

Exemples d'application du Traitement Automatique des Langues en SHS

**Identifier dans un corpus des acteurs,
des concepts et les relations entre eux**

Pablo Ruiz Fabo, LATTICE

*Collecter et produire des données pour la recherche en SHS
Fréjus, 15-18 novembre 2016*

1. Trouver des Acteurs et Concepts dans un corpus
2. Trouver les relations entre eux

- Technologie
- Comment évaluer les résultats
- Démo d'intégration avec interface utilisateur

- **Trouver des Acteurs et Concepts dans un corpus :**

Technologie : Résolution référentielle des entités nommées et concepts (Entity Linking)

Entity Linking (Résolution référentielle des entités)

- **Tâche**: Repérer des **concepts** d'une base de connaissances (comme Wikipédia) dans un corpus à travers de **mentions** qui font référence à ces concepts
 - Ex. acteurs du corpus, sujets abordés ...

- Établir des **liens** entre les éléments du corpus à travers leurs concepts en commun *malgré la variabilité dans les mentions*
 - Navigation par concept, visualisation en réseau
- Point d'entrée au **web sémantique** à travers des annotations DBpedia
- **Complément** à extraction de terminologie (prés. précédente (Ferguth))

Entity Linking : Étapes

- **Reconnaissance** des mentions
 - Reconnaissance des entités nommées
- **Désambiguïsation** des mentions vers leur concept (entité) dans la base de connaissances (ou ontologie)
- **Ontologie** ? Hiérarchie de concepts et les relations entre ceux-ci

Reconnaissance des entités nommées (REN)

- **Entités ? Concepts ?**

- Séquences textuelles appartenant à des types spécifiques (ex. *personnes, organisations, lieux, produits, substances, autres*)
- Exemple d'un inventaire de types courant: **Ontologie DBpedia** (classes)
- Inventaire classique de types:
Entités étendues de **Sekine** (2003)
(Poibeau, 2011)

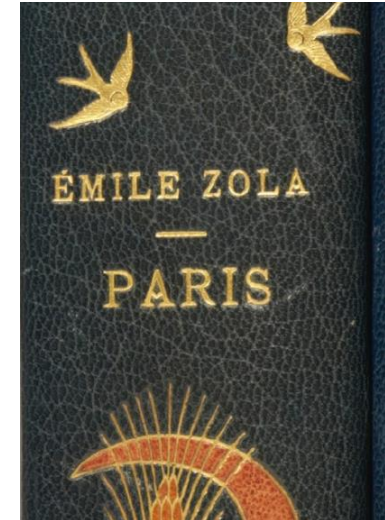
1. **Mentions ambiguës** entre plusieurs entités possibles (*homonymie*)



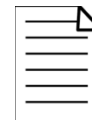
1. Mentions ambiguës entre plusieurs entités possibles (*homonymie*)

1 Liste des villes s'appelant Paris

- 1.1 France
- 1.2 Danemark
- 1.3 États-Unis
- 1.4 Canada
- 1.5 Gabon
- 1.6 Panama
- 1.7 Russie
- 1.8 Suède



Paris



Entity Linking : Défis

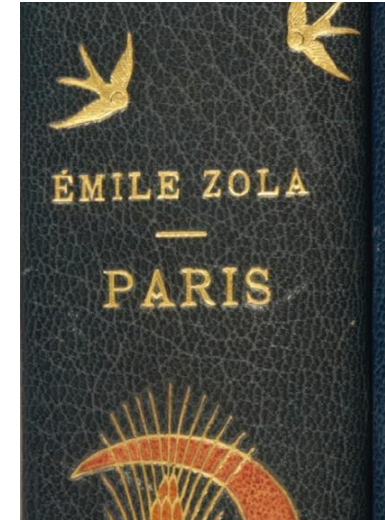
1. Mentions ambiguës entre plusieurs entités possibles (*homonymie*)

1 Liste des villes s'appelant Paris

- 1.1 France
- 1.2 Danemark
- 1.3 États-Unis
- 1.4 Canada
- 1.5 Gabon
- 1.6 Panama
- 1.7 Russie
- 1.8 Suède



album de The Cure?



roman de Zola?

une ville?

Paris

Entity Linking : Défis

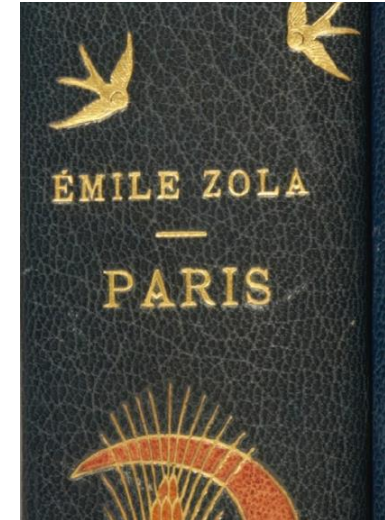
1. Mentions ambiguës entre plusieurs entités possibles (*homonymie*)

1 Liste des villes s'appelant Paris

- 1.1 France
- 1.2 Danemark
- 1.3 États-Unis
- 1.4 Canada
- 1.5 Gabon
- 1.6 Panama
- 1.7 Russie
- 1.8 Suède



album de The Cure?



roman de Zola?

une ville?

Paris

2. Plusieurs mentions possibles pour chaque entité (*synonymie*)

Barack Hussein Obama, Barack Obama, Obama, M. Obama, Président Obama, POTUS, #POTUS

Entity Linking : Un autre exemple d'homonymie

1. Exemple de Hoffart et al. (2011) Mentions ambiguës entre plusieurs entités possibles

Hoffart et al. 2011



When **Page** played **Kashmir** at Knebworth, his Les Paul was uniquely tuned



When **Page** played **Kashmir** at **Knebworth**, his **Les Paul** was uniquely tuned

Page: Larry vs. Jimmy

Kashmir: Région vs. chanson

Knebworth: Festival vs. ville

Les Paul: guitarre vs. musicien

wikipedia**miner**

2008

TAGME

2010

DBpedia**Spotlight**

2011

aida

2011

de**x**ter

2014

Babel**fy**

2014

- Outils publics qui désambigüisent vers des **ontologies générales** (DBpedia, YAGO, Babelnet)
- Outils à **code ouvert** (peut promouvoir l'interprétabilité des résultats)
[Note: Le code de Babelfy n'est pas ouvert, la base Babelnet oui]
- **Services web** pour déploiement local, et, en général, services web hébergés publiquement

