

# Classification d'articles encyclopédiques et reconnaissance d'entités nommées : Application à l'Encyclopédie de Diderot et d'Alembert

Equipe projet GEODE :  
L. Moncla, D. Vigier, K. McDonough, A. Brenon, T. Joliveau

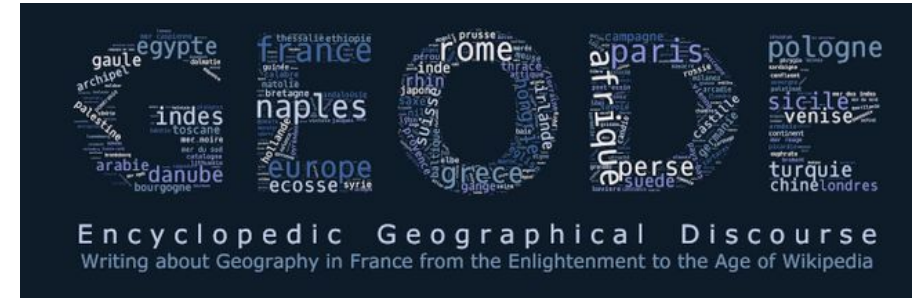


# SOMMAIRE

1. Introduction
2. Classification des articles encyclopédiques
  - 2.1. Problématique
  - 2.2. Expérimentations
  - 2.3. Discussion & Perspectives
3. Reconnaissance et classification des entités nommées
  - 3.1. Indices linguistiques forts
  - 3.2. Indices linguistiques faibles
  - 3.3. Expérimentations et évaluation

# Introduction

## Contexte



## Projet GEODE

- Projet interdisciplinaire financé par le LabEx ASLAN

<https://geode-project.github.io/>

## Objectif

- Analyse des discours géographiques dans les encyclopédies françaises

## Tâches

- Préparation et homogénéisation des différents corpus (EDdA, LGE, Universalis, Wikipedia)
- Conception des méthodes et d'outils d'analyse
  - Classification supervisée, génération de modèles de langues, repérage automatique de routines discursives

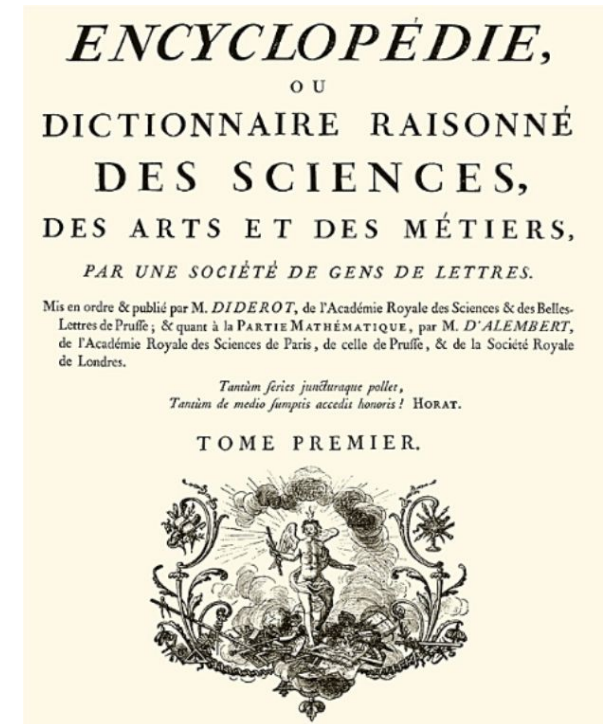
# Introduction

## Le corpus

- Édition numérique du premier tirage de l'édition originale dite « de Paris » de l'EDdA communiquée par l'*American and French Research on the Treasury of the French Language* (ARTFL) de l'Université de Chicago
- 17 tomes de texte (11 de planches) : environ 74k articles

## Les travaux en cours

- Classification automatique des articles
  - repérer les articles de géographie parmi les articles non classés
  - analyser l'évolution des domaines au cours du temps
- Reconnaissance et classification automatique des entités nommées
  - identifier les noms de lieux et de personnes
- Repérage des coordonnées géographiques



# SOMMAIRE

1. Introduction
2. Classification des articles encyclopédiques
  - 2.1. Problématique
  - 2.2. Expérimentations
  - 2.3. Discussion & Perspectives
3. Reconnaissance et classification des entités nommées
  - 3.1. Indices linguistiques forts
  - 3.2. Indices linguistiques faibles
  - 3.3. Expérimentations et évaluation

# Classification des articles encyclopédiques

## Problématique

### Classification en domaines

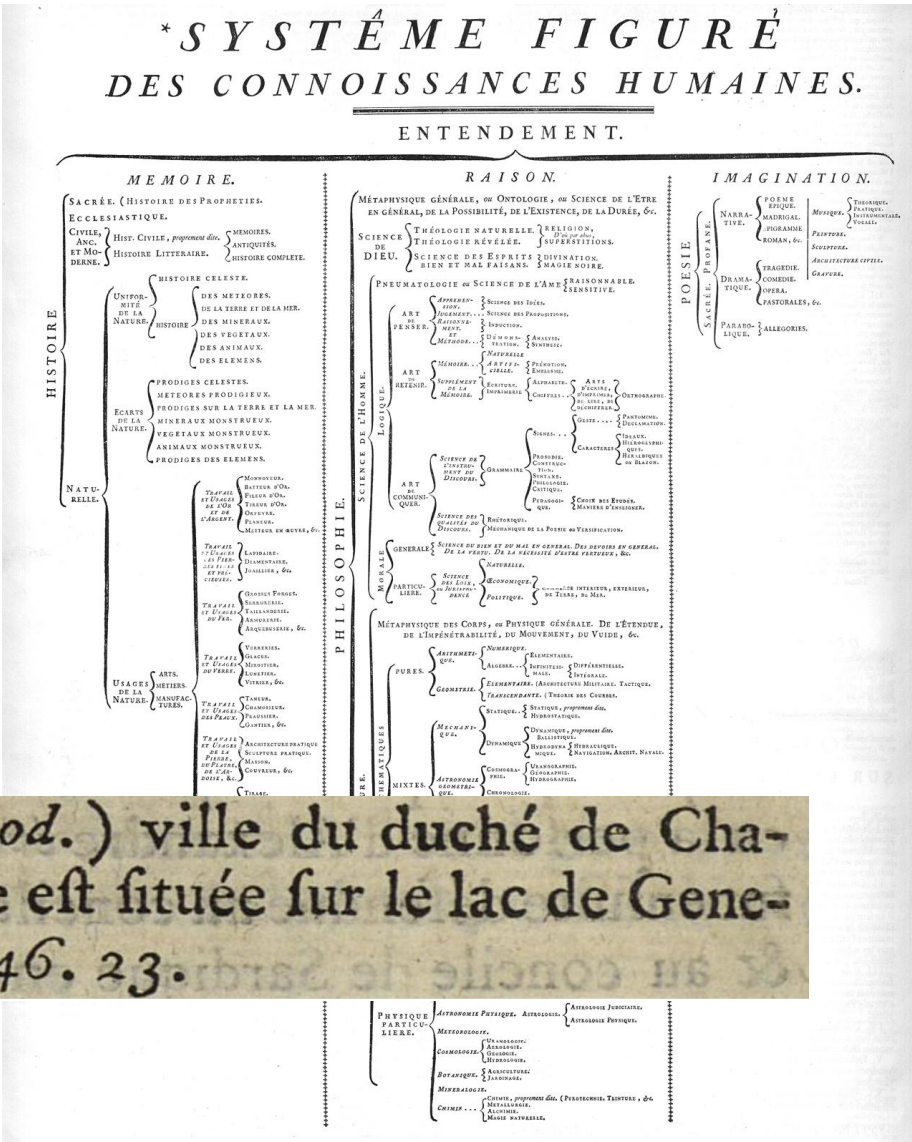
- Liste des domaines non figée
- Classification multi-domaines
- 2 sources : ARTFL et ENCCRE

### ARTFL

- 2 620 classes normalisées (normalisation orthographique)
- 12 685 articles non classés

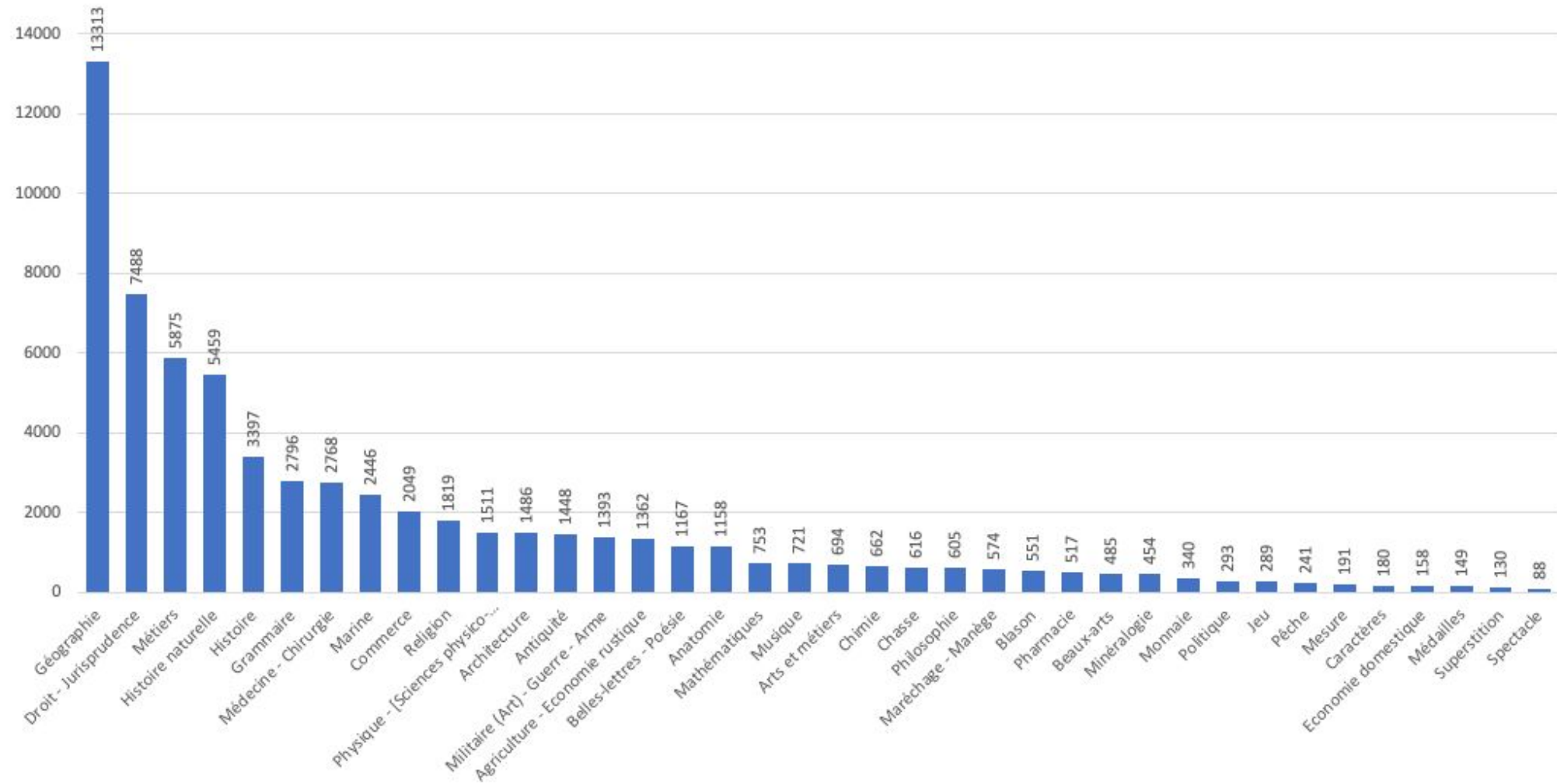
### ENCCRE

- > 7 000 désignants
- 44 ensembles de domaines (327 domaines)
- 2 392 articles non classés



