Quelques expériences autour du flux de dépêches AFP

Éric de la Clergerie

<Eric.De_La_Clergerie@inria.fr>

et travaux de B. Sagot, R. Stern, Y. Nakamura, D. Nouvel, . . .

http://alpage.inria.fr

INRIA Paris-Rocquencourt / Univ. Paris Diderot

Journée “Information, Médias et Informatique”

IRISA, 15 Mars 2016

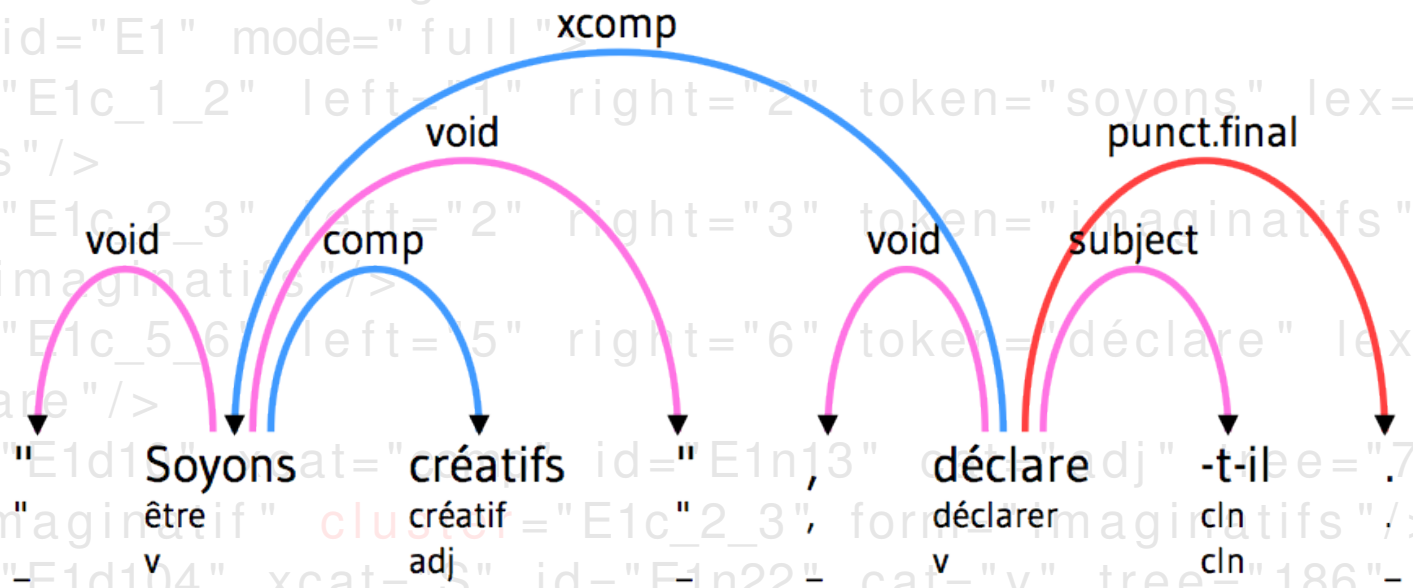
Analyse Linguistique Profonde À Grande Échelle

Outils + Ressources linguistiques

Grammaire

FRMG

Lexique LEFFF



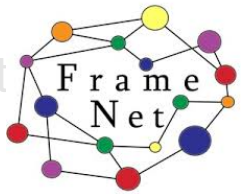
à tester sur `http://alpage.inria.fr/frmgwiki`

Entités Nommées

ALEDA

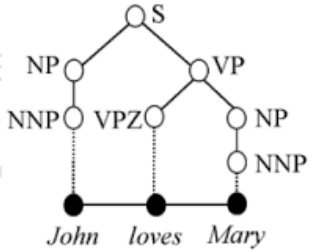
WordNet

WOLF



FrameNet

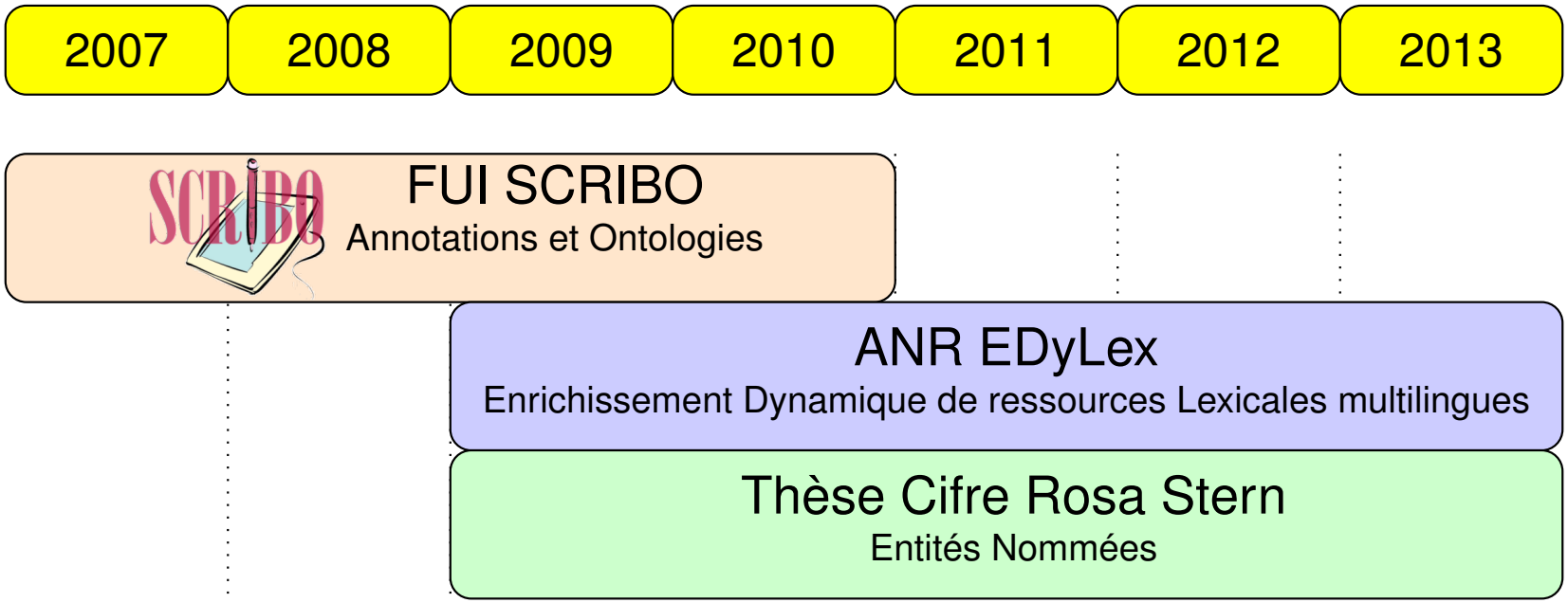
ASFALDA

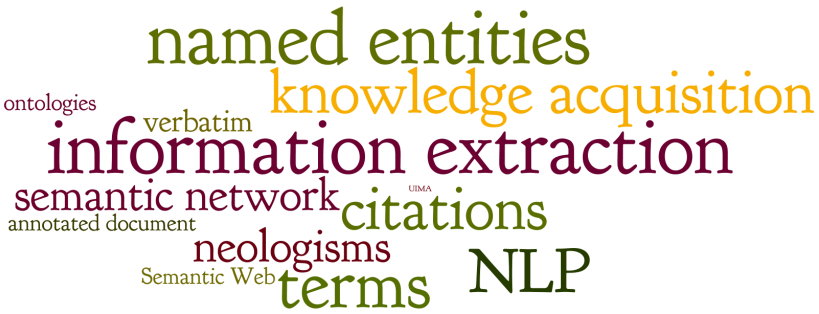


treebanks

Une longue et fructueuse collaboration avec l'AFP

MediaLab de l'**Agence France Presse** :
volontaire pour exploiter leur flux de dépêches
(riche, multilingue et multimédia)





