Exploring the Impact of Transliteration on NLP Performance: Treating Maltese as an Arabic Dialect

Kurt Micallef[1]
kurt.micallef@um.edu.mt

Fadhl Eryani[2,3]
fadhl.eryani@nyu.edu

Nizar Habash[3]
nizar.habash@nyu.edu

Houda Bouamor[4]
hbouamor@cmu.edu

Claudia Borg[1]
claudia.borg@um.edu.mt

[1]Department of Artificial Intelligence, University of Malta
[2]University of Tübingen
[3]Computational Approaches to Modeling Language Lab, New York University Abu Dhabi
[4]Carnegie Mellon University Qatar

Abstract

Multilingual models such as mBERT have been demonstrated to exhibit impressive cross-lingual transfer for a number of languages. Despite this, the performance drops for lower-resourced languages, especially when they are not part of the pre-training setup and when there are script differences. In this work we consider Maltese, a low-resource language of Arabic and Romance origins written in Latin script. Specifically, we investigate the impact of transliterating Maltese into Arabic scipt on a number of downstream tasks: Part-of-Speech Tagging, Dependency Parsing, and Sentiment Analysis. We compare multiple transliteration pipelines ranging from deterministic character maps to more sophisticated alternatives, including manually annotated word mappings and non-deterministic character mappings. For the latter, we show that selection techniques using n-gram language models of Tunisian Arabic, the dialect with the highest degree of mutual intelligibility to Maltese, yield better results on downstream tasks. Moreover, our experiments highlight that the use of an Arabic pre-trained model paired with transliteration outperforms mBERT. Overall, our results show that transliterating Maltese can be considered an option to improve the cross-lingual transfer capabilities.

1 Introduction

The availability of multilingual models has facilitated the development of NLP tools for many low-resource languages. Their appeal not only comes from this universal language representation but also through leveraging data from related languages (Kondratyuk and Straka, 2019; Wu and Dredze, 2019; Conneau et al., 2020). Despite this, multilingual models may fall short especially for lower-resourced languages (Wu and Dredze, 2020; Muller et al., 2021). In particular, Muller et al. (2021) show that the cross-lingual transfer capabilities are hampered due to script differences between the target language and the related language which is part of the multilingual model pre-training. However, they show that performance is dramatically improved by transliterating the target language to the same script as the related language. Unlike translation, transliteration is a relatively cheap process which maps characters in a given script to another.

In this work, we focus on Maltese, a low-resource hybrid/mixed language of Semitic origin but written in Latin script. Although it retains a strong Semitic, specifically Arabic, component in its grammar, it borrows heavily from Romance (Italian) and English. This motivates us to explore the impact that transliterating Maltese into Arabic script could have on a number of downstream tasks. Besides using large language models based on Arabic text, this experimental setup opens up interesting research questions about the impact of high ambiguity from Arabic orthographic choices, such as dropping diacritics which may lower out-of-vocabulary, but also increase ambiguity.

Despite its Arabic roots, transliterating Maltese to Arabic script is not trivial because of the strong non-Arabic influences on Maltese and its evolution independent of the Arab world for a significant number of years (Sutcliffe, 1936; Borg and Azzopardi-Alexander, 1997). That said, Arabic-transliterated Maltese can be deemed as an Arabic dialect with a higher degree of Italian code-switching. As such, unlike Muller et al. (2021), we do not just rely on multilingual language models for cross-lingual transfer, but also make use of Arabic language models, specifically CAMeLBERT (Inoue et al., 2021). We compare its performance

Proceedings of the Workshop on Computation and Written Language (CAWL 2023), pages 22–32
July 14, 2023 ©2023 Association for Computational Linguistics

to multilingual BERT (Devlin et al., 2019) and monolingual Maltese BERT (Micallef et al., 2022). Empirically, we show that there are differences in the cross-lingual transfer capabilities of Arabic models and multilingual models for Maltese.

The main contributions of this work are as follows. We present various transliteration pipelines from Maltese to Arabic script ranging from simple one-to-one character maps to more sophisticated alternatives that explore multiple possibilities or make use of manually annotated linguistic constructions. We show that the sophisticated systems are consistently better than simpler systems, quantitatively and qualitatively. We also show that despite its hybrid nature, transliterating Maltese can be considered as an option to improve the cross-lingual transfer capabilities.

The rest of the paper is organized as follows. We present a discussion of motivating linguistic background (Section 2) followed by our approach for transliterating Maltese to Arabic script (Section 3). We present our evaluation in Sections 4 and 5.

2 Linguistic Background

2.1 Arabic and Maltese

Arabic and Maltese are closely related Semitic languages. Arabic is the national language of ∼360 million people across 22 countries (Eberhard et al., 2022), while Maltese is the national language of Malta and is spoken by ∼500,000 people (Rosner and Borg, 2022).

