Rencontres Minutes

STIH EA 4509

UFR de Sociologie et d'Informatique pour les Sciences Humaines

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

Qui suis-je ?

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

Mes Intérêts en recherche

Contraintes sur 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 ?

Modéliser la transmission de l'information

- Intention de communication (cf. Sperber & Wilson)
- Stratégies du récepteur (cf. J.Coursil)
- Rôle de l'éco-système (cf. H.Déjean ou F.Rastier)

Les principes exploités

- Le type de discours guide la communication
- Le sens est une interprétation
- La saillance facilite la réception
- Les **observables** doivent être conservés

Sur quoi je l'applique ?

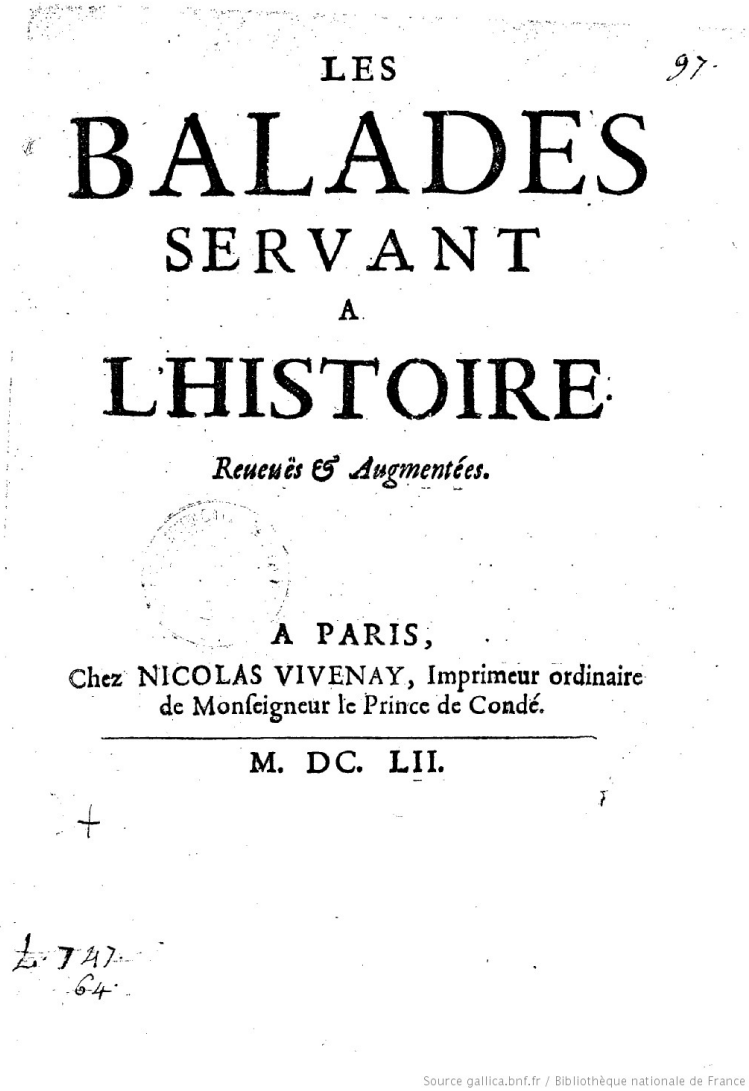
Des travaux récents

- Classification (veille, diagnostic de langue, datation)
- Stylométrie (attribution d'auteur, extraction de structure)
- DEFT (Indexation, Appariement, Polarité ...)

Projets en cours

- MEMES (2019-2021) Memes : Extraction automatique et analyse par Myriadisation d'Expressions Semi-figées (A.Gautier, K.Fort, L.Zhu)
- ANTONOMAZ (2018-2022) ANalyse auTOmatique et NumérisatiOn des MAZarinades (K.Abiven, A.Baledent, N.Hiebel, JB.Tanguy, G.Roe)

Datation automatique de textes anciens (stage Antonomaz)



LES

BALADES

SERVANT

A

L'HISTOIRE

Reueuës & Augmentées.

A PARIS,

Chez NICOLAS VIVENAY, Imprimeur ordinaire de Monſeigneur le Prince de Condé.

M. DC. LII.

Source gallica.bnf.fr / Bibliothèque nationale de France

- Récupération d'un corpus de documents anciens numérisés
- Paramétrage de classifieurs de *machine learning*
- Mise en ligne de l'outil de datation

Thèse (à partir d'octobre)

Évaluation des annotations par des mesures d'accord inter-annotateurs : le cas des flux (texte, audio, vidéo)

sous la direction de Yann Mathet et Antoine Widlöcher, au GREYC (Caen)

- Annotation d'*unitizing*
- Extension et généralisation la mesure γ : relations entre annotations et des attributs/valeurs
- Adaptation aux flux multi-dimensionnels
- Meilleure compréhension des désaccords d'annotation

LETTRES SORBONNE UNIVERSITÉ

Systèmes multi-classifieurs pour la reconnaissance des données médicales

Amina BOUBELNZA

Speed dating

Jeudi le 12/09/2019

Université Abou Bekr Belkaid
Faculté des Sciences
Département d'informatique

Laboratoire de Génie Biomédicale

Équipe CREDOM (Caractérisation et reconnaissance des données médicales)

Laboratoire des sciences de l'information et des systèmes

Équipe OASIS (Ontologies, Agents and Services for Information Systems)
Marseille - France

