

# Le Traitement des Entités Nommées

DEFINITION, RESOURCES, MÉTHODES, APPLICATIONS

---

Maud Ehrmann<sup>1</sup> Sophie Rosset<sup>2</sup>

11 Juillet 2018

**Session 1:** Cours (1h30)

**Session 2:** TP, Elaborer un système de reconnaissance d'EN

**Session 3:** TP, Evaluer un système de reconnaissance d'EN

<sup>1</sup>EPFL-DHLAB, Lausanne, Switzerland

<sup>2</sup>LIMSI-ILES, Orsay, France

# Plan du cours

1. Contexte et Applications
2. Définition
3. Resources
4. Reconnaissance et classification
5. Liaison
6. Evaluation

# 1. Contexte et Applications

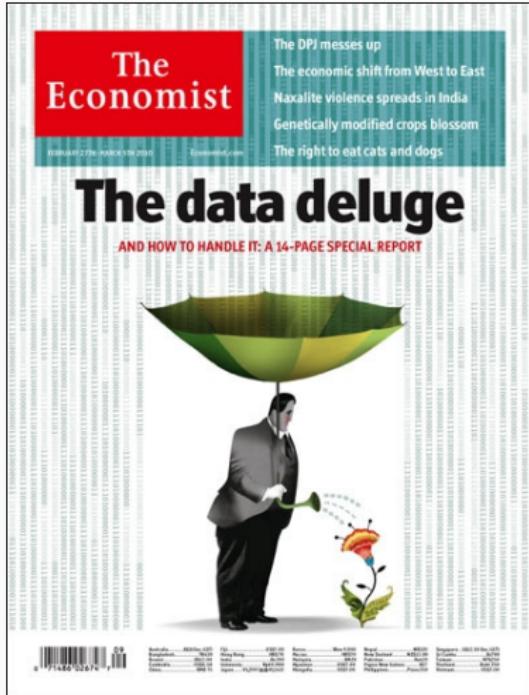
---

# **1. Contexte et Applications**

---

## **1.1 Introduction**

# Contexte



# Données

**Quoi:** TOUT

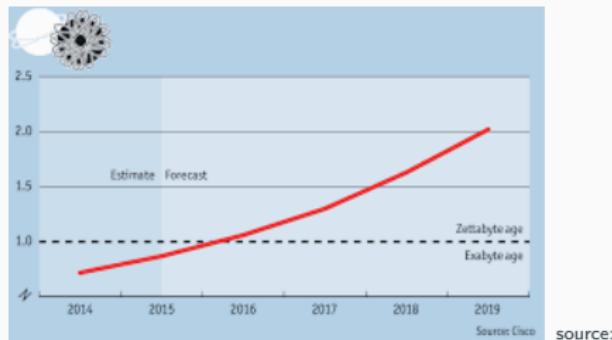
i.e. texte, images, audio publiés via sites de news, médias sociaux, plateformes collaboratives, smartphones, capteurs, etc.



# Profil (1/2)

**Combien:** croissance astronomique

- quantité: double tous les 2 ans
- trafic: entré dans l'ère des zettabyte en 2016 (1 trillion gigabytes)
- stockage: prévision de 44 zettabytes en 2020



## Profil (2/2)

**Nature:** 80 à 90% des données sont **non structurées**, i.e. sans modèle ni format pré-définis. **Défis**

- stockage de plus en plus coûteux
- mais surtout: exploiter les données, **extraire l'information utile**

# Mise au point

Donnée	Information	Connaissance
description élémentaire d'une réalité	données avec un sens construisant une représentation de la réalité	informations avec une vérité
<i>mesure des températures</i>	<i>courbe sur l'évolution des minima et maxima moyens en un lieu donné, par mois</i>	<i>fait que la température sur terre augmente du fait de l'activité humaine</i>
<i>série d'articles journalistiques</i>	<i>nom de personnes et leurs polarités</i>	<i>opinion des médias vis-à-vis de personnalités</i>

inspiré de: [http://www.college-de-france.fr/site/serge-abiteboul/\\_inaugural-lecture.htm](http://www.college-de-france.fr/site/serge-abiteboul/_inaugural-lecture.htm)

# Données semi-structurée

## Cannes Film Festival

From Wikipedia, the free encyclopedia

Coordinates: 43°33'03.10"N 7°01'02.10"E

The Cannes Festival ([/kænʃ/](#)) (French: *Festival de Cannes*), named until 2002 as the International Film Festival (*Festival international du film*) and known in English as the Cannes Film Festival, is an annual film festival held in Cannes, France, which previews new films of all genres, including documentaries, from all around the world. Founded in 1946, the invitation-only festival is held annually (usually in May) at the Palais des Festivals et des Congrès.<sup>[1][2][3]</sup>

On 1 July 2014, co-founder and former head of French pay-TV operator Canal+ Pierre Lescure took over as President of the festival. The Board of Directors also appointed Gilles Jacob as Honorary President of the festival.<sup>[4][5][6]</sup>

The 2016 Cannes Film Festival took place between 11 and 22 May 2016. Australian film director George Miller was the President of the Jury. *I, Daniel Blake*, directed by British director Ken Loach, won the Palme d'Or.

In 2017, The Festival de Cannes will celebrate its 70th anniversary edition from May 17 to 28.

### Contents [hide]

- 1 History
- 2 Impact
- 3 Programmes
- 4 Juries
- 5 Awards
- 6 See also
- 7 References
- 8 Further reading
- 9 External links



Festival de Cannes

@Festival\_Cannes



Follow

In French theaters today, testimonies from Ugandan ex-child soldiers : Wrong Elements by Jonathan Littell #SpecialScreening in #Cannes2016

### Cannes Film Festival



Location	Cannes, France
Founded	September 20, 1946
Awards	Palme d'Or, Grand Prix
Website	<a href="http://festival-cannes.com">festival-cannes.com</a>

*mais la plupart du temps,  
l'information est ‘cachée’ dans les textes*

# Données non-structurées

“On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar.”

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th Festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar.

Monica Belucci's friendship with the Festival de Cannes goes back a long way: in 2000, she walked up the steps for the first time to present *Under Suspicion* by Stephen Hopkins. She returned two years later with Gaspar Noé's steamy *Irreversible* which entranced the Croisette with its unforgettable violence.

Monica Belucci was a member of the Jury in 2006 under the presidency of Wong Kar-wai. In the following years, Belucci returned to Cannes for the Official Selection with Marco Tullio Giordana's *What About Us?* and *Don't Look Back* by María de Varn. In 2014, she was back on the Croisette to present *The Wonders* by Italian director Alice Rohrwacher, which picked up the Jury Grand Prix.

Belucci's film career demonstrates her ease across a range of genres with outstanding performances in both comedy and drama, based on eclectic and daring artistic choices. She has films for a number of prestigious directors including Bertrand Blier, Danièle

source: [www.festival-cannes.com](http://www.festival-cannes.com)

## Information ‘cachée’ dans les textes

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar.

PERSON, ORGANIZATION, TIME-EXPR, EVENT

# Extraction d'Information (EI)

L'objectif est d'extraire des informations structurées à partir de textes non structurés, c-a-d:

- identifier et catégoriser des fragments d'information
- les relier avec des bases de connaissances
- les aggréger pour extraire d'autres informations

# Principales tâches en EI

- **traitement des entités nommées:**

reconnaissance, catégorisation et désambiguisation

- *Monica Belluci* et *Pedro Almodovar* sont des PERSON.
  - *Monica Belluci*  $\xrightarrow{\text{reference}}$  [http://dbpedia.org/page/Monica\\_Bellucci](http://dbpedia.org/page/Monica_Bellucci)

- **traitement des expressions temporelle:**

extraction et normalisation

- *from 17 to 28 May 2017* est une DURATION
  - *from 17 to 28 May 2017* → [17-05-2017, 28-05-2017]

- **extraction d'événements**

- *70th Festival de Cannes* est un FACTUAL, RECURRING EVENT
  - *70th Festival de Cannes*  $\xrightarrow{\text{instance\_of}}$   
[https://en.wikipedia.org/wiki/Cannes\\_Film\\_Festival](https://en.wikipedia.org/wiki/Cannes_Film_Festival)

- **extraction de relations:**

- *70th Festival de Cannes, tookPlace, [17-05-2017, 28-05-2017]*

# 1. Contexte et Applications

---

## 1.2 Un peu d'histoire

# De la compréhension à l'extraction

- **1980s:** objectif **compréhension automatique** de textes
- Un projet **trop ambitieux** face à des difficultés techniques et théoriques:
  - faible couverture des grammaires
  - trop d'ambiguïtés non résolues
  - difficultés à collecter, représenter et manipuler les connaissances

→ approche générique de la compréhension de textes est encore une **utopie**
- **1990s:** **décomposition de la tâche** de compréhension
  - se focalise sur des éléments précis d'intérêt
  - un modèle est défini à l'avance en fonction de l'application
  - analyse locale (10-20% du texte nécessaire).

# La série des conférences MUC

- *Message Understanding Conference*
- Cycle de 7 campagnes d'évaluation entre 1987 et 1998
- Initié par la Division pour la Recherche et le Développement de la Marine américaine
- Financé par le DARPA (Defense Advanced Research Project Agency)

# Evolution des conférences MUC

## Phase 1: cycle exploratoire

- **1987 (MUC-1)** pas de tâche précise, rapports militaires sur des opérations navales en style télégraphique;
- **1989 (MUC-2)** définition de formulaires prédéfinis (*templates*) devant être complétés (**10** champs); définition de mesures d'évaluation (precision et rappel).

## Phase 2: remplissage de formulaires de plus en plus complexes

- **1991 (MUC-3)** dépêches de presse sur événements terroristes en Amérique centrale et du sud; formulaire avec **18** champs.
- **1992 (MUC-4)** idem, **24** champs
- **1993 (MUC-5)** tâches plus complexes, test sur domaines nouveaux, 2 langues, 11 formulaires **48** champs hiérarchisés

## MUC 3: Définition de la tâche de compréhension

Etant donné un document, il était demandé de:

- repérer des événements,
- repérer les éléments s'y rattachant,
- "normaliser" ces éléments,
- remplir un formulaire descriptif.

e.g. pour chaque événement, il fallait trouver son type, sa date, les agents impliqués, le lieu etc.

