

2nd MEDAR International Conference on Arabic Language Resources & Tools,
Cairo, 22-23 April 2009

Minimal Resources for Arabic Parsing: an Interactive Method for the Construction of Evolutive Automata

Claude Audebert (MMSH)

Christian Gaubert (IFAO)

André Jaccarini (MMSH)



Maison Méditerranéenne des Sciences de l'Homme – Aix-en-Provence, France

Institut français d'archéologie orientale au Caire, Egypt



Information retrieval and reported discourse

- Information retrieval IR
- Minimal resources
- Illustrate our approach with a few grammars
 - Reported speech and citation extraction
 - Judgment (minimal-)
 - Some morphology

Reported discourse

- Marks and external signs can help detect
- Proper names with titles/position
- These names are a source of difficulty
- Root based names or silences (non derivative)
- Titles can be put in a lexicon
- Punctuation is not dependable

A minimal approach

- We chose a surface approach
- 1 – based on the morphology of the Arabic noun, without lexicon
- 2 – minimal set of rules
- 3 – rules represented by automata :
conciseness and reflect the very nature
of Arabic (Audebert, Jaccarini 1985)

Reported discourse : syntactic features

- Specific verbs : declarative verbs
- Government : *inna* ^وإِنَّ or *anna* ^وأَنَّ
- Only قال *qāla* (should) takes *inna*
- Different preposition are used by those verbs:

^ع*abbara* ^ع*an*, ^ع*aḥbara-hu* *bi*, ^ع*ašara* *ilā*
عبر عن ، أخبره بـ ، أشار الى

Points of view

- Different implementations are possible :
 - - an automaton for each preposition
 - - a single automaton branching for the conjunctions
- These different choices reflect some *points of view* on the language ...
- To consider their adequacy to a given aim

Patterns to detect (1)

- We limit ourselves to the most probable conjugation
- We have to take into account the ambiguity of the *alif nūn* script, which may be also: *in* introducing negative sentence or *an* introducing a subordinate conjunctive clause

Patterns to detect (2)

- Removal of *anna* -> *Maṣdar*
ʾašāra ilā anna-hā waṣalet ʾamsi
ʾašāra ilā wuṣūlihā
- Not a real quotation or reported discourse : will be a noise
yuʾakkidu ḍalika anna ...

***Sarfiyya*, a tool for automata interactive designing**

- Sarfiyya contains a specially designed parser for FST (using trees, chained lists and stacks), evolving to ATN
- Parse categorized FST
- Uses deterministic computation and pre-parsing (acceptor)
- Implements processes for morphological analysis

Sarfiyya

- Automata are regex and/or graphically specified with the support of a specific automata toolbox
- Allows piping between processes or automata
- Has a debugger mode

Sarfiyya

- Contains Arabic linguistic resources besides automata
- Now completely written in Java / XML / SVG with design patterns under Eclipse and open-source technologies
- ... on-going optimization !

srf.Sarfiyya Fichier Édition Morphologie Recherche Grammaires Utilitaires

Sarfiyya dev 0.5

ah07 nv 0.933 s.

انخفاض العائد الذي يحصل عليه العاملون في هذه المجالات بل قد يترتب علم ذلك تخفيض في حجم العمالة الفعلية، وبالطبع فإن كل ذلك سينعكس بصورة سلبية على مدى السيولة المتاحة للعاملين في هذه القطاعات حيث سيترتب على نقص السيولة حدوث نوع من الكساد.

وأوضح أن هناك عدة أساليب مقترحة نها إخطار البنوك بالقطاعات الصناعية التي يوجد فائض كبير في إنتاجها مقارنة بحجم الطلب وبالتالي فإنه عند التقدم بطلبات لإقامة مشروعات جديدة في هذه المجالات كمشروعات المطاحن مثلا فإن البنوك تضم ذلك في اعتبارها وتوقف تمويل مثل هذه المشروعات باعتبار أنها لن تحقق عائدا اقتصاديا مناسبيا يتيح سداد أقساط القروض المصرفية لمثل هذه المشروعات، كما يمكن أيضا وقف أي مزايا وحوافز تقدمها الدولة للمشروعات الجديدة التي قد يطلب بعض المستثمرين إقامتها، وذلك على أساس أن الهدف من هذه الحوافز الاستثمارية التشجيع علم، الاستثمار، ومادام يوجد فائض كبير في الطاقات الإنتاجية للمطاحن فإن هذه النوعية من المشروعات لا تحتاج إلى تشجيع وبالتالي لا يوجد سبب لنحها حوافز. ومن ضمن المقترحات في هذا المجال أيضا أن يتم إصدار أي موافقات جديدة لإقامة مشروعات للمطاحن في مصر لفترة محددة يمكن مثلا أن تكون في حدود 5 سنوات.

