion words

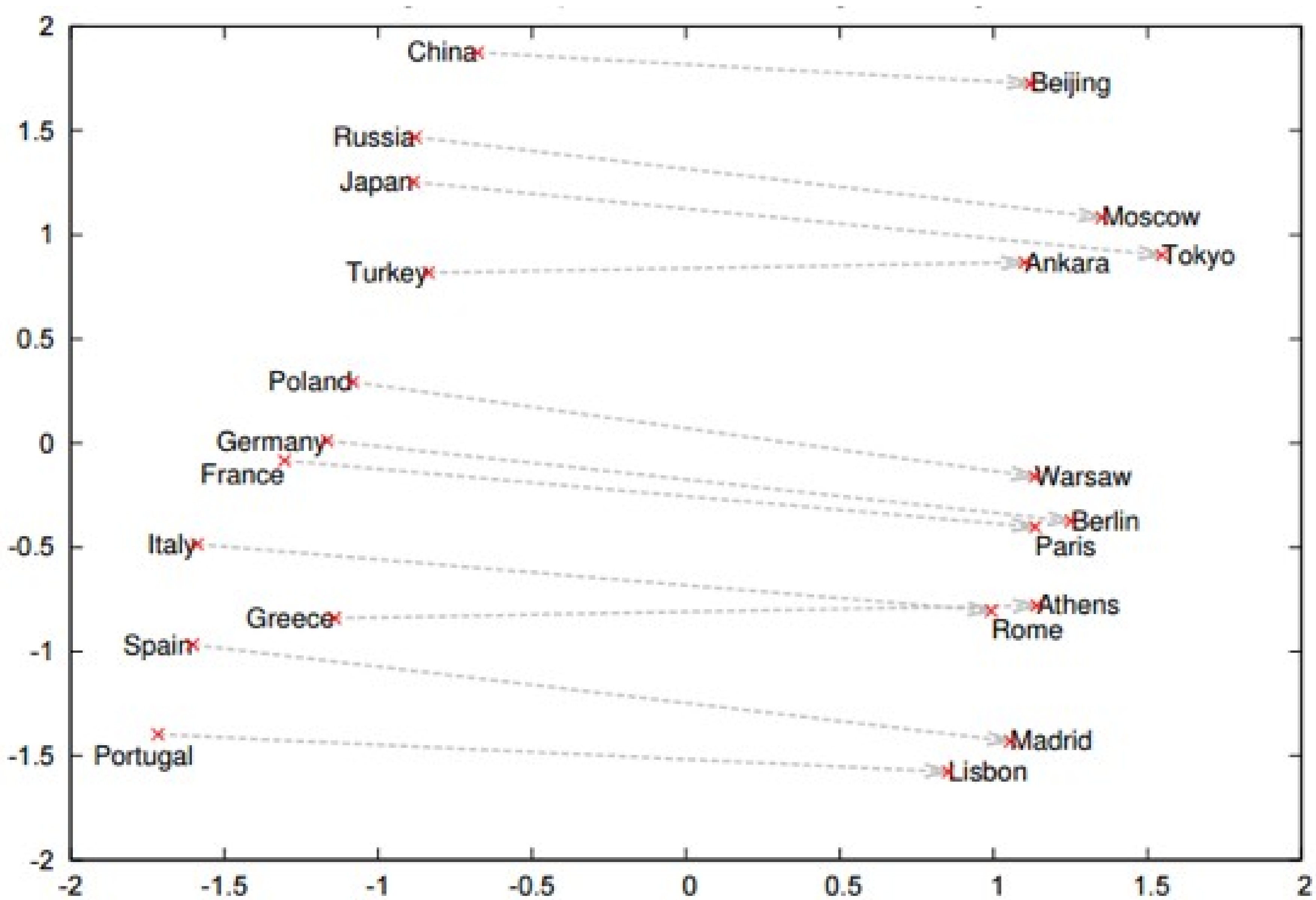
Input: running	Cosine similarity
runnin	0.758099
runing	0.702119
Running	0.69014
runnning	0.669039
sprinting	0.587385
runnung	0.578426
run	0.576671
walking/running	0.563114
runin	0.556682
walking	0.542137

Input: :)	Cosine similarity
:))	0.885355
=)	0.836011
:D	0.818340
;))	0.814380
(:	0.809806
:)))	0.808298
:-)	0.798115
:))))	0.777765
;)	0.772422
:-))	0.758584

