



# DEMO-net

## D14.3 The role of Natural Language Processing in eParticipation

DEMO-net Consortium

## Editor details

Author Name	Organisation	Email
Claudia Soria	CNR-ILC	claudia.soria@ilc.cnr.it
Francesca Bertagna	CNR-ILC	francesca.bertagna@ilc.cnr.it

## Author details

Author Name	Organisation	Email
Claudia Soria	CNR-ILC	claudia.soria@ilc.cnr.it
Francesca Bertagna	CNR-ILC	francesca.bertagna@ilc.cnr.it

We thank Colin Fraser for his input to the first draft of the Booklet



IST Network of Excellence Project  
FP6-2004-IST-4-027219  
Thematic Priority 2: Information Society Technologies

**DEMO-net**  
**The Democracy Network**

## **DEMO-net : D14.3 The role of Natural Language Processing in eParticipation**

**Editor:** Istituto di Linguistica  
Computazionale

**Revision:**

**Dissemination Level:** PU

**Author(s):** Claudia Soria, Francesca Bertagna

**Due date of deliverable:** 31.12.2007

**Actual submission date:**


**Start date of project:** 01 January 2006

**Duration:** 4 years

**WP no.:** 14

**Organisation name of lead contractor for this deliverable:** CNR-ILC

**Abstract:** Aim of this document is introducing Natural Language Processing (NLP), a wide range of technologies that take the analysis of human language as their focus. The document is addressed to stakeholders of eParticipation and developers of eParticipation systems, to inform them about the availability of instruments and tools for the analysis of language that can be incorporated in more complex systems to capture information that is hidden in documents and speech. The booklet is organized in three main sections: the first one introduce what we call the "basic technologies", i.e. the different tools that are conceived to perform the analysis of specific linguistic phenomena at different level of linguistic description and complexity. The second section is dedicated to the analysis of the "final applications", i.e. the families of systems which can exploit NLP "basic technologies" to perform more complex and advanced task. In the last section we try to sketch some conclusions, in particular by analysing the needs and requirements of eParticipation more close to NLP.



Project funded by the European Community under the FP6 IST Programme  
© Copyright by the DEMO-net Consortium

## History

Version	Date	Modification reason	Modified by
1	1-1-2007	First overview of technologies and applications	Claudia Soria and Colin Fraser
2	1-9-2007	Added content on description of technologies and application. Added sections on eParticipation exploitation	Francesca Bertagna
3	26-10-2007	Revision and added analysis of D5.3 questionnaire	Claudia Soria and Francesca Bertagna
4	26-11-2007	Revision	Claudia Soria and Francesca Bertagna

# Table of Contents

<b><i>HISTORY</i></b>	<b>5</b>
<b><i>TABLE OF CONTENTS</i></b>	<b>6</b>
<b><i>EXECUTIVE SUMMARY</i></b>	<b>8</b>
<b>1 INTRODUCTION</b>	<b>10</b>
<b>2 NLP BASIC COMPONENTS</b>	<b>11</b>
2.1 TECHNOLOGIES FOR TEXTUAL ANALYSIS	12
2.1.1 Part Of Speech Tagging	12
2.1.2 Lemmatization	14
2.1.3 Stemming	14
2.1.4 Syntactic Parsing	15
2.1.5 Word Clustering	16
2.1.6 Multiword Recognition and Extraction	18
2.1.7 Named Entity Recognition	18
2.1.8 Word Sense Disambiguation	19
2.1.9 Coreference and Anaphora Resolution	20
2.2 SPEECH TECHNOLOGIES	21
2.2.1 Speech Recognition	21
2.2.2 Speech Synthesis	21
2.3 LANGUAGE RESOURCES	22
2.3.1 Written Language Corpora	23
2.3.2 Spoken Language Corpora	24
2.3.3 Computational Lexicons	25
<b>3 OUTLINE OF MAIN APPLICATION AREAS</b>	<b>28</b>
3.1 MACHINE TRANSLATION	28
3.1.1 Current exploitation in eParticipation	31
3.2 INFORMATION RETRIEVAL	34
3.2.1 Basic Architecture of a IR system	36
3.2.2 NLP and IR	38
3.2.3 Cross-Language Information Retrieval	40
3.2.4 Document and Text Classification	41
3.2.5 Summarization	41
3.2.6 Question Answering	42
3.2.7 Information Extraction	45
3.2.8 Fact Extraction	46
3.2.9 Extraction of Temporal Information	47
3.3 CURRENT IR EXPLOITATION IN ePARTICIPATION	48
3.4 TEXT MINING	50
3.4.1 Opinion Mining	52
3.4.2 Ontology Acquisition	53
3.4.3 Text Mining Current Exploitation in eParticipation	54
3.5 LANGUAGE RESOURCES AND TOOLS INFRASTRUCTURE	55



**4 CONCLUSION: PARTICIPATION AREAS AND TECHNOLOGICAL CHALLENGES 57**

4.1 NARROWING THE LANGUAGE GAP 57

4.2 MULTILINGUALISM 58

4.3 "GOING STRAIGHT TO THE POINT" 59

4.4 RETRIEVING TRENDS, OPINIONS AND SENTIMENTS 59

4.5 NLP TECHNOLOGIES IN THE DEMO-NET D5.3 QUESTIONNAIRE 59

**REFERENCES 61**

## Executive Summary

Aim of this document is introducing Natural Language Processing (NLP), a wide range of technologies that take the analysis of human language as their focus. NLP is a highly interdisciplinary field, close to Artificial Intelligence, Cognitive Sciences, Philosophy of Language, Linguistics, Computer Science, Statistics, and Engineering. It is not a new area of study, yet it is very lively and constantly evolving. NLP encompasses a very wide range of applications; it is aimed to analyse the many levels of linguistic description, from phonology to pragmatics, and to deal with spoken and written language.


Giving a full description of such a complex, rich and varied field of study is well beyond the scope of this document. Instead, what we would like to do is introducing the potentialities of the tools for the analysis of language in the emerging area of eParticipation.

Language plays a fundamental role in eParticipation, since it is the medium through which all the communication takes place: it is the language we find in institutional sites to explain to citizens how to obtain a particular service, it is the language of political discourse, the language of people expressing political opinions on a non official forum.

NLP can be an instrument to deal with all these types of messages in an automatic or semiautomatic way. This document is addressed, in this sense, not to the NLP scientific community, which would find in it neither an in-depth analysis of tools and data, nor a survey of the most cutting edge solutions. The document is instead addressed to stakeholders of eParticipation and developers of eParticipation systems, to inform them about the availability of technologies for the analysis of language that can be incorporated in more complex systems to capture information that is hidden in documents and speech. We conceived the document as a booklet and we tried to keep its size, for what was possible given the complexity and the maturity of the field, "under control". We hope that the internal organization of the booklet, with its distinction in application and tool typologies, may help the reader to easily identify the area(s) more close to his/her interests and needs.

We also would like to highlight the fact that many of the presented solutions are "sought-after" by more of one field of study and expertise: this is true, for example, for the Text Mining technologies, which is, with good reason, also a branch of Knowledge Mining. In this sense, the contribution we want to provide with this document goes in the direction to enrich the continuum which keeps together and links the many families of technologies surveyed and presented in Demo-Net WP5 and WP14.

The booklet is organized in three main sections: the first one introduce what we call the "basic technologies", i.e. the different tools that are conceived to perform the analysis of specific linguistic phenomena at different level of linguistic description and complexity: tools for Part of Speech Tagging, for lemmatization, for detection of poly-lexical units or Named Entity units, for syntactic analysis etc. At the same time, we



introduce the data, in the form of lexicons or corpora, which can be used for sustaining the analysis of natural language.

The second section is dedicated to the analysis of the “final applications”, i.e. the families of systems which can exploit NLP “basic technologies” to perform more complex and advanced task: Machine Translation, Information Retrieval and Text Mining. For these applications we provide a description of their aims and of their generic architecture. At the same time, when possible, we indicate, as examples, some well-know existing systems. For each of them, we analyse the potentiality for eParticipation and, when possible, their actual use in that applicative scenario.

In the last section we try to sketch some conclusions, in particular by analysing the needs and requirements of eParticipation more close to NLP. This should provide a “flavour” of the potentialities, yet not fully exploited, of NLP technologies. In the conclusions, we also provide a brief analysis of the result of the questionnaire prepared and filled during Demo-Net task5.3. The analysis seems to indicate that NLP is still little used in eParticipation.

# 1 Introduction

Natural Language Processing is the name given to a wide range of technologies which take the analysis of human language as their focus. Examples of the successful application of these technologies are web search engines, automatic message processing, web based machine translation services, voice recognition on mobile phones and on dialogue systems for railways, cinemas and banks, amongst others, speech synthesis and dialogue systems which are used over the telephone. Although these technologies are not yet perfected, they are good enough to be used in a great many applications by many millions of citizens every day. The research challenges are great; but then so are the rewards – what is at stake are more natural, language based interfaces to the wide range of information that is available to us in current society.

Our strategy in this report is to give the reader a flavour of the basic technologies which are used in this area and then go into an in-depth account of how these are utilised in the major research and application sub-areas which constitute modern NLP research.

We briefly describe:

- tools for the analysis of language at phonological, morphological, syntactic, semantic and also discourse level;
- resources or repositories of linguistic information that are needed by tools and instruments for language analysis, basically lexicons and corpora.

These tools and data repositories are called “basic” technologies since they should be seen as modules that can be combined and integrated in final applications able to perform more complex tasks. As a matter of fact, NLP technologies are often combined in multimodal dialogue and interaction systems, and to a great extent NLP is often considered to be merely a component which processes language either in order to extract information of some sort or to present this within larger applications. A second chapter introduces the more advanced technologies which constitute the applicative environment where eParticipation tools and systems can be developed and tested. Last chapter is dedicated to some concluding remarks, by introducing the challenges of eParticipation and discussing how NLP can meet the eParticipation areas described in Deliverable 5.1.

This deliverable is complementary to the work presented in Demonet deliverable 5.2, in particular for what Knowledge Management and Ontologies are concerned. Many are the links among these two sectors and NLP and a kind of continuum encompassed the three fields.

## 2 NLP basic components

Before introducing the various technologies, we would like to briefly list the major problems which represent the challenges for the discipline. The key notion is "ambiguity".

In lexical ambiguity, analysis tools have to handle the fact that many words have more than one meaning (a "bank" can be a financial institution or a part of a river.); thus, the meaning that makes the most sense in context has to be selected. The task of find the right sense within a context is called Word Sense Disambiguation.

Syntactic ambiguity is due to the inherent ambiguity of the grammar for natural languages: there are often multiple possible syntactic analysis for a given sentence. ("The man saw the boy with the telescope": is the man or the boy who has a telescope?) Choosing the most appropriate one usually requires semantic and contextual information.

Understanding the context is even more necessary to solve semantic ambiguities. For instance, incorrect anaphora resolution can lead to misunderstanding (in "every farmer who owns a donkey beats it", is the pronoun "it" a reference to "farmer" or "donkey"?). Identifying the logical subject is also sometimes difficult (in "John asks his mother to do that", is John or his mother that is supposed to do the action?).

In what follows, we present an overview of the technologies which are used, in final applications, to try to resolve these different levels of ambiguity when analysing language at various levels of linguistic description and complexity.

NLP technologies are classified on the basis of the fundamental distinction between tools and instruments for the analysis and generation of written language and technologies dedicated to the understanding and generation of natural language speech. It will be possible to recognize that the technologies presented concern different levels of linguistic descriptions, with an increasing level of complexity: from the morpho-syntactic and syntactic level of lemmatization, POS tagging and parsing to the semantics of WSD and Semantic analysis to the discourse and pragmatic level required by Anaphora resolution.

The set of technologies taken into account is not meant to be exhaustive; instead, it represents a selection of the most salient and well-established methods adopted in NLP.

It is worth remembering here that NLP has gone, in recent years, through an important modification: the discipline and its main sub-areas have witnessed the strong prevalence of statistical, example-based research over traditional symbolic, rule-based, deterministic approaches. There are many explanations for this prevalence:

- lack of high precision scalable parsers and sufficient lexical resources;
- necessity to maintain large lexicons and resources and adapt them to different domains;
- complexity of required software development etc.

Statistical systems, on the other hand, are highly flexible, language-independent and can be re-targeted to a new subject area or a new language in a matter of weeks. In what follows, many basic technologies and final applications are introduced and described: for all of them, we have to keep in mind that different approaches are followed, usually based on this fundamental, almost paradigmatic distinctions. However, the time is mature to recognize that statistical methods alone are not always able to produce high precision results and that a "third way" would be welcomed, based on the combination of both approaches.

## 2.1 Technologies for Textual Analysis

Written language has an enormous importance in human communication. Even though speech is somehow more natural to humans than writing, understanding written signs, at various level of complexity, is surely a fundamental task, since it enables the access to an incredible amount of information that is stored day by day in many different forms. Web pages, books, electronic documents, newspapers, hand-written personal communications: all these sources of knowledge are nowadays available for being analysed, decomposed and understood.

One of the aims of computational linguistics is to analyse textual material by finding how larger textual units of meaning arise out of the combination or smaller ones. This is why the work of a computational linguist often seems composed by incremental steps, which allow the understanding of information of more and more complex nature. Tools and technologies should be considered, in this sense, like pyramid levels that in theory may in the end give access to the full understanding of language.

### 2.1.1 Part Of Speech Tagging

A Part of Speech (POS) may be considered to be a word class category such as a noun, a verb, a preposition or a determiner, and the process of POS Tagging is to assign a Part of Speech to all words (and usually to punctuation) in a corpus. The input to a POS tagger is generally a natural language text and set of POS tags.

There are a number of a problems associated with giving a tag to a particular word. Consider the text [Jurafsky & Martin, 2000]:

*Book that flight*

VB DT NN

(VB: verb, DT: determiner, NN: Noun)

The word book is ambiguous – it could easily also be a noun if we didn't know the context. That is similar – it can operate as a relative pronoun (eg. I think that Donald is smaller than Chris) or a determiner. The key to a successful POS tagger is to resolve these ambiguities in a process of disambiguation, an incredibly important process in NLP.

For English, there are commonly 8 parts of speech distinguished: noun, verb, adjective, preposition, pronoun, adverb, conjunction, and interjection. However, there are clearly many more categories and sub-categories. In POS tagging by computer, it is typical to distinguish from 50 to 150 separate parts of speech for English (see the POS tags used in the Brown Corpus [see e.g. <http://icame.uib.no/brown/bcm.html>]), and up to 1.000 parts of speech for other languages.

The task of POS tagging corresponds to the morphological and syntactical analysis of a text. Both these analyses have to deal with ambiguity of natural language on several levels. In general, a set of hypotheses of candidate POS tags is generated for each word (token) on the morphological level, and the ambiguity is solved on the syntactical level, taking into account the context and syntagmatic relations of particular words in analysed text.

However, solving the ambiguities on the level of syntax by rule-based NLP approaches without understanding the semantics or even the pragmatics of the context is extremely expensive, especially because analyzing the higher levels is much harder when multiple part-of-speech possibilities must be considered for each word. To automatize the POS tagging and make it useful for large corpora, a set of statistical approaches and methods was developed since mid 1980s [Charniak, 97].

A popular statistical (stochastic) methods for POS tagging is Hidden Markov models (HMMs). HMMs involve counting cases (such as from the Brown Corpus), and making a table of the probabilities of certain syntagmatic sequences, by presupposition of next N words according to the analysis of M previous words. More advanced ("higher order") HMMs learn the probabilities not only of pairs, but triples or even larger sequences. The stochastic taggers of second type use decision trees or maximum entropy models to combine probabilistic features.

Nowadays there is a large suite of POS taggers, mostly implemented on statistical principles. Among others, we mention:

Brill's Tagger (<http://www.cs.jhu.edu/~brill/code.html>), the Stanford POS tagger (based on Penn Treebank tag set [Toutanova et al., 2003], <http://nlp.stanford.edu/software/tagger.shtml>), TreeTagger ([Schmid,

1994], <http://www.lsi.upc.edu/~nlp/SVMTool/>) (decision tree based tagger, language independent), SVMTool (based on SVMs), Maximum Entropy part of speech tagger with sentence boundary detector included, ACOPOST (implements maximum entropy, HMM trigram, and transformation-based learning), fnTBL (implements Transformation-Based Learning), and many others.

