Hahooa Arabic Web Directory

Natural Language Processing for Arabic Information Retrieval

ACL Workshop on Arabic Language Processing July 6 2001

Malek Boualem, France Telecom R&D

Régina Sneifer, Wanadoo

Contents

1. Arabic Web
2. Arabic directories and portals
3. Arabic search engines

4. Hahooa Arabic Web directory

5. Some problems with Arabic information retrieval
6. Arabic linguistic features

7. Natural language processing
8. Arabic NLP activities at France Telecom R&D

1. Arabic Web

- Internet users (2002) :

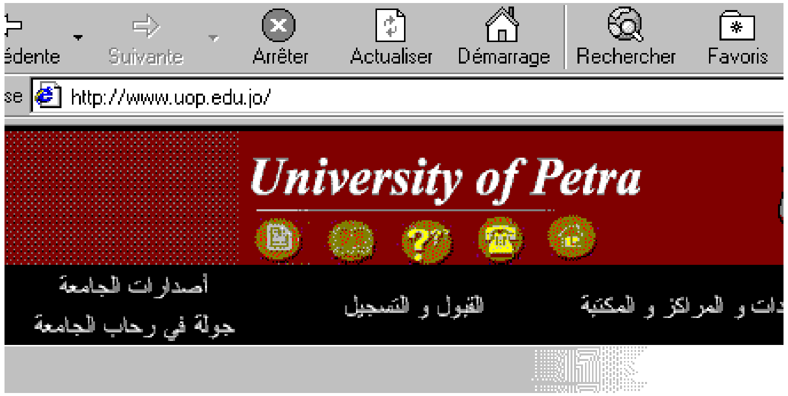
12 M / 240 M

- Arabic Web :

High speed development.

MS, Word, FrontPage, IE.

- Character encoding :
  - Microsoft CP1256
  - ISO-8859-6
  - UNICODE-UTF8



تأسست الجامعة في أيلول
سبتمبر) 1991 ، وتقع في)
ضاحية سكنية من ضواحي
مدينة عمان على شارع مطار
الملكة علياء الدولي ، وعلى
بعد 7 كيلومترات من الدوار
السابع . ويبلغ عدد الطلاب
الملتحقين بالجامعة في العام
الحالي حوالي (1600) طالب وطالبة من (30) دولة عربية وأجنبية مختلفة مما يوفر
لهؤلاء الطلبة جواً متنوع الثقافات ومتميزاً بالحيوية والانفتاح

ويوجد في الجامعة خمس كليات تقدم (15) تخصصاً مختلفاً يمكن
للطالب أن يحصل على درجة البكالوريوس في أي منها . وقد حرصت هذه الكليات
على توفير أحدث التجهيزات
التعليمية والأجهزة المختبرية بما يكفل حصول طلابها على أقصى ما يمكن من
الخبرات النظرية والعملية

2. Arabic directories and portals

Directories and Portals (Arabic & English) :

- *http://www.albawaba.com*
- *http://www.arabia.com*
- *http://www.arabicseek.com*
- *http://www.ayna.com*
- *http://www.konouz.com*
- *http://www.maktoob.com*
- *http://www.naseej.com*

