Graphes linguistiques multiniveau pour l'extraction de connaissances : l'exemple des collocations

Vincent Archer
(GETALP – Laboratoire d'informatique de Grenoble)
Dir. : Gilles Sérasset et Christian Boitet

24 septembre 2009
Thèse en informatique - Université Joseph Fourier (Grenoble 1)

Contexte

- Traitement automatique des langues naturelles
  - Tâches linguistiques
    - Traduction automatique
    - Analyse, génération de langage naturel, reconnaissance de parole, etc.
- Résultats de qualité insuffisante
  - Ambiguïtés
  - Phénomènes linguistiques particuliers
    - Ex. : « pluie battante » (collocation) ne doit pas être traduit mot-à-mot
  - Manque de ressources linguistiques

Comment obtenir des ressources ?

- Produire des ressources linguistiques
- Rapidité de production
  - Processus automatiques
  - Qualité moindre
- Précision des ressources
  - Nécessite le travail de spécialistes (linguistes)
  - Prend beaucoup de temps
- = 2 volontés opposées

Outils pour les linguistes

- Combiner automatique et humain
  - Outils automatiques pour aider le linguiste
  - Permettre au linguiste de guider le processus
- Outils existants non adaptés
  - Trop de connaissances en programmation requises
  - Trop spécialisés
- Besoin d'outils génériques

Plan

- Étude du problème de l'extraction
  - Premières expérimentations
  - Cahier des charges
  - Modèle de graphes linguistiques existants
- Solution : MuLLinG
  - Modèle de graphe linguistique multiniveau
  - Opérations
  - Implémentation
- Utilisation : expérimentations
  - Extraction de collocations/bicollocations
  - Pondération de traductions lexicales

Collecte de ressources

- Compléter les ressources linguistiques
  - Produire des candidats (à valider)
- Approche basique : concordanciers
  - Contexte d'emploi des mots
  - Utilisé en lexicographie
  - L'informatique peut être utilisée davantage

the affects of acid	**rain**	. But I think this
YOUR WINDOW: TROPICAL	**RAIN**	FORESTS, EMERALD
Sunday night's heavy	**rain**	, was The Moody Blues.
near Lyndhurst. Heavy	**rain**	in early April
CAPTIONS [/c] RECENT	**RAIN**	produced perfect
starts and the real	**rain**	begins. Here we are
Dirty Mind and Purple	**Rain**	, and three from
happening to tropical	**rain**	forests because it
hampered by heavy	**rain**	and low cloud cover
by torrential	**rain**	during the last few

Collecte de ressources : extraction

Sous une **pluie battante**, les deux formations mettaient vingt bonnes minutes à entamer les hostilités.

(...)

La faute à une **pluie battante** qui est venue gâcher un match pourtant commencé sur un excellent tempo.

(...)

En raison de la **pluie battante** qui s'est abattue hier sur Flushing Meadows, aucun match n'a pu être joué.

Collecte de ressources : extraction

Sous une **pluie battante**, les deux formations mettaient vingt bonnes minutes à entamer
(...)
La faute à t
venue gâc
commenc
(...)
En raison de la **pluie battante** qui s'est abattue hier sur Flushing Meadows, aucun match n'a pu être joué.

« pluie battante »
Nombre d'occurrences : 3
WMI = 0,6

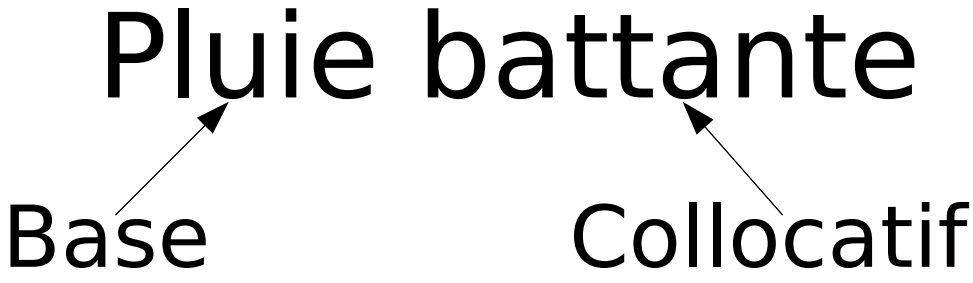
Collecte de ressources : extraction (2)

- Statistique : classer les candidats
  - Mesure doit refléter les propriétés du phénomène
- Linguistique : patrons syntaxiques
  - Reconnaissance des candidats
  - Ex. : *Verbe* suivi d'un *adverbe*
- Autres approches (hors thèse)
  - Apprentissage automatique
  - Contribution

Étude du problème

- Travaux d'*informaticien* linguiste
- But : fournir des outils favorisant l'extraction
  - Pas la description d'une meilleure approche d'extraction
- Étude
  - État de la pratique
  - Mes propres expérimentations

Collocations



~~Beating~~ rain

Driving

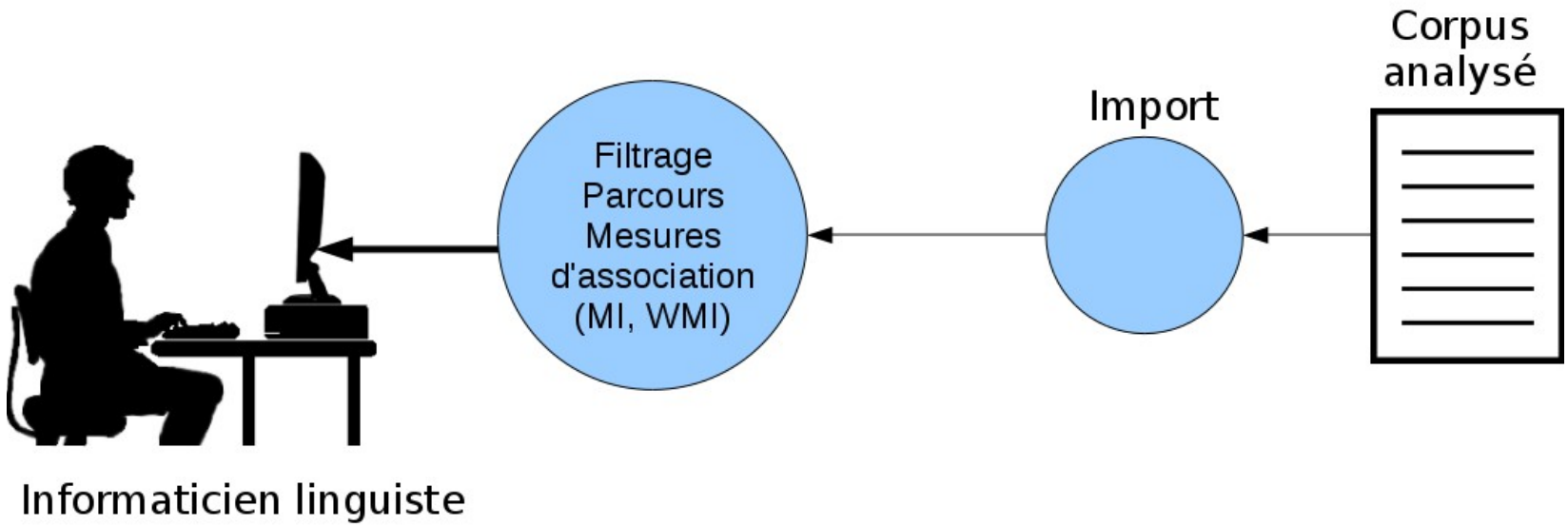
- Collocatif choisi arbitrairement
  - Exprime un sens donné en fonction de la base
  - Expression *semi-figée*
- Problème pour la traduction

Extraction de collocations

- Verbe modifié par adverbe
  - Ex. « *Ignorer superbement* »
  - Intensification : filtrage sémantique
- Mesure d'association
  - Plus souvent que par hasard

$$MI(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1)\,P(w_2)}$$

$$WMI(w_1, r, w_2) = P(w_1, r, w_2) \times \log \frac{P(w_1, r, w_2)}{P_g(w_1|r)\,P_d(w_2|r)\,P(r$$



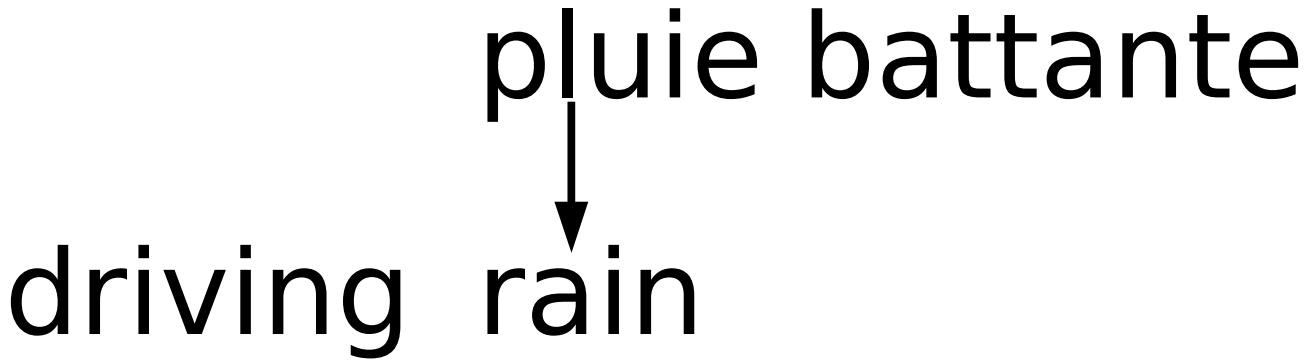
Extraction de collocations : résultats

	1	2	3	4
Contexte	Dépendance	Dépendance	Linéaire	Linéaire
Filtrage des verbes	Aucun	Verbes d'action	Verbes d'action	Verbes d'action
Filtrage des adverbes	Aucun	Vecteurs conceptuels	Vecteurs conceptuels	Vecteurs conceptuels +raffinement
Précision	**17%**	**41%**	**44%**	**83%**

10 premiers candidats (expérimentation 4)
régner sans partage
changer radicalement
réduire considérablement
devoir beaucoup
ignorer superbement
assumer pleinement
parler beaucoup
aimer beaucoup
exercer pleinement
montrer particulièrement

- Corpus LeMonde95
  - 25 millions de mots
- Besoin de beaucoup de filtrage
  - Élimine les collocations les moins décodables
    - « défendre bec et ongles »

Extraction bilingue : bicollocations



- Bicollocations : couple de collocations traductions l'une de l'autre
  - Les 2 bases sont traductions l'une de l'autre
  - Les collocatifs expriment (pour leur base) exactement le même sens

Extraction bilingue : expérimentations

- Mesures d'association basées sur :
  - Collocabilités monolingues
  - Apparition des collocations dans des documents associés

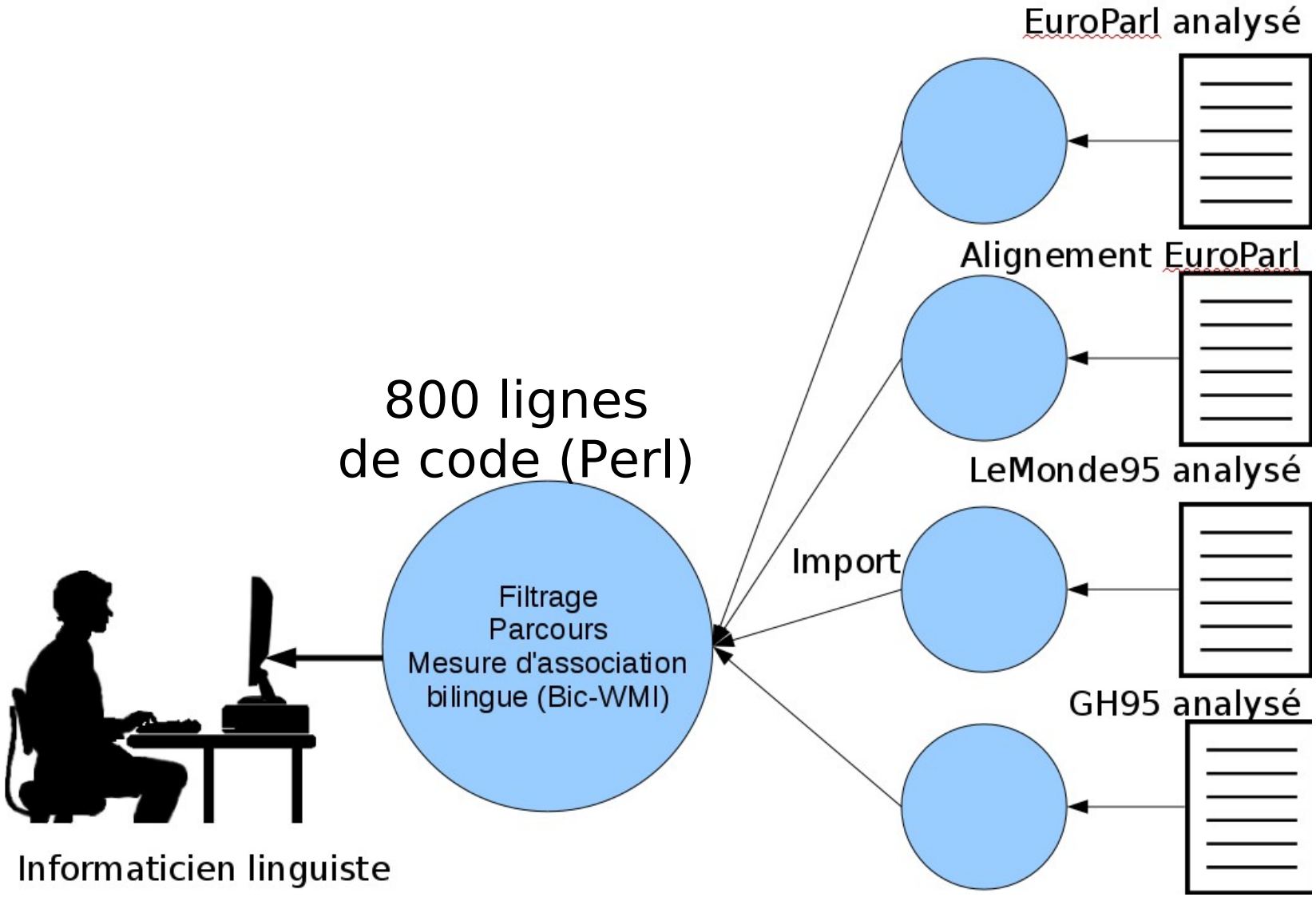
$$\cos_{bilingue}(c_1, c_2) = \frac{|AE(c_1, c_2)|}{\sqrt{|C_1| \times |C_2|}}$$

$$Bic_{WMI}(\langle b_1, c_1 \rangle, \langle b_2, c_2 \rangle) = [WMI(\langle b_1, c_1 \rangle) + WMI(\langle b_2, c_2 \rangle)] \times \cos_{bil}(\langle b_1, c_1 \rangle, \langle b_2, c_2 \rangle)$$

	Langue		Documents	Phrases	Mots
LeMonde95	FR	Comparables	47 646	1 016 876	24 730 579
GH95	EN		56 472	1 321 323	28 122 780
EuroParl-fr v2	FR	Parallèles	495	1 089 670	31 115 677
EuroParl-en v2	EN		491	1 064 462	25 089 232

  - LeMonde95, GH95 : analyse de dépendances
  - EuroParl : étiquetage grammatical

Extraction de bicollocations



- Faire la glu entre les différentes entrées

Extraction de bicollocations : résultats

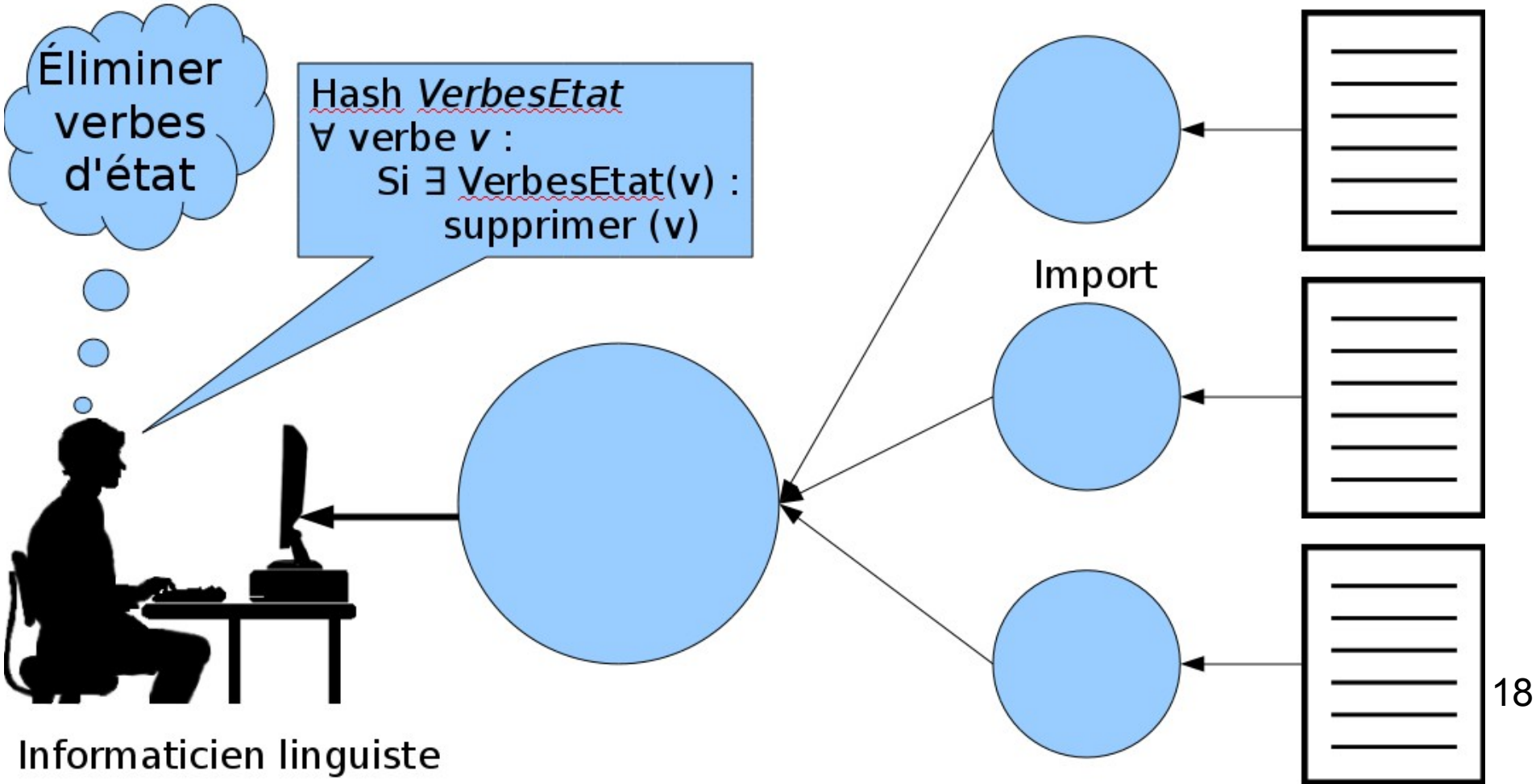
Expérimentations	Nombre de candidats positifs selon le filtrage			Précision
	Sans	Filtrage de base	Raffinement manuel	
comparable	80 298	3 973	201	**10%**
parallèle	15 583	1 995	43	**36,5%**

- Beaucoup de filtrage
  - Monolingue
- Silence

10 premiers candidats (expé. parallèle)	
soutenir pleinement	support wholly
jouer pleinement	work together
changer radicalement	***change radically***
modifier radicalement	***modify radically***
jouer pleinement	play right
travailler intensivement	*work hard*
contribuer grandement	***contribute significantly***
changer radicalement	***change dramatically***
contribuer grandement	***contribute considerably***
accepter partiellement	*accept fully*

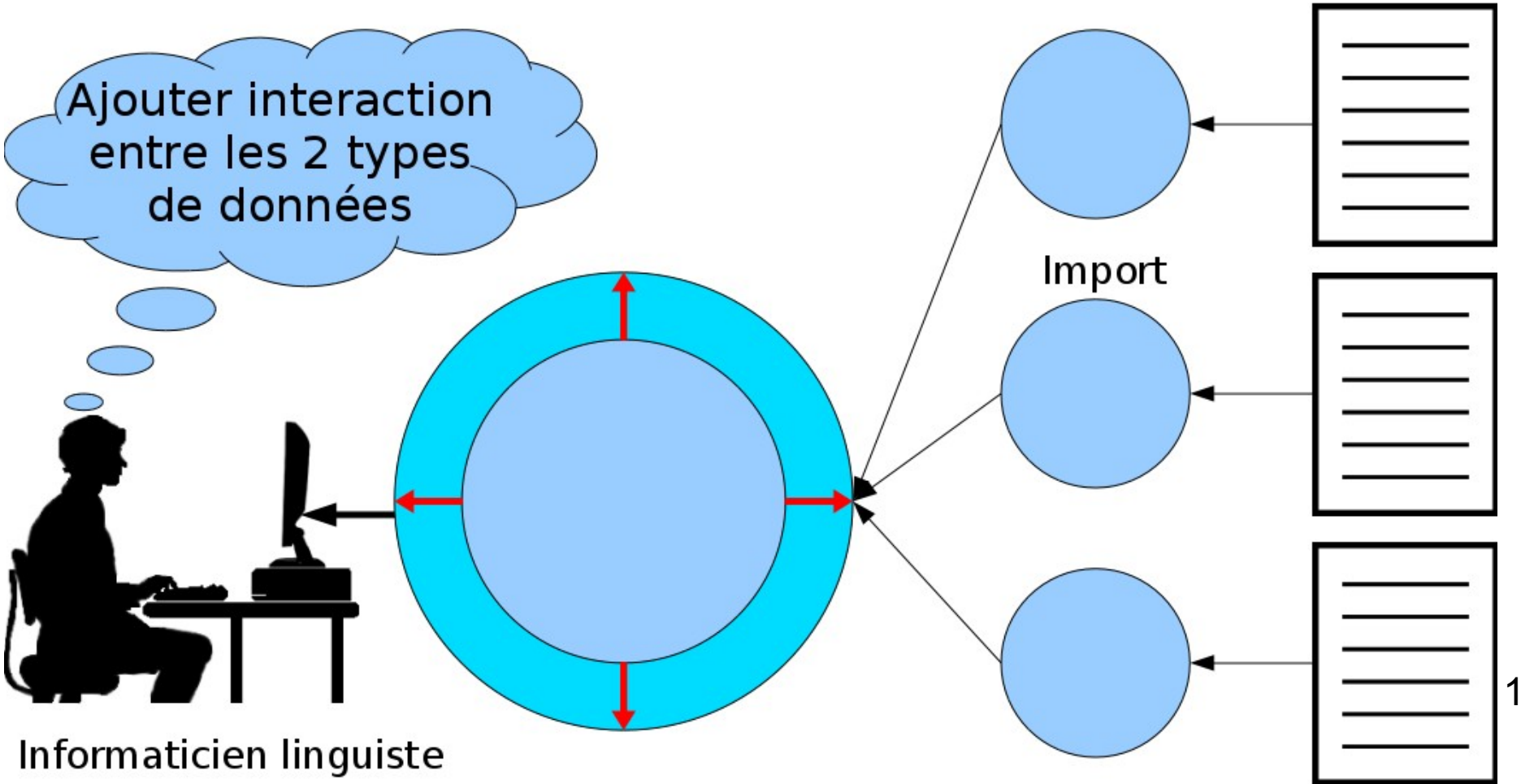
Programmation lourde et difficile

- Nécessite une double compétence



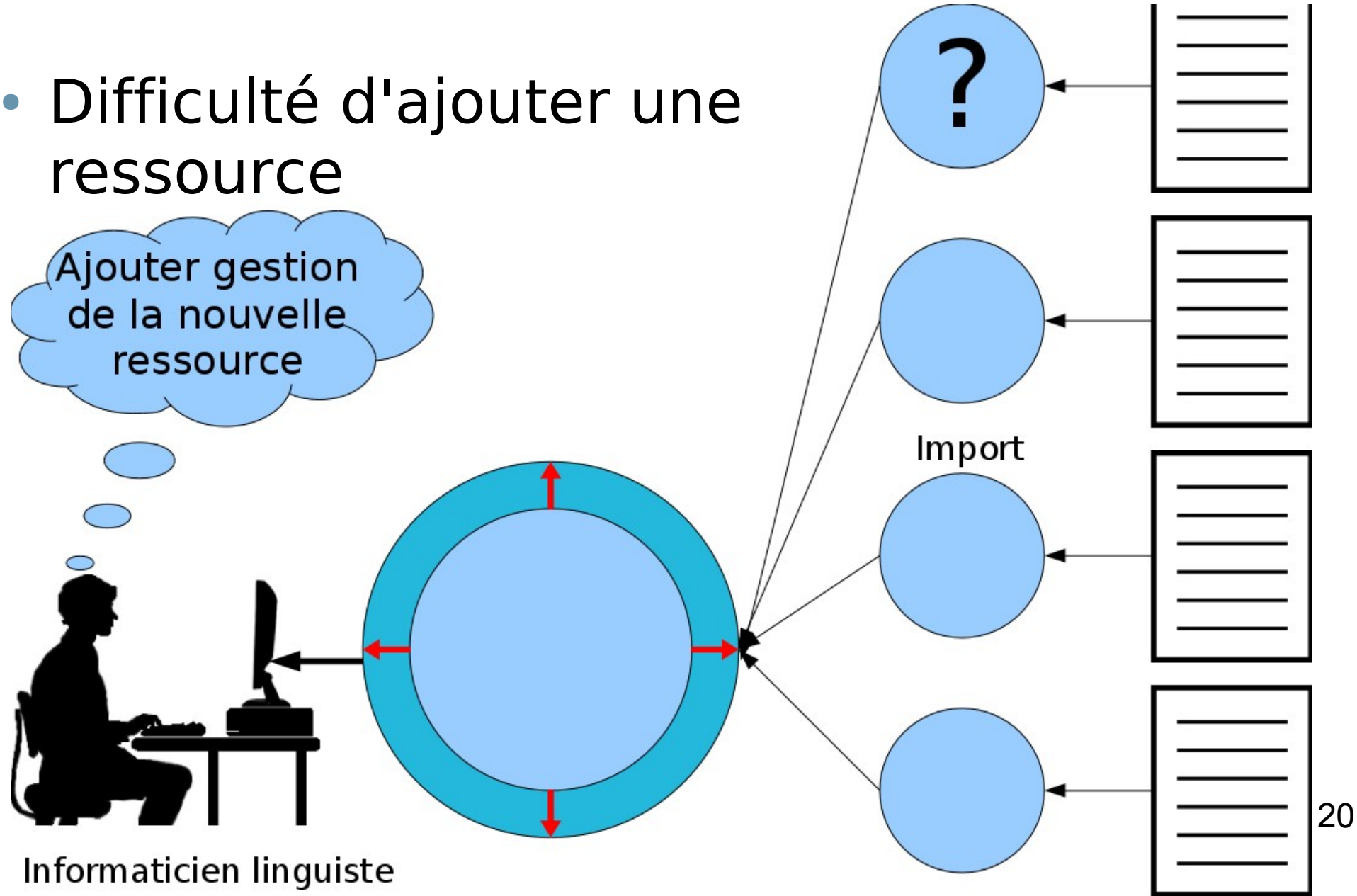
Programmation lourde et difficile

- Différents modèles de données
  - Corpus analysé, alignement, dictionnaire, etc.



