

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

Outils + Ressources linguistiques



Grammaire



Lexique LEFFP



Entités Nommées

ALEDage



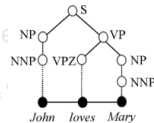
WordNet

WOLF

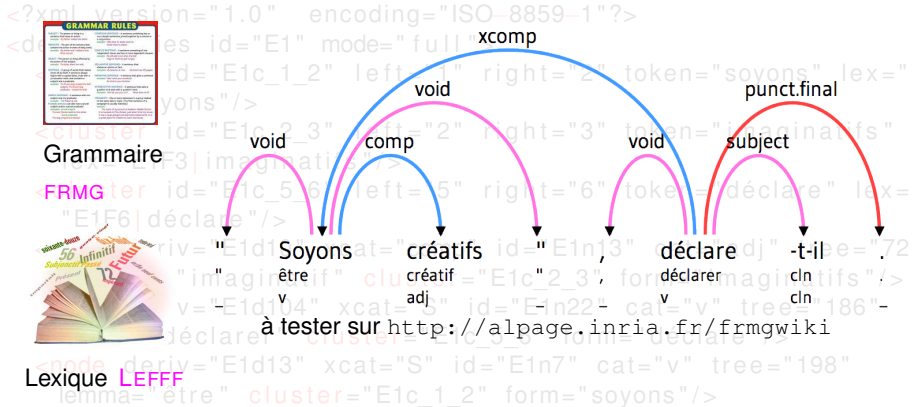


FrameNet

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)

2007

2008

2009

2010

2011

2012

2013



FUI SCRIBO

Annotations et Ontologies

ANR EDyLex

Enrichissement Dynamique de ressources Lexicales multilingues

Thèse Cifre Rosa Stern

Entités Nommées

named entities
knowledge acquisition
information extraction
semantic network
terms NLP
neologisms
citations
ontology
verbatim
annotated document
Semantic Web

- 1 Les dépêches en tant que documents
- 2 Les dépêches en tant que corpus
- 3 Les dépêches en tant que flux

Une dépêche (et enrichissements sémantiques)

Organisation des Nations unies
 United Nations
 الأمم المتحدة
 聯合國
 Организация Объединённых Наций
 Organización de las Naciones Unidas



Devise : Aucune officiellement



Carte des États membres de l'ONU.
 À noter que cette carte ne représente pas le point de vue de ses membres ou de l'ONU concernant le statut juridique des pays, ni ne relie que les zones gouvernementales qui ont un représentant de l'ONU.

Région	International
Création	26 juin 1945 : signature de la Charte des Nations unies, entrée en vigueur effective le 24 octobre 1945.
Type	Organisation intergouvernementale
Siège	Siège des Nations unies (Manhattan, New York) États-Unis
Coordonnées	40° 45' 00" N 73° 58' 03" W
Langue(s)	Anglais Arabe Chinois Espagnol Français Russe
Budget	Biennal 2010-2011 : 5,648 milliards de dollars des États-Unis
Membre(s)	193 États
Effectifs	Environ 44 000 (juin 2010)
Secrétaire général des Nations unies	Ban Ki-moon
Personne(s) clé(s)	Gladwyn Jebb
Site web	http://www.un.org/fr/

```
<metadata type="afp.com/metadata/concepts/Person"
resource="encyclo.org/Ban_Ki-moon">Ban Ki-moon</metadata>
```

```
<metadata type="afp.com/metadata/concepts/Organization"
resource="encyclo.org/ONU">ONU</metadata>
```

Ban Ki-moon (prononcé [ban ki] ; est un diplomate et homme d'État sud-coréen, né le 13 juin 1944 à Suwon) ; il est depuis le 14 août 2007, le 6^e actuel et huitième secrétaire général des Nations unies depuis le 17 janvier 2007. Il a été réélu à son poste le 21 juin 2011, jusqu'au 31 décembre 2016. Précédemment, il a été ministre des affaires étrangères du commerce, de son pays, de janvier 2004 au 10 novembre 2006. Il entre dans les services diplomatiques l'année où il est diplômé de son université, acceptant son premier poste à l'ONU, en Irak. Au sein du ministère des affaires étrangères il est chargé d'une réputation d'un homme modeste et compétent (en français). De langue maternelle coréenne, Ban parle couramment l'anglais, il parle également français et



