

Application de la résolution référentielle d'entités (entity linking) au domaine des Humanités numériques

Journée Sciences des données et
Humanités numériques

Thierry Poibeau, Pablo Ruiz — LATTICE



Résumé

- **Technologie** de résolution référentielle des entités (entity linking)
 - **Description** de la technologie
 - **Outils** publiques
 - **Implications** des choix d'outil pour la modélisation des données textuelles
- **Exemple** d'application sur corpus

Entity Linking

- **Tâche:** Repérer des **concepts** d'une ontologie dans un corpus à travers de **mentions** qui font référence à ces concepts
 - Ex. acteurs du corpus, sujets abordés ...
- **Utilité:** Établir des **liens** entre les éléments du corpus à travers leurs concepts en commun *malgré la variabilité dans les mentions*
 - Navigation par concept
 - Visualisation sous format réseau

Entity Linking : Étapes

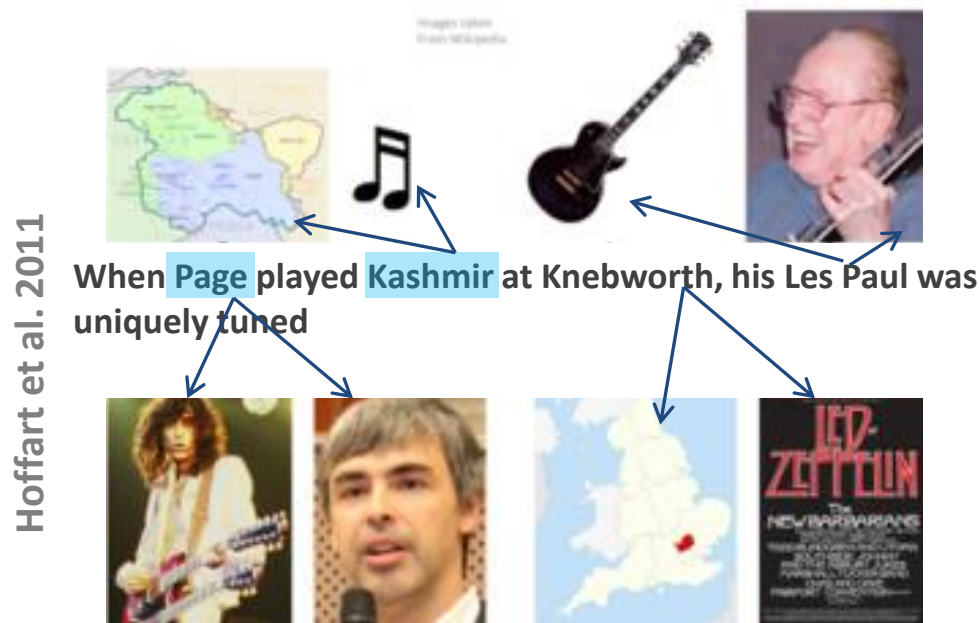
- **Reconnaissance** des mentions
 - Reconnaissance des entités nommées
- **Désambiguïsation** des mentions vers leur concept (entité) dans l'ontologie
 - Entity Linking

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)

Entity Linking : Défis

1. Mentions ambiguës entre plusieurs entités possibles



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

Page: Larry vs. Jimmy

Kashmir: Région vs. chanson

2. Plusieurs mentions possibles pour chaque entité

Barack Hussein Obama, Barack Obama, Obama, Mr. Obama, President Obama, POTUS, #POTUS

REN

- Techniques: dictionnaires et/ou étiquetage de séquences (Tjong Kim Sang et al.(2003); Huang et al., (2015))

Le Monde, 20 nov 2015



Et pour cause : son prénom, homonyme du nom anglais de l'organisation Etat Islamique (ISIS, acronyme d'*Islamic State in Irak and Syria*), est motif de suspension pour la firme de Mark Zuckerberg.

REN

- Techniques: dictionnaires et/ou étiquetage de séquences (Tjong Kim Sang et al.(2003); Huang et al. (2015))

Le Monde, 20 nov 2015



Et pour cause : son prénom, homonyme du nom anglais de l'organisation Etat Islamique (ISIS, acronyme d'*Islamic State in Irak and Syria*), est motif de suspension pour la firme de Mark Zuckerberg.

