Sens Texte Informatique Histoire

LETTRES SORBONNE UNIVERSITÉ

Rencontres Minutes de l'équipe de Linguistique Computationnelle aka Speed Dating

10 septembre 2020

STIH, Sorbonne University

Le tiercé dans l'ordre

1. Karën Fort
2. Carlos González
3. Lichao Zhu
4. Vincent Lully
5. Imed Laaridh
6. Yoann Dupont
7. Zijian Wang
8. Julien Bezancon
9. Caroline Parfait
10. Gaël Lejeune
11. Fouad Aounti
12. Jean Baptiste Tanguy

Un séminaire pour quoi ?

- Des rencontres régulières (1 fois/mois)
- Pas trop formelles

Un séminaire pour quoi ?

- Des rencontres régulières (1 fois/mois)
- Pas trop formelles
- Comprendre l'éco-système
- Connaître les collègues
- Partager des idées
- Collaborer (en enseignement et en recherche)

Les rencontres minutes d'aujourd'hui : faire un tour d'horizon rapide et boire des cafés

Le Wiki `http://stih-sorbonne-universite.fr/dokuwiki/doku.php?id=seminaire`
La "FAQ" `https://gitlab.com/YoannDupont/accueil-su/-/blob/master/guide-ac`

Karën Fort

Création de ressources langagières et éthique pour le TAL

Karën Fort

karen.fort@sorbonne-universite.fr / https://www.schplaf.org/kf/

Séminaire d'équipe, 10 septembre 2020

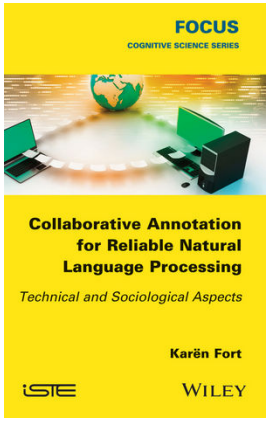
Loria
Laboratoire lorrain de recherche en informatique et ses applications

LETTRES SORBONNE UNIVERSITÉ

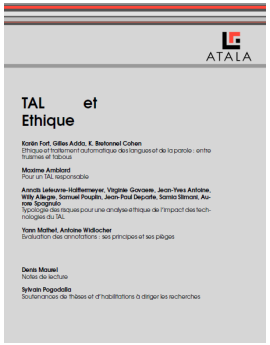
D'où je parle

Voir https://www.schplaf.org/kf/

- Création de ressources langagières pour le TAL



- Ethique et TAL



Production participative (*crowdsourcing*)

Portail de jeux pour les langues et atelier récurrent :

Games4NLP

Projets locaux :

- MEMES : A. Gautier, G. Lejeune et G. Sioufi
- CAMARADERIE : M. Avanzi et A. Thibault

Conseils

- Conseil national des universités (CNU) 27 (informatique) : https://cnu27.univ-lille.fr/
  - Qualifications
  - Promotions, CRCT, etc
- Comité d'éthique de la recherche (CER) de SU
- GDR LIFT (linguistique informatique, formelle et de terrain)
- actions européennes COST :
  - European Network for Combining Language Learning with Crowdsourcing Techniques (enetCollect)
  - Language In The Human-Machine Era (LITHME)

Carlos González

LETTRES
SORBONNE
UNIVERSITÉ

Séminaire Linguistique Computationnelle

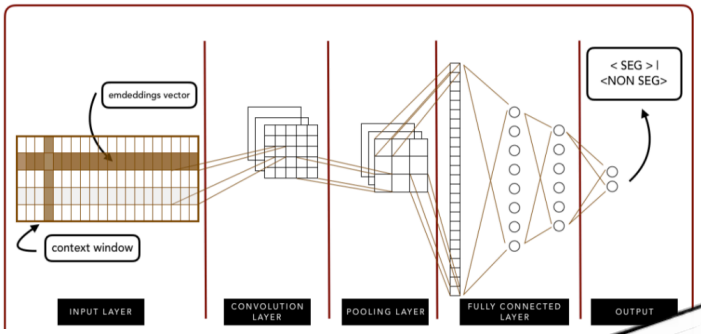
Carlos González

10/09/2020

Axes de recherche

Y a-t-il des phrases dans le langage oral ?

	Mots	Segmentation manuelle ref_1	ref_2	ref_3	Moyenne
Total	8 080	502	485	**281**	422±100

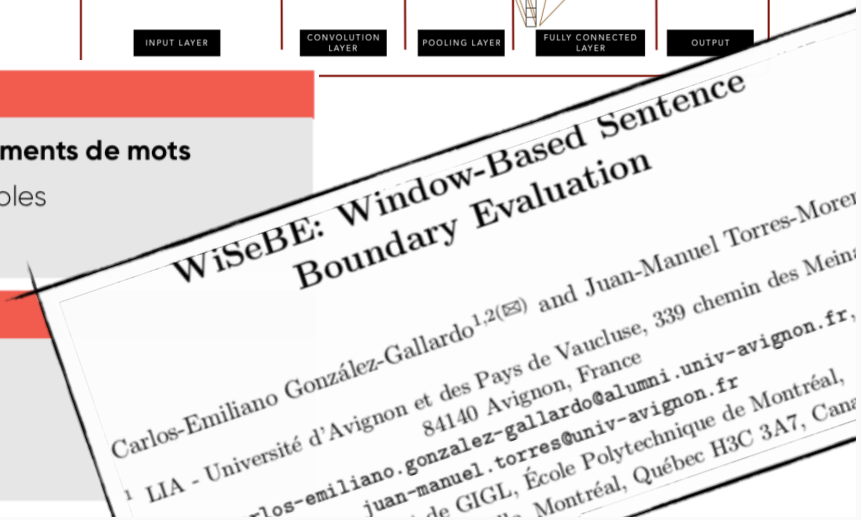


Information textuelle des transcriptions

- **Sac-de-mots, n-grammes de mots/caractères, plongements de mots**
- ✓ Information textuelle, ressources linguistiques disponibles
- ✗ erreurs de transcription

Signal audio et informations extraites

- **Pauses, durée des mots, niveau du signal, énergie**
- ✓ Indépendante de l'étape de transcription
- ✗ découpage des phrases incomplètes


WiSeBE: Window-Based Sentence Boundary Evaluation

Carlos-Emiliano González-Gallardo[1,2(✉)] and Juan-Manuel Torres-More

[1] LIA - Université d'Avignon et des Pays de Vaucluse, 339 chemin des Mei
84140 Avignon, France
{carlos-emiliano.gonzalez-gallardo@alumni.univ-avignon.fr,
juan-manuel.torres@univ-avignon.fr}
[2] GIGL, École Polytechnique de Montréal,
Montréal, Québec H3C 3A7, Cana