# Formulaire MUC-3

*19 March - A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb - allegedly detonated by urban guerrilla commandos - blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).*

INCIDENT TYPE	bombing
DATE	March 19
LOCATION	El Salvador : San Salvador (city)
PERPETRATOR	urban guerrilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	-
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

todo: ajouter ref

# Evolution des conférences MUC

## Phase 3: reformulation des objectifs et définition de sous-tâches

- **1995 (MUC-6)**

- technologies et composants indépendants
    - définition de la sous-tâche dédiée aux "entités nommées"
  - systèmes portables
    - définition de template génériques
  - considération des "briques de base de la compréhension"
    - co-reference, désambiguisation lexicale, structure argument-prédicat

- **1997 (MUC-7)** continuation

## MUC 6: point de départ des travaux sur les EN

- Définition de la notion d'entité nommée
- Définition de la tâche de reconnaissance des EN

**Ensuite:** MET, IREX, CONLL, ACE, ESTER, ETAPE, HAREM, EVALITA,  
GERMEVAL, TREC, TAC, etc.

## **1. Contexte et Applications**

---

### **1.3 Définition courante**

## Entités nommées: première définition (TAL)

- des éléments "d'intérêt", généralement de type *Personne*, *Organisation*, *Lieu*
- des unités référentielles qui sous-tendent la sémantique des textes.

## Entités nommées: différentes tâches

1. **reconnaissance**: détecter, repérer des entités nommées dans les flux textuels (on pose les frontières dans le texte)
2. **classification**: catégoriser les éléments reconnus selon des catégories sémantiques pré-définies (on affecte un type)
3. **désambiguïsation/liaison**: lier les mentions d'entités à une référence unique (on lie à une référence)
4. **extraction de relation**: découvrir des relations entre entités (*father-of, born-in, alma mater*)

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar. [...] Monica Bellucci's friendship with the Festival de Cannes goes back a long way: in 2000, she walked up the steps for the first time to present *Under Suspicion* by Stephen Hopkins.

# Application de Stanford NER

On the invitation of the **Festival de Cannes**, the Italian actress **Monica Belucci** has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de **Cannes** to be held from 17 to 28 **May 2017**, under the presidency of Spanish filmmaker **Pedro Almodovar**. [...] **Monica Bellucci**'s friendship with the **Festival de Cannes** goes back a long way: in **2000**, she walked up the steps for the first time to present *Under Suspicion* by **Stephen Hopkins**.

PERSON, ORGANIZATION, LOCATION, DATE

## Plus d'information?

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar. [...] Monica Bellucci's friendship with the Festival de Cannes goes back a long way: in 2000, she walked up the steps for the first time to present Under Suspicion by Stephen Hopkins.

# Désambiguisation et relations

On the invitation of the **Festival de Cannes**, the Italian actress **Monica Bellucci** has agreed to play the role of Mistress of the Ceremonies of the 70th festival de **Cannes** to **May 2017**, under the presidency of Spanish **Carles Puigdemont**. [...] **Monica Bellucci**'s friendship with **Stephen Hopkins** goes back a long way: in **2000**, she walked up the red carpet to present *Under Suspicion* by **Stephen Hopkins**.



DBpedia

About: [Monica Bellucci](#)

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

WIKIDATA

Item Discussion

Monica Bellucci (Q81819)

Italian actress

# 1. Contexte et Applications

---

## 1.4 Applications

# Applications ‘internes’ au TAL (1/4)

- Étiquetage morpho-syntaxique et analyse syntaxique de surface
  - HyOx, Inc.
  - Seat and Porsche has fewer registrations in July 1996.
- Analyse syntaxique
  - *He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to Egypt and Jordan.*
  - *He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to Egypt and Likud party.*

# Applications ‘internes’ au TAL (2/4)

- **Analyse en dépendances**

*They met in **Bagdad**.* → LOCATION(*met*, *Bagdad*)

- **Coreference**

*John bought a new computer. **It** was able to train the model.*

- **Traduction**

*Jack London was an american writer. London is a busy city.*

- **Désambiguisation lexicale**

- *It is difficult to leave Paris on Friday evenings.*
  - *Some wonder if they will leave the Socialist party.*

Quelle est la signification de 'leave' ?

## Applications 'internes' au TAL (4/4)

- Désambiguïsation lexicale

*It is difficult to leave Paris on Friday evenings.*

→ leave = "go away from a place" (#1 WordNet)

*Some wonder if they will leave the Socialist party.*

→ leave = "remove oneself from an association with or participation in" (#8 WordNet)

- **Extraction d'information et 'media monitoring'**
  - population de bases de connaissances avec des informations relatives à des entités
  - alertes sur certains sujets ou entités
- **Clustering de documents cross-lingue**

Les documents mentionnant les mêmes entités ont de fortes chances d'être reliés.
- **Résumé automatique**

Les EN sont des 'ancres' informationnelles aidant à identifier les éléments clés d'un texte

## **contexte: à retenir**

- La notion d'EN est apparue dans les **années 90** lors de campagnes d'évaluations sur la **compréhension de documents**
- Les EN ont rapidement pris une place importante et sont devenues un **pivot central** pour les systèmes d'analyse automatique des textes.

## 2. Définition

---

*A quoi correspondent vraiment les entités nommées? Comment les définir?*

## **2. Définition**

---

### **2.1 Difficultés d'appréhension**

# Les EN dans le monde : le problème de la catégorisation

- Le choix des catégories

TRIADE UNIVERSELLE :  
Personne,  
Lieu,  
Organisation



DIVERSIFICATION :  
Bâtiment, Arme,  
Produit, Divers  
Véhicule, ...

- La détermination de ce qu'elles recouvrent

Catégorie PERSONNE :

Lionel Jospin	les Démocrates	Bison Futé
les Windsors	les Talibans	le Prince Charmant
la famille Kennedy	Zorro	l'épouse Chirac
les frères Cohen	St Nicolas	...

→ catégorisation instable

# Les EN dans le texte : le problème de l'annotation

- **Combinaisons de syntagmes : une ou plusieurs entités ?**

*Les Banques centrales américaine et européenne ont décidé...*

*Bill et Hillary Clinton*

*l'Université de Corte*

- **Un syntagme : quelles frontières ?**

*la candidate Ségolène Royal, Professeur Paolucci*

*George W. Bush Jr., La Mecque, l'Abbé Pierre*

- **Une entité : quelle unité lexicale ?**

*Jacques Chirac, Monsieur Chirac, le Président Jacques Chirac,*

*le Président français, le Président de la République française, Chichi*

→ **caractérisation imprécise, diversité des mentions**

# Les EN dans la langue : le problème des « polysémies »

- **Homonymie**

*Orange a invité M. Hollande.*

- **Métonymie**

*Leclerc a fermé ses magasins en Rhône-Alpes.*

- **“Facettes”**

*Le candidat Sarkozy, devenu chef de l'Etat, a changé de position sur la présence française au sein de la force internationale.*

→ **polyréférentialité**

- **Hétérogénéité des réalisations**

Les entités nommées ne se limitent pas à une catégorisation, une mention, une interprétation.

- **Hétérogénéité des points de vue**

- Formules définitoires sous la forme d'énumérations
- Caractérisations diverses (sens, forme)

→ **Question** : Que sont les entités nommées ?

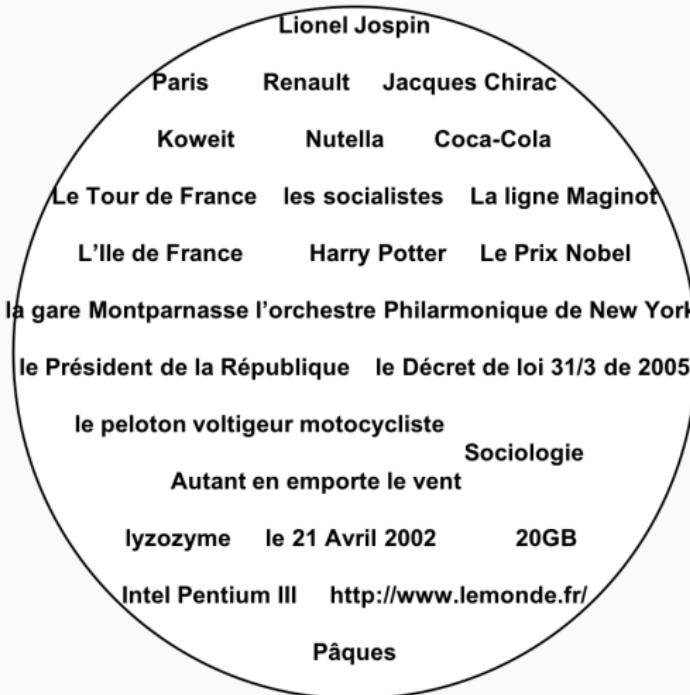
## **2. Définition**

---

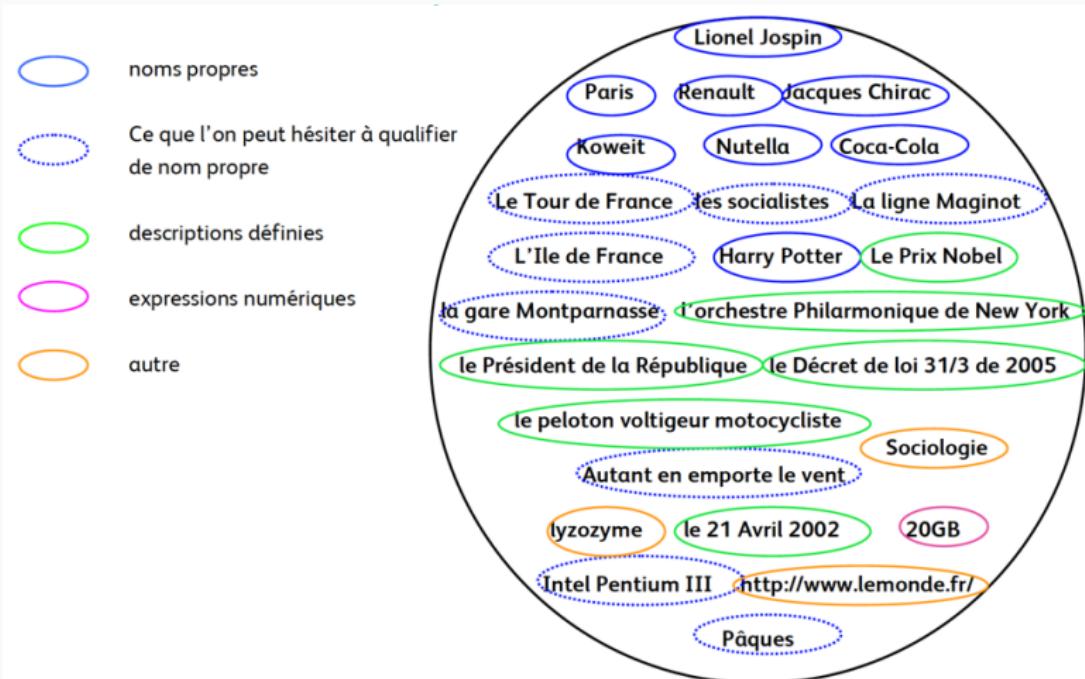
### **2.2 Vers une définition des entités nommées**

# Le “matériau” de départ

Unités lexicales  
considérées  
comme des entités  
nommées



# Le “matériau” de départ



## Proposition de définition

**Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.**

Questions que l'on s'est posées :

- Comment définir un objet TAL ?
- Que sont les noms propres et les descriptions définies ?
- Que devient le cadre linguistique du sens et de la référence en TAL ?