Plan

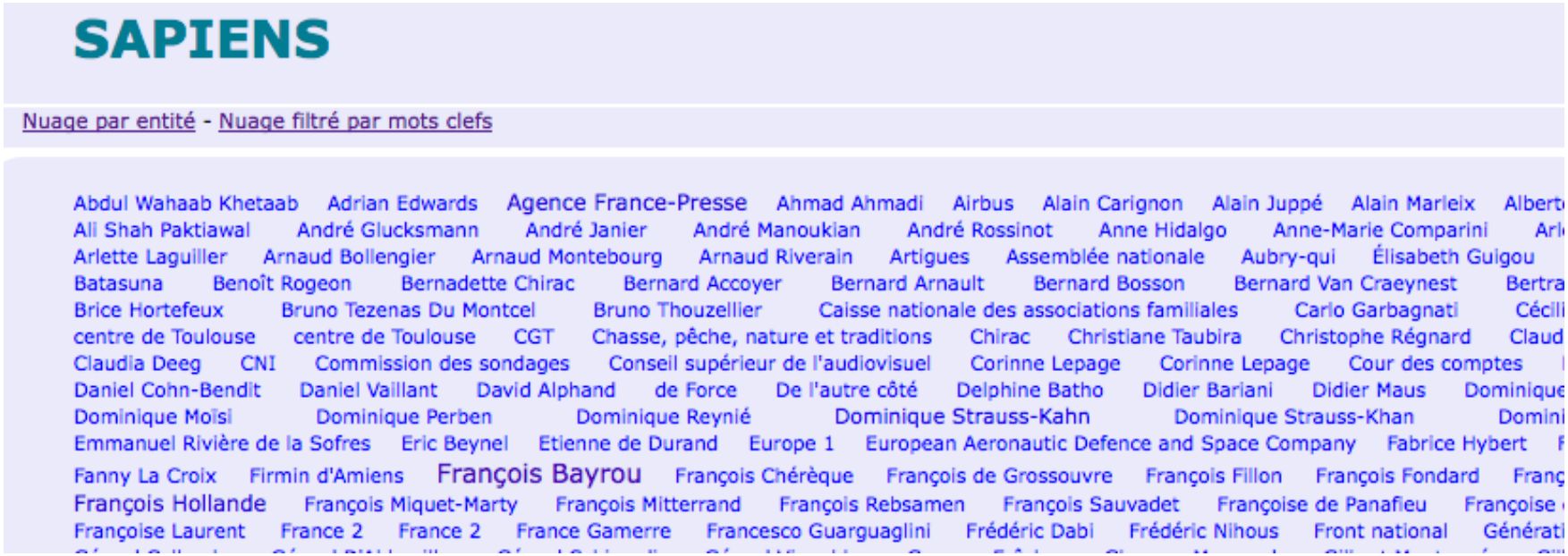
Une dépêche (et enrichissements sémantiques)

Crédit : Rosa Stern

SAPIENS : qui a dit quoi ?

Démonstrateur WEB (`http://alpage.inria.fr/sapiens`)
sur un lot de dépêches concernant la présidentielle 2007

Objectif : retrouver les citations des uns et des autres
(précurseur fact-checking !)



SAPIENS : qui a dit quoi ?

Citations de **François Bayrou** (homme politique français actuel député des Pyrénées-Atlantiques (Bordères, 25 mai 1951)) :

Dépêches : 43	**Citations de François Bayrou : 72**
AFP, 2007-04-01 **Sarkozy et Royal continuent l'affrontement sur la sécurité (CHAPEAU)** PARIS, 1 avr 2007 (AFP) - Nicolas Sarkozy et Ségolène Royal ont continué dimanche leur affrontement sur la sécurité alors que l'ensemble des candidats s'exprimaient tous azimuts dans les meetings ...	- "rien ne peut l'abattre!"
AFP, 2007-04-03 **Bayrou prône une République apaisante, Royal contre les frais bancaires (CHAPEAU)** PARIS, 3 avr 2007 (AFP) - François Bayrou, candidat UDF à la présidentielle, a présenté son programme officiellement mardi, prônant une République "apaisante", tandis que Ségolène Royal a proposé ...	- "Les crédits de la défense nationale ne sauraient être la variable d'ajustement de notre politique budgétaire" - il voulait être un président
AFP, 2007-04-05 **Les candidats planchent pour Elle sur les questions liées aux femmes (PAPIER GENERAL) Par Frédéric DUMOULIN =(PHOTO+VIDEO)=** PARIS, 5 avr 2007 (AFP) - A l'invitation de l'hebdomadaire féminin Elle, la plupart des candidats à la présidentielle planchaient jeudi à Paris sur la condition féminine mais également les autres ...	- "faire un service public de plus, avec la crise de l'Etat, ne (lui) paraît pas la solution idéale"
AFP, 2007-04-05 **François Bayrou opposé à un "service public de la petite enfance"** PARIS, 5 avr 2007 (AFP) - François Bayrou s'est opposé jeudi à la proposition de sa rivale socialiste, Ségolène Royal, de créer "un service public de la petite enfance", en la jugeant "irréaliste ...	- il fallait "chercher le mieux-disant européen" - "Rien que la menace de ces sanctions pourra faire bouger les choses"
AFP, 2007-04-05 **Service public de la petite enfance, droit opposable: des idées controversées (ENCADRE)** PARIS, 5 avr 2007 (AFP) - L'instauration d'un "service public de la petite enfance" ou d'un "droit opposable à la garde d'enfant", proposée jeudi par des candidats à la présidentielle, a été jugé ...	- Des expérimentations pourraient
AFP, 2007-04-06 **Bayrou, toujours dans le trio de tête, peine à reprendre de l'élan (DOSSIER, PAPIER D'ANGLE) Par Pascale JUILLIARD** PARIS, 6 avr 2007 (AFP) - A deux semaines du premier tour, François Bayrou reste dans le trio de tête dans les sondages mais a du mal à perturber le duel Sarkozy/Royal, en raison notamment du tou ...	- "Le vote utile, c'est moi" - "le seul à pouvoir battre Nicolas Sarkozy au second tour".

SAPIENS : qui a dit quoi ?

SAPIENS

Nuage par entité - Nuage filtré par mots clefs

Retourner aux citations de François Bayrou

Service public de la petite enfance, droit opposable: des idées controversées (ENCADRE)

PARIS, 5 avr 2007 (AFP) - L'instauration d'un *"service public de la petite enfance"* ou d'un *"droit opposable à la garde d'enfant"*, proposée jeudi par des candidats à la présidentielle, a été jugée peu réalisable à court terme par le Conseil d'analyse stratégique (CAS) dans un récent rapport.Pour résoudre le problème du manque d'offres de garde et rendre plus égalitaire leur accès, Ségolène Royal (PS), Marie-George Buffet (PCF) et Dominique Voynet (Verts) se sont prononcées pour la mise en place d'un *"service public de la petite enfance"*, jugé en revanche *"irréaliste"* et *"fallacieux"* par **François Bayrou** (UDF). Nicolas Sarkozy (UMP) s'est engagé pour *"un droit opposable à la garde d'enfant"*, d'ici 5 ans.

La notion de *"service public de la petite enfance reste encore très floue et peu opératoire"*, tout comme celle de *"droit opposable"*, qui implique un recours possible devant le tribunal, a jugé le Centre d'analyse stratégique (CAS) dans un rapport remis à sa demande au Premier ministre Dominique de Villepin en février.

Pourtant, dans un pré-rapport, le CAS avait envisagé la création d'un tel service public de la petite enfance, avec garantie à terme d'une solution d'accueil pour tous les moins de 3 ans, sans décider cependant quelle collectivité terrritoriale ou organisme en serait le maître d'oeuvre, une question pourtant cruciale.

Pour expliquer son renoncement, le CAS avançait des *"raisons de coût et de faisabilité matérielle"*, et préconisait certaines mesures : recensement des besoins et structuration de l'offre de garde dans les départements, mise en place d'un service individualisé d'information des familles. Des expérimentations pourraient, disait-il, être lancées pour créer un numéro unique d'enregistrement des demandes des familles.

"Il faut être réaliste, on n'a pas la possibilité aujourd'hui de répondre à chaque Français confronté à une difficulté de garde d'enfant", a déclaré à l'AFP Jean-Louis Deroussen, président de la Caisse nationale des allocations familiales (Cnaf).

En revanche, l'Union nationale des associations familiales (Unaf) est favorable au principe d'un service public de la petite enfance. *"On sait qu'en matière de garde d'enfant, on manque dans certains territoires cruellement de solutions. Cela aura un coût pour la collectivité territoriale et l'Etat, mais ce sera positif pour les familles"*, a pour sa part estimé François Fondard, président de l'Unaf.

