Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier
LIRMM

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Lexique et analyse sémantique de textes

structures, acquisitions, calculs, et jeux de mots

M. Lafourcade

Christian Boitet	Université Joseph Fourier - Grenoble 1 - LIG	Père spirituel
Vladimir Fomichov	Higher School of Economics, Moscou	Rapporteur
Marianne Huchard	Université Montpellier 2 - LIRMM	Modèle
Violaine Prince	Université Montpellier 2 - LIRMM	Mère spirituelle
Christian Retoré	Université Bordeaux 1 - LaBRI	Rapporteur
Eric Wehrli	Université de Genève - LATL	Examinateur
Michael Zock	Université de Marseille - LIF	Examinateur
Pierre Zweigenbaum	LIMSI - CNRS	Rapporteur

Contexte

… et problématique

Analyse de textes

structures de représentation du sens

- thématique : vecteurs ? mots-clés ?
- sémantique : quelles relations entre syntagmes ?

désambiguïsation lexicale - inventaire de sens ?

fonctions lexicales ?

- magn(fièvre) = forte fièvre (?) magn(maison) = villa

référence et substitution ?

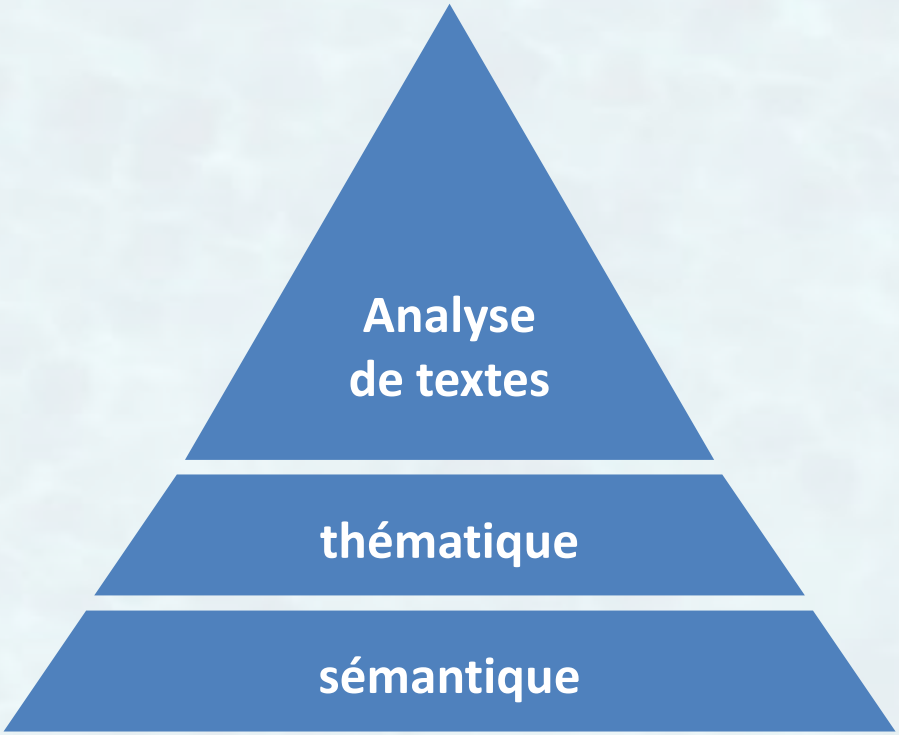
- chat ⇒ matou / animal (?) mouche ⇒ drosophile / insecte

Calcul / acquisition des structures

vecteurs / réseaux lexico-sémantiques

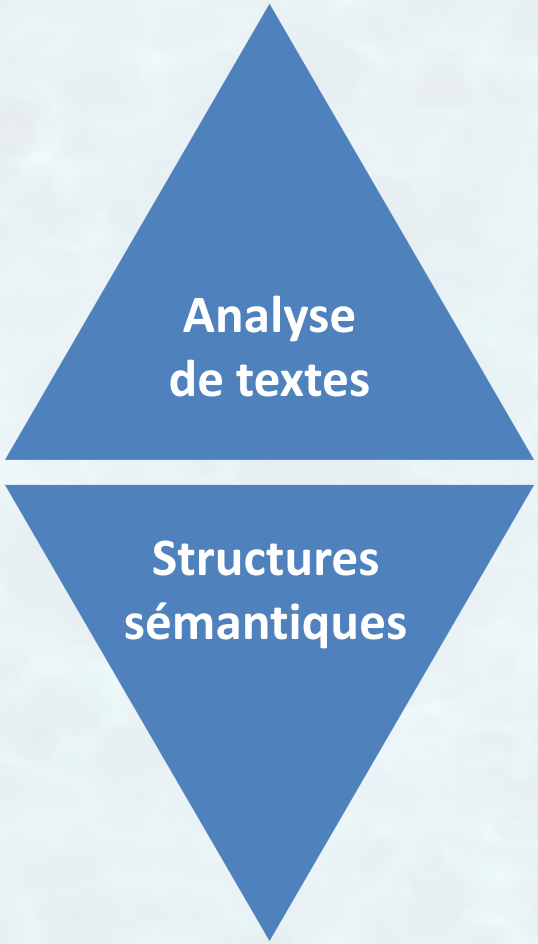
Contexte

structures, acquisitions, calculs, et jeux de mots



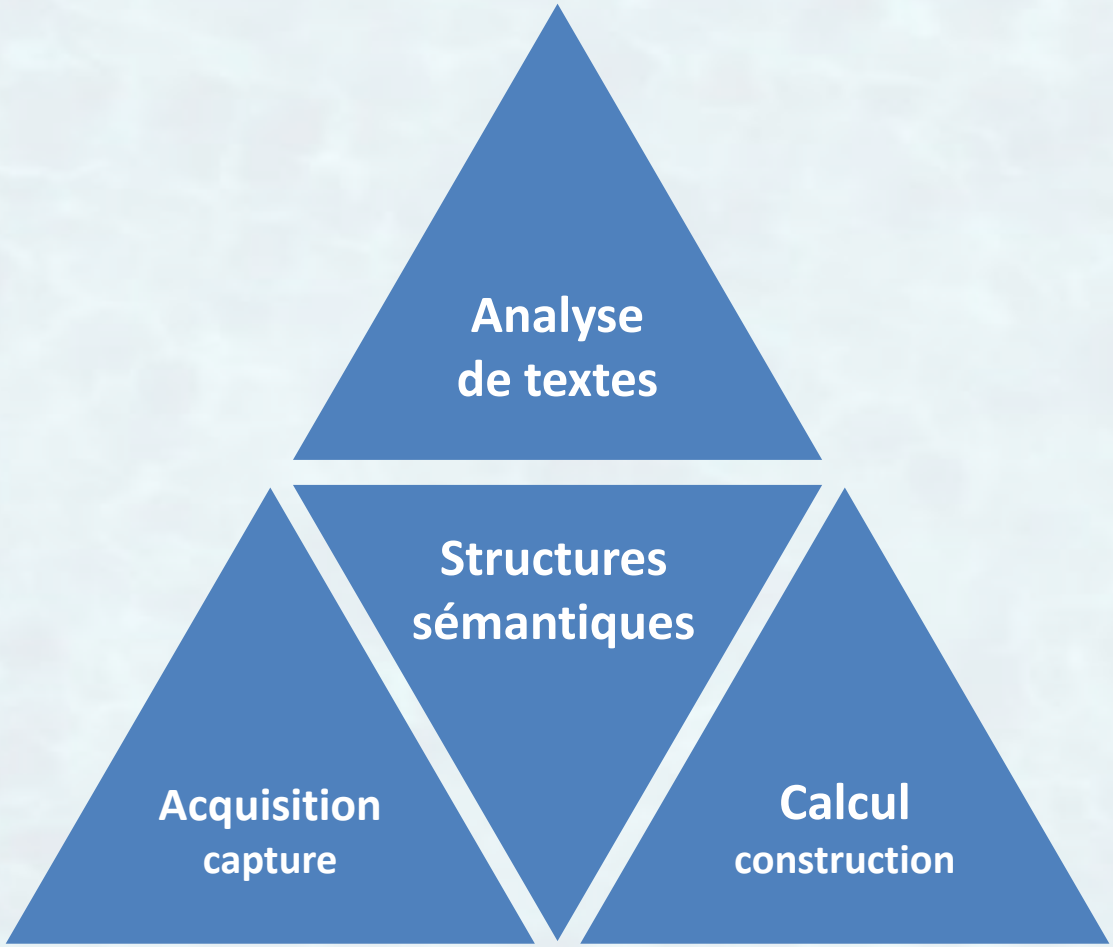
Contexte

structures, acquisitions, calculs, et jeux de mots



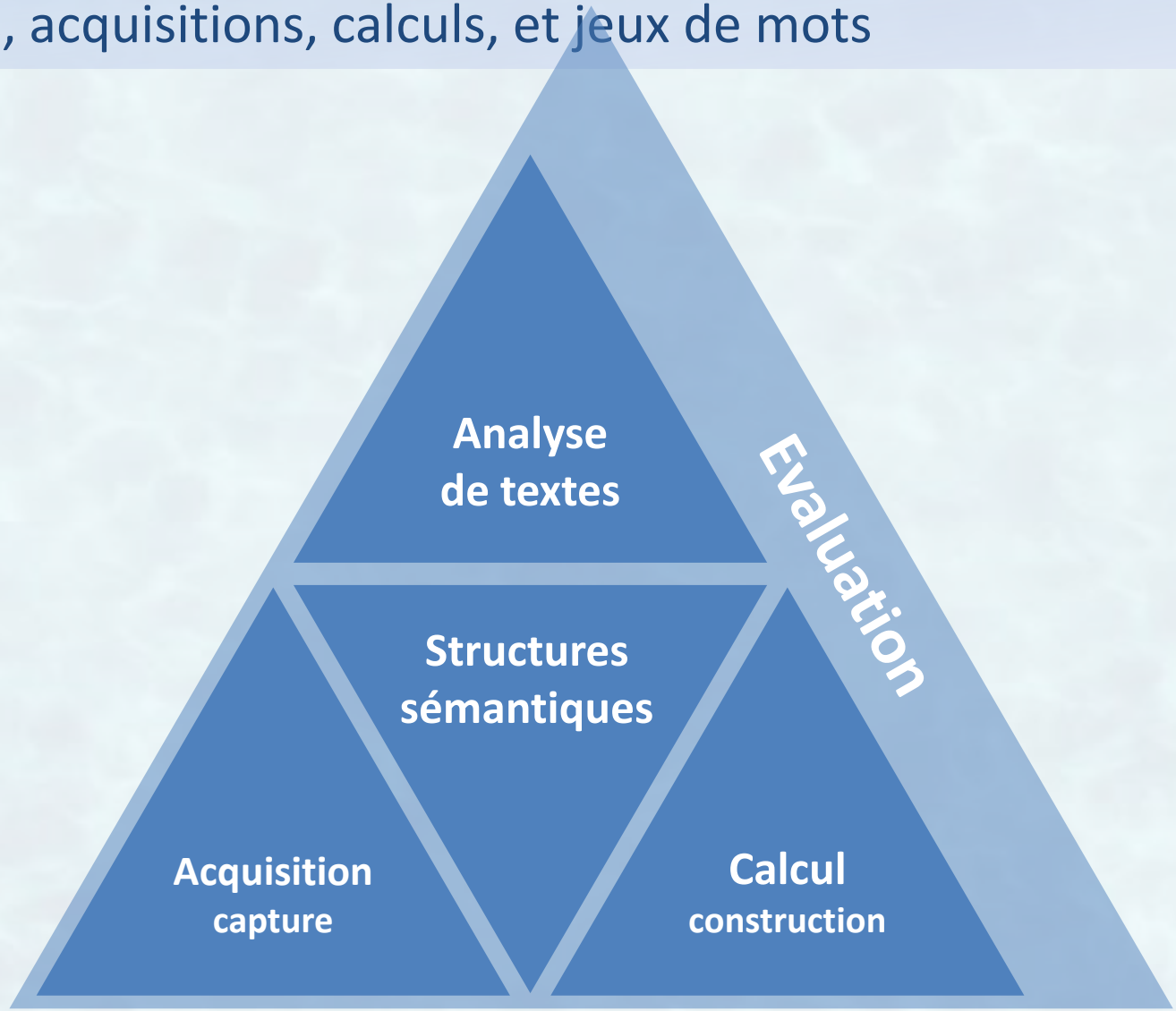
Contexte

structures, acquisitions, calculs, et jeux de mots



Contexte

structures, acquisitions, calculs, et jeux de mots



Plan

Calcul de vecteurs d'idées

- Principes, variétés et construction
- Fonctions lexicales
- Pliage et dépliage

bouclage

permanence de l'acquisition

Acquisition de relations lexicales

- JeuxDeMots : principe et résultats
- Inventaire de sens / usages
- Tentative d'évaluation qualitative : AKI

consensus populaire

raffinement des structures

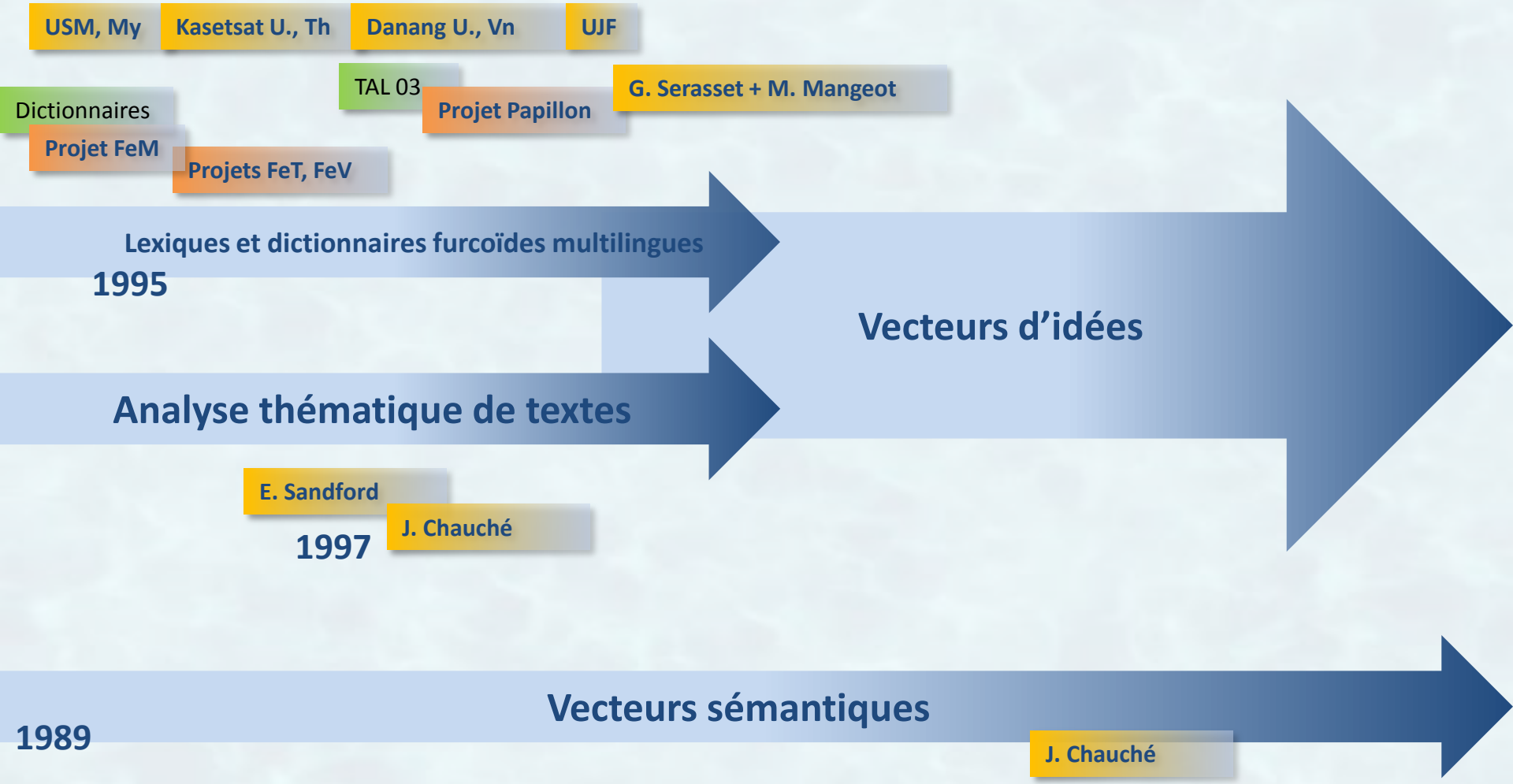
Analyse de textes

- Thématique par vecteurs
- Thématique par mot-clés
- Métaheuristique bioinspirée