Thèse soutenu le 22 Février 2017

Bachelors & Masters

Contexte

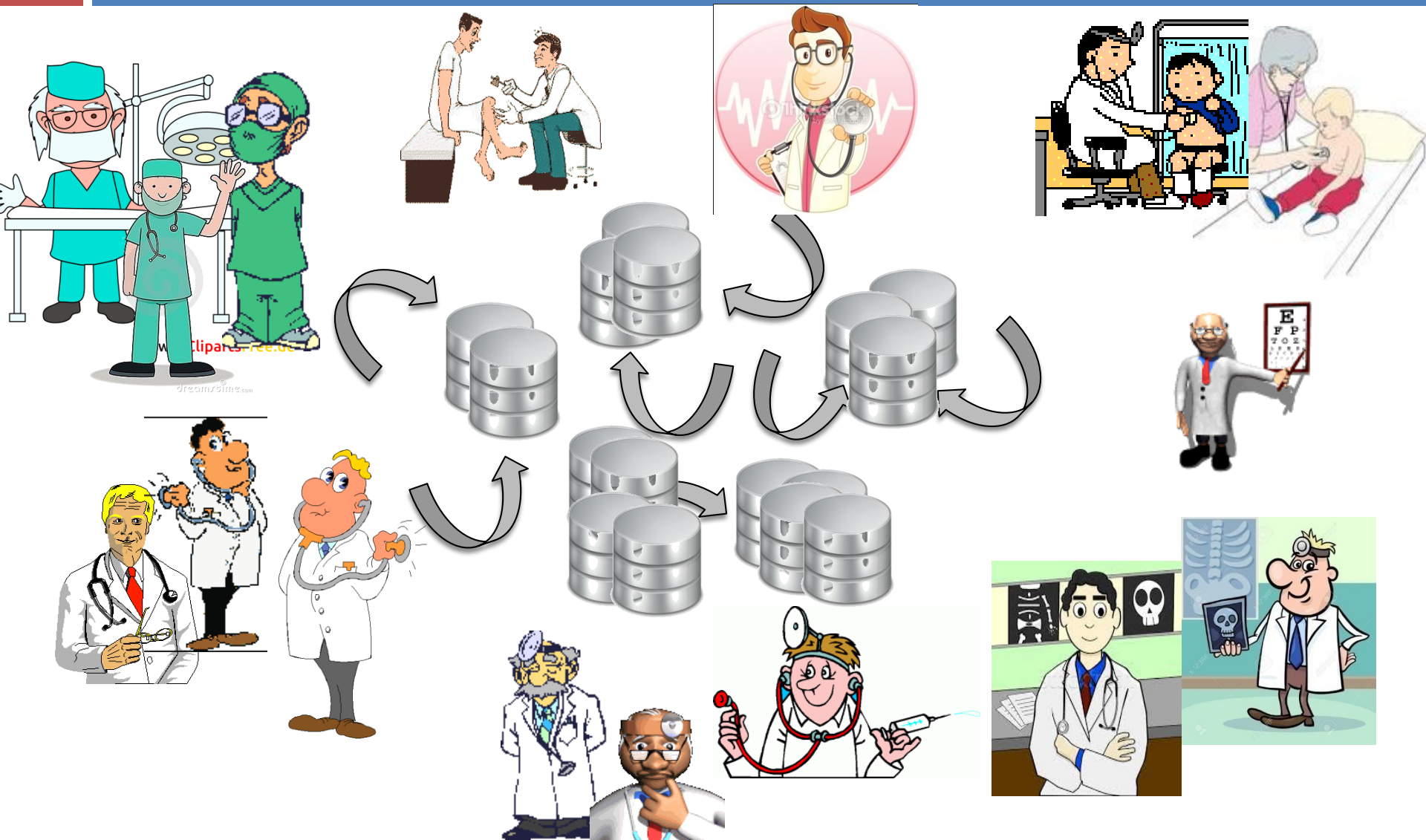
Diagnostic médical

- Démarche pour déterminer l'affection dont souffre le patient, et qui va permettre de proposer un traitement.
- Repose sur la recherche des causes (étiologie) et des effets (symptômes) de l'affection

Système d'aide au diagnostic médical

- Système fournissant aux cliniciens, les informations liées aux patients, intelligemment filtrées et présentées à des moments appropriés, afin d'améliorer les soins de santé.