Ban Ki-moon
반기문

En fonction depuis le 17 janvier 2007
 38 ans, 3 mois et 15 jours
 12 octobre 2006
 Election 21 juin 2011
 Prédateur KIM Annam
 50P ministre des Affaires étrangères de Corée du Sud
 17 janvier 2004 - 10 novembre 2006

L'Info | AFP.com
www.afp.com/fr/info/news/2013/01/08/1301081661.html

NEW YORK (Nations unies , 8 Jan 2013) (AFP) - Bahreïn: Ban déplore la confirmation des peines contre des opposants

Le secrétaire général de l'**ONU Ban Ki-moon** "regrette profondément" la confirmation par la justice des "peines sévères" prononcées contre des dirigeants de l'opposition à **Bahreïn**, a déclaré mardi le porte-parole de l'**ONU Martin Nesirky**.

M. **Ban**, a-t-il ajouté, "réaffirme qu'il considère que la seule manière de promouvoir la paix, la stabilité, la justice et la prospérité à **Bahreïn** est par un dialogue national qui réponde aux aspirations légitimes de tous les Bahreïnais et auquel toutes les communautés puissent participer librement, sans crainte ni intimidation".

Le secrétaire général des **Nations unies** appelle aussi le gouvernement de **Bahreïn** à tenir sa promesse de mener une "réforme judiciaire".

```
<metadata type="afp.com/metadata/concepts/Location"
resource="geo.org/Bahrein">Bahrein</metadata>
```



Kingdom of Bahrain (ca. 3 m)
 Al Bahrain, Al Bahrain, Bah, Bah, Ba, on, Baereini, Bc

Bahrain independent political entity
 population: 738000
 N 26° 0' 0" E 50° 30' 0"
 26.17 50.5
 GeoNameID: 239201

2000 2005 2010 2015 2020
 2010 2015 2020
 2010 2015 2020

Legend: Name, Country, Feature, Size in meters
 1 Kingdom of Bahrain, Bahrain, independent political entity, 0 km

Crédit : Rosa Stern

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

[Nuage par entité](#) - [Nuage filtré par mots clefs](#)

Abdul Wahaab Khetaab Adrian Edwards Agence France-Presse Ahmad Ahmadi Airbus Alain Carignon Alain Juppé Alain Marieix Alberto
Ali Shah Paktiawal André Glucksmann André Janier André Manoukian André Rossinot Anne Hidalgo Anne-Marie Comparini Ari
Ariette Laguilleur Arnaud Bollengier Arnaud Montebourg Arnaud Riverain Artigues Assemblée nationale Aubry-qui Élisabeth Guigou
Batasuna Benoit Rogeon Bernadette Chirac Bernard Accoyer Bernard Arnault Bernard Bosson Bernard Van Craeynest Bertra
Brice Hortefeux Bruno Tezenas Du Montcel Bruno Thouzellier Caisse nationale des associations familiales Carlo Garbagnati Cécili
centre de Toulouse centre de Toulouse CGT Chasse, pêche, nature et traditions Chirac Christiane Taubira Christophe Régnard Claud
Claudia Deeg CNI Commission des sondages Conseil supérieur de l'audiovisuel Corinne Lepage Corinne Lepage Cour des comptes
Daniel Cohn-Bendit Daniel Vaillant David Alphand de Force De l'autre côté Delphine Batho Didier Bariani Didier Maus Dominique
Dominique Moïsi Dominique Perben Dominique Reynié Dominique Strauss-Kahn Dominique Strauss-Khan Domini
Emmanuel Rivière de la Sofres Eric Beynel Etienne de Durand Europe 1 European Aeronautic Defence and Space Company Fabrice Hybert f
Fanny La Croix Firmin d'Amiens François Bayrou François Chérèque François de Grossouvre François Fillon François Fondard Franç
François Hollande François Miquet-Marty François Mitterrand François Rebsamen François Sauvadet Françoise de Panafieu Françoise
Françoise Laurent France 2 France 2 France Gamerre Francesco Guarguaglini Frédéric Dabi Frédéric Nihous Front national Générati

