# NEW TRENDS IN LESS-RESOURCED LANGUAGE PROCESSING: CASE OF AMAZIGH LANGUAGE

Fadoua Ataa Allah and Siham Boulaknadel

Royal Institute of Amazigh Culture/Center of computer sciences studies, information system and communication, Rabat, Morocco

***ABSTRACT***

*The coronavirus (COVID-19) pandemic has dramatically changed lifestyles in much of the world. It forced people to profoundly review their relationships and interactions with digital technologies. Nevertheless, people prefer using these technologies in their favorite languages.*

*Unfortunately, most languages are considered even as low or less-resourced, and they do not have the potential to keep up with the new needs. Therefore, this study explores how this kind of languages, mainly the Amazigh, will behave in the wholly digital environment, and what to expect for new trends.*

*Contrary to last decades, the research gap of low and less-resourced languages is continually reducing. Nonetheless, the literature review exploration unveils the need for innovative research to review their informatization roadmap, while rethinking, in a valuable way, people's behaviors in this increasingly changing environment. Through this work, we will try first to introduce the technology access challenges, and explain how natural language processing contributes to their overcoming. Then, we will give an overview of existing studies and research related to under and less-resourced languages' informatization, with an emphasis on the Amazigh language. After, based on these studies and the agile revolution, a new roadmap will be presented.*

***KEYWORDS***

*Language Informatization, Informatization Roadmap, Amazigh Language, NLP, Less-Resourced Languages, Green Computing*

## 1. INTRODUCTION

Since 1992, linguists have sounded the alarm bell for languages in danger. As a result, the Endangered Languages Committee was formed. This latter created the International Clearing House for Endangered Languages (ICHEL) research center. In collaboration with UNESCO, ICHEL collected the regional expert reports and published the 'UNESCO Red Book of Endangered Languages', which has been called 'Atlas of Endangered Languages' in the next editions. These studies have noted that 6,700 languages of the 7,000 currently spoken, in the world, are indigenous and the most endangered. Unfortunately, even after three decades after the first alert, this situation is still alarming [1].

Therefore, it is so important that states, communities, and researchers pool their efforts towards a common goal of preserving and promoting native and indigenous languages. To this end, many studies have contributed to the endowment of under and less-resourced languages with resources and tools for ensuring their survival in the new world invaded by technology. Nevertheless, for great results, few of them treat the subject as a whole.
In this context, the present paper is interested in less-resourced languages' informatization process. More specifically, it focuses on existing roadmaps with the aim to propose careful

planning taking into consideration the recent challenges. Thus, it will introduce, in Section 2, the technology challenges confronting less-resourced language communities and how natural language processing (NLP) could contribute to overcoming them. In section 3, it will present a part of the-state-of-the-art of less-resourced languages' informatization studies, and discuss the strengths of each. While in Section 4, it will give a brief overview of the Amazigh language situation, and propose new directives, based on the agile revolution. Finally, in Section 5, the paper will draw conclusions.

## 2. HOW NLP COULD OVERCOME TECHNOLOGY ACCESS CHALLENGES?

Since its beginning in the second half of the 20th century, the Digital Revolution, known also as the Third Industry Revolution and the Information Age, has impacted work practices and consequently lifestyles, especially after the appearance of the Internet, mobile devices, and social networking. Nevertheless, the emergence of this impact did not affect all populations. When developed countries experienced the Fourth Industrial Revolution (4IR), by blurring the borders between the physical, digital, and biological worlds; others still viewed technology as a luxury.

Fortunately, this vision did not last for long. During the COVID-19 pandemic, several communities have felt the real need to integrate technologies into daily life. Since, technologies became an interesting tool for inclusion, communication, learning, productivity, creativity, etc., even in developing and least developed countries. Hence, many communities in these countries have tried to create plans with the aim to weave digital technology throughout their life.

### 2.1. Technology Access Challenges

To lead the development of a technology plan, in this context, communities faced several challenges. These challenges guide them to specify the roadmap steps. The first and main challenge is encouraging adjustment to the evolving process. Given that humanity is living in its 3rd and 4th industrial revolutions, the world is changing exponentially. Thus, the individual, during his(er) career, is required to change his work practices, unlike in previous centuries, where the changes were really minimal. Therefore, nowadays, the individual is supposed to be highly agile.

The second is providing literacy in computational thinking to let technology more user-friendly. On the other hand, let individuals at all levels using technology in many areas of their lives: from communication through social media to shopping online, and from working remotely assisted by web applications to entertainment using on-demand services. Thus, these systems not only will facilitate the various aspects of the individual life but, furthermore, (s)he will contribute to building reliable shared resources and progressively will get an active role in creating technology. As a result, computational thinking became a lens and a set of categories for understanding the algorithmic fabric of today's world [2].

The third challenge is ensuring high-quality and equitable education, whatever the circumstances. The school closures during the COVID-19 pandemic exposed deep inequities within education due to the lack of infrastructure and special materials for online education, to name just a few. So, it is important to help students being independent learners, in a way that they become the leaders of their own learning by giving them more agency, and disrupting the old routines of receiving knowledge from the teacher. On the other side, help schools to complete their digital transformation by getting software solutions that enable active learning or at least blended one (mixed-mode instruction).

