

Towards Automatic Normalization of the Moroccan Dialectal User Generated Text

Ridouane Tachicart & Karim Bouzoubaa

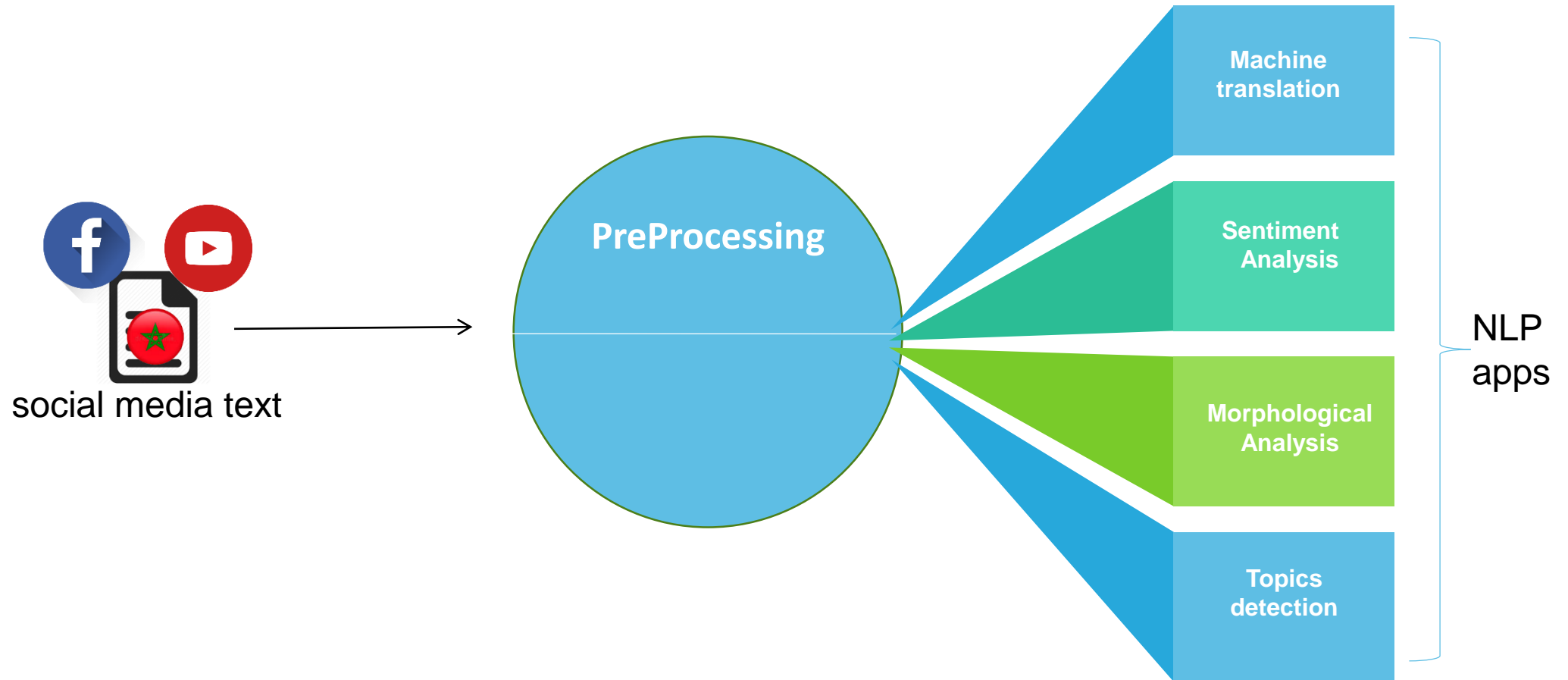
Arabic Language Engineering and Learning Modeling – ALELM Lab
Mohammed V University in Rabat - Morocco

Motivation	Problem Description	Related works	Proposed Approach	Experimental Results	Summary
	 <p>Rich information High amount of text</p> <p>5 comments</p> <p>Like Share</p> <p>Most relevant ▾</p> <p>اللي بغا نوصلوووو شي *** كوومند مرحبااا</p> <p>x5 اددرااي شكون كايعرف شي لييفروور في الرباط</p> <p>محتاج لي وصليا واحد كومونض ضروري**</p> <p>Chkon t3amel m3a amana igolina kidayra livraison w service</p> <p>جوميا مطلعاا ثمن 10 دلمرات ف الـ بزاف لا علاقة هههههه</p> <p>MLP opportunities</p>				

Market trend is based on the content of the online news articles, sentiments, and events

Several opportunities to understand consumer through text analysis (promote products, reach potential consumers...)

How to **automatically** examine Moroccan social media text in order to generate **new** and **useful** information ?



Motivation	Problem Description	Related works	Proposed Approach	Experimental Results	Summary
------------	---------------------	---------------	-------------------	----------------------	---------