SOCCER –JAPAN GET LUCKY WIN, CHINA IN SURPRISE DEFEAT

```
{
  "end": 13,
  "title": "Japan national football team",
  "spot": "JAPAN",
  "start": 8,
  "rho": "0.42172",
  "dbpedia_categories": [
    "Japan national football team",
    "AFC Asian Cup-winning countries",
    "Asian national association football teams",
    "National sports teams of Japan"
  ],
  "id": 993546
},
```

SOCCER – *JAPAN* GET LUCKY WIN, CHINA IN
SURPRISE DEFEAT

```
{  
  "end": 13,  
  "title": "Japan national football team",  
  "spot": "JAPAN",  
  "start": 8,  
  "rho": "0.42172",  
  "dbpedia_categories": [  
    "Japan national football team",  
    "AFC Asian Cup-winning countries",  
    "Asian national association football teams",  
    "National sports teams of Japan"  
  ],  
  "id": 993546  
},
```

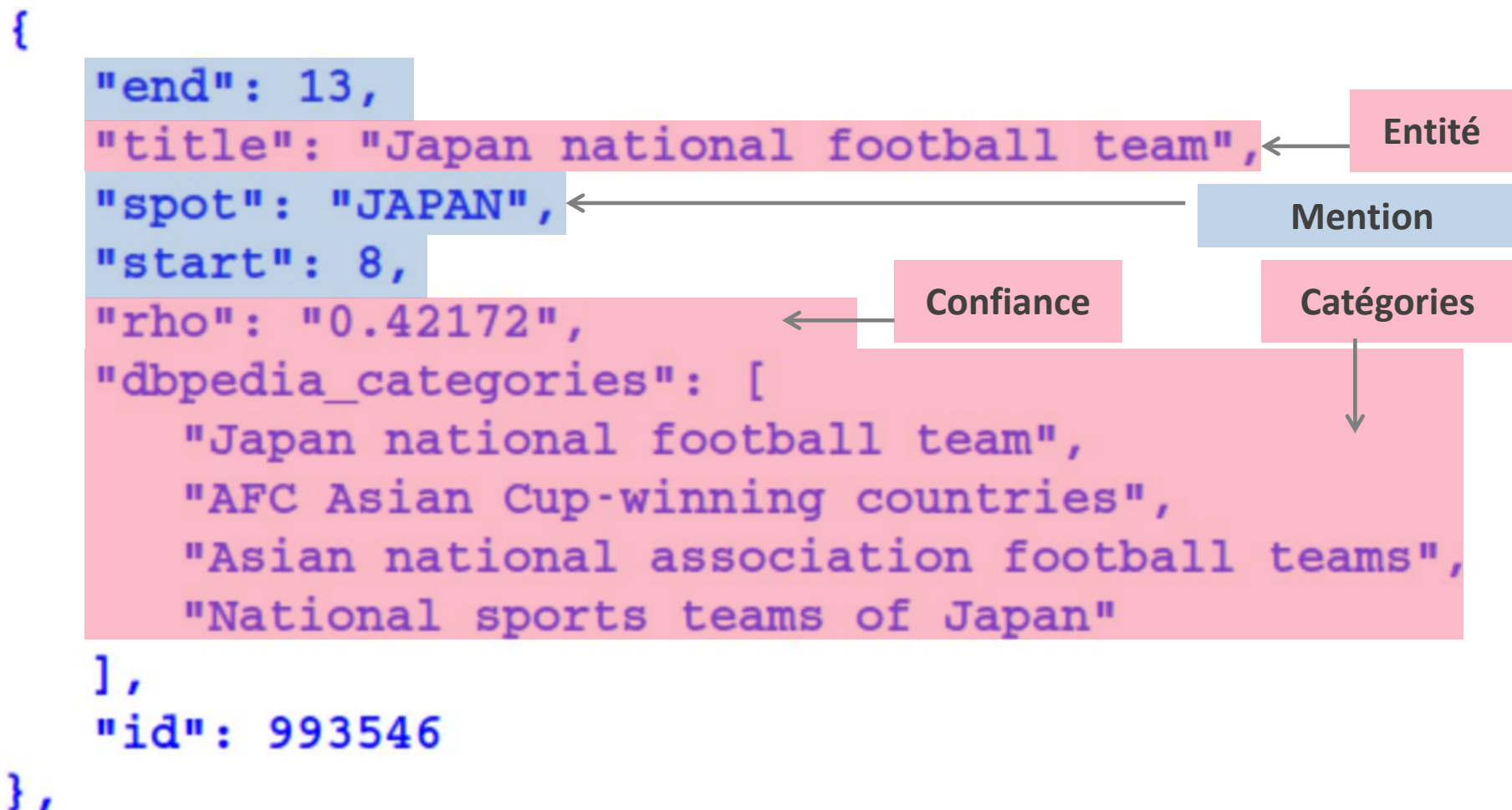
← Mention

SOCCER –JAPAN GET LUCKY WIN, CHINA IN SURPRISE DEFEAT

```
{
  "end": 13,
  "title": "Japan national football team",
  "spot": "JAPAN",
  "start": 8,
  "rho": "0.42172",
  "dbpedia_categories": [
    "Japan national football team",
    "AFC Asian Cup-winning countries",
    "Asian national association football teams",
    "National sports teams of Japan"
  ],
  "id": 993546
},
```

← Mention

SOCCER – **JAPAN** GET LUCKY WIN, CHINA IN SURPRISE DEFEAT



- Comparaison des **contextes** d'une mention dans le corpus et du texte relié aux entités dans la base de connaissance (tokens, liens)
- Mesure de **cohérence** entre les désambiguïsations proposées à l'intérieur d'un document (*liens communs Wikipédia*)
 - Milne & Witten (2008)
 - Ferragina et al. (2010)
 - Hoffart et al. (2011)
 - Moro et al. (2014)

wikipedia**miner**



aida

Babelfy

- Tester avec le corpus visé (la qualité des résultats pour chaque outil varie selon le corpus, Cornolti et al. 2013, Usbeck et al. 2015)
- **Métriques** de TAL. Outils :
 - Neleval : <https://github.com/wikilinks/neleval>
 - BAT Framework : github.com/marcocor/bat-framework
- Évaluation par les **experts du domaine** : Impact (ou pas) des erreurs pour les chercheurs les utilisant
- **Aspects pratiques** : licence, déploiement ...

