The interplay between morphological typology and script on a novel multi-layer Algerian dialect corpus.

Anonymous ACL-IJCNLP submission

Abstract

Recent years have seen a rise in interest for cross-lingual transfer between languages with similar typology, and between languages of various scripts. However, the interplay between morphological typology and difference in script on cross-lingual transfer is a less studied problem. We explore this interplay on cross-lingual transfer for two supervised tasks, namely part-of-speech tagging and sentiment analysis. We introduce a newly annotated corpus of Algerian user-generated comments comprising parallel annotations of Algerian written in Latin, Arabic, and code-switched scripts, as well as annotations for sentiment and topic categories. We perform baseline experiments by fine-tuning multi-lingual language models. We further explore the effect of script vs. morphological typology in cross-lingual transfer by fine-tuning multi-lingual models on languages which are a) morphologically distinct, but use the same script, b) morphologically similar, but use a distinct script, or c) are morphologically similar and use the same script. We find there is a delicate relationship between script and typology for part-of-speech, while sentiment analysis is less sensitive.

1 Introduction

Cross-lingual transfer has shown promising results for several tasks, however the effect of and the interplay between typologically related languages and languages that do not share the same script has seen less focus. This is especially true for under-resourced vernacular languages and dialects. In this paper, we focus our work on the Algerian language, a non-standardized vernacular Arabic variety, characterized by the heavy use of both code-switching and borrowings. The existing code-switching can be anything from local Algerian dialects (e.g. region based Algerian or Berber), French, English, Spanish, Modern Standard Arabic (MSA), or other Arabic dialects. The borrowings depend on the speakers' background, but is usually heavily French-based.

Algerian is a spoken language with no standardized writing, and with the rise of social media, it has become a language extensively used to communicate online. Algerian can be written in both Arabic and Latin scripts, and code-switching can therefore occur in a mixture of scripts, or within one same script. Arabic varieties written in Latin script are referred to as Arabizi, with north African languages referred to as North African Arabizi, NArabizi in short (Seddah et al., 2020). For the remainder of the paper, we will refer to Algerian written in Latin script as NArabizi (NA) and Algerian written in Arabic script as Algerian Arabic (DZ).

The broad usage of Algerian results in large amounts of data, with no resources or tools to automatically process them. To address this issue and further investigate which of scripts and typological differences influence the results the most, we use a corpus of user comments that reflect the nature of the Algerian vernacular dialect: with heavy use of non-standardised spellings and code-switching.

Our main contributions are (i) a new layer of annotations (transliteration, sentiment analysis, topic classification) that build on the Algerian NArabizi treebank corpus (Seddah et al., 2020), (ii) we investigate the interplay of script and typology on cross-lingual transfer for the two tasks part-of-speech (POS) tagging and sentiment analysis (SA); (iii) we give a baseline model for topic categorization for Algerian. All of the data, annotations, and models will be made freely available[1].

To the best of our knowledge, the corpus we present in this work is the first dataset of parallel

[1] https://www.AnonymizedForReview.com

Algerian texts written in NArabizi and DZ, annotated on the morphological and syntactic levels, and for which the interplay between typology and script can be investigated. We also believe that it can help developing approaches to tackle the heavy code-switched nature of the language.

In what follows, in Section 2, we give a brief overview of related works. In Section 3, we describe our dataset and annotations, the annotation processes, and give detailed statistics of the data. We start with some benchmark experiments in Section 4, and present in Section 5 our experiments for POS tagging, SA, and topic classification. In Section 6, we summarize and discuss our results, and conclude in Section 7 with our main findings and future plans.

2 Related work

The vernacular Algerian language is under-resourced, and few freely available corpora and tools exist. Despite work in recent years on this language (Adouane et al., 2020; Moudjari et al., 2020; Adouane et al., 2018; Adouane and Dobnik, 2017; Cotterell et al., 2014), there is only one corpus manually annotated for morphological and syntactical analysis (Seddah et al., 2020).