# Classification des articles encyclopédiques

## Problématique



# SOMMAIRE

1. Introduction
2. Classification des articles encyclopédiques
  - 2.1. Problématique
  - 2.2. Expérimentations
  - 2.3. Discussion & Perspectives
3. Reconnaissance et classification des entités nommées
  - 3.1. Indices linguistiques forts
  - 3.2. Indices linguistiques faibles
  - 3.3. Expérimentations et évaluation



# Classification des articles encyclopédiques

## Méthodologie

- Préparation des données
  - tokenisation, lemmatisation, suppression mots vides
  - équilibrage des classes
- Vectorisation des articles
  - bag of words, TF-IDF, Doc2Vec, FastText
- Classification supervisée
  - Naive Bayes, Logistic Regression, Random Forest, SGD, SVM
  - CNN et LSTM
  - BERT et CamemBERT

# Classification des articles encyclopédiques

## Comparaison des méthodes de classification (f-mesure)

Classifieur	Vectorisation	Test		
		(1)	(2)	(3)
Naive Bayes	Bag of Words	0.72	0.68	0.61
	TF-IDF	0.74	0.59	0.37
Logistic Regression	Bag of Words	0.85	0.85	0.86
	TF-IDF	<b>0.88</b>	<b>0.88</b>	<u>0.88</u>
	Doc2Vec	0.39	0.39	0.44
Random Forest	Bag of Words	0.50	0.49	0.17
	TF-IDF	0.48	0.48	0.16
	Doc2Vec	0.28	0.29	0.37
SGD	Bag of Words	0.85	<u>0.86</u>	0.86
	TF-IDF	<b>0.88</b>	<b>0.88</b>	<u>0.88</u>
	Doc2Vec	0.43	0.42	0.44
SVM	Bag of Words	0.85	0.85	<u>0.88</u>
	TF-IDF	<u>0.86</u>	<u>0.86</u>	0.87
	Doc2Vec	0.32	0.32	0.43
CNN	FastText	0.04	0.05	0.09
LSTM		0.10	0.10	0.12
BERT Multilingual ( <i>fine-tuning</i> )	-	0.84	<b>0.88</b>	<b>0.89</b>
CamemBERT ( <i>fine-tuning</i> )	-	0.82	<u>0.86</u>	<u>0.88</u>

(1) : < 500 articles par classe

(2) : < 1500 articles par classe

(3) : sans échantillonnage

Moncla, L., Chabane, K., & Brenon, A. (2022). Classification automatique d'articles encyclopédiques. *Conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC)*. Blois, France.

# Classification des articles encyclopédiques

## Comparaison des méthodes de classification (f-mesure)

Ensemble de domaines	#	(1)	(2)	(3)	Ensemble de domaines	#	(1)	(2)	(3)
Géographie	2 870	<u>0.98</u>	0.22	<u>0.99</u>	Arts et métiers	132	0.45	0.00	0.51
Droit - Jurisprudence	1 452	<u>0.92</u>	0.39	<u>0.94</u>	Blason	126	<u>0.93</u>	0.00	<u>0.93</u>
Métiers	1 220	<u>0.87</u>	0.07	<u>0.89</u>	Chasse	124	<u>0.92</u>	0.01	<u>0.92</u>
Histoire naturelle	1 130	<u>0.92</u>	0.06	<u>0.95</u>	Maréchage [...]	118	<u>0.90</u>	0.00	<u>0.88</u>
Histoire	726	0.76	0.08	<u>0.80</u>	Chimie	115	0.75	0.02	0.72
Grammaire	575	0.77	0.08	<u>0.81</u>	Philosophie	115	0.75	0.01	0.69
Médecine [...]	535	<u>0.87</u>	0.07	<u>0.87</u>	Beaux-arts	103	<u>0.86</u>	0.00	<u>0.84</u>
Marine	454	<u>0.93</u>	0.03	<u>0.94</u>	Monnaie	74	<u>0.81</u>	0.00	0.79
Commerce	437	<u>0.85</u>	0.04	<u>0.85</u>	Pharmacie	75	0.65	0.00	0.58
Religion	389	<u>0.89</u>	0.02	<u>0.90</u>	Jeu	67	<u>0.85</u>	0.00	<u>0.87</u>
Architecture	326	<u>0.88</u>	0.01	<u>0.88</u>	Pêche	48	<u>0.93</u>	0.00	<u>0.90</u>
Antiquité	321	<u>0.80</u>	0.01	<u>0.82</u>	Mesure	43	0.65	0.00	0.74
Physique	309	<u>0.85</u>	0.04	<u>0.86</u>	Economie domestique	31	0.75	0.00	0.58
Militaire [...]	304	<u>0.92</u>	0.01	<u>0.92</u>	Médailles	28	0.84	0.00	0.79
Agriculture [...]	259	<u>0.80</u>	0.04	<u>0.80</u>	Caractères	27	0.67	0.00	0.51
Belles-lettres - Poésie	246	0.75	0.01	0.74	Politique	27	0.31	0.00	0.00
Anatomie	245	<u>0.92</u>	0.02	<u>0.91</u>	Minéralogie	26	0.68	0.00	0.65
Mathématiques	164	<u>0.88</u>	0.00	<u>0.89</u>	Superstition	26	<u>0.81</u>	0.00	0.73
Musique	163	<u>0.94</u>	0.01	<u>0.94</u>	Spectacle	11	0.17	0.00	0.00

(1) : TF-IDF + SGD

(2) : FastText + LSTM

(3) : BERT

Moncla, L., Chabane, K., & Brenon, A. (2022). Classification automatique d'articles encyclopédiques. *Conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC)*. Blois, France.

# SOMMAIRE

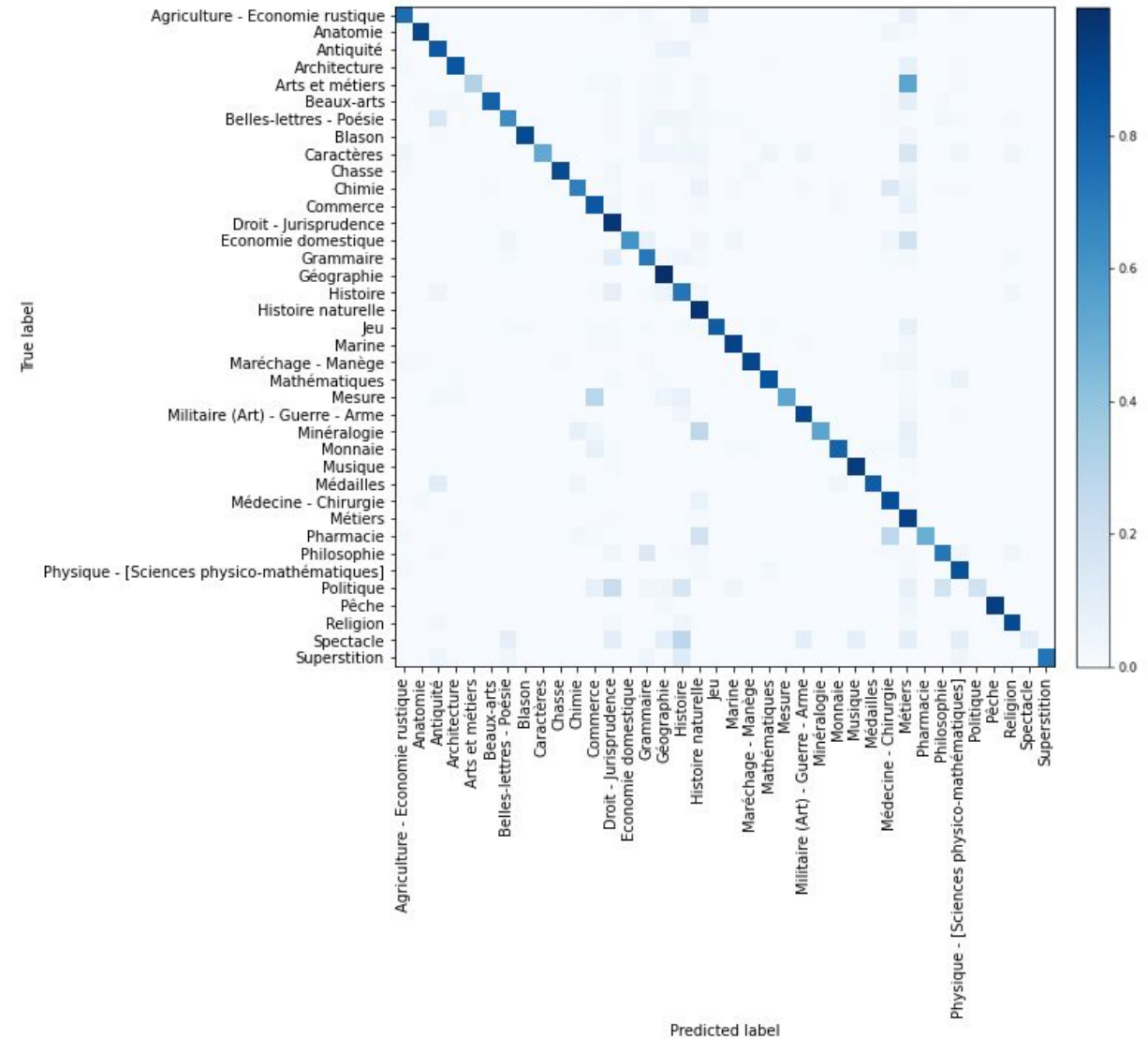
1. Introduction
2. Classification des articles encyclopédiques
  - 2.1. Problématique
  - 2.2. Expérimentations
  - 2.3. Discussion & Perspectives
3. Reconnaissance et classification des entités nommées
  - 3.1. Indices linguistiques forts
  - 3.2. Indices linguistiques faibles
  - 3.3. Expérimentations et évaluation

# Classification des articles encyclopédiques

## Discussion

### Erreurs de classification ?

- Représentation faible
  - Spectacle (88)
  - Politique (293)
- Proximité lexicale / sémantique
  - Arts et Métier -> Métiers : 0.55
  - Pharmacie -> Médecine/Chirurgie : 0.28
  - Mesure -> Commerce : 0.29