The fourth is struggling with the lack of technology for indigenous and local languages in addition to languages in use. Technology has made it easier to access and transmit information with minimal effort, to enjoy exchanging conversations from anywhere in the world at an extremely low rate, and to ensure better participation in society on an equal basis with others. Nevertheless, how can the end users avail of this technology when they could not know how to use it or understand its content? Thus, the communities need to address issues facing their population, and remove barriers that prevent them from fully benefiting from technological advancements. In this perspective, they could encourage developers and operators to localize their local cultures, by adapting technological products to their languages in a manner that they integrate at least their specific graphics and local currencies, in addition to translate the end users' interfaces.

While the last challenge is enabling infrastructure enhancements. The ideal idea is to implement smart city initiatives. However, at a minimum, communities must provide network infrastructure to underserved areas; connect facilities; and offer secure, flexible, and redundant infrastructure, with the aim to meet current and future data, video, and voice communications needs. Furthermore, they should control the technology costs, and could choose to partner with sponsors for particular services for the benefit of fragile families.

## 2.2. NLP a Solution to Challenges

Natural language processing (NLP) has existed for more than 50 years. It refers to the branch of artificial intelligence that combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these approaches enable machines to get, in a smart and useful way, the ability to analyze, understand, and derive meaning from written and spoken languages, in much the same way human beings can. If the human has different sensors, such as eyes to see and ears to hear, machines have programs to read text and microphones to collect audio. While, the human brain processes these inputs, machines' programs do the same, and at a specific step, they convert their respective inputs to digital codes that the machines understand.

The goals pursued with NLP can be classified into two superordinate categories: Understanding language and generating language [3]. These categories are so elementary for human-machine interaction technologies that, in the last few years, NLP is no longer foreign to this kind of technology. Indeed, it deals with the different forms of information created, modified, or exchanged with humans. Since that, the current language technologies have changed the fate of different languages, and are contributing to facing many challenges.

Certainly, personal development is the hope of everyone. Nevertheless, it is a lifelong process, improving the quality of life, and achieving deep aspirations. It can affect different aspects of life, such as personal, familial, professional, spiritual, and financial. Furthermore, the observed actions and progress in one aspect have, generally, positive repercussions in the others. Nowadays, technology has facilitated this process. Based on self-knowledge, it is better that individuals tend to realize their aspirations themselves through specific objectives: improving skills, talents, potential, employability, etc. To achieve these goals, they can look online for the respective training, workshops, and best practices. So, necessarily, they will use search engines based on natural language processing to find relevant information.

This technology practice cannot in any way be anchored in the users' attitudes or rooted in their habits if they are not really satisfied with its use. In other words, the availability of interesting functionalities will encourage the use of technology in daily life, especially if they are presented and their user manual is explained to individuals. Fortunately, at present this is made through

online advertisements. Nonetheless, the central question is how to guarantee the functionalities' importance and fulfill their interest?

Before answering this focal question, it should be noted that the majority of functionalities that the individual needs mainly concern communication through voice over internet protocol services, instant messaging applications, and social media platforms for a better exchange; studies and work by means of meeting tools and video conferencing applications for more collaboration between coworkers and competitiveness of businesses; in addition to divertissement via digital games and virtual reality for more fun. Based on the study of these mediums, it appears clear that NLP is the core success of most of them. Whatever the aim of these needs, all are related at least to one of the human language types: verbal, non-verbal, written, or visual. Consequently, once an individual is using these mediums, (s)he is no longer just a simple beneficiary but furthermore (s)he becomes an active actor in the process of data creation.

One type of interaction with these functionalities is learning, where an individual can even be a learner or an educator. With the technology evolution, many platforms have emerged for both cases. They could be arranged into three categories: adaptive platforms that typically provide activities according to learner's level and improvement areas; collaborative ones allowing group collaboration, and focusing on learners' interactions; and gaming platforms which offer the possibility of learning through gaming. Obviously, for the three categories, NLP has a great part in the process of these materials building.

For better exploitation of these materials, it is judicious that individuals can find their native, local languages and those in use already integrated. Therefore, operators must undertake localization technology in platform building. This technology ensures translation, editing, and proofreading that inevitably calls on a set of processes, starting from language codification up to advanced tools. Nevertheless, platform building needs to be closer to the agile approach, allowing a progressive addition of new modules, mainly for non-integrated languages.

## 3. WHAT HAS BEEN DONE TO PROCESS LESS-RESOURCED LANGUAGE?

In this age of information, NLP techniques are increasingly applied to new technologies. By taking into consideration the widely and daily used 'Internet', everyone can notice that its content is even processed or created by NLP techniques. However, the majority of NLP advancements, in the world, to date have been restricted to English as a language. Yet, there is still a glaring mismatch between English and non-English NLP models. Joshi et al. [4] pointed out that more than 88% of the world's languages, even spoken by around 1.2 billion people, have been and are still ignored in the language technology aspect. Unfortunately, the lack of NLP technology support for these languages will reduce the degree to which users are exposed to them, and implicitly will generate a downward spiral: less contribution to the content creation in these languages, will inevitably lead to the scarcity of resources, which will in its turn hinder the development of NLP-based technologies.

