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

كلية العلوم
Faculté des Sciences

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

Nadia khlif nadia.khlif@ilc.cnr.it
Azzedine Mazroui azze.Mazroui@gmail.com
Ouafae Nahli ouafae.nahli@ilc.cnr.it

08/11/2024

PLAN

- Introduction
- Arabic language & Darija Dialect
- Linguistic features
- Darija's Challenge
- 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.

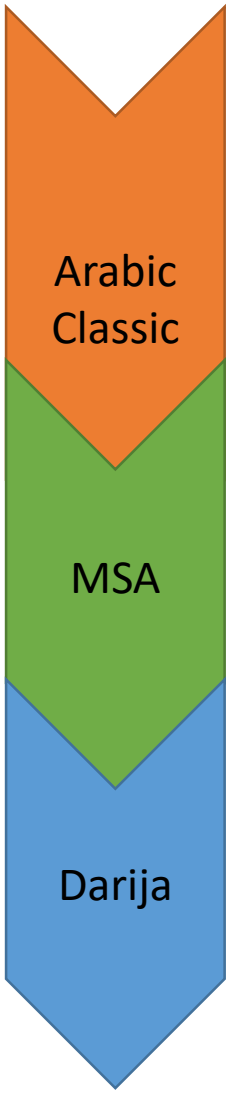
Modern Standard Arabic (MSA)