Problems

😊اللي بغا نوصلووو شي *** كوومند مرحباا

❑ Noisy data *

x5اددرااي شكون كايعرف شي لييفروور في الرباط

❑ Spelling normalization

😞محتاج لي وصليا واحد كومونض ضروري**

❑ Arabizi

Chkon t3amel m3a amana igolina kidayra livraison w service

جوميا مطلعاا ثمن10 دلمرات ف الكوموند هدشي بزاف لا علاقةهههههه

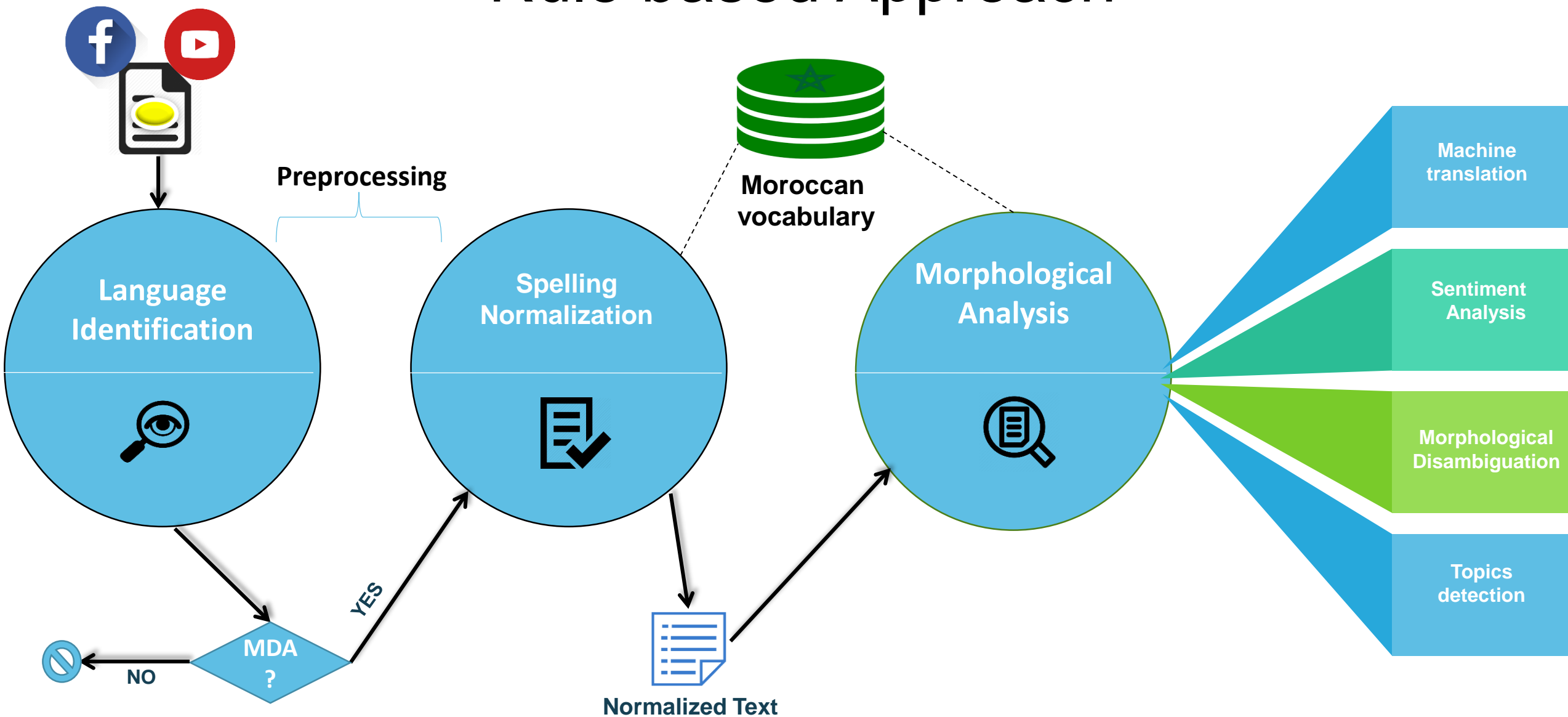
* 37% of Moroccan User generated Text is noisy
Tachicart & Bouzoubaa, 2019. An Empirical Analysis of Moroccan Dialectal User-Generated Text. ICCCI, Hendaye 2019

Existing spelling normalization Works

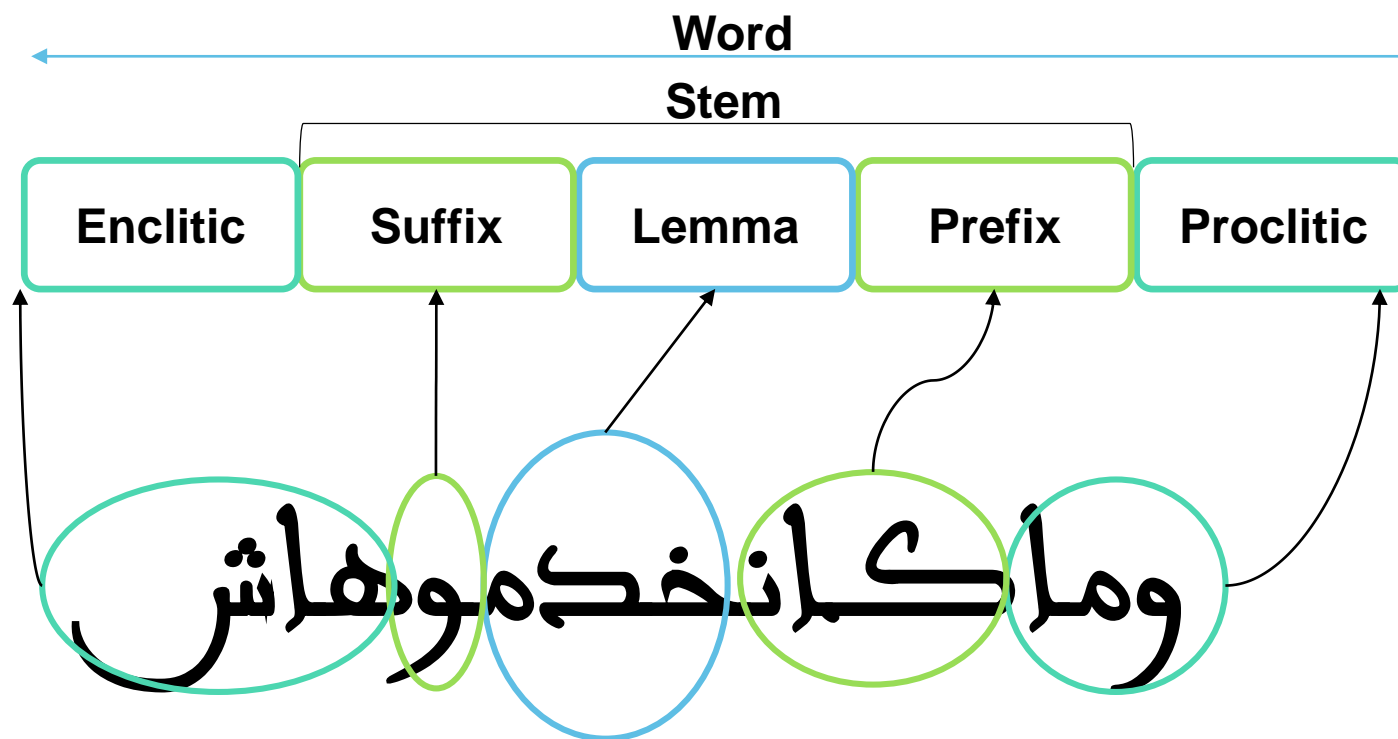
System	Authors	Description	Claimed accuracy
CODA	Habash et al.	Writting standards for Arabic dialects	-
CODAFy	Eskander et al.	(tool) converts EGY to CODA	-
Tuni CODA	Boujelbane et al.	(tool) converts TUN to CODA	86%
MADARi	Obeid et al.	(tool) annotation & spelling correction of GULF	-
UGT	Afli et al.	(tool) error correction system for Arabic UGT machine translation	68%

Different solutions are proposed for the spelling inconsistency
 cannot be extended to Moroccan
 Moroccan Arabic has not been targeted yet.