EL : Démo Corpus PoliInformatics



2010



2011

wikipedia**miner**

2008

aida

2011

Babelfy

2013

résultats affichés sur la démo

pas affichés

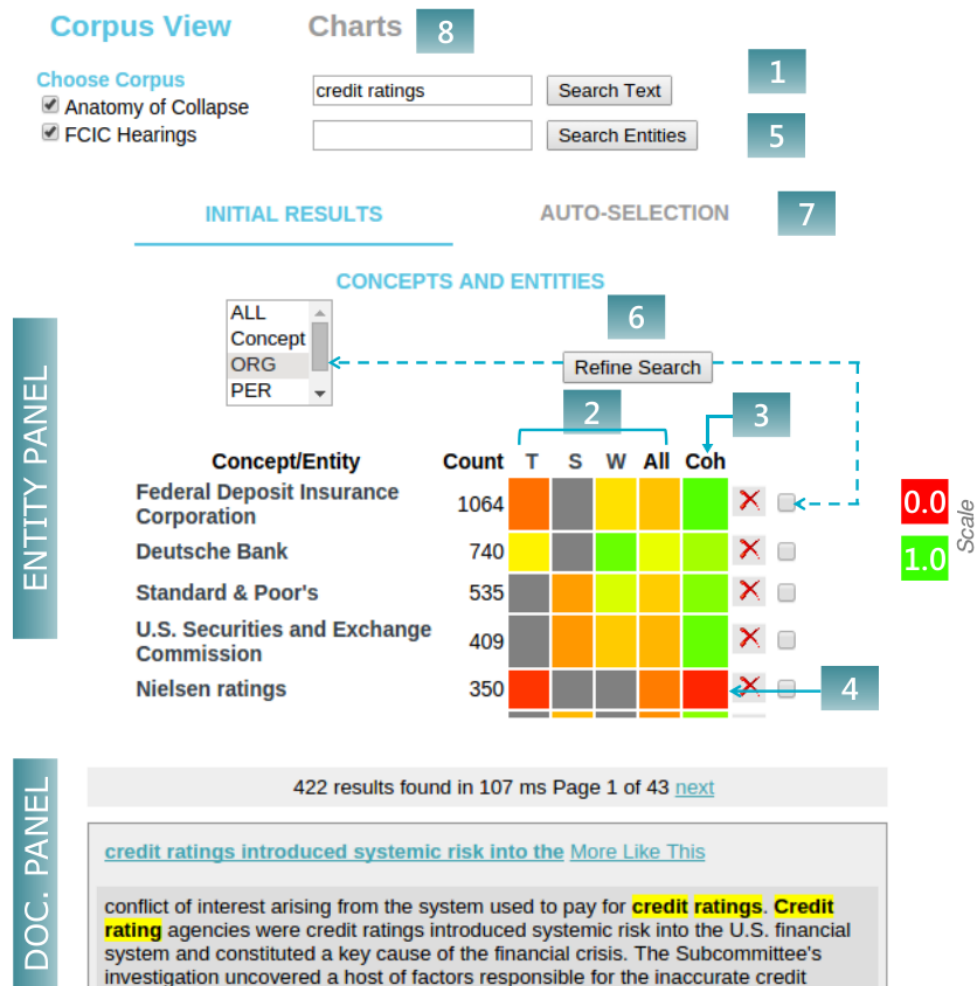
Ruiz, Mélanie, Poibeau (2015), NAACL Démo

- **Combinaison d'outils** pour aborder variabilité des corpus
- **Aborder besoins** des utilisateurs:
 - Veulent garder un certain contrôle sur sortie des outils et sélectionner des concepts à la main =>
 - **Informations** pour guider un **filtrage** manuel des entités.

EL : UI démo Corpus PoliInformatics

Corpus: subset of PoliInformatics (Smith et al. 2014), about 2008 US financial crisis

- (1) Query via **Search Text** displays:
 - **Document Panel**: Documents matching the query
 - **Entity Panel**: Entities extracted in the documents matching the query displayed on doc. panel, plus:
 - (2) Confidence Scores for each annotator, normalized to a 0-1 range. (T=Tagme, S=Spotlight, W=Wiki Miner).
 - (3) Coherence score between the entity and a representative subset of the corpus entities.
 - (4) Entities not coherent with the corpus flagged in red.
- (5) Query via **Search Entities** displays:
 - **Entity Panel**: Entities matching the query.
 - **Document Panel**: Documents containing one of the entities displayed on the entity panel.
- (6) **Refine Search**: Entities can be selected with a list of types (like **ORG**) or selected individually with checkboxes.
- (7) The **Auto-Selection** tab shows the output of an automatic filtering via weighted voting of annotations.
- (8) **Charts**: examples of co-occurrence networks, created offline exploiting workflow information (sentence number, confidence, ...)