The Arabic language is a collection of coexisting varieties. While Classical Arabic (CA) still survives in Muslim religious ceremonies, Modern Standard Arabic (MSA) is the official national language of the media and formal education, but neither CA nor MSA is the *mother tongue* of Arabs today. A number of dialectal Arabic (DA) varieties are primary spoken (and increasingly informally written varieties). The coexistence of MSA and DA is described as diglossia (Ferguson, 1959). Foreign languages, particularly French (in the Maghreb) and English (in the Middle East) have a strong presence in DA and result in common code-switching (Hamed et al., 2020).

The Maltese language traces its origins to medieval Sicilian Arabic. In its current form, Maltese contains elements from Arabic, Italian, and most recently English. This reflects its geography and history: the Mediterranean island of Malta is halfway between Tunisia and Italy, and historically it was under Arab rule from 870 to 1090 CE. Being the language of a Christian nation, Maltese has no CA influences, unlike other Arabic dialects with diglossia (Sutcliffe, 1936; Borg and Azzopardi-Alexander, 1997). In some simplifying respects, Maltese can be seen linguistically as a dialect of Arabic with a higher degree of code-switching to Italian. Čéplö et al. (2016) reports that mutual intelligibility between Maltese, Tunisian Arabic (TA), and Libyan Arabic (LA) ranges between 30% and 40%, with TA having the highest level of mutual intelligibility with either of the other two varieties.

2.2 Script and Orthography

The most important difference between Arabic and Maltese is the use of Arabic and Latin scripts, respectively. Furthermore, the two languages use different orthographic philosophies in how to map linguistic features (phonology and morphology) to script letters. Given the topic of this paper and to facilitate the presentation of the remaining sections, we will start with a discussion of the two scripts and the orthographic philosophies they use.

The Arabic script is used to write a number of languages from different language families, e.g. Arabic (Semitic), Persian (Indo-European), and Uyghur (Turkic). The Arabic script is mostly used as an Abjad where diacritical marks represent short vowels and consonantal doubling, although there are exceptions such as Uyghur's Arabic alphabet. Since most diacritics are optionally written in the context of the Arabic *language* (Abjad) (Habash, 2010), this leads to a high degree of ambiguity. It should be noted, however, that initial vowels are always marked by having a word initial Alif ا *A* as a diacritic carrier; and final vowels typically reflect some deeper morphological feature of the word such as a weak verbal root radical or a nominal feminine ending. Arabic effectively relies on its strong templatic morphology that allows readers to limit the ambiguity space. Arabic orthography also tends toward morphophonemic spelling which abstracts away from allomorphy, e.g. the Arabic definite determiner proclitic ال *Al* has a number of allomorphs that assimilate with word-initial coronal consonants (so-called Sun Letters) but is always written in the morphemic form: الشمس *Al+šms* /*aš+šams*/ 'the+sun'. Finally, while MSA has standard rules for orthography, DAs do not. Arabic NLP researchers have developed conventions for writing DA to allow studying spelling

		(a)	(b)	(c)	(d)	(e)	(f)	(g)
(i)	**Maltese**	Min	ma	jġarrabx	il-ħażin	ma	jafx	it-tajjeb
	Arabic	مين *myn*	ما *mA*	يجربش *yjrb$*	الحزين *AlHzyn*	ما *mA*	يفش *yf$*	الطيب *AlTyb*
	Gloss	who	not	he-experiences-not	the-bad	not	he-knows-not	the-good
	Origin	AR	AR	AR	AR*	AR	AR*	AR
	English	He who has not experienced what is bad cannot know the worth of what is good						

(ii)	**Maltese**	Il-bnedmin	kollha	jitwieldu	ħielsa	u ugwali	fid-dinjità	u d-drittijiet
	Arabic	البنادمين *AlbnAdmyn*	كلها *klhA*	يتوالدوا *ytwAldwA*	خالصة *xAlSp*	واجوالي *wAjwAly*	فالدنيتا *fAldnytA*	والدريتيات *wAldrytyAt*
	Gloss	the-humans	all	they-are-born	free	and equal	in-the-dignity	and the-rights
	Origin	AR	AR	AR*	AR*	AR IT	AR IT	AR IT*
	English	All human beings are born free and equal in dignity and rights						

Table 1: Two Maltese examples paired with their idealized Arabic script orthography. Example (i) is a traditional proverb, and example (ii) is the first sentence in the Universal Declaration of Human Rights. The tags in Origin are AR=Arabic, IT=Italian, and *=*modified*. Arabic is presented from left to right to align with Maltese. Arabic Romanization is in the Buckwalter scheme (Habash et al., 2007).

varieties (Zribi et al., 2014; Habash et al., 2018).