Sélectionner toutes les entités

Masquer les citations

- Agence France-Presse
- Caisse nationale des associations (1)
- Dominique Voynet
- Dominique de Villepin
- **François Bayrou (1)**
- François Fondard (1)
- Jean-Louis Deroussen
- Marie-George Buffet
- Nicolas Sarkozy (1)
- Parti communiste français
- Parti socialiste
- Premier ministre
- Ségolène Royal
- Union nationale des associations f
- Union pour la démocratie franç
- Union pour un mouvement pop

SAPIENS : les suites

Les idées de SAPIENS utilisées par l'AFP pour :

- une application WEB pour les élections de 2012 sur le site de Libération
- une application sur smartphone

Derrière la scène : NewsProcess

NEWSPROCESS : un service HTTP de traitement de dépêches

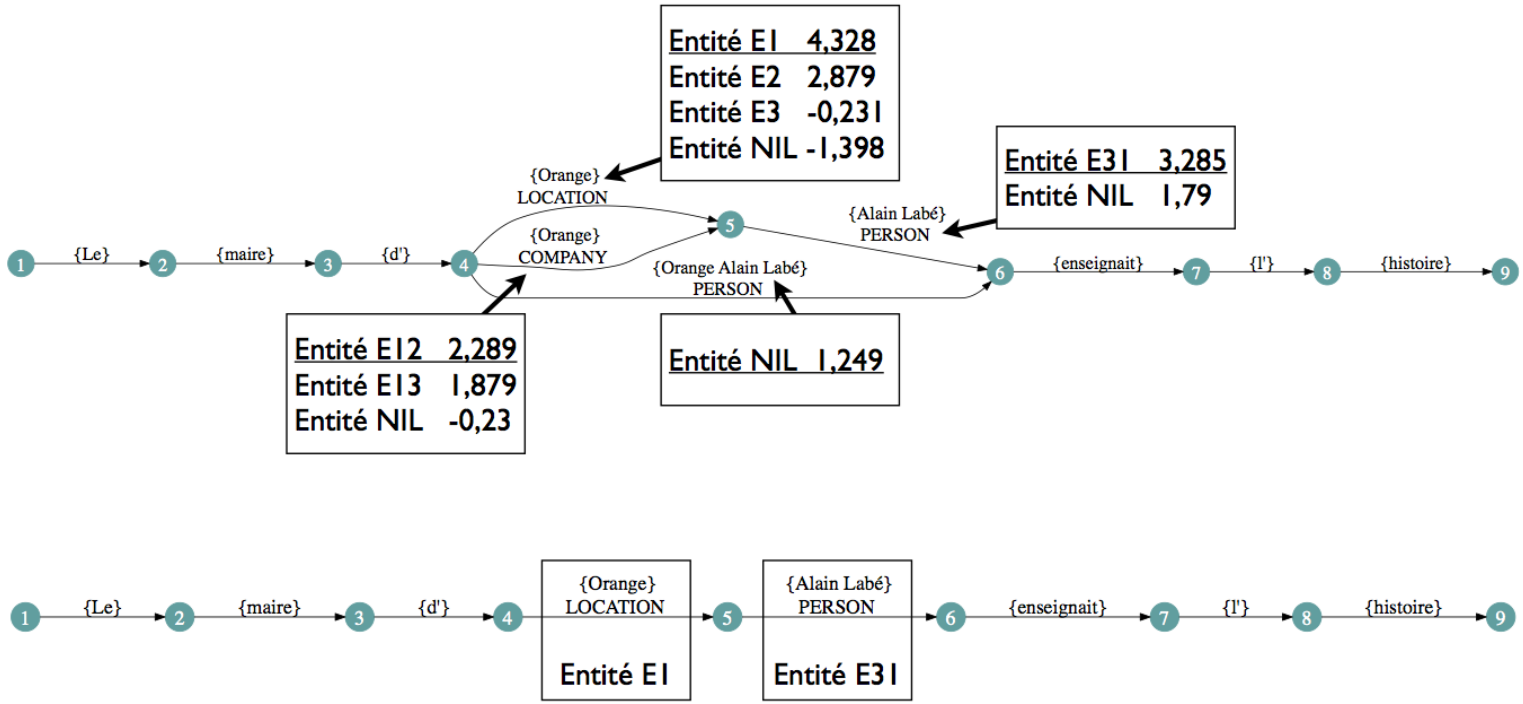
- implanté en Perl comme un pipeline de traitements
- chaque traitement ajoute une couche d'annotations (XML)
  et exploite les annotations des couches précédentes
  1. la dépêche sous format NewsML
  2. Segmentation (phrase/mots) et Reconnaissance Entitées Nommées (REN)
     avec SXPIPE et ALEDA
  3. (Alignement token/texte)
  4. Analyse Syntaxique avec FRMG
  5. Extraction des *verbatim* (segments quotés)
  6. Résolution des co-références (pronoms-entités)
  7. Liage (global) des entités (entités-entités)
  8. Extraction des citations

Inspiré de UIMA (en plus simple !), de GATE et de Tipster
L'architecture facilite l'ajout de nouveaux traitements

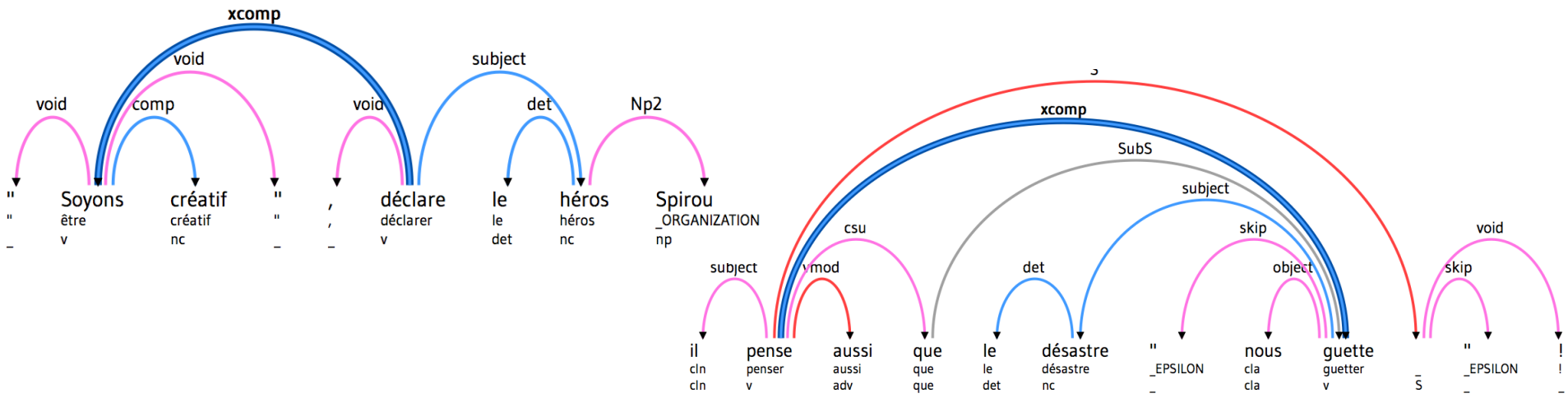
Tokenisation et Entitées Nommées

- Découpage du texte en phrases et mots (SXPIPE)
  en conservant les ambiguités de segmentation
- Détection et liage (local) des entités nommées
  avec résolution de certaines ambiguïtés (NPNORMALIZER, Sagot)
  - expressions régulières et règles CFG (Nombres, Dates, Adresses, Noms, . . .)
  - listes d'EN (ALEDA)

Le maire d'Orange Alain Labé enseignait l'histoire



Analyse syntaxique et extraction des citations



Utilisation de requêtes DPath pour extraire les citations et leurs actants :

```
dpath is_xcomp
  .( source is_active { $citation_verbs->{$_->lemma} })
```

⇝ travail linguistique sur les verbes de citations et modes d'expression
e.g. : **selon X, S – S, poursuivit X**

Note : Citations et **Verbatim** sont deux notions distinctes

Plan

Un corpus important

Les dépêches AFP représentent un volume important de textes (de qualité) :

