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].

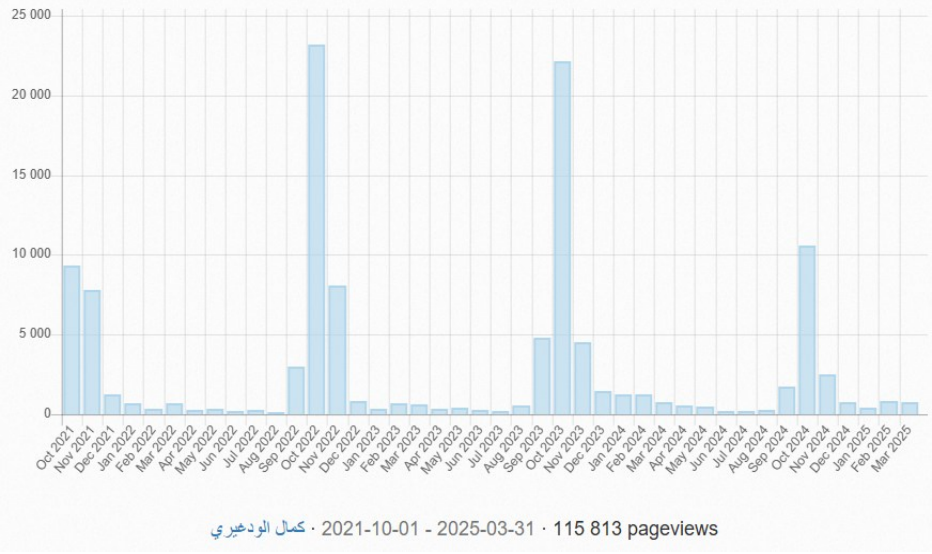


Figure 1: Number of pageviews for Kamal Oudrhiri's Moroccan Darija Wikipedia page

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.