Rule based Approach



Moroccan Arabic Morphology



We do not process it

Concatenative Morphology

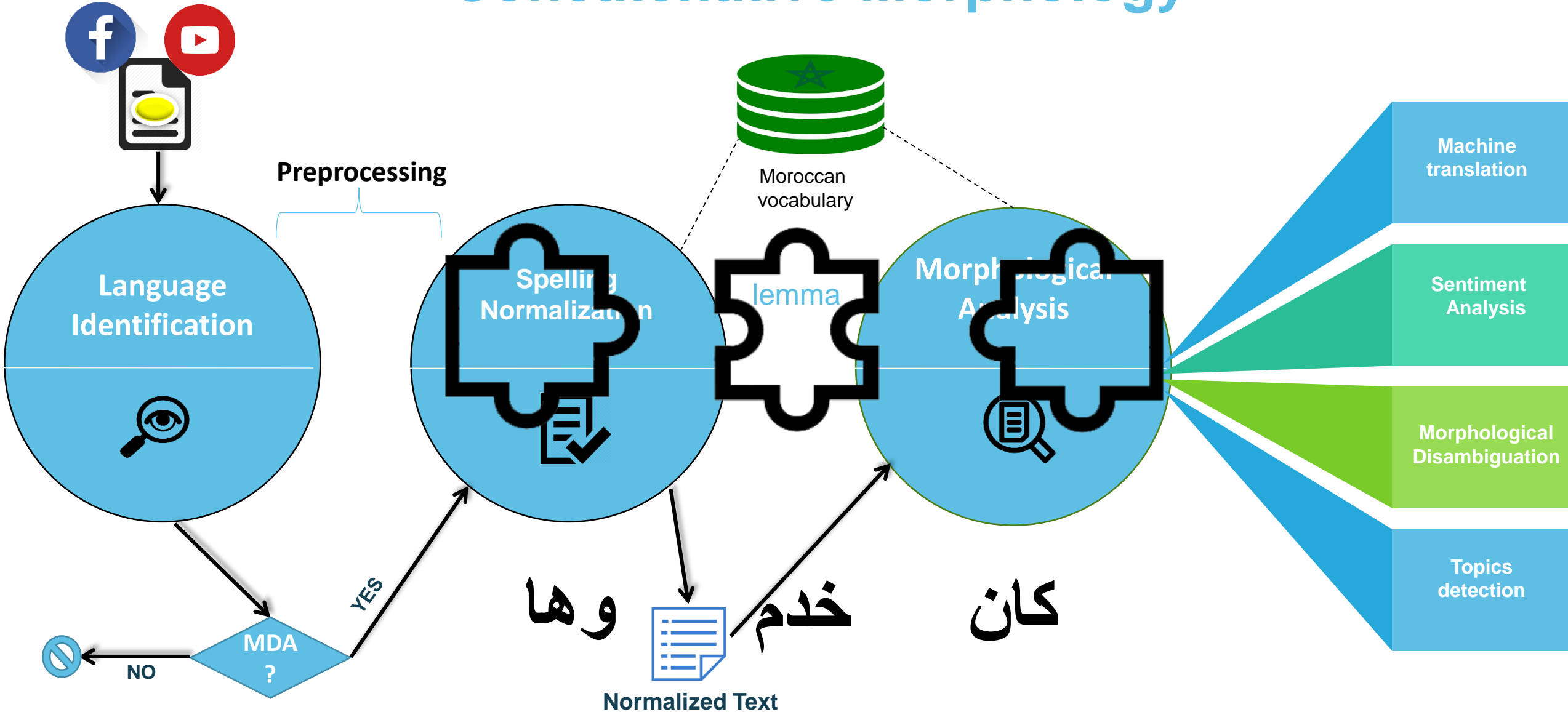


Templatic Morphology



خكم
كانفعلوها كان
خكم
كانخكموها
وها

Concatenative Morphology



Moroccan Reference Vocabulary

Prefix=وكان+v+pr+m+s+0

Suffix=ها+v+all+all+all+0

affixes + clitics


lexicon of lemmas

سطاسيونا+verb

Generator



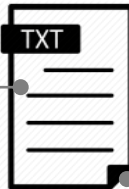
MRV



4.5M words

Word	POS	tense	gender	num	negation
وكانسطاسيونيهـا	Verb	Present	Masculine	Singular	-

