

# Les entités nommées

## Module TAL - Master HN PSL

Carmen Brando

[carmen.brando@ehess.fr](mailto:carmen.brando@ehess.fr)

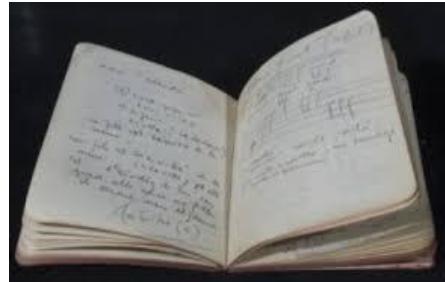
Ingénieure de recherche en humanités numériques

Ecole des hautes études en sciences sociales

Collaboration avec le **Laboratoire LATTICE**

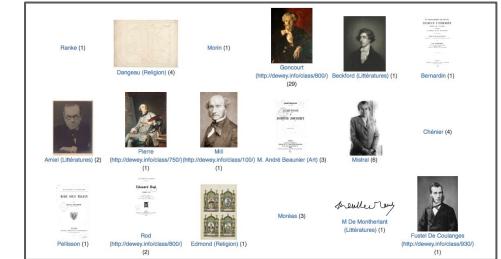
Paris, 22 février 2021

# Des archives et sources en passant par des textes numériques annotés jusqu'aux données à explorer sous l'angle de noms de personnes, de lieux, ...



```
@Begin
@Languages: fra
@Participants: AV16 Target_Adult, CH_EE1 Investigator, CH_OD1 Investigator,
AV16FIE Child
@ID: fra[Entretien|AV16||||Target_Adult||]
@ID: fra[Entretien|CH_EE1||||Investigator||]
@ID: fra[Entretien|CH_OD1||||Investigator||]
@ID: fra[Entretien|AV16FIE||||Child||]
@Medium: AV16
*AV16 je m'appelle AV16 <chein> et je suis né aux <lieu id=3114710> Asturias </lieu>
d'une famille nombreuse par la suite mais au moment où je suis né
nous n'étions que deux mon frère ainé et moi et par la suite y a
une petite sœur qui est arrivée et nous étions déjà trois aux
<lieu id=3114710> Asturias </lieu> mais le mouvement de ma famille a été très bouleversé
parce que :
mon père qui était fils <conv> de propriétaire terrien très
important dans la <lieu id=3336902> Galice </lieu> a quitté la terre qu'il n'aimait pas pour
aller dans les <lieu id=3114710> Asturias </lieu> dans les mines pour voir comment on
<revision> je ce ce </revision> gagnait <repetition> la
</repetition> la vie les mineurs de fond et cela ne lui convenant
pas il a commencé *
à faire <repetition> des des </repetition> des études parce qu'il en
avait déjà fait par l'église qui le protégeait là-bas où il est né
```

< Quant au rythme, si Victor Hugo a dépassé Lamartine, il n'a pas été plus loin que Vigny.  
Après lui il a pratiqué la césure mobile et l'enjambement... il n'a pas inventé de mètres nouveaux. Il s'est borné à faire consciemment ce que Lamartine avait fait par négligence, et Vigny par souci d'harmoniser la forme avec la pensée qu'elle traduisait. Que le sens du rythme soit infiniment plus puissant chez Victor Hugo que chez Vigny, cela peut-il seulement être discuté ? Ce n'est pas après Vigny (dont le vers est assez classique) que Victor Hugo a pratiqué l'enjambement, c'est après Chénier qui avait déjà influé sur Vigny. Victor Hugo, il est vrai, n'a pas inventé de mètres nouveaux, mais d'une part le symbolisme lui-même a montré par ses essais que le champ ouvert à l'invention métrique est fort limité, et d'autre part, Victor Hugo a dépassé de loin Ronsard dans l'invention de combinaisons métriques nouvelles, de strophes ou plutôt d'associations de strophes selon le mouvement oratoire ou poétique (ce qui est en somme de l'invention métrique). La négligence de Lamartine est une demi-légende, créée par lui-même ; elle ne s'applique qu'à ses vers faibles et à sa prose ; ses belles pièces, dont nous avons quelques-unes les brouillons, travaillées longuement, sont au contraire de magnifiques victoires sur sa facilité.



# Les lieux parisiens d'Apollinaire dans « Le passant de Prague »

*“Voilà ! J'avais eu affaire, **rue de la Pépinière**, près de la **place Saint-Augustin**, et je revenais par le **boulevard Malesherbes** en l'intention de prendre l'omnibus à la **Madeleine**. Tout à coup, au coin de la **rue des Mathurins**, un homme se dressa devant moi en criant : “Madame ou mademoiselle, [...]. ””*

## Visualization

The following map shows distribution of places mentioned in the input TEI-XML file, geo-coordinates are obtained via [French DBpedia](#).

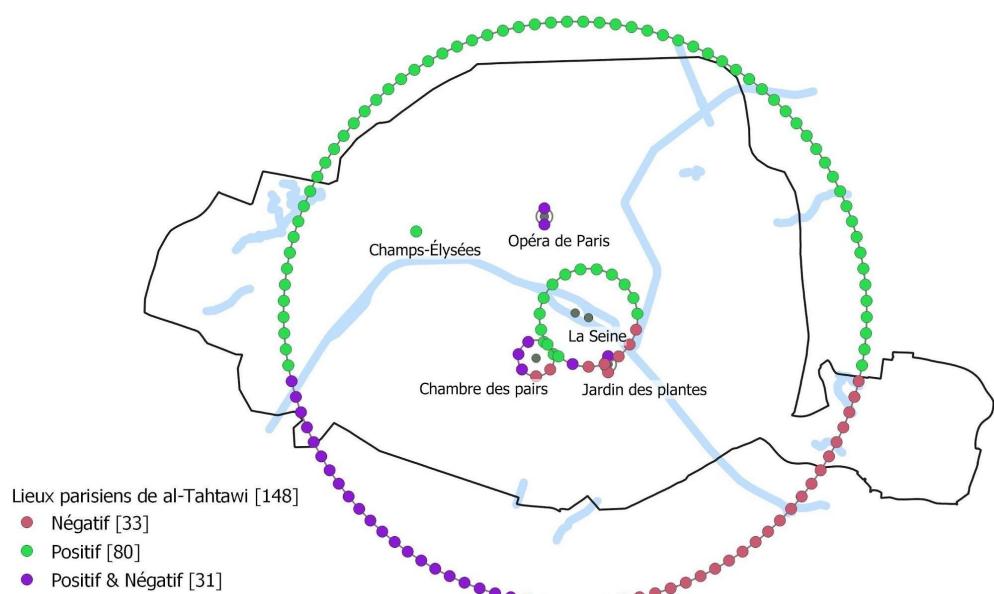


137 places are displayed on the map.

17 places were not included on the map because geo-coordinates were unavailable, these are: Berlin, Rhin, Bohème, montagnes Rocheuses, Queensland, royaume de Juda, La Nouvelle-Orléans, Provence, Neckar, empire des Habsbourg, Danube, Moldau, Ile-de-France, Hambourg, Bavière, Savoie, Amsterdam

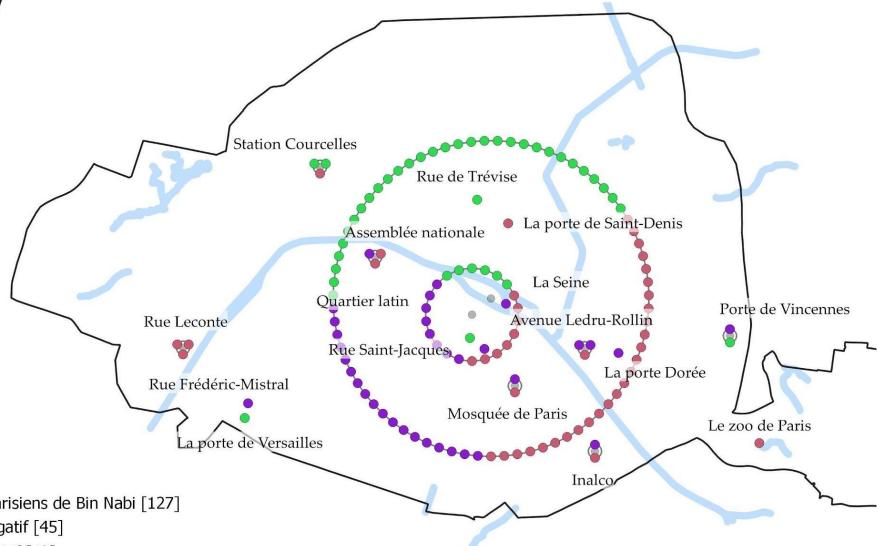
You can download the resulting annotated XML-TEI file [here](#)

Alrahabi et al 2021, "Paris dans les récits de voyage d'écrivains arabes : repérage, analyse sémantique et cartographie de toponymes"



Lieux parisiens de al-Tahtawi [148]

- Négatif [33]
- Positif [80]
- Positif & Négatif [31]



Lieux parisiens de Bin Nabi [127]

- Négatif [45]
- Positif [43]
- Positif & Négatif [36]

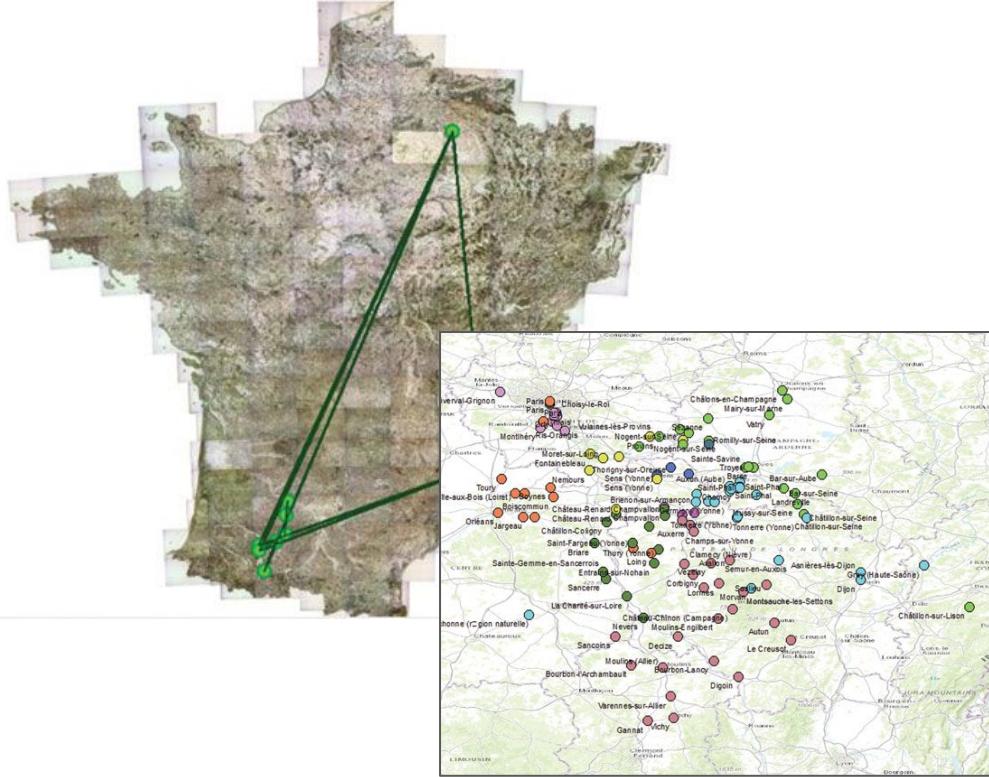
# La dimension spatiale de « La grande peur de 1789 » de Georges Lefebvre

## Extrait sur la Champagne et le Sud-Ouest :

*“Sa puissance émotive, qui fut grande, demeura intacte jusqu'à la fin. Elle partit, le 28, de Ruffec, dans les circonstances qu'on connaît. Vers l'Ouest, elle gagna les forêts de Chizé et d'Aulnay, semble-t-il, à moins que celles-ci n'aient constitué un centre d'émotion locale.”*

(Paris et al 2017)

<https://dx.doi.org/10.1080/15420353.2017.1307306>



# La TEI intègre l'annotation d'entités nommées dans ses consignes de balisage

.... c'est ainsi que les curieuses expériences de  
**<persName ref="http://www.idref.fr/035763655">H. de Vries</persName>**,  
par exemple, en montrant que des variations importantes peuvent se produire brusquement et se transmettre régulièrement ....

... prédire par exemple l'état de la faune de la  
**<placeName**  
**ref="http://fr.dbpedia.org/page/Royaume-Uni">Grande-Bretagne</placeNa**  
**me>** en 1868 ...

[Bergson - L'évolution créatrice]

Télécharger : tei, epub, kindle, texte brut, iramuteq, html.

