LINGUA POSNANIENSIS
LXVII
2025 (1)

THE POZNAŃ SOCIETY FOR THE ADVANCEMENT OF THE ARTS AND SCIENCES
PHILOLOGICAL AND PHILOSOPHICAL SECTION
THE COMMITTEE OF LINGUISTICS
in co-operation with
ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ

LINGUA
POSNANIENSIS

REVIEW OF GENERAL
AND COMPARATIVE LINGUISTICS

LXVII (1)

POZNAŃ 2025

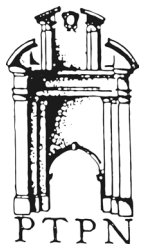
PUBLISHING HOUSE OF THE POZNAŃ SOCIETY FOR THE ADVANCEMENT
OF THE ARTS AND SCIENCES

P O Z N A Ń S K I E T O W A R Z Y S T W O P R Z Y J A C I Ó Ł N A U K
WYDZIAŁ FILOLOGICZNO-FILOZOFICZNY
KOMISJA JĘZYKOZNAWCZA
we współpracy
z UNIWERSYTETEM IM. ADAMA MICKIEWICZA W POZNANIU

LINGUA
POSNANIENSIS

CZASOPISMO POŚWIĘCONE JĘZYKOZNAWSTWU
PORÓWNAWCZEMU I OGÓLNEMU

LXVII (1)



POZNAŃ 2025

WYDAWNICTWO POZNAŃSKIEGO TOWARZYSTWA PRZYJACIÓŁ NAUK

Adres Redakcji / Editorial Office
Uniwersytet im. Adama Mickiewicza
al. Niepodległości 4, 61-874 Poznań, Poland
e-mail: linguapo@amu.edu.pl

Informacje dotyczące wskazówek dla autorów są dostępne online pod adresem:
https://pressto.amu.edu.pl/index.php/linpo/

All information regarding notes for contributors is available online at:
https://pressto.amu.edu.pl/index.php/linpo/

Publikacja sfinansowana przez Uniwersytet im. Adama Mickiewicza w Poznaniu.

This publication has been financed by Adam Mickiewicz University in Poznań.

ISSN 0079-4740
eISSN 2083-6090

Wersja drukowana czasopisma jest wersją pierwotną.

TABLE OF CONTENTS

DOI: 10.14746/linpo.2025.67.1.1

Moroccan Arabic: New advances in researching its multilingual practices and digital spaces

Sarali Gintsburg[1] & Adil Moustaoui[2]

[1]Institute of Languages and Cultures of the Mediterranean and the Near East, Madrid
Spanish National Research Council (CSIC)
sarali.gintsburg@cchs.csic.es | ORCID: 0000-0003-2962-9534
[2]Universidad Complutense de Madrid
adilmous@ucm.es | ORCID: 0000-0002-0770-943X

Introduction

The last fifteen years have brought many new developments to Arabic philology – first of all, we finally have access to large databases of linguistic data, consisting of the press, materials provided by video-hosting sites, as well as posts on social networks, both in Standard and dialectal varieties of Arabic. The possibility of working with such materials has sparked interest in linguistic theories of a more general character, and those based on other languages and cultures. Thus, specialists in Arabic language and, in our case, Moroccan Arabic, have had the unprecedented opportunity to go from descriptive linguistics to more universal trends and to see whether it would be possible to fit their Arabic-language based conclusions and observations into the big picture.

The Maghreb countries in general, and Morocco in particular, have traditionally been a point of attraction for Arabists, specialising in sociolinguistics. This is first and foremost due to its multilingualism: Morocco alone is home to a range of varieties of Moroccan Arabic, Standard Arabic, as well as three varieties of Amazigh (Tarifit, Tashelhit, and Tamazight). In addition, French and Spanish are also traditionally widely used, and,

recently, English has started to gain popularity. While not so long ago sociolinguists working with Morocco were to conduct fieldwork, as well as design surveys and questionnaires, today they are increasingly turning to the study of sociolinguistic trends using digital/online material.

Therefore, it is not surprising that last year, in 2024, we dedicated the 8th edition of the International Conference on Moroccan Arabic to the digitalisation of Morocco and to new trends in the study of the country's linguistic situation. This conference took place at the Institute for Culture and Society (the University of Navarra) under the title *Moroccan Arabic: New generations and new practices in the digital era*, and this special issue is the fruit of that conference. The special issue focuses on advances, approaches, and emerging trends in the study of Moroccan Arabic, its writing, standardisation, and use in the digital era and its spaces. This special issue therefore seeks to address the following questions: 1. What new approaches could we use to better situate Moroccan Arabic in its historical and contemporary contexts?; 2. How is the use and status of Moroccan Arabic changing in both traditional and digital environments, and what ideologies are involved?; 3. In the digital environment, what would be the interplay of different communicative needs and linguistic affordances?; 4. How does the digital environment transform linguistic and discursive practices?; and, finally, 5. What ethnolinguistic, political, and social subjectivities emerge in the new digital communication in Moroccan Arabic?

To tackle these questions, the articles included in this special issue use a wide array of approaches and combine theories and methodologies from applied linguistics, sociolinguistics, linguistic engineering, and computer-mediated communication.

The issue opens with an article by Sarali Gintsburg and Mike Baynham, titled "A thousand years of translanguaging in the multilingual Maghreb". In their study, the authors use the theory of translanguaging to build a bridge between the linguistic situation in the Muslim Andalusia and in the contemporary Maghreb. The authors argue that the peripheral character of the Maghreb dialects, as well as their linguistic permeability, allows these dialects to easily accommodate different varieties of linguistic switching. This allows the authors to draw a parallel to switching in Andalusian Arabic, which was characteristic of poetry and, in some respects, daily communication in al-Andalus. To this end, parallels are drawn between the Andalusian *kharjas* and *zajals* on the one hand and data readily available online – an episode from the comic show *al-Kamira lakum* by the Moroccan actress Hanane el-Fadhili and the song "Partir loin" by the Algerian singer Reda Talyani on the other.

The next article in this special issue is "A Darija and the global multilingual digital landscape" by Mena Lafkioui. In her study, the author offers readers a general sociolinguistic analysis of the Moroccan linguistic digital landscape and comes to the conclusion that Moroccan Arabic is strengthening its position and the author even applies a new term for her observation – *darijation*. We can compare this with the rather well-studied phenomenon of secondary orality coined by Walter Ong, with the interesting nuance that in the case of Morocco, while Moroccan Arabic (that is, traditionally perceived as oral variety) is playing an increasingly important role in written communication, the positions of its rivals in the digital field, primarily French and literary Arabic, are weakening.

The third article in the collection is produced by a group of three authors – Anass Sedrati, Reda Benkhadra, and Mounir Afifi – and represents a kind of technical report

compiled by the authors who were behind the birth and development of the Wikipedia in Moroccan Arabic. In this report, titled "Standardising Darija: Collaborative approaches in the Moroccan Darija Wikipedia", the authors share their experiences and also draw attention to the technical difficulties they inevitably have to face. This report will undoubtedly be of interest to linguists and experts in Moroccan Arabic: in addition to presenting a history of Moroccan Arabic Wikipedia, it outlines a platform-specific proposed orthographic standard for Moroccan Arabic and compares it to actual user practice. Viewed as such, it provides a clear and thorough – indeed, rather lengthy – account.

The next paper in the issue, "Digital age and polycentricity in Moroccan Arabic: Complexity and heterogeneity in linguistic practices" by Adil Moustaoui addresses the topic of polycentricity in the context of Moroccan Arabic. The author explores how digitalization of the Moroccan society is central to conceptualizing the concept of polycentricity coined by the late Jan Bloemmart (2005). The analysis provided by the author is based on a language corpus taken from various social media and digital platforms that includes influencer profiles as a highly revealing manifestation of polycentricity in Moroccan Arabic. Moustaoui concludes that Moroccan Arabic is not only essentially polycentric but that its polycentricity is highly complex and therefore there exist no guidelines that would determine standards or norms of its use.

Rosa Pennisi's article entitled "(Moroccan) mixed Arabic in digital media: A comparative analysis of oral and written practices in Moroccan digital platforms and newspaper" continues the conversation about the practices of using Moroccan Arabic online. To run her comparative analysis of oral and written styles, the author uses several online resources – a newspaper, episodes from a talkshow and a podcast. Similar to Lafkioui, Pennisi concludes that Moroccan Arabic is starting to play a more prominent role in online spaces (in this case, in formal media communication).

In her turn, Samera Abdelati offers a rather fresh perspective on the practices and norms of Moroccan Arabic – she chose advertising products as her research material. In her paper entitled "Moroccan Arabic in advertising context: An analysis of oral and written messages", Abdelati explores the extent to which different languages and registers in Morocco are used in contemporary advertising, providing compelling examples of their application. Additionally, the author offers a valuable historical overview of the role these linguistic varieties have played in advertising over time. The study also highlights key strategies employed by advertisers to target specific audiences, as well as the linguistic factors that justify the writing conventions they adopt.

Our special issue concludes with another technical report entitled "*Sawtone*: A universal framework for phonetic similarity and alignment across languages and scripts", in which the developer of *Sawtone*, Omar Kamali, shares his experience in creating integrated framework designed to enable consistent cross-script phonetic alignment and text normalisation aimed at addressing the inherent challenges in processing text across diverse writing systems. Kamali's report contains not only step-by-step description but also a case study on preprocessing Moroccan Arabic data for Large Language Model (LLM) training. The author concludes by emphasising the importance of such frameworks in the context of rapidly growing digital communication.

DOI: 10.14746/linpo.2025.67.1.2

A thousand years of translanguaging in the multilingual Maghreb

Sarali Gintsburg[1] & Mike Baynham[2]

[1]Institute of Languages and Cultures of the Mediterranean and the Near East (ILC), Madrid
Spanish National Research Council (CSIC)
sarali.gintsburg@cchs.csic.es | ORCID 0000-0003-2962-9534
[2]University of Leeds
m.baynham@education.leeds.ac.uk | ORCID 0000-0002-6259-7936

Abstract: In this paper we explore the interaction between Maghrebi Darijas and Romance languages from the perspective of both historical and contemporary evidence. As Caubet 2002, following Lahlou 1991 argues, among contemporary educated populations in urban centres across the Maghreb, bilingual/multilingual interaction is the norm not the exception. Historical evidence tells us this was also the case in 11th century al-Andalus, though it is of course impossible down the centuries to reconstruct the actual interaction. We will however argue that certain surviving texts can provide an indication when analysed in terms of the constraints on conversational codeswitching such as is provided by Aabi. We start from the position that Maghrebi Darija is a special case of linguistic permeability due to its politico-geographic location on the frontier and given its thousand year history of close contact with Romance. To investigate this phenomenon in both its historical and contemporary manifestations we draw on the current construct of translanguaging, an alternative perspective on multilingual interaction to code switching as expounded in Baynham & Lee (2018) and the notion of convivencia as elaborated by Bossong in his study of linguistic conviviality and coexistence in mediaeval Andalusian poetry (Bossong 2010). We then go on to analyze this in two time slices: examining evidence of the productive convivencia/coexistence of romance and dialectal Arabic i) in the kharjas of 11th century al-Andalus as discussed by Bossong and others and ii) in the modern Maghreb music and performance scene (cf. Caubet 2002; Baynham & Gintsburg 2022). We do this here through analysis of a song by the Algerian singer Talyani and a performance of the Moroccan comedian Hanane el-Fadhili, using in both time slices translanguaging and Bossong's notion of convivencia in our analysis. We then conclude by arguing as Heath (2020) does that for effective research into such varieties as Maghrebi Arabic, both currently and historically, it is necessary for cross disciplinary work between researchers in Arabic and its Romance contact languages, in order to fully address its sociolinguistics. We understand this as a form of disciplinary translanguaging to be undertaken in order to establish the dynamics of the convivencia/coexistence of Arabic and Romance elements in this type of data.

Keywords: Maghrebi Arabic, Andalusi Arabic, performance, translanguaging, periphery, linguistic hybridity, linguistic permeability, recognizability, convivencia

1. Introduction

1.1. Situating Maghrebi Arabic in Performance

Maghrebi Arabic and especially Moroccan Arabic has long enjoyed the reputation of being "different" from other varieties of Arabic language. As early as in the 10th century A.D. Arabic spoken in the Maghreb was described by the Syrian geographer al-Muqaddasi as distant and incomprehensible compared to the Arabic spoken in Iraq, Syria, Egypt and Arabia (Zavadovskiy 1962: 7). Five centuries later, Ibn Khaldoun characterized people from the Maghreb as "incapable of mastering the linguistic habit" (Ibn Khaldûn 2015: 616).

Today Maghrebi Arabic continues to enjoy the reputation of being non-conformist" and even "deviant" (cf. Lafkioui, in this issue). So what are these characteristics that make Maghrebi Arabic and, in a broader sense, the local linguistic landscape so special even in the eyes of their Arabic-speaking neighbours?

The most obvious explanation would be its peripheral status together with characteristics of the arrival of Arabic to Maghreb. Arabic was initially brought to North Africa by the natives of the Arabian Peninsula quite early (between 7th-8th centuries). However, this wave of Arabisation was somewhat patchy and affected only some territories (mainly some cities, such as Tunis, Kairouan, Sfax, Fes). A full-scale Arabisation of the Maghreb and, consequently, the Iberian Peninsular, took place already in the 11th century. However the variety of Arabic brought in was not enough to cover local communication needs. Moreover, the linguistic and cultural centre of the Arab world was too far and, consequently, the linguistic substrate began to play an important role in the elaboration of the day to day ways of speaking in Maghreb. In addition, this shift to Arabic caused phonetic and morphological changes. Another important factor to take into consideration is its sociolinguistic situation: Morocco is often described by linguists as a "linguistic palette", a "plurilingual culture". Indeed, in Morocco, the following languages are spoken: Moroccan Arabic (one dialect with numerous sub-dialects), Standard Arabic, Amazigh (and its three dialects – Tachelhit, Tamazight and Tarifit), and also French, Spanish (in the north), and more recently, English. Some researchers even use the term "pentaglossia" to describe the linguistic diversity of this country (cf. Moscoso 2010). It can be said, then, that in the case of the Maghreb (or especially Morocco), the development of the Arabic language has always depended on foreign borrowings – either in the form of substrate lexemes or in the form of borrowed lexemes from neighbouring languages.

This "otherness" of Maghrebi Arabic, based on its peripheral status, was also reflected in local literary tradition. Written literature in dialectal Arabic existed in Morocco from at least the 16th century when Sufi mystic Abderrahman el-Mejdoub produced his famous quatrains in that variety (de Prémare 1985). Majdoub's quatrains very soon became popular not only in Morocco but also in Algeria and Tunisia and are often considered as the basis for many local performance arts. This happened much earlier than in the cultural centre of the Arab world. From the 17th century, el-Mejdoub's written quatrains were circulating not only in Morocco, but also throughout the Maghreb. This is

radically different from the situation in Egypt, where for example, the Egyptian writer Yūsuf al-Širbīnī published his satirical work *Hazz al-kuhūf* (1686), written in Egyptian Arabic, just to show to his readers the impossibility of using Arabic dialect, the Arabic of the streets, in literature.

Like anywhere else, performance arts of Maghreb played and continue playing an important role in shaping local artistic language (Darija). Of special importance is the role of the audience, as any author – real, or anonymous, or collective – must always think of their audience. In the case of peripheral and plurilingual societies of Maghreb and especially Morocco, the audience will be mixed, that is culturally diverse and plurilingual. Since the performer will have in mind such a mixed audience, their audience will play an important role in shaping and transforming already existing performance canons. This is, for example, the case in the contemporary tradition of the Jbala, an ethnic group in the northwest of Morocco and another excellent example of the peripherality of Maghrebi culture. There local poets produce their poetry trying to satisfy mixed audiences by integrating elements that belong to other genres – the *malhoun*, *chaâbi*, but also *rai* and *charki* (Gintsburg 2020). As Lahlou points out, multilingual language production, both everyday and artistic, is the norm rather than the exception in this context:

> Code-switching is their "default mode" of conversation, a mode which is in the middle of their linguistic continuum, with Moroccan Arabic at one end of the continuum and French at the other. ... It is when they do NOT code-switch that the question as to why should be raised, not when they code-switch. (Lahlou 1991: 182).

Such linguistic permeability and generic hybridity are not recent in origin. The discovery of the kharjas with their bilingual texts shows us that about a thousand years ago Andalusi Arabic demonstrated an impressive capacity to incorporate elements of Romance, such as its phonetic system, first and foremost its stress system, that led to changes in the rhythmical structure of Andalusi Arabic, as well as its morphological and syntactic structure (Vicente 2020: 232-234). Importantly, Andalusi Arabic was also known for incorporating lexical elements from the vernacular Romance language[1]. Similarly, contemporary Maghrebi Arabic (first and foremost, Moroccan Arabic) is known for its exceptional ability to absorb elements from contact languages at phonetic, lexical and morphological levels (Heath 2020: 213-224).

1.2. Translanguaging

As will be clear from the words of Lahlou above, the study of bilingual/multilingual interaction in Maghrebi Arabic has so far been addressed through the construct of code-switching (cf. Bossong 2003, Caubet 2002, Aabi 2020 etc). Here we adopt the perspective of translanguaging, which takes a rather different focus. While code-switching as the name suggests focuses on the linguistic code or system, translanguaging is speak-

[1] Another important source of lexical borrowings was Berber language. However, due to scarcity of information on the subject (see, for instance, Ferrando 1997 and Vicente 2020) and space considerations we will focus only on borrowings from Romance

er-oriented and focuses on the resources deployed by the speaker, their repertoire and linguistic creativity. First of all, the notion of translanguaging reflects a dynamic approach to language: as it uses the *languaging* not *language*. The element *trans* also implies language that does not respect linguistic borders and can contains a rather transgressive meaning. As defined by Baynham & Lee:

> Translanguaging is the creative selection and combination of communication modes (verbal, visual, gestural and embodied) available in a speaker's repertoire. Translanguaging practices are locally occasioned, thus influenced and shaped by context but also by the affordances of particular communication modes or combinations thereof in context. Translanguaging practices are typically language from below and are liable to be seen as infringing purist monolingual or regulated bilingual language ideologies and hence can be understood as implicitly speaking back to those ideologies (Baynham & Lee 2019: 24-25).

As such translanguaging is a contribution to the theory of the speaker not the linguistic code or system. For this reason, in our view code-switching exists in a relationship of theoretical complementarity with translanguaging and we will throughout this chapter draw on the insights of the code switching literature, while re-interpreting them where necessary through the translanguaging lens with its focus on the speaker and their repertoire. Canagarajah for example writes of "the ability of multilingual speakers to shuttle between languages, treating the diverse languages that form their repertoire as an integrated system (Canagarajah 2011: 401). Translanguaging is thus a dynamic, processual, performative alternative to the reification of linguistic codes (Baynham & Lee 2019: 35-36). Translanguaging always involves a selection from available resources in a speaker/writer's repertoire, indeed this is how we can define translanguaging. Translanguaging, playfully or seriously, involves the mixing, and blending of communicative resources and the crossing and transgressing of boundaries. This approach, with the emphasis on the creativity of the individual speaker in performance is particularly suited to the focus on artistic and comedic performance in this paper, though it must be emphasized that as Baynham & Lee (2019) show, the approach is also applicable to understanding the ordinary everyday linguistic creativity of speakers who routinely shuttle between two or more languages. It can be further argued that the impact of the performances we analyze lies precisely in their connection to and play with everyday language behaviour. The audience responds because they see themselves and others in the verbal play of the performer. We would therefore argue that the language use we examine is not just an artefact of the performance itself, but draws on recognizable sociolinguistic norms and usage. Translanguaging can also be seen as an enactment in language of the convivencia [2] that Bossong identifies in the multilingual environment of 11th century al-Andalus and as we and others do in the contemporary Maghreb.

[2] As Aabi points out however (personal communication), the notion of convivencia has to be understood in the context of the marked inequalities that can be reconstructed in Andalusi society. Indeed our analysis below, which proposes an interaction between a fluent Arabic speaking male poet and a Romance speaking girl not fluent in Arabic confirms this. Convivencia is co-existence that does not imply equality of all parties.

2. Approach

It might seem rather eccentric to want to compare and contrast languaging practices with a thousand years in between them, but here we note Ferrando's paper "On some parallels between Andalusi and Maghrebi Arabic" (1998) in which he does just that. In our case we are interested in the incidence of switching/translanguaging in artful texts from the Andalusi and contemporary Maghreb time/space. In his later work Ferrando reviews the conflicting theories concerning the sociolinguistic situation in al-Andalus proposing that no single one of them fits all circumstances and different contexts, periods, regions and social classes which would throw up different language ecologies:

> One theory claims that Arabic quickly and completely replaced Latin and Romance, and a second postulates that the use of Arabic was restricted, and Romance continued to be used on a large scale. A third theory points to the coexistence of the two languages without real bilingualism. Finally, a fourth theory maintains widespread Arabic/Romance bilingualism everywhere in the country. Nevertheless, rather than such global approaches, it seems necessary to distinguish between the very different periods, regions and social classes which are the basis and context of the claims formulated, as it is evident that situations of bilingualism in al-Andalus clearly differed from each other according to the time, place and society concerned. (Ferrando 2000: 45)

Following Ferrando we would not claim of course that we can generalize across all the dimensions of al-Andalus any more than we would in the contemporary Maghreb. Transposing Lahlou's (1991) argument that among contemporary educated populations in urban centres across the Maghreb, bilingual/multilingual interaction is the norm not the exception, we can note historical evidence that tells us this was also the case in 11th century al-Andalus, though it is of course hard down the centuries to reconstruct the interaction or the variety of contexts in which interaction occurred. Bossong and others explicitly warn against reading the sociolinguistic environment from the poetic text:

> *Evidentemente, debemos evitar el error de tomar estos poemas como testimonios sociolinguísticos directos*.
> Clearly, we should avoid the mistake of taking these poems as direct sociolinguistic evidence. (Bossong 2010: 296)

Yet in our view it is possible to undertake a certain amount of reconstruction of the sociolinguistic environment, while taking into account the essential dimension of aesthetic transformation. It is incontrovertible that the kharjas and certain parts of Ibn Quzman's poetry seem to provide evidence that they were produced in sociolinguistic environments where bilingualism was a feature:

> *...no sería justificado tomar estos textos como documentos primarios; la lengua hablada está transformada al haber sido integrada, injertada al texto literario. Pero lo que se nos ha conservado no puede ser totalmente ajeno a la vida lingüística cotidiana, no es completamente*

> *estilizado; algo deben reflejar estos poemas de lo que realmente se hablaba en las calles de Granada, Sevilla y Córdoba en los siglos XI y XII.*
> ...it would not be justifiable to take these texts as primary documents; the spoken language has been transformed by being integrated and inserted into the literary text. But what has been preserved cannot be totally alien to everyday linguistic life; it is not completely stylised. These poems must reflect something of what was actually spoken on the streets of Granada, Seville and Cordoba in the 11th and 12th centuries. (Bossong 2010: 282)

So how far does the switching in the kharjas conform to what is known about conversational switching? In the absence of access to the spoken varieties of 11th century al-Andalus, we draw on the work of Aabi, who synthesizes previous work on the syntax of code switching, to formulate a functional constraint which he argues holds both within and across languages. Working within a version of principles and parameters theory[3], he posits a very general principle that selectional properties (i.e. what elements can combine in an utterance) must be met in code switching and monolingual constructions alike. If properties are parametric (i.e., cannot be satisfied from the other language), code switching will be blocked. Here we use this approach as a test or proxy of the extent that the switching observed in the kharjas and zajals conforms to the expectations of conversational switching. If we observe conformity with Aabi's principle, this is at least an indication that the utterance is oriented to general principles of conversational switching and is not simply an artefact of the poet.

In our analysis of contemporary data we will also be alluding from time to time to insights from Caubet's "*Jeux de langues*: humour and codeswitching in the Maghreb" (2002). In her analysis of codeswitching in comical production from Algeria and Morocco, Caubet distinguishes six major uses of codeswitching to create comical/humorous effect: phonological games, calque translations, playing with the separate elements of idiomatic expressions by using their literal and not figurative meaning, playing with different meanings of separate lexemes, making *translinguistic* puns, and using elements from foreign languages out of place (p. 234). It is important to remember that due to numerous well documented difficulties associated with reading and deciphering the kharjas and zajals, it is impossible to identify some of the translanguaging artistic devices described by Caubet. This is the case, for instance, with translinguistic puns: words written in a "wrong" way might be nothing more than the poet's attempt to create a comical effect by imitating someone's pronunciation, changing morphological structure by mixing together Arabic and Romance elements, or playing on the similarity of sounds. An example of this might be the phrase *mio habibi*, discussed below, hard to explain in terms of Aabi's constraints on switching since the first person is marked twice. Given that the kharja composer is by definition fluent in Arabic, we suggest that this might be evidence of a comic treatment of a Romance speaker who is not fluent in Arabic, but has picked up and uses the pervasive formula *habibi*.

[3] Simply stated, in this approach principles refer to characteristics shared by all languages, parameters are features that make them different. The feature PRO DROP is a parameter: Spanish and Arabic permit subject pronoun dropping, French and English do not.

3. Analysis

3.1. Historical data

In our discussion we will present data from the kharjas and Ibn Quzman's zajals. In both cases whatever we can reconstruct as corresponding to everyday language usage is shaped aesthetically through incorporation into the verse form. In our discussion of the kharjas and the zajal poetry of Ibn Quzman, we will discuss in more detail the connection that can be identified between translanguaging data and everyday language use. In the discussion of extracts from Ibn Quzman the focus is more on its aesthetic incorporation. This provides a transition to our consideration of contemporary data such as Talyani's song.

3.1.1. Translanguaging in the kharjas

The kharjas, that were discovered in the 1940s, caused a lively controversy, during the following decades about the origin of lyrical romance and the opposing Romanist theories with those of the Arabists, which should not concern us here[4]. To develop our arguments we cite without prejudice both Arabist and Romance authorities. With less controversy but in a no less interesting way from the sociolinguistic point of view, the kharjas offer us an insight into the texture of this Andalusian Arabic of yesteryear and the Romance vernacular of the time, a theme addressed by Bossong (2003, 2010). Given the paleographic difficulty of arriving at precise versions of the original manuscripts without diacritics, which have exercised scholars over the years, in our analysis we will use versions of the kharjas from Corriente's "The kharjas: An updated survey of theories, texts, and their interpretation" (2009), as well as evidence from the zajals from the Diwan of Ibn Quzman edited and translated by James Monroe (2017). As discussed above, in order to test the correspondence between this artful language use and the everyday, albeit indirectly, we will check the instances of translanguaging against the constraints on switching identified by Aabi (2020).

3.1.2. Situating the kharjas textually

As a generically hybrid form, the muwaššaḥs as is well known are composed of a central section of stanzas in fuṣḥa with a kharja or 'envoi', defined by Corriente as follows:

> *kharja* (pl. *kharjāt*), is literally on 'outing', that is a technical term in Arabic literature, that is, synchronically speaking, the last *qufl* of a *muwaššaḥ* (pl. *muwaššaḥāt*), namely, a kind of stanzaic poem in which the stanzas begin with segments (*simṭ*, pl. *asmāṭ*) in a rhyming sequence that is different for each stanza, and end with segments (*qufl*, pl. *aqfāl*) sharing the same rhyming sequence in ever. (Corriente 2009: 110)

[4] Although we fully realise the importance of the rhyme pattern of kharjas and zajals, this will not be of our concern in this paper.

The muwaššah also sometimes has a preface – *maṭla'*, sometimes defined as refrain, that introduces the rhyme. The *muwaššaḥ* with its *maṭla'* (optional) and *kharja* (obligatory) constitute a characteristic mix of Standard Arabic, vernacular Arabic and the vernacular Romance language. In the sociolinguistic whole of al-Andalus of the 11th or 12th century, for example, one can therefore postulate a dialectal variety of Arabic (Andalusi Arabic) that is deeply influenced by the Romance and the vernacular language that in turn are profoundly influenced by its counterpart (Andalusi Arabic), with a whole set of distinctive varieties and idiolects.

3.1.3. Translanguaging in kharjas as a mirror of everyday speech

Of course, as already noted, one must be careful of simplistic assumption that there is a necessary correspondence between everyday street bilingualism and its representation in different artistic forms. Fortunately, our point of comparison is not Andalusi Arabic vs Darija as it is spoken on the streets and in the markets, or in Cordoba in the eleventh century, or indeed in Tangiers in the twenty-first century, but in the texts of popular literature, abundant in the digital spaces of contemporary Maghreb: a song by Reda Talyani (Algeria), a comic show by Hanane el-Fadhili (Morocco). Of course, there are stylistic exaggerations that make up their artfulness. One must of course work on the assumption that their connection to the language spoken on the streets of eleventh-century Cordoba as in the port at Tangiers today would have been more indirect. But we reject the notion that there was a strong division between the language that was spoken on the streets and the literary language, even including the literary representations of the so called "vulgar" languages. The literary effect produced by the kharjas on the one hand and the Talyani's song, as well as the comic sketch by Hanane el-Fadhili (Morocco), must depend in some measure on the echoes which will be recognized by its listeners from the language of the street. Through the centuries we recognize that voice that says[5]:

> *Gare*[6] *sos devina y devinas* ***bi-l haqq****, garme cuand me vernad mio* ***habibi ishaq***
> Since you are a fortune teller and your predictions are true, tell me when my friend Ishaq will come to me.[7]

This kharja obviously belongs to the Romance kharja since the matrix sentence is Romance with Arabic insertions. Its entire grammatical structure is in the vernacular

[5] In this article, in order to illustrate examples of translanguaging, we used bold italic to highlight instances of using the language different from the main language of the text. For instance, if the main language is Romance, then Andalusi Arabic will be in bold, if the main language is French, then Moroccan /Algerian Arabic will be in bold. If Moroccan Arabic is the main language, then instances of using French will be in bold. The same is applied to mixing different registers of Arabic in one text. Finally, Moroccan/Algerian Arabic-influenced pronunciation of French is also given in bold italic.

[6] In this article, we decided to keep the original transcription used in the sources we cite.

[7] Although this kharja by Yehuda Halevi belongs to the Jewish branch of Andalusian poetry, our position is in line with the following view, expressed by Samuel Stern: "Hebrew poets, when they wrote muwasššaḥs, doubtless merely imitated their Arabic models now lost" (1974: 129).

Romance style. Embedded in it, like jewels, are two Arabic formulae, the first the simpler *bi-l-haqq*, the second more complex *mio habibi Ishaq*. Why do we include the Romance mio in the sentence? Simply because it is connected to it by a grammatical logic that is difficult to dissolve. But we can clearly see that the possessive adjective is already marked in the Arabic first person morpheme *-i*, which suggests that *habibi* is an unanalysed formula.

In terms of Aabi's constraints on switching, if we treat *habibi* as a formula, both observe the constraints on switching identified by Aabi. An intriguing possibility, mentioned above, is that the poet is animating the voice of the speaker addressing the fortune teller as a non fluent Romance speaker of Arabic, who uses the formula *habibi* combined with *mio* in a way that flouts the selectional constraints of fluent switching. Could it be, as we suggested above, that the poet wanted to mimic the switches from Romance to Andalusi Arabic and back and amuse his audience? Such similar examples suggest to us that the linguistic choices of the kharjas, without conforming in every detail to the spoken language of al-Andalus, would offer a mirror, perhaps artistically heightened or exaggerated, which would resonate with its listeners and it is exactly in this resonance that the artistic effect would reside. Now let us take a closer look at several kharjas as they were read and decoded by Corriente:

> Al. *Vén* ***sídi abráhim****, / **ya+ndá min thálj**e, / vént+ a(d)míb de nókhte, // o nón, śi non kéres, / virém+ a(d)tíb, / garré(d)me ób liqárte*
> Come, my lord Ibrahim, oh you who are fresher than the snow, come to me at night or else, if you do not want to, I shall come to you; tell me where I shall find you (Corriente 2009: 120)

This kharja like the others below, contains Arabic phrases that on the one hand go beyond the formula (***ya+ndá min thálj**e*), on the other towards complete integration within the verb phrase through a creative fusing of the Arabic root (***liqá***) in a romance structure (***liqár**te*). Again these structures correspond to data analysed by Aabi, for example in the switch between the Romance imperative verb vén and the Arabic vocative, ***sídi abráhim***

> A6. ***assaʕáma min kháli / múdhi ḥáli*** *qerbáre: // ké faréyo,* ***yámmi****, / ya non pódo lebáre!*
> My darling's ennui hurts me to the point of shattering me: what shall I do, mother? I cannot bear it any longer. (Corriente 2009: 121)

Here the kharja begins in Arabic and ends in Romance. There is a balance between the phraseology of Arabic origin and that of Romance origin, both are seemingly evidence of a creative and unformulated use, although ***yámmi*** seems formulaic to us. Once again "ké faréyo, yámmi" observes Aabi's constraints on switching.

> A9. *Non temptaréy* ***illá*** *kon+* ***ashshárṭi // an tijammáʕ khalkháli maʕ qúrṭi***
> I shall not even try it unless you [make love to me and] raise my anklets up to my earrings. (Corriente 2009: 120)

In this example there is also a certain balance between Romance/Arabic phraseology, and this kharja that starts in Romance and then shifts to Arabic. The combination of (***illá***

kon+ ***ashshárṭi***) shows a structure that is typically Arabic (la… illa). Here the switching is interestingly divergent from Aabi's approach as he would discount a switching across discontinuous negatives (Non…. Illá), which would not be predicted in his model.

A13. *Non kéro bóno* ***ḥall****éllo // **illá assamré**llo*
I want no handsome little thief [of hearts] but the little dark-skinned one. (Corriente 2009: 121)

Just like A9, kharja A13 is also built using the *Non… illá* structure we just discussed. Further, we also see the interesting combination of Arabic roots with the Romance diminutive, that is linguistically intricate, and which in our opinion demonstrates the creativity of a poet, plausibly plundering sentences from the street and reanimating them in his poem to dazzle his listeners. Again the evidence as to contemporary sociolinguistic use is indirect, but Farida Abu Haidar in her study of the use of diminutives in Ibn Quzman's zajals (1989), concludes that in the 10th-12th centuries the diminutives were a distinctive feature of Romance and were obviously a part of Mozarabic speech. While in Eastern Arabic poetry, according to Abu Haidar, the use of diminutives was scarce, and were typically used pejoratively, this was not the case in Andalusian poetry, where, just like in Romance, diminutive suffix gives the noun an affectionate hue.

Today extensive use of diminutive forms is one of the highlights of the Arabic dialects spoken in Maghreb, something which exists but is considerably less current in the Eastern Arabic dialects. We can therefore assume that the use of the Romance diminutive suffix *ello*/*ella*, which we find in both kharjas and zajals, reflects the norm of the everyday speech of that epoch, which then passed into Maghebi dialects, where Arabic and sometimes Berber suffixes are used to give the affectionate hue to the word. It can again be pointed out here that this switching within the word is in accordance with constraints on switching identified by Aabi.

A24. ***qúltu ásḥ taḥyíni****, bokélla /* ***ḥúlwa mithl+ ásh!***
I said: how exciting you are for me, little mouth, what! (Corriente 2009: 122-123)

This example is entirely in Arabic except for one word – *bokélla* – 'little mouth', which regularly appears in the kharjas in its diminutive form (A11, A14, A20, A25, etc.) and seems to be charged with a specific meaning associated with romantic poetry and therefore can be classified as a formula. Our observation can also be supported by this example from Ibn Quzman (zajal 67), where Romance lexemes are used to increase the romantic context of the poem. Another interesting characteristic of this example is that the poet here also used Arabic lexeme *ḥajal* (partridge) in its diminutive form, thus echoing Abu-Haidar's observations:

ya**d̲**d̲a** collo d̲e **l-ġazālah
'i bukillah d̲e ***ḥujaylah***
The one with a neck like a gazelle
And a mouth like a little partridge (Monroe 2017: 398-399)[8]

[8] In Ibn Quzman's *Diwan* the Romance *collo* and *bukillah* are used only once, in other instances the Arabic lexemes ʿunq and fumaymah are preferred.

Again the switching here seems to observe the constraints on conversational switching identified by Aabi.

A38. *Mamm+ eśt+* ***alghulám // la bud kullu líyya, / halál aw ḥárám***
Mother, this boy has to be mine alone, lawfully or unlawfully (Corriente 2009: 124)

This kharja starts with the Romance *Mamm+eśt* and shifts to Arabic. As we demonstrated in A6, Romance *mámma*/*mamm* is interchangeable with Arabic *yimma*, so we can suggest the decision to start the line in Romance in this case is deliberate and can be explained as poetic creativity perhaps again animating a voice that habitually blends Romance and non fluent Arabic. Consider however the following examples of using Romance demonstrative pronoun *eśte*/*eśt* followed by an Arabic noun with definite article: *eśt* ***arraqí*** (A10), *eśt* ***alḥarakí*** (A10), *eśt+* ***az- zaméne*** (H1), *eśte* ***alkhalláq*** (H6), *eśt* ***alḥabíb*** (H15). In contrast, in the only instance we found when eśt is followed by the word in Romance, no article is used: *eśta díya* (A22). Of interest here is that these forms all incorporate the double subject as identified by Aabi, characteristic of Arabic, but not of Romance.

A40. *Ké faréyo o ké+n śerád de míbe, //* ***ḥabíbi*** */ non te+ mṭálya de míbe*
What shall I do or what will become of me, my darling? Do not break up with me! (Corriente 2009: 124)

Here we see the case identical to A24 with the difference that the text of the kharja is entirely in Romance with the exception of one word in Arabic, the formulaic ***ḥabíbi***. Just as I the case of A24, in A40 one foreign lexeme is inserted into the text decorating it and highlighting/accentuating its romantic essence. Again the switching at the vocative corresponds with the constraints on switching identified by Aabi.

In conclusion we can say that, since almost all the instances of switching we have identified in the kharjas quoted observe Aabi's constraints on conversational switching, it is plausible to argue that this is not simply a question of literary invention, but a conscious echoing for artistic purposes of the translanguaging characteristic in the speech of the time.

3.1.4. Translanguaging as aesthetic device in the *zajals* of Ibn Quzman

Before we start analyzing our contemporary data, let us further discuss translanguaging for aesthetic purposes. In order to do it, we turn to zajals from the famous *Diwan* of Ibn Quzman, the poet, who, according to Corriente, was bilingual, although his first and main language was Arabic (2008: 82), and look at several instances of shifting from Andalusi Arabic to Romance. The language of the *Diwan* is predominantly Arabic, but it also contains various instances of using Romance vocabulary. These uses are mimicking and reproducing the language of Christians and slaves, as well as bilingualism of Andalusi women (Corriente 2008: 81). This observation further supports our analysis of "mio habibi" and the analysis of diminutives above.

Zajal 84 is also of special interest to us because of the double meaning that also allows for its satirical reading. The text of this zajal classified as panegyric is built based on the principle of ring composition and its main theme is introduced by the refrain *fī ḍamānī ʾin tuʿṭà ʾal-ẖiyār lam tarā mā rayt [yaḏḏa] min al-asfār* (I guarantee that, even if you were given cucumbers,You would not see the travels I have seen), where the lexeme *ẖiyār* is a deliberate pun as it can be interpreted both as 'cucumbers' and 'choice'. If 'choice' implies the serious meaning, where life events are dictated by fate, 'cucumbers', give the poem a comical effect, as the entire idea of travelling, even if imposed by fate, discussed in the poem is then nothing more than an involuntary trip to the bathroom provoked by the laxative effect of cucumbers[9]. This zajal is made of three parts and its main protagonist and narrator is the poet-trickster. In the first part of the poem the poet promises the readers to tell them an exceptionally interesting story from his life but then decides not to do it. In the second part, the poet describes his encounter with his neighbour, a woman, who, as we understand from the context, is also a trickster and who reads his palm and predicts that he will become famous and rich if he goes to a certain Abu-l-Ala. This is the part that contains switching from Arabic to Romance in the interactions between a fluent and non-fluent speaker of Arabic. Finally, in the third part of the zajal the poet decides to travel but fails to do so because the mule he had previously rented turned out to be epileptic. Below are the three strophes, where the poet uses both Arabic and Romance. In the first strophe, the only Romance lexeme used by the poet is the adjective ***ya***.

The next two strophes contain the dialogue between the poet and his neighbour, a fortune-teller, for whom, as we know from Bossong's study from 2010, it is typical to speak Romance or a mix of Romance and Andalusi Arabic (2010: 296-298). Indeed, Ibn Quzman's fortune-teller also uses in her replies Romance with Arabic, which contributes to creating comic effect (zajal 84):

anā ʾay kunt naẓartu mā taʿmal
min fulān sīr w-abṣir fulān muqbal
wa-l-ḥubūb kulli marrah tatbaddal
aš naqul lak ***yā*** *lam naʿud ẖummār*
Wherever I was, I observed what she was up to:
"So-and-so, get lost!—Look, here comes So-and-so!"
For she was constantly changing her lovers.
What can I say? I'll not be made a laughing stock again

qultu lah ba-llah anẓur ṯamm aš yakūn
naẓarat kafff[ayya] wa-qālat lī ***bōn***
fāṭaš albaš *narāk bi-ḥāl al-quṭūn*
*[aww]aḏā l-jāh [*eš de*]* ***nōn akabbār***
I asked her: "By God, look here. What do you see?"
She gazed at [my] palm and said: "Good!
Propitious fairies! I see you [white] as cotton:
Such glory [is] unending."

[9] See Monroe's detailed analysis of this poem, its ring structure and several levels of reading it in Monroe (2017: 1140-1170).

qult aš al-ḥīlah innamā ḏā ġalā
las narà [f]a-d-dunyā [li]-man ḏāb malā
ʾillā law kān mawlā-nā ʾabū l-ʿalā
qālat ***ešte keriya ew nom[n]ār***
I said: "What shall I devise? Prices are high these days;
I see no one in the world who is prosperous now,
Unless it be our lord Abū l-ʿAlāʾ."
She replied: "Just the one I wanted to name!" (Monroe 2017: 516-519)

We have established that there is a comic trope in the al-Andalusi period of interactions between Arab speakers, typically male and Romance speaking women. This gives further support to our argument, suggesting that there is some element of sociolinguistic appropriacy here and in the kharjas in the rendering of mixed speech, the recognizability of which, as we will see in the contemporary data, must inform the comedy.

3.2. Contemporary data

Let us now turn to the *Al-Kamira lakum* (2016), a show produced by Hanane el-Fadhili, a professional Moroccan comedian actress and her brother Adil el-Fadhili, a screenwriter. Each episode of the show which lasts about fifteen minutes, is framed by a title: a proverb or a popular saying. The episode we are going to analyse is titled *Qaṭrān blādi wulla ʕasal əl-buldān*, and is focused on the subject of emigration, a subject of great importance in the Moroccan/Maghreb context. The choice of this proverb introduces another genre into the textual hybridity we are exploring – the proverb is a very important and widespread genre in the popular literature of the Maghreb. This title frames what is to come, much as Ibn al-Mulk characterized the function of the *kharja* in the *muwaššsaḥ*: it expresses in an anticipatory manner the essence of what is to come, its salt, its sugar, its musk. There is a second element that shapes the entire episode – the song by Algerian singer Reda Talyani *Partir loin*. *Partir loin* was released in 2007 and immediately became popular not only in Algeria but also in Morocco and France.

This song becomes central for the *Qatrān blādi wulla ʕasal əl-buldān*: its first, opening part of the episode starts and ends with the first lines from *Partir loin*. This quotation of others' poetic text, indeed, reminds us of Ibn Quzman's quotation of fragments of earlier poetry, sometimes with unidentified authorship, in the form of *kharja*, as described by Monroe (2017: 37, footnote IV). We will therefore start our analysis with the song and then focus on each of the three parts of the episode.

3.2.1. *Partir loin*

This song represents a mix of genres that became typical of the Maghreb music scene in the mid 1990s – here we clearly define a few fragments of the rai – such as the first lines: *Yal babour ya mon amour*/ *Kharejni mel la misère* (*partir loin*), next to some elements of hip-hop, and, through all the songs, the traditional Algerian dance melody. At the genre level, *Partir loin* is a hybrid that consists of various elements and traditions.

To be clear, however, we are not going to draw a straight line of descent between the kharjas and the popular songs of the Maghreb of today. The connection must be more indirect, but analytically we will consider the two influences that we have already noted in the kharjas: the influence of the spoken language and that of the shaping processes of poetic composition. Here is the first verse of the song sung by Talyani:

Yal babour ya [Oh, boat, oh,] *mon amour*
Kharejni mel [take me out of] *la misère (partir loin)*
Fi bladi rani mahgoure [in my country I am lost]
3yit 3yit [I'm tired, tired] *tout j'en ai marre (c'est bon)*
Ma nratich [I won't miss] *l'occasion (on est là)*
Fi bali [It's on my mind] *ça fait longtemps*
Hada nessetni [It made me forget] *qui je suis*
Nkhdem aliha [I work on it] *jour et nuit*
Yal babour ya [Oh, boat, oh,] *mon amour*
Kharejni mel [take me out of] *la misère*

We first see a strange balance between the lines (from the third line), each line starts in Arabic and ends in French[10]. This seems to us to be attributable to the poetic shaping rather than an echo of the spoken language. If we turn to the kharjas we analyzed in the previous section, we will notice that structurally this stanza echoes A38, where we guessed that the choice to start the line with the Romance *Mamm+eśt* and continue it in Arabic was deliberate and not spontaneous. But nowadays there is nothing easier than to explore the link between poetic expressions and spoken language, using corpus linguistics, an option that was obviously not open in the eleventh-century in al-Andalus[11]. Here, then, we see an exemplification of the creative tension between poetic formation and spoken language.

As noted above in the discussion of translanguaging, the key concept in sociolinguistics is the sociolinguistic or communicative repertoire. A recent definition of sociolinguistic repertoire is:

> [people] performing repertoires of identities through linguistic-semiotic resources acquired over the course of their life trajectories through membership or participation in various sociocultural spaces in which their identities are measured against normative centres of practice (Blommaert & Spotti 2017: 171).

What is the repertoire that Talyani's song exhibits? In the first stanza we encountered that linguistic coexistence that goes under the name of translanguaging, while noting the contribution of its poetic formation, where the tropes of parallelism and repertoire are

[10] Bossong (2010) similarly identifies a tendency for translanguaging in the kharjas to start in Romance and end in Arabic and in Ibn Quzman for it to start in Arabic and end in Romance.

[11] While there are of course corpora of contemporary language use, such as Aabi draws on, there can be none for the 11th century data. For this reason we draw on Aabi's principles and parameters approach which expressly formalizes the characteristics of conversational switching. We are not aware that it has been used so far in the kind of historical reconstruction we are attempting here.

important (remember the repetition in the kharjas, for example the words in Arabic with Romance diminutive suffixes, such as:

Non kéro bóno ***ḥall****éllo // **illá assamré**llo* (A13).

In the second stanza we note, as sung by Rim-K, leaving aside certain sentences – ***habsini maalich*** or ***ya hmar*** – very little Arabic but a lot of cultural references on the one hand to the culture of Maghreb or on the other to the European culture (Robinson Crusoe and his sheep), European and global culture in the sense that *Alf Laila wa Laila* is global. The line is used to introduce the identity of the protagonist of this song and it is done with the help of what Caubet termed *translinguistic* pun: he (the protagonist) is from *Kabyle Fornie*, a clearly playful allusion to Californie:

Moi, je suis de Kabyle Fornie
On fumait 350 benji
Sur les bords de la corniche
Habsini maalich [you can arrest me, I don't care]
Rien à perdre, Rim-K le malade mentale
Plus connu que le Haj Mamba, je mens pas
Je voudrais passé le henné à ma bien aimée
Avant que je taille
Comme Cheb Hasni je suis sentimental
Partir loin, rien à perdre
Fih [to] *Boston,* ***wulla*** [or] *je n'sais pas*
Laissez moi de toi
Comme Robinson sur une île
Mon mouton, je l'appellerai Mercredi
Dès que l'avion atterrit j'applaudis
Comme les ***chibanie****s* [old immigrants]*, je vous rends la carte de résidence*
Un moment d'évasion, ***ya hmar*** [you, idiot]*, lève-toi et danse*

The first stanza repeats as a chorus, sung by Talyani, then the song resumes, sung again by Rim-K in a French more or less full of cultural references. The name *blédard* is a derogatory term in origin (derived from the Arabic *bled*, a term dating from the colonial era that was more recently adopted by diasporic youth as of term of identity as seen here:)

Je reste blédard, débrouillard, j't'annonce
Amène moi loin de la misère
Mon plus fidèle compagnon
En route pour l'eldorado
Tellement plein, c'est quoi? ***Dirou*** [take]*, le sac à dos*
Partir loin, sans les cousins
Le plein toujours les carages, c'est dur
Je me considère chanceux d'être en vie
Pourvu que ça dure
J'ai grandis qu'avec des voleurs

J'aurais toujours les youyous qui résonne
Dans ma tête être à la quête du bonheur

A life dominated by the extended family, seeking a life away from the cousins (leaving far, without the cousins) a life of fantasy where he is brought away from the misery, the backpack.

This largely French stanza is followed by a short stanza that is entirely in Algerian Darija:

Yal bledi nti fik el khir
Yediha elli andou zhar
Y3ich li 3Andou lktef
watzidilou mel lebhar
My country, you have treasures in you,
But only those who have luck will have them
Those who have connections, live well,
And you help the rich to get even wealthier[12]

The song *Partir Loin* shows our theme of linguistic permeability that is generic: shared between Talyani, who sings the refrain in the style of rai and Rim-K, who sings the rap, here one finds a generic hybridity that indexes the cultural practices that are at the same time local/global. This is precisely the phenomenon of Jbala (Gintsburg 2020). Obviously there is also the practice of translanguaging: the song's composition and performance is a translinguistic practice.

This practice plays out in an interesting way, when it comes to transcribing this song. It is assumed that this is not a composed song, written and then sung. There is a lot of improvisation there. So the "lyrics" found on the internet are at times very variable, with this lack of some of the variants favouring a French version and others an Arabic version:

Version 1: *C'est Boston, nous lâché pas*
Version 2: ***Fih*** [to] *Boston,* ***wulla*** [or] *je n'sais pas*

The first version erases the Arabic element of the sentence which is revealed in the second version as being translinguistic. This is possibly the result of an automatic translation, but we also noted these dilemmas in our manual translation of Hanane's video. Translinguistic practices seem to produce a certain significant fluidity. The lack of norms in the lyrics is saying something, especially when compared to the counter-normativity that was noted in the discussion of translanguaging. These lyrics are aimed at young people who consider themselves to be "outside the norm" or on the margins of the norm. They do not adhere to transcription standards, nor those of writing in the dominant language (French or Arabic).

[12] In this article, only texts in Arabic language are accompanied by their translation into English. Texts in French are given without English translation. Although in our English translation we tried to be as close to the Arabic original as possible, we decided not to resort to literal translation.

3.2.2. The *beau gosse*

Turning to Hanane's monologue, the first part represents the direct continuation with the song of Reda Talyani, indeed this part represents a development or a variation of the subject of Partir loin. The scene opens with a young man singing the first lines of this song and telling his story. This character, who introduces himself as "handsome" is yet another incarnation of the picaresque hero – his dream is to emigrate to Italy, no matter what, even if it means marrying an old woman, or, as our hero puts it, an expired woman (pirīmi). But why Italy? –because he is tired of his life in Morocco, because he feels very connected to Italian culture through some films with de Niro and Al Pachino and because there are many Moroccans in Italy[13].

> *w kēyn wāḥəd əl-fiʔa waḥəd eṭ-ṭabaqa dyāl š-šabāb ktār bḥāli dīprīsyō w-ṭfīq mʕa ṣ-ṣabāḥ tsanna n-nhār yəṭīr bāš bāš yəji l-līl w-l-līl ja tsannā yəṭīr bāš yəji n-nhār… **ainsi de suite et ainsi de suite**… rā f škəl. ana l-ḥulm dyāli huwa nəmši l-Ṭalyā ā nəmši l-Ṭalyā.. bəzzāf də-nnās ki-ygūlu ʕalāš Ṭalyā ʕalāš Ṭalyā..Ṭalyā fī ṣ-ṣarf a ṣaḥbi w-zād waḥəd lli ka-yəhbaṭ f **tirīṭwār** dyāl Ṭalyān ka-təlqa wlād əl-blād fhəmti ka-yəməddū lək īd əl-musāʕada..*
> There is a group, a numerous class of young men like me, who are depressed: they wake up in the morning and [can't] wait for the day to fly away, so, so that the night could come. When the night comes, they wait for the day to come … ainsi de suite et ainsi de suite … This is hard! My dream is to go to Italy, yep, to go to Italy.. lots of people tell me: Why Italy? Why Italy? Italy has money, my friends, plus if you 'fall' on the Italian territory, you will meet people from your country, you know what I mean, and then they will help you … (Baynham & Gintsburg 2022: 163).

However the handsome young man's adventures end where they started – in Morocco but we see that he is already preparing his next trick. Linguistic analysis reveals that this text is a typical example of translanguaging. There are regular shifts from Arabic to French and vice versa, something that is typical as we have seen of everyday discourse in urban areas of the Maghreb. Insertions in French can serve as examples of what Caubet described as *phonological games* – they are pronounced with an exaggerated Moroccan accent, so that, for instance, the French ***territoire*** becomes ***tirīṭwār*** – a foreign land that is, however, full of Moroccans who are going to help out.

There are, in addition, more multilingual incorporations, cleverly inserted fragments in Italian – *prōnto la māma* – or in English when the character writes his message to Barbara towards the end of his story. Here Hanane resorts to quotational switching, the linguistic device that was also used by Yehuda Halevi in the kharja we discussed in the beginning of our analysis and by Ibn Quzman in zajal 84 – to imitate the direct speech, with the difference that Halevi and Ibn Quzman used Romance and Hanane produces a cocktail of French and English, accompanied by its translation into Darija along with commentaries:

> *Chère and Barbara how are you?*
> ***kī dāyra ma thəmmnī š ġīr awrāqi*** (I don't care about you, I only care about my papers)

[13] For a more profound discussion of the episode, please see our earlier work Tar or honey? Space and time of Moroccan migration in a video sketch comedy 'l-kāmīra la-kum' (2022).

I am very well
ana ṣāfā w-ṣāfi ma bġītš nbəyyen lha rāni ṣāfi gāʕ nkərha
(I'm fine. And that's all, I didn't want to show her I hated her)
But my situation very difficult
l-waḍʕīya fə š-škəl
(This situation is a problem)
no gārrō..wālū… ma kēyn la gārrō w-la wālū
(I don't [even] have cigarettes… no cigarettes, no nothing)
But my God looks my brave heart Barbara
zaʕma rabbi rā ʕāləm qalbi š-šujjāʕ
(that is God knows that I have a brave heart)
a Barbara

3.2.3. The French mother

We saw in the first part of the halqa the desire to leave/emigrate as a fundamental aspect of the migratory narrative. In the second part we see other dimensions: that of living *là-ba*s and the return. In addition, we also see how living abroad can change one. In this part Hanane turns into a mother with eight children, of Moroccan origin living in France. This part is not really structured as a story of emigration, here the character is giving an opinion, maybe a little naïvely on the subject of emigration, on preserving Moroccan culture abroad, making sure that her eight children, who live under the constant threat of the American culture of MacDonald's, rap and hip hop (50 cents, Eminem), will remain Moroccans. The mother therefore encourages them to stick with the tajine and not the hamburger, to listen to traditional, even slightly outdated Moroccan musicians, such Snaji, Daoudi and Jedwane. As if she was addressing the camera and her future audience, she ends up with some kind of generalized advice for future immigrants.

Most of this monologue is in French, but this variety of French is explicitly influenced by Arabic at all levels – phonetics, morphology and grammar. This, we conclude, is in itself a translinguistic influence. In conformity with Aabi's constraints, the first phrase starts with the Arabic preposition ***ʕind***, an equivalent of the European auxiliary verb that transmits the meaning of possession (English have, French avoir, etc.). This beginning conditions the rest of the phrase, which, although it is mostly made of French words, is essentially Arabic in its structure – the mother differentiates between her female and male children by producing the female form *enfante*, although in French *enfant* is an gender neutral noun. In contrary, in Arabic language (in this case, in Moroccan Arabic) the difference between a girl (*bent*) and a boy (weld) is always emphasized in a conversation. The mother then goes on and explains *that* her children receive a 'very progressive' education (*éducation developpée*), another calque from Arabic (i.e., according to Caubet, switching used to create a comical effect) that prepares the audience to the idea that everything that will be said about the education after that won't have much sense. Indeed, we soon learn that this education makes children open up towards the outer world and, continues the French mother, at home her children are not allowed to speak French:

> ***ʕindi*** *(I have) huit enfants* ***u-ʕindi*** *(and I have) sept enfantes, garçons et filles. l'éducation qu'on a donné à nos enfants* ***l-ḥamdilla*** *(all praise be to God), l'éducation très développée – ouverture le monde, les enfants défendés, j'ai dis défendés parler français à la maison – 'parle français avec la* ***mitrēz*** *parle français avec tes copains, la maison tu parles arabe! Voilà* ***répondī*** *en arabe, c'est tout hein? Moi défendé les enfants: écoute music de 50 cent ou écoute la music* ***dī*** *Eminem* ***wulla*** *(or) écoute la music* ***dī*** *113'... oui défendé' J'ai dis: écoute moi très bien, t'écoute SNaji, t'écoute Daoudi, t'écoute Jedwane, écoute Najat Aâtabou, écoute la musique* ***marocān****, c'est tout'.*

We notice the same cross-linguistic combination, already noted in the kharja and the song of Talyani (***ʕindi*** *8 enfants* ***u-ʕindi*** *7 enfantes, garçons et filles)*, with the Arabic formulas (***lḥamdilla***). Her speech similarly positions her, in the way we have suggested is evident in the verse of Ibn Quzman and the kharjas.

The monologue of the French mother reminds us of the period of more organized migration, when one left as a migrant, regulated by inter-governmental agreements, not as ḥarraga without papers. This is then contrasted with the current migration chaos:

> *maintenant quand tu vois* ***š-šubbān*** *(these young folks), tu vois les gens hein?* ***Rīskē, rīskē*** *ça* ***vait*** *comme ça, dans la mer et la plage, c'est pas bien, il faut pas faire ça, attendez mariage, attendez quelque chose, de.. de bien hein? C'est pas, c'est pas comme ça, les gens n'est partent… Moi je suis pas d'accord, je pas du tout d'accord* (Baynham & Gintsburg 2022: 166).

She then tells a little story to support her opinion. A young Moroccan girl wants to run away her country, and our mother of eight tries to dissuade her of doing this illegally. However, the young girl still decides to do that but has to make the trip to Paris in a washing machine, perhaps alluding to the dangers of the boat crossing. Confronted by the customs officer, she just explains that she arrived in that washing machine. As the mother end the story, she repeats her disapproval ("I was really angry when I was told the story"):

> *La pauvre, elle a foutue dans le car, comme le car était bien remplis, elle a mise dans une machine à laver.. les gens, c'est même pas que j'raconte la misère, la misère: quelqu'un appuie sur le bouton* ***da*** *marche et la fille məskīna (the poor thing) tournait, tournait, passait par l'essorage, par le lavage, par le rinçage et elle été arrivée à la douane française məskīna tout essorée, le douanier demande: 'Madame d'ou vous* ***sortē****?' et elle luis dit 'monsieur dans une machine à laver' et il lui dit retournisez elle n'as pas bien* ***répondē****.* ***Ma ʕarfat š tjāwbu ma.. u-kūn kunt ana kunt nqūl lu*** *(*she didn't know how to reply properly [to him], but if it was me I would have said to him): *monsieur retournisez vous, toi-même* ***tī****, toi pour que tu t'apprends! J'étais vraiment énervée quand on m'a raconté l'histoire* (Baynham & Gintsburg 2022: 166-167).

Just as the *beau gosse* glosses his letter to Barbara with his thoughts and comments in Arabic, here the mother glosses her story with her comment in Arabic on what the girl should have done. We note how the transition between the thought in Arabic and the quoted words in French addressed to the customs officer is again in accord with Aabi's constraints on switching. In contrast the phrase *la fille* ***məskīna*** the adjective ***məskīna*** agrees in gender with the noun *la fille*, again in accord with Aabi's constraints on switching.

3.2.4. Yūmmu ʕAbdərraḥīmu

In the third and the last part of the episode Hanane plays an elderly, most probably Berber (as it suggests her tattooed chin) mother, whose son has supposedly gone to Italy and the poor mother continues to wait for him. At some point, we start to realize that she is most likely the mother of our *beau gosse* from the first part. In this manner, the last part of Hanane's three-part episode is tightly related to its first part through the figure of *beau gosse* echoing the ring composition used by Ibn Quzman in his zajals. Just like the Andalusian poet from zajal 84, the *beau gosse* didn't go anywhere.

Although entirely in darija, this humorous text represents what Baynham & Lee defined as intralingual translanguaging (2019: 93), where the actress brings together two different registers of Moroccan Arabic – everyday language and vernacular poetry. The first register is intended to mimic the speech of an illiterate woman from a rural area. This is achieved by using two tools: linguistic and stylistic choices. In terms of the former, Register 1, the variety of Moroccan Arabic used by *Yūmmu ʕAbdərraḥīmu* which is almost free from lexical borrowings (there are exceptions – borrowings that are already considered by speakers of Moroccan Arabic as words in Arabic – *kartōna*, *kāmīra*, etc) and, similarly to the borrowings from Romance in the zajals of Ibn Quzman that became completely assimilated by Arabic and therefore can't be treated as cases of switching (Corriente 2008: 80). In terms of the latter, Register 2, the text is built stylistically on numerous repetitions: chunks of everyday speech are artfully mixed with lines of vernacular poetry exhibiting traces of different rhythm and elements of rhyme[14]. The result of this artistic languaging reminds us of a kind of *chaâbi* song – it has the structure of a song, where stanzas are made of non-rhymed text (Register 1) and are followed by more or less the same refrain (Register 2), as shown in the excerpt below:

> *s-salām ana yūmmu ʕAbdərraḥīmu ʕarəftu ʕAbdərraḥimu* ***ma ʕarəftu š ʕAbdərraḥimu a wīli ʕAbdərraḥimu ṭ-ṭwīl z-zīn w-l-ḥəjbu məgrūnīn w-ʕaynīnu mbəllgīn*** *a tšūfū howa tšūfūni anāya yəšbuh līya wlīdi wlīdi ḥnīn wlīdi kbīdti twaḥḥaštu bəzzāf wlīdi wlīdi mhājər aw wāš ngūl l-kum huwa lli bġa yəhājer ayya mʕāya mmi ana bāġi nhājər bāġi nhājər mšīt xtībṭ lu Hājar bənt si Smāʕīl wālu ma bġā š yeglis līya w-ma bqa līya wlīdi w-nəbqa nətfakkər f wlīdi ngūl wlīdi wāš wākəl wlīdi wāš ma wākəl š wlīdi wāš nāʕəs fōg n-nāmūsīya wulla fōg kartōna alla ya wlīdi alla* ***ma ʕarəftū š ʕAbdərrḥimu a wildi ʕAbdərrḥim ṭ-ṭalyānu ṭ-ṭwīl z-zīn w-l-ḥəjba məgrūnīn w-l-ʕaynīn mbəllgīn*** *ʕAbdərraḥimu wlīdi ya wlīdi alla w-aš ġādi ndīr aš ġādi ndīr jbədt lli quddāmi w-lli mrāya rākum ʕAbdərrḥimu bāš nseyfṭu l-Ṭalyān beʕt l-ḥaṭṭa w-d-dahab w-beʕt ṣ-ṣəddari.*
>
> Hello, I am Abderrahimu's mother, do you know Abderrahimu? [What?] **You don't know Abderrahimu? Aw, Abderrahimu is tall and handsome, with arched eyebrows and bright eyes**. Look, he, no look at me, he looks like me a lot, oh, my son, my son, I long for him so much, he is my heartbeat, I miss him so much! My son left the country, what can I tell you – he was the one who wanted to emigrate. He said: mother, I want to emigrate, I want to emigrate, I want to emigrate. So I went and found for him Hajar, the daughter of Si Smail, however, he didn't

[14] Compare this refrain to the following lines from Ibn Quzman's zajal 87: *[raqbatan] šaṭṭah bayḍa miṯla l-quṭūn / ʿaynan akḥal wa-ḥājiban maqrūn* (I beheld a slender neck, white as cotton; / A collyrium-dark eye, and an eyebrow joined to its twin) (Monroe 2017: 538-539).

> want to stay. And so I had no child anymore and I started thinking about him, I was saying: my son, has he eaten or not? Or my son, does he sleep in bed or on some cardboard? Oh, my son, don't you know Abderrahimu? **Oh, my son Abderrahimu, the Italian, is tall and handsome, with arched eyebrows and bright eyes**. My son Abderrahimu, my son, what should I do, what should I do? I collected everything I had to send Abderrahimu to Italy, I even sold my gold and the sofa (El-Fadhili 2016).