## Considération des aspects linguistiques

Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute **expression linguistique qui réfère à une entité unique** du modèle de manière **autonome** dans le corpus.

# Sens et référence en linguistique

- La **référence** désigne le lien qui existe entre une expression linguistique et l'élément du réel auquel elle renvoie.
- Le **sens** détermine les caractéristiques qu'une entité doit satisfaire pour pouvoir être désignée par telle ou telle expression.
- Un **modèle hétérogène** du sens (G. Kleiber)
  - sens descriptif (*oursin, table*)
  - sens instructionnel (*je*)
- **Comprendre** grâce à :
  - des connaissances lexicales
  - des connaissances sur le contexte
  - des connaissances sur le monde

- **Le nom propre réfère à un particulier**
  - **nomination d'un particulier** (Felix) vs. nomination d'une classe conceptuelle (chat)
  - **unicité** : une individualité considérée comme unique au sein d'une catégorie d'exsistants
  - **unité** : une individualité considérée comme formant un tout reconnaissable
- **Les descriptions définies**
  - présupposition d'existence et d'unicité  
*le président de la République, le père de Charles II, le marronnier*  
Une description de la forme "le tel et tel" présuppose qu'il existe une et une seule entité qui soit telle et telle

## Comment s'opère la référence à une entité unique ?

### Noms propres

- sens instructionnel dénominatif → connaissance d'une convention
- dénomination non contingente → désignateur rigide
- dénomination plus ou moins descriptive (*Massif Central*)

### Descriptions définies

- sens descriptif
- descriptions définies complètes et incomplètes  
*le président, le président de la République française en 2003*

- **L'ensemble 'entités nommées' n'est pas réductible à une catégorie linguistique**

'Plus que les noms propres et moins que les descriptions définies'

- **Caractérisation d'un comportement référentiel**

Référence à une entité unique et autonomie référentielle

*Jacques Chirac, le Président de la République, le costume bleu du président*

→ La perspective linguistique ne suffit pas

## Considération des aspects liés au TAL

Etant donné un **modèle applicatif** et un **corpus**, on appelle entité nommée toute **expression linguistique qui réfère à une entité unique** du modèle de manière **autonome dans le corpus**.

# Sens et référence en TAL

- **Caractérisation de la référence en TAL**

- restriction
- représentation

La référence en TAL désigne le lien qui existe entre une expression linguistique et l'élément du modèle auquel elle renvoie.

- **Comprendre en TAL grâce à**

- des ressources lexicales
- des ressources encyclopédiques
- des informations sur le contexte (issues du corpus)

- **Articulation sens–référence en TAL**

- entre le langage et le modèle
- trois mécanismes : segmentation, classification et reformulation

# Illustration

Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.

		le président de la République en 2005
Laguna	Jacques Chirac	Napoléon III
le président	je	30°
	l'Empereur des Français	2028hPa
Ivan	le président de la République en 2007	l'été 2004
	l'ouragan	Louise Colet

Application : générique « typique »

Modèle : Personnes, Lieux, Organisations

Corpus : journalistique français de 1998 à 2008

# Illustration

Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.

Laguna	le président de la République en 2005		
	Jacques Chirac	Napoléon III	
le président	je		30°
	l'Empereur des Français		2028hPa
Ivan	le président de la République en 2007		
	l'ouragan	Louise Colet	l'été 2004

Application : étude sur le climat

Modèle : températures, mesures atmosphérique, ouragan, dates, périodes, ...

Corpus : totalité des observations météorologiques sur une période données

# Illustration

Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.

		le président de la République en 2005
Laguna	Jacques Chirac	Napoléon III
	le président	je
		30°
	l'Empereur des Français	2028hPa
Ivan		le président de la République en 2007
	l'ouragan	Louise Colet
		l'été 2004

Application : « littéraire »

Modèle : personnes, lieux, événements

Corpus : correspondance de Flaubert

## De la linguistique au TAL, spécification d'un cadre théorique pour les EN :

- perspective linguistique : non réductibles à une catégorie mais caractérisables par un comportement référentiel
- perspective TAL : existent relativement à un modèle applicatif précis

→ Pas d'entité nommée « en soi », seulement des critères linguistiques et un modèle.

## Conséquences

- point de vue général : explication de l'hétérogénéité et de la variabilité de l'ensemble 'entités nommées'
- point de vue pratique : critères de décision pour annoter
- point de vue méthodologique : besoin impératif d'expliciter le modèle

### **3. Resources**

---

De quoi a-t-on besoin pour traiter les entités nommées?

1. **Typologies**, pour définir d'un cadre sémantique
2. **Corpus annotés**, pour servir de référence (évaluation) et d'illustration
3. **Lexiques et bases de connaissances**, pour donner des informations sur les éléments à traiter (entraînement)

## 3. Resources

---

### 3.1 Typologies

# Typologies: une façon de structurer

- Une typologie (ou *tagset*) est une **formalisation descriptive** des catégories d'EN à prendre en compte:
  - quoi reconnaître (cibler des éléments appartenant à des catégories spécifiques)
  - comment le représenter (pour un élément, choisir une catégorie parmi d'autres)
- De **multiples variations** en fonction des domaines et des applications
  - différences de catégories
  - différences de structure
  - différences sur la définition de ce que recouvrent les catégories

# Catégorie : une notion centrale

## Définitions:

- *une classe dans laquelle on range des objets de même nature* (Petit Robert)
- *un ensemble de choses qui ont un certain nombre de caractères en commun* (TLFi)

## Pour les EN:

- les traits communs entre les objets sont de nature sémantique.
- la détermination des catégories revient donc à spécifier des classes sémantiques.
- concrètement, les catégories sont les étiquettes utilisées pour annoter les EN, c'est à dire leur *type*

# Comment déterminer des catégories?

## Approches:

- *top-down*: on a une idée, on définit, et voilà.
- *bottom-up*: les catégories émergent des données
- mixte: on a des idées, on confronte au données, on remanie.
- utilisation de ressources : catégories issues des infobox de Wikipedia, de son équivalent sémantique DBpedia (200 classes), plus récemment de Wikidata.

## Dans les faits:

- très peu d'explications données sur l'élaboration des typologies
- influence thématique des financeurs
- seul Sekine [SSN02] a détaillé sa méthodologie de définition de sa typologie (200 !).

# Typologie MUC

- **noms propres** (ENAMEX) : lieux, personnes, organisations,
- **expressions numériques** (NUMEX) : dates et heures (expressions absolues), montants monétaires et pourcentages.

Types	Exemple	Contre-exemple
ORG	<b>DARPA</b>	our university
PERS	<b>Harry Schearer</b>	St. Michael
LOC	<b>U.S.</b>	53140 Gatchell Road
MONEY	<b>19 dollars</b>	ça en coûte 19
TIME	<b>8 heures</b>	la nuit dernière (+ MUC7)
DATE	<b>en juillet</b>	en <b>juillet</b> dernier (+ MUC7)

# Typologie ACE

- 4 nouvelles catégories par rapport à MUC :
  - **Geo-political Entity** (gpe)
  - **Facility** (fac)
  - **Vehicle** (veh)
  - **Weapon** (wea)
- introduction d'une **hiérarchie** parmi les types et sous-types (pers = individus, groupes, indéfinis) ;
- **distinction** entre les expressions numériques (NUMEX) et les expressions temporelles (TIMEX).

**N.B:** il n'y a pas une mais des typologies ACE (bcp d'évolutions)

# Typologie ACE

Types	Sous-types
PERS	individu, groupe, indéterminé
ORG	gouvernementales, commerciales, education, non gouvernementales, divertissement, media, religieuses, médical et sciences, sports,
GPE	continent, nation, état ou province, département ou région, villes, groupement de gpe, spécial, ainsi que des types comme pers, loc, org
LOC	adresses, frontières, objets astronomiques, plans d'eau, région géographique, région internationale, région autre
FAC	aéroports, usines, constructions, portion de construction
VEH	air, terre, eau, portions de véhicule, non spécifié
WEA	contondantes, explosives, coupantes, chimiques, biologiques, armes à feu, munitions, nucléaires, non spécifiés

## NOMBREUSES AUTRES TYPOLOGIES S'INSPIRANT DE MUC ET ACE:

- **CoNLL**: inspiration MUC, ajout d'une catégorie MISC
- **HAREM**: inspiration ACE, ajout de différentes catégories (Idée, Objet, Autre, Groupe)
- **ESTER-2**: encore plus de sous-types (e.g. pers.hum, pers.anim, loc.geo, loc.admin, etc) et traitement de l'imbrication

Refs: [TKSDM03, SSCV06, Est07]

## Imbrication d'entités (*nested entities*)

Au delà de la structuration en type et sous-types, il y a la **notion d'imbrication** :

- une entité peut en contenir une autre.
- *The <pers> president of <org> Ford </org> </pers>*

Structuration très utilisée dans des domaines de spécialité,  
e.g. la typologie GENIA (domaine bio-médical) [KOTT03].

