Acquisition et validation de connaissances à partir de sorties syntaxiques

Éric de la Clergerie

INRIA Paris-Rocquencourt
http://alpage.inria.fr

Séminaire TEXTE
Montpellier, 28 Mai 2015

Cheminement

- Développement de FRMG, une grammaire TAG à large couverture
  - s'appuyant sur une méta-grammaire [Candito,Crabbé]
  - initiée en 2004 sur quelques mois pour la campagne EASy
  - largement étendue et améliorée depuis

- Conçue pour valider analyse syntaxique par charte pour les TAGs

2 grandes questions:

1. À quoi sert un analyseur syntaxique ?
   - Extraction d'information (`http://passage.inria.fr/SAPIENS`)
   - Questions-Réponses
   - **Acquisition de connaissances sur corpus**
   - ...

2. Comment continuer à améliorer un analyseur syntaxique ?
   par **injection de connaissances** (attachement prépositionnel)

Outline

Modularité des méta-grammaires

FRMG: issue d'une description modulaire d'une hiérarchie de classes

- exprimant des contraintes sur les noeuds (précédence, dominance, . . .) et leurs traits
- pouvant fournir ou requérir des fonctionnalités (ressource)

Lors de la phase de compilation:

- Des classes neutres sont obtenues par croisement des classes productrices/consommatrices
- Des arbres TAGs (minimaux) sont produits à partir des contraintes des classes neutres

Modularité $\Rightarrow$ facilite l'ajout ou l'enrichissement de phénomènes syntaxiques

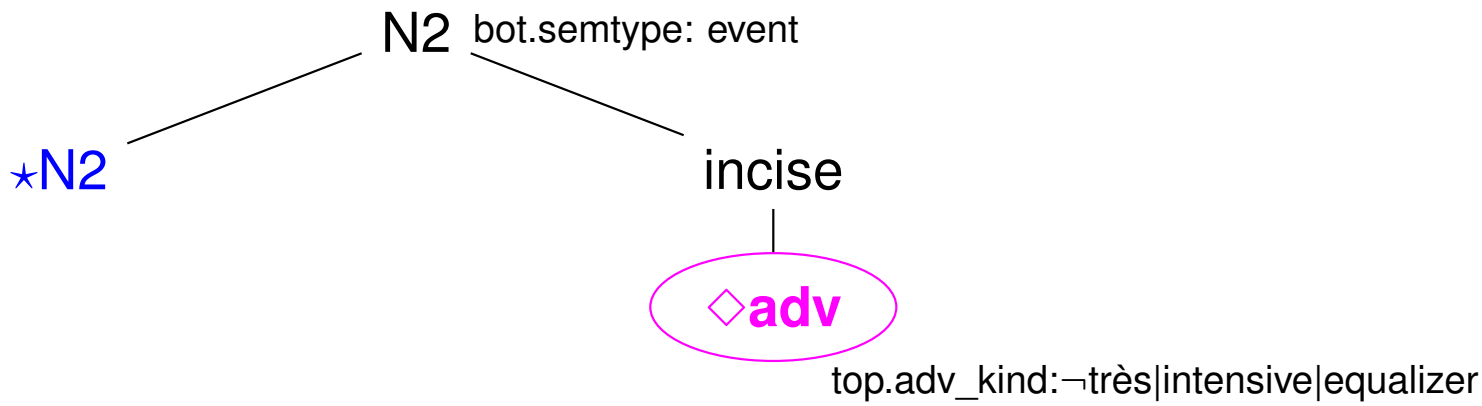
- gestion des arguments verbaux
- *qu'est-ce qu'une pomme ?* : ajout d'une classe, enrichissant les clivées
- gestion des "*adverbiaux*" flottants
  type de flottants $\times$ type d'incise $\times$ positions $\times$ nature modifiés

Exemple: noms évenementiels

Accroche de modifieurs flottants (adverbiaux) après les noms évenementiels:

> *L'annonce, **ce matin**, d'un remaniment a surpris tous les commentateurs*

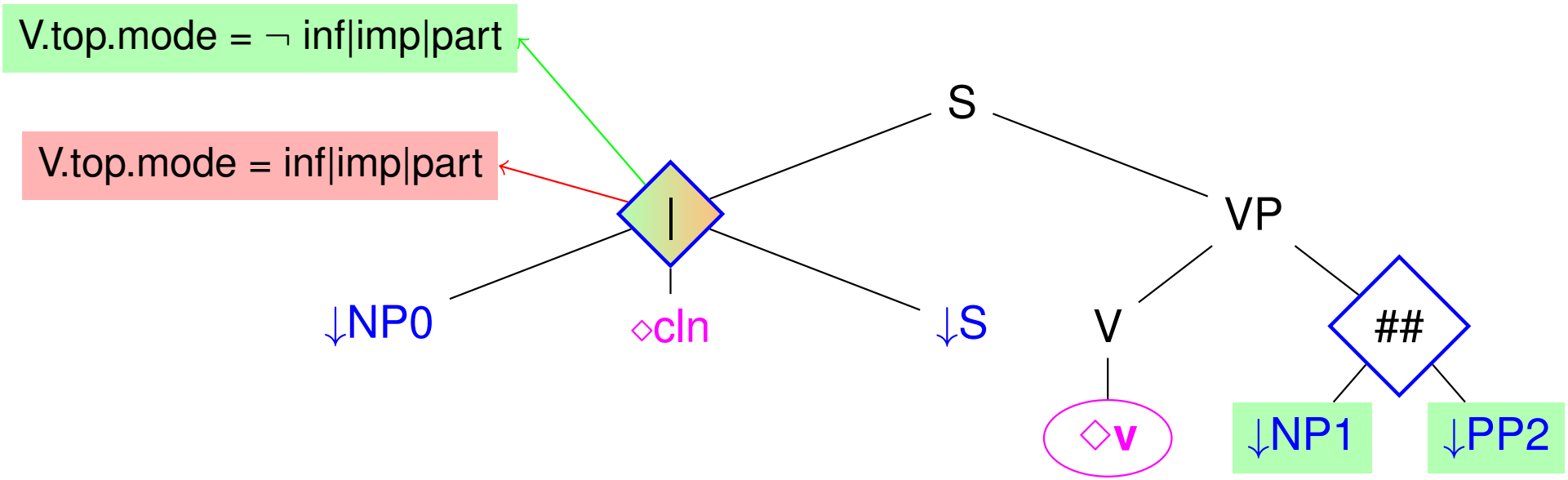
```
class mod_on_N2 {
  + s_modifier;
  - x_modifier;
  - s_adj_pos;
  node(Root).cat = value(N2);
  node(Root).bot.semtype = value(event); %% semantic property
}
```



Note: Comment renseigner les noms événementiels !

Arbres factorisés

Le croisement des classes peut normalement produire beaucoup d'arbres
$\Rightarrow$ utilisation d'opérateurs réguliers pour des arbres factorisés



Défactorisation: $(1_{\text{no subj}} + 3_{\text{subj}}) * (1_{\text{no arg}} + 2_{\text{1 arg}} + 2_{\text{2 args}}) = 20$ arbres

actuellement (2015) : 368 arbres factorisés
expansion FRMG (2007) $\sim$ 200 arbres $\rightsquigarrow \sim$ 2 millions d'arbres !

Analyser avec FRMG

- Utilisation SxPipe (tokenisation, EN) et Lefff (lexique) Sagot
  $\Rightarrow$ treillis de mots (DAG)
- analyse par charte, hybride TAG/TIG, relation coin-gauche, filtrage par lexicalisation, ...
- Analyse complète des phrases
  si échec, essai de corrections sur la phrase (eg, accord sujet-verbe)
  si échec, analyses partielles couvrant la phrase
  $\Rightarrow$ très peu de phrases sans analyse (<1% timeout, 5 à 20% partielles)
- Retourne l'ensemble des analyses (totales ou partielles)
  sous forme de forêts partagées
- Utilisation de clusters de machines pour traiter de gros corpus

Désambiguïser & Convertir

Par défaut, l'analyseur rend une forêt partagée de toutes les dérivations
mais nécessaire de pouvoir sortir la ou les meilleures analyses !

Conversion des arbres de dérivations en arbres de dépendances

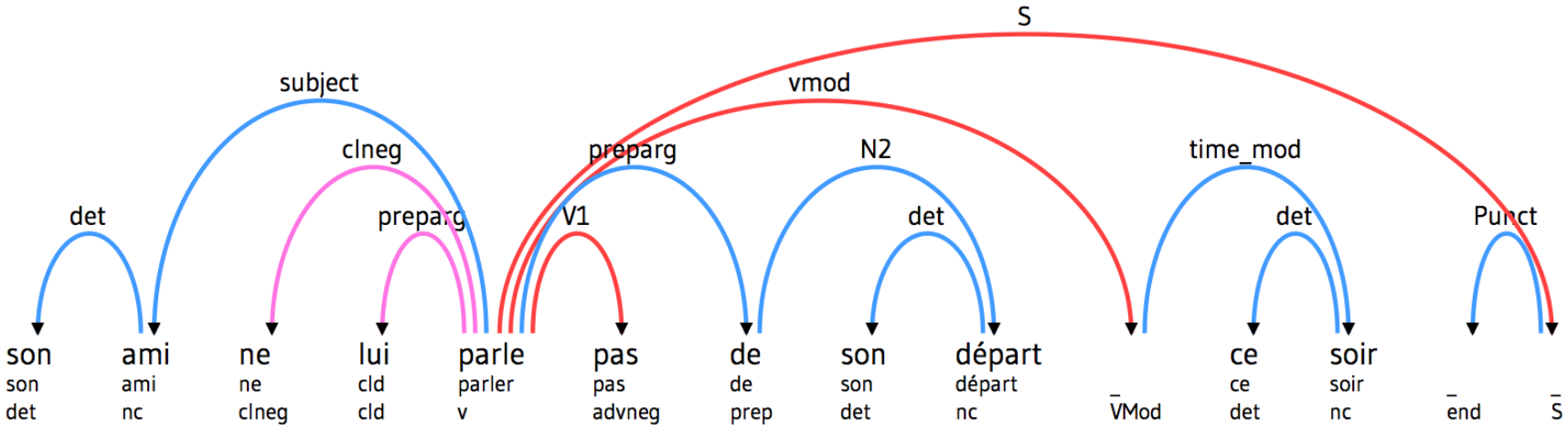
Utilisation de règles de désambiguïsation sur les dépendances pour rechercher la meilleure analyse

- initialement, règles et poids manuellement définis
  exemple: favoriser les arguments verbaux, pénaliser inversion sujet, ...
- maintenant, apprentissage des poids appris sur FrenchTreeBank (FTB)