The Maltese script is based on the Latin script with some extension ns (*ċ*, *ġ*, *ħ*, and *ż*). The Maltese orthographic philosophy is in some way diametrically opposed to Arabic's more abstracting orthographic philosophy: Maltese tries as much as possible to reflect the phonological form of the words. There are a few exceptions to this principle, which are felicitous for our task. First, Maltese marks the form of the definite determiner with a hyphen. The number of determiner variants is quite large (∼150) due to allomorphy from phonetic assimilation and proclitics, e.g., *il-*, *l-*, *ix-*, *x-*, *is-*, *s-*, *it-*, *t-*, *id-*, *d-*, are just a few forms of the definite determiner, all of which map to Arabic ال *Al*; this is in addition to many cliticized forms such as *lill-*, *lix-*, *lis-*, *lit-*, *lid-* (all with the preposition *lil* 'for') or *tal-*, *tax-*, *tas-*, *tat-*, *tad-* (all with the preposition *ta'* 'of') (Sutcliffe, 1936). Second, Maltese writes some consonants to reflect their etymological link to Arabic, e.g. *għ* which is mostly silent and corresponds to Arabic ع/غ *E/g* (Fabri et al., 2014). Third, Maltese spells the commonly used conjunction *u* 'and' separately from the word, whereas Arabic attaches it to the following word. Finally, Maltese has access to capital letters, a concept that has no parallel in Arabic script. Capital letters are used in Maltese in similar ways to English, marking proper nouns. We leave the use of capitalization as an additional modeling feature to future work.

2.3 Phonological Differences

Maltese lost many Arabic phonological features. These include all emphatic consonants (Walter, 2006), e.g. the *s* letters in Maltese *sejf* 'dagger' and *sajf* 'summer', correspond to two letters in their Arabic cognates, سيف *sayf* and صيف *Sayf*. Other examples include the Arabic voiced pharyngeal (ع *E*) and voiced uvular (غ *g*) merging into Maltese *għ*; and the voiceless versions of both (ح *H* and خ *x*) merging into Maltese *ħ*, among others. Many of the Arabic cognates in Maltese with the Qaf consonantal variable ق *q* are spelled in Maltese with *q* although pronounced as a glottal stop (as in Urban Levantine and Egyptian Arabic), e.g. Maltese *triq* 'street' طريق *Tryq*. Due to Italian influences, the Maltese phonetic inventory has acquired a number of non-Arabic sounds such as *p*, and *v*.

Maltese has six vowels (*a*, *e*, *i*, *o*, *u*, *ie*), the first five of which may be shortened in some contexts. Standard Arabic has three short and three long vowels; while most dialects expand the set to five short and long. Arabic short vowels are generally written with diacritical marks with exceptions due to underlying derivation, or word position (initial/final) (Habash et al., 2018).

2.4 Morphological Differences

Maltese morphology shares a lot of features with Maghreb Arabic morphology, and Arabic/Semitic morphology in general. This include a rich inflectional space (person, gender, number, aspect) and many clitics that both Arabic and Maltese write as part of the word form. For example, Maltese *dar* 'house', a cognate of Arabic دار *dAr*, can be inflected into *id-dar* 'the house', *tad-dar* 'of the house', *f'dar* 'in a house', and *darha* 'her house' which correspond to Arabic الدار *AldAr*, تاع الدار *tAE AldAr*, فدار *fdAr*, and دارها *dArhA*, respectively. While Maltese marks the determiner with a hyphen, which provides a strong morphological signal (Section 2.2), it does not mark pronominal clitics or negation particles similarly. In this paper, we do not use any morphological analysis and disambiguation tools to help in transliteration. We leave this direction to future work.

2.5 Lexical Differences

While there are many shared words between Maltese and Arabic (especially Maghreb and Tunisian Arabic), there are important differences. Table 1 highlights some examples of both kinds, but these can be further broken down. First are words that undergo a major phonological shift, e.g. *jafx* 'does not know' comes from Arabic يعرفش *yErf$*. Second are words that went through a semantic shift, e.g. *ħażin* 'bad' is related to Arabic حزين *Hzyn* 'sad'. Thirdly, Maltese has many univerbation constructions which Arabic generally avoids, e.g. Maltese *waranofsinhar* 'afternoon' corresponds to three Arabic cognate words ورا نفص النهار *wrA nfS AlnhAr* 'after middle [of] the day'. Finally, Italian-origin words are all distinct from Arabic, although in some cases they may have cognates in the dialects, e.g. *kċina* 'kitchen' corresponds to TA كوجينه *kwjynh* (Aquilina, 1987, 1990).

2.6 NLP Conventions in Arabic and Maltese

NLP research conventions have developed independently in Arabic and Maltese, posing challenges to working on them jointly. For example, basic tokenization in the Universal Dependency Treebanks for Arabic follows a relatively deep morphological tokenization that separates all clitics (except the determiner), and normalizes the form of the baseword (Nivre et al., 2017; Taji et al., 2017). In contrast, Maltese tokenizes the determiner but not much more, leaving all other enclitics attached to the word and proclitics attached to the determiner (Čéplö, 2018). We follow the Maltese decisions here to simplify our training and evaluation.

3 Our Transliteration System

We present a Maltese-to-Arabic transliteration system with a number of variants that we evaluate in Sections 4 and 5. The transliteration system contains two operations: **mapping** and **ranking**. Maltese text tokens and characters are mapped from Latin script to one or more alternatives in Arabic script (Section 3.2). Then, a separate component ranks the choices or uses a deterministic hardcoded baseline (Section 3.3).

