International Conference NooJ 2020 Zagreb, Croatia, June 5 - 7

Automatic Recognition and Translation of Tunisian Dialect Named Entities into Modern Standard Arabic

Roua Torjmen and Kais Haddar

rouatorjmen@gmail.com kais.haddar@yahoo.fr

Miracl Laboratory, University of Sfax, Tunisia

Outline

- Introduction
- Previous work
- Linguistic study
- Proposed method
- Experimentation and evaluation
- Conclusion and perspectives

Introduction

Previous work

Linguistic study

Proposed method

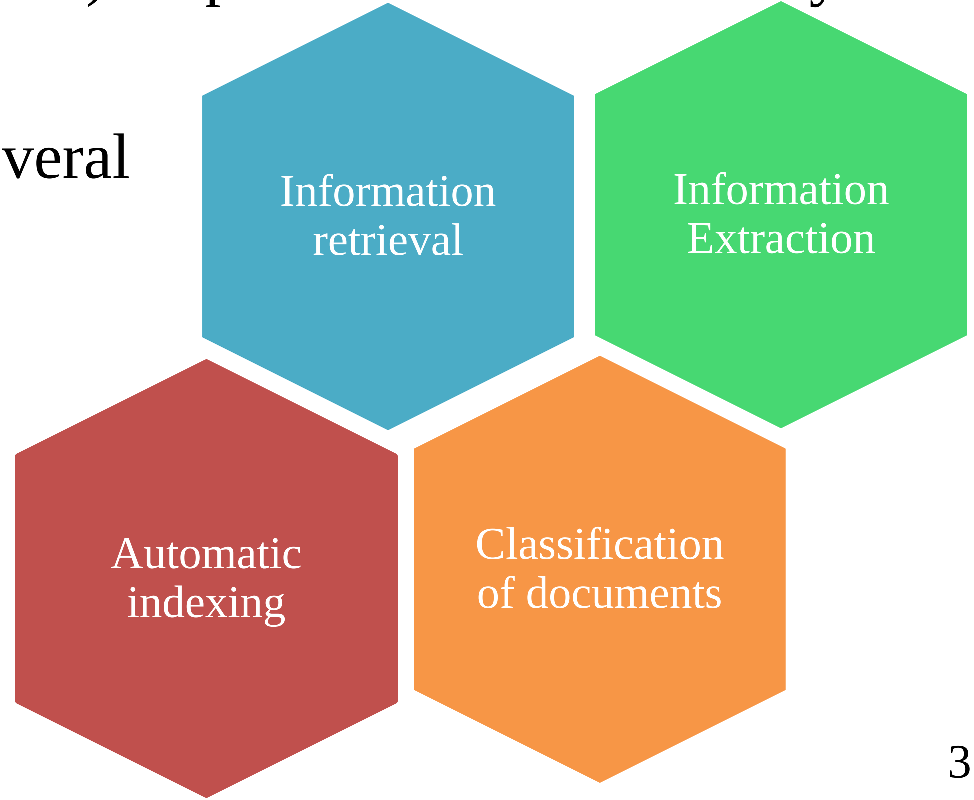
Experimentation and evaluation

Conclusion and Perspectives

Introduction

Tunisian Dialect (TD) corpus contains enormously NEs

ANER helps in several fields:



Information retrieval

Information Extraction

Automatic indexing

Classification of documents

Introduction
Previous work
Linguistic study
Proposed method
Experimentation and evaluation
Conclusion and Perspectives

Introduction

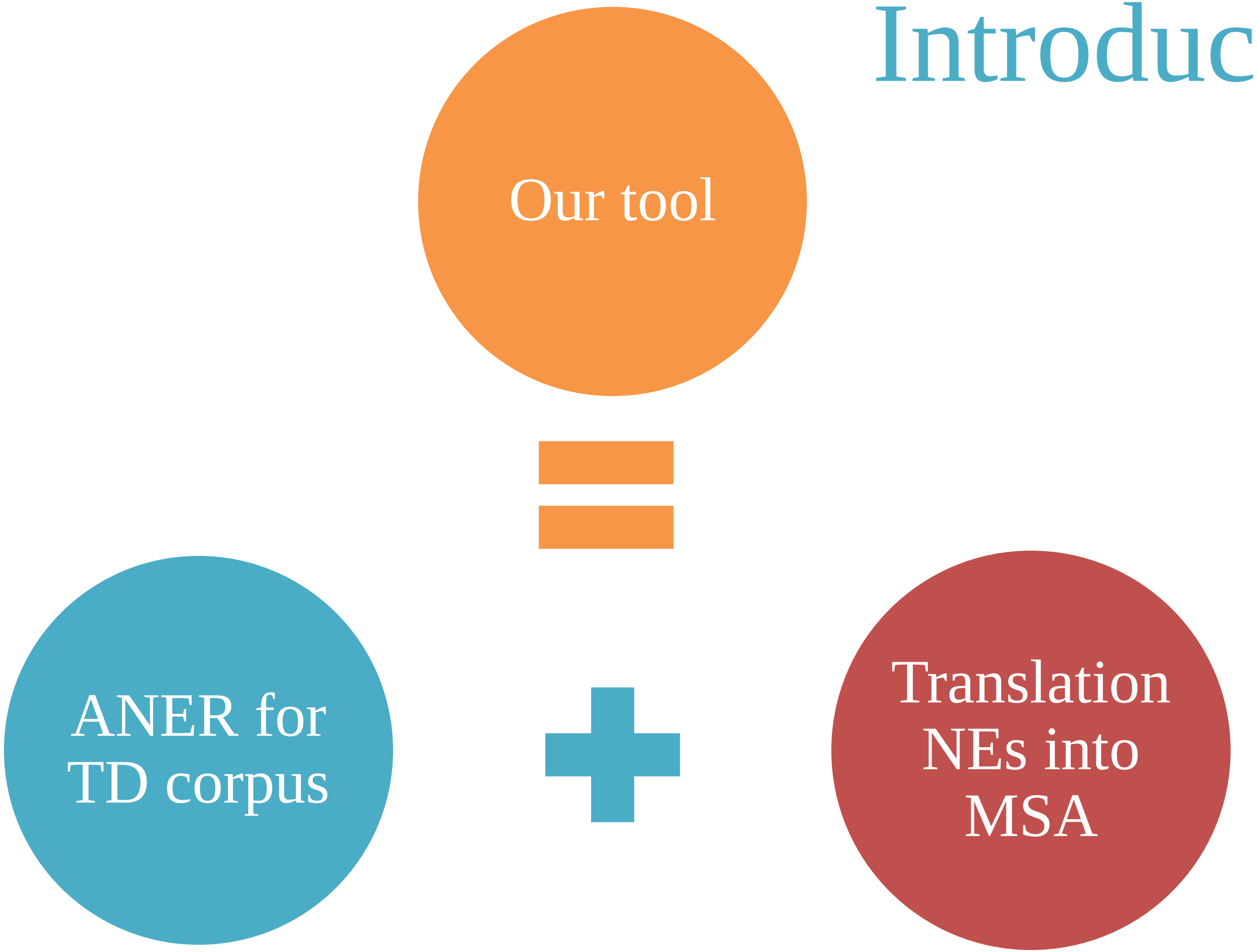
TD does not have a standard spelling

TD words not only of Arab origin but of different origins

No capital letters at the beginning of a word for TD proper noun

TD proper nouns can be written in different manners

Introduction



Introduction

Provide a linguistic study for TD

Reuse a bilingual dictionary TD-MSA

Establish a set of syntactic grammars that offers NEs recognition and their translation

