Séminaire de rentrée : Recontres Minute Science

08 septembre 2022

STIH, Sorbonne Université

Programme

1. Les rencontres minute
2. Alexandre Bartz
3. Caroline Parfait
4. Corina Chutaux
5. Eva Lacroix
6. Gael Lejeune
7. Ibtihel Ben Ltaifa
8. Julien Bezancon
9. Karen Fort
10. Laurie Acensio
11. Nour El Houda N E

Les rencontres minute

Un séminaire pour quoi ?

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

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

Alexandre Bartz

Alexandre Bartz
Ingénieur de projet - Antonomaz

Séminaire de rentrée - Linguistique Computationnelle

Le projet Antonomaz

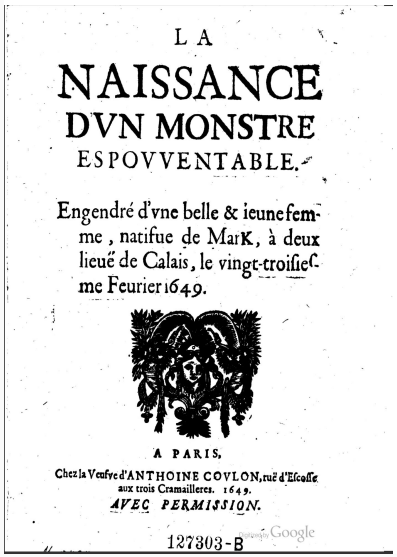
Les mazarinades, c'est quoi ?

- Corpus d'environ 5.000 textes publiés sous la Fronde (1648-1653), principalement contre (ou pour) Mazarin ;
- Textes courts (8 pages env.), imprimés sur du papier de mauvaise qualité ;
- Production très importante pour l'époque : textes d'actualités.

Les objectifs du projet :

- Rendre le maximum de textes accessibles aux chercheuses et chercheurs ;
- Proposer des exploitations automatiques des données (métadonnées + texte) ;
- Remettre les mazarinades dans leur contexte.

OCR et encodage automatique

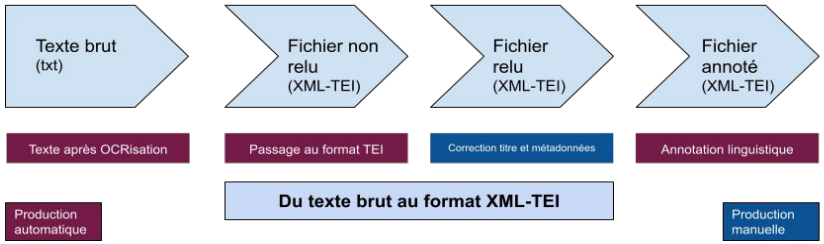


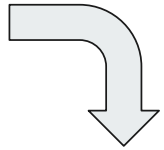
LA
NAISSANCE
DVN MONSTRE
ESPOVVENTABLE.

Engendré d'vne belle & ieune femme, natifue de Mark, à deux lieuë de Calais, le vingt-troiſiefme Feurier 1649.

A PARIS,
Chez la Veufve d'ANTHOINE COVLON, ruë d'Eſcoſſe aux trois Cramailleres. 1649.
AVEC PERMISSION.

127303-B





```
<text>
  <body>
    <p><pb n="5"/>
      <lb/>LA
      <lb/>NAISSANCE
      <lb/>DVN MONSTRE
      <lb/>ESPOVVENTABLE.
      <lb/>Engendré d'vne belle &amp; ieune fem¬
      <lb/>me, natifue de Mark, à deux
      <lb/>lieuë de Calais, le vingt-troiſief¬
      <lb/>me Feurier 1649. <figure type="decoration"/>
      <lb/>A PARIS,
      <lb/>Chez la Veufve d'ANTHOINE COVLON, ruë d'Eſcoſſe
      <lb/>aux trois Cramailleres. 1649.
      <lb/>
      <imprimatur>AVEC PERMISSION.</imprimatur>
      <pb n="6"/>
```

Rendre accessibles les données

- Plus de 2.600 textes encodés en XML-TEI ;
- Choix de l'outil TEI Publisher pour publier ces données sur le web.

Application (toujours en cours de conception) accessible en ligne :

https://antonomaz.huma-num.fr

Caroline Parfait

Reconnaissance des Entités Nommées spatiales dans un corpus littéraire bruité : robustesse des systèmes ?

Caroline Koudoro-Parfait (1,2,3), Glenn Roe (1), Gaël Lejeune (2), Motasem Alrahabi (1)

Rencontres minutes Science - STIH

8 septembre 2022

caroline.parfait@sorbonne-universite.fr
(1) OBTIC, Sorbonne Université, Paris, France
(2) STIH, Sorbonne Université, Paris France
(3) SCAI, Sorbonne Center for Artificial Intelligence, Paris, France

Contexte et enjeux

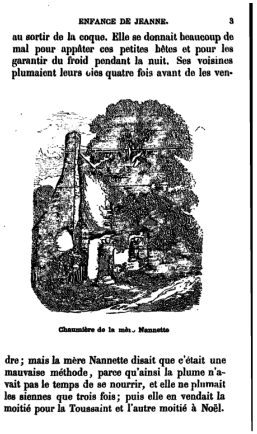
Question Comment rendre les systèmes de REN plus robustes face aux variations dans les données qui leur sont soumises par les utilisateurs ?