Montréal la « verte » s'apprête à déverser ses égouts dans le Saint-Laurent

LE MONDE | 09.10.2015 à 11h13 • Mis à jour le 11.10.2015 à 13h33 |

Par Anne Pélouas (Montréal, correspondance)

Abonnez vous à partir de 1 €



Réagir



Classer



f Partager (1 119)



Tweeter



Le projet risque fort de ruiner sa réputation de « ville verte » : la ville de Montréal s'apprête à déverser volontairement huit milliards de litres d'eaux usées (rejets d'égouts et de toilettes) dans le fleuve Saint-Laurent. Or, la majeure partie de la population du Québec vit, comme 64 espèces d'animaux, au bord de cette artère maritime longue de plus de 1 000 kilomètres. Elle abrite de nombreuses espèces marines, dont des baleines et des bélugas dans son estuaire, 80 sortes de poissons et près de 400 d'oiseaux. Le Saint-Laurent fournit 45 % de l'eau potable consommée par les 8 millions de Québécois.



Le maire de Montréal, Denis Coderre, a annoncé que l'opération se déroulerait du 18 au 25 octobre pour permettre des travaux sur un échangeur autoroutier sous lequel se trouve une conduite majeure acheminant les eaux usées vers une station d'épuration.

Le Monde, oct 2015

Conjonction **or** mal reconnue comme candidat d'entité

...

Montréal la « verte » s'apprête à déverser ses égouts dans le Saint-Laurent

LE MONDE | 09.10.2015 à 11h13 • Mis à jour le 11.10.2015 à 13h33 |

Par Anne Pélouas (Montréal, correspondance)

Abonnez vous à partir de 1 €



Réagir



Classer



Partager (1 119)



Tweeter



Le projet risque fort de ruiner sa réputation de « ville verte » : la ville de Montréal s'apprête à déverser volontairement huit milliards de litres d'eaux usées (rejets d'égouts et de toilettes) dans le fleuve Saint-Laurent. Or, la majeure partie de la population du Québec vit, comme 64 espèces d'animaux, au bord de cette artère maritime longue de plus de 1 000 kilomètres. Elle abrite de nombreuses espèces marines, dont des baleines et des bélugas dans son estuaire, 80 sortes de poissons et près de 400 d'oiseaux. Le Saint-Laurent fournit 45 % de l'eau potable consommée par les 8 millions de Québécois.



Le maire de Montréal, Denis Coderre, a annoncé que l'opération se déroulerait du 18 au 25 octobre pour permettre des travaux sur un échangeur autoroutier sous lequel se trouve une conduite majeure acheminant les eaux usées vers une station d'épuration.

Le Monde, oct 2015

Conjonction **or** mal reconnue comme candidat d'entité

...

Et ensuite reliée (erreur) avec le produit **Or** du marché boursier

Outils Entity Linking (EL)



2010



2011

wikipedia**miner**

2008

aida

2011

Babelfy

2014

- Outils publics qui désambiguïsent vers des **ontologies générales** (DBpedia, YAGO, Babelnet)
- Des **évaluations** publiques ont montré que les forces de chaque outils varient selon certains traits des textes:
 - Cornolti et al. (2013) BAT Framework
 - Usbeck et al. (2014) GERBIL

Outils EL : Différences selon textes

- **Corrélations** entre nombre d'occurrences d'un **trait** textuel et performance de l'outil
- Évaluation (plateforme GERBIL) sur 10 corpus divers (web, nouvelles, tweets ...)

CORRELATIONS	Nbr. Personnes	Nbr. Organisations	Nbr. Lieux	Nbr. AUTRES
Babelfy	0.769	-0.376	0.254	-0.431
Spotlight	0.217	-0.480	-0.461	0.26
TagMe	0.257	-0.272	-0.194	0.036
WikipediaMiner	0.082	-0.679	-0.632	0.497

Données du 20 nov, plateforme GERBIL (gerbil.aksw.org/gerbil/overview), tâche A2KB/Ma

Outils EL : Différences selon textes

- Évaluation : comment définir résultat correct?
 - Découpage mention + lien avec BC, lien seul ...
 - Ji et al., 2014 (Campagne *TAC-KBP*): outil *neleval*

Données de Cornolti et al. (2013)

OUTIL	CORPUS					
	AIDA/CoNLL (nouvelles, sports)			IITB (web, divers)		
	P	R	F1	P	R	F1
Spotlight	31.2	40.4	35.2	46.2	50.0	48.0
TagMe	61.4	55.5	58.3	45.2	42.0	43.6
WikipediaMiner	46.9	52.8	49.7	56.8	48.2	43.6
AIDA	63.3	29.1	39.8	65.7	4.1	7.6

