Séminaire de rentrée : Recontres Minute Science

09 septembre 2021

STIH, Sorbonne University

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

Alexandre Bartz

Alexandre Bartz

Sorbonne Université

9 septembre 2021

Projet Antonomaz

- Master TNAH (*Technologies numériques appliquées à l'histoire*) de l'École nationale des chartes (Paris)
- Depuis avril 2021, ingénieur du projet Antonomaz

Antonomaz

ANalyse au**TO**matique et **N**umérisati**O**n des **MAZ**arinades

- Exploiter un corpus d'environ 5.000 libelles rédigés au milieu du XVIIe siècle, sous la Fronde
- But : analyser automatiquement ce corpus et le mettre à la disposition des chercheuses et chercheurs

Mes missions

- Encodage des documents en XML-TEI, contrôle qualité grâce à des schémas d'encodage (ODD)
- Travail sur les métadonnées
- Mettre à disposition les données : publication sur Github, aussi bien de la chaîne de production que des textes encodés

La suite

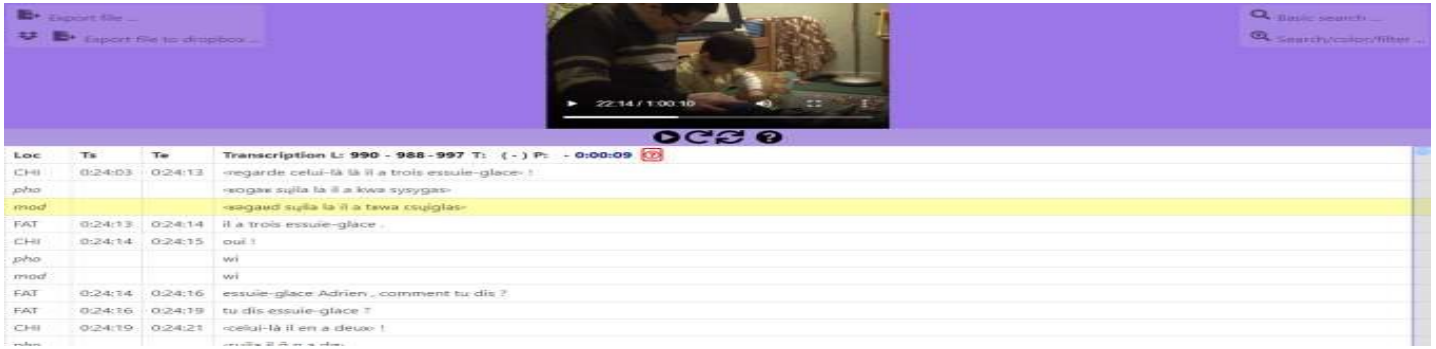
- Rendre les données textuelles produites exploitables et interrogeables par tout.e.s : lemmatisation, annotation morphosyntaxique, normalisations et *NER* (en collaboration avec l'INRIA, Simon Gabay et Philippe Gambette)
- Proposer une application web pour mettre les données en valeur
- Projet secondaire : reconnaissance automatique de l'imprimeur grâce à la casse utilisée (voir `https://github.com/Imprinters` en collaboration avec Simon Gabay)

Andrea Briglia

Andrea Briglia, PhD en Linguistique, 10/2017- 3/2021 (cotutelle Messina & Montpellier)

« Approches statistiques et computationnelles à l'acquisition du français L1 chez l'enfant »

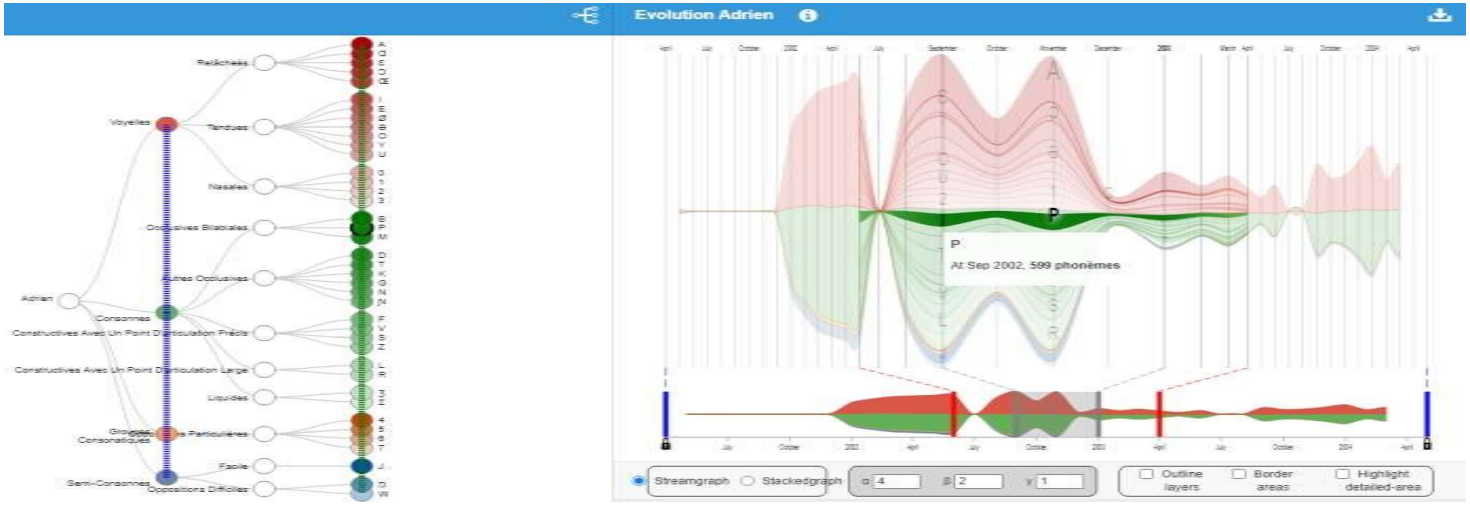
Données CoLaJE (ANR, Morgenstern & Parisse, 2012). 6 enfants enregistrés une heure par mois, tous les mois, d'un an jusqu'à cinq ans. Transcriptions en trois types différents.