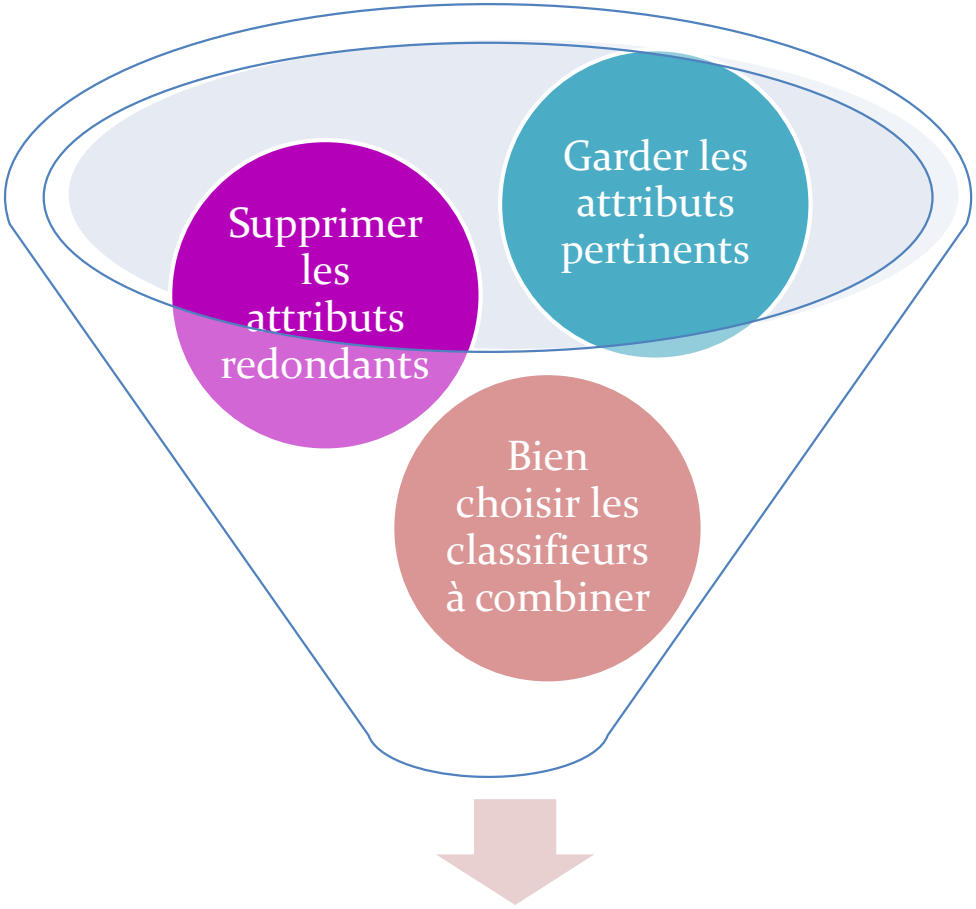
Les faits



Problématiques

- Comment faire face à la croissance exponentielle des données médicales?
- Quelles solutions face à la malédiction de dimensionnalité : beaucoup de variables, peu d'exemples, Redondance… ?
- Existe-t-il un modèle de classification qui traite n'importe quelle distribution de données ?
- Quel est le meilleur classifieur ?

Résumé



Améliorer la classification

Minimiser les erreurs médicales

Création de ressources langagières et éthique pour le TAL

Karën Fort

karen.fort@paris-sorbonne.fr

Séminaire d'équipe, 12 septembre 2019

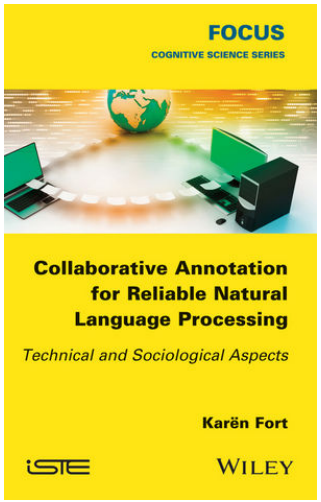
cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

LETTRES SORBONNE UNIVERSITÉ

D'où je parle

Voir `http://karenfort.org/`

- Création de ressources langagières pour le TAL



- Ethique et TAL

Production participative (*crowdsourcing*)

Jeux ayant un but que j'ai participé à créer :

Portail de jeux pour les langues et atelier récurrent :

Games4NLP

Dhaou GHOUL, Attaché temporaire d'enseignement et de recherche en Informatique, Sorbonne Université,Paris, France.

dhaou.ghoul@sorbonne-universite.fr / dhaou.ghoul@gmail.com
http://cereli.fr/membres/dhaou-ghoul/
06 52 75 98 33

12 septembre 2019

LETTRES SORBONNE UNIVERSITÉ

Thèse de doctorat

- Titre : Classifications et grammaires des invariants lexicaux arabes : construction d'un modèle théorique de l'arabe.
- Sotenu le 07 Décembre 2016 à Sorbonne université.

Mes axes de recherche + Actualités

Mes axes de recherche :

- Linguistique de corpus.
- Traitement autmatique de la langue arabe.
- Etiquetage morphosyntaxique de la langue arabe.
- Classification de dialèctes arabes.
- Analyse et classification de documents.

Actualités :

- 02/07/19 : Deft 2019 @TALN.
- 27/05/19 : MICHAEL, système de reconnaissance de dialectes arabes @MADAR shared task 2019.

Quelques publictaions

- Dhaou Ghoul,Gaël Lejeune,Lichao Zhu "Evaluating Lightweight text classification and Information Extraction for Arabic texts". CicLing'19. La Rochelle, France, Avril 2019.
- Dhaou Ghoul, André Jaccarini, Amr Helmy Ibrahim. "Cassification and grammars of simple Arabic lexical invarinats in anticipation of an automatic processing of this language : "the temporal invariants"". CicLing'17, Budepest, Hungary, Avril 2017.
- Dhaou Ghoul, Amr Helmy Ibrahim, Claude Audebert "Rules-based grammatical and semantic disambiguation of the token "hattā" in Arabic". ICTA'2015, Marrakech, Maroc, December 2015
- Dhaou Ghoul,"Développement de ressources pour l'entraînement et l'utilisation de l'étiqueteur morphosyntaxique TreeTagger sur l'arabe".conférence TALN-RECITAL, Sables d'Olonnes, France,juin 2013.