Word2Vec : Interesting Results

- V('king') – V('man') + V('woman') ≈ V('queen')
- W('woman')−('man') ⍰ W('aunt')−W('uncle')
- V('Rome') – V('Italy') = V('France') – V('Paris')
- V('Iraq') – V('Violence') = V('Jordan')
- V('Human') -V('Animal') = V('Ethics')
- V('President') – V('Power') = V('Prime Minister')
- V('Library') – V('Books') = V('Hall')
- Analogy: V('Stock Market') ≈ V('Thermometer')

Word2Vec : Interesting Results



Word2Vec : Interesting Results

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Pre-training Language Representations

Overview

- Models are pretrained on very large corpora of text
  - Capture many aspects of the input text that are universally meaningful.
  - Allow downstream models to leverage linguistic information learned from larger datasets.
- The learned parameters are then applied to downstream tasks:
  - Feature-based approach
  - Fine-tuning approach
- Current state of the art in many NLP tasks.
- Most prominent works:
  - ELMo (Peters et al. 2018): best paper award at NAACL 2018.
  - BERT (Devlin et al. 2018): best paper award at NAACL 2019.
  - XLNet (Yang et al. 2019): published on arXiv in June 2019, current state of the art.

ELMo - Deep Contextualized Word Embeddings

Model architecture

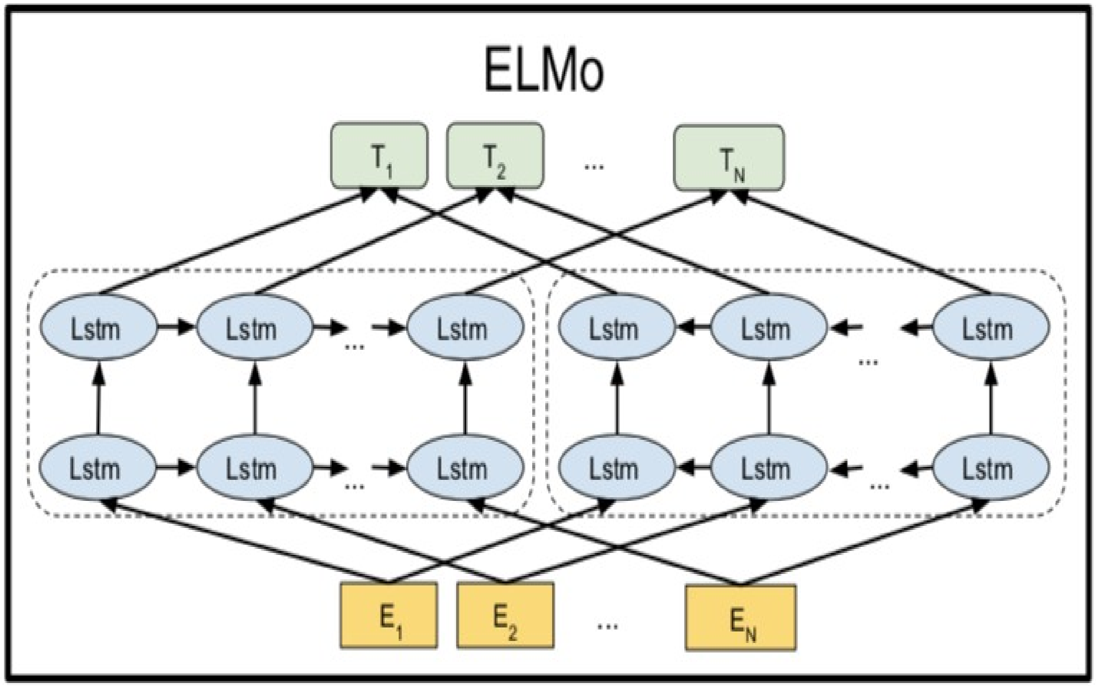


Figure from Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al.)