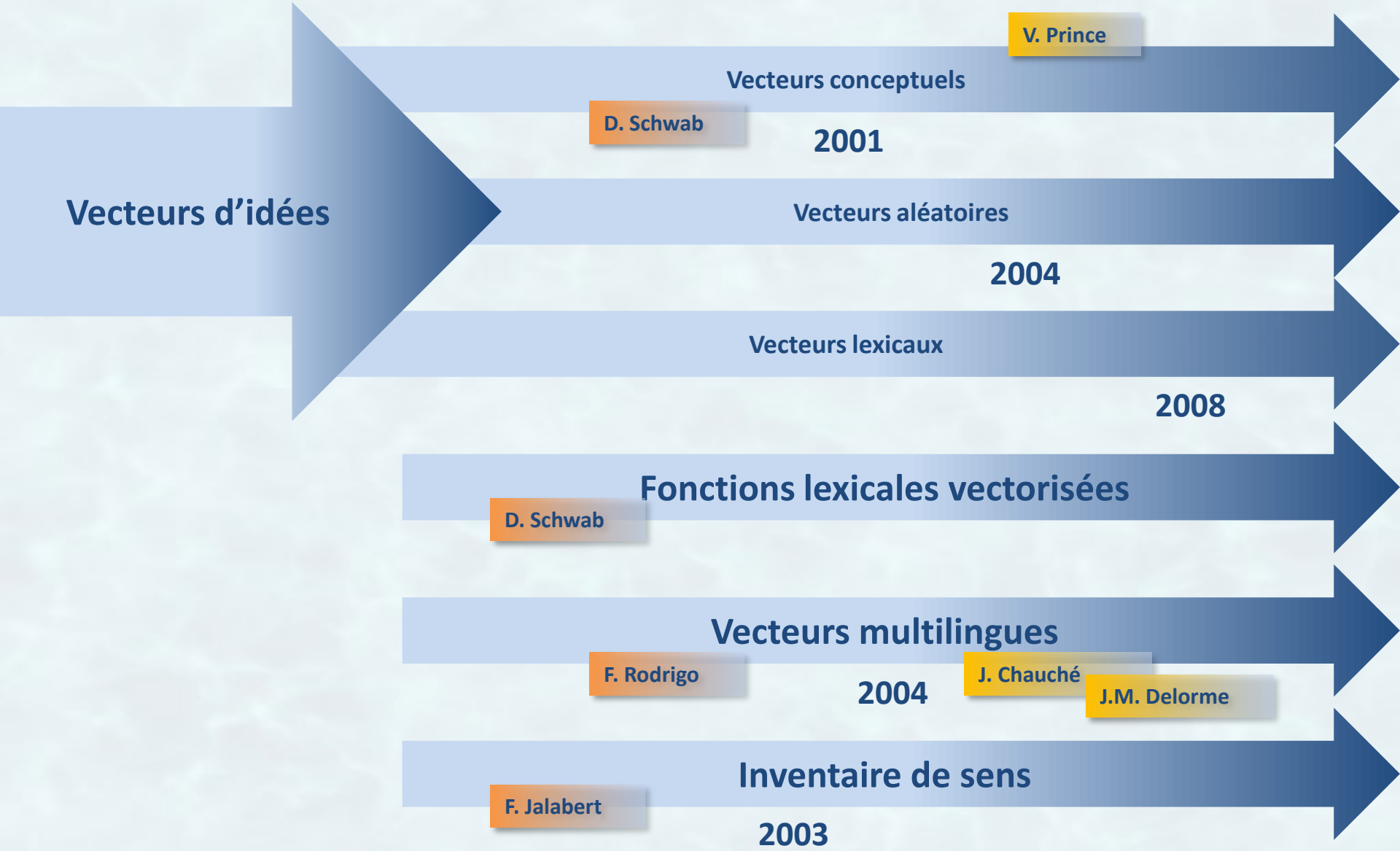
propagation et diffusion

activation et inhibition

Chronologie



Chronologie



Chronologie

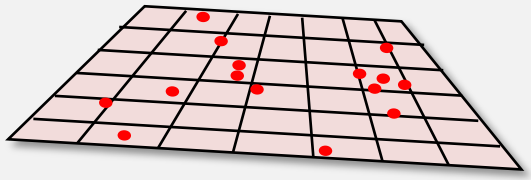
Vecteurs d'idées

vecteurs conceptuels

Une idée
= une combinaison de concepts = un vecteur
⇒ un espace vectoriel (en fait, une famille génératrice)

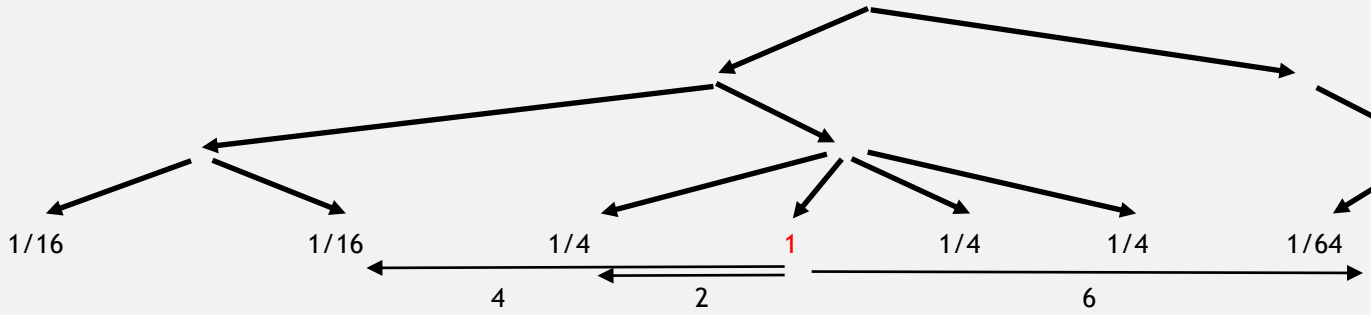
Un concept
= une idée importante/récurrente/élémentaire = un vecteur (ancrage)
= combination de lui-même + voisinage

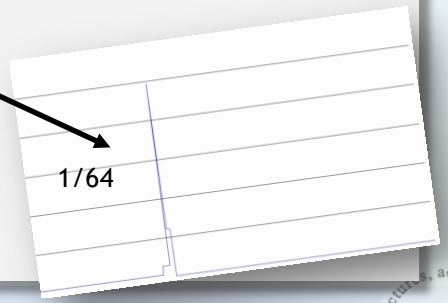
Un Espace des sens
= une collection de vecteurs (un espace peuplé)



Exemple : thésaurus Larousse = 873 concepts
⇒ des vecteurs de dimension 873

Exploiter la hiérarchie du thésaurus pour représenter le voisinage





Vecteurs conceptuels

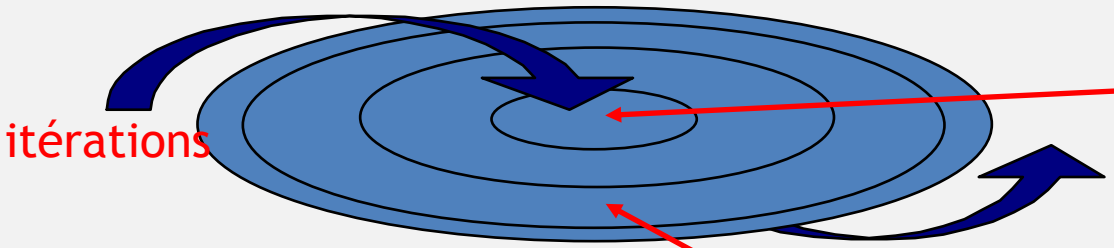
Construction à partir d'une ressource

Comment peupler un espace?

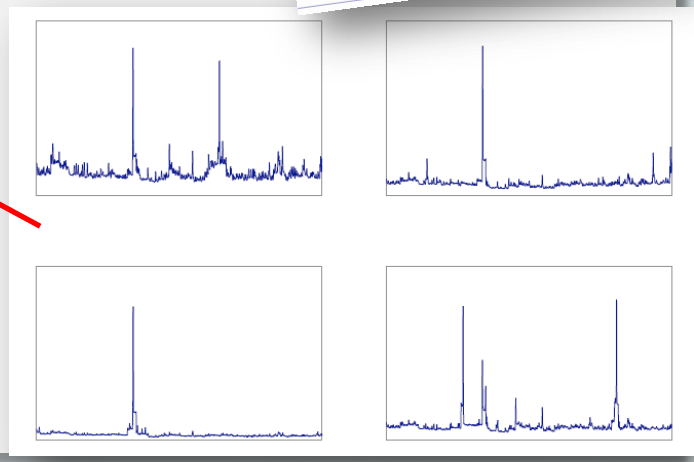
Par exemple, par analyse thématique

- Définitions dictionnairiques (Larousse, Hachette, etc.)
- Définition = genre + différence
- Vecteur nul pour un terme non encore indexé

permanence de l'acquisition



noyau

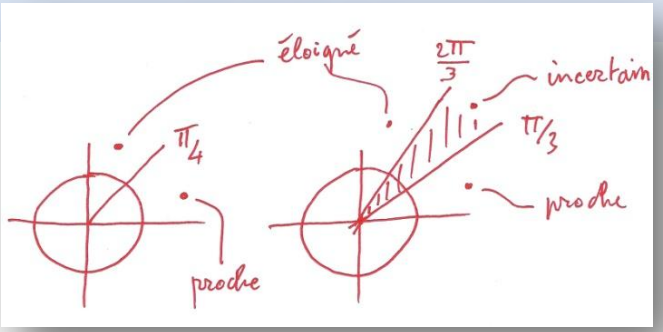


Autres types de ressources possibles

- Liste de termes
- Réseau lexical

Vecteurs d'idées

opération de comparaison



Distance angulaire $D_A(X, Y)$ = angle (X, Y)

$0 \leq D_A(X, Y) \leq \pi/2$
si 0 alors colinéaire – même idée
si $\pi/2$ alors rien en commun

La norme n'intervient pas !

$$D_A(X,X) = 0$$
$$D_A(X,Y) = D_A(Y,X)$$
$$D_A(X,Y) + D_A(Y,Z) \geq D_A(X,Z)$$

$$D_A(X \oplus X, Y \oplus X) = D_A(X, Y \oplus X) \leq D_A(X,Y)$$
$$D_A(X \oplus Z, Y \oplus Z) \leq D_A(X,Y)$$

Interprétation géométrique

$Sim(X,Y) = \cos(D_A(X, Y))$ $\quad$ $Disim^2(X,Y) = 1 - Sim^2(X,Y)$

$Ponderation(X,Y) = Sim(X,Y) \,/\, Disim(X,Y) = \cot(D_A(X, Y))$

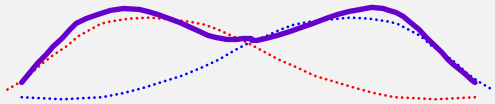
Vecteurs d'idées

opérations

Somme

$V = X \oplus Y \Rightarrow v_i = x_i + y_i$

Élément neutre : **0 (vecteur nul)**

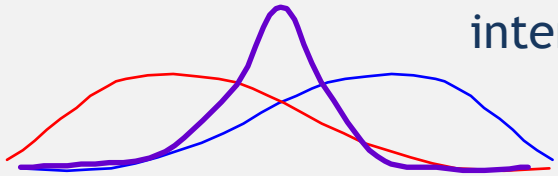


moyenne

Produit terme à terme

$V = X \otimes Y \Rightarrow v_i = x_i * y_i$

Élément neutre : **1 (vecteur unitaire)**



intersection

Contextualisation faible

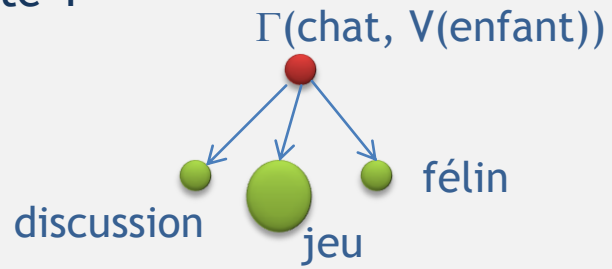
$V = \gamma(X, Y) = X \oplus (X \otimes Y)$

Sélection des idées de X présentes dans le contexte Y

Contextualisation forte du terme w avec le contexte Y

$V(w) = \Gamma(w, Y) = \beta_1 V(w_1) \oplus \ldots \oplus \beta_k V(w_k)$

et β_i = Ponderation(Y, $V(w_k)$)



