2nd MEDAR International Conference on Arabic Language Ressources & 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 (min al-)
  - 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* إنّ or *anna* أنّ
- Only قال *qāla* (should) takes *inna*
- Different preposition are used by those verbs:

ʿ*abbara ʿan*, *aḫbara-hu bi*, ʾ*ašara ilā*

عبّر 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 ḏalika 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

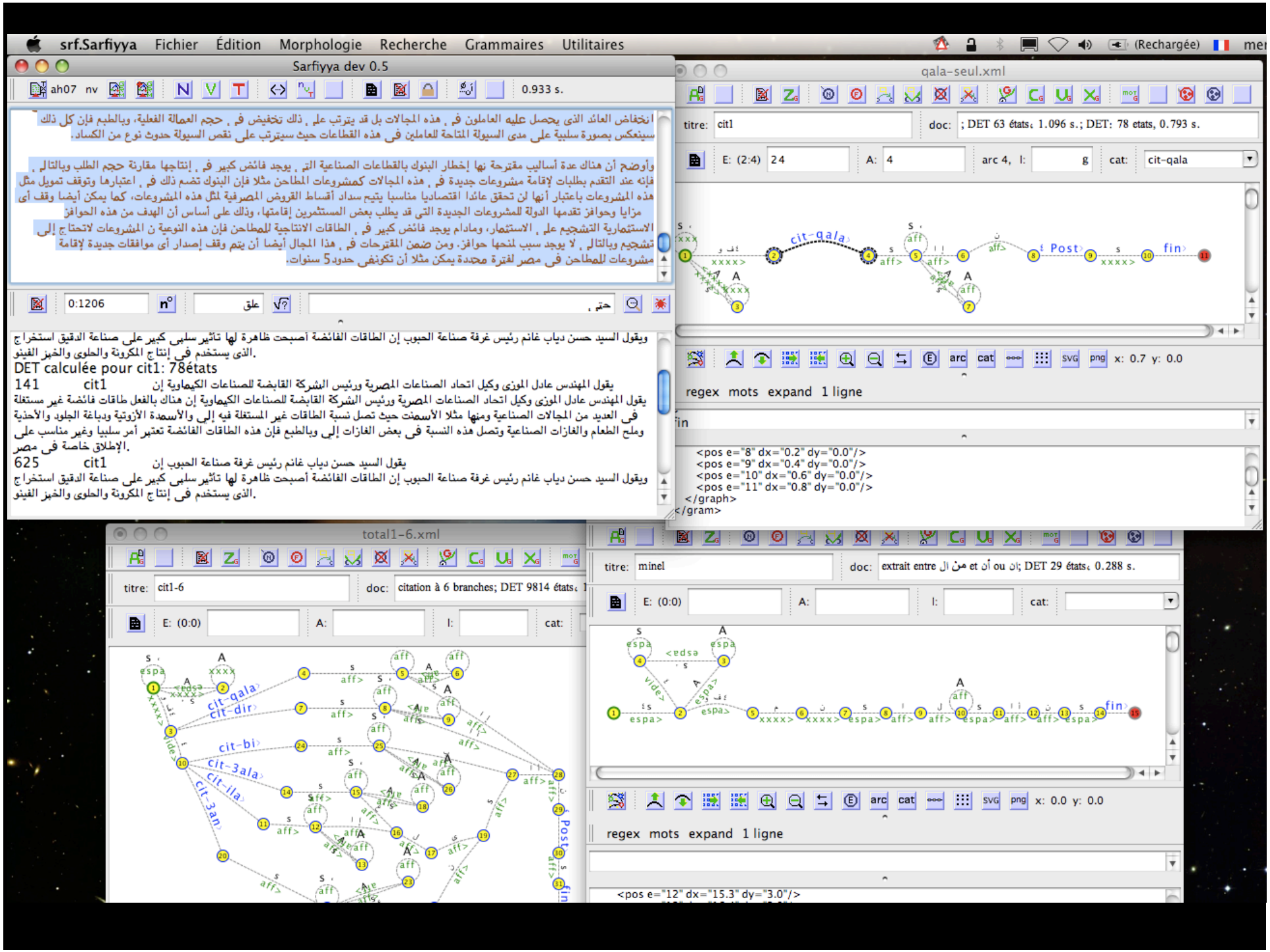
Sarfiyya

- 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
qala-seul.xml
total1-6.xml
cit-qala
cit-dir
cit-bi
cit-3ala
cit-ila
cit-3an
regex mots expand 1 ligne