### 2.1.2 Lemmatization

Lemmatization is a process where the inflectional and variant forms of a word are reduced to their lemma: their base form, or dictionary look-up form. When lemmatizing a text, each individual word in that text is replaced with its lemma. For example, in English, the verb "to walk" may appear as "walk", "walked", "walks" and "walking". The base form, "walk", is called the lexeme for the word. The combination of the base form with the part of speech is often called the lemma of the word.

The process of lemmatization requires usage of various morphological analysis techniques, which are significantly dependent on particular language. For non-flective languages (e.g. English) analysis on morphological level is quite easy and can be done by simple rule-based mechanism, covering transformations of regular and irregular verb forms to infinitive, transformations of noun forms to nominative singular, and transformations of superlative adjectival forms to its base form.

In flective languages (e.g. Slavic, German languages, etc.), the morphological analysis is much more tricky and complex. Corpora-based statistical methods, combined with manual annotation of training sets, are used for identification of complex morphologic, syntactic, lexical and semantic properties of words that serve as a criterion for their classifying into parts of speech.

### 2.1.3 Stemming

Stemming is the process and the algorithm which removes the morphological information from a word, thus obtaining the root instead of the inflected form of the lemma. A stemmer, thus, is able to derive the root elect- from the different forms electing, election, elected. Stemming has a long tradition in document retrieval, and a variety of stemmers are available, see [Hull, 1996] for an overview. Probably, the most used and well-known stemmer is the Porter stemmer [Porter, 1980], which is rule-based. Stemming is used in Information Retrieval and its sub-areas as a device to enhance recall, often as an alternative to lemmatization.

A specific language, Snowball, is available and presented at <http://snowball.tartarus.org/>, in which stemmers can be exactly defined, and from which fast stemmer programs in ANSI C or Java can be generated. A range of stemmers is presented in parallel algorithmic and Snowball form, including the original Porter stemmer for English.

#### 2.1.4 Syntactic Parsing

The parsing of natural language is one of the more established areas in NLP technology. The basic process in any NLP parser is to take some natural language input and produce some representation of the syntactic structure of that input.

The representation of syntactic (grammatical) structure is given by the grammar - a description of language plus a set of structural constraints - according to which the parser attempts to analyse the symbol sequences presented to it. The input sequence is ungrammatical if there is no grammatical structure available for the sequence. Again, as it always happens when dealing with NLP, the main problem is ambiguity, for which any particular parser must have strategy to deal with. There will be more than one grammatical structure if the input is ambiguous with respect to the grammar, i.e. if the grammar permits more than one analysis of the input. The parser is incomplete if it fails to find all of the structures the grammar permits. In NLP, parsing may be of word sequences, part-of-speech tag sequences, syntactic structures, or of sequences of complex symbols such as feature bundles (e.g. where a word may have been replaced by a set of features including its orthographic form, part-of-speech, inflectional class, etc.).

Approaches can be divided into the three groups:

- Grammatical approach,
- Deterministic / non-deterministic method,
- Control strategy.

For grammatical approach, many grammatical theories have been advanced and for most of them some sort of automatic parser has been implemented. Some of them, which have been most influential in the computational analysis of language, are phrase structure grammar, tree adjoining grammar, categorial grammar, dependency grammar, transformational grammar, government and binding/principles and parameters.

The ambiguity of language is handled in parsers by adopting some deterministic or non-deterministic methods. These may range from calculating and retaining all combinatorially possible structures allowed by the grammar to discarding all but one possibility, by pruning either on probabilistic evidence (probabilities are associated with grammar rules and the parser manipulates these) or psycholinguistic evidence.

The control strategy is given by a starting point and a "direction" of browsing the space of grammatically possible structures. A sentence can be parsed left-to-right or right-to-left. Parsing may be bottom-up in a data-driven mode or top-down in an expectation-driven mode. The exploration of search space can use the depth-first and breadth-first strategies.

Despite the number of grammar theories, the mostly used and applied parsers for identification of phrase structure are based on finite state / chunking approaches. For identifying simple phrase structure a regular phrase structure grammar is sufficient, and can be implemented by a finite state automaton which is extremely efficient. Such devices, or cascades of such devices, have recently attracted interest because they provide robust, high-speed mechanisms for detecting basic phrase structure in large volumes of text.

Phrase structure approaches based on context-free grammars augmented with feature structures have also been pursued at some length in applied systems as they offer greater expressive capacity without necessarily becoming computationally infeasible.

Applications of parsing include everything from simple phrase finding, e.g. for proper name recognition, to full semantic analysis of text, e.g. for information extraction, information retrieval, or machine translation.

On top of the syntactic representation it is possible to build another layer of description, i.e. the semantic one, by assigning a particular "meaning structure" to a linguistic input of some kind. This more advanced task is known as semantic parsing, or semantic analysis, and requires a wide range of linguistic resources which attempt to use knowledge from lexicons and grammars to provide various forms of "meanings".

The list here below groups some of the instruments available.

- CASS [Abney, 1996]: a fast, robust partial parser developed by Steven Paul Abney (<http://www.vinartus.com/spa/>)
- Marcaus BaseNP chunker (<http://www.dcs.shef.ac.uk/~mark/index.html?http://www.dcs.shef.ac.uk/~mark/phd/software/chunker.html>)
- YamCha (<http://chasen.org/~taku/software/yamcha/>) chunker
- Stanford parser [Klein and Manning, 2003] (<http://nlp.stanford.edu/software/lex-parser.shtml>)
- MINIPAR [Lin, 2001] (<http://www.cs.ualberta.ca/~lindek/minipar.htm>)
- CHUNK-IT and IDEAL, a chunker [Lenci, 2001] and a dependency parser [Bartolini et al., 2002] for Italian language developed at CNR-ILC.

### 2.1.5 Word Clustering

Word clustering is a technique for partitioning sets of words into subsets of semantically similar words and is increasingly becoming a major technique used in a number of NLP tasks ranging from word sense or structural disambiguation to information retrieval and filtering.

Two main different types of similarity are considered:



1. paradigmatic (or substitutional) similarity (or also "semantic similarity"): two words that are paradigmatically similar may be substituted for one another in a particular context. For example, in the context "I read the book", the word "book" can be replaced by "magazine" with no violation of the semantic well-formedness of the sentence, and therefore the two words can be said to be paradigmatically similar;

2. syntagmatic similarity (also called "semantic relatedness"): two words that are syntagmatically similar significantly occur together in text. For instance, "cut" and "knife" are syntagmatically similar since they typically co-occur within the same context.

Both types of similarity, computed through different methods, are used in the framework of a wide range of NLP applications.

For the semantic similarity, two methods are usually exploited, namely:

1. on the basis of taxonomical relationships such as hyperonymy and synonymy; it presupposes prior availability of independent hierarchically structured sources of lexico-semantic information such as WordNet (see paragraph 2.3.3),

2. through distributional evidence; the semantic similarity between two words  $W_1$  and  $W_2$  is computed on the basis of the extent to which their typical contexts of use overlap.

Concerning semantic relatedness, a typical co-occurrence of two (or more) is the major indicator for the syntagmatic similarity. The word co-occurrence patterns are instrumental for identifying clusterings of semantically similar words not only for distributionally-based semantic similarity, but also for the semantic relatedness of syntagmatically similar structures.

- A large number of potentially important applications exists for word clustering techniques, including:
- helping lexicographers in identifying normal and conventional usage;
- helping computational linguists in compiling lexicons with lexico-semantic knowledge;
- parsing highly ambiguous syntactic structures (such as noun compounds, complex coordinated structures, complements attachment, subject/object assignment for languages like Italian);
- sense identification;
- retrieving texts and/or information from large databases;
- constraining the language model for speech recognition and optical character recognition (to help disambiguating among phonetically or optically confusable words).

### 2.1.6 Multiword Recognition and Extraction

Multiwords are units composed by two or more lexical forms. Types of multiwords vary greatly since they comprehend idioms, phrasal verbs, lexical and grammatical collocations, compounds.

Particularly important, especially for recognition and extraction of terminology, is the identification of noun phrases in text. To reach this goal, systems adopt linguistics- or statistics-based approaches.

In the first type of method, systems (such as [Justeson and Katz, 1995], [Daille et al., 1994]) rely on syntactic criteria, trying to identify significant noun phrases by recurring to language-specific regular expressions. Examples of patterns for English are Adjective-Noun (musical instrument) and Noun-Noun (wave length), while for Italian examples are Noun-Adjective (strumento musicale) and Noun-Prep-Noun (lunghezza d'onda).

Statistical approaches involve the evaluation of the frequency of occurrence of the potential multiword term ([Justeson & Katz, 1995], [Daille et al., 1994], [Enguehard & Pantera, 1994]). More sophisticated measures can be taken into account, such as mutual information, loglike coefficient, diversity, F2 coefficient.

Particularly significant is the notion of mutual information, an association ratio used to measure the degree of cohesiveness of two words (an in this sense, well suited to extract frozen compounds).

### 2.1.7 Named Entity Recognition

In the Named Entity Recognition (NER) task, systems try to identify elements in text to be mapped onto predefined categories such as names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. As all NLP technologies, also NER systems can be developed by adopting rule- or statistics-based approaches; in the first case, hand-crafted grammars are used while statistical NER systems typically require a large amount of manually annotated training data that can be, differently from what happens with grammars, easily ported to other languages and domains.

Many are the fields interested by this kind of analysis and important results were obtained in different branches such as bioinformatics or medicine, where techniques were adopted to extract for example names of genes and drugs. NER, a task closely related to Information Extraction [3.2.7] and to Automatic Term Recognition (ATR) in general, is extremely important to allow effective content and information access, as it happens when a NER module is incorporated in Question Answering applications.

### 2.1.8 Word Sense Disambiguation

One of the first problems that is encountered by any NLP system is that of lexical ambiguity (polysemy or synonymy), be it syntactic or semantic. The problem is that words often have more than one meaning, sometimes fairly similar and sometimes completely different. Actual meaning of a word in a particular usage can only be determined by examining its context. The resolution of a word's syntactic ambiguity has largely been solved in language processing by part-of-speech taggers which predict the syntactic category of words in text with high levels of accuracy [Brill, 95]. The problem of resolving semantic ambiguity is generally known as word sense disambiguation (WSD) and has proved to be more difficult than syntactic disambiguation.

All disambiguation work involves matching the context of the instance of the word to be disambiguated with either information from an external knowledge source (knowledge-driven WSD), or information about the contexts of previously disambiguated instances of the word derived from corpora (data-driven or corpus-based WSD). Any variety of association methods is used to determine the best match between the current context and one of these sources of information, in order to assign a sense to each word occurrence.

In the knowledge-driven approach, disambiguation is carried out using information from an explicit lexicon or knowledge base. The lexicon may be a machine-readable dictionary, thesaurus, or it may be hand-crafted. This is one of the most popular approaches to word sense disambiguation and amongst others, work has been done using existing lexical knowledge sources such as WordNet, machine-readable dictionary (LDOCE), and Roget's International Thesaurus.

The corpus-based approach attempts to disambiguate words using information which is gained by training on some corpus, rather than taking it directly from an explicit knowledge source. This training can be carried out on either a disambiguated or raw corpus, where a disambiguated corpus is one where the semantics of each polysemous lexical item is marked and a raw corpus one without such marking.

There is a third, so called hybrid approach, which is a combination of the two previous methods. The hybrid approach can be neither properly classified as knowledge or corpus based but uses part of both approaches. A good example of this is Luk's system [Luk, 1995] this uses the textual definitions of senses from LDOCE to identify relations between senses. A corpus is then used for calculation of mutual information scores between these related senses in order to discover the most useful. Another example of this approach is the unsupervised algorithm of Yarowsky [Yarowsky, 1995]. This takes a small number of seed definitions of the senses of some word (the seeds could be WordNet synsets or definitions from some lexicon) and uses these to classify the "obvious" cases in a corpus. Decision lists are then used to make generalisations based on the corpus instances classified so far and these lists are then re-applied to the

corpus to classify more instances. The learning proceeds in this way until all corpus instances are classified. Yarowsky reports that the system correctly classifies senses 96% of the time.

Sense disambiguation is an "intermediate task" which is not an end in itself, but rather is necessary at some level to accomplish most NLP tasks. It is obviously essential for applications such as message understanding, man-machine communication, machine translation, information retrieval, content and thematic analysis, speech and text processing, etc.

### 2.1.9 Coreference and Anaphora Resolution

Another key task in NLP is Coreference and Anaphora Resolution. Anaphora is the linguistic phenomenon of pointing back to a previously mentioned item in text while coreference can be described as referring to the same referent in the real world. The two tasks are closely related: in coreference, the system has to identify referring expressions and deciding when and if these referring expressions co-refer. A referring expression may be an expression in Natural Language like "Mary" or "she" which refers to some entity, and co-reference resolution is when you try to determine that two referring expressions refer to the same thing. For example, in the sentence

"Sally and Mary were at the restaurant but Mary was worried: she had forgotten to check if the door was closed"

there are four referring expressions "Sally", "Mary", "she" and "door", but only three entities: "Sally", "Mary" and "door". The "she", however, is operating as an anaphor, that is as a referring expression which refers to some entity which has already been introduced into the discourse. In this case it is fairly simple to understand that "she" refers to Mary and that "she" and "Mary" co-refer. But we could know something about the context that would lead us to more plausibly link "she" to Sally instead. In computational discourse research, anaphora resolution studies the context beyond a sentence and often requires the use of world knowledge; that is why anaphora resolution is an extremely difficult and laborious task, in particular when the antecedent is located beyond the boundaries of the immediate sentence in which the anaphor appears.

The Lancaster Anaphoric Treebank [Garside, 1993] (100.000 words) is annotated for noun phrases, pronominal anaphora coreferences and ellipsis while a similar effort is going on at Xerox Parc in Grenoble.

## 2.2 Speech Technologies

This sub-area concerns the understanding and generation of natural language speech. Understanding and generating spoken language means to be able not only to analyse the morphological, syntactic and semantic levels but also to deal with speech signal, independently from the device, speaker or the environment. Many final applications can be foreseen (and are actually available), usually connected to the automation of complex operator-based tasks: customer care, dictation, form filling applications, provisioning of new services, customer help lines, e-commerce, etc.

### 2.2.1 Speech Recognition

Consider the difference between the following semantically and syntactically different sentences which sound similar:

“Recognise speech” vs. “Wreck a nice beach”

“Give me a new display” vs. “Gimmick a nudist play”

Research in the area of Automatic Speech Recognition (ASR) mainly attempts to address these issues using computational means by creating algorithms and systems which try to take an acoustic signal as input and then output a (hopefully valid) string of words. The task within the more complex area of speech understanding is to figure out which is the more likely interpretation, given context and a whole host of other factors.


Speech recognition systems are generally based on Hidden Markov Models (HMMs), a statistical approach whose output is a sequence of symbols or quantities and that is based on the following parameters: i) a set of states, ii) transition probabilities, and iii) observation likelihood [Jurafsky and Martin, 2000].

Speech recognition has emerged over the last few years as a technology exploitable in the fields of telephony and in many “hands-busy” or “eyes-busy” applications, such as the ones where the user has objects to manipulate or equipment to control. Examples of final applications are: automatic translation, automotive speech recognition, speech biometric recognition, dictation, voice command recognition computer user interface, home automation, interactive voice response, medical transcription, mobile telephony, pronunciation evaluation in computer-aided language learning applications, robotics.

### 2.2.2 Speech Synthesis

Speech synthesis (or Text-to-Speech) has at its heart the attempt to produce plausible generated speech from a textual input. Most people are familiar with basic speech synthesis, such as that used by Stephen Hawking to communicate from his wheelchair or systems used by blind people to read out text from a document. However, they may also be used in dialogue systems, where together with a speech recognition system, they may be used over the telephone to talk with “conversational agents” which engage in dialogues with citizens. There have been great





innovations in the development of believable speech synthesis in the past five years, although it is almost always possible to recognise that it is a computer generating the speech rather than a human.

The essential goal of any speech synthesis system is to create an acoustic waveform from some form of input – this may be a textual input (as in systems which read out text) or some other form of input (ordinarily a dialogue system will output something which is not a text input but some form of internal representation to give to a synthesis module).