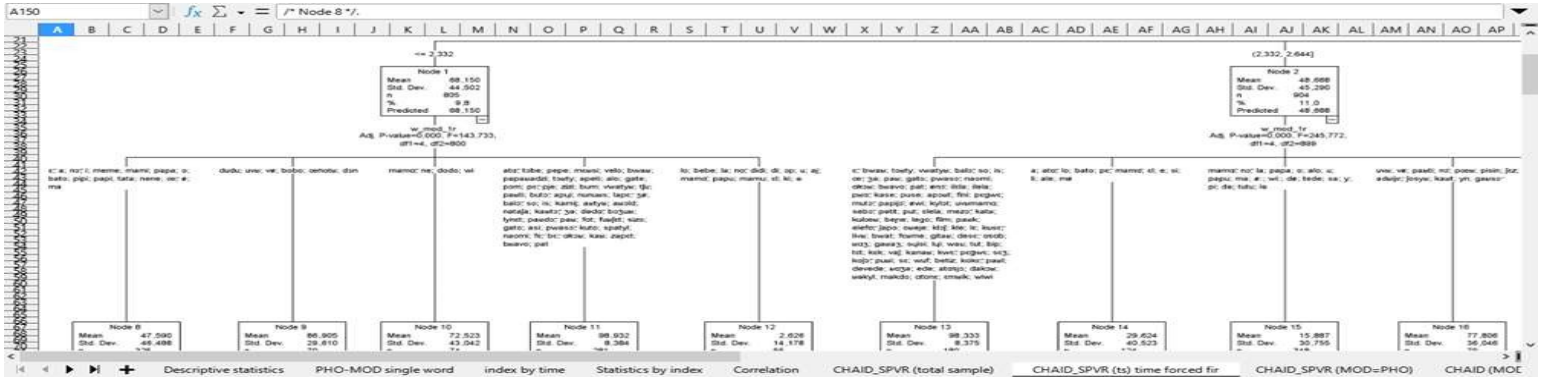
En total, environ 8'000 énoncés pour chaque enfant. Test de représentativité statistique (capture rate) 1%

HP Suite des variation phonético/phonologiques, étapes acquisition du fr L1, tendances et contraintes, comparaison *intra*-enfant et *inter*-enfants. Deux niveaux d'analyse : phonétique et syntaxique.

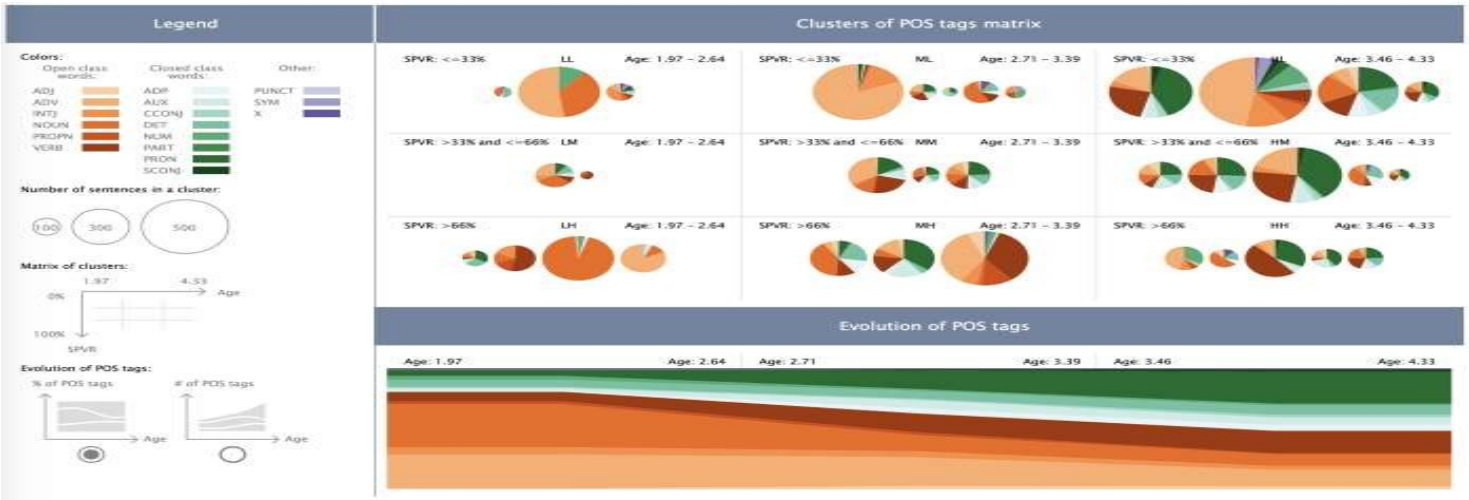
Méthodes Chi2, POS tagging (stanza, UD), clusterisation avec EM, visualisation avec Multistreamgraph, modélisation et réproduction avec un CNN