Angle Expé. : "Impact du bruit des transcriptions OCR sur la REN ?" Comment évaluer la qualité des résultats de REN sur des corpus OCR bruités ?

Données Corpus ELTeC (Réf.) [1] : Litt. Française 19ème siècle, 11 oeuvres, 3195 p.

OCR Kraken, Tesseract (Modèle français et de base)

Modèles REN fra. Spacy, Stanza, SEM, CasEN


ENFANCE DE JEANNE. 3

au sortir de la coque. Elle se donnait beaucoup de mal pour appâter ces petites bêtes et pour les garantir du froid pendant la nuit. Ses voisines plumaient leurs oies quatre fois avant de les ven-

Chaumière de la mère Nannette

dre ; mais la mère Nannette disait que c'était une mauvaise méthode, parce qu'ainsi la plume n'avait pas le temps de se nourrir, et elle ne plumait les siennes que trois fois ; puis elle en vendait la moitié pour la Toussaint et l'autre moitié à Noël.

Kraken
Ses voisines plumaient leurs oioes quatre fois avant de les ven- LL I II II I I IIF M ii I I II E E g Chamnnlhrs de t a mn Mamnnetta dre ; mais la mere Nannette disait que eetait une mauvaise m6thode, paree qu'ainsi la plume ...

Table – Transcriptions OCR d'une illustration et de sa légende.(Vert : illustration, bleu : légende de l'illustration)

Figure – Illustration et légende.

1. European Literary Text Collection, `https://www.distant-reading.net/eltec/`

Evaluer l'impact de la contamination OCR sur la REN

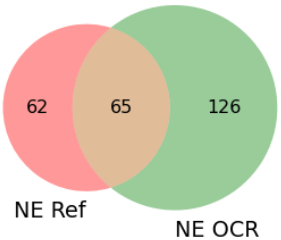
➜ **OCR de meilleure qualité → Moins de Faux Négatifs/Faux Positifs ?**

Version	**Context**	`Spacy_lg`	`Stanza`	`SEM`	`CasEN`
Ref.	*les États [...] de Guadalajara*	Guadalajara	Guadalajara	Guadalajara	Guadalajara
Kraken	*les Etats [...] de Guadalazara*	Guadalazara	Guadalazara	Guadalazara	()
Tess fr	*les États [...] de Guadalaæw*a*	Guadalaæw*a	Guadalaæw*a	()	()

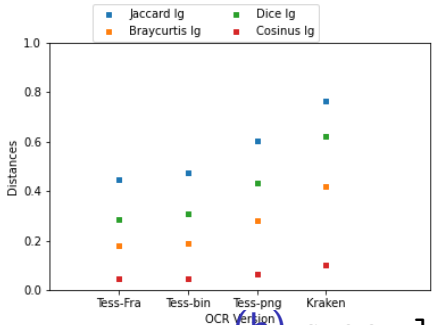
Table – Evaluation manuelle des formes contaminées des entités nommées (EN).

Version	#Entités		Évaluation par NERVAL			
	Version OCR	Référence	Intersection	Précision	Rappel	F_1 mesure
Kraken	*1391*	965	576	*0.414*	*0.597*	0.489
Tess fr	980	965	713	0.728	0.739	**0.733**

Table – Comparaison des résultats de la reconnaissance d'EN avec `spacy_lg` sur différentes versions de *"Le petit chose"*, Daudet, 1868, après alignement avec NERVAL



(a) `SEM` - Kraken



(b) `spacy_lg`

Figure – (a) et (b) Intersections *"Une vie"*, Maupassant, 1883. (c) et (d)Résultats des distances de Jaccard, Bray-Curtis, Dice et Cosinus entre les Entités Nommées "LOC" Réf. et OCR, *"Le petit chose"*, Daudet,1868.

Contributions & Perspectives

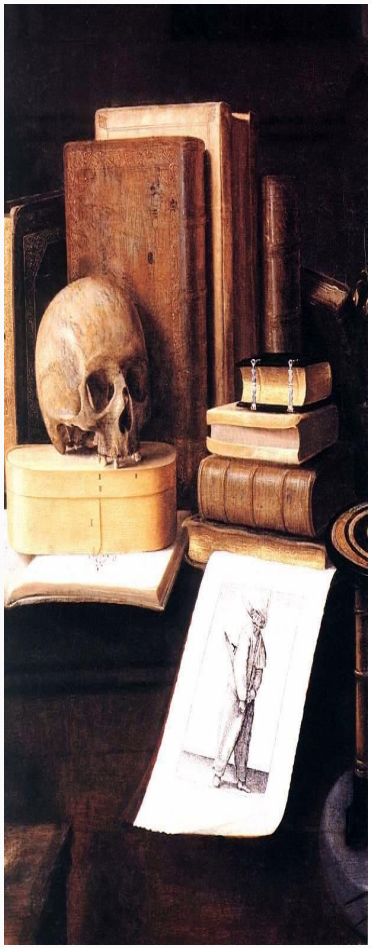
- Toutes les **erreurs d'OCR ne se valent pas :**
  ➜ fautes d'orthographe $\neq$ mots collés ensemble
- La **comparaison** de plusieurs **distances** n'est **pas triviale :**
  ➜ Cos. sous-estime les distances ? Jaccard les surestime ?
- Résolution des **problèmes de liage** (diachronie, formes contaminées) :
  ➜ **distance, cluster,** liage avec des bases de données ou embeddings ?
- La **représentation cartographique** comme **évaluation** :
  ➜ EN récupérée par +ieurs outils est potentiellement un VP ?
- Une géographie littéraire du 19ème siècle

(a) Réf. vert clair : 1, vert moy. : 2, vert foncé : 3.