EL : Infos de sortie

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
},
```

EL : Infos de sortie (étape *REN*)

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  
},
```

EL : Infos de sortie (REN + *EL*)

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  
},
```


EL : Désambiguïsation

- Comparaison des **contextes** d'une mention dans le corpus et du texte relié aux entités dans la base de connaissance (tokens, liens)
- Méthodes graphiques avec une mesure de **cohérence** (ex. *liens entrants communs*)
 - Milne & Witten (2008)
 - Ferragina et al. (2010)
 - Hoffart et al. (2011)
 - Moro et al. (2014)
- Spotlight (Mendes et al. 2011) exploite la comparaison de contextes (sans mesure de cohérence).

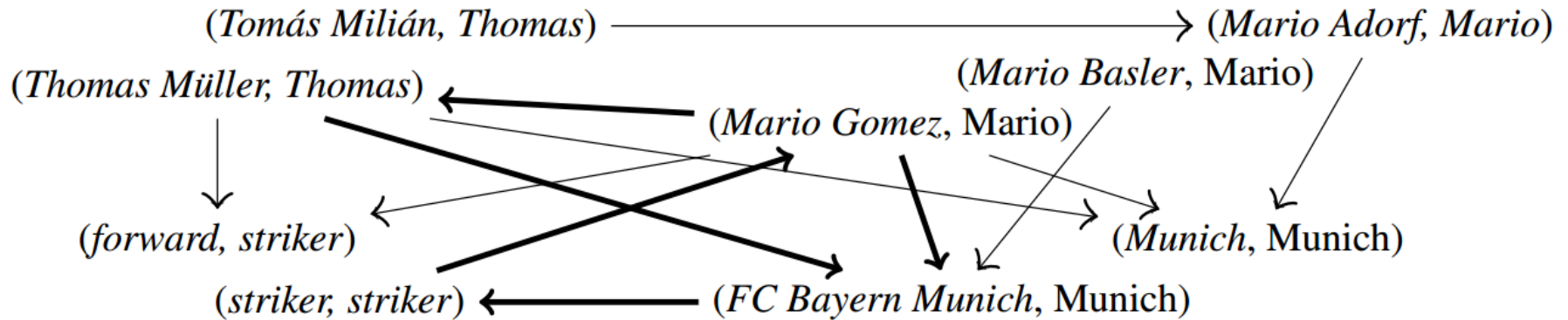
wikipedia**miner**



aida

Babelfy

EL : Cohérence



Thomas and Mario are strikers playing in Munich

(Moro and Navigli, 2014)

Application : Corpus Bentham

Jeremy Bentham: Philosophe, réformateur
(1748-1832, Londres)



Transcribe
Bentham

