



# Regards croisés sur la linguistique informatique

Denis Maurel et Karën Fort

karen.fort@sorbonne-universite.fr, denis.maurel@univ-tours.fr

13 septembre 2019





# Regards croisés sur la linguistique informatique

Karën Fort et Denis Maurel

karen.fort@sorbonne-universite.fr, denis.maurel@univ-tours.fr

13 septembre 2019



D'où parlons-nous ?

Linguistique (et) informatique

Linguistique et TAL, en 2019

Au cœur de la linguistique informatique : la catégorisation

Conclusion

# Karën Fort : ressources langagières pour le TAL

entre la linguistique et l'apprentissage (machine)

- ▶ Thèse (2012) sur l'annotation manuelle pour le TAL
- ▶ Depuis 2014 :
  - ▶ production participative (crowdsourcing) de ressources langagières pour le TAL
  - ▶ éthique et TAL
- ▶ Avant :
  - ▶ industrie du TAL (Xerox, TEMIS)
  - ▶ traductrice : multilinguisme (pas abordé ici)

# Denis Maurel : linguistique et utilisation de règles

entre la linguistique et les algorithmes à nombre fini d'états

- ▶ Thèse (1989) sur les adverbes de date sous la direction de Maurice Gross
- ▶ Depuis :
  - ▶ Implantation d'un système de cascades de graphes sous Unitex (CasSys)
  - ▶ Études des entités nommées (CasEN)
  - ▶ Constitution de ressources dictionnairiques autour des noms propres (Prolexbase)
  - ▶ Recherche d'informations dans les textes scientifiques par analyse Prédicat-Arguments (Abliss)

D'où parlons-nous ?

Linguistique (et) informatique

Une peu d'histoires

Des influences réciproques

Le TAL en France

Linguistique et TAL, en 2019

Au cœur de la linguistique informatique : la catégorisation

Conclusion

D'où parlons-nous ?

**Linguistique (et) informatique**

Une peu d'histoires

Des influences réciproques

Le TAL en France

Linguistique et TAL, en 2019

Au cœur de la linguistique informatique : la catégorisation

Conclusion

## Le TAL : enfant de la guerre (froide)



Expérience Georgetown-IBM (janvier 1954)

Dan - Flickr : IBM 701 / CC BY-SA 2.0



## Dates (et personnes) clés en linguistique informatique

- ▶ Zellig S. Harris : *Mathematical Structures of Language*, 1968
- ▶ Noam Chomsky : *Syntactic Structures*, 1957
- ▶ Igor Mel'čuk : *Dependency syntax : theory and practice*, 1988

D'où parlons-nous ?

**Linguistique (et) informatique**

Une peu d'histoires

Des influences réciproques

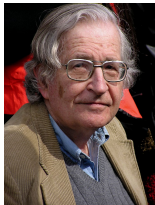
Le TAL en France

Linguistique et TAL, en 2019

Au cœur de la linguistique informatique : la catégorisation

Conclusion

# Noam Chomsky



© Duncan Rawlinson

## Des influences réciproques

- ▶ langage Algol 60
- ▶ Maurice Gross : *Méthodes en syntaxe*, 1975

D'où parlons-nous ?

## Linguistique (et) informatique

Une peu d'histoires

Des influences réciproques

**Le TAL en France**

Linguistique et TAL, en 2019

Au cœur de la linguistique informatique : la catégorisation

Conclusion

# Des dénominations qui évoluent, une communauté active

Création d'une association savante, l'Atala :

- ▶ 1959 : Association pour l'étude et le développement de la traduction automatique et de la linguistique appliquée
- ▶ 1965 : Association pour le traitement automatique des langues

Et de sa revue :

- ▶ 1960 : La Traduction Automatique
- ▶ 1965 : TA Informations, Revue internationale des applications de l'automatique au langage
- ▶ 1993 : Traitement automatique des langues

## Des dénominations qui évoluent, une discipline qui se perd ?

- ▶ Traduction automatique
- ▶ Linguistique informatique
- ▶ Traitement automatique des langues  
avec une forte influence de
  - ▶ La reconnaissance de la parole
  - ▶ La recherche d'information

→ *Hybride* quand la description linguistique s'allie à l'algorithmique

→ *Chimère* quand on oublie qu'on traite de la langue

D'où parlons-nous ?

