

# White Paper Lingway



**Pas de Web sémantique sans  
traitement automatique des langues**

OCTOBRE 2004

**>lingway**

Solutions in language processing

# White Paper Lingway

<b>SOMMAIRE</b>	<b>1</b>
<b>1 INTRODUCTION</b>	<b>2</b>
<b>2 LE WEB SEMANTIQUE</b>	<b>3</b>
2.1 EVOLUTION OU ALTERNATIVE ?	3
2.2 LES LANGAGES DU WEB SEMANTIQUE	4
2.2.1 RDF	4
2.2.2 OWL	4
<b>3 LE TRAITEMENT AUTOMATIQUE DES LANGUES : TAL</b>	<b>5</b>
3.1 LES NIVEAUX DE L'ANALYSE LINGUISTIQUE	5
<b>4 TAL ET WS : UN CERCLE VERTUEUX</b>	<b>7</b>
4.1 UTILISATION D'ONTOLOGIES PAR LES SYSTEMES DE TAL	8
4.1.1 Interopérabilité des ontologies	8
4.1.2 Utilisation d'ontologies dans un système de TAL	9
4.2 UTILISATION DU WEB COMME SOURCE D'APPRENTISSAGE POUR LE TAL	10
4.3 UTILISATION DU TAL POUR « STRUCTURER LE WEB »	11
4.3.1 Extraction d'entités nommées	11
4.3.2 Extraction de descripteurs thématiques	12
4.3.3 Associer des attributs aux descripteurs	13
4.4 UTILISATION DU TAL POUR FACILITER LA RECHERCHE	14
<b>5 CONCLUSION</b>	<b>15</b>

# White Paper Lingway

## 1 Introduction

La notion de « Semantic Web », que l'on n'ose pas traduire par « Toile Sémantique », prend de plus en plus d'importance et devient l'une des voies d'évolution les plus prometteuses du Web, tel qu'on le connaît aujourd'hui. La définition la plus connue est celle donnée par Tim Berners-Lee dans un article paru dans Scientific American en Mai 2001, et citée sur le site du W3C (<http://www.w3.org/2001/sw/>).

**« Le Web sémantique est une extension du Web actuel, dans laquelle l'information reçoit une signification bien définie, améliorant les possibilités de travail collaboratif entre les ordinateurs et les personnes »**

Cela semble donc signifier, dans l'esprit des auteurs, que l'information n'a pas de signification bien définie dans le Web actuel. C'est effectivement vrai d'un point de vue d'informaticien puisque la majeure partie de l'information sur le Web est sous forme textuelle, très peu structurée et donc inutilisable pour faire des traitements de calcul ou d'inférences. Il est pourtant bien évident que l'information disponible sur le Web actuel a une signification, mais qu'elle n'est accessible aujourd'hui qu'à des lecteurs humains.

Or, si les travaux et normes en cours d'élaboration sur le Web Sémantique proposent de structurer, coder, catégoriser l'information, en utilisant des langages XML particuliers comme RDF ou OWL, ils ne disent rien en revanche sur la façon de transposer dans ces formalismes l'information contenue dans les textes.

Si l'on peut penser qu'une partie de l'information nouvelle qui sera publiée dans les prochaines années sur le Web sera effectivement « indexée » dès sa production dans ces formalismes du Web Sémantique, au moins, pour certains types d'informations plus ou moins factuelles et donc facilement codables, il n'en reste pas moins que la majeure partie de l'information restera sous forme textuelle et que la question de son indexation dans les langages du Web Sémantique deviendra cruciale.

C'est justement l'objectif Traitement Automatique des Langues (TAL) que de comprendre et modéliser le langage humain, en quelque sorte de « rendre le texte calculable ».

LINGWAY [www.lingway.com](http://www.lingway.com) est spécialisée dans le développement de logiciels de traitement de l'information basés sur le TAL. Ce document présente son approche par rapport à cette problématique, pour indexer automatiquement des documents textuels, par extraction d'entités nommées, extraction de descripteurs thématiques ou catégorisation par rapport à une taxonomie donnée. Pour illustrer notre propos, cette approche sera mise en correspondance avec les fonctionnalités de Lingway KM, l'offre logicielle de Lingway.