https://marine27.github.io/TER/site_aquisition_du_langage/multistream.html réf Cuenca et al. « Multiresolution streamgraph approach » IEEE Transactions on visualisations and computer graphics, 2018



https://hal.archives-ouvertes.fr/tel-03319126



http://advanse.lirmm.fr/EMClustering/

Réseau neuronal de type convolutionnel : https://colab.research.google.com/drive/1fIa0ak1k-yWFmsCx1FZpl6VdYEY_PpwS

Limites données bruitées, prosodie et non-verbal, langage adulte, distinction attaque/coda et coarticulations, difficulté de représenter la phonologie en utilisant des séquences des caractères.. et en plus, on constate un grande variabilité entre les enfant (exceptions, retour en arrière..)

Conclusions « n'importe quelle variation ne varie pas en n'importe quelle autre » (Sauvage, 2015), cadre intérpretatif ; théorie des traits phonologique de N. Clements.

- /gR/ → /dR/
- /dR/ → /gR/
- /tR/ → /kR/
- /kR/ → /tR/
- « 52 Q : ben attends, on essaie de/ de l'**touver**, si on le **trouve** pas ze/ ben c'est pas grave hein, ça c'est un gros euh c'est bien… si on le **krouve** pas, alors c'est pas grave…donc, [..] »
- Jérémi « c'est [dra] ça ? » prononcé exprès à la place de « c'est grand ça ? » dans le but d'étudier la réaction de l'enfant, qui répond ainsi « on dit pas [dra], on dit [dra] !! ». (Sauvage, 2015)

Là, on trouve toute la difficulté d'étudier la conscience phonologique en cours de développement

Caroline Parfait

Reconnaissance des Entités Nommées (REN) spatiales dans un corpus littéraire : robustesse des systèmes existants ?

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

Speed Dating (Vos travaux de recherche en 3mn), 2021

9 Septembre 2021

caroline.parfait@sorbonne-universite.fr
gael.lejeune@sorbonne-universite.fr

REN : Vocabulaire & application concrète

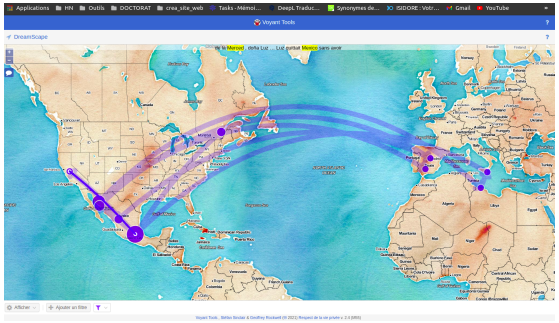


Figure: Voyant Tools, *"Les trappeurs de l'Arkansas (4e édition) "*,Gustave Aimard, 1858.

- Humanités Numériques (HN)
- Entités nommées spatiales (ENS)= Toponymes
- Reconnaissance Optique de Caratères (OCR)
- Bruit(s) :
  - OCR = Ajout, transformations de caractères, autres (?)
  - REN = NON ENS
- Silence(s?) :
  - OCR = Suppression de caractères (?)
  - REN = NON-Reconnaissance ENS (VP, FP, FN, etc.)

Problématique(s) et Méthode

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

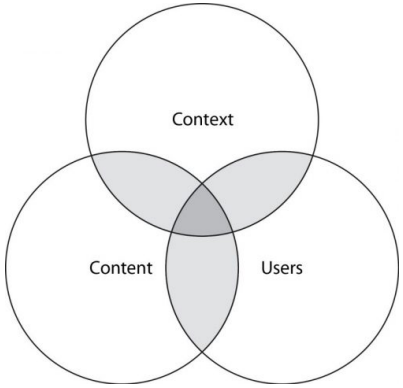


Figure: Peter Morville, *"Three Circles of Information Architecture"*, 2004

- **Utilisateurs** : chercheurs SHS, HN, TAL ... Interdisciplinarité
- **Données** : Corpus ELTeC (Romans Français 19/20ème s.)
- **Variabilités** : diachronie, diatopie, qualité de la transcription OCR ?

Réalisations 2020/2021

- Enquêtes **Utilisateurs** quantitative et qualitative avec J-B. Tanguy
- **Évaluation des systèmes** SPACY et STANZA
- **Angle Expé.** : "Impact du bruit des transcriptions OCR sur la REN ?"