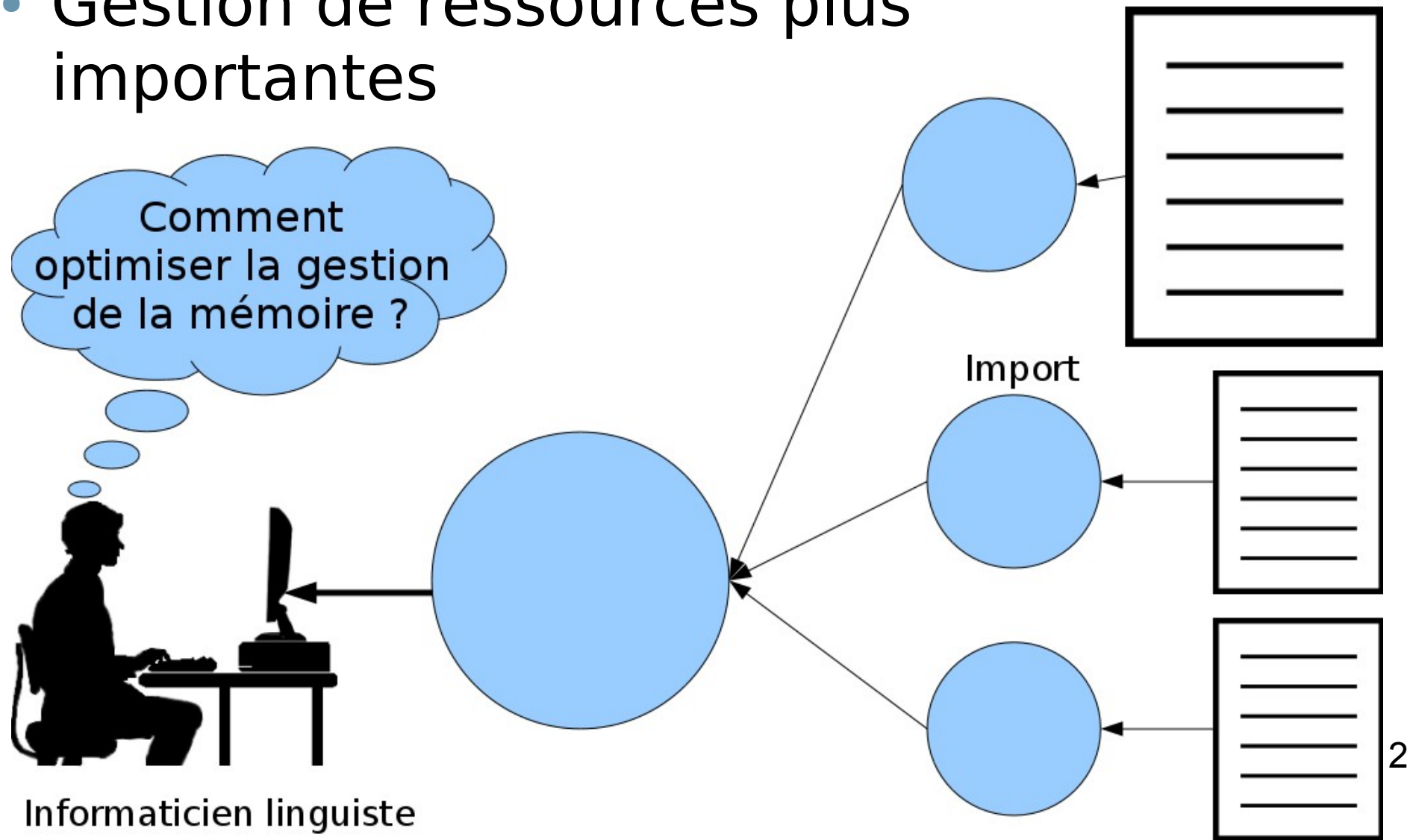
Programmation lourde et difficile

- Difficulté d'ajouter une ressource

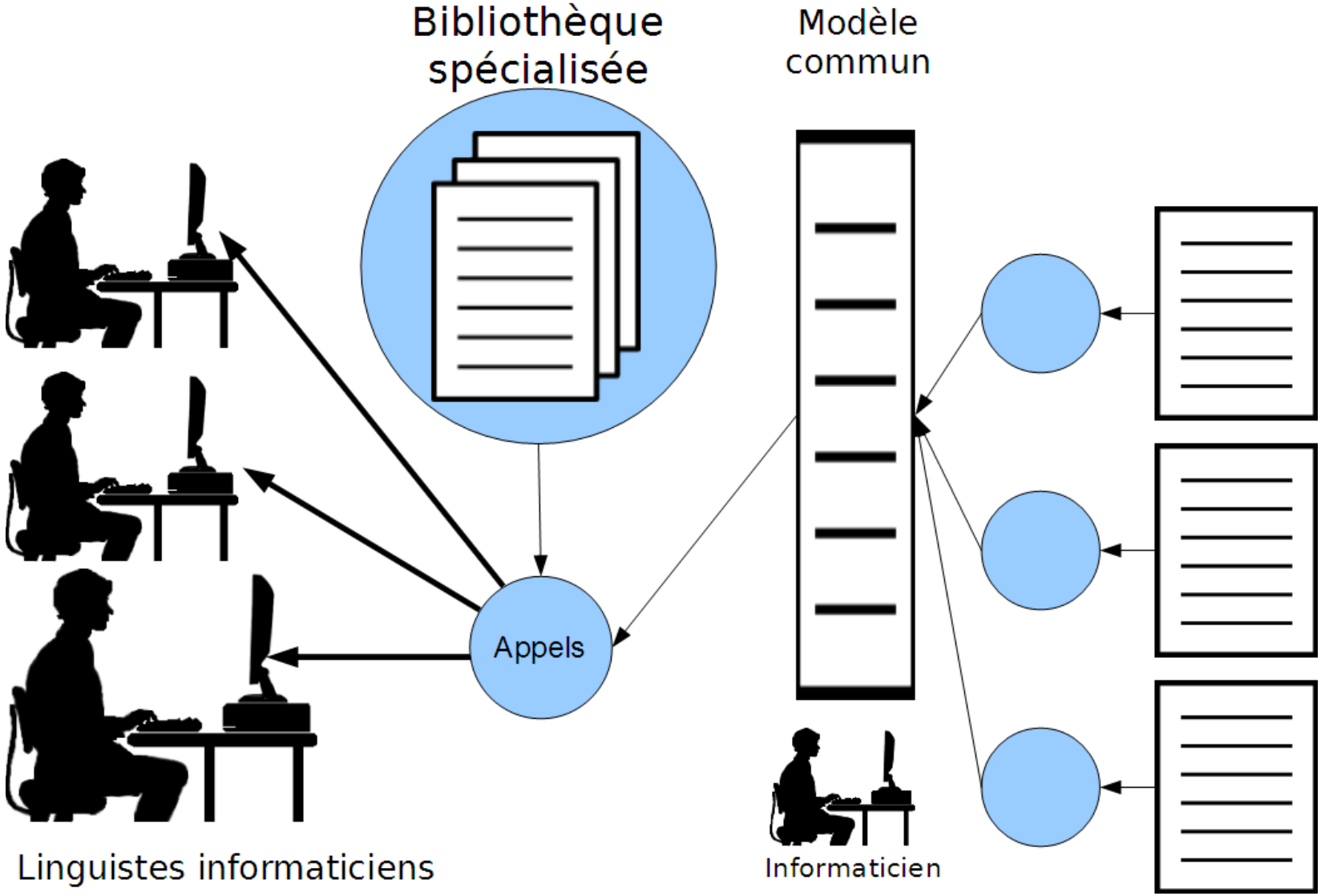


Programmation lourde et difficile

- Gestion de ressources plus importantes

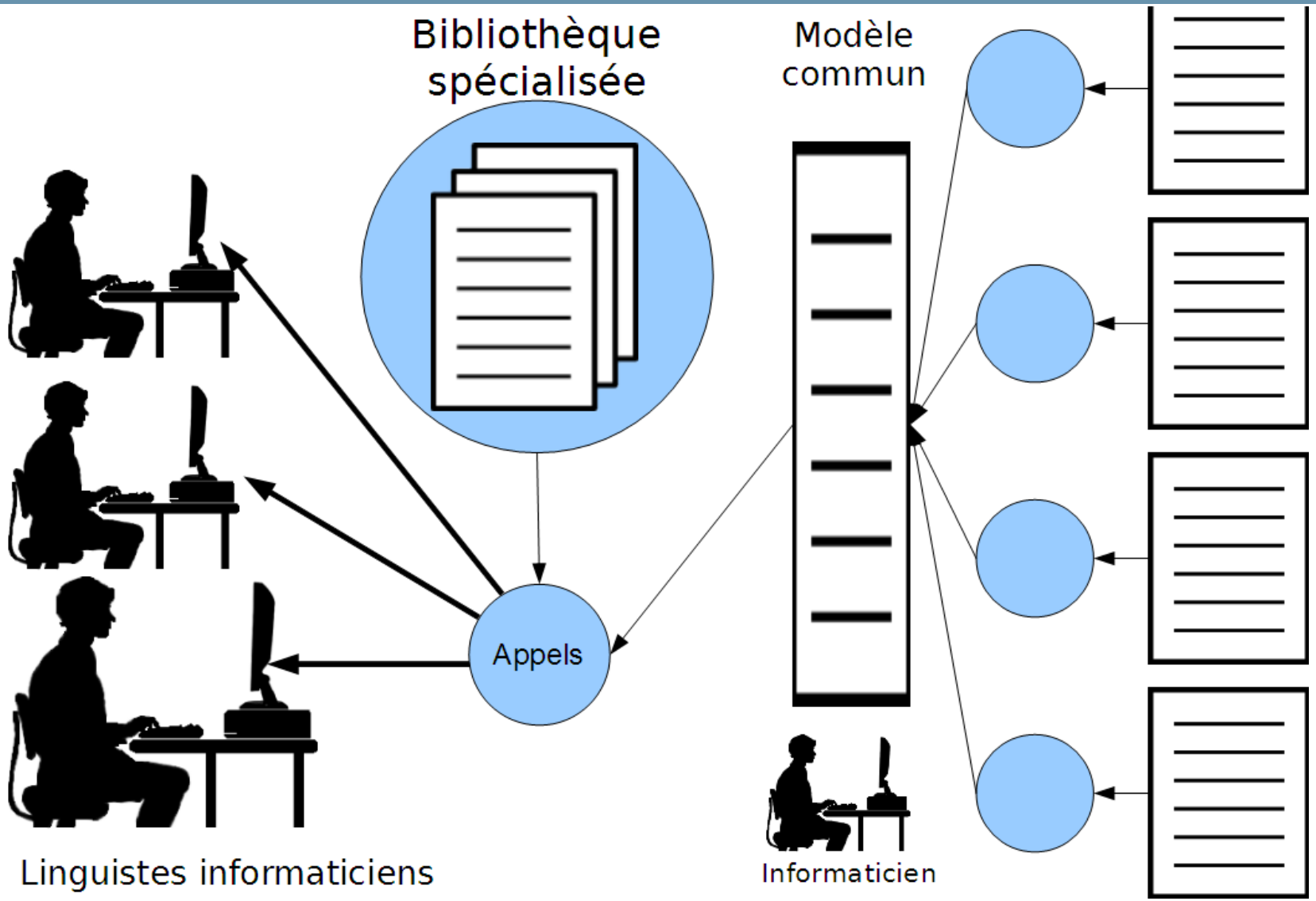


Outil générique



- Moins de compétences en programmation

Outil générique



- Ajout d'une nouvelle ressource facilité

Besoin : un outil générique d'extraction

- Langage spécialisé ?
  - Intéressant
  - Prématuré
- Bibliothèque spécialisée
  - Donne les outils
  - À privilégier dans un premier temps
- Interface graphique : pas utile pour le moment
- ⇒ Bibliothèque spécialisée

Cahier des charges : Données

- Simplicité
  - De la représentation
- Expressivité
  - Représenter des données plus complexes
- Généricité
  - Corpus monolingues, bilingues (analysés ou non)
  - Dictionnaires
  - etc.

Cahier des charges : Processus

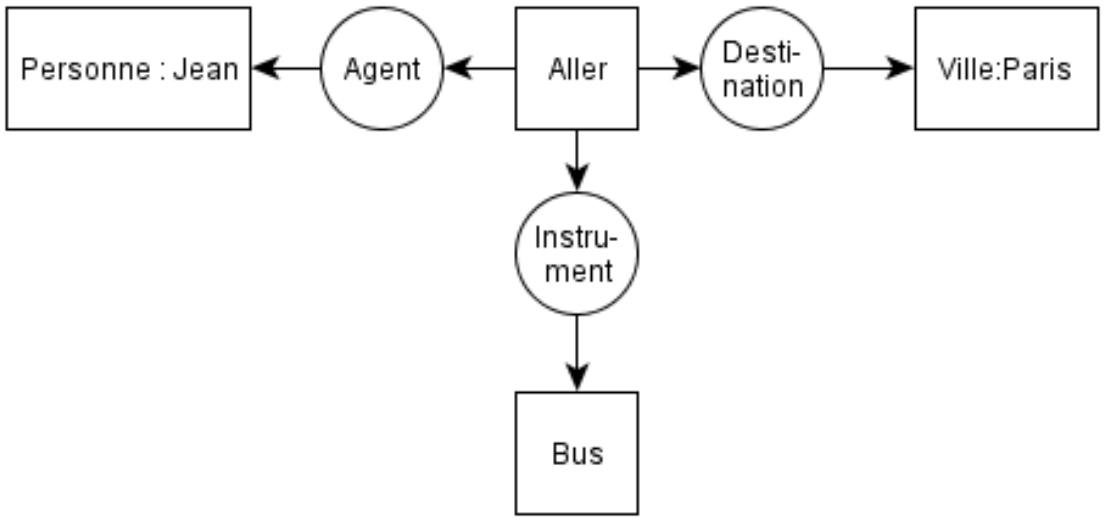
- Opérations simples...
  - Haut niveau
  - Combiner des opérations simples, plutôt qu'une opération complexe
- ...et génériques
  - Par rapport à la tâche
  - Par rapport au type de connaissances
  - Paramétrables

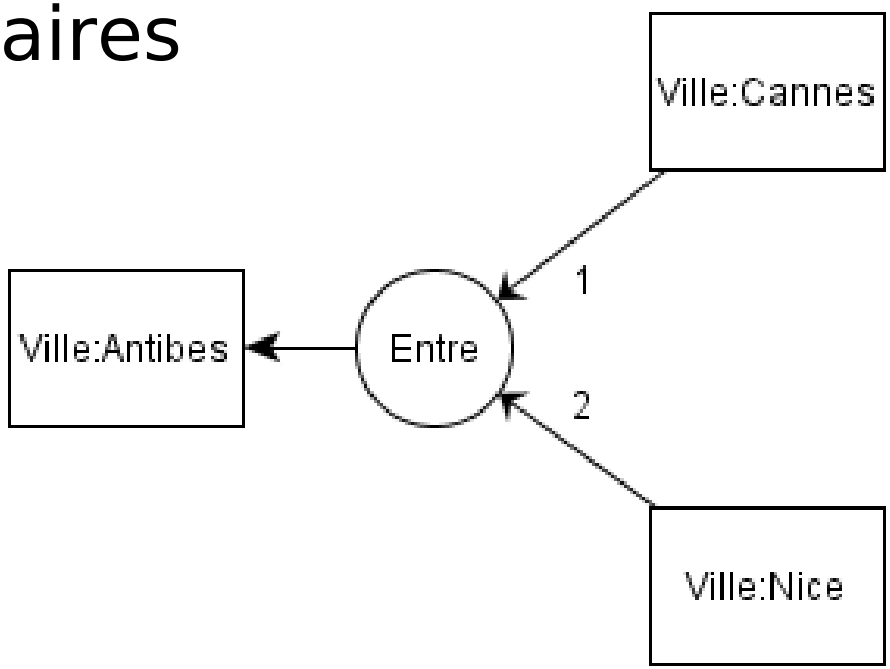
Choix des graphes

- Représentation familière pour les linguistes
- Compréhensibles rapidement par un humain
- Facilement utilisables par un processus automatique
- Assez souples pour représenter divers types de données
  - Corpus : relations entre occurrences de mots
  - Dictionnaires : relations entre mots

Inspiration : Graphes conceptuels

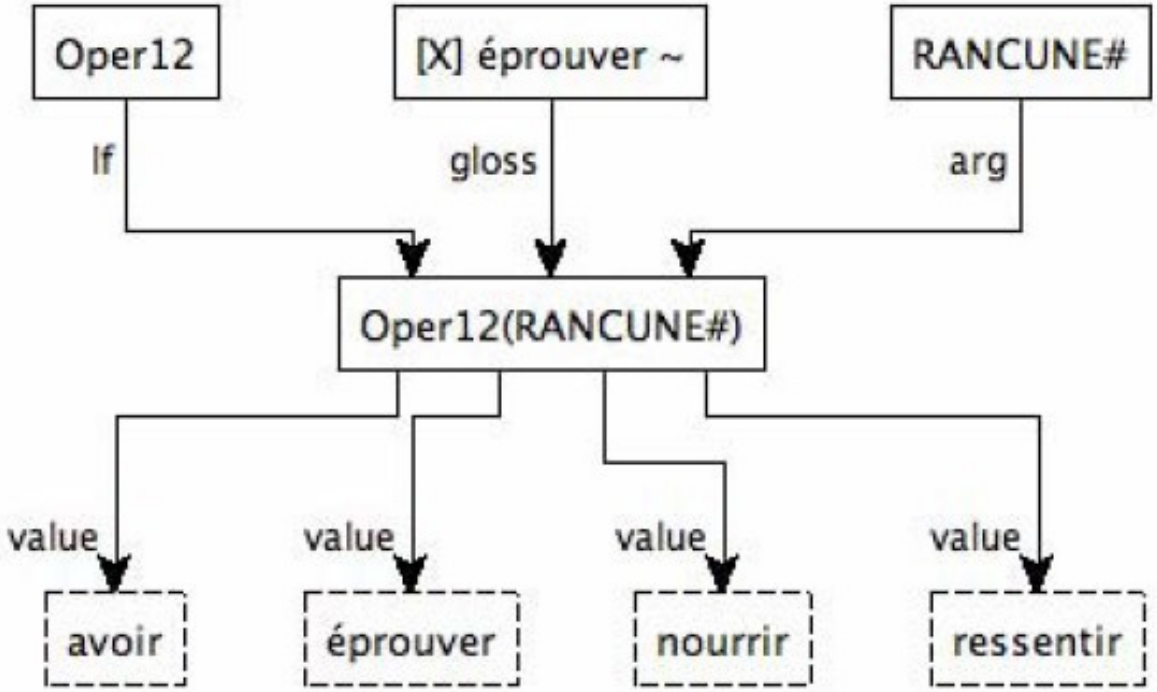
- **Matérialisation des relations** sous forme de nœuds
  - Inspiré par les graphes existenciels (Peirce)
  - Gestion des relations n-aires
- (Sowa, 1976)