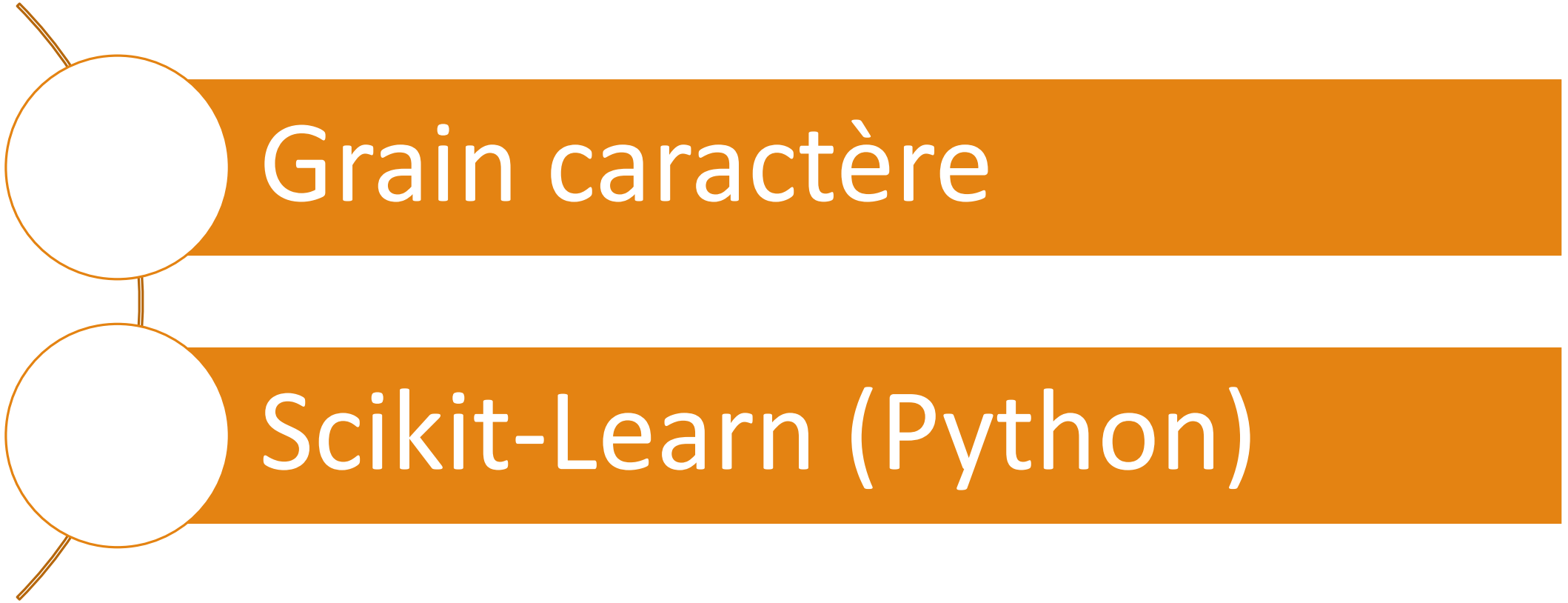
Identification de dialectes arabes

REPRISE DU ''MADAR CHALLENGE''

Le corpus

- Constitué pour le défi
- 26 dialectes
- 46 600 documents

Méthode



Expériences et résultats

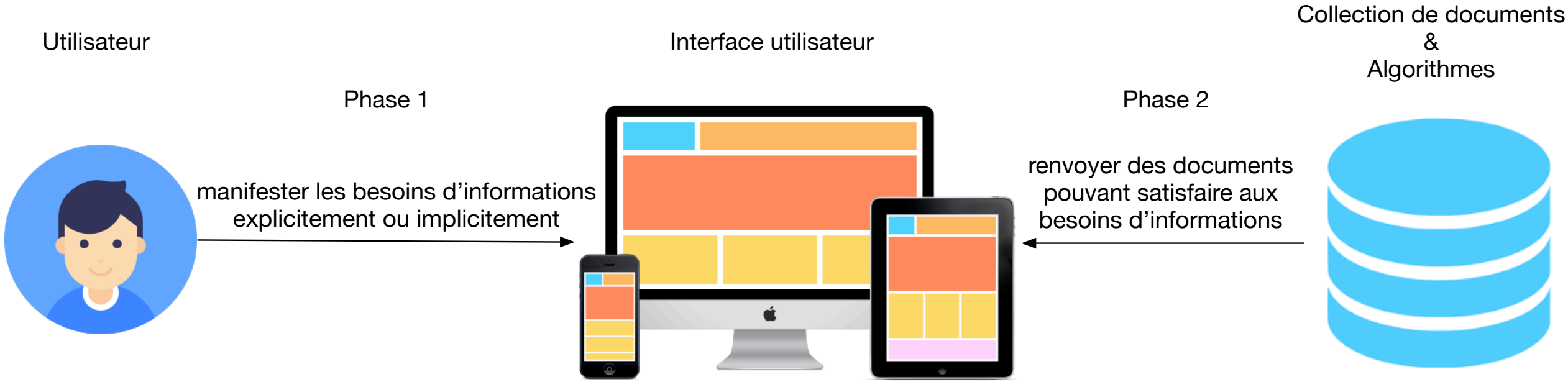
- Variation du nombre de caractères
- Variation de l'effectif
- Application à d'autres corpus