As pointed out by Seddah et al. (2020), Algerian is a non-codified spoken Semitic language. It is a morphologically-rich language (Tsarfaty et al., 2010), although less so than MSA (Saadane and Habash, 2015). Similarly to other north African languages, it uses heavy code-switching and borrowings, which can either be lexicalized borrowings that receive Arabic-like morphology, or borrowings that remain invariant or take the morphology of the borrowings' original language (*e.g.,* French). Furthermore, Algerian exhibits high variance at the morphological and phonological levels, as well as the lexicon and conventions (Seddah et al., 2020). As shown in Table 1, the Arabic name of the country "*Algeria*" can be written in various ways in both NArabizi and DZ scripts.

As in other North African languages written in Latin script, phonemes that do not exist in the Latin alphabet are represented by digits that are visually similar. For example Table 2 shows how the digits *3* and *9* are used to represent the Arabic letters "*ayin*" and "*qāf*" respectively. The nature of the language makes it therefore an interesting avenue to explore the interplay between morphological typology and differences in script on cross-lingual transfer.

NArabizi	DZ
al-dzayer	الدزاير
dzayer	دزاير
jazayer	جزاير
al-jazayer	الجزاير
al-jazaair	الجزائر

Table 1: Lexical variations of the word "*Algeria*" in Algerian written in NArabizi and DZ scripts.

Gloss	NArabizi	Arabic	D	Letter
why	we3lach	وعلاش	3	ع (ayin)
he said	9alli	قالي	9	ق (qāf)

Table 2: Example of non-Latin phonemes represented as digits in NArabizi.

The script of NArabizi differs from the more resourceful MSA and French languages, which can be seen as its culturally closest languages. However, Muller et al. (2020) show that transfer learning approaches can be used on NArabizi, both for POS-tagging and dependency parsing. They show that multilingual BERT (Xu et al., 2019) trained on Maltese, French, and English can successfully transfer to NArabizi, despite not being included in pretraining. This shows the potential for multilingual language models to transfer to unseen dialects across scripts.

The effect of morphological typology on NLP tasks is well known (Ponti et al., 2019), with several dedicated workshop series (Nicolai et al., 2020; Zampieri et al., 2018). More recently, attention has turned to larger scale analyses of morphological typology effects on language modeling (Gerz et al., 2018; Cotterell et al., 2018; Mielke et al., 2019). Cross-lingual transfer between languages with related morphological typology is more successful than between languages that do not share similar scripts (Murikinati et al., 2020; Anastasopoulos and Neubig, 2019), especially for the study of morphological inflection. Finally, regarding difference in script, Murikinati et al. (2020) find that using high-quality transliteration as preprocessing can improve the accuracy of such models.

However, in contrast to these previous works, we are interested in the interplay between morphological typology and difference in script on *cross-*

NArabizi	*ycombati la misere li las9at fina welat kiste*
Arabic transliteration	يكومباطي لا ميزار لي لسقت فينا ولات كيست
Code-switched transliteration	*kyste* ولات فينا لسقت لي *la misère* يكومباطي
English translation	he fights the misery that sticks to us and which has become a cyst

Table 3: Example of transliteration annotations into Arabic and code-switched scripts. The NArabizi is from (Seddah et al., 2020). The translation to English is added for readers' comprehension.

lingual transfer for two *supervised tasks*, namely POS tagging and sentiment analysis. More precisely, we are interested in investigating if there are differences in performance based on the various Algerian scripts.

3 Data and Annotations

The underlying dataset we use is the NArabizi treebank presented in (Seddah et al., 2020). This dataset comprises approximately 1500 sentences: 1300 NArabizi sentences extracted from an Algerian newspaper's web forum (Cotterell et al., 2014), and 200 sentences from lyrics of songs collected manually from the web. Each NArabizi sentence has five annotation layers: tokenization, morphology, identification of code-switching, syntax, and translation to French (Seddah et al., 2020). The corpus is in *conllu* format, and is freely available[2].

To investigate the interplay between script and typology for cross-lingual transfer on POS tagging and SA, we extend the annotations of (Seddah et al., 2020) by adding two levels of annotations:

Token level: for each token of the NArabizi sentences we:

1. transliterate each NArabizi token to Arabic script (*i.e.,* DZ).
2. transliterate each NArabizi token to code-switched scripts (Arabic or Latin) based on the origin of the token (and the code-switch annotation label of the treebank).

