Moroccan dialect NLP resources for Data Engineering and Intelligent Systems

Safâa EL OUAHABI[1], *Sara* EL OUAHABI[2], *Mohamed* ATOUNTI[1], *Issam* SEDDIK[1], *EL Wardani* DADI[2]

[1] Laboratory of Applied Mathematics and Information Systems, Multidisciplinary Faculty of Nador, Mohamed First University Oujda, Morocco
[2] SOVIA team, LSA, National School of Applied Sciences Hoceima, Abdelmalek Essaadi University of Tangier, 90000, Tangier, Morocco

Abstract. Dialectal resources can provide valuable information for scientific research in many domains. They can play an important role in scientific research, especially in the fields of linguistics, sociology, anthropology, psychology, digital transformation, and artificial intelligence. NLP can also play an important role in decision-making by enabling the analysis of large volumes of textual data to extract relevant information. Data relevance is a key factor in decision-making, and companies wishing to join the trend must have all NLP resources at their disposal. In this paper, we will present the different resources and systems developed for Data Engineering and Intelligent Systems for Moroccan dialects.

Index Terms— Moroccan Dialects, Natural Language Processing (NLP), Artificially Intelligence, Data Engineering.

1. Introduction

Natural Language Processing (NLP) resources have a significant impact on decision-making processes by enabling the analysis of large volumes of textual data to extract relevant information. These resources find applications in various domains such as information retrieval, virtual assistance, machine translation, sentiment analysis, task automation, cybersecurity, automatic summarization, medical assistance, data analysis, automatic dialect translation, and dialect speech recognition. NLP algorithms facilitate search engines in understanding user queries, chatbots in interacting with users, and sentiment analysis tools in assessing customer opinions. Additionally, NLP enables the automation of repetitive tasks, detection of cybersecurity threats, extraction of key information from texts, interpretation of medical records, and analysis of textual data for business insights. Moreover, NLP aids in dialect translation and recognition, enhancing communication and accuracy in multilingual contexts.

The development of NLP resources for dialects is contingent on several conditions. Firstly, data availability plays a crucial role. Sufficient amounts of dialect-specific text and speech data need to be collected and annotated for training language models and building dialect-specific resources. Additionally, linguistic expertise is essential to accurately annotate and label the dialect data, ensuring high-quality training materials.

Another condition is the involvement of language communities. Collaborative efforts between researchers, linguists, and native speakers of the dialect are vital for understanding

the linguistic nuances, identifying dialect-specific features, and validating the developed resources. Engaging language communities ensures that the resources accurately reflect the dialect's characteristics and are relevant to its speakers.
Furthermore, technological infrastructure and computational resources are necessary for processing and analyzing large amounts of dialect data. High-performance computing systems and efficient algorithms are crucial for training complex language models and running computationally intensive NLP tasks.
The development of NLP resources for dialects involves exhaustive data collection, meticulous annotation, lexicon construction, language model development, the adaptation of existing models, and performance evaluation. Data collection encompasses gathering representative written texts, audio recordings, and annotated corpora specific to the dialect. Corpus annotation involves labeling linguistic features like parts of speech and named entities. Dialect-specific lexicons are constructed to capture word information and relationships. Language models are developed using the collected data. Existing models can be adapted through transfer learning or multitask learning. Performance evaluation assesses the resources' accuracy, reliability, and suitability for NLP application.
The following sections describe the NLP applications and resources (Amazigh and Darija) built in the domains of speech recognition, handwriting recognition, and sentiment analysis for Moroccan dialects. Finally, the paper summarizes ongoing tasks and perspectives.

