Langues en contact et créolisation : de la grammaticalisation au traitement automatique des langues créoles

Sibylle KRIEGEL
(Professeure – Laboratoire LPL – Aix-Marseille Université)
sibylle.kriegel@univ-amu.fr

Nicolas DAVID
(Doctorant – Laboratoire LIDILEM – Université Grenoble Alpes)
davidnic@univ-grenoble-alpes.fr

Séminaire transversal / DéLiCorTAL – Université Grenoble Alpes

30 septembre 2022

Sommaire

1. Catégoriser les langues créoles
2. Langues créoles et TAL
3. Construction et exploitation d’un corpus arboré
4. Pistes d’amélioration et perspectives de progression

1. Catégoriser les langues créoles

- Langues créoles à base lexicales indo-européennes
- Langues « orales » à fort dynamisme « scriptural »
- Langues en plein essor
- Langues « encadrées »
- Langues vivantes

2. Langues créoles et TAL

- Les langues créoles → un défi pour le TAL (Traitement Automatique des Langues)
- Langues peu documentées (au niveau de la description linguistique)
- Langues régies par un degré de dotation en ressources électroniques et applications informatiques

2. Langues créoles et TAL

Degré de dotation d'une langue

- Évaluation du degré d'informatisation d'une langue (Berment, 2004)
  - Attribution d'un niveau de criticité (C_k) et d'une note (N_k) à cinq catégories de services ou ressources
  - Obtention de l'indice-σ

2. Langues créoles et TAL

Degré de dotation d'une langue

➢Les 3 niveaux d'indice-σ (Berment, 2004)

- Langues-π : moyenne qui se situe entre 0 et 9,99 → langue peu dotée
- Langues-μ : moyenne qui se situe entre 10 et 13,99 → langue moyennement dotée
- Langues-τ : moyenne qui se situe entre 14 et 20 → langue bien dotée

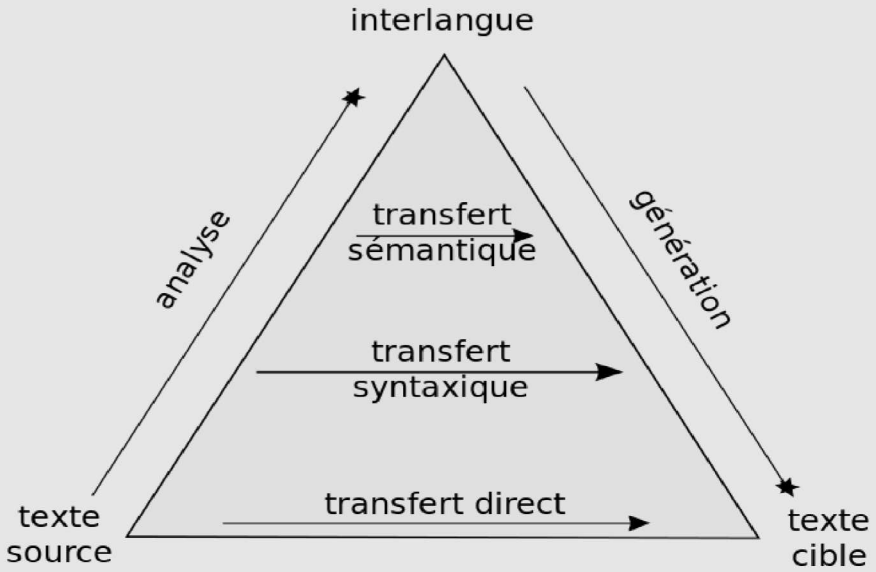
2. Langues créoles et TAL

Degré de dotation d'une langue

« En Traitement Automatique des Langues (TAL), une langue peu dotée est généralement une langue qui souffre d'un manque de ressources lexicales informatisées, de textes électroniques et de corpus annotés. »

(David, 2019, p.110)