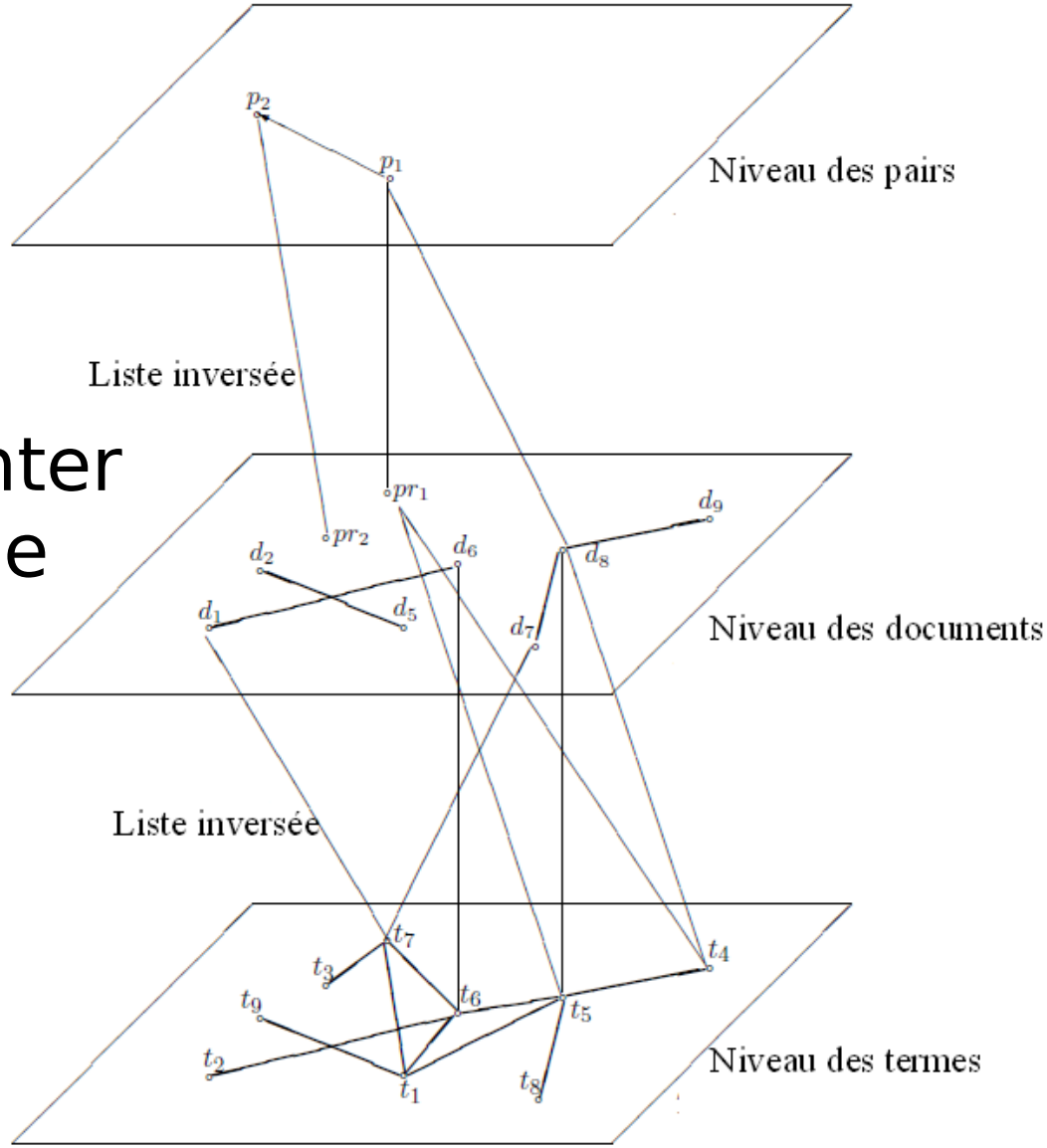
Inspiration : Systèmes lexicaux

- Graphes orientés, pondérés, non hiérarchiques
- **Hétérogénéité**
  - peut représenter des termes, des collocations, des sens, etc.
- (Polguère, 2006)



Inspiration : MLAG

- **Multiniveau**
  - Arcs entre nœuds de niveaux successifs
  - Permet de représenter plusieurs niveaux de connaissances
- (Witschel, 2007)

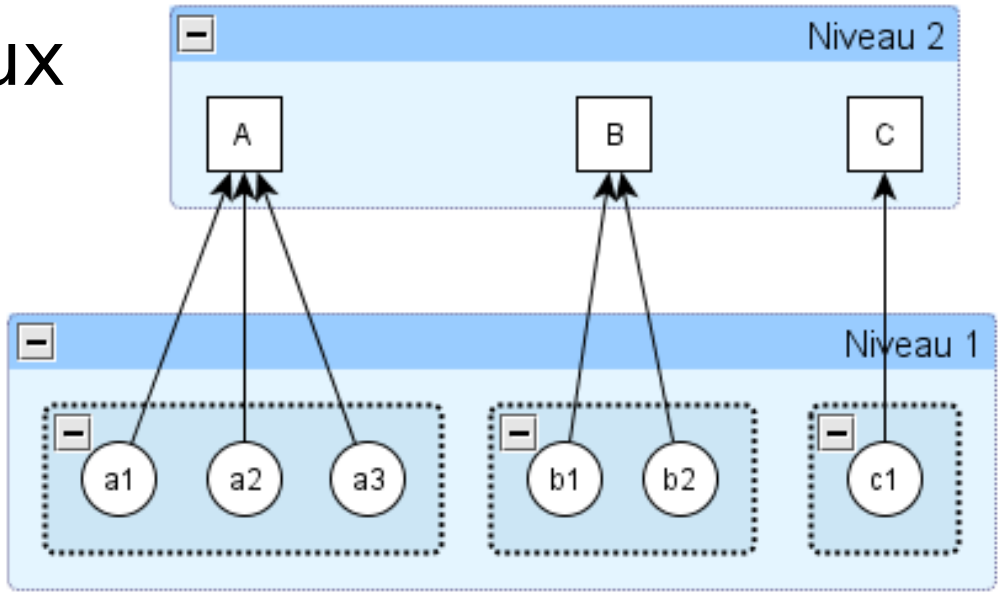


Plan

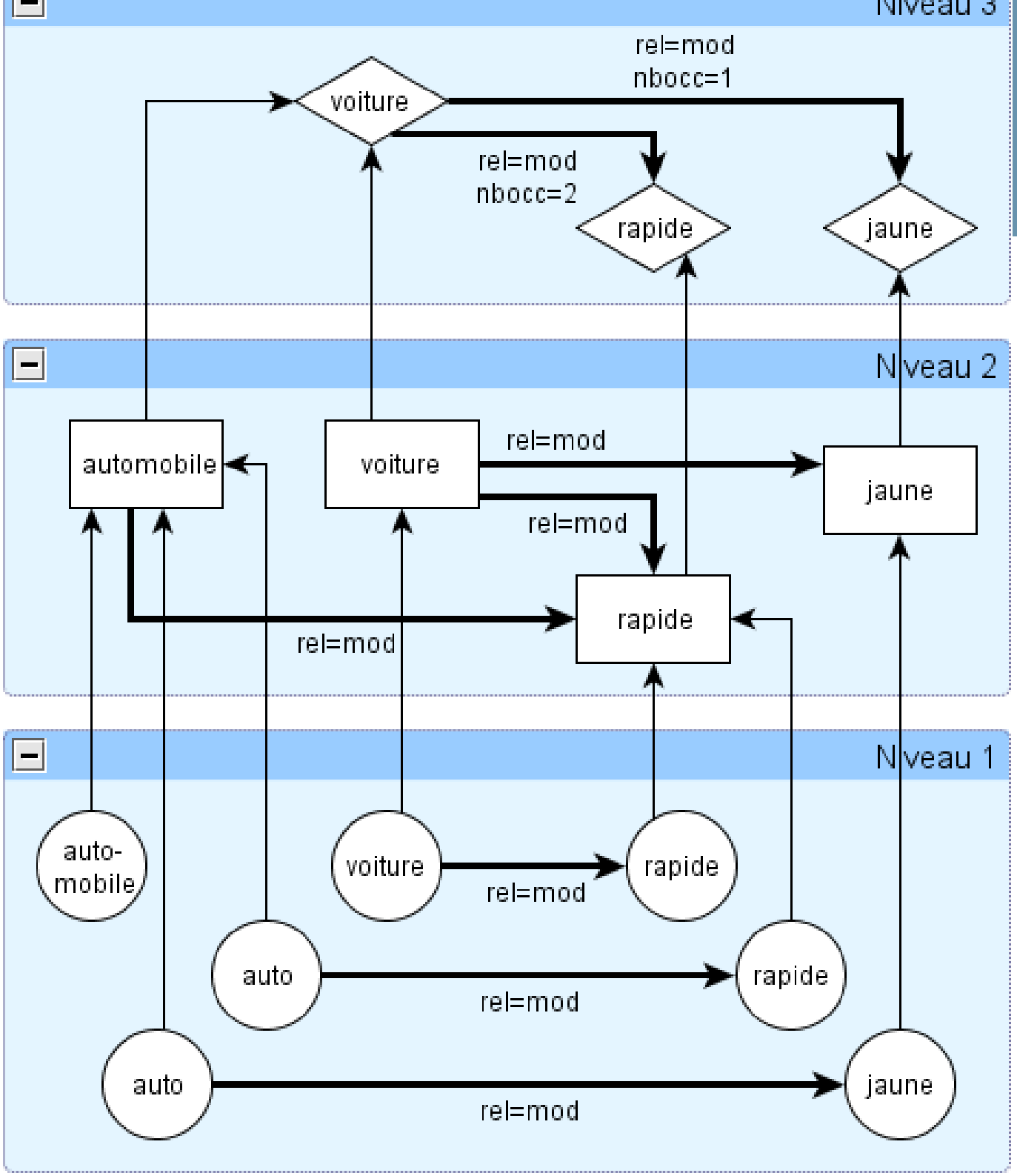
- Étude du problème de l'extraction
  - Premières expérimentations
  - Cahier des charges
  - Modèle de graphes linguistiques existants
- Solution : MuLLinG
  - Modèle de graphe linguistique multiniveau
  - Opérations
  - Implémentation
- Utilisation : expérimentations
  - Extraction de collocations/bicollocations
  - Pondération de traductions lexicales

MuLLinG : structure de graphe multiniveau

- Niveaux distincts = vues différentes
- Regroupement par classes d'équivalence
  - 1 classe d'équivalence = 1 nœud au niveau supérieur
  - Hiérarchie de niveaux
- Arcs interniveau
  - Lien entre un nœud et sa classe d'équivalence



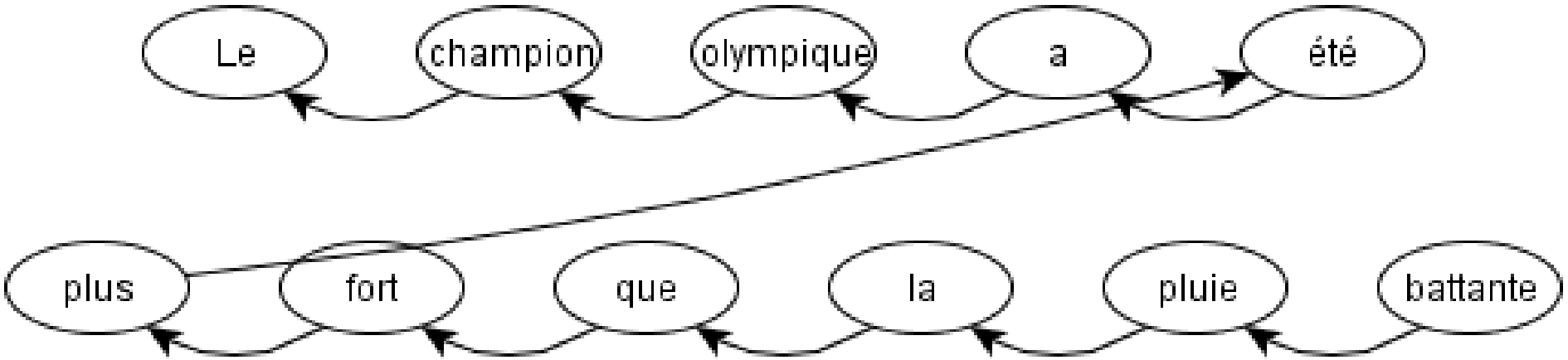
- Attributs libres



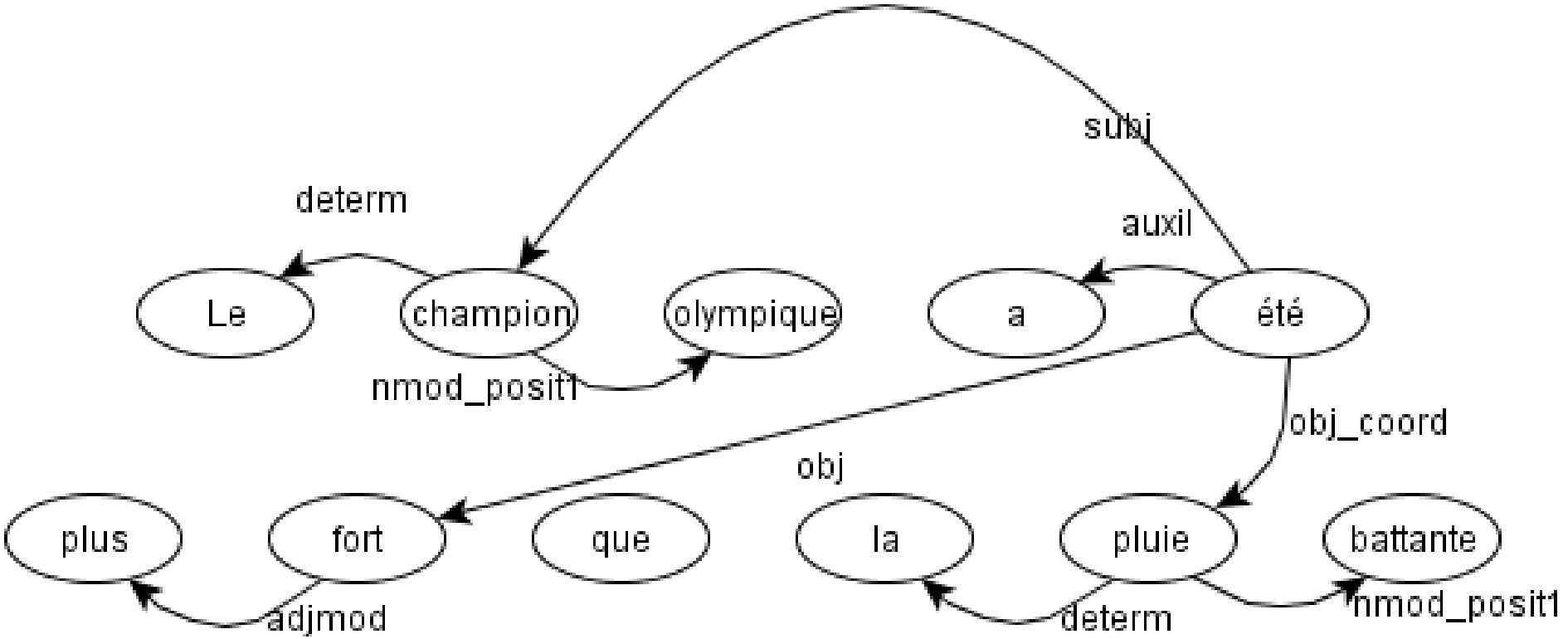
Utilisation typique : corpus de documents

Le champion olympique a été plus fort que la pluie battante. Comme à Pékin, il y eut Bolt, sculpture moulée dans son maillot jaune, et les autres. Détenteur du record du monde en 19 sec 30, la foudre a infligé un écart de 82/100 à son suivant, l'Américain LaShawn Merritt, médaillé d'or du 400 m aux derniers JO.

Utilisation : contexte linéaire



Utilisation : contexte de dépendance



Opérations de base

- Ajouter/Supprimer un nœud ou un arc (interniveau, intraniveau)
  - Nettoyer un nœud (supprimer les arcs dont il est extrémité)
  - Supprimer un nœud et sa descendance
- Appliquer une fonction (avec condition)
  - Ex. : suppression conditionnelle de nœuds/ d'arcs
- Calcul de mesures

Opérations : principe général

- Toute opération modifiant le graphe prend comme paramètres, **à fixer par l'utilisateur** :
  - Niveau
  - Fonction de filtrage
  - Fonctions de calcul sur les attributs des objets du graphe (ex. : nombre d'occurrences)
- ⇒ Séparation des préoccupations
  - Système : cohérence du graphe, parcours

Émergence

- Passage au niveau supérieur
  - Basé sur les classes d'équivalence
  - Avant : relations entre objets
  - Après : relations (regroupées) entre objets regroupés
- Paramètre **fixé par l'utilisateur** :
  - Fonction renvoyant l'identifiant de la classe d'équivalence d'un nœud/d'un arc

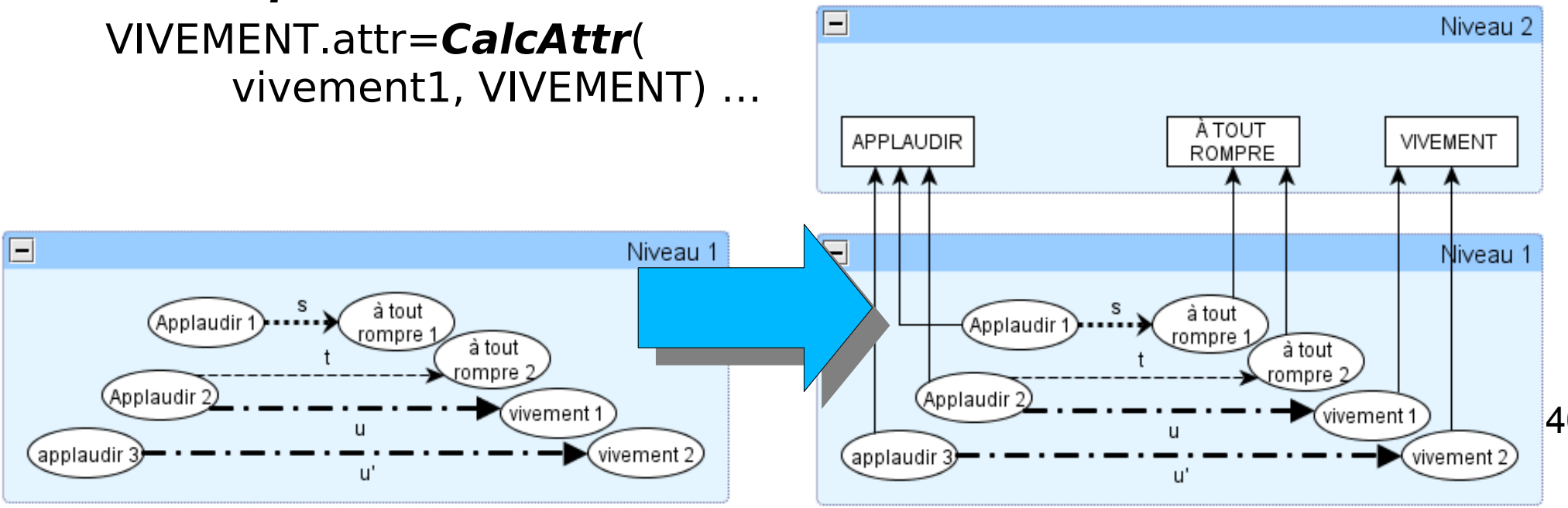
Opérateurs : émergence de nœuds

- 1 nœud = 1 classe d'équivalence
  - Reliés aux nœuds du niveau inférieur appartenant à la classe qu'ils représentent

Niveau : ***1*** Filtrage : ***vrai***

ClassÉquiv(vivement1)=VIVEMENT ...

VIVEMENT.attr=***CalcAttr***(
vivement1, VIVEMENT) ...



Opérateurs : émergence d'arcs intraniveau

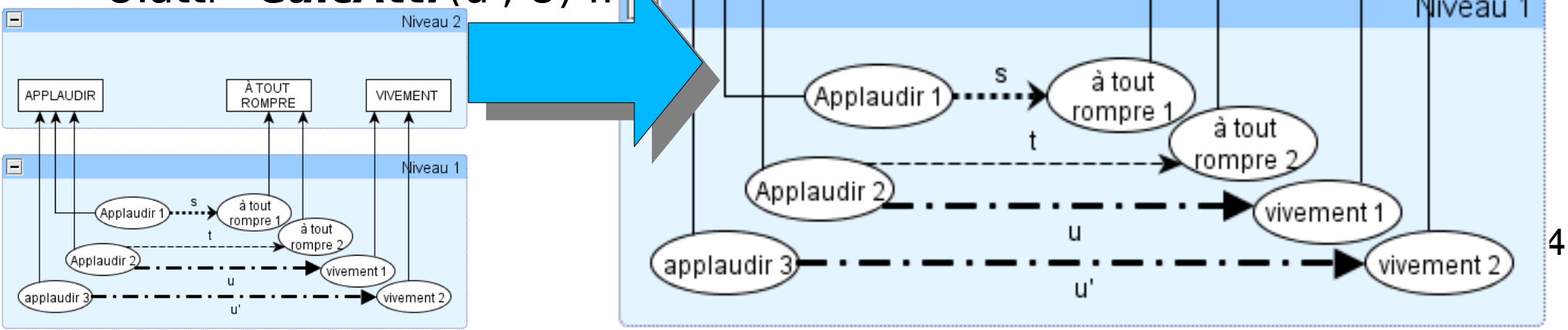
- 1 arc entre A et B = 1 ensemble d'arcs
  - Entre un élément de la classe A et un élément de la classe B
  - Équivalents

Niveau : ***1***

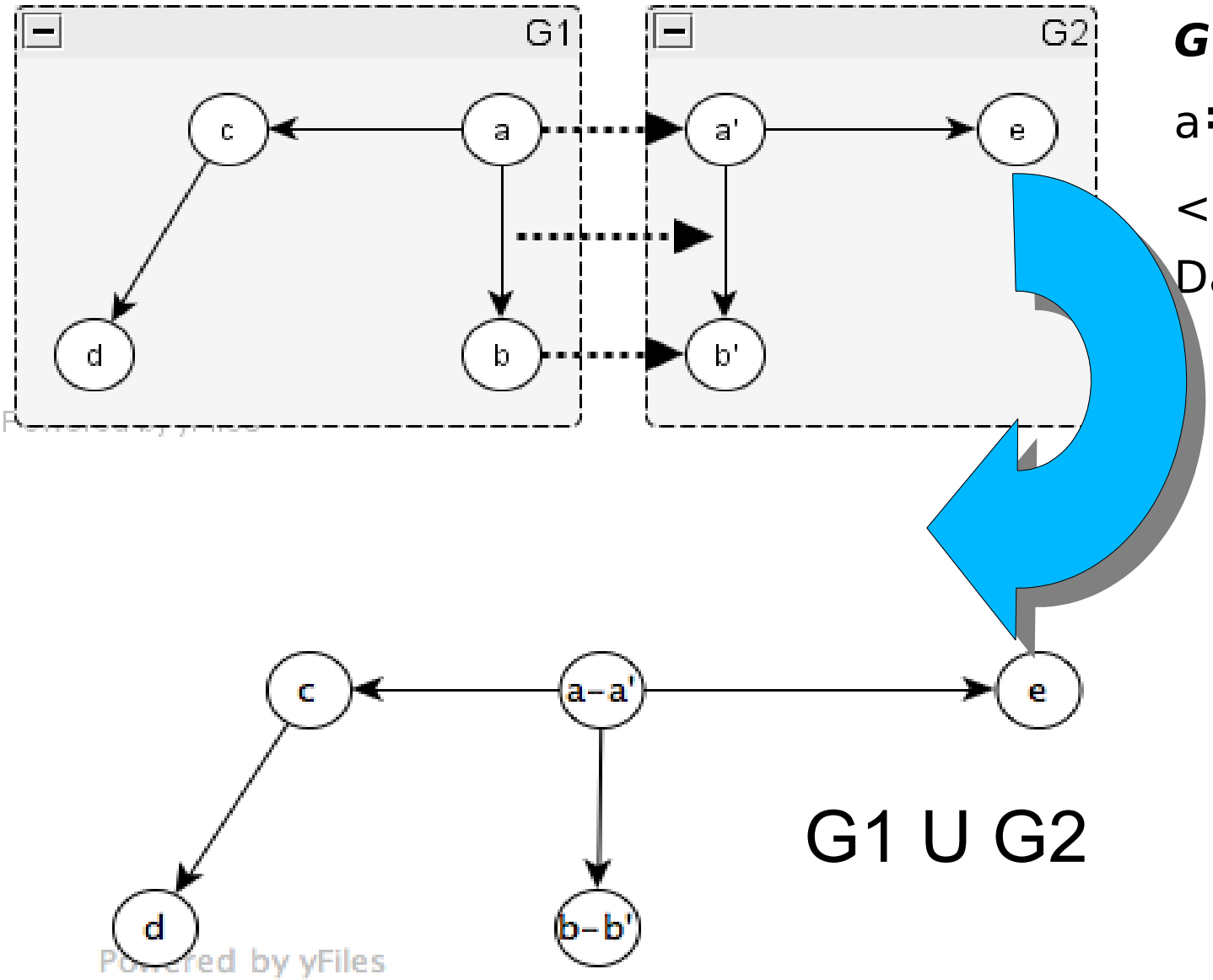
Filtrage : ***vrai***

ClassÉquiv(u')=U ...

U.attr=***CalcAttr***(u', U) ...