The way Hanane weaves into the mother's monologue these repetitions reminds us of the stylistic differentiation from the rest of the text of refrains from strophic Andalusi poetry, meant to bring the audience's attention to a particular theme (Monroe 2017: 1047--1048). Just like the poetic texts created by Andalusian poets in vulgar Arabic, Hanane's text is artistic creation, based on exaggeration, with sociolinguistic recognizability as a key factor in its artfulness.

4. Discussion

Our starting point for this paper was a sense of Morocco and its Darija as being in some way peripheral in the Arab world over a thousand year span through its geographical positioning between Africa and Europe and on the frontier of the Arab world. Using insights derived from both linguistics and literary studies we have examined the interaction of Maghrebi Arabic and Romance languages in artful texts from the perspective of both historical and contemporary evidence using the notion of translanguaging. The language features we have been examining can be seen in one sense as an enactment of this peripheral, border crossing positioning. We can see this intimate linguistic and cultural engagement, which Bossong describes as *convivencia*, crossing linguistic borders within utterances, phrases, within words even. To do so, we have drawn on two sets of data, historical and contemporary. While focusing on translanguaging as our theoretical framing we have not ignored the important insights from studies of code-switching, indeed Aabi's analysis has been able to confirm for us that the switches we observe in the kharjas and the zajal of Ibn Quzman, as well as in the song of Talyani and the monologue of Hanane are not some arbitrary literary invention as has sometimes been argued, but grounded in everyday language usage. Translanguaging, with its emphasis on the speaker and their creativity is particularly apt as an approach to the analysis of artful language use, poems, songs, dramatic monologue, but we would also want to assert, as others have, the creativity of everyday language. What we have tried to demonstrate in our analysis is the crossover between artful language use and the everyday. This is of course not to reduce the artful to the everyday. What we encounter in the texts we have examined is a synergy between both, the artfulness of the text drawing energy and strength from its engagement with everyday language use. Of course we have no way of knowing how the actual language of the street played out in 11th century al-Andalus, but we believe that applying Aabi's analysis suggests that the 11th century audiences would have found the multilingual language use in the kharjas and zajals recognizable and that the artfulness of the poets known and unknown would have been in part to draw on that recognizability.

In addition to finding similarities in the multilingual language use, we also identified some evidence of literary continuity between the two sets of data we used. This is particularly convincing when comparing the legacy of Ibn Quzman and contemporary data. Thus, in Quzman's zajals, as well as in the Talyani's lyrics and Hanane el-Fadhili's monologues, there emerge the common features of the main protagonist – a trickster and a rogue. It is around this protagonist that the plot is built. While comparing the data from two time slices, we also found structural similarities: to start her monologue, Hanane uses a small fragment from Talyani's *Partir loin*. By doing this, Hanane sets the theme for the whole episode, so, in literary terms, the role of *Partir loin* in Hanane's text is comparable to the role the kharja had in the Andalusian muwaššaḥ. Finally, as if confirming the assumption made by James Monroe about the continuity between Ibn Quzman's zajals and the oral poetry of North Africa (2017: 1102-1103), we also spotted a certain parallel between stock phrases used to describe a handsome young man in Ibn Quzman's Diwan and Hanane's comical sketch.

5. Conclusion

In this paper we have explored the interaction between Maghrebi Arabic and Romance languages from the perspective of both historical and contemporary evidence. For our analysis we applied the notion of translanguaging understood as an enactment in language of *convivencia*, to artful texts from two time slices – kharjas and zajals from the 11th-13th century al-Andalus, a song by the Algerian singer Reda Talyani and a comical sketch by the Moroccan actress Hanane el-Fadhili. While framing our analysis in the translanguaging approach, we drew on insights from earlier research on code-switching in both Andalusi Arabic and Maghrebi Arabic as well as relevant data from literary studies and demonstrated that the artfulness of the texts we examined was informed by everyday language use. In addition, we were able to demonstrate that there exists a certain continuity over the centuries that links literary production from al-Andalus to the literary production of contemporary Maghreb.

Finally, we would endorse from our perspective Heath's argument on the need for interdisciplinary work between Arabic and Romance scholars, a kind of intellectual translanguaging, itself represented in this paper, written by two authors whose backgrounds embody the Arabist (Gintsburg) and the Romanist (Baynham) perspective. Future research might potentially involve further cross disciplinary work between researchers in Arabic and its Romance contact languages to will enable us to fully address the sociolinguistics of Moroccan/Algerian Darija.

References

Aabi, Mustapha. 2020. *The syntax of Arabic and French code switching in Morocco*. London: Palgrave Macmillan.

Abu-Haidar, Farida. 1989. The diminutives in the Diwan of Ibn Quzman: A product of their Hispanic milieu? *Bulletin of the School of Oriental and African Studies* 52(2). 239-254.

Baynham, Mike & Gintsburg, Sarali. 2022. Tar or honey? Space and time of Moroccan migration in a video sketch comedy 'l-kāmīra la-kum'. In Breeze, Ruth & Gintsburg, Sarali & Baynham, Mike (eds.), *Narrating migrations from Africa and the Middle East: A spatio-temporal approach*, 157-174. London: Bloomsbury.

Baynham, Mike & Tong King Lee. 2019. *Translation and translanguaging*. London: Routledge.

Blommaert, Jan & Max Spotti. 2017. Bilingualism, multilingualism, globalization and superdiversity: Toward sociolinguistic repertoires. In Garcia, Ofelia & Flores, Nelson & Spotti, Massimiliano (eds.), T*he Oxford handbook of language and society*, 161-178. Oxford: Oxford University Press.

Bossong, Georg. 2003. El cambio de código árabo-románico en las kharaǧāt e Ibn Quzmān. In Temimi, Abdeljelil (ed.). *Hommage à l'École d'Oviedo d'Études Aljamiado (dédié au fondateur Álvaro Galmés de Fuentes)*, 129-149. Zaghouan: Fondation Temimi pour la Recherche Scientifique et l'Information.

Bossong, Georg. 2010. *Poesia en convivencia: Estudios sobre la lírica árabe, hebrea y romance en la España de las tres religiones*. Gijón: Ediciones Trea.

Canagarajah, Suresh. 2011. Codemeshing in academic writing: Identifying teachable strategies of translanguaging. *Modern Language Journal* 95. 401-417.

Caubet, Dominique. 2002. Jeux de langues: Humor and codeswitching in the Maghreb. In Rouchdy, Aleya (ed.). *Language contact and language conflict in Arabic*, 233-255. London: Routledge.

Corriente Córdoba, Federico. 2008. *Code-switching and code-mixing in Ibn Quzman revisited*. In Döhla, Hans-Jörg & Montero Muñoz, Raquel & Báez de Aguilar González, Francisco (eds.). *Lenguas en diálogo: El iberorromance y su diversidad lingüística y literaria: Ensayos en homenaje a Georg Bossong*, 65-86. Madrid: Iberoamericana – Vervuert.

Corriente Córdoba, Federico. 2009. The kharjas: An updated survey of theories, texts and their interpretation. *Romance Philology* 63(1). 109-129.

El-Fadhili, Hanane. 2016. *Al-Kāmīra lakum: Al-Awla*. Rabat. (https://www.youtube.com/watch?v=9vxoYPKz7WU) (Accessed 25-03-17.)

Ferrando, Ignacio. 1997. G. S. Colin y los berberismos del árabe andalusí. *Estudios de Dialectología Norteafricana y Andalusí* 2. 105-146.

Ferrando, Ignacio. 1998. On some parallels between Andalusi and Maghrebi Arabic. In Cressier, Patrice & Aguadé, Jordi & Vicente, Ángeles (eds.), *Peuplement et arabisation au Maghreb occidental: Dialectologie et histoire*, 59-73. Madrid–Zaragoza: Casa de Velázquez-Universidad de Zaragoza.

Ferrando, Ignacio. 2000. The Arabic language among the Mozarabs of Toledo. In Fishman, Joshua A. (ed.), *Arabic as minority language*, 45-63. Berlin–New York: De Gruyter.

Frenk Alatorre, Margit. 1975. *La jarchas mozárabes y los comienzos de la lírica románica*. Mexico: El Colegio De Mexico.

Gintsburg, Sarali. 2020. Living through transition: The poetic tradition of the Jbala between orality and literacy at a time of major cultural transformations. Rilce 36(4) (*Transitional texts: Drifting between the oral and the written*, invited eds. Gintsburg, Sarali & Ford, John C. & Asier Barandiaran Amarika). 202-222.

Heath, Jeffrey. 2020. Moroccan Arabic. In Lucas, Christopher & Manfredi, Stefano (eds), *Arabic and contact-induced change*, 213-223. Berlin: Language Science Press.

Ibn Khaldûn. 2015. *Al-Muqaddimah: An introduction to history: The classic Islamic history of the world* (translated by Franz Rosenthal). New Jersey: Princeton University Press.

Lafkioui, Mena. 2025. Darija and the global multilingual digital landscape. *Lingua Posnaniensis* 67(1). 35-53.

Lahlou, Moncef, 1991. *A morpho-syntactic study of code-switching between Moroccan Arabic and French*. Austin: The University of Texas. (Doctoral dissertation.)

Monroe, James. 2017. *The mischievous Muse: Extant poetry and prose by Ibn Quzmān of Córdoba* (d. AH 555/AD 1160). Leiden: Brill.

Moscoso García, Francsico. 2010. La pentaglosia en Marruecos: Propuestas para la estandarización del árabe marroquí. *Miscelánea de Estudios Árabes y Hebraicos* 59. 45-61.

Prémare de, Alfred-Louis. 1985. *Sidi 'Abd-er-Rahman el-Mejdûb: Mysticisme populaire, société et pouvoir au Maroc au 16e siècle*. Rabat – Paris: Centre National de la Recherche Scientifique–SMER Éditions–Diffusion.

Stern, Samuel. 1974. *Hispano-Arabic strophic poetry* (selected and edited by L. P. Harvey). Oxford: Clarendon Press.

Talyani, Reda. 2007. *Partir loin*. (https://www.youtube.com/watch?v=DLMkUr_GIIc) (Accessed 2025-03-17).
Vicente, Ángeles. 2020. Andalusi Arabic. In Lucas, Christopher & Manfredi, Stefano (eds), *Arabic and contact-induced change*, 225-244. Berlin: Language Science Press.
Zavadovskiy, Yuriy. 1962. *Arabskiye dialekty Magriba* [*Arabic dialects of the Maghreb*]. Moskva: Izdatelstvo vostochnoy literatury.

DOI: 10.14746/linpo.2025.67.1.3

Darija and the global multilingual digital landscape

Mena B. Lafkioui

École des hautes études en sciences sociales – CNRS-LIER-FYT, Paris
m.lafkioui@ehess.fr | ORCID 0000-0002-1016-4071

Abstract: The present study investigates how Darija, within a complex multilingual and digital context, is reshaping the roles of traditionally dominant languages like Standard Arabic and French. It highlights a shift towards a more symmetrical sociolinguistic system, where local interactions redefine linguistic functions. The research also explores how Darija interacts with the Tamazight languages in the global digital sphere, addressing conflicts and competitions. It delves further into the concept of 'Darijation', an unintended result of North African language policies, and reveals that Darija is increasingly displacing other languages, creating a new linguistic landscape.

Keywords: Darija, Darijation, Tamazight, multilingualism, digitalisation, globalisation, conventionalised heteroglossia, intertextuality

1. Darija from an "integrating interactional perspective"

In this study, I examine how Darija, within its complex multilingual, globalised, and digital landscape, reframes and reshapes the roles of dominant languages like Standard Arabic and French. Traditionally, these languages have held superior and normative sociolinguistic positions. However, recently, Darija is increasingly facilitating a shift towards a more symmetric interactional system where sociolinguistic functions are locally negotiated and assessed (Lafkioui 2013, 2019, 2021, 2024). The study also focuses on how Darija-interactants establish language and cultural norms and accommodations in relation to the Tamazight languages and their local varieties within the global digital environment, examining the conflicts and competitions among them. Accordingly, the study inquires further into the concept of "Darijation" (Lafkioui 2024), an unintended byproduct of North African language policies driven by "Institutional Arabisation". This concept is pivotal for comprehending the political and sociocultural dynamics of contemporary North Africa and its diaspora. The findings reveal that Darija is encroaching upon and supplanting

the functions and linguistic practices of other languages, even those that have been traditionally dominant, leading to the emergence of a new linguistic landscape, as will be shown in the subsequent sections.

The study adopts an interactional sociolinguistic approach to meticulously look at the complex interplay between language, identity, and power (Goffman 1981; Gumperz 1982; Lafkioui 2019, 2024), particularly in the context of globalisation. At the heart of this approach is the focus on "interactants" – the individuals engaged in social interactions – rather than viewing language as a detached abstract concept. Interactants collaboratively generate and regenerate meaning, thereby producing and perpetuating cultural values, identities, and ethnicities.

Significantly, this study considers both linguistic and extralinguistic features of interactions, which are intertwined with historical, social, cultural, and political contexts. This "integrating interactional paradigm" (Lafkioui 2013, 2024) incorporates concepts from linguistic ethnography and anthropology, focusing on the dynamics of power and its manifestation through language, whether in practice or in theory (Blommaert 2010; Bourdieu 1982; Fairclough 1989; Gal 2006). Consequently, Lafkioui's "integrating interactional paradigm" emphasises the necessity of combining linguistic and extralinguistic perspectives to fully comprehend human interaction and, by extension, human nature. The linguistic perspective encompasses the study of various dimensions – from prosody to syntax, semantics, and pragmatics – and pertains to all levels of interaction. These range from the minimal unit, the speech act, to the maximal unit, the interaction paragraph, whose structure is tied to the extralinguistic context, often conveyed through prosody.

This paradigm has informed my research on language and culture from the outset, shaped by extensive fieldwork in North Africa and Europe since the mid-1990s. The data and analyses presented in this study were gathered from various offline and online settings, resulting in a substantial ecological, multilingual, and multimodal corpus from Africa and Europe.

The structure of the study is as follows: Section 2 addresses Darija's position within North Africa's landscape of layered and stratified multilingualism. Section 3 focuses on the concept of "Darijation" and its role and impact within this linguistic landscape. Section 4 discusses recent developments that have led to the perception and representation of Darija as part of the Tamazight heritage. Section 5 examines how Darija is framed and reframed in global and digital contexts. The study concludes by presenting the overall findings.

2. Darija within North Africa's "layered and stratified multilingualism"

North Africa today presents an intricate sociolinguistic landscape marked by what has been termed "layered and stratified multilingualism" (Lafkioui 2008, 2013, 2024) while referring to the setting in which "the various languages in use do not hold equal sociolinguistic status nor serve identical sociocultural functions. Instead, the sociolinguistic hierarchy of languages is primarily determined by national and local policies. Both offline and online, the activation or non-activation of different linguistic resources inevitably

signifies variation in interactive functions and the social categories associated with them by the interactants" (Lafkioui 2024: 20-21).

In this diversified sociolinguistic landscape, which reflects the complex historical, social, and political interactions of North Africa, the Tamazight languages (Afroasiatic) stand out as the only endogenic languages (Section 2.1) alongside numerous exogenic ones. Darija also stands out as it is an "endogenised" contact language which has Tamazight as one of its main components, as will be explained in Section 2.2.

Among the various languages attested in North Africa, there are sub-Saharan African languages such as Songhay (Nilo-Saharan), Fula and Wolof (Niger-Congo), and Hausa (Afroasiatic), which are regularly used as contact languages among the Zenaga and Tuareg Amazigh peoples in the Sahara and northern Sahel regions.

Arabic, in its classical, standard, and vernacular forms (Semitic, Afroasiatic), was introduced to predominantly Tamazight-speaking North Africa through Islamic conquests mainly starting from the 7th century. These conquests initiated the process of Arabisation, which gained significant impetus many centuries later, particularly following independence from Western colonial powers in the 20th century when the newly established nation-states adopted Arabisation policies as a chief precept, coined "Institutional Arabisation" in Lafkioui (2013, 2024). Institutional Arabisation operates as a cyclical process, closely aligned with the shifting dynamics of local and global hegemonic conjunctures. It is primarily driven by nationalist governance policies that impose language changes from the top down and that "has persistently aimed at establishing Standard Arabic as the national language, often invoking Islam as justification for this endeavour" (Lafkioui 2024: 19). The Institutional Arabisation policy is influenced by both French centralist Jacobinism and Nasserist and Baathist pan-Arabism (i.e., *ʿurūba*), blending elements from both ideologies to promote linguistic and cultural uniformity.

Among the diverse Indo-European languages introduced to North Africa primarily through Western colonisation, French and Spanish still play significant roles in the region's power dynamics. Additionally, English functions prominently as the international lingua franca.

2.1. Tamazight

Tamazight, the endogenic language family of North Africa, comprises around forty distinct languages and their local varieties, all of which form a specific branch of the Afroasiatic phylum. These languages are only mutually intelligible among neighbouring varieties or those within the same subgroup or type. Otherwise, effective communication typically requires formal education or extensive exposure to the different languages. Even within a single Tamazight language, there can be significant variations that hinder mutual understanding among speakers of different varieties.

Overall, the Tamazight languages in North Africa form a linguistic continuum, with no clear-cut boundaries separating one language from another. Instead, there is a gradual transition from one language to the next, reflecting the intricate and overlapping nature of this language family (Lafkioui 2018, 2024).

Tamazight encompasses ancient language forms, historically known as Libyan or Numidian, which date back to the 5th-10th century BCE. These early forms evolved into

both ancient and modern Tifinagh scripts. Tifinagh remains the endogenic writing system for the Amazigh peoples and is still actively used by the Tuaregs, who primarily live in the Sahara and northern Sahel regions, collectively also known as southern Tamazgha. Over time, Tifinagh, particularly its Neo-Tifinagh version, has been adapted from its original form. In northern Tamazgha, especially Morocco, Neo-Tifinagh saw a development after Tamazight was incorporated into the official education system in 2003.

Despite the increasing use of Neo-Tifinagh and the recent official recognition of Tamazight alongside Standard Arabic in Morocco and Algeria, the adoption process remains inconsistent and imprecise. Both countries are currently working on developing a standardised form of Tamazight. In Morocco, the official standardisation uses the Tifinagh script, while in Algeria, it employs the Latin script. These efforts aim to unify the various Tamazight languages at the national level. However, these standardisation initiatives often face significant resistance from Tamazight-practicing communities. The primary concerns stem from the subpar outcomes of these initiatives and their limited practical impact on key areas such as education and administrative functions.

A critical issue is that the standardisation process fails to adequately consider the regional and local variations of the Tamazight languages, which reflect significant demographic, sociocultural, and historical diversity. This oversight undermines the effectiveness and acceptance of the standardised forms among native speakers.

Moreover, the implementation of the Tamazight language project has experienced significant delays, especially within the education sector. In Morocco, for instance, the initiative to expand Tamazight education, which began in 2003, has seen sluggish progress. Originally, there was a promise that by 2010, Tamazight would be taught at all educational levels – from primary schools to universities – throughout the country, including in predominantly Darija-speaking regions. However, the current state of Tamazight education in Morocco fails to deliver. It is confined to the primary grades and suffers from inadequate quality. This deficit is partly due to a shortage of qualified teaching staff and insufficient appropriate pedagogical materials. As a result, the promise of comprehensive Tamazight education at all levels is far from being fulfilled.

Instead, the current situation highlights how Tamazight and its activism have been heavily instrumentalised since its recognition as a "national" and later "official" language in Algeria and Morocco starting in the 1990s. Despite the formal acknowledgment, the practical implementation and genuine support for Tamazight remain inconsistent, mostly leveraged for political and economic purposes than for true linguistic and cultural preservation and development.

One effect of this instrumentalisation is the noticeable decline in the use of Tamazight attested across North Africa and its diaspora, even in regions with substantial Tamazight-speaking populations, such as Southern Morocco where Tashelhit is prominent. Darija is progressively replacing Tamazight across all social classes. Additionally, Standard Arabic is displacing French and Spanish, particularly among the educated middle class. These trends underscore the challenges faced by Tamazight-interactants in maintaining their linguistic and cultural heritage amidst broader linguistic shifts and societal changes in the region.

Consequently, those Imazighen who have resisted the instrumentalisation of Tamazight, often referred to as *ḫubza* ('loaf of bread' in Darija), meaning clientelism, continue to

pursue their struggle for language, cultural, and identity rights through non-governmental networks. Amidst these trials, the vigorous advocacy by numerous non-governmental organisations and platforms has significantly elevated the social and political prominence of Tamazight languages and cultures in recent years. This renaissance is further bolstered by a remarkable surge in scholarly inquiry and cultural output dedicated to the Tamazight linguistic and cultural legacy. Notably, digital media contributed to the formation and expansion of "Amazighness" or *Tamuzgha*, the "trans-local (pan-)Amazigh collective identity", in which both Tamazight and Tifinagh serve as icons (Lafkioui 2008, 2013, 2024).

2.2. Darija

Darija or *Darğa* (or variants) is:

> a gradually varying language continuum that spans North Africa and functions as a lingua franca, emerging from the interaction between Tamazight, its substratum and sole endogenous component, and Arabic since the 7th century. In addition to the substantial influences of Latin and Greek on Darija, adstrata of Tamazight since Antiquity, the impact of Portuguese, Spanish, and French is even more pronounced, with the latter two still actively contributing to its development, along with other pluricentric languages like English. Consequently, Darija encompasses more than the commonly understood translation of 'Arabic dialect' or its national equivalents, like e.g., Moroccan Arabic, Tunisian Arabic, Libyan Arabic or their abbreviated counterparts like e.g., Moroccan, etc. Hassaniyya is also part of this continuum, forming its peripheries not only geographically but also linguistically. Its distinctive features arise from contact with various sub-Saharan languages, such as Wolof (Niger-Congo). Hassaniyya is principally practiced in Mauritania, Morocco, Algeria, Burkina Faso, Mali, Niger, Senegal, and the Western Sahara (Lafkioui 2024: 21).

The use of Darija is almost unavoidable when attempting to spontaneously speak (Modern) Standard Arabic, or *al-Fuṣḥā*, in North Africa. Switching between Standard Arabic and Darija has become so routine that a kind of "intermediate language variety" quickly emerged following the introduction of Standard Arabic as the official language of the newly formed nation-states post-independence. This hybrid form is now frequently employed in formal and semi-formal educated settings. While it is plausible to categorise this "intermediate" variety of Darija as a distinct form, akin to what Youssi (1995) refers to as "Middle Moroccan Arabic", it remains debatable whether this is a language in its own right (as in e.g., Ennaji 2001; al-Midlāwī 2019; Youssi 1995) or rather a register or set of registers of Darija – such as an "educated register" – with its own genres and styles, like an "artistic style", for instance.

In recent years, several initiatives have emerged attempting to standardise Darija or engage in related debates, particularly in the realm of orthography (e.g., Aguadé 2006; Caubet 2017; Durand 2004; Hoogland 2014; Michalski 2019; al-Midlāwī 2019; Miller 2017; Moscoso 2009; Moustaoui Srhir 2016). These standardisation efforts, while unofficial – since Darija lacks any official status in North Africa – often take this "intermediate" variety of Darija as a starting point, typically considering Standard Arabic as a reference (e.g., al-Midlāwī 2019; Youssi 1995). This approach is evident even in the

way "Darija" is written, often with a long vowel *ā* as in *Dārija* or *Dāriğa*, despite the absence of long vowels in Darija, a trait it shares with Tamazight. This trend is especially prevalent in academic circles, whose suggestions are increasingly picked up by stakeholders in the political and business sectors aiming to instrumentalise Darija, as will be discussed in Section 3. These stakeholders usually have no genuine interest in Darija as a language or its practice as cultural capital; rather, they view it as a blemish, a reminder of the failure of their Arabisation project or as a means to amend it.

A typical example of how this "intermediate" variety of Darija, also known as *ad-Dārija al-Wusṭā*, is used to enhance the linguistic and cultural competencies in Standard Arabic among North Africans, thereby advancing the Arabisation project, is reflected in the efforts of the Zakoura Foundation. Established in 1997 in Casablanca, this foundation published a dictionary in Darija with the explicit aim of perpetuating, renewing, and expanding Standard Arabic and the culture it represents, as advocated in Chekayri (2018).

The paradox lies in the fact that institutions like Zakoura, which claim to promote rural development, use the local mother tongue, Darija, not to sustain it but as a means to introduce the exogenous and dominant Standard Arabic through an intermediate linguistic form, *Dārija al-Wusṭā*. They do not hesitate to employ other dominant languages, such as French and English, for broader exposure and economic facilitation. For instance, Zakoura's current website is almost entirely in French, with no Darija presented – only its speakers are depicted through typical rural images. Thus, Darija, like other interactionally "dominated" languages, becomes merely a tool for obtaining and maintaining power, both politically and economically.

Figure 1. Maroc Telecom advertising[1]

[1] Source: https://www.iam.ma/index.aspx (Accessed 2024-10-01).

It is hardly surprising then that one of the first sectors to adopt Darija in public spaces after independence was the telecommunications industry; an industry that continues to do so for the same neo-capitalistic objectives.

For instance, Figure 1 illustrates how nowadays stakeholder *Maroc Telecom* draws on Darija, often framed within a multilingual setting, to attract customers; e.g., the expression عِش الفُرجة *ʿiš al-furža* 'Live the spectacle' in Darija sets the focus of the attention, while la fibre 'the fiber' and Méga in French, together with the Standard Arabic إلى غاية *ʾilā ġāya* 'up to' provide more practical details.

Conversely, there is an emerging interest group that seeks to "organically" standardise Darija through various forms of creative and educational expression, including writing. Digital media have been particularly instrumental for this purpose, as they facilitate heteroglossic practices, which refer to "multilingual interactions relating to diverse intersubjective voices construed from diverse sociocultural interactional positions" (Lafkioui 2021). In the case of Darija, this is supported by heterography based on either the Arabic or Latin alphabet (Section 5).

3. Darijation

Darija, as a lingua franca in North Africa, is increasingly infiltrating all areas of interaction, including those traditionally dominated by other languages, such as French and Spanish. Language choices and usages, often featuring jargon specific to contexts like academia where French was once prevalent, now frequently include Darija. This often involves code-switching with other dominant languages such as Spanish or English, and occasionally with Tamazight as well. In fact, switching between Darija and Tamazight on social media is a common practice among certain groups, serving various interactional purposes, including playfulness.

The ascendancy of Darija across diverse spheres of interaction is a direct outcome of Institutional Arabisation, which established Standard Arabic as the exclusive official language.

This policy, rigorously enforced in the 1980s and further entrenched in the 1990s, precipitated a profound transformation in the educational landscape of North Africa. Subjects formerly taught in French, such as science, are now predominantly conducted in Standard Arabic within national public education systems. Notwithstanding, this shift frequently involves a dynamic interplay with Darija, French, and English, reflecting the complex and evolving linguistic and cultural landscape of the region.

Despite the apparent failure of the Arabisation process, particularly evident in public education and research, authorities persist in advocating for the use of Standard Arabic throughout society. However, many argue that a more effective approach might involve preserving local identities through Darija and Tamazight, potentially alongside a pragmatic re-engagement with languages like French or a shift towards English, which offer greater international influence and visibility (see also e.g., Bouziane & Saoudi 2021).

English is not new to North Africa, especially in Morocco, where American influence – both civil and military – have been significant since the early twentieth century. This

influence is increasingly visible through various private educational networks, such as the American Institute for Maghrib Studies (AIMS), established in 1984, with affiliated partners in Tangier (TALIM), Oran (CEMA), and Tunis (CEMAT).

Although Arabisation remains a significant sociopolitical effort deeply embedded in Arab-Islamic culture, this policy has not fully supplanted the widespread use of Darija. Instead, the current situation highlights a shift from Arabisation to what Lafkioui (2024: 23) terms "Darijation" – "the systematic adoption and proliferation of Darija across all levels of society, including formal interactional settings". Darija has gained substantial traction in North African society, especially in Morocco, where it has increasingly overshadowed Tamazight. Traditionally spoken by the majority, particularly in rural areas where it was the sole language for many, Tamazight is now being eclipsed by the rise of Darija. The aggressive Arabisation campaigns of the 1990s in Morocco and Algeria profoundly reshaped the sociolinguistic landscape, triggering a swift shift from Tamazight to Darija.

This transformation, intimately linked with the propagation of Sunni Islam, deftly benefits from religious institutions like the Institut Mohammed VI pour la formation des Imams Morchidines et Morchidates, established by the Moroccan monarchy in 2013. These hubs of authority equip imams with specialised training, moulding them into key figures within a broader strategy that blends linguistic and religious objectives. Through the prism of an Islamic framework, these imams are tasked with advancing Standard Arabic – a strategic move aimed not only at countering the influence of Shiism but also at subtly pressuring Tamazight speakers to relinquish their ancestral language, often via the intermediary of Darija when necessary. As a result, Darija, typically downplayed by policymakers as a mere dialectal branch of Standard Arabic, emerges as a deliberate tool for Arabisation, reinforcing this calculated agenda. Institutions such as Zakoura, which push the adoption of Standard Arabic under the guise of Darija, encapsulate this orchestrated effort (see Section 2.2).

Consequently, a significant portion of the Tamazight-speaking community has transitioned to Darija, adopting it not only as their first language (L1) but as the educational foundation for their children – a decision shaped by the desire for academic success and social mobility, and, in some cases, guided by religious undercurrents. Ironically, some religious leaders, including state-appointed imams, publicly denounce Tamazight and its cultural practices while paradoxically utilising the very languages they seek to undermine, whether Tamazight or Darija, in their own discourse (Lafkioui 2024).

Darija faces disparagement not only from policymakers but also from its own speakers, who exhibit ambivalent attitudes toward it. Often ridiculed in comparison to Standard or Classical Arabic and other dominant languages like French – whose prestige remains high and continues to serve as the lingua franca among the elite – Darija is frequently undervalued. Nonetheless, it remains crucial for conveying emotion, particularly in verbal interactions, unless one is entirely immersed in the "select international bubble" where French and English predominate, such as in expatriate communities and their international schools.

Even within the realm of vernacular Arabic varieties, i.e., the so-called "Arabic dialects", Darija is perceived as anomalous, as deviant even. This is illustrated by the meme in Figure 2, which depicts Darija as the sole nonconformist in an otherwise harmonious

and traditional family of Arabic and its varieties. This deviation from norms does not inherently carry a negative connotation. In reality, the significance of this portrayal is highly context-dependent. On social media, where the meme is frequently recontextualised, Darija's unique characteristics are often embraced positively. It is used constructively to foster collective identities, such as Moroccan identity, highlighting how Darija's distinctiveness can contribute to cultural cohesion and pride.

Another striking feature of Figure 2's meme lies in the attribution of Classical Arabic to the father and Standard Arabic to the mother. This portrayal defies the typical belief that women tend to be more conservative in preserving and transmitting language, particularly since Standard Arabic is a contemporary evolution of Classical Arabic. By assigning the father the role of guardian of Classical Arabic – viewed as the pinnacle of linguistic purity – the meme reinforces established Arabic-Islamic cultural (including gender) norms. In this context, Darija seems conspicuously detached from the lineage represented in the meme, almost as though it is marginalised from the patriarchal heritage entirely.

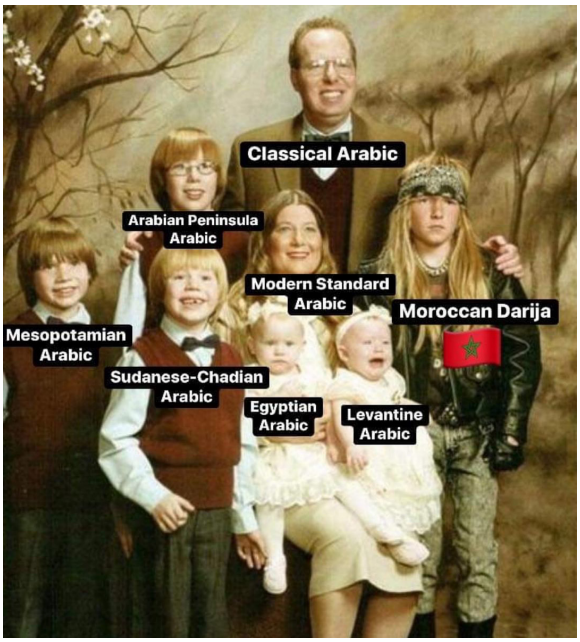


Figure 2. Meme of Darija amidst Arabic varieties[2]

An additional remarkable illustration of Darija's distinctiveness can be seen in a popular YouTube video (https://www.youtube.com/watch?v=blrGmR4-qaY) titled "Students Speak Different Arabic Dialects" and wherein young people compare their primary languages, including Darija and various Middle Eastern Arabic varieties. Despite the laughter

[2] Source: https://ifunny.co/picture/classical-arabic-arabian-peninsula-arabic-modern-standard-arabic-moroccan-darija-mR1MNQou7?s=cl

and teasing directed at his "translations" into Darija – some of which incorporate French and Spanish – the young Darija-interactant remains unphased. His confidence and even pride in Darija reflect a relatively recent phenomenon, particularly noticeable on social media platforms. This video has generated thousands of comments, in English mainly!

Here are some to illustrate the phenomenon; quoted comments are retaken as such here: @NUNS posts as comment "The moroccan didn't even bring the deepest vocabulary of Darija and they're still confused", to which replies, for instance, @aaabatteries5576 by saying "the only one I understood I'm Algerian", while @die4race says "That's because Darija is a Language on itself, its not just a Dialect" and @nanaa428 "dude i dont even think darija is 3arabi", to which @omarfilali6659 replies "it's not Arabic, definitely not, as a moroccan, I think moroccans are not arabs, either arabiszed moroccans(aerobi), riffi, shloh, Amazigh (berber), so basically the most common darija you hear from moroccans, is either, barbarized arabic or arabized amazigh and some morocconized french and spanish depending on the…". Comments from non Darija-interactants, like that of @cooldiamondgamer611 in "I am arab and everything he said was gibberish to me", are also common.

The recent shift in the representation of Darija, relative to how Tamazight is represented, marks a significant development in North Africa's current sociolinguistic landscape. This transformation will be addressed in detail in the following Section 4, while focusing on Morocco.

4. Darija as Tamazight heritage

In North Africa, the recognition of Amazigh identity is closely tied to the use of Tamazight, emphasising the deep connection between language and ethnocultural belonging. As a result, the Imazighen's struggle for greater rights revolves largely around the acknowledgment and preservation of Tamazight. Tamazight, along with its endogenic script, Tifinagh, serve as powerful icons of "Amazighness", i.e., the "translocal Amazigh collective identity", referred to as *Tamuzgha* (or variants) in Tamazight (Lafkioui 2024). Despite the everyday dominance of other – often pluricentric – languages, Tamazight remains a defining feature of Amazigh identity and plays a pivotal role in shaping social and institutional power, especially in Morocco, which counts the largest numbers of Tamazight-interactants. In other words, discussions of Amazigh identity frequently focus on language, reflecting how ethnic and cultural identities in North Africa are closely intertwined with linguistic choices (Lafkioui 2013, 2024).

Traditionally, Tamazight-interactants, much like Darija-interactants, have regarded Darija as a degraded form of Standard or Classical Arabic. It has often been dismissed as the vernacular of uneducated rural descendants of the Arab invaders who swept into North Africa – the so-called *ʿrubiyya*. The term "Arab" frequently encompassed Arabised Amazigh populations who had lost their connection to their ancestral Tamazight tongue.

However, perceptions among Tamazight-interactants have begun to evolve in recent years. Darija is no longer seen solely as the native language of Arabic speakers or Arabs, nor is it scorned as the language of "lost" Imazighen. Instead, a growing movement –

largely fuelled by social media – is actively reframing Darija's identity. It is increasingly recognised not just as a dialect of Arabic, but as a language deeply intertwined with Tamazight, reflecting a unique linguistic fusion. Many now see Darija as distinctly North African, heavily influenced by Tamazight's linguistic and cultural heritage. This perspective is gaining traction even among Darija-interactants with no command in Tamazight, who share in this revaluation.

This is instanced in Figure 3, which illustrates this phenomenon perfectly through an excerpt of a post about Darija on a website *Framed à la Tamazight*, as understood in Lafkioui (2013: 142) and so "indicating the overall pro-Amazigh intersubjective viewpoint and, hence, offering a general template to interpret the online discourses". The site is hosted by *Imazigheninusa*, one of the most active Facebook groups and its associated social media outlets, primarily Instagram and Twitter. Stances similar to those expressed in this post are increasingly common among Darija-interactants, particularly among younger generations who have grown up in an era of globalisation and digital media. These individuals are acutely aware of the impact of these forces on society and their own Amazigh heritage, though they have not directly experienced the intense sociopolitical repression of earlier times. While repression still exists in different forms, linguistic and cultural rights have seen some recent advancements for Tamazight, which has been recognised as an official language of Morocco since 2016.

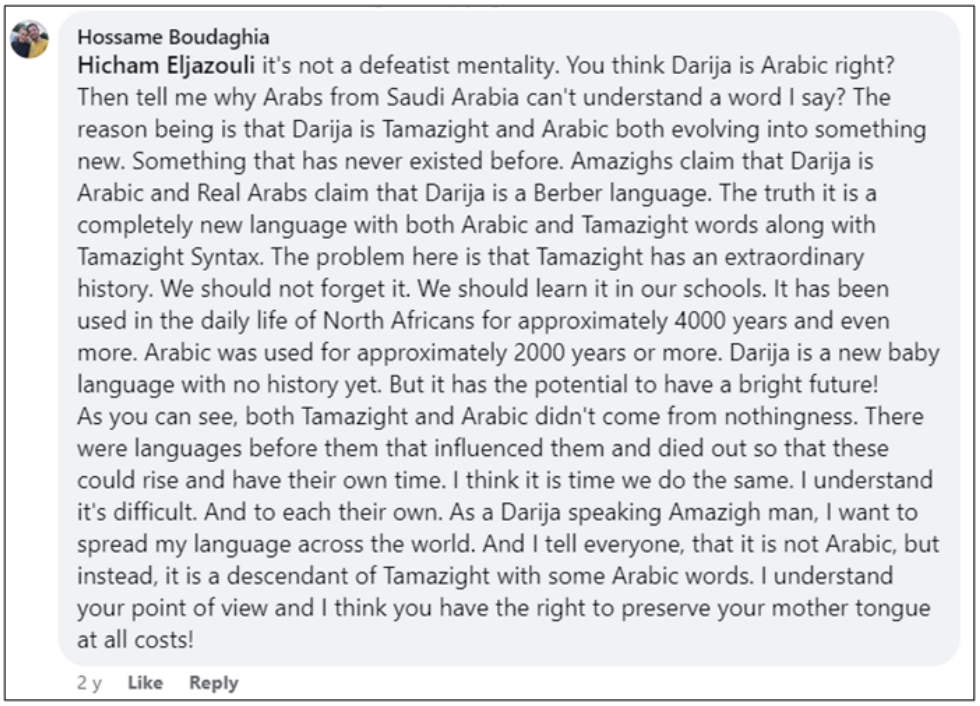


Figure 3. Post about Darija on a website *Framed à la Tamazight*[3]

[3] Source: www.facebook.com/groups/Imazigheninusa (Accessed 2024-10-01).

A major catalyst for the evolving perception of Darija is the increasing recognition among the Amazigh people of the pivotal transition from Tamazight to Darija. This change drives Darija-interactants, particularly those who have lost their Tamazight skills or never acquired them, to seek a deeper understanding. Concurrently, there is a rising consciousness among Darija-interactants, including within the *ʿrubiyya*, of their Amazigh heritage. This has led quite some of them to assert their identity more visibly, even via methods like DNA testing, which has recently become a popular topic on social media. For many, this evolving perspective presents a strategic opportunity to leverage the recognition of Tamazight as an official language for a range of sociopolitical, administrative, and economic advantages.

For those dedicated to preserving the Tamazight language and culture, this emerging trend is deeply troubling. If Darija is increasingly seen as part of Tamazight heritage, it could potentially threaten the survival of Tamazight itself. The expansion and entrenchment of "Darijation" could lead to the further erosion of Tamazight, particularly if this process unfolds organically from the grassroots level. Paradoxically, the very force capable of halting this organic process of Darijation – should such an intervention be desired – might be the formal recognition of Darija as an official language, which would necessitate its standardisation through top-down mandates. Such an institutionalisation could potentially undermine the fluid, adaptive nature of Darija, stifling its grassroots evolution and transforming it into a regulated linguistic form, disconnected from its dynamic, living roots (Section 5).

Thus, a profound sociolinguistic tension exists between Darija and Tamazight, a reality deeply felt by many Tamazight-interactants in their daily lives, particularly those who hold a strong connection to their ancestral heritage. Aware of the risks posed by a complete shift to Darija, these interactants grapple with the preservation of their linguistic roots. Memes, such as the one depicted in Figure 4, emerge as cultural tools, sparking dialogue and engaging in this ongoing debate surrounding language, identity, and heritage.

In this codeswitched meme, Tarifit (i.e., the Tamazight language of northern, northwestern, and northeastern Morocco) is written in Latin script, while Darija is rendered in Arabic script. The distinction between the two writing systems is significant, as it highlights the cultural differences they represent. In other words, the meme portrays the competing languages distinctly – illustrated through the representation of a healthy versus a disabled girl – and their use differs not only in content but also in form, with Latin script for Tarifit and Arabic script for Darija, as the Latin script is often regarded as the most "modern" choice among Tamazight-interactants (Lafkioui 2013, 2024).

So, the meme depicts Tarifit as the dominant force, with a girl speaking Tarifit pushing another girl, who speaks Darija, in a wheelchair. The girl steering the wheelchair asks in Tarifit, "Aren't you of Rif origin?", to which the girl in the wheelchair replies in Darija, "Yes, my father is a Rif Amazigh, but I don't speak Tarifit". In response, the Tarifit-speaking girl pushes her into an abyss, shouting "Then off you go into the abyss".

Interestingly, this meme also highlights the digital creator's greater proficiency in Darija compared to Tarifit, evident through the differences in orthographic and grammat-

ical quality and coherence between the two languages. Ironically, this linguistic imbalance may be the driving force behind the "dramatic" stance taken in Tarifit, emphasising the urgent need to preserve and revitalise Tamazight.

Additionally, the intersentential code-switching within the meme plays a crucial role in managing the interaction, functioning as a tool for turn-taking while also signalling shifts in stance. In so doing, this complex meme – like most memes – instantiates the concept of "conventionalised heteroglossia" (Lafkioui 2019, 2021), which relies heavily on "intertextuality" and will be addressed in the following Section 5.

Figure 4. Codeswitched meme (Tarifit – Darija) about language and identity[4]

5. Framing and Reframing Darija globally and digitally

Darija was represented by online content right from the beginning of the technological revolution and the creation of the World Wide Web, mainly through the commonly known chat rooms, fora, and blogs, along with various sites showcasing their edited content. However, it was with the advent of Web 2.0 that the landscape truly shifted for Darija. The transition from professionally edited content to user-generated material, facilitated by various online creative communities, significantly expanded its opportunities for creation and sharing.

While digital media enable Darija interactants to freely share, discuss, and develop content as they wish with minimal interference, some oversight is still exercised by platform moderators, who may refute or modify content when deemed necessary. After

[4] Source: https://www.facebook.com/groups/anessentarifit (Accessed 2024-10-01).

all, digital platforms remain institutionalised spaces that regulate, to some extent, the nature and function of interactions and shared material (Lafkioui 2008, 2013). Nonetheless, online exchanges are typically negotiated in a relatively symmetrical manner.

A notable example of how Darija has been digitally – and consequently globally – reframed is reflected in its evolving writing practices. One of the most significant milestones in the bottom-up standardisation and dissemination of Darija is its recent integration into Wikimedia, particularly Wikipedia (for a general overview of wiki-Darija, see Sedrati & Ait Ali 2019). Although these grassroots initiatives face challenges stemming from the heterography inherent to Darija as a non-standardised language, the interactive nature of such platforms offers the best chance for its preservation. These efforts encompass a wide range of informal expressions, covering informative, educational, and creative content. Some exceptional and promising examples come from authors like Mourad Alami, Hamid El Mahdaoui and Farouk El Merrakchi, who challenge the belief held by some Moroccan intellectuals that Darija is ill-suited for conveying "higher" cultural expressions – a view reflected, for instance, in Abdellah Laroui's interview for alyaoum24.com on 21 November 2013 [5]. Their works serve as counterexamples, demonstrating Darija's potential in articulating complex and culturally significant ideas.

Although online platforms typically *Framed à la Tamazight* are increasingly incorporating Darija due to the ongoing process of Darijation, the reverse is also true. Platforms and settings centred around Darija are gradually referencing and even using Tamazight. In some cases, this includes artistic expressions, such as the many comedic sketches in Darija that have long integrated Tamazight – often in a stylised, mocking manner. However, following Tamazight's official recognition in Morocco, its presence in such contexts has become more overt.

Darija's interaction with Tamazight, as well as with vernacular French and potentially other languages, has been ongoing for a long time, not only in North Africa but also in the diaspora (Lafkioui 1998, 2006). The key difference is that with the globalisation and digitalisation of communication, the degree of hybridisation – and the resulting shifts in the interactive landscape – has taken a significant leap. This is evident in how people express themselves and the increasing heteroglossia of the language forms they use, a phenomenon that is termed "conventionalised heteroglossia".

This latter concept stands for

> multilingual interactions relating to diverse intersubjective voices construed from diverse sociocultural interactional positions within specific, yet dynamic, conventionalised multilingual interactional frameworks. Accordingly, "conventionalised" refers to the joint management of polyphony within these interactions, contingent upon the nature of their heteroglossia, the framework in which they occur, and the extent to which they have become routinised (Lafkioui 2024: 31).

In other words, in the context of globalisation, especially in migration, Darija is increasingly contributing to multilingual interactions, testifying to the fluidity and adaptabil-

[5] Al-ʿarwī: Ḥāwalt al-kitāba bi-d-dāriža, https://alyaoum24.com/167927.html (Accessed 2025-06-05).

ity of language, and so showcasing how individuals navigate diverse linguistic landscapes in "glocal" interactions, interrogating the traditional correlation between languages and social as well as ethnocultural identities.

In this intricate sociolinguistic landscape, Darija may facilitate the accommodation, socialisation, and emancipation of multilingual interactants, whether of North African descent or not. As a matter of fact, the linguistic interactions of Darija-interactants – especially in the diaspora – are shaped by a complex interplay of cultural identity, historical background, and sociopolitical dynamics. In certain interactional contexts, this can lead to sophisticated multilingual code-switching between structurally distinct languages, both genetically and typologically, a phenomenon referred to as "incongruous multilingual code-switching" in Lafkioui (2021), of which the following example is retaken.

wa	**3ayyeqti**,	*die bal,*	*die bal,*	*daar*	*moete*	*zijn*,		
INTJ	exaggerate.PFV.2SG,	DIST ball,	DIST ball,	DIST	must.AUX.PRES.2SG	be,		
wa	HANDICAP **3ayyeqti**		*zijde*	*gij*	***ṯeddaɣreḏ***	*of*	*zo*	*jong*
INTJ	handicap exaggerate.PFV.2SG		be.PRES.2SG	2SG	be.blind.PFV.2SG	or	what	young

'Come on, that ball, that ball, there you should be, oh you nitwit come on, are you blind or what, man.'

In this complex configuration, the Darija expression *wa 3ayyeqti* ('oh you exaggerate' > 'oh come on') is alternated a few times with vernacular Dutch (specifically of Ghent, Belgium), culminating in the hybrid switch *wa HANDICAP 3ayyeqti* ('oh you nitwit, come on'). In this instance, the Dutch noun *handicap* is inserted as a frozen interjection, carrying various discursive functions with a high indexical value, often emphasised prosodically. Interestingly, placing *handicap* in the preverbal position does not align with Darija's word order, which would favour *wa 3ayyeqti a handicap* (preceded by the vocative *a*) if *handicap* were used as a vocative interjection. In fact, this configuration would not be adequate in Dutch either, but it would be in Tarifit (Tamazight of the Rif; North, Northwest, Northeast Morocco), indicating that Tarifit may be the base language of the hybrid switch. Additionally, the final switch to vernacular Dutch involves a sentence-internal switch, where the conjugated verb phrase *ṯeddaɣreḏ* ('you are blind') in Tarifit is used in place of an adverb, which vernacular Dutch morphosyntax would typically require in such a construction. Complex instances like this one serve various interactive purposes, including conveying intense emotions such as excitement and frustration, as is the case here.

These phenomena of "incongruous multilingual code-switching", which are becoming increasingly common in superdiverse, globalised contexts – especially in youth language – challenge our understanding of language evolution. They compel us to pay greater attention to the impact of language contact and digital mediation, which has opened the door to hybridisation on a wide scale. Despite the widespread cross-pollination driven by globalisation and digitalisation, much of it is fleeting. What remains, relatively speaking, is the result of "conventionalised heteroglossia", a process of routinisation that anchors these hybrid forms in both sociocognitive and substantial spatial dimensions.

In addition to being one of the languages of socialisation within the diaspora – used not only in intra-ethnic but also inter-ethnic contexts such as neighbourhood work and

youth outreach – Darija is increasingly becoming part of the multilingual landscape, both in urban settings and at national and transnational levels. For instance, expressions like *iwa* 'so, come on' and *iwa d-drari* 'come on guys' (VOC DEF-boys.PL), originally from Darija, have been adopted by children in Belgium and are now gaining national recognition and spreading (Lafkioui 2021).

A more prominent example, which recently caused a stir in France, is the Darija word *waš*, often rendered as *wesh*. Initially a question word, *wesh* has evolved into a filler word in French slang, serving a variety of interesting interactional functions, including signalling agreement or dissent. This is illustrated by the meme in Figure 5, where it precedes the vernacular French expression "Wesh vous faites quoi en Syrie" ('So what the heck are you doing in Syria.').

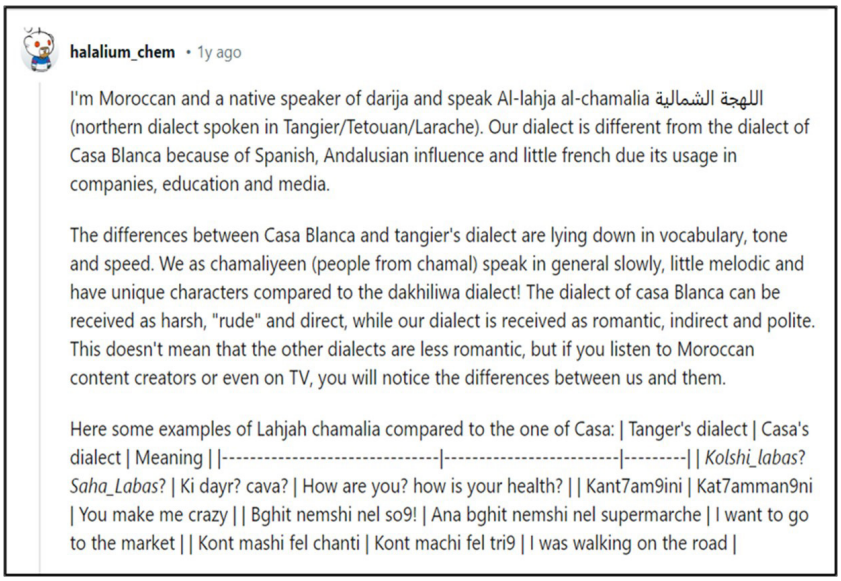
Figure 5. *WESH*-meme in vernacular French[6]

A significant amount of "conventionalised heteroglossia" related to "Digital Darija" – i.e., Darija shaped by interaction on digital platforms – can be observed in metalinguistic interactions, where speakers engage in reflective, metacognitive awareness of linguistic structures and variation. These instances often highlight that competition exists not only between Darija and Tamazight but also among the numerous varieties of Darija. This rivalry is evident not only at the national level, as seen in the classic examples of Moroccan Darija versus Algerian Darija, but also at the local level, as demonstrated by the following example, which echoes certain well-known offline stereotypes, such as rural (e.g., Casablanca) versus urban (e.g., Tangier) varieties.

Ultimately, Darija cannot escape the influence of globalisation, which, under neo-capitalist pressure, imposes socio-cultural templates worldwide. These templates, shaped by Anglo-Saxon culture from the Global North and its various "Englishes", dictate how individuals – socially interactive and adapted to the digital age – should present themselves and behave. The subsequent excerpts in Figures 7 and 8 exemplify this phenomenon, which manifests not only in the visual composition and semiotic features of these digital posts but also in their discursive content. This is particularly apparent in the

[6] Source: https://x.com/(Accessed 2024-10-01).

stereotypes depicted in Figure 8, including those pertaining to gender (e.g., the theme of "girls and love") and ethnicity (e.g., the themes of "red flags" and the problematisation of Moroccan identity).



halalium_chem • 1y ago

I'm Moroccan and a native speaker of darija and speak Al-lahja al-chamalia اللهجة الشمالية (northern dialect spoken in Tangier/Tetouan/Larache). Our dialect is different from the dialect of Casa Blanca because of Spanish, Andalusian influence and little french due its usage in companies, education and media.

The differences between Casa Blanca and tangier's dialect are lying down in vocabulary, tone and speed. We as chamaliyeen (people from chamal) speak in general slowly, little melodic and have unique characters compared to the dakhiliwa dialect! The dialect of casa Blanca can be received as harsh, "rude" and direct, while our dialect is received as romantic, indirect and polite. This doesn't mean that the other dialects are less romantic, but if you listen to Moroccan content creators or even on TV, you will notice the differences between us and them.

Here some examples of Lahjah chamalia compared to the one of Casa: | Tanger's dialect | Casa's dialect | Meaning | |----------|----------|---------| | *Kolshi_labas? Saha_Labas?* | Ki dayr? cava? | How are you? how is your health? | | Kant7am9ini | Kat7amman9ni | You make me crazy | | Bghit nemshi nel so9! | Ana bghit nemshi nel supermarche | I want to go to the market | | Kont mashi fel chanti | Kont machi fel tri9 | I was walking on the road |

Figure 6. Metalinguistic excerpt from Reddit[7]

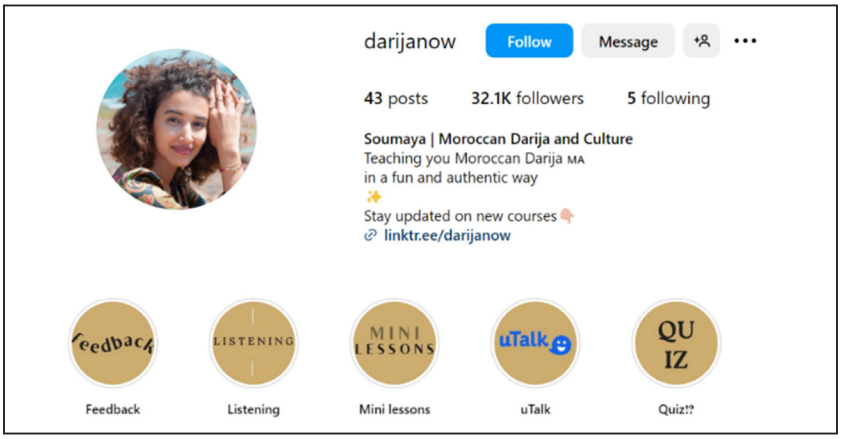


Figure 7. "Darijanow" by Soumaya on Instagram[8]

[7] Source: https://www.reddit.com/r/learn_arabic/comments/zed1gu/moroccan_darija_some_resources_for_beginners/ (Accessed 2024-10-01).

[8] Source: https://www.instagram.com/darijanow/ (Accessed 2025-06-05).

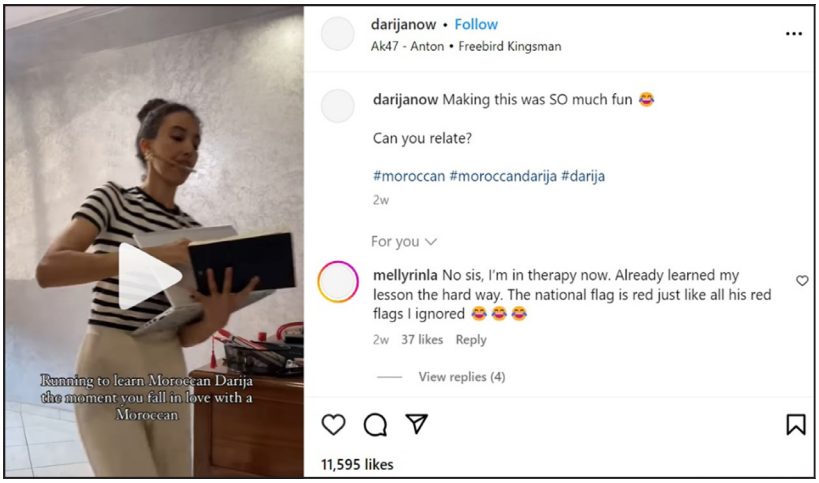


Figure 8. Comment & video excerpts of "darijanow" on Instagram[9]

6. Conclusion

Darija, particularly what is referred to here as "Digital Darija", exemplifies the concept of "vernacular globalisation" (Appadurai 1996: 10), enabling the reframing of dominant languages and shifting them away from an asymmetric system where they hold normative positions. In this multilingual, digital, and globalised context, Darija heavily relies on "conventionalised heteroglossia", utilising intertextuality extensively – not only in typical examples of memes (as seen in Figures 2, 3, and 5) but also in much of the content shared on platforms like Instagram, including the instances presented in Figures 7 and 8.

As a lingua franca, Darija finds itself in fierce competition with Tamazight, rapidly usurping its position. Moreover, as an unintended consequence of Institutional Arabisation, it is encroaching upon and replacing the functions and linguistic practices not only of Tamazight but also of traditionally dominant languages, such as Standard Arabic and French. This shift paves the way for a new linguistic landscape characterised by "Darijation". This grassroots phenomenon, supported by various online networks dedicated to preserving and developing Darija and its cultural expressions, poses a significant challenge to North Africa's sociopolitical policies, which continue to promote Standard Arabic as a linguistic and cultural icon despite its overall shortcomings.

Unrestrained neo-capitalism (exemplified in excerpts 1, 7, and 8) lies at the core of the instrumentalisation of politically non-dominant languages, such as Darija and Tamazight, as demonstrated in this study and supported by Lafkioui (2024). However, dominant interest groups have overlooked the impact and breadth of this instrumentalisation, particularly in the case of Darija, within the context of globalisation and digitalisation. This

[9] Source: https://www.instagram.com/p/C5V4KFxNdLi/ (Accessed 2025-06-05).

has influenced current policymaking, especially in education, a key area of contention. Certain stakeholders now advocate for the officialisation and integration of Darija into national education – a debate that gained momentum following the official recognition of Tamazight in Morocco in 2011.

One potential explanation for Darija's transnational success, reflected in the growing phenomenon of "Darijation", lies in its "ethnographically integrative" nature (Mufwene 2004: 206), which sets it apart from other lingua francas such as French or Standard Arabic. Unlike these more "segregative" languages (Mufwene 2004: 206), which primarily serve the elite and exclude the majority of North African populations, Darija facilitates socioeconomic integration and upward mobility for the broader populace by functioning as an endogenised lingua franca. Her success, however, comes at a cost, borne by her older endogenous sister, Tamazight. As Darija continues to rise in prominence, Tamazight faces the challenge of maintaining its sociolinguistic space, often overshadowed by the growing influence of Darija in both local and transnational contexts.

References

Aguadé, Jordi. 2006. Writing dialect in Morocco. *Estudios de dialectología norteafricana y andalusí* 10. 253-274.

Appadurai, Arjun. 1996. *Modernity at large: Cultural dimensions of globalization*. Minneapolis: University of Minnesota Press.

Blommaert, Jan. 2010. *The sociolinguistics of globalization*. Cambridge: Cambridge University Press.

Bourdieu, Pierre. 1982. *Ce que parler veut dire*. Paris: Fayard.

Bouziane, Abdelmajid & Saoudi, Mohamed. 2021. The status of English in Morocco: Lessons from spontaneous debates in social media. *English Studies at NBU* 7(2). 187-208.

Caubet, Dominique. 2017. Morocco: An informal passage to literacy in Dārija (Moroccan Arabic). In Høigilt, Jacob & Mejdell, Gunvor (eds.), *The politics of written language in Arab world*, 116-141. Leiden: Brill.

Chekayri, Abdellah 2018. L'enseignement en arabe dialectal pour consolider l'acquisition de l'arabe standard In Chekayri, Abdellah & Mgharfaoui, Khalil (eds.), *Réflexions sur le lexique et l'enseignement de l'arabe marocain*, 13-51. Casablanca: Centre de la promotion du darija.

Durand, Olivier. 2004. *L'arabo del Marocco: Elementi di dialetto standard e mediano*. Roma: Università degli Studi La Sapienza.

Ennaji, Moha. 2001. De la diglossie à la quadriglossie. *Languages and Linguistics* 8. 49-64.

Fairclough, Norman. 1989. *Language and power*. London: Longman.

Gal, Susan. 2006. Minorities, migration and multilingualism: Language ideologies in Europe. In Mar-Molinaro, Clare & Stevenson, Patrick (eds.), *Language ideologies, practices and polices: Language and the future of Europe*, 13-27. London: Palgrave.

Goffman, Erving. 1981. *Forms of talk*. Philadelphia: University of Pennsylvania.

Gumperz, John J. 1982. *Discourse strategies*. Cambridge: Cambridge University Press.

Hoogland, Jan. 2014. Toward a standardized orthography of Moroccan Arabic based on best practices and common ground among a selection of authors. In Santillán, Paula & Pérez, Luis Miguel & Moscoso, Francisco (eds.), *El árabe marroquí: De la oralidad a la escritura*, 59-76. Ciudad Real: Publicaciones de la Universidad de Castilla la Mancha.

Lafkioui, Mena B. 1998. Les Berbères et leur langue: Le cas des immigrés berbères en Belgique. In Canut, Cécile (ed.), *Attitudes, représentations et imaginaires en Afrique*, 119-130. Paris: L'Harmattan/Langues O'.

Lafkioui, Mena B. 2006. Pratiques et représentations linguistiques en contexte multilingue : Le cas des Berbères en Belgique. *Quaderni del Dipartimento di Linguistica (Unical)* 24. 73-84.

Lafkioui, Mena B. 2008. Identity construction through bilingual Amazigh-Dutch "digital" discourse. In Lafkioui, Mena & Brugnatelli, Vermondo (eds.), *Berber in contact: Linguistic and sociolinguistic perspectives*, 217–231. Köln: Köppe Verlag.

Lafkioui, Mena B. 2013. Multilingualism, Multimodality, and Identity Construction on French-Based Amazigh (Berber) Websites. *Revue Française de Linguistique Appliquée* XVIII (2). 135–151.

Lafkioui, Mena B. 2018. *Berber languages and linguistics*. Oxford: Oxford Bibliographies.

Lafkioui, Mena B. 2019. Le français face à la « super-diversité » dans la ville métropole de Gand. In Gadet, Françoise (ed.), *Les Métropoles francophones européennes*. 185-202. Paris: Garnier.

Lafkioui, Mena B. 2021. Codeswitching and artistic performance among multilingual minorities in Flanders. *Belgian Journal of Linguistics* 35. 34–50.

Lafkioui, Mena B. 2024. Pluricentricity, iconisation, and instrumentalisation of language in North Africa and its diaspora. In Huber, Maté & Meisnitzer, Benjamin (eds.), *Pluricentric languages in Africa. Multilingualism and Linguistic Dehegemonisation in Africa and Around the World*, 15-38. Graz: PCL-Press.

Michalski, Marcin. 2019. *Written Moroccan Arabic: A study of qualitative variational heterography*. Poznań: Adam Mickiewicz University Press.

al-Midlāwī, Muḥammad. 2019. *Al-ʿarabiyya ad-dāriğa: Imlāʾiyya wa naḥw: Al-aṣwāt, aṣ-ṣarf, at-tarkīb, al-muʿğam* (Darija Arabic: Spelling and grammar: Sounds, conjugation, structure, vocabulary). Markaz Tanmiyat al-Dāriğa. Zakoura. Ad-Dār al-Bayḍāʾ.

Miller, Catherine. 2017. Contemporary dārija writings in Morocco: Ideology and practices. In Høigilt, Jacor & Mejdell, Gunvor (eds.), *The politics of written language in the Arab world*, 90-115. Leiden: Brill.

Moscoso García, Francisco. 2009. Comunidad lingüística marroquí en los foros y chats. Expresión escrita, ¿norma o anarquía? *Al-Andalus Magreb* 16. 209-226.

Moustaoui Srhir, Adil. 2016. *Sociolinguistics of Moroccan Arabic*. Main Frankfurt & Berlin: Peter Lang.

Mufwene, Salikoko S. 2004. *The Ecology of language evolution*. Cambridge: Cambridge University Press.

Sedrati, Anass & Ait Ali, Abderrahman. 2019. *Moroccan Darija in online creation communities: Example of Wikipedia.* Al-Andalus Magreb: Estudios árabes e islámicos 26. 1-14.

Youssi, Abderrahim. 1995. The Moroccan triglossia: Facts and implications. *International Journal of the Sociology of Language* 112. 29-43.

