

Recherche d'information et Traitement automatique de la langue arabe



Institut du Monde Arabe, Paris
Juin 2001

Malek Boualem, FTR&D - DMI/GRI

This document contains information which belongs to France Telecom. The acceptance of this document by the addressee implies on behalf of the recipient, the recognition of the content's confidentiality and the agreement of making no reproduction, transmission to third parties, disclosure and commercial use without prior written authorization from France Telecom R&D.

Contenu de la présentation



1. **Marché arabe de l'Internet et des télécommunications**
2. **Le Web arabe**
3. **Annuaire et portails arabes**
4. **Moteurs de recherche en langue arabe**
5. **Problèmes de la recherche d'information en langue arabe**
6. **Caractéristiques morphologiques et syntaxiques de la langue arabe**
7. **Traitement automatique du langage naturel**
8. **Activités de TALN à France Télécom R&D**

Contact

1. Marché arabe de l'Internet et des télécommunications



6ème langue

(Chinois, Anglais, Russe, Espagnol, Hindi, Arabe, ...)

Population arabophone : plus de 220.000.000

Chaque pays possède son propre dialecte parlé.
L'arabe écrit est strictement le même.

L'informatique, Internet et les télécommunications sont devenus populaires dans le monde arabe.

2. Le Web arabe

- Sites Web arabes :
 - disponibilité d'outils pour la création et la navigation (MS, Word, FrontPage, Explorer ...).

- Systèmes plus stables pour le codage des documents arabes :
 - **Microsoft CP1256**
 - **ISO-8859-6**
 - **UNICODE-UTF8**

- Nombre de sites arabes en forte croissance.



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 جوّاً متنوع الثقافات ومتميزاً بالحياة والانفتاح

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



3. Annuaires et portails arabes

Arabe et anglais :

- <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>

Arabe, anglais et français :

- <http://www.hahooa.com>



4. Moteurs de recherche en langue arabe



- **Moteurs sur des annuaires**

- **Extense-Voila** (Echo-Wanadoo, France)
<http://www.hahooa.com>
- **Ayna** (Ayna, USA)
<http://www.ayna.com>
- **Al-Idrisi** (Sakhr, Egypte)
<http://www.sakhr.com/>
- **Konouz** (Alladin, Australie)
<http://www.konouz.com/>

- **Moteur web grand public**

- **Arabvista** (Emirates Internet & Multimedia and COMPAQ)
<http://www.arabvista.com>

5. Problèmes de la recherche d'information en langue arabe



Quelques tests effectués sur un moteur de recherche :

Mot clé "مجتمع" (société) :	3012	réponses
Mot clé "المجتمع" (la société) :	20695	réponses
Mot clé "مجتمعات" (sociétés) :	599	réponses
Mot clé "علم" (science) :	12410	réponses
Mot clé "علوم" (sciences) :	3925	réponses
Mot clé "إمرأة" (femme) :	193	réponses
Mot clé "نساء" (femmes) :	11970	réponses
Mot clé "كتب" (livre/écrire) :	44133	réponses
Mot clé "كتابة" (écriture) :	2990	réponses
Mot clé "تعليم" (enseignement) :	46728	réponses
Mot clé "تدريس" (enseignement) :	770	réponses