Sentence level: we annotate each sentence of the NArabizi corpus for:

1. sentiment: each sentence is annotated as POS (positive), NEG (negative), NEU (neutral), or MIX (a mix of two or more of the three previous classes).
2. topic: each sentence is annotated as belonging to *one* of the following topics: Politics, Prayer, Religion, Societal, Sport, or NONE.

All the annotations were carried out by native speakers of Algerian, Arabic, and French. Two annotators worked on the token-level annotations, and three annotators for the sentence-level annotations. Before starting the annotations, we did a common annotation round to agree on the guidelines, and discuss possible issues. During this, we identified a set of errors in the NArabizi treebank, we therefore started by preprocessing the data and correct some of the recurring errors. More details about our preprocessing of the dataset is given in Section 3.1, the transliteration annotations are described in Section 3.2, and sentiment and topic annotations are described in respectively Section 3.3 and Section 3.4.

3.1 Annotation Preprocessing

The NArabizi treebank dataset (Seddah et al., 2020) contains duplicates both in document IDs and in sentences (strings), both across splits and within splits. Duplicate IDs refer to the same sentences, and therefore duplicate IDs imply duplicate sentences. However, duplicate sentences represent same strings with different IDs. There are far more sentence duplicates than ID duplicates.

All duplicates were removed. However, as the corpus is already quite small, we attempt to avoid removing duplicates from the dev and test splits. For the inter-split duplicates, we identified 9 duplicated IDs and 46 (12 unique) duplicated sentences. Intra-split duplicates were only present in train split, with 9 duplicated IDs, and 28 (8 unique) duplicated sentences. We kept one occurrence of each as it seems that most of these duplicates come from the chorus of the song lyrics, and short common utterances as *e.g.,* "viva Algeria".

[2] https://parsiti.github.io/NArabizi/

Letter	Transliteration
v	ف (*f*)
p	ب (*b*)
g	ڤ (*gu*)

Table 4: Normalization of some of the Latin characters that do not have equivalent phonemes in Arabic.

3.2 Transliteration to Arabic and code-switched scripts

Two annotators expanded the annotations of the NArabizi treebank by Seddah et al. (2020) by adding for each token of each sentence a transliteration into Arabic script, and a code-switched version that includes both Latin and Arabic scripts. The Latin script is used for tokens that originate from Latin-scripted languages.

For example, Table 3 shows how the NArabizi sentence is transliterated into the corresponding DZ and code-switched scripts. The first word, "يكومباطي", is actually a borrowing from French. However, borrowings that are integrated into the Algerian language lexicon, and that are influenced by Arabic verbal inflections, were not written in Latin script in the code-switched annotations.

The two annotators were given 300 overlapping sentences to transliterate (due to the lack of codification, we do not compute any inter-annotator agreement). The overlapping sentences were mainly used to set the annotation guidelines, and were extensively discussed by the annotators. We decided to normalize some of the Latin characters that do not have equivalent pronunciations in Arabic, these were transliterated into what the native annotators deemed to be the corresponding Arabic characters. In Table 4 we show the Latin letters and the Arabic form they were transliterated to. Even so, we decided to transliterate the last letter (phoneme *gu*) into a non-native Arabic letter. This letter is vastly used in various Algerian dialects, it represents the dialectal pronunciation of *qāf*, and is also used in names of places and persons.

We are aware of the various efforts to develop guidelines for conventional orthography of Algerian and other Arabic dialects (Saadane and Habash, 2015; Habash et al., 2018; Adouane et al., 2019), but we decided to keep the transliterations as identical as possible to the original NArabizi pronunciations and spellings, to reflect the distinctiveness of the language and its use in normal settings on social media.

During the transliteration annotations, several issues were identified in the original NArabizi treebank by Seddah et al. (2020). However, since our annotators were not trained to alter the dependency treebanks, only a small selection of the identified errors were corrected.

The first problem encountered is a lack of consistency in the tokenization. For example, the definite article "*el*" can be found both as a stand-alone token, or attached to a word. The same applies to the adposition "*in*" where it can be found both as a stand-alone token, and attached to the next word. For example, it was kept with the token in "*f'doute*" ("*in doubt*"), while it was tokenized as "*f+almarikhe*" for the word "*falmarikhe*" ("*in Mars*"). All tokenization errors were not corrected, as this would lead to altering the dependency trees, and as previously mentioned, our annotators were not trained for this task.