0:1206 n° علق √? حتم

ويقول السيد حسن دياب غانم رئيس غرفة صناعة الحبوب إن الطاقات الفائضة أصبحت ظاهرة لها تأثير سلبي كبير على صناعة الدقيق استخراج الذي يستخدم في إنتاج المكرونة والعلوى والخبز الفينو.

DET calculée pour cit1: 78états

141 cit1 يقول المهندس عادل الموزي وكيل اتحاد الصناعات المصرية ورئيس الشركة القابضة للصناعات الكيماوية إن هناك بالفعل طاقات فائضة غير مستغلة في العديد من المجالات الصناعية ومنها مثلا الأسمدة حيث تصل نسبة الطاقات غير المستغلة فيه إلى والأسمدة الأزوتية وديباجة الجلود والأحذية وعلع الطعام والغازات الصناعية وتصل هذه النسبة في بعض الغازات إلى وبالطبع فإن هذه الطاقات الفائضة تعتبر أمر سلبيا وغير مناسب على الإطلاق خاصة في مصر.

625 cit1 يقول السيد حسن دياب غانم رئيس غرفة صناعة الحبوب إن الطاقات الفائضة أصبحت ظاهرة لها تأثير سلبي كبير على صناعة الدقيق استخراج الذي يستخدم في إنتاج المكرونة والعلوى والخبز الفينو.

gala-seul.xml

titre: cit1 doc: ; DET 63 états. 1.096 s.; DET: 78 états, 0.793 s.

E: (2:4) 24 A: 4 arc 4, l: g cat: cit-qala

regex mots expand 1 ligne

```

<pos e="8" dx="0.2" dy="0.0"/>
<pos e="9" dx="0.4" dy="0.0"/>
<pos e="10" dx="0.6" dy="0.0"/>
<pos e="11" dx="0.8" dy="0.0"/>
</graph>
</gram>

```

total1-6.xml

titre: cit1-6 doc: citation à 6 branches; DET 9814 états.

E: (0:0) A: l: cat:

titre: minel doc: extrait entre ال من et أن ou إن; DET 29 états. 0.288 s.

E: (0:0) A: l: cat:

regex mots expand 1 ligne

```

<pos e="12" dx="15.3" dy="3.0"/>

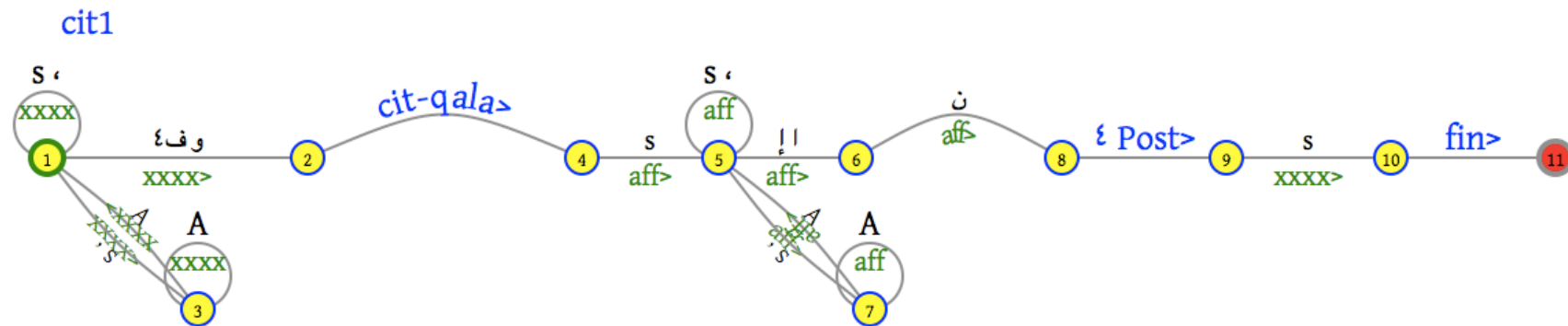
```