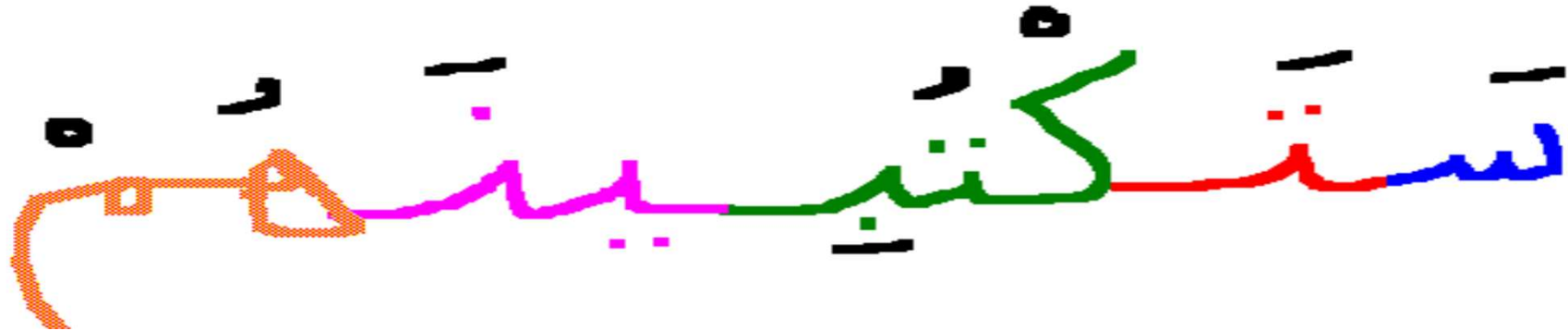


Quel est le problème au juste ?

Contrairement aux moteurs de recherche pour l'anglais ou le français, où les traitements linguistiques peuvent être contournés, des traitements linguistiques sophistiqués sont nécessaires pour **l'indexation** et la **recherche d'information** en langue arabe.

Cette nécessité est liée aux propriétés graphématiques, morphologiques, syntaxiques, sémantiques et même pragmatiques de la langue arabe.

6. Caractéristiques morphologiques et syntaxiques de la langue arabe



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

Tu (féminin) vas les (masculin) écrire

Exemples de phénomènes morphologiques de l'arabe



Morphologie flexionnelle par préfixe (ou pour la détermination) :

مجتمع (société) : المجتمع (la société)

Morphologie flexionnelle par suffixe :

مجتمع (société) : مجتمعات (sociétés)

Morphologie flexionnelle par infixe :

علم (science) : علوم (sciences)

Phénomène du pluriel brisé :

إمرأة (femme) : نساء (femmes)

Morphologie dérivationnelle :

كتب (livre/écrire) : كتابة (écriture)