Séminaire linguistique computationnelle

Le 12 septembre 2019

3 diapositives de Vincent LULLY

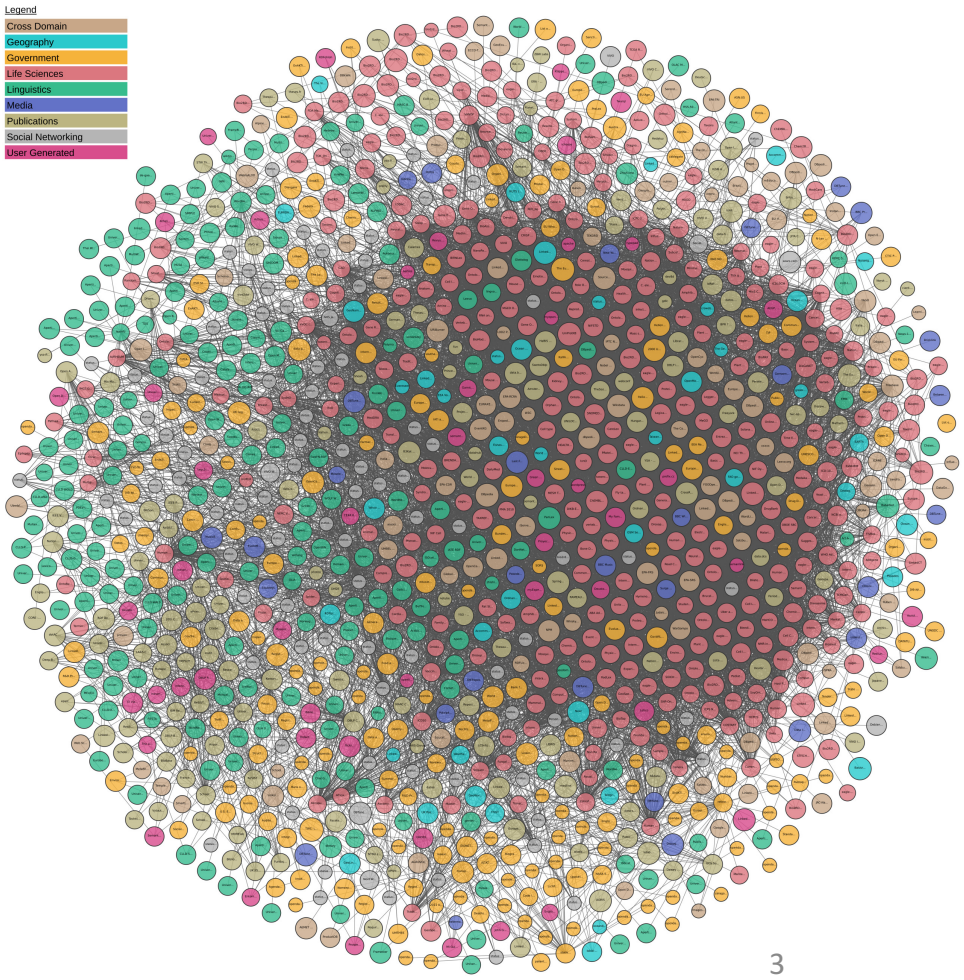
Accès à l'information pertinente



1. Aide insuffisante à la manifestation des besoins d'informations explicites
2. Manque de pertinence des documents sélectionnés
3. Documents sélectionnés pas assez bien mis en valeur

Web sémantique : le remède ?

- Outils et standards définis par le Web sémantique permettant l'interopérabilité des données hétérogènes
- Progrès plus récents sur le nuage LOD : 1,239 jeux de données en mars 2019 couvrant des domaines divers
- Nouvelles capacités d'analyse des textes grâce aux extracteurs d'entités sémantiques tels que DBpedia Spotlight et Babelfy



Domaines d’application

- E-tourisme
- Ville intelligente
- Patrimoine culturel

Construction de ressources langagières par myriadisation pour le traitement automatique des langues peu dotées : le cas des langues de France

Alice Millour, encadrée par Karën Fort et Claude Montacié

September 12, 2019

Sorbonne Université

LETTRES SORBONNE UNIVERSITÉ

Constats et enjeux

les technologies en TAL ne se développent pas à la même vitesse pour toutes les langues, pourquoi ?

- les ressources nécessaire à l'outillage de certaines langues sont insuffisantes

Constats et enjeux

les technologies en TAL ne se développent pas à la même vitesse pour toutes les langues, pourquoi ?

- les ressources nécessaire à l'outillage de certaines langues sont insuffisantes
  - BRUTES et **représentatives** des pratiques linguistiques actuelles qu'on souhaiterait pouvoir "traiter" (indexer, annoter, traduire, corriger, résumer etc.)

Constats et enjeux

les technologies en TAL ne se développent pas à la même vitesse pour toutes les langues, pourquoi ?

- les ressources nécessaire à l'outillage de certaines langues sont insuffisantes
  - BRUTES et **représentatives** des pratiques linguistiques actuelles qu'on souhaiterait pouvoir "traiter" (indexer, annoter, traduire, corriger, résumer etc.)
  - ENRICHIES de métadonnées spécifiques (annotations linguistiques, variante(s) linguistique(s) présente(s) dans le corpus, convention(s) orthographique(s) suivie(s) etc.)

Constats et enjeux

les technologies en TAL ne se développent pas à la même vitesse pour toutes les langues, pourquoi ?

- les ressources nécessaire à l'outillage de certaines langues sont insuffisantes
  - BRUTES et **représentatives** des pratiques linguistiques actuelles qu'on souhaiterait pouvoir "traiter" (indexer, annoter, traduire, corriger, résumer etc.)
  - ENRICHIES de métadonnées spécifiques (annotations linguistiques, variante(s) linguistique(s) présente(s) dans le corpus, convention(s) orthographique(s) suivie(s) etc.)
- développer des outils pour des langues non majoritaires demande une justification / est plus difficile à valoriser

Constats et enjeux

les technologies en TAL ne se développent pas à la même vitesse pour toutes les langues, pourquoi ?

