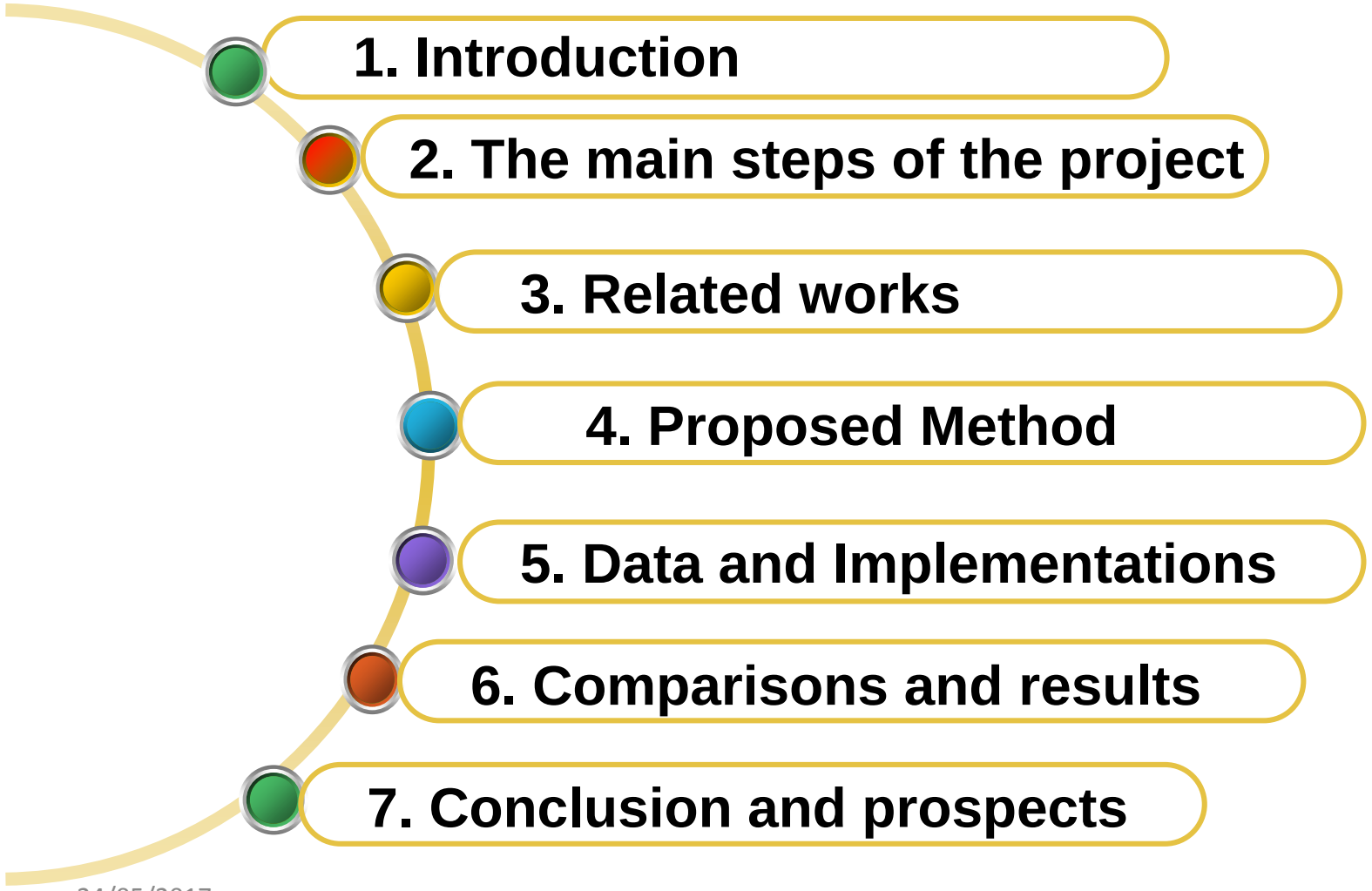
The formalization of the Maçdar category (V-n) by their patterns using NooJ platform

International Conference Nooj 17
Kenitra May 18-20 2017

by:

- BOUNOUA Ahmed - Dr. ZINEDINE Ahmed - Dr. EL HANNACH Mohamed & Dr. KASMI Rachida

Outline


1. Introduction
2. The main steps of the project
3. Related works
4. Proposed Method
5. Data and Implementations
6. Comparisons and results
7. Conclusion and prospects

Introduction

- Linguistic resources

- Today, there is a huge amount of Arabic linguistic resources, but the major problem is the absence of a formal framework that formalizes and exploits these resources
- The main goal of our project is the elaboration of an electronic dictionary in a theoretical framework

The main steps of project

The main steps of project

Corpora analysis by patterns

- morphological, syntactical and semantical information are reserved in patterns.