## Réflexions sur la littérature

- [1. Une thèse sur le symbolisme](#)
- [2. Un livre sur Ronsard](#)
- [3. Le cinquantenaire de Vigny](#)
- [4. La pitié des églises par Barrès](#)
- [5. Anthologie des avocats par Payen](#)
- [6. Cristallisations](#)
- [7. Le masque de Shakespeare](#)
- [8. Les spectacles dans 1 fauteuil](#)
- [9. Le centenaire d'H. Spencer](#)
- [10. Discussion sur le moderne](#)
- [11. Les Goncourt](#)
- [12. L'idée de génération](#)
- [13. Les chapelles littéraires](#)

« Quant au rythme, si Victor Hugo a dépassé Lamartine, il n'a pas été plus loin que Vigny.

Après lui il a pratiqué la césure mobile et l'enjambement... il n'a pas inventé de mètres nouveaux. Il s'est borné à faire consciemment ce que Lamartine avait fait par négligence, et Vigny par souci d'harmoniser la forme avec la pensée qu'elle traduisait. Que le sens du rythme soit infiniment plus puissant chez Victor Hugo que chez Vigny, cela peut-il seulement être discuté ? Ce n'est pas après Vigny (dont le vers est assez classique) que Victor Hugo a pratiqué l'enjambement, c'est après Chénier qui avait déjà influé sur Vigny. Victor Hugo, il est vrai, n'a pas inventé de mètres nouveaux, mais d'une part le symbolisme lui-même a montré par ses essais que le champ ouvert à l'invention métrique est fort limité, et d'autre part, Victor Hugo a dépassé de loin Ronsard dans l'invention de combinaisons métriques nouvelles, de strophes ou plutôt d'associations de strophes selon le mouvement oratoire ou poétique (ce qui est en somme de l'invention métrique). La négligence de Lamartine est une demi-légende, créée par lui-même ; elle ne s'applique qu'à ses vers faibles et à sa prose ; ses belles pièces, dont nous avons quelquefois les brouillons, travaillées longuement, sont au contraire de magnifiques victoires sur sa facilité.

# Le TAL pour automatiser le balisage d'EN dans des éditions numériques TEI

## Savoirs EHESS



### LE CORPUS

La bibliothèque Savoirs contient des textes relevant de l'histoire et de l'anthropologie des sciences et des savoirs. Ils sont choisis dans une perspective interdisciplinaire et comparatiste, toutes périodes et aires culturelles confondues.

[EN SAVOIR +](#)

### LE THÉSAURUS

Le théâtre Savoirs est une cartographie conceptuelle du champ de l'histoire et de l'anthropologie des savoirs. Il a été construit dans une perspective interdisciplinaire large, qui traverse les domaines des



## Testaments de Poilus Édition numérique

1914-1918

RECHERCHER Les testaments Les testateurs Les lieux Les unités militaires EXPLORER MON ESPACE

Etat de la recherche

Exploitation du corpus

Chacun des testaments édités permet d'entendre la voix singulière de son auteur et entretient ainsi le souvenir du disparu qui n'aura peut-être pas laissé d'autre trace écrite. À côté de cet usage mémoriel, apanage des familles, une exploitation historique de ces testaments est possible malgré leur caractère bref et laconique.

Lire la suite

Actualités

Quelques exemples de testaments

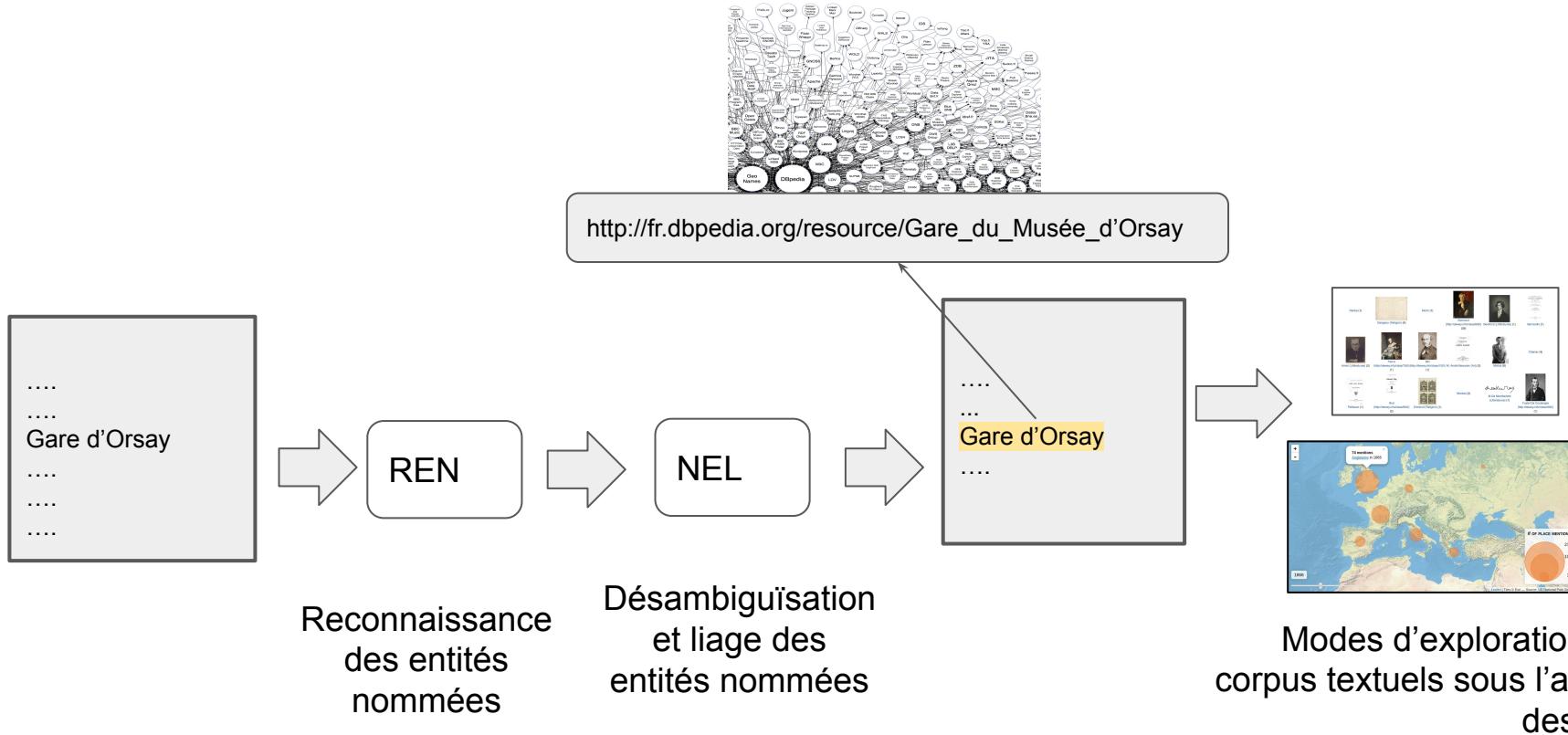
Arcade Eugène BACOUR Cote MC/ET/XXIII/1677

Louis Raymond BOVERAT Cote MC/ET/XXXIV/1641

Emile Marcel Désiré GOUPY Cote 3E13 345

Georges Marcel FLEURET Cote 3E38 818

# Chaîne TAL pour identifier les EN dans un corpus textuel en vue d'une étude outillée linguistique et spatiale



@Begin  
@Languages: fra  
@Participants: AV16 Target\_Adult, CH\_EE1 Investigator, CH\_OD1 Investigator,  
AV16FIE Child  
@ID: fra|Entretien|AV16||||Target\_Adult|||  
@ID: fra|Entretien|CH\_EE1||||Investigator|||  
@ID: fra|Entretien|CH\_OD1||||Investigator|||  
@ID: fra|Entretien|AV16FIE||||Child|||  
@Media: AV16

\*AV16 : je m'appelle AV16 <hein> et je suis né aux <lieu id=3114710> Asturias </lieu>  
d'une famille nombreuse par la suite mais au moment où je suis né  
nous n'étions que deux mon frère aîné et moi et par la suite y a  
une petite sœur qui est arrivée et nous étions déjà trois aux  
<lieu id=3114710> Asturias </lieu> mais le mouvement de ma famille a été très bouleversé  
parce que \*  
mon père qui était fils <conv> de propriétaire terrien très  
important dans la <lieu id=3336902> Galice </lieu> a quitté la terre qu'il n'aimait pas pour  
aller dans les <lieu id=3114710> Asturias </lieu> dans les moins  
<revision> je ce ce </revision> gagnait <repetition> la  
</repetition> la vie les mineurs de fond et cela ne lui convenait  
pas il a commencé \*  
à faire <repetition> des des </repetition> des études parce qu'il  
avait déjà fait par l'église qui le protégeait là-bas où il est né

## Récits de vie (transcription d'entretien oral)

## Critique littéraire

« Quant au rythme, si Victor Hugo a dépassé Lamartine, il n'a pas  
été plus loin que Vigny.

Après lui il a pratiqué la césure mobile et l'enjambement... il n'a  
pas inventé de mètres nouveaux. Il s'est borné à faire consciemment  
ce que Lamartine avait fait par négligence, et Vigny par souci  
d'harmoniser la forme avec la pensée qu'il traduisait. Qu le sens du  
rythme soit infiniment plus puissant chez Victor Hugo que chez  
Vigny, cela peut-il seulement être discuté ? Ce n'est pas après Vigny  
(dont le vers est assez classique) que Victor Hugo a pratiqué  
l'enjambement, c'est après Chénier qui avait déjà influé sur Vigny.  
Victor Hugo, il est vrai, n'a pas inventé de mètres nouveaux, mais  
d'une part le symbolisme lui-même a montré par ses essais que le  
champ ouvert à l'invention métrique est fort limité, et d'autre part,  
Victor Hugo a dépassé de loin Ronsard dans l'invention de  
combinaisons métriques nouvelles, de strophes ou plutôt  
d'associations de strophes selon le mouvement oratoire ou poétique  
(ce qui est en somme de l'invention métrique). La négligence de  
Lamartine est une demi-légende, créée par lui-même ; elle ne  
s'applique qu'à ses vers faibles et à sa prose ; ses belles pièces, dont  
nous avons quelques fois brouillons, travaillées longuement, sont au  
contraire de magnifiques victoires sur sa facilité.

## AVIS

Toute personne désirant être inscrite dans le présent Annuaire devra justifier DE SON TITRE DE PROPRIÉTAIRE.

Les inscriptions et les changements de domicile seront inscrites gratuitement, mais devront être adressées à la Direction avant le 1<sup>er</sup> NOVEMBRE.

## LISTE ALPHABÉTIQUE DES PROPRIÉTAIRES DE PARIS

A
M.M. Aurore, Paris, r. Faub. 5 (XIV). <small>ancien</small> Abadie, Paris, r. Bellière, 64 (XV). <small>ancien</small> Abadie, Paris, r. Boissière, 10 (XV). <small>ancien</small> Abadie, grv. Amst., quart Cent., 5. <small>ancien</small> Abadie-Neaudig, Paris, r. Clot, 27. <small>ancien</small> Abadie, Paris, loc. Bourg-Morlaix, 35 (IX). <small>ancien</small>
Abadie (M.), Paris, r. Vaugirard, 28. <small>ancien</small>
Abadie, Fontenay (M.), r. Gare d'Orléans, 8. <small>ancien</small>
Abadie (M.), Paris, r. Washington, 32 (XVII). <small>ancien</small>
Abadie (M.), Paris, r. Championnet, 112 <sup>e</sup> (XVII). <small>ancien</small>
Abadie (M.), Paris, r. Faub. 142 (XVII). <small>ancien</small>
Abadie, Paris, r. St-Honoré, 27 (XVII). <small>ancien</small>
Abadie (M.), Paris, av. Brune, 92 (XVII). <small>ancien</small>
Abel, Isidore, baron de St-Germain (P-de-l.). C. 1800 (M), Paris, r. Luxembourg, 16. <small>ancien</small>
Abelard, Paris, av. Brune, 46 (XVII). <small>ancien</small>
Abundance, Paris, pl. Constitue-Pepper, 4 (XVII). <small>ancien</small>
Abou-Saleh, Paris, r. Abel, 7 (XVII). <small>ancien</small>
Abot, St-Sauveur (Gout, was). Aboucaya, Paris, r. Mornay, 40. <small>ancien</small>
Abraham, Paris, r. Honfleur, 38. <small>ancien</small>
Abrial (C.), Paris, r. Louvois, 39 (XVII). <small>ancien</small>
Abreu, Paris, imp. Grasset, 5 (XX). <small>ancien</small>
Abrenwick, Paris, r. Maré-Chr., 4. <small>ancien</small>

## L'ASSURANCE GÉNÉRALE DES EAUX ET ACCIDENTS

PARIS, 102, Rue de Richelieu — Tel. 258-10

Responsabilité civile des assureurs des assurances sur incendie, vol et cambriolage.  
Assurance contre les dégâts des eaux, incendie, électricité, etc.