- 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



/qiTar/قطار
Camel Walking in sequence

/sayya:rat/ سيارة
Camel Walking in simultaneous

Word Time Shifts

/qiTar/قطار
Train

/sayya:rat/ سيارة
Car

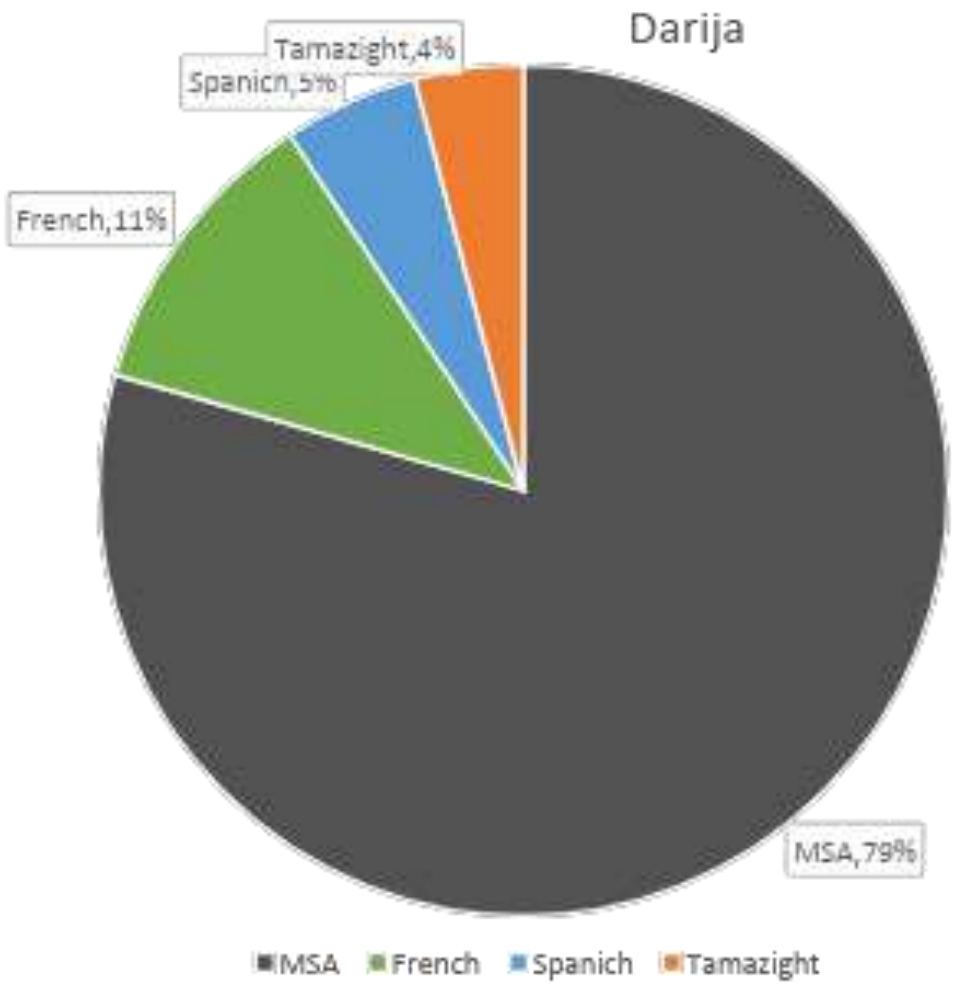
/tora:n/ تران
Train

/Tu:mu:bi:ɬ/ طوموبيل

/sayya:rat/ سيارة
Car

Lexical borrowing

Arabic Language & Darija dialect



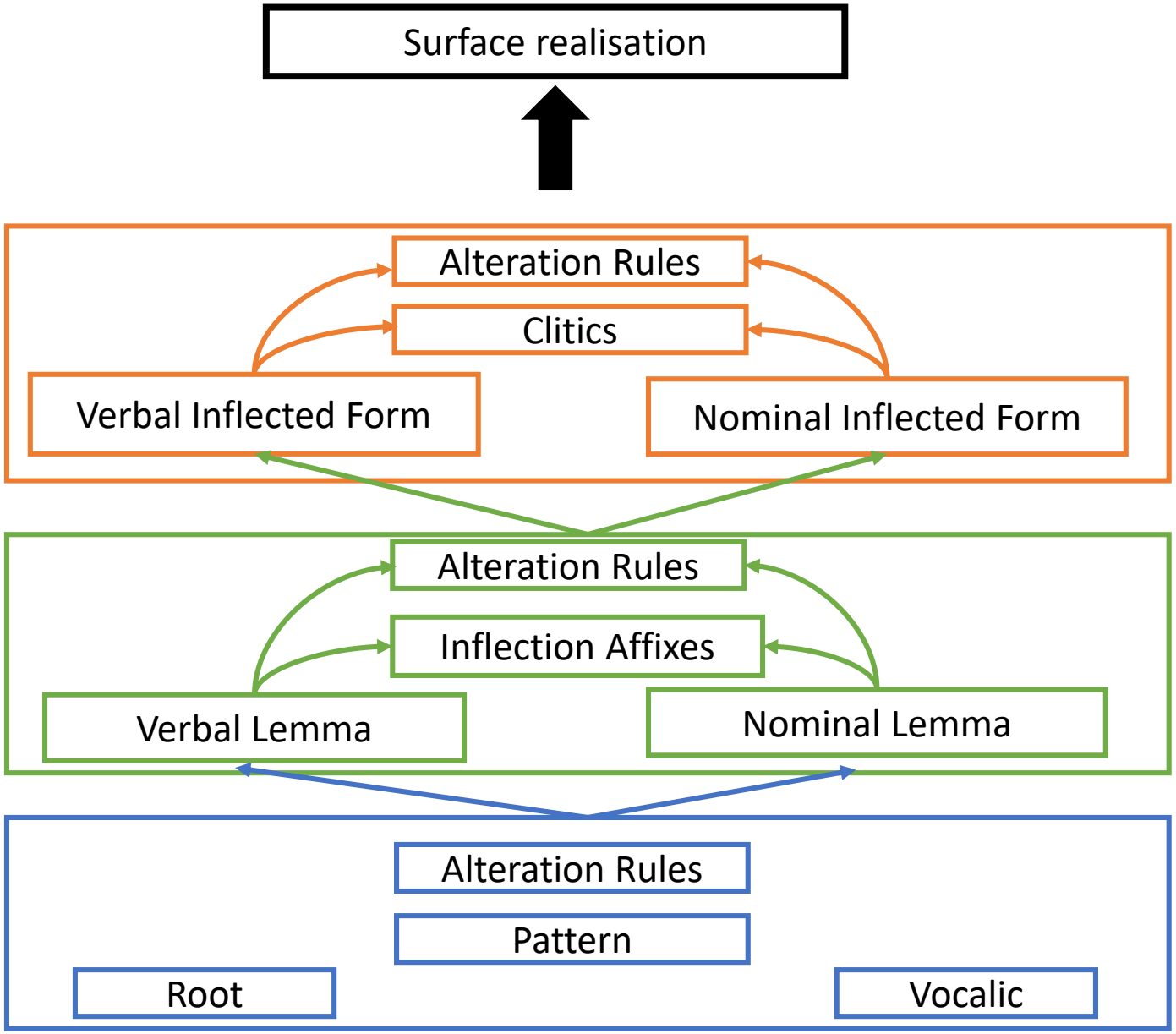
Article de reference:
Tachicart, Ridouane, Karim Bouzoubaa, and Hamid Jaafar. 2016. “Lexical differences and similarities between moroccan dialect and Arabic”.