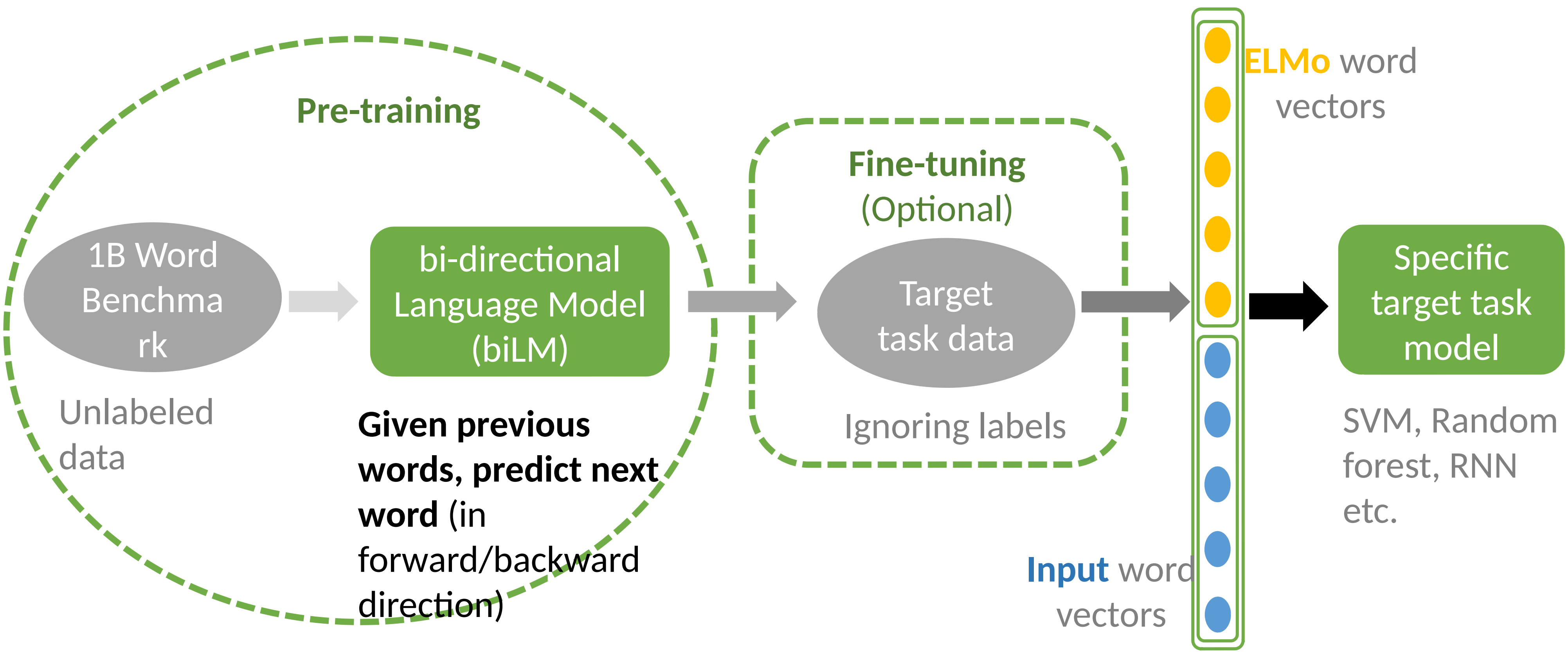
LSTM: Long short-term memory (Hochreiter and Schmidhuber, 1997)

The model learns to predict next token given the history in both direction:

- Forward: the history contains words before the target token
- Backward: the history contains words after the target token