rules



concatenation

orthography

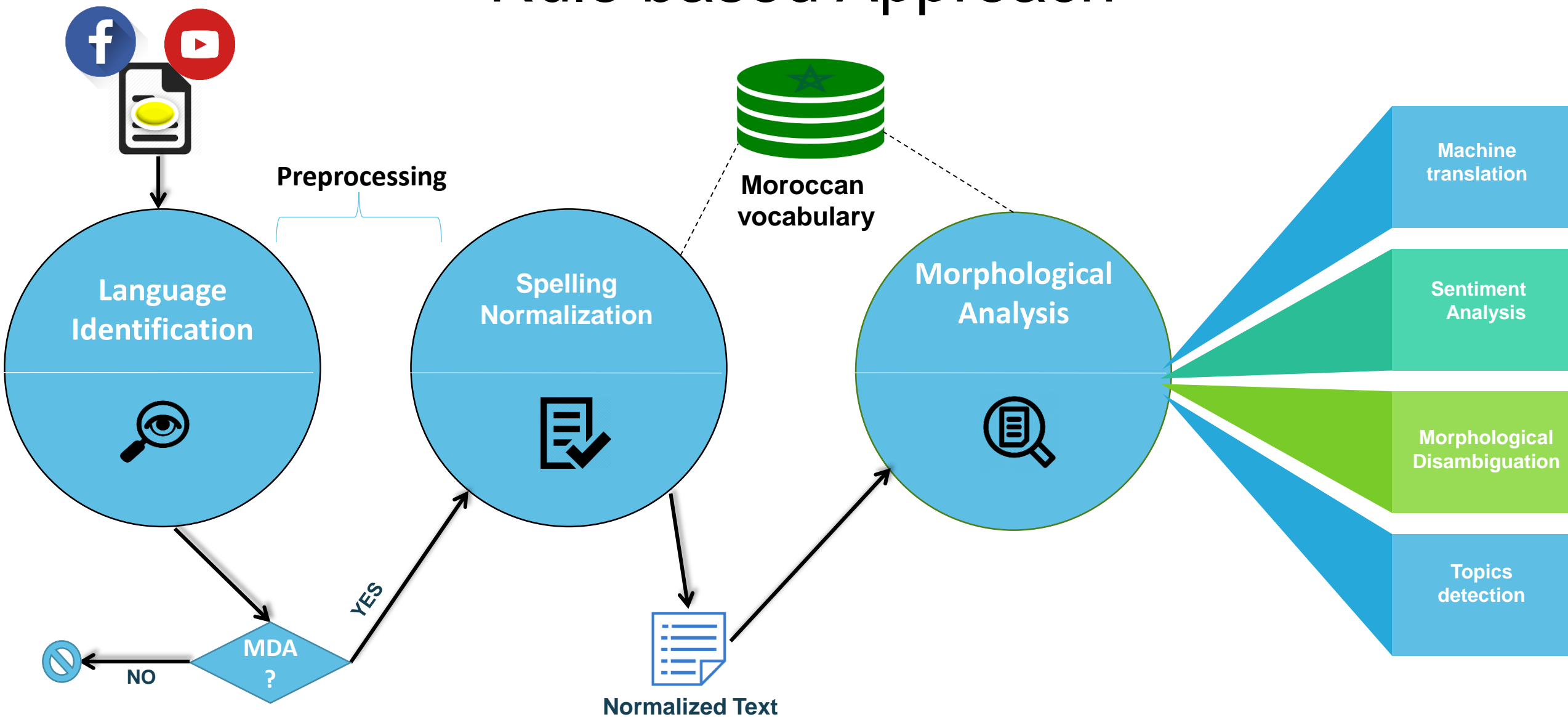
compatibility

feature	pref	Lemma	suff	word
tense	present	verb	-	present
gender	m	verb	-	m
num	s	verb	-	s
negation	0	verb	0	0