To overcome this situation, many researchers and companies made several efforts to lower barriers confronting languages, mainly the less-resourced ones. In this section, we focus on existing research that proposes and studies the process of language informatization in its entirety. Furthermore, we discuss the strengths of each.

According to the literature, to the best of our knowledge, this kind of reflection was first conducted by individual researchers or small groups, such as Sarasola [5], Agirre et al. [6], Diaz de Ilarraza et al. [7], Streiter and De Luca [8], and Berment [9]. Nevertheless, due to the importance of the problem, many networks were funded, both in the USA and in Europe, uniting organizations from several countries. These organizations, even academic institutes or private

companies, are all active in language and speech technology. One of the first projects piloted by these networks was the BLARK roadmap [10] that have been followed by many others over the world. They comprehend, among others, the Technology Development of Indian Languages (TDIL) program [11], African Languages Technology Initiative (ALT-i) [12], the Digital Language Diversity Project (DLDP) [13], and the Canadian Indigenous Languages Technology (ILT) project [14].

In regard to the high number of studies on this topic, we decided to present in the remainder of this section the first experience of Sarasola strategy [5], the oldest launched project by the European Commission - BLARK roadmap [10], the project concerning Amazigh language, the object of this work - Ataa Allah and Boulaknadel strategy [15], one of the European Commission recent project - DLDP project [13], and recent work in the field - Zhang et al. proposal [16].

### 3.1. Sarasola Strategy

Thanks to twelve years of experience in the IXA group with the automatic processing of Basque, Sarasola defined a strategy for the development of language technology in minority languages [5]. The strategy distinguished three levels that, according to the author, have to be progressively developed in a parallel and in concert. The first level ‘Applications’ concerns commercial systems intended for nonspecialized users. The second one ‘Tools’ focuses on systems oriented to application developers. While the third level ‘Language foundations’ includes essential research to create any tool or application. Nevertheless, at the same time, the author takes into account some tools and applications that are helpful in research and improving language foundations. Each item of the three levels is placed in different columns with respect to the linguistic knowledge it needs: lexicon, morphology, syntax, semantic, and speech.

Furthermore, these items are presented in five scalar phases: The initial phase ‘Laying foundations’ includes corpus I, composed of raw text without any tagging mark; lexical database I, which could be a list of lemmas and affixes; machine-readable dictionaries, such as monolingual and bilingual dictionaries, beside thesaurus; morphological description, formalizing morphological phenomena; speech corpus I, consisting of speech recordings; and description of phonemes. The second phase ‘Basic tools’ contains statistical tools for the treatment of corpus, like bigram and trigram frequencies, word count, collocations, and co-occurrences; morphological analyzer/generator; lemmatizer/tagger; speech processing at word level; corpus II, incorporating word-forms tagged with their corresponding part of speech and lemma; in addition to lexical database II that includes in a second version part of speech and morphological information. The third phase ‘Tools of medium complexity’ comprises an environment for tool integration to allow the integrated use of the available tools; spelling checker and corrector; web crawler; surface syntax to recognize simple syntactic constituents such as verbs, noun phrases, or prepositional phrases; structured versions of dictionaries, allowing sophisticated queries; as well as lexical database III, extending the previous version with multiword lexical units. The fourth phase ‘Advanced tools’ contains corpus III, composed of syntactically tagged text; grammar and style checkers; integration of dictionaries in text editors; lexical-semantic knowledge base, ensuring taxonomy of concepts representation; word-sense disambiguation; speech processing at sentence level; and language learning systems. The fifth and the last phase ‘Multilinguality and general applications’ includes corpus IV, dealing with semantically tagged text; information retrieval and extraction; translation aids and dialog systems.

Besides these works, the author noted that there is a previous and necessary phase for languages not using Latin characters. In this phase, the written representation of words needs to be defined.

## 3.2. BLARK Roadmap

BLARK is the abbreviation of the Basic Language Resource Kit. The concept of this project was defined as a joint initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association) [17]. However, the first party that adopted this concept was the Dutch Language Union. They produced a rather detailed definition of the BLARK, with respect to the linguistic components, the priorities for its completion, and with respect to proposals for a management, maintenance, and distribution structure [5].

As a roadmap for all languages, regardless of their size or importance, it was intended to be language-independent, helping to improve starting circumstances conditions for research, education, and development in language and speech technology. Furthermore, it aimed to reduce the efforts for defining required resources for any work, and to ensure interoperability and interconnectivity.

The roadmap has the shape of a collection of objects (challenges, milestones) in different categories (resources, technologies, and applications). In the beginning, the roadmap activities covered some topic areas, such as speech technology, machine translation, multimodality, and knowledge management. The resource roadmap's data collection and milestones has been initiated in close cooperation with the ENABLER (European National Activities for Basic Language Resources) network. Within this network, ELDA (Evaluations and Language resources Distribution Agency) implemented and improved the original BLARK matrices to be available for as many languages as possible, and highlight the gaps regarding language resources (LRs) needed for Human Language Technologies (HLT) [18]. The BLARK matrices cross-link a list of potential applications and modules with the LRs needed and corresponding languages. They are of the order of four matrices: Applications versus Modules and Language Resources versus Modules for both spoken and written language. Furthermore, they specify the importance of resources for modules and the importance of modules for applications. The different resources, modules, and applications proposed in these matrices are compiled in Table 1.