- les ressources nécessaire à l'outillage de certaines langues sont insuffisantes
  - BRUTES et **représentatives** des pratiques linguistiques actuelles qu'on souhaiterait pouvoir "traiter" (indexer, annoter, traduire, corriger, résumer etc.)
  - ENRICHIES de métadonnées spécifiques (annotations linguistiques, variante(s) linguistique(s) présente(s) dans le corpus, convention(s) orthographique(s) suivie(s) etc.)
- développer des outils pour des langues non majoritaires demande une justification / est plus difficile à valoriser
- l'adaptation d'une technologie d'une langue mieux dotée à une langue moins dotée n'est pas immédiate (ni garantie d'être efficace)

Constats et enjeux

les technologies en TAL ne se développent pas à la même vitesse pour toutes les langues, pourquoi ?

- les ressources nécessaire à l'outillage de certaines langues sont insuffisantes
  - BRUTES et **représentatives** des pratiques linguistiques actuelles qu'on souhaiterait pouvoir "traiter" (indexer, annoter, traduire, corriger, résumer etc.)
  - ENRICHIES de métadonnées spécifiques (annotations linguistiques, variante(s) linguistique(s) présente(s) dans le corpus, convention(s) orthographique(s) suivie(s) etc.)
- développer des outils pour des langues non majoritaires demande une justification / est plus difficile à valoriser
- l'adaptation d'une technologie d'une langue mieux dotée à une langue moins dotée n'est pas immédiate (ni garantie d'être efficace)
- pour certaines langues, l'absence de standard est une difficulté ***immédiate*** au traitement de la langue : la prise en compte en première instance de la variation est **nécessaire**

les communautés linguistiques **connectées** sont les premières à pâtir du manque d'outils

les communautés linguistiques **connectées** sont les premières à pâtir du manque d'outils ... en contrepartie, elles sont **accessibles** *via* le Web

Crowdsourcing / ressources pour le TAL / variation linguistique

les communautés linguistiques **connectées** sont les premières à pâtir du manque d'outils ... en contrepartie, elles sont **accessibles** *via* le Web

- investir le locuteur dans le traitement automatique de sa langue ?
- permettre la production de ressources (brutes et enrichies) ?
- intégrer la variation en amont du traitement automatisé ?

LETTRES SORBONNE UNIVERSITÉ

L'accessibilité et l'exploitation des documents textuels numérisés : un enjeu pour les bibliothèques numériques d'Île-de-France

Observatoire de la vie littéraire

Directeur : Glenn Roe
Encadrants : Karine Abiven et Gaël Lejeune

Partenaire : Bibliothèque Mazarine

îledeFrance

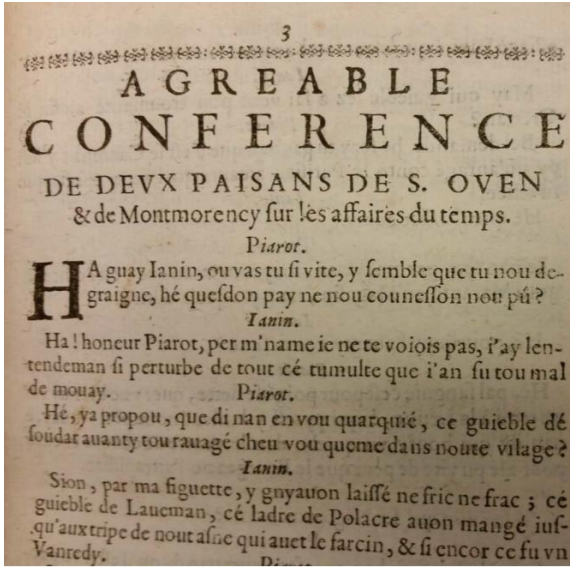
Financement : Région Île-de-France (PhD2)

Le corpus de travail

Les Mazarinades
- Pamphlets, poésies burlesques, lettres fictives…
- Imprimés au sujet du Cardinal Mazarin pendant la Fronde (XVIIè siècle)
- La bibliothèque Mazarine en possède environ 250000 exemplaires

Une première numérisation à venir
- 500 exemplaires sélectionnés
- Définition d'un ensemble de données sur ces exemplaires (ex. : nombre de vues)
- Numérisation par le prestataire (financé par le projet Antonomaz)

Plusieurs itérations prévues



3

AGREABLE
CONFERENCE
DE DEVX PAISANS DE S. OVEN
& de Montmorency sur les affaires du temps.

Piarot.

HA guay Ianin, ou vas tu ſi vite, y ſemble que tu nou degraigne, hé queſdon pay ne nou couneſſon nou pú ?

Ianin.

Ha ! honeur Piarot, per m'name ie ne te voiois pas, i'ay lentendeman ſi perturbe de tout cé tumulte que i'an ſu tou mal de mouay.

Piarot.

Hé, ya propou, que di nan en vou quarquié, ce guieble dé ſoudar auanty tou rauagé cheu vou queme dans noute vilage ?

Ianin.

Sion, par ma figuette, y gnyauon laiſſé ne fric ne frac ; cé guieble de Laueman, cé ladre de Polacre auon mangé iuſqu'aux tripe de nout aſne qui auet le farcin, & ſi encor ce fu vn Vanredy.