Linguistique (et) informatique

**Linguistique et TAL, en 2019**

Les deux (r)évolutions du TAL

Qu'est-ce qu'annoter ?

Au cœur de la linguistique informatique : la catégorisation

Conclusion



D'où parlons-nous ?

Linguistique (et) informatique

**Linguistique et TAL, en 2019**

Les deux (r)évolutions du TAL

Qu'est-ce qu'annoter ?

Au cœur de la linguistique informatique : la catégorisation

Conclusion

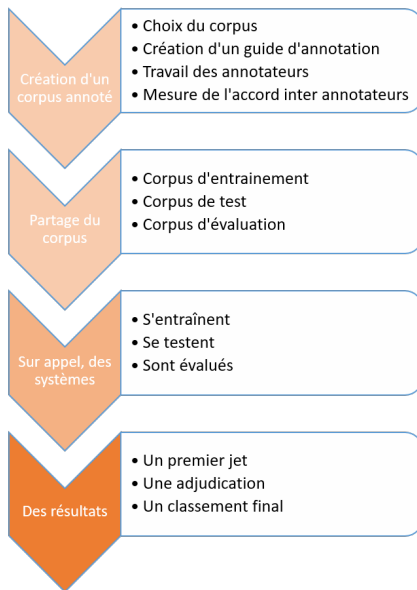
# La révolution de l'évaluation

Ré-apparition dans les années 90, après le calamiteux rapport ALPAC [Paroubek et al., 2007] :

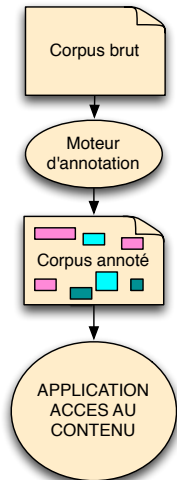
- ▶ influence de la **parole** (1987)
- ▶ projet DARPA TIPSTER (1991) : Message Understanding Conferences (MUC)
- ▶ devenu une **tradition** en TAL [Escartin et al., 2017] :
  - ▶ Conférence A\*, ACL 2016 : 9 nouvelles *shared tasks*
  - ▶ Conference on Machine Translation 2016 : 10 *shared tasks*

... très liée à l'apprentissage

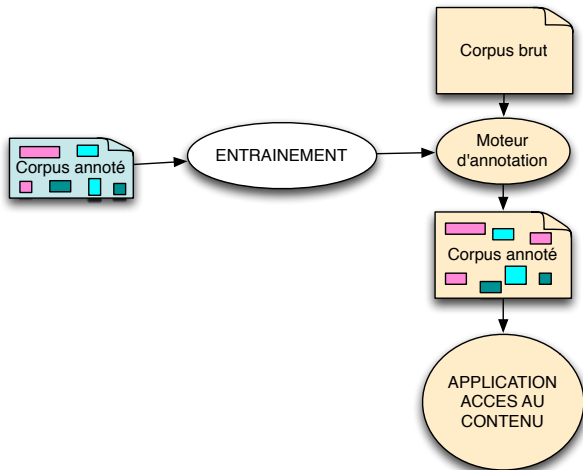
# Qu'est-ce qu'une *campagne d'évaluation* ?