Corpus	**#phrases (millions)**	**#mots (millions)**	**Description**
AFP	14.0	248.3	400K dépêches (30mois)
Wikipedia (fr)	18.0	178.9	504K pages encyclopédiques
Wikisource (fr)	4.4	64.0	12.8K textes littéraires
EstRepublicain	10.5	144.9	journalistique
JRC	3.5	66.5	directives européennes
EP	1.6	41.5	débats parlementaires
Total	52.0	744.2	

$\Rightarrow$ source utile pour acquérir des connaissances linguistiques et *ontologiques*

Acquisition de connaissances

Collecte et décompte de motifs récurrents dans les analyses syntaxiques ensuite utilisés dans 2 grandes directions :

Concepts

- Extraction de terminologie
- Construction de réseau de mots (proximité sémantique entre mots)
- Regroupement de mots en cluster (*synset*), plus regroupement hiérarchique
- extraction de relations ontologiques (par ex. hypéronymie)

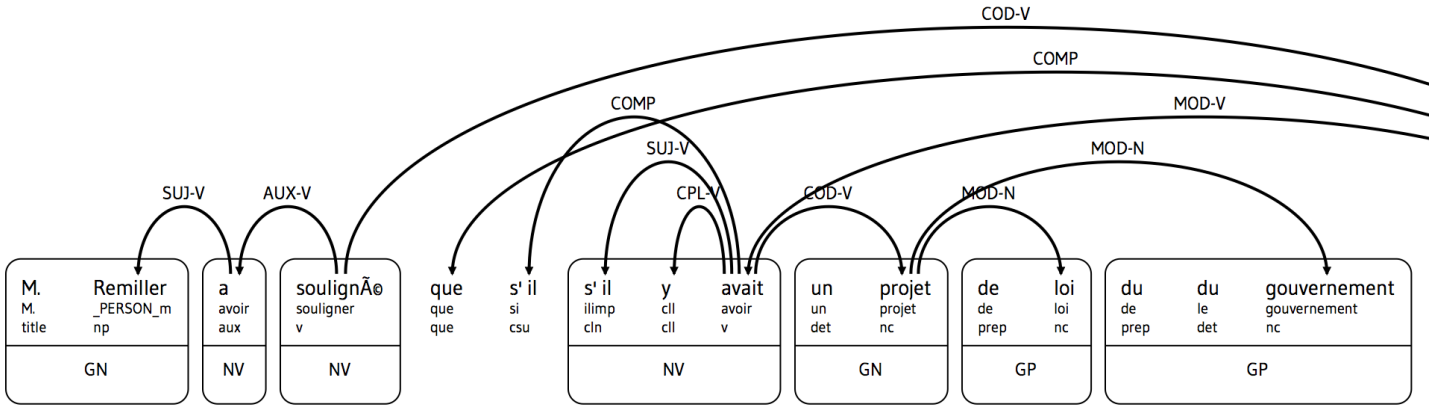
Évènements

- 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 entre entités François Bayrou président de Modem

Extraction terminologique

M. Remiller a souligné que s'il y avait un projet de loi du gouvernement, "nous voulons qu'une partie de nos propositions puisse être retenue"

`afp200701_02:E3484`



Facteurs (plus ou moins classiques) gouvernant l'extraction de termes :

- **structure** (chunks bien adaptés) `(GN)(GR*GA|GP|PV)+`
- forte **fréquence** (f=20989)
- forte **cohésion** interne (information mutuelle, mi=0.54)
- forte maximalité et **autonomie** : un terme doit exister comme un GN nu i.e., pas toujours modifié, présent dans des rôles objects, sujets, ...

$\Rightarrow$ plus de 100k termes potentiels (ordonnés) issus des dépêches AFP
pas de seuil, bruit : organisations, évènements (entités nommées), ...

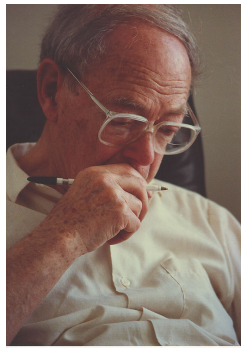
Les candidats termes les mieux notés

coupe du monde
projet de loi
ministre des affaires étrangères
émissions de gaz à effet de serre
chef de l'etat
secrétaire d'etat
cour d'appel
élection présidentielle
programme nucléaire
régimes spéciaux de retraite
président de la république
ministère des affaires étrangères
gaz à effet de serre
assemblée nationale
quarts de finale
organisation de défense des droits de l'homme
ministre des affaires
service de renseignement

Réseau de mots

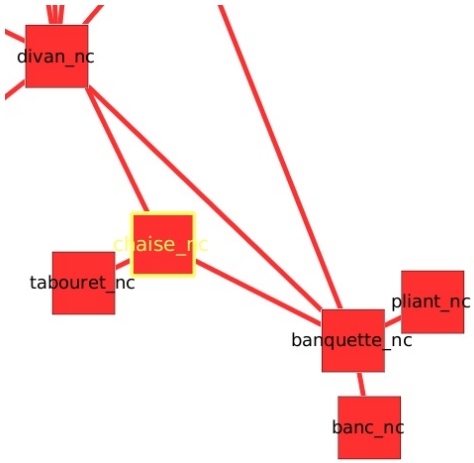
Meanings of words are (largely) determined by their distributional patterns (Harris 1968)

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



À quoi sert une chaise ?

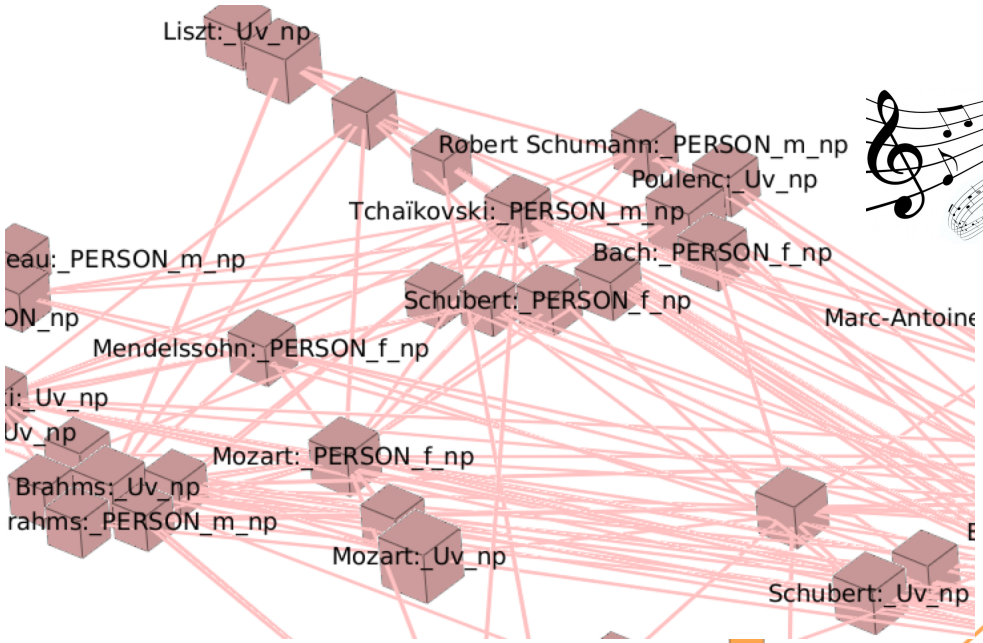
Rapprochement des mots en fonction de la similarité de leurs contextes