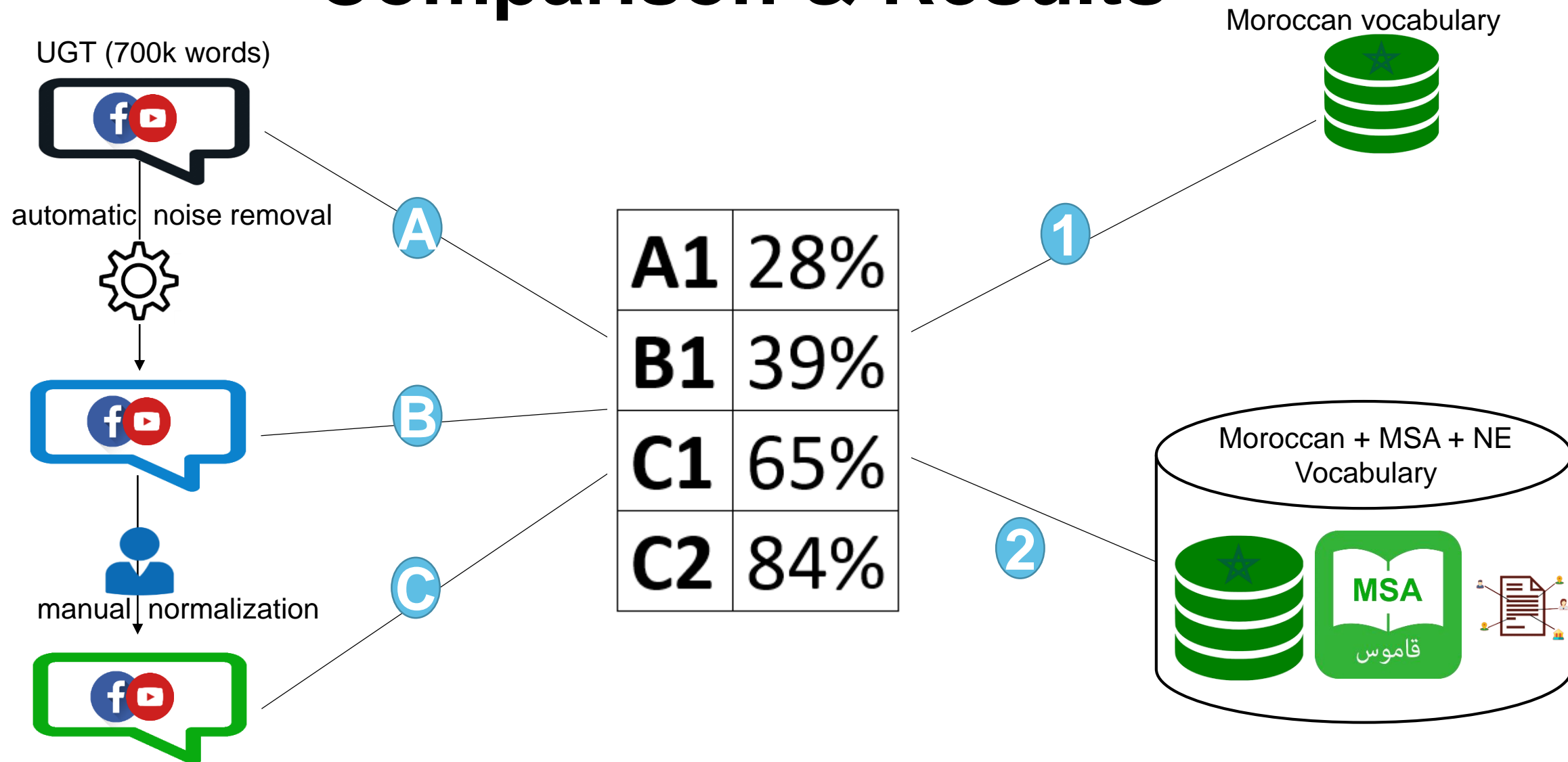
وكان + ها = true

Verb(cat2)+ present= replace ا by ي

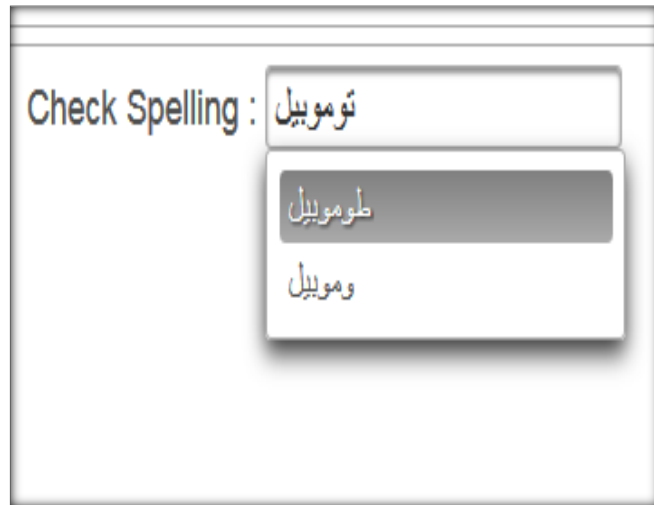
Rule based Approach



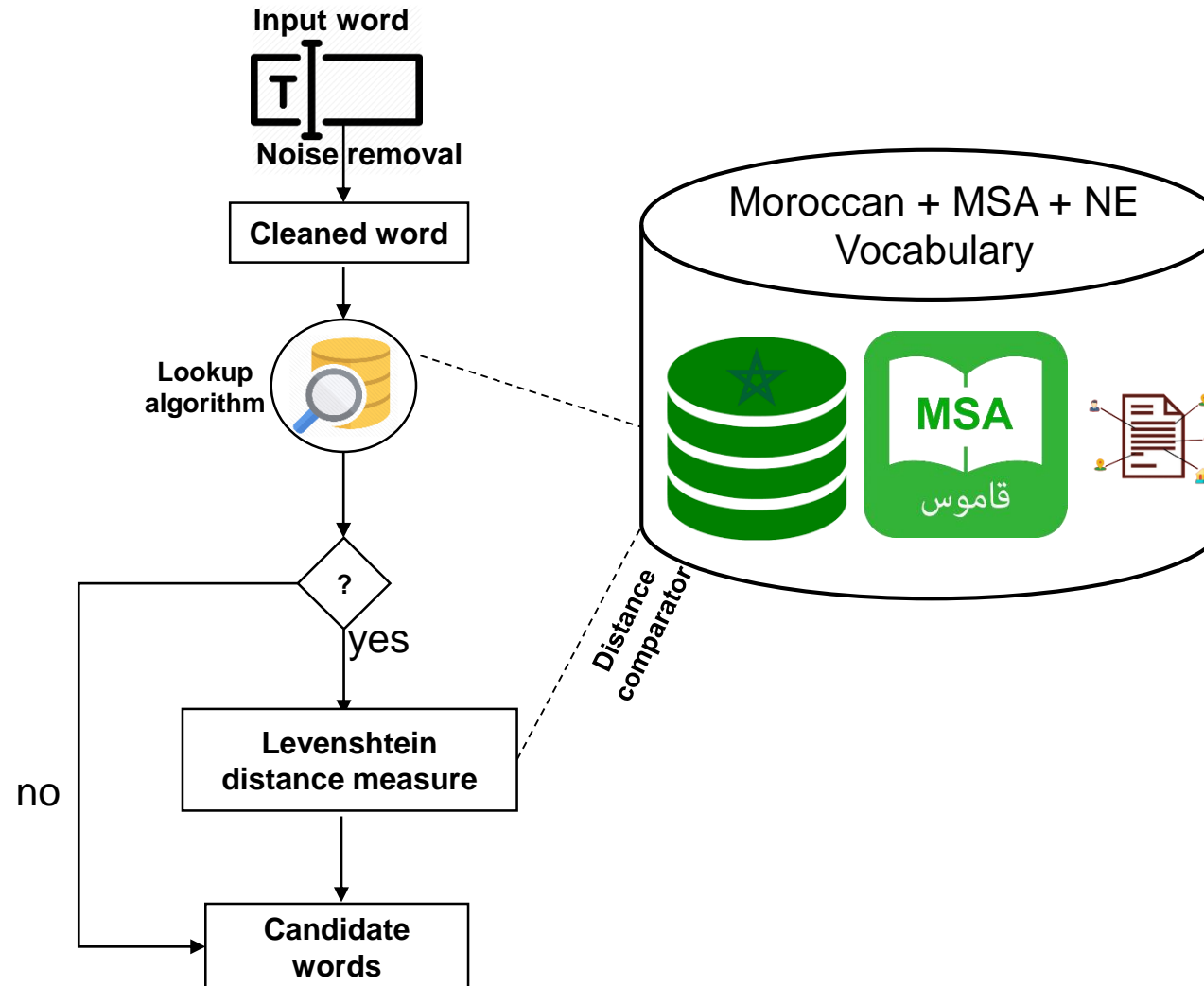
Comparison & Results



Spelling Normalization



Spelling Normalizer web interface



Spelling Normalization Evaluation

	Recall	Precision	F-measure
Proposing candidates	50%	69%	58%

Test Corpus =

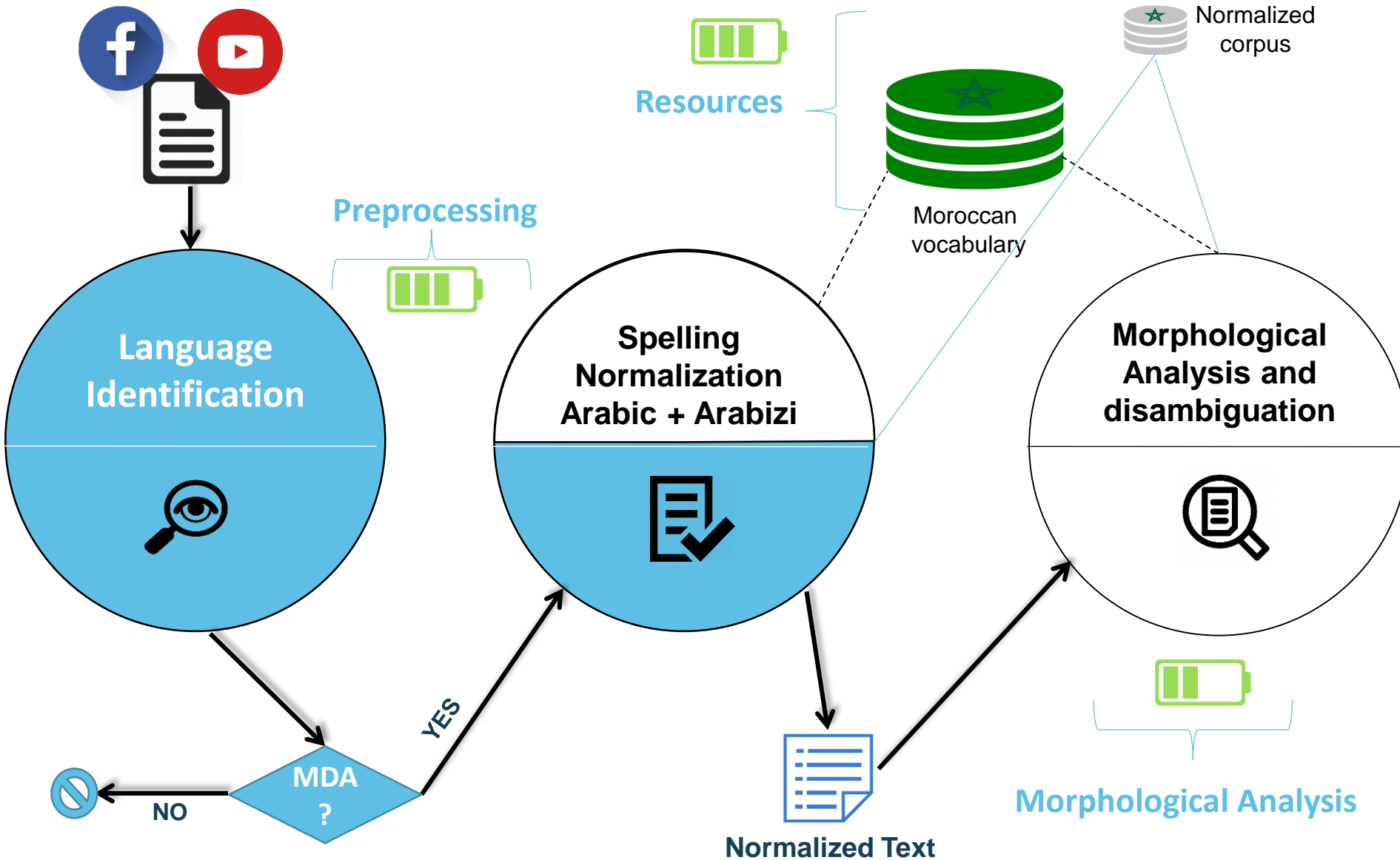
3682 words

400 sentences

على الله ايوعا و حكامنا من هاد الفديو
 الرياضة او الفن من أهم الحوايج الي تاخليو الشباب دماغهم عامر بشي حاجة ايجابية
 و هاد دراري تبارك الله عليهم دراري عضمة تانعرفهم شخصيا بالصح الرياضة خلقت منهم ناس
 ستنتجت واحد الحقيقة الي دائما تانستنجه مع راسي او تقدر تبان لكم غبية
 ستنتجت انه ما فاهمين في هاد الحياة حتى زفتة او عند راسنا ما عرفت شكون أو عيب تموت او
 برئكم اشناهم اهم النقط الي خاص المواطن يعرفهم في البرامج الانتخابية ديال الاحزاب
 متلا معدل النمو فاش **تايقوليك** غادي نطلعو
 الحلقات المقبلة من أجي تفهم أشنو فهمت غادي دوي على القانون الجنائي أو اشنو أهم
 مهم هادشي بديت فيه لهدف قبل انتخابات ديال هاد العام نكون لحت الحلقات
 حرب عشاق ابيضار ضد النكافات ديال سارة موضوع الساعة وا الله ما تكسابو شي عقل
 المشكل هو علاش الطاكسيات كايلاقوا لهاد الناس رغم ان الناس كايستأغلوا هاد التطبيق
 هاد الجوج بجوج بيهم عندهم تروة تقدر مليار دولار
 دابا رجع للتصويرة ما فيها لا ساعة روليكس لا تريكو ديال التمساح والو من داكشي
 قليل الي تلقا عندو تيليكموموند في الدار بلا سكوتش

- ❑ Investigated the problem of text normalization
- ❑ Formalized the problem as a task of UGT standardization that involves the building of a reference vocabulary
- ❑ Proposed an approach to spelling normalization
- ❑ Spelling Normalization is a crucial stage towards processing the MDA UGT

