Moroccan Darija to English: NLP’s Struggles, Failures and Insights in Machine Translation

Ikram Ait Taleb Naser
i.aittalebnaser@student.vu.nl

1 Introduction

This work marks an attempt to tackle the gaps in NLP tasks to perform Machine Translation for Moroccan Arabic to English. Moroccan Arabic, also known as Darija, despite being spoken by 40 million people remains low-resource as MSA is used in official domains in Morocco, while Darija, a blend of MSA, Amazigh, French, and Spanish, is widely spoken in daily life. Substantial advancements have been made in Moroccan Arabic NLP, such as POS tagging, dependency parsing, and code-switched text analysis, however, most tools and frameworks—like SAFAR, CAMeL Tools, and Farasa—primarily support Darija in its Arabic script. This limitation poses significant challenges, as the internet increasingly sees non-Roman script languages, like Darija, being romanized into Arabizi. This trend makes standard NLP tools ineffective to process non-standardized orthography in Darija.

The main task of this work focuses on exploring and evaluating large language models (LLMs) using various prompting methods for Moroccan Darija-English machine translation (MT). The evaluations aim at examining sentence level translation, idiomatic expressions, POS tagging and dependency parsing implementation on code-switched text. Through a combination of automatic and human evaluations, a comprehensive overview of the translation task was achieved.

2 Related Work

I begin by reviewing NLP resources developed for Arabic and its dialects. **Arabic and Darija NLP Frameworks: *Farasa*** (Institute, 2023) a package for Arabic Language Processing, performs segmentation, lemmatization, POS tagging, NER, Diactritization, constituency parsing and dependency parsing. ***CAMeL Tools*** (Obeid et al., 2020) a suite of Arabic and its dialects natural language processing tools. It is most robust in its processing of MSA and includes support for dialect identification and analysis of several major varieties, but it does not cover Darija extensively. ***SAFAR*** (SAFAR, 2023) a platform dedicated to ANLP (Arabic Natural Language Processing). It includes basic levels modules of language, especially those of the Arabic language, namely morphology, syntax and semantics. It supports Arabic dialects but not in Arabizi script.

Arabic and Darija NLP Models: *DarijaBERT* (Gaanoun et al., 2023), is currently the only “LLM” dedicated to the Moroccan Arabic dialect. The model was trained on 100M tokens. However, DarijaBERT is encoder-only, and no decoder-only models have been developed for Darija. ***DarijaBERT Arabizi***, (Gaanoun et al., 2023) trained on a dataset issued from Youtube comments, this model is the Arabizi specific version of DarijaBERT and it was trained on a total of $\tilde{4}$.6 Million sequences of Darija dialect written in Latin letters.

3 Linguistic Analysis

3.1 challenging linguistic phenomena

Within the Human Language Technology (HLT) research community literature, low density refers to languages “for which few online resources exist” (Megerdoomian and Parvaz, 2008) or “for which few computational data resources exist”. The terms underresourced or low resource seem to have similar semantics. However, as Hammarström (Hammarström, 2009) explains, it is unclear whether this is measured in absolute terms or relative to some other language. Hammarström also introduces the alternative low-affluence defined via the metric of Gross Language Product (GLP) which is the product of the number of native speakers of the language in any country and the country’s per capita Gross National Product.

For the remainder of this paper I will use the term low resource languages to refer to those that have fewer technologies, especially data sets relative to some measure of their international importance.

3.1.1 Phonetic Substitution Using Numbers

Arabizi, also known as Latinized Arabic or Arabic chat Alphabet, is a form of writing Arabic that uses Latin letters and Arabic numbers. Emerged in the early ‘90s as a mean of communication among young people who were using new technologies, it is still widely used today. It uses an encoding system that leverages a combination of the Latin script and Arabic numbers instead of Arabic letters as a solution due to the similarity in pronunciation between some Arabic and Latin characters. Each English letter represents an Arabic phoneme that matches it. For instance, Darija speakers use "3" for ع , “9” for ق , and "5" for خ. This creates ambiguity not only in the computational linguistic context, but also for non-Arabic speakers, as the numbers can be both interpreted as phonetic markers or literal numbers, depending on the context.