ELMo - Deep Contextualized Word Embeddings

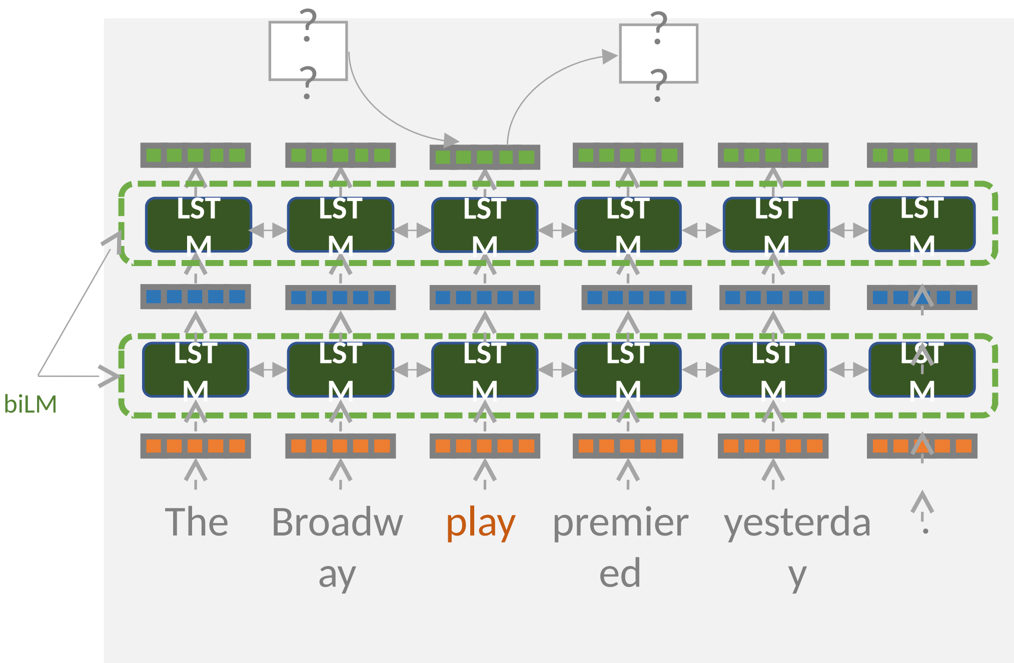
Training pipeline



ELMo - Deep Contextualized Word Embeddings

Pre-training & Fine-tuning

Pre-training → **Fine-tuning on specific tasks**



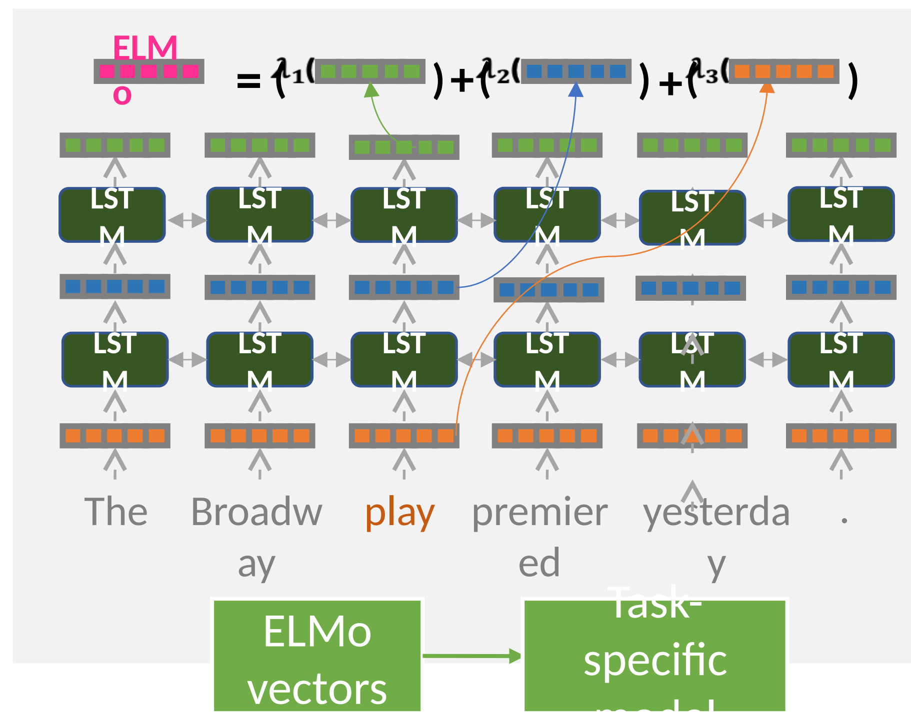


Figure recreated based on oral presentation of authors at NAACL 2018.

BERT - Pre-training of Deep Bidirectional Transformers for Language Understanding