# Comment la machine peut accéder au « sens » d'un texte ?

- ❖ Par des différentes couches d'analyse:
  - Lexique
  - Syntaxe
  - Prosodie
  - **Sémantique**

# Entité nommée (EN)

- ❖ “Il s’agit de types d’unités lexicales particulières qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques, et qui est désignée par son nom” (convention Ester)
  - Ex : *Barack Obama, Hugo, rue de Rivoli*
- ❖ associée aux **noms propres** et parfois à des **descriptions définies** (un groupe nominal, ex : *le chat noir*)
- ❖ Les trois catégories (dites *coarse-grained*) : noms de personnes, de lieux, d’organisations
- ❖ extension de la typologie d’EN pour inclure : les fonctions de personnes, les productions humaines, les expressions numériques, temporelles, etc.

# Une définition précise du nom propre ?

- ❖ Du point de vue de la linguistique, la catégorie nom propre est difficile à définir, car il y a des nombreuses exceptions
- ❖ Les critères traditionnels :
  - sur la forme des mots : la marque de la majuscule (mais ce n'est pas une règle comme dans : *la gare de Montparnasse*)
  - d'ordre factuel : la non traduction et l'absence de dictionnaires de la langue
  - d'ordre morphosyntaxique : l'absence de déterminant et de flexion en théorie car il y a des nombreuses exceptions dans les usages : *la Seine, la Rochelle, le Paris d'après-guerre* (emploi figuré), “J'ai acheté trois *Picassos*”

# Nom propre : La marque de la majuscule ?

- ❖ n'est pas translinguistique (usage différent d'une langue à l'autre, ex : l'allemand)
- ❖ n'est pas valide en diachronie (usage inexistant dans les corpus anciens, notion qui apparaît avec l'imprimerie) : besoin de normaliser un corpus
- ❖ n'est pas appréciable à l'oral (aussi, besoin de normaliser)

# Types d'ambigüités

- ❖ le même nom est utilisé pour plusieurs entités
  - *Paris* (France) et *Paris* (Texas)
- ❖ une même entité peut avoir plusieurs noms
  - *Paris, Paname*
- ❖ le nom d'une entité peut être utilisé pour désigner une entité en catégories différentes (le cas de la métonymie)
  - la *Sorbonne*, la *France*, ... une organisation mais peut être aussi un lieu selon le contexte
  - “Le *prix Nobel de la Paix* s'est montré digne devant une telle épreuve”

# Difficultés à définir précisément une EN

- les bornes/limites de l'EN
  - la rue de Strasbourg, il peut s'agir selon le contexte d'une rue qui porte ce nom ou bien d'une rue dont le nom n'est pas précisée et qui est localisée dans la ville de Strasbourg
  - le comité exécutif de l'Union des Associations Européennes de Football, il s'agit d'un syntagme complexe, et il y a deux niveaux de granularité possibles..
- l'imbrication de l'annotation de plusieurs EN
  - le président de la France, il s'agit d'une personne, celle qui occupe la fonction, et aussi d'un pays ("France")
  - l'Église Saint-Pierre-et-Saint-Paul
- un référent parfois flou, collectif ou historique :
  - les côtes de la Guyane, le nord de l'Europe, La Bohême

# Les typologies d'EN et la notion de campagne d'évaluation en TAL

Types	Exemple	Contre-exemple
ORG	<i>DARPA</i>	our university
PERS	<i>Harry Schearer</i>	St. Michael
LOC	<i>U.S.</i>	53140 Gatchell Road
MONEY	<i>19 dollars</i>	en dollars ? ça fait 19
TIME	<i>8 heures</i>	la nuit dernière (*)
DATE	<i>le 23 juillet</i>	en juillet dernier (*)

Typologie des EN dans le cadre de MUC-7

A ce stade,  
aucune expression  
imbriquée n'est  
marquée.

# Les typologies d'EN et la notion de campagne d'évaluation

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

Typologie des EN dans le cadre d'ACE

Types	Exemples
FAC	L'aéroport <i>Charles de Gaulle</i> est grand.
GPE	<i>Andorre</i> se situe dans les montagnes.
LOC	<i>M42</i> est une nébuleuse magnifique.
ORG	Le <i>LDC</i> est un laboratoire de recherche.
PER	<i>Pierre</i> roule sur la mousse avec la voiture.
VEH	les <i>hélicoptères militaires</i> ont ... ; l' <i>USS Alabama</i> est un navire de ligne ...
WEA	des <i>missiles sol-air</i> ont été tirés... ; le <i>gaz sarin</i> ...

(Nouvel, Ehrmann, Rosset, 201)

Voir aussi :

<https://damien.nouvelets.net/resourcesen/typologies.html>

Types	Sous-types
pers	pers.hum, pers.anim
fonc	fonc.pol fonc.mil fonc.admi fonc.rel fonc.ari
org	org.pol org.edu org.com org.non-profit org.div org.gsp
loc	loc.geo loc.admi loc.line loc.addr (+3) loc.fac
prod	prod.vehicule prod.award prod.art prod.doc
time	time.date (+ 2 abs et rel) time.hour (+ 2 abs et rel)
amount	amount.phy.age amount.phy.dur amount.phy.temp amount.phy.len amount.phy.area amount.phy.vol amount.phy.wei amount.phy.spd amount.phy.other amount.cur

Typologie des EN proposée par Ester-2  
(similaire à son successeur ETAPE)

## Les typologies d'EN et les campagnes d'évaluation françaises

Quelques exemples présentés par le guide Ester-2

- (a) Le [ent=org.pol-] *Parti Communiste* [-ent=org.pol] a peu de chance d'être au second tour.
- (b) Le [ent=org.pol-] *RPR* [-ent=org.pol] est dissous en 2002.
- (c) La course à la [ent=org.pol-] *Mairie de* [ent=loc.admi-] *Paris* [-ent=loc.admi] [-ent=org.pol] a commencé entre les deux principaux candidats.
- (d) La [ent=org.pol-] *CIA* [-ent=org.pol] est chargée de l'acquisition du renseignement à l'étranger.
- (e) Pendant la Guerre froide, le [ent=org.pol-] *KGB* [-ent=org.pol] joua un rôle crucial dans la survie de l'[ent=org.gsp-] État soviétique [-ent=org.gsp]

(Nouvel, Ehrmann, Rosset, 201)

Voir aussi :

<https://damien.nouvelets.net/resourcesen/typologies.html>

# Les typologies d'EN et les campagnes d'évaluation françaises\*

Personne			Fonctions		
<i>pers.ind</i> (personne individuelle)		<i>pers.coll</i> (groupe de personnes)	<i>func.ind</i> (fonction individuelle)		<i>func.coll</i> (ensemble de fonctions)
Lieu			Produit		
administratif <i>loc.adm.town</i> , <i>loc.adm.reg</i> , <i>loc.adm.nat</i> , <i>loc.adm.sup</i>	physique <i>loc.phys.geo</i> , <i>loc.phys.hydro</i> , <i>loc.phys.astro</i>	bâtiments <i>loc.fac</i> , odonymes <i>loc.orp</i> , adresse <i>loc.add.phys</i> , <i>loc.add.elec</i>	<i>prod.object</i> (produit manufacture)	<i>prod.serv</i> (transport)	<i>prod.fin</i> (produit financier)
			<i>prod.doctr</i> (doctrine)	<i>prod.rule</i> (loi)	<i>prod.soft</i> (logiciel)
			<i>prod.art</i>	<i>prod.media</i>	<i>prod.award</i>
Organisation			Temps		
<i>org.adm</i> (administration)		<i>org.ent</i> (services)	<i>time.date.abs</i> (date absolue), <i>time.date.rel</i> (date relative)		<i>time.hour.abs</i> (heure absolue), <i>time.hour.rel</i> (heure relative)
Montant					
<i>amount</i> (avec des unités officielles ou des objets), incluant la durée					

Typologie des EN proposée par QUAERO

Types	Exemples
PERS	le <i>socialiste Bertrand Delanoë</i> , <i>Astérix</i> , la <i>diaspora argentine</i> , les <i>Beatles</i>
LOC	la <i>ville de Paris</i> , la <i>Lune</i> , l' <i>autoroute A6</i> , la <i>région Atlas</i>
ORG	la <i>société Peugeot</i> , la <i>police française</i> , le <i>syndicat FSU</i>
AMOUNT	<i>trois pompiers</i> , une <i>dizaine de voitures</i> , <i>quelques minutes</i>
TIME	<i>jeudi 16 avril</i> , en <i>1945</i> , les <i>années 1970</i> , <i>hier matin</i> , <i>il y a 3 jours</i>
PROD	<i>AK 47</i> , <i>Le malade imaginaire</i> , <i>Firefox 36.0.4</i> , la <i>palme d'or</i>
FONC	le <i>maire de Paris</i> , le <i>pompier</i> , ...

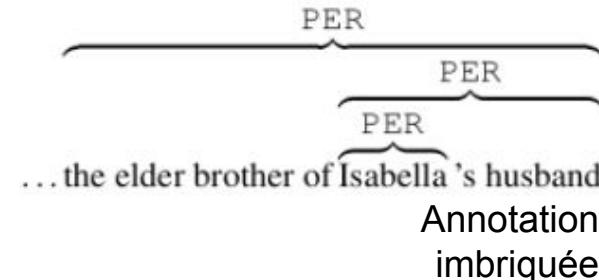
\*Le Pevedic et Maurel (2016) reviennent sur leur compatibilité avec la TEI  
<https://doi.org/10.4000/corela.4644>

# Les campagnes d'évaluation dans les HN : le début

- Named Entity Processing on Historical Newspapers (2020)
  - challenge HIPE dans le cadre de la conférence CLEF
  - Site web : <https://impresso.github.io/CLEF-HIPE-2020/>
  - 5 catégories de premier niveau (PERS, ORG, LOC, PROD, TIME), nouvelle catégorie transverse NAME, avec des sous-catégories respectivement
  - Consignes dérivées de Quaero
  - Corpus diachronique de presse (1798-1888) en français, allemand et anglais

# Les campagnes d'évaluation dans les HN : le début

- The Lit Bank: an Annotated Dataset of Literary Entities (2019)
  - Site Web : <https://github.com/dbamman/litbank>
  - On considère les EN (ex : Tom Sawyer) ainsi que les noms communs (le policier),
  - Consignes d'annotation dérivées d'ACE, les catégories retenues sont :
    - People (PER): *Tom Sawyer, her daughter*
    - Facilities (FAC): *the house, the kitchen*
    - Geo-political entities (GPE): *London, the village*
    - Locations (LOC): *the forest, the river*
    - Vehicles (VEH): *the ship, the car*
    - Organizations (ORG): *the army, the Church*
  - Corpus littéraire de fiction (XVIII-XX) des auteurs anglophones (projet Gutenberg)



# La REN dans les infrastructures de recherche pour les HN (CLARIN)

The screenshot shows the CLARIN website's "Tools for named entity recognition" page. At the top, there is a navigation bar with links for About, Language Resources, Learn & Exchange, Events, News, Contact, and a search bar. Below the navigation bar, a breadcrumb trail shows Home / Language Resources / Resource families / Tools for named entity recognition. The main title is "Tools for named entity recognition". A sub-section title "Introduction" is followed by a detailed description of Named Entity Recognition (NER) and its applications. Another section, "Tools for named entity recognition in the CLARIN infrastructure", lists various tools with their descriptions. A sidebar on the right contains a "TABLE OF CONTENTS" with sections for Introduction, Publications, and a link to "Tools for named entity recognition in the CLARIN infrastructure". At the bottom right, there is a small text box with the date "mercredi 23 février 2022".

## Tools for named entity recognition

### Introduction

Named entity recognition (NER) is an information extraction task which identifies mentions of various named entities in unstructured text and classifies them into predetermined categories, such as person names, organisations, locations, date/time, monetary values, and so forth. They can, for example, help with the classification of news content, content recommendations and search algorithms.

The CLARIN infrastructure offers 24 tools for NER. 15 tools are aimed at normalizing texts within a single language (4 Dutch, 2 English, 1 Finnish, 2 German, 1 Greek, 1 Hungarian, 3 Polish, 1 Portuguese), while the rest have a very broad multilingual scope. While 16 tools are in terms of their functionality dedicated exclusively to NER, 8 are part of tool pipelines that also provide functionalities such as PoS-tagging, lemmatisation and syntactic parsing.

For comments, changes of the existing content or inclusion of new tools, send us an [email](#).

This website was last updated on 20 October 2021.

### Tools for named entity recognition in the CLARIN infrastructure

Tool	Language	Description
CTexTools 2	Afrikaans, English, South Ndebele, Xhosa, Zulu, German	This is a corpus query and manipulation tool primarily for the official South African languages. The tool supports the creation of
<b>Functionality:</b> tokenization, sentence segmentation, PoS-		

<https://www.clarin.eu/resource-families/tools-named-entity-recognition>

# L'annotation manuelle d'un texte pour définir une EN dans un “nouveau” domaine

- ❖ Quelles sont les catégories d'entités nommées d'intérêt ?
- ❖ Quels sont les textes qui seraient les plus représentatifs de la diversité des entités nommées dans le corpus ?
- ❖ Combien d'annotateurs (humains) sont disponibles et quel est leur degré d'expertise du domaine et de la langue concernés ?