(1) ntlakaw m3a 3
PREP meet-1PL at 3
‘Let’s meet at 3’

(2) 3tini 5 dkai9
PREP give-2SG-IMP 5 minutes
‘Give me 5 minutes’

3.1.2 Non-Standardized Spelling

Arabic is a language that exhibits diglossia, where MSA is being used in official settings such as media communication, newspapers, and education, while the dialects dominate daily conversations and informal settings, especially on social media. The written form of the Moroccan dialect is, in fact, predominantly found on the internet and is expressed in either Arabic script or a Latinized version called Arabizi. For this reason, Darija as an informal and colloquial spoken dialect lacks standardized spelling, resulting in multiple ways of spelling the same words. Here are some examples:

a. *bghit* / *bghit* / *b3it* / *bghyt*
want.1SG
‘I want’

b. *ma3reftch* / *ma 3rftch* / *ma3raftch* / *ma3refch*
neg.know.1SG
‘I don’t know’

c. *kaydayr* / *kidair* / *kidayer* / *kidayr*
how-do.SG.M
‘How are you?’

d. *hbes* / *hbess* / *habess* / *7bes*
stop
‘Stop’

Since Darija has no standardized orthography, speakers are accustomed to seeing words represented in various forms, especially in Romanized scripts. Multiple spellings often produce the same or similar sounds, allowing readers to interpret the intended word through phonetic familiarity rather than visual uniformity.

3.1.3 Idiomatic Expressions

As a low resource language, special attention should be given to idioms when creating a lexicon for Moroccan Arabic, whose meaning is non-compositional, as the meaning of the expression is different from the meaning of its individual components. Idioms are often opaque and culturally specific. Moreover, the wording of these idioms generally does not allow for vocabulary substitutions or omission. Baker’s four strategies have been employed in several idiom translation studies (Laoudi et al., 2018):

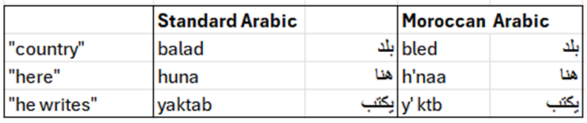
- Translation to an idiom of similar meaning and form, either because idioms are compositional and transparent in both languages; or because the idioms are both non-compositional and opaque yet conventionalized in both languages.

- Translation to an idiom of similar meaning but dissimilar form; when idioms in the starting language and target language possess the same meaning but utilize different lexical items to convey that meaning.

- Translation by paraphrase, when a match between the languages does not exist. Often used when idioms are opaque and non-compositional semantically, this is the case for “pure idioms”

- Translation omitted, when there are no equivalent expressions between the two languages.

Translation by paraphrase is the most common strategy used, also for Moroccan Darija. However, this approach often results in the loss of subtle and original nuances of the idiomatic expressions. And, in the specific case of Moroccan Darija, the optimal translation would require a deep understanding of cultural references. The lack of diverse idiomatic examples means that models are trained on a limited set of data, which can restrict



	Standard Arabic		Moroccan Arabic	
"country"	balad	بلد	bled	بلد
"here"	huna	هنا	h'naa	هنا
"he writes"	yaktab	يكتب	y' ktb	يكتب

Figure 1: Key differences in phonology between MSA and Moroccan Darija

their ability to recognize and efficiently translate idiomatic expressions.

3.1.4 Root-and-pattern morphology

difference with Arabic What makes Darija less intelligible for speakers of other dialects is its phonology, specifically its stress and the reduction of vowels. As shown in Figure 1, in Arabic words the stress is on the first syllable, while in Darija the stress moves to the final vowel and an earlier vowel is dropped.

The reflexive form in Darija:

(3) Ahmed shaf
Ahmed saw-PST-3SG
rasu flmraya
head/himself-REFL-POSS-3SG in the mirror
'Ahmed saw himself in the mirror'

'Rasu' means literally 'his head', it's originated from Amazigh, since in SA it's *'nafs'* not *ras.* We can observe the Amazigh influence also in the grammatical construction of the passive form, which is formed by adding /t/ before the verb, while in SA this happens by changing the root's vowels of the verb.

(4) Standard Arabic:

a. Kataba al-kitab
Wrote-3SG-MASC the book
'He wrote the book.'

b. K**u**t**i**ba al-kitab
Written-3SG-MASC-PASS the book
'The book was written.'

(5) Darija:

a. Ktb l ktab
Wrote-3SG-MASC the book
'He wrote the book.'

b. L ktab **t**ktb
The book written-3SG-MASC-PASS
'The book was written.'