Secondly, there were also errors in the translations from NArabizi to French. This is likely due to non-native Algerian speakers translating some parts of the NArabizi treebank. We only corrected the translations that did not alter the tree, *i.e.,* the POS did not change. Some examples of these types of errors can be found in Table 10 in Appendix A.

Finally, we also found some errors in the marker for code-switching (label *lang* in the data). Some Algerian tokens were marked as French, and vice-versa. This also happened with other languages present in the data (as Spanish, English, and MSA). One of the typical errors was the acronyms of football clubs which were all labeled as Algerian. These were corrected to French, since the acronyms come from their names in French. For example the football club "*MCA*" stands for "*Mouloudia Club d'Alger*", while the Arabic name is "*Nadi mouloudiat al-jazair*".

3.3 Sentiment annotations

The sentences were classified based on their polarities into four different classes: POS (positive), NEG (negative), NEU (neutral), and MIX (mixed). The annotation guidelines were quite simple, and annotators were asked to use POS and NEG in clear positive and negative cases respectively. If a sentence do not express any kind of polarity, then NEU was assigned. When sentences express a combi-

		Train	Dev	Test
Sentiment	POS	291	32	59
	NEG	274	44	34
	NEU	191	21	20
	MIX	242	40	31
Topic	NONE	300	34	36
	Politics	80	11	16
	Prayer	38	9	9
	Religion	17	4	1
	Societal	204	25	31
	Sport	359	54	51

Table 5: Distribution of sentiment and topic annotations.

		NA	DZ	CS
Sentiment	BOW	47.5	45.1	**49.5**
	AVE	52.5	43.2	36.7
	CNN	50.2	**50.4**	46.2
	BiLSTM	**53.9**	45.9	45.6
Topic	BOW	25.8	34.9	**38.1**
	AVE	40.9	44.6	22.8
	CNN	24.4	33.4	27.4
	BiLSTM	**49.4**	**57.0**	36.4

Table 6: Benchmark results for Sentiment Analysis and Topic classification on the three varieties of the dataset: NArabizi (NA), Algerian Arabic (DZ), and a code-switched version (CS). Sentiment and Topic are both Macro F_1.

nation of two or more of the POS, NEG, or NEU polarities, annotators were asked to assign the MIX label. The inter-annotator agreement using *Cohen's kappa coefficient* κ is 0.71 on the doubly annotated subset of 300 sentences. Table 5 shows the distribution of the four labels across the training, development, and test sets. The distribution is unbalanced, and the large amount of sentences categorized as MIX can be problematic as it can contain all other polarities. However, the difference between the POS and NEG classes is relatively small, which we believe should be suitable for binary sentiment classification tasks.

3.4 Topic annotations

After a first round of common analysis in collaboration with the annotators, we identified five topics. However, some sentences were difficult to classify and we therefore decided to include the category "*NONE*". The final dataset is annotated for the following six categories: (1) *Politics*: contains all sentences referring or discussing political events or issues; (2) *Prayer*: all sentences representing prayers; (3) *Religion*: sentences discussing religious issues or issues related to religion in general; (4) *Societal*: societal related discussions. Covers everything from schools and teaching, to terrorism and extremism; (5) *Sport*: mainly covering football events, but spans all types of sports and related events; (6) *NONE*: sentences that were impossible to categorize. This was mainly due to the lack of context, as some sentences were comments responding to either articles or other comments. The final κ score for the triply annotated 300 sentences was 0.70.

Table 5 also shows the distribution of topics across the three splits. Most sentences were classified as "Societal" and "Sport". A large amount of sentences could not be categorised, and few sentences were related to "Religion" and "Prayer". Due to the size of the two latter, one could argue that they could be collapsed into a single topic, as done in our benchmarking experiments (see Section 4). However, we decided to keep them separate in the annotations, to facilitate further annotations in the future.