Building the *cit* grammar

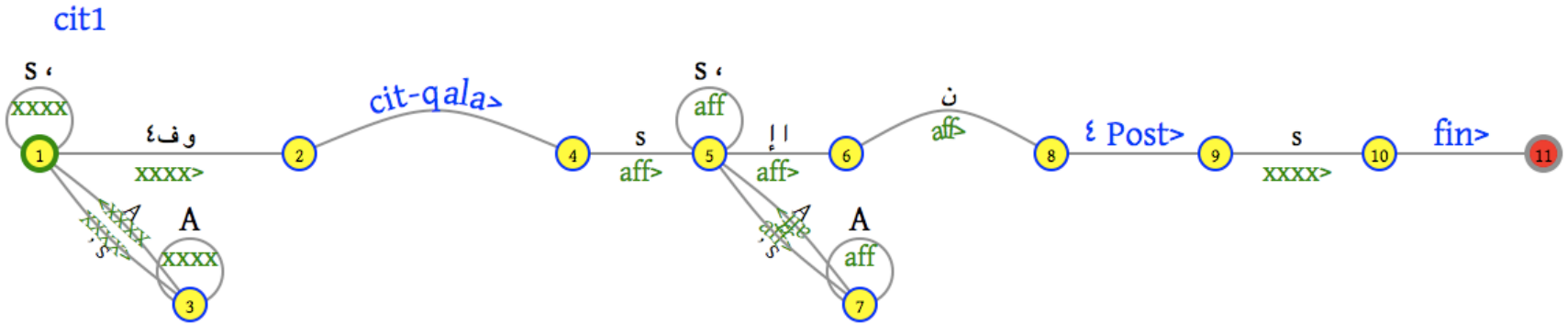
- Phrase-qala = words (ε+ف+و) cit-qala words إن (Post+ε) words end