Implement them in NooJ linguistic platform

Previous work

	Approach	Tool
El Bazi et al., 2019	Statistic	Bidirectional Long Short-Term Memory (LSTM)
Hkiri et al., 2017	Hybrid	ANNIE/GATE and CRF
Zaghouani, 2012	Linguistic	Language-independent rules
Fehri et al., 2011 Hasni et al., 2009 Khalfallah et al., 2009 Mesfar, 2007	Linguistic	Finite state transducers

Previous work

No work deals with ANER for TD corpus

Finite transducers have given encouraging results in:

- MSA ANER (Mesfar)
- MSA-French (Fehri)

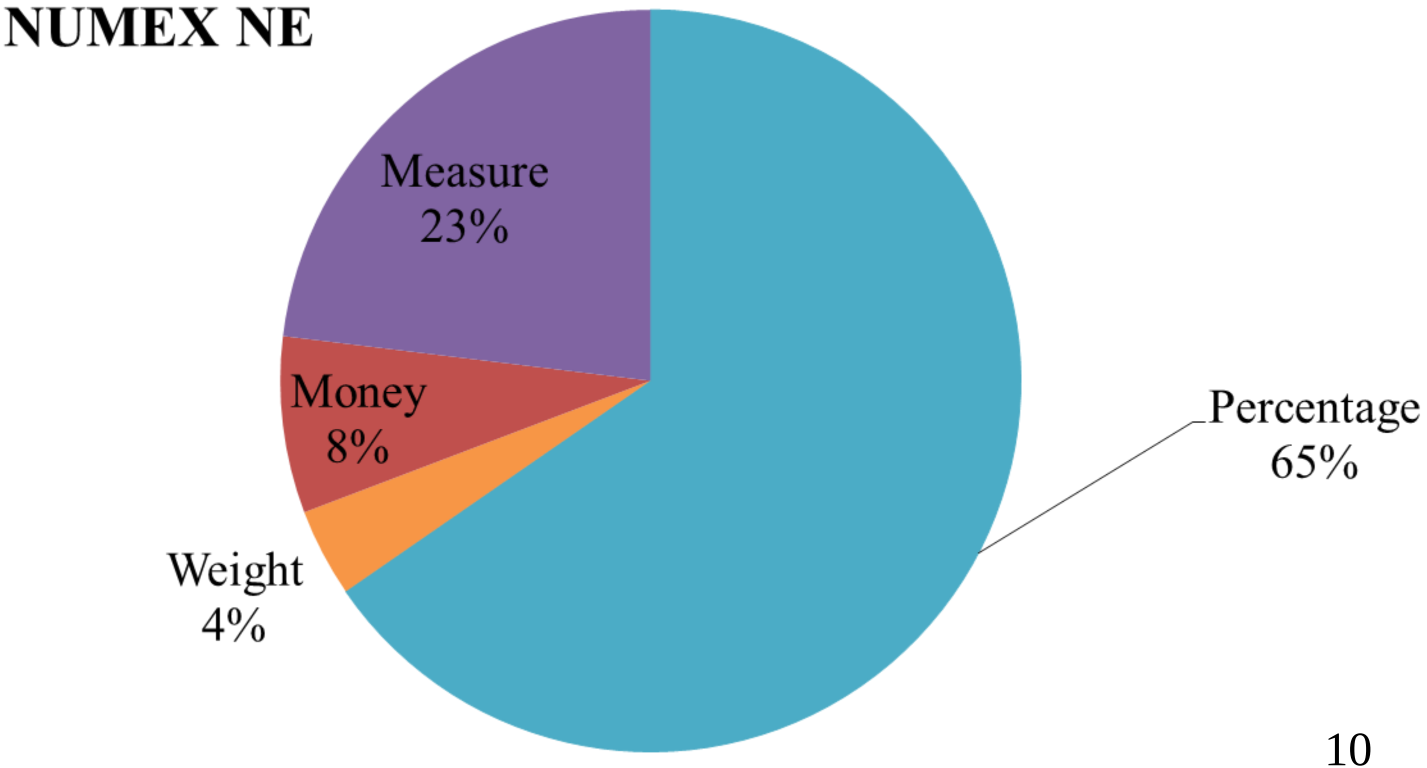
Linguistic study

ENAMEX: People, organizations, locations

TIMEX: Date, time. . .

NUMEX: Money, percentages. . .

Linguistic study



Introduction

Previous work

Linguistic study

Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Linguistic study

Percen-tage

سبعين بالمية

sab'In bilmiyah
(seventy percent)

سبعين في المية

sab'In filmiyah
(seventy percent)

% 70

Rule:
(number|digit$^+$)(preposition definite_noun|%)

Rule:
(number|digit$^+$) weight_unit

weight

70 كيلو

70kIlO
(70 kilo)

سبعين كيلو

sab'iin kIlO
(seventy kilo)

Linguistic study

Measure

سبعين كتاب

sab'iin ktAb
(seventy books)

6 كتب

6 ktob
(6 books)

Rule:
(number|digit^{+}) indefinite_noun

Money

70 دولار

70 dOlAr
(70 dollars)

$ 70

سبعين دولار

sab'iin dOlAr
(seventy dollars)