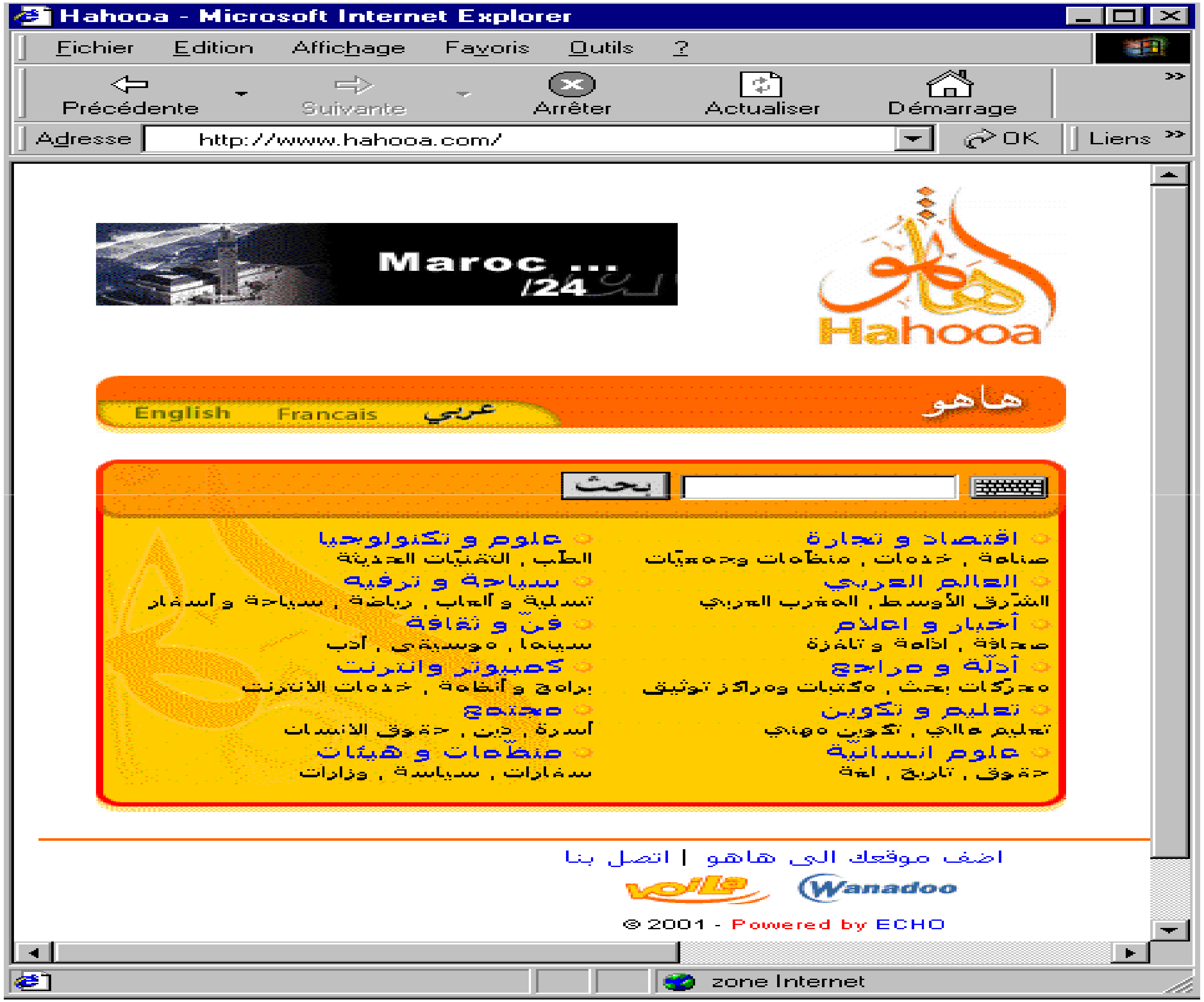
Directories and Portals (Arabic, English & French) :

- *http://www.hahooa.com*

3. Arabic search engines

- **Search engines for directories and portals**
  - **Extense-Voila** (Echo-Wanadoo, France)
    http://www.hahooa.com
  - **Ayna** (Ayna, USA)
    http://www.ayna.com
  - **Al-Idrisi** (Sakhr, Egypt)
    http://www.sakhr.com/
  - **Konouz** (Alladin, Australia)
    http://www.konouz.com/
- **Search engines for the Web**
  - **Arabvista** (Emirates Internet & Multimedia and Compaq)
    http://www.arabvista.com

4. Hahooa Arabic Web directory



Adresse: http://www.hahooa.com/

هاهو

اقتصاد و تجارة
صناعة , خدمات , منظمات وجمعيّات

العالم العربي
الشّرق الأوسط , المغرب العربي

أخبار و اعلام
صحافة , اذاعة و تلفزة

أدلّة و مراجع
محرّكات بحث , مكتبات ومراكز توثيق

تعليم و تكوين
تعليم عالي , تكوين مهني

علوم انسانيّة
حقوق , تاريخ , لغة

علوم و تكنولوجيا
الطب , التقنيّات الحديثة

سياحة و ترفيه
تسلية و ألعاب , رياضة , سياحة و أسفار

فنّ و ثقافة
سينما , موسيقى , أدب

كمبيوتر وانترنت
برامج و أنظمة , خدمات الانترنت

مجتمع
أسرة , دين , حقوق الانسان

منظّمات و هيئات
سفارات , سياسة , وزارات

اضف موقعك الى هاهو | اتصل بنا

voila Wanadoo

© 2001 - Powered by ECHO

France Telecom R&D

Hahooa : *Here it is*
http://www.hahooa.com

- Echo-Wanadoo, France Telecom (Voila search engine *http://www.voila.fr*).

- Hahooa is trilingual : Arabic, English and French.