4 Benchmarking experiments

We perform benchmark experiments for SA and topic classification. Specifically, we use the setup from Barnes et al. (2017), who perform experiments with a logistic regression classifier with bag-of-words features (BOW) and averaged embedding features (AVE), as well as a CNN and BiLSTM. We use their default value for hyperparameters (c=1, hidden dimension = 100, dropout = 0.3) and train for 20 epochs, finally testing the best model on the dev set. As the label distribution for both tasks is highly skewed, we use Macro F_1 to evaluate. Given the size of the categories "Prayer" and "Religion", we collapse them to a single topic, converting the topic classification task into a 5-class multi-class problem.

For the NArabizi and code-switched experiments, we create 100-dimensional fasttext embeddings (Bojanowski et al., 2017) on the unlabeled NArabizi data made available by Seddah et al. (2020). For the DZ experiments, we use avail-

		UPOS		Sentiment			
group	language	# sents.	avg. len.	# docs.	avg. len.	pos	neg
Original	NArabizi	1,276	16.1	731	14.4	380	351
`Script`	Persian	5,997	26.3	879	49.6	419	460
`Script`	Urdu	5,130	27.0	980	17.5	480	500
`Typology`	Hebrew	6,216	30.6	12,434	24.0	8,512	3,922
`Typology`	Maltese	2,074	21.8	719	18.7	237	482
`Both`	MSA	7,664	42.3	51,051	60.8	42,828	8,223

Table 7: Statistics of UPOS and binarized sentiment data.

able 300-dimensional MSA fasttext embeddings[3] trained on Wikipedia articles.

Table 6 shows the results. On the sentiment task, the BiLSTM performs best on NArabizi, the CNN best on DZ, and BOW best on the code-switched data. For topic classification the performance is similar, but BiLSTM is also best on DZ. The fact that BOW performs best on code-switched data is largely due to the large amount of out-of-vocabulary words for all other methods, which require embeddings. These baseline experiments show that the dataset is challenging, and the variation means that no single model is always best. The code-switched setting is particularly challenging.

5 The interplay between morphological typology and script

The transliteration and further sentiment and topic annotations allow us to explore what interplay there is between morphological typology and script in cross-lingual transfer. Muller et al. (2020) perform experiments on zero-shot cross-lingual transfer for POS tagging on NArabizi. They find that the best transfer language is Maltese, a Semitic language which is written in Latin script, rather than MSA, which performs poorly. This begs the question: *is it mainly morphological similarity or a shared script that leads to this result?* The transliterated dataset, along with the further sentiment annotations allows us to investigate this question in more depth, as we are able to control for the script choice.

We choose Persian and Urdu, languages written in Arabic script, but morphologically distinct from DZ (we refer to this group as `Script`), Hebrew and Maltese, two Semitic languages written in other scripts (`Typology`), and MSA, which is both morphologically similar and written in Arabic script (`Both`). These languages are both available in UD (Zeman et al., 2020) and also have available sentiment analysis datasets (Hebrew (Amram et al., 2018), Maltese (Dingli and Sant, 2016), MSA (Nabil et al., 2015; Abdulla et al., 2013), Urdu (Khan and Nizami, 2020), Persian (Hosseini et al., 2018)). As not all sentiment datasets have the same labels as the NArabizi dataset, we remove all neutral and mixed labels and create binary sentiment data for all languages.

Table 7 gives an overview of the statistics of the POS and SA datasets, respectively. The NArabizi data is the smallest POS data (1,276 sentences), followed by Maltese (2,074), Urdu (5,130), Persian (5,997), Hebrew (6,216) and finally MSA (7,664). The average sentence lengths in tokens range between 16.1 for NArabizi and 42.3 in MSA. The sentiment datasets have a larger variance, ranging from 719 sentences for Maltese to 51,051 for MSA. The distribution of polarity is also skewed to a different degree in each dataset.

5.1 Modeling

We model universal POS (UPOS) tagging as a sequence labeling task and SA as a classification task using multilingual BERT (Xu et al., 2019). We fine-tune each model on the available training data in each language, using a shared set of hyperparameters which were selected from recommended values according to the characteristics of our data. We set the learning rate to 2e-5, max sequence length of 256, batch size of 8 or 16[4], and perform early stopping once the validation score has not improved in the last epochs, saving the model that performs best on the dev set. We then test each model on its own

[3] Available at https://fasttext.cc/docs/en/pretrained-vectors.html.