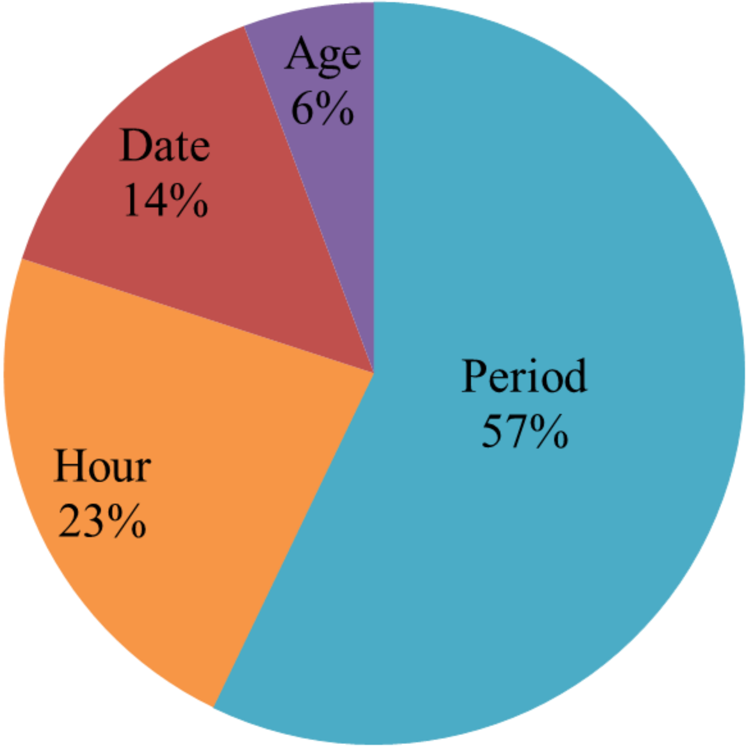
Rule:
(number|digit^{+})(devise|symbol)

Translation keeps the same syntax structure for NUMEX NEs

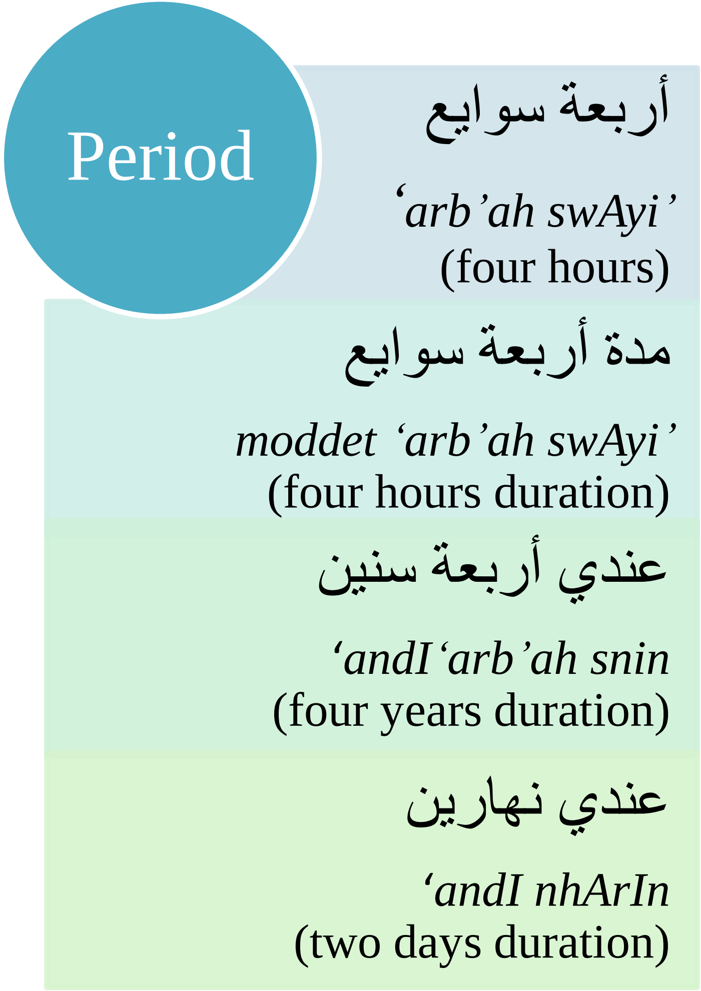
➔ Word for word translation

Linguistic study

TIMEX NE



Linguistic study



Period

أربعة سوايع

‘arb’ah swAyi’
(four hours)

مدة أربعة سوايع

moddet ‘arb’ah swAyi’
(four hours duration)

عندي أربعة سنين

‘andI ‘arb’ah snin
(four years duration)

عندي نهارين

‘andI nhArIn
(two days duration)

Rule:
(prep suffix|indefinite_noun)? (number|digit^{+}) * time_unit

Linguistic study

Hour

نهار التلاث مع الثمنية

nhar iltlAth m'A ilthmanya
(Tuesday at 8 am)

غدوة الصباح/غدا صباحا

ghodwah ilsbAh
(tomorrow morning)

التسعة متاع الصباح/
تاسعة صباحا

iltes'ah mtA' ilsbAh
(at 8 am)

Rule:
((indefinite_noun day_name)? (prep)? (number|digit^{+}) (prep part_day)?)|(noun part_day)

Introduction

Previous work

Linguistic study

Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Linguistic study

Age

عمري 5 سنين

‘omrI 5 snIn (my age is 5 years)

قفلت ستة سنين

qfalt sittah snIn (my age is 6 years)

Rule:

(عمر suffix|قفل_conjugated) (number|digit$^+$) time_unit

Translation keeps the same syntax structure for TIMEX NEs except for hours

➔ Word for word translation & adjustment translation

Introduction

Previous work

Linguistic study

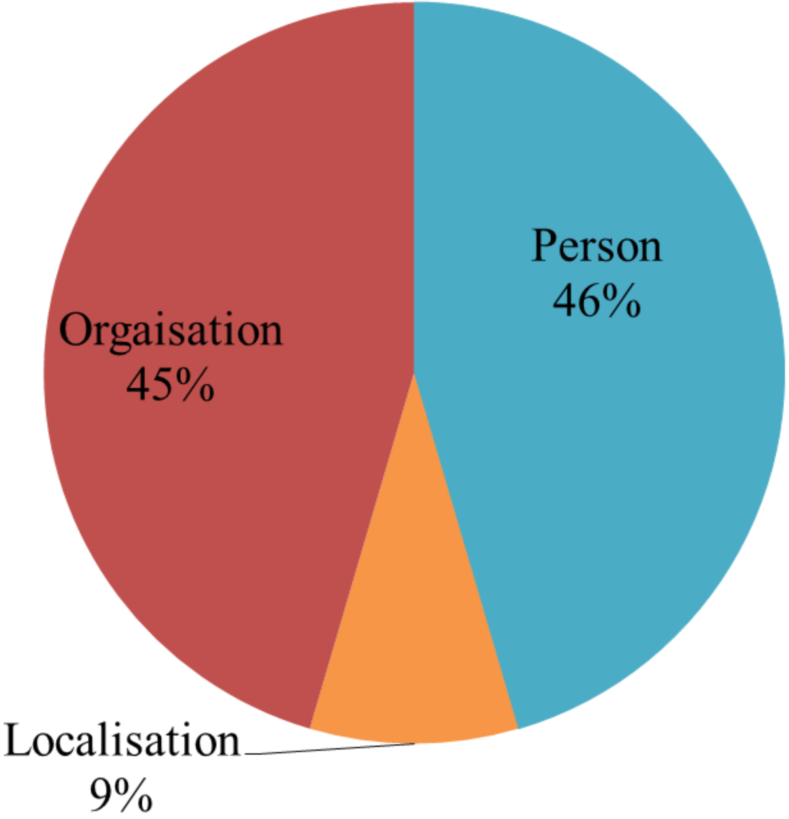
Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Linguistic study

