Putting Figures on Influences on Moroccan Darija from Arabic, French and Spanish using the WordNet

Khalil Mrini
Ecole Polytechnique Fédérale de Lausanne
Switzerland
khalil.mrini@epfl.ch

Francis Bond
Nanyang Technological University
Singapore
bond@ieee.org

Abstract

Moroccan Darija is a variant of Arabic with many influences. Using the Open Multilingual WordNet (OMW), we compare the lemmas in the Moroccan Darija Wordnet (MDW) with the standard Arabic, French and Spanish ones. We then compared the lemmas in each synset with their translation equivalents. Transliteration is used to bridge alphabet differences and match lemmas in the closest phonological way. The results put figures on the similarity Moroccan Darija has with Arabic, French and Spanish: respectively 42.0%, 2.8% and 2.2%.

1 Introduction

Locally known as Darija and referred to as a dialect, the Moroccan variant of the Arabic language is spoken by the overwhelming majority of Moroccans (HCP, 2014) with small regional differences. The Moroccan Darija Wordnet (MDW) (Mrini and Bond, 2017) was released as part of the Open Multilingual WordNet (OMW) (Bond and Foster, 2013), thereby linking all the languages in the OMW to Moroccan Darija.

Morocco has a complex language situation. Its two official languages are Arabic, the basis of Moroccan Darija, and, since 2011, Tamazight. The North African Kingdom has gained its independence in 1956 from colonial France and Spain, and both countries have had linguistic influence on Moroccan Darija through loanwords.

Moroccan Darija is used in day-to-day informal communication (Ennaji, 2005) and doesn't have the prestige associated with Arabic or French, which are the languages used in education. In the 2014 census (HCP, 2014), it was reported that Morocco's literacy rate is at 67.8%, making it one of the lowest in the Arab World. A 2010 study (Magin, 2010) found that although the reasons of high illiteracy rates in the Arab World are varied and subject to controversy, one of them was the "*disconnect between high Arabic used as the medium of instruction in schools and the various dialects of Arabic spoken in Arab region*".

This paper aims at putting figures on the influences of the Arabic, French and Spanish languages on Moroccan Darija. Accordingly, on top of the MDW, the Arabic (Black et al., 2006; Abouenour et al., 2013), French (Sagot and Fišer, 2008) and Spanish (Gonzalez-Agirre et al., 2012) wordnets were used.

To gauge the influence of these languages on Moroccan Darija, the word distance between each Moroccan lemma and its corresponding lemma in the other language was computed. This is done using the sense-based property of the WordNet. Transliteration helped to bridge the difference between the Arabic and Latin alphabets and the difference of use of the Latin alphabet between the European languages studied. The language-dialect similarities were computed for both the automatically linked Moroccan synsets, as well as the manually validated ones.

2 Related Work

In this section, we describe relevant aspects of the MDW. Then, we provide a review of studies on language-dialect and dialect-dialect similarity, as well as a review of methods to compute word-to-word similarity.

2.1 The Moroccan Darija Wordnet

The Moroccan Darija Wordnet (Mrini and Bond, 2017) was developed using an *expand* approach with the vocabulary being extracted from a bilingual Moroccan-English dictionary (Harrell, 1963). There were 12,224 Moroccan synsets automatically connected to the Princeton WordNet (Fellbaum, 1998), with 2,319 of them being manually verified.

During the process of developing the MDW, a Latin-based Moroccan alphabet was set, using as basis the one used in the bilingual dictionary, as well as the colloquial alphabet used in daily written communication between Moroccans. The MDW alphabet assigns one sound to one letter, and this facilitates transliteration to other languages.

2.2 Linguistic Similarity and Dialects

A similar case to Moroccan Darija is the Maltese language. Brincat (2005) recounts that it was first considered an Arabic dialect, but an etymological analysis of the 41,000 words of a bilingual Maltese-English dictionary shows that 32.41% of them are of Arabic origin, 52.46% of Sicilian or Italian orgin and 6.12% of English origin. This heterogeneous mix is probably one of the reasons that Maltese is the only colloquial Arabic dialect that emancipated to become a full-fledged language (Mallette, 2011). Aquilina (1972) established a wordlist of Maltese Christian words of Arabic origin, as well as an earlier detailed etymological study comparing Maltese and Arabic (Aquilina, 1958).