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



[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 territoriale 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 familiales (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 familiales
- ▶ Union pour la démocratie française
- ▶ Union pour un mouvement populaire

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

The screenshot shows the AFP website's search interface for election citations. At the top, a blue banner contains the AFP logo and the title "Elections 2012 : Recherche de Citations". Below this, there are two search input fields: "Recherche par auteurs" (with a subtext "Saisissez un prénom puis un nom") and "Recherche par mot-clé" (with a subtext "Saisissez un ou plusieurs mots clés"). A "Rechercher" button is positioned to the right of these fields. Below the search area, a "Recherche par date" section features a calendar-style interface with tabs for "2011" and "2012", and a range from "Jan" to "Déc". A "Comparateur" button is located on the right side of the interface. The main content area, titled "citations", displays a list of search results. The first result is dated "Du 01/01/2011 au 31/12/2012" and includes a snippet: "Deux plaintes contre des affiches 'piégées' de NKM dans l'Essonne 22/06/2012. Selon lui, ces affiches étaient apposées 'sur des poteaux électriques', sur des abribus dans la rue mais aussi à proximité de la mairie." The second result is dated "22/06/2012" and includes a snippet: "Deux plaintes contre des affiches 'piégées' de NKM dans l'Essonne 22/06/2012. 'Ca aurait pu paraître anodin, on enlève l'affiche on n'en parle plus, malheureusement il y avait un système de piège qui était actionné par une espèce de relais par une corde', a expliqué l'édile joint par l'AFP confirmant une information du Parisien.fr." The third result is dated "22/06/2012" and includes a snippet: "Deux plaintes contre des affiches 'piégées' de NKM dans l'Essonne 22/06/2012". On the left side, a sidebar titled "les principaux auteurs" lists several names, with "François Hollande" highlighted. On the right side, a vertical stack of three portrait photos is visible, with the top one being a woman and the others men.

NEWSPROCESS : un service HTTP de traitement de dépêches

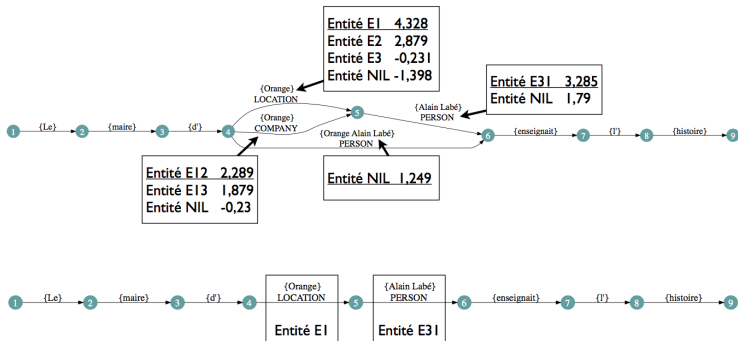
- implanté en Perl comme un pipeline de traitements
- chaque traitement ajoute une couche d'annotations (XML) et exploite les annotations des couches précédentes
 - 1 la dépêche sous format NewsML
 - 2 Segmentation (phrase/mots) et Reconnaissance Entités Nommées (REN) avec **SXPIPE** et **ALEDA**
 - 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

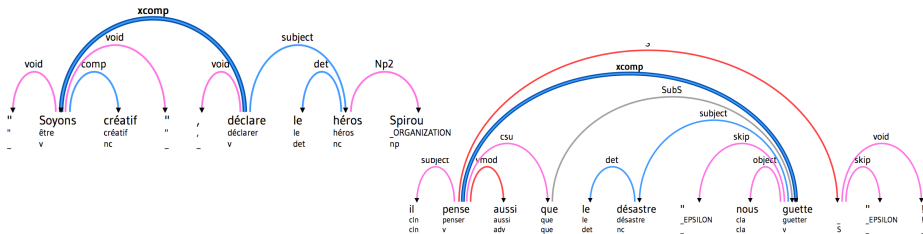
Tokenisation et Entités Nommées

- Découpage du texte en phrases et mots (SXPipe) en conservant les ambiguïtés 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, poursuit X**