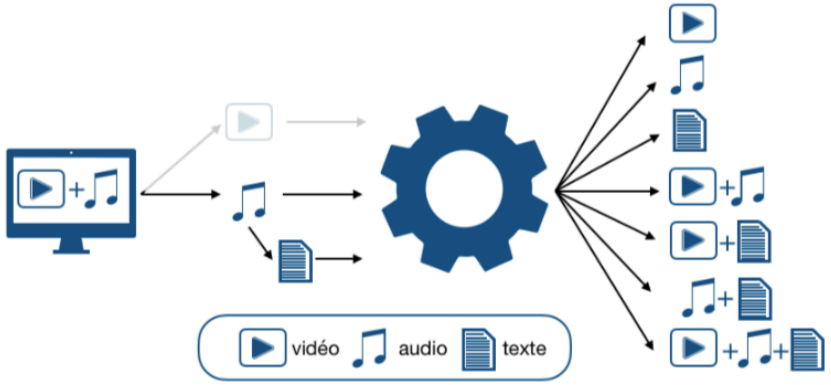
Axes de recherche

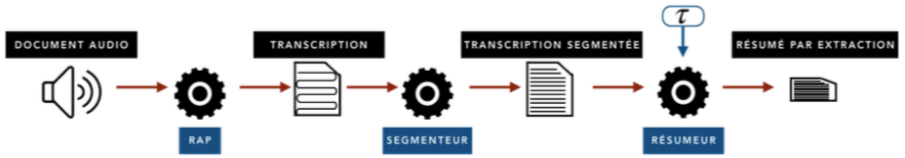
Résumer pour informer

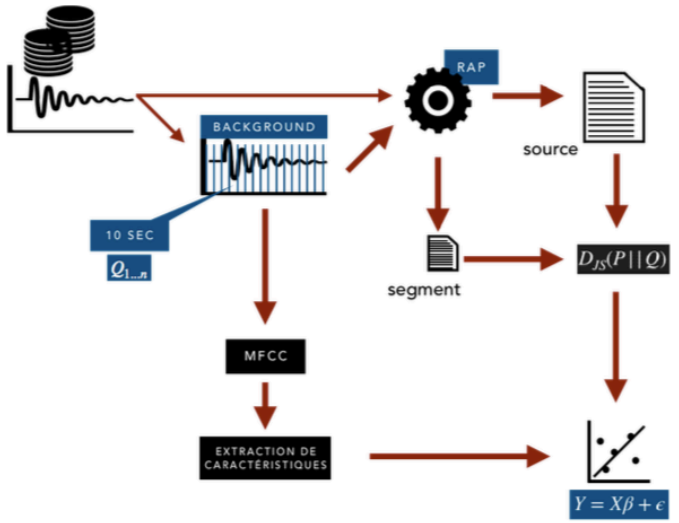
ACTUALITÉS / TÉLÉREPORTAGES

multimédia & multilingue

résumé

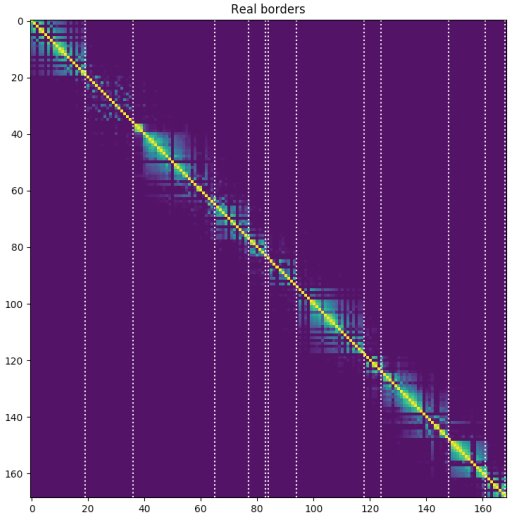


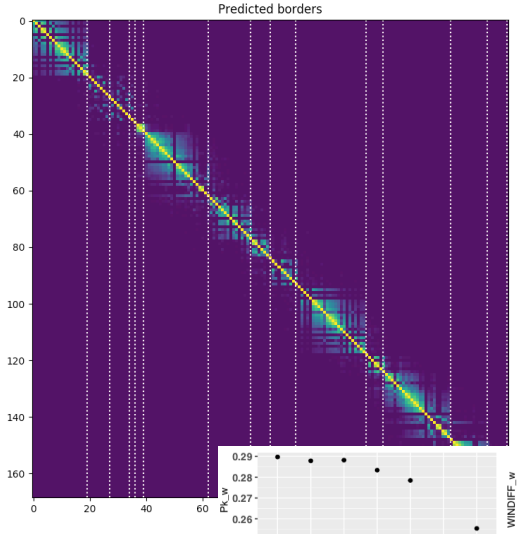


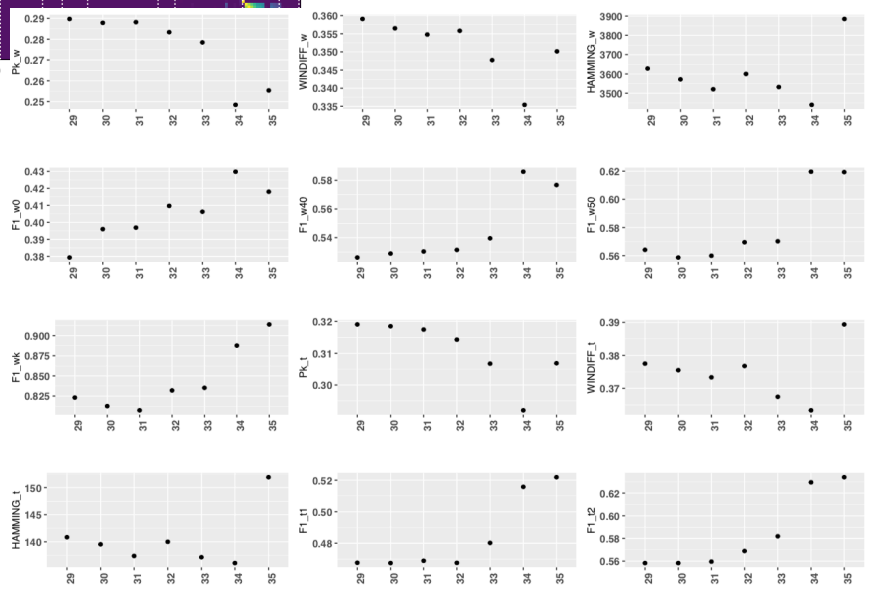


Axes de recherche

Diviser et régner







carlos-emiliano.gonzalez-gallardo@paris-sorbonne.fr

Lichao Zhu

MÈMES pas mèmes

Lichao ZHU

10 septembre 2020

Qu'est-ce ?

- Mèmes : Extraction automatique et analyse par Myriadisation d'Expressions Semi-figées (Karën, Antoine, Gaël et bibi)
- Annotations et modélisations des candidats mèmes

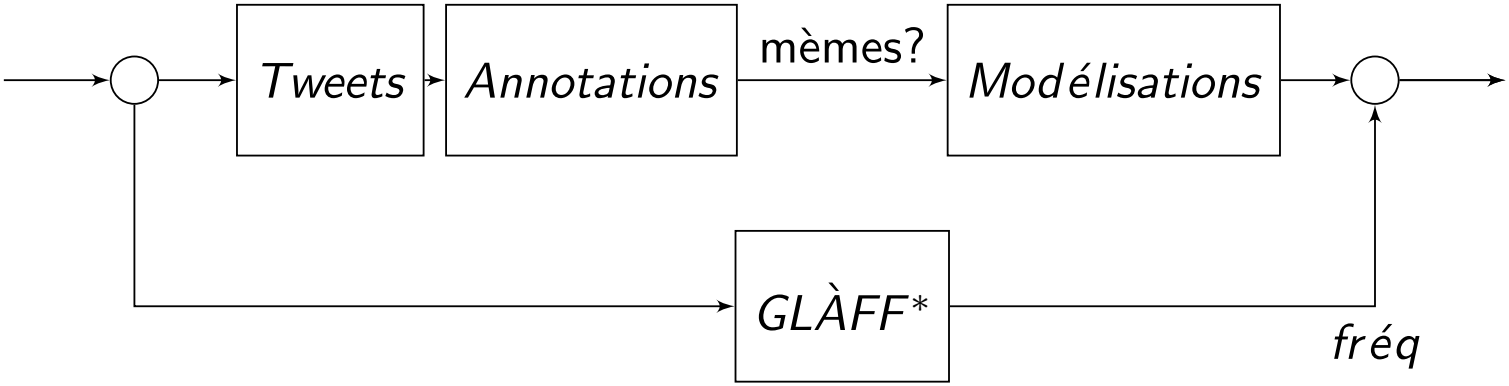
Exemples

- Roubaix, ton univers (politique) impitoyable ... !
- @PiotrSzut URSSAF, ton univers impitoyable

Questions

- Quels sont des traits formels des mèmes (et leur origine) ?
- Sur quels critères formels peut-on les repérer ?

Comment ?



Méthode

- annotations des candidats mèmes ;
- typologies phono-morpho-syntaxiques des cas avérés et non avérés ;
- modélisations formelles des mèmes avérés ;
- comparaisons avec les données du GLÀFF.

* : Gros Lexique À tout Faire du Français

Observations

Dans un mème,

- la structure syntaxique est souvent banale ;
- le nombre de mots outils est souvent supérieur à 2 ;
- au moins un mot est de moyenne ou de forte fréquence lexicale ;
- la ponctuation joue un rôle important.

Vincent Lully

- Vincent LULLY
- Master ILGII -> Thèse CIFRE -> ATER -> MCF
- Web sémantique et ses applications
  - Système de recommandation !!!
  - Moteur de recherche !!
  - Patrimoine culturel !
- Comparaisons intrinsèques et extrinsèques des plongements sémantiques et lexicaux
- Auto-complétion dans des systèmes de questions-réponses

Imed Laaridh

Imed Laaridh, PhD – Traitement automatique de la parole

Formation universitaire

2013-2017	**Doctorat en Informatique** : Laboratoire LIA, Bourse Brain and Language Research Institut (BLRI/ILCB). CERI, Université d'Avignon, Avignon, France.
2010-2013	**Ingénieur en Informatique** École Nationale des Sciences de l'Informatique (ENSI), Tunisie.
2008-2010	**Études préparatoires en Mathématiques et Physique** Institut Préparatoire aux Études d'Ingénieurs de Tunis (IPEIT), Tunis, Tunisie

Positions académiques

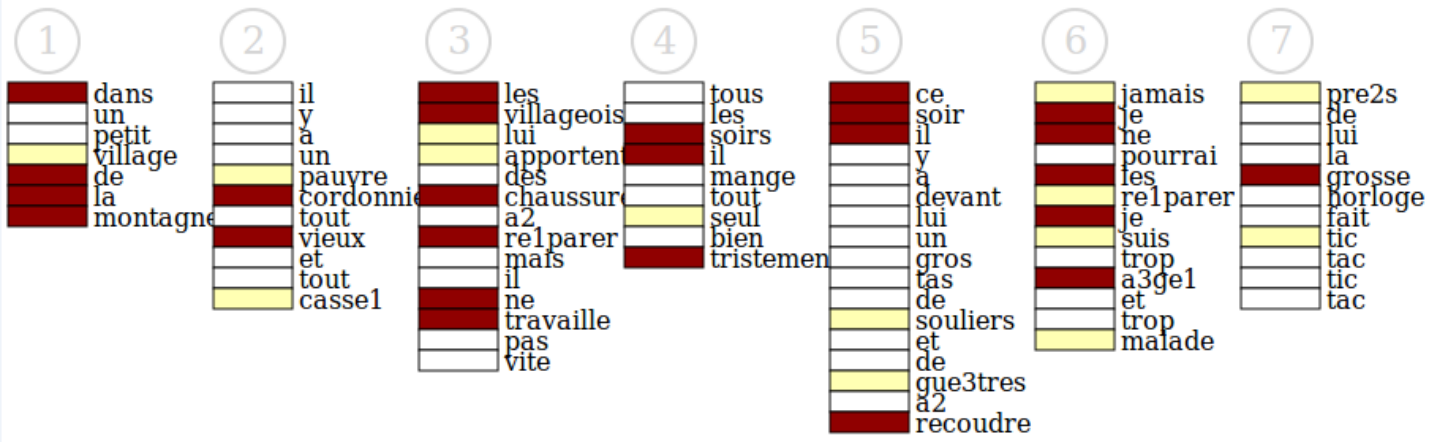
2017-2020	**Post-doc** : Laboratoire IRIT, Financement projet région & ANR. IRIT, Université de Toulouse/CNRS
2017-2017	**Ingénieur de recherche** : Laboratoire Informatique d'Avignon (LIA) LIA-CERI, Université d'Avignon, Avignon.