Vecteurs d'idées

fonctions lexicales vectorielles

Distance de synonymie relative

$Syn_R(X, Y, Z)$ — Z est une référence

D. Schwab

V. Prince

$$Syn_R(X, Y, Z) = D_A(\gamma(X, Z), \gamma(Y, Z))$$

TAL 02

NLPRS

D_A(charbon, nuit) = 0.9
Syn_R(charbon, nuit, couleur) = 0.4
Syn_R(charbon, nuit, noir) = 0.35

utile en structuration terminologique et analyse de métaphore

Travail sur l'antonymie

complémentaire
scalaire
duale

lexicalisation croissante

D. Schwab (DEA)

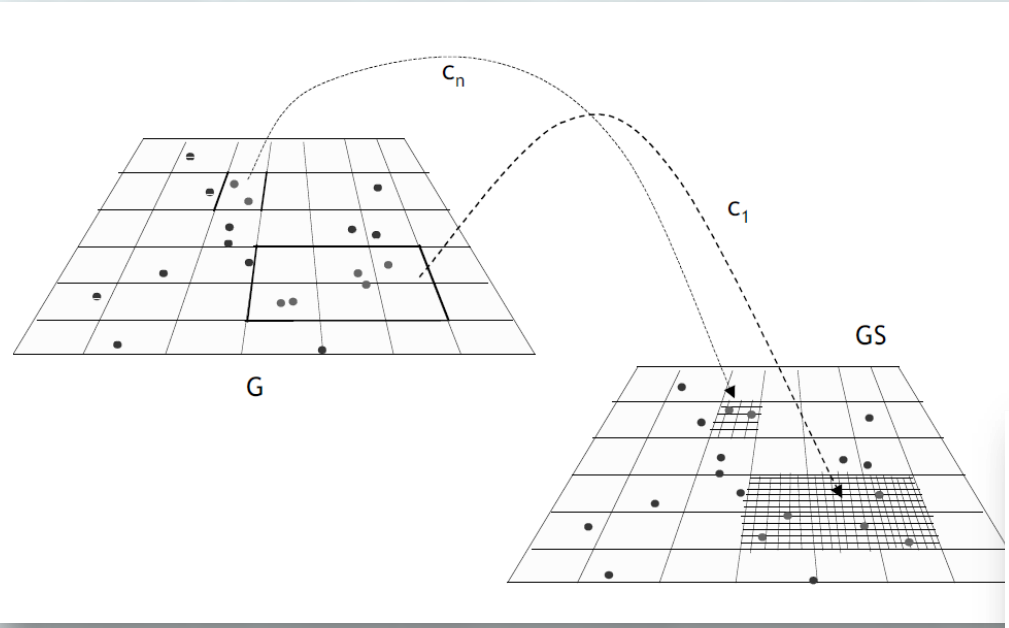
COLING

utile en gestion de la négation dans l'analyse thématique

Vecteurs d'idées

pliage et dépliage

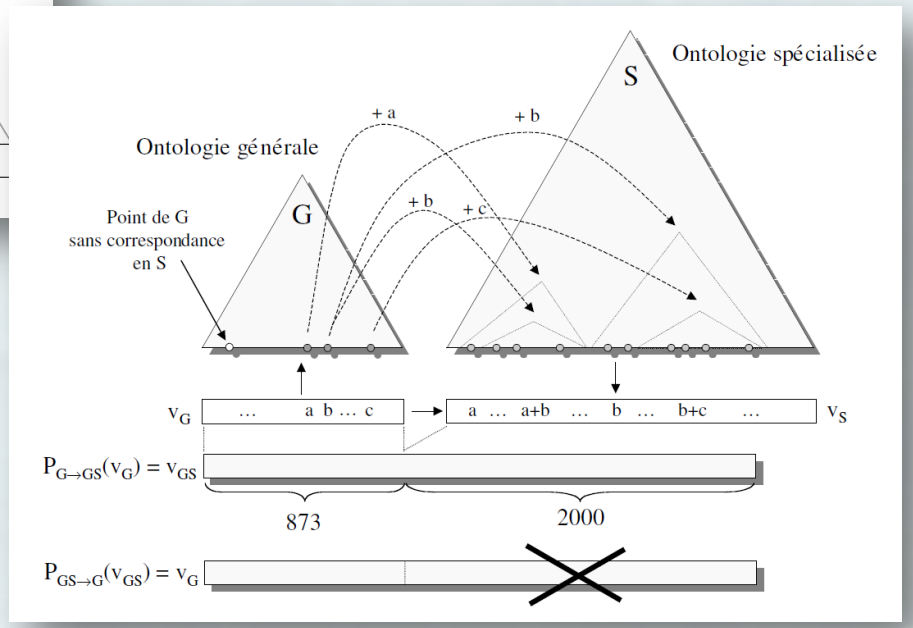
Comment étendre un espace à un domaine de spécialité ?



séparation dans S
si X,Y ∈ GS : $D_{A,G}(X, Y) > D_{A,GS}(X, Y)$

D. Schwab

V. Prince



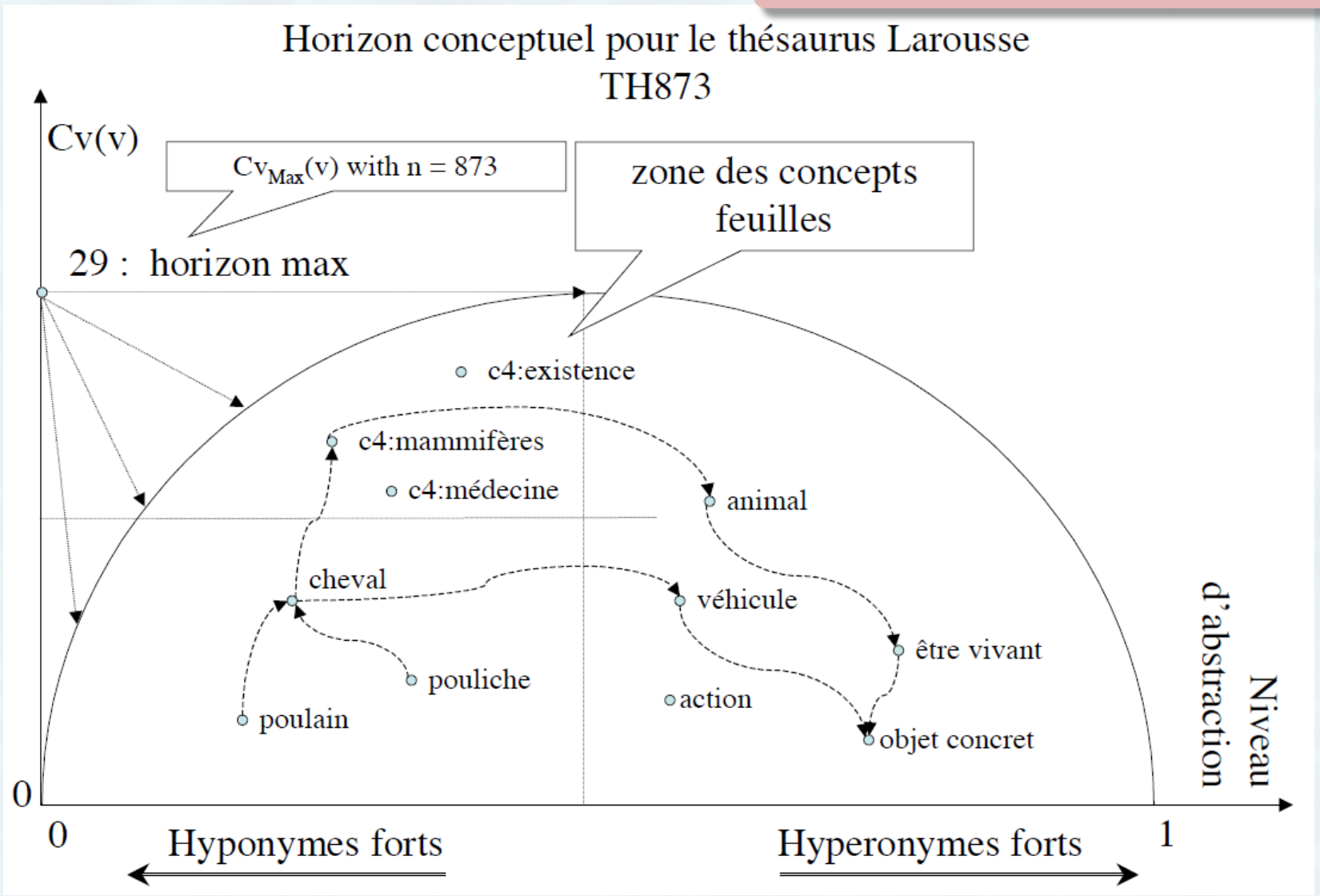
myopie hors S ?
si X,Y ∉ GS : $D_{A,G}(X, Y) < D_{A,GS}(X, Y)$

IEEE Trans. on Systems, Man, and Cybernetics

Vecteurs d'idées

horizon conceptuel

une limite indépassable ?



Vecteurs d'idées - autres formes

vecteurs aléatoires/émergents

Choix d'une taille de vecteurs

Placement initial aléatoire

LREC

Random indexing - Sahlgren

Fonction d'agglomération
⇒ analyse de définitions
⇒ moyenne pondérée de termes

Fonction de répulsion nécessaire
car pas d'ancrage

Bonne répartition sur l'espace
Ajustable avec recalcul
Composantes non décodables

vecteurs lexicaux

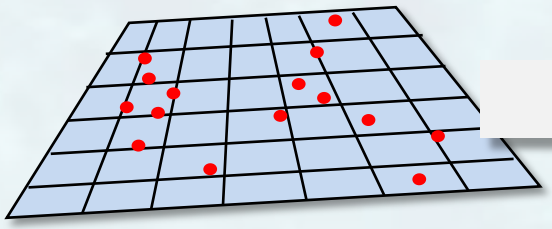
Compilation locale d'un réseau lexical
⇒ liste de mots pondérée

≈ Salton ?