Conversion vers divers schémas d'annotations:
DepXML, **FTB/CONLL** & **PASSAGE** (+ variantes)

Schéma DepXML

Schéma natif FRMG pour les arbres (et forêts) de dépendances

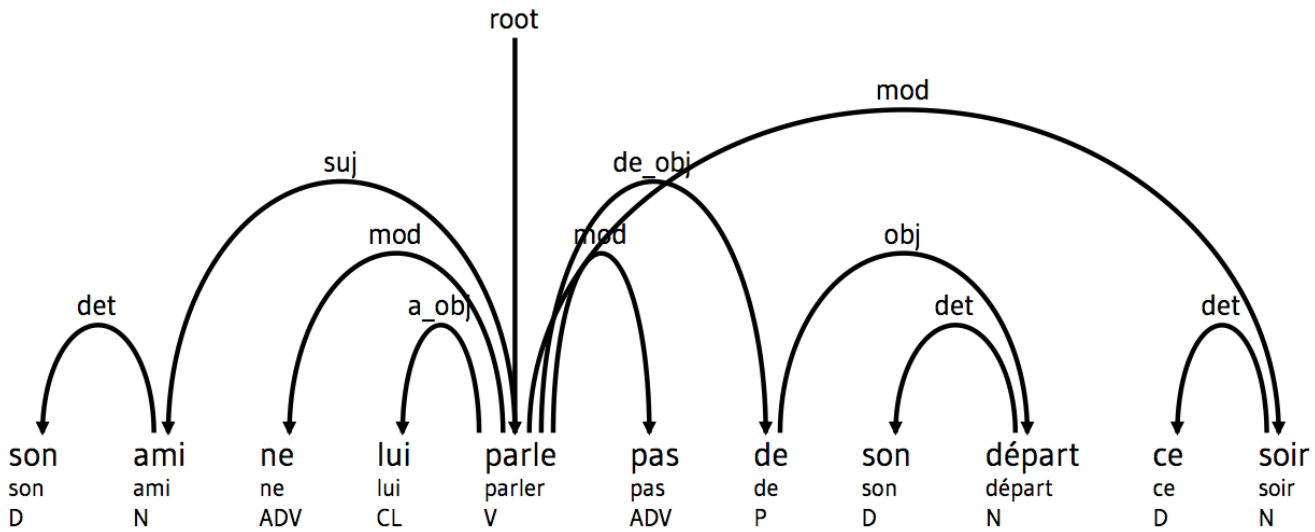


Derrière la vue graphique, un format XML riche en information

- tokens
- noeuds avec formes, lemmes, catégories et arbre TAG ancré
  pseudo-noeuds pour les arbres non lexicalisés
- dépendances entre noeuds (label, type, dérivation)
- constituants “maximaux” (couverts par les arbres) avec traits
- info syntaxiques (hypertag): sous-catégorisation, contrôle, diathèse, . . .

Schéma FTB-UC/CONLL

Schéma d'annotation utilisé pour la version en dépendance du French TreeBank (FTB) [Candito & Seddah]



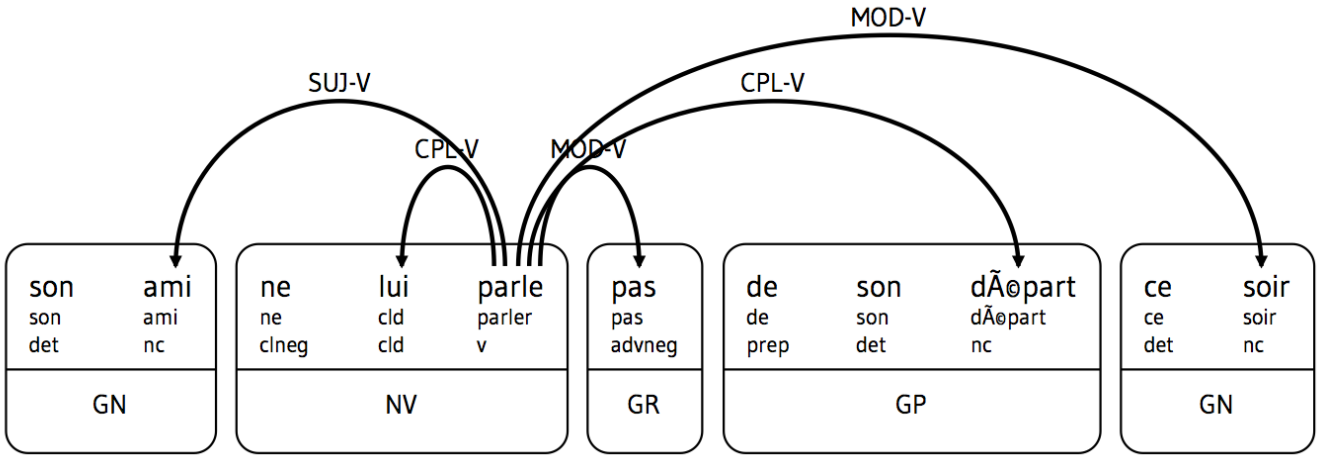
Représentation tabulaire (CONLL)

```
1 son    son     D    DET n=s|s=poss                  2 det
2 ami    ami     N    NC  g=m|n=s|s=c                 5 suj
3 ne     ne      ADV  ADV s=neg                       5 mod
4 lui    lui     CL   CLO n=s|p=3                     5 a_obj
5 parle  parler  V    V   g=m|m=ind|n=s|p=3|t=pst    0 root
...
```

Schéma Passage

Un schéma issu des campagnes d'évaluation EASy/Passage avec

- 6 types of *chunks* (GN, NV, GA, GR, GP, PV)
- 14 types de dépendances (SUJ-V, COD-V, CPL-V, ATB-SO, MOD-N, . . .)
- format XML



Moins riche en information que DepXML mais "standard"
$\Rightarrow$ utilisé pour les xp d'acquisition

Quelques résultats

Sur FTB et Sequoia (FTB/CONLL, LAS sans ponctuation)
et EASy (PASSAGE, F-mesure sur les relations [90.5% sur les chunks])

système	dev	test	sequoia	easy
FRMG init	80.85	82.08	81.13	65.92
FRMG +tuning	86.76	87.95	86.41	69.89
BKY	86.50	86.80	–	–
MALT	86.90	87.30	–	–
MST	87.50	88.20	–	–
TALISMANE (b=20)	88.10	88.50	–	–
Le Roux (2012)	–	89.20	–	–
DYALOG-SR (b=8)	88.17	89.01	85.02	–
DYALOG-SR+FRMG (b=10)	89.02	90.25	87.14	–

Constat:

- Performances FRMG similaires à celles des parseurs statistiques
- plus stable sur du hors domaine (SEQUOIA)
- meilleur sur phénomènes profonds
  (Ribeyre, article RECITAL: contrôle et ellipses)

Outline

Du langage au sens (et vice-versa)

“Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!”

Il était grilheure; les slictueux toves
Gyraient sur l’alloinde et vriblaient:
Tout flivoreux allaient les borogoves;
Les verchons fourgus bourniflaient.

Paul s’est cassé la **binti**.
Sa fracture à la **binti** a été correctement réduite.
Il a des douleurs dans la **binti**.

un tremblement de terre de forte magnitude
il observe un homme avec un télescope

Motivations

Point de départ: des expérimentation menées dans le cadre du projet SCRIBO

- Traitement syntaxique de gros corpus (avec FRMG)
- Exploitation des sorties syntaxiques
- Acquisition de connaissances lexico-sémantiques, dont terminologie, réseau de mots, . . .
  - ré-injection pour la désambiguisation syntaxique
    tremblement de terre de forte magnitude
  - aide à la construction de glossaires, thésaurus, ontologies lexicalisées (domaines spécialisés)

Obtention de grosses ressources (entre 1K à 100K *entrées*)
$\Rightarrow$ problématiques de visualisation, évaluation, et validation

Mise en place d'une interface de visualisation et validation collaborative

Quelques corpus

Un large corpus "*général*" hétérogène (CPL+AFP)

Corpus	**#Mphrases**	**#Mmots**	**Description**
Wikipedia	18.0	178.9	504K pages encyclopédie
Wikisource	4.4	64.0	12.8K textes littéraires
EstRepublicain	10.5	144.9	journalistique
JRC	3.5	66.5	directives européennes
EuroParl	1.6	41.5	débats parlementaires
Total **CPL**	38.0	495.8	
AFP	14.0	248.3	400K dépêches
Total **ALL**	52.0	744.2	

Mais aussi des corpus spécialisés plus petits (certains en Juridique)

Corpus	**#Mphrases**	**#Mmots**
fiscal	7.2	145.2
social	6.8	127.5
civil	2.6	40.9
affaires	7.2	133.8

Et de nombreux autres: botanique (1Mmots), automobile, récits de voyage, . . .

Méthode générale

Phase0 Analyser les corpus, sur cluster

Phase1: Collecter et compter

- Algorithme MapReduce (popularisé par Google et HADOOP)
  version distribuée de `grep|sort|uniq -c|sort -nr`
- collecter des éléments d'information, étant donné un ou des motifs
- trier les éléments
- et les compter

Phase2: utiliser les décomptes pour prendre des décisions
clustering, classification, ranking, . . .

(**Phase3**): Valider !

(**Phase4**): Injecter dans les phases 0, 1, ou 2

Extraction Terminologique

Un ensemble plus ou moins standard de critères guidant l'extraction de termes

- structure (multi-mots) : les chunks Passage sont bien adaptés
  `(GN) (GR*GA|GP|PV|NV)+`
  base: Chunks nominaux GN (et GP/Prep), avec modifieurs syntaxiques
- forte fréquence
  note: pas (encore) d'information de contraste pour des domaines spécialisés
- forte cohésion interne (information mutuelle)
- (**nouveau**) autonomie: un terme doit apparaître nu
  i.e., sans modifieurs, dans des fonctions sujets ou objets, ...
- (**nouveau**) diversité: un terme apparaît dans des contextes diversifiés (évite collocations, phrases dupliquées, ...)
- variants : un peu de variation (mais pas trop) sur même lemmes
  élimination de certaines entités nommées

$\rightsquigarrow$ plus de 100k termes potentiels (ordonnés) extraits des dépêches AFP
pas de seuil (favorise rappel)
bruit: entités nommées (organisations, événements), VP mal coupés, ...