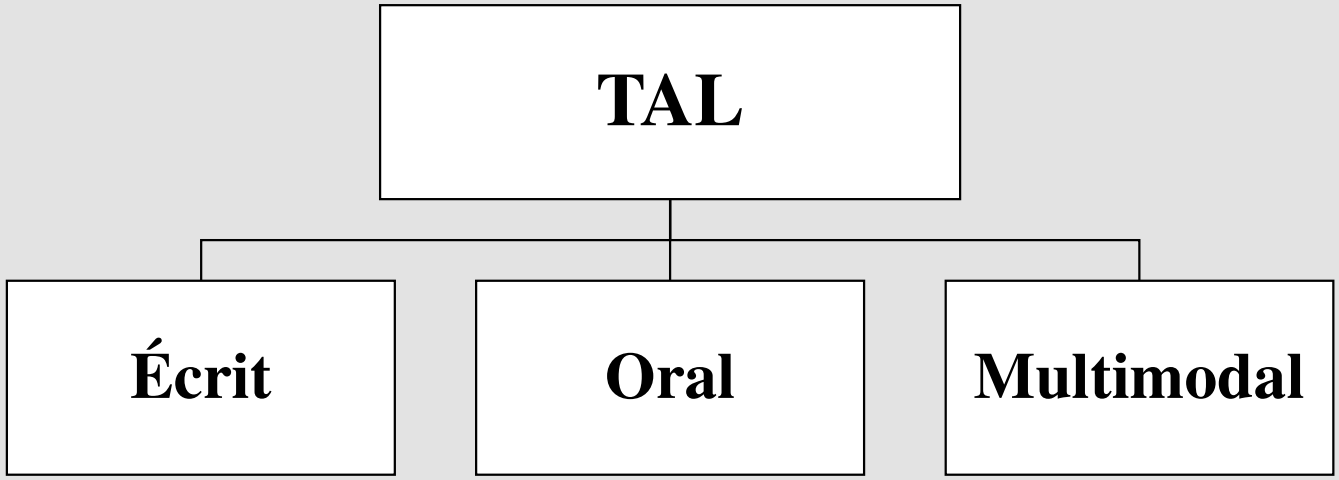
2. Langues créoles et TAL



Triangle de Vauquois
(Source du graphique : *Wikipedia*)

2. Langues créoles et TAL

Domaines associés au TAL



2. Langues créoles et TAL

Tentative de définition du TAL

« Tout processus qui tend vers la modélisation informatique d'une langue naturelle, de ses ressources linguistiques, afin de permettre à une machine d'accomplir des tâches linguistiques et d'assurer le développement d'applications et de modules, dans l'optique de promouvoir et de faciliter les interactions humain-machine. »

(David, 2022)

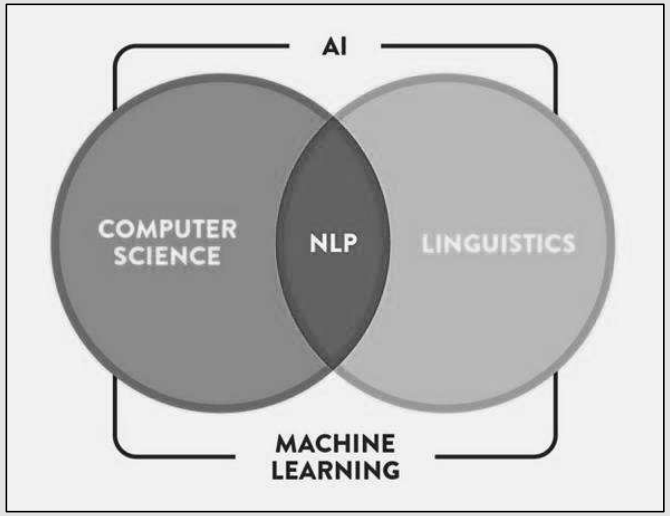
2. Langues créoles et TAL

Niveaux de traitement en TAL

- Phonétique, phonologie et prosodie
- Segmentation / Tokénisation
- Morphologie et lexique
- Syntaxe
- Sémantique
- Pragmatique