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.

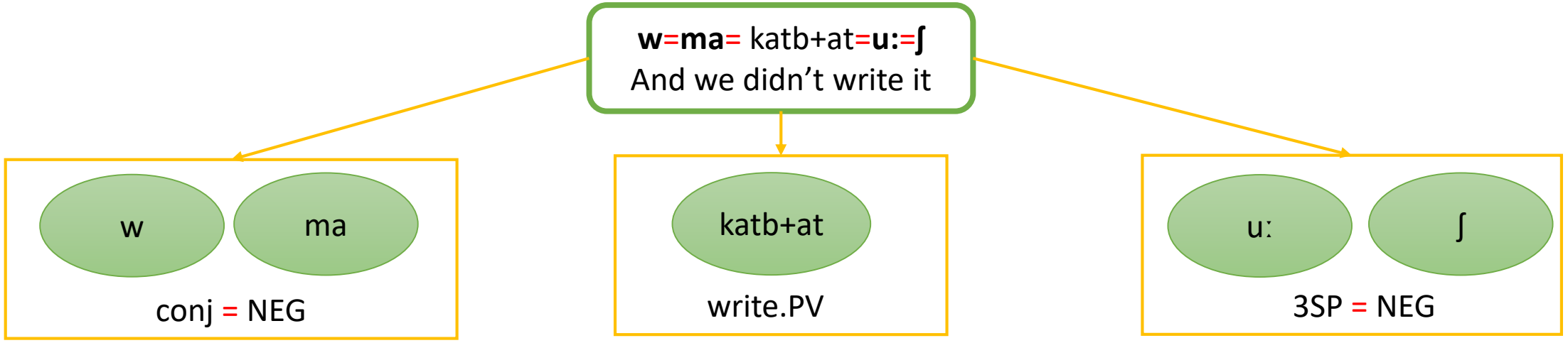
- The inflectional layer is the one where the lemma combines with inflectional affixes to give inflectional forms.

- 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.