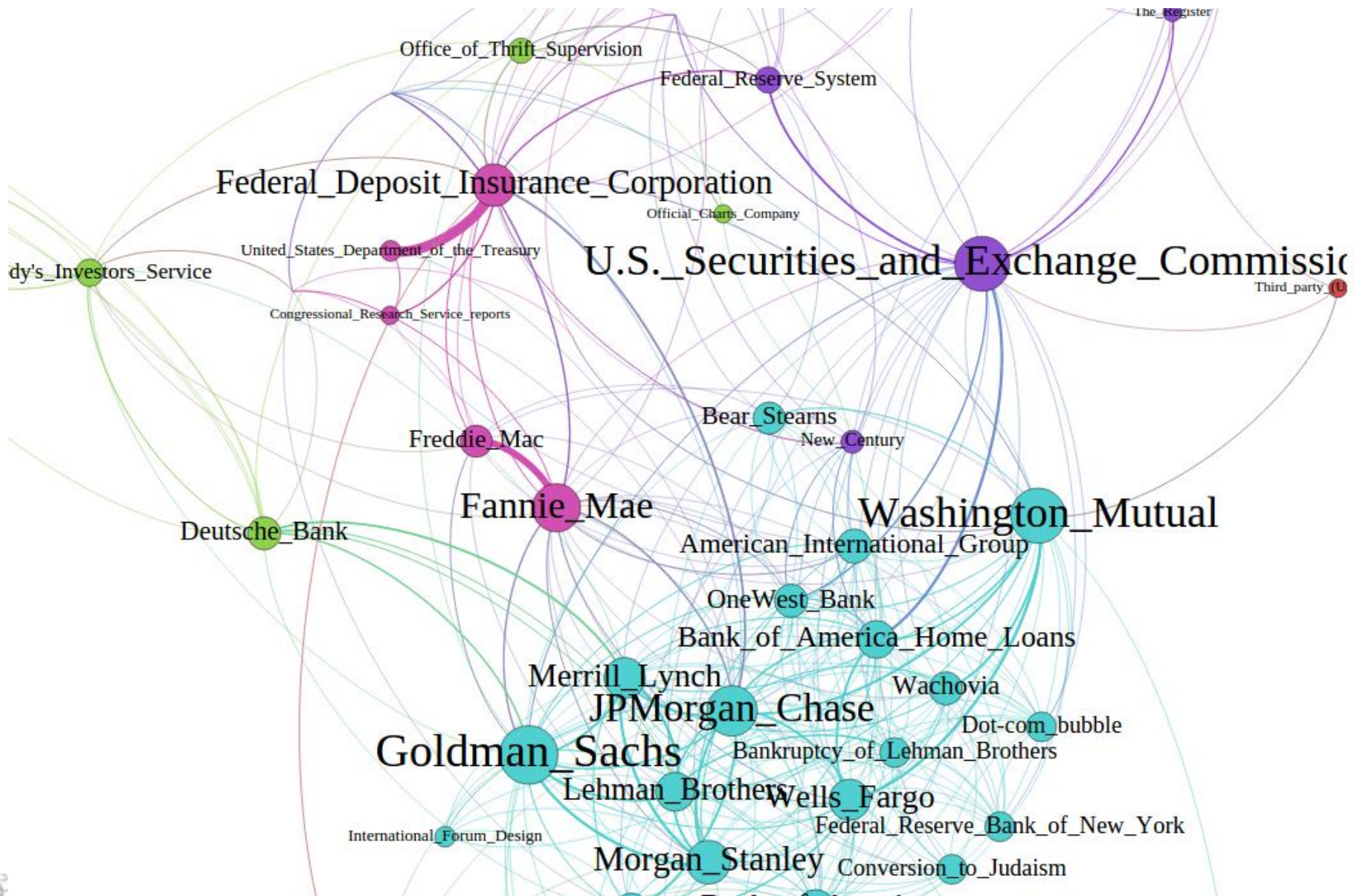


Ruiz, Mélanie, Poibeau (2015), NAACL Démo [<http://apps.lattice.cnrs.fr>]

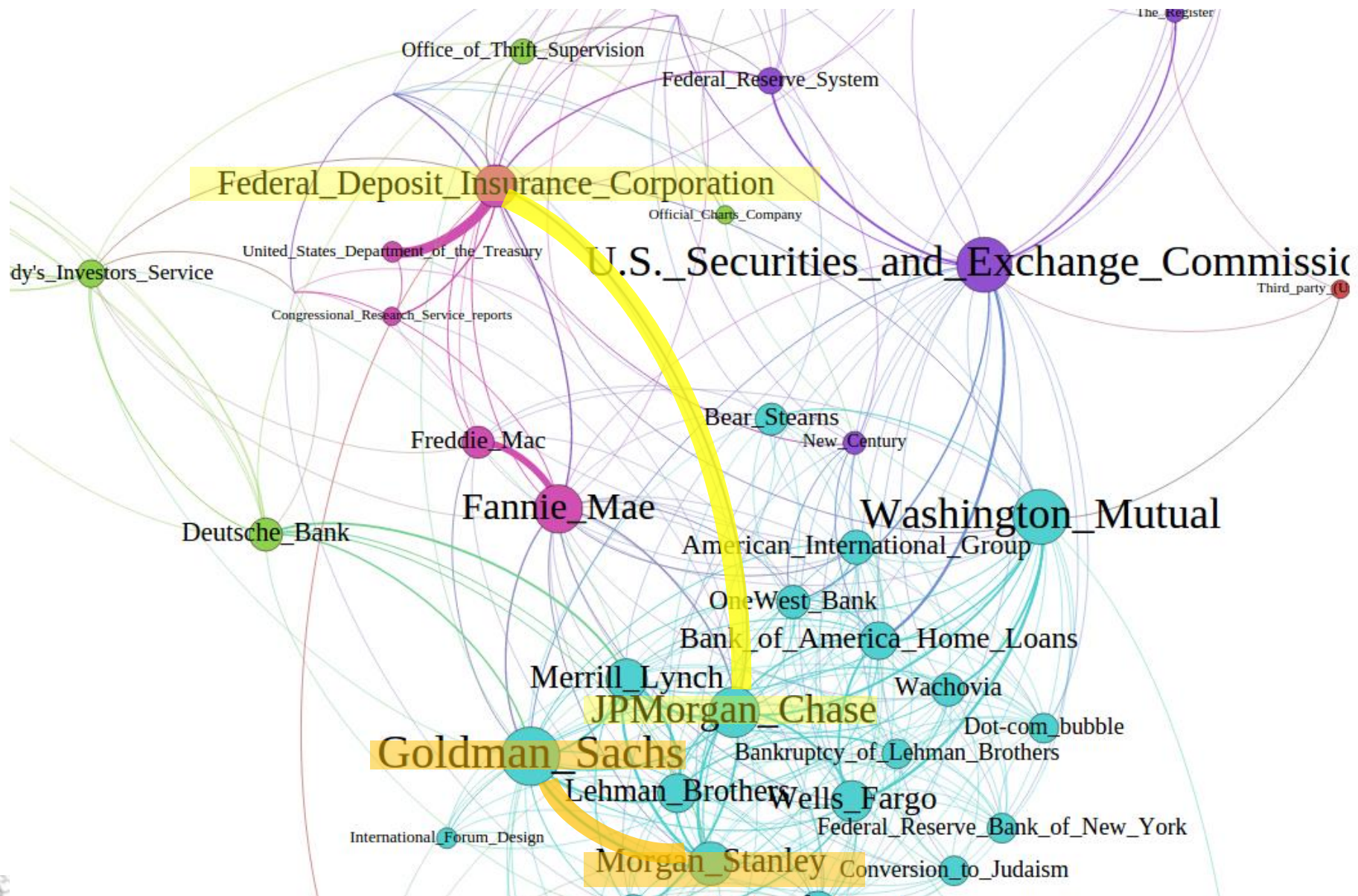
1. Trouver des Acteurs et Concepts dans un corpus
2. Trouver les relations entre eux

- Technologie
- Comment évaluer les résultats
- Démo d'intégration avec interface utilisateur

Relations entre acteurs et concepts



Relations entre acteurs et concepts



Relations entre acteurs et concepts : Extraction de propositions

- **Tâche**
 - Identification des **concepts** et **acteurs** en cooccurrence, et des **prédicats** qui relient ces concepts et auteurs
- **Corpus** : *Earth Negotiations Bulletin*
 - **Sujet** : Négociations internationales sur le changement climatique (comme la COP-21 qui a eu lieu à Paris)
- **Utilité** : Aide à identifier les options proposées par les acteurs dans les négociations, et des éléments auxquels les acteurs s'opposent

Qui a dit quoi et comment ?