Voyelles en langue arabe

Le mot سلم

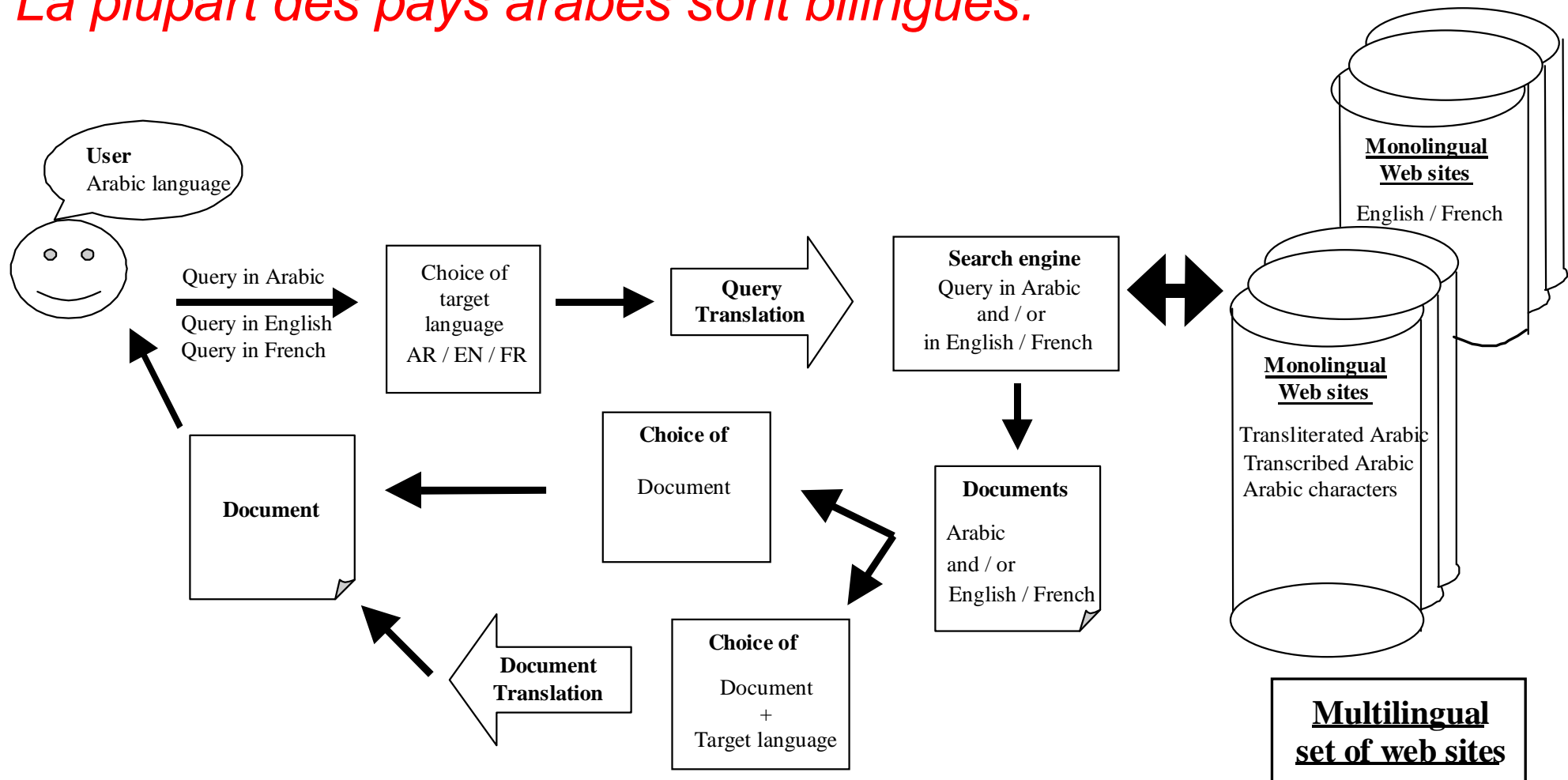
- signifie "paix" lorsqu'il est voyellé ainsi : سلم
- signifie "échelle" lorsqu'il est voyellé ainsi : سلم
- signifie "a transmis" lorsqu'il est voyellé ainsi : سلم
- signifie "est guéri" lorsqu'il est voyellé ainsi : سلم

Les documents arabes publiés sur le Web peuvent être non voyellés, partiellement ou entièrement voyellés. A priori, les requêtes sur les moteurs de recherche ne sont pas voyellées. Les moteurs de recherche devraient traiter cet aspect morpho-syntaxique important de l'arabe.



Recherche d'information inter-langues (Cross-Language Information Retrieval)

La plupart des pays arabes sont bilingues.





7. Traitement automatique du langage naturel

La mangge souris lle chat

Orthographe : **NON**

➡ **Lexique et morphologie**

La mange souris le chat

Orthographe : **OUI**

Syntaxe : **NON**

➡ **Grammaire**

La souris mange le chat

Orthographe : **OUI**

Syntaxe : **OUI**

Sémantique : **NON**

➡ **Sens**

(animal ou informatique ?)

La souris mange le chat

Orthographe : **OUI**

Syntaxe : **OUI**

Sémantique : **OUI**

Pragmatique : **NON**

➡ **Domaine, monde réel**

(la souris ne mange pas le chat !)