DOI: 10.14746/linpo.2025.67.1.4

Standardizing Darija: Collaborative approaches in the Moroccan Darija Wikipedia

Anass Sedrati[1] & Mounir Afifi[2] & Reda Benkhadra[3]

[1]Wikimedia Morocco – KTH, Royal Institute of Technology, Stockholm
anass@kth.se | ORCID: 0000-0003-2763-572X
[2, 3]Wikimedia Morocco
prebirthtime@gmail.com | ORCID: 0009-0004-6859-1609
m.benkhadra@aui.ma | ORCID: 0009-0001-2707-0063

Abstract: This paper examines the development of Moroccan Darija Wikipedia since its launch in July 2020. It details the strategies employed by the Wikimedia Morocco user group, focusing on bot automation and editing contests, to foster growth within this low-resource language Wikipedia. The paper highlights the opportunities Darija Wikipedia presents for Artificial Intelligence research, particularly in Natural Language Processing, given its status as the largest online Darija dataset. It also explores how the standardization efforts undertaken by the user group enable valuable collaboration between volunteers, experts, and researchers, potentially setting a precedent for other similar language communities. Furthermore, the paper addresses key challenges, including ensuring community sustainability and mitigating vandalism, and analyzes the manifestation of diverse spelling conventions (phonetic, etymological) within the encyclopedia's content.

Keywords: Artificial Intelligence, Darija, growth, Morocco, standardization, Wikipedia

Introduction

As similar as they might seem, each language version of Wikipedia has its own background, rules, and community, which affect its structure and functions. While some Wikipedias were created by online editors who do not know each other, others emerged as a consequence of structured work planned offline.

Darija refers to various forms of dialectal Arabic used in Morocco that share common features. As its first speakers were Arabized Berbers, its pronunciation is substantially diffe-

rent from the Middle Eastern Arabic vernaculars (Heath 1997: 206). Darija is in a diglossic relationship with Standard Arabic (Chtatou 1997: 101). Since it is neither codified nor standardized, Darija is considered to have a lower status than Fusha, or Standard variety of Arabic, which is used for religious and official matters (Ennaji et al. 2004: 1). Darija is also considered to be an oral language and is rarely used in written form due to the reasons mentioned earlier (as well as claims that it does not have a standard, that the writing is already done in Standard Arabic, and that there are considerable regional differences in Darija) (Miller 2017: 90).

Darija has 28 consonant phonemes and four vowel phonemes and is the dominant vernacular language in Morocco strongly influenced by different varieties of Arabic, Berber, French and Spanish (Mrini & Bond 2018: 1). Given the diverse origins of Darija, it is then not unusual that the same object can be referred to in different words, depending on the speaker and his region of origin.

The current paper presents the state of art of the Darija Wikipedia (ary) in 2025, which now contains over 10,500 encyclopedic articles. It is structured as follows: Section 2 provides a short background to introduce Wikipedia in general, its vision, and the process that needs to be followed to create a new language version, in addition to a literature review. Section 3 then dives deeper into Darija Wikipedia, detailing its governance and community processes, before introducing in section 4 how current editors participate in standardization efforts and policy creation. In the next section (5), an overview of technical tools used in this Wiki are presented to the reader. These include bots, interface translation and namespaces. Following that, several strategies used by Wikimedia Morocco to encourage editing Darija Wikipedia, are presented in section 6. Section 7 provides a high-level description of challenges still to be addressed, either in terms of processes, of community sustainability or vandalism. The latter aspect is further analyzed in section 8, where statistics about vandalism and spelling tendencies in Darija Wikipedia are shared, together with an analysis of the findings. Finally, section 9 presents opportunities to be explored for this young Wiki, which can be investigated in future work, before ideas for next steps conclude the paper.

This research includes three supporting appendices. Appendix 1 presents statistics on articles about males and females in selected Wikipedia versions, providing additional context regarding gender representation in the compared languages. Appendix 2 documents the distribution of letters used in ary Wikipedia, which is relevant to our linguistic analysis, as well as their chosen Latin transcriptions throughout the paper. Finally, Appendix 3 lists the 100 most frequently used words and their spelling forms, as introduced by various editors, offering insights into common vocabulary patterns. These appendices are included to provide detailed supplementary information that may be of interest to readers concerned with the full methodological aspects of the research.

1. Background

Wikipedia is an online written encyclopedia and is considered to be the largest in the world in terms of reading, traffic, and content volume (The Economist 2021), with over 64

million articles in 341 languages (Meta Wikimedia 2025). Open and free to edit, it allows anyone to edit or create content respecting its five pillars (English Wikipedia 2025a) and using reliable sources.

Wikipedia is written and maintained by a community of volunteers known as Wikipedians. Each language version of Wikipedia has its own volunteers who gather in a "language community" (Massa & Scrinzi 2011: 213). Any interested person can freely join any language community of their choice. The Wikipedia model is fully decentralized, even in times of growth (Forte et al. 2009: 65). It is the community that manages the content of Wikipedia, although the Wikimedia Foundation has the legal responsibility related to the hosting of the website without interfering with its content (Wikimedia Foundation 2025).

Founded in 2001 by Jimmy Wales and Larry Sanger, Wikipedia began in English and expanded rapidly. At that time, there were no processes in place for having a new language in Wikipedia, and all requests were handled on an informal basis. Soon, other Wikimedia projects saw the light, such as Wiktionary, Wikinews, Wikivoyage, among others, many of which likewise can be available in several languages.

On June 2, 2006, the Wikimedia Incubator was founded. This project, hosted by the Wikimedia Foundation, formalized processes for new Wikipedia language editions. It serves as "a platform where anyone can build up a community in a certain language edition of a Wikimedia project that does not yet have its own subdomain" (Wikimedia Incubator 2007).

The same year (2006), the Wikimedia Foundation Language Committee was created. Its role is to make decisions on requests for new languages that are currently in the Incubator. For a language to be eligible for a full Wikipedia version, the Language Committee has established a set of criteria. These include having a valid ISO 639 code, being "sufficiently unique", and having "sufficient number of fluent users" (Meta Wikimedia 2007).

The first request for a Moroccan Darija Wikipedia was made in January 2008[1], following which an Incubator test page was opened for the project[2]. After several years of relatively low and scattered activities in the page, along with a number of challenges (Sedrati & Ait Ali 2019: 8-11), Wikimedia MA User Group (created in 2015) took the responsibility of activating the project, with the goal of launching a Darija Wikipedia.

On July 20, 2020, the Wikimedia Language Committee approved the request of having a Wikipedia in Moroccan Darija, which now has its own domain (ary.wikipedia.org). This Wikipedia version was launched with the support of Wikimedia Morocco User Group[3], which took the responsibility of taking it outside of the Wikimedia Incubator[4], in a joint effort with interested online users and Darija enthusiasts. The aim of this initiative was to enable Darija speakers to have their own version of Wikipedia, to be able to produce and read knowledge in their native tongue.

[1] Requests for new languages/Wikipedia Moroccan – January 2008 – https://w.wiki/CM6M.

[2] Darija Wikipedia Incubator Project – Archived in July 2020 – https://w.wiki/CM6R.

[3] https://w.wiki/8HA.

[4] https://w.wiki/d6i.

As of April 2025, the Darija Wikipedia edition has over 10,500 articles, 4 human administrators, and an average of over 250 000 page views per month by human users (Wikimedia Statistics 2025).

2.1. Literature review

The Moroccan Darija Wikipedia has been the subject of several scientific studies, as well as less formal comparisons with other Wikipedias. For example, three publications by Alshahrani et. al. (2022, 2023, 2024), which, although mainly focused on Egyptian Arabic Wikipedia, compared the Egyptian Arabic Wikipedia to the Moroccan Darija and the Standard Arabic Wikipedias in terms of the quality of their content for Natural Language Processing (NLP) applications and training Large Language Models (LLMs). These studies investigated the impact of the usage of automated articles, the cultural and linguistic representativity of the content, as well as its quantity, compared with the overall quality of Arabic Wikipedia editions for the aforementioned applications. For the Moroccan Darija Wikipedia, this study revealed that, although its text corpus is small in comparison to the Egyptian Wikipedia, it shows a similar pattern of content type distribution, editor types, and cultural representativity of content to the Standard Arabic and the English Wikipedias, while the Egyptian Wikipedia does not show such patterns. The small size of the Darija Wikipedia, however, makes its usability for NLP applications and training LLMs quite limited.

Further, Alshahrani et al. (2024: 9) found that Moroccan Darija Wikipedia displays more lexical richness and diversity than both the Standard Arabic and the Egyptian Arabic Wikipedias based on the 'Measure of Textual Lexical Diversity', introduced by McCarthy and Jarvis (2010: 381). Additionally, a recent informal statistical study of African language Wikipedias found that Moroccan Darija Wikipedia exhibits the highest editing depth[5] among all of them, with a value of 190 in October 2024 (Gilfillan 2024).

With this paper we take one more step and delve into those aspects of the Moroccan Darija Wikipedia that were not studied before, such as administrative, linguistic and community-related ones, highlighting potential limitations, pitfalls and opportunities for further studies and collaborations.

3. Governance and community

Moroccan Darija Wikipedia is managed by volunteer editors – كتاتبيا (*ktātbiyā)*,[6] who contribute to content creation, policy formulation, and page maintenance. There are two main types of editors, with different access levels: Anonymous users (IP users) and registered users. In addition to these human contributors, there is also another type of editors who

[5] Wikipedia Article Depth - https://w.wiki/H7Z

[6] In this paper, we are using a modified British Standard (BS 4280) as a transcription system, with some adaptations, namely: ة => a or t (instead of h or t), أ => ʾa, گ => g, *kasra* => i, *fatḥa* => a, *ḍamma* => u, *šadda* (doubling the letter), *schwa* (e). The full transcription system can be found in Appendix 2.

perform automated tasks and edits, called bots – بوتات (*būtāt*). They will be discussed further in section 5.

IP users – خدايميا ب أيبي (*ḫdāymiyā b-ʾāypī*) can edit and create most articles, participate in discussions, and preview edits to minimize errors without having an account on Wikimedia projects. They can also participate in discussions, either related to policies or specifically related to an article, but they cannot take part in community votes. However, they cannot vote or edit protected pages.

Registered users – خدايميا مقيدين (*ḫdāymiyā mqiyydīn*) are logged in with their Wikimedia accounts. They have additional privileges, such as maintaining watchlists, personal pages, uploading media, and seeking adminship status. Administrators – إمغارن (*imgāren*)[7] have renewable mandates. There are other higher-access roles, such as bureaucrats, stewards, and check users, but they are not tied to a specific Wikipedia project, and as of now, no editor on Moroccan Darija Wikipedia has them.

All users from the Darija Wikipedia *community* – جماعة *(ǧmāʿa)* collaborate to enrich the content and help advance the standardization of the language. They write rules empirically, allowing flexibility in early stages while mandating adherence to agreed-upon rules for uniformity. Spelling and grammar standards are discussed on ميزان لكلام (*mīzān le-klām)* and formalized in كناش لقواعد (*kennāš le-qwāʿd*), with rules categorized as *directive* – توجيهية (*tūǧīhiya*) or *imperative* – إلزامية (*ilzāmiya*).

For new words, contributors can propose neologisms on a dedicated page – طلب كلمة ؤلا تعبير (*ṭalab kelma ulā taʿbīr*), drawing from various language sources like Arabic, French, English, Tamazight, or Darija. Accepted terms are recorded in كناش لكلمات الجداد (*kennāš l-kelmāt ǧ-ǧdād)* for future use. The Darija Wikipedia community operates on a consensus-based, bottom-up approach, with all users participating in discussions while administrators implement decisions.

4. Standardization plan

4.1. Writing system

The writing style of Darija, its writing rules, and its spelling represent a major challenge in Wikipedia, given that this language does not have a unified standard form. On the societal level, there have been several attempts to standardize the orthography of Darija (Srhir 2012: 61), but none of the developed orthographies is used universally, i.e., throughout the country. In the context of Wikipedia, the goal of writing is to convey information to the reader in the simplest way possible, and without confusion, which can sometimes be challenging as some Darija words can have multiple meanings, and several of them can have the same orthography in the writing system used in the text.

[7] Plural of أمغار (*ʾamgār*), which means in the Moroccan culture a tribal leader or chieftain. The word is of Amazigh origin (Šafīq 1999: 58).

Given the dialectal variations of Darija, and the influences of foreign languages such as Standard Arabic, French, and Spanish, we have many possibilities for how to choose a word that expresses one meaning, and how to write that word (Caubet 2018: 388). In this context, we have two main conflicting tendencies:

- Conservative or etymological writing – Writing that attempts to preserve the form and orthography of a word as it is in its original language. This applies especially to words that are originally from Standard Arabic. This form of writing is convenient for someone who is familiar and comfortable with the orthography of Modern Standard Arabic.
- Phonetic writing – Writing that attempts to write words as they are or could be pronounced by speakers. There are two ways to represent phonetic writing: diacritization, such as "الْمَغْرِبْ بْلَادْ جَاتْ فْشَمَالْ إِفْرِيقْيَا وْقْرِيبَة لْأُوروبَا، ضَايْرْ بِهَا الْبَحْرْ الشَّامِي والْمُحِيطْ الْأَطْلَسِي", or by marking vowels using *matres lectionis* (Michalski 2016: 392-393), as in "لمغريب بلاد جات ف شمال إفريقيا و قريبة ل ؤروپا، ضاير بيها لبحر شّامي و لموحيط لأطلاسي". Diacritization is not practical in Wikipedia, given the difficulty and time needed to vocalize long texts with a keyboard, so the second method is the one that is commonly used and will henceforth be referred to as "phonetic spelling" in this paper. This form of writing is convenient for someone who is not accustomed to Standard Arabic writing, for example an adult who did not receive a high level of education in Morocco, or a young child, or a Moroccan who grew up abroad.

The majority of Darija editors write in a syncretic system, although they often prefer one of the two spelling approaches. Syncretic or reconciliatory writing attempts to reconcile the two tendencies, benefitting from their advantages and minimizing their drawbacks, so that writing and reading texts in Darija is relatively easy for any reader of any level and so that there is less confusion.

For example, the sentence "شفنا مدافع الجيش ف المغرب"[8] (see Table 1 for transcription and translation) can be read in 4 different and semantically correct ways, depending how the words مدافع (*mdāfʿ* or *mūdāfiʿ*) and المغرب (*lmeġrib* or *lmūġreb*) are pronounced, resulting in 4 correct sentences with different pronunciations and meanings. A solution suggested to avoid this potential confusion is to use a phonetic spelling system to distinguish between words that traditionally have the same spelling forms, but different pronunciations and meanings (al-Midlāwī al-Mnabbhi 2019: 18-20). Table 1 below summarizes the possible transcriptions and translations of that sentence using phonetic spelling.

[8] Inspired by a similar example from al-Midlāwī al-Mnabbhi (2019: 19). In the Darija spellings in this table, we used the common definite form *al-* ال to not distract from the main point which is the variations of pronunciations and meanings ensued from words that have otherwise identical spellings in their Standard Arabic origin.

Table 1. Possible transcriptions and translations of شفنا مدافع الجيش ف المغرب

Darija (phonetic spelling)	**Transcription**	**English translation**
شفنا مدافع الجيش ف الموغرب.	*šfnā mdāfʿ l-ǧīš f l-mūġreb*	We saw the army cannons during sunset.
شفنا مودافيع الجيش ف الموغرب.	*šfnā mūdāfīʿ l-ǧīš f l-mūġreb*	We saw the Army[9] defender during sunset.
شفنا مودافيع الجيش ف المغريب.	*šfnā mūdāfīʿ l-ǧīš f l-meġrīb*	We saw the Army defender in Morocco.
شفنا مدافع الجيش ف المغريب.	*šfnā mdāfʿ l-ǧīš f l-meġrīb*	We saw the army cannons in Morocco.

On the other hand, applying a phonetic spelling system in some cases, especially for words that are less prone to confusion but also rich in vowels, can result in spelling forms that are lengthier and less recognizable by native speakers who are familiar with Standard Arabic. These spelling forms may tend to be mocked or rejected, as shown by comments on Darija Wikipedia, and on social media. Examples include: موجتاماع *(mūǧtāmāʿ* – etymology: مجتمع), and بارلامان (*bārlāmān* – etymology: برلمان).

Practical wisdom therefore dictates using the phonetic spelling system only when confusion of meaning is likely or demonstrably possible, through the existence of two or more commonly used words that share the same etymological spelling. This approach is perhaps more suited to the Darija Wikipedia, as it is an experimental approach, and is constantly evolving as the encyclopedia develops. This form of writing reduces the effort required for someone who is not used to writing Standard Arabic to understand what is written, and at the same time requires less adaptation effort than phonetic writing so that a person with a high level of command of Standard Arabic can understand and follow what is written. It remains an open question whether this approach will result in a consistent and viable spelling system of Moroccan Darija. Furthermore, since the Darija Wikipedia is still at its start, quantity, variety and understandability of the content have currently a higher priority compared with the consolidation of the spelling and grammar rules.

4.2. Phonology

The Darija Wikipedia project adopts Arabic script as a writing system to represent existing sounds. As for sounds that do not exist in Standard Arabic (like /g/) or that come from Romance languages (like /p/ and /v/), variant letters of the Arabic alphabet were used, similar to Persian and Urdu which rely on alphabet systems derived from the Arabic alphabet. The

[9] "The Army" (*l-ǧīš*) is a term commonly used to refer to AS FAR (Association sportive des Forces armées royales), a football club based in Rabat.

letter *bāʾ* with three dots at the bottom represents the sound /p/; the letter *fāʾ* with three dots above represents the sound /v/; and the letter *kāf* with three dots above represents the sound /g/. Table 2 below shows the glyph forms used for these sounds.

Table 2. Selected representations of /p/, /g/ and /v/ in Darija Wikipedia

/p/	**/g/**	**/v/**
پ	ڴ	ڤ

It is worth noting that the interdentals existing in the Arabic language – i.e. the letters ث (*ṯ*), ظ (*ẓ*), ذ (*ḏ*) – are not widely used in Darija, except in some dialects in North-Eastern Morocco (Behnstedt & Benabbou 2005: 17). In the scope of Moroccan Darija Wikipedia, most of the written text is in the so-called "Moroccan Koine", spoken in big cities and very present online. Thus, interdentals are often absent and are represented by the corresponding apico-alveolar consonants, respectively ت (*t*), ض (*ḍ*), د (*d*). There is however no restriction on using other varieties of Moroccan Darija on Wikipedia, so long as the text is understandable to everyone.

The project makes use of ء (*hamza*) placed over أ (*'alif*) to represent the sound sequence /ʾa/ or ؤ *(wāw*) for the sequence /ʾu/, or under إ (*'alif*) for /ʾi/.

For example, instead of writing أوروپا (*'ūrūppā*) or أوكرانيا (*'ūkrānyā*), the letter ؤ is used to represent the sound *ū* instead of أو which can be read as *aw*. The same applies to the sound sequence /'i/, the letter إ is used alone to represent this sound, instead of using the letter *yā'* as well as in Arabic to write إيطاليا (*'iṭālyā*) or إيبيريا (*'ībīryā*), for instance. Table 3 provides a summary of this aspect.

Table 3. Selected representations of the sequences /'a/, /'u/ and /'i/ in Darija Wikipedia

/'a/	**/'u/**	**/'i/**
أتاي *'atāy* 'tea'	ؤروپا *'urūppā* 'Europe'	إباون *'ibāwn* 'beans'

As for the *hamza* at the end of words of Arabic origins, it shall not be written if its pronunciation in Darija is common without *hamza*, for example: سما *smā* 'sky', فقها *feqhā* 'religious scholars', ما *mā* 'water'.

The *hamza* can be written in exceptional cases if the word is not used at all by speakers without a hamza, and it may not be understood or cause confusion if the *hamza* is not written, and it does not have an equivalent in Darija, such as فضاء *faḍā'* 'space, outer space'.

There are words that are acceptable in common usage, even though they have equivalents without a *hamza*, because this equivalent is not very common. In this case, both forms are acceptable. For example: جزء *ğuz'* 'part' and ڴزو *gzū* 'part' and their corresponding plural forms أجزاء *ʾağzā*' 'parts' and ڴزوّات *gzuwwāt* 'parts'.

When it comes to the use of ة (*tā marbūṭa*), its use is recommended due to the morphological roles that it plays, making it difficult to abandon. An example is when it shows that the word is feminine, or it distinguishes between the plural and the feminine in some cases (صيادة = fisherwoman, صيادا = fishermen) - both pronounced *ṣiyyāḍa*, or it is pronounced and/or becomes a ت (*tā mabsūṭa*), in case of a pronoun or genitive, as in:

سمية د لبلاصة	*smiyt l-blāṣā*	= سميت لبلاصة	*smiyya d l-blāṣa*	'the name of the location/place'
مدينتي	*mdīntī*	= مدينة ديالي	*mdīna dyālī*	'my city'

4.3. Verb conjugation

We have developed a simplified conjugation table (Table 4) designed to be easily comprehensible for the average user, avoiding unnecessary linguistic complexities. This conjugation system mainly relies on clustering and grouping verbs by their correspondence to a Wikimedia template (see templates in the Section 5.2) based on their conjugation patterns. Groups within the same verb cluster share the same method of conjugation in the perfect (past) form, and either display only minor differences in other patterns, or have the same dictionary form but differing conjugation patterns. Furthermore, verbs within the same group share conjugation patterns in the imperfect forms (present and future), in addition to the perfect form, as they belong to the same cluster. In other words, each group has a Wikimedia template, whereby inputting the root letters of a verb results in a pre-constructed table of conjugation for that verb as output. It does not follow from this that these clusters correspond to linguistically meaningful verb categories. These clusters are:

Cluster 1 – Regular verbs, without '*alif* in the last or penultimate position, are grouped into one group, which is group 1.

Cluster 2 – Verbs without '*alif* at the end, including verbs with 2 letters (هز *hezz* 'to carry', حط *ḥeṭṭ* 'to put down', دز *dezz* 'to shear') and verbs with 4 letters or more that have '*alif* in the penultimate position. There is no 3-letter verb in this cluster (given that the stress is ignored). This cluster contains two groups: Group 2 and Group 2*.

Cluster 3 – Verbs with '*alif* in the penultimate position (in the middle for verbs with 3 letters) and do not follow the rule of cluster 2 in the past form. This cluster has three groups 3, 4 and 5.

Cluster 4 – Verbs with '*alif* at the end, regardless of the number of letters. There are 4 groups here, group 6 to 9, but two of them are very rare (Group 6 and Group 9).

Table 4. Suggested conjugation table for Darija, clustering verbs in different groups

Cluster	Group	Description	Example verbs
Cluster 1	Group 1	Verbs without *ʻalif* at the end or penultimate, including verbs with 3, 4, 5 or 6 letters. E.g. أنا هربت *ʾanā hrebt* 'I escaped' أنا كانهرب *ʾanā kānhreb* 'I am escaping' أنا كركبت *ʾanā kerkebt* 'I rolled' أنا كانكركب *ʾanā kānkerkeb* 'I am rolling'	هرب *hreb* 'escape' قتل *qtel* 'kill' مثّل *mettel* 'act' نقّز *neqqez* 'jump' كركب *kerkeb* 'roll' ستعمل *ste ʿmel* 'use' تستعمل *teste ʿmel* 'be used' صاوب *ṣāwb* 'make' تصاوب*tṣāwb* 'be made'
Cluster 2	Group 2	Verbs with 2 letters without *ʻalif* at the end, others with 3 or 4 letters with *ʻalif* in the penultimate position, all ending with a *shadda*. E.g. أنا كبّيت *ʾanā kebbīt* 'I poured' أنا كانكبّ *ʾanā kānkebb* 'I am pouring' أنا قادّيت *ʾanā qāddīt* 'I adjusted' أنا كانقادّ *ʾanā kānqādd* 'I am adjusting'	كبّ *kebb* 'pour' هزّ *hezz* 'carry' قادّ *qādd* 'adjust' تقادّ *tqādd* 'be adjusted'
	Group 2*	Verbs with 4 letters with *ʻalif* in the penultimate position (which follow the verb template فْعَالْ). E.g. أنا تقاليت *ʾanā tqālit* 'I got heavy' أنا كانتقال *ʾanā kāntqāl* 'I am getting heavy' أنا عواجيت *ʾanā ʿwāǧīt* 'I got bent' أنا كانعواج *ʾanā kān ʿwāǧ* 'I am getting bent'	تقال *tqāl* 'get slow/heavy' عواج *ʿwāǧ* 'get bent' كبار *kbār* 'get big' صغار *ṣǧār* 'get small' سمان *sman* 'get fat' طوال *ṭwāl* 'get tall' قصار *qṣār* 'get short'
Cluster 3	Group 3	Verbs with 3 or 4 letters, with *ʻalif* in the penultimate position, keep *ʻalif* in the present tense, and is absent the past form of the 1st and 2nd person singular (not following the rule of group 2). E.g. أنا خفت *ʾanā ḫeft* 'I got afraid' أنا كانخاف *ʾanā kānḫāf* 'I am getting afraid' أنا بنت *ʾanā bent* 'I appeared' أنا كانبان *ʾanā kānbān* 'I am appearing'	خاف *ḫāf* 'be afraid' سال *sāl* 'owe' بان *bān* 'appear' تباع *tbāʿ* 'be sold' تدار *tdār* 'be done' تخاد *tḫād* 'be taken' تكال *tkāl* 'be eaten'
	Group 4	Verbs with 3 letters, with a *ʻalif* in the middle, becoming *wāw* in the present tense and is absent in the past form of the 1st and 2nd person singular. E.g. أنا گلت *ʾanā gelt* 'I said' أنا كانگول *ʾanā kāngūl* 'I am saying' أنا متّ *ʾanā mett* 'I died' أنا كانموت *ʾanā kānmūt* 'I am dying'	گال *gāl* 'say' مات *māt* 'die' فات *fāt* 'pass' بال *bāl* 'pee' كان *kān* 'be'

	Group 5	Verbs with 3 letters, with a *ʻalif* in the middle, becoming *ya'* in the present tense, and is absent in the past form of the 1st and 2nd person singular. E.g. أنا بعت *ʾanā beʿt* 'I sold' كانبيع *ʾanā kānbīʿ* 'I am selling' أنا درت *ʾanā dert* 'I did' أنا كاندير *ʾanā kāndīr* 'I am doing'	باع *bāʿ* 'sell' دار *dār* 'do' عاف *ʿāf* 'be disgusted' سال *sāl* 'flow'
Cluster 4	Group 6	Verbs with 3 letters with *ʻalif* at the end, whose position changes to the beginning of the verb in the present tense. E.g. أنا كليت *ʾanā klīt* 'I ate' أنا كاناكل *ʾanā kānākel* 'I am eating' أنا خديت *ʾanā ẖdīt* 'I took' أنا كاناخد *ʾanā kānāẖed* 'I am taking'	كلا *klā* 'eat' خدا *ẖdā* 'take'
	Group 7	Verbs with *ʻalif* at the end, changing into *ya'* in the present tense. E.g. أنا عطيت *ʾanā ʿṭīt* 'I gave' أنا كانعطي *ʾanā kānʿṭī* 'I am giving' أنا جيت *ʾanā ǧīt* 'I came' أنا كانجي *ʾanā kānǧī* 'I am coming'	جا *ǧā* 'come' دا *ddā* 'get' شرا *šrā* 'buy' عطا *ʿṭā* ''give' ميزا *mīza* bet' أنسطالا *ʾanstalā* 'install' كومونيكا *kūmūnīkā* 'communicate'
	Group 8	Verbs with *ʻalif* at the end, keeping *ʻalif* in the present tense. E.g. أنا شقيت *ʾanā šqīt* 'I toiled' أنا كانشقا *ʾanā kānšqā* 'I am toiling' أنا بريت *ʾanā brīt* 'I healed' أنا كانبرا *ʾanā kānbrā* 'I am healing'	بقا *bqā* remain شقا *šqā* toil برا *brā* heal لقا *lqā* find تلاقا *tlāqā* meet تشرا *tešrā* be bought تسطّا *tseṭṭā* be crazy
	Group 9	Verbs with *ʻalif* at the end, changing into *wāw* in the present tense. E.g. أنا عفيت *ʾanā ʿfīt* 'I forgave' أنا كانعفو *ʾanā kānʿfū* 'I am forgiving' أنا حبيت *ʾanā ḥbīt* 'I crawled' أنا كانحبو *ʾanā kānḥbū* 'I am crawling'	عفا *ʿfā* forgive حبا *ḥbā* crawl

In determining the letter count of verbs, we disregarded the *shadda* (doubling), where a letter beneath it is considered as one letter, not two.

For verbs starting or ending with a *tā*' or a *nūn*, no distinct linguistic or template rule exists, but due to assimilation, these verbs interact with prefixes or suffixes such as:

- نقّز *neqqez* 'to jump' + كان *kān-* => كانّقّز *kānneqqez* 'I jump'
- تلاقا *tlāqā* 'to meet' + كات *kāt-* => كاتاكل *kāttlāqā* 'you meet'
- فات *fāt* 'to pass' + تي *-tī* => فتّي *fettī* 'you passed'
- بان *bān* 'to appear' + نا *-nā* => بنّا *bennā* 'we appeared'

Consequently, the *tā* or *nūn* must assimilate with a similar letter. For instance, for the verb كان *kān* 'to be', it is more correct to write كنّا *kunnā* 'we were' rather than كننا.

4.4. Pronouns

The pronouns are used to designate someone or something. We distinguish four different types: personal, possessive, objective and demonstrative pronouns.

Table 5. Personal pronouns

هي *hiyya* 'she'	هو *huwwa* 'he'	نتي *ntī* 'you (f. sing)	نتا *ntā* 'you (m. sing)'	أنا *ʾanā* 'I'
هوما *hūmā* 'they'		نتوما *ntūmā* 'you (pl)'		حنا *ḥnā* 'we'

Table 6. Possessive prepositional phrases

ديالها *dyālhā* 'hers'	ديالو *dyālū* 'his'	ديالك *dyālek* 'yours (sing)'	ديالي *dyālī* 'mine'
ديالهم *dyālhum* 'theirs'		ديالكم *dyālkum* 'yours (pl)'	ديالنا *dyālnā* 'ours'

When pronouns are linked to feminine nouns, the *tā marbūṭa* becomes *tā mabsūṭa* and is pronounced with the following suffixes (see Table 7). In general, *tā marbūṭa* in a noun is pronounced when the noun is the first (or not last) in a construct state.

Table 7. Possessive pronouns

ها- *-hā* 'her'	و- *-ū* 'his'	ك- *-k* 'your (sing)'	ي- *-ī* 'my'
هم- *-hum* 'their'		كم- *-kum* 'your (pl)'	نا- *-nā* 'our'

Table 8. Objective pronouns

ها- *-hā* 'her'	ه- *-h* 'him'	ك- *-k* 'you (sing)'	ني- *-nī* 'me'
هم- *-hum* 'them'		كم- *-kum* 'you (pl)'	نا- *-nā* 'us'

Table 9. Demonstrative pronouns

هادو (*hādū*) 'these'	هادي (*hādī*) 'this' (f. sing)	هادا (*hādā*) 'this' (m sing)	هاد (*hād*) 'this'
هادوك(*hādūk*) 'those'	هاديك (*hādīk*) 'that' (f sing)	هاداك (*hādāk*) 'that' (m sing)	

4.5. Annexation particles

Table 10. Annexation particles in Darija

متاع *mtāʿ*	نتاع *ntāʿ*	تاع *tāʿ*	د *d*	ديال *dyāl*

The word د *d* is the contraction of [10]ديال *dyāl*, while نتاع *ntāʿ* and متاع *mtāʿ* are regional variations of تاع *tāʿ*. All these words are considered synonymous (meaning "of" or used to express possessiveness) and should be written separately from the next word.

4.6. Prepositions

Prepositions establish relationships between nouns or pronouns and other words in a sentence, indicating direction, time, place, and spatial relationships. Each preposition may be used in different contexts. In Wikipedia, Darija prepositions, presented in Table 11 are written separately from the words following them, similar to connectors[11].

Table 11. Prepositions

تال/حتال *tāl/ḥtāl* 'until'	بحال/فحال *bḥāl/fḥāl* 'like'	معا *m ʿā* 'with'	ل *l-* 'to'	من *men* 'from'
كي *kī* 'like'	كيفما/كيما *kīfmā/kīmā* 'like'	على/عل *ʿlā/ʿl* 'on, about'	ف/في *f/fī* 'in'	ب *b* 'with'

5. Technical aspects

All Wikimedia projects run on a free and open license software called MediaWiki[12], used also by tens of thousands of websites, and thousands of organizations and companies (Barrett 2008: 4). To offload the processing power from Wikimedia servers, scripts with special privileges, called the bot flag, are run by users on their local computers, or on Wikimedia servers dedicated to such tools (such as Toolforge). Bots edit Wikimedia pages as if they were human editors.

[10] There are several hypotheses for the origin and etymology of this word. It might originate from the Andalusi dialect brought by the Moriscos, being formed by the fusion of the Latin word *di* or *de* with the Arabic definite article (ال), similarly as del in Spanish (Ouhalla 2015). Heath (2015 & 2020, p. 218) considers ديال (*dyāl*) a back-formation from ديالو (*dyālū* 'his') and ديالها (*dyālhā*, *dyālā* 'hers'), which he in turn derives from Vulgar Latin **di ellu* and **di ella*, from Classical *de* + *illum*, *illa*.

[11] The prepositions listed can have other meanings or usages, depending on context.

[12] https://w.wiki/VtJ.

5.1. Bots

A bot is a script that automates repetitive and time-consuming tasks on the Internet. It is used to automate routine Wikipedia tasks (MediaWiki 2010), allowing human editors to focus on complex content creation activities. Darija Wikipedia uses bots to support, *not replace*, editors, preserving cultural authenticity (تامغرابيت *tāmġrābīt*). The bot policy limits bot-created articles to 30% of total content, which are tracked for human oversight. This ensures quality and local cultural relevance.[13] [14]

There are currently 8 bots on Darija Wikipedia:

- ***Menobot*** – The first bot approved by the community, working on format and technical adjustments, such as removing extra spaces.
- ***MediaWiki default*** and ***MediaWiki message delivery*** – Used by the Wikimedia Foundation and affiliates to write announcements on talk pages, or on the community page – ساحة د الجماعة (*sāḥa d jjmāʿa*).
- ***DarijaBot*** – Handles tasks such as creating articles, managing categories and templates, generating statistics, and maintaining pages. So far it has created over 3,500 articles.[15] [16]
- ***PGVBot*** – Standardizes Darija characters for the letters P, G, and V by replacing various Unicode alternatives with the community-approved defaults and redirecting to them.
- ***Sa7bot*** – Fact-checking and spelling correction bot, correcting mainly dates of birth and death, and other factoids.
- ***InternetArchiveBot*** – Maintains web references used in articles, archives urls, and maintains reference tags.
- ***AmgharBot*** – A bot with administrator privilege. It can perform administrative tasks (such as protecting and deleting pages).

5.2. Namespaces

Wikipedia content is divided into namespaces – مجالات سمياتية (*majālāt smiyātiya*), each serving a specific content type and handled differently by the MediaWiki software. On September 25, 2021, the Darija Wikipedia community renamed many namespaces from the Standard Arabic defaults and introduced two new namespaces along with their corresponding talk pages (Darija Wikipedia 2025a)[17]. Table 12 presents the main existing namespaces in ary Wikipedia.

[13] Bot Policy – Discussion Page – Moroccan Darija Wikipedia - https://w.wiki/AiZW.

[14] Content Policy – Mass Content – Moroccan Darija Wikipedia - https://w.wiki/AiZZ.

[15] User "DarijaBot" contributions – Moroccan Darija Wikipedia – https://w.wiki/AiZi.

[16] List of articles created by DarijaBot – https://w.wiki/Bhgx.

[17] The full description of namespaces in Darija Wikipedia can be found here: https://w.wiki/CM5W.

Table 12. Darija Wikipedia main namespaces

Darija Wikipedia namespace	English Wikipedia equivalent
رئيسي (مقالات)	Main (articles)
تصنيف *teṣnīf*	Category
موضيل *mūḍīl*	Template
مودول *mūdūl*	Module
واساخ *wāsāḫ*	Draft
ويكيپيديا *wīkīpīdyā*	Wikipedia
معاونة *mʿāwna*	Help
خدايمي *ḫdāymī*	User
فيشي *fīšī*	File
قيسارية *qīsāriya*	Portal
ميدياويكي *mīdyāwīkī*	MediaWiki
خاص *ḫāṣ*	Special
ميديا *mīdyā*	Media
سوتيتر *sūtītr*	TimedText
مجالات سمياتية ديال لمداكرة (بحال: مداكرة، مداكرة د ويكيپيديا، ...) *majālāt smiyātiya dyāl l-mdākra* *(bḥāl: mdākra, mdākra d wīkīpīdyā, ...)*	Talk namespaces (e.g. Talk, Wikipedia Talk, etc)

6. Strategies and activities

6.1. *Ḥerkat* or editing campaigns

The community uses editing campaigns, locally known as حركات (*ḥerkāt*),[18] to develop content systematically around selected themes. Initially, these campaigns aimed to create a snowball effect by exploring a central theme and its related topics but faced challenges with diluted focus, such as straying from historical monuments to unrelated subjects like movies. To address this, the campaigns adopted a list-based approach, prioritizing specific

[18] Plural of حَرْكة (*ḥerka*), which means in the history of Morocco a military campaign led by the Sultan or other State notables for political, military, or financial purposes.

articles for development. Themes are chosen based on context, current events (e.g., football tournaments), or comprehensive subjects like Morocco's territorial organization, covering topics such as communes (municipalities), douars (villages), and national team players in FIFA World Cups.

Campaigns also consider reader trends, like increased interest in Moroccan scientist Kamal Oudrhiri during September–November due to his inclusion in school textbooks (see figure 1). Consequently, the community has decided to extract the topics mentioned in these textbooks, such as Moroccan dynasties, biographies, and geographical locations, to enhance the articles surrounding them[19].

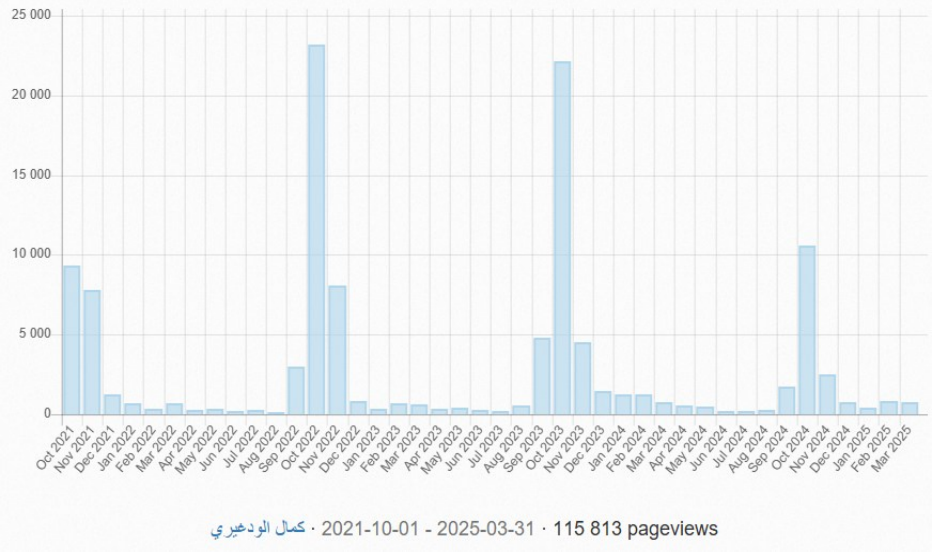


Figure 1: Number of pageviews for Kamal Oudrhiri's Moroccan Darija Wikipedia page

6.2. Contests

As part of the effort to extend the topic coverage in Wikipedia, editing contests with prizes are organized. Contests have proven to be powerful tools for adding new content to Wikipedia. On Moroccan Darija Wikipedia, several contests have been organized so far:

- **WikiForHumanRights 2021**[20] featured Darija as one of the 4 languages of the contest and resulted in 18 new articles written by contestants[21] (Wikimedia Morocco 2021).

[19] More details can be found here: https://w.wiki/Bj8m.

[20] WikiForHumanRights 2021 in Morocco – https://w.wiki/37Xd.

[21] How can editing contests support smaller Wikipedias? (Arctic Knot Conference 2021) – https://youtu.be/SxcTLnMCwkA.

- **Wikimedia Morocco contest 2023**,[22] which ran over 45 days and resulted in 34 new and 19 improved articles, with 7 participants (Wikimedia Morocco 2023).
- **Wikipedia Darija Birthday Contest July 2024**,[23] where 19 editors participated, resulting in 16 new and 39 improved articles, over 8 days[24].
- **Wikipedia Darija Contest August 2024**,[25] which stretched over 2 weeks, had 52 participants and resulted in 52 new and 82 edited articles[26].

6.3. Outreach

Raising awareness about the Darija Wikipedia is manifested in several ways. Besides a small Facebook page[27] dedicated to this Wiki (with ca. 900 followers), there is also a podcast[28], where longer articles from the encyclopedia are read and recorded. This recording tradition comes from the fact that Darija itself is considered to be more oral. Since Wikipedia is a written encyclopedia, editors are obliged to create "written" articles. To combine both approaches, the community works actively in providing the so-called spoken articles – مقالات مسموعة (*maqālāt mesmūʿa*) – as well. These are audio recordings where a volunteer reads the content of an article and uploads the audio file alongside the written version. As of April 2025, there were over 564 spoken articles in Moroccan Darija Wikipedia[29].

Both the podcasting and audio recording alternatives are also a direct answer to the argument stating that Darija is mainly an oral language that should not be written, as these tools provide oral encyclopedic content to any person wishing to listen to it.

7. Challenges

7.1. Implementation/Respect of existing processes (e.g. new words)

Despite the community's efforts to establish standards and processes for language development within the project, these guidelines are often little implemented, particularly by new editors who may be unaware of their existence. As a result, variations in writing styles persist within the Darija Wikipedia. In response to this challenge, bots (particularly DarijaBot and PGVBot) have been employed for spelling corrections.[30] Additionally, several resources

[22] Wikipedia Morocco Contest 2023 – Moroccan Darija Wikipedia – https://w.wiki/6cR8.

[23] Darija Wikipedia contest of July 2024 – https://w.wiki/AjGM.

[24] Dashboard of Darija Wikipedia contest of July 2024 – https://tinyurl.com/yc7rzv2j.

[25] Darija Wikipedia contest of August 2024 – https://w.wiki/AwEr.

[26] Dashboard of Darija Wikipedia contest of August 2024 – https://tinyurl.com/3hb9jthe.

[27] Moroccan Darija Facebook Page – https://www.facebook.com/wikipedia.darija.

[28] Wikipedia b Darija – Podcast on Spotify – https://open.spotify.com/show/7JiFdWCBz7BPA2KsZzEATu.

[29] List of spoken articles in Moroccan Darija (ie. articles having an audio recording) – https://w.wiki/9Xxm.

[30] Both bots operate through a deterministic system, using key-value data (either python dictionaries or json files), replacing the dictionary key (current value) with the dictionary value (target value). The keys and values

have been created to facilitate the introduction of new users into the project for a better understanding of its mode of operation and procedures. These resources include welcoming messages containing essential links, an FAQ page, and a contact page, all designed to enhance newcomers' understanding and engagement with the project.

7.2. Community sustainability

Like many small and relatively new communities, the Moroccan Darija Wikipedia relies on a limited number of volunteers, making its sustainability vulnerable to fluctuations in their availability. Since editing is unpaid and done in free time, activity levels directly impact the project. Contributors often leave due to burnout (Konieczny 2018) or shifts in motivation, such as seeking social status, impact, belonging, or skill development (Baytiyeh & Pfaffman 2010: 132). This pattern has led to the closure of several Wikimedia projects, including 13 Wikipedias, due to prolonged community inactivity.[31]

By April 2025, there were 4 human administrators in the Darija Wikipedia (Darija Wikipedia 2025b), and 12 active editors[32] in the Wiki, which are not alarming numbers. However, the community is aware of this strong dependency on a small number of people, therefore more efforts are expected in outreach, to retain new volunteers who can ensure the continuity of the project even if the current active community members move on to other tasks and interests in their lives.

7.3. Vandalism

English Wikipedia defines vandalism as "editing (or other behavior) deliberately intended to obstruct or defeat the project's purpose" (English Wikipedia 2025b). This is particularly relevant for Darija Wikipedia, because it does not only impact the content, but also language and spelling. It can sometimes be difficult to decide what counts as vandalism vs what is essentially a difference in opinion or mere misunderstanding, unless a clear behavior emerges, such as emptying a whole page.

Wikipedia encourages editors to assume good faith and provides technical tools to fight vandalism. For example, confirmed users can revert edits with a single click. Also, administrators have additional privileges: they can revert multiple edits, protect pages from unauthorized editing, and block users by time, page, or IP range.

In addition to these tools, strategies and community-based rules are under development to limit the impact of vandalism, such as daily patrols and discussing ambiguous situations case by case.

can take the form of a character, a word, a sentence or a regular expression (a sequence of characters that specifies a match pattern in text) – https://w.wiki/3jKQ

[31] Closed and read-only Wikis - https://w.wiki/BbrK

[32] An active editor in Darija Wikipedia is defined as an editor making at least 5 edits per month

8. Observations and findings

8.1. Vandalism statistics

To gain a general understanding of vandalism on Darija Wikipedia, statistics were collected on reverted disruptive edits and deleted pages created by both anonymous users (IPs) and registered users. They are presented in the figures below: Mapping disruptive edits made by anonymous users (Figure 2), disruptive edits made by all users, including the registered ones (Figure 3), devices used for disruptive edits (mobile vs others, Figure 4), comparison in number and size of vandalism (between anonymous and registered users, Figure 5), and the size of disruptive edits (Figure 6).

Note that: (1) The data was limited to the main namespace (articles), as a deeper analysis of vandalism and disruptive behavior falls outside the scope of this paper, (2) Defining vandalism can be subjective, as it depends on the editor's intent, which can only be inferred from behavior, not confirmed, and (3) The deletion logs can have missing data points, especially the page creator's usernames. The change type for reverted edits is also sometimes unknown, likely because some of these edits have been hidden by an administrator.

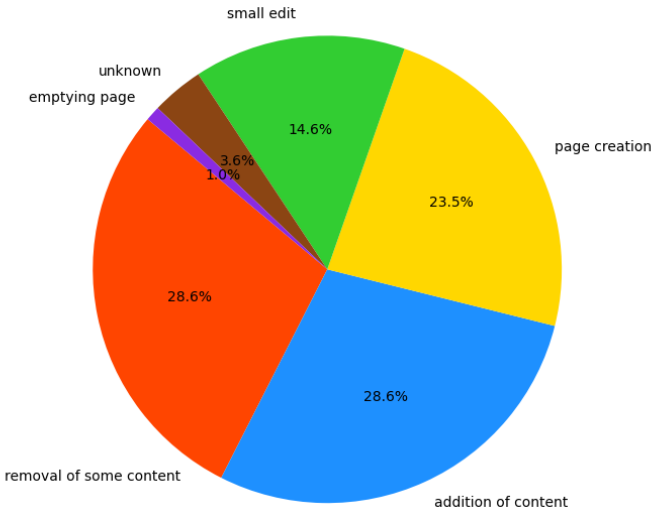


Figure 2: Mapping of different types of disruptive edits by anonymous users

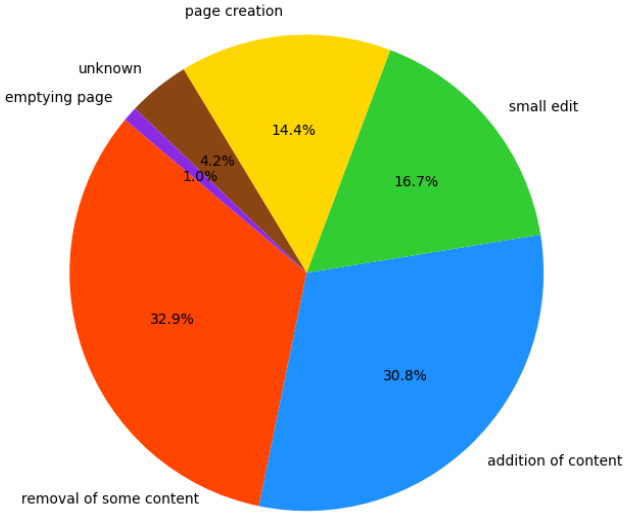


Fig 3. Mapping of different types of disruptive edits by all users

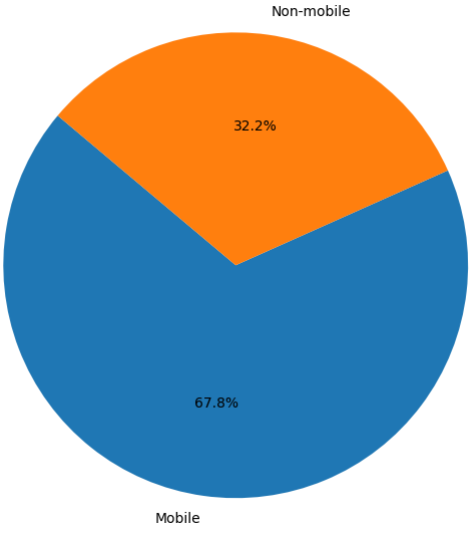


Figure 4: Percentage of disruptive edits made from a mobile device vs other devices for all users. Anonymous and account specific disruptive edits show a similar trend (68.6% and 64.5% respectively)

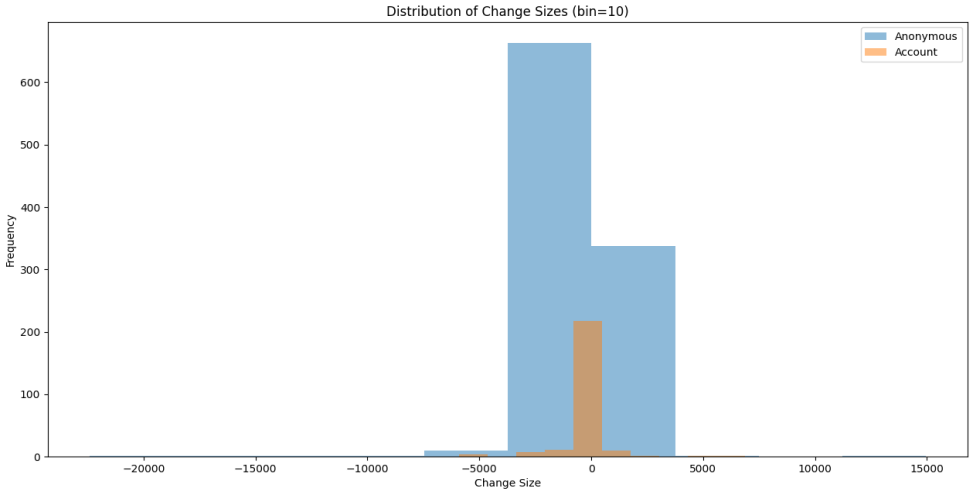


Figure 5: While disruptive edits by registered users are centered around 0 bytes and are smaller in size, anonymous disruptive edits tend to be larger and lean towards the negative (removal of content)

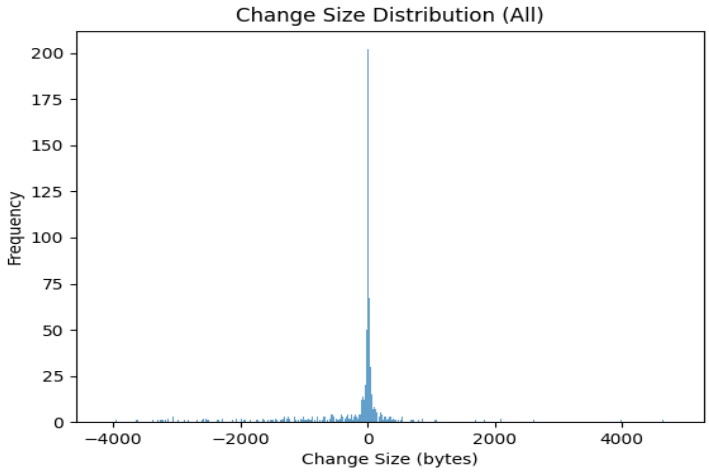


Figure 6: The majority of disruptive edits are small edits (close to 0 bytes added or removed)

8.2. Spelling tendencies

To understand word choice and spelling preferences of users, we collected the words used in the bare text of the creation edit of each page in the main namespace (articles). This amounted to almost 80,000 unique words and spellings from 6,806 non-bot articles (out of

a total of 10,312 articles at the time the data was collected, October 20, 2024).[33] All statistics in this section refer to the first version of each article in the Darija Wikipedia, not the current public-facing version of these articles. The purpose was to understand the "naive" and "spontaneous" language-related choices of editors, not tainted by later corrections, additions or improvements made by experienced and more active users. Articles created by bots (3,506) were excluded from this investigation, as they usually represent the typographic and spelling choices of a specific user (the bot operator), which given their high number, would skew the statistics. There are however some caveats with this approach, as the first edit may contain spelling mistakes, and it may also contain irrelevant or faulty content either added intentionally or due to a misunderstanding of how Wikipedia operates. These would likely become statistically insignificant as the dataset becomes larger. Furthermore, the dataset does not contain any contextual information, such as the full sentence, topic, and date of page creation, which can offer better insight into the reasons for the word and spelling choices, and the relationship to a given community spelling rule (which may not have existed when the word was used in the article). Finally, experienced and very active users are highly represented in the dataset, in comparison to less active users who made less contributions, and therefore a deeper analysis would require normalizing the data and using more advanced statistical methods to get better insights. But this is out of scope for this paper, which is restricted to a plain descriptive approach. The approach employed herein is indeed descriptive or indicative (showing examples of spelling tendencies in the Wiki), but not prescriptive (giving definite and final results regarding spelling tendencies in the Wiki).

The dataset consists of words embedded in a python dictionary that contains basic statistics about each word, namely the editors who wrote that word in an article and how often they did so. From this now-generated dataset, we tried to extract more insights related to letter usage, as well as word spelling frequency and their relationship with spelling rules that have been obtained through community consensus. To give an approximate idea about actual spelling preferences, we also indicated the number of unique editors who employed one spelling form or another[34].

Below are the general statistics for the dataset:

- Number of unique spelling forms: 79,529
- Total number of words in raw first edits: 454,903
- Number of unique characters: 1,492
- Number of unique Arabic-like characters: 69[35]
- Total number of unique editors: 189[36]

[33] Darija Wikipedia statistics of unique words. Link to the Google Sheet: https://tinyurl.com/j8r2698s.

[34] For more details on how the data was collected, see the code on Github: https://tinyurl.com/phtfwr8b.

[35] Includes Arabic characters as well as other characters for Darija-specific sounds, such as /g/ ڭ, /v/ ڤ, /p/ پ and emphatic /z/ ڜ.

[36] This refers to the total number of unique editors who wrote an article, not total number of editors who contributed to Darija Wikipedia in general, as the dataset is restricted only to the first edit of each article, i.e. only article creators whose articles have not been deleted. Editors who only made changes in articles, but never created a new one, would not be included.

8.3.1. Letter-level spelling tendencies

From the raw dataset of words, we extracted the number of unique editors and the total uses of letters in these words. The "total uses" here is defined as the total number of times a letter was used in all words by all editors. For example, if we have the word لمغريب (*l-meġrīb*) used 10 times by user1, and 3 times by user2, and the word مراكش (*merrākš*) used 5 times by user3 and 8 times by user1, we obtain the "total uses" of the letter م (*mīm*) as follows: 10+3+5+8 = 26. If a letter is used more than once in the same word (for example the word معمر *mʿemmer* has the letter م twice), it will be counted as in- word frequency multiplied by the frequency of the word (i.e. for the word معمر the frequency for the letter م will be multiplied by 2). The full statistics of the frequency of usage of all characters can be found in a Google sheet linked below[37]. The statistics for Darija letters are detailed in Appendix 2.

Our main goal from this exercise is to understand specific letter choices, in regards to the letters for /g/, /p/ and /v/ (which have no equivalent in Standard Arabic), as well as the Standard Arabic dental fricative letters ث (*ṯ*), ظ (*ẓ*), ذ (*ḏ*) whose Darija pronunciations are commonly equivalent to ت (*t*), ض (*ḍ*), د (*d*) respectively. To explore these topics, we extracted the target words that contain these specific letters or spelling forms, then compared them, in terms of frequency, with their alternatives.

8.3.1.1. /g/, /p/ and /v/

/g/ *characters*[38]

We collected the following statistics for the usage of various representations of /g/ (pronounced as in English *gap*) in the first version of each article:

Table 13. Occurrence of different representations of /g/ in Darija Wikipedia

Letter	UTF-32	Unique editors	Total instances
ڭ	u+000006ad	49	3,473
ݣ	u+00000763	38	1,740
گ	u+000006af	22	781
ڜ	u+00000686	7	19
ﮓ	u+000006b4	1	1
ﮕ	u+0000063b	1	1

[37] Frequency of arywiki Arabic-like letters (sheet 1) and all characters (sheet 2). Link to the Google Sheet: https://tinyurl.com/4r4ck7se.

[38] As noted by al-Midlāwī al-Mnabbhi 2019, the sound /g/ in Darija has multiple sources, such as *q* ق (e.g. *qāl* قال => *gāl* گال 'he said'), *ǧ* ج (e.g. *ǧles* جلس => *gles* گلس 'he sat down'), or may come directly from a foreign word (e.g. *gāwrī* گاوري 'foreigner', from Ottoman *gavur*; گيدون *gīdūn* 'steering wheel', from French *guidon*).

Interestingly, even though the letter ڤ is used in Algeria and Tunisia as an equivalent to the phoneme /g/,[39] in the Moroccan Darija Wikipedia in all 19 instances of its usage it represented the phoneme /v/ (see section for V below). In addition to that, many editors use the letters ك (*k*), ج (ǧ), or غ (ḡ) to represent this phoneme, as is common in Standard Arabic. For example, the word form كاع for *gāʿ* (meaning "*all*"), was introduced by 21 different editors, 47 times, in a variety of forms (e.g. base form or attached to a particle or suffix)[40]. Table 14 below shows the statistics of different spelling varieties of this word:

Table 14. Occurrence of different representations of the word *gāʿ* in Darija Wikipedia

Base form	Character	Unique editors	Total instances
كاع	(00000643+u) ك	21	47
ڭاع	(u+000006ad) ڭ	18	157
ڴاع	(00000763+u) ڴ	13	70
گاع	(u+000006af) گ	8	14
غاع	(u+0000063a) غ	2	4
جاع	(u+0000062c) ج	1	5
Total		**45**	**297**

/p/ *characters*

We found the following statistics for the usage of various representations of /p/ in the first version of each article:

Table 15. Occurrence of different representations of /p/ in Darija Wikipedia

Letter	UTF-32	Unique editors	Total instances
پ	u+0000067e	55	4,253
ٻ	u+0000067b	0	0

We note the absence of the character ٻ (UTF-32: u+0000067b) in the first versions of Darija Wikipedia articles. This character was also rare in later revisions, and all of its instances had been replaced by the more common پ (UTF-32: u+0000067e) using PGVBot.

[39] https://w.wiki/BhPT.

[40] The statistics in the table indeed represent aggregated counts for different usage forms of the same. spelling. For example, the statistics for بكاع, فكاع, كاعما, كاع, etc were all aggregated and represented by their base form كاع in the table. See the code of the script for more details: https://tinyurl.com/5b3fk7nc.

/v/ *characters*

We found the following statistics for the usage of various representations of /v/ in the first version of each article:

Table 16. Occurrence of different representations of /v/ in Darija Wikipedia

Letter	UTF-32	Unique editors	Total instances
ڤ	u+000006a4	41	1,912
ۋ	u+000006a8	6	19

We note the low frequency of alternative P and V characters, in comparison to the main character for each that has been adopted in Darija Wikipedia by consensus. This could be due to their availability in *Lexilogos* Arabic keyboard[41] which is one of the most commonly used external keyboards, and one of the recommended ones in Wikipedia help pages, to use for writing in Darija Wikipedia.[42] For the same reason, the character ڭ (UTF-32: u+000006ad) appears more frequently in the first edit, in comparison to alternatives like ک (UTF-32: u+00000763) and the Farsi گ (UTF-32: u+000006af), as well as the Farsi tch چ (UTF-32: u+00000686), or even the much less frequent ګ (UTF-32: u+0000063b). The character ک (UTF-32: u+00000763) is in fact, as already noted, the one that has been adopted by consensus, and all other forms of /g/ should be converted to it in subsequent edits. The common practice in Standard Arabic of using available letters like ب (*b*), ف (*f*), or ك (*k*) to represent /v/, /p/ and /g/ respectively, seems to continue among some editors (at least at page creation). This variety may reflect the types of input systems available for writing in Darija Wikipedia and their options and limitations, or conscious choices by some editors.

8.3.1.2. Dental fricatives letters

Table 17 below shows the statistics of the usage of dental fricative letters in Darija Wikipedia on the creation of articles.

Table 17. Occurrence of dental fricative letters in Darija Wikipedia

Letter	Unique editors	Total instances
ث (t̲)	126	2,706
ذ (d̲)	90	1,478
ظ (z̲)	84	917

The use of these letters is therefore by no means marginal, even though many editors replace them with non-fricative letters, in line with common pronunciation of Koine Darija.

[41] https://www.lexilogos.com/clavier/araby.htm.

[42] Wikipedia tools page: https://w.wiki/BhkM.

Letter ذ (ḏ)

To get a better idea about the distribution of usage, and similarly to the word /gāʿ/, we surveyed the usage distributions of the three most common words, with a dental fricative, in Darija Wikipedia's first page revisions, whose Darija equivalent can be written with its non-fricative equivalent.[43] These were the forms of *hada* 'this', *dheb* 'gold' and *dker* 'male', as shown in Table 18[44].

Table 18. Comparison of occurrence of use vs non-use of the dental fricative letter ذ (ḏ) for *hada* 'this', *dheb* 'gold' and *dker* 'male'

Base form	Character	Unique editors	Total instances
هادا (hādā)	د (d)	94	2,413
هذا (hḏa)	ذ (ḏ)	42	203
Total		**106**	**2,616**
ذهب (ḏhb)	ذ (ḏ)	27	61
دهب (dhb)	د (d)	21	93
Total		**38**	**154**
ذكر (ḏkr)	ذ (ḏ)	26	91
دكر (dkr)	د (d)	16	80
Total		**34**	**171**

Many editors seem to prefer using the non-fricative د (d) instead of the dental fricative ذ (ḏ), but there is also an overlap, with some editors sometimes using one form or another. Noting that many uses of ذ may come from incomplete translations of Standard Arabic articles into Darija (for example using the Content Translation Tool[45]). Nonetheless, the usage of the dental fricative letter ذ (ḏ) in written Darija remains significant.

Letter ث (ṯ)

The statistics for ث (*ṯ*) vs ت (*t*) are generally close among unique editors, but the ت (*t*) has a significant advantage in terms of number of instances, which reflects the preferences of the most active users.

[43] The word الذي *alladī* for instance was excluded, since neither it nor الدي *alladī* are used in Darija, and it has alternative equivalents that do not include the letter د (d).

[44] For example, *haḏihi* or *hadi* هذه, *hādūk* هادوك, *lihaḏā* لهذا, etc are all included in the statistics and aggregated with their corresponding form with (ḏ) ذ or (d) د. For a full list of the forms, check Set 2 in the list "spelling_variants" in the code https://tinyurl.com/5b3fk7nc.

[45] A tool which assists editors in translating existing Wikipedia articles from one language to another, and can involve automated translations generated by AI (See: https://w.wiki/CM6Y). As of April 2025, the automated translation using MinT does not work very well for Darija (see Phabricator ticket: https://w.wiki/CM6Z).

Table 19. Comparison of occurrence of use vs non-use of the dental fricative letter ث (ṯ) for *tani* 'second', *ḥit* 'because and *kter* 'more'

Base form	Character	Unique editors	Total instances
تاني (tānī)	ت (*t*)	47	617
ثاني (ṯānī)	ث (*ṯ*)	46	181
Total		**67**	**798**
حيت (ḥīt)	ت (*t*)	53	560
حيث (ḥīṯ)	ث (*ṯ*)	36	106
Total		**67**	**666**
كتر (ktr)	ت (*t*)	52	550
كثر (kṯr)	ث (*ṯ*)	48	206
Total		**77**	**756**

Letter ظ (ẓ)

In the case of the letter ظ (*ẓ*), it seems that the editors' preference tilts stronger towards a more etymological rather than a phonetic spelling, in comparison to other dental fricative letters, at least for the most common words.