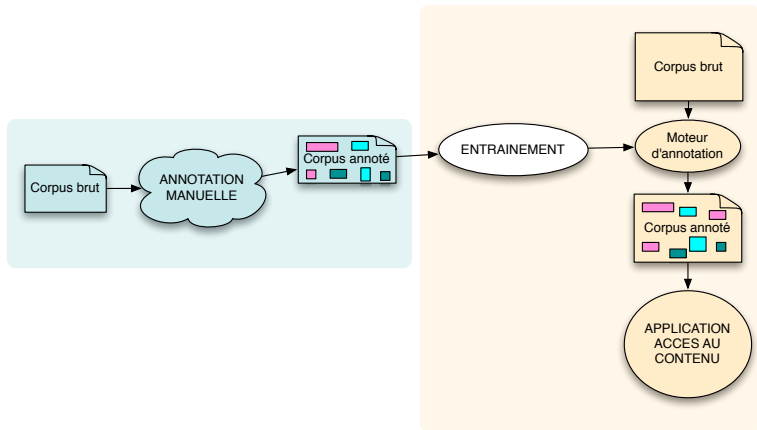
# Le TAL par l'exemple (apprentissage)



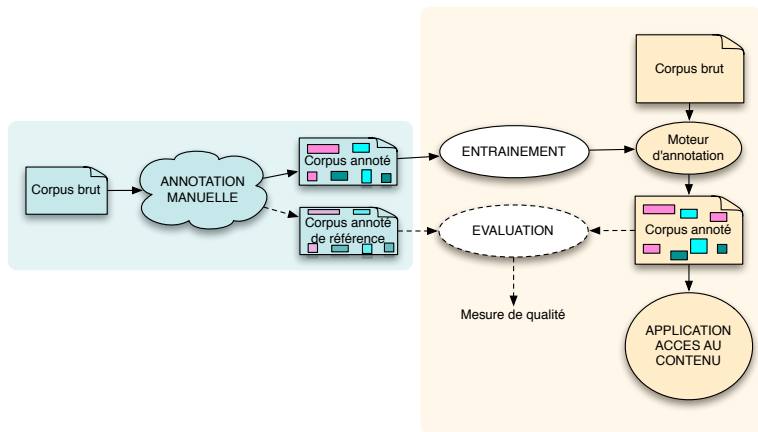
# Le TAL par l'exemple (apprentissage)



# Le TAL par l'exemple (apprentissage)



# Le TAL par l'exemple (apprentissage)



# La révolution du TAL par l'exemple (apprentissage)

F. Yvon, citant P. Norvig (<http://norvig.com/chomsky.html>)

- ▶ **Moteurs de recherche** : 100 % des principaux acteurs sont à base d'apprentissage
- ▶ **Reconnaissance de parole** : 100 % des principaux systèmes sont à base d'apprentissage
- ▶ **Traduction automatique** : 100 % des concurrents leaders dans les compétitions de type NIST sont à base d'apprentissage
- ▶ **Question-réponse** : cette application est moins développée, mais de nombreux systèmes dépendent lourdement d'approches à base d'apprentissage. Le système Watson, d'IBM, qui a gagné au Jeopardy devant des humains est largement basé sur de l'apprentissage



## La révolution du TAL par l'exemple (apprentissage)

Puisque les algorithmes sont rois, quel travail pour le linguiste computationnel ?

Un travail indispensable, même si il est transparent pour l'utilisateur final !

- ▶ La création de corpus annotés pour l'apprentissage
- ▶ La définition et le raffinement des annotations

*(Est-ce la fin des souris ? J'en souris, bien qu'elles me soient souvent utiles.)*

# En TAL, les corpus sont omniprésents, *via* l'apprentissage

*(Est-ce la fin des souris ? J'en souris, bien qu'elles me soient souvent utiles.)*

- ▶ **Désambiguïstation lexicale** : 100 % des concurrents leaders dans la compétition SemEval-2 ont utilisé des techniques statistiques.  
['souris' / animal ou 'souris' / outil ?]
- ▶ **Résolution d'anaphore** : la grande majorité des systèmes actuels sont à base d'apprentissage  
['elles' fait référence à ?]
- ▶ **Étiquetage morpho-syntaxique** : la plupart des systèmes actuels sont à base d'apprentissage  
['souris' / N ou 'souris' / V ?]
- ▶ **Analyse syntaxique** : approches multiples.  
[subj('utile') = ?]