(b) Tess fra. bleu : 1, violet moy. : 2, violet foncé : 3.

Figure – 3 outils de REN, *"Le petit chose"*, Daudet,1868.

Corina Chutaux



Corina CHUTAUX

Formation
Double Master

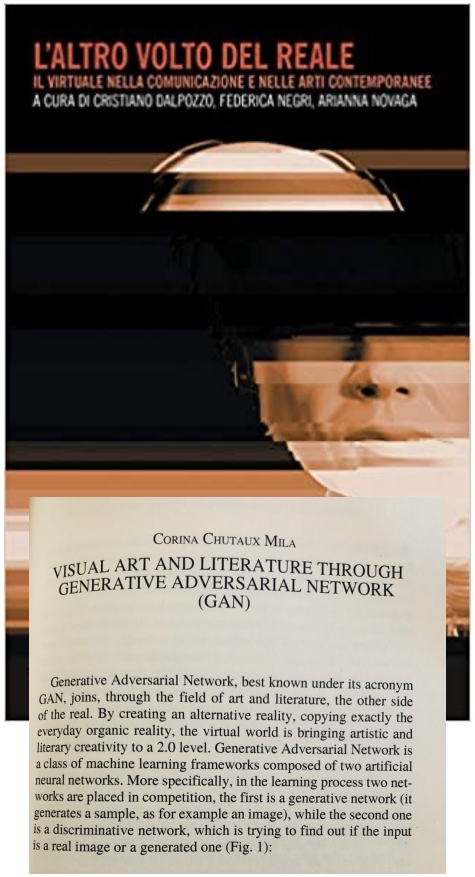
Marché de l'art (ICART) & Littérature Générale et comparée (SORBONNE-NOUVELLE)

Thèse de doctorat (Sorbonne-Nouvelle)
Titre *: Dématérialisation de l'art et de la littérature à l'aube de la digitalisation*

> ***Sujet*** *:* Analyse des œuvres littéraires et artistiques générées par des Intelligences artificielles et les conséquences qui en découlent dans ce que j'ai appelé, « le siècle de la dématérialisation »

Problématiques thèse *:*

Qu'est-ce que la dématérialisation ? Comment l'art et la littérature se délestent de la tradition pour créer une esthétique et une poétique de la dématérialisation ? Annonce-t-elle, l'Intelligence Artificielle la désuétude du statut de l'écrivain et de l'artiste ? Comment analyser la littérarité d'un texte sans auteur humain ? Comment interpréter une œuvre d'art si elle n'est plus issue de l'intention d'un artiste mais qu'elle est le fruit d'une programmation arbitraire ou réfléchie ? Et finalement, comment interpréter cette dématérialisation totale : de l'œuvre, de l'artiste/écrivain, du processus de production et du processus de mise à disposition du public d'un point de vue socio-économique ?



Publications ouvrages

burning base of sky
half the hemisphere high
a fire hour
crept upon by night

back on its golden hinges
the gate of memory swings
and my heart goes into the garden
and walks with the olden things

Bei L., Jianlong F., Makoto P. K., Masatoshi Y., *Beyond narrative descrip- tion: Generating Poetry from Images by Multi-Adversarial Training*, 2018

Collectiv Obvious, *Comte de Belamy, 2018*

Ouvrage collectif, en collaboration avec l'Université de Vérone, paru en 2020

Chapitre intitulé

Visual Art and literature trough Generative Adversarial Networks

Résumé

Analyse des textes littéraires et des œuvres d'art générés par des IA de type GAN et explication du processus de programmation derrière les créations.

Publications ouvrages

Ouvrage théorisant pour la première l'art invisuel : un art sans œuvre d'art, sans spectateur traditionnel et parfois sans auteur,

Ouvrage paru en octobre 2021

Le lancement a eu lieu au Palais de Tokyo

Résumé

Un art sans œuvre d'art, sans auteur et sans spectateur.
Dans l'Invisuel il ne s'agit pas de créer une œuvre mais de faire œuvre, de renoncer au spectateur traditionnel, réduit à sa fonction d'observateur, et exclu par-là complètement de l'activité artistique, pour un « spectateur » qui prend part à la pratique et qui est parfois en osmose avec celle-ci. Il s'établit ainsi une relation de partage entre l'artiste et le spectateur.

Eva Lacroix

EVA SCHAEFFER-LACROIX

Maitresse de conférences HDR (sections 7 et 12)

Département d'allemand de l'Inspé de Paris (école interne de Sorbonne Université)

Ce qui me charactérise

- Recherche à dominante engagée
- Expérience des usagers : observer ce que les personnes font et ce qu'elles en disent
- Outillage des apprentissages et analyses scientifiques
- Création et analyse de corpus numériques