Quelques exemples

Diversité des motifs (et existence de long termes)

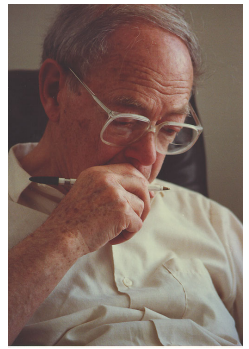
dioxyde de carbone *carbon dioxid*	[dioxyde/nc]$_{GN}$ [de/prep carbone/nc]$_{GP}$
hockey sur glace *ice hockey*	[hockey/nc]$_{GN}$ [sur/prep glace/nc]$_{GP}$
téléphone portable *mobile phone*	[téléphone/nc]$_{GN}$ [portable/adj]$_{GA}$
lait écrémé *skimmed milk*	[lait/nc]$_{GN}$ [écrémer/v]$_{NV}$
permis de conduire *driving license*	[permis/nc]$_{GN}$ [de/prep conduire/v]$_{PV}$
procréation médicalement assistée *medically assisted procreation*	[procréation/nc]$_{GN}$ [médicalement/adv]$_{GR}$ [assisté/adj]$_{GA}$
implant chirurgical non actif *non active chirurgical implant*	[implant/nc]$_{GN}$ [chirurgical/adj]$_{GA}$ [non/adv]$_{GR}$ [actif/adj]$_{GA}$

Limites: *fils et filles de*

Rapprocher les mots: l'approche distributionnelle

Meanings of words are (largely) determined by their distributional patterns (Harris 1968)

You shall know a word by the company it keeps (Firth 1957)



1. attacher à chaque mot un vecteur (pondéré) de contextes, ici **syntaxiques**
2. utiliser ces vecteurs pour mesurer la similarité entre paires de mots
3. exploiter les similarités pour organiser/regrouper les mots

Nombreuses variantes sur ces 3 points (Lin, Pantel, Pedersen, Bourrigault, . . .)
Mais souvent: boîte noire, pas d'explication, classes *dures* (pas de polysémie), . . .
$\Rightarrow$ envie d'explorer une approche flexible

Collationner et compter les dépendances

<gouverneur>	<rel>	<gouverné>	<freq>
chaise_nc	et	table_nc	235
asseoir_v	sur	chaise_nc	227
chaise_nc	modifieur	long_adj	168
chaise_nc	de=	poste_nc	115
tomber_v	sur	chaise_nc	103
chaise_nc	modifieur	musical_adj	102
se_asseoir_v	sur	chaise_nc	93
prendre_v	cod	chaise_nc	87
chaise_nc	modifieur	électrique_adj	82
chaise_nc	modifieur	vide_adj	80
chaise_nc	à=	porteur_nc	80
dossier_nc	de	chaise_nc	78
avoir_v	cod	chaise_nc	71
table_nc	et	chaise_nc	62
chaise_nc	de=	paille_nc	56

Abstraire les dépendances

À partir des dépendances Passage, avec un peu de correction et d'abstraction:

- rectification des passifs
- ajout de *se* pour les verbes pronominaux
- relation directe entre un sujet et un attribut
  (*pomme rouge* dans *la pomme est rouge*)
- abstraction du verbe dans les arguments sententiels
  pouvoir cod sentence $\rightsquigarrow$ *pouvoir cod manger*
- distribution entre éléments coordonnées
- ajout des attachements ambiguës

```
terre_nc          de=     magnitude_nc    344
tremblement_nc    de=*    magnitude_nc    357
```

- injection des termes candidats

```
qualité_nc de= président_du_conseil  189
tremblement_de_terre  de=*    magnitude_nc
```

Des dépendances aux contextes

Une dépendance (*se_asseoir*, *sur*, *chaise*) fournit l'association

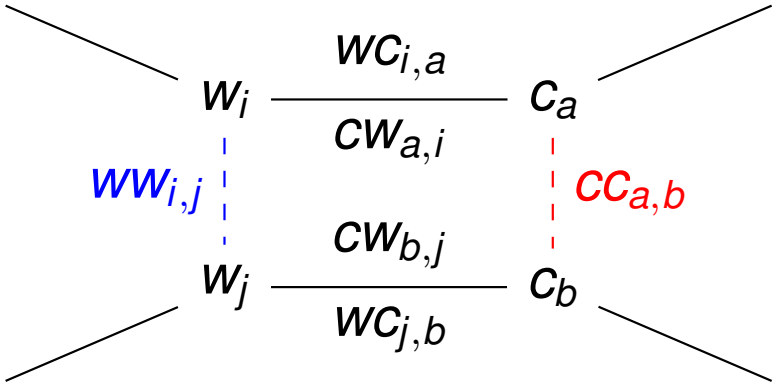
- d'un contexte syntaxique <*se_asseoir sur* •> avec le mot *chaise*
- et, symétriquement, <• *sur chaise*> avec *se_asseoir*

Corpus	#dep (millions)	#formes distinctes (milliers)	# contextes distincts (millions)
CPL	170	1149	4
AFP	93	378	2
Total **ALL**	263	1366	5

Algorithme

Inspiré de Markov clustering [MCL, van Dongen] cherchant, dans un graphe,

- à favoriser les fortes densités de chemins courts
- à pénaliser les chemins longs



$$ww_{i,j} = \frac{1}{Z_i}\left(\sum_{a,b}(wc_{i,a})(cc_{a,b})(wc_{j,b})\right)^{\alpha}$$

$$cc_{a,b} = \frac{1}{Z_a}\left(\sum_{i,j}(cw_{a,i})(ww_{i,j})(cw_{b,j})\right)^{\alpha}$$

avec inflation $\alpha > 1$ (défaut: 2) et normalisation $\frac{1}{Z}$
$\Rightarrow$ renforce les forts coefficients, réduit les faibles !

Formulation matricielle

Les équations précédentes s'expriment sous forme matricielle compacte:

$$\begin{cases} W = \Gamma_\alpha(F^t C F) \\ C = \Gamma_\alpha(G^t W G) \end{cases}$$

avec l'opérateur d'inflation et normalisation Γ_α

où:

- $W = (ww_{i,j})$ et $C = (cc_{a,b})$ sont les matrices de similarités à calculer
- $F = (wc_{i,a})$ et $G = (cw_{a,i})$ matrices de paramètres
  - $wc_{i,a}$: importance (pondérée) du contexte c_a pour le mot w_i
  - $cw_{a,i}$: importance (pondérée) du mot w_i pour le contexte c_a

Formulation récursive, attaquable par un algorithme itératif de point-fixe en démarrant d'une matrice initiale $W^{(0)}$

Pondération des contextes

Variation de tf.idf, avec un ratio de normalisation:

$$wc_{i,a} = \frac{ln(u_{ai}) * \eta_a}{Z_i} \text{ with } \eta_a = ln\left(\frac{\#\text{mots distincts}}{\sqrt{|\{w_j | u_{aj} > 0\}|}}\right) \quad (1)$$

$$cw_{a,i} = \frac{ln(u_{ai}) * \eta_i}{Z_a} \text{ with } \eta_i = ln\left(\frac{\#\text{contextes distincts}}{\sqrt{|\{c_b | u_{bi} > 0\}|}}\right) \quad (2)$$

où u_{ai} nombre de co-occurrences de c_a avec w_i

Transferts entre mots et contextes

Intuition: si w_1 et w_2 sont similaires, alors le sont aussi les contextes:

- <w_1 r •> et <w_2 r •>
- <• r w_1> et <• r w_2>

e.g. **chaise** $\sim$ **tabouret** $\rightsquigarrow$ <• *sur chaise*> $\sim$ <• *sur tabouret*>
Idem des contextes vers les mots

Formalisation par des matrices de transfert:

$$\begin{cases} W = \Gamma_\alpha(F^t C F + \tau(C)) \\ C = \Gamma_\alpha(G^t W G + \rho(W)) \end{cases} \tag{3}$$

avec

$$\tau(C) = \sum_r \beta_r T_r^t C T_r \text{ and } \rho(W) = \sum_r \beta_r T_r W T_r^t \tag{4}$$

et $T_r = (t_{r,ia})_{ia}$ avec $t_{r,ia} = 1$ si $c_a = r.w_i$ et 0 autrement

Par défaut, $\beta = 0.2$

Bonus et Malus

Possibilité d'injecter à chaque itération une matrice L de bonus/malus:

$$W = \Gamma_1((I + L) \circ \Gamma_\alpha(F^t C F + \tau(C))) \quad \text{avec } (A \circ B)_{ij} = A_{ij}.B_{ij}$$

Utilisation:

- forcer l'auto-similarité d'un mot avec lui-même
- Coordination (en fonction du nombre d'occurrences)
- Distance d'édition: des mots proches (si assez longs) tendent à être proches sémantiquement
- différences de productivités et fréquences
  $\Rightarrow$ émergence d'une hiérarchie sur les mots avec des *hubs*
- Random indexing : distance cosinus entre vecteurs de taille 1000
  projection des vecteurs de contextes
  élargissement *qualitatif* des vecteurs de contextes
- similarités germes, par exemple venant de WordNet
  (peu testé pour l'instant)

Corrélation sur les rangs

On peut estimer la contribution d'un contexte c_a pour grouper w_i et w_j:

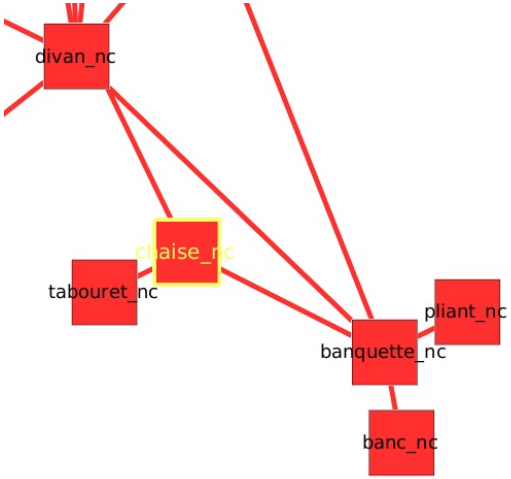
$$\langle ww_{i,j}, c_a \rangle = \sum_b (wc_{i,a})(cc_{a,b})(wc_{j,b})$$

$\Rightarrow$ permet de mesurer la corrélation entre les rangs des contextes groupant w_i et w_j, avec ceux pour w_i (et pour w_j):

$$\Delta_{ij} = \sum_{c_a, wc_{i,a}>0} \left(1 + \left|1 - \frac{\text{rank}_{aij}}{\text{rank}_{ai}}\right|\right)^{-1} \left(1 + \left|1 - \frac{\langle ww_{i,j}, c_a \rangle}{wc_{i,a}}\right|\right)^{-1} \quad (5)$$

À quoi sert une chaise ?