3.1 Preprocessing

As discussed in Section 2.6, we operate on tokenized Maltese to allow us to maintain label alignments from the training data of the downstream tasks. As such, Maltese texts are first tokenized using the MLRS tokenizer.[1] Next, all Maltese texts are lower-cased, since there is no casing information in Arabic; and all Latin script diacritics are removed, excluding those relevant to Maltese, namely: *ċ*, *ġ*, *ħ*, and *ż*. For example, *soċjetà* 'society', which is a remnant of Italian, is mapped to *soċjeta*, reflecting a common form of spelling such words in standard Maltese.

3.2 Character and Token Mappings

Character Mappings We list all of the Maltese-to-Arabic character mappings we consider in Table 2. Most are letter-to-letter mappings such as *k* to ك *k*, but also include the Maltese multi-character letters *ie* and *għ*. The **Basic** column indicates the most expected letter mapping based on our observations considering etymology, phonology, and Arabic letter frequencies. The additional columns in the table indicate conditional mappings as well as non-deterministic additional mappings. For vowels, we include word-initial and word-final conditional mappings; and for consonants, the second of doubled letters may be mapped to a Shadda (Arabic gemination diacritic). All Arabic diacritics are deleted after the mapping step since they are often absent in the language model training data (Habash, 2010). We use the character mappings in two ways: (a) **deterministic** mappings using only the Basic column and its associated Doubling column in Table 2, and (b) **non-deterministic** mappings using all the columns in Table 2. Our deterministic mapping does not apply context specific word-initial and word-final rules.

[1] https://mlrs.research.um.edu.mt/

Vowels								
Maltese	Basic		Additional		Word Initial		Word Final	
ie	ا	*A*						
a	◌َ	*a*	ا	*A*	ا، آ	*A* , |	ا، ى، ة، ه	*A* , *Y* , *p* , *h*
e	◌َ	*a*	ا	*A*	ا	*A*	ي	*y*
i	◌ِ	*i*	ي	*y*	ا	*A*	ي	*y*
o	◌ُ	*u*	و	*w*	ا	*A*	و	*w*
u	◌ُ	*u*	و	*w*	ا	*A*	و، وا، ه	*w* , *wA* , *h*

Consonants						
Maltese	Basic		Additional		Doubling	
għ	ع	*E*	غ	*g*		
'	ع	*E*				
b	ب	*b*			◌ّ	~
ċ	تش	*t$*			◌ّ	~
c	ك	*k*			◌ّ	~
d	د	*d*	ذ، ض، ظ	**, D, Z*	◌ّ	~
f	ف	*f*			◌ّ	~
ġ	ج	*j*			◌ّ	~
g	ج	*j*			◌ّ	~
h	ه	*h*			◌ّ	~
ħ	ح	*H*	خ	*x*	◌ّ	~
j	ي	*y*			◌ّ	~
k	ك	*k*			◌ّ	~
l	ل	*l*			◌ّ	~
m	م	*m*			◌ّ	~
n	ن	*n*			◌ّ	~
p	ب	*b*			◌ّ	~
q	ق	*q*			◌ّ	~
r	ر	*r*			◌ّ	~
s	س	*s*	ص	*S*	◌ّ	~
t	ت	*t*	ط	*T*	◌ّ	~
v	ف	*f*			◌ّ	~
w	و	*w*			◌ّ	~
x	ش	*$*			◌ّ	~
y	ي	*y*			◌ّ	~
ż	ز	*z*			◌ّ	~
z	دز	*dz*			◌ّ	~

Table 2: Maltese to Arabic Character Mappings (shaded regions are romanizations in the Buckwalter scheme (Habash et al., 2007)). Additional, Word Initial, and Word Final are the conditional alternatives considered for the non-deterministic mappings.

Token Mappings We augment the character-level mappings with **closed class** token-level mappings that exploit known features of Maltese orthography such as the spelling of the definite determiner, as well as the Zipfian head distribution of closed class words which we expect to help transfer learning in downstream tasks. In the settings where they are used, the token mappings take precedence over the character mappings since they match a token exactly.

We extracted all the closed class tokens from the training set portion of the MLRS POS data (Gatt and Čéplö, 2013), by filtering over the part-of-speech tag. For each of these tokens, a manual transliteration is performed by native Arabic speakers following a consistent interpretation of the Conventional Orthography for Dialectal Arabic (CODA) guidelines (Habash et al., 2018). To facilitate the interpretation of the token, annotators were provided with the POS tag, the IPA transcription (extracted using a grapheme-to-phoneme system by Borg et al. (2014)), and a sample sentence where the token is used. Native Maltese speakers were consulted for ambiguous cases.

In total, we have 691 mappings (henceforth, **Full** closed-class).[2] Examples include *fuq* `PREP` 'over', *ftit* `QUAN` 'some', *kellu* `VERB_PSEU` 'he had', and *mhux* `PRON_PERS_NEG` 'he is not', which map to فوق *fwq*, فتيت *ftyt*, كان له *kAn lh*, and ماهوش *mAhw$*, respectively. Additionally, we consider a subset of 135 mappings (henceforth, **Small** closed-class) which we restrict to the tokens containing - and/or ', such as *il-* 'the', *fis-* 'in the', and *t'* 'of', which map to ال *Al*, فال *fAl*, and تاع *tAE*, respectively. We designate not using the token mappings as **None**.