30 NOVEMBRE - 11 DÉCEMBRE 2015

21^E CONFÉRENCE DES NATIONS UNIES SUR LE CHANGEMENT CLIMATIQUE

TOUS ENSEMBLE POUR LE CLIMAT

30 NOVEMBER - 11 DECEMBER 2015

21ST UNITED NATIONS CLIMATE CHANGE CONFERENCE

UNITED FOR CLIMATE ACTION

cop21.gouv.fr

The EU, with NEW ZEALAND and opposed by CHINA, MALAYSIA and BHUTAN, supported including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."

www.iisd.ca/enbvol/enb-background.htm



Earth Negotiations Bulletin

A Reporting Service for Environment and Sustainable Development Negotiations

Acteurs (ou pays)

The **EU**, with **NEW ZEALAND** and opposed by **CHINA**, **MALAYSIA** and **BHUTAN**, supported including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."

Message (objet de la négociation)

The EU, with NEW ZEALAND and opposed by CHINA, MALAYSIA and BHUTAN, supported

including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."

Message (objet de la négociation)

The EU, with NEW ZEALAND and opposed by CHINA, MALAYSIA and BHUTAN, supported *including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."*

Prédicats (soutien/opposition)

The EU, with NEW ZEALAND and **opposed** by CHINA, MALAYSIA and BHUTAN, **supported** including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."

Acteurs + prédicats + message

The EU, with NEW ZEALAND and **opposed** by CHINA, MALAYSIA and BHUTAN, **supported** *including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."*

Acteur + Prédicat + Message = Proposition

ACTEURS	PRÉDICATS	MESSAGE
European_Union	supported	including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."
New_Zealand		
China	~supported	
Malaysia		
Bhutan		

- Exemple avec prédicat nominal

	ACTEURS	PRÉDICATS NOMINAUX	MESSAGE
1	Group_of_77 / China	proposal	to include research and development in the transport and energy sectors in the priority areas to be financed by the SCCF.

- Pipeline de TAL (<http://ixa2.si.ehu.es/ixa-pipes/>)
- Modèle du domaine : Base de connaissance contenant
 - Noms des acteurs
 - Série de prédicats d'énonciation

PLUS DE DÉTAILS

<https://sites.google.com/site/nlp4climate/>

- Pipeline de TAL (IXA Pipeline):
 - **Analyse syntaxique (parsing) en dépendances**
 - Sujets pour prédicats verbaux
 - **Étiquetage des rôles sémantiques (SRL)**
 - *Agent Prédicat Message* [PropBank et NomBank]
 - Rôle *Négation* (complémenté avec de règles)
 - **Résolution des anaphores pronominales**

With Benin and Zimbabwe, China cautioned against the COP losing focus on the Convention. **He** opposed the argument that ...

- Messages enrichis avec des **keyphrases**, **DBpedia** et un **thésaurus** pour le climat

- Pipeline de TAL (IXA Pipeline):
 - Analyse syntaxique (parsing) en dépendances
 - Sujets pour prédicats verbaux
 - Étiquetage des rôles sémantiques (SRL)
 - *Agent Prédicat Message* [PropBank et NomBank]
 - Rôle *Négation* (complémenté avec de règles)
 - **Résolution des anaphores pronominales**

With Benin and Zimbabwe, China cautioned against the COP losing focus on the Convention. **He** opposed the argument that ...

- Messages enrichis avec des **keyphrases**, **DBpedia** et un **thésaurus** pour le climat

- Pipeline de TAL (IXA Pipeline):
 - Analyse syntaxique (parsing) en dépendances
 - Sujets pour prédicats verbaux
 - Étiquetage des rôles sémantiques (SRL)
 - *Agent Prédicat Message* [PropBank et NomBank]
 - Rôle *Négation* (complémenté avec de règles)
 - **Résolution des anaphores pronominales**

With Benin[?] and Zimbabwe[?], China[?] cautioned against the COP losing focus on the Convention. He[?] opposed the argument that ...
- Messages enrichis avec des **keyphrases**, **DBpedia** et un **thésaurus** pour le climat

- Pipeline de TAL (IXA Pipeline):
 - Analyse syntaxique (parsing) en dépendances
 - Sujets pour prédicats verbaux
 - Étiquetage des rôles sémantiques (SRL)
 - *Agent Prédicat Message* [PropBank et NomBank]
 - Rôle *Négation* (complémenté avec de règles)
 - **Résolution des anaphores pronominales**

With Benin and Zimbabwe, China cautioned against the COP losing focus on the Convention. **[He => China]** opposed the argument that ...

- Messages enrichis avec des **keyphrases**, **DBpedia** et un **thésaurus** pour le climat

- Pipeline de TAL (IXA Pipeline):
 - Analyse syntaxique (parsing) en dépendances
 - Sujets pour prédicats verbaux
 - Étiquetage des rôles sémantiques (SRL)
 - *Agent Prédicat Message* [PropBank et NomBank]
 - Rôle *Négation* (complémenté avec de règles)
 - Résolution des anaphores pronominales

With Benin and Zimbabwe, China cautioned against the COP losing focus on the Convention. [He => China] opposed the argument that ...