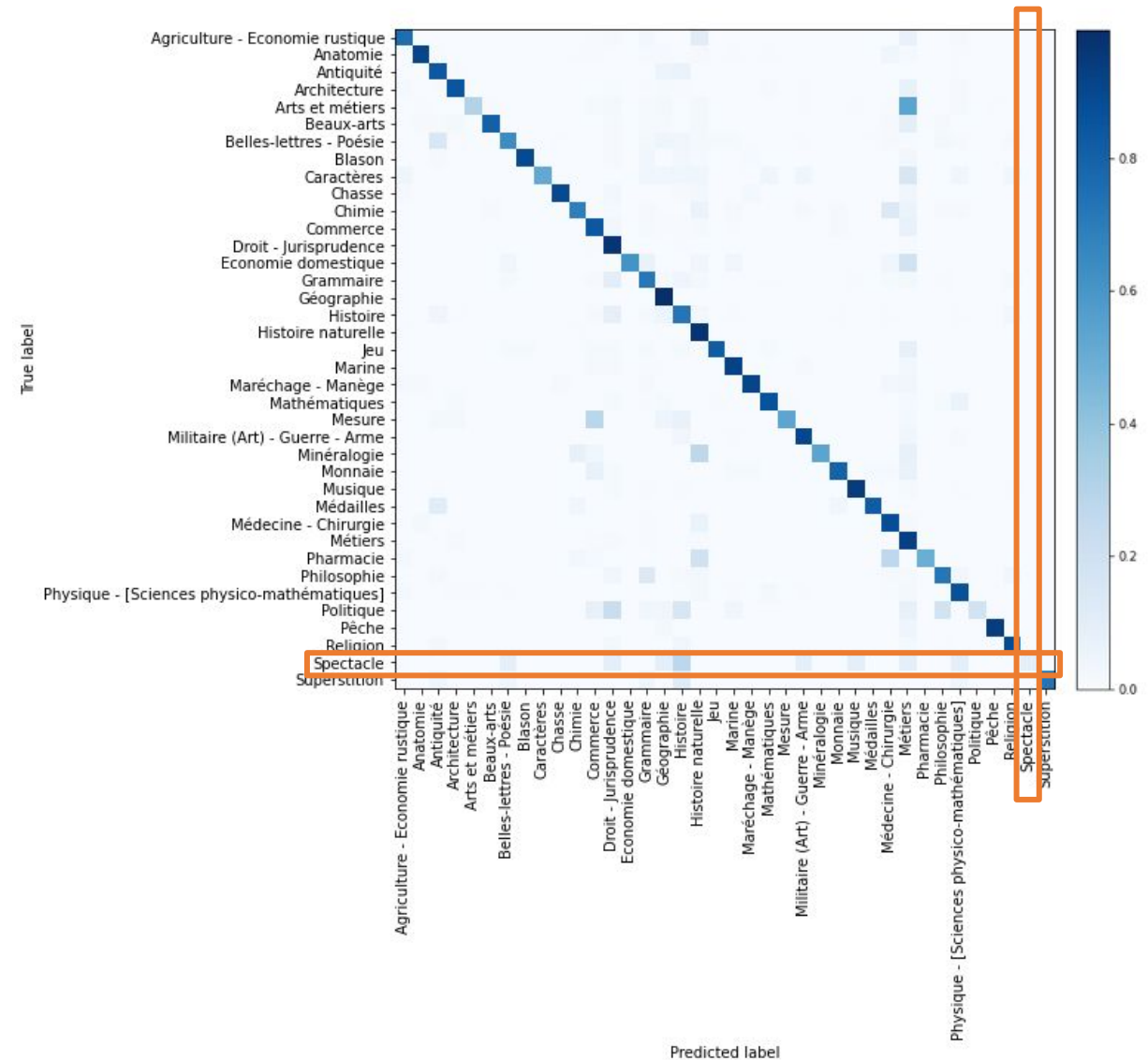


# Classification des articles encyclopédiques

## Discussion

### Erreurs de classification ?

- Représentation faible
  - Spectacle (88)
  - Politique (293)
- Proximité lexicale / sémantique
  - Arts et Métier -> Métiers : 0.55
  - Pharmacie -> Médecine/Chirurgie : 0.28
  - Mesure -> Commerce : 0.29



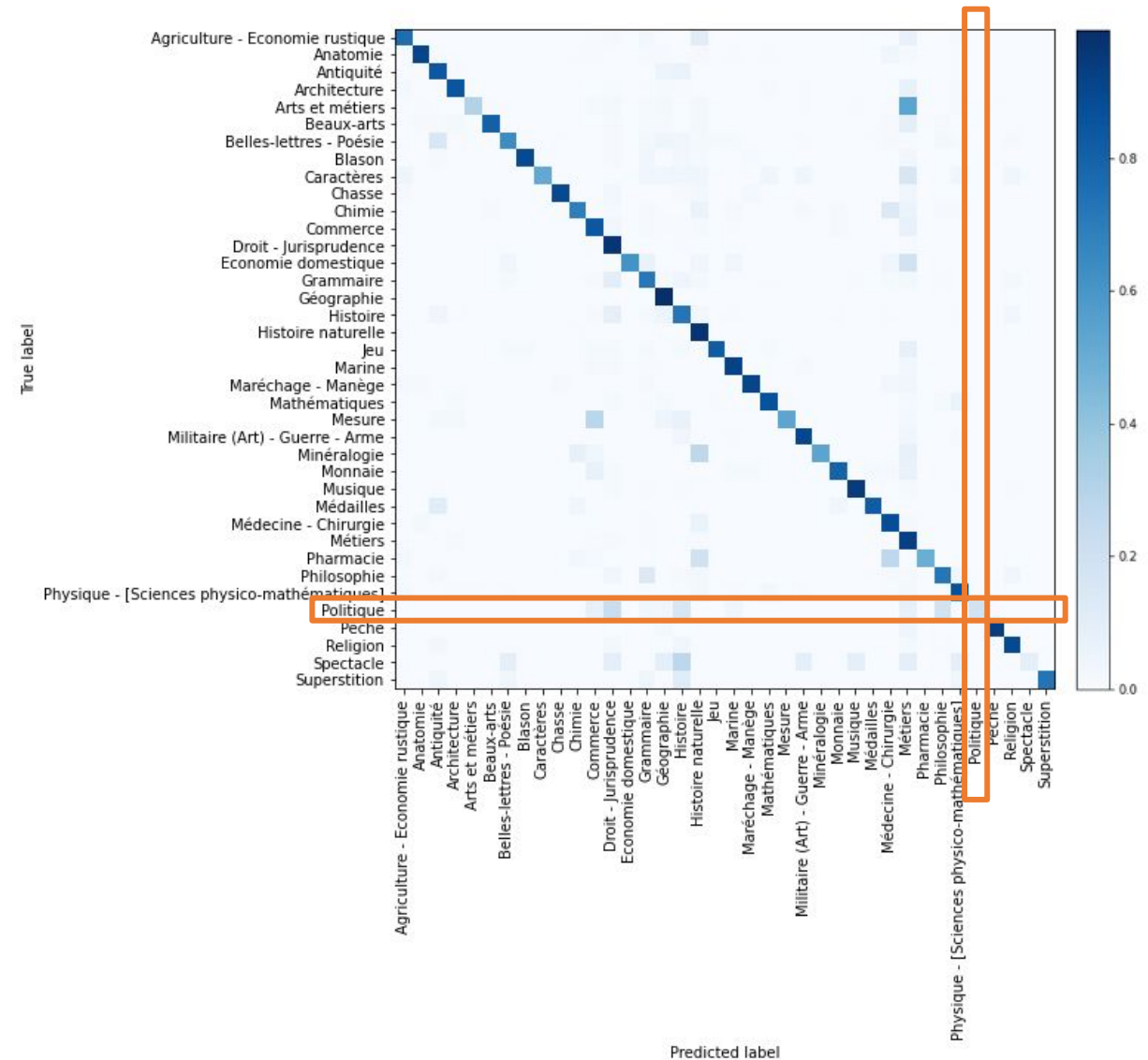


# Classification des articles encyclopédiques

## Discussion

### Erreurs de classification ?

- Représentation faible
  - Spectacle (88)
  - Politique (293)
- Proximité lexicale / sémantique
  - Arts et Métier -> Métiers : 0.55
  - Pharmacie -> Médecine/Chirurgie : 0.28
  - Mesure -> Commerce : 0.29

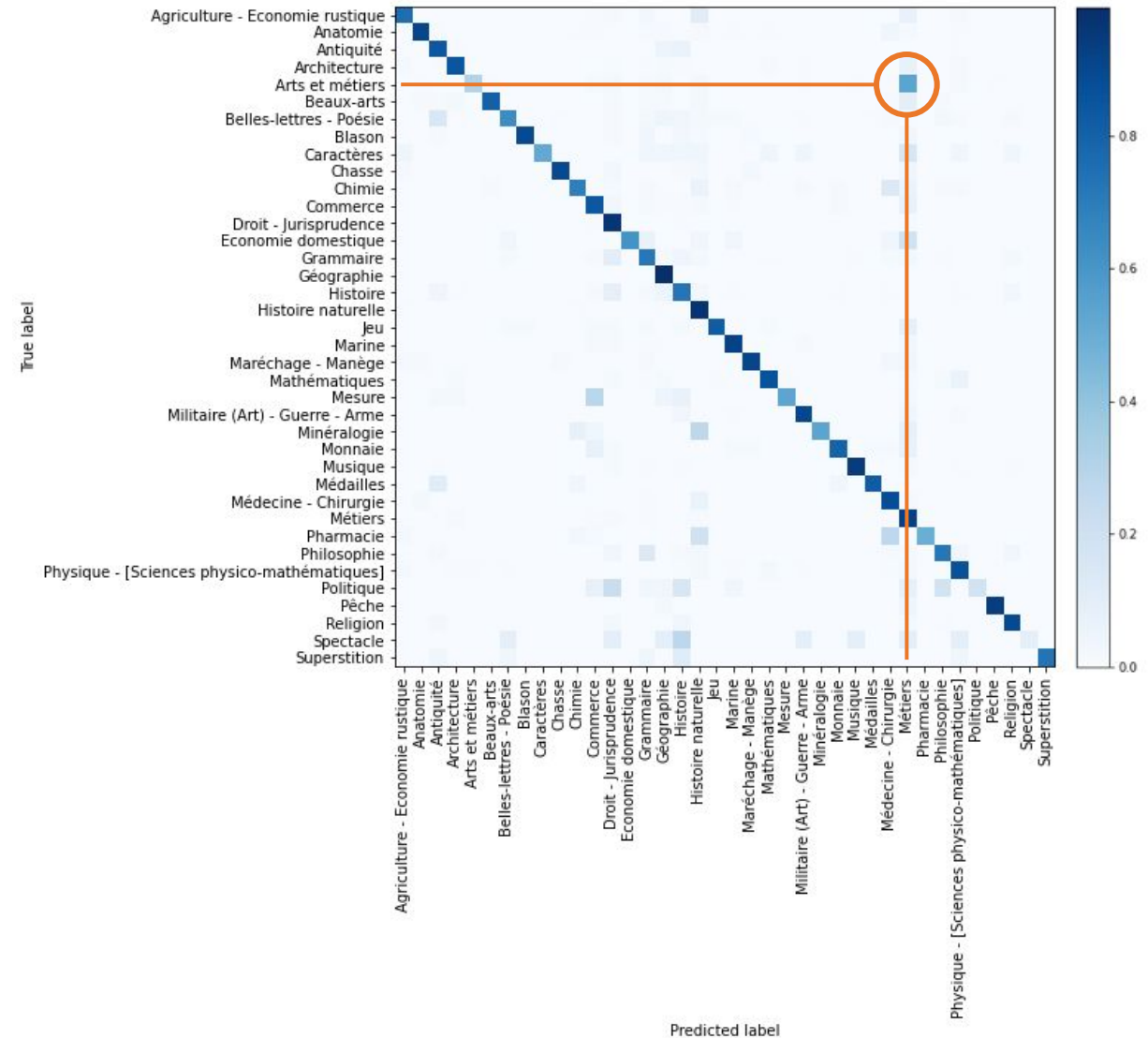


# Classification des articles encyclopédiques

## Discussion

### Erreurs de classification ?

- Représentation faible
  - Spectacle (88)
  - Politique (293)
- Proximité lexicale / sémantique
  - Arts et Métier -> Métiers : 0.55
  - Pharmacie -> Médecine/Chirurgie : 0.28
  - Mesure -> Commerce : 0.29



