NLP: Franco-Arabic Corpus and Challenges

Written by Zina Abohaia ([20194621@ug.buid.ac.ae](mailto:20194621@ug.buid.ac.ae)) and Yousef Mamdouh (20196448@ug.buid.ac.ae)

**Abstract** Romanized Arabic, or Arabic written in the Roman script, is commonly known as Franco-Arabic or Arabizi. Studies have shown that the use of Franco is on the rise, particularly in its written form, on social media. In this paper, we cover some of the issues in processing Franco, one of them being that there are no set rules, and the spelling of an Arabic word differs from dialect to dialect, and context to context. Therefore, we sought to build a Franco-Arabic corpus and use machine translation to see if it would be successful in translating Franco to Arabic. We used the 3,500 most used Arabic words online, as a starting point for the corpus, evaluated it manually, and achieved an accuracy of 92.4%.

**Keywords**: NLP, Romanized Arabic, Franco, Franco-Arabic, Arabizi

## Introduction

Franco-Arabic was once a solution to technology only supporting the roman script,and now it has evolved to being a preferred mode of communication, especially by the arab youth. Social Media is filled with Franco text, and it is usually excluded when researchers parse tweets or facebook posts. This is problematic because not only is it widely used, but it is growing, and a way to structure it is required.

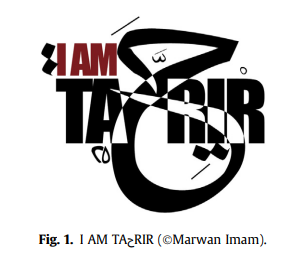
Building a Franco corpus is one way to solve the problem, and from the papers we have found and covered in the literature review, we see that it has been attempted to set rules for franco, and annotate for sentiment analysis but translation has not been covered.

In this paper, we build a Franco corpus, and test various machine learning and translation techniques to see if this is a viable solution. The challenges in building a Franco corpus, and dealing with Franco, generally, are also covered.

## Literature Review

Panović (2018) discusses Arabic-English script-fusing, and compares it to Franco, romanized arabic or writing arabic in the roman script. According to Panović (2018), Script-fusing is the “insertion of at least one Arabic letter into an Arabic word or name written in Roman script” as seen in Fig 1, which sprang from a movement of activism during the Egyptian uprising in 2011, as well as creativity. Franco's roots hail as a solution to technological restraints, when only English script was supported on phones. So users wrote Arabic in English script, and it also allowed them to send the same message with less space, saving them the cost of that space. However, even after Arabic script was adapted, despite the negative public opinion, franco grew in popularity, especially with younger generations.