Table 1. Overview of resources, modules, and applications proposed in the BLARK matrices.

| **Data** | **Language module** | **Speech module** | **Applications** |
|---|---|---|---|
| Monolingual lexicon | Diacritizer | Acoustic models | Text-to-Speech |
| Multi/bilingual lexicon | Morphological composition | Language identification | Summarization |
| Thesauri, ontologies | Named entity recognition | Emotion identification | Machine Assisted Translation |
| Unannotated corpora | Pos disambiguator/tagger | Language models | Machine Translation |
| Annotated corpora | Semantic analysis | Lexicon adaptation | Information retrieval/filtering |
| Speech corpora | Sentence boundary detection | Lips movement reading | Indexing |
| Phonetic lexicon | Sentence synthesis and generation | Phoneme alignment | Information extraction |
| Audio data with markers | Shallow parsing | Pronunciation lexicon | Document Production |
| Telephony | Syntactic analysis compound | Prosody prediction | Dialog systems |
| Parallel multilingual corpora | Term extraction | Prosody recognition | Classification |
| Onomastica (proper names) | Grapheme recognition (for handwritten OCR) | Segmenter speech/silence | Automatic speech recognition/dictatio n |
| Multimodal corpora for (hand) OCR | Grapheme recognition (for typewritten OCR) | Sentence boundary detection | Transcription of conversational speech |
| Multimodal corpora for (typed) OCR | Alignment | Speaker adaptation | Transcription of broadcast news |
| BNSC | Word sense disambiguation | Speaker recognition/identific ation | Topic detection, segmentation, boundaries |
| Desktop/Microphone & high quality | Transfer tool (software) | Speech unit selection | Telephony speech |
| Visual data (faces, lips, etc.) | | Speech/non-speech music detection | Speaker recognition, adaptation |
| | | Word boundary identification | Speaker-to-speaker mapping |
| | | | Lips movement reading |
| | | | Lips movement generation |
| | | | Emotion/prosody output |
| | | | Emotion identification |
| | | | Embedded speech |

## 3.3. Ataa Allah and Boulaknadel strategy

Due to the transformation that the world underwent from the industrial economy into the information economy, the survival of many languages and their associated cultures was becoming

threatened. The tendency in investigating applying NLP technology to a language is, generally, driven by its economic and political importance. Therefore, with the aim to reduce the language technology gap between developed and endangered languages, Ataa Allah and Boulaknadel proposed a revitalization strategy [15]. Inspired by the Amazigh language situation, this latter is based on eight keys:

(1) Linguistic contribution: This point involves matching or modeling language skills by discovering and presenting explicitly the rules governing this language to aid in its processing. Furthermore, it advises letting this work be shared and collaborative, to avoid reduplication and wastage of efforts and resources.
(2) Resource recycling: This action recommends standardizing released resources in a suitable format that will allow their reuse in different research, tools, and applications. Moreover, it will enable data enrichment.
(3) Adapting computational language processing (CLP) techniques: Based on existing techniques, this operation makes it possible to build specific tools for other languages, taking advantage of the similarity of linguistic features.
(4) Extensibility focused: According to this direction's guiding idea, any project should be designed in such a way that it can be easily expanded to another level or to another language.
(5) Open-source focused: Since computational language processing involves lots of funds and most endangered languages are economically poor, adopting an open-source strategy will get around this obstacle and cut down on the financial issues. Furthermore, it will allow the implementation of the two previous directions.
(6) Professional documentation: Documentation will also greatly help in the continuation and extension of projects. This documentation will assist people who may be interested in the use of a project, or permit them to access any phase of the work and continue its development.
(7) Evaluation system: This action is the process allowing to measure the gap between fixed objectives and attained results. To do, three evaluation steps have to be undertaken: (1) before realizing the project, to determine its objectives and prerequisites; (2) during development, to guide and direct the development advancement; and (3) after the implementation, to yield the satisfaction level and relevance, as well as its durability, continuity and extensibility.
(8) Roadmap: Conscious of the key role that can search engines, machine translation, human-machine dialogue, and e-learning play in the survival of endangered and less-resourced languages, the authors undertook a deep study on them. This latter shows that the four projects are mutually related to each other, in a way that one can act as a part of the other, and they are based on various processing. Therefore, the authors identified a list of the needed processes and resources. Additionally, they represented them by a chain starting from elementary processing, passing by the constitution of linguistic resource bricks, and going towards real applications. Thus, the authors provided a roadmap arranging them chronologically in short, medium, and long terms.