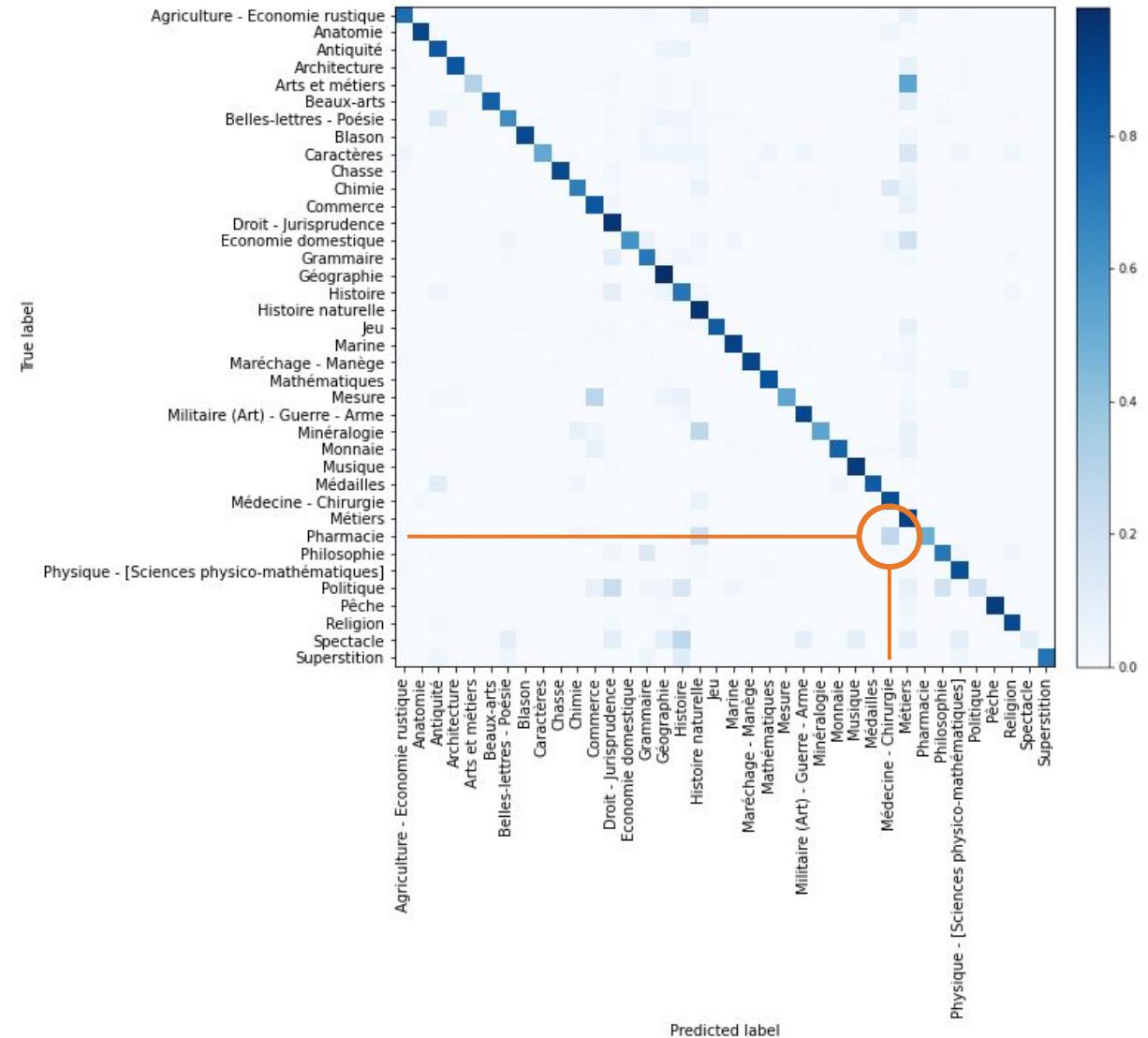


# Classification des articles encyclopédiques

## Discussion

### Erreurs de classification ?

- Représentation faible
  - Spectacle (88)
  - Politique (293)
- Proximité lexicale / sémantique
  - Arts et Métier -> Métiers : 0.55
  - Pharmacie -> Médecine/Chirurgie : 0.28
  - Mesure -> Commerce : 0.29

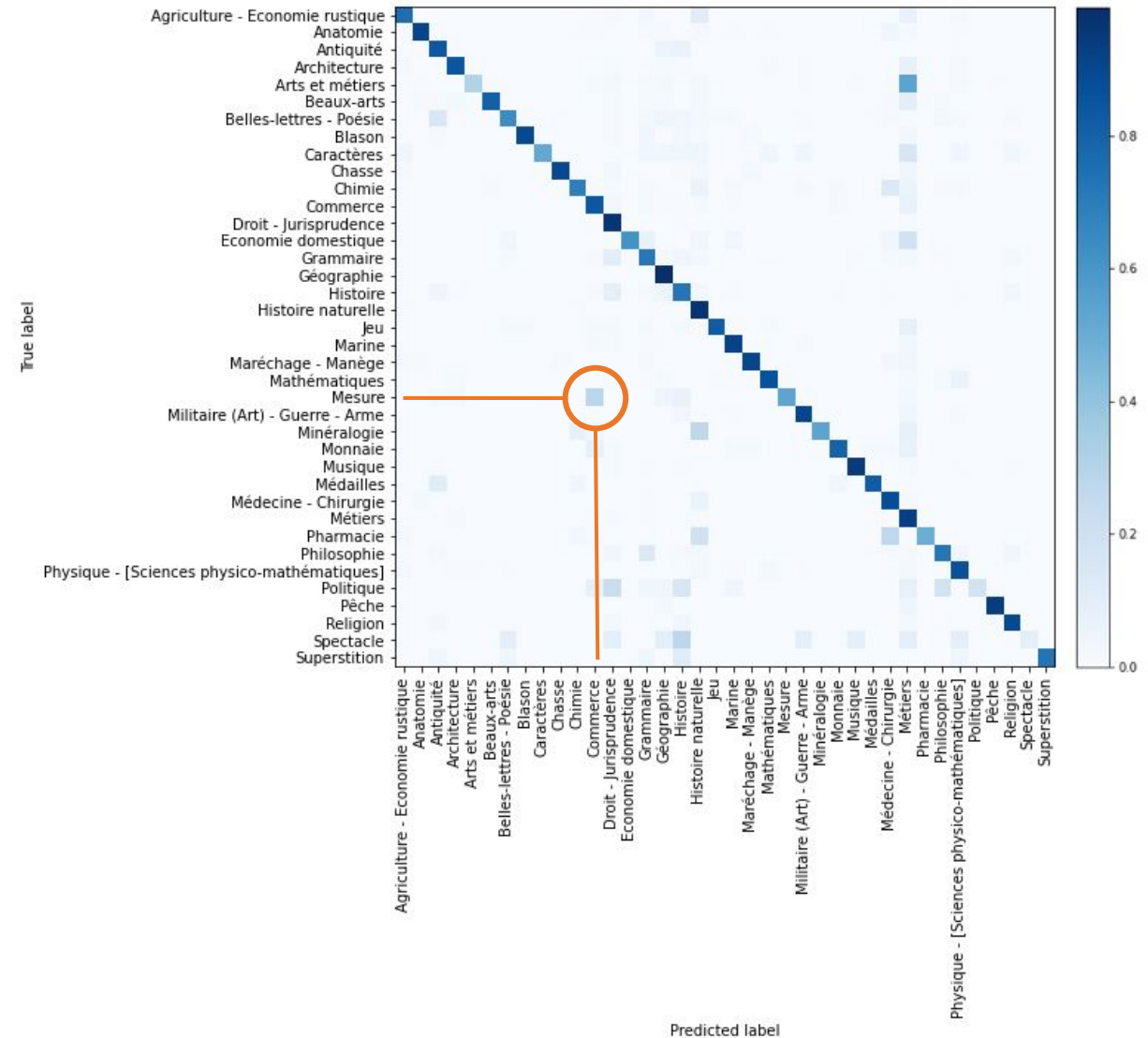


# Classification des articles encyclopédiques

## Discussion

### Erreurs de classification ?

- Représentation faible
  - Spectacle (88)
  - Politique (293)
- Proximité lexicale / sémantique
  - Arts et Métier -> Métiers : 0.55
  - Pharmacie -> Médecine/Chirurgie : 0.28
  - Mesure -> Commerce : 0.29



