Istituto di Linguistica Computazionale
"Antonio Zampolli"
Consiglio Nazionale delle Ricerche

كلية العلوم
Faculté des Sciences

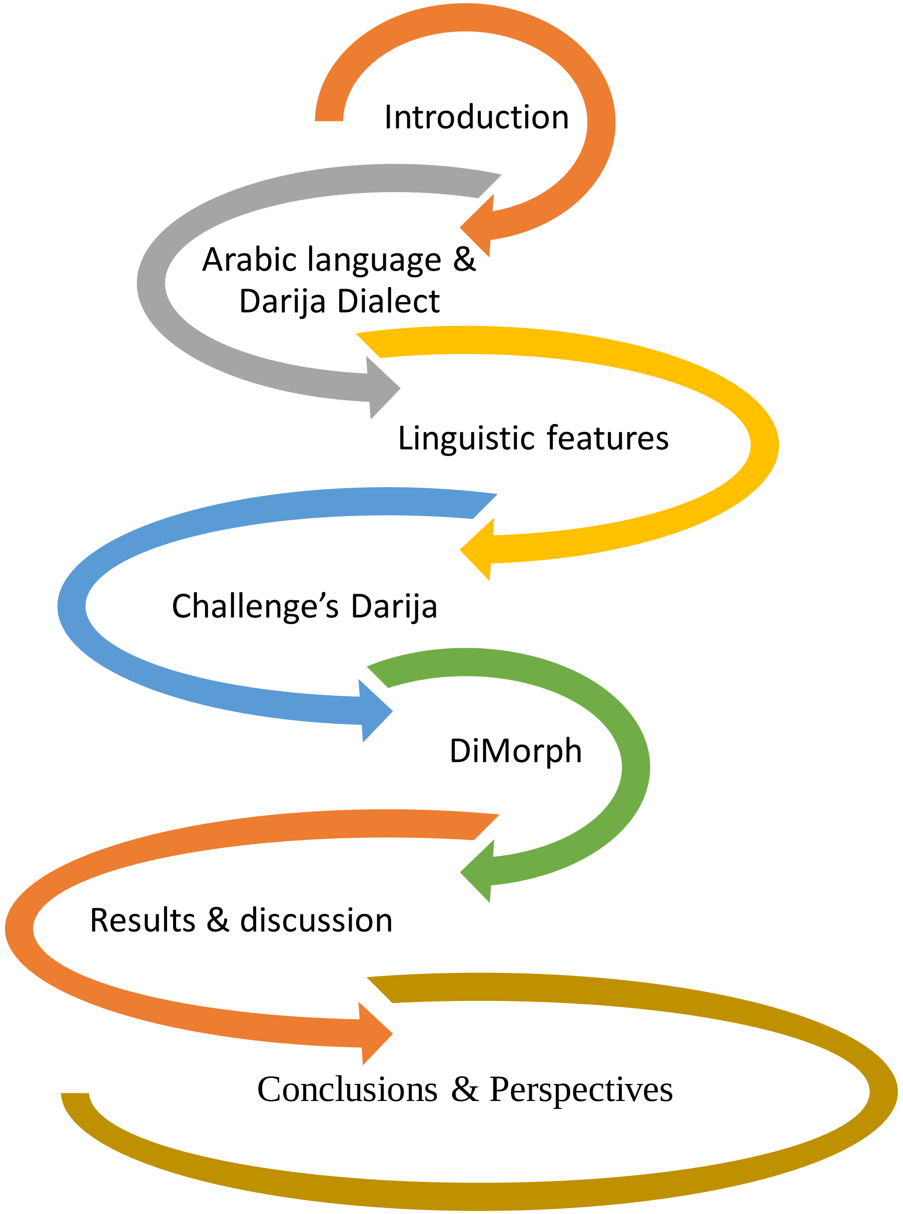
Challenges and Advances in Constructing Arabic Dialect Corpora and Linguistic Tools for the Moroccan Dialect

Nadia khlif
nadia.khlif@ilc.cnr.it

AIUCD2024

07/05/2024

PLAN



Introduction

Arabic language & Darija Dialect

Linguistic features

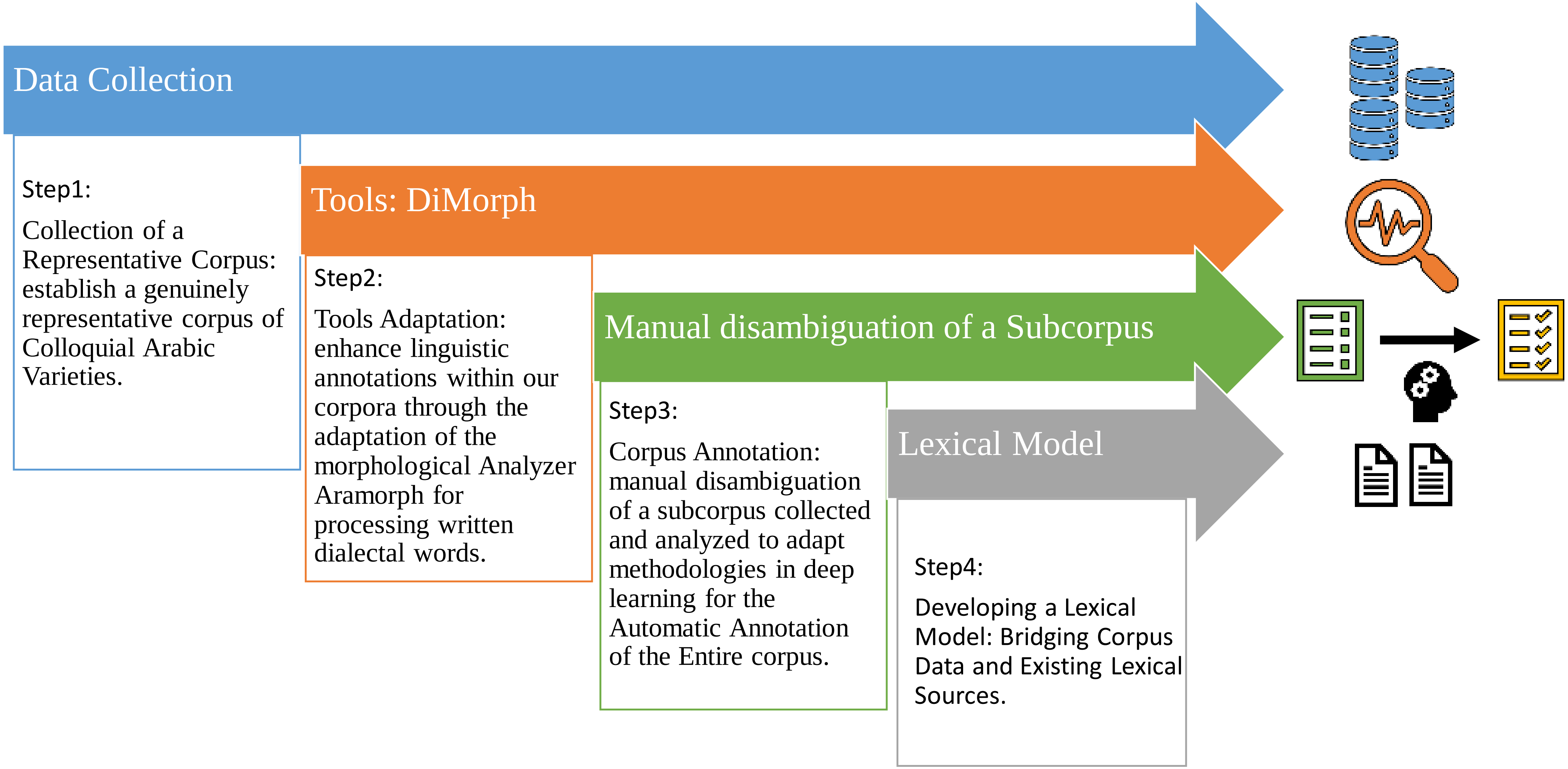
Challenge’s Darija

DiMorph

Results & discussion

Conclusions & Perspectives

Introduction



Data Collection

Step1:

Collection of a Representative Corpus: establish a genuinely representative corpus of Colloquial Arabic Varieties.

Tools: DiMorph

Step2:

Tools Adaptation: enhance linguistic annotations within our corpora through the adaptation of the morphological Analyzer Aramorph for processing written dialectal words.

Manual disambiguation of a Subcorpus

Step3:

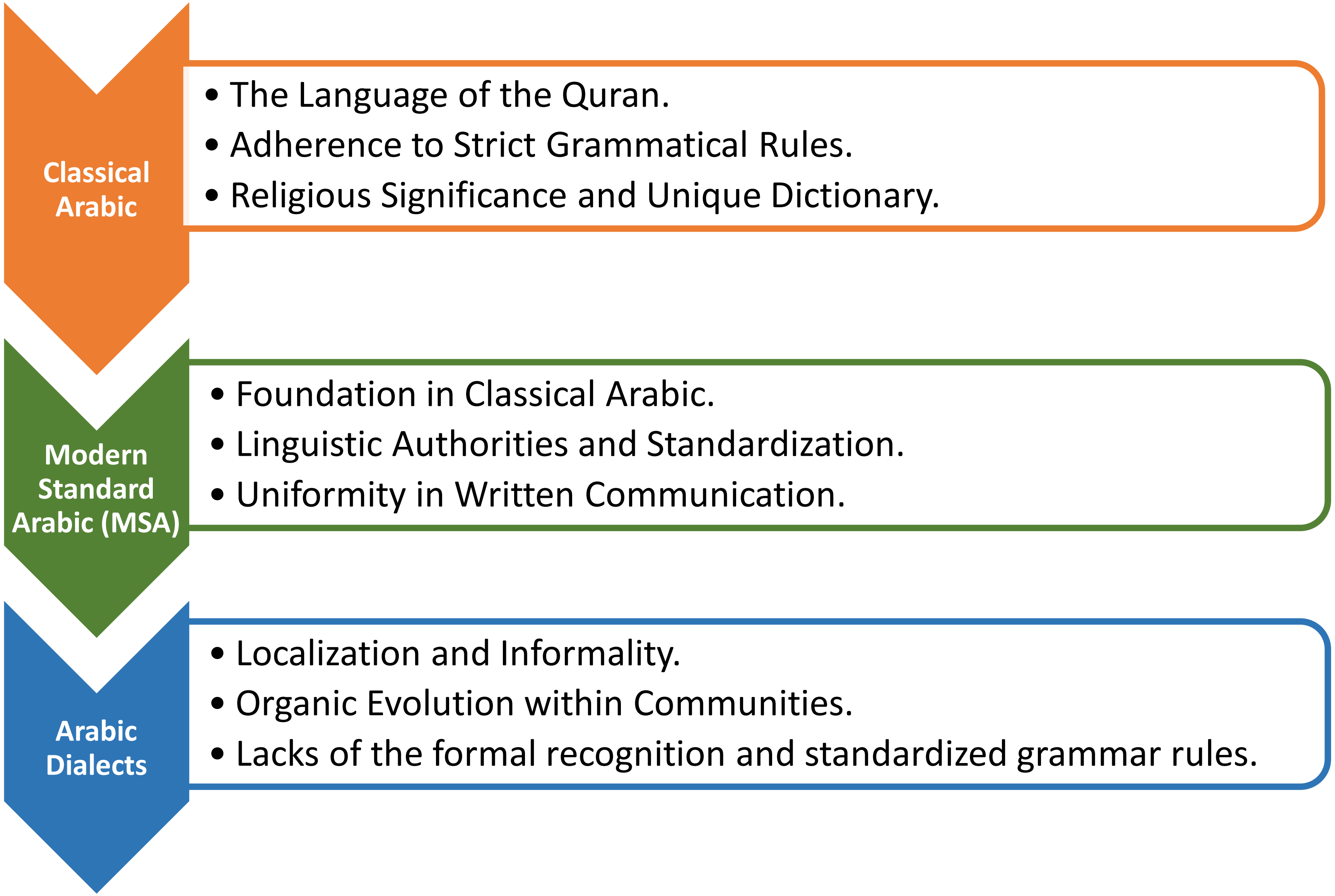
Corpus Annotation: manual disambiguation of a subcorpus collected and analyzed to adapt methodologies in deep learning for the Automatic Annotation of the Entire corpus.

Lexical Model

Step4:

Developing a Lexical Model: Bridging Corpus Data and Existing Lexical Sources.