Table 20. Comparison of occurrence of use vs non-use of the dental fricative letter ظ (ẓ) for *niḍam* 'system, order', *ḥfeḍ* 'he learned' and *naḍariya* 'theory'

Base form	Character	Unique editors	Total instances
نظام (nẓām)	ظ (ẓ)	32	81
نضام (nḍām)	ض (ḍ)	18	74
Total		**40**	**155**
حفظ (ḥfẓ)	ظ (*ẓ*)	30	102
حفض (ḥfḍ)	ض (ḍ)	15	70
Total		**40**	**172**
نظرية (nẓrya)	ظ (ẓ)	11	27
نضرية (nḍrya)	ض (ḍ)	9	42
Total		**17**	**69**

8.3.2. Word-level spelling tendencies

Below is an analysis of spelling forms and their relationship to Wikipedia spelling rules. Additionally, we will investigate spelling tendencies within the most used words in the first revision of each non-bot article in Darija Wikipedia.[46]

Rule 1: Prepositions of possession

As detailed in section 4, rule 1 in كناش لقواعد (*kennāš l-qwāʿd*) deals with prepositions of possession (equivalent to English "of"). The rule does not account for the particle ت (t), whose usage is not addressed, but which shows up in the data.

Following, we explore the usage distributions of prepositions of possession in the first revisions of the Wiki. There are two groups of these prepositions: ديال (dyāl) and its reduced form د (d) (attached or separate from the next word – but not with a suffixed pronoun), and متاع (mtāʿ) / نتاع (ntāʿ) / تاع (tāʿ) and their reduced form ت (t) (attached or separate). To ensure that the attached prepositions were not actually words that started with the letters د (d) or ت (t), the data had to be cleaned up manually.

By a significant margin, the first group is more represented in the dataset (see Table 21). This may be reflective of the regional and dialectal backgrounds of the editors who are active on the Wiki and may also reflect socio-economic and/or linguistic realities in Morocco and among the Moroccan diaspora (for instance, access to the Internet, urban vs rural dialects, etc.). Furthermore, a big number of editors prefer to attach the prepositions د (d) and ت (t) to the next word, even though the rule clearly states that prepositions should be written separately.

Table 21. Occurrence of different prepositions of possession in Darija Wikipedia

Base form	Unique editors	Total instances
ديال (dyāl)	137	6,349
د (d)	83	3,064
د (d) (att.)	57	1,353
تاع (tāʿ)	17	172
نتاع (ntāʿ)	8	27
ت (t)	8	18
ت (t) (att.)	7	82
متاع (mtāʿ)	2	6
Total	**148**	**11,071**

[46] See the list of the most used 100 words and spelling forms in the Appendix 3

Rule 2: Connectors

Rule 2 lists some prepositions and connectors, and examples of their usage, and suggests that they should be written separately from the next word. These prepositions are among the most common words in Darija Wikipedia, as shown in Appendix 3, listing the 100 words with the highest usage frequency. Here below, we present three tables (22, 23 and 24) with raw comparisons of various prepositions that are etymologically related and/or functionally equivalent and may reflect dialectal varieties or personal preferences.

Table 22. Occurrence of connectors *kī* vs *kīf* 'as'/'like' in Darija Wikipedia

Base form	Unique editors	Total instances
كي (kī)	34	144
كيف (kīf)	23	66
Total	**43**	**210**

Both prepositions على (ʿlā) 'on' and بحال (bḥāl) 'like' are used significantly more than their equivalent alternative spellings or closely related forms. Other aspects that could have been investigated include the attachment and separateness of the prepositions ب (b), ف (f) and ل (l), as well as the assimilation and attachment/detachment of من and عل. The same could be said about حتى (ḥtā) / حتى (ḥtā) / تا (tā), all meaning 'until'. Due to the limited scope of this paper and insufficient time resources, the answers to these questions were not pursued, but they could be part of another paper that focuses on the linguistic aspects of the Wiki and spelling preferences (see Section 9).

Table 23. Occurrence of the synonymous connectors *bḥal* vs *fḥal* in Darija Wikipedia

Base form	Unique editors	Total instances
بحال (bḥāl)	78	1,126
فحال (fḥāl)	12	21
Total	**81**	**1,147**

Table 24. Occurrence of difference representations of *ʿlā* 'on' in Darija Wikipedia

Word	Unique editors	Total instances
على (ʿlā)	108	3,343
علا (ʿlā)	16	270
عل (ʿl)	13	33
Total	**110**	**3646**

Rule 3: ‘and’ / ‘or’

Rule 3 concerns the accepted forms for ‘and’ and ‘or’. This is a crucial issue, since some editors use the word form أو which can be pronounced as *u* by some (meaning ‘and’) and *ʾaw* by others (meaning ‘or’). Therefore, أو (ʾāw) is not accepted, and is systematically replaced by either و/ؤ (*w/ū/o/u*) ‘and’ or ولا (*wellā*) ‘or’. Some editors also use ؤلا and أولا which are accepted in practice, despite not being addressed by rule 3. The rule also specifies that these particles should be separated from the next word. Following, we investigate the usage distributions of these forms. Not included are the forms where و/ؤ are attached to the next word, since they are too numerous (estimated between 5,000 and 6,000 unique word forms).

Note that ولا (*wlā*) could also represent the verb *wellā* ‘to become, and أولا (*ʾawlā*) could be the Standard Arabic term for *ʾawwalā* ‘first of all’, which can sometimes be used in Darija. This shows the limitations of this comparative approach, especially using a dataset of words without their original context. The results showcase the limited effectiveness of enforcing spelling rules in changing writing habits of editors.

Table 25. Occurrence of different representations of the words meaning ‘and’ and ‘or’ in Darija Wikipedia

Base form	Unique editors	Total instances
و (w; ū)	122	3,882
ولا (welā)	71	3,177
أو (ʾaw)	50	601
أولا (ʾawlā)	32	180
ؤ (o; u)	30	1,104
ؤلا (ulā)	12	642
Total	**137**	**9,586**

Rule 4: The *tā marbūṭa*

Rule 4 deals with the *tā marbūṭa* (ة) and recommends its usage for words of Arabic etymology, even in noun groups (for example, محلبة الحسين *maḥlabat lḥūsīn*, not محلبت الحسين *maḥlabat lḥūsīn*). Some editors prefer to replace it with *ʾalif* at the end of the word (ا). Table 26 presents a general comparison of the usage statistics of the 5 most used words that have

a form with *tā marbūṭa*[47], whereas in Table 27, we compare the usage of *tā marbūṭa* with *tā mabsūṭa* in noun groups[48].

Table 26. Comparison of usage of *tā marbūṭa* vs *'alif* in the 5 most used words with *tā marbūṭa* in Darija Wikipedia

Base form	Unique editors	Total instances
ـة	104	5,510
ـا	18	80
Total	**105**	**5,590**

Table 27. Comparison of usage of *tā marbūṭa* vs *tā mabsūṭa* in noun groups

Base form	Unique editors	Total instances
ـة	87	2,476
ـت	10	28
Total	**89**	**2,504**

Rule 5: Definite/Indefinite forms

In rule 5, two possible ways to write the definite form are prescribed:

- ال (āl) for the solar letters and ل (l) for the lunar letters[49],
- only a *shadda*[50] on the first letter for solar letters and ل (*l*) for lunar letters.

Given that this is an encouraged not a mandatory rule, many users fall back on the Standard Arabic rule of adding ال (āl) in both cases of solar and lunar letters. Table 28 shows the statistics for the spelling of the definite form for lunar letters, whereas Table 29 presents the distribution of *shadda* vs *āl* for solar letters.

[47] These words are مدينة *mdīna* 'city, كبيرة *kbīra* (big, f.), دولة *dūla* or *dawla* 'state', كاينة *kāyna* 'existing',مجموعة *meğmuʿa* 'group' and their various word form occurrences. Excluded from the statistics was وحدة *weḥda* 'one', which is among the top 100 most frequently used words, but its '*alif* spelling form وحدا (weḥdā) can be interpreted as وحدا = *w+ ḥdā* meaning 'and next to'.

[48] For example, مدينت الدارلبيضا vs مدينة الدارلبيضا *mdīnt ddār lbīd*ā 'the city of Casablanca'. Surveyed were دولة *dūla* or *dawla* 'state', مجموعة *meğmuʿa* 'group', مدينة *mdīna* 'city', and their various word form occurrences. To make the comparison fairer, the forms with tā marbūṭa had their definite word forms removed from the investigation, since the definite form would never occur for the first noun in a noun cluster (e.g. مدينة الدارلبيضا not المدينة الدارلبيضا). Nonetheless, given the lack of context, it remains uncertain if the spelling forms with tā marbūṭa are the first word in a noun cluster or not.

[49] The solar or sun or shamsi letters are letters that, when they occur at the beginning of a noun, eclipse the pronunciation of the *lam* ل in the definite particle āl- ال (equivalent to *the* in English). The moon or lunar or qamari letters are the opposite, where the āl- ال is pronounced fully. See An-Nassir 1985: 79.

[50] The *shadda* شدة is a diacritic symbol which indicates doubling of a consonant and is represented with ّ. For example, برا (*brā*) 'to get healed, and برّا (*berrā*) 'outside'.

Table 28. Occurrence of the two ways of representing definite form for lunar letters

Base form	Unique editors	Total instances
ال (āl)	77	601
ل (l)	59	3,032
Total	**97**	**3,633**

Table 29. Occurrence of using *shadda* vs ال (āl) for solar letters

Base form	Unique editors	Total instances
ال (āl)	76	1,350
shadda	15	4,246
Total	**79**	**5,596**

Rule 6: The *hamza*

Rule number 6 deals with the character ء (*hamza*) at the end of words. Following we investigate the usage distributions of words that can be written with or without *hamza* in Darija.

Table 30. Comparison of use vs non-use of the *hamza* character when it is at the end of a word

Base form	Unique editors	Total instances
with ء	47	161
without ء	43	810
Total	**61**	**971**

Forms of *lmeġrīb* 'Morocco'

As shown in Table 31, orthographic variation is so widespread when writing that Darija Wikipedia editors use four distinct spellings of the word *lmeġrīb* 'Morocco'.

Table 31. Occurrence of the four main representations of the word 'Morocco' in Darija Wikipedia

Base form	Unique editors	Total instances
المغرب	48	209
لمغريب	34	1,020
المغريب	18	67
لمغرب	13	30
Total	**65**	**1,326**

The variety of spellings of *lmeġrīb* 'Morocco' in Darija highlights the two main spelling tendencies in the Wiki (and in written Darija in general), namely the etymological vs phonetic spelling. The etymological spelling (using *āl-* and without the explicit vowel ي instead of *kasra*), seems to have a non-negligible advantage in terms of choice made by unique editors, while the phonetic spelling seems to be more widespread in terms of number of instances, reflecting the personal choices of more active editors, as opposed to casual and less active editors; though there is also some overlap that could be indicative of either uncertainty or hesitancy on the part of some editors, evolution of preferences over time, or simply a desire for variation and experimentation. This is a general observation that can be drawn more or less from the various examples and statistics for all rules and spelling forms investigated, and it merits deeper and wider investigation in the Wiki itself, as well as comparison with other sources of written Darija, such as printed literature, blogs and social media.

8.4. Trends

Since the launch of Darija Wikipedia in July 2020, the project has attracted dozens of contributors. From then until April 2025, there has been an average of 67 contributors participating each month. Peaks are particularly observed during editing contests (such as in April-June 2023 and July-August 2024), as shown in Figure 7.

Figure 7: Number of editors of Darija Wikipedia per month – August 2020 - March 2025[51]

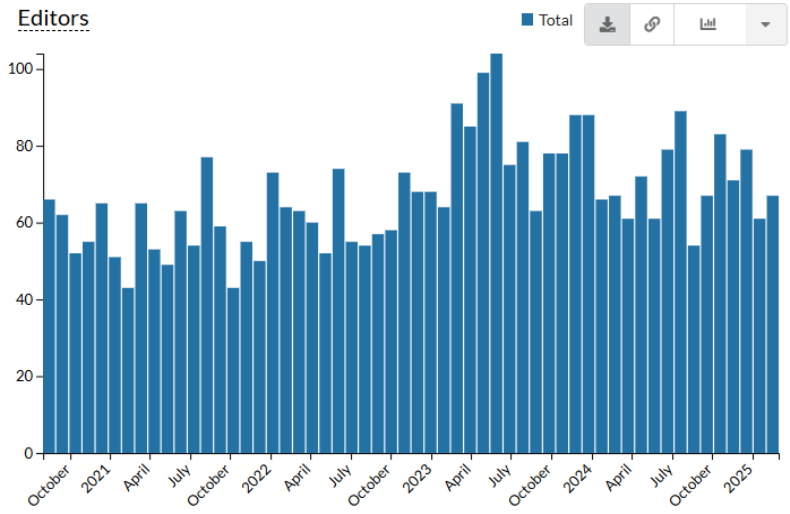
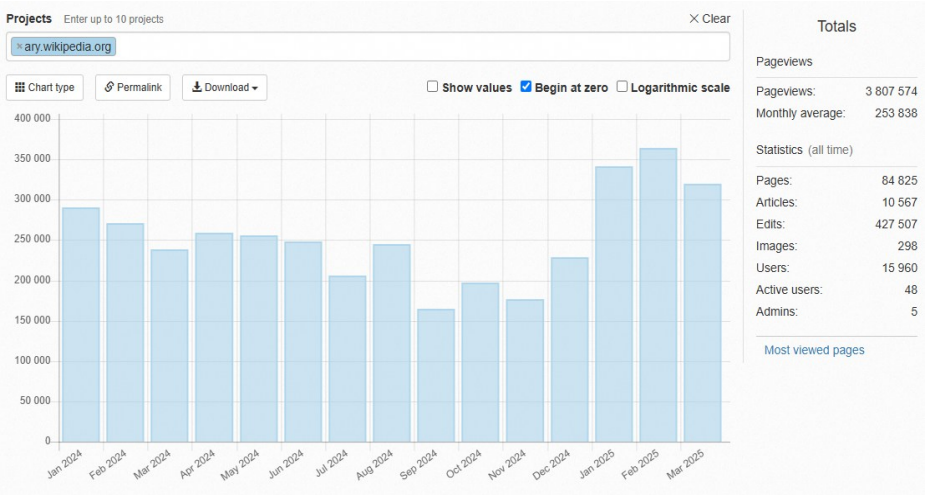


Figure 8 presents page views of Darija Wikipedia in 2024-2025. While variating from one month to another, they still show a slight but consistent increase on average. The peak activity in some periods may be due to the contests that have been organized and attracted many occasional editors.

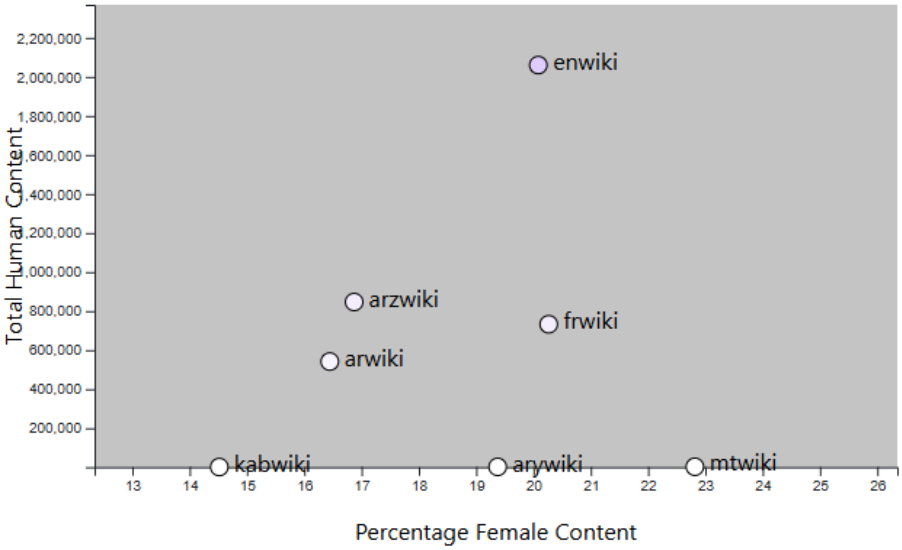
[51] Wikimedia Statistics - https://w.wiki/CM6b

Figure 8: Pageviews per month (Darija Wikipedia) – January 2024 – March 2025[52]



In terms of gender content gap,[53] Darija Wikipedia (ary), although having fewer articles than the well-established versions, has a similar ratio to French (fr) and English (en) (nearly 19.3%) and performs much better than the versions in Arabic (ar) and Egyptian (arz). Figure 9 below, as well as Appendix 1 detail this information in different forms.

Figure 9: Gender gap by language editions of Wikipedia (All time, as of 2025-04-14)



Source: humanikidata.org powered by Wikidata, CC BY-SA 3.0

[52] Wikimedia Siteviews Analysis (Moroccan Darija) – https://pageviews.wmcloud.org/siteviews/?platform=all-access&source=pageviews&agent=user&start=2024-01&end=2024-09&sites=ary.wikipedia.org

[53] Gender content gap refers to the difference in coverage of topics about women vs men. https://w.wiki/DsZ7

9. Opportunities

Darija Wikipedia is certainly one of the biggest structured datasets in Darija online, and most probably the largest open one[54]. In this regard, it can be considered for several implementations in artificial intelligence (AI). One can, for example, cite its use in NLP (presented at AMLD Africa 2021[55]), named-entity recognition (Moussa & Mourhir 2023), chatbots (Shang et al. 2024), or the growing interest by the private sector in potential applications, such as *Sawalni* (Ask Me)[56].

Being mainly a spoken language, Darija can also benefit from systems enhancing speech synthesis (text to speech) and speech recognition (speech to text). Besides making this Wikipedia version more popular, this could also free volunteers from manually recording articles, who might then use this time to create more content on other subjects. Automatization is already a reality in this Wiki, where bots are used to perform many tasks, including writing some articles, but there is always a potential to develop even more, especially with the active support of Wikimedia Morocco User group.

As a relatively small Wikipedia, the Darija version provides an opportunity to become a reference for other Wikis that are in a similar situation. Although Wikipedias are independent from each other, there are many common aspects that can be developed in one that can then be successfully deployed in others. With the presence of several technically skilled volunteers in the team, Darija Wikipedia can pave the way for other small and minority language communities, by developing standard generic templates and modules that can be used later for other languages as well, since Wikis follow generally the same structure. The Moroccan Darija Wikipedia can therefore capitalize on these arguments to bring even more volunteers on board and become a reference in its category.

Finally, research is also an important area providing several opportunities for Darija Wikipedia. On the one hand, collaboration with researchers and experts will raise awareness and promote research about Darija in general in academia, and on another, it will enrich content about this subject and improve the overall quality of the encyclopedia. One concrete example of research work that can be applied in the Darija Wikipedia is application of graph theory to understand connections between different stakeholders of Wikipedia (readers, users, administrators, etc.), in addition to analyzing interactions between users on talk pages, and their effects on editor productivity and retention.

10. Conclusion

Today, four years after its launching, the Moroccan Darija Wikipedia has over 10,500 articles, 4 administrators and an average of 250,000 monthly pageviews. These numbers

[54] https://w.wiki/CM6i

[55] Moroccan Darija Wikipedia: Basics of Natural Language Processing for a Low-Resource Language – AMLD Africa 2021 – https://appliedmldays.org/events/amld-africa-2021/workshops/moroccan-darija-wikipedia-basics-of-natural-language-processing-for-a-low-resource-language

[56] *Sawalni*, the first AI chatbot 100% in Darija – https://sawalni.com/

were reached through efforts of different volunteers collaborating online with the support of the Wikimedia Morocco User Group.

After providing a short background introduction to Wikimedia and its pillars, a description of the process of the Wikipedia Darija creation was presented. It was then followed by diving into that community and its governance, by detailing the different levels and types of users, and explaining how standardization and policies are created between existing editors. Examples of phonology and verbs conjugation were used to provide concrete cases community works on.

A particularity of the Darija Wikipedia among other smaller wikis is its technical aspects, that are used on a big scale. In that regard, the 8 bots operating on ary.wikipedia.org were explained, as well as interface translation, and the various types of namespaces active in this version.

Operationally, strategies used by Wikimedia Morocco to promote Darija Wikipedia (campaigns, contests and outreach) were explained, before detailing challenges still to be addressed, either in terms of processes, of community sustainability or vandalism. The latter aspect was further analyzed, with statistics mapping its different types, devices used, and size. Interesting findings from these statistics were that most disruptive edits (67.8%) come from mobile devices, and that anonymous accounts tend to have larger vandalism, leaning more towards removing content. Spelling variations among editors were also explored, showing a wide variety, but also some strong tendencies, which do not always conform to general community consensus.

Finally, the paper is concluded by investigating opportunities for future work, to further develop the Darija Wikipedia, and produce more research around it. Ideas for next steps would be to work on AI applications, advance in standardization, develop tools to explore spelling and grammar, enhance speech synthesis and speech recognition, as well as developing guidelines to become a reference for similar Wikis.

References

Al-Nassir, Abdulmunim Abdulamir. 1985. *Sibawayh the phonologist: A critical study of the phonetic and phonological theory of Sibawayh as presented in his treatise Al-Kitab.* York: University of York. (Doctoral dissertation.)

Alshahrani, Saied & Wali, Esma & Matthews, Jeanna. 2022. Learning from Arabic corpora but not always from Arabic speakers: A case study of the Arabic Wikipedia editions. In Bouamor, Houda etc. (eds.), *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, 361-371. Abu Dhabi: Association for Computational Linguistics.

Alshahrani, Saied & Alshahrani, Norah & Dey, Soumyabrata & Matthews, Jeanna. 2023. Performance implications of using unrepresentative corpora. In Sawaf, Hassan etc. (eds.), *Arabic Natural Language Processing. Proceedings of Arabic NLP 2023*, 218-231. Singapore: Association for Computational Linguistics.

Alshahrani, Saied & Haroon, Hesham & Elfilali, Ali & Njie, Mariama & Matthews, Jeanna. 2024. Leveraging corpus metadata to detect template-based translation: An exploratory case study of the Egyptian Arabic Wikipedia edition. In Al Khalifa, Hend & Darwish, Kareem & Mubarak, Hamdy (eds.), *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, 31-45. Torino: *LREC--COLING*

Barrett, Daniel J. 2008. *MediaWiki*. Beijing etc.: O'Reilly.

Baytiyeh, Hoda & Pfaffman, Jay. 2010. Volunteers in Wikipedia: Why the community matters. *Journal of Educational Technology & Society* 13(2). 128-140.

Behnstedt, Peter & Benabbou, Mostafa. 2005. Données nouvelles sur les parlers arabes du Nord-Est marocain. *Zeitschrift für Arabische Linguistik* 44. 17-70.

Boumans, Louis. 2006. The attributive possessive in Moroccan Arabic spoken by young bilinguals in the Netherlands and their peers in Morocco. *Bilingualism: Language and Cognition* 9(3). 213-231.

Caubet, Dominique. 2018. New elaborate written forms in Darija: Blogging, posting and slamming in Morocco. In Benmamoun, Elabbas & Bassiouney, Reem (eds.), *The Routledge handbook of Arabic linguistics*, 387-406. London: Routledge.

Chtatou, Mohamed. 1997. The influence of the Berber language on Moroccan Arabic. *International Journal of the Sociology of Language* 123. 101-118.

Darija Wikipedia. 2025a. *Discussion Page: Namespace*. (https://w.wiki/AiaQ) (Accessed 2025-04-24.)

Darija Wikipedia. 2025b. *List of Administrators*. (https://w.wiki/AipB) (Accessed 2025-04-24.)

English Wikipedia. 2025a. *The five pillars of Wikipedia*. (https://w.wiki/5) (Accessed 2025-04-24.)

English Wikipedia. 2025b. *Vandalism*. (https://w.wiki/mrS) (Accessed 2025-04-24.)

Ennaji, Moha & Makhoukh, Ahmed & Es-Saiydi, Hassan & Moubtassime, Mohamed & Slaoui, Souad. 2004. *A grammar of Moroccan Arabic.* Fès: Faculty of Letters Dhar El Mehraz.

Forte, Andrea & Larco, Vanesa & Bruckman, Amy. 2009. Decentralization in Wikipedia governance. *Journal of Management Information Systems* 26(1). 49-72.

Gilfillan, Ian. 2024. October 2024 African language Wikipedia update. (https://www.greenman.co.za/blog/?p=2944) (Accessed 2025-04-24.)

Heath, Jeffrey. 1997. Moroccan Arabic phonology. *Phonologies of Asia and Africa (including the Caucasus)* 1. 205-217.

Heath, Jeffrey. 2015. D-possessives and the origins of Moroccan Arabic. *Diachronica* 32(1). 1-33.

Heath, Jeffrey. 2020. Moroccan Arabic. In Lucas, Christopher & Manfredi, Stefano (eds.), *Arabic and contact-induced change*, 213-223. Berlin: Language Science Press.

Konieczny, Piotr. 2018. Volunteer retention, burnout and dropout in online voluntary organizations: Stress, conflict and retirement of Wikipedians. In Coy, Patrick G. (ed.), *Research in social movements, conflicts and change*, vol. 42, 199-219. Bingley: Emerald Publishing Limited

Massa, Paolo, & Scrinzi, Federico. 2011. Exploring linguistic points of view of Wikipedia. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration,* 213-214. New York: Association for Computing Machinery.

McCarthy, Philip M. & Jarvis, Scott. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods* 42(2). 381-392.

MediaWiki. 2010. *Manual:Bots*. (https://w.wiki/DsYp) (Accessed 2025-04-24.)

Meta Wikimedia. 2007. *Language proposal policy*. (https://w.wiki/5RsC) (Accessed 2025-04-24.)

Meta Wikimedia. 2025. *List of Wikipedias*. (https://w.wiki/7iw) (Accessed 2025-04-24.)

Michalski, Marcin. 2016. Spelling Moroccan Arabic in Arabic script: The case of literary texts. In Grigore, George & Bițună, Gabriel (eds.), *Arabic varieties – far and wide: Proceedings of the 11th International Conference of AIDA – Bucharest, 2015*, 385-394. București: Editura Universității din București.

al-Midlāwī al-Mnabbhi, Muḥammad. 2019. *Al-ʿArabiyya al-dāriǧa: Imlāʾiyya wa-naḥw: Al-aṣwāt, al-ṣarf, al-tarkīb, al-muʿǧam* (Darija Arabic: Spelling and grammar: Sounds, conjugation, structure, vocabulary). Zākūra: Markaz Tanmiyat al-Dāriǧa.

Miller, Catherine. 2017. Contemporary dārija writings in Morocco: Ideology and practices. In Høigilt, Jacob & Mejdell, Gunvor (eds.), *The politics of written language in the Arab world: Writing change*, 90-115. Leiden: Brill.

Moussa, Hanane Nour & Mourhir, Asmaa. 2023. DarNERcorp: An annotated named entity recognition dataset in the Moroccan dialect. *Data in Brief* 48. 109234.

Moustaoui Srhir, Adil. 2012. Language planning, standardization and dinamics of change in Moroccan Arabic. *Dialectologia* 9. 53-69.

Moustaoui Srhir, Adil. *Sociolinguistics of Moroccan Arabic: New topics*. Frankfurt/Berlin: Peter Lang.

Mrini, Khalil & Bond, Francis. 2018. Putting figures on influences on Moroccan darija from Arabic, French and Spanish using the Wordnet. In Bond, Francis & Piek, Vossen & Fellbaum, Christiane (eds.), *Proceedings of the 9th Global Wordnet Conference*, 372-377. Singapore: Global Wordnet Association.

Ouhalla, Jamal. 2015. The origins of Andalusi-Moroccan Arabic and the role of diglossia. *Brill's Journal of Afroasiatic Languages and Linguistics* 7(2). 157-195.
Šafīq, Muḥammad. 1999. *Al-Dāriğa al-maġribiyya: Maǧāl tawārud bayn al-amāzīǧiyya wa-al-ʿarabiyya*. Rabat: Academy of the Kingdom of Morocco.
Sedrati, Anass & Ait Ali, Abderrahman. 2019. Moroccan Darija in online creation communities: Example of Wikipedia. *Al-Andalus Magreb* 26(1). 1-14.
Shang, Guokan & Abdine, Hadi & Khoubrane, Yousef & Mohamed, Amr & Abbahaddou, Yassine & Ennadir, Sofiane & Momayiz, Imane & Ren, Xuguang & Moulines, Eric & Nakov, Preslav & Vazirgiannis, Michalis & Xing, Eric. 2024. *Atlas-Chat: Adapting Large Language Models for low-resource Moroccan Arabic dialect*. *arXiv preprint.* (https://arxiv.org/pdf/2409.17912) (Accessed 2025-04-24.)
The Economist. 2021. *Wikipedia is 20, and its reputation has never been higher.* (https://www.economist.com/international/2021/01/09/wikipedia-is-20-and-its-reputation-has-never-been-higher) (Accessed 2025-04-24.)
Wikimedia Foundation. 2025. *About us.* (https://wikimediafoundation.org/about/) (Accessed 2025-04-24.)
Wikimedia Incubator. 2007. *Incubator: About.* (https://w.wiki/3Sav) (Accessed 2025-04-24.)
Wikimedia Morocco. 2021. *Annual Report.* (https://w.wiki/DsY$) (Accessed 2025-04-24.)
Wikimedia Morocco. 2023. *Annual Report.* (https://w.wiki/Cg2t) (Accessed 2025-04-24.)
Wikimedia Statistics. 2025. *Moroccan Darija Monthly Overview.* (https://w.wiki/DsZ4) (Accessed 2025-04-24.)

Appendices

Appendix 1. Statistics of articles about males and females on selected Wikipedia versions

LANGUAGE Enter LANGUAGE...	Total	female	female Percent	male	male Percent	∑ Other Genders	∑ Other Genders Percent
Kabyle Wikipedia	675	98	14.519%	577	85.481%	0	
Arabic Wikipedia	541 226	89 010	16.446%	451 786	83.475%	430	0.079%
Egyptian Arabic Wikipedia	846 262	142 780	16.872%	702 796	83.047%	686	0.081%
Moroccan Arabic Wikipedia	1 507	292	19.376%	1 213	80.491%	2	0.133%
English Wikipedia	2 060 228	413 796	20.085%	1 643 542	79.775%	2 890	0.140%
French Wikipedia	732 349	148 423	20.267%	582 831	79.584%	1 095	0.150%
Maltese Wikipedia	2 108	481	22.818%	1 625	77.087%	2	0.095%

Appendix 2. Occurrence of the letters used in ary Wikipedia

	Letter (Transcription)	Unique editors	Total uses
1	ي (y; ī)	182	199,232
2	ر (r)	182	109,652
3	ا (ā)	181	257,815
4	ن (n)	181	105,730
5	ل (l)	180	200,779
6	م (m)	180	121,839
7	ب (b)	179	82,318
8	د (d)	179	82,023
9	و (w; ū)	178	137,273
10	ه (h)	178	40,628
11	ت (t)	176	85,149
12	س(s)	175	59,050
13	ف (f)	175	49,394
14	ج (ǧ)	174	29,295
15	ة (a; t)	173	80,935
16	ك (k)	171	60,811
17	ع (ʿ)	170	60,455
18	ق (q)	164	43,109
19	ز (z)	161	16,968
20	ح (ḥ)	159	34,349
21	ش (š)	157	29,294
22	ط (ṭ)	157	28,643
23	أ (ʾa)	155	20,754
24	ص (ṣ)	155	19,167
25	غ (ḡ)	151	12,749
26	خ (ẖ)	143	23,982
27	ض (ḍ)	137	10,059
28	ى (ā)	134	6,648

	Letter (Transcription)	Unique editors	Total uses
29	إ (i)	127	9,720
30	ث (ṯ)	126	2,706
31	ئ (i)	120	3,141
32	ء (ʾ)	104	3,016
33	ذ (ḏ)	90	1,478
34	ظ (ẓ)	84	917
35	ؤ (o; u)	77	4,066
36	آ (ʾā)	76	2,080
37	پ (p)	55	4,253
38	ڭ (g)	49	3,473
39	ڤ (v)	41	1,912
40	ک (g)	38	1,740
41	گ (N/A)	22	781
42	ک (N/A)	8	11
43	چ (N/A)	7	19
44	ڤ (N/A)	6	19
45	اً (N/A)	2	4
46	ھ (N/A)	2	4
47	ڒ (N/A)	2	3
48	ڪ (N/A)	1	1
49	ډ (N/A)	1	1
50	ڜ (N/A)	1	1
51	ب (N/A)	1	1
52	ك (N/A)	1	1
53	گ (N/A)	1	1
54	ۏ (N/A)	1	1
55	ۋ (N/A)	1	1
56	ێ (N/A)	1	1

Appendix 3. List of the 100 most used words and spelling forms

	Word	Unique editors	Total uses
1	من	142	3,137
2	ديال	128	2,894
3	هو	127	1,986
4	هي	123	1,535
5	و	122	3,882
6	في	110	1,978
7	على	108	3,343
8	ف	106	4,799
9	بزاف	98	1,575
10	واحد	89	2,670
11	لي	88	4,558
12	باش	87	1,069
13	هاد	85	1,554
14	عام	85	975
15	د	83	3,064
16	كان	83	1,351
17	مع	80	691
18	فيها	79	2,664
19	ديالو	79	1,571
20	بحال	78	1,066
21	فيه	75	711
22	بعد	75	630
23	بين	74	788
24	ديالها	72	1,159
25	ولا	71	3,177
26	ما	71	2,182
27	كانت	71	847

	Word	Unique editors	Total uses
28	حتى	68	290
29	اللي	67	2,939
30	شي	64	693
31	ب	63	2,079
32	كل	63	226
33	قبل	62	544
34	كاين	62	515
35	عندو	58	616
36	مدينة	58	542
37	ل	56	1,267
38	عندها	56	570
39	عند	56	391
40	كبير	56	338
41	الناس	55	663
42	عليها	55	278
43	نهار	54	394
44	غير	54	355
45	جوج	53	508
46	ولكن	53	220
47	محمد	52	545
48	أول	52	484
49	وهي	52	167
50	عبد	51	357
51	أو	50	601
52	بن	50	346
53	عليه	49	325
54	وهو	49	162

55	دار	48	539
56	حيت	48	498
57	معروف	48	315
58	سميتو	48	276
59	مليون	48	248
60	الله	48	247
61	المغرب	48	209
62	كانو	47	440
63	لا	47	232
64	إلى	47	105
65	منها	46	463
66	تزاد	46	186
67	وحدة	45	320
68	كبيرة	45	244
69	ومن	45	153
70	ناس	44	2,912
71	دولة	44	1,537
72	بلي	44	433
73	حساب	44	395
74	بدا	44	296
75	ليه	44	246
76	ديالهم	44	245
77	تحت	44	207

78	سنة	44	124
79	عدد	43	512
80	بدات	43	260
81	مرة	43	239
82	ضد	43	234
83	بعض	43	147
84	عن	43	100
85	العالم	43	97
86	فاش	42	294
87	غادي	42	268
88	تاريخ	42	243
89	يكون	42	233
90	كاينة	41	1,773
91	هوما	41	348
92	كاينين	41	341
93	هادشي	41	293
94	تقريبا	41	268
95	ماشي	41	244
96	ليها	41	241
97	يونيو	41	224
98	مجموعة	41	221
99	بيها	41	199
100	سميتها	41	193

DOI: 10.14746/linpo.2025.67.1.5

Digitalisation and polycentricity in Moroccan Arabic: Complexity and heterogeneity in linguistic practices

Adil Moustaoui Srhir

Universidad Complutense de Madrid
adilmous@ucm.es | ORCID: 0000-0002-0770-943X

Abstract: The main aim of this paper is to analyse the notion of polycentricity (Blommaert 2005; 2010) and its application to Moroccan Arabic as a possible polycentric language. I understand polycentricity in the sense that any environment in which human beings gather and communicate is by definition polycentric and may therefore contain more than one identifiable centre of power (Blommaert 2005). A second objective is to explore how the digitalisation of Moroccan society and the increased use of social media and digital platforms are central to the phenomenon of polycentricity in Moroccan Arabic. In order to do so, I will examine a corpus of uses in various social media and digital platforms as a highly revealing manifestation of polycentricity and heteroglossia in Moroccan Arabic. The analysis reveals how practices and voicing are not subject to constraints or a single norm that defines what is 'irregular' or plain, thereby creating a rather complex polycentric system in Moroccan Arabic that affects both the language itself as linguistic system and its patterns of use.

Keywords: Moroccan Arabic, polycentricity, digitalisation, (de)normalisation and heteroglossic linguistic practices

1. Introduction

The advent of digital platforms, including social media, instant messaging, and online communities, has transformed language dynamics in Moroccan society, giving rise to novel modes of expression and communication styles using Moroccan local languages. This new model of communication has particularly favoured the use of Moroccan Arabic (hereinafter MA) and Amazigh in both private and public media. This is partly attributable to the privatisation of broadcast media and telecommunications, and has affected the way Moroccan society conceives mother tongues as both a language and a resource. This paper focuses especially on MA, the use of which has expanded over the last decade on

national TV channels and mediated computer communication, and which is now used intensively on social media such as Facebook, YouTube, TikTok, Instagram and other virtual platforms. These dynamics trigger linguistic, social, political and identitarian changes that need to be explored and analysed through research that looks beyond the large-scale of social and linguistic elements and a micro gap related to individual interactions, speech patterns and communicative dynamics and by addressing the focus on complexity, diversity, multiplicity and polycentricity in multilingual Morocco. The main aim of this paper is therefore to analyse how the digitalisation of Moroccan society and the increased use of social media and digital platforms are central to the phenomenon of polycentricity in MA. I investigate how the creation of polycentricity in MA could play a role in the (de)normalisation and (de)centralisation (Milroy 2001) of this variety in a new Moroccan language market and regime. My interest lies essentially in the conceptualisation of polycentricity in sociolinguistic and digital terms applied to MA. My analysis is based on a language corpus taken from various social media and digital platforms that includes influencer profiles as a highly revealing manifestation of polycentricity in MA. My analysis follows the concept of polycentricity (Blommaert et al 2005; Blommaert 2010) and its application to MA. A further objective is to consider the potential utility of the concepts of scale and discourses in digital media in relation to both, the polycentric character of MA and subsequently its multi-layered repositioning in the sociolinguistic regime. In this sense, polymedia analysis (Madianou 2015) provides an insightful tool and method for understanding and interpreting today's digital interaction, considering digital media not as isolated platforms but rather as an interconnected multimodal (Kress 2010) and polycentric environment. For the purpose of this article, MA represents not just a linguistic variety or local language but also a semiotic resource with several uses: (i) to communicate and build a translocal and transnational linguistic community; (ii) as a commodity in material, virtual and transnational markets; and (iii) as a tool for creative identities and subjectivities performance (Hachimi 2016).

This paper is structured as follows: Section 2 describes the theoretical background of the notion of language as a resource connecting scale and sociolinguistic regime. Section 3 focuses on polycentricity and complexity applied to the context of Morocco and MA. Section 4 discusses digital platforms as spaces for interaction and communication in digital environment. Section 5 offers quantitative data related to the use of the internet and social media in Morocco. Section 6 provides an overview of the research methodology and the data collection process. Section 7 is dedicated to the analysis with a special focus on polycentricity in MA associated with heteroglossic repertoires and its role in the development of complex linguistic practices and norms. The article ends with a series of conclusions in Section 8.

2. Theoretical starting points: Language as a resource and digitalisation

Language is a political and social force and is therefore managed, regulated, regimented and policed (cf. Blommaert 1999; Kroskrity 2000; Heller 2010; Costa 2019). Changing the position of a language or language variety in a specific sociolinguistic regime

involves the repositioning of this variety within a language hierarchy, usage domains, and user profiles, and thus has ideological, sociopolitical and economic implications. In this sense, the idea of repositioning evokes a particular social and political conceptualisation of language in the sense that it is considered a resource and cultural capital (Bourdieu 1982; Heller 2006; Blommaert 2007a). As Pietikäinen (2010: 81) argued, "this view of language moves away from the notion that languages are whole, bounded, complete entities and leans toward a more heteroglossic understanding that speakers draw on linguistic and semiotic resources." Heller (2010: 350) pointed out that "the emergence of the idea of language as resource, and of the new forms of its production and circulation in current market conditions, does challenge dominant ideologies of language in important ways [...] we can draw on Bakhtinian notions of heteroglossia, Foucauldian ideas of discourse, and Bordieuan ideas of markets to reimagine language as communicative resources socially constructed in uneven, unequal distributed social spaces." In addition, and as Blommaert (2010: 155) argued, languages are also placed resources, that is, a given language may be perfectly functional and valuable in certain positions but may become dysfunctional elsewhere. Furthermore, languages are structured in ways that make sense not only in a specific social and ideological context, but also in particular chronotopes. Sociolinguistic dynamics of change also create differences in the values of languages because several functions are reallocated to them as accommodated resources. However, the social value of a language is a social construct that involves a series of historical conditions, and hence historical power relations and hierarchies between languages are strongly bound to social, political and economic factors. Indeed, the study of MA provides an insight into the longitudinal processes of change in the role and function of this variety, as well as its current values in Moroccan society and its sociolinguistic regime (Moustaoui 2023).

The relocation of MA as a new linguistic resource has led to a repositioning of this variety within the Moroccan sociolinguistic order (Moustaoui 2016). This repositioning marks a decisive shift in the power relations established by the state's sociolinguistic, political and economic regime, but also in the historical relations of dominance and subordination between languages, known historically as diglossia, in the Moroccan linguistic market (Ennaji 2011) with Modern Standard Arabic (hereinafter MSA) being the high variety and MA the low variety. Consequently, MA as a language variety and a resource is used not only over a large territory, spaces and sites, but also by diverse and polycentric communities with considerable linguistic variation and several centres.

3. Polycentricity and complexity in MA practices

Blommaert et al. (2005) described the situation of a language with several interacting centres, each with their own codified norms by using the term polycentricity and they considered it as a theoretical metaphor based on the idea that linguistic practices are shaped by various sociolinguistic norms, authorities or centres. A centre is thereby defined not by its physical materiality, but by the way community connectivity is organized. Blommaert et al. (2005) have associated these normative centres with particular places

in the urban space, such as schools, mosques or local shops. Blommaert (2010) also points out that every environment in which human beings meet and communicate with each other is almost by definition polycentric. Consequently, all language practices, in the case of both standard and non-standard languages, are polycentric in nature. Blommaert (2010: 40-41) also points out that "polycentricity is a key feature of interactional regimes in human environments: even though many interaction events look 'stable' and monocentric (e.g. exams, wedding ceremonies), there are as a rule multiple – though never unlimited – batteries of norms to which one can orient […]. This multiplicity has been previously captured under terms such as 'polyphony' or 'multivocality'." However, languages unite through their use, and they separate through the development of linguistic norms and orders of indexicality and linguistic variability that the same speakers can identify. As Blommaert (2007b: 117) argued, "Indexical order, thus, is the metapragmatic organising principle behind what is widely understood as the 'pragmatics' of language."

In a particular kind of polycentric environment, people have a range of overlapping navigating points in which they can choose from the various varieties, accents, registers, and discourses that can be found and inscribed on the physical and virtual linguistic landscape. To investigate this multi-layered and polycentric character of MA, a conceptual move of this article is the consideration of MA as a space of spaces, linked not just to a single geographical place, but rather to a ranking of registers classified in terms of value or adequacy that are constructed situationally in interaction and that emerge in multiple, differentiated contexts. Polycentric environments like MA also involve multiple scales, each ascribing different values and functions to this variety. For the purpose of this article, scale can be helpful in seeing how the process of repositioning MA is accompanied by processes of hierarchical ordering: shifts in scale trigger shifts in value and function (Blommaert 2007b: 126) in which different discourses and their significations are networked together. Yet, in a polycentric environment such as MA, any extension of the use of a language into areas where it was not previously used (education, the media, literature, the institutional domain, the linguistic landscape, etc) leads us to reflect on the relationship of this language with the other varieties present in the linguistic market. This also raises important questions regarding language policy, planning and management: i) How does polycentricity in MA relate to heteroglossic language practices and the construction of local subjectivities and new forms of belonging? ii) How does this phenomenon work in digital social networks? iii) How the speakers mobilise distinct resources from their polycentric language repertoires?

In a polycentric MA context, where the logics and regimes of different regions, trans-local communities, domains and contexts overlap, people have needs and opportunities, or desires to use several genres, registers, and accents in their everyday lives. This is increasingly related to economic, cultural, and sociolinguistic features, generating a language commodity. In the Moroccan context, the strengths that influence the cultural and linguistic needs of MA linguistic communities are linked not only to subjectivities and socio-political identities such as digitalization and the emergence of social media and digital communication. The resulting kind of contemporary polycentric environments present

new challenges to MA in digital social media and the digital polycentricity and linguistic practices they have applied to date.

The other interrelated shift is how the transformation from oral to written form created new contexts and spaces, especially within the media, indicating the modern value of MA as a resource and its fit with the requirements of the modern information society (Michalski 2019). In this sense, these spatial scales move MA from its traditional, geographic and institutional places into new, often hybrid spaces, in which semiotic practices are created, linked to Morocco's new economy (Moustaoui 2019), activism (Suárez Collado & Moustaoui 2023), urban music (Boum, 2012; Moreno Almedia, 2017) and popular music (Hachimi 2022). Within this scale, at least two changes are taking place: (i) from local to global and (ii) from the use of mainly one variety or register to polycentric and hybrid practices, thereby guaranteeing what Heller (2003) calls the 'commodification of languages'. This commodification has created new values for MA in social media.

4. Digital language communities of practice

Digital platforms are spaces of performance, creativity, linguistic innovation and diversity, where communities of practice and speakers experiment with new styles and functions of language and express different forms of belonging and identities, attitudes, and opinions. Digital platforms are also considered spaces of linguistic contact and negotiation in which several users interact and communicate with other linguistic communities and ideologies and they navigate within linguistic norms and pressures that they may contest or renegotiate. In addition, by substituting physical interaction for a digital alternative, several platforms have introduced a novel avenue for building a sense of community and regimes of communication, transcending established political and administrative boundaries (Quercia et al. 2012; Takhteyev et al. 2012) in order to diffuse opinions, information and ideas across space. This has led to the emergence of a new type of geospatial information that defies the conventions of authoritative or volunteered geographic information and introduces a new virtual order of interaction. In this sense, Danesi (2017) examined the nuanced impact of digital communication on interpersonal relationships and language usage, shedding light on how digital interactions influence language patterns and shape communication dynamics. Within these new dynamics, various connected virtual communities are formed around issues related to a specific virtual space such as the polycentric one. This process generates a large polycentric structure, which comprises not only speakers from a specific state within its established legal or political norms and borders but also other communities, people and speakers outside and under the jurisdiction of a particular state (Androutsopoulos & Vold Lexander 2021). In a rapidly changing linguistic, virtual and human landscape, the projection of linguistic practices onto the virtual space thus spans multiple zones, sites, contexts and semiotic cartographies.

So, my aim in this paper is to identify the spatial distribution of global Moroccan communities in social media and the manner in which MA is regimented, used and

capitalised within and from its digital polycentric environment. In this sense, the notion of 'digital polycentricity' (Androutsopoulos & Vold Lexander 2021) will help us to examine how participants use digital media and choose a range of varieties, registers, accents and discursive practices that are related to their multiple forms of belonging and subjectivities.

5. Digital platforms and multilingualism in Morocco

The use of the internet in Morocco is increasing steadily, with 34.47 million users in January 2024, reflecting a penetration rate of 90.7%, according to the *Digital 2024: Morocco report*. Social media are also widely used, with 21.30 million users, 55.7% of the population and a total of 51.36 million cellular mobile connections were active in Morocco in early 2024. A notable demographic trend shows that the majority of social media users are men (58.3%), while women make up 41.7%. The same report notes that the number of social media users in Morocco is equivalent to 55.7 percent of the total population and 61.5 percent of the total internet user base used at least one social media platform (Datareportal 2024).

The growing use of the internet and social media in Morocco, driven by mobile connectivity, provides young users with considerable opportunities and highlights the importance of digital media as a key avenue for engaging Moroccan consumers. In addition, the rise of the internet, along with the country's complex multilingual landscape, has significantly transformed the way information is constructed and consumed, particularly through digital and social media. Young Moroccans – the most digitally literate segment of the population – are at the forefront of this transformation, actively engaging with various digital platforms using local languages (Zaid & Ibahrine 2011).

This has amplified the role of digital media in shaping public opinion and spreading information, making it a critical factor in the broader social and political dynamics of Moroccan society and serving as a means "to counter the hegemonic discourse of traditional media" (El-Issawi 2016: 6). Furthermore, Morocco's local multilingualism adds an additional layer of complexity to the digital landscape. Traditionally, news, information and media were available principally in MSA and French, the languages of power and education in Morocco. However, the proliferation of new media technologies has brought about a process of language democratisation, resulting in an increased use of MA or Darija[1] and Amazigh. This linguistic shift in the digital space not only reflects Morocco's social and cultural plurality but also has significant implications for accessibility and inclusivity. As more digital platform outlets offer content in MA and Amazigh, populations who may not be as fluent in MSA or French now have greater access to news and information. This is particularly important for rural and less literate segments of the population, as well as for the Amazigh-speaking community, which has historically been marginalised in terms of its linguistic representation in the media (Lafkioui 2024). Furthermore, the spread of digital networks and social media allows Moroccan diaspora

[1] Moroccan Arabic is also called Darija.

communities to stay connected with information from home, wherever they may be. Digital platforms also foster spaces for political discourse and social mobilisation, especially among younger generations who are more adept at using these tools in which the local languages are relocated and capitalised. This trend underscores the growing role of digital technologies in defining the public sphere in contemporary Moroccan society. In the digital area in particular, MA has gained significant traction in the media landscape, reflecting a greater leaning towards embracing the linguistic realities of everyday communication in Morocco. One of the key reasons for MA's growing presence in digital platforms is its ability to connect with a broader audience on a more personal and relatable level. By using MA as a lingua franca, media outlets and content creators can communicate directly with Moroccans, breaking down the barriers created by the more formal and less commonly spoken MSA or the often elite-driven use of French. In addition, a number of intellectuals and media figures have also emerged as proponents of MA in digital platforms. So, the growing presence of MA in the media reflects its practical role as a language of mass communication and social cohesion and as a potential semiotic resource. By gaining traction, MA not only empowers a broader spectrum of Moroccan society but also challenges traditional linguistic hierarchies that have long shaped the country's media and intellectual landscape.

6. The present study

In recent years, I have observed and participated in several social media. In particular, I have observed the accounts of a number of influencers that have a notable presence in the Moroccan public sphere and on digital platforms. Furthermore, I have noted how the social and linguistic context has influenced the way Moroccan influencers build and consolidate their social media presence.

This study employs two methodological tools: polymedia analysis and digital observation. As I argued previously, my analysis touches on two critical concepts in sociolinguistics: complexity and polycentricity. The ongoing diversification of interpersonal media is effectively theorised with the concept of polymedia (Madianou & Miller 2013). To this purpose, I will conduct a polymedia analysis – an insightful tool and method for understanding and interpreting today's digital interactions that considers digital media not as isolated platforms but as an interconnected polycentric environment. Polymedia analysis also refers to the diverse media choices available for communication, from text to video and voice, and how individuals use these to manage social/digital interaction and relationships. As Madianou and Miller (2013: 170) state: "It is an 'integrated structure' within which each individual medium is defined in relational terms in the context of all other media." This methodology will help us analyse how MA linguistic communities use the collective landscape of tools to fit their specific communicative needs, often combining various semiotic artefacts to create meaningful ways of connecting and interacting in the digital space. Polymedia analysis also implies a shift in agency and responsibility for users, highlighting how linguistic choices within this environment can reflect social relationships, cultural norms, and individual ideologies. This methodology is espe-

cially relevant for exploring how Moroccan linguistic communities balance complex social interactions using MA and its various accents and registers by blending various platforms to communicate appropriately across different contexts and spaces, including transnational ones. Lastly, a polymedia analysis also focuses our attention on how MA usage patterns and linguistic choices shape communicative norms and expectations in a particular kind of polycentric environment. As for the digital observation, it should be noted that social media are considered a research site of data collection method for the study combined with a content analysis of the profiles of numerous Moroccan influencers. This approach considers social media as a particular type of hybrid site, replete with videos, texts, visual artefacts and/or connections between speakers, communities and entities. For the purpose of this study, I have examined the profiles of five influencers: two women and three men, whose details are given in Table 1.

Table 1. Influencer profile information

Influencer's Name	Sex and age	Digital platform	Profession
Hicham Chaibi	Male 32	https://www.tiktok.com/@hicham_chaibi93	Creator content in MA
Mayssa Salama Ennaji	Female 54	https://www.facebook.com/elMayssa	President of @NeoMoroccans ThinkTank
Said Abarnous	Male 49	https://www.facebook.com/Abarnouss	Artist
Choumicha Chafay	Female 53	https://vm.tiktok.com/ZGe2b4Dor/	Chef, TV presenter and content creator
Ahmed Asid	Male 63	https://www.tiktok.com/@ahmed.aassid	Amazigh and political activist. Professor of Philosophy

I chose a diversified linguistic corpus composed of a number of videos representing examples of spoken MA in terms of registers and subvarieties. It should also be noted that some of the examples also contain words from other linguistic varieties, namely MSA, Amazigh or French. I have decided to choose a phonetic spelling for the MA script that is more in line with the pronunciation and linguistic realisation of the speakers.[2]

7. Results

7.1. MA as individual talent and accent

The first example is from TikTok. It is a video posted by the Moroccan influencer Hicham Chaibi, who lives in Spain. The video reports an act of racism he and his friends experienced in the centre of Madrid. The video received 44,000 views and 115 comments.

[2] The audios are available under this link: https://drive.google.com/drive/folders/1vC3ffJ3RVDgxo1ONnoTUBHQDcYthgiP7.

Most of the comments are written in Arabizi, an informal writing method that uses Latin characters and numbers to represent Arabic sounds. The title of the video, which is written in MA, reads "what happened to me with racism and the police" and features a police sticker and a Spanish flag (Figure 1 and Example 1).

Figure 1. Title of the video posted on TikTok[3]

Example 1

اسمع اسمع هاد اللقطة وقعت لي أنا ف هاد البلاد هادي. خارج كانتمشى أنا وْالدري ف واحد الپلاصة سميتها الپارطِ بيخا، بحال إلا گلتي المدينة القديمة ف كازا، غا هي داك الشي، داك الشي عندهم زايد داك الشي ناضي. وْهاد الپلاصة ف الليل كاتكون عامرة فيها الديسكوطيكات، الناس كايكونو شاربين خارجين كانتمشاو دزنا من واحد الزنقة، تكروازينا مع خمسة ديال الدراري معهم درية. واحد الشوية هو الدّري صاحبها هو يمشي عند الدّري گال له علاش كاتشوف لي ف صاحبتي؟ حنا عارفين ديك الپلاصة شوية فيها الزكا. منين وصلت عندو أنا هو بدا كايغوت هاديك اللعيبة علاش كاتشوف لي ف صاحبتي. أنا بغيت نبرد وْگلت له لا أنا ماشفتش لك ف صاحبتك وْلا والو. واحد الشوية وأنا نشوفو كايطليق لي واحد الضربة هنايا، جاب الله ما جابهاش لي ف عيني. أنا مع ضربني وأنا نجهل. تا گلت ديك اللعيبة واش تا انا جاي محشي تحت باطو جاي قاطع خمسطاش الكيلومتر مصفطّني رفولي، أنا راه عايش معه عا ديسير أنايا راه متشحال هادي غا مستنف. عاخليتو عازكيتو واحد الشوية وتلاحيت عليه وأنا نعطيه واحد الكروشي صافي وْهي تنوض تمك، تبلبلات. لدراري معه كلشي هزو القراعي ديال الزجاج واحد عكاز القراعي بداو كاي من حدا وجهي پاق پاق. كانسمع عا الدري گال لي ديك اللعيبة علّق.

asmaʿ asmaʿ hād l-laqṭa ūqʿat lī āna f hād l-blād hādī. ḫārəž kā-ntməššа āna ʾu d-dərrī f wāḥəd l-blāṣa smīthā alpārṭe byeḥā, bḥāl ilā gultī l-mdīna l-qadīma f kāzā, ġā hiyya dāk š-šī, ʿəndhum zāyəd dāk š-šī nāḍī. ʾu hād l-blāṣa f l-līl kā-tkūnʿāmra fīhā d-dīskūṭēkāt, n-nās kā-ykūnū šārbīn ḫāržīn kā-ntməššāw dəznā min wāḥəd z-zənqa, tkrwāzīnā mʿa ḫamsa dyāl d-drārī mʿāhum dərrīya. (...) wāḥəd š-šwīya huwwa d-dərrī ṣāḥəbhā huwwa yimšī ʿənd d-dərrī gāl-lih ʿlāš kā-tšūf lī f ṣāḥəbtī? ḥnā ʿārfīn dīk l-blāṣa šwīya fīhā z-zkā. (...) mnīn wṣalt ʿəndū āna huwwa bdā kā-yġuwwət hādīk l-lʿība ʿlāš kā-tšūf lī f ṣāḥəbtī. āna bġīt nbərrəd ʾu gəlt-lih lā āna mā-šuftš lik f ṣāḥəbtək wullā wālū. wāḥəd š-šwīya ʾu āna nšūfū kā-yṭlīq lī wāḥəd ḍ-ḍarba hnāyā, žāb llah mā žābhāš lī f ʿīnī āna mʿa ḍrabnī ʾu āna nəžhəl. tā gelt dīk l-lʿība wāš tā āna žāy məḥšī taḥt bāṭū žāy qāṭəʿ ḫəmsṭāš əl-kīlūmetr mṣafṭenī refūlī, āna rāh ʿāyəš mʿah ʿā dīsīr anāyā rāh mətšḥāl hādī ġā məstanəf. (...)ʿā ḫəllītū ʿā zəkītū wāḥəd š-šwīya ū-tlāḥīt ʿlīh ʾu āna nəʿṭīh wāḥəd l-krōšē ṣāfī ʾu hiyya tnūḍ təmmak, tbəlblāt. d-drārī mʿa kull šī hazzū l-qrāʿī dyāl z-zžāž wāḥəd ʿukkāz l-qrāʿī bdāw kāy mən ḥdā ūžhī pāq pāq. kā-nəsməʿ ʿā d-dərrī gāl-lī dīk l-lʿība ʿalləq.

[Listen, listen – this incident (scene) happened to me here in this country. I was out walking with my friend in a place called la Parte Vieja, kind of like the Old Medina in Casa-

[3] https://www.tiktok.com/@hicham_chaibi93/video/7293279855792098593

blanca, but there it's even wilder and more intense. At night this area is packed, full of nightclubs, drunk people out on the streets. We were walking and passed through an alley, ran into five guys with a girl. After a bit, the girl's boyfriend goes up to my friend and says, "Why are you looking at my girlfriend?" You know that place has a reputation, a bit rough. When I got there, he started shouting that same thing: "Why are you looking at my girlfriend?" I wanted to cool things down, so I told him, "No, I didn't look at your girlfriend or anything." A moment later, I see him throw a punch at me here -thank God it didn't land in my eye. He hit me, and I lost it. I thought to myself: "Man, I came here high, under the boat, I crossed the strait during 15 kilometers, my nerves are fried, I've been living off nothing but dessert for a while, I'm exhausted." So, I let him swing a little, then I lunged at him and gave him a hook. That's when everything blew up. His friends all grabbed glass bottles and one of those bottle sticks, and they started swinging right by my face, bang, bang. I could only hear my buddy saying that line: "Comment on it."]

In this video, Hicham Chaibi speaks the Darija variety from Casablanca, which is also his native variety. Casablanca is Morocco's largest city and economic centre, where modernity and cultural diversity are the norm. The city's accent has become a symbol of these positive qualities and is commonly perceived as representative of open-mindedness and connectedness to the world of modernity. In Moroccan communication culture, the Casablanca accent is associated with cosmopolitan lifestyle and Moroccan youth who are exposed to international influences and contacts and are aware of global trends. The identity of these young, cosmopolitan groups is marked by modernity, yet at the same time retains a connection to the country's roots and customs, giving it an authentic and unique character. The accent also reflects this cultural balance: while retaining the basis of MA, it integrates many words and expressions from other languages, especially French, as seen in the example:

(1)

- الكروشي *l-krōšē* Fr. *le crochet* 'the punch'
- الديسكوطيكات *d-dīskūṭēkāt* Fr. *les discotèques* 'the discotheques'
- تكروازينا *tkrwāzīnā* Fr. *se croiser* 'to cross'
- باطو *bāṭō* Fr. *bateau* 'ship'
- رفولي *refūlī* Fr. *refouler* 'to expel'

Indeed, the way Casablanca speaks symbolises the evolution of a city that is at the same time a translocal and super urban space. The youthful and pragmatic nature of the accent has proved to be highly influential, spreading all over Morocco through music, film, and the media and it is becoming a reference for the outward-looking lifestyle that Casablanca represents. In addition, communication using this variety and its accent is perceived as direct and pragmatic. Indeed, the Casablanca variety of MA is characterised by its focus on common expressions and clarity, as well as its uninhibited and distinctive style:

(2)

- هاد اللقطة *hād l-laqṭa* 'this incident'
- داك الشي عندهم زايد *dāk š-šī ʿandhum zāyed* 'what they have is cool'
- البلاصة شوية فيها الزكا *blāṣa šwīya fīhā z-zkā* 'a place with a lot of racket'

- عازكيتو واحد الشوية *ʿā zəkītū wāḥəd š-šwīya* ‘I made him stop’
- گال لي ديك اللعيبة *gāl-lī dīk l-lʿība* ‘he told me the plan’

This facilitates localised and situated communication in the digital space, based on immediacy and closeness to speakers using an accent that conveys a strong sense of belonging to Casablanca and its large language speech community. The Casablanca variety is also perceived as ‘cool’ due to its connections with other global settings, reflecting progress and new opportunities. In this example, we can see how the use of MA as a linguistic practice is an action that rejects any essentialist conception in the representation of MA, in terms of both its spelling as we can see in figure 1 and voicing. In this sense, language processing based on this example is seen as an individual talent (Heller 2010), in line with Duchêne and Heller’s (2012) vision of language as a word of work and as a wordforce (parole d’oeuvre). This cultural model and linguistic ideology also see language as a resource that serves the pragmatic nature of communication imposed by the new local, regional, national or globalised economy. Finally, Hicham’s use of a recognizable Casablanca accent that reinforces the image of modernity that appeals to the younger generation, thereby enabling him to attract more followers and secure more views of the posts and videos he shares on his profile.

7.2. MA and the linguistic expertise

The second example is a video posted by the journalist and president of @NeoMoroccans think-tank Mayssa Salama Ennaji on her Facebook page. In recent years, Mayssa has built up a reputation in traditional and social media as a young journalist and think-tanker who opposes the government of Akhnouch. She is especially noted for her oratory skills, both in MSA and MA, attracting considerable attention from outside Morocco and earning her a large social media following.

Example 2

فطبعاً هاد الفيديو درتو، أولاً باش نمسيكم كاملين بالخير وباش نگول شكرا إل ساكنة العيون وأهلنا ف الجهات الجنوبية اللي تفاعلو مع الفيديو الأخير ديالي اللّي، جبدت فيه غير نقيطة من بحر د الفساد اللي كاين تما ومسكوت عنه. راه حياتنا كاملة و حنا كانهضرو على هاد المناطق. تانهضرو على أخنوش أربعة وعشرين ساعة على أربعة وعشرين ساعة، على الغلاء أربعة وعشرين ساعة على أربعة وعشرين ساعة، حتى خرجات مجلة جون أفريك سمتني رأس حربة معارضة أخنوش. يعني الفراغ ف البلاد فراغ المعارضة وصل إل درجة سماو مدوّنة هي رأس حربة المعارضة. يعني ماتگولوش لي ناس ديال الداخل وَهضري علينا. نوبة الناس ديال الصحرا. علاش؟ لأن هاداك الفاسدين تمّا السياسة ديالهم هي عزل الساكنة. هادوك الناس ديال المناطق الجنوبية معزولين. حنا ما كانهضروش عليهم. ما تانهضروش على العيون وطانطان و أسا الزاگ والسمارة وغيرو وغيرو من الجهات ديال الأقاليم الجنوبية. علاش هادوك الناس تعزلو حتى إل ديك الدرجة؟ خوفا لا نهضرو على معاناتهم إلا يگول لك انفصاليين. لا ما يمكنش. ما يمكنش تكون هادي تهمة جاه تاتعلقوها إل عباد الله. حنا وطنيين وكاين فساد. أنا بغيت نجاوب على شي بعضين حيت اللي مضرور و اللي واكل الدّق واللي واكل العصا واللي واصلة له إل العضم.

fā-ṭabʿan hād l-vīdyō ʾawwalan dərtū bāš nmassīkum kāmlīn bi-l-ẖēr ʾu bāš ngūl šukran l-sākīnat l-ʿīyūn ʾu ʾahəlnā f əl-ğihāt l-ğanūbīya llī ṭfāʿlū mʿa l-vīdyō l-ʾaẖīr dyālī llī žbədt fīh ġēr nqīṭa min bḥar d əl-fāsād llī kāyin təmmā uʾ məskūt ʿanhu. (...) rāh ḥiyāt-

nā kāmla uʾ ḥnā kā-nhəḍrū ʿla hād l-manāṭəq. (...) tā-nhəḍrū ʿala aḫənnūš arbʿa ʾu ʿəšrīn sāʿa ʿala arbʿa ʾu ʿəšrīn sāʿa, ʿla l-ġalāʾ arbʿa ʾu ʿəšrīn sāʿa ʿla arbʿa ʾu ʿəšrīn sāʿa, ḥətta ḫaržāt mağallat žēn afrīk səmmātnī raʾs ḥarbat mūʿārāḍat aḫənnūš. yaʿnī l-farāġ f əl-blād farāġ əl-muʿāraḍa ūṣel l-daražat səmmāw mudawwena hīya raʾs ḥarbaṭ al-mūʿārāḍa. yaʿnī mā-tgūlūš lī nās dyāl d-dāḫil wa-ḥḍrī ʿlīnā. nuwbat n-nās dyāl ṣ-ṣaḥrā. ʿlāš? liʾənna hādāk l-fāsdīn təmma s-sīyāsa dyālhum hiya ʿazl s-sākīna. (...) hādūk n-nās dyāl l-manāṭəq l-ğanūbīya məʿzūlīn. ḥnā mā kā-nhəḍrūš ʿlīhum. mā tā-nhəḍrūš ʿla l-ʿīyūn ʾu ṭānṭān ʾu əssā z-zāg ʾu s-smāra wa-ġērū wa-ġērū mə-l-ğīhāt dyāl l-ʾaqālīm l-ğanūbīya. ʿlāš hādūk n-nās tʿəzzlū ḥətta l-dīk d-dārāğa? ḫawfan lā nhaḍrū ʿala mūʿānāthum ilā īgūl-lək infiṣālīyīn. lā mā yimkənš, mā yimkənš llī ḫruğ dwā ʿla ḥaqqū infiṣālī. mā yimkənš tkūn hādī tuhmat žāh tā-tʿəllqūhā l-ʿibād llāh. ḥnā wāṭānīyīn ʾu kāyn fāsād. āna bġīt nžāwəb ʿla šī baʿḍīn ḥīt llī məḍrūr ʾu llī wākəl d-daqq ʾu llī wākəl l-ʿaṣā ʾu llī wāṣla lih l-l-ʿaḍam.