    The short-term phase, considered as an initial step, mainly consists on the identification of the language encoding; the primary resources building, such as keyboard, fonts, basic lexical database, and elementary corpora; besides basic tools and applications, like encoding converter, sentence and token splitter, basic concordancer, web search engine, morphological generator, and optic character recognition system. While the medium-term phase focuses on advanced tools and applications that could be even rule-based or statistical, based on the size and the representativity of the elaborated resources. The tools specified in this step are a stemmer or lemmatizer, part-of-speech tagger, chunker, syntactical analyzer, and speech recognition. In their turn, these tools will enable to build

spell-checkers, terminology extractors, text generators, and human-machine dialogue applications. They will also enable the improvement of the first phase tools and applications. Simultaneously, in this phase, the required resources for the following step can be prepared, including multilingual dictionaries, multilingual aligned corpora, and semantic annotated corpora. Whereas the authors considered the long-term phase of the roadmap as the synthesis phase of the realized work. Besides the creation of a pronunciation lexicon, Word Net, word-sense disambiguator and a speech synthesizer, this phase focuses also on multilingual applications.

### 3.4. Digital Language Diversity Project

To build a sustainable policy for safeguarding and promoting languages, policies for digital development have to be involved. Thus, the Digital Language Diversity Project (DLDP) was proposed for European regional and minority languages. This project is a partnership set up by Consiglio Nazionale delle Ricerche, ELEN, Elhuyar, SOMU at the University of Mainz, and the Karelian Language Society. Its assignment is to advance the sustainability of their languages in the digital world, by empowering speakers with the knowledge and skills to produce and share content on digital devices by means of their native languages.

The core activities of the DLDP project is to survey: the current range, availability, and accessibility of digital content in European regional and minority languages, and the usability of those languages over digital media and tools [12]. It serves as a starting point for the development of an instrument – generically applicable – for assessing the level of digital fitness of a language. After identifying the types of actions that must to be taken in accordance with the different levels of digital fitness, a training program was developed, to give concrete, hands-on advice about how to perform those actions.

Then, the DLDP identified the ‘Digital Language Survival Kit’, which is an instrument allowing languages speakers and communities to (1) self-assess the degree of digital capability of their language, by locating current gaps and areas where action necessities to be taken; (2) learn about what tangible actions and enterprises can be implemented based on the specific digital fitness level recognized. For example, a minimal degree of digital fitness will involve a level of digital: safeguarding connectivity; adopting a standardized encoding; elaborating a standardized orthography and some basic language resources, at least a corpus, a spellchecker, and a lexicon. While greater levels of digital fitness will necessitate other types of measures, such as creating or enriching a Wikipedia in the language, and pushing for having a localized version of important sites, main operating systems, and social media interfaces.

### 3.5. Zhang et al. proposal

Aware of the importance of languages in maintaining world cultural diversity, Zhang et al. proposed a roadmap to revitalize endangered languages [16]. This roadmap addresses three steps:

(1) Before NLP: In this step, the authors advise NLP practitioners, especially those who are often outsiders of the language communities, to be conscious of three important principles when contacting indigenous speakers: (1) respect, reciprocity, and understanding cultural practices and social norms; (2) decolonizing research in a manner to not privileging NLP familiar methodologies to the detriment of community knowledge and kinship practices; (3) building an open community supported by technologies before performing the research, in order to foster common attitudes, to find common interests, and to set up common goals, which will help to support wide and long-lasting cooperation between indigenous speakers, language learners, and NLP practitioners.

(2) NLP for language education: the second step focuses on applying NLP techniques in assisting language education through three approaches: (1) automated quiz generation to assist instructors, that are generally few, in designing quizzes, and to increase playfulness and language learning games; (2) automated assessment through some easier and feasible assessments, such as language learning quizzes based on automatic error analysis and providing suitable learning plans based on readability approaches; (3) community-based language learning using online and collaborative language learning platform.

(3) Language-specific NLP research: The last step presents NLP resources and tools that seem appear beneficial for community members and might be able to expand the language's usage domains.. For resource collection, the authors suggest collaborative manners: platforms to share individual resources with permission; applications for game-with-purpose, increasing fun and engaging contributors in data collection, in condition to focus on what interests the community more; in addition to automatic data mining methods to crawl and structure resources. Besides, the authors recommend open-source NLP tools for free and wide use by the community. In accordance with native speakers, they are interested in machine translation and its training data, optical character recognition, and its evaluation sets, speech recognition and synthesis, tokenization and morphology parsing, and additionally to POS-tagging and dependency parsing.

### 3.6. Discussion

Endangered languages' development has aroused the interest of several researchers, teams, and organizations. However, each of them has specific guidelines. With reference to the strategies presented above, it could be noted that Sarasola roadmap [5] has the advantage to be illustrated according to three dimensions: the first refers to the category of the performed product - Application, Tool, and Language foundation -. The second dimension concerns linguistic knowledge - Lexicon, Morphology, Syntax, Semantic, and Speech. While the third dimension is related to the project lifetime that is scaled into five phases. Whereas the BLARK roadmap ensures a wide and depth coverage of the nature of language resources, modules, and applications proposed in its matrices [18]. Furthermore, it specifies the level of need of each language resource into specific modules, and of each module into specific applications, either in the spoken or written field. Regarding Ataa Allah and Boulaknadel strategy [15], the roadmap describes the process of endangered and less-resourced languages' informatization from scratch, where it includes the alphabet codification and the keyboard layout creation. In addition, it is accompanied by a set of directives helping to ensure the extension and reuse of products, as well as their evaluation. Concerning the DLDP project, it helps to assess the degree of digital capability of a language and to undertake the appropriate actions according to its digital fitness level. In regard the last roadmap sample [16], it is distinguished by the proposal of an interesting step that they called 'Before NLP'. In this step, they focus on some pieces of advice for the project practitioners, especially for those outside the language communities. Moreover, they suggest building an open community and encouraging collaborations between indigenous speakers, language learners, and NLP practitioners. Even the importance of all these strategies, within a few years, they need to be updated to deal with the new challenges, specifically the ones related to recent developments in technology. In this context, the next section will present the proposed updates and discuss the reasons behind them.