Arabic Language & Darija dialect



Classical Arabic

- The Language of the Quran.
- Adherence to Strict Grammatical Rules.
- Religious Significance and Unique Dictionary.

Modern Standard Arabic (MSA)

- Foundation in Classical Arabic.
- Linguistic Authorities and Standardization.
- Uniformity in Written Communication.

Arabic Dialects

- Localization and Informality.
- Organic Evolution within Communities.
- Lacks of the formal recognition and standardized grammar rules.

Arabic Language & Darija dialect

Arabic Classic

MSA

Darija

/kataba/ كتب
Write

/kataba/ كتب
Write

/kotabo/ كتب
Write

/ʈˤayɂir/طائر
Bird

/ ʈˤayɂirt / طائرة
Plane

/ ʈˤayaːr /طيار
Plane

Variant phonetic

/qiʈˤar/قطار
Walking in sequence

/qiʈˤar/قطار
Train

/toraːn/ تران
Train

/sayaːrat/ سيارة
Walking in different way

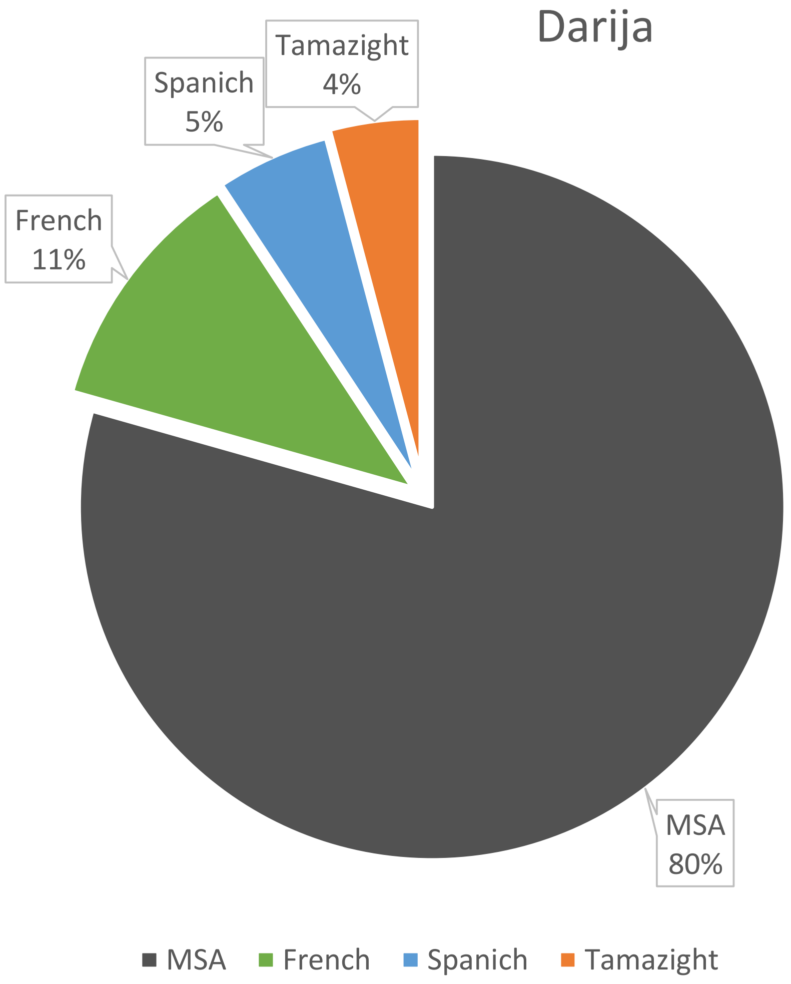
/sayaːrat/ سيارة
Car

/ʈˤuːmuːbiːɫ/ طوموبيل
Car

Word Time Shifts

Lexical borrowing

Arabic Language & Darija dialect



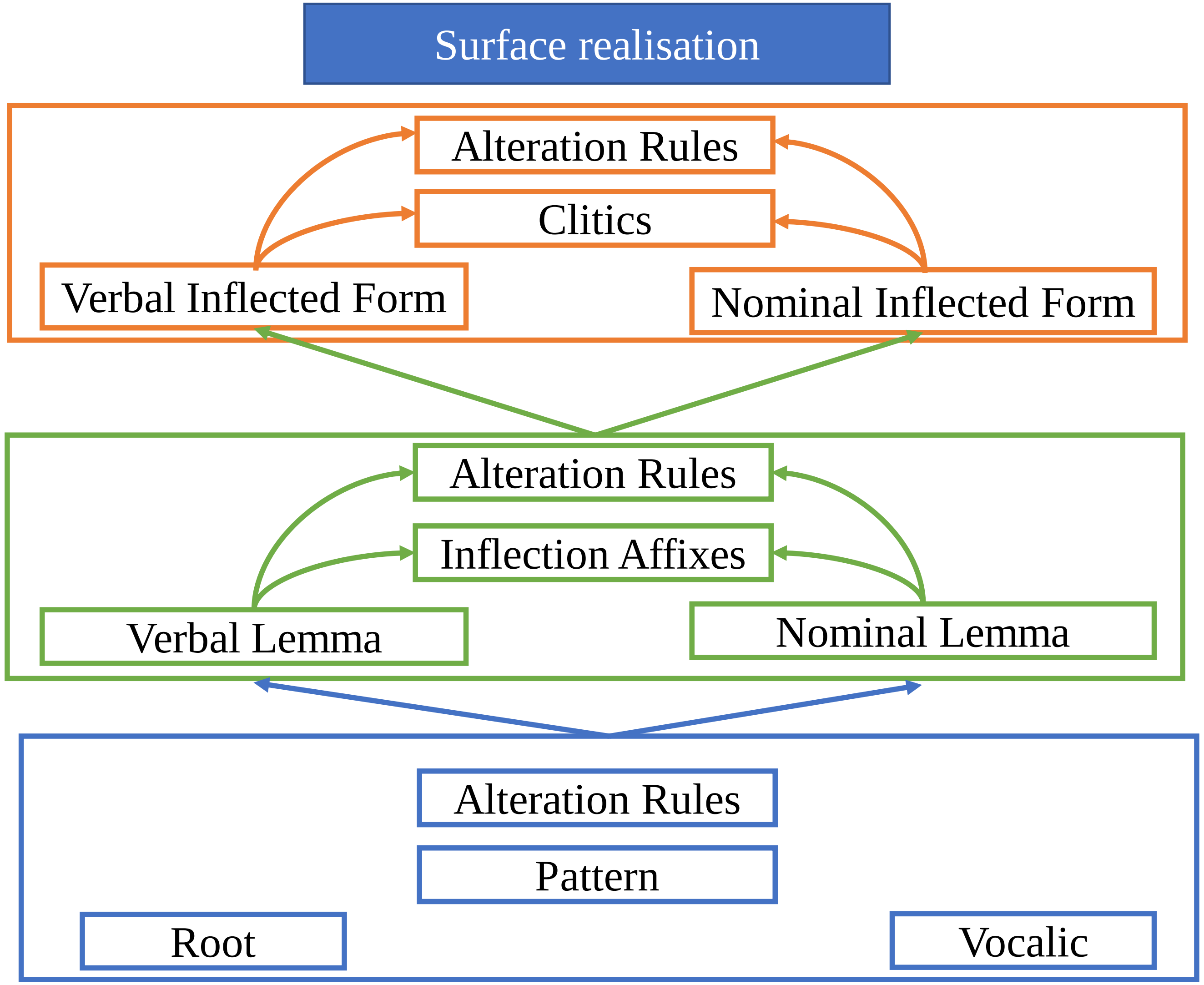
Article de reference:
Tachicart, Ridouane, Karim Bouzoubaa, and Hamid Jaafar. 2016. “Lexical differences and similarities between moroccan dialect and Arabic”.

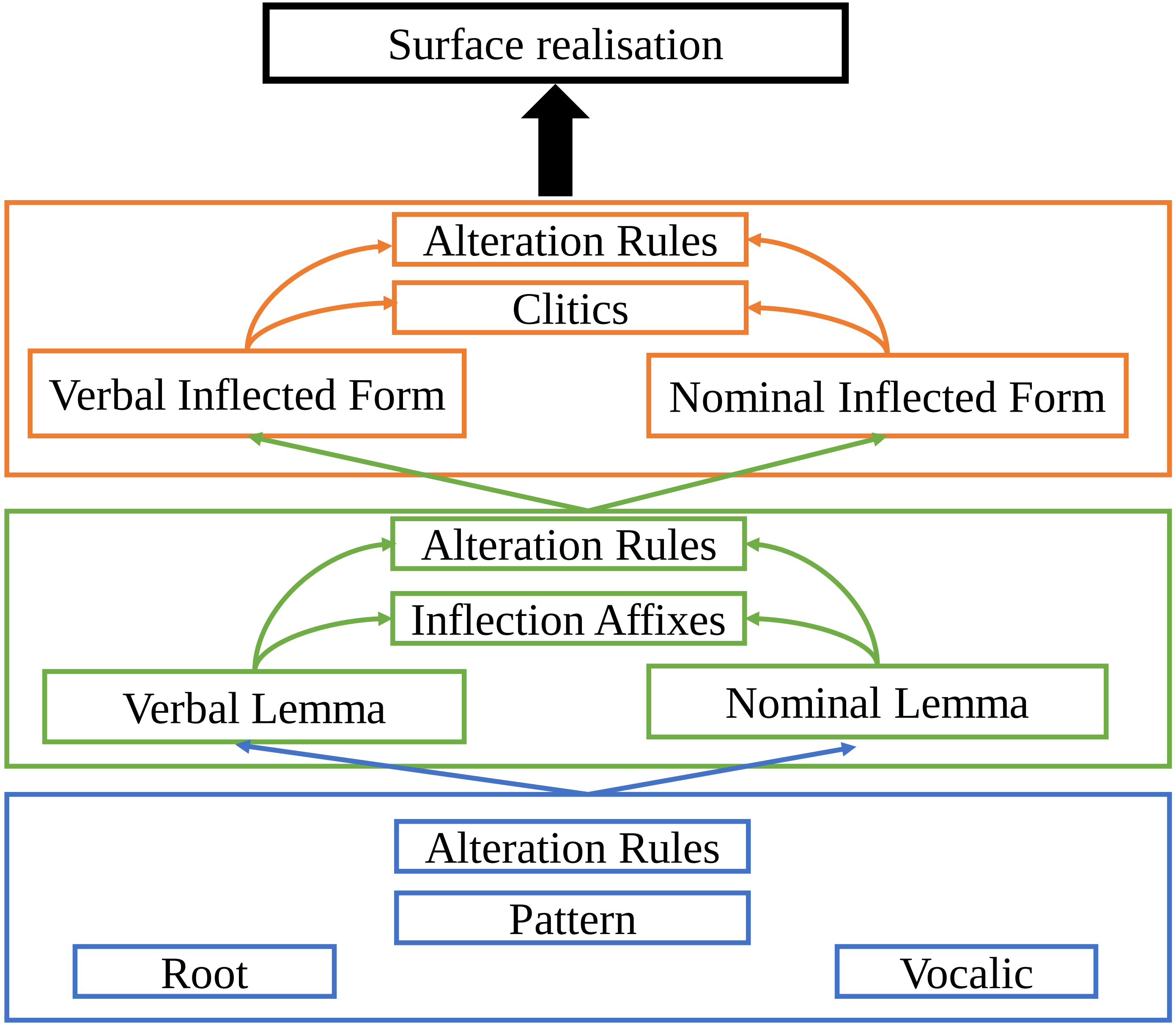
Linguistic features

- The morpho-syntactic layer combines the inflected form with clitics (prepositions, conjunctions, definite articles, etc.) to shape a rich and complex surface form.

- The inflectional layer is the one where the lemma combines with inflectional affixes to give inflectional forms.

