

## Proposition pour une extension de la définition des entités nommées *De la définition à l'évaluation*

Sophie Rosset<sup>α</sup>  
Pierre Zweigenbaum<sup>α</sup>

Cyril Grouin<sup>α</sup>  
Karën Fort<sup>β,γ</sup>

Olivier Galibert<sup>δ</sup>  
Ludovic Quintard<sup>δ</sup>

<sup>α</sup>LIMSI-CNRS, France



<sup>β</sup>INIST-CNRS, France



<sup>γ</sup>LIPN, France



<sup>δ</sup>LNE, France



Journée ATALA  
Reconnaissance d'Entités Nommées  
Nouvelles Frontières et Nouvelles Approches  
Paris, 20 juin 2011



- 1 Définition des Entités Nommées
- 2 Entités Nommées Étendues
- 3 Annotations
- 4 Évaluation
- 5 Conclusions et Perspectives
- 6 Remerciements

## Les entités nommées

Entités Nommées : éléments d'un texte que l'on peut classer sur un plan sémantique.

## Objectif

Répondre à des questions de base : *Qui ? Quoi ? Où ? Quand ? Comment ? Pourquoi ?*

## Différents Types (cf. MUC6)

- noms de personnes (*Zina Aïtlamin Nesrin El-Hamdaoui*) ;
- noms de lieux (*Brest*) ;
- noms d'organisations (*ONU*) ;
- montants et quantités (*900 millions de dirhams*) ;
- expressions temporelles comme les dates et les durées.

## Nos références

- ESTER2 (Ester2, 2009) et son groupe de travail ;
- Thèse de Maud Ehrmann (Ehrmann, 2008) ;
- Définition des entités temporelles de TIMEX3 (Timex, 2009) ;
- Thèse de Michaël Tran (Noms Propres) (Tran, 2006) ;
- Hiérarchie proposée par S. Sekine (Sekine, 2004a; Nadeau & Sekine, 2007).

- 1 Définition des Entités Nommées
- 2 Entités Nommées Étendues**
- 3 Annotations
- 4 Évaluation
- 5 Conclusions et Perspectives
- 6 Remerciements

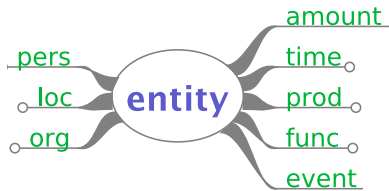
## Objectifs

- Constitution d'une base de connaissances :
    - Extraction d'information dans des données de type *nouvelles* ;
    - Détection de relations entre les informations (entités).
- ⇒ Les entités nommées étendues (ENE) constituent le pivot de cette base de connaissances.

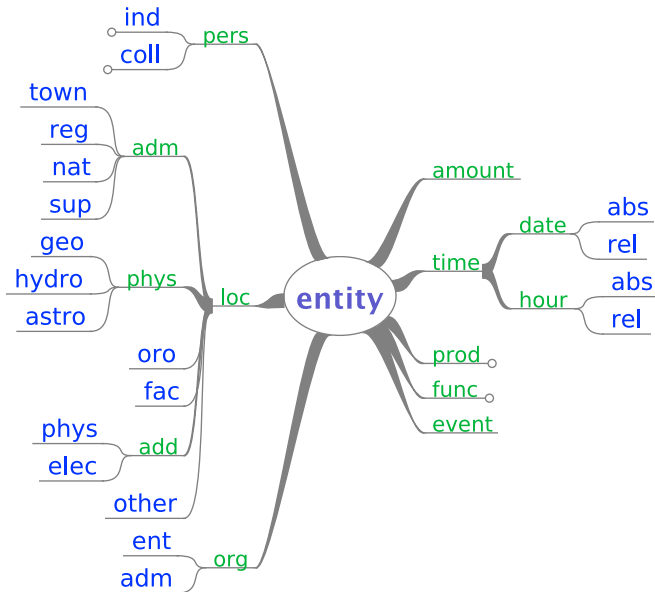
## Extension de la définition

- Extension des entités nommées à de nouveaux types (*produits* (Sekine, 2004b), *fonctions* (Galliano *et al.*, 2009)) et à des expressions sans nom propre ;
  - Structuration des entités nommées :
    - **hiérarchie** (types et sous-types) et
    - **composition** (composants transverses et spécifiques).
- ⇒ Une entité a un type et un sous-type associé et chacun de ses éléments est typé.