3.3 Ranking Techniques

While token mappings are essentially deterministic, character mappings produce a large lattice of combinations, e.g. Maltese *ħielsa* 'free' results in 20 forms, including خالصة *xAlSp*, حالصى *HAlSY*, خالسا *xAlsA*, حالصه *HAlSh*, and حلس *Hls*. In this section we present various ranking techniques used to select one of the alternatives for a given input.

Deterministic The use of deterministic character mappings yields a single alternative for each word, thereby not requiring any ranking.

Random A random choice is made by selecting the first alphabetically sorted token from the list of combinations. We sort to keep this technique stable across different runs.

Sub-Token Count The BERT-based language models used in the downstream task may split a given word into multiple sub-tokens (Devlin et al., 2019). We choose the mapping combination with fewer resulting sub-tokens. This idea is based on the evidence that a tokenizer that splits tokens into fewer sub-tokens correlates with better downstream performance (Rust et al., 2021).

[2] We do not use the part-of-speech information in the actual mapping process. Although there are a number of tokens that appear with different part-of-speech tags, only one of these tokens (*m'*) resulted in different transliterations: `NEG` ما *mA*, `COMP` من *mn*, and `PREP` مع *mE*. Since the `NEG` reading is the most common in our data by far, we ignore the other two.

N-gram Language Model Scores We use the word and character n-gram language models from Baimukan et al. (2022) to get **word-level** and **character-level** scores on each generated token, respectively. As highlighted in Section 2.1, due to the similarity with Maltese, we consider both the country-level Tunisian (TUN) and region-level Maghrebi (MAG) models, ending up with two sets of scores.

Some of these scores are bound to produce ties, occasionally ranking more than one token in first place. In our analysis, the word n-gram model tends to produce ties whenever the word is out of vocabulary while the sub-token count model is much more sporadic with ties. We observed that the character n-gram score almost never produced ties on the data, and so we used it as a fallback to resolve ties. Ties can further be resolved randomly if need be.

3.4 Implementation

The various mapping settings we consider in the rest of the paper select for a token-mapping setup and a ranking setup. Conceptually, putting together all of these components results in a pipeline where a token is mapped using the closed-class token mappings, backing off to the character mappings whenever the token is not found in the token mappings. This is followed by a ranking step which selects among the various options produced by the mapping component.

We implement all mappings from Section 3.2 using finite-state machinery in Pynini (Gorman, 2016). The seven basic FSTs we implement are the following: full token mappings, small token mappings, non-deterministic and deterministic multi-character mappings,[3] non-deterministic and deterministic single character mappings, and dediacritization mappings. These are then composed in succession on the fly, based on the experimental setup being used.

The generated alternatives are ranked as described in Section 3.3. The sub-token count metric is implemented using the Transformers library (Wolf et al., 2020) while the n-gram language model scores are obtained using KenLM (Heafield, 2011).

We make the code publicly available.[4]

4 Downstream Task Evaluation

The transliteration system is evaluated on three downstream tasks: Part-of-Speech Tagging (**XPOS**),[5] Dependency Parsing (**DP**), and Sentiment Analysis (**SA**). Input tokens in the datasets are transliterated as discussed in Section 3, with their corresponding labels/tags remaining unchanged. Further details on the dastaset sources and processing is given in Section 4.1.

We consider three setups of token mappings, all of which use the character mappings as described in Section 3.2: the entire set of the full closed-class mappings (**Full**), the small closed-class mappings (**Small**), and no token mappings (**None**). For each of these setups, we use the ranking techniques from Section 3.3. This creates 24 distinct transliteration pipelines (3 mapping options by 8 ranking techniques), which we explore in this Section. Every dataset is transliterated through each of these pipelines, which are then used to fine-tune the language model following the setup used by Micallef et al. (2022). Each fine-tuned model is evaluated on the corresponding transliterated test set.

We systematically compare the pipelines on CAMeLBERT-Mix (Inoue et al., 2021) due to its training on dialectal data. We also fine-tune BERTu, a monolingual Maltese model (Micallef et al., 2022), and multilingual BERT (mBERT) (Devlin et al., 2019) on the datasets in the original script (untransliterated). Additionally, we consider another setup for mBERT where it is fine-tuned on transliterated Maltese. We report accuracy for XPOS Tagging, Labelled Attachment Score (LAS) for DP, and macro-averaged F1 for SA.

4.1 Datasets

We use the MUDT (Čéplö, 2018) dataset as is for the DP task. For the XPOS task, we use the MLRS POS dataset (Gatt and Čéplö, 2013), but with different splits from Micallef et al. (2022) to ensure that the instances overlapping with the MUDT data are in the same splits.

The SA dataset used (Martínez-García et al., 2021) is preprocessed and tokenized using the MLRS tokenizer. Although this task involves classifying a whole sentence, this preprocessing is done because the transliteration system operates on tokens rather than sentences. Once each token is transliterated these are joined back as a single text,

[3] This includes multi-character letters (*ie* and *għ*) along with geminates (**Doubling**), and **Word Initial** and **Word Final** vowels from Table 2.