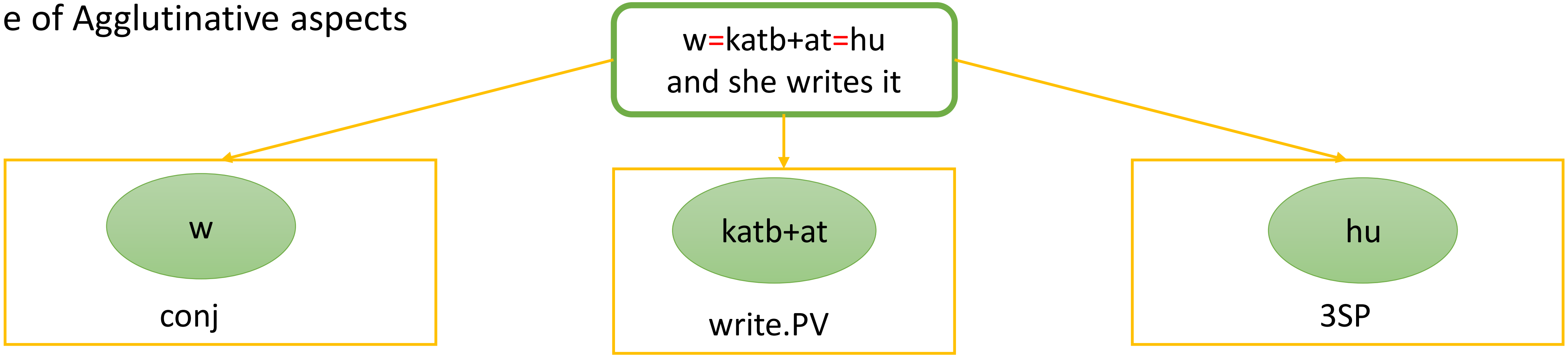
- The derivation layer is the deepest one. At this level, the root combines with the vowels, according to determined patterns, to produce a verbal or a nominal lemma.




Surface realisation
Alteration Rules
Clitics
Verbal Inflected Form
Nominal Inflected Form
Alteration Rules
Inflection Affixes
Verbal Lemma
Nominal Lemma
Alteration Rules
Pattern
Root
Vocalic

Linguistic features

Example of Agglutinative aspects



w=katb+at=hu
and she writes it

w

conj

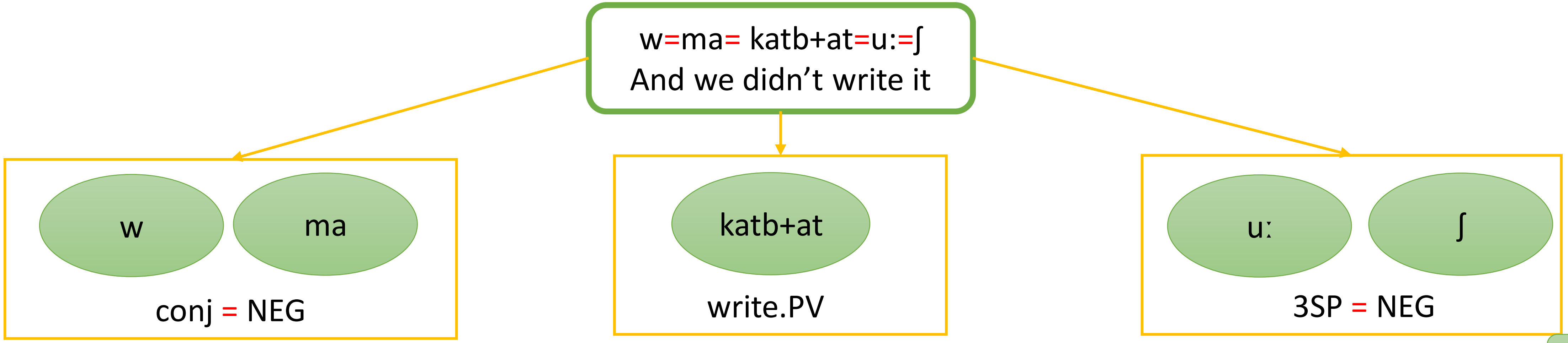
katb+at

write.PV

hu

3SP

Example of Grammatical Aspects



w=ma= katb+at=u:=ʃ
And we didn’t write it

w

ma

conj = NEG

katb+at

write.PV

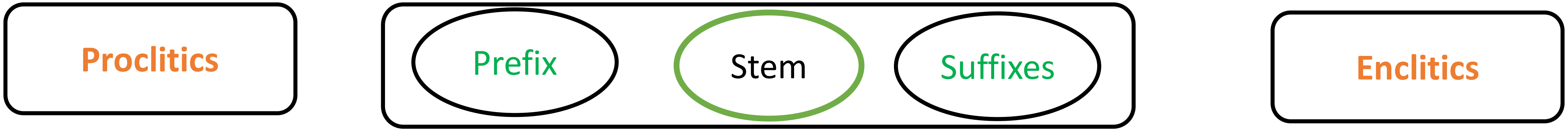
uː

ʃ

3SP = NEG

Linguistic features

In the two examples, the inflected form is surrounded by clitics and the morphological structure is:



Proclitics

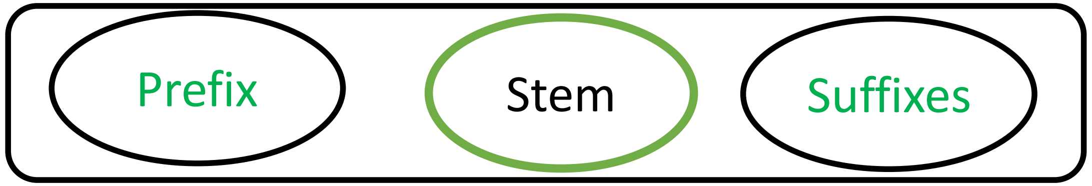
Prefix

Stem

Suffixes

Enclitics

By removing clitics, the remaining word form is a minimally autonomous inflected form whose structure consists of:

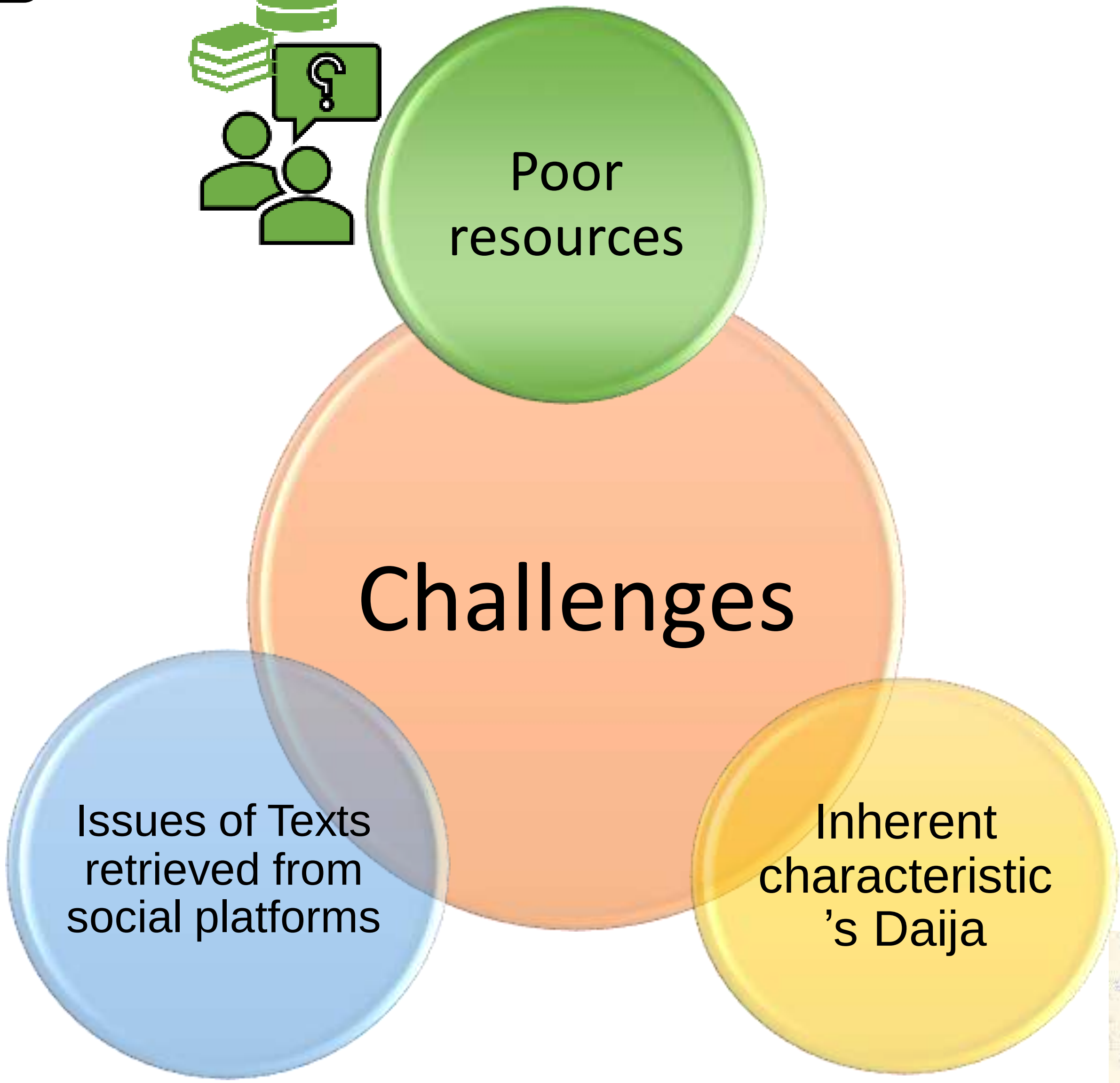


Prefix

Stem

Suffixes

Challenge’s Darija


Poor resources
Challenges
Issues of Texts retrieved from social platforms
Inherent characteristic ’s Daija

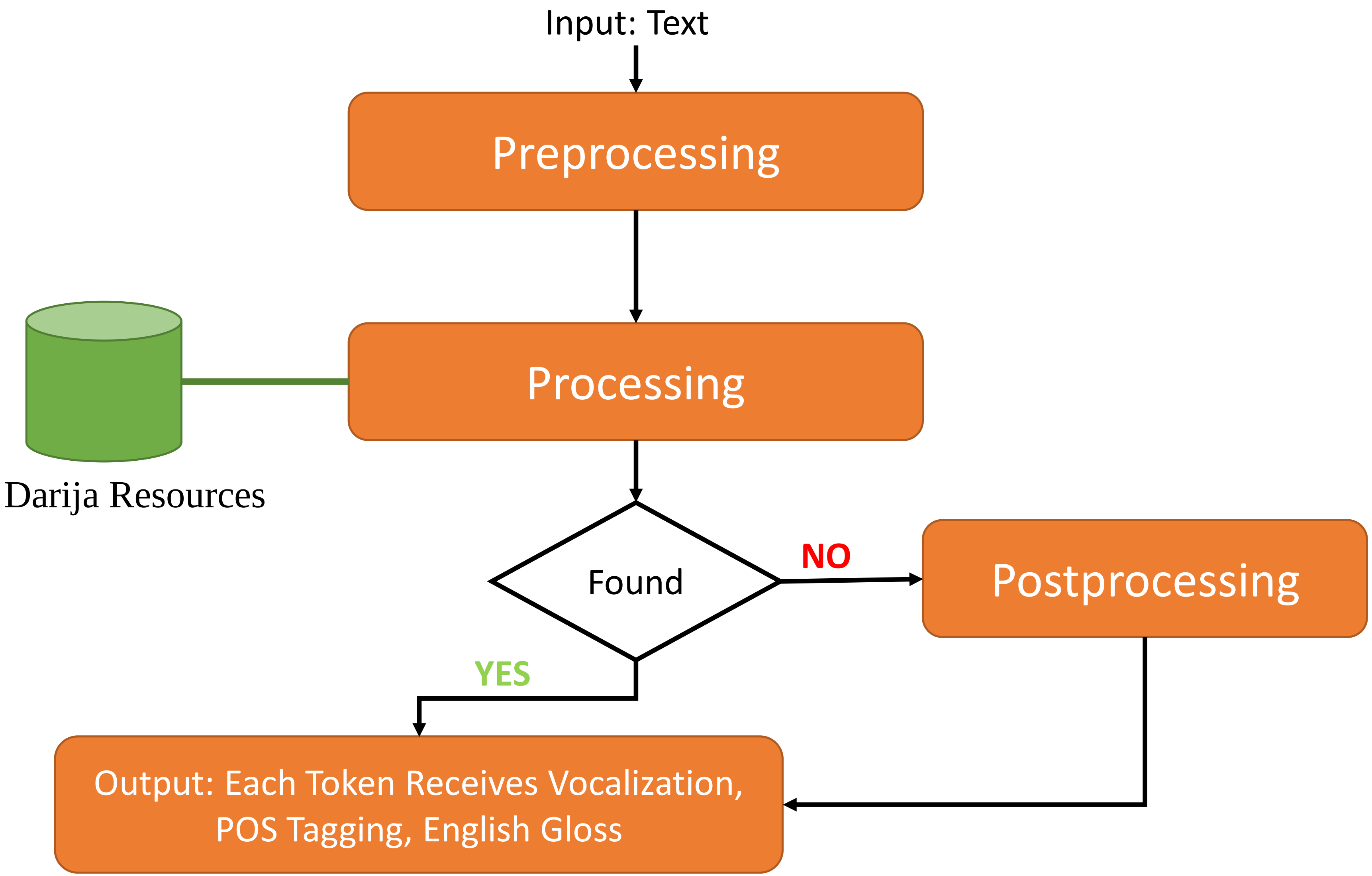
The picture can't be displayed.