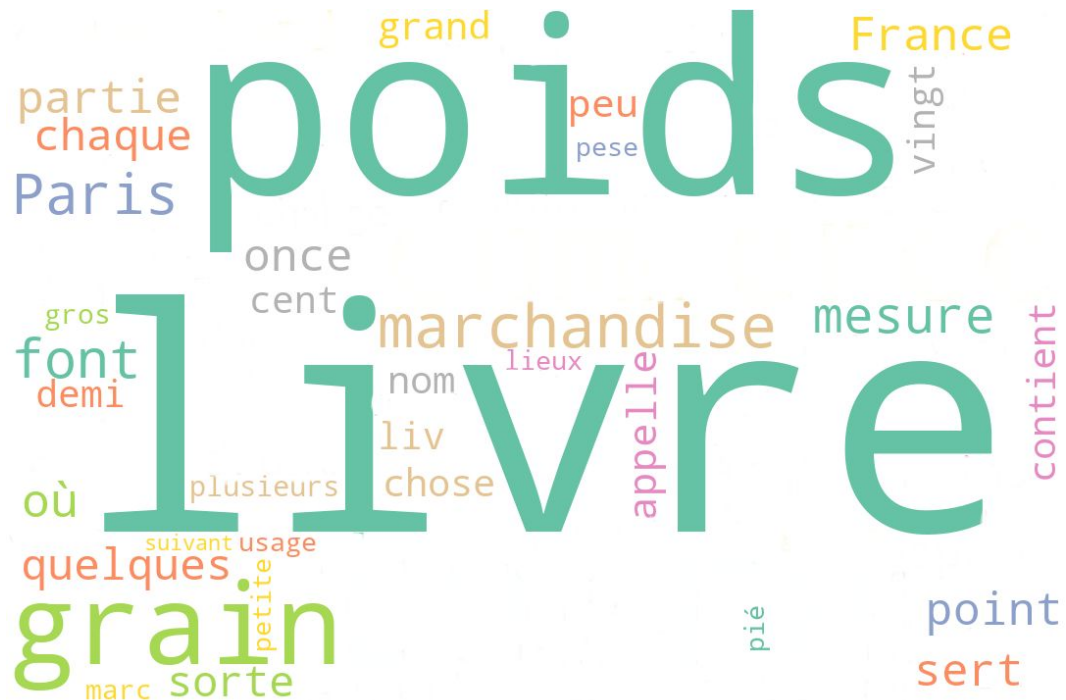


# Classification des articles encyclopédiques

Discussion : proximité lexicale

## Commerce ou mesure ?

- 35 “mots” en commun parmi les 100 plus fréquents



Commerce



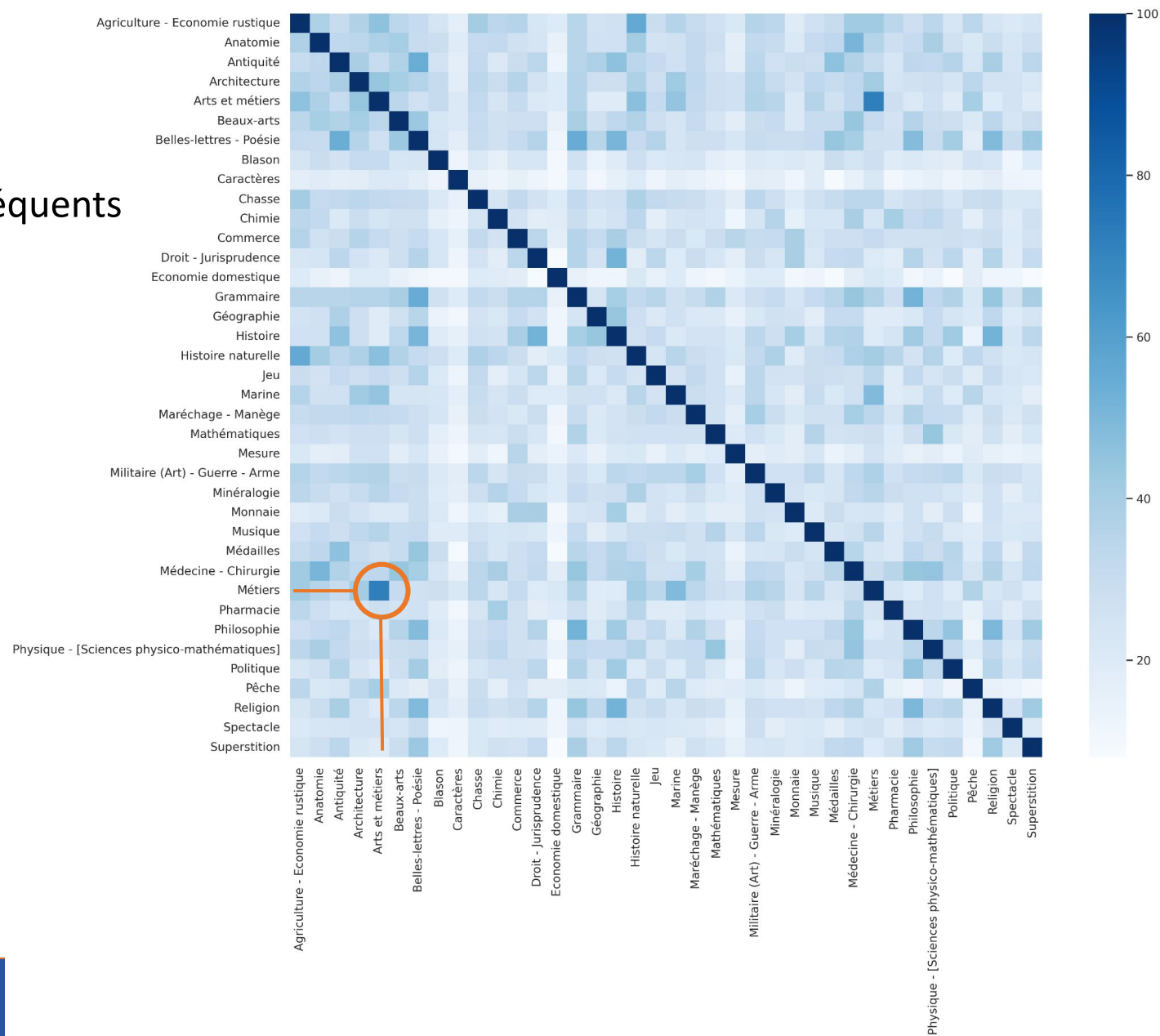
Mesure



# Classification des articles encyclopédiques

## Discussion : proximité lexicale

- Mots en commun parmi les 100 plus fréquents
- 'Arts et métiers' et 'Métiers' : 72







# SOMMAIRE

1. Introduction
2. Classification des articles encyclopédiques
  - 2.1. Problématique
  - 2.2. Expérimentations
  - 2.3. Discussion & Perspectives
3. Reconnaissance et classification des entités nommées
  - 3.1. Indices linguistiques forts
  - 3.2. Indices linguistiques faibles
  - 3.3. Expérimentations et évaluation



# Reconnaissance et classification des entités nommées

## Approche symbolique

### Méthodologie

- Recherche d'« indices » linguistiques dans un corpus annoté pour la détection et la classification des EN au moyen de la plateforme **TXM**
- Restriction au **sous-corpus** des articles de « géographie » dans EDdA (20% du total des articles)
- Implémentation de règles au moyen de transducteurs **Unitex** dans PERDIDO
- Evaluation des performances et feed-back

Moncla, L., Vigier, D., McDonough, K., Brenon, A., & Joliveau, T. (2021). Combinaison d'approches qualitative et quantitative pour le repérage et la classification des entités nommées dans l'Encyclopédie de Diderot et d'Alembert (1751-1772). *International Symposium of Theoretical linguistics in the light of the interaction of qualitative and quantitative approaches*. Neuchâtel, Switzerland.

Vigier, D., Moncla, L., Brenon, A., McDonough, K., & Joliveau, T. (2020). Classification des entités nommées dans l'Encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers par une société de gens de lettres (1751-1772). *7e Congrès Mondial de Linguistique Française (CMLF)*. Montpellier, France.

# Les indices linguistiques forts (ILFo)

Détection et classification des EN vedettes dans EDdA

- Une décision peut être prise à partir d'eux avec une probabilité proche de 100%
- Format « type » des articles dans EDdA :

[VEDETTE] [ponctuation\*] [(nom de domaine placé entre parenthèses)\*]  
[ponctuation\*] [texte] [coordonnées géographiques \*][signature(s) de l'auteur/des auteurs\*]

- Exemple : article n° 3001 volume 17

**ZONZEN**

ZONZEN, (*Géog. mod.*) ville de Perse dans la province de Mazanderan. *Long. 85. 15. latit. 35. 59. (D. J.)*

# Les indices linguistiques forts (ILFo)

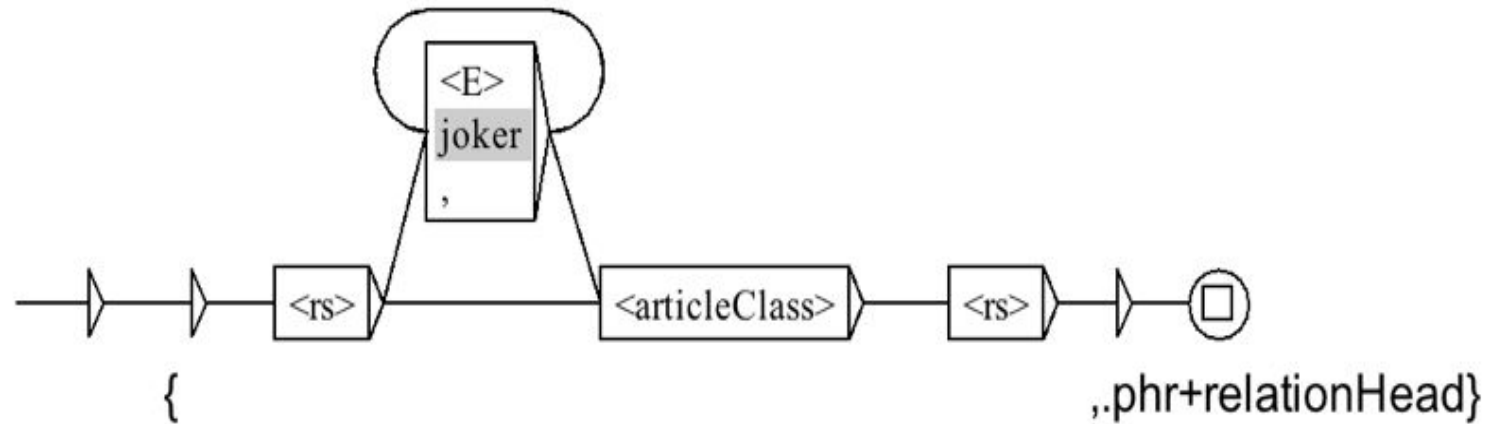
## Les « noms classifieurs »

- Dans 9 082 articles (soit 95,5%), un nom classifieur apparaît parmi les trois premiers mots qui suivent la vedette qui permet de classer [ENG/ENH] et sous-classer l'EN
- Exemples :
  - HAINGEN, (*Géogr.*) petite ville d'Allemagne (...) (VIII, 26)
  - ZERGUE, (*Géog. mod.*) petite riviere de France (...) (XVII, 706)
  - VINDERIUS, (*Géogr. anc.*) fleuve de l'Hibernie. (...) (XVII, 307)
- Sous-classes d'ENG :
  - *ville (5696), rivière (753), île (521), province (346), fleuve (206), pays (202), bourg (194), royaume (194), montagne (190), lieu (122), comté (90), port (85), village (66), lac (65), contrée (60), bourgade (55), promontoire (55), golfe (51), forteresse (47), duché (38), château (34), canton (29), cap (28), place (24), capitale (22), palatinat (20), vallée (18), forêt (14), principauté (14), fontaine (11), chaîne (10), maison (10)*

# Les indices linguistiques forts (ILFo)

## Implémentation

- Transducteur implémenté pour le repérage et l'annotation du motif 1



# SOMMAIRE

1. Introduction
2. Classification des articles encyclopédiques
  - 2.1. Problématique
  - 2.2. Expérimentations
  - 2.3. Discussion & Perspectives
3. Reconnaissance et classification des entités nommées
  - 3.1. Indices linguistiques forts
  - 3.2. Indices linguistiques faibles
  - 3.3. Expérimentations et évaluation

# Les indices linguistiques faibles (ILFa)

## Détection et classification des entités nommées dans EDdA

- Indices non-décisifs
  - les ILFa seuls ne permettent pas de prendre une décision avec une certitude de 100%
- Nous leur associons un degré de confiance (faible, moyen, fort) fondé sur :
  - les performances des indices (cf *infra*)
  - la combinaison de plusieurs indices lorsque le cas se produit

# Les indices linguistiques faibles (ILFa)

Détection et classification des entités nommées dans EDdA

Élaboration de règles testées au moyen du concordancier sur TXM

- Analyse des **cotextes préférés d'apparition** des NPr désignants des ENG / ENH ( $f > 100$ ) dans le sous-corpus Géographie

## Objectif

- Identifier en amont / aval des séquences préférées pour l'occurrence de ces NPr.
  - Recours au calcul des spécificités sur TXM (<https://txm.gitpages.huma-num.fr/textometrie/index.html>)

# Les indices linguistiques faibles (ILFa)



## Détection et classification des entités nommées dans EDdA

- Ces cotextes ont permis d'élaborer 34 règles
  - 9 règles qui mettent en jeu la préposition **de**,
  - 8 la préposition **à**,
  - 8 règles la préposition **dans**,
  - 5 règles la préposition **par**,
  - La préposition **en** et le couple **chez/parmi** donnant lieu chacun à 1 règle,
  - La préposition **sur** et le couple **chez/parmi** donnant lieu ensemble à 2 règles.