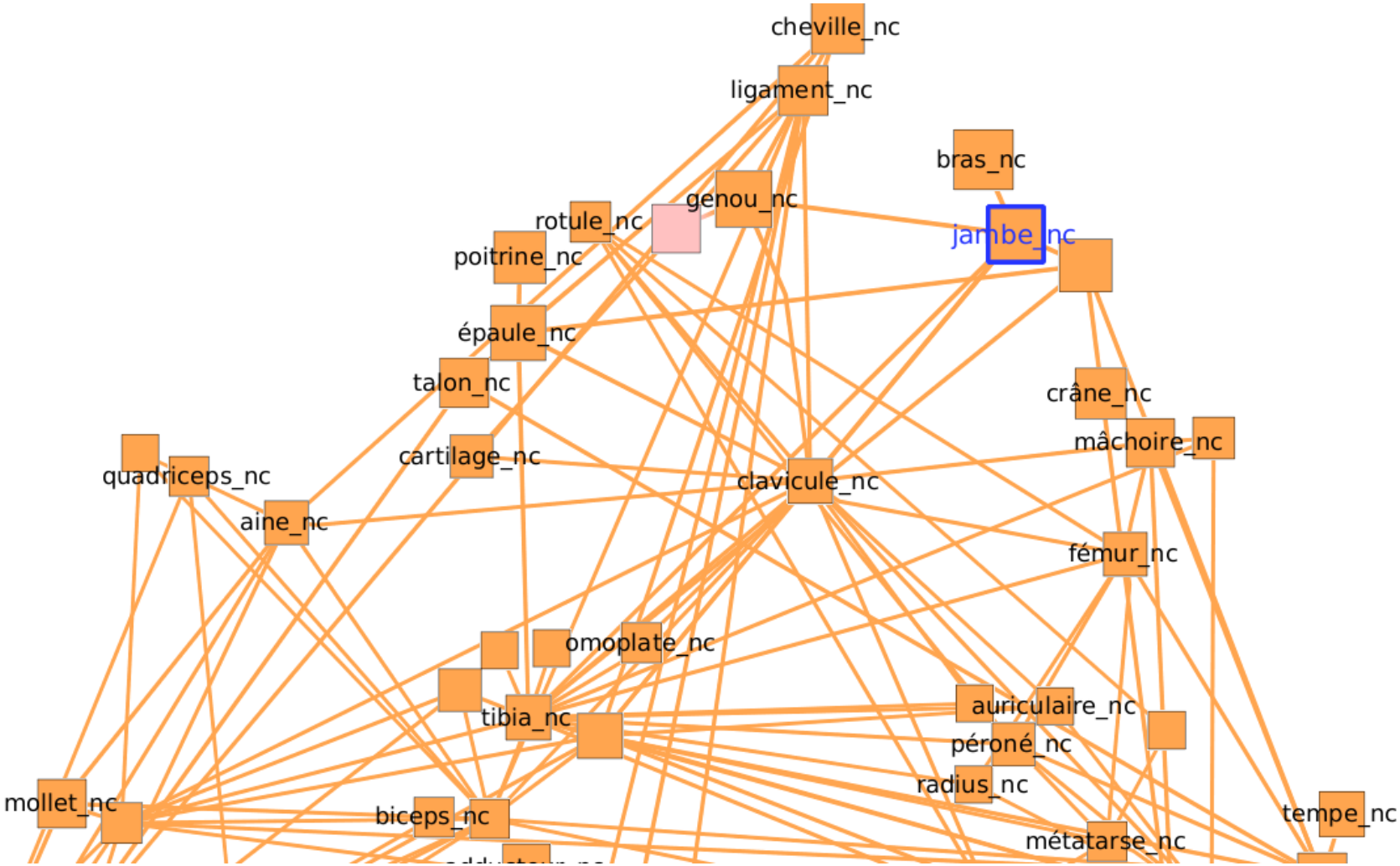
Les regroupements sont motivés par des contextes syntaxiques $\langle ww_{i,j}, c_b \rangle = \Sigma_a(wc_{i,a})(cc_{a,b})(wc_{j,b})$



	chaise divan	chaise tabouret	banquette divan	banquette canapé	banquette chaise
se asseoir sur [•]	●	●	●	●	●
asseoir sur [•]	●	●	●	●	●
allonger sur [•]	●		●	●	
dormir sur [•]	●		●	●	●
tomber sur [•]	●		●	●	●
monter sur [•]		●			●
place sur [•]					
grimper sur [•]		●			●
installer sur [•]		●			●
poser sur [•]		●			●

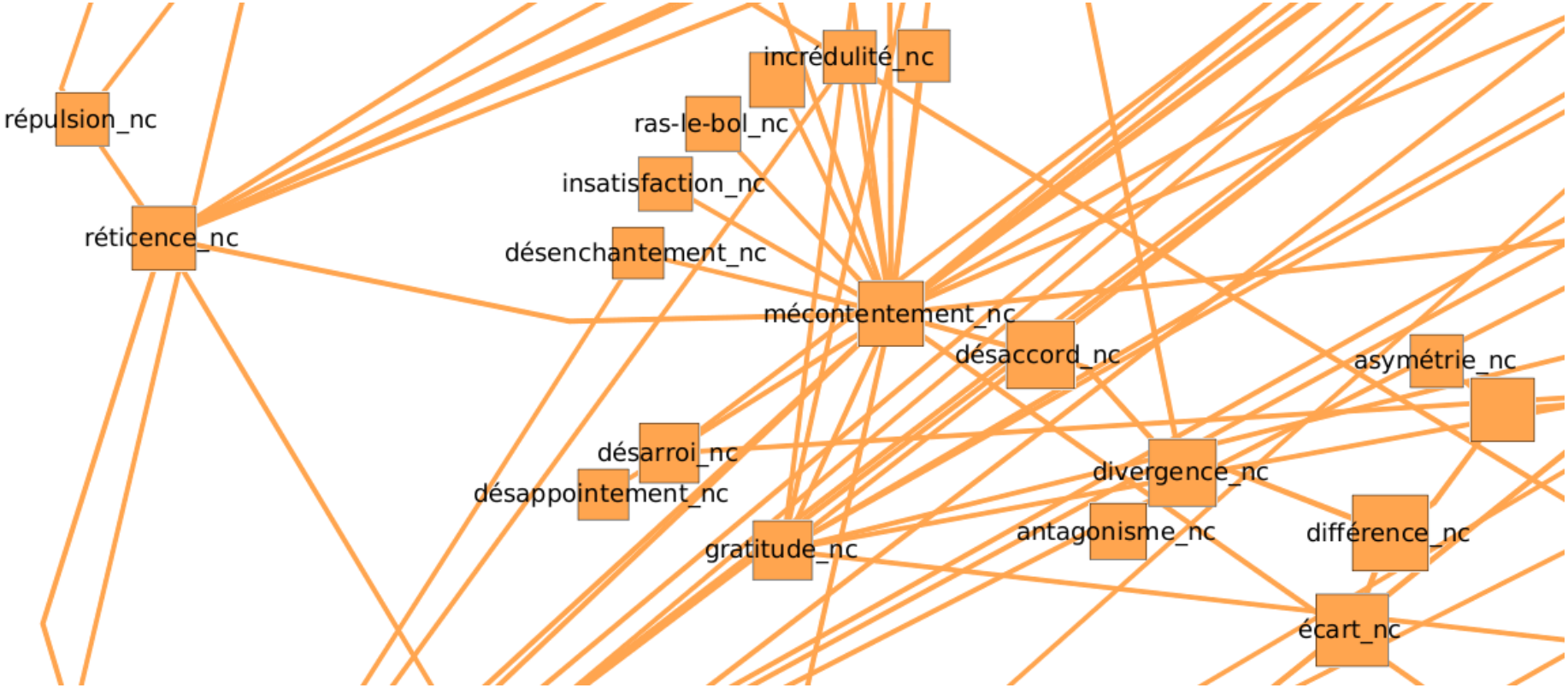
Visualisation: que d'os, que d'os !

Graphe d'environ 40K connections
Visualisation avec TULIP (`http://tulip.labri.fr/`), layout *BubbleTree*