Exemples :

Version	**Context**	**Output**
Ref.	*prendre la diligence de Châlons*	Châlons
Kraken	*prendre la diligence de Ch_lons*	Ch_lons
Tess	*prendre la diligence de Chalons*	Chalons
Tess fr	*prendre la diligence de Châlons*	Châlons

Eva Lacroix

Eva Schaeffer-Lacroix

Maitresse de conférences HDR (sections 7 et 12)

Inspé de Paris (Sorbonne Université)

elacroix@inspe-paris.fr
http://didaktik.hautetfort.com/

Mon parcours

- Magister artium Germanistik & Romanistik, Universität Konstanz (1992)
- Professeure agrégée d'allemand (1994-2011)
- Master en didactique des langues et cultures, Paris 3 (2004-2005)
- Thèse en sciences du langage, sections 7 et 12, Paris 3 (2006-2009) :

  Corpus numériques et production écrite en allemand langue étrangère. Une recherche avec des apprenants d'allemand.
- Maitresse de conférences à l'Inspé de Paris depuis 2011
- Habilitation en sciences du langage, sections 7 et 12, Université de Strasbourg (2020) :

  Recherches et pratiques outillées en allemand et français - Didactique, linguistique de corpus et traduction.

Mes recherches

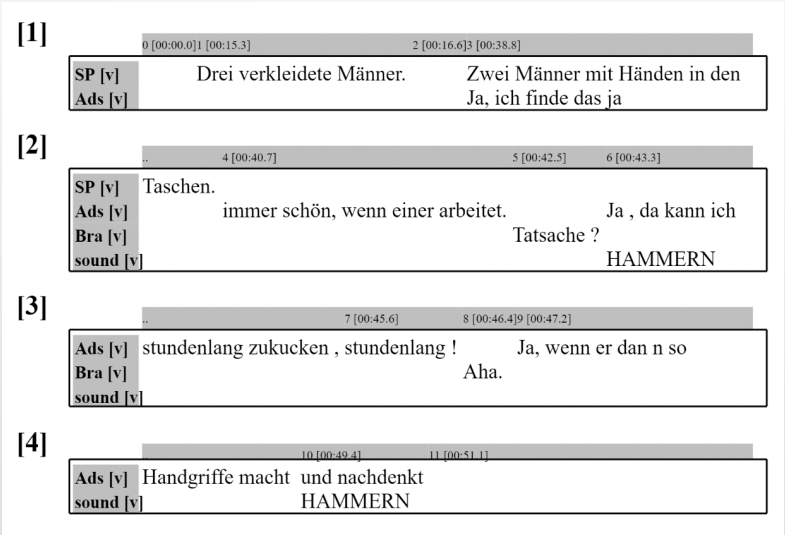
- Mymap
- Recherche à dominante engagée
- Expérience "usager" : observer les actions des personnes et ce qu'elles en disent
- Outillage des apprentissages et analyses
- Collaboratrice scientifique au département d'allemand de l'université de Genève : linguiste de corpus & projet FNS (Fonds national suisse) sur la politique de la langue en Suisse (mars 2015 – octobre 2016)
- Initiatrice d'une collaboration franco-allemande ayant mené à un dépôt de projet ANR-DFG sur la traductibilité de scripts d'audiodescription en mars 2021

Projet *TADS* (Traduction de scripts d'audiodescription)

Annotation XML-TEI

element	***tag* XML-TEI**	**exemple *Buettenwarder* (épisode 12)**
indications de temps	<time>	10: 06: 50
parties à enregistrer (*speaker*)	<sp>	Brakelmann fährt durchs Dorf. [Brakelmann traverse le village.]
instructions concernant les actions à faire ou à éviter pendant l'enregistrement	<stage>	(Rest übersprechen) [Chevaucher le reste]
débit	<stage type="delivery">	s, ss, n [rapide, très rapide, normal]
légendes qui s'affichent à l'écran (titres, textes de pancartes, etc.)	<caption>	Schild am Straßenrand. "Goethe Eier 5 Euro 10" [Panneau au bord de la route. "Oeufs à la manière de Goethe 5 euros 10]
(fins de) répliques	<prompt>	"Goethe gehört mir!" [Goethe est à moi !]

EXMARaLDA



Nous recherchons une personne prête à nous aider à créer une feuille de style XSL afin de pouvoir associer les vidéos et les fichiers XML, annotés (entre autres) avec des indications de temps.

Ibtihel Ben Ltaifa

Speed Dating Recherche

Équipe de Linguistique Computationnelle

Ibtihel BEN LTAIFA

Doctor en Informatique

Ibtihel.BEN_Ltaifa@paris-sorbonne.fr

Domaine de recherche

- Fouille de Données
- Réseaux Sociaux
- Intelligence artificielle

Mots clés : Traitement du langage naturel, Extraction de connaissances à partir de données, Indexation sémantique, Annotation sémantique, Classification, Clustering, Recherche d'Information, Détection de communautés, Apprentissage supervisé/non supervisé

***Cybernetics and Systems* 2020** (Impact Factor = 1.880)

[J1] Ben Ltaifa Ibtihel, Lobna Hlaoua, and Lotfi Ben Romdhane. **Hybrid deep neural network-based text representation model to improve microblog retrieval**. Cybernetics and Systems, 51(2):115–139, 2020.