Une pièce du corpus : "Agréable conférence de deux paysans de Saint-Ouen et de Montmorency sur les affaires du temps" (dialogue imitant le patois de l'Ile de France, 1649), Réserve de la Bibliothèque Interuniversitaire de la Sorbonne

Les objectifs (1/2) Accessibilité

Améliorer les données textuelles issues d'OCR (reconnaissance optique de caractères)

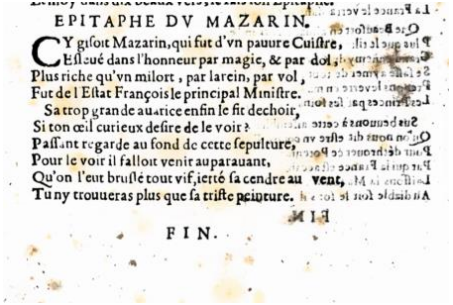
- Langue « non-standard » : lettrines, qualité variable des imprimés, etc.
- Outil envisagé : *Kraken*

  Simon Gabay de l'Université de Neuchâtel déjà entraîné sur un corpus du français du XVIIè siècle
  Corpus : œuvres littéraires (principalement pièces de théâtre), environ 110000 mots
  Précision de 98%

Diagnostic de la qualité des résultats d'OCR

- Comprendre ce sur quoi l'OCR butte
- Détecter quels documents nécessitent une révision manuelle

IMAGE



EPITAPHE DV MAZARIN.

CY gifoit Mazarin, qui fut d'vn pauure Cuiſtre,
Efleué dans l'honneur par magie, & par dol,
Plus riche qu'vn milort, par larein, par vol,
Fut de l'Eſtat François le principal Miniſtre.
Sa trop grande auarice enfin le fit dechoir,
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.

***ABBYY fine reader* (payant)**

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.

Kraken modèle S Gabay (gratuit)

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.

Les objectifs (2/2) Exploitation

L'accessibilité et l'exploitation des documents textuels numérisés : un enjeu pour les bibliothèques numériques d'Île-de-France

Analyse des besoins avérés des chercheurs et utilisateurs et des problèmes

Création de métadonnées

- Pour les documents transcrits manuellement
  - tokenisation/lemmatisation/NER automatiques
- Pour les données bruitées issues d'OCR
  - Lissage manuel : très coûteux → beaucoup d'attente
  - Paradigme : analyse au grain caractère (ex. : avec tokenisation « non-supervisée »)
- Ex. : datation automatique, identification d'auteur [K. Abiven, G. Lejeune, 2019]

Accès

- Un corpus transcrit (pour les chercheurs et le grand public)
- Une bibliothèque numérique enrichie (*Mazarinuum*)