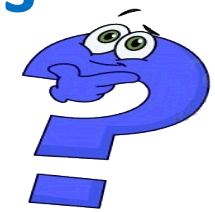
Adoption of lexicon-grammar framework

- Researchers all over the world have adopted the Gross model of description, which serves as a computational model for any language.
(NooJ is based on this framework)

The main steps of project

Process all Arabic grammatical categories

- Masdars المصادر (Verbal Noun)
- Verbs
- Adjectives
- Nouns



Why Masdars?

- The most productive category :
  136 patterns
  29 000 forms
  Millions inflected forms
- Category not systematic
- Arabic researchers don’t give importance to this category : the most researchers have confusion between nouns, adjectives and Masdars.

Masdar and Inflections :

- Basic form V-n (المصدر الأصلي : asl) : action without time or aspect. (ضَرْبٌ)

Inflected forms :

- V-n mer : (مصدر المرة) : number of actions. (twice : ضَرْبتان)
- V-n hay : (مصدر الهيئة) : represented by one pattern فِعْلَة and refers to the manner of the action (ضِرْبَة)
- V-n mim: (المصدر الميمي) : is the same of basic Masdar , the unique difference between those two items is the Pref = حرف الميم
- V-n sin : (المصدر الصناعي) : (profession) It is any masdar or noun with suffix=الياء, for example:

صناعة >>>>>>> صناعية

Statistics: Arabic lexicon entries

Category	Patterns	Basic items	Inflected items
V-n	136	33 600	168 000
V	60	20 000	650 000
V-a	120	60 000	720 000
N	160	80 000	620 000

Roots	
3 Letters	9600
4 Letters	1800

Morphological Arabic System

Arabic features:

- Morphology: 2,500,000 basic words
- Fusionist system :
  Arabic word = Prefix + Infix + Suffix >> 320,000,000 words
  The infix have several formats : irregular
- Each Arabic word should be associated to a pattern الوزن

Morphological Arabic System

- Infix

Infix	Lemma	Pattern
كتابتان	كتابة	فِعَالَة
رمايات	رماية	
كِنايَة	كناية	
فِراسَة	فراسة	
رعايتنا	رعاية	

In this example one pattern represent 5 lemmas.

Related works

Related works

EL-DICAR: (Mesfar, 2008)

- EL-DICAR (ELectronic DICtionary for ARabic) :
  the only Arabic dictionary available for NooJ platform, which allows to link all the inflexional, morphological, syntactico-semantic information to the list of lemmas.
- Including more than 52 000 lexical entries : nouns (N), verbs (V), adjectives (ADJ), particles (PREP, ADV, REL, DEM), localizations (N+LOC), First names (N+Prenom)

Related works

- The morphological Analyzer AlKhalil (BOUDLAL, 2010)

  - The system was developed in collaboration with the Arab League Educational, Cultural and Scientific Organisation (ALECSO) and King Abdul Aziz City for Science and Technology (KACST).
  - The system can process non vocalized texts as well as partially or totally vocalized ones.
  - It is based on modelling a very large set of Arabic morphological rules, and also on integrating linguistic resources that are useful to the analysis, such as the root database, vocalized patterns associated with roots, and proclitic and enclitic tables.

Related works

- Arabic WordNet (AWN) (2006)

  - The ArabicWordNet effort started in 2006 through a collaboration of several universities and companies .

  - Arabic words in AWN are represented in terms of their lemmas ; this representation is inspired from English WordNet.