Forte précision mais rappel moindre

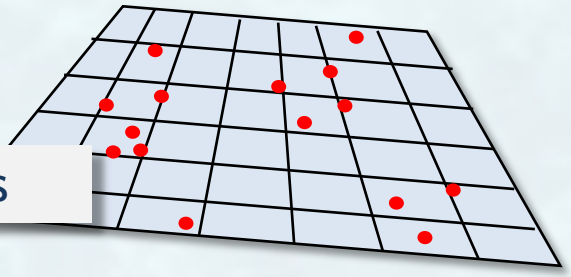
Ajustable sans recalcul

Fortement décodables mais déconceptualisés
⇒ chat, matou, minet

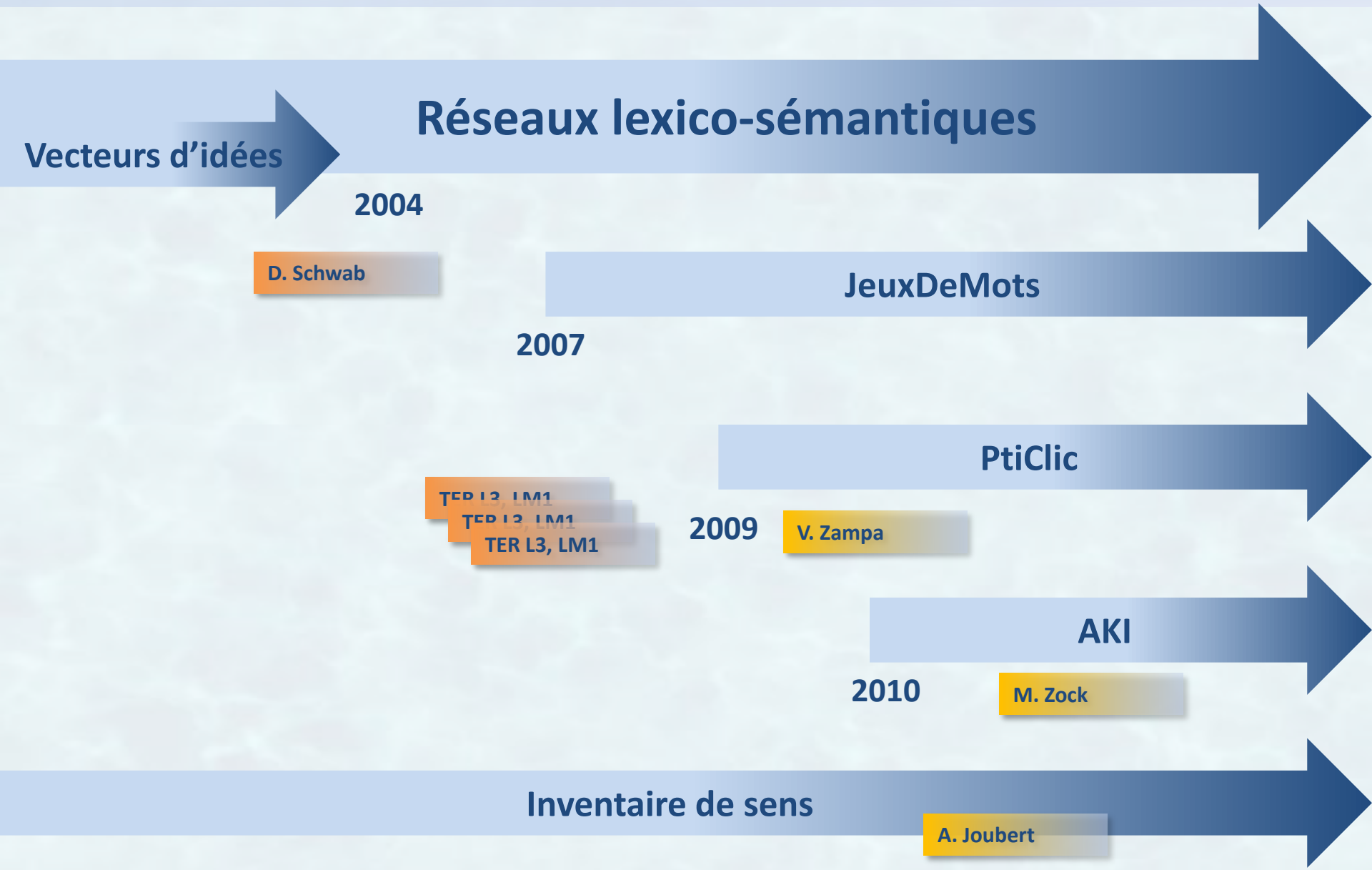


invariance globale des voisinages

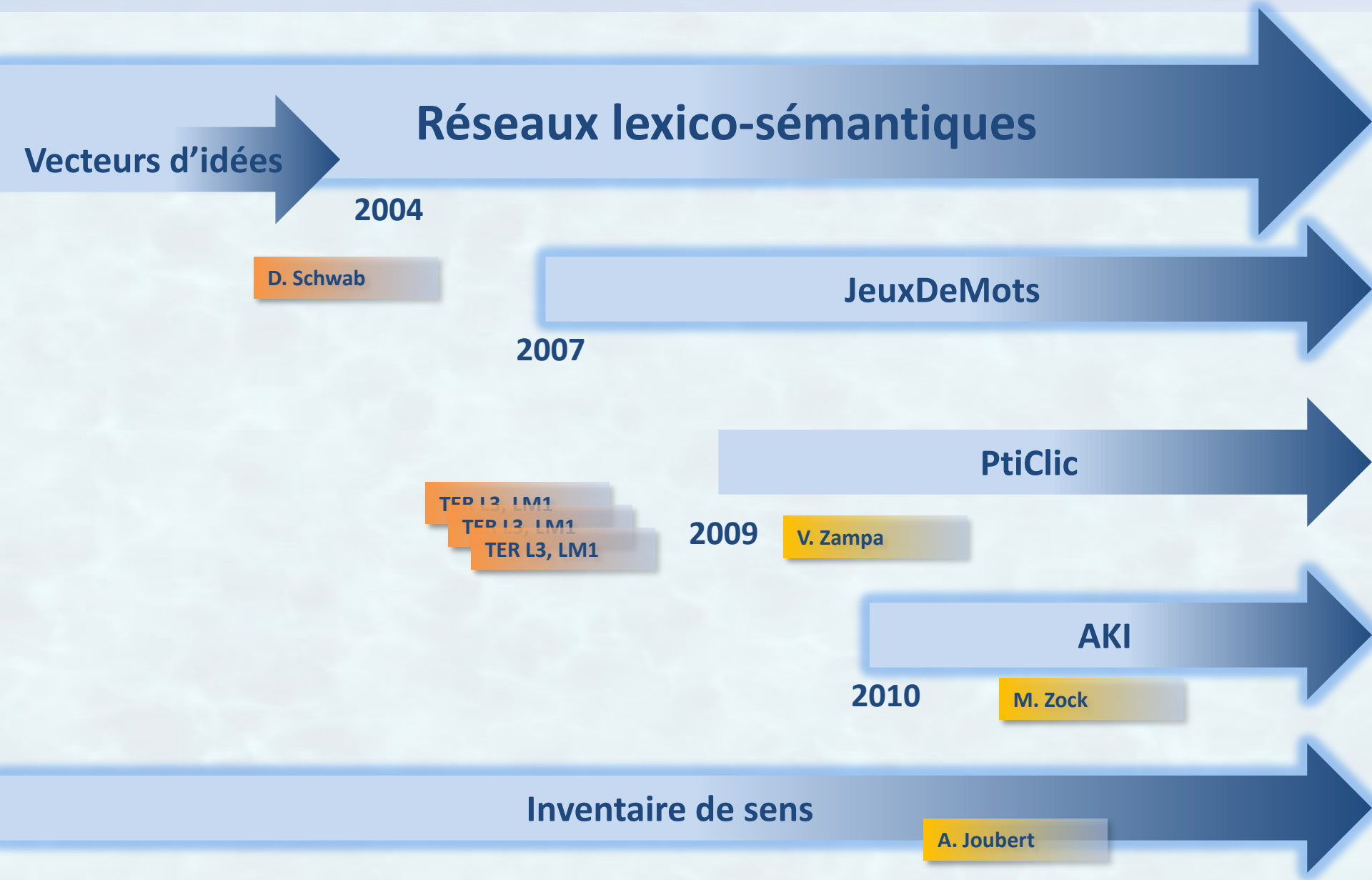
J. Chauché – TAL89



Chronologie



Chronologie



Réseaux lexico-sémantiques

Vers une structure générique…

Pour quoi faire ?

- Utiles pour des applications nécessitant
  - des informations terminologiques
  - des représentations du monde
  - des informations linguistiques
- … et en TAL : Analyse de textes, TAO, résumé, RI, ToT, correction
  - vers la compréhension automatique du discours, de modèles, etc.
  - $\Rightarrow$ Analyse sémantique

Comment ?

- À la main…
- Automatiquement (extraction à partir de textes)
  - mais les ‘savoirs ’ ne sont pas toujours dans les textes
    - Ni exclusivement
    - Ni totalement

Réseaux lexico-sémantiques

une structure de graphe

Des nœuds = des éléments du lexiques

Typés : terme – partie du discours (N, V, Adj, Adv, …) – des concepts …

Pondérés : interprétable comme { fréquence d'usage ; importance conceptuelle }

Des liens entre nœuds = des relations lexico-sémantiques

Typés : information portée par la relation

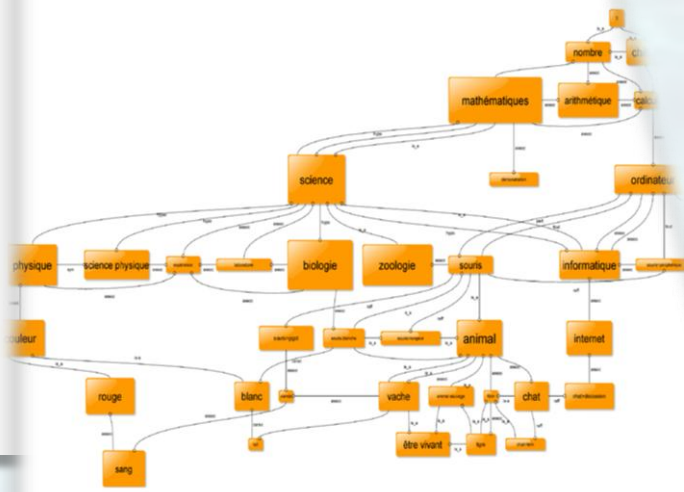
Orientés : chat → félin ≠ félin → chat

Pondérés : force de la relation

Hypothèse : plus un poids est fort

⇒ plus la relation est active

⇒ plus elle a des chances d'être valide



Réseaux lexico-sémantiques

toutes les relations potentiellement utiles

Associations libres : chat $\Rightarrow$ animal, chien, souris, …

Relations ontologiques : hyperonymes – hyponymes – partie de – tout – susbtance/matière …

Relations lexicales : synonymie – contraire – locutions – même famille lexicale – magn/antimagn …

Relations de typicalité/rôles sémantiques : agent – patient – instrument – lieu – rôle télique (fonction/but) – rôle agentif (mode de création)…

Usages : plus général qu'acception/sens

sapin (arbre), sapin (sapin de Noël),
sapin (bois), sapin(cercueil), sapin (fiacre)

Réseaux lexico-sémantiques

autres réseaux existants

Quelques exemples

WordNet

(25 ans) Miller et al.- *The 2006 version contains 155,287 words organized in 117,659 synsets for a total of 206,941 word-sense pairs.*

coût = 6 M$

HowNet

(début fin des années 80) Dong et Dong – Anglais-Chinois – environ 100k termes dans chaque langue – 2600 concepts

Wolf

(2006) Sagot et al. – Adaptation au **Français** de WordNet (et EuroWN)

BabelNet et WordNet++

(2010) Navigli et al. – croisement automatique de WordNet et Wikipedia

Construction manuelle et/ou par extraction
+ vérification manuelle

Acquisition de relations

par consensus populaire – jeux sérieux

Peut-on construire un tel réseau autrement ?

Avec des jeux proposés à des non-spécialistes

Von Ahn

Hypothèse : rapide – gratuit – efficace

TAL

TALN

Expérience de faisabilité : projet JeuxDeMots

Principe : demander à des joueurs de renseigner une fonction *f* pour un terme *x*

Rendre la tâche

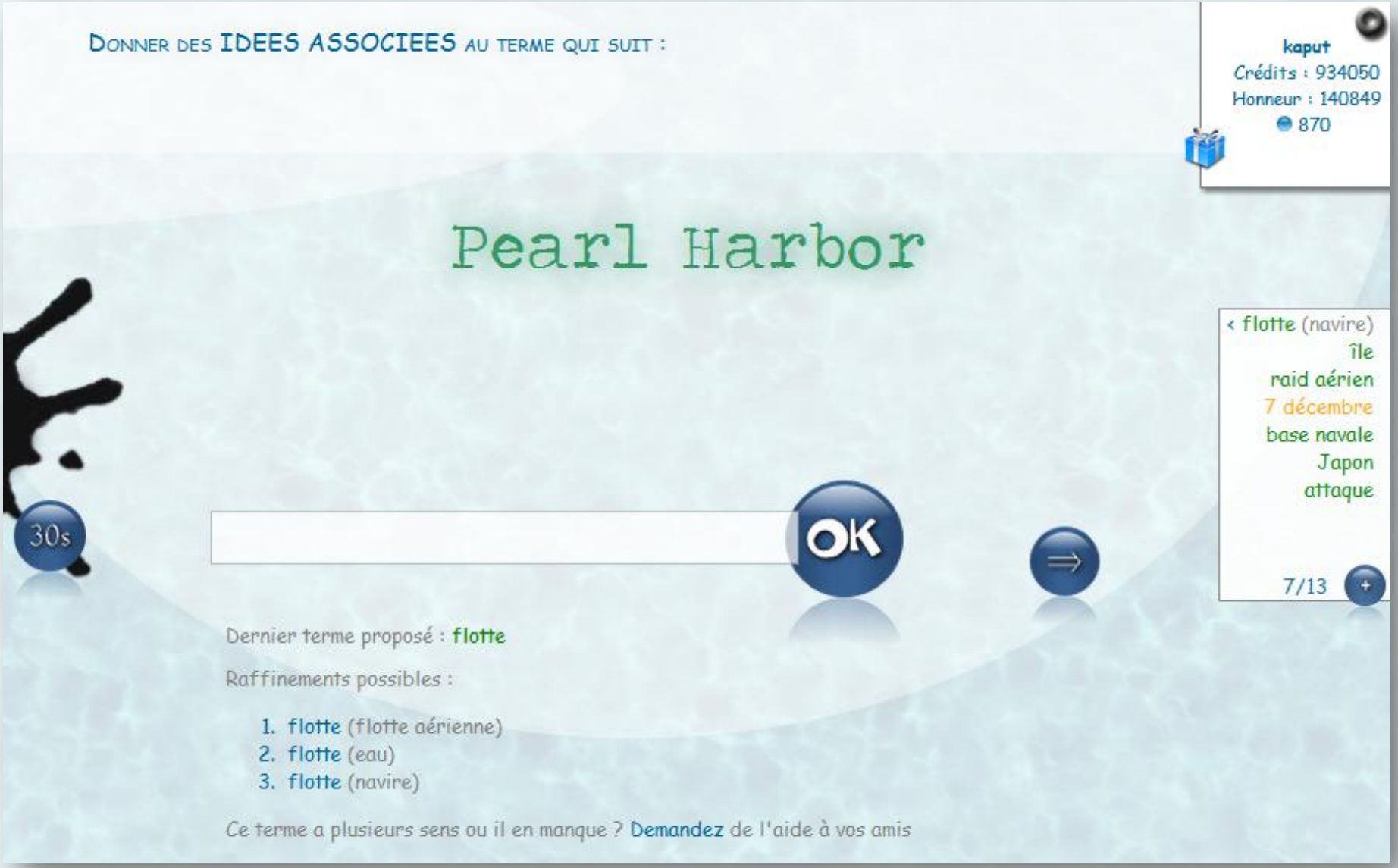
amusante en permettant la comparaison avec les propositions d'un autre joueur

communautaire /compétitive par le gain de points – classement - forum

appropriable par les joueurs avec des possibilités de contrôle partiel sur le jeu