## 4. WHAT NEW DIRECTIVES TO TAKE?

Language informatization is a lifelong process. It begins with the integration of the language into the digital world, toward the alphabet codification and the keyboard layout creation; continues to develop, with the aim to use the language in daily life; and accompanies continuously recent

innovations, to confront new challenges and to ensure language survival. To guarantee the success of this project, its management basically involves the following six keys:

(1) Make an inventory: This involves carrying out a state-of-the-art before launching the project. It is essential to identify the context in which the project is launched and its challenges. It is obvious that the treatment of an under or less-resourced language is not necessarily like the treatment of a well or high-resourced one.

(2) Determine human resources: Any project requires resources that must be diagnosed upstream. Resource verification is a key step of the project since the presence of a skill in the project team does not necessarily imply its availability. The absence of certain skills is likely to block the project or even significantly delay its progress. So, the project manager must ensure that each member of the team knows his role as well as those of the other members to establish fluid communication.

(3) Plan all the tasks: It should be noted that certain kinds of projects require both macro and micro-planning. In our case, the macro-planning concerns roadmap fulfillment. While micro-planning is about scheduling tasks related to the realization of one item on that roadmap. In the latter case, some managers plan all the tasks upstream, while their associates have good practices that should be taken advantage of during the planning phase to ensure the feasibility of a task and optimize its execution. Moreover, whatever the project and its duration, it is necessary to set intermediate objectives, broken down into sub-objectives, to separate challenges and give a better rhythm to the progress of the project.

(4) Put in place appropriate communication: The success of a project often depends on the good flow of information within the team. In this sense, it is often necessary to bring the project team members together to allow them to communicate easily with each other, and to be able to make proposals. All useful information must be easily accessible to each member. Knowledge sharing must be ensured so that everyone can build skills for the success of the project, or at least get a more precise vision of everyone's work.

(5) Set up a follow-up: If planning is fundamental, the ability to follow the progress of each member of the team and the success of each task is essential. It can allow everyone to organize themselves and consider an unexpected constraint in the progress of the project. The fact to make the problems encountered easily explicit to let each member of the team can visualize them, help to solve them, and gain experience for future projects.

(6) Knowing how to solve problems: The advantage of the project mode is its collaborative form. Thus, problem-solving should not be restricted to the hierarchy, but it is preferable to seek solutions with varied profiles with different expertise.

(7) The following subsections give a brief overview of the Amazigh language processing state-of-the-art, and present the new strategy for the informatization of this language, based on the advantages of the strategies presented and discussed in Section 2.

## 4.1. Amazigh Language Processing

Over the past several years, numerous resources and tools for both written and spoken Amazigh have been created, varying in size and breadth of scope. These tools and resources were developed as a result of activities within the research community, especially at the Royal Institute of Amazigh Culture [19], as well as by private individuals, sometimes in collaboration with national universities. These resources and tools can be classified as follows:

- Lexicons: To date, numerous lexical resources have been developed, ranging from small-scale wordlists to large-scale collections of lexical entries with associated morphological information.
- Corpora: Now, corpora include a large corpus of written Amazigh in a variety of genres. In addition, other text collections have also been created for specific purposes. These corpora have different modes of access (including, for example, publicly searchable web interfaces) as well as different levels of annotation (including, for example, part of speech tagging).
- Tools and applications: There is also a significant number of products that handle Amazigh, ranging from text annotation tools, such as tokenizer and part-of-speech tagger, and outgoing to authorship support applications, such as spellchecker.

## 4.2. The New Strategy

We note the importance of technology for the future of a language, and the necessity for the technology to support Amazigh, in order to be used in as many situations as possible. This strategy recognizes technology as a priority area to be addressed to ensure a place for Amazigh in our future lives. It is not starting from scratch; it builds on foundations already laid by Ataa Allah and Boulaknadel work [15]. It gathers the main language technology components and projects to produce a national ambition for language technology. This ambition will care for current and future Amazigh speakers in education and in their professional and social lives.

The initial strategy proposed eight keys: linguistic contribution, resource recycling, adapting CLP techniques, extensibility focused, open source focused, professional documentation, evaluation system, and roadmap. However, in this version, a new key is added, which is green computing. The latest reports from the International Energy Agency draw attention to the excessive consumption of digital devices [20]. Most of this consumption occurs during the “sleep” periods of the devices to maintain their connection to the network. Thus, this action suggests favoring offline applications as much as possible to optimize connectivity to networks and consequently reduce energy consumption.

