Introduction (i)

Dans le projet LiSe, pour la partie ***Data et Sense Mining***, nous traitons l'information dans le domaine de la sécurité

Dans cette application du TAL, deux méthodologies fondamentales sont appliquées : la méthodologie utilisant des mots-clés et une méthodologie reposant sur notre modélisation linguistique. Cette dernière dépend des structures linguistiques et d'une base de données « intelligente » permettant d'extraire l'information et d'analyser son contenu afin de pouvoir l'évaluer.

Introduction (ii)

L'information extraite est l'union de plusieurs éléments : le verbe considéré comme la source d'information fondamentale avec ses satellites (éléments qu'il régit)

Exemple d'information :

"Le mois dernier, la bande a acheté des armes en Iraq clandestinement".

verbe(évènement, temps, endroit, manière)

De cette information on peut extraire les indices suivants :
- l'évènement (achat d'armes),
- le temps (mois dernier),
- l'endroit (Iraq),
- la manière (clandestinement).

Modélisation

Niveaux de la modélisation :

- Étude et analyse du système verbal ;
- Étude des structures phrastiques simples et complexes;
- Analyse sommative tenant compte de l'ensemble des particules linguistiques (indicatives de modes et aspects…).

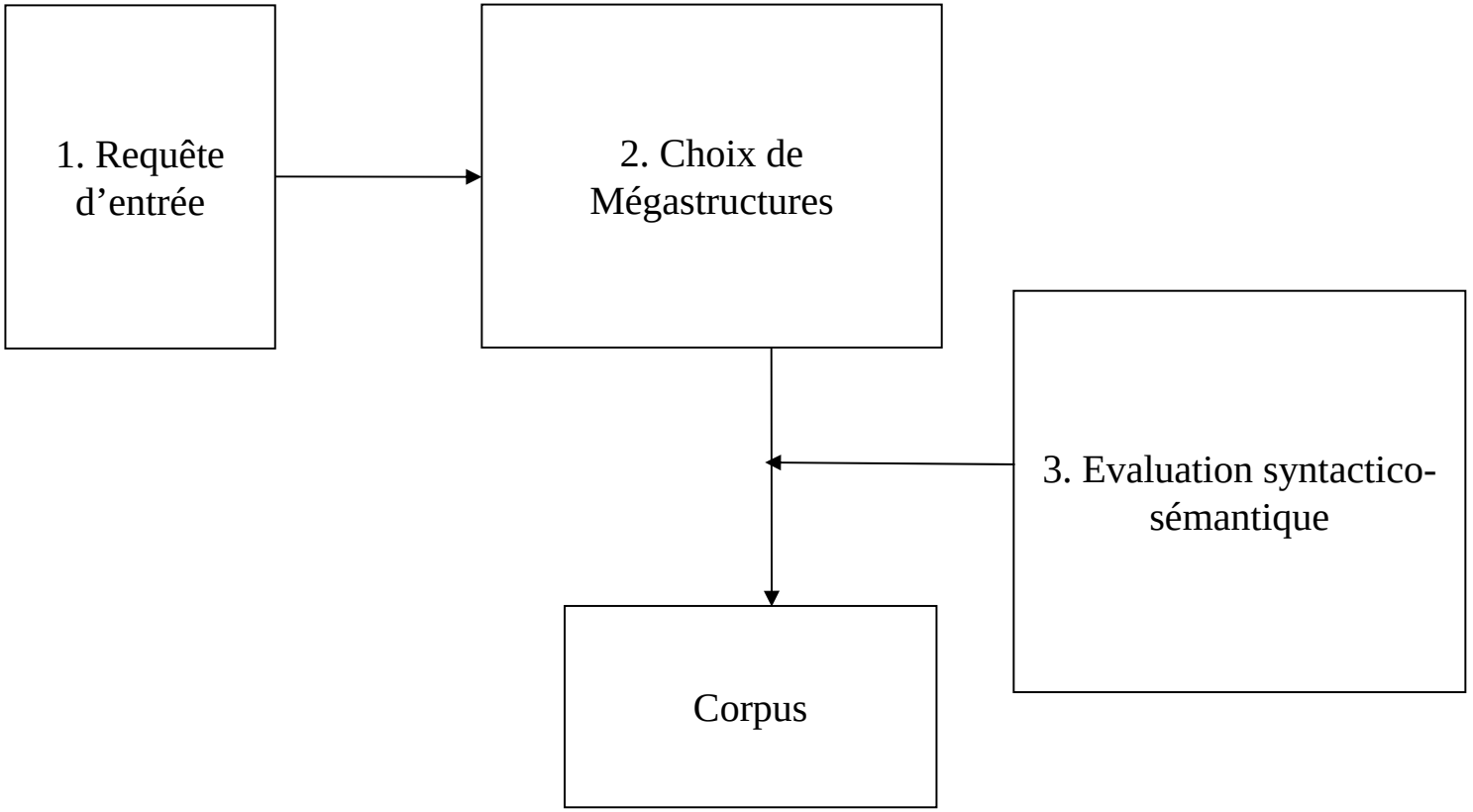
Σ particules linguistiques (contextuelles, modales, aspectuelles) → évaluation sommative (syntactico-sémantique)