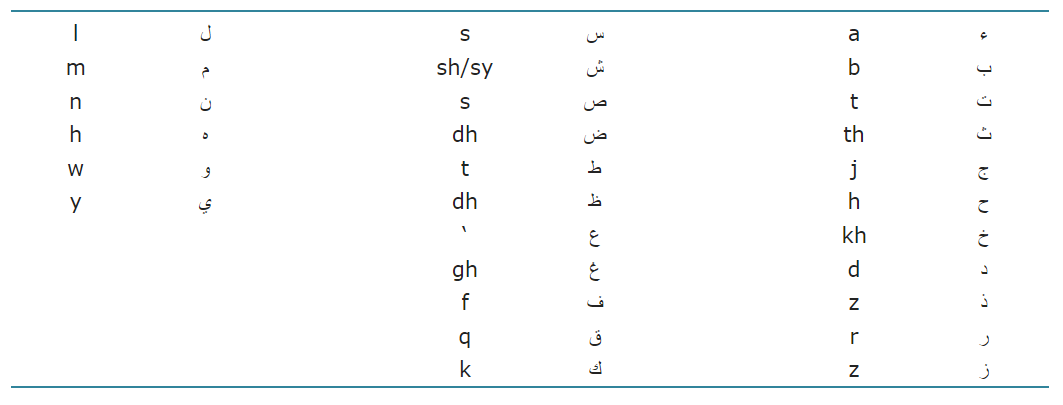
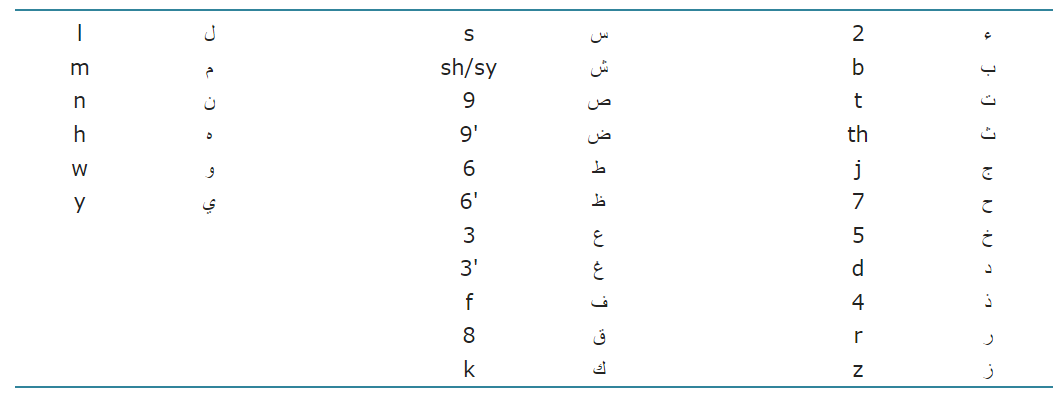
Furthermore, franco, or “arabizi”, the other term used to describe franco, grew to also encompass Arabic-French in English Script, particularly prevalent in Morocco, Tunisia, and Algeria. The researcher sees that the expansion of Franco is unavoidable, and that it is commonly used in social media.



**Fig 1**. I am Taحrir (copyrighted Marwan Imam) Panović (2018)

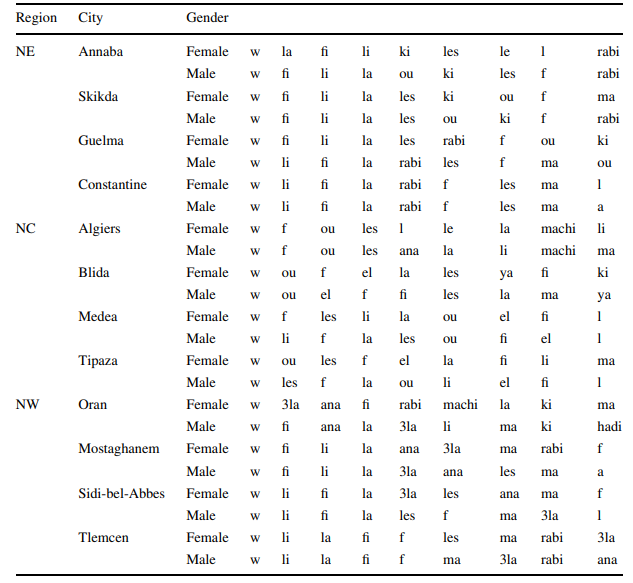
Akbar, Taqi and Sadiq (2020) investigated the prevalence of arabizi in kuwaiti youth, through a digital questionnaire and a 12h whatsapp discussion group. They found that arabizi is seen to be viewed in two main, contradicting views, by the kuwaiti youth. The group from English private schools see it to be a symbol of “friendliness and modesty”, and see bilinguality and westernization as a positive attribute as it helps them understand others, and communicate with them effectively. While the group from the public school saw it as “being show-offs, disloyal, unfriendly, effeminate and silly”, viewing bilinguality and westernization as a negative attribute that causes the loss of national identity. Finally, Akbar, Taqi and Sadiq (2020) see that Kuwait is “is on its way of establishing a newly diagraphic situation that may lead to the standard Arabic language attrition”.