Note : Citations et Verbatim sont deux notions distinctes

- 1 Les dépêches en tant que documents
- 2 Les dépêches en tant que corpus
- 3 Les dépêches en tant que flux

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	

⇒ source utile pour acquérir des connaissances linguistiques et *ontologiques*

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)

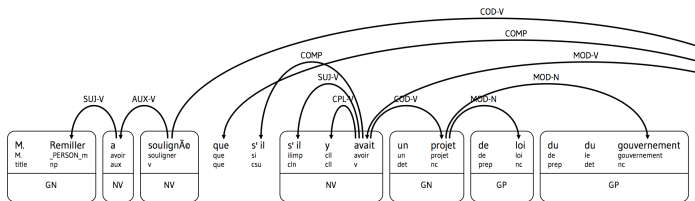
Évènements

- Regroupement de verbes, dénotant un type d'évènement
 - ▶ /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 objets, sujets, ...

⇒ 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'état
secrétaire d'état
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

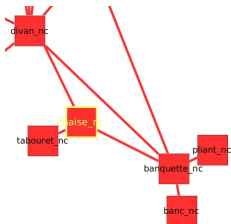
Meanings of words are (largely) determined by their distributional patterns (Harris 1968)

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

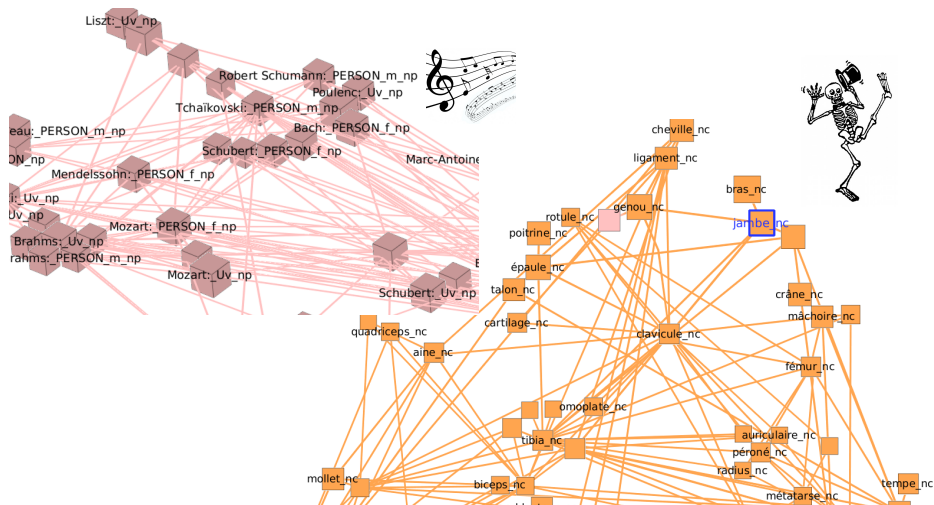


	chaise divan	chaise tabouret	banquette divan	banquette canapé	banquette chaise
se asseoir sur [•]	●	●	●	●	●
asseoir sur [•]	●	●	●	●	●
allonger sur [•]	●		●	●	
dormir sur [•]	●		●	●	●
tomber sur [•]	●		●	●	●
monter sur [•]		●			●
place sur [•]					
grimper sur [•]		●			●