# L'ogre a faim !

Nécessité de grandes masses de données **annotées** pour

entraîner

et

évaluer les systèmes

Exemple :

→ 100 000 mots annotés pour entraîner un tagger :

*Il/CLS est/V sain/ADJ et/CC sauf/P ./PONCT*

⇒ besoin d'**annotateurs humains**, ce qui coûte cher (600 000 \$ pour le Prague Treebank)

⇒ beaucoup de langues restent "peu dotées"

⇒ certaines entreprises continuent à utiliser des systèmes à base de règles

D'où parlons-nous ?

Linguistique (et) informatique

**Linguistique et TAL, en 2019**

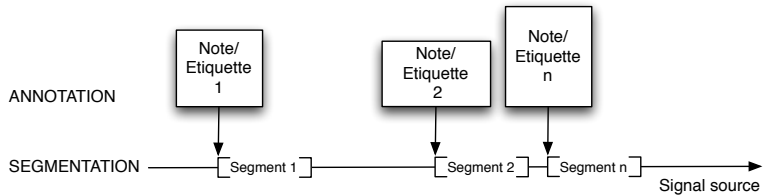
Les deux (r)évolutions du TAL

Qu'est-ce qu'annoter ?

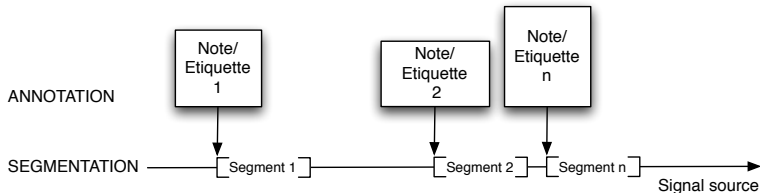
Au cœur de la linguistique informatique : la catégorisation

Conclusion

# Définition



# Définition



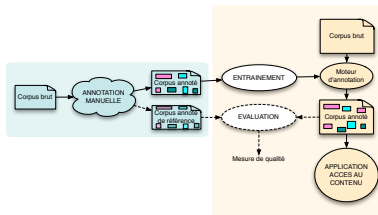
Ajout d'informations [interprétatives](#) [Leech, 1997, Habert, 2005]

# L'application : horizon de l'annotation

*Une annotation est toujours orientée par une tâche [Habert, 2000].*

- ▶ visée applicative directe (résumés de matchs pour la campagne football)
- ▶ application intermédiaire ou interne au TAL (étiquetage morpho-syntaxique)

*Une annotation est d'autant plus utile qu'elle a été conçue en fonction d'une application spécifique [Leech, 2005].*



D'où parlons-nous ?

Linguistique (et) informatique

Linguistique et TAL, en 2019

**Au cœur de la linguistique informatique : la catégorisation**

La catégorisation, cette (presque) inconnue

L'exemple des noms propres et des entités nommées

Conclusion



D'où parlons-nous ?

Linguistique (et) informatique

Linguistique et TAL, en 2019

**Au cœur de la linguistique informatique : la catégorisation**

La catégorisation, cette (presque) inconnue

L'exemple des noms propres et des entités nommées

Conclusion

# Le consensus, au cœur de l'annotation

Il faut «convenir pour mesurer »[Desrosières, 2008]

L'annotation est de l'ordre de la **quantification**

Mesurer vs quantifier [Desrosières, 2008] :

- ▶ **mesurer** : implique une forme mesurable (par ex. la hauteur du Mont Blanc)
- ▶ **quantifier** : suppose des conventions d'équivalences préalables

Outiller le consensus :

- ▶ guide d'annotation (12 p. pour le football)
- ▶ réunions avec les annotateurs et le gestionnaire de la campagne
- ▶ **évaluer** le consensus (la cohérence)

## Ce n'est pas nouveau : l'acceptabilité *est une* annotation

- (a') *Certains libraires vendent ces livres*
- (b') *Ces livres, certains libraires les vendent*
- (a'') *?Des libraires vendent ces livres*
- (b'') *\*Ces livres, certains libraires vendent*

[Guentchéva and Desclés, 1991]

Annotation insérée en début de phrase, 3 catégories possibles :

- ▶ acceptable (aucune note)
- ▶ non acceptable : \*
- ▶ incertain : ?