## 2.3 Language Resources

[Godfrey and Zampolli, 1997] define the term *linguistic resources* as language data and descriptions in machine readable form to be used in building, improving or evaluating natural language and speech algorithms or systems. Linguistic resources are written and spoken corpora, lexical databases and terminologies, even if sometimes the term is used also to refer to software and tools that work on such resources.

A corpus is a collection of text or speech collected in machine readable form. Generally speaking, a corpus should be both *representative* of the domain under study and *balanced*. It is representative if the linguistic information found within the sample also holds for the general population of the domain under investigation. It is balanced if attempts to cover as many textual styles as possible by incorporating samples from various genres, such as poetry, fiction, prose, email etc., of different length and referring to different periods of time.

Nevertheless, the properties a corpus should have highly depend on the application and final use they are created for. For statistical NLP, the size of the corpus is much more important than its balance; in the same way, if corpora are created for a specific purpose, they should maybe not include different genres. We thus can say that the notion of balance is somehow relative and not absolute: a corpus should always be balanced within the scope of a given domain. For specific guidelines for corpus building we refer to [Atkins et al., 1992].

The current trend in computational linguistics is enriching corpora with explicit semantic and syntactic information which i) facilitates retrieval and analysis and ii) allows the learning of features and parameters by statistical algorithms.

Today, many laboratories have hundreds of millions or even billions of words under the form of corpora. These collections are becoming widely available, thanks to data collection efforts such as the following:

Association for Computational Linguistics' Data Collection Initiative (ACL/DCI), the Linguistic Data Consortium (LDC), the Consortium for Lexical Research (CLR), the Japanese Electronic Dictionary Research (EDR), the European Corpus Initiative (ECI), International Computer Archive of Modern English (ICAME), the British National Corpus (BNC), the

French corpus Frantext of Institut National de la Langue Francaise (INaLF-CNRS), the German Institut für deutsche Sprache (IDS), the Dutch Instituut voor Nederlandse Lexicologie (INL), the Danish Dansk Korpus (DK), the Italian Istituto di Linguistica Computazionale (ILC-CNR), the Spanish Reference Corpus Project of Sociedad Estatal del V Centenario, Norwegian corpora of Norsk Tekstarkiv, the Swedish Stockholm-Umea Corpus (SUC) and corpora at Sprakdata, and Finnish corpora of the University of Helsinki Language Corpus Server.

Data collections exist for many languages in addition to these, and new data collection efforts are being initiated. There are also standardization efforts for the encoding and exchange of corpora such as the Text Encoding Initiative (TEI).

Computational lexicons are language resources where lexical knowledge, i.e. knowledge about individual words in the language, is stored and organized. In the last decade the availability of large-scale lexical resources, with broad coverage and basic types of information has become a reality and many are nowadays the lexicons available for the most different languages.

### 2.3.1 Written Language Corpora

Written Language Corpora are collections of text in electronic form that has been brought together according to a certain set of predetermined criteria. They are used for extracting statistical and linguistic information and for testing hypotheses about natural language. Their traditional use is in lexicography to study word use, to associate uses with meanings and to find interesting associations among words (collocations). Nowadays they have an enormous importance in NLP, since they sustain statistical approaches based on training and testing of probabilistic algorithms. Language teachers use on-line corpora in the classroom to help learners distinguish central and typical uses of words from mannered, poetic, and erroneous uses. Corpora are also important for terminologists, since they are exploited to build glossaries and to assure consistent and correct translations of difficult terms. Commercial applications, in particular word-processors, already integrate corpora to perform spelling correctors, hyphenation routines and grammar checkers.

One of the first major collections used in computational linguistics was the Brown corpus [Kucera and Francis, 1967], designed as a representative sample of written American English and including about 1 million tagged words. Its British English counterpart was developed during the 70s with the name of Lancaster-Oslo-Bergen corpus (Johansson et al., 1978). The most important corpus of British English is however the British National Corpus (BNC), completed in 1994 and containing 100 million words. A similar effort across the ocean is the American National Corpus (ANC) project, whose result will be a massive electronic collection of American English, including texts of all genres and transcripts of spoken data produced from 1990 onward. When completed, the ANC will contain a core corpus of at least 100 million words.

A special mention deserves the Penn TreeBank, a large corpus of about 4.5 million words. What distinguishes a tree bank is a rich annotation,

usually organized in layers: this is the case of the Penn TB, that contains PoS tags and labelled brackets marking syntactic analysis.

### 2.3.2 Spoken Language Corpora

Spoken language, a trait so essential to human communication, is shaped by many factors, such as i) the phonological, syntactic and prosodic structure of the language being spoken, ii) the acoustic environment and context in which it is produced and iii) the communication channel.

Speech is produced differently by each speaker, which gives his/her own signature to the signal (determined by dialect, accent and speaking rate, social status, emotional and physical state).

Large amounts of annotated speech data are needed to model the effects of these different sources of variability on linguistic units such as phonemes, words, and sequences of words. Current recognition techniques require large amounts of training data to perform well on a given task and, to respond to this need, many efforts are underway worldwide to collect, annotate and distribute speech corpora in many languages. These corpora allow scientists to study, understand, and model the different sources of variability, and to develop, evaluate and compare speech technologies on a common basis.

The availability of corpora of spoken language is what has ignited important advances in speech and language recognition by enabling comparative system evaluation. One of first corpora used for common evaluation, the TI-DIGITS corpus, recorded in 1984, has been (and still is) widely used as a test base for isolated and connected digit recognition [Leonard, 1984].

In the United States, the Department of Defense supports the production of two early corpora: Road Rally and the King Corpus. The Advanced Research Projects Agency (ARPA) has funded TIMIT, a phonetically transcribed corpus of read sentences used for modelling phonetic variability and for evaluation of phonetic recognition algorithms, and task related corpora such as Resource Management (RM) and Wall Street Journal (WSJ) for research on continuous speech recognition, and ATIS (Air Travel Information Service) [Price, 1990], [Hirschmann, 1992] for research on spontaneous speech and natural language understanding.

More than twenty speech corpora, comprising many hundreds of hours of speech, are distributed by the American LCD (Linguistic Data Consortium). Another important activator in the field is the Center for Spoken Language Understanding (CSLU) at the Oregon Graduate Institute collects, annotates and distributes telephone speech corpora. CSLU's Multi-Language Corpus (also available through the LDC), is the NIST standard for evaluating language identification algorithms, and is comprised of spontaneous speech in eleven different languages.

Europe is by nature multilingual, with each country having their own language(s), as well as dialectal variations and lesser used languages. Corpora development in Europe is thus the result of both National efforts and efforts sponsored by the European Union (typically under the ESPRIT (European Strategic Programme for Research and Development in Information Technology), LRE (Linguistic Research and Engineering), and TIDE (Technology Initiative for Disabled and Elderly People) programs,

and now for Eastern Europe under the PECO (Pays d'Europe Centrale et Orientale)/Copernicus programs).

The European SPEECHDAT project were responsible for the creation of the European infrastructure of spoken resources while several ESPRIT projects have attempted to create multilingual speech corpora in some or all of the official European languages. The first multilingual speech collection action in Europe was in 1989, consisting of comparable speech material recorded in five languages: Danish, Dutch, English, French, Italian. Many are the corpora in Europe resulting from National efforts.

There have also been some recent efforts to record everyday speech of typical citizens. One such effort is part of the British National Corpus in which about 1500 hours of speech representing a demographic sampling of the population and wide range of materials has been recorded ensuring coverage of four contextual categories: educational, business, public/institutional, and leisure.

Other major efforts in corpora collection have been undertaken in other parts of the world.

### 2.3.3 Computational Lexicons

Computational lexicons may contain a wide range of word-specific information, depending on the structure and task of the natural language processing system. A basic lexicon will typically include information about morphology, either in a form enabling the generation of all potential word-forms associated with pertinent morpho-syntactic features, or as a list of word-forms, or as a combination of the two. On the syntactic level, it will include in particular the complement structures of each word or word sense. Very important today, for the implication they have on the possibility to move NLP towards semantic analysis, are computational lexicons which include semantic information, such as a classification hierarchy, selectional patterns, case frames stated etc. For Machine Translation and all the types of multilingual applications (such as Cross Language (CL) Information Retrieval, Question Answering, Information Extraction etc.) the availability of multilingual computational lexicons where correspondences between lexical items in the source and target language have to be recorded and stored is essential. For speech understanding and generation, lexicons listing information about the pronunciation of individual words are required.

Strictly related to the types of information connected with each lexical entry are two other issues: (i) the overall lexicon architecture, and (ii) the representation formalism used to encode the data.

In general, a lexicon will be composed of different modules, corresponding to the different levels of linguistic descriptions, linked to each other according to the chosen overall architecture.


As for representation, almost all the lexicons available are presented at least in XML format, while many are now converted in RDF and OWL representations that can be read, analysed and treated by software for the inferential analysis, such as Protégé. The TEI (Text Encoding Initiative) has developed a model for representing machine readable dictionaries.

Manual creation of computational lexicons is an extremely laborious and expensive task. This is why two areas of research have recently deserved much effort and attention, i.e. the ones dedicated to: i) the development of common, shared and standard lexical representations and co-operative lexicon development to make greater use of already existing resources, and ii) automatic learning of lexical characteristics from instances in textual corpora by means of statistical approaches.

For the first line of research, we mention the advancements in harmonization achieved in the past by a series of projects like GENELEX [Antoni-Lay et al.,1994], EAGLES [Zampolli & Calzolari 1994], MULTEXT, PAROLE, SIMPLE, ISLE [Calzolari et al., 2003]. More recently, these initiatives towards the standardization of lexical resources have been endorsed at the level of ISO/TC37/SC4, a committee dedicated to the specification of a full family of standards for language resources. Today, the most advanced standard is the Lexical Markup Framework – LMF ISO 24614 – that has the purpose of defining high level specifications which provide a meta-model to for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources [Francopoulo et al., 2006].

The first lexicons to be seen as shared resources for computational linguistics were the machine-readable versions of published dictionaries. Major efforts involved machine-readable versions of selected information from Merriam-Webster's 7th Collegiate Dictionary, from Longman's Dictionary of Contemporary English and Oxford Advanced Learner's Dictionary. Then, over the years, a number of projects to create large lexical resources for general use provided important results. Perhaps, the most successful experience was that of the WordNet family (Fellbaum, 1998, Ide et al., 1998) thanks to the ampleness of its uses, its notoriety and the numerous versions in languages other than English (and also due to the fact that it is free of charge). Always remaining in the field of WordNet, we mention also the eXtended WordNet resource (<http://xwn.hlt.utdallas.edu/>). Many other lexicons, designed according to the most different theoretical frameworks (or even supposedly theory-free) are available: the Frame-based lexicons (Baker et al. 1998), the SIMPLE lexicons (Lenci et al., 2000), the CYC ontology (Lenat, 1995), lexicons based on Lexical Conceptual Structures (Dorr, 1994), nominalization lexicons (such as NOMLEX, McLeod at al., 1998), Lexical-Semantic Databases directly obtained by MRDs (such as the Collins-Robert database, cf. Fontenelle, 1997), the Japanese Electronic Dictionary Research (EDR).

Providing an exhaustive list of CLs available for the languages of the world goes beyond the aim of this brief overview. The interested reader can refer to [Grishman and Calzolari, 1997] and [Sanfilippo et al., 1999] for an overview of the most important lexicons, their design and the type and quantity of information they store. We think however it may be useful to provide a (non-exhaustive) list of the major phenomena and information types that we can find represented in most advanced type of lexicons, i.e. the semantic one. The list is in line with the EAGLES guidelines for Lexical



Semantic Standards [Sanfilippo et al., 1999] and with the general grid for evaluating the content and structure of lexical resources proposed in the ISLE Survey of Existing Lexicons [Calzolari et al., 2001]. We do not include in the list information of a morphosyntactic or syntactic nature, even if it can also be encoded in a semantic lexical entry.

In general, we can say that the lexical entry in semantic computational lexicons can be encoded with the following information types:

- Semantic Type: reference to an ontology of types which are used to classify word senses (for example Living entities, Human, Artefact, Event etc.)
- Domain: information concerning the terminological domain to which a given sense belongs.
- Gloss: a lexicographic definition.
- Semantic relations: different types of relations (meronymy, hyperonymy, Qualia Roles, etc.) between word senses.
- Lexical relations: synonymy, antonymy.
- Argument structure: argument frames (possibly with semantic information identifying the type of the arguments, selectional constraints, etc.).
- Regular Polysemy: representation of regular polysemous alternations.
- Equivalence relations: relations with corresponding lexical entries in another language (for multilingual and bilingual resources).
- Usage: the style, register, regional variety, etc.
- Example of use.

### 3 Outline of main application areas

We present here the applications which can be considered more in line with the eParticipation aims and scope. As we said, either they are already used in existing systems or they are technologies which can have an important role in the future.


We decided to classify all the applications in three very general domains: Machine Translation, Information Retrieval and Text Mining. The distinction and classification used to introduce the various topics have obviously do not obviously have an absolute value; this is true in particular for Information Retrieval, a very general class of applications which encompasses Question Answering, Document Classification, Summarization etc. We decide to classify all the applications in three very general domains: Machine Translation, Information Retrieval and Text Mining. Different classifications may be chosen.

#### 3.1 Machine Translation

Machine Translation (MT) is the sub-area of NLP which is concerned with enabling automated or semi-automated ways of translating texts in one language to another. MT was one of the first arenas for computational linguistics in mid 50s and received great attention as a promising and extremely useful technology; the dream of the universal, global machine translation able to overcome all problems in human communication clashed soon with reality: language revealed itself to be an object far too complex to be described with a finite set of rules and heuristics and lexical, syntactic and semantic ambiguity, lexical, syntactic and semantic, revealed to be a real true bottleneck for any real-world application. In few years, very poor results were obtained.

Disappointment led to the Automatic Language Processing Advisory Committee (ALPAC) report of 1966, which sent the clear message that MT was hopeless. As a result, large-scale funding of MT research in the United States came to a virtual standstill for the next twenty years.

Now scientists, companies and institutions are aware that automatically translating an arbitrary text from one language to another and producing a high-quality output is such a difficult task to be far too hard to be completely automated. Nevertheless, in the last years we are witnessing a revival of interest for MT: the need for translation continues to increase significantly due to the web advent and translation technology has improved since it first appeared in the mid-50s. As a matter of fact, more advanced NLP tools are now available (in particular syntactic parsers that are today able to analyze complex grammatical forms), together with different supports for a variety of fields of expertise (such as dictionaries in law, medicine and information technology for many language pairs). Despite recent improvements, translation technology is not very accurate when used as a standalone solution. At the same time, human translation



cannot keep pace with demand and a solution is MT is thus considered as an aid to translators (as it happens for Computer-aided translation, CAT).

In general, MT systems focus on the following tasks:

- for which a rough translation is enough
- where a human post-editor can improve the output
- limited to narrow domains in which good quality is achievable

In developing MT systems, three main approaches are usually followed:

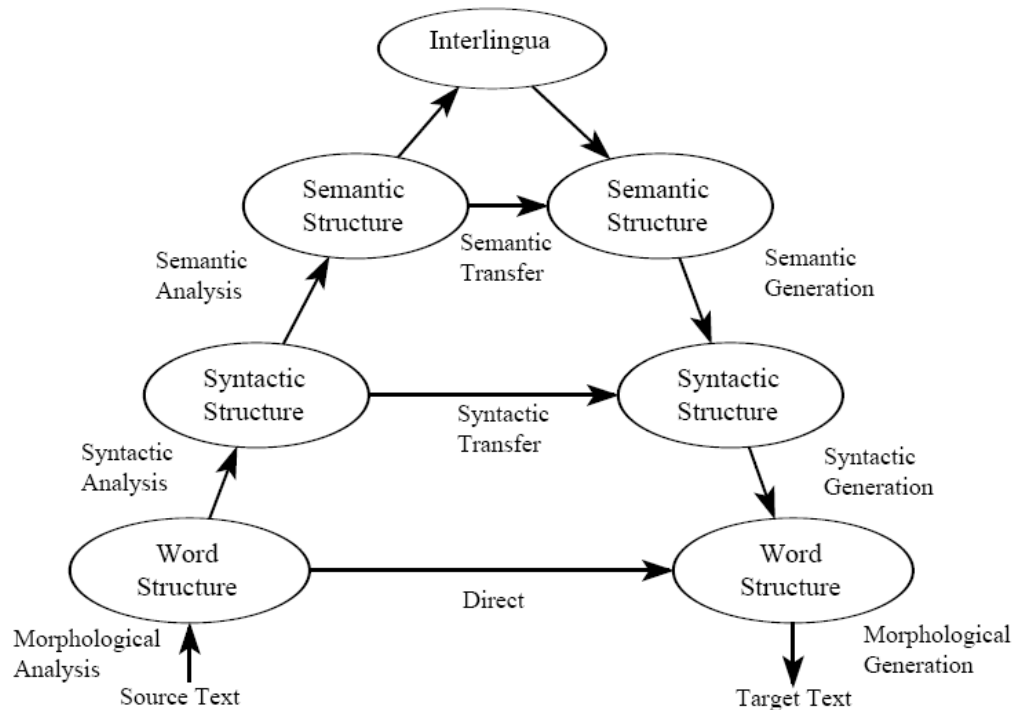
- Direct
- Transfer
- Interlingua

Fig 1 (from [Dorr et al., 1998]) gives an idea of the different levels of complexity involved in the three approaches.