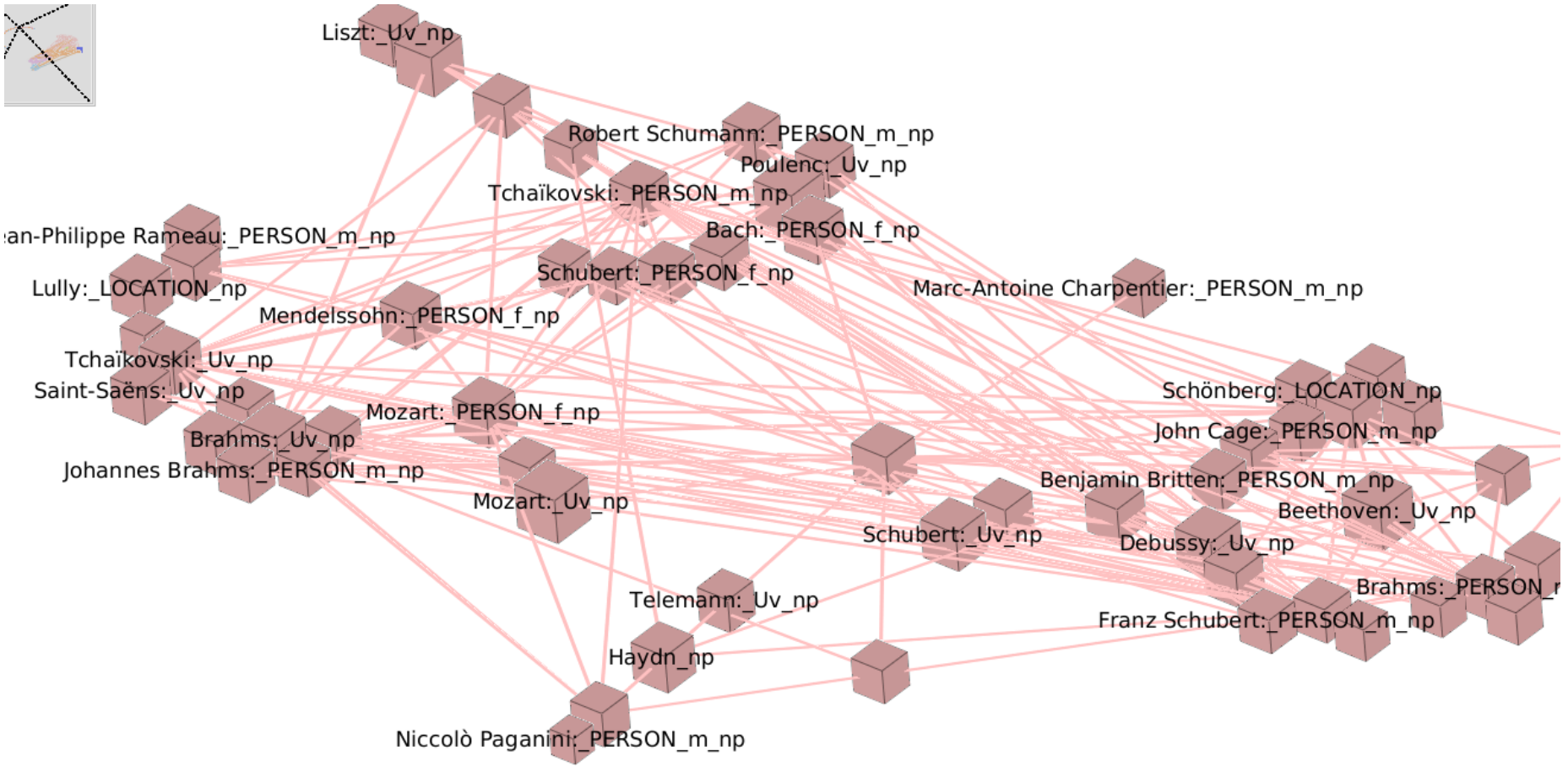


d'autres fragments sur `http://alpage.inria.fr/~clerger/wnet/wnet.html`

Visualisation: sentiments négatifs



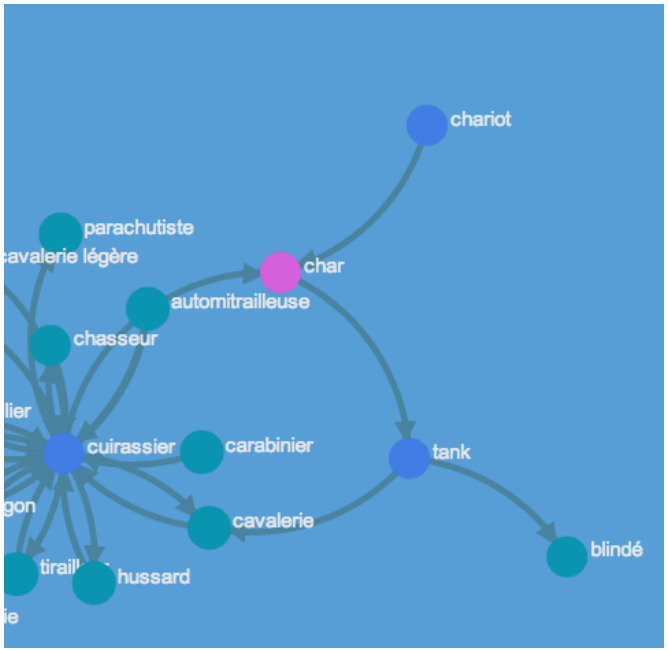
Visualisation: compositeurs



Structures topologiques

Vue à gros grain déjà utile pour repérer certaines structures topologiques:

- **buissons** fortement connectés: très proches de classes sémantiques
- **filaments**: glissement de sens
- **étoiles**: un centre avec de nombreux satellites
  parfois pertinents, souvent mauvais !
- des mots polysémiques à la jonction de buissons



char et **chariot**
<• *modifieur atteler*>, <*promenade en* •>

char et **tank**
<• *de combat*>, <*régiment de* •>

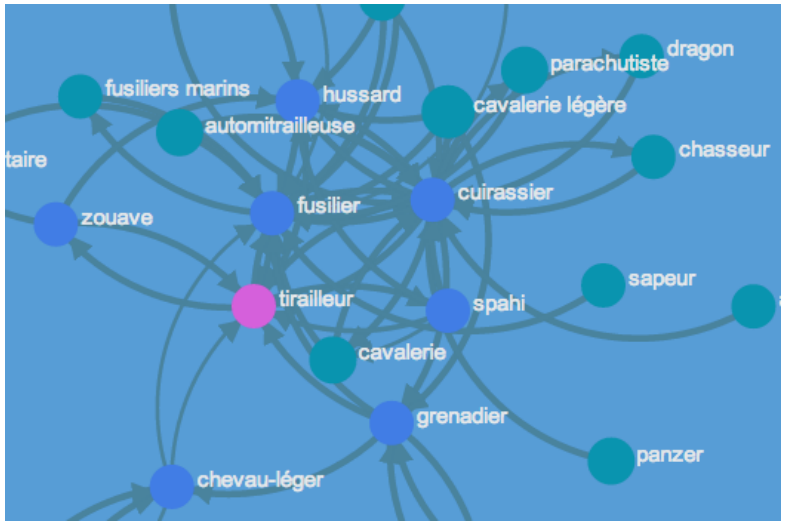
Quelques classes topologiques

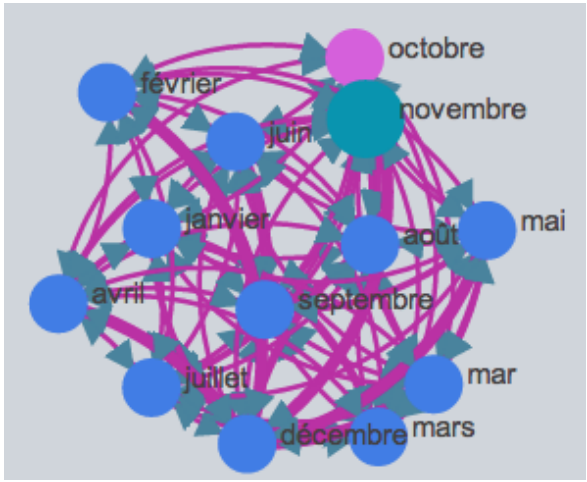
Les *buissons* permettent l'extraction d'environ 4000 classes (ALL)

<79> *sulky malinois fox-terrier setter cocker colley chiot fox labrador ratier griffon caniche teckel épagneul*

<80> *arrière-garde canonnier cavalerie carabinier tirailleur hussard panzer voltigeur blindé grenadier cuirassier avant-garde zouave lancier*

<83> *pneumonie paludisme diphtérie pneumopathie variole dysenterie malaria botulisme poliomyélite septicémie varicelle polio rougeole méningite*





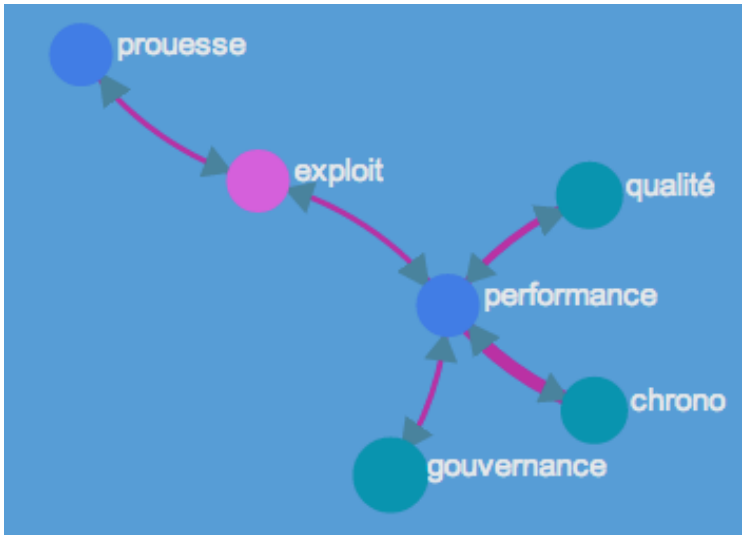
Évaluation

Plusieurs types d'évaluation, dont une par des tests TOEFL construits à partir de wordnets du français (FrenchWordNet, Wolf, JAWS)

Génération aléatoire de tests à partir des synsets

toutefois	*néanmoins*	complètement	progressivement	sensiblement
exploit	*prouesse*	offset	plie	bit

Sélection du candidat le plus proche ($d < 10$) pour répondre
+ choix aléatoire autrement (*baseline hasard*=25%)

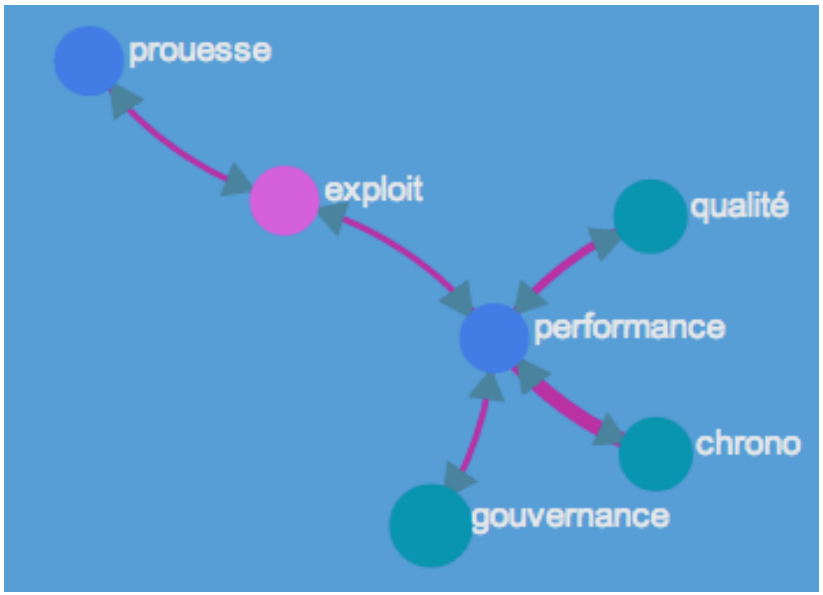


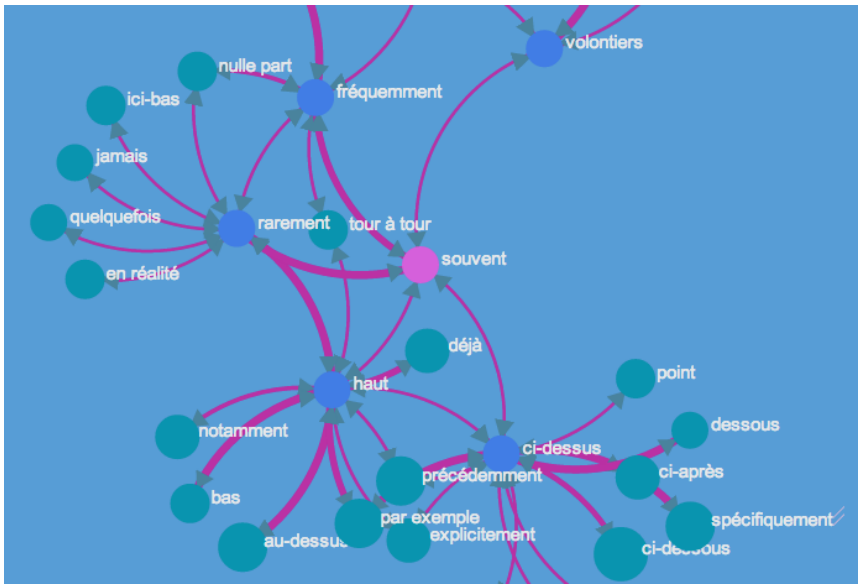
	fwn		wolf	
network	**%ok**	**#tests**	**%ok**	**#tests**
all	51,5	4,121	42,1	7,674
fiscal	46,1	104	37,0	493
business	35,1	248	43,2	1,055
social	39,4	274	37,7	1,345
wolf	64,5	1,076	-	-