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

Graphe d'environ 40K connections

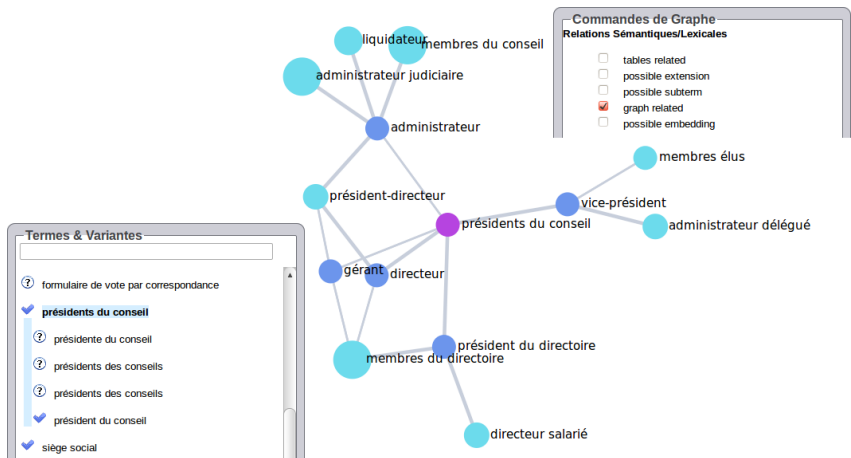
Visualisation avec TULIP (<http://tulip.labri.fr/>),



Vu les volumes, complexité et erreurs,
vrai besoin de :

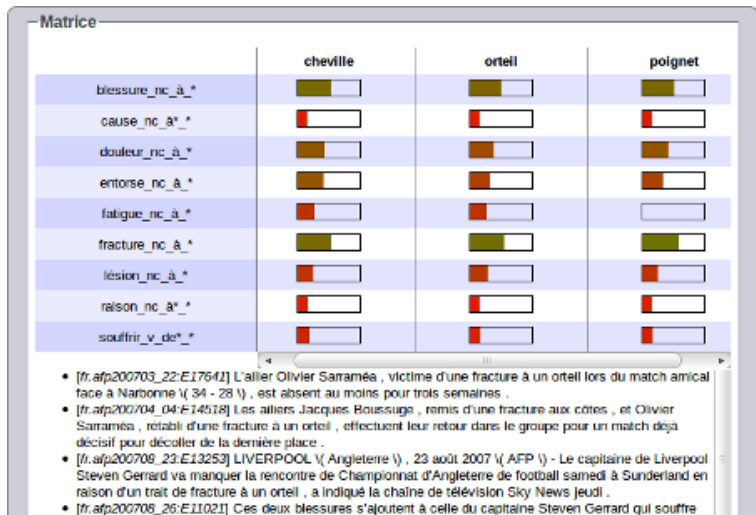
- visualisation riche, locale (zoom), mais sans surcharge
- mécanismes simples de navigation et de recherche
- accès à des explications et exemples
- validation collaborative
 - ▶ ressources imparfaites
 - ▶ maintenance et évolutions
 - ▶ effort de validation important devant être distribué
 - ▶ échanges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

⇒ développement d'une interface WEB sur la plateforme **LIBELLEX**
en collaboration avec Lingua et Machina.

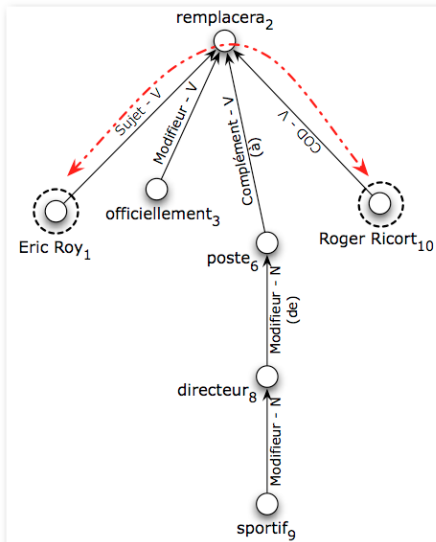


<http://alpage.inria.fr/Lbx> (guest/guest)

Les rapprochements explicables par des contextes pondérés



- Extraction des chemins de dépendances entre ENs
- Hypothèse** : les chemins représentent la relation entre les ENs
~> 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



chemins \rightsquigarrow Couples EN

http://alpage.inria.fr/~nakamura/afp0520_0729_5_ch
http://alpage.inria.fr/~nakamura/afp0520_0729_5_chemins.xml

Chemin : [41] ==> (MOD-N) président (nc) <== (MOD-N(de))

