

Bénédicte Pierrejean

E-MAIL benedicte.pierrejean@univ-tlse2.fr • Toulouse • France

PROFIL

DOCTORANTE EN TRAITEMENT AUTOMATIQUE DES LANGUES

EDUCATION

- Depuis 2016 **Doctorat en Traitement Automatique des Langues**
Sujet de thèse : Etude des variations qualitatives entre des bases de données distributionnelles construites sur des corpus différents, dirigée par Ludovic Tanguy.
Contrat Doctoral Unique (CDU).
- 2013 **Master Sciences du Langage spécialité Ergonomie Cognitive et Traitement Automatique des Langues**
Université Toulouse-II, Toulouse, France.
Mémoire M2 : Amélioration d'un système de génération automatique de textes pour les commerces locaux, dirigé par Ludovic Tanguy.
Mémoire M1 : L'encyclopédie d'exobiologie vue par les sciences du langage, dirigé par Anne Condamines.
Major de promotion. Mention Bien.
- 2011 **Master Arts, Lettres et Langues mention Langues et Cultures Etrangères spécialité Etudes Anglophones**
Université Toulouse-II, Toulouse, France.
Mémoire M2 : Eléments pour une sociophonologie de l'anglais à St. Louis, dirigé par Jacques Durand et Anne Przewozny.
Mémoire M1 : A PAC Based Sociolinguistic Study of the Vowel System in the English of St. Louis, dirigé par Anne Przewozny.
Mention Bien.
- 2008 **Licence Arts, Lettres et Langues, mention Langues, Littératures, Civilisations Etrangères**
Université Toulouse-II, Toulouse, France.

CERTIFICATS

- 2016 **Certification en Apprentissage Automatique**
Cours dispensé sur 3 mois.
Coursera – Andrew Ng, Stanford University.

EXPERIENCE

Maluuba – Linguiste Informaticienne (Octobre 2013 – Août 2016), Waterloo, ON (Canada)

Mise en place d'un système d'assistant personnel pour le français

- Conception de solutions en français destinées à être intégrées dans différents appareils électroniques.
- Collecte, tri et étiquetage de données.

- Interprétation des demandes de l'utilisateur grâce à des outils d'apprentissage automatique (classification, reconnaissance d'entités nommées).
- Conception d'outils pour la génération automatique de lexiques.
- Définition et implémentation de traits linguistiques spécifiques au français.

Conception d'une librairie Java multilingue pour la correction de la reconnaissance de la parole

- Mise au point d'un algorithme de correction pour la reconnaissance de la parole modulable et adaptable selon le domaine et la langue.
- Portage de l'algorithme de correction pour la reconnaissance de la parole en C++.

Nomao (groupe Ebuzzing) – Stage (Mars – Août 2013), Toulouse (France)

Contribution à la conception d'un prototype de Recherche d'Information

- Amélioration de 207% en moyenne du contenu associé à des lieux locaux.
- Taux d'erreur d'extraction très bas (4%).

Amélioration d'un système de génération de langage naturel

- Textes générés rallongés de 51%.

CNES (Centre National d'Etudes Spatiales) - Stage (Janvier – Juin 2012), Toulouse (France)

Enrichissement d'une encyclopédie d'exobiologie grâce à la linguistique de corpus (YaTeA, Antconc).

PUBLICATIONS ET BREVETS

Pierrejean, Bénédicte (2012). *A Study of the vowel system of the English of St. Louis, Missouri*. PAC 2012, The Phonology of Contemporary English : Variation and Change. Toulouse : 29 February – 2 March 2012.

Brevet déposé avec Maluuba : algorithme pour la correction des noms propres mal identifiés par les systèmes de reconnaissance vocale.

COMPETENCES

Langues :	Anglais (courant)
Langages de programmation :	Python, Java, C#, C++ (bases), Ruby, Perl
Bases de données :	MongoDB
Outils Linguistiques :	Stanford POS Tagger, YaTea (Yet Another Term Extractor), Antconc, TreeTagger, Praat
Apprentissage automatique :	SVM, CRF, Naive Bayes, embeddings (w2v et w2vf), réseaux de neurones