Perspectives



Questions

Demo

1	class	pref3	pref1	suff3	suff2	suff1	pe	genderPref	glossrSuffi	gend	nu	tense	negati	codePrefix	codeSuffix	pref	suff	orig	root	trn	objec	po	md	msa	id	word	
885667	##case3 - 2##	na	na	p_wa_02	na	s_w_61	s_t_2001	3	all	she	f	s	past	0	p_wa_02	s_at_11102	و	تو	A	شاف	2	3	v	شاف	رأى	6549	وشافتو
885668	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_w_61	s_t_2001	3	all	she	f	s	past	1	p_ma_01	s_at_11107	ما	توش	A	شاف	2	3	v	شاف	رأى	6549	ماشافتوش
885669	##case3 - 2##	na	p_ma_C	p_wa_02	s_ch_132	s_w_61	s_t_2001	3	all	she	f	s	past	1	p_ma_01_w	s_at_11107	وما	توش	A	شاف	2	3	v	شاف	رأى	6549	وماشافتوش
885670	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_ny_95	s_t_2001	3	all	[fem.s.]_you	f	s	past	1	p_ma_01	s_at_128	ما	ثليش	A	شاف	2	3	v	شاف	رأى	6549	ماشافتنيش
885671	##case3 - 2##	na	p_ma_C	p_wa_02	s_ch_132	s_ny_95	s_t_2001	3	all	[fem.s.]_you	f	s	past	1	p_ma_01_w	s_at_128	وما	ثليش	A	شاف	2	3	v	شاف	رأى	6549	وماشافتنيش
885672	##case3 - 2##	na	na	p_wa_02	na	s_ny_95	s_t_2001	3	all	[fem.s.]_you	f	s	past	0	p_wa_02	s_at_129	و	ثي	A	شاف	2	3	v	شاف	رأى	6549	وشافتي
885673	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_ha_89	s_t_2001	3	all	3FS	f	s	past	1	p_ma_01	s_at_ha_88	ما	تهاش	A	شاف	2	3	v	شاف	رأى	6549	ماشافتهاش
885674	##case3 - 2##	na	p_ma_C	p_wa_02	s_ch_132	s_ha_89	s_t_2001	3	all	3FS	f	s	past	1	p_ma_01_w	s_at_ha_88	وما	تهاش	A	شاف	2	3	v	شاف	رأى	6549	وماشافتهاش
885675	##case3 - 2##	na	na	p_wa_02	na	s_ha_89	s_t_2001	3	all	3FS	f	s	past	0	p_wa_02	s_at_ha_89	و	تها	A	شاف	2	3	v	شاف	رأى	6549	وشافتها
885676	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_hm_87	s_t_2001	3	all	3MP	f	s	past	1	p_ma_01	s_at_hm_86	ما	تهمش	A	شاف	2	3	v	شاف	رأى	6549	ماشافتهمش
885677	##case3 - 2##	na	p_ma_C	p_wa_02	s_ch_132	s_hm_87	s_t_2001	3	all	3MP	f	s	past	1	p_ma_01_w	s_at_hm_86	وما	تهمش	A	شاف	2	3	v	شاف	رأى	6549	وماشافتهمش
885678	##case3 - 2##	na	na	p_wa_02	na	s_hm_87	s_t_2001	3	all	3MP	f	s	past	0	p_wa_02	s_at_hm_87	و	تهم	A	شاف	2	3	v	شاف	رأى	6549	وشافتهم
885679	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_k_91	s_t_2001	3	all	2MS:2FS	f	s	past	1	p_ma_01	s_at_k_90	ما	نكش	A	شاف	2	3	v	شاف	رأى	6549	ماشافتكش
885680	##case3 - 2##	na	p_ma_C	p_wa_02	s_ch_132	s_k_91	s_t_2001	3	all	2MS:2FS	f	s	past	1	p_ma_01_w	s_at_k_90	وما	نكش	A	شاف	2	3	v	شاف	رأى	6549	وماشافتكش
885681	##case3 - 2##	na	na	p_wa_02	na	s_k_91	s_t_2001	3	all	2MS:2FS	f	s	past	0	p_wa_02	s_at_k_91	و	تك	A	شاف	2	3	v	شاف	رأى	6549	وشافتك
885682	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_km_93	s_t_2001	3	all	2MP	f	s	past	1	p_ma_01	s_at_km_92_ch	ما	نكمش	A	شاف	2	3	v	شاف	رأى	6549	ماشافتكمش
885683	##case3 - 2##	na	p_ma_C	p_wa_02	s_ch_132	s_km_93	s_t_2001	3	all	2MP	f	s	past	1	p_ma_01_w	s_at_km_92_ch	وما	نكمش	A	شاف	2	3	v	شاف	رأى	6549	وماشافتكمش
885684	##case3 - 2##	na	na	p_wa_02	na	s_km_93	s_t_2001	3	all	2MP	f	s	past	0	p_wa_02	s_at_km_93	و	نكم	A	شاف	2	3	v	شاف	رأى	6549	وشافتكم
885685	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_na_142b	s_t_2001	3	all	she	f	s	past	1	p_ma_01	s_at_na_92_ch	ما	ثناش	A	شاف	2	3	v	شاف	رأى	6549	ماشافتناش

Table 9. Sample of the MDA affixes and clitics

morpheme	value	composition	pos	tense	pers	neg	num	gen
clitic	و	simple	verb	all	all	all	all	all
prefix	وكان	و+كان	verb	present	1	0	p	all
prefix	بال	ب+ال	noun	-	-	0	all	all
suffix	ين	simple	noun	-	-	0	p	all
proclitic	وما ب	و+ما ب	noun	-	-	1	all	all
suffix	ات	simple	verb	all	3	0	s	f
enclitic	كش	ك+ش	verb	all	all	1	s	all

Concatenation constraints

feature	Morph1	lemma	Morph2	result
gender	m	verb	na	m
gender	m	verb	all	m
tense	present	verb	na	present
tense	present	verb	all	present
num	s	verb	all	s
num	s	verb	na	s

Affixes & clitics

code	position	affixe	type	example	compositi	content	negation	tense	combinaiso	num	gender	gloss	clitic
p_al_9	prefix	ال	noun	الدار	0	p_al_9	0	na	0	all	all	the	1
p_bi_3	prefix	ب	noun	بيدا	0	p_bi_3	0	na	1	all	all	by:with	1
p_al_11	prefix	بال	noun	بالموس	1	p_bi_3:p_al_9	0	na	0	all	all	with:by_+_the	1
p_ma_add1	prefix	مايل	noun		1	p_bi_3:p_al_9	1	na	1	all	all	not	1
p_kat_25	prefix	كات	verb	كاتكتب	0	p_kat_25	0	present	1	s	all	you	0
p_to_psv_21	prefix	كات	verb		1	p_kat_25:p_to_psv	0	present	1	all	all	you	2
p_kat_25b	prefix	كات	verb	كاتكتب	0	p_kat_25b	0	present	1	s	f	she	0
p_to_psv_25	prefix	كات	verb		1	p_kat_25b:p_to_psv	0	present	1	s	f	She	2
p_kat_25c	prefix	كات	verb	كاتكتبو	0	p_kat_25c	0	present	1	p	all	you	0
p_kat_67	prefix	تات	verb	تاتكتب	0	p_kat_67	0	present	1	s	all	you	0
p_kat_67b	prefix	تات	verb	تاتكتب	0	p_kat_67b	0	present	1	s	f	she	0
p_kat_67c	prefix	تات	verb	تاتكتبو	0	p_kat_67c	0	present	1	p	all	you	0
p_kay_18	prefix	كاي	verb	كايكتب	0	p_kay_18	0	present	1	s	m	he	0
p_to_psv_5	prefix	كاي	verb		1	p_kay_18:p_to_psv	0	present	1	p	all	it	2
p_to_psv_5_	prefix	كاي	verb		1	p_kay_18:p_to_psv	0	present	1	s	m	it	2
p_kay_18b	prefix	كاي	verb	كايكتبو	0	p_kay_18b	0	present	1	p	all	they	0
p_kay_68	prefix	تان	verb	تانكتب	0	p_kay_68	0	present	1	s	all	I	0
p_kay_68b	prefix	تان	verb	تانكتبو	0	p_kay_68b	0	present	1	p	all	we	0
p_kay_69	prefix	كان	verb	كانكتب	0	p_kay_69	0	present	1	s	all	I	0
p_to_psv_17	prefix	كان	verb		1	p_kay_69:p_to_psv	0	present	1	all	all	I	2
p_kay_69b	prefix	كان	verb	كانكتبو	0	p_kay_69b	0	present	1	p	all	we	0
n_kav_70	prefix	تات	verb	تاتكتب	0	n_kav_70	0	present	1	s	m	he	0

rule	correct	incorrect
if a word is originated from MSA then it is written as origin	الأرض	لرض
Negation tool is attached to verb or noun	ما ملعوبش	ما ملعوبش
Negation tool is attached to verb or noun	ما كايكملش	ما كايكملش
stop word as in MSA	في	ف
stop word ب is attached to noun	بالماء	ب الماء
كا (origin is) attached to verb	كايتمشي	كا يتمشي
replace ض by ظ		
replace ت by ث		
replace د by ذ		
ب P		
ف V		
ك G		
foreign verb ا alif	سطاسيوننا	سطاسيونني
ا is converted to ي	تكا	تكي
articles for french nouns لا , لي , لو attached	لاصال	لاصال
separate verbs	شربناها لكم	شربناها لكم
if a word is pronounced in more than way, we retain the closest to ا	بقرة قال	بكرة كال