Thèmes de recherche

- Traitement automatique de la parole pathologique
- Évaluation de la qualité de la prononciation au niveau phonème (projet TypALoc)
  - Alignement de la parole lue des patients au niveau phonème
  - Décision au niveau phonème de la normalité ou non de la production

Thèmes de recherche

- Traitement automatique de la parole pathologique
- Évaluation de la qualité de la prononciation au niveau phonème (projet TypALoc)
  - Alignement de la parole lue des patients au niveau phonème
  - Décision au niveau phonème de la normalité ou non de la production

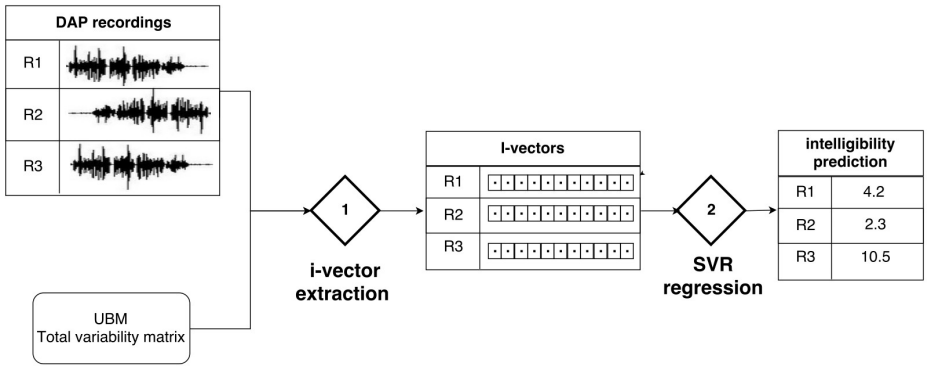


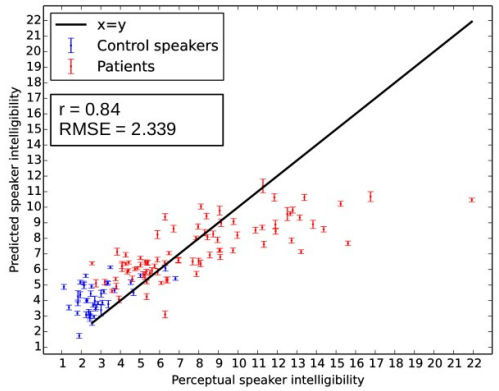
Thèmes de recherche

- Prédiction automatique de l’intelligibilité de patients atteints de la dysarthrie et de cancer de la tête et du cou (projet C2SI)
  - Utilisation des i-vecteur adaptés de la reconnaissance du locuteur pour prédire intelligibilité des patients et la sévérité de la maladie
  - Utilisation de différentes annotations automatiques (reconnaissance au niveau phonème, score de vraisemblance, confusion phonémique) pour la prédiction de l’intelligibilité

Thèmes de recherche

- Prédiction automatique de l'intelligibilité de patients atteints de la dysarthrie et de cancer de la tête et du cou (projet C2SI)
  - Utilisation des i-vecteur adaptés de la reconnaissance du locuteur pour prédire intelligibilité des patients et la sévérité de la maladie
  - Utilisation de différentes annotations automatiques (reconnaissance au niveau phonème, score de vraisemblance, confusion phonémique) pour la prédiction de l'intelligibilité





Thèmes de recherche

- Utilisation de la reconnaissance automatique de la parole pour le réglage des prothèses auditives (projet Phonics)
  - Simulation des effets des traumatismes sonores sur la parole

- Recherche de paramètres distinctifs automatique entre les patients atteints de la maladie de Parkinson et l'Atrophie Multi-Systématisée AMS (projet Voice4PD-MSA)
  - Étude du rythme de la parole chez les patients
  - Étude de la confusion phonémique et la qualité de prononciation comme marqueur de la pathologie

Yoann Dupont

Entités nommées : systèmes et données

Yoann Dupont

Les entités nommées (EN)

Japon : la droite au pouvoir remporte les élections

22 octobre 2017. – Le Parti libéral-démocrate de Shinzō Abe a remporté largement les élections législatives organisées ce dimanche. Le Premier ministre pourrait ainsi théoriquement occuper ce poste jusqu'en 2021, alors qu'il est au pouvoir depuis 2012. S'il y parvient, il battra alors le record de longévité en tant que chef du gouvernement, dans un pays habitué à l'instabilité politique. Shinzō Abe a réussi son pari d'orienter la campagne électorale vers les menaces étrangères, à savoir la Chine et la Corée du Nord, dans un contexte de tensions militaires très fortes. Le Premier ministre entend réformer la constitution japonaise pour supprimer son caractère pacifique, et développer l'armée nationale.

Shinzō Abe le 18 août 2017.

Son parti de la droite conservatrice réussit à décrocher 311 sièges sur les 465 que compte la Chambre des représentants. S'il perd quelques sièges, il conserve une large majorité absolue. Deux coalitions s'opposaient à lui : l'une de gauche et pacifiste, qui obtient 67 sièges soit presque deux fois plus que dans la législature sortante. La coalition libérale « Koike » obtient elle 58 sièges, treize de moins qu'auparavant.

Figure 1: source: `https://fr.wikinews.org`

Les entités nommées (EN)

Japon : la droite au pouvoir remporte les élections

22 octobre 2017. – Le Parti libéral-démocrate de Shinzō Abe a remporté largement les élections législatives organisées ce dimanche. Le Premier ministre pourrait ainsi théoriquement occuper ce poste jusqu'en 2021, alors qu'il est au pouvoir depuis 2012. S'il y parvient, il battra alors le record de longévité en tant que chef du gouvernement, dans un pays habitué à l'instabilité politique. Shinzō Abe a réussi son pari d'orienter la campagne électorale vers les menaces étrangères, à savoir la Chine et la Corée du Nord, dans un contexte de tensions militaires très fortes. Le Premier ministre entend réformer la constitution japonaise pour supprimer son caractère pacifique, et développer l'armée nationale.

Shinzō Abe le 18 août 2017.

Son parti de la droite conservatrice réussit à décrocher 311 sièges sur les 465 que compte la Chambre des représentants. S'il perd quelques sièges, il conserve une large majorité absolue. Deux coalitions s'opposaient à lui : l'une de gauche et pacifiste, qui obtient 67 sièges soit presque deux fois plus que dans la législature sortante. La coalition libérale « Koike » obtient elle 58 sièges, treize de moins qu'auparavant.

Figure 1: source: `https://fr.wikinews.org`

Apprentissage : système entraîné à reproduire une tâche à partir d'exemples.

Les entités nommées (EN)

Japon : la droite au pouvoir remporte les élections

22 octobre 2017. – Le Parti libéral-démocrate de Shinzō Abe a remporté largement les élections législatives organisées ce dimanche. Le Premier ministre pourrait ainsi théoriquement occuper ce poste jusqu'en 2021, alors qu'il est au pouvoir depuis 2012. S'il y parvient, il battra alors le record de longévité en tant que chef du gouvernement, dans un pays habitué à l'instabilité politique. Shinzō Abe a réussi son pari d'orienter la campagne électorale vers les menaces étrangères, à savoir la Chine et la Corée du Nord, dans un contexte de tensions militaires très fortes. Le Premier ministre entend réformer la constitution japonaise pour supprimer son caractère pacifique, et développer l'armée nationale.

Shinzō Abe le 18 août 2017.

Son parti de la droite conservatrice réussit à décrocher 311 sièges sur les 465 que compte la Chambre des représentants. S'il perd quelques sièges, il conserve une large majorité absolue. Deux coalitions s'opposaient à lui : l'une de gauche et pacifiste, qui obtient 67 sièges soit presque deux fois plus que dans la législature sortante. La coalition libérale « Koike » obtient elle 58 sièges, treize de moins qu'auparavant.

Figure 1: source: `https://fr.wikinews.org`

Apprentissage : système entraîné à reproduire une tâche à partir d'exemples.

Exemples d'applications :

- extraction d'informations (qui ? quoi ? quand ? où etc...)
- systèmes de question-réponse
- anonymisation

Travaux en cours / prévus : structuration

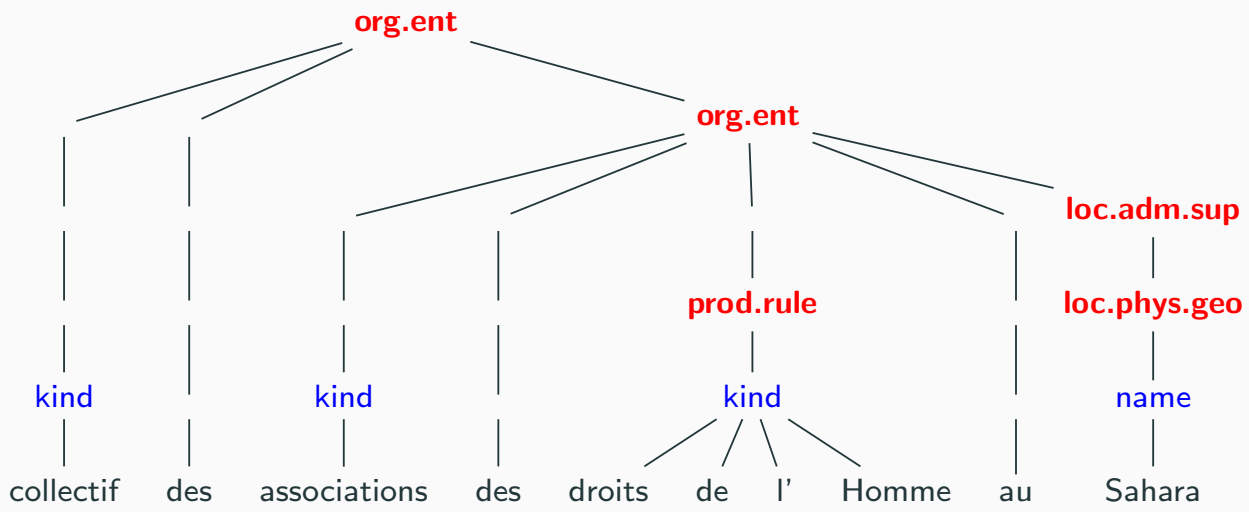
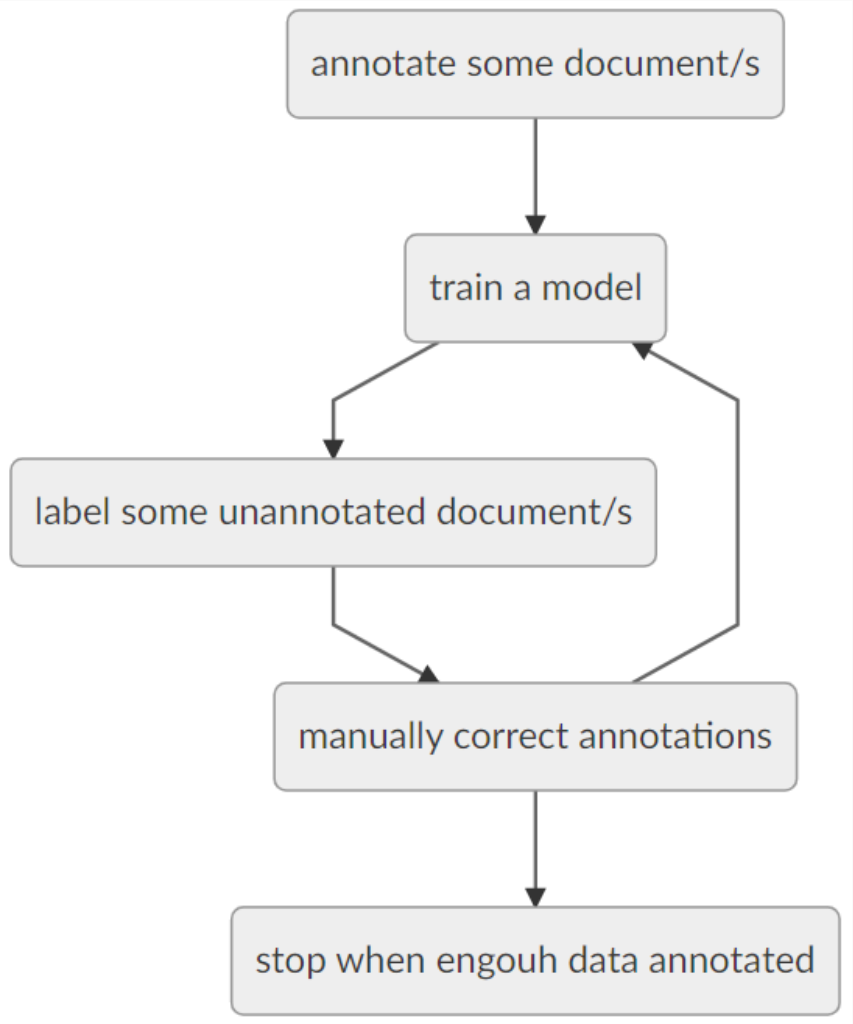