Root morphology Darija, like other Arabic dialects, employs a non-concatenative morphology. In these languages, words are formed based on a consonantal unit called the root. On its own, the root is unpronounceable. Only when combined with a pattern of vocal elements does it become pronounceable. For example, the root *"k-t-b"* in Darija can yield words such as *kteb (he wrote), maktub (written), ktaba (writing), mektaba (library), ktuba (books), katib (writer), …* Understanding the root- morphology can help forming causative verbs as they are derived directly from their corresponding roots in the lexicon, instead of other surface forms. Causatives in MA are formed by doubling (geminating) the second segment of the base form.

(6) Root and Causative Forms:

a. brd 'cold'
b**rr**d
'make (something) cold'

b. n3s 'sleep'
ne**33**as
'make (someone) sleep'

c. sbr 'be patient'
se**bb**ar
'cause to be patient'

d. frh 'be happy'
fe**rr**han
'make happy'

e. dab 'melt'
de**ww**ab
'melt (something)'

f. faq 'wake up'
fe**jj**aq
'wake (someone) up'

g. qra 'read'
qa**rr**i
'make (someone) read / teach'

French influence As for French, Darija constructs the negative form in a double prepositions format "*ma … sh*" and does not follow the Arabic form in which it is used the negative preposition *"laysa"* (not).

(7) Ma3andish ta shi
Not I have-NEG-POSS-1SG
matesha, walakin 3andi bezzaf
any-NEG-INDEF tomatoes but I
dial formage
have-POSS.1SG a lot of cheese

‘I don’t have any tomatoes, but I have a lot of cheese’

The passive participle in Darija is obtained through the prefixation of the morpheme m- to a verb root. For example: *m-kerfes* ‘in a deteriorated state’, *m-barqe3* ‘stained’ and *m-xerbeq* ‘messy’. This is also applied to the verbs adapted from the French language:

(8) Passive Participles:

a. bgs beau gosse ‘handsome’
m-buges
‘handsome’

b. bdr poutre ‘physically strong person’
m-buder
‘physically strong person’

(9) Causatives:

a. mzk musique ‘music’
m**zz**ek
‘to make music / play music’

3.1.5 Code-Switching with MSA, French, and English

Morocco’s ethnical and linguistic richness had a huge impact in the rise of the code-switching phenomenon, where the multilingual speakers alternate between multiple languages, often within single sentences. Darija is the language of popular culture; Amazigh, the language of the native people of Morocco; Modern Standard Arabic serves as the formal written language used in official contexts and French is prevalent mainly in higher education and among certain social classes. Throughout the years, these languages started to being used across multiple contexts resulting in an overlap of languages. Another interesting phenomenon worth mentioning is that a recognisable amount of French verb stems is adapted into the colloquial Moroccan Arabic verb morphology patterns, for example:

(10) Borrowed Verbs:

a. yivalidaw *(valider)*
validate-3PL
‘to validate’

b. ghatchanger *(changer)*
FUT-change
‘to change’

c. kaymetriziw *(maîtriser)*
PROG-master-2PL
‘to master’

From a young age, Moroccan speakers develop the ability to navigate code-switching based on domain and contextual norms. For instance, a speaker might switch to Berber for an expression tied to local culture, to MSA for religious references such as “*InchAllah*”, “*Alhamdullilah*” or to reference verses of the Quran. French is often used to convey more technical and formal terms particularly in professional or academic contexts especially since Darija lacks specific vocabulary. Among younger generations, code-switching to English has also become common, driven by social media influence and the integration of trending terms into everyday conversation.

3.2 analysis of findings

N-gram analysis. By using n-grams, we can detect patterns of spelling variations at the character level, helping our model identify equivalent words across different spellings. In the context of machine translation, identifying these spelling variations is important to ensure the model does not treat them as completely distinct words, which could degrade translation quality. To investigate this property, I first tokenize the words into character-level n-grams, such as trigrams. For each word in the dataset, extract all possible trigrams and calculate their frequency across the entire corpus. By analyzing the most common trigrams, I can detect recurring patterns across different spellings as the goal is to connect the most frequent trigrams that appear across multiple spelling variations around the same words. Figure 2 illustrates the co-occurrence of character-level n-grams for 4 different spelling variations of the same word. Highlighting the common trigrams, this analysis serves as a foundation for understanding the frequency and distribution of spelling variations, guiding a more efficient data augmentation strategy, as detailed in the GitHub repository.