[4] `https://github.com/MLRS/malti_arabi_fst`

[5] XPOS refers to the language-specific tagset as opposed to UPOS, the universal tagset (Nivre et al., 2017).

Task	Dataset	Training	Validation	Testing
XPOS	MLRS POS	4,935	616	616
DP	MUDT	1,123	518	433
SA	Sentiment	595	85	171

Table 3: Dataset sizes in terms of sentences

separated by spaces. Admittedly, this results in different spacing compared to the source sentence, particularly for tokens with determiners and punctuation symbols in general. However, we fine-tune the baselines which use the original Latin script with this same pre-processing strategy.

The tokens in the MUDT and MLRS POS datasets are kept as is since these are consistent with the MLRS tokenizer. A summary of the dataset sizes is given in Table 3. To address the discrepancy in the data sizes, we also consider a lower-resourced setup where the training and validation (but not test) sets of each tasks are reduced to the smallest dataset size used in this evaluation (SA). This allows us to control for size when analysing the cross-lingual transfer capabilities.

4.2 Results

The results shown in Table 4 highlight that a combination of the full closed-class mappings and any of the non-deterministic systems achieve the best performance across all tasks. As expected, the deterministic system without any token mappings performs the worse, generally. Analysing the token and character mappings as different dimensions, reveals some interesting trends.

Token Mappings The inclusion of closed-class mappings consistently yields improvements over using no token mappings in all scenarios. In the deterministic case, the full token mappings are beneficial to surpass the non-deterministic counterpart with random ranking, in the DP and SA tasks, and are competitive against the other non-deterministic non-random scores in all tasks. The small closed-class mappings also generally improve over using no token mappings, although this is slightly detrimental in a few cases. Inspecting further the relationship between the full, small, and no token mappings, it is evident that the jumps in performance are more pronounced in the lower-resourced setup compared to the whole data setup. These findings indicate that while linguistic annotations are generally helpful, they are most useful in setups when data is scarcer.

Ranking Techniques The random ranking performs the worst of all the techniques considered for the non-deterministic character mappings but does better than the deterministic counterpart in cases where no additional tokens are present. All techniques, apart from random, perform comparably, with the word language model ranker achieving the best scores on the syntactic tasks while the character model and sub-token rankers give the best results in the semantic task.

Ranking with the Tunisian word model scores yielded the best result in 3 out of 5 task-data setups. This is likely due to the high degree of mutual intelligibility between Maltese and Tunisian Arabic as detailed in Section 2.1. Moreover, this ranking tends to give significant boosts in performance just by using the small token mappings as evidenced by the similar results obtained by the system with the full token mappings. Conversely, the Maghrebi models tend to give worse scores without any token mappings and gave a worse result than the deterministic system in one particular case in SA.

Pre-trained LMs and Transliteration Comparing the best result from each task and data size setup from Table 4 against the baselines shown in Table 5, it is evident that mBERT fine-tuned without transliteration is only better than the best transliteration pipeline in XPOS when the entire data is used. For SA and lower-resourced DP, the difference between mBERT and the best transliteration pipeline on CAMeLBERT is found to be statistically significant, using a 1-tailed t-test with a p-value of < 0.05.

In Table 5, we compare transliteration with the Tunisian word model ranking with full token mappings. It is clear that fine-tuning BERTu with the original (untransliterated) data yields the best performance overall, owing to the Maltese corpora that this model is pre-trained on. Inspecting the results obtained for mBERT, transliteration does not always improve performance compared to untransliterated fine-tuning, and can result in significant degradations as evidenced in the SA task. Since, this is counter to what Muller et al. (2021) reported, it could be attributed to the hybrid nature of Maltese. However, we posit that this is also due to the fact the mBERT was solely pre-trained on MSA. Conversely, CAMeLBERT-Mix was trained on 5.8 billion DA data, making up around a third of the entire pre-training corpus (Inoue et al., 2021). In fact, when fine-tuning with transliterated Mal-

		XPOS						DP						SA		
		Large Training			Small Training			Large Training			Small Training			Small Training		
LM	Mapping	None	Small	Full	None	Small	Full	None	Small	Full	None	Small	Full	None	Small	Full
N/A	Deterministic	94.4	94.2	95.2	89.8	89.7	91.5	68.8	69.7	76.4	63.4	64.3	70.9	62.7	61.2	67.0
	Random	95.6	95.6	95.9	90.7	91.0	91.8	74.0	75.1	76.0	67.2	69.0	70.3	64.3	64.2	64.7
TUN	Char Model	95.5	95.7	95.9	91.1	91.7	92.4	75.1	76.1	77.3	69.5	70.5	**72.4**	64.5	63.9	65.5
	Word Model	95.6	96.0	**96.1**	91.3	92.5	**92.6**	75.1	76.5	77.3	68.8	71.5	**72.4**	64.0	65.9	66.7
	Sub-Tokens	95.6	95.6	95.9	91.0	91.5	92.3	75.2	76.4	**77.5**	69.0	70.7	72.0	66.3	67.6	67.3
MAG	Char Model	95.5	95.5	95.9	90.7	91.3	92.4	74.9	75.6	76.7	68.6	69.5	71.7	65.1	65.3	**69.7**
	Word Model	95.3	95.7	95.7	91.2	92.2	92.4	75.2	76.7	77.0	68.8	70.8	71.8	62.5	62.4	68.7
	Sub-Tokens	95.5	95.5	95.9	90.8	91.4	92.2	74.9	76.4	77.4	68.8	70.2	71.9	66.1	65.4	69.4