## Vers une structuration fine des mentions

À la fin des années 2000, le programme Quaero définit une nouvelle typologie, utilisée dans la campagne ETAPE:

- inspiration ACE pour les catégories principales
- décomposition de la typologies (et de la tâche) en deux niveaux:
  1. caractérisation des types et sous-types (*type*)
  2. caractérisation des mots composants la mention (*composant*)

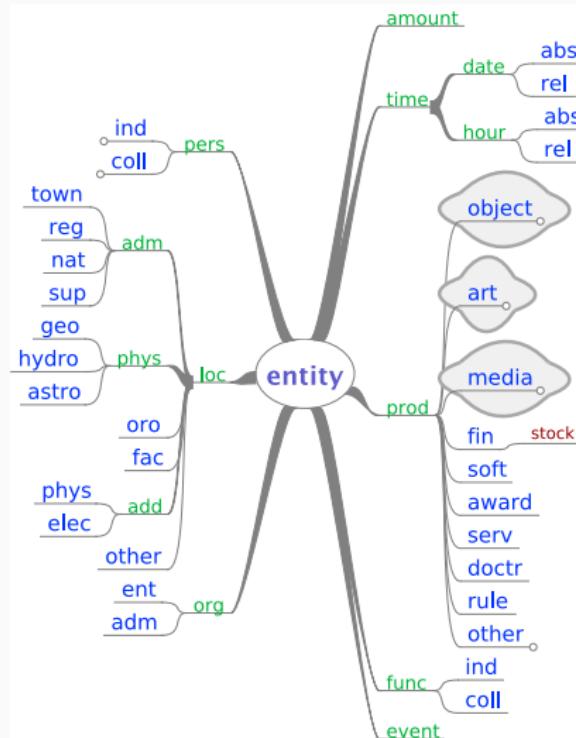
→ typologie hiérarchique et entités compositionnelles

Ref: [GRZ<sup>+</sup>11, RGZ11]

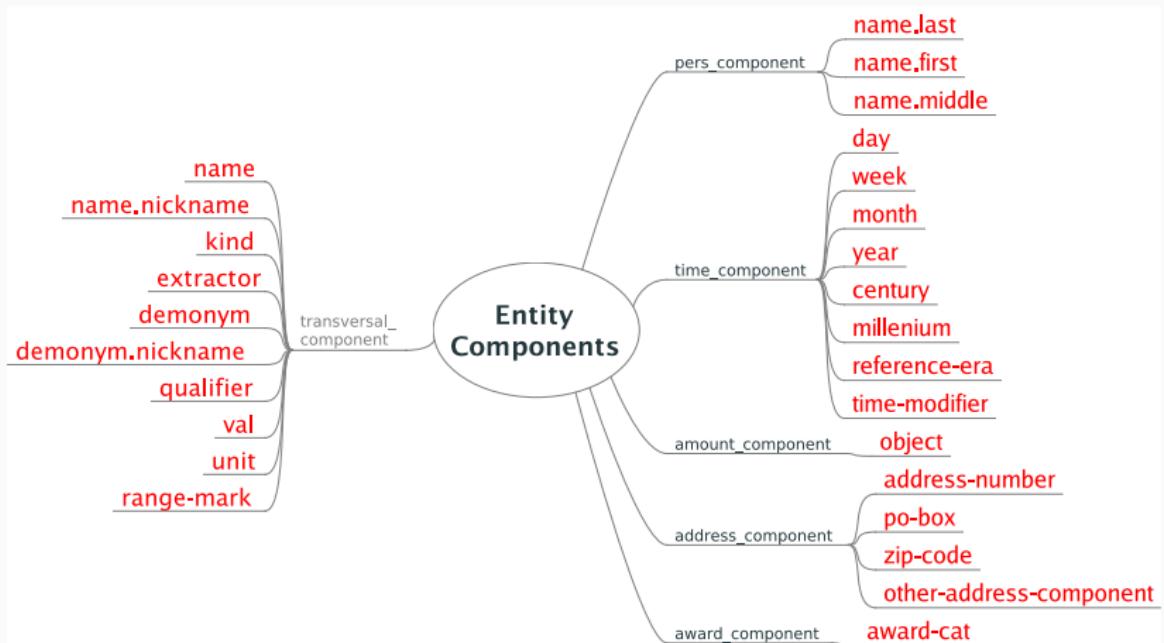
## 8 catégories principales

- **Personne:** personne individuelle, groupe de personnes;
- **Lieu:** lieu administratif, lieu physique, constructions, odonymes, adresse;
- **Organisation:** administration, service;
- **Expression temporelles:** dates/heures absolues et relatives;
- **Montants;**
- **Produits:** objet manufacturé, routes, produits financiers, doctrine, loi, software, art, media, récompense;
- **Fonction:** individuelle ou collective;

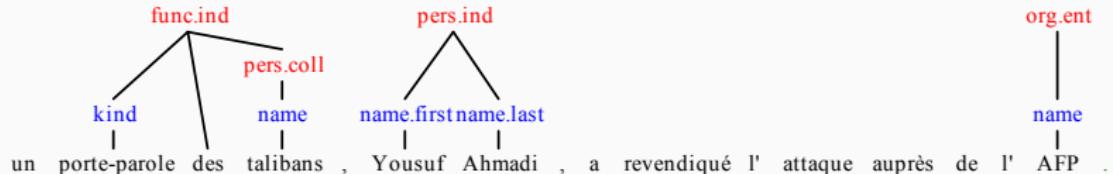
# Typologie Quaero: sous-types



# Typologie Quaero: composants d'entités



# Quaero: composants d'entités



Les composants permettent :

- d'avoir, par compositionnalité, de nombreux types sans les multiplier
- d'aider au suivi et à la liaison, au moins intra-documents (l'usine Renault → l'usine)

## Comparaison de typologies par l'exemple

MUC d'après le Bureau du recensement des LOC[Etats-Unis] , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE[2011] .

ACE d'après le ORG[Bureau du recensement des Etats-Unis] , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE[2011] .

ESTER d'après le ORG[Bureau du recensement des LOC[Etats-Unis]] , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE[2011] .

QUA d'après le ORG [name [Bureau du recensement] des LOC [ name[Etats-Unis] ] ] , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE[ year[2011]] .

# Principales questions et points de divergences

1. **Gestion de la métonymie:** doit-on annoter le **sens en contexte** ou le **sens absolu** d'une entité?

e.g. *France* peut référer au pays, à l'organisation politique, à une équipe sportive.

2. **Gestion des entités imbriquées**

3. **Gestion des entités coordonnées**

*Jacques et Bernadette Chirac*: une ou deux entités?

Les différentes typologies apportent des réponses très variées.

# Depuis 2009 : Text Analysis Conference Knowledge Base Population (TAC-KBP)

Pour une entité donnée, il importe de trouver de **nombreux attributs**.

E.g. pour une entité de type PERS:

- des **noms** : les autres noms que porte ou a porté cette personne (alias, faux noms, noms de scène, etc.) ;
- des **fonctions et activités** : ses emplois, ses occupations, etc. ;
- des **dates** (ou âge) : de naissance, de mort, des différents événements, son âge ;
- des **lieux** : les lieux en rapport avec des événements de sa vie comme la naissance, la mort bien entendu, mais aussi les différents emplois etc. ;
- des **personnes liées** : conjoint(e), enfants, autres membres de sa famille, etc. ;
- **autres informations** : les écoles et universités fréquentées, les pays visités, etc.

→ **un retour à la compréhension !**

## Un retour à la compréhension

- La tâche reste très complexe malgré les progrès (importants)
  - en 2010 le meilleur système ne dépassait pas 0.30 de F-mesure, en 2014 le meilleur score était de 0.36 [SJ14]
- Les EN restent au cœur du processus :

*This year's slot filling evaluation represented an effort at continuity (...) It remains difficult to achieve F-measure higher than 30%. Reaching competitive performance on this task requires a fairly mature NLP system, such as **high-quality name tagging**, coreference resolution and syntactic analysis. [JGDG10]*

## typologies: à retenir

- **Indispensable** à la tâche de NERC
- Fort héritage de **MUC** et **ACE**
- Grande **diversité** (plus de 20 typos inventoriées en 2016)...
- ...mais toujours **triade universelle** (person, organisation, lieu)
- Tendance à la **complexification** (imbrication, composants, knowledge base population)

Les typologies définissent le cadre d'action.

Elles sont indispensable à la création de *corpus*.

### **3. Resources**

---

#### **3.2 Corpus annotés**

# Corpus annoté

Un ensemble de documents textuels dont le texte est enrichi, lors d'une campagne d'annotation, par un marquage des entités nommées respectant une typologie donnée.

## Typologies —> Manuel d'annotation

- exemplification des catégories
- règles pour permettre à l'annotateur de faire des choix
- souvent, définition en parallèle de la typologie et de guide d'annotation

# Campagne d'annotation

- à partir d'outils dédiés (BRATT, GLOZZ, WEBANNO)
- importance de la mesure de la qualité et de la cohérence des annotations
- publication du corpus avec des informations: sources, accord inter-annotateur, mesures utilisées, typologie et guide d'annotation.
- à faire avec soin: time and resource consuming !

## Exemples de corpus français

- ESTER 2: broadcast news transcrives manuellement et automatiquement
- QUAERO: broadcast news / broadcast conversation (parole spontanée)
- ETAPE

## Vue d'ensemble des corpus existants

Recensement d'environ 160 corpus en 2016,  
avec différent(e)s:

- langues (mais prédominance de l'**anglais**)
- domaines (mais prédominance du **général**)
- modalités (mais prédominance de l'**écrit**)
- typologies
- formats
- licenses
- méthodes de construction

## **corpus: à retenir**

- indispensable pour entraîner et évaluer
- lien étroit avec les typologies
- coûteux à élaborer
- dominance des langues d'Europe de l'Ouest et de la modalité écrite

De quoi a-t-on besoin pour traiter les entités nommées?

1. **Typologies**, pour définir d'un cadre sémantique
2. **Corpus annotés**, pour servir de référence (évaluation) et d'illustration
3. **Lexiques et bases de connaissances**, pour donner des informations sur les éléments à traiter (entraînement)

### **3. Resources**

---

#### **3.3 Lexiques et bases de connaissances**

# Lexiques et bases de connaissances

Objectif: fournir des informations relatives à des entités, en général ou dans des domaines de spécialité, sur lesquelles les systèmes automatiques peuvent s'appuyer afin de les reconnaître, les catégoriser et les désambiguïser.

2 types d'informations:

- **lexicales**, sur les unités composant les EN
- **encyclopédiques**, sur les référents des EN

Un élément central pour la reconnaissance et la classification des EN (mais évolution avec le deep learning).

Evolution importante de ce type de ressource depuis l'apparition de la tâche:

simple 'gazetteers' → encodage de plus en plus d'information

Encodent 2 types d'information:

- des **noms ou parties de noms d'entités** avec leurs types associés  
→ directement utilisés pour reconnaître des unités équivalentes dans les textes
- des **mots amorces**, également avec leurs types associés. E.g. *Justin Trudeau*  
→ des unités indiquant avec une forte probabilité la présence d'une entité d'un certain type. E.g. *Monsieur*

# Constitution de bases lexicales

- forte **dépendance v-a-v du domaine** d'application  
e.g. liste de mots amorces pour le domaine général vs. bio-médical
- défi 1: privilégier la **qualité** sur la quantité:  
un petit nombre d'entrées suffit à reconnaître une majorité d'entités
- défi 2: se conformer à l'**évolution rapide** des entités nommés qui  
sont une classe ouvert

## Quelques exemples de ressources lexicales

- **WordNet** (Princeton): éloigné du monde de EN, mais utile pour l'intégration de resources
- **PROLEX** (Univ. Tours): base d'EN multilingue (en, fr, pl)
- **Geonames**: toponymes et assimilés, 7M d'entités et 10M d'entrées lexicales

- un 'by-product' d'un système de veille médiatique:  
7000 sources, 300k articles par jour, 70 langues, dont 21 avec traitement fin des entités nommées
- ca. 340,000 entités uniques (PERS et ORG)
- 1,7 million de variantes de noms (lexicalisations) dans 170 langues
- 32 millions relations cross-lingue, y compris entre différents jeux de caractères
- jusqu'à 400 variantes pour une entité



## emm



ESTD. 2007

## Top Stories

UPDATED EVERY 10 MINUTES, 24 HOURS PER DAY.