2. Langues créoles et TAL

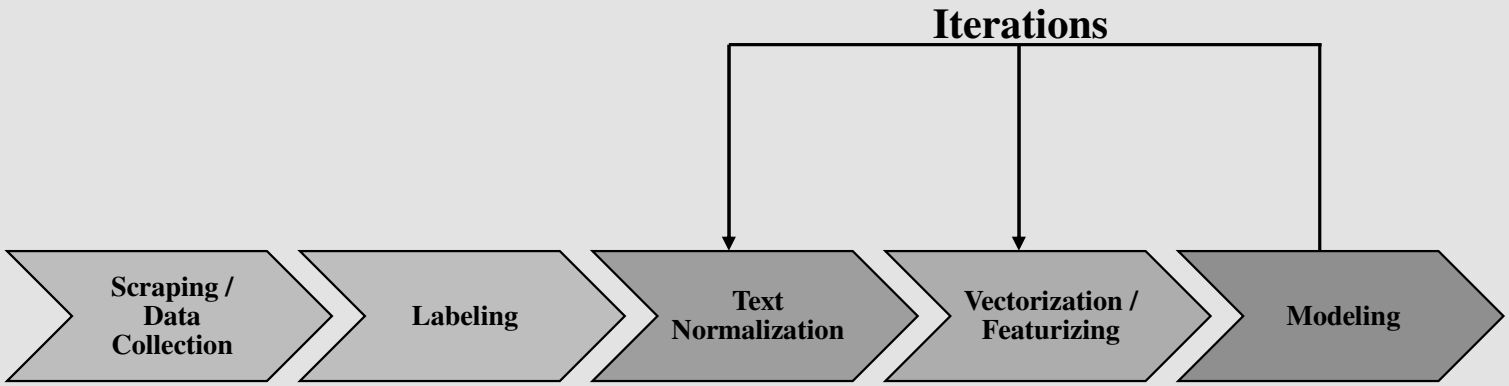
Évolution des niveaux de traitement en TAL
(apports de l'intelligence artificielle et de l'apprentissage automatique)



(Source du graphique : *CleverTap*)

2. Langues créoles et TAL

Text Processing Workflow
(Bansal, 2021)



2. Langues créoles et TAL

Bibliothèques logicielles pour le TAL

- spaCy
- NLTK
- Stanford NLP : CoreNLP et Stanza
- Gensim
- AllenNLP

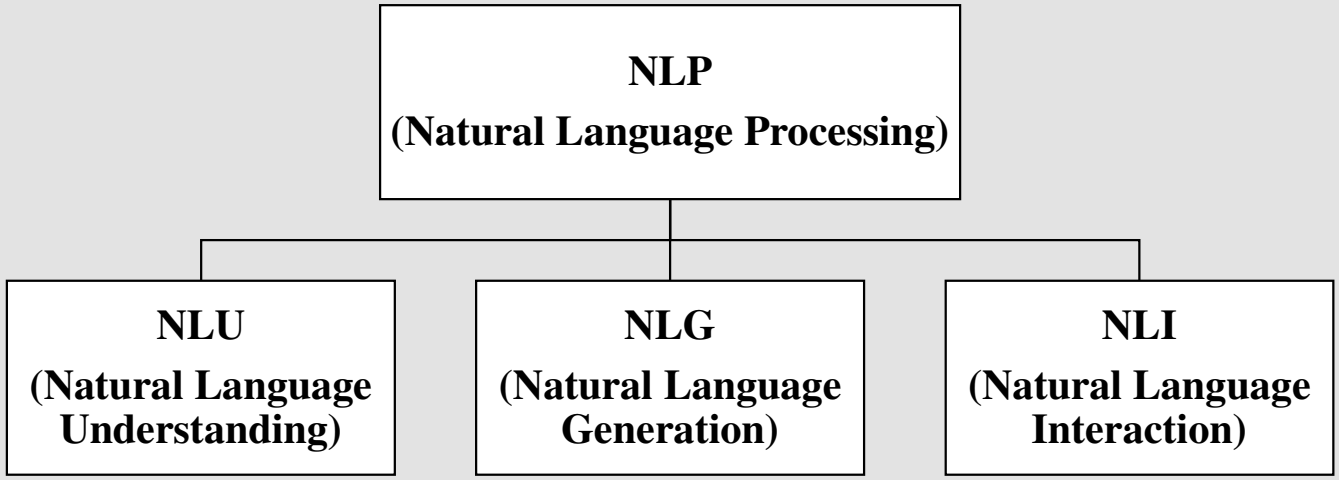
2. Langues créoles et TAL

Modèles et architectures pour le TAL

- RNN / CNN
- LSTMs / BiLSTMs
- CRFs
- TF-IDF
- BERT / Transformers

2. Langues créoles et TAL

Le NLP et ses sous-domaines



3. Construction et exploitation d'un corpus arboré

Cadre du travail doctoral

Construction et exploitation d'un corpus arboré en dépendances pour le créole mauricien

3. Construction et exploitation d'un corpus arboré

Thématiques de recherche pluridisciplinaires

- Langues créoles à base française
- Langues peu dotées et peu documentées
- Analyse syntaxique en dépendances
- TAL : linguistique et informatique appliquées

