الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للاعلام الالي 8 ماي 1945
École Supérieure en Informatique
8 Mai 1945 Sidi Bel Abbès

MÉMOIRE

En vue de l'obtention du diplôme de **Master**
Filière: **Informatique**
Spécialité: **Ingénierie des Systèmes Informatiques (ISI)**

Thème

A Smart NLP System for Tamasheq – Arabic Bilingual Corpus Building

---

Présenté par:
Benlaria Ayyoub Yassine
Dafi Adel

Soutenu le : **17/09/2023** devant le jury composé de :

Dr. Chaib souleymane	Président
Dr. Bousmaha Rabab	Examinatrice
Dr. Mediani Mohammed	Encadrant
Dr. Khaldi Belkacem	Encadrant

Année Universitaire : 2022-2023

Thanks

First and foremost, heartfelt gratitude and praises go to the Almighty ***Allah****.*

We would like to extend our heartfelt appreciation to our esteemed supervisors, ***Dr Belkacem Khaldi****,* ***Dr Mohammed Mediani****, and* ***Pr Jan Nieheues****. Their unwavering guidance, unwavering support, and invaluable insights have been instrumental in the successful completion of our thesis.*

We are also deeply thankful for our ***parents*** *and* **family** *whose unending love and support have been our constant motivation throughout this academic journey. Their encouragement has been a driving force behind our achievements. We would like to extend our thanks to our friends, especially* ***Guerrout Mohammed Salem****,* ***Hiri El-Wannes*** *and* ***Ghaitaoui Moulay El-Hadj*** *for the great help they have assisted us with.*

Lastly, our sincere thanks go to all the ***participants*** *in our study. Their willingness to share their knowledge was pivotal to the success of our research. Without their contributions, this work would not have been possible.*

ABSTRACT

This thesis explores the challenges and innovative approaches in the domains of speech recognition and machine translation, with a particular focus on addressing the complexities associated with low-resource languages. It delves into various methodologies, including Multilayer Perceptrons (MLPs), Hidden Markov Models (HMMs), Transfer Learning, Prior Models, and Multilingual Learning, providing a comprehensive overview of the state-of-the-art techniques. Through in-depth analysis and comparative studies, this thesis reveals the strengths and limitations of each approach and lays the foundation for further research and development in the field. The study culminates with an examination of low-resource scenarios in the Tamasheq language, shedding light on the challenges faced and potential solutions for improving automatic speech recognition (ASR) and speech translation (ST) in such environments.

Keywords: Natural language processing, Automatic Speech recognition, Machine translation, low resource language, deep learning, Tamasheq, Tuareg, Wav2Vec2, M2M.

تستكشف هذه الأطروحة التحديات والأساليب المبتكرة في مجالات التعرف على الكلام والترجمة الآلية، مع التركيز بشكل خاص على معالجة التعقيدات المرتبطة باللغات محدودة البيانات. تمت مراجعة منهجيات مختلفة، بما في ذلك Multilayer Perceptrons (MLPs)، و Hidden Markov Models (HMMs)، و Transfer Learning، و Preior Models ، و Multilingual Learning، مما يوفر نظرة شاملة على أحدث التقنيات. ومن خلال التحليل المتعمق والدراسات المقارنة، تكشف هذه الأطروحة عن نقاط القوة والقيود في كل نهج وتضع الأساس لمزيد من البحث والتطوير في هذا المجال. وتتوج الدراسة بدراسة سيناريوهات محدودية البيانات في لغة تماشق، وتسليط الضوء على التحديات التي تواجهها والحلول المحتملة لتحسين التعرف التلقائي على الكلام (ASR) وترجمة الكلام (ST) في مثل هذه البيئات.

Cette thèse explore les défis et les approches innovantes dans les domaines de la reconnaissance vocale et de la traduction automatique, avec un accent particulier sur la résolution des complexités associées aux langues à faibles ressources. Il explore diverses méthodologies, notamment les perceptrons multicouches (MLP), les modèles de Markov cachés (HMM), l'apprentissage par transfert, les modèles antérieurs et l'apprentissage multilingue, offrant un aperçu complet des techniques de pointe. Grâce à une analyse approfondie et des études comparatives, cette thèse révèle les forces et les limites de chaque approche et jette les bases de recherches et de développements ultérieurs dans le domaine. L'étude se termine par un examen de scénarios à faibles ressources dans la langue tamasheq, mettant en lumière les défis rencontrés et les solutions potentielles pour améliorer la reconnaissance automatique de la parole (ASR) et la traduction vocale (ST) dans de tels environnements.

ACRONYMES

- **NLP**: Natural Language Processing
- **ST**: Speech Translation
- **SSL**: Self-Supervised Learning
- **ASR**: Automatic Speech Recognition
- **ML**: Machine Learning
- **DL**: Deep Learning
- **MT**: Machine Translation
- **BoW**: Bag Of Words
- **TF-IDF**: Term Frequency-Inverse Document Frequency
- **LDA**: Latent Dirichlet Allocation
- **AI**: Artificial Intelligence
- **ANN**: Artificial Neural Network
- **FNN**: Feed-Forward Neural Network
- **CNN**: Convolutional Neural Network

- **DNN**: Deep Neural Network
- **RNN**: Recurrent Neural Network
- **ReLU**: Rectified Linear Unit
- **MLP**: Multilayer Perceptron
- **LRL**: Low-Resourse Language
- **HRL**: High-Resourse Language
- **MLP**: Multi-Layer Perceptron
- **HMM**: Hidden Markov Models
- **GMM**: Gaussian Mixture Model
- **PEM**: Phoneme Error Rate
- **WER**: Word Error Rate
- **CER**: C Error Rate
- **SHL**: Shared Hidden Layer
- **TDNN**: Time Delay Deep Neural Network
- **LSTM**: long-short term memory
- **LM**: Language Model
- **NMT**: Neural Machine Translation
- **TL**: Transfer Learning
- **BLEU**: BiLingual Evaluation Understudy
- **ZST**: Zero Shot Translation