Main Menu

- Top Stories
- 24 Hours Overview
- Events Detection
- Most Active Themes
- Help about EMM
- Overview
- Advanced Search
- Sources list
- Web Site Map

EU Focus

EU Policy Areas

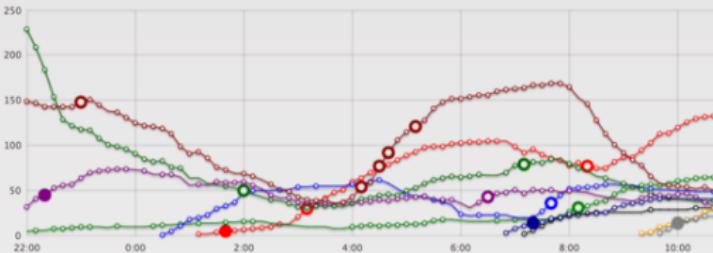
Themes

The World

Offices & Agencies

Current top 10 stories

Language: en Period: Jun 15, 2018 10:40 PM – Jun 16, 2018 10:40 AM



Multimillion-pound restoration hit by blaze at Mackintosh Building – fire chief Fire has caused "extensive" damage at Glasgow's famed Mackintosh Building...  
enCA | Fire rages through historic Scottish school of art ↗  
enCA Saturday, June 16, 2018 10:44:00 AM CEST | Info [other]  
Entities: Peter Capaldi[1]; Harry Potter[1]; James Bond[1]; Robbie Coltrane[1]; Nicola Sturgeon[1]; Paul Sweeney[1]; Simon Starling[1]; Martin Boyce[1]; Franz Ferdinand[1]; David Mundell[1]; Richard Wright[1]; Charles Rennie Mackintosh[2];  
LONDON – Fire ripped through one of the world's top art schools, the Glasgow School of Art in Scotland, late on Friday. The historic building -- designed by Art Nouveau architect Charles Rennie Mackintosh -- was undergoing major restoration work following a blaze four years ago....  
More articles...

Saudi-led forces seize airport in Yemen port city of Hodeida  
expressindia Saturday, June 16, 2018 10:21:00 AM CEST | Info [other]

Tools

Saturday, June 16, 2018 10:57:00 AM CEST

RSS | MAP

Facebook

subscribe | manage

info

Available on the App Store ANDROID APP ON Google play

Languages

Select top stories in other languages.

ar	bg	cs	da	de	el
en	es	et	fi	fr	hr
hu	it	lt	lv	mt	nl
pl	pt	ro	ru	sk	sl
sv	sw	tr	zh		

Show additional languages

Interface: en - English

Legend

Country Watch

The country most in the news at the moment.

<http://emm.newsbrief.eu>

Ehrmann, Rosset

Ecole thématique 'Big Data and Speech', Roscoff, Juillet 2018

80

## Main Menu

[Top Stories](#)  
[24 Hours Overview](#)  
[Events Detection](#)  
[Most Active Themes](#)  
[Help about EMM](#)  
[Overview](#)  
[Advanced Search](#)  
[Sources list](#)  
[Web Site Map](#)

## EU Focus

## EU Policy Areas

## Themes

## The World

## Offices & Agencies

# Nicola Sturgeon

Last updated on 2018-02-21T08:07+0100.



ABOUT THIS IMAGE  
LICENSING UNKNOWN  
AUTHOR: THE SCOTTISH GOVERNMENT

### Extracted quotes from

Nicola Sturgeon said : "not listened to, who is responsible and how are we going to ensure individuals are accountable?" [\[link\]](#)  
thecourier Thursday, June 14, 2018 6:35:00 PM CEST

Nicola Sturgeon said : "Yesterday morning I was spending my time in two primary schools, as well as a secondary school and an early years centre. And I was talking to a range of primary school children including some five-year-olds. "I didn't meet any of them in tears, it didn't see any of them that looked crushed. What I saw were confident, bright enthusiastic young people - some of those were showing me computer coding and some were speaking Mandarin, that is how confident they were" [\[link\]](#)  
bbc Thursday, June 14, 2018 4:32:00 PM CEST

Nicola Sturgeon said : "As local MSP for the Gorbals, I'm in close contact with NewGorbalsSHA who are on site." [\[link\]](#)  
dailystar Thursday, June 14, 2018 1:53:00 PM CEST

Nicola Sturgeon said : "As local MSP for the Gorbals, I'm in close contact with NewGorbalsSHA who are on site." [\[link\]](#)  
dailystar Thursday, June 14, 2018 12:21:00 PM CEST

### Key Titles and Phrases (Last 30)

### Names (Top 30)

KEY TITLES AND PHRASES	COUNT	LANG	LAST SEEN
minister	47.38%	EN	15/06/2018
leader	9.97%	EN	15/06/2018
première ministre écossaise	4.26%	FR	15/06/2018
ministre écossaise	3.87%	FR	15/06/2018
first minister of scotland	1.72%	EN	14/06/2018
minister of scotland	1.06%	EN	14/06/2018

### Related entities (Top 30)

### Associated entities (Top 30)

TYPE	ENTITY NAME	COUNT
EU	EU	7.23%
EU	Glasgow School	5.40%
EU	Charles Rennie Mackintosh	5.30%
EU	Ian Blackford	4.18%
EU	Theresa May	4.18%
EU	Paul Sweeney	3.97%

Articles published more than 12 hours ago

## Tools

Saturday, June 16, 2018  
10:58:00 AM CEST

Facebook

manage



## Languages

Select your languages

am	ar	az	be	bg	bs
ca	cs	da	de	el	en
eo	es	et	fa	fi	fr
ga	ha	he	hi	hr	hu
hy	id	is	it	ja	ka
km	ko	ku	ky	lb	lo
lt	lv	mik	ml	mt	nl
no	pap	pl	ps	pt	ro
ru	nw	si	sk	sl	sq
sr	sv	sw	ta	th	tr
uk	ur	vi	zh		
all					

Interface:

en - English

Legend

### Explore Relations



### Extracted quotes about

Adam Tomkins said ( about Nicola Sturgeon ) : "This is a remarkable report which exposes Nicola Sturgeon's secret Scotland. "People will see this report and

- **Médical:** Meta-thesaurus [UMLS](#)  
unifie environ 200 terminologies (*National Library of Medicine*)
- **Biologie:** [UniProt](#), [SwissProt](#), bases lexicales sur les gènes et les protéines

## Biologies

- bases lexicales sur les gènes et les protéines
- **UniProt**<sup>1</sup> (*Universal Protein Resource*) [C+10]
- avec des entrées validées manuellement:  
**SwissProt**, 548 454 entrées
- et d'autres obtenues automatiquement:  
**TrEMBL**, 47 452 313 entrées.

---

<sup>1</sup><http://www.uniprot.org/help/about>

# Bilan sur les bases lexicales

- Ne sont pas toutes répertoriées ou publiées
- Au départ: décrire les réalisations linguistiques possibles d'entités
- Evolution: enrichissement avec d'autres informations, e.g. date de naissance d'une personne, population d'une ville, etc.
- 3 directions d'enrichissement:
  - couverture plus large
  - multilinguisme
  - information encyclopédique

→ structures de données plus complexes et volumineuses

# Les bases de connaissances (survol rapide)

- **Wikipedia** (initiée en 2001)
  - utile pour extraire et intégrer des lexiques d'EN
  - constitution semi-automatique de corpus annotés
  - acquisition de relations entre entités
- **DBpedia** (équivalent RDF de Wikipedia)
- **YAGO** (Wikipedia, WordNet, plus infos spatiales et temporelles)
- **BabelNet**
- **Wikidata**
- **OpenCyc** (partie libre du Cyc), information de 'sens commun'

## **base lexicales et de connaissances: à retenir**

- troisième pilier des ressources pour les EN
- information lexicales et sémantiques
- difficile à acquérir, représenter, stocker et utiliser jusqu'au milieu 2000
- aujourd'hui: explosion d'information, principalement pour le domaine général

# SESSION 2

## **4. Reconnaissance et classification**

---

# Traitement des EN - rappel des tâches

1. **reconnaissance**: détecter, repérer des entités nommées dans les flux textuels (on pose les frontières dans le texte)
2. **classification**: catégoriser les éléments reconnus selon des catégories sémantiques pré-définies (on affecte un type)
3. **désambiguisation/liaison**: lier les mentions d'entités à une référence unique (on lie à une référence)
4. **extraction de relation**: découvrir des relations entre entités (*father-of, born-in, alma mater*)

# Objectifs

**Construire des systèmes logiciels qui effectuent ces tâches de manière automatique.**

Exigences:

- **qualité**: ne pas faire trop d'erreurs
- **exhaustivité**: ne pas manquer trop d'entités
- **robustesse**: ne pas échouer face à des cas non canoniques

En pratique:

- difficile de répondre à ces 3 exigences simultanément
- recherche du **meilleur compromis** en fonction des ressources et de l'application.

## Matériaux de base

texte, représenté comme une **structure linéaire**, i.e. une séquence de mots pouvant être découpée en **segments**

## Objectif

indices portés par les segments  $\iff$  présence d'une entité nommée  
 $\iff$  d'une certaine catégorie  
 $\iff$  référant à une certaine entité

attention: pas de correspondance systématique entre un ensemble de propriétés et une classe d'entités

## Vue d'ensemble des méthodes

- méthodes symboliques à base d'automates
- méthodes numériques (heuristiques ou pondérations)
- méthodes statistiques, *i.e.* apprentissage automatique, supervisé, semi-supervisé ou non supervisé

## **4. Reconnaissance et classification**

---

### **4.1 Indices**

# Représentation du texte

La **représentation des textes** comme séquences de mots donne 2 niveaux de granularité:

- les **caractères**, qui forment un mot
- les  **mots**, qui composent une séquence (un texte)

Les **indices** peuvent être caractérisés au niveau:

- des caractères: **indices morphologiques**
- des mots eux-mêmes: **indices lexicaux**
- de la séquence de mots: **indices contextuels**

# Indices morphologiques

C'est à dire:

- les caractères qui constituent les mots
- l'existence de différentes classes de caractères
- l'existence de régularités dans l'agencement des caractères, dont certaines sont utiles pour les EN

# La majuscule

- très utilisée dans les jeux de caractères occidentaux pour annoncer un nom propre
- très facile à tester pour détecter des EN

MAIS

- majuscules également présentes en début de phrase, dans les sigles
- utilisée pour les noms communs en allemand
- n'aide pas à la classification
- la notion de casse est minoritaire en orient (n'existe pas en chinois, hindi, arabe, etc.)