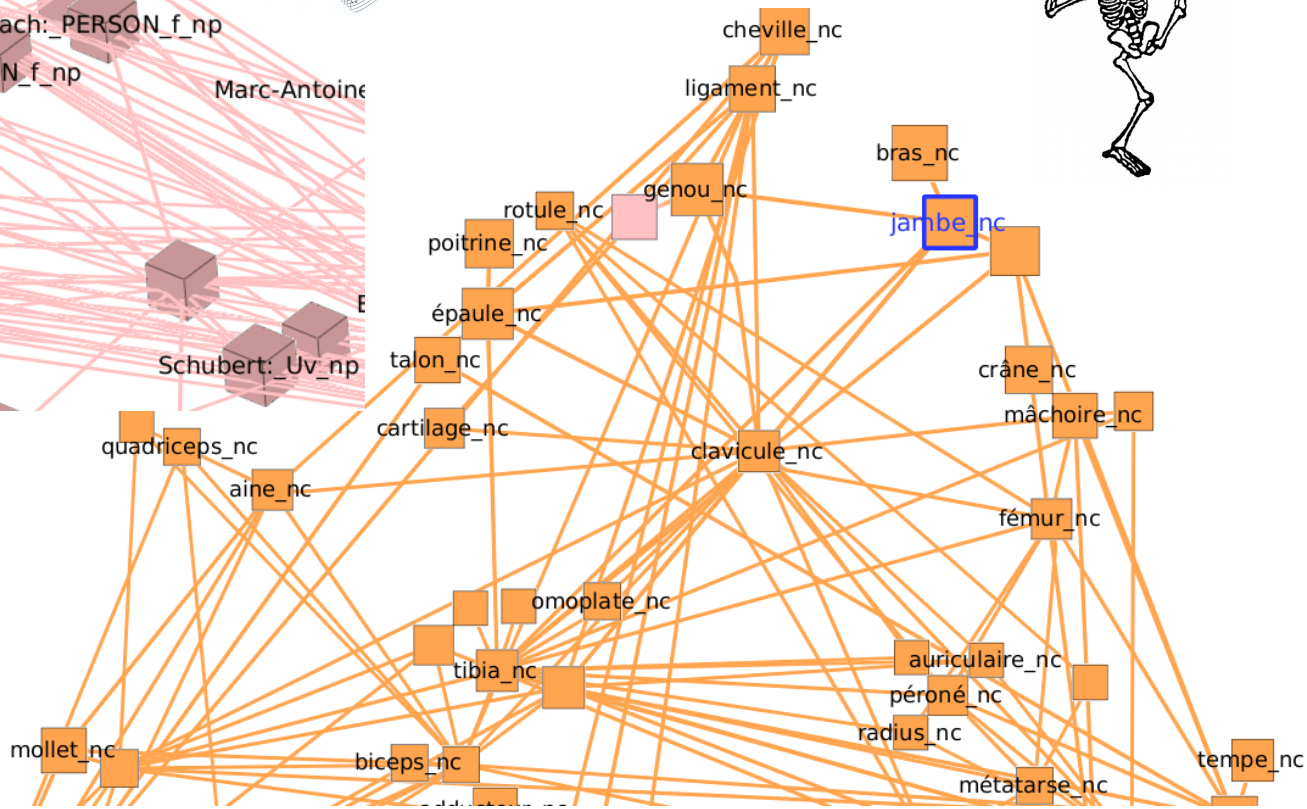


	chaise divan	chaise tabouret	banquette divan	banquette canapé	banquette chaise
se asseoir sur [•]	●	●	●	●	●
asseoir sur [•]	●	●	●	●	●
allonger sur [•]	●		●	●	
dormir sur [•]	●		●	●	●
tomber sur [•]	●		●	●	●
monter sur [•]		●			●
place sur [•]					
grimper sur [•]		●			●

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

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





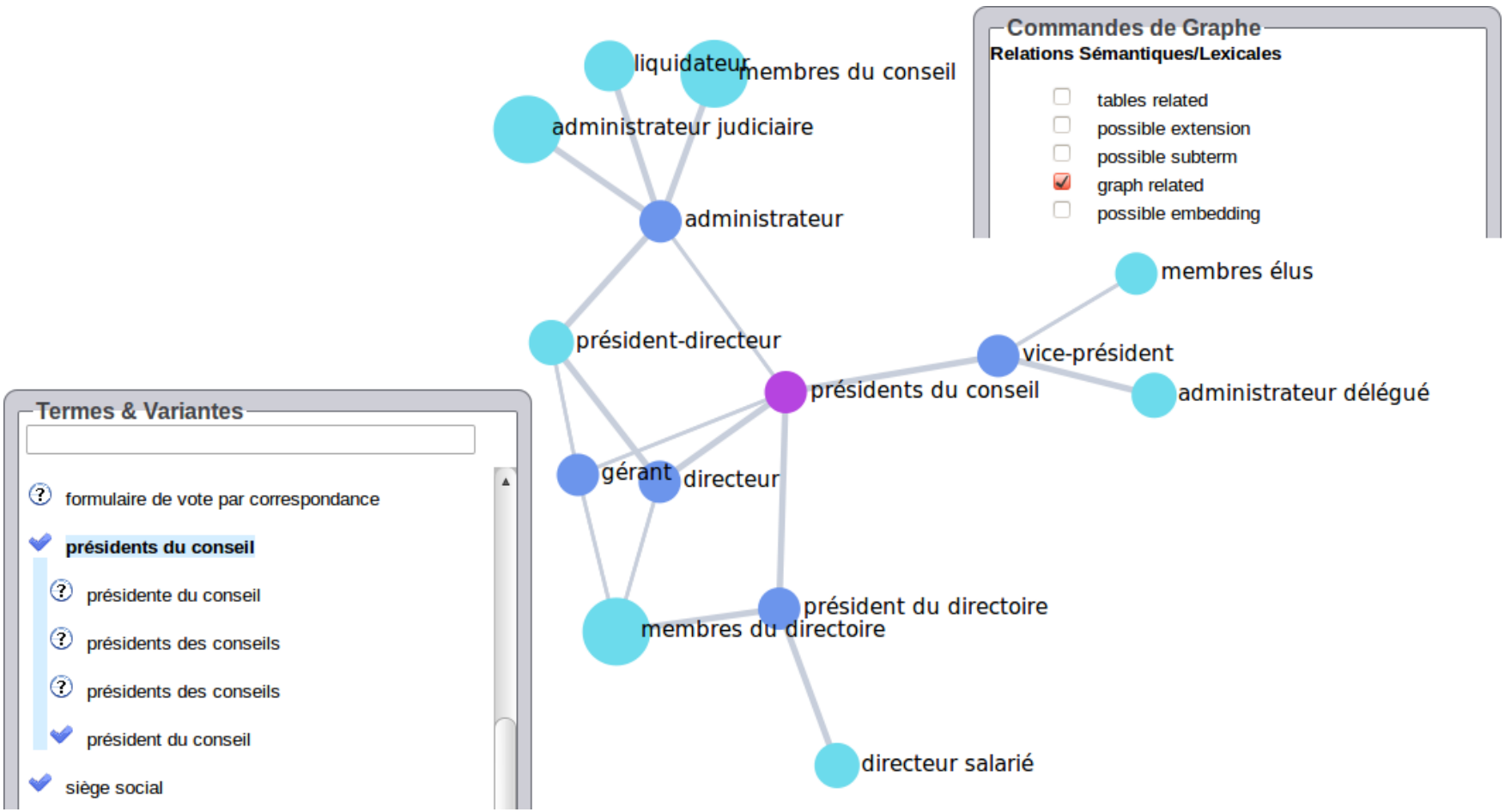
Vers une interface de visualisation et validation

Vu les volumes, complexité et erreurs,
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
en collaboration avec Lingua et Machina.