Kenali, Yusoff, Kenali and Kamarudin (2016) aimed to identify how much spoken and written arabizi is used among Arab students in Saudi Arabia, by measuring it on the Cronbach’s Alpha scale. They found that Spoken Arabizi consumption is low, while written Arabizi consumption is moderate and that Arabizi is considered to have experienced a great expansion in a small period. They also found, just as in the two previous papers, that Arabizi is growing, and they urge experts to prevent it from “being an alternative or a replacement for Arabic in the future”. They contributed to the research by writing a basic, and advanced, arabizi code, as seen in Fig 2, and Fig 3, below.



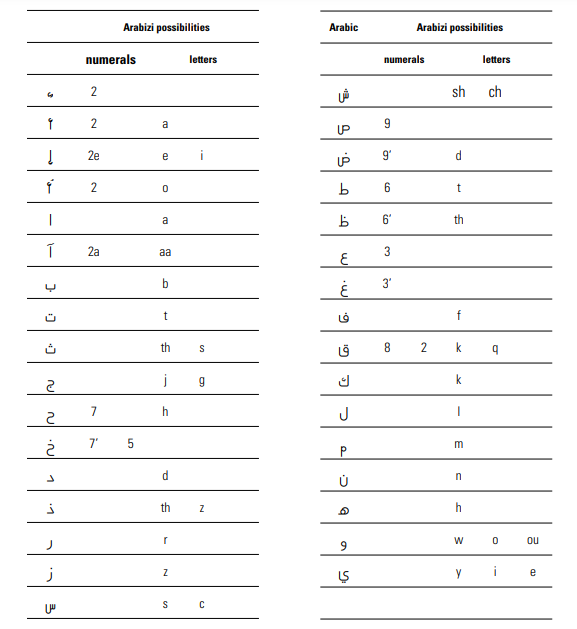
**Fig 2, 3**. Basic and Advanced Rules for Franco Kenali, Yusoff, Kenali and Kamarudin (2016)

Abainia (2019) Provides a multi-purpose parallel corpus for various natural language processing purposes: Text crawled from facebook was tokenized, organized by gender, region, and city, transliterated into arabic script, written in modern standard arabic, and annotated for emotion and abuse detection. Also, they used various machine learning techniques (langid.py, LangDetect, NB and SVM) to evaluate whether sub-algerian dialects can be identified correctly. They found that for Algerian sub-dialect identification, LangDetect did the best with an accuracy of 0.272, while for Algerian dialect identification, it also did the best with 0.589. Code-switching, the overlapping of dialects, and classification based on the text’s topic, all proved to be challenges to the correct identification of dialects. In the following figure, Abainia (2019) lists the common stop words used by each gender in each algerian city, to show how different and similar they can be.



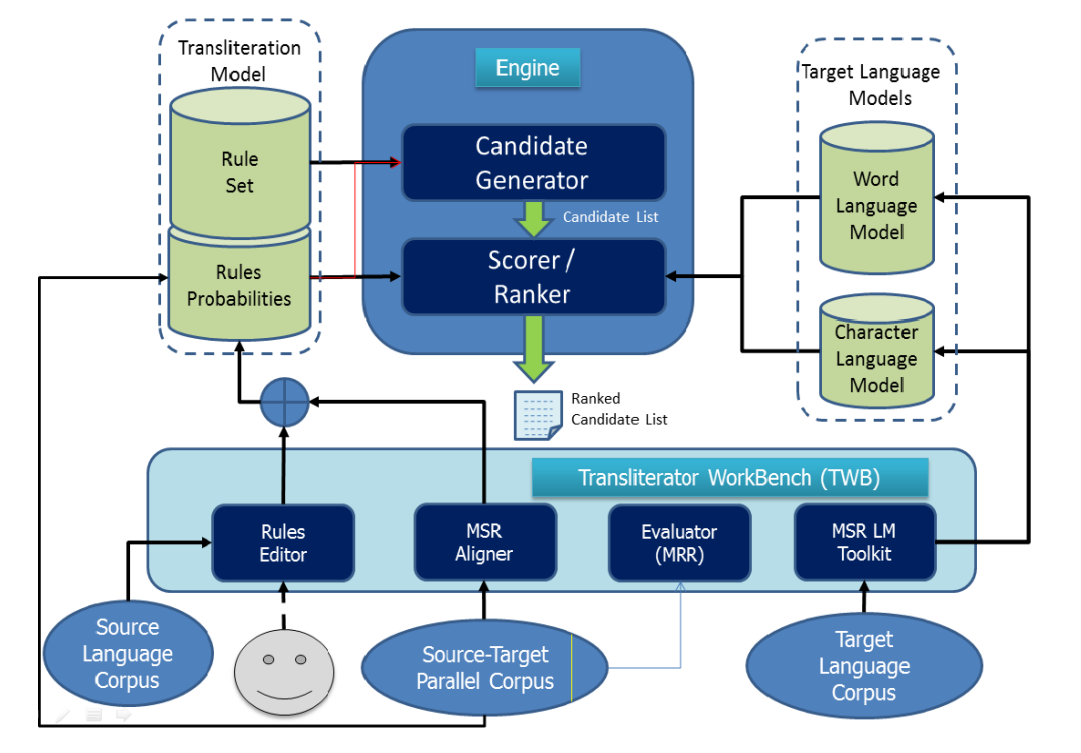
**Fig 4**. The different Franco spellings depending on gender and the sub-dialects of each region Abainia (2019)