## Ce n'est pas nouveau : l'acceptabilité *est une* annotation

- (a') *Certains libraires vendent ces livres*
- (b') *Ces livres, certains libraires les vendent*
- (a'') *?Des libraires vendent ces livres*
- (b'') *\*Ces livres, certains libraires vendent*

[Guentchéva and Desclés, 1991]

Obtention d'un consensus d'acceptabilité [Habert, 2008] :

- ▶ jugement éduqué, informé, soumis à un apprentissage
- ▶ suppose un travail collectif

# Jugement d'acceptabilité vs annotation

cependant...

En France [Habert, 2008] :

- ▶ pas de guide d'acceptabilité : pas de "trace" globale des acceptabilités sur tel ou tel phénomène (sauf LADL/M. Gross)
- ▶ pas de travail en largeur ou systématique (sauf LADL/M. Gross)
- ▶ travail sur des énoncés simplifiés [Milner, 1989]
  
- ▶ l'annotation traite d'une très (plus?) large variété de phénomènes

D'où parlons-nous ?

Linguistique (et) informatique

Linguistique et TAL, en 2019

**Au cœur de la linguistique informatique : la catégorisation**

La catégorisation, cette (presque) inconnue

L'exemple des noms propres et des entités nommées

Conclusion

# Entités nommées

La première *définition* ? est due à Nancy Chinchor qui utilise des listes d'évidence :

- ▶ *Enamex* : personnes, lieux et organisations
- ▶ *Timex* : dates et heures
- ▶ *Numex* : pourcentages et valeurs monétaires

D'après Maud Ehrmann [Ehrmann, 2008] :

*Étant 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.*

# Noms propres

Entre autres : [Molino, 1982, M. Grevisse, 1986, Kleiber, 1996]...

Kerstin Jonasson [Jonasson, 1994] :

*Toute expression associée dans la mémoire à long terme à un particulier en vertu d'un lien dénominatif conventionnel stable*



## De nombreuses ontologies des noms propres existent

Entre autres : [Zabeeh, 1968, Bauer, 1985, Grass, 2000, Paik et al., 1996, Sekine et al., 2002]...

Ainsi que celle de Prolexbase [Tran and Maurel, 2006]

## Exemple

Nom propre			
Anthroponyme			Ergonyme
Individuel	Collectif		
	Groupe		
Célébrité	Dynastie	Association	Objet
Patronyme	Ethnonyme	Ensemble	Œuvre
Prénom		Entreprise	Pensée
Pseudo-anthroponyme		Institution	Produit
		Organisation	Vaisseau

Table – Prolexbase (1)

## Exemple

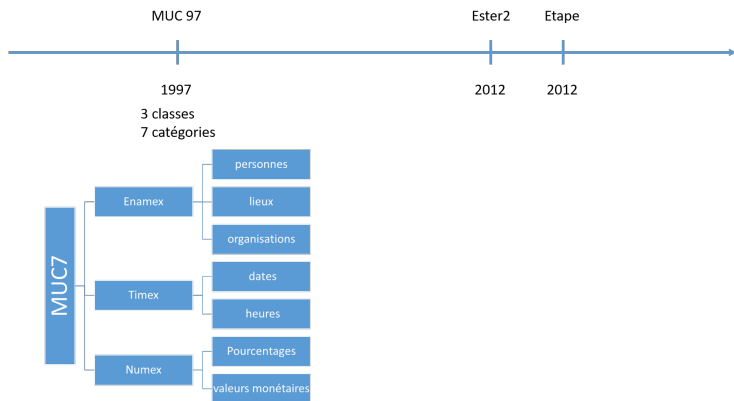
Nom propre		
Pragmonyme	Toponyme	
		Territoire
Catastrophe	Astronyme	Pays
Fête	Édifice	Région
Histoire	Géonyme	Supranational
Manifestation	Hydronyme	
Météorologie	Ville	
	Voie	