Séminaire de Linguistique Computationnelle

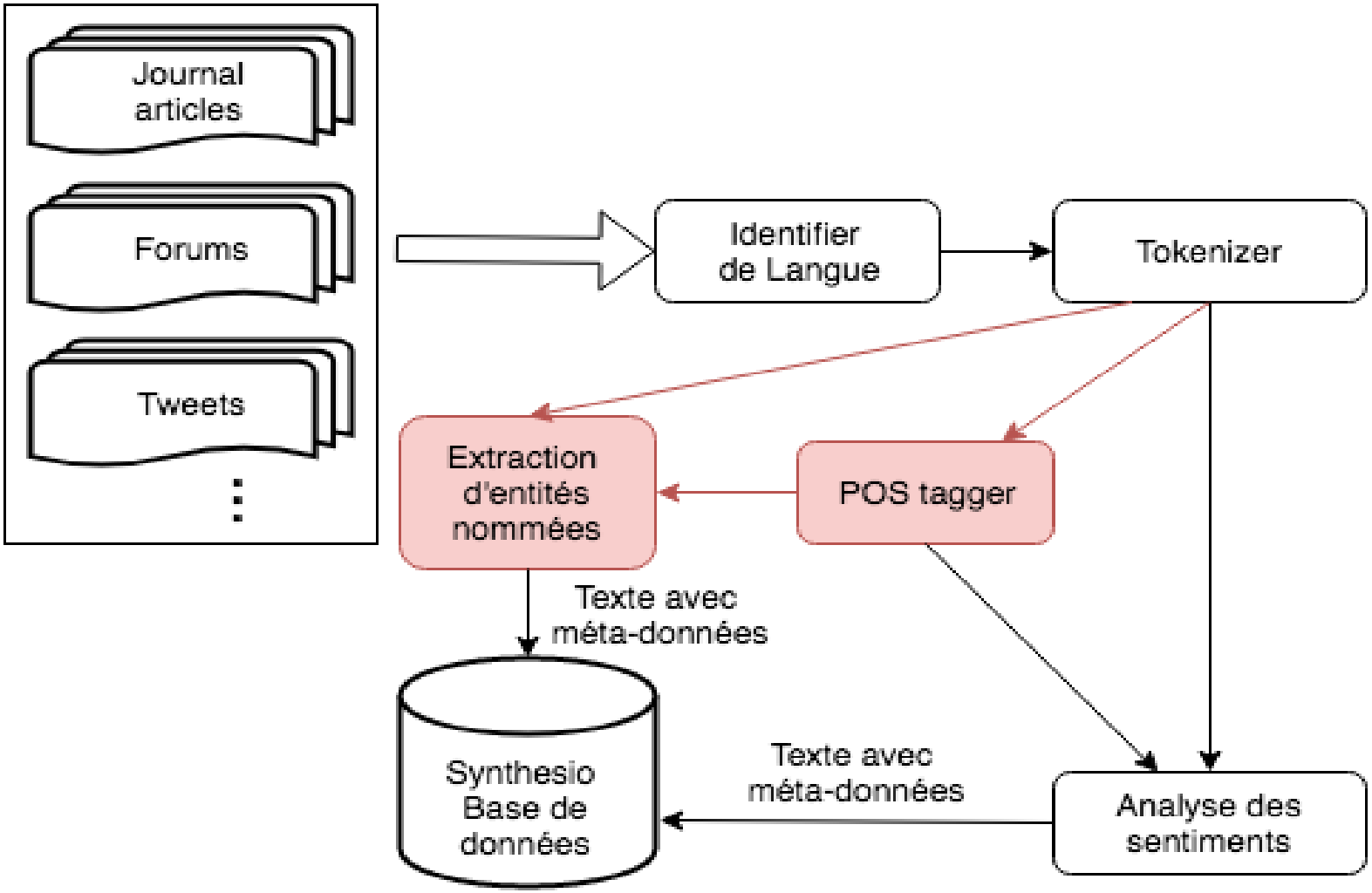
Tian TIAN

12-09-2019

TIAN Tian 田甜

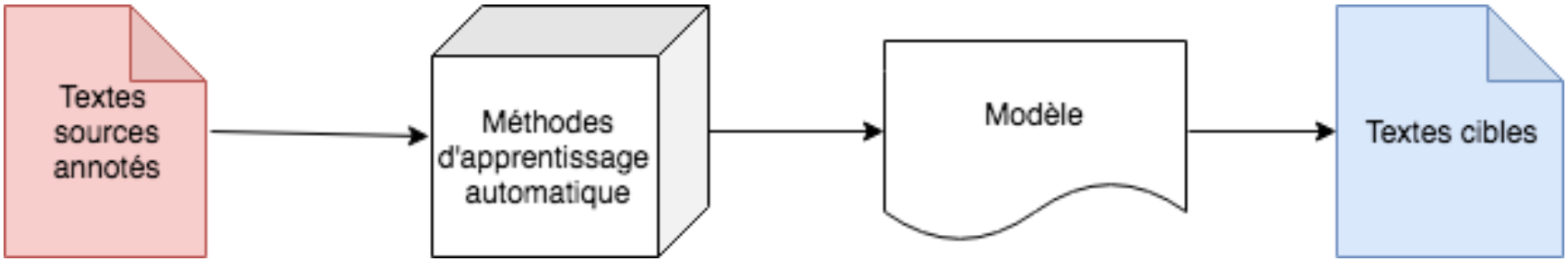
- ATER à Paris IV depuis fin janvier 2019
- Doctorante en Science du Language de Paris III
- Thèse CIFRE chez Synthesio 2014-2017 (Ipsos)
- Directeurs de thèse
  - Isabelle Tellier
  - Marco Dinarelli (LIG)
  - Thierry Poibeau
- Soutenance prévue le 16 octobre 2019

Chaîne de traitement de textes



Méthodes d'apprentissage automatique

Utilisation de méthodes d'apprentissage automatique et de domaine d'adaptation pour extraction d'entités nommées et POS tagging dans les textes multi-domaines et multi-sources



Lucie Vercruyssen **Activités de recherche / d'enseignement**

2019 **Post Doctorante** ENS, Laboratoire Lattice, Experimentations et analyses de données.

2017-18 **Analyste Linguiste** Google, Expertise linguistique, annotation et vérification de données linguistiques.

2016-17 **Post Doct. / Assistante de recherche en Traitement Automatique des Langues** Université Catholique de Louvain, Expertise linguistique, conception d'une base de données et d'un site.

2015-16 **Chercheur externe** Université Catholique de Louvain

2012-15 **Assistante Doctorante** Université de Neuchâtel

2018 (3m.) **Vacataire Enseignement secondaire** Professeure de Lettres modernes et d'Histoire-Géographie, filières technologiques et professionnnelles. *240* heures de cours au total.

2012 - 15 **TP de recherche, 8 ECTS** Enseignement pour Master Orthophonie - 1h30/sem. soit 135 heures de cours.

Prosodie, Syntaxe, et Référence: processus cognitifs et marqueurs linguistiques

- Étude de la variation d'emploi des marqueurs de référence, en combinant indices syntaxiques et prosodiques, produits lors de narrations d'images séquentielles.
- Objectifs :
  - déterminer l'effet des paramètres situationnels
  - mettre en évidence les compétences (socio)cognitives sous-jacentes
  - étudier l'impact du vieillissement.
- Matériel : tâche de narration d'images séquentielles + tests (socio)cognitifs
- Participants : 30 jeunes adultes (19-39 ans) et 30 séniors (59-79 ans).
- Résultats :
  - marqueurs syntaxiques et prosodiques varient en fonction des étapes de discours et du contexte référentiel.
    - Accessibilité moindre : SN défini et Allongement de la DMP
  - accessibilité cognitive du référent -> compétences (socio)cognitives -> marquage référentiel
    - accessibilité cognitive moindre : + flexible, + SN défini
  - Effets du vieillissement sur le marquage référentiel : l'emploi des marqueurs de référence chez les séniors est différent de l'emploi des jeunes adultes.
    - Accessibilité cognitive faible : - SN indéfinis chez les séniors

Post-doc Democrat

- Etude des chaînes de référence
  - Distance entre les mentions : effet sur l'hétérogénéité lemmatique
    - narratifs : Plus la chaîne de référence est instable, plus l'interdistance est élevée
    - non narratif : Plus la chaîne de référence est instable, moins l'interdistance est élevée.

- Expériences de perception en ligne sur la résolution du « on » flou

Je m’appelle …

- ZHU, Lichao
- atterri à l’université de Caen et migré à l’université Paris 13

Je suis …

beaucoup linguiste, de plus en plus informaticien (j'ai l'impression), traducteur assumé, littéraire de formation, un peu geek (…mais pas trop)

Je m'intéresse …

- à l'écriture, au lexique (dont la phraséologie) et à la sémantique
- à l'automatisation des traitements liés au langage et aux connaissances

Mais j'aime aussi
- les jeux d'échecs, la cuisine, Rimbaud, Nietzsche, WANG Yangming, les Inconnus, quelques genres de musique, etc.