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

كلية العلوم
Faculté des Sciences

A Robust Morphological Analysis System for the Moroccan Dialect: Design and Evaluation

Nadia Khlif nadia.khlif@ump.ac.ma
Azzeddine Mazroui azze.Mazroui@gmail.com
Ouafae Nahli ouafae.nahli@ilc.cnr.it

25/04/2025

Introduction

Arabic Language & Darija Dialect

Darija’s Challenge

DiMorph

Results & Discussion

Conclusions & Perspectives

Introduction

- Arabic NLP faces significant challenges due to diglossia with Modern Standard Arabic (MSA) used in writing and dialects in spoken communication.
- With social media, dialects now appear widely in written form, but their lack of standardization and limited resources hinder NLP development.
- To address this gap, our work within the *CWALM* project focuses on developing tools for Contemporary Written Arabic.
- As part of this effort, we have developed **DiMorph**, a morphological analyzer specifically designed to dialectal Arabic.

Arabic Language & Darija dialect

Classical Arabic

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

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
- Lack of the formal recognition and standardized grammar rules

Arabic Language & Darija dialect

Classical Arabic

MSA

Darija

/qiTar/قطار
Camel walking in sequence

/sayya:rat/سيارة
Camel walking in simultaneous

Word Time Shifts

/qiTar/قطار
Train

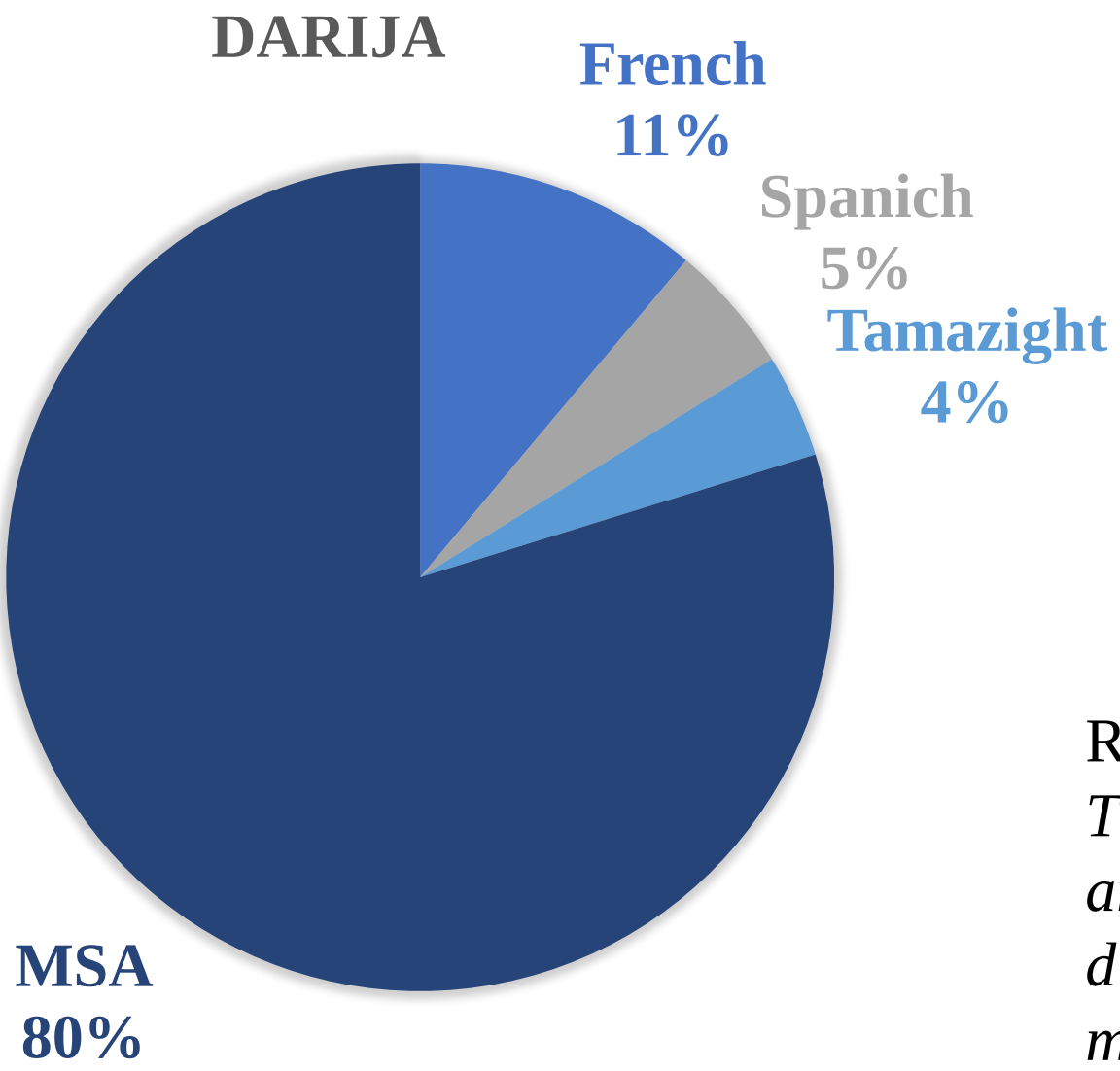
/tora:n/ تران
Train

/sayya:rat/ سيارة
Car

/Tu:mu:bi:ł/ طوموبيل /sayya:rat/ سيارة
Car

Lexical borrowing

Arabic Language & Darija dialect



Reference article:
Tachicart, Ridouane, Karim Bouzoubaa, and Hamid Jaafar. 2016. "Lexical differences and similarities between moroccan dialect and Arabic".

Darija's Challenge

Poor resources

Challenges

Issues of Texts retrieved from social platforms

Inherent characteristic's Daija

Darija's Challenge

Poor resources

- The lack of available linguistic resources and annotated corpora significantly hinders progress.

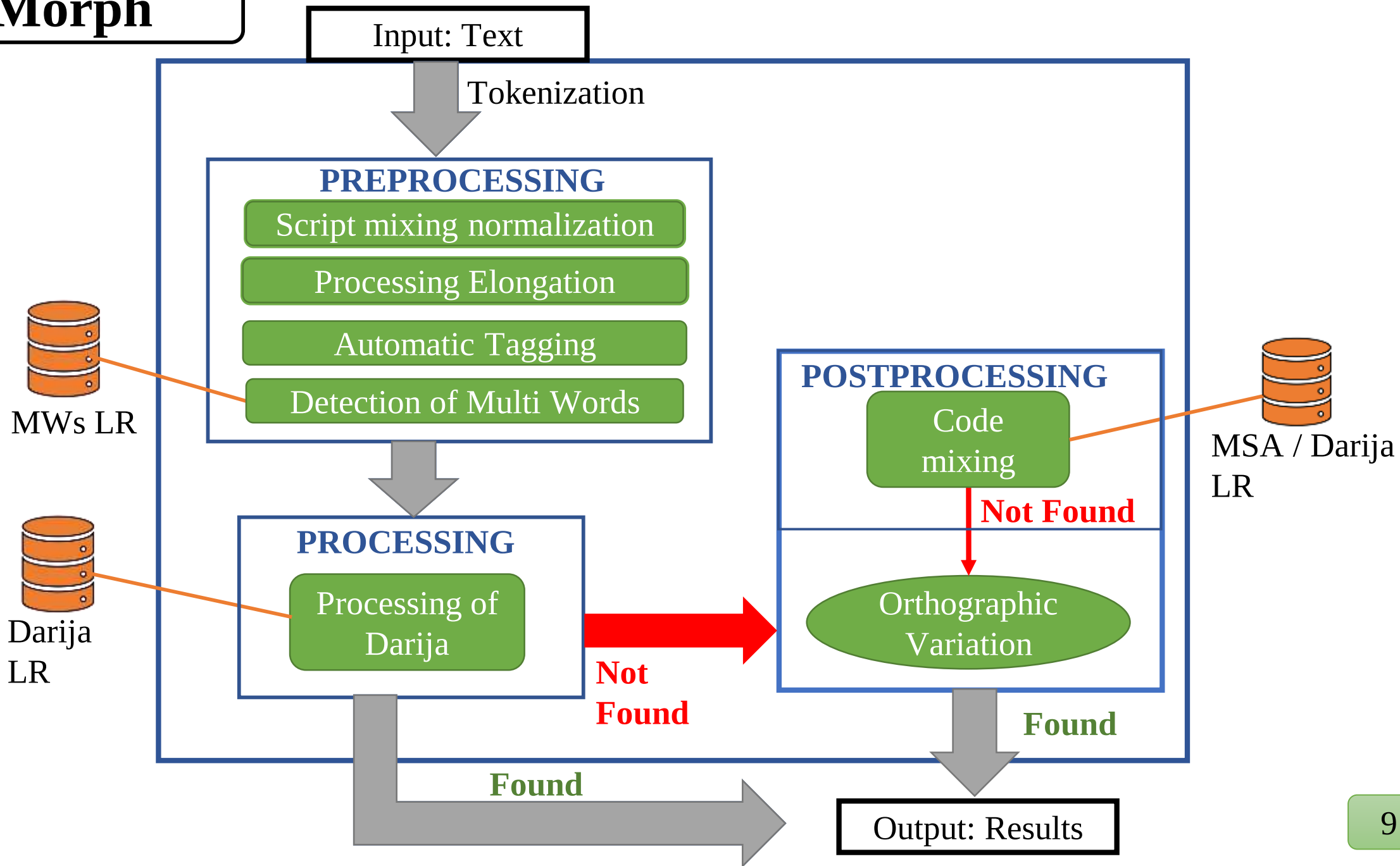
Issues of Texts retrieved from social platforms