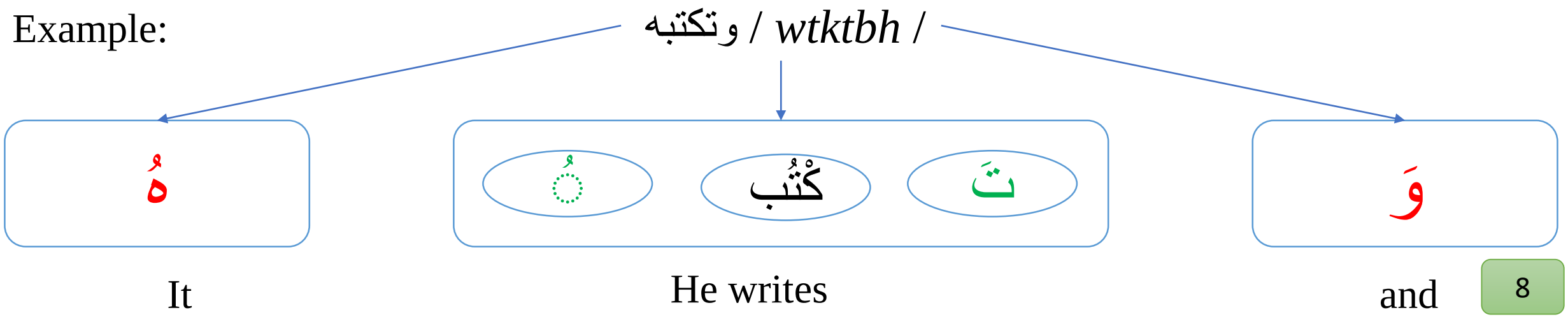


Linguistic features

Example of Grammatical Aspects

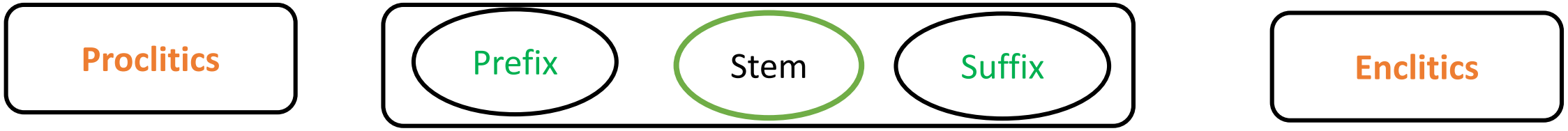


Example:

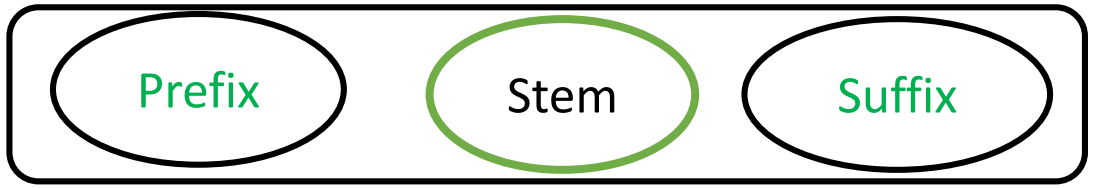


Linguistic features

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



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



Darija's Challenge

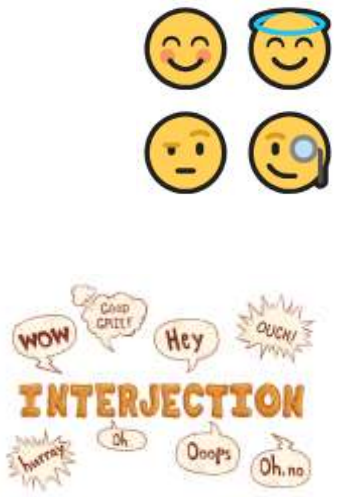
Poor resources

Challenges

Issues of Texts retrieved from social platforms

Inherent characteristic 's Daija

Darija’s Challenge



Issues of Texts retrieved from social platforms

- مبروووك /mabru:u:u:u:k/ ‘congratulation’