In the direct model, a “primitive form of transfer” is adopted, since simply word-for-word replacement is foreseen.

In the actual transfer model, the system has to provide syntactically correct target language text by transforming source-language representations into suitable target-language representations. Both the transfer and the interlingual approaches require linking rules that map the surface text to some form of internal representation. In case of transfer-based MT systems, however, this internal representation is assumed to vary widely from language to language: this determines that for each source-language/target-language pair specific sets of rules have to be created. This is the major drawback of this type of approach. All the systems of this type exploit some form of syntactic analysis while many of them also move to semantic analysis in order to obtain more accurate results.

The interlingual approach is based on the idea that the analysis of the source-language text is somehow independent from the source-language. The target-language text is then generated from a neutral, interlingual representation, which works as a sort of pivot among different languages. The very interesting aspect of such an approach is that the representation development and linking rules have to be created only once for each language.



**Fig. 1. Types of MT systems (from [Dorr et al., 1998])**

Moreover, different paradigms of MT research may be identified. Those that:

- rely on linguistic techniques: for these approaches, MT is grounded on principles of linguistic theory. Constraints at syntax, lexical and semantic level are used to identify the target language equivalent. Under this type of approach, [Dorr et al., 1998] classifies Constraint-based MT, Knowledge-based MT, Lexical-based MT, Rule-based MT, Principle-based MT, Shake and bake MT.
- rely only on statistics without recurring to linguistic knowledge: these approaches are enabled by the availability of great computational power and huge corpora which are used for training and for storing examples of translation. Under this type of approach, [Dorr et al., 1998] classifies Statistical-based MT, Example-based MT, Dialogue-based MT, Neural network based MT
- use a combination of the two approaches, adopting hybrid methods.

MT is maybe the most representative application of NLP. Many are the resources available, as well as papers and books. A good starting point for the interested reader is the Machine Translation Archive (<http://www.mt-archive.info/>), Electronic repository and bibliography of articles, books and papers on topics in machine translation and computer-based translation tools, which store more than 3000 items.

Moreover, many are the International Associations dedicated to the topic. Their web sites are actual portals to navigate the huge quantity of material available:

- Asia-Pacific Association for Machine Translation (<http://www.aamt.info/>)
- Association for Machine Translation in the Americas (<http://www.amtaweb.org/>)
- European Association for Machine Translation (<http://www.eamt.org/>)
- British Computer Society Natural Language Translation Group (<http://www.bcs-mt.org/index.htm>)


### 3.1.1 Current exploitation in eParticipation

The interest of MT is self-evident in eParticipation. Different aspects need to be discussed:

- challenges connected to content provision for eGovernment services (typical of institutional, public web sites, in a mostly vertical, unidirectional model of communication);
- tools for professionals of PAs, officials and politicians, mainly used or usable in back-office tasks (for example drafting tools which allow the preparation of legal text in native language and then request a machine translation step followed by manual revision);
- necessity to translate more spontaneous content, created by citizens and people around the world in a more citizen-to-citizen, horizontal model of communication.

The first two situations are connected. The most common strategy for handling multilingualism in public institutional web sites consists in making documents available in multiple languages. Large political bodies benefit of translation services which also make use of MT. The European Commission's Directorate-General for Translation, for example, offers a machine translation service for 28 language combinations. The service is known as EC Systran, which is based on SYSTRAN™ technology. Within the Commission, administrators use the system as a browsing, translation and drafting tool, whereas the Translation Service employs it almost exclusively as a translation aid. MT was not only developed in order to cope with the Commission's fast-growing internal demand for translation but also with a view to overcoming the language barriers between EU Member States, in line with Information Society initiatives. Consequently, EC Systran is freely available to EU public administrations, including EU institutions and bodies, government ministries, national parliaments, regional authorities and universities.

To this service, more of hundreds of thousands of pages are submitted every year, around 80% of which come from the Commission itself. The remaining 20% is accounted for by other EU bodies and the Member States. Spanish and German authorities (at both national and regional



level) are the main national users, but demand is also increasing in France.

The main reference for the MT effort in European eGovernment is surely the MT section of the IDABC (the Community Programme managed by the European Commission's Directorate General for Informatics). IDABC-MT covers a series of projects aimed at providing more effective and user-friendly access to the European Commission's machine translation service for the interchange of multilingual data between European public administrations. A significant achievement of IDABC-MT is the Feasibility Study, carried out within the IDA framework (Interchange of Data between Administrations) whose purpose was to:

- collect information on the principal MT needs of European public administrations by means of a survey; and
- define the most appropriate and cost-effective ways of satisfying those needs in terms of access, language and terminology coverage, and translation quality.

The feasibility study is interesting also because shows that many national administrators were unaware that MT technology was at their disposal and that only a third of them had used MT before. Other efforts of IBAC-MT were addressed to:

- Integration of terminology with around 6700 terms provided by national administrations.
- Improvement of access allowing real-time translation and a MT via a Web Services platform that can be integrate in national administrations web sites.
- Study on MT products for the "new" EU languages (Bulgarian and Romanian)

Machine translation to analyse textual content created by citizens is usually performed in a different ways: non-official, non-institutional forums and web sites often put at users' disposal links to web-based machine translation systems. Machine translation in blogs or websites is probably the easiest way of helping readers from different countries communicate. Even if a blogger writes in Spanish and a reader speaks English, or vice versa, both of them can still understand each other. Free language translation tools let bloggers post messages simultaneously in English, German, French, Spanish and other languages. The most popular free Web-based language translation tools are offered by Google, AltaVista Babelfish and WorldLingo. Using these services, web audiences who speak English, French, Spanish, German, Italian, Portuguese, Dutch, Greek, Chinese (Traditional and Simplified), Japanese, Korean or Russian will be able to translate websites into their native language. With over 65% of web users speaking a language other than English, providing the means of translating websites with English content to another language is very essential.



**Alta Vista Babel Fish:** Babel Fish Translate is available in English, German, Spanish, French, Italian and Portuguese.

**Google Translate:** Available in English, German, Spanish, French, Italian, Portuguese, Japanese, Korean and Chinese (Simplified).

**WorldLingo URL Translator:** Google indexes the static URLs translated by World Lingo. Thus, a website can be found on Google even while searching for non-english terms. World Lingo supports English, French, Spanish, German, Italian, Portuguese, Dutch, Greek, Chinese (Traditional and Simplified), Japanese, Korean and Russian.

These free-of-charge functionalities have rarely accurate results; nevertheless, they provide a fast translation which is often enough to "understand" the topic at hand and to form an idea on the positions in the field.

A particularly interesting application would be the one concerning social networking technologies (such as Skype, Youtube, Facebook, MySpace etc). But we know that MT is not currently used in these environments, even if the portion of non-English content is growing. Skype, for example, use a mix of Internet calling technology and personal (human) interpreter services, to give its users the ability to talk instantly to anyone, anywhere in the world, regardless of language ([www.languageline.com/skype](http://www.languageline.com/skype)). Also the other common social networking site, Facebook, is working to make its popular social-networking site available in languages other than English. Nevertheless, Facebook has been unclear about timing, and has not released any kind of official "coming soon" teaser. Multilingualism is a very important feature of the Wikipedia world: the content is however translated by human in a collaborative way (this does not exclude that humans can use some translation software if they want).

Even if machine translation is not used yet in collaborative environments, its exploitation would be of great interest. One project which shows the level of attention on the subject is the Language Grid Project [Ishida, 2006] (<http://langrid.nict.go.jp/publicatione.htm>), studied and developed at the Japanese National Institute of Information and Communications Technology (NICT). The aim of the project is supporting intercultural collaboration by developing an infrastructure that is built on the top of the Internet. It allows a better understanding of Internet contents written in different languages and by people from different countries. In addition, the Language Grid allows users to easily develop new language services by combining existing ones to satisfy their needs. Within the project, a number of different tools (which combine specialized dictionaries and machine translators) are envisaged, such as:

- Langrid Chat

A chat tool with multilingual translations.

- **Langrid Blackboard**

A tool for summarizing and sharing information. Users can create cards in their first languages and put them on a shared workspace. The texts on the cards are immediately translated into multiple languages. This tool is useful for international conferences, where people always summarize discussions for international attendees, or seminars for exchange students who need language support.

- **Langrid Input**

A multilingual input interface for existing collaboration tools, such as BBS. Input texts are translated in real time into various languages and sent to existing tools. Users can multilingualize existing tools by attaching the Langrid Input to them. Users can also edit specialized dictionaries by using the Langrid Input.

Results of this project would represent an important move forward in the field of intercultural collaboration and machine translation. Nevertheless, if we look at the situation of current exploitation of this kind of technologies, we have to say that many fields are not covered yet by an extensive use of MT. This gives an idea of a certain immaturity of the technology, even after all these years.

### **3.2 Information Retrieval**

IR is a wide-ranging field which encompasses the representation, storage and retrieval of diverse kinds of media. IR, especially document retrieval, is often not considered to be one of the main NLP areas, because often different techniques may be used to conceptualise information, usually statistically based, which do not necessarily look at the linguistic features of what it is they are dealing with. However, there has recently been considerable interest in the NLP community concerning statistically motivated methods for describing textual information, and there is an increasing convergence in the methods used by those within the IR and NLP communities. It should be said that it is usually the case that those working in the area of IR will be in a different research group to those working in NLP, and will tend to go to different conferences. But there are obvious overlaps and more and more NLP researchers are getting interested in getting involved in the IR community.

NLP techniques may be used in assisting the retrieval of mainly text based documents, but also to augment the representation and retrieval of different forms of media – such as film, photography and music.

Traditionally speaking IR conceives of the task of retrieving information as retrieving a document from a set of documents given a certain search query. Many are the definitions that can be found in the literature. Here we report three of them:

- *IR is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. (. . . ) IR embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques, and machines that are employed to carry out the operation. [Mooers, 1951]*
- *IR is the discipline concerned with the storage and retrieval of documents; its goal is the realization of computer systems that allow the storage of large quantities of documents in a document base, in such a way as to allow the efficient retrieval of the documents relevant to users' (information) needs.*
- *IR deals with computer-supported access to data with poorly understood semantics. [Lewis, 1991]*

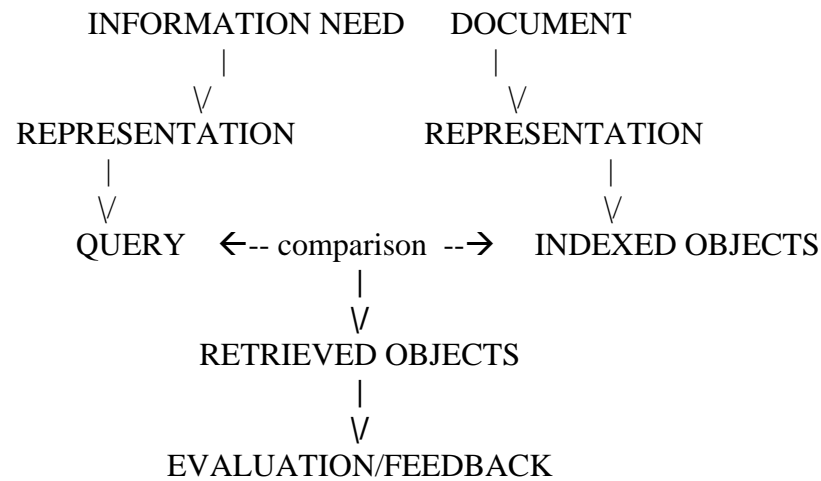
The last definition encompasses several tasks performed on several media. Among the tasks are:

- Search (retrieval)
  - > text search coincides with IR "narrowly conceived", and is also known as ad hoc retrieval
  - > a particular case of search is known-item search, when the target is a particular document that the searcher knows already to exist in the document base
- Metasearching
- Passage retrieval
- Categorization
- Filtering
- Collaborative filtering
- Clustering
- Summarisation
- Question answering
- Information extraction

The definition of IR that we espouse here is the largest one, the one that consider IR as something which encompassed several different sub-fields. For this reason, we choose to introduce Question Answering, Classification and Summarization within this same section. Other classifications may be possible, of course, but we think that all these applications have in common the satisfaction of specific informative needs of users. We first introduce the basic structure of an IR systems and then the other sub-fields.

### 3.2.1 Basic Architecture of a IR system

Most IR systems depend on the notion of a query and a set of document (see Fig. 2):




**Fig. 2: the IR process (from [McCallum, 2000])**

We may characterise the information retrieval process as constituting 3 parts

1. Question formulation
2. Answer provision
3. Assessment of answer (from [Belew 2000])

We should not confuse this process of constructing an answer with the field of Question Answering (see section 3.2.6), because the "answer" is not directly given by the IR system; it is up to the user to find the answer from the document. If the answer is not sufficient, the user must revise the query that they have made.

The first step is initiated by particular individuals and their own questions. Typically, a search engine, because of its ordinary ad-hoc nature, knows nothing about the user in question – and it cannot "read" the mind of the user either. All that can be done is to look at how individuals represent their question in the form of a query. However, this is the first hurdle in an information retrieval system – sometimes users of a system are unable to characterise their own query because they are insufficiently familiar with the domain in question – this is often called their "knowledge gap". This may inhibit them from coming to the answer, or the possible information sources which will provide an answer. Part of the strange name of posing questions is that if you need to know enough about the domain in order to get the appropriate answer that you want – but in fact sometimes you want to be able to ask general questions when you aren't that familiar with a domain which would allow you to make a more specific



query. Either way, the way that any current information retrieval interfaces with a user is in the taking of queries from them which are constructed using some sort of query language. Since the advent of Google and other web-based IR systems, the use of controlled vocabulary to construct queries has become less popular, and this has been replaced with more natural "free text" queries – these are queries where you just type in some words and try to get the IR system to retrieve the best match.

The second step is the construction of an answer (or, more typically, "answers"). The main problem here is that not all "questions", indeed not all questions as formulated using the queries which IR systems allow, really have "answers". Document retrieval typically views the process of "answering" a query as one that requires the provision of a set of documents, more usually an ordered set.

The third step, closely allied to the second and the first, is the assessment of the set of documents which have been retrieved. This is typically considered to be a "closing of the loop" between asker and answerer, where the asker, the user, is allowed to say how relevant the documents retrieved for their own requirements. The process of communicating whether or not a document is relevant or not is called "relevance feedback", and the way this may be realised in a system can vary quite considerably.

Most IR systems may be considered, then, as consisting of 3 things – a document set (that is, the set containing all possible candidate members for each set of answers), a user with an information need, or a "question", and a means of specifying this information need as a query to the document set. The way we describe documents in some way orients the way we retrieve them. Although in this subchapter we are dealing with IR in itself, we deal with the specifics of different forms of document classification in a different section, as these are not just about retrieval but also about allowing users to orient themselves around particular documents. However, here we will deal with the forms of document classification which constitute classical approaches to IR. There are three main ways of doing this:

- controlled vocabulary indexing
- free text indexing -- a free text query: a query where there are no search operators, considering a query as merely a set of terms. The query is then considered to be a "document" and a similarity score is computed, which tries to match the most similar representation of the query to a set of hopefully relevant documents
- full text indexing (McCallum 2006)

The way the query is specified to some extent guides how we match the set of relevant documents, which a user typically nowadays is unlikely to be able to go through on their own. Thus the rank ordering of documents to matching any particular query is essential. An IR system will compute a score for every document which is returned as the result of a query. It

can do this by not only taking information from the document itself, but by also looking at other elements such as its title, the name of the author, the date of creation – and looking at structural aspects of the document, such as whether it has an “introduction”, a “conclusion”, a series of subsections, whether it has pictures and whether these are labelled.