3. Construction et exploitation d'un corpus arboré

Principaux objectifs de la thèse

- Construction d'un corpus écrit et annoté syntaxiquement
  - Privilégier l'annotation syntaxique en dépendances
- Modélisation des ressources linguistiques du créole mauricien
  - Optimiser les traitements de base en TAL
- Développement d'une chaîne de traitement
  - Aboutir à un modèle pouvant être implémenté en tant que module de TAL

3. Construction et exploitation d'un corpus arboré

Problématique « provisoire »

Dans quelle mesure est-il possible, grâce aux apports du TAL et de l'apprentissage profond, de modéliser les ressources linguistiques du créole mauricien afin d'aboutir à la construction et l'exploitation d'un corpus écrit annoté syntaxiquement en dépendances ?

3. Construction et exploitation d'un corpus arboré

Modalités de recueil du corpus

- Usage d'une application OCR
- *Océrisation* de textes essentiellement littéraires
- Volume brut : ± 100 000 tokens
- Volume traité : 25 638 tokens / 1 457 phrases

3. Construction et exploitation d'un corpus arboré

Processus de normalisation

- **Écart à la norme et forme existante**
  - Ex. *nu** (*mfe.* nou ; *fr.* nous)
- **Écart à la norme et forme inexistante**
  - Ex. *inifikasyon** (*mfe.* inifikasion ; *fr.* unification)
- **Forme normée, mais inexistante**
  - Ex. *rezeton* (*fr.* rejeton)
- **Emprunt**
  - Ex. *wage slavery* (*fr.* esclavage salarié)

3. Construction et exploitation d'un corpus arboré

Entreprise des traitements de base en TAL

- Tokénisation / Segmentation
  - Expression régulière
- Analyse lexicale (non désambiguïsée)
  - Script *Python*
- Étiquetage morphosyntaxique
  - Modèle statistique à base de CRFs (module *CRF++*)

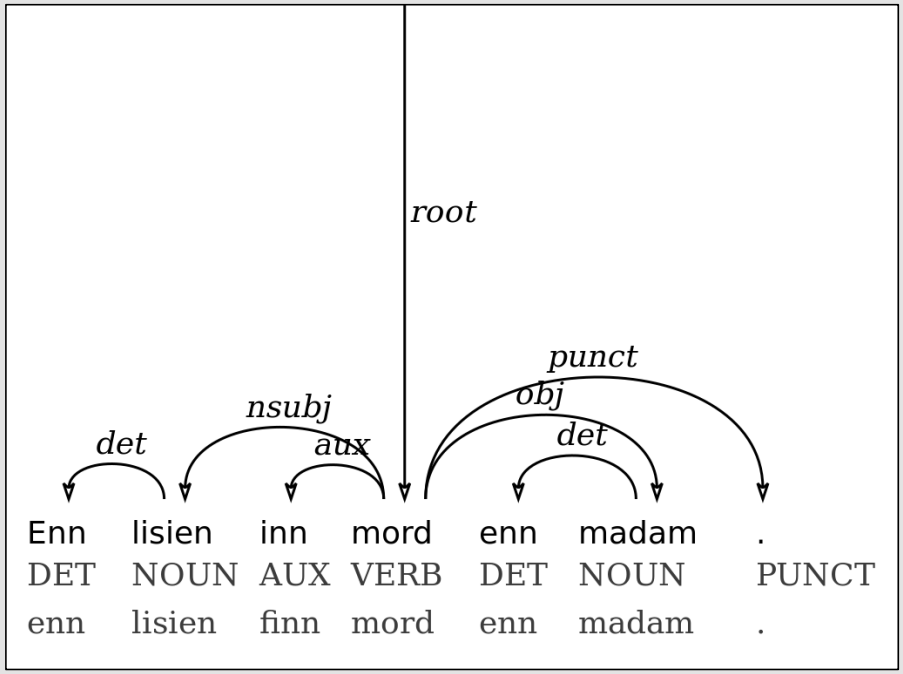
3. Construction et exploitation d'un corpus arboré

Annotation syntaxique en dépendances

- Annotation basée sur les *Universal Dependencies* (UD)
- Adoption du format CoNLL-U
- Apprentissages et tests au moyen de *pipeline* de référence (*UDPipe* et *spaCy*)
- Possibilité de générer des arbres de dépendances (représentation graphique)

3. Construction et exploitation d'un corpus arboré

Représentation arborescente en dépendances
(graphique généré au moyen d'*Arborator*)



3. Construction et exploitation d'un corpus arboré

Esquisse d'une évaluation

- *TAG Accuracy* (moyenne) : 90, 32 %
- *Unlabeled Attachment Score* (UAS) : 72,51 %
- *Labeled Attachment Score* (LAS) : 67,38 %

3. Construction et exploitation d'un corpus arboré

Contraintes liées aux cas de « dépendances éloignées » (Rothman, 2021)

"Alice, whose husband went jogging every Sunday, liked to go to a dancing class in the meantime."