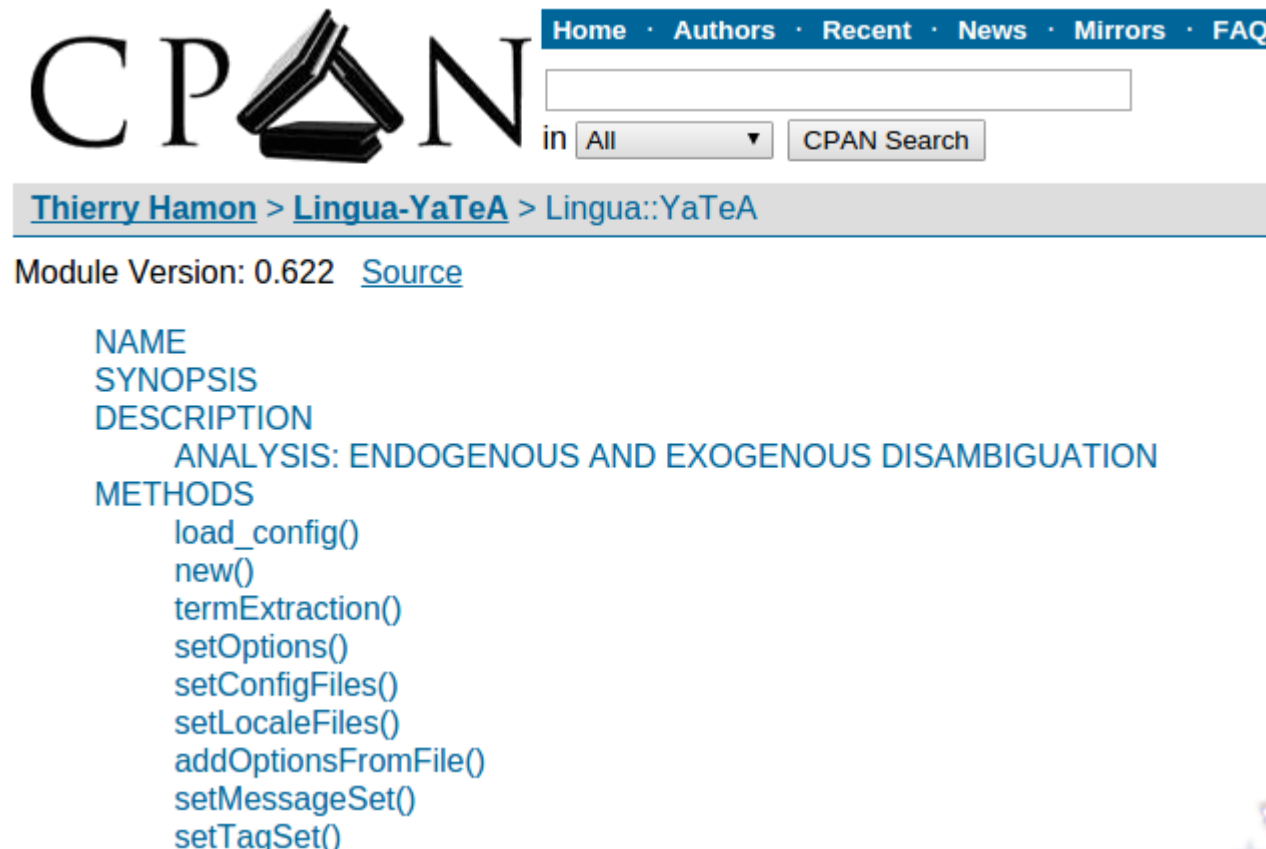
- Messages enrichis avec des **keyphrases**, **DBpedia** et un **thésaurus** pour le climat

- Modèle du domaine
 - Acteurs dans la négociation
 - Prédicats d'énonciation
 - Nominaux (NomBank)
 - Verbaux (VerbNet, chez PropBank)

PLUS DE DÉTAILS

<https://sites.google.com/site/nlp4climate/>

- YaTeA (Aubin et Hamon, 2006)



The screenshot shows the CPAN (Comprehensive Perl Archive Network) website interface. At the top, the CPAN logo is displayed with a navigation bar containing links: Home, Authors, Recent, News, Mirrors, and FAQ. Below the logo is a search bar with a dropdown menu set to 'All' and a 'CPAN Search' button. The main content area shows the breadcrumb path: [Thierry Hamon](#) > [Lingua-YaTeA](#) > [Lingua::YaTeA](#). Below this, the module version is listed as 0.622, with a link to the [Source](#). The module's metadata is listed in a tree structure: NAME, SYNOPSIS, DESCRIPTION (with the text 'ANALYSIS: ENDOGENOUS AND EXOGENOUS DISAMBIGUATION'), and METHODS. The methods listed are: load_config(), new(), termExtraction(), setOptions(), setConfigFiles(), setLocaleFiles(), addOptionsFromFile(), setMessageSet(), and setTagSet().

- Métriques **TAL**
- Utilité pour les **experts du domaine**

- Annoté à la main deux ensembles d'approx. 300 et 550 propositions
- Résultat correct : Coïncidence exacte des trois composantes de la proposition (acteur, prédicat, message)

Corpus	Phrases	Props	F1	P	R
ENB-COP	100	311	0.69	0.687	0.693
ENB-IPCC	283	566	0.718	0.714	0.722

Évaluation par les experts

- Évaluation **qualitative**
- **Entretiens** avec des experts du domaine
 - Montrer exemples utilisation UI
 - Demander d'utiliser UI pour tester des hypothèses ou trouver des renseignements sur le corpus
- Aspects évalués
 - UI utile pour fournir **vue d'ensemble** du corpus ?
 - Extractions automatiques **factuellement correctes** ?
 - **Nouvelles idées** obtenues grâce à usage UI ?

Évaluation par les experts

- **Vue d'ensemble :**
 - Oui, mais UI plus utile quand on a une question de départ
 - Niveau de détail très satisfactoire
- **Qualité factuelle :**
 - Extractions ont un sense, qualité suffisante pour l'exploration du corpus
- **Nouvelles idées ?**
 - Il y a eu quelques exemples

Démo : Extraction de propositions

canada

1

Actions...

2

Points...

5

5

1995

2015

Point Only

energy

3

☒ support

☐ oppose

☐ report

☒

☐

3 messages [p 1 / 1]

ActorView

ActionView

AgreeDisagree

Actor

Action

Point

COP

Year

Conf

Canada	called	for recognition of the potential contribution of other measures, including the export of energy with low carbon content	3	1997	5
Canada	Recommending	a focus beyond trade issues	9	2003	5

3 sentences | 2

Sentences

Docs

KeyPhrase

DBpedia

ClimTag

Label

Count

carbon content	1
export of energy	1
low carbon content	1
potential contribution	1
recognition of the potential contribution	1

EXAMPLE USER QUERY (Main View)

Propositions with **Actor Canada** (1) voicing **support** (2), in sentences containing **energy** (3), with **confidence 5** (4), **time-range 1995-2015**, and **keyphrases** (7) extracted from the propositions' points

Keyphrase extraction and Entity Linking / Wikification (7, 8) on the messages allows:

- › Identifying **issues supported or opposed** by actors
- › Comparing actor profiles: *What issues do actors agree or disagree about?*

Access to Sentence and Document (6) helps verify annotations

Démo : Extraction de propositions

MainViews
AgreeDisagree

Actor 1

European U

Actor 2

Group of 77

RelationType

disagreeeme

✓

✗

9

Sentences

10

Sentence

The G-77/CHINA, opposed by the EU and NORWAY, stressed the need for predictable and adequate funding levels.