However, the most popular way of assigning a score is known as tf-idf (term frequency/inverse document frequency), where a document,  $d$ , is given a series of term weight in document that is


1. highest when  $t$  occurs many times within a small number of documents (thus lending high discriminating power to those documents);
2. lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
3. lowest when the term occurs in virtually all documents. (from Schutze, 2007)

A popular representation of a set of documents is the vector space model, where documents are considered as vectors in a common vector space – it is probably the most popular current method for achieving a number of IR operations such as scoring documents based on the given query, classification of documents and the attempt to sort documents into “clusters” which sort them into thematic groups. Typically a query is considered as a “bag of words”, where it itself is considered to be a document. Thus a similarity score may be deduced between the query, as represented as a vector, and the documents which are also represented as vectors. Using a cosine similarity measure, it is possible to see which documents most closely resemble the “query” – thus providing a simple and effective way of having a common representation for the queries and the documents that can allow for the best scoring of similarities between a user’s “question” and the set of possible answers which are presented to him or her.

A popular method for displaying the results of a query is by the user of “snippets” – it is likely that a user doesn’t want to have to go through all the possible candidate answers that are retrieved for them, but would rather have some sort of short summary or some short “snippet” of the text in question that can allow them to decide whether or not they want to actually look through the body of the document itself. A snippet normally has the document title and a short summary, which is automatically extracted. These will ordinarily be constructed in relation to the query which the user had made – these are called “dynamic summaries”, which are to be considered different from static summaries.

### 3.2.2 NLP and IR

As we said, IR is not necessarily a NLP task. Nevertheless, NLP can play a role in IR, in particular for what concerns the translation of potentially ambiguous natural language queries and documents into unambiguous



internal representations on which matching and retrieval can take place. An "ideal" IR system should "understand" the underlying meaning of the query, in this way allowing a conceptual matching of queries and documents. Nevertheless, many are the reason for which NLP techniques are not very much applied in common Search Engines. Above all, there is a general lack of promising empirical results which tease out the individual contributions of each of the levels of processing. Moreover, the complexity of processing required by some of these higher levels of language understanding generates some concern in developers. Nevertheless, NLP offers to IR a very broad range of tools which can sustain the IR process providing analysis at various levels of linguistic description and complexity.


The phonological level is useful in speech recognition systems which accept spoken queries or even provide spoken documents.

The morphological level is the level of linguistic processing most commonly incorporated in IR systems: often terms in documents and queries are stemmed [no description of such technology in this deliverable] so that morphological variants between query and document will match. There have been differing empirical results on the impact of stemming in English, most current IR systems support stemming to avoid the potential for obviously missed relevant documents. For other languages that have a richer morphology, the attention to morphological processing offers a much more obvious and larger pay-off for IR than it does for English.

A very disputed issue concerns the lexical level: since the dawning of IR, the role of lexical knowledge provided by thesauri and other similar resources has been considered as a source of improvement. At the very beginning, the lexical knowledge consisted in manual consultation tools for both indexers and searchers. They were utilized to ensure that a common vocabulary was used in selecting appropriate indexing or searching terms. Now, with the advent of statistical approaches, the use of such resources is really not relevant.

A very interesting application concerns the syntactic level of linguistic processing, which is used to obtain a better indexing of documents via the recognition of phrases instead of single-word indexing which frequently introduces ambiguity into the representation and resultant retrieval.

Use of the *semantic* level of language in IR includes interpretation of the meaning of sentences as the unit of understanding, as opposed to processing at the individual word or phrase level. This level of processing can include the semantic disambiguation of words with multiple senses. More often in the past, knowledge resources were considered important to carry out query expansion, i.e. the addition to the terms in the query of synonymous keywords. Term expansions can be obtained from lexical sources such as WordNet or IR-style thesauri. Query expansion, in particular in open domain, is rarely useful, since the challenge is to add just those terms which are expansions of the particular sense of the word



intended in the query. Another usage of semantic processing is the production of semantic vectors to represent both queries and documents, but this also requires that the appropriate sense of each term has been determined and the appropriate semantic category selected for inclusion in the semantic vector.

In IR, the discourse level structure can be utilized to understand what the specific role of a piece of information is in a document, for example - is this a conclusion, is this an opinion, is this a prediction or is this a fact? Additionally, anaphora recognition and resolution would provide an improved representation of both documents and queries since it would enable the implicit presence of concepts to be more completely accounted for at the lexical level and an integrated representation of the contents of a query or document to be produced at both the semantic and discourse levels.


While experiments of NLP exploitation have been carried within research systems (see Liddy and Myaeng, 1994], rarely linguistic enhancements have reached commercial search engines, in which the exploitation of NLP is often rudimentary if not absent at all. Stemming is the most used NLP solutions but very interesting is the automatic identification of proper nouns and other Named Entities by means of Name Entity Recognizers. Also phrase and multiword identification, as well as concept identification by means of statistical co-occurrence measure, is very useful.

Nowadays we can state that IR remains mainly a coarse task, while NLP exploitation is much more common in sub-areas of IR, such as Question answering and Information Extraction. We will give a description of these fields in following sections.

### **3.2.3 Cross-Language Information Retrieval**

Though initially the WWW was dominated by English speakers; now less than half of existing web pages are in English. Accessing information in a host of languages is clearly important for many uses. In monolingual retrieval, the queries are in the same language as the collection being accessed. The purpose of CLIR is to support queries in one language against a collection in other languages. Cross-lingual document retrieval performs essentially as accurately as monolingual retrieval. The result above indicates that the initial milestone for CLIR has been achieved. The technology may be applied to other sub-applications: Cross-lingual Question Answering allows posing factoid questions in one language and receive answers in the same language, on the basis of a multilingual set of documents. Also Cross-lingual summarization would provide short summaries in a specific language starting from documents in other languages.

One major contributor to progress in CLIR has been formal evaluations and data that have been made available through the Text Retrieval Conferences (TREC), the Cross-Language Evaluation Forum (CLEF), and the NII-NACISIS Test Collection for IR Systems (NCCTR).



The main approach used in CLIR is the probabilistic translation model. The techniques use a mix of manual bilingual dictionaries and statistical bilingual dictionaries automatically derived from corpora of translated documents. It is interesting that machine translation systems for the document language are not a prerequisite.

### 3.2.4 Document and Text Classification

As we said, IR encompassed several different sub-fields. One of them is Classification, which is often undertaken in order to make the representation of the document in question more amenable to retrieval through a common document and query representation. The first reason behind text classification is distinguish between different kinds of documents in order to facilitate the retrieval of the information we are looking for. If we have a set of 100 documents, and we know that some are about a topic and some about another, it would be useful for us to distinguish between the two different kinds of documents, in order to retrieve only the documents pertaining the topic we are interested in. Or we may want to be able to distinguish between the Spam documents in our Inbox and the ones that are genuine mails. Classification may also be a way of assisting users with identifying documents which are similar to ones which they have been looking for, or to imposing a structure on a set of documents that wasn't there before that can allow for easier navigation between them.

There are a number of different approaches to text classification:

- single category text classification, where each text belongs to exactly one category.
- multi-category text classification, where each text can have zero or more categories.
- clustering, where the set of categories is not predefined.

The actual approaches to these tasks vary according to the different techniques one can currently use. Maximum entropy, vector space models and Latent Semantic Indexing are three popular ways of approaching these problems. Classification can be carried out also on the basis of language, genre and author.

See [Sebastiani, 2001] for a survey of the machine learning approaches to classification.

### 3.2.5 Summarization

Automatic Summarization [Many and Maybury, 1999] is the sub-area of NLP concerned with the creation of shortened versions, i.e. summaries, of either one document (document summarisation) or of a series of documents (multi-document summarisation), from either one language or many. Such summary still contains the most important points of the original text. The notion of what is the essential information in a text is relative to the goals of the user who wishes to use the summary as a surrogate for the original(s), and varying types of summaries may be produced to meet the needs to various users. This dependency implies a

requirement of a model of human comprehension and production of language and involves a deep processing of both aspects of language – understanding and generation.

Application areas of Automatic Summarization are:

- web-content production from existing textual resources,
- web search engines – providing extended summaries of retrieved documents according to the user's needs,
- creation of knowledge bases / ontologies from a set of documents describing an application domain,
- customization of information for different channels and formats (e.g. paper newspaper, web, WAP, SMS message, radio, and a spoken newspaper for the visually handicapped), especially if it involves shortening of original texts,
- multilingual document production – increasing availability of various documents across language borders, e.g. legal ones, by first summarizing them before translation,
- preparation of information for use in small mobile devices, which may need considerable reduction of content,
- the techniques used in Automatic Summarization interesting spin-off effects in the area of advanced search engine technologies in form of stemming, query expansion, algorithms for discourse segmentation, the use of synonym dictionaries, as well as spell checking of the query.

### 3.2.6 Question Answering

Question Answering is an application that allows the user to obtain brief and concise answers (instead of whole documents) in response to written natural language questions. Today, a very large number of implemented QA systems is available for English while the number of new systems dedicated to languages other than English is constantly growing. Every year, the results of literally dozens of QA systems are presented to the two conferences that host a QA track, i.e. the TREC ([www.trec.nist.gov](http://www.trec.nist.gov)) and the CLEF ([www.clef-campaign.org](http://www.clef-campaign.org)) campaigns.

The TREC in particular has defined the "borders" and the characteristics of the so-called Open-Domain Question Answering, i.e. the task of identifying, among large collections of documents, text snippet where the answer to a natural language question lies. In this view of QA, the answer is usually constrained in a given text span (for example 50 bytes) and the system incorporated an index of the collection and a paragraph retrieval mechanism.

But these definitions mainly hold to current QA systems that submit their results to the TREC (and now CLEF) evaluation campaigns, while the QA concept is in general much wider and comprehensive of different sub-tasks and approaches.

[Hirschman and Gaizauskas, 2001] indicate the following set of dimensions of the "QA problem": i) applications, ii) user, iii) question types, iv) answer types, v) evaluation and vi) presentation.

The applications of QA can vary depending on many factors, such as the source of the answer (structured, semi-structured data or free text), the type of textual collection (a single text, the fixed set of documents typical of the TREC and CLEF campaigns, encyclopaedias, the open-ended Web), the topics covered by the questions (close-domain or open-domain QA) etc.

Different users, on the basis of their specific expertise and aims, could require different types of answers, of different granularity and depth: the requirements and skills of a professional analyst and of an average InterNet user are surely different.

The type of question is probably one of the most important factors effecting performance of QA. [Hirschman and Gaizauskas, 2001] distinguish questions on the basis of the possible answers, thus identifying factual, opinion and summary questions, yes/no questions, Wh-questions, commands.

Also the type of answer plays an important role in approaching the QA: answers can be extracted (cutting pertinent snippets of text) or generated, can constitute a list and can also be intended as a summarization of a longer text.

For presentation we intend the modalities that the systems adopts when presenting the answer to the user. The answer can be released for each question without any connection with previous answers but it may be the case that a sort of dialogue is engaged between the user and the system. We can also suppose that, if the system can handle speech input and dialogue, a true conversational access to information (for example to content of web pages) could be achieved.

All this dimensions cut across the QA problem, determining a very large variety of possible instantiation of the same "application type".

This extreme variety of possible results is just what makes of QA such an interesting application from an industrial perspective: as a matter of fact, in recent years we have witnessed an exponential growth of the interest in QA, in particular since the availability of huge document collections (e.g. the web itself) has ignited the demand for better information access. But Question Answering is not new: researchers have always been fascinated by the idea of answering natural language questions and first Question Answering systems date back to the 1960s.

In order to provide a brief historical account of QA, we refer to the surveys presented in [Hirschman and Gaizauskas, 2001] and [Monz, 2003]. But [Simmons, 1965], by the middle of the 60s, already illustrated

about fifteen implemented English language question-answering systems built over the previous five years.

The historical account in [Hirschman and Gaizauskas, 2001] provides a coarse classification of typologies that interprets and “declines” the general notion of QA in different final applications, constituted by: i) conversational question answerers, ii) front-ends to structured data repositories and iii) extractors of answers from text sources (as encyclopaedias). Front-ends systems are interface to a structured database, according to the assumption that it would be useful to provide the final user with the possibility of accessing vast amounts of highly detailed information using natural language rather than a specialized query language (e.g. SQL). In this sense, these systems represent a mechanism to negotiate between the natural language of the user and the formal language of the database. Examples of this type of QA are the well-known systems BASEBALL [Green et al., 1961], STUDENT [Winograd, 1977], LUNAR [Woods, 1973] etc. What marks this approach to QA (i.e. the source of knowledge) is also its strongest limit: it is unrealistic to consider the possibility of scaling-up this type of system from very narrow and specific domains to open-domain. The recent interest for QA is motivated by the necessity to access the content of vast amount of unrestricted texts and answering questions over the web is a kind of Holy Grail that tows all the research efforts in this field. This necessity is behind the current research, driven by programs such as AQUAINT and evaluation exercises such as TREC, NTCIR and CLEF, all of which focus on open-domain question answering. The availability of large volumes of data (e.g. documents extracted from the World Wide Web) has prompted the development of systems that focus on shallow text processing. Nevertheless, also in view of our interest in application in eParticipation domain, we have to highlight that there are many document sets in restricted domains that are potentially valuable as a source for question answering systems. There is a wealth of information in technical documentation such as software manuals, car maintenance manuals, and encyclopediae of specific areas such as medicine. Users interested in these specific areas would benefit from QA systems targeted to their areas of interest. In this sense, a distinction has to be made between:

- Restricted(/Closed)-Domain Question Answering, which deals with questions under a specific domain (for example, medicine or law), and can exploit very detailed and domain-specific knowledge such as dedicated ontologies.
- Open-Domain Question Answering, which deals with questions about nearly everything.

Restricted domains typically have limited data available and therefore conventional techniques based on data redundancy, as the ones used in Open-Domain, can simply not be applied in an effective way. The scarcity of data available seems to prompt for a more targeted, NLP-intensive approach to QA. As a matter of fact, many are the existing systems for specific domains and work has been dedicated also for the development of legal QA, i.e. QA specifically addressing the interrogation of legal text [Saías and Quaresma, 2002].

### 3.2.7 Information Extraction

Information Extraction (IE) is about extracting specific information from documents. As such it can be a more specialised and domain specific process than IR. A number of basic technologies from NLP are utilised in order to provide the structure to obtain facts and then extract these from the text in question. Typically the input into an IE system is text (although speech is sometimes used) and the output is a structured form of data that can either be used immediately or which can be passed on for further processing in a variety of tasks and applications. In IE, the desired knowledge is described by a relatively simple and fixed template, or frame, with slots that need to be filled in with material from the text, and only a small part of the information in the text is relevant for filling in this frame, while the rest can be ignored.

The main difference between IR and IE is that in IR the user is expected to find the information they are looking for from the documents that are returned to them, whereas IE will actually provide the information they are looking for without (necessarily) providing the document.

IE uses more fine-grained analysis that requires the basic technologies of NLP and also requires a user to specify quite clearly what it is that they are looking for. The main disadvantage of such an approach is that it is a more knowledge-intensive process, is more computationally expensive and can be fairly domain specific. For ad-hoc, non-replicable searches, it is probably better to use IR and then search for the information yourself. But for replicable searches over large text collections, IE is likely to be a more efficient solution in the long term due to the reduced need for the user to spend their time consuming the information in the document.

IE applications allow to typically retrieve specific information such as the one you can look for in business news, weather reports, news about what happened today on the stock market etc. So, the sort of information that IE tends to be able to extract from texts are:

- Companies
- People
- Genes
- Attributes of documents, such as title or section headings
- Sums of money
- Relationships between some kinds of named entities
- Events, (natural disasters, elections, football matches etc.)

IE tends not to require the full gamut of NLP analysis resources in order to fulfil many of its tasks, although this depends on the exact nature of the task in hand, the type of texts being used and the sort of domain in which the extraction is likely to occur. Performance levels tend to be affected and the domain specificity becomes greater when you look for more

complex entities, relationships or even notions of an event. [Cunningham 2005].