3. Construction et exploitation d'un corpus arboré

Apports de l'apprentissage profond

Simple BERT Models for Relation Extraction and Semantic Role Labeling
(Shi et Lin, 2019)

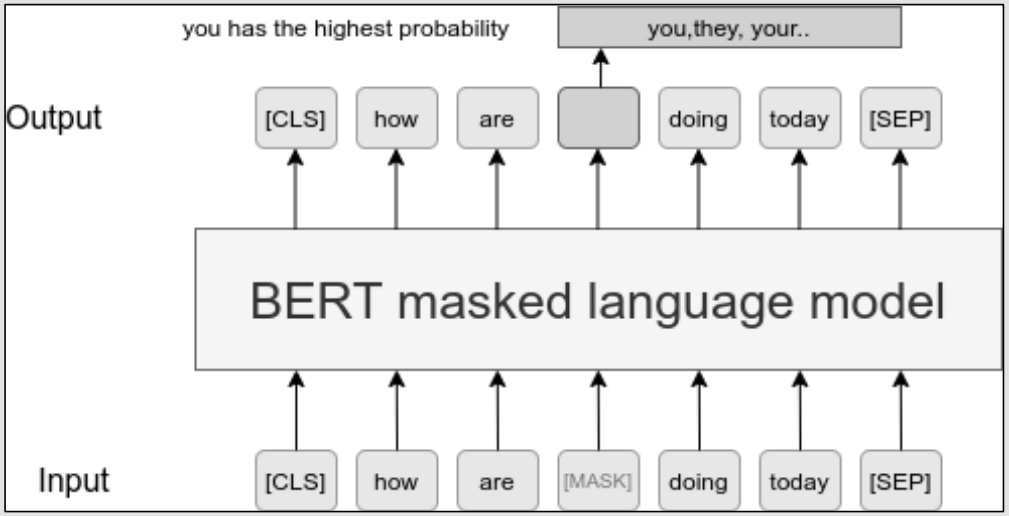
3. Construction et exploitation d'un corpus arboré

Apports de l'apprentissage profond

"Shi and Lin (2019) start their paper by asking if preliminary syntactic and lexical training can be skipped. Can a BERT-based model perform SRL without going through those classical training phases? The answer is yes!"
(Rothman, 2021, p.295)

3. Construction et exploitation d'un corpus arboré

BERT Language Model
(***B**idirectional **E**ncoder **R**epresentations from **T**ransformers*)