Type-Token Ratio analysis. Type-Token Ratio (TTR) analysis provide insights into the root pattern morphology of Moroccan Darija. These metrics will help examining the reason of length disparities and vocabular diversity. More specifically, Darija tends to produce shorter sentences compared to English, largely due to its use of root-

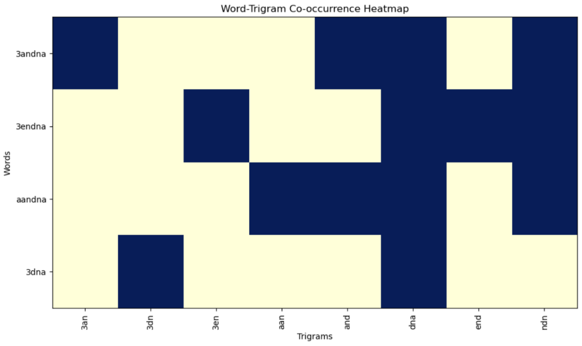


Figure 2: N-gram analysis on 4 variations of the same word

pattern morphology and derivational processes, which results in the expression of meaning in fewer words. TTR analysis further highlights Darija's lexical diversity as the morphological system generates varied word forms. Investigating these properties will inform targeted data augmentation strategies, such as balancing sentence lengths, generating morphologically diverse tokens, and creating examples with varied switching patterns. To analyze average sentence length, the sentences are tokenized for both the Darija-to-English parallel dataset and the code-switched song lyrics dataset. The absolute differences in sentence length are calculated for each parallel pair, enabling a comparison between Darija and English sentence structures. TTR is then computed for each dataset by dividing the number of unique words by the total word count. Finally, as shown in Table 1 the TTR values are compared between the Darija-English dataset and the code-switched dataset to highlight variations in lexical diversity and morphological patterns. Darija sentences tend to be shorter due to the compact morphological features characteristic of Arabic and its dialects, where meaning is expressed through affixes rather than auxiliary words. The TTR analysis confirm this hypothesis as Darija has a higher TTR due to its richer morphology and syntax, whereas English has a lower TTR, reflecting repetitive auxiliary constructions.

Language	**TTR**
Darija	0.3748
English	0.1743

Table 1: TTR Values for Darija and English

4 Prompting with LLMs

State-of-the-art LLMs such as LLaMA (Touvron et al., 2023) mainly come from public repositories, web-crawled text, academic papers, and other multilingual corpora. While Arabic as a language is likely included, Darija's presence is sparse or incidental. Moreover, open-source LLMs tailored for the Arabic language such as Jais (Sengupta et al., 2023), that includes 20 models ranging from 590 million to 70 billion parameters and trained on up to 1.6 trillion tokens of Arabic, English, and code data and AceGPT (Huang et al., 2023) are also unsuitable for Darija. This is primarily because such models do not support the dialectal varieties written in Arabizi. Therefore, mainstream models such as ChatGPT and Claude were chosen for this study for a comparative analysis of their translation performance.

4.1 methods

Few-shot technique. Few-shot prompting refers to the practice of providing a pre-trained model with a small set of examples of a task (often fewer than 10) in order to guide the model's response to similar inputs. The model is expected to infer the pattern or structure of the task from these few instances and apply it to new, unseen examples (Brown et al., 2020)

Chain of thoughts technique. Chain-of-thought prompting has several attractive properties as an approach for facilitating reasoning in language models. First, it allows models to decompose multi-step problems into intermediate steps, which means that additional computation can be allocated to problems that require more reasoning steps. Second, a chain of thought provides an interpretable window into the behavior of the model, suggesting how it might have arrived at a particular answer and providing opportunities to debug where the reasoning path went wrong. chain-of-thought reasoning can be used for tasks such as math word problems, commonsense reasoning, and symbolic manipulation. Finally, chain-of-thought reasoning can be readily elicited in sufficiently large off-the-shelf language models simply by including examples of chain of thought sequences into the exemplars of few-shot prompting. (Wei et al., 2022)

4.1.1 Experiments

For translating idioms, I employed the few-shot prompting technique (Figure 6), which leverages in-context learning by including examples directly within the prompt. This approach provides the model with demonstrations to guide and align its performance with the task's requirements. Since idiom translation requires precise but straightforward reasoning, 2-shot prompting proved to be already very effective.