Related works

- Drawbacks:

EL-DICAR	AlKhalil	Arabic WordNet
• The lemma is considered as the base of each lexical entry, it is the same principle of Latin and European languages. • Absence of the category Vn	• It is based on a set of linguistic resources. These resources are available just for AlKhalil. • No explicit dictionary	• Arabic words in AWN are represented in terms of their lemmas. • This representation is inspired from the representation of English WordNet

Proposed method

Proposed method

We propose a new method for the formalization of the category Masdar (V-n) by :

- The construction of a new dictionary for Masdars with NooJ format.
- The focus of this formalization is root +pattern.
- Linking each Masdar with its associated verb (s).
- Associating each V-n with its flexional derivation:

V-n Asl, V-n mim, V-n mer, Vn- hay, & V-n sin

Proposed method

- Advantages :

  - Reduce the size of dictionary (reduce number of NooJ operators)
  - Search words by pattern
  - Have a link between the category Masdar and verbs, which will be very important at syntactic analysis level

Data and Implementations

- Excel format:

الرقم الترتيبي	الجذر	وزن الفعل	المصدر	وزن المصدر	وزن الهيئة	وزن المرة	وزن المصدر الميمي	المصدر الصناعي
155	كتب	فَعَلَ*يَفْعُلُ	كَتْب_كِتاب_كِتابَة	1_8_15	فِعْلَة	فَعْلَة	مَفْعَل	المصدر+يَّة
		فَعَّل	تَكْتِيب	19	_	ة+	مُفَعَّل	المصدر+يَّة
		افْتَعَلَ	اكْتِتاب	22	_	ة+	مُفْتَعَل	المصدر+يَّة
		اسْتَفْعَلَ	اسْتِكْتاب	26	_	ة+	مُسْتَفْعَل	المصدر+يَّة
		فاعَلَ	مُكاتَبَة	20	_	واحدة+	مُفاعَل	المصدر+يَّة
		أَفْعَلَ	إكْتاب	17	_	ة+	مُفْعَل	المصدر+يَّة
		تَفَعَّلَ	تَكَتُّب	24	_	ة+	مُتَفَعَّل	المصدر+يَّة
		تَفاعَلَ	تَكاتُب	23	_	ة+	مُتَفاعَل	المصدر+يَّة

Data and Implementations

- NooJ format (.dic):

```
1 بِنَايَة,VN+as1+فِعَالَة+FLX=FlexT
2 بِنَايَة,كتب,NW+DRV=فِعْلَة000:FlexT+DRV=فِعْلَة000:FlexT+DRV=مَفْعَل000:Flex
3 بِنَايَة,VN+DRV=3صناعي:FlexT
```

Data and Implementations

- NooJ format (.nof):

  - Derivations are defined according to the pattern and the type of root. We get the type of root by the type of each letter of the root. Each letter is represented by one number from 0 to 3 (0 for healthy letter, 1 for Hamza, 2 for letter و, 3 for letter ي) and the number 4 to mention if there is a doubled letter in the last of root.

```
مُفعَل001=
<LW>مُ<R>ا<R><S>أ<SW>/VN+mim+مُفعَل;
```

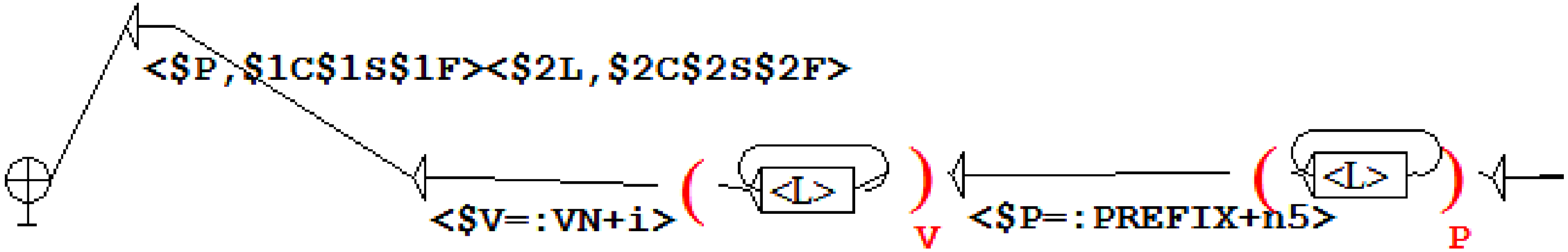
Data and Implementations

- Grammar (.nom):