Figure 2: une entité nommée structurée. Source : corpus Quaero Grouin *et al.* (2011)

Travaux en cours / prévus : apprentissage actif



Travaux en cours / prévus : BERT Devlin et al. (2018)

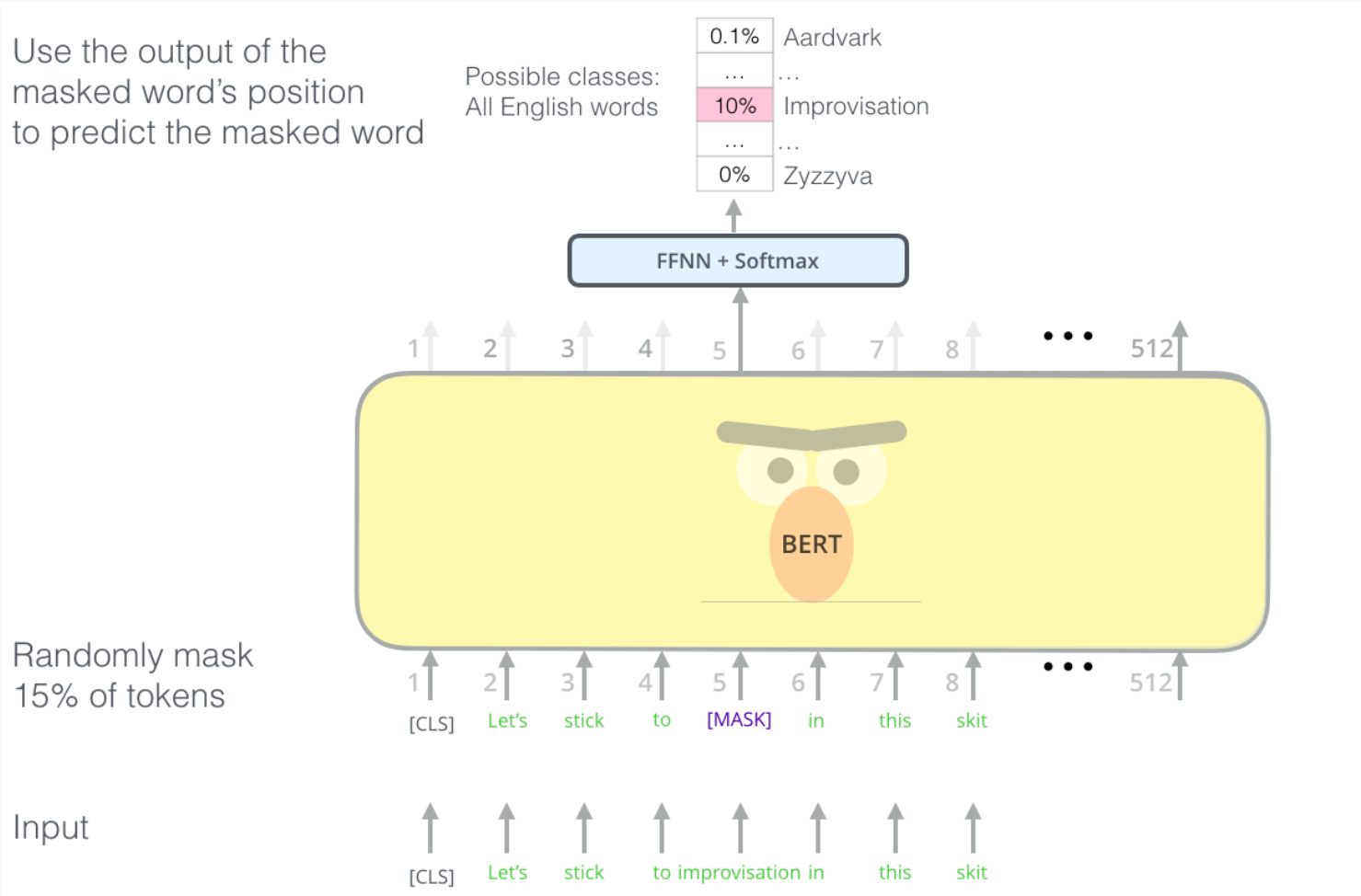


Figure 3: source : `https://jalammar.github.io/illustrated-bert/`

References

Devlin J., Chang M.-W., Lee K. & Toutanova K. (2018). **Bert: Pre-training of deep bidirectional transformers for language understanding**. *arXiv preprint arXiv:1810.04805*.

Grouin C., Rosset S., Zweigenbaum P., Fort K., Galibert O. & Quintard L. (2011). **Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview**. In *Proceedings of the 5th Linguistic Annotation Workshop*, p. 92–100: Association for Computational Linguistics.

Zijian Wang

Comparaison des résultats de différents outils pour la détection des entités nommées

Zijian WANG Master 2 Langue et Informatique

September 9, 2020

LETTRES
SORBONNE
UNIVERSITÉ

Évaluation des résultats

- *Détection des entités nommées par des outils existants*

- *Ré-entraînement des modèles*

- *Evaluation par diverses façons*

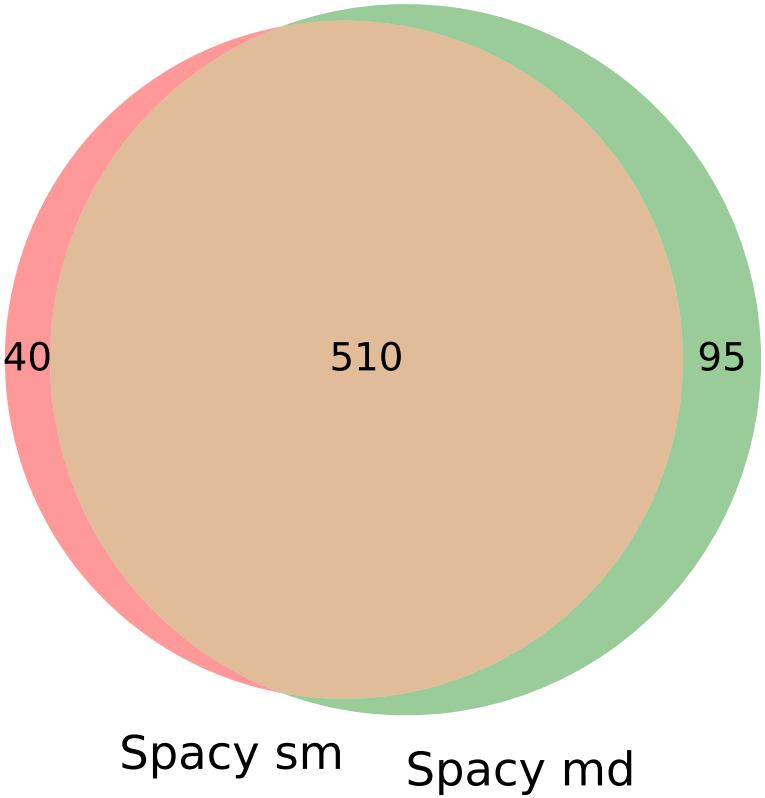
- *Visualisation des résultats pour ces évaluations*