[4] Depending on the size of the training set, model architecture, and available GPU memory.

dev and test data, the NArabizi test set, and finally the transliterated data. We use accuracy as our metric for POS and macro F_1 for sentiment, as the latter often contains unbalanced classes, and define a baseline as the result of predicting the majority class.

6 Results and Discussion

In order to quantify the zero-shot loss, we define a measure of average transfer loss between a group in Equation 1:

$$TL_{x \to y} = S_{x \to x} - S_{x \to y} \quad (1)$$

where $TL_{x \to y}$ is the transfer loss experienced by a model fine-tuned in language x when transferring to language y and $S_{x \to y}$ is the score[5] achieved when testing a model fine-tuned in language x on language y. Thus, it is a measure of the performance lost in the transfer process.

We also define its averaged variant:

$$\overline{TL}_{A \to B} = \frac{1}{N_A} \sum_{i \in A} \overline{TL}_{i \to B} \quad (2)$$

where $\overline{TL}_{A \to B}$ refers to the average transfer loss experienced by languages from any group A to languages from group B (*group-to-group* transfer loss) and N_A is the number of languages included in the experiment that belong to group A (in our case, either languages that are morphologically similar, or have the same script).

6.1 POS

Table 8 shows the results for the POS tagging. For completeness, we compare with the results from Seddah et al. (2020), who use a feature-based alVWTagger, described in more detail in de La Clergerie et al. (2017) and Muller et al. (2020), who use mBERT and the StanfordNLP tagger (Qi et al., 2018).

Hebrew has the best test accuracy (96.8) and Maltese the worst (93.8), while the others are somewhere between. All models perform better on the transliterated data than the original NArabizi, although training on Urdu performs lower than the majority baseline. This suggests that even though mBERT was not pretrained on NArabizi or DZ, there is a preference for DZ. This is likely due to the fact that at least *some* of the words have been seen in pretraining, *i.e.,* through MSA. An analysis of the tokenization shows that mBERT splits NArabizi words at a much higher rate than DZ (see Figure 1), breaking it into smaller pieces, which may account for some of the differences between the two. The fact that training on Maltese achieves the best score on both NArabizi (37.8) and DZ (38.4), however, suggests that there is still an effect of typology.

The monolingual model trained and tested on DZ performs better (82.5 acc.) than the one trained and tested on NArabizi (76.3). When each of these models is tested on the other, they have significant transfer losses (32.7 for NArabizi → DZ, and 42.6 for DZ → NArabizi). Here too, transfer to DZ script seems easier.

On POS, the effect of language typology is

[5]The score metric will depend on the task: accuracy in POS and macro F_1 in sentiment analysis.

	Dev	Test	NA	DZ
Maj.	–	–	19.9	19.9
1-NArabizi		–	80.4	–
2-NArabizi	–	–	81.6	–
2-Maltese	–	–	35.1	–
NArabizi	77.1	–	76.3	43.6
Algerian (DZ)	83.2	–	39.9	82.5
Persian	95.8	95.5	22.7	26.5
Urdu	94.0	93.4	18.7	21.6
Hebrew	97.4	96.8	32.7	38.2
Maltese	93.6	93.0	**37.8**	**38.4**
MSA	97.0	96.7	20.0	30.5

Table 8: POS accuracy when training on Train Lang. Dev Acc. and Test are in-language, while Test Acc. on Narabizi and Algerian (DZ) is zero-shot cross-lingual. 1-Seddah et al. (2020), 2-Muller et al. (2020).

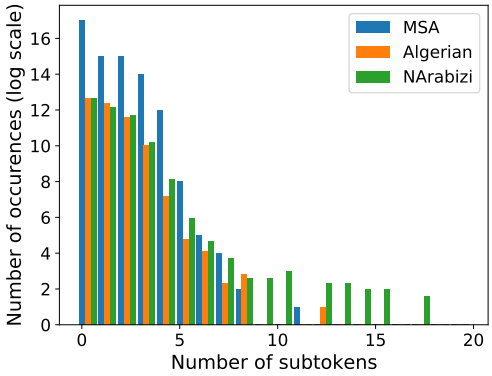

Figure 1: The effect of mBERT tokenization on MSA, Algerian Arabic (DZ), and NArabizi.

stronger than script, with the best results achieved by training on Maltese and Hebrew. The average transfer loss from Persian and Urdu to DZ is 70.4 while for Hebrew and Maltese to NArabizi it is 59.6, showing less transfer loss from `Typology`. MSA has higher transfer loss on NArabizi (76.7) than DZ (66.2). The differences between average transfer loss on NArabizi and DZ are also slightly larger for `Script` (3.4) compared to `Typology` (3.1) or MSA (3.1).