2. NLP resources for Amazigh speech recognition

2.1 Voice database for Amazigh speech recognition

A database of Amazigh speech recognition [1] was developed as part of a research thesis, which consists of two datasets. One dataset contains 10 digits and 33 standard letters with 19,500 recordings, and the other dataset contains 520 isolated Amazigh words spoken by 50 speakers with a repetition of 5 times for each word, making up a total of 130,000 recordings. The process of creating this database went through several stages, starting with voice recording, noise reduction, storage, and labeling of audio files. This database is also intended to serve other researchers in various fields and standardize research on the Amazigh language in Morocco. The database has not yet been published due to licensing and copyright constraints. The tables below present some characteristics of this database, such as examples of words recorded in English with its Tarifyt pronunciation and writing (See Table 1), the number of speakers participating in word recording by age categories (see Table 2), and Parameters used in recording phase (See Table 3).

Table 1. Examples of words recorded in English with their Tarifyt pronunciation and writing.

English word	Tarifiyt pronunciation	Tifinagh writing
ablution	ËuÄu	ⵇⵗⵄ
to bark	Zu	ⵣⵄ
drink	Ssu	ⵙⵙⵄ
absent	Iveyyeb	ⵉⵖⵄⵔⵔⵄⵜ
accept	Qber	ⵥⵜⵄⵙ
to accompany	Mun	ⵎⵄⵏ
to welcome	Areppeb	ⴰⵙⵄⵃⵃⵄⵜ
donut	Sfenj	ⵙⴼⵄⵏⵊ
profit	lfayda	ⵍⴼⴰⵢⴷⴰ
benzoin	Rebxur	ⵙⵄⵜⵅⵄⵙ
cradle	ddup	ⵃⵃⵄⵃ
corn	irden	ⵉⵙⵃⵄⵏ
injured	amejËup	ⴰⵎⵄⵊⵍⵄⵃ
hurt	jaËep	ⵊⴰⵍⵄⵃ
injury	ajaËip	ⴰⵊⴰⵍⵉⵃ
blue	Aziza	ⴰⵣⵉⵣⴰ
beef	tafunast	ⵜⴰⴼⵓⵏⴰⵙⵜ
able	Izemmar	ⵉⵣⵎⵄⵎⵎⴰⵙ
capacity	Tizemmar	ⵜⵉⵣⵎⵄⵎⵎⴰⵙ
captive	Amepbus	ⴰⵎⵄⵃⵜⵄⵙ

Table 2. Number of speakers participating in word's recording by age categories.

N°	Age Category	Gender		Total
		H	F	
1	Less than 30 years	20	20	40
2	30 years an above	5	5	10
Total		25	25	50

Table 3. Parameters used in recording phase.

Recording attribute & corpus details	Value
Sampling rate (kHz)	16
Bit-depth (bits)	16
Channels	1(Mono)
Audio data file format	.wav
Corpus	520 isolated words
Accent	Moroccan Tarifit Berber

2.2 Speech Recognition Systems (Digits and Letters)

An automatic speech recognition system for Amazigh language was developed using HTK [2-4]. The system is built with context-independent phonetic models, and several choices are made, such as the number of states of the models, the type of emission probability densities associated with the states, and the representation of the signal by cepstral coefficients and differential coefficients. The Amazigh digit and letter recognition system is based on Hidden Markov Models. From the preparation phase of the training data to the testing and analysis of results according to different experiments, the system achieves a recognition rate of 91.31% for MMC = 3 with a combination of 32 Gaussian mixtures (See Table 4), which is a very satisfactory rate [2-4].

Table 4. P Optimal parameters selected for automatic recognition of spoken Amazigh digits and letters using Hidden Markov Model Toolkit [12].

	GMMs				
HMMs	2	4	8	16	32
3	80.83	85.01	89	89.82	91.31
4	82.99	85.86	88.13	88.54	88.66
5	85.46	86.95	88.30	88.28	88.43
6	85.77	87.61	88.11	87.95	87.47
7	86.53	88.25	88.53	87.98	88.43

2.3 Tarifit Word Speech Recognition Systems

A system for automatic recognition of Amazigh-Tarifit speech was developed using CMU Sphinx, through different stages of preparation of the Amazigh word corpus, feature extraction, learning process, and recognition using Hidden Markov Models. Finally, the results were tested and analyzed. The training database used in the different experiments had a size of 187 words (46,750 recordings). The system achieved a satisfactory recognition rate of 92.2%, (See Fig.1 and Fig.2) compared to other systems developed for other languages. This is the only system developed so far containing such a large Amazigh corpus.