Évaluation (suite)

Forte différence en fonction des catégories syntaxiques et pb de couverture.

wordnet	**pos**	**#tests**	**%ok**	**%bad**	**%missing**	**%ok/(ok + bad)**
wolf	v	3,876	35,5	30,9	33,6	53,4
	nc	1,078	33,5	2,1	64,4	94,0
	adj	2,085	22,3	11,3	66,4	66,3
	adv	1,533	36,9	41,9	21,7	46,8





Analyse de la connectivité (POS)

corpus	partie du discours	#mots	#conservés	%c/m
cpl	np	1,779,848	8,749	0,5
	nc	35,417	5,782	16,3
	v	11,224	2,480	22,1
	adj	10,198	3,108	30,5
	adv	2,693	776	28,8
fiscal	terme	50,479	4,981	9,9
	nc	18,760	1,976	10,5
	v	4,783	593	12,4
affaires	terme	65,138	5,142	7,9
	nc	23,506	2,095	8,9

Évaluation: intersection de chemins

Principe:

- les ressources (même WordNet) sont vus comme des graphes.
- étant donné une connection (a, b) dans le graphe G_1,
  vérifier si (a, b) 10-connectable dans G_2
  baseline hasard: $\sim$ 1%; résultats 33% (prec. 80%, rappel: 21%)

Faible couverture de notre réseau
mais en fait également faible couverture des ressources de référence !

Exemples de paires non 10-connectables dans French WordNet:

écrivain pianiste	caverne abîme	caverne gouffre
canne betterave	karaté yoga	facteur enjeu
dirigeant activiste	typhus leucémie	conjuration conspiration
éminence promontoire	harpe violoncelle	cigare cigarette
tendon auriculaire	banderole pancarte	opportunité intention
informatique biologie	nationalisation privatisation	scooter fourgonnette
cancer hémorragie	électricien ajusteur	**mise application**
provocation agression	phoque baleine	militant dirigeant
haschisch cocaïne	informatique génétique	cancer infarctus

...

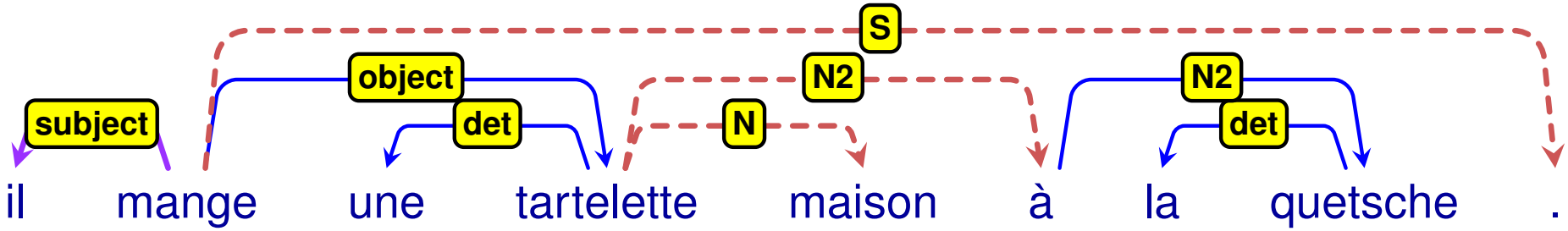
Évaluation par la tâche

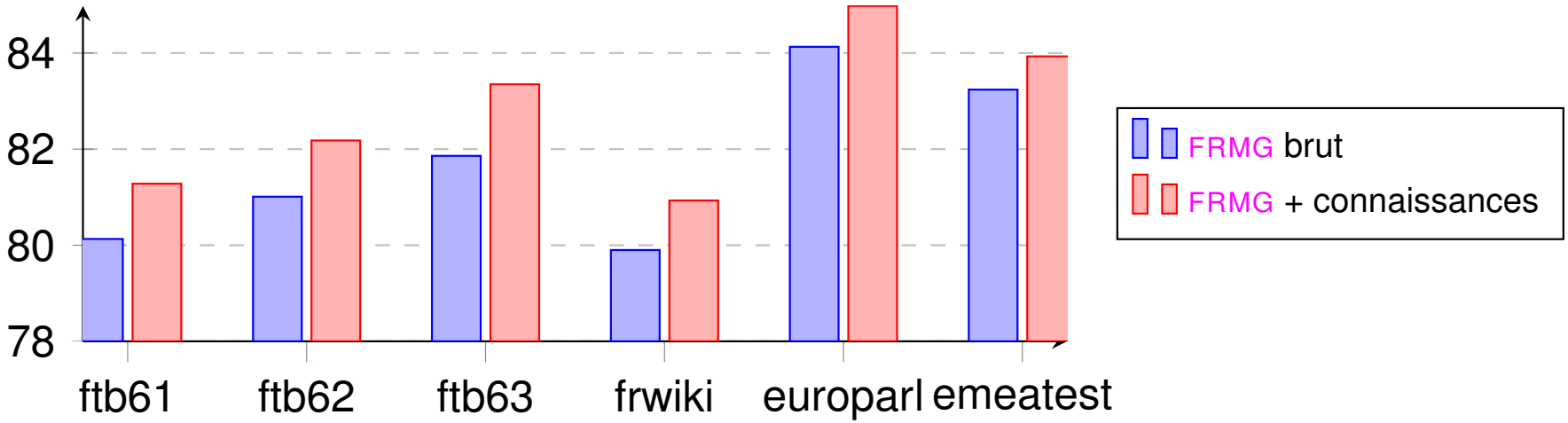
Injection des connaissances acquises dans FRMG (similarité + contextes)

il mange une tartelette maison à la quetsche.

tartelette proche de **tarte**
quetsche sorte de **fruit**
`aux_fruits` contexte fréquent sur **tarte**

$\Rightarrow$ **tartelette à la quetsche**





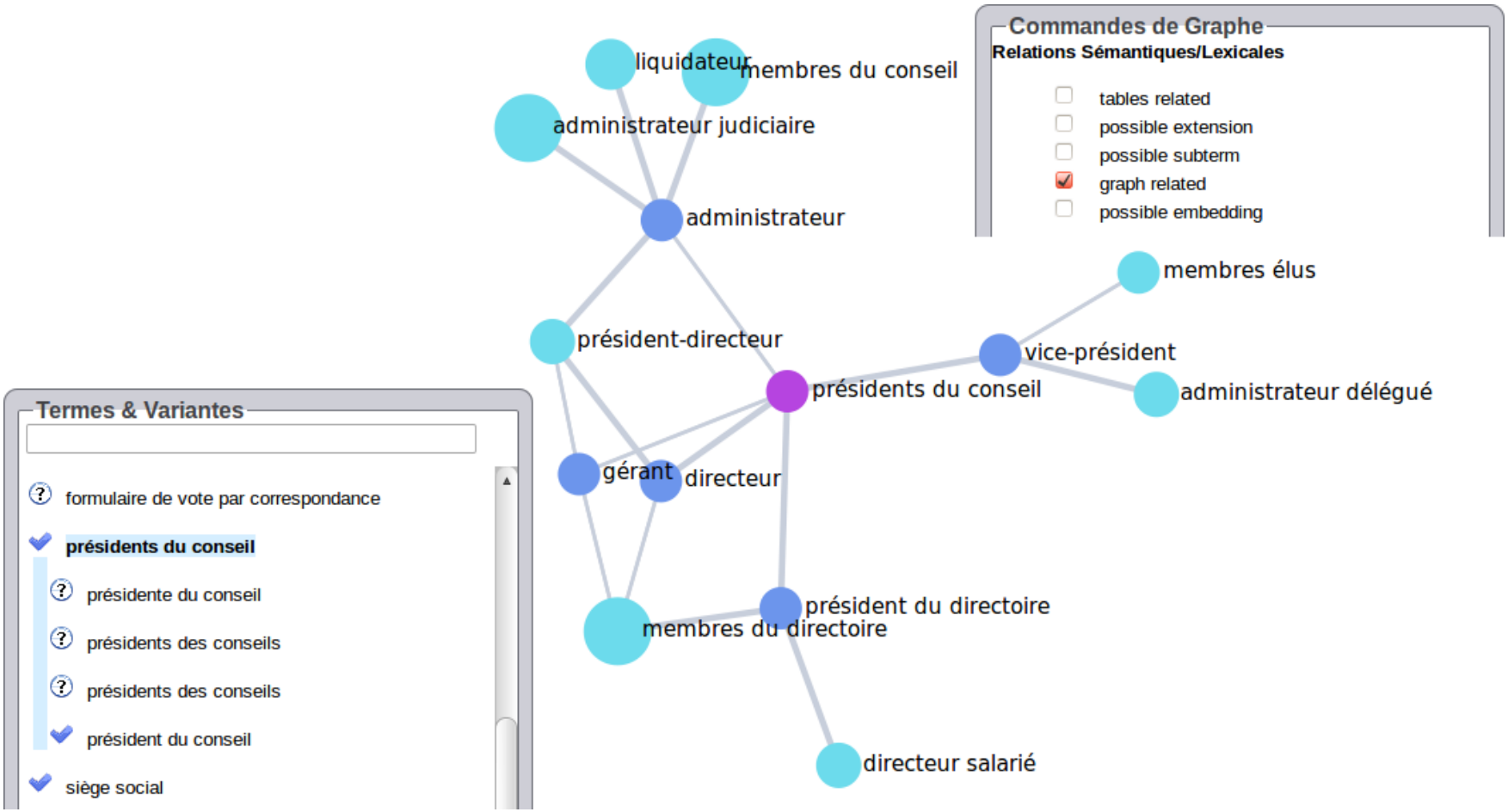
Vers une interface de visualisation et validation

Vrai besoin de:

- visualisation riche, locale (zoom), mais sans surcharge
- mécanismes simples de navigation et de recherche
- accès à des explications et exemples
- validation collaborative
  - ressources imparfaites
  - maintenance et évolutions
  - effort de validation important devant être distribué
  - échanges et discussions