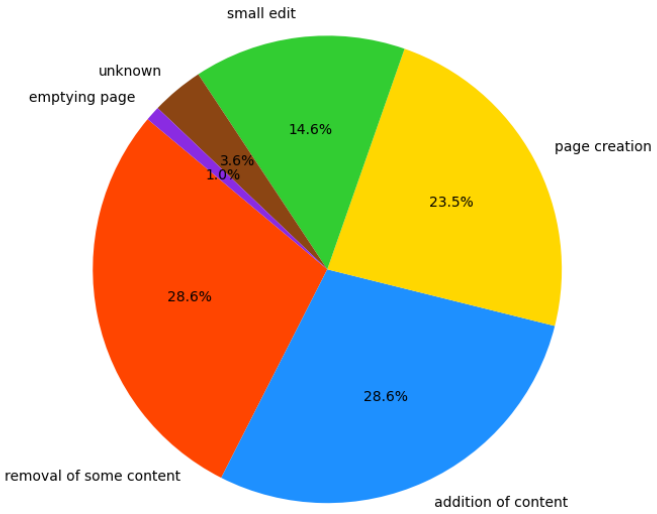


Figure 2: Mapping of different types of disruptive edits by anonymous users

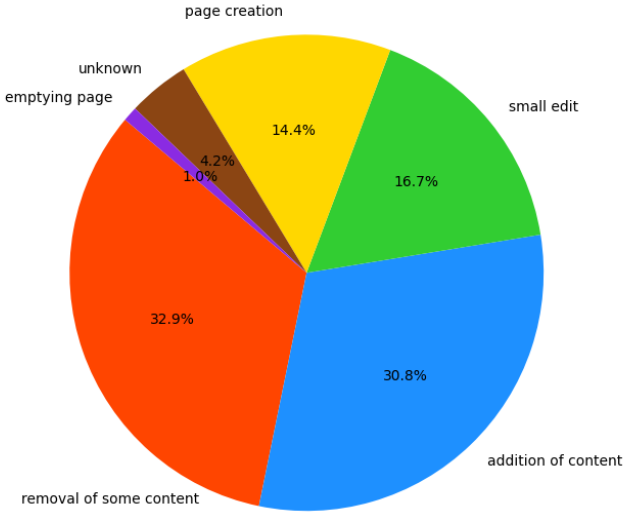


Fig 3. Mapping of different types of disruptive edits by all users

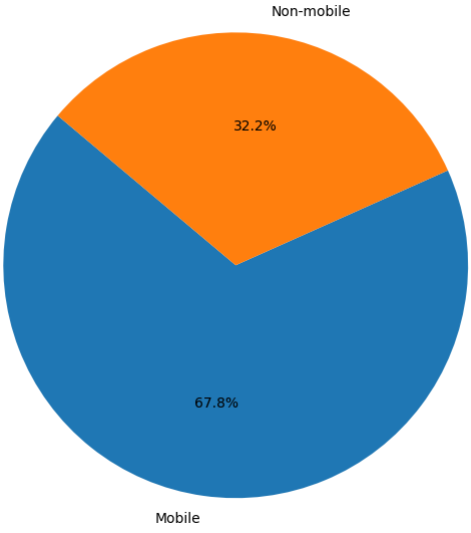


Figure 4: Percentage of disruptive edits made from a mobile device vs other devices for all users. Anonymous and account specific disruptive edits show a similar trend (68.6% and 64.5% respectively)

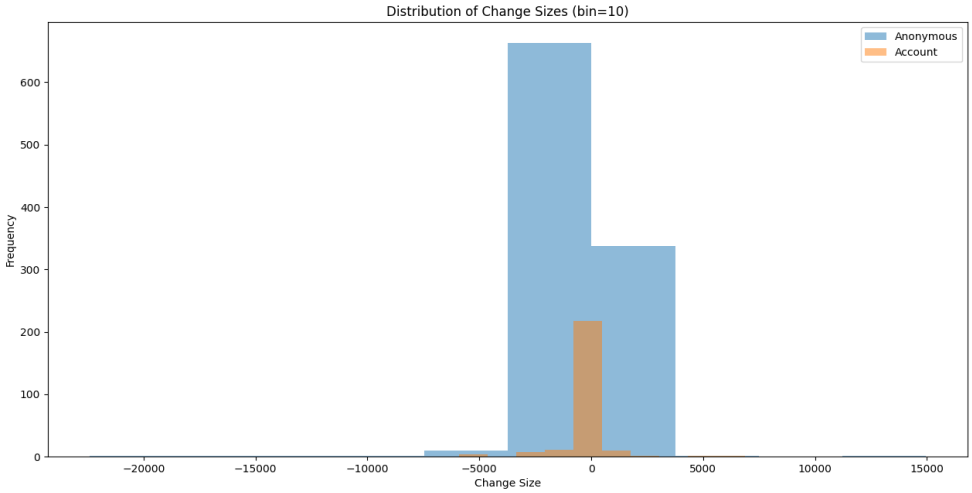


Figure 5: While disruptive edits by registered users are centered around 0 bytes and are smaller in size, anonymous disruptive edits tend to be larger and lean towards the negative (removal of content)

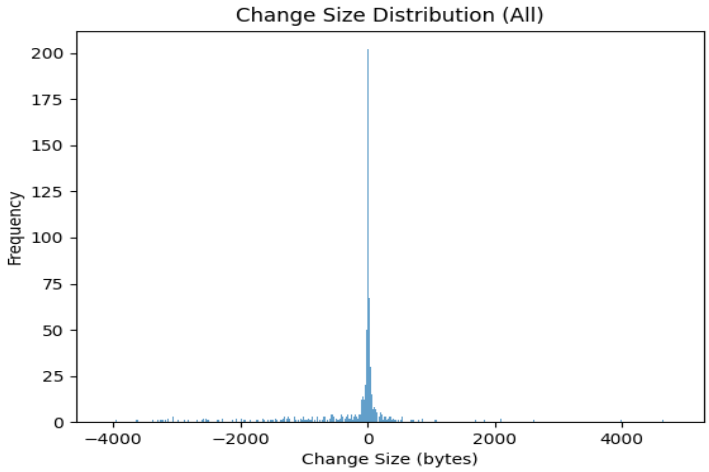


Figure 6: The majority of disruptive edits are small edits (close to 0 bytes added or removed)

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]