All of this points to a complicated relationship between script and typology on POS. First of all, it is clear that mBERT prefers the Arabic script seen in pretraining. At the same time, morphological similarity also plays a strong role in cross-lingual transfer in POS, although even in this case, the best scores are found on DZ.

6.2 Sentiment

Table 9 shows the results for sentiment analysis. Training in-language again produces the best results (72.1 and 80.3 on NArabizi and DZ, respectively). The transfer loss from NArabizi to DZ is relatively low (9.0), while inversely it is immense (52.6).

Like on POS, most models perform better on DZ and the best zero-shot results do not come from training on `Typology`. In fact, quite the opposite, as these lead to the worst scores and have the highest average transfer loss (34.8/27.9). The best models are MSA for NArabizi (62.4) and Urdu for DZ (63.9), which curiously performs better than NArabizi $\rightarrow$ DZ. MSA has transfer losses of 12.8/25.1, while `Script` have the lowest average transfer loss (9.2/4.2). This suggests that cross-lingual transfer for a more semantic task, *e.g.,* sentiment analysis, is less reliant on both typological and script similarities.

	Dev	Test	NA	DZ
Maj.	–	–	39.0	39.0
NArabizi	78.8	–	72.1	63.1
Algerian (DZ)	84.9	–	27.7	80.3
Persian	65.9	66.2	56.9	56.2
Urdu	59.0	62.4	53.3	**63.9**
Hebrew	88.4	88.7	47.2	52.2
Maltese	63.7	61.8	33.8	42.5
MSA	74.2	75.2	**62.4**	50.1

Table 9: Macro F_1 on the zero-shot cross-lingual sentiment task. Note that these results are not comparable to the benchmark experiments, as the data has been converted to binary sentiment classification in order to perform the cross-lingual experiments.

7 Conclusion and Future work

In this paper we have described the process of annotating an available Algerian corpus with sentiment and topics, as well as the transliteration to Arabic and code-switched scripts, and finally some aspects of corpus cleanup. We performed benchmark experiments on the three script varieties and show that they are a challenging testbed for future experiments.

We used this new resource to explore a valuable research question in cross-lingual transfer: namely, what is the interplay between morphological typology and script when choosing a source language? We found there is a delicate interplay between morphological typology and script for transfer in part-of-speech tagging, where morphology is more important, but having seen the script in pretraining also influences results. Sentiment analysis, on the other hand, is less sensitive to morphological differences, while still preferring the script seen in pretraining. This suggests that choice of transfer language is task-specific and that surprising differences can appear from one task to another.

In the future, we would like to address data related issues, and correct the tokenization and translation issues discussed in Section 3.1. Moreover, we plan to focus more concretely on the code-switching aspect of our dataset. The challenges of code-switched data to NLP techniques are numerous, and we would like to focus on the syntactic analysis of our code-switched data, and to explore in more details language modeling approaches to processing it.

References

Nawaf Abdulla, Nizar A. Ahmed, Mohammed Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, AEECT 2013*, pages 1–6.

Wafia Adouane, Jean-Philippe Bernardy, and Simon Dobnik. 2018. Improving neural network performance by injecting background knowledge: Detecting code-switching and borrowing in Algerian texts. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*,

pages 20–28, Melbourne, Australia. Association for Computational Linguistics.

Wafia Adouane, Jean-Philippe Bernardy, and Simon Dobnik. 2019. Normalising non-standardised orthography in Algerian code-switched user-generated data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 131–140, Hong Kong, China. Association for Computational Linguistics.

Wafia Adouane and Simon Dobnik. 2017. Identification of languages in Algerian Arabic multilingual documents. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 1–8, Valencia, Spain. Association for Computational Linguistics.

Wafia Adouane, Samia Touileb, and Jean-Philippe Bernardy. 2020. Identifying sentiments in Algerian code-switched user-generated comments. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2698–2705, Marseille, France. European Language Resources Association.