[enb12226e.txt-3]

KeyPhrase	Count	DBpedia	Count	ClimTag	Count
co-chairs	2	Climate change	1	implementation	3
adequate funding levels	1	Decision making	1	Global Environment Fund	1
adjustments of estimates	1	Developing country	1	Special Climate Change Fund	1
agenda item	1	Greenhouse gas	1	assumptions	1
alternative text	1	Methodology	1	baseline	1
annex i parties	1				

AgreeDisagree: Keyphrases and ClimTag concepts in proposition points where *the EU* and the *Group of 77 disagree (9)*, with access to the sentences (10)

- Rodrigo Agerri, Josu Bermudez and German Rigau (2014). IXA Pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of LREC 2014, the 9th Language Resources and Evaluation Conference*. Reykjavik, Iceland.
- Sophie Aubin and Thierry Hamon. Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing: 5th International Conference on NLP, FinTAL 2006*, pp. 380-387. LNAI 4139. Springer.
- Sören Auer et al. (2007). DBpedia: A nucleus for a web of open data. *The Semantic Web*. Springer.
- Trani, S., D. Ceccarelli, C. Lucchese, Orlando, R. Perego. (2014). Dexter 2.0 - an open source tool for semantically enriching data. In *Proceedings of the 13th International Semantic Web Conference*, Riva Del Garda, Italy
- Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. (2013). A framework for benchmarking entity-annotation systems. In *Proc. of WWW*.
- Paolo Ferragina and Ugo Scaiella. (2010). Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proc. of CIKM'10*, 1625–1628.
- Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. (2011). Robust disambiguation of named entities in text. In *Proc. of EMNLP*, 782–792.
- Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. (2011). DBpedia spotlight: shedding light on the web of documents. In *Proc. of the 7th Int. Conf. on Semantic Systems, I-SEMANTICS'11*, 1–8.
- David Milne and Ian H. Witten. (2008a). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proc. of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, 25–30.
- Andrea Moro, Alessandro Raganato, and Roberto Navigli. (2013). Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the ACL*, 2, 231–244.
- Thierry Poibeau. (2011). *Traitement automatique du contenu textuel*. Lavoisier.
- Thierry Poibeau, Horacio Saggon, Jakub Piskorski, and Roman Yangarber, eds. (2012). Multi-source, multilingual information extraction and summarization. Springer Science & Business Media.
- Pablo Ruiz, Thierry Poibeau, Frédérique Mélanie. (2015). Entity Linking with corpus coherence combining open source annotators. In *Proc. NAACL-HLT Demos*
- Pablo Ruiz, Clément Plancq, Thierry Poibeau (2016). More than Word Cooccurrence: Exploring Support and Opposition in International Climate Negotiations with Semantic Parsing. In *Proc. LREC*.
- Satoshi Sekine, Kiyoshi Sudo and Chikashi Nobata. (2002). Extended Named Entity Hierarchy. In *Proc. LREC*.
- Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proc. CoNLL*, pp. 159–177. Association for Computational Linguistics.
- Ricardo Usbeck et al. (2015). GERBIL – General Entity Annotator Benchmarking Framework. In *Proc. WWW*.
- Seth Van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle. (2013). Exploring entity recognition and disambiguation for cultural heritage collections. In *Digital Scholarship in the Humanities*, Oxford: Oxford University Press.
- Tommaso Venturini and Daniele Guido. (2012). Once upon a text: an ANT tale in Text Analytics. *Sociologica*, 3:1-17. Il Mulino, Bologna.



Avec le soutien d'une allocation doctorale de la Région Île-de-France

pablo.ruiz.fabo@ens.fr

<http://www.lattice.cnrs.fr/Pablo-Ruiz-Fabo,541>

<http://apps.lattice.cnrs.fr>

***Collecter et produire des données pour la recherche en SHS
Fréjus, 15-18 novembre 2016***

Slides supplémentaires

Nom	Liens
Tagme2	https://tagme.d4science.org/tagme/tagme_help.html
Spotlight	https://github.com/dbpedia-spotlight/dbpedia-spotlight
AIDA	https://github.com/yago-naga/aida
Dexter	https://github.com/dexter/dexter
AGDISTIS	https://github.com/AKSW/AGDISTIS
FOX	https://github.com/AKSW/FOX http://fox-demo.aksw.org/#!/home

- Toutes les librairies ont des services web (même hébergés publiquement).
- Déployer les outils en local facilite la reproductibilité des résultats

	Nom	Liens
Java	DKPro	https://dkpro.github.io/
	IXA	http://ixa2.si.ehu.es/ixa-pipes/
	Pipeline	http://ixa2.si.ehu.es/ixa-pipes/third-party-tools.html
	CoreNLP	https://github.com/stanfordnlp/CoreNLP
	Mate	https://code.google.com/archive/p/mate-tools/wikis/ParserAndModels.wiki
Scala	UW	https://github.com/knowitall/nlptools https://github.com/allenai/openie-standalone
Python	NLTK	http://www.nltk.org/ http://www.nltk.org/nltk_data/
C++	FreeLing	http://nlp.lsi.upc.edu/freeling/node/1 http://nlp.lsi.upc.edu/freeling/demo/demo.php
	LIMA	https://github.com/aymara/lima/wiki

- La plupart des librairies ont des services web


```
<nafHeader>  
  <linguisticProcessors layer="text">  
  <linguisticProcessors layer="terms">  
  <linguisticProcessors layer="constituency">  
  <linguisticProcessors layer="coreferences">  
  <linguisticProcessors layer="deps">  
  <linguisticProcessors layer="srl">  
</nafHeader>
```

- Exemple avec la librairie IXA Pipeline (mais étapes communes à d'autres librairies de TAL)
- Format de sortie NAF (XML) : NLP Annotation Format (NLP = Natural Language Processing $\approx$ TAL)