auto-évaluation à l'aune des autres joueurs – satisfaction/frustration

JeuxDeMots - exemple de partie

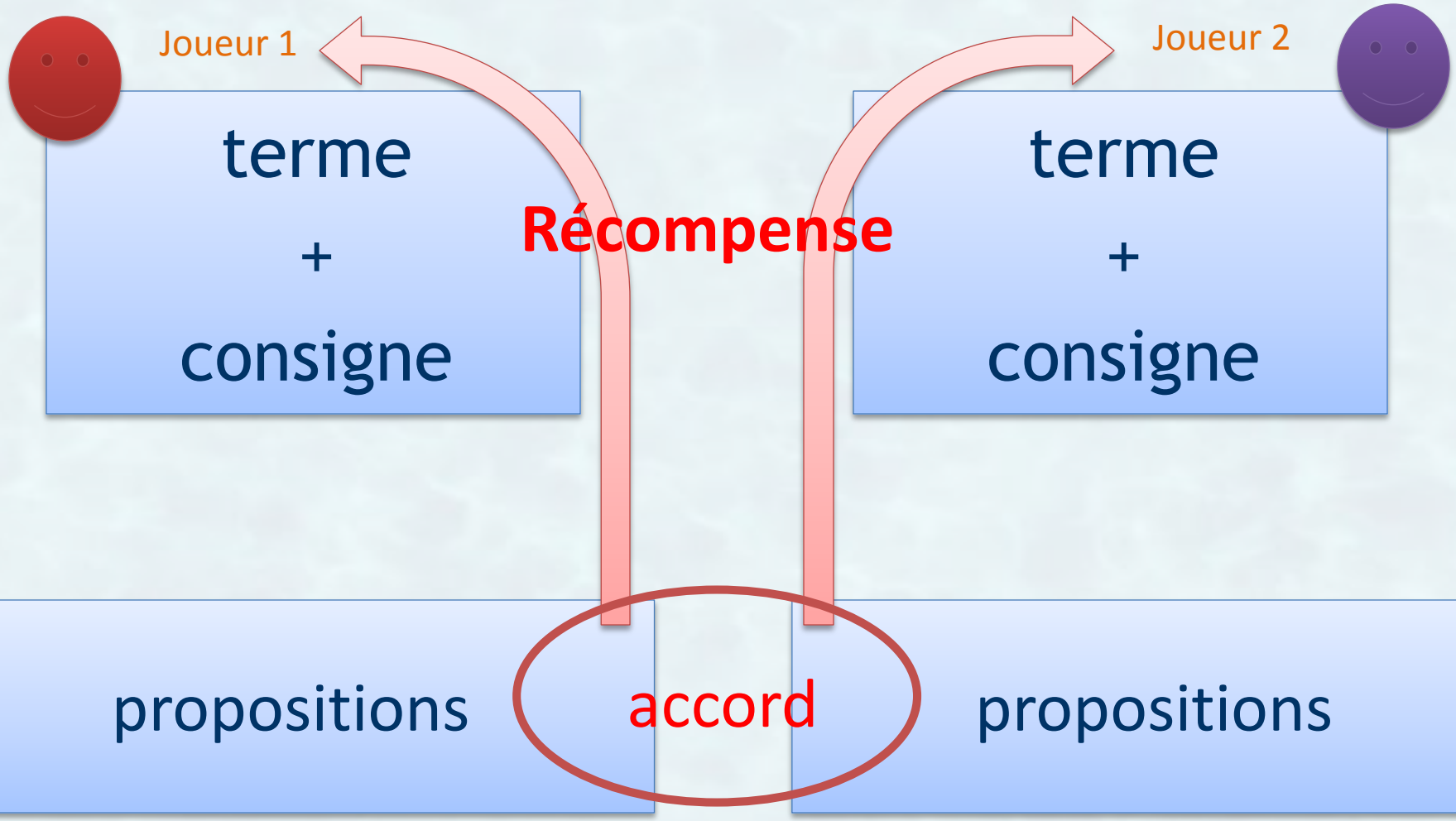


JeuxDeMots

exemple de partie : résultat

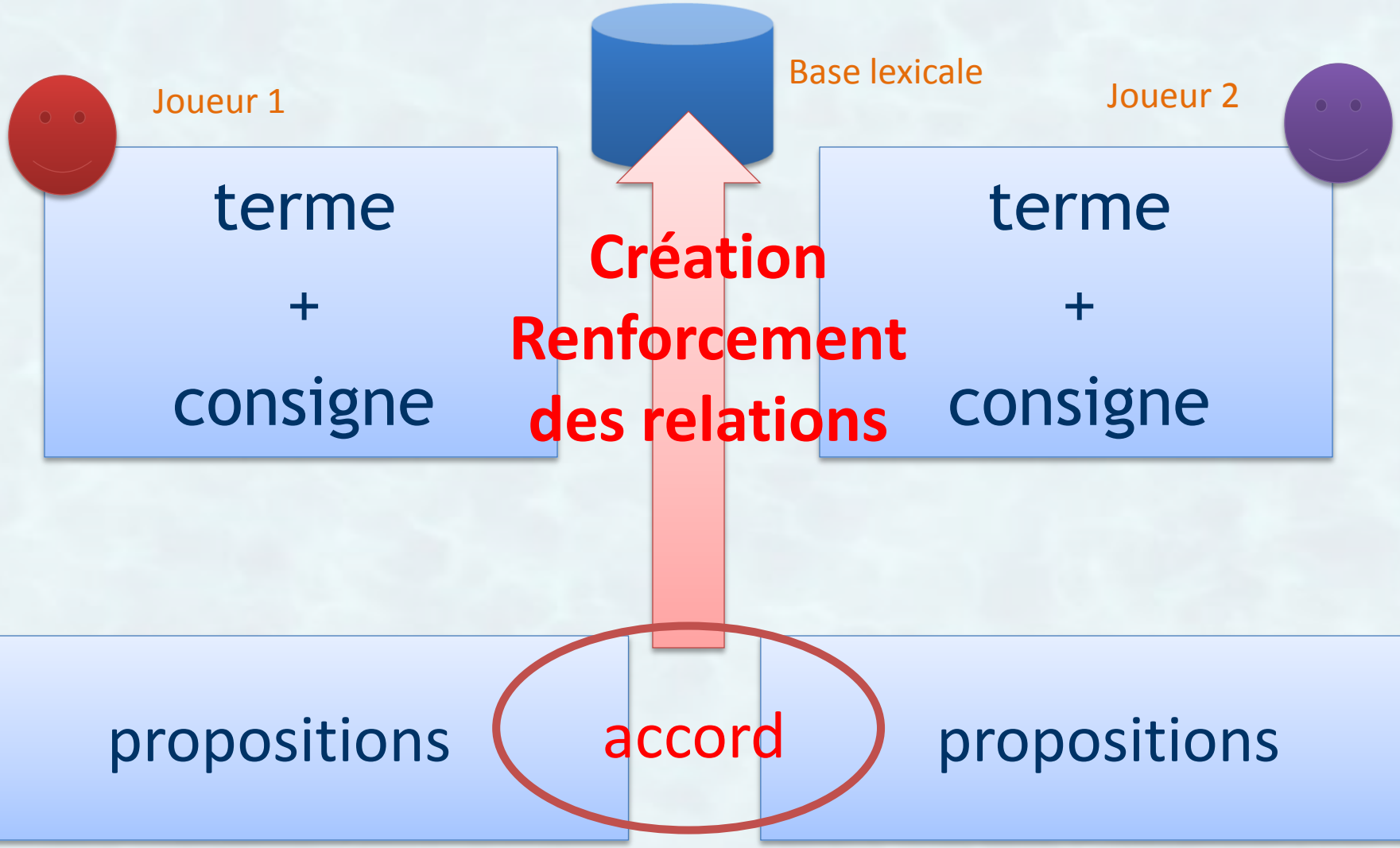
JeuxDeMots

modèle d'interaction



JeuxDeMots

modèle d'interaction



Réseau lexical

modèle d'interaction – filtrage

Accord par paires

- Consensus minimal itéré (pondération)
- Minimiser le bruit, maximiser le rappel (longue traîne)

Caractéristiques

- Mot tiré au hasard
- Partenaire inconnu durant la partie
- Partie asynchrone

Points

⇒ calibrés pour développer la longue traîne (augmenter le rappel)

Bcp si relation peu pondérée	peu si relation très pondérée	rien si relation taboue

JeuxDeMots

résultats quantitatifs

Depuis septembre 2007

Plus de 2500 inscrits

Environ 1 million de parties jouées - 16574h de temps de jeu cumulé (690 j)

1 250 000 relations entre 230 000 termes

(amorçage avec 0 relation et 150 000 termes)

632835 r_associated
83349 r_domain
152463 r_syn
41674 r_isa
12111 r_anto
10098 r_hypo
12158 r_has_part
9322 r_holo
9346 r_locution
11306 r_agent
…

coût = 1 mcf/année

Ressource accessible librement

sortie complète mensuelle (pseudo rdf)

sortie terme à terme (pseudo XML)

Des JeuxDeMots dans d'autres langues

English (anglais)

ภาษาไทย (thaï)

ភាសាខ្មែរ (khmer)

العربية (arabe)

日本語 (japonais)

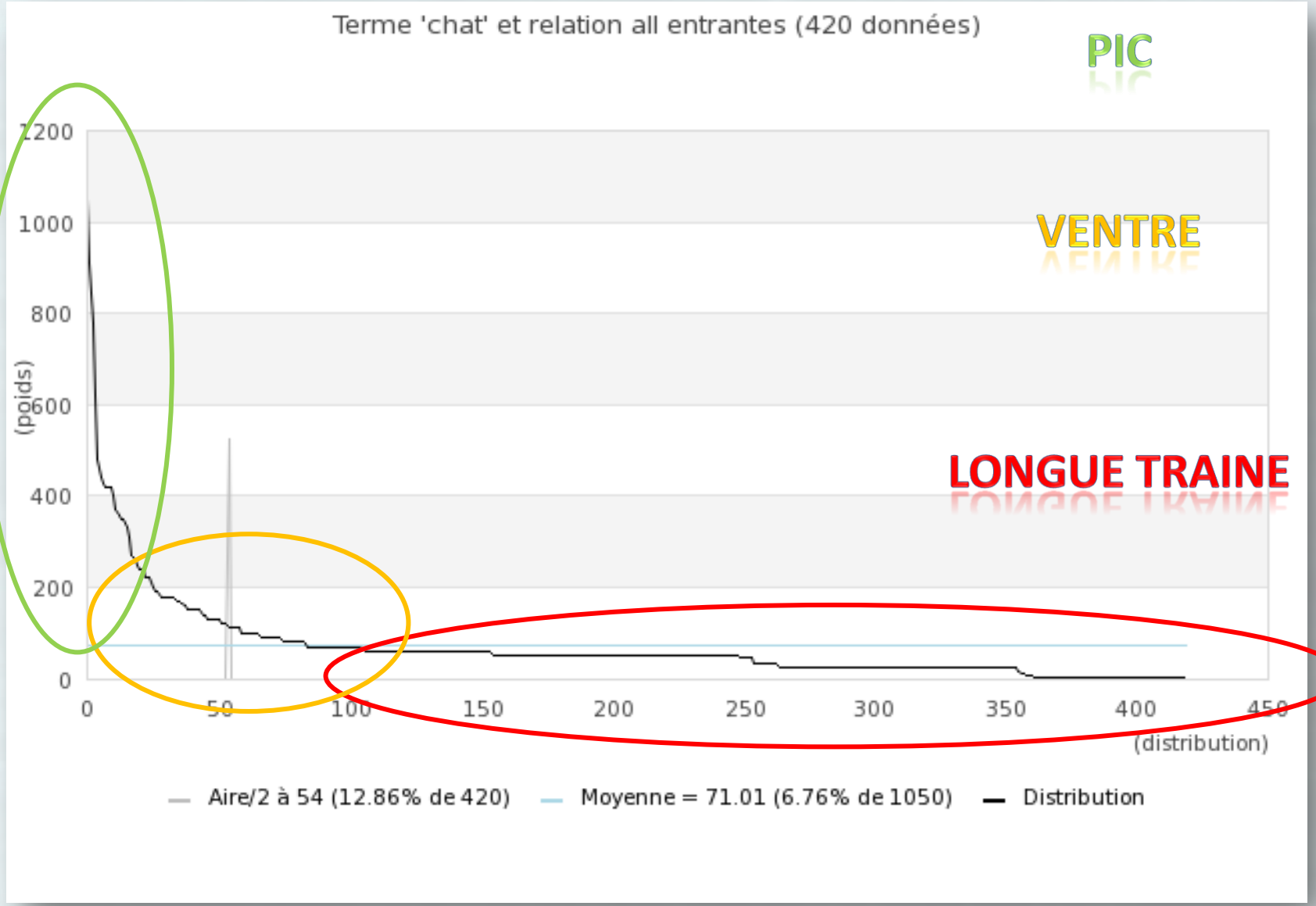
português (portugais)

español (espagnol)

shikomori (comorien)

Tiếng Việt (vietnamien)

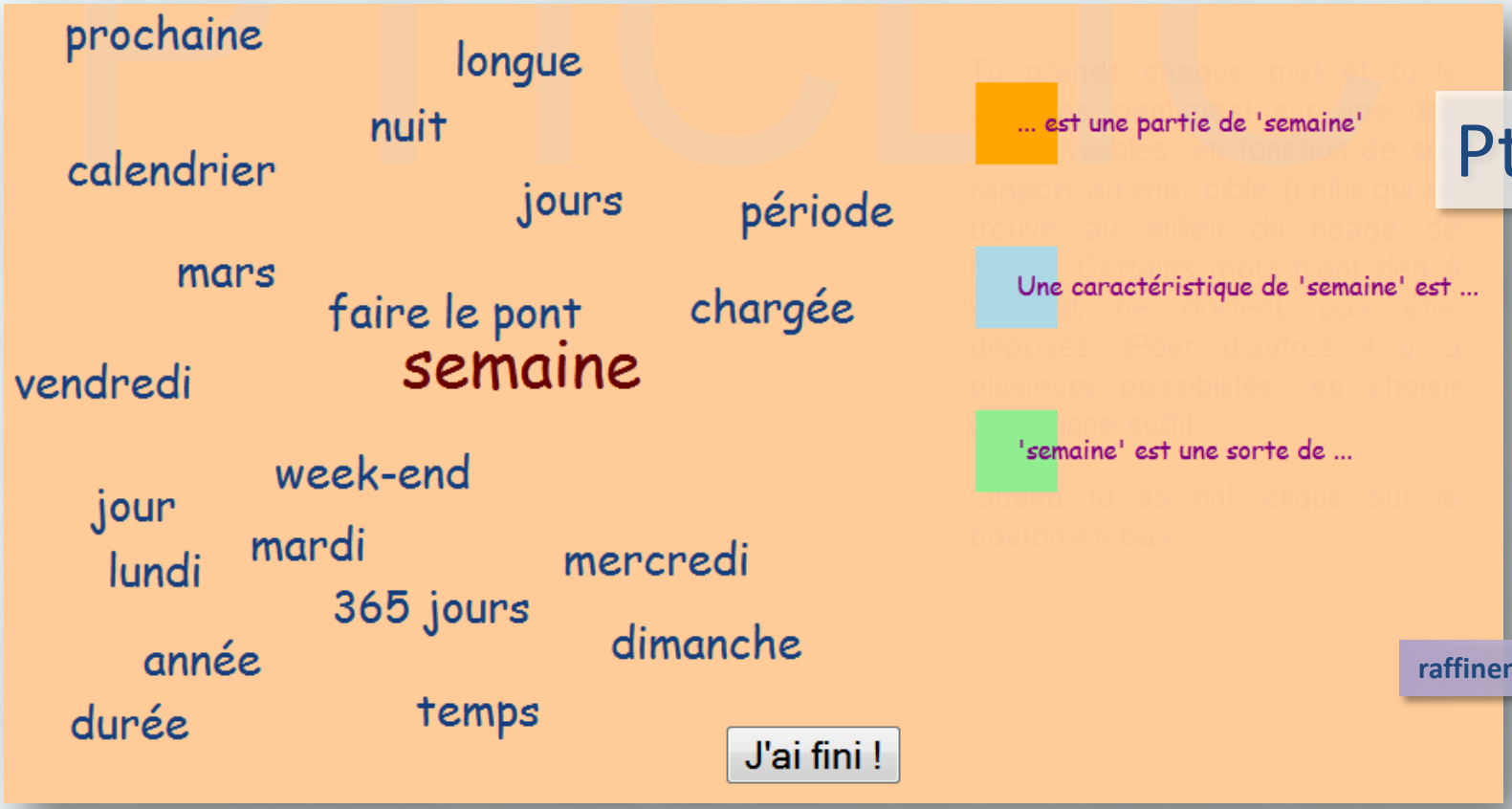
Réseau lexical – distribution pour *chat* - zipfien



Réseau lexical

autres jeux

Jeu de renforcement / raffinement des relations



PtiClic

V. Zampa

STICEF

raffinement des structures

Réseau lexical

autres jeux

Jeu de propagation sur les usages
et création de relations négatives / inhibition

Askit

activation et inhibition

Réseau lexical

identification d'usages

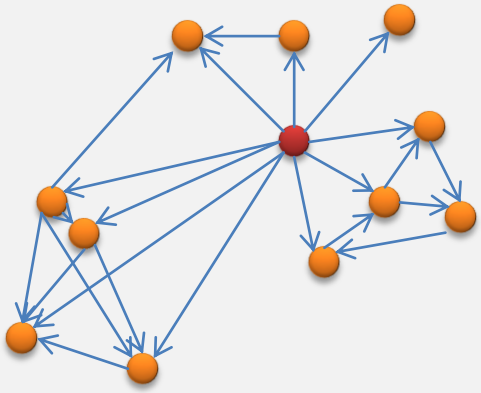
A. Joubert | TAL | JADT

Peut-on constuire un inventaire de sens / usages ?