Opérateurs : union



G1 ***G2***

a=a'

<a,b>=<a',b'>

Dans le nouveau graphe :

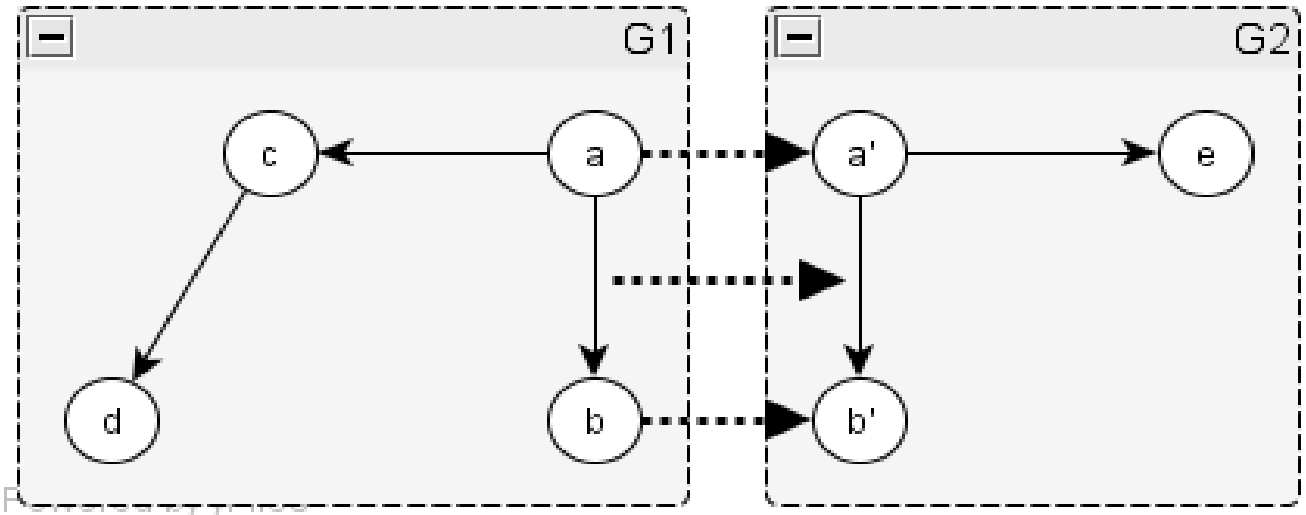
c.attr= ***cop-noeud***(G1.c)

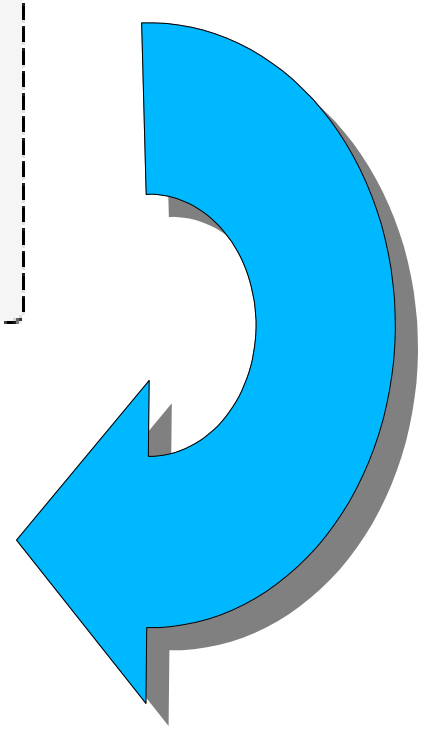
<c,d>.attr= ***cop-arc***(G1.<c,d>)

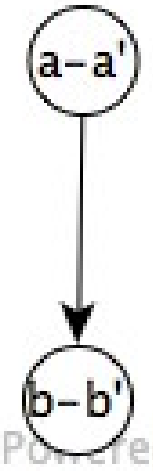
a-a'.attr= ***fus-noeud***(G1.a, G2.a')

<a-a',b-b'>.attr= ***fus-arc***(G1.<a,b>, G2.<a',b'>)

Opérateurs : intersection

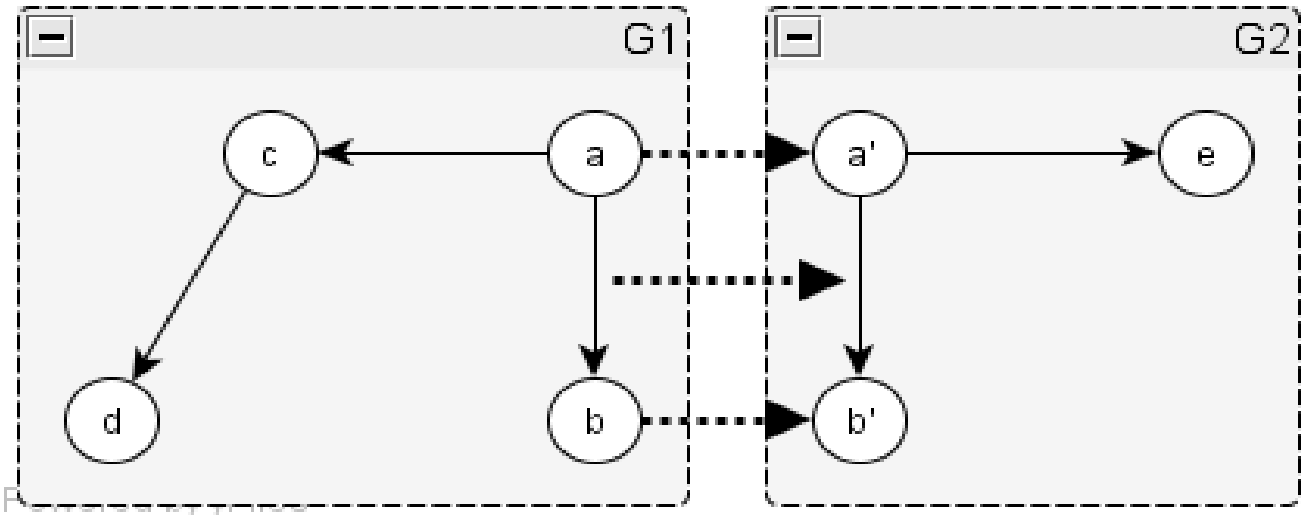


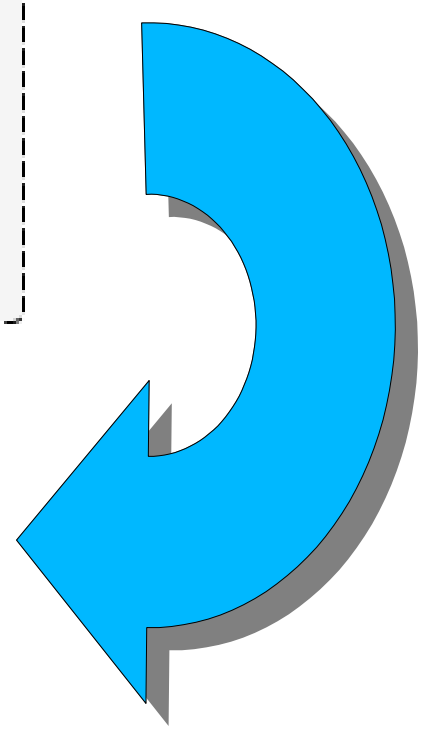


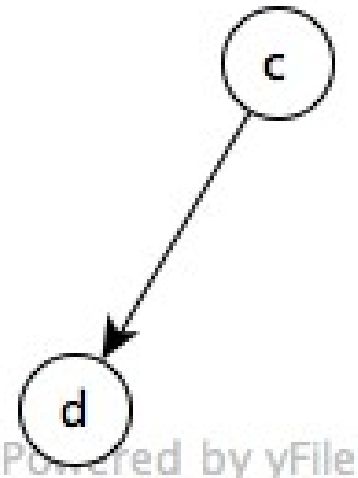


G1 ∩ G2

Opérateurs : complément (différence)



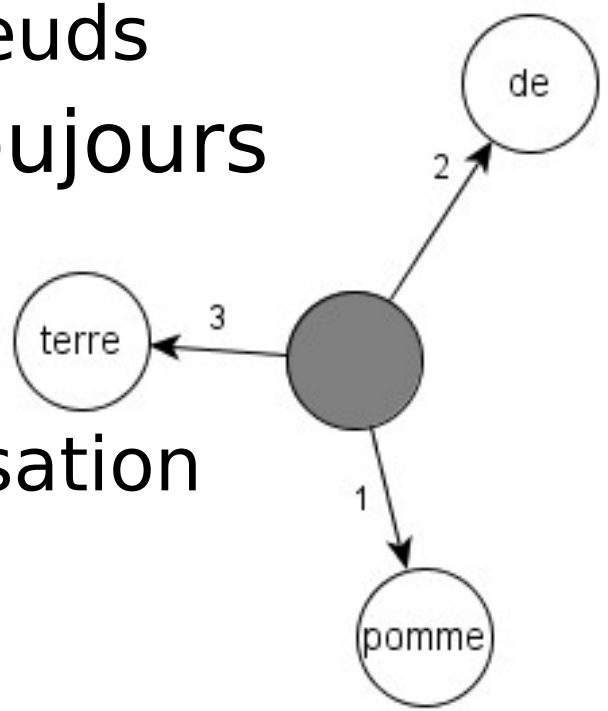




G1 \ G2

Représentation « complexe »

- Représentation simple (jusqu'ici)
  - 1 relation = 1 arc entre les nœuds
- Mais une relation n'est pas toujours binaire
- Représentation **complexe**
  - 1 relation = 1 nœud matérialisation de la relation + des *arcs arguments* numérotés
  - Le nœud matérialisant la relation est un nœud standard

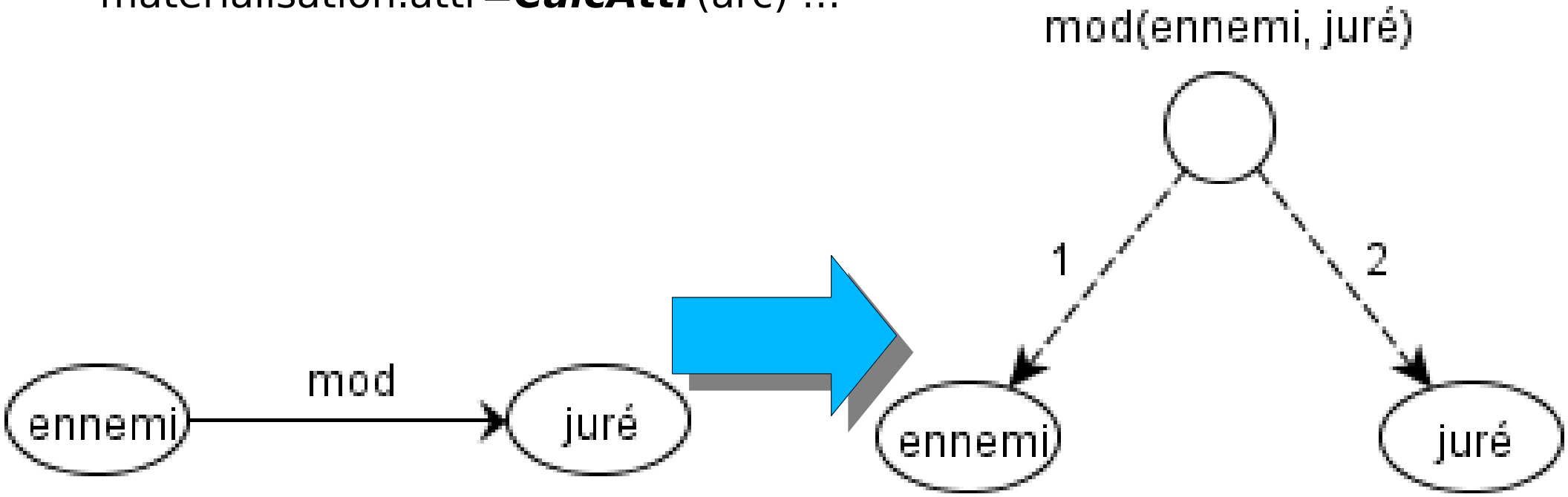


Nouvel opérateur : matérialisation

Niveau : ***1***

Filtrage : ***typerel=mod***

materialisation.attr=***CalcAttr***(arc) ...



Opérateurs adaptés

- Émergence
- Calcul de mesures
- Union, intersection, différence
  - Ajout de la copie ou fusion des arcs arguments

Implémentation

- Bibliothèque C++
  - Utilise *Boost Graph Library* (BGL), open-source : gestion des accès, des parcours
  - Attributs
- GraphML format d'entrée/sortie
  - XML
  - Standard de la description de graphe (BGL, éditeurs *YEd Graph Editor*, *JUNG*)

Plan

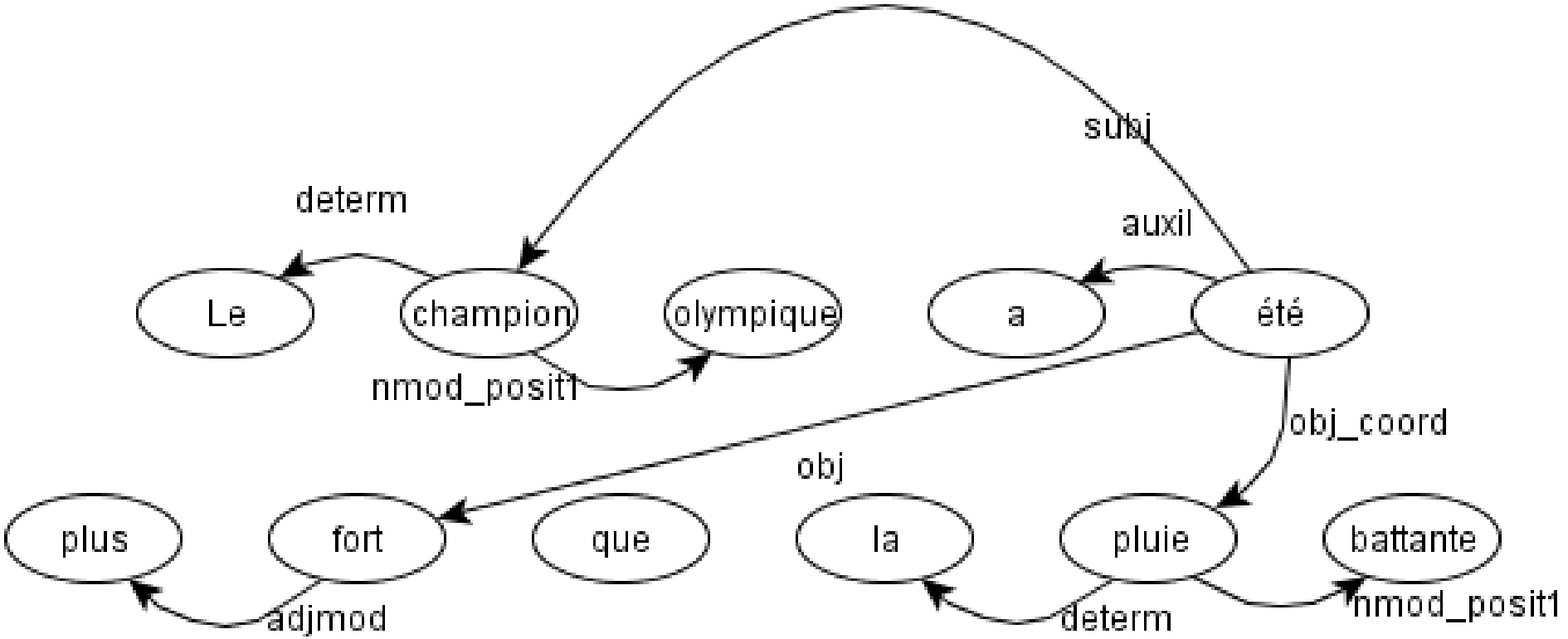
- Étude du problème de l'extraction
  - Premières expérimentations
  - Cahier des charges
  - Modèle de graphes linguistiques existants
- Solution : MuLLinG
  - Modèle de graphe linguistique multiniveau
  - Opérations
  - Implémentation
- Utilisation : expérimentations
  - Extraction de collocations/bicollocations
  - Pondération de traductions lexicales

Expérimentations basées sur le modèle

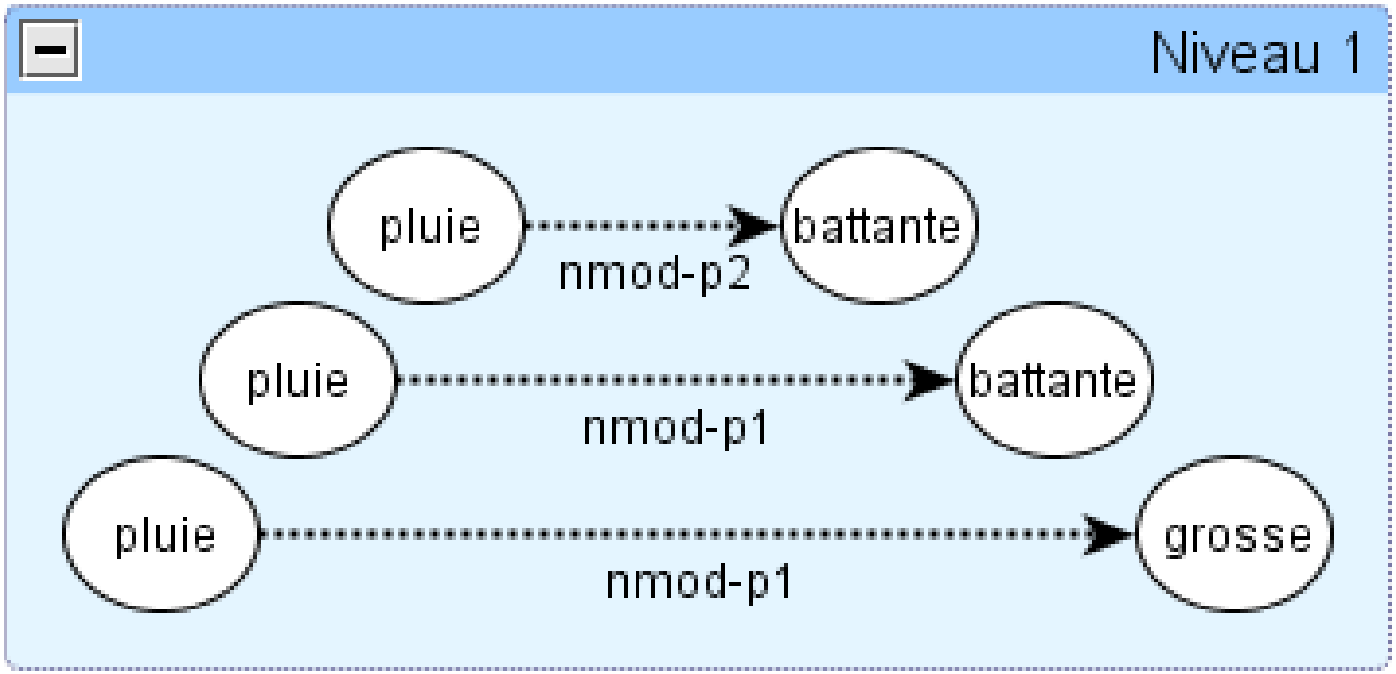
- Le but de nos travaux n'est pas de proposer la méthode la plus efficace...
  - Qualité des résultats produits dépend des ressources (et leur analyse), des mesures
- ...mais de proposer les outils pour mettre en œuvre les méthodes

Expérimentations : extraction de collocations

- Reprise des expérimentations initiales
- Graphe de départ

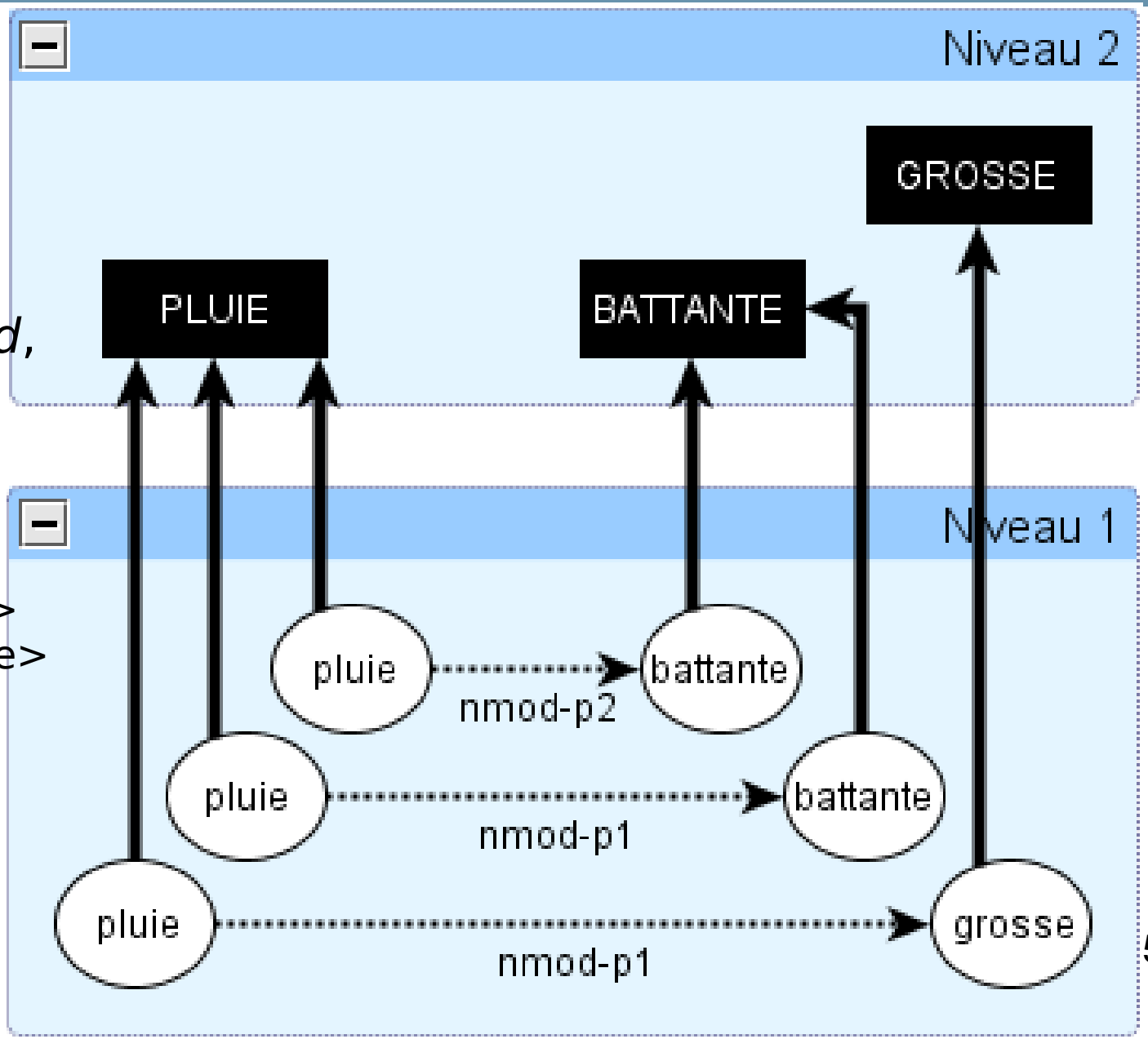


Extraction monolingue : filtrage



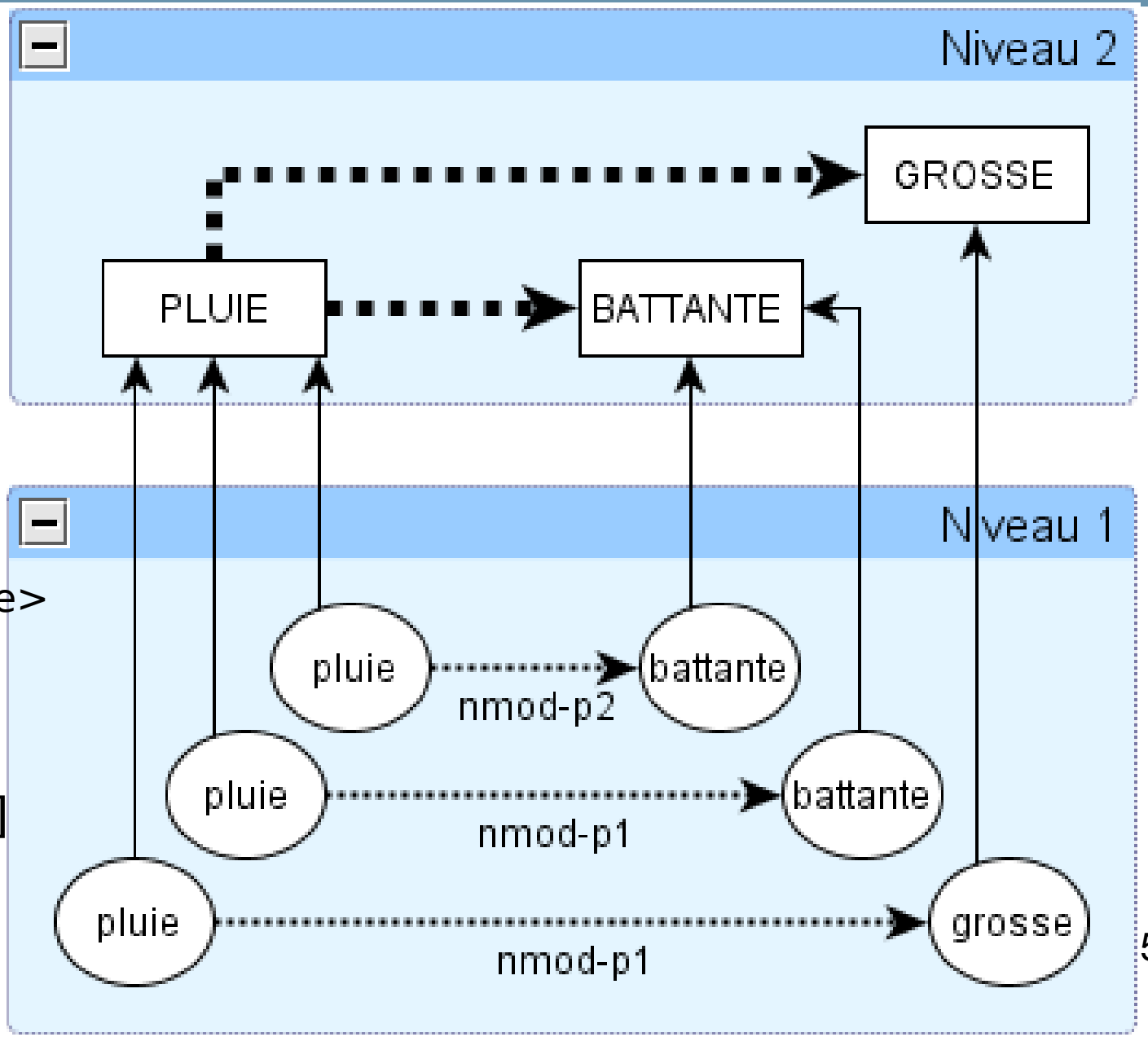
Extraction monolingue : émergence de nœuds