Composantes typiques du traitement TAL

- **Tokénisation** : délimitation en mots
- **Lemmatisation** : ramener aux formes non-fléchies (ex. infinitif pour les verbes)
- Étiquetage de **catégories grammaticales**
- Analyse en **constituants syntaxiques**
- **Coréférence**
- Analyse en **dépendances syntaxiques**
- Analyse en **rôles sémantiques**

```
<wf id="w1758" sent="48" para="1" offset="10213" length="1">.</wf>
<wf id="w1759" sent="49" para="1" offset="10215" length="2">He</wf>
<wf id="w1760" sent="49" para="1" offset="10218" length="6">stated</wf>
<wf id="w1761" sent="49" para="1" offset="10225" length="4">that</wf>
<wf id="w1762" sent="49" para="1" offset="10230" length="8">advanced</wf>
<wf id="w1763" sent="49" para="1" offset="10239" length="8">research</wf>
<wf id="w1764" sent="49" para="1" offset="10248" length="3">and</wf>
<wf id="w1765" sent="49" para="1" offset="10252" length="9">knowledge</w
<wf id="w1766" sent="49" para="1" offset="10262" length="4">must</wf>
<wf id="w1767" sent="49" para="1" offset="10267" length="2">be</wf>
<wf id="w1768" sent="49" para="1" offset="10270" length="11">accompanied
<wf id="w1769" sent="49" para="1" offset="10282" length="2">by</wf>
<wf id="w1770" sent="49" para="1" offset="10285" length="9">increased</w
<wf id="w1771" sent="49" para="1" offset="10295" length="9">resources</w
```

MATE-SHS Lemmes et catégories grammaticales (pos)

```
<!--stated-->
<term id="t1760" type="open" lemma="state" pos="V" morphofeat="VBD">
  <span>
    <target id="w1760" />
  </span>
</term>
<!--that-->
<term id="t1761" type="close" lemma="that" pos="P" morphofeat="IN">
  <span>
    <target id="w1761" />
  </span>
</term>
<!--advanced-->
<term id="t1762" type="open" lemma="advanced" pos="G" morphofeat="JJ">
  <span>
    <target id="w1762" />
  </span>
</term>
<!--research-->
<term id="t1763" type="open" lemma="research" pos="N" morphofeat="NN">
  <span>
    <target id="w1763" />
  </span>
</term>
```

Analyse en dépendances

```

... <!-- SBJ(stated, He) -->
... <dep from="t1760" to="t1759" rfunc="SBJ" />
... <!-- OBJ(stated, that) -->
... <dep from="t1760" to="t1761" rfunc="OBJ" />
... <!-- NMOD(research, advanced) -->
... <dep from="t1763" to="t1762" rfunc="NMOD" />
... <!-- SBJ(must, research) -->
... <dep from="t1766" to="t1763" rfunc="SBJ" />
... <!-- COORD(research, and) -->
... <dep from="t1763" to="t1764" rfunc="COORD" />
... <!-- CONJ(and, knowledge) -->
... <dep from="t1764" to="t1765" rfunc="CONJ" />
... <!-- SUB(that, must) -->
... <dep from="t1761" to="t1766" rfunc="SUB" />
... <!-- VC(must, be) -->
... <dep from="t1766" to="t1767" rfunc="VC" />
... <!-- VC(be, accompanied) -->

```

Chaînes de coréférence

```

... <coref id="co467">
... <!--President González García-->
... <span>
... <target id="t2186" />
... <target id="t2187" />
... <target id="t2188" />
... </span>
... <!--He-->
... <span>
... <target id="t1759" />
... </span>
... <!--he-->
... <span>
... <target id="t12170" />
... </span>
... <!--President González García-->
... <span>

```


Chaînes de corréférence

```

... <coref id="co467">
...   <!--President González García-->
...   <span>
...     <target id="t2186" />
...     <target id="t2187" />
...     <target id="t2188" />
...   </span>
...   <!--He-->
...   <span>
...     <target id="t1759" />
...   </span>
...   <!--he-->
...   <span>
...     <target id="t12170" />
...   </span>
...   <!--President González García-->
...   <span>

```

Rôles sémantiques : Prédicat

```
<!--t1760 stated : A0[t1759 He] A1[t1761 that]-->
<predicate id="pr324">
  <!--stated-->
  <span>
    <target id="t1760" />
  </span>
  <externalReferences>
    <externalRef resource="PropBank" reference="state.01" />
    <externalRef resource="VerbNet" reference="say-37.7" />
    <externalRef resource="VerbNet" reference="say-37.7-1" />
    <externalRef resource="FrameNet" reference="Statement" />
    <externalRef resource="PropBank" reference="state.01" />
    <externalRef resource="EventType" reference="communication" />
    <externalRef resource="WordNet" reference="ili-30-01009240-v" />
  </externalReferences>
```

- Texte enrichi automatiquement avec des références à des bases lexicales sur le web

```

</externalReferences>
<role id="rl611" semRole="A0">
  <!--He-->
  <span>
    <target id="t1759" head="yes" />
  </span>
  <externalReferences>
    <externalRef resource="VerbNet" reference="say-37.7@Agent" />
    <externalRef resource="FrameNet" reference="Statement@Speaker" />
    <externalRef resource="PropBank" reference="state.01@0" />
  </externalReferences>
</role>
<role id="rl612" semRole="A1">
  <!--that advanced research and knowledge must be accompanied by increased res
that commitments from all members of society are needed-->
  <span>
    <target id="t1761" head="yes" />
    <target id="t1762" />
  </span>
  <span>
    <target id="t1763" />
  </span>
  <externalReferences>
    <externalRef resource="VerbNet" reference="say-37.7@Topic" />
    <externalRef resource="FrameNet" reference="Statement@Message" />
    <externalRef resource="FrameNet" reference="Statement@Topic" />
    <externalRef resource="PropBank" reference="state.01@1" />
  </externalReferences>
</role>

```

```
</externalReferences>
<role id="rl611" semRole="A0">
  <!--He-->
  <span>
    <target id="t1759" head="yes" />
  </span>
  <externalReferences>
    <externalRef resource="VerbNet" reference="say-37.7@Agent" />
    <externalRef resource="FrameNet" reference="Statement@Speaker" />
    <externalRef resource="PropBank" reference="state.01@0" />
  </externalReferences>
</role>
<role id="rl612" semRole="A1">
  <!--that advanced research and knowledge must be accompanied by increased res
that commitments from all members of society are needed-->
  <span>
    <target id="t1761" head="yes" />
    <target id="t1762" />
  </span>
  <span>
    <target id="t1763" />
  </span>
  <externalReferences>
    <externalRef resource="VerbNet" reference="say-37.7@Topic" />
    <externalRef resource="FrameNet" reference="Statement@Message" />
    <externalRef resource="FrameNet" reference="Statement@Topic" />
    <externalRef resource="PropBank" reference="state.01@1" />
  </externalReferences>
</role>
```