Until now limited application-related resources have been shown to be more effective in applied systems since systems have been applied to very restricted domains. The necessity to scale up the IE technology to wider domains (e.g. search in the Net) would necessarily need the use (or re-use) of extensive resources. The main information needed in IE is mainly related to sub-categorization frames, nominalization, adjectivization. Most part of this information is not provided by WordNet-like resources thus have to be acquired from text, also in order to make the resource more tightly adhere with the domain at hand.

As a matter of fact, applicative domains often show deviations with respect to the normal use of language in terms of:

- the kind of subcategorization frame: the frame may change according to specific uses; for example the verb "indicare" in standard Italian is a normal intransitive verb, while in the financial domain has an additional argument related to the value introduced by the preposition "a" (e.g. "i titoli sono stati indicati al 2%"); also role restrictions can change;
- meaning: very often words assume additional meanings in specific domains; for example the verb "indicare" in Italian means "to point"; but in the financial domain it is used to introduce prices for listed shares;
- familiarity; for example the verb "to index" is considered rare in standard English (is not even listed in WordNet), but it is very familiar in finance and computer science.

The use of generic resources in analysing texts in restricted domains also introduce the problem of the relation between the domain description available or needed by the system (in order to reason on the extracted information; e.g. the knowledge base) and the generic lexical semantic definition given by the generic resources. Partial overlaps can be found, but the domain specific description is likely to be more precisely defined and reliable.

### 3.2.8 Fact Extraction

A particularly interesting sub-area of IE is the Fact Extraction (FE) where the facts that give evidence of some domains are extracted from textual corpora of different type. Important references for this type of application are [Siegel and McKeown, 2000], (Filatova and Hovy, 2001), (Filatova and Hatzivassiloglou, 2003).

Defining what is an event is rather difficult; nevertheless, it can be defined on the basis of the ultimate goal any application has set (relative Vs absolute definition). For example, within the Topic Detection and Tracking

framework<sup>1</sup>, event was defined as "some unique thing that happens at some point in time". In addition, one of the recent efforts in information extraction is the ACE project (<http://www ldc.upenn.edu/Projects/ACE/>), whose objective is to "develop extraction technology to support automatic processing of source language data". Thus, with respect to this objective the research efforts pertain to the detection and characterization of entities, relations and events.

Often FE modules attempt to retrieve important content from the available resources, taking advantage of multi-layer annotation comprising: i) dependency-based syntactic representations, ii) information from shallow ontologies (event types), iii) information concerning the semantic arguments of each fact. Facts are defined as the most significant events that characterize the data of a specific domain.

Definitions can be considered particular type of "facts". There is a huge work on definition identification in different areas such as Question Answering (CLEF initiative in Europe) and in some European projects like LT4eL .

Output of a FE module is a list of predicate-argument structures involving predicates of interest and their syntactic and semantic arguments.

### 3.2.9 Extraction of Temporal Information

The extraction of temporal information has become a critical component for any robust system for information retrieval and extraction or for language understanding. Temporal information in a document can be identified by two distinct kinds of data: i) the temporal metadata, regarding when the document was created, published, distributed, received, revised, etc., and ii) the temporal properties of the contents of the document. Temporal objects and relations have been studied from logical and ontological points of view [Hobbs and Pan, 2004], [Mani et al., 2005]. A number of efforts have taken place in order to develop ontologies of time, for expressing the temporal content of web pages and temporal properties of web pages and web services (ONTOLINGUA , SUMO , CYC , ONTOTEXT among others). The latest collaborative efforts is the OWL-Time [Hobbs-Pan, 2004] which covers the basic topological temporal relations on instants and intervals, measures of duration, clock and calendar units, months and years, time and duration stamps, including temporal aggregates (every morning for the last four years), deictic time (now) and vague temporal concepts (recently, soon, a little while). OWL-Time maps easily onto other temporal theories/ontologies (e.g., Cyc, SUMO), it connects with various temporal resources and supports reasoning about time. A mapping between TimeML and OWL-Time is presented in (Hobbs-Pustejovsky, 2003). Moreover, TIMEX2 [Ferro et al, 2001] can be indicated as the most used annotation scheme and TimeML (TimeML Working Group, 2005) as the de facto standard for this task (also becoming ISO standard). TimeBank (2002) is an annotated corpus for temporal expressions, events and their relations.

---

<sup>1</sup> (<http://www.nist.gov/speech/tests/tdt/index.htm>)

### 3.3 Current IR Exploitation in eParticipation

Information Retrieval is very important for Content Provision in eParticipation. Any access to the content available in the WWW starts with a query submitted to one of the many Search Engines now available. The use of Google to search the web is widespread, but there are plenty of different solutions available and exploited by users. Search Engines can be general purpose portals allowing the access to (usually very large) portions of indexed web documents. At the same time, many are the sites which, among the other information, allow the user to navigate through their content by means of an internal search engine.

Many web sites dedicate to public communication have a search engine which allow citizens to submit a query and to obtain the documents that is supposed to satisfy their needs. If we look at the templates gathered for the DEMO-net: D5.3 (Report on experiments at regional, national and EU) we see that the Search Engine, as ICT tool, is exploited by 12 of the almost 70 surveyed projects. The problem, in eGovernment and eParticipation, is that is much more difficult for users to arrive at a useful answer. When using Google for searching general purpose information, we are very likely to obtain the information we are looking for. As a matter of fact, often information searched by users in their every day life is pretty much reachable: which actress won the Oscar in 1999, where can I find a Greek Restaurant in Budapest, what's the title of Britney Spears' last song?

For this kind of queries, quite simple methods to retrieve the document are often enough to provide the answer; this for different reasons: i) the more a topic is known, the more we are going to find the information, ii) for this kind of "general", "popular" knowledge, usually redundant information is available, iii) often an easy mapping between named entities present in the query and in documents is enough to find the answer (there is not much lexical distance between the keywords of the query and of the documents).

The situation with information sensible to eParticipation is very difficult, in particular for what the access to laws, regulations and administrative issues is concerned.

The problem is that, in this cases, the gap between the form of the query and the answer tend to be larger: quite often, when interested in specific information about, for example, the set of steps needed to obtain a service, the citizen uses words in the query that will not be present in the answer; this because often the language of PA and its peculiar way to organize content do not correspond with the language of citizens: a married couple asks about adopting a "baby" while in the law we will probably find the word "minor"; at the same way, when they want to know specific information about the age they should have before submitting their dossier, rarely they will put the word "adopter" in their query, while that's the word used in the law.

But it is not only a problem of "lexical" choices; this is particularly clear when looking at the problems that would be encountered by a Question Answering system when dealing with "administrative" questions. In Open-Domain, the vast majority of the accepted questions are the so-called factual questions, (often introduced by the stems Who, When, Where, What) that usually expect as answer a Named Entity (a location, a human name, the name of a company, a measure etc.). If we think at the questions that a citizen could submit to a system, we could think that more often they will be Yes/No questions (*Does an incentive exist in the case that one young couple expects a son? Can I recover the days of the leave I did not enjoy due to illness?*) or questions expecting as answer a detailed description of a procedure (*What should I do to undertake the military career? How can I sign myself at an Italian school after having attended the school abroad?*). Answering such questions is really difficult for state-of-the-art IR techniques. We think that NLP can play an important role, in the future, to make IR progress, in particular in domains and contexts for which a bag-of-words approach cannot be enough. A reference model has to capture the main concepts in the domain and their relationships as the basis for development, analysis and management of e-government systems, processes and structures. As we say, the key word seems to be "semantics": going from the lexical, superficial level, to the conceptual one. In this sense, information extraction, summarization, QA and also the use of common and shared representations (ontologies) is something that may be of great help in the next future, in particular if connected with core technologies such as semantic web services and web 2.0 (cf. the other booklets presented within the 14.3 series).

We analyse all the projects surveyed in DEMO-net Deliverable 5.3. Few are the experiences of use of NLP to strengthen the retrieval capability of the eParticipation systems. Two examples of interesting use of NLP are the HANDS and TELE\_P@b projects. HANDS ([www.hands-online.org](http://www.hands-online.org), derived by the EDEN project) shows the advantages of using Natural Language Processing (NLP) over keyword-based question-answering techniques as a means of increasing the potential for take-up by all citizens. In HANDS, specific modules, empowered by NLP tools, help the user to find the information he is looking for: the Answer Tree, a FAQ-based QA module, which manages FAQ lists and retrieve the questions and answers most similar to the user's question; the Natural-language Map, which retrieve texts and maps through a natural-language interface [Carenini et al., 2007]. Similar instruments are also available in the DemOracolo platform [DemOracolo, 2005].

In the TELE\_P@b Project, among the different modules foreseen, we find the service for the presentation of the final budget of the municipalities involved in the project. The idea is providing citizens with the possibility to really interpret and understand the data of the final balance, with the help of information extraction procedures which allow to link, to each item in the balance, the relevant information presented in programmatic documents and in specific laws and regulations. Moreover, in TELE\_P@b, contributions coming from civic society (texts about the final balance, opinions and ideas expressed on forums and blogs by non-institutional associations and individual citizens) are analysed and processed with



techniques of IE and categorizations and can be browsed and searched by means of a Search Engine [TELE\_P@b, 2006].

Particularly interesting is the development of systems specifically addressing the interrogation of legal text (legal QA). [Saías and Quaresma, 2002] presents an interesting approach to the problem. Among the projects analysed in D5.3, the E-POWER one ([www.belastingdienst.nl/epower](http://www.belastingdienst.nl/epower)) pursues the same final aim, by developing methods and supporting tools to convert legislation and regulations into executable, formal representation.

The challenge of future research is to develop proper methods, technologies and tools to achieve objectives like these. Technologies and tools are not yet mature enough for their wide application in real contexts. Core technology and applied research is needed in these areas in order to overcome the limitations of current emerging technologies and to bring forward new concepts that better suit the needs of advanced and intelligent public services supported by information and knowledge management.

### 3.4 Text Mining

Text Mining, or Knowledge Discovery, is a method of using NLP techniques to make inferences and to obtain extra knowledge from a particular text or series of texts. It is a multidisciplinary field, with very strong links with information retrieval, text analysis, information extraction, clustering, categorisation, visualisation, database technology, machine learning, and data mining. Text Mining typically consists of two phases:

- Text refining, a pre-processing phase in which free text documents is converted into a chosen intermediate form
- Knowledge distillation that deduces patterns or knowledge from the intermediate form.


Intermediate form (IF) can be semi-structured such as the conceptual graph representation, or structured such as the relational data representation (e.g. in a form of term-document matrix). Moreover, the intermediate form can store document-based information (i.e. each entity represents a document), or concept-based information (entity represents an object or concept of interests in a specific domain).

Application areas of Text Mining can also be divided to two groups, according to the preferred mining task as well as the intermediate form adopted. First group focuses on document organisation, visualisation, and navigation. The general approach is to organise the documents based on their similarities and present the groups or clusters of the documents in certain graphical representation. Some of products belonging to this group are:

- Cartia's ThemeScope is an enterprise information mapping application that presents clusters of documents in landscape representation. [<http://www.cartia.com>]
- Canis's cMap is a document clustering and visualisation tool based on Self-Organizing Map. [<http://www.canis.uiuc.edu/projects/interspace/highlight.html>]
- IBM's Technology Watch is a text mining application in the scientific domain. It performs document clustering plus visualisation in the form of maps for patent databases and technical publications. [<http://www.synthema.it/tewat/maine.htm>]
- VizControls is a visualisation tool that performs value-added post-processing of search results by clustering the documents into groups and displaying based on a hyperbolic tree representation. [<http://www.dcc.uchile.cl/~rbaeza/cursos/visual/ix/index.html>]
- Entrieva's SemioMap employs a three-dimensional graphical interface that maps the links between concepts in the document collection. The SemioMap is concept-based in the sense that it explores the relationships between concepts whereas most other visualisation tools are document-based. [<http://www.entrieva.com/entrieva/html%5Fsite/semiomap.htm>]

Second group of Text Mining applications, mainly based on NLP techniques, focuses on text analysis functions, notably, information retrieval, information extraction, categorisation, and summarisation. Some of applications are listed here:

- Inxight's LinguistX is a platform that provides advanced text analysis capabilities in several languages, making it the solution of choice for search engines, data mining applications, indexing applications, and text categorisation and routing tools. [<http://www.inxight.com/products/sdks/lx/>]
- IBM's Intelligent Miner is probably one of the most comprehensive text mining products around. It offers a set of text analysis tools, including a feature extraction tool, a set of clustering tools, a summarisation tool, and a categorisation tool. Also incorporated are the IBM's text search engine, NetQuestion Solution and the IBM web crawler package. [<http://www-306.ibm.com/software/data/iminer/>]
- TextWise, an R&D company based in Syracuse University, offers various text mining products: DR-LINK is an information retrieval system based on automatic concept expansion; CINDOR is its cross lingual version; CHESS is a text analysis and information extraction tool. [<http://www.textwiselabs.com>]
- Data Junction's Cambio is an information extraction tool that extracts data in the form of relational attributes from text. [[http://www.datajunction.com/products/cambio\\_technical.html](http://www.datajunction.com/products/cambio_technical.html)]
- Megaputer's TextAnalyst uses a semantic net representation of documents and performs automated indexing, topic assignment, text abstraction, and semantic search. [<http://www.megaputer.com/products/tm.php3>]



In addition to the tools and applications mentioned above, there exists a broad suite of standards, toolkits, environments, and platforms for Text Mining. Most of them are online applications or freely available open source projects, e.g. GATE, UIMA standard, YALE, Bow, Jbowl, Topicalizer, Textalyzer, etc.

Text Mining techniques are nowadays applied in many areas, most notably in the security, commercial, and academic fields – in all the areas in which highly specific information is often contained within written text. In the WWW, Text mining will enable searches that can be directly answered by the semantic web. Text mining is also the technique used for fighting email spam.

### 3.4.1 Opinion Mining

Perhaps one of the most exciting area in NLP is that of opinion or sentiment classification. Sentiment and subjectivity in text constitute a problem that is orthogonal to typical topic detection tasks in text classification. Despite the lack of a precise definition of sentiment or subjectivity, headway has been made in matching human judgments by automatic means. Such systems can prove useful in a variety of contexts. In many applications it is important to distinguish what an author is talking about from his or her subjective stance towards the topic. If the writing is highly subjective, as for example in an editorial text or comment, the text should be treated differently than if it were a mostly objective presentation of facts, as for example in a newswire. Information extraction, summarization, and question answering can benefit from an accurate separation of subjective content from objective content. Furthermore, the particular sentiment expressed by an author towards a topic is important for "opinion mining", i.e. the extraction of prevalent opinions about topics or items from a collection of texts. It cuts across many applications such as text classification, information retrieval, information extraction, text mining etc. It covers several topics including the learning of semantic orientation of terms, sentiment analysis, opinion and attitude analysis etc. In this context, various approaches have been introduced: [Pang et al., 2002], [Turney, 2002], [Wiebe et al., 1999], [Riloff et al., 2003], [Yu and Hatzivassiloglou, 2003], [Kim and Hovy, 2006]. Standard machine-learning classification techniques, implementing text-categorization techniques such as SVMs can be applied to the entire documents. One of the most recent efforts in opinion mining systems evaluation is TREC 2006 Blog Opinion Mining task . In order to support the development of relevant systems, various lexical resources are also being developed such as: MPQA corpus of opinion annotations , Senti-WordNet , General Inquirer , etc.