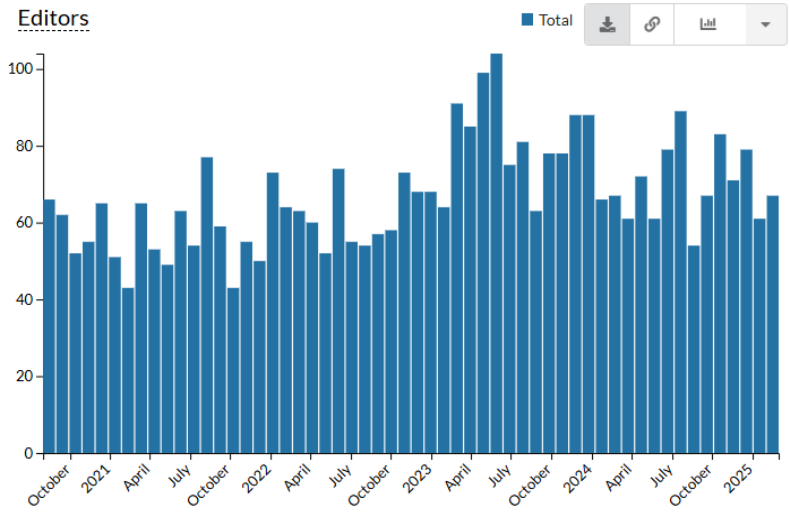
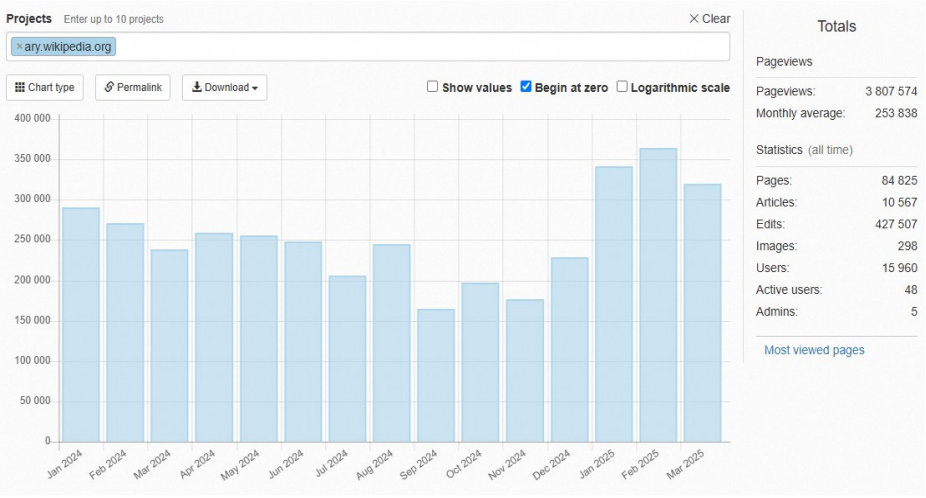


Figure 8 presents page views of Darija Wikipedia in 2024-2025. While variating from one month to another, they still show a slight but consistent increase on average. The peak activity in some periods may be due to the contests that have been organized and attracted many occasional editors.

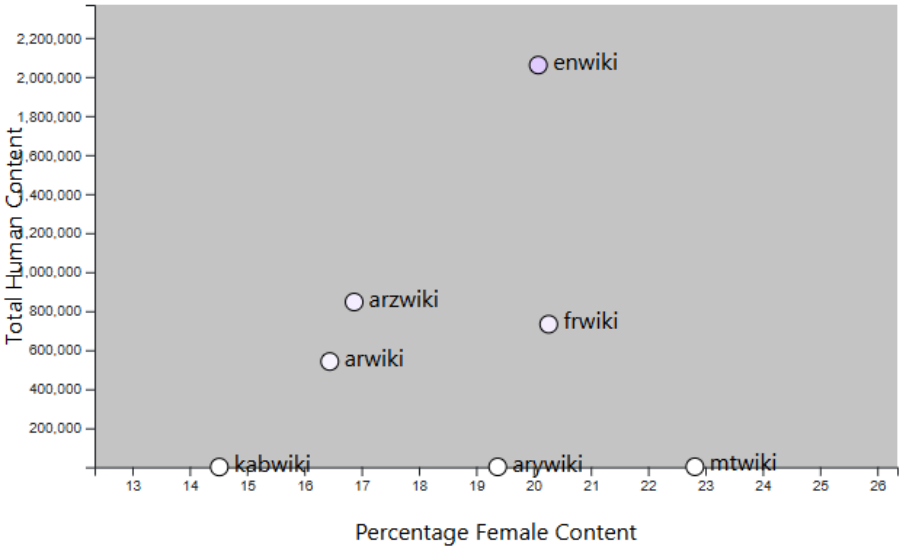
[51] Wikimedia Statistics - https://w.wiki/CM6b

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