Table 4: Test set results for CAMeLBERT-Mix fine-tuned on transliterated Maltese, grouped by token and character mappings. The ranking techniques are further grouped by the language model used by the primary and/or fall-back technique: no language model (N/A), Tunisian (TUN), Maghrebi (MAG). Each value is an average of 5 runs with different random seeds. The best score in a task is **bolded**, while the best results per token mappings are underlined. Color shading is done with respect to the best and worst values of each task and training size setup.

		XPOS		DP		SA
Script	Model	Large	Small	Large	Small	Small
Arabic	CAMeLBERT	96.1	92.6	77.3	72.4	66.7
Arabic	mBERT	95.9	92.1	77.7	72.0	61.6
Latin	mBERT	96.7	92.4	77.3	71.1	67.3
Latin	BERTu	98.3	97.4	88.1	86.3	83.1

Table 5: Comparison of fine-tuning on raw and transliterated Maltese using different language models. The transliteration pipeline used is the Tunisian Word Model Ranking with Full token mappings. Large and Small refers to Large Training and Small Training set ups as in Table 4.

tese, CAMeLBERT performs better than mBERT in most task-data setups. CAMeLBERT is also able to surpass or be very competitive with mBERT fine-tuned on raw Maltese. This finding gives further evidence that there is some level of mutual intelligibility between transliterated Maltese and dialectal Arabic. Moreover, making use of monolingual models should be considered for cross-lingual transfer, whenever this is available.

5 Human Readability Evaluation

In this section, we investigate how readable transliterated Maltese is to native Arabic speakers. We compare four settings: the deterministic system against the non-deterministic system (Tunisian Arabic Word Model) with both None and Full token mappings. We sample 50 instances from the MUDT training set (Čéplö, 2018). For each example, we provide the original sentence, a translation extracted from Google Translate as a reference, and each of the alternative transliterations (see example in Table 6). We hide the transliteration system information and shuffle the order in which each alternative is displayed to prevent biases. For this study we ask the evaluator, a native speaker of Tunisian Arabic, who is fluent in French and familiar with Italian, to rank the transliterations in the order of how readable the text is, where 1 is best.

Table 7 shows the average readability rank for different combinations. The results show that using the Word Model is better than the Deterministic model, and that using the token mappings is helpful. These results correlate with our empirical evaluation from Section 4.

The evaluator reported that reading Maltese written in the Arabic script allowed them to easily recognize shared words between Maltese and Tunisian Arabic. For instance, the evaluator did not recognize the Maltese word *kien*, but when transliterated into the Arabic script as كان ‘he was’, it was evident. However, Italian-origin words presented a reading challenge, e.g. Maltese *akkuża* (Italian *accusa*) was garbled. The evaluator also pointed out that none of the transliteration models were capable of handling

	Text	Rank
Maltese	Illum waranofsinhar il-Maġistrat Miriam Hayman iddikjarat li hemm biżżejjed provi biex il-ħames aħwa tan-negozjant George Farrugia jitqiegħdu taħt att ta' akkuża.	
English	This afternoon Magistrate Miriam Hayman has stated that there is enough evidence to put the five brothers of businessman George Farrugia under indictment.	
Word Model + Full CC	الم ورنفسنهر ال ماجسترات مريم هيمان اذكيارات اللي هم بالزايد بروفي باش ال خمس اخوه تاع ال نجودزينت جورجي فرجة يتقاعده تحت ات تاع ا كزة .	**1**
Word Model + None	الم ورنفسنهر ال ماجسترات مريم هيمان اذكيارات لي هم بزايد بروفي باش ال حماس اخوه تن نجودزينت جورجي فرجة يتقاعده تحت ات تاع ا كزة .	2
Deterministic + Full CC	لم ورنفسنهر ال مجسترت مرم هيمن دكيرت اللي هم بالزايد برف باش ال خمس حو تاع ال نجدزينت جرج فرج يتقاعد تحت ت تاع كز.	3
Deterministic + None	لم ورنفسنهر ل- مجسترت مرم هيمن دكيرت ل هم بزيد برف باش ل- حمس حو تن- نجدزينت جرج فرج يتقاعد تحت ت تع كز.	4

Table 6: An example of a Maltese sentence along with its English translation and the output of the four transliteration models ranked by their readability level.

Mapping	Average Rank
Word Model + Full	1.1
Deterministic + Full	2.3
Word Model + None	2.5
Deterministic + None	4.0

Table 7: Average Readability Rank

Maltese univerbations, such as *waranofsinhar* 'afternoon' (see Table 6), which made it challenging to recognize and read them accurately.