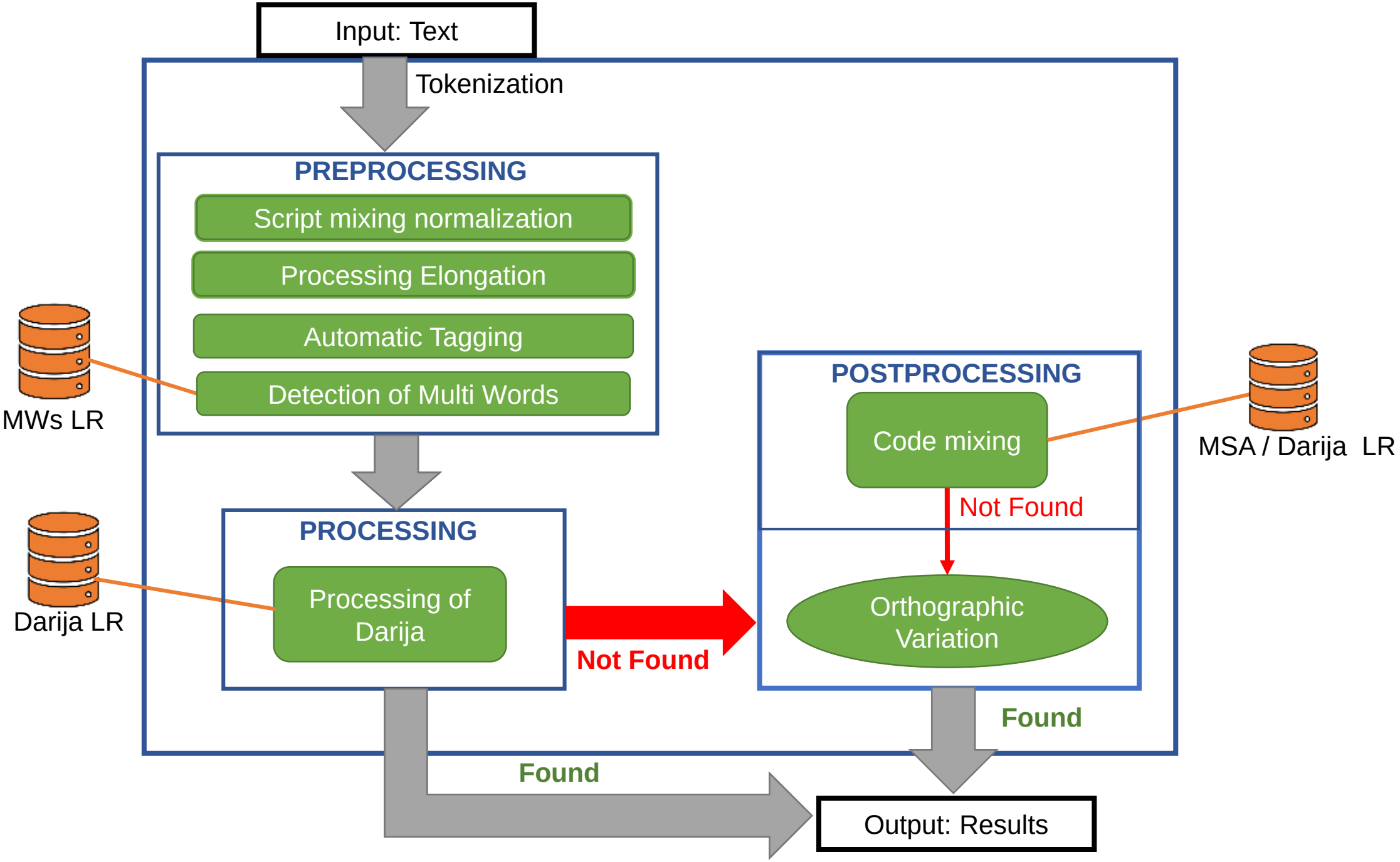
Darija’s Challenge

Inherent characteristic’s Daija

- **Lack of Orthographic and Grammatical Rules**: Due to the absence of established orthographic and grammatical rules, words are often written as pronounced, leading to variations.
  - راس , رءس: instead of the standardized رأس

- **Early Technology Limitations**: The lack of Arabic keyboards led users to adopt the Latin alphabet with numbers (Arabizi) to represent unique Arabic sounds, such as:
  - ع -> 3.
  - ح -> 7.

- **Use of Arabic Script with Limitations**: With Arabic keyboards now available, people primarily use Arabic script. However, some sounds in Darija, like **/g/** in **‘واطو g’** **“**cake" (a borrowed word from French "gateau“), aren't represented in standard Arabic script, leading to script mixing.


Input: Text
Tokenization
PREPROCESSING
Script mixing normalization
Processing Elongation
Automatic Tagging
Detection of Multi Words
MWs LR
PROCESSING
Processing of Darija
Darija LR
Not Found
POSTPROCESSING
Code mixing
Not Found
Orthographic Variation
MSA / Darija LR
Found
Found
Output: Results

DiMorph

Script mixing normalization

Processing of numeral characters:

When each number can be directly replaced by an Arabic letter, we apply automated substitution.

For example:

- فت7ها => فتحها /*ftaHħaː*/ **`he opened it'**

Processing of literal characters:

In case where the Latin letter "g" represents the phoneme /g/, which does not exist in standard Arabic script, we apply automated substitution.

For example:

- واطg => گاطو /gaːTuː/ **`cake'**

DiMorph

Processing elongation

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

DiMorph

Automatic tagging

Implementation of an automatic tagging system to identify:

- Punctuation/ number.
- Emoticons.
- Interjections.
- Word Foreign.