Model Architecture

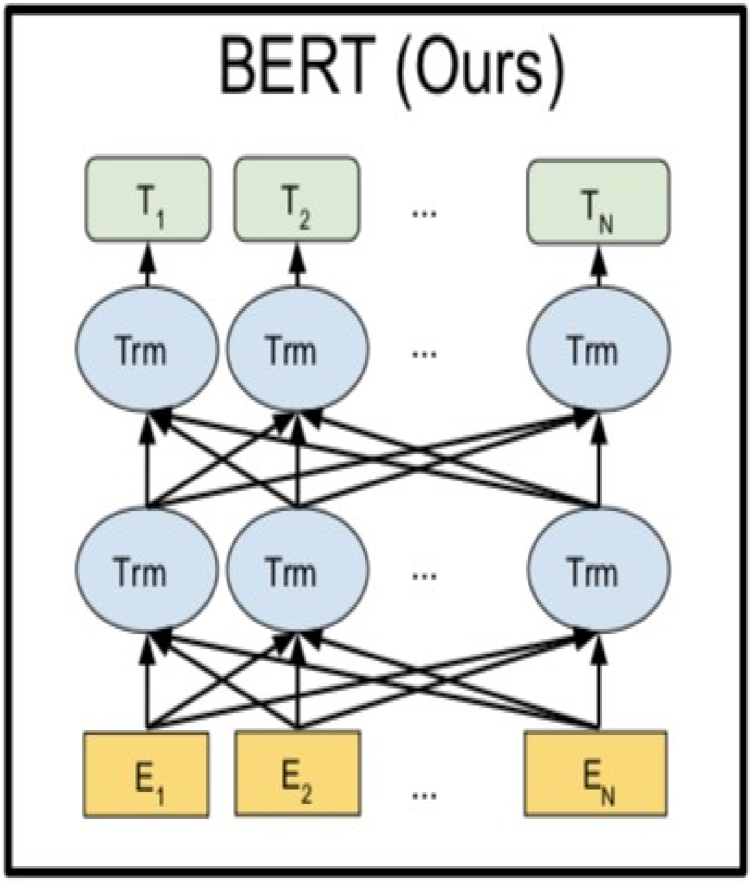


Figure from Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al.)

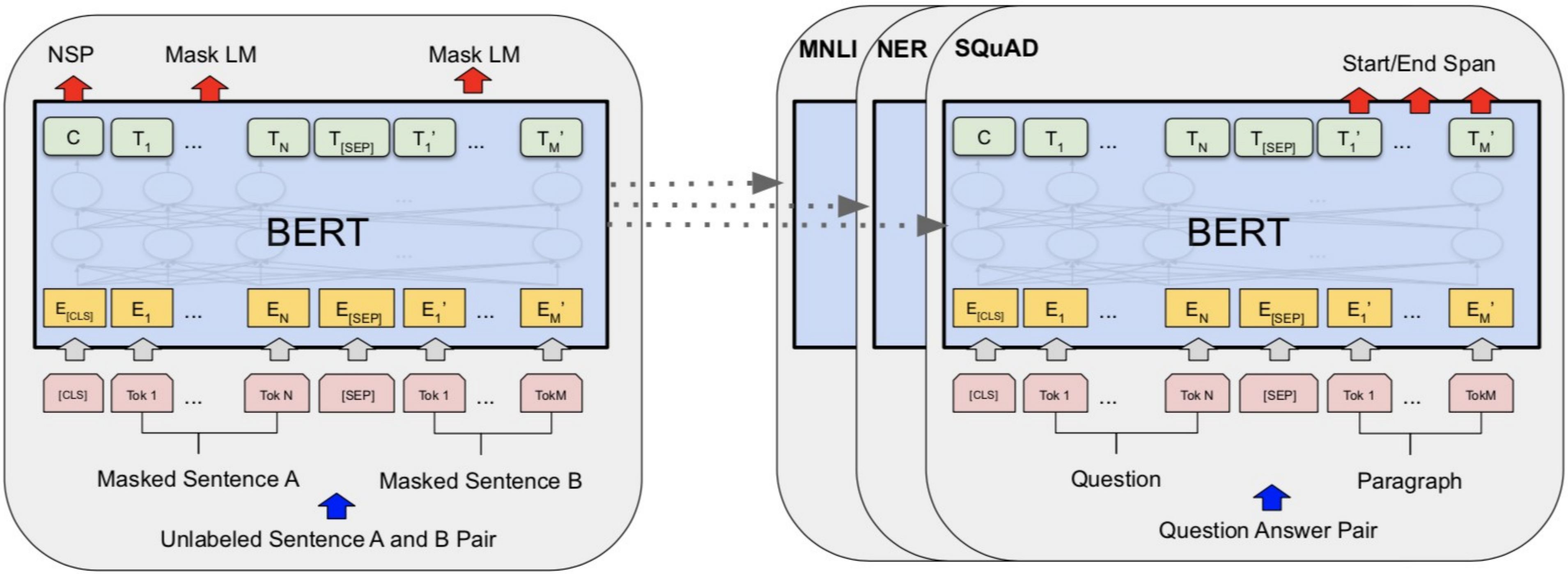
Trm: Transformer (Vaswani et al.)