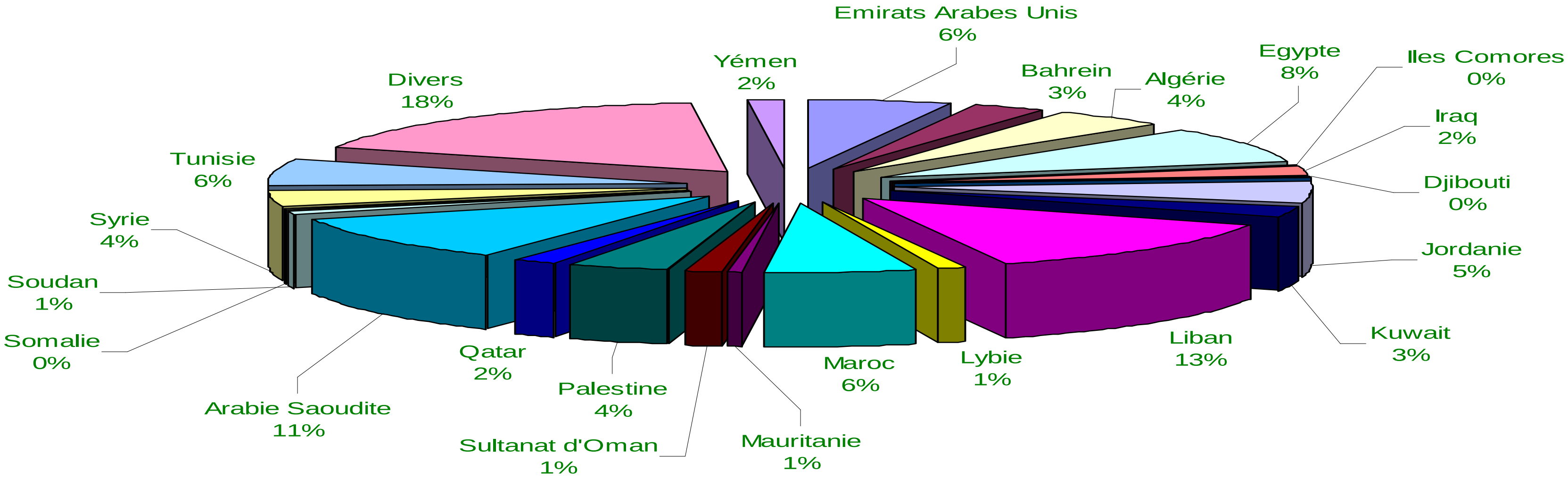
- More than 7600 URLs (Arabic around 3000).
- 418 themes (commerce and economy : more than 42%).

- Navigation languages : Arabic 49% , English, French
- Users : France, Saudia, Lebanon, Morocco, Jordan, Canada, Belgium, etc.
- Frequent themes : Arts & Culture (music, …), Arab world, News & Media.

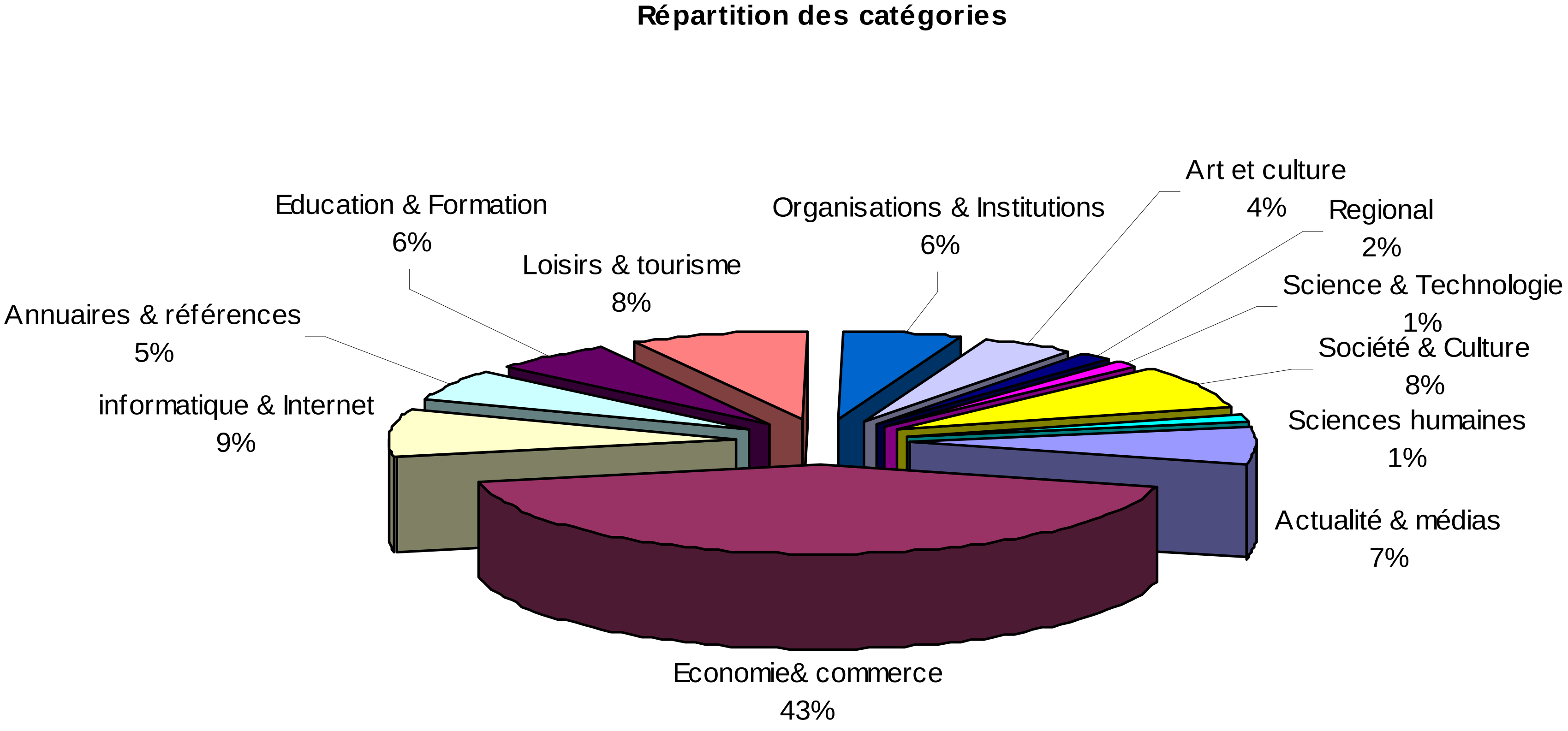
- Hahooa is regularly updated.