Projets

Mymap

- TADS (Traduction de scripts d'audiodescription) : départ de ma partenaire principale de l'université de Hildesheim (Allemagne) ➔ projet interrompu/abandonné ? (reste le colloque en octobre et un numéro spécial publié ensemble)
- Groupe d'intérêt "corpus multilingues pour soutenir la traduction semi-automatique d'audiodescriptions". Créé ensemble avec Nina Reviers et Anna Jankowska de l'université d'Anvers, ouvert à d'autres personnes (prospection en cours).
- Participation à l'ANR Tractive dirigé par Virginie Juillard : création d'un corpus d'audiodescription pour analyser le *male gaze* (tbc).
- Projet FanTALES (*digital storytelling* ; projet subventionné par Erasmus+ 2017-2020) : développement d'une alternative pour l'outil Twine avec Frederik Cornillie (chercheur à la Katholieke Universiteit de Louvain, Belgique) et d'autres partenaires européens (Allemagne, Espagne) (tbc).
- Projet participatif AudYD

Projet participatif AudYD

Création d'audiodescriptions par des bénévoles à l'aide de YouDescribe.org

- Projet diffusé via le site Science Ensemble
- Participation de trois étudiantes (SHS) + une jeune retraitée (ancienne journaliste de documentaires et conseillère artistique d'une association défendant la cause du handicap mental)
- Deux séances ont eu lieu à distance (juin et juillet 2022) avec sessions filmées
  - ➔ Introduction à l'audiodescription
- Troisième séance cet après-midi à la maison de la recherche
  - ➔ Visite INSEAD (Centre multidisciplinaire des sciences comportementales)
  - ➔ Prise en main de l'outil de création d'audiodescriptions YouDescribe.org
- Quatrième séance (contenu tbc, date tba)
  - ➔ Option A : Tester la faisabilité d'une recherche avec des personnes aveugles à qui on proposera de comparer des audiodescriptions enregistrées par des humains avec celles mises en voix par une voix de synthèse (film *Dans les allées*) ; lieu : locaux de l'INSEAD
  - ➔ Option B : Rédiger une réponse à l'appel à projets ANR Science avec et pour la société – Recherches participatives (deux dates limites possibles : 30 septembre ou décembre 2022). Partenaire envisagé : Association Valentin Haüy (dispose d'un réseau de bénévoles pour le soutien des aveugles, mais pas encore de missions liées à l'accès aux films).

Publications récentes

- (Abstract proposé) Demonstration as a method to teach corpus use. In Fiona Farr & Agnieszka Leńko-Szymańska (Eds.). Special Issue of Second Language Teacher Education: *Corpus Linguistics in Second Language Teacher Education*. Publication planifiée pour 2023.
- (à paraître en octobre 2022). Automatisch erkannte Adverbien und Adverbiale in deutschsprachigen Audiodeskriptionsskripten [Reconnaissance automatique d'adverbes et adverbiaux dans des scripts d'audiodescription allemands]. In Pierre-Yves Modicom (dir.). *Adverbien und Adverbiale im Deutschen – Grenzen und Gliederung einer syntaktischen Kategorie im Deutschen*, 18 pages. Winter-Verlag.
- Berkling, Kay, Roger Gilabert Guerrero, et Eva Schaeffer-Lacroix. 2022. An Overview to Games and Gaming for Native and Foreign Language Learning. *Alsic. Apprentissage Des Langues et Systèmes d'Information et de Communication*, n° Vol. 25, n° 1. https://journals.openedition.org/alsic/6160.
- "Es ist vieles getan, es bleibt vieles zu tun." – Production assistée par corpus d'un discours de fin d'année en cursus LEA allemand. *Alsic (Apprentissage des Langues et Systèmes d'Information et de Communication)*. https://journals.openedition.org/alsic/5900.
- Schaeffer-Lacroix, Eva, et Kirsten Berland. 2022. Dealing with Variation in Audio Description Scripts. *Journal of Audiovisual Translation* 5 (1): 150-65. https://doi.org/10.47476/jat.v5i1.2022.181.

Merci pour votre attention !

- elacroix@inspe-paris.fr
- Orcid
- CV HAL
- Blogue académique

Gael 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
- Humanités Numériques

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
- Humanités Numériques

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
- Humanités Numériques

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 ?

De manière transversale : comment tirer avantage de la variation ?

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-) Memes : Extraction automatique et analyse par Myriadisation d'Expressions Semi-figées (K.Fort, A.Gautier, L.Zhu) → Thèse de Julien Bezançon
- ANTONOMAZ (2018-2022) ANalyse auTOmatique et NumérisatiOn des MAZarinades (K.Abiven, G.Roe, A. Bartz) Thèse de Jean-Baptiste Tanguy
- OBVIL-NER (2020-2023) Humanités Numériques et Entités Nommées (C. Parfait, M. Alrahabi, G. Roe)
- WADDLE et DANIEL (2018-...) Données Textuelles hétérogènes (A. Barbaresi, E. Giguet) et multilingues (E.Boros, A.Doucet) Thèse de Steve Mutuvi
- CERES Centre d'expérimentation en Méthodes Numériques pour les SHS (V. Julliard, C. Marti, C.Guillotel, T. Bottini, E. Bouté, F. Allié) + collaboration GEMASS

Ibtihel Ben Ltaifa

Ltaifa.pdf

Speed Dating Recherche

Équipe de Linguistique Computationnelle

Ibtihel BEN LTAIFA

Docteur en Informatique

ATER à l'UFR de sociologie et d'informatique pour les sciences humaines

Ibtihel.BEN_Ltaifa@paris-sorbonne.fr

Domaine de recherche

- Fouille de Données
- Réseaux Sociaux
- TAL (Traitement du langage naturel)
- Computer vision
- Intelligence artificielle

Mots clés: Extraction de connaissances à partir de données, Indexation sémantique, Annotation sémantique, Recherche d'Information, Apprentissage supervisé/non supervisé, Apprentissage profond

Travaux de recherche menés dans le cadre de la thèse de doctorat

- **Sujet:** A new Representation and Ranking Approaches based on Deep Learning to improve the Semantic Information Retrieval in Microblogs.