Yaghan (2008) Developed rules through discussions with second-year graphic design students and explored a couple of Arabizi typefaces. The following “Arabizi Rules” were produced: “(1) In general, the Arabizi system is contextual. (2) The traditional Arabic vocalization marks are substituted by vowels. The “fatha” by a; the “kasra” by i or e; and the “damma” by u, ou, or o. (3) The use of the vowels is optional in Arabizi, and they could be omitted. Three factors determine this omission: the reader’s background; the contextual clarity of the word; and the allowable number of characters per message. (4) As with slang Arabic, which borrows English words and phrases (for example “please,” “OK,” “nice meeting you,” “thanks,” etc.), Arabizi uses English within the text. Common World Wide Web and cellular phone message abbreviations are used (“plz” for “please,” “thnx” for “thanks,” etc.). (5) The use of capital letters indicates yelling, excitement, emotions, or calls for special attention (as with most messaging systems). (6) There are many ways of representing the same situation and conveying the same meaning. (7) Besides the English abbreviations, there are many abbreviations regarding some word endings in Arabic. For example @ is used for the affix added for certain types of plurality in Arabic (…aat   
ط ;(and 8 as an affix to indicate the firstperson past tense of certain verbs in slang Arabic (…eet h). (8) The Arabizi system differs for every Arabic country, depending on the local dialect. (9) The Arabic language uses a special mark when stressing a consonant instead of doubling it (Figure 1). In Arabizi, it is written twice unless it was a compound letter. Then it is left to the context to be understood. (10) Some combinations of English letters are used to draw the actual shape of an Arabic word. For example, the combination oLI I for the Arabic a|ا”. They also mapped arabic letters to their english counterparts, as Kenali, Yusoff, Kenali and Kamarudin (2016) did but had a different result as seen below:



**Fig 5**. Franco Rules Yaghan (2008)

Chalabi and Gerges (2022) built a framework for a romanized Arabic transliteration engine that they later scaled to cover other scripts. Their framework had a number of models, and parts, which is included in the below figure, based on “the basic phrase-based statistical machine translation approach” (Chalabi and Gerges, 2022). They used a variety of sources: blog data, phone call transcripts, websites, and forums. Using the Topmost technique for evaluation, they achieved 85% accuracy using the top candidate only. Furthermore, they found that Framework efficiently produced models for different languages when human expertise was high, and there was a low amount of data,as well as when human expertise was low, and the data high.



**Fig 6**. Romanized Arabic Framework Chalabi and Gerges (2022)

Sabby, Sharaf and Abdennadher (2018) aimed to Build a prototype to perform sentiment-analysis on Franco-Arabic text. They tagged Text from facebook posts to five classes, and used Sickit Learn classifiers to classify the text. Using a testing sample of 50 posts, they achieved an accuracy of 85%. Some of the challenges they faced were variance in deciding the emotion implied, detection of sarcasm, and having the text be based on dialects and not modern standard arabic. Further work would be to use a bi-gram model, instead of a uni-gram model.