S . Ploux & B Victorri - synonymes

Identification de cliques et quasi-cliques
Classification hiérarchique ascendante
Nommage/glosage descendant

- chat



Réseau lexical

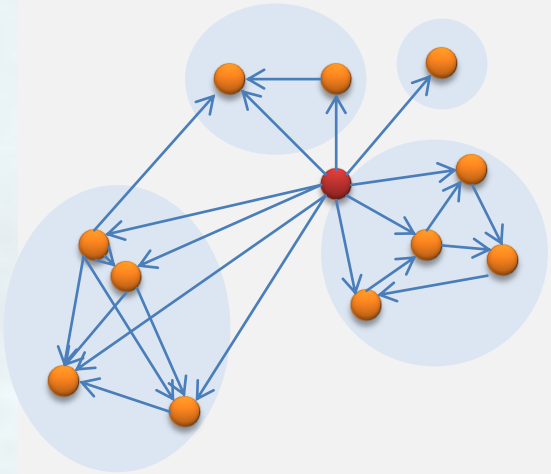
identification d'usages

A. Joubert | TAL | JADT

Peut-on constuire un inventaire de sens / usages ?

S . Ploux & B Victorri - synonymes

Identification de cliques et quasi-cliques
Classification hiérarchique ascendante
Nommage/glosage descendant



Réseau lexical

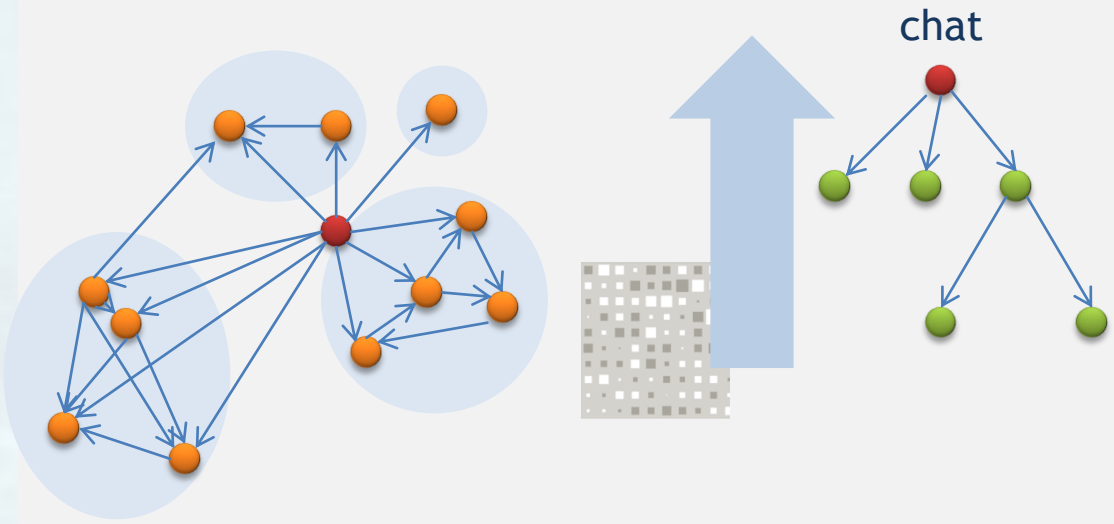
identification d'usages

A. Joubert | TAL | JADT

Peut-on constuire un inventaire de sens / usages ?

S . Ploux & B Victorri - synonymes

Identification de cliques et quasi-cliques
Classification hiérarchique ascendante
Nommage/glosage descendant



Réseau lexical

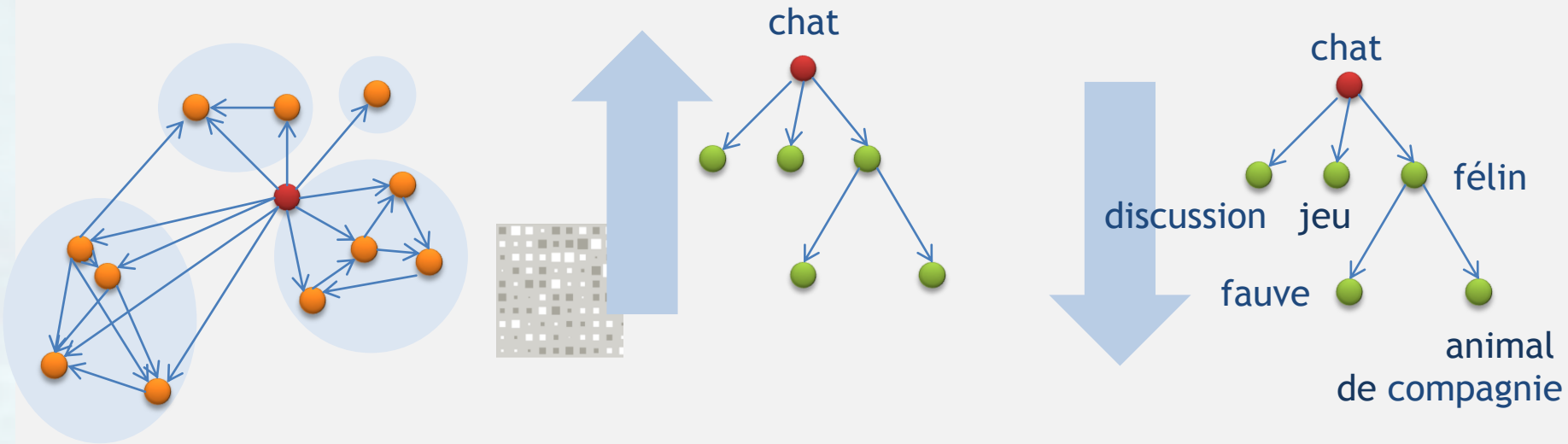
identification d'usages

A. Joubert | TAL | JADT

Peut-on constuire un inventaire de sens / usages ?

S . Ploux & B Victorri - synonymes

Identification de cliques et quasi-cliques
Classification hiérarchique ascendante
Nommage/glosage descendant



Injection dans le réseau lexical
raffinements utilisables / jouables par les joueurs

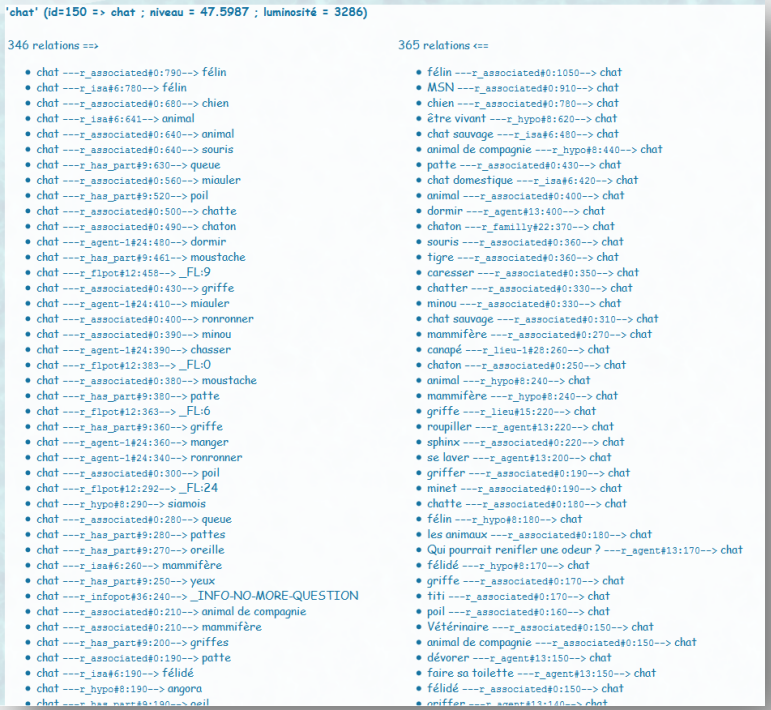
bouclage

Réseau lexical

évaluation qualitative

Comment évaluer *qualitativement* le réseau ?

Problème difficile
comment évaluer cela ?



Quel échantillonage ?

⇒ Approche indirecte via une tâche simple

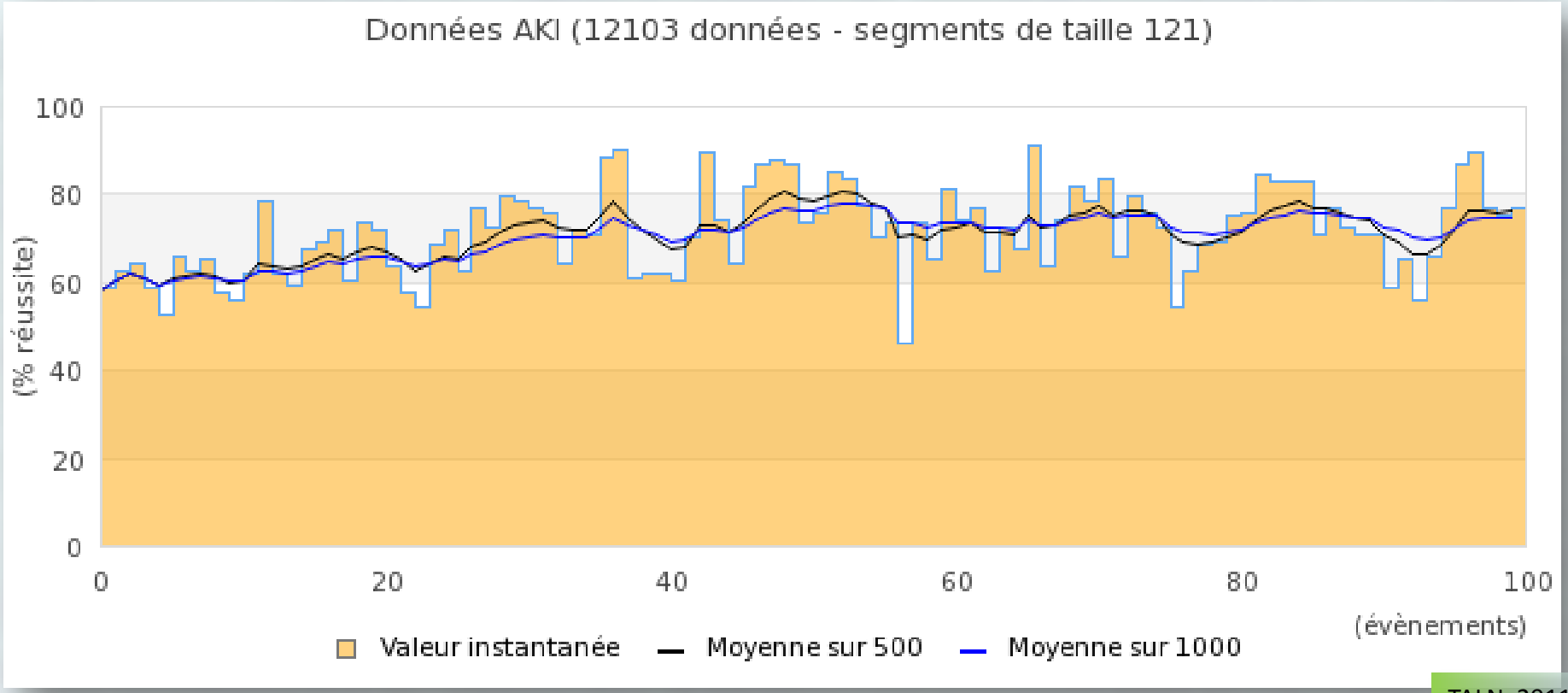
AKI – un jeu de devinette : faire trouver un terme à partir d'indices