- **Contexte**: Recherche d'Information Sociale dans les microblogs

-Retrouver les microblogs répondant à un besoin d'information spécifié par un utilisateur.

-La prise en compte des données des réseaux sociaux (facteurs sociaux) dans le processus de recherche d'information.

Travaux de recherche menés dans le cadre de la thèse de doctorat

Partie « linguistique »

- **La phase de prétraitement**: du texte aux données: Nettoyage, Normalisation des données, Term-Frequency (TF) …

- **La phase d'Extraction d'information**: trouver les descripteurs linguistiques pertinents

  -Exploitation des bases de connaissances pour l'Extraction de connaissances, identifier les relations

  -Identification des entités nommées

  -Word Sense Disambiguation

Travaux de recherche menés dans le cadre de la thèse de doctorat

Partie « apprentissage automatique : des données au modèle

- **Modèles classiques d'apprentissage** *:*

-**Supervisé** (classification) : SVM, Naive Bayes

-**Non supervisé** (clustering): K-means

- **Modèles d'apprentissage en profondeur:** couches de neurones peuvent être perçues comme des extracteurs automatiques de features (feature engineering):

-Réseaux de neurones récurrents (RNN), BiLSTM, réseaux à propagation avant (feed-forward), autoencoder

Travaux de recherche post-thèse

Domaine: Informatique Médicale, TAL, Computer Vision

Tâche 1: Question-Réponses Visuelle (VQA, en anglais) dans laquelle un agent doit répondre à des questions posées sur des images.

- Développer de nouvelles méthodes basées sur des réseaux de neurones artificiels pour améliorer les performances des modèles de VQA.

Travaux de recherche post-thèse

Domaine: Fouille de Données , Informatique Médicale, Computer Vision

Tâche 2: traitement et analyse des images médicales pour l'aide au diagnostic

- Utilisation des méthodes d'analyse basées sur des réseaux de neurones pour:

- transformer les données en informations utiles pour la prise de décision,

-aider les professionnels de santé à identifier les traitements les plus efficaces pour certaines pathologies .

Travaux de recherche post-thèse

Domaine: Fouille de textes, TAL

Tâche 3 : DEFT 2022 (DÉfi Fouille de Textes 2022): Avec l'équipe STIH

- **Stylo@DEFT2022**: Notation automatique de copies d'étudiants par combinaisons de méthodes de similarité:
  - Modèles d'Extraction de caractéristiques: sentence BERT
  - Mesures de similarité: cosinus
  - Notation par régression

Publications

- Journaux Internationaux

-Olfa hrizi, Karim gasmi, Ibtihel Ben ltaifa, Hamoud Alshammari, Hanen Karamti, Moez Krichen, Lassaad Ben Ammar, Mahmood A. Mahmood : **Tuberculosis Disease Diagnosis Based on an Optimized Machine Learning Model**. Journal of Healthcare Engineering (2022)

-Karim Gasmi, Ahmed Kharrat, Ibtihel Ben ltaifa, Moez Krichen, Hamoud Alshammari, Mohamed Osman Abdelhadi, Lassaad Ben Ammar: **Classification of MRI Brain Tumors Based on Registration Preprocessing and Deep Belief Networks**. Journal of Healthcare Engineering (2022)

-Karim Gasmi, Ben Ltaifa Ibtihel, Gaël Lejeune, Hamoud Alshammari, Lassaad Ben Ammar, Mahmood A. Mahmood : **Optimal Deep Neural Network-based model for answering Visual Medical Question**. Cybernetics and Systems (2021): 1-22.

-Ben Ltaifa Ibtihel, Lobna Hlaoua, Lotfi Ben Romdhane: **Hybrid Deep Neural Network-Based Text Representation Model to Improve Microblog Retrieval**. Cybern. Syst. 51(2): 115-139 (2020)

-Conférences Internationales & Ateliers :

-Ben LTAIFA, Ibtihel, BOUBEHZIZ, Toufik, BRIGLIA, Andrea, et al. Stylo@ DEFT2022: **Notation automatique de copies d'étudiant· e· s par combinaisons de méthodes de similarité**. Atelier DÉfi Fouille de Textes (DEFT). ATALA, 2022. p. 11-22.

-Ben Ltaifa Ibtihel, Hlaoua Lobna, Lotfi Ben Romdhane: **A Deep Learning-based Ranking Approach for Microblog Retrieval.** KES 2019: 352-362

-Ben Ltaifa Ibtihel, Hlaoua Lobna, Maher Ben Jemaa: **A Semantic Approach for Tweet Categorization**. KES 2018: 335-344