LETTRES
SORBONNE
UNIVERSITÉ

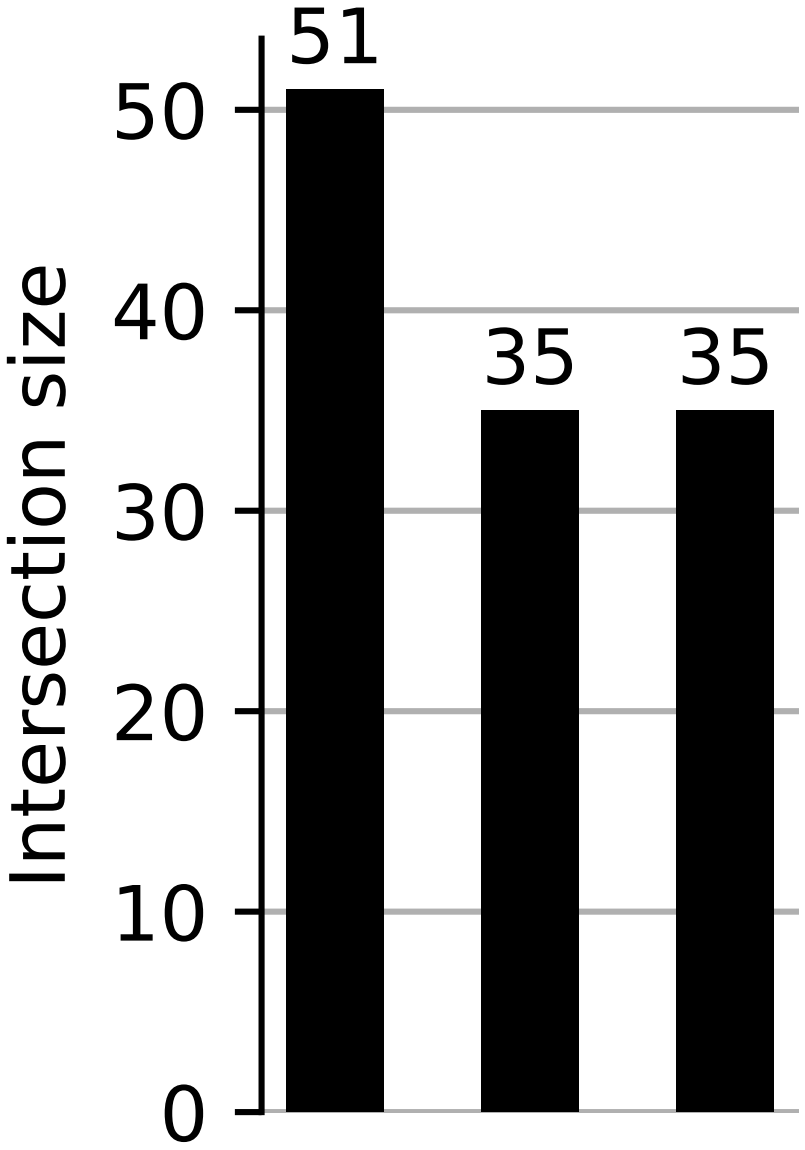
Quelle est votre destination? Paris LOC . Quels sites voulez-vous visiter? La Tour Eiffel LOC .

- Ce sont des syntagmes ne portant qu'un sens dans la langue.

- Ce sont des attributs pertinents pour l'apprentissage d'une phrase.


40
510
95
Spacy sm
Spacy md

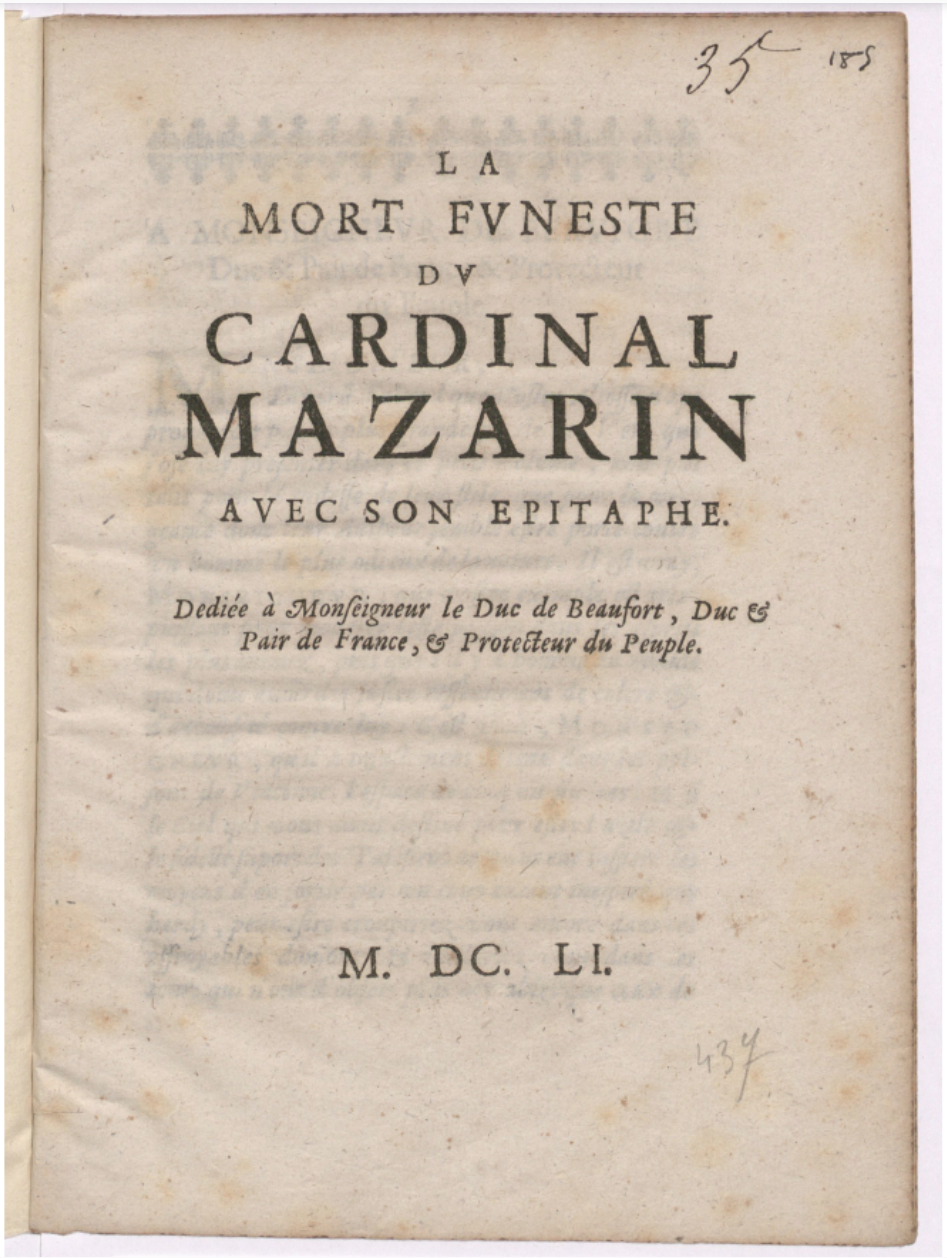
Which is the best for ORG?



Julien Bezancon

stage.pdf

Julien
Bezançon



LA
MORT FVNESTE
DV
CARDINAL
MAZARIN
AVEC SON EPITAPHE.

Dediée à Monſeigneur le Duc de Beaufort, Duc & Pair de France, & Protecteur du Peuple.