Variantes avec plusieurs groupes de TER

Hypothèse : si le terme est trouvé alors il est bien indexé

Réseau lexical

AKI - Évaluation



TALN 2011

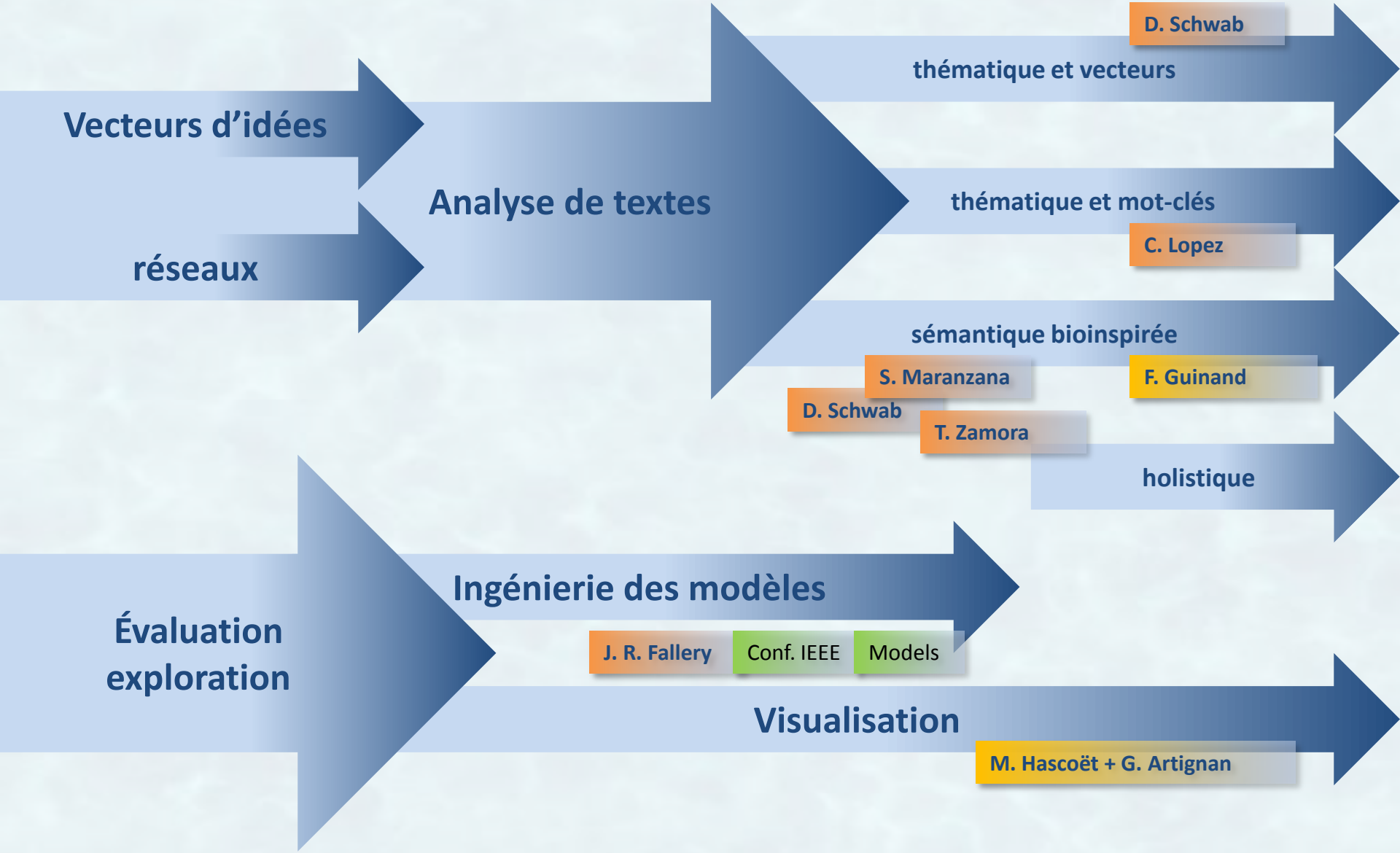
AKI ~ 75%

Test sur > 10000 parties dont les termes sont choisis par les utilisateurs

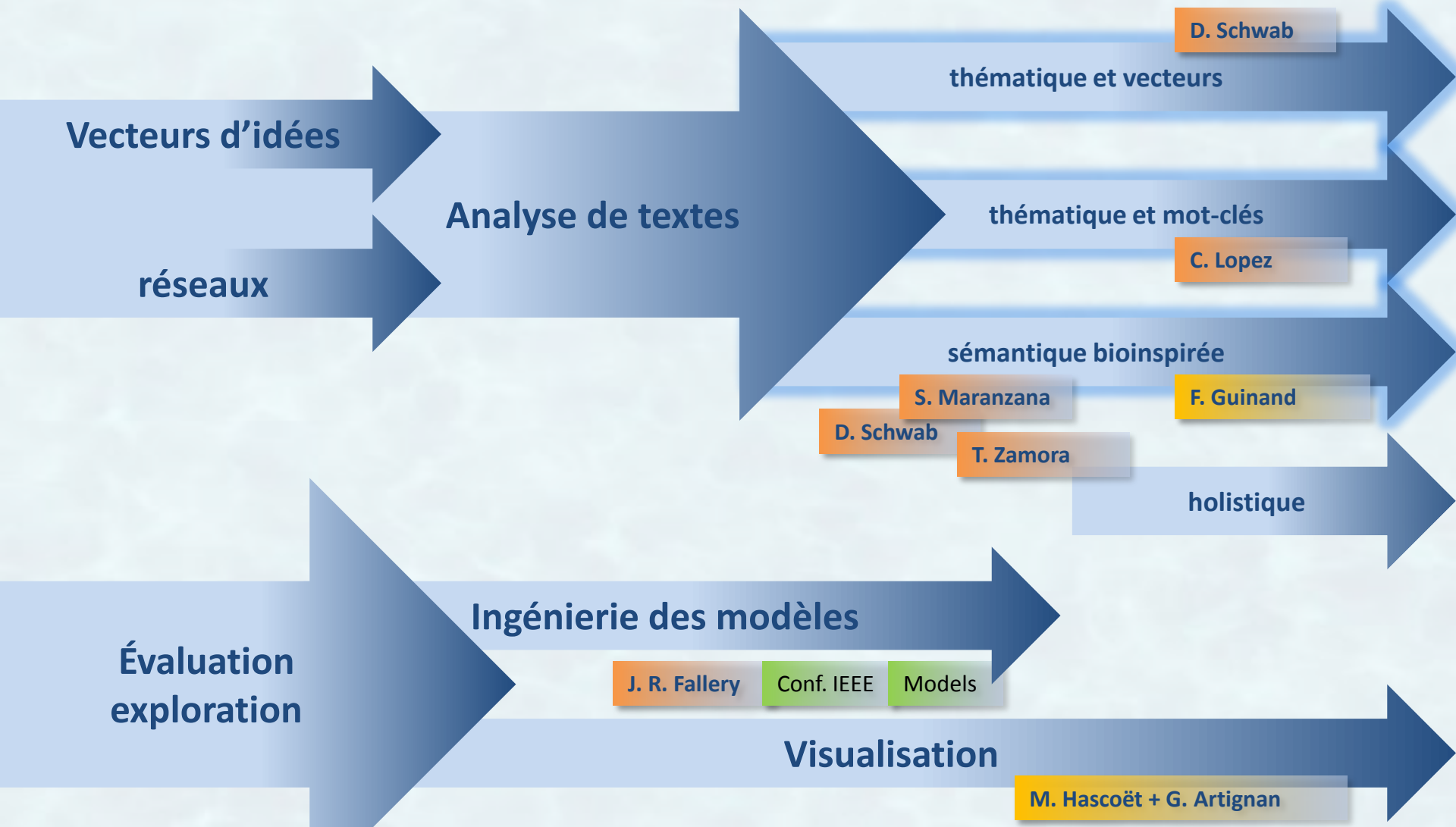
Utilisateur ~ 46%

Test sur environ 200 entrées et les termes tabous de AKI donnés comme indices

Chronologie



Chronologie



Analyse de texte

thématique et vectorielle - propagation

Construire des vecteurs
pour des textes et des définitions ?

Support : arbre d'analyse morpho-syntaxique

J. Chauché

⇒ Sygfran
- ⇒ syntagmes
- ⇒ gouverneurs/têtes
- ⇒ fonctions syntaxiques

} différence de pondération des vecteurs pour $\oplus$

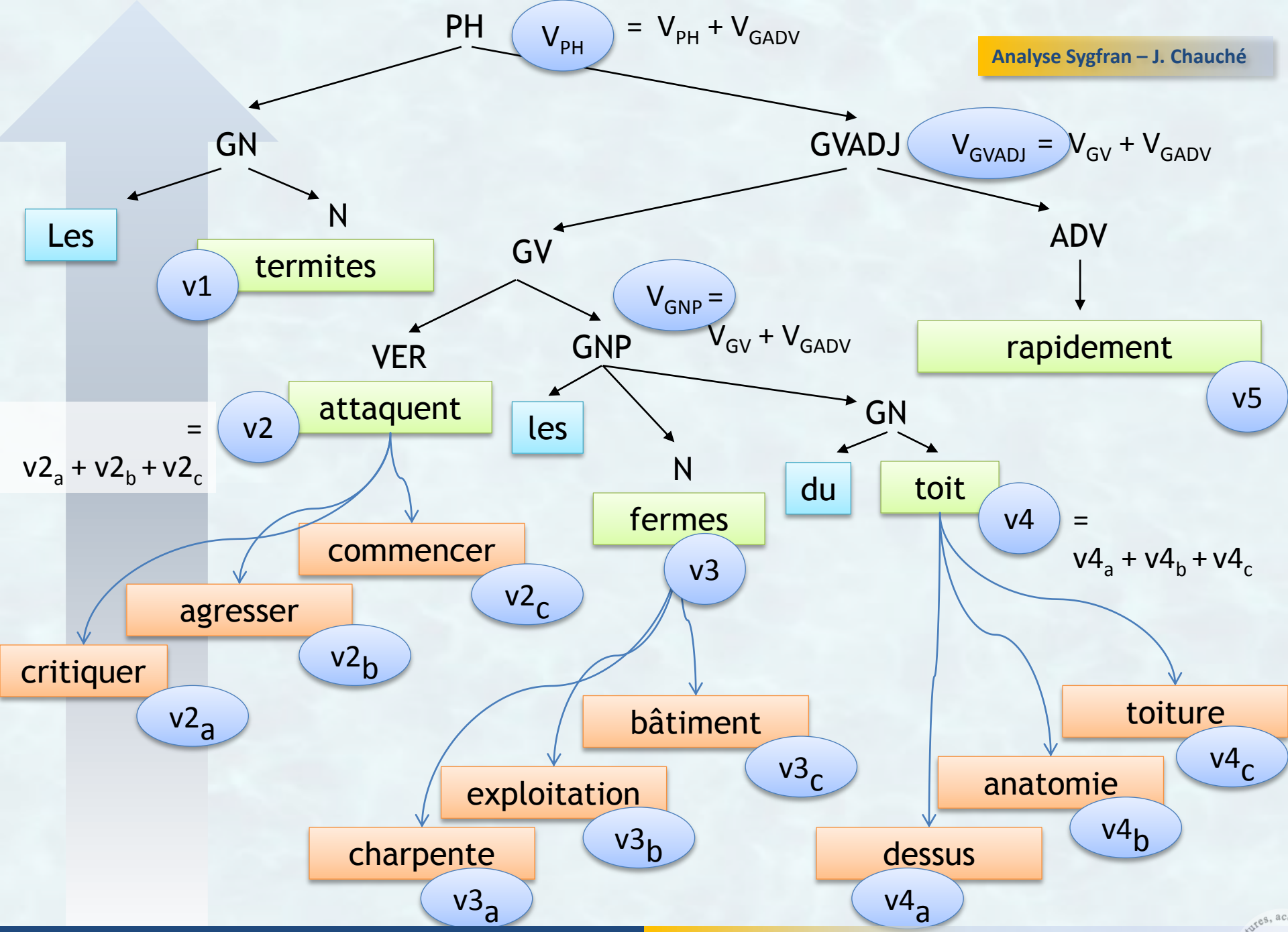
Principe : propagation de vecteurs

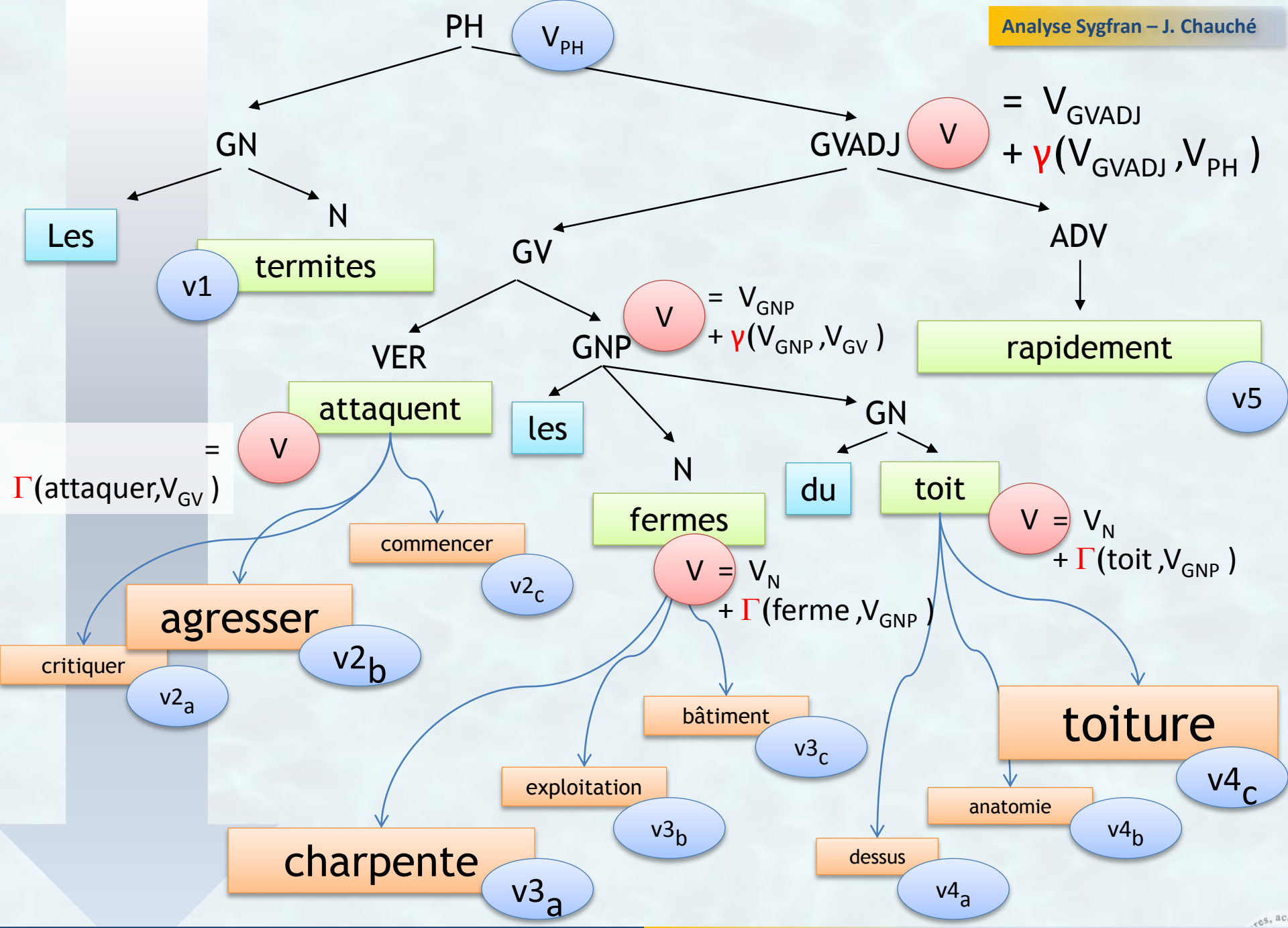
D. Schwab

TAL

itération (3-4 cycles de remontée/descente suffisent en général)
convergence globale (vecteur de la racine)

contextualisations faible γ et forte Γ


PH
V_{PH}
= V_{PH} + V_{GADV}
GN
GVADJ
V_{GVADJ} = V_{GV} + V_{GADV}
Les
N
v1
termites
GV
ADV
V_{GNP} = V_{GV} + V_{GADV}
VER
GNP
rapidement
v5
=
$v2_a + v2_b + v2_c$
v2
attaquent
les
N
GN
du
toit
v4
=
$v4_a + v4_b + v4_c$
commencer
fermes
v3
agresser
$v2_c$
critiquer
$v2_b$
$v2_a$
bâtiment
toiture
$v4_c$
$v3_c$
exploitation
anatomie
$v4_b$
$v3_b$
charpente
dessus
$v3_a$
$v4_a$