Julien Bezancon

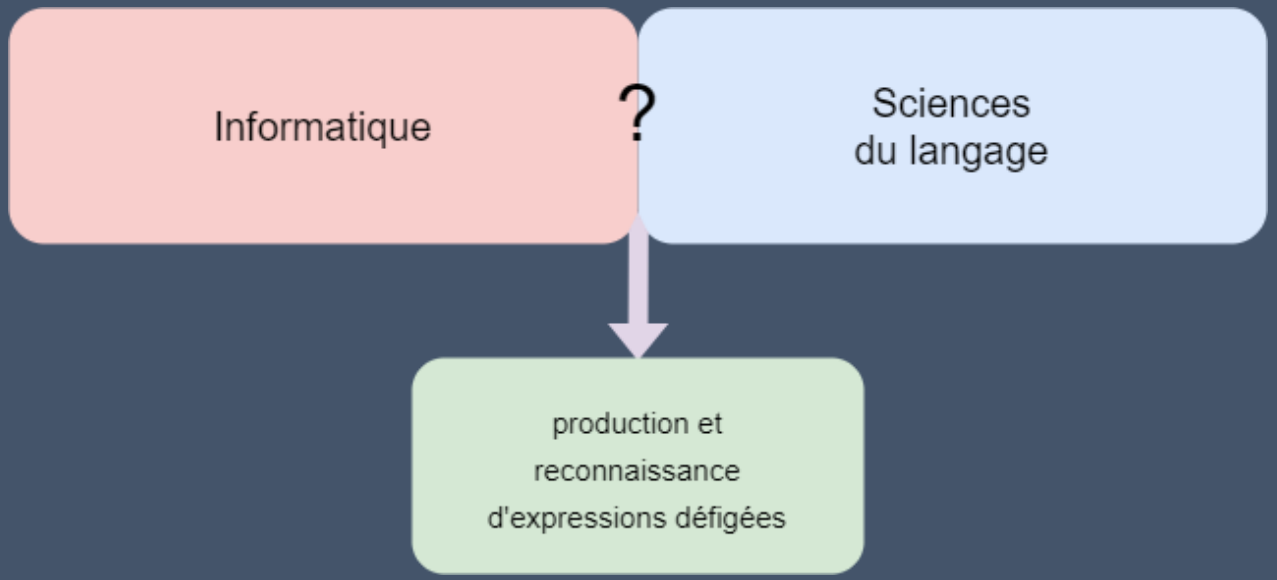
SORBONNE UNIVERSITÉ

Sujet de thèse
C.E.R.E.S

Détection et production de défigements linguistiques dans les réseaux sociaux assistées par les sciences participatives : fertilisation croisée entre traitement informatique et analyse linguistique

Julien Bezançon
encadrement : Gilles Siouffi, Antoine Gautier et Gaël Lejeune

Sujet de thèse



Sujet de thèse

@User ben oui, c'est le deuxième effet kiss pascool

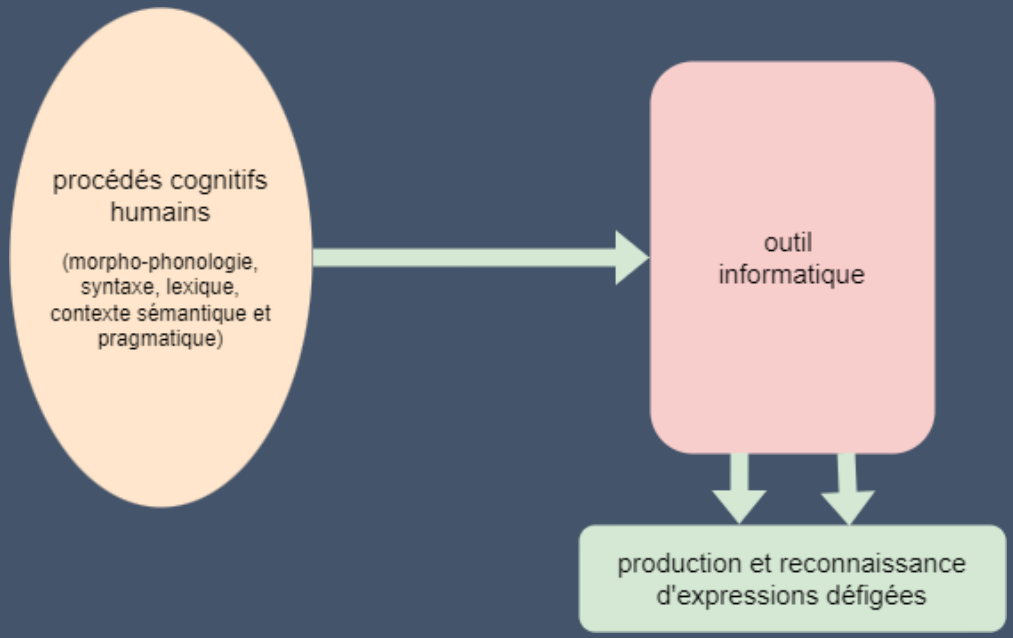
Sujet de thèse

@User Que la force qui a renversé la finale de coupe de France 2019 soit avec toi !

@User Que la chance soit avec toi 🍀🤞😎

@User Que la force, avec toi soit 😎

Sujet de thèse



Karen Fort

Création de ressources langagières et éthique pour le TAL

Karën Fort

karen.fort@sorbonne-universite.fr / https://members.loria.fr/KFort/

Séminaire d'équipe, Speed dating, 8 septembre 2022

Loria
Laboratoire lorrain de recherche en informatique et ses applications

LETTRES SORBONNE UNIVERSITÉ

Production participative (*crowdsourcing*)

Portails de science participative et de jeux pour les langues :

Atelier récurrent : Games4NLP (LREC 2018-2020-2022)