- [27] François Bayrou - MoDem
- [2] Jacques Rogge - Comité international olympique
- [3] Jacques Delors - Commission
- [1] Philippe de Villiers - MPF
- [9] Jean-Claude Trichet - BCE
- [1] Fahey - de l'AMA John
- [1] Elio Di Rupo - PS
- [2] Allain Bougrain-Dubourg - LPO
- [6] Hervé Morin - Nouveau Centre
- [4] François Bayrou - Modem
- [1] Mouammar Kadhafi - l'Union
- [1] François Bayrou - Mouvement Démocrate
- [4] Jean-Claude Trichet - Banque centrale
- [1] Jérôme Lejeune - Fondation
- [14] Jean-Paul Bailly - La Poste
- [4] Poul Nyrup Rasmussen - Parti socialiste
- [27] Dominique Sopo - SOS Racisme
- [1] Pascal Colombani - conseil d'administration
- [2] Jean-Luc Mélenchon - Parti de gauche
- [2] Romano Prodi - Commission
- [6] Jean-Paul Huchon - conseil régional
- [3] José Sarney - Sénat
- [2] Roger Karoutchi - l'Assemblée

couples EN \rightsquigarrow Chemins

http://alpage.inria.fr/~nakamura/afp0520_0729_5_ENs.xml
http://alpage.inria.fr/~nakamura/afp0520_0729_5_ENs.xml

Relation : Ban Ki-moon - l'ONUL (254)

organization

[244] ==>

[1] <==

[4] ==>coréen<==

[2] ==>sud<==

[1] ==>appelle<==Conseil<==

[1] ==>lettre==>secrétaire<==

[1] ==>fait<==unanimité<==patron<==

Relation : José Manuel Barroso - Commission (168)

[3] <==

[41] <==président<==

[14] ==>président<==

[1] ==>président<==

- 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
⇒ 149 correctes (83.7%)



http://alpage.inria.fr/~nakamura/afp0520_0729_resultat.xml

http://alpage.inria.fr/~nakamura/afp0520_0729_resultat.xml

Couples EN en relation d'appartenance : (rouge = individual, bleu = organization)

- Xavier Bertrand - UMP
- Nicolas Sarkozy - UMP
- Nicolas Sarkozy - New Delhi
- Mahamane Ousmane - CDS
- Calderon - Parti d'action
- Lazarus Murendo - tribunal de première instance
- Fredrik Reinfeldt - Commission
- Roland Koch - Deutschlandfunk
- Franco Frattini - Rai Uno
- Devedjian - BFM
- Sébastien Delahaye - CFDT
- M. Berlusconi - Rai
- Ballack - Bild
- M. Berlusconi - Rai Uno
- Malliy - BFM
- Batho - UMP
- M. Roche - BBC
- M. Obama - CBS News
- Moscovici - BFM
- Pascal Baudouin - CGT
- Fred Irwin - Handelsblatt
- Dubus - Paris
- Huchon - Paris
- René Raimondi - PS
- Biden - ABC

- 1 Les dépêches en tant que documents
- 2 Les dépêches en tant que corpus
- 3 Les dépêches en tant que flux

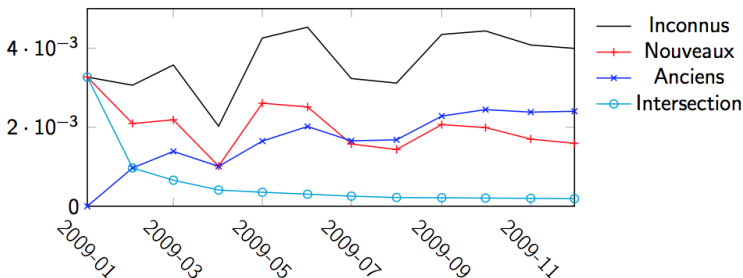
Flux pour suivre les évolutions du langage

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



↪ un inconnu distinct tous les ~ 1000 tokens

Comment distinguer les nouveaux termes/entités de types ?

distance d'édition, quotes, répétitions, sources externes, paradigmes, emprunts, ...