## Présentation générale : Entités traditionnelles et additionnelles

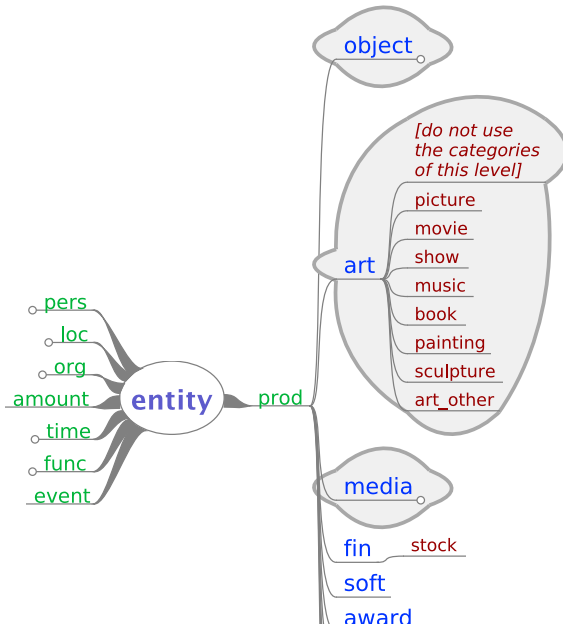


# Présentation générale : Entités traditionnelles

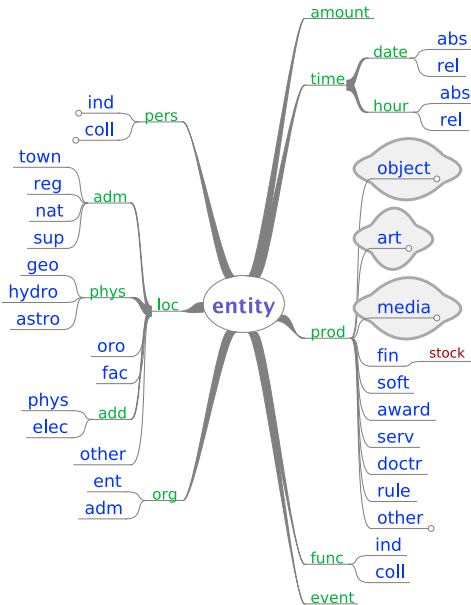




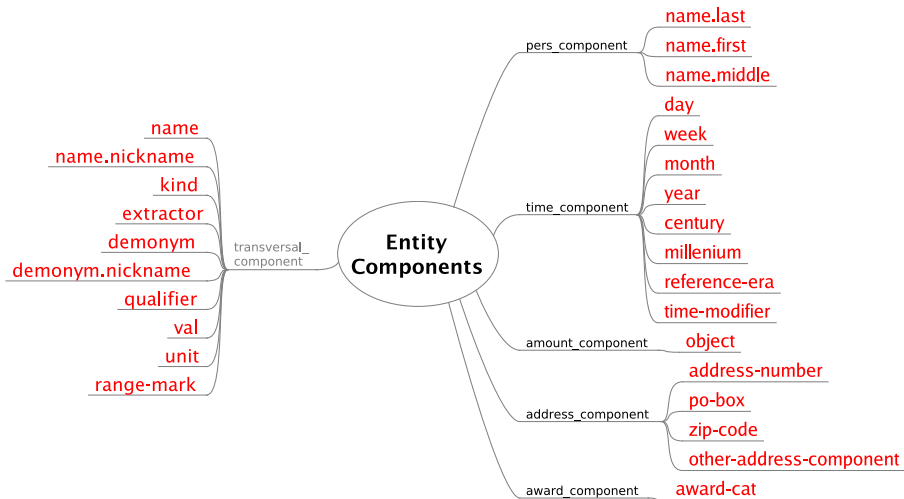
# Présentation générale : Entités additionnelles