Éthique et TAL

- Création de données pour un TAL plus éthique :
  - corpus du français pour l'évaluation des biais stéréotypés des modèles de DL [Névéol et al., 2022]
  - corpus de dossiers médicaux synthétiques en français [Hiebel et al., 2022]
  - Extension de CrowsPairs pour l'allemand, le maltais, l'italien et l'espagnol (M. Mieskes, C. Borg, Sergio Zanotto et Wolfgang Sebastian Schmeisser Nieto)
- Éthique déontologique :
  - Évaluation empirique de l'application de la #RègleDeBender [Ducel et al., 2022]
  - Proposition d'une nouvelle description de la participation de l'humain dans les systèmes [Anderson and Fort, 2022]
  - Analyse des sections éthiques de la conférence NAACL 2021 (E. Bender, M. Mitchell et E. van Miltenburg)
  - Analyse de l'impact des BigTech sur le TAL (F. Ducel, A. Névéol, S. Mohammad, M. Abdalla, T. Lima Ruas et J-P. Wahle)

Projets en cours

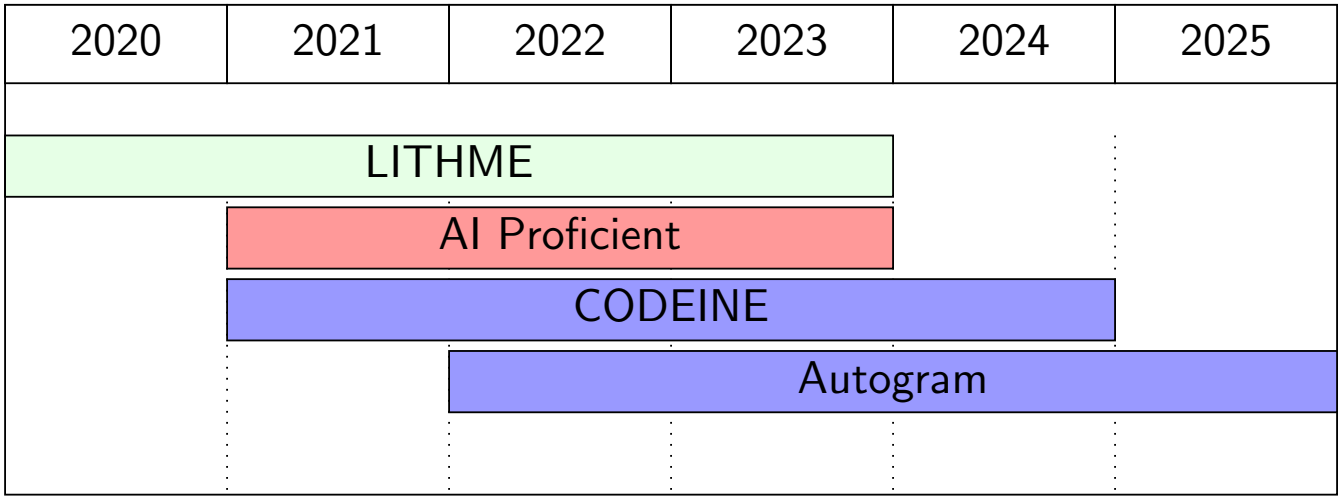


FIGURE – Projets en cours (en vert les actions COST, en rouge les projets européens, en bleu les projets ANR)

Doctorée / Doctorants

- Alice Millour : thèse soutenue en 2020, aujourd'hui MCF à Paris 8.

- Heesoo Choi (avec Mathieu Constant et Bruno Guillaume) : *Lier des ressources lexicales du français en vue d'une interopérabilité entre niveaux linguistiques*. (Univ. de Lorraine).
- Nicolas Hiebel (avec Aurélie Névéol et Olivier Ferret) : *Création éthique de données textuelles artificielles : application au domaine biomédical*. (CODEINE).

Responsabilités

- Co-chair du comité d'éthique d'ACL (Association for Computational Linguistics)
- Conseil national des universités (CNU) 27 (informatique) : https://cnu27.univ-lille.fr/
  - Qualifications
  - Promotions, CRCT, etc
- GDR :
  - LIFT (linguistique informatique, formelle et de terrain) : École d'été sur l'annotation en 2022 (à venir en 2024)
  - TAL : École d'été en 2023
- Actions européennes COST :
  - Language In The Human-Machine Era (LITHME)
  - UniDive (démarre)


Anderson, M. and Fort, K. (2022).
Human where ? a new scale defining human involvement in technology communities from an ethical standpoint.
IRIE - International Review of Information Ethics, 31.


Ducel, F., Fort, K., Lejeune, G., and Lepage, Y. (2022).
Do we name the languages we study ? the #benderrule in LREC and ACL articles.
In Proceedings of International Conference on Language Resources and Evaluation (LREC) 2022, Marseille, France. European Language Resources Association (ELRA).


Hiebel, N., Ferret, O., Fort, K., and Névéol, A. (2022).
CLISTER : A Corpus for Semantic Textual Similarity in French Clinical Narratives.
In
LREC 2022 - International Conference on Language Resources and E
Marseille, France.


Névéol, A., Dupont, Y., Bezançon, J., and Fort, K. (2022).

French crows-pairs : Extending a challenge dataset for measuring social bias in masked language models to a language other than english.
In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Irlande.