Exemple de modélisation

Phrase arabe	Structure de la phrase
Requête d'entrée : أصاب + التلوثُ + الماءَ ←→ eau+ pollution + contamination	
أصاب التلوثُ الماءَ	complément(eau) + nom(pollution) + verbe(contaminer).
La pollution a contaminé l'eau (probabilité : information certaine - temporalité: passé)	
ما أصاب التلوثُ الماءَ	complément(eau) + nom(pollution) + verbe(contaminer) + Adverbe_négation(ne…pas).
La pollution n'a pas contaminé l'eau (probabilité : information rejetée - temporalité : passé)	
ربما أصاب التلو ثُ الماءَ	complément(eau) + nom(pollution) + verbe(contaminer) + Adverbe_probabilité(peut-être)
La pollution a peut-être contaminé l'eau (probabilité : information probable - temporalité: passé)	

Partie portant sur l'arabe

Architecture arabe



Atouts du système arabe

L’avantage de notre système est d’être beaucoup plus performant qu’un simple moteur de recherche par mots-clés. En effet, au moment où la requête de l’utilisateur est découpée en constituants, le système, via les mégastructures phrastiques définies, essaie tous les équivalents de la requête. Ces équivalents peuvent être parfaits s’ils contiennent les mêmes constituants que ceux de la requête (même dans un ordre différent que celui de la phrase trouvée dans le corpus).