8. Activités de traitement automatique des langues naturelles à France Télécom R&D

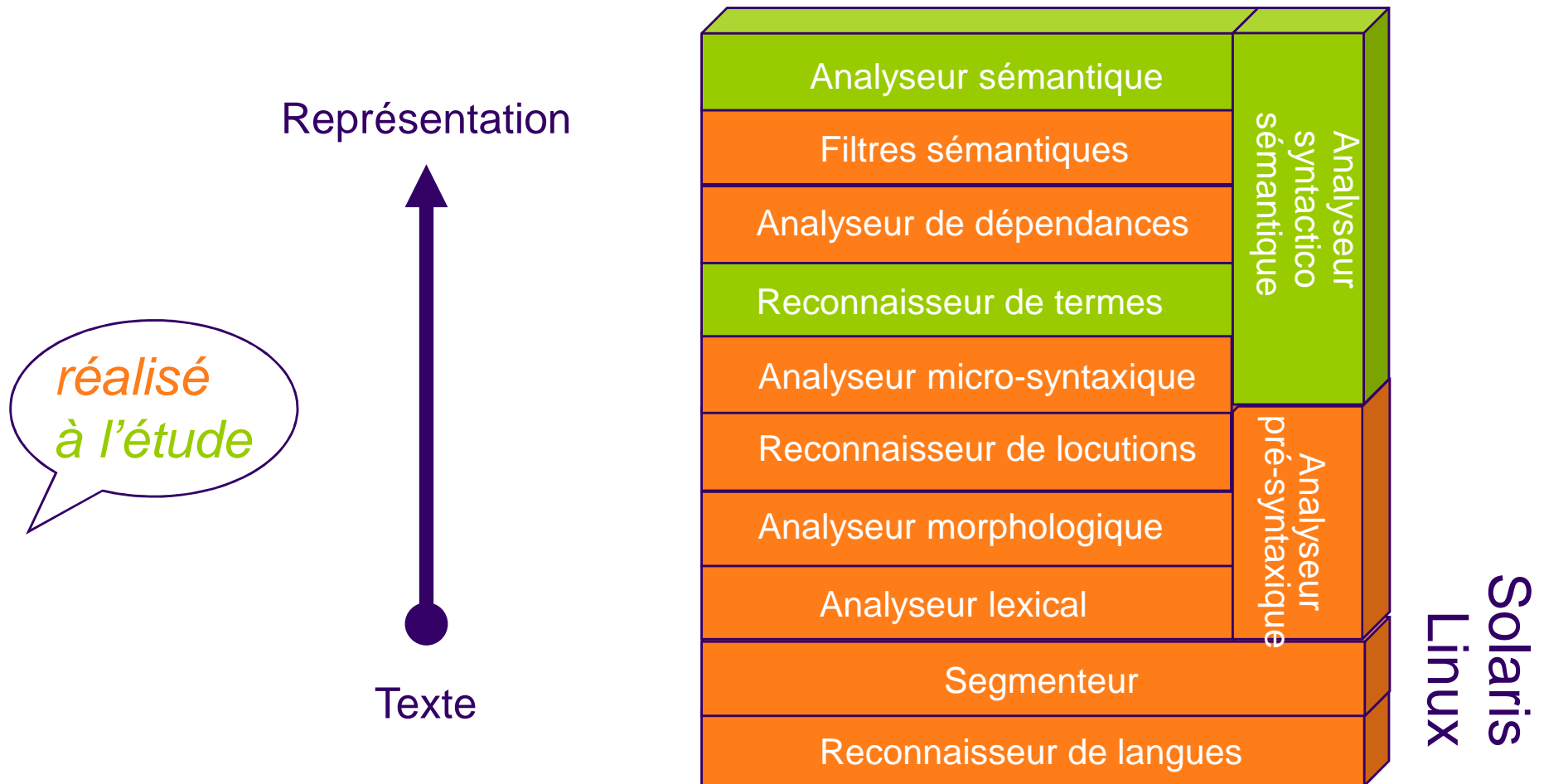
Équipe Traitement automatique des langues naturelles
14 permanents + 10 non permanents
Responsable : Jérôme Vinesse

France Télécom R&D
Direction DMI
Laboratoire GRI (Gestion et Recherche de l'Information)