**Intérêt à définir des consignes d'annotation dans un guide, c'est un processus itératif.**

# Cas exemple : Les entités nommées et les textes littéraires

Cas d'étude : romans français du XIXe siècle rattachés aux courants réalistes et naturalistes, qui font figure de modèle dans l'histoire du genre. Annotation de plusieurs échantillons et création d'un guide inspiré d'Ester-2.

- Qu'est-ce qu'une “entité littéraire” ?
- Y a t il de traits caractéristiques (orthographiques, stylistiques, historiques, ..) ?
- Faudrait il étendre la définition des EN en dehors des noms propres, si oui, comment les définir ?
- Quelle analyse et modes de visualisation ciblons-nous : cartes et analyse de réseaux de personnages ?

# Cas exemple : Les entités nommées et les textes littéraires

<i>la mère Chantemesse</i> <i>le comte Muffat</i> <i>les filles Méhudin</i>	Les descriptions définies (DD) contenant un NP et des rôles (familiaux, amicaux, métiers, ...).
<i>nom de Dieu !</i> <i>Beethoven</i>	Parfois présents dans des expressions.
<i>La mère de Louise</i> <i>l'ami de Claude Lantier</i>	Ne pas annoter les rôles (famille, amis) mais on annote le noms de personnages y références si c'est le cas.
<i>le marchand de vin</i> <i>le jeune homme</i>	On n'annote pas les désignations génériques souvent capitalisées et trouvées sous forme de groupes nominaux contenant aucun NP

## Extraits du guide d'annotation - exemples des EN

# Accord inter-annotateur pour assurer l'homogénéité de la définition d'EN

- Il s'agit d'un **ensemble de métriques** pour déterminer la cohérence des annotations car il n'y a pas de "vérité terrain", les catégories linguistiques sont donc déterminées par le jugement humain
- Une fois le corpus annoté, il convient de mesurer la qualité et la cohérence des annotations produites, c'est-à-dire d'assurer que **chaque annotateur** aura bien eu la même compréhension de la tâche et interprétation du **guide d'annotation**
- Les indicateurs **Kappa de Cohen** pour mesurer l'accord attendu en tenant compte du hasard sont les plus répandus dans le cas de deux annotateurs. La variante **Kappa de Fleiss** permet de mesurer l'accord en présence de trois annotateurs
  - Voir un bon exemple du calcul de la méthode du Kappa proposé par Wikipédia :  
[https://fr.wikipedia.org/wiki/Kappa\\_de\\_Cohen](https://fr.wikipedia.org/wiki/Kappa_de_Cohen)

# Reconnaissance d'entités nommées (REN)

- ❖ **Détection et repérage** : déterminer les frontières, autrement dit, quels segments de texte sont concernés
- ❖ **Classification** : déterminer le type à partir d'une typologie prédéfinie

# Caractéristiques les plus souvent utilisées pour la REN

- Caractéristiques au niveau des mots,
- Listes d'entités (*gazetiers/gazetteer*, dictionnaires, ..)
- Caractéristiques des documents et corpus

# Caractéristiques au niveau des mots

- **Case** : - commence par une majuscule - le mot est tout en majuscules - le mot est en majuscules et en minuscules (par exemple, ProSys, eBay)
- **Ponctuation** : - se termine par un point, - il y a un point interne (par exemple, St., I.B.M.) - un apostrophe, trait d'union ou esperluette interne (par ex, O'Connor)
- **Chiffre** : - succession de chiffres - cardinal et ordinal - nombre romain - mot avec chiffres (ex : W3C, 3M)
- **Caractère** : - marque possessive (ex : Esther's family), pronom à la première personne
- **Morphologie** : - Préfixe, suffixe, singulier, racine du mot - fin commune
- **Catégorie grammaticale** : - nom propre, verbe, nom, mot étranger - combinaison récurrente de catégories, ex : madame François -> NOM+NAM

# Caractéristiques au niveau des mots (2)

- ❖ **Contexte**
  - local : mots qui précèdent ou suivent l'EN, ex : “Il a vu Hollande à la télévision” vs. “Son voyage en *Hollande* s'est bien passé”
  - parfois besoin de contexte plus large (phrase, phrase proche), ex : “Je me documente sur *Washington* pour mon travail”
  - coût de calcul computationnel non négligeable
- ❖ Les indices contextuels viennent en complément des indices présentés précédemment

# Listes d'entités

- **Liste générale** : - dictionnaire général - mot fonction (mots vides) - noms en majuscules - Abréviations courantes
- **Liste des entités** : - organisation, gouvernement, établissement d'enseignement - prénom, nom, célébrité - continent, pays, état, ville
- **Liste des indices d'entité** : - mots typiques dans l'organisation - titre et honorifiques de la personne, préfixe du nom - Mot générique de lieux, points cardinaux

# Caractéristiques des documents et corpus

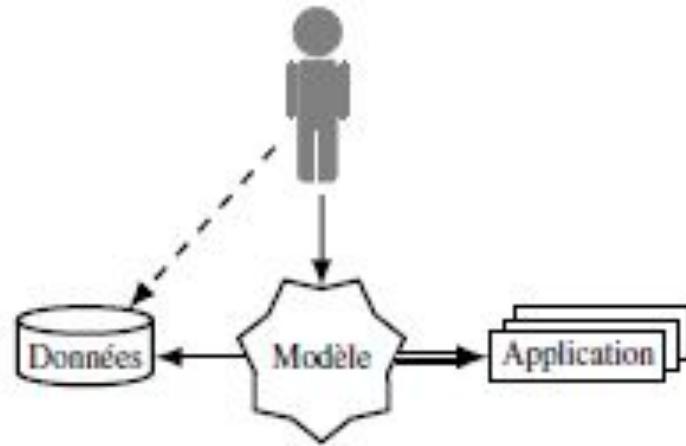
- **Occurrences multiples** : autres entités dans le contexte - occurrences en majuscules et en minuscules - anaphore/chaînes de coréférence
- **Syntaxe** : - énumération, apposition - position dans la phrase, dans le paragraphe et dans le document
- **Méta-information** : - Uri, en-tête de courriel, entête XML - listes à puces/numérotées, tableaux, figures
- **Nombre d'occurrences** : - des mots et des phrases - co-occurrences - expressions polylexicales

# REN : types d'approches

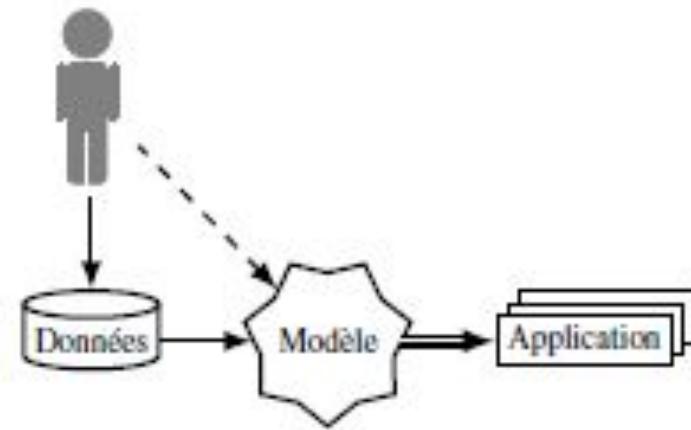
- ❖ Les approches à base de méthodes symboliques, elles reposent sur des règles élaborées par un expert et des dictionnaires (listes)
- ❖ **Les approches guidées par les données et l'apprentissage :**
  - Ces courants en croisement des mathématiques, statistiques et sciences cognitives, cherchent à déterminer les paramètres d'un modèle à partir de données
  - On distingue l'apprentissage automatique “classique” qui comprend trois modes d'apprentissage : **supervisé**, non supervisé et semi-supervisé et l'apprentissage **profond** avec ses **modèles de langue**

# REN : types d'approches

(Nouvel et al 2015)



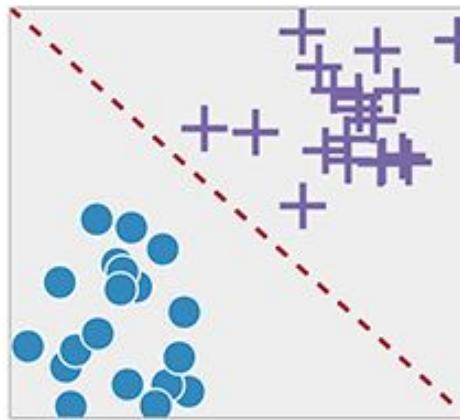
Système symbolique



Système guidé par les données

- interact majoritairement
- visualise, évalue, paramètre

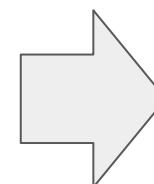
En apprentissage automatique, la REN est modélisée comme un problème de classification



A partir de données, l'algorithme vise à déterminer de **valeurs discrètes (catégories)** à attribuer à une **séquence de mots** donnée en entrée en calculant la décision par combinaison linéaire des échantillons.

Mot	Catégorie
Lucy	PER
qui	O
descend	O
...	O
dit	O
la	PER
Faloise	PER
à	O
Fauchery	PER

Pour la REN, besoin d'au moins au moins 3 catégories (PER, LOC, ORG) mais il est nécessaire de distinguer une entité polylexicale de deux EN contigues du même type, il y a donc davantage de catégories à décider.



Mot	Catégorie	Mot	Catégorie
Lucy	PER	Lucy	B-PER
qui	O	qui	O
descend	O	descend	O
...	O	...	O
dit	O	dit	O
la	PER	la	B-PER
Faloise	PER	Faloise	I-PER
à	O	à	O
Fauchery	PER	Fauchery	B-PER

Format BIO -> 2N + 1 catégories

Format BILOU -> 4N + 1 catégories

# Les approches guidées par les données et l'apprentissage

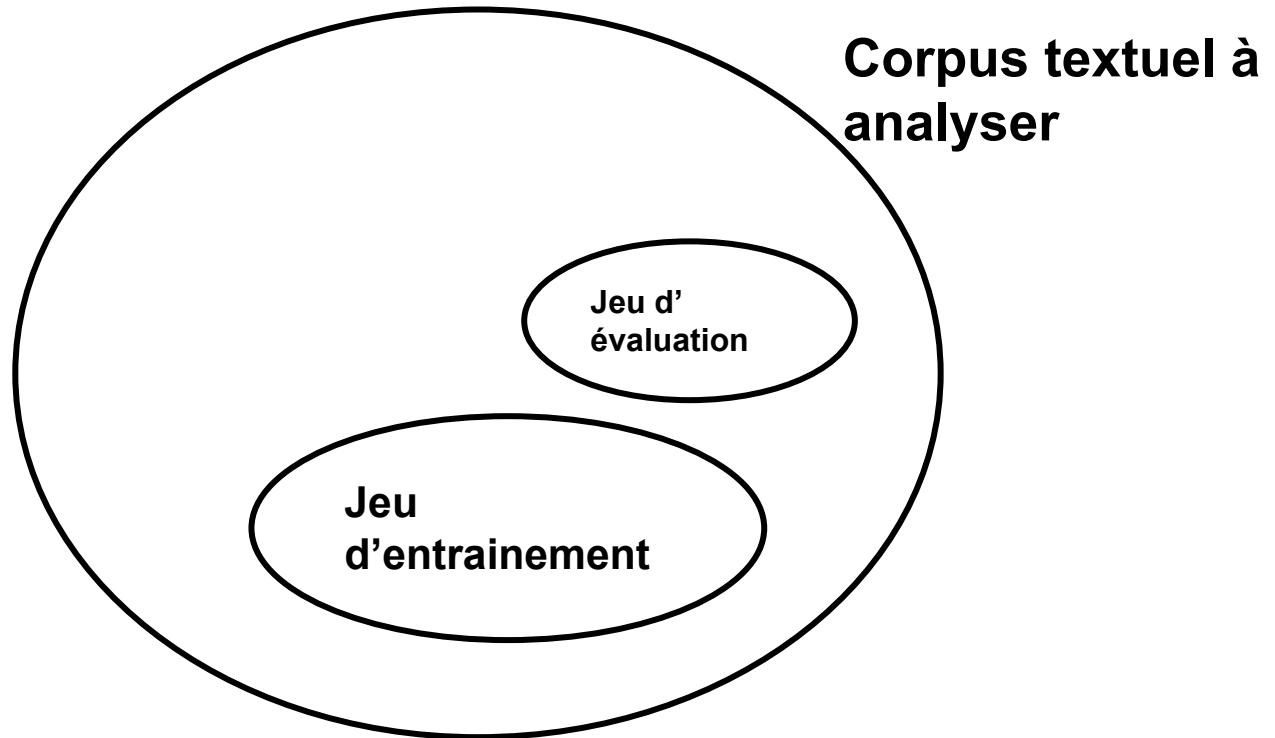
La manière dont les données sont présentées au système joue un rôle prédominant parce que :