Emergence
DeNoeuds(
1,
vrai,
ClassEquivNœud,
//lemme
+pos
CalcAttrNœud,
//[<id, Classe>
<type, TypeTerme>
<nbocc, Incremente>
]
)

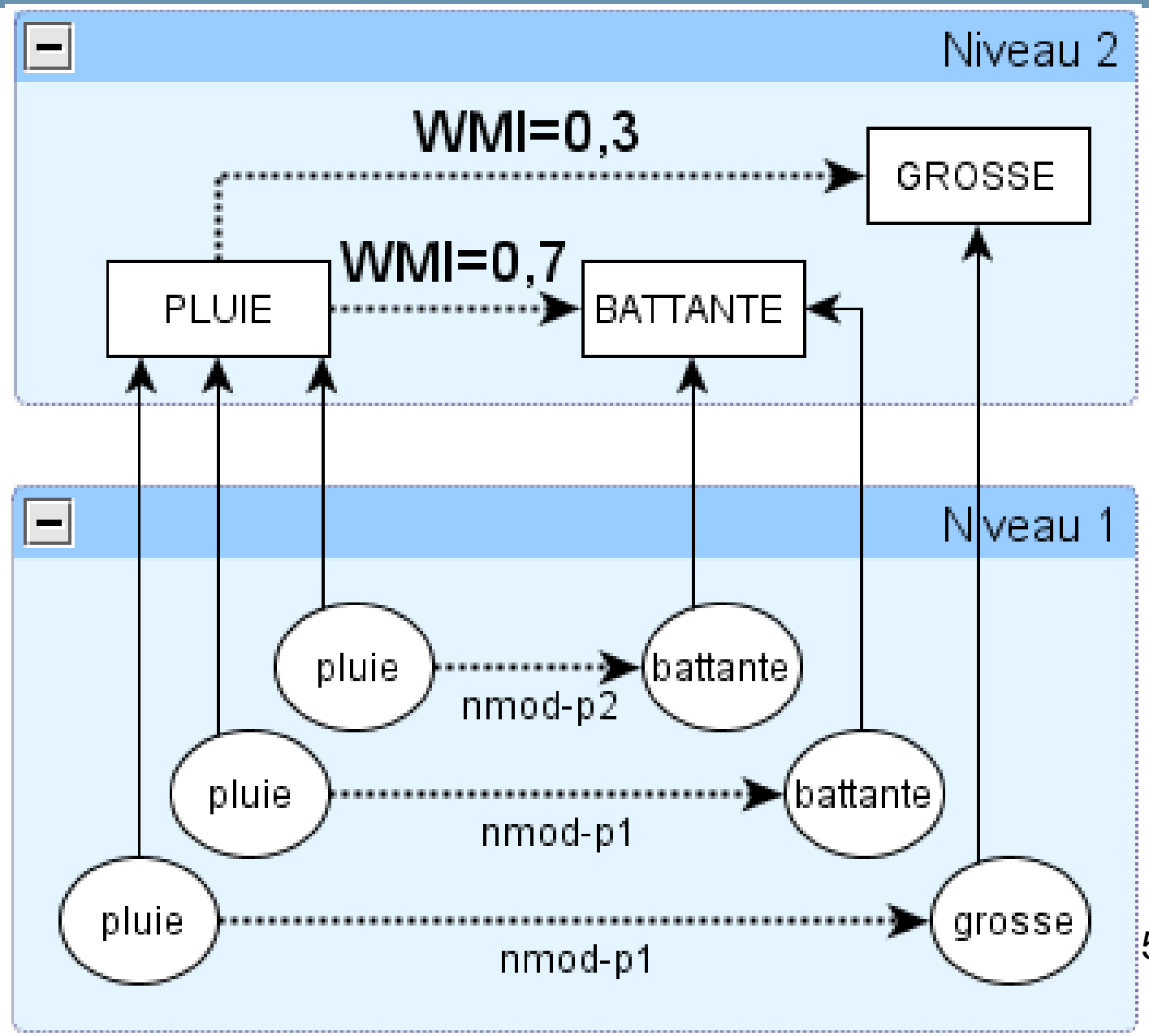


Extraction monolingue : émergence d'arcs

Emergence
DArcs(
1,
vrai,
ClassEquivArc,
// *type*
CalcAttrArc,
//[<id, Classe>
<type, TypeMod>
<nbocc, Incremente>
]
CalcAttrSource,
//[
<d+, incrémente>]
CalcAttrCible,
//[
<d-, incrémente>]



Extraction monolingue : calcul de mesures



Extraction monolingue : expérimentation

- Mesures d'association utilisées :
  - Information mutuelle
  - WMI
- Corpus utilisé : *LeMonde95*

Expérimentations	*Niveau 1*		*Niveau 2*	
	Nœuds	*Arcs*	*Nœuds*	*Arcs*
verbe-adverbe	1 155 824	1 780 759	6 813	**144 586**
nom-adjectif	1 319 474	2 009 051	33 132	**273 655**

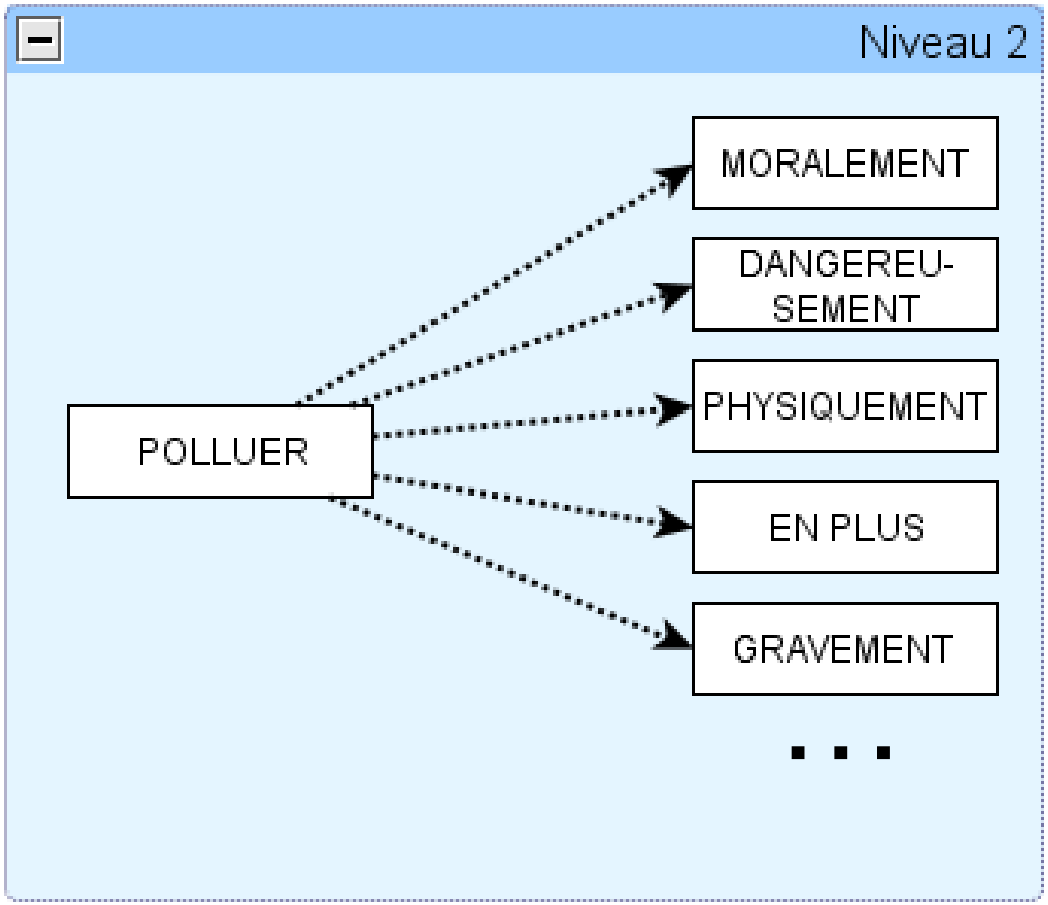
Extraction monolingue : résultats

- Les résultats sont cohérents avec l'expérimentation initiale

10 premiers candidats (WMI+filtrage)
pouvoir difficilement
travailler ensemble
renvoyer dos à dos
chercher en vain
répéter à l'envi
devoir impérativement
accueillir favorablement
dépasser largement
participer activement
régner sans partage

Classement des collocatifs par base

- Représentation graphique

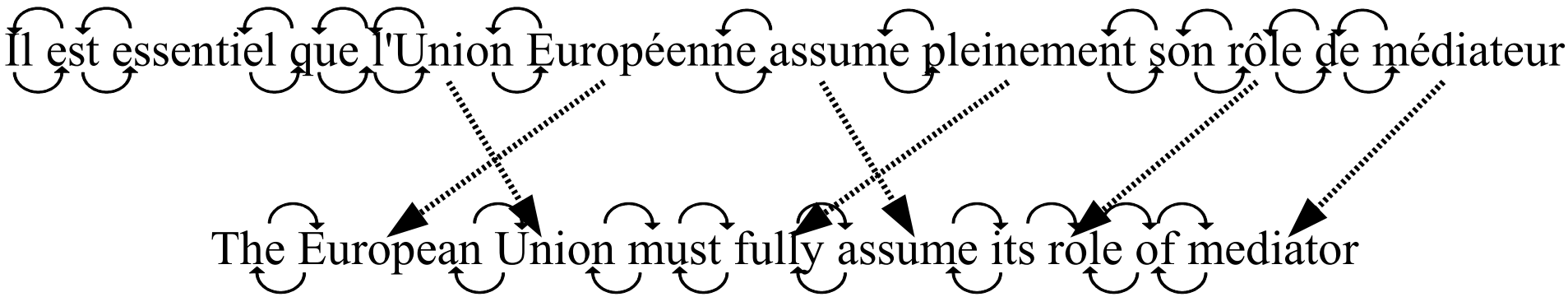


Collocatifs de « polluer » (WMI)
moralement
dangereusement
physiquement
en plus
gravement
considérablement
un peu plus
voire
tant
davantage
mieux
beaucoup
peut-être
moins
même
donc
aussi
aujourd'hui
encore

Collocatifs de « foisonnement » (WMI)
concomitant
instrumental
génial
sympathique
documentaire
parallèle
associatif
incroyable
technologique
riche
présenté
court
immense
vrai
tel
public

Expérimentations : extraction de bicollocations

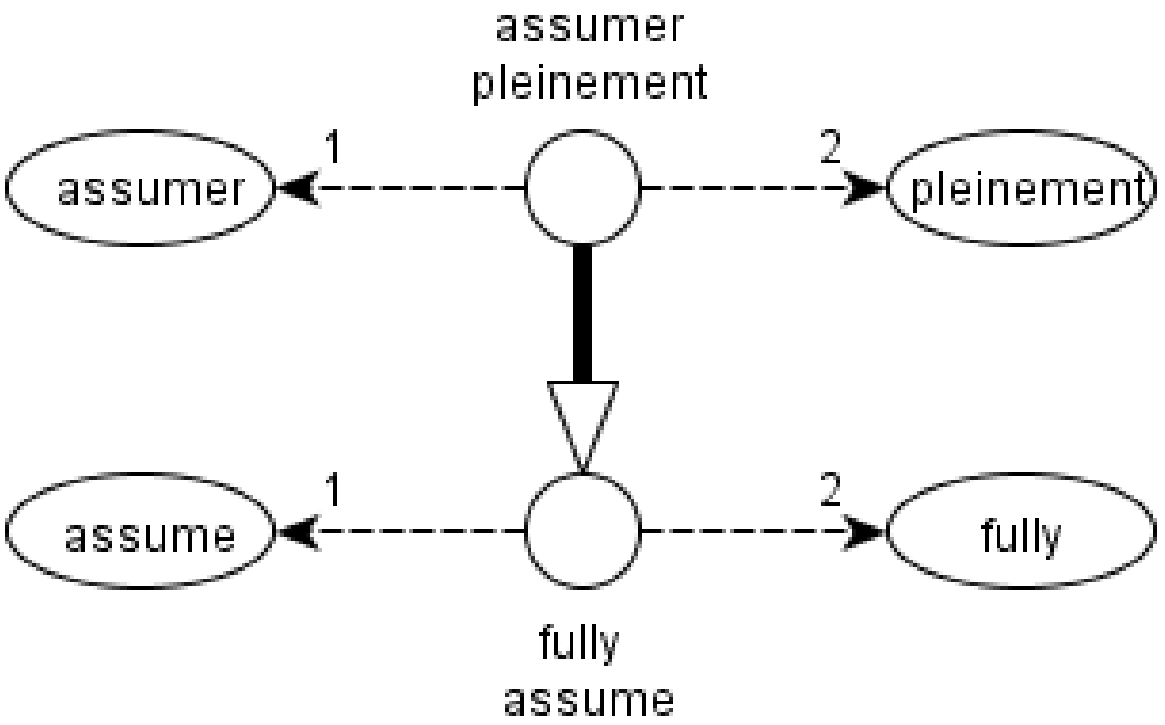
- Graphe de départ :



- Relations de traduction
  - Produites d'après l'alignement des phrases, en utilisant un dictionnaire

Utilisation de la représentation complexe

- Bicollocations = relations bilingues entre collocations monolingues
  - Représentation complexe permet une description claire

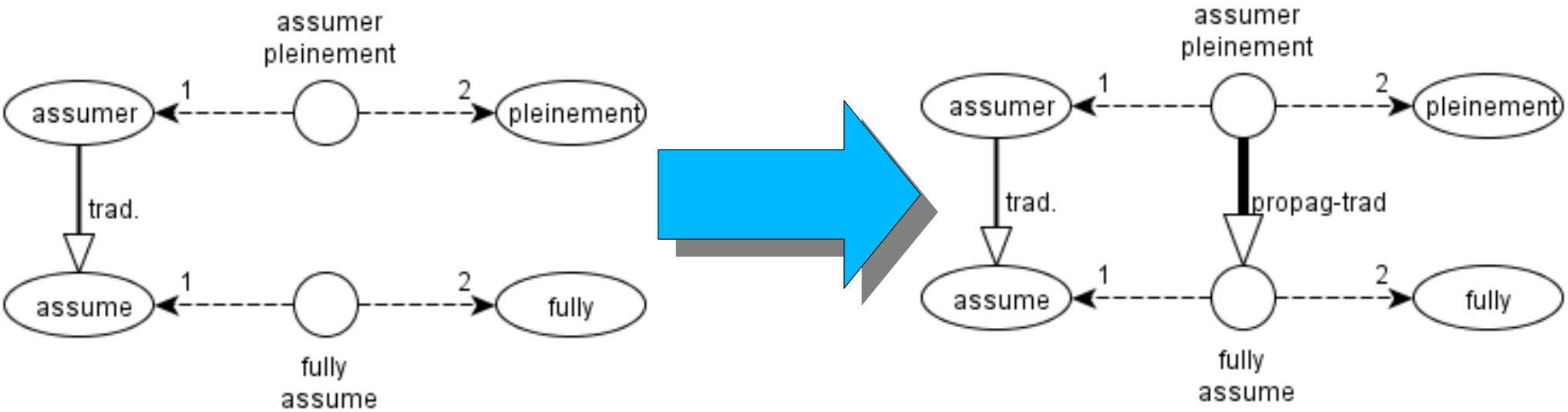


Opérateur spécifique : propagation

- Lien entre deux matérialisations si lien entre les 1ers arguments de chacune.

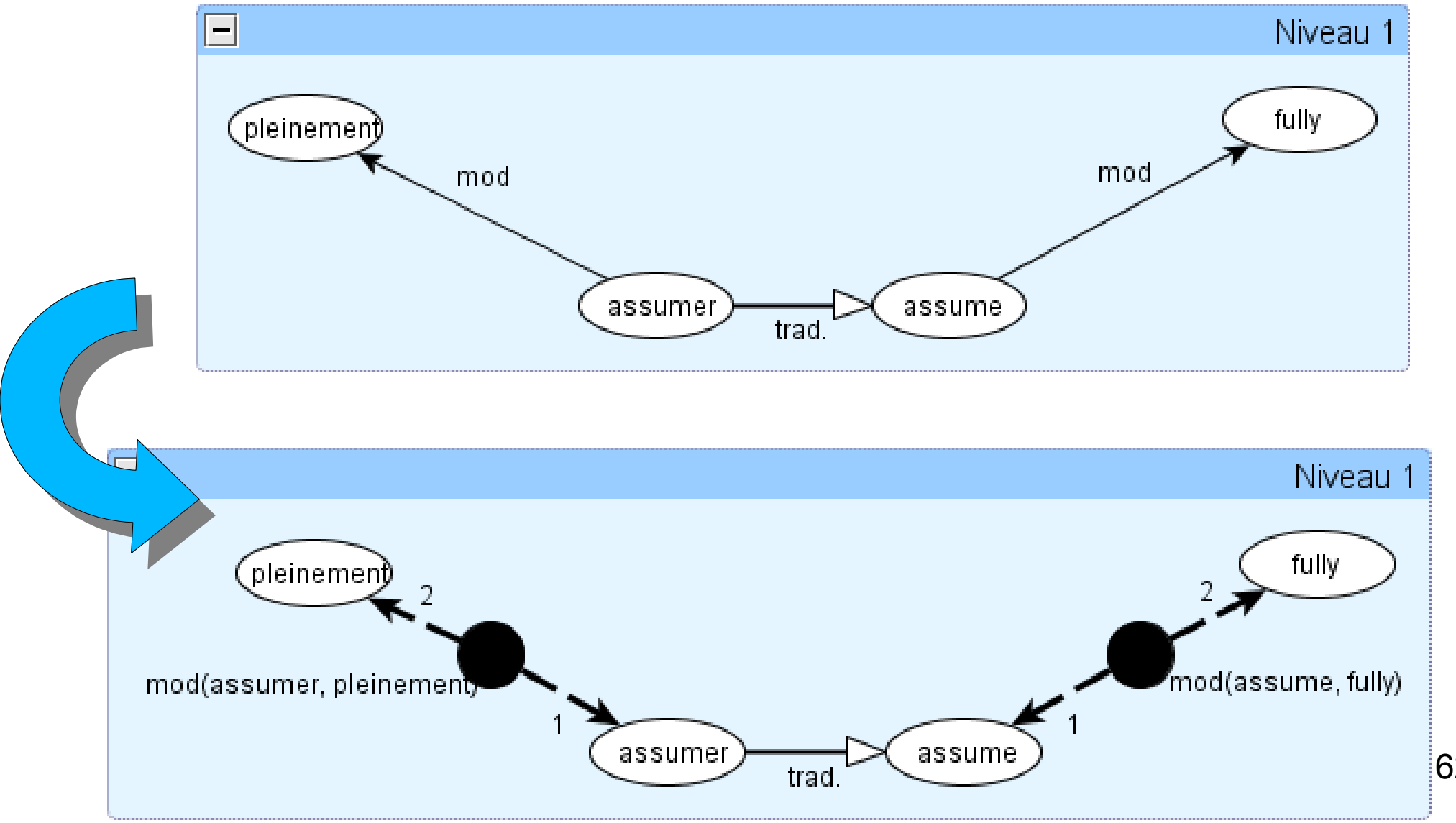
Niveau : ***1*** Filtrage : ***typerel=trad.***

nouvelarc.attr=***CalcAttr***(arc) ...