Scherrer (2012) proposes a simple metric to measure the similarity between different dialects of Swiss German in a corpus. It is based on the Levenshtein distance (Levenshtein, 1996), also known as the edit distance. The latter is a string metric measuring the difference between two sequences. It represents the minimum number of characters to modify to make both sequences identical. Those modifications can be single-character insertions, substitutions or deletions. Heeringa et al. (2006) propose an evaluation of distance measurement algorithms for dialectelogy, in which they use a normalised version of the edit distance so that it is comprised between 0 and 1. Inkpen et al. (2005) propose normalising the Levenshtein distance by dividing by the length of the longest string. That is the method that we will be using to compute word-to-word similarity.

3 Estimating Influences

To compare Moroccan Darija with Arabic, French and Spanish, we will perform word-to-word comparisons. These comparisons must be phonologically accurate despite alphabet differences.

3.1 Computing Word Distance

The similarity is assessed through looking at lemmas across different languages that share the same sense. That means that in a given OMW synset with one or more Moroccan lemmas, the latter will be compared to the synset's Arabic, French and Spanish lemmas.

Spaces and underscores in multiword expressions would count as letters, and there may be word order differences, thus resulting in inaccuracies. Therefore, multiword expressions were excluded from the comparisons.

We compare the three languages first to all the Moroccan synsets that were automatically linked in Mrini and Bond (2017), and then only to the ones manually validated and included in the first release of the MDW.

3.2 Transliteration of the MDW Alphabet

We can compute the normalised distance between two words, but we also need to bridge the difference of alphabets between the languages studied. We do this through a process of transliteration. The transliteration used from Moroccan Darija was specific to each language to which it was compared and proposed phonological correspondences. The purpose of these phonological correspondences is to be able to recognise what words are cognates or borrowings. It is at this point difficult to distinguish if a pair of similar lemmas are cognates or the result of a borrowing.

The transliteration process is complicated, as, if it is strict, its accuracy may be low. This is why numerous options were considered for the transliteration of each Moroccan letter. This way, all the possible transliterations of a word are considered for comparison. The number of possible transliterations for a lemma is the product of the lengths of each set of possible transliterations of each letter contained in the lemma. The flexibility of the transliteration process is optimistic, as the smallest distance resulting from any transliteration option is the one considered for computing the overall cross-lingual average distance.

3.2.1 Transliteration to Arabic

The Arabic WordNet (Black et al., 2006) has words in the Arabic alphabet with irregular diacritics, meaning that a short vowel in a word may or may not be illustrated by diacritics. Each diacritic is considered to be a separate character in the string that represents the Arabic lemma. Therefore, two

	Transliterations		
Darija	Arabic	French	Spanish
a	ا, ى, ة, ϕ	a	a, á
b, ḅ	ب	b, p, v	b, p, v
d	د, ض, ظ, ذ	d	d
ḍ	ض, ظ	d	d
e	ϕ	e, é, è, ê	e, é
ă	ا, ϕ	a, e, é, è, ê	a, e, é
f	ف	f, ph	f
g	گ, ق	g	g
8	غ	r	r
h	ه	h	h
7	ح	h, ϕ	h, j, ϕ
i	ي, ىء, ϕ	i	i, í
ĭ	ي, ϕ	i	i, í
j	ج	j	y
k	ک	k, c	k, c
l, ḷ	ل	l	l
m, ṃ	م	m	m
n	ن	n	n
o	و, وء, ϕ	o	o, ó
q	ق	q, k, c	q, k, c
r, ṛ	ر	r	r
s	س, ص	s	s, c, z
ṣ	ص	s	s, c, z
š	ش	ch	ch
t	ت, ط, ث, ظ	t	t
ṭ	ط, ظ	t	t
u	و, وء, ϕ	ou, u	u, ú
w	و, وء, ϕ	w, ou	u, ú
x	خ	kh	j
y	ي, ϕ	y	y, i, í, ll, ϕ
z, ẓ	ز	z	z
2	ϕ	ϕ	ϕ
3	ع	a, ϕ	a, ϕ