ENAMEX NE



Introduction

Previous work

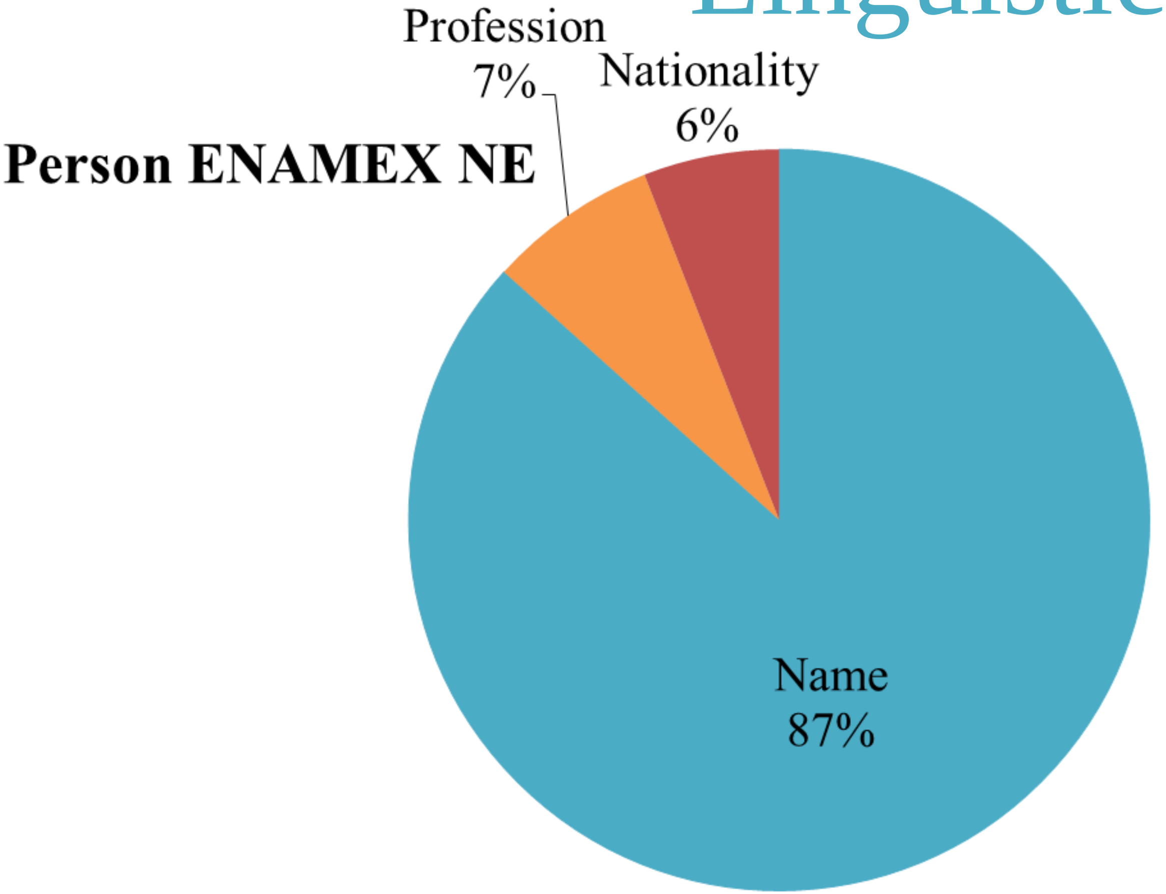
Linguistic study

Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Linguistic study



Introduction

Previous work

Linguistic study

Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Linguistic study

Name

عصام الشوالي

Issam Alchawali

عصام

Issam

أنا عصام

ana issAm
(i am Issam)

اسمي عصام

ismI issAm
(my name is Issam)

سي عصام الشوالي

sI issAm alchawAlI
(Mr Issam Alchawali)

Rule:
(Pronoun| اسم suffix|civility_noun)? first_name last_name?

Linguistic study

Profession

الفنان عاصي الحلاني

ilfannAn ‘AssI alhallAnI
(the singer Assi AlHallani)

وزير الثقافة محمد النزالي

wazIr ilthakAfah muhammad ilnzAli
(Culture Minister Mohamed AlNzali)

نخدم رئيس تحرير

Nikhdim ra’Is tahrIr
(I work Chief Editor)

Rule: (Profession_noun definite_noun? first_name last_name?) | (خدم_conjugacated Profession_noun definite_noun?)

Introduction

Previous work

Linguistic study

Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Linguistic study

Nationality

أحنا التوانسة

Ahna twAnsah
(we Tunisians)

هند صبري التونسية

hind sabrI iltUnsiyah
(Tunisian Hind Sabri)

الفرنساوي جاك

IfransAwI jAk
(French Jack)

Rule:
((first_name last_name?| pronoun) nationality_noun) | (nationality_noun first_name last_name?)