PH
V_{PH}
GN
GVADJ
V
$= V_{GVADJ} + \gamma(V_{GVADJ}, V_{PH})$
Les
N
v1
termites
GV
ADV
VER
GNP
V
$= V_{GNP} + \gamma(V_{GNP}, V_{GV})$
rapidement
v5
attaquent
V
$= \Gamma(\text{attaquer}, V_{GV})$
les
N
GN
du
toit
fermes
V = V_N + Γ(toit, V_{GNP})
commencer
$v2_c$
agresser
$v2_b$
critiquer
$v2_a$
V = V_N + Γ(ferme, V_{GNP})
bâtiment
$v3_c$
toiture
$v4_c$
exploitation
$v3_b$
anatomie
$v4_b$
charpente
$v3_a$
dessus
$v4_a$

Analyse de texte

thématique et lexicale - diffusion

C. Lopez

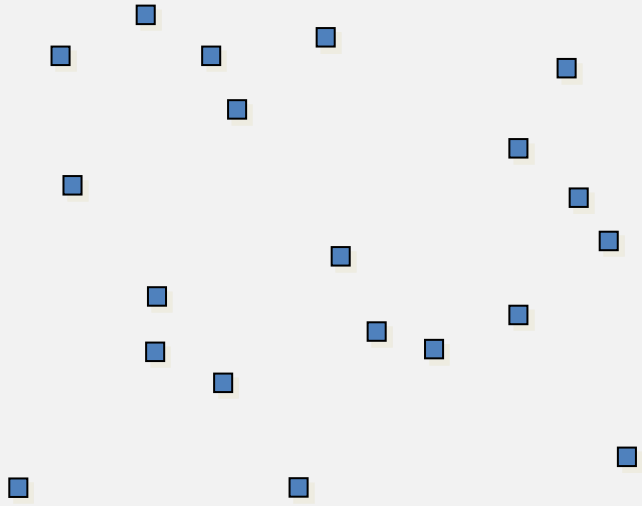
Extraire les termes *pertinents* d'un texte

Amorçage d'un noyau
- mot-clés centraux
- par TF-IDF

Itération
- mot-clés périphériques
- réduction de la distance seuil

Capture de mot-clés proches dans le réseau
- proche au sens de DA sur le vecteur lexical contruit

Exemple : *carambolage sur l'A7*



Analyse de texte

thématique et lexicale - diffusion

C. Lopez

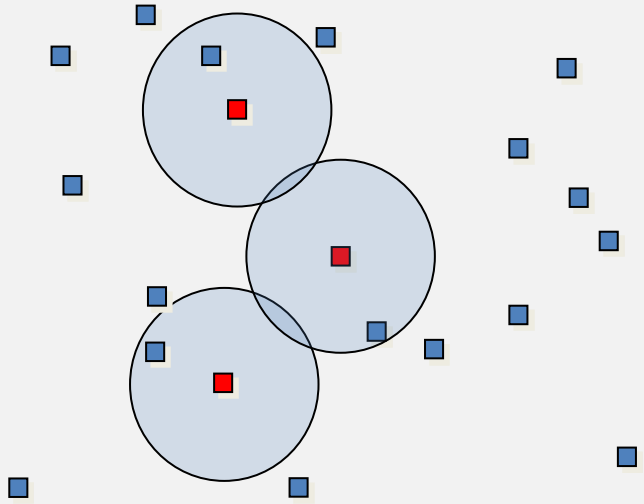
Extraire les termes *pertinents* d'un texte

Amorçage d'un noyau
- mot-clés centraux
- par TF-IDF

Itération
- mot-clés périphériques
- réduction de la distance seuil

Capture de mot-clés proches dans le réseau
- proche au sens de DA sur le vecteur lexical contruit

Exemple : *carambolage sur l'A7*



Analyse de texte

thématique et lexicale - diffusion

C. Lopez

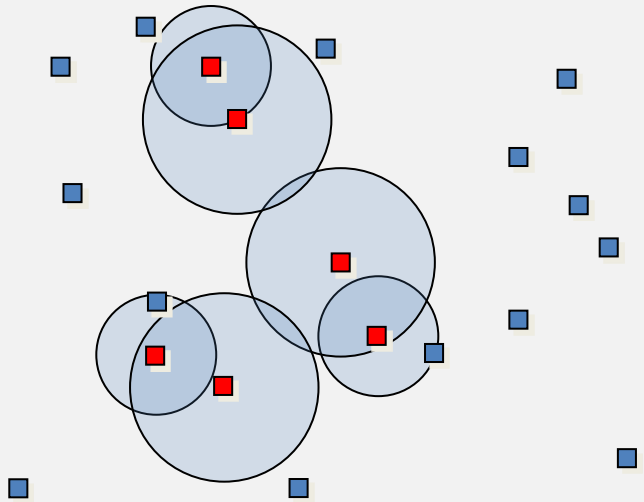
Extraire les termes *pertinents* d'un texte

Amorçage d'un noyau
- mot-clés centraux
- par TF-IDF

Itération
- mot-clés périphériques
- réduction de la distance seuil

Capture de mot-clés proches dans le réseau
- proche au sens de DA sur le vecteur lexical contruit

Exemple : *carambolage sur l'A7*



Analyse de texte

thématique et lexicale - diffusion

C. Lopez

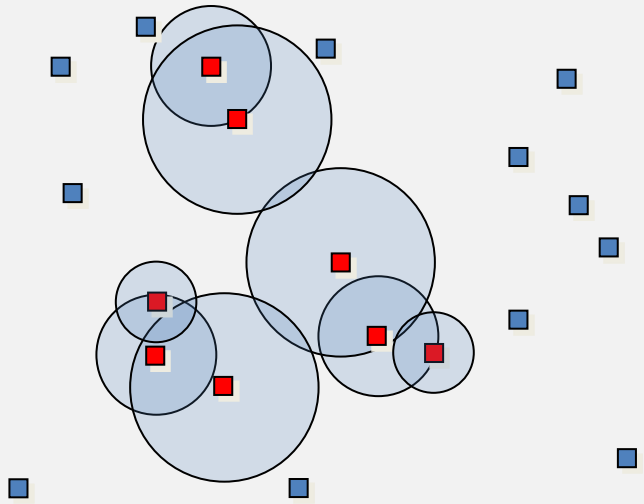
Extraire les termes *pertinents* d'un texte

Amorçage d'un noyau
- mot-clés centraux
- par TF-IDF

Itération
- mot-clés périphériques
- réduction de la distance seuil

Capture de mot-clés proches dans le réseau
- proche au sens de DA sur le vecteur lexical contruit

Exemple : *carambolage sur l'A7*

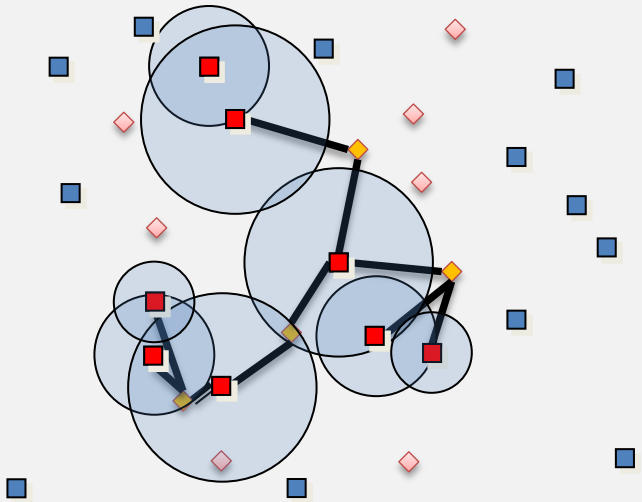


Analyse de texte

thématique et lexicale - diffusion

C. Lopez

Extraire les termes *pertinents* d'un texte



Amorçage d'un noyau
- mot-clés centraux
- par TF-IDF

Itération
- mot-clés périphériques
- réduction de la distance seuil

Capture de mot-clés proches dans le réseau
- proche au sens de DA sur le vecteur lexical contruit

Exemple : *carambolage sur l'A7*

carambolage, A7, automobile, transport par route, accident de la route, autoroute, voiture>automobile

Diffusion dans un espace

Propagation dans le réseau

Analyse de texte

sémantique, relationnelle et bioinspirée

Comment être moins localiste ?
Expliciter l'interprétation ?

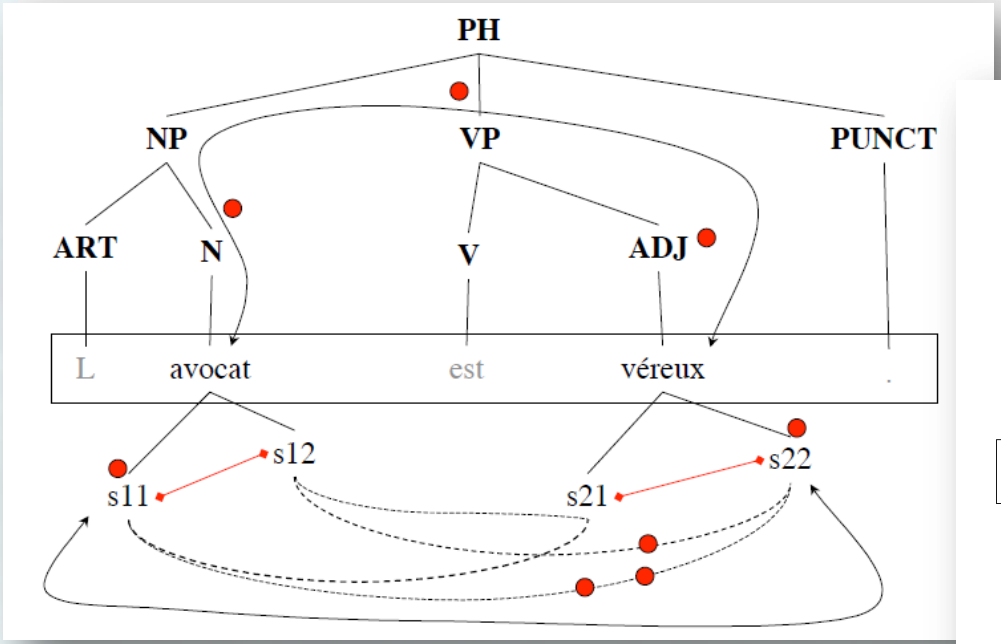
Multi-agents réactifs : transporteurs d'informations
Support : réseau (arbres, graphes UNL)
Déplacements pseudo-aléatoires
stygmergie – le contrôle par le marquage temporaire de l'environnement
Création de ponts : court-cicuits
typés pour rendre explicite la (les) relation(s) entre les segments
Utilisation (recopie locale) du réseau lexico-sémantique
Rattachement de groupes prépositionels

LREC + chap. livre **Ch. Boitet**

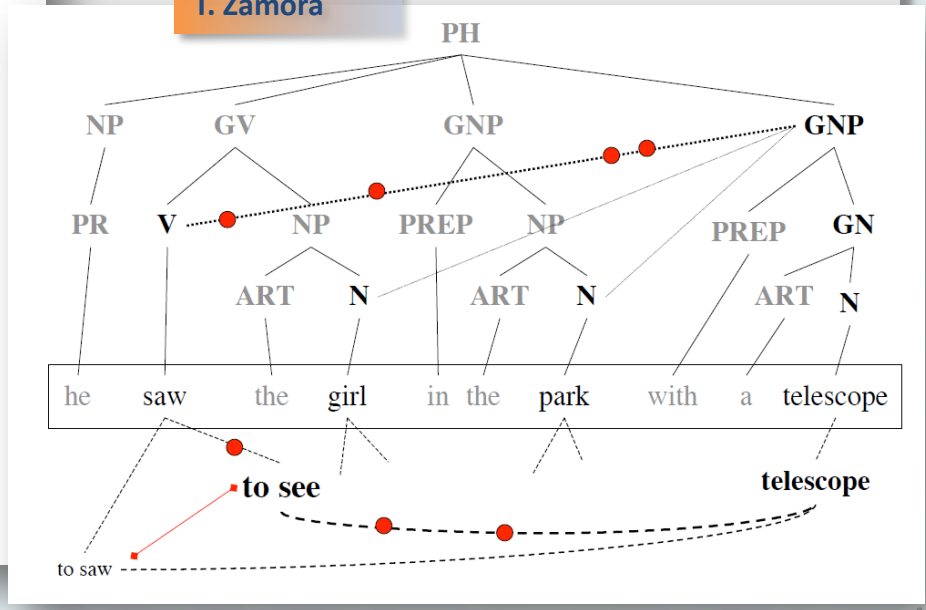
2 chap. de livre **F. Guinand**

ECTI **N. Gala**

S. Maranzana



T. Zamora



Conclusions - apports

Analyse de textes

Thématique et sémantique

abordée de façon généraliste

Construction de ressources lexico-sémantiques

Vecteurs et relations : complémentaires

rappel (concepts) et précision (vocables)

Acquisition par consensus populaire – vers du *sens commun*

les rendre disponibles et contribuables

Valorisation :

Lingua & Machina
Namae Concepts
Succeed Together
Cemagref
…

Calcul

Propagation (et diffusion) - Bouclage

Activation (et inhibition)

mise en concurrence d'objets lexicaux : sélection d'usages en contexte

Application et évaluation

Sortir du TAL (par exemple, ingénierie des modèles)

Exploration des ressources

Oracle lexical (AKI) : un outil ludique

M. Huchard

B. Gaume

M. Hascoët

M. Zock

Conclusions - perspectives

Vers des mécanismes d'inférence pour les réseaux sémantiques

Endogène & Exogène (le contenu des textes comme amorçage)

Renforcer / créer certaines relations

Découvrir des règles d'inférences

Renforcer le couplage avec l'analyse

Objets lexicaux agrégés & meta information

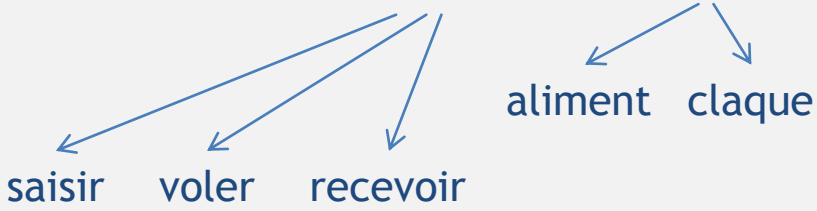
(embarcation ⇒**carac**⇒ surchargée) ⇒**conséquence**⇒ chavirer

Réifier les occurrences de relations

r ⇒**info**⇒ raciste, sexiste, humoristique, etc.

r ⇒**info**⇒ tjrs vrai, probable, possible, déduite

Le commis / le vagabond / l'enfant a pris une tarte.



Vers du raisonnement et de l'explication

Restaurer au TAL son désir d'intelligence

Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier

LIRMM

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Merci