Adam Amram, Anat Ben David, and Reut Tsarfaty. 2018. Representations and architectures in neural sentiment analysis for morphologically rich languages: A case study from modern Hebrew. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2242–2252, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, page 984–996, Hong Kong, China,. Association for Computational Linguistics.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–12, Copenhagen, Denmark. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Ryan Cotterell, Adithya Renduchintala, Naomi Saphra, and Chris Callison-Burch. 2014. An algerian arabic-french code-switched corpus. In *Workshop on free/open-source Arabic corpora and corpora processing tools workshop programme*, page 34.

Alexiei Dingli and Nicole Sant. 2016. Sentiment analysis on maltese using machine learning. In *Proceedings of The Tenth International Conference on Advances in Semantic Processing (SEMAPRO 2016)*, pages 21–25.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al-Shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pedram Hosseini, Ali Ahmadian Ramaki, Hassan Maleki, Mansoureh Anvari, and Seyed Abolghasem Mirroshandel. 2018. Sentipers: A sentiment analysis corpus for persian. *CoRR*, abs/1801.07737.

M. Y. Khan and M. S. Nizami. 2020. Urdu sentiment corpus (v1.0): Linguistic exploration and visualization of labeled dataset for urdu sentiment analysis. In *2020 International Conference on Information Science and Communication Technology (ICISCT)*, pages 1–15.

Éric de La Clergerie, Benoît Sagot, and Djamé Seddah. 2017. The ParisNLP entry at the ConLL UD shared task 2017: A tale of a #ParsingTragedy. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 243–252, Vancouver, Canada. Association for Computational Linguistics.

Sebastian J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Leila Moudjari, Karima Akli-Astouati, and Farah Benamara. 2020. An Algerian corpus and an annotation platform for opinion and emotion analysis. In *Proceedings of the 12th Language Resources*

and Evaluation Conference, pages 1202–1210, Marseille, France. European Language Resources Association.

Benjamin Muller, Benoit Sagot, and Djamé Seddah. 2020. Can multilingual language models transfer to an unseen dialect? a case study on north african arabizi.

Nikitha Murikinati, Antonios Anastasopoulos, and Graham Neubig. 2020. Transliteration for cross-lingual morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–197, Online. Association for Computational Linguistics.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. ASTD: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, Lisbon, Portugal. Association for Computational Linguistics.

Garrett Nicolai, Kyle Gorman, and Ryan Cotterell, editors. 2020. *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Online.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Houda Saadane and Nizar Habash. 2015. A conventional orthography for Algerian Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 69–79, Beijing, China. Association for Computational Linguistics.

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150, Online. Association for Computational Linguistics.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Yannick Versley, Marie Candito, Jennifer Foster, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (spmrl) what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–12.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Nikola Ljubešić, Jörg Tiedemann, Shervin Malmasi, and Ahmed Ali, editors. 2018. *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Angelina Aquino, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Ethan Chi, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig,

Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-apińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. Universal dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Appendices

id – Sentence (En)	Translation	Correct (En)	Explanation	Status
1 – *Mabrouk ya lafhal* 1 (congratulations oh brave)	Lafhal	courageux (brave)	annotated as PROPN, should be NOUN.	X
2 – *el hamdou lilah ya rabi alla 3awdat chawchi* (thanks God for the return of Chawchi)	Allah	alla (for)	this is the word على annotated as PROPN, should be DET.	X
3 – *vive toi mbolhi* (long live you Mbolhi)	fou (crazy)	Mbolhi	Mbolhi is the name of a football player. It is not an ADJ, should be PROPN.	X
4 – *mabka fiha ghure sehab elderaham* (the only ones remaining are those with money)	pas pleurer (not cry)	mabka (only remain)	this is the word ما بقى (only remain) and not ما بكى (not cry).	✓
5 – *al mou3ak fil jazair mayakdarch yakhrouj* (the handicapped in Algeria can't go out)	obstacle (obstacle)	handicapé (handicapped)	the word *mou3ak* in this context means handicapped.	✓

Table 10: Examples of errors present in the NArabizi treebank. Status "X" means not corrected, while status "✓" means corrected.