M. DC. LI.

Compte-rendu
de stage de L3

stage.pdf

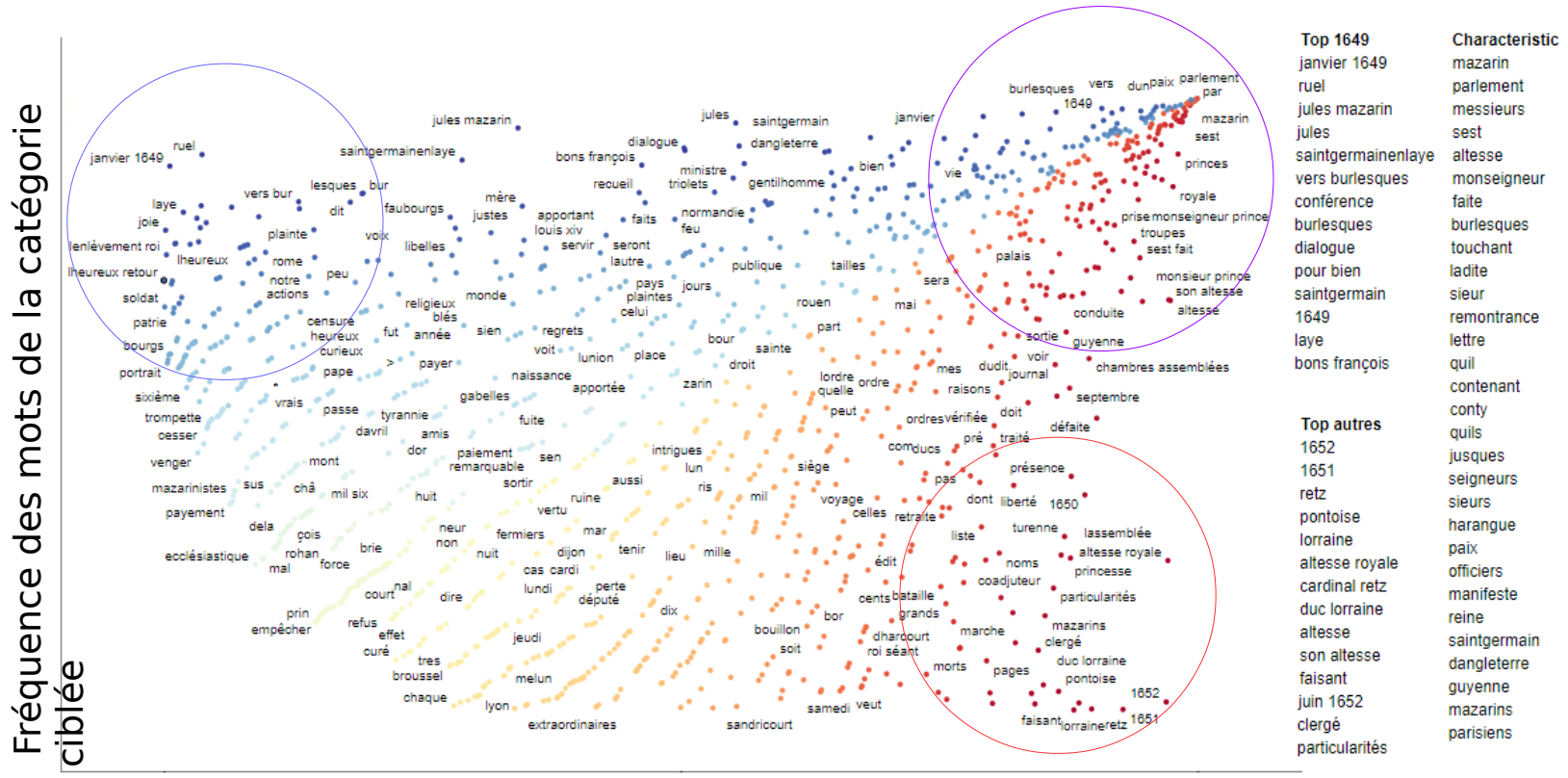
Extrait du corpus informatisé

Titres
Dates
Lieux
Pages
Textes

- Lecture du tableur
- Nettoyage des textes
- Création d'un graphique

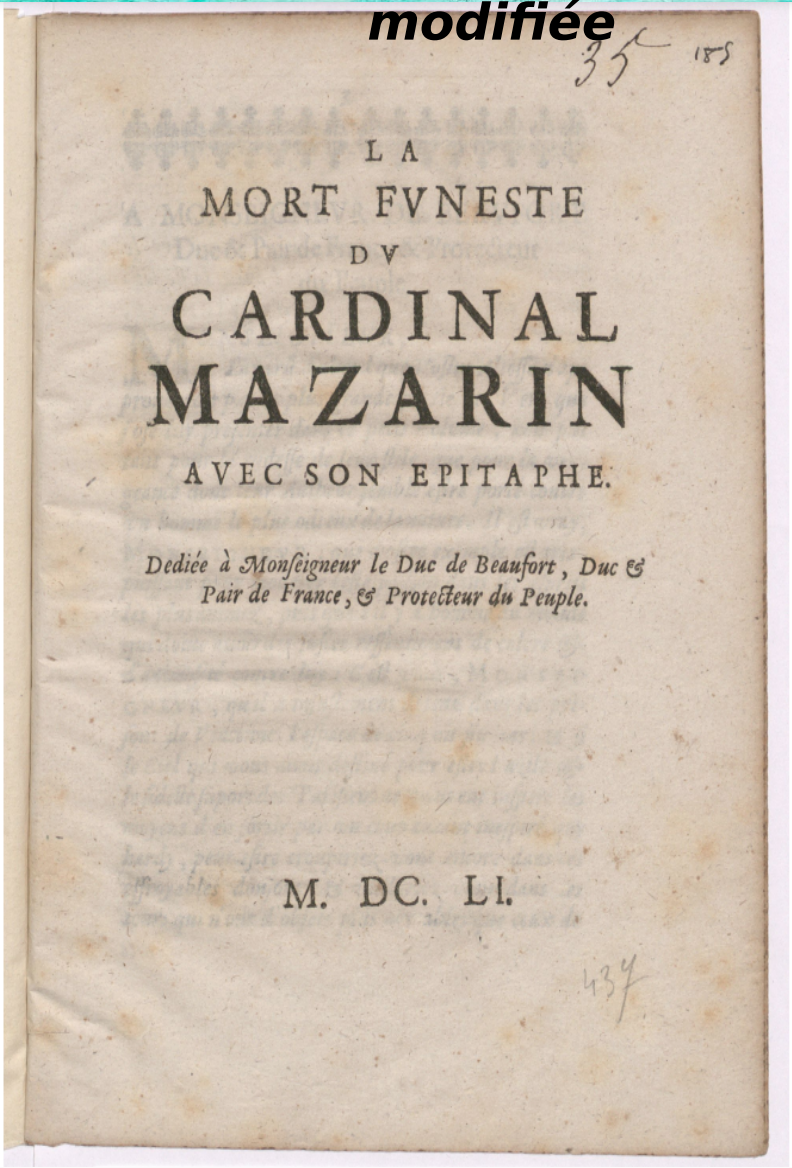
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID Moreau	titre	date	lieu	nb pages	texte									
2	1	A la reine, pa	1652		19	1. A la reine, par un ecclésiastique. « Vade , quoniam vas electionis est mihi , ut portet no- « men meum coram gentibus et regibus et filiis Is- «									
3	2	A messieurs	(s. d.)	(s. l.)		2. A messieurs du parlement. (S. l. n. d.), [1650], A pages. Rare. Pour la liberté des princes , après leur transfert au Mavre.									
4	3	A monseigne	1690		4	3. A monseigneur Charles de Lorraine, duc d'Elbeuf, généralissime des armées du roi. (S. 1. n. d.), 4 pages. Cette pièce est signée nu Bos (Ma									
5	4	À monseigneur le Coadjuteur sur sa retrai			1	4. À monseigneur le Coadjuteur sur sa retraite du 5 avril 1651, Sonnet. (S. 1. n. d.), 1 page. Rare. Mauvais vers où l'on met le coadjuteur au-dess									
6	5	À monsieur d	1649	Paris	4	5. À monsieur de Broussel , conseiller du roi au parle- ment de Paris. Paris, François Noël, 1649, 4 pages. { enefena)									
7	6	À nos seigne	(s. d.)	(s. l.)	7	6. À nos seigneurs du parlement. (S. l. n. d.), [1650], 7 pages Requête en addition pour la récusation du premier président par Beaufort, Gondy e									
8	7	À nos seigne	1649		2	7. À nos seigneurs du parlement. (S. 1. n. d.), [1649], 2 pages. C'est une requête d'opposition, faite par les propriétaires des A maisons, aux arré									
9	8	À qui aime la	1649		4	8. À qui aime la vérité. (S. 1. n. d.), [1649], 4 pages. Signé : Le désintéressé à Paris. C'est le second des deux billets que le chevalier de la Vallet									
10	9	À savoir (l') s	1652	Paris	19	9. À savoir (l') si nous aurons la paix et si notre guerre civile s'achèvera bientôt. Paris, 1652, 19 pages. Il y a un mot de bon sens dans cette pièc									
11	10	À tous les év	1660		47	10. À tous les évêques, prêtres et enfants de l'Église, Jean-François-Paul de Gondy, cardinal de Retz, ar- chevêque de Paris. (S. 1. n. d.), [1660									
12	11	À tous les ha	1652		47	11. À tous les habitants de la terre, l'heurcux génie sa- lut. Les advenues du bien souverain de l'homme, c'est à savoir le traité de la paix entre le									
13	12	A très-haut p	1652		3	12. A très-haut prince, nostre très-cher et amé cousin , le seigneur Armand de Bourbon, prince de Conty, du sang royal. (S. 1. n. d.), [1652], 3 pa									
14	13	A un ministre	1649			13. A un ministre d'État sur les œufs. (S.1.),1649, 7 pag. Toutes les rimes sont en eæx. Voilà l'explication du titre.									
15	14	Abolition (l') d	1652	Paris	8	14. Abolition (l') de M®. le duc de Beaufort au parle- ment de Paris, avec les dernières nouvelles de l'ar- mée de M. le prince, et l'ordre que M. le									
16	15	Abrégé de l'A	(s. d.)	(s. l.)	4	15. Abrégé de l'Arsenal de la foi qui est contenu en cette copie de la conclusion d'une lettre d'un secré- taire de Saint-Innocent, par lui écrite à sa									
17	16	Abrégé vérita	1652		4	16. Abrégé véritable de ce qui s'est fait pour le soula- gemnent des pauvres des villages du diocèse de Paris; la nécessité de soutenir cette entr									
18	17	Acclamation	1649	Paris	8	17. Acclamations (les) de joie des bons Parisiens sur l'heureuse arrivée de la paix. Paris, Nicolas de La Vigne, 1649, 8 pages. Avec permission.									
19	18	Accord (l') pa	1649	Paris	8	18. Accord (l') passé entre les quatre empereurs de l'Orient et les empereurs, rois et princes de l'Oc- cident, pour venger la mort du roi d'Anglet									
20	19	Accouchée (	1652	Paris	23	19. Accouchée (l') espagnole, avec le caquet des po- litiques, ou le frère et la suite du politique lutin sur les maladies de l'État, par le sieur de Sa									
21	20	Accueil (l") fa	1650		8	20. Accueil (l") fait à Son Éminence par les Borde- lois. (S. 1.), [1650], 8 pages. Ce sont les Larmes Mazarines accommodées , tant bien que m									
22	21	Achat (1') de	1649			21. Achat (1') de Mazarin en vers burlesques. (S. l.), 1649, 1% pages. Cette pièce a paru pendant le siége de Cambrai.									
23	22	Acte de la ré	1656			22. Acte de la révocation du vicariat général de M. Du Saussay, en l'archevêché de Paris. Placard in-folio, daté du 15 mai 1656. Trés-rare.									
24	23	Acte (l') d'opp	1649		7	23. Acte (l') d'opposition de M. le duc de Chaulnes, gouverneur et lieutenant général pour le roi du Haut et Bas (sic) Auvergne, à la proposition fa									
25	24	Acte portant	1656			24. Acte portant la nomination faite par M. le cardinal de Retz, archevêque de Paris, d'un official et d'un vice-gérant en l'Officialité de Paris au lieu									
26	25	Action de grâ	1649	Paris		25. Action de grâce à nos seigneurs de parlement par les habitants de la ville et faubourgs de Paris, pour l'ac- quittement et la décharge des loua									
27	26	Actions de gr	1650		16	26. Actions de grâces à Dieu pour la paix, par un prêtre bourdelois. (S. l. n. d.), [1650], 16 pages. Très-rare. « Figurez-vous une ville, recomman									
28	27	Actions de gr	1649	Paris	8	27. Actions de grâce de la France au prince de la paix, monseigneur le duc de Beaufort, par le sieur D. P. Paris, veuve À. Musnier, 1649, 8 page									
29	28	Actions de gr	1640		8	28. Actions de grâces de toute la France à monseigneur le prince de Condé, touchant son consentement à la paix, fait par un Bourguignon. (S. 1									
30	29	Actions (les)	1649	Paris	8	29. Actions (les) de grâces des bourgeois et habitants de la ville de Paris faictes au roy, à la reyne et aux princes après l'herex (sc) retour de Sa									
31	30	Actions de gr	1649	Paris	16	30. Actions de grâces des pauvres paysans de l'élection de Paris pour le soulagement des tailles que la reine leur a promis par la déclaration de									
32	31	Ad Annam au	1649		4	31. Ad Annam austriacam Ludovici XTV, Gallorum regis christianissimi, reginam matrem{de Julio Mazarino so- lemni senatûs consulto fuger 3us									
33	32	Ad christiani	1652	Paris		32. Ad christianissimum Francorum et Navarræ regem Ludovicum XIV, a Dco datum, felicem, invictum, cle- mentem, Parrhisios, post civicos tu									
34	33	Ad Deum pre	(s. d.)			33. Ad Deum prepotentem, optimum, maximum, filium- que ejus Jesum Christum, Dominum Nostrum, pro pace generali, deprecatio hebraicè, l									
35	34	Ad regem Lu	1649	Paris	7	34. Ad regem Ludovicum XTV ut in urbem regni prin- cipem redire velit. Ode. Parisis, Dyonisius Lan- gl:£us, 1649, 7 pages. Cette pièce est sign									
36	35	, Addition que	1649	Paris	15	35. Addition que les Conseillers secrétaires du roi font aux r'emontrances et avis qu'ils ont ci-devant dressés pour ce qui regarde le payement de									
37	36	Adieu (l) burl	1649	Paris		36. Adieu (l) burlesque de la France à la guerre. Paris, Pierre du Pont, 1649, S pages. Plaisante description de l'armée royale.									
38	37	Adieu (l') de J	1649		6	37. Adieu (l') de Jules Mazarin à la France, à Paris et à messieurs du Parlement. Pans, Mathurm Henault, 1649, 6 pages. Signé Francese Corne									
39	38	Adieu (l') de l	1652	Paris	6	38. Adieu (l') de la France au cardinal Mazarin. Paris, 1652, 6 pages. Pièce assez plaisante mais surtout trés-rare.									