Over the past several years, there has been an increasing number of publications focused on the detection and classification of sentiment and subjectivity in text, which feeds into the area of individual differences and the reliability of particular pieces of information, aspects which have

hitherto been extremely difficult to use computational techniques to analyse. Key areas of interest of are:

- the attempt to determine the “strength” of a particular sentiment,
- the attempt to automatically identify bias, opinion or subjectivity in a text by automated or semi-automated means,
- tracking opinion about a particular topic over time (such as citizen’s attitudes to a particular product or a to a particular person) time
- automated analysis of survey responses
- summarisation of the opinions of various different people about a particular topic

### 3.4.2 Ontology Acquisition

In DEMO-net Deliverable D5.2, we read, about “information and knowledge extraction”:

*This area includes the use of language techniques in combination with ontologies for the clear interpretation of public opinion as it is captured e.g. in discussion forums, blogs, wikis or other forms of common discussion spaces. [...] Ontologies will offer a common understanding of the heterogeneous data and through the implementation of knowledge extraction and statistical techniques upon the data, we will extract meaningful messages that will represent to the greater extent public opinion and serious concerns or strong arguments in the public life.*

Moreover:

*A lot of research in the field of ontologies and knowledge and information extraction may be exploited in order to build the appropriate ontologies and design discourse analysis techniques for analysing large-scale information sources of political discourse. This type of research will be significantly useful in the promotion of a democracy of a superior quality.*

These premises allow us to understand that there is a very strong link between NLP and Ontologies, which is based on the exploitation of NLP to mine textual materials in order to support the creation, maintenance and population of common representational frameworks. Part of D5.1 was dedicated to Ontologies and Knowledge Management, while in this booklet we want to introduce the more NLP-based field known as Ontology Acquisition, Building or Population. It is a very active research field, as witnessed by the increasing number of Knowledge Management applications based on automated routines for ontology navigation and update. Electronic texts still represent the most accessible and natural repositories of specialised information worldwide and different methodologies have been proposed to automatically extract information from them and provide a structured organisation of extracted knowledge in as diverse domains/sectors as bio-informatics, health-care, public administration and company document bases. The situation in the legal domain, particularly important in view of exploitation in the eParticipation

and eGovernment sectors, is in line with this general trend and probably made even more critical by the fact that laws are invariably conveyed through natural language.

The last few years have seen a growing body of research and practice in constructing legal ontologies and applying them to the law domain. A number of legal ontologies have been proposed in different research projects: yet, most of them focus on a upper level of concepts and were mostly hand-crafted by domain experts (for a survey of legal ontologies, see [Valente, 2005]). It goes without saying that realistically large knowledge-based applications in the legal domain need more comprehensive ontologies incorporating up-to-date knowledge: ontology-learning from texts could be of some help in this direction.

Relatively few attempts have been made so far to automatically induce legal domain ontologies from texts: this is the case, for instance, of [Lame, 2005], [Saias and Quaresma, 2005] and [Walter and Pinkal, 2006], [Lenci et al., 2007]. In [Lenci et al., 2007], an approach to Ontology Acquisition is described by introducing the T2K (Text-to-Knowledge) tool (jointly designed and developed by the Institute of Computational Linguistics (CNR) and the Department of Linguistics of the University of Pisa). The system offers a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine language learning, which are dynamically integrated to provide an accurate representation of the content of vast repositories of unstructured documents in technical domains. Text interpretation ranges from acquisition of lexical and terminological resources, to advanced syntax and ontological/conceptual mapping. Interpretation results are annotated as XML metadata, thus offering the further bonus of a growing interoperability with automated content management systems for personalised knowledge profiling. Prototype versions of T2K are currently running on public administration portals and have been used for indexing E-learning and E-commerce materials.

### **3.4.3 Text Mining Current Exploitation in eParticipation**

Text Mining is already applied on a variety of different sectors and fields. A typical example of a well investigated area is the domain of bio-medicine, for which the work of the NacTem research centre (<http://www.nactem.ac.uk/>) represents an important reference. An interesting application on the domain of social sciences, is the work carried out within the ASSERT project (<http://www.nactem.ac.uk/assert/projectSummary.php>), dedicated to developed Text Mining tools for the analysis and summarization of systematic reviews.

Under the 5th and 6th Framework Program of the Information Society Technologies (IST), the European Commission has funded a number of projects to advance eGovernment developments in Europe. Nowadays, the focus of current research is on data and knowledge management and interoperability. The main knowledge-oriented research areas that emerge

from the aforementioned projects include complex knowledge mining, flexibility and adaptation capability, in particular via search, retrieval and exploitation of heterogeneous and fragmented sources of knowledge within public administrations, using semantic annotation and mining tools at real time. Recent European Research projects in Semantically Enriched, Knowledge-Based eGovernment and eBusiness Systems include:

**Access eGov** - Access to e-Government Services Employing Semantic Technologies

**SAKE**: Semantic-enabled Agile Knowledge-based e-Government

**FIT**: Self-adaptive e-government Service Improvement with Semantic Technologies

**SemanticGov** - Providing Integrated Public Services to Citizens at the National and Pan-European level with the use of Emerging Semantic Web Technologies


### 3.5 Language resources and tools infrastructure

In the conclusions of DEMO-net Deliverable D5.1 (Report on current ICTs to enable Participation) we read:

*[...] a reflection of preconditions for a successful deployment has been given. Such aspects are integration, interoperability [...].*

*In order to realize all the phases of a public dialogue, from information to decision making and evaluation, a mixture of tools is required. [...] Comprehensive and powerful eParticipation solutions require also smooth integration with administrative solutions and contexts, interoperability among heterogeneous systems and automated information and workflow processing. Unfortunately, investigations in interoperability and standards for eParticipation are not yet addressed extensively in research and practice.*

Key notions are thus interoperability, integration, common representations, standardization. These aspects have been the focus of an entire research thread of the Language Resources field since mid 80s. Many have been the actions undertaken by the NLP and LR communities, often EU funded, which have aimed at defining standard representation for language resources and tools. Among the others, we remember EAGLES, ISLE etc. The last initiative in this sense is CLARIN, which can be seen as an ultimate, long-term action aimed at creating an infrastructure that makes language resources and technology available and readily usable to scholars of all disciplines, in particular, the humanities and social sciences. Linguists and computer scientists have created a large and invaluable set of resources such as annotated recordings, texts, lexica and ontologies and tools such as speech recognizers, lemmatisers, parsers and summarizers and information extractors during the last decades, yet



almost nothing fits together and only a small selection can be used easily by non-experts to the benefits of their research. Smooth access to ready-to-use language resources, language technologies and appropriate advice on how to apply them is of vital importance in an era where the broad accessibility to language resources and technologies becomes essential to master the information explosion and where we face increasing challenges presented by multicultural and multilingual societies such as in Europe. Therefore, CLARIN will focus on making the existing components available through the creation of a Europe-wide federation of archives and repositories, which will take the shape of network of stable centres offering highly available services to everyone via secure networks. European researchers have always been at the forefront of standardization and infrastructure building efforts on the international scene. CLARIN will maintain the leading position of European researchers and will build on what has already been achieved with the help of former national and European initiatives. A new type of advanced research infrastructure will allow eParticipation stakeholders to use and exploit NLP technologies.

## 4 Conclusion: participation areas and technological challenges

In the previous chapter we presented an overview of the final applications that are more used in eParticipation or that can be more interesting as emerging technologies. In this final section we discuss the situation of actual and current exploitation of NLP in systems for eParticipation and we show that Human Language Technologies appear not to be very much used in this field. Nevertheless, we foresee the great potentiality of these technologies, in more than one area of Participation and in many of the citizen-centric services we imagine for future development. By analysing the "Participation" areas described in the DEMO-net deliverable 5.1, it is possible to circumscribe some fields that could be more impacted by NLP. They are<sup>2</sup>:

Information Provision	ICT to structure, represent and manage information in participation contexts
Deliberation	ICT to support virtual, small and large-group discussions, allowing reflection and consideration of issues
Discourse	ICT to support analysis and representation of discourse
Mediation	ICT to resolve disputes or conflicts in an online context
Polling	ICT to measure public opinion and sentiment


The analysis of this list allows us to highlight a set of very general needs that cut across the different areas and to which NLP can provide an answer. We present these core needs and introduce the NLP applications which can satisfy them. It is important to always keep in mind that the nature of "participation" varies greatly: citizens participate by expressing opinions in a public forum, by providing contributes to deliberative democratic processes, but also by benefiting from a public service in an active way (a typical eGovernment topic). All these different situations determine different communication models, which change the way NLP can contribute to the technological innovation.

### 4.1 Narrowing the Language Gap

Is language a barrier to participation? Of course it is. First of all, there is the need to narrow the gap between the language expressed by PA and the language of citizens. We know that most citizens do not have special confidence with any of the formalized languages adopted in dealing with legal and administrative activities. Moreover, also the "horizontal" dialogue among citizens needs encouragement, since people often do not

---

<sup>2</sup> We refer to the list of Participation Areas presented in (Thorleifsdottir and Wimmer, 2006)



share a common parlance or do not have equal skills in speaking and using a non-native language.

This aspect is particularly important in Content Provision, where NLP may be exploited to improve the performance of **Information Retrieval** and **Question Answering** systems but also to classify and categorize content by means of **Document and Text Classification** systems which may help navigation and access. Moreover, NLP can also help professionals who created laws, regulations and web pages with public information to use a language that is closer to the one used and understood by people. The two examples show two different ways to overcome the problem of language gap: in the first case, technology is something which that automatically draws up the content expressed in PA documents and the informative need of the citizen by working at the level of conceptual and semantic representation and classification. In the second case, NLP drives the modality of content creation by suggesting how to write in an understandable way (by using, for example, instruments showing the overall **readability** of the written texts). NLP is also very important for all the aspects concerning **conversational technologies** and **vocal interfaces**, which can sustain human-computer interaction and an effective content provision, also for granting access to many disabled people, especially those who are partially or totally unable to see or read.

## 4.2 Multilingualism

A particular, structural source of language diversity is obviously multilingualism. In the business field, English has become the dominant language within global organizations and it is the de-facto answer to the language barrier problem. It is obvious that such a response cannot be considered a solution for eGovernment and eParticipation, since it cannot be chosen as a means to overcome exclusion of citizens from the democratic arena.

Effective Content Provision is heavily influenced by the actual possibility for European citizens to access content expressed in languages different from their idiom. Language can be a real barrier for democratic transparency. Some may object that, at in the end, politics is mainly local and that the scope of democracy is often not larger than a country. This is not true, of course, in contemporary Europe, where super-national policies, legislation, discussions are becoming everyday more and more important. Moreover, also restricting the scope of the problem at a local level, multilingualism has a strong impact on *inclusion* (which is based on the possibility to reach everyone), in particular when the communication involves immigrant citizens or residents. If Content Provision is the area most affected by this kind of problem, also Mediation and Deliberation, which concern opinion forming, are interested. With this premises, it is obvious that the availability of **Machine Translation** systems would be important to develop actual multilingual services. Also **Information Extraction, Information Retrieval, Summarization** and **Question Answering** can be "declined" in a cross-language way, allowing systems to deal with different languages at the same time.

### 4.3 “Going straight to the point”

Another important and somehow “transversal” issue is represented by the necessity to derive information from the bulk of data available on the Internet, in order to sustain the formation of opinions and ideas. The idea is that who consults the web prefers not spending too much time in reading very long documents, blog and forum entries to form an opinion on a specific subject. The key-word is “going straight to the point”: what is the salient information, which are the positions in the field and what are their mutual differences? This issue affects Deliberation and Mediation and can be dealt with by recurring to **Information Extraction** and **Summarization** systems. Information Extraction is very promising also in view of the analysis and organization of the content of the entries available in Wikipedia (or other sites based on Wiki software). The ambitious vision of eParticipation relies on the possibility to capture semantics embedded in documents and other content objects by means of innovative information extraction and knowledge mining tools. The ultimate aim is reaching an effective content provision and to overcome the problems that hamper it.

### 4.4 Retrieving Trends, Opinions and Sentiments

A very sensitive issue is the necessity, for citizens but also for professionals of politics and consensus formation, to measure, in an easy and fast way, public opinions and sentiments on specific subjects in an easy and fast way. This need has been ignited by the availability of huge quantity of spontaneous information, constantly changing and dislocated on a great number of web sites (institutional sites, on-line newspapers, non-official blogs, forums etc.). Opinions and sentiments detection mainly concerns the Participation area of Polling but affects also other fields, such as Deliberation and Mediation. Current Information Retrieval Systems are not suited to work on opinions and subjective expressions, since they are designed to retrieve *facts* more than *ideas*. Search Engine and Question Answering systems are keyword-based, while opinions can hardly be expressed by keywords; moreover, the ranking strategy itself is not something suitable for searching opinions. For this reason, in the last few years, ad-hoc techniques for **Opinion Mining** have been developed.

In general, however, there is the need to know salient facts and features, hidden in very large quantity of data, which stand out for their frequency: this allows deriving interesting and constantly updated information about trends, tendencies, most important topics in a given period etc. For this kind of exigencies, **Text Mining** techniques are very useful and promising.

### 4.5 NLP technologies in the Demo-Net D5.3 questionnaire

Even if NLP technologies seem very promising for the field of eParticipation, only few projects actually use them. We carried out an

analysis of the “Report on experiments at regional, national and EU level” (Demo-Net Deliverable 5.3)<sup>3</sup> to highlight the projects and initiatives making any use of NLP technologies, and discovered that: on among almost 70 European projects which have been surveyed, only eight of them state to exploit NLP technologies (and, quite surprisingly, many of them are Italian funded projects). They are listed below:

<b>Project Name</b>	<b>Period</b>	<b>Funding</b>
partecipaPUG	2006-07	Italian
Concerto	2006-07	Italian
DemOracolo	2005-07	Italian
Telep@b	2006-08	Italian
EDEN	2001-03	EU
Safir	2003-04	EU
E-POWER	2001-03	EU
HANDS	2005-07	EU

The main interest in NLP is motivated by the possibility and necessity to move, in eGovernment, towards semantic analysis and conceptual representation and retrieval of information. This same objective is what can be found in many EU funded projects, not surveyed in D55.3, which aims at a semantically enriched, knowledge-based eGovernment. These projects are: Access eGov, SAKE, FIT and SemanticGov.

The scarceness of actual prototypes incorporating NLP solutions and the interest in semantics and knowledge-based approaches gives us a quite precise measure of how much we need interdisciplinary research to stimulate the creation of a real common ground of competences and expertises to develop new and more advanced tools to strengthen and enlarge citizenship.

---

<sup>3</sup> We refer here to a preliminary version of the deliverable.

## References

Abney, S., (1996) Tagging and Partial Parsing. In: Ken Church, Steve Young, and Gerrit Bloothoof (eds.), *Corpus-Based Methods in Language and Speech*. Kluwer Academic Publishers, Dordrecht.

[Antoni-Lay et al., 1994]

Antony-Lay M.E., Francopoulo, G., Zaysser, L., (1994) A Generic Model for Reuseable Lexicons: The Genelex Project. *Literary and Linguistic Computing* 1994 9(1), 47-54.

[Bartolini et al., 2002]

Bartolini, R., Lenci, A., Montemagni, S., and Pirrelli, V., (2002) Grammar and Lexicon in the Robust Parsing of Italian: Towards a Non-Naïve Interplay. In *Proceedings of COLING 2002 Workshop on Grammar Engineering and Evaluation*, Taipei, Taiwan.

[Baker et al, 1998]

Baker, C.F., Fillmore, C.J., Lowe, J.B., (1998) The Berkeley FrameNet Project, in *Coling-ACL 1998: Proceedings of the Conference*, 86-90.

[Belew, 2000]

Belew, R.K., (2000) *Finding Out About: Search Engine Technology from a cognitive Perspective*. Cambridge University Press.

[Brill, 1995]

Brill, E., (1995) Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. In *Computational Linguistics*, 21(4), 543-566.

[Calzolari et al., 2001]

Calzolari, N., Grishman, R., Palmer, M. (eds.) (2001) Survey of major approaches towards Bilingual/Multilingual Lexicons, ISLE Deliverable 2.1-3.1, WP2-3.

[Calzolari et al., 2003]

Calzolari, N., Bertagna, F., Lenci, A., Monachini M. (eds.) (2003) Standards and Best Practice for Multilingual Computational Lexicons, MILE (the Multilingual ISLE Lexical Entry), ISLE Deliverable 2.2 & 3.2 WP2-3.

[Carenini et al., 2007]

Carenini, M., Whyte, A., Bertorello, L., Vanocchi, M., (2007) Improving Communication in E-democracy Using Natural Language Processing. In IEEE Intelligent Systems Issue pp. 20-27

[Charniak, 1997]

Charniak, E., (1997) Statistical parsing with a context-free grammar and word statistics. In Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI '97), 598-603.

[Cunningham, 2005]

Cunningham, H., (2005) Information Extraction, Automatic. In Keith Brown (ed.), Encyclopedia of Language and Linguistics, vol. 1-14, 2nd Edition, Elsevier Science Publishers, 665-677.

[Daille et al., 1994]

Daille, B., Gaussier, É., Langé, J.-M., (1994) Towards Automatic Extraction of Monolingual and Bilingual Terminology. In: Proceedings of COLING 94. 515-521.