In contrast, for dependency parsing in code-switched text, I combined 1-shot prompting (Figure 7) with chain-of-thought techniques (Figure 8) to address the task's complexity on the code switched lyrics text. Chain-of-thought prompting leverages intermediate reasoning steps, allowing the model to generate structured outputs. The demonstrations provided in the prompt as the "step by step thinking " serve as conditioning for subsequent examples, guiding the model to produce consistent responses. However, reasoning chains generated using chain-of-thought prompting can sometimes contain errors. To mitigate the impact of such mistakes, I ensured the demonstrations were supported by detailed explanations.

5 Machine Translation Task

5.1 Results

BLEU (Bilingual Evaluation Understudy) is an automatic evaluation metric for machine translation that measures the similarity between a machine-generated translation and one or more human reference translations. It calculates the overlap of n-grams between the machine translation and references, penalizing outputs that are too short with a brevity penalty. BLEU scores range from 0 to 1 (often scaled to 0–100), where higher values indicate higher similarity to the reference translations. (Papineni et al., 2002)

The results demonstrate that both models achieved similar levels of alignment with the reference translations as shown in Table 2. Both GPT-4 and Claude achieved similar BLEU scores—**0.3678** and **0.3732**, respectively when evaluated on the translation of the 1000 sentence pairs. It is important to note that the accuracy of BLEU scores could improve if multiple reference translations were provided for each Darija sentence as this would account for the variability in human translations, as there are often multiple valid ways to express the same idea in English.

Model	BLEU Score
GPT-4	0.3678
Claude	0.3732

Table 2: BLEU Scores for GPT-4 and Claude

In addition to BLEU scores, the models' ability to translate 63 idiomatic expressions was evaluated through **mean accuracy** scores, as reported in Table 3. The results demonstrate that GPT-4 outperformed Claude in this aspect, achieving a mean accuracy score of **2.25**, compared to Claude's **1.76**. Despite this difference, both models struggled to capture the full nuances of idiomatic and sarcastic expressions, reflecting a limitation in their understanding of context-specific meanings within the Darija dialect.

Model	Mean Accuracy
GPT-4	2.25
Claude	1.76

Table 3: Mean Accuracy for Idiom Translation

The translation results for the idioms are summarized in the idioms-evalaution file, which compares the translations from both models alongside their respective evaluations. For GPT-4, **25.39%** of the translations received the highest score, while **38.09%** received the lowest. In contrast, for Claude, **23.81%** of the translations received the highest score, while **58.73%** received the lowest.

6 Discussion

One interesting phenomenon I noticed is that sarcasm detection is an area of improvement for my models. In the idiom "*Lah i3Tik matakol*" the literal translation is "May God give you something to eat" in the sense of ”May God provide for you”. In Moroccan culture, this phrase is used sarcastically, typically to wish bad luck or curse someone (e.g., in the context of someone who's been troublesome). Both of my models provided a correct literal translation but missed the sarcastic nature of the phrase. Therefore I refined my evaluations introducing more questions for the evaluation:

Sarcastic Intent: Did the translation preserve the sarcastic tone of the original idiom?

Naturalness: Does the translation sound like a natural sarcastic phrase in the target language?

Fluency: Does the translated phrase flow naturally in the target language?

Cultural Relevance: Does the translation reflect the Moroccan cultural context and the use of sarcasm in this specific situation?

7 Conclusion

To deepen my understanding of the grammatical structures of Moroccan Darija—elements that, as a native speaker, I may have overlooked—I took the initiative to attend lectures focused on Moroccan Darija grammar. These lectures proved invaluable, as they highlighted the challenges that non-native learners encounter. I gained a more comprehensive understanding of the language's rules but also identified specific areas where learners typically face difficulties. This experience inspired me to creatively incorporate these insights into my project, ensuring that my work is not only linguistically accurate but also sensitive to the challenges faced by learners of Moroccan Darija.

In this work I presented these unique challenges and linguistic phenomena inherent to Moroccan Arabic (Darija) and its implications for NLP tasks, particularly in the context of machine translation and code-switched data. By examining large language models through various prompting techniques, I have demonstrated the difficulties posed by Darija's non-standardized orthography, morphological complexity, and heavy influence from languages like Amazigh, French, and Modern Standard Arabic. Additionally, phenomena such as phonetic substitutions using numbers, code-switching, and idiomatic expressions have been shown to further exacerbate translation and analysis complexities for current models.

