

# White Paper Lingway

## Natural language Processing : The Key to the Semantic Web

### Table of Content

TABLE OF CONTENT .....	1
1 INTRODUCTION .....	2
2 THE SEMANTIC WEB .....	3
2.1 EVOLUTION OR ALTERNATIVE? .....	3
2.2 THE LANGUAGES OF SEMANTIC WEB .....	4
2.2.1 RDP .....	4
2.2.2 OWL .....	4
3. THE NATURAL LANGUAGE PROCESSING: NLP .....	5
3.1 THE LEVELS OF THE LINGUISTIC ANALYSIS .....	5
4. NLP AND WS: A VIRTUOUS CIRCLE .....	7
4.1 USE OF ONTOLOGIES BY THE NLP <small>Document actif</small> <a href="#">CTRL+clik pour suivre le lien</a> .....	8
4.1.1 Interoperability of ontologies .....	8
4.1.2 Use of ontologies in a NLP systems .....	8
4.2 USE OF THE WEB AS LEARNING SOURCE FOR THE NLP .....	9
4.3 USE OF THE NPL TO « STRUCTURE THE WEB » .....	11
4.3.1 Extracting named entities.....	11
4.3.2 Extracting thematic descriptors .....	12
4.3.3 Associating attributes to descriptors.....	13
4.4 USE OF THE NLP TO MAKE THE SEARCH EASIER .....	13
5. CONCLUSION .....	15

# 1 Introduction

The concept of « Semantic Web » gains more and more in importance and becomes one of the most promising ways of the Web evolution, such as we know it today. The most famous definition is the one given by Tim Berners-Lee in an article published in Scientific American of May 2001 and quoted on the W3C website (<http://www.w3.org/2001/sw/>).

« *"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."* »

Thus the authors seem to think that information do not have a well defined meaning in the actual Web. What is effectively true according to the computer engineer's point of view: most of the information available on the Web is a textual one, very less structured and therefore unusable for carrying out calculation or inferences processing. However it is obvious that this information has a meaning, but that only human readers can understand it today.

Now if the works and standards in progress on the Semantic Web specify how to structure, code, categorize the information using specific XML languages, such as RDF or OWL, they omit on the other hand to indicate the way of transposing the textual information into these formalisms.

If we can assume that a part of the new information which will be published in the next years on the Web will be effectively "indexed" from its production in the Semantic Web formalisms - at least, for certain types of more or less factual information which will be easily coded - nevertheless most of the information will keep its textual form and the indexation issue in Semantic Web languages will become crucial.

It is exactly the objective of the Natural Language Processing (NLP) which aims at understanding and modelling the human language to "*makes the text computable*".

**LINGWAY** [www.lingway.com](http://www.lingway.com) is focused on the development of software dedicated to information processing based on the NLP. This document presents its approach of this issue which aims at automatically index textual documents by extraction of named entities, thematic descriptors or categories related to a specific taxonomy. To put this presentation in concrete form, this approach will be related with the functionalities of Lingway KM, the software offer of Lingway.

We will also point out the role of the tools for multilingual query and reading assistance such as summarization and visualization.

## 2 The semantic Web

### 2.1 Evolution or alternative?

One question which arises regarding the semantic Web is the following: would the current Web be developed and changed by the integration of more and more structure, or would “another Web” appear parallel to the current one which could carry on to live a very long time with its own standards?

In some ways, the « purists » of the semantic Web re-create with a new technology what has been done since always with relational DBMS: definition of a data model, management rules, etc. According to this point of view, to write a RDF definition is finally rather similar to describe a table in SQL.

This approach is obviously interesting to process relatively restricted, finished and entirely model-able fields. It is typically the scope of “e-commerce” applications which will certainly be the first users of this type of evolution (and which are the initial motivation besides). However it cannot be applied without adaptation to the modelling of the vast quantity of non-structured information available in the texts of the current Web, which are a source of great importance for other applications such as the information Watch or the “Knowledge Management”.

It thus seems legitimate to think that besides a structured semantic Web based on XML-RDF suitable for factual and model-able information will still carry on, and for a long time, a Web as the one which exists today meaning non-structured, widely textual, based on HTML pages. Obviously, both of them will interpenetrate each other and all kinds of hybrid situations will appear.

We can compare this evolution with the one of databases for 25 years: the globalization of SQL databases did not take the place of all textual databases, managed with “documentary systems”. However both databases interpenetrated each other and the actual relational DBMS offer functions for search and textual storage whereas documentary systems, firstly based on textual search functions, gradually added functions of processing of related factual data.

The interesting point according to the point of view of the use of technologies related to language processing, is double:

- On the one hand, how will be processed the existing mass of textual information to be indexed, even in a rudimentary way, within the semantic Web formalisms and to make it suitable for new information which will be created in new formats?
- On the other hand, how will the linguistic technologies take part into the structuring of the new information which will be published on the Web in the next years?

## **2.2 The languages of semantic Web**

Several publications are available on this theme. Only very simplified definitions are reminded here.

### **2.2.1 RDF**

The RDF (Resource Description Framework) definition, such as published by W3C, specifies that RDF is a language used to present the information on Web resources, especially suitable for presenting metadata related to documents available on the Web, such as title, author, modification date, copyright, etc. In a wider sense, the same language makes it possible to present the information about objects identified on the Web, even if they are not directly available there.

A RDF term includes a « subject », a « predicate » (or « verb ») and an « object ». This type of model is far from being new. It allows representing complex facts by analyzing them in elementary facts. The innovation here is that subject, object and predicate are described by URI (Uniform Resource Identifiers) which makes it possible to give precise meanings, available on the Web, to concepts and predicates.

### **2.2.2 OWL**

The RDF problem is that it does not imply any standardization on the semantic itself, and that various users could give different names to entities having the same meaning, which prevents from carrying out any aggregation between data coming from different Web sources.

The solution is provided by the development of ontologies, which expressly describe the terms and the relations between them.

OWL (Web Ontology Language) is a language allowing to publish and to share ontologies on the Web. It is a RDF extension. OWL makes it possible to give equivalences, from a world to another (meaning from an ontology to another) between objects descriptions. For instance, we could describe that the object "ZipCode" in the application 1 is the counterpart of the object "CodePostal" in the application 2.

## 3. The Natural Language Processing: NLP

Before pointing out in what the NLP can help with the realization of the semantic Web, we will make a short reminder on these technologies.

### 3.1 *The levels of the linguistic analysis*

Progress could be made in the field of NLP (Natural Language Processing) from the day onwards the problem, too wide to be approached overall, could have been broken down in several sub-problems corresponding with sub-tasks articulated between them. This analysis is carried out at the same time on the level of linguistic resources (dictionaries and grammatical rules) and on the level of algorithms (analyzers, transducers, etc.).

The main traditional levels are:

- the morphological level: to identify the words of a sentence
- the syntactic level: to identify the constituents and the functions of a sentence
- the semantic level: to identify the meaning of the words and the logical structure of a sentence
- the text level: to identify the structure of a text and the relations between sentences
- the corpus level: to identify the structure of a set of texts

Each task (level) can also be broken down in sub-tasks.

Thus the morphological level can be re-broken down in:

- the tokenisation: to identify the frontiers of words (simple and compound) and of sentences.
- the tagging: to identify the category – noun, verb, adjective – of each word.
- the lemmatisation: to identify the canonical form in the dictionary.

The syntactic level is often broken down in:

- the chunking: to identify the frontiers of great importance of constituents (group of nouns, verbs, etc.) and/or relations of great importance between the words.
- the functional tagging: to assign grammatical functions to constituents.
- the parsing: to build a tree representing the structure of the complete sentence.

The semantic level is also broken down in:

- the selection of meaning (WSD « word sense disambiguation »): to select the meaning of each word.
- the logical structuring: to identify arguments of each predicate and their semantic role (agent, purpose, place, etc.).

The text level can also be broken down in:

- resolving anaphors (antecedents of the pronouns, ellipses, references).
- defining the rhetoric structure (comments, explanations, causalities, etc.).
- defining the thematic structure (what about is the text?)

Finally the corpus level can be broken down in:

- defining the nature of documents (press article, technical article, lawful text, commercial booklet, etc.)
- the thematic structure of the corpus (what about is the corpus?).

Most of these tasks are related to sub-disciplines of linguistics. According to the NLP engineering point of view, they correspond with modules, with specific programs, each one of them needing a type of linguistic resources and a specific algorithmic type.

### *Achievement in Lingway KM*

All these modules are implemented and used in Lingway KM.

In particular, Lingway KM is an integrated product offering:

- **A semantic search engine** capable of translating a natural language query to Boolean languages and/or to a taxonomy (classification plan) in several languages.
- **An indexing and categorization engine** based on a XML structuring module allowing the mark-up of the text on a semantic level.
- **A reading assistance tool engine** (automatic summarization, text colorization and visualization) making it possible to tackle the content of a document or a set of documents.

Lingway KM is based on a linguistic model broken down in 3 levels:

- On the morphological level, the information related to the syntactic category of the word and its flexional features are coded: gender, number, flexion mode.
- On the semantic level, the possible meanings of the words are specified by semantic attributes (belonging to a semantic class) and semantic relations (for instance, derived relations between a verb and a noun).
- On conceptual level, the various meanings are related to a concept, independently of the language.

All these data make out a « dictionary » - in reality a real knowledge database – of about 150,000 concepts combined with five languages: English, French, German, Spanish, and Dutch.

## 4. NLP and WS: a virtuous circle

The NLP and semantic Web technologies tally one with the other in several ways. They will develop simultaneously while being reinforced mutually in a virtuous circle.

On the one hand, the Semantic Web will, progressively with its development, become itself a resource for the development of NLP technologies:

- The large number of ontologies already published, and which will inevitably be multiplied, represent an invaluable knowledge source for the tools based on NLP. However, these resources practically never specify the linguistic terms described by ontologies, which remain “simple” character strings. To make them usable by NLP systems, it is necessary to create a link between the ontologies, such as available on the Web and a language dictionary suitable for NLP.
- The more the Web is structured, the more it becomes a knowledge source for NLP. This is especially true regarding categorization applications, which are often based on learning technologies and which will take advantage of the indexation by easily useable metadata. The great numbers of naturally aligned multilingual resources which are made out by the versions in various languages of great numbers of websites, especially trade ones, also provide an important resource for assisting the compilation of dictionaries and more widely multilingual and “cross-language” applications.

On the other hand, NLP technologies will contribute towards the semantic Web setting up which will be gradually structured while benefiting from their regular and undoubtedly increasingly fast improvement:

- To contribute towards the progressive Web structuring, on the one hand by adding metadata to pages already published, in a “retro-conversion” logic, and on the other hand by providing structuring assistance tools for the future pages.
- To contribute towards the ontologies and multilingual resources compilation, by using information extracting and structuring techniques from the texts, and by combining these new data with partial ontologies already agreed by human experts.
- To contribute towards textual information search, by adding linguistic and semantic methods in the search engines, such as they exist today, which will appreciably improve the searches quality, especially by managing the multilingual access.

The issue related to each of the above-mentioned points will be presented, and then completed by an outline of the solutions provided by Lingway.

## 4.1 Use of ontologies by the NLP systems

A lot of various ontologies will be – and already are – published on the Web. OWL being very recent, little are still in this format, but precursor formats were already used and by now a great numbers of resources are available. Some authors think that millions of elements coded under the form of RDF triplets will be created and attainable, which raises immediately two new problems: first how to ensure a certain coherence and interoperability between these ontologies and then how to integrate them in the NLP systems?

### 4.1.1 Interoperability of ontologies

One of the OWL important primitives is « same-as » which makes it possible to specify that two URI indicate in fact the same “object”. This mechanism is central in the semantic Web, since the same object could be named differently in various ontologies.

For instance, if an ontology on a given website includes an input worded “*Mr Jacques Chirac*” and another input worded “*The president Chirac*”, they could not be unified except clearly binding them together by a “same-as” link. Assuming that efforts of millions of Web surfers – private individuals or organizations – will gradually converge to create such links, it is still likely to be an extremely long process. And even longer if the identity of both information elements is more complex to identify than the one of the previous example: for instance, let us imagine the term “*majority poll*” described in a specific ontology and “*election in the majority*” in another. Both of these descriptions could not be shared as long as a “same-as” link will not be created between them.

### Achievement in Lingway KM

The Lingway tools provide methods of ontologies comparison and linking. It makes it possible on the basis of product’s semantic dictionaries, which know the link between “poll” and “election” to automatically bind both of the terms, and thus to create a link between both ontologies.

Moreover, Lingway is capable of establish the same correspondences between ontologies expressed in different languages.

### 4.1.2 Use of ontologies in a NLP systems

The systems based on NLP, especially search and « text-mining » engines may find a benefit of using the information modelised within these ontologies. Nevertheless, they are usually not directly usable because they do not include any linguistic information of the lexical level.

For instance, if an ontology belonging to IT field specifies that « *mouse* » is related to « *peripheral equipment* », this information becomes ambiguous in a linguistic system, the word « *mouse* » having several meanings, and that is likely to generate interpretations indicating a link with “*rat*” (related to the “animal mouse”).

### **Achievement in Lingway KM**

The Lingway tools make it possible this correspondence between an external ontology and the product internal linguistic ontology. To carry it out, the Lingway ontology provides each word with several meanings or “concepts”, each one of them being specified by semantic domains and classes.

The concept of semantic domains covers the concept of area in which the word is used with a specific meaning. For instance, the concept @MOUSE: EQUIPMENT belongs to IT domain. There are about twenty domains with hierarchical links to more than 400 sub-domains on a hierarchy of 6 levels.

In addition, the words meanings may belong to one or several classes, which purpose is the description of the object intrinsic properties. For instance, *mouse* meaning “peripheral equipment” belongs to the class TOOL whereas mouse meaning animal is included in the class MAMMAL. Nearly 500 classes are pre-defined in our dictionary model.

Various analysis processes, including heuristics of semantic disambiguation enables to analyse the terms of an ontology and to integrate them in Lingway KM, into various functions of search engine, extraction-categorization or reading assistance tool.

### **4.2 Use of the Web as learning source for the NLP**

At the opposite, the information mass still available on the Web becomes a source for the NLP systems learning.

The methods based on learning are focused on *annotated* corpus and aim at setting up automatically the systems capable of re-producing similar annotations on non-annotated texts.

The annotations can be more or less precise: on the roughest level, the classification of a document in a specific category of taxonomy is the first level of annotation. This level is often the one used by the systems known as “categorization”. At the other end of the spectrum are the corpus annotated on a very fine level – even on the level of each word – usually carried out by university teams who compile invaluable resources for the training.

For instance, the following elements can be suitable for the training:

- Websites in several languages: a great numbers of websites are translated in various languages. The e-trade websites are usually written in their national language and in English. It provides a great source of *aligned multilingual* corpus which can be used to compile multilingual terminologies. This natural alignment can be considered as an annotation form usable by a learning system.
- Categorized websites: great numbers of websites already classified their content within their own classification plans and taxonomies which are sometimes coming from the famous international classifications (for instance, MeSH in medicine).

### **Achievement in Lingway KM**

Usually the learning techniques from great-sized corpus are largely used by Lingway KM.

- To compile monolingual terminologies by using the terms extractors of Lingway KM.
- To compile multilingual dictionaries: from terms extracted in a language, an automation automatically generates the translation possibilities and then checks, in a target language corpus, the agreed terms which provide the right translations. This method is particularly efficient if 3 languages and more must be compiled. If A in the first language is the translation of B in a second language, which is translated by C in a third one being the right translation of A in the first language, the translations triplet can be regarded as reliable.
- To learn the categorization systems: the difference with a “traditional” system, which is focused on learning from character strings, lies in this that the learning starts from linguistically significant terms, and thus “semantically” of better quality.
- To assist the ontologies working out: by combining the linguistic technologies with the « clustering » methods, relations between terms appear and can be used by experts for the ontologies working out. These methods are also applied for information watch issues to localize associations between objects, events, persons, organizations, etc.

### **4.3 Use of the NPL to « structure the Web »**

The techniques of information extraction (“IE” for Information Extraction), based on NLP technologies makes it possible to structure a textual information which is at the beginning devoid of any logical structure. The IE field is often broken down in several sub-problems:

- Extracting named entities,
- Extracting thematic descriptors,
- Extracting attributes and facts.

More specific issues, such as the resolution of co references are often related to the IE. Even if this type of problem does not have a global solution yet, Lingway takes part in research projects on these subjects.

#### **4.3.1 Extracting named entities**

The identification of « named entities » has been identified as a separated task during the MUC (Machine Understanding Conference) evaluation campaign conducted ten years ago. Three sub-tasks are usually differentiated: recognition of named entities (organizations, persons, and places), temporal expressions (dates and other temporal markers) and numerical expressions (measurable importance, quantities, percentages, etc.).

#### ***Achievement in Lingway KM***

Lingway makes it possible to automatically extract the following entities:

- Persons: essentially on the basis of rules of type “first name + X” or “X chairman of Y”. The extractor of persons’ name identifies:
  - ✓ the surname
  - ✓ the first name,
  - ✓ the position if available in the text
  - ✓ the title or position in a company
- Places: essentially based on files of places name. The extractor identifies:
  - ✓ the name of the place
  - ✓ its type (country, town, state, etc.)
- Organizations: based on lists and rules. The extractor identifies:
  - ✓ the name of the organization
  - ✓ its type (public limited company, association, etc.)
- Dates and other temporal markers.

### 4.3.2 Extracting thematic descriptors

We usually distinguish several kinds of thematic indexation:

- Controlled indexation: only terms belonging to a pre-established standardized list known as “authority list” can be retained. A natural resource for such list can obviously be an ontology specified in OWL format. The word “categorization” is also often used to indicate this kind of processing.
- Free indexation: the terms selected as thematic descriptors can be freely selected independently of any authority list.

The problems related to automatic indexation of a text with thematic descriptors are quite different according to the use of one or the other of these indexation methods. However in both of them is first carried out a phase of identification of the “themes”, then a possible processing for regrouping or normalizing these terms and finally a relevance calculation specifying the significance of the theme in the document.

The traditional methods which do not rely on linguistic techniques tend to find “too many” descriptors, especially because descriptors having the same meaning, which are derived words, are not unified.

#### *Achievement in Lingway KM*

In Lingway, the extractor of themes is based on the recognition of terms which are localized by their own linguistic structure: for instance, a term formed in English on the structure “Noun-preposition-Adjective (optional)-Noun” such as “*raising of wages*” or “*raising of minimum wages*” is a good candidate for being a thematic descriptor. It will be actually retained or not to index the document on the basis of additional statistical criteria.

If the extractor finds two descriptors such as « *action of the government* » and “*governmental action*”, they will be automatically unified under one of the two forms, which contributes obviously to a much more effective indexation.

This unification can be carried out on the basis of morpho-syntactic criteria, such as in the above-mentioned example (« *governmental* » is an adjective derived from the noun “*government*”), but can also be executed on semantic criteria from data of dictionary (“*action of the White House*” could be unified with “*action of government*”).

This normalization can be carried out in the frame of a free or controlled indexation.

### 4.3.3 Associating attributes to descriptors

The extraction of entities makes it possible to identify persons, organizations, places, etc. However we usually want to go further and to associate attributes extracted from the text to these entities: the *position* of a person (director, project manager, etc.), its *role* in a document (author, quoted person, person responsible for a file, addressee, etc.) or the *nature* of an organization (company, association, trust, etc.) its *turnover*, etc.

#### *Achievement in Lingway KM*

In Lingway, contextual rules allow to extract these associations and to generate an appropriate RDF description.

For instance, regarding an application in the e-trade field, Lingway makes it possible to analyse within the texts of commercial offers and to automatically generate a table with each attribute. Concerning a cloth, for instance, the size, the material, the colour, etc. will be identified. Regarding a job offer, the company, the activity sector, the state, the wages, etc will be identified.

### 4.4 Use of the NLP to make the search easier

The different issues and approaches presented up to now all tend to a more important structuring of information, thus aiming at improving the search, the navigation, the interoperability and the IT use of this information available on the Web, which is at present widely un-usable by programs.

However the NLP techniques are also often applied to this non-structured textual information, especially to make easier the search without modifying the modes of indexation or information structuring, which is very often impossible because the textual backgrounds are already generated and already indexed by traditional methods and because it is out of question to change these databases.

Therefore other methods of query expansion are used which will generate a traditional Boolean request from a query in natural language – or from a text from which the relevant descriptors terms would have been automatically extracted, which enables to carry out a *search by similarity* – and by submitting it to a traditional search system.

#### *Achievement in Lingway KM*

Lingway KM includes a search module in natural language based on this principle. It makes it possible to query “full-text” databases, directly on the text. Obviously, this search can also be combined with a search on metadata or existing taxonomies, thus combining the advantages of the various approaches.

Lingway KM includes its own full text search engine, built with the *Lucène* open source, and therefore can easily be connected to various search engines available on the market. The only condition is to generate the query in the syntax specific to each search engine.

Moreover, because of its architecture, Lingway KM integrates « natively » possibilities of multilingual search since it is easy to generate a Boolean query in another language than the one of the initial request. Finally, the expansion can be modified by the user, providing him with a good monitoring level on the system.

## 5. Conclusion

The evolution towards the Semantic Web appears inescapable. It goes in the global direction of the IT field from its origin, which always tends to structure the information more and more. The Web, such as we know it today, will be deeply modified by this evolution which will however be done slowly.

The natural language processing will play a significant part in this evolution, the semantic Web not being able to only be a transposition in new technologies of the traditional models of traditional databases. The language analysis provides solutions to three major stakes of the Web, which are:

- The information structuring,
- The interoperability of ontologies,
- The multilinguism, major technical, cultural and politic stake.

Lingway which has been built by a team working in the linguistic engineering field for more than twenty years, provides **an unique expertise**, with software tools, linguistic data, and – what is perhaps the most important – a rigorous and tested methodology which will make it possible the companies to integrate , without clashes, this major evolution.