- le suffixe *-ville* ou le préfixe *Saint-* en français
- suffixes réguliers pour les noms de personnes
  - en russe, le suffixe *-vitch*
  - en suédois, le suffixe *-sson*
  - en islandais, *-dóttir*
  - en Afrique du nord, les préfixes *Ben-* ou *Aït-*
  - en japonais, le suffixe *-san*
- mots issus de conventions ou de normes :  
*Inc.* en anglais, *S.A.* en français, *GmbH* en allemand pour les ORG

- les dates, montants ou mesures, contiennent typiquement des nombres (écrits en chiffres ou lettres)
- chiffres plus ou moins soudés (*10 000, quatre-vingt-treize, cent dix-huit*) et/ou avec des caractères spécifiques (*10,38, 24/03*).
- mélange de caractères numériques et alphabétiques (*100km, 10h30, etc.*).
- sigles et d'abréviations (*A380, ISO-9000, Canon EOS 70D*).

- **En général:** attribution d'*étiquettes* aux mots pour distinguer e.g. les noms propres, chiffres, adverbes, déterminants, noms communs, etc.
- **Appliquée aux EN:**  
utilisation de motifs morphologiques et des motifs morphologiques résumés:  
i.e. quelles combinaisons de classes de caractères (alphabétiques, ponctuations, chiffres) sont utilisées pour former quels mots  
→ mécanismes de reconnaissance à l'aide d'*expressions régulières*

# Bilan indices morphologiques

Détection d'entités caractérisées par des régularités, pour un certain nombre de langues.

Mais ils restent insuffisants:

- ne couvrent que les formes très normées d'EN
- permettent de *déetecter*, mais pas de *catégoriser*

**Principe:** confronter les textes à des listes d'entités de ou de composants d'entités.

- mécanisme très précis si entrées lexicales contrôlées
- lexiques souvent organisés selon les types d'EN ou le degré d'ambiguïté
  - e.g. *Hollande vs. Obama*
- attention à trouver le bon compromis entre quantité et efficacité

**En pratique:** les algorithmes retournent la liste des segments correspondant aux occurrences des entrées des lexiques.

Attention: mécanisme est ambigu, plusieurs entrées peuvent être retournées.

*Aujourd'hui, François Hollande a rencontré Obama à Washington.*

- la Personne *François Hollande*
- le Pays *Hollande*
- la Personne *Obama*
- la Personne ou le Lieu *Washington*.

**Défi:** les EN sont une classe **ouverte**, impossible d'être exhaustif

- de nouveaux noms propres se créent continuellement  
[McD96, Fri02]
- certaines parties de descriptions définies sont substituables

→ tenir à jour un lexique d'EN avec toutes leurs formes exactes et leurs variations est une tâche coûteuse et complexe.

Souvent, **prise en compte conjointe** d'indices morphologiques et lexicaux:

- **Personnes** : le premier mot est un prénom, le second un nom propre
- **Dates** : le premier et le dernier mot sont composés de chiffres, le second mot fait partie de la liste des noms de mois (*5 juillet 2012*)
- **Lieux** : contient *sur* ou *en* suivi d'un nom de cours d'eau (*Montlouis sur Loire*)
- etc.

L'examen des mots qui composent les entités ne dit pas tout.  
Les indices morpho et lexicaux peuvent être absents ou ambigus.

→ besoin d'indices complémentaires, à proximité:

- **contexte local:** mots qui précèdent ou suivent l'entité.
- **contexte global:** phrase, phrases proches, paragraphe, document.

# Importance des indices contextuels

1. *Il a vu Hollande à la télévision.*
2. *Son voyage en Hollande s'est bien passé.*
3. *Il a acheté une Renault Clio.*
4. *La muse Clio chante le passé des hommes et des cités.*
5. *Je me documente sur Washington pour mon travail.*

La graphie de l'entité étant identique, seul un appel au contexte permet de classifier.

Facile et intuitif pour l'humain, plus compliqué pour une machine.

- traitement appliqué non aux mots, mais à des pans de textes  
→ coût computationnel plus élevé
- souvent besoin de s'appuyer sur des analyses préliminaires:  
syntaxique, coréférences, thématique du document
- plus économique: sélectionner, *a priori* ou *a posteriori* les indices contextuels les plus discriminants et leurs combinaisons
- analyses en contexte importante lorsque la typologie des EN est fine

## **indices: à retenir**

- de nature morphologique, lexicale ou contextuelle
- possibilité d'indices composites, par conjonction ou disjonction
- ce sont les 'ingrédients' des systèmes automatiques de traitements d'EN

## **4. Reconnaissance et classification**

---

### **4.2 Approches symboliques**

# Techniques à base d'automates

- **Objectif:** insertion de balises dans les textes indiquant où se trouvent les ENs
- **Principe:** conception de *règles* formant un *grammaire locale*
- De nombreuses **boîtes à outils**:
  - GATE
  - LingPipe
  - NooJ
  - OpenNLP
  - OpenCalais
  - Unitex
  - WMatch

Possibilité d'avoir des prétraitements:  
segmentation en mots, en phrases, étiquetage morphosyntaxique.  
  
→ indices supplémentaires fort utiles,  
mais qui impactent les performances si bruités.

## Basculement vers les approches statistiques

Au début des années 2000, grâce à la mise à disposition de jeux de données volumineux.

Mais les approches symboliques sont toujours présentes:

- combinées avec des méthodes statistiques
- prédominent pour les langues ou les typologies sans corpus de données suffisants
- gardent l'avantage pour le contrôle et de l'ingénierie: plus compréhensibles, modulables, possibilités de réglages fins.
- majoritaires dans le milieu industriel.

## **4. Reconnaissance et classification**

---

### **4.3 Modèles guidés par les données et apprentissage**

# Le paradigme de l'apprentissage automatique

**Objectif:** déterminer les paramètres d'un modèle à partir de données,  
d'où le terme *apprentissage*

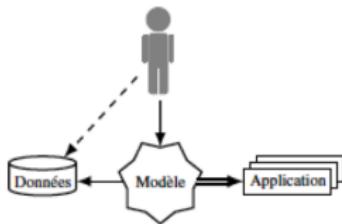
Ces paramètres et ce modèle sont ensuite utilisés pour prendre les décisions les plus probables (ou vraisemblables) sur de nouvelles données à traiter.

Il s'agit, simultanément, de spécifier le modèle et de généraliser les données.

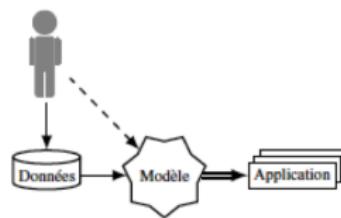
# Le paradigme de l'apprentissage automatique

**Systèmes symboliques:** le concepteur du système interagit majoritairement avec le modèle (l'automate), et n'utilise les données que pour visualiser ou d'évaluer.

**Systèmes guidés par les données:** le concepteur agit sur les données, la structure du modèle est prédéfinie et rigide et les paramètres ajustés automatiquement à partir des données.



Système symbolique



Système guidé par les données

# Approches existantes

La détection et reconnaissance des EN revient à une **tâche de classification**.

- arbres de décision
- modèles probabilistes
- réseaux neuronaux

Déterminer la classe d'un mot à partir de la classe qui lui est majoritairement associée dans le corpus d'apprentissage.

Formulation à l'aide de probabilités:

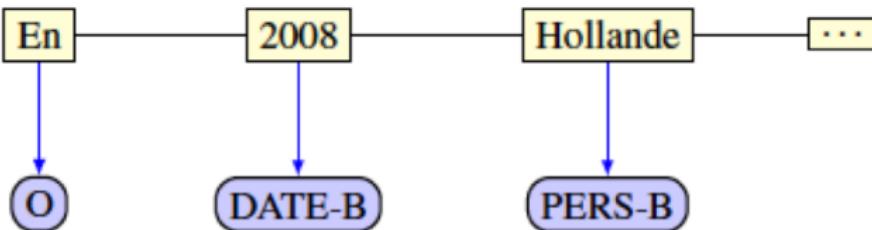
- fréquence du mot  $F(m)$
- fréquence d'une étiquette  $F(e)$
- fréquence de la présence jointe du mot et de l'étiquette  $F(m, e)$

La formule de Bayes et l'estimation statistique permettent de calculer la probabilité d'une étiquette sachant le mot :

$$P(E_i = e|M_i = m) = \frac{P(M_i = m, E_i = e)}{P(M_i = m)} = \frac{F(e, m)}{F(m)}$$

Probabilité d'une étiquettes pour un mot donné =  
ratio entre la fréquence dans le corpus annoté du mot avec une étiquette  
et la fréquence dans ce même corpus du mot (quelque soit l'étiquette)

## Modèles par classes majoritaires



**Figure 4.3. Modèle par classes majoritaires**

(l'orientation des flèches indique quelles dépendances sont prises en compte par le modèle)

**Objectif:** tenir compte de la vraisemblance d'**étiquettes contiguës**

*François Hollande*

- *Hollande*: Lieu ou Personne ?
- *François*, annoté comme Personne, peut conditionner l'annotation du mot *Hollande*

# Modèles à décisions contextuelles (HMM)

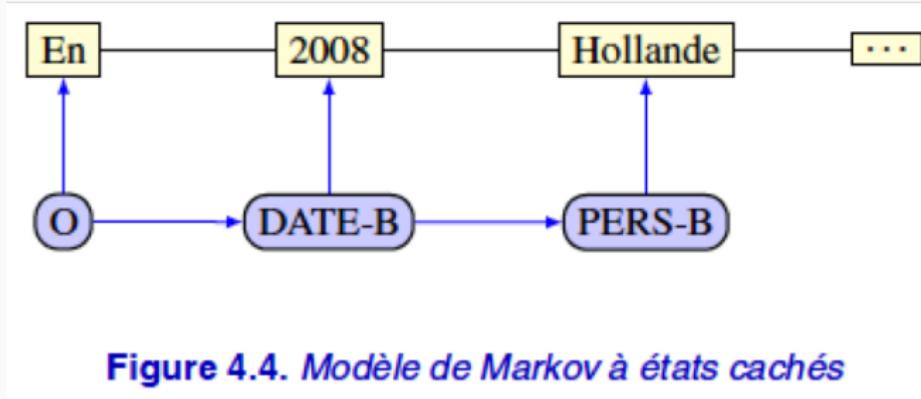
**Option:** modèles génératifs comme les modèles de Markov à états cachés.