Through evaluations that combine automatic and human assessments it was possible to indicate that both GPT-4o and Claude are aligned with reference translations, but they fail to capture nuances such as idiomatic and sarcastic intent. Furthermore, this study underscores the need for tailored tools and datasets that can address Darija's use in informal settings, including Arabizi. Future work must prioritize the development of dedicated frameworks, culturally-aware idiomatic translations, and enhanced multilingual models that include low-resource varieties like Darija.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei, and OpenAI. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2023. Darijabert: a step forward in nlp for the written moroccan dialect.

Harald Hammarström. 2009. A survey of computational morphological resources for low-density languages. *Journal of the NEALT*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2023. Acegpt, localizing large language models in arabic.

Qatar Computing Research Institute. 2023. Farasa - a comprehensive arabic nlp toolkit. Accessed: 2023-12-15.

Jamal Laoudi, Claire Bonial, Lucia Donatelli, Stephen Tratz, and Clare Voss. 2018. Towards a computational lexicon for Moroccan Darija: Words, idioms, and constructions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 74–85, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Karine Megerdoomian and Dan Parvaz. 2008. Low-density language bootstrapping: the case of tajiki Persian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Equipe SAFAR. 2023. Safar - a toolkit for morphological analysis of arabic. Accessed: 2023-12-15.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

8 Appendix

8.1 Data statement

8.1.1 Curation Rationale

Darija Open Dataset (DODa) is an open-source project for the Moroccan dialect, and the largest open-source collaborative project for Darija-English translation. DODa contains more than 10 thousand entries covering verbs, nouns, adjectives, verb-to-noun, singular to-plural correspondences, conjugations of hundreds of verbs in different tenses. The Dataset is also divided into specialized subcategories such as food, animals, human body, health, education, family, environment, economy and many others. Besides semantic categorization, it presents words under different spellings, offers verb-to-noun and masculine-to-feminine correspondences, extra categories for idioms, proverbs and short expressions, as well as more that 86,000 translated sentences. My curation prioritized linguistic phenomena relevant to my project goals namely investigating spelling variations, and syntactic challenges in code-switching. Specifically, the dataset in my GitHub Repository includes 65 idioms and proverbs, 241 words under different spellings and 46k translated sentences.

To investigate code-switching phenomena, I constructed a custom corpus using online sources. To extract song lyrics, I scraped the website genius.com through the Genius API, used multiple queries to extract song lyrics by Moroccan artists who are known to include many instances of Code Switching in their words. Moroccan Darija was the main (matrix) language while other languages such as French and English were present as guest (embedded) languages.

8.1.2 Language Variety

The language of my dataset is Moroccan Darija written in Latin script: ary-Latn-MA. For the sake of simplicity and consistency, the dataset focuses on the standard Darija spoken and understood by the majority of Moroccans. And to align with the real-world usage and informal communication, only the Darija written in Latin script will be taken into account.

8.1.3 Speaker demographic

NA

8.1.4 Annotator demographic

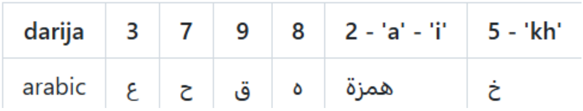
NA

8.1.5 Speech situation

NA

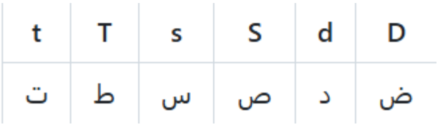
8.1.6 Text characteristics

The Arabic script column from the original dataset was excluded to focus solely on the Latin-script representation of Darija. Additionally, idioms and proverbs were consolidated into a single dataset. To maintain phonetic clarity and consistency, the dataset adheres to the correspondences and norms detailed in the following Tables 3, 4, 5:



darija	3	7	9	8	2 - 'a' - 'i'	5 - 'kh'
arabic	ع	ح	ق	ه	همزة	خ

Figure 3: Encoding numbers representations as letters

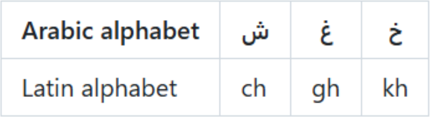


t	T	s	S	d	D
ت	ط	س	ص	د	ض

Figure 4: Correspondences latin and arabic alphabet

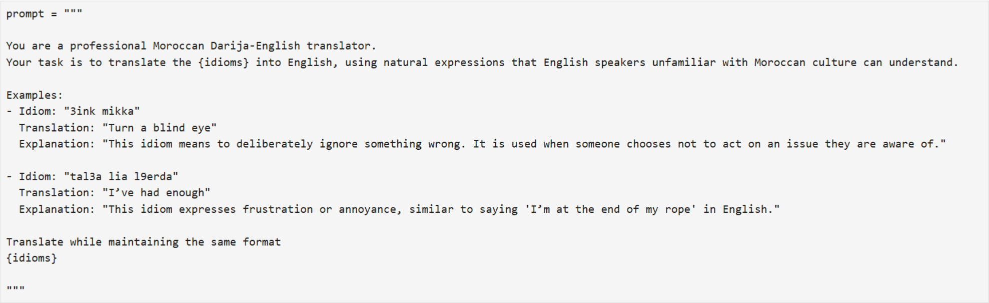
Furthermore, other principles are included:

- doubling characters for phonetic emphasis (7mam – pigeons) (7mmam – bathroom)
- avoid appending the French “e" at the end of words, such as "louz" instead of "louze."
- Letters like "Z" or "th" are not used for Arabic phonemes ظ ، ذ ،ث as these sounds are rare.



Arabic alphabet	ش	غ	خ
Latin alphabet	ch	gh	kh

Figure 5: Phonetic correspondences latin and arabic alphabet



```
prompt = """

You are a professional Moroccan Darija-English translator.
Your task is to translate the {idioms} into English, using natural expressions that English speakers unfamiliar with Moroccan culture can understand.

Examples:
- Idiom: "3ink mikka"
  Translation: "Turn a blind eye"
  Explanation: "This idiom means to deliberately ignore something wrong. It is used when someone chooses not to act on an issue they are aware of."

- Idiom: "tal3a lia l9erda"
  Translation: "I've had enough"
  Explanation: "This idiom expresses frustration or annoyance, similar to saying 'I'm at the end of my rope' in English."

Translate while maintaining the same format
{idioms}

"""
```