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

For example:

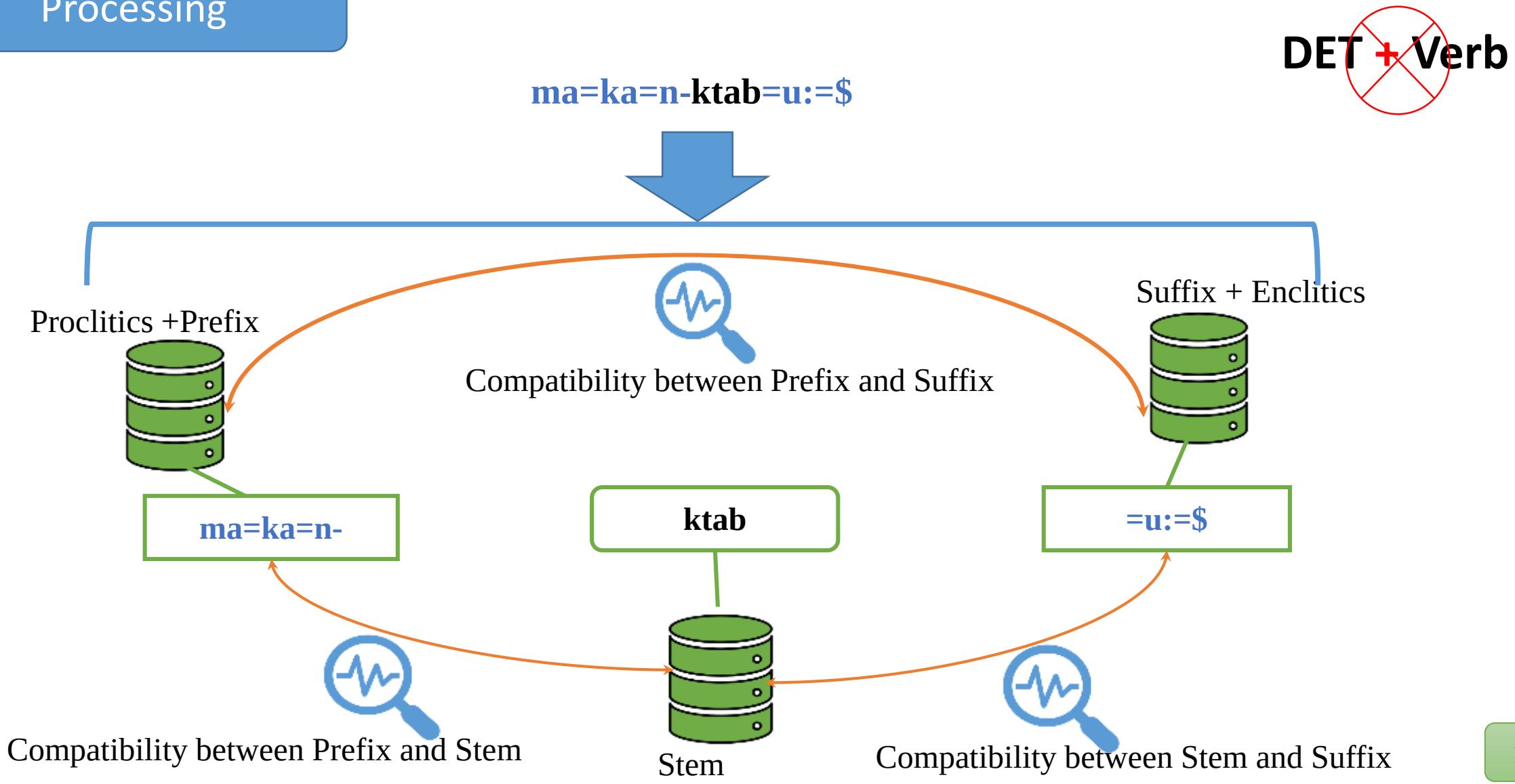
- لPaola /l=Paola/ : ل/PREP+WORD_FOREIGN `for Paola'.
- ب5000 /b=5000/ : ب/PREP+NUMBER `with 5000'.

DiMorph

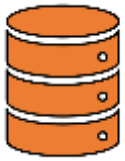
Detection Multi Words

DiMorph

Processing



DiMorph



Linguistic Resources & Processing

DictStem	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 + enclitics	Prefix=Stem	Stem=Suffix	Prefix=Suffix
297	570	770	780	1067

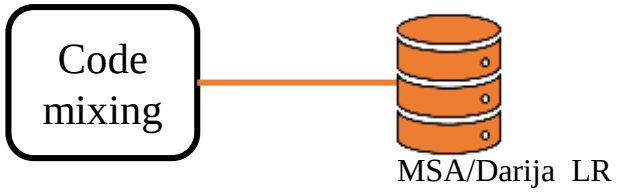
DiMorph

Postprocessing

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

- Standardize written forms according to orthographic rules, ensuring that spelling variations are unified into a single, consistent form.