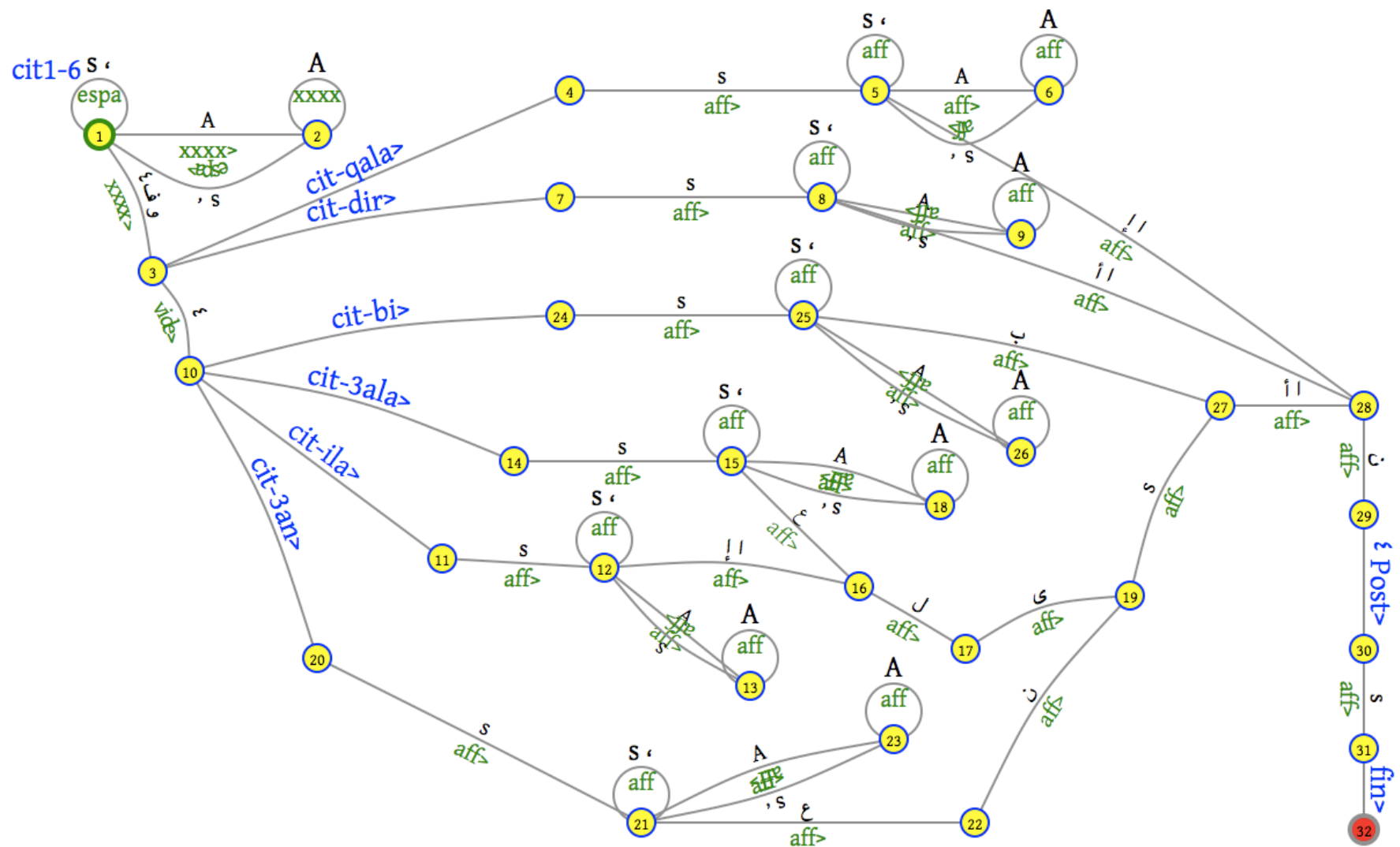
Building the *cit* grammar

- Phrase-qala = words (ε+ف+و) cit-qala words إن (Post+ε) words end
- + for disjunction; s for space; * Kleen star
- Words = (A A* (s+،) (s+،) *)*
- cit-qala = قال+قالت+قالوا+يقول+تقول etc.