Challenge’s Darija

Examples :

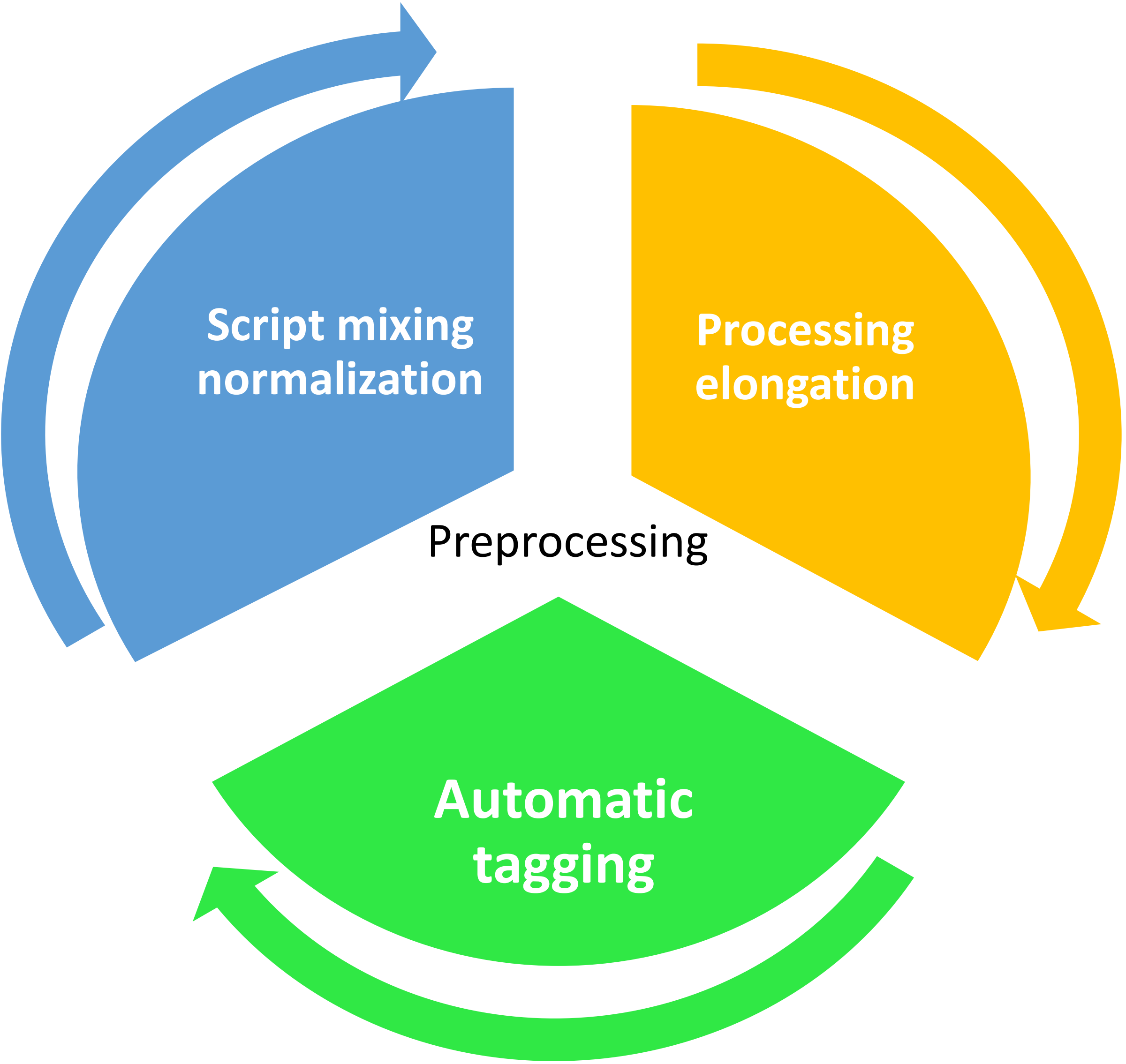
	Phenomenon	Examples	English
Script mixing	Is a mixing letter between Arabic and Latin	gال قال /qa:l/	‘say’
Code switching	Is a switching between Moroccan dialect and French words ‘les valises’.	كنخوي لي فاليز /kanaxu:i: li: fa:li:z/	‘I unload the luggage’
Code mixing	Is a French word ‘La valise’ with MSA enclitic ‘ال’ and feminine suffix ‘ة’.	الفاليزة /a:=lfa:li:z-t/	‘The luggage’

DiMorph


Input: Text
Preprocessing
Processing
Darija Resources
Found
NO
Postprocessing
YES
Output: Each Token Receives Vocalization, POS Tagging, English Gloss

DiMorph

Preprocessing



DiMorph

Processing elongation

Elongated word	Normalized word	English translation
مبرووووك mabru:u:u:u:k	مبروك mabru:k	`congratulation'
بزااف bazza:a:f	بزاف bazza:f	`much'

DiMorph

Automatic tagging

Implementation of an automatic tagging system to identify:

- Punctuation/ number.
- Emoticons.
- Interjections.

Where the foreign number or word is preceded by a dialectal prefix (e.g., ب/b/ - ف/f/ - و/w/ - ال/Al/).

For example:

- لPaola /l=Paola/ : ل /PREP+WORD_FOREIGN `for Paola’.
- ب5000 /b=5000/ : ب/PREP+NUMBER `with 5000’.

DiMorph

Script mixing normalization

Automated Substitution for One-to-One Correspondences: When each number or Latin letter can be directly replaced by an Arabic letter, we implement automated substitution. For example:

- **فت7ها => فتحها** ***/ftaHħa:/*** **`he opened it’**

Script mixing normalization

Semi-Automated Substitution for One-to-Many Correspondences: The specific case involving the Latin letter "g" or the Persian letter “گ”, both representing the phoneme /g/, isn't automatically handled due to the varied correspondences with Arabic phonemes such as “ق” /q/ or “ج” /j/. Instead, these cases are managed at the lexical level, considering the multiple possible phonetic mappings of the previous letter /g/.For example:

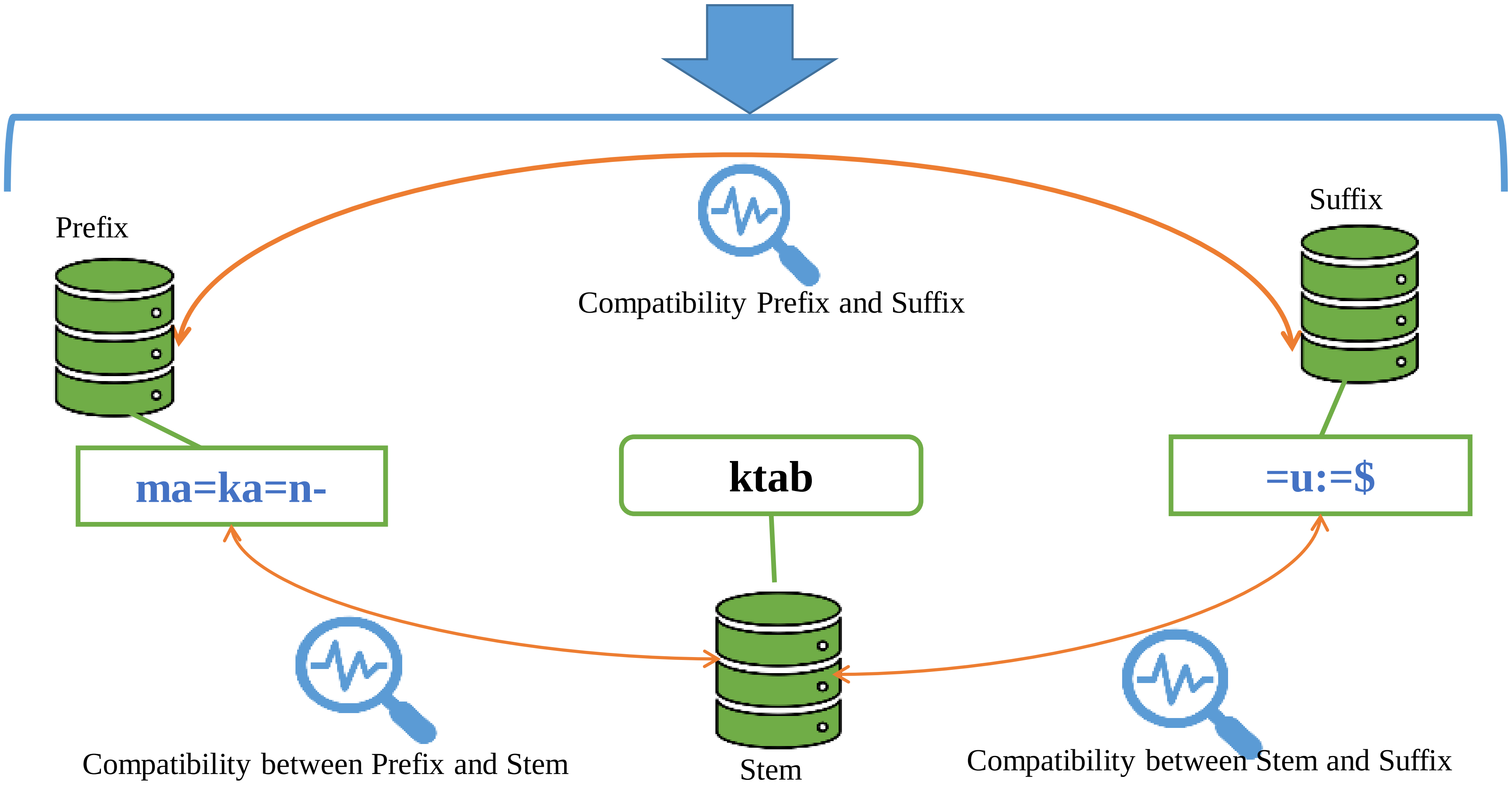
origin	Term in Darija	Cognate Arabic	Script in Arabic and Latin alphabet	Script in Arabic and Persian alphabet	English translation
Arabic	/gaːl/	قال/qaː/	gال	گال	‘say’
Arabic	/glas/	جلس/ʒlas/	gلس	گلس	‘sit’
Foreign	/gaːtˤuː/		gاطو	گاطو	‘cake’
Foreign	/gaːʕ/		gاع	گاع	‘never’

DiMorph

Processing

DET ~~+~~ Verb

ma=ka=n-**ktab**=u:=$


Prefix
Suffix
Compatibility Prefix and Suffix
ma=ka=n-
ktab
=u:=$
Compatibility between Prefix and Stem
Stem
Compatibility between Stem and Suffix