Table 1: Transliteration from Moroccan Darija to Arabic, French and Spanish

transliterations were necessary. On the one hand, the diacritics on Arabic WordNet lemmas were erased. On the other hand, each Moroccan character was transliterated as per Table 1. The emphatic Arabic characters were included in both emphatic and dotless *b*'s, *d*'s, *s*'s and *t*'s, but non-emphatic Arabic characters were not included in the possible transliterations of *ḅ*'s, *ḍ*'s, *ṣ*'s and *ṭ*'s. Diacritics having been removed, transliteration of Moroccan vowels to short vowels was represented by the possibility of removing them (ϕ).

3.2.2 Transliteration to French and Spanish

The transliterations of Moroccan Darija to lemmas of the French (Sagot and Fišer, 2008) and Spanish (Gonzalez-Agirre et al., 2012) wordnets were also made to be as flexible as possible. In French, all accents on *e*'s were considered. Likewise, the accents used for stressed syllables in Spanish were considered for all vowels. For both languages, the Moroccan *b* could be transliterated as either *b*, *p* or *v*, as there is a near-absolute absence of *p* and *v* in Moroccan Darija. The Spanish pronunciation of *ll*, *z* and *j* differs from the French one and therefore they were mapped to different letters in the Moroccan alphabet. Furthermore, some Moroccan letters were matched to two French letters, as the French pronunciations of *ou* and *ph* respectively match the Moroccan *u* and *f*, and Morocco's official transliteration of the */x/* sound is *kh*. Like the Arabic Transliteration code above, each character was transliterated individually. The transliterations to French and Spanish are also in Table 1.

4 Results and Discussion

The aggregated results of the word-to-word comparisons gave an estimation of the linguistic influences on Moroccan Darija. We obtained results for the Moroccan synsets that were automatically linked and the ones that were manually validated.

Each of the Arabic, French and Spanish wordnets had a certain number of links to Moroccan synsets for which they had at least one available single-word lemma. To these synsets, a certain number of lemmas were associated. In both comparisons, the Moroccan lemmas were matched for pairwise comparisons. Based on that number of synsets matched, an average normalised Levenshtein distance was given for both languages. Examining the results, we decided that synsets should only be counted as a match if they had at least 60% similarity.

Comparison with:	Arabic	French	Spanish
Number of links to Moroccan synsets	7,958	11,605	10,167
– excluding synsets with only multi-word expressions	6,702	9,954	8,612
Average normalised Levenshtein distance	0.4619	0.7337	0.7521
Number of synsets with one or more word pairs at least 60% similar	2,816	278	188
Percentage of synsets with one or more word pairs at least 60% similar	42.02%	2.79%	2.18%

Table 2: Results of the comparisons of automatically linked synsets of Moroccan Darija with Arabic, French, Spanish

4.1 Comparison based on Automatically Linked Moroccan Synsets

The results of the comparisons of the automatically linked Moroccan synsets with each language are given in Table 2.

4.1.1 Cross-lingual Similarity Scores

The results show that, on average, a Moroccan Darija word is 53.81% similar to its Arabic translation, 26.63% similar to its French translation and 24.79% similar to its Spanish translation, the similarity being 1 minus the distance. The similarity method used is akin to related work on semantic similarity (Ciobanu and Dinu, 2014). The average normalised distance was computed by averaging the lowest normalised Levenshtein distance found in any lemma pair in each comparison of a Moroccan synset to the WordNet synset matches, with all Moroccan synsets having equal weights in the average.