**Table 1:** Summary of reviewed Papers

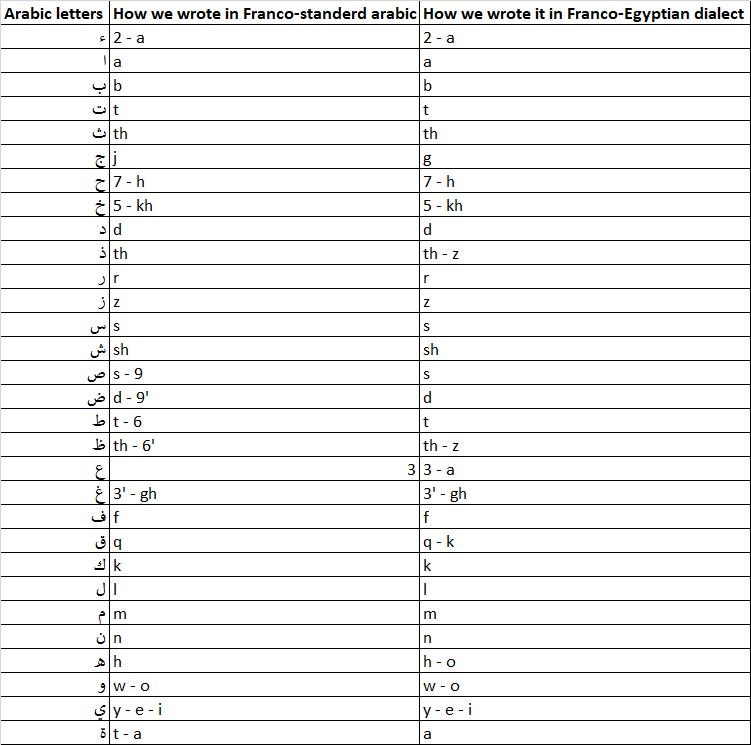
| Paper | Aim | Methodology | Findings | Limitations |
| --- | --- | --- | --- | --- |
| Panović (2018) | Offers a historicized interpretive framework to understand Arabic-English script-fusing | Compares script-fusing to Franco, and analyzes the design choices of visual script-fusing | Franco and Script-fusing popularity is on a rise | Analyses a limited number of user’s opinions on scrip-fusing |
| Akbar, Taqi and Sadiq (2020) | Investiages Arabizi in amongst Kuwaitis who have different educational backgrounds | Digital questionnaire, and Whatsapp group discussions | Two nonstandard language versions, and a development of a sociolinguistic setting where Arabizi should be put across Kuwait’s linguistic profile | From the 212 participants, only two females were included from the public school group, resulting in the results being biased to males. |
| Kenali, Yusoff, Kenali and Kamarudin (2016) | Identify how much spoken and written arabizi is used among Arab students in Saudi Arabia | Used Cronbach’s Alpha scale to measure the use of spoken and written arabizi | Spoken Arabizi consumption is low, while written Arabizi consumption is moderate. They also add that Arabizi is considered to have experienced a great expansion in a small period. | The sample of 1454 students was not varied, and the study was not applied to the general population. THey also had difficulty circulating the survey due to Saudi Arabian etiquette. |
| Abainia (2019) | Provides a multi-purpose parallel corpus for various natural language processing purposes | Text crawled from facebook was tokenized, organized by gender, region, and city, transliterated into arabic script, written in modern standard arabic, and annotated for emotion and abuse detection. Also, Uses various machine learning techniques to evaluate whether sub-algerian dialects can be identified correctly. | Using various machine learning techniques to evaluate whether sub-algerian dialects can be identified correctly. | The identification of sub-algerian dialects using the subject or topic of the sentence, as well as the overlapping between the dialects proved to be challenging for successful identification. |
| Yaghan (2008) | Aimed to write rules for arabizi, and explore how they were used in typefaces | Developed rules through discussions with second-year graphic design students | Arabizi should have a dictionary, and more of its typefaces explored, as well as explored as an official system to translating Arabic to English | Explored only a few typefaces, and did not expand on public opinion on them, mainly students' opinions. |
| Chalabi and Gerges (2022) | Build a romanized arabic transliteration engine | Combined both rule-based and SMT-like techniques, as well as developed parallel corpus | Framework efficiently produced models for different languages when human expertise was high, and there was low data, and vice versa |  |
| Sabby, Sharaf and Abdennadher (2018) | Build a prototype to perform sentiment-analysis on Franco-Arabic text | Text from facebook posts was tagged then analysed using machine learning techniques | With 50 posts, they had a success rate of 85%, with the 15% lost due to the variation of dialects in the training dataset | Mainly focuses on the egyptian dialect, and should be tested by actual users. |