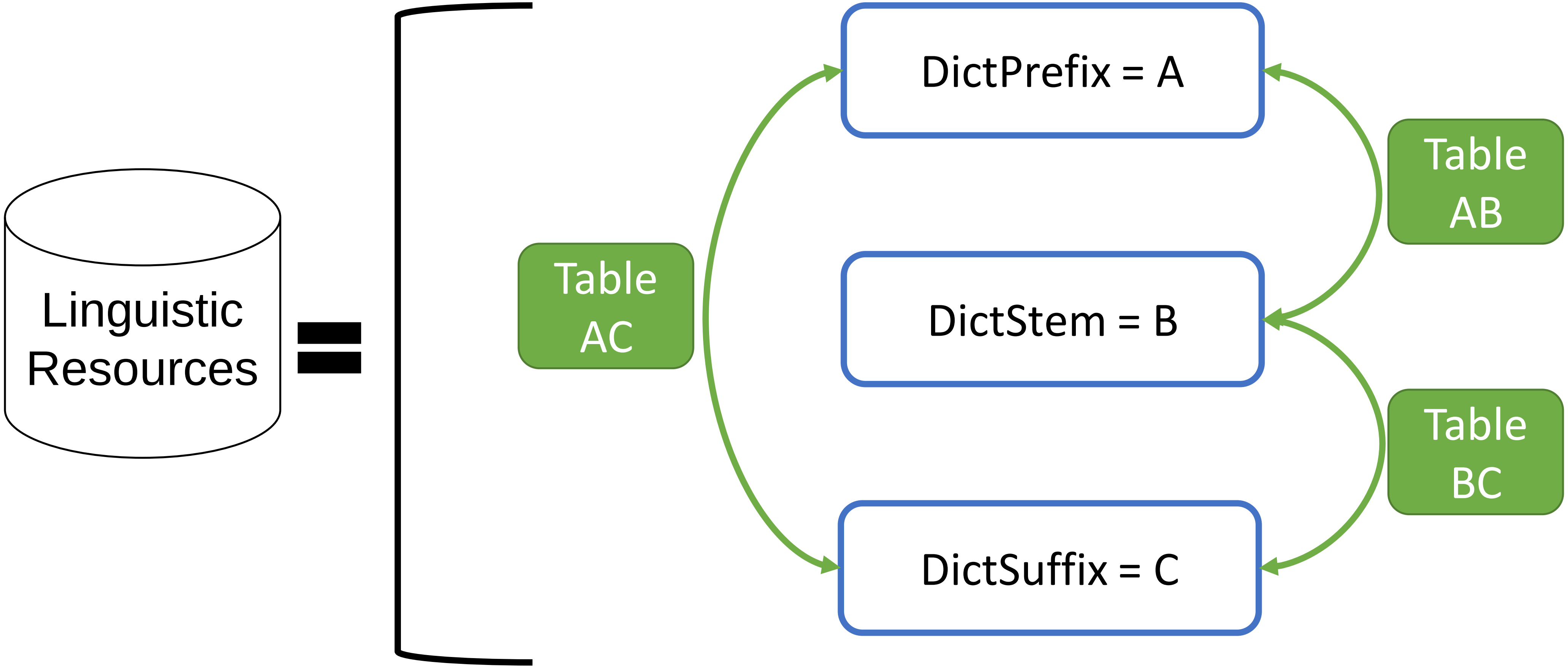
DiMorph

Linguistic Resources

Minimal word			**Prefix**	**Stem**	**Suffix**		
Maximal word	Proclitic1	Proclitic2	Prefix	Stem	Suffix	Enclitic1	Enclitic2
	DictPrefix			**DictStem**	**DictSuffix**		

DiMorph

Linguistic Resources



DiMorph

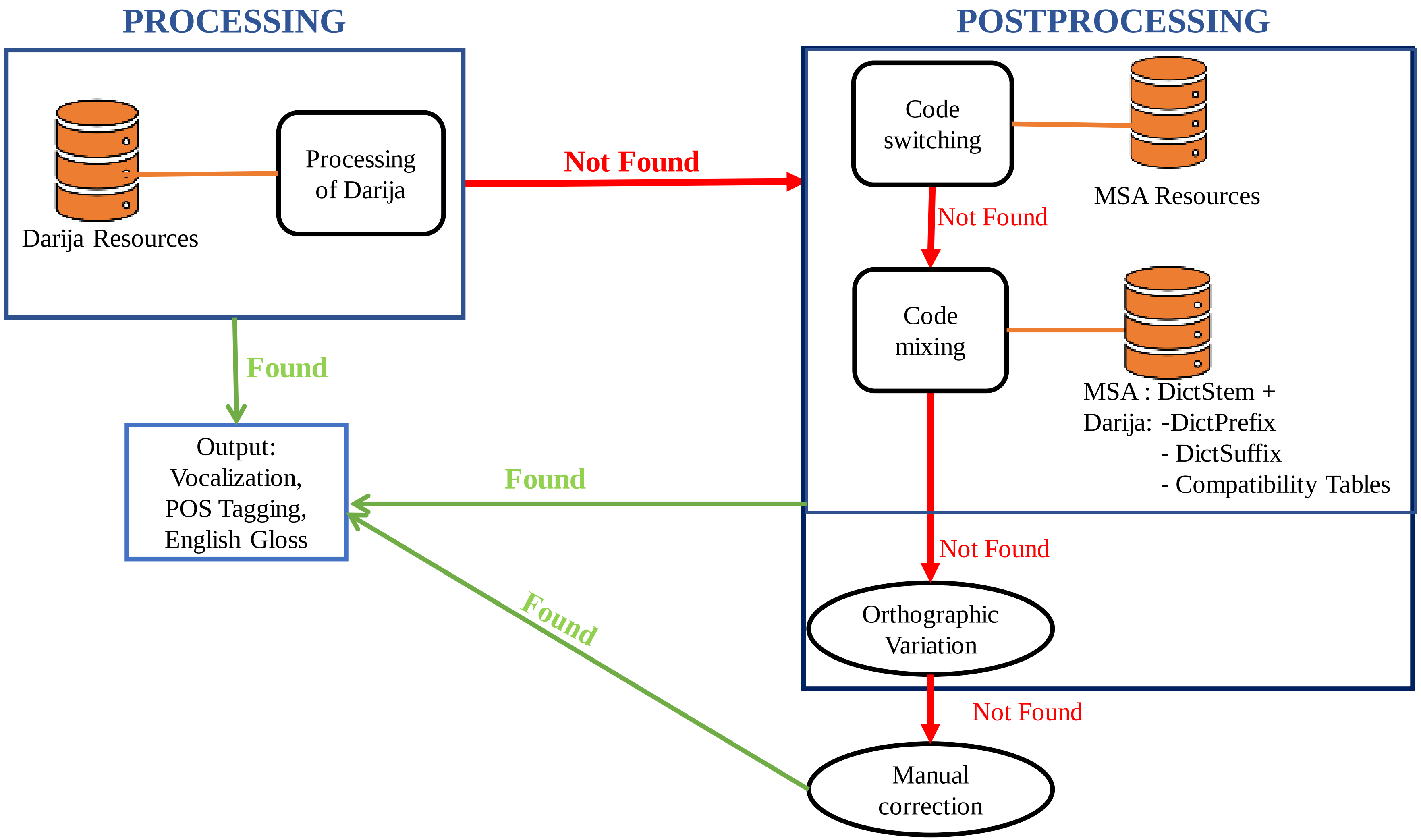
Linguistic Resources & Processing

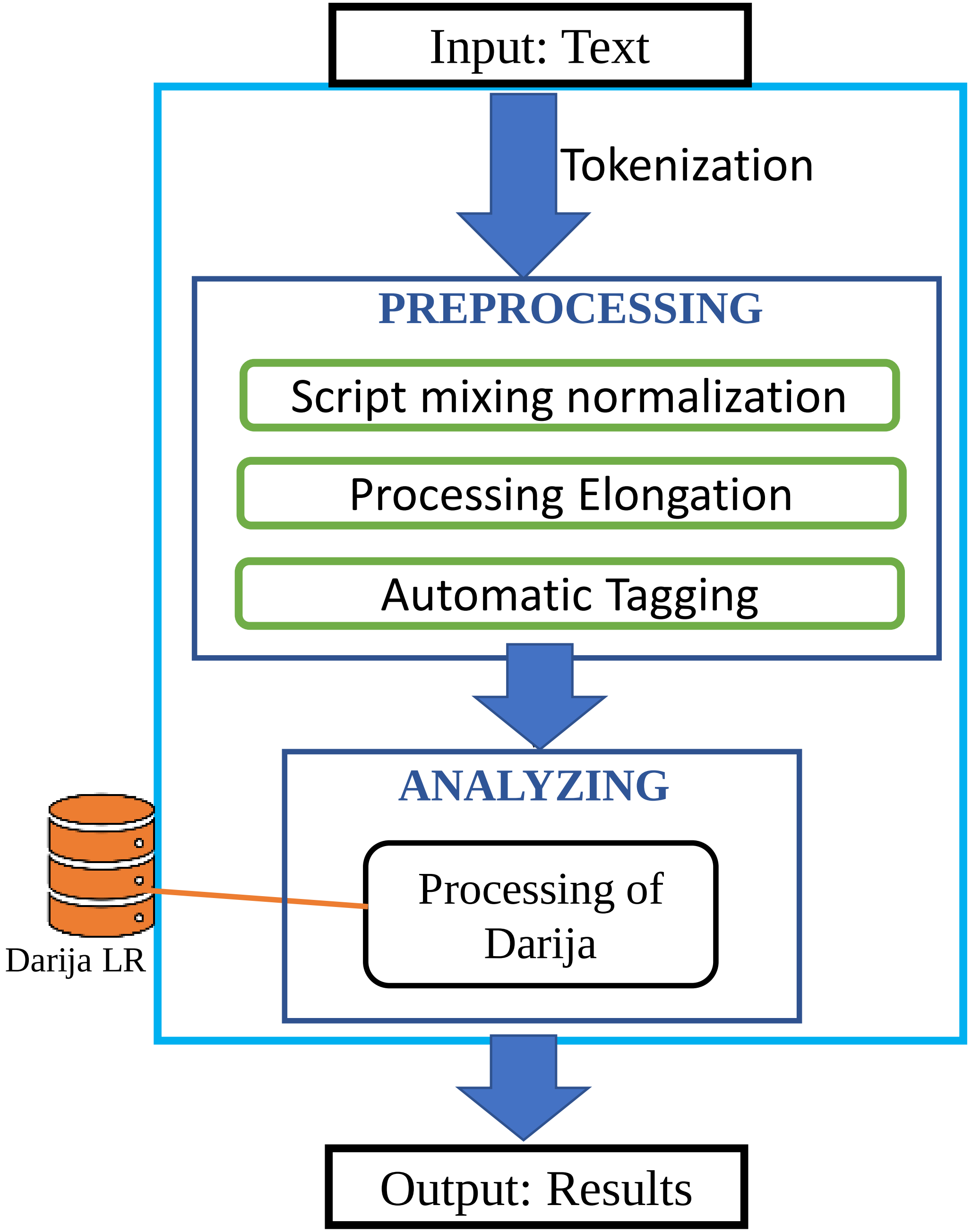
	NOUNS	VERBS	ADJECTIFS	ADVERBS	PRONOMS	FUNCTION WORDS
Darija	5169	3132	1128	146	28	156
Foreign	295	28	16	6	-	4

DictPrefix	DictSuffix	Compatibility Tables		
proclitics + prefixes	suffixes + proclitics	AB	BC	AC
297	570	770	780	1067