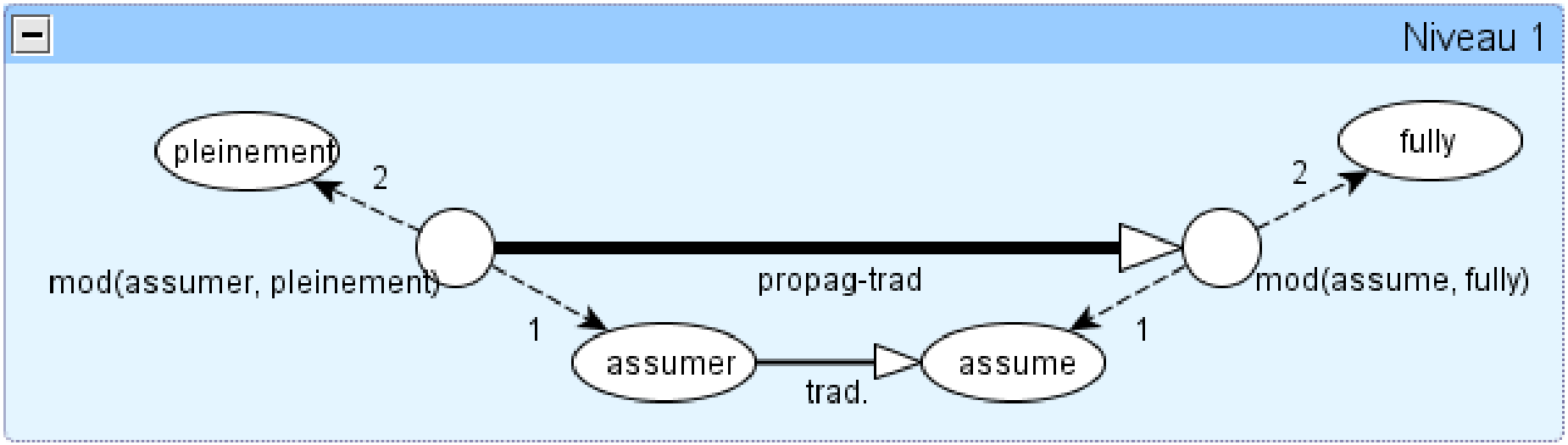


  - Application : Lien entre collocations produit à partir d'un lien entre bases

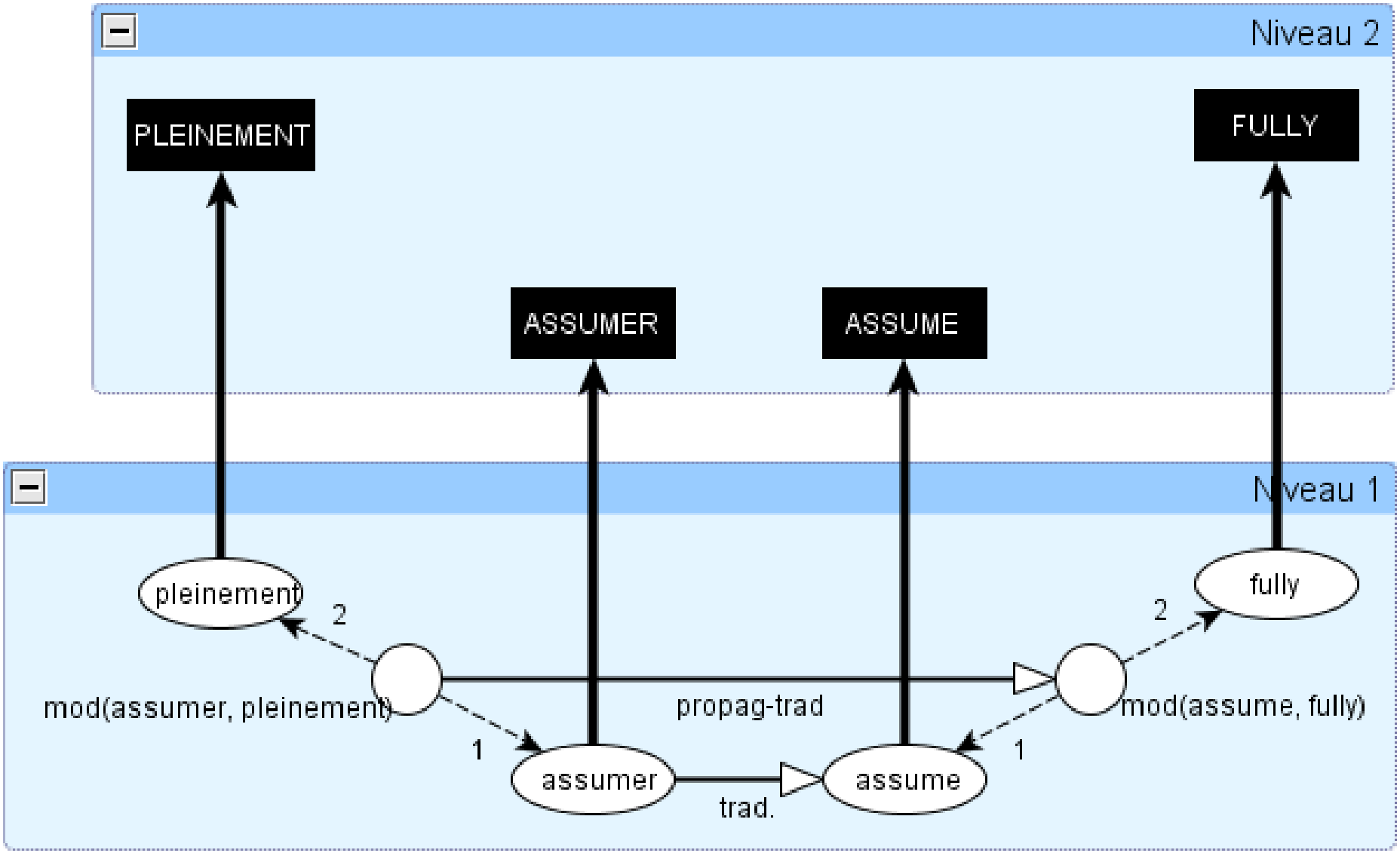
Extraction bilingue : matérialisation des arcs



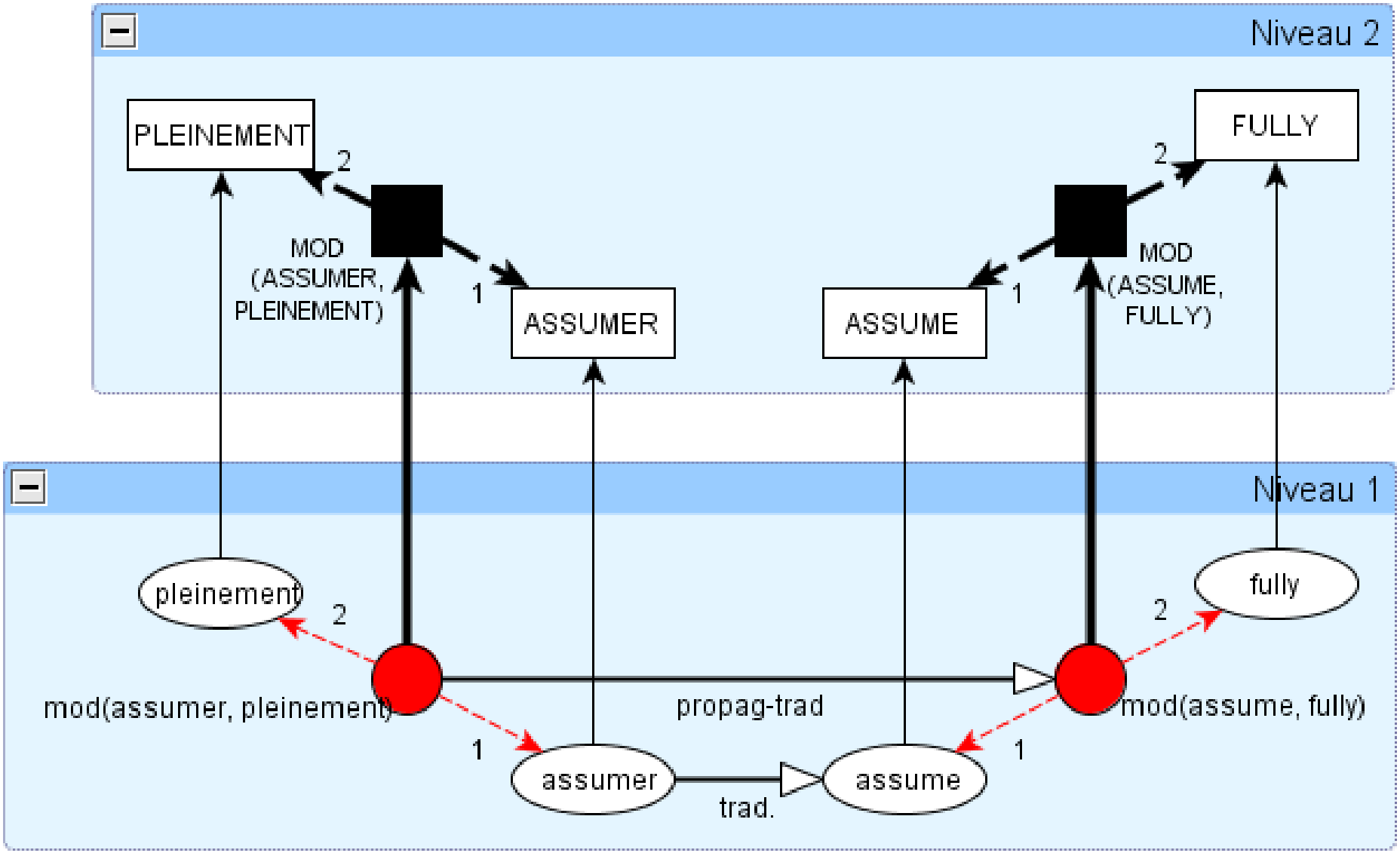
Extraction bilingue : propagation



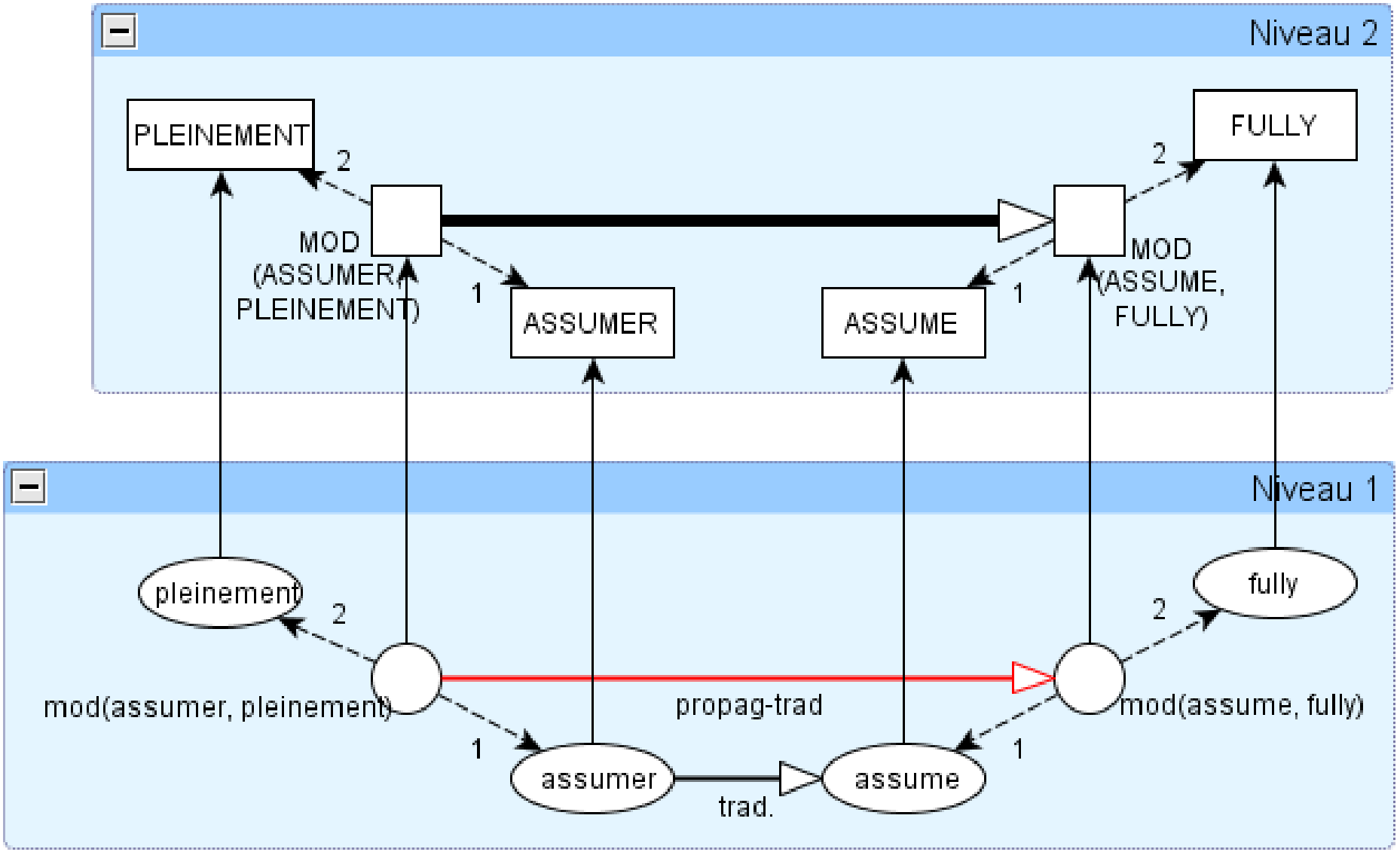
Extraction bilingues : émergence de termes



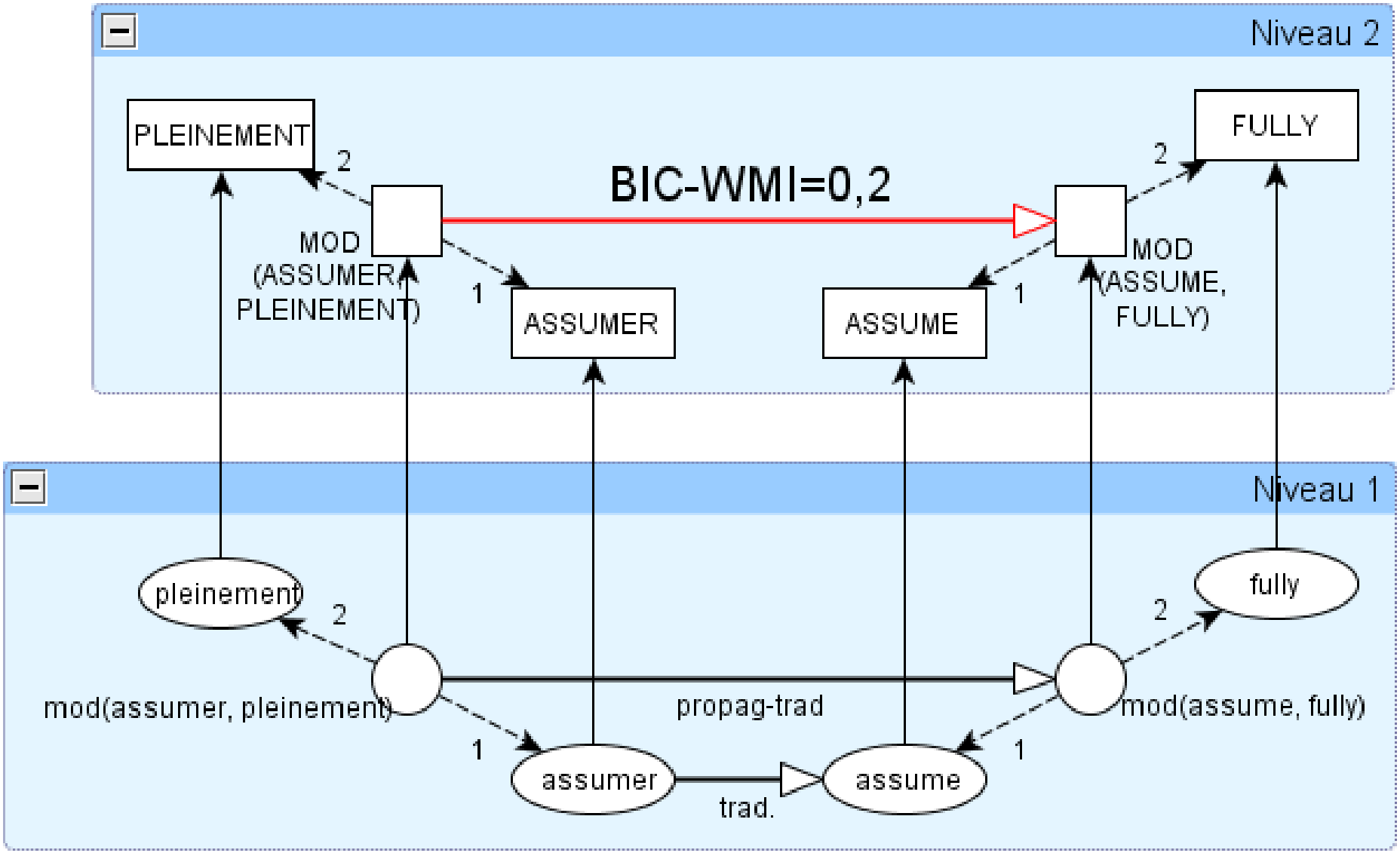
Extraction bilingues : émergence de collocations



Extraction bilingue : émergence de bicollocations



Extraction bilingue : calcul de la mesure bilingue



Extraction de bicollocations : corpus

- Mesures d'association utilisées : Bic-WMI, Bic-MI
- Corpus utilisé : EuroParl (parallèle)

Expérimentations	Niveau 1		Niveau 2	
	Nœuds	Arcs	Nœuds	Arcs
verbe-adverbe	670 998	1 120 868	66 202	120 319
verbe-adverbe sans repr. complexe :	*447 332*	*673 536*	*6 385*	***685***
adjectif-adverbe	1 301 054	2 203 121	93 258	170 262
verbe-adverbe sans repr. complexe :	*862 628*	*1 326 269*	*10 089*	***3 924***

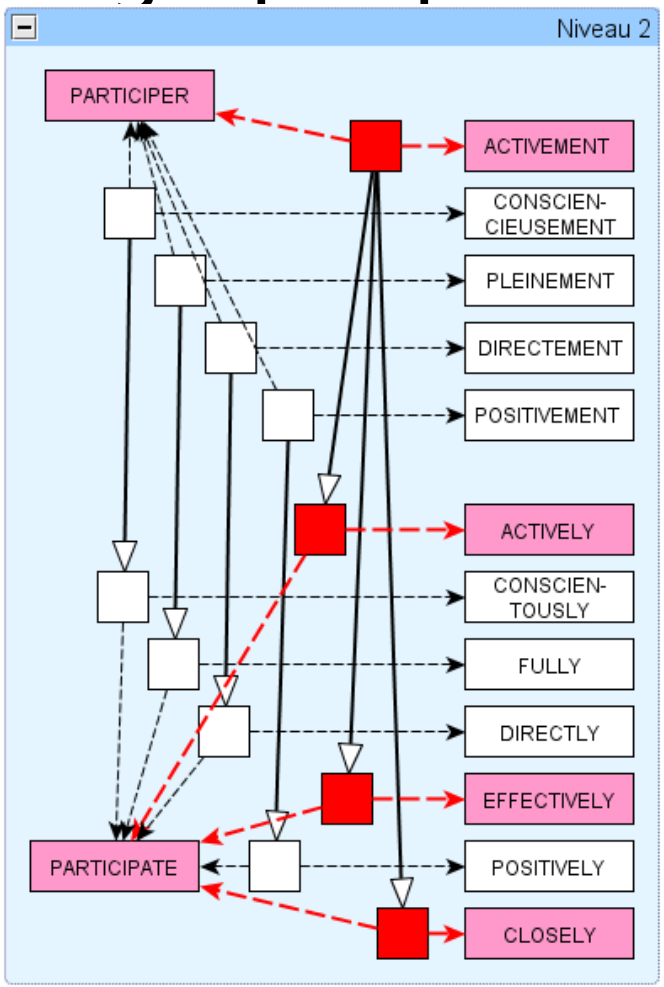
Extraction bilingue : résultats

- Résultats cohérents avec ceux de l'expérimentation initiale

10 premiers candidats (Bic-MI)	
mesurer concomitamment	*measure concomitantly*
prier publiquement	pray ostentatiously
attendre impatiemment	*wait impatienlty*
placer alternativement	*place alternatively*
renvoyer ad	postpone indefinitely
prolonger indéfiniment	***prolong indefinitely***
attendre patiemment	*wait patiently*
payer cher	pay dearly
traiter bestialement	*treat abominably*
frapper durement	***hit hard***

Classement des bicollocations par base

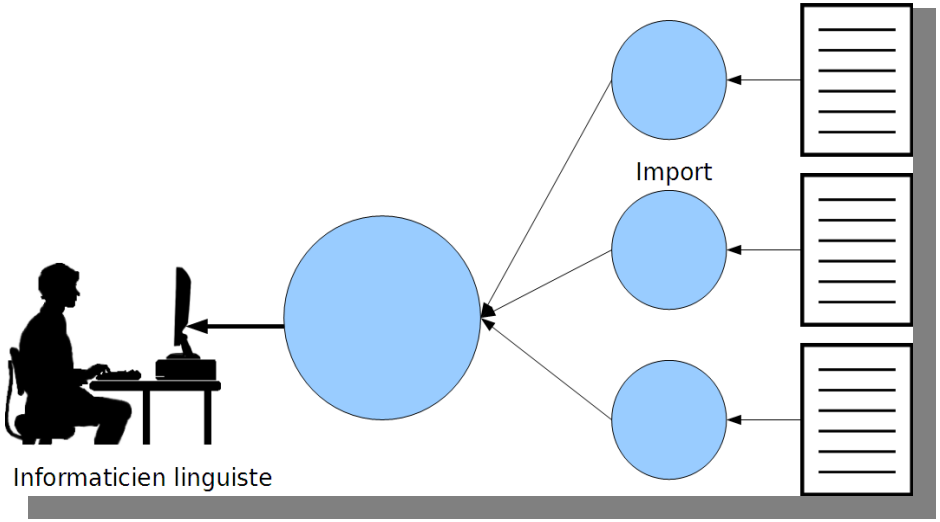
- Représentation graphique



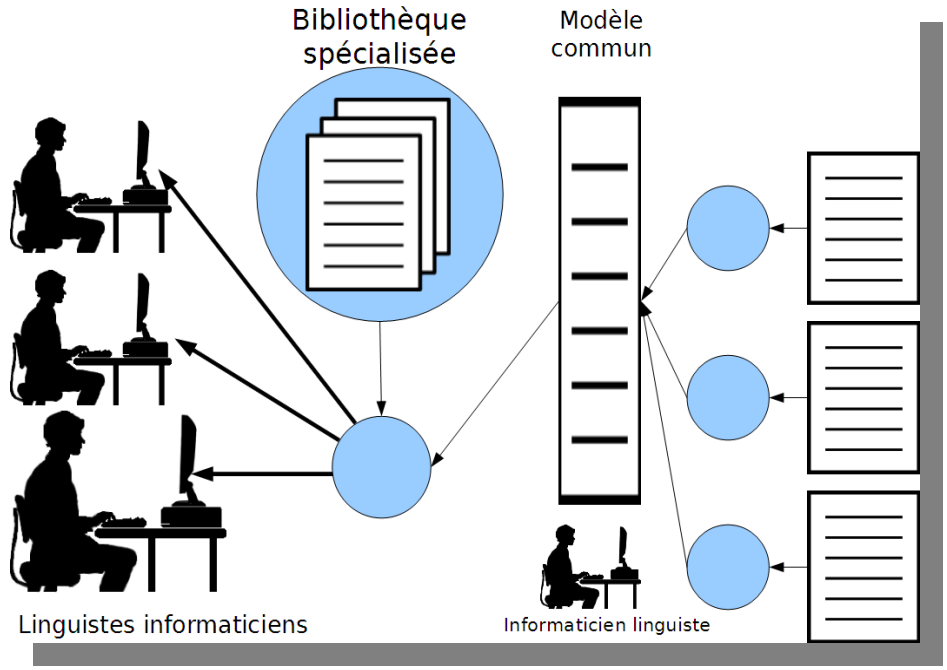
Bicollocations candidates avec base française « participer » (Bic-MI)	
participer activement	***participate actively***
participer consciencieusement	*participate conscientously*
participer pleinement	***participate fully***
participer directement	*participate directly*
participer activement	***participate effectively***
participer positivement	*participate positively*
participer activement	participate closely

Bicollocations candidates avec base française « correct » (Bic-MI)	
formellement correct	***formally correct***
politiquement correct	*politically correct*
juridiquement correct	morally correct
certes correct	politically correct
juridiquement correct	*legally correct*
absolument correct	***entirely correct***
juridiquement correct	not correct
vraiment correct	***quite correct***
très correct	***quite correct***
très correct	***very proper***
plus correct	more correct
plus correct	most correct

Comparaison entre les expérimentations : code



- Ad hoc : 800 lignes
  - Parcours des données
  - +200 lignes (import)



- MuLLinG : 200 lignes
  - Appels des fonctions de la bibliothèque
  - Perfectible
  - +250 lignes (import)
    - + aéré

Comparaison entre les expérimentations (2)

