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

THESIS

En vue de l'obtention du diplôme d'**ingénieur d'état**
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 study explores the development of Natural Language Processing (NLP) tools and models for the Tamasheq language, a low-resource and underrepresented language spoken in North Africa. The research focuses on Automatic Speech Recognition (ASR) and translation tasks, utilizing fine-tuning and data augmentation techniques. Additionally, a user-friendly TamaLinguist Platform is introduced, providing easier access and utilization of the developed models. The findings indicate that fine-tuning ASR models with mixed datasets and integrating features from other languages and models can significantly enhance accuracy. Moreover, the research demonstrates progress in Tamasheq text translation, although challenges remain due to limited data and language structural complexities. This work highlights the potential of NLP advancements in revitalizing and preserving endangered languages like Tamasheq.

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

تستكشف هذه الدراسة تطور أدوات ونماذج معالجة اللغات الطبيعية (NLP) للغة تماشق، وهي لغة محدودة البيانات يتحدث بها في بعض مناطق شمال إفريقيا. يركز البحث على التعرف على الكلام (ASR) و الترجمة، وذلك باستخدام Fine-Tuning و Data-Augmentation. بالإضافة إلى ذلك، تم تقديم منصة TamaLinguist سهلة الاستخدام، مما يوفر سهولة الوصول إلى النماذج المطورة واستخدامها. تشير النتائج إلى أن الضبط الدقيق لنماذج ASR مع مجموعات البيانات المختلطة ودمج بعض المميزات من لغات ونماذج أخرى يمكن أن يعزز الدقة بشكل كبير. علاوة على ذلك، يوضح البحث التقدم المحرز في ترجمة نصوص تماشق، على الرغم من استمرار التحديات بسبب محدودية البيانات والتعقيدات اللغوية. يسلط هذا العمل الضوء على إمكانات التقدم في البرمجة اللغوية العصبية في تنشيط اللغات المهددة بالانقراض والحفاظ عليها مثل تماشق.

Cette étude explore le développement d'outils et de modèles de traitement du langage naturel (NLP) pour la langue Tamasheq, une langue à faibles ressources et sous-représentée parlée en Afrique du Nord. La recherche se concentre sur la reconnaissance automatique de la parole (ASR) et les tâches de traduction, en utilisant des techniques de Fine-Tuning et d'augmentation des données. De plus, une plate-forme TamaLinguist conviviale est introduite, facilitant l'accès et l'utilisation des modèles développés. Les résultats indiquent quun Fine-Tuning des modèles ASR avec des ensembles de données mixtes et lintégration de fonctionnalités dautres langages et modèles peuvent améliorer considérablement la précision. De plus, la recherche démontre des progrès dans la traduction de textes en Tamasheq, même si des défis subsistent en raison du nombre limité de données et de la complexité structurelle de la langue. Ce travail met en évidence le potentiel des progrès de la NLP dans la revitalisation et la préservation de langues en voie de disparition comme le Tamasheq.

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
- **MMS**: Massively Multilingual Speech