DiMorph

Postprocessing





Input: Text

Tokenization

PREPROCESSING

Script mixing normalization

Processing Elongation

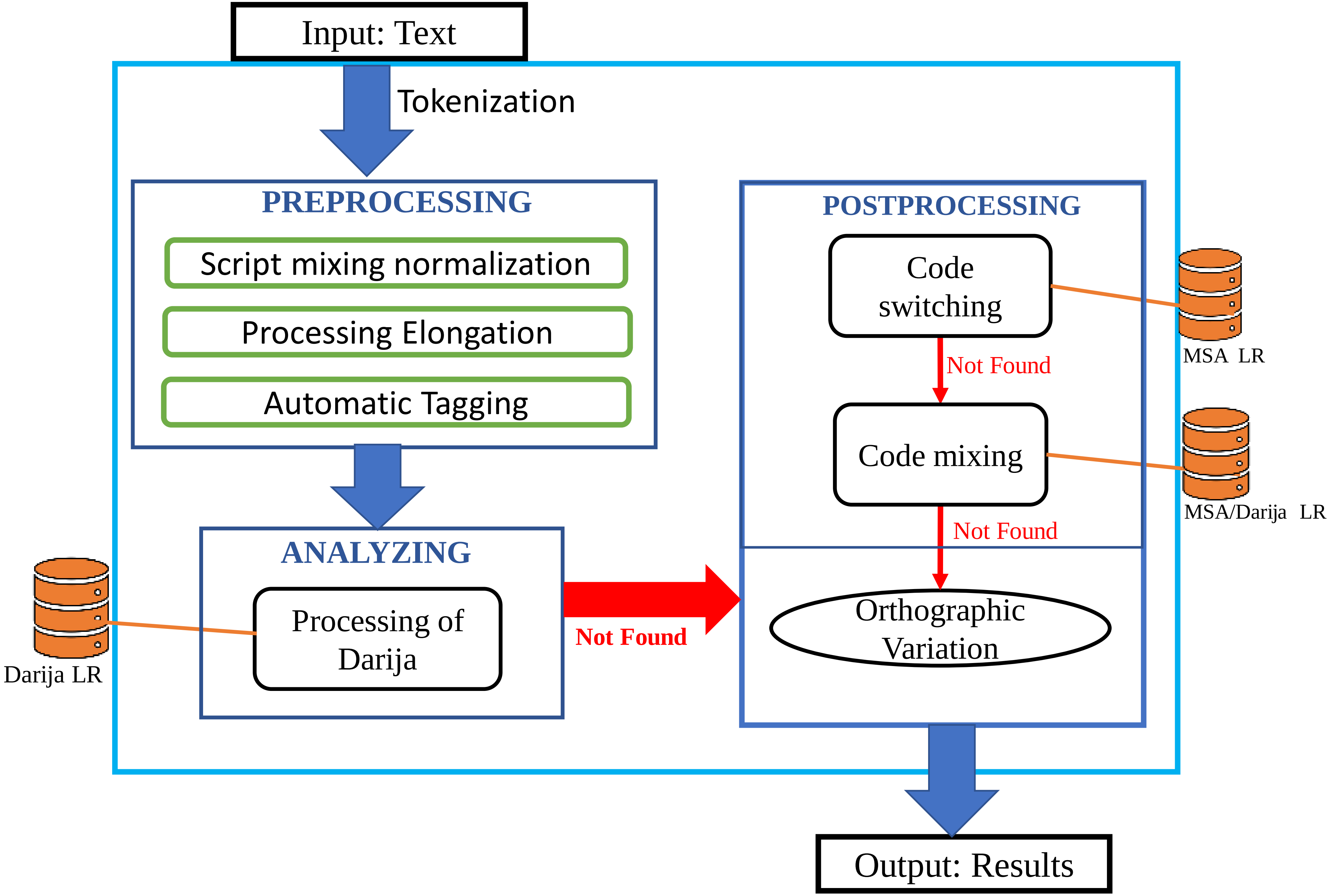
Automatic Tagging

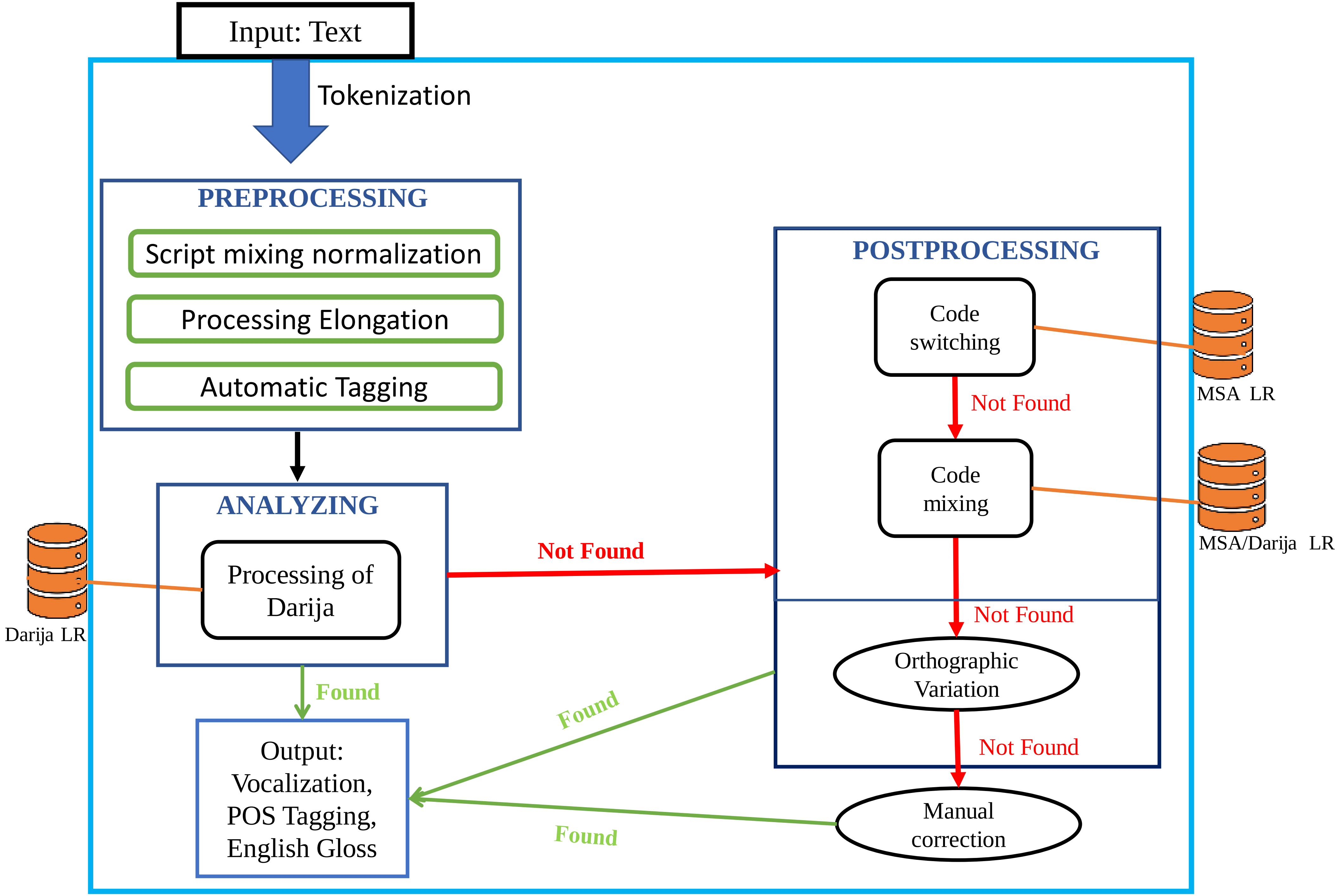
ANALYZING

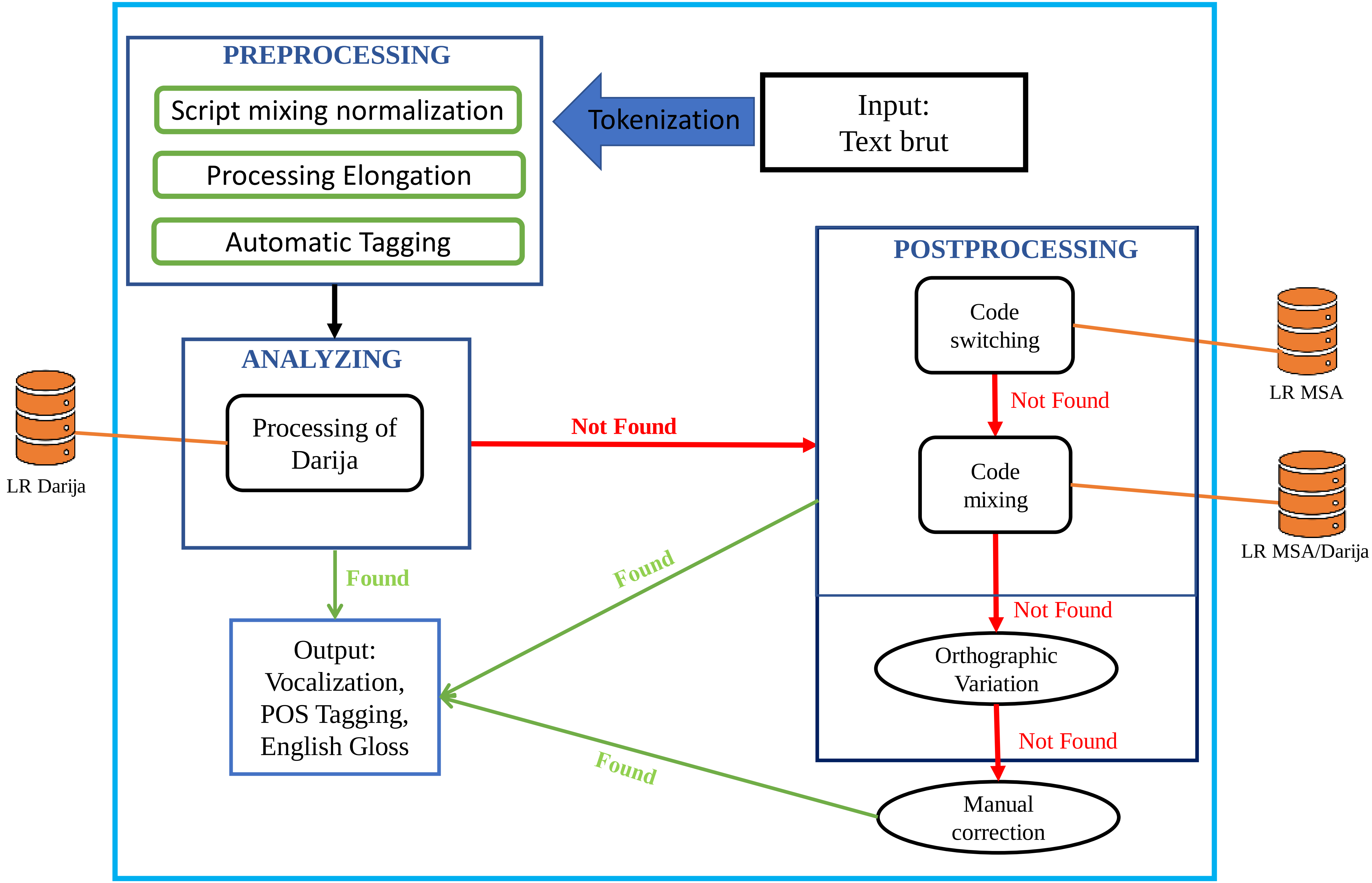
Processing of Darija

Darija LR

Output: Results


Input: Text
Tokenization
PREPROCESSING
Script mixing normalization
Processing Elongation
Automatic Tagging
ANALYZING
Processing of Darija
Darija LR
Not Found
POSTPROCESSING
Code switching
MSA LR
Not Found
Code mixing
MSA/Darija LR
Not Found
Orthographic Variation
Output: Results


Input: Text
Tokenization
PREPROCESSING
Script mixing normalization
Processing Elongation
Automatic Tagging
ANALYZING
Processing of Darija
Darija LR
Found
Output: Vocalization, POS Tagging, English Gloss
Not Found
POSTPROCESSING
Code switching
MSA LR
Not Found
Code mixing
MSA/Darija LR
Not Found
Orthographic Variation
Found
Not Found
Manual correction
Found


PREPROCESSING
Script mixing normalization
Processing Elongation
Automatic Tagging
Tokenization
Input:
Text brut
ANALYZING
Processing of Darija
LR Darija
Not Found
POSTPROCESSING
Code switching
LR MSA
Not Found
Code mixing
LR MSA/Darija
Not Found
Orthographic Variation
Not Found
Manual correction
Found
Found
Found
Output:
Vocalization,
POS Tagging,
English Gloss

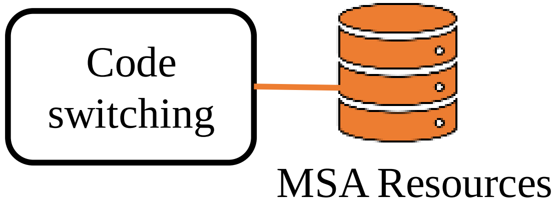
DiMorph

Postprocessing

MSA Detection in DiMorph: Identifying Code-Switching Tokens Not Analyzed as MSA.

For example:

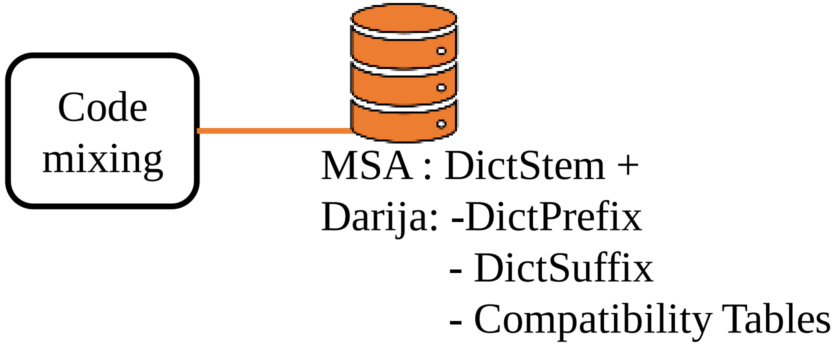
- **سنستورد /sa=na-stawrid-u/ “We will import”.**



MSA-Darija Identification: Detecting Code-Mixing Tokens in DiMorph through Analysis of Clitics in Darija and Stem in MSA.