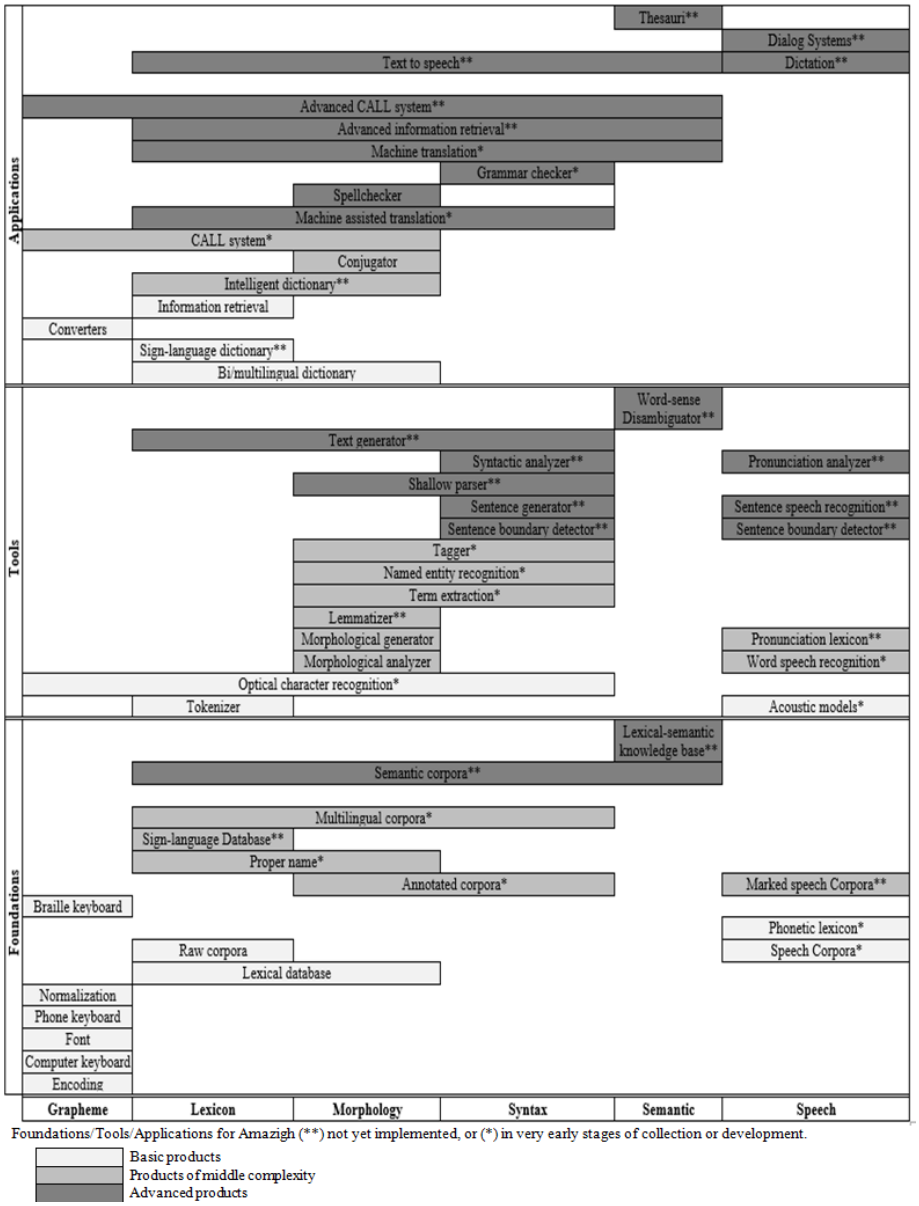


Figure 1. Amazigh language revitalization roadmap

Based on the initial roadmap, the new one, illustrated in Fig. 1, takes into consideration the advantages of our collective experiences at the Royal Institute of Amazigh Culture and those of the discussed strategies. Hence, it is envisaged to be integral, displaying the computerization process in its entirety and considering persons with disabilities' needs. Furthermore, it will structure the different products within three dimensions, namely: the dimension of realization complexity - Basic products, Products of middle complexity, and Advanced products; the dimension of the products' nature - Foundation, Tool, and Application; in addition to the linguistic knowledge dimension - Grapheme, Lexicon, Morphology, Syntax, Semantic, and Speech. It's crucial to keep in mind that this roadmap is based on (1) end-user applications that would be of most practical benefit, according to the Amazigh language status and use inside Morocco; (2) tools that are oriented to application developers and would provide the most reliable benefit with respect to the required resources and development efforts; and (3) foundation

based on digital linguistic resources and processing that would be necessary to build, improve, or evaluate tools and applications.

In this roadmap, frameworks and approaches that could be used for collecting resources and developing processing are not discussed. However, according to the adapting CLP technology, it is fruitful to exploit recent methods. Collecting resources started with modest individual initiatives constrained by a limited digital infrastructure; then, it was improved by technological evolution, which allowed, nowadays, a large and collaborative collection via crowdsourcing. For processing, the rule-based approaches were the only ones used at the beginning; then, the statistical methods become widely used; while today, deep learning approaches take over. In the future, collecting and processing must follow new trends.