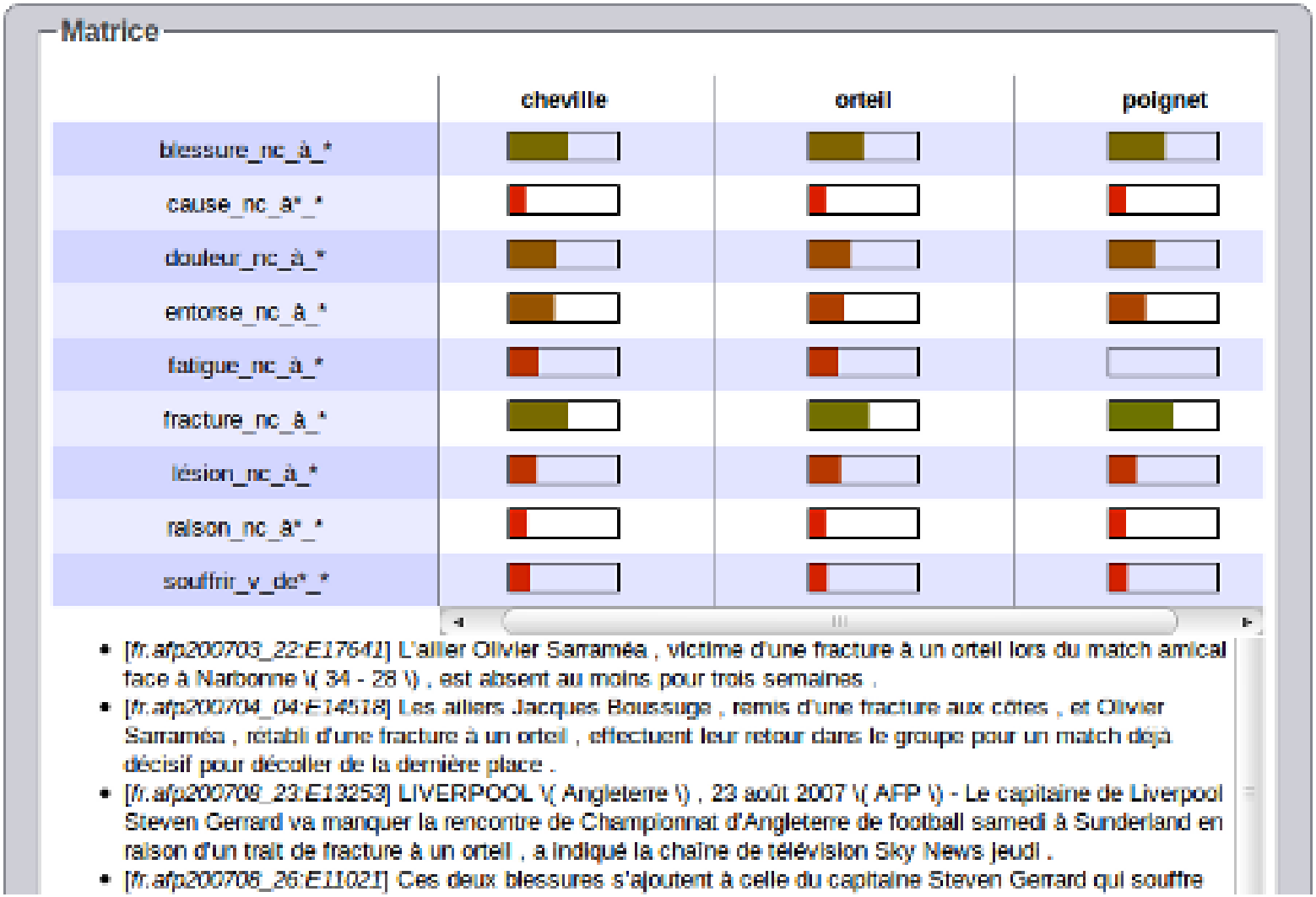
Interface sous Libellex



http://alpage.inria.fr/Lbx (guest/guest)

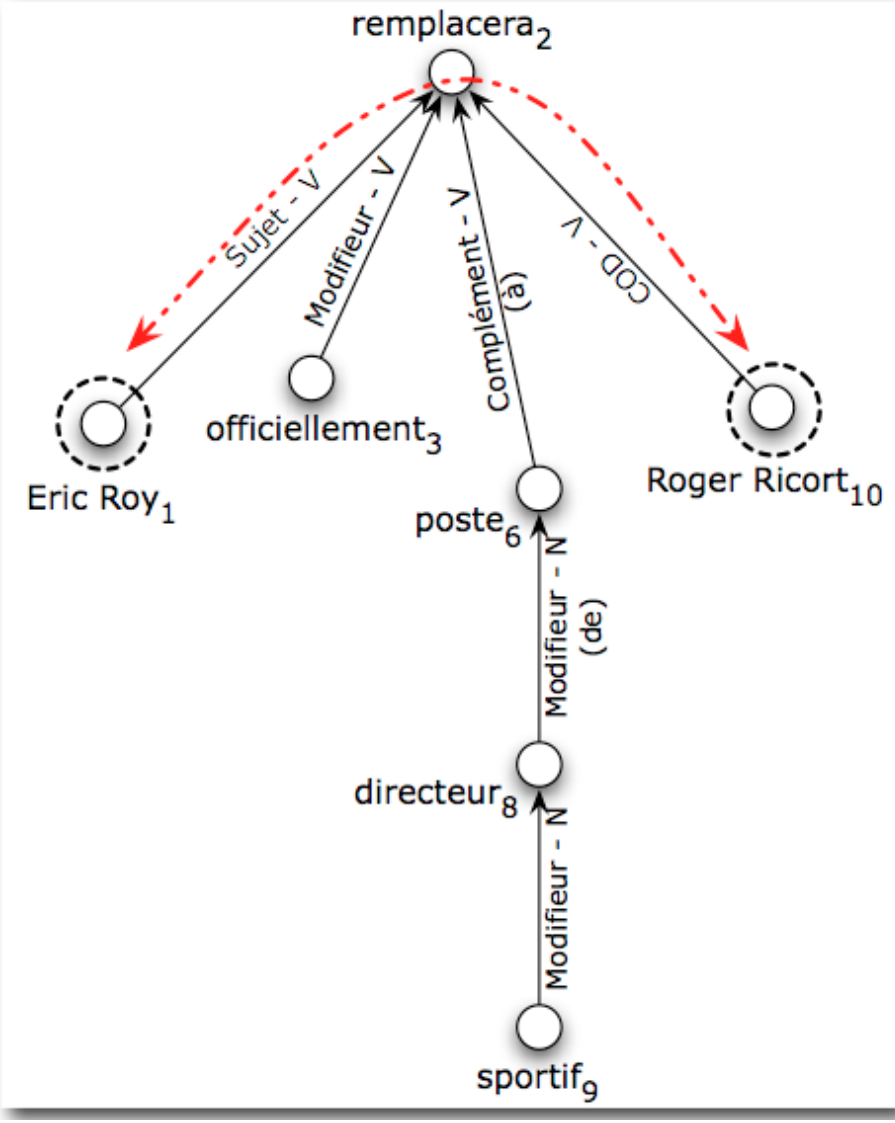
Matrices explicatives

Les rapprochements explicables par des contextes pondérés



Acquisition de relations entre entités Nakamura

- Extraction des chemins de dépendances entre ENs
- **Hypothèse** : les chemins représentent la relation entre les ENs
  $\rightsquigarrow$ patrons pour cette relation
- Étape 1 : **Regroupement**
  - des couples d'ENs selon les chemins
  - des chemins selon les couples d'ENs
- Étape 2 : **Acquisition** (2 méthodes)
  - semi-supervisée : acquisition par induction
  - non-supervisée : classification selon les contextes partagés



Regroupement

chemins $\rightsquigarrow$ Couples EN

couples EN $\rightsquigarrow$ Chemins

Acquisition semi-supervisée par induction

- Acquisition par induction à partir de quelques exemples
- expérience sur la relation d'appartenance
- 2 mois de dépêches AFP (Mai-Juillet 2009)
- Résultats (pour l'appartenance) :
  - 136 chemins extraits
  - 1469 paires de ENs extraites
  - vérification manuelle de 178 paires
    $\Rightarrow$ 149 correctes (83.7%)