[So of course, I made this video first to greet you all, and to say thank you to the people of Laayoune and our kinsfolks in the southern regions who reacted to my last video – where I only touched on a single drop from the ocean of corruption that exists there and is being silenced. During all our lives, we've been talking about these regions. We talk about Akhannouch 24/7, about high prices 24/7, to the point where Jeune Afrique magazine called me the "spearhead of the opposition to Akhannouch." That means the political void in the country, the void of opposition, reached the point where they called a blogger the spearhead of the opposition. So don't say to me, "people from the interior, speak about us." Now it's the turn of the Saharan people. Why? Because those corrupt ones over there, their policy is to isolate the population. Those people in the southern regions are isolated. We don't talk about them. We don't talk about Laayoune, Tan-Tan, Assa-Zag, Samara, and so on from the southern provinces. Why have those people been isolated to that extent? Out of fear that if we talk about their suffering, they'll accuse us of being separatists. No, that's impossible. That can't be a ready-made accusation to hang on people. We are patriots, and there is corruption.]

The first aspect to note in this example is the level of MA used, which navigates between two scales, MSA and urban MA, the last associated with the cities of Rabat and Sale in particular as we can see in these examples:

(4)

▪ أولاً	*ʾawwalan*	'first'
▪ ساكنة العيون	*sākīnat l-ʿīyūn*	'the population of Laayoune'
▪ مسكوت عنه	*məskūt ʿanhu*	'the silenced'
▪ فراغ المعارضة	*farāġ əl-muʿāraḍa*	'the opposition vacuum'

which reflect a normative character of the use of lexicon and manifests a far more national and less regional or local character of the variety used. On the other hand, it should be noted that linguistic expertise is a fundamental element in the treatment of language and patterns of language use.

To a certain degree, this indicates that there could be a process of linguistic levelling, manifested in Mayssa's use of typical MSA expressions such as:

(5)

▪ تهمة جاه	*tuhmat ǧāh*	'false accusation'
▪ رأس حربة المعارضة	*raʾs ḥarbat al-mūʿārāḍa*	'opposition leader'
▪ خوفا	*ẖawfan*	'fearing'

This tendency to linguistic levelling is used to bring MA closer to MSA. Indeed, as a content creator in Arabic, Mayssa adopts these traits and linguistic features close to MSA which reflect authenticity and an ideological stance indicates her alignment with the language of intellectuals in Morocco (Irvine & Gal 2000) and a cultural resonance of her voice as a Moroccan intellectual woman (Hachimi 2017). This kind of practice not only generates a sense of differentiation but also strengthens the creation of a register with norms and a sense of identity and belonging among her followers, who identify with the shared linguistic practices, choices and a single cultural model; in other words, the implicit ways in which the knowledge of society's language is created, organised, and managed at various stages. The other aspect in this example is the notion of stance or positionality from which Mayssa produces her discourse. Despite being physically located in the southern city of Layoune, Mayssa is attempting to connect and interact with a large audience located all over the country because as she said her posted video has feedback from various people situted in different regions. This positionality creates a spatiotemporal and translocal scale that connects her with many other communities throughout Morocco. As she points out in the video, she has received comments asking why she does not condemn the corruption that exists in other regions of Morocco. This information demonstrates the ability of digital communication to overcome physical barriers and attract followers from various territories, not only because of the quality or relevance of the content, but also because of the way people communicate and connect with their audience and the sense of proximity it generates.

7.3. Heteroglossia in action and MA

The third example is a video posted by the journalist and humourist Said Abarnous on his Facebook page. He defines himself as a digital creator, journalist and artist. Said Abarnous is originally from the city of Al Hoceima in the Rif region. He has a university degree in linguistics and communication, and his PhD thesis was dedicated to the history of theatre in the Riffian variety of Amazigh. In his posts and videos, this uses several languages, first and foremost, MA, MSA, and Riffian and occasionally French. However, as I was analyzing the content of his accounts on Facebook, YouTube and Instagram, I noted that most of his posts and videos were in MA. Said is also famous for his videos and posts on Algerian politics, the country's president and relations with Morocco. The transcript below (Example 3) is from the video featuring the Algerian national football team and its poor performance during the latest edition of the African Cup of Nations (2024).

Example 3

السلام عليكم. المنتخب الجزائري بالأمس يسجل نتيجة التعادل واحد الواحد. تماما بحال المنتخب المصري. دارو ديك الدخلة ديال حنا واعرين ؤ ف التالي وقع التعادل. بزاف تاع الخوت من الجزائر، جزائريين ما عجبتهومش النتيجة، ؤمازال ما لقاوش شي تفسيرات. أنا دبا ف إطار أنني نعاونكم أخوتي خوتي لزرجليان جبت لكم شوية متاع الأسباب. أولا بالماضي ف إطار التضامن مع المصريين ؤ المنتخب نتاع مصر گالك ما يمكنش أنا نربح ؤهم يتعادلو. المصريين تعادلو ؤ فلتو من الزندقة إذا حتى يانا غادي نفلت من الأنگالا ؤ غادي نجيبها تعادل ؤمريضنا ما عندو باس. تمتا من يذهب أبعاد ؤ كايگولك بأن القضية عندها علاقة مع الأرضية ديال الملعب. كيفاش؟ كاتعرفو بأنه تقارير صحفية، قادة من قناة الشروق ؤصحفيين آخرين گالو بأن الكاف نبهت بالماضي باش ما يبقاش يمشي يزور الملعب دبا وساعة، لأن هادي عشر أيام ولا تلت أيام ولا خمس أيام ، بالماضي كان كايمشي مع الصباح كايضرب دورة كايطّل على الملعب ؤ مع الضهر ؤ مع العُصر، يعني كان كايمشي دبا وساعة.

s-salāmu ʿalaykum. al-muntāḫab əl-ğazāʾīrī bi-l-ʾams yusəžžil nātīžat t-taʿādul wāḥəd l-wāḥəd. tāmāman bḥāl l-muntāḫab əl-miṣrī. dārū dīk d-dəḫla dyāl ḥnā wāʿrīn ʾu fə-t-tālī wqəʿ t-taʿādul. bəzzāf tāʿ əl-ḫūt mə-l-ğazāʾir ğazāʾīrīyīn mā ʿəžbāthumš n-nātīža, ʾu māzāl mā lqāwš šī tafsīrāt. āna dāba fʾīṭār ənnī nʿāwənkum a-ḫūtī ḫawtī lezaržəlyɛ̃ žəbt līkum šwīya mtāʿ l-ʾasbāb. ʾawwalan bəlmāḍī fʾīṭār t-tāḍāmun mʿa l-maṣrīyīn ʾu l-muntāḫab ntāʿ maṣr gāllək mā yimkənš āna nərbəḥ uʾ humma yitʿādlū. l-maṣrīyīn taʿādlū uʾ fəltū mən əz-zandāqā iḍān ḥətta iyyānā ġādī nəflət mən əl-ʾangālā ʾu ġādī nžībhā taʿādul ʾu mərīḍnā mā ʿəndū bās. (...) ṯəmmatā man yaḍhabu ʾabʿād ʾu kā-ygūllik biʾanna l-qāḍīya ʿəndhā ʿalāqa mʿa l-ʾarḍīya dyāl l-məlʿab. kīfāš? kā-tʿərfū biʾanna taqārīr ṣaḥafīya, qāda mən qanāt š-šurūq ʾu ṣaḥafīyīn āḫrīn gālū biʾanna l-kāf nəbbhāt bəlmāḍī bāš mā yibqāš yimšī yizūr l-məlʿab dābā w-sāʿa, liʾənna hādī ʿəšr iyyām wullā təlt iyyām wullā ḫəms iyyām, bəlmāḍī kān kā-yimšī mʿa ṣ-ṣbāḥ kā-yḍrəb dūra kā-ytəll ʿla l-məlʿab ʾu mʿa ḍ-ḍhur ʾu mʿa l-ʿaṣər, yəʿnī kān kā-yimšī dābā w-sāʿa.

[Peace be upon you. Yesterday, the Algerian national team recorded a 1–1 draw, exactly like the Egyptian national team. They made that big entrance of "we're strong," and in the end, it was a draw. Many brothers from Algeria, Algerians, weren't happy with the result and still haven't found clear explanations. Now, in the spirit of helping you, my Algerian brothers, I've brought you a few reasons. First, Belmadi, out of solidarity with the Egyptians and their national team, said to himself: "I can't win while they draw. The Egyptians drew and escaped trouble, so I'll also escape the trap and bring home a draw – and no harm done." Then there are those who go further and say the matter has to do with the pitch conditions. How so? You know that according to press reports – even from El Chourouk channel and other journalists – CAF warned Belmadi not to keep visiting the stadium constantly. Because for ten days, or maybe three or five days, Belmadi had been going there morning, noon, and afternoon, making the rounds and checking out the field.]

The most striking feature of this example is how the use of MA makes conversations immediate and authentic, demonstrating how the language is adapted to Morocco's football culture.

[4] All French words are transcribed the way they sound in Moroccan Arabic and are in bold.

In this sense, MA includes a repertoire of expressions and specific jargon that allows speakers to communicate experiences and emotions immediately with different degrees of humour such as:

(6)

▪ مريضنا ما عندو باس	*mərīḍnā mā ʿəndū bās*	'the sick person gets better'
▪ فلتو من الزندقة إذا حتى يانا غادي نفلت من الأنگالا	*fəltū mən əz-zandāqā iḍān ḥətta iyyānā ġādī nəflət mən əl-ʾangālā*	'they were saved and I too was saved from the stumble'
▪ نجيبها تعادل	*nžībhā taʿādul*	'I draw a tie'
▪ دارو ديك الدخلة ديال حنا واعرين	*dārū dīk d-dəḫla dyāl ḥnā wāʿrīn*	'they started at the beginning by "we're strong"'

This football jargon, common in sports media, uses slang, expressions, and informal grammatical structures. Said's choice of register reflects a level of social inclusion: by using MA with its idiomatic nuances and jargon, he constructs a collective identity that not only celebrates the sport but also underpins a sense of belonging. This creates a space for exchange in which speakers and followers can express both their knowledge and their sense of humour, all in a register and discourse that is familiar and accessible to most Moroccans. Secondly, the video is a parody of the Algerian national team's elimination from the African Cup of Nations, and attempts to satirise the Algerian coach's explanations and excuses. This type of content has great potential to resonate with followers and communities, as it connects with common emotions and expectations surrounding football in Morocco. Said employs comical exaggeration, referring to absurd factors or unusual excuses to justify the result. This type of humour frequently resorts to the use of gestures, expressions and an exaggerated tone, and plays with Darija slang and idioms to add greater authenticity and proximity. Significantly, in this parody, MA acts as a potential resource that guarantees a kind of collective catharsis, allowing followers to share and laugh at the disappointment, while exploring the exaggerated reasons. This is also an example of heteroglossia in action: MA becomes a tool for expressing multiple voices within a playful and open context such as a football parody. Heteroglossia emerges in the ability of the language to incorporate and interweave various norms, registers, accents, jargons and cultural nuances within a single act of communication, reflecting its polycentric and diverse character. In this specific context, the heteroglossic and polycentric character of MA not only adds to the humorous content but also allows for a communication that escapes the control of normativity as a purist and monoglossic ideology.

7.4. MA between neutrality and switching & mixing

The fourth example is a video posted by the famous Moroccan chef Choumicha on her TikTok page. The video is the preparation of a pasta recipe with salmon and vegetables. The video was viewed by more than 17,000 people, shared by more than 6,000 and had 290 comments, mostly written in MA with Arabic script and a Latin transcription, as well as a number of comments in French. In her TikTok profile, Choumicha defines herself as a TV presenter of cultural programmes.

Example 4

مؤخرا دقت واحد المقارونيا طايبة بالسلمون أو متحضرة بالسلمون. أنا كنت طلبتها غير **بالصمون** ولكن منين جابوها لي تفاجأت باللي فيها بعض المكونات ؤلا بعض الخضر اللي زادو داك **الصمون** واحد النكهة رائعة: بورو خزو ؤ البسباس. شفتو معيَ درنا داك البورو مع شوية ديال الزبدة حتى تشحر ؤ تعسل، من بعد حكينا خزو، كانحكوه ف الجهة ديال الحكاكة الغليضة، كانحكو البسباس. بالنسبة **لپاط** كانخدو طنجرة ديال الما كانديرلو الما حتى كايغلى ؤ نديرو معه الملحة ؤ من بعد كانزيدو الكراڤاطَ أولا لفارفيل. نخليوهم كايطيبو تقريبا تلاتة أرباع الوقت المحدد على العلبة أو على البواطة، ؤ كانحيدوهم. داك الشي اللي بقى لهم راه حنا غادي نكملوه ف لصوص. بالنسبة لصوص شفتو معيَ درنا البورو حتى طاح مع ديك الزبيدة، درنا معه البسباس ؤمن بعد ضفنا عليه خيزو، غادي نزيدو دبا شوية ديال الما ديال السليق، شوفو عندكم زوج ديال الإختيارات إما إلا عندكم المرق ديال الدجاج كاديرو المرق ديال الدجاج، إلا ماكانش المرق ديال الدجاج متوفر كاديرو الما ديال السليق ديال الماقارونية. أنا غادي نستعمل المرق حيت أنا من عشاق المرق سواء كان ديال الدجاج أو ديال اللحم ؤ لا ديال السمك. نخليو هاك دا غير يغلي غادي نرجّع له الكمية ديال الماقارونية اللي أنا باغي نحضر، غادي نضيف الكمية اللي خاصني ديال **الپاپيون** أولا **الكراڤاطَ** غادي نحركوها ؤ نضيفو لها كمية ديال **الصمون**، نحركو مزيان نخليو داك النكهة ديال **لفومي** أو ديال الدخان اللي كاين ف الحوت تطلق ف المرق أولا ف لصوص. من بعد كانضيفو لكريم فرش القشدة الطرية مهمة جداً بالنسبة لهاد الوصفة هادي. ؤ ف الأخير كانديرو شوية ديال الفروماج ، كانحركو باش يلاه كاتشرب الماقرونية داك المكونات كلها ؤ كاتبدا كاتبان لنا **لصوص** أنها كاتختر. صافي دبا غادي نزلها ف طابق ديال التقديم، غادي نرش عليها شوية ديال الجبن مبشور أولا **الفروماج** محكوك ![1]

*muʾaḥḥaran dəqt wāḥəd l-maqārōnīyā ṭāyyiba bə-ṣ-ṣalamōn aw məṭḥaḍḍra bə-ṣ-ṣalamōn. āna kunt ṭlabthā ġēr bə-**ṣ-ṣomõ** wālākin mnīn žābūhā lī tfāžaʾt bəllī fīhā baʿḍ l-mūkaw-wīnāt wullā baʿḍ l-ḥuḍar llī zādū dāk **ṣ-ṣomõ** wāḥəd n-nəkha rāʾiʿa: bōrō ḫizzū ʾu l-bəs-bās. šəftū mʿāya dərnā dāk l-bōrō mʿa šwīya dyāl z-zəbda ḥətta tšḥər ʾu tʿəssəl, mən bəʿd ḥəkkīnā ḫizzū, kā-nḥəkkūh f ž-žīha dyāl l-ḥəkkāka l-ġlīḍa, kā-nḥəkkū l-bəsbās. Bə-n-nisba l-**lēpāṭ** kā-naḫdū ṭanžara dyāl l-mā kā-ndīr lū l-mā ḥətta kā-yiġla ʾu ndīrū mʿah l-məlḥa ʾu mən bəʿd kā-nzīdū l-krāvāṭa awlā le-fārfēl. nḫəllīhum kā-yṭībū təqrīban tlāta arbāʿ əl-wuqt əl-muḥaddad ʿla l-ʿulba aw ʿla l-bwāṭa, ʾu kā-nḥiyydūhum. dāk š-šī llī bqā lhum rāh ḥnā ġādī nkammlūh f lāṣōṣ. Bə-n-nisba lāṣōṣ šəftū mʿāya dərnā l-bōrō ḥətta ṭāḥ mʿa dīk z-zbīda, dərnā mʿāha l-bəsbās ʾu mən bəʿd ḍəfnā ʿlīh ḫīzzū, ġādī nzīdū dāba šwīya dyāl l-mā dyāl s-slīq, šūfū ʿəndkum žūž dyāl l-iḫtīyārāt immā ilā ʿəndkum l-maraq dyāl d-džāž kā-ddīrū l-maraq dyāl d-džāž, ilā mā kānš l-maraq dyāl d-džāž mətwəffər kā-ddīrū l-mā dyāl s-slīq dyāl l-māqārōnīya. āna ġādī nəstaʿməl l-maraq ḥīt āna mən ʿušāq l-maraq sawāʾan kān dyāl d-džāž awlā dyāl l-lḥəm awlā dyāl s-samak. nḫālliw hākdā ġēr yiġlī ġādī nrəžžəʿ lih l-kəmmīya dyāl l-māqārōnīya llī āna bāġa nḥəḍḍər, ġādī nḍīf kəmmīya llī ḫāsṣānī dyāl **l-pāpīyõ** ʾawlā **l-krāvāṭa**, nḥərrkūhā ʾu nḍīfū lihā kəmmiya dyāl ṣ-ṣomõ, nḥərrkū məzyān nḫālliw dāk n-nəkha dyāl **lfymɛ** awlā dyāl d-duḫḥān llī kāyn fə l-ḥūt ṭṭləq fə l-maraq awlā f **lāṣōṣ**. mən bəʿd kā-nḍīfū **lākrēm frēš** l-qišda ṭ-ṭarīya mūhīmma židdan bə-n-nisba l-hād l-wuṣfa hādī. ʾu fə l-ʾāḫīr kā-ndīrū šwīya dyāl l-frōmāž, kā-nḥərrkū bāš yəllāh kā-tšrəb l-māqārōnīya dāk l-mūkawwīnāt kullhā ʾu kā-təbdā kā-tbān linā **lāṣōṣ** ənnāhā kā-təḫtər. ṣāfī dābā ġādī nnəzzəlhā f ṭābaq dyāl t-təqdīm, ġādī nrəšš ʿlīhā šwīya dyāl l-žubn məbšūr awullā **l-frōmāž** maḥkūk.*

[Recently, I tried some pasta cooked with salmon, or more precisely prepared with salmon. I had ordered it just with salmon, but when they brought it to me, I was surprised to find some added ingredients, some vegetables that gave the salmon a wonderful flavor: leek, carrot, and fennel. So, as you saw with me, we sautéed the leek with a bit of

butter until it softened and caramelized. Then we grated the carrot – using the coarse side of the grater – and also grated the fennel. For the pasta: we take a pot of water, bring it to a boil, add salt, and then add farfalle or penne. We let them cook about three-quarters of the time indicated on the package, then remove them. The rest of the cooking will be completed in the sauce. For the sauce, we started with the leek sautéed in butter, added the fennel, then the carrot. After that, we pour in some of the pasta cooking water. You have two options here: either use chicken stock if you have it, or if not, just use the pasta cooking water. I'm going to use stock, because I'm a big fan of it – whether chicken, beef, or fish. We let it simmer a little, then add back the amount of pasta we want to prepare. We stir it in, then add the salmon pieces. We mix well so that the smoky, fish flavor blends into the stock or sauce. After that, we add fresh cream – very important for this recipe. Finally, we add some grated cheese, stir so that the pasta absorbs all the flavors, and the sauce begins to thicken. That's it – now we transfer it to a serving dish, sprinkle some more grated cheese on top.]

What is striking in this video is the level of neutrality of the MA used by Choumicha in lexical, morphosyntactic and even phonetic terms in the sense that the register used seems to be unrelated with any region or city. This feature also confers a degree of homogeneity on MA that distances it from any ethnolinguistic or regional territoriality or features. This video also clearly displays Choumicha's ability to switch registers by creating a norm based on repeating expressions and words from different varieties: MA, MSA and French:

(7)

▪ متحضرة بالسلمون	*məthaḍḍra bə-ṣ-ṣalamōn*	'prepared with salmon'
▪ أنا كنت طلبتها غير **بالصمون**	*āna kunt ṭləbthā ġēr bə-**ṣ-ṣomõ***	'I ordered it with salmon'
▪ من بعد كانضيفو **لكريم فرش** القشدة الطرية	*mən baʿd kā-nḍīfū **lākrēm frēš** l-qišda ṭ-ṭarīy*a	'then we add the fresh cream' 'I draw a tie'
▪ الجبن مبشور **أولا الفروماج** محكوك	*l-žubn məbšūr awullā **l-frōmāž** maḥkūk*	'grated cheese' 'grated cheese'

In this example, we can see how the words 'salmon', 'cheese' and 'cream' are used in both Arabic and French. This practice is attractive and distances itself from any purist view of the language, eliminating boundaries and creating a unique style. It is a language choice that can indicate a high social status in digital spaces. Nevertheless, this style generates comments and even interpellations from the audience. Below are such examples in MA:

(8)

الحمد لله وليت كنسمع القشدة الطرية وأنا فرحانة ماشي كيف بكري كنت غير كنسمعها كتشدني السخانة

al-ḥamdu li-llāh wullīt ka-nsməʿ l-qišda ṭ-ṭarīya ʾu āna farḥāna māšī kīf bəkri ġēr ka-nsməʿha kā-tšəddnī s-saḥāna

'Thank God now that I listen to l-qišda ṭ-ṭarīya (crème fraiche) I get happy; before I listened to it and I would get hot flushes'.

la crem frech 😂 = *chomicha*.
'fresh cream 😂 = chomicha'.

The latter expression refers, in a humorous way, to Choumicha's use of French when speaking of fresh cream.

Switching between or mixing varieties and registers as in these examples reflects a particular tone:

(9)
بعض الخضر اللي زادو داك **الصمون** واحد النكهة رائعة: بورو خزو ؤ البسباس.
baʿḍ l-ḫuḍar llī zādū dāk ṣ-ṣomõ wāḥəd n-nəkha rāʾiʿa: bōrō ḫizzū ʾu l-bəsbās
'some vegetables that gave the salmon a very good flavour: leek, carrot and fennel'

It attempts to highlight a personal style that manifests a distinct and hybrid Moroccan modern identity manifested through a cultural fusion in everyday life and in social relations, yet which is also relatable to diverse audiences and followers who are comfortable with translanguaging, crossing and mixing languages. This is clear from the reactions to the video in the comments. This form of expression enables Choumicha to engage with a global digital culture and is also a way of asserting her distinct voice and discourse as a Moroccan woman, within broader Arab-speaking digital spheres. In addition, switching between languages allows Choumicha to communicate in a more precise, open and expressive manner. Certain concepts or emotions may be easier to convey in MA, while others might be clearer or more widely understood if they are in MSA or French, e.g. *les pattes* 'the pasta', *fumé* 'kipper'. This flexibility enables users to communicate efficiently, opting for the language that best conveys their intent and thereby accommodating a particular style in a digital polycentric communication environment. Finally, this model of managing Moroccan languages has a strong implication for content creation, marketing, and policy-making in the sense that content creators must strategically mix languages and also follow a multilingual content strategy which can expand potential new markets and spaces.

7.5. MA and *luġat al-muṯaqqafīn* 'the language of intellectuals'

The final example is a video posted by Ahmed Assid on his TikTok account. Assid is an Amazigh activist, intellectual and High school teacher of philosophy. He is famous for his discourse as an advocate of a secular society and as well as an individual rights activist. In recent years, Ahmed Assid became even more prominent in the public media sphere for criticising the re-Islamisation of Moroccan society and the degradation of ethical principles. The video is titled: Why is it that a talent from Morocco is restricted in his own country and is a shining success in other countries?

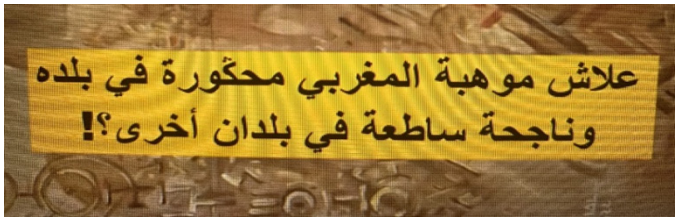


Figure 2: Title of the video posted on TikTok[5]

Example 5

يتما ستما، السيدات والسادة المشاهدات وٱلمشاهدين الأعزاء. كانحييكم ؤ كانرحب بكم ف هاد الڤيديو اللي الموضوع ديالو واحد السؤال دائما كايطرحوه عليّ المغاربة ؤ بالخصوص الشباب. علاش المغربي لمّا كايكون الداخل، الموهبة ديالو ماكاتبانش؟ لما كايكون الداخل ديال بلادو ف المغرب ؤ كايكون محگور كايكون مهمش. ولكن لمّا كايخرج من البلد لأي بلد آخر خارج الوطن كاتبان الموهبة ديالو ؤ كاينجح ؤ إلا غير دلك. أشنو هو السبب؟ السبب هو مانسميه تحرير الطّاقات. ؤ لهدا ما كانوش كايعتبرو أن هناك تلميد كسول ؤ تلميد مجتهد. هاد التنائيات ما عندهومشاي. إذا لمّا الطفل مند الطفولة المبكرة كايتعلم يحرر الطاقات ديالو كايكبر بهاد التربية المتحررة اللي ماكايقمع فيها حتى حاجة الداخل ديالو بالعكس كل ما عندو كايعطيه كايعطيه إل بلادو إل الأسرة ديالو ؤ المجتمع ديالو. حنا للأسف هدا هو العطب اللي عندنا بالضبط ، لمّا كاتكون ميول ف الطفل كانقمعوها فيه، كانرضوها فيه الداخل. أو أننا ما كانعطيوهش الإطار أو المحيط اللي يقدر يعبر فيه على هاديك الطاقات بل كايخليها عندو الداخل. هادا بالإضافة إلى نسبة الهدر المدرسي الكبيرة اللي كاتجعل أنه واحد العدد د الناس منين كايتلاحو خارج المدرسة على أنه ما صالحين ال والو ؤ ماعندهم تكوين ما عندهم عمل كايضيعو. كايضيعو لأنه كايعتاقدو بأنه ما عندهم أي طاقة ؤ أي شيء لفعل والو لأن المجتمع ابتداء من ؤ الأسرة المدرسة علمهوم بأنهم لا شيء أنهم والو بينما هم لا كل فرض ف المجتمع إلا وعندو شي حاجة اللي يمكن يعطيها. خاصنا غير نقلبو عليها ؤ نعرفوها منذ الطفولة، ما كانديروش هاد المجهود.

a-yitmā stmā s-sayyidāt u-s-sāda l-mūšāhidāt ʾu l-mūšāhidīn l-ʾaʿizzāʾ. kā-nhəyyīkum ʾu kā-nrəḥḥəb bikum f hād l-vīdyō llī l-mawḍūʿ dyālū wāḥəd s-suʾāl dāʾiman kā-yiṭərḥūh ʿlīya l-maġārba ʾu bə-l-ḫuṣūṣ š-šabāb. ʿalāš l-məġrībī ləmmā kā-ykūn l-dāḫil, l-mawhiba dyālū mā kā-tbānš? ləmmā kā-ykūn l-dāḫil dyāl blādū f l-məġrib ʾu kā-ykūn məḥgūr kā-ykūn muhammaš. (...) wālākin ləmmā kā-yiḫruž min al-bālād l-ʾayy bālād āḫar ḫāriž al-wāṭān kā-tbān l-mawhiba dyālū ʾu kā-yinžaḥ ʾu ilā ġēr dālik. šnū huwwa s-sābāb? s-sābāb huwwa mā nusammīh taḥrīr aṭ-ṭāqāt. l-buldān n-nāhīḍa ʾu l-buldān l-mūtāqād-dīma kā-tḥrəṣ bəzzāf ʿla ənnāh tḥarrar aṭ-ṭāqa dyāl l-farḍ mundu ṭ-ṭūfūla. (...) wa li-hādā mā kānūš kā-yiʿtabrū anna hunāka tilmīd kasūl ʾu tilmīd muztāhīd. hād t-tunāʾīyāt mā ʿəndhumšāy. idan ləmmā ṭ-ṭifl mundu ṭ-ṭūfūla l-mūbakkīra kā-yitʿəlləm yiḥarrar ṭ-ṭāqāt dyālū kā-yikbər b-hād ət-tərbīya l-mutaḥarrira llī mā kā-yiqmaʿ fīhā ḥətta ḥāža l-dāḫəl dyālū bə-l-ʿaks, kull mā ʿəndū kā-yəʿṭīh, kā-yəʿṭīh l-blādū l-l-ʾusra dyālū ʾu l-mūztāmaʿ dyālū. ḥnā lə-l-ʾasaf hadā huwwa l-ʿāṭāb llī ʿəndnā bə-ḍ-ḍabṭ, ləmmā kā-tkūn muyūl f ṭ-ṭifl kā-nqəmʿūhā fīh, kā-nrəḍḍūhā fīh lə-d-dāḫel. ʾaw ənnanā mā kā-nəʿṭīwəhš l-ʾīṭār ʾaw l-mūḥīṭ llī yiqdər yiʿabbar fīh ʿla hādīk ṭ-ṭāqāt bal kā-yḥəllīhā ʿəndū l-dāḫel. hādā bi-l-iḍāfā ilā nisbat əl-hadr əl-madrāsī l-kəbīra llī kā-təžʿəl ənnu wāḥəd l-ʿādād də n-nās mnīn kā-yitlāḥū ḫārəž l-madrāsa ʿla ənnahu mā ṣālḥīn l-wālū ʾu mā ʿəndhum təkwīn ʾu mā ʿəndhum ʿamal kā-yḍīʿū. kā-yḍīʿū li-ʾənnahu kā-yiʿtāqdū ʾənna mā ʿəndhum ayyi

[5] https://www.tiktok.com/search/user?q=ahmed%20assid&t=1749219564573.

ṭāqa ʾaw ayyi šəyʾ li-fiʿl wālū li-ʾənna l-mūžtāmaʿ ibtidāʾan mə-l-ʾusra ʾu-l-madrāsa ʿəllmūhum bi-ʾənnahum lā šayʾ ənnahum wālū baynamā humma lā, kull farḍ f əl-mūžtāmaʿ ʾillā wa-ʿəndū šī ḥāža llī yimkən yiʿṭīhā. ḫəṣṣnā ġēr nqəllbū ʿlīhā ʾu nʿərfūhā mundu ṭ-ṭūfūla, mā kā-ndīrūš hād l-məžhūd.

[Ladies and gentlemen, dear viewers, greetings and welcome to this video. The topic today is a question that Moroccans – especially young people – often ask me: Why is it that when a Moroccan is inside the country, their talent doesn't show? When they're in Morocco, they're ignored, marginalized. But when they leave the country, go abroad, suddenly their talent emerges, they succeed, and it's a different story. What's the reason? The reason is what I call the releasing of talent. That's why, in those places, they don't consider there to be a "lazy student" versus a "hard-working student." These dualities don't exist. When a child, from early childhood, learns to free their potential, they grow up with this liberated upbringing where nothing inside them is repressed. On the contrary, everything they have to give, they give – to their country, their family, and their society. Unfortunately, this is exactly the flaw we have: when a child shows an inclination, we suppress it. We bury it inside them. Or we don't give them the framework, the environment where they can express those energies – instead, they keep them bottled up. On top of this, the high rate of school dropouts means many people get thrown out of the education system with no skills, no training, no job – and they end up lost. They're lost because they've come to believe they have no potential, nothing to offer, since society – starting with the family and school – taught them they are nothing, they're worthless. But in reality, every individual in society has something they can contribute. What's needed is simply to look for it and recognize it from childhood. And we don't make that effort.]

Assid begins the video in Amazigh in order to say hello يتما ستما *a-yitma stma* 'friends' before switching to MSA to welcome:

(10)
السادات والسادة المشاهدات ؤالمشاهدين الأعزاء
s-sayyidāt u-s-sāda l-mušāhidāt ʾu l-mušāhidīn l-ʾaʿizzāʾ
'ladies and gentlemen and dear viewers'

Starting the post in Amazigh is not just a communicative act but also a political one in the sense that the speaker revealed his Amazigh linguistic and socio-cultural identity. This practice situates Assid as a bi/multi-lingual speaker addressing a broad ethnolinguistic and cultural communities that includes Amazigh, MSA and MA. Secondly, Assid's example shows us that the register he used in MA is that of an expert linguist, as he chooses a variety that represents what is called the language of intellectuals. This notion of 'the language of the intellectuals' implies in itself the presence of one or more standard linguistic references. In the case of MA, this is MSA and the prestigious Moroccan linguistic varieties.

(11)
السبب هو مانسميه تحرير الطّاقات
s-sābāb huwwa mā nusammīh taḥrīr aṭ-ṭāqāt
'the motive is what we call releasing the talent'

إذا لمّا الطفل منذ الطفولة المبكرة
idan ləmmā ṭ-ṭifl mundu ṭ-ṭūfūla l-mubakkira
'then when the child from an early age'

This tendency in linguistic choices tends to shape vocabulary, pronunciation, grammar, and even cultural norms within this frame, thereby exerting an element of attraction, providing central standards or prestigious linguistic varieties. At the same, this practice often holds a higher social or cultural status. Assid also adopts terms, expressions, or a neutral accent that will raise his social prestige, which draws audiences from other language backgrounds and other linguistic communities:

(12)

▪ تلميد كسول ؤتلميد مجتهد	*tilmīd kasūl ʾu tilmīd mužtahid*	'bad student or diligent student'
▪ التربية المتحررة	*ət-tərbīya l-mutaḥarrira*	'liberating education'
▪ نسبة الهدر المدرسي	*nisbat əl-hadr əl-madrāsī*	'the school dropout rate'
▪ الإطار أو المحيط	*l-ʾīṭār ʾaw l-mūḥīṭ*	'the setting or environment'
▪ لا شيء	*lā šayʾ*	'nothing'

Similarly, the fact that MSA is still the reference language, will lead to a clear reduction in the linguistic differentiations between MA varieties and accents. This linguistic choice is related to a distinct economic and cultural social class, which affects the main selection criteria for a variety that represents a large Arabic-speaking area inside and outside Morocco.

8. Conclusion

In this paper I have re-conceptualised the notion of polycentricity from the physical setting to complex digital media environments for the purpose of identifying 'centres', that is, different patterns of language use and orientations for participants' digital language and practices. The examples analysed demonstrate a social and linguistic complexity in the way the Moroccan digital linguistic community manages the various spoken forms, varieties and accents of MA in digital media. The use of MA in a digital and polycentric environment such as social media, where English and other dominant languages often dominate, adds a layer of originality and authenticity that can be perceived as attractive and genuine. This reshapes the assumption that each individual navigates purposefully from a multitude of communicative options in their everyday media use. Rather than simply being a matter of technological convenience, choices of different semiotic artefacts, including text and voice reflect relational, cultural, ideological and situational preferences and the way these choices shape several communicative norms and expectations.

These choices act as signs that index particular social relationships, identities, or emotions. For instance, a post message might signal a casual or routine interaction, while a video voice could imply a need for immediacy or intimacy. In all examples, one of the norms employed is the conscious choice of the language variety, accent or register and its representation in terms of both individual and collective identity. Through varying combinations of communicative and linguistic practices, we have observed how the influencers manage resources to create an integrated style where MA as lingua franca plays an essential role in connecting with their audience and building a Moroccan digital and cultural identity.

If we look further at these examples, we can observe the diversity and polycentric character of the MA language system, which in turn ensures that multiple voices are heard and repositioned in the public and virtual space, thus generating language management and policies that are more inclusive and open, from the bottom-up scale.

Another important aspect for consideration is whether and how MA in digital communication is considered a distinct, independent and autonomous linguistic construction in which people have active agency. This means that the choices available to individuals for their own initiatives, visibilising their varieties or accents and mobilising them in the digital space, correspond to linguistic repertoires they know and use in their everyday interactions.

Furthermore, our polymedia analysis reveals that Morocco's norm of mixed language use on digital platforms is both a reflection of the nation's linguistic diversity and a creative expression of its dynamic identity. Through the flexible use of MA, MSA, French or Amazigh, Moroccans have crafted a distinct online language ecosystem that balances cultural pride, social status, and accessibility. This digital and heteroglossic multilingualism demonstrates the adaptability and resilience of Moroccan identity in the global digital era, affirming that language is not just a means of communication but a key element of cultural connection, personal expression, and social interaction.

Finally, the absence of a standard variety and an orthographic homogenised norm, or any national, state-imposed standard leaves MA a situation of free and open linguistic communication, which is uncontrolled, unmanaged and unplanned by institutions. This means that oral and written practices in MA are not subject to constraints or norms that define what is 'irregular' or simply 'wrong' from a linguistic and normative point of view, thereby creating polycentric, de (centralised) and de (normalised) environments for this variety.

References

Androutsopoulos, Janis, & Vold Lexander, Kristin. 2021. Digital polycentricity and diasporic connectivity: A Norwegian-Senegalese case study. Journal of Sociolinguistics 25(5). 720-736. https://doi.org/10.1111/josl.12518

Blommaert, Jan (ed). 1999. Language and ideological debates. Berlin – New York: Mouton de Gruyter.

Blommaert, Jan. 2005. Discourse: A critical introduction. Cambridge. Cambridge University Press.

Blommaert, Jan. 2007a. Sociolinguistic scales. Intercultural Pragmatics 4(1). 1-19.

Blommaert, Jan. 2007b. Sociolinguistics and discourse analysis: Orders of indexicality and polycentricity. Journal of Multicultural Discourses 2. 115-130. https://doi.org/10.2167/md089.0

Blommaert, Jan. 2010. The sociolinguistics of globalization. Cambridge: Cambridge University Press.

Blommaert, Jan & Collins, James & Slembrouk, Stephen. 2005. Spaces of multilingualism. Language & Communication 25, 197-216. https://doi.org/10.1016/j.langcom.2005.05.002

Boum, Aomar. 2012. Youth, political activism and the festivalization of hip-hop music in Morocco. In Weitzman, Maddy & Zisenwine, Daniel (eds.), Contemporary Morocco: State, politics and society under Mohammed VI, 161-177. London – New York: Routledge.

Bourdieu, Pierre. 1982. Ce que parler veut-dire. Paris: Éditions Fayard.

Costa, James. 2019. Introduction: Regimes of language and the social, hierarchized organization of ideologies. Language & Communication 66. 1-5. https://doi.org/10.1016/j.langcom.2018.10.002

Danesi, Marcel. 2017. Language, society, and new media: Sociolinguistics today. London – New York: Routledge.

Datareportal. 2024. Digital 2024: Morocco. (https://datareportal.com/reports/digital-2024-morocco) (Accessed 2025-05-04.)

Duchêne, Alexandre & Heller, Monica (eds.). 2012. Language in late capitalism: Pride and profit. London – New York: Routledge.

El-Issawi, Fatima. 2016. Arab national media and political change: Recording the transition. New York: Palgrave McMillan.

Ennaji, Moha. 2011. The promotion of Moroccan Arabic: Successes and failures. In Fishman, Joshua & Garcia, Ofelia (eds.), Handbook of language and ethnic identity, vol. 2: The success-failure continuum in language and ethnic identity efforts, 46-53. New York: Oxford University Press.

Hachimi, Atiqa. 2016. Moroccan artists 'blacklisted': Dialect loyalty and gendered national identity in an age of digital discourse. In Davis, Stuart & Soltan, Usama (eds.), Perspectives on Arabic linguistics 27, 123--150. Amsterdam – Philadelphia: John Benjamins.

Hachimi, Atiqa. 2017. Moralizing stances: Discursive play and ideologies of language and gender in Moroccan digital discourse. In Høigilt, Jacob, & Mejdell, Gunvor (eds.), The politics of written language in the Arab world: Writing change. 239-265. Leiden: Brill.

Hachimi, Atiqa. 2022. In the Middle East, it's cool to 'sing Moroccan': Ideologies of slang and contested meanings of Arabic popular music on social media. International Journal of the Sociology of Language 278. 107-131. https://doi.org/10.1515/ijsl-2022-0042

Heller, Monica. 2003. Globalization, the new economy, and the commodification of language and identity. Journal of Sociolinguistics 7(4). 473-492. https://doi.org/10.1111/j.1467-9841.2003.00238.x

Heller, Monica. 2006. (ed.) Bilingualism: A social approach. Hampshire: Palgrave.

Heller, Monica. 2010. Language as resource in the globalised new economy. In Coupland, Nicolas (ed.), Handbook of language and globalization, 349-365. Oxford: Blackwell.

Irvine, Judith T. & Gal, Susan. 2000. Language ideology and linguistic differentiation. In Kroskrity, Paul V. (ed.), Regimes of language: Ideologies, polities, and identities, 35-85. Santa Fe: University of New Mexico Press.

Kress, Gunther. 2010. Multimodality: A social semiotic approach to contemporary communication. London – New York: Routledge.

Kroskrity, Paul V. (ed.). 2000. Regimes of language: Ideologies, polities, and identities. Santa Fe: University of New Mexico Press.

Lafkioui, Mena. 2024. Pluricentricity, iconisation, and instrumentalisation of language in North Africa and its diaspora. In Máté, Huber & Meisnitzer, Benjamin (eds.), Pluricentric languages in Africa: Multilingualism and linguistic dehegemonisation in Africa and around the world. 15-38. Graz: PCL-Press.

Madianou, Mirca. 2015. Polymedia and ethnography: Understanding the social in social media. Social Media + Society 1(1). 1-3. https://doi.org/10.1177/2056305115578675.

Madianou, Mirca & Miller, Daniel. 2013. Polymedia: Towards a new theory of digital media in interpersonal communication. International Journal of Cultural Studies 16. 169-187.

Michalski, Marcin. 2019. Written Moroccan Arabic: A study of qualitative variational heterography. Poznań: Wydawnictwo Naukowe UAM.

Milroy, James. 2001. Language ideologies and the consequences of standardization. Journal of Sociolinguistics 5(4). 530-555.

Moreno Almedia, Cristina. 2017. Rap beyond resistance: Staging power in contemporary Morocco. New York: Palgrave Macmillan.

Moustaoui, Adil. 2016. Sociolinguistics of Moroccan Arabic: New topics. Frankfurt am Main & Berlin: Peter Lang.

Moustaoui, Adil. 2019. Transforming the urban public space: Linguistic landscape and new linguistic practices in Moroccan Arabic. Linguistic Landscape 5(1). 80-102. https://doi.org/10.1075/ll.18008.mou

Moustaoui, Adil. 2023. Landscaping in Moroccan Arabic: Language regimentation, practices and ideologies. In Al Rashdi, Fathiya, & Rao Mehta, Sandhya (eds.), Language and identity in the Arab world, 29-50. London: Routledge.

Quercia, Daniele & Capra, Licia & Crowcroft, Jon. 2012. The social world of Twitter: Topics, geography, and emotions. In Proceedings of the Sixth International Conference on Weblogs and Social Media, 298-305. Palo Alto, CA: AAAI Press.

Pietikäinen, Sari. 2010. Sámi language mobility: Scales and discourses of multilingualism in a polycentric environment. International Journal of the Sociology of Language 202. 79-101. https://doi.org/10.1515/ijsl.2010.015

Suárez-Collado, Ángela, & Moustaoui, Adil. 2023. Diáspora y crisis del Hirak: Análisis de las dinámicas de movilización y el discurso del activismo rifeño en Madrid. In Azaola, Bárbara & Desrues, Thierry & de Larramendi, Miguel H. & Planet, Ana I. & Ramírez, Ángeles (eds.), *Cambio, crisis y movilizaciones en el Mediterráneo Occidental*, 391-408. Granada: Editorial Comares.

Takhteyev, Yuri & Gruzd, Anatoli & Wellman, Barry. 2012. Geography of Twitter networks. Social Networks 34(1). 73-81.

Zaid, Bouziane & Ibahrine, Mohamed. 2011. Mapping Digital Media: Morocco. A report by the Open Society Foundations. London: Open Society Foundation.

DOI: 10.14746/linpo.2025.67.1.6

(Moroccan) Mixed Arabic in digital media: A comparative analysis of oral and written practices in Moroccan digital platforms and newspapers

Rosa Pennisi

University of Catania
rosa.pennisi@unict.it | ORCID 0000-0001-8653-7877

Abstract: This study aims to analyze the transition of (Moroccan) Mixed Arabic from oral to written productions in digital communication. The phenomenon of literacy in written Moroccan Arabic is an issue already observed by scholars, especially concerning the developments of standardization from below (Caubet 2017a-b, 2018; Miller 2017; Pennisi 2025, among others). Discussing the traditional perspective of the functionalist diglossic continuum of the Arabic language (Youssi 1992), this study aims to compare the morphosyntactic characteristics of (Moroccan) Mixed Arabic in oral and written productions through speakers/writers practices emerging from Moroccan digital media production. While several studies have already focused on the descriptions of dialectal elements occurring in different textual typologies, especially in Moroccan traditional media (Hoogland 2013, 2018; Brigui 2016, among others), the present study aims to analyze how a (Moroccan) Mixed Arabic style contributes to (informally) conventionalize a journalistic register, which is continuously employed in online newspapers and digital multimodal platforms. In order to achieve those goals, a corpus of written and oral data from 1) a digital Moroccan newspaper (*Goud*), 2) episodes of a talk-show, and 3) podcast, i.e. premeditated and unpremeditated communication in written and oral communication, has been contextually analyzed. The comparative analysis of oral and written practices shows that a mixed register of contemporary (Moroccan) Mixed Arabic is spreading in formal media Moroccan communication, serving variable discursive strategies.

Keywords: Mixed Written Media Arabic, Moroccan Arabic, digital Moroccan media, stylistic variation, Formal Non-Standard Arabic, Middle/Educated Moroccan Arabic

1. Introduction[1]

Abdellah Tourabi, a Moroccan journalist and previously the editorial director of Moroccan magazines *TelQuel* and *Zamane*, describes the journalistic language used in the talk show he presents, broadcasted on the semi-public free-to-air channel *2M*, *Ḥadīṯ*^{un} *maʿa aṣ-ṣaḥāfa*[2] / *Confidence de presse* ('Talking to the Press / Press confidences'), as follows:

> […] c'est impossible de tenir toute une émission en *darija pure*, c'est impossible. C'est-à-dire que *darija* c'est une langue qui se traite pas parfois à la fonction, qui se traite pas à décrire des concepts… par exemple si je vais utiliser l'expression « ḥuqūq l-marʾa » ça n'existe pas en *darija*. En effet, *ça fait partie de la darija*, je pense *pas qu'il y a une frontière qui sépare*. Alors, dans la presse il y a ce qu'on appelle la troisième langue, «*al-luġa at-tālita*». La troisième langue, c'est-à-dire un mélange entre *darija* et *fuṣḥā*, l'arabe classique. On essaie de trouver *une phase intermédiaire* entre les deux.[3]
> (Interview with Abdellah Tourabi, 2M journalist, February 22, 2018, Casablanca, in Pennisi 2025: 529).

From this brief excerpt, emerges the point that will be the central object of the present study. Tourabi, in fact, defines journalistic language as "*al-luġa al-ṯāliṯa*" 'the third language', or a mixture of *Fuṣḥā* (Classical/Standard Arabic) and *Darija*, Moroccan Arabic (MA). Tourabi understands the journalistic language, which is also used in the TV program he presents, as an intermediary stage between *Fuṣḥā* and *Darija*, and points out that the two cannot be clearly separated: in other words, there would be no clear boundaries that in journalistic language would delimit *Darija* from *Fuṣḥā* and vice versa. The present study considers the expression (Moroccan) Mixed Arabic what Tourabi called the intermediary stage between *Fuṣḥā* and *Darija*, and in particular the mixed language of the formal journalistic register widespread in Morocco.

Specifically, this study aims to analyze the linguistic and stylistic characteristics[4] of the Mixed Arabic used in the Moroccan media landscape, especially online, through

[1] The present article contributes to the ongoing research of the SABIRANET project (CUP E63C24001920006; ID SOE2024_0000078), funded by the European Union – NextGenerationEU under Italy's National Recovery and Resilience Plan (PNRR), Young Researcher 2024 – SoE line, administered by the Italian Ministry of University and Research (MUR).

[2] Transcriptions from oral data report the speeches as they are produced, including the superscript end inflection, when occurring. The program name *Ḥadīṯ*^{un} *maʿa aṣ-ṣaḥāfa*, is transcribed here including the nunation because it is pronounced the same way by the journalist Abdellah Tourabi during his talk-show.

[3] '[...] it's impossible to do a whole program in pure *Darija*, it's impossible. This means that *Darija* is a language that is not necessarily used to describe concepts... for example, if I use the expression "ḥuqūq l-marʾa", it doesn't exist in *Darija*. Indeed, it's part of *Darija*, I don't think there's a border separating it. Then, in the press there's a so-called third language, "al-luġa at-tālita". The third language, a mix between *Darija* and *Fuṣḥā*, Classical Arabic. We try to find an intermediate phase between the two.' Emphasis added. See the entire interview in Pennisi (2025: 525-538).

[4] n the present study we differentiate "linguistic characteristics" – which refer to morphosyntactic items – from "stylistic characteristics", by which we mean, instead, the ways of being of discourse, borrowing Fairclough's analytical approach in analyzing social and ideological representations of discourse (Fairglough

a comparative analysis of oral and written, premeditated and unpremeditated practices[5]. A comparative analysis of linguistic practices in both oral and written communicative modes allows us to move beyond the functionalist view of diglossia, which traditionally sees Standard Arabic as a variety used only in written productions and separated from colloquial varieties used only in ordinary oral communications, as originally pointed out by Ferguson (1959). The functionalist perspective of diglossia marks most of the later studies, albeit in terms of a linguistic continuum (Blanc 1960; Badawī 1973; Meiseles 1980; Walters 1996; Hudson 1992, 1994, 2002; Haeri 2000, 2003; Kaye 2001; Bousso-fara-Omar 2006; Khalil 2018, 2022), and even ideologically in the linguistic representations of Arabic-speaking speakers (Brustad 2017).

The present study, therefore, aims to analyze the communicative strategies implemented by speakers of Mixed Arabic in the context of premeditated and unpremeditated journalistic communication in Morocco. The oral data were analyzed taking into consideration Youssi's (1992) description of the variety of "Arabe Marocain Moderne", i.e. Modern Moroccan Arabic (MMA henceforth), as a reference model of the formal oral language in Morocco. As for written Moroccan Arabic, the topic of transition from oral to written *Darija* has been well researched (Aguadé 2006, 2012; Benítez-Fernández 2008; Hoogland 2013, 2018; Brigui 2016; Kebede & Hinds 2016; Caubet 2017 a, 2017 b, 2018; Miller 2017; Pennisi 2025), especially in the context of the most famous *Darija* editorial productions. This study makes one step further, since it approaches the issue of non-Standard varieties of Arabic in Morocco through the perspective of "mixed styles" which Mejdell (2006) used to describe the linguistic situation in Egypt, comparing both written and oral productions. This comparative approach makes it possible to observe communicative strategies, as well as the morphosyntactic and lexical features that have become conventionalized in the formal oral journalistic register (Youssi 1992). These features, along with their associated communicative strategies, tend to be reproduced and/or represented in written texts as well, where they are now increasingly widespread, particularly in digital communication.

Prior to presenting the results of the comparative analysis, this study contextualizes the theoretical framework of linguistic variation in Arabic, with attention also to pragmatic and stylistic aspects in both oral and written production. The study then outlines the discourse analysis methodology and the description of selected corpus, before presenting the results of the analysis.

2003:26-28). Through this perspective, in fact, it is possible to analyze the different registers and styles of Arabic, taking into consideration both the strictly linguistic dimension (description and distribution of the morphosyntactic features being examined) and the stylistic dimension (registers of Arabic) as well as its social meaning (i.e. what and how a given linguistic and stylistic choice impacts and creates social meaning). Moving beyond the almost always dichotomous distinction between standard Arabic and colloquial varieties, Mixed Arabic is here considered not as a variety of Arabic, but a style that represents a way of being of speech.

[5] While premeditated practices involve planned and structurally controlled language use, unpremeditated practices rely on spontaneous and more fluid speech patterns (Eagleson 1958: 153-154).

2. Theoretical contextualization

The concept of "Third Language" (*al-luġa al-ṯāliṯa*) evoked earlier by the Moroccan journalist's words, can be situated within the framework of linguistic variation. It was adopted by Arab writers in the 1950s and 1960s, particularly by the Egyptian playwright Tawfīq al-Ḥakīm (Avallone 2017). According to him, texts can be deliberately written in a Third Language appearing uniform to the norms of Classical Arabic, but performed as colloquial Arabic (Badawi 1985: 15-16). In other words, the written Third Language text is bivalent, because it can be read as Classical Arabic, but performed as non-Standard/colloquial Arabic as well. This sort of intermediary level, has long been described by scholars, not only as *al-ʿarabiyya al-wusṭā*, but also as ESA, or Educated Spoken Arabic (Mitchell 1986), which, however, was applied to only oral communication. As Badawi states:

> The very label itself, ESA, is variously applied to (a) spoken Arabic used by an educated Arab while conversing with an Arab from a different country, (b) spoken Arabic used by educated nationals of the same Arab country on subjects pertaining to their level of education and culture, and (c) the variety used by educated Arabic speakers coming from different Arab countries or from the same country to communicate with one another. (Badawi 1985:16)

Before digital communication, written (and journalistic) texts in non-Standard forms were linked only to sporadic ideologically and politically motivated editorial productions[6] (Zack 2014). Nowadays, however, texts written in non-Standard Arabic are extremely popular online and thus enjoy greater visibility. These texts show "a new, variable, pluralistic, multilayered concept of (one) Arabic in which boundaries are erased" (Mejdell 2022: 119).

In the case of Morocco, Youssi (1992: 23-25) described this intermediate register as "Modern Moroccan Arabic", which, according to him, represents the: "model of the Standard variety of orality […] the variety of formal situations of oral exchange […] the medium used on radio and television". Building on this definition, it is possible to identify at least three registers of Arabic in Morocco: Classical Arabic and/or Modern Literal Arabic for the communicative modality of the written language, Moroccan Arabic for the informal oral modality, and MMA, the intermediate variety, for the formal oral modality. His approach reflects a traditional functionalist perspective, which does not conceive of MMA as a written language. Today, however, digital Moroccan space is also witnessing a growing use of non-Standard Arabic for writing not only informal, but also formal texts.

Regarding the distinction between oral and written communication, the concept of "permanence" (Sebba 2012) is relevant. Sebba states:

> '*Permanence*', though the term is not completely satisfactory, is a factor which distinguishes many written texts from spoken ones. Texts in spoken genres by their nature tend not to be

[6] See for example, among the most famous ones, the editorial production of *Abù Naḍḍāra*, lit. 'the man with glasses', a satirical newspaper founded by Yaʿqūb Ṣanūʿ (1839-1912), where Egyptian Arabic was used in writing (Zack 2014). For more details in other countries see also Langone (2016) and Miller (2017).

> permanent, while texts in written genres, up to the age of the Internet, mainly had some degree of permanence. […] to fully understand language mixing in written texts we need to know not only by whom and for whom they are produced, but how they are produced and how they will be read. (Sebba 2012: 7-8)

The concept of "Permanence" from Sebba is, therefore, relevant for observing whether or not elements of oral practices are maintained in written practices in (Moroccan) Mixed Arabic, and is therefore useful for interpreting pragmatic functions through a "contextual analysis" (van Dijk 2008), in which the communicative context of language mixing is taken into account.

3. Methodology and corpus description

The comparative analysis was conducted taking into account what Adam, in the context of journalistic texts, defines as the three levels of pragmatic organization:

> Trois plans de l'organisation pragmatique peuvent être distingués : la visée illocutoire (valeur et force des actes de discours), la prise en charge énonciative des propositions et la représentation construite ou « monde » du texte.[7] (Adam 1997: 16)

These levels provide insight into how a text constructs meaning according to its communicative intentionality and its enunciative configuration. In Adam's model, the *visée illocutoire* refers to the overarching communicative purpose of the text, i.e. what guides the overall construction of the text, orienting the choice of discursive genres, linguistic registers and rhetorical strategies. The second level analyzes the position of the speaker with respect to the statements in the text (e.g., fully assuming an utterance, distancing oneself, delegating responsibility to someone else), influencing the credibility, apparent neutrality, or involvement of the text in relation to its content. The third level concerns the semantic-discursive construction of the content, i.e. the world represented by the text, presenting facts, actors, events, places, temporalities, causes, consequences, causal relations, implicit evaluations, and possible future scenarios.

These three aspects have been observed in the selected corpus of the present study to distinguish how linguistic variation impacts the illocutionary force of speech acts in (premeditated and unpremeditated) oral communication. This influence is mediated by different communicative strategies – such as 'grounding' (Khalil 2000), i.e. the process by which a text signals which information is central, new, or relevant and which information is incidental, taken for granted or background – and rhetorical tactics, including repetition, we/you polarization, and the use of religious or patriotic language (Mazraani 2008). Furthermore, the analysis shows how Mixed Arabic styles function as communicative tools to enhance the effectiveness of discourse in (written and oral) premeditated text.

[7] 'Three levels of pragmatic organization can be distinguished: the illocutionary aim (value and force of speech acts), the enunciative assumption of responsibility for propositions, and the constructed representation or "world" of the text.'

The corpus used in this study comprises three sub-corpora, each representing a different source of data. The first one includes the recording of the episode of the TV talk show broadcasted on the channel *2M*, *Ḥadīṯ*^{un} *maʿa aṣ-ṣaḥāfa*, titled *Waḍʿiyyat ḥuqūq al-ʾinsān wa-l-ḥurriyyāt bi-l-Maġrib rifqat Muṣṭafā al-Ramīd*, 'The situation of human rights and freedoms in Morocco, with Mustafa Ramid'.[8] The program is presented by Abdellah Tourabi who interviews the main guest of the episode, with the support of two other journalists, always different in each episode. The episode transcript represents the oral corpus of the unpremeditated communication (one hour of recording).

The second one consists of the oral corpus of premeditated communication and covers 11 episodes of the podcast produced and disseminated on the multimodal information platform *Hawāmiš* (hawamich.info), selected from the column, featuring cultural life *ʿalā hāmiš al-ṯaqāfa* (lit. 'On the margins of culture'). The total length of the episodes is about one hour and thirty-five minutes. *Hawāmiš* podcasts are published alongside their corresponding written articles or reports. The data collected in this section of the corpus is considered premeditated because the narrative voice of the podcast episodes reads/interprets the corresponding written articles. The written texts of the articles from which the narrative voice records the podcasts were also analyzed and compared along with the premeditated oral data.

Finally, the third source was the corpus of written texts (premeditated communication). It includes 76,026 words, or 346 articles selected by the online journal *Goud* from all columns (three articles per month from January to December 2016).[9]

All data were previously analyzed quantitatively: the most frequent morphosyntactic features clearly belonging to the repertoire of non-Standard Arabic were subsequently analyzed in terms of their function and distribution across different textual and communicative contexts. The non-Standard Arabic features include: verbal morphology (the occurrences of suffixal and prefixal conjugation of *Darija*' repertoire), nominal negation (occurrences of *māšī* 'it is not') and verbal negation (occurrences of discontinuous negation *mā- -š*), possession and annexation through the use of the preposition *dyāl* 'of', and the use of the relative sentence introduced by the relative pronoun *llī* (invariable pronoun in *Darija*) in variation with *allaḏī* and its morphological variants (from Standard Arabic). The following comparative analysis presents the MMA traits in their textual and discursive context, in both oral and written modes of communication.

4. Analysis

4.1. Oral corpus: unpremeditated speech

The first part of the analysis concerns the unpremeditated oral texts from the talk show *Ḥadīṯ*^{un} *maʿa aṣ-ṣaḥāfa*. Journalist Abdellah Tourabi opens the broadcast as shown in (1) and (2), below:

[8] The episode is available on the official YouTube page of the 2M television network, at the following link https://youtu.be/lN1yCo2XKyg?si=DKyqzFDiXy-OEFUJ (last accessed on 10/10/2023).

[9] See Pennisi (2025: 151-154).

(1) mušāhidī-nā l-kirām ʾahl[an] wa-marḥab[an] bi-kum ka-kull yawm ʾaḥad maʿa barnāmiǧ-kum l-ḥiwārī Ḥadīt[un] maʿa ṣ-ṣaḥāfa. ḍayfu-nā li-**hādā** l-yawm huwa s-sayyid Muṣṭafā r-Ramīd, wazīr ad-dawla al-mukallaf bi-ḥuqūq l-ʾinsān, li-l-ḥadīt ʿan waḍʿiyyat ḥuqūq l-ʾinsān wa-l-ḥurriyāt bi-l-Maġrib wa-kayfa tanwī ḫuṭṭat al-ʿamal l-waṭanī fī mağāl ad-dīmūqrāṭiyya wa-ḥuqūq l-ʾinsān tarsīḫa-hā wa-taqwiyata-hā wa-madā ənsiǧām al-ʾaġlabiyya l-ḥukūmiyya wa-hal **tuʾattir** al-iḫtilāfāt bayna mukawwināti-hā ʿalā ʾadāʾ al-ḥukūma? ʾiḍāfat[an] bi-ṭabīʿat[i] l-ḥāl ʾilā mawāḍīʿ ʾuḫrā. **s-sī** r-Ramīd **marḥbā bī-k**…

(min: 00:26-00:54)[10]

Dear viewers, welcome, like every Sunday, to your talk-show *Ḥadīṯ[un] maʿa aṣ-ṣaḥāfa*. Our guest today is Mr. Mustafa Ramid, Minister of State for Human Rights, to talk about the situation of human rights and freedoms in Morocco and how human rights and fundamental freedoms are respected in Morocco. How does the national action plan for democracy and human rights intend to strengthen and reinforce them, and to what extent is the governmental majority harmonious, and do the differences between its components affect the government? Mr. Ramid, welcome…

(2) bidāyyat[an] **qbəl mā ndəḫlū** fə-t-tafāṣīl, ʾa **s-sī** r-Ramīd **nta kəntī** muḥāmī li-sanawāt ṭawīla, **kəntī** ʾayḍ[an] wazīr l-ʿadl wa-l-ʾān **nta** wazīr d-dawla f-mağāl ḥuqūq l-ʾinsān wakabtī **t-taṭawwur dyāl l-ḥurriyyāt** w-ḥuqūq l-ʾinsān f-l-Maġrīb, **qbəl mā ndəḫlū fə-t-tafāṣīl dyāl hād l-qaḍāyā dyāl ḥuqūq l-ʾinsān** w-l-ḥurriyyāt f-l-Maġrīb suʾāl s-**sī** r-Ramīd lī-k šaḫṣiyy[an] min ḫilāl hād **l-masār dyālə-k** f-mağāl l-ḥurriyyat w-ḥuqūq l-ʾinsān, **wāš nta** rāḍī šaḫṣiyy[an] ʿalā waḍʿ ḥuqūq l-ʾinsān ḥāliyy[an] fī-l-Maġrīb? **kifāš ka-ybān** lī-k **l-waḍʿ?**

(min: 01:13-01 :37)[11]

Before going into detail, Mr. Ramid, you were a lawyer for many years, you were also Minister of Justice, and you are now Minister of State for Human Rights, responsible for the development of freedoms and human rights in Morocco. Before going into the details of this issue of human rights and freedoms in Morocco, a personal question, Mr. Ramid: based on your career in the field of freedoms and human rights, are you personally satisfied with the human rights situation in Morocco today? How do you see this situation?