If the confidence scores were used as weights in the average normalised Levenshtein distance, then Moroccan Darija would be on average 52.99% similar to Arabic, 24.02% similar to French and 22.25% similar to Spanish. Some of the similarities may be random, this is why a threshold must be empirically established, such that word pairs which similarity has crossed the threshold are visibly similar. On establishing a threshold of 60% similarity, the similarity numbers dwindle faster for French and Spanish than for Arabic.

4.1.2 Similarity with Arabic

Moroccan Darija and Arabic share an average normalised Levenshtein distance of around 0.4619. This number puts a figure on the similarity between Moroccan Darija and Arabic.

For comparison, the same method of comparison can be applied to other pairs of languages. This way, it can be determined that Portuguese (de Paiva et al., 2012) and Galician (Gonzalez-Agirre et al., 2012) are the closest case to Moroccan Darija and Arabic with an average Levenshtein distance of 0.4760. The former two languages are considered independent languages. These comparisons show how blurry the line is between a dialect or variant and an independent language, especially within the continuum of Arabic dialects (Greene, 2013). From these results, Moroccan Darija can be seen as distinct enough from Arabic to possibly be considered a language of its own.

4.1.3 Similarity with French and Spanish

Out of the 278 synsets that were more than 60% similar to Moroccan Darija for French and the 188 ones for Spanish, there were 95 common synsets. Therefore, some non-negligible part of the similarity of French and Spanish with Moroccan Darija is due to the similarity between French and Spanish. Future work would allow to distinguish the linguistic influence represented by each of these common synsets.

4.1.4 Moroccan Lemmas of Unknown Origin

Taking the Moroccan synsets connected to the Arabic, French and Spanish wordnets, the lemmas that were among any of the lists of word pairs that were more than 60% similar were eliminated. Therefore, this resulted in a set of 2,736 Moroccan synsets of unknown origin. Among these, there are words of Arabic origin such as "*deqq*" (from the Arabic verb for "*to block*") and "*nzel*" (from the Arabic verb for "*to go down*"). Some words are of French or Spanish origin such as "*serbisa*" (from the Spanish noun for "*beer*"). These were probably due to errors in linking the Moroccan synsets to the WordNet.

A sizeable proportion is of Tamazight origin,

	Average distance			At least 60% similarity		
Comparison with	Arabic	French	Spanish	Arabic	French	Spanish
The 12,224 synsets that form the total	0.4619	0.7337	0.7521	42.02%	2.79%	2.18%
The 617 manually validated synsets	0.4393	0.7544	0.7721	47.00%	3.08%	2.92%

Table 3: Comparison of average Levenshtein distances between different sets of synsets and comparison of percentage of number of synsets that are at least 60% similar between different sets of synsets

such as "*degdeg*" (from the Tamazight verb for "*to smash*") and "*seqsi*" (from the Tamazight verb for "*to ask*"). The influence of Tamazight is very visible on the words that start with "*ta-*" and end in "-*t*" such as "*tazellajt*" and "*tabennayet*". The study of the Tamazight influence will most likely require the creation of a Tamazight WordNet.

4.2 Comparison based on Manually Validated Moroccan Synsets

In order to investigate the effect of linking errors, we perform the same comparison on the 2,319 manually verified synsets contained in the current release of the MDW. Then we filtered them to obtain the synsets with links to each of the Arabic, French and Spanish wordnets. Therefore this set used for validation contains 617 Moroccan synsets.

The average Levenshtein distances and the percentages of synsets that are at least 60% similar are in Table 3. The difference in figures between the manually validated Moroccan synsets and the automatically linked ones proved small enough to say that the linking noise was not an issue.

5 Summary

In this paper, we attempted to put figures on the similarity between Moroccan Darija and each of Arabic, French and Spanish.

Transliteration was used to bridge the alphabet gap and perform phonological comparisons. The methods used were flexible and the comparisons exploited all possible transliterations for each letter. Transliteration was one-way from the Moroccan Darija Wordnet (Mrini and Bond, 2017) for the French (Sagot and Fišer, 2008) and Spanish (Gonzalez-Agirre et al., 2012) wordnets, but was both ways for the comparison with the Arabic WordNet (Black et al., 2006). The word-to-word distance was computed using Levenshtein distance (Levenshtein, 1996), which was normalised (Heeringa et al., 2006) using the biggest word length in the word pair (Inkpen et al., 2005). Multiword expressions were ignored for the comparisons.