# Les indices linguistiques faibles (ILFa)

## Illustration de deux règles

- Lemmes nominaux « *confluent* / *embouchure* / *source* » +  
« *du* / *de la* / *de l'* » + nom propre  hydronyme  
[100% réussite]
- Lemmes verbaux « *achever* / *bâtir* / *chasser* / *découvrir* /  
*décrire* / *détruire* / *donner* / *ériger* / *fonder* / *fortifier* /  
*habiter* / *nommer* / *occuper* / *placer* / *posséder* / *rapporter*  
/ *ruiner* / *subjuguier* / *tuer* / *vanter* / *vaincre* » + *par* +  
article défini *les* + nom propre  nom de peuple  
[96% réussite]

# Les indices linguistiques faibles (ILFa)

Classification des EN à partir des IFLo et ILFa

## Remarques générales


- Un indice ne vaut que pour l'occurrence de l'EN auquel il est attaché
- Un ILFa ne peut être combiné à une ILFo mais peut l'être avec un autre indice affecté par PERDIDO.
  - Si les deux indices sont convergents, ils se renforcent. S'ils divergent, les degrés de confiance de l'ILFa et de l'annotation PERDIDO sont pris en compte

# Les indices linguistiques faibles (ILFa)

Classification des EN à partir des IFLo et ILFa

- Chaque occurrence de Np n'est pas forcément associée à un indice après application des règles.

Exemple : nom propre *Paris*

- 6 892 occ. dans EDdA
- 6 101 occ. non-classées ( $\approx 90\%$ )
- 761 occ. classées ENG
- 30 occ. classées ENH  classées grâce aux ILFa

# Les indices linguistiques faibles (ILFa)

## Méthode ILFa+

### Extension des ILFa pour la classification des occurrences non classées

- Appliquée aux EN dont un sous-ensemble d'occurrences se sont vu assigner un ILFa
- Catégorie majoritaire (ex. ENG) étendue à toutes les occ. non-classées
- Evaluation nécessaire de cette approche + degré de confiance à affecter
  - Le degré de confiance peut être fonction du nombre de classes et de leur proportion

# Les indices linguistiques faibles (ILFa)

## Implémentation

- Les règles ont été implémentées dans l'outil PERDIDO afin d'améliorer la classification des EN.
- Cet outil se compose de deux cascade de transducteurs (analyse et synthèse) permettant le repérage de motifs et l'ajout d'informations sémantiques.
- 2 phases :
  - ILFa : annotés au sein de la cascade d'analyse
  - ILFa+ : annotés en post-traitement (nécessite une vue globale sur l'ensemble du corpus)

# SOMMAIRE

1. Introduction
2. Classification des articles encyclopédiques
  - 2.1. Problématique
  - 2.2. Expérimentations
  - 2.3. Discussion & Perspectives
3. Reconnaissance et classification des entités nommées
  - 3.1. Indices linguistiques forts
  - 3.2. Indices linguistiques faibles
  - 3.3. Expérimentations et évaluation

# Premières évaluations

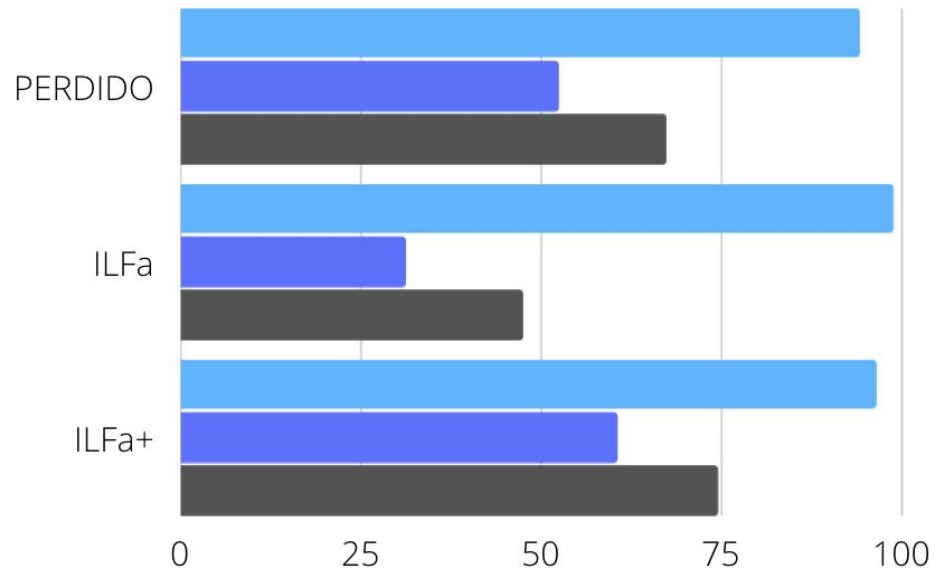
Nombre d'occurrences repérées par les ILFa et ventilées par macro et sous-catégories

Classification	Effectifs	Classification	Effectifs	Classification	Effectifs
ENG	66 854	toponyme	62 123	¬ ville	31498
				ville	116
		hydronyme	3 571	classifieur	882
ENH	19 922	non classé	1 160	non classé	29627
		individu	14 957		
		collectivité	3 287		
non classé	453 871	non classé	1 678		

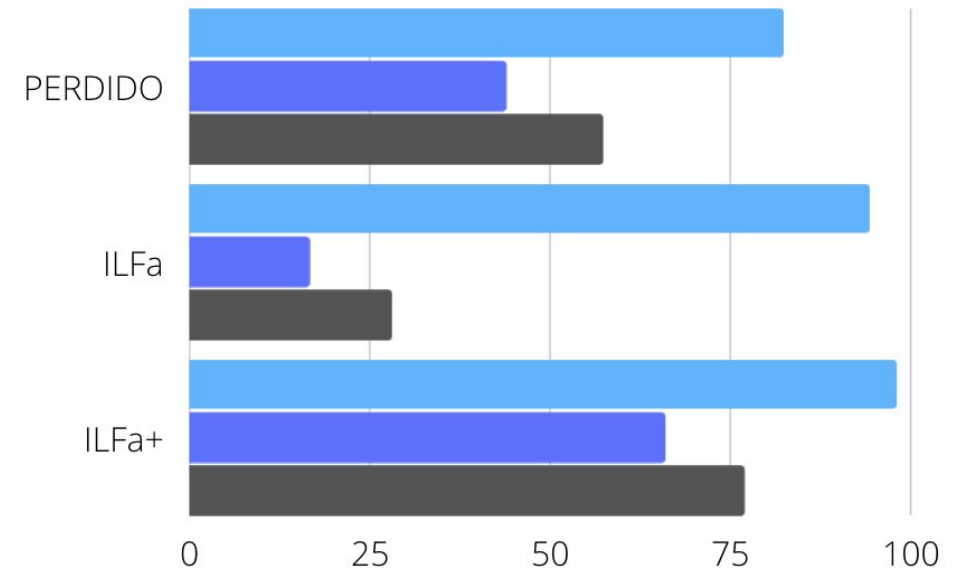
# Premières évaluations

## Évaluation automatique

ENG



ENH



Précision Rappel F-mesure



# Premières évaluations

## Évaluation manuelle de la précision

Répartition du nombre d'erreurs de classification répartis par ILFa (TP = vrais positifs)

Règle	Erreurs	TP	% erreur	Règle	Erreurs	TP	% erreurs
1	212	802	26,43	12	77	2 545	3,03
7	231	1 063	21,73	8	4	132	3,03
26	85	480	17,98	34	62	2 160	2,87
5	170	1 031	16,49	17	4	154	2,60
2	7	45	15,56	28	19	843	2,25
4	99	675	14,67	11	26	1 211	2,15
20	39	320	12,19	33	21	1 066	1,97
24	38	327	11,62	19	8	443	1,81
21	275	3 056	9,00	22	6	563	1,07
16	3	37	8,11	3	16	1 571	1,02
23	15	214	7,01	27	2	235	0,85
32	801	13 326	6,01	25	4	766	0,52
9	1 189	22 283	5,34	14	2	498	0,40
29	210	4 334	4,85	31	10	2 876	0,35
15	63	1 504	4,19	13	0	607	0,00
6	12	298	4,03	18	0	31	0,00
30	1	26	3,85	10	0	103	0,00

# Conclusion

## Perspectives

- Amélioration des approches et méthodes présentées
- Annotation automatique des entités nommées étendues
  - expérimentation de méthodes d'apprentissage profond (comparaison)
- Repérage et interprétation des informations géographiques
- Extension de ces travaux aux autres encyclopédies de notre corpus

Merci pour votre attention

### Contact

Ludovic Moncla

[ludovic.moncla@insa-lyon.fr](mailto:ludovic.moncla@insa-lyon.fr)

Denis Vigier

[denis.vigier@ens-lyon.fr](mailto:denis.vigier@ens-lyon.fr)

Les auteurs remercient le LABEX ASLAN (ANR-10-LABX-0081) de l'Université de Lyon pour son soutien financier dans le cadre du programme français "Investissements d'Avenir" géré par l'Agence Nationale de la Recherche (ANR).