Translation keeps the same syntax structure for Person ENAMEX NEs

➔ Word for word translation

Introduction

Previous work

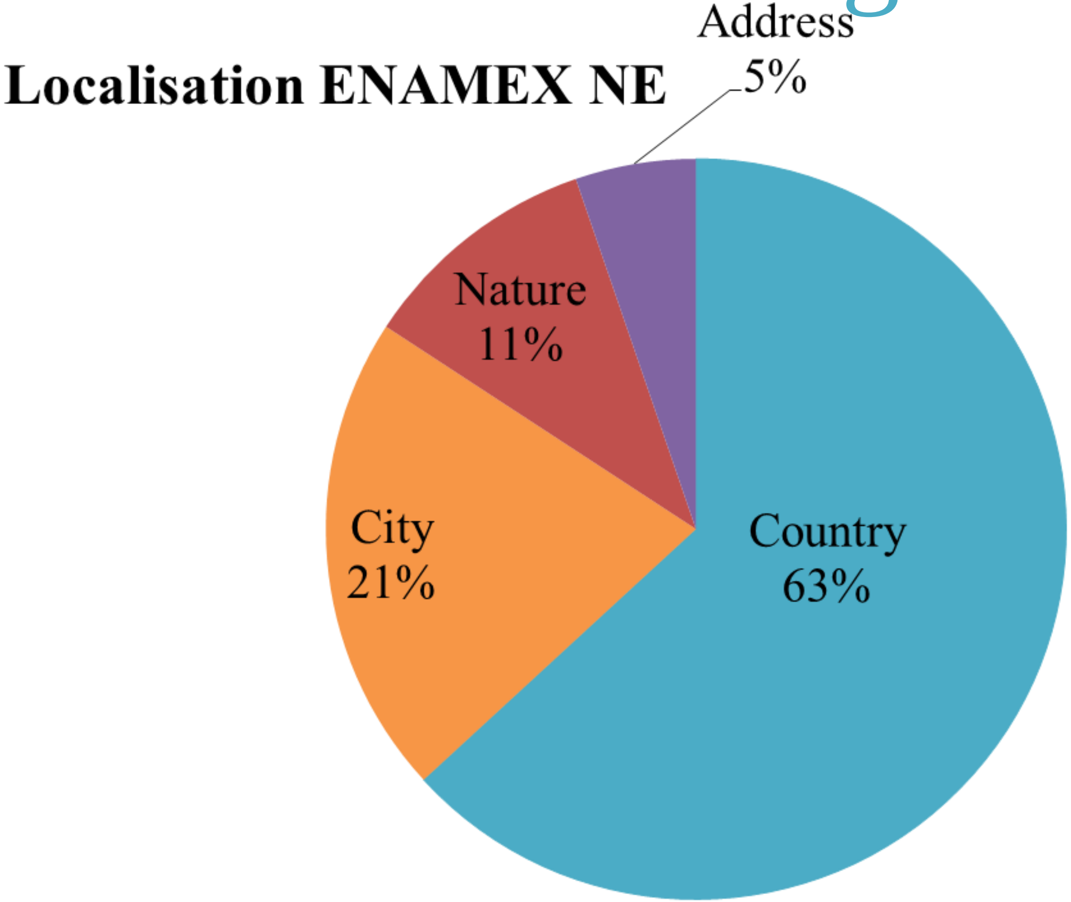
Linguistic study

Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Linguistic study



Introduction

Previous work

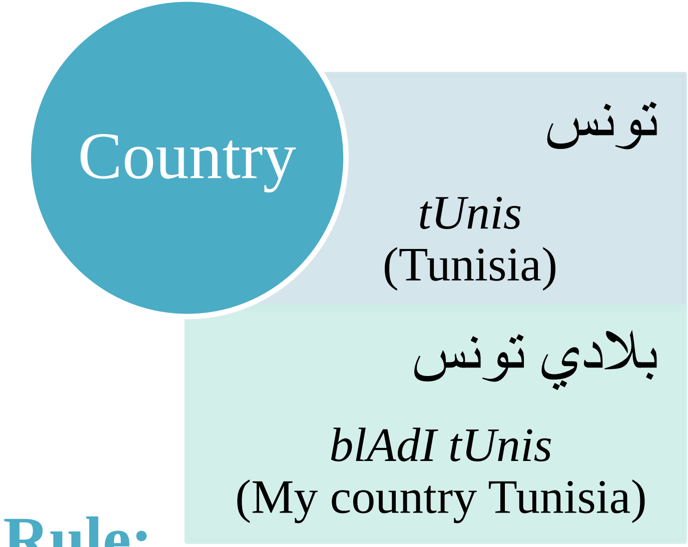
Linguistic study

Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Linguistic study



Country

تونس

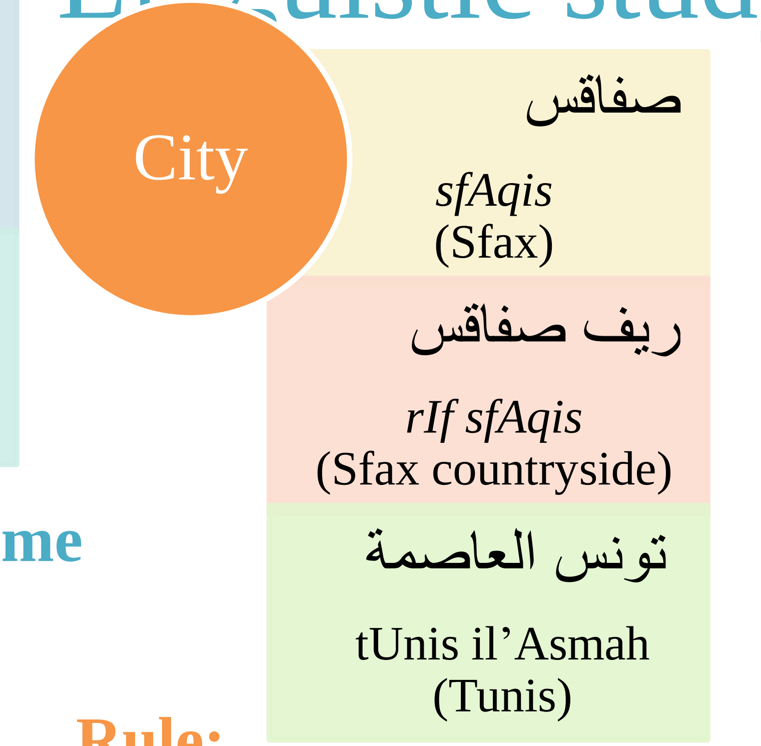
tUnis
(Tunisia)

بلادي تونس

blAdI tUnis
(My country Tunisia)

Rule:

(بلاد suffix)? Country_name



City

صفاقس

sfAqis
(Sfax)

ريف صفاقس

rIf sfAqis
(Sfax countryside)

تونس العاصمة

tUnis il’Asmah
(Tunis)

Rule:

(ريف)? City_name (العاصمة)?

Introduction

Previous work

Linguistic study

Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Linguistic study

Nature

جبل بوقرنين

jbal bUgarnIn
(Bougernin Mountain)

جزيرة الاحلام جربة

jazirah al’ahlAm jirbah
(Island of dreams Djerba)

واد مجردة

wAd majrdah
(Majrda Valley)

Rule:
(جبل|جزيرة|واد) City_name

Translation keeps the same syntax structure for location ENAMEX NEs

➔ Word for word translation

Introduction

Previous work

Linguistic study

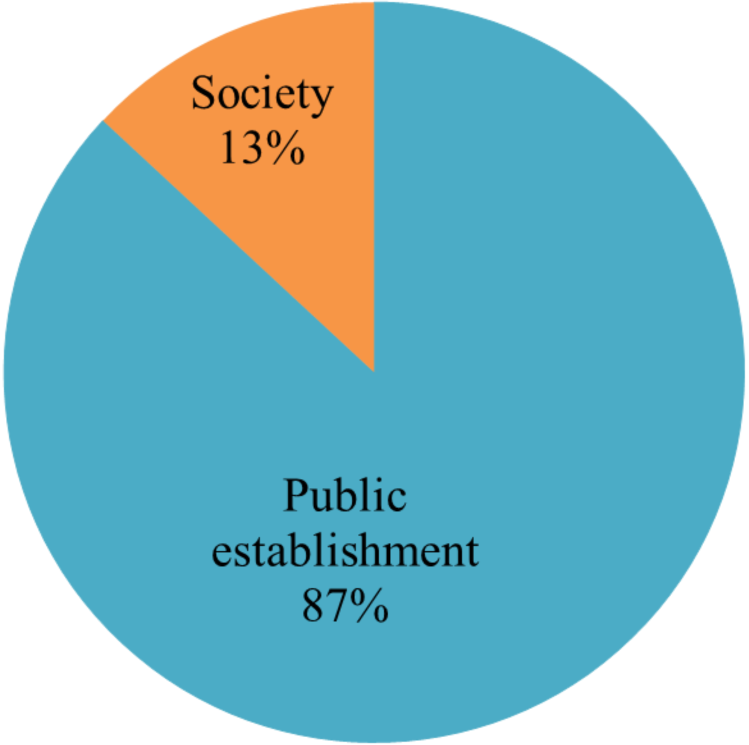
Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Linguistic study