These first two excerpts represent a type of unpremeditated speech, insofar as they would not be explicitly read. Although Tourabi uses more or less the same introductory formulas to begin his program and introduce guests with premeditated questions, his speech act shows typical characteristics of unpremeditated communication, such as the use of repetition[12] (Eagleson 1958: 149-150). See especially the repetition of *qbəl mā ndəḫlū fə-t-tafāṣīl* 'Before going into the details' in (2). The sentence is used by Tourabi to introduce the first question for the guest. The question, however, is preceded by a digression contextualizing the guest; therefore, the repetition of the above sentence echoes the introduction of the question and serves, at the communicative level, to make the (oral)

[10] https://www.youtube.com/watch?v=lN1yCo2XKyg&list=PL6tDa8neN6tHKH0Ny7Ta7l2SCO-ApNz6N&index=53 (Accessed 23-10-10). This study includes both longer excerpts and (glossed) shorter examples drawn from them. Long excerpts are presented in regular font. Within these, terms and expressions in bold indicate elements that are explicitly in Moroccan Arabic, while italics are used to mark interjections and bivalent words. By contrast, (glossed) shorter examples are consistently presented in italics.

[11] *Ibidem*.

[12] Repetition, on the other hand, is regularly used as a means of persuasion, see in particular Johnstone (1991).

message clear and straightforward for the audience (the direct addressee, Mr. Ramid, and the larger audience, in the studio and at home), alternating between different foreground and background content (Khalil 2000: 121-126).

Linguistically, Tourabi uses a mixed style of Arabic. In excerpts (1) and (2), words in bold represent occurrences of words and expressions pronounced by the journalist in non-Standard Arabic. Phonetically, several words follow the typical Moroccan Arabic pronunciation, such as, for example, the merger of interdental fricatives with dental occlusives (Heath 1997: 206). Note in (1) the variation t̲ ~ t (*tuʾattir*, and not *tuʾaṯṯir* 'to affect' in Standard Arabic), or the variation of d̲ ~ d (*hādā* and not *hāḏā* 'this' in Standard Arabic). Note, too, the tendency to elide short vowels in open syllables (Versteegh 1997: 164-169), one of the typical features of Maghrebi Arabic (Aguadé 2018: 33). Indeed, observe in (1) *marḥbā* (and not *marḥab^{an}*, 'welcome', in Standard Arabic), and in (2) *qbəl* (and not *qabla*, 'before', in Standard Arabic).

On the morphosyntactic level Tourabi's speech act is characterized by the alternation of Standard and non-Standard elements. See, for example, the systematic use of the prefix conjugation, with or without the preverb *ka-*, as in (2) the verb 'to enter' *ndəḫlū*, 1st p.m. (*nadḫulu* '[we] enter' in Standard Arabic), and the verb *ka-ybān*, 'to appear', 3rd s.m., in the sentence entirely constructed in Moroccan Arabic, given in (2-a), below:

(2)
a) *kifāš* *ka-ybān* *lī-k* *l-waḍʿ?*[13]
how NPST-appear.PRS.3SG.M to-2SG.M DEF-situation
'How do you see this situation?' (Lit. 'How does the situation appear to you?')

Another morphosyntactic construction in Moroccan Arabic repertoire is the analytic annexation via the preposition *dyāl*, as in (2b-c) given below:

(2)
b) *t-taṭawwur* *dyāl* *l-ḥurriyyāt*
DEF-development of DEF-freedoms
'The development of freedoms'

c) *f-t-tafāṣīl* *dyāl* *hād* *l-qaḍāyā* *dyāl* *ḥuqūq* *l-ʾinsān*
into-DEF-details of DEM DEF-issues of rights DEF-human.being
'Into the details of this issue of human rights'

In (2-b) the analytic annexation is single, i.e. only the first term placed before the preposition *dyāl* 'of' (*t-taṭawwur* 'the development', which is a noun defined by the definite article), and the second term postposed after the preposition *dyāl l-ḥurriyyāt* ('of the freedoms'), are part of the annexation. In contrast, in (2-c), the analytic annexation

[13] All examples listed in alphabetical order refer to the previously numbered excerpts and the sources listed therein previously. Glossing rules follow the Leipzig Glossing Rules for interlinear morpheme-by-morpheme glosses.

consists of more elements, and in particular the last part of the chain, *l-qaḍāyā dyāl ḥuqūq l-ʾinsān*, 'the issue of human rights', is constructed with a noun determined by the definite article *l-qaḍāyā* ('the issues') before the preposition *dyāl*, and by an *iḍāfa* (synthetic annexation, used in both Standard Arabic and *Darija*), *ḥuqūq l-ʾinsān* ('human rights', lit. 'rights of the human'), after the preposition *dyāl*.

It should also be noted that the preposition *dyāl* in *Darija*, also serves the function of a possessive adjective when suffixed by personal pronouns as in (2-d) below:

(2)
d) *l-masār* *dyālə-k*
DEF route of-2SG.M
'Your route.' (Lit. 'the route of you').

Other elements of the lexical and morphosyntactic repertoire of *Darija* also appear in extracts (1) and (2), such as the suffix conjugation to express the past tense (see the conjugation of *kəntī*, '[you] were'), the use of interrogative pronouns, such as *kifāš*, 'how', already mentioned, and *wāš*, 'do you/are you', in (2-e), below:

(2)
e) *wāš* *nta* *rāḍī* *šaḥṣiyy^{an}?*
Q 2SG.M satisfied.SG.M personally
'Are you personally satisfied?'

Finally, on the lexical level, the Moroccan appellation *s-sī*, lit. 'Sir, or Mr.', a shortened form of *al-sayyid*, 'Sir/Mister' (in Standard Arabic), is used in this case by the journalist as a formula of respect to address Mr. Ramid.

As shown in (1) and (2), therefore, morphosyntactic features of Moroccan Arabic alternate and merge with constructions and expressions in Standard Arabic. All the morphosyntactic and lexical features shown above are part of what Youssi (1992) called MMA, and which in this study corresponds to (Moroccan) Mixed Arabic, i.e., a formal oral register of non-Standard Arabic in Morocco. This register also represents what Tourabi termed *al-luġa al-ṯāliṯa*, i.e. a mixed style in which the contours of what is traditionally called Standard Arabic are more fluid, and where morphosyntactic features of the repertoires of non-Standard varieties, in this case Moroccan Arabic, are merged with the Standard structures of the Arabic language. This is not simply the use of dialectal elements in Standard Arabic speech (such as *dyālək,* 'your' in the excerpt *d*, or *wāš,* 'are you?', in the excerpt *e*), but Standard Arabic also becomes part of this third language through the use of now bivalent terms and expressions (Mejdell 2006), such as *l-masār,* 'route', (excerpt *d*) and *šaḥṣiyy^{an}*,'personally', (excerpt *e*). Although the practice observed in (1) and (2) reflects the uncodified norm prevalent even in the oral practices of contemporary journalistic communication, it is interesting to note that there is stylistic variation within this mixed style, where the alternation between the use of expressions closer to the style of Standard Arabic and expressions closer to Moroccan Arabic reflect a pragmatic strategy. In fact, returning to examples (1) and (2), Tourabi's

introductory speech is initially constructed, both phonetically and morpho-syntactically, on an almost total adherence to the style of Standard Arabic. Instead, elements in Moroccan Arabic mark passages to the focal points of the communicative functions of his speech. In particular, note in (1) that at the end of his introduction to the topic of the episode and the announcement of the guest, both in Standard Arabic, Abdellah Tourabi welcomes his guest by switching to Moroccan Arabic, stating: *s-sī r-Ramīd marḥbā bī-k*, 'Mr. Ramid, welcome.' This abrupt switch to the Moroccan Arabic repertoire marks on the pragmatic level the performative act of "welcoming", using the colloquial register, *Darija*, which symbolically represents closeness and unity with Moroccans and at the same time a direct, clear, and concise means of communication.[14] This first shift in style allows the journalist's speech as a whole to shift and focus attention on the guest, Mr. Ramid. A second stylistic shift also occurs in (2), where Tourabi directly addresses his guest by using Mixed Arabic. Note, in particular, the final part of his speech, i.e. how the journalist structures the question for his guest. Initially, Tourabi asks:

(2)

f)	*wāš*	*nta*	*rāḍī*	*šaḥṣiyy^{an}*	*ʿalā*	*waḍʿ*	*ḥuqūq*	*l-ʾinsān*
	Q	2SG.M	satisfied.SG.M	personally	on	situation	rights	DEF-human.being

ḥāliyy^{an} *fī-* *l-Maġrīb?*
currently in DEF-Morocco
'Are you personally satisfied with the human rights situation in Morocco today?'

In this first question, the journalist alternates morphosyntactic elements of Moroccan Arabic with lexical choices in Standard Arabic, see in fact in (2-f) the interrogatives and the personal pronoun in Moroccan Arabic and the rest of the sentence in Standard Arabic.[15] Later, the journalist switches to Moroccan Arabic, more abruptly, as indicated in example (2-a) above. Thus, on the pragmatic level, the reiteration of the question, first in Mixed and more detailed Arabic, then in Moroccan Arabic, reflects a conscious discursive strategy of the journalist aimed at making his communicative intention explicit, i.e. inviting his guest to answer explicitly, clearly and directly. Moreover, Tourabi stated:

> Dans mon émission, j'essaie même de tirer l'invité vers *darija*, parce que vers *darija* il peut être plus clair, plus direct et loin de la langue de bois. L'arabe classique tel qu'il est pratiqué au Ma-

[14] Caubet (2017a, 2017b, 2018), likewise, shows that the *Darija* used by Moroccan artists in their written productions is considered a more effective linguistic tool allowing them to represent their communicative intentions and artistic creations.

[15] The lexical choices in Standard Arabic are bivalent in (2-f), insofar the lexical repertoire of *Darija* has now acquired these kinds of expressions, which are common to the journalistic lexicon. Moreover, with the exception of the first two elements of the sentence, the pronunciation of Tourabi in (2-f) reflects the phonetics of standard Arabic, but as it is usual in Media Arabic, the case ending of each individual element of the sentence is dropped. The exceptions in this case are *šaḥṣiyy^{an}*, 'personally' and *ḥāliyy^{an}*, 'currently', where the case ending of adverbs in *-an* (indefinite accusative) is in common use.

roc, c'est beaucoup plus propice à la langue de bois, c'est-à-dire qu'on dit tout et on dit rien.[16]
(Interview with Abdellah Tourabi, 2M journalist, February 22, 2018, Casablanca)[17]

The minister's reply is also characterized by stylistic variation, as shown in the following long extracts:

(3) bi-smī llāh ar-raḥman ar-raḥīm, al-ḥaqīqa ʾanna l-masār l-ḥuqūqī li-l-bilād ʾiḏā ʾanta ʾaḥaḍta-hu **f-waḥḥəd** l-madā zamanī wāsiʿ ***matal**[an]* ***nqūlū*** mən ʾalf-**wu-tsāʿ-miya-wu-tsʿīn** ***matal**[an]* l-ʾ**al-fayn-wu-tsaʿṭāš**, sa-tarā bi-ʾanna hunāk **waḥḥəd l-masīra**[18] taṣāʿudiyya wa-ʾanna-hu hunāk muktasabāt mutawāliyya, ʾiḏā **dḏəltī f-t-tafāṣīl wu-bdītī ka-ddaqqaq f-**l-waqāʾiʿ wa-l-muʿṭayāt sa-taġid *yaʿnī* l-ʿadīd mən l-mašākil wa-l-ḫilālāt wa-li-ḏālik huwwa bi-šakl[in] mulḫaṣ ʾaqūl la-k ʾanā *yaʿnī* rāḍī nisbiyy[an] ʿalā l-masār l-ḥuqūqī fī l-bilād wa-lakinna-nī ʾayḍ[an] *yaʿnī* ġayr rāḍī ʿalā kaṭīr mən at-tafāṣīl….
[Tourabi: **matal**[an] ?]
matal[an] **dākšī kulšī, māšī kulšī**, wa-lakin l-ʿadīd mimmā yuṭraḥ mimmā huwwa maʿrūf, ʿandī **matal**[an] **l-mawḍūʿ dyāl**… taʾsīs l-ġamʿiyyāt **llī** fī **katīr** min l-ʾaḥyān **kāyn** ġamʿiyyāt **mā ka-ttwəṣṣəl-š bi-t-tawāṣīl dyāl-hā**, *ʿandī matal*[an] **šī marrāt ka-ykūn baʿḍ** t-tadaḫḫulāt *yaʿnī* l-ʾamniyya allatī tutīr niqāš[an], *ʿandī matalan* baʿḍ l-muwāṭinīn alladīn **yaštakūn** mən madā tawaffur ḍamānāt l-muḥākama l-ʿādila...
(min : 02:34-03:13)[19]
In the name of God, the merciful and the omnipotent... the truth is that the route of rights in the country, if you take it in a broad time dimension, let's say from the 1990s to 2019, you will see that there are progressive steps and that there is increasing progress. If you get into the details and start looking at the facts and figures, you'll find plenty of problems and irregularities. To sum up, I can say that I'm relatively satisfied with the country's human rights progress, but I'm also dissatisfied with several details...
[Tourabi: For instance?]
Like all this, not everything, but a large part of what is known and considered... I have, for example, the question of associations which, in many neighborhoods, are not connected. I have, for example, some security interventions that provoke debate. I sometimes have citizens complaining about the lack of guarantees of fair trial…

(4) […] ʾid[an] **ḥnāyā**… **ka-nqūlū** dāʾim[an], taḥt hād l-ʿunwān, ḥuqūq l-ʾinsān f-l-Maġrīb tataqaddam[u] taqaddum[an] mustamirr[an] wa-muḍṭarid[an] [sic], lakinna-h[u] taqaddum[un] baṭiʾ[un] wa-muḍṭarib. ʾid[an], ʿand-nā *yaʿnī* l-ğānib, kamā **ka-nqūl** dāʾim[an], l-mamlūʾ min l-kaʾs, **w-**ʿ**and-nā** l-ğānib *yaʿnī* l-fāriġ min l-kaʾs, fa-alladī yurakkiz ʿalā l-ğānib l-mamlūʾ, **ka-ybān lī-h mzyān dakšī**, alladī yurīd[u] ʾan yanḍur ʾilā l-ʾumūr f-ğānib-hā l-fāriġ, **rāh ġādī yšūf** ***yaʿnī*** **fi-hā kāyn** ***l-*****mašākil**.
(min : 04:04-04:37)[20]

[16] 'In my broadcast, I even try to draw the guest into *Darija*, because in *Darija* he can be clearer, more direct and far from the "langue de bois. Classical Arabic, as it is spoken in Morocco, is much more conducive to "langue de bois", namely you say everything and you say nothing.'

[17] See the entire interview in Pennisi (2025: 520-538).

[18] The usual phonetic realization of the indefinite article in Moroccan Arabic is *waḥd əl-*. All transcriptions from oral texts reproduce dialogues as they were performed.

[19] https://www.youtube.com/watch?v=lN1yCo2XKyg&list=PL6tDa8neN6tHKH0Ny7Ta7l2SCO-ApNz6N&index=53 (Accessed 23-10-10).

[20] *Ibidem*.

> […] So we always say, under this formula, that human rights in Morocco are progressing steadily and continuously, but this progress is slow and disturbed.
> So we have, as I always say, the full half of the glass... and the empty half of the glass. To those who focus on the (glass-half) full side, everything looks good; those who want to see the facts on the (glass-half) empty side will see everything as a problem.

Unlike the journalist, the minister's speech is characterized by a more extensive use of colloquial style, or rather, a more pronounced stylistic variation. Note, for example, not only the verbal morphology of the *Darija* repertoire (as in (3) *ka-nqūlū*, '[we] say', prefix conjugation, or *dḫəltī*, lit. '[you] entered', suffix conjugation), or possession through the preposition *dyāl* (in (3) *t-tawāṣīl dyāl-hā* lit. 'their connection'), but also nominal and verbal negation, through the use, respectively, of *māšī kulšī*, lit. 'it is not everything', and *mā ka-ttwəṣṣəl-š*, lit. '[it]doesn't get to'.

Moreover, the minister's speech is built on a stylistic alternation which is most evident insofar as, in addition to using the bivalent terms of the semantic field related to human rights and politics (such as *muktasabāt*, 'progress', *tafāṣīl*, 'details', *l-masār l-ḥuqūqī*, '[human] rights progress', *ḥuqūq l-ʾinsān*, 'human rights', among others), he alternates lexical and morphosyntactic choices that are part of Standard Arabic and *Darija*, respectively. Note, for example, at the beginning of his speech in (3), *sa-tarā* '[you] will see', a verb conjugated in the future tense in line with the verbal morphology of Standard Arabic, i.e., the morpheme of the future *sa-* prefixed to the prefixal conjugation *tarā*, 2nd s.m., and the verbal root which is part of the lexical repertoire of Standard Arabic. In contrast, in (4) he uses *ġādī yšūf*, '[he] will see', which is a verbal root from the repertoire of *Darija* (*šāf* / *yšūf* 'to see'). Again, stylistic variation represents a precise communicative strategy, corroborated by the expressions used within his speech. Indeed, looking at his performative act, it is evident that in both (3) and (4) the minister's speech tends to be extremely polarized: on the one hand, the lexical choices and morphosyntactic constructions are rigidly tied to the norms of Standard Arabic. Note, for example, the use of the *basmala* (*bi-smī llāh ar-raḥman ar-raḥīm*, 'In the name of God, the merciful and the omnipotent') to emphasize his performative act, i.e. to formalize his speech in compliance with Muslim precepts[21]; moreover, unlike the journalist, he emphasizes certain passages of his speech by systematically marking the case and mood endings, such as:

(4)

a) *ḥuqūq l-ʾinsān ḥāliyy^{an} f-l-Maġrīb tataqaddam^{u} taqaddum-^{an} mustamirr-^{an}*
rights DEF-human.being in-DEF-Morocco progress.PL.NPST progress-ACC continuous-ACC
wa-muḍṭarid-^{an}, lakinna-h^{u} taqaddum-^{un} *baṭīʾ-^{un}* *wa-muḍṭarib*
and-steady-ACC but-SG.M progress-NOM slow-NOM and-disturbed
'Human rights in Morocco are progressing steadily and continuously, but this progress is slow and disturbed.'

In (4-a) the key terms of his speech are uttered with the case endings (in superscript characters). On the other hand, this sentence, entirely in Standard Arabic - and with the

[21] See for instance, the use of verses and extracts from the Coran in Geddafi's political discourses (Mazraani 2008: 665).

case ending -, precedes the last part of his speech in which the minister alternates idiomatic expressions, lexical and morphosyntactic choices that are typically part of the *Darija* repertoire. Note, in particular, his metaphorical discourse: the metaphor of the full or empty glass is expressed in Standard Arabic (as illustrated in 4-b and 4-d below), while the minister's interpretation of the metaphor is expressed in *Darija* (as illustrated in 4-c and 4-e below):−

(4)

b) *fa-* *alladī* *yurakkiz* *ʿalā* *l-ǧānib l-mamlūʾ*
and who focus.PRS.3SG.M on DEF-side DEF.full
'those who focus on the (glass-half) full side' (in Standard Arabic entirely)

c) *ka-ybān* *lī-h* *mzyān* *dakšī*
appear.PRS.3SG.M to-3SG.M good all.that
lit.'everything appears to him good'.

d) *alladī* *yurīdᵘ* *ʾan* *yanḍur* *ʾilā* *l-ʾumūr* *f-ǧānib-hā* *l-fāriġ*
who want.PRS.3SG.M COMP see.PRS.3SG.M to DEF-facts DEF-side-3PL DEF-empty
'those who want to see the facts on the (glass-half) empty side'

e) *ġādī* *yšūf* *yaʿnī* *fi-hā* *kāyn* *l-mašākil*
FUT see.PRS.3SG.M that.is in.3PL there.is DEF-problems
'he will see everything as a problem'.

On a pragmatic level, this mechanism serves a double purpose: the first concerns communicating the purely semantic meaning of the message, i.e., the slow advancement of human rights in Morocco can be interpreted positively and negatively; the second purpose, on the other hand, concerns the communicative intention, (the illocutionary force of his speech), i.e. what his speech intends to communicate. Alternating in his discourse Standard Arabic and *Darija* allows the minister to symbolically reinforce the feeling of closeness and empathy with his audience. Therefore, his communicative intention is linguistically constructed through stylistic alternation, i.e. in (Moroccan) Mixed Arabic.

4.2. Oral corpus: premeditated speech

For practical reasons, the present analysis focuses on one episode of the *Hawāmiš* podcast, entitled *Maġallat ʾAnfās tatanaffasu muğaddadan fī ʾAmrīkā... bāḥiṯūna maġāriba yuḥayyūna taġribat Al-Laʿabī wa-rifāqi-hi* ('Anfas magazine breathes again in America... Moroccan researchers breathe new life to the experience of Laabi and his companions').[22] As indicated earlier, the text of the podcast is published on hawamich.info in two versions, namely version one through a written journalistic article (whose author is Imad Stitou), and

[22] See https://hawamich.info/6468/ – :~:text = بعد أزيد من 50 سنة,عرفه مغرب سبعينات القرن الماضي . See also Pennisi (2025: 350-361).

version two through the reading of the same text (which is interpreted by Issam Belgana). The incipit of the article/podcast is shown in Table 1,[23] presenting both versions:

Table 1. Incipit Podcast/Article Hawamich.info

Translation	Podcast	Article
After more than 50 years, Anfas is back with a new breath, a group of researchers who decided to revive the experience of a magazine that characterized Morocco's cultural and political history before it was halted due to the political climate of the 1970s.	**mən ba**ʿ**d ktar mən ḥəmsīn** sana **ka-tarğa**ʿ *mğallat* ʾAnfās *b-nafas ğdīd*, mağmūʿa **mən** l-bāḥiṯīn qarrarū **y**ʿ**āwdū** ʾiḥyāʾ *təğribat* l-*mağella* **llī** ṭabʿāt t-tārīḫ at-taqāfī w-s-siyāsī f-l-Maġrib **qbal mā twaqqaf** *b-sbab* l-munāḫ s-siyāsī **llī** ʿ**arfū** Maġrib **səb**ʿ**īniyāt** l-qarn l-māḍī.	**بعد أزيد من** 50 سنة، **تعود** مجلة "أنفاس" بنفس جديد، مجموعة من الباحثين قرروا **إعادة** إحياء تجربة مجلة طبعت التاريخ الثقافي والسياسي في المغرب **قبل أن تتوقف**، بسبب المناخ السياسي **الذي عرفه** مغرب **سبعينات** القرن الماضي. **ba**ʿ**da azyad min** ḥamsīn sana, **ta**ʿ**ūdu** mağallat "Anfās" bi-nafas ğadīd, mağmūʿa min al-bāḥiṯīn qarrarū **i**ʿ**ādat** iḥyāʾ taġribat mağalla ṭabaʿat at-tārīḫ aṯ-ṯaqāfī wa-s-siyāsī fī al-Maġrib **qabla an tatawaqqaf**, bi-sabab al-munāḫ as-siyāsī **allaḏī** ʿ**arafa-hu** Maġrib **sab**ʿ**īnāt** al-qarn al-māḍī.
This article is published on hawamich.info, written by Imad Stitou and played in the audio version by Issam Belgana. You can listen to all episodes of the Hamawich podcast on all podcast platforms.	**hād** l-maqāl **ka-yğī-kūm** ʿalā Hawāmiš *point info*, **mən** ʾiʿdād ʿImād Stīttū *f-n-nusḫa l-masmūʿa*, **ka-yrāfəq-kum** ʿIṣām Belgana **w-yəmkən li-kum tsəm**ʿ**ū** ğamīʿ ḥalaqāt podcast Hawāmiš ʿalā kull minṣṣāt l-*podcast*.	2023 5.18 HAWAMICH عماد استيتو [date of publication, platform, and the author's name]
In 1966, in the Moroccan capital Rabat, a group of Moroccan poets who called themselves "linguistic fedayeen" decided to launch a unique experiment through which they intended to bring about a cultural revolution in a country just emerging from colonialism. This was the French-language literary magazine Anfas, which soon took on a political dimension, becoming indignant against existing conditions, the dominant system, and colonial and capitalist powers.	sanat ʾalf **w-ts**ʿ**āmi**ʾ**a w-stā w-stīn f-**l-ʿāṣima l-maġribiyya ar-Ribāṭ, **ğādī yqrarū** ṭulla **mən** š-šuʿarāʾ l-maġāriba **llī ṭəlqū** ʿ**lā** *nfas-hum* **smīyya dyāl** "ʾal-fidāʾi-yyīn ʾal-luġawiyyīn", qarrarū taʾsīs *təğriba* farīda **mən** nawʿ-hā, **bğāw mən** ḥilāl-hā ʾiḥdāt ṭawra ṭaqāfiya **f-**balad **llī** yā allāh **ḫrəğ mən** l-stiʿmār fa-**kānt** *mağellat* ʾAnfās l-ʾadabiyya ʾan-nāṭiqa b-l-faransiyya **lli** surʿāna-mā **ḥḍāt** buʿd siyāsī sāḥiṭ ʿ**al** ʾawḍāʿ l-qāʾima wa-l-*niḍām* s-sāʾid w-ḍidd **l-qiwā** l-istiʿmāriyya w-raʾsmāliyya.	سنة 1966، **في** العاصمة المغربية الرباط، **سيقرر** ثلة **من** الشعراء المغاربة **ممن أطلقوا** على نفسهم **تسمية** "الفدائيين اللغويين" تأسيس تجربة فريدة **من** نوعها، **أرادوا من** خلالها إحداث ثورة ثقافية في بلد **خارج للتو من** الاستعمار، **فكانت** مجلة "أنفاس" الأدبية الناطقة بالفرنسية، **التي** سرعان ما **أخذت** بعدا سياسيا ساخطا على الأوضاع القائمة، والنظام السائد، وضد **القوى** الاستعمارية والرأسمالية. sanat alf wa-tisʿu miʾa wa-sitta wa-sittūn, **fī** al-ʿāṣima al-Maġribiyya ar-Ribāṭ, **sa-yuqar-riru** ṭulla min aš-šuʿarāʾ al-Maġāriba **mimman** ʾ**aṭlaqū** ʿalā nafsi-him **tasmiyat** "al-fadāʾiyyīn al-luġawiyyīn" taʾsīs taġriba farīda **min** nawʿi-hā, **arādū min** ḥilāli-hā iḥdāṯ ṭawra ṭaqāfiyya fī balad **ḥāriğ li-t-tawwi min** al-istiʿmār, fa-**kānat** mağallat "ʾAnfās" al-ʾadabiyya an-nāṭiqa bi-l-faran-siyya, **allatī** surʿān mā ʾ**aḥaḍat** buʿdan siyāsiyyan sāḥiṭan ʿalā al-awḍāʿ al-qāʾima, wa-n-niẓām as-sāʾid, wa-ḍidda l-**quwā** al-istiʿmāriyya wa-r-raʾsmāliyya.

[23] *Ibidem*.

Unlike unpremeditated communication, *Hawāmiš*'s podcast does not contain spontaneous speech (there are no repetitions and interjections, for example); rather, the podcast follows the written text. However, the reading differs from the written text in some lexical choices and morphosyntactic structures. In Table 1, the elements that differentiate these choices in the podcast and the article are highlighted in bold, while the elements highlighted in italics represent the loan words and/or non-Arabic terms pronounced in foreign languages, but also the bivalent terms and expressions (Standard Arabic/ *Darija*) that in the podcast are pronounced according to Moroccan Arabic practices or inserted in non-Standard constructions. The most striking divergences between the two texts mainly concern morphosyntactic constructions, as for instance the realization of relative sentences shown in Table 1, as indicated in the examples 5(a-b)-9(a-b) below.

Whereas in the written article the relative pronoun *allaḏī* (and its morphological variants) is used consistently with the norms of Standard Arabic, in the podcast the only relative used is *llī*, i.e., the invariable relative of the *Darija* repertoire, established in MMA (Youssi 1992: 265-273). The following examples show the variation between the relative pronoun *allaḏī* (and its morphological variants) and the use of *llī* in the written and oral text:

5a)

qarrarū	*iʿādat*	*iḥyāʾ*	*taġribat*	*maǧalla*	*ṭabaʿat*
decide.PRF.3PL.M	repetition	giving.life	experience	magazine	mark.PRF.3SG.F

al-tārīḫ *al-ṯaqāfī* *wa-l-siyāsī* *fī-l-Maġrib*
DEF-history DEF-cultural and-DEF-political in-DEF-Morocco
'[They]decided to revitalize the experience of a magazine [that] marked the cultural and political history in Morocco'

5b)

qarrarū	*yʿāwdū*	*ʾiḥyāʾ*	*təǧribat*	*l-maǧella*	*llī*	*ṭabʿāt*
decide.PRF.3PL.M	repeat.NPST.3PL.M	giving.life	experience	DEF-magazine	REL	mark.PRF.3SG.F

t-tārīḫ *at-taqāfī* *w-s-siyāsī* *f-l-Maġrib*
DEF-history DEF-cultural and-DEF-political in-DEF-Morocco
'[They]decided to revitalize the experience of a magazine that marked the cultural and political history in Morocco'

6a)

al-munāḫ	*al-siyāsī*	*allaḏī*	*ʿarafa-hu*	*al-Maġrib*
DEF-atmosphere	DEF-political	REL	know.PRF.3SG.M-3SG.M	Morocco

'The political atmosphere that Morocco has known'

6b)

l-munāḫ	*s-siyāsī*	*llī*	*ʿarf-ū*	*l-Maġrib*
DEF-atmosphere	DEF-political	REL	know.PRF.3SG.M-3SG.M	Morocco

'The political atmosphere that Morocco has known'

7a)
ṯulla min al-šuʿarāʾ al-maġāriba mim-man aṭlaqū ʿalā nafs-i-him
group of DEF-poets DEF-Moroccan.PL.M among-REL attach.PRF.3PL.M on self-GEN-3PL.M
tasmiyat "al-fadāʾiyyīn al-luġawiyyīn"
name DEF-*Fedayeen*.PL.M DEF-linguistic.PL.M
'A group of Moroccan poets who called themselves *linguistic Fedayeen*'

7b)
ṯulla mən š-šuʿarāʾ l-maġāriba llī ṭəlqū ʿlā nfas-hum
group of DEF-poets DEF-Moroccan.PL.M REL attach.PRF.3PL.M on self-3PL.M
smiyya dyāl "ʾal-fidāʾiyyīnʾ al-luġawiyyīn"
name of DEF-*Fedayeen*.PL.M DEF-linguistic.PL.M
'A group of Moroccan poets who called themselves *linguistic Fedayeen*'

8a)
fī balad ḫāriğ li-t-tawwi min al-istiʿmār
in country coming.out just out.of DEF-colonization
'In a country just coming out of colonization'

8b)
f-balad llī yā-allāh ḫrəğ mən l-stiʿmār
in-country REL just come.out.PRF.3SG.M from DEF-colonization
'In a country that has just come out of colonization'

9a)
fa-kānat maǧallat "Anfās"[...] allatī surʿāna-mā aḫaḏat buʿd-an siyāsiyy-an
and-be.PRF.3SG.F magazine Anfas REL quickly take.PRF.3SG.F dimension-ACC political-ACC
'So, *Anfas* was the magazine that quickly took on a political dimension'

9b)
fa-kānt maǧellat ʾAnfās [...] llī surʿāna-mā ḫḏāt buʿd siyāsī
and-be.PRF.3SG.F magazine Anfas REL quickly take.PRF.3SG.F dimension political
'So, *Anfas* was the magazine that quickly took on a political dimension'

Examples 5(a-b)-9(a-b) demonstrate the occurrences of the relative pronoun in Table 1 and highlight the divergences between the text of the written article and the podcast. Specifically, the relative pronoun appears in both the written article and the podcast in 6a and 6b, and in 9a and 9b; on the other hand, the syntactic structure of the podcast text in 5b and in 8b diverges from the text of the written article in 5a and 8a. Moreover, in the written article in 7a, the indefinite relative *man* 'whom' occurs affixed to the preposition *min*, lit. 'from', i.e. *mimman* 'among whom'; whereas in the podcast the corresponding syntactic structure is built with the specific (and invariable) relative *llī*

in 7b. Such variation between written and oral texts (article/podcast), both premeditated, shows and corroborates that in formal oral production the relative pronoun *llī* is the conventionally accepted and widespread form (Youssi 1992: 265-273). In particular, as Youssi (1992) asserted, while the forms of the relative in MMA have been simplified through the use of the invariable relative pronoun *llī*, the relative syntax has, on the other hand, undergone a phenomenon of syntactic complexification, i.e. "dans bon nombre d'occurrences de la relativation en MMA, il s'agit plutôt d'emplois descriptifs ou appositifs dont la suppression n'affectera en rien les éléments préexistants, constitutifs de la proposition dite principale"[24] (Youssi 1992:270). He also adds : "Il semble même que, plus la situation est empreinte de formalisme, plus le locuteur recourt à la relativation d'élément d'énoncés qui seraient autrement dans une relation prédicative par rapport au syntagme nominal employé comme antécédent"[25] (Youssi 1992 : 271). So, the example 8b[26] corroborates Youssi's (1992) assertions regarding the prevalence of the pronoun *llī* in MMA.

Finally, it must be pointed out that the style represented by the podcast's premeditated text thus tends to reproduce and conventionalize the formal and journalistic (oral) register, namely (Moroccan) Mixed Arabic. Indeed, if one looks globally at the podcast's premeditated oral text, the morpho-syntactic traits described by Youssi (1992) and typical of MMA for oral communication (including the relative sentences just observed) are blended with lexical choices more in keeping with the lexicon of Standard Arabic. Examples 5(a-b)-9(a-b) clearly illustrate this phenomenon. The lexical choices in the podcast, in fact, slavishly follow the lexicon used in the written article; extremely evident, for example, is the use of expressions and collocations attested in Standard Arabic and not usual in *Darija*, such as *ṯulla mən* 'a group of' in 7b, or *surʿāna-mā* 'quickly' in 9b that would certainly have correspondents in more current uses in Moroccan Arabic, but not necessarily less formal ones.

The uniqueness of the oral and premeditated text of the podcast lies precisely in its mixed and bivalent nature: mixed because Standard and non-Standard syntactic and morphological elements are mixed in the same text; bivalent because the lexicon (which also includes the verbal roots of almost all the verbs used in the podcast that are conjugated according to the Moroccan Arabic system) is derived from the shared repertoire (Standard and non-Standard) of Arabic.

[24] 'In many occurrences of relative sentences in AMM, these are rather descriptive or appositive uses whose deletion will in no way affect the pre-existing elements making up the so-called main proposition'.

[25] 'It even seems that, the more formal the situation, the more the speaker relies on the relativization of elements of statements that would otherwise be in a predicative relation to the nominal phrase used as an antecedent'.

[26] Note that, in contrast, in example (5-b) the relative is used because unlike the text of the written article, in the podcast the antecedent, *l-mağella* 'the magazine' is determined by the definite article, and therefore the use of the relative *llī*, would be grammatically required. Note, however, that in (5-h), on the other hand, the antecedent *f-balad* 'in a country' is indefinite, and therefore the use of *llī* would not have been grammatically required, but more importantly, as already mentioned, it represents a syntactic complexification, otherwise replaceable by the active participle *ḫāriğ*, 'coming out', as used in parallel in the written article in (5-g).

Moving beyond the dichotomous view of diglossia and its functionalist approach, the textual types analyzed and compared in the present study show that it is not completely possible to divide two distinct varieties of Arabic, but that the linguistic and stylistic variations found in the (premeditated and unpremeditated) practices highlight the tendency to systematically use Mixed Arabic. The difficulty of classifying Arabic into different varieties emerges more clearly when observing the practices of non-Standard written and premeditated production, will be looked at in detail at in the next section.

4.3. Analysis of written production: premeditated written discourse

This section focuses on the use of (Moroccan) Mixed Arabic in premeditated journalistic writing on the online newspaper *Goud*. Many *Goud* journalists claim to express themselves in Moroccan Arabic[27]. However, when carefully analyzing their texts, what emerges is actually (Moroccan) Mixed Arabic. Despite the fact that they claim to write in *Darija* (a non-codified language for formal written purposes), their non-Standard texts (written in Arabic characters) bring out more of the bivalent and mixed nature of *Goud*'s articles, especially in terms of lexical choices. As already demonstrated on the oral data in the previous section, the *Goud*'s texts written in (Moroccan) Mixed Arabic are also characterized by an overt syntactic structuring borrowed from the repertoire of Moroccan Arabic, and a lexical choice more closely adherent to Standard Arabic. In Opinions articles, however, stylistic variation is further emphasized by linguistic choices in order to mark the illocutionary force of such argumentative texts.

See, in particular, the Opinions article by Mohamed Socrates[28] in which the former militant of the Feb. 20 movement criticizes the repression of Rif protesters by police forces in 2016. Table 2 below shows part of the incipit (on the right)[29] and its translation on the left:

Table 2. Opinions article from *Goud*

Translation	Arabic source
Of course, we're known, but even we, we know you, one by one. We know you from generation to generation, dating back many centuries. […] […] We know that you know us, and that all our movements are carried out according to this mutual knowledge that we have built up. So, as you've seen, and as the world has seen, our movements were peaceful, and you can't use violence with your acquainted people. we respect the knowledge that [there is] between us, but, unfortunately, it is you who have not respected it.	طبعا نحن معروفون ، ولكن **حتى حنا كنعرفوكم** واحد واحد ، **كنعرفوكم** أبا عن جد وطيلة قرون[...] [...]**إحنا عارفين** أنكم **كتعرفونا** وأي تحرك **ديالنا راه كيتم** وفق **هاد** المعرفة المسبقة **لي بيناتنا** ، لذا فكما رأيت ورأى العالم تحركاتنا كانت سلمية ، **مايمكنش تدير** العنف مع المعارف ، ونحن نحترم هذه المعرفة التي بيننا ، ولكن للأسف **نتومة لي مكتحترموهاش** [...] ṭabʿan naḥnū maʿrūfūn, wa-lakin **ḥəṭṭā ḥnā ka-nʿərfū-kum** wāḥd wāḥd, **ka-nʿərfū-kum** abā ʿan ǧadd wa-ṭīla qurūn [...] [...] **ḥnā** ʿ**ārfīn** ʾanna-kum **ka-taʿrfūnā** wa-ʾayyi taḥarruk **diyāl-nā** rāh **ka-yatimm** wifqa **hād** l-maʿrifa l-musbiqa **lī bayanāt-nā**, liḏā fa-kamā raʾayt wa-raʾā al-ʿālam taḥarrukāt-nā kānat silmiyya, **mā yimkinš tədīr** al-ʿunf maʿ al-maʿārif, wa-naḥnu naḥtarim haḏihi al-maʿrifa allatī bayna-nā, wa-lakin li-l-asaf, **ntūma llī ma-ka-təḥtārmū-hā-š** […]

[27] See for instance the interview to Ahmed Najim (Editorial Director of *Goud*) and to Mohamed Socrate (*Goud*'s collaborator), in Pennisi (2025: 294-335).

[28] https://www.goud.ma/250402-طحنتو-سحقتونا-حيث-مكتعرفوناش-ي-،/. See Pennisi 2020 concerning the linguistic and stylistic expression of resistance in Goud's media discourse.

[29] https://www.goud.ma/250402-طحنتو-سحقتونا-حيث-مكتعرفوناش-أخاي-،/.

Socrates builds his speech by mixing expressions closer to Standard Arabic (معروفون طبعا نحن *ṭabʿan naḥnū maʿrūfūn* ‘of course, we’re known’), with expressions closer to Moroccan Arabic (ولكن حتى حنا كنعرفوكم *wa-lakin ḥəṭṭā ḥnā ka-nʿərfū-kum* ‘but even we, we know you’), alternating between Standard and non-Standard linguistic features and morphosyntactic structures. For instance, he alternates between prefix conjugation according to the norms of Standard Arabic with conjugation norms typical of Moroccan Arabic, as showed in example 10 below:

(10)

wa-naḥnu	*naḥtarim*	*haḏihi*	*al-maʿrifa*	*allatī*	*bayna-nā,*	*wa-lakin*	*li-l-asaf,*
and-1.PL	respect.PRS.1PL	this	DEF-knowledge	REL	between-1.PL	and-but	for-DEF-sorrow
ntūma	*llī*	*ma*	*ka-təḥtārmū-hā-š*				
2.PL	REL	NEG	NPST-respect-PRS.2PL-3SG.F-NEG				

In example 10, Socrate uses *naḥtarim* (‘we respect’), affirmative prefixal conjugation, in Standard Arabic, *versus* negative prefixal conjugation, in *Darija*, such as *ma-ka-təḥtārmū-hā-š* (‘you don’t respect’), or again, the variable use of the relative pronoun, such as *allatī* (‘that’), in Standard, *versus llī* (‘who’), in *Darija*, in the same sentence.

His entire article, as well as the underlying discourse it conveys, is structured on an antithetical construction, which reflects morphosyntactic and stylistic variation, emphasizing, as shown in (10), the illocutionary force of his message. Indeed, in (10), Socrates uses the personal pronoun *ntūma* (‘you’)[30] in *Darija* topicalizing it, where in fact, *ntūma* is the antecedent of the relative. This syntactically marked construction through topicalization, as occurs in other language systems as well, is also stylistically marked: the first part of the sentence (*naḥnu naḥtarim*) is in Standard Arabic, while the second part of the sentence, which is syntactically marked, is in Moroccan Arabic (*ntūma llī ma-ka-təḥtārmū-hā-š*). The shift to *Darija* reflects a concern to forcefully express a message that the author considers important and central to his discourse, namely, in his view, it is the Power that has no respect for the Moroccan people. Expressing it in *Darija* allows him to construct (allegorically) his closeness to and identification with the people, i.e. with all those repressed demonstrators that the former militant of the February 20 movement defends in his article. The illocutionary force of his message is based, therefore, not only on a syntactically marked construction, but also linguistically marked on the level of stylistic choices.

The stylistic mixing and alternation within his discourse makes his written text linguistically and stylistically identifiable with (Moroccan) Mixed Arabic. Indeed, even the constructions and expressions that in writing might visually appear to be in Standard are, in fact, bivalent depending on how the reader reads them in their overall syntactic and stylistic context.

Socrates’ premeditated text is functionally constructed as an argumentative text with its communicative strategies designed to criticize repression, but also as a message of

[30] Note also the alternation of personal pronouns and deictics used as a political discourse strategy. For more details, see La Rosa (2018), Maalej (2013), Pennycook (1994), Manetti (2015) and Holes (1993).

solidarity and identification with the repressed Moroccan people. The Mixed Arabic of his text thus allows him to express his message more incisively.

5. Conclusion

The analysis conducted in this study has demonstrated that (Moroccan) Mixed Arabic is emerging as an increasingly stabilized register within Morocco's formal and digital media communication, both in oral and written forms. By comparing unpremeditated oral texts (talk shows), premeditated oral texts (podcasts), and written texts (digital journalism), the study reveals the stability of morphosyntactic and lexical features characteristic of formal oral *Darija* – defined as MMA by Youssi (1992) – blended with structures drawn from Standard Arabic, even in written productions. Notably, the variety of Mixed Arabic employed in the online newspaper *Goud* mirrors, in written form, the mixed linguistic practices observed in oral communication. One illustrative example is the use of the relative pronoun *llī*: while variation between *llī* (*Darija*) and *allaḏī* (Standard Arabic) was observed in the spontaneous oral speech of Minister Ramid and participating journalists, all instances of relative clauses in the podcast (a premeditated oral text) were constructed using *llī*. This systematic preference in the podcast may suggest an emerging process of informal standardization of a journalistic-style formal register within (Moroccan) Mixed Arabic. The same stylistic features are observable in the written corpus, with the exception of the *Goud*'s opinion articles, where texts, such as the one authored by Socrate, exhibit similar stylistic variation and discursive strategies to those found in unpremeditated oral speech. Rather than being isolated or informal phenomena, these discursive practices indicate that (Moroccan) Mixed Arabic is functioning as a *de facto* vehicular language within the Moroccan journalistic sphere. This mixed style enables the expression of authority, emotional engagement, and conceptual clarity. As demonstrated, the stylization of such mixed forms serves specific pragmatic purposes: it facilitates modulation of register, marks thematic transitions, and fosters a sense of proximity and accessibility in communication with audiences.

These findings contribute to ongoing sociolinguistic and ideological discussions regarding the nature of linguistic variation in Arabic. While early models, such as Ferguson's (1959) foundational diglossia framework, proposed a binary distinction between "high" and "low" varieties, subsequent scholarship – particularly Badawī's (1973) stratified model of Arabic levels, Meiseles' (1980) continuum-based approaches, and Mejdell's (2006) work on mixed styles – has increasingly emphasized a more dynamic, gradated understanding of variation, especially within spoken registers. Expanding on this line of research, the present study proposes an extension of the diglossic continuum model through the inclusion of written communicative practices – an area that has often received limited attention in studies on diglossia. By adopting a comparative methodology that spans unpremeditated and premeditated oral data as well as written digital media, this study demonstrates how morphosyntactic features associated with formal spoken varieties – such as those described by Youssi (1992) in his definition of MMA – are increasingly being stylized and conventionalized in written journalistic discourse. Importantly, the

systematic use of (Moroccan) Mixed Arabic in established and professional media outlets – such as the digital newspaper *Goud*, the multimodal platform *Hawāmiš*, and the television broadcaster 2M – suggests a form of informal or implicit legitimacy attributed to these mixed practices. Although not formally codified, the consistent uses of such mixed styles in high-visibility, formal communicative domains indicate their growing normative status. In this regard, the study not only reinforces but also expands current understandings of Arabic's sociolinguistic landscape, by highlighting how mixed language practices are contributing to a broader reconfiguration of the boundaries between Standard and non-Standard varieties – not only in speech, but also in writing.

Nevertheless, the study was conducted using limited data. For instance, the corpus analyzed is temporally limited (with only one year of written data) and geographically/sociologically bounded (focused on a limited number of platforms and programs). Future research could expand the dataset to encompass a broader range of sources, including additional discourse genres such as social media interactions, sports commentary, and institutional communication, in order to further assess the scope and variability of the phenomenon. Furthermore, ethnographic research of the reception of such linguistic practices among Moroccan audiences would yield valuable insights into perceptions of legitimacy, acceptability, and the social indexicality of mixed forms.

Lastly, (Moroccan) Mixed Arabic emerges from this study not only as a descriptive linguistic object, but also as an index of broader sociocultural transformations reshaping the relationship between language, identity, and media in contemporary Maghreb societies. If the practices documented here reflect a wider trend, it may be argued that the boundaries between Standard and colloquial Arabic are not merely being blurred, but are actively being reconfigured through deliberate, contextually anchored communicative practices – practices that merit full recognition in the study of modern and contemporary Arabic.

References

Adam, Jean-Michel. 1997. Unités rédactionnelles et genres discursifs: Cadre général pour une approche de la presse écrite. *Pratiques linguistique, littérature, didactique* 94. 3-18.

Aguadé, Jordi. 2006. Writing dialect in Morocco. *Estudios de dialectología norteafricana y andalusí* 10. 253--274.

Aguadé, Jordi. 2012. Monarquía, dialecto e insolencia en Marruecos: El caso Nichane. In Meouak, Mohamed & Sánchez, Pablo & Vicente, Ángeles (eds.), *De los manuscritos medievales a internet: La presencia del árabe vernáculo en las fuentes escritas*, 441-464. Zaragoza: Universidad de Zaragoza.

Aguadé, Jordi. 2018. The Maghrebi dialects of Arabic. In Clive, Holes (ed.), *Arabic historical dialectology. Linguistic and sociolinguistic approach*, 29-63. Oxford: Oxford University Press.

Avallone, Lucia. 2017. Neither *fuṣḥā* nor '*āmmiyya*: How to reach a simplified Arabic language writing for theatre: Linguistic devices in Tawfīq al-Ḥakīm's theory and practice of the Third Language. *Vicino Oriente* 12. 47-54.

Badawī, al-Saʿīd Muḥammad. 1973. *Mustawayāt al-ʿarabiyya al-muʿāṣira fī Miṣr*. Cairo: Dār al-Maʿārif bi-Miṣr.

Badawi, El-said, Muhammad. 1985. Educated spoken Arabic: A problem in teaching Arabic as a foreign language. In Jankowsky, Kurt R. (ed.) *Scientific and humanistic dimensions of language*, 15-22. Amsterdam: John Benjamins Publishing.

Benítez-Fernández, Montserrat. 2008. Arabe marroquí como proyecto editorial: Es una experience posible? In Abu-Shams, Leyla (ed.), *Actas del III Congresso Internacional de Árabe Marroquí: Estudio, enseñanza y aprendizaje*, 37-54. Bilbao: Universidad del País Vasco.

Blanc, Haim. 1960. Stylistic variation in spoken Arabic: A sample of interdialectal educated conversation. In Ferguson, A. Charles (ed.), *Contributions to Arabic linguistics*, 81-156. Cambridge, Mass: Harvard University.

Boussofara-Omar, Naima. 2006a. Diglossia. In Versteegh, Kees & Eid, Mushira & Elgibali, Alaa & Woidich, Manfred & Zaborski, Andrzej (eds.), *Encyclopedia of Arabic language and linguistics* 1, 629-637. Leiden: Brill.

Boussofara-Omar, Naima. 2006b. Neither third language nor middle varieties but diglossic switching. *Zeitschrift für Arabische Linguistik* 45. 55-80.

Brigui, Fouad. 2016. De l'usage de l'arabe dialectal dans la presse écrite marocaine. In García Moscoso, Francisco & Moustaoui Srhir, Adil (eds.), *Identidad y Conciencia Lingüística, VI Congreso de Árabe Marroquí*, 249-264. Madrid: UAM Ediciones.

Brustad, Kristen. 2017. Diglossia as Ideology. In Høigilt, Jacob & Mejdell, Gunvor (eds.), *The politics of written language in the Arab world: Writing change*, 41-67. Leiden – Boston: Brill.

Caubet, Dominique. 2017a. Morocco: An Informal passage to literacy in dārija (Moroccan Arabic). In Høigilt, Jacob & Mejdell, Gunvor (eds.), *The politics of written language in the Arab world: Writing change*, 116-141. Leiden and Boston: Brill.

Caubet, Dominique. 2017b. Darija and the construction of 'Moroccanness'. In Bassiouney, Reem (ed.), *Identity and dialect performance: A study of communities and dialects*, 99-124. Abingdon, Oxon / New York: Routledge.

Caubet, Dominique. 2018. New elaborate written forms in Darija: Blogging, posting, and slamming in Morocco. In Benmamoun, Elabbas & Bassiouney, Reem (eds.), *The Routledge handbook of Arabic linguistics*, 387-406. London – New York: Routledge.

Eagleson, D. Robert. 1958. Premeditated and unpremeditated speech: The nature of the difference. *English Studies* 39(1-6). 145-154. https://doi.org/10.1080/00138385808597011

Fairglough, N. 2003. *Analysing discourse. Textual analysis for social research*. London – New York: Routledge.

Ferguson, A. Charles. 1959. Diglossia. *Word* 15. 325-340.

Haeri, Niloofar. 2000. Form and ideology: Arabic socio-linguistics and beyond. *Annual Reviews of Anthropology* 29. 61-87.

Haeri, Niloofar. 2003. *Sacred language, ordinary people*. New York: Palgrave Macmillan.

Heath, Jeffrey. 1997. Moroccan Arabic phonology. *Phonologies of Asia and Africa (including the Caucasus)* 1. 205-217.

Holes, Clive. 1993. The uses of variation: A study of the political speeches of Gamal Abd al-Nasir. In Eid, Mushira & Holes, Clive (eds.), *Perspectives on Arabic linguistics V*, 13-45. Amsterdam: John Benjamins Publishing.

Hoogland, Jan. 2013. L'arabe marocain langue écrite. In Benítez Fernandez, Montserrat & Miller, Catherine & de Ruiter, Jan Jaap & Tamer, Youssef (eds), *Evolution des pratiques et représentations langagières dans le Maroc du XXIème siècle*, 175-188. Paris: L'Harmattan.

Hoogland, Jan. 2018. Darija in the Moroccan press: The case of the magazine *Nichane*. *Sociolinguistic Studies* 12(2). 273-293.

Hudson, Alan. 1992. Diglossia: A bibliographic review. *International Journal of the Sociology of Language* 21. 611-674.

Hudson, Alan. 1994. Diglossia as a special case of register variation. In Biber, Douglas & Finegan, Edward (eds.), *Sociolinguistic perspectives on register*, 294-314. New York: Oxford University Press.

Hudson, Alan. 2002. Outline of a theory of diglossia. *International Journal of the Sociology of Language* 157. 1-48.

Johnstone, Barbara. 1991. *Repetition in Arabic discourse: Paradigms, syntagms, and the ecology of language*. Amsterdam – Philadelphia: John Benjamins Publishing.

Kaye, S. Alan. 2001. Diglossia: The state of the art. *International Journal of the Sociology of Language* 152. 117-129.

Khalil, N. Esam. 2000. *Grounding in English and Arabic news discourses*. Amsterdam – Philadelphia: John Benjamins Publishing.

Khalil, Saussan. 2018. *Fuṣḥá, ʿāmmīyyah, or both? Towards a theoretical framework for written Cairene Arabic.* Leeds: University of Leeds. (Doctoral dissertation.)

Khalil, Saussan. 2022. *Arabic Writing in the digital age: Towards a theoretical framework*. New York: Routledge.

Langone, A. D. 2016. Lingua araba in vecchi e nuovi media: Riflessioni sull'intrusione dell'arabo dialettale come lingua scritta in epoca contemporanea. *Annali Sezione Orientale* 76(1-2). 51-76.

La Rosa, Cristina. 2018. Alcune strategie retoriche nel discorso politico tunisino: Uso dei deittici e ripetizione lessicale. *La rivista di Arablit* 8(15). 67-92.

Maalej, Zouheir A. 2013. Framing and manipulation of person deixis in Hosni Mubarak's last three speeches: A cognitive-pragmatic approach. *Pragmatics* 23(4). 633-659.

Manetti, Giovanni. 2015. Il noi tra enunciazione, indessicalità e funzionalismo. In Janner Maria, Chiara & Della Costanza, Mario & Sutermeister, Paul (eds.), *Noi, Nous, Nosotros: Studi romanzi – Études romanes – Estudios románicos*, 23-44. Frankfurt: Peter Lang.

Mazraani, Nathalie. 2008. Political discourse and language. In Versteegh, Kees & Eid, Mushira & Elgibali, Alaa & Woidich, Manfred & Zaborski, Andrzej (eds.), *Encyclopedia of Arabic language and linguistics* 3, 663-671. Leiden: Brill.

Meiseles, Gustav. 1980. Educated spoken Arabic and the Arabic language continuum. *Archivum Linguisticum* 11. 118-148.

Mejdell, Gunvor. 2006. *Mixed styles in Spoken Arabic in Egypt: Somewhere between order and chaos*. Leiden – Boston: Brill.

Mejdell, Gunvor. 2022. Erasing boundaries in contemporary Written Mixed Arabic (Egypt). In Jérôme, Lentin & Grand'Henry, Jacques (eds.), *Middle and Mixed Arabic over time and across written and oral genres*, 181-194. Louvain-la-Neuve: Peeters.

Miller, Catherine. 2017. Contemporary dārija writings in Morocco: Ideology and practices. In Høigilt, Jacob & Mejdell, Gunvor (eds.), *The politics of written language in the Arab world: Writing change*, 90-115. Leiden – Boston: Brill.

Mitchell, Terence, F. 1986. What is educated spoken Arabic? *International Journal of the Sociology of Language* 61. 7–32.

Pennisi, Rosa. 2020. Expressions of resistance, "Goud" and stylistic variation in Moroccan digital newspapers. *La rivista di Arablit* 20. 79-98.

Pennisi, Rosa. 2025. *Arabe Mixte 2.0: Pratiques et représentations linguistiques dans les journaux et les médias numériques marocains.* Roma: Istituto per l'Oriente C.A. Nallino.

Pennycook, Alastair. 1994. The politics of pronouns. *English Language Teaching Journal* 48(2). 173-178.

van Dijk, Teun A. 2008. *Discourse and context: A sociocognitive approach*. Cambridge: Cambridge University Press.

Versteegh, Kees. 1997. *The Arabic language*. Edinburgh: Edinburgh University Press.

Walters, Keith. 1996. Diglossia, linguistic variation, and language change. In Mushira, Eid (ed.), *Perspectives on Arabic linguistics*, 157-197. Amsterdam and Philadelphia: John Benjamins Publishing.

Youssi, Abdelrahhim. 1992. *Grammaire et lexique de l'arabe marocain moderne*. Casablanca: Wallada.

Zack, Liesbeth. 2014. The use of the Egyptian dialect in the satirical newspaper *Abu naddāra zarʾa*. In Durand, Olivier & Daiana Langone, Angela & Mion, Giuliano (eds.), *Alf lahǧa wa lahǧa: Proceedings of the 9th Aida Conference*, 465-478. Wien – Münster: Lit Verlag.

DOI: 10.14746/linpo.2025.67.1.7

Moroccan Arabic in advertising context: Analysis of oral and written messages

Samera Abdelati

University of Naples L'Orientale
s.abdelati@unior.it | ORCID: 0009-0001-8918-3549

Abstract: The outputs of the contact due to the presence and interaction of multiple languages in Morocco have been tackled by numerous sociolinguistic studies over the years (Ennaji 2002, Chekayri 2006, Caubet 2017b). This research attempts to illustrate the way this complex linguistic landscape affected the language of oral and written Moroccan advertising, wherein the alternation of registers and the functional expansion of the *dārija* became more detectable.
To this end, in this study I aim to explore the correlation between language use and target audience, after providing a general overview of the first advertising materials disseminated in Morocco since the beginning of the 20th century. The discussion on linguistic innovations in this field will be accompanied by an analysis of commercials that have been aired on *2M*, Morocco's most popular television channel, since the 1960s-70s. This focus will allow us to emphasise the peculiarities of mostly oral advertising messages based on the target audience they seem to address. Further, the second part of the paper will be devoted to the reflection on the orthographic representation strategies used by speakers of Moroccan Arabic, by means of the analysis of billboards, which represent the so-called 'outdoor advertising'. Furthermore, such billboards, whose pictures have been captured in various parts of Morocco, provide us with an opportunity to observe contact phenomena of alternation, insertion, and code-mixing (Appel & Muysken 1987), and to evaluate their role in the effectiveness of the advertising message, as well as to provide some observations in their orthographic treatment.

Keywords: Moroccan Arabic, multilingualism, advertising, orthography, target audience

1. Introduction

The juxtaposition of different languages in the Arab countries is a concrete and persistent phenomenon which can be defined as "diglossia" when it pertains to the relationship between classical Arabic and dialectal varieties of Arabic, and as "multilingualism" when it involves additional languages, frequently of European origin or, in the context of the Maghreb, also Berber varieties. As Morocco is one of the countries where it is

inevitable to encounter a multitude of linguistic realities, this research aims at demonstrating how the coexistence and the alternation of various linguistic codes are reflected, both in written and oral forms, across all spheres of life, with a particular emphasis on advertising. Given that language serves as a primary tool for establishing a relationship between the advertiser and the advertisee, this study examines how Moroccan Arabic[1] is employed in advertising communications across television, radio, and billboards. Therefore, an initial overview of the role of Moroccan Arabic within the sociolinguistic landscape of Morocco, along with the early advertising tools that emerged in the country, is followed by a contemplation on the nature of advertising communications broadcasted through radio and television. The last section of the paper examines the strategies employed in the transliteration of *dārija*, enabling an analysis of the various writing systems that are adopted within this context. The study concludes with an examination of the factors influencing the selection of one writing system over another, as well as an analysis of the interrelationship between content, language register, and the target audience, in order to get a clearer understanding of how multilingual advertising communication is handled in Morocco.

2. Methodology

This study emerges from the recognition that Moroccan Arabic, although traditionally considered as a colloquial, and hence predominantly oral, variety of Arabic, is nowadays being increasingly adopted in various contexts, both formal and informal. The initial paragraphs of the study, indeed, provide a more descriptive analysis of the linguistic choices of the audiovisual advertising, concentrating on a total of 34 advertisements from *Med Radio*, 23 from *Hit Radio*, and 37 TV commercials from *2M*. However, the functional expansion of the *dārija* is also particularly evident in the streets of Morocco, where one can observe the frequent graphical representation of Moroccan Arabic on a variety of advertising billboards. Thus, recognizing the diverse transliteration systems employed, I began to capture billboards during my recent travels in Morocco. Consequently, this research emerges from an examination and a reflection on the various strategies adopted to write Moroccan Arabic, on the basis of 41 advertising billboards captured between Meknés, Ksar El Kebir, Rabat, Casablanca and Mohammedia in February and then in August/September 2024.

3. Moroccan linguistic panorama

Historical developments and contact between Arabs and indigenous populations, as well as the social changes that followed the process of urbanization that has affected North Africa in recent decades, contributed to the hybridization and further complication of the Moroccan sociolinguistic profile. Thus, the employment of a particular termino-

[1] In this study, the term “Moroccan Arabic” is used interchangeably with “dārija”.

logical framework to elucidate the sociolinguistic context of Morocco presents a considerable degree of complexity. Indeed, the phenomenon cannot merely be categorized as simple diglossia, in which a colloquial variety coexists with what Ferguson (1959: 336) calls 'a super-posed variety' i.e. Standard Arabic. The sociolinguistic panorama observed in Morocco is significantly more heterogeneous and intricate, characterized by the confluence of foreign languages, including French and, to a lesser degree, Spanish and English, alongside the nation's official languages, Arabic and Berber. Terms such as 'multilingualism' (Ennaji 2005) and 'transglossia' (Durand 2018: 95) have recently been adopted to describe the expansion of languages that are extraneous to the cultural heritage of the community witnessing their spread. Indeed, this is a mainstream phenomenon that leads people to mix more languages and varieties almost instinctively. Within this framework, these languages enjoy different statuses and degrees of use. Moroccan Arabic and Amazigh varieties have been mostly limited to daily communication and excluded, at least until a few decades ago, from the intellectual and academic fields. On the contrary, Standard Arabic and French have served as languages of prestige, deemed appropriate for formal settings including educational and administrative functions.

Nonetheless, due to the emergence of technological innovations and the adoption of new communication tools, a noticeable linguistic dynamism is currently evident, which is reflected in the integration of informal registers in broader contexts and, occasionally, in the combination and/or alternation of various languages, as is the case in everyday exchanges. Morocco is presently navigating through diverse social and political dynamics that markedly shape the evolution of its linguistic landscape. As a result, an ongoing process of reconfiguration regarding the hierarchical positions and the ideologies linked to the diverse languages that, for various reasons, constitute the Moroccan linguistic landscape is currently taking place.

In order to gain a comprehensive understanding of the nature of this new sociolinguistic reality that characterises Morocco, the following paragraphs will outline the key strategies adopted to promote the use of Moroccan Arabic in the private and public sectors, with the aim of integrating it into various facets of the society.

3.1. The expansion of Moroccan Arabic

Moroccan Arabic, commonly referred to as *dārija*[2], serves as the primary language for the majority of the Moroccan population and is generally learned as a second language by Berber speakers. It is not only a means of everyday communication but also a crucial element in the cultural identity and social interaction within Moroccan society. It can in fact be considered as a *lingua franca* in Morocco, even though it has no legal status. Indeed, the contemporary sociolinguistic reality in Morocco is facing an extension of the use of the vernacular to different spheres. This phenomenon arises from a diverse array of needs and objectives, and is manifested through various channels and mediums, each tailored to meet specific communicative demands. The emergence of technology has un-

[2] For an explanation of the origin of the term *dārija* and its subsequent diffusion within Moroccans, see Caubet (2017: 99-100).

deniably played a pivotal role in reshaping national communication practices. However, it is essential to recognise that even prior to the Independence (1956), Moroccan audio-visual landscape was never characterised by monolingualism (Miller 2013: 92). Radio broadcasting was introduced in Morocco in 1928 with the first public radio (*Radio Maroc*), followed by many other radio stations that emerged during the Protectorate, and all of them used to broadcast in different languages[3] (Miller 2013). Similarly, the first advertising messages were transmitted through different channels and languages, as we will see below.

That being said, the use and alternation of different codes and languages has never constituted a major novelty in the oral Moroccan audio-visual sector. What does, however, distinguish between the communicative practices of the first half of the 20th century and those of the more recent decades is the expansion of Moroccan Arabic into written context.