KES 2019

[C2] Ben Ltaifa Ibtihel, Hlaoua Lobna, and Ben Romdhane Lotfi. **A deep learning-based ranking approach for microblog retrieval**. *Procedia Computer Science, 159:352–362, 2019*.

KES 2018

[C1] Ben Ltaifa Ibtihel, Hlaoua Lobna, and Ben Jemaa Maher. **A semantic approach for tweet categorization**. *Procedia Computer Science, 126:335–* 344, 2018.

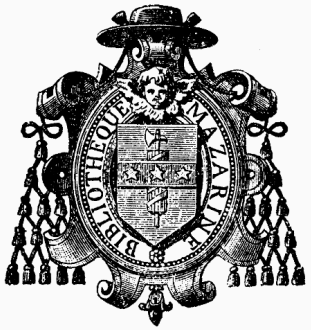
-Approche linguistique pour l'extraction des connaissances dans un texte Arabe en utilisant l'apprentissage automatique profond.

Jean Baptiste Tanguy

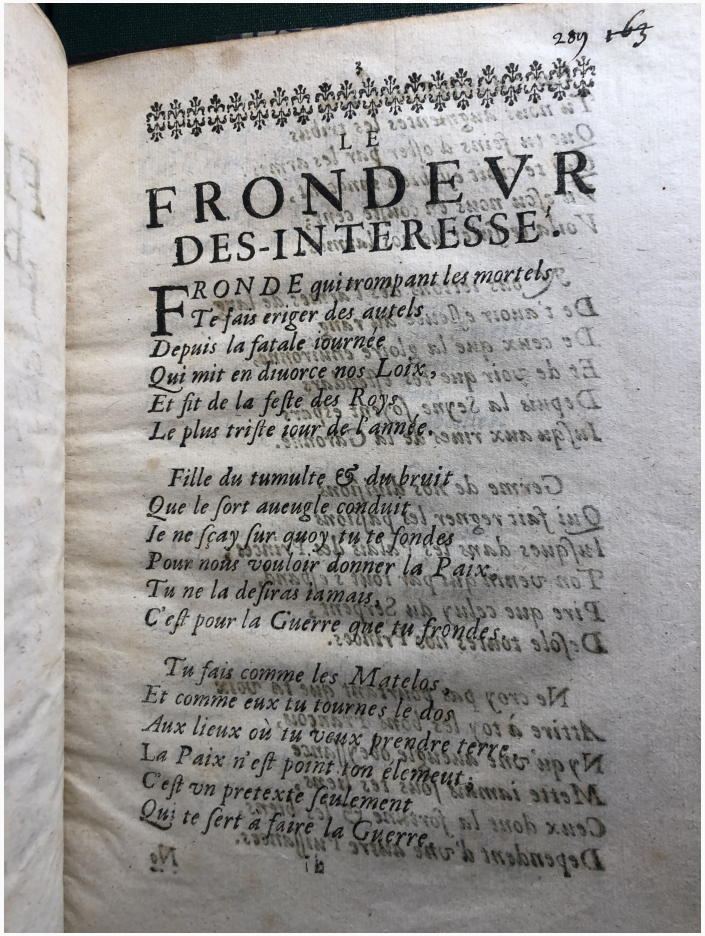
Une thèse en humanités numériques

Du bruit dans les données textuelles océrisées : mesure, impact et évaluation. Le cas des mazarinades (1648-1653).

Directeur : Glenn Roe
Encadrants : Karine Abiven et Gaël Lejeune
Partenaire : Bibliothèqe Mazarine
Financement : Région Île-de-France



Un premier objet d'étude : le "corpus" des mazarinades



LE

FRONDEVR

DES-INTERESSE'.

FRONDE qui trompant les mortels
Te fais eriger des autels
Depuis la fatale iournée
Qui mit en diuorce nos Loix,
Et fit de la feſte des Roys
Le plus triſte iour de l'année.

Fille du tumulte & du bruit
Que le ſort aueugle conduit
Ie ne ſçay ſur quoy tu te fondes
Pour nous vouloir donner la Paix.
Tu ne la deſiras iamais,
C'eſt pour la Guerre que tu frondes.

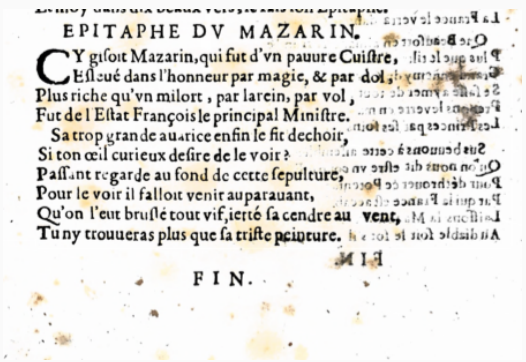
Tu fais comme les Matelots
Et comme eux tu tournes le dos
Aux lieux où tu veux prendre terre.
La Paix n'eſt point ton element;
C'eſt vn pretexte ſeulement
Qui te ſert à faire la Guerre.