# Présentation générale : Hiérarchie

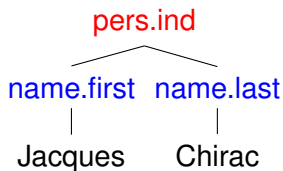


# Présentation générale : Composition (généralité)



## La composition

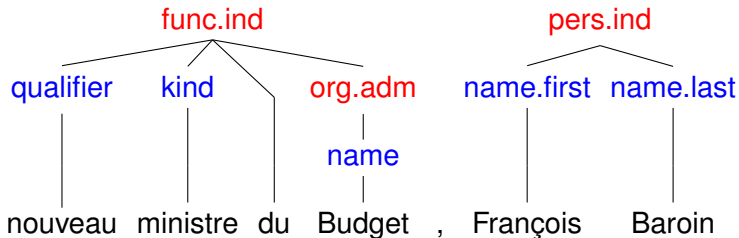
❶ Un sous-type contient un composant :



⇒ chaque **sous-type** contient au moins un **composant**.

## La composition

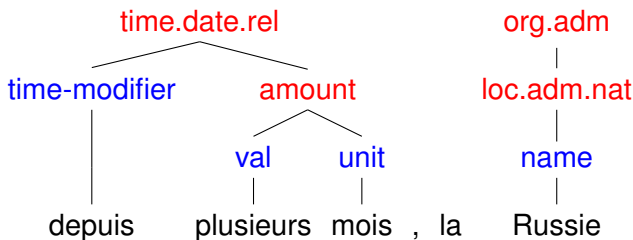
② Un type inclut un autre type : un sous-type peut-être utilisé comme un composant.



⇒ le type **fonction** (**func.ind**) intègre un type **organisation** (**org.adm**).

## La composition

③ Métonymie et antonomase : un type d'entité est utilisé pour référer à un autre type d'entité :



- 1 Définition des Entités Nommées
- 2 Entités Nommées Étendues
- 3 Annotations**
- 4 Évaluation
- 5 Conclusions et Perspectives
- 6 Remerciements

## Procédure

**Corpus** : 188 émissions pour apprentissage et 18 pour le test (broadcast news et broadcast conversations) ;

**Double annotation** par 4 annotateurs (étudiants) (10 homme-mois) ;

**Évaluation régulière** durant toute la campagne d'annotation et analyse des retours et corrections du guide (problèmes rencontrés, incohérences, erreurs résiduelles) ;

**Corpus utilisé** pour la campagne d'évaluation 2011 du projet Quaero.



## Quelques chiffres

Inf. \ Données	Entraînement	Test
# émissions	188	18
# lignes	43,289	5,637
# mots	1,291,225	108,010
# types d'entités	113,885	5,523
# types distincts	41	32
# composants	146,405	8,902
# composants distincts	29	22

## Mini référence

- Mini corpus de référence (400 lignes tirées au hasard) pour évaluer la qualité de l'annotation générale ;
- Annotation effectuée par 4 chercheurs (2 LIMSI et 2 INIST)
- Calcul d'un coefficient  $\kappa$  selon différents *marquables* (le marquage de base aléatoire étant difficile à évaluer avec ce type d'annotation complexe) (Grouin *et al.*, 2011) :
  - Annotations LIMSI/INIST :  $0.82 < \kappa < 0.91$  ;
  - Corpus global vs. mini référence :  $0.71 < \kappa < 0.85$ .