## 5. CONCLUSION

In the modern digital world, technology plays a central role. This paper briefly described technology access challenges and delineates how natural language is tackling them, especially for supporting under and less-resourced languages. This endeavor is so important and needs to deploy an effective strategy to succeed in the informatization language process. In this context, the paper described some relevant strategies and roadmaps that have been proposed to revitalize endangered languages and discuss the advantages of each one.

From these experiences, the present paper (1) proposes new directives concerning the management of the informatization language process, respecting agile methodologies; (2) suggests the addition of the green computing key to the first strategy for revitalizing a language; and (3) updates its roadmap. This roadmap would be a benefit for native speakers as well as for generations of learners who are striving to find a place for indigenous languages in their modern lives.

## REFERENCES

[1] IESALC Homepage, https://www.iesalc.unesco.org/en/2022/02/21/a-decade-to-prevent-the-disappearance-of-3000-languages/, last accessed 2022/11/02.

[2] Lodi, M. & Martini, S., (2021) “Computational Thinking, Between Papert and Wing”, Science & Education, Vol. 30, No. 4, pp. 883–908. https://doi.org/10.1007/s11191-021-00202-5

[3] Devi, M. I. & Purkayastha, B. S. (2018) “Advancements on NLP applications for Manipuri language”, International Journal on Natural Language Computing, Vol.7, No.5, pp.47-58. DOI: 10.5121/ijnlc.2018.7505 47

[4] Joshi, P., Santy, S., Budhiraja, A., Bali, K. & Choudhury, M., (2020) “The state and fate of linguistic diversity and inclusion in the nlp world”, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6282–6293. Online.

[5] Sarasola, K., (2000) “Strategic priorities for the development of language technology in minority languages”, Proceedings of Developing language resources for minority languages Workshop: re-usability and strategic priorities, LREC’00, pp. 106–109, Athens, Greece.

[6] Agirre, E., Aldezabal, I., Alegria, I., Arregi, X., Arriola, J. M., Artola, X., Diaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Sarasola, K. & Soroa A., (2002) “Towards the definition of a basic toolkit for HLT”, Proceedings of the Workshop Portability issues in Human Language Technologies, LREC’02, pp. 42–48. Las Palmas de Gran Canaria, Spain.

[7] Diaz de Ilarraza, A., Gurrutxaga, A., Hernaez, I., Lopez de Gerenu & N. Sarasola K., (2003) “HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities”, Streiter O. (eds.), Proceedings of the Workshop Traitement automatique des langues minoritaires et des petites langues, 10th conference TALN, pp. 243–252. Batz-sur-Mer, France.

[8] Streiter, O. & De Luca, E. W., (2003) “Example-based NLP for Minority Languages: Tasks, Resources and Tools’”, Streiter, O. (eds.): Workshop Traitement automatique des langues minoritaires et des petites langues, Conference TALN, pp. 233–242. Batz-sur-Mer, France.

[9] Berment, V., (2004) Méthodes pour informatiser des langues et des groupes de langues « peu dotées », Ph.D. Thesis at Joseph Fourier University, Grenoble, France.
[10] Krauwer, S., (2003) "The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap", Proceedings of the International Workshop Speech and Computer, SPECOM'03. Moscow, Russia.
[11] Vikas, O., "Language Technology Development in India. Ministry of Information Technology, New Delhi", India. http://www.emille.lancs.ac.uk/lesal/omvikas.pdf
[12] ALT-i Homepage, http://www.alt-i.org/, last accessed 2022/11/08.
[13] DLDP Homepage, http://www.dldp.eu/en/content/project, last accessed 2022/11/08.
[14] Kuhn, R., Davis, F., Désilets, A. & et al., (2020) "The Indigenous Languages Technology project at NRC Canada: an empowerment-oriented approach to developing language software", Proceedings of International Conference on Computational Linguistics, pp. 5866–5878.
[15] Ataa Allah, F. & Boulaknadel, S. (2012) "Toward Computational Processing of less Resourced Languages: Primarily Experiments for Moroccan Amazigh Language", Advanced Text Mining, pp. 197–218. Rijeka: InTech.
[16] Zhang, S., Frey, B. E. & Bansal, M., (2022) "How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language", Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Vol. 1: Long Papers, pp. 1529–1541. Dublin, Ireland.
[17] Krauwer, S., (1998) "ELSNET and ELRA: Common past, common future. ELRA", Newsletter Vol. 3, No. 2.
[18] BLARK homepage, http://www.blark.org/aim.php, last accessed 2022/11/08.
[19] Ataa Allah, F. & Bouhjar, A., (2019) "The IRCAM Realizations for the Amazigh Preservation and Revitalization in Morocco", Proceedings of International Conference on Language Technologies for All. Paris, France.
[20] https://fr.wikipedia.org/wiki/Efficacit%C3%A9_%C3%A9nerg%C3%A9tique_(%C3%A9conomie)#:~:text=Un%20rapport%20de%20l'Agence,en%202020%20et%20500%20milliards