Hahooa by countries

Répartition des sites par pays

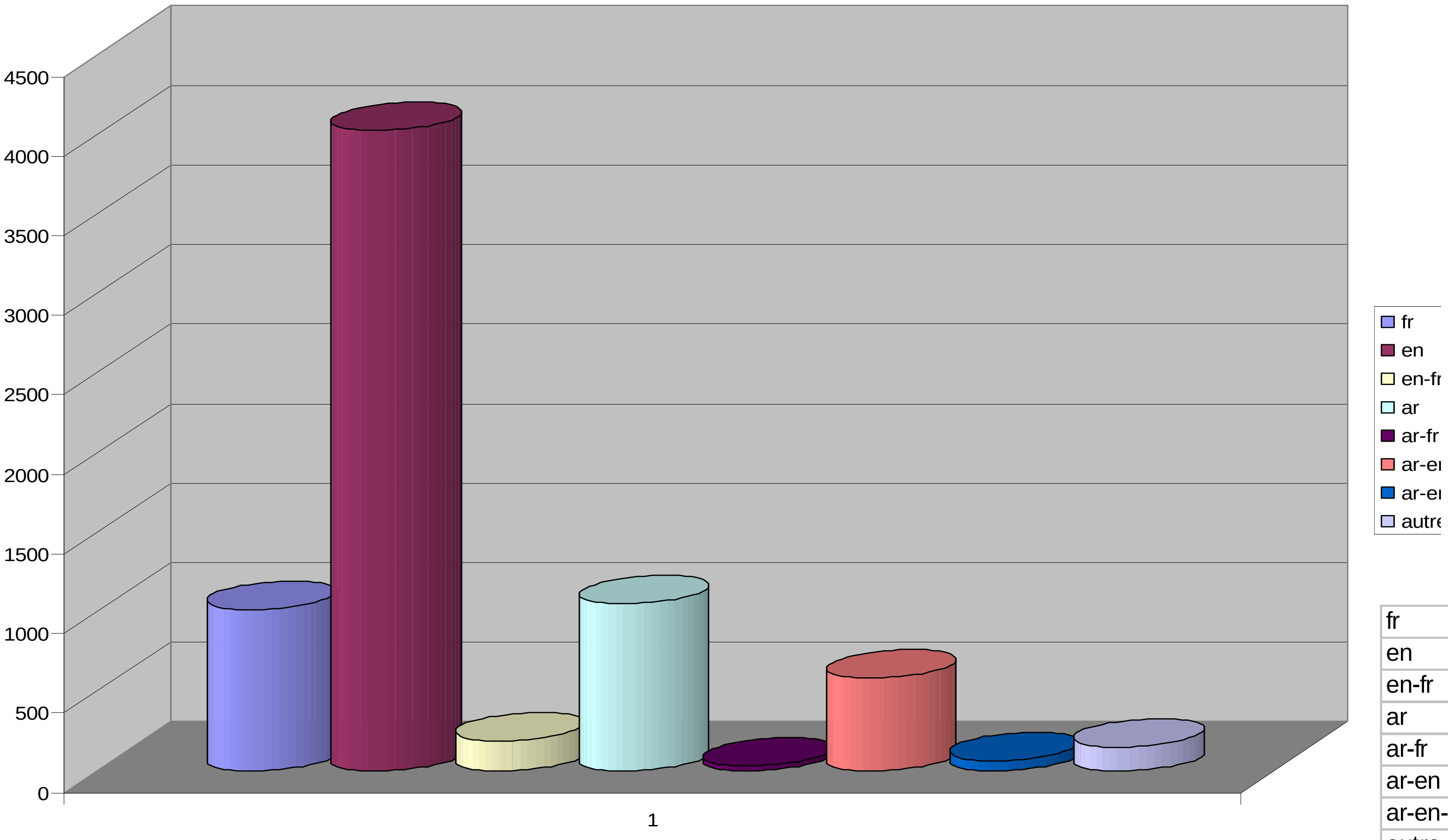


Hahooa by themes



Hahooa by languages

Répartition des sites pa langue



fr	1019
en	4028
en-fr	192
ar	1054
ar-fr	32
ar-en	586
ar-en-fr	69
autre	154

No copy nor circulation of this document without written authorization by France Telecom R&D

5. Some problems with Arabic information retrieval

Keyword	Answers or documents
مجتمع (society)	3012
المجتمع (the society)	20695
مجتمعات (societies)	599
علم (science)	12410
علوم (sciences)	3925
إمرأة (woman)	193
نساء (women)	11970
كتب (book/write)	44133
كتابة (writing)	2990
تعليم (education)	46728
تدريس (education)	770

Some tests using one of the search engines