- راس /ra:s/
- رءس /raas/

'Head'

رأس /raa:s/

'Head'

Results & discussion

Evaluation

- **INV Rate (In-Vocabulary Rate):** Measures the percentage of tokens successfully analyzed by the system.
- **OOV Rate (Out-of-Vocabulary Rate):** Measures the percentage of tokens the system could not analyze.

DiMorph using	Total Tokens	INV rate	OOV rate
With Preprocessing and Postprocessing	105 064	**96%**	3,95%

Results & discussion

Evaluation

Conclusion

DiMorph's Analysis Capacity

Average Analyses per Token:

- DiMorph provides, on average, 2.45 possible analyses for each token.
- This metric highlights the system's current capability to generate multiple solutions per token, indicating flexibility in interpretation.

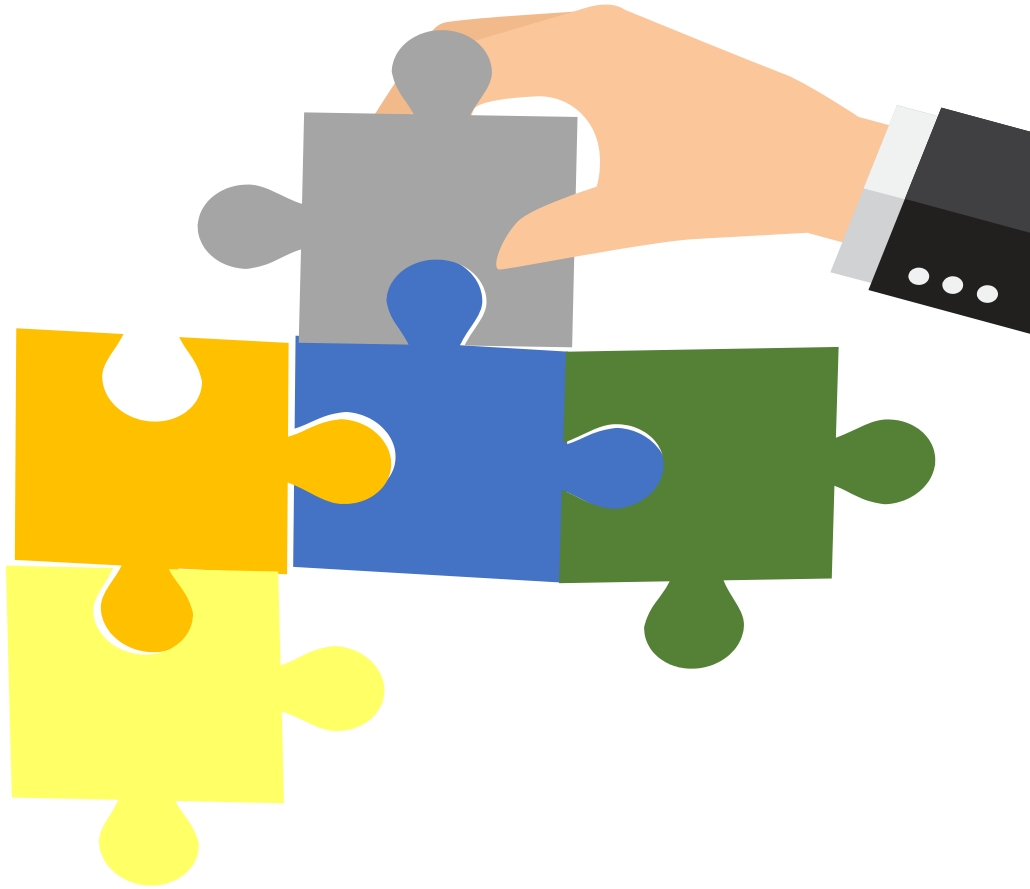
Challenges:

- Out-of-Context Issue: While DiMorph generates multiple analyses, it often lacks the ability to determine the correct analysis in context, leading to potential ambiguity.

Perspectives

Enrich the Moroccan DiMorph linguistic resources.

Apply deep learning models to provide context-aware solutions and accurately 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
Azzedine Mazroui azze.Mazroui@gmail.com
Ouafae Nahli ouafae.nahli@ilc.cnr.it