Nous montrerons également le rôle que peuvent jouer les outils de recherche multilingue et d'aide à la lecture comme le résumé et la cartographie.

# White Paper Lingway

## 2 Le Web sémantique

### 2.1 Evolution ou alternative ?

Une des questions qui se pose à propos du Web Sémantique est de savoir s'il s'agira d'une évolution du Web actuel, qui se transformera en se dotant de plus en plus de structure, ou si ce sera un « autre Web », parallèle à l'actuel, qui pourrait continuer à vivre très longtemps avec ses propres standards.

Par certains côtés, les « puristes » du Web sémantique réinventent avec une nouvelle technologie ce que l'on fait depuis toujours avec des SGBD relationnels : définition d'un modèle de données, règles de gestion, etc. : écrire une définition RDF est de ce point de vue finalement assez similaire à décrire une table en SQL.

Cette approche est évidemment intéressante pour traiter de domaines relativement restreints, finis et entièrement modélisables. C'est typiquement le champ d'action des applications de « e-commerce » sur Internet qui seront certainement les plus utilisatrices de ce type d'évolution (et qui en sont d'ailleurs la motivation initiale). Elle ne peut cependant pas s'appliquer sans adaptation à la modélisation de l'immense quantité d'informations non structurée présente dans les textes du Web actuel, qui est une source majeure pour d'autres applications comme la Veille ou le « Knowledge Management ».

Il paraît donc légitime de penser que, à côté d'un Web sémantique, structuré, à base de XML-RDF orienté vers des informations factuelles modélisables, va continuer à exister, et pour longtemps, un Web du type de celui que l'on connaît aujourd'hui, pas structuré, largement textuel, à base de pages HTML. Bien évidemment, ces deux mondes vont s'interpénétrer et toutes sortes de situations hybrides vont apparaître.

On peut sur ce point faire un parallèle avec l'évolution des bases de données depuis 25 ans : la généralisation des bases SQL n'a pas pour autant fait disparaître les bases de données textuelles, gérées avec des « systèmes documentaires ». Mais il est vrai que les deux mondes se sont interpénétrés, et que les SGBD relationnels d'aujourd'hui proposent des fonctions de recherche et stockage du texte alors que, parallèlement, des systèmes documentaires, initialement centrés sur des fonctions de recherche textuelle, ont petit à petit ajouté des fonctions de traitement de données factuelles associées.

Le point qui nous intéresse alors, du point de vue de l'utilisation des technologies du traitement de la langue, est double :

- d'une part, comment va-t-on reprendre la masse d'information textuelle existante pour l'indexer, même de façon rudimentaire dans les formalismes du Web sémantique, pour la rendre compatible avec les informations nouvelles qui seront créées dans les nouveaux formats ?
- d'autre part, comment les technologies linguistiques vont-elles aider à structurer les nouvelles informations qui seront publiées sur le Web dans les prochaines années ?

# White Paper Lingway

## 2.2 Les langages du Web sémantique

De nombreuses publications sont disponibles sur ce thème. Nous ne donnons ici que des définitions très simplifiées.

### 2.2.1 RDF

La définition du RDF (Resource Description Framework), telle que publiée par le W3C, précise que RDF est un langage pour représenter l'information sur les ressources du Web, particulièrement adapté pour représenter des méta données attachées à des documents disponibles sur le Web comme le titre, l'auteur, la date de modification, le copyright, etc. Par extension, ce même langage permet de représenter de l'information sur des objets identifiés sur le Web, même s'ils n'y sont pas directement retrouvables.