To write
A book
An office
A writer
books

كتب
كتاب
مكتب
كاتب
كتب



affixes + clitics



MDA	MSA	pos	root	Origin	English
ماكلة	طعام	Noun	كلا	MSA	Food
شحال	كم	particle	شحال	MSA	How much
سطاسيونا	ركن	Verb	سطاسيون	French	To park
رجل	رجل	Noun	رجل	MSA	man
رجّالة	رجال	NounBP	رجل	MSA	men
رجال	رجال	NounBP	رجل	MSA	men

Vocabulary Evaluation

	Verbs	Nouns	Particles
# words	932	1453	615
OOV	11%	23%	7%

1	classe	pref3	pref1	pref1	suff3	suff2	suff1	pe	genderPref	glossrSuffi	gend	nu	tense	negativ	codePrefix	codeSuffix	pref	suff	orig	root	trm	objec	po	md	msa	id	word
885667	##case3 - 2##	na	na	p_wa_02	na	s_w_61	s_t_2001	3	all	she	f	s	past	0	p_wa_02	s_at_11102	و	ثو	A	شاف	2	3	v	شاف	رای	6549	وشافتو
885668	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_w_61	s_t_2001	3	all	she	f	s	past	1	p_ma_01	s_at_11107	ما	توش	A	شاف	2	3	v	شاف	رای	6549	وماشافتوش
885669	##case3 - 2##	na	p_ma_c	p_wa_02	s_ch_132	s_w_61	s_t_2001	3	all	she	f	s	past	1	p_ma_01_w	s_at_11107	وما	توش	A	شاف	2	3	v	شاف	رای	6549	وماشافتوش
885670	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_ny_95	s_t_2001	3	all	[fem.s.]_you	f	s	past	1	p_ma_01	s_at_128	ما	تنبش	A	شاف	2	3	v	شاف	رای	6549	ماشافتنبش
885671	##case3 - 2##	na	p_ma_c	p_wa_02	s_ch_132	s_ny_95	s_t_2001	3	all	[fem.s.]_you	f	s	past	1	p_ma_01_w	s_at_128	وما	تنبش	A	شاف	2	3	v	شاف	رای	6549	وماشافتنبش
885672	##case3 - 2##	na	na	p_wa_02	na	s_ny_95	s_t_2001	3	all	[fem.s.]_you	f	s	past	0	p_wa_02	s_at_129	و	نبی	A	شاف	2	3	v	شاف	رای	6549	وشافنبی
885673	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_ha_89	s_t_2001	3	all	3FS	f	s	past	1	p_ma_01	s_at_ha_88	ما	نهاش	A	شاف	2	3	v	شاف	رای	6549	ماشافتهاش
885674	##case3 - 2##	na	p_ma_c	p_wa_02	s_ch_132	s_ha_89	s_t_2001	3	all	3FS	f	s	past	1	p_ma_01_w	s_at_ha_88	وما	نهاش	A	شاف	2	3	v	شاف	رای	6549	وماشافتهاش
885675	##case3 - 2##	na	na	p_wa_02	na	s_ha_89	s_t_2001	3	all	3FS	f	s	past	0	p_wa_02	s_at_ha_89	و	نها	A	شاف	2	3	v	شاف	رای	6549	وشافتها
885676	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_hm_87	s_t_2001	3	all	3MP	f	s	past	1	p_ma_01	s_at_hm_86	ما	نهمش	A	شاف	2	3	v	شاف	رای	6549	ماشافتنهمش
885677	##case3 - 2##	na	p_ma_c	p_wa_02	s_ch_132	s_hm_87	s_t_2001	3	all	3MP	f	s	past	1	p_ma_01_w	s_at_hm_86	وما	نهمش	A	شاف	2	3	v	شاف	رای	6549	وماشافتنهمش
885678	##case3 - 2##	na	na	p_wa_02	na	s_hm_87	s_t_2001	3	all	3MP	f	s	past	0	p_wa_02	s_at_hm_87	و	نهم	A	شاف	2	3	v	شاف	رای	6549	وشافنهم
885679	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_k_91	s_t_2001	3	all	2MS:2FS	f	s	past	1	p_ma_01	s_at_k_90	ما	نکش	A	شاف	2	3	v	شاف	رای	6549	ماشافتنکش
885680	##case3 - 2##	na	p_ma_c	p_wa_02	s_ch_132	s_k_91	s_t_2001	3	all	2MS:2FS	f	s	past	1	p_ma_01_w	s_at_k_90	وما	نکش	A	شاف	2	3	v	شاف	رای	6549	وماشافتنکش
885681	##case3 - 2##	na	na	p_wa_02	na	s_k_91	s_t_2001	3	all	2MS:2FS	f	s	past	0	p_wa_02	s_at_k_91	و	نک	A	شاف	2	3	v	شاف	رای	6549	وشافتنک
885682	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_km_93	s_t_2001	3	all	2MP	f	s	past	1	p_ma_01	s_at_km_92_ch	ما	نکشم	A	شاف	2	3	v	شاف	رای	6549	ماشافتنکشم
885683	##case3 - 2##	na	p_ma_c	p_wa_02	s_ch_132	s_km_93	s_t_2001	3	all	2MP	f	s	past	1	p_ma_01_w	s_at_km_92_ch	وما	نکشم	A	شاف	2	3	v	شاف	رای	6549	وماشافتنکشم
885684	##case3 - 2##	na	na	p_wa_02	na	s_km_93	s_t_2001	3	all	2MP	f	s	past	0	p_wa_02	s_at_km_93	و	نکم	A	شاف	2	3	v	شاف	رای	6549	وشافتنکم
885685	##case3 - 2##	na	na	p_ma_01	s_ch_132	s_na_142b	s_t_2001	3	all	she	f	s	past	1	p_ma_01	s_at_na_92_ch	ما	نناش	A	شاف	2	3	v	شاف	رای	6549	ماشافتنناش