For example:

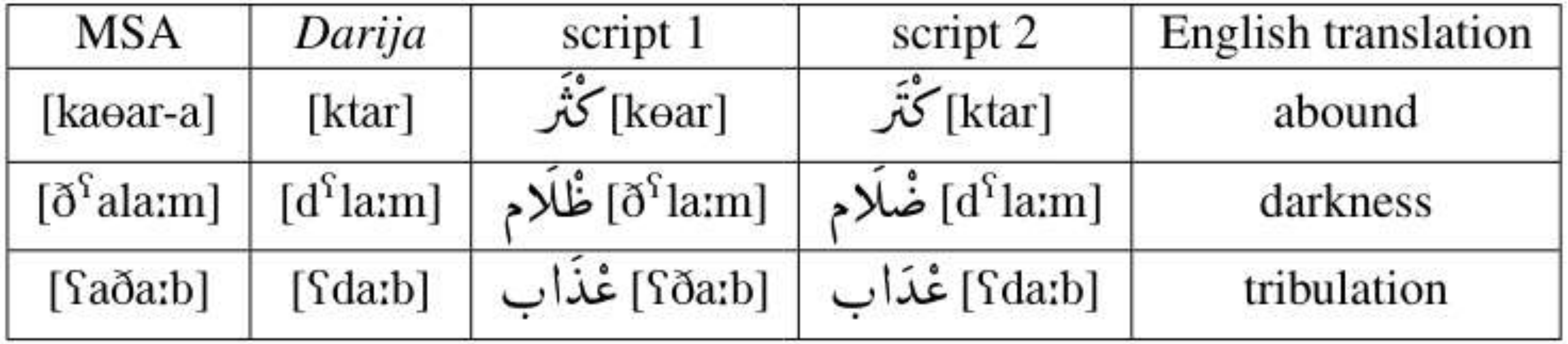
- **غنستوردو /ʁa=nstawrd-uː/ “We will import”.**



Postprocessing

Orthographic variation

Phonological simplification: This refers to the process whereby a language alters its sound system, resulting in the reduction or elimination of certain phonetic features.



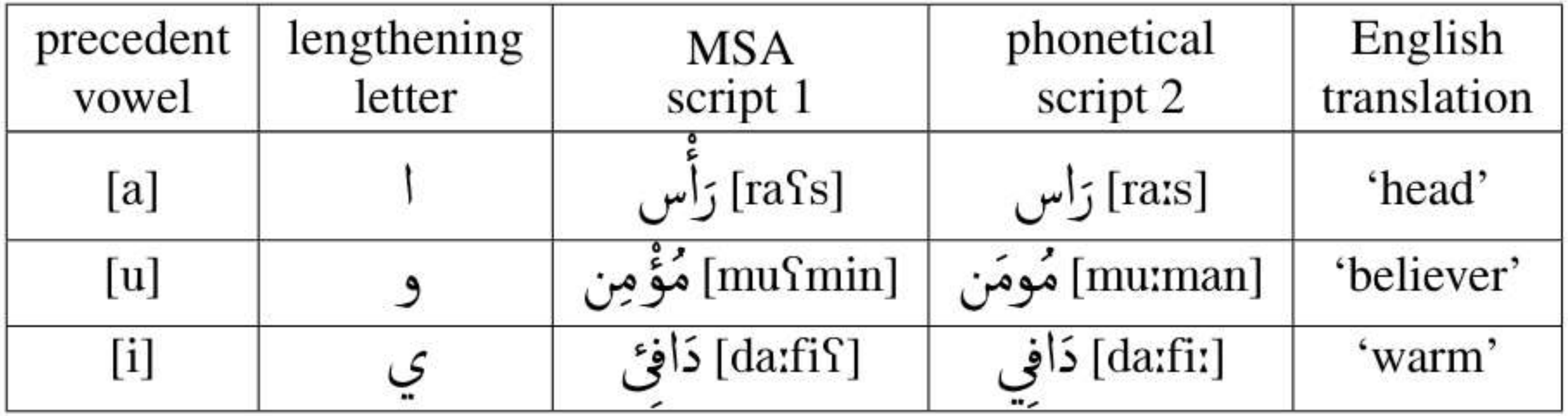
MSA	*Darija*	script 1	script 2	English translation
[kaθar-a]	[ktar]	كْثَر [kθar]	كْتَر [ktar]	abound
[ðˤala:m]	[dˤla:m]	ظْلَام [ðˤla:m]	ضْلَام [dˤla:m]	darkness
[ʕaða:b]	[ʕda:b]	عْذَاب [ʕða:b]	عْدَاب [ʕda:b]	tribulation

DiMorph

Postprocessing

Orthographic variation

Strict spelling conventions: include the representation of the glottal stop, known as the ***hamza***.



precedent vowel	lengthening letter	MSA script 1	phonetical script 2	English translation
[a]	ا	رَأْس [raʕs]	رَاس [ra:s]	‘head’
[u]	و	مُؤْمِن [muʕmin]	مُومَن [mu:man]	‘believer’
[i]	ي	دَافِئ [da:fiʕ]	دَافِي [da:fi:]	‘warm’

DiMorph

Postprocessing

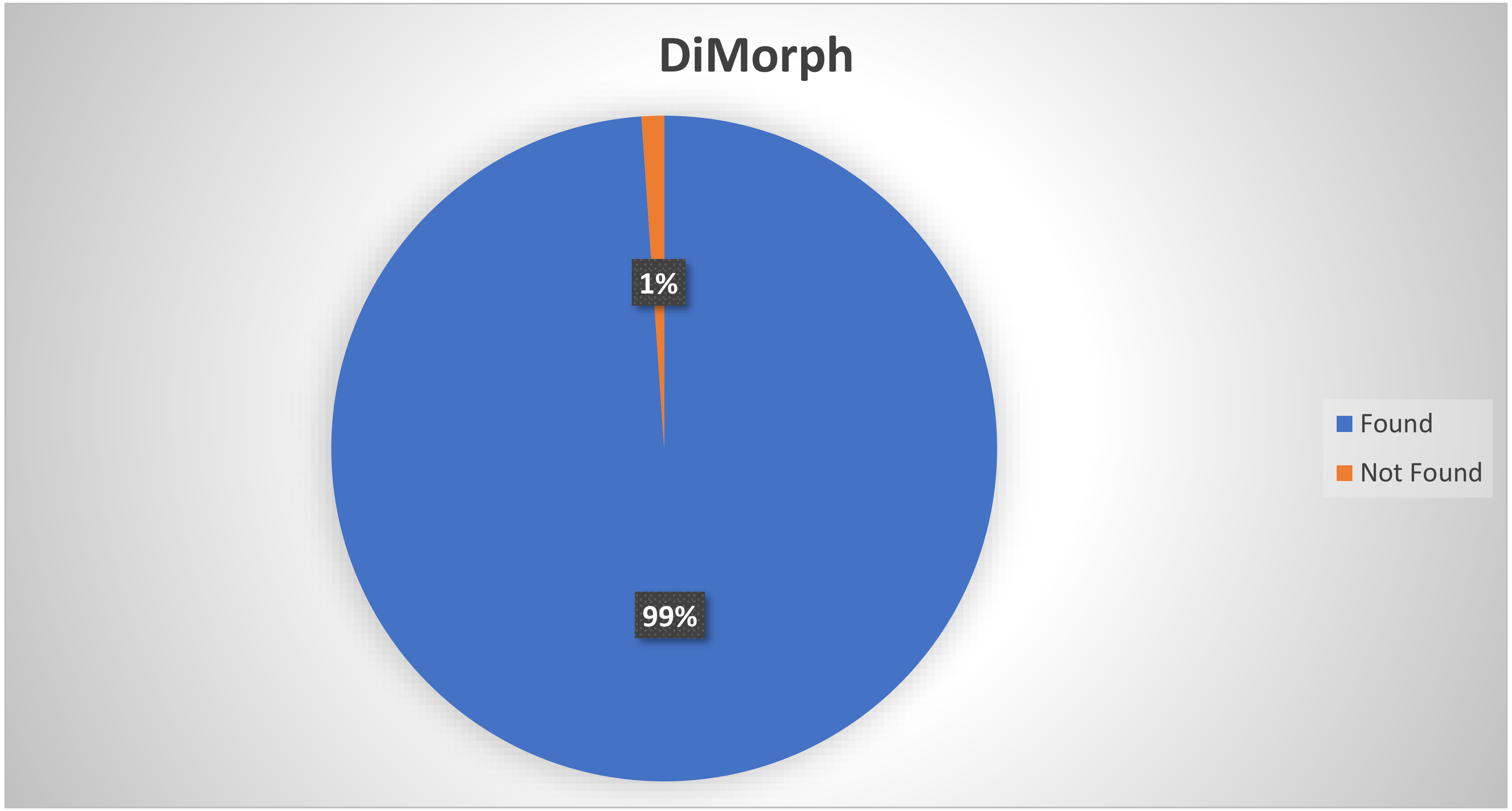
Manual correction

The spelling errors: were identified as 'Not Found' during analysis and subsequently they were corrected by the annotators.
For example:

- **تدحك /tadħak/** is corrected in **تضحك /tadˤħak/ `she laughs’.**
- **عندومشاكيل /ʕanduːmaʃaːkiːl/** is corrected in **عندو مشاكيل / ʕanduː maʃaːkiːl / `he has a problem’.**