	GMM					
Tied state	**2**	**4**	**6**	**8**	**16**	**32**
200	70	75,9	77,9	79,1	82,4	83,9
300	76,3	81,5	83,1	84,8	86,3	87,2
400	80,3	85,2	86,7	87,4	88,4	89,1
500	82,7	86,6	87,5	88,2	89,7	89,6
600	84,4	87,8	88,9	89,1	90	90
700	85,4	88,5	89,2	89,7	90,5	90
800	86,3	89,7	90	90,3	90,6	90,3
900	87,1	90	90,5	90,9	91,4	90,4
1000	87,6	90,4	90,8	91,5	**91,8**	90,4
2000	89,4	91,7	**91,8**	**91,8**	90,8	87,8
3000	89,4	91,7	**91,8**	**91,8**	90,8	87,8
4000	89,4	91,7	**91,8**	**91,8**	90,8	87,8

Fig. 1. In reference to tied triphone and GMM.

Word Accuracy Rate (%)
92,2

Fig. 2. Using clustering tree decision.

3. NLP Resources for Sentiment Analysis of Darija Maghribya (Under Construction)

In recent years, the advent of social media has created a flood of textual data on the World Wide Web. The shared data is voluminous, fast-moving, and diverse, offering new opportunities and posing numerous challenges for machine learning and natural language processing (NLP) in particular. The poor quality, informality, and noise in this data present several challenges. Most of the efforts in sentiment analysis for text classification have focused on English, while research on Arabic, presents many challenges. Arabic is one of the most difficult Semitic languages to handle due to its complex morphology. In this work, a new contribution to Arabic resources is presented as a large set of Moroccan data extracted from different social networks and carefully annotated. To our knowledge, this dataset is the largest Moroccan dataset for sentiment analysis. It stands out for its size and quality. The dataset is multidomain, the figures (Fig.3 and Fig.4) bellows show a few examples of the most word used in the domain of sport and society.

Fig. 3. Most words in sport domain.

Fig. 4. Most words in social domain.

3.1 Role of Created Dataset and Sentiment Analysis System for Darija Maghribiya

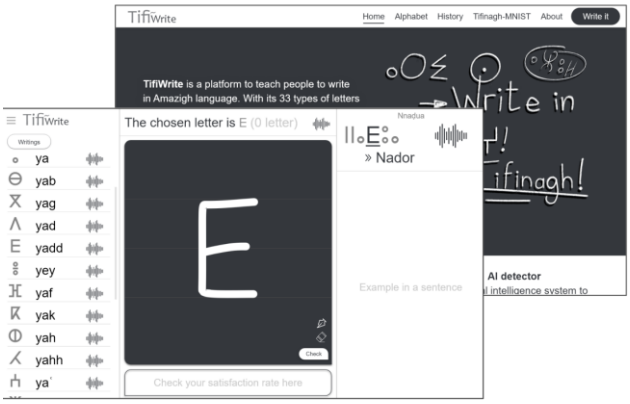
Data plays a crucial role in sentiment analysis. Datasets provide examples of comments or texts as well as labels that indicate the sentiment associated with each text. Here are some of the key roles that datasets play in sentiment analysis:

- Model training: Datasets are used to train sentiment analysis models. Models learn to identify the language features and patterns that are associated with each sentiment (e.g., positive or negative). The more varied and representative the data, the more accurate the model will be.
- Model evaluation: Datasets are also used to evaluate sentiment analysis models. Models are tested with unseen data, and their performance is measured by comparing the model's predictions to the actual labels.
- Model improvement: Datasets are used to improve sentiment analysis models. Model prediction errors can be analysed to identify common errors and improve the models.
- Domain-specific adaptation: Datasets can be adapted to a specific domain, such as beauty products, restaurants, or hotels. This allows for the creation of models that are specific to a domain and can provide more accurate results for that domain.
- Trend analysis: Datasets can be used to track sentiment trends over time. User comments can be collected over a given period, and models can be used to identify changes in sentiment and trends over time.