Until two decades ago, indeed, the role of the *dārija* was essentially that of an oral language, and the writing of it was very limited; the few examples of written production in *dārija* derived from old poetic traditions, such as *malḥūn* and *zajal*[4]. We also know that there existed so-called *majdubiyāt* – poetic quatrains that were presumably produced by the famous Sufi mystic Abderrahman El Majdoub (1506-1568). These poems were composed and written down in a colloquial Arabic that was perfectly comprehensible across the Maghreb, and even today separate proverbs keep circulating across the Greater Maghreb, that is from Morocco to Libya. More recently, Moroccan vernacular resurfaced in dramatic compositions with Tayeb Saddiki and Ahmed Taieb El Alj, the first two dramatists that wrote modern theatre in Moroccan Arabic during the second half of the 20th century[5]. In this regard, it also worth mentioning that a new kind of poet started emerging within the local, until very recent, predominantly oral poetic traditions:

> The artist, in his/her attempt to reflect the new reality of a society that now extends far beyond local tribal affairs increasingly begins experimenting – simultaneously endeavoring to attract a more worldly new audience without alienating the older traditional one – creating different kind of songs and, ultimately, poems. We thus now have a new generation of poetry within the tradition: authored, written texts. Authors now write down their poetry in a copy book, or *kunnash* (*kunnāš*), and, therefore see themselves as part of the literary Arab tradition or, perhaps more likely, part of the Magreb's centuries-old literary and music tradition. (Gintsburg 2022: 209)

However, as I already mentioned, it was mostly in the last decades that the use of written Moroccan Arabic experienced a significant shift, transcending its traditional boundaries and permeating wider cultural spheres. Nowadays, this phenomenon manifests

[3] See Jaidi (2000) for a detailed analysis of the audiovisual media diffusion in Morocco.

[4] Forms of dialectal poetry particularly widespread in Morocco and Algeria, and to a lesser extent in Tunisia a Libya since the 15th century. The term *Malḥūn* can be applied to sung and recited oral traditions (Gintsburg 2020: 206). See Pellat (1987: 247-257) for more details.

[5] On the role of *dārija* in the theatre, see Amine & Carlson (2011).

itself in an appropriation of written *dārija* in diverse contexts[6], such as social media and advertising, but also poetry and novels, as Aguadé (2003: 253) points out:

> In many novels written in Standard Literary Arabic, the authors use the dialect in all the dialogues, looking for more realism.

An example of alternation between Standard Arabic and *dārija* can be observed in Muhammad Berrada's *Luʕba-t al-nisyān* (The game of forgetting), one of the most important novels of Arabic literature. In Morocco, indeed, this novel was included in the secondary education curriculum from 1995 to 2005 by decision of the Ministry of Education. In the field of poetry, it is important to mention Ahmed Lemsyeh, one of the founding fathers of contemporary Moroccan *zajal*, as well as Driss Mesnaoui (b. 1948), who employs the dialect to convey what Standard Arabic cannot express[7], and many others[8].

This evolution illustrates that the dialect is no longer intrinsically linked to notions of educational deficiency or social stigma; rather, it has emerged as 'a key element for the definition of a new Moroccan identity, or "Moroccanness"' (Caubet 2017b: 99). In the following paragraphs we will explore the transformation of linguistic practices within the field of advertising, examining the different forms and strategies through which advertising messages have been issued over the years.

4. The origin of advertising in Morocco

The concept of advertising traces its origins to the medieval Latin verb *advertere*, which means 'to turn or direct something toward'. Thus, the etymological origin of the term underscores the primary function of advertising: to capture and direct the attention of the audience (Danesi 2015). In essence, advertising encompasses various forms of public announcements aimed at informing potential customers about the availability, characteristics, and pricing of specific goods or services. Advertisers use a variety of media sources to reach customers effectively. Traditionally, advertisements were predominantly disseminated through print media such as newspapers and magazines. Another way adopted to convey advertising messages involved the utilization of billboards and posters, the so called 'outdoor advertising', which is still very widespread nowadays. However, with the advent of technology, the spectrum of advertising has expanded significantly. Television and radio remain powerful platforms providing audio-visual content that can engage audiences on a deeper level. Moreover, in the last decades, direct mail has transformed

[6] For a comprehensive examination of the key orthographic features used by Moroccans when writing in dialect, refer to Aguadé (2006).

[7] For further information see the interview with Driss Mesnaoui conducted by Deborah Kapchan (Kapchan 2022).

[8] For a more detailed study, refer to Moscoso et al. (2024).

with the integration of personalised content, while social media platforms and websites have emerged as vital channels for targeting advertising.

In a multilingual country such as Morocco, the introduction of innovative tools has undeniably influenced the linguistic choices employed in the dissemination of various advertising messages. Because of this, to fully comprehend the evolution that has occurred within this domain, it is necessary to examine the historical context and origins of the initial mediums utilized for broadcasting advertisements.

Until the end of the Middle Ages, in many countries, the primary strategy employed for transmitting advertising communications was the oral repetition of information in crowded public spaces, so that the message could reach as many people as possible. In Morocco, this practice was very common among the merchants, and the figure of a barker, called *bərrāḥ*[9], was often employed to perform this function (Boutahri 2018: 2). The utility of barkers can be attributed to the significant illiteracy rates among the population to whom the information was addressed. Orality, indeed, served as a mechanism to ensure that messages were accessible to everyone. For the same reason, government institutions frequently employed the figure of the *bərrāḥ* as a means of conveying public announcements and essential messages.

Alongside this figure, that persisted well after the Industrial Revolution, a new era of possibilities emerged in the 19th century. Although the immediate impacts of this phenomenon were predominantly experienced in Europe and North America, its influences gradually permeated other regions, including North African countries. In Morocco, factors such as trade dynamics and colonialism in the beginning of the 19th century played crucial roles in shaping the region's adaptation to industrial practices. Therefore, the introduction of newspapers at the conclusion of the 19th century marked an important evolution in the landscape of advertising media, gradually leading to the decline of the traditional role of the *bərrāḥ*. During the French Protectorate, indeed, this figure continued to operate predominantly within the rural regions of the country, highlighting the importance of oral communication in less urbanized areas, even as modern advertising methods began to take root (Boutahri 2018: 3).

As mentioned above, the latter half of the 19th century witnessed a series of initiatives aimed at establishing newspapers in Morocco. The primary objective of these endeavours was to persuade Moroccans of the advantages associated with the colonial occupation and modernism, in view of the forthcoming invasion. Thus, on November 7, 1904, Morocco saw the appearance of its first newspaper featuring advertisements, titled *Aṣ-Ṣabāḥ* (el-Ganbūrī 2010). This newspaper used the Arabic language in order to disseminate news about France, as well as to promote French commercial products, through advertisement often accompanied by illustrative imagery. The advertisements, indeed, were not solely focused on promoting colonialism. During the early 1900s, there were also announces for medical practices, oils, and other beauty products. This naturally fostered the establishment of a communication network connecting Moroccan population with foreign producers.

Moreover, film, hotel and travel companies advertising posters were also quite prevalent, particularly during the period in which Morocco had the status of French Protec-

[9] Literally, 'public crier'.

torate[10]. At that time "one of the oldest advertising materials were in the French newsletters coming from France and distributed in Casablanca and Rabat streets. The majority of those advertisements were offers of the hotels established by the colonizer in Casablanca" (Attar 2017: 7). The aim of these practices was above all that of convincing French people to move to Morocco.

Consequently, unlike the newspaper advertisements that targeted a Moroccan audience and employed Arabic language to promote foreign products, the posters and the newsletters distributed in the early 20th century predominantly used French as a means of expression, since they were directed at a French-speaking audience, in the form of an invitation to visit Morocco. Actually, roughly one-third of the posters from this era promoted the shipping line *Compagnie de Navigation Paquet*, which used to link Marseille to Tangiers and Casablanca twice a month (Lebouq 1911), and were headed by the sentence "Visitez le Maroc".

The final years of the colonial rule in Morocco aligned with the emergence of the initial radio and television stations. As will be discussed subsequently, these developments contributed to significant transformations within various facets of the advertising industry, including the linguistic strategies employed.

4.1. The advent of the radio in Morocco

The development of the audiovisual tools represented a significant advancement in Morocco, as the high illiteracy rate meant that journalistic contents were only available to a small portion of the population. For this reason, one could argue that a true mass communication only emerged with the introduction of radio and television. However, that's not completely accurate either. In reality, the first radio station established in Morocco, named *Radio Maroc*, can be traced back to 1928, and it primarily broadcasted its French-language programs to foreign residents living in the country. In the 1930s *Radio Maroc* began offering programs in Arabic, and from 1947 onward, it expanded its broadcasts to include both French and Arabic across two separate stations (Miller 2013: 92). In the meantime, however, several new private radio stations began broadcasting in various Moroccan cities (Jaïdi 2000), employing different languages for communication. Then, following Independence, there was a notable Arabization process that entailed an increased use of the Arabic language across various contexts, including on the radio. As may be expected, this phenomenon did not result in the homogenization of the linguistic landscape but instead emphasized the multilingual nature of Morocco. This has given rise to the establishment of new radio stations that incorporated both French and Arabic (Standard and/or Moroccan dialect), as well as Berber[11]. Meanwhile, the use of Spanish and English remained relatively limited. Nowadays, even with the proliferation of social media and other technological devices, the radio remains a vital advertising medium for Moroccans, because of its capacity to react to emerging trends. Among 39 radio stations

[10] The French colonial rule in Morocco lasted from 1912 to 1956.

[11] For a more comprehensive overview of the different radio stations that arose after the Independence and the language practices they adopted, refer to Miller (2013).

currently operating in Morocco[12], *Med Radio* and *Hit Radio* stand out as the most popular in terms of listener numbers. These can also be accessed online, which is how I was able to observe the typology and the linguistic nature of the advertisements broadcast. According to an analysis of the languages most frequently used for advertising content in *Med Radio* and *Hit Radio*, I could detect that, from a linguistic perspective, these two differ in some respects and are similar in others. It is important to note, first of all, that neither radio station chooses to adopt a linguistic homogenisation method. The three languages that are the most frequently used in advertisements are French, Moroccan Arabic and Standard Arabic. Linguistic alternations and instances of code-switching between these languages are not uncommon. However, while in *Med Radio* there is a prevalence of Arabic, both Moroccan and standard, the advertising phrases of *Hit Radio* prominently feature a significant incorporation of French. Moreover, while advertisements in *dārija* are certainly present in *Hit Radio*, Standard Arabic is used only sporadically. Anyway, the predominant style observed in both stations is a multilingualism, where announces can blend up to three languages, that is French, *dārija* and Standard Arabic. The predominance of French in *Hit Radio* could be explained by the fact that this station, launched in Morocco in 2006, is nowadays reaching audiences in twelve African nations where French is spoken as a second language[13].

Finally, English is represented by only a small number of lexical loanwords or expressions, and Berber remains completely outside the sphere of linguistic interaction. What distinguishes the two radio stations is the disparity in the quantitative distribution of these languages. Indeed, *Med Radio's* commercials are more frequently produced in *dārija* or in a mixed Arabic, with frequent shifts between *dārija* and Standard Arabic. I registered only few advertisements that were produced entirely in Standard Arabic or in French.

Thus, the multilingual advertising system is nothing more than an accurate reflection of the sociolinguistic landscape in Morocco. Not surprisingly, French and Standard Arabic, the languages of the administration and/or formal communication, are employed to disseminate information regarding political debates, national and international events or festivals organised under the direction of the monarchy or governmental authorities. On the other hand, advertisements in Moroccan Arabic frequently incorporate French insertions, particularly in the context of technical terminology, mirroring real-life linguistic interactions. These lexical insertions, indeed, are not regarded as foreign terms; rather, they are considered integral components of the Moroccan lexicon, worthy of inclusion in dialectal communication.

4.2. The spread of the television

Morocco was among those Arab countries who were the first ones to launch television broadcasting, even though the first television channel in Morocco had a rather short

[12] Morocco has currently sixteen public and twenty three private radio stations.

[13] Hit Radio broadcasts in Morocco, Central African Republic, Burkina Faso, Congo Brazzaville, Senegal, Togo, Gabon, Ivory Coast, Burundi, Chad, Niger and Comoros.

period of activity. The francophone channel *TELMA*, indeed, started broadcasting on February 28, 1954 but, due to financial issues, ceased its activities a little over a year later, in 1955. While the broadcasts on the channel were in French, there remained opportunities for public service announcements or advertisements endorsing commercial businesses operating in Morocco.

After this first experience, the launch of a public television station in Morocco did not occur until 1962. It was marked by a speech from King Hassan II during the Throne Day celebrations on *Al Oula* channel, known at that time as *TVM* (Télévision marocaine). During that period, Morocco initiated the policy of Arabisation, which resulted in the predominance of Standard Arabic in both radio and television. However, this approach did not last long, because as Hassa (2023: 263) points out:

> In the early 1980s there was a gradual shift toward more bilingual Arabic and French media as the vision of Morocco as a monolingual Arabic country seemed inadequate to allow Morocco to be competitive in the emerging global market.

This discourse, nonetheless, pertains more to television shows than to advertising. Indeed, since advertising serves also as a medium for establishing specific cultural and social connections with the audience, the use of Moroccan Arabic in TV commercials has never been a novel concept. Consequently, since the 1960s, TV advertising, which at that time were in black and white, have been predominantly delivered in the form of dialogues or parodies in Moroccan Arabic. This did not imply a rejection of Standard Arabic or French; rather, it was common for these commercials to feature an accompanying illustration of the product that included a brief description in either Standard Arabic or French, even when the language orally used was *dārija*.

Furthermore, in the context of television advertisements for international clothing or technology brands, automobiles, or financial institutions, the preference frequently leaned toward French. This selection was undoubtedly influenced by social factors. It is noteworthy, in fact, that French and Standard Arabic were mainly featured in commercials targeting the middle or upper classes of the society, whereas Moroccan Arabic was the most employed register for the promotion of food and hygiene products. This means that the choice of the language depended heavily on the target audience.

The situation today has not changed much. The primary languages used in TV commercials continue to be French, Standard Arabic and *dārija*, with the latter being the most dominant. However, it seems that the primary linguistic distinction between the oldest advertisements and the contemporary ones lies in the way these languages are employed. In modern advertisements, indeed, the occurrences of code-switching and code-mixing[14] are significantly more prevalent. Consequently, rather than producing different commercials in different languages, it's more likely to encounter multiple languages within the same advertisement. Numerous written messages are also presented during the commercials, maintaining a similar informal tone, as illustrated by examples (1) and (2).

[14] For a better understanding of the code-switching and code-mixing model to which I refer, see Appel & Muysken (1987: 118).

(1) بغيتو سمارتفون جديد؟
[*bġītu smārtfūn ždīd?*]
'Do you want a new smartphone?'

(2) Magix يرضيك و فحياتك يهنيك
[*Mažīks yrḍī-k w-f ḥyāt-ək yhannī-k*]
'Magix pleases you and eases your life'

As shown in the above examples, in TV commercials the *dārija* is predominantly represented through Arabic script, and this equally applies to linguistic insertions, which are regarded as components of Moroccan Arabic. Moreover, it is evident that both in auditory and visual advertisements, the majority of instances involving code-switching occur when promoting technological products or highlighting innovations that were previously non-existent, such as telephone lines and Wi-Fi. On the other side, similar to the trends observed in radio communications, a more formal linguistic style, characterised by the intersection of *dārija* and Standard Arabic, or by the use of French, emerges in the advertisements of national or financial institutions, as well as in health awareness messages. Once again, it seems that Standard Arabic and French are being used to give greater seriousness to the communication.

From this analysis, it can be said that the advertising communications presented on television exhibit a linguistic behaviour that parallels that of radio, except that TV commercials frequently incorporate brief written messages that compensate the oral content, maintaining the same register as the spoken language used in the communication. The subsequent sections will delve into the characteristics of written advertisements, emphasizing methods for effectively representing Moroccan Arabic in a written format.

5. Graphic representation of Moroccan Arabic in the billboards

From the linguistic point of view, Moroccan billboards exhibit a considerable diversity. The innovation in this context is not much caused by the introduction of written *dārija*, but rather by the various graphical representations through which it is expressed. The absence of an official standardization of Arabic dialects, indeed, results in the use and mixing of different spelling systems for their transliteration. This coincides with the observations of Caubet (2018: 400), who noted that

> more than fifteen years of experience in writing Darija, in Latin or Arabic script, have led to a situation where most connected Moroccans have now acquired fluidity in reading and writing Darija, through collective national effort".

In the 41 billboards analysed for this study, *dārija* emerges as the predominant linguistic register. Only ten of these billboards lack any representation of Moroccan Arabic, featuring French and Standard Arabic mainly for the promotion of foundations, fitness

centres, automobiles, and insurance services. In numerous instances, instead, Standard Arabic is juxtaposed with *dārija* on billboards.

I was able to identify four systems that were used for transcribing *dārija*, where the predominant method was characterised by the use of Arabic letters, as is the case in examples examples (3)-(5).

(3) كارط ڴيشي خلص وتيري فأي بلاصة بغيتي
[*kārṭ gīšī xallaṣ w-tīri f ʔay blāṣa bġīti*]
'Credit card, pay and withdraw wherever you want'

(4) الفيبر ديال أورنج وصلات لباب داركم
l-fībər dyāl ʔorōnž waṣlāt l bāb ḍār-kum]
'The Orange fiber has reached your door'

(5) البنك الشعبي معاك خطوة بخطوة حتى تشري دارك وتحقق أحلامك
[*l-bank əš-šaʕbi mʕā-k xaṭwa b-xaṭwa ḥta tšri ḍār-ək w-tḥaqqaq ʔaḥlām-ək*]
'The Popular Bank is with you step by step until you buy your house and your dreams become reality'

Various strategies are implemented in these messages, aimed at providing a written language as closely aligned as possible with the oral register. In (3), a common occurrence in *dārija* transliteration is observed, where the Persian character ڴ is employed to address the lack of the *g* phoneme in Arabic. Indeed, this latter exists in various dialects of Arabic and is not only associated with lexical borrowing, as in the above example, but can also arise from a different realisation of the Arabic etymological *q*. Staying on the subject of lexical loans, their integration in a written text reveal that they are regarded as components of Moroccan Arabic. For instance, the verb تيري *tīri*, which comes from the French *retirer*, and the nouns كارط *kārṭ*, بلاصة *blāṣa* and فيبر *fībər*, from *carte*, *place* and *fibre*, respectively, are normally inserted within a communication in *dārija*. Additionally, the presence of emphatic *ṭ* and *ṣ* in كارط *kārṭ* and بلاصة *blāṣa* highlights that these loans are transcribed according to the pronunciation they acquired in Moroccan Arabic. Consequently, despite the absence of a distinction between pharyngeal and non-pharyngeal phonemes in the European languages, the word *carte* and *place*are graphically adapted to the phonology of Moroccan Arabic. In بلاصة *blāṣa* we can note also the replacement of *p*, absent in the phonemic inventory of Classical Arabic, by *b*.

The will of preserving a language as colloquial as possible manifests not only at the lexical level but also in the orthographic representation of Arabic prepositions. Thus, in Example (3), the preposition ف *f-* 'in' is not separated by a space from the subsequent lexeme, as its counterpart في *fī* is in Standard Arabic. Its vowel, which is hardly heard, is not marked in writing. For the same reason, the Moroccan Arabic preposition *l-* 'to', written as لـ, is directly linked to باب *bāb* in (4). In the last example, the preposition *mʕa* مع 'with', followed by a second person suffix pronoun, marks the lengthening of *a*, as it happens in *dārija*.

These examples appear to highlight a degree of inconsistency in the transliteration of Moroccan Arabic. At times, the transliteration adheres to the conventions and rules od Standard Arabic, while in other cases, it aims to reflect the authentic pronunciation of Moroccan Arabic. Moreover, while lexical loans are adapted to the phonology of Moroccan Arabic in spoken contexts, a similar phenomenon occurs in the written form, which demonstrates that also in written advertisements, Moroccan Arabic can be a valid linguistic choice for those who want to expand their message to a wider audience. The use of *dārija*, which reflects the everyday habits of the majority of Moroccans, fosters a familiar environment with the target audience, which is a crucial factor in the advertising sector.

However, the Arabic script does not constitute a norm for the graphic representation of the *dārija*. As frequently observed on messaging platforms and on social media, the use of the Latin alphabet for writing in Moroccan Arabic is also an alternative. The primary reason for this decision stems from the fact that, with the spread of the first technological tools, keyboards lacked Arabic letters (Innaccaro & Tamburini 2021: 33). This absence compelled Arabs to resort to Latin letters for their communication needs. Subsequently, even after the incorporation of Arabic script into the keyboard, a lot of people, due to habit, have continued to use Latin script for the written expression of their dialect. Nowadays, when it comes to communication or conversation in dialect, this practice is still evident among Moroccans, especially young people. Seven of the advertising panels I captured feature Latin script used for conveying messages in Moroccan Arabic (examples 6-12):

(6) *Kenzup dima f' jibek* (Kenz'up)
'Kenzup always in your pocket'

(7) *Khalik hani* (Inwi)
'Take it easy'

(8) *Wajdine la Fibre Inwi?* (Inwi)
'Are you ready for Inwi fiber?'

(9) *Siiir b3id fine ma kenti* (Inwi)
'Go far away wherever you are'

(10) *Excelo: berrrrra3 rassek* (Excelo)
'Excelo: treat yourself well'

(11) *Partagi La7da, Machi Melts* (Pizza Hut)
'Share the moment, not Melts'

(12) *Kayn sahd? Kayn McDo!* (Mc Donald's)
'Is it hot? There is McDo!'

At first glance, it is evident that these messages are predominantly aimed at a youthful audience, who exhibit a notable interest in the advertising products. More precisely, Kenz'up is a multi-brand loyalty program that allows people to shop and earn points; the following three advertisements are basically internet offers, while the last two messages are aimed at promoting a new burger and an ice-cream proposed by the chains Pizza Hut and Mc Donald's, respectively. The recognition that the primary users of the advertised services are predominantly young individuals has resulted in the decision to employ Latin characters for these Moroccan Arabic messages. Consequently, this choice reflects the spelling system that is most commonly employed by young Moroccans and stems from the awareness that it will not pose any comprehension issues for the intended audience.

Here, too, different strategies for writing the dialect are put into action. First, it is important to highlight that an *e*, which has not to be pronounced, appears at the end of the word *wajdine* and *fine*, in Examples (8) and (9), respectively. This convention is employed to prevent ambiguity. Indeed, Latin characters coincide with the French ones, and since the French phonological system permits the realization of the nasal [n] only when the latter is followed by a vowel, the word *fin*, which is also part of the French lexicon, could be pronounced as [fɛ̃] in (9). In this sense, the inclusion of the letter *e* at the end of the word facilitates the correct pronunciation of the nasal sound, ensuring that the Moroccan term is phonetically articulated as [fĩn].

In order to convey the same emphasis that certain words and phrases possess in oral communication, the term *siiir* in (9) is represented with multiple instances of the vowel *i*. This repetition of *i* is deliberately designed to evoke the popular Moroccan football chant that gained significant recognition during the 2020 FIFA World Cup (Berrada 2022).

Further, in Examples (9) and (11) we can observe another particular feature: the incorporation of numerical graphemes to represent certain phonemes that are absent in the Latin alphabet, yet are present in both Standard Arabic and Moroccan Arabic[15]. For this reason, *b3id* and *la7da* have numbers representing ʕ and *ḥ*, respectively. The choice to include these numbers is in line with the principle that these messages are addressed primarily to young people, since "Romanized script slow down reading and obstruct comprehension especially in the case of adult readers who are not familiar with the numbers used to substitute phonemes for which English graphemes do not exist" (Al-Jarf 2021: 26). Moreover, it is noteworthy to observe that, while both Examples (7) and (9) are advertisements from Inwi, Example (7) does not employ the numerical grapheme "5" to denote the sound [x]. This distinction likely arises from the prevalent practice on social media, where the sound [x] is more commonly represented by the letters *kh* rather than by a numeral, although the latter remains a viable alternative. This phenomenon illustrates that advertisements often favor the most widely recognised and familiar spelling conventions, rather than adhering to a singular, fixed orthographic system, thereby catering to the preferences of their target audience.

[15] Apart from "7" indicating *ḥ* and "3" representing ʕ, there are other numerical graphemes used to compensate the lack of some Arabic phonemes in the Latin alphabet (Durand 2009: 32). In particular, "2" for the glottal stop *ʔ* and "9" for the uvular *q* are very frequent. In some rare cases, instead, "4" and "5" represent *ġ* and *x*, respectively.

From the lexical perspective, a colloquial register predominates, and nominal and verbal insertions from French are adapted to the morphological structure of Moroccan Arabic, as illustrated in Example (11) with the verb *partagi*. A further linguistic phenomenon to observe is the inclination to drop the *l-* of the article or the Arabic *li-* preposition, also reduced to *l-* in *dārija*, before a noun starting with *l*. In Example (8), the preposition *l-* would typically follow the participle *wajdine*; however, due to its position before the French article, it is omitted. This practice similarly applies to Example (11), where it is probable that the term *la7da* incorporates an assimilated article not graphically represented. This phenomenon may arise from the observation that, in spoken language, the article or preposition *l-* would have been naturally assimilated, thus remaining unperceived prior to a word beginning with *l*.

These examples demonstrate that there are no established conventions for writing in *dārija*. Instead, the objective is to grant familiarity by considering the predominant writing systems employed by the intended audience.

In certain instances, the use of multiple languages is accompanied by a change in spelling, resulting in each language being represented by a distinct writing system (examples 13-16):

(13) *WiFi Fibre, L'Max debit* [*l-kull maġribi*] لكل مغربي (Inwi)
'For every Moroccan the maximum speed, WiFi fiber'

(14) الهمزة [*əl-hamza*] + *Fibre, #inwi_m3ak avec la vitesse supérieure de la fibre* (Inwi)
'The fortune + fiber, #inwi_with you with the fiber's highest speed'

(15) *KitKat* [*xud l-ək*] خد لك *break* [*xud l-ək*] خد لك (KitKat)
'Take a break, take a KitKat'

(16) *La fibre d'Orange* [*kull-na mbbarʕīn b*] كلنا مبرعين ب (Orange)
'We are all doing very well with Orange fiber'

These messages incorporate an alternative spelling of a foreign noun or phrase. In Example (14), it is noteworthy that the slogan of Inwi (inwi_m3ak), although presented in Moroccan Arabic, is rendered in Latin script, adhering to the aforementioned writing conventions. In other cases, the Latin orthography is employed to preserve the original spelling of foreign terms incorporated into these sentences. In (13), an explanation for the fall of *e* in the French masculine article *le*, could be found in the fact that often in the French oral language, one tends to lose the pronunciation of the schwa, so that [lə] becomes [l]. Thus, the French article comes to coincide with the Moroccan article *l-*. In Example (15), instead, the introduction of *break* appears to be primarily driven by the intention to establish two analogous phrases within the message, creating thus a parallelism. Naturally, the interdental sound present in the Arabic verb أخذ *ʔ-x-ḏ* transforms into an alveolar occlusive sound, as is typical in Moroccan Arabic[16].

[16] Interdentals have undergone a phonological merger in Moroccan Arabic, whereby /t̪/ -> /t/, /d̪/ -> /d/ and /ḏ̪/ -> /ḍ/. Some exceptions can be found in the north-eastern Morocco. For further details, see Guerrero (2023).

This system arises from the desire to maintain the Latin spelling system for the languages that use it, such as French and English, while for *dārija*, which is an Arabic dialect, Arabic letters are employed. In this way, the phenomenon of code switching, which is very common in Moroccan linguistic landscape, becomes even more prominent, particularly when it comes to contents related to technology.

6. Conclusion

In this paper I examined the evolution of linguistic choices that emerge in the Moroccan advertising oral and written communication that took place over several decades. The findings demonstrate that advertising messages exhibit considerable linguistic diversity, thereby highlighting the intricate sociolinguistic environment of Morocco, where multiple languages coexist and interrelate. However, since for advertising communications it is essential to forge a connection with the target audience, the linguistic choices made within this context are never random. It is important to highlight that, although Berber is recognized as an official language, its presence in advertising remains nearly nonexistent. In the context of television commercials from *2M*, Berber appears solely in the daily schedule summary to denote graphically the days on which specific programs will air and is accompanied by Arabic and French. The analysis revealed that neither radio, nor TV used Berber for oral advertising. Moreover, none of the 41 billboards analysed in this study used Berber alphabet for written messages. Undoubtedly, the reason for this lies in the fact that the use of Berber would limit the accessibility of the message, as this language is not understood by all Moroccans. Indeed, the primary objective of advertising is to influence a broad audience; consequently, as this study clearly indicates, Moroccan Arabic is the dominant language utilized in the advertising sector. This language serves as a conduit for Moroccan culture and is comprehended by nearly the entire Moroccan population. It also fosters a sense of familiarity among the addressees.

Conversely, throughout the period of the French protectorate and its aftermath, Morocco adopted the French language within its administrative and educational environments. The current circumstances remain nearly unchanged, which explains the preference for French and Standard Arabic in highly formal settings. Thus, the present study highlights that advertisements endorsing activities organized by governmental entities or national institutions, as well as communications disseminated by the Ministry of Health and/or Education, aim to uphold a degree of seriousness and formality, which is achieved through the use of Standard Arabic and/or French. Also, as far as written messages are concerned, both on TV and billboards, Standard Arabic and French do not pose any problem, as they have their own officially recognised writing system.

The dynamics change when one aims to compose a message in Moroccan Arabic. Given that this dialect lacks any official status and doesn't have a standardised writing system, it leaves the possibility of adopting and/or mixing different systems. This lack of a standardised written system for Moroccan Arabic, however, does not diminish its presence on billboards. In this study, out of the 41 advertising panels that were captured, *dārija* was found in 31 of them, which represents 75% of the cases. However, what this

analysis revealed is that the selection of an orthographic system is intricately connected to the target audience of the advertisements. The representation of *dārija* in Latin script is evidently prevalent in advertisements targeting products that appeal to young people. This approach is grounded in the understanding that it will not hinder comprehension among the intended audience. Indeed, among young Moroccans, the use of Latin characters in messaging applications and social media platforms is very common. Consequently, the use of the Latin alphabet for writing *dārija* can be interpreted as a means to establish a better connection with the target audience, as well as to attract their attention, as this is a fairly recent practice in the advertising context. In contrast, when it comes to marketing products that may appeal to a broader audience, Arabic script is predominantly used for writing in *dārija*. This form of writing is particularly accessible to individuals who may not be accustomed to engaging with social media platforms.

The acknowledgment of the significant role that *dārija* plays in Morocco is undoubtedly a contributing factor to its increasing popularity in written contexts, in advertising and elsewhere. In recent years, indeed, the recognition of *dārija*'s status as a vehicle of Moroccan identity led to a notable expansion in its functionality. Consequently, written Moroccan Arabic is now an essential communication tool that can be utilized across various domains of social and cultural life. Within this framework, this study confirms that there is a redistribution of roles among the different languages that shape the socio-linguistic landscape of Morocco.

Appendix

The appendix illustrates the advertising messages featured on the 41 billboards that were captured and used for the present study.

Messages without any representation of Moroccan Arabic	
Message	Translation
Shoppez malin avec votre carte CIH *Bank chez* Celio.	Shop smart with you CIH Bank card at Celio.
Fitness Park*: rejoins-nous! 190 dhs le premier mois, puis 290 dhs/mois.*	Fitness Park: join us! 190 dhs for the first month, then 290 dhs per month.
Allianz*: Partenaire Mondial d'Assurances. Ensemble pour aller plus loin!*	Allianz: Global Insurance Partner. Together to go further!
Al boraq*: découvrez le confort inégalé à bord de nos trains* al boraq.	Al boraq: experience the unmatched comfort on board our al boraq trains.
Volkswagen*: le future démarre aujourd'hui. Nouveau* Touareg *3.OTDI 259cv BVA.*	Volkswagen: the future starts today. New Touareg 3.OTDI 259hp BVA.
Simplifiez vos paiements avec Attijariwafa Bank *et* Google Pay.	Simplify your payments with Attijariwafa Bank and Google Pay.
Aswak assalam: des promos bien étudiées pour la rentré.	Aswak assalam: well-designed promos the back-to-school.

Inwi*: les forfaits entreprises les plus généreux en internet. Sans engagement*.	Inwi: the most generous internet corporate packages. Without commitment.
حوّلوا أموالكم نحو حسابكم في الحين مع بنك أفريقيا. [*ḥawwilu ʔamwāla-kum naḥwa ḥisābi-kum fi l-ḥīn maʕa bank ʔifrīqya*]	Transfer immediately your money to your account with Bank of Africa.
مؤسسة محمد الخامس للتضامن تساهم في عملية مرحبا 2024. سعداء بإستقبالكم. [*muʔassasat muḥammad al-xāmis li-l-taḍāmun tusāhim fi ʕamalīyat marḥaba 2024. suʕadāʔ bi-ʔistiqbāli-kum*]	Mohammed V Solidarity Foundation contributes to the Welcome 2024 process. Happy to receive you.

Messages featuring Moroccan Arabic	
Message	Translation
ALSA الحل الاسرع للكوتي ديالك [*Alsa l-ḥall l-ʔasraʕ l l-gūti dyāl-ək*]	Alsa is the fastest solution for your snack
بالقليل...شري الكثير (BIM) [*Bīm, b-əl-qlīl…šri l-kṯīr*]	With less, buy more (BIM)
كارط كَيشي خلص وتيري فأي بلاصة بغيتي [*kārṭ gīšī xallaṣ w-tīri f ʔay blāṣa bġīti*]	Credit card, pay and withdraw wherever you want
الفيبر ديال أورنج وصلات لباب داركم [l-fībər dyāl ʔorōnž waṣlāt l bāb ḍār-kum]	The Orange fiber has reached your door
البنك الشعبي معاك خطوة بخطوة حتى تشري دارك وتحقق أحلامك [*l-bank əš-šaʕbi mʕā-k xaṭwa b-xaṭwa ḥta tšri ḍār-ək w-ṭḥaqqaq ʔaḥlām-ək*]	The Popular Bank is with you step by step until you buy your house and your dreams become reality
وفاسلف: إيلا لقيتي ما رخص نردو ليك الفرق [*wafasalāf: ʔīla lqīti ma rxaṣ nruddu l-ək əl-farq*]	Wafasalaf: If you find something cheaper, we'll refund the difference
جواز: أش كتسناو؟ دوز...وبراحتك فوز [*žawāz: ʔaš ka-tsənnāw? dūz...w-b-rāḥt-ək fūz*]	Jawaz: what are you waiting for? Go by and get your peace
البريد بنك: خليك قريب ليهم. مع البريد بنك وريا توصلوا بفلوسكم بسرعة وأمان [*al-barīd bank: xallī-k qrīb lī-hum. mʕa l-barīd bank w-rīya ttwaṣṣlu b-flūs-kum b-surʕa w-ʔamān*]	Al Barid Bank: stay close to them. With *Al Barid Bank* and *Ria* you get your money fast and safe
لافاش كيري ديما مرافقانا [*la-vāš-ki-rī dīma mrāfqā-na*]	*LaVacheQuiRit* is always with us.
اتصالات المغرب: عيط لحبابك مع نجمة 22 [*ʔittiṣālat al-maġrib: ʕayyṭ l-ḥbāb-ək mʕa nəžma 22*]	Maroc Telecom: call your loved ones with star 22.
Ikea: %80 تخفيضات حتال كل ما غتحتاجو لداركم بسعر منخفض [*ʔikīya: taxfīḍāt htā-l 80%. kull ma ġa-təḥtāžu l-ḍār-kum b-siʕr munxafiḍ*]	Ikea: discounts up to 80%. everything you need for your home at a discount.
MT cash خلص فواتيرك وخلي بالك هاني مع [*xallaṣ fawātīr-ək w-xalli bāl-ək hāni mʕa ʔem ti kāš*]	Pay your bills and let your mind rest with *MT cash.*

CIH Bank: خلص الضريبة ديالك فابور وغير بكليك [*se-i-āš bank: xallaṣ əḍ-ḍarība dyāl-ək fabūr w-ġīr b-klīk*]	CIH Bank: pay your fee for free and with just one click.
Markoub.ma: قطع الكار بسهولة من دارك [*markūb.ʔem-ʔa qaṭṭaʕ əl-kār b-suhūla mən dār-ək*]	Markoub.ma: book your coach with ease from home.
ONCF Yalla Morocco هاد العطلة غاتشوف كلشي مع بطاقة [*hād əl-ʕoṭla ġa-tšūf kull-šī mʕa biṭāqa-t yalla morōkko*]	ONCF During this holiday you will see everything with *Yalla Morocco* card.
Cote&Sport:النيڤو طلع. كاين لعب 1x2 وكاين لعب العباقرة [*kotē-ē-spōʁ: ən-nīvū ṭlaʕ. kāyn laʕb 1x2 w-kāyn laʕb əl-ʕabāqira*]	Cote&Sport: the level has risen. There are 1x2 games and there are games for geniuses.
الهمزة [*al-hamza*] + Fibre, #inwi_m3ak avec la vitesse supérieure de la fibre	The fortune + fiber, #inwi_withyou with the fiber's highest speed.
KitKat [*xūd l-ək*] خد لك break [*xūd l-ək*] خد لك	Take a break, take a *KitKat*.
Dalaa كبّرت بها وليداتي كاملين [*dalāʕ kabbart bi-ha wlīdāt-i kāmlīn*]	I raised all my children with *Dalaa*.
Orange: la fibre d'Orange كلنا مبرعين ب [*ʔoʁōnž: kull-na mbbarrʕīn b la fībʁ d orōnž*]	Orange: We are all doing very well with Orange fiber.
اتصلات المغرب: scroolé بلا حدرد [*ʔittiṣālāt al-maġrib: skrōlē blā ḥudūd*]	Maroc Telecom: scroll without limits.
Inwi: WiFi Fibre, L'Max débit [*l-kull maġribi*] لكل مغربي	For each Moroccan the maximum speed, WiFi fiber.
Inwi: Khalik hani	Inwi: Take it easy.
Wajdine la Fibre Inwi?	Are you ready for Inwi fiber?
Inwi: Siiir b3id fine ma kenti	Inwi: Go far away wherever you are.
Pizza Hut: Partagi La7da, Machi Melts	Pizza Hut: share the moment, not *Melts*.
Mc Donald's: kayn sahd? Kayn McDo!	Mc Donald's: is it hot? There is *McDo*!
Excelo: berrrrra3 rassek	Excelo: treat yourself well
Kenz'up dima f' jibek	Kenz'up always in your pocket.
Richbond: 7it testahlou ahssan ne3sa	Richbond: as you deserve the best sleep
Lamacom امممم...تبرعكوووم [*ʔammm...tbarraʕ-kum*] N°1 des ustensiles de cuisine au Maroc	*Lamacom* treats you very well. Number 1 kitchen utensils in Morocco.

References

Aguadé, Jordi. 2006. Writing dialect in Morocco. *Estudios de dialectología norteafricana y andalusí* 10. 253--274.

Al-Jarf, Reima. 2021. Impact of Social Media on Arabic language deterioration. *Eurasian Arabic Studies* 15. 16–32.

Amine, Khalid & Carlson, Marvin. 2011. *The theatres of Morocco, Algeria and Tunisia: Performance traditions of the Maghreb*. New York: Palgrave Macmillan.

Appel, René & Muysken, Pieter. 1987. *Language contact and bilingualism*. London – Baltimore: Edward Arnold.

Attar, Amina. 2017. *Media and advertising in Morocco: A brief history*. Fez, Morocco: Department of English Studies Faculty of Letters & Human sciences, Dhar El Mehrez. (Doctoral dissertation.)

Bassiouney, Reem. 2010. *Arabic and the media: Linguistic analyses and applications*. Leiden: Brill.

Berrada, Mohamed. 2022. Sir, Sir, Sir, le chant des supporteurs marocains devenu viral. *SNRT News*. (https://snrtnews.com/fr/article/sir-sir-sir-le-chant-des-supporteurs-marocains-devenu-viral-61443) (Accessed 2024--12-01.)

Boumans, Louis. 1998. *The syntax of code-switching: Analysing Moroccan Arabic/Dutch conversation*. Tilburg: Tilburg University Press.

Boutahri, Fairouz. 2017. La communication publicitaire au Maroc: Une naissance sous le masque du modernisme. *Sciences, language et communication* 2. 1-15.

Bouziane, Zaid. 2009. *Public service television policy and national development in Morocco*. Florida: University of South Florida. (Doctoral dissertation.)

Caubet, Dominique. 2008. From Movida to Nayda in Morocco: The use of darija (Moroccan Arabic) in the artistic creation at the beginning of the 3rd Millenium. In Prochazka, Stephan & Ritt-Benmimoun, Veronika (eds.), *Between the Atlantic and the Ocean, Proceedings of the 7th International Conference of AIDA*, 113-124. Vienna: LIT Verlag

Caubet, Dominique. 2017a. Darija and the construction of 'Moroccanness'. Bassiouney Reem (ed.), In *Identity and dialect performance: A study of communities and dialects*, 99-124. Abingdon – New York: Routledge.

Caubet, Dominique. 2017b. Morocco: An informal passage to literacy in Darija. *The politics of written language in the Arab world: Writing change*, Høigilt, Jacob & Mejdell, Gunvor (eds.), 116-141. Leiden: Brill.

Caubet, Dominique. 2018. New elaborate written forms in Darija: Blogging, posting and slamming in Morocco. In Benmamoun, El Abbas & Bassiouney, Reem (eds.), *Routledge handbook on Arabic linguistics*, 387-406. New York: Routledge.

Danesi, Marcel. 2015. Advertising discourse. In Tracy, Karen (ed.), *The international encyclopedia of language and social interaction*, 2-10. Hoboken, NJ: John Wiley & Sons.

Durand, Olivier. 2018. *Dialettologia araba*. Roma: Carocci.

Ennaji, Moha. 2002. Language contact, Arabization policy and education in Morocco. In Rouchdy, Aleya (ed.), *Language contact and language conflict in Arabic: Variations on a sociolinguistic theme*, 70-88. New York: Routledge.

Ennaji, Moha. 2005. *Multilingualism, cultural identity, and education in Morocco*. New York: Springer.

Ferguson, Charles. A. 1959. Diglossia. *Word* 15. 325-340.

el-Ganbūrī, Idrīs. 2010. *Al-ʔIqāma al-ʕĀmma al-Faransīya tuʔassis jarīdat* al-Saʕāda *li-turawwij ʔuṭrūḥatahā al-ʔistiʕmārīya wasaṭ al-Maġāriba*. Al-Masāʔ, 01-07-2010 (https://www.maghress.com/almassae/111740) (Accessed 2024-11-08)

Gintsburg, Sarali. 2020. Living through transition: The poetic tradition of the Jbala between orality and literacy at a time of major cultural transformations. *Rilce* 4(36). 202-222.

Guerrero, Jairo. 2023, Dialect variation across generations in Berkane (north-eastern Morocco): The case of interdental fricatives. *Miscelánea de Estudios Árabes y Hebraicos* 72. 69-86.

Hassa, Samira. 2023. French, a local language of radio and podcasts in Morocco. *Contemporary French Civilization* 48. 257-282.

Hoogland, Jan. 2013. L'Arabe marocain, langue écrite. In Benítez Fernández, Montserrat & Miller, Chaterine & de Ruiter, Jan Jaap & Tamer, Youssef (eds.), *Évolution des pratiques et representations langagières dans le Maroc du XXIe siècle* 1175-1188. Paris: L'Harmattan,.

Jaïdi, Moulay Driss. 2000. *Diffusion et audience des médias audiovisules*. Rabat: Al-Majal.

Kapchan, Deborah. 2022. Zajal: the Darija poets of Morocco. *The Markaz Review*. (https://themarkaz.org/zajal-the-darija-poets-of-morocco/) (Accessed 2025-04-17.)

Leboucq, Charles. 1911. Nos relations avec le Maroc. *Le Journal*. Cited from: "La compagnie de navigation Paquet", 2014, www.entreprises-coloniales.fr (https://www.entreprises-coloniales.fr7afrique-du-nord/paquet_la_cie.pdf) (Accessed 2024-11-12).

Michalski, Marcin. 2019. *Written Moroccan Arabic: A study of qualitative variational heterography.* Poznań: Wydawnictwo Naukowe UAM.

Miller, Catherine. 2012. Observations concernant la présence de l'arabe marocain dans la presse marocaine arabophone des années 2009-2010. In Meouak, Mohamed & Sánchez, Pablo & Vicente, Ángeles (eds.), *De los manuscritos medievales a internet: La presencia del árabe vernáculo en las fuentes escritas*, 419--440. Zaragoza: Universidad de Zaragoza.

Miller, Catherine. 2013. Evolution des usages linguistiques dans les nouvelles radios marocaines. In Benítez Fernández, Montserrat & Miller, Catherine & de Ruiter, Jan Jaap & Tamer, Youssef (eds.), *Evolution des pratiques et des représentations langagières dans le Maroc du 21ème siècle*, vol. 1, 89-118. Paris: L'Harmattan.

Moscoso García, Francisco & Aragón Huerta, Mercedes & Boutakka, Hassan & Gintsburg, Sarali. 2024. *Antología del zéjel contemporáneo marroquí*. Sevilla: Editorial Universidad de Sevilla.

Moustaoui, Srhir, Adil. 2019. La lengua árabe marroquí de los cartels: Análisis desde los estudios del paisaje lingüístico y la politica de la lengua. *Miscelánea de Estudios Árabes y Hebraicos* 68. 231-262.

Pellat, Charles. 1987. Malḥūn. In Bosworth, C. E. & van Donzel, E. & Lewis, B. & Pellat, Ch. (eds.), *Encyclopaedia of Islam*, 2nd edn, vol. 6, 247-257. Leiden: Brill.

Tamburini, Elena & Iannaccaro, Gabriele. 2021. "Siftlikom msg f tel opostitha hna rj3 l history": Mescolanze di codice nella comunicazione mediale in Marocco. *Annali di Ca' Foscari. Serie occidentale* 55. 29–64.

Ziamari, Karima. 2009. *Le code switching au Maroc: L'arabe marocain au contact du français*. Paris: L"Harmattan.

DOI: 10.14746/linpo.2025.67.1.8

Sawtone: A universal framework for phonetic similarity and alignment across languages and scripts

Omar Kamali

Omneity Labs
omar@omneitylabs.com | ORCID 0009-0006-5354-0328

Abstract: Processing text across different scripts presents significant hurdles in natural language processing, especially when dealing with non-standardized orthographies and informal writing systems common in low--resource languages. To address this, we introduce *Sawtone*, an integrated framework designed to enable consistent cross-script phonetic alignment and text normalization. At its heart is an architecture built for interoperability, combining a unified phonological feature space rooted in linguistic principles with modular, language-specific adapters. This structure allows for robust mapping and comparison between any pair of scripts. Crucially, it enables diverse adapters—developed using different methods or data—to work together cohesively for cross-language tasks. The framework readily supports alloglottographic text and is designed to function with minimal resource requirements. We demonstrate its practicality through implementations for transliteration, cross-script sequence alignment, and text normalization, further illustrated by a case study on preprocessing Moroccan Arabic data for Large Language Model (LLM) training. Initial results are encouraging: transliteration reached an 88% BLEU score, phonetic-based text sequence alignment achieved 87-95% accuracy across various language and script pairs, and text normalization significantly reduced variations in spelling. *Sawtone* offers a structured, interoperable foundation for advancing phonetic-aware NLP across linguistic boundaries.

Keywords: phonetics, cross-script alignment, low-resource languages, text normalization, transliteration, NLP, phonological alignment, phonological similarity, LLM training, Moroccan Arabic

1. Introduction

While the digital age connects us globally, bridging linguistic and cultural barriers, it also underscores the difficulty in processing text across different contexts and writing systems. This is particularly true for low-resource languages that often lack standardized spellings (Bird 2020). Many such languages exhibit alloglottography, where native

languages are written using borrowed scripts, resulting in diverse written forms (Crystal 2011; Unseth 2005).

These challenges affect several areas. Communication on social media is hampered by non-standard conventions and mixed scripts (Crystal 2011). Low-resource languages struggle with limited digital tools and competing writing systems (Bird 2020). Furthermore, digital preservation efforts face difficulties with cross-script search and retrieval of cultural heritage materials (Naji & Allan 2016). Addressing these issues requires robust phonetic handling and solutions designed for interoperability.

We introduce *Sawtone* (derived from Arabic *ṣawt* 'sound' plus *tone*), an integrated framework created for consistent cross-script phonetic alignment and text normalization, with a strong emphasis on interoperability. Its core architecture (Section 3) brings together:

- A unified phonological feature space (Section 3.2) grounded in linguistic principles, allowing for language-neutral sound comparison.
- A consistent phonetic similarity metric (Section 3.3) operating within this space.
- Modular, language-specific adapters (Section 3.4, Section 4.1) that map diverse orthographies—including non-standard and alloglottographic text—to the universal space.

This design separates the universal representation from language-specific processing. It allows adapters developed independently (using different methods or data) to function together seamlessly and therefore facilitating cross-script and cross-language phonetic processing. *Sawtone* is designed to work with minimal resources, making it suitable for low-resource scenarios. We demonstrate its utility through applications like transliteration (Section 5.2), cross-script alignment (Section 5.1), and text normalization (Section 5.3).

The paper unfolds as follows: Section 2 reviews related work. Section 3 outlines the *Sawtone* framework. Section 4 details the implementation methodologies. Section 5 showcases practical applications. Section 6 presents a case study involving Moroccan Arabic for LLM training. Section 7 discusses strengths, limitations, and future directions.

2. Literature review

Processing text across diverse languages and scripts, particularly non-standard varieties like dialects and informal text, presents significant challenges for Natural Language Processing (NLP). *Sawtone* aims to provide a unified framework grounded in phonetic similarity to tackle these issues. Here, we review related work in phonological representation, grapheme-to-phoneme conversion, text normalization, and the specific challenges posed by cross-script and low-resource contexts.

2.1. Phonological representation and feature systems

Achieving consistent sound representation across languages is crucial. Linguistics utilizes phonological features—abstract properties that differentiate phonemes. The Sound Pattern of English (SPE) (Chomsky 1968) introduced an influential set of binary features.

Subsequent theories have refined feature organization, such as Feature Geometry (George N. Clements 1985), and explored their grounding in phonetics (Ladefoged 1996). While linguistically rich, these systems can be complex to implement computationally. They often require expert knowledge and may struggle to capture gradient phonetic details or non-standard pronunciations.

More recent computational approaches often learn representations directly from data (e.g., phonetic word embeddings (Sharma 2021)), but these may lack explicit phonetic grounding. They can also face difficulties generalizing across different scripts or encountering unseen phonological phenomena. *Sawtone* bridges this gap by proposing a computationally tractable, multi-dimensional feature space (Section 3.2) inspired by established phonetic and phonological principles (Chomsky 1968). This provides an interpretable, quantitative foundation for comparing sounds.

2.2. Grapheme-to-Phoneme conversion and cross-script processing

Grapheme-to-Phoneme (G2P) conversion involves mapping written text to its sound representation. Existing tools like Epitran (Mortensen 2018) and Transphone (Li 2022) offer G2P capabilities but often assume standardized orthographies or lack contextual awareness, facing challenges with informal text or dialectal writing. Traditional approaches frequently lack unified mechanisms for handling arbitrary combinations of scripts (Karimi 2011).

Transliteration—converting text between scripts while preserving pronunciation (Knight 1998)—encounters similar hurdles. *Sawtone* addresses these through its adaptable "adapter" mechanism (Section 3.4, Section 4.1). These adapters map diverse orthographies into *Sawtone*'s universal phonetic space, thereby facilitating G2P, transliteration (Section 5.2), and cross-script alignment (Section 5.1) within a single, unified framework.

2.3. Text normalization for non-standard varieties

The prevalence of non-standard orthographies in user-generated content necessitates text normalization: converting variant forms into a canonical representation (Sproat 2016). Existing techniques include graph-based methods (Sonmez 2014), sequence-to-sequence models (Lourentzou 2019), and nearest neighbor approaches that leverage phonetic or contextual similarity (Ansari 2017; Elgeish 2019).

Normalizing highly variable text remains a difficult task, especially in languages like Arabic with complex morphology or diglossia (Darwish 2014; Habash 2010). Rule-based systems often struggle with the sheer scale of variation, while data-driven models might over-correct or fail when encountering unseen variants. *Sawtone*'s approach (Section 5.3) utilizes phonetic similarity within its universal space (Section 3.2) to cluster orthographic variants. This enables normalization based on phonetic equivalence, proving particularly beneficial for variations stemming from phonetic approximation or alloglottography, as demonstrated in the Moroccan Arabic case study (Section 6).

2.4. Low-resource languages and alloglottography

Many languages lack extensive digital resources and standardized orthographies (Bird 2020). Alloglottography – writing one language using another's script (Unseth 2005), common in digital communication like Arabeezi – often results in inconsistent, phonetically-driven spellings where standard NLP tools tend to perform poorly. *Sawtone* is specifically designed for these scenarios. It requires minimal resources and employs flexible adapters (Section 3.4, Section 4.1) capable of handling non-standard input. The "Crescendo Adapter Refinement" strategy (Section 4.3.4) further assists development in low-resource contexts.

2.5. Positioning *Sawtone*'s contribution

While previous work has addressed aspects of phonetic processing, there remains a need for unified, flexible frameworks that can handle diverse scripts, languages (including low-resource ones), and varying orthographic standards. Existing systems often rely on language-pair-specific rules, struggle with extensive variation, or demand large datasets. Although phonetic similarity measures exist, they are less commonly integrated into comprehensive, adaptable frameworks (Kondrak 2003).

Sawtone's contribution therefore lies in:

- Proposing a universal, quantitative phonological feature space (Section 3.2) that is both linguistically grounded and computationally tractable.
- Introducing a flexible adapter architecture (Section 3.4, Section 4.1) that accommodates various implementation strategies (rule-based, machine learning, hybrid) suitable for different languages and resource levels.
- Demonstrating practical applications in transliteration, cross-script alignment, and robust text normalization (Section 5), particularly for non-standard varieties like Moroccan Arabic (Section 6).

By bridging theory and practice, *Sawtone* aims to provide a more general and robust solution for phonetic-aware text processing across linguistic boundaries

3. Framework

In this section, we present the *Sawtone* Framework (Fig. 1), a flexible, modular framework for phonetic-aware text processing across linguistic boundaries with various downstream applications.

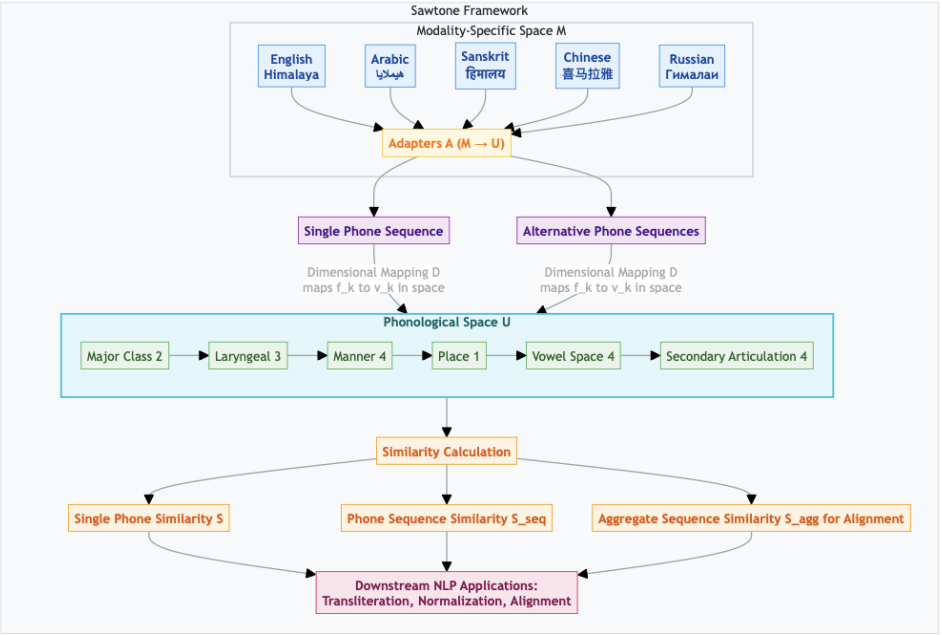


Figure 1: *Sawtone* framework architecture

3.1. Component architecture

The *Sawtone* framework facilitates cross-script phonetic alignment through interacting components centered around a universal phonological representation space.

We define the *Sawtone* framework Φ as follows:

$$\Phi = (U,S,A)$$

where:

- U: Represents the universal phonological space, defined by a feature set and a dimensional mapping .
- S: $U \times U \rightarrow [0,1]$: Is a similarity function that serves as a distance metric between phones within the space .
- A: Denotes a set of modality-specific adapters. These adapters map between the universal space and specific modality spaces M (such as text): $A_{(M \rightarrow U)}: M \rightarrow U$ (encoding) and (decoding).

3.2. Universal phonological representation space (*U*)

At the core of *Sawtone* lies a universal phonological representation space. This is a multi-dimensional vector space designed to quantitatively capture phonetic properties across different languages. Its primary goal is to provide a uniform, comparable representation where spatial proximity directly correlates with perceived phonetic similarity. This structure enables meaningful comparison of sounds originating from different languages or scripts—processed via potentially diverse adapters (Section 4.1)—using standard distance metrics.

The space integrates principles from phonetics and phonology, drawing upon concepts like distinctive features (Chomsky 1968), articulatory phonetics (Ladefoged 1996, 2011), Feature Geometry (George N. Clements 1985, 1995), IPA classifications (IPA 1999), acoustic phonetics (Stevens 1998), and insights from language-specific studies (De Premare 1998; Watson 2002). Features are selected for their orthogonality and are mapped onto a continuous scale suitable for distance calculations. Consequently, each phoneme or sound segment is represented as a vector within this space. We define 18 core feature dimensions based on this synthesis, aiming for comprehensiveness and strong phonetic grounding.

3.2.1. Phonological features (*F*)

The 18 dimensions, categorized below, are numerically represented on a scale from 0.0 to 1.0 (binary or continuous/multi-valued as detailed in Section 3.2.3 Dimensional Mapping (*D*)):

- **Major class (2 dims):** Consonantal, Sonorant (Chomsky 1968).
- **Laryngeal (3 dims):** Voice, Spread Glottis (reflecting aspiration/breathy voice), Constricted Glottis (reflecting glottalization/ejectives) (Gordon 2001; Ladefoged 1996, 2011).
- **Manner (4 dims):** Stricture (ranging from stop to vowel), Nasality, Laterality, TrillTap (Ladefoged 1996, 2011; Stevens 1998).
- **Place (1 dim):** PlaceArticulation (scaled from 0.0 for Bilabial to 1.0 for Glottal) (George N. Clements 1995; Ladefoged 1996, 2011).
- **Vowel space (4 dims):** VowelHeight, VowelBackness, LipRounding, TongueRoot (ATR – Advanced Tongue Root) (Archangeli 1994; International Phonetic Association 1999; Ladefoged 1996, 2011).
- **Secondary articulation (4 dims):** Labialized, Palatalized, Velarized, Pharyngealized (often associated with emphatic sounds) (Ladefoged 1996, 2011; Watson 2002).

Figure 2: *Sawtone* Phonological feature space

This 18-dimensional vector provides a rich, quantitative profile for each sound. Refer to Fig. 2 and Table 1 for an overview.

Table 1: Overview of *Sawtone* phonological feature dimensions

Feature Category	Dimensions	Specific Features	Primary References
Major Class	2	Consonantal, Sonorant	(*Chomsky, 1968*)
Laryngeal	3	Voice, Spread Glottis, Constricted Glottis	(*Gordon, 2001*; *Ladefoged, 1996*)
Manner	4	Stricture, Nasality, Laterality, TrillTap	(*Ladefoged, 1996*, *2011*; *Stevens, 1998*)
Place	1	PlaceArticulation (Scaled 0-1)	(*George N. Clements, 1995*; *Ladefoged, 1996*, *2011*)
Vowel Space	4	VowelHeight, VowelBackness, LipRounding, ATR	(*Archangeli, 1994*; *International Phonetic Association, 1999*; *Ladefoged, 1996*, *2011*)
Secondary Articulation	4	Labialized, Palatalized, Velarized, Pharyngealized	(*Ladefoged, 1996*, *2011*; *Watson, 2002*)
Total	**18**		

3.2.1.1. Suprasegmental features

Features like stress, pitch, and tone (Lehiste 1970), known as suprasegmentals, are generally absent in written text. Therefore, they are excluded from the current feature space. However, future extensions supporting speech modalities could incorporate them.

3.2.1.2. IPA lookup table (LUT)

The IPA lookup table serves as a critical bridge in the *Sawtone* framework, mapping International Phonetic Alphabet (IPA) symbols to their corresponding 18-dimensional feature vectors in the universal phonological space. This mapping is deterministic and language-agnostic, functioning as the foundation that enables cross-linguistic interoperability within our framework. During our work, the LUTs were constructed following the IPA (1999) and Ladefoged's (1996) guidelines.

Each IPA symbol (e.g., [p], [a], [ʃ]) is assigned a precise vector representation based on its phonetic properties across all 18 dimensions. For example, the voiceless bilabial plosive [p] would have high values for Consonantal and Stricture features, a value of 0 for PlaceArticulation (bilabial), and 0 for Voice. This standardized representation ensures that phonetically similar sounds across different languages occupy proximate positions in the universal space.

The lookup table is particularly valuable for:

- **Interpretability**: The LUT as an intermediate step allows for more interpretable adapters
- **Cross-script compatibility**: Enabling meaningful comparisons between sounds from distant writing systems (e.g., Arabic, Latin, Cyrillic)

- **Adapter interoperability**: Allowing adapter outputs to be comparable through a shared representation, regardless of their implementation details
- **Phonological analysis**: Providing a quantitative basis for studying sound patterns across languages

This standardized mapping layer ensures that all phonetic information, regardless of source language or script, can be represented in a consistent format within the universal phonological space, forming the backbone of *Sawtone*'s language-agnostic approach to phonological representation.

3.2.1.3. Phonetic inventory

Each language has its own phonetic inventory—the set of distinct sounds (phonemes) that can be expressed in that language. These inventories are typically defined using IPA symbols and can be sourced from multilingual dictionaries (De Premare 1998) or introductory linguistic materials for specific languages.

While collecting a language's complete inventory is valuable, it isn't always mandatory within our framework. When parallel text-to-IPA data is available, the inventory can be reconstructed automatically by extracting the IPA symbols that appear in the data. Regardless of the collection method, the inventory should contain only valid IPA phonemes to ensure accurate representation in the universal phonological space.

3.2.2. Universal phonological space (*U*)

The universal phonological space encompasses all possible 18-dimensional feature vectors where each dimension ranges from 0 to 1 ($[0,1]^{18}$). This space forms the domain for the similarity function:

$$U = \{\phi \in [0,1]^{18} \mid \phi = D\ (f)\ f \text{ or } f \in F\}$$

Representing sounds as vectors in this shared space allows for the calculation of phonetic distance (e.g., using weighted cosine distance, Section 3.3). The underlying hypothesis is that sounds perceived as similar will be located closer together within this space. Adapters (Section 4.1) map diverse orthographies into comparable locations within , ensuring cross-linguistic applicability. Notably, IPA features can be deterministically mapped to and from the *Sawtone* space, reinforcing its universality. The space itself is language-agnostic, accommodating language-specific variations through adapters and optional similarity weights. Special partial or superposition phones handle incomplete information (Section 3.4.1).

3.2.3. Dimensional mapping (*D*)

This process translates the 18 abstract features (f_k) into numerical values v^k within the range [0, 1], forming the vector $\phi_p = (v^1,...,v18)$ for a given phone p.

- **Binary features** (e.g., Consonantal): Mapped directly to either 0.0 or 0.1.
- **Categorical features** (e.g., VowelHeight, Stricture): For features with ordered categories (indexed $i = 0$ to $N = 1$), these are mapped to uniformly spaced points: $v^k = i/(N–1)$. For instance, 4 categories would map to .
- **Gradient features** (e.g., PlaceArticulation): Mapped linearly onto the [0, 1] range based on an established phonetic scale (e.g., Bilabial maps to 0.0, Glottal to 1.0).
- **Standard implementation:** Unless otherwise specified, these standard mappings are used. Range bounding is applied to adapter outputs to ensure numerical stability.
- **Limitations & future work:** Assuming uniform spacing for categorical features is a simplification. Future research could explore non-uniform mappings derived from acoustic or perceptual data (Kondrak 2003), or through data-driven learning, balancing enhanced accuracy with interpretability. This paper concentrates on the framework architecture using the standard mapping.

3.3. Similarity metric (*S*)

To compare individual phones and , we use their vector representations v_{p1} and v_{p2} with a **weighted cosine similarity** measure (Kondrak 2003) illustrated in Fig. 3:

$$S(p_1, p_2) = \frac{\sum_{k=1}^{18} w_k v_{p_1}^k v_{p_2}^k}{\sqrt{\sum_{k=1}^{18} w_k (v_{p_1}^k)^2} \sqrt{\sum_{k=1}^{18} w_k (v_{p_2}^k)^2}}$$

Here:
- v_p^k is the k-th feature value for phone *p*.
- w_k represents the weight for the k-th feature ($0 \leq w_k \leq 1$), reflecting its relative importance.

This calculation yields a score between 0 and 1, indicating the phonetic closeness of the two phones.

Feature weights (w_k): These weights allow for nuanced control over the similarity calculation. Using uniform weights (w_k =1 for all k) provides a strong baseline. Optimizing these weights, perhaps based on acoustic or perceptual data, remains an area for potential future work. For this paper, we assume uniform weights unless stated otherwise.