Localisation :
France Télécom R&D
2, avenue Pierre Marzin
22307 Lannion
France

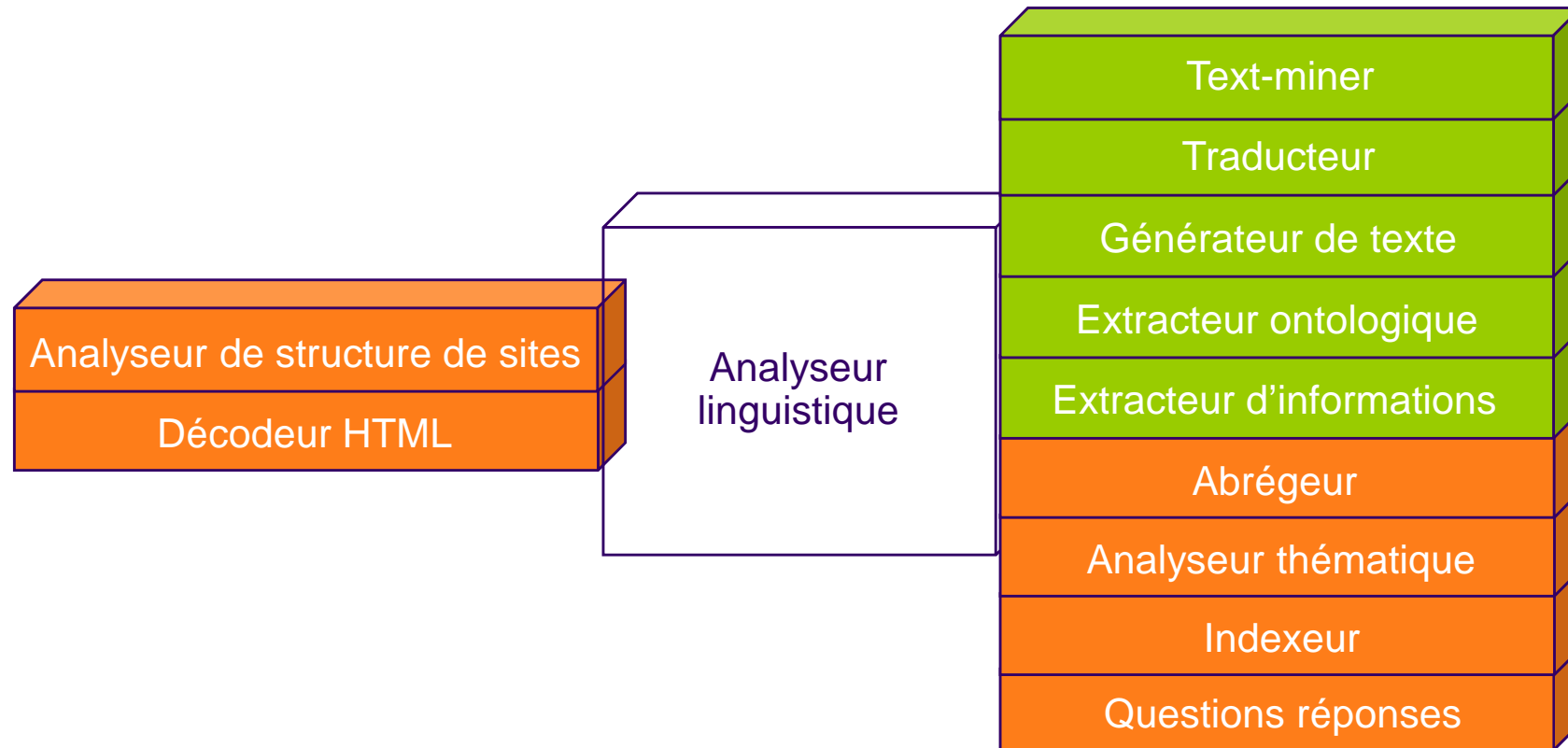


Analyseur linguistique TiLT





Modules dédiés



Document



Information

Catalogue



**Identification
de la langue**



**Classement
thématique**



**Génération
d'exercices
linguistiques**



**Filtrage
linguistique**



**Questions
Réponses**



**Localisation
linguistique**



**Résumé
(abrégé)**



**Analyse du contenu
et de la structure
de pages HTML**

Langues couvertes ou à l'étude



français

anglais

espagnol

allemand

polonais

arabe



Malek Boualem

Equipe Traitement automatique des langues naturelles

France Télécom R&D

Laboratoire DMI/GRI

2, avenue Pierre Marzin

22307 Lannion

France

Tel: (33)(0)2.96.05.29.83

Fax: (33)(0)2.96.05.32.86

Email: malek.boualem@rd.francetelecom.com