Table – Prolexbase (2)

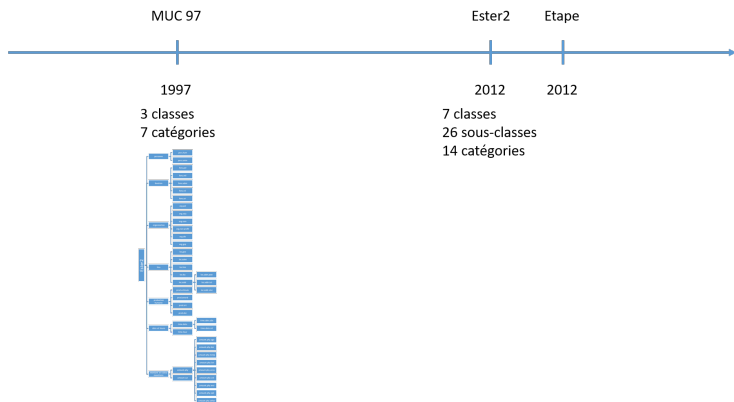
# Les campagnes d'évaluation des entités nommées



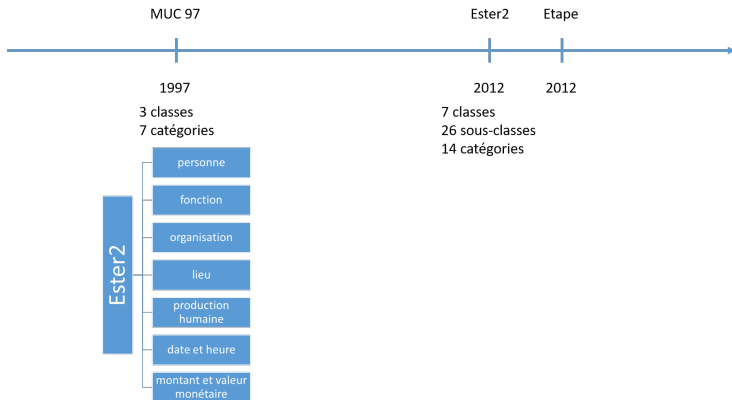
# Les campagnes d'évaluation des entités nommées



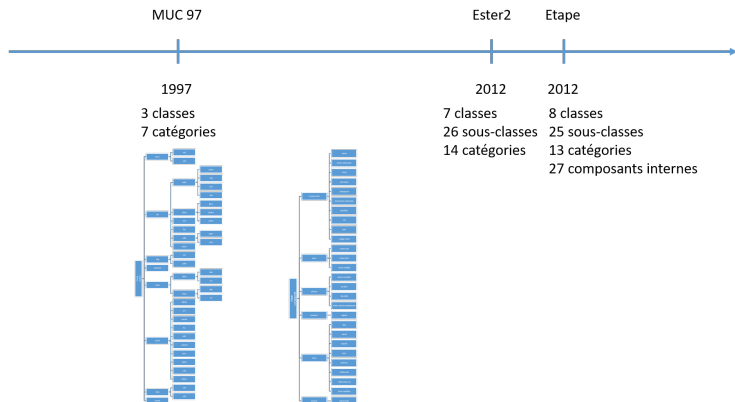
# Les campagnes d'évaluation des entités nommées



# Les campagnes d'évaluation des entités nommées

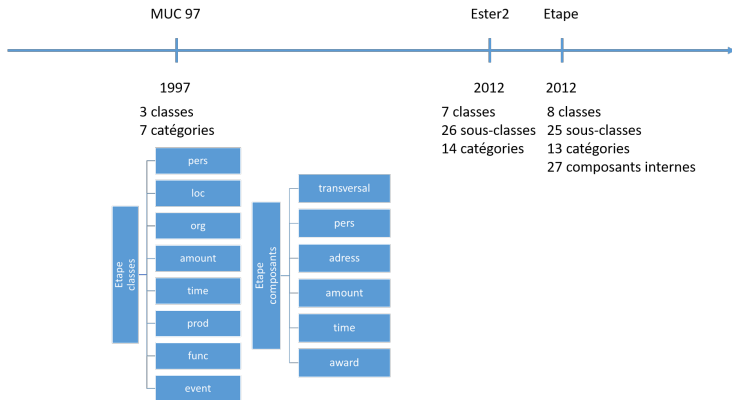


# Les campagnes d'évaluation des entités nommées





# Les campagnes d'évaluation des entités nommées



## Exemple

Exemple extrait du site du journal Le Monde, consulté le 2 juillet 2012 :