Exemples (janvier 2013) :

- (termes) cryothérapie, lumino-technique, narcoterroriste, co-attribué, pro-putsch, graffitiste, galvanisante
 - ▶ **cryothérapie** : Wiktionary (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 ?

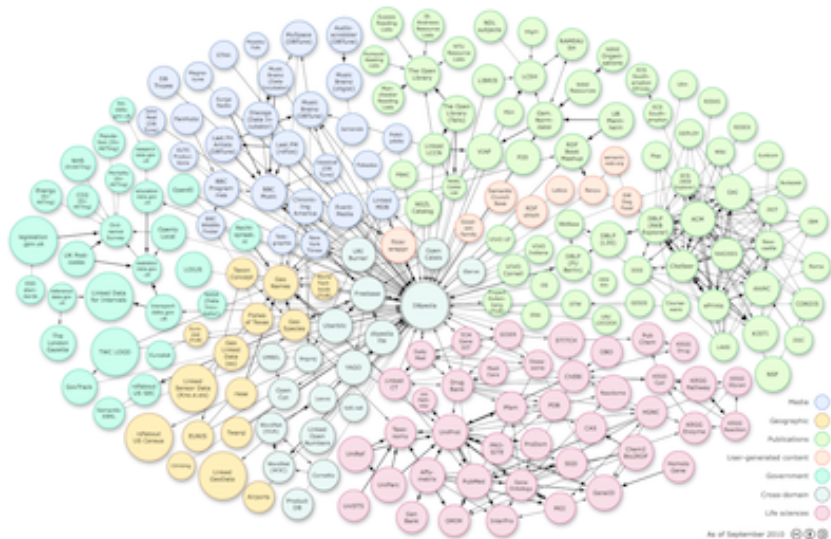
Référentiel AFP d'Entités Nommées

Class hierarchy:



besoin de constituer et maintenir un référentiel

LOD : des sources de données connectées et accessibles



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
304158	41543	18109	465926
PRODUIT	ŒUVRE	FICTIONCHAR	Total
2526	83713	6729	922704

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 \cap SXPIPE) + NOMOS	92.95	72.82	81.66	91,96	75,10

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

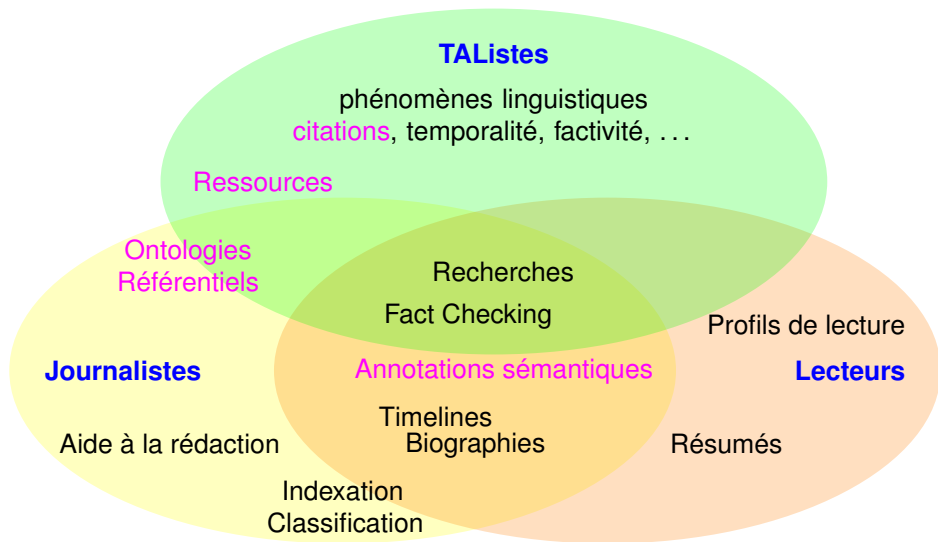
TABLE 1 – Nombre de nouveaux mots détectés par Alpage en fonction de la période sélectionnée et du seuil d'occurrence minimal.

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 adaptés, calculé sur le jeu de test AFP.

Au final, quelques pistes

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



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 !