Graphique obtenu avec le programme créé



1649 document count: 1,577; word count: 20,580
autres document count: 2,789; word count: 47,225

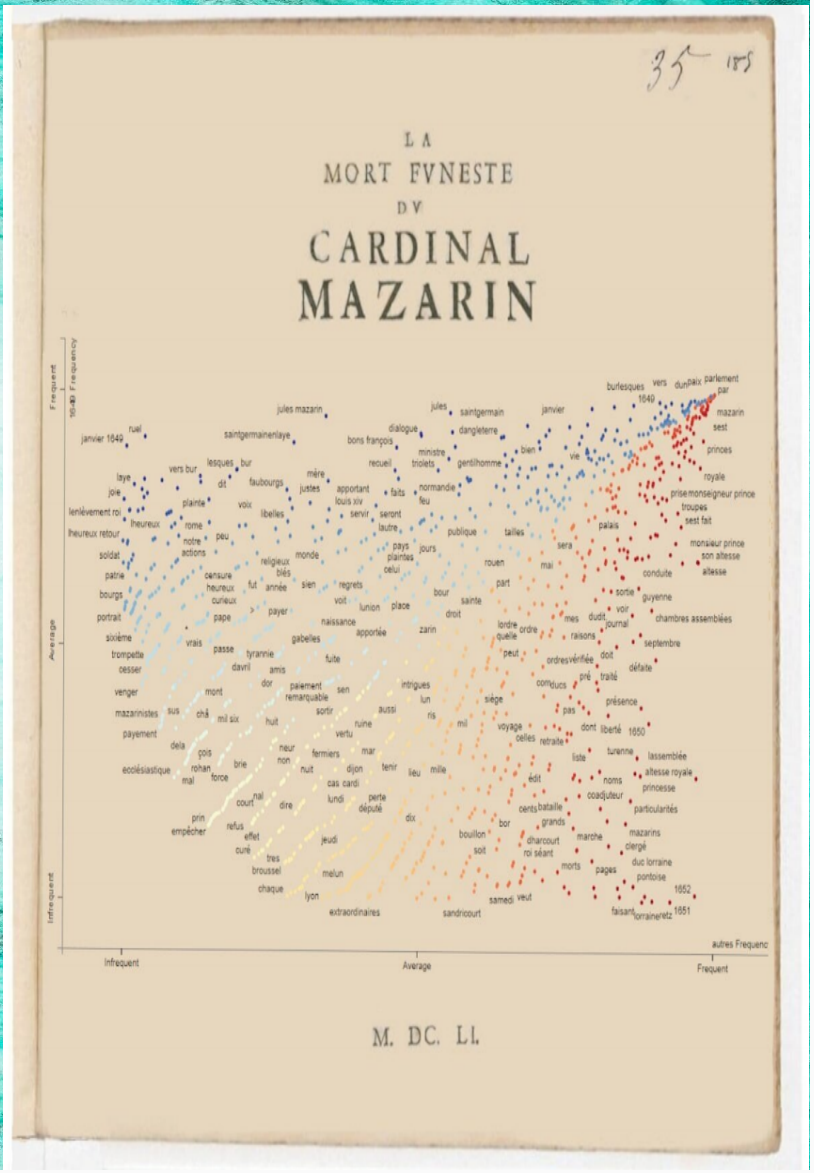
Couverture originale / couverture modifiée



LA
MORT FVNESTE
DV
CARDINAL
MAZARIN
AVEC SON EPITAPHE.

Dediée à Monſeigneur le Duc de Beaufort, Duc & Pair de France, & Protecteur du Peuple.

M. DC. LI.



LA
MORT FVNESTE
DV
CARDINAL
MAZARIN

M. DC. LI.

Caroline Parfait

Speed dating recherche, Septembre 2020

contrat doctoral SCAI

LETTRES SORBONNE UNIVERSITÉ

Analyse de l'espace littéraire : apprentissage automatique et évaluation des systèmes de reconnaissance des entités nommées

Caroline Parfait, Humanités Numériques

Centre d'études supérieures de la Renaissance

Intelligence Patrimoines

Objectifs

- Dépasser les problèmes de **variabilités**
- Utiliser le potentiel des outils de **machine learning pour la REN** dans la Littérature

Livrables :

- Un **modèle adaptatif** d'outils numériques pour la REN spatiales
- Un **corpus standard** annoté
- un **guide d'annotation** pour les Humanités Numériques littéraires

Méthodologie proposée : “conception centrée utilisateur”

- **Identification des besoins** et des contextes d’utilisation → penser l’interdisciplinarité
- **Évaluer** les modèles **existants** pour la REN spatiale :
  - Corpus : Récit de voyages du XIXème et XXème siècle
    - ***ANR Chapitre***, Paris 3 Sorbonne Nouvelle → 3000 documents
    - Corpus de la ***Très Grande Bibliothèque*** → 35000 documents
- **Modéliser** un nouvel outil
- **Évaluation** auprès des **utilisateurs**

Perspectives

- **Applicabilité transversale** -> extraction de concepts médicaux
- Entériner des **analyses littéraires à partir des résultats** (stylométrie, analyses comparatives, analyses des sentiments et des émotions…)
- Exploiter les **technologies du web sémantique pour la désambiguïsation** des EN (data linking)

Merci de votre attention

Gaël Lejeune

Qui suis-je ?

- Gaël Lejeune MCF en informatique
- Thèse en 2013 : Veille Epidémiologique Multilingue

Et depuis ?

Différentes tâches :

- Extraction de Contenu/de structure (PDF, PNG, HTML...)
- Classification (polarité, émotion, dialectes, datation ...)
- Extraction d'Information et Reconnaissance d'Entités Nommées
- Indexation/Terminologie
- Evaluation Intrinsèque VS Extrinsèque (plongements...)

Qui suis-je ?

- Gaël Lejeune MCF en informatique
- Thèse en 2013 : Veille Epidémiologique Multilingue

Et depuis ?

Différentes tâches :

- Extraction de Contenu/de structure (PDF, PNG, HTML...)
- Classification (polarité, émotion, dialectes, datation ...)
- Extraction d'Information et Reconnaissance d'Entités Nommées
- Indexation/Terminologie
- Evaluation Intrinsèque VS Extrinsèque (plongements...)

Dans un contexte de variation dans les données

- Multilinguisme : comment analyser n langues ?
- Hétérogénéité : comment traiter n états de textes ?
- Massification : comment travailler sur n To de textes ?

Qui suis-je ?

- Gaël Lejeune MCF en informatique
- Thèse en 2013 : Veille Epidémiologique Multilingue

Et depuis ?

Différentes tâches :

- Extraction de Contenu/de structure (PDF, PNG, HTML...)
- Classification (polarité, émotion, dialectes, datation ...)
- Extraction d'Information et Reconnaissance d'Entités Nommées
- Indexation/Terminologie
- Evaluation Intrinsèque VS Extrinsèque (plongements...)

Dans un contexte de variation dans les données

- Multilinguisme : comment analyser n langues ?
- Hétérogénéité : comment traiter n états de textes ?
- Massification : comment travailler sur n To de textes ?

Qu'est-ce qui m'intéresse ?

Il n'existe pas de telle chose qu'une donnée parfaite

Tout pré-traitement amène son lot de désagréments

- Enlever les ponctuations, pourquoi ?
- Découper en mots, pourquoi ?
- Découper en phrases, pourquoi ?
- $\rightarrow$ Tendance de l'informatique à javelliser/uniformiser les données et les approches

Sur quoi je travaille ?

Projets en cours

- MEMES (2019-2021) Memes : Extraction automatique et analyse par Myriadisation d'Expressions Semi-figées (K.Fort, A.Gautier, L.Zhu)
- ANTONOMAZ (2018-2022) ANalyse auTOmatique et NumérisatiOn des MAZarinades (K.Abiven, G.Roe , JB.Tanguy*et al.*)
- OBVIL-NER (2020-2023) Apprentissage automatique et évaluation des systèmes de reconnaissance des entités nommées (C. Parfait, M. Alrahabi, G. Roe)
- WADDLE et DANIEL (2018-...) Exploitation de Données Textuelles hétérogènes (A. Barbaresi, E. Giguet)
- SINNER (2020-...) Extraction d'Entités Nommées en contexte Multilingue et Bruité (Y.Dupont, P. Ortiz, T. Tian)

Fouad Aounti

Détection d'objets dans des images médiévales

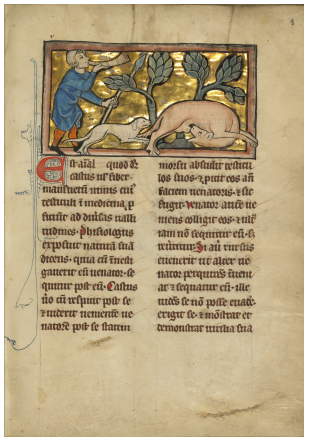
Fouad AOUINTI

Maison de la Recherche
Sorbonne-Université, Paris

10 septembre 2020

Préparation de données

- Dataset original, environ de 1000 images :
  - train set : 80% d'images ;
  - test set : le reste 20%.