Migaud : "Il faut trouver de l'ordre de 33 milliards d'euros pour 2013". Premier président de la Cour des comptes et ancien député PS, Didier Migaud a remis...

## Exemple

Exemple extrait du site du journal Le Monde, consulté le 2 juillet 2012 :

`<enamex type="person">` Migaud `</enamex>` : "Il faut trouver de l'ordre de `<numex type="money">` 33 milliards d'euros `</numex>` pour `<timex type="date">` 2013 `</timex>`". Premier président de la Cour des comptes et ancien député PS, `<enamex type="person">` Didier Migaud `</enamex>` a remis...

## Exemple

Exemple extrait du site du journal Le Monde, consulté le 2 juillet 2012 :

`<entity type="pers.hum">Migaud</entity>` : "Il faut trouver de l'ordre de `<entity type="amount.cur">33 milliards d'euros</entity>` pour `<entity type="time.date.abs">2013</entity>`". `<entity type="fonc.admi">Premier président de la <entity type="org.pol">Cour des comptes</entity></entity>` et `<entity type="fonc.pol">ancien député PS</entity>`, `<entity type="pers.hum">Didier Migaud</entity>` a remis...

## Exemple

Exemple extrait du site du journal Le Monde, consulté le 2 juillet 2012 :

`<pers.ind><name>`Migaud`</name></pers.ind>` : "Il faut trouver `<amount><qualifier>`de l'ordre de`</qualifier> <val>`33 milliards`</val>` d'`<unit>`euros`</unit></amount>`  
`<time.date.abs><time-modifier>`pour`</time-modifier>`  
`<year>`2013`</year></time.date.abs>`".  
`<fonc.ind><name>`Premier président`</name>` de la  
`<org.adm><name>`Cour des  
comptes`</name></org.adm></fonc.ind>` et  
`<fonc.ind><qualifier>`ancien`</qualifier><name>`député`</name>`  
`<org.ent>`PS`</org.ent></fonc.ind>`,  
`<pers.ind><name.first>`Didier`</name.first>`  
`<name.last>`Migaud`</name.last></pers.ind>` a remis...

## Entités nommées : un exemple de raffinement

- ▶ L'exemple des entités nommées n'est peut-être pas des plus pertinents pour un linguiste
- ▶ On aurait pu prendre un autre exemple, l'analyse de sentiments
- ▶ Mais cet exemple montre que le raffinement des concepts est possible la porte est donc ouverte pour les linguistes en TAL...

D'où parlons-nous ?

Linguistique (et) informatique

Linguistique et TAL, en 2019

Au cœur de la linguistique informatique : la catégorisation

**Conclusion**

# Besoin de formalisationS linguistiques

Maurice Gross [Gross, 1975] :

- ▶ Eviter les observations isolées dans des régions différentes de la structure linguistique
- ▶ Systématiser les observations
- ▶ Long travail d'accumulation systématique des données
- ▶ Tester l'acceptabilité d'une séquence, c'est procéder à une expérience
- ▶ Il est donc fondamental :
  - ▶ Que l'acceptabilité et l'interprétation ne pose pas de problème lors d'une éventuelle reproduction
  - ▶ Que les exemples ne soient pas ambigus

→ Besoin de linguistes qui font de la linguistique !



# Vers un retour de la linguistique en TAL ?

Des efforts, notamment. . .