- Elongation: مبرووووك'congratulation'
- Interjection: اوف'ohf'
- Emoticon.

Inherent characteristic's Daija

- Lack of Orthographic and Grammatical Rules:
  راس, رءس, رأس
- Early Technology Limitations: 7 <- ح ,3 <- ع
- Code/Script-mixing: واطg

DiMorph


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
MSA / Darija LR
Not Found
Orthographic Variation
Found
Found
Output: Results

DiMorph

Script mixing normalization

Numeral character processing:

For example:

- فت7 ها => فتحها/*ftaHha:*/ **`he opened it'**

Literal character processing:

For example:

- طاg => طاگ/ga:Tu:/ **`cake'**

Elongation processing

- **بزاااف bazza:a:f => بزاف '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

Multi-Word detection

Detection and segmentation

قولي اسي أحمد اش طرى وجرا في قضية الفقيه بن صالح

Tell me Mr Ahmed what happened concerning the Sidi Kacem case

الفقيه بن صالح

Fquih Ben Salah

Compound proper noun

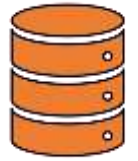
اش طرى وجرا

what happened

Multi-word expression

DiMorph

Multi-Word detection



MWs LR

	Multi-Word expressions	Compound proper nouns
Number of Units	153	511

DiMorph

MWs LR

Multi-Word detection

الفقيه بن صالح
Fquih Ben Salah