4. NLP Resources for Amazigh Handwriting Recognition

Generation and classification of the Tifinagh handwritten letter corpus using deep learning models ANN, CNN & GANs. The work presents a new dataset named Tifinagh-MNIST: handwritten letters of the Tifinagh alphabet, which is used to write Tamazight languages. The presented dataset contains 82,500 grayscale images of size 28 × 28 pixels belonging to 33 classes (or letters) (See Fig. 5), with 2500 images per class. Specifically, the training set consists of 66,000 images while the test set contains 16,500 images. Tifinagh-MNIST is intended for the development of AI tools aimed at processing handwritten Tifinagh characters. We also provide use cases of this corpus through neural network models for classification and data generation. The corpus will be made available to the public to promote the development of AI solutions for processing Tamazight texts. The dataset is available at https://github.com/iseddik/Tifinagh-MNIST. An example of use case of this dataset is on developing, it consists of an innovative approach to teaching and learning Tifinagh script through the use of state-of-the-art AI technology (See Fig. 6).

Fig. 5. Tifinagh-MNIST Dataset.



Tifiwrite platform.

Fig. 6: Web-based Tifinagh handwriting learning platform.

4.1 Role of the Created Database and Handwriting Recognition System for the Amazigh Language

Handwriting recognition can be a useful tool for the development of natural language processing (NLP) systems. Here are some examples of how handwriting recognition can be used in the context of NLP:

- Handwritten text recognition: Handwriting recognition can be used to transcribe handwritten text into machine-readable text that can be analyzed by NLP systems. This is particularly useful when dealing with historical documents or archives that are only available in handwritten form.
- Handwriting-based text analysis: Handwriting recognition can be used to analyze the characteristics of handwriting, such as stroke width, angle, and pressure, which can provide additional information about the text beyond its content. This can be particularly useful in forensic analysis and document authentication.
- Handwriting-based language models: Handwriting recognition can be used to develop handwriting-specific language models, which can be used to improve the accuracy of handwriting recognition and generate text with a more natural handwriting-like tone.
Personalization: Handwriting recognition can be used to recognize individual users' writing styles, allowing NLP systems to better understand their preferences and improve recognition accuracy over time.
- Multilingual handwriting recognition: Handwriting recognition can be used to recognize and transcribe handwriting in multiple languages, which can be useful in multilingual NLP applications.
Overall, handwriting recognition can be a valuable tool in the development of NLP systems as it enables the processing of handwritten text and provides additional information beyond the text itself, which can improve the accuracy and usefulness of NLP analysis.

5. Conclusion

This paper summarizes the overall NLP resources and systems developed for the Moroccan Dialect, including the Amazigh speech corpus and automatic speech recognition systems, including standard numbers and letters, and several words in Tarifyt. The main goal is to make these resources available for academic research and use, such as developing speaker-independent speech recognition systems, speech synthesis, speaker recognition, and others. A significant part of this database was used in our research project. A second contribution is in the domain of sentiment analysis for Darija lmaghribya. In this work, a new contribution to Arabic resources is presented as a large set of Moroccan data extracted from different social networks and carefully annotated, which will serve the integration of our Darija into the field of artificial intelligence and decision-making based on social networks. Finally, we described a third contribution in the domain of Amazigh handwriting recognition. Our work is ongoing, and it is directed toward integrating Moroccan dialects into Data Engineering and intelligent systems for Moroccan dialects.

References

1. S. E. Ouahabi, M. Atounti and M. Bellouki, "A database for Amazigh speech recognition research: AMZSRD," 2017 3rd International Conference of Cloud

Computing Technologies and Applications (CloudTech), Rabat, Morocco, 2017, pp. 1-5, doi: 10.1109/CloudTech.2017.8284715.