General structure

Data and Implementations

- A piece of Grammar :

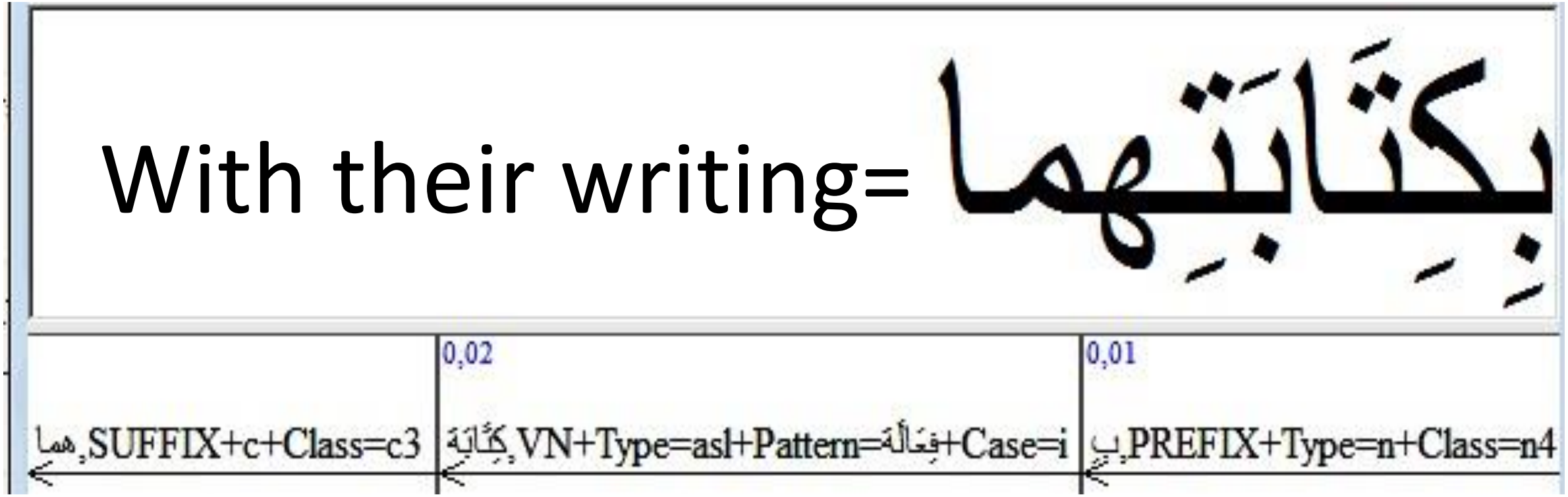


- This short graph means that the prefixes of category n5 (وَبِالْ, وَلِلْ , أَبِالْ) admit just the genitive case of Masdar (VN+i) and without any suffix.

Data and Implementations

- Test example

With their writing= بِكِتَابَتِهما



Evaluation

- Due to the unavailability of a test corpus which represents words with the category Masdar, it is actually hard to evaluate our dictionary (Erfan).

- Therefore, we chose to do many comparisons with El-DicAr dictionary and the morphological analyzer AlKhalil.

Results and Comparison

- The new dictionary (Erfan) contains about 29 000 original Masdars.
- El-DicAr recognize just 36.45 % (all words recognized are annotated as Nouns or Adjectives)
- AlKhalil recognize 86% of Masdars.

Results and Comparison

- We also analyze three different text corpus:
  - Vocalized corpus (Quran: 17712 different words)
  - non-vocalized corpus:
    corpus 1: 47635 different words.
    corpus 2 (Watan-2004): 260 000 different words
- Link of corpus :
  https://sites.google.com/a/aucegypt.edu/infoguistics/directory/Corpus-Linguistics/arabic-corpora

Results and Comparison

Corpus		Erfan	AlKhalil
Vocalized		25%	23%
Non-Vocalized	Corpus 1	41.48%	36.71%
	Corpus 2	41.35%	37.48%

Conclusion and Prospects

We have elaborated an electronic dictionary in the context of lexicon-grammar approach by exploiting the characteristic of the pattern in the Arabic.

As perspectives, we hope to rich the dictionary by others categories (Verbs, nouns, Adjectives) to prepare the ground towards the syntax

Thank you for your listening