## Methodology

Since franco arabic is mainly used on social media, we started by getting a dataset of the to 50,000 used arabic words on social media and online(Modern Standard Arabic, 2017), the dataset had the 50,000 words alongside their english translation. Then we started translating the words from arabic and writing them in the franco-arabic script, using both our knowledge of arabic, english and franco-arabic scripts. We translated around 3,500 words to Franco-Arabic and manually put the possible franco equivalent of the words.

Since franco-arabic is mainly used on social media it’s usually not standard arabic written in roman scripts, but rather arabic dialects written in roman scripts, we mostly focused on translating the words into standard arabic written in franco-arabic for a baseline to work on but we also translated to different dialects, leaning towards the egyptian dialect.

As we have seen in the literature review, franco-arabic does not have standard rules but rather different people use it differently, so for our corpus these are the main set of rules we used:



**Fig 7.** Our Franco Rules

## Findings: Challenges

While working on the corpus we faced multiple challenges, these challenges were mainly caused by the fact that franco-arabic is not a standard language or a standard way of writing, it was made as a tool to solve a problem, with no rules or one correct way to use the tool, and some of these challenges stimmed from some ANLP challenges that already exist since franco arabic is just a different way of writing in arabic:

* **Writing based on dialect**

As mentioned previously, franco arabic is a way of typing used mainly by young age groups and mainly on social media, and when people are typing on social media they are not typing formally, standard arabic is rarely used on social media, mainly used by formal accounts or websites, like news accounts for example, so while translating to franco-arabic the first problem we faced was that we can translate standard arabic to franco-arabic but most people type based on their dialect, and that’s why we started focusing on the dialect we're most familiar with, the Egyptian dialect, alongside standard arabic.

* **Sub-dialects**

Arabic dialects branching from standard arabic was not the end of that problem, we realised that every dialect has sub-dialects, and while the main dialect is used more than the sub-dialects especially when it comes to typing in franco-arabic, a lot of people would still type in their sub-dialect, for example in egypt the word “setting” or in standard arabic “جالس” is said differently in different parts of egypt, the main way is “aa’ed” or “ءاعد” but it can also be “gaa’ed” or “قاعد”, and the same goes for every other arabic dialect, different areas in the country that uses a certain dialect use that dialect differently.

* **Different use of vowels in romanizing**

This problem can also be partially considered as an ANLP problem, as arabic letters can be read differently, especially vowels, they can also be notated which can make the word be read completely different or even have a completely different meaning, and this is specifically a problem when it comes to typing arabic letters and notations in roman scripts. The letter “و” for example, in roman script it can be written in multiple different ways, o, w, u or even oo, and the letter “ي” can be written as y, i, e or ee, then when it comes to notations, some arabic notations can also be considered vowels, for example the arabic notation “ ُ “ known as “damma” is usually pronounced like the letter “o” but it can be written or not written depending on the person typing.

* **Multiple ways of writing in one dialect based on preference**

The last three problems could be considered ANLP problems as well, however from here on the problems are more specific to franco-arabic, since it is not a standardised way of writing, even if two people are using the same dialect and same sub-dialect they can still type the same word in different ways, as you can see in the previous table we had to include the different ways of writing the same letter that people can use, for example the name “خالد” can be written the same way it’s written in english “khaled” but the “kh” can also be replaced with the number 5, which in franco represents the arabic letter “خ” , and it all comes down to preference since there are no set rules.