Un terme RDF comprend un « sujet », « un prédicat » (ou « verbe ») et « un objet ». Ce type de modèle est loin d'être nouveau. Il permet de représenter des faits complexes comme par décomposition en faits élémentaires. La nouveauté est ici que sujet, objet et prédicat sont décrits par des URI (Uniform Resource Identifiers) qui permettent de donner des significations précises, accessibles par tous sur le Web, aux concepts et aux prédicats.

### 2.2.2 OWL

Le problème de RDF est qu'il n'introduit aucune normalisation sur la sémantique proprement dite, et que divers utilisateurs pourront donner des noms différents à des entités de même sens, ce qui empêche d'imaginer toute agrégation entre des données provenant de sources Web différentes.

La solution vient du développement d'ontologies, qui décrivent formellement les termes et les relations entre eux.

OWL (Web Ontology Language) est un langage permettant de publier et partager des ontologies sur le Web. C'est une extension de RDF. OWL permet de donner des équivalences, d'un monde à l'autre (i.e. d'une ontologie à l'autre) entre des désignations d'objets. Par exemple, on pourra décrire que l'objet « ZipCode » dans l'application 1 est l'équivalent de l'objet « CodePostal » dans l'application 2.

# White Paper Lingway

## 3 Le Traitement Automatique des Langues : TAL

Avant de discuter en quoi le TAL peut aider à la réalisation du Web sémantique, nous ferons un bref rappel sur ces technologies.

### 3.1 Les niveaux de l'analyse linguistique

Des progrès ont pu être faits dans le domaine du TAL (Traitement Automatique des Langues) ou NLP en anglais (Natural Language Processing), à partir du moment où l'on a pu décomposer le problème, trop vaste pour être abordé globalement, en plusieurs sous problèmes, correspondant à plusieurs sous tâches articulées entre elles. Cette décomposition est faite à la fois au niveau des ressources linguistiques (dictionnaires et règles de grammaire) et au niveau des algorithmes (analyseurs, transducteurs, etc.).

Les principaux niveaux classiques sont :

- le niveau morphologique : identification des mots d'une phrase,
- le niveau syntaxique : identification des constituants d'une phrase et de leur fonction,
- le niveau sémantique : identification du sens des mots et de la structure logique d'une phrase,
- le niveau du texte : identification des relations entre les phrases et de la structure d'un texte,
- le niveau du corpus : identification de la structure d'un ensemble de textes.

Chaque niveau peut à son tour être décomposé en sous-niveaux.

Ainsi le niveau morphologique peut se re-décomposer en :

- la tokenisation : identification des frontières de mots (simples et composés), et de phrases,
- le tagging : identification de la catégorie - nom, verbe, adjectif - de chaque mot,
- la lemmatisation : identification de la forme canonique dans le dictionnaire.

Le niveau syntaxique est souvent décomposé en :

- le chunking : identification des frontières majeures de constituants (groupe nominal, verbal, etc.) et/ou des relations majeures entre les mots,
- le tagging fonctionnel : affectation de fonctions grammaticales aux constituants,
- le parsing : construction d'un arbre représentant la structure de la phrase complète.

Le niveau sémantique se décompose également en :

- la sélection de sens (WSD « word sense disambiguation ») : choix du sens de chaque mot,
- la structuration logique : identification des arguments de chaque prédicat et de leur rôle sémantique (agent, but, lieu, etc.).

# White Paper Lingway

Le niveau texte peut aussi se décomposer en :

- la résolution des anaphores (antécédents des pronoms, ellipses, références),
- la détermination de la structure rhétorique (commentaires, explications, causalités, etc.),
- la détermination de la structure thématique (de quels sujets le texte traite-t-il ?).

Enfin le niveau du corpus peut se décomposer en :

- la détermination de la nature des documents (article de presse, article technique, texte réglementaire, brochure commerciale, etc.),
- la structure thématique du corpus (de quels sujets le corpus traite-t-il ?).

La plupart des ces tâches correspondent à des sous disciplines de la linguistique. Du point de vue de l'ingénierie du TAL, elles correspondent à des modules, des programmes spécifiques qui ont chacun besoin d'un type de ressources linguistiques et d'un type d'algorithmiques particuliers.

## Réalisation dans Lingway KM

Tous ces modules sont implémentés et utilisés dans Lingway KM.

En particulier, Lingway KM est un produit intégré proposant :

- **un moteur de recherche sémantique** permettant de combiner l'interrogation des bases de texte intégral en langage naturel et par taxonomie (plan de classement) en plusieurs langues,
- **un moteur d'indexation et de catégorisation** basé sur un module de structuration XML permettant le marquage du texte à un niveau sémantique,
- **un moteur d'aide à la lecture** (résumé automatique, colorisation du texte et cartographie) permettant d'appréhender le contenu d'un document ou d'un ensemble de documents.

Lingway KM est basé sur un modèle linguistique à 3 niveaux :

- au niveau morphologique sont codées les informations relatives à la catégorie syntaxique du mot et ses caractéristiques flexionnelles : genre, nombre, mode de flexion,
- au niveau sémantique, on distingue les sens possibles des mots avec des attributs sémantiques (appartenance à une classe sémantique) et des relations sémantiques (relations de dérivation entre un verbe et un nom par exemple),
- au niveau conceptuel, les différents sens sont reliés à un concept, cette fois indépendant de la langue.

L'ensemble de ces données constitue un « dictionnaire » - en fait une véritable base de connaissances - de 150.000 concepts environ, reliés à 5 langues : anglais, français, allemand, espagnol, néerlandais.

# White Paper Lingway

## 4 TAL et WS : un cercle vertueux

Les technologies du TAL et du Web sémantique se recoupent de plusieurs manières. Elles vont se développer simultanément en se renforçant mutuellement dans un cercle vertueux.

D'une part, le Web Sémantique, au fur et à mesure de son développement, devient en lui-même une ressource pour le développement des technologies du TAL :

- les nombreuses ontologies déjà publiées, et qui vont forcément se multiplier, représentent une source de connaissance précieuse pour les outils basés sur le TAL. Cependant, ces ressources ne décrivent pratiquement jamais l'aspect linguistique des termes décrits dans les ontologies, qui restent de « simples » chaînes de caractères. Pour les rendre utilisables par des systèmes de TAL, il faut donc établir le lien entre les ontologies telles qu'elles sont disponibles sur le Web et un dictionnaire de langue adapté au TAL,
- plus le Web est structuré, plus il devient une source de connaissances pour le TAL. Cela est particulièrement vrai pour les applications de catégorisation, qui sont souvent basées sur des technologies d'apprentissage et qui tireront parti de l'indexation par des méta données facilement utilisables. Les nombreuses ressources multilingues naturellement alignées, qui sont constituées par les versions en différentes langues de très nombreux sites, notamment commerciaux, constituent également une ressource importante pour l'aide à la construction de dictionnaires, et plus généralement d'applications multilingues et « cross-language ».

D'autre part, les technologies du TAL vont contribuer à la construction du Web sémantique qui se structurera progressivement, en profitant de leur amélioration régulière et sans doute de plus en plus rapide :

- Pour aider à structurer progressivement le Web, d'une part en ajoutant des méta données aux pages déjà publiées, dans une logique de « retro-conversion », et d'autre part en proposant des outils d'aide à la structuration pour les pages à venir,
- pour aider à construire des ontologies et des ressources multilingues, en utilisant des techniques d'extraction et de structuration d'informations à partir des textes, et en combinant ces données nouvelles avec des ontologies partielles déjà validées par des experts humains,
- pour aider à rechercher l'information textuelle, en ajoutant des méthodes linguistiques et sémantiques dans les moteurs de recherche tels qu'on les connaît aujourd'hui, ce qui améliorera sensiblement la qualité des recherches, notamment en gérant l'accès multilingue.

Pour chacun de ces points une présentation rapide de la problématique est complétée par un aperçu des solutions apportées par Lingway pour y répondre.

# White Paper Lingway

## 4.1 Utilisation d'ontologies par les systèmes de TAL

Des quantités d'ontologies diverses vont être – et sont déjà – publiées sur le Web. OWL étant très récent, peu le sont encore dans ce format, mais des formats précurseurs ont déjà été utilisés, et il existe déjà de très nombreuses ressources disponibles. Certains auteurs estiment que ce sont des milliards d'éléments d'informations codés sous la forme de triplets RDF, qui vont être créés et accessibles, ce qui fait surgir immédiatement deux nouveaux problèmes : comment assurer une certaine cohérence et interopérabilité entre ces ontologies d'une part, et comment les intégrer dans des systèmes de TAL ?

### 4.1.1 Interopérabilité des ontologies

Une des primitives importantes d'OWL est « same-as » qui permet de spécifier que deux URI désignent en fait le même « objet ». Ce mécanisme est central dans le Web sémantique, puisque le même objet aura toutes les chances d'être nommé différemment dans diverses ontologies.

Par exemple, si une ontologie sur un site donné contient une entrée libellée « Mr Jacques Chirac » et une autre une entrée libellée « Le président Chirac », il ne sera pas possible de les unifier sauf à les lier explicitement par un lien « same as ». On peut toujours supposer que les efforts de millions d'internautes – particuliers ou organisations – vont petit à petit converger pour tisser de tels liens, mais cela risque d'être un processus fort long. D'autant que si dans l'exemple précédent repérer l'identité des deux éléments d'information paraît assez simple, on arrivera vite à des cas plus complexes : par exemple, imaginons dans une ontologie donnée une description de « scrutin majoritaire » et dans une autre de « élection à la majorité ». Ces deux descriptions ne pourront pas être partagées tant qu'un lien « same-as » n'aura pas été créé entre les deux.

### Réalisation dans Lingway KM

Les outils Lingway offrent des méthodes de comparaison et de rapprochement d'ontologies. Il est ainsi possible sur la base des dictionnaires sémantiques du produit, qui connaît le lien entre « scrutin » et « élection » d'une part, et entre « majoritaire » et « majorité » d'autre part, de rapprocher automatiquement ces deux termes, et donc d'établir un lien entre les deux ontologies.

De plus, Lingway est capable d'établir ces mêmes correspondances entre des ontologies exprimées dans des langues différentes.

# White Paper Lingway

## 4.1.2 Utilisation d'ontologies dans un système de TAL

Les systèmes basés sur le TAL, notamment les moteurs de recherche et de « text-mining » ont intérêt à réutiliser l'information modélisée dans ces ontologies. Néanmoins, elles ne sont généralement pas directement utilisables car elles ne contiennent aucune information linguistique du niveau lexical.

Par exemple, si une ontologie dans le domaine informatique indique que « souris » a une relation avec « périphérique », cette information devient ambiguë dans un système linguistique, ces deux mots ayant plusieurs sens, et cela risque de générer des interprétations indiquant une relation entre un rat (par voisinage avec l'animal souris) et le boulevard périphérique.

### Réalisation dans Lingway KM

Les outils Lingway permettent cette mise en correspondance entre une ontologie externe et l'ontologie linguistique interne au produit. Pour cela, l'ontologie Lingway éclate chaque mot en plusieurs sens ou « concepts », qui sont chacun décrits par des domaines et des classes.

La notion de domaine recouvre celle de domaine d'activité dans lequel un mot est utilisé avec un sens donné. Par exemple, le concept @SOURIS : APPAREIL appartient au domaine informatique. Il existe une vingtaine de domaines, hiérarchisés en plus de 400 sous domaines, sur une hiérarchie à 6 niveaux.

Par ailleurs, les sens des mots peuvent appartenir à une ou plusieurs classes, dont l'objet est de décrire des propriétés intrinsèques de l'objet. Par exemple, souris au sens de « périphérique » a pour classe INSTRUMENT, tandis que souris au sens d'animal a pour classe MAMMIFERE. Il existe près de 500 classes prédéfinies dans notre modèle de dictionnaire.

Diverses procédures d'analyse, incluant des heuristiques de désambiguïsation sémantique permettent alors d'analyser des termes d'une ontologie et de les rendre utilisables dans Lingway KM, dans les diverses fonctions de moteur de recherche, d'extraction-catégorisation ou d'aide à la lecture.

# White Paper Lingway

## 4.2 Utilisation du Web comme source d'apprentissage pour le TAL

Inversement, la masse d'informations actuellement disponible sur le Web devient une source pour l'apprentissage de systèmes de TAL.

Les méthodes basées sur l'apprentissage se basent sur des corpus annotés, et vont rechercher à construire automatiquement des systèmes capables de reproduire des annotations similaires sur des textes non encore annotés.

Les annotations peuvent être plus ou moins fines : au niveau le plus grossier on peut considérer que le classement d'un document dans une catégorie donnée d'une taxonomie constitue un premier niveau d'annotation. C'est souvent le niveau effectivement utilisé par des systèmes dits de « catégorisation ». A l'autre bout du spectre, on trouve des corpus annotés à un niveau très fin – même au niveau de chaque mot- , généralement réalisés par des équipes universitaires qui constituent des ressources précieuses pour l'apprentissage.

A titre d'exemple, on peut utiliser pour un apprentissage :

- des sites en plusieurs langues : de très nombreux sites sont en plusieurs langues. Pour ce qui est des sites commerciaux, ils sont très souvent dans leur langue nationale et en anglais. Cela constitue un important réservoir de corpus multilingues alignés qui peuvent être utilisés pour aider à construire des terminologies multilingues. Cet alignement naturel peut être vu comme une forme d'annotation utilisable par un système d'apprentissage,
- des sites catégorisés : de nombreux sites ont d'ores et déjà classé leur contenu dans des plans de classements, taxonomies qui leur sont généralement propres, mais qui parfois sont issus de grandes classifications internationales (MeSH en médecine, par exemple).

### Réalisation dans Lingway KM

Les techniques d'apprentissage à partir de grands corpus en général sont largement utilisées par Lingway KM :

- pour la construction de terminologies monolingues, en utilisant les extracteurs de termes de Lingway KM,
- pour établir des dictionnaires multilingues : à partir de termes extraits dans une langue, un automate génère automatiquement des possibilités de traduction et va ensuite vérifier, dans un corpus en langue cible, les termes attestés qui permettent de retenir les bonnes traductions. Cette méthode est particulièrement efficace si l'on a 3 langues ou plus à construire : si A dans la première langue se traduit B dans une deuxième, qui se traduit C dans une troisième, qui lui-même se traduit A dans la première, le triplet de traductions peut être considéré comme fiable,
- pour établir l'apprentissage de systèmes de catégorisation : la différence avec un système « classique », qui fait un apprentissage sur des chaînes de caractères, est essentiellement que l'apprentissage se fait ici à partir de termes linguistiquement significatifs, et donc « sémantiquement » de meilleure qualité,
- pour l'aide à l'établissement d'ontologies : en couplant les technologies linguistiques avec des méthodes de « clustering », on fait apparaître des relations entre termes qui peuvent être utilisées par des experts lors de la construction d'ontologies. Ces méthodes sont également utilisées dans des problématiques de veille pour repérer des associations entre objets, événements, personnes, organisations, etc.

# White Paper Lingway

## 4.3 Utilisation du TAL pour « structurer le Web »

Les techniques de l'extraction d'information (« IE pour Information Extraction »), basées sur des technologies du TAL permettent de structurer une information textuelle au départ dépourvue de toute structure logique. Le champ de l'IE est souvent décomposé en plusieurs sous problèmes qui sont :

- l'extraction d'entités nommées,
- l'extraction de descripteurs thématiques,
- l'extraction d'attributs et de faits.

Des problématiques plus « pointues », comme la résolution de coréférences sont maintenant souvent rattachées à l'IE. Bien que ce dernier type de problème n'ait pas encore de solution générale, Lingway participe à des travaux de recherche sur ces sujets.

### 4.3.1 Extraction d'entités nommées

L'identification d'«entités nommées» a été identifiée comme une tâche à part entière lors des campagnes d'évaluation MUC (Machine Understanding Conference) il y a une dizaine d'années. On distingue généralement trois sous tâches : reconnaissance des noms d'entités (organisations, personnes, lieux), des expressions temporelles (dates et autres désignations temporelles) et des expressions numériques (grandeurs mesurables, quantités, pourcentages, etc.).

#### Réalisation dans Lingway KM

Lingway KM permet d'extraire automatiquement les entités suivantes :

- **Personnes** : essentiellement sur la base de règles de type « prénom + X » ou « X président de Y ». L'extracteur de nom de personnes identifie :
  - ✓ le nom patronymique
  - ✓ le prénom
  - ✓ une fonction quand elle est présente dans le texte
- **Lieux** : basé essentiellement sur des dictionnaires de noms de lieux, l'extracteur identifie :
  - ✓ le nom du lieu
  - ✓ son type (pays, ville, région, etc.)
- **Organisations** : basé sur des listes et des règles, l'extracteur identifie :
  - ✓ le nom de l'organisation
  - ✓ son type (société anonyme, association, etc.)
- **Dates et autres marqueurs temporels.**

# White Paper Lingway

## 4.3.2 Extraction de descripteurs thématiques

On distingue habituellement plusieurs types d'indexation thématique :

- indexation contrôlée : seuls des termes appartenant à une liste normalisée dite « d'autorité » pré-établie peuvent être retenus. Une ressource naturelle pour une telle liste peut évidemment être une ontologie décrite en format OWL. On utilise souvent aussi le vocable de « catégorisation » pour désigner ce type de traitement,
- indexation libre : les termes choisis comme descripteurs thématiques peuvent être librement choisis, sans contrainte par rapport à une liste d'autorité.

Les problèmes qui se posent quand on veut automatiser l'indexation d'un texte en descripteurs thématiques sont sensiblement différents selon qu'on utilise l'un ou l'autre de ces modes d'indexation. Mais dans les deux cas on a d'abord une phase d'identification des « thèmes », un éventuel traitement de regroupement ou de normalisation de ces termes, et enfin un calcul de pertinence indiquant l'importance du thème dans le document.

Les méthodes traditionnelles qui ne reposent pas sur des techniques linguistiques ont tendance à trouver « trop » de descripteurs, notamment parce que des descripteurs ayant le même sens, qui ne sont que des variantes l'un de l'autre, ne sont pas unifiés.

### Réalisation dans Lingway KM

Dans Lingway KM, l'extracteur de thèmes se base sur la reconnaissance de termes qui sont repérés par leur structure linguistique propre : par exemple en français un terme formé sur la structure « Nom + préposition + Nom + Adjectif (optionnel) » comme « augmentation des salaires » ou « augmentation du salaire minimum » est un bon candidat à être un descripteur thématique. Il sera effectivement retenu ou non pour indexer le document sur la base de critères statistiques complémentaires.

Si l'extracteur trouve deux descripteurs comme « action du gouvernement » et « action gouvernementale », ils seront automatiquement unifiés sous une des deux formes, ce qui constitue évidemment une indexation beaucoup plus efficace.

Cette unification peut être faite sur la base de critères morpho-syntaxiques, comme dans l'exemple ci-dessus (« gouvernemental » est un adjectif qui dérive du nom « gouvernement »), mais peut également être faite sur des critères sémantiques, à partir des données du dictionnaire (« action de Matignon ») pourrait ainsi être unifié avec « action du gouvernement ».

Cette normalisation peut se faire dans le cadre d'une indexation libre ou contrôlée.

# White Paper Lingway

## 4.3.3 Associer des attributs aux descripteurs

L'extraction d'entités permet d'identifier des personnes, des organisations, des lieux, etc. Mais on veut généralement aller plus loin et associer à ces entités des attributs extraits du texte : la fonction d'une personne (directeur, chef de projet, etc.), son rôle dans un document (auteur, personne citée, personne en charge d'un dossier, destinataire, etc.). Ou encore la nature d'une organisation (société, association, groupement, etc.), son chiffre d'affaires, etc.

### Réalisation dans Lingway KM

Dans Lingway KM, des règles contextuelles permettent d'extraire ces associations et de produire une description RDF appropriée.

Par exemple, pour une application dans le domaine du commerce électronique, Lingway permet d'analyser des textes d'offres commerciales et de construire automatiquement une table avec chaque attribut. Pour un vêtement, par exemple, on identifiera la taille, la marque, la matière, la couleur, etc. Pour une annonce d'emploi, on identifiera la société, le secteur d'activité, la région, le salaire, etc.

# White Paper Lingway

## 4.4 Utilisation du TAL pour faciliter la recherche

Les différentes problématiques et approches présentées jusqu'ici tendent toutes vers une plus grande structuration de l'information, visant ainsi à améliorer la recherche, la navigation, l'interopérabilité et l'usage informatique de cette information, aujourd'hui largement inexploitable par des programmes, disponible sur le Web.

Cependant, les techniques du TAL sont également souvent utilisées sur cette information textuelle non structurée, notamment pour faciliter la recherche, sans pour autant avoir à toucher aux modes d'indexation ou de structuration de l'information, ce qui est très souvent impossible parce que les fonds textuels sont déjà existants, déjà indexés par des méthodes traditionnelles et qu'il n'est pas question de modifier ces bases de données.

On utilise alors des méthodes d'expansion de questions (« query expansion ») qui vont construire une requête booléenne classique à partir d'une question en langage naturel – ou à partir d'un texte dont on aura automatiquement extrait les termes descripteurs importants, ce qui permet alors de faire de la recherche par similarité, et en la soumettant à un système de recherche classique.

### Réalisation dans Lingway KM

Lingway KM inclut un module de recherche en langage naturel construit sur ce principe. Il permet d'interroger des bases « full-text », directement sur le texte. Evidemment, cette recherche est également combinable avec une recherche sur des méta-données ou des plans de classements s'il en existe, combinant ainsi les avantages de diverses approches.

Lingway KM comprend son propre moteur full text, construit à partir de l'open source Lucène, mais peut donc également facilement se connecter à divers moteurs disponibles sur le marché. Il suffit pour cela de générer une requête dans la syntaxe particulière de chaque moteur.

De plus, compte tenu de son architecture, Lingway KM intègre « nativement » des possibilités de recherches multilingues, puisqu'il est facile de générer une requête booléenne dans une autre langue que celle de la requête initiale. Enfin, on notera que l'expansion est modifiable par l'utilisateur, ce qui lui donne un bon niveau de contrôle sur le système.

# White Paper Lingway

## 5 Conclusion

L'évolution vers le Web Sémantique paraît inéluctable. Cela va dans le sens général de l'informatique depuis son origine, qui a toujours été de tendre à structurer de plus en plus l'information. Le Web tel qu'on le connaît aujourd'hui sera profondément modifié par cette évolution, qui se fera pourtant lentement.

Le traitement automatique du langage jouera un rôle important dans cette évolution, le Web sémantique ne pouvant pas être seulement une transposition dans des technologies nouvelles des modèles classiques des bases de données traditionnelles. L'analyse du langage apporte des solutions à trois enjeux majeurs du Web que sont :

- la structuration de l'information,
- l'interopérabilité des ontologies,
- le multilinguisme, enjeu technique, culturel et politique majeur.

Lingway, qui est construit autour d'une équipe travaillant dans le domaine de l'ingénierie linguistique depuis plus de vingt ans, apporte une expertise unique, avec des outils logiciels, des données linguistiques, et – ce qui est peut-être le plus important – une méthodologie rigoureuse et éprouvée qui va permettre aux entreprises d'intégrer, sans heurts, cette évolution majeure.