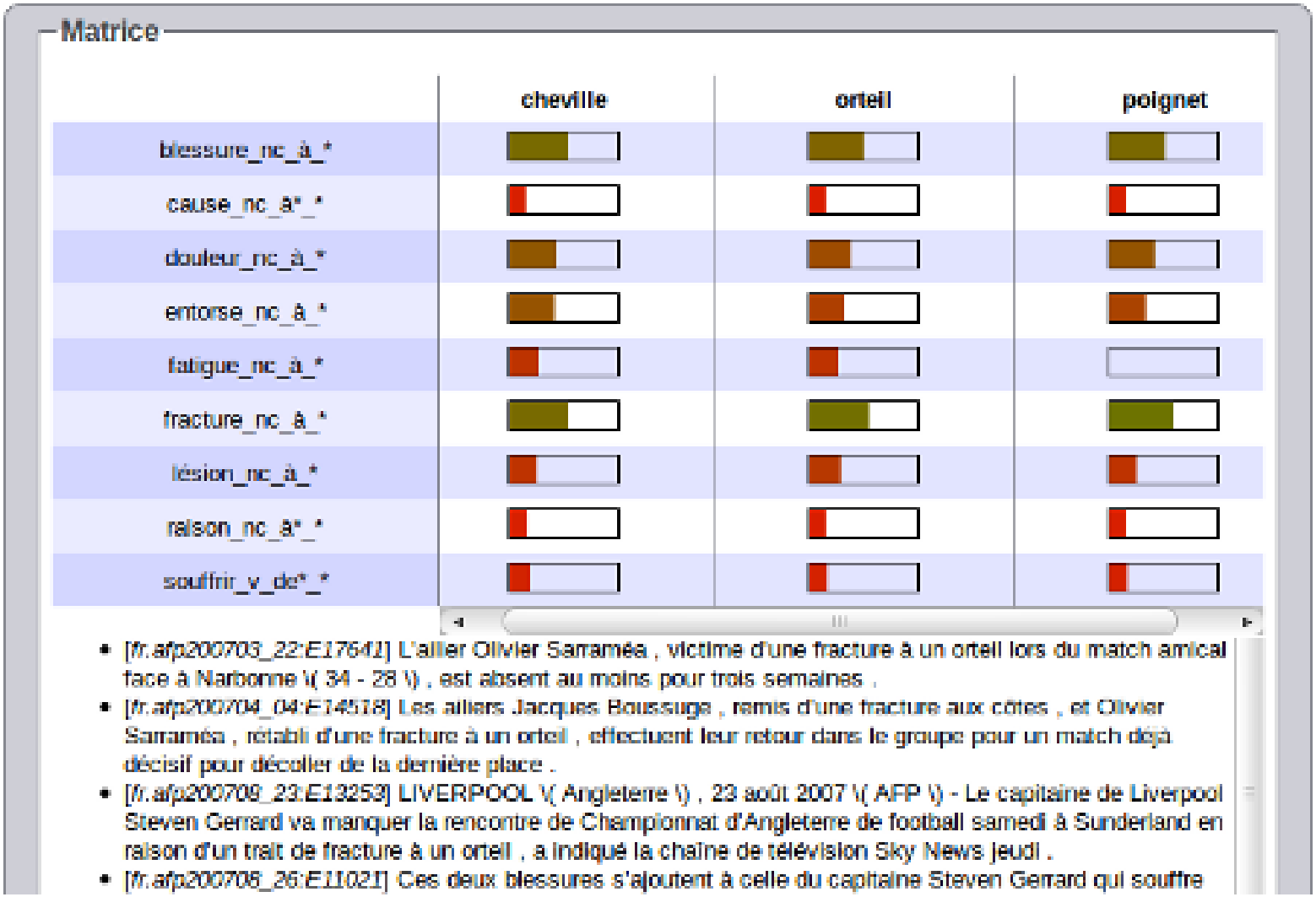
$\Rightarrow$ développement d'une interface WEB sur la plateforme LIBELLEX dans le cadre d'une collaboration avec Lingua et Machina.

Interface sous Libellex [Morado]



Matrices explicatives

Les rapprochements explicables par des contextes dont l'importance fournie par $\langle ww_{i,j}, c_b \rangle = \Sigma_a(wc_{i,a})(cc_{a,b})(wc_{j,b})$



Outline

Autres expériences

D'autres expériences préliminaires menées dans le cadre de SCRIBO sur des sorties d'analyse de corpus:

Concepts

- Extraction de terminologie
  garde à vue
- Construction de réseau de mots
- Regroupement de mots en cluster (*synset*), plus regroupement hiérarchique
- extraction de relations ontologiques (par ex. hypéronymie)
  navire de guerre: *destroyer*, *aviso*

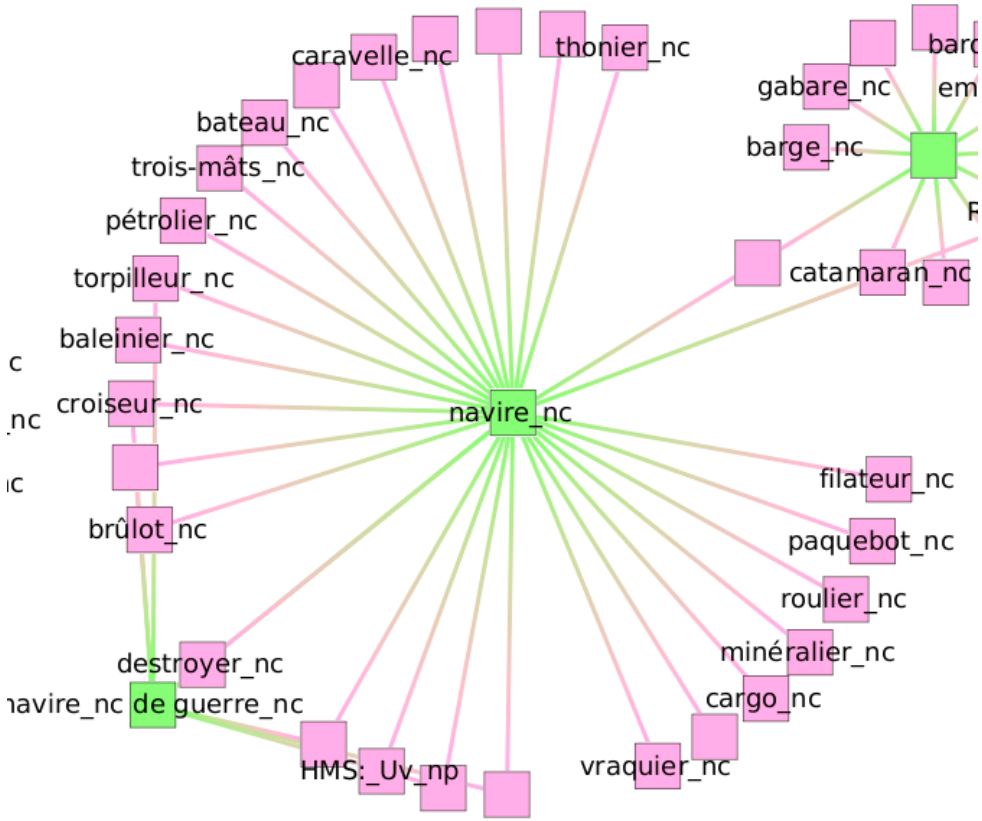
Évènements

- Regroupement de verbes, dénotant un type d'évenement
  - `/transfer/` *donner*, *offrir*, *céder*
  - `/communication act/` *annoncer*, *indiquer*, *affirmer*
- Identification de paires reliées verbe-nom
  - *déclarer*/*déclaration* ;
  - *identifier*/*identification* ;
  - *commencer*/*commencement*/*début*
- Découverte de chemins de dépendances caractéristiques entre des paires d'entités d'un certain type

Assigner des labels aux classes [van de Cruys]

Retrouver et analyser les définitions des mots (dans Wikipédia), avec des patrons d'extraction: *X est (une sorte de|une forme de|. . .)* Y_{genus} . . . differentia

Un porte-avions est un navire de guerre permettant de transporter et de mettre en oeuvre des avions de combat



Regroupement de verbes [Kata Gabor]

Motivation: indentifier des verbes correspondant au mêmes types d'évenements

Collecte et décomte des *frames remplis*

```
12 accepter_v_active subject:cln_cln object:cla_cla
 9 accepter_v_active object:poste_nc
 7 accepter_v_active subject:_LOCATION_np de_xcomp:reprendre_v
 5 accepter_v_active subject:cln_cln object:quelque chose_pro
```

- Selection des 1000 verbes les plus fréquents (AFP)
- **algorithme**: unsupervised hierarchical clustering
- **measure de similarité**: Jaccard index

Résultats: groupes de verbes proches sémantiquement (max. 4 élements) :

- partageant des cadres de sous-catégorisation
- avec des arguments proches sémantiquement dans ces cadres
- **pb principal:** la dispersion des données et les arguments non essentiels

Exemples

- **synonymes**

  accepter approuver adopter voter
  attirer séduire ravir impressionner
  clore clôturer boucler couronner

- **synonyms/antonyms**

  chuter diminuer baisser augmenter
  rassurer décevoir choquer surprendre

- **"scenarii"**

  poser aborder résoudre régler (un problème)
  déposer examiner étudier rejeter (une loi)
  blesser tuer interpeller libérer (violence action)