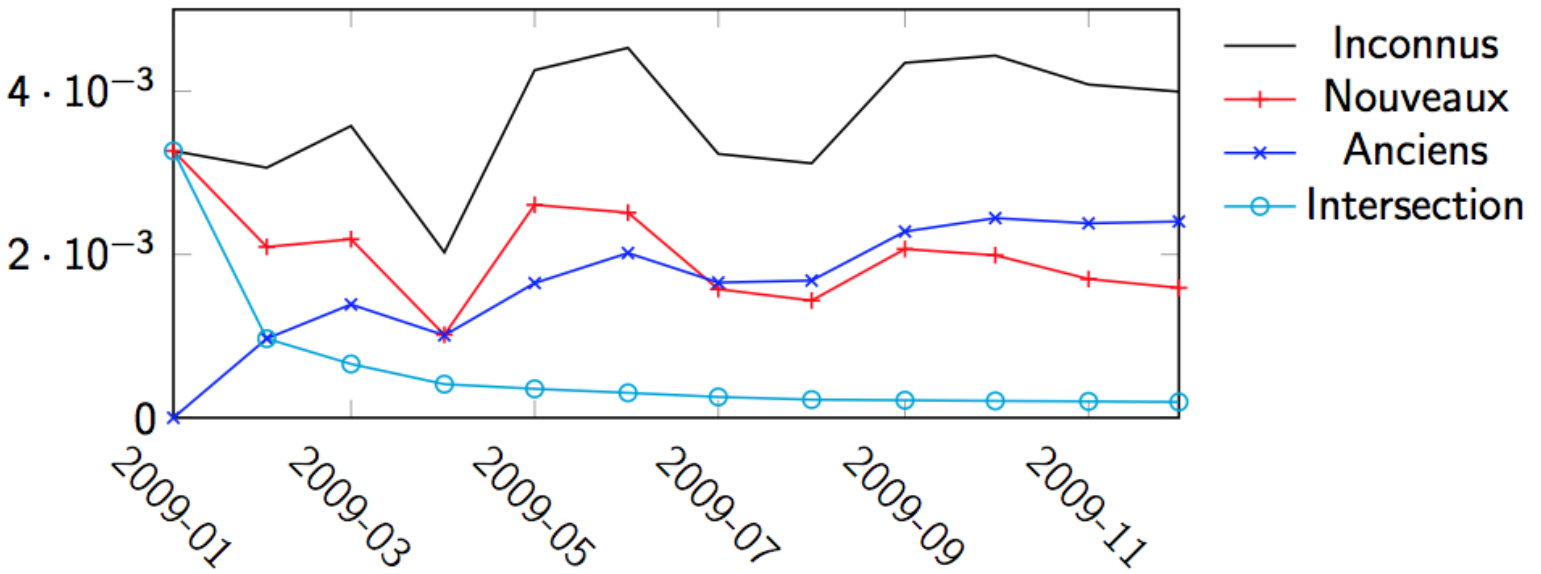
Plan

Flux pour suivre les évolutions du langage

EDyLex : Enrichissement Dynamique de ressources Lexicales multilingues
Partenaires : AFP - Alpage - LIF - LIMSI - Syllabs - Vocapia Research

Constat : en permanence de nouvelles entités et de nouveaux termes

sur les 311 981 dépêches de l'année 2009, après annotation automatique



$\rightsquigarrow$ un inconnu distinct tous les $\sim$1000 tokens

Détecter et gérer la nouveauté ?

Comment distinguer les nouveaux termes/entités de typos ?
distance d'édition, quotes, répétitions, sources externes, paradigmes, emprunts, ...

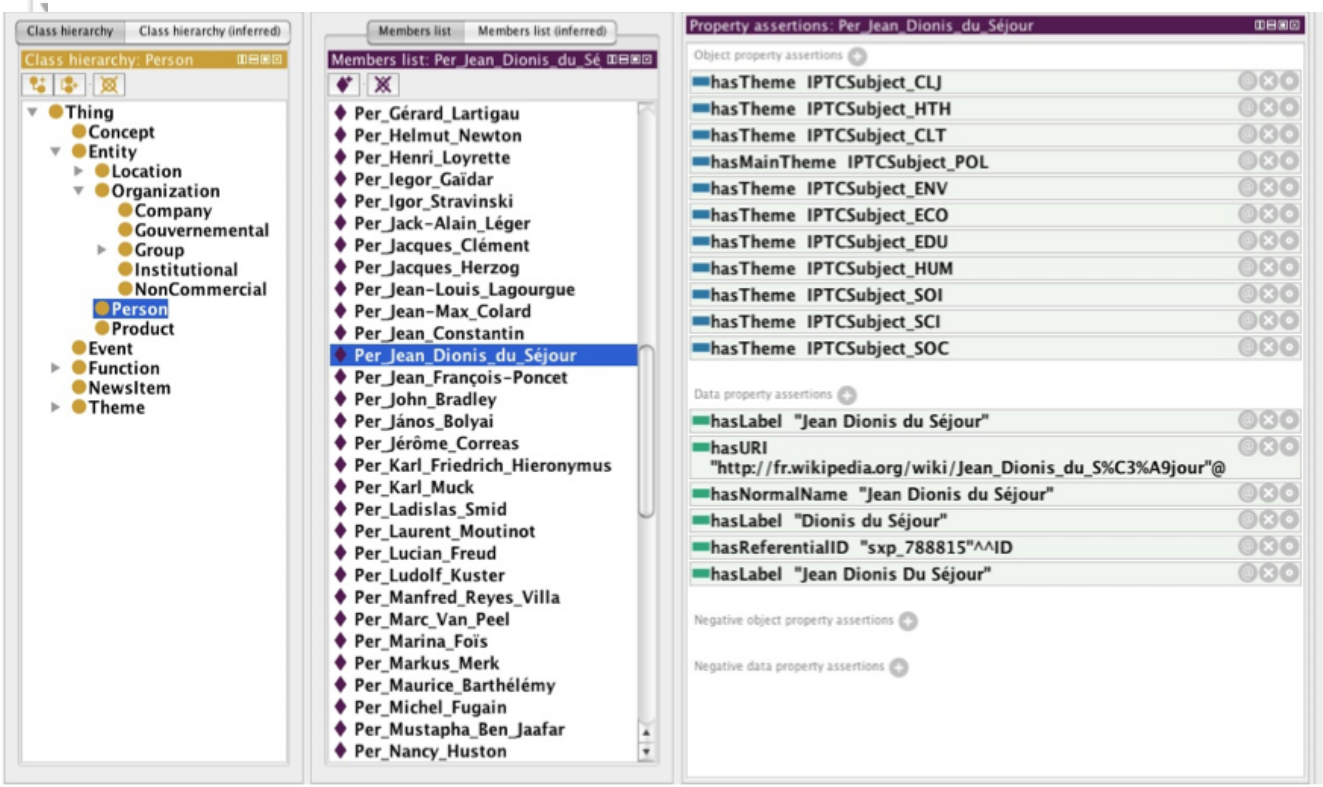
Exemples (janvier 2013) :

- (termes) cryothérapie, lumino-technique, narcoterroriste, co-attribué, pro-putsch, graffitiste, galvanisante
  - **cryothérapie** : Wiktionnary (NC)
  - **galvanisante** : dérivation analogique ; galvaniser +ant(e)(PP)
  - **narcoterroriste** : dérivation préfixale ; narco- + terroriste (NC)
  - **lumino-technique** : composition ; lumineux + technique (ADJ)
  - (Robert 2014) **kéké**, **kriek**, **choupinet** : non analysés !

- (entités) Gamède, Konna, Sévaré, MISMA
  - comment leur donner du sens ?
  - comment les écrire et les prononcer ?