The model learns to:

- Predict masked words in sentences
- Predict next sentences

BERT - Pre-training of Deep Bidirectional Transformers for Language Understanding

Pre-training & Fine-tuning



Pre-training

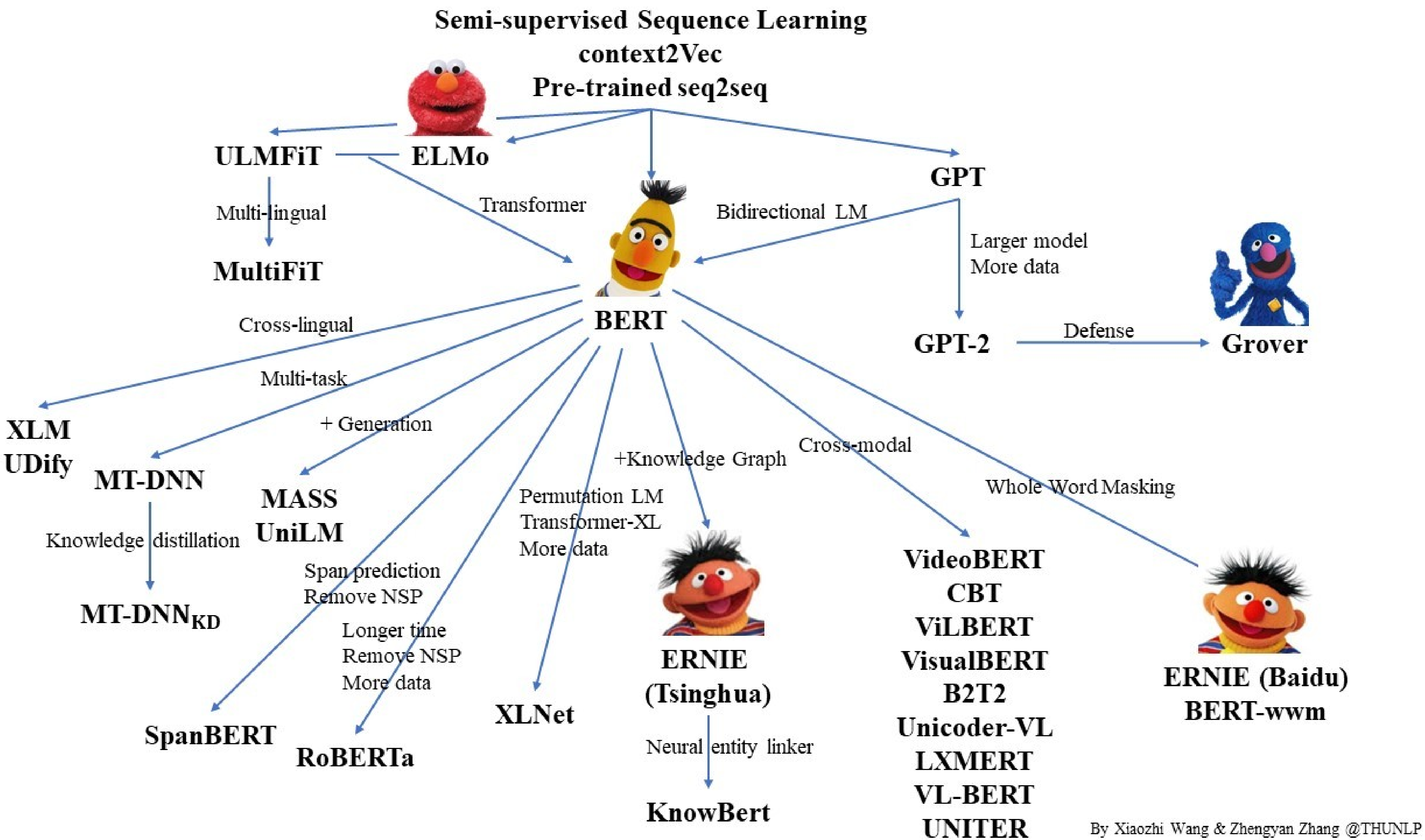
Fine-Tuning

Learn to predict masked words and next sentences.

Add a single output layer for specific tasks.

Figure from Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2019)

BERT and relatives
- Pre-trained Language Models



https://github.com/thunlp/PLMpapers

Glue benchmark
- General Language Understanding Evaluation (GLUE)

- https://gluebenchmark.com
- [Wang et al, 2019]
- A benchmark of nine sentence- or sentence-pair language understanding tasks built on established existing datasets and selected to cover a diverse range of dataset sizes, text genres, and degrees of difficulty,
- A diagnostic dataset designed to evaluate and analyze model performance with respect to a wide range of linguistic phenomena found in natural language, and
- A public leaderboard for tracking performance on the benchmark and a dashboard for visualizing the performance of models on the diagnostic set.