Surface syntactic Universal Dependencies [Gerdes et al., 2019] :

- ▶ garder la proposition universelle et les dépendances
- ▶ en étant plus proche de la tradition linguistique

GDR LIFT : Linguistique Informatique, Formelle et de Terrain

Merci de votre attention



-  Bauer, G. (1985).  
Namenkunde des Deutschen.  
Germanistische Lehrbuchsammlung Band 21, Berlin.
-  Desrosières, A. (2008).  
Pour une sociologie historique de la quantification :  
L'Argument statistique I.  
Presses de l'école des Mines de Paris.
-  Ehrmann, M. (2008).  
Les entités nommées, de la linguistique au TAL : statut  
théorique et méthodes de désambiguïsation.  
PhD thesis, Université Paris 7.
-  Escartin, C. P., Reijers, W., Lynn, T., Moorkens, J., Way, A.,  
and Liu, C.-H. (2017).  
Ethical considerations in nlp shared tasks.  
In First Workshop on Ethics in Natural Language Processing.
-  Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2019).

Improving Surface-syntactic Universal Dependencies (SUD) :  
surface-syntactic relations and deep syntactic features.

In

International Workshop on Treebanks and Linguistic Theories,  
Paris, France.



Grass, T. (2000).

Typologie et traductibilité des noms propres de l'allemand vers  
le français.

TAL, 41(3) :643–669.



Gross, M. (1975).

Méthodes en syntaxe.

Hermann.



Guentchéva, Z. and Desclés, J.-P. (1991).

Test et acceptabilité.

Histoire Épistémologie Langage, 13(2) :9–25.



Habert, B. (2000).

Corpus. Méthodologie et applications linguistiques, chapter  
Détournements d'annotation : armer la main et le regard,  
pages 106–120.

Champion and Presses Universitaires de Perpignan.



Habert, B. (2005).

Portrait de linguiste(s) à l'instrument.

Texto !, vol. X(4).



Habert, B. (2008).

Observer, aujourd'hui, c'est manipuler.

In François, J., editor,

Observations et manipulations en linguistique : entre concurrence et c

volume 16 of Mémoires de la Société de linguistique de Paris.

Nouvelle série, pages 33–53. Peeters, Paris, France.



Jonasson, K. (1994).

Le nom propre. Constructions et interprétations.

Duculot, Paris.



Kleiber, G. (1996).

Noms propres et noms communs : un problème de dénomination.

In Meta, volume 41-4, pages 567–589.



Leech, G. (1997).

Corpus annotation : Linguistic information from computer text corpora, chapter Introducing corpus annotation, pages 1–18.  
Longman, Londres, Angleterre.



Leech, G. (2005).

Developing Linguistic Corpora : a Guide to Good Practice, chapter Adding Linguistic Annotation, pages 17–29.  
Oxford : Oxbow Books.








M. Grevisse, A. G. (1986).

Le Bon Usage.  
Duculot, Gembloux, Belgique.



Milner, J. (1989).

Introduction à une science du langage.  
Des travaux. Editions du Seuil.

-  Molino, J. (1982).  
Le nom propre dans la langue.  
Langage, n66 Paris Larousse.
-  Paik, W., Liddy, E. D., Yu, E., and McKenna, M. (1996).  
Categorizing and Standardizing Proper Nouns for Efficient  
Information Retrieval.  
Corpus Processing for Lexical Acquisition, pages 61–76.
-  Paroubek, P., Chaudiron, S., and Hirschman, L. (2007).  
Principles of Evaluation in Natural Language Processing.  
Traitement Automatique des Langues, 48(1) :7–31.
-  Sekine, S., Sudo, K., and Nobata, C. (2002).  
Extended Named Entity Hierarchy.  
In Proceedings of the third International Conference on  
Language Ressources and Evaluation (LREC'2002), volume 5,  
pages 1818–1828.
-  Tran, M. and Maurel, D. (2006).

Prolexbase : un dictionnaire relationnel multilingue de noms propres.

TAL, 47(3) :115–139.



Zabeeh, F. (1968).

What's in a Name, An Inquiry into the Semantics and Pragmatics of Proper Names.

La Haye, Martinus Nijhoff.