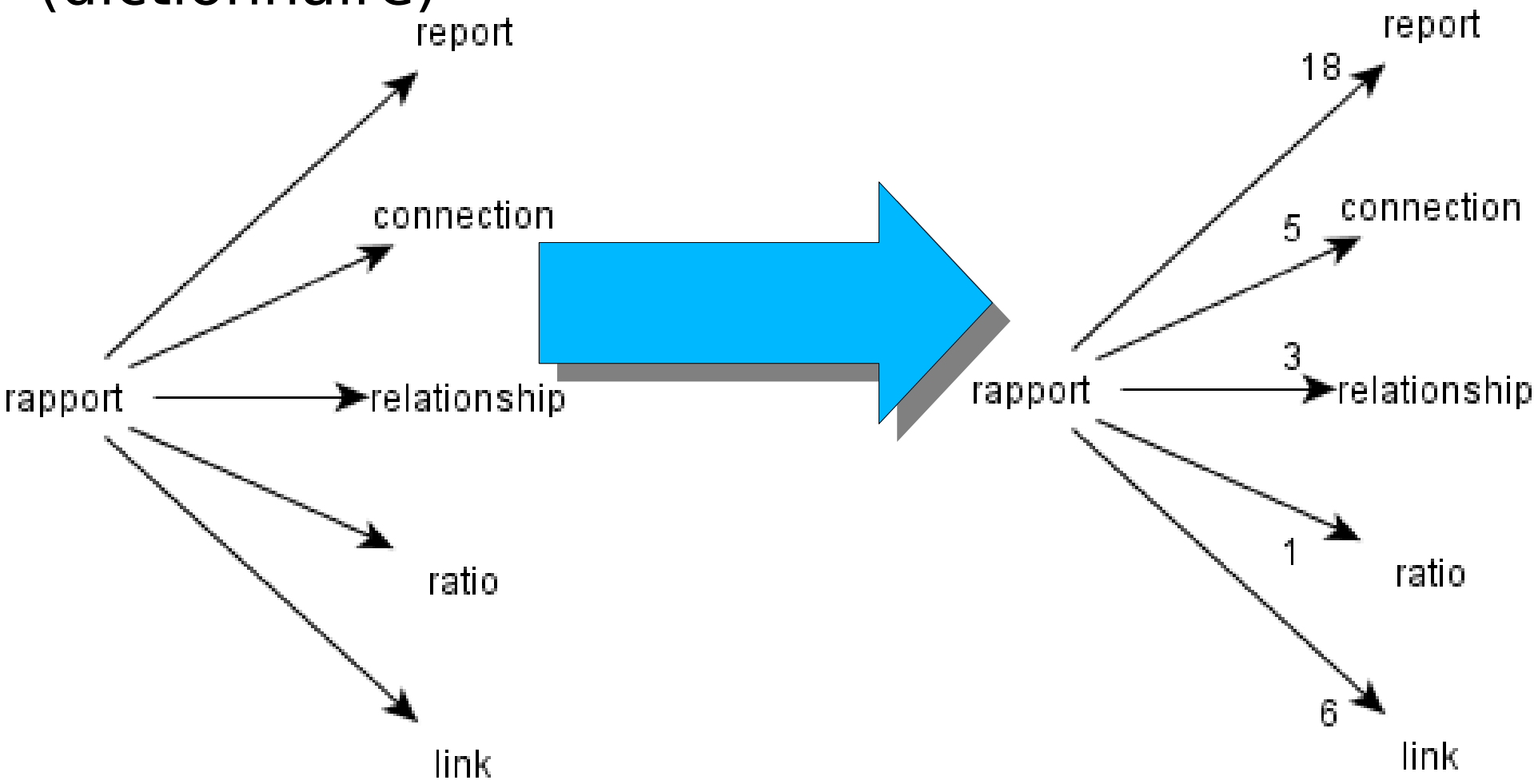
- MuLLinG vs. programme *ad hoc*
- Résultats similaires
  - Avec les mêmes problèmes
- MuLLinG un peu plus lent (BGL)
- *Mais* :
  - Pas de contraintes sur le type de ressource en entrée
  - Description du processus bien plus rapide
  - Réutilisable avec n'importe quel type de relations

Nouvelle expérimentation

- MuLLinG peut être utilisé pour d'autres tâches que l'extraction

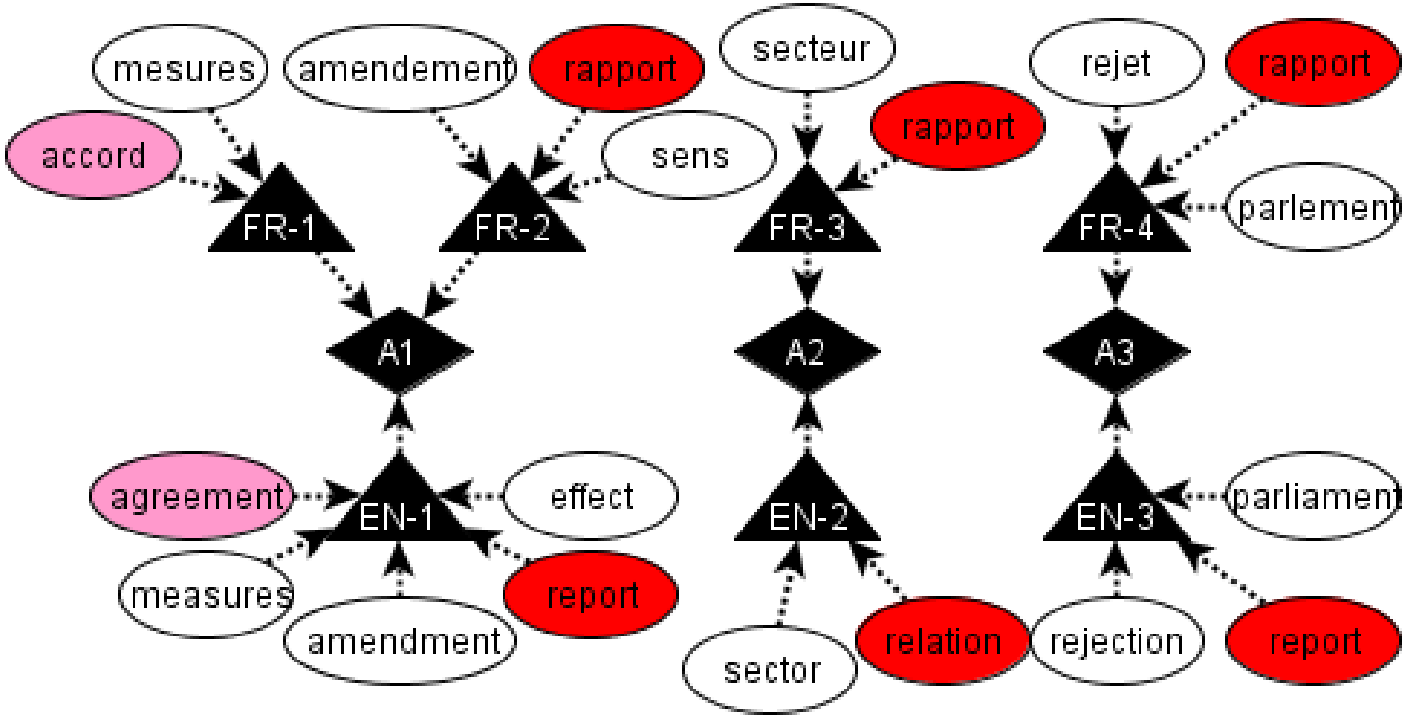
Pondération de traductions lexicales

- Relations de traduction (dictionnaire)
- Relations pondérées

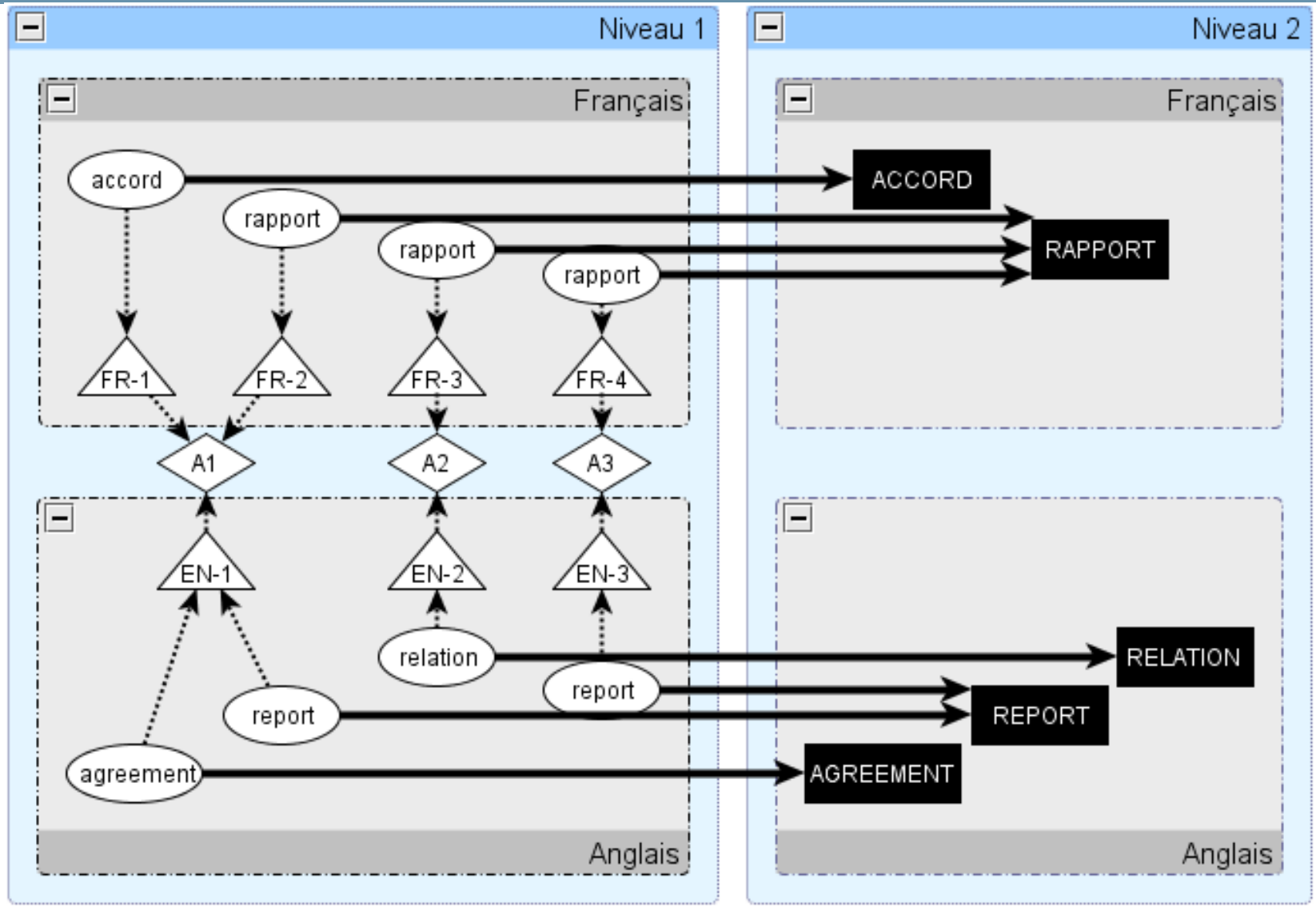


Pondération : graphe du corpus

Un accord environnemental permet de prendre des mesures efficaces et de réduire le nombre de mesures législatives et administratives. J'ai moi-même proposé un amendement au **rapport** qui va dans ce sens	Through an agreement , appropriate measures are taken which mean that legal and administrative measures can be reduced , and I have tabled an amendment to the **report** to this effect .
Nous ne pouvons pas pénaliser ce secteur par **rapport** aux autres.	We cannot penalize this sector in **relation** to the others.
Le rejet par le Parlement de ce **rapport** est scandaleux	The rejection of this **report** by Parliament is shameful.

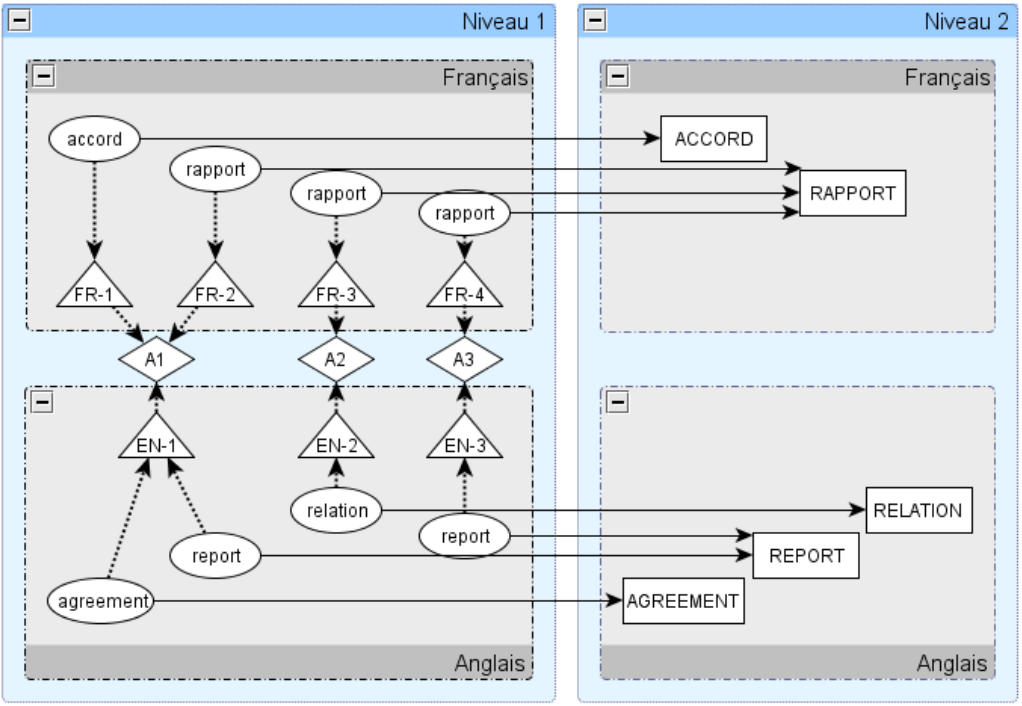


Pondération : émergence des lemmes

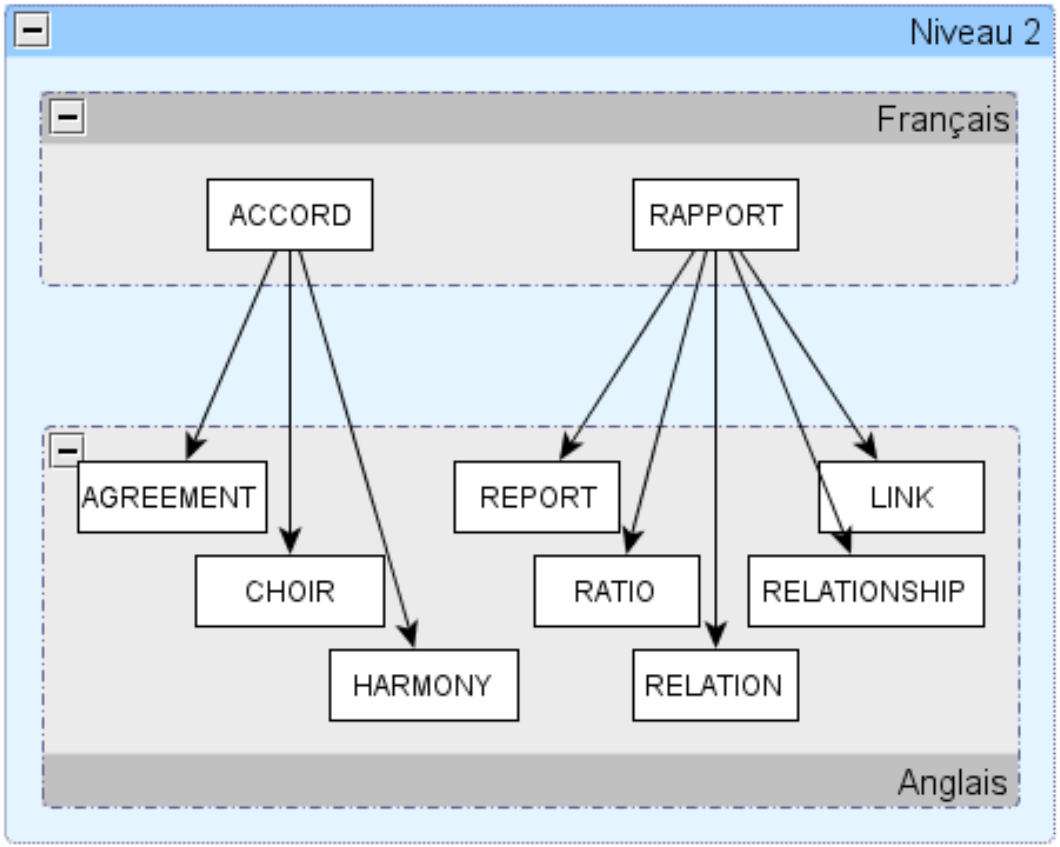


Pondération : union (avec graphe du dictionnaire)

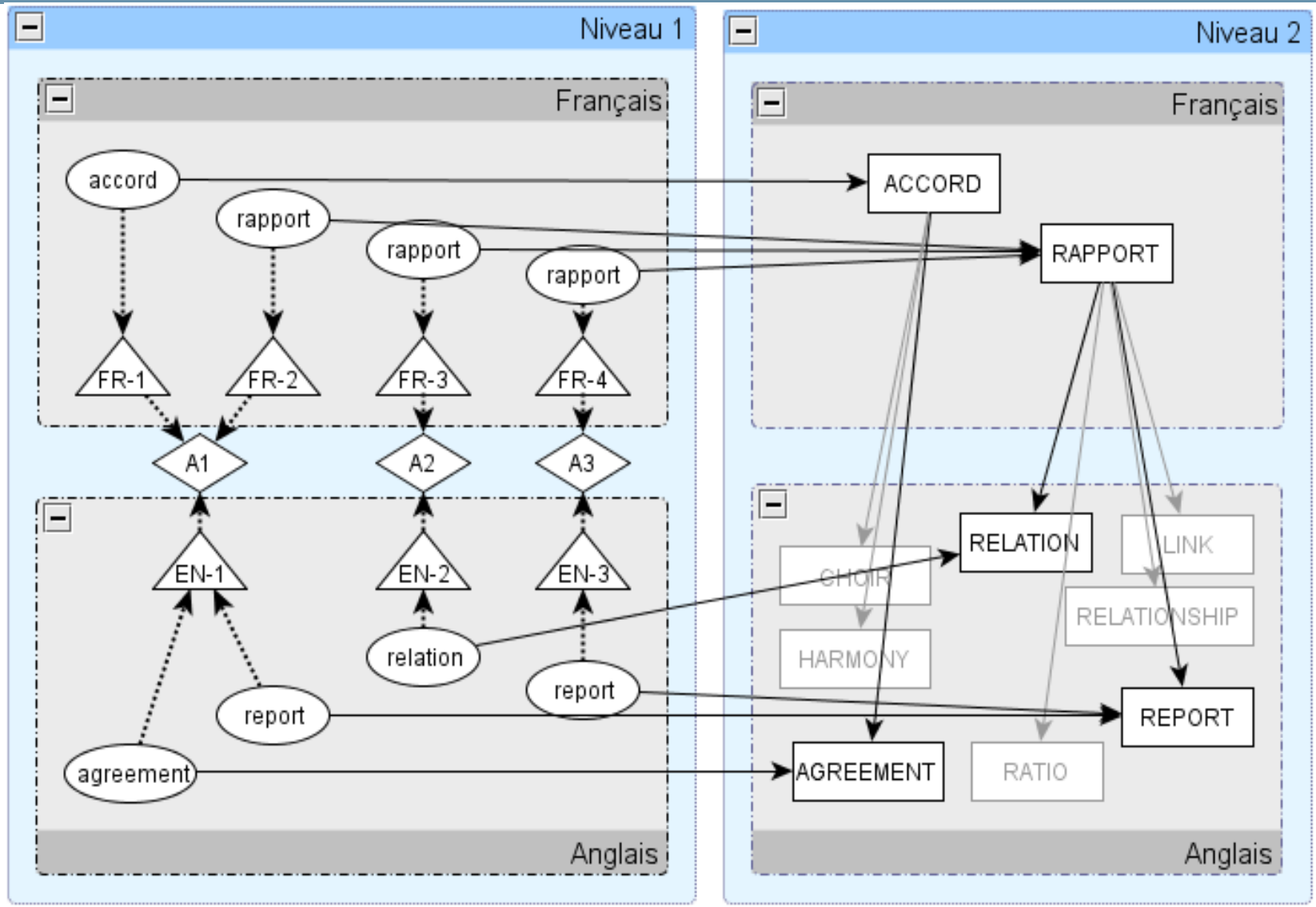
- Graphe du corpus (lemmes)