Glue benchmark

- Leaderboard (16/10/2019 – 9:30 UTC) – Rank 1 - 24

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
	1	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3	91.0	99.2	89.2	91.8	50.2
+	2	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.0	69.2	97.1	93.6/91.5	92.7/92.3	74.4/90.7	90.7	90.2	99.2	87.3	89.7	47.8
	3	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	91.1	90.7	98.8	88.7	89.0	50.1
	4	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	98.9	88.2	89.0	48.7
	5	XLNet Team	XLNet-Large (ensemble)		88.4	67.8	96.8	93.0/90.7	91.6/91.1	74.2/90.3	90.2	89.8	98.6	86.3	90.4	47.5
+	6	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
	7	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-
	8	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9	87.6	87.2	93.9	80.9	65.1	39.9
	9	XLM Systems	XLM (English only)		83.1	62.9	95.6	90.7/87.1	88.8/88.2	73.2/89.8	89.1	88.5	94.0	76.0	71.9	44.7
	10	Zhuosheng Zhang	SemBERT		82.9	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6	86.3	94.6	84.5	65.1	42.4
	11	Danqi Chen	SpanBERT (single-task training)		82.8	64.3	94.8	90.9/87.9	89.9/89.1	71.9/89.5	88.1	87.7	94.3	79.0	65.1	45.1
	12	Kevin Clark	BERT + BAM		82.3	61.5	95.2	91.3/88.3	88.6/87.9	72.5/89.7	86.6	85.8	93.1	80.4	65.1	40.7
	13	Nitish Shirish Keskar	Span-Extractive BERT on STILTs		82.3	63.2	94.5	90.6/87.6	89.4/89.2	72.2/89.4	86.5	85.8	92.5	79.8	65.1	28.3
	14	Jason Phang	BERT on STILTs		82.0	62.1	94.3	90.2/86.6	88.7/88.3	71.9/89.4	86.4	85.6	92.7	80.1	65.1	28.3
	15	廖亿	RGLM-Base (Huawei Noah's Ark Lab)		81.3	56.9	94.2	90.7/87.7	89.7/89.1	72.2/89.4	86.1	85.4	92.1	78.5	65.1	40.0
+	16	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hidden		80.5	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7	85.9	92.7	70.1	65.1	39.6
	17	Neil Houlsby	BERT + Single-task Adapters		80.2	59.2	94.3	88.7/84.3	87.3/86.1	71.5/89.4	85.4	85.0	92.4	71.6	65.1	9.2
	18	Zhuohan Li	Macaron Net-base		79.7	57.6	94.0	88.4/84.4	87.5/86.3	70.8/89.0	85.4	84.5	91.6	70.5	65.1	38.7
	19	蘇大鈞	SesameBERT-Base		78.6	52.7	94.2	88.9/84.8	86.5/85.5	70.8/88.8	83.7	83.6	91.0	67.6	65.1	35.8
+	20	MobileBERT Team	MobileBERT		78.5	51.1	92.6	88.8/84.5	86.2/84.8	70.5/88.3	84.3	83.4	91.6	70.4	65.1	34.3
	21	Linyuan Gong	StackingBERT-Base		78.4	56.2	93.9	88.2/83.9	84.2/82.5	70.4/88.7	84.4	84.2	90.1	67.0	65.1	36.6
	22	Huawei Noah's Ark Lab	TinyBERT (4-layers; 7.5x smaller and 9.4x faster than BERT-base)		75.4	43.3	92.6	86.4/81.2	81.2/79.9	71.3/89.2	82.5	81.8	87.7	62.9	65.1	33.7
	23	shijing si	bert+pos6		74.9	52.9	93.9	88.8/84.6	83.8/85.5	71.4/89.2	84.4	83.3	90.4	66.9	34.9	0.0
	24	GLUE Baselines	BiLSTM+ELMo+Attn	[illegible]	[illegible]	[illegible]	[illegible]	84.4/78.0	74.2/72.3	63.1/84.3	74.1	74.5	79.8	58.9	65.1	21.7