# Les indices linguistiques forts (ILFo)

## Identification des articles *via* la plateforme TXM

Requête :  Pivot: word

Clés de tri : #1  #2  #3  #4

|< < 1 - 100 / 11902 > >|

text_id	Contexte gauche	Pivot	Contexte droit
volume01-1643		ALENCON, (Géog.)	ville de France dans la basse Normandie sur la Sarthe, grossie par la Briante. Lon. 17. 45. lat 48. 25.
volume01-1645		ALENTAKIE (Géog.)	Province de l'Esthonie, sur le Golfe de Finlande.
volume01-1646		ALENTÉJO, (Géog.)	Province de Portugal, située entre le Tage et la Guadiana.
volume01-1648		ALEP, (Géog.)	grande ville de Syrie, en Asie, sur le ruisseau Marsgras ou Coié. Long. 55. lat. 35. 50.
volume01-1662		ALESSIS (Géog.)	ville d'Albanie dans la Turquie Européenne, proche l'embouchure du Drin. Long. 37. 15. lat. 41. 48.
volume01-1667		ALEXANDRETTE (Géog.)	ville de Syrie en Asie, à l'extrémité de la mer méditerranée, à l'embouchure d'un petit ruisseau appelé Belum ou Soldrat, sur le golfe d'
volume01-1764		ALLASSAC, (Géog.)	ville de France, dans le Limosin et la Généralité de Limoges.
volume01-1782		ALLEGANIA, (Géog.)	petite isle d'Afrique, l'une des Canaries, au nord de la Gracieuse, au nord - ouest de Rocca, et au nord - est de Sainte
volume01-1933		ALPUXARRAS, (Géog.)	hautes montagnes d'Espagne dans le Royaume de Grenade au bord de la Méditerranée.
volume01-2409		ANAB, (Géog. anc.)	montagne dans la Tribu de Juda, au pied de laquelle il y avoit une ville du même nom, entre Dabet et Istamo. V. Jos. xj
volume01-2410		ANABAGATHA, (Géog. anc.)	ancienne ville d'Asie, sous le Patriarchat d'Antioche. Voyez Aubert le Mire, in Géog. eccles. not.
volume01-2411		ANABAO, (Géog. mod.)	une des Îles Moluques, au sud - ouest de Timor. Anabao et Timor sont séparées par un canal qui peut recevoir tous les vaisseaux. Il y a
volume01-2426		ANACHIMOUSSE, s. m. (Géog. mod.)	peuple de l'île Madagascar, dont il occupe la partie méridionale, située au nord de Manamboule.
volume01-2446		ANAFE ou AFFA, (Géog. mod.)	ville de la province de Temesne, au Royaume de Fez en Afrique, sur la côte de l'Océan atlantique. Alfonso Roi de Portugal, la ruina,
volume01-2449		ANAGARSKAIE, (Géog. mod.)	ville des Moscovites de la grande Tartarie, dans la province de Dauria, à l'orient du lac Baycal, aux sources de la rivière d'Amur. Long
volume01-2450		ANAGHELOME, (Géog. mod.)	petite ville d'Irlande, dans la Province d'Ulster ou d'Ultonie, Comté de Dowane, sur le Ban.
volume01-2459		ANAGYRUS, (Géog. et Myth.)	bourg de l'Attique en Grece dans la tribu Erechide. On dérive son nom ou de l'anagyris, plante ; ou d'un Anagyrus, demi - dieu
volume01-2460		ANAHARATH, (Géog. anc.)	ville de la tribu d'Issachar, dont il est fait mention dans Josué xix. 19.
volume01-2483		ANAN ou ANNAND (Géog. mod.)	fleuve d'Écosse, dans sa partie méridionale, province d'Anandal ; il prend sa source près du Cluid et se décharge dans un golfe de la mer
volume01-2485		ANANDAL (Géog. mod.)	Province de l'Ecosse méridionale, entre la contrée d'Eskédale au couchant, et celle de Nitheisdale à l'orient.
volume01-2488		ANAPE, s. m. (Géog. et Myth.)	aujourd'hui l'Alfeo, fleuve de Sicile, près de Syracuse ; les Poètes l'ont fait amoureux de Cyané, et Protecteur de Proserpine, contre l'at
volume01-2490		ANAPHE, s. f. (Géog. et Myth.)	île de la mer Egée qu'on dit s'être formée insensiblement comme Delos, Hiera, et Rhodes. C'est du culte particulier qu'on y rendoit à
volume01-2516		ANATORIA, (Géog.)	petite ville de Grece, anciennement Tanagra. Voyez Tanagra.
volume01-2519		ANAZZO ou TORRE - D'ANAZZO, (Géog. mod.)	ville de la province de Bari au royaume de Naples. On croit que c'est l'ancienne Egnatia ou Gnatia. Quelques Modernes la nomment Gnaz
volume01-2520		ANBAR, (Géog. mod.)	ville de la province de Chaldée ou Iraque Arabique, sur l'Euphrate. Elle s'est appelée Haschemiah.
volume01-2523		ANCAMARES ou ANTAMARES, (Géog. mod.)	peuples de l'Amérique méridionale, qui habitent le long du fleuve Madere, qui se perd dans la rivière des Amazones.
volume01-2525		ANCARANO, (Géog. mod.)	petite ville de l'Etat ecclésiastique dans la Marche d'Ancone.
volume01-2527		ANCENIS, (Géog. mod.)	ville de France dans la Bretagne sur la Loire. Long. 16. 28. lat. 47. 22.

# Les indices linguistiques forts (ILFo)

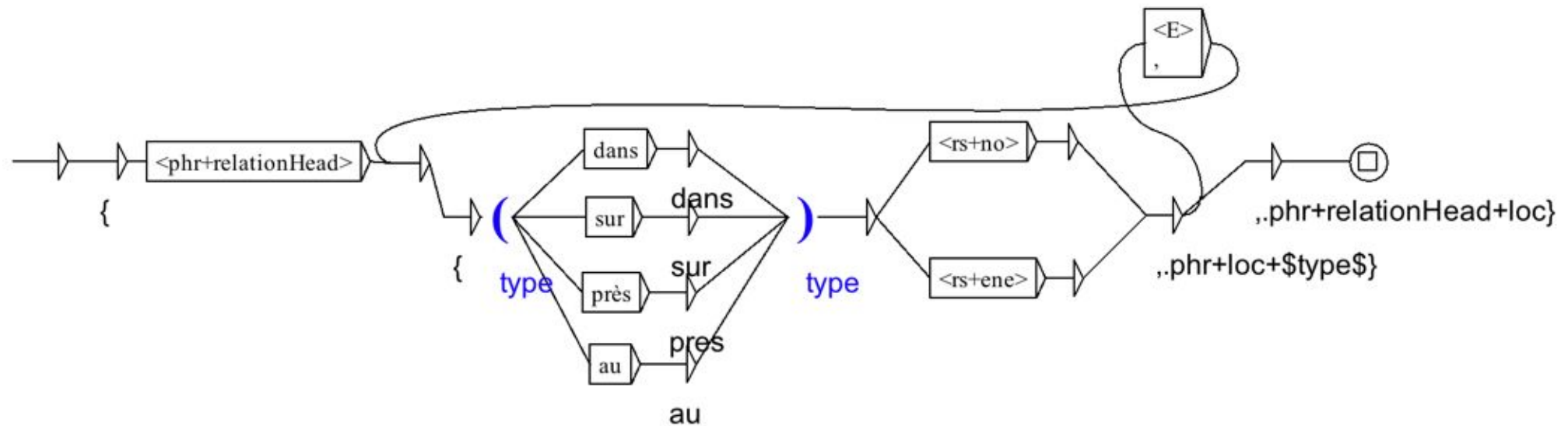
## Entités Nommées Humaines (ENH)

- 406 articles soit 4,5%
- Exemples :
  - TZCHALATZKI les, & les TZUKTZCHI, (*Géog. mod.*) nom de deux peuples barbares & alliés (...) (XVI, 788)
  - BARBETS, s. m. pl. (*Géog.*) habitants des vallées du piémont (...) (II, 73)
  - OPHIOGENES les, (*Géog. anc.*) race particuliere d'hommes (...) (XI, 502)
- Sous-classes d'ENH :
  - *peuple, communauté, habitant, race, peuplade*

# Les indices linguistiques forts (ILFo)

## Implémentation

- Transducteur implémenté pour le repérage et l'annotation des motifs 2 à 5





# Les indices linguistiques forts (ILFo)

## Les motifs

D 'autres indices forts : le cas des ENG

Motifs	Noms	% ss-corpus Géo
N1 [classifieur] de N2	motif 1	52,9 %
N1 de N2 (,) Prép SN	motif 2	24,9 %
N1 de N2 (,) Prép SN (,) Prép SN	motif 3	9,2 %
N1 de N2 (,) Prép SN (,) Prép SN (,) Prép SN	motif 4	2,3 %
N1 de N2 (,) Prép SN (,) Prép SN (,) Prép SN (,) Prép SN	motif 5	0,5 %

Exemple : motif 5

DENAT, (*Géog. mod.*) [petite **ville** de France][au diocèse d'Alby] [dans le Languedoc], [sur l'Assore], [à trois lieues d'Alby.] (IV, 824)

# Les indices linguistiques forts (ILFo)

## Les motifs

### Pour les ENH

- Les motifs sont moins étendus : max. motif 3
  - Motif 1 : 2,6 % du ss-corpus Géographie
  - Motif 2 : 0,6 %
  - Motif 3 : 0,2 %

### Exemple : motif 3

OZAGES, (*Géog.*) [**peuple** de l'Amérique septentrionale] [dans la Louisiane], [au couchant du fleuve Mississippi]. (XI, 730b)



# Les indices linguistiques forts (ILFo)

## Implémentation

### Annotation PERDIDO

- Format XML-TEI
- Ajout de règles (transducteurs)

### Résultat avant implémentation des indices

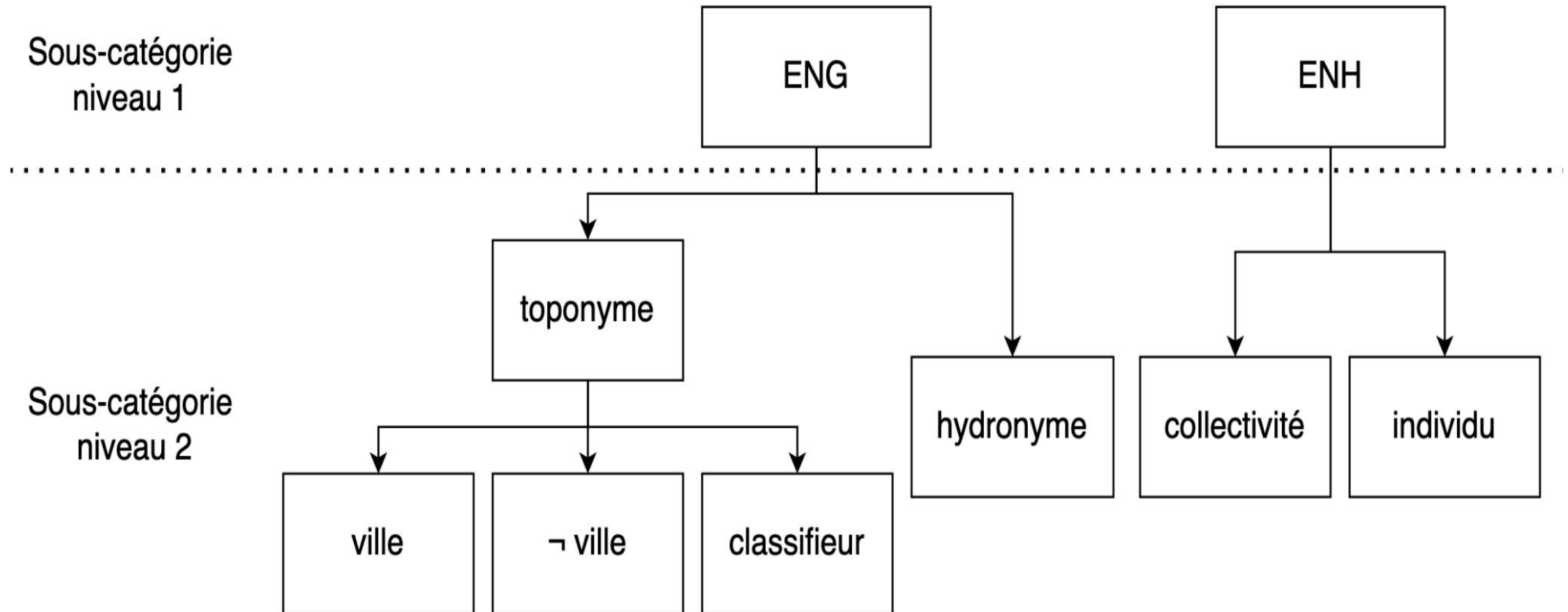
#### **CANGOXUMA**

CANGOXUMA, (*Géog.*) ville d'Asie de l'empire du Japon, dans l'île de Ximo, au royaume de Bungo.

```
<s>
  <rs type="place">
    <name type="place" subtype="edda">CANGOXUMA</name>
  </rs> ,
  <term type="articleClass">(Géog.)</term>
  <rs type="place">
    <rs type="place">
      <term type="place">ville d'</term>
      <rs type="unknown">
        <name type="unknown">Asie</name>
      </rs>
    </rs>
  </rs>
  <rs type="unknown">
    <term type="unknown">de l'empire du</term>
    <rs type="place">
      <name type="place" subtype="edda">Japon</name>
    </rs>
  </rs>
  </rs> , dans
  <rs type="place">
    <term type="place">l'île de</term>
    <name type="unknown">Ximo</name>
  </rs> , au
  <rs type="place">
    <term type="place">royaume de </term>
    <rs type="place">
      <name type="place" subtype="edda">Bungo</name>
    </rs>
  </rs> .
</s>
```