Transcribe Bentham

(Causer & Terras, 2014)

- UCL (Londres)
- Manuscrits inédits, transcrits par des volontaires (crowdsourcing)
- 30,000 pages

Image: blogs.ucl.ac.uk/transcribe-bentham/

Application : Corpus Bentham

Défi

- Accès relevant et structuration du contenu

Méthode

- Entity linking vers DBpedia avec Spotlight, choix des concepts les plus robustes
- Basé sur ces concepts, création de réseaux avec Cortext Manager (gexf)
- Visualisations avec Gephi



Application : Corpus Bentham

Résultats

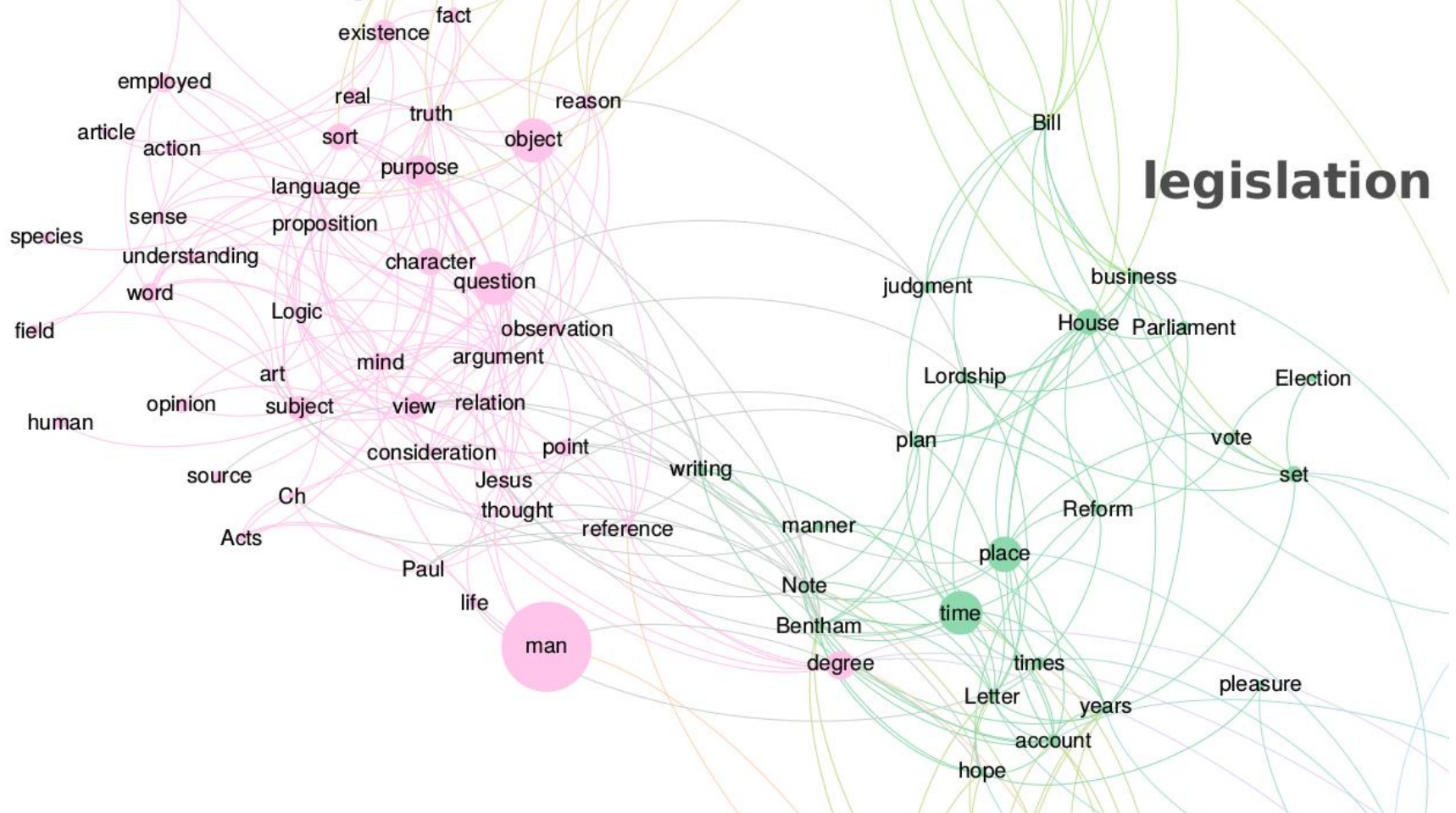
- Grands groupes thématiques de l'œuvre de Bentham visibles sur réseau, comme confirmé par des experts du domaine

Réflexions

- Intérêt / validité d'appliquer une ontologie publiée en 2007 à des textes de 1800 ?
- Ontologies générales et spécifiques à un domaine ; enjeux « politiques » (van Hooland et al., 2013)

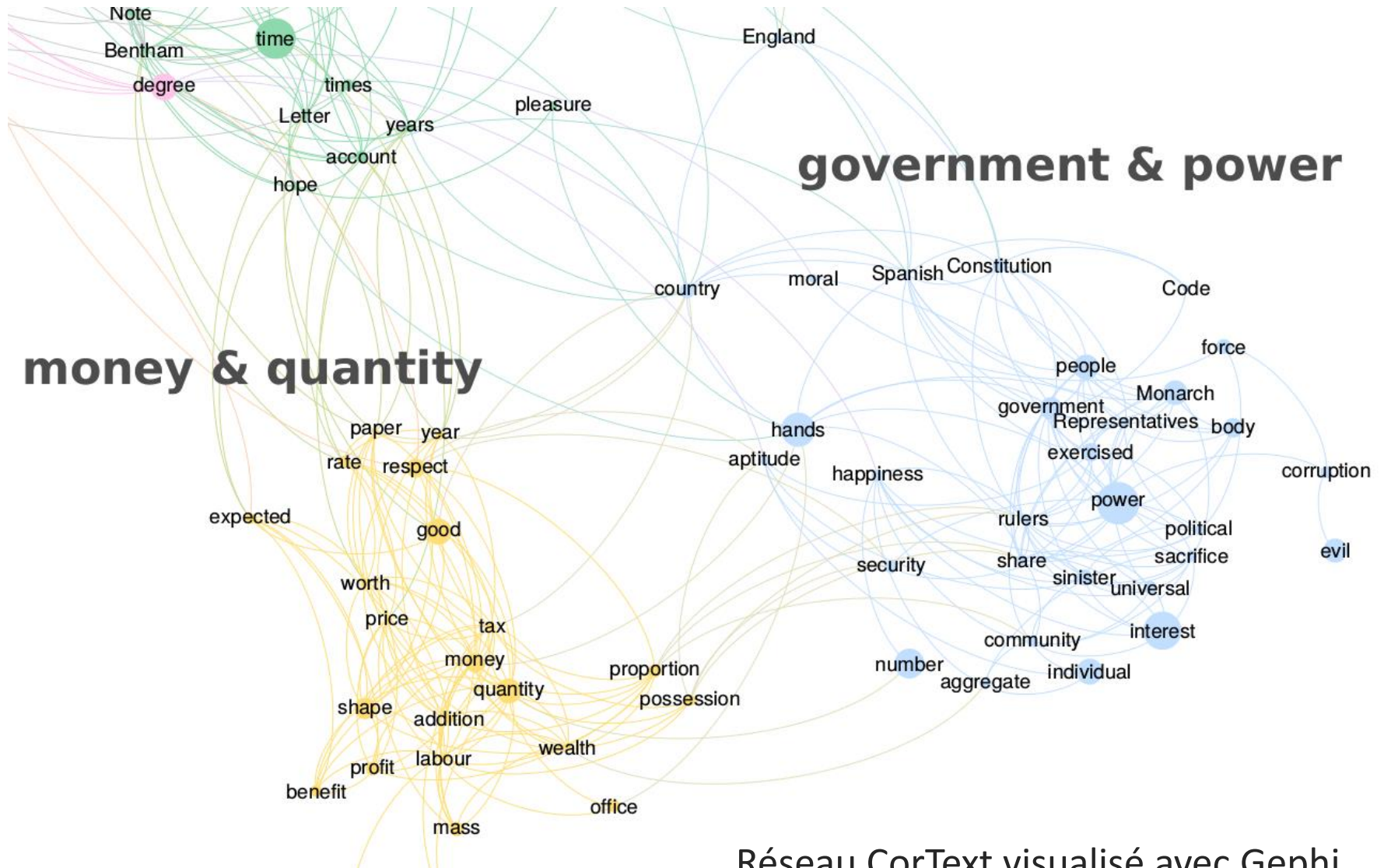
Clusters réseau Bentham

discourse & argument



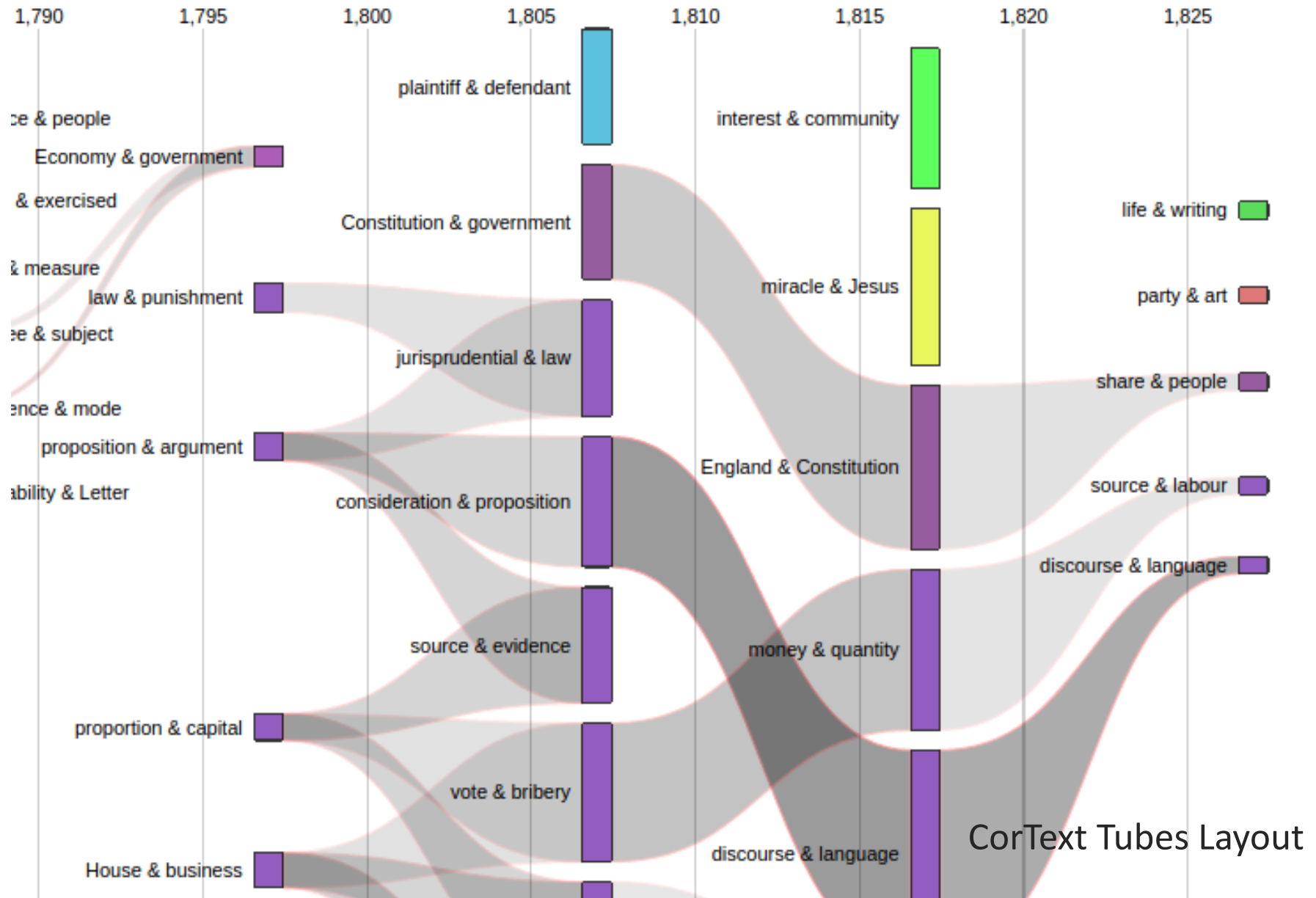
Réseau CorText visualisé avec Gephi