* **Context-based**

Since franco-arabic is not a language but rather a way of writing a language, we face challenges like this one, in standard arabic the letter “ء” can be written over other letters to be pronounced differently, like “ئ” or “ؤ” or “إ” these are all pronounced differently and only sharing the how you start the sound of the latter which is “ a’ ”, when it comes to franco arabic the letter “ؤ” for example can be written as “2o” or “o”, the letter “إ” can be written as “2” or “e” or even “2e”, all depending on context as well as preference as mentioned previously.

* **Fluid**

The arabic language like any other language can change and evolve with time, which can lead to multiple ANLP challenges, franco-arabic, as a way of writing arabic, end up facing the same challenges, with an extra layer added to them being that it’s written in roman scripts, for example nowadays it is uncommon to find someone writing in arabic and using arabic notations, but they are still pronounced, so when it comes to translating to franco-arabic it can be difficult to decide if the notation should be written as its own letter. Different words and terms can also be added to the arabic language depending on world events, for example the term “social distancing” or in arabic “التباعد الإجتماعي” was uncommon in the arabic language before the COVID-19 pandemic, but now it is used more. Another major factor is that more people are learning and using the english language, so it is common to find people on the social media typing in a mix of arabic and english, especially people who type in franco arabic since the letters are already english letters, they find it easy to mix between franco-arabic and english words in the same sentence.

* **Differences between arabic and english**

Other than the differences already mentioned like the arabic notation for letters, there are major differences between writing in arabic and writing in english, one of them being that arabic letters are connected together to make words while english letters aren’t connected. For example the word “the” in english is a prefix in arabic and it’s the two letters “ا” and “ل” so if we take the term “ the life” in arabic it’s written as one word “الحياة” so when writing arabic in english letters some people write it as one word as well “al7yah” and some people write as two words like in english “al 7yah”, which could lead to multiple challenges.

## Evaluation

For evaluation, we decided to use manual evaluation since it would be more reliable than automatic evaluation but also because our dataset was not big enough for automatic evaluation; making a model, and using it to translate. Before this decision, we did attempt using a neural network but with our limited computational power, it took too long, and did not produce an output.

Therefore, we decided to evaluate it by the target of our study: The internet. To check our spellings and translations of words in the dataset, we searched them online to see if they have been used at least once, in the correct meaning, and context. We searched mainly on google, but also checked instagram, and Twitter. When the word was found, the number 1 was put in the row, and if it was not found we used 0, since it is common and convenient to use binary for boolean problems.

For this evaluation, we ran it on the 250 words in the aforementioned dataset, which are also the most used Arabic words on the internet. As seen in the table below, from the 250 words, only 19 we did not find online, and thus our accuracy is 92.4%. The 7.6% are lost due to a number of factors that came up in our evaluation:

* **The Franco word has a homonym**

The franco and its equivalent Arabic word did come up with our search, however, the arabic word was not the same one as the one we originally translated from Franco. As highlighted in our findings, this is also due to how limited Franco is compared to Arabic. An example of this is the Franco word “mata”, we translated from “متى”, meaning “when”, but came up as “ماتا”, meaning “he died”.

* **The Franco word has the same spelling of a word in another language**

Since Franco uses the romanized script, which is used in multiple languages, at least 100: English, French, Spanish, etc. (Latin Alphabet. 23 Magic Letters The Most Widely Used Around The World, 2022). Therefore in translating the arabic word “إننا”, meaning “we”, to Franco, it became “enana” which is also a spanish word meaning dwarf. We faced this same issue slightly differently with English, where the Franco word would actually be a name for a person, such as “ethan” which in Arabic is “إذن” meaning “then”. But it didn’t just stop there, the word “مثل”, meaning “like” or “for example” in Arabic, could be written as “methl” or “metl” in Franco. The former google search outputs the chemical compound Methyl, while the latter outputs Metal.