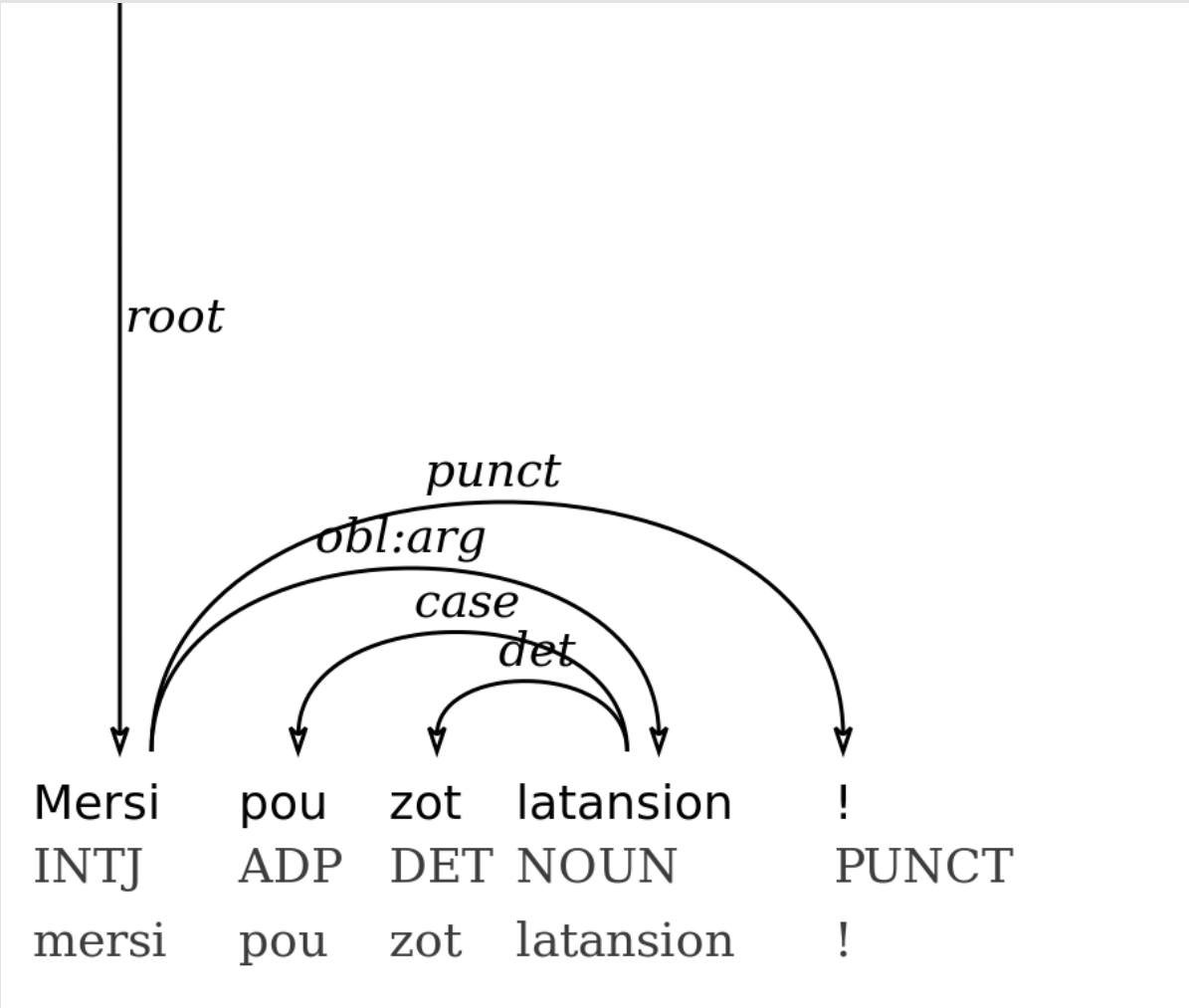
(Source du graphique : *SBERT*)

3. Construction et exploitation d'un corpus arboré

Résolution des cas de « dépendances éloignées » (Rothman, 2021)

4. Pistes d'amélioration et perspectives de progression

- Tendre vers un *gold standard corpus*
- Harmoniser la chaîne de traitement
- Exploiter les modèles à base de *transformeurs*, en particulier le modèle BERT
- Améliorer le UAS et le LAS
- Parvenir à l'implémentation d'un modèle en tant que module de TAL


root
punct
obl:arg
case
det
Mersi pou zot latansion !
INTJ ADP DET NOUN PUNCT
mersi pou zot latansion !

Références bibliographiques

- Bansal, A. (2021). *Advanced Natural Language Processing with TensorFlow 2.* Birmingham : Packt Publishing Ltd.

- Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues « peu dotées ».* Grenoble : Université Joseph-Fourier – Grenoble I.

- David, N. (2019). *Modélisation des ressources linguistiques du créole mauricien en vue de son traitement automatique.* Grenoble : Université Grenoble Alpes.

- Duong, L. T. (2017). *Natural Language Processing for Resource-Poor Languages.* Melbourne : University of Melbourne.

Références bibliographiques

- Fuchs, C., Danlos, L., Lacheret-Dujour, A., Luzzati, D. et Victorri, B. (1993). *Linguistique et Traitements Automatiques des Langues*. Paris : Hachette.

- Lê, V. B. (2006). *Reconnaissance automatique de la parole pour des langues peu dotées*. Grenoble : Université Joseph-Fourier – Grenoble I.

- Pellegrini, T. (2008). *Transcription automatique de langues peu dotées*. Paris : Université Paris-Sud – Paris XI.

- Rothman, D. (2021). *Transformers for Natural Language Processing*. Birmingham : Packt Publishing Ltd.