# Les indices linguistiques faibles (ILFa)

## Taxonomie des sous-catégories d'EN



# Les indices linguistiques faibles (ILFa)

Contexte gauche	Pivot
Albanie dans la Turquie Européenne, proche l'	embouchure du Drin
d'Espagne dans le Royaume de Grenade au	bord de la Méditerranée
vers les monts S. Pierre et la	source du Buria
Bentheim et de Tecklembourg ; ou sur les	bords de la Sala
, séparées par une petite rivière proche les	bords de la Forth
même des ruines qu'on rencontreroit sur les	bords du Nil
de Bievre, à une lieue de l'	embouchure du Rhone
dans le Brésil. Ils habitent à la	source du Ganabara
.) royaume maritime des Indes proche l'	embouchure du Gange
Géog.) vallée des Pyrénées à la	source de la Garonne
, dans l'Amérique méridionale, à la	source du Xanxa
Andalousie, sur le côté oriental de l'	embouchure de la Guadiana
, dans la petite Tartarie, à l'	embouchure du Don
vis l'île de Corfou, à l'	embouchure de la Calamou
) ville du bas Languedoc, sur le	bord du Rhone
ville d'Egypte, à l'une des	embouchures du Nil
et le val d'Aost, à la	source de la Drance
vallée de Mazara en Sicile, entre la	source du Biccari
et port de la Biscaye, à l'	embouchure du Nervio
dans l'électorat de Mayence, sur le	bord du Rhin
, vis - à - vis de l'	embouchure de la Niera
Espagne, dans la Galice, à l'	embouchure du Minho
dans la principauté de Furstemberg, vers la	source du Danube
est au pied d'une colline sur les	bords du Rhin
dans le royaume de Valence, sur le	bord de la Méditerranée
province de Bretagne au - dessous de l'	embouchure de la Loire
que les Suisses y possèdent, entre les	sources du Rhin

Vue sur une concordance TXM obtenue à partir de la requête CQL :

```
[lemma="confluent|embouchure|source"][word="du|de"][word="la"]?[type="NPr"]
```



# Les indices linguistiques faibles (ILFa)

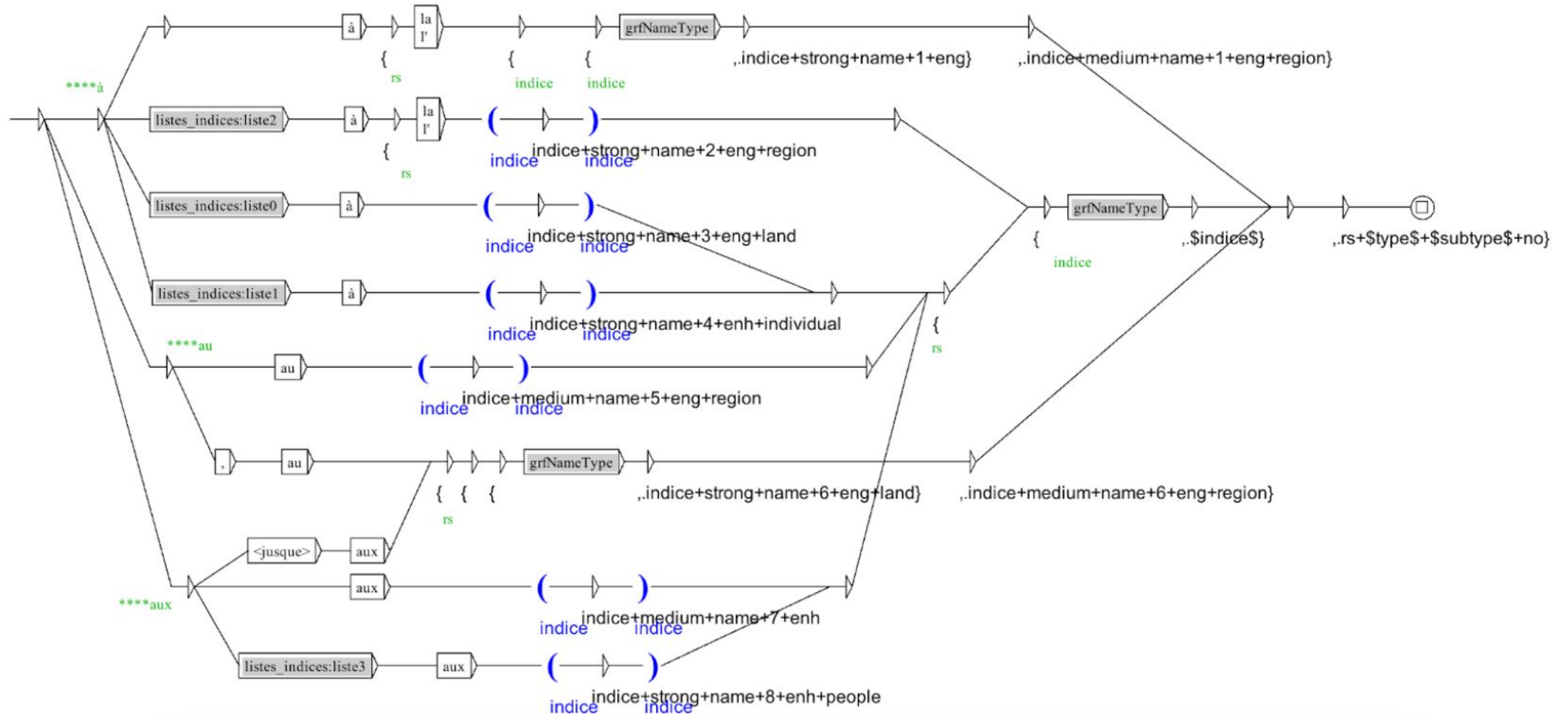
Vue sur une concordance TXM obtenue à partir de la requête CQL :

```
[lemma="achever|bâtir|chasser|découvrir|décrire|détruire|donner|ériger|fonder|fortifier|habiter|nommer|occuper|placer|posséder|rapporter|ruiner|subjuguier|tuer|vanter|vaincre"] [word="par"] [word="les"] [type="NPr"]
```

Contexte gauche	Pivot
levant de l'île de Curaçao, et	occupée par les Hollandais
au midi du détroit de le Maire,	découvert par les Hollandais
la côte d'Afrique ; il a été	découvert par les Portugais
est ainsi que l'on nomme le pays	habité par les Curdes
est un aujourd'hui à étudier. Il est	habité par les Cophtes
Ante, entre Caën et Seez, et	bâtie par les Normans
un temple de la fortune qui y fut	bâti par les Romains
rapport de Dapper ; son extrémité occidentale est	nommée par les Portugais
sous la ligne, et qui ont été	découvertes par les Espagnols
Géogr.) ile de l'Océan ainsi	nommée par les Hollandais
Valence ; ce nom qui lui a été	donné par les Maures
150 degrés, on y aperçoit une montagne	nommée par les Hollandais
la province d'Anossi. Il a été	bâti par les François
changé de nom ; les Cimbriens en furent	chassés par les Pelasges
plus de son tems ; qu'elle fut	détruite par les Romains
dans l'isle de la grande Java,	détruite par les Hollandais
la riviere de même nom ; elle fut	bâtie par les Espagnols
l'Amérique, capitale de la Jamaïque,	bâtie par les Espagnols
26 minutes de latitude septentrionale. Elle est	habitée par les Jakutes
empire russe dans la province de Daurie,	habitée par les Tonguses
orient. Les Jazyges Méthanastes, qui furent	subjugués par les Romains
a dans ce canton quatre villes murées,	bâties par les Numides
M. de Marca ; Iluro ayant été	détruite par les Mores
grande des îles de Salomon, elle fut	découverte par les Espagnols
, au midi de cette péninsule, et	placée par les Hollandais
culte du pays : ceux - ci furent	chassés par les Curdes
, dans la Natolie. Cet endroit ainsi	nommé par les Francs
, les, (Géog.) nom	donné par les Hollandais
, sur la rive droite du Rhin,	bâti par les François

# Les indices linguistiques faibles (ILFa)

Transducteur Unitex pour l'annotation des ILFa : la préposition à (règles 1 à 8)



# Premières évaluations

Description des règles et distribution  
du nombre d'occurrences d'**ENG**  
classées par chaque ILFa

Règle	Description	Ss-catégorie	Nbre occ
9	.   liste 4 + de/d' + NPr	toponyme	27715
32	en + NPr	¬ ville	17925
21	NPr + virgule   Ø , dans + le   la   les   l' + NP	¬ ville	3863
5	au + NPr	¬ ville	2895
15	liste 9 + de + la   l' + NPr	¬ ville	1900
3	liste 0 + à + NPr	toponyme	1879
1	à + la   l' + NPr	¬ ville	1532
11	liste 6 + de   d' + NPr	¬ ville	1434
33	sur + le / la / l' + NPr	hydronyme	1321
26	par + le   la   les   l' + NPr	Ø	1160
25	dans + le   la + liste 15 + de   d'	classifieur	882
20	dans + les + NPr	hydronyme	857
13	liste 8 + du   de + Ø   la   l' + NPr	hydronyme	723
22	liste 14 + dans + le   la   les   l' + NPr	hydronyme	670
14	liste 9 + du + NPr	¬ ville	594
6	virgule + au + NPr	¬ ville	508
24	virgule + dans + le   la   l' + NPr	¬ ville	447
23	liste 13 + dans + le   la   l' + NPr	¬ ville	290
10	liste 5 + de   d' + NPr	ville	116
2	liste 2 + à + la   l' + NPr	¬ ville	63
16	liste 10 + des + NPr	¬ ville	47
30	passer + par + NPr	toponyme	33



# Premières évaluations

Règle	Description	Ss-catégorie	Nbre occ
29	par + NPr	individu	5604
31	suivant   selon + NPr	individu	3588
12	liste 7 + de   d' + NPr	individu	3166
34	chez les + NPr	collectivité	2675
7	aux + NPr	Ø	1678
28	liste 16 + par + NPr	individu	1033
4	liste 1 + à + NPr	individu	951
19	liste 12 + NPr	individu	576
27	liste 17 + par + les + NPr	collectivité	269
17	liste 11 + des + NPr	collectivité	184
8	liste 3 + aux + NPr	collectivité	159
18	dans + NPr + virgule + liv	individu	39

Description des règles et distribution  
du nombre d'occurrences d'**ENH**  
classées par chaque ILFa