Organisation ENAMEX NE



Introduction

Previous work

Linguistic study

Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Linguistic study

Company

بنك الامان

Bank il'amAn
(Amen bank)

مقاولات شعبان

muqAWlAt cha'bAn
(Chaaben Contracting)

مغازة كارفور

maghAzah kArfUr
(carrefour market)

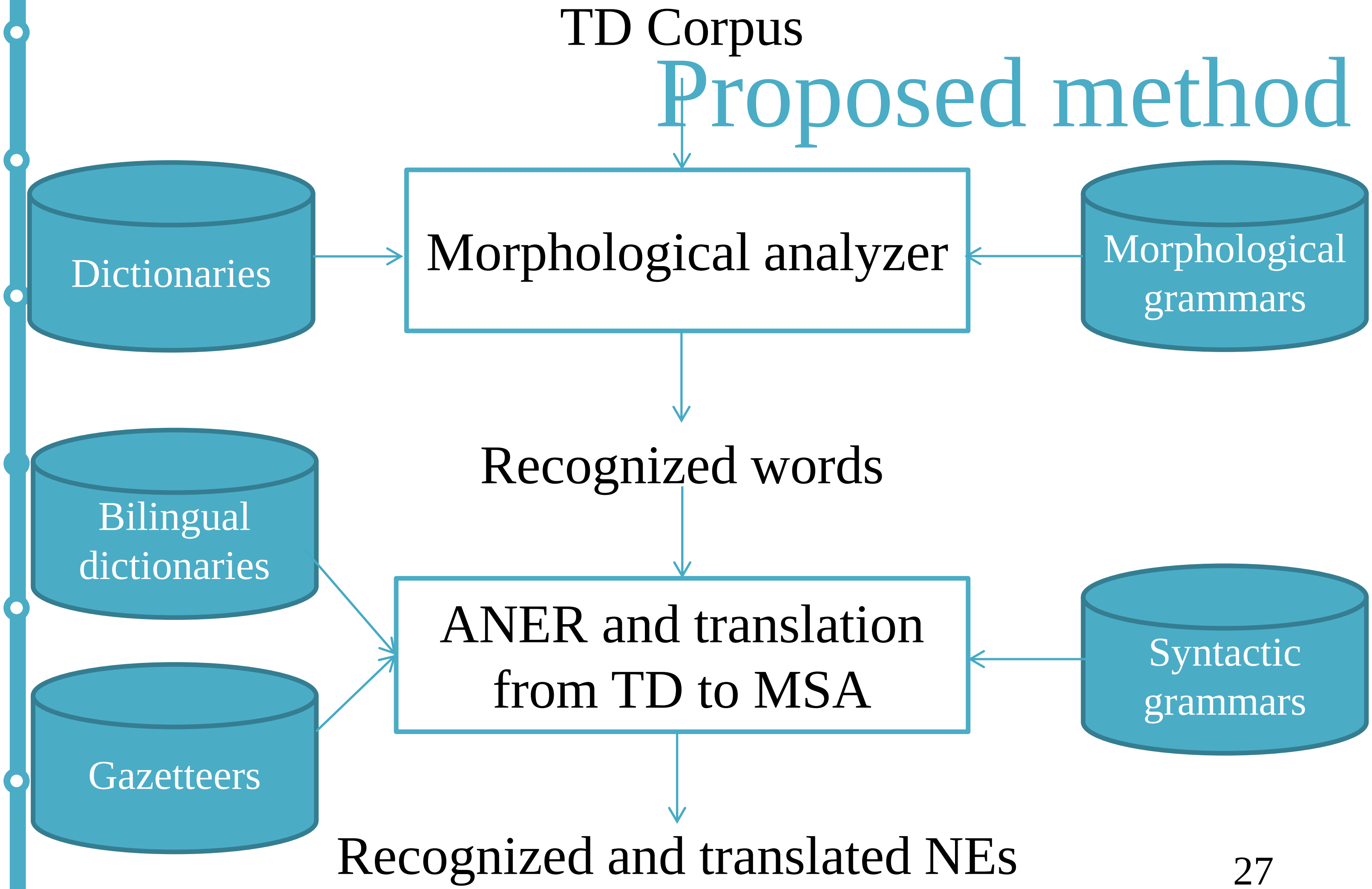
Rule: company_noun (first_name? Last_name) definite noun))

Translation keeps the same syntax structure for organization ENAMEX NEs

➔ Word for word translation

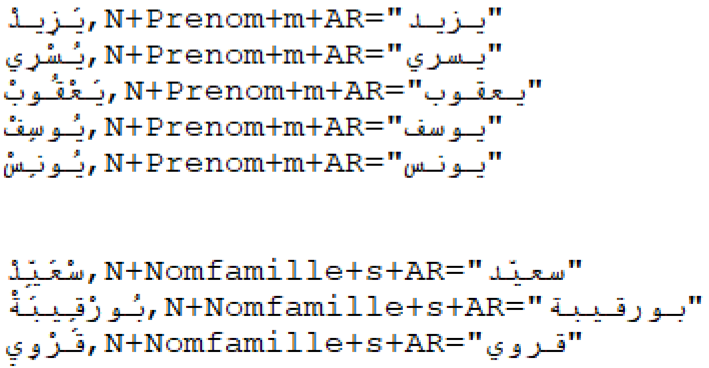
Introduction
Previous work
Linguistic study
Proposed method
Experimentation and evaluation
Conclusion and Perspectives

Proposed method



Proposed method

Bilingual dictionary thanks to + AR



```
يَزِيدْ,N+Prenom+m+AR="يـزيـد"
يُسْرِي,N+Prenom+m+AR="يـسري"
يَعْقُوبْ,N+Prenom+m+AR="يـعقـوب"
يُوسِفْ,N+Prenom+m+AR="يـوسف"
يُونِسْ,N+Prenom+m+AR="يـونـس"

سْعَيِدْ,N+Nomfamille+s+AR="سعيد"
بُورْقِيبَةْ,N+Nomfamille+s+AR="بـورقـيـبة"
قَزْوِي,N+Nomfamille+s+AR="قـروي"
```