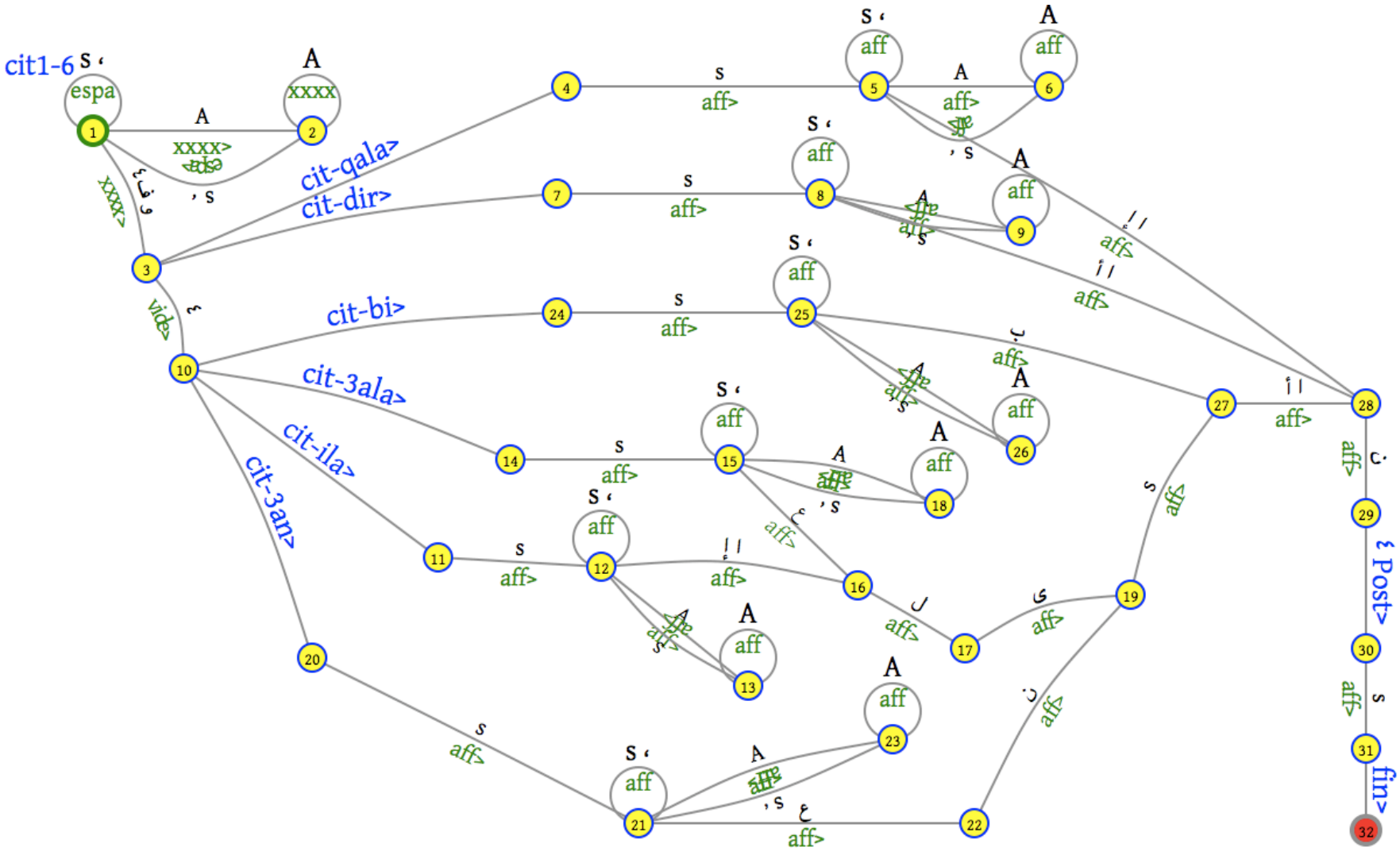
- + for disjunction; s for space; * Kleen star
- Words = (A A* (s+،) (s+،) *)*
- cit-qala = تقول+يقول+قالوا+قالت+قال+ etc.

The *cit* grammar

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

- cit1 = cit (cit-qala, ε, إن)





$Cit_{1\text{-}6}$= words (ε+ف+و) (Σcit_i) (Post+ε) 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 separatly 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*, *ʾakada*, *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 الـــ أن

- من الممكن، من الواضح...
- ومن غير المنطقى أن
- words (من s ال (ε +غير) A* s ان s من (e+و+ف
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 pro * كواكب برو
http://localhost:8080/kawakib/pro.html?lg=en
كواكب
Kawâkib bêta كواكب
Français English عربية
Arabic Text : enlarge hide erase
كواكب
قمح
صحة
إعلام
شهد الأسبوع الماضى الإعلان عن عدة ظواهر فلكية مهمة،وهى اكتشاف ١٠ كواكب جديدة خارج المجموعة الشمسية، وحدوث 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 القومى للبحوث الفلكية - أن تحدث عدة انفجارات هائلة بالشمس نظرا لمرورها بمرحلة النهاية العظمى لدورة نشاطها التى تتكرر كل ١١ سنة وتنتهى 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 you choice in Arabic (taken for instance from the Arabic newspapers websites like الأهرام ,المصري اليوم, etc. or any other source of texts) or choose one of the proposed exemples on the left-side list. You can enlarge this text, hide it to work in 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 و 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 inital page
Terminé
S3 Fox commission • لجن

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