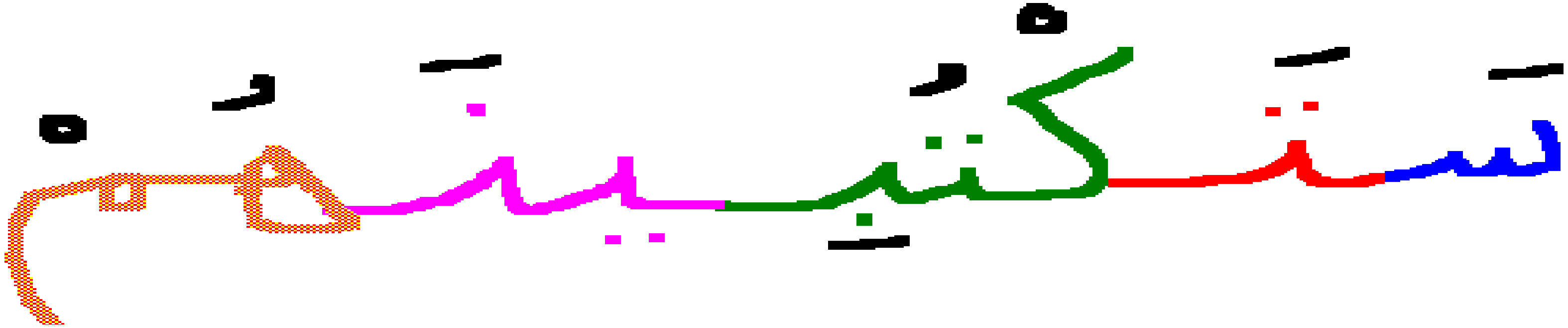
6. Arabic linguistic features

Morphology

Syntax

Vowels

…



Enclitic	Suffix Morphem	Radical	Prefix Morphem	Proclitic
<hom’>	<yna>	< ktobi>	<ta>	<sa>
Object Masculine Plural	Subject Feminine singular	Root "ktb"	Subject-Pronoun "you"	Temporal affix

You (féminine) will write them (masculine)

Examples of morphological features

Inflected morphology by prefix (or by determination) :

مجتمع (society) : المجتمع (the society)

Inflected morphology by suffix :

مجتمع (society) : مجتمعات (societies)

Inflected morphology by infix :

علم (science) : علوم (sciences)

Irregular plural :

إمرأة (femme) : نساء (femmes)

Derivational morphology :

كتب (book/write) : كتابة (writing)