Introduction

Previous work

Linguistic study

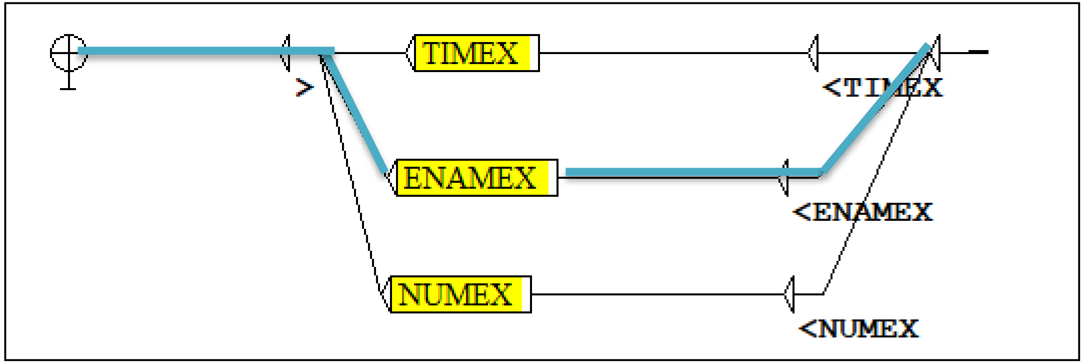
Proposed method

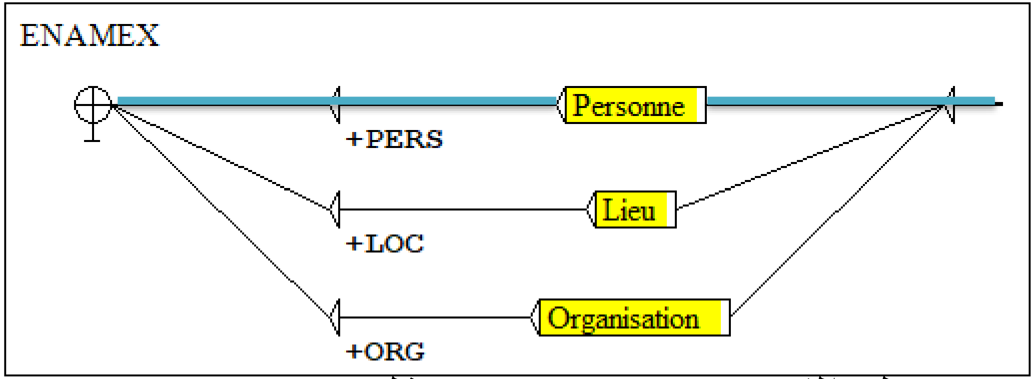
Experimentation and evaluation

Conclusion and Perspectives

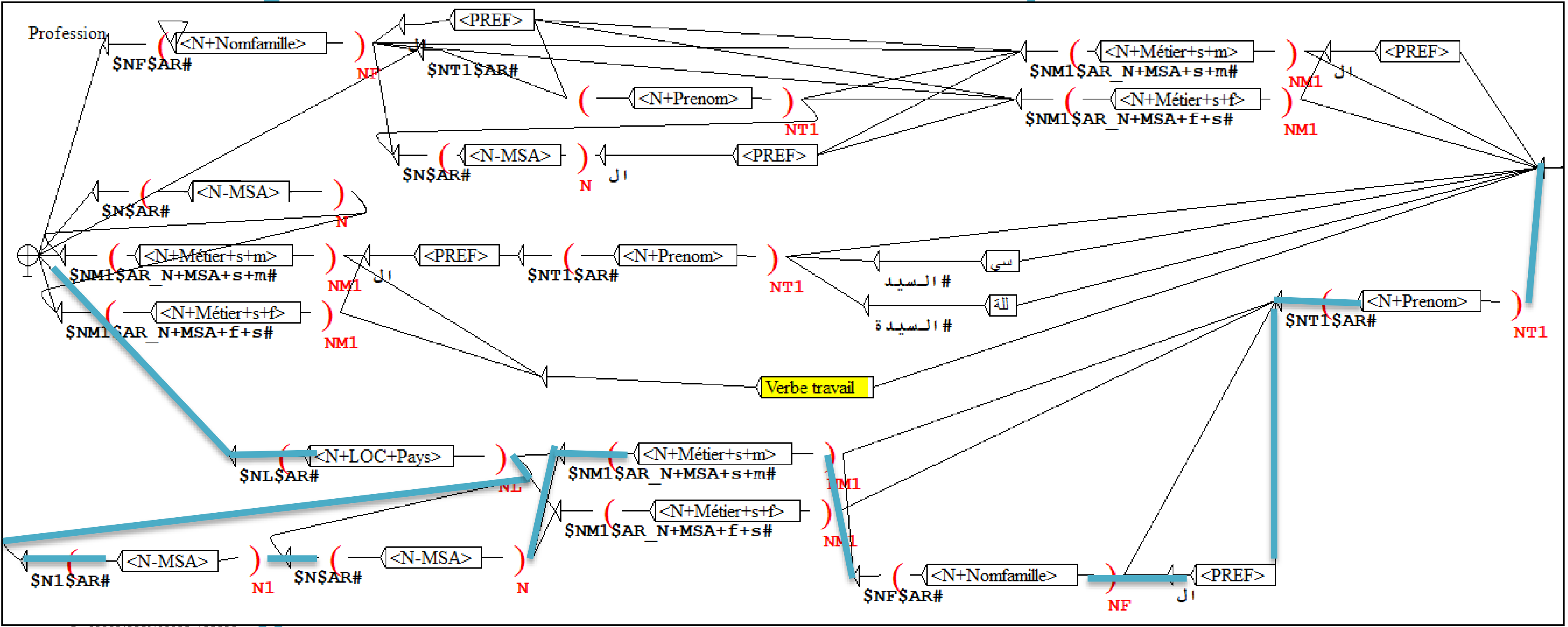
Proposed method

Grammar of ANER and translation from TD to MSA





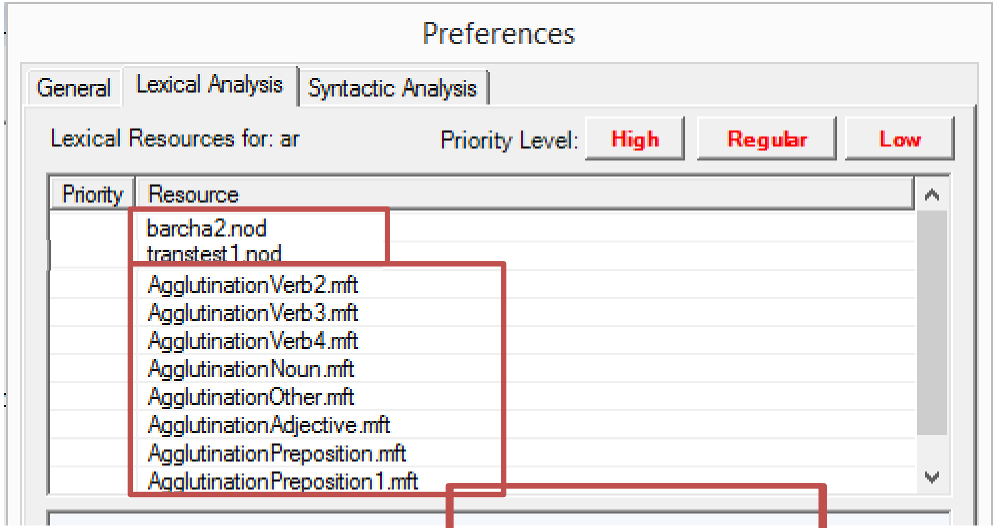
Proposed method



نبيل القروي رئيس حزب قلب تونس

Introduction
Previous work
Linguistic study
Proposed method
Experimentation and evaluation
Conclusion and Perspectives

Experimentation and evaluation



Preferences

Lexical Resources for: ar

Priority Level: High Regular Low

Priority	Resource
	barcha2.nod
	transtest1.nod
	AgglutinationVerb2.mft
	AgglutinationVerb3.mft
	AgglutinationVerb4.mft
	AgglutinationNoun.mft
	AgglutinationOther.mft
	AgglutinationAdjective.mft
	AgglutinationPreposition.mft
	AgglutinationPreposition1.mft

Allow the Bilingual agglutination dictionaries treatment

Introduction

Previous work

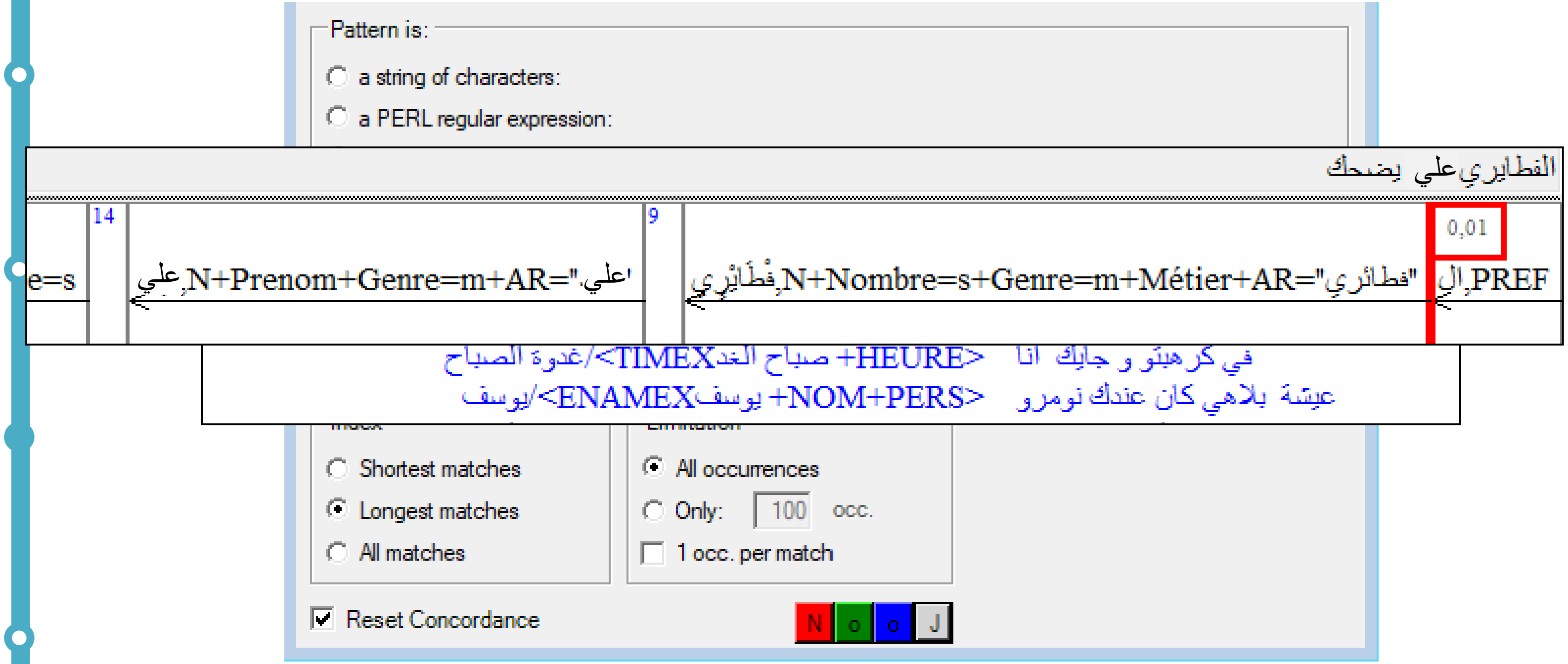
Linguistic study

Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Experimentation and evaluation



Pattern is:

○ a string of characters:

○ a PERL regular expression:

الفطايري علي يضحك

0,01

ال,PREF

فطائري,N+Nombre=s+Genre=m+Métier+AR="فطائري"

9

علي,N+Prenom+Genre=m+AR="علي"

14

e=s

في كرهبتو و جايك انا <HEURE+ صباح الغدTIMEX>/غدوة الصباح

عيشة بلاهي كان عندك نومرو <NOM+PERS+ يوسفENAMEX>/يوسف

○ Shortest matches

● Longest matches

○ All matches

● All occurrences

○ Only: 100 occ.

□ 1 occ. per match

☑ Reset Concordance

Introduction

Previous work

Linguistic study

Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Experimentation and evaluation

Test corpus contains around 20000 words

	Precision	Recall
TIMEX	96%	94%
NUMEX	98%	91%
ENAMEX	91%	83%

Our dictionary treat words of different origin

Our syntactic grammars use adjustment translation

Noise is present due to lack of disambiguation rules

Introduction

Previous work

Linguistic study

Proposed method

Experimentation and evaluation

Conclusion and Perspectives

Conclusion and Perspectives

NE Linguistic study

ANER and Translation grammars based on transducers and implemented in NooJ

Word for word and adjustment translation

Conclusion and Perspectives

Increased dictionary coverage

Building translation grammars for other NEs

Disambiguation of words

Thank you for your attention