Compound proper noun

اش طرى وجرا
what happened

Multi word expression

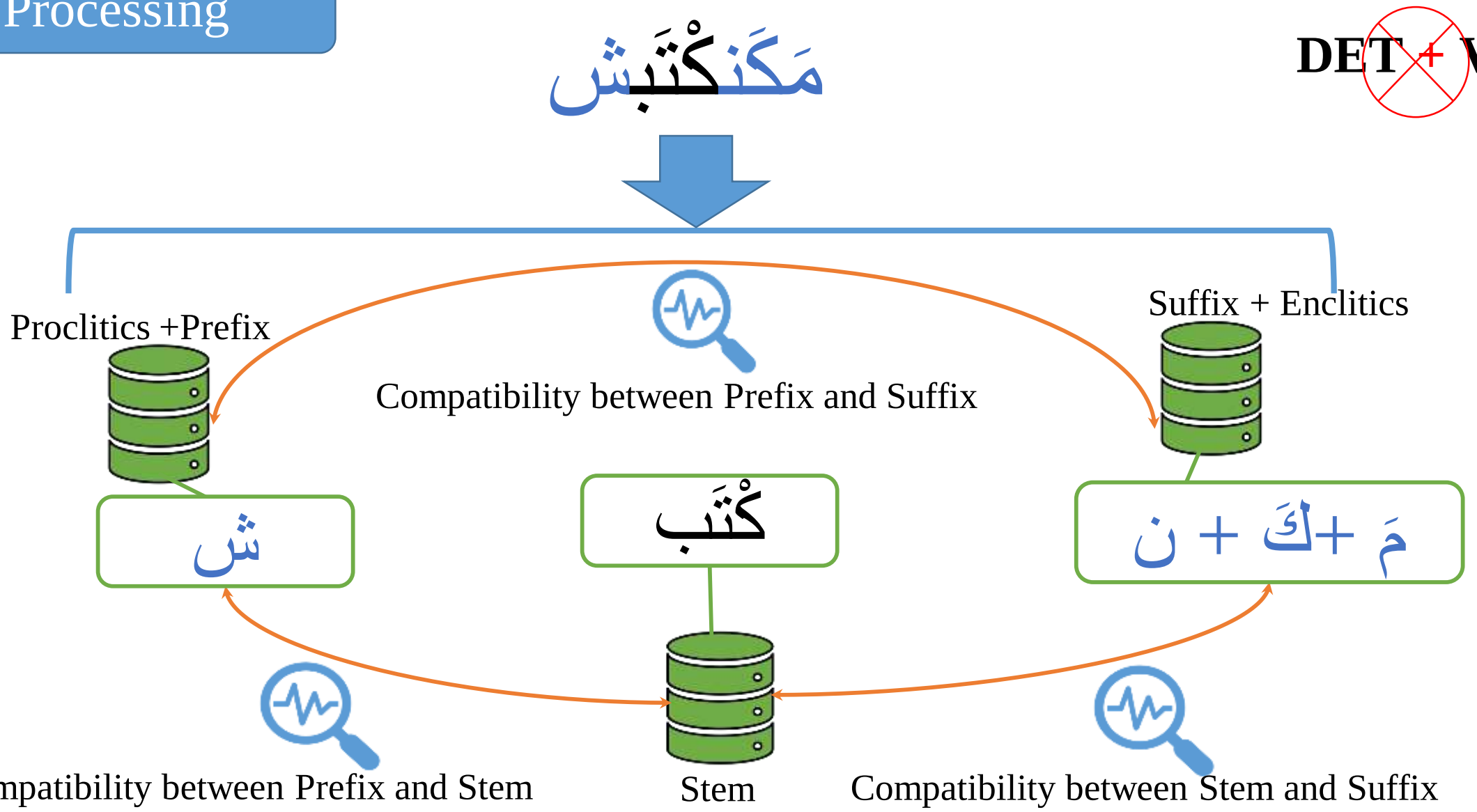
Analyzing

```
TOKEN:  الفقيه بن صالح  Alfqyh bn SAlH
    Arabic-msd: =الفْقِيهْ/B-GPE:CITY=/NSUFF+
    Buckwalter-msd: =lafoqiyh/B-GPE:CITY=/NSUFF+
    Arabic-msd: =بَنْ/I-GPE:CITY=/NSUFF+
    Buckwalter-msd: =ban/PROPN_I-GPE:CITY=/NSUFF+
    Arabic-msd: =صَالَحْ/I-GPE:CITY=/NSUFF+
    Buckwalter-msd: =SAlaH/I-GPE:CITY=/NSUFF+
Vocalised PROPN:    lafoqiyh ban SAlaH
                    الفْقِيهْ بَنْ صَالَحْ
Glosses:    Fquih Ben Salah+

TOKEN:  اش طرا وجرا A$ TrA wjrA
Vocalised Token:    أش طْرَا وَجْرا
Arabic-msd:     أش طْرَا وَجْرا/MULTIWORD+
Buckwalter-msd: >a$ ToraA wajorA/MULTIWORD+
Glosses:    what happen?
```

DiMorph

Processing



DiMorph

Darija LR

Linguistic Resources & Processing

DictStem	NOUNS	VERBS	ADJECTIFS	ADVERBS	PROPER NOUNS	PRONOMS	FUNCTION WORDS
Darija	8760	11647	1674	214	931	92	153
Foreign	675	122	33	8	100	-	3

DictPrefix	DictSuffix	Compatibility Tables		
proclitics + prefixes	suffixes + enclitics	Prefix=Stem	Stem=Suffix	Prefix=Suffix
394	641	770	780	1067

DiMorph

Postprocessing


MSA / Darija LR

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

For example:

كيستفزها /kayastafazzha/ “He is provoking her”

DiMorph

Postprocessing

Orthographic Variation

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

Type	Variation	Example
Hamza Normalization	unified form ← ء, أ, ئ, إ, ؤ	سأل ← سءل, سال
Phonological Shifts	ث ← ت	ثلاثة ← تلاتة
	ذ ← د	ذهب ← دهب
	ظ ← ض	ظلام ← ضلام
Word-Final Changes	و ← ه	عنده ← عندو
	ة ← ا	ركبة ← ركبا

Results & discussion

Evaluation

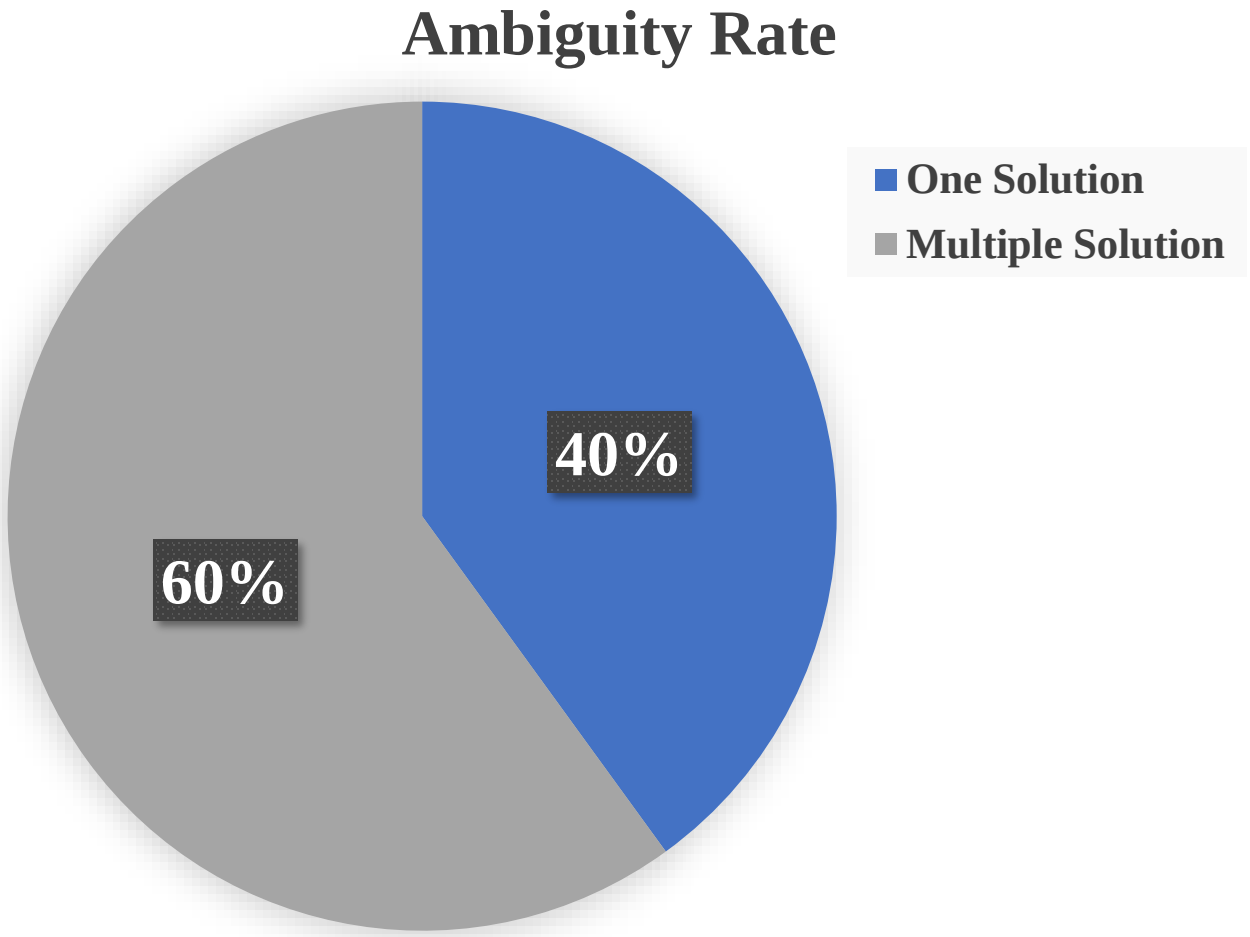
- **In-Vocabulary Rate (INV Rate):** Measures the percentage of tokens successfully analyzed by the system.
- **Out-of-Vocabulary Rate (OOV Rate):** Measures the percentage of tokens the system could not analyze.