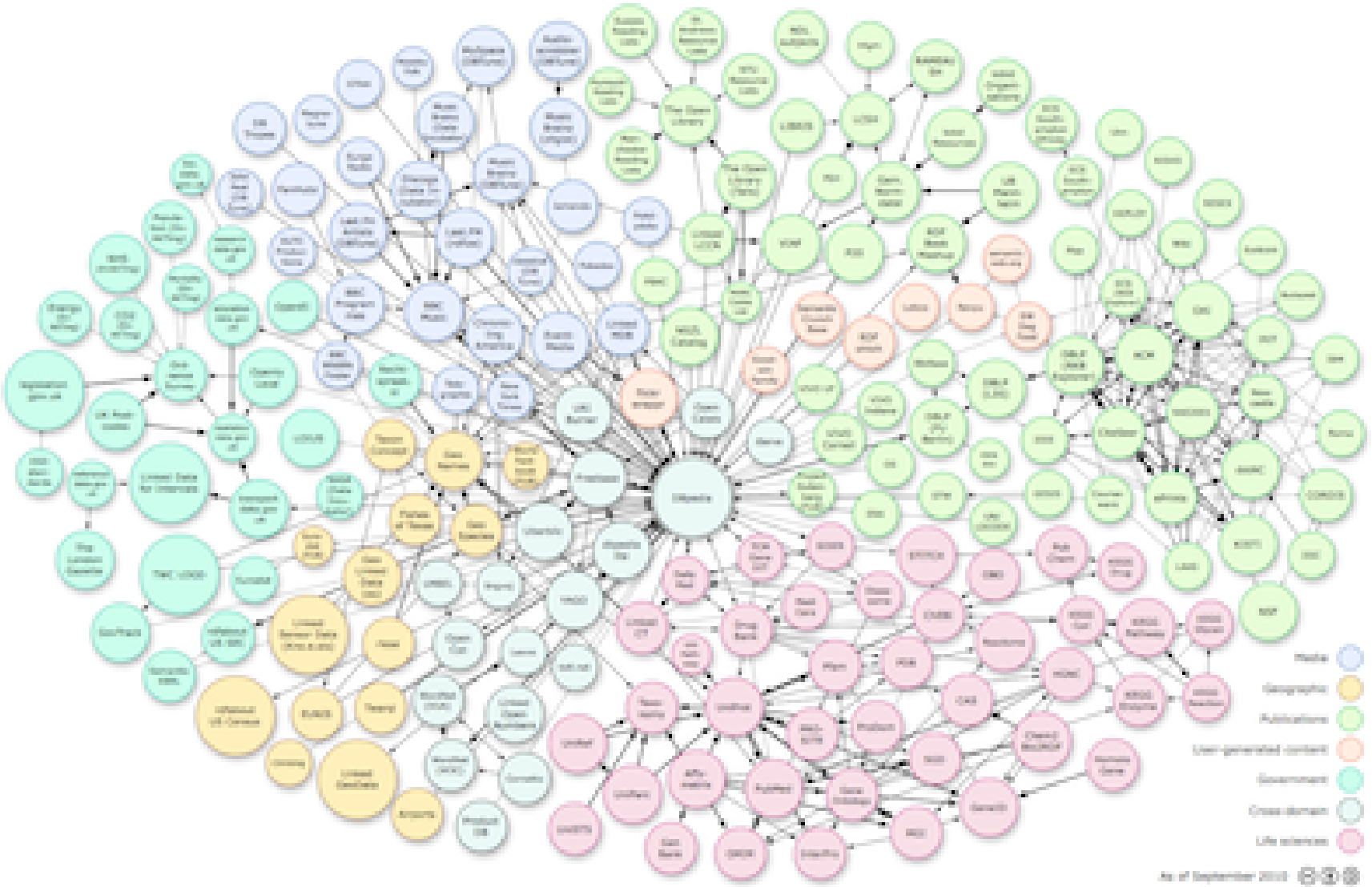
Référentiel AFP d'Entités Nommées

besoin de constituer et maintenir un référentiel



Linked Open Data

LOD : des sources de données connectées et accessibles



Base Aleda et apprentissage

Base Aleda :

- constituée (entre autres) à partir de Wikipedia et GeoNames
- pondération par popularité (taille) et niveaux geonames
- régulièrement mise à jour

PERSONNE	ORGANISATION	ENTREPRISE	LIEU
304 158	41 543	18 109	465 926
PRODUIT	**ŒUVRE**	**FICTIONCHAR**	**Total**
2 526	83 713	6 729	**922 704**

NOMOS (Stern) : apprentissage supervisé pour lier les entités

Évaluation (corpus GAFP : 96 dépêches sur Mai/Juin 2009)

- 1535 occurrences pour 641 entités ;
- 2 liages possibles par occurrence (moyenne)

	P_{REN}	R_{REN}	F_{REN}	Acc_{liage}	F_{liage}
SXPIPE+ NPNORMALIZER	87,75	78,22	82,71	88,78	73,43
LIANE + NOMOS	87.64	83,88	85,71	87,60	75,08
(LIANE ∩ SXPIPE) + NOMOS	92.95	72.82	81.66	91,96	75,10

Expérience

Détection de mots et entités inconnu(e)s dans 6 mois de dépêches, et injection des plus fréquents dans un reconnaisseur vocal :

mois	04	05	06	04-06	01-06
seuil d'occurrence	>1	>1	>1	>2	>3
nouveaux mots	2382	2666	568	2349	4232
match	4/12	6/12	3/12	6/12	8/12

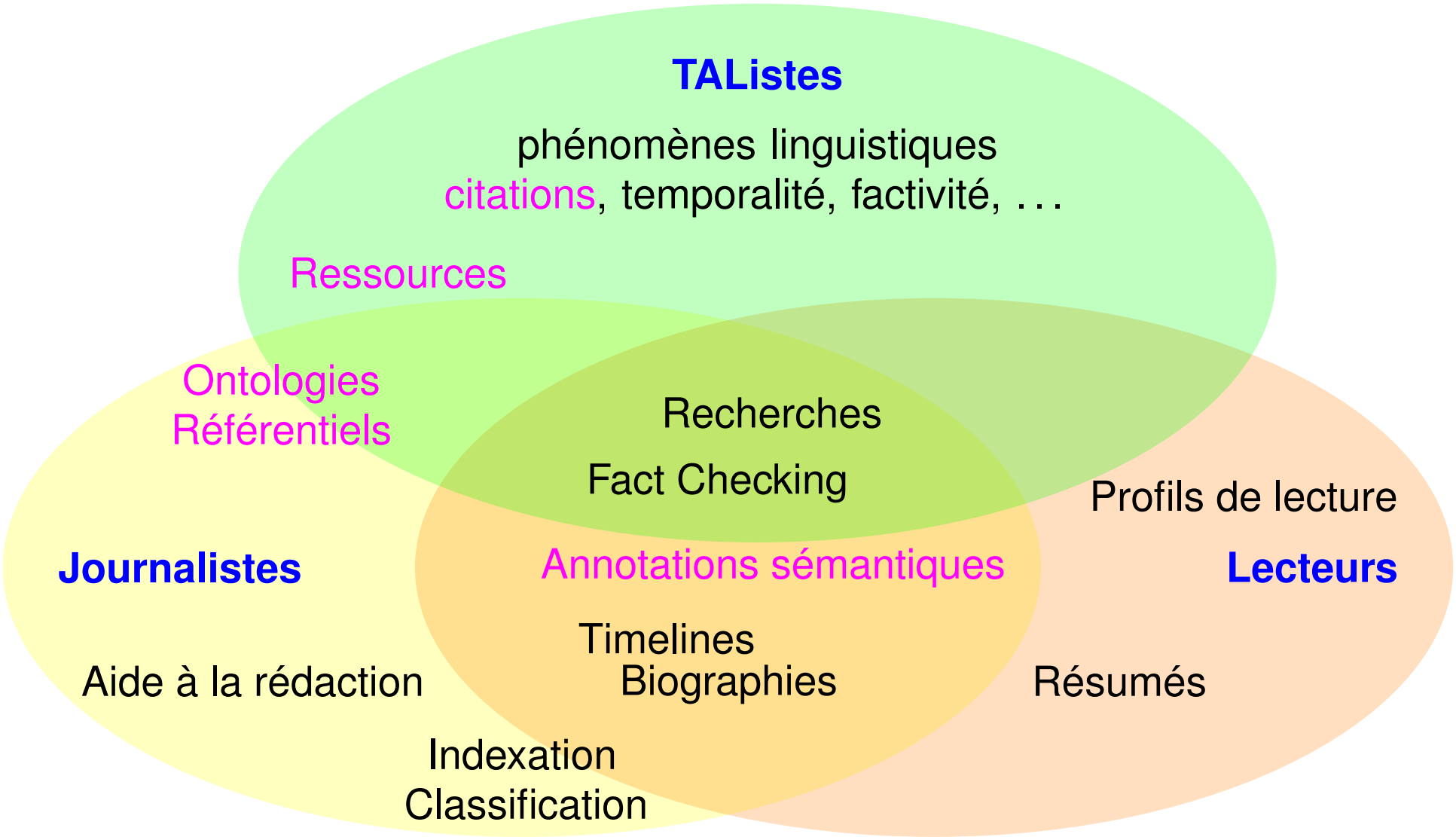
TABLE 1 – Nombre de nouveaux mots détectés par Alpage en fonction de la période selectionnée et du seuil d'occurrence minimal.

Système	4.1	4.1 Alpage	4.1 Oracle
WER	5.43%	5.16%	5.14%

TABLE 2 – Taux d'erreur de mot du système de base et des systèmes adapté, calculé sur le jeu de test AFP.

Au final, quelques pistes

Des applications potentielles vers au moins trois grands « *publics* » cibles



Conclusions et perspectives

Travailler sur le fond AFP : une magnifique opportunité

- une mine d'or pour les linguistes et TAListes
- de nouveaux services possibles pour les journalistes et lecteurs

NEWSPROCESS bien adapté pour traiter des flux (1K dépêches/j)
mais néanmoins arrêt des traitements des dépêches :

- question de la finalité du traitement d'un flux ?
  - pas uniquement pour remplir un disque avec des données !
  - doit s'inscrire dans le cadre d'une **plateforme** (API) pérenne pour ancrer diverses (et nombreuses) applications

- question des droits (diffusion et reproductibilité)
  dans l'idéal : possibilité de libérer des jeux de données
  (au moins pour évaluation et/ou usages académiques)

Merci de votre attention !