This experiment highlighted some of the many challenges in reading Maltese written in Arabic script and provided insights into the limitations of different transliteration models, and issues to consider addressing in the future.

6 Conclusion and Future Work

We presented a Maltese-to-Arabic transliteration system as a tool to leverage cross-lingual transfer from Arabic. As evidenced by our empirical results, a non-deterministic system with signals from the target language helps in choosing a better transliteration alternative, especially in ambiguous cases. Moreover, incorporating human-annotated transliterations of a set of closed-class of words is beneficial in downstream performance, especially in lower-resource settings.

Our experimental setup exploited an Arabic language model for cross-lingual transfer, instead of a multilingual model such as mBERT. Results show promising results, giving better performance than multilingual models. This echoes the findings by Wu and Dredze (2020), and we encourage further research to investigate ways to effectively leverage resources from linguistically related languages.

Future work should investigate the use of large sentence-level contexts in mapping selections. Exploring cross-lingual transfer from Italian and English is also an interesting direction, including the use of few-shot and zero-shot learning. It would be interesting to also investigate these cross-lingual transfer techniques through transliterated Arabic.

Limitations

In this work, we transliterate all Maltese words in the same manner. Given the hybrid nature of Maltese, it might be optimal to handle words which do not have an Arabic origin in a different way. Similarly, we do not treat named-entities any differently.

Moreover, we assume that the Maltese text is written using the standard orthographic rules. In turn, the system might produce spurious transliterations for cases with spelling errors. This issue also exists when the text is in raw form, but may be further exacerbated with transliteration. The character mappings could be expanded to handle dropped Maltese diacritics, such as writing *c* instead of *ċ*, but there are other cases where silent letters such as *għ* are dropped altogether, making the problem non-trivial.

Acknowledgements

We acknowledge LT-Bridge Project (GA 952194) and DFKI for access to the Virtual Laboratory. We further acknowledge funding by Malta Enterprise.

References

Joseph Aquilina. 1987. *Maltese-English Dictionary Vol. I, A-L.* Midsea Books, Valletta, Malta.

Joseph Aquilina. 1990. *Maltese-English Dictionary Vol. II, M-Z.* Midsea Books, Valletta, Malta.

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical aggregation of dialectal data for Arabic dialect identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4586–4596, Marseille, France. European Language Resources Association.

Albert Borg and Marie Azzopardi-Alexander. 1997. *Maltese: Descriptive Grammars*. Routledge, London and New York.

Mark Borg, Keith Bugeja, Colin Vella, Gordon Mangion, and Carmel Gafà. 2014. Preparation of a free-running text corpus for Maltese concatenative speech synthesis. *Perspectives on Maltese Linguistics, Studia Typologica*, 14:297–318. Berlin:Akademie Verlag.

Slavomír Čéplö. 2018. *Constituent order in Maltese: A quantitative analysis*. Ph.D. thesis, Charles University, Prague.

Slavomír Čéplö, Ján Bátora, Adam Benkato, Jiří Milička, Christophe Pereira, and Petr Zemánek. 2016. Mutual Intelligibility of Spoken Maltese, Libyan Arabic, and Tunisian Arabic Functionally Tested: A Pilot Study. *Folia Linguistica*, 50(2):583–628.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. 25 edition. SIL International. [link].

Ray Fabri, Michael Gasser, Nizar Habash, George Kiraz, and Shuly Wintner. 2014. *Linguistic Introduction: The Orthography, Morphology and Syntax of Semitic Languages*, pages 3–41. Springer Berlin Heidelberg, Berlin, Heidelberg.

Charles A. Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Albert Gatt and Slavomír Čéplö. 2013. Digital Corpora and Other Electronic Resources for Maltese. In *Proceedings of the International Conference on Corpus Linguistics*, pages 96–97. UCREL, Lancaster, UK.

Kyle Gorman. 2016. Pynini: A Python library for weighted finite-state grammar compilation. In *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*, pages 75–80, Berlin, Germany. 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).

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4237–4246, Marseille, France. European Language Resources Association.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Antonio Martínez-García, Toni Badia, and Jeremy Barnes. 2021. Evaluating morphological typology in zero-shot cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3136–3153, Online. Association for Computational Linguistics.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 90–101, Hybrid. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Mike Rosner and Claudia Borg. 2022. *Report on the Maltese Language*. Language Technology Support of Europe's Languages in 2020/2021. Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne (Series Editors). Available online at https://european-language-equality.eu/deliverables/.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Edmund F. Sutcliffe. 1936. *A grammar of the Maltese language, with chrestomathy and vocabulary*. Oxford University Press: Humphrey Milford, London.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal dependencies for Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, Valencia, Spain.

Mary Ann Walter. 2006. Pharyngealization effects in Maltese Arabic. In *Perspectives on Arabic Linguistics: Papers from the annual symposium on Arabic linguistics. Volume XVI:, Cambridge, March 2002*, volume 266, pages 161–178. John Benjamins Publishing.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Belguith, and Nizar Habash. 2014. A conventional orthography for Tunisian Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2355–2361, Reykjavik, Iceland. European Language Resources Association (ELRA).