Arabic vowels

Le mot سلم

- signifie "paix" lorsqu'il est voyellé ainsi : سِلْمٌ

- signifie "échelle" lorsqu'il est voyellé ainsi : سُلَّمٌ

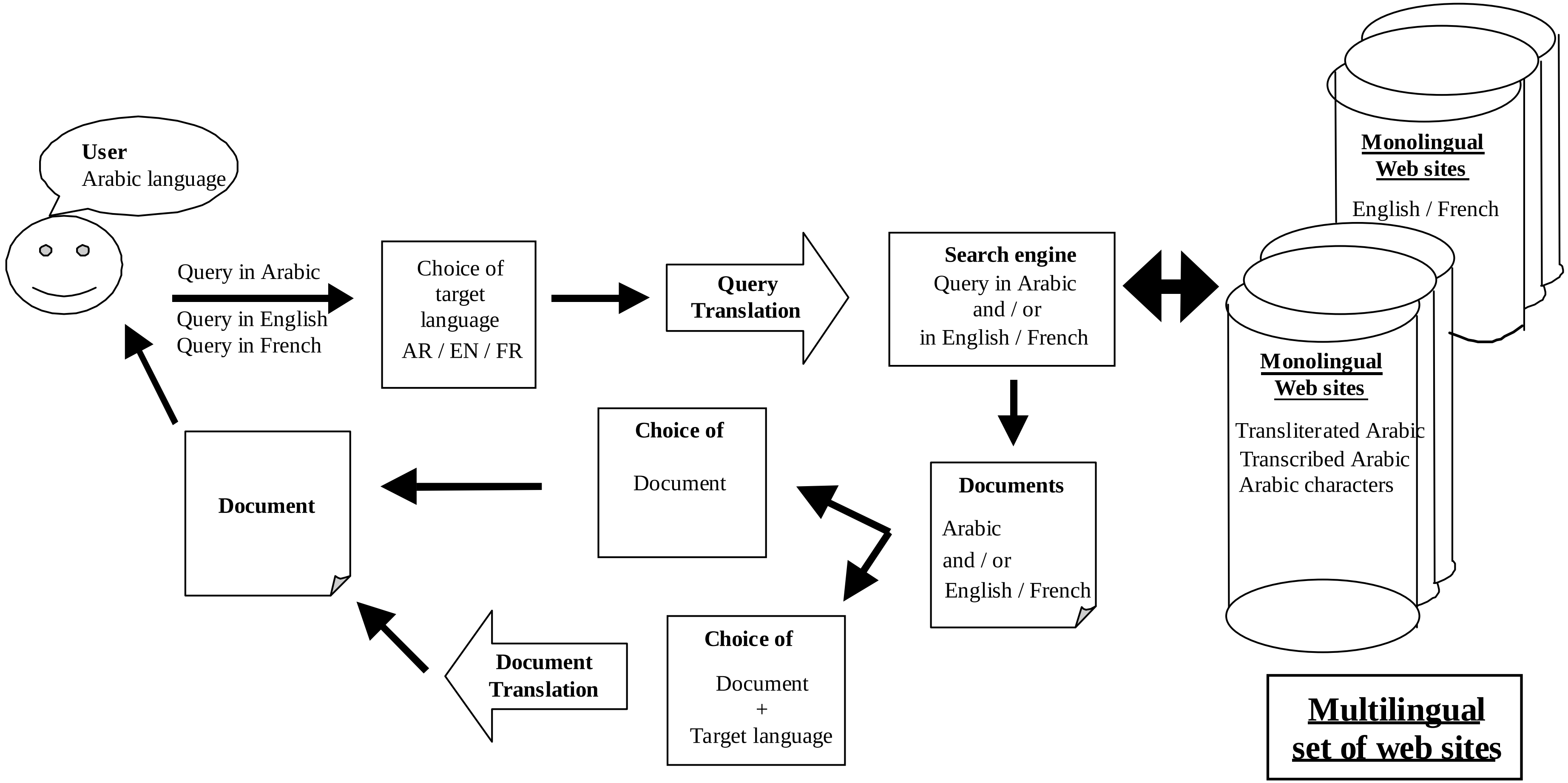
- signifie "a transmis" lorsqu'il est voyellé ainsi : سَلَّمَ

- signifie "est guéri" lorsqu'il est voyellé ainsi : سَلِمَ

- *Arabic texts that are published on the Web could be with/without/partially-with vowels.*
- *Queries on search engines should not contain vowels.*
- *Search engines should take this important morpho-syntactic feature into account.*

Cross-Language Information Retrieval

Most of the Arabic countries are bilingual



7. Natural language processing

The eaats mouse tthe cat

Orthography : NO ➨ **Lexicons and morphology**

The eats mouse the cat

Orthography : YES

Syntax : NO ➨ **Grammar**

The mouse eats the cat

Orthography : YES

Syntax : YES

Semantic : NO ➨ **Meaning**

(animal or computer ?)