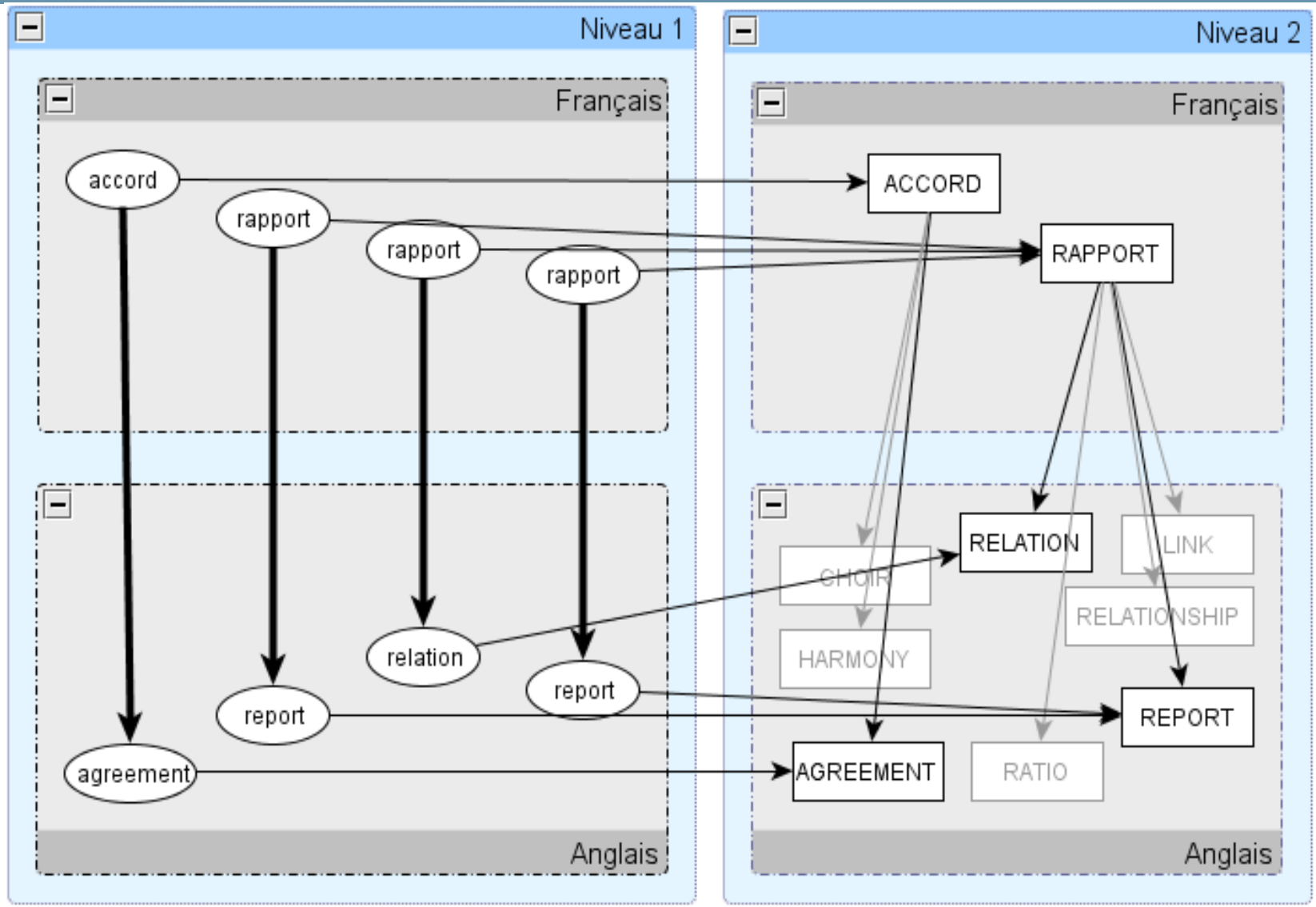
- Graphe du dictionnaire



Pondération : résultat de l'union

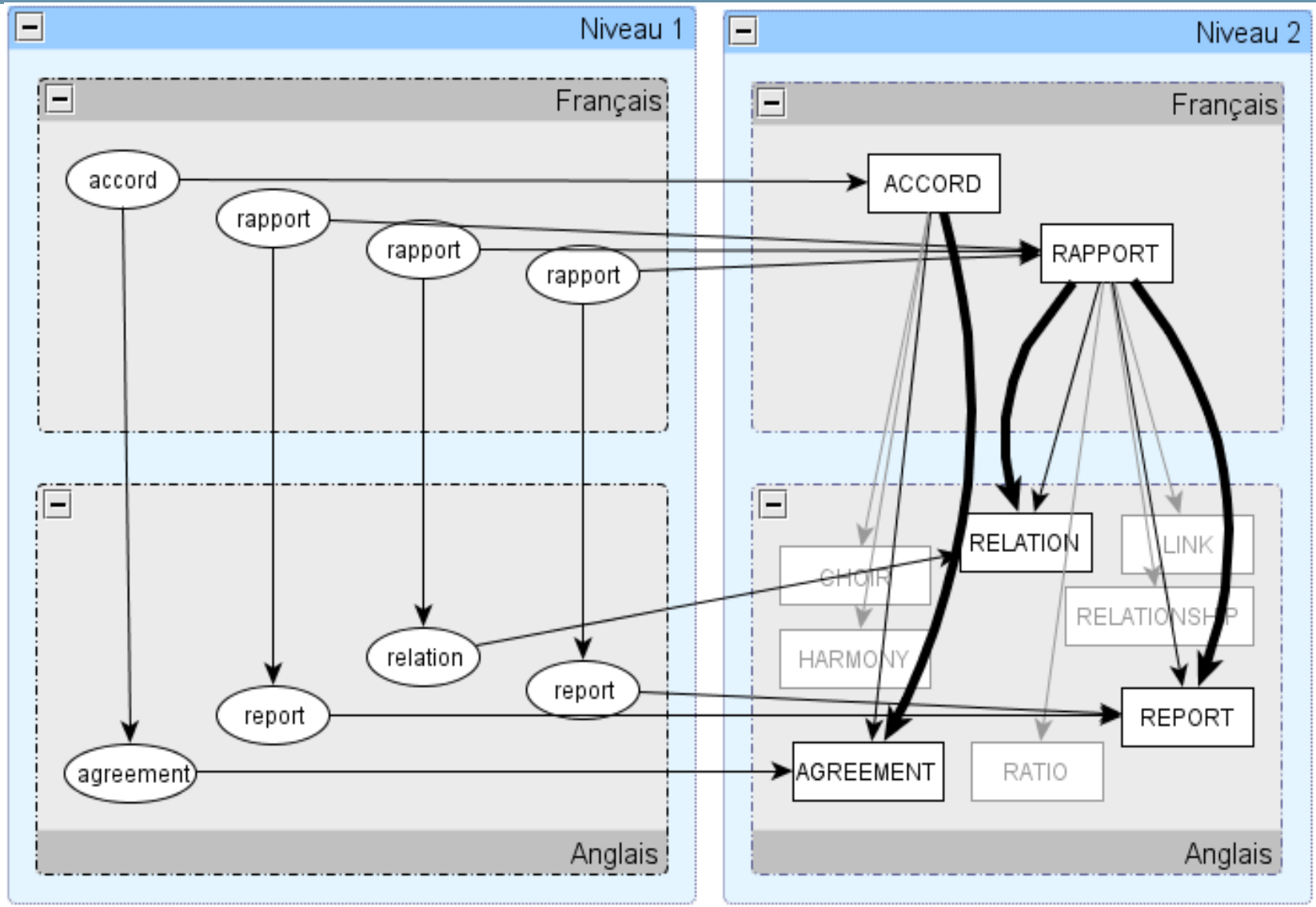


Pondération : liens d'apparition conjointe

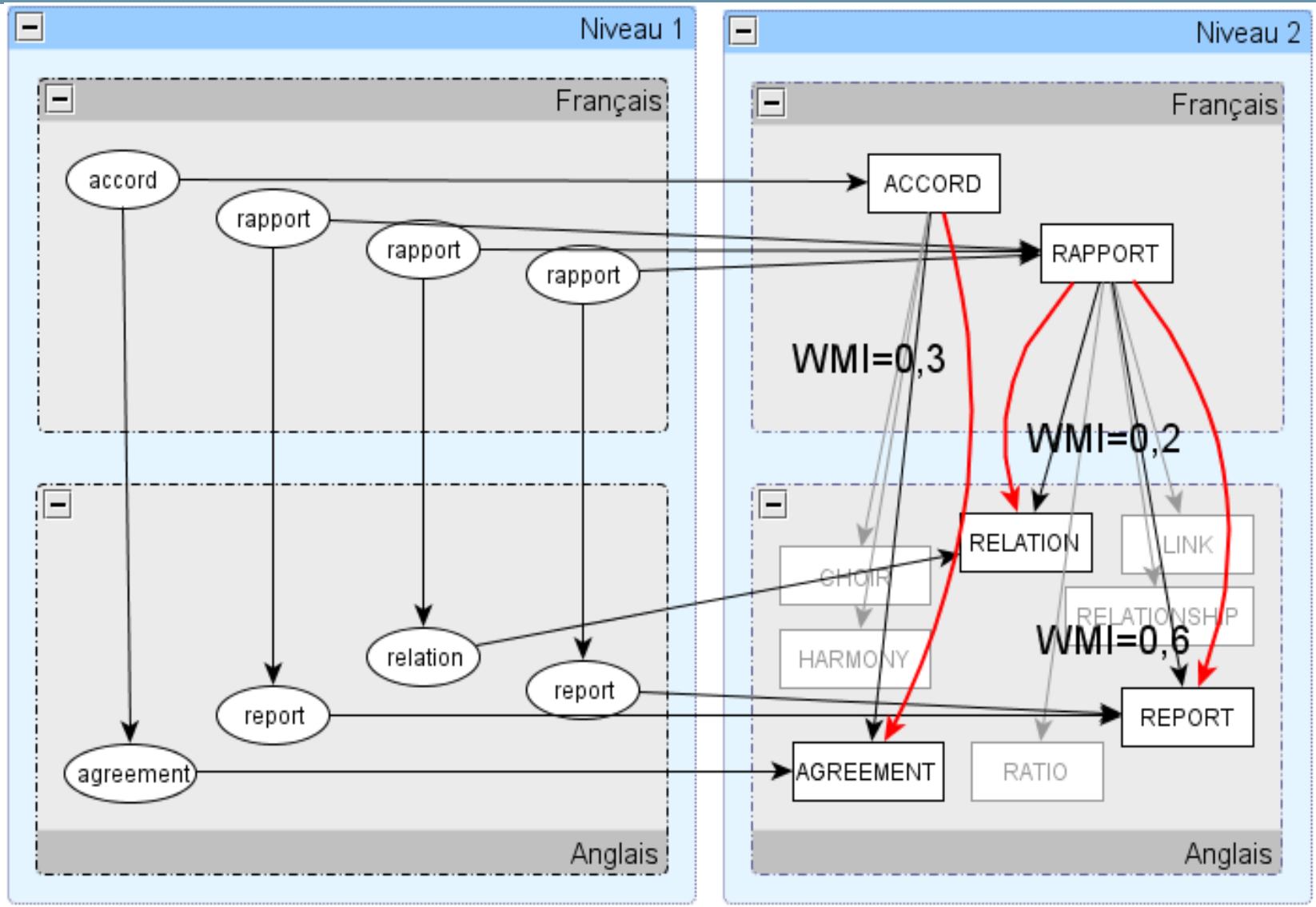


- Opérateur spécifique (conditionnel)

Pondération : émergence des liens de traduction



Pondération : calcul de la pondération



Pondération : expérimentation (1)

- Corpus utilisé : EuroParl
- Dictionnaire : à partir des wordnets
  - Anglais : PWN
  - Français : EWN, Wolf

Wordnet utilisé	Noms	Verbes	Adjectifs		Adverbes	
	Wolf	EWN	Wolf	EWN	Wolf	Wolf
Termes français	39 335	24 178	2 430	8 282	1 959	941
Termes anglais	59 518	34 420	4 025	13 146	3 260	1 365
Couples de traduction	87 103	51 244	6 857	24 155	4 887	2 146

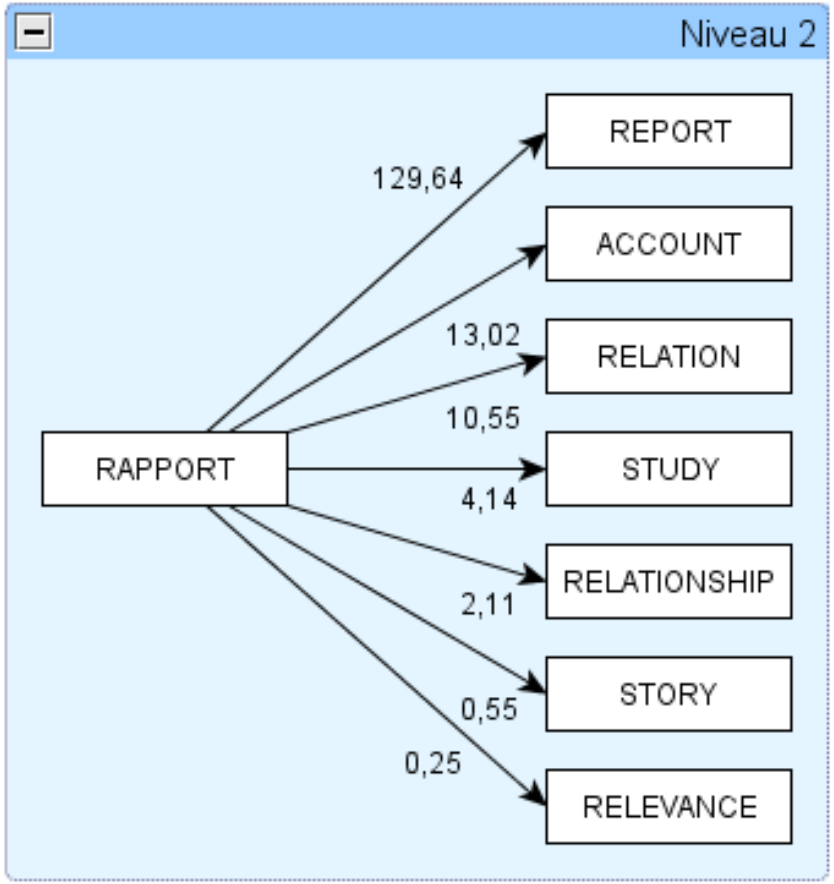
Pondération : expérimentation (2)

- Association : 2 termes apparaissant conjointement dans des entités alignées
  - Mesures d'association : Information mutuelle et WMI

Expérimentations		Niveau1		Niveau 2	
		Nœuds	Arcs	Nœuds	Arcs
Noms	Wolf	1 746 356	45 694	2 478	2 427
	EWN	1 746 356	45 356	2 866	2 735
Verbes	Wolf	1 222 064	317 480	1 068	1 528
	EWN	1 222 064	166 249	1 376	1 846
Adjectifs	Wolf	1 513 787	12 215	685	525
Adverbes	Wolf	1 078 995	40 872	378	385

Pondération : résultats

- Représentation graphique
  - Classement

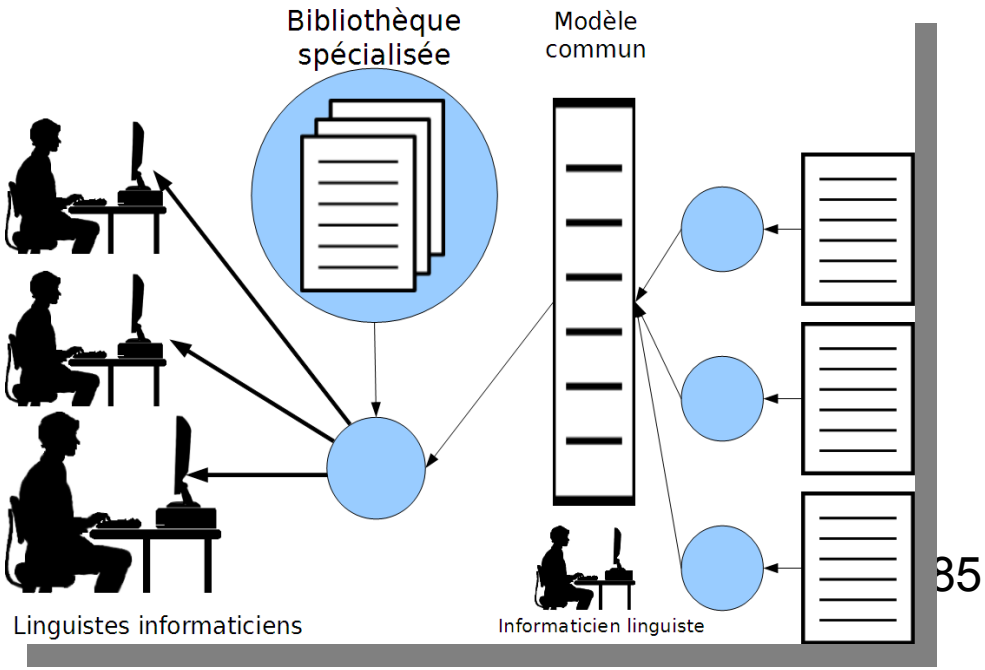


Rapport			
Avec Wolf		**Avec EuroWordNet**	
WMI		**WMI**	
Report	***129,64***	***Report***	***126,43***
Account	***13,02***	***Account***	***11,62***
Relation	10,55	Relation	9,98
Study	4,14	Statement	4,46
Relationship	2,11	Study	3,63
Story	0,55	Connection	2,84
Relevance	0,25	Argument	2,06
		Relationship	1,99
		Link	1,41
		Theme	1,34
		Paper	1,25
		Composition	1,13
		Story	0,55
		Relevance	0,24

Pondération : observations

- Montre l'utilité de notre modèle
- Gère différents types de données
  - Liens entre termes / liens entre occurrences
  - Union de 2 graphes

- 120 lignes de code
  - Appels des fonctions de la bibliothèque
  - +import (300 lignes)



Conclusion : modèle

- Extraction de connaissances lexicales
  - Manque d'outils
- Cahier des charges d'un tel outil :
  - Simplicité
  - Généricité du modèle et des opérations
- Réponse : MuLLinG
  - Modèle de graphe
  - Multiniveau (différentes vues de l'information)
  - Sans contrainte sur le contenu représenté
  - Opérations simples de manipulation
  - Repr. complexe pour matérialiser les relations

Conclusion : outil

- Programmation de tâches linguistiques
  - Plus simple
  - Plus efficace
  - Non dépendante des ressources
- Pour des linguistes informaticiens
- Plus facile de réaliser des expérimentations
  - Amélioration des résultats
- La représentation complexe permet d'envisager de nouvelles expérimentations

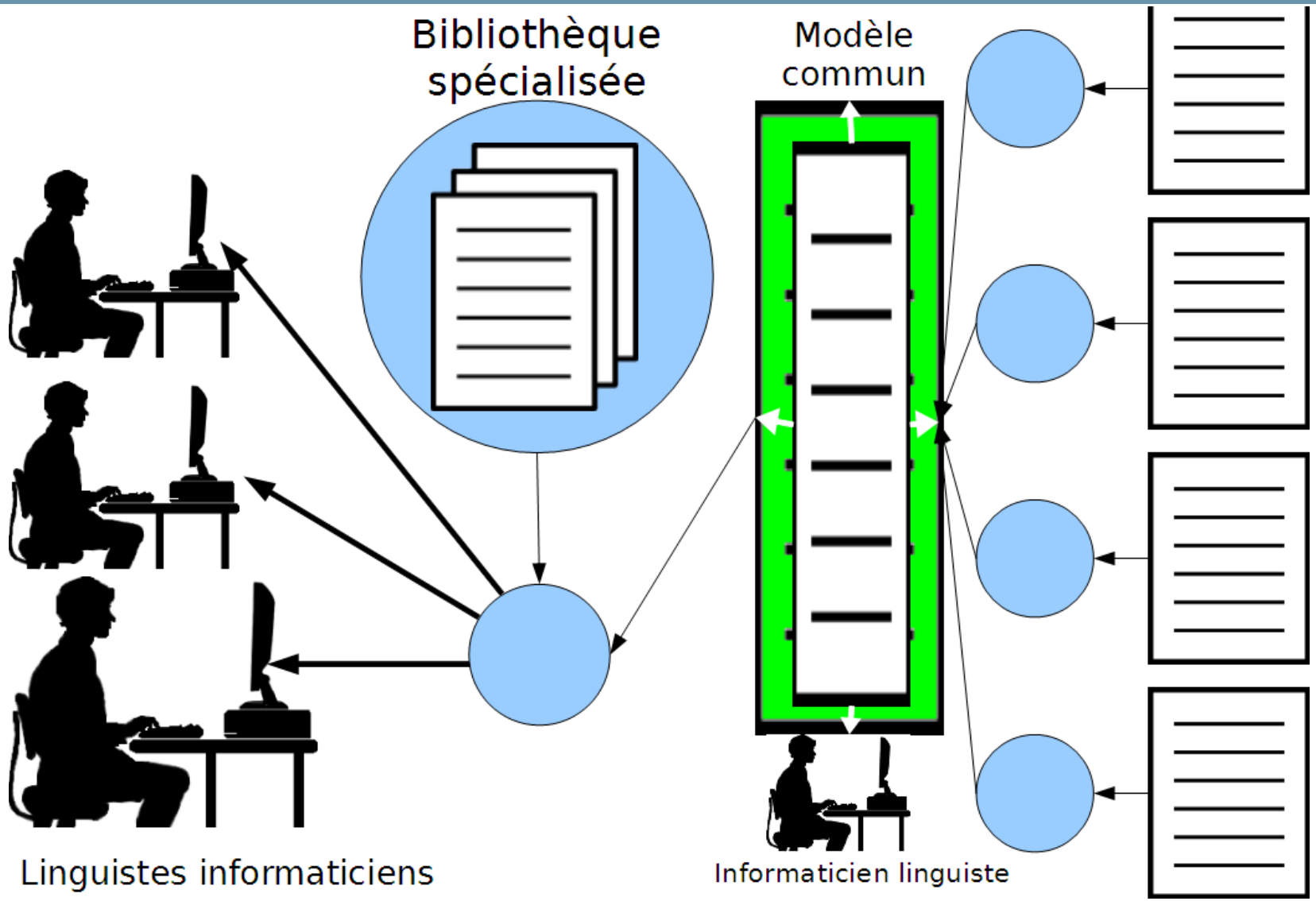
Perspectives : prise en main

- Langage de plus haut niveau
  - Langage de requête
- Interface graphique ?
  - Œil de poisson ?
- Intégration (Gate, LinguaStream, ?)
- Diffusion à d'autres informaticiens

Perspectives : extensions

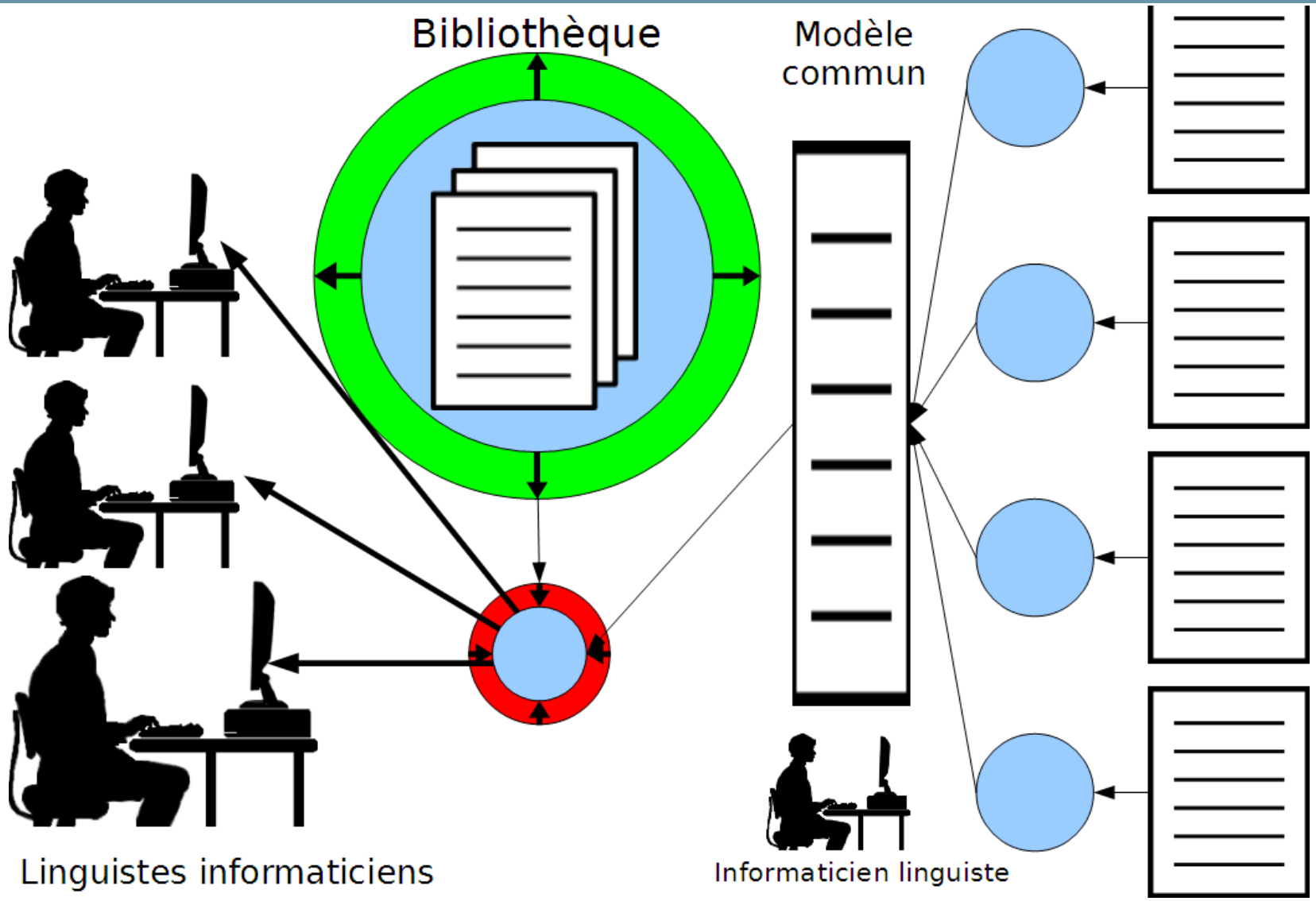
- Faciliter l'import
- Plus d'opérations de base
  - Élargir le champ d'application
- Extraction de connaissances (sous forme de graphes) pas limitée à la linguistique
  - Recherche d'informations
- Évaluation
- Intégration à des outils du web sémantique
  - RDF/SPARQL

Perspectives : efficacité



- Optimisation (gestion de la mémoire)

Perspectives : efficacité



- Implémentation efficace d'opérateurs plus complexes

Merci de votre attention !

Archer Vincent (2007). Using conceptual vectors to get collocations. *Proceedings of MTT 2007, Klagenfurt, Autriche*. pp.57-66

Archer Vincent (2006). Acquisition semi-automatique de collocations à partir de corpus monolingues et multilingues comparables. *Actes de RECITAL 2006, Leuven, Belgique*. pp.

Archer Vincent (2005). *Acquisition semi-automatisée de fonctions lexicales à partir de corpus monolingues et multilingues comparables*. Rapport de M2R Informatique, UJF & INPG, Grenoble. 90 p.