- la quantité et la qualité des données peuvent rendre le système **plus ou moins précis** (peu de reconnaissances erronées), **couvrant** (peu de reconnaissances manquées), **robuste** (résistance au bruit),
- les modèles **ne s'adaptent toujours pas** aisément et le **genre** du texte sur lequel est réalisé l'apprentissage conditionne **l'applicabilité** du modèle à d'autres genres de texte,
- Les **pré-traitements** (segmentation de mots) et le **codage** des entités peuvent influencer la manière dont les entités sont reconnues
- Aujourd'hui, les **modèles de langue pré-entraînés** (BERT) sont indispensables et permettent de tirer avantage de l'énorme quantité de textes bruts disponibles, il faut ensuite affiner en ajoutant une couche supplémentaire; le modèle peut alors être **entraîné sur un jeu de données annotées** pour une tâche particulière (ici, la REN)

# Apprentissage supervisé

- ❖ Il s'agit de systèmes de classification qui traitent un grand corpus annoté et apprennent à partir des exemples de **textes annotés par des humains**, un modèle est donc entraîné,
- ❖ A partir du corpus d'apprentissage, ces systèmes mémorisent des listes d'entités et créent des règles de désambiguïsation basées sur des **caractéristiques discriminatoires** (comme celles listés dans le dernier cours),
- ❖ La performance du système REN :
  - dépend du **transfert de vocabulaire**, qui est la proportion de mots, sans répétitions, apparaissant dans le corpus d'entraînement et de test
  - est influencée par la **quantité** et le **format** des données annotées ainsi que par le **nombre de catégories** à apprendre (assez d'instances par classes)

# Adaptation d'un système REN : apprentissage à partir d'un corpus annoté par un utilisateur



# Evaluation d'un système d'annotation automatique

manuel : *elle est allée dans la banlieue de Barcelone elle allait diriger une usine de textile*

automatique : *elle est allée dans la banlieue de Barcelone elle allait diriger une usine de textile*

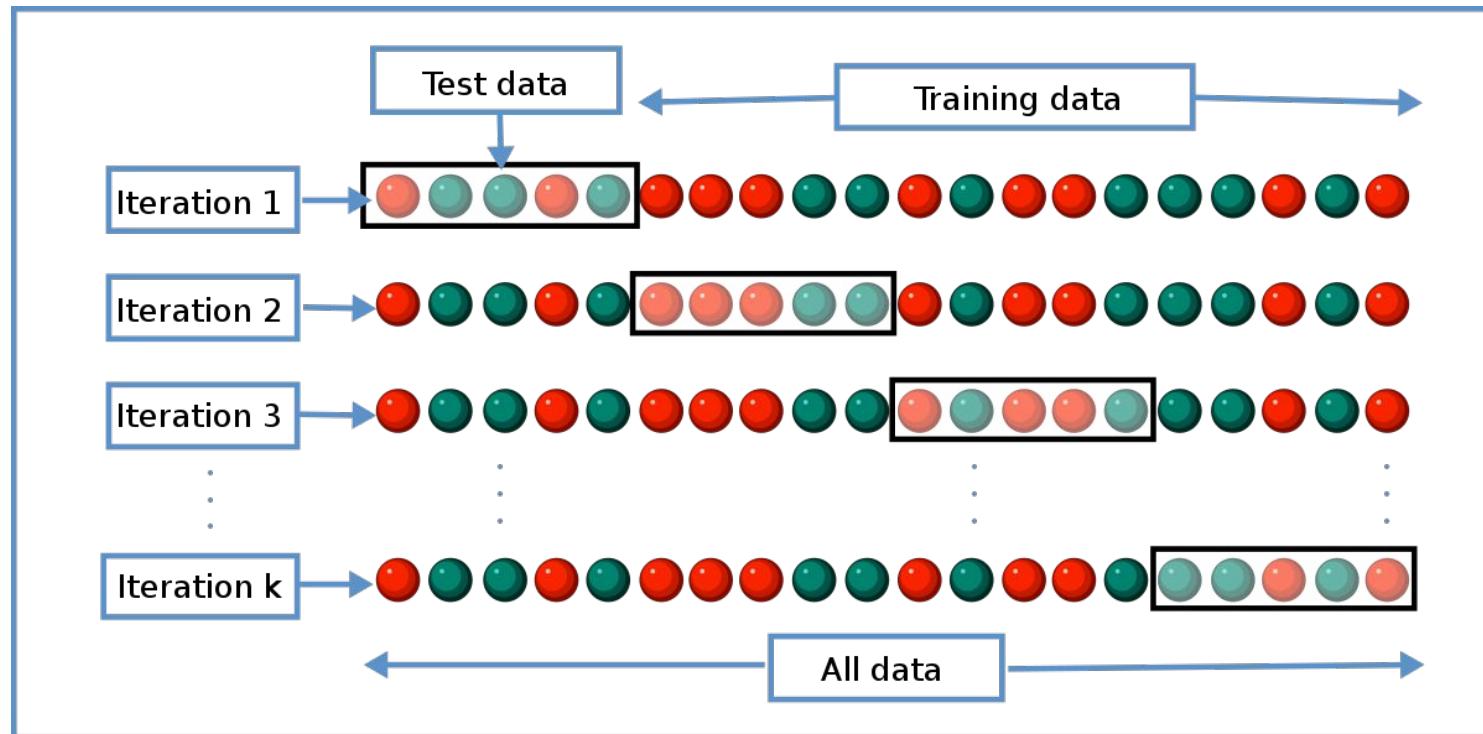
*faux positif (FP)*

*vrai positif (VP)*

*faux négatif (FN)*

- ❖ **rappel** : les réponses justes retournées par rapport à ce qui aurait dû être retourné  
=  $\text{VP} / (\text{VP} + \text{FN})$
- ❖ **précision** : les réponses justes retournées par rapport à ce qui a été retourné  
=  $\text{VP} / (\text{VP} + \text{FP})$

# Choix de sous-corpus pour entraîner : Approche de validation croisée



# Quelques outils d'annotation d'EN avec interface graphique

# GATE : intégration ad-hoc des approches REN à base de règles et apprentissage dans un outil d'annotation :

- lieux noms propres (*gazetteers*)
- lieux noms communs  
(apprentissage automatique – Stanford NER)

The screenshot shows the GATE Developer 8.3 build 5704 interface. The main window displays a text document with various entities highlighted in green, such as "Catalogne", "village", "Barcelone", "banlieue de Barcelone", "usine de textile", "Badalona", "Barcelone", "usine de textile", "cave", "terrasse", "grande terrasse", "cave", "assists", "DCA", "hôpital", "école", "rue", "difficiles", "pas", "parents", "vivaient", "plupart", "temps", "gosses", "m'ont", "appris", "vivre", "adversité", "moments", "très", "difficiles", "survivre", "conditions", "extremes", "m'a", "mûri", "six", "entre", "tout", "ce", "j'ai", "vécu", "bombardements", "et", "tout", "ça", "et", "ce", "que", "j'ai", "connu", "après". Below the text, a table lists "Annotations" with columns for Type, Set, Start, End, Id, and Features. A sidebar on the right shows a color-coded legend for entity types like Geonames (green), B-LOC (blue), and I-LOC (orange). The bottom navigation bar includes tabs for Document Editor, Initialisation Parameters, and Relation Viewer.

Type	Set	Start	End	Id	Features
Chefieu		46	54	153918	{1-source=BDNameWGS84, 2-entityType=loc, 3-theme=political, 41-vague=no}
Geonames		194	200	153920	{1-source=geonames, 2-entityType=loc, 3-theme=political, 41-vague=no}
PositiveSentiment		226	233	160528	{StringBearingPositivePolarity=plaisir}
Geonames		288	295	153922	{1-source=geonames, 2-entityType=loc, 3-theme=political, 41-vague=no}
Negation		414	418	154196	{language=fr, majorType=neg, minorType=explicit, subtype=c}
Negation		545	548	154203	{language=fr, majorType=neg, minorType=explicit, subtype=c}
B-LOC		625	632	153703	{}
B-LOC		663	670	153704	{}
Negation		781	784	154228	{language=fr, majorType=neg, minorType=explicit, subtype=c}
B-LOC		790	797	153705	{}
Negation		902	905	154235	{language=fr, majorType=neg, minorType=explicit, subtype=c}



01\_Macon\_01.txt

## 01\_Macon\_01.txt

IRHT - projet VELUM

54 Annotations · No Other Contributors · CC0 1.0

EDITION: NORMAL SAISIE RAPIDE ▾ RELATIONS COULEUR: PAR TYPE D'ENTITÉ PAR STATUS DE VÉRIFICATION PAR MARQUEUR

Omnia que processu temporis solida debent persistere apicum characteribus opportet adnotare ut valeant inconvulta per diurna tempora subsistere.

Hujus rei gratia notum habeatur omnibus huic deciduo cosmo degentibus, quia dum resideret dominus Berno episcopus secus basilicam Alme Marie Virginis Genitricis eterni authoris, expetiit ab eo dominus Albericus, comes illustris, ex rebus aliquid terrenis subtrahi Sancti Vincentii Matisconensis, fiscum videlicet unum in pago Lugdunensi conjacente ecclesia Sanctorum Amoris et Viatoris, et quitquid in Ventiaco villa cernebat habere, ministerium-ve prepositure dignitatis, quecumque ibidem subjecta sunt et cum capellis inibi adjacentibus, inquisitis et inquirendis, universa sibi conferri sub integritate oppido postulavit. At dominus Berno pretaxatus pontifex annuens precibus prelibati comitis contulit illi ecclesiam geminorum Sanctorum Amoris et Viatoris cum omnibus appenditiis sicut superius inserta sunt. Ut ipse Leutaldus atque et Umbertus filii prenominati Alberici, temporibus vite illorum ipsas res suis adaptare usibus, firmiter sine ullius controversia valerent et ut liberius hoc tenere quiverint sine ulla contradictione precario quia facta erant ista, et in helemosina Sancto Vincentio fuerant collata, donaverunt Sancto Vincentio in pago Matisconensi capellam unam in honore Beati Bartholomei apostoli dicatam et in villa Fabricas sitam, et quitquid in ipsa villa visi erant habere cum cunctis rebus ibi adherentibus, ut abhinc et deinceps sine ulla contradictione Sanctus Vincentius et rectores ejus teneant et possideant, et in pago Scodingensi villam



Active Learning

Named entity

Recommendation

Text

Sugg estio n

Scor e

value

Annotate

History

demo: LitEN/chapitre1-5eshuf.txt 1-75 / 75 lines [doc 1 / 1]

1

3 PER gendarme ! Hector crut qu'il devait chercher une phrase aimable.

4 Mais un léger frémissement agita la salle. PER Rose Mignon venait d'entrer, en Diane. (Name)

Bien quelle n'eût ni la taille ni la figure du rôle, maigre et noire, d'une laideur adorable. Son air d'entrée, des paroles bêtes à pleurer, où elle se plaignait de Mars, qui était en PER

train de la lâcher pour Vénus, fut chanté avec une réserve pudique, si pleine de PER

sous-entendus égrillards, que le public s'échauffa. Le mari et Steiner, coude à coude, riaient complaisamment. Et toute la salle éclata, lorsque Prullière, cet acteur si aimé, se montra en général, un Mars de la Courtille, empanaché d'un plumet géant, traînant un PER

sabre qui lui arrivait à l'épaule. Lui, avait assez de Diane ; elle faisait trop sa poire. Alors PER, Diane jurait de le surveiller et de se venger. Le duo se terminait par une tyrolienne PER

Layer

Annotation

No annotation selected



cvbe-ehess



A

Save

Confirm



1 of 1

A neuf heures, la salle du théâtre des Variétés était encore vide. Quelques personnes, au balcon et à l'orchestre, attendaient, perdues parmi les fauteuils de velours grenat, dans le petit jour du lustre à demi-feux. Une ombre noyait la grande tache rouge du rideau; et pas un bruit ne venait de la scène, la rampe éteinte, les pupitres des musiciens débandés. En haut seulement, à la troisième galerie, autour de la rotonde du plafond où des femmes et des enfants nus prenaient leur volée dans un ciel verdi par le gaz, des appels et des rires sortaient d'un brouhaha continu de voix, des têtes coiffées de bonnets et de casquettes s'étagaient sous les larges baies rondes, encadrées d'or. Par moments, une ouvreuse se montrait, affairée, des coupons à la main, poussant devant elle un monsieur et une dame qui s'asseyaient, l'homme en habit, la femme mince et cambrée, promenant un lent regard.

Deux jeunes gens parurent à l'orchestre. Ils se tinrent debout, regardant.

-- Que te disais-je, **Hector**? s'écria le plus âgé, un grand garçon à petites moustaches noires, nous venons trop tôt. Tu aurais bien pu me laisser achever mon cigare.

Une ouvreuse passait.

-- Oh! monsieur **Fauchery**, dit-elle familièrement, ça ne commencera pas avant une demi-heure.