	Total Tokens	**INV rate**	**OOV rate**
DiMorph with Preprocessing and Postprocessing	11 085	**97.84%**	2.16%

Results & discussion

Evaluation

Further statistical analysis reveals that:



- 89.71% of the cases corresponds to instances of homography.
- 9.31% of the cases corresponds to instances of polysemy.
- 0.98% of the ambiguity is due to the absence of vocalization.

Conclusion

DiMorph's Analysis Capacity

Average Analyses per Token:

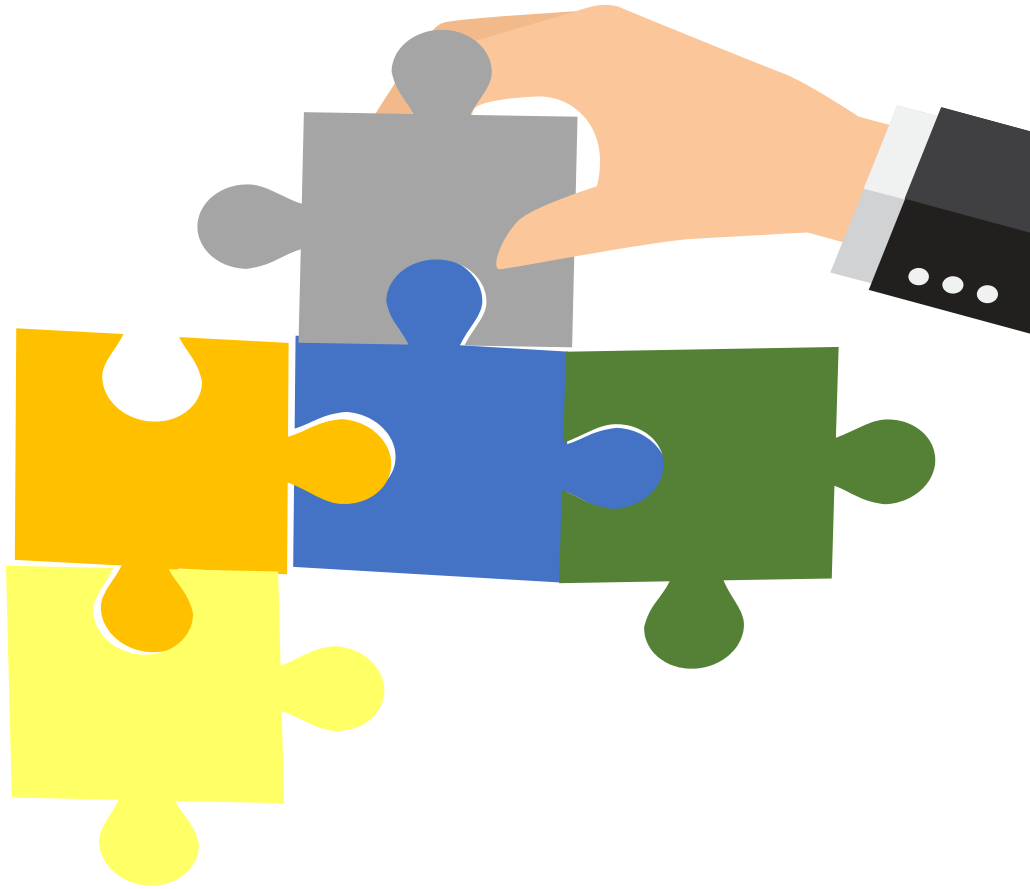
- DiMorph provides, on average, 1.45 possible analyses for each token.
- This metric highlights the system's current capability to generate multiple solutions per token, indicating flexibility in interpretation.

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@ump.ac.ma
Azzeddine Mazroui azze.Mazroui@gmail.com
Ouafae Nahli ouafae.nahli@ilc.cnr.it