Click on a submission to see more information

Glue benchmark
- Leaderboard (16/10/2019 – 9:30 UTC) – Rank 16 - …

16	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hidden	80.5	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7	85.9	92.7	70.1	65.1	39.6
17	Neil Houlsby	BERT + Single-task Adapters	80.2	59.2	94.3	88.7/84.3	87.3/86.1	71.5/89.4	85.4	85.0	92.4	71.6	65.1	9.2
18	Zhuohan Li	Macaron Net-base	79.7	57.6	94.0	88.4/84.4	87.5/86.3	70.8/89.0	85.4	84.5	91.6	70.5	65.1	38.7
19	蘇大鈞	SesameBERT-Base	78.6	52.7	94.2	88.9/84.8	86.5/85.5	70.8/88.8	83.7	83.6	91.0	67.6	65.1	35.8
20	MobileBERT Team	MobileBERT	78.5	51.1	92.6	88.8/84.5	86.2/84.8	70.5/88.3	84.3	83.4	91.6	70.4	65.1	34.3
21	Linyuan Gong	StackingBERT-Base	78.4	56.2	93.9	88.2/83.9	84.2/82.5	70.4/88.7	84.4	84.2	90.1	67.0	65.1	36.6
22	Huawei Noah's Ark Lab	TinyBERT (4-layers; 7.5x smaller and 9.4x faster than BERT-base)	75.4	43.3	92.6	86.4/81.2	81.2/79.9	71.3/89.2	82.5	81.8	87.7	62.9	65.1	33.7
23	shijing si	bert+pos6	74.9	52.9	93.9	88.8/84.6	83.8/85.5	71.4/89.2	84.4	83.3	90.4	66.9	34.9	0.0
24	GLUE Baselines	BiLSTM+ELMo+Attn	70.0	33.6	90.4	84.4/78.0	74.2/72.3	63.1/84.3	74.1	74.5	79.8	58.9	65.1	21.7
		BiLSTM+ELMo	67.7	32.1	89.3	84.7/78.0	70.3/67.8	61.1/82.6	67.2	67.9	75.5	57.4	65.1	21.3
		Single Task BiLSTM+ELMo+Attn	66.5	35.0	90.2	80.2/68.8	55.5/52.5	66.1/86.5	76.9	76.7	76.7	50.3	65.1	27.9
		Single Task BiLSTM+ELMo	66.4	35.0	90.2	80.8/69.0	64.0/60.2	65.6/85.7	72.9	73.4	71.7	50.1	65.1	19.5
		GenSen	66.1	7.7	83.1	83.0/76.6	79.3/79.2	59.8/82.9	71.4	71.3	78.6	59.2	65.1	20.6
		BiLSTM+Attn	65.6	18.6	83.0	83.9/76.2	72.8/70.5	60.1/82.4	67.6	68.3	74.3	58.4	65.1	17.8
		BiLSTM	64.2	11.6	82.8	81.8/74.3	70.3/67.8	62.5/84.2	65.6	66.1	74.6	57.4	65.1	20.3
		InferSent	63.9	4.5	85.1	81.2/74.1	75.9/75.3	59.1/81.7	66.1	65.7	72.7	58.0	65.1	18.3
		Single Task BiLSTM	63.7	15.7	85.9	79.4/69.3	66.0/62.8	61.4/81.7	70.3	70.8	75.7	52.8	62.3	21.0
		Single Task BiLSTM+CoVe	63.6	14.5	88.5	81.4/73.4	67.2/64.1	59.4/83.3	64.5	64.8	75.4	53.5	61.6	20.6
		BiLSTM+CoVe+Attn	63.1	8.3	80.7	80.0/71.8	69.8/68.4	60.5/83.4	68.1	68.6	72.9	56.0	65.1	18.3
		Single Task BiLSTM+CoVe+Attn	63.1	14.5	88.5	79.7/68.6	57.2/53.6	60.1/84.1	71.6	71.5	74.5	52.7	64.4	23.8
		BiLSTM+CoVe	62.9	18.5	81.9	78.7/71.5	64.4/62.7	60.6/84.9	65.4	65.7	70.8	52.7	65.1	17.6
		Single Task BiLSTM+Attn	62.8	15.7	85.9	80.3/68.5	59.3/55.8	62.9/83.5	74.2	73.8	77.2	51.9	55.5	24.9
		DisSent	61.9	4.9	83.7	81.7/74.1	66.1/64.8	59.5/82.6	58.7	59.1	73.9	56.4	65.1	15.9
		Skip-Thought	61.3	0.0	81.8	80.8/71.7	71.8/69.7	56.4/82.2	62.9	62.8	72.9	53.1	65.1	12.2
		CBOW	58.6	0.0	80.0	81.5/73.4	61.2/58.7	51.4/79.1	56.0	56.4	72.1	54.1	62.3	9.2