Figure 6: prompt for idioms

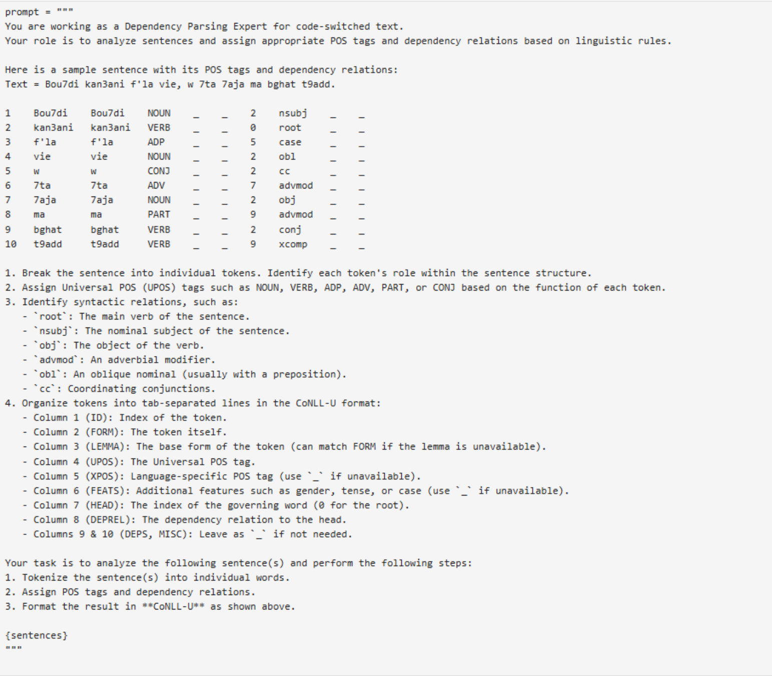
8.1.7 Recording quality

NA

8.1.8 Other

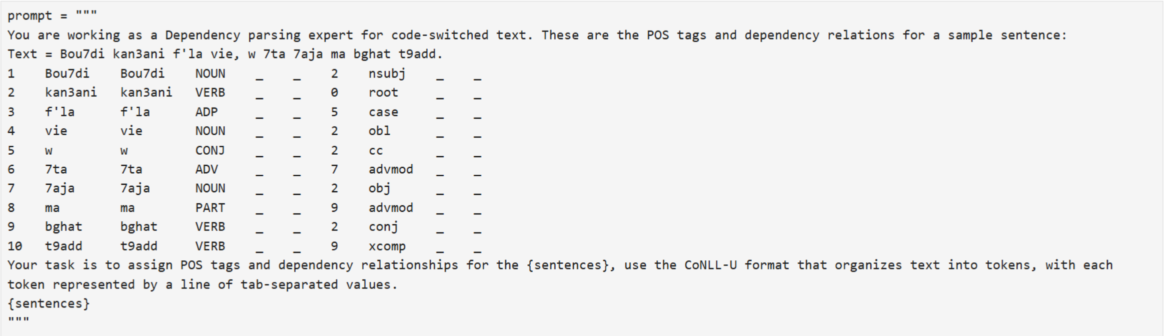
NA

9 prompts



```
prompt = """
You are working as a Dependency Parsing Expert for code-switched text.
Your role is to analyze sentences and assign appropriate POS tags and dependency relations based on linguistic rules.

Here is a sample sentence with its POS tags and dependency relations:
Text = Bou7di kan3ani f'la vie, w 7ta 7aja ma bghat t9add.

1    Bou7di    Bou7di    NOUN    _    _    2    nsubj    _    _
2    kan3ani   kan3ani   VERB    _    _    0    root     _    _
3    f'la      f'la      ADP     _    _    5    case     _    _
4    vie       vie       NOUN    _    _    2    obl      _    _
5    w         w         CONJ    _    _    2    cc       _    _
6    7ta       7ta       ADV     _    _    7    advmod   _    _
7    7aja      7aja      NOUN    _    _    2    obj      _    _
8    ma        ma        PART    _    _    9    advmod   _    _
9    bghat     bghat     VERB    _    _    2    conj     _    _
10   t9add     t9add     VERB    _    _    9    xcomp    _    _

1. Break the sentence into individual tokens. Identify each token's role within the sentence structure.
2. Assign Universal POS (UPOS) tags such as NOUN, VERB, ADP, ADV, PART, or CONJ based on the function of each token.
3. Identify syntactic relations, such as:
   - `root`: The main verb of the sentence.
   - `nsubj`: The nominal subject of the sentence.
   - `obj`: The object of the verb.
   - `advmod`: An adverbial modifier.
   - `obl`: An oblique nominal (usually with a preposition).
   - `cc`: Coordinating conjunctions.
4. Organize tokens into tab-separated lines in the CoNLL-U format:
   - Column 1 (ID): Index of the token.
   - Column 2 (FORM): The token itself.
   - Column 3 (LEMMA): The base form of the token (can match FORM if the lemma is unavailable).
   - Column 4 (UPOS): The Universal POS tag.
   - Column 5 (XPOS): Language-specific POS tag (use `_` if unavailable).
   - Column 6 (FEATS): Additional features such as gender, tense, or case (use `_` if unavailable).
   - Column 7 (HEAD): The index of the governing word (0 for the root).
   - Column 8 (DEPREL): The dependency relation to the head.
   - Columns 9 & 10 (DEPS, MISC): Leave as `_` if not needed.

Your task is to analyze the following sentence(s) and perform the following steps:
1. Tokenize the sentence(s) into individual words.
2. Assign POS tags and dependency relations.
3. Format the result in **CoNLL-U** as shown above.

{sentences}
"""
```

Figure 8: Chain of thought prompting for parsing



```
prompt = """
You are working as a Dependency parsing expert for code-switched text. These are the POS tags and dependency relations for a sample sentence:
Text = Bou7di kan3ani f'la vie, w 7ta 7aja ma bghat t9add.
1    Bou7di    Bou7di    NOUN    _    _    2    nsubj    _    _
2    kan3ani   kan3ani   VERB    _    _    0    root     _    _
3    f'la      f'la      ADP     _    _    5    case     _    _
4    vie       vie       NOUN    _    _    2    obl      _    _
5    w         w         CONJ    _    _    2    cc       _    _
6    7ta       7ta       ADV     _    _    7    advmod   _    _
7    7aja      7aja      NOUN    _    _    2    obj      _    _
8    ma        ma        PART    _    _    9    advmod   _    _
9    bghat     bghat     VERB    _    _    2    conj     _    _
10   t9add     t9add     VERB    _    _    9    xcomp    _    _
Your task is to assign POS tags and dependency relationships for the {sentences}, use the CoNLL-U format that organizes text into tokens, with each
token represented by a line of tab-separated values.
{sentences}
"""
```

Figure 7: 1-shot prompting for parsing