The mouse eats the cat

Orthography : YES

Syntax : YES

Semantic : YES

Pragmatic : NO ➨ **Domain, real word**

(mousses do not eat cats !)

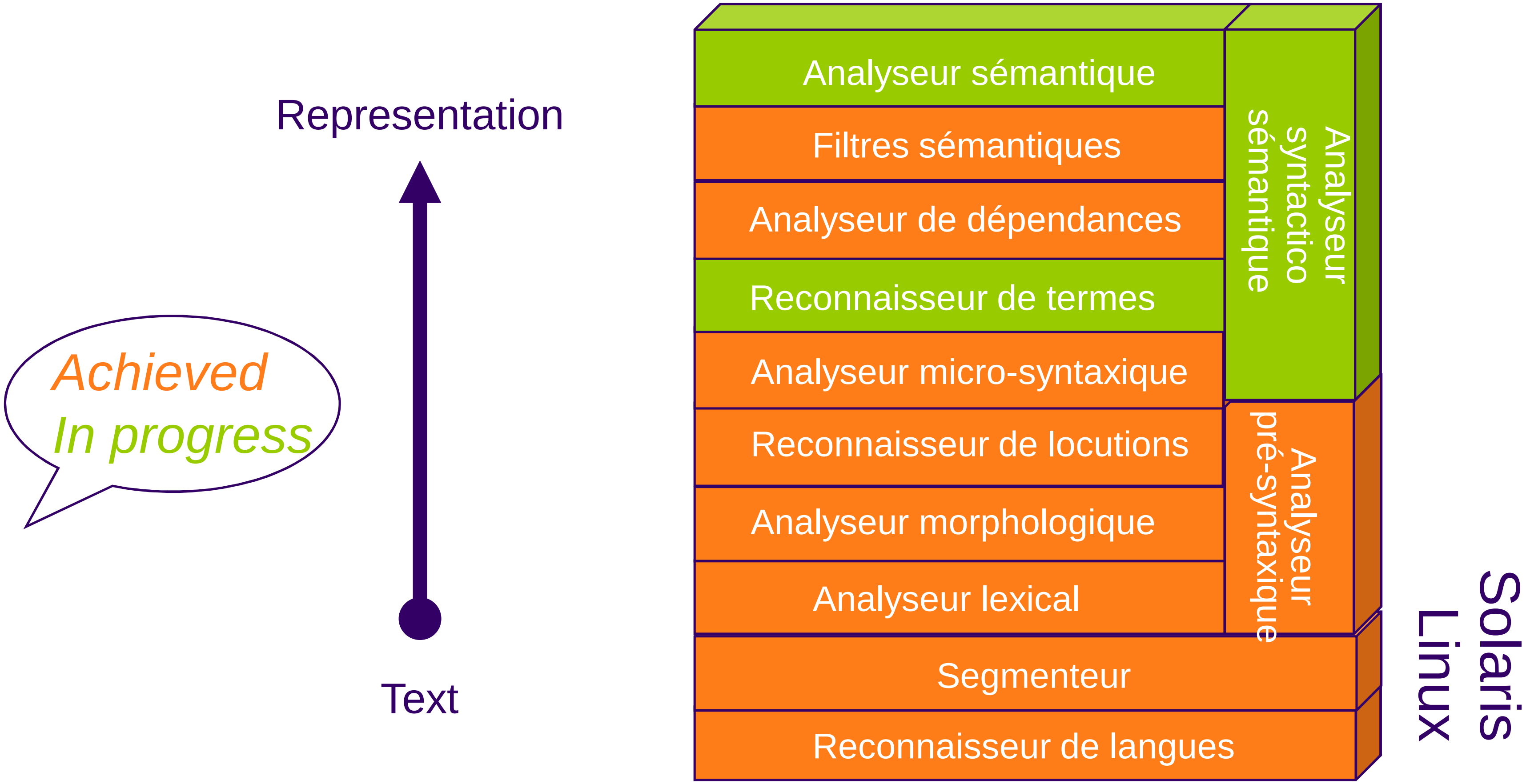
8. Arabic NLP activities at France Telecom R&D

Natural Language Processing Group
14 members + 10 non members
Responsible : Jérôme Vinesse

France Telecom R&D
DMI Division
GRI laboratory (Data management and information retrieval)