Clusters réseau Bentham



Réseau CorText visualisé avec Gephi

Bentham : vue dynamique



Références

- Tim Causer and Melissa Terras (2014). Crowdsourcing Bentham: Beyond the traditional boundaries of academic history. *International Journal of Humanities and Arts Computing*, vol. 8(1), pp. 46-64.
- Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. (2013). A framework for benchmarking entity-annotation systems. In *Proc. of WWW*, 249–260.
- 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.
- Heng Ji, Joel Nothman and Ben Hachey. (2014). Overview of TAC-KBP2014 Entity Discovery and Linking Tasks. In *Proc. Text Analysis Conference*.
- Huang et al. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv preprint arXiv:1508.01991
- 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.
- Pablo Ruiz, Thierry Poibeau, Frédérique Mélanie. (2015). Entity Linking with corpus coherence combining open source annotators. In *Proc. NAACL-HLT: Demonstrations*.
- Satoshi Sekine, Kiyoshi Sudo and Chikashi Nobata. (2002). Extended Named Entity Hierarchy. In *Proc. LREC*.
- Eric F. Tjong Kim Sang and Fien De Meulder. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proc. CoNLL*. (ACL)
- Ricardo Usbeck et al. (2015). GERBIL – General Entity Annotator Benchmarking Framework. In *Proc. of 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.

Merci !



fr.dbpedia.org/page/Merci_(mot)

About: Merci (mot)

An Entity of Type : [Thing](#), from Named Graph : <http://fr.dbpedia.org>, within Data Space : [fr.](#)

Property	Value
dbpedia-owl:wikiPageExternalLink	http://atilf.atilf.fr/tlf.htm
dcterms:subject	dbpedia-fr:Catégorie:Expression_française_populaire
rdfs:label	Merci (mot)
http://www.w3.org/ns/prov#wasDerivedFrom	http://fr.wikipedia.org/wiki/Merci_(mot)?oldid=93658311
foaf:isPrimaryTopicOf	http://fr.wikipedia.org/wiki/Merci_(mot)
is dbpedia-owl:wikiPageDisambiguates of	dbpedia-fr:Merci
is dbpedia-owl:wikiPageWikiLink of	dbpedia-fr:Merci
is foaf:primaryTopic of	http://fr.wikipedia.org/wiki/Merci_(mot)