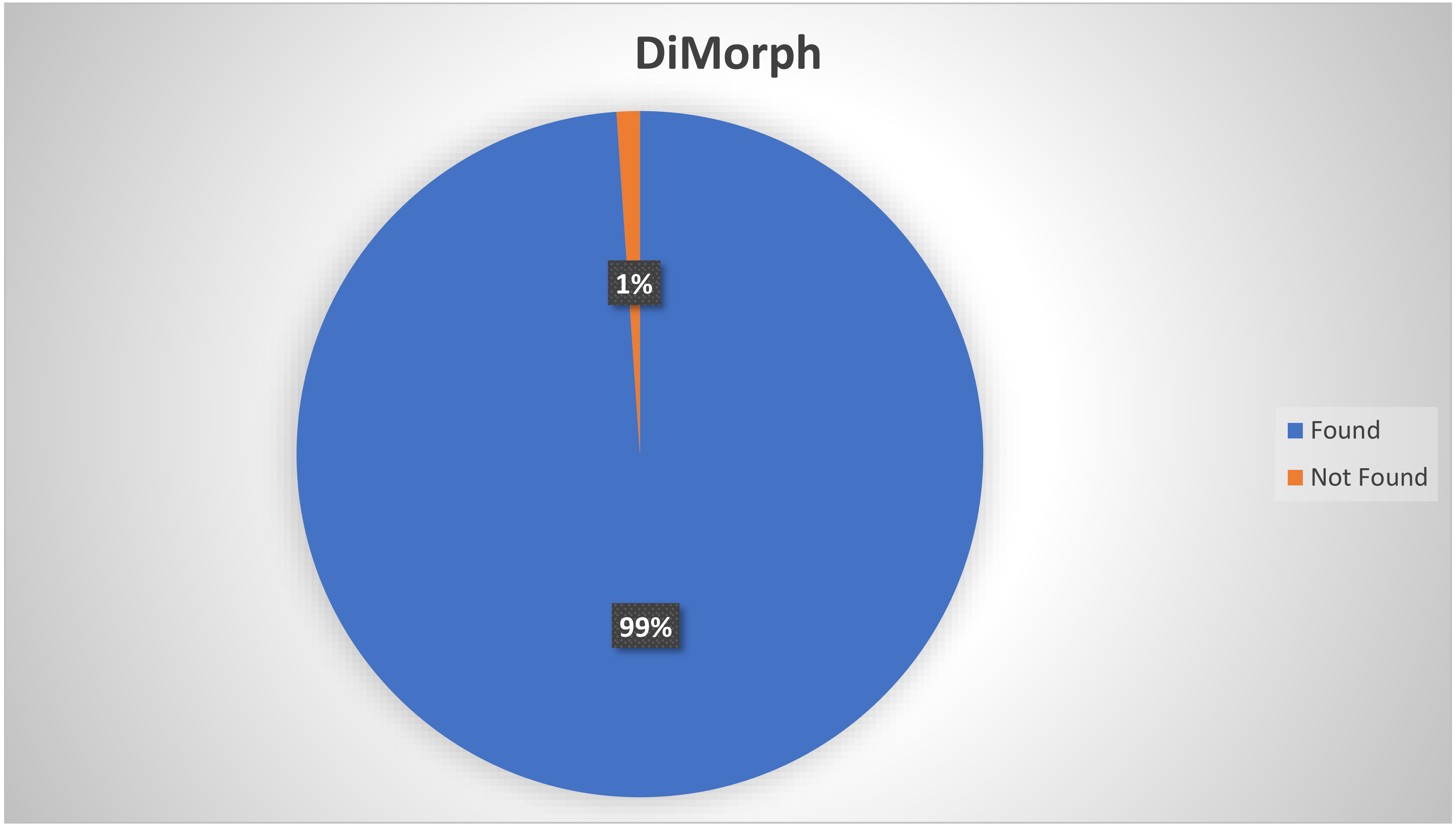
Results discussion

Size of the corpus is: 91.592 tokens



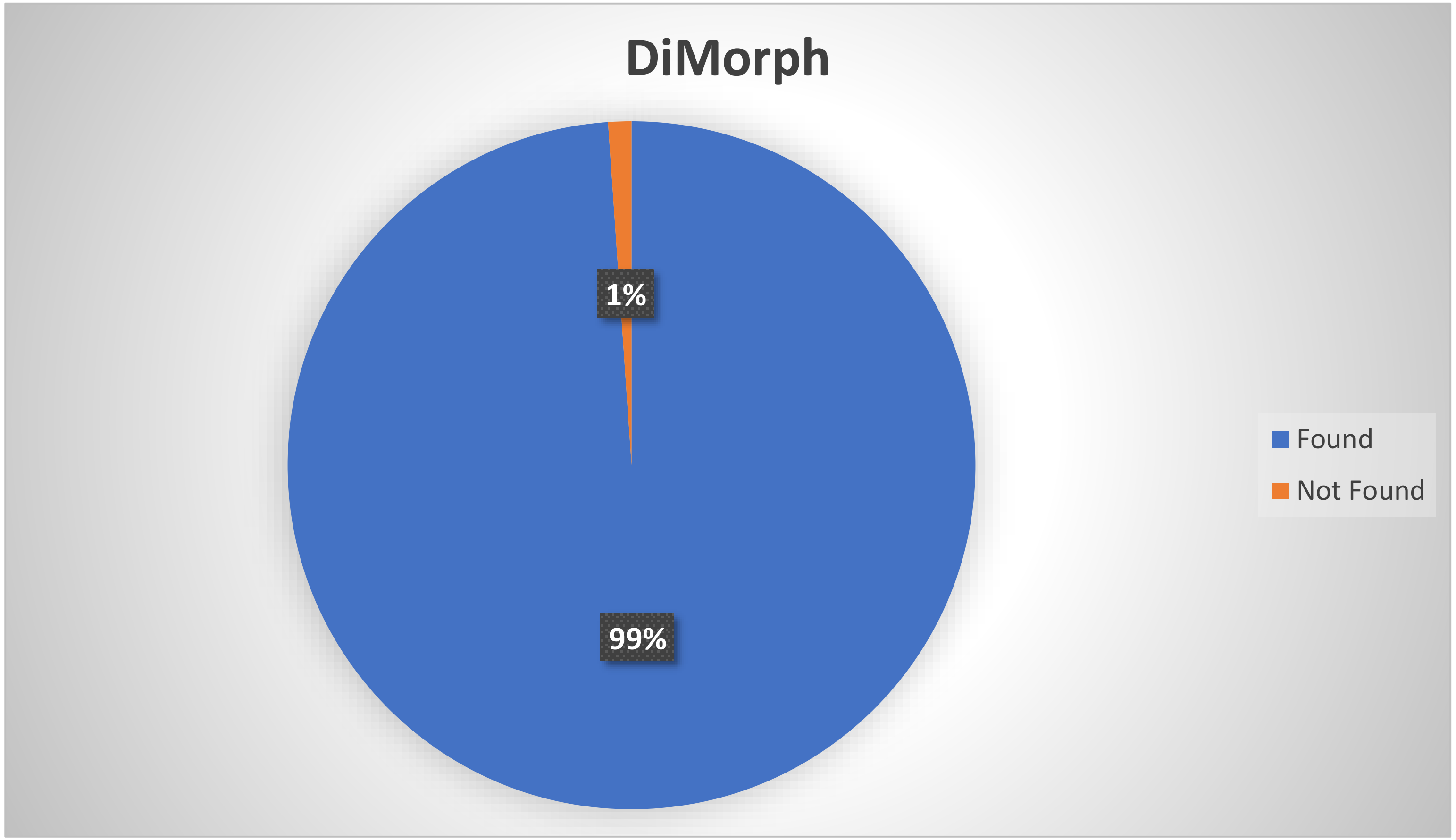
Results discussion

Size of the corpus is: 91.592 tokens

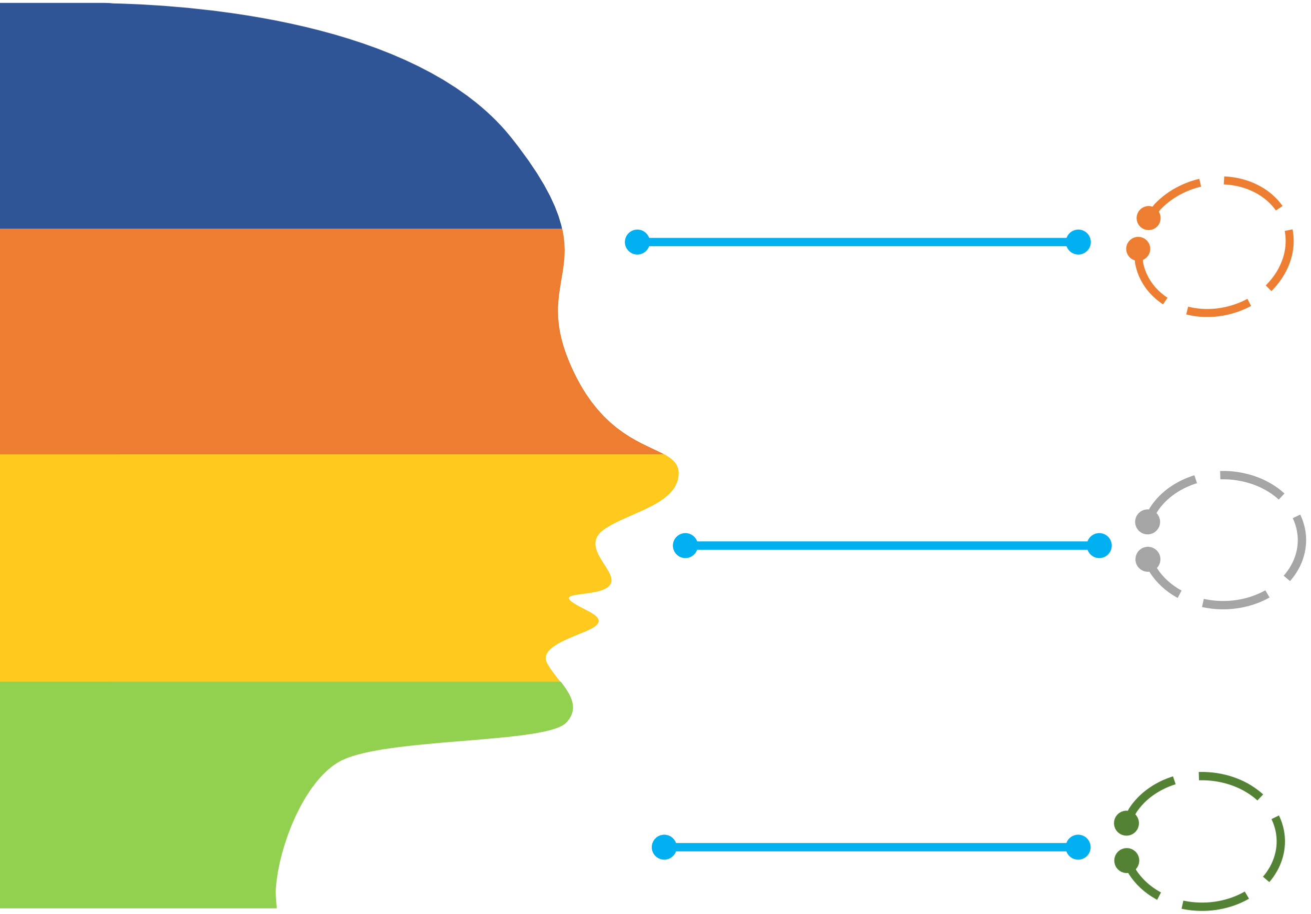


Results discussion

Size of the corpus is: 91.592 tokens



Conclusion



The creation of a morphological analyzer customized to dialectal characteristics.

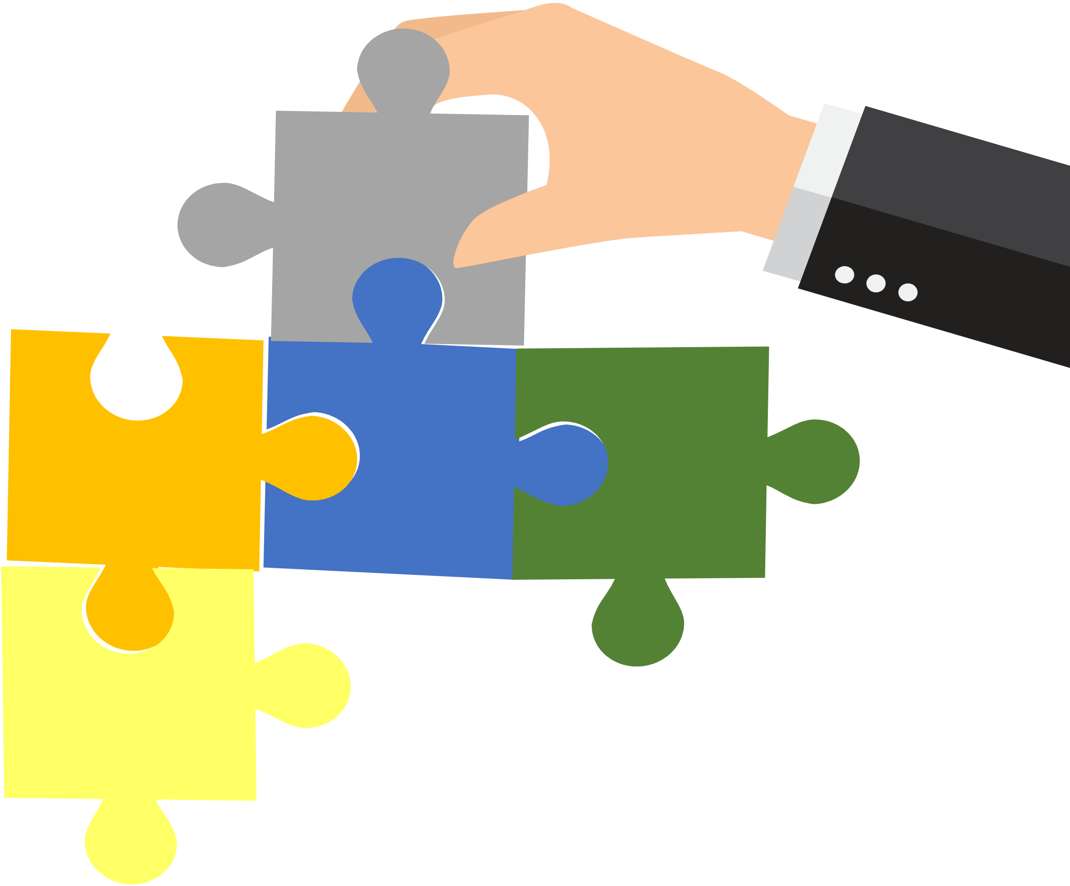
The construction of linguistic resources specifically for dialects.

The construction of corpus annotated and vocalized.

Perspectives

- Enrich the Moroccan DiMorph linguistic resources.

- Apply deep learning strategies to annotate the Moroccan corpus.



Istituto di Linguistica Computazionale "Antonio Zampolli"
Consiglio Nazionale delle Ricerche

كلية العلوم
Faculté des Sciences

Thank you for your attention

Nadia khlif	nadia.khlif@ilc.cnr.it
Giulia Benotto	giulia.benotto@ilc.cnr.it
Ouafae Nahli	ouafae.nahli@ilc.cnr.it