- Pamphlets, poésies burlesques, actes légaux, etc., imprimés entre 1648 et 1653 (La Fronde)
- Ensemble d'environ 6000 pièces (+ existence d'éditions et d'états)
- Français pré-classique (ex. : u=v, i=j, présence du s long, etc.)
- Exemplaires datant du XVIIème siècle, donc : tâches, restaurations, etc.

Environ 500 pièces sélectionnées pour la première vague de numérisation

Acquisition automatique des données textuelles (OCR) sans correction

OCR = reconnaissance optique de caractères $\neq$ HTR

Image	ABBYY (payant)	Kraken (gratuit)
	EPITAPHE DV MAZARIN. CY gifoitMazarin.quifutd'vn pauure Cuiftre, dii.il sua aul^ Efleué dans l'honneur par magic, ÔZ par doi Plus riche qu'vn milort, par larcin, par vol â; 5 ? Fut de l Ellat François le principal Miniftre. ' Sa trop grande auance enfin le ht déchoir, 1 Si ton œil curieux defire de le voir? .-j.in shs-j émouuüdtù? • Paffiint regarde au fond de cette fepulturé, nv ' U?<Ln" Pour le voir il falloir venir auparauant, >f, n . Qu'on l'eut bruOé tout vif,iette fa cendre au venc^ .d.ii.i indÉia Tuny trouueras plus que fa trifte peinture, h < :oï si nul jldsibuA FIN. MH.	EPITAPHE DV MAZARIN. CY gifoit Mazarin, qui fut d'vn pauure Cuiftre, Cfleué dans l'honneur par magie, & par dol, Plus riche qu'vn milort, par larein, par vol, Fut de l'Eftat François le principal Miniftre. Sa trop grande auarice enfin le fit déchoir, Si ton œil curieux defire de le voir? Paſſant regarde au fond de cette ſepulture, Pour le voir il falloit venir auparauant, Qu'on l'eut bruſlé tout vif, ietté ſa cendre au vent, Tu ny trouueras plus que ſa triſte peinture. FIN.

- *Observer* le bruit $\rightarrow$ mesurer le bruit
- Comprendre ce sur quoi les logiciels d'OCR buttent
- Erreurs d'OCR et évaluations extrinsèques : dépendance à la tâche, au modèle, aux données ?
- De telles erreurs excluent-elles toute étude en HN ? (ressenti vs. réel)

Lejeune Gael

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 (2019-2021) Memes : Extraction automatique et analyse par Myriadisation d'Expressions Semi-figées (K.Fort, A.Gautier, L.Zhu) $\rightarrow$ Thèse à Venir
- ANTONOMAZ (2018-2022) ANalyse auTOmatique et NumérisatiOn des MAZarinades (K.Abiven, G.Roe , JB.Tanguy, A. Bartz)
- 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 (S. Mutuvi, E.Boros, A.Doucet)
- CERES Centre d'expérimentation en Méthodes Numériques pour els SHS (V. Julliard, C. Marti, C.Guillotel, T. Bottini, E. Papinot, F. Allié) + collaboration GEMASS

Manuela Yapomo

Clustering de textes multilingues pour l'extraction de néologismes

Manuela Yapomo

STIH - Sens Texte Informatique Histoire

09 septembre 2021

Contexte

- Création semi-manuelle d'un petit corpus de textes multilingues et multithématiques structuré ;
- Élargissement de ce corpus au moyen du clustering ;
- Extraction de néologismes (candidats) et leurs traductions (candidates) à partir des clusters formés.

Extraits d'articles comparables (biogaz – Wikipédia)

fr

Le biogaz est le gaz produit par la fermentation de matières organiques animales ou végétales en l'absence d'oxygène. Cette fermentation appelée aussi méthanisation se produit naturellement (dans les marais) ou spontanément dans les décharges contenant des déchets organiques, [...].

en

Biogas typically refers to a mixture of gases produced by the breakdown of organic matter in the absence of oxygen. Biogas can be produced from regionally available raw materials such as recycled waste.

de

Biogas ist ein brennbares Gas, das durch Vergärung von Biomasse jeder Art entsteht. Es wird in Biogasanlagen hergestellt, wozu sowohl Abfälle als auch nachwachsende Rohstoffe vergoren werden.

Motivation scientifique de la recherche

- Vérification de la plus-value du clustering pour l'extraction de néologismes bilingues dans un domaine ;
- Contribution à la science terminologique : documentation des phénomènes néologiques de ce domaine de spécialité ; contraste de langues.