- 1 Définition des Entités Nommées
- 2 Entités Nommées Étendues
- 3 Annotations
- 4 Évaluation**
- 5 Conclusions et Perspectives
- 6 Remerciements

## Mesures d'évaluation classiques – 1

- **Rappel** (mesure de quantité) :

$$\text{Rappel} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux négatifs}}$$

- **Précision** (mesure de qualité) :

$$\text{Précision} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs}}$$

- **F-mesure** (moyenne pondérée du Rappel et de la Précision) :

$$\text{F-mesure} = \frac{(1 + \beta^2) \times \text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}}$$

## Mesures d'évaluation classiques — 2

Le **Slot Error Rate** (Makhoul, 1999) complète ces mesures et prend en compte :

- **Suppression (D)** : entités dans Ref mais pas dans Hyp ;
- **Insertion (I)** : entités dans Hyp mais pas dans Ref ;
- **Types et frontières (TF)** : entités de Hyp avec erreur de type ET de frontière ;
- **Types (T)** : entités de Hyp avec erreur de type ;
- **Frontières (F)** : entités de Hyp avec erreur de frontière ;
- **Entités attendues (R)** : entités présentes dans Ref.

$$\text{Slot Error Rate} = \frac{D + I + TF + 0,5 \times (T + F)}{R}$$

## Protocole d'évaluation

- Corpus de transcriptions de journaux et d'émissions ;
- Annotations de **référence** sur la **transcription manuelle** ;
- Annotations des **systèmes des participants** sur :
  - la **transcription manuelle**
  - **4 transcriptions automatiques** :
    - 1 sorties de 3 systèmes différents (ASR1, ASR2, ASR3) ;
    - 2 sortie enrichie du premier système (ASR1+) : ponctuation et majuscule en début de phrase.

⇒ Besoin d'aligner référence et annotations des participants.

## Alignement

- Appariement entre les entités de Hyp et de Ref ;
- Définition d'un coût d'appariement (SER) ;
- Association entre les entités de Hyp et de Ref au plus faible coût ;
- Algorithme : Viterbi (Programmation dynamique).

⇒ Cadre générique flexible où le coût dirige l'évaluation.

## Projection sur une transcription automatique — 1

- Problème : on n'a pas d'annotation des entités nommées étendues sur une sortie ASR ;
- Question : *voulons-nous annoter les mots reconnus ou les mots effectivement prononcés ?*
- *Ce qui a été dit* a davantage de sens d'un point de vue applicatif.

⇒ besoin de projeter la référence manuelle sur une sortie ASR.



## Projection sur une transcription automatique — 2

- Construire un alignement temporel de la transcription manuelle ;
- Extraire à partir de cet alignement le temps pour les annotations de référence ;
- Sélectionner une taille d'intervalle (médiane des longueurs des mots reconnus).

⇒ Nous avons l'information temporelle, nous devons retourner aux mots reconnus.

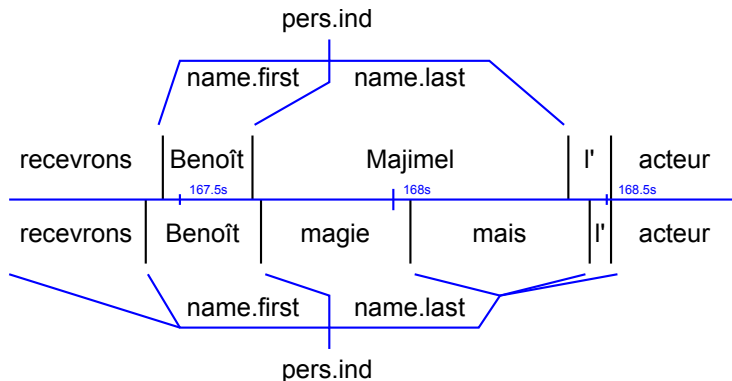
- Trouver les frontières des mots reconnus entrant dans les intervalles (début et fin).

⇒ Il y a souvent de multiples frontières.

- Construire une référence *floue* où des frontières multiples sont possibles.

## Projection sur une transcription automatique — 3

Exemple d'une référence floue obtenue par alignement temporel de la référence propre (haut) sur ASR (bas) :



## Résultats

- 3 participants (P1, P2, P3)
- Données d'entraînement = Ester1 + Ester2
- Données de test : DEV+Test évaluation Quaero ASR 2010

Perte absolue en terme de SER.