Laurie Acensio

Présentation 1/3

- **Laurie Acensio**
- **Formation**

  -Master TAL (Université Paris-Sorbonne)

  -Doctorat Informatique (Université Lille)

  Equipe NOCE (Luigi Lancieri/Frédéric Hoogstoel)

- **Thèse**

  -Système d'aide à la gestion et planification des groupes de formation en formation continue (Juin 2021)

  -Partenariat Public/Privé

Activités scientifiques 2/3

- **Domaines de recherche**

-Informatique : aide à la décision multicritère, optimisation combinatoire, analytique de données (RH, formation), base de données (Sql/NoSql)

-Science de l'éducation : dynamique du groupe de formation, adulte apprenant

Objectif : semi-automatiser la composition des groupes de formation homogènes

-**TAL :** Analyse des questions ouvertes pour le paramétrage de l'AG

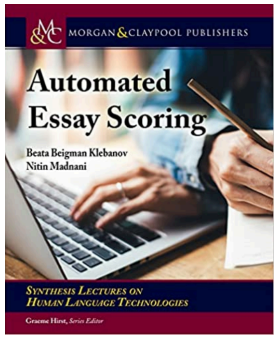
Analyse thématique, lexicale, assignation des réponses selon une polarité positive/négative

Activités scientifiques 3/3

- **Intérêts de recherche**

-TAL : Linguistique de corpus, recherche d'information sémantique (annotation, analyse des relations…)

-Formation/Education : Analyse des contenus (interactions en ligne *via* Mooc), système de notation automatisée (ou *Automated Essay Scoring*)



Apprentissage automatique/TAL :

-Evaluation multidimensionnelle de la qualité de l'essai (grammaire, argumentation, cohérence, structuration…)

-Rétroaction pédagogique

Klebanov, B. B., & Madnani, N. (2021). Automated Essay Scoring. *Synthesis Lectures on Human Language Technologies*, *14*(5), 1-314.

Nour El Houda N E

Séminaire de l'équipe Linguistique Computationnelle

Nour El Houda BEN CHAABENE

Docteure en Informatique de l'Institut Polytechnique de Paris

Nour-El-Houda.Ben_Chaabene@paris-sorbonne.fr

08/09/2022

Axe de recherche

Détection d'utilisateurs violents et de menaces dans les réseaux sociaux

Mots-clés : Intelligence Artificielle, Machine Learning, Deep Learning, Apprentissage Supervisé/Non-supervisé, Réseau de Neurones Artificiels, Réseau de Neurones Convolutionnels, Extraction de Features, Analyse des Réseaux Sociaux, Détection des Anomalies, Traitement du Langage Naturel, Détection des Communautés

Défis

- **QR 1 :** Quel est l'impact sur la détection des comportements anormaux en combinant la topographie du réseau et les activités d'un individu ?
- **QR 2 :** Comment exploiter les graphes multidimensionnels et les communautés pour la modélisation des comportements anormaux ?
- **QR 3 :** Comment considérer l'évolution des comportements dans le temps ainsi que la dynamité du réseau social ?
- **QR 4 :** Comment exploiter les différents types de données pour garantir l'extraction d'une information pertinente et complète ?
- **QR 5 :** Comment extraire des données réelles de plusieurs réseaux sociaux ? Et comment les synchroniser afin de garantir une modélisation multidimensionnelle ?

Contributions

Contribution 1	**Modèle de détection et de prédiction des comportements anormaux sur Twitter**
Contribution 2	**Méthode de détection des comportements anormaux sur la base de l'analyse des relations dans une structure multidimensionnelle**
Contribution 3	**Framework hybride de détection des comportements anormaux sur un réseau multidimensionnel utilisant des données multimodales**

Perspectives

- Appliquer les fonctionnalités et les critères extraits des réseaux sociaux dans d'autres domaines liés à l'évaluation des sources tels que la fraude, le spam et les rumeurs.

  ➥ Génération des réseaux synthétiques avec des caractéristiques du monde réel afin d'étudier leur évolution et leur influence

- Etudier le comportement des groupes plutôt que les individus en détectant des changements dans l'évolution de l'activité d'un groupe.

  ➥ Définition des concepts de base de l'évolution de l'activité des communautés en appliquant des caractéristiques historiques

Publications

- **N.E.H Ben Chaabene,** and R. Guetari. *Semantic Annotation for the "on demand graphical representation" of variable data in Web documents.* In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 3417–3422, **2016**.
- **N.E.H Ben Chaabene,** and M. Mallek. *Learning statistics from raw text documents.* In 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT), pages 321–326, **2018.**
- **N.E.H Ben Chaabene**, A. Bouzeghoub, R. Guetari, S. Balti, and H. Hajjami Ben Ghezala. *Detection of users' abnormal behavior on social networks*. In International Conference on Advanced Information Networking and Applications (AINA), Advanced Information Networking and Applications, volume 1151, pages 617–629, **2020.**
- **N.E.H Ben Chaabene**, A. Bouzeghoub, R. Guetari, and H. Hajjami Ben Ghezala. *Deep learning methods for anomalies detection in social networks using multidimensional networks and multimodal data: a survey*. Multimedia Systems, pages 1–11, **2021.**
- **N.E.H Ben Chaabene**, A. Bouzeghoub, R. Guetari, and H. Hajjami Ben Ghezala. *Applying Machine Learning Models for Detecting and Predicting Militant Terrorists Behaviour in Twitter*. In IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 309–314, **2021.**
- **N.E.H Ben Chaabene**, A. Bouzeghoub, R. Guetari, and H. Hajjami Ben Ghezala. *New Deep Learning Framework for Detecting the Behavior of a Terrorist Group on a Multidimensional Network Using Multimodal Data.* Expert Systems With Applications, **2022.**