The *cit* grammar

- *cit* (verbs, preposition, conjunction) = Verbs words* preposition conjunction (Post+ ϵ)
- *cit1* = *cit* (*cit-qala*, ϵ , إن_ء)





$Cit_{1-6} = \text{words } (\varepsilon + \text{و} + \text{ف} + \varepsilon) (\sum \text{cit}_i) (\text{Post} + \varepsilon) \text{ words}$

Computing DET(*cit*)

- Using the classical DET algorithm (Aho 86, etc.)
- Exponential complexity :
- cit1-4 : 1900 states; cit1-5 : 4200 states; cit1-6 near 10 000 states within hours!
- Requires optimized structures
- Using workarounds (each branch being separately determinized)

Testing *cit*

- Using a 20 000 words corpus of al-Ahram contained in a larger one
- Retrieves 48 quotations and reported speech, mainly from *qāla*, *'akada*, *awḍaḥa*, *aḍāfa*, about 1000 words/s
- Also retrieves noise coming from long sentences with multiples *anna / inna*

Troubleshooting *cit*

- Introducing a complement capacity into the parser
- Split some sentences into overlapping sequences to be parsed and reassembled
- Develop a non greedy branch of the parser
- A hard limit to limiting some cycles into automata

من الـ أن

- من الممكن، من الواضح...
- ومن غير المنطقي أن
- words (ف+و+e) من s ال (ε + غير) A* s ان s words
- Fighting some noise with a small lexicon

Revisiting morphology

- From the very simple automata of 6 states to 13 states for the determinist version : Cohen then Jaccarini 94
- We define general morphology as a *transduction* of a basic system (Jaccarini 98)
- To associate micro-lexicon data and morphological categories can help to cover 98% of the needs, with unavoidable but measurable noise (Gaubert 01)

Kawâkib bêta كواكب

[Français](#) [English](#) [عربية](#)

Arabic Text : [enlarge](#) [hide](#) [erase](#)

كواكب
قمح
صحة
إعلام

شهد الأسبوع الماضي الإعلان عن عدة ظواهر فلكية مهمة، وهي اكتشاف ١٠ كواكب جديدة خارج المجموعة الشمسية، وحدث عدة انفجارات شمسية وصفها العلماء بعطسة الشمس أثرت بصورة مباشرة على الأقمار الصناعية والشبكات الكهربائية بالأرض، ومن المتوقع - كما يقول الدكتور محمد أحمد سليمان رئيس معمل أبحاث الشمس بالمعهد القومي للبحوث الفلكية - أن تحدث عدة انفجارات هائلة بالشمس نظرا لمرورها بمرحلة النهاية العظمى لدورة نشاطها التي تتكرر كل ١١ سنة وتنتهي عام ٢٠٠١.

Settings and Actions : [arabic keyboard help](#)

Schemes

Root نظر (to look at)

3 ر

2 ظ

1 ن

Frequent roots

Empty phrase

Tokens suites

Tokens

Sequence

رئيس

Reported speech

Frequencies

Repetitions

NVT

Result : [enlarge](#) [right to left](#)

Test me!

Copy a text of your choice in Arabic (taken for instance from the Arabic newspapers websites like [المصري اليوم](#), [الأهرام](#), etc. or any other source of texts) or choose one of the proposed examples on the left-side list. You can enlarge this text, hide it to work in a blindly way or erase it to replace it.

Then achieve one of the following actions:

- Search for the 20 most frequently used roots
- Search for a trilitar root like ن ظ ر using for the input the right side (consonnants) of the virtual [arabic keyboard](#) if you dont have a hardware keyboard; in R3, the letter ر must be used for ي (enter شفو for شفى for instance)
If you erase one of the radical letters Kawakib will show a list of roots using the other letters
- Search a character string
- Read the text with its tokens (tool words)
- Click on settings and actions/help or on the Cygnus constellation (Deneb, [ذنب الدجاجة](#)) to reload the initial page

Conclusion

- Necessity of the feed-back method
- To evaluate grammars themselves not only results
- Ambiguity not necessarily due to the absence of lexicon and should be organized into a hierarchy
- Automats helps clarifying the transition from the declarative to the operational