3.3.1. Phonetic sequence similarity (S_{seq})

To compare sequences of phones, $A = (p_{a_1},...,p_{a_m})$ and $B = (p_{b_1},...,p_{b_n})$, we employ **sequence alignment**, specifically the Needleman-Wunsch algorithm (Needleman, n.d.) illustrated in Fig. 3:
- **Match/Mismatch score:** Determined by the pairwise *Sawtone* similarity $S(p_{a_i}, p_{b_j})$ (eq. 3) between aligned phones.

- **Gap penalty (g):** A constant penalty (e.g., $g = -0.5$) applied when aligning a phone with a gap. This is a tunable hyperparameter.

This process yields a raw alignment score $S_{raw\_seq}(A, B)$. For comparison across sequences of different lengths, this score can be normalized, for instance, by the average sequence length:

$$S_{seq}(A, B) = \frac{S_{raw\_seq}(A, B)}{(m + n)/2}$$

This provides a robust measure of the overall phonetic resemblance between two sequences.

3.4. Modality-specific adapters

Adapters serve as the bridge between the universal phonological space and modality-specific spaces , such as text or potentially speech. They perform two key functions: encoding representations from a specific modality into , and decoding representations from back into a modality (Mortensen, 2018). See Methodology (Section 4.1) for implementation details.

$$A_{U \to M} : U \to M(Decode)$$

$$A_{M \to U} : M \to U(\text{Encode})$$

3.4.1. Handling phonetic ambiguity

Writing systems often provide incomplete phonological information, meaning the mapping performed by adapters may not be perfectly reversible (bijective) and can sometimes be lossy (Bird, 1994). To manage this, we introduce two special elements within :

1. Partial phone (ϕ_p): Represents a placeholder where certain features are unspecified (nullified) due to missing information. Formally, a partial phone contains undefined values for one or more feature dimensions in the vector. We use an underscore character to represent Partial Phones in phone sequences in this paper.

2. Superposition phone (ϕ_s): Represents multiple possible phonological interpretations simultaneously, useful for cases like heteronyms or inconsistent spelling conventions. Formally, a superposition phone is a weighted distribution over a finite set of possible phones $\{\phi_1, \phi_2, ..., \phi_n\}$ with associated weights $\{p_1, p_2, ..., p_n\}$ where $\sum_i p_i = 1$ for adapters that support it.

These elements allow the framework to represent ambiguity explicitly, carrying it through calculations, albeit potentially at the cost of some precision.

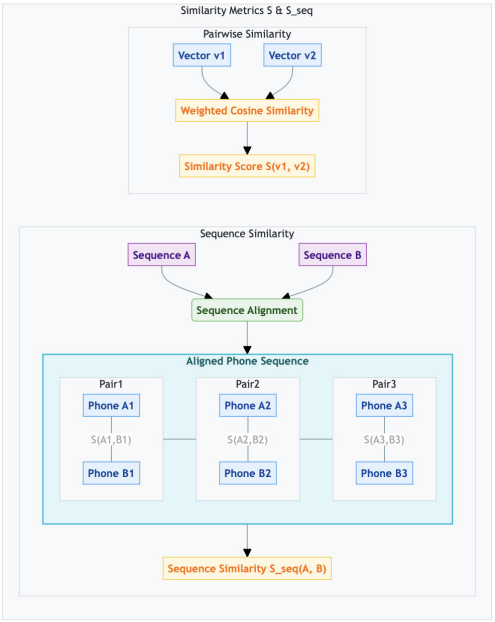


Figure 3: Similarity metrics S & S_{seq}

3.4.2. Calculating similarity with ambiguous phones (S_{agg})

Due to ambiguity, an adapter might generate multiple potential phonological sequences for a given input (e.g., a set $A = \{A_1, ...\}$ for input X_A, and $B = \{B_1, ...\}$ for input X_B). In such cases, we need a method to calculate an aggregated similarity score between these *sets* of sequences.

We first compute the pairwise sequence similarities S_{seq} (A_i, B_j) (Section 3.3.1) for all combinations and then aggregate these scores:

1. Average similarity: $S_{avg}(A,B) = \frac{1}{|A||B|}\sum_{i,j} \quad S_{seq}(A_i, B_j)$

2. Maximum similarity: $S_{max}(A,B) = max_{i,j} S_{seq}(A_i, B_j)$

3. Weighted average similarity: If adapters provide confidence scores (A_i) and (B_i) for each potential sequence:

$$S_{weighted\_avg}(A,B) = \frac{\sum_{i,j} \quad w(A_i)w(B_j)S_{seq}(A_i,B_j)}{\sum_{i,j} \quad w(A_i)w(B_j)}$$

Computational approximation: When the number of potential sequences $|A|$ and $|B|$ is large, computing all pairwise similarities can be expensive. We can approximate A_{agg} by:

1. Selecting representative subsets $R_A \subseteq A$ and $R_B \subseteq B$ (of size and , respectively) through sampling or clustering (e.g., using k-medoids).

2. Computing S_{seq} only for pairs within the Cartesian product $R_A \times R_B$.

3. Aggregating these $N_A \times N_B$ scores using one of the methods above.

The sizes N_A and N_B control the trade-off between computational cost and approximation accuracy.

3.5. Extension points

The framework's architecture is designed to be extensible. It allows for the integration of new components, such as different distance metrics or adapters for modalities like speech (Goldsmith 1990). Furthermore, individual components can potentially be used independently; for instance, the universal space could be employed solely for phonetic similarity calculations in Information Retrieval (IR).

Additional features (like suprasegmental or prosodic features) can be incorporated into the *Sawtone* space, and new adapters can be developed to handle them without altering the core framework structure.

4. Methodology

This section delves into the practical aspects of *Sawtone*, detailing the construction of its universal phonological space (Section 3.2) and outlining flexible strategies for implementing the adapters (Section 3.3). These adapters are crucial for mapping text to and from this universal space, accommodating diverse languages and varying resource levels. A primary consideration during adapter development is feasibility, especially given the resource constraints often associated with low-resource languages.

4.1. Modality-specific adapters ($A_{M\to\Phi}$ and $A_{\Phi\to M}$)

The effectiveness of *Sawtone* hinges on reliably mapping between diverse orthographies (graphemes) and the universal phonological space (encoding via $A_{M\to\Phi}$, (Fig. 4)), and potentially mapping back to a target script (decoding via $A_{\Phi\to M}$, (Fig. 5)). Writing systems exhibit significant variation—from alphabetic to syllabic, with spelling regularities ranging from highly consistent to quite irregular. Likewise, the availability of resources like dictionaries, corpora, and linguistic expertise varies widely. Consequently, no single adapter implementation strategy proves universally optimal.

Sawtone's flexibility allows for the integration of various techniques. We explored several complementary approaches:

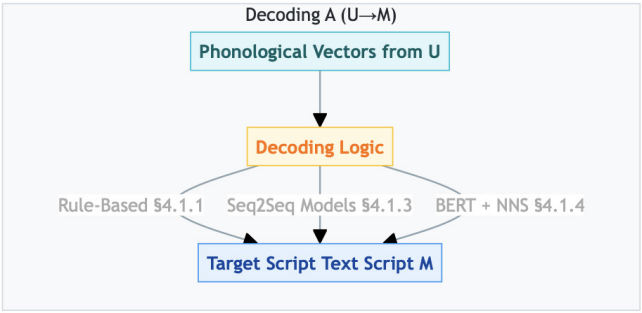


Figure 4: Encoding adapter architecture

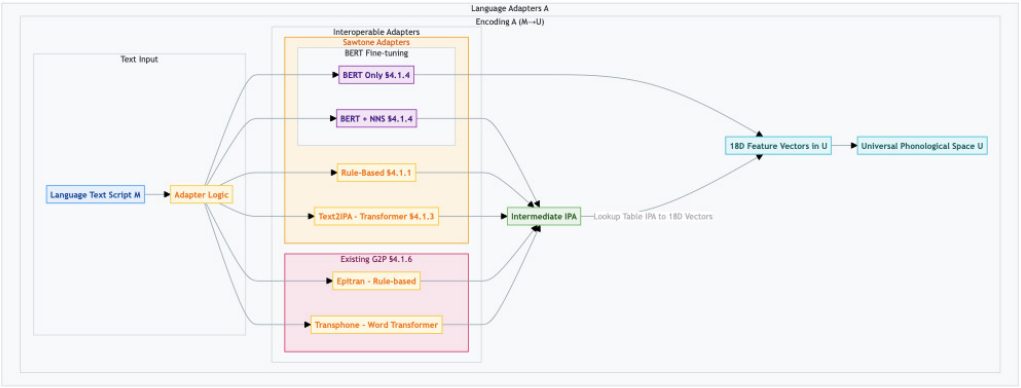


Figure 5: Decoding adapter architecture

4.1.1. Rule-based systems

- *Principle*: These systems encode expert linguistic knowledge into sets of context-sensitive rewrite rules (Fig. 6). These rules map sequences of graphemes to an intermediate phonetic representation (e.g., IPA), which is then converted to 18D feature vectors using a lookup table. We take inspiration from FST (Finite State Transducer) algorithms (Koskenniemi 1983) in the implementation of our rule-based adapters (Fig. 7).
- *Implementation*: Typically involves multi-layered, ordered rules formatted like *[ContextBefore, MatchSeq, ContextAfter, Replacement]*, performing text-to-text or text-to-IPA transformations. Successive layers handle progressively more complex phenomena, starting with basic mappings and moving to context-dependent rules and phonological processes. These systems can generate multiple interpretations to manage ambiguity (Section 3.4.1).

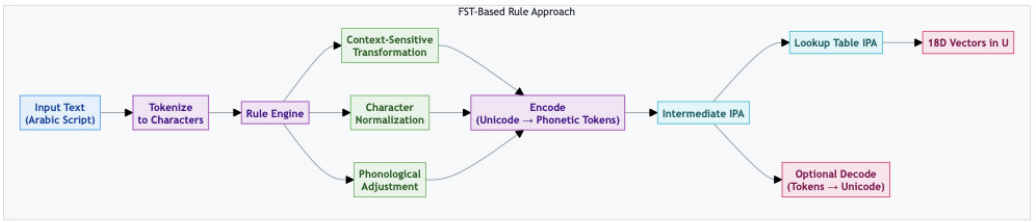


Figure 6: Rule-based adapter architecture

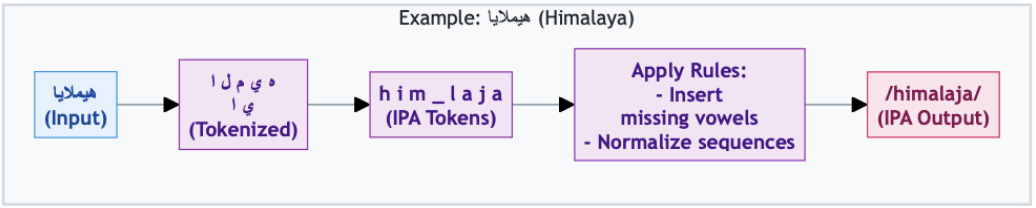


Figure 7: Rule-based adapter example (هيملايا → /himalaja/)

- *Decoding*: While rule-based decoding (IPA back to graphemes) is feasible, it can be a lengthy and costly process, especially for non-standard orthographies. Data-driven methods are often preferred for decoding.
- *Applicability*: Best suited for languages with regular orthographies or where linguistic expertise is readily available. Ensures outputs align well with the IPA-based features of the *Sawtone* space. However, development is labor-intensive, and the rules can be brittle when faced with irregular or noisy text.

4.1.2. A note on tokenization

Statistical methods generally require explicit text tokenization. For scalability, we focus on statistical tokenizers. Tokenization itself is complex due to the variable mapping between graphemes and phonemes (e.g., 'sh' mapping to a single phoneme /ʃ/ versus 's' and 'h' representing separate sounds in *misheard*) (Ni 2018). This presents a trade-off:

1. Single-character tokens: Risk splitting multi-character representations of single phonemes (like 'sh' into 's' and 'h').

2. Multi-character tokens: Risk merging adjacent characters that represent distinct phonemes (like 's' and 'h' in *misheard*).

Achieving perfect resolution typically requires language-specific contextual tokenizers, which is often intractable for language-agnostic systems. We approximate phoneme boundaries using multi-character tokenization, preferring potential ambiguity (merging) over definite information loss (splitting), relying on the adapters' contextual capabilities to interpret the resulting tokens correctly.

Since phonemes rarely span more than 2-3 characters, while standard tokenizers like BPE or WordPiece might create overly long tokens, we trained custom Unigram tokenizers for each language. We limited the maximum token length (typically 3 characters; 1 for ideographic scripts) and pruned rare tokens. This approach balances vocabulary size (around to tokens) with the ability to capture common multi-character phonemes. While cross-lingual tokenizers are possible, per-language tokenizers reduce the demands placed on the adapter models. Our Unigram approach offers a simple solution that mitigates the risk of splitting phonemes.

4.1.3. Sequence-to-Sequence (Text2IPA) models

- *Principle*: These models (Fig. 8) leverage standard transformer architectures (like T5 (Raffel, 2020), GPT-2 (Radford, 2019), Llama 3 (L. 3. Team, 2024), Qwen 2.5 (Q. Team, 2024)) trained on parallel corpora mapping text to IPA sequences. Large Language Models (LLMs) generally demonstrated better handling of out-of-domain input and performed better on longer sequences.
- *Implementation*: Typically involves training one model per language or language pair. The model predicts an IPA sequence from the input text; these IPA symbols are then mapped to 18D *Sawtone* vectors via a LUT (Section 3.2.1.2). Training data often comes from parallel corpora (e.g., Wikipedia, social media) automatically converted to IPA using phonetic dictionaries (Doherty, n.d.). Our largest dataset comprised 10 million sentences.

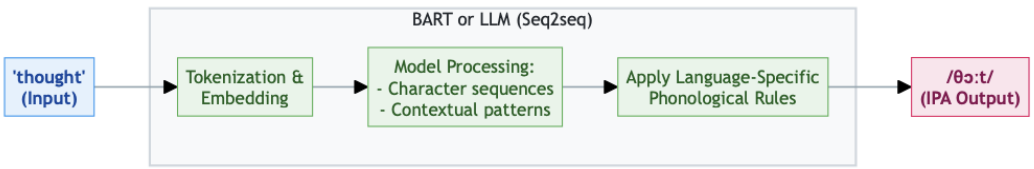


Figure 8: Text2IPA Adapter Example (‘thought’ → /θɔ:t/)

- *Decoding*: A reverse model (IPA -> Text) can be trained using the same data and methodology. Performance varies depending on the language pair and dataset quality.
- *Applicability*: Highly effective when sufficient aligned text-IPA data is available. These models can learn complex mappings and handle irregularities present in the training data. Outputs naturally align with the IPA-based *Sawtone* features via the lookup table mapping. Requires less manual effort than rule-based systems but is data-dependent and less interpretable.

4.1.4. BERT embedding regression and fine-tuning

- *Principle*: This approach refines existing text embeddings by training on pairs of phonetically related strings, such as homophones, transliterations, or rhymes. It aims to learn phonetic exceptions and nuances from large volumes of such paired data.

Languages with strict poetic conventions (like Arabic *buḥūr*, French metres, Sanskrit *chandas*) can particularly benefit. This method is empirically driven and capable of capturing subtle phonetic variations.

- *Decoding*: Decoding often involves a nearest neighbor search on the resulting vectors using the *Sawtone* phonetic similarity metric, selecting the closest phoneme from the target language's phonetic inventory.
- *Implementation*: We utilized a compact BERT-like model (based on MicroBert (Gessler & Zeldes 2022), with 64/128 dimensions), assuming phonetic information is less complex than full semantics. The model was first pretrained using Masked Language Modeling (MLM) on target language data. It was then fine-tuned using one of two strategies:
  1. *Embedding regression*: Training the model to project its BERT embeddings into the 18D *Sawtone* space using Mean Squared Error (MSE) loss against target vectors generated by another adapter (e.g., a rule-based or Text2IPA model acting as a "teacher"). The quality depends on the teacher model, but this allows for manual correction of teacher errors.
  2. *Semantic Textual Similarity (STS) adaptation*: Fine-tuning the model on pairs known to be phonetically close, such as homophones, transliterations, or lines from poetry/lyrics (e.g., Moroccan *zajal*/*melhoun*, Chinese *ci*, English rap). Using pairs that sound similar (not just identical) provides a potent signal for learning subtle phonetic relationships.
- *Applicability*: Useful for enhancing the phonetic awareness of embeddings, especially when large annotated datasets are scarce but lists of phonetically related pairs are available.

4.2. Comparative overview of adapter strategies

All adapter types ultimately map text to vectors in the universal phonological space, either directly or via an intermediate IPA representation. Rule-based and Text2IPA adapters typically perform this mapping via IPA. Text2IPA models are generally preferred for their ability to handle context effectively. BERT fine-tuning focuses on refining embeddings using phonetic similarity signals derived from paired data, which can be valuable for bootstrapping high-quality Text2IPA adapters. Table 2 summarizes key properties of these adapter strategies:

Table 2: Alignment with IPA-based features, resource requirements and performance

Method	Aligned with IPA Features?	IPA-Interpretability	Expert Rules?	Text-IPA Data?	Pair Lists?	Hours to Train	Converted Tokens/s
Rule-Based	Yes	Yes	Yes	No	No	Low	0.5M
Text2IPA (Seq2seq)	Yes	Yes	No	Yes	No	High	200
BERT Fine-tuning	Yes	Partial	No	Yes (as vectors)	Yes	High	15K

Computational resources: Figures are based on a single NVIDIA A100 GPU (SXM 80GB). Rule-based execution measured on a single core of an AMD EPYC 7551 CPU. **Performance**: Values provide a rough guide and depend heavily on data quality and specific implementation details.

4.2.1. Using existing G2P tools

Any method capable of producing IPA sequences from graphemes can be integrated with *Sawtone*. We explored Epitran (Mortensen, 2018) (which uses rules and lookup tables) and Transphone (Li 2022) (a word-level transformer trained on Wiktionary). Epitran behaves similarly to our rule-based adapters. Transphone resembles Text2IPA adapters but operates with limited context and wasn't trained on noisy text. Both are compatible with the *Sawtone* framework. A comprehensive evaluation across many languages would require suitable cross-lingual, phonetically-annotated datasets that reflect real-world variation, which is a direction for future work.

4.3. Practical considerations

4.3.1. Universality of the IPA features

Ensuring that the outputs from different adapters align consistently within the universal, IPA-based space requires two conditions:

1. The vector dimensions must be interpretable as (combinations of) IPA features.
2. The value ranges used for these features must be comparable across different adapters and languages.

This consistency is achieved by using a common LUT (Section 3.2.1.2) that maps IPA symbols to predefined *Sawtone* feature vectors. This LUT is used by methods involving an intermediate IPA representation (rule-based, Text2IPA), ensuring consistency. BERT vectors capture phonetic similarity implicitly but are not directly aligned with specific IPA features and cannot be interpreted as such.

4.3.2. Handling orthographic noise

Real-world text often contains typos and non-standard spellings. We employed an unsupervised filtering technique: clustering words with high orthographic similarity (low edit distance) that appear in similar contexts, and then filtering out low-frequency variants within these clusters as likely noise. This approach aims to balance normalization benefits with the preservation of legitimate rare forms.

4.3.3. What about diacritics?

Languages like Arabic utilize optional diacritics for determining correct pronunciation. Their absence leads to a loss of phonetic information. We evaluated several strategies to handle this:

- *Deterministic recovery*: Requires sophisticated POS tagging and morphological analysis, often infeasible or inaccurate for noisy, informal text.
- *Statistical inference*: Requires large, diacritized corpora matching the target text register; we attempted this for Arabic, finding it promising but unstable.
- *Treat absence as ambiguity (partial/superposition phone)* (Section 3.4.1): This was our choice due to its simplicity and effectiveness at handling various scenarios. It trades accuracy for robustness.

4.3.4. Crescendo adapter development strategy

The choice of adapter strategy depends on the target language and available resources. Rule-based systems (Section 4.1.1) are a good starting point if linguistic descriptions exist. Text2IPA models (Section 4.1.3) are viable if phonetic dictionaries or aligned corpora are available. BERT fine-tuning (Section 4.1.4) becomes an option if homophones, transliterations, or relevant poetic data can be compiled.

These methods can be employed iteratively in what we term the "Crescendo" strategy. For instance, simple rules can generate initial phonetic data to bootstrap BERT training; the improved embeddings from BERT can then be used to train more sophisticated Text2IPA models. This strategy allows for robust adapter development even for low-resource languages by progressively leveraging different types of data and techniques, starting with human expert knowledge. This contrasts with approaches requiring large, perfectly annotated datasets from the outset. Our development process involved iterative refinement and qualitative validation.

5. Applications of *Sawtone* to NLP tasks

This section illustrates how the *Sawtone* framework can be applied to core NLP tasks that involve cross-script interactions and non-standard text, namely: sequence alignment, transliteration and normalization.

5.1. Cross-script sequence alignment

Task Definition: This task involves aligning text sequences written in different scripts based on their underlying phonetic similarity, identifying corresponding sound segments (Figs. 9 & 10). It is fundamental for determining how phonetically related two strings are across script boundaries. The objectives are twofold: 1) Align sequences of phones, potentially introducing gaps where necessary. 2) Calculate an overall phonetic similarity score based on the alignment.

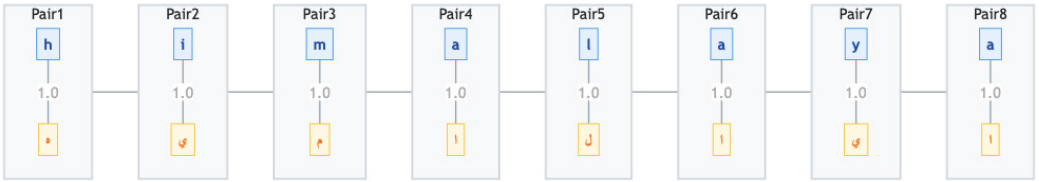


Figure 9: *Perfect Alignment, Score 1.0:* Comparing *Himalaya* with its Arabic transliteration هيملايا (himalaya).

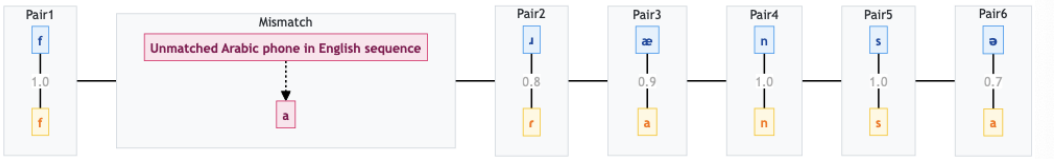


Figure 10: *Imperfect Alignment, Score < 1.0:* Comparing *France* with its common Arabic transliteration فرنسا (faransa). Note the vowel mismatch and the different 'r' sounds.

***Sawtone* Application**:

- *Encoding*: Encode the input sequences (Sequence X in Script A, Sequence Y in Script B) into *Sawtone* sequences using their respective adapters.
- *Sequence-Aligned Scoring*: Apply the Needleman-Wunsch algorithm (Section 3.3.1), using the *Sawtone* similarity metric (eq.3) to calculate match/mismatch scores between pairs of phone vectors with gap penalty .
- *Similarity Score*: Calculate the average similarity of the aligned vector pairs (excluding pairs involving gaps).

5.1.1. Experimental setup

Datasets: We used manually curated word pairs covering various languages and challenges:

- ParaNames (Sälevä, 2022) (Arabic/English names)
- Google Transliteration Dataset (*Rosca, 2016*) (Arabic/English general words)
- ANETAC (Ameur, 2019) (Moroccan Arabic/Arabeezi named entities)
- Internal French Homophone List
- Internal Japanese/English Homophone List
- Sentence pairs were generated where possible. These datasets typically averaged around 500 pairs per language pair.

5.1.2. Note on choosing adapters

For most languages, we utilized Text2IPA adapters (Section 4.1.3) as parallel script-to-IPA corpora were available for training. However, for Moroccan Arabic (MA), characterized by high variation and lack of initial parallel IPA data, we employed the Crescendo strategy (Section 4.3.4). This involved starting with a rule-based adapter (Section

4.1.1) and progressively refining it towards a Text2IPA model. Text2IPA adapters generally offer good flexibility and resistance to noise.

5.1.3. Moroccan Arabic adapters: Application of the Crescendo strategy

Developing effective adapters for MA was challenging due to its diglossic nature, non-standard spellings, and noisy data sources (Kamali & Abchir 2024). The goal was to map text written in both standard Arabic script and varied Latin-based Arabeezi scripts to the *same* underlying IPA sequence for consistent alignment. Since parallel script-IPA corpora were unavailable, the Crescendo strategy (Section 4.3.4) was essential. The multi-step development process (Rule-based -> BERT -> Text2IPA) is an iterative process that allowed us to generate progressively better training data (MA/Arabeezi -> IPA pairs), ultimately enabling the training of a robust Text2IPA adapter capable of handling the complexities of MA.

5.1.4. Adapters for other languages (English, Arabic, French, Japanese)

For other languages like English, Standard Arabic, French, and Japanese, we used Text2IPA adapters (Section 4.1.3) trained directly on existing parallel script-IPA corpora (averaging around 10 million words per language). These adapters proved effective, although their quality is inherently limited by the training data (e.g., word-level IPA conversion might miss phonetic changes occurring between words). Despite these minor limitations, they achieved good results. Further fine-tuning with homophone lists could offer marginal gains but was not deemed necessary for these initial experiments.

Evaluation: We assessed the system's ability to assign high similarity scores to phonetically related pairs (like transliterations and homophones) and low scores to unrelated pairs. Alignment accuracy was measured by calculating the edit distance between the predicted alignment and a ground truth alignment (counting gaps, normalized by the length of the longer sequence, with vowel/silent letter gaps potentially down-weighted).

5.1.5. Metrics and results

Alignment accuracy: Measured as 1 minus the normalized edit distance between the predicted and ground truth alignments. **F1 score**: The harmonic mean of precision and recall, evaluating the correctness of matched phone pairs in the alignment.

Table 3: Alignment system evaluation results across language pairs

Language pair	Alignment accuracy	F1 score
Arabic/English	0.92	0.94
Moroccan Arabic/Moroccan Arabeezi	0.90	0.92
Japanese/English	0.87	0.88
French/French (Homophones)	0.95	0.97

Discussion: The high accuracy and F1 scores reported in (Table 3) across diverse language pairs demonstrate *Sawtone*'s effectiveness in identifying phonetic relatedness, both across different scripts and within the same script (for homophones), by interpreting the underlying sounds represented by the graphemes. The examples below (Tables 4-8) illustrate the nuanced similarity scores produced by the system.

5.1.5.1. Arabic/English

Table 4: Arabic/English homophone pairs with Sawtone similarity scores

Native	Romanized	English	Sawtone similarity
سيل	sail	Sail	1.00
كمال	kamal	Camel	0.96
قرين	qarin	Karen	0.74
فم	fam	Fam	0.98
بر	bar	Bar	0.92
قلم	qalam	Column	0.63
كلب	kalb	Calf	0.75
دارت	darat	Dart	0.86
زهر	zahr	Tsar	0.78
لوم	lawm	Loom	0.83
أمنيتي	umniyati	Omneity	0.89

5.1.5.2. Moroccan Arabic/Arabeezi

Similarity scores between MA words written in Arabic script and their Arabeezi counterparts, including various spelling variants.

Table 5: Phonetic similarity between Arabeezi spellings of *tbarkellah* (تبارك الله) 'God bless'

Word	tbarkellah	tabarklah	tbraklah	tbrklh	tabaraka allah	tbar9ela7
tbarkellah	1.00	0.95	0.95	0.92	0.92	0.76
tabarklah	0.95	1.00	0.92	0.87	0.92	0.71
tbraklah	0.95	0.92	1.00	0.96	0.86	0.77
tbrklh	0.92	0.87	0.96	1.00	0.79	0.80
tabaraka allah	0.92	0.92	0.86	0.79	1.00	0.67
tbar9ela7	0.76	0.71	0.77	0.80	0.67	1.00

Table 6: Phonetic similarity between spellings of *ghanmchi* (غانمشي) 'I will go' in MA and Arabeezi

Word	hanmchi	ghanmchi	ghnmchi	ghannamchi	ghanemechi	غانمشي	غامشي	غنمشي
hanmchi	1.00	0.94	0.60	0.83	0.81	0.94	0.46	0.60
ghanmchi	0.94	1.00	0.61	0.87	0.87	1.00	0.53	0.61
ghnmchi	0.60	0.61	1.00	0.48	0.58	0.61	0.67	1.00
ghannamchi	0.83	0.87	0.48	1.00	0.90	0.87	0.45	0.48
ghanemechi	0.81	0.87	0.58	0.90	1.00	0.87	0.50	0.58
غانمشي	0.94	1.00	0.61	0.87	0.87	1.00	0.53	0.61
غامشي	0.46	0.53	0.67	0.45	0.50	0.53	1.00	0.67
غنمشي	0.60	0.61	1.00	0.48	0.58	0.61	0.67	1.00

5.1.5.3. Japanese/English

Table 7: Japanese/English homophone pairs with *Sawtone* similarity scores

Native	Romanized	English	Sawtone similarity
ユーセイ	yusei	You say	0.85
飴	ame	Amy	0.90
箸	hashi	Hushy	0.88
蜜	mitsu	Meets	0.87
待つ	matsu	Match	0.83
買う	kau	Cow	0.89
家	ie	Yeah	0.86
足	ashi	Ashy	0.91
波	nami	Nummy	0.82
夢	yume	You May	0.83
鳥	tori	Tory	0.80
花	hana	Honor	0.81
山	yama	Yammer	0.79
音	oto	Auto	0.92
竹	take	Tacky	0.75
餅	mochi	Mushy	0.76
見る	miru	Mirror	0.78
来る	kuru	Crew	0.77

5.1.5.4. French Homophones

Table 8: French homophone pairs with *Sawtone* similarity scores

Word 1	Word 2	Sawtone similarity
mer	mère	0.99
maire	mère	1.00
sain	saint	0.97
sein	ceint	0.93
vers	verre	0.92
vert	ver	0.94
saut	seau	0.94
sot	seau	0.91
sang	sans	0.97
cent	sans	0.88
fois	foie	0.93
foi	foie	0.98
pair	père	0.95
paire	père	0.98
pot	peau	0.97
champ	chant	0.89
compte	conte	0.91
comte	conte	0.96

Task Definition: Converting text from one writing script to another (Fig. 11) while aiming to preserve the original pronunciation as closely as possible (Knight 1998). It's crucial for tasks like Named Entity Recognition (NER), Cross-Lingual Information Retrieval (CLIR), and handling alloglottography.

5.2.1. *Sawtone* application

Transliteration using *Sawtone* involves two stages:

- **Encoding ($A_{(M\rightarrow\Phi)}$)**: The source text (in Script A) is first processed by its corresponding *Sawtone* adapter (Section 3.3) to generate a sequence of phonological vectors (Section 3.2).
- **Decoding ($A_{(\Phi\rightarrow M)}$)**: These phonological vectors are then fed into the adapter for the target script (Script B), which generates the corresponding grapheme sequence. The quality of the final transliteration depends on the accuracy of both the encoding and decoding adapters.

5.2.2. Experimental Setup

Dataset: We used the ParaNames dataset (Sälevä 2022), filtering it to obtain 1000 direct Arabic-to-English transliteration pairs, primarily focusing on named entities. **Adapter**:

Language-specific Text2IPA adapters were employed for both encoding (Arabic -> IPA/*Sawtone* vectors) and decoding (IPA/*Sawtone* vectors -> English), trained as described in the Methodology section.

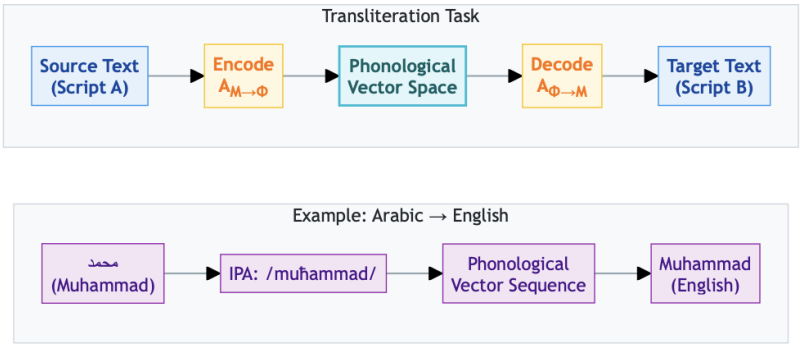


Figure 11: Transliteration with *Sawtone*

We allowed the decoding adapter to sample multiple potential outputs, ranked by perplexity. **Evaluation**: We generated English transliterations for the Arabic names in the dataset and compared them against the provided reference transliterations.

5.2.3. Metrics and results

Performance was assessed using several standard metrics (summarized in Table 9):

- **WER (Word error rate)**: Calculated using Levenshtein distance at the word level (treating each name as a single word). Lower scores indicate better performance.
- **CER (Character error rate)**: Levenshtein distance calculated at the character level. Lower is better.
- **BLEU score (Mean)**: Measures the overlap of character n-grams between the generated transliteration and the reference (treating the name as a sentence). Higher scores are better.
- **MRR (Mean reciprocal rank)**: Evaluates the quality of the ranked list of potential transliterations generated by the decoder. It measures how high up the correct reference appears in the ranked list, on average. Higher is better.

Table 9: Results for Arabic to English transliteration on ParaNames

Metric	Score
Word Error Rate	0.24
Character Error Rate	0.15
Mean BLEU Score	0.88
Mean MRR	0.67

Discussion: The results in Table 9 indicate reasonable performance. The high character-level similarity (reflected in BLEU and CER) suggests the phonetic mapping is largely successful, although there is room for improvement in generating the exact target word form consistently (reflected in WER and MRR).

5.3. Text normalization

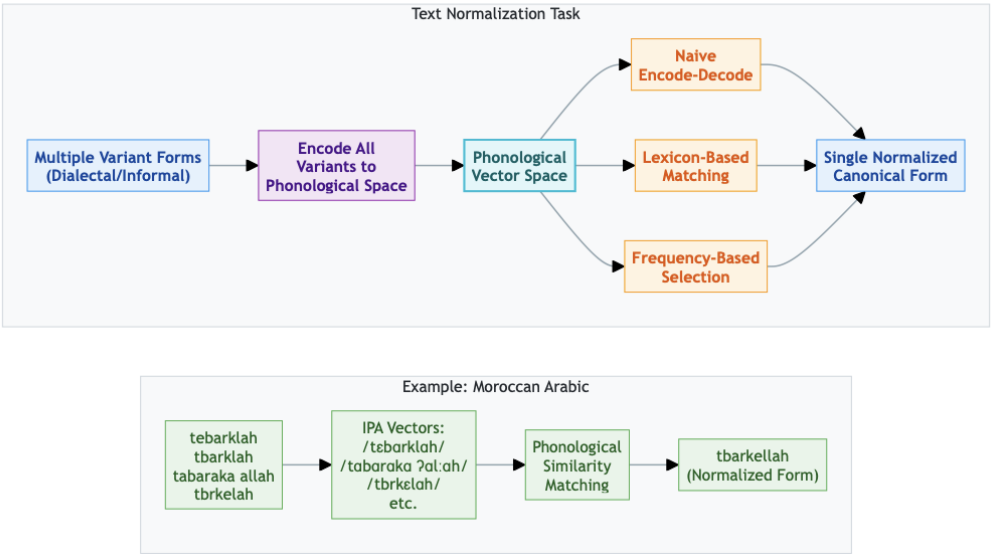


Figure 12: Text normalization and an example in Moroccan Arabic

Task definition: Converting non-standard text forms—arising from typos, dialectal spellings, informal abbreviations, etc.—into a standardized or canonical representation (Sproat 2016) as illustrated in Fig. 12.

***Sawtone* application**: *Sawtone* enables phonetic-based normalization. We encode words into the *Sawtone* space and then map phonetically similar variants to a single, chosen canonical form. We explored three approaches:

- *Naive encoding-decoding*: Simply encode the input word using its script's adapter ($A_{M\to\Phi}$) and immediately decode it back using the same adapter ($A_{\Phi\to M}$). The output serves as the normalized form. This is akin to transliterating a script onto itself via the intermediate phonetic space.
- *Lexicon-based*: Encode the input word. Search a predefined lexicon of canonical forms to find entries that are phonetically similar (above a certain threshold) using the *Sawtone* similarity metric (Elgeish 2019). Replace the input word with the most similar canonical form found.
- *Frequency-based*: First, build a frequency dictionary from a relevant corpus. Then, encode the input word. Find phonetically similar words (above a threshold) within the frequency dictionary. Replace the input word with the most frequent word among the similar candidates.

5.3.1. Experimental setup

Datasets:

- *Naive encoding-decoding*: We used a manually created dataset of 200 Moroccan Arabic (MA) sentence clusters. Each cluster contained 10 sentences (3-5 words each) representing the same sentence written differently, along with a ground truth canonical sentence form.
- *Lexicon*: We used a standard Arabic social media dataset (manually annotated) and a Standard Arabic lexicon compiled from various sources (containing ~100K words) (al-Khalīl 8th century; Ibn Ḥammād al-Jawharī 10th-11th century; Ibn Sīda 11th century).
- *Frequency*: We utilized a subset of a 1-million-word MA corpus gathered from online sources (Section 5.2) to build the frequency list.

Adapter: The same adapters developed for previous tasks were used here. **Thresholds**: Similarity thresholds for the lexicon and frequency-based methods were determined empirically via grid search. **Evaluation**: Output compared against ground truth (for Encoding/Decoding and lexicon) or against the most frequent word in the corpus (for Frequency).

5.3.2. Metrics & results

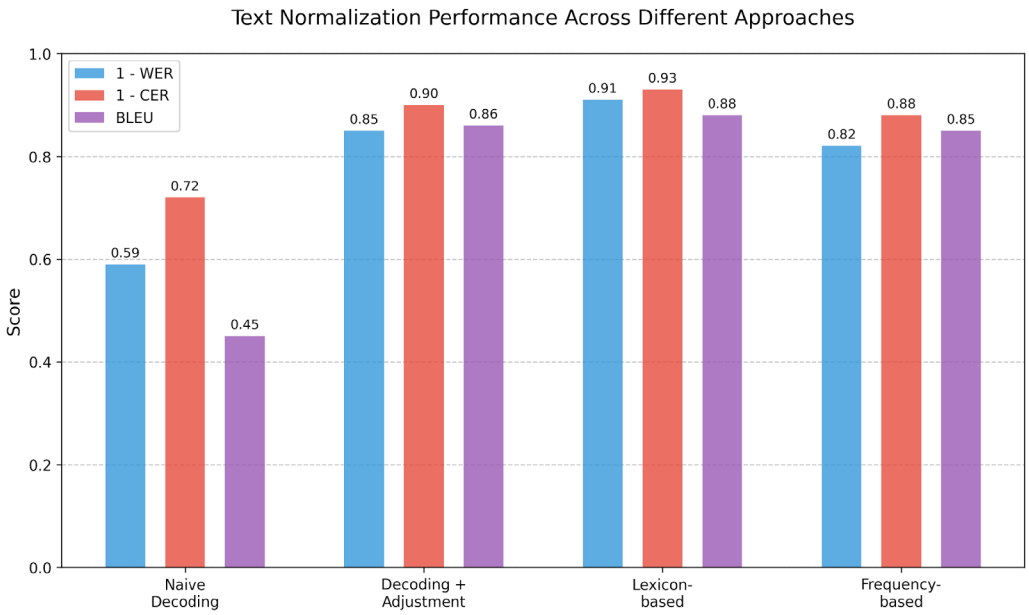


Figure 13: Comparison of accuracy for different normalization approaches

WER/CER: Word/character differences compared to the reference. Lower is better. **BLEU score**: Character sequence overlap compared to the reference. Higher is better.

Coverage (Lexicon/Frequency): The percentage of input words for which a phonetically similar variant was found above the threshold, allowing normalization to be attempted. **Processing speed**: Measured in tokens processed per second (after the initial build phase for lexicon/frequency methods).

Interpretation: It is important to note that the "ground truth" reference differs between methods (manual annotation vs. corpus frequency), making direct comparisons across methods challenging and the choice of methodology a circumstantial choice. Contextual normalization, which considers the surrounding sentence context, was not implemented due to dataset limitations and remains an area for future work. (Fig. 13) provides a visual comparison.

5.3.2.1. Normalizing by encoding and decoding

Comparing the output of the naive encode-decode method to the manual ground truth yielded less than satisfactory results (Table 10). This is likely because the adapter's inherent writing style might differ significantly from typical human writing conventions—a limitation of the specific adapter implementation rather than the *Sawtone* framework itself.

Table 10: Performance metrics for decoding-based normalization approach (vs. Human Ground Truth)

Metric	Value
Word Error Rate	0.41
Character Error Rate	0.28
BLEU Score	0.45
Processing Speed	15,000 tokens/second

However, evaluating the adapter against its *own* writing style as the ground truth reveals its intrinsic consistency (Table 11). Further rule-tuning could align the adapter's output more closely with human style.

Table 11: Performance metrics using adapter's writing style as ground truth

Metric	Value
Word Error Rate	0.15
Character Error Rate	0.10
BLEU Score	0.86
Processing Speed	15,000 tokens/second

5.3.2.2. Lexicon-based text normalization

This method encodes words and then searches a lexicon for the phonetically closest canonical form using the *Sawtone* similarity measure. It can be robust if a high-quality lexicon is available, although this is often not the case for low-resource languages (Bird 2020). Results on a Standard Arabic dataset are shown in (Table 12):

Table 12: Results for lexicon-based normalization on Standard Arabic dataset[1]

Metric	Value
Character Error Rate	0.07
Word Error Rate	0.09
Coverage	0.78
BLEU Score	0.88
Processing Speed	80,000 tokens/second

5.3.2.3. Frequency-based text normalization

This approach does not need a formal lexicon but is sensitive to the clustering threshold, noise within the corpus, and potential ambiguities (e.g., multiple common spellings, or morphologically related words that are phonetically similar but semantically incorrect matches). Results on the 1M-word MA corpus are presented in (Table 13):

Table 13: Performance metrics for frequency-based normalization[2]

Metric	Value
Character Error Rate	0.12
Word Error Rate	0.18
Coverage	0.91
BLEU Score	0.85
Processing Speed	80,000 tokens/second

5.3.3. Error analysis

We manually analyzed 100 errors from each normalization method. Table 14 shows an estimated breakdown of error types *within* each method, aggregated across the methods:

[1] Coverage indicates the percentage of input words matched to a lexicon entry above the similarity threshold.

[2] Speed measured after the initial frequency list construction.

Table 14: Distribution of error types within each normalization method

Error type	Encoding/Decoding (%)	Lexicon-Based (%)	Frequency-Based (%)
Ambiguous (Context Needed)	24	46	36
Legitimate Spelling Variations	14	13	39
Incomplete Phonological Mappings	51	20	15
Other Factors	11	21	10

- **Ambiguous (Context needed)**: Cases where correct normalization requires sentence-level context (constituting roughly 35% of total errors across methods).
- **Legitimate spelling variations**: Situations where multiple spellings are considered valid or common, making the choice of a single canonical form problematic (around 22% of total errors).
- **Incomplete phonological mappings**: Errors arising because the input orthography lacks sufficient phonetic information (e.g., missing diacritics, very terse spellings) (approximately 28.6% of total errors).
- **Other factors**: Including Out-Of-Vocabulary (OOV) words, unfortunate phonetic collisions between distinct words, length mismatches, etc. (about 14% of total errors).

6. Case study: Moroccan Arabic LLM training using *Sawtone*

To showcase *Sawtone*'s practical utility in a demanding, low-resource, high-variation setting, we applied it to preprocess data for training *Sawalni*, a Large Language Model (LLM) specifically designed for Moroccan Arabic (MA).

6.1. Challenges in processing Moroccan Arabic data

MA presents a unique set of challenges for NLP (Habash 2010):

- **Diglossia and script variation**: MA exists alongside Modern Standard Arabic (MSA) and is frequently written informally using both the Arabic script and various Latin-based Arabeezi systems, often mixed within the same text. Neither script has a standardized orthography for MA (Bouamor 2018).
- **Orthographic inconsistency**: The same MA word can be spelled in numerous ways, leading to extensive variation (Darwish 2014).
- **Phonological complexity**: MA possesses distinct phonetic features (like emphatic consonants and vowel reduction) that are often inconsistently represented in writing (Watson 2002).
- **Data scarcity**: Large, clean, and standardized corpora for MA are scarce (Bird 2020; Kamali & Abchir 2024).

Standard text preprocessing techniques often fall short in this context. An approach capable of handling cross-script variation and normalizing inconsistency based on underlying linguistic principles is highly desirable.

6.2. Data collection and initial state

We compiled a diverse text corpus for MA by gathering data from various online sources, including social media platforms, forums, and blogs. This resulted in a heterogeneous dataset containing MA written in both Arabic and Arabeezi scripts, often interspersed with MSA and French, and exhibiting extreme orthographic inconsistency.

6.3. *Sawtone*-powered preprocessing pipeline

We integrated *Sawtone* into our preprocessing pipeline specifically to tackle the challenges inherent in the MA data:

- **Adapter development**: We developed a *Sawtone* adapter tailored for MA, capable of mapping text from both Arabic and Arabeezi scripts into the universal phonological space (Section 3.2). This was achieved using the Crescendo strategy: starting with an initial rule-based adapter, using its output to train a BERT-based adapter, and further fine-tuning this adapter on MA/Arabeezi pairs along with MA-French and MA-MSA homophone pairs via an STS task.

- **Cross-script alignment and data augmentation**:
  - We encoded both Arabic-script and Arabeezi-script MA texts into the common *Sawtone* phonological space using the dedicated MA adapter.
  - We leveraged phonetic similarity (Section 3.4.2) to identify and align parallel text fragments across the two scripts within our corpus.
  - We trained a transliteration model (capable of converting between Arabic MA and Arabeezi) using *Sawtone*'s phonetic mappings as an intermediate representation.
  - We augmented the dataset by generating script-converted versions of monolingual texts, increasing the amount of parallel data available.
- **Phonetic normalization**: To address the pervasive orthographic variation, we applied *Sawtone'*s phonetic clustering capability (Section 4).
  - *Method*: Words were encoded into *Sawtone* vectors. Words whose vectors were closer than an empirically determined threshold were clustered together. A canonical form was selected for each cluster (typically based on frequency). All variants within a cluster were then mapped to this canonical form. We incorporated mechanisms to consider context for particularly challenging cases involving similar-sounding but distinct words.
  - *Example*: This process successfully reduced over 100 different observed spellings for the word *tbarkellah* (تبارك الله, see Table 5) down to a single canonical form. It also helped manage challenging cases like distinguishing *mchina* 'we went' from *machina* 'train', which are sometimes spelled identically in informal contexts.

6.4. Observed impacts on LLM training

Applying *Sawtone*-based preprocessing led to observable changes in the dataset characteristics and suggested potential positive impacts on the LLM training process:

- **Vocabulary reduction**: The number of unique word types in the dataset significantly decreased, from approximately 600,000 in the raw data to around 150,000 after processing. This reduction helps alleviate input layer sparsity for the LLM.
- **Improved data consistency**: Mapping diverse spelling variants to canonical forms based on phonetic equivalence resulted in a more consistent input stream for the LLM. This allows the model to better focus on learning underlying semantic and syntactic patterns rather than grappling with superficial spelling variations (Lourentzou 2019).
- **Enhanced cross-script handling**: The alignment and augmentation steps likely improved the model's ability to understand and generate MA across both Arabic and Arabeezi scripts by exposing it to explicitly linked parallel or phonetically related representations during training.
- **Training observations**: Preliminary comparisons between models trained on the raw dataset versus the *Sawtone*-normalized dataset showed an approximate 8% reduction in perplexity for the model trained on normalized data. This suggests potentially faster convergence and better generalization. We also observed more stable training dynamics, characterized by less variance in the loss function and fewer gradient spikes.

6.5. Discussion: Trade-offs in normalization

While phonetic normalization proved effective in managing orthographic chaos and reducing vocabulary size, it inevitably involves trade-offs. Aggressively normalizing text can lead to the loss of potentially meaningful sociolinguistic or stylistic information conveyed through non-standard spellings (Crystal 2011).

Our decision to normalize heavily was a pragmatic one for this initial LLM development effort, prioritizing model convergence and the capture of core linguistic patterns over the preservation of fine-grained orthographic detail. Future work could explore more nuanced normalization strategies, perhaps preserving particular variation types or training representations that remain sensitive to orthographic choices. Furthermore, phonetic ambiguity where normalization might merge similarly spelled but otherwise distinct words requires deeper integration of contextual information (Elgeish 2019).

This case study demonstrates the significant value *Sawtone* can bring to NLP development for low-resource, high-variation languages like Moroccan Arabic. By leveraging phonetic principles, it provides a principled way to bridge script differences and normalize text variation, offering crucial preprocessing capabilities, although the balance between normalization and information preservation warrants ongoing consideration.

7. Discussion

In this section, we reflect on *Sawtone*'s contributions, acknowledge its limitations, and outline potential directions for future research and development.

7.1. Strengths and contributions

Sawtone's primary contribution is its integrated architecture designed to enable consistent and interoperable phonetic-aware text processing across diverse languages and scripts. Its key strengths include:

1. Interoperability foundation: By decoupling the unified phonological space (Section 3.2) from modular, language-specific adapters (Section 3.4 Modality-specific adapters, Section 4.1), *Sawtone* creates a foundation where adapters developed independently—potentially using different methodologies or data sources—can function cohesively within the framework. This facilitates complex cross-script and cross-language tasks.

2. Consistent phonetic representation: The linguistically grounded, quantitative feature space (Table 1) provides a stable foundation for comparing sounds across languages, promoting consistency regardless of the specific adapters used.

3. Demonstrated applicability: We showcased *Sawtone*'s practical utility in core NLP tasks, including transliteration (Section 5.2, Table 9), cross-script sequence alignment (Section 5.1, Table 3, and text normalization (Section 5.3, Tables 10 & 14). The Moroccan Arabic case study (Section 6) further highlights its value, particularly for low-resource languages exhibiting high orthographic variation, by improving data quality for downstream tasks like LLM training.

4. Low-resource and non-standard focus: *Sawtone* is explicitly designed with low-resource languages, non-standard orthographies, and informal digital text (Section 2.4, Section 6.1) in mind. It offers a principled approach to handling variation based on phonetic equivalence, which is crucial for broadening the reach of NLP technologies.

In essence, *Sawtone* provides a structured, interoperable, and phonetically grounded approach that facilitates robust, comparable, and potentially collaborative NLP research and development, especially benefiting work on less-resourced languages and non-standard text varieties.

7.2. Limitations and future work

This initial presentation of *Sawtone* has several limitations that point towards important avenues for future work:

1. Comparative evaluation: The scope of this work did not include extensive benchmarking between the different adapter implementation strategies (rule-based, Text2IPA, BERT-based, see Table 2). Future research should conduct direct comparisons, focusing particularly on aspects like robustness to noise and sensitivity to context.

2. Evaluation rigor: The evaluations presented (Section 5) are preliminary. There is a need for larger, standardized cross-script datasets suitable for benchmarking. More rigorous evaluation protocols, including significance testing, are required. Furthermore, metrics for tasks like alignment and normalization need refinement; for instance, normalization evaluation (Tables 10-13) would benefit greatly from consistent datasets annotated against human-standardized ground truths.

3. Adapter depth and strategy: We did not deeply optimize the specific architectures of the adapters used, nor did we rigorously validate the effectiveness of the Crescendo

strategy (Section 4.3.4) across different scenarios. Further research is needed on optimal adapter implementations and practical testing of interoperability between adapters developed using different methods.

4. Feature space optimization: While the proposed 18-dimensional space (Section 3.2) proved effective for the tasks explored, a deeper analysis is warranted. This includes investigating the optimality of the chosen dimensional mapping, comparing it systematically to alternative feature systems, and conducting sensitivity analyses regarding feature choices and weights.

5. Normalization trade-offs: As discussed (Section 6.5), aggressive normalization can lead to information loss. Developing more nuanced, potentially context-aware normalization strategies (Section 5.3.3) that strike a better balance between achieving consistency and preserving meaningful variation is an important direction.

6. Ambiguity handling: The current mechanisms for handling phonetic ambiguity (Section 3.4.1, Section 3.4.2) are functional but could be improved. Exploring more sophisticated ranking or selection strategies for multiple phonetic interpretations could enhance accuracy.

7. Theoretical extensions: The framework could be extended theoretically, for example, by refining the feature space (incorporating dynamic weights, hierarchical structures, exploring phonological universals (Chomsky 1968), adding suprasegmentals (Lehiste 1970)), investigating advanced neural architectures (end-to-end models (Graves 2014), self-supervised learning (Baevski 2020; Conneau 2020; Devlin 2018), transfer learning (Pan 2020)), integrating audio modalities (for Speech-to-Text (Chan 2015; Wu 2016) or direct speech alignment (Watanabe 2017)), and developing standardized evaluation frameworks and benchmarks specific to cross-script phonetic tasks (Graham 2013; Karimi 2011; Wang 2019).

***. Technical improvements**: Practical enhancements could include optimizing computational performance – parallelization, efficient search algorithms, distributed processing – improving usability through better tooling, such as IDE plugins, web services, mobile app integration, and creating more readily available resources, such as annotated corpora (Kunchukuttan 2018), diverse datasets (Wang 2019), pretrained adapter models (Devlin 2018).

Addressing these limitations will require focused research efforts and likely community collaboration. *Sawtone* aims to provide a solid foundation upon which such interoperable phonetic processing advancements can be built.

8. Conclusion

In this paper, we introduced *Sawtone*, a universal framework designed for cross-script phonetic alignment and normalization, aimed at addressing the inherent challenges in processing text across diverse writing systems. By integrating principles from phonological theory with practical engineering considerations, *Sawtone* offers a flexible and robust system capable of handling arbitrary script pairs through the use of a shared, universal phonological representation space.

Its effectiveness has been demonstrated across multiple languages and applications, proving particularly valuable for low-resource languages and systems involving alloglottography. The case study focused on normalizing Moroccan Arabic data for LLM training (Section 6) highlights its potential for real-world impact. Key contributions include the universal feature space itself, the flexible adapter architecture promoting interoperability, the capacity for robust handling of non-standard text based on phonetic principles, and its design for minimal resource requirements.

As digital communication continues to expand and more of the world's languages gain a digital presence (Crystal 2012), frameworks like *Sawtone* become increasingly vital. They offer a path towards more effective cross-script text processing and play a role in preserving linguistic diversity in the digital realm (Bird 2020). We hope this work encourages further research and collaboration to expand *Sawtone*'s capabilities and broaden its impact.

Acknowledgments

This research was conducted by the *Sawalni Team* at *Omneity Labs* (echoing أمنيتي, Arabic for 'my wish'). We extend our gratitude to colleagues who provided valuable insights and to members of the Moroccan Arabic speaking community for their input during adapter development. We also thank the anonymous reviewers for their constructive comments and acknowledge the open-source community for the indispensable tools and libraries utilized in this work.

References

Ameur, M. S. H. & Meziane, F. & Guessoum, A. 2019. ANETAC: Arabic named entity transliteration and classification dataset. (arXiv:1907.03110).

Ansari, Z. 2017. Improving text normalization by optimizing nearest neighbor matching. https://doi.org/10.48550/arXiv.1712.09518.

Archangeli, D. B. & Pulleyblank, D. 1994. *Grounded phonology*. Cambridge, MA: MIT Press.

Baevski, A. & Zhou, H. & Mohamed, A. & Auli, M. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. https://doi.org/10.48550/arXiv.2006.11477.

Bird, S. 2020. Digital support for threatened languages: Progress and challenges. *Computer* 53(4). 82-85.

Bird, S. & Klein, E. 1994. Phonological analysis in typed feature systems. *Computational Linguistics* 20(3). 455-491.

Bouamor, H. & Hassan, S. & Habash, N. 2018. The MADAR Arabic dialect corpus and lexicon. In Calzolari, Nicoletta & Choukri, Khalid & Cieri, Christopher & Declerck, Thierry (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 3387-3396. Miyazaki: European Language Resources Association.

Chan, W. & Jaitly, N. & Le, Q. & Vinyals, O. 2015. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. (arXiv:1508.01211).

Chomsky, N. & Halle, M. 1968. *The sound pattern of English*. Chicago: The University of Chicago Press.

Clements, George N. 1985. The geometry of phonological features. Phonology 2(1). 225-252. https://doi.org/10.1017/S0952675700000440.

Clements, George N. & Hume, E. V. 1995. The internal organization of speech sounds. In Goldsmith, J. A. (ed.), *The handbook of phonological theory*, 245-306. Cambridge, MA: Blackwell.

Conneau, A. & Khandelwal, K. & Goyal, N. & Chaudhary, V. & Wenzek, G. & Guzmán, F. & Stoyanov, V. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, 8440-8451. Abu Dhabi: Association for Computational Linguistics.

Crystal, D. 2011. *Internet linguistics: A student guide*. London: Routledge.

Crystal, D. 2012. *English as a global language*. Cambridge: Cambridge University Press.

Darwish, K. 2014. Arabizi detection and conversion to Arabic. In Habash, N. & Vogel, S. (eds.), *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*. 217-224. Doha: Association for Computational Linguistics.

De Premare, A.-L. 1998. *Dictionnaire arabe-français (dialecte marocain)*. Paris: L'Harmattan.

Devlin, J. & Chang, M.-W. & Lee, K. & Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J. & Doran, C. & Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, vol 1, 4171-4186, Minneapolis: Association for Computational Linguistics.

Doherty, L. n.d. Ipa-dict – monolingual wordlists with pronunciation information in IPA. (https://github.com/open-dict-data/ipa-dict) (Accessed 2025-05-25).

Elgeish, M. 2019. Learning joint acoustic-phonetic word embeddings for speech recognition. (arXiv:1908.00493).

Gessler, L. & Zeldes, A. 2022. MicroBERT: Effective training of low-resource monolingual BERTs through parameter reduction and multitask learning. In Ataman, D. etc. (eds.), *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, 86-99. Abu Dhabi: Association for Computational Linguistics.

Goldsmith, J. A. 1990. *Autosegmental and metrical phonology*. Oxford: Basil Blackwell.

Gordon, M. K. & Ladefoged, P. 2001. Phonation types: A cross-linguistic overview. *Journal of Phonetics* 29(4). 383-406.

Graham, Y. & Baldwin, T. & Moffat, A. & Zobel, J. 2013. Continuous measurement scales in human evaluation of machine translation. In Pareja-Lora, A. & Liakta, M. & Dipper, S. (eds.), *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 33-41. Sofia: Association for Computational Linguistics.

Graves, A. & Jaitly, N. 2014. Towards end-to-end speech recognition with recurrent neural networks. In Xing, E. P. & Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*. 1764--1772. Beijing: PMLR.

Habash, N. Y. 2010. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies* 3(1). 1-187. https://doi.org/10.2200/S00277ED1V01Y201008HLT010.

Ibn Ḥammād al-Jawharī, Ismāʿīl. n.d. *Al-Ṣiḥāḥ fī al-lughah*. Bayrūt: Dār al-Fikr.

Ibn Sīda. 2000. *Al-Muḥkam wa-al-muḥīt al-aʿẓam*. Bayrūt: Dār al-Kutub al-ʿIlmiyya.

International Phonetic Association. 1999. *Handbook of the IPA: A guide to the use of the international phonetic alphabet.* Cambridge: Cambridge University Press.

Kamali, Omar. & Abchir, M. 2024. Finding Moroccan Arabic (Darija) in Fineweb 2. (https://huggingface.co/blog/omarkamali/gherbal-multilingual-fineweb-moroccan-arabic) (Accessed 2025-05-25).

Karimi, S. 2011. Machine transliteration survey. ACM Comput. Surv. 43. 17. https://doi.org/10.1145/1922649.1922654.

al-Khalīl, Ibn Aḥmad al-Farāhīdī. 2003. *Kitāb al-ʿAyn*. Bayrūt: Dār al-Kutub al-ʿIlmiyya.

Knight, K. & Graehl, J. 1998. Machine transliteration. *Computational Linguistics* 24(4). 599-612.

Kondrak, G. 2003. Phonetic alignment and similarity. *Computers and the Humanities* 37(3). 273-291. https://doi.org/10.1023/A:1025071200644.

Koskenniemi, K. 1983. Two-level model for morphological analysis. IJCAI'83. *Proceedings of the Eighth international joint conference on Artificial intelligence*, vol. 2, 683-685. Karlsruhe: Morgan Kaufmann Publishers.

Kunchukuttan, A. & Mehta, P. & Bhattacharyya, P. 2018. The IIT Bombay English-Hindi parallel corpus. In Calzolari, N. & Choukri, Kh. & Cieri, C. & Declerck, T. (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association.

Ladefoged, P. & Johnson, K. 2011. *A course in phonetics*. Boston, MA: Cengage Learning.

Ladefoged, P. & Maddieson, I. 1996. *The sounds of the world's languages*. Oxford: Blackwell Publishing.

Lehiste, I. 1970. *Suprasegmentals*. Cambridge, MA: MIT Press.

Li, X. & Metze, F. & Mortensen, D. & Watanabe, S. & Black, A. 2022. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In Muresan, S. & Nakov, P. &Villavicencio, A. (eds.),

Findings of the Association for Computational Linguistics: ACL 2022, 2106-2115, Dublin: Association for Computational Linguistics.

Llama 3 Team. 2024. The llama 3 herd of models. (arXiv:2407.21783).

Lourentzou, I. & Manghnani, K. & Zhai, C. X. 2019. Adapting sequence to sequence models for text normalization in social media. In Calzolari, N. & Choukri, K. & Cieri, C. & Declerck, T. (eds.), *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)*, 335-345. Munich: AAII.

Qwen Team. 2024. Qwen2.5 technical report. (arXiv:2412.15115).

Mortensen, D. 2018. Epitran: Precision G2P for many languages. In Calzolari, N. & Choukri, K. & Cieri, C. & Declerck, T. (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2710-2714. Miyazaki: European Language Resources Association (ELRA).

Naji, N. & Allan, J. 2016. *On Cross-Script Information Retrieval*. 9626. 10.1007/978-3-319-30671-1_70.

Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3). 443-453. DOI:10.1016/0022-2836(70)90057-4.

Ni, J. 2018. Multilingual grapheme-to-phoneme conversion with global character vectors. *Interspeech 2018*. 2823-2827. https://doi.org/10.21437/Interspeech.2018-1626.

Pan, L. 2020. Multilingual BERT post-pretraining alignment. In Toutanova, K. & Rumshisky, A. & Zettlemoyer, L. & Hakkani-Tur, D. & Beltagy, I. & Bethard, S. & Cotterell, R. & Chakraborty, T. & Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 210-219. Abu Dhabi: Association for Computational Linguistics.

Radford, A. & Wu, J. & Child, R. & Luan, D. & Amodei, D. & Sutskever, I. 2019. Language models are unsupervised multitask learners. (https://api.semanticscholar.org/CorpusID:160025533) (Accessed 2025-05-25)

Raffel, C. & Shazeer, N. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. (arXiv:1910.10683).

Rosca, M. & Breuel, T. 2016. Sequence-to-sequence neural network models for transliteration. (arXiv: 1610.09565).

Sälevä, J. & Lignos, C. 2022. ParaNames: A massively multilingual entity name corpus. (arXiv:2202.14035).

Sharma, D. 2021. Learning phonetic word embeddings. (arXiv:2109.14796).

Sonmez, O. 2014. Graph-based text normalization. In Moschitti, A. & Walter, B. & Daelemans, W. (eds.), *Proceedings of the 2014. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 313-324, Doha: Association for Computational Linguistics.

Sproat, R. & Jaitly, N. 2016. RNN approaches to text normalization: A challenge. (arXiv:1611.00068).

Stevens, K. N. 1998. *Acoustic phonetics*. Cambridge, MA: MIT Press.

Unseth, P. 2005. Sociolinguistic parallels between choosing scripts and languages. *Written Language & Literacy* 8(1). 19-42.

Wang, A. & Pruksachatkun, Y. & Nangia, N. & Singh, A. & Michael, J. & Hill, F. & Levy, O. & Bowman, S. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In Wallach, H. & Larochelle, H. & Beygelzimer, A. & d'Alché-Buc, F. & Fox, E. & Garnett, R. (eds.), *Advances in neural information processing systems, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2190-2194. Vancouver: Association for Computational Linguistics.

Watanabe, S. & Hori, T. & Kim, S. & Hershey, J. R. & Hayashi, T. 2017. Hybrid CTC/attention architecture for end-to-end speech recognition. *Journal of Selected Topics in Signal Processing* 11(8). 1240-1253. https://doi.org/10.1109/JSTSP.2017.2763455.

Watson, J. C. E. 2002. *The phonology and morphology of Arabic*. Oxford: Oxford University Press.

Wu, Y. & Schuster, M. & Chen, Z. & Le, Q. V. & Norouzi, M. & Macherey, W. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. (arXiv:1609.08144).