Partie portant sur le chinois

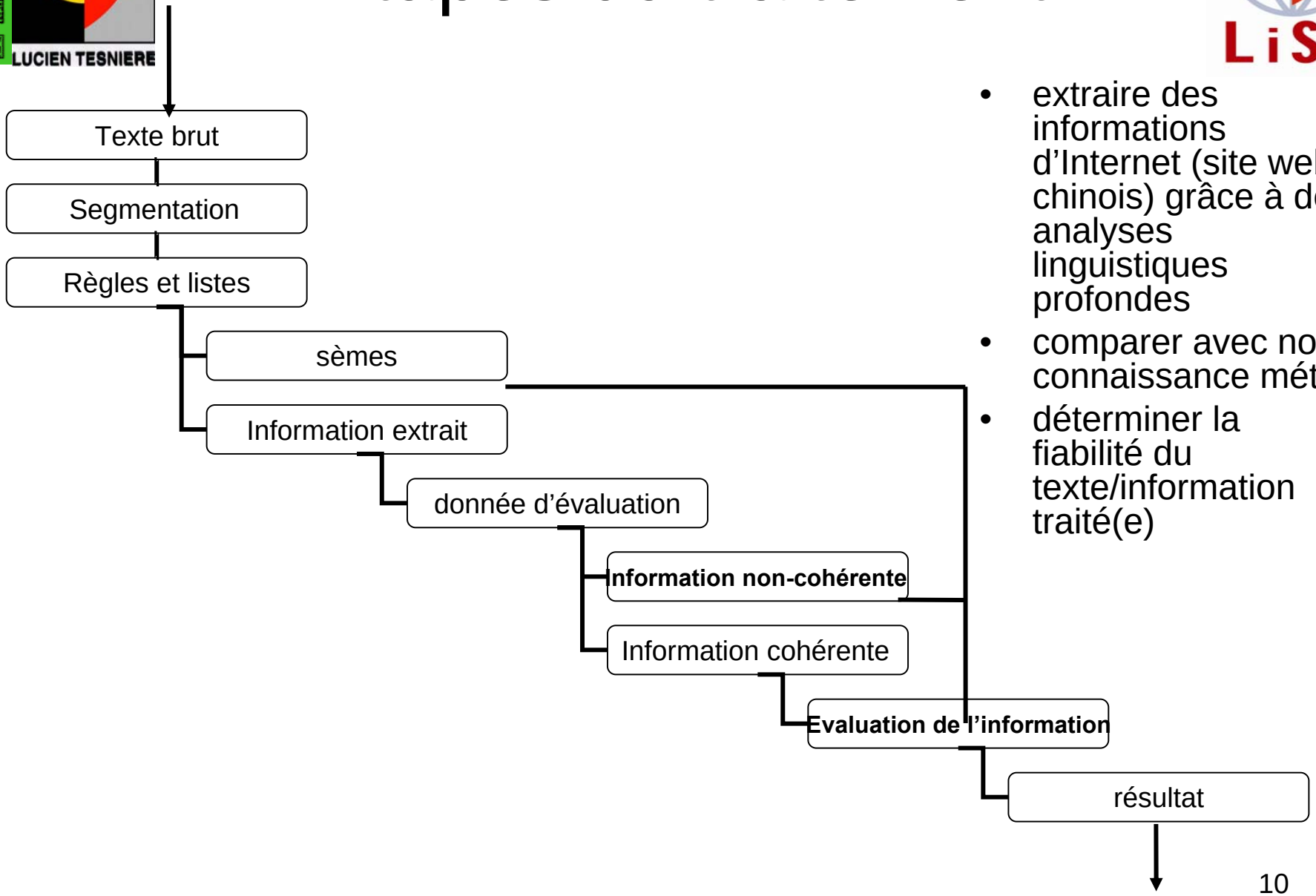
Objectifs principaux de la recherche sur le chinois

1. Déterminer la précision du texte/information dans un contexte de crise avec connaissance du domaine.
2. La surveillance de l'activité hostile: sécurité des états (crises nucléaires dans la Corée du Nord) et compagnies (boycottage de Carrefour)

Contexte d'Internet en Chine :

- Un cadre favorable pour les discussions politiques (démocratisation à la chinoise selon certains politiciens).
- La présence du public général avec un intérêt pour les sujets militaires sur tous les principaux sites.

Etapes du traitement



- extraire des informations d'Internet (site web chinois) grâce à des analyses linguistiques profondes
- comparer avec notre connaissance métier
- déterminer la fiabilité du texte/information traité(e)

Exemple de corpus

- (…)
- 朝鲜于当地时间2009年4月5号上午11时30分发射了光明星2号试验通信卫星，卫星还是导弹应该能够予以判断。朝鲜公布的光明星2号卫星轨道近地是228公里，远地是400多公里。而且在UHF波段以470兆赫的频率播放《伟大领袖 金正日之歌》和《金日成将军之歌》。所以包括中国在内都在收听，但从4月5号到现在，还没有一个国家收听到这两首革命歌曲。
- (…)
- *(Traduction: 11H30 temps local, 05/04/2009, Corée nord a lancé le satellite 'vedette clair' No.2, nous pouvons déterminer que c'est un missile guidé et pas un satellite. Parce que la Corée du Nord a annoncé que l'emplacement le plus proche de l'orbite du satellite 'vedette clair' No.2 est de 228km-Ray, l'emplacement lointain de l'orbite du satellite 'vedette clair' No.2 est de 400km-Ray. Et (le Satellite) a diffusé " La chanson pour notre grand chef Kim Jong-il" et "La chanson pour notre grand général Kim Jong-il" en fréquence 470mhz sur la bande l'UHF. Tout le monde essaie de localiser les chansons. Mais du 5 avril jusqu'à ce moment (16/04/2009), personne ne peut entendre les 2 chansons.)*
-
- « 朝鲜发射了卫星还是导弹，是成功还是失败？ » *2009-04-16 09:15:01.*
- (*Traduction: La Corée du Nord a lancé un satellite ou un missile guidé, succès ?*)
- *2668 caracters.*
- *http://military.china.com/zh_cn/critical3/27/20090416/15434727.html*