Address :
France Telecom R&D
2, avenue Pierre Marzin
22307 Lannion
France

Linguistic analysis tool : TiLT



Representation

Text

Achieved

In progress

Analyseur sémantique

Filtres sémantiques

Analyseur de dépendances

Reconnaisseur de termes

Analyseur micro-syntaxique

Reconnaisseur de locutions

Analyseur morphologique

Analyseur lexical

Segmenteur

Reconnaisseur de langues

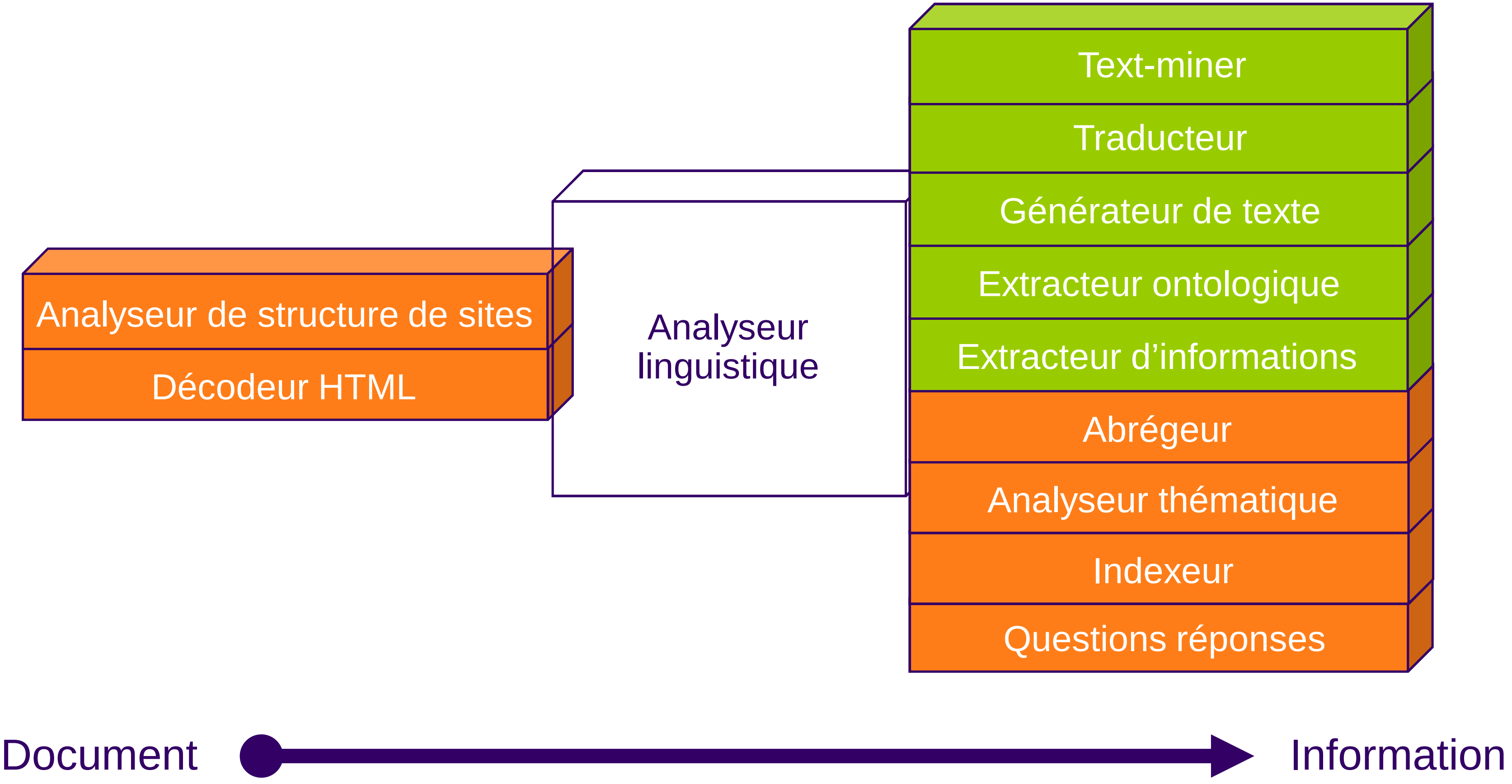
Analyseur syntactico sémantique

Analyseur pré-syntaxique

Solaris

Linux

Application modules



Catalogue

Language and charset Identification

Thematic classification

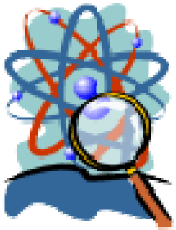
Automatic generation of linguistic exercises

Linguistic filtering

Question / Answering

Linguistic localization

Summarization



HTML parsing

Languages

French

English

Spanish

German

Polish

Arabic

Contact

Malek Boualem

Natural Language Processing Group
France Télécom R&D
DMI/GRI Laboratory
2, avenue Pierre Marzin
22307 Lannion
France

Phone: (33)(0)2.96.05.29.83
Fax: (33)(0)2.96.05.32.86
Email: malek.boualem@rd.francetelecom.com