Nicolas Hiebel

Stagiaire M2 et futur doctorant

Nicolas Hiebel

9 septembre 2021

Présentation

- Master Langue et Informatique en Sorbonne
- Stage au Laboratoire Interdisciplinaire des Sciences du Numérique (LISN)
- Encadrement
  - Aurélie Névéol (LISN)
  - Olivier Ferret (CEA)
  - Karën Fort (LORIA / STIH)
- Thèse prévue à partir d'octobre

Travaux en cours

Stage M2 : Similarité phrastique dans des corpus médicaux

Problématiques

- Similarité
- Domaine médical

Thèse en octobre : Création éthique de données textuelles artificielles : application au domaine biomédical

Nikola Lakovic

Thèse de Nikola Lackovic

≪ De la voix à la génération de métadonnées sur la relation client : Application à l'amélioration de la satisfaction client ≫

(Thèse CIFRE Malakoff Humanis, Dir. C. Montacié)

- Actes de Dialogue appliqués à la reconnaissance de la parole
- Combiner les informations du texte (retranscription) et de la parole (paralinguistique)
- Objectif 1 : Modèle améliore dé reconnaissance pour le français
- Objectif 2 : Meilleure "collaboration" TAL/TAP

Solveig Poder

Séminaire de rentrée : linguistique computationnelle

Clustering de corpus et topic modeling

Objectifs de la méthode

- Alternative aux outils de textométrie
- Pistes d’analyses pour corpus diachroniques
- Méthode non supervisée et généraliste
- Visualisation claire des résultats

Données

- Deux corpus d’articles de presse (Europresse)
  - Affaire Sackler (crise des opioïdes)
  - PMA
- Un corpus littéraire : 681 Mazarinades

- Formatage des données :
  - Conversion en JSON (scrapping de HTML)
  - Suppression des doublons (Levenshtein)

Méthode

- **Clustering** : propagation d’affinité
- **Topic modeling** : Latent Dirichlet Allocation
- Représentation des données :
  - Plusieurs types de vecteurs (fréquences, TF-IDF)
  - Plusieurs tailles de n-gram (1 à 3 mots)
- Clustering par périodes de 1 à 3 ans
- 1 cluster --> 5 descripteurs (n-grams)

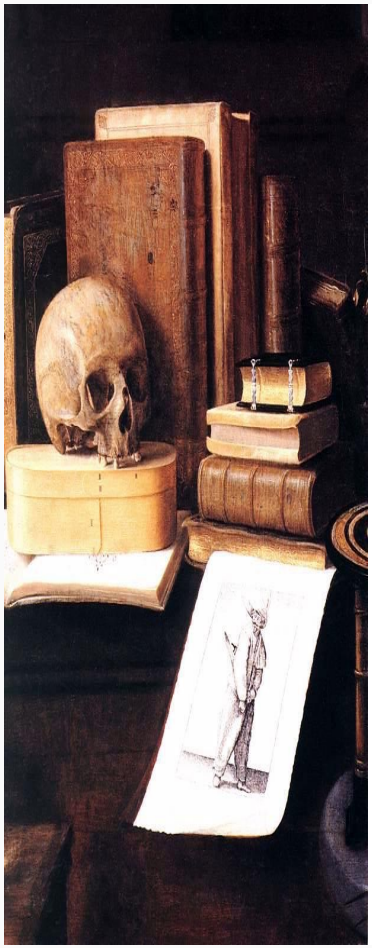
Visualisation (API Flask)

https://clustertool.herokuapp.com/

Formulaire :

Visualisation (API Flask)

Corina Chutaux



Corina CHUTAUX MILA

Formation
Double Master

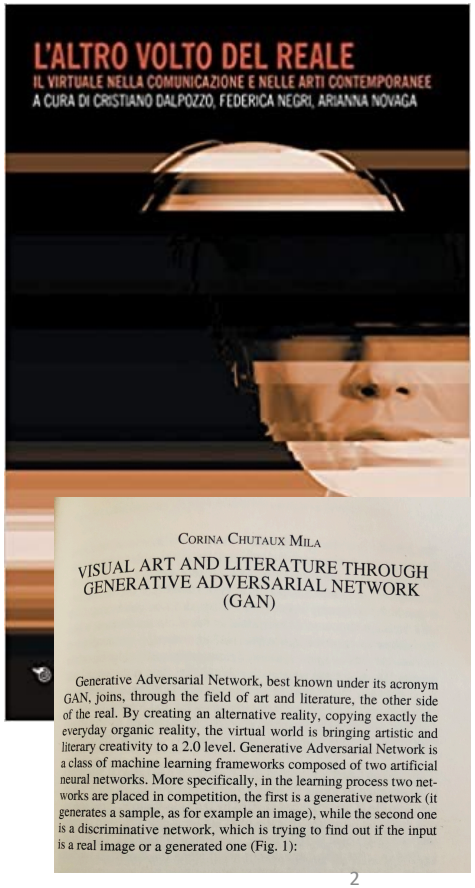
Marché de l'art (ICART) & Littérature Générale et comparée (SORBONNE-NOUVELLE)

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 »

Publications

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



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

Chapitre intitulé *Visual Art and literature trough Generative Adversarial Networks,*

Expériences professionnelles antérieures dans l'enseignement

ENDA

Cours :
- Histoire de l'art
- Art Invisuel vs Art transhumaniste

UNIVERSITÉ SORBONNE NOUVELLE PARIS 3

Cours :
- Méthodologie de la Recherche documentaire
- Communication digitale

SciencesPo. PARIS

Cours :
- Art, Littérature et Politique du 18e au 21e siècle
- Enseignante-Référente

Heesoo Choi

Séminaire STIH - Linguistique computationnelle

Hee-Soo Choi

9 septembre 2021

Qui je suis

- Précédemment en Licence Langue Française et Techniques Informatiques (LFTI)
- Actuellement en Master 2 Langue et Informatique et en stage de recherche au LORIA sous la direction de Karën Fort et Bruno Guillaume
- Prochainement en Thèse...