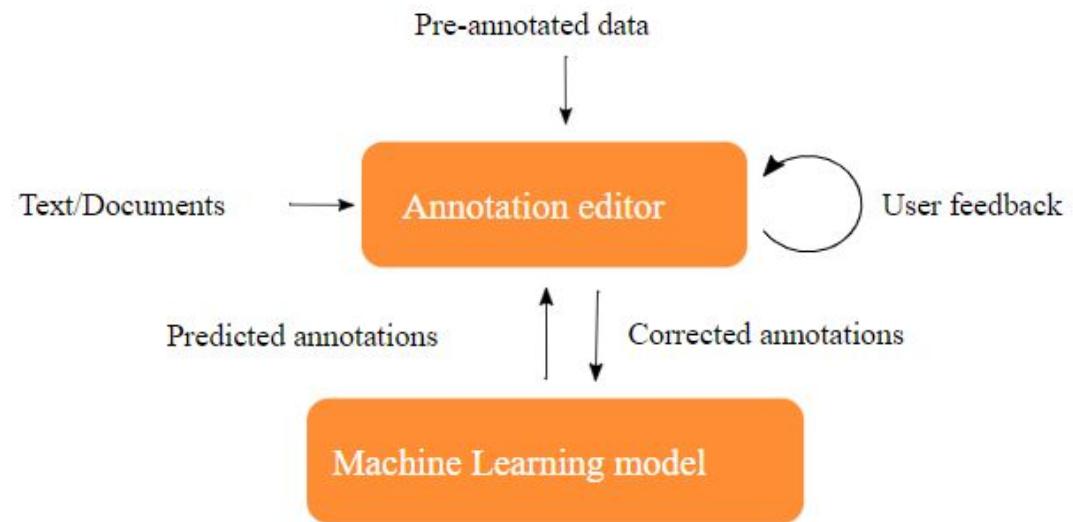
## Entities

total **2216** not normalized **2216**

Group/filter entities ▾

	Personnages	2215	◀
	14	⚡ 2201 (0.00%) IAA:	
	NANA	2	▼
	Hector	13	▼
	Fauchery	249	▼
	Lucy	66	▼
	Bordenave	157	▼
	Nana	866	▼
	la Faloise	97	▼

# Exercice guidé



<https://github.com/cvbrandoe/coursTAL/tree/master/2022>



cvbe-ehess



A

faisait un peristyle de temple en carton, de hautes affiches jaunes s'étaisent violemment, avec le nom de **Nana** en grosses lettres noires. Des messieurs, comme accrochés au passage, les lisaien; d'autres, debout, causaient, barrant les portes; tandis que, près du bureau de location, un homme épais, à large face rasée, répondait brutalement aux personnes qui insistaient pour avoir des places.

-- Voilà **Bordenave** dit **Fauchery**, en descendant l'escalier.

Mais le directeur l'avait aperçu.

-- Eh! vous êtes gentil! lui cria-t-il de loin. C'est comme ça que vous m'avez fait une chronique... J'ai ouvert ce matin le Figaro. Rien.

-- Attendez donc! répondit **Fauchery**. Il faut bien que je connaisse votre **Nana**, avant de parler d'elle... Je n'ai rien promis, d'ailleurs.

Puis, pour couper court, il présenta son cousin, M. **Hector** de la Faloise, un jeune homme qui venaitachever son éducation à **Paris**. Le directeur pesa le jeune homme d'un coup d'oeil. Mais **Hector** l'examinait avec émotion. C'était donc là ce **Bordenave**, ce monteur de femmes qui les traitait en garde-chiourme, ce cerveau toujours fumant de quelque réclame, criant, crachant, se tapant sur les cuisses, cynique, et ayant un esprit de gendarme! **Hector** crut qu'il devait chercher une phrase aimable.



Save



Confirm



1 of 1

## Entities

total 2318 not normalized 2318

Group/filter entities

Personnages 85%

2215

14 ⚡ 2201 (0.00%)

NANA

2

Hector

13

Fauchery

249

Lucy

66

Bordenave

157

Nana

866

la Faloise

97

l a Faloise

48



## Projects / CoursENTAL / doc1.txt

Cours sur les entités nommées du module TAL pour le master humanités numériques de PSL

Settings Documents Metrics Downloads

admin



pool

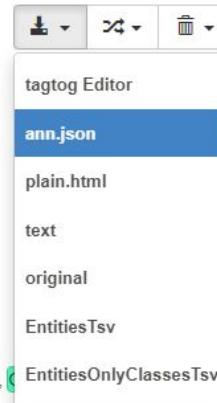
PREMIÈRE PARTIE.

Folder actions

+ Add new

|

Quand la caissière lui eut rendu la monnaie de sa pièce de cent sous,



Comme il portait beau, par nature et par pose d'ancien sous-officier, il cambra sa taille, frisa sa moustache d'un geste militaire et familier, et jeta sur les dîneurs attardés un regard rapide et circulaire, un de ces regards de joli garçon, qui s'étendent comme des coups d'épervier.

Les femmes avaient levé la tête vers lui, trois petites ouvrières, une maîtresse de musique entre deux âges, mal peignée, négligée, coiffée d'un chapeau toujours poussiéreux et vêtue d'une robe toujours de travers, et deux bourgeoises avec leurs maris, habituées de cette aeroote à prix fixe.

 Save  Confirm

1 of 3



## Entities

total 72 not normalized 72

Group/filter entities

Personnages 34  
9 25 (80.00%)

Georges Duroy 3

Duroy 10

Forestier 13



## Examples: import text pre-annotated by spaCy

Python

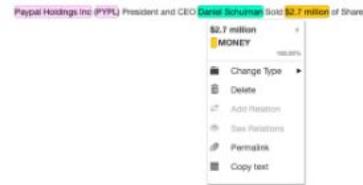
This example shows how to generate a set of annotations with a spaCy model and send the pre-annotated text to tagtog. The model used is `en_core_web_sm`. We want to do NER and extract `PEOPLE`, `ORG`, and `MONEY` entities (see "Label Scheme").

For more details, check out this step-by-step guide: [Integrating tagtog and spaCy & the full GitHub repository](#).

```
import spacy
import json
import requests
import os

def get_class_id(label):
    """
    Translates the spaCy label id into the tagtog entity type
    - label: spaCy label id
    """
    choices = {'PERSON': 'e_1', 'ORG': 'e_2', 'MONEY': 'e_3'}
    return choices.get(label, None)

def get_entities(spans, pipeline):
    """
    Translates a tuple of named entity Span objects (https://list-of-tagtog-entities)
    into tagtog entities (https://docs.tagtog.net/anndoc.html)
    """
    ann = []
    for span in spans:
        ann.append({
            'label': get_class_id(span.label_),
            'start': span.start,
            'end': span.end,
            'text': span.text
        })
    return ann
```



The resulting pre-annotated document visualized in tagtog editor

## project.yml

The `project.yml` defines the data assets required by the project, as well as the available commands and workflows. For details, see the spaCy projects documentation.

### Commands

The following commands are defined by the project. They can be executed using `spacy project run [name]`. Commands are only re-run if their inputs have changed.

Command	Description
<code>train</code>	Train a named entity recognition model
<code>devuate</code>	devuate the model and export metrics
<code>package</code>	Package the trained model so it can be installed
<code>visualize-model</code>	Visualize the model's output interactively using Streamlit
<code>visualize-data</code>	Explore the annotated data in an interactive Streamlit app

### Workflows

The following workflows are defined by the project. They can be executed using `spacy project run [name]` and will run the specified commands in order. Commands are only re-run if their inputs have changed.

Workflow	Steps
<code>all</code>	<code>train</code> → <code>devuate</code>

Exemple projet Spacy en mode CLI :

[https://github.com/PSIG-EHESS/SavoirsEN/tree/main/ner\\_savoirs](https://github.com/PSIG-EHESS/SavoirsEN/tree/main/ner_savoirs)

The agreement percentage near the title of each annotation task represents the average agreement for this annotation task.

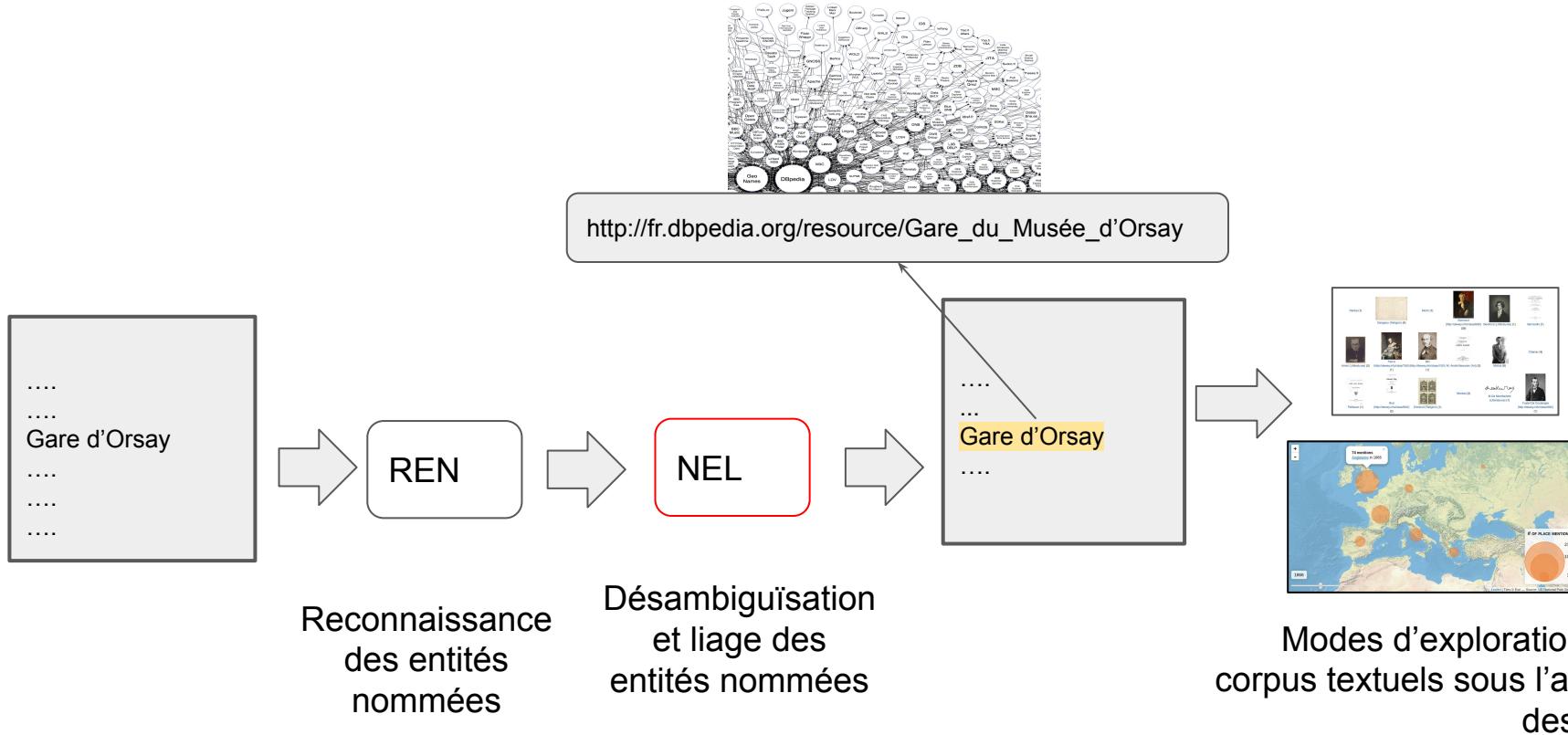
Prosecutor-Request e\_16 59.30%

	Mary	Gerard	Vega	Joao	Linda
Mary		25.00%	75.00%	66.67%	90.12%
Gerard	25.00%		47.18%	35.00%	18.00%
Vega	75.00%	47.18%		87.33%	95.65%
Joao	66.67%	35.00%	87.33%		66.67%
Linda	90.12%	18.00%	95.65%	66.67%	

**Inter-annotator agreement matrix.** It contains the scores between pairs of users. For example, Vega and Joao agree on the 87% of the cases. Vega and Gerard on the 47%. This visualization provides an overview of the agreement among annotators. It also helps find weak spots. In this example we can see how Gerard is not aligned with the rest of annotators (25%, 47%, 35%, 18%). A training might be required to have him aligned with the guidelines and the rest of the team. On the top left we find the annotation task name, id and the agreement average (59.30%).

## Accord inter-annotateur sur TagTog

# Chaîne TAL pour identifier les EN dans un corpus textuel en vue d'une étude outillée et d'une cartographie



# La chaîne complète dans le contexte de la TEI

*“Je rentre à Paris.”*

- Reconnaissance
  - Je suis à <placeName>Paris</placeName>.
- Disambiguation
  - Je suis à <placeName ref="#ParisFrance">Paris</placeName>.
- Référencement
  - Je suis à <placeName ref="<http://www.geonames.org/2988506>">Paris</placeName>.
- Représentation
  - 

# Lier les EN aux référentiels

Aussi connu sous le nom de **Named Entity Linking**, il s'agit d'un moyen d'établir un lien explicite entre les mentions d'EN cités dans le texte et les objets du monde auxquels elles réfèrent. Créer ce lien est important afin de lever toute ambiguïté sur l'identité de l'EN. Il y a donc deux objectifs :

- Désambiguer

"Goncourt" - Edmond de Goncourt ou Jules de Goncourt ?