- Techniques d'augmentation : scale, rotation, zoom, flip, etc.
- 2 classes : *graphic*, *no-graphic*.
- Deep Learning en Python : Keras avec Tensorflow.

Classification d'images

- Modèles CNN [a] pré-entrainés : VGG16, ResNet, etc.
- Stratégie du Transfer Learning : entrainer tout le modèle.

a. Convolutional Neural Network

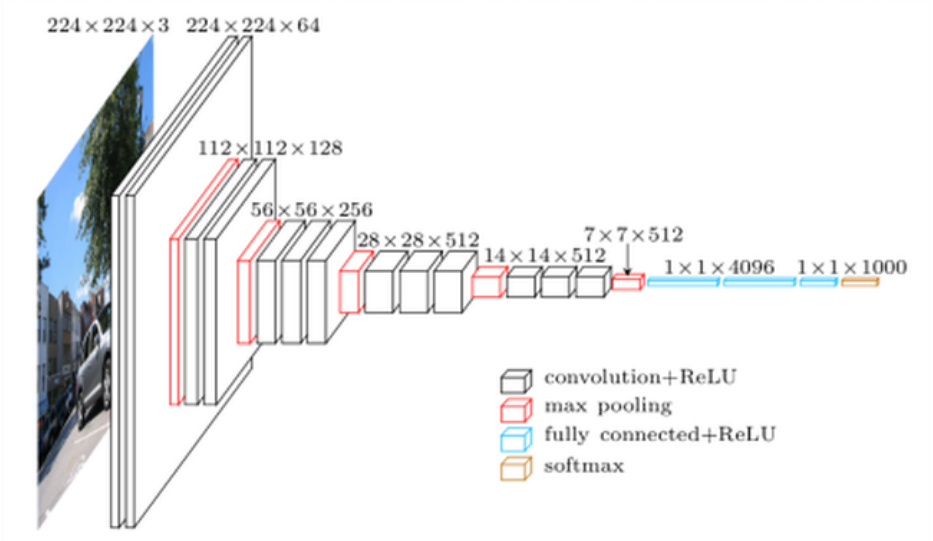
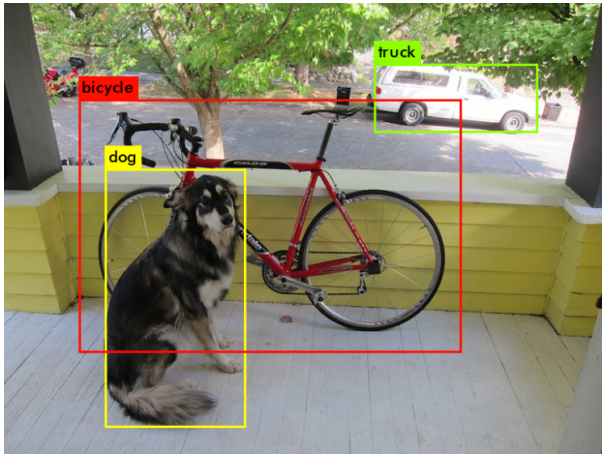
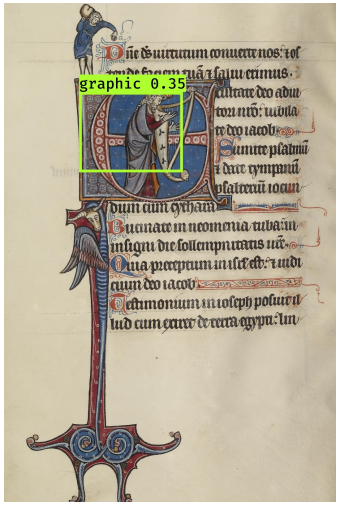


FIGURE 1 – Architecture du VGG16.

Détection d'objets



(a)



(b)

FIGURE 2 – (a) Image réelle, (b) Images médiévales.

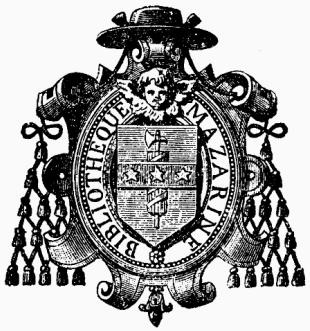
Jean Baptiste Tanguy

Une thèse en humanités numériques

Encrages linguistiques de la politisation de l'imprimé pendant la Fronde (1648-1653) : la masse polémique des Mazarinades

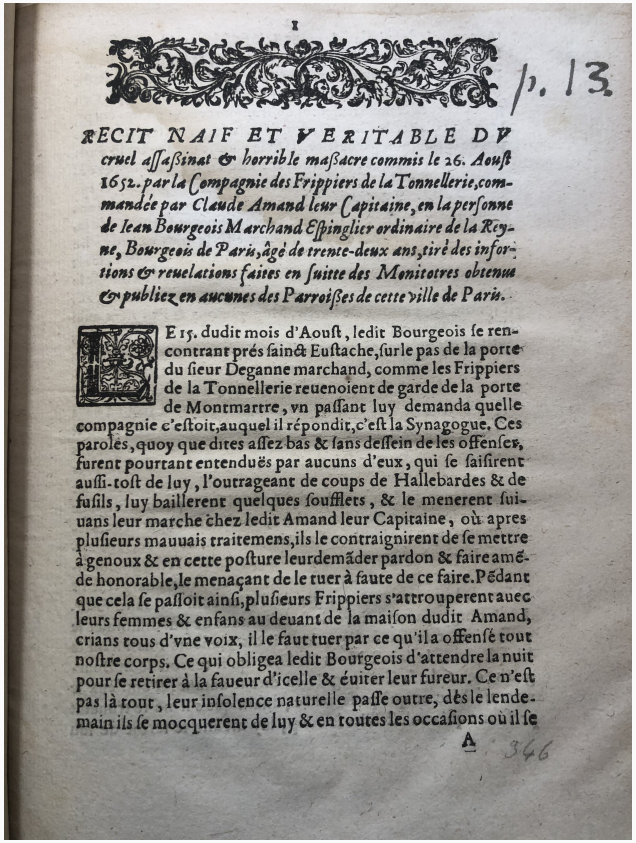
Directeur : Glenn Roe
Encadrants : Karine Abiven et Gaël Lejeune
Rattachements : LabEx OBVIL et équipe STIH
Partenaire : Bibliothèqe Mazarine
Financement : Région Île-de-France

Observatoire de la vie littéraire



îledeFrance

Objet d'étude : les Mazarinades



RECIT NAIF ET VERITABLE DV cruel aſſaſsinat & horrible maſsacre commis le 26. Aouſt 1652. par la Compagnie des Frippiers de la Tonnellerie, commandée par Claude Amand leur Capitaine, en la perſonne de Iean Bourgeois Marchand Eſpinglier ordinaire de la Reyne, Bourgeois de Paris, âgé de trente-deux ans, tiré des informations & reuelations faites en ſuitte des Monitoires obtenus & publiez en aucunes des Parroiſſes de cette ville de Paris.

LE 15. dudit mois d'Aouſt, ledit Bourgeois ſe rencontrant prés ſainct Euſtache, ſur le pas de la porte du ſieur Deganne marchand, comme les Frippiers de la Tonnellerie reuenoient de garde de la porte de Montmartre, vn paſſant luy demanda quelle compagnie c'eſtoit, auquel il répondit, c'eſt la Synagogue. Ces paroles, quoy que dites aſſez bas & ſans deſſein de les offenſer, furent pourtant entenduës par aucuns d'eux, qui ſe ſaiſirent auſſi-toſt de luy, l'outrageant de coups de Hallebardes & de fuſils, luy baillerent quelques ſoufflets, & le menerent ſuiuans leur marche chez ledit Amand leur Capitaine, où apres pluſieurs mauuais traitemens, ils le contraignirent de ſe mettre à genoux & en cette poſture leur demãder pardon & faire amẽde honorable, le menaçant de le tuer à faute de ce faire. Pẽdant que cela ſe paſſoit ainſi, pluſieurs Frippiers s'attrouperent auec leurs femmes & enfans au deuant de la maiſon dudit Amand, crians tous d'vne voix, il le faut tuer par ce qu'il a offenſé tout noſtre corps. Ce qui obligea ledit Bourgeois d'attendre la nuit pour ſe retirer à la faueur d'icelle & éuiter leur fureur. Ce n'eſt pas là tout, leur inſolence naturelle paſſe outre, dés le lendemain ils ſe mocquerent de luy & en toutes les occaſions où il ſe

A

Figure 1 – *Récit naïf et véritable...*, s.l.n.d., 7 pages, Bibl. Maz. M. 12232

- Entre 5000-6000 Mazarinades : plusieurs bibliographies existent mais aucune n'est/ne peut être exhaustive
- Elles nous sont mystérieuses : date, lieu, auteur, conditions de parution, de tirage, de lecture, contexte...
- Pamphlets, actes de la cours, récits de guerre ou de fête, lettres... parus en France pendant la Fonde (1648-1653) : politisation du petit imprimé (C. Jouhaud)
- Une multitude de formes qui se **voient** et s'**entendent** dans la ville à l'actualité affolée
- Français pré-classique (ex. : u=v, i=j, présence du s long, etc.) mais aussi latin, italien, patois
- Exemplaires datant du XVIIème siècle, donc : tâches, restaurations, etc.

Axes de recherche

- Identifier des petits corpus cohérents :
  - Des "faisceaux" (C. Jouhaud), des réseaux, etc. : travail des historiens
  - Réaliser du *clustering* pour rapprocher des pièces jusqu'alors restées isolées
- Construire des corpus numériques :
  - Acquisition des données (sur documents historiques) : numérisation, OCR, évaluation (transcription)
  - Mise à disposition pérenne des données : consortium CAHIER et FAIRisation des données
- Étude des phénomènes linguistiques mis en jeu dans la politisation de l'imprimé :
  - Néologie, isotopie, phraséologie, rhétorique...
  - Linguistique sur corpus, à partir de données textuelles océrisées, donc bruitées

Merci à tous

Figure 1 – Source : Jeff Stahler 2014