The comparisons using the automatically linked Moroccan synsets gave that Moroccan Darija has an average normalised Levenshtein distance of 0.4619 with Arabic, 0.7337 with French and 0.7521 with Spanish. The percentage of synsets with word pairs that were at least 60% similar is 42.02% for Arabic, 2.79% for French and 2.18% for Spanish. There remained 2,763 Moroccan synsets of unknown origin out of those linked to the OMW. Some have origins in Arabic, French or Spanish due to errors in linking, whereas others were found to have links to Tamazight.

The comparisons using the manually validated Moroccan synsets yielded an average normalised Levenshtein distance of 0.4393 with Arabic, 0.7544 with French and 0.7721 with Spanish, with the percentage of synsets with word pairs that were at least 60% similar is 47.00% for Arabic, 3.08% for French and 2.92% for Spanish. The results for the normalised Levenshtein distance can be considered as a validation, but the number of word pairs that were at least 60% similar is too small to give a clear validation.

The similarity between Moroccan Darija and Arabic is closest to the one between Portuguese (de Paiva et al., 2012) and Galician (Gonzalez-Agirre et al., 2012), that are two independent languages. This shows that Moroccan Darija may be considered a language of its own. Nonetheless, there is no case of WordNet dialect-language or variant-language comparison to confirm this hypothesis.

References

Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2013. On the evaluation and improvement of Arabic wordnet coverage and usability. *Language Resources and Evaluation*, 47(3):891–917.

Joseph Aquilina. 1958. Maltese as a mixed language. *Journal of Semitic Studies*, 3(1):58.

Joseph Aquilina. 1972. Maltese christian words of arabic origin.

W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, M. Bertran, and C. Fellbaum. 2006. The arabic wordnet project. In *Proceedings of LREC 2006*.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013. Sofia*, page 1352–1362.

Joseph M Brincat. 2005. Maltese–an unusual formula. *MED Magazine*, 27.

Alina Maria Ciobanu and Liviu P. Dinu. 2014. An etymological approach to cross-language orthographic similarity. application on romanian. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1047–1058.

Valéria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: an open Brazilian Wordnet for reasoning. EMAp technical report, Escola de Matemática Aplicada, FGV, Brazil.

Moha Ennaji. 2005. *Multilingualism, cultural identity, and education in Morocco*. Springer Science & Business Media.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. MIT Press.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.

Robert Lane Greene. 2013. Arabic: A language with too many armies and navies? *The Economist*.

Richard S. Harrell. 1963. A dictionary of moroccan arabic: Moroccan-english. *Georgetown University Press*.

Haut Commissariat au Plan du Maroc HCP. 2014. Recensement de la population.

Wilbert Heeringa, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In *Proceedings of the workshop on linguistic distances*, pages 51–62. Association for Computational Linguistics.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in french and english. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, volume 9, pages 251–257.

Vladimir Levenshtein, editor. 1996. *Binary codes capable of correcting deletions, insertions, and reversals*.

Shawn Magin. 2010. Illiteracy in the arab region: A meta study. *GIA Lens*, 2.

Karla Mallette. 2011. *European Modernity and the Arab Mediterranean: Toward a New Philology and a Counter-Orientalism*. University of Pennsylvania Press.

Khalil Mrini and Francis Bond. 2017. Building the moroccan darija wordnet (mdw) using bilingual resources. In *Proceedings of the International Conference on Natural Language, Signal and Speech Processing (ICNLSSP), Casablanca, Morocco*.

Benoît Sagot and Daria Fišer. 2008. Building a free french wordnet from multilingual resources. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco*.

Yves Scherrer. 2012. Recovering dialect geography from an unaligned comparable corpus. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS and UNCLH, Avignon, France*, pages 63–71.