**Calcul des probabilités inversé** : déterminer, pour une suite d'étiquettes, la probabilité qu'elle génère un texte donné.

$$P(M_1, M_2 \dots M_n | E_1, E_2 \dots E_n) = \prod_{i=1}^n P(M_i | E_i) * P(E_i | E_{i-1})$$

Soit le produit des probabilités de génération  $P(M_i | E_i)$  et de transition  $P(E_i | E_{i-1})$ .

# Modèles à décisions contextuelles (HMM)

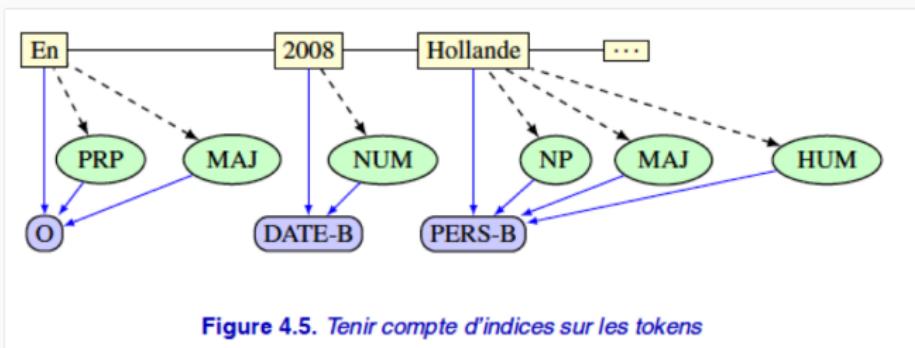


**Figure 4.4. Modèle de Markov à états cachés**

Décisions non indépendantes : la solution la plus vraisemblable est choisie en fonction des étiquettes préalablement choisies.

# Modèles utilisant des indices multiples (softmax, MaxEnt)

**Objectif:** considérer plus d'indices que les mots, i.e. prendre en compte la morphologie, les indices lexicaux, le contexte, etc.



# Champs markoviens conditionnels (CRF)

Les CRF (*Conditional Random Fields* ou champs markoviens conditionnels) combinent les deux aspects précédents :

- **tenir compte du contexte** pour prendre des décisions  
(une décision sur un mot influence la décision pour le mot suivant)
- **tenir compte de multiples indices**  
(analyses en prétraitements)

Modèle qui obtient de très bonnes performances pour la reconnaissance d'EN.

Ref: Lafferty, MacCallum, Pereira, 2001.

# Champs markoviens conditionnels (CRF)

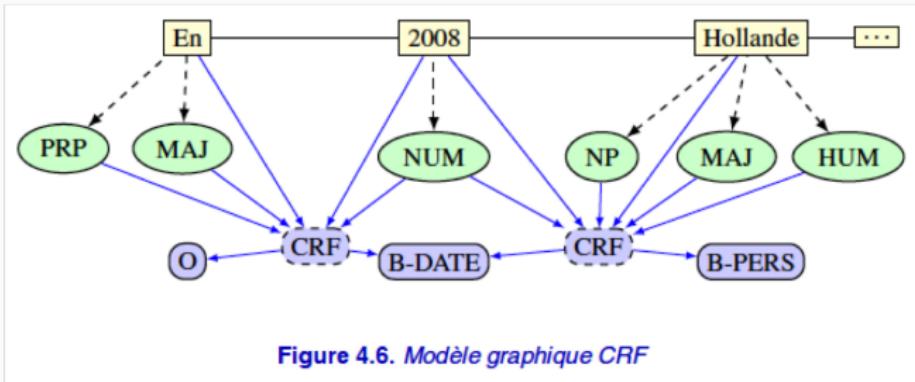


Figure 4.6. Modèle graphique CRF

$$G(e, m, f_1 \dots f_k) = \exp \left( \sum_{p=1}^k \alpha_{ep} * f_p \right)$$

## **Reconnaissance et classification d'en: à retenir**

- possibilité d'utiliser de nombreux indices
- via des méthodes diverses qui peuvent être combinées.
- si plus d'indices, alors complexité grandissante et besoin de plus de données annotées
- importance de la sélection d'indices

## 5. Liaison

---

## Où en sommes-nous

- nous savons reconnaître et catégoriser des segments textuels:  
des *mentions* d'entiés qui font référence à un objet du monde.
  - ce qu'il reste à faire: établir le lien entre les mentions et les objets auxquels elles réfèrent
- objectif: désambiguïsation, résolution, liaison

# Des mentions aux référents

- **Catégoriser n'est pas désambiguer:**

*G. Bush et F. Mitterrand* sont des PERSON

Mais lequel des 2 réfère au *43ème président des États-Unis*?

- **Le problème des homonymes:**

*F. Mitterrand* est une PERSON

Mais *François Mitterrand* ou *Frédéric Mitterrand* ?

*Bush* est une PERSON

Mais *G. W. Bush* ou *G. Bush* ?

- **Le problèmes des variantes:**

*Jean-Claudem Junckerem, Juncker, Jean-Cluade Juncker et le président de la Commission Européenne* réfèrent-elles à la même entité?

# Le point sur les tâches

- **Résolution de co-référence:**  
au sein d'un même document, identifier que *Frédéric Mitterrand*, *Mitterrand*, *FM* ont le même référent (quel qu'il soit)
- **Clustering de mentions:**  
pour une collection de documents, identifier que *Frédéric Mitterrand*, *Mitterrand*, *FM* ont le même référent (avec ou sans référentiel)
- **Liaison d'entités:**  
étant donnés des documents, identifier les mentions d'entités et lier chacune d'elles à un référent d'une base de connaissances

## 6. Evaluation

---

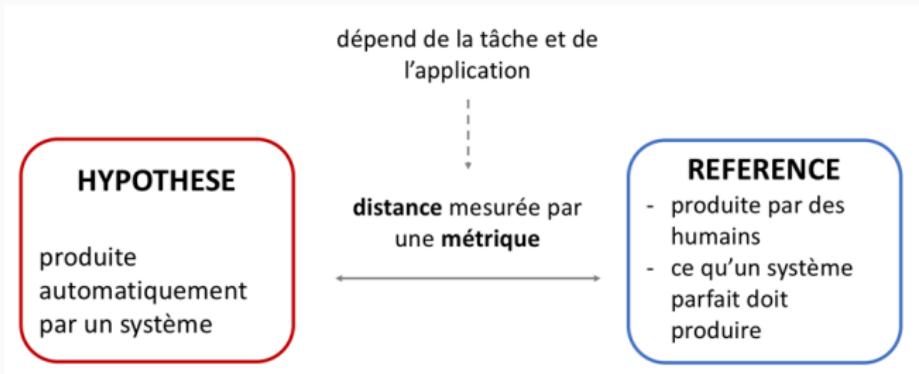
## **6. Evaluation**

---

### **6.1 Introduction**

- Première formalisation de la procédure d'évaluation: MUC3 [Sun91]
- Motivation: avoir des éléments de comparaison stables et effectifs entre hypothèses et références

# Protocole d'évaluation



**Objectif:** mesurer à quel point le système trouve les “bonnes réponses”

Quelle “bonnes réponse”?

- traduction ou le résumé automatique : bonne réponses multiples
- EN: on peut supposer une seule et unique “bonne réponse”

- **Transparence:** "règles du jeu" connues par tous
- **Coût:** réduit par rapport à une évaluation manuelle pour chaque hypothèse des systèmes ;
- **Reproductibilité:** réutilisation au delà des campagnes permettant une comparaison des résultats dans la production scientifique

# Ce qu'il faut pour évaluer

---

1. Une **métrique** mesurant la distance entre une référence et une hypothèse ;
2. Un **algorithme d'alignement** de la référence et de l'hypothèse.
3. Un **algorithme de projection** des entités annotées sur la transcription manuelle de référence vers la transcription automatique

## 6. Evaluation

---

### 6.2 Les mesures classiques

## Précision

Ratio entre le nombre de **réponses correctes** et toutes les **réponses données** par un système

$$P = \frac{C}{C + S + I} \quad (1)$$

- $C$  : nombre d'objets **corrects** dans l'hypothèse;
- $I$  : nombre d'**insertions** par le système ;
- $S$  : nombre de **substitutions** par le système (entités mal typées).
- soit  $C + S + I$  : nombre total d'objets dans l'hypothèse.

## Rappel

Ratio entre le nombre de **réponses correctes** et le nombre des **réponses attendues** (i.e. présentes dans la référence)

$$R = \frac{C}{C + S + D} \tag{2}$$

- $D$  : nombre total d'**omission** (*deletions*) opérées par le système (entités non détectées) ;
- $C + S + D$  : nombre total d'objets dans la référence.

## Exemple 1

REF: <pers> Bertrand Delanoë </pers> a été élu maire de <loc>  
Paris </loc>

HYP1 : <pers> Bertrand Delanoë </pers> a été élu <pers> maire  
</pers> de <loc> Paris </loc>

- Precision =  $\frac{2}{3} = 0,67$
- Rappel =  $\frac{2}{2} = 1$

→ ici HYP1 produit du **bruit**

## Exemple 2

**REF:** <pers> Bertrand Delanoë </pers> a été élu maire de  
<loc> Paris </loc>

**HYP2:** <pers> Bertrand Delanoë </pers> a été élu maire de  
Paris

- Precision = 1
- Rappel =  $\frac{1}{2} = 0.5$

→ HYP2 produit du **silence**

- La **précision** tient compte des **insertions** et **substitutions**
- Le **rappel** tient compte des **omissions**

Comment combiner les 2 en une seule mesure?

**F-mesure**, définie comme la **moyenne harmonique entre Précision et Rappel**:

$$F = (1 + \beta^2) \times \frac{P \times R}{\beta^2 P + R} \quad (3)$$

Où  $\beta$  est un **poids** permettant d'ajuster l'importance de P ou R (si 1, égale importance).