**Table 2:** Evaluated words analysis

| **Number of words evaluated** | **Number of words found online** | **Number of words not found** | **Accuracy** |
| --- | --- | --- | --- |
| 250 | 231 | 19 | 92.4% |

The column named “Arabic” had the word spelled in arabic, while the column named “Franco” has the arabic word translated to Franco. The column “Franco 2” has another possible spelling for the Arabic word, but like we can see not all arabic words have more than one spelling. Finally, the “Evaluation” column had only inputs of 0 and 1, as seen in the screenshots below:





**Fig 8, 9**. Screenshots from our dataset

## Discussion & Conclusion

**Discussion:**

Franco-arabic is a way of typing that is getting increasingly used, and this study is a contribution to the NLP field focusing on franco-arabic. Since franco-arabic is not standardised and has no set rules, standardising it could solve some of the challenges we faced and make it easier to do research on this subject, on the other hand it can also lead to different problems, since franco-arabic is not a set language, it depends mainly on the user of this way of typing, so standardising it could also lead different sub-writing-languages or it could have no use at all since people will continue to type using their preference.

**Future work:**

We have faced multiple challenges in this study and we found a lack of research done on this subject. We hope to expand the corpus to include more dialects to help solve some of the challenges, as well as sub dialects. We also aim to annotate the corpus so it can be used for sentiment analysis purposes, as well as finding and using a suitable machine translation method to evaluate our work more.

**Conclusion:**

In this paper we present a foundation to including franco-arabic in NLP applications, by providing a corpus based on the top 50,000 arabic words used online(Modern Standard Arabic, 2017), we were able to translate around 3500 words from arabic script to franco-arabic. We started with reviewing the literature done on this subject and we found that there is a lack of research done on this subject, then we present our methodology of translating the words and making the corpus, we then present the evaluation of our results which was done manually, then we present the challenges we faced while working on the corpus and how we dealt with these challenges.

## References

Panović, I., 2018. ‘You don't have enough letters to make this noise’: Arabic speakers' creative engagements with the Roman script. *Language Sciences*, 65, pp.70-81.

Akbar, R., Taqi, H. and Sadiq, T., 2020. Arabizi in Kuwait: An Emerging Case of digraphia. *Language &amp; Communication*, 74, pp.204-216.

Kenali, A., Yusoff, N., Kenali, H. and Kamarudin, M., 2016. Code-Mixing Consumptions among Arab Students. *Creative Education*, 07(07), pp.931-940.

Abainia, K., 2019. DZDC12: a new multipurpose parallel Algerian Arabizi–French code-switched corpus. *Language Resources and Evaluation*, 54(2), pp.419-455.

Yaghan, M., 2008. “Arabizi”: A Contemporary Style of Arabic Slang. *Design Issues*, 24(2), pp.39-52.

Chalabi, A. and Gerges, H., 2022. Romanized Arabic Transliteration. In: *COLING*. [online] Mumbai: Proceedings of the Second Workshop on Advances in Text Input Methods (WTIM 2), pp.89–96. Available at: <https://aclanthology.org/W12-4808.pdf> [Accessed 10 June 2022].

Sabby, C., Sharaf, N. and Abdennadher, S., 2018. Mood Extraction from Franco-Arabic Facebook Posts. In: *19th International Conference on Computational Linguistics and Intelligent Text Processing*. [online] Available at: <https://www.researchgate.net/publication/327282118\_Mood\_extraction\_from\_Franco-Arabic\_Facebook\_Posts> [Accessed 10 June 2022].

Modernstandardarabic.com. 2017. *Top 50000 Arabic Words | Modern Standard Arabic*. [online] Available at: <http://www.modernstandardarabic.com/top-50000-arabic-words/> [Accessed 11 June 2022].

Learning Mole. 2022. Latin Alphabet. 23 Magic Letters The Most Widely Used Around The World. [online] Available at: <https://learningmole.com/latin-alphabet/> [Accessed 18 June 2022].