- **même évenement, autre point de vue**

  subir entraîner provoquer causer
  vendre acheter racheter acquérir

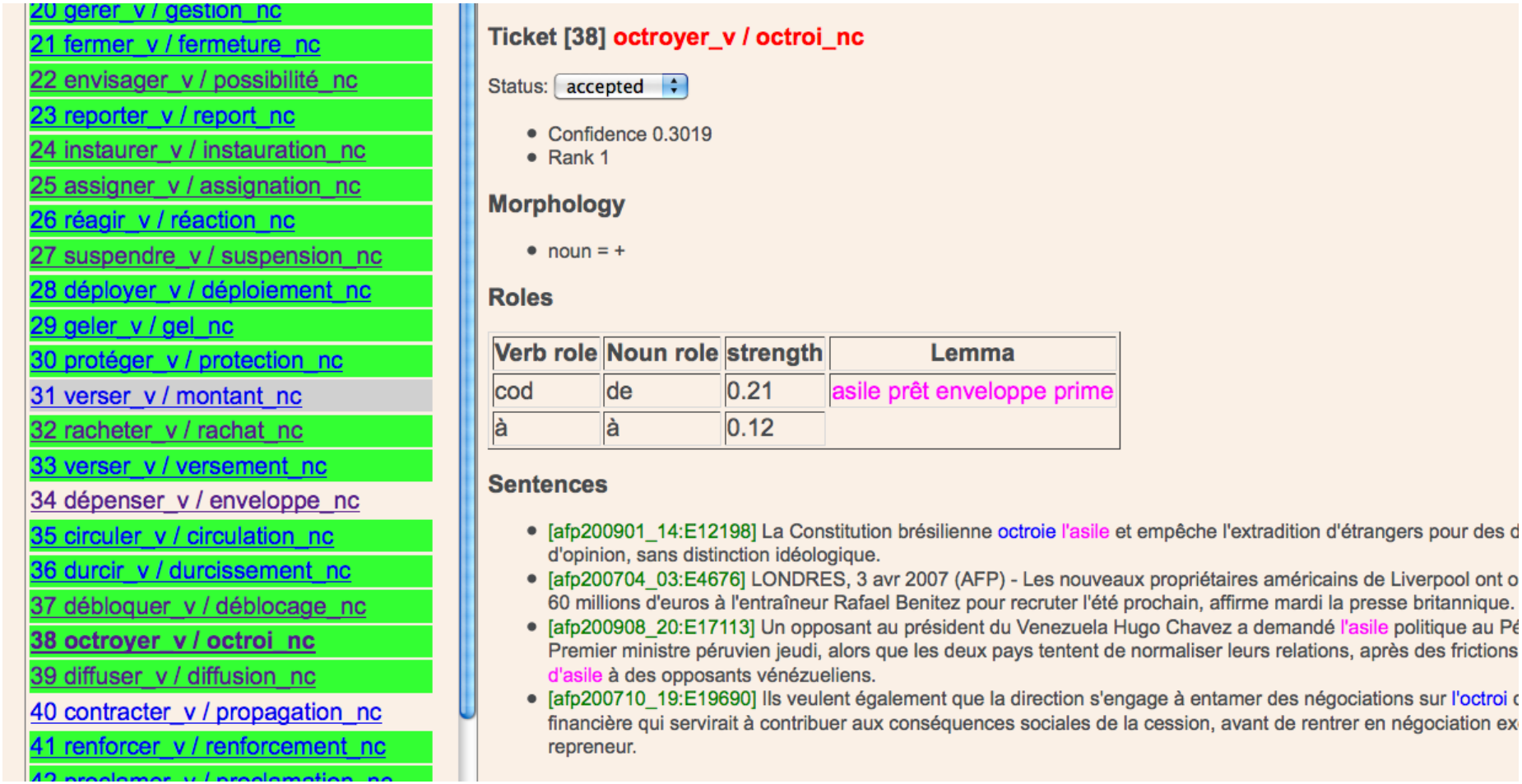
Paires verbe/nom

- **objectif**: identifier des paires verb/noms, où le nom nominalise le verbe:

  X *manifester* contre Y – *manifestation* de X contre Y
  ? les *manifestants* contre Y

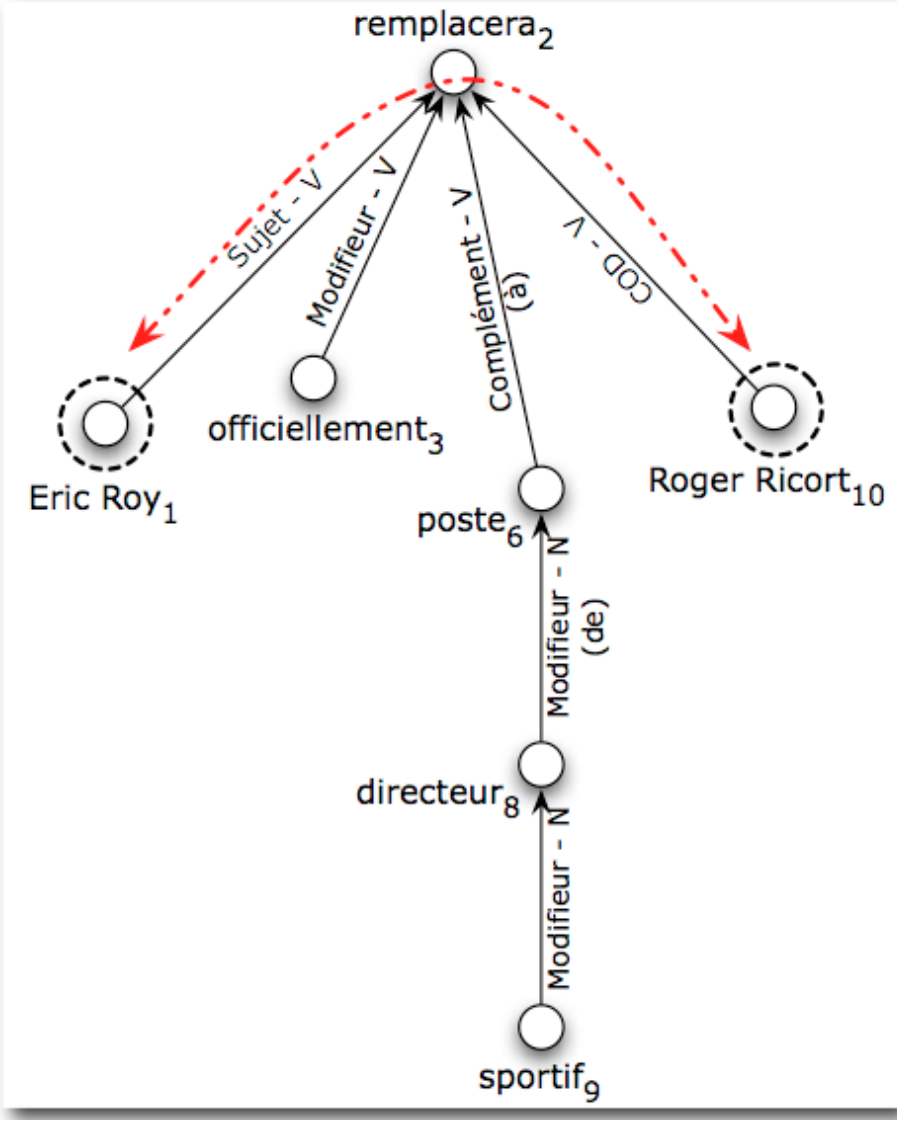
- **algorithme**: combiner
  - similarité distributionnelle: arguments partagés entre verbes et noms (évenementiels), Jaccard Index avec PMI
  - distance d'éditions: forme faible de morphologie dérivationelle

Visualisation et validation



Acquisition of relations between entities [Nakamura]

- Extraction of dependency paths between NEs
- **Hypothesis**: the paths denote the relation between NEs
  $\rightsquigarrow$ patterns for the relation
- Step 1: **Clustering**
  - NE pairs given paths
  - paths given NE pairs
- Step 2: **Acquisition** (2 methods)
  - partially supervised: acquisition by induction
  - non-supervised: classification based on the shared contexts



Clustering

NE pairs $\rightsquigarrow$ paths

paths $\rightsquigarrow$ NE pairs

Semi-supervised acquisition by induction

- Acquisition by induction given a few seed examples
- experiments for the membership relation
- 2 month of AFP news
- Results:
  - extraction of 136 paths, for 1469 NE pairs
  - manual validation of 178 pairs
    $\Rightarrow$ 149 OK; 29 wrong

Outline

Conclusions

- Confirmation de l'intérêt de données syntaxiques pour l'acquisition de connaissance lexicales
  mais la problématique de l'évaluation reste !

- Un environnement assez complet avec une interface de visualisation utile
  Nombreux composants échangeables
  (chaîne linguistique, schéma d'annotation, algo. acquisition, . . .)
  $\Rightarrow$ nombreuses expériences à mener

- Validation des termes faites (domaine juridique)
  $\Rightarrow$ (prévu:) amélioration outils et apprentissage classification/reranking

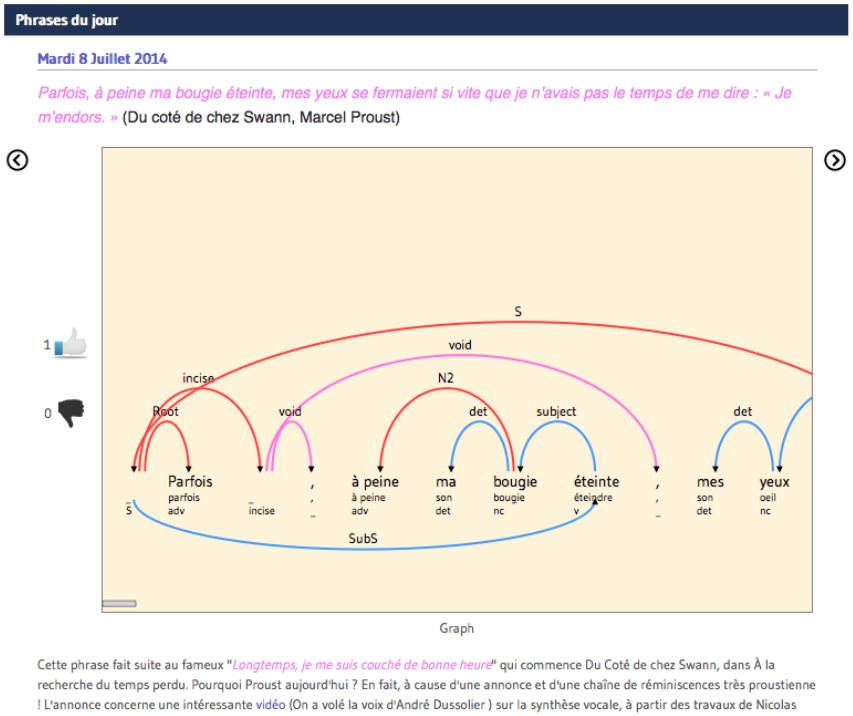
- évolution vers la gestion de ressources plus riches (WOLF)
  mais aussi extraction de relations plus riches

- Système consultable en ligne
  `http://alpage.inria.fr/Lbx` (*guest*/*guest*)
  ok pour des traitements de corpus / chargement de ressources

Découvrir et utiliser FRMG [Bui-Quang]

FRMGWIKI un wiki linguistique aux fonctionalités étendues sur

`http://alpage.inria.fr/frmgwiki`



- découvrir et jouer en ligne avec FRMG
- discuter de phénonèmes syntaxiques avec des phrases d'exemples
- accéder à un service de traitement de corpus (1Mmots)

- disponibilité de gros corpus analysés (Wikipédia, Wikisource, . . .)
- indexation et requêtage sur corpus
- à terme, couplage FRMGWIKI et LIBELLEX