	<b>Manuel</b>	<b>ASR1</b>	<b>ASR1+</b>	<b>ASR2</b>	<b>ASR3</b>
	<b>Broadcast News</b>				
WER		16.32%	16.32%	18.77%	24.06%
P1	42.7%	-12.5%	-10.0%	-15.8%	-18.7%
P2	29.7%	-18.8%	-24.1%	-22.5%	-23.8%
P3	39.1%	-16.5%	-15.4%	-21.2%	-22.7%
	<b>Broadcast Conversations</b>				
WER		23.34%	23.34%	22.99%	29.18%
P1	55.3%	-32.6%	-34.6%	-23.0%	-33.9%
P2	37.0%	-36.9%	-42.0%	-29.6%	<b>-36.0%</b>
P3	43.0%	-46.3%	-40.3%	-38.2%	<b>-41.1%</b>

## Résultats

- Sous-types plutôt bien annotés ( $F_1$ -mesure) :

<b>Sous-type</b>	<b>P1</b>	<b>P2</b>	<b>P3</b>	<b>Référence</b>
pers.ind	0.834	0.893	0.900	0.991
loc.adm.town	0.759	0.745	0.797	0.947

- Sous-types problématiques ( $F_1$ -mesure) :

<b>Sous-type</b>	<b>P1</b>	<b>P2</b>	<b>P3</b>	<b>Référence</b>
func.ind	0.461	0.552	0.599	0.966
org.adm	0.418	0.448	0.360	0.911
org.ent	0.313	0.521	0.509	0.864
pers.coll	0.446	0.557	0.478	0.788

## Conclusions et perspectives

- Modifications de structure :
  - transformation de func.\* en composants ;
  - fusion de org.adm et org.ent ;
  - problèmes avec org.ent et pers.coll : que faire ?
- Nouveauté
  - annotation en événements ;
    - cf. Présentation de B. Arnulphy ;
  - annotation des relations entre entités (plus tard).

- 1 Définition des Entités Nommées
- 2 Entités Nommées Étendues
- 3 Annotations
- 4 Évaluation
- 5 Conclusions et Perspectives
- 6 Remerciements**

## Remerciements

Nous remercions les annotateurs pour leur travail ainsi que Sabine Barreaux (INIST–CNRS) pour son implication sur le corpus de référence.

## Financements

Ce travail a été réalisé dans le cadre du programme Quaero, financé par Oseo, agence française pour l'innovation et la recherche, et par le projet ANR Etape.

# Références I

- 📄 EHRMANN M. (2008).  
*Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation.*  
PhD thesis, Université Paris 7 - Denis Diderot.
- 📄 ESTER2.  
*Annotation Entités Nommées, dates, heures et montants.*
- 📄 GALLIANO S., GRAVIER G. & CHAUBARD L. (2009).  
The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts.  
*In Proc of Interspeech 2009.*
- 📄 GROUIN C., ROSSET S., ZWEIGENBAUM P., FORT K., GALIBERT O. & QUINTARD L. (2011).  
Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview.  
*In Proc. of the Fifth Linguistic Annotation Workshop (LAW-V), Portland, OR: Association for Computational Linguistics.*
- 📄 NADEAU D. & SEKINE S. (2007).  
A survey of named entity recognition and classification.  
*Linguisticæ Investigationes*, **30(1)**.



## Références II

- 📄 [SEKINE S. \(2004a\).](#)  
Definition, dictionaries and tagger of extended named entity hierarchy.  
In *LREC'04*, Lisbon, Portugal.
- 📄 [SEKINE S. \(2004b\).](#)  
Definition, dictionaries and tagger of extended named entity hierarchy.  
In *LREC'04*, Lisbon, Portugal.
- 📄 [TimeML Working Group.](#)  
*Guidelines for Temporal Expression Annotation for English for TempEval 2010.*
- 📄 [TRAN M. \(2006\).](#)  
*Prolexbase. Un dictionnaire relationnel multilingue de noms propres : conception implantation et gestion en ligne.*  
PhD thesis, Université Francois Rabelais de Tours.