2. S. El Ouahabi, M. Atounti and M. Bellouki (2016), " Building HMM Independent Isolated Speech Recognizer System for amazigh Language," Europe and MENA Cooperation Advances in Information and Communication Technologies, 2016, pp. Europe and MENA Cooperation Advances in Information and Communication Technologies Volume 520 of the series Advances in Intelligent Systems and Computing pp 299-307 . doi: 10.1007/978-3-319-46568-5_31.Isbn :978-3-319-46568-5. url :http://dx.doi.org/10.1007/978-3-319-46568-5_31.
3. S. El Ouahabi, M. Atounti and M. Bellouki (2016), "amazigh Isolated-Word speech recognition system using Hidden Markov Model toolkit (HTK)," 2016 International Conference on Information Technology for Organizations Development (IT4OD), Fez, 2016, pp.1-7. doi: 10.1109/IT4OD.2016.7479305
4. S. El Ouahabi, M. Atounti and M. Bellouki (2016), "Automatic amazigh Recognition System : An Approach using HMM" 2016, Ecole de recherché CIMPA & Workshop sous le thème : Modélisation, analyses mathématique et numérique pour les problèmes aux dérivés partielles du 09 au 19 mai 2016 à la Faculté pluridisciplinaire de Nador.
5. S. EL Ouahabi, M. Atounti and M. Bellouki (2019), "amazigh Speech Recognition using Triphone Modeling and Clustering Tree Decision", Annals of the University of Craiova, Mathematics and Computer Science Series, Volume 46(1), 2019, Pages 56{65}, ISSN: 1223-6934
6. S. EL Ouahabi, M. Atounti and M. Bellouki (2019), "Toward an automatic speech recognition system for amazigh-tarifit language", International Journal of Speech Technology, 22(2), 421-432, doi:10.1007/s10772-019-09617-6
7. S. EL Ouahabi, M. Atounti and M. Bellouki (2018), " Contribution à la Reconnaissance Automatique de la parole amazighe à base des Modèles de Markov Cachés", 4ème édition scientifique du laboratoire de mathématiques appliqués et systèmes d'information (MASI), 06 et 07 décembre 2018 à la Faculté pluridisciplinaire de Nador.
8. S. EL Ouahabi, M. Atounti and M. Bellouki (2018), "Contribution au développement de corpus et système de reconnaissance vocal pour la langue amazighe", TICAM 2018 : La conférence internationale sur les Technologies d'Information et de Communication pour l'amazighe, IRCAM, Rabat, (2018).
9. S. EL Ouahabi, M. Atounti and M. Bellouki (2017), " Vers une base d'apprentissage pour les systèmes de reconnaissance automatique de la parole en amazighe", 3ème édition scientifique du laboratoire de mathématiques appliqués et systèmes d'information (MASI), 21 et 22 Novembre 2017 à la Faculté pluridisciplinaire de Nador.
10. S. El Ouahabi, M. Atounti and M. Bellouki (2016), "Système de reconnaissance automatique de la parole pour la langue amazighe basé sur les HMM", 2ème édition scientifique du laboratoire de mathématiques appliqués et systèmes d'information (MASI), 16 et 17 décembre 2016 à la Faculté pluridisciplinaire de Nador.
11. S. El Ouahabi, M. Atounti and M. Bellouki (2015), "Système de reconnaissance automatique de la langue amazigh", 1ère édition scientifique du laboratoire de mathématiques appliqués et systèmes d'information (MASI), 16 et 17 décembre 2015 à la Faculté pluridisciplinaire de Nador.
12. Safâa, El Ouahabi & Mohamed, Atounti & Bellouki, Mohamed. (2020). Optimal parameters selected for automatic recognition of spoken Amazigh digits and letters

using Hidden Markov Model Toolkit. International Journal of Speech Technology. 23. 10.1007/s10772-020-09762-3.

13. Safâa, El Ouahabi & Mohamed, Atounti & Bellouki, Mohamed. (2020). HMM-GMM based Amazigh speech recognition system. International Journal of Signal and Imaging Systems Engineering. 12. 47. 10.1504/IJSISE.2020.113564.