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

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

Francis Bond
Nanyang Technological University
bond@ieee.org

Introduction

Context

Moroccan Darija, local Arabic variant, is spoken by 90.9% of population

Major influences:

- 🟩 **Arabic:** official, Afro-Asiatic
- **Tamazight:** official, Afro-Asiatic

Minor, Colonial influences:

- **French:** widespread, Indo-European
- **Spanish:** limited, Indo-European

- **Objective:** automatic estimation of linguistic influences on Moroccan Darija
- **Data:** Open Multilingual WordNet (OMW) (Bond and Foster, 2013), which contains the Moroccan Darija (MDW) (Mrini and Bond, 2018), Arabic (Black et al., 2016; Abouenour et al., 2013), French (Sagot and Fišer, 2008), Spanish (Gonzalez-Agirre et al., 2012) wordnets

Results and Discussion

- Comparisons based on automatically linked synsets of the MDW and manually validated ones.

Comparison with automatically linked synsets

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%

- If confidence scores are used as weights: average normalised Levenshtein distance of Moroccan Darija is 0.4701 with Arabic, 0.7598 with French, 0.7775 with Spanish.
- To diminish randomness of similarity, a threshold is established empirically at 60%.
- **Arabic and Moroccan Darija:** Closest pair of languages in similarity are Portuguese and Galician (average Levenshtein distance of 0.4760), two independent languages.
- **Moroccan Darija with French and Spanish:** Out of the synsets more than 60% similar, 95 are common. Future work to allow distinction of origin of influence.
- **Moroccan lemmas of unknown origin:** 2,736. They include:
  - Lemmas originating from Arabic (*nzel*, to go down), Spanish (*serbisa*, beer), not detected due to errors in linking.
  - Sizeable proportion of Tamazight origin (*seqsi*, to ask). Influence particularly visible on words starting with *ta-* and ending in *–t* (*tazellajt*, *tabennayet*, etc.).

Comparison with manually validated synsets

	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%

- Differences in figures small enough to prove linking noise was not an issue. However, the number of lemma pairs at least 60% similar is too small to give clear validation.

Estimating Influences

- Linguistic Influence = Linguistic Similarity = Average of lowest lemma-to-lemma normalised Levenshtein distance in pairs of aligned synsets
- Moroccan Darija is written in the MDW in a modified Latin alphanumeric alphabet.
- **Transliteration** bridges alphabet differences in a phonological way.
- Many possibilities of transliterations are proposed to give flexible correspondences:

	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

References

OMW: 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.
MWN: 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.*
Arabic WordNet (2006): 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.*
Arabic WordNet (2013): 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.
French WordNet: 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.*
Spanish WordNet: Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository ver- sion 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.