Mémoire de recherche

Vérification d'universaux linguistiques sur des corpus multilingues annotés en syntaxe de dépendances

- 4 universaux de Greenberg testés
- 141 corpus, 74 langues d'Universal Dependencies
- *Investigating Dominant Word Order on Universal Dependencies With Graph Rewriting*, Hee-Soo Choi, Bruno Guillaume, Karën Fort, Guy Perrier (RANLP2021)

.grew

Thèse en Sciences du Langage

Lier des ressources lexicales du français en vue d'une interopérabilité entre niveaux linguistiques

- Encadré par Mathieu Constant, Karën Fort et Bruno Guillaume
- Laboratoires : ATILF et LORIA
- Financement : École doctorale SLTC, Université de Lorraine

Thèse en Sciences du Langage

Lier des ressources lexicales du français en vue d'une interopérabilité entre niveaux linguistiques

- Lier automatiquement des ressources hétérogènes mais complémentaires
- Développer des méthodes de liage cohérentes
- Utiliser une représentation par graphes
- Obtenir un ensemble lié de ressources pour les outils de TAL

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, 9 septembre 2021

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

Éthique et TAL

- Création de données pour un TAL plus éthique :
  - corpus du français pour l'évaluation des biais stéréotypés des modèles de DL (A. Névéol, Y. Dupont et J. Besançon) : https://languagearc.com/projects/19 (Participez !)
  - corpus de dossiers médicaux synthétiques en français (A. Névéol, O. Ferret, N. Hiebel)
- Analyse des sections éthiques de la conférence NAACL 2021 (E. Bender, M. Mitchell et E. van Miltenburg)
- Grille d'analyse déontologique pour le TAL (M. Amblard)

Projets en cours

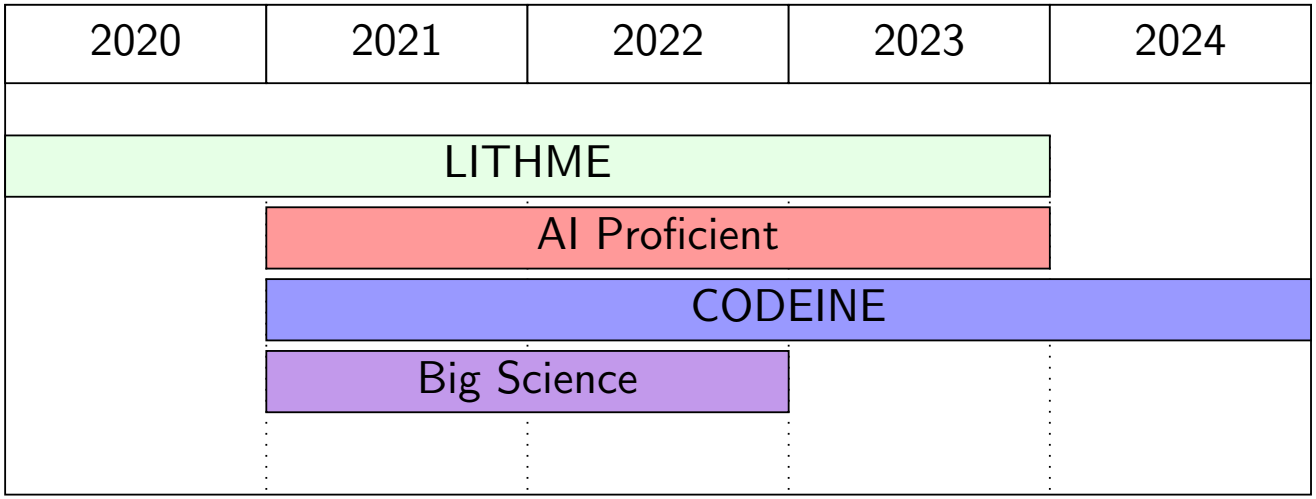


FIGURE – Projets en cours (en vert les actions COST, en rouge les projets européens, en bleu les projets ANR, en violet les projets avec une entreprise)

Encadrements de thèses

- Lucas Lambrey (avec Gilles Siouffi et Antoine Gautier) : *Détection et production de défigements linguistiques dans les réseaux sociaux assistées par les sciences participatives : fertilisation croisée entre traitement informatique et analyse linguistique*. (CERES)
- 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

- Chair du comité d'éthique d'ACL (Association for Computational Linguistics)
- Comité d'éthique de la recherche (CER) de SU
- 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 confirmer)
- Actions européennes COST :
  - Language In The Human-Machine Era (LITHME)