## Exemples

**REF:** <pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris </loc>

**HYP1 :** <pers> Bertrand Delanoë </pers> a été élu <pers> maire </pers> de <loc> Paris </loc>

**HYP2:** <pers> Bertrand Delanoë </pers> a été élu maire de Paris

$$F(HYP1) = (1 + 1^2) \times \frac{0,67 \times 1}{1^2 \times 0,67 + 1} = 0,80 \quad (4)$$

$$F(HYP2) = (1 + 1^2) \times \frac{1 \times 0,5}{1^2 \times 1 + 0,5} = 0,67 \quad (5)$$

## Inconvénients des mesures classiques

- Fusionner P et R minimise le poids des erreurs d'insertion et d'omission par rapport aux erreurs de substitution, quel que soit  $\beta$  [MKS99]
- Avec les typologies fines et complexes, besoin d'une métrique différenciant les erreurs.

# Différents types d'erreur

**REF**: the <pers.ind> president of Ford </pers.ind>

**HYP1** : the <pers.ind> president </pers.ind> of Ford  
→ erreur de frontière

**HYP2** : the <pers.coll> president of Ford </pers.coll>  
→ erreur de sous-type

**HYP3** : the <pers.coll> president </pers.coll> of Ford  
→ erreur de sous-type et de frontière.

## **6. Evaluation**

---

### **6.3 Les mesures basées sur le décompte d'erreurs**

## SER: Slor Error Rate

- proposée par [MKSW99]
- identique au WER utilisé en RAP
- utilisée lors de ACE, ESTER-2, QUAERO et ETAPE
- suppression du nombre d'insertion ( $I$ ) du dénominateur:

$$SER = \frac{S + D + I}{C + D + S} = \frac{S + D + I}{R} \quad (6)$$

où  $R$  = nombre total d'entités de la référence.

Possibilité d'affiner l'importance relative des erreurs:

$$SER = \frac{\alpha_1 S_t + \alpha_2 S_f + \beta D + \gamma I}{R} \quad (7)$$

- $S_t$  et  $S_f$ : nombre total de substitution de type et de frontières ;
- D et I: nombre total d'omissions et insertions ;
- $\alpha_1$   $\alpha_2$   $\beta$  et  $\gamma$  : poids affectées à chaque catégorie d'erreur.

## **6. Evaluation**

---

### **6.4 Zoom sur données de parole**

# Corpus et campagnes d'évaluation

- Assez peu de corpus et de campagnes d'évaluation
  - en France : ESTER 1 et 2, ETAPE (+ QUAERO), REPERE (pour les personnes, multimodal)
  - à l'international : campagne ACE (2000-2008)
- Difficile de comparer REN sur textes et REN sur parole car on ne dispose pas de corpus et campagnes comparables (types de données + typologies)
- Résultats nettement différents entre transcriptions manuelles et transcriptions automatiques

## REN sur transcriptions automatiques et manuelles

Pour comparer simplement, utilisée par [BJ15] :

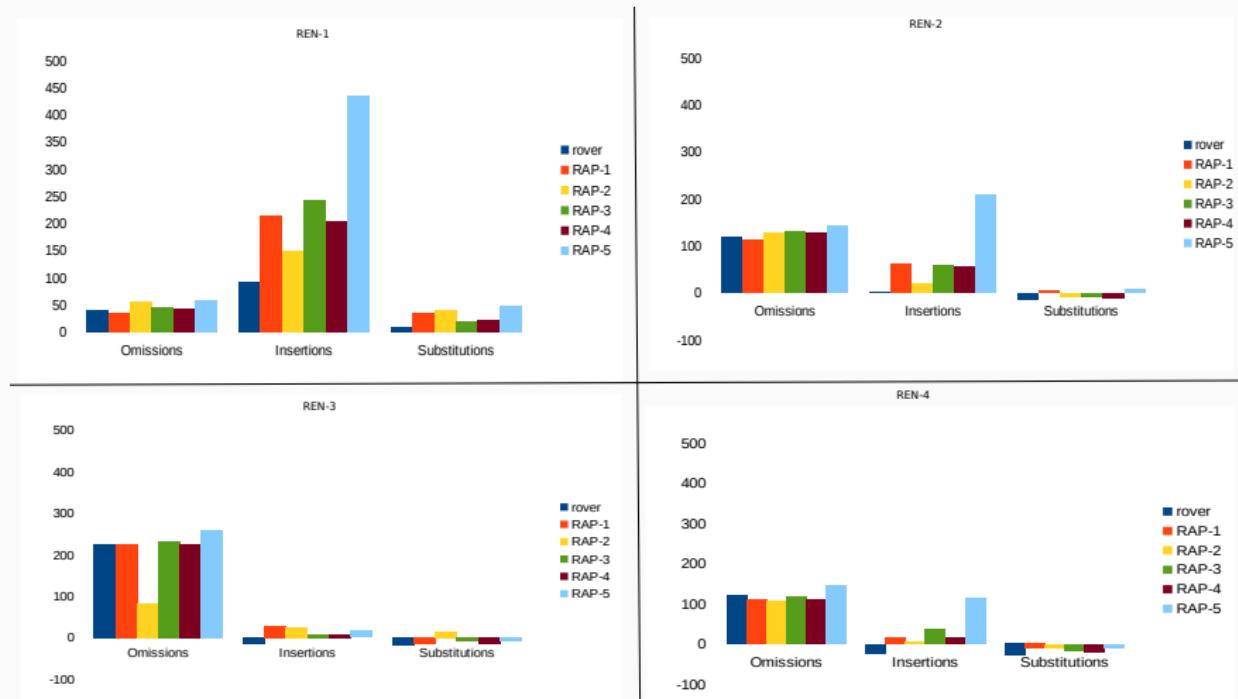
$$PAE(e) = 100 * \frac{NB_A(e) - NB_M(e)}{NB_M(e)} \quad (8)$$

Avec :

- $e$  une erreur de REN de type omission, insertion ou substitutions;
- $NB_A$  le nombre des erreurs de type  $e$  sur les transcriptions automatiques;
- $NB_M$  le nombre des erreurs de type  $e$  sur les transcriptions manuelles.

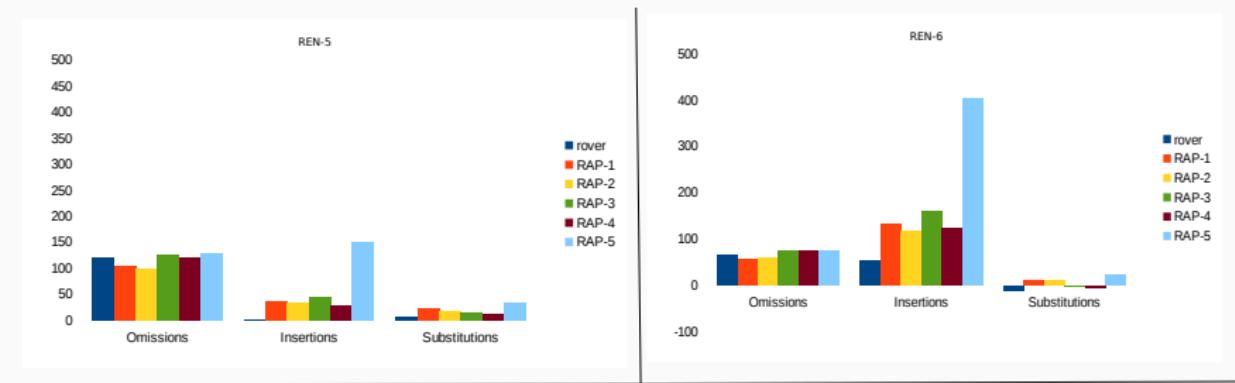
# REN sur transcriptions automatiques et manuelles

## PAE ETAPE-1



# REN sur transcriptions automatiques et manuelles

## PAE ETAPE-2



## Quelques constats

1. sur ETAPE plus un système ASR insère de mots, plus un système REN insère d'entités (pas observées sur les données QUAERO)
2. impact fort des erreurs (notamment omissions et insertions) sur la mention de l'entité
3. impact non nul des erreurs des mots qui introduisent une EN

Point 2 et 3 en lien direct avec la façon dont les systèmes REN sont développés.

# Contacts

Maud Ehrmann  
EPFL-DHLAB  
[maud.ehrmann@epfl.ch](mailto:maud.ehrmann@epfl.ch)

Sophie Rosset  
LIMSI  
[sophie.rosset@limsi.fr](mailto:sophie.rosset@limsi.fr)



-  Mohamed Ameur Ben Jannet, *Évaluation adaptative des systèmes de transcription en contextes applicatifs*, Ph.D. thesis, Université Paris Sud, octobre 2015.
-  UniProt Consortium et al., *The universal protein resource (uniprot) in 2010*, Nucleic acids research **38** (2010), no. suppl 1, D142–D148.
-  Ester2, *Entités nommées, dates, heures et montants*, 2007.
-  Nathalie Friburger, *Reconnaissance automatique des noms propres: application à la classification automatique de textes journalistiques*, Ph.D. thesis, Tours, 2002.

## References ii

-  Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard, *Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview*, Proceedings of the Fifth Linguistic Annotation Workshop (LAW-V) (Portland, OR), Association for Computational Linguistics, June 2011, pp. 92–100.
-  Heng Ji, Ralph Grishman, Hoa Trang Dang, and Joe Griffitt, Kira and Ellis, *Overview of the tac 2010 knowledge base population track*, Third Text Analysis Conference (TAC 2010), vol. 3, 2010, pp. 3–3.
-  J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, *Genia corpus: a semantically annotated corpus for bio-textmining*, no. suppl 1, i180–i182.

## References iii

-  David McDonald, *Internal and external evidence in the identification and semantic categorization of proper names*, Corpus processing for lexical acquisition (1996), 21–39.
-  John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel, *Performance measures for information extraction*, In Proceedings of DARPA Broadcast News Workshop, 1999, pp. 249–252.
-  Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum, *Entités nommées structurées : guide d'annotation quaero*, LIMSI-CNRS, notes et documents limsi n° : 2011-04 ed., 2011.
-  Mihai Surdeanu and Heng Ji, *Overview of the english slot filling track at the tac2014 knowledge base population evaluation*, Proc. Text Analysis Conference (TAC2014), 2014.

-  Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela, *HAREM: An Advanced NER Evaluation Contest for Portuguese*, Irec (Genoa), May 2006, pp. 1640–1643.
-  Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata, *Extended named entity hierarchy*, LREC, 2002.
-  Beth M. Sundheim, *Overview of the third message understanding evaluation and conference*, THIRD MESSAGE UNDERSTANDING CONFERENCE (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991, 1991.

-  Erik F. Tjong Kim Sang and Fien De Meulder, *Introduction to the conll-2003 shared task: Language-independent named entity recognition*, Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (Stroudsburg, PA, USA), CONLL '03, Association for Computational Linguistics, 2003, pp. 142–147.