"Voltaire", "François-Marie Arouet" - manières de désigner la même personne

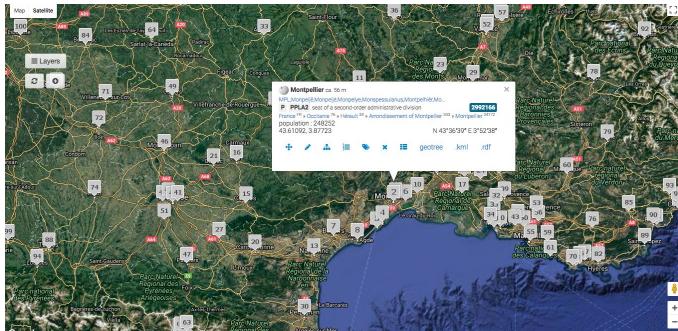
- Lier à un référentiel

"Voltaire" - associer l'entrée correspondante dans Wikidata  
(<https://www.wikidata.org/wiki/Q9068>)

# Lier les EN aux référentiels

- La **construction d'un référentiel** n'est pas une tâche triviale : il doit être aussi **couvrant** possible, mais il est impossible de connaître en amont toutes les entités cités ni toutes leurs désignations possibles,
- Il s'agit donc d'établir une liste aussi large que possible de référents auxquels lier les expressions linguistiques, chacun disposant d'un **identifiant unique**,
  - cette tâche effectuée souvent de manière semi-automatique car besoin de choisir les données en accord avec le genre de texte à traiter,
  - Il est important de mesurer la complétude du référentiel vis-à-vis du texte pour éventuellement l'enrichir.

# Les bases de connaissances en tant que référentiel pour le NEL



geonames.org

Typologie de lieu

Noms alternatifs  
Noms dans autres langues

Hierarchie

2992166

N 43°36'39" E 3°52'38"

Population

Identifiant unique

Geo-localisation

Collections Search 

## Paris, L'Avenue du Bois de Boulogne

 NO COPYRIGHT - UNITED STATES    DOWNLOAD

This image is available for download, without charge, under the Getty's Open Content Program.

## avenue Foch (Q790562)



avenue in Paris, France

 In more languages

Configure

Language	Label	Description	Also known as
English	avenue Foch	avenue in Paris, France	
French	avenue Foch	avenue de Paris, France	Avenue du Bois-de-Boulogne Avenue du Bois Avenue de l'Impératrice Avenue du Général-Uhrich
Spanish	Avenida Foch	No description defined	Avenue Foch
German	Avenue Foch	Prachtstraße in Paris	

All entered languages

## Statements

instance of

avenue

 0 references add reference

# Bases de connaissances historiques

Click the  icon to view the hierarchy.

[Semantic View \(JSON, JSONLD, RDF, N3/Turtle, N-Triples\)](#)

ID: 7658922

## Cenon-sur-Vienne (inhabited place)

Coordinates:

Lat: 46 46 27 N degrees minutes Lat: 46.7743 decimal degrees  
Long: 000 32 13 E degrees minutes Long: 0.5370 decimal degrees

Names:

Cenon-sur-Vienne (preferred)  
Cenon (C,V)

Hierarchical Position:

-  World (facet)
-  ... Europe (continent) (P)
-  ..... France (nation) (P)
-  ..... Île-de-France (region)
-  ..... Paris (inhabited)

Additional Parents:

-  World (facet)
-  ... Europe (continent) (P)
-  ..... France (nation) (P)
-  ..... Nouvelle-Aquitaine
-  ..... Cenon-sur-Vie

Place Types:

inhabited place (preferred, i

Sources and Contributors:

Cenon..... [VP]  
..... NGA/NIMA data  
Cenon-sur-Vienne..... [VP]  
..... NGI

Subject: .... [VP]

..... NGA/NIMA database (2003-) -1417396

 **Lutetia**  
a Pleiades place resource

Creators: C. Haselgrove, J. Kunow  
Contributors: DARMC, Sören Stark, R. Talbert, Sean Gillies, Johan Åhfeldt, Jeffrey Becker, Tom Elliott  
Copyright © The Creators. Sharing and remixing permitted under terms of the Creative Commons Attribution 3.0 License (cc-by).  
Last modified Feb 26, 2017 08:56 AM — History

tags: dare.ancient=1, dare.feature=settlement, dare.major=0

Lutetia (modern Paris) was the capital of the Parisii, a tribe of ancient Gaul.

Canonical URI for this page:  
<https://pleiades.stoa.org/places/109126> 

Representative Point (Latitude, Longitude):  
48.0511447, 2.34702675 

 Locations:

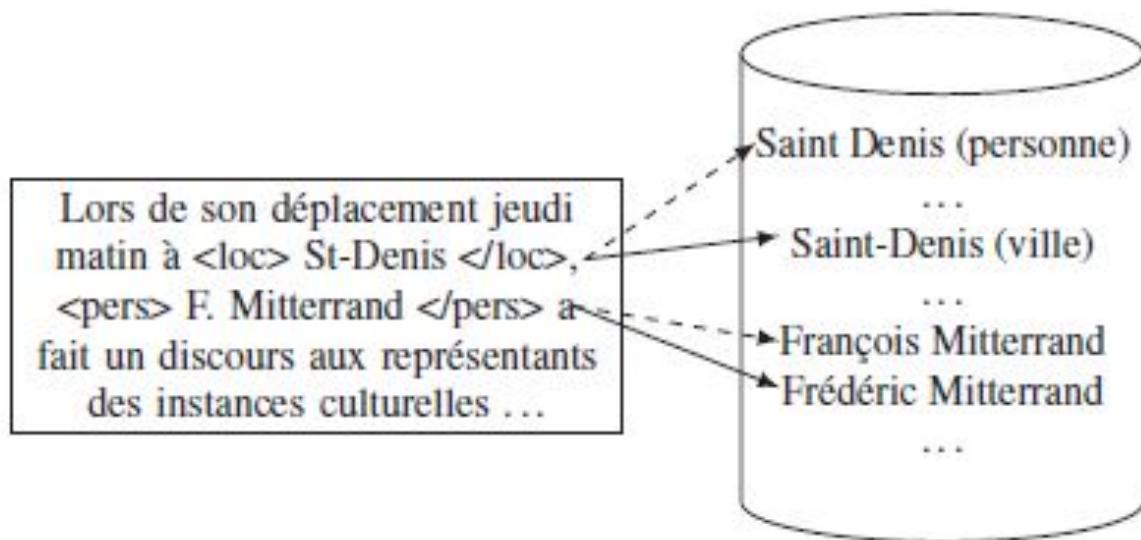
- DARE Location (330 BC - AD 640)
- DARMC location 13544 (330 BC - AD 640)
- location of Arènes de Lutèce (30 BC - AD 300)
- location of Thermes de Cluny (30 BC - AD 300)

Names:

- Loukorokia (Loutokolia; 30 BC - AD 300)
- Luteci (AD 300 - AD 640)
- Lutecia (330 BC - AD 640)
- Lutetia (330 BC - AD 640)
- Lutetia Parisiorum (330 BC - 30 BC)
- Luticia (330 BC - AD 640)
- Paris (modern)



# Lier les EN aux référentiels



Plusieurs entités candidates pour une mention d'EN (graphe bipartite)

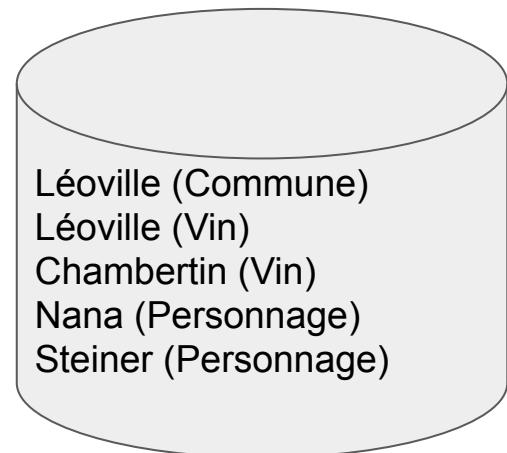
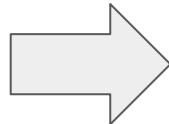
La mention étudiée hors de son contexte est ambiguë. Un humain peut typiquement lever l'ambiguïté par utilisation conjointe de **la connaissance du monde** dont il dispose et d'**indices contextuels**.

# Lier les EN aux référentiels :

## Besoin de constituer de référents pertinants pour une application

**Léoville ou Chambertin ?**  
murmura un garçon, en allongeant la tête entre **Nana et Steiner**, au moment où celui-ci parlait bas à la jeune femme.

Texte annoté en EN :  
Nana (Zola)



Référents possibles

# NEL : les étapes et les approches

Les phases du processus :

- Rechercher de mentions d'EN dans un texte (REN)
- (1) Sélection de candidats pour chaque mention
- (2) Choix du meilleur candidat pour chacune, ou attribuer NIL

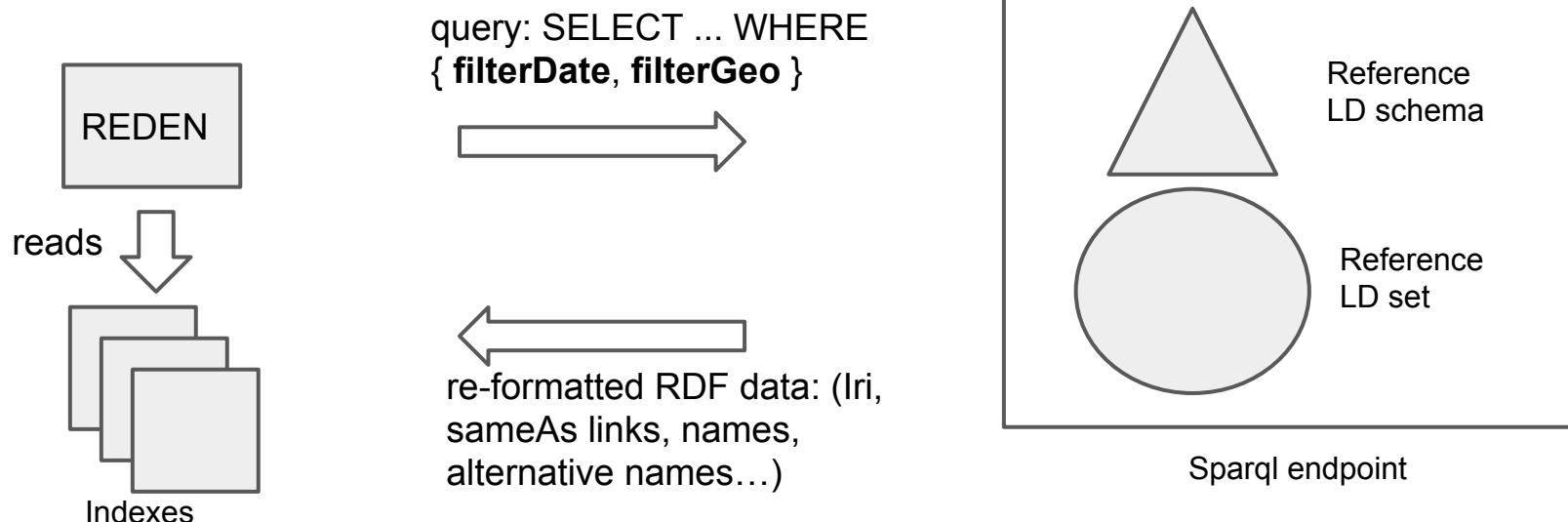
Les approches :

- **à base de textes** : utilisation conjointe de mesures de distances de chaîne de caractères et de fréquence de mots, avec parfois des prétraitements linguistiques (racinisation, lemmatisation),
- **à base de graphes** : utilisation de la structure de graphes sous-jacente (données RDF, hyperliens Wikipedia).

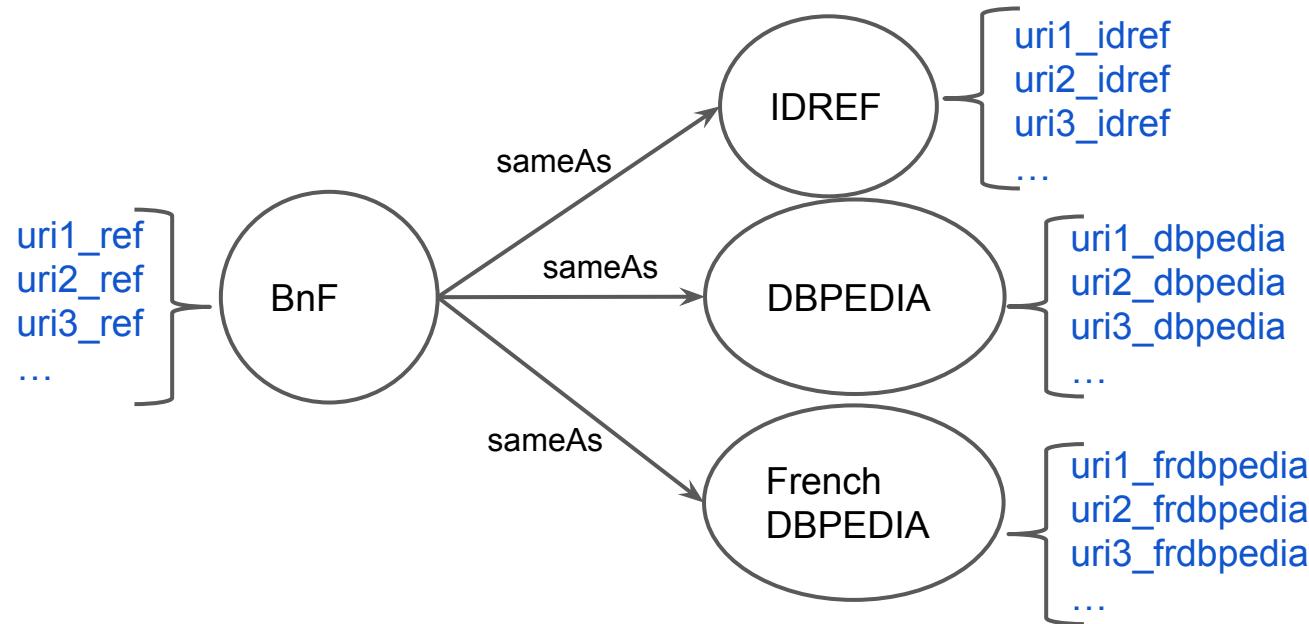
# REDEN : une approche NEL à base de graphes

- ❖ Approche fondée sur les graphes qui s'appuie sur :
  - la connaissance des **données liées** (plusieurs bases),
  - la notion de **degré de centralité** d'un graphe,
- ❖ Selon le **type d'entités**, une **base de connaissances** (DBPEDIA, BNF, Geonames, LGD, ...) peut être pertinente, REDEN peut en principe être utilisé pour toute source à condition de disposer d'un **point d'interrogation SPARQL**.

# REDEN : constitution d'index & sélection de candidats

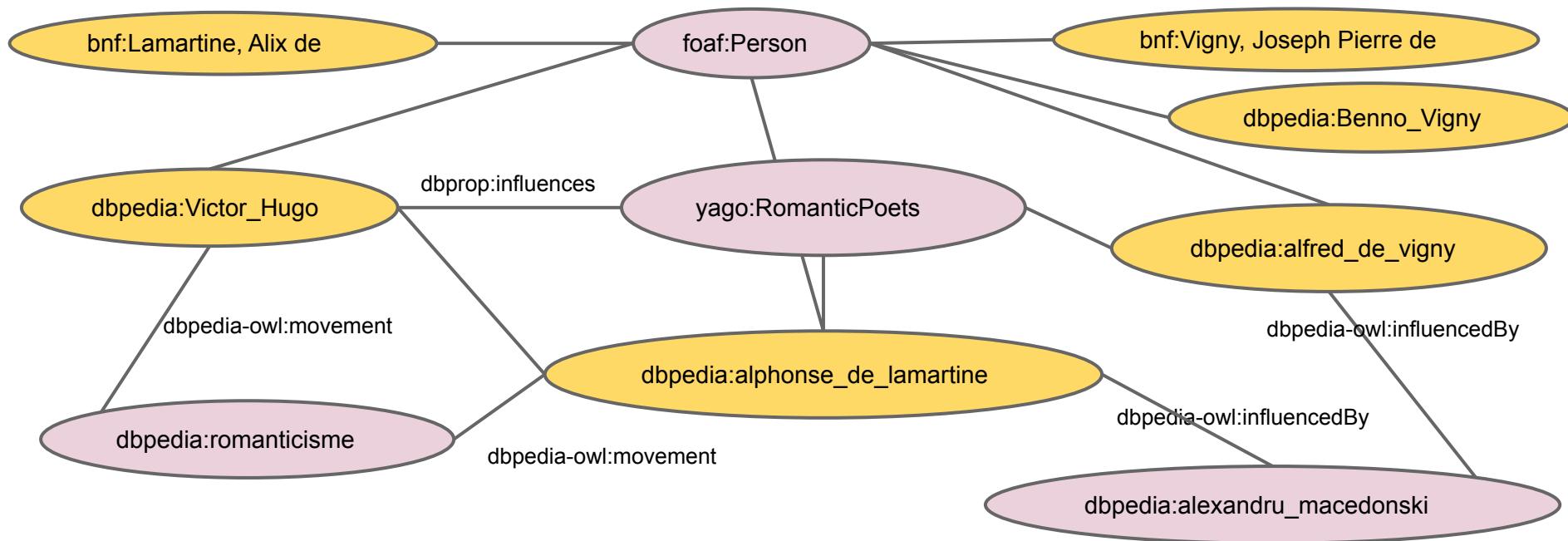


# REDEN : constitution d'index & sélection de candidats (2)



# REDEN - désambiguïsation et liage d'EN

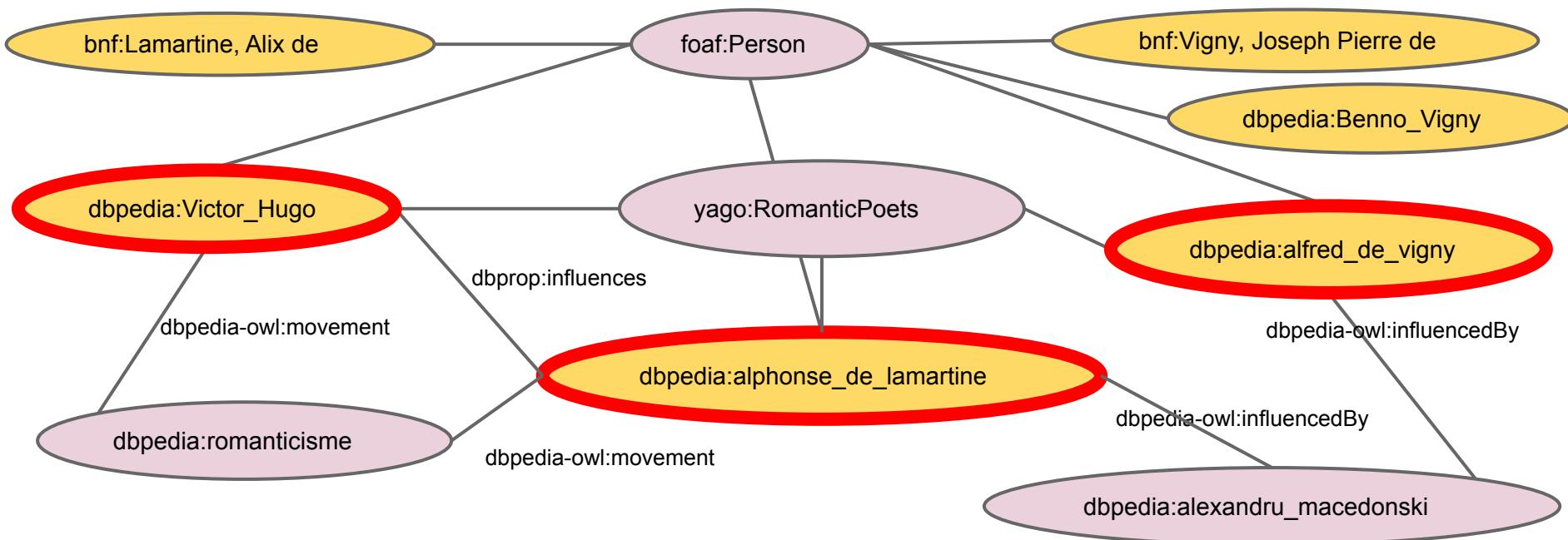
Recherche de sous-graphes connexes et calcul de la centralité du graphe lors du choix des candidats (le plus “populaire” dans le contexte)



“Quant au rythme, si Victor Hugo a dépassé Lamartine, il n'a pas été plus loin que Vigny.”

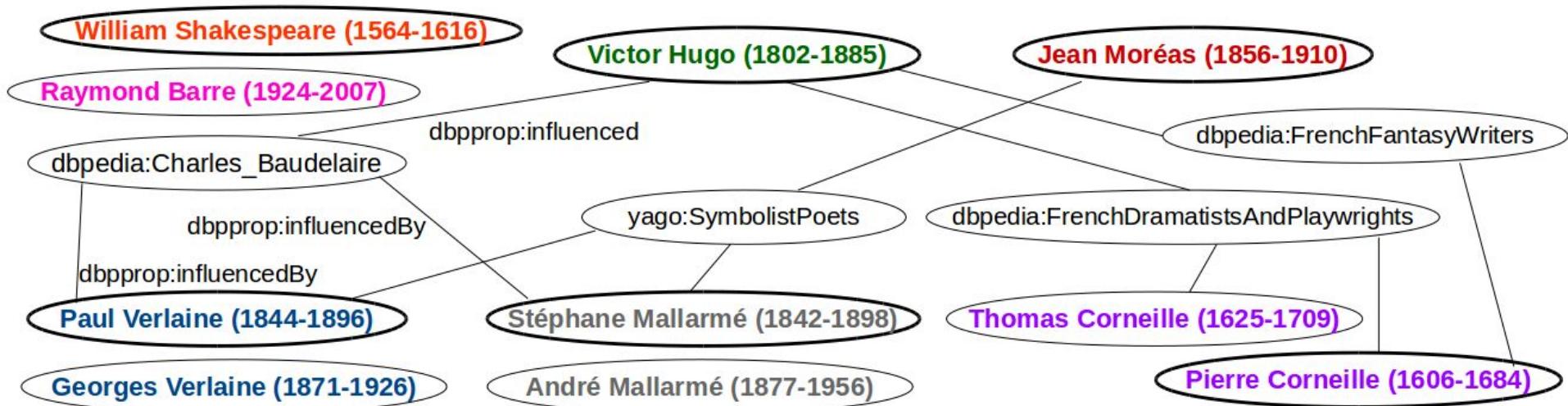
# REDEN - désambiguïsation et liage d'EN

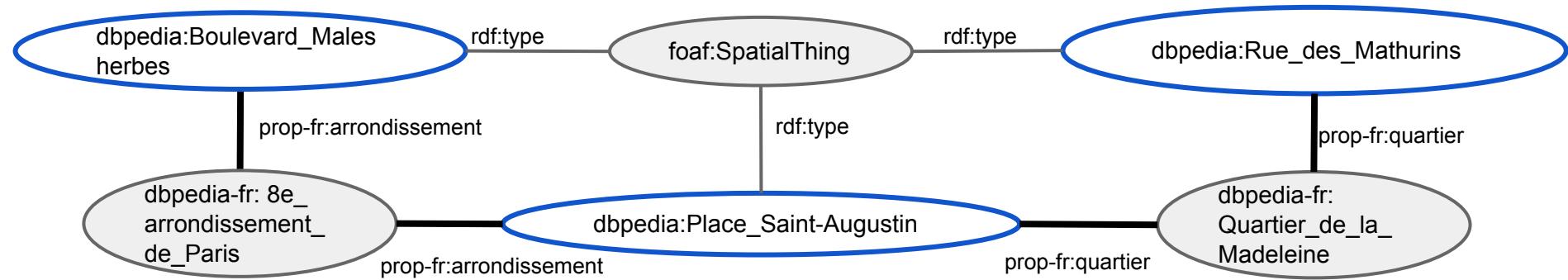
Recherche de sous-graphes connexes et calcul de la centralité du graphe lors du choix des candidats (le plus “populaire” dans le contexte)



“Quant au rythme, si Victor Hugo a dépassé Lamartine, il n'a pas été plus loin que Vigny.”

Mais, en somme, ... , lorsqu'il s'est mis à réfléchir sur le sens de son œuvre. C'est à cette place que l'on situerait par exemple, chez **Corneille** ou **Victor Hugo**, les discours sur le poème dramatique, ou **William Shakespeare**. La troisième partie du livre est consacrée aux maîtres du symbolisme, qui sont, d'après **M. Barre**, **Verlaine**, **Mallarmé** et **Moréas** (Réflexions sur la littérature, Albert Thibaudet).





“ Voilà ! J'avais eu affaire, **rue de la Pépinière**, près de la **place Saint-Augustin**, et je revenais par le **boulevard Malesherbes** en l'intention de prendre l'omnibus à la **Madeleine**. Tout à coup, au coin de la **rue des Mathurins**, un homme se dressa devant moi en criant : “Madame ou mademoiselle, [...]. ”  
(Le passant de Prague, Guillaume Apollinaire)