[Demoracolo, 2005]

DemOracolo, Project Description, 2005. Available at [http://portale.comune.verona.it/.../allegati/Demoracolo\\_sintesi.pdf&title=La%20sintesi%20del%](http://portale.comune.verona.it/.../allegati/Demoracolo_sintesi.pdf&title=La%20sintesi%20del%20)

[Dorr et al., 1998]

Dorr, B. J., Jordan, P. W., Benoit, J.W., (1998) Surveys of current Paradigms in Machine Translation. Technical Report: LAMP-TR-027, UMIACS-TR-98-72, CS-TR-3961, University of Maryland, College Park.

[Dorr, 1994]

Dorr B.J., Machine Translation: A view form the Lexicon, Cambridge, MA: The MIT Press, 1994.

[Elhadad, 1992]

Elhadad, M., (1992) Using Argumentation to Control Lexical Choice: a Funtional Unification-Based Approach. Ph.D. Thesis, Computer Science Department, Columbia University,

[Enguehard & Pantera, 1994]

Enguehard, C. & Pantera P., (1994) automatic manual acquisition of terminology. In *Journal of Quantitative Linguistics*, 2(1), 27-32.

[Fellbaum, 1998]

Fellbaum, C. (ed.) (1998) *WordNet: An Electronic Lexical Database and Some of its Applications*, MIT Press.

[Ferro et al., 2001]

Ferro, L., Mani, I., Sundheim, B., and Wilson, G., (2001) *TIDES Temporal Annotation Guidelines Draft - Version 1.02*. MITRE Tech. Rep. MTR 01W000004. The MITRE Corp., McLean, VA.

[Filatova & Hatzivassiloglou, 2003]

Filatova, E., Hatzivassiloglou, V., (2003) *Domain-Independent Detection, Extraction, and Labeling of Atomic Events* In *Proceedings of RANLP*, Borovetz, Bulgaria.

[Filatova & Hovy, 2001]

Filatova, E., Hovy, E.H., (2001) *Assigning Time-Stamps to Event-Clauses*. In *Proceedings of ACL Workshop on Temporal and Spatial Reasoning at the Conference*. Toulouse.

[Fontenelle, 1997]

Fontenelle, T., (1997) *Turning a bilingual dictionary into a lexical-semantic database*, Max Niemeyer Verlag, *Lexicographica Series Maior 79*, Tübingen.

[Francopoulo et al., 2006]

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C. (2006). *Lexical Markup Framework (LMF)*. *Proceedings of LREC2006*, Genoa, Italy.

[Garside 1993]

Garside, R., (1993) *The Marking of Cohesive Relationships: Tools for the Construction of a Large Bank of Anaphoric Data*. *ICAME Journal* 17, 5-27.

[Godfrey and Zampolli, 1997]

Godfrey, J.J., Zampolli, A. (1997) Language Resources. In A. Zampolli, G.B. Varile (Managing Editors), Survey of the State of the Art in Human Language Technology. *Linguistica Computazionale*, Cambridge University Press, 381-384.

[Green et al., 1961]

Green, B. F., Wolf, A. K., Chomsky, C., Laughery, K., (1961) BASEBALL: an Automatic Question Answerer. In Proceedings of the Western Joint Computer Conference.

[Grishman and Calzolari, 1997]

Grishman, R., Calzolari, N., *Lexicons* (1997) Chapter on Language Resources. In R. Cole et al. (eds.) Survey of the State of the Art in Human Language Technology, Cambridge University Press.

[Hirschman and Gaizauskas, 2001]

Hirschman, L., Gaizauskas, R., (2001) Natural Language Question Answering: The View from Here. In *Natural Language Engineering* 7(4).

[Hirschmann, 1992]

Hirschmann, L., (1992) Multi-site data collection for a spoken language corpus. In Proceedings of the Fifth DARPA Speech and Natural Language Workshop. Defence Advanced Research Projects Agency, Morgan Kaufmann.

[Hobbs and Pan, 2006]

Hobbs, J.R., Pan, F., (2006) Time Ontology in OWL. *Ontology Engineering Patterns Task Force of the Semantic Web Best Practices and Deployment Working Group*, World Wide Web Consortium (W3C) notes.

[Hobbs and Pustejovsky, 2003]

Hobbs, J., Pustejovsky, J. (2003) Annotating and reasoning about time and events. Proc. Of AAI Spring Symposium on Logical Formalizations of Common Sense Reasoning, Stanford, CA.

[Hovy, 1988]

Hovy, E.H., (1988) *Generating Natural Language under Pragmatic constraints*. Lawrence Erlbaum, Hillsdale, New Jersey.

[Hull, 1996]

Hull, D., (1996) Stemming algorithms—a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1) pp. 70–84.

[Ide et al., 1998]

Ide, N., Greenstein, D., Vossen, P. (eds.) (1998) Special Issue on EuroWordNet, in *Computers and the Humanities*, Volume 32, Nos. 2-3 1998, Kluwer Academic Publishers, Dordrecht.

[Ishida, 2006]

Ishida, T., (2006) Language Grid: An Infrastructure for Intercultural Collaboration. *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pp. 96-100.

[Johansson et al., 1978]

Johansson, S., Leech, G.N., Goodluck, H. (1978), *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*, Department of English, University of Oslo.

[Juraksy & Martin, 2000]

Juraksy, D. and Martin, J., (2000) *Speech and Language Processing – An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall.

[Justeson and Katz, 1995]

Justeson, J.J., Katz M., (1995) Technical terminology: some linguistic properties and an algorithm for identification in text. In *Natural Language Engineering Vol.1(1)*, 9-27.

[Kim and Hovy, 2006]

Kim, S.M. and E.H. Hovy, E.H., (2006) Identifying and Analyzing Judgment Opinions. Full paper. *Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006)*. New York, NY.

[Klein and Manning, 2003]

Klein, D., and Manning, C. D, (2003) Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

[Kucera and Francis, 1967]

Kučera, H., Francis, W.N., (1967) Computational Analysis of Present-day American English, Brown University Press.

[Lame, 2005]

Lame, G., (2005) Using NLP techniques to identify legal ontology components: concepts and relations. Lectures Notes in Computer Science, Vol., 3369, pp. 169-184.

[Lenat, 1995]

Lenat, D.B., (1995) CYC: A Large-Scale Investment in Knowledge Infrastructure, Communication of the ACM.

[Lenci et al., 2000]

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowsky, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., and Zampolli, A., (2000) SIMPLE: A General Framework for the Development of Multilingual Lexicons, in International Journal of Lexicography, 13 (4), 249-263.

[Lenci et al., 2001]

Lenci, A., Montemagni, S., Pirrelli, V., (2001) CHUNK-IT. An Italian Shallow Parser for Robust Syntactic Annotation. In Linguistica Computazionale, Istituti Editoriali e Poligrafici Internazionali, Pisa-Roma, ISSN 0392-6907.

[Lenci et al., 2007]

Lenci, A., Montemagni, S., Venturi, G., (2007) NLP-based ontology learning from legal texts. A case study. In Proceedings of LOAIT 07, Workshop on Legal Ontologies and Artificial Intelligence Techniques, pp.113-129.

[Leonard, 1984].

Leonard, R.G., (1984) A database for speaker-independent digit recognition. In Proceedings of the 1984 International Conference on Acoustic, Speech, and Signal Processing, Volume 3, Institute of Electrical and Electronic Engineers.

[Lewis, 1991]

Lewis, D., (1991) Learning in intelligent information retrieval. In Proceedings of the Eighth International Workshop on Machine Learning, pages 235-239.

[Liddy and Myaeng, 1994]

Liddy, E., and Myaeng, S. H., (1994) DR-LINK: A System Update for TREC2. In Harman, Donna K. (editor). The Second Text REtrieval Conference (TREC-2).NIST Special Publication 500-215, National Institute of Standards and Technology, Gaithersburg, MD.

[Lin, 2001]

Lin, D., (2001) Latat: Language and Text Analysis Tools. In Proceedings of Human Language Technology Conference, CA, USA.

[Luk, 95]

Luk, A., (1995) Statistical sense disambiguation with relatively small corpora using dictionary definitions. In Proceedings of the 33rd Meetings of the Association for Computational Linguistics (ACL-95), 181-188, Cambridge, M.A.

[Macleod et al., 1998]

Macleod, C., Grishman, R., Meyers, A., Barrett, L., Reeves, R., (1998) NOMLEX: A Lexicon of Nominalizations, Proceedings of EURALEX'98, Liege, Belgium.

[Mani and Maybury, 1999]

Mani, I., Maybury M., (eds.) (1999) Advances in Automatic Text Summarization. Computational Linguistics, Cambridge, MA: The MIT Press, 1999,

[Mani et al., 2005]

Mani, I., Pustejovsky, J., Gaizauskas, R., (eds.) (2005) The Language of Time. Oxford Univ. Press.

[Matthiessen, 1983]

Matthiessen C.M.I.M., (1983) Systemic grammar in Computation: the Nigel case. In Proceedings of the First Conference of the European Chapter of the Association for Computational Linguistics, Pisa, Italy.

[McDonald, 1980]

McDonald, D.D., (1980) Natural Language Production as a Process of Decision Making Under constraint. PhD Thesis, Department of Computer Science and Electrical Engineering, Massachusetts Institute of Technology.

[Mooers, 1950]

Mooers, C., (1950) Information retrieval viewed as temporal signalling. Proceedings of the International Conference of Mathematicians, Cambridge, US, pp.572-573.

[Monz, 2003]

Monz, C., (2003) Theoretical and Practical approaches to Question Answering, in "From document Retrieval to Question Answering", PhD thesis, ILLC, U. of Amsterdam.

[Moore, 1989]

Moore J.D., (1989) A reactive Approach to Explanation in Expert and Advice-giving Systems. PhD Thesis, University of California at Los angeles.

[Pang et al., 2002]

Pang, B., Lee, L., Vaithyanathan S., (2002) Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10.

[Porter, 1980]

Porter, M., (1980) *An algorithm for suffix stripping*. Program, 14(3), pp 130-137.

[Price, 1990]

Price, P., (1990) Evaluation of spoken language systems: The ATIS Domain. In Proceedings of the Third DARPA Speech and Natural Language Workshop, Hidden Valley, Pennsylvania. Defence Advanced Research Projects Agency, Morgan Kaufmann.

[Saias and Quaresma, 2002]

Saias, J., and Quaresma, P., (2002) Semantic Enrichment of a Web Legal Information Retrieval System. In T.J.M. Bench-Capon, A. Daskalopulu and R.G.F. Winkels (eds.), Legal Knowledge and Information Systems. Jurix 2002: The Fifteenth Annual Conference. Amsterdam: IOS Press, pp. 11-19.

[Saias and Quaresma, 2005]

Saias, J., and Quaresma, P., (2005) A Methodology to Create Legal Ontologies in a Logic Programming Based Web Information Retrieval System. Lecture Notes in Computer Science, Vol. 3369, pp 185-200.

[Sanfilippo et al., 1999]

Sanfilippo, A. et al. (1999). EAGLES Recommendations on Semantic Encoding. <http://www.ilc.pi.cnr.it/EAGLES96/rep2>

[Schmid, 1994]

Schmid, H., (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of the International Conference on New Methods in Language Processing, UK.

[Sebastiani, 2002]

Sebastiani, F., (2002) Machine learning in automated text categorization. ACM Computing Surveys, 34(1), pp. 1-47.

[Siegal and McKeown, 2000]

Siegel, E.V., McKeown, K., (2000) Learning Methods to Combine Linguistic Indicators: Improving Aspectual Classification and Revealing Linguistic Insights. Computational Linguistics 26(4): 595-627.

[Simmons, 1965]

Simmons, R. F., (1965) Answering English Questions by computers: A Survey. In Communications of the ACM, 8(1).

[TELE\_P@b, 2006]

TELE\_P@b Project description. Available at: [www.un-cemtoscana.it/upload/TELE\\_P@B-documento%20di%20progetto%20rev2\\_54D.pdf](http://www.un-cemtoscana.it/upload/TELE_P@B-documento%20di%20progetto%20rev2_54D.pdf)

[Thorleifsdottir and Wimmer, 2006]

Thorleifsdottir, A., and Wimmer, M., (eds.) Report on current ICTs to enable Participation. DEMO-net: Deliverable 5.1, January 2006

[Toutanova et al, 2003]

Toutanova, K., Klein, D., Manning, C., Singer, Y., (2003) Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, 252-259.

[Turney, 2002]

Turney, P. D., (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings 40th

Annual Meeting of the Association for Computational Linguistics (ACL'02), 417-424.

[Uszkoreit, 1997]

Uszkoreit, H., (1997) Language Generation. Book Chapter in R. Cole et al. (eds.) Survey of the State of the Art in Human Language Technology, Cambridge University Press.

[Valente, 2005]

Valente, A., (2005) Types and Roles of Legal Ontologies. Lecture Notes in Computer Science, Vol. 3369, pp. 65-76.

[Wiebe and Riloff, 2003]

Wiebe, J., Riloff, E., and Wilson, T. (2003) "Learning Subjective Nouns Using Extraction Pattern Bootstrapping", Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)..

[Wiebe and Riloff, 2005]

Wiebe, J. and Riloff, E. (2005) "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts", Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-05), Invited Paper, Springer LNCS Vol. 3406 © Springer-Verlag .

[Wiebe et al., 1999]

Wiebe, J., Bruce, R. F., O'Hara, T. P., (1999) Development and use of a gold-standard data set for subjectivity classifications." In Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99).


[Winograd, 1977]

Winograd, T., (1977) Five Lectures on Artificial Intelligence. In A. Zampolli (ed.) Fundamental Studies in Computer Science. North Holland. 5, 399-520.

[Yarowsky, 1995]

Yarowsky, D., (1995) Unsupervised word-sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95), 189-196, Cambridge, MA.

[Yu and Hatzivassiloglou, 2003]



Yu, H., and Hatzivassiloglou, V., (2003) Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of EMNLP'03, 2003.

[Zampolli and Calzolari ,1994]

Zampolli, A., Calzolari, N., (1994). Towards Sharable Linguistic Resources for Language Engineering in Europe. In N. Calzolari, C. Guo (eds.), International Workshop on Directions of Lexical Research, Proceedings of the PostColing'94. Tsinghua University, (Beijing), China, 112-121

[Walter and Pinkal, 2006]

Walter, S., and Pinkal, M., (2006) Automatic Extraction of definition from German court decisions. In Proceedings of the COLING-2006 Workshop on Information Extraction Beyond the Document, pp. 20-28.

[Woods, 1973]

Woods, W., (1973) Progress in Natural Language Understanding - an Application to Lunar Geology. In AFIPS Conference Proceedings.

### ***The Demo-Net Consortium consists of:***

■ County of North Jutland - Digital North Denmark	Coordinator	Denmark
■ University of Leeds	Coordinator	United Kingdom
■ Örebro University	Partner	Sweden
■ University of Koblenz-Landau	Partner	Germany
■ Fraunhofer Gesellschaft zur Förderung der angewandten Forschung e.V.	Partner	Germany
■ Institut für Informationsmanagement Bremen GmbH	Partner	Germany
■ University of Macedonia	Partner	Greece
■ Institute of Communication and Computer Systems	Partner	Greece
■ Copenhagen Business School	Partner	Denmark
■ Aalborg University	Partner	Denmark
■ Fondation National des Sciences Politiques	Partner	France
■ Technical University of Košice	Partner	Slovakia
■ Consiglio Nazionale delle Ricerche	Partner	Italy
■ University of Bergamo	Partner	Italy
■ Yorkshire and Humber Assembly	Partner	United Kingdom
■ European Projects and Management Agency (EPMA)	Partner	Czech Republic
■ Napier University	Partner	United Kingdom
■ University of Iceland	Partner	Iceland
■ University of Helsinki	Partner	Finland
■ Institute of Technology Assessment, Austrian Academy of Sciences (ITA)	Partner	Austria
■ University of Southern California	Partner	U.S.A.

### ***DEMO-net contact information:***

■ **Strategic Research Coordinator: Professor Ann Macintosh,**

University of Leeds, Tel.: +44 (0) 113 343 5806, E-mail: A.Macintosh@leeds.ac.uk

■ **Dissemination Leader: Dr. Efthimios Tambouris**